IBM Data Analyst Capstone Project

Sachin Rathi

11-August-2024

https://github.com/sachinrathi1997/IBM_Data-Analyst_Capstone_Project

OUTLINE

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

EXECUTIVE SUMMARY

- **Project Overview:** Analyzed current and future trends in programming languages, databases, and job postings.
- **Data Sources:** Used survey data, job API, and web scraping for comprehensive analysis.
- **Data Sources:** Used survey data, job API, and web scraping for comprehensive analysis.
  - Survey data covered technology usage trends.
  - Job API provided real-time job postings data.
  - Web scraping captured salary trends for popular languages.
- **Key Insights:**
  - Identified top 10 programming languages and databases for the Current and next year.
  - Visualizations created for clear trend representation.
- **Outcome:** Developed an interactive dashboard to explore findings, offering valuable insights for aspiring data analysts.

INTRODUCTION

- Overview of the project’s purpose and significance.
- Relevance of analyzing technology trends in the data analytics field.
- Explanation of the data sources used for analysis.
- Objectives of the study.
  - Identify current trends in programming languages and databases.
  - Predict future demands and skill requirements in the industry.

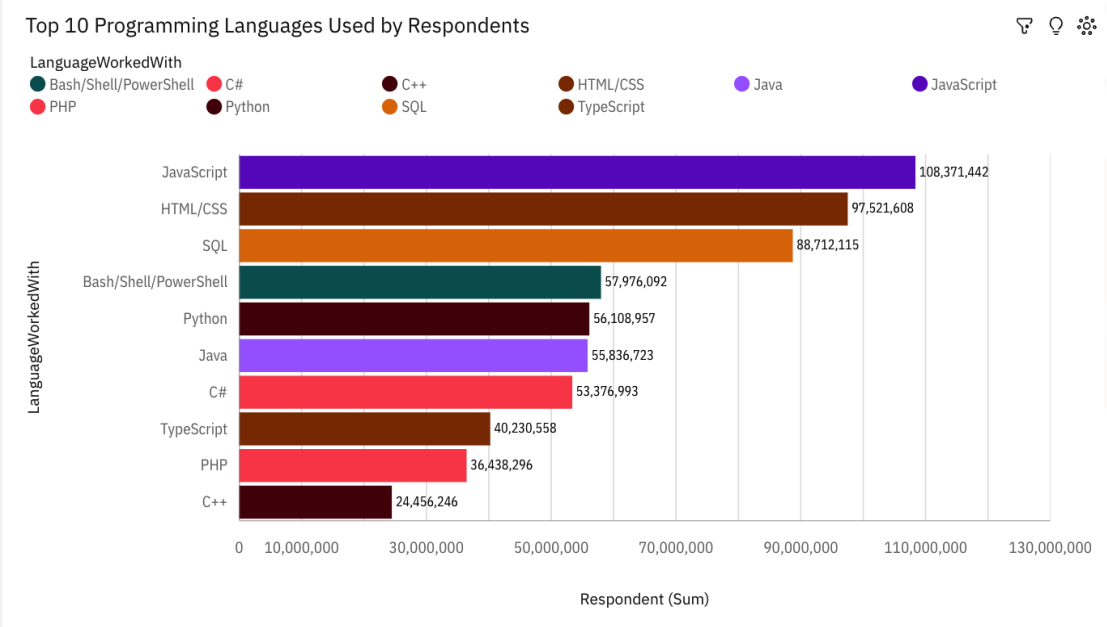
METHODOLOGY Part 1

- **Data Collection**
  - Sources: Job postings API, web scraping, Kaggle survey datasets.
  - Tools: APIs, web scraping libraries (BeautifulSoup, Scrapy).
- **Data Cleaning and Preparation**
  - Cleaning: Handle missing values, remove duplicates, standardize formats.
  - Transformation: Normalize data, encode categorical variables, merge datasets.
- **Data Analysis**
  - EDA: Summary statistics, data distributions, patterns, and anomalies.
  - Techniques: Statistical tests, data visualizations (e.g., bar charts, line graphs).
- **Dashboard Creation**
  - Design: Layout and visualization principles.
  - Tools: IBM Cognos Analytics, Google Looker Studio.
  - Visualizations: Bar charts, pie charts, line graphs, maps.
- **Validation and Testing**
  - Verification: Ensure accuracy and reliability.
  - Validation: Cross-check with other data sources.
- **Challenges and Solutions**
  - Challenges: Data quality issues, technical limitations.
  - Solutions: Workarounds and problem-solving strategies.
- **Conclusion**
  - Summary: Effectiveness of methodology in achieving project goals.

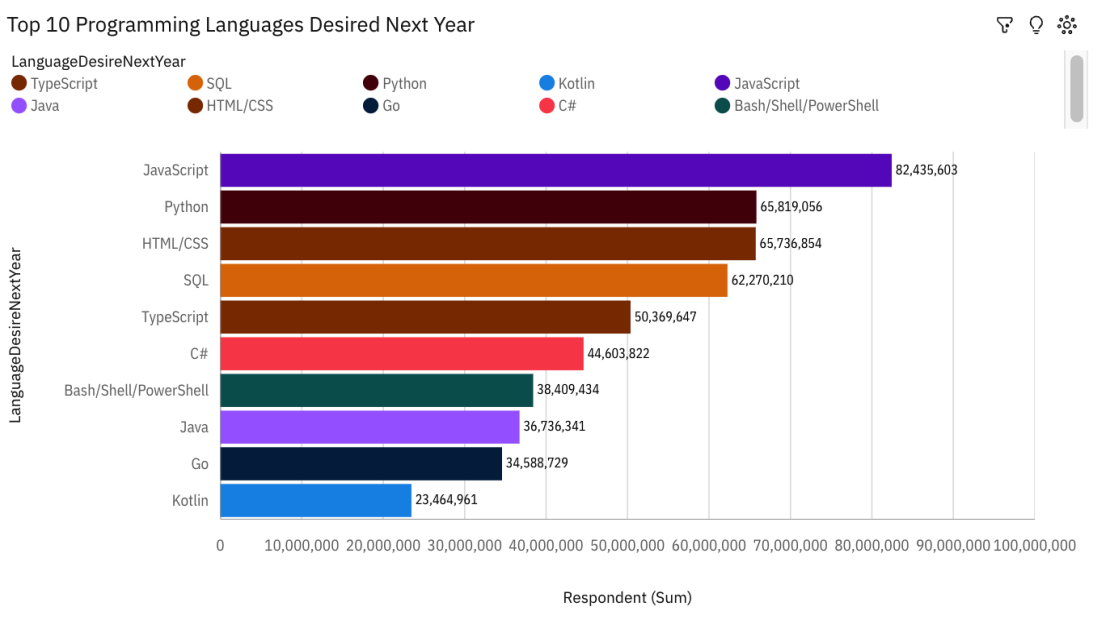
RESULTS

PROGRAMMING LANGUAGE TRENDS

Current Year



Next Year



PROGRAMMING LANGUAGE TRENDS - FINDINGS & IMPLICATIONS

Findings

- LanguageWorkedWith JavaScript has the highest total respondent count, reflecting its strong current usage.
- The average number of respondents is approximately 61.9 million (61,902,903).
- Respondent count is unusually high when LanguageWorkedWith is JavaScript.
- The summed values of respondent counts range from over 24 million to over 108 million.
- For LanguageWorkedWith, the most significant values are JavaScript, HTML/CSS, and SQL, adding up to nearly 295 million (47.6% of the total).
- Across all values of LanguageWorkedWith, the sum of respondents is over 619 million.

Implications

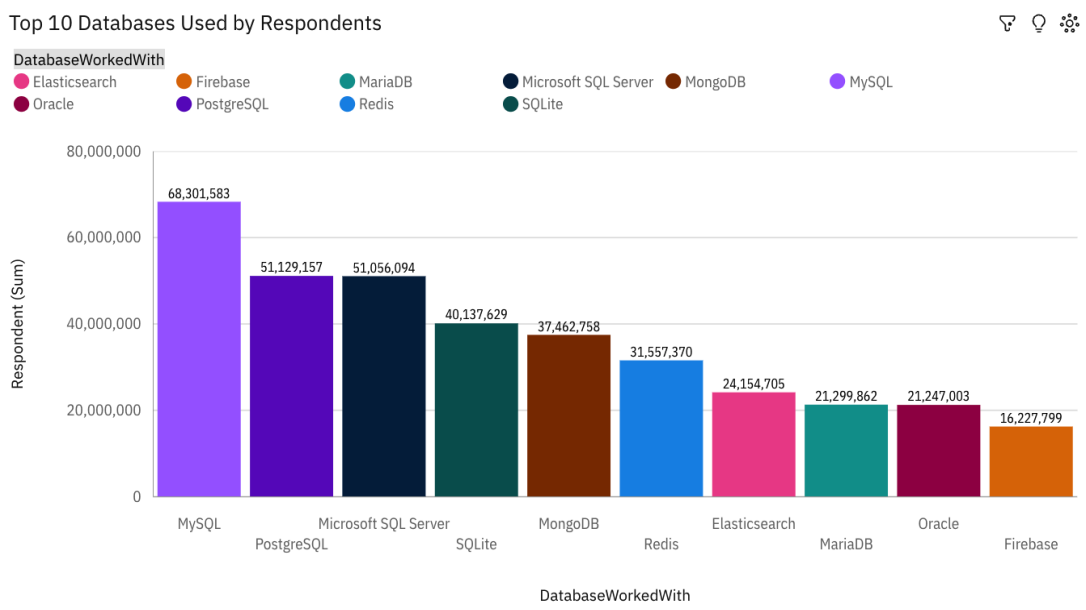
- LanguageDesireNextYear JavaScript has the highest total respondent count, indicating continued strong interest in the future.
- The average number of respondents for future interest languages is about 50.4 million (50,443,466).
- Respondent count is most unusual when LanguageDesireNextYear is JavaScript and Kotlin.
- The summed values of respondent counts range from over 23 million to over 82 million.
- For LanguageDesireNextYear, JavaScript is the most significant value, with over 82 million respondents (16.3% of the total).
- Across all values of LanguageDesireNextYear, the sum of respondents is over 504 million.

PROGRAMMING LANGUAGE TRENDS - COMPARISON

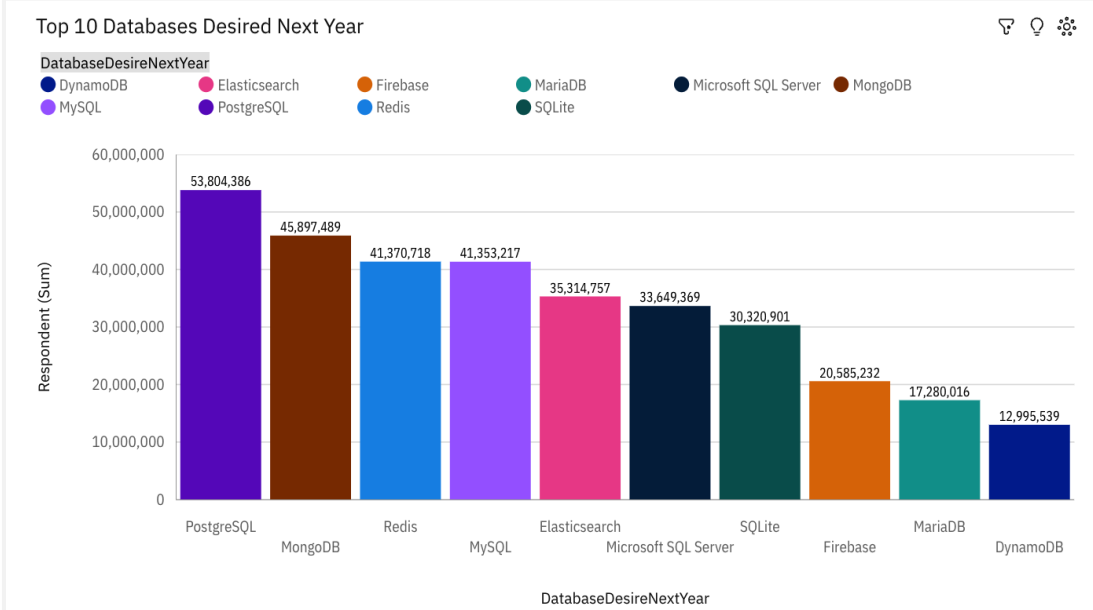
- **Top Respondent Language**: JavaScript leads in both current usage and future interest, underscoring its importance in both contexts.
- **Average Respondent Count**: The average respondent count for current languages is higher than for future interest, suggesting a higher engagement with current technologies compared to future interests.
- **Unusual Respondent Count**: High respondent counts for JavaScript in current usage and for JavaScript and Kotlin in future interest highlight areas of significant focus.
- **Range of Respondent Values**: The range of values is slightly higher for current languages, indicating broader engagement compared to future languages.
- **Top Values by Language**: JavaScript remains the most significant language in both current usage and future interest, though its relative impact is greater in current usage.
- **Total Respondents**: There is a higher total respondent count for current languages compared to future languages, reflecting broader engagement with current technologies.

DATABASE TRENDS

Current Year



Next Year



DATABASE TRENDS - FINDINGS & IMPLICATIONS

Findings

- **Respondent is unusually high when DatabaseWorkedWith is MySQL.**
- The average number of respondents is approximately 36.26 million(36,257,396)
- **LanguageWorkedWith HTML/CSS has the highest Respondent at almost 41 million, out of which DatabaseWorkedWith MySQL contributed the most at nearly 17 million.**
- **DatabaseWorkedWith MySQL has the highest total Respondent due to LanguageWorkedWith HTML/CSS.**
- **Microsoft SQL Server has a Respondent of nearly eighteen million for LanguageWorkedWith C#.**
- **The summed values of Respondent range from over sixteen million to over 68 million.**
- **For Respondent, the most significant value of DatabaseWorkedWith is MySQL, whose respective Respondent values add up to over 68 million, or 18.8 % of the total.**
- **Over all values of DatabaseWorkedWith and DatabaseWorkedWith, the sum of Respondent is nearly 363 million.**
- **HTML/CSS LanguageWorkedWith accounted for 47% of MySQL Respondent compared to 22% for Microsoft SQL Server.**

Implications

- **Respondent is most unusual when DatabaseDesireNextYear is PostgreSQL, DynamoDB and MariaDB.**
- The average number of respondents for future interest languages is about 33,26 million(33,257,162)
- LanguageWorkedWith HTML/CSS has the highest Respondent at over 34 million, out of which DatabaseDesireNextYear MongoDB contributed the most at almost 9.9 million.
- DatabaseDesireNextYear Elasticsearch has the highest total Respondent due to LanguageWorkedWith Bash/Shell/PowerShell.
- Microsoft SQL Server has a Respondent of over eleven million for LanguageWorkedWith C#.
- The summed values of Respondent range from nearly thirteen million to almost 54 million.
- For Respondent, the most significant value of DatabaseDesireNextYear is PostgreSQL, whose respective Respondent values add up to almost 54 million, or 16.2 % of the total.
- Over all values of DatabaseDesireNextYear and DatabaseDesireNextYear, the sum of Respondent is almost 333 million.

DATABASE TRENDS - COMPARISON

- **Current Usage**: MySQL is the most used database, especially among HTML/CSS developers. It accounts for a significant portion of the respondent base.
- **Future Interest**: PostgreSQL, DynamoDB, and MariaDB are notably desired for future use. MongoDB also shows high future interest, especially among HTML/CSS developers.
- **Respondent Counts**:
  - Average respondents: ~36.26 million overall; ~33.26 million for future interest.
  - MySQL has a higher total respondent count compared to PostgreSQL, but PostgreSQL shows strong future interest.
- **Proportional Representation**: HTML/CSS users account for 47% of MySQL respondents, while Microsoft SQL Server has a smaller share (22%).

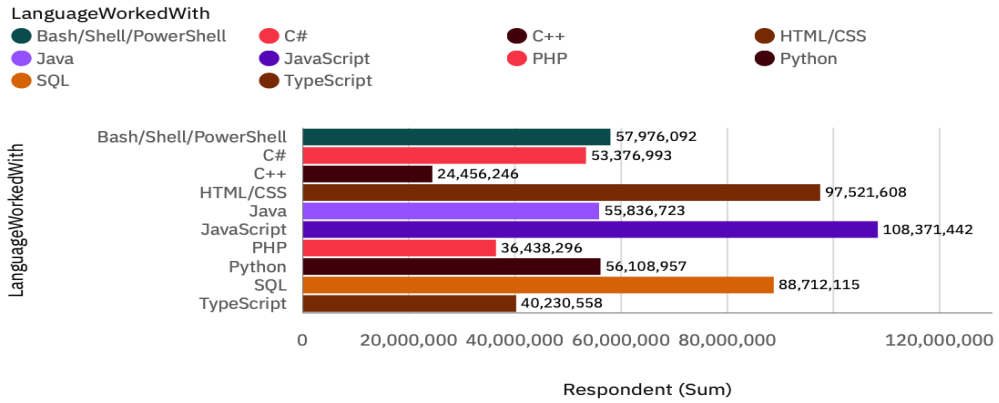
DASHBOARD

https://github.com/sachinrathi1997/IBM_Data-Analyst_Capstone_Project/blob/main/IBM_Data_Analyst_Capstone_Project/Module%205/Graded%20Assignment_%20Dashboards.pdf

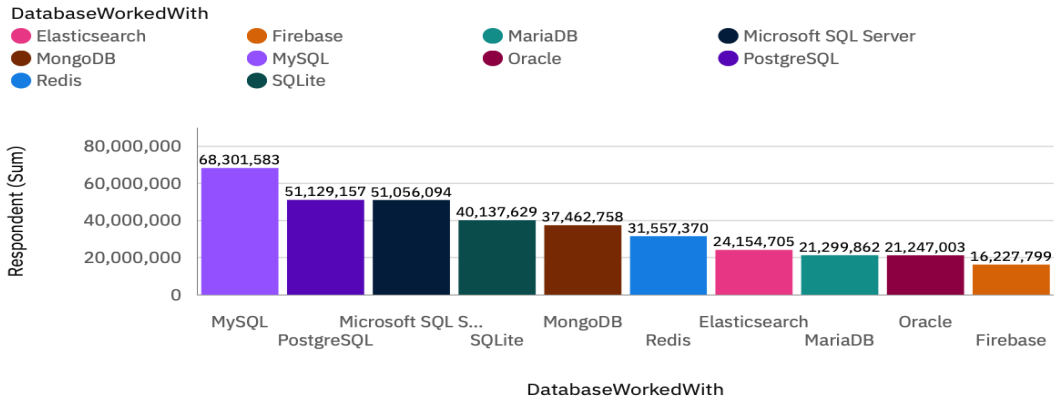
DASHBOARD TAB 1

Current Technology Usage

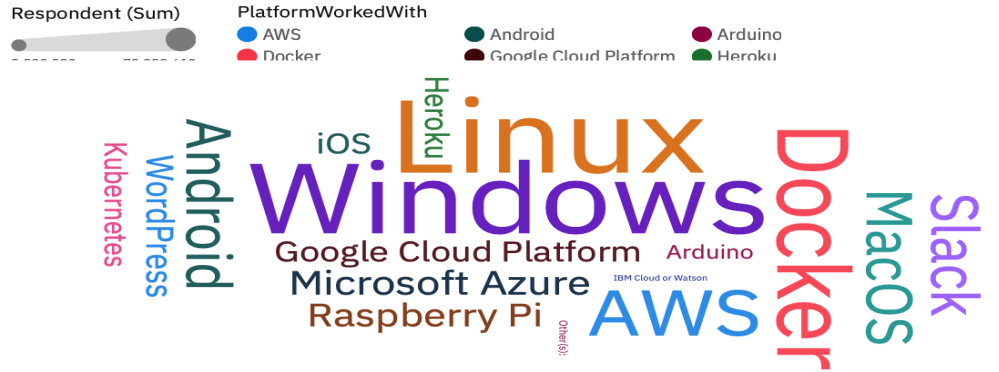
Top 10 Programming Languages Used by Respondents



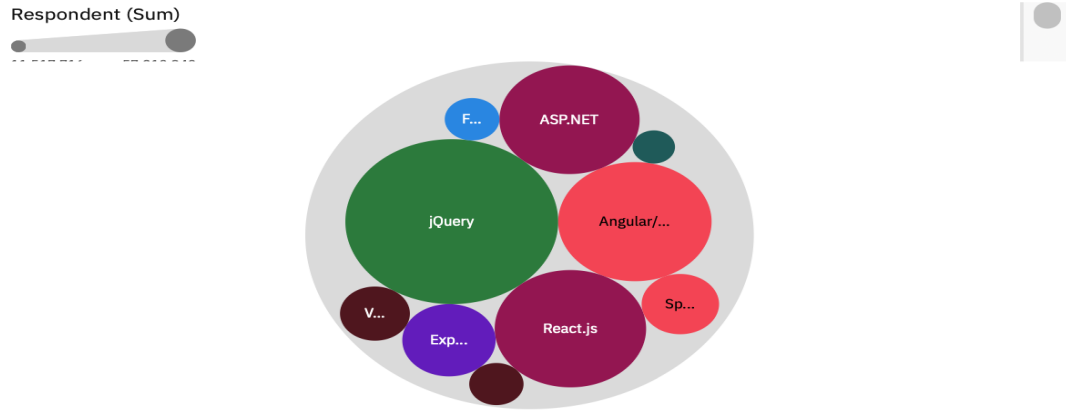
Top 10 Databases Used by Respondents



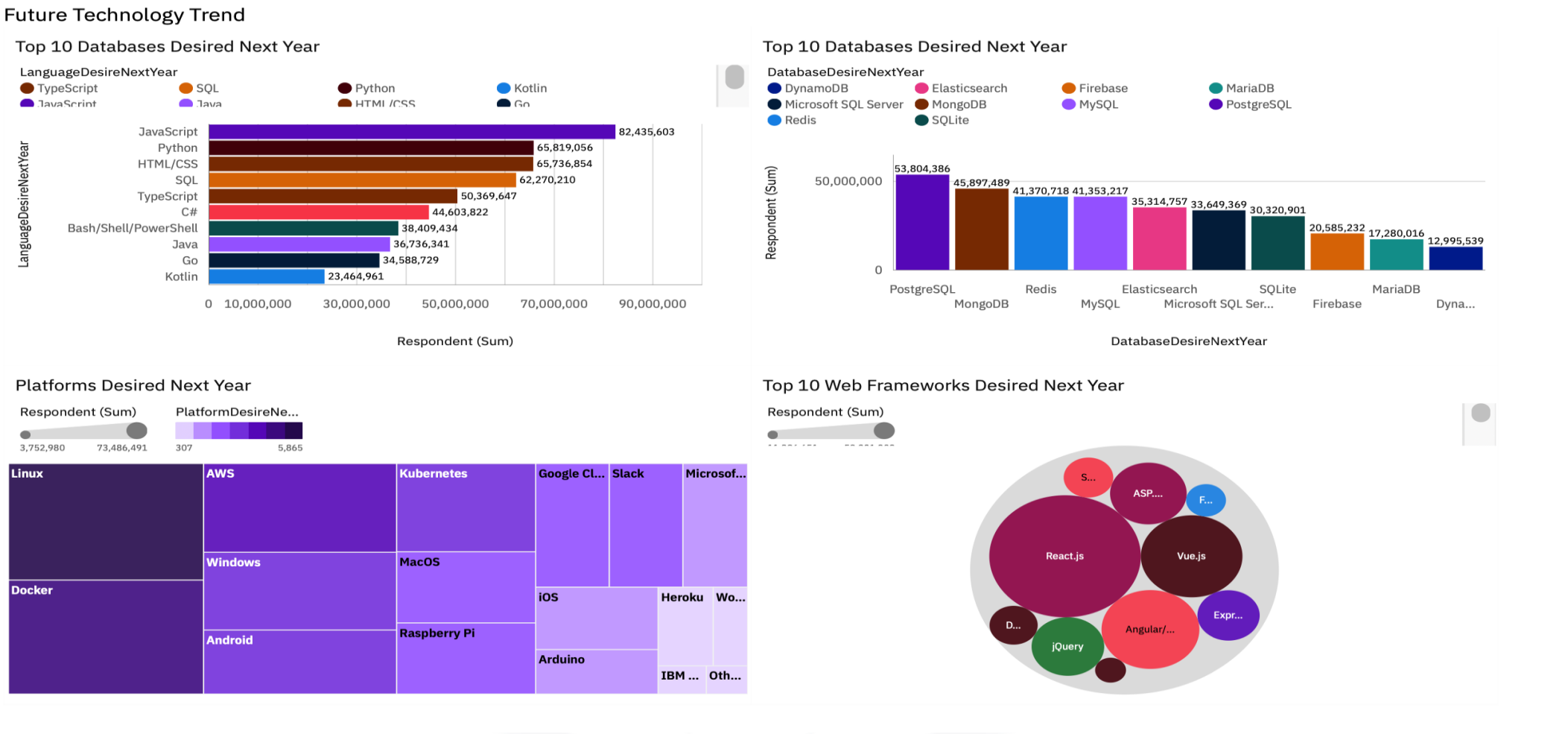
Platforms Used by Respondents



Top 10 Web Frameworks Used by Respondents



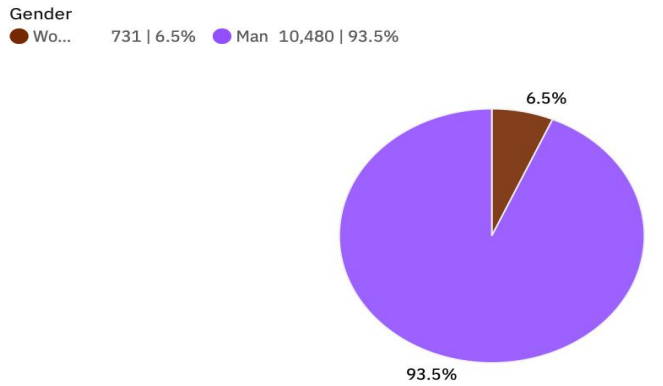
DASHBOARD TAB 2



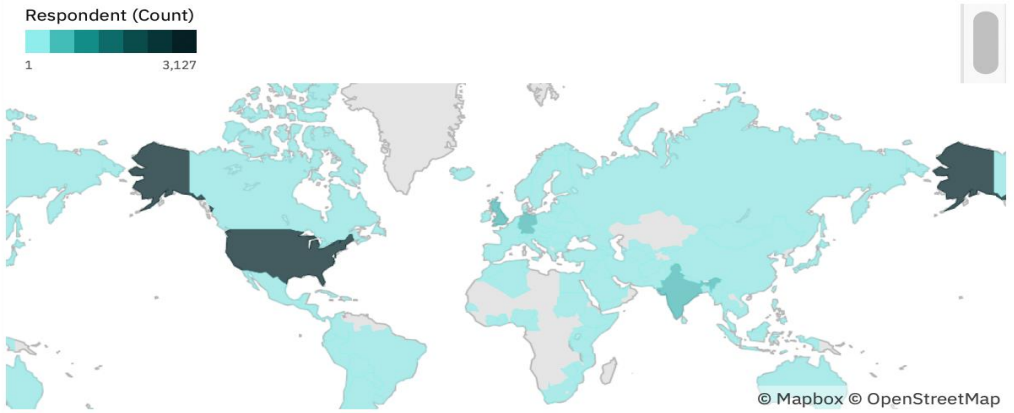
DASHBOARD TAB 3

Demographics

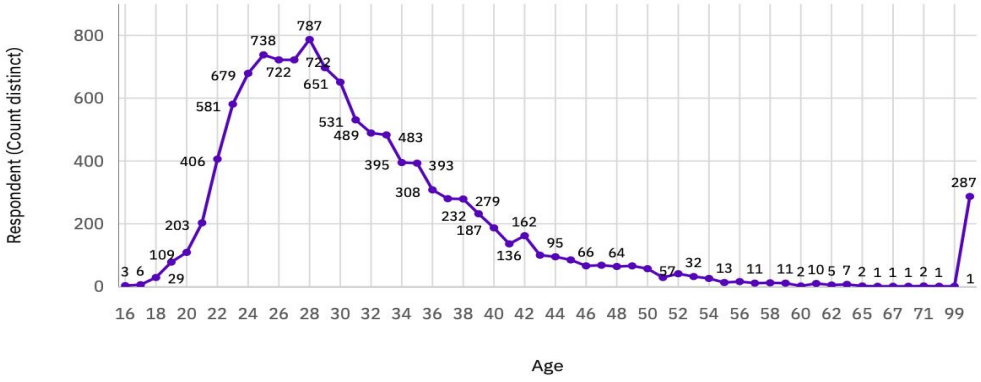
Respondent by Gender



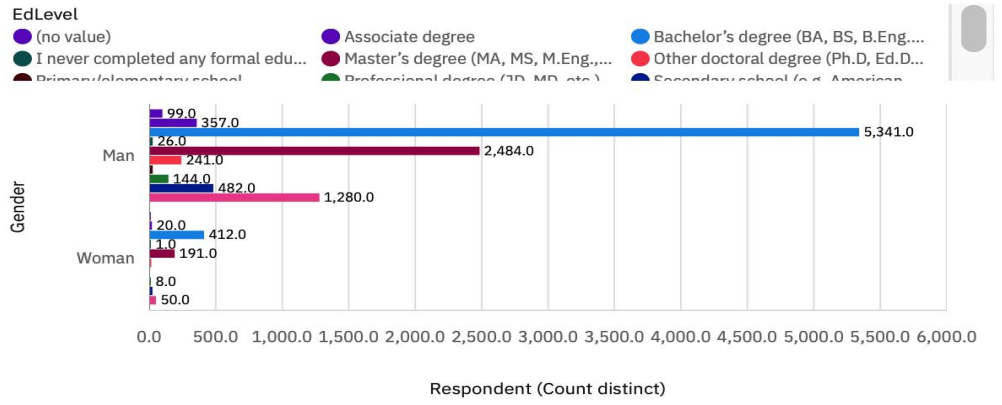
Respondent Count by Country



Respondent Count by Age



Respondent Count by Gender and Education Level



DISCUSSION

https://www.coursera.org/learn/ibm-data-analyst-capstone-project/discussions

OVERALL FINDINGS & IMPLICATIONS

Findings

1. **Geographical Focus**: **United States** dominates in respondent count but has a lower average age compared to other countries.
2. **Gender Disparity**: **Men** overwhelmingly outnumber women in the dataset, highlighting a gender imbalance.
3. **Educational Background**: **Bachelor's degrees** are the most common among respondents, especially in the U.S.
4. **Age Trends**: **28 years** old is the most common age group among respondents, particularly in the U.S.
5. **Technology Preferences**: **JavaScript** is the most desired technology, followed by **SQL** and **HTML/CSS**.
6. **Database & Platform**: **MySQL** and **PostgreSQL** are the most sought-after databases. **Linux** and **Docker** are the most preferred platforms.
7. **Frameworks**: **React.js** and **jQuery** are the most desired web frameworks.

Implications

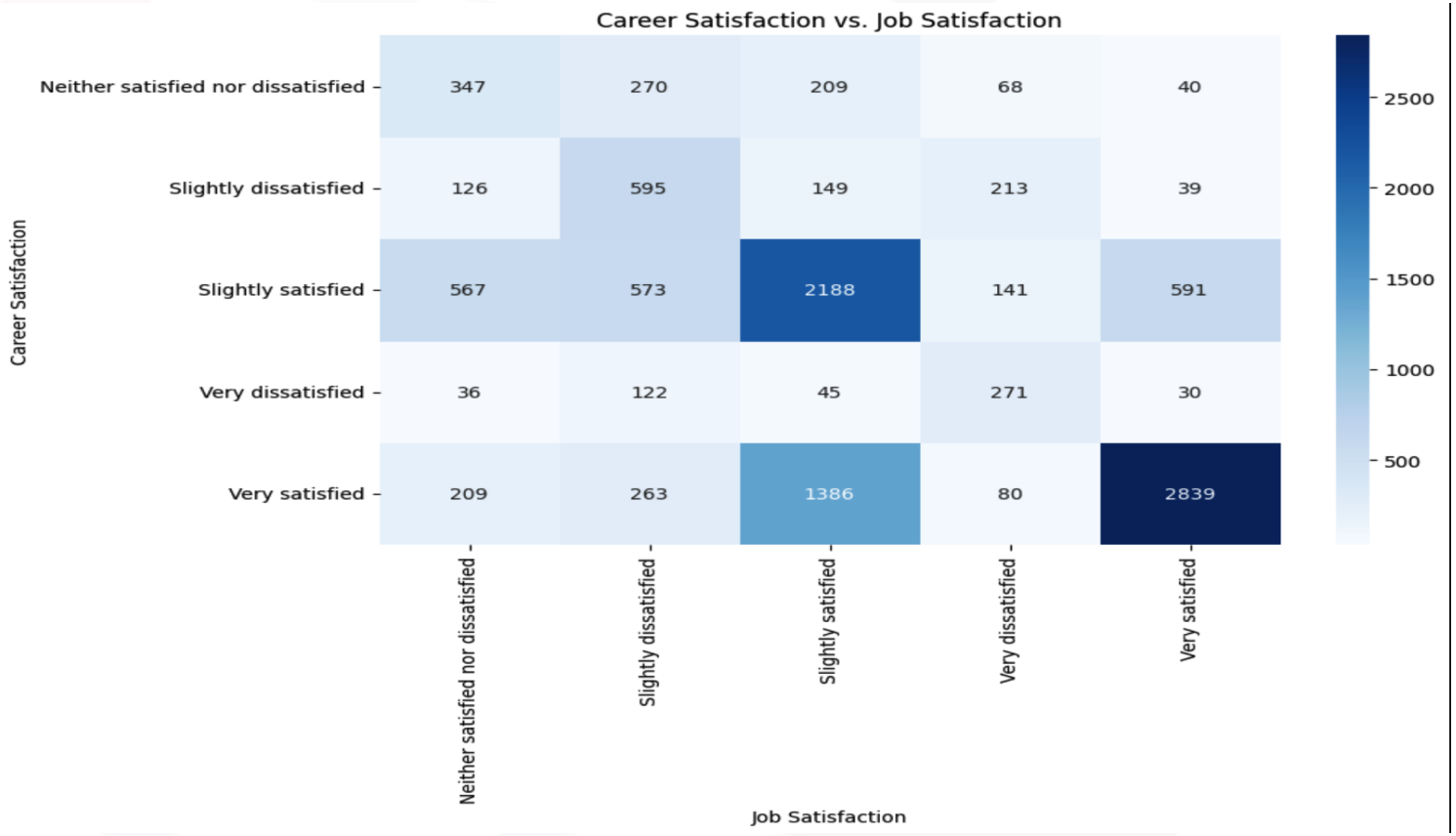
Focus on U.S. trends but address global diversity.

- Improve gender balance in tech surveys.
- Align tech investments with popular technologies and platforms.
- Consider age and educational background when analyzing tech trends.

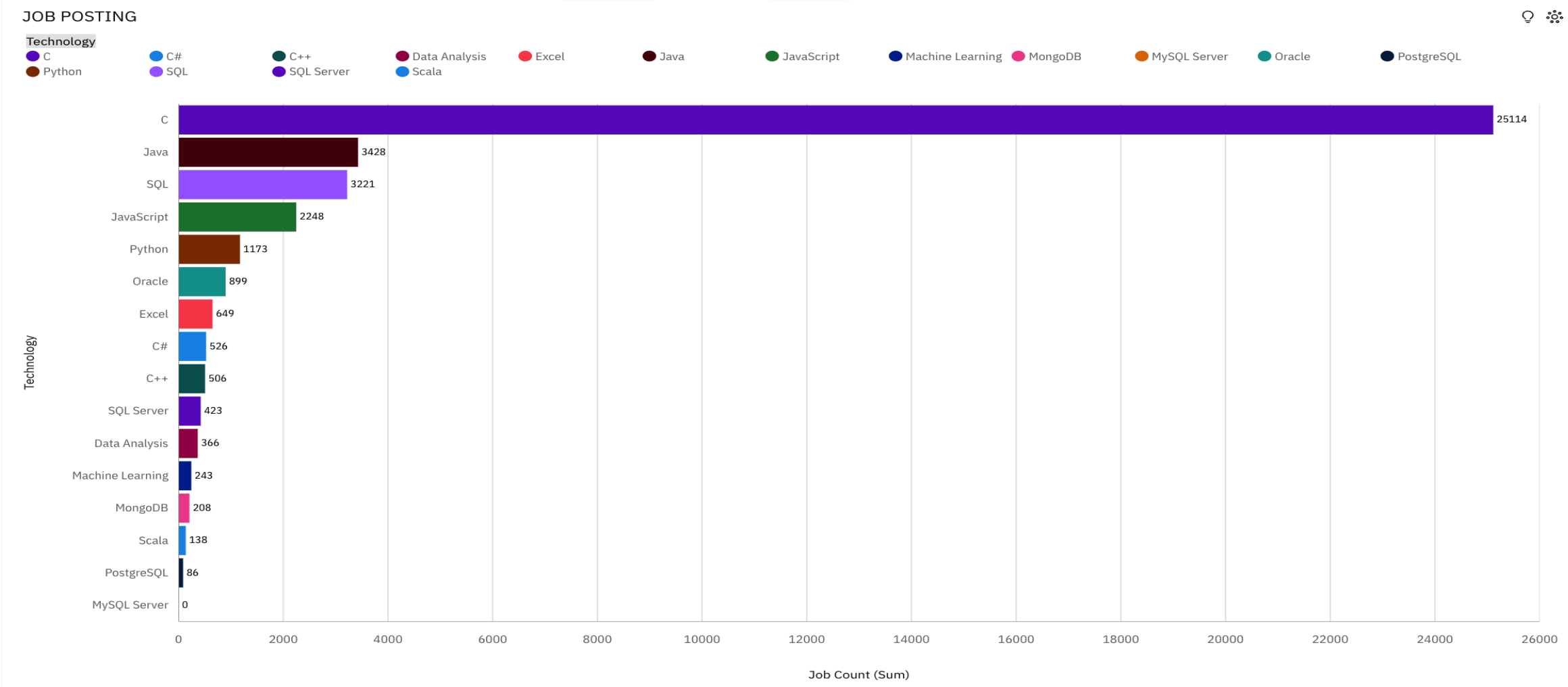
CONCLUSION

- The survey data reveals significant trends and patterns that can guide future technology investments and research. The dominant presence of the United States in respondent count highlights the need to focus on this key market while also addressing global diversity to ensure a comprehensive understanding of tech trends worldwide. The pronounced gender imbalance suggests a need for initiatives to encourage greater female participation in tech-related surveys.
- Educationally, the prevalence of respondents with Bachelor's degrees indicates that this group is most engaged with technology, which can inform targeted outreach and development strategies. The age trend, with 28-year-olds being the most represented, points to a core demographic that is actively engaged in technology and likely to influence future trends.
- In terms of technology preferences, JavaScript, SQL, and HTML/CSS stand out as key areas of interest, indicating where future tech developments and investments should be concentrated. The strong interest in MySQL, PostgreSQL, Linux, and Docker underscores the importance of these technologies in the current landscape.
- Lastly, the preference for frameworks like React.js and jQuery suggests these will continue to be crucial in web development. Overall, aligning strategies with these insights will help organizations stay ahead in the rapidly evolving tech industry and address both current and emerging needs effectively.

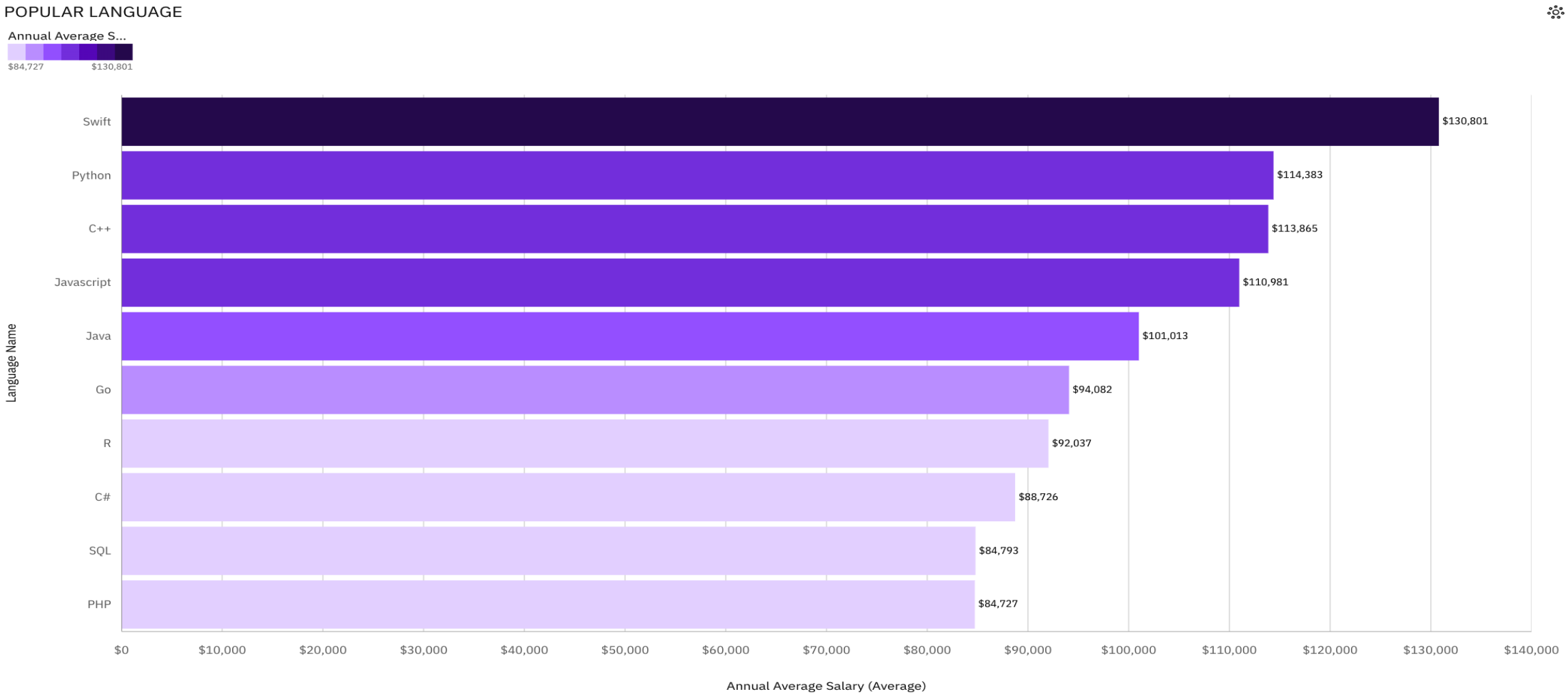
APPENDIX


Career Satisfaction vs. Job Satisfaction
Neither satisfied nor dissatisfied
Slightly dissatisfied
Slightly satisfied
Very dissatisfied
Very satisfied
347 270 209 68 40
126 595 149 213 39
567 573 2188 141 591
36 122 45 271 30
209 263 1386 80 2839
Career Satisfaction
Job Satisfaction
Neither satisfied nor dissatisfied
Slightly dissatisfied
Slightly satisfied
Very dissatisfied
Very satisfied
2500
2000
1500
1000
500

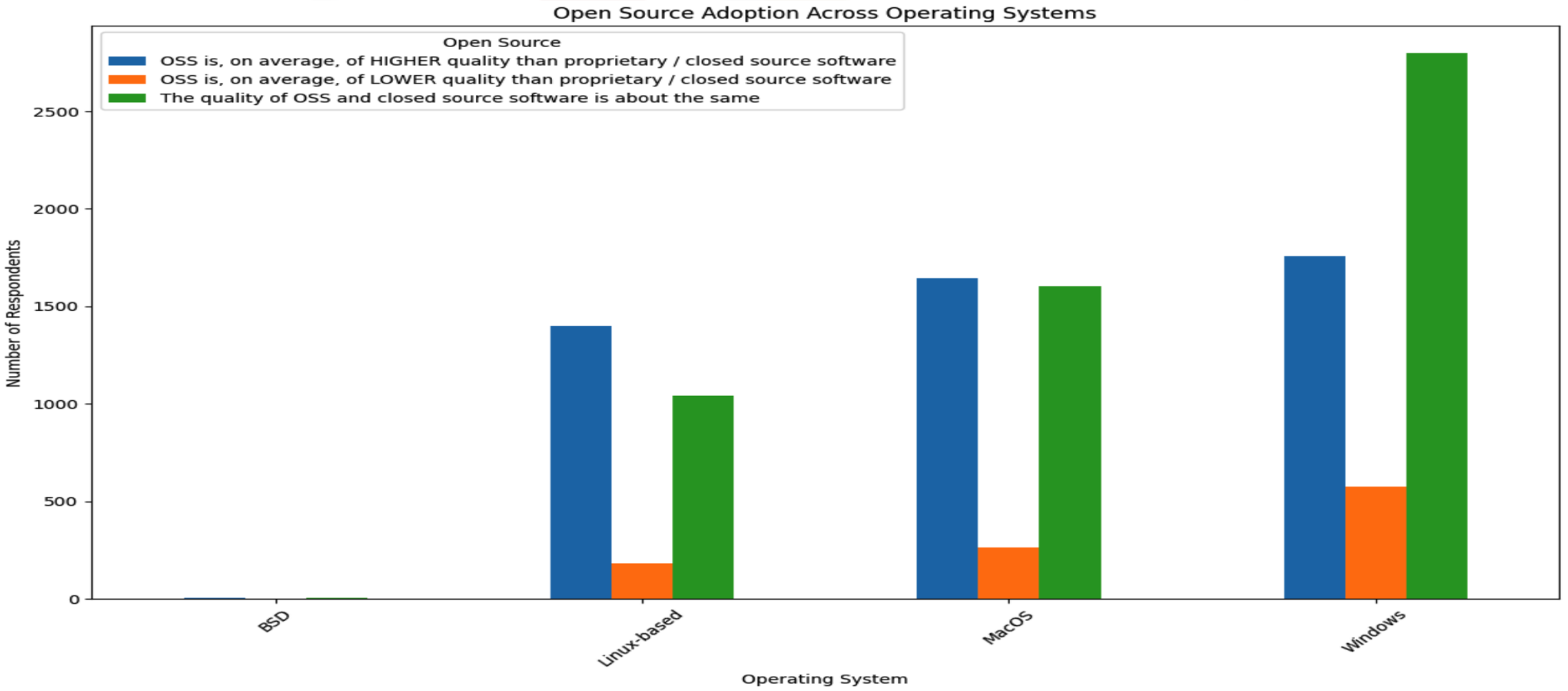
JOB POSTINGS



POPULAR LANGUAGES



OPEN SOURCE ADOPTION ACROSS OS



Job Satistefction and Career Aspirations by Employment Status