IBM Data Analyst Capstone Project Report

2019 Stack Overflow Developer Survey Analysis

Ong Wei Quan

5th July 2024

IBM Developer

SKILLS NETWORK

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- References
- Appendix

EXECUTIVE SUMMARY

- The average age of respondents (total 11,398) was 30.78 years. 93.5% of respondents were men (6.5% were women).
- JavaScript and HTML/CSS were the top 2 most frequently used and sought-after language.
- Remarkably, Python was the fifth-most used language, but it was the third most sought-after language, indicating the increasing trend in usage of Python.
- Software developers should pick up C due to its remarkably high demand in job postings.
- MySQL, Microsoft SQL Server and PostgreSQL were the top 3 most used database.
- PostgreSQL was the most sought-after database, indicating the increasing trend in using PostgreSQL.
- Nearly half of the respondents based on gender had a Bachelor's Degree.

INTRODUCTION

- In 2019, a worldwide online survey of software professionals conducted by **Stack Overflow** was released as open-source data.
- The 20-minute online survey recorded approximately **90,000 responses** by developers.
  - Note: For this survey analysis, only **around one-tenth (~10%) of the survey data was used** as outlined by the IBM Data Analyst course.
- In this report presentation, valuable insights will be explored from the 2019 Stack Overflow Developer Survey, including certain trends and patterns.
- The following report presentation is aimed towards organisations and its employers as well as employees and job seekers regarding insights of emerging skills in the software field.

INTRODUCTION – Survey Analysis Questions

1. What are the top programming languages in demand?
2. What are the top database skills in demand?
3. What are the education levels of software professionals?

METHODOLOGY

- Dataset used for survey analysis:
  - Stack Overflow Developer Survey 2019 (link)
    - Note: The above link to the survey data is only around one-tenth (~10%) of the survey data of the total survey data as outlined by the IBM Data Analyst course (11,398 samples used).
  - Job Opening JSON file (link)
    - Note: For discussion purpose.
  - Popular Programming Languages HTML (link)
    - Note: For discussion purpose.
- Platforms used:
  - Python using Jupyter Notebook
  - Python Libraries (Pandas, NumPy, Matplotlib, Seaborn, Requests, BeautifulSoup, SQLite)
  - IBM Cognos Analytics
  - Microsoft Excel

Exploratory Data Analysis and Data Visualisation was conducted to derive insights.

RESULTS

- Visualization – Charts
  - Demographics (Background) – Age & Gender of Respondents
  - Programming Language Trends
  - Database Trends
  - Education Levels
- Dashboard
  - Current Technology Used
  - Future Technology Used
  - Demographics of Respondents

DEMOGRAPHICS OF RESPONDENTS

Majority of respondents were men (93.5%, total 10,480 men).

The **average** age (mean) of respondents was **30.78 years**.

The **median** age (50%) of respondents was **29 years old**.

	Age
count	11111.000000
mean	30.779318
std	7.393680
min	16.000000
25%	25.000000
50%	29.000000
75%	35.000000
max	99.000000

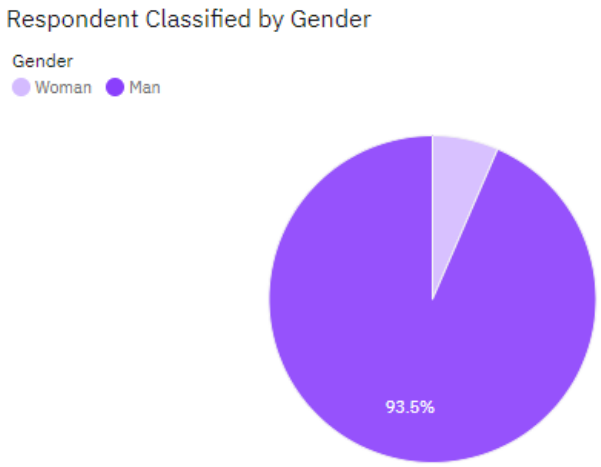


Figure 1: Percentage of respondents classified by genders (Note: Only Man and Woman were included as instructed in Module 5).

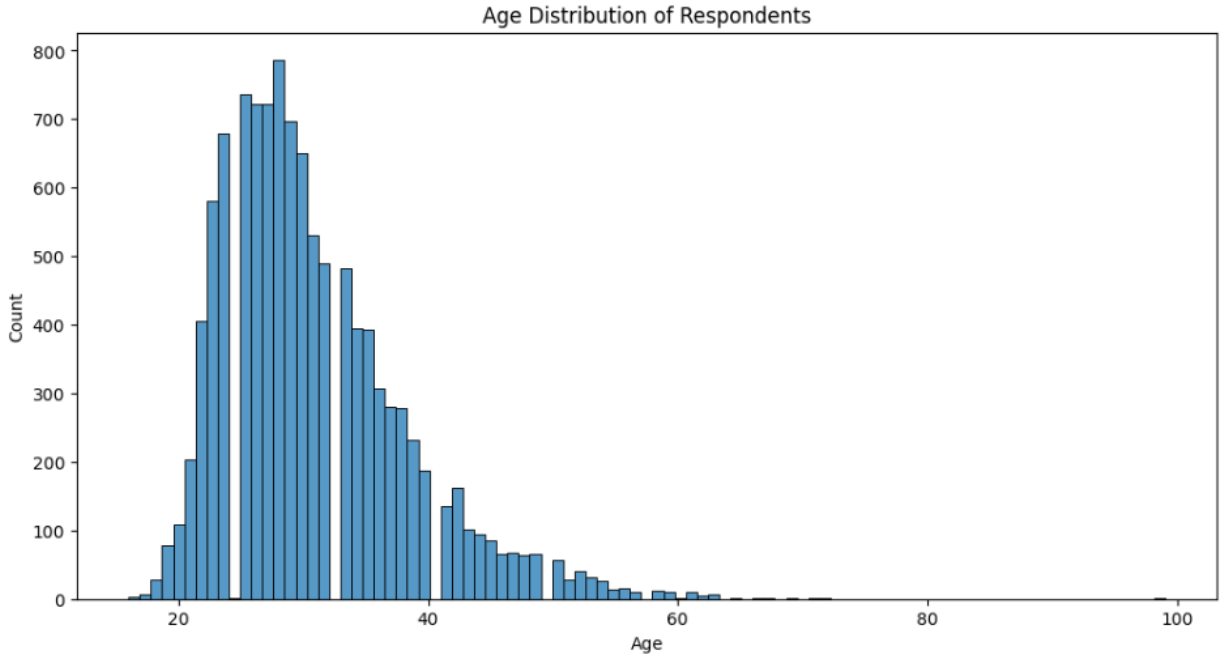


Figure 2: Age distribution of respondents.

PROGRAMMING LANGUAGE TRENDS

Current Year

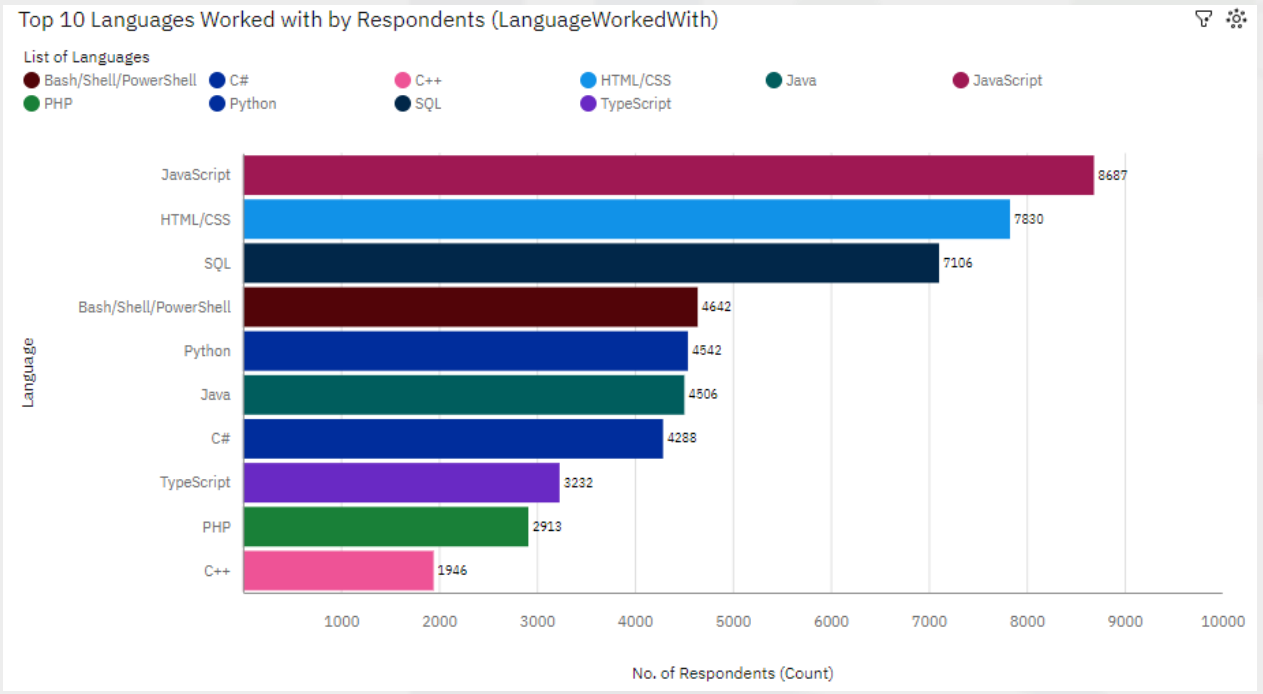


Figure 3: Top 10 Languages worked with by respondents.

Next Year

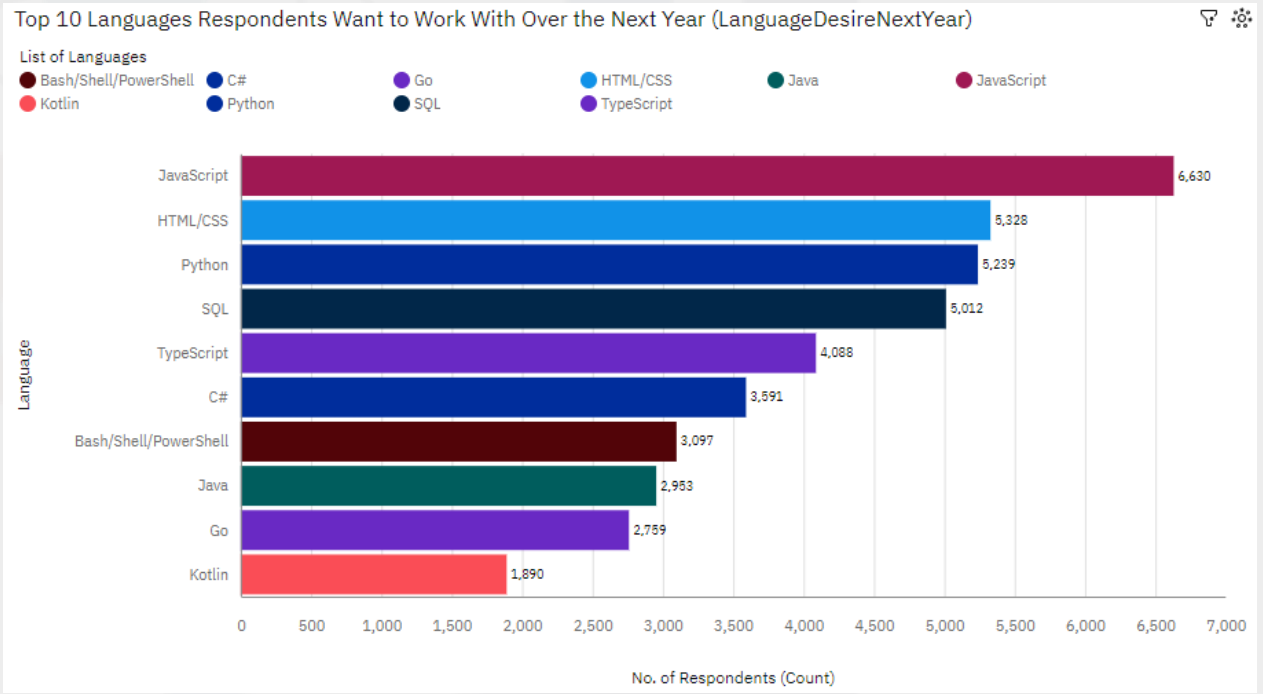


Figure 4: Top 10 Languages respondents want to work with over the next year.

PROGRAMMING LANGUAGE TRENDS – FINDINGS & IMPLICATIONS

Findings:

- JavaScript was the **most frequently used and the top programming language respondents want to learn over the next year**, followed by HTML/CSS.
- Python was the **fifth-most frequently used** programming language, but it was the **third-highest programming language respondents want to learn over the next year**.

Implications:

- JavaScript and HTML/CSS are programming languages that aspiring software developers should learn.
- Python may be starting to gain importance in the software industry, hence more software developers should start learning Python if they have not.

DATABASE TRENDS

Current Year

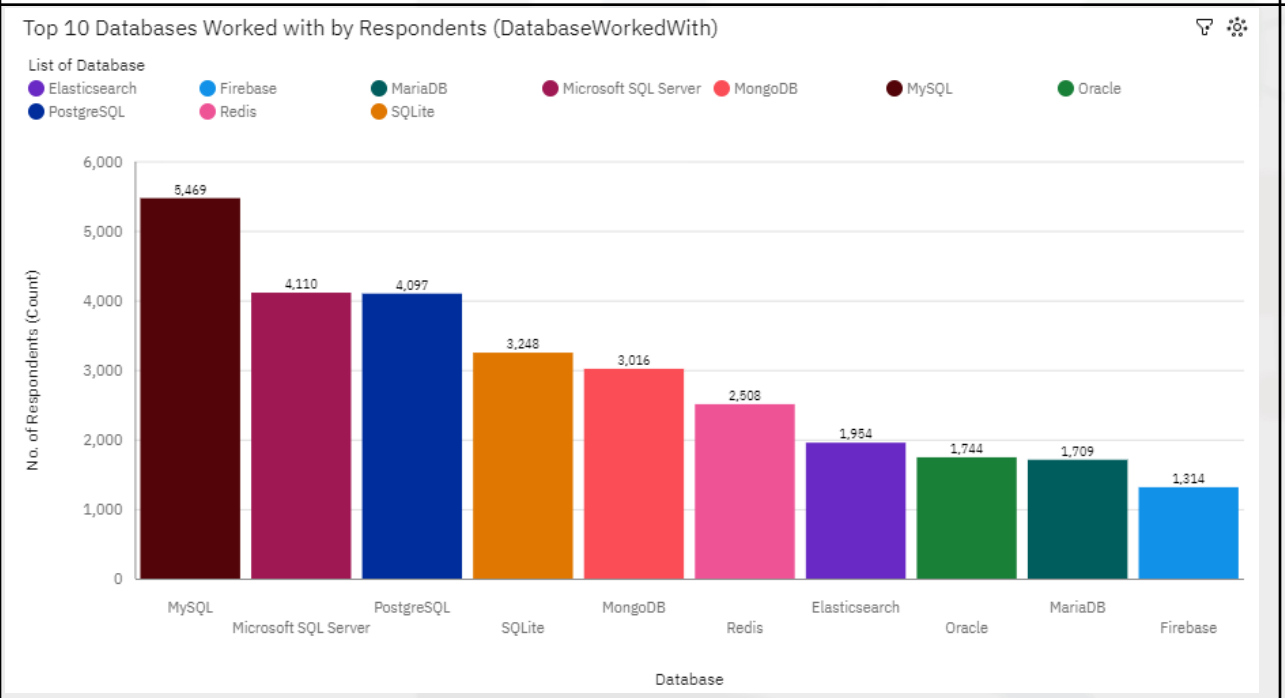


Figure 5: Top 10 Databases worked with by respondents.

Next Year

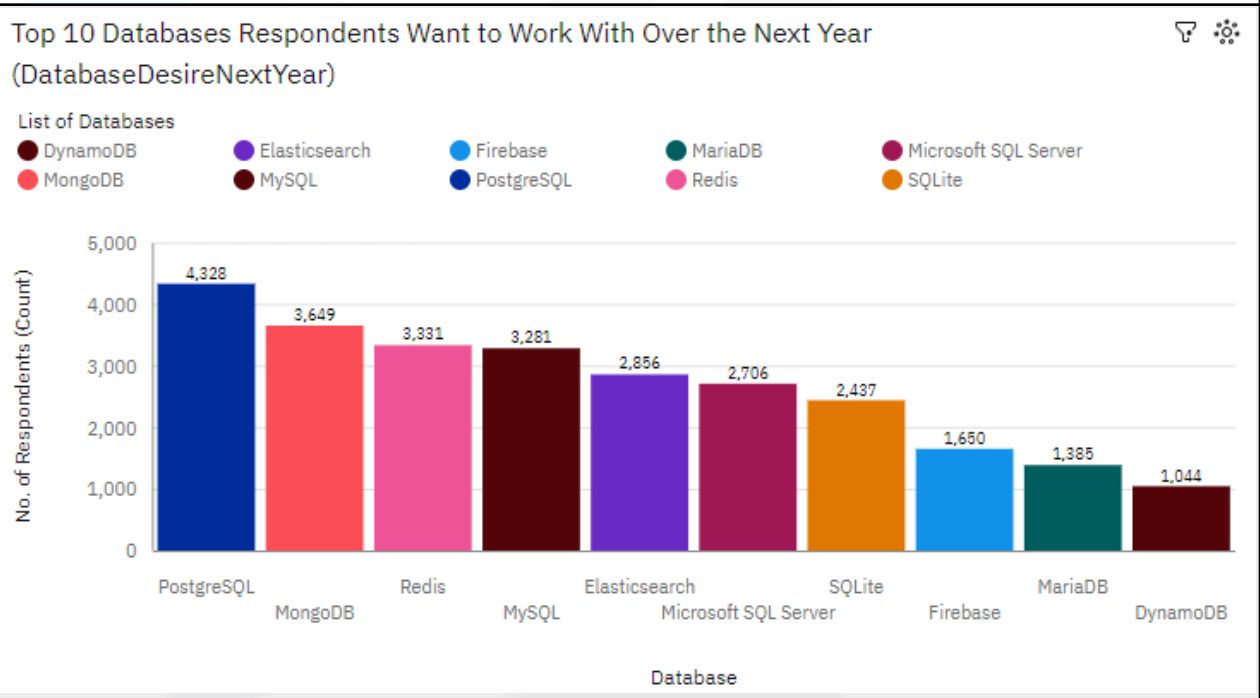


Figure 6: Top 10 Databases respondents want to work with over the next year.

DATABASE TRENDS – FINDINGS & IMPLICATIONS

Findings:

- MySQL was the **most frequently used** database, followed by Microsoft SQL Server and PostgreSQL.
- PostgreSQL was the **top database respondents would like to learn over the next year**, followed by MongoDB and Redis.

Implications:

- The first database aspiring software professionals should learn is MySQL due to its popularity.
- PostgreSQL may be starting to get popular in the industry, thus many respondents may want to pick up PostgreSQL over the next year.

EDUCATION LEVELS – FINDINGS & IMPLICATIONS

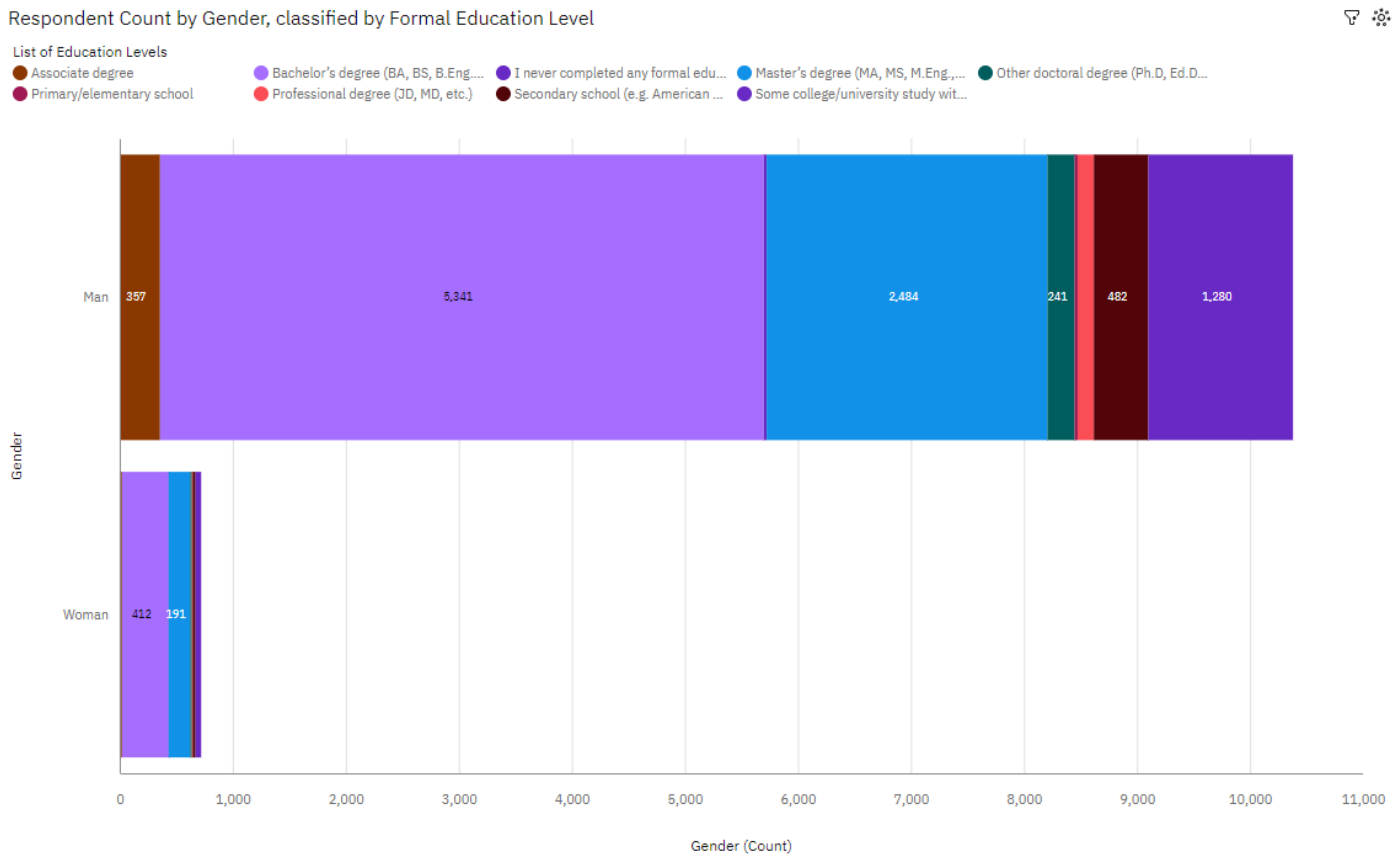


Figure 7: Respondent Count by Gender, classified by Formal Education Level.

EDUCATION LEVELS – FINDINGS & IMPLICATIONS

Findings:

- **50.96% of men** (total 10,480 men) and **56.36% of women** (total 731 women) had a Bachelor's Degree.
- **23.70% of men** and **26.13% of women** had a Master's Degree.

Implications:

- Employers may prefer software developers with higher (tertiary) education level as they can offer greater depth of knowledge, competency and understanding in the software field. Hence, higher education level, as highlighted in the bar chart, may be recommended to work in the software field.

DASHBOARD

GitHub link for access to dashboard:

https://github.com/OngWQ/IBM-Data-Analyst-Capstone/blob/main/IBM%20Data%20Analyst%20Capstone%20Project%20-%20Building%20A%20Dashboard%20With%20IBM%20Cognos%20Analytics%20(Updated).pdf

DASHBOARD TAB 1

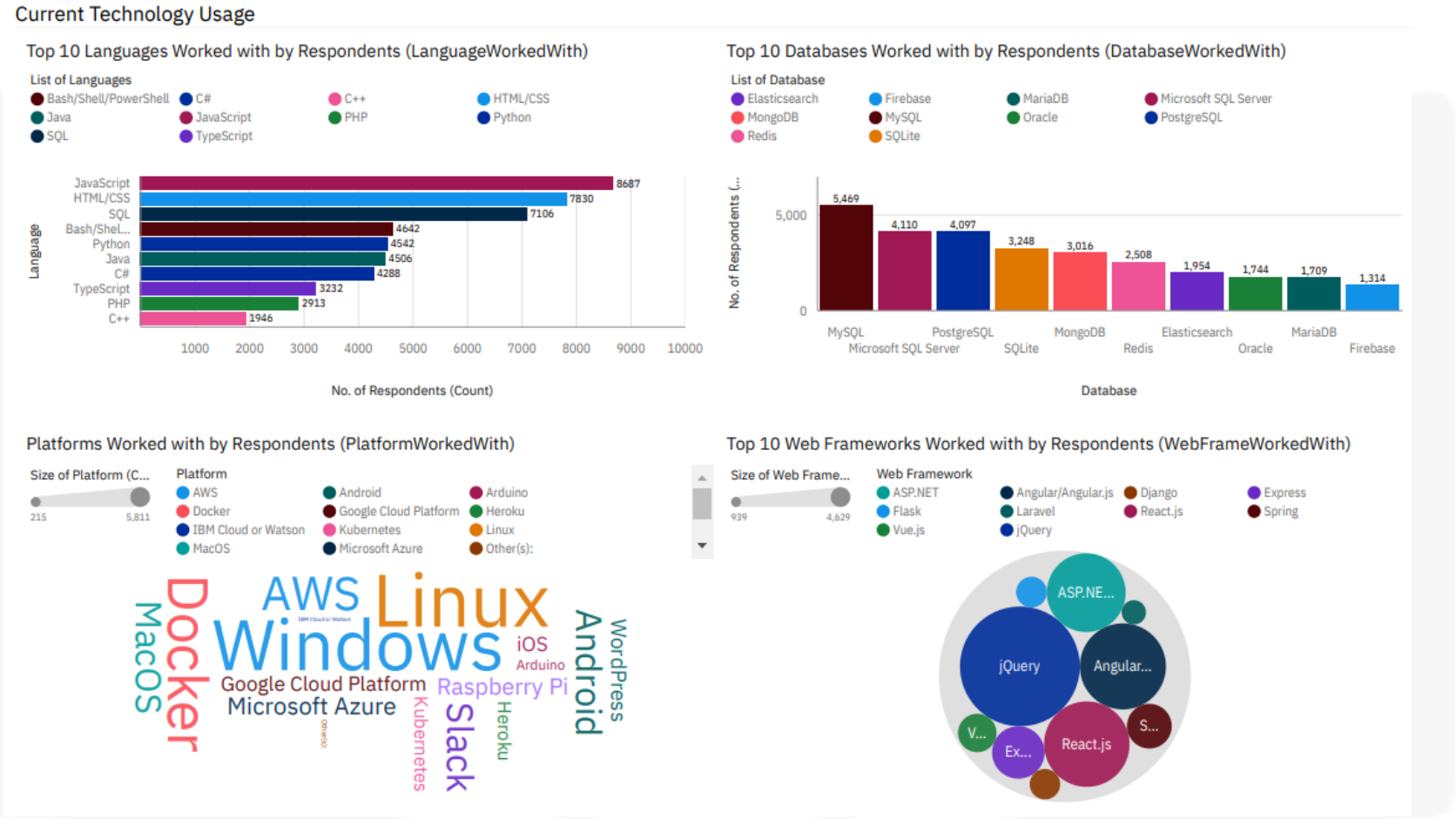


Figure 8: First tab of Dashboard displaying data visualisation pertaining to Current Technology Used.

DASHBOARD TAB 2

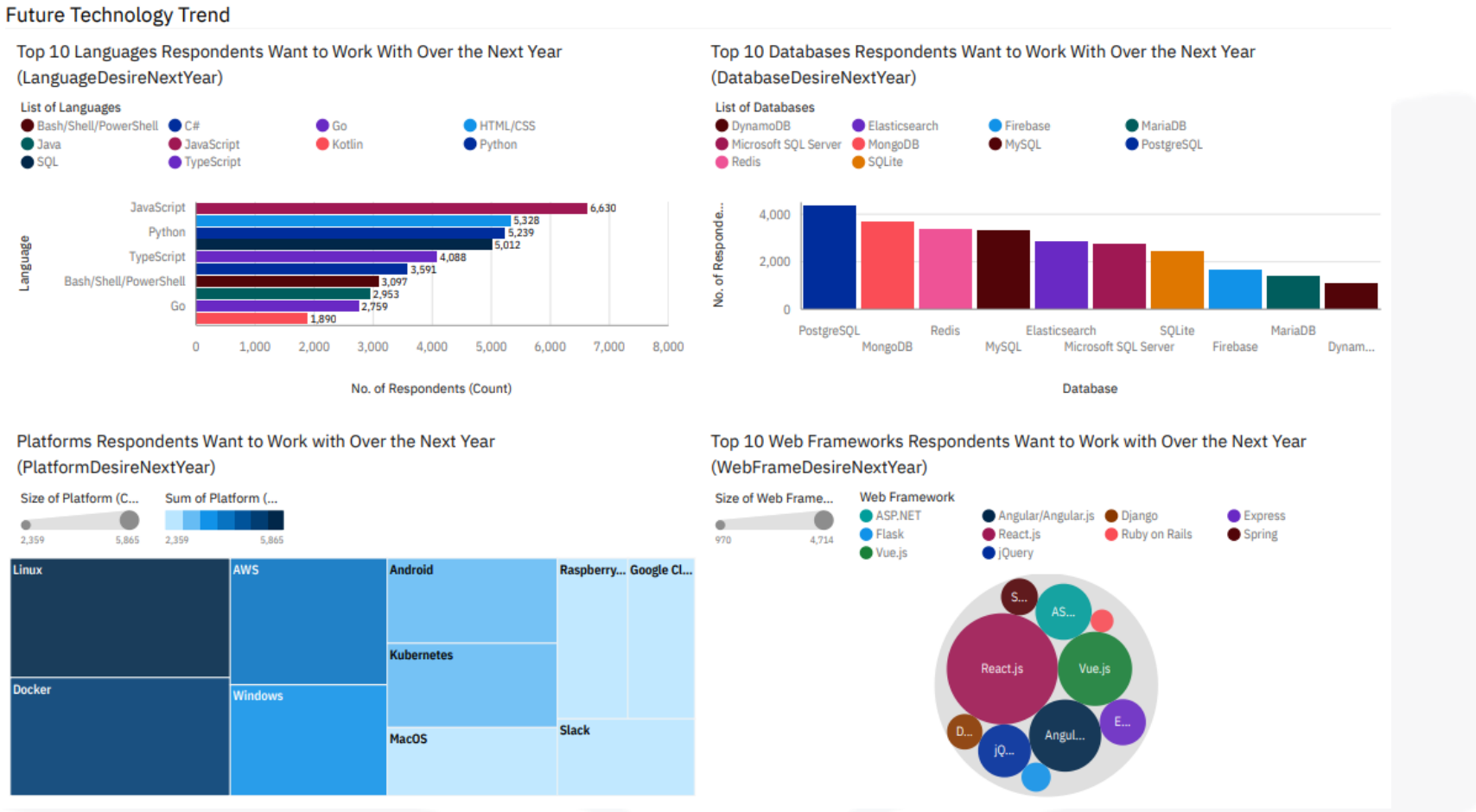


Figure 9: Second tab of Dashboard displaying data visualisation pertaining to Future Technology Used.

DASHBOARD TAB 3

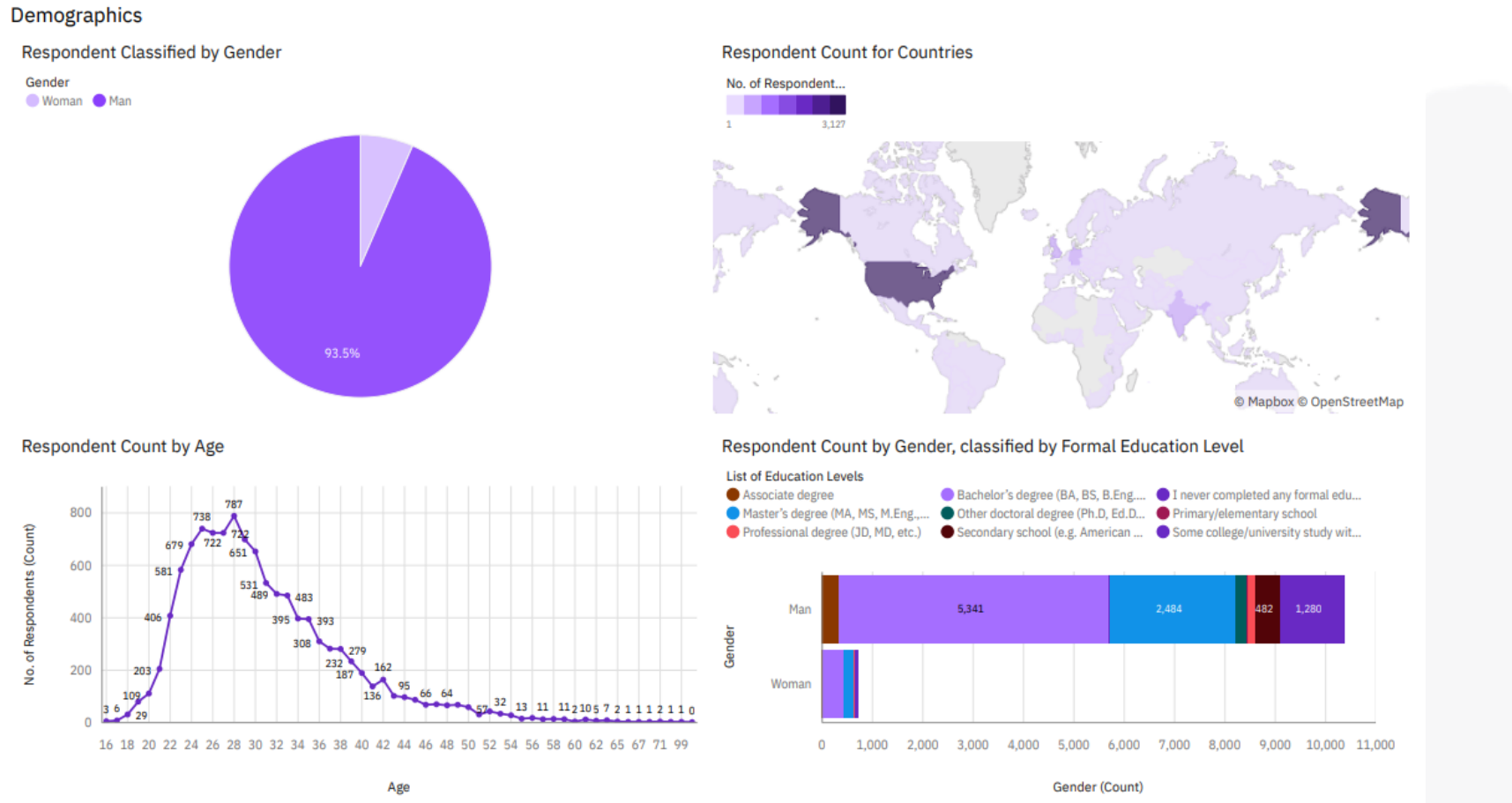


Figure 10: Third tab of Dashboard displaying data visualisation pertaining to demographics of respondents.

DISCUSSION

1. Increasing trend in usage of Python and PostgreSQL
2. Comparison of developers' preference of technologies to the demand of technologies in job postings
3. Further implications of education levels of software developers

DISCUSSION – Increasing trend in usage of Python and PostgreSQL

Findings:

- Remarkably, Python and PostgreSQL had a **higher ranking in technologies** (languages and databases, respectively) **wanted to learn by respondents over the next year** compared to their respective ranking in technologies (languages and databases, respectively) worked with by respondents.

Implications:

- Python: This may be attributed to its easier learning curve, quick execution of codes and extensive applications in many aspects of the software and data field (OnlineITGuru, 2023). Python is projected to become the fastest growing programming language (Robinson, 2017).
- PostgreSQL: This may be due to the flexibility of the database in which it can support many programming languages, as well as its robustness and security (Linux Polska, 2023; Riggs, 2021).

DISCUSSION – Comparison of developers' preference of technologies to the demand of technologies in job postings

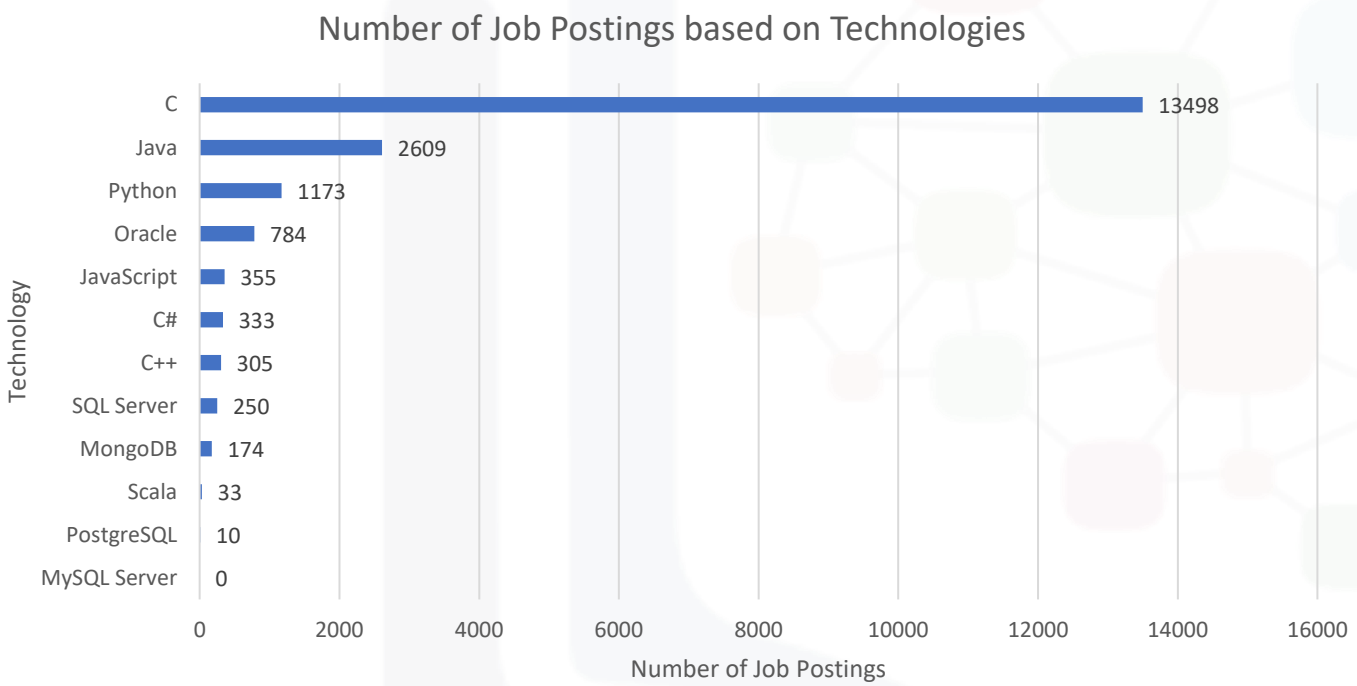


Figure 11: Number of Job Postings based on Technologies.

Findings:

- Although JavaScript, HTML/CSS, and Python are the top 3 most sought-after programming languages, C appeared to be the **most wanted programming language employers are looking for** in job postings by a wide margin.
- C was not among the top 10 languages worked with or to learn.

Implications:

- Software developers should also consider picking up C, primarily as it is an efficient, open-source programming language that can be used in developing internet browsers, developing operating systems (OS), and machine learning, despite being one of the oldest programming language used (Herrity, 2023; Motiso, 2024).

DISCUSSION – Further implications of education levels of software developers

Findings:

- There may be misconceptions in the software or computer field that many aspects such as programming languages and databases can be self-taught, and tertiary education is optional (Jalli, 2022).
- However, **over half of the respondents** based on gender had a Bachelor's Degree, and many respondents also obtained a Master's Degree.

Implications:

- This may be attributed to dynamic nature of the software or computer field, where there will be many changes to the demands of certain technologies. Example: Artificial Intelligence (AI) and Machine Learning (ML) (Kumar et al., 2023).
- Thus, further studying in universities can help in broadening the knowledge in the software field as many universities offer courses in those specialised area.
- Employers may also prefer candidates with a degree than those without due to proven competence with a degree.

Nevertheless, certain aspects in the software field like programming can still be self-taught by having a passion and desire to learn (Jalli, 2022).

CONCLUSION

- The most frequently used and sought-after language: JavaScript.
- Learning Python should be considered as it was the third-most sought-after language despite being the fifth-most used language.
- Contrary to the top languages, software developers should pick up and master C due to its high demand in job postings.

- The most frequently used database: MySQL.
- The most frequently sought-after database: PostgreSQL. Thus, learning PostgreSQL should be considered.

- Over half of the respondents based on gender have a Bachelor's Degree – higher education may still be required by employers.

REFERENCES

Herrity, K. (2023, February 4). *A Guide To C Programming: Definition, Uses and Benefits*. Indeed. https://www.indeed.com/career-advice/career-development/c-programming

Jalli, A. (2022, May 3). *10 Myths About Programming and Software Development*. Built In. https://builtin.com/software-engineering-perspectives/10-myths-about-programming

Kumar, K., Kumar, P., Deb, D., Unguresan, M.-L., & Muresan, V. (2023). Artificial Intelligence and Machine Learning Based Intervention in Medical Infrastructure: A Review and Future Trends. *Healthcare*, *11*(2), 207. https://doi.org/10.3390/healthcare11020207

Linux Polska. *PostgreSQL – the database most frequently chosen by developers and the future of DBMS*. (2023, March 14). Linux Polska. https://linuxpolska.com/en/knowledge-base/blog/postgresql-the-database-most-frequently-chosen-by-developers-and-the-future-of-dbms/

Motiso, D. (2024, February 13). *9 of the Most In-Demand Coding Languages (With Tips)*. Indeed. https://www.indeed.com/career-advice/career-development/most-in-demand-coding-languages

OnlineITGuru. *Why Python Is The Fastest-Growing Programming Language*. (2023, March 28). OnlineITGuru. https://onlineitguru.com/blogger/why-python-is-the-fastest-growing-programming-language

Riggs, S. (2021, January 11). *4 Reasons PostgreSQL was Named Database Management System of the Year 2020*. EnterpriseDB. https://www.enterprisedb.com/blog/4-reasons-postgresql-was-named-database-management-system-year-2020

Robinson, D. (2017, September 27). *The Incredible Growth of Python*. The Stack Overflow Blog. https://stackoverflow.blog/2017/09/06/incredible-growth-python/

APPENDIX

- Job Postings – Location
- Popular Languages
- Working Hours per week
- Annual salary distribution

Job Postings – Location

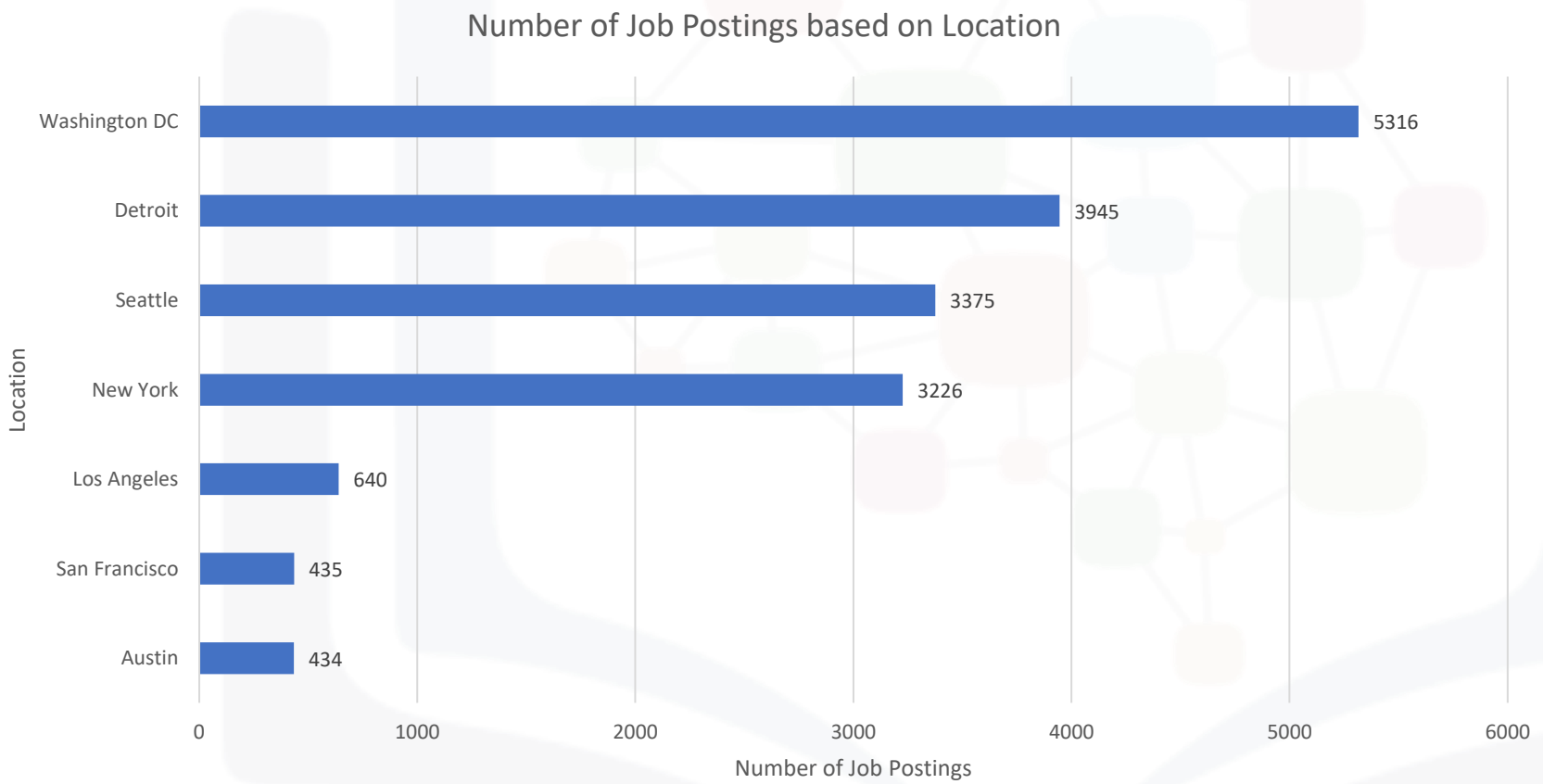


Figure 12: Number of Job Postings based on Locations.

Finding: Washington DC had the highest number of job postings, followed by Detroit, Seattle and then New York.

Implication: Job seekers in United States that are not in these locations may need to be ready to move to these locations to have a higher likelihood to land a job.

Popular Languages

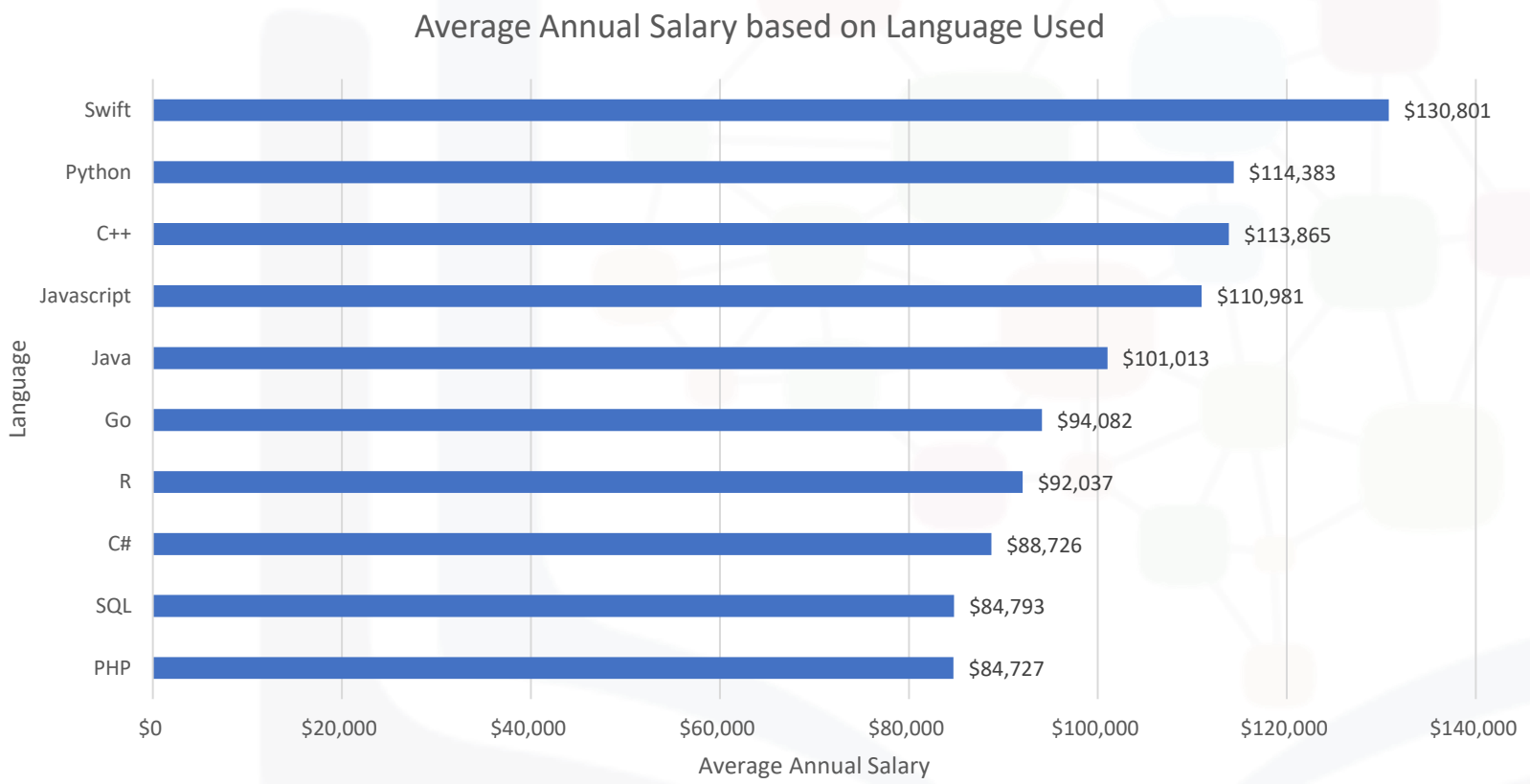


Figure 13: Average annual salary based on language used.

Finding: Swift had the highest average annual salary, followed by Python and then C++.

Implication: Employees and job seekers can expect an average annual salary around the listed values based on the language used.

Working Hours per week

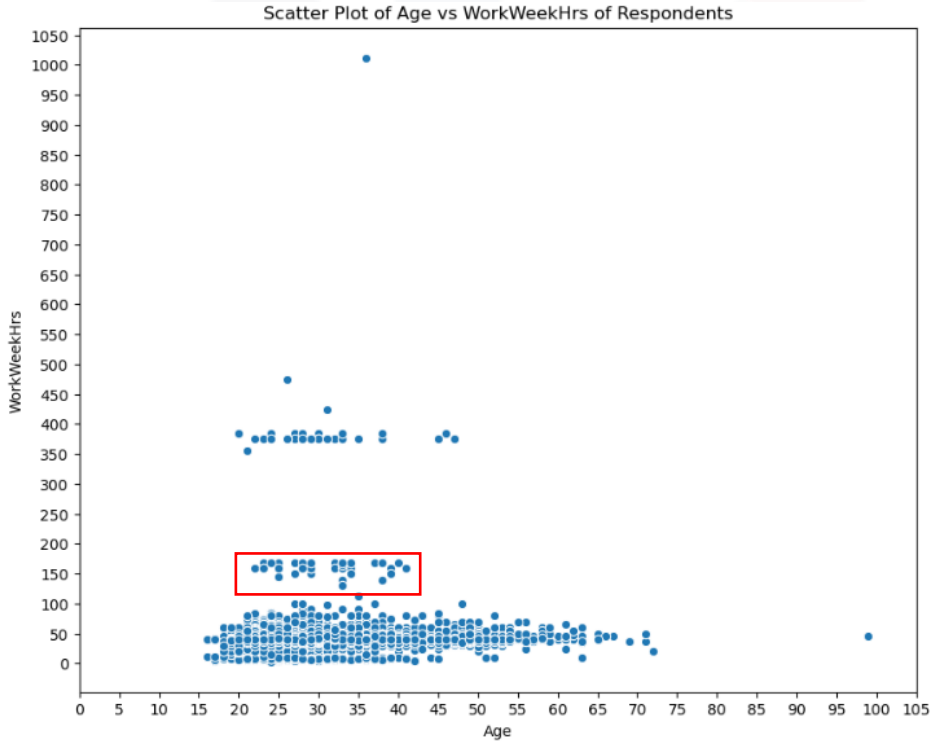


Figure 14: Scatter plot of age against working hours per week reported by respondents.

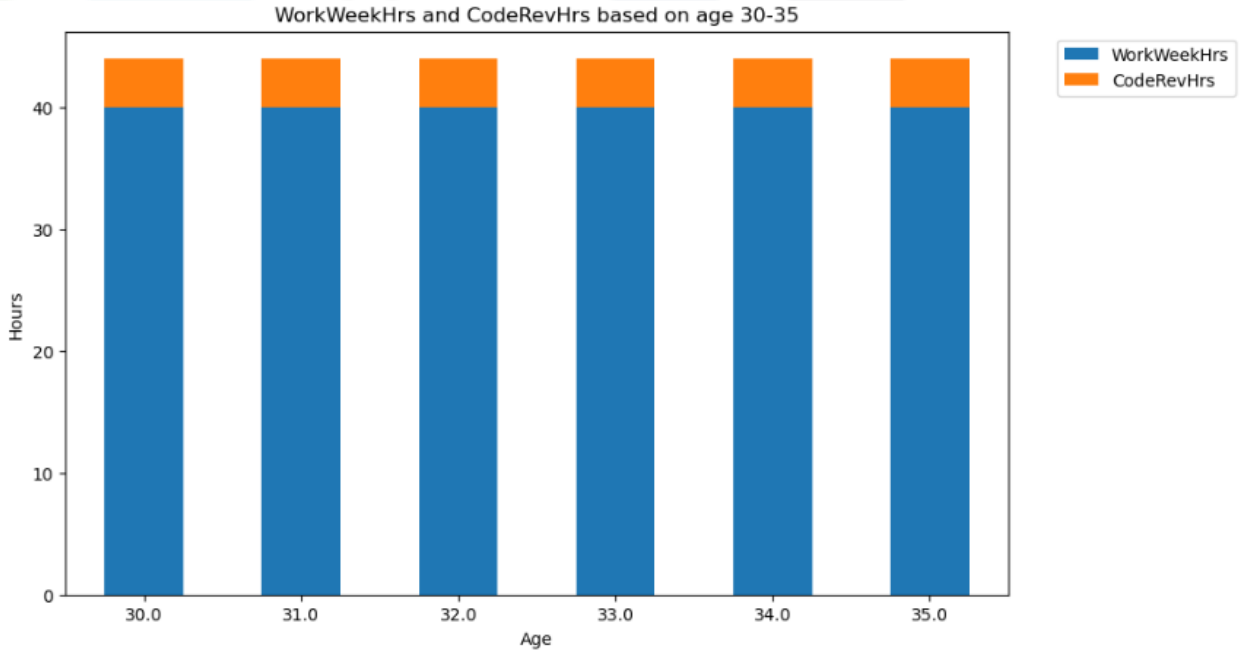
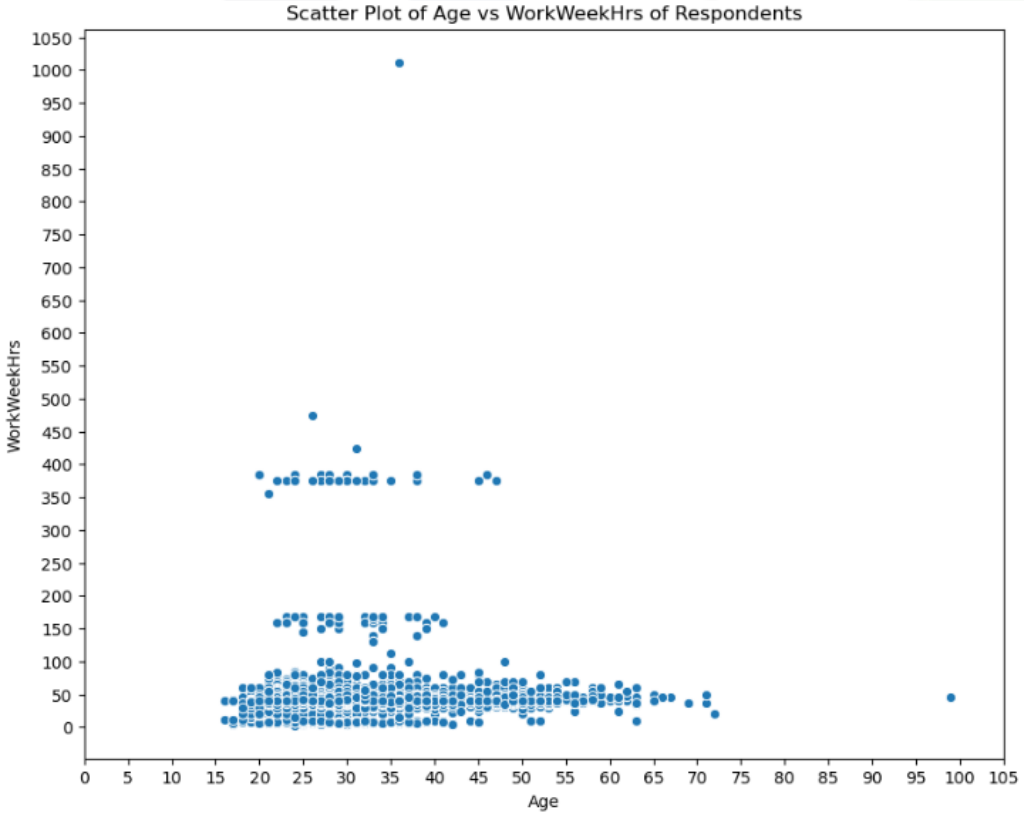


Figure 15: Working Hours per week and Code Review Hours per week in age 30-35.

Working Hours per week



Finding: Although median working hours per week for software developers aged 30 (which is the average age of respondents) was 40 hours, some data points the scatter plot show some respondents having unusually high working hours (more than 40 hours).

Implication: Although certain data points in the scatter plot are outliers, the unusually high working hours may be a good research topic to investigate more on the working hours of software professionals to advocate the well-being of software professionals.

Note: 1 week only has 168 hours.

Annual salary distribution

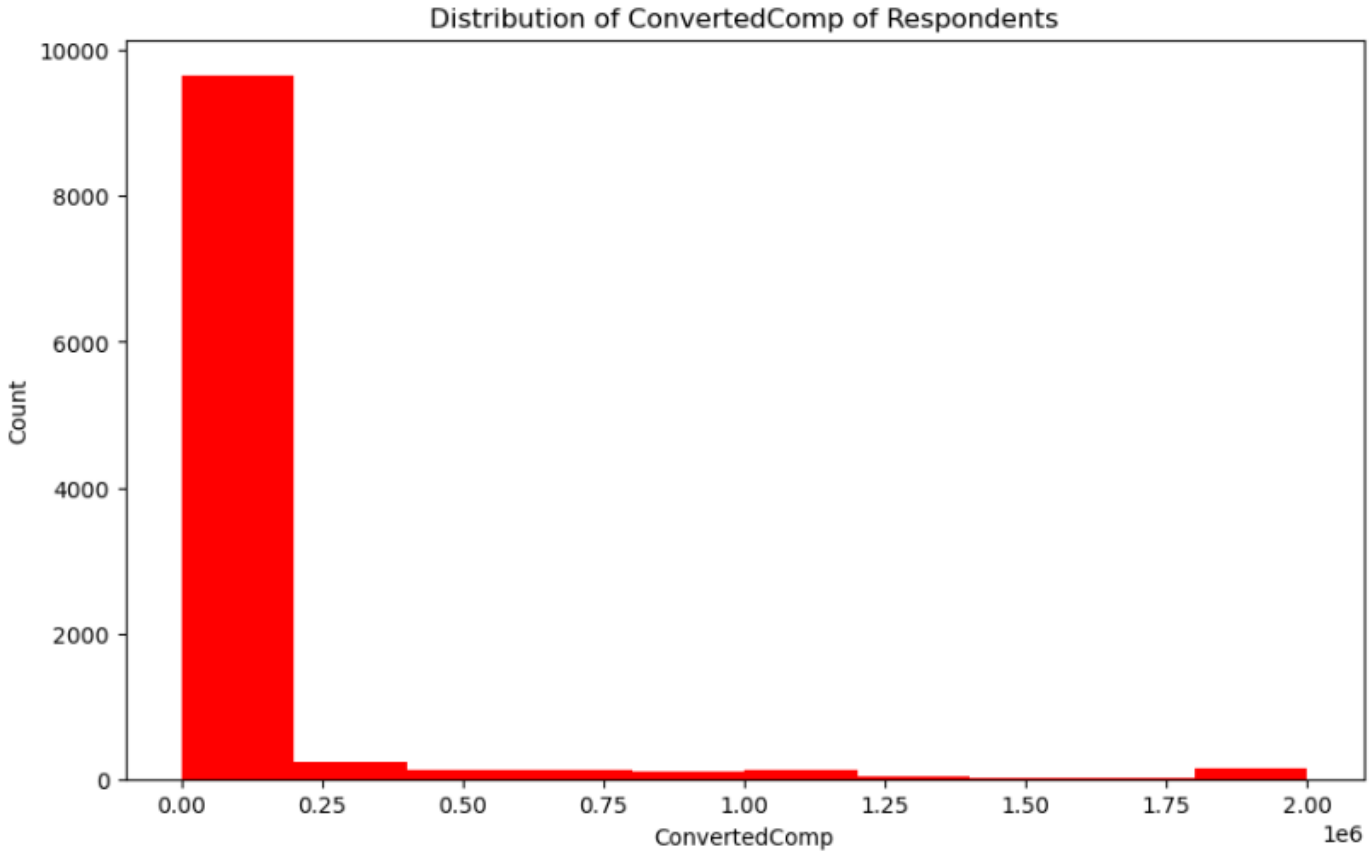


Figure 16: Distribution of annual salary (converted to USD) of respondents.

What is the median of the column `ConvertedComp` ?

```
[9]: # your code goes here
     df['ConvertedComp'].median()

[9]: 57745.0
```

Finding: The annual salary of respondents are commonly distributed at around the median (which is USD 57,745).

Implication: Software professionals should expect an annual salary around the median value.

IBM Developer

SKILLS NETWORK