



University of
Zurich^{UZH}

Exploring Important Factors of Crowdsourcing Data Science Projects

Master Thesis May 3, 2016

Frida Juldashewa

Student-ID: 10-739-217
juldashewa.frida@gmail.com

Advisor: **Michael Feldman**

Prof. Abraham Bernstein, PhD
Institut für Informatik
Universität Zürich
<http://www.ifi.uzh.ch/ddis>

Acknowledgements

First of all, I would like to thank all interviewees who dedicated their valuable time and shared their insights about data science with me. Their contributions build the essential basis for this thesis. I am also thankful to all participants of the survey study who constitute an important part of the thesis.

Furthermore, I would like to thank my advisor, PhD student Michael Feldman, and my supervisor, Prof. Abraham Bernstein, PhD, for their valuable insights, guidance, and support, and for giving me the opportunity to realize this interesting project in the DDIS Group.

Last but not least, special thanks go to my family and my friends who always support me and provide helpful comments. Especially, I would like to thank Samuele for his enduring support and motivation, and David and Oliver for numerous inspiring discussions during our lunch breaks.

Abstract

To overcome the growing shortage of data scientists and to accommodate the simultaneously increasing demand for data analysis experts, various ways have to be explored to find people with the required skill sets. One such way is outsourcing data analysis tasks to freelancers available on online labor markets. The objective of this research is to gain an understanding of factors essential for this endeavor. Specifically, we intend 1) to learn the skills required from freelancers, 2) to collect information about the skills present on major freelance platforms, and 3) to recognize the main hurdles to freelance data analysis. This exploratory research study adopts a sequential mixed-method approach consisting of an interpretive case study, i.e. interviews with 20 data analysis experts, followed by a web survey with 80 respondents from various freelance platforms. Together, the qualitative and quantitative study results provide comprehensive information about the research goals: Not only commonly known skills like technical or mathematical capabilities were mentioned but interviewees emphasized various factors such as understanding the domain, having an eye for aesthetics when visualizing data, being able to communicate clearly, and having a natural understanding of the possibilities and limitations of data. These skills were found to be existent on various freelance platforms, which suggests that outsourcing data analysis projects, or parts of them, to online freelancers is indeed feasible. However, there are several hurdles, including e.g. communication issues, knowledge gaps, quality of work, and confidentiality of data, which may limit the possibilities and the willingness of outsourcing data analysis to freelancers. Nevertheless, these limitations can be overcome by taking certain precautions, which will be discussed in this thesis as well.

Zusammenfassung

Um den zunehmenden Mangel an Data Scientists zu überwinden und die gleichzeitig steigende Nachfrage nach Experten für Datenanalyse auszugleichen, müssen verschiedene Möglichkeiten in Betracht gezogen werden, um Personen zu finden, die die erforderlichen Fähigkeiten besitzen. Ein möglicher Weg ist das Outsourcen von Datenanalyse-Aufgaben an Freelancer auf Online-Arbeitsmärkten. Das Ziel dieser Forschungsarbeit ist es, ein Verständnis der Faktoren zu gewinnen, die für dieses Unterfangen wesentlich sind. Insbesondere wollen wir 1) die Fähigkeiten erforschen, welche von Freelancern erforderlich sind, 2) Informationen über die Fähigkeiten sammeln, welche auf den grössten Freelancing Plattformen vorhanden sind, und 3) die primären Hürden beim Outsourcen von Datenanalyse identifizieren. Diese explorative Studie nutzt einen sequentiellen Mixed-Method Ansatz, der aus einer interpretativen Fallstudie (Interviews mit 20 Datenanalyse-Experten) besteht, gefolgt von einer Onlineumfrage mit 80 Befragten von verschiedenen Freelancing Plattformen. Gemeinsam bieten die qualitativen und quantitativen Ergebnisse der Studie umfassende Informationen über die Forschungsziele: Nicht nur allgemein bekannte Fähigkeiten wie technische oder mathematische Kenntnisse wurden erwähnt, sondern die Befragten betonten verschiedene Faktoren, wie z.B. dass man die Domäne verstehen muss, dass man ein Gefühl für Ästhetik hat, wenn man Daten visualisiert, dass man in der Lage ist, klar zu kommunizieren, und dass man ein natürliches Verständnis für die Möglichkeiten und die Grenzen von Daten hat. Diese Fähigkeiten wurden auf diversen Freelancing Plattformen gefunden, was darauf schliessen lässt, dass das Outsourcen von Datenanalyse-Projekten, oder Teile davon, an Freelancer in der Tat möglich ist. Allerdings gibt es mehrere Hindernisse, z.B. Kommunikationsprobleme, Wissenslücken, die Qualität der Arbeit und die Vertraulichkeit der Daten, die die Möglichkeiten und die Bereitschaft, Datenanalyse an Freelancer auszulagern, begrenzen. Diese Einschränkungen können jedoch überwunden werden, indem man bestimmte Vorsichtsmassnahmen berücksichtigt, welche ebenfalls in dieser Arbeit diskutiert werden.

Table of Contents

1	Introduction	1
2	Related Work	5
3	Theoretical Background	9
4	Research Approach	13
4.1	Qualitative Research	14
4.2	Quantitative Research	15
5	Results	19
5.1	Analysis of Qualitative Data	19
5.1.1	The Process of Data Analysis	19
5.1.2	Skills and Knowledge Required from Data Scientists	21
5.1.3	Most Tedious Tasks and Tasks to Be Outsourced	24
5.1.4	Skills and Knowledge Required from Freelancers	26
5.1.5	Difficulties with Outsourcing Data Analysis	28
5.2	Analysis of Quantitative Results	31
6	Discussion	39
7	Limitations and Future Research	43
8	Conclusions	45
A	Appendix from Qualitative Research	53
A.1	Interview Guideline	53
A.2	Flyer for Recruiting Interviewees	55
A.3	Email for Recruiting Interviewees	56
A.4	Summaries of Interviewees' Background	57
A.5	Tools and Programming Languages used by Interviewees	60
A.6	Statistical Methods used by Interviewees	61
A.7	Most Tedious Tasks and Tasks (Not) to be Outsourced	62
A.8	Visualizations of Interviewees' Answers	65

A.9	Tabular Summary of Possible Tasks to Outsource, Their Difficulties & How to Solve Them	67
A.10	Tabular Summary of Difficulties	70
B	Appendix from Quantitative Research	71
B.1	Pilot Test Questions	71
B.2	Online Survey	72
B.3	Market Share of Upwork and Freelancer	92
B.4	Survey Published on Upwork and Freelancer	93
B.5	Visual Summaries of Survey Data	94

Introduction

In the past years it has become evident that there is a continuously growing demand for data scientists and people able to systematically interpret data. Since the availability of data is growing faster than the availability of experts with the relevant skill set to interpret it, finding competent experts for data analysis tasks is becoming increasingly challenging due to the variety of required skills. While in the past benefiting from data products was a prerogative of big companies, having bulk servers and in-house teams of technicians supporting the hardware, currently, the development of cloud computing allows for on-demand usage of computational resources at reasonable costs. Therefore, also small up to mid-tier companies have started collecting various data about their customers, business transactions, and other records related to their business. However, the need to analyze the collected data is hampered with the growing shortage of data experts capable of analyzing the data and producing comprehensive insights that are intelligible to the wide audience of population and decision makers [Davenport and Patil, 2012].

Lacking internally available talent, companies are compelled to seek for external solutions that will allow to make sense of their data. Recognizing the need for on-demand data analysis, multiple software companies such as Microsoft and Google started developing and offering cloud-based data analytics products that allow to run various machine learning products on a big scale in the cloud. Some of these products make the premise of reducing the complexity threshold of data analysis by allowing to plug the data to their services and subsequently to run data analysis as a black box process. However, this approach leads to questionable results given the limited control over the data analysis process and restricted flexibility to preprocess the data and tailor the models for concrete needs. Therefore, even though the variety of available tools to analyze data has made it much more accessible, there is an apparent need for skilled experts to conduct and orchestrate the process of data analysis.

The lack of experts available in the geographical proximity can be resolved with online labor markets, that overcome multiple drawbacks and allow to hire experts in a flexible manner. The potential of such platforms has been recognized by the industry and the earnings of such platforms experienced sharp increase during past years, attracting growing numbers of freelancers and employers. The recruitment process is crucial to the success of online labor markets and could be theoretically supported with a substantial research body on the personnel selection theory well known in the organization psychology. However, the nature of recruitment on online labor markets is different from

traditional candidate selection due to its remote character and high level of uncertainty with regard to workers' skills and abilities. Unfortunately, research on the skills necessary for potentially crowdsourced data analysis still remains in a blind spot and has not been sufficiently addressed. The Information Systems (IS) community has surely addressed the interdependencies between tasks and different psychological and environmental aspects of individual candidates, however, in the online labor market context, no theoretical frameworks have been proposed, so far, to describe employer-employee interactions in the selection process. In our study we rely on the Person-Job (P-J) fit framework to understand how requirements for data experts match the existing talent on freelance platforms. P-J outlines the importance of the needs-supplies conceptualization as an important factor for good fit, and therefore, we explore the skills available online. We argue for the importance of this phenomenon and hope that our work will promote the discussion on the major constructs composing the selection process on online labor markets. Our current work explores the explicit factors in online employee selection by adopting the needs-supplies dimension of the Person-Job theory.

Various data analysis tasks require diverse knowledge and skills. While some tasks are fairly straightforward and effortless, others require proficiency in multiple disciplines and hands-on experience. To illustrate, descriptive statistics require a somewhat shallow statistical background, while other, predictive methods, such as neural networks or support vector machine learning, require in-depth knowledge. Therefore, our first research question is:

- **RQ1** *What are the skills required in data analysis?*

Identifying the skills required for a certain task is the first step in finding suitable workers to perform it. For successful assigning of freelancers to tasks we also need to understand the capabilities of the workers. Hence, gaining an insight about the skills of freelancers will allow to design tasks taking into account the constraints imposed by the existing talent pool in the online labor market. Therefore, the second research question is:

- **RQ2** *What are the relevant skills and characteristics that freelancers possess, and do they match the required skills for data analysis (identified for RQ1)?*

The last research question deals with various hurdles to outsource data analysis to freelancers and gain insights on potential ways to resolve these problems. Hence, our third research question is:

- **RQ3** *What kind of data analysis tasks can be outsourced to freelancers and what are the hurdles in doing so?*

To answer these questions we conducted an exploratory study including 20 interviews with data scientists, followed by a survey research with 80 freelancers to learn about the talent existing on major freelance platforms. Our contribution is twofold: 1) we fill the research gap concerning the endeavor of outsourcing data analysis to online labor markets by means of systematic research on needs vs. supplies of explicit skills and 2) provide the necessary basis for future theorization on employee selection in an online setting.

The remainder of this paper is structured as follows: In the next section we explore the related work in the domains of crowdsourcing and online labor markets, and give an overview of the theoretical work related to our study. Then we describe our research approach and outline both the qualitative and quantitative outputs of our study. Following, we discuss the results of our study and present limitations and possible future research. Finally, the thesis closes with conclusions.

Related Work

Crowdsourcing has gained increased relevance as an accepted approach for outsourcing activities to an online community. In recent years, the business community embraced the approach of outsourcing some activities to a crowd by means of evolved specialized, web-based platforms. Jobs are mostly partitioned into groups of simplified sub-tasks and distributed to crowd workers in an open call manner [Howe, 2008, Alam and Campbell, 2014]. Even though crowdsourcing has a long history (e.g. Longitude prize, Oxford English Dictionary), its current popularity is largely accounted to Amazon’s establishment of the first crowdsourcing Internet marketplace - MTurk¹. This platform provides a wealth of paid micro-tasks that require minimal time and cognitive effort, but aggregates results in major accomplishments [Kamar et al., 2012]. As for today, MTurk has preserved its leading role as the most popular platform for micro-tasks. However, the phenomenon of online work has expanded, encompassing now also platforms supporting laborious expert work in various domains. For instance, platforms such as Upwork² and Freelancer³ offer the service of matchmaking employers with freelancers based on reported expertise. Other platforms such as TopCoder⁴, 99designs⁵, or Kaggle⁶ offer contest-based participation, while yet other platforms like InnoCentive⁷ or Idea Bounty⁸ are crowdsourcing ideas for solving challenging problems. Encouraged by human computation potentials, attempts have been made recently to extend the types of human computation tasks beyond relatively simple and non-demanding ones, e.g. [Haas et al., 2015]. As a consequence, the crowdsourcing domain is faced with an emerging need to find concepts and paradigms such that complex tasks, requiring a wide spectrum of human abilities and talents, can be successfully assigned to suitable crowd workers. Finding appropriate crowd workers, however, is non-trivial due to the human motivational, cognitive, and error diversity [Bernstein et al., 2012]. Moreover, the remote and unstable character of most crowd markets, where the ability to track

¹www.mturk.com/mturk/welcome

²www.upwork.com

³www.freelancer.com

⁴www.topcoder.com

⁵www.99designs.com

⁶www.kaggle.com

⁷www.innocentive.com

⁸www.ideabounty.com

and profile workers is largely limited, gives rise to an even greater challenge to establish trustful and robust recruiting policies [Ipeirotis, 2010].

As to the freelancer platforms, even though plethora of studies are devoted to deriving optimal hiring policies based on freelancer characteristics, e.g. [Verroios et al., 2015, Kokkodis and Ipeirotis, 2015], to the best of our knowledge there is no up-to-date comprehensive study on freelancer demographics. However, there is a broad classification of the freelancers provided by the Freelancers' Union and Upwork survey with 7,000 US respondents from mid-2015 [Horowitz and Rosati, 2014]. According to this survey there are five general types of freelancers: 1) Independent Contractors (36%) - traditional freelancers that are not employed and do freelance work on a project basis, 2) Moonlighters (25%) - workers with a primary traditional job that also work as freelancers in their free time, 3) Diversified Workers (26%) - individuals with multiple income sources from part-time jobs, 4) Temporary Workers (9%) - those who work in temporary job, and 5) Freelance Business Owners (5%) - freelancers that employ workers to do their freelance projects.

Evidently, platforms such as MTurk are primarily designed for micro-tasks and do not support complex and ill-defined tasks that require well established coordination and trust relationships between the requesters and crowd workers. The rise of freelancer platforms aims to fill this gap through supporting both employers and crowd workers/freelancers with a user interface and a workflow that allows to deliver complex projects in a remote manner. In contrast to the micro-task labor market that has reached saturation and even some decline (see Ipeirotis' analysis on MTurk⁹), the online freelancer market has grown substantially during the last years, including now millions of freelancers and employers [Agrawal et al., 2013]. This shift exemplifies the transition of online labor markets from simple, short-term, and low effort jobs, as they were originally common in online labor markets, to complex and long-term tasks that are typical to traditional work settings. The growing number of white collar workers switching to online labor markets is due to the advantages online work brings along to some communities such as students, homemakers, retired experts, and single parents, while the professions to be frequently found on freelance platforms are graphic design, copywriting, data entry, and programming [Mill, 2011].

As for the data analysis professionals, similarly to the general shortage of data scientists [Davenport and Patil, 2012], i.e. experts able to provide comprehensive data-driven solutions, it is fairly uncommon to find specialists of that kind on a crowdsourcing platform. It is, however, not unusual to find workers possessing some partial knowledge and/or willing to learn a new topic. A survey of 153 data scientists, conducted by [Crowdfunder, 2015], has revealed that data scientists mostly refer to themselves as researchers (54%), computer scientists (52%), BI analysts (36%), and mathematicians (19%). Additionally, most of the respondents mentioned they are working with Excel (56%), R (43%), and Tableau (26%). The majority of the respondents consider data cleaning and organizing as the most time consuming task (67%), while 53% say that collecting data sets is the most laborious. This matches with [Kurgan and Musilek, 2006]

⁹www.behind-the-enemy-lines.com/2016/02/a-cohort-analysis-of-mechanical-turk.html

that surveyed multiple papers evaluating the relative effort in different data analysis activities and concluded that data preparation is by far the most time consuming activity with estimates ranging between 45 and 60%. As to the skills, the most in demand skills are programming and statistics, while proficiency in Python and R are by far the most prominent. According to the Accenture Institute for High Performance [Harris et al., 2013], a data scientist has to be capable of 1) designing statistical models, 2) creating machine learning and text mining algorithms, 3) cleaning and converting raw data, 4) carrying out quality assurance testing to ensure the quality of models, and 5) communicating the results through clear data visualization. Other supplementary skills such as communication, collaboration, and creativity are also mentioned as key success factors. [Chatfield et al., 2014] have analyzed a body of literature from six major academic databases and derived a set of six data science attributes commonly perceived by academic researchers: 1) Entrepreneurship and business domain knowledge, 2) Computer scientist, 3) Effective Communication skills, 4) Create valuable and actionable insights, 5) Inquisitive and curious, and 6) Statistics and modeling.

Theoretical Background

In this section we first overview existing research on hiring policies common in online markets and then review theoretical frameworks relevant to the online matchmaking of freelancers with employers. We then support our choice to adopt the Person-Job fit theory as a theoretical lens for our study.

Evaluating skills of job candidates is one of the major challenges in both online and traditional labor markets. Even though some platforms allow job candidates to perform various online tests to assess their competence in a variety of topics, cheating and tests' leakage hamper reliable evaluation. Moreover, technological advancement requires tests to be frequently updated and reliably evaluated, promising that the performance on these tests adequately reflects a candidate's skills [Christoforaki and Ipeirotis, 2015]. One major difference of online compared to traditional labor markets is the highly heterogeneous workforce composed of a crowd with different skills spanning across a variety of different professions. While candidates in traditional markets share some similarities such as common cultural and geographical specifications, in online labor markets candidates are coming from all around the world and exhibit high variance in qualities and skills. [Kokkodis and Ipeirotis, 2015] assume these skills to be latent, however, possible to be measured through the worker's available characteristics on a platform such as employee ratings, accomplished projects, hours worked, and wages. Utilizing these characteristics, they present a number of models that estimate the workers' latent skills and their evolution over time. [Feldman and Bernstein, 2014] propose that cognitive abilities of freelancers are another latent factor that, to a large extent, predefines the performance on various crowd-tasks. They examine the performance on various crowd-tasks with different setups to predict task performance where cognitive abilities, performance on previous crowd-tasks, or both of them are partially known, and show that cognition-based task assignment leads to an improvement in task performance prediction. [Suzuki et al., 2016] propose to support the skill development of workers by introducing a concept of micro-internships. According to the proposed solution, micro-internships allow workers to learn, improve, and develop new skills, while at the same time employers could evaluate the skills of a candidate. [Verroios et al., 2015] are grouping employers on Odesk (today Upwork) with respect to their hiring criteria and learn the hiring preferences for each cluster. Results show that while some groups of clients are positively biased to freelancers that are new to a marketplace, others ignore their reputation and focus primarily on a person's job fit. In this context, [Pallais, 2013]

uses a field experiment on Odesk to show that awarding new freelancers with a first job benefits the market with information about their abilities and increases freelancers' average earnings. However, the study outlines that companies will tend to hire fewer inexperienced freelancers if hiring is costly and workers cannot reimburse them in a case of failure.

The challenge of finding a good match between a person and a job has been studied from different disciplines. Typically, such problems are considered under the perspectives of task-related and social aspects [Jackson, 1996, Zhang et al., 2012], human and social capital, and person-environment fit [Sandelands et al., 1988, Werbel and DeMarie, 2005]. According to [Malinowski et al., 2006], building on these theories, an IS-supported approach for the selection of candidates needs to consider the matching of individuals to tasks as well as to other peers with whom the person will collaborate. In the same vein, [Hough and Oswald, 2000] review the personnel selection research from 1995 through 1999 and conclude that taxonomies linking between the candidates' traits and skills with jobs/tasks characteristics have to be developed in order to improve the candidates' fit for the predetermined range of tasks.

The theoretical foundations of our exploratory study may be seen in the Person-Job fit element of the well-studied Person-Environment fit theory (P-E). The concept of P-E fit originates from the interactionist theory behavior [Chatman, 1989]. This theory proposes that neither personal nor the environment characteristics on their own can fully explain the compatibility between people and organizations, but rather the interaction between these two yields to a better understanding of the fit between the person and the environment. P-E has three common conceptualizations: 1) *Supplementary vs. complementary fit*: while supplementary fit occurs when persons possess characteristics that are similar to the other individuals in the environment, complementary fit arises when the traits and skills possessed by the individual complete the missing characteristics. 2) Further, the complementary fit can be distinguished between *needs-supplies vs. demands-abilities fit*. Needs-supplies fit occurs when an organization satisfies individuals' needs, desires, or preferences. On the other hand, the demands-abilities perspective suggests that fit occurs when an individual has the abilities required to meet organizational demands. 3) The last conceptualization is *objective vs. subjective fit*. Subjective (or perceived) fit is conceptualized as the assessment whether an individual fits well in the environment, while objective (or actual) fit is the congruence between the person and the environment that is externally measured with indirect measures such as outside evaluators [Kristof, 1996, Sekiguchi and Huber, 2011].

Person-Job (P-J) fit is in the core of employee selection studies and is primarily concerned with finding suitable candidates by analyzing the demands of the job and assessing candidates' skills and abilities. Good P-J fit leads to high job proficiency and, therefore, work is likely to be accomplished quickly and with higher quality than in circumstances when job proficiency is low [Werbel and Johnson, 2001]. Originating with fairly simple interpretation sparked by Taylor's monograph on the principles of scientific management [Taylor, 1911], the process has gained sophistication with formulation of statistically reliable processes that can be used to determine P-J fit [Werbel and Gilliland, 1999, Brkich et al., 2002]. The main conceptualizations relevant

to P-J fit are *needs-supplies* and *demands-abilities*, and, as discussed earlier, both are extended conceptualization of *complementary fit*. The needs-supplies conceptualization incorporates the desires of individuals, and the characteristics and attributes of the job that may satisfy those desires. The desires can be seen as a mixture of goals, psychological needs, interests, and values attributed to the person, whereas the job supplies are described as general characteristics of pay, occupation, and job attributes. The demands-abilities perspective consists on the one hand of the job demands that are required in order to carry out the tasks of the job and on the other hand of the abilities that the individual has and can use to meet the job requirements. Job demands typically consist of the knowledge, skills, and abilities required to perform at an acceptable level in the job. Abilities include education, experience, knowledge, and skills. In employee selection practices, strategies to assess P-J fit include resumes, tests, interviews, reference checks, and a variety of other selection tools [Sekiguchi, 2004]. There is considerable evidence that a high level of P-J fit has a number of positive outcomes. The review of the P-J fit literature by [Edwards, 1991] identified job satisfaction, low job stress, motivation, performance, attendance, and retention as outcomes that are positively affected by P-J fit. When P-J fit is assessed as the congruence between an individual's desires and benefits received from performing the job, it leads to improved job satisfaction, commitment, and reduced intentions to quit.

Research Approach

The objective of this research is to gain an understanding of factors principal for outsourcing data analysis to freelancers available on online labor markets. Specifically, we are interested in learning the skills necessarily to be inherent to online freelancers, recognizing the main hurdles to freelance data analysis, and collecting information about the skills present on various freelance platforms. In our study we adopt a sequential mixed-method approach [Teddlie and Tashakkori, 2009], harnessing both the power of qualitative research as well as of quantitative research. The former comprises an interpretive case study [Myers et al., 1997, Walsham, 1995, Walsham, 2006], in which 20 data analysis experts are interviewed, whereas the latter consists of a web survey [Dillman, 2011], following a descriptive and cross-sectional research design [Pinsonneault and Kraemer, 1993]. The results of the first research phase inform the design of the second research phase. Using such a multiple-method approach leads to higher robustness of results due to triangulation [Kaplan and Duchon, 1988]. In particular, based on the various types of triangulation formulated by [Denzin, 1973], our approach allows on the one side for data triangulation - the use of a variety of sources in a study, and on the other side for methodological triangulation - the use of multiple methods to study a research problem. As qualitative research is especially appropriate for studying complex phenomena [Johnson and Onwuegbuzie, 2004], we apply this method first to explore the domain of data scientists and to gain in-depth descriptions and understanding of their environment. After having gained a thorough understanding, quantitative research is well suited to apply the learned content onto a broader population and obtain quantitative data to generalize research findings [Johnson and Onwuegbuzie, 2004]. Overall, our process consists of four steps: first, qualitative data is collected through interviews with data scientists and data analysts; secondly, the qualitative data is used to identify necessary skills and other factors for outsourcing data analysis tasks to freelancers; thirdly, an online survey is designed based on the previously identified skills and distributed on various freelancing platforms; and, last but not least, the generated quantitative data is analyzed to compare the skills available on online labor markets with the desired skills.

4.1 Qualitative Research

We have approached potential participants for interviews through personal contacts, professional online business networks (e.g. LinkedIn, XING, Data Science Central¹), and professional meetups² in Zurich, Switzerland (see Appendix A.2 and A.3), as suggested by [Ko et al., 2015]. By exploring various sources, we counted on composing a sample of individuals with diverse backgrounds, spanning over different industries and positions. The main prerequisite for participation in our study was the profession criterion; specifically, we aimed at individuals who either held positions of data scientists or data analysts, or were mainly involved in data analysis in their daily work. Whereas job titles and job requirements vary throughout organizations, stemming from the fact that “a data scientist represents an evolution from the business or data analyst role” [Zikopoulos et al., 2012], they still have a common foundation and work with data to answer business questions [Kandel et al., 2012]. Therefore, all selected participants were experts in data analysis and were able to provide valuable insights about the domain of data analysis.

In total, 20 semi-structured interviews, lasting between 30 and 60 minutes, were conducted during a seven-week period in December 2015 and January 2016. The researcher thereby played the role of an outside observer [Walsham, 1995]. Based on the participant’s preference, 14 interviews were conducted in German and 6 in English, while the interviewer is fully proficient in both languages³. This follows [Myers and Newman, 2007]’s suggestion to create a friendly environment, noting the importance of interviewees being able to use their own language and therefore increase the likelihood of disclosure. Due to geographical and time limitations of some participants, two interviews were conducted via Skype, while the others were performed face-to-face in locations preferred by the participants. All interviews were recorded and then transcribed, while notes were taken during the interviews to capture complementary non-verbal insights.

Overall, we followed closely the principles proposed by [Klein and Myers, 1999] as well as those of [Myers and Newman, 2007], in which a dramaturgical model for conceptualizing interviews is proposed. In this model, the interview is seen as a drama with a stage (e.g. an office), props such as notes or tape recorders, actors, and an audience. Interviewer and interviewee can thereby enact both roles. Furthermore, the model includes a script, i.e. the question script, an entry, an exit, and the overall performance, which is affected by all above mentioned aspects. We also followed [Walsham, 2006]’s recommendations which provided valuable inputs for collecting interview data. To ensure the coverage of all important questions, we conducted semi-structured interviews using a question script (see Appendix A.1), which allowed on the one hand for free development of the dialogue and at the same time assured a similar structure between all interviews. The interviewer first introduced herself, explained the purpose of the research and guaranteed confidentiality to the interviewees, which is important according

¹www.linkedin.com, www.xing.com, www.datasciencecentral.com

²www.meetup.com

³Quotations taken from the German interviews and used in the Results chapter were faithfully translated into English language.

to [Walsham, 2006]. Then, to begin with the questions, interviewees were asked about their occupational background and their experience with data analysis. Furthermore, they were asked to describe tasks in an analysis project they often perform and the skills and tools required for these tasks. Progressing to the main topic of the interviews, we asked what tasks they would outsource to freelancers and what kind of skills and knowledge freelancers would need to perform these tasks. Finally, we discussed difficulties and few other general conditions in the endeavor of outsourcing data analysis to freelancers. All questions were open-ended, i.e. designed in a way that participants could specify their own answers, without influence of the interviewer and thus minimizing the possibility of biases occurring in the data collection process [Fontana and Frey, 2000]. Closing the conversation, the researcher offered participants to receive the study results upon finish, as suggested by [Myers and Newman, 2007].

After all interview records were transcribed, the available data was iteratively analyzed using coding technique [Miles and Huberman, 1984]. Following this technique, we first applied open coding where the entire data was explored and broken apart to create codes. Subsequently, applying axial coding, we identified possible connections between codes and concepts [Corbin and Strauss, 2008]. This iterative process implied repeated examination of the interview data which gradually led to the elaboration of generalizations, i.e. factors and skills necessary for outsourcing data analysis tasks to online labor markets. Eventually, since no new insights were developed after 15 interviews, we concluded that data saturation [Guest et al., 2006] has been reached and stopped interviewing after 20 interviews.

4.2 Quantitative Research

After analyzing all interview transcripts, we developed an online questionnaire based on the results of the conducted interviews. In order to ensure traceability of results throughout the entire research project, and thus also support credibility, the original meanings of the identified concepts from the qualitative analysis have been transferred to the survey [Cronholm and Hjalmarsson, 2011]. Consequently, we included every factor identified in the interviews into the survey questions. Following a descriptive and cross-sectional research design [Pinsonneault and Kraemer, 1993], the purpose of the questionnaire was to study about the distribution of skills, expertise, and knowledge in the population of most prominent freelance platforms. The adopted cross-sectional design implies that data was collected once and thus represents the population at that one point in time.

We have selected freelance platforms based on their size and on the availability of freelancers with data science or data analysis experience. After a thorough analysis of currently available freelancing platforms, Upwork and Freelancer were selected as they constitute the biggest online workforce to date in general, and a large pool of freelancers specialized in different kinds of data analysis⁴ (see Appendix B.4). For each platform we

⁴Upwork reports over 10 million freelancers, Freelancer reports over 18 million users which can be both freelancers and employers. Together, they account for approximately 70% of the market share in crowdsourcing and freelancing markets (see Appendix B.3)

received 40 reliable survey submissions that have passed the quality assurance checks, making it a total of 80 participants, of which each was rewarded with 5 US dollars. Other platforms such as Amazon Mechanical Turk, Guru, or Crowdfunder were disregarded because of various reasons. MTurk and other similar crowdsourcing platforms were neglected because they concentrate on micro-tasks, which require less time and engagement with the subject matter, and thus do not fit to our intended target group for data analysis tasks. Guru on the other hand was among our initially selected platforms due to its size, however we were unable to publish the survey there because of its terms and policies that restrict tasks with a reward lower than 25 US dollars. Crowdfunder was not considered as, despite its focus on data scientists, it is rather a platform for micro-tasks which helps data scientists to alleviate their work.

For the survey study we followed guidelines taken from [Dillman, 2011] and from [Fowler, 2013]. After the questionnaire was fully designed, it was iteratively pilot-tested with seven persons (see Appendix B.1). Short discussions with each person led to improvements and helped to refine the final version of the web survey. The welcome page of the questionnaire stated the purpose and relevance of the research, guaranteed confidentiality to the respondents, and thanked them in advance for their time. To improve data quality, participants had to - despite guaranteed confidentiality - input their name on the survey so that identification of dishonest participants was possible. The final survey consisted of 29 questions, spanning a mixture of mainly Likert-type questions and some open-ended, multiple-choice, and single-choice questions (see Appendix B.2). The Likert-type questions were designated to the freelancers' skills, knowledge, and expertise with various tools, programming languages, and statistical methods. These questions were grouped into several matrices, had five-point response scales [Dawes, 2008], and were all constructed in a similar manner, i.e. starting with "To what extent...". Other questions aimed at learning about the demographics, educational and occupational background, and experience with data analysis of the respondents. In order to compare opinions of interviewees and freelancers, we asked them which tasks of a data analysis project they could or could not imagine being outsourced to freelancers, and what difficulties they see in this undertaking. Since surveys, particularly online surveys, are subject to careless or inattentive responses [Meade and Craig, 2012], the issue of reliability had to be addressed. In order to detect such malicious, careless, or random behavior of participants, several quality assurance questions were integrated into the survey, that were explicitly verifiable. These questions include "humor-evoking attention check questions" [Marshall and Shipman, 2013] which we placed into the matrix questions. For instance, we asked whether they had aliens as friends, whether they had been to the moon or planet Mars, and how their knowledge in "Apache Cow" - in the style of "Apache Pig", in "Ingression" - in the style of "Regression", and in "Random Mountain" - in the style of "Random Forest" were. Furthermore, we asked participants to create several keywords that they associated with the survey [Kittur et al., 2008]. This would show whether they paid attention to the content. Also highly effective in detecting careless respondents are open-ended questions that inquire participants about recent incidents [Gadiraju et al., 2015], in which case we asked them to state their previous tasks on the freelance platform. Additionally, we posed this question twice, in the beginning and

in the end of the survey, to cross-check if answers were identical. Stemming from the same idea, we included several skills twice to see whether responses matched. Moreover, a time tracker was added to sort out respondents who had unacceptable little time to complete the survey honestly.

During the conduction of the web survey it became clear that the above mentioned quality assurance question about having aliens as friends was sometimes misinterpreted as being friends with strangers or foreigners. Thus, although this question was meant to sort out respondents who filled in the survey randomly, it could not be used as a fully reliable catch question. Nevertheless, as a result of the implemented quality assurance checks, 10 of 90 submissions were identified as careless and random responses and thus have been excluded from the analysis. The remaining 80 submissions were analyzed with SPSS and Excel, performing one-sample two-tailed t-tests [De Winter and Dodou, 2010] in order to identify statistically significant skills and expertise in tools and statistical methods. Similarly to [Schlauderer and Overhage, 2013], we regarded those items whose mean values were significantly larger or smaller than 3 (the answer containing “to a medium extent”) as present, respectively, absent on these platforms. Moreover, for further analyses, we conducted descriptive statistics analysis as well as Spearman rank sum correlations.

Results

In this section we first present our qualitative results from the interpretive case study, i.e. the interviews, and then outline our quantitative results from the web survey study.

5.1 Analysis of Qualitative Data

Our 20 interviewees comprised a diverse group of individuals that work in various positions dealing with data science or data analysis, ranging from junior to senior positions. Their experience varied between four and 22 years, with a median of 6.5 years (Figure 5.1). By having such a diverse range of professionals, elite bias in the interviews was avoided [Miles et al., 2013]. Various industries were represented including insurances (4), financial services (4), digital analytics companies (5), analytics consultancies (3), telecommunication providers (1), retail companies (2), and Internet broadcasting companies (1). The interviewees were all experts in data analysis and 95% stated to have acquired their knowledge about it in their university studies including mainly Economics (6), Statistics (3), Physics (3), and other areas of study encompassing mathematics and computer science classes. The majority, 60%, held a master’s degree, 25% held a doctorate degree, and 15% a bachelor’s degree. 70% stated to have continued to learn on the job and 35% additionally made use of online courses, books, or individual programming tasks to improve their skills (see Appendix A.4).

5.1.1 The Process of Data Analysis

To bring the researcher and the interviewees down to a common denominator, we asked them to list and describe typical tasks or phases that they performed in data analysis projects. After summarizing the answers received from participants following workflow chart emerged, which comes close to the data science process established by [Schutt and O’Neil, 2013]:

(I) Problem Definition → (II) Data Collection → (III) Data Cleansing → (IV) Data Modeling/Analysis → (V) Visualization → (VI) Presentation

The problem definition (I) includes the requirement analysis to specify what the customer

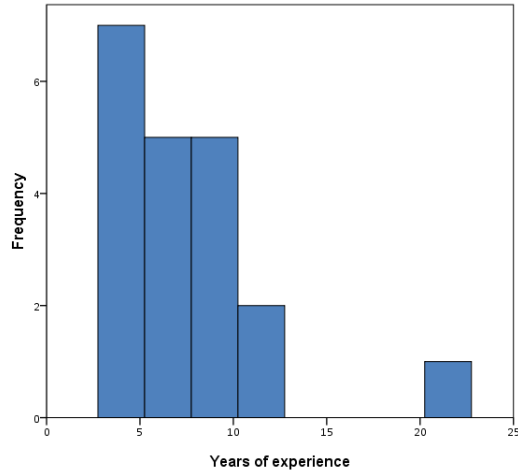


Figure 5.1: Interviewees’ years of experience depicted as a histogram: The majority of interviewees had between 4 and 10 years of experience with data analysis. Only few (3) had more than 10 years of experience.

wants and what the aim of the analysis project is. This step can be either hypothesis-driven or hypothesis-free, where the research is rather driven by serendipitous discovery than by a previously stated hypothesis. E.g. P5, participant number 5, said: *“It often happens that people come to you and say ‘Tell us something we don’t know about our data’ and that is a very vague problem proposition.”* After that, relevant data has to be collected (II) for the specified purpose. This can be done by retrieving data from a database, by crawling it from the web, or by buying the data. P20 criticized current practices: *“Although theory tells us that we need a golden source of truth of data, it’s still the case with companies that they have a broad landscape of databases, or systems, i.e. the data that is usually needed does not lie in one single place, not under one single form.”* The next step comprises the data cleansing (III) in which the analyst has to deal with messy, wrong, or missing data. First, the data has to be understood in order to decide what is necessary to ensure data quality and how the data should be prepared and cleaned. Possible activities include combining different data sets, imputing missing data, eliminating outliers, and reshaping data. According to the interviewees, this part takes up most of their time, up to 80%, and includes many feedback-loops with the person from whom they received the data. After the structure of the data is established, the actual analysis can take place (IV). As a first step, analysts use some visualization tools to explore the data, referred to as the *“playground”* phase by P16, and to understand the data and its structure better. They often use statistical software to build models, to aggregate, filter, sort, and segment data by different kinds of dimensions, to look at patterns, and to determine *“why”*. This process is iterative, as once a model is built, the analyst might realize he needs to get additional data or to cleanse the available data further. When it comes to the visualization of the results (V), analysts have

several possibilities including plotting figures of e.g. KPIs, ROI, or trends, developing dashboards, and designing reports and documentations. This phase also includes many review phases with the customer. As a last step, there is usually a presentation and communication of the results (VI) to the management and customers in order for them to draw conclusions. P7 described the distribution of steps in following way: *“What you really see of the analysis is the visualization on top, that’s the typical iceberg, and all the data preprocessing is the big bulk that you don’t see.”*

5.1.2 Skills and Knowledge Required from Data Scientists

Whereas the main focus of the interviews was to extract the necessary skill set for freelancers, we first asked interviewees to name the skills they themselves needed in their jobs (Figure 5.2). This would provide data to partially answer RQ1. All interviewees declared technical skills to be absolutely necessary. These include programming skills (75%), database skills (55%), machine learning skills (30%), and general IT affinity like understanding how servers, software, and apps work (30%). Also statistical and mathematical skills were mentioned by almost every interviewee (90%) which includes to be able to conduct statistical analyses with various methods and tools and to have a general flair for numbers. P5 explained the necessity for statistical and mathematical skills in following way: *“You need statistics and math. And we really hold that near and dear to our hearts. Normally when we talk to stakeholders, to people who provide us with the use cases and the data, they don’t really agree or they don’t really understand why do you need a big mathematics skill set, but I would argue you do even nowadays when most of the algorithms are already coded and packaged into libraries, you still need a good mathematical background to make your conscious decision of the techniques you’re using. And then to evaluate them, oftentimes you might need to bridge algorithms from the library by coding yourself a technique. So definitely, definitely good statistics and mathematics background.”* Also P18 argued that mathematical and statistical understanding is very important as it is necessary to *“understand in what direction you can develop the problem at its best.”* Regarding the order of importance, P19 and some other interviewees declared programming skills as most important: *“I would say without statistical skills, but with coding skills you can do something, but vice versa, if you have statistical skills but no coding skills, you cannot do anything.”* However, those skills do not have to be very deep, as P6 summarized his and some other interviewees’ opinion: *“It’s more about having some high level programming skills and query skills.”* Furthermore important, stated by 55%, is domain knowledge in the respective field in which the data analysis project is situated, e.g. in finance, as mentioned by P20: *“If I analyze data from the financial industry and I personally have no idea what finance is, what a bond is, what an obligation is, what a stock is, then it will go wrong very fast”*, or marketing, as mentioned by P14: *“If you do an analysis for marketing, you need to understand how a campaign works and what is relevant for a campaign.”* It is necessary to understand the context and to know the company’s business goals, as this influences the direction of an analysis project. Interdisciplinarity was mentioned to be important in this context, as data analysts or scientists often have to tackle problems from diverse areas of a company. P10

stated “*You need to understand the process behind it*” and P14 said “*You need to have a pretty broad understanding, not always super deep, but I think it is important in breath, in order for you to understand what you are actually analyzing.*”

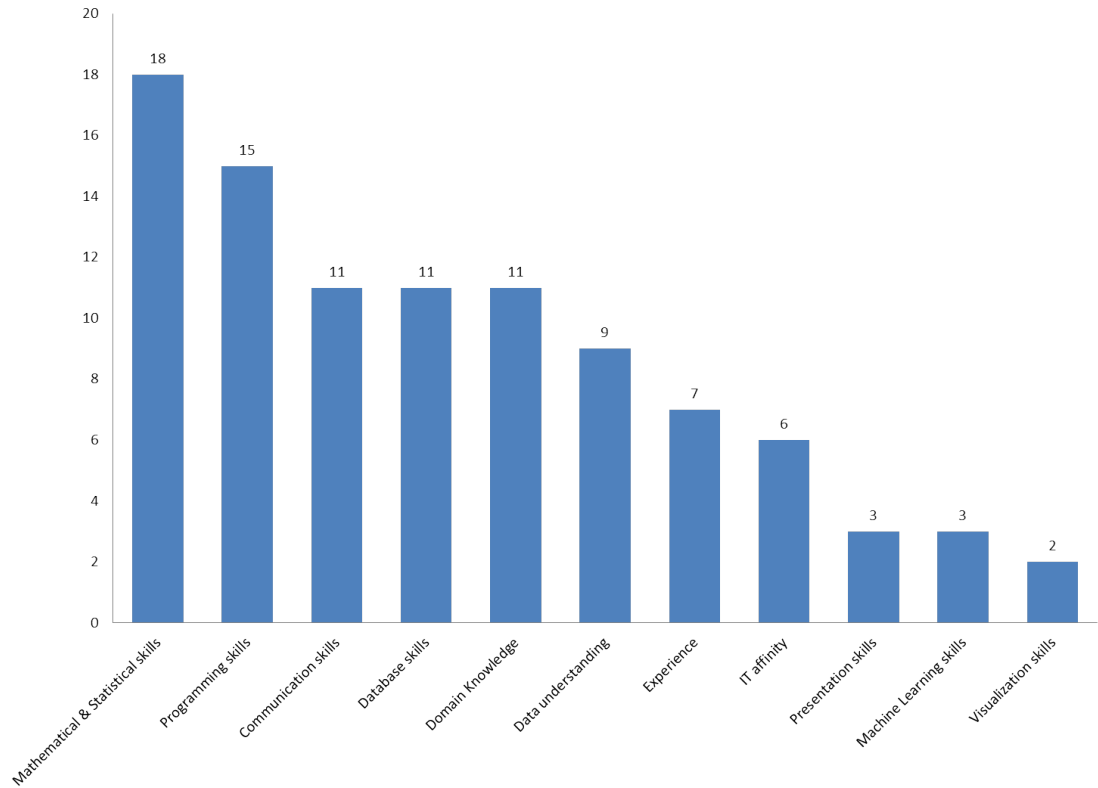


Figure 5.2: Interviewees’ skills: Almost every interviewee (18 of 20) said to have mathematical and statistical skills, followed closely by programming skills (15). More than half of interviewees (11) also stated to have communication, database, and domain knowledge, which are important for being a good data scientist.

Another very important skill is communication, mentioned by 55% of the interviewees. Being able to communicate with different groups of interest and present results in a clear and simple way is essential for the success of an analysis project. If communication fails, the effort for the analyses could be in vain, like P3 stated: “*Without communication it’s maybe also a waste of work. [...] It doesn’t help to just sit at the computer, you can be the best programmer, but at the end you need to be able to communicate the results clearly, and in the beginning you need to understand, what is important for the other person.*” Hand in hand with communication skills goes also the flair for consulting and mentoring, mentioned by 25%. This includes working well with customers and keeping an open mind towards their needs.

Having a mindset that is oriented towards data is another aspect interviewees men-

tioned several times (45%), meaning to understand data and its structure, to be sensitive to what is possible with data analysis and where its limits lie. P15 illustrated her experience with clients who approach her with analyses problems and to whom she has to explain the limits of data analysis: *“I have to explain sort of the actual explanatory power here of the data isn’t very great, because I can only tell you that there was this big jump, maybe I can break it down like where it happened, geographically, or what time. I can try to get out an answer for you with the characteristics of the change, but really, the causation, I cannot give you a definitive answer.”* Also P1 talks about similar experiences: *“Sometimes you also have to say: ‘That is the problem, I have the data, but it doesn’t work. I cannot answer it with the data. Period. Doesn’t work. We can do 30 more iterations, it won’t work.’ That’s also difficult to sell, but it simply doesn’t work.”* This skill of data understanding intersects with the aspect that data analysts or scientists need experience (35%), in their job in order to be successful, e.g. to know what method to use best in what case and how to interact with the given data.

Their way of working and approach to problems needs to be structured, and they need to be able to break problems down, abstract, and think in steps. Logical thinking, including function-oriented thinking, logical deduction, and reasoning, are further stated points. These aspects were mentioned by 45% of interviewees, e.g. by P9: *“I would say you need to have very very good analysis skills, logical deduction. You need to understand that if $A + B = C$, it doesn’t always equal C , that it can also be $A + B = D$, because of the fact that you have different influences and different context.”* Also P5 said: *“Because nobody will ever come to you and say, here is a set of numbers and please do a prediction on this set of numbers. They’re gonna come with a very fuzzy problem, with very dirty data, with the very unclear requirements. And you have to make sense of the chaos before you apply any of the pretty techniques we know from, you know, academic papers. So it’s a part of abstraction.”* Others stated: *“You need a structured approach, because if you’re not a tiny bit structured, you start losing yourself in the data. So you should really always keep in mind, where do we want to go or which variable or function we want to optimize. And then anything we do, we target towards this goal. Otherwise we end up basically analyzing data which is not relevant for us”* (P19), and: *“Thinking in scripts certainly helps, thinking in steps, I do this and then that and out of that I make this, and then it branches off and then I put it together again and aggregate it, so this flowchart kind of thinking, that is extremely important”* (P7).

Oftentimes, data analysts or scientists are confronted with new problems in a new field, or have to apply new methods, new tools, or new software. For that, they need to be willing to learn new things (20%) and to keep a curious attitude (20%). P20 specified this as *“willingness to learn something new, to come out of one’s own world, out of the comfort zone.”* P7 mentioned both aspects in his statement: *“You need to be extremely curious, you need to be an inventor, a developer, you must not have fear of new things”*. Curiosity was not only mentioned in connection with learning new things, but also as an aspect of being curious about what can be found in the available data, as stated by P16: *“This analytical curiosity, that you can simply recognize certain structures in the data itself, or also exercise patience to play around a little bit and to look in what direction the whole issue will develop.”* Also, data scientists and analysts need to be

patient and enduring, as analysis project often require exhausting examination of data, like P4 said: *“I call it digging into the stuff. Because sometimes you spend hours on some tools or some data and you don’t find anything. So it’s very hard sometimes to keep you motivated.”* Thinking about the problem deeply and trying to get to the bottom of it, thus having investigative skills, is therefore an important trait of data analysts. This includes paying attention to details and having an intuitive and inventive mind. This was mentioned by several interviewees (25%), e.g. by P7: *“You need to learn to pay attention to the last, to the smallest detail. You will, and that’s guaranteed, stumble over those at the end”*, or by P11: *“There is also systematics, but there is also a lot of intuition in the process.”* At the same time they need to be skilled in project management (15%) as data analysis projects require good time management, mentioned e.g. by P10: *“Oftentimes you work under time pressure. [...] You oftentimes also don’t get the data that you need at the right time. So sometimes you have to prepare scripts blindly, then you receive the data and apply the scripts onto the raw data. [...] Because there are bigger latencies, from the moment of ordering the data until you receive them. And you don’t know the quality a priori. Yes, you need good time management.”*

Several interviewees (30%) noted the fact that a data scientist cannot have all skills mentioned above. Data science or data analysis projects require teamwork, *“it would be optimal to have three people”* (P8), and thus each analyst has a specific role with an according skill set, *“usually you just pick one or two max, that you try to perfect in a sense”* (P4). This skill set can be composed of the previously discussed skills.

As to the software, a great majority of the interviewed data analysts and data scientists work with R (85%) and Python (75%) in their jobs. In general, the opinion prevailed that one programming language is required but that it is irrelevant which one. Rather, it is important, as stated by P2, to *“having acquired all the programming logic in one or another language.”* Also SQL (55%) and Excel (45%) are commonly used among the interviewees. Other tools that were mentioned several times include Java, JavaScript, Tableau, Hadoop technologies, ETL tools, and Oracle Database. A full overview of all mentioned tools can be found in Appendix A.5. Regarding statistical methods, P11 said: *“It’s a full zoo of methods, over which you need to have a little overview.”* Mainly, interviewees utilize on the one hand descriptive statistics and on the other hand regressions, clustering, classifications, predictions and a number of machine learning algorithms (for a more detailed overview see Appendix A.6). P16 looks at this topic with grief: *“90% of our jobs are descriptive. And all the cool stuff, that is really fun, is unfortunately done way too rarely.”*

5.1.3 Most Tedious Tasks and Tasks to Be Outsourced

The following data provides partial answer to RQ3. Although 70% of interviewees stated that data cleansing is one of the most tedious tasks in data analysis, only 50% mentioned of their own accord to outsource data cleansing to freelancers. Additionally, they expressed doubts about it, saying that specifications have to be absolutely clear and that the process of outsourcing data cleansing would be very difficult. For example, P12 stated: *“You have to make absolutely sure that you know what has been done during*

data cleansing, and which assumptions have been made during data cleansing, which data has deliberately not been added, which data was changed in what way.” Essentially, after asking participants directly at the end of the interview whether data cleansing could be outsourced to freelancers, four additional interviewees answered affirmatively, making it 70% who would see data cleansing as a potential task to be outsourced. The overall concern was that data cleansing is a very important step in a data analysis project and thus, if outsourcing it to a freelancer, no guarantee is given that the task is performed correctly, and hence the following analysis could be in vain. P1 stressed the importance of data cleansing by saying: *“If you put garbage in, you will get garbage out.”* And P18 added how important it is to perform this step very cleanly because *“it’s what has to be done in order to be able to continue with the work.”* Also P10 argued: *“If something goes wrong, all following steps are dependent on it.”* That’s why he would rather outsource *“those parts that come chronologically at the end of the process chain.”* Nevertheless, despite all concerns, data cleansing remains the most frequently mentioned task interviewees would like to outsource to freelancers. Furthermore, it is interesting to notice that the most tedious task is also the one which they would like to outsource the most, regardless of the complexity it contains.

Also data acquisition was mentioned as a very tedious task by 40% of interviewees, and as a potential task to be outsourced by 30%. One reason why data acquisition is seen as tedious, was stated by P10: *“Because, in a nutshell, you never get what you want. You always have requirements about the data that you would like to have, and what you get is very very far away from that.”* Again, the second most tedious task is also the second most mentioned task interviewees would like to outsource (for further information see Appendix A.9). Further tasks that have been declared to be tedious, repetitive, or frequent are writing reports (20%), testing (15%), and working with business or management (15%). The latter is first of all due to the fact that business requirements change constantly and data has to be adjusted accordingly, which results in constant maintenance work; furthermore, when presenting in front of management there is no guarantee that their wishes have been met; and last but not least, one problem mentioned by P12 is that business sees data science as a crystal ball: *“The hope towards data science is that there is a crystal ball that can be looked into and that answers all kinds of questions that you have. And the understanding, that you need data to answer such questions, and that there are questions which cannot be answered with a certain absoluteness, this understanding is a big challenge.”*

Other tasks that are seen as potential tasks to be outsourced by freelancers are the actual analysis or modeling (25%), the visualization of data (20%), the implementation and testing of analysis tools (20%), and the writing of reports (15%). Furthermore, 10% of the interviewees declared the presentation of results, the system and database maintenance, the implementation of the infrastructure such as a production environment, and the development of recommendation systems as possible tasks to be outsourced. Data profiling, data exploration, the ETL process, model validation, and operational support were mentioned each by one interviewee. 20% stated that only the entire process as a whole could be outsourced, since tearing apart the process would be too difficult, as e.g. explained by P1: *“Because not everybody thinks in the same way, and one person*

maybe doesn't know what the other person did. Even if you comment it, you didn't do it yourself, you don't know what exactly has been going on. [...] It is difficult to isolate this in individual steps without losing the understanding for the overall process."

Interestingly, without being asked, most interviewees also mentioned tasks they would certainly not like to outsource. They include again data cleansing mentioned by 30% (versus 70% who would outsource it), as it requires very specific knowledge about the client, the business and its processes, and is an iterative process that has many review loops and requires a lot of customer contact. Furthermore, the actual analysis or modeling was mentioned by 30% as a task they would not outsource, on the one hand because it also requires too much knowledge of the company and its systems, and on the other hand because the knowledge about the created model needs to be in-house in case clients have questions about it. 15% mentioned they would not outsource the specification process, 10% the visualization of the data, and 5% the data acquisition and the documentation of results. To summarize, P20 said: *"The less the process is definable, the less repetitive, [...] the more data quality problems arise, the more it is difficult to outsource the whole thing"* (see Appendix A.7).

5.1.4 Skills and Knowledge Required from Freelancers

Logically, each task requires different skills from freelancers. Since we are interested in the entire skill set that should be present on a freelance platform for any given possible task, we asked interviewees for the desired skills of freelancers for their mentioned possible tasks, and combined all answers to receive a full overview of necessary skills (Figure 5.3). This provides complementary data to answer RQ1.

Statistical and mathematical skills as well as database skills are the most necessary skills freelancers should have according to the interviewees, each stated by 80%, e.g. by P10: *"Solid mathematical basic knowledge is compulsory."* Next, programming skills and domain knowledge were mentioned each by 65%. P20 explained the necessity for domain knowledge in following way: *"Data analysis per se doesn't exist. It's always data analysis in a context: logistics, medicine etc. So the know-how in this context, in this domain, is absolutely necessary. Without it, it's difficult. You need to explain first what the data means, signifies. If people don't understand it, they cannot identify the data quality problems etc. So that is really important."* Data understanding, i.e. to understand data and its structure and how to work with it, *"data thinking"* as called by P12 was as well categorized into important skills (50%). Followed by 45% are communication skills that encompass being able to talk to different groups of interest, being an *"interface person"* as called by P7, and communicate results in a clear and straightforward way. P3 stressed the importance of communication: *"I think when it's external, it's even more important that there is good communication, because the person doesn't know the company well and the company doesn't know the person well. That's why you have to pay attention to communication all the more."* Furthermore important are visualization skills (40%), i.e. to visualize data in a meaningful way and have an eye for simple and clear design. P5 illustrated: *"You need to understand what speaks to a person, what appeals at a person, when they look at the chart or at the plot or, you know, some statistics."* And P13 added

that freelancers should have *“maybe also some creative mind, I’m not saying artistic, but something in that direction, in order to visualize it meaningfully.”* Paying attention to details in visualizations, e.g. making sure *“that you don’t make it red and green, but maybe orange and blue, because people have a red-green deficiency”* is important as stated by P14.

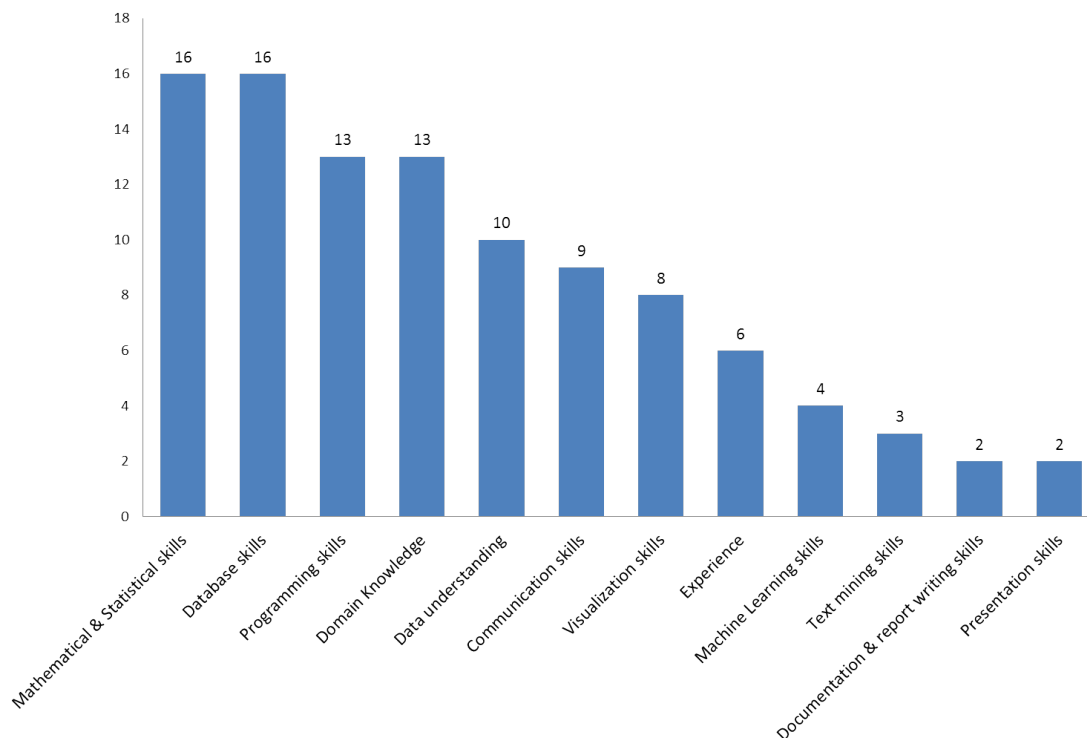


Figure 5.3: Skills freelancers should have: According to interviewees, freelancers necessarily need mathematical and statistical skills (16), technical skills such as database (16) and programming skills (13), and domain knowledge (13).

Having experience, mentioned by 30%, is an advantage since it helps to make the right decisions e.g. about the appropriate method, and also reassures the hiring employer that the freelancer has done something similar before. Machine learning, text mining, and documentation and reporting skills were also mentioned as necessary skills for freelancers by 10 to 20%. Moreover, some interviewees (25%) noted that freelancers should have an algorithmic and logical way of thinking, which includes breaking problems down into smaller parts and deduct, as well as maintain a big picture view throughout their work. This was e.g. mentioned by P1 who said *“You need to think algorithmically, you need to break problems down into sub-problems, and then handle them one after another, and in doing so nevertheless not lose focus of the big whole thing, and capture the problem in its entirety.”* Also P14 said in the same direction: *“You need quantitative logical thinking, simplify, break down problems to a great extent. You must not lose yourself in too high*

complexity, you always need to somehow keep the big picture.” Furthermore, freelancers should be trustworthy, accurate, reliable, thorough, patient, and willing to learn new things.

Almost every interviewee said freelancers should know R (90%), followed by Python (65%). In general, similarly to the responses about their own skills, they regarded one programming language as necessary for freelancers, but meaningless which language (40%), because on the one hand learning a new language is feasible and on the other hand a problem can usually be solved in various languages, as stated by P1: *“If you’re good in some programming language, you will get your problems worked out.”* Also P13 said: *“I would say that it doesn’t matter which programming language, because someone who knows one programming language, learns a new one very quickly.”* And P2 also reasoned: *“You should have learned some language at some point, so that you can adapt the actual language that you need for that easier.”* However, interviewees mentioned that Java, JavaScript, and HTML are advantageous to know (each 15%). Other tools that freelancers should be familiar with are SQL (45%), Excel (40%), and ETL tools such as Talend (25%). Between 5% and 20% mentioned SPSS, SAS, Stata, Matlab, Hadoop technologies including Apache Pig, HDFS, MapReduce, and Apache Spark, Oracle Database, Tableau, Microsoft Visio and Powerpoint, QlikView, Google Analytics, Webtrends, and Adobe Analytics. Furthermore, freelancers should be experienced with diverse data formats such as JSON, CSV, and XML. In fact, P1 even stated: *“They have to be able to handle every data format that exists.”* Also, they should know different kinds of operating systems in order to be able to work with the operating system of the client.

5.1.5 Difficulties with Outsourcing Data Analysis

The following data provides answers to why outsourcing data analysis tasks to freelancers is not necessarily an easy endeavor (RQ3). As Figure 5.4 shows, almost every interviewee was concerned about communication issues (80%). To begin with, a common language and shared understanding of the matter is necessary. This implies very clear requirements and well defined tasks. One problem here is that knowledge is sometimes assumed to be implicit and thus is not explicitly communicated between the freelancer and the customer, which can result in misunderstandings, inefficiencies, and non-transparency. Transparency is very important, i.e. it is essential that the client knows how the freelancer came to the results of the data analysis e.g. during data cleansing, because otherwise the auditability of data is no more guaranteed. This was a big concern by P7: *“If you don’t know where the raw data is coming from, up to the graph, it’s no more auditable. If you don’t know what your outsourcing colleagues are doing, it’s no more auditable. And auditability is extremely important.”* Furthermore, because of the long-distance work relationship, communication can take longer as one cannot simply go over to a colleague’s desk but rather has to wait for e.g. an email response. Also, cultural differences can lead to communication problems, as noted by two interviewees, e.g. P2: *“When you have foreign-language or diverse cultures, that maybe all speak English, but that maybe have a completely different understanding of a task, how to do it, and I would say the further away the cultures are from each other the more difficult it*

is.” The fact that information can be lost during intermediate steps of communication, called “*Chinese Whisper Effect*” by P8, is another possibly occurring problem.

The initial briefing is a related issue that was mentioned as a difficulty by 30% of the interviewees: “*That’s the tricky part, getting the briefing right*” (P4). Freelancers need to know exactly what they have to do, so the briefing has to be very precise which means additional time and cost. And this in return raises the question if it is worth to outsource. “*I fear a little bit, that exactly this means so much more effort, that in the end it doesn’t pay off to outsource*” (P16). Also P14 raised the concern: “*Oftentimes you ask yourself, should I rather do it myself or work somebody in, after all it’s an investment.*” High setup costs and time, not only regarding the briefing, but also the infrastructure, was mentioned by 15% as a barrier to outsourcing data analysis tasks to freelancers. The effort has to be “*justified*” as stressed by P3. Connected to this whole issue is also the knowledge gap (40%) that results from the complex IT environment of a company and the entire domain knowledge that freelancers first have to familiarize with.

Another big issue mentioned by 55% of the interviewees is privacy and confidentiality of data. Similarly to other respondents, P8 said: “*The problem with outsourcing is always that you actually don’t want to give away the data, primarily for us they are customer data.*” This is also one of the reasons why some companies have not utilized outsourcing services so far and would feel uncomfortable giving out their data to freelancers. Ways to deal with this problem is anonymizing data or signing non-disclosure agreements. Still, this problem remains a sensitive issue. P6 pointed out the difficulty of finding a trade-off when he has to “*anonymize the sensitive data but to retain the utility of the data.*” Furthermore, even if data is anonymized, it can happen that conclusions can be drawn about the actual identities of persons. P8 exemplified this with a recent scandal, where anonymized medical data was publicly available, however somebody could draw conclusions through the use of some data analysis approach, and published the entire medical history of a senator online.

Monitoring and control for the quality of work is another difficulty seen by 40% of the interviewees. “*If you were doing crowdsourcing, you have no guarantee, whatsoever, on the quality of the code or the piece of analytics that you get. So [...] somebody has to go and verify afterwards. And then it remains to be assessed, you know, how much you benefit from crowdsourcing if you need to check afterwards what happened*” (P5). P7 stated: “*You rely on something. Some filter is set falsely, a join wrongly, then it’s a construct on sand, gone with the next wave.*” Trust into freelancers (20%), the vulnerability of data whenever it is passed around (20%), the meeting of deadlines (15%), and the danger of data manipulation (15%), be it intentional or accidental, are other issues that were mentioned several times. P20 asked himself during the interview: “*Does the guy have a huge incentive to disappear with the data and supply the competition with the analyses? Can I prosecute him? Can I find him at all?*” P13 sees following danger with data analysis in general and even more with freelancers: “*I can always steer data analysis a little bit in a certain direction. So if you pursue some interests and know some statistics, I’m not saying cheating, but bending statistics in a way that they claim something even if it’s not really true.*” P5 mentioned her concern: “*You need to know where your data is going. You need to make sure it’s not being utilized to other*

purposes than what you gave it for. Because once you give the data, you know, copies can happen anytime anyhow, you don't watermark it, it's not like money. So you might get your results back but your data may go elsewhere as well." And P14 pointed out: *"If something comes to light or is sold, then it can get really ugly for a company very fast."* Thus, safeguards need to be applied. Further two interviewees argued that the problem to be outsourced has to be self-contained, i.e. to have a clear beginning and end. Another issue is that the freelancer and the client need to agree on a common interface, be it a tool or a system, which is necessary in order to guarantee reproducibility. Also the fact that the amount of data is oftentimes very big leads to difficulties due to technical reasons. Furthermore, it is more difficult for a freelancer to have influence in a company's project than for an internal employee. One problem which is not directly concerned with freelancers but with the outsourcing company was mentioned by 25% of interviewees: *"I think it's a huge problem that companies often don't even know what they can do in the first place"* (P3). Thus, they do not understand how they can draw meaningful insights through data science and hence have to be guided in the initial phase of exploring the possibilities of data analysis.

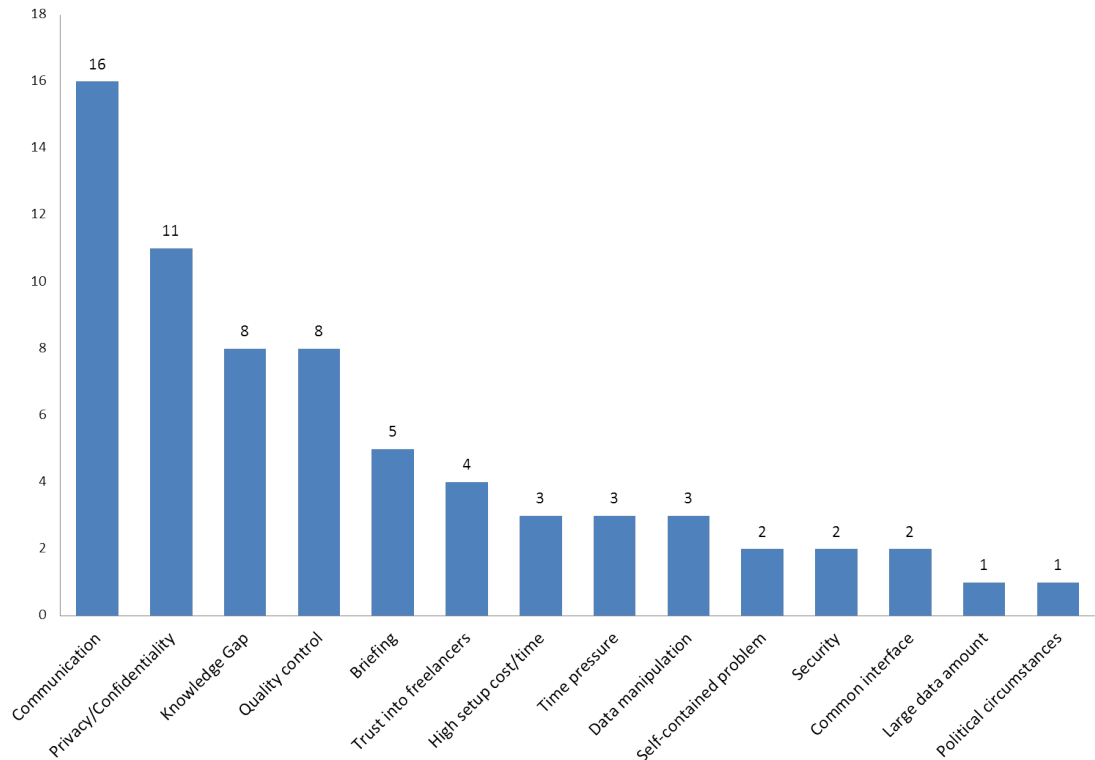


Figure 5.4: Difficulties: Interviewees see communication issues (16) as the biggest hurdle in outsourcing data analysis tasks, followed by privacy and confidentiality issues (11).

Other questions that were asked during the interviews are if they or their companies had outsourced data analysis to external specialized companies before, for which 55% answered “Yes” and 45% “No”. Also they were asked if freelancers need to understand the structure and the meaning of the data, for which 65% answered “Yes”, reasoning e.g. like P9: *“If you don’t understand the meaning, then you will not be able to draw any meaningful conclusion.”* 35% answered that it would depend on the task and that if freelancers did not have to understand it, then they would need very exact instructions, like a *“cooking recipe”* (P18) to follow in order to execute the task. Furthermore, they were asked whether they would feel comfortable giving out their data to freelancers, for which 30% answered “Yes”, 15% “No”, and 55% with a mixed answer. Those whose answers were mixed argued that it depended on the data. With sensitive data or data concerning the core business they would not feel comfortable, whereas with anonymized or open access data they would agree doing so. The last question inquired whether they had experienced difficulties in their job where freelancers could have helped them out, for which 55% answered “Yes” and 45% “No”. Reasons for the affirmative answers were time-related bottlenecks and lack of own resources and expertise (see Appendix A.8).

5.2 Analysis of Quantitative Results

We received 80 survey submissions from freelancers of the platforms Upwork and Freelancer, which provide insights to answer RQ2. Their experience with data analysis ranged from under a year to 45 years, with a median of 4 years. Noteworthy, a lot of beginners, i.e. freelancers with experience of just one year, were among the participants. Since data science is an emerging field, this could indicate that a lot of people are eager to get started in this profession. Participants rated their expertise on average with 3.82 and median of 4 on a five-point Likert scale (Figure 5.5). They stated to have learned about data analysis on the job (22%), through university courses (22%), through the Internet (17%), books (16%), online courses (15%), and teaching videos (8%).

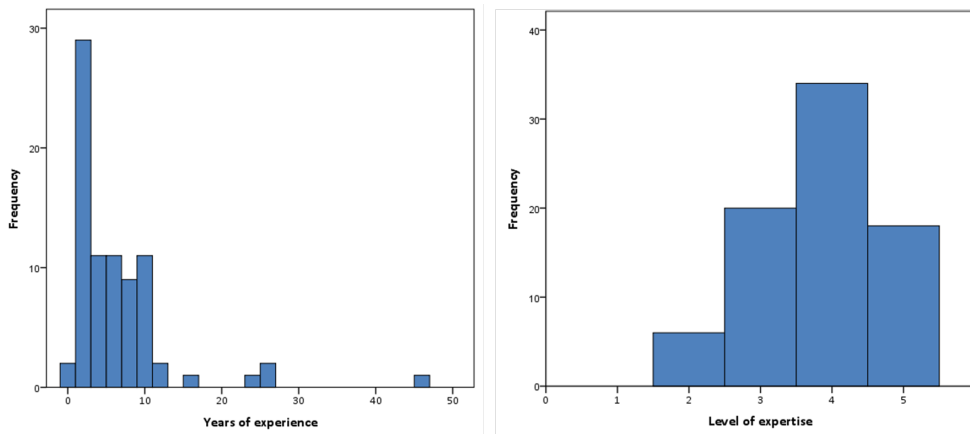


Figure 5.5: Participants’ years of experience and level of expertise in data analysis

Most freelancers, 76%, were male, whereas 24% were female. 51% were between the age of 25 and 34, 24% between 18 and 24, and 13% between 35 and 44; the remaining 12% spread between 45 and over 65 years. Almost half of participants were living in European countries, a quarter in Asia and the remaining in America, Africa, or Australia. All participants either had a university degree or were enrolled as beginning students. The level of education was thus quite high, with 28% holding a doctorate degree, 40% a master's degree, and 25% a bachelor's degree. Their field of studies encompassed mainly Computer Science, Mathematics and Statistics, Engineering Sciences, Natural Sciences, and Business and Management. The majority, 59%, was employed in full or part-time jobs, 19% were currently looking for jobs, and 18% were students. Due to the high employment rate, 45% spent less than 10 hours per week on freelance work, whereas the rest spread more or less evenly over more hours. For more details see Appendix B.5.

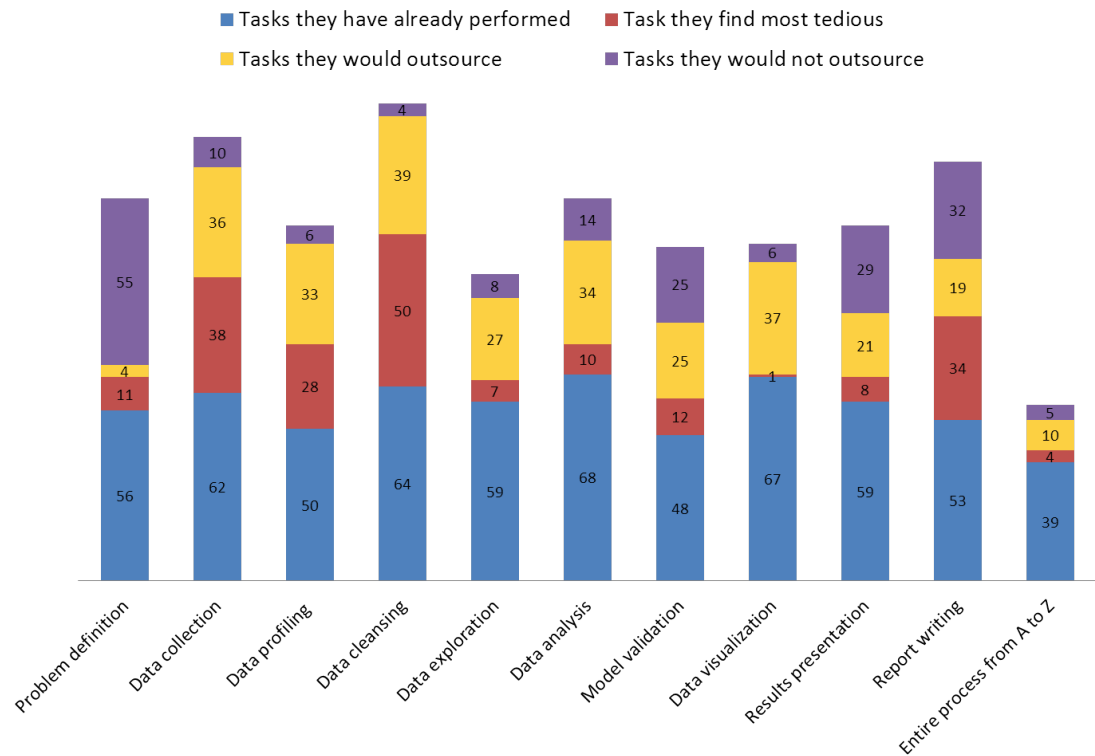


Figure 5.6: What tasks survey participants have already performed, regard as the most tedious, and would (not) outsource

In order to see to what extent participants were acquainted with the steps of data analysis, we asked them - in accordance with the process we established from the interview analyses - what tasks in an analysis project they have already performed. Results were more or less equally distributed, with data collection, cleansing, analysis, and visu-

alization being selected the most. Following up, we were interested in finding out what they regard as the most tedious tasks, which resulted in data cleansing being selected by 62.5% of freelancers, data collection by 47.5%, and report writing by 42.5%. This supports our qualitative findings in which data cleansing was stated by 70% to be the most tedious task, followed by data collection (40%) and report writing (20%). Asking which of the tasks could potentially be outsourced in their opinion resulted in data cleansing being again the most commonly chosen answer (48.75%), which also reflects our qualitative findings, where data cleansing was mentioned by most interviewees (70%) as a task to be outsourced. Following closely are many other steps such as visualization, collection, analysis, and profiling. Only the definition of the problem was selected by almost no participant. The tasks freelancers do not see as possible to outsource mirror the least chosen answer in the previous question: the problem definition was chosen by most freelancers as a task not to outsource (68.75%). Some also do not see the last few steps of an analysis project, i.e. model validation, results presentation, and report writing as possible tasks to be outsourced. All in all, as can be seen in Figure 5.6, data cleansing stands out as the task that is on the one hand the most tedious and on the other hand the most possible to be outsourced to freelancers. Following, data collection and data profiling are further tedious tasks that were mentioned to be possible to be outsourced. Although report writing is as well a tedious task, survey participants would rather not outsource it. Also the problem definition is definitely a task they would not outsource. On the contrary, data visualization, which is not a very tedious task according to the participants, was stated by many as a task possible to be outsourced.

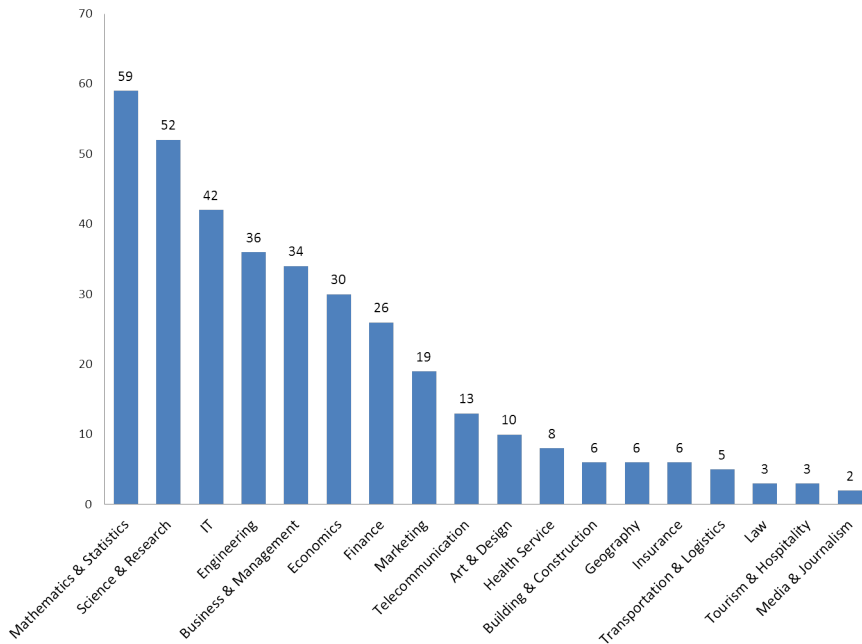


Figure 5.7: Domain expertise of survey participants

Since our interviewees mentioned that freelancers should have expertise in the field where the analysis project is situated, we asked freelancers in what domains they were experienced. Consequently, Mathematics & Statistics were chosen as the most common domains, followed by Science & Research, IT, Engineering, Business & Managements, Economics, and Finance (Figure 5.7).

In order to test all skills from the survey for statistical significance we performed t-tests by comparing the mean values of the responses to the test value 3, which corresponds to having skills to a medium extent. Results show that all general skills, i.e. statistical, mathematical, database, programming, communication, presentation, visualization, machine learning, text mining, documentation and report writing skills, and data understanding were significantly different from 3. Thus, all mean values were larger, ranging from 3.28 in machine learning to 4.45 in data understanding, indicating that freelancers had a rather high level of skills (Table 5.1).

T-test and Descriptive Statistics for General Skills									
Variable	N	T-test			Descriptive Statistics				
		t-value	p-value	mean \neq 3	min.	max.	mean	median	std.dev.
Data understanding	80	21.116	.000***	Yes	3	5	4.45	5	.614
Communication skills	80	14.367	.000***	Yes	2	5	4.23	4	.763
Documentation/Report skills	80	11.075	.000***	Yes	2	5	4.08	4	.868
Presentation skills	80	9.786	.000***	Yes	1	5	4.00	4	.914
Visualization skills	80	9.649	.000***	Yes	2	5	3.96	4	.892
Mathematical skills	80	9.494	.000***	Yes	1	5	3.91	4	.860
Statistical skills	80	7.980	.000***	Yes	1	5	3.83	4	.925
Programming skills	80	6.749	.000***	Yes	1	5	3.80	4	1.060
Database skills	80	3.717	.000***	Yes	1	5	3.40	3	.963
Text Mining skills	80	2.895	.005*	Yes	1	5	3.34	4	1.043
Machine Learning skills	80	2.085	.040**	Yes	1	5	3.28	3	1.180

Legend: ***: 0.001, **: 0.01, *: 0.05 significance level

Table 5.1: T-test and descriptive statistics of survey participants' skills

Performing the same tests for programming languages, tools, and data formats, we could obtain following insights (Table 5.2). Almost every item was statistically significantly different from 3, however, having means lower than 3. This implies that a lot of tools, programming languages, and data formats are not widely known by the surveyed freelancers. R, Python, and SQL, three items that were mentioned the most by our interviewees to be important for freelancers to have (90%, 65%, 45%), were found to be not significantly different from 3. This suggests on the one hand that these three tools are slightly better known than other tools, which had means lower than 3. But on the other hand it also indicates that know-how for these tools is not highly present on the freelance platforms. Since particularly R and Python are some of the core skills required for data scientists as found by our interviewees and by [Kurgan and Musilek, 2006], our findings from the web survey support the fact that the widely discussed shortage of data scientists [Davenport and Patil, 2012] exists on online labor markets as well. The only items having a statistically significant mean greater than 3 were Excel, PowerPoint, and CSV. Excel, with a mean of 4.21, is widely known among freelancers and they feel highly

proficient in it. This is an important insight, as 40% of our interviewees stated Excel as a necessary skill to be known by freelancers.

T-test and Descriptive Statistics for Programming Languages, Tools, and Data Formats									
Variable	N	T-test			Descriptive Statistics				
		t-value	p-value	mean \neq 3	min.	max.	mean	median	std.dev.
PowerPoint	80	14.267	.000***	Yes	2	5	4.26	4	.791
Excel	80	12.952	.000***	Yes	2	5	4.21	4	.837
CSV	80	4.388	.000***	Yes	1	5	3.70	4	1.427
SQL	80	.845	.401	No	1	5	3.11	3	1.191
Python	80	-.553	.582	No	1	5	2.91	3	1.416
R	80	-.883	.380	No	1	5	2.85	3	1.519
MATLAB	80	-1.178	.242	No	1	5	2.81	2	1.424
HTML	80	-1.485	.141	No	1	5	2.79	3	1.280
XML	80	-4.855	.000***	Yes	1	5	2.38	2	1.151
Microsoft Visio	80	-3.994	.000***	Yes	1	5	2.38	2	1.400
Google Analytics	80	-4.307	.000***	Yes	1	5	2.36	2	1.324
JSON	80	-4.585	.000***	Yes	1	5	2.26	2	1.439
Java	80	-5.659	.000***	Yes	1	5	2.25	2	1.185
CSS	80	-5.522	.000***	Yes	1	5	2.24	2	1.235
Javascript	80	-7.257	.000***	Yes	1	5	2.10	2	1.109
SPSS	80	-6.266	.000***	Yes	1	5	2.01	1	1.410
Oracle Database	80	-7.318	.000***	Yes	1	5	2.00	1.5	1.222
ETL Tools	80	-7.422	.000***	Yes	1	5	1.89	1	1.341
Tableau	80	-8.611	.000***	Yes	1	5	1.81	1	1.233
MapReduce	80	-9.624	.000***	Yes	1	5	1.76	1	1.150
SAS	80	-10.090	.000***	Yes	1	5	1.75	1	1.108
Apache Hadoop	80	-10.477	.000***	Yes	1	5	1.69	1	1.121
HDFS	80	-11.058	.000***	Yes	1	5	1.65	1	1.092
Stata	80	-13.392	.000***	Yes	1	5	1.48	1	1.018
Apache Spark	80	-14.098	.000***	Yes	1	5	1.48	1	.968
QlikView	80	-16.000	.000***	Yes	1	4	1.40	1	.894
Apache Pig	80	-18.321	.000***	Yes	1	4	1.36	1	.799
Adobe Analytics	80	-20.904	.000***	Yes	1	4	1.31	1	.722
Webtrends	80	-25.784	.000***	Yes	1	4	1.20	1	.624
Talend	80	-30.552	.000***	Yes	1	5	1.14	1	.545

Legend: ***: 0.001, **: 0.01, *: 0.05 significance level

Table 5.2: T-test and descriptive statistics of survey participants' tool, programming, and data format expertise

Again, we performed the same t-tests, this time for freelancers' knowledge of statistical methods (Table 5.3). Even if freelancers are not very knowledgeable in statistical tools as indicated by the previous table (5.2), they stated to have a high level of skills in statistical methods, particularly in descriptive and most inferential statistics methods. This is implied by the fact that most tested statistical methods had a mean value that is statistically significantly greater than 3. Machine learning techniques on the other hand had mean values that were lower than 3, partly statistically significant and partly not. This could show that these methods are not yet widely adopted by freelancers.

T-test and Descriptive Statistics for Statistical Methods									
Variable	N	T-test			Descriptive Statistics				
		t-value	p-value	mean \neq 3	min.	max.	mean	median	std.dev.
Descriptive statistics	80	12.834	.000***	Yes	1	5	4.40	5	.976
Regressions	80	8.608	.000***	Yes	1	5	4.10	4	1.143
Correlation	80	9.344	.000***	Yes	1	5	4.05	4	1.005
Prediction	80	9.014	.000***	Yes	1	5	3.98	4	.968
Estimation	80	7.892	.000***	Yes	1	5	3.86	4	.978
Data mining	80	6.620	.000***	Yes	1	5	3.85	4	1.148
Classifications	80	7.062	.000***	Yes	1	5	3.84	4	1.061
Hypothesis testing	80	6.290	.000***	Yes	1	5	3.78	4	1.102
Forecasting	80	6.356	.000***	Yes	1	5	3.78	4	1.091
Clustering	80	6.130	.000***	Yes	1	5	3.74	4	1.076
Extrapolation	80	5.137	.000***	Yes	1	5	3.69	4	1.197
Time series	80	3.985	.000***	Yes	1	5	3.58	4	1.290
Inferential statistics	80	3.601	.001**	Yes	1	5	3.51	4	1.273
Interpolation	80	3.494	.001**	Yes	1	5	3.51	4	1.312
Nearest neighbor	80	1.136	.259	No	1	5	3.19	3	1.476
Dimension reduction	80	.760	.449	No	1	5	3.13	3	1.470
Hidden pattern recognition	80	.366	.715	No	1	5	3.05	3	1.221
Principal component	80	.289	.774	No	1	5	3.05	3	1.550
Likelihood	80	.169	.866	No	1	5	3.03	3	1.321
Naïve Bayes classifier	80	-1.170	.246	No	1	5	2.80	2	1.529
Random forest	80	-1.278	.205	No	1	5	2.78	3	1.575
Support vector machine	80	-1.721	.089	No	1	5	2.71	2.5	1.494
Neural network	80	-2.192	.031*	Yes	1	5	2.66	2.5	1.377
Natural language processing	80	-4.581	.000***	Yes	1	5	2.43	2	1.123
Bayesian network	80	-4.297	.000***	Yes	1	5	2.40	2	1.249
Network theory	80	-5.263	.000***	Yes	1	5	2.28	2	1.232
Deep learning	80	-5.504	.000***	Yes	1	5	2.21	2	1.280

Legend: ***: 0.001, **: 0.01, *: 0.05 significance level

Table 5.3: T-test and descriptive statistics of survey participants' statistical methods expertise

To test whether correlations exist between any surveyed items such as freelancers' experience, skills, and expertise in various tools, programming languages, and statistical methods, we performed Spearman rank sum correlations. As expected, the more years of experience freelancers had, the higher they ranked their level of expertise in data analysis. In turn, with increasing level of expertise they rated almost every general skill higher, except database, programming, and machine learning skills. Also most inferential statistical methods were positively correlated with their level of expertise, again, excluding most machine learning techniques.

The higher freelancers ranked themselves in statistical skills, the higher they also rated their expertise in tools such as R, Excel, SPSS, and Stata, and in statistical methods, including all descriptive and inferential, and many machine learning methods. Also mathematical skills, being highly positively correlated with statistical skills, show that freelancers have high expertise in all descriptive and inferential statistical methods. Programming skills were highly correlated with machine learning skills, which also manifests in the responses about machine learning techniques and several other statistical methods.

Furthermore, all tested programming languages, all data formats, mathematical skills, Matlab and Hadoop technologies were positively correlated with programming skills.

Additionally, the higher freelancers ranked themselves in data understanding, the higher they also perceived their level in almost every other skill and in most inferential and descriptive statistical methods. Moreover, communication, presentation, visualization, and report writing skills were all positively correlated with each other. The higher freelancers rated their machine learning skills, the higher they also ranked their mathematical, statistical, programming, and data understanding skills; also programming languages, Hadoop technologies, statistical methods and all individually tested machine learning techniques were ranked accordingly.

Those freelancers who stated to have high skills in Python also rated their skills in R and other programming languages, in Hadoop technologies, and in most statistical methods including machine learning methods higher. On the contrary, Python and Excel were negatively correlated, i.e. there is a tendency that freelancers are either skilled in Python or in Excel. Freelancers who rated themselves highly in R, show a similar distribution, however, instead of programming languages they were rather skilled in several other statistical tools such as SAS and Stata. Moreover, with increasing years of experience and level of expertise, freelancers also had higher skills in Excel, whereas Python and Hadoop technologies were in turn negatively correlated to the years of experience freelancers had in data analysis. This could suggest that these tools and technologies are rather adopted by people who are newer in the field than by people who are already set on more traditional technologies, such as Excel.

To compare interviewees' and freelancers' opinions on the obstacles that exist when outsourcing data analysis, we also asked freelancers open-ended questions about possible hurdles. One difficulty they see is that the problem has to be very well defined and requirements and specifications have to be clearly set (40%). Also, it has to be made sure that freelancers understand the problem and the goal of the analysis project (8.75%). Otherwise, as mentioned by one respondent, following problem could arise: *"Freelancers might misunderstand the main objective of the project, thus building different models or using less-satisfactory techniques to solve the problem at hand."* In this regard, briefing (23.75%) was mentioned as an essential and also difficult phase of the outsourcing process, in which *"providing maximum information regarding the problem to be solved"* is necessary. Accordingly, also communication was mentioned by many participants (36.25%) as a difficulty when outsourcing data analysis to online freelancers. Knowledge gaps (13.75%) were also often identified obstacles in the outsourcing process, as e.g. mentioned by one respondent: *"Freelancers might not have the knowledge in the specific domain to conduct any meaningful interpretation of the results."* This is why it is the more important to find and choose freelancers that possess the necessary skills for a given project, and are as well reliable (21.25%). One participant regarded this as rather impossible: *"Data analysis is quite a difficult science branch demanding lots of devotion, scrupulousness, time, knowledge, and, last but not least, responsibility. In my modest opinion, the majority of people possessing these qualities do not freelance."* Furthermore, quality of work was seen as a problem and thus appropriate monitoring and control needs to be applied (16.25%). Confidentiality of data was stated as a problem

by solely 6 participants (7.5%). Time zone differences, language barriers, and providing an accurate scope regarding time and price were seen as hurdles each by 5 participants.

Discussion

Together, the qualitative and quantitative study results provide comprehensive information both about expected skills from freelancing data analysts as well as about the talent existing on major freelance platforms. Moreover, the interview results contribute to a better understanding of the obstacles towards outsourcing entire or parts of data analysis projects to online labor markets. Interestingly, the skills identified by the interviewed data scientists are not only limited to concrete competencies picked up throughout studies such as math or coding, e.g. [Kurgan and Musilek, 2006], but go much beyond it, encompassing various skills required for data analysis. In the following, we discuss the answers to our previously stated research questions.

- **RQ1** *What are the skills required in data analysis?*

The most prominent skills data scientists should have, in accordance with literature, are mathematical/statistical skills and technical affinity such as database and programming capabilities. However, in addition to those, our interviewees emphasized the importance of domain knowledge and communication skills which highly influence the success of an outsourced analysis project. Also having an eye for aesthetics and details when visualizing data is a trait that is not necessarily associated with data science, but was mentioned by many of our interviewees as very important. Moreover, possessing the above mentioned skills does not immediately lead to being a good data scientist. What else is crucial, as highlighted by our interviewees, is a mindset that is oriented towards data. Thus, having a combination of hard and soft skills, understanding of data and knowing how to get the most out of it, are important factors, that altogether represent a good data scientist. This includes understanding the limits of what can be achieved with the data at hand and the ability to communicate those limitations to the clients. The clients in turn, according to the interviewed data scientists, do not have a thorough understanding of data analysis and see data science as an oracle, capable of answering any kind of questions. These excessively high expectations could be attributed to the spread of data science buzz through the mainstream in recent years.

- **RQ2** *What are the relevant skills and characteristics that freelancers possess, and do they match the required skills for data analysis (identified for RQ1)?*

All skills that interviewees expected freelancers to have were statistically significant when tested (with means ranging from 3.28 to 4.45 on a 5-point Likert scale), suggesting the existence of necessary skills for outsourcing data analysis on these platforms. Table 6.1 is arranged in decreasing order of freelancers' self-reported skills (last column in the table). Interestingly, the most highly ranked skills are abilities attributed to general data understanding, communication, and documentation - skills going into the direction of so called "soft skills". This can be explained with the subjective character of these capabilities and might hint to the need to find additional ways to evaluate these skills. On the other hand, freelancers feel most unconfident about basic skills such as text mining and machine learning. This can be explained with long specialization required in order to be proficient in these topics. We also asked data scientists what skills they have in order to see whether they project their own set of skills to those required from online freelancers or seek for experts with complementary abilities (first column in the table). The skills that data scientists expected from freelancers much more than they had themselves are documentation, visualization, and database skills. On the other hand, data scientists did not expect freelancers to be as good as they are in advanced topics that require mathematics, statistics, and machine learning. It seems like data scientists might be interested in workers with a complementary set of skills that could perform tasks which do not require advanced knowledge but rather skills that allow to perform general tasks such as extracting data or preprocessing. All in all, concluding from the data, the necessary skills to perform data analysis projects exist on freelance platforms and outsourcing them, or parts of them, to online freelancers is a feasible task with regard to the skills.

Skills...	...Data scientists have	...Data scientists think freelancers should have	...Freelancers have (on a 5-point scale)	
			Mean	Std.Dev.
Data understanding	45%	50%	4.45***	.614
Communication skills	55%	45%	4.23***	.763
Documentation/Report skills	-	10%	4.08***	.868
Presentation skills	15%	10%	4.00***	.914
Visualization skills	10%	40%	3.96***	.892
Mathematical skills	90%	80%	3.91***	.860
Statistical skills	90%	80%	3.83***	.925
Programming skills	75%	65%	3.80***	1.060
Database skills	55%	80%	3.40***	.963
Text Mining skills	-	15%	3.34**	1.043
Machine Learning skills	30%	20%	3.28*	1.180
Domain Knowledge	55%	65%	Was asked directly	
Experience	35%	30%	Was inferred from years of experience and self-reported expertise on a 5-point scale	

Legend: ***: 0.001, **: 0.01, *: 0.05 significance level

Table 6.1: Comparison of data scientist skills, freelancer skills required according to data scientists, and existing skills on freelance platforms

- **RQ3** *What kind of data analysis tasks can be outsourced to freelancers and what are the hurdles in doing so?*

However, not only the compliance of necessary skills determines the success of an outsourced data analysis project, but also several other factors. Particularly, both interviewees and freelancers saw communication issues as the biggest hurdle when outsourcing data analysis. This includes the necessity to have clear requirements about the project, conducting precise briefing with the freelancer, establishing shared understanding of tasks, and maintaining good communication throughout the project. Also quality assurance and knowledge gaps are aspects that were mentioned by both interviewees and freelancers as hurdles in an outsourced project, whereas the latter laid even more emphasis on finding freelancers with appropriate knowledge and skills. Privacy and confidentiality of data on the other hand were mostly a concern of the interviewees, and not as much by the surveyed freelancers. All difficulties may lead to the fact that, although one hopes to save own resources in terms of time, money, and employees, outsourcing the project could mean additional effort in terms of high setup costs and loss of time through additional communication, briefing, and performing quality assurance checks (see Appendix A.10).

Furthermore, interviewees and surveyed freelancers stated preprocessing data the most as a possible task to outsource to online freelancers. However, it is also the task which entails the most difficulties when outsourced. Additionally, it was also stated as the most tedious task, suggesting that data scientists would like to outsource the most tedious tasks, regardless of the complexities they contain. Problems identified during the interviews for the data cleansing process were possibly confidential data, the necessity of domain knowledge for freelancers to understand what the data represents, the fact that a lot of assumptions are put into data cleansing which is thus subjective, and that trust into freelancers is necessary in order to hand out data to them. Moreover, data cleansing is a complex and iterative process which requires a lot of customer contact and has to be iteratively repeated. Hence, the effort for coordinating with the freelancer is very big as communication has to be constantly maintained in both ways and specifications and results have to be clearly conveyed. The next most mentioned task to outsource to online freelancers, as mentioned by the interviewees, is data collection. But also in this step several difficulties arise such as that data is spread over various sources and that a freelancer first needs to have an aggregated overview in terms of where the needed data is stored, and how to get access to them. Also, data collection is error-prone and needs to be auditable, i.e. exchange of knowledge between the freelancer and the client about the details of the executed processes has to take place. Furthermore, several interviewees stated to outsource only the entire process as a whole to a freelancer, since all steps within the project are connected to each other and breaking the process apart would lead to loss of knowledge and complicate the project. Definition of the problem or specification of the project, on the other hand, was rather mentioned as a task not to outsource, by interviewees as well as by freelancers. This is due to the fact that this step entails several difficulties such as that freelancers need a lot of knowledge about the company, the domain, and the data. Interestingly, opinions about the actual analysis

of data or the modeling were divided almost equally among interviewees who see it as a task to outsource and those who do not. In any case, they attributed many difficulties to this step such as that, again, domain, background, and data knowledge is necessary. Furthermore, to handle large volumes of data freelancers would need a lot of storage and computational power. Another issue is that data analysis can be manipulated in terms of bending statistics, for which on the one hand trust into freelancers is necessary, but on the other hand also some control mechanism, as proposed by our interviewees. Also, similarly to previously mentioned difficulties, this step requires a lot of communication and exchange of knowledge, requirements, and results, and thus means additional effort (see Appendix A.9).

- *Theoretical Justification*

The needs-supplies conceptualization of the Job-Fit theory provides us with theoretical justification to examine the skills required for online data analysis as opposite to the existing talent pool (i.e. supplies) in online labor markets. However, to support future theorization on a job's fit in online labor markets, it is necessary to better understand the skills and abilities required in this setting. Our study takes the first step towards reaching this goal through thorough methodological exploration of the skills required for freelance data analysis. We have chosen data analysis as a prototypical job of online labor markets for two reasons: 1) It is a common domain existing on all major online labor markets and 2) Data science is a sought after domain with substantial shortage of experts. Therefore, our study - focusing on this domain - has potential to contribute in addition to practitioners not only by shedding light on the existing talent in online labor markets but also by discussing the hurdles in outsourcing data analysis to such platforms.

Limitations and Future Research

It is to say that our approach has the following limitations. We have conducted an exploratory study, mainly bounded to the explicit factors related to the needs-supplies conceptualization of Job-Person fit. However, behavioral factors such as cognitive abilities, personal desires, or satisfaction as well as other psychological needs have not been taken into account. Further research is expected to be developed touching on the topics related to the behavioral specifications of candidates in the future years. Moreover, the cross-sectional design we followed in this study can be expanded to a longitudinal design [Pinsonneault and Kraemer, 1993] in order to see how abilities, skills, and expertise develop over time.

Another limitation of our study is the fact that the skills of freelancers have been self-reported and might therefore be biased. An alternative way of assessing freelancers' knowledge would be through exhaustive examination with various tests. While this is a more reliable way of assessment, in this study we primarily aimed at collecting comprehensive data from a big sample of freelancers active on different major online labor market platforms, and thus receiving an initial overview of the skills. Future research in this direction might sacrifice the generalizability over different platforms for the sake of an improved evaluation of skills.

Lastly, even though the geographical proximity of interviewees and the fact that they reside in Switzerland and Germany might be seen as a potential argument for lack of generalization, we tried to overcome this limitation through a relatively big number of experts that have been interviewed and thanks to the fact that many of them work in international companies. We, of course, hope to see other comprehensive studies in this domain conducted in other countries as well.

Conclusions

Due to the shortage and the increasing demand of data scientists, various ways have to be explored to find competent workers to perform data analysis tasks. One way to approach this problem is through online labor markets, especially, since crowdsourcing and freelancing platforms have gained considerable relevance in today's business environments. However, research regarding necessary factors for outsourcing data analysis to online labor markets is barely existent. Therefore, we made a first step towards closing this research gap by exploring various factors such as required skills, knowledge, and expertise, as well as by discussing possible hurdles in this endeavor.

Our exploratory study, which consisted of a sequential mixed-method research, included an interview study with data science experts, followed by a web survey with numerous freelancers. The results of the study revealed what skills are required for data analysis and whether freelancers' skills match those requirements. Furthermore, we could answer what obstacles exist that hinder outsourcing data science projects, how they are related to individual data analysis tasks, and what potential solutions they encompass. In addition, through the adoption of the needs-supplies dimension of the Person-Job fit theory, we provide answers on the needs vs. supplies of explicit skills, and therefore build the basis for future theorization on employee selection in online settings.

It could be shown that the skills required for data analysis are existent on major freelance platforms and that outsourcing data science projects or parts of them to online labor markets is thus feasible. These skills include data understanding, communication, documentation, presentation, and visualization, as well as mathematical/statistical and technical abilities like programming, database, text mining, and machine learning. Furthermore, although various hurdles exist in this endeavor such as communication issues, knowledge gaps, quality of work, and confidentiality of data, they can be overcome, making outsourcing data analysis tasks possible.

References

- [Agrawal et al., 2013] Agrawal, A., Horton, J., Lacetera, N., and Lyons, E. (2013). Digitization and the contract labor market: A research agenda. Technical report, National Bureau of Economic Research.
- [Alam and Campbell, 2014] Alam, S. and Campbell, J. (2014). Examining cultural volunteer crowdsourcing technology: An appropriation perspective. In *35th International Conference on Information Systems (ICIS 2014)*, Auckland, New Zealand.
- [Bernstein et al., 2012] Bernstein, A., Klein, M., and Malone, T. W. (2012). Programming the global brain. *Communications of the ACM*, 55(5):41–43.
- [Brkich et al., 2002] Brkich, M., Jeffs, D., and Carless, S. A. (2002). A global self-report measure of person-job fit. *European Journal of Psychological Assessment*, 18(1):43.
- [Chatfield et al., 2014] Chatfield, A. T., Shlemoon, V. N., Redublado, W., and Rahman, F. (2014). Data scientists as game changers in big data environments. ACIS.
- [Chatman, 1989] Chatman, J. A. (1989). Improving interactional organizational research: A model of person-organization fit. *Academy of management Review*, 14(3):333–349.
- [Christoforaki and Ipeirotis, 2015] Christoforaki, M. and Ipeirotis, P. G. (2015). A system for scalable and reliable technical-skill testing in online labor markets. *Computer Networks*, 90:110–120.
- [Corbin and Strauss, 2008] Corbin, J. and Strauss, A. (2008). *Basics of qualitative research*. Thousand Oaks, CA: SAGE, 3rd edition.
- [Cronholm and Hjalmarsson, 2011] Cronholm, S. and Hjalmarsson, A. (2011). Experiences from sequential use of mixed methods. *The Electronic Journal of Business Research Methods*, 9(2):87–95.
- [Crowdflower, 2015] Crowdflower (2015). Crowdflower 2015 data scientist report. Technical report.
- [Davenport and Patil, 2012] Davenport, T. H. and Patil, D. (2012). Data scientist: The sexiest job of the 21st century. *Harvard business review*, 90:70–76.

- [Dawes, 2008] Dawes, J. G. (2008). Do data characteristics change according to the number of scale points used? an experiment using 5 point, 7 point and 10 point scales. *International journal of market research*, 51(1).
- [De Winter and Dodou, 2010] De Winter, J. C. and Dodou, D. (2010). Five-point likert items: t test versus mann-whitney-wilcoxon. *Practical Assessment, Research & Evaluation*, 15(11):1–12.
- [Denzin, 1973] Denzin, N. K. (1973). *The research act: A theoretical introduction to sociological methods*. Transaction publishers.
- [Dillman, 2011] Dillman, D. A. (2011). *Mail and Internet surveys: The tailored design method–2007 Update with new Internet, visual, and mixed-mode guide*. John Wiley & Sons.
- [Edwards, 1991] Edwards, J. R. (1991). *Person-job fit: A conceptual integration, literature review, and methodological critique*. John Wiley & Sons.
- [Feldman and Bernstein, 2014] Feldman, M. and Bernstein, A. (2014). Cognition-based task routing: Towards highly-effective task-assignments in crowdsourcing settings. In *35th International Conference on Information Systems (ICIS 2014)*, Auckland, New Zealand.
- [Fontana and Frey, 2000] Fontana, A. and Frey, J. H. (2000). The interview: From structured questions to negotiated text. *Handbook of qualitative research*, 2(6):645–672.
- [Fowler, 2013] Fowler, F. J. (2013). *Survey research methods*. Sage publications.
- [Gadiraju et al., 2015] Gadiraju, U., Kawase, R., Dietze, S., and Demartini, G. (2015). Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1631–1640. ACM.
- [Guest et al., 2006] Guest, G., Bunce, A., and Johnson, L. (2006). How many interviews are enough? an experiment with data saturation and variability. *Field methods*, 18(1):59–82.
- [Haas et al., 2015] Haas, D., Ansel, J., Gu, L., and Marcus, A. (2015). Argonaut: macro-task crowdsourcing for complex data processing. *Proceedings of the VLDB Endowment*, 8(12):1642–1653.
- [Harris et al., 2013] Harris, J. G., Shetterley, N., Alter, A. E., and Schnell, K. (2013). The team solution to the data scientist shortage. *Accenture Institute for High Performance*.
- [Horowitz and Rosati, 2014] Horowitz, S. and Rosati, F. (2014). Freelancing america: A national survey of the new workforce. *Freelancers Union & Elance o-Desk*.

- [Hough and Oswald, 2000] Hough, L. M. and Oswald, F. L. (2000). Personnel selection: Looking toward the future—remembering the past. *Annual review of psychology*, 51(1):631–664.
- [Howe, 2008] Howe, J. (2008). *Crowdsourcing: How the power of the crowd is driving the future of business*. Random House.
- [Ipeirotis, 2010] Ipeirotis, P. G. (2010). Analyzing the amazon mechanical turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students*, 17(2):16–21.
- [Jackson, 1996] Jackson, S. E. (1996). The consequences of diversity in multidisciplinary work teams. *Handbook of work group psychology*, pages 53–75.
- [Johnson and Onwuegbuzie, 2004] Johnson, R. B. and Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational researcher*, 33(7):14–26.
- [Kamar et al., 2012] Kamar, E., Hacker, S., and Horvitz, E. (2012). Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 467–474. International Foundation for Autonomous Agents and Multiagent Systems.
- [Kandel et al., 2012] Kandel, S., Paepcke, A., Hellerstein, J. M., and Heer, J. (2012). Enterprise data analysis and visualization: An interview study. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2917–2926.
- [Kaplan and Duchon, 1988] Kaplan, B. and Duchon, D. (1988). Combining qualitative and quantitative methods in information systems research: a case study. *MIS quarterly*, pages 571–586.
- [Kittur et al., 2008] Kittur, A., Chi, E. H., and Suh, B. (2008). Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456. ACM.
- [Klein and Myers, 1999] Klein, H. K. and Myers, M. D. (1999). A set of principles for conducting and evaluating interpretive field studies in information systems. *MIS quarterly*, pages 67–93.
- [Ko et al., 2015] Ko, A. J., LaToza, T. D., and Burnett, M. M. (2015). A practical guide to controlled experiments of software engineering tools with human participants. *Empirical Software Engineering*, 20(1):110–141.
- [Kokkodis and Ipeirotis, 2015] Kokkodis, M. and Ipeirotis, P. G. (2015). Reputation transferability in online labor markets. *Management Science*.
- [Kristof, 1996] Kristof, A. L. (1996). Person-organization fit: An integrative review of its conceptualizations, measurement, and implications. *Personnel psychology*, 49(1):1–49.

- [Kurgan and Musilek, 2006] Kurgan, L. A. and Musilek, P. (2006). A survey of knowledge discovery and data mining process models. *The Knowledge Engineering Review*, 21(01):1–24.
- [Malinowski et al., 2006] Malinowski, J., Keim, T., Wendt, O., and Weitzel, T. (2006). Matching people and jobs: A bilateral recommendation approach. In *System Sciences, 2006. HICSS’06. Proceedings of the 39th Annual Hawaii International Conference on System Sciences*, volume 6. IEEE.
- [Marshall and Shipman, 2013] Marshall, C. C. and Shipman, F. M. (2013). Experiences surveying the crowd: Reflections on methods, participation, and reliability. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 234–243. ACM.
- [Meade and Craig, 2012] Meade, A. W. and Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological methods*, 17(3):437.
- [Miles and Huberman, 1984] Miles, M. B. and Huberman, A. M. (1984). *Qualitative data analysis: A sourcebook of new methods*. JSTOR.
- [Miles et al., 2013] Miles, M. B., Huberman, A. M., and Saldana, J. (2013). *Qualitative data analysis: A methods sourcebook*. SAGE Publications, Incorporated.
- [Mill, 2011] Mill, R. (2011). Hiring and learning in online global labor markets. *Stanford University*.
- [Myers et al., 1997] Myers, M. D. et al. (1997). Qualitative research in information systems. *Management Information Systems Quarterly*, 21(2):241–242.
- [Myers and Newman, 2007] Myers, M. D. and Newman, M. (2007). The qualitative interview in is research: Examining the craft. *Information and organization*, 17(1):2–26.
- [Pallais, 2013] Pallais, A. (2013). Inefficient hiring in entry-level labor markets. Technical report, National Bureau of Economic Research.
- [Pinsonneault and Kraemer, 1993] Pinsonneault, A. and Kraemer, K. (1993). Survey research methodology in management information systems: an assessment. *Journal of management information systems*, 10(2):75–105.
- [Sandelands et al., 1988] Sandelands, L. E., Brockner, J., and Glynn, M. A. (1988). If at first you don’t succeed, try, try again: Effects of persistence-performance contingencies, ego involvement, and self-esteem on task persistence. *Journal of Applied Psychology*, 73(2):208.
- [Schlauderer and Overhage, 2013] Schlauderer, S. and Overhage, S. (2013). Exploring the customer perspective of agile development: Acceptance factors and on-site customer perceptions in scrum projects. In *34th International Conference on Information Systems (ICIS 2013), Milan, Italy*.

- [Schutt and O’Neil, 2013] Schutt, R. and O’Neil, C. (2013). *Doing data science: Straight talk from the frontline*. O’Reilly Media, Inc.
- [Sekiguchi, 2004] Sekiguchi, T. (2004). Person-organization fit and person-job fit in employee selection: A review of the literature. *Osaka keidai ronshu*, 54(6):179–196.
- [Sekiguchi and Huber, 2011] Sekiguchi, T. and Huber, V. L. (2011). The use of person–organization fit and person–job fit information in making selection decisions. *Organizational Behavior and Human Decision Processes*, 116(2):203–216.
- [Suzuki et al., 2016] Suzuki, R., Salehi, N., Lam, M. S., Marroquin, J. C., and Bernstein, M. S. (2016). Atelier: Repurposing expert crowdsourcing tasks as micro-internships. In *CHI 2016, San Jose, California, USA*.
- [Taylor, 1911] Taylor, F. W. (1911). *The Mathematics Teacher*, 4(1):44–44.
- [Teddle and Tashakkori, 2009] Teddle, C. and Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. Sage Publications Inc.
- [Verroios et al., 2015] Verroios, V., Papadimitriou, P., Johari, R., and Garcia-Molina, H. (2015). Client clustering for hiring modeling in work marketplaces. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2187–2196. ACM.
- [Walsham, 1995] Walsham, G. (1995). Interpretive case studies in is research: nature and method. *European Journal of information systems*, 4(2):74–81.
- [Walsham, 2006] Walsham, G. (2006). Doing interpretive research. *European journal of information systems*, 15(3):320–330.
- [Werbel and DeMarie, 2005] Werbel, J. D. and DeMarie, S. M. (2005). Aligning strategic human resource management and person–environment fit. *Human Resource Management Review*, 15(4):247–262.
- [Werbel and Gilliland, 1999] Werbel, J. D. and Gilliland, S. W. (1999). Person–environment fit in the selection process. In *Ferris, Gerald R. (Ed). Research in human resources management*. Elsevier Science/JAI Press.
- [Werbel and Johnson, 2001] Werbel, J. D. and Johnson, D. J. (2001). The use of person–group fit for employment selection: A missing link in person–environment fit. *Human Resource Management*, 40(3):227–240.
- [Zhang et al., 2012] Zhang, H., Horvitz, E., Chen, Y., and Parkes, D. C. (2012). Task routing for prediction tasks. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 889–896. International Foundation for Autonomous Agents and Multiagent Systems.

- [Zikopoulos et al., 2012] Zikopoulos, P. C., Eaton, C., DeRoos, D., Deutsch, T., and Lapis, G. (2012). Understanding big data. *New York et al: McGraw-Hill*.

A

Appendix from Qualitative Research

A.1 Interview Guideline

This interview guideline was used to ensure a similar structure between all interviews.

1. Introduction of myself and the purpose of the interview

- Thesis: part of PhD project
 - developing a crowdsourcing platform for data analysis tasks, where every person, even non-experts, can contribute to the ongoing research as much as he or she can; enabling laypeople to run collaborative data-driven research on any topic of their interest; in order to collaboratively work on data analysis, it is necessary to know what skills various data analysis tasks require and whether crowd workers/freelancers possess these skills
- First conduct interviews
 - use interview results to create a survey for crowdsourcing platforms
- Publish results in master thesis and conference paper
- Guaranteed confidentiality
- Will voice record and might use some quotes
- Can I proceed with the questions?

2. Questions

The underlying RQ of the interview study: What are the skills required for different data analysis tasks?

1. Could you briefly explain what your job is in this organization and how your occupational background looks like?
 - What is your exact job title?
2. How and how long did you learn about data analysis?
3. For how long have you been conducting data analysis?

4. What are typical tasks or phases that you perform in data analysis?
 - What kind of tasks did you perform this/last week?
 - What does a typical day at work look like for you?
5. What skills do you need as a data scientist or data analyst to perform well in these tasks?
6. What specific tools and techniques do you use?
7. What programming languages do you use in your daily job?
8. What are the most frequent, tedious, and repetitive tasks in data analysis?
9. Have you/your company ever outsourced some data analysis?
 - If yes, what?
10. What tasks or phases of data analysis can be potentially outsourced in your opinion?
 - Give an example of such a possible outsourced task. How would you picture such a collaboration?
 - What skills would freelancers need to have to perform this task?
 - What knowledge would be necessary to have?
 - Which tools and techniques would they need to know?
 - Which programming languages would they need to know?
 - You mentioned several skills freelancers would have to possess in order to solve these tasks. Rank them in importance.
11. What are the barriers/difficulties when outsourcing data analysis to freelancers? How do you think these difficulties can be solved?
12. Do freelancers need to understand the structure and the meaning of the data?
13. Do you feel comfortable giving out your data to freelancers?
 - If you were the owner of a data-driven project and it's up to you to decide what to do, would you involve freelancers in a job?
14. Have you experienced difficulties in your job as a data analyst or data scientist and do you think freelancers could have helped you out? How?
15. Could you imagine outsourcing the task of pre-processing data to freelancers?
 - How would you picture such a collaboration?
 - What skills would freelancers need to have to perform this task?
 - What knowledge would be necessary to have?

- Which tools and techniques would they need to know?
- Which programming languages would they need to know?

3. Conclusion

- Thank you
- Any questions left?
- Interested in receiving final research report?

A.2 Flyer for Recruiting Interviewees

This flyer was handed to potential interviewees at Meetups.



Figure A.1: Flyer for recruiting interviewees

A.3 Email for Recruiting Interviewees

This email was sent to potential interviewees through online business networks.

Dear XXX

I am doing my master thesis at the University of Zurich and I am looking for data analysts or data scientists to interview. I found your profile on "Data Science Central".

My thesis is part of a research project envisioning a crowdsourcing platform for data analysis tasks where every person, even non-experts, can contribute to the ongoing research as much as he or she can. We want to enable laypeople to run collaborative data-driven research on any topic of their interest. In order to collaboratively work on data analysis, it is necessary to know what skills various data analysis tasks require and whether crowd workers/freelancers possess these skills.

I am interested in learning about your typical tasks in your daily job and what skills and knowledge are necessary. The interviews will take place in December and January. It would be of great help if you could spare some time (30-40 minutes) for my interview and contribute to our research. All participants have the chance of winning 2 x 2 cinema ticket vouchers.

I would greatly appreciate if I could interview you.

Thank you for your response and kind regards,
Frida Juldasczewa

A.4 Summaries of Interviewees' Background

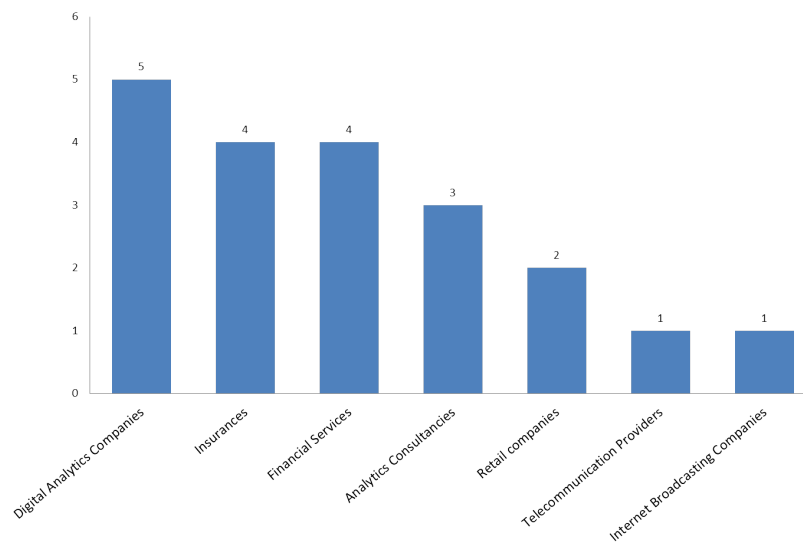


Figure A.2: Affiliations of interviewees, e.g. 5 interviewees were affiliated with Digital Analytics Companies

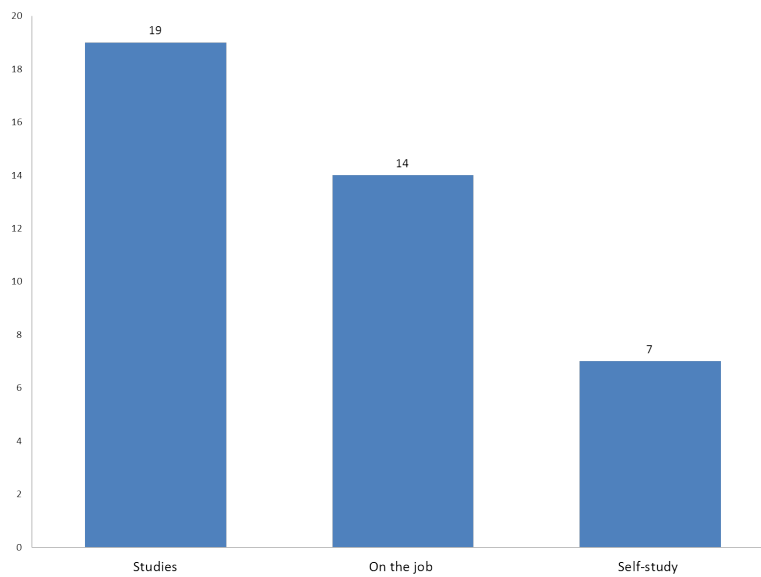


Figure A.3: How interviewees learned about data analysis: Almost all interviewees (19 of 20) learned data analysis in their studies. Many (14) continued to learn on the job, and several (7) made use of e.g. books, online courses, teaching videos, or other resources on the Internet

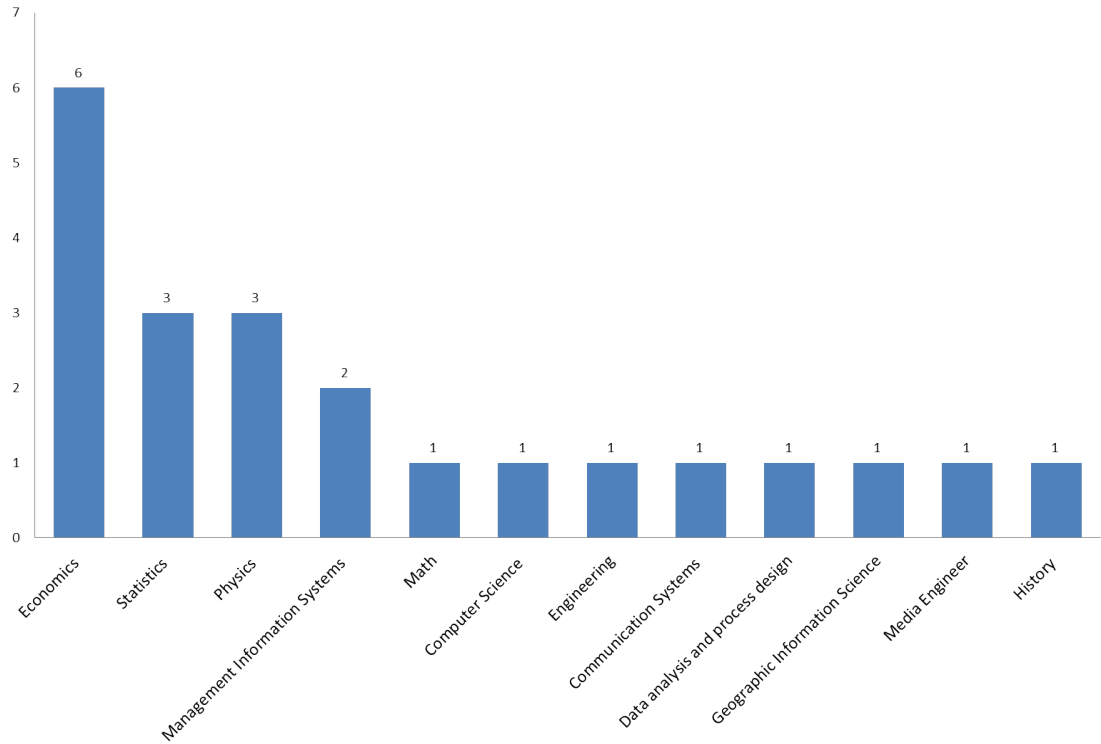


Figure A.4: What interviewees studied to learn about data analysis: Most interviewees (6) studied Economics, followed by Statistics (3) and Physics (3). Other field of studies included often mathematics and computer science classes. History e.g. encompassed many statistics and methodology classes about quantitative and qualitative methods, and how to make inferences from data.

P	Field of Study	Highest Degree
1	Computer Science	PhD
2	Management Information Systems	Lic.oec.publ.BI
3	Statistics	Master
4	Economics (focus Econometrics)	Master
5	Communication Systems	Master
6	Mathematics	Master
7	Economics, Statistics	Master
8	Physics, Economics	PhD
9	Hospitality Management	Bachelor
10	Data Analysis and Process Design	Dipl.Ing.
11	Physics	Master
12	Economics, Statistics	Master
13	Geographic Information Science	Master
14	Economics (focus Statistics)	Master
15	History	PhD
16	Economics (focus Statistics)	PhD
17	Media Engineer	Bachelor
18	Physics	PhD
19	Engineering	Master
20	Management Information Systems, Finance	Master

Table A.1: Educational background of interviewees: Field of study and earned degree

A.5 Tools and Programming Languages used by Interviewees

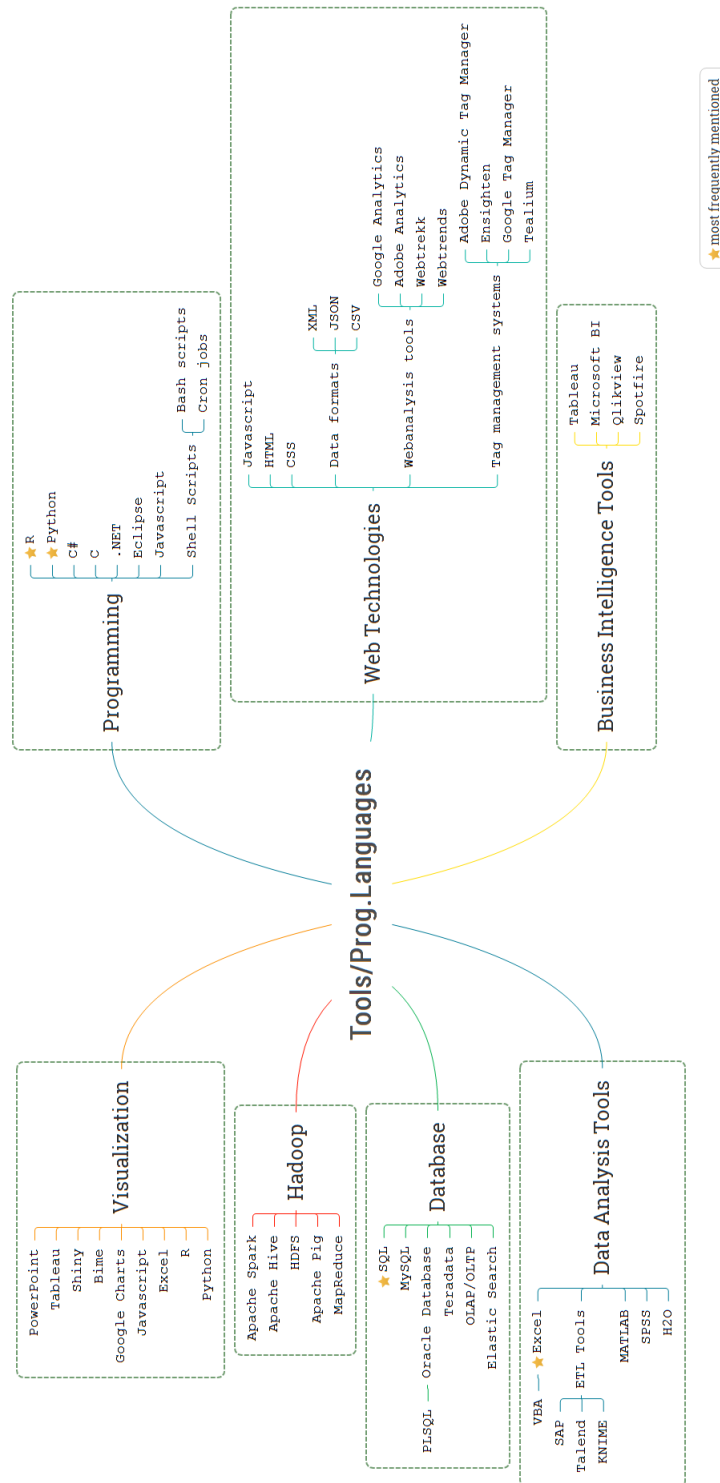


Figure A.5: Tools and programming languages

A.6 Statistical Methods used by Interviewees

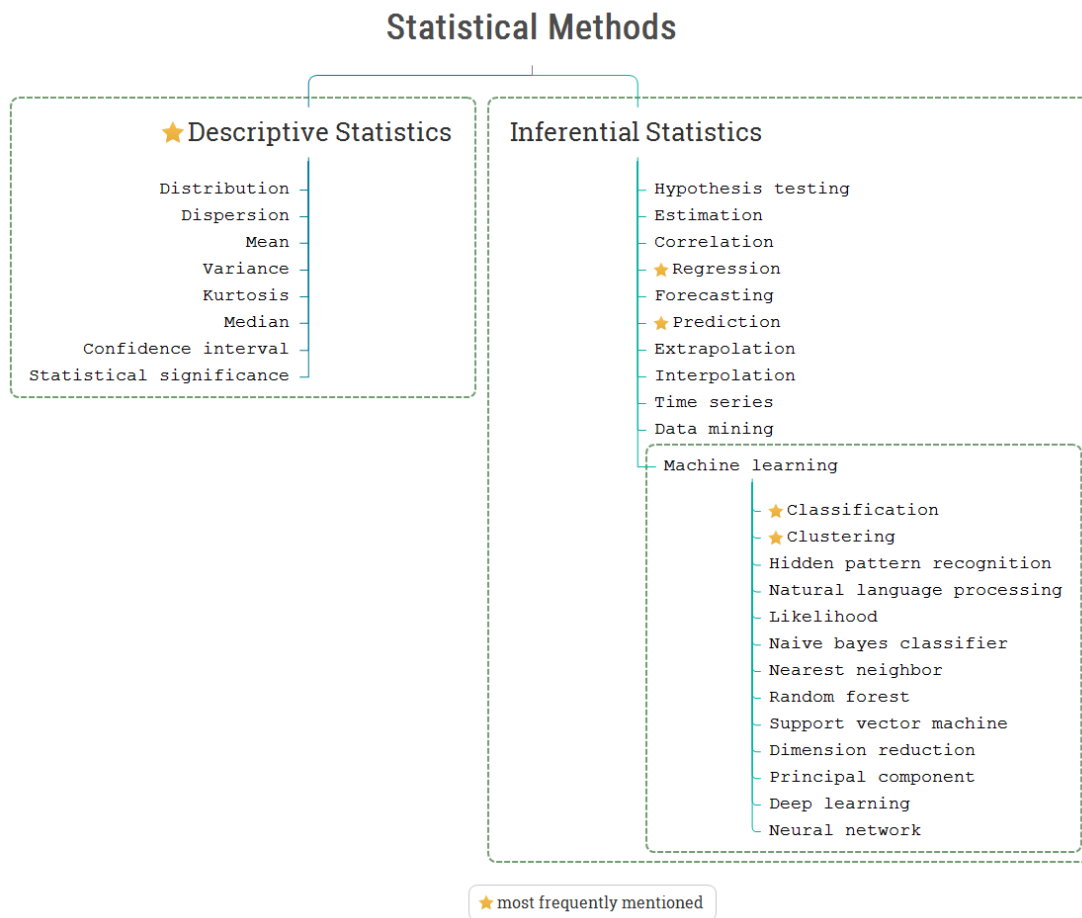


Figure A.6: Statistical methods

A.7 Most Tedious Tasks and Tasks (Not) to be Outsourced

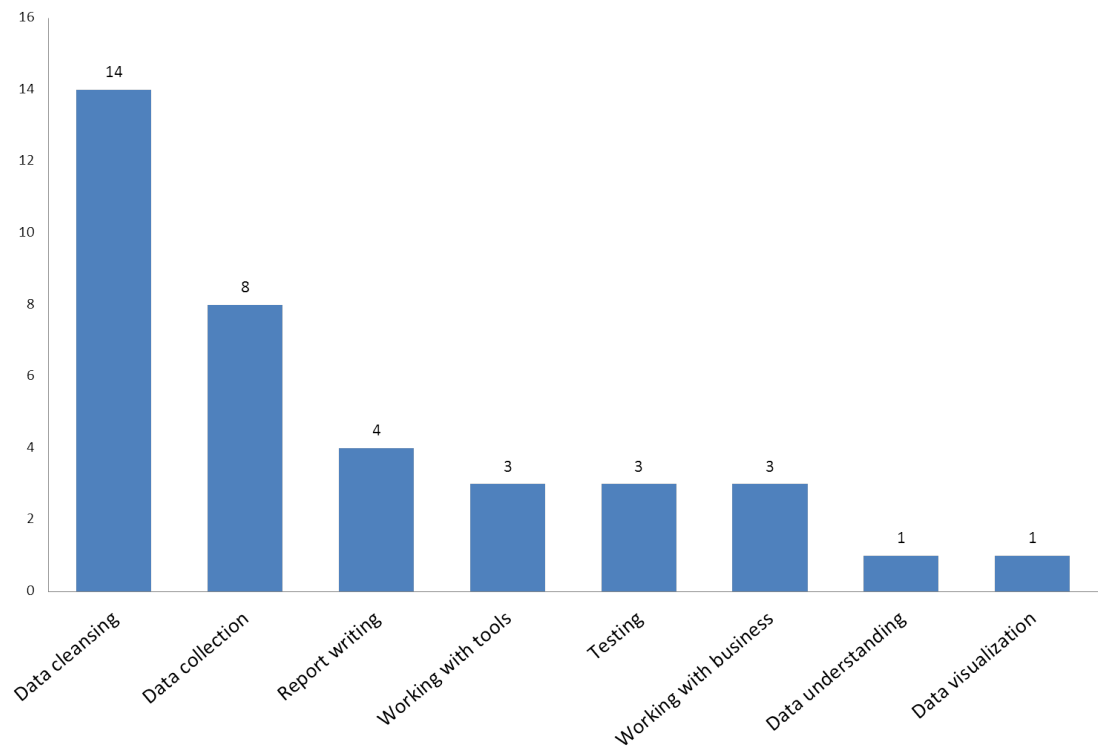


Figure A.7: Tasks interviewees regard as the most tedious: The by far most mentioned tedious task is data cleansing, as stated by 14 interviewees.

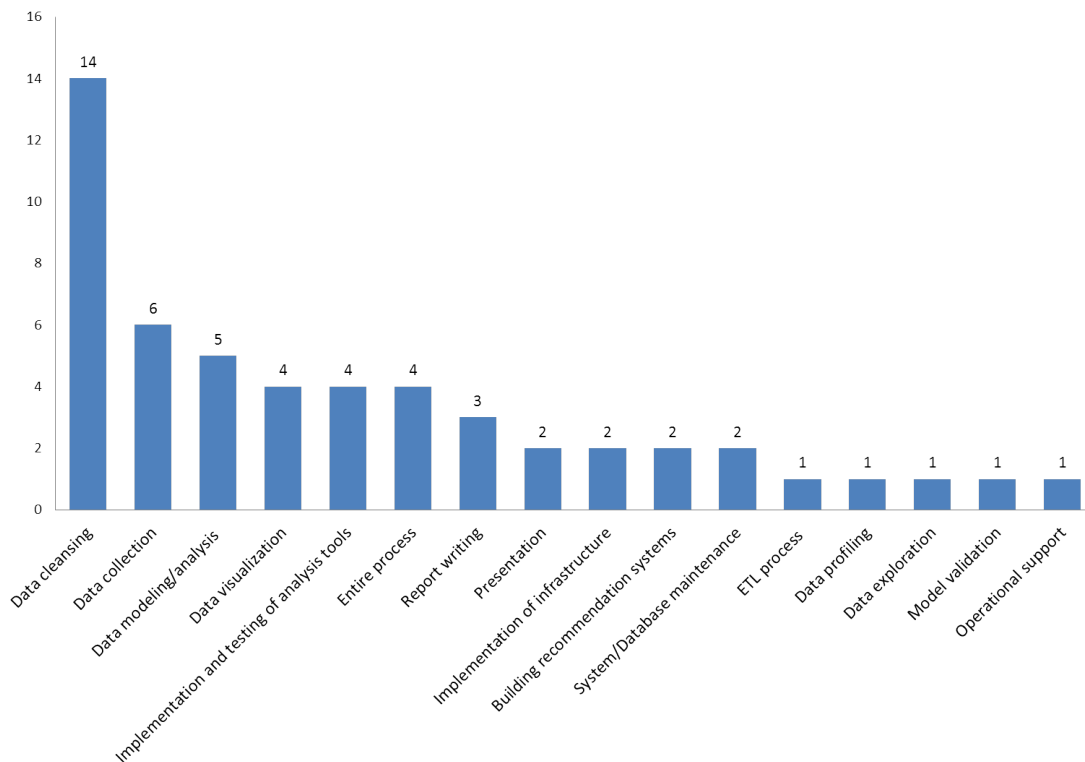


Figure A.8: Tasks interviewees regard as possible to outsource: Again, data cleansing is the most mentioned answer by interviewees (14).

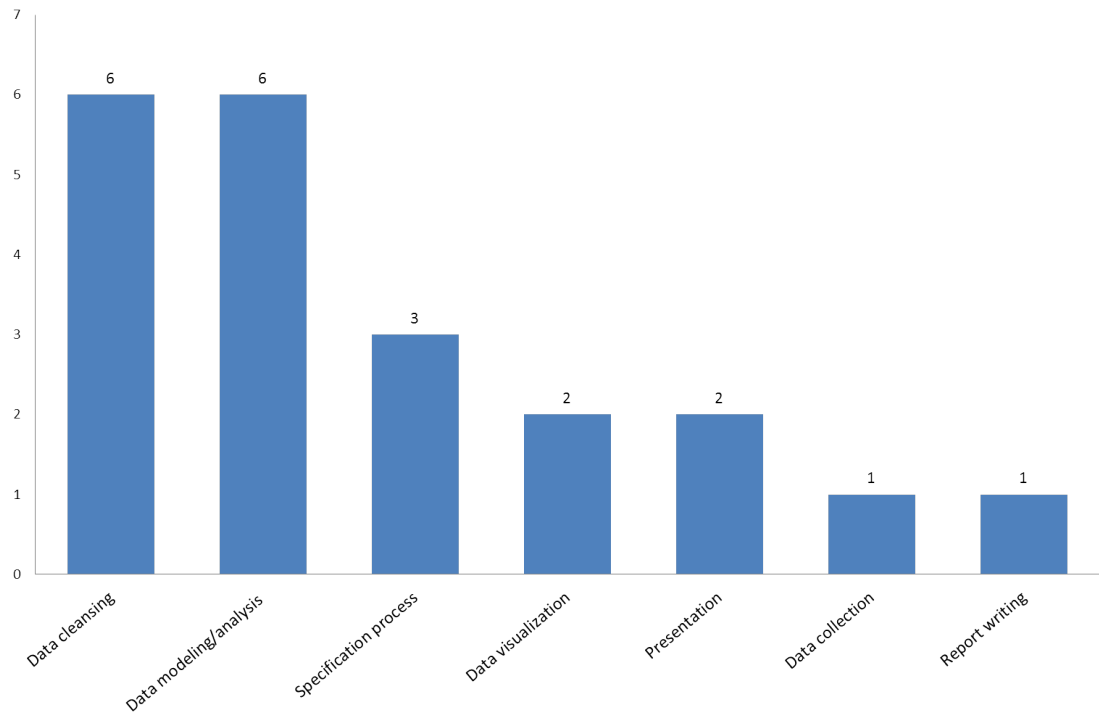


Figure A.9: Tasks interviewees would not outsource: Data cleansing and data modeling were mentioned by most interviewees (6 each) not to outsource.

A.8 Visualizations of Interviewees' Answers

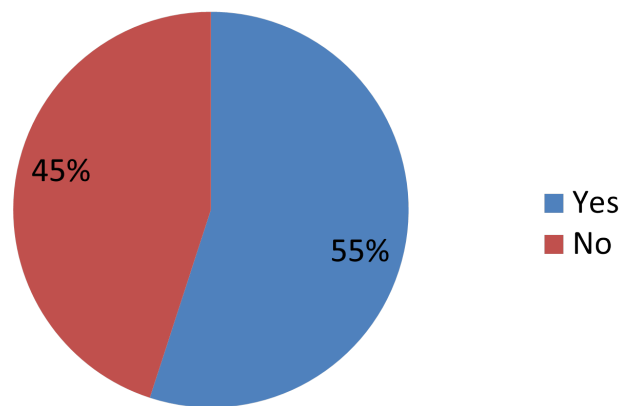


Figure A.10: Interviewees' answers to whether they or their companies had outsourced data analysis before. Reasons for negative answers were e.g. confidentiality issues of the data.

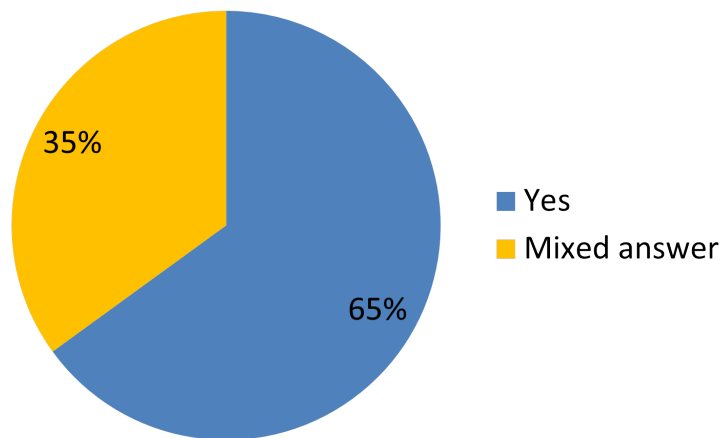


Figure A.11: Interviewees' answers to whether freelancers need to understand the structure and meaning of the data. Mixed answers refer to answers such as that it would depend on the task.

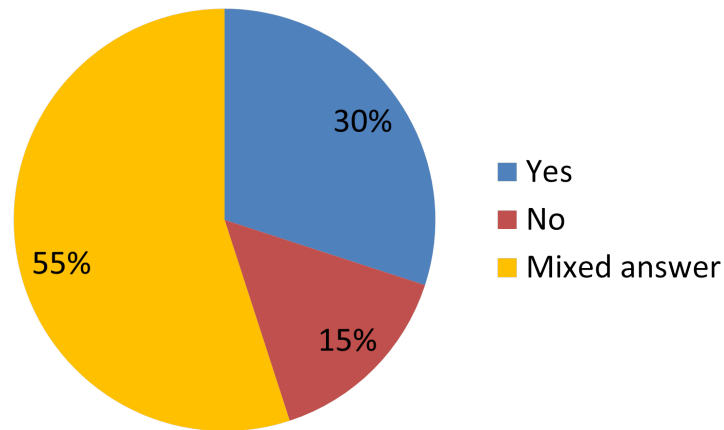


Figure A.12: Interviewees' answers to whether they would feel comfortable giving out their data to freelancers. Mixed answers refer to answers such as that it would depend on the data, e.g. with sensitive data they would not feel comfortable whereas with anonymized data they would.

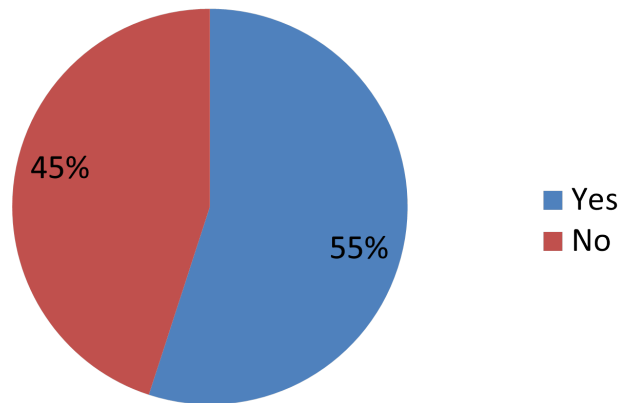


Figure A.13: Interviewees' answers to whether they had experienced difficulties in their job where freelancers could have helped them out. Reasons for affirmative answers were e.g. time-related bottlenecks and lack of own resources and expertise.

A.9 Tabular Summary of Possible Tasks to Outsource, Their Difficulties & How to Solve Them

Task	Mentioned by...		Problematics when crowdsourcing this task according to interviewees	How to solve these problems according to interviewees
	...to be out-sourced	...not to be out-sourced		
Data cleansing	70%	30%	1. Garbage in, garbage out 2. Data could be sensitive, confidential (4) 3. A lot of assumptions are put into data cleansing, it's subjective (2) 4. Data cleansing is a complex and iterative process in terms of: - a lot of customer contact (4) - once the first data cleansing is completed, again data cleansing in the modeling part has to be done (2) - when requirements change, new data has to be cleansed (2) → communicating with freelancer is time consuming and might be even more effort than doing it yourself 5. Data might not be clearly understandable, freelancer has questions 6. Data cleansing has to be done manually, because things have to be first spotted 7. Knowledge gap: Freelancer needs to understand the domain and what the data represents (6) 8. Data amount could be very large 9. Data cleansing is error-prone 10. Data auditability is not guaranteed 11. Freelancer needs a lot of experience 12. If something goes wrong, all following steps are dependent on it 13. Effort for coordination is big: - because client must know exactly how data was cleansed and what assumptions have been made (2) - because client must describe the specifications and process very clearly to the freelancer (3) 14. Trust into freelancers, if they are reliable, deliver on time, data manipulation etc. (3)	- → anonymize data - → there needs to be a tight connection and good communication between the person doing the data cleansing and all other parties → having a contact person at the company for the freelancer - → freelancer needs to have good communication skills and ask client or other various people in order to understand the data → give freelancer only a small batch of data and then apply the algorithm in the big data environment → trust into freelancers and control → process should be standardized and client needs to know how data was cleaned, how it was aggregated etc. → allow freelancer more time to get acquainted with the project → hire freelancer not just for short-term but for longer → outsource only parts that are at the end of the data analysis process → there needs to be an easy way to describe the task accurately, e.g. provide a template of what is expected → close collaboration and good communication between all parties (business, data cleansing person, data analysis person...) → monitoring of freelancers

Data collection	30%	5%	1. Data is spread throughout various sources and freelancers need to have an aggregated overview	-
			2. Data collection is error-prone	→ trust into freelancers and control
			3. Data auditability is not guaranteed	→ process should be standardized and client needs to know where data came from
			4. It's a complex task getting the exactly right data that is required, freelancer might not have access to various systems or people from whom to get the data	→ provide access for freelancer and establish connection to the necessary persons
Analysis/Modeling	25%	30%	1. Run analytics on large volumes of data near real time or real time	→ freelancer needs a lot of computer power, a lot of storage
			2. Background knowledge about the business (why this business is as it is), about data, about the domain (3)	→ freelancer needs business understanding
			3. Close collaboration necessary	→ decide together on methods to be applied, do first analyses and then evaluate and review; also constant meetings & communication
			4. Clear specifications necessary	-
			5. Effort to separate modeling part from the rest of data analysis is big	→ business and data analysis person have to collaborate closely and elaborate requirements together
			6. Data analysis can be directed, statistics can be bend in order to have a statement that is not entirely true	→ control mechanisms
			7. Model and code need to be understandable to the client	→ freelancer needs to communicate clearly on the model
			8. There are various ways to conduct analysis and the understanding about the used methods needs to stay within the company, what assumptions have been made, knowledge about the created model needs to be in-house in case clients have questions about it	→ precise communication
			9. Security of data, trust into freelancer	→ secure with contract
Data visualization	20%	10%	1. Visualization of large amounts of data	→ freelancer needs a lot of computer power, a lot of storage
			2. Good knowledge of what the customer wants and needs in the visualization is necessary	→ good communication necessary
			3. Concept how to visualize data should be precisely defined	→ precise communication

Implementation and testing of analysis tools	20%	-	1. Specifications need to be defined a priori and outsourcing partner needs to be briefed very well and abide to the specifications 2. Time pressure	→ control and final test by the client
Writing reports	15%	5%	Exact specifications what is expected in the report	→ freelancer needs to be flexible in time
Presentation of results	10%	-	Data needs to be understood very well by the freelancer	→ precise briefing and communication → allow time and brief well
System and database maintenance	10%	-	It's a time-consuming and complex task	-
Implementation of infrastructure	10%	-	-	-
Development of recommendation systems	10%	-	Freelancer needs excellent knowledge in machine learning	→ choose freelancer with appropriate skills
Data profiling	5%	-	-	-
Data exploration	5%	-	Needs to happen together with the customer	→ close collaboration
ETL process	5%	-	-	-
Model validation	5%	-	1. Corruption: "here is my money, can you validate the model?" 2. Freelancer needs background knowledge on the model	-
			3. Complex task	→ provide document explaining methodology and applied procedures
			4. Very time-consuming also for the client because everything has to be explained: the data, the model etc.	→ freelancer needs a lot of experience with models and with the business itself → do this only for big projects and not for small projects
Operational support	5%	-	-	-
Specification process/Problem definition	5%	15%	1. Freelancer needs to know a lot about the company, about the data, about the core business, what's most interesting for them 2. Outsourcing this would take away the core key business, the creative part	→ freelancer would have to work very closely with business → company needs to think about its concept itself
			3. Defining the research question is very important for the project	→ close collaboration with business
Only as entire process together	20%	-	1. Difficult to tear apart process because of the understanding of data and making interpretations 2. Very complex process, constantly changing requirements	→ very close collaboration → outsource to only one freelancer

Figure A.14: Difficulties mentioned by interviewees for each task and how to solve them

A.10 Tabular Summary of Difficulties

General problems	Mentioned by...	How to solve these problems according to interviewees
Communication issues (incl. misunderstandings, inefficiencies, non-transparency etc.)	80%	<ul style="list-style-type: none"> Common language Shared understanding Clear requirements and well defined tasks Communicate implicit knowledge explicitly Communicate assumptions clearly Constant communication
Privacy and confidentiality of data	55%	<ul style="list-style-type: none"> Anonymize data Sign non-disclosure agreements Internalize freelancers and don't let them take out data out of the company's environment Trust Give freelancer only part of the data
Knowledge gap	40%	<ul style="list-style-type: none"> Invest more time Give freelancer more time to understand project and gain an overview Freelancer needs to talk to various people Provide a contact person for freelancer Provide background information sheets to freelancers
Quality	40%	<ul style="list-style-type: none"> Quality checks Monitor freelancers Contracts Find and choose good freelancers Trust
Briefing	30%	<ul style="list-style-type: none"> Clear requirements and very precise briefing
Companies don't know what they can do with data analysis	25%	<ul style="list-style-type: none"> Companies need to be guided in the initial phase of exploring the possibilities of data analysis
Trust into freelancers	20%	<ul style="list-style-type: none"> Having a reliable rating system Reliable web profiles and conducting hiring interviews Contracts Safeguards
Vulnerability of data when passed around	20%	<ul style="list-style-type: none"> Safeguards Contracts Don't allow data to leave enterprise environment (e.g. access through VPN)
Meeting of deadlines	15%	<ul style="list-style-type: none"> Good time management
High setup costs and time	15%	-
Danger of data manipulation	15%	<ul style="list-style-type: none"> Safeguards
Outsourced problem needs to be self-contained	10%	-
Reproducibility	10%	<ul style="list-style-type: none"> Agree on common tool to be used by freelancer and client
Large amount of data	10%	<ul style="list-style-type: none"> Provide interface for freelancer to the company's network and working environment, e.g. to the database
Political circumstances in companies	5%	<ul style="list-style-type: none"> Difficult for a freelancer to have influence in a company

Figure A.15: General difficulties mentioned by interviewees and how to solve them

B

Appendix from Quantitative Research

B.1 Pilot Test Questions

The survey was pre-tested with 7 persons. Each person was asked following questions, which were thoroughly discussed:

- How much time did you need to fill out the survey?
- Is the questionnaire understandable and logical?
- Did you encounter any confusing questions?
- Did you encounter any difficult questions?
- Did you have any other problems with some questions?
- Did you encounter missing answer options?
- Did you encounter confusing answer options?
- Do you have any comments on the format and the design of the questionnaire?
- Did you encounter any grammatical or spelling errors?
- Is the introductory text clear?

B.2 Online Survey



University of
Zurich^{UZH}

Freelancer Survey

Dear participant

Thank you in advance for taking the time to complete this questionnaire. Your answers in this academic research experiment will help us better understand the available skills on different crowdsourcing platforms. Therefore, we would like to ask you to take this questionnaire seriously and answer the questions as accurately as possible.

Your responses, along with those from other participants, will be used only for scholarly purposes and will be kept completely confidential.

The survey consists of 29 questions and will not take more than 20 minutes to answer. You will be rewarded with 5 US-Dollars.

Please provide the name of the freelancing platform on which you are performing this survey *

Please provide your username on this freelancing platform *

Please provide a short description (1–2 sentences) of your last task on this freelancing platform *

Have you conducted data analysis or parts of data analysis before? *

- ☐ No
- ☐ Yes

Next

How did you learn about data analysis? *

- ☐ On the job
- ☐ Through university or college courses
- ☐ Through books or e-books
- ☐ Through online courses (for example Coursera or Udacity)
- ☐ Through teaching videos
- ☐ Through the Internet (for example articles or blogs)
- ☐ Other

For how many years have you been conducting data analysis? *

How do you rate your level of expertise in the field of data analysis? *

1 2 3 4 5

Very poor ☐ ☐ ☐ ☐ ☐ Excellent

Back

Next

Which of the following tasks have you already performed during data analysis projects? *

☐ Problem definition

(defining what is the aim of the project, what is the leading question)

☐ Data collection

(gathering data, for example by crawling it from the web, buying it, or retrieving it from a database)

☐ Data profiling

(examining available data and collecting information about it, for example adding keywords and descriptions, giving metrics on data quality)

☐ Data cleansing

(removing or correcting incomplete, incorrect, duplicated, or improperly formatted data)

☐ Data exploration

(summarizing the main characteristics of a dataset to familiarize with it)

☐ Data analysis

(applying statistical software, building models, or aggregating, filtering, and segmenting data)

☐ Model validation

(determining if the built model is correct and appropriate, and where its faults lie)

☐ Data visualization

(plotting figures or developing dashboards)

☐ Results presentation

(presenting results of the analysis project to the customer or management)

☐ Report writing

(creating documents for the customer or management with results of the analysis project)

☐ Entire process from A to Z

(from problem definition/data collection to presentation/reports)

Back

Next

Which of the following tasks in a data analysis project are the most tedious in your opinion? *

☐ Problem definition

(defining what is the aim of the project, what is the leading question)

☐ Data collection

(gathering data, for example by crawling it from the web, buying it, or retrieving it from a database)

☐ Data profiling

(examining available data and collecting information about it, for example adding keywords and descriptions, giving metrics on data quality)

☐ Data cleansing

(removing or correcting incomplete, incorrect, duplicated, or improperly formatted data)

☐ Data exploration

(summarizing the main characteristics of a dataset to familiarize with it)

☐ Data analysis

(applying statistical software, building models, or aggregating, filtering, and segmenting data)

☐ Model validation

(determining if the built model is correct and appropriate, and where its faults lie)

☐ Data visualization

(plotting figures or developing dashboards)

☐ Results presentation

(presenting results of the analysis project to the customer or management)

☐ Report writing

(creating documents for the customer or management with results of the analysis project)

☐ Entire process from A to Z

(from problem definition/data collection to presentation/reports)

☐ I don't know

☐ Other

Back

Next

Which of the following tasks in a data analysis project could be potentially outsourced to online freelancers in your opinion? *

☐ Problem definition

(defining what is the aim of the project, what is the leading question)

☐ Data collection

(gathering data, for example by crawling it from the web, buying it, or retrieving it from a database)

☐ Data profiling

(examining available data and collecting information about it, for example adding keywords and descriptions, giving metrics on data quality)

☐ Data cleansing

(removing or correcting incomplete, incorrect, duplicated, or improperly formatted data)

☐ Data exploration

(summarizing the main characteristics of a dataset to familiarize with it)

☐ Data analysis

(applying statistical software, building models, or aggregating, filtering, and segmenting data)

☐ Model validation

(determining if the built model is correct and appropriate, and where its faults lie)

☐ Data visualization

(plotting figures or developing dashboards)

☐ Results presentation

(presenting results of the analysis project to the customer or management)

☐ Report writing

(creating documents for the customer or management with results of the analysis project)

☐ Entire process from A to Z

(from problem definition/data collection to presentation/reports)

☐ I don't know

☐ Other

Back

Next

Which of the following tasks in a data analysis project should NOT be outsourced to online freelancers in your opinion? *

☐ Problem definition

(defining what is the aim of the project, what is the leading question)

☐ Data collection

(gathering data, for example by crawling it from the web, buying it, or retrieving it from a database)

☐ Data profiling

(examining available data and collecting information about it, for example adding keywords and descriptions, giving metrics on data quality)

☐ Data cleansing

(removing or correcting incomplete, incorrect, duplicated, or improperly formatted data)

☐ Data exploration

(summarizing the main characteristics of a dataset to familiarize with it)

☐ Data analysis

(applying statistical software, building models, or aggregating, filtering, and segmenting data)

☐ Model validation

(determining if the built model is correct and appropriate, and where its faults lie)

☐ Data visualization

(plotting figures or developing dashboards)

☐ Results presentation

(presenting results of the analysis project to the customer or management)

☐ Report writing

(creating documents for the customer or management with results of the analysis project)

☐ Entire process from A to Z

(from problem definition/data collection to presentation/reports)

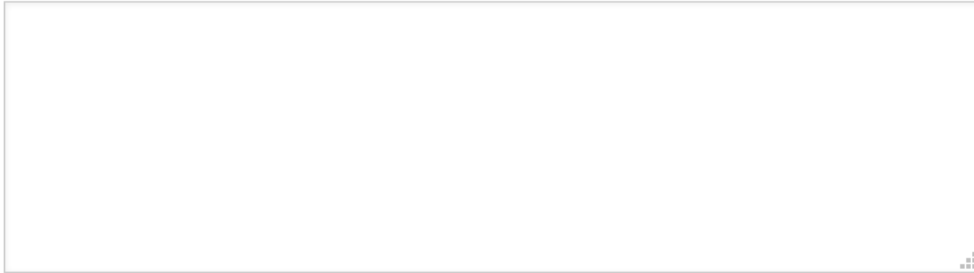
☐ I don't know

☐ Other

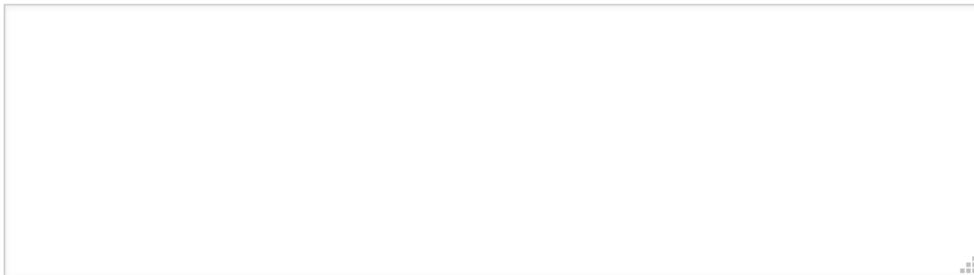
Back

Next

What are the difficulties when outsourcing data analysis tasks to online freelancers? *

A large, empty rectangular text input box with a thin gray border. In the bottom right corner, there is a small icon consisting of three small squares arranged in a triangle.

How can the above mentioned difficulties be solved? *

A large, empty rectangular text input box with a thin gray border. In the bottom right corner, there is a small icon consisting of three small squares arranged in a triangle.

Back

Next

Below is a list of skills.

To what extent do you consider yourself to have these skills? *

	Not at all	To a low extent	To a medium extent	To a high extent	To a very high extent
Statistical skills	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Mathematical skills	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Database skills	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Programming skills	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Data understanding <i>(being able to work with data, having the mindset of knowing what to do with data)</i>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Communication skills <i>(being able to communicate with different groups of interest)</i>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Presentation skills	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Visualization skills <i>(being able to visualize data in a meaningful and simple way)</i>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Machine Learning skills	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Mathematical skills	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Text Mining skills <i>(being able to analyze unstructured information)</i>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Understanding English language	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Documentation and report writing skills	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Below is a list of software tools and programming languages.

To what extent do you consider yourself to be skilled in each of them? *

	Not at all	To a low extent	To a medium extent	To a high extent	To a very high extent
Python	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
R	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Java	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Javascript	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
HTML	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
CSS	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
XML	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
JSON	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
CSV	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
LMNO	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Excel	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Word	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
MATLAB	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
SPSS	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
SAS	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Stata	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Internet	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
ETL Tools	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Talend	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

[Back](#)
[Next](#)

Below is a list of software tools and programming languages.

To what extent do you consider yourself to be skilled in each of them?

(continued) *

	Not at all	To a low extent	To a medium extent	To a high extent	To a very high extent
PowerPoint	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Tableau	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Microsoft Visio	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Microsoft Word	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
ClickView	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Google Analytics	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Adobe Analytics	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
HTML	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Webtrends	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Apache Hadoop	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Apache Pig	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
HDFS	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
MapReduce	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Apache Spark	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Apache Cow	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Excel	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
SQL	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Oracle Database	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Back

Next

To what extent are you able to work with the following operating systems? *

	Not at all	To a low extent	To a medium extent	To a high extent	To a very high extent
Windows	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Linux	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Mac OS	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

[Back](#)[Next](#)

Below is a list of statistical methods.

To what extent do you consider yourself to be skilled in each of them? *

	Not at all	To a low extent	To a medium extent	To a high extent	To a very high extent
Descriptive statistics (for example median or variance)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Inferential statistics	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Hypothesis testing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ingression	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Estimation	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Correlation	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Regressions	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Forecasting	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Prediction	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Extrapolation	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Interpolation	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Time series	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Data mining	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

[Back](#)
[Next](#)

Below is a list of statistical methods.

To what extent do you consider yourself to be skilled in each of them?

(continued) *

	Not at all	To a low extent	To a medium extent	To a high extent	To a very high extent
Classifications	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Clustering	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Correlation	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Hidden pattern recognition	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Natural language processing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Going to the moon	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Likelihood	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Naïve Bayes classifier	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Nearest neighbor	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Network theory	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Bayesian network	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Random forest	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Support vector machine	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Random mountain	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Dimension reduction	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Principal component	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Deep learning	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Neural network	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

[Back](#)
[Next](#)

To what extent do you consider your way of working, your approach to problems to include... *

	Not at all	To a low extent	To a medium extent	To a high extent	To a very high extent
...algorithmic thinking?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
...logical thinking?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
...breaking problems down into smaller parts?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
...deduction?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
...going to planet Mars?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
...keeping the big picture in mind?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

[Back](#)
[Next](#)

In what domains do you have expertise, for example through previous work, studies, or personal interest? *

- ☐ Art & Design
- ☐ Building & Construction
- ☐ Business & Management
- ☐ Economics
- ☐ Engineering
- ☐ Finance
- ☐ Geography
- ☐ Health Service
- ☐ Insurance
- ☐ IT
- ☐ Law
- ☐ Marketing
- ☐ Mathematics & Statistics
- ☐ Media & Journalism
- ☐ Science & Research
- ☐ Telecommunication
- ☐ Tourism & Hospitality
- ☐ Transportation & Logistics
- ☐ Other

To finish, we just need some demographic information, and then you're done.

What is your gender? *

- ☐ Female
- ☐ Male
- ☐ Other

What is your age? *

- ☐ 0 – 17
- ☐ 18 – 24
- ☐ 25 – 34
- ☐ 35 – 44
- ☐ 45 – 54
- ☐ 55 – 64
- ☐ 65 or more

Where do you currently live? *

What is the highest level of school you have completed or the highest degree you have received? *

- ☐ Less than high school diploma
- ☐ High school degree or equivalent (for example Matura or Abitur)
- ☐ Some college but no degree
- ☐ Associate degree
- ☐ Bachelor's degree
- ☐ Master's degree
- ☐ Doctorate degree
- ☐ Other

If you have an Associate, Bachelor's, Master's, or Doctorate degree, what was your field of studies?

- ☐ Architecture
- ☐ Business & Management
- ☐ Communications
- ☐ Computer Science
- ☐ Economics
- ☐ Engineering Sciences (for example Mechanical Engineering, Civil Engineering, Electrical Engineering)
- ☐ Finance
- ☐ Humanities & Arts (for example Languages, Literature, Philosophy, Music)
- ☐ Law
- ☐ Mathematics & Statistics
- ☐ Medicine & Health Sciences
- ☐ Natural Sciences (for example Biology, Chemistry, Physics, Geography)
- ☐ Social Science (for example History, Psychology, Politics)
- ☐ Other

Which of the following categories best describes your employment status? *

- ☐ Employed, working 40 or more hours per week
- ☐ Employed, working 1–39 hours per week
- ☐ Not employed, looking for work
- ☐ Not employed, NOT looking for work
- ☐ Student
- ☐ Homemaker
- ☐ Military
- ☐ Retired
- ☐ Disabled, not able to work

If you are employed, please state your current job title and explain shortly what you are doing in your job

How many hours per week do you spend on freelance work? *

- ☐ 40 or more hours per week
- ☐ 30–39 hours per week
- ☐ 20–29 hours per week
- ☐ 15–19 hours per week
- ☐ 10–14 hours per week
- ☐ 5–9 hours per week
- ☐ 2–4 hours per week
- ☐ 0–1 hour per week
- ☐ Other

What was your last task on this freelancing platform? *

Please identify 5 keywords or tags that represent this survey (separated by commas) *

Please leave any feedback that you consider relevant for this survey

Submit Survey

B.3 Market Share of Upwork and Freelancer

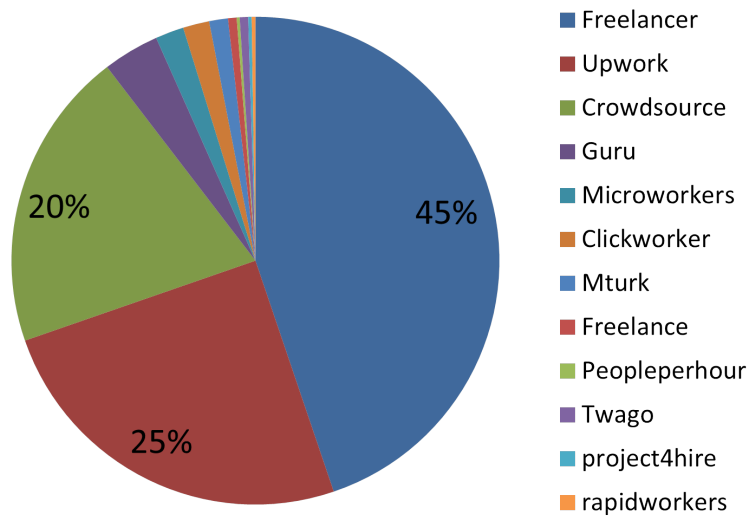
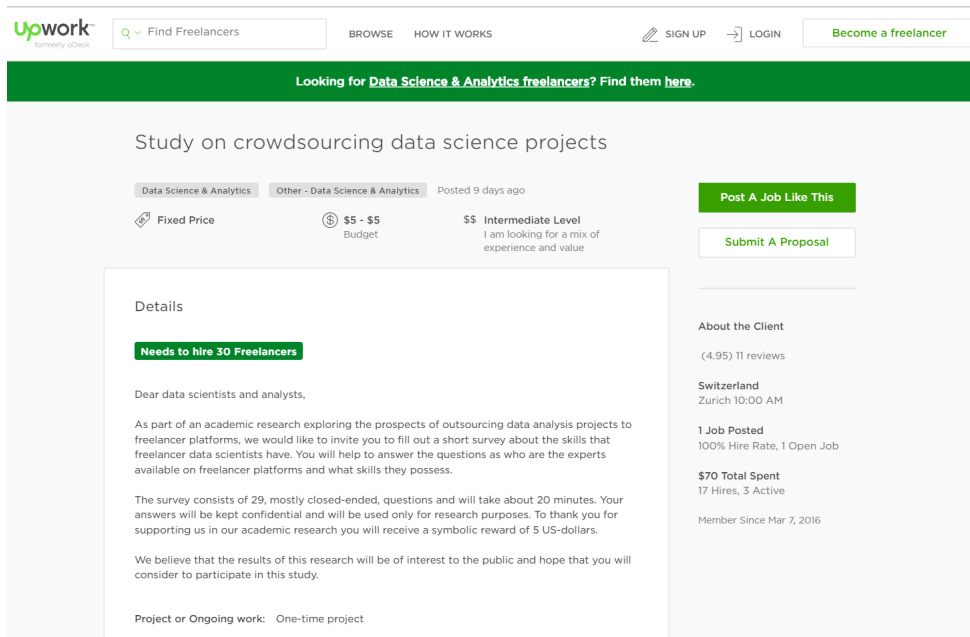


Figure B.1: Upwork and Freelancer account for 70% of the available crowdworkers and freelancers, and thus constitute the biggest online workforce up to date.

B.4 Survey Published on Upwork and Freelancer



The screenshot shows the Upwork interface for a job post titled "Study on crowdsourcing data science projects". The post is categorized under "Data Science & Analytics" and "Other - Data Science & Analytics", posted 9 days ago. It is a fixed-price job with a budget of \$5 - \$5 and an intermediate level. The client is from Switzerland, Zurich, and has a 100% hire rate with 1 open job. The total spent is \$70, with 17 hires and 3 active jobs. The client has been a member since March 7, 2016.

Details

Needs to hire 30 Freelancers

Dear data scientists and analysts,

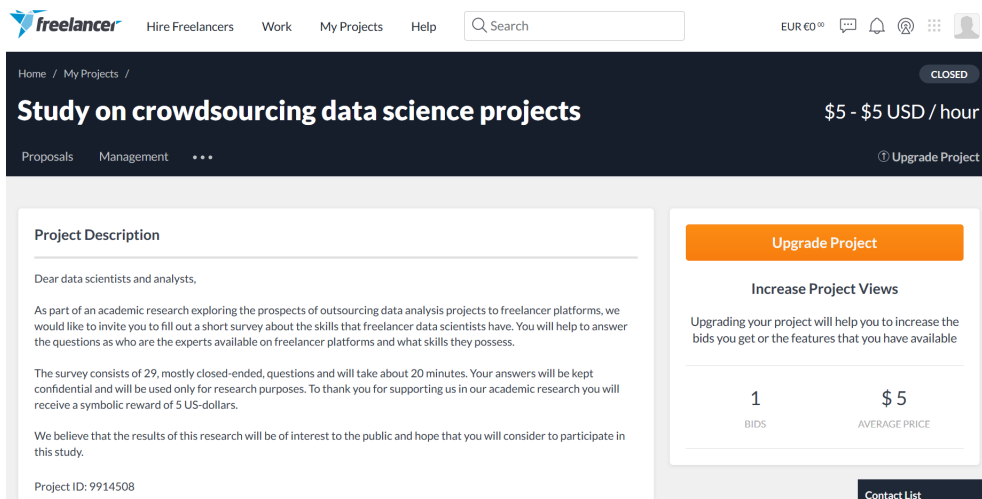
As part of an academic research exploring the prospects of outsourcing data analysis projects to freelancer platforms, we would like to invite you to fill out a short survey about the skills that freelancer data scientists have. You will help to answer the questions as who are the experts available on freelancer platforms and what skills they possess.

The survey consists of 29, mostly closed-ended, questions and will take about 20 minutes. Your answers will be kept confidential and will be used only for research purposes. To thank you for supporting us in our academic research you will receive a symbolic reward of 5 US-dollars.

We believe that the results of this research will be of interest to the public and hope that you will consider to participate in this study.

Project or Ongoing work: One-time project

Figure B.2: Survey on Upwork



The screenshot shows the Freelancer interface for a job post titled "Study on crowdsourcing data science projects". The post is categorized under "Data Science & Analytics" and "Other - Data Science & Analytics", posted 9 days ago. It is a fixed-price job with a budget of \$5 - \$5 and an intermediate level. The client is from Switzerland, Zurich, and has a 100% hire rate with 1 open job. The total spent is \$70, with 17 hires and 3 active jobs. The client has been a member since March 7, 2016.

Project Description

Dear data scientists and analysts,

As part of an academic research exploring the prospects of outsourcing data analysis projects to freelancer platforms, we would like to invite you to fill out a short survey about the skills that freelancer data scientists have. You will help to answer the questions as who are the experts available on freelancer platforms and what skills they possess.

The survey consists of 29, mostly closed-ended, questions and will take about 20 minutes. Your answers will be kept confidential and will be used only for research purposes. To thank you for supporting us in our academic research you will receive a symbolic reward of 5 US-dollars.

We believe that the results of this research will be of interest to the public and hope that you will consider to participate in this study.

Project ID: 9914508

Upgrade Project

Increase Project Views

Upgrading your project will help you to increase the bids you get or the features that you have available

1	\$5
BIDS	AVERAGE PRICE

Contact List

Figure B.3: Survey on Freelancer

B.5 Visual Summaries of Survey Data

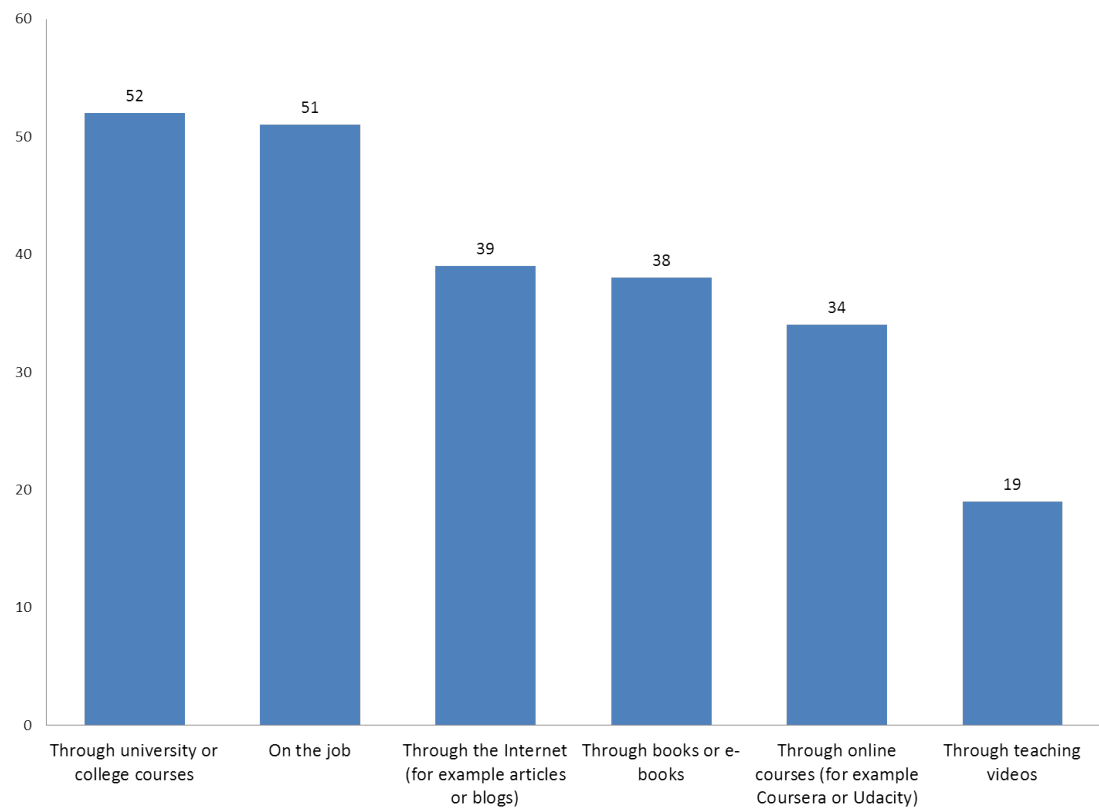


Figure B.4: How survey participants learned about data analysis: Most freelancers learned it through university (52 of 80) and on the job (51 of 80)

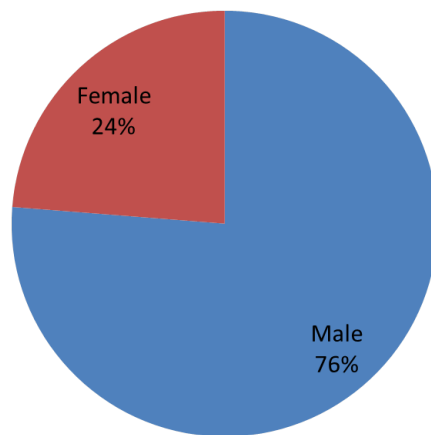


Figure B.5: Gender of participants

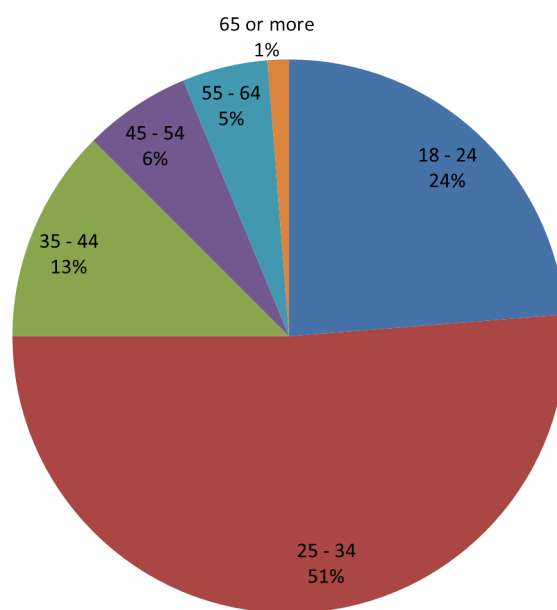


Figure B.6: Age of participants

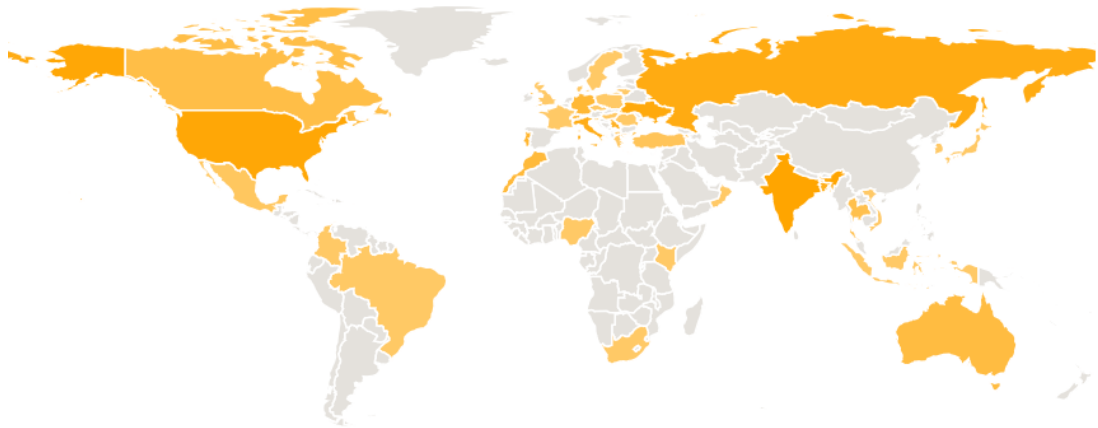


Figure B.7: Choropleth map showing survey participants' distribution in the world

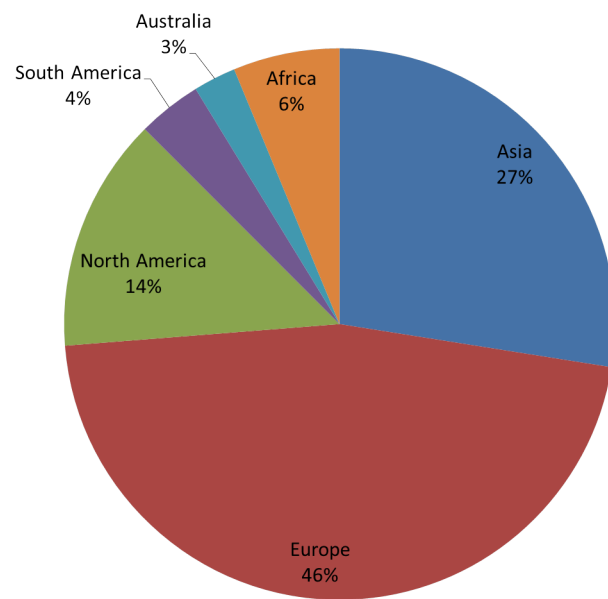


Figure B.8: Country residencies of survey participants

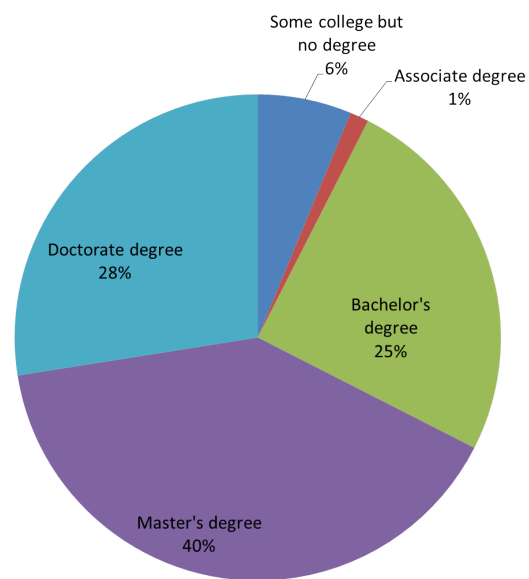


Figure B.9: University degrees of participants

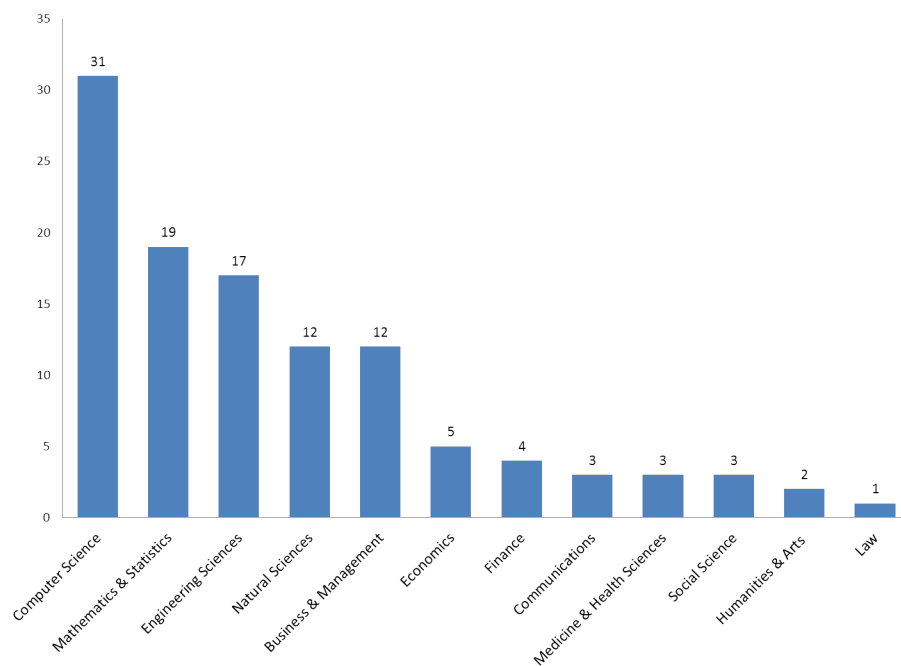


Figure B.10: What survey participants studied: Most freelancers studied Computer Science (31), followed by Mathematics & Statistics (19) and Engineering Sciences (17).

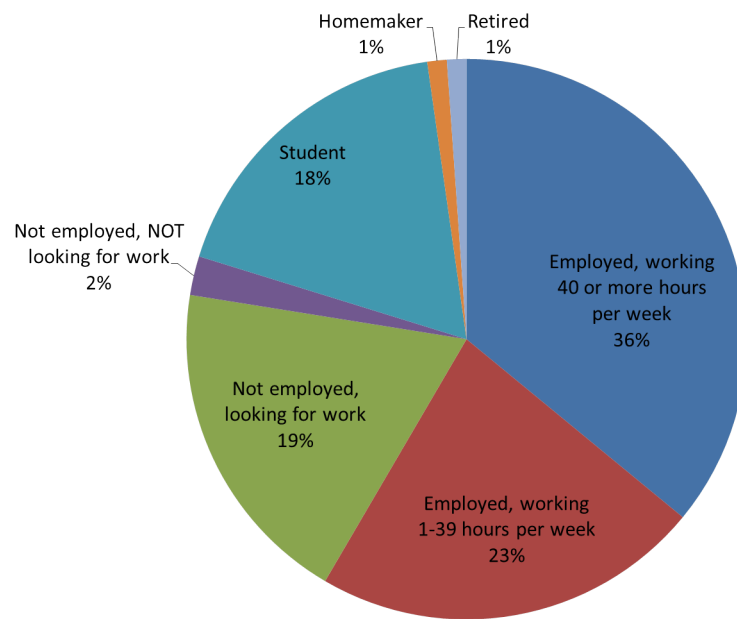


Figure B.11: Employment status of participants

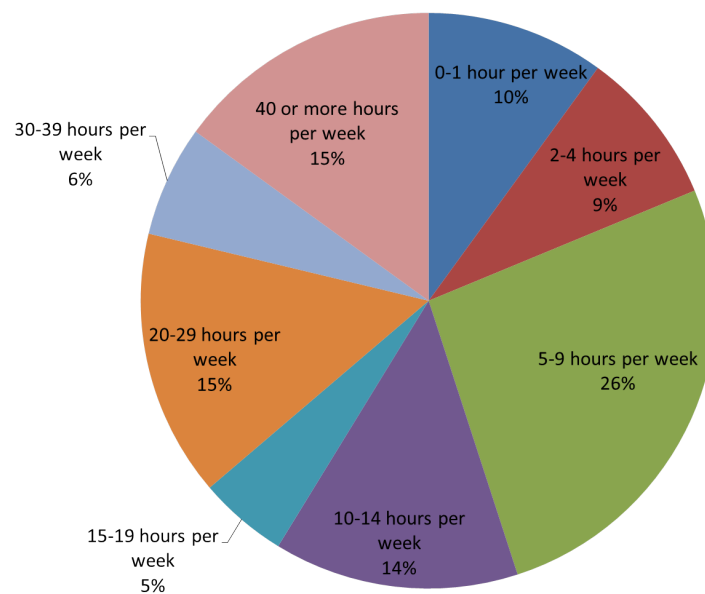


Figure B.12: How many hours participants spend on freelance work

List of Figures

5.1	Interviewees' years of experience depicted as a histogram: The majority of interviewees had between 4 and 10 years of experience with data analysis. Only few (3) had more than 10 years of experience.	20
5.2	Interviewees' skills: Almost every interviewee (18 of 20) said to have mathematical and statistical skills, followed closely by programming skills (15). More than half of interviewees (11) also stated to have communication, database, and domain knowledge, which are important for being a good data scientist.	22
5.3	Skills freelancers should have: According to interviewees, freelancers necessarily need mathematical and statistical skills (16), technical skills such as database (16) and programming skills (13), and domain knowledge (13).	27
5.4	Difficulties: Interviewees see communication issues (16) as the biggest hurdle in outsourcing data analysis tasks, followed by privacy and confidentiality issues (11).	30
5.5	Participants' years of experience and level of expertise in data analysis	31
5.6	What tasks survey participants have already performed, regard as the most tedious, and would (not) outsource	32
5.7	Domain expertise of survey participants	33
A.1	Flyer for recruiting interviewees	55
A.2	Affiliations of interviewees, e.g. 5 interviewees were affiliated with Digital Analytics Companies	57
A.3	How interviewees learned about data analysis: Almost all interviewees (19 of 20) learned data analysis in their studies. Many (14) continued to learn on the job, and several (7) made use of e.g. books, online courses, teaching videos, or other resources on the Internet	57
A.4	What interviewees studied to learn about data analysis: Most interviewees (6) studied Economics, followed by Statistics (3) and Physics (3). Other field of studies included often mathematics and computer science classes. History e.g. encompassed many statistics and methodology classes about quantitative and qualitative methods, and how to make inferences from data.	58
A.5	Tools and programming languages	60

A.6	Statistical methods	61
A.7	Tasks interviewees regard as the most tedious: The by far most mentioned tedious task is data cleansing, as stated by 14 interviewees.	62
A.8	Tasks interviewees regard as possible to outsource: Again, data cleansing is the most mentioned answer by interviewees (14).	63
A.9	Tasks interviewees would not outsource: Data cleansing and data modeling were mentioned by most interviewees (6 each) not to outsource.	64
A.10	Interviewees' answers to whether they or their companies had outsourced data analysis before. Reasons for negative answers were e.g. confidentiality issues of the data.	65
A.11	Interviewees' answers to whether freelancers need to understand the structure and meaning of the data. Mixed answers refer to answers such as that it would depend on the task.	65
A.12	Interviewees' answers to whether they would feel comfortable giving out their data to freelancers. Mixed answers refer to answers such as that it would depend on the data, e.g. with sensitive data they would not feel comfortable whereas with anonymized data they would.	66
A.13	Interviewees' answers to whether they had experienced difficulties in their job where freelancers could have helped them out. Reasons for affirmative answers were e.g. time-related bottlenecks and lack of own resources and expertise.	66
A.14	Difficulties mentioned by interviewees for each task and how to solve them	69
A.15	General difficulties mentioned by interviewees and how to solve them . . .	70
B.1	Upwork and Freelancer account for 70% of the available crowdworkers and freelancers, and thus constitute the biggest online workforce up to date. .	92
B.2	Survey on Upwork	93
B.3	Survey on Freelancer	93
B.4	How survey participants learned about data analysis: Most freelancers learned it through university (52 of 80) and on the job (51 of 80)	94
B.5	Gender of participants	95
B.6	Age of participants	95
B.7	Choropleth map showing survey participants' distribution in the world . .	96
B.8	Country residencies of survey participants	96
B.9	University degrees of participants	97
B.10	What survey participants studied: Most freelancers studied Computer Science (31), followed by Mathematics & Statistics (19) and Engineering Sciences (17).	97
B.11	Employment status of participants	98
B.12	How many hours participants spend on freelance work	98

List of Tables

5.1	T-test and descriptive statistics of survey participants' skills	34
5.2	T-test and descriptive statistics of survey participants' tool, programming, and data format expertise	35
5.3	T-test and descriptive statistics of survey participants' statistical methods expertise	36
6.1	Comparison of data scientist skills, freelancer skills required according to data scientists, and existing skills on freelance platforms	40
A.1	Educational background of interviewees: Field of study and earned degree	59