Mining the Insights of Stack Overflow Developer Survey

Mohammad Imrul Jubair
University of Colordao Boulder
Boulder, USA
mohammad.jubair@colorado.edu

Mohsena Ashraf
University of Colordao Boulder
Boulder, USA
mohsena.ashraf@colorado.edu

Adejoro Onimisi Cornelius
University of Colordao Boulder
Boulder, USA
adejoro.cornelius@colorado.edu

Abstract

There has always been a high demand for software development and tech-related jobs in market and the programming community plays a vital role to comprehend this employment sector. While speaking of a tech community, Stack Overflow is the most popular one since programmers and developers from diverse categories participate here very frequently. Over the past few years, Stack Overflow has been conducting yearly surveys on its users to collect information about how they learn and upgrade themselves, which tools they are utilizing, and what they enjoy in development. In this project proposal, we plan to apply an exploratory data analysis on this survey-based dataset—called the *Stack Overflow Developer Survey (SODS)*—to investigate correlations between job market and the programming technology along with corresponding cultures. We aim to find answers to few relevant questions and to test a hypothesis that articulates gender diversity in tech organizations. We believe, our findings will help job seekers to understand contemporary situations in tech-based job field. In addition, employers can also exploit our outcomes to examine their hiring strategies.

CCS Concepts: • **Social and professional topics** → *Computing occupations.*

Keywords: Stack Overflow Developer Survey, Data Mining Techniques

1 Problem Domain and Motivation

From 2021 to 2031—according to the report from *U.S. Bureau of Labor Statistics* [10]—it is anticipated that overall employment in computer and information technology occupations would expand by 15%, which is much faster than the average for all occupations and will add roughly 682, 800 new jobs. From this information, we can realize the demand and rise of technology related job in the market. In this era of technology, the world is witnessing tremendous digitization trends that appear to be on steady increase following continuous innovations, and the emerging application domain of these technologies. However, it has often been said that available software developers are still grossly inadequate; as we can observe from *IDC Market Perspective* report that says—the global shortage of full-time developers will increase from 1.4M to 4.0M for the period of 2021 to 2025; hence the full-time developer labor force will perform 90.8% capacity in 2021 and 84.9% capacity in 2025 [4].

However, The job market is still extremely competitive and of interest to all prospective candidates. It is important to comprehend and analyze current employment condition, locally or globally, so that one can act accordingly. For instance, If John—a research-oriented job seeker—finds *Django*[1] framework as a very demanding tool in contemporary tech profession, he can plan to improve corresponding skills through any platform, i.e., *Coursera*, *edX*, etc. However, it is very difficult for John to extract these insights, because the market is very broad, and the corresponding population is a complex field to study. One practical way is to deal a smaller sample group or community within this population and consider that for further mining. Without a question, Stack Overflow is today's most well-known community and a potential field for market analysis. It is an autonomous online Question and Answer forum. Stack Overflow's success is largely due to the engaged and active user community that collaboratively manages the site [8]. For the past few years, Stack Overflow (SO) has been surveying its members—known as *Stack Overflow Developer Survey (SODS)*[2]—to discover more about their learning process and progress, what technologies they employ, and what fascinates them. In SODS 2022, more than 70, 000 SO users from different background and experiences participated and the forum has been sharing publicly these responses as datasets for the past few years. Some sample questions are listed below.

- `Which programming, scripting, and markup languages have you done extensive development work in over the past year, and which do you want to work in over the next year?`
- `What is the primary operating system in which you work?`

In our project, we consider this survey dataset as a representative sample of the enormous tech-related community and want to investigate for intriguing insights. We plan to apply exploratory data analysis EDA) [11] that will help future developers better understand the industry. We propose to answer numerous research questions because our goal is

[1] https://www.djangoproject.com/
[2] https://survey.stackoverflow.co/2022/

to provide an analytical picture of the tech-job market with the most important ones being mentioned below.

- **Trends related questions:** *(i)* Can we model the current popularity of different programming languages, OS, tools and frameworks over the past years? *(ii)* Can we observe the temporal behavior of salary in tech-industries for different countries? *(iii)* What is the current impact of women in the tech-culture?
- **Correlation related questions:** What is the correlation between: *(i)* programming language vs. employement status? *(ii)* coding experience vs operating system? *(iii)* online course platform vs. job status? *(iv)* gender/ ethnicity vs. organization size?
- **Frequent Pattern related questions:** Given a minimum support, can we find frequently preferred programming frameworks?
- **Clustering related questions:** Can we cluster the samples to groups based on: *(i)* educational level and salaries? *(i)* age groups and salaries?
- **Classification related questions:** Given the preferred programming language and education information, can we classify a user to estimate salary range?

In addition to the above questions, we also plan to test the following hypothesis through our exploratory data analysis on SODS.

> **Hypothesis:** *Having the same job position, educational background and technical skill, the female employers DO NOT obtain same remuneration as male employers.*

We believe, providing answers to the aforementioned questions and testing the hypothesis can reveal some fundamental insights of the current job market. The results of our experiment can be used both by companies and job applicants to change their policies and attitudes respectively.

Report Organization. We reviewed several literatures and presented their overviews in Section 2. We describe important details of SODS dataset in Section 3 followed by our proposed plan for this project in Section 4. Section 5 concludes the report with estimated timeline and milestones.

2 Related Work

This section contains earlier studies that examined Stack Overflow or similar platform related data. Numerous research articles examined user participation on the SO platform and drew conclusions from it. *Fu et al.* [6] discovered that the interest change rate of suers follows a power-law distribution, which is different from the research-based interest change. According to the author, this phenomenon indicates that the community is more inclined to exploration strategy. In the technical forum, the study of *Brooke* [2] reveals a dominant male influence. *Moutidis et al.* examines the migration of the users between communities [8]. Note that, the aforementioned studies considered their dataset by scraping Stack Exchange forum [3]. Moroever, Some publications analyze their own conducted surveys of SO users, such as [1, 12, 13]. *Georgiou et al.* [7] incorporated SODS (2020) and scraped data to find COVID-19 impact on the Stack Overflow. Similarly, *Ford et al.* [5] used SODS from the year of 2017 with scraped data and found that women were more likely to start interacting with one another in SO than those who did not when they came across another woman. In 2020, *Nivala et al.* [9] analyzed SODS (2011–2018) and reported their findings. They concluded that, the novices are becoming more involved in the community, while number of expert users who provide good answers has decreased. According to them, developing countries are becoming more and more visible in SO. However, the apparent gender gap in programming has not been addressed by the Stack Overflow community. In [3] by *Dada et al.*, the authors exhibited different visualizations of different attributes from the SODS (2020) and measure the frequencies. For example, *JavaScript* was the top programming skills used in various IT roles. The authors identified and ranked the top popular IT tools for 23 distinct IT professions, including operating systems (OS), databases, programming languages, and collaborative applications.

We can see from the aforementioned analysis of the literature that none of the studies—to the best of our knowledge—addressed the research questions we are interested in, and we think our project will help to close this gap.

3 Overview of the SODS Dataset

The SODS dataset consists of 73268 samples where there are different types of questions with different answering formats, such as multiple choice (single answers or multiples answers), Likert, Writing Text in the box. Some questions are mandatory to response, while some are optional (which may cause "NA" in the entries). To better comprehend the SODS dataset and design the pattern mining technique, we separate the dataset into three ***conceptual*** entities:

- ***personal.*** These are the questions related to user demographics, and basic information such as—age, etc. There are total 11 questions.
- ***organizational.*** These are the questions related to workplace type, educational background, salary range, work experience, etc. This category contains a total of 14 questions.
- ***technical.*** These are the questions related to programming languages, development tools, operating system, version controls, etc. This set also contains 14 queries.

Our study will attempt to give a thorough examination of each of these entities separately as well as an analysis of their connections. Note that, there are other types of questions in the SODS that targets SO Usage, Community and the survey itself; we avoid this types from the scope of our project since

[3] https://stackexchange.com/

they are irrelevant to the job-market analysis. We also ignore the questions that requires answers as texts in the boxes.

4 Proposed Work and Tools

Having the motivation behind our project and the collected SODS dataset, we designed the following tasks to reach to our final goal.

- **Data Pre-processing.** At first, we plan to perform data preprocessing. We want to measure data quality by using statistical description and perform data cleaning (e.g. dealing with "NA" entries). We also need data integration for multiple years in order to execute temporal trend analysis. If needed, we also plan to apply data transformations.
- **Data Warehousing and visualization.** We want to store the dataset in an efficient way for better investigation. Thus, we plan to apply data warehousing. Moreover, we will present the important aspects of the data pictorially using graphs and plots.
- **Analysis.** This is one of the crucial tasks our project. This task includes several sub-tasks—*Trend Analysis*, *Frequent Pattern Analysis* and *Correlation Analysis*.
- **Clustering.** Using supervised and/or unsupervised method, we plan to apply clustering algorithms, such as k-means clustering. This approach will help us extracting meaningful clusters from the SODS dataset, e.g., education level vs. salaries.
- **Classification.** We would like to examine different attributes of SODS dataset and to find if any of those can be treated as labels. We wish to implement classification algorithms, i.e., decision tree, kNN classifier. One possible application of this task is classifying salary group based on given experience, work-mode preference, and education information of a test user.

4.1 Evaluation

We plan to perform different evaluation matrices to measure the performance of our system. The tentative matrices are: *Accuracy*, *Precision*, *Recall*, and *F1 Score*. Moreover, we plan to use *Speed*, *Robustness*, *Scalability*, *Interpretabilty*, and *Goodness of rules*.

We will use `Python` programming language for implementation with the few packages, such as `Numpy`, `Scipy`, `Pandas`, `Matplotlib`, `PostgreSQL`, `Keras`, etc.

5 Conclusion and Project Milestone

In this project, we aim to perform exploratory data analysis on the Stack Overflow Developer Survey. Our target is to provide a picture of the current scenario in the tech job market. We want to start answering to some relevant questions to mine the dataset and to perform a hypothesis testing. A tentative milestone of our project is shown in Table 1. Our plan is to finish as much tasks as possible for the project by sixteenth week of Fall 2022.

Table 1. Tentative milestone of our project.

Project Work	Accomplishment week
Data Preprocessing	Week 8
Data Warehousing	Week 9
Data Visualization	Week 9
Trend Analysis	Week 10
Correlation Analysis	Week 12
Frequent Pattern	Week 14
Clustering	Week 15
Classification	Week 16

References

[1] Tanveer Ahmed and Abhishek Srivastava. 2017. Understanding and evaluating the behavior of technical users. A study of developer interaction at StackOverflow. *Human-centric Computing and Information Sciences* 7, 1 (2017), 1–18.

[2] SJ Brooke. 2021. Trouble in programmer's paradise: gender-biases in sharing and recognising technical knowledge on Stack Overflow. *Information, Communication & Society* 24, 14 (2021), 2091–2112.

[3] Oluwaseun Alexander Dada, George Obaido, Ismaila Temitayo Sanusi, Kehinde Aruleba, and Abdullahi Abubakar Yunusa. 2022. Hidden Gold for IT Professionals, Educators, and Students: Insights From Stack Overflow Survey. *IEEE Transactions on Computational Social Systems* (2022).

[4] Arnal Dayaratna. 2021. Quantifying the Worldwide Shortage of Full-Time Developers. https://www.idc.com/getdoc.jsp?containerId=US48223621 Accessed: 2022-09-28.

[5] Denae Ford, Alisse Harkins, and Chris Parnin. 2017. Someone like me: How does peer parity influence participation of women on stack overflow?. In *2017 IEEE symposium on visual languages and human-centric computing (VL/HCC)*. IEEE, 239–243.

[6] Chenbo Fu, Xinchen Yue, Bin Shen, Shanqing Yu, and Yong Min. 2022. Patterns of interest change in stack overflow. *Scientific reports* 12, 1 (2022), 1–10.

[7] Konstantinos Georgiou, Nikolaos Mittas, Alexandros Chatzigeorgiou, and Lefteris Angelis. 2021. An empirical study of COVID-19 related posts on Stack Overflow: Topics and technologies. *Journal of Systems and Software* 182 (2021), 111089.

[8] Iraklis Moutidis and Hywel T. P. Williams. 2021. Community evolution on Stack Overflow. *PLoS ONE* 16 (2021).

[9] Markus Nivala, Alena Seredko, Tanya Osborne, and Thomas Hillman. 2020. Stack Overflow–Informal learning and the global expansion of professional development and opportunities in programming?. In *2020 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, 402–408.

[10] U.S. Bureau of Labor Statistics. 2022. Occupational Outlook Handbook. https://www.bls.gov/ooh/computer-and-information-technology/computer-programmers.htm Accessed: 2022-09-28.

[11] Prasad Patil. 2022. What is Exploratory Data Analysis? https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15 Accessed: 2022-10-5.

[12] Chaiyong Ragkhitwetsagul, Jens Krinke, Matheus Paixao, Giuseppe Bianco, and Rocco Oliveto. 2019. Toxic code snippets on stack overflow. *IEEE Transactions on Software Engineering* 47, 3 (2019), 560–581.

[13] Yuhao Wu, Shaowei Wang, Cor-Paul Bezemer, and Katsuro Inoue. 2019. How do developers utilize source code from stack overflow? *Empirical Software Engineering* 24, 2 (2019), 637–673.