Visualizations of Stack Overflow Developers' Survey

Ang Wei Xuan Dion, David Chow Jing Shan, Peh Anqi

Abstract— Stack Overflow is arguably the biggest online community for developers all around the world. Each year, Stack Overflow fields a survey with questions ranging from developers' favourite technologies to their job preferences. This is done to allow Stack Overflow to better understand its active users. Although the official Stack Overflow website provided a series of visualizations for the result of their survey questions, however, most of them are static and do not allow further exploration and in-depth analysis. Therefore, there is a need for a more in-depth and flexible visualization tool, so that any user can explore the data easily and acquire different insights according to their own needs. After much deliberation, we have narrowed down to three core aspects for in-depth analysis: Technology, Salary and Job analysis. This research paper aims to analyse the available data, explore some of the key findings of this survey and provide some insights related to the active developers of the Stack Overflow community. The visualization tools – Choropleth Map, Age-gender pyramid, Network diagram, Scatter plot, Divergent bar chart and etc. are developed to aid users in data exploration and analysis.

Index Terms—Stack Overflow, Survey results, Developers, Programming, Job factors, Salary, Technologies

1 Introduction

Stack Overflow's annual Developer Survey is the largest and most comprehensive survey of people who code from around the world. Each year, they field a survey covering everything from developers' favourite technologies to their job preferences. 2019 marks the ninth year that they've published their annual Developer Survey results and nearly 90,000 developers took the 20-minute survey earlier this year.

Despite the survey's broad reach and capacity for informing valuable conclusions, the results are not equally representative of everybody in the developer community [1]. However, it is still worthwhile to look at what are some of the main characteristics of these active developers in Stack Overflow as it can still reveal some potentially useful information/trends for references.

Stack Overflow themselves has prepared various visualizations for their survey finding. However, many are either static or are displayed in multiple tabs that do not allow for easy comparison and further exploration and in-depth analysis. For instance, the visualizations of the distribution of developer type and gender are only confined to either all respondents or the respondents from the United States, with no option for filtering the results by other countries. Therefore, we present DevBuzz – a dynamic visualization platform specially designed for users to understand the findings of Stack Overflow's annual survey in a deeper context.

Consisting of 8 main sections, this research paper reports our research and development efforts from brainstorming, designing and implementing the comprehensive and interactive application that aids in visualizing the results of the Stack Overflow 's annual survey.

2 MOTIVATION AND OBJECTIVES

Our development efforts were motivated by the lack of interactive visualization provided to visualize the survey findings. Currently, many useful filter options such as country and gender are not available for users to further explore and analyse the survey results. Additionally, users will have to tediously scroll through many visualizations just to reach the one that they may be interested in. Hence, we aim to provide a more comprehensive and simple way for users to do more exploration and in-depth analysis.

After studying the survey questions and results, our team has narrowed down to three main areas (excluding demographic) that we think most general users can benefit from this survey:

1. Technology - To identify commonly used tools (e.g. programming languages and platforms) and their potential complementary/related tools. Additionally, we will also explore what are the most loved, dreaded and desired tools according to the respondents.
2. Salary - To find out the salary distribution of different developer types and how they may be influenced by their undergraduate major and how it differs from country to country.
3. Job analysis - To identify prioritized job factors and how it differs between Man and Woman and varies from country to country. The same will be done for job satisfaction and working hours.

To help us expedite the building of these interactive visualizations, our team has decided to use R which gives us access to a wide variety of libraries and tools for data pre-processing and building user-friendly dashboards.

3 RELATED WORKS

As mentioned previously, existing visualization for the survey findings are either static or have been charted in multiple tabs, that does not provide users with the flexibility to conduct further exploration and in-depth analysis.

Figure 1 is a proportional symbol map to represent the demographic of the survey respondents. The visualization has a tooltip that allows users to view the country and the percentage of respondents from that country.

[1] https://insights.stackoverflow.com/survey/2019

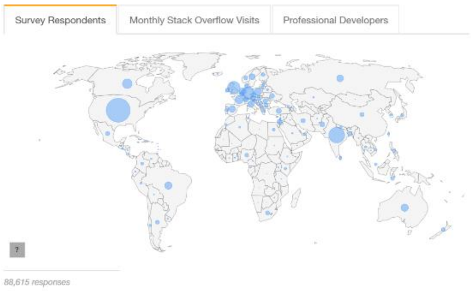


Fig. 1. Geography of Survey Respondents

However, the proportional symbol map does not provide a clear comparison between the countries especially for countries around Germany, UK, and Poland, where there are many overlaps.

Looking at Figure 2, we can see the correlations between the different types of technologies. This provides a good overview of which technologies are highly used together in the ecosystem.

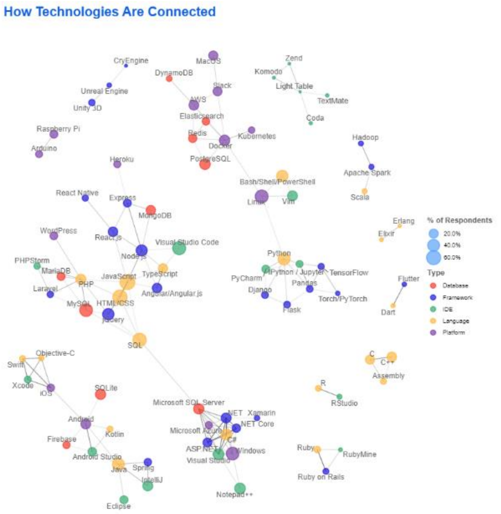


Fig. 2. Correlation of technologies

However, this visualization is a static chart. Users would not be able to find out more information from the visualization, such as which skills are used more by which developer types, or have the flexibility to select the type of technology that they wish to look at.

From Figure 3, we can see a comparison of the median salary for the different developer types. We are also able to see which developer jobs have higher salaries even though the developers have similar coding experience in terms of year.

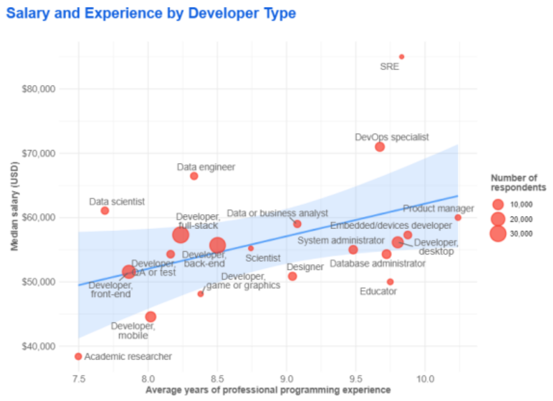


Fig. 3. Salary and experience by Developer Type

However, this visualization is also static. This can make smaller dots unclear. For example, it is difficult to find out how many respondents working as SRE provided information for this data. Our team also believes that there can be other hidden factors that can affect salary such as country due to differing standards of living. Hence, we feel that would be something interesting for the users to explore.

Looking at Figure 4, we can see the top programming languages loved by the respondents.

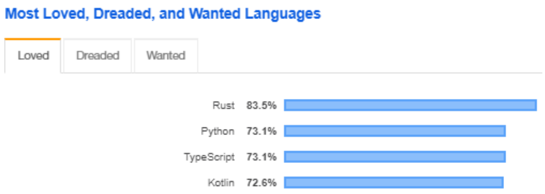


Fig. 4. Most Loved, Dreaded and wanted Languages

However, the bar chart for Dreaded and Wanted programming languages are in other tabs. This makes it difficult for the readers to compare across the three tabs. We feel that this chart is not only relevant for the programming language, this also can be plotted for the other technology types. (I.e. Database, Framework and Platform)

In summary, our group gained a deeper understanding of the strengths and limitations of the visualization currently available. We have brainstormed and design our application to provide more interactivity with the help of filters and tooltip to allow users to conduct more exploration and in-depth analysis.

4 VISUALIZATION APPROACH

4.1 Exploratory Data Analysis (EDA)

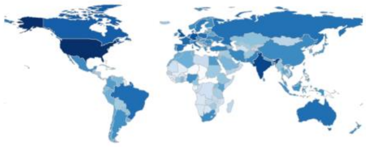
Apart from the survey schema, all the survey results are recorded in the same excel file. There is a total of 85 columns (survey questions). We categorize the questions into three main groups: Background, Job Prospects, and Skills. Through EDA performed in Microsoft Excel and Tableau, we learnt that most of the respondents are men and most of them are from the United States, India, and Canada.

Next, our team also began exploring the most used languages and we learnt that the top languages used are JavaScript, HTML/CSS, SQL, and Python. We have also explored the declared salaries earned by various developers and the top earners usually come from Site Reliability Engineers and DevOps Specialists. Additionally, we also investigated the general job priorities and found out that the technologies that they will be working with and good office environment are among the top job factors that they look out for.

4.2 Brainstorming and design consideration

After EDA, the team began researching various visualization ideas and r packages that we could use to produce the necessary dashboards. Our team discussed each graph's suitability and ensure that every visualization that we choose is the best in terms of clarity and aesthetic. We matched the data type to the graphs to ensure data is being accurately presented. Furthermore, we also deliberated on the suitable interactivity and filters for each graph. Below are the main graphs that we have cherry-picked to visualize the survey results.

4.2.1 Choropleth Map

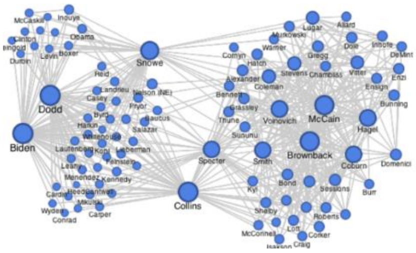


We have decided to use a Choropleth map[2] instead of a Proportional symbol map to showcase the distribution of survey respondents. This is because we have found it better at visualizing the variability of the respondents' locations across the different regions. Unlike the Proportional symbol map, the Choropleth map will not have the problem of symbols overlapping each other, allow users to view the distribution of respondents on the entire world map clearly.

4.2.2 Bar Diagram

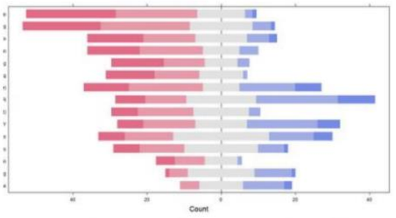
Even though it is simple, the bar diagram is very effective in showing the proportion of data and allow easy comparison. As many of our data involve some sort of ranking order, we feel that the bar diagram is the clearest way for users to view them. Additionally, we have decided to use horizontal bar diagram over vertical bar diagram because many factors have names that are too long and/or complicated to abbreviate. Thus, using a horizontal bar diagram will provide more spacing for the longer names.

4.2.3 Network Graph



Network graphs are useful for visualizing relationships between entities[3] and our team feels that it is perfect to show off the potential linkages between technologies. The size of the nodes reflects the size of the respective technology's userbase and the edges will link to their potential complementary/related tools.

4.2.4 Divergent Bar Chart



As we have Likert scale data (Job Satisfaction) that we would like to visualize, our team feels that the divergent bar chart is the best for this task. It positions the responses horizontally such that positive responses are stacked to the right of a vertical baseline and negative responses are stacked to the left of this baseline. Meanwhile, the reference line is positioned at the neutral responses. This is a very effective way for users to get an overview of Likert scale data and it will be easy for them to find out which are the areas that most respondents feel strongly about.

4.2.5 Scatter plot

A Scatter plot is simple yet effective in depicting relationships between variables (Salary and years of professional coding in our case). Together with the regression line, scatter plot can potentially show how strong the relationship between the two variables is, and if there are any unusual points (developer types).

4.3 Data Cleaning, Data Pre-processing, and Implementation

To fit the data into the desired graph, we used Microsoft Excel to remove redundant information such as columns/questions that are not useful for our visualizations. Next, we used Pandas library in Python to convert columns with multiple answers into their own individual columns (binary).

During the implementation stage in R, we used libraries such as tidyverse, dplyr to filter, subset, mutate and aggregate data as needed by the respective graphs before plotting them using Plotly or ggplot2.

5 DATA VISUALIZATION WALKTHROUGHS

There are 5 main dashboards available in our application, which is the Home page, Demographic, Technologies, Salary and Job factors.

5.1 Home

The homepage provides overall information for readers to understand the problem and motivation and objective of the visualization.

5.2 Demographic

The demographic page provides readers with the overall demographic information about Stack Overflow developers.

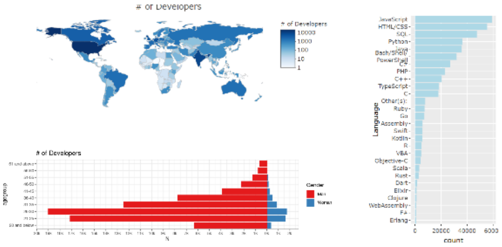


Fig. 5. Demographics of respondents

The choropleth map at the top shows the distribution of respondents. The age-gender pyramid at the bottom shows the distribution by age group and gender. By clicking on a country on the choropleth map, the age-gender pyramid will change according to the country selected.

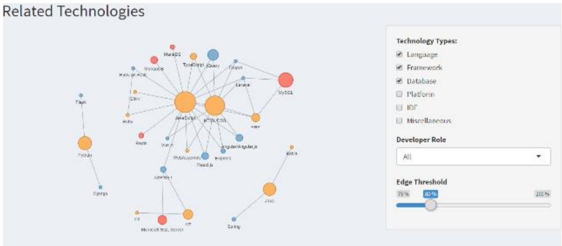
On the right side of the dashboard is a bar chart that shows the number of developers who are using the programming languages. The bar chart is sorted in descending order and act as a filter for both the choropleth map and the age-gender pyramid.

[2] https://www.arcgis.com/apps/MapJournal/index.html?appid=75eff041036d40cf8e70df99641004ca

[3] https://www.analyticsvidhya.com/blog/2018/04/introduction-to-graph-theory-network-analysis-python-codes/

5.3 Technologies

The technologies page provides 3 visualizations. Firstly, it is Related technologies and secondly is Loved technologies. Thirdly, it is Desired Technologies.



Technology Types:
☑ Language
☑ Framework
☑ Database
☐ Platform
☐ IDE
☐ Miscellaneous

Developer Role: All

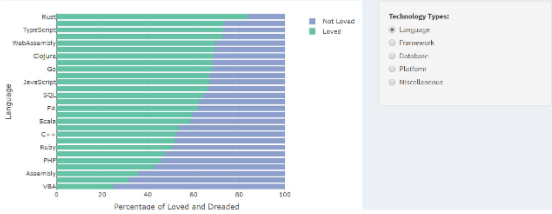
Edge Threshold: 75% 80% 100%

Fig. 6. Related Technologies

This visualization shows the correlation between the different technology types. If two nodes are connected, this would mean that they are often used together.

The size of the node is the number of respondents using the technology and line between two nodes indicates that there is a correlation between the two technologies. Readers can use the filter on the right-hand side to pick the technology types (each has its unique node colour) they wish to include in the visualization. They also can filter by Developer Role and Edge Threshold. Edge Threshold is used to draw the correlation between technologies. The default, 80%, would mean that the edges draw between nodes at least have 80% of the correlation between them. Through trial and error, our team decided to set the default to 80% as the network graph is not too crowded and is still able to highlight the major correlated technologies to the user.

From this visualization, readers will be able to find out which technologies are mostly used together. This can be helpful for them when deciding which technologies should they learn, given that they are already skilled in one or more technology.



Technology Types:
● Language
○ Framework
○ Database
○ Platform
○ Miscellaneous

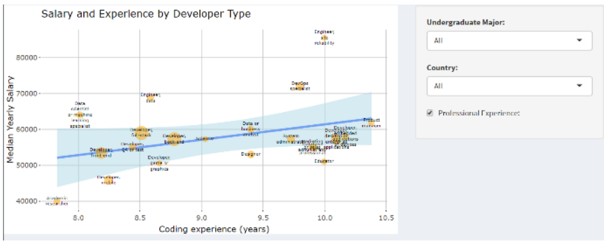
Fig. 7. Loved Technologies

The 100% stacked bar chart in Fig 7 shows the loved against not loved distribution for the different programming languages. The chart is sorted in descending order with the language with highest percentage of loved at the top and the lowest percentage of loved at the bottom. The filter at the right allows users to change the technology type. They can pick between language, framework, Database, Platform or Miscellaneous. The bar chart wants to show the user which are the technologies that are most popular and "in trend" among the respondents.

Using a similar bar chart, the chart of Desired Technologies will show the technologies that most developers would like to work with next.

5.4 Salary

The salary page provides users with a scatter plot to view the median salary for different Developer Type.



Undergraduate Major: All

Country: All

☑ Professional Experience:

Fig. 8. Salary and Experience by Developer Type

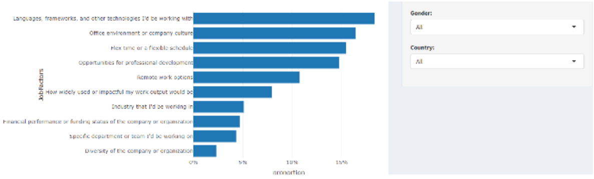
The scatter plot in Fig 8 shows the median salary (USD) against the average years of coding experience for the different Developer Type. The linear regression line aims to help users understand which developer jobs are higher paying and on average how many years of coding experience do developers in that job type have.

The filters on the right allow users to filter the chart base on Undergraduate Major and Country. As we believe that different country's economy can affect the Salary, this would be an interesting area for the users to explore.

By default, the average years of coding experience are professional experience. However, many developers also spend years learning how to code and code as a hobby. When the user unchecks the box for professional experience, the chart would change the x-axis in the chart to become years of coding experience, which includes coding professionally and non-professionally.

5.5 Job factors

The Job Factors page provides 3 visualizations. Firstly, is Job Priorities, secondly is Job Satisfaction and lastly is working hours.


Languages, frameworks, and other technologies I'd be working with
Office environment or company culture
Flex time or a flexible schedule
Opportunities for professional development
Remote work options
How widely used or impactful my work output would be
Industry that I'd be working in
Financial performance or funding status of the company or organization
Specific department or team I'd be working on
Diversity of the company or organization
JobFactors
0%
5%
10%
15%
proportion

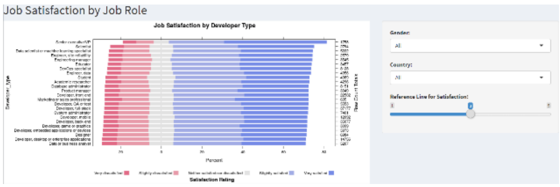
Gender: All

Country: All

Fig. 9. Job Priorities

The bar chart in Fig. 9 shows factors that respondents will consider when applying or accepting a job. Overall, we can see that the top factor is language, framework and other technologies that they will be working with. Hovering over the chart would also provide a detailed percentage of how many percent of respondents voted for the factor that is important to them.

The filter on the right allows the user to filter the chart by gender and country. This is because we think that the priorities of individuals might defer due to these two factors.



Gender: All

Country: All

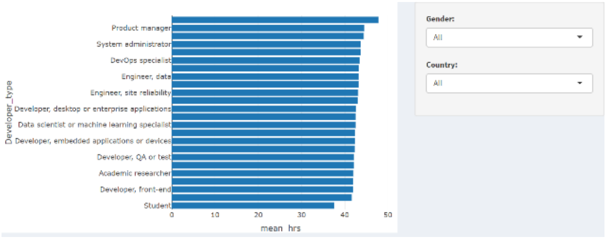
Reference Line for Satisfaction: 1 3 5

Fig. 10. Job Satisfaction

The diverging stacked bar chart in fig 10 provides an overview on which developer has higher job satisfaction based on the developer's job type. The count of respondents for each developer type is displayed at the right of the chart.

The users can select which value on the Likert scale: Very Dissatisfied (1), Dissatisfied (2), Neutral (3), Satisfied (4) or Very Satisfied (5) as the reference point on the chart. The reference line will then shift accordingly.

The user can also filter the chart by gender and country for further analysis.



Gender:
All

Country:
All

Fig. 11. Working Hours

The bar chart in fig 11 shows the mean working hours for each developer type. We can see that the overall Product manager has the longest working hours, followed by System administrator.

The user can also filter the chart by gender and country.

Overall, these 3 visualizations aim to help users, especially students who aim to work as a developer understand more about each developer type and provide them some information to help with deciding which developer position to apply for.

6 KEY FINDINGS AND OBSERVATIONS

6.1 Salary distribution of different developer types

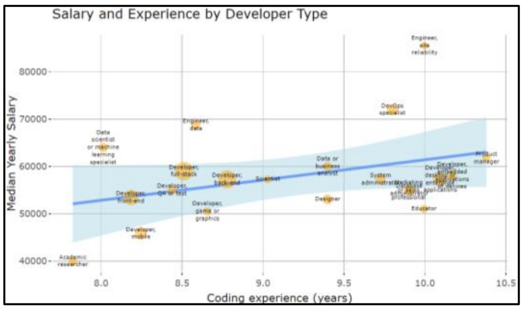


Fig. 12 Overall salary and experience by Developer Type

Overall, the site reliability engineer has the highest median salary, followed by DevOps Specialist with more than 9.5 years of professional coding experience. Naturally, professions with higher coding experience should have a higher salary. Data scientists and Data engineers are high earners for their level of experience. While application developers (mobile and desktop) are low earners for their high level of experience.

From the demographic, we can tell that most of the respondents are from the US and India. Hence, we would like to compare the salary of the two countries.

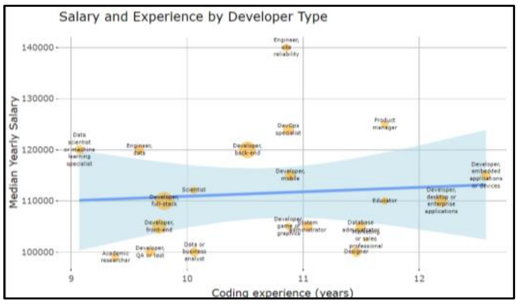


Fig. 13 Salary of developers in the United States

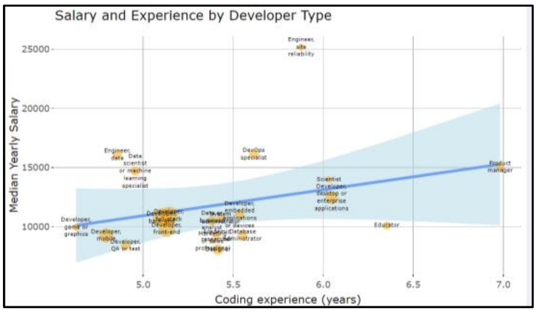


Fig. 14 Salary of developers in India

Comparing both figures above, we can tell that developers in the US are paid a lot more than the developers in India. The salary across the different developer types in India is also much more drastic than in the US.

In both countries, high earners are Site reliability developers, DevOps specialists, Data Scientists, and Data Engineers.

In the US, the average years of professional experience for Mobile and Game developers are more than 10 years. However, interestingly, the average years of professional experience for Mobile and Game developers in India are less than 5 years.

6.2 Identify commonly used technological tools and their related/complementary technologies

Looking at the highest-earning developer type, Site Reliability engineer, we found out that their most-used languages are Javascript, HTML/CSS, SQL, BASH and Python. While Javascript and HTML are at the center of the cluster, and many technologies are linked to them, Python and Bash have far fewer links. In terms of database, most of them used MySQL and PostegreSQL. As for platforms, most preferred Linux, Docker, and AWS.

After comparing with the technologies used by other top-earning developer types such as DevOps Specialists, most of the technologies are consistent with that of Site Reliability engineer. Hence, apart from the necessary job domain knowledge, developers might also want to consider mastering the tools and technologies that are at the center of these clusters for a potentially better job prospect.

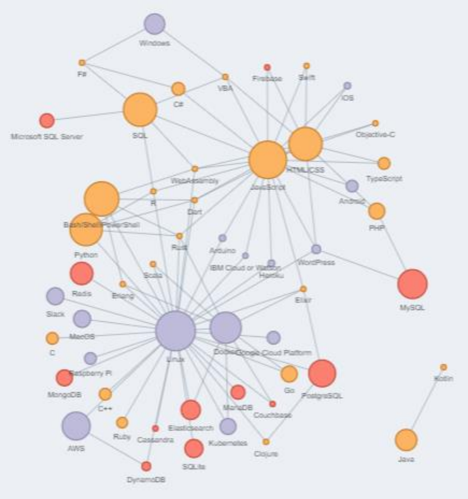


Fig. 15 Common and related technologies used by Site Reliability Engineer

6.3 Prioritized job factors and Job satisfaction

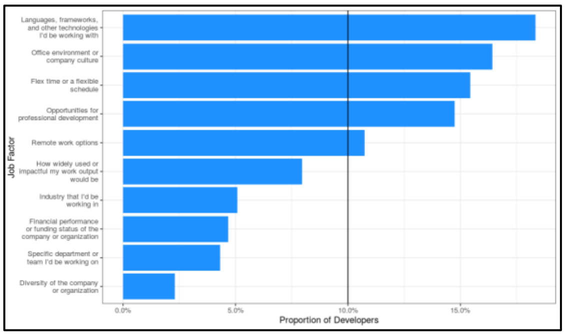


Fig. 16 Overall Prioritized Job Factors

The top three job factors that respondents prioritized are technologies that they will be working with, office environment and having a flexible schedule.

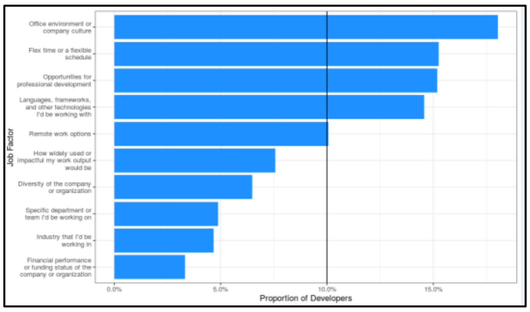


Fig. 17 Prioritized Job Factors for Female

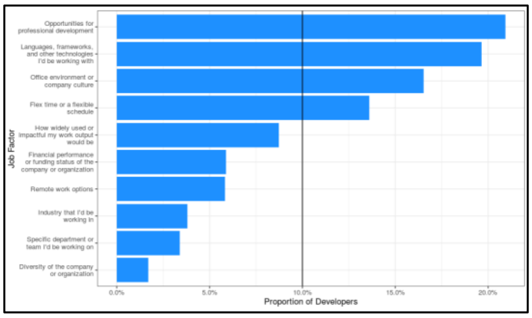


Fig. 18 Prioritized Job Factors for Female in India

Males have the same top priorities as Fig. 16. Between the two genders, Females prioritize opportunities for professional development more than males. This is especially important to the females in India, where more than 20% of them rank it as the top priority when looking for a job, as shown in Fig. 18.

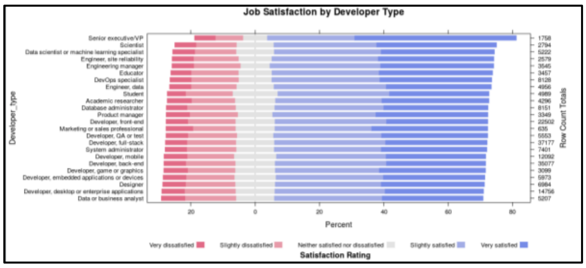


Fig. 19 Overall Job Satisfaction by Developer Type

From Fig. 19, we can see the overall job satisfaction for each developer type. Developer Type that we have identified as high earners in section 6.1, Site reliability engineer, DevOps Specialist Data scientist, and Data engineer have a high job satisfaction rate. Application developers (mobile and desktop) that we have identified as low earners are ranked the 2nd and 4th jobs with the least job satisfaction out of all the developer types. This could indicate that salary have a big part to play in job satisfaction.

7 LIMITATIONS

Once again, the survey data is not representative of the entire Stack Overflow community. The results and observations of this data can only be used as representative of the entire Stack Overflow community or the general coding community.

8 CONCLUSION

Overall, our team found it challenging to analyse and visualize the general developer community. Further research and efforts are required to understand the trends in the developers' vast community. However, throughout the process of design and development, our team learnt many valuable lessons. For the technical aspect, we learnt how to do data pre-processing and use multiple R packages such as Tidyverse, Plotly, ggplot2, and Likert. From our developer topic, we had the chance to know a little bit more about the active developers within the Stack Overflow community.

For future work, our team could explore: 1. Discover features that might affect the employability of developers 2. Uses past data to work on a prediction model that can predict the employability of a developer when applying for a specific role given his/her credentials and experience

ACKNOWLEDGMENTS

We would like to thank Professor Kam Tin Seong from Singapore Management University for his guidance and advice during class and consultation sessions. His valuable insights had helped my team to remain on the right track and implement changes that improved our application.

REFERENCES

[1] Stack Overflow (2019). Developer Survey Results 2019. Retrieved fromhttps://insights.stackoverflow.com/survey/2019

[2] DeLorenzo N., Dugger A. (n.d.). Choropleth Map. Retrieved fromhttps://www.arcgis.com/apps/MapJournal/index.html?appid=75eff041036d40cf8e70df99641004ca

[3] Srivatsa (19 Apr 2018). An Introduction to Graph Theory and Network Analysis (with Python codes). Retrieved from https://www.analyticsvidhya.com/blog/2018/04/introduction-to-graph-theory-network-analysis-python-codes/