University of
Zurich^UZH

Understanding the population bias of Stack Overflow survey respondents

Tanbir Mann
Zurich, Switzerland
16-744-807

Supervisor: Prof. Aniko Hannak & Dr. Johannes Wachs
Date of Submission: September 15th, 2020

University of Zurich
Department of Informatics (IFI)
Andreasstrasse 15, CH-8050 Zurich, Switzerland

MASTER THESIS – Social Computing Group, Prof. Aniko Hannak

Master Thesis
Social Computing Group
Department of Informatics (IFI)
University of Zurich
Andreasstrasse 15, CH-8050 Zurich Switzerland
URL: https://www.ifi.uzh.ch/en/scg.html

Abstract

Stack Overflow (SO) is recognized as a technical knowledge-sharing market where goods and services are merely based on asking questions and providing answers. The majority of the questions are related to technology and coding problems. Each year SO publishes a survey with an idea to reach out to coders across the world and to gain more insight into its users and their experience on the platform. SO does everything to serve the needs within the developers' community. The data obtained from the annual survey helps to make changes and set goals to improve the environment and make it more welcoming and inclusive of the SO community. The community does only include visitors to SO, but also everyone who codes or does some coding in their work or studies. The survey questionnaire starts with questions about user demographics, coding interests, current company experience, preferences for different coding languages, and getting feedback on leading technologies of the time. To maximize the accuracy of results, the platform has a minimum threshold for total time spent by each candidate in completing the survey. The platform provides a census badge to its users after completing the survey. The badge falls under the silver badge category and exhibits high reputation scores. This study is a quantitative attempt to understand the differences among the users who participate in the developers' survey to the ones who do not participate. We wanted to identify the key factors that may influence participation in the survey to gain better understanding of the population that takes the survey and how they differ - if at all - from the rest of the community. It also aimed to help us understand the attitude of underrepresented groups such as women and non-active users towards the developers' survey. Our findings suggested that the majority of survey respondents belonged to the community of users with high reputation scores on the website. The users with high tenure on the website were also more likely to participate in the survey. The self-promoters - users who actively promote themselves on the website, and also on other social media platforms such as LinkedIn, Twitter, and Github - were among the majority of survey participants. In terms of user activity on the website, 85% of the survey participants were active answer providers. We also aggregated the participation from the level of participation of the users to the level of geographical regions and learned that the users from the continent of Oceania were the principal contributors in the survey, followed closely by those from Africa and South America; Europe came a distant fourth, and the contribution rate of users from Asia was the lowest of all. We could not find statistically significant results for users from North America. It is inquisitive for us see if the census badge leads to more participation. Our study suggests that in 2017, 40% of the respondents claimed the badge, followed by 50% in 2018 and 60% in 2019. The participation-to-badge-claim ratio has increased by 10% each year from 2017 to 2019.

Acknowledgments

I would like to thank my project supervisors Johannes Wachs and Prof. Aniko Hannak for unconditional help during the thesis which has been essential for its successful completion. I would also like to take the opportunity and thank my family for their support during my studies.

Contents

Chapter 1

Introduction

This chapter gives a short introduction about the topic of interest and research focus along with the motivation behind the selection of this topic, followed by already existing work in the field and thesis outline.

1.1 Motivation

It is fair to say that online communities and social networking sites are becoming the primary source of knowledge sharing in the open-source world. Some of the best-known names in these communities are GitHub and Bitbucket (providing the ability to share content such as source code fragments), Flickr (images), Reddit (news posting) and Stack Overflow (SO). SO is recognized as the most prominent hub of knowledge sharing. It is based on a give-and-take mechanism: one gives some of their knowledge to the platform and takes some from the platform via questions and answers. SO is very useful to the people who seek help from tech experts in the IT industry. When one publishes a post (asks a question) on the platform, it is a request for support from other contributors of the community. SO is also known to be a hub of IT experts and developers and is a useful source for solving coding-related problems.

Maintaining a position in this fast-paced and fast-changing IT world, it is important for platforms like SO to know their users and to have a thorough knowledge of their demands and the level of interaction with the platform. This approach helps them to identify whether or not the platform is welcoming for everyone or if there are any urgent issues to address. SO continues to be built by developers for developers and to serve the needs of the developer community. Whether that is helping visitors find a new product or job, or improving communication with Stack Overflow for Teams, the aim is to serve the developer community as a whole. In order to do this the platform requires a full understanding of developers' needs. In order to accomplish this, the platform publishes a web survey called Stack Overflow Developer Survey. A web survey is a simple means of gaining access to a large group of potential respondents. Questionnaires can be distributed at a very low price with no mailing or printing costs. The distribution range of web surveys is also very wide. They can be launched very quickly, and the waiting time until the questionnaire is ready for fieldwork is also very short.

SO conducts its web survey - the Developer Survey - every year, which is the largest and most comprehensive survey in the world for people who code. The survey started as an attempt to learn more about the community and how its representatives work. Now, it has evolved into an asset used by many professionals and publications around the world to know more about the prevailing technologies and their popularity among developers. It also helps to analyze the happiness of technical employees, and even to address the gender and cultural gaps in the world of coding. The survey questionnaires cover important aspects from a developers' point of view. The survey includes questions related to technology, coding habits, and respondents' preferred work environments - for example, questions about favourite technologies, knowledge of computer languages, favourite coding languages, and the years of experience in the IT field.
From 2017 onwards, SO has adopted a reward system: the approach focuses on attracting more people to participate in the survey. For every successful completion of the survey, participants receive the 'census badge' as a reward. To claim this badge, a user has to log in or register themselves on the website, and this badge will be shown on the user's profile along with other badges in their SO account. The badge also adds to one's scores or so-called 'reputation' on the website. The badge-receiving mechanism and the reward system of SO is explained exhaustively by A. Marry, J. Wachs and A. Hannak in 'Gender differences in participation and reward on Stack Overflow' [2].

Despite the survey's broad reach and capacity for forming valuable conclusions, the website acknowledges that their results do not equally represent everyone in the developers community (e.g., new developers, female developers, experienced developers, and experts in the field), which brings us to the goal of this thesis. In this work, we are trying to understand the characteristics of survey participants. We want to know about the difference between profiles of users who participate in the survey and those who do not. The study is also an attempt to understand the response of under-represented groups (whether defined by gender, race, or geo-location) [1] and to know if the survey results suffer from sampling bias. We want to explore the similarities and differences in the characteristics and behaviour of survey respondents and non-respondents, classify the key factors that may influence user participation, and ascertain whether participation varies concerning socio-demographics. Ultimately, the main motivation is to confirm if the Developer Survey is helping to serve the sole purpose of the SO goal - reaching out to every developer in the world and gathering relevant data - or whether the findings suffer from any social biases.

1.2 Related Work

In this section, we outline related work on motivation to participate in online platforms. We have gathered studies related to gender differences and barriers on platforms like SO and how it influences interaction with the platform. We have also gathered studies related to activity comparison of users with other functional social accounts and without any other functional social accounts.

Motivation to contribute in online platforms The contributions of users to SO originate around the world, with Europe and North America being the principal and almost equal contributors, followed by Asia at distant third, mainly represented by India, and Oceania contributing less than Asia but more than South America and Africa together [21].The motivation behind contributing in online question-answers platforms is analyzed in one of the studies related to Wikipedia; findings suggests that the primary reason for contributing is to 'educate humanity or boost awareness', followed by the motivation to feel like the contributor is making a difference the society. Very few respondents cited a willingness to give back to the community and to build a credible online reputation for themselves. The virtual realm of Wikipedia commends its participants in ways that are unrivalled by most formulations in the non-virtual world. Wikipedia contributors (or 'Wikipedians') enjoy a sense of accomplishment, community, and altruism while working with exceptional freedom and ease, and further they suggest that the primary reason for not contributing is the lack of time or knowledge of subject matter [9].
In an attempt to understand the participation of website users, Adaji I and Vassileva. J have suggested that most users do not complete their online profile; yet, the average reputation scores are higher for users that have complete profiles than those with incomplete profiles. Users with complete profiles tend to post high-quality answers to questions and are more valuable to the community [10]. The study measured the impact of the badge system, an approach to rewarding users for their contribution based on gamification fundamentals; it confirmed that badge value and gamification effectively stimulated voluntary participation [14].

Gender behaviour and barriers on online platforms In a study called 'Gender differences in Professional Self-Promotion', findings suggested that women are less likely to utilize the data fields and write a summary about their job interest and professional experience. Most leave their summary and job description fields empty [5]. Women, in general, have been found to suffer from more barriers than men - specifically, in five barriers at a rate significantly higher than that of males. Some of these barriers included doubts in the expertise field, doubts in the level of knowledge needed to contribute, feeling overwhelmed when competing with a large number of peers, and limited awareness about the features of the website. There are further barriers that equally impacted all SO users and affected particular groups, such as industry programmers [1]. In a study on Wikipedia, women's probability of contributing was lower than that of men. This was found to be due to a lack of self-confidence; female users reported thinking they do not have enough knowledge or expertise to inform both theory and practice [6]. Men provided more answers than women on the site and were rewarded more on average for their answers, even when controlling for possible confounding variables such as tenure. Women asked more questions on the site and gained more rewards per question [2]. The participation and engagement with SO of females was found to be greatly influenced by whether or not they encountered contributions from other females, a phenomenon known as peer parity [7].

Reputation building and its effects on behaviour of the users One of the studies showed that activities that helped to build a reputation quickly on SO included answering questions related to tags with lower expertise density, answering questions promptly, being the first to answer a question, being active during off-peak hours, and contributing to diverse areas [4]. Studies have shown that the communities differ concerning the personality properties of authors with top, medium, and low reputations. As shown in a study called 'On the Personality Traits of StackOverflow Users', authors with top, medium, and low reputations differed in terms of neuroticism, extroversion, openness, agreeableness, and conscientiousness. The findings suggest that users with top reputations are less neurotic, more extroverted, and more open compared to those with medium and low reputations. Furthermore, users associated with posts tagged 'Android' exhibited more neuroticism than authors with posts tagged as 'Java', 'JavaScript' and 'PHP'. Authors with tags 'C#' and 'PHP', also exhibited less neuroticism and extroversion in comparison to authors of 'Java', 'JavaScript' and 'Android' [11].

Activity comparison on SO platform and findings from the developers survey
Previous research on the association between software development and crowd-sourced knowledge on SO and GitHub has proposed that users on SO with active GitHub activities and development process provide more answers and ask fewer questions. Furthermore, the active SO question-posters have been found to administer their work in a comparatively less uniform way than developers that do not ask questions. There is a high correlation between the SO and GitHub activity rates [12]. Another study on the contribution revealed that knowledge is formed and administered in two forms: one is participation where various users collaborate to create and build knowledge. The second is crowd-sourced, where users mainly work autonomous of one another. The participation impression suggests that the users who are progressive in both channels are the majority of answer providers, serving as a hub of knowledge providers [13]. The Developer Survey from SO gathered data about the participants and made the anonymized results available for download. A study that gathered insights from the SO developers' survey found that diversity in a company is not an attentive means when it comes to decision-making considerations for developers who assess a new job opening. The survey respondents from underrepresented groups tended to believe that they were not as good as their fellow peers, resulting in conscious bias [8].

1.3 Thesis Outline

Chapter 2 describes the design of the research approach, along with the selected hypothesis for the study. Chapter 3 details the data collection processes and creation of features out of the raw data, while Chapter 4 defines various methodologies and algorithms used to provide the final results. In Chapter 5, the results and findings from the applied methods and algorithms are analyzed. Chapters 6 and 7 conclude with possible directions for future research.

Chapter 2

Research Design

This section demonstrates the design of this thesis following the Goal-Question-Metric (GAM) approach [15] as used in Bogdan. V and Andrea. C [16].

2.1 Goal

The objective of this work is to understand the demographic differences among SO users and SO survey respondents.

Rationale: The rationale for this is that studies have shown that the web survey is a promising, attractive, and robust instrument for gathering data in a fast and efficient way, but also that it often suffers from various biases. These factors could include limited internet accessibility resulting in certain groups being under-represented. The participation in such surveys is also based on individuals' self-selection. Both of these two factors could strongly influence the results of a survey. The web survey has also been found to have methodological problems, as in Jelke Bethlehem's *Selection Bias in Web Surveys* [17]. The paper acknowledges that the web survey is a promising, attractive and stable means for gathering data, but that there are various factors that could dominate the estimates from a study.

Our aim is to identify the factors that may lead to unreliable results due to self-selection and other potential problems with uneven coverage. Therefore, to fully understand which factors may influence the respondents of the SO survey, qualitative and quantitative studies are needed.

2.2 Questions

The following research questions are addressed in this thesis.

RQ$_1$. ***What socio-demographic factors influence the participation of users in the survey?***

Rationale: the identification of socio-demographic factors is crucial to identify significant differences between the two populations: SO website users and SO survey respondents. We were interested in knowing the differences in gender, experience, and geography of our two sample groups. This segmentation might offer insights that would have been missed by only looking at the aggregate data. One could miss the divergence in the data; therefore, demographics had to be considered in order to account for diversity. We would be able to address factors that may have influence over the survey results.

RQ$_2$. *What is the participation ratio of under-represented groups in the survey? Is the survey biased towards certain groups of users?*

Rationale: we care about understanding the accuracy of the survey results. The participation of respondents is often based on self-selection, but it is always challenging to identify the factors that lead to under-representation of a particular community of users on online platforms. To identify such problems with uneven coverage, we have created certain features that might have an effect on participation in the survey. One of these features is the reputation of the user on the website. We are also interested in knowing if the group of users who actively participate on the website are the core respondents of the survey. Addressing this research question also helps to measure the response of under-represented groups on the SO website such as women [2].

RQ$_3$. *How motivational is the "Census Badge" for the participation?*

Rationale: since 2017, SO has included the option of obtaining a census badge: after completing the survey, this badge can be claimed and added to the participant's SO user profile, improving their scores on the website. We are interested in knowing if this gamification has influenced participation in the survey and in quantifying the ratio among the number of users who participated and the number of users who obtained the badge.

2.3 Metrics

Following are the metrics we use to represent the differentiation among our two groups.

- Users: the number of users registered on the SO website;
- Respondents: the number of users who have participated in the SO survey and are census-badge holders. Participation is signalled by the census badge;
- Non-Respondents: the number of users who never participated in any of the SO surveys. Non-participation is signalled by lack of census badge;
- Active Users: the number of users who are active participants on the website. Participation occurs when a user proposes a new question or attempts to answer an existing one.

With respect to the questions and metrics formulated above, we postulate a number of null hypotheses (reported in **table 2.1**)), to be tested statistically. The most relevant and important test for a non-parametric statistical analysis is the Mann-Whitney test, also known as the Wilcoxon rank-sum test; we applied the Mann-Whitney test on a non-parametric statistical hypothesis test for quantifying whether one of the two samples of independent observations tends to have higher values than the other [3]. The test works by calculating a test value U and comparing the calculation with the distribution which is known under the null hypothesis. This comparison results in a p-value. If the p-value was lower than the predefined threshold (we used the traditional threshold of 0.05) than we could reject the null hypothesis and accept the alternative hypothesis. If this threshold was larger than 0.05 then our null hypothesis was not rejected, and we instead rejected the alternative hypothesis.

$\mathbf{H_0}$ **(Null)**	$\mathbf{H_1}$	**RQ**
H_{10}: The participation of users who promote themselves on the website is same as the users who do not promote themselves on the website	H_{11}: The participation of users who promote themselves on the website is different then the users who do not promote themselves on the website	RQ_1
H_{20}: The participation of women who took the survey is the same as women's participation on SO website	H_{21}: Women's participation in the survey is different to the participation of women on SO website	RQ_1 and RQ_2
H_{30}: The reputation of users does not influence participation in the survey	H_{31}: The reputation of users does influence participation in the survey	RQ_2
H_{40}: The experience of the users does not influence participation in the survey	H_{41}: The experience of the users does influence participation in the survey	RQ_1
H_{50}: User's participation in the survey does not change with respect to the user's geographical location	H_{51}: User's participation in the survey does change with respect to the user's geographical location	RQ_2
H_{60}: The number of participants that participated in the survey is equal to the number of participants that received the Census badge for the same survey	H_{61}: The number of participants that participated in the survey is different to the number of participants that receive the Census badge for the same survey	RQ_3

Table 2.1: Null Hypothesis to be tested, and their relationship to the Research Questions

Chapter 3

Data Collection and Feature Creation

In this section, a comprehensive description is given of how the Stack Overflow website and its Developer Survey works along with the process of receiving the census badge. We provide further details on how we gathered data about the users and the participants of the survey and how we distinguished these two groups of users. We also provide a detailed outline of the features that we created, which more clearly demonstrates the basis of our imminent analyses. First, we represent the data collection process, after which we show our dataset and how it splits into small subsets addressing features such as gender, geography, and users with a self-promotion index.

3.1 StackOverflow Website and Developers Survey

3.1.1 Website Workflow

There are various activities that users who have an account on the SO website can perform to make full use of the platform, including posting a question and providing answers to already existing questions, commenting on the posts, or accepting a question or an answer. All these activities help users to increase their score on the website in the form of reputation. The more one engages in the platform, the more one can benefit from it. All these features are not available for users at the beginning, such as accepting a question or editing posts; instead, one has to gain specific reputation scores to have access to these features. For example, to accept a question, one must have a minimum reputation score of 15, and the scores are even higher for editing someone's posts.
The platform also rewards its users by assigning various achievement badges once users start utilizing the features provided by the website. The first badge received is the 'Informed' badge after reading the tour page or after reading the instructions on the workflow of the platform. If one asks a question and accepts an answer, one receives a 'Scholar' badge; it is also possible to get more badges depending on the activities one performs and how actively one engages.

3.1.2 Developers Survey

The platform continuously attempts to improve itself in order to achieve maximum engagement from its users. In order to know more about them and to make the platform more user-friendly, and to engage coders around the world, SO publishes a survey every year called the *Stack Overflow Developer's Survey.* The survey consists of various categories of questions. In the following points, we attempted to categorize the questions according to their types.

- Questions regarding demographics: this includes participant's age, gender, race, sexuality, ethnicity, education, and experience in coding.
- Questions regarding work environment: attitude towards fellow peers, working environment preferences, and preferences for a new job if seeking one.
- Questions regarding technology: technologies users like to work with, challenges encountered while adopting new technologies, questions about the willingness of their current company to adapt to new trends, how much influence on their organization they feel they have in purchasing new technology tools.
- Questions about stack overflow: how satisfied a user is from the website or whether they found an answer that solves their coding problem; if they have copied the coding example and pasted their solution, seen a job listing they were interested in, researched a potential employer by visiting its company page, searched for a job, asked a new question, written a new answer to someone else's question, or participated in a community discussion on meta or in chat.

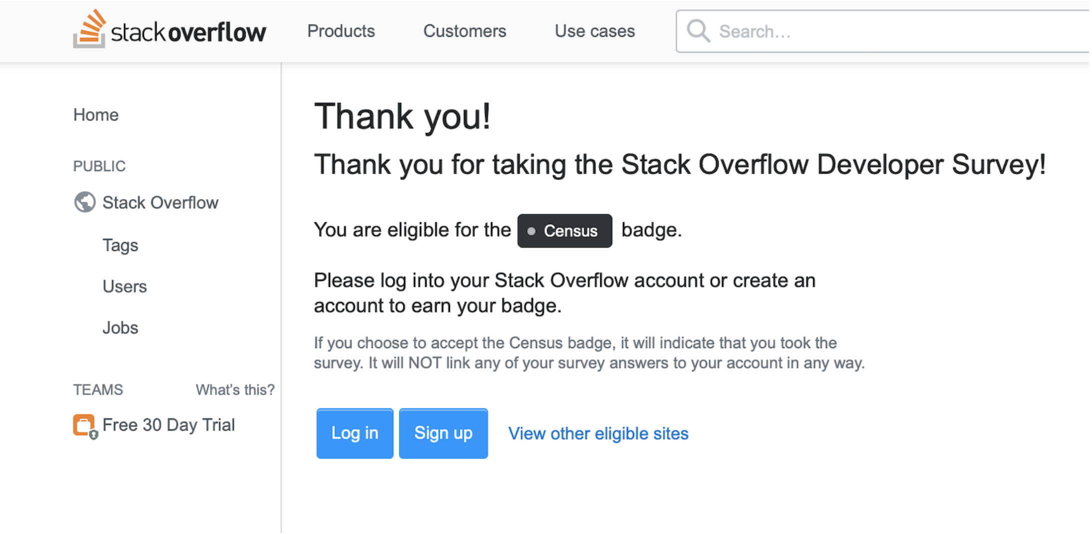
More information on the survey and its results can be found at *Stack Overflow Blog.*[1] The more participants a study has, the more beneficial it is to produce useful findings. In 2017, SO started rewarding its respondents by assigning them with a badge named 'Census'. This badge falls under the silver badge category identified as class 2. One can claim this badge after successfully completing the survey, as shown in **figure 5.3** . Non-registered users have to register themselves first on the website to get this reward, whereas registered respondents must log in to their account to reflect the badge on their profiles. This approach helps users to improve their SO reputation and profile. The aim is to attract more people to fill the survey; it is interesting for us to know if the platform is benefiting from this gamification approach.

3.1.3 Data Collection

We used the Stack Overflow data dump from SOTorrent Dataset Version 2019-12-02 [2] to collect the available information on users. Since we were comparing the survey respondents to non-survey respondents, we distinguished our two groups by separating the users who

[1] https://stackoverflow.blog

[2] https://empirical-software.engineering/projects/sotorrent/



Thank you!

Thank you for taking the Stack Overflow Developer Survey!

You are eligible for the ● Census badge.

Please log into your Stack Overflow account or create an account to earn your badge.

If you choose to accept the Census badge, it will indicate that you took the survey. It will NOT link any of your survey answers to your account in any way.

Figure 3.1: Claim the census-badge.

had received census badge from the ones who had not. We could not gather information about all the respondents of the survey as it is anonymous and open for everyone, even those who do not have an account on the website. Therefore, the best way to gather the information about the survey respondents was to quantify users signalling from Census Badge. In total, we collected data on 11,0376,305 users, out of which we identified 104,079 users who participated in the survey (as our survey respondents) signalled by their census badges.

3.2 Feature Creation

Most features used in our analysis were retrieved from user profiles. We started by analyzing these by looking into available fields including the name, creation date, location, ‘about me’, number of profile views and number of up-votes and down-votes cast by the user. The features that we created for the analyses are as follows:

3.2.1 Self-Promotion

From the ‘AboutMe’ field, we were able to learn more about the user and their other social networking accounts such as LinkedIn, Twitter, and Github. We utilized this information to create a self-promotion index for our model. This feature took a binary value of either 0 or 1, assigning 1 to the users who had promoted either of the three social networking sites mentioned above in their ‘AboutMe’ section of their SO profile and 0 to the rest. This helped us identify if any of our two sample groups were proactively promoting themselves on the website and if any of the two groups were more open and willing to provide links to their professional and personal lives. Accordingly, we could understand more about the behavioural differences among these groups on the platform. The following table shows the data shown after filtering keywords such as ‘LinkedIn’, ‘Twitter’ and ‘GitHub’:

Category	Self-promotion	No self-promotion	Total
Respondents	2,854	101,225	104,079
Non-Respondents	19,189	11,253,037	11,272,226

Table 3.1: Respondents and non-respondents with self-promotion index

As shown by the figures in **table 3.1**, we found that out of all respondents only 2.7% had promoted themselves on the website. Conversely, among the non-respondents group, only 0.17% had provided links to their other social accounts. We acknowledge that it could be possible that users might not have an account on one of the three social platforms; despite this, we would still argue that the data we have is the best possible data to address our hypothesis and perform the analysis. It also helps us to be more confident in the reliability of the data we have by limiting it to the users for whom we have a higher level of certainty. The statistical significance of these numbers was shown by applying the Mann-Whitney U test detailed in Chapter 5.

3.2.2 Experience

We were also interested in the age of the user's account, as it provides us with the number of years a user has spent on the website and how this influences their participation in the survey. To quantify a user's experience on the website, we created an experience index from the sign-up date (account's creation date), and we added the numbers in the feature named 'Experience'. Since the platform became public in 2008, the oldest user accounts date from the year 2008, and the newest account was created in the year 2019. We subtracted the account creation year from the year 2020 in order to calculate the years of experience. The resulting number is the tenure of a user on the website. The maximum account age was therefore 12 years, and the minimum was 1.

3.2.3 Reputation

We took this feature directly from user profiles as it provides us with the scores a user holds as reputation points on the website. We were interested in this feature as we wanted to find out if there was any correlation among users with a high reputation and those who took part in the survey.

3.2.4 Gender

Since there was no labelled field about the gender of the users in the SO dataset, we inferred the gender of individuals from their profile's name. Gender deduced from profile names is a complicated task because there are some regions where males and females can have the same names; for example, 'Mapreet' is both a male and female name in India. To ascertain the gender, we therefore used the same approach applied in another study (Gender differences in participation and reward on Stack Overflow, 2019)[2]. We used *GenderGusser* version 0.4.0 to identify the gender of users, which is a sophisticated approach, deducing gender according to first name and location of inputs.

The result of the tool corresponded to one of the following variables: 'unknown' (name not found), 'andy' (androgynous), 'male', 'female', 'mostly male', or 'mostly female'. The difference between andy and unknown is that the former is found to have the same probability of being male than female, while the latter means that the name was not found in the database [18]. The tool provided gender for 41,177 survey respondents out of 104,079. Selecting only the users identified as male or female, we ended up with a smaller but more accurate sample size of 37,486. Out of these respondents, we had 2,673 female and 34,813 male respondents. Meanwhile, for the non-respondents, we ascertained the gender of 3,921,283 users by further narrowing it down to male and female, resulting in a sample of 513,754 users identified as female and 2,946,631 as male, as shown in **table 3.2**.

Gender	**Users**	**Respondents**	**Non- Respondents**
Females	516,427	2,673	513,754
Males	2,981,444	34,813	2,946,631
Mostly-females	8,0126	582	79,544
Mostly-male	260,792	2,491	258,301
Andy	12,3671	618	123,053
Unknown	7,413,845	62,902	7,350,943

Table 3.2: Gender Inference

It is worth mentioning various constraints to our approach to inference. First, it was assumed that men and women were equally likely to have user names that would directly reflect their gender or real identity, and that this would not affect our hypothesis. Various studies have shown that anonymity is a key factor when it comes to surveys and online platforms [19]. It is possible that users have anonymous names to build an independent and separate identity; such behaviour from the individuals on the online platforms is well documented in the study 'Gender-swapping in the Internet' [20].

Despite these constraints, we believe that all the identified names gave us the best possible data to inform our analyses and run the model. It also helped us to be more certain about the data we had by limiting it to the users with a higher level of certainty. For the analysis, we were only interested in male and female genders, assigning a binary value 1 to male and 0 to female.

3.2.5 Geography

Stack Overflow users come from around the world, and the majority of contributors are from Europe and North America, with Asia third (mainly represented by India) and contributors from Oceania comparatively higher than South America and Africa together with Asia [21, 23]. The skewed results can be explained by various factors identified in previous research. One of the critical factors is the language of the site, which is predominantly English; SO also has a policy of question-answers only being in English [24], acting as a roadblock for contributors from non-English speaking parts of the world [25].
The platform is also available in a local version for some countries, such as a Russian language version in Russia and a Portuguese one in Portugal. Along with language barriers, there were many other factors that could influence the contribution, such as limited access to online resources. In some countries and regions, internet speed could act as a barrier limiting the involvement on the platform [22]. The unequal distribution of online infrastructure across different countries or within countries could also be a significant factor shaping the interaction of the users on the platform in comparison to peers [26]. All these factors are potential candidates for shaping an individual's experience on the platform, and the level of influence is also different for each individual with respect to these barriers when it comes to improving their profiles or keeping up with their peers.

In this study, we inferred the geographical location of the users from the free-text provided in the location field of our dataset. We used Python Library *Geotext*[3] to gather the information about the location of respondent and non-respondent groups. Geotech is a free software under MIT Licence; the tool reads a string and retrieves the values for cities and countries along with the count. In practice, if a string is something like 'London is a great city', applying a cities method produces 'London' as our output, whereas if we have a string 'New York, Texas, and also China' and we apply a countries method, it produces (u 'US', 2), (u 'CN', 1) as our output with a count [27]. This approach has few limitations as it is biased towards language-specific strings given as an input - for example, if the location name is in English, i.e. Italy instead of Italia. We were only able to infer the location of 823,271 users in total out of 3,497,871 after filtering for gender, resulting in only 23.53% of users. We grouped the geographies of users into seven categories according to the seven continents: Asia, Africa, Europe, Oceania, North America, South America, and Antarctica. For the sake of the analysis and to save the model computation time in selecting the matches for around 200 different geographical variables, we decided to get users' locations by continents in contrast to countries. Since we had no users from Antarctica, we reduced our analysis to the remaining six categories. The data we produced after filtering users was the most accurate data possible according to our applied approaches.

[3] https://geotext.readthedocs.io/en/latest/

Chapter 4

Methods

In this section, an introduction of various methods is provided that have been used to build the model for the analysis. We discuss the techniques and provides the reasoning behind the selection of model and matching algorithm that aligns best with the requirements to perform the analysis.

4.1 Propensity Score Matching

We first introduce the Propensity Score Method (PSM), also known as one of the remedies for reducing the selection bias in the surveys, but with a dependency on the selected variables. If the selected variables for the analysis are associated with the selection recorded, PSM is an approach to deal with the selection bias [28]. Figure **4.1** gives a clear and short understanding of which methods to consider in different scenarios to reduce selection bias. We discuss more about PSM and its use along with the regression model selection in the following section.

Propensity score methodology was introduced by Rosenbaum and Rubin in 1983. It is used to design observational studies with non-parametric data in an analogous way as randomized experiments are designed. Traditional models, e.g., least square regression and difference in difference are used to analyze observational data resulting in outcomes that should not be used as they are not designed for observational studies. The reasoning behind this claim is explained in *Rubins* words - “propensity score is a function only of covariates, not outcomes, repeated analyses attempting to balance covariate distributions across treatment groups do not bias estimates of the treatment effect on outcome variables” [29].

The propensity score is the probability of being treated (Wi = 1 vs Wi = 0), where i indexes the units in the study (i = 1, . . ., N) and Wi is the indicator of received treatment assigning 1 to the units being treated and 0 to the units which did not receive the treatment. These scores are used to reduce the selection bias by appending balance in the covariates (the characteristics of participants), helping in matching the individual’s

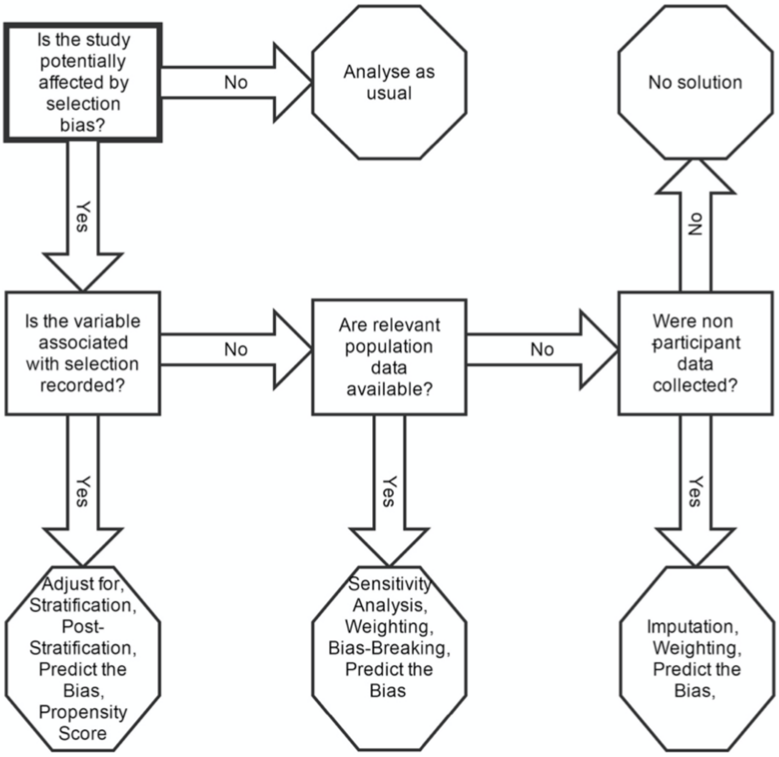

Is the study potentially affected by selection bias?
No
Analyse as usual
No solution
Yes
No
Is the variable associated with selection recorded?
No
Are relevant population data available?
No
Were non participant data collected?
Yes
Yes
Yes
Adjust for, Stratification, Post-Stratification, Predict the Bias, Propensity Score
Sensitivity Analysis, Weighting, Bias-Breaking, Predict the Bias
Imputation, Weighting, Predict the Bias,

Figure 4.1: A flowchart tool for researchers: Which methods are suitable to reduce selection bias? [28]

characteristics between the two groups [31].Propensity score matching forms group of participants for treatment and control units. From these groups, a matching set is created which consists of at least one individual from each group that have similar propensity scores. The aim is to remove the conflicts that comes with observational studies and their data analysis. These conflicts may arise due to the differences in the baseline characteristics of the treated or exposed subjects and control or unexposed subjects. According to a study, prerequisite to yield high-quality grounds to inform decision-making is the ability to minimize the effect of confounding [32]. The propensity score is the basis and more frequent approach to reduce bias while estimating casual treatment effects [33].

The design of the matching is either bipartite or non-bipartite. In bipartite matching, sampling of groups is done without replacement meaning that if one individual from a group is selected at first and matched with the individual from another group, then that particular individual cannot be matched again. Whereas, non-bipartite matching is simply the opposite of bipartite matching. It is done with replacement; individuals can be matched again even if they have found a match in the first place. In practice, bipartite matching is more common when, in fact, non-bipartite is only usable for the rarefied cases when there is no other option, and the same member must be reused. For example, if the same control is used as a match for two or more treatment group participants [31]. In the following section, we briefly review different algorithms to form pairs of treated and untreated subjects matched on the propensity score.

4.1.1 Matching Algorithms

The section describes seven different matching methods; exact matching, subclassification, optimal matching, greedy nearest neighbour matching, full matching, genetic matching, and coarsened exact matching [34].

Exact Matching

This method implements matching both with or without replacement. The methods which are based on matching without replacement are the ones that match each untreated unit to at most one treated unit. Once an untreated unit has been matched to a treated subject, that untreated unit is no longer eligible to consider as a match for other treated units. Exact matching is known as the simplest and most common method of matching. The technique matches each sample of treatment subject to all the possible control subjects with precisely the same values of all of its covariates, resulting into subclasses such that every treatment and control unit have the values of the same covariates within each subclass.

Subclassification

Often there are cases with many covariates with large values that finding exact matching could be impossible. Instead of exact values of covariates, subclassification aims to provide as similar matches as possible for treatment and control units forming different subclasses. There are several subclassification schemes, including the one based on a scalar distance measure, in this the propensity score is estimated using the distance option. The methods forms number of sub-classes based on the distance measure (the propensity score) using any of the suitable regression models. By default, each subclass has approximately the same number of treated units. The method may also be used in association with nearest-neighbour matching. Performing matching by forming an association of both the methods work by selecting matches using nearest neighbour matching and after finding the matches, the subdivision of matches takes place forming subclasses and adds a variable to the output object indicating subclass membership.

Nearest Neighbor Matching

Most of the tools like MatchIt uses “greedy” matching as a default nearest neighbour matching [34]. It selects the control matches that are best suited to each subject in the treatment group. Same as the subclassification method, the nearest neighbour method uses a distance measure (propensity score) established by the distance option (logit by default). Matches are selected from control units to each unit of the treatment group one by one; the matching order can be specified in either ascending or descending order. At each matching step, the control unit is selected that is not yet matched but is closest to the treated unit on the distance measure.

A modification of the method is greedy nearest neighbour matching in a range of specified calliper widths. In this modified method, matching treated and untreated units is conceivable if the absolute difference in their propensity scores is in the range of specified maximal distance (the calliper distance). For implementing matching with a calliper, the logit model is used over the propensity score using a calliper width defined as a proportion of the standard deviation of the logit and of the propensity scores [35, 36]. For instance, the calliper of value 0.5 matches the treatment units to the control subjects within the range of 0 - 0.5 with respect to their propensity scores. If more closely related matches are desired, one can flexibly decrease the width of the calliper.

Optimal Matching

In optimal matching, the matched pairs are formed in order to minimize the average within-pair difference in propensity scores. In contrast, the greedy nearest matching method selects the closest match in the control unit for each treatment unit (in case of several untreated units are equally close to the treated subject, one of these untreated units is selected randomly) [32]. On the other hand, the resulting matching sets of treatment and control units are generally the same for both the methods (nearest and optimal), but optimal matching performs more delicate work in minimizing the distance within each matched pair [37]. Optimal matching is also beneficial when the matches for the treatment units are scarce and inappropriate.

Full Matching

Full matching is the specific form of subclassification; the creation of subclasses takes place optimally. The matched sample in this matching consists of fully matched sets, these fully matched sets contain one treatment unit matched with one or more than one control unit or vice versa (several treatment units for one control unit) [38]. The units that are outside the range of standard support are discarded and have no place in the subclass, this approach is same as in the subclassification method.

Genetic Matching

The genetic matching method automates the process of finding a good matching solution [39]. The concept behind the genetic matching is to optimize the genetic search algorithm in order to assign a weight set for all the covariates producing optimally balanced form after completion of the algorithm. As of now, the matching is done with replacement, and balance among the units is obtained using paired t-test and Kolmogorov-Smirnov test for continuous variables with the flexibility of changing these options.

Coarsened Exact Matching

Coarsened Exact Matching (CEM) is a Monotonic Imbalance Bounding (MIB) matching method - a user selects the balance among the treatment and control units instead of identifying via onerous ways and continuous reestimation. This approach helps to accommodate the imbalance, which might occur in one variable while adjusting for the other. The method is tightly constrained over the user preference for the average treatment effect estimation error and degree of dependence of the model. In addition it eliminates the requirement for independent conduct to bound the data to standard empirical support. These methods are also potent to any type of measurement error, have extremely fast computation speed, fits congruence principle, performs very well with multiple imputation methods for missing data [34].

After carefully analyzing the above-mentioned matching algorithms, we work with nearest-neighbour matching and exact matching as these two matching algorithms suit best with our data and objective of this study. The matching results from the nearest neighbour match were comparatively loose then the exact match method, the implementation, and matching results from the methods are explained in more details in chapter 5.

4.2 Regression model

It's well known fact that simple regression analysis is not enough and could lead to misleading results when the co-variate distribution in the sample groups is not normally distributed and are very different from each other. The three conditions that must be fulfilled in order to get trustworthy results from the regression analysis are well explained in rubins work [29]. Whereas when the co-variate are approximately normally distributed, the difference in means of the propensity score in the two groups must be less than the half of standard deviation apart. The ratio among the treated and control group is close to 1, and the ratio of variance of the residuals among the co-variates after adjusting for the propensity score must be close to 1 [30].

We use the Logistic regression model for the analysis. In this section, we introduce the logistic regression model, reasoning for the model selection, its usage and requirements, functionalities along with conducting a simple logistic regression and interpreting the results.

4.2.1 Logistic Model

Logistic regression is a tool that could be used for building models with categorical data or a response variable with two levels. It is a class of generalized linear model (GLM). In order to deal with response variables where multiple linear regression models fails to get meaningful results, logistic regression is used. Notably, the response variables in such scenarios frequently take a form for which the distribution of residuals is not normally distributed. GLMs can be anticipated as a two-stage modelling approach. In the first stage, we model the response variable using a probability distribution which could be binomial or poison's distribution. In the second stage, all the parameters of the distribution are modelled via a collection of predictors and a particular form of multiple regression [41]. The fundamental mathematical approach that inhibits logistic regression is the logit model, also known as the natural logarithm of an odds ratio. In general, logistic regression is an excellent fit for interpreting the categorical data and addressing hypotheses about the relationships among the categorical variables outcome and for one or several continuous predictor variables [40].

Notion for Logistic model: The outcome variable for the GLM model is denoted by Yi, where the index i is used to represent observation i. In the application, Yi is used to represent whether a user is a survey respondent (Yi = 1) or not (Yi = 0). The name of our dependent variable is 'Participation' , if a user is a participant our outcome variable i.e participation is 1 (Yi = 1) otherwise it is 0 (Yi = 0). The predictor variables are represented as follows: x1: is the value of variable 1 for observation i, x2: is the value of variable 2 for observation i, and so on.
The outcome, Yi, takes the value 1 (in our application, this represents the participation of a user in the survey) with probability pi and the value 0 with probability 1 - pi. It is the probability pi that we model in relation to the predictor variables.

The logistic regression model relates the probability a user is a participant of the survey (pi) to the predictors x1, i, x2, i ..., xk, i through a framework equivalent to multiple regression:

$$transformation(p_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... \beta_k x_{k,i} \tag{4.1}$$

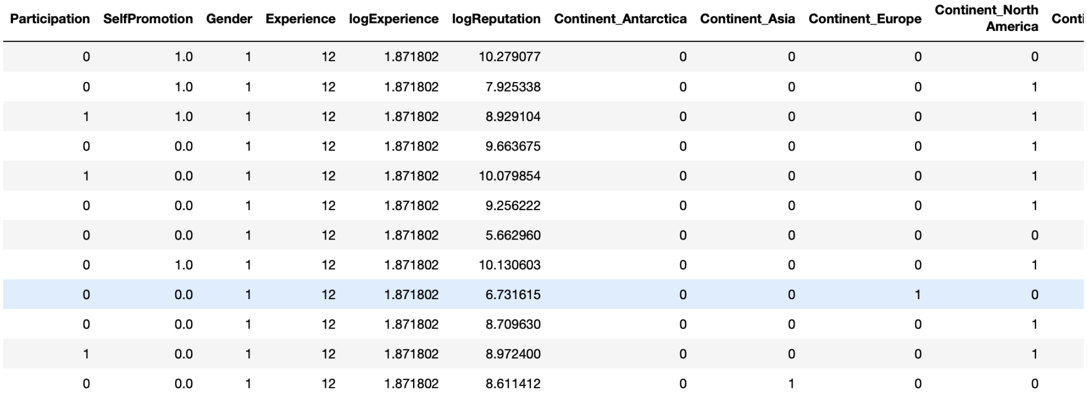
We want to transform the equation mentioned above such that it is mathematically acceptable. For example, we want to transform the range of possibilities in L.H.S of the equation to be greater than or equal to the range of possibilities on R.H.S of the equation, resulting into values either 0 or 1 for the left side of the equation. Whereas, the right side of the equation could take values in any range of numbers. The most common transformation for pi is the logit transformation, which may be written as :

$$logit(p_i) = log_e(\frac{p_i}{1 - p_i}) \tag{4.2}$$

Equation 4.2 can be rewritten as below using log transformation for p_i

$$log_e(\frac{p_i}{1 - p_i}) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... \beta_k x_{k,i} \tag{4.3}$$

In figure **4.2**, we give a glimpse of our data and the variables we selected for the analysis. For the continuous variables such as experience and reputation, we take the log transformation of the data. The reason for log transforming the data is to deal with skewness or to get closer to a normal distribution whereas it also helps to verify the data against validity, additivity, and linearity which are typically much more critical. The log transformation is particularly relevant when the data vary a lot on the relative scale as can be seen in figure **4.3 and 4.4**, the distribution of data before and after log transformation.



Participation	SelfPromotion	Gender	Experience	logExperience	logReputation	Continent_Antarctica	Continent_Asia	Continent_Europe	Continent_North America	Conti
0	1.0	1	12	1.871802	10.279077	0	0	0	0	
0	1.0	1	12	1.871802	7.925338	0	0	0	1	
1	1.0	1	12	1.871802	8.929104	0	0	0	1	
0	0.0	1	12	1.871802	9.663675	0	0	0	1	
1	0.0	1	12	1.871802	10.079854	0	0	0	1	
0	0.0	1	12	1.871802	9.256222	0	0	0	1	
0	0.0	1	12	1.871802	5.662960	0	0	0	0	
0	1.0	1	12	1.871802	10.130603	0	0	0	1	
0	0.0	1	12	1.871802	6.731615	0	0	1	0	
0	0.0	1	12	1.871802	8.709630	0	0	0	1	
1	0.0	1	12	1.871802	8.972400	0	0	0	1	
0	0.0	1	12	1.871802	8.611412	0	1	0	0	

Figure 4.2: Data-set glimpse and selected features for the analysis.

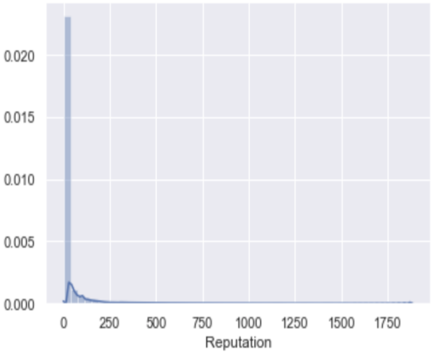


Figure 4.3: Distribution before log transformation

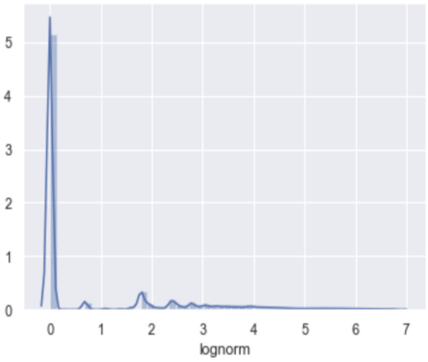


Figure 4.4: Distribution after log transformation

After scaling the data and getting rid of missing value, we run the simple logit regression on the variable reputation to check its effects on participation in the survey. The regression model looks like in equation 4.4.

$$y(participation) = \beta_0 + \beta_1 x(reputation) \tag{4.4}$$

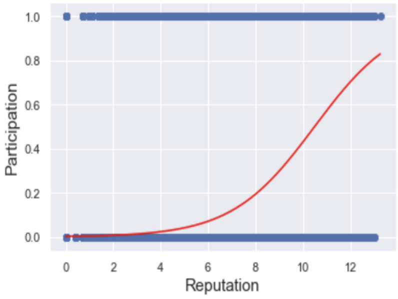
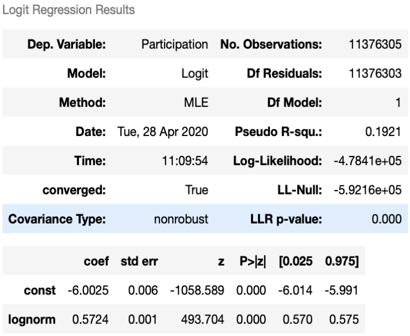


Figure 4.5: Logistic regression curve



Logit Regression Results

Dep. Variable:	Participation	**No. Observations:**	11376305
Model:	Logit	**Df Residuals:**	11376303
Method:	MLE	**Df Model:**	1
Date:	Tue, 28 Apr 2020	**Pseudo R-squ.:**	0.1921
Time:	11:09:54	**Log-Likelihood:**	-4.7841e+05
converged:	True	**LL-Null:**	-5.9216e+05
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-6.0025	0.006	-1058.589	0.000	-6.014	-5.991
lognorm	0.5724	0.001	493.704	0.000	0.570	0.575

Figure 4.6: Logistic regression summary table

The value of our $\beta_0 = -6.0025$ and $\beta_1 = 0.5724$, the prediction for the model is such that one unit increase in the log reputation, increases the likelihood of a user participating in the survey by 0.5724 units. We used the MLE method known as maximum likelihood estimation; the likelihood function estimates how likely it is that the model at hand describes the real underlying relationship of the variables. The bigger the likelihood, the higher the probability that our model is correct [42, 43], MLE tries to maximize the log-likelihood function. Log-likelihood is always negative, and the regression always gets the most optimized value for the model by going through different until it gets the value for which the log-likelihood is the maximum, and then it stops optimizing. Next, we have McFadden's r-square, it's value is 0.2 (after rounding off) which is in the range of 0.2 - 0.4 identified as good value for any logistic regression [43], the p-value is also below 0.05 which makes the variable statistically significant for the model. The results of the regression are described in the summary table in figure **4.6** and logistic regression curve is described in figure **4.5**.

In this section, we ran the simple regression to get familiar with the method, it's functioning, and interpretation. In **Chapter 5**, we perform the analysis with all the relevant variables for the analysis by first calculating the propensity score, then matching the individuals in the two groups (treatment and control), and rerunning the logistic model to interpret the final results.

Chapter 5

Analysis and Results

In this section, implemention of the above-mentioned methods is formulated. First the describtion of the variables is given and then estimation of the propensity scores using GLM is provided. On these propensity scores and variables at hand, implemention of the nearest neighbour matching is done to form the matched pairs within the two groups and later exact matching has been performed to get more optimal matches for our treatment and control groups.

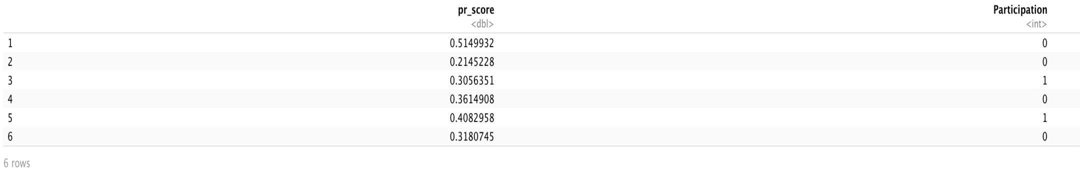
5.1 Analysis

The data collected from the user profile were analyzed using participation as a dependent variable and gender, reputation, experience, self-promotion, and continent (geography) as independent variables, mentioned in section 3.2. To examine the effect of variables on Participation = 1 (Treated) and Participation = 0 (Control) we go through the following steps:

1. Estimate the propensity score (the probability of being treated given a set of pre-treatment covariates).
2. Examine the region of common support.
3. Choose and execute a matching algorithm : nearest neighbour and exact matching
4. Examine covariate balance after matching.
5. Estimate treatment effects.

5.1.1 Propensity Score Estimation

We use the logit model to estimate the propensity score with a binary outcome variable reflecting the treatment status. The model looks as follows :



	pr_score	Participation
	<dbl>	<int>
1	0.5149932	0
2	0.2145228	0
3	0.3056351	1
4	0.3614908	0
5	0.4082958	1
6	0.3180745	0

6 rows

Figure 5.1: Propensity score dataframe

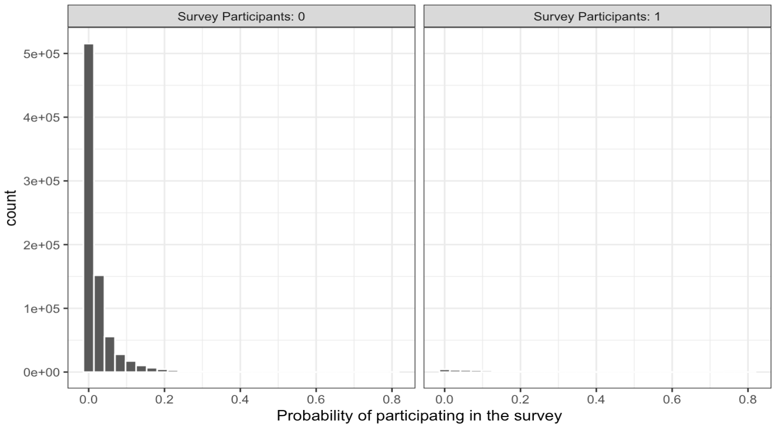


Figure 5.2: Probability of participating in the survey

$$y_{participation} = \beta_0 + \beta_1 x_{experience} + \beta_2 x_{reputation} + \beta_3 x_{selfpromotion} + \beta_4 x_{gender} + \beta_5 x_{continent}$$

Using the above model, we can now calculate the propensity score for each user. It is simply the user's predicted probability of being treated, given the estimates from the logit model. We calculate the propensity score using $predict()$ and create a data frame describe in figure **5.1**, it contains the values for propensity scores and user's actual treatment status. After estimating the propensity score, we plot histogram of the estimated propensity scores by treatment status via examining the region of common support, as shown in figure **5.2**.

5.1.2 Executing Matching Algorithm

A simple method for estimating the treatment effect is to restrict the sample observations within the region of standard support, and then divide the sample under this region of standard support into five quintiles based on the estimated propensity score. Within each of these five quintiles, we can then estimate the mean difference by treatment status. However, most matching algorithms adopt more sophisticated methods. The method we use is to find pairs of observations that have very similar propensity scores, but that differ in their treatment status.

We use the package MatchIt in R for performing the matching algorithm. The package estimates the propensity score in the background and then matches observations based on the method of choice ("nearest" in this case). After running the nearest-neighbour matching on our 24,013 treatment observations, we found 24,013 matches in the control group, matching each observation in the treatment group with the control group and ultimately resulting in 48,026 total number of observations for the model. The balance between the covariates in the matched samples is visualised in figures **5.3, 5.4, 5.5, 5.7,**

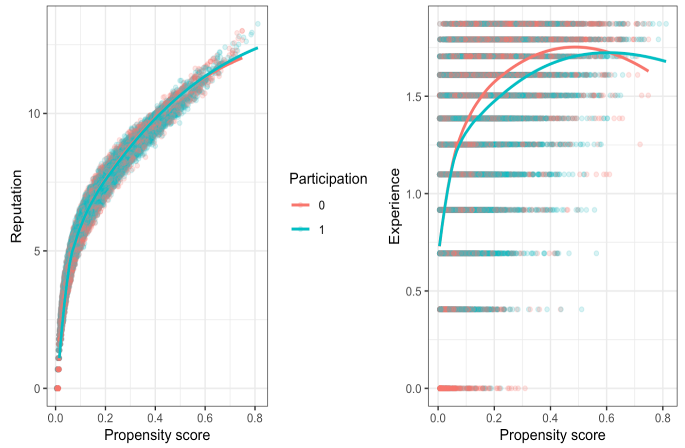


Figure 5.3: Treatment and control matching for reputation and experience

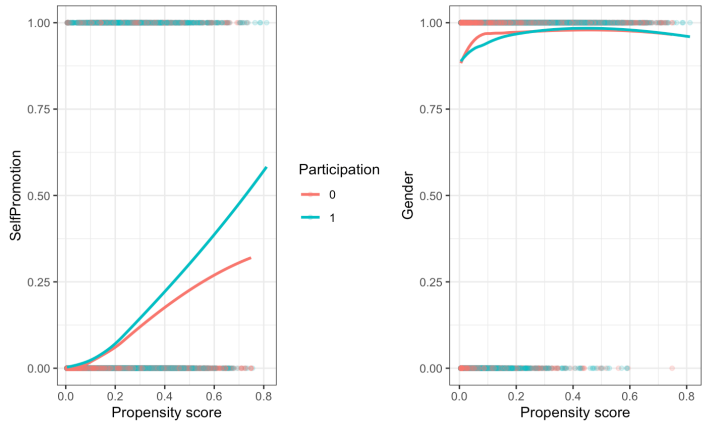


Figure 5.4: Treatment and control matching for gender and selfpromotion

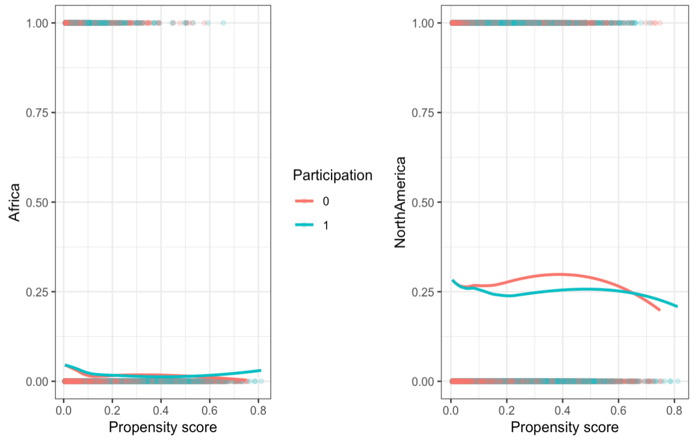
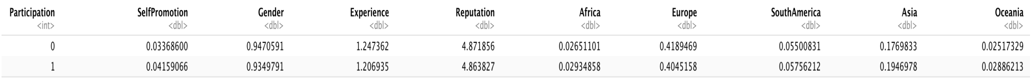


Figure 5.5: Treatment and control matching for continent africa and north-america



Participation <int>	SelfPromotion <dbl>	Gender <dbl>	Experience <dbl>	Reputation <dbl>	Africa <dbl>	Europe <dbl>	SouthAmerica <dbl>	Asia <dbl>	Oceania <dbl>
0	0.03368600	0.9470591	1.247362	4.871856	0.02651101	0.4189469	0.05500831	0.1769833	0.02517329
1	0.04159066	0.9349791	1.206935	4.863827	0.02934858	0.4045158	0.05756212	0.1946978	0.02886213

Figure 5.6: Difference in mean for treatment and control groups.

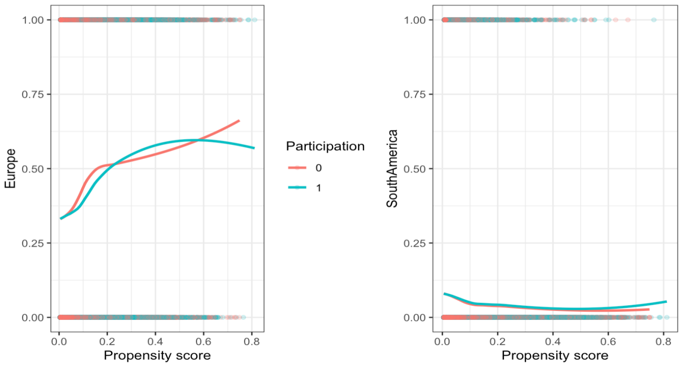


Figure 5.7: Treatment and control matching for continent europe and south-america

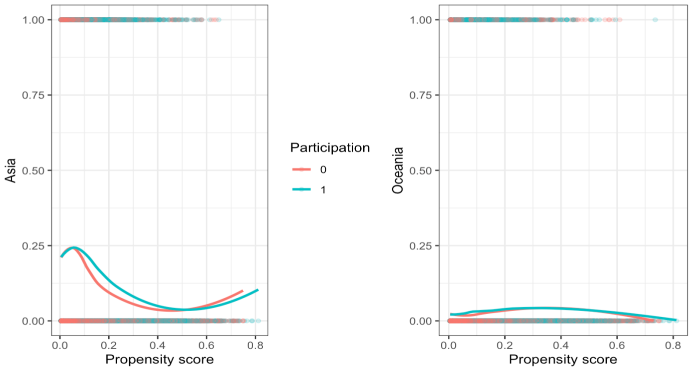


Figure 5.8: Treatment and control matching for continent asia and ocenia

5.8.

It is useful to plot the mean of each covariate against the estimated propensity score, separately by treatment status. If matching is done well, the treatment and control groups have (near) identical means of each covariate at each value of the propensity score. Figure **5.6** shows the results in the mean difference between our treatment and control groups after performing the nearest neighbour matching. We found almost identical mean values for all the covariates with some difference in covariates self-promotion, continent-Asia, Africa, and North America.

As nearest-neighbour matching successfully found a match for each observation in the treatment group but leaving us with only 48,026 total number of observations for the model. Therefore, we execute "exact" matching to find if the exact matching algorithm improves the total number of observations for the model. After implementing the exact matching, we found matches for 14,686 observations out of 24,013 in the treatment group, resulting in finding 646,365 exactly matched observations in the control group leaving us with 661,051 total number of observations. We also found better (nearly) identical mean values for all the covariates in both the treatment and control group using exact matching as compared to nearest neighbour matching.

5.1.3 Estimating Average Treatment Effect

We fit the resulting data from each algorithm concerning its corresponding weights and subclass into a new data-set to estimate the treatment effect in the two models. The summary table **5.1** shows the regression results. Model 1 corresponds to a simple regression model on the data before estimating the propensity scores and executing matching algorithms, whereas Model 2 corresponds to the regression results after fitting the data for nearest-neighbour matching and Model 3 for exact matching. All the three models are executed using generalized-least-square or logit model.

Table 5.1

	Before Matching	After Matching	
		logit model	
		(nearest-neighbour)	**(exact)**
	Model 1	Model 2	Model 3
$Intercept$	-4.744^{***}	0.425^{***}	-5.054^{***}
	(0.031)	(0.046)	(0.041)
$SelfPromotion$	0.098^{**}	0.187^{***}	1.864^{***}
	(0.038)	(0.049)	(0.216)
$Gender$	0.144^{***}	-0.237^{***}	-0.170^{***}
	(0.027)	(0.040)	(0.038)
$Experience$	-0.185^{***}	-0.331^{***}	0.265^{***}
	(0.019)	(0.028)	(0.024)
$Reputation$	0.475^{***}	0.035^{***}	0.671^{***}
	(0.003)	(0.004)	(0.005)
$Continent : Africa$	-0.047	0.165^{***}	0.417^{***}
	(0.040)	(0.057)	(0.054)
$Continent : Asia$	-0.284^{***}	0.082^{***}	-0.157^{***}
	(0.019)	(0.026)	(0.024)
$Continent : NorthAmerica$	-0.217^{***}	-0.005	0.016
	(0.017)	(0.023)	(0.022)
$Continent : SouthAmerica$	0.022	0.047	0.411^{***}
	(0.030)	(0.041)	(0.038)
$Continent : Oceania$	-0.219^{***}	0.174^{***}	0.575^{***}
	(0.042)	(0.057)	(0.076)
Observations	823,226	48,026	661,051
Akaike Inf. Crit. (AIC)	177,956.100	66,349.270	111,118.000
McFadden Pseudo R^2	0.180	0.003	0.211

Significance thresholds:$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Selection of final model for the analysis

Model 1 is a simple logistic regression model without matching. Since for this study we want to compare the matched observations of the treatment and control groups; hence we focus on the results from the remaining two models. In Model 2, after performing the logistic regression, we get significant results for almost all the covariates, but due to very few observations in the model, the significance of the model itself is not a good fit for predicting results as indicated by the very low value of MacFadden r-square 0.003. Hence, Model 3 with exact matching is the best fit for the prediction of results with MacFadden r-square 0.211.

5.1.4 Interpretation of Results

Reading the results from table 5.1 for variable **gender**, we first took the exponential of the units to calculate its effect i.e. **exp**(-0.170) = 1.185. This value indicates that given all the other variables are identical, a user who is identified as male is 1.2 times (or 120%) less likely to participate in the survey as compared to females.

Addressing Hypothesis $\mathbf{H}_1$

The variable **self-promotion** helps to address our hypothesis H_1. We calculate its effect same as for gender by taking exp (1.864) = 6.44, we found that given all the variables are identical for a user, if the users promote themselves on the website, they are 6.5 times (or 650%) more likely to participate in the survey as compared to the users who do not promote themselves on the website. Hence, we reject our hypothesis H_{10} and accept H_{11}.

Addressing Hypothesis $\mathbf{H}_2$

We are interested in comparing the participation ratio of women in the survey with the ratio of women who were active users on the website. We filtered these groups within the range of 1 year time period; we gathered data from 2019 Stack-Overflow Developers Survey to compare the ratio of women who participated in the 2019 survey to the women who were active on the website during the year 2018 (as the survey takes place at the start of the year annually, we compared the ratio with active users in 2018 because the website advertises about its next upcoming survey to the website users). To filter the active females on the website w.r.t year 2019, we selected all the females with ‘Last Access Date’ in the year 2019. The feature ‘Last Access Date’ gives the most recent time-stamp of a user when they were online on the website. After filtering for both the groups, we found that only 1.4% of registered females participated in the 2019 survey as compared to the ones who were active on the site as shown in table **5.2**.

Category	Females
Survey Respondents (2019)	1,478
Active on Website (2019)	103,920

Table 5.2: Female Participation

To test the hypothesis H_2, we performed the Mann-Whitney Wilcoxon Test on the filtered dataset. The mean and standard deviation for survey respondents is 0.013 and 0.112, respectively, whereas the mean and standard deviation for active females is 0.987 and 0.112, respectively. The p-value is 0.0001 with statistics = 70684776.000 resulting in deferential distribution among the two groups. Hence, we reject our hypothesis H_{20} and accept H_{21}. It is very interesting to know that out of 7.23% of total female respondents of survey 2019, only 1.4% of them have an account on the website. We can further say that the survey suffers from sampling bias as only 1.4 % of the female participants are general representatives of the whole female community.

Addressing Hypothesis H_3

The variable **reputation** helps us to address our hypothesis H_3. We calculate its causal effect by taking exp (0.671) = 1.95, we found that given all the variables are identical for a user, one unit increase in the reputation increases the chances of participating in the survey by 1.95 times or if the odds of reputation increases by 1 unit, the odds of participation in the survey increases by 195%. Hence, we reject our hypothesis H_{30} and accept H_{31}.

Addressing Hypothesis H_4

The variable **experience** helps us to address our hypothesis H_4. We calculate its causal effect by taking exp (0.265) = 1.30, we found that given all the variables are identical for a user, if the odds of experience increases by 1 unit, the odds of participating in the survey increases by 1.3 times or we can say that one unit increase in experience increases the probability of participating in the survey by 130%. Hence, we reject our hypothesis H_{40} and accept H_{41}.

Addressing Hypothesis H_5

To address our hypothesis H_5, we divide users' location into six categories (continents) as mentioned in section **3.2.5**. We choose continent Europe as our reference category for the comparison with other categories. Users from continent Africa are 1.5 times or 150% (exp(0.417) = 1.51) more likely to participate in the survey as compared to users from Europe. In comparison, users from Asia are 15% (exp(-0.157) = 0.85) less likely to participate in comparison with users from Europe. We get statistically insignificant results for continent North-America, this variable does not affect the results in the analysis, on

the other hand, users from continent South-America are 1.5 times or 150% (exp(0.411) = 1.50) more likely to participate in the survey as compared to users from Europe. Users from continent Oceania are also 1.7 times or 177% (exp(0.575) = 1.77)) more likely to participate in the survey as compared to users from Europe. The proportion of participation is very different concerning users geographic location.
We found that geographically, the user's participation ratio from continent Oceania is higher, followed by Africa and South-America with nearly the same results. Users participants in the survey from Europe are 15% higher in comparison to Asia, keeping Europe in the fourth position, whereas participants from Asia are least likely to participate. Hence, we reject our hypothesis H_{50} and accept H_{51}.

Addressing Hypothesis H_6

We are interested in finding out the ratio among the participants and 'census' badge receivers.

We are interested in identifying the difference between the number of participants who took the survey and the number of 'census' badge issued for that specific survey. As one can claim the census badge after participating in the survey, this helps us to measure the gamification effects on participation in the survey. It is of our concern to know if it leads to more participation in the survey. In order to compare the ratio, first, we count the total number of participants who participated in the Stack-Overflow Developers Survey, and then we count the number of census badges issued for that specific year. We compared the results, as shown in table **5.3**.

Year	Respondents	Census Issued	Ratio
2019	88,883	56,123	63.14%
2018	98,855	50,116	50.69%
2017	51,392	20,775	40.42%

Table 5.3: Participation and Census Badge Ratio

Out of all respondents in 2019, only 63.14% claimed the census badge, nearly 37% of the participants did not take the badge. We see a 10% decrease in the respondents of 2019 as compared to 2018. In contrast, the number of census badges receivers increases by nearly 10% every year. We again performed Mann-Whitney Wilcoxon Test to test our hypothesis. We get the p-value is 0.0019 with statistics = 68074776.000 resulting in differential distribution among the two groups. Hence, we reject our hypothesis H_{20} and accept H_{21}.

5.1.5 Analyzing Patterns in Activities of Users

We are further interested in analysing the difference in activities among the survey respondents and non-respondents. There are various kinds of activities a user can perform,

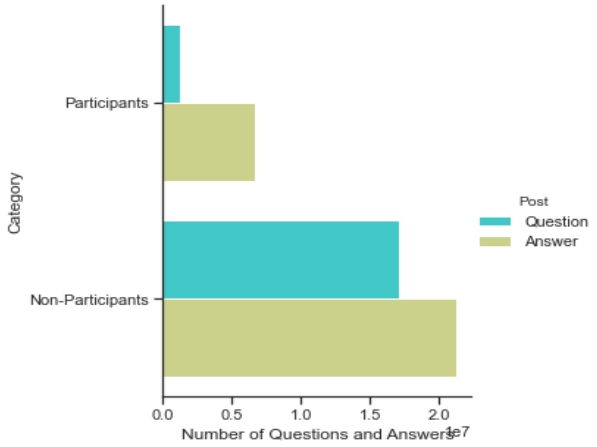


Figure 5.9: Difference in Post Types

starting from asking questions, providing answers, accepting questions, accepting answers, editing a question, and commenting on the posts. First, we wanted to find the number of users in our two groups who are post owners, and then we wanted to compare the type of posts in these two groups. There are eight different types of posts in the data-set: Question, Answer, Orphaned tag wiki, Tag wiki excerpt, Tag wiki, Moderator nomination, Wiki placeholder (seems to only be the election description) and Privilege wiki. Out of all these types, we are only interested in the question, and answer posts as they are the most dominating post types, around 90% posts are either questions/answers posts.

- Out of 104,079 the survey-respondent group, 84,732 users are posts owner.

  81.41 % of the Respondents have Posts i.e. they have either asked a question or have provided an answer.

  84.29 % of the Respondents have answer post

  15.26% of the Respondents have question post

- Out of 11,272,226 the non-Respondent group, 4,448,869 users are post owner.

  39.46% of the Non-Respondents have Posts i.e. they have either asked a question or have provided an answer.

  55.41 % of Non-Respondents have answer post

  44.40 % of non- Respondents have question post

In figure **5.9** we see the difference in post type among our participants(i.e respondents) and non-participants (non-respondents). Therefore, we can say that most of the survey respondents belong to the group of people who actively provides answers to the questions on the website. As the difference in the ratios is very high, 85% of the respondents are answer providers on the website, whereas the difference in the non-respondent group is 50% - 40% (answer-question respectively).

Chapter 6

Conclusion

This study quantitatively investigated the biases in the Stack Overflow Developer Survey. In the regression model, we found evidence that the survey is biased towards the groups of users who promote themselves on the SO website by providing links to other social networking sites like LinkedIn, Twitter, and Github. The account age and reputation of the users on the SO platform also had a significant effect on participation in the survey. The older accounts and higher-reputed users were more likely to participate in the survey, resulting in under-representation of the users who may have low reputation scores and are new users of the platform. It was also interesting for us to discover that male participants were less likely to complete the survey compared to female participants regardless of other factors from their profiles and activities on the site.
The survey respondents were from all over the world; we found that the users from Oceania had a higher rate of participation than users from other continents. The participation rate of users from Africa and South-America was nearly the same and included the second-most contributors after Oceania. Participants from Europe were the second-to-last, whereas participation from Asia was least in comparison to participants from other regions.

This contradicts the study of Dennis Schenk and Mircea Lungu [21]. The evidence there showed that Europe and North America were the principal and almost equal contributors; Asia came a distant third, mainly represented by India; and Oceania contributed less than Asia but more than South America and Africa together. We believe that the differences in these results could be related to the privacy of users and their willingness to provide data about themselves concerning cultural or regional differences. Apart from these differences, the participation in the survey could also be affected by its reach and level of advertisement in a specific geographic region; this may have differed according to geography.
We found that only 1.4% of the participants were registered as women, evidencing their minority status and leading to sampling biases in the survey. Ninety percent of the survey respondents were men, biasing the results towards the male community members. The majority of survey respondents were those who provided answers to the questions on the website, leading to an under-represention of users who asked questions. Hence, we can say that the survey is biased towards the group of users who had more experience on the website, held high reputation scores, and were answer providers. Moreover, we found that participation in the survey from the registered users increased from 2017 to 2018 by 51% and decreased by 11% from 2018 to 2019 with a constant 10% increase in census badge

receivers every year from 2017 to 2019.
From the results it is worth acknowledging that the ratio of users' participation to badge claim is increasing every year. We conclude that the population of the survey differs significantly from the overall population on the website; hence, specific communities of users are more likely to participate then the rest of the population. This results in under-representation of the users who have low reputation points, have recently became a part of the SO community, have more question posts, and belong to a specific geographical region. To improve the results and accuracy of findings, an interesting possibility would be to conduct a small reference survey to collect data from these under-represented communities. Publishing a reference survey specifically for users who are less likely to participate would help to validate the results from the Developer Survey. It is hard to make sure that users from the targeted group will respond to such a survey. Hence, to encourage the participation, it could be best to send out the reference survey via emails to the target group, and rewarding participants with greater rewards than the census badge on the website. It would also be interesting to investigate the reasons why they did not participate in the survey in the first place, with questions related to what holds them back from participating, their awareness towards the annual survey, and their overall experience on the website. Another worthwhile finding would be the number of users who would care about responding to this reference survey.

Chapter 7

Future Considerations

Future work should expand on the current notion of research to find out the reasons for differences in contribution to the website and participation in the survey concerning the geography of users. It would also be interesting to know whether there were any other types of biases in the survey - for example, if certain communities like the ones related to coding languages (python, java, R, etc.) have an impact on participation in the survey. More community-related research could be done by distinguishing communities by their tags (hashtags). It would be intriguing to learn if there are specific batch holders belonging to a specific group of people and to compare their contribution to the survey - for example, categorizing people concerning batch categories gold, silver and brown and comparing their participation in the survey and on the website. Therefore, as a possible direction for future work, we consider going beyond community and geographic differences and learning the reasons behind these differences.

Bibliography

[1] Denae Ford, Justin Smith, Philip J. Guo, Chris Parnin. Paradise Unplugged: Identifying Barriers for Female Participation on Stack Overflow. FSE 2016. Pages 846 - 857. Avialable at: https://dl.acm.org/doi/abs/10.1145/2950290.2950331. Last visited : 17 June 2020.

[2] Anna May, Johannes Wachs, Aniko Hanna. Gender differences in participation and reward on Stack Overflow. Empirical Software Engineering. Available at: https://doi.org/10.1007/s10664-019-09685-x. Last visited on: 13.06.2020

[3] H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. The Annals of Mathematical Statistics, 18(1):50-60, 1947.

[4] A. Bosu, C. S. Corley, D. Heaton, D. Chatterji, J. C. Carver and N. A. Kraft, ”Building reputation in StackOverflow: An empirical investigation,” 2013 10th Working Conference on Mining Software Repositories (MSR), San Francisco, CA, 2013, pp. 89-92, doi: 10.1109/MSR.2013.6624013.

[5] Kristen M. Altenburger, Rajlakshmi De, Kaylyn Frazier, Nikolai Avteniev, Jim Hamilton.Are There Gender Differences in Professional Self-Promotion? An Empirical Case Study of LinkedIn Profiles Among Recent MBA Graduates. AAAI . Available at https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/viewPaper/15615. Last Visited: 17 June 2020.

[6] Collier, Benjamin & Bear, Julia. (2012). Conflict, confidence, or criticism: An empirical examination of the gender gap in wikipedia. Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW. 383-392. 10.1145/2145204.2145265.

[7] D. Ford, A. Harkins and C. Parnin, “Someone like me: How does peer parity influence participation of women on stack overflow?,” 2017 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), Raleigh, NC, 2017, pp. 239-243, doi: 10.1109/VLHCC.2017.8103473.

[8] Silveira, Karina & Musse, Soraia & Manssour, Isabel & Vieira, Renata & Prikladnicki, Rafael. (2019). Reinforcing Diversity Company Policies: Insights from StackOverflow Developers Survey. 119-129. 10.5220/0007707901190129.

[9] Yang, Heng-Li & Lai, Cheng-Yu. (2010). Motivations of Wikipedia content contributors. Computers in Human Behavior. 26. 1377-1383. 10.1016/j.chb.2010.04.011.

[10] Adaji I., Vassileva J. (2016) Towards Understanding User Participation in Stack Overflow Using Profile Data. In: Spiro E., Ahn YY. (eds) Social Informatics. SocInfo 2016. Lecture Notes in Computer Science, vol 10047. Springer, Cham

[11] B. Bazelli, A. Hindle and E. Stroulia, "On the Personality Traits of StackOverflow Users," 2013 IEEE International Conference on Software Maintenance, Eindhoven, 2013, pp. 460-463, doi: 10.1109/ICSM.2013.72.

[12] B. Vasilescu, V. Filkov and A. Serebrenik, "StackOverflow and GitHub: Associations between Software Development and Crowdsourced Knowledge," 2013 International Conference on Social Computing, Alexandria, VA, 2013, pp. 188-195, doi: 10.1109/SocialCom.2013.35.

[13] Zagalsky, A., German, D.M., Storey, M. et al. How the R community creates and curates knowledge: an extended study of stack overflow and mailing lists. Empir Software Eng 23, 953-986 (2018). https://doi.org/10.1007/s10664-017-9536-y

[14] Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing, February 2015, Pages 171-174 https://doi.org/10.1145/2685553.2698999

[15] Vctor R. Basili and David M. Weiss. A Methodology for Collecting Valid Software Engineering Data. IEEE Transactions on Software Engineering, VOL. SE-10,NO. 6, November 1984

[16] Bogdan Vasilescu, Andrea Capiluppi and Alexander Serebrenik. Gender, Representation and Online Participation: A Quantitative Study of StackOverflow.

[17] Jelke Bethlehem. Selection Bias in Web Surveys. International Statistical Review (2010), 78, 2, 161-188

[18] gender-guesser 0.4.0. Available at: https://pypi.org/project/gender-guesser/0.4.0/. Last visited on: 13.06.2020

[19] Ronald E. Robertson1 Felix W. Tran1 Lauren N. Lewark1 Robert Epstein1. Estimates of Non-Heterosexual Prevalence: The Roles of Anonymity and Privacy in Survey Methodology.Arch Sex Behav 47:1-16. https://doi.org/10.1007/s10508-017-1044-z

[20] Bruckman A (1996) Gender swapping on the internet. High noon on the electronic frontier: conceptual issues in cyberspace pp 317-326

[21] Dennis Schenk, Mircea Lungu. Geo-Locating the Knowledge Transfer in Stack Overflow. 8 (2013), 21-24 https://doi.org/10.1145/2501535.2501540

[22] Janes R, Arroll B, Buetow S, Coster G, McCormick R, Hague I. Rural New Zealand health professionals' perceived barriers to greater use of the internet for learning. Rural and Remote Health 2005; 5: 436. Available: www.rrh.org.au/journal/article/436 Last visited on: 13.07.2020

[23] Nigini Oliveira, Nazareno Andrade, and Katharina Reinecke. 2016. Participation differences in Q&A sites across countries: opportunities for cultural adaptation. In Proceedings of the 9th Nordic Conference on Human-Computer Interaction. 1-10.

[24] Stack Overflow. 2009. Non-English Question Policy. https://stackoverflow.blog/2009/07/23/non-english-question-policy/[Online; accessed 17-June-2020].

[25] Philip J Guo. 2018. Non-native english speakers learning computer programming: Barriers, desires, and design opportunities. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM, 396.

[26] Wenhong Chen and Barry Wellman. 2004. The global digital divide-within and between countries. It&Society 1, 7 (2004), 39-45

[27] Geotext. geotext documentation Available at https://geotext.readthedocs.io/en/latest/ Last visited: 13 June 2020.

[28] Keeble, C. , Law, G. , Barber, S. and Baxter, P. (2015) Choosing a Method to Reduce Selection Bias: A Tool for Researchers. Open Journal of Epidemiology, 5, 155-162. doi: 10.4236/ojepi.2015.53020.

[29] Rubin, D.B. Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation. Health Services & Outcomes Research Methodology 2, 169-188 (2001). https://doi.org/10.1023/A:1020363010465

[30] Jason W. Osborne. Best Practices in Quantitative Methods. SAGE. Available at: https://books.google.ch/books?id=M5_FCgCuwFgC&pg=PA176&dq=Matching+using+estimated+propensity+scores:+relating+theory+to+practice+by+Rubin+D.B&hl=en&sa=X&ved=0ahUKEwjrnKSyo6ToAhX9wcQBHUQyB2oQ6AEIKDAA#v=onepage&q=Matching%20using%20estimated%20propensity%20scores%3A%20relating%20theory%20to%20practice%20by%20Rubin%20D.B&f=false

[31] Statistics How To. Propensity Score Matching: Definition & Overview. https://www.statisticshowto.com/propensity-score-matching/ [Online; accessed 17-June-2020]

[32] Austin PC. A comparison of 12 algorithms for matching on the propensity score. Stat Med. 2014; 33(6):1057-1069. doi:10.1002/sim.6004

[33] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983; 70:41-55.

[34] Daniel E. Ho, Kosuke Imai, Gary King, Elizabeth A. Stuart. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. May 2011, Volume 42, Issue 8. http://www.jstatsoft.org/

[35] Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician 1985; 39:33-38.

[36] Cochran WG, Rubin DB. Controlling bias in observational studies: a review. Sankhya: The Indian Journal of Statistics 1973; 35:416-466.

[37] Gu, X. and Rosenbaum, P. R. (1993), “Comparison of multivariate matching methods: structures, distances, and algorithms,” Journal of Computational and Graphical Statistics, 2, 405-420

[38] Rosenbaum, P. R. (2002), Observational Studies, 2nd Edition, New York, NY:Springer Verlag.

[39] Diamond, A. and Sekhon, J. (2005), “Genetic Matching for Estimating Causal Effects: A New Method of Achieving Balance in Observational Studies,” `http://jsekhon.fas.harvard.edu/` Last visited: 17 June 2020.

[40] Peng, J., Lee, L., & Ingersoll, M. (n.d.). An Introduction to Logistic Regression Analysis and Reporting. The Journal of Educational Research, 96(1), 3-14. `https://doi.org/10.1080/00220670209598786`, Last visited: 18 June 2020.

[41] Lumin, Introduction to Statistic : “Introduction to logistic regression” . `https://courses.lumenlearning.com/introstats1/chapter/introduction-to-logistic-regression/`, Last Visited : 18 June 2020.

[42] Hosmer, D. & Lemeshow, S. (2000). Applied Logistic Regression (Second Edition). New York: John Wiley & Sons, Inc.

[43] Long, J. Scott (1997). Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA: Sage Publications.

List of Figures

List of Tables