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

Cornelius Onimisi Adejoro
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, 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. This report exhibits the check points of our work and presents intermediate outcomes.

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* [11]—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 [5].

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 [9]. 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) [12] that will help future developers better understand the industry. We propose to answer numerous research questions because our goal is to provide an analytical picture of the tech-job market with the most important ones being mentioned below.

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

- **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 (1):** *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. We present intermediate results in Section 5 with insights. Section 6 concludes the report with future plans.

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.* [7] 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* [3] reveals a dominant male influence. *Moutidis et al.* examines the migration of the users between communities [9]. 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 [2, 13, 14]. *Georgiou et al.* [8] incorporated SODS (2020) and scraped data to find COVID-19 impact on the Stack Overflow. Similarly, *Ford et al.* [6] 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.* [10] 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 [4] 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]https://stackexchange.com/

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 "`NaN`" in the entries).

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 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*.

5 Intermediate Results

In this section we show some of our results as a milestones of our progress. At this moment, we have several modules which we believe we have completed almost and some have been partially finished. Table 1 illustrates a status of our accomplishment for different module. For implementation, we used `Python` programming language for implementation with the few packages, which are: `numpy`, `scipy`, `pandas`, `matplotlib`, `sklearn`, `mlxtend`, `pycountry_convert`, `geopy`, and `geopandas`, etc. A brief discussion on the checkpoints are described below.

Table 1. Progress of our accomplishment.

Project Checkpoints	Status
Data Visualization [5.1]	mostly
Data Preprocessing [5.5]: cleaning, reduction & transformation	partially
Exploratory Data Analysis [5.6]	mostly
Frequent Pattern Mining [5.7]	mostly
Correlation Analysis [5.8]: χ^2 test and lift	mostly
Clustering [5.9]	partially

5.1 Data Visualization

Starting to work with our dataset, as our dataset was large and contained various range of values per sample, first we decided to visualize different properties of the dataset. Visualizations from our different exploration are shown below.

5.2 Attribute wise Unique Values

We attempted to visualize how many unique values we have in each attribute which will help us to be able to choose interesting attributes to work with. Figure 1 shows the demonstration of the number of unique values each attribute. See Fig. 1. This is to assist us have an easier view of the stories surrounding data into a form easier to understand.

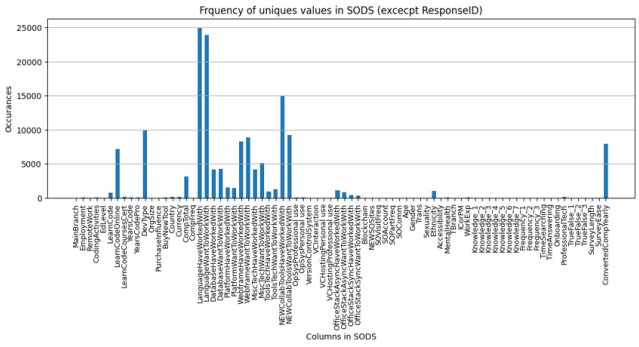

Figure 1. Frequency of unique values in SODS Dataset.

5.3 Gender wise Age Distribution

Focusing specifically on the gender group, we wanted to visualize the distribution of age in each gender to check the percentage in software development. We found an interesting finding here which shows that the age range is similar for both men and women i.e., most of the people involved in software development have an age of 25-34 years old. For the male persons, the percentage is 39.6%, for females it is 46.8%, and for other genders it is 29.5% which can be seen in Fig. 2.

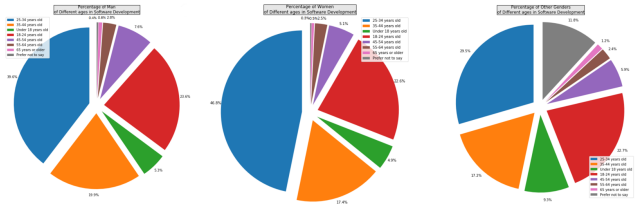

Figure 2. Gender wise Age Distribution. *Left to right:* Male, Female and Other gender.

5.4 Gender wise Participation with average Salary

In Fig. 3, we can see the gender ratio in software development along with each gender's average salary. A significant finding can be seen from this figure that, the participation of men in software development is almost 70000, whereas only around 5000 women are involved in this sector. Also, the average salary for male members is remarkably higher than the female members which is depicted in the figure.

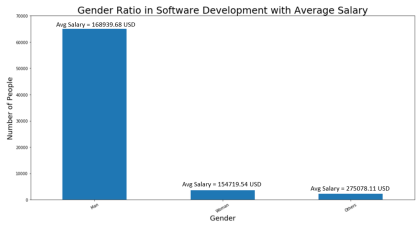

Figure 3. Gender wise participation in software Development with their average salary.

5.4.1 Yearly Salary Distribution. Fig. 4 portrays the information of the yearly distribution of salary of the people in software development. As we got a skewed distribution in 4(a), we performed a log-based transformation in 4(b) to find out a well distributed range which will help us to identify noise, and outliers.

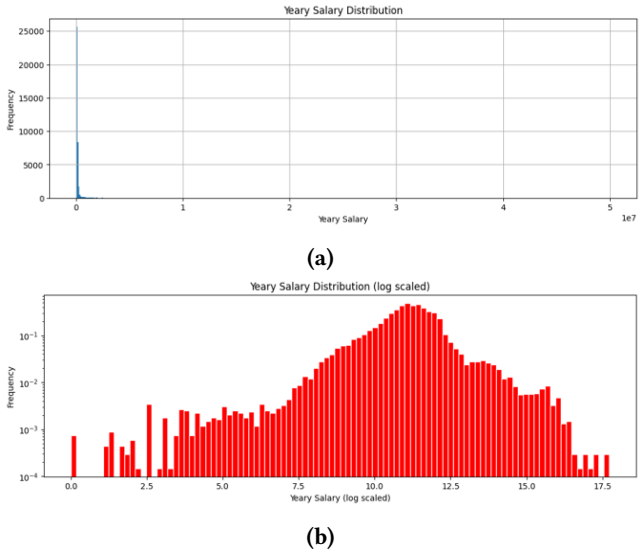

Figure 4. (a) Yearly salary distribution of people in software development, (b) Log-scaled visaulization of (a).

5.5 Data Preprocessing

In order to enhance the performance of the dataset used for this project for all metrics, a preprocessing procedure has been implemented to ensure and to "flesh out" NaN values which represent undefined or unrepresentable data as depicted in Fig. 5. This *cleaning* step is important as to avoid missing values, noisy data, and unwanted inconsistencies in the SOSD dataset. It was at this point that we noted the entire attributes will not be needed for this project.

We also implemented a *reduction* technique that extracts attribute based on a minimum amount of allowed NaN values. Therefore, we can filter out a subset of SODS with a minumum percentage on NaN values, e.g., 5% NaN values (Fig. 6).

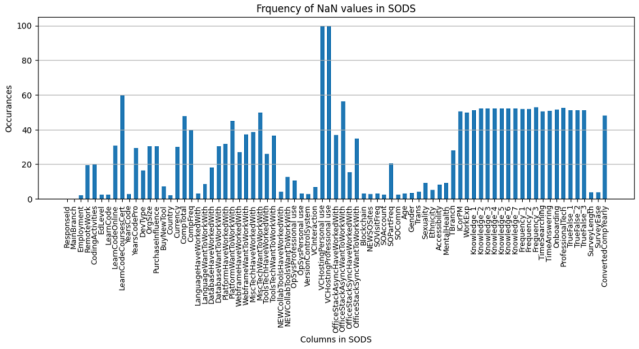

Figure 5. Distribution of NaN values in SODS.

We can also work with a sebset which does not contain any NaN value. The idea is to apply AND operation on the attributes to keep entries which has no NaNs in the rows.

- ResponseId : 0.000
- MainBranch : 0.000
- Employment : 2.128
- EdLevel : 2.316
- LearnCode : 2.304
- YearsCode : 2.644
- Country : 2.043
- LanguageHaveWorkedWith : 3.130
- NEWCollabToolsHaveWorkedWith : 3.987
- OpSysPersonal use : 3.146
- VersionControlSystem : 2.578
- Blockchain : 2.999
- NEWSOSites : 2.597
- SOVisitFreq : 3.149
- SOAccount : 2.315
- SOComm : 2.539
- Age : 3.169
- Gender : 3.296
- Trans : 4.030
- SurveyLength : 3.854
- SurveyEase : 3.767

Figure 6. Attributes which has less than 5% of NaN values in SODS. It shows the name of the attributes with percentages of NaN values.

5.6 Exploratory Analysis

Geolocation based analysis was carried out on one hundred and eighty (181) countries. We investigated to determine the few research questions, i.e., *Developers from which countries get more salaries?*. We faced several challenges to encounter this analysis, for example, *(i)* Not all the countries were in common format, e.g., *North Macedonia* is entered as *The former Yugoslav Republic of Macedonia* in SODS. *(ii)* Lattitude and longitude of the countries were necessary to plot multivariate information on world map for better visualization. Therefore, we performed necessary *transformation* for converting country names. Fig. 7 shows geolocation based analysis of the salary (`ConvertedCompYearly`). It shows the maximum salary (converted to USD) for individual countries. We can observe that, maximum salary occurs in Europe.

5.7 Frequent Pattern Analysis

Since we aim to mine job market related insights from SODS, it is very important to observe the most popular programming language among the programmer. Therefore we attempted to find most frequently used programming language from SODS. Instead of extract one popular language, we worked on extracting a set of frequently used programming language. For the survey, SO allowed the participants to enter multple entries of language as their favorite ones. Therefore,

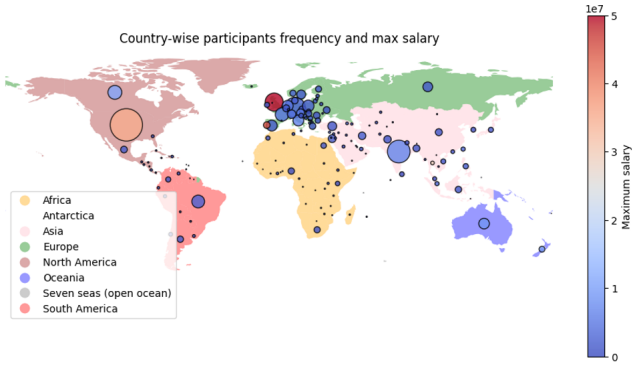

Figure 7. Geolocation based analysis of the salary (`ConvertedCompYearly`). The radius of the circles shows the number of participants in SODS for an individual location. The colors represent the maximum salary with *cool-to-warm* colormap.

we applied *Apriori* algorithm [1] for frequent pattern mining of the ttribute (`LanguageHaveWorkedWith`).

For this purpose we performed preprocssing since the `LanguageHaveWorkedWith` attribute contains values seperated by semicolon (see Fig. 8a), while our implementation required them to be stored as `python`'s list structure. Therefore, we applied simple string splitting and list appending to obtain desired format.

Fig. 8 shows frequent pattern analysis performed to find the most frequently used programming language using the *Apriori* algorithm. The result showed that the frequently used programming languages are `HTML/CSS` and `JavaScript` when minimum support is set for 0.4.

5.8 Correlation

We performed correlation measurement between different attributes of our dataset to find out if there any correlation exists between them. We employed both lift measurement and χ^2 Analysis.

5.8.1 Lift measurement. We computed the lift measurement value calculation to find out if the female members are more likely to work in-person, remote, or in hybrid mode. For this reason, we used the gender and remote work status (`RemoteWork`) attribute. However, a challenge we faced in this case was that our gender attribute contained a lot of noise, e.g., some of the samples contained multiple genders which we needed to handle. At first, we generated the contingency table using the `Gender` (Fig. 9a) and `RemoteWork` attributes and then we computed the lift values. From the lift values, we discovered a positive correlation between female persons and fully remote method of work (Fig. 9b). Therefore, it is evident that female persons prefer to work in fully remote mode, rather than working in in-person or hybrid mode.

	LanguageHaveWorkedWith
1	JavaScript;TypeScript
2	C#;C++;HTML/CSS;JavaScript;Python
3	C#;JavaScript;SQL;TypeScript
4	C#;HTML/CSS;JavaScript;SQL;Swift;TypeScript
5	C++;Lua
...	...
73263	Bash/Shell;Dart;JavaScript;PHP;Python;SQL;Type...
73264	Bash/Shell;HTML/CSS;JavaScript;Python;SQL
73265	HTML/CSS;JavaScript;PHP;Python;SQL
73266	C#;Delphi;VBA
73267	C#;JavaScript;Lua;PowerShell;SQL;TypeScript

(a) A snapshot of `LanguageHaveWorkedWith` attribute

	support	itemsets
0	0.551490	(HTML/CSS)
1	0.654357	(JavaScript)
2	0.481226	(Python)
3	0.494921	(SQL)
4	0.490525	(JavaScript, HTML/CSS)

(b) Result of *Apriori* algorthim for `minimum support` of 0.4.

Figure 8. Frequent Pattern Analysis with *Apriori* Algorithm.

RemoteWork Gender	Full in-person	Fully remote	Hybrid
Female	415	1341	1183
Male	7870	23057	22957
Others	192	668	611

(a) Contingency table for `Gender` vs `RemoteWork`.

```
lift( 0 , 0 ) = 0.9710 [-ve corr]
lift( 0 , 1 ) = 1.0611 [+ve corr]
lift( 0 , 2 ) = 0.9480 [-ve corr]
lift( 1 , 0 ) = 1.0044 [+ve corr]
lift( 1 , 1 ) = 0.9951 [-ve corr]
lift( 1 , 2 ) = 1.0034 [+ve corr]
lift( 2 , 0 ) = 0.8976 [-ve corr]
lift( 2 , 1 ) = 1.0561 [+ve corr]
lift( 2 , 2 ) = 0.9783 [-ve corr]
```

(b) Lift values based on the contingency table.

Figure 9. Frequent Pattern Analysis with *Apriori* Algorithm.

5.8.2 The χ^2 test. To obtain more insights from the previous analysis, we performed a Chi-Square Analysis using the contingency table found in Fig. 9a. For the significance level of 0.05 and degree of freedom, $d = 4$, we used the chi-square table[4] and found that these two attributes are correlated as the computed chi-square value was greater than the critical value. After that, we calculated the expected values (shown

[4]https://people.richland.edu/james/lecture/m170/tbl-chi.html

in matrix 1) and mined some significant insights.

$$e_a = \begin{bmatrix} 427.383 & 1263.748 & 1247.867 \\ 7835.706 & 23169.731 & 22878.561 \\ 213.909 & 632.519 & 624.570 \end{bmatrix} \quad (1)$$

For instance, as we can see from expected values, $o_{11} < e_{11}$, $o_{12} > e_{12}$, and $o_{13} < e_{13}$, we can say that female persons are more likely to work in fully remote mode, rather than working in-person or in hybrid mode which makes the claim from lift analysis stronger as the claims are same in both cases.

Similarly, we executed another chi-square test on the attributes remote work status (`RemoteWork`) and education level (*EdLevel*) to find out if the PhD persons prefer to work remotely or in-person. Here, the attribute `EdLevel` had five unique values and we categorized them as PhD or non-PhD. Then, we generated a contingency table using these two attributes (Fig. 10). From the contingency table, for $d = 2$ and significance level 0.05, we get our computed χ^2 value greater than the critical value. Also, from the expected values shown in matrix 2, we see, $o_{21} > e_{21}$, $o_{22} < e_{22}$, and $o_{23} > e_{23}$, it reflects that the PhD persons are more likely to work in person or hybrid, rather than working remotely.

RemoteWork	Full in-person	Fully remote	Hybrid
EdLevel			
Non-PhD	8294	24720	23900
PhD	302	621	1121

Figure 10. Contingency table for `EdLevel` vs `RemoteWork`.

$$e_b = \begin{bmatrix} 8297.98 & 24462.45 & 24153.55 \\ 298.01 & 878.54 & 867.44 \end{bmatrix} \quad (2)$$

5.9 Clustering

As an experiment, we implemented clustering algorithms. We considered *yearly salary* (`ConvertedCompYearly`) and *years of coding* (`YearsCode`). We applied k-means clustering algorithm for this two attributes and result is presented in Fig. 11. The clustering output may help us in future to determine noise. For example, the green points can be considered as noise since higher salaries for less number of years coding seems impractical.

6 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. Our future plan is to complete the rest of our plans, which are classification, evaluation anf more exploratory analysis. We also want to test the hypothesis (1) using Naïve Bayesian classifier.

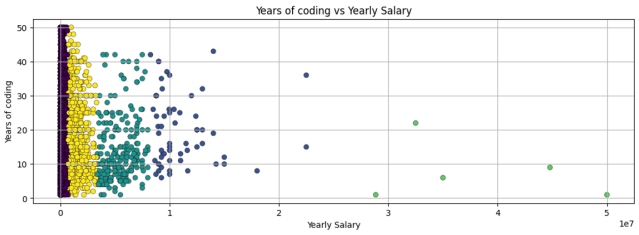

Figure 11. K-means clustering on *yearly salary* vs *years of coding* for $k = 5$.

References

[1] Rakesh Agrawal, Ramakrishnan Srikant, et al. 1994. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, Vol. 1215. Santiago, Chile, 487–499.

[2] 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.

[3] 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.

[4] 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).

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

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.