Data Analysis to identify trends for emerging technologies

Krithika Navaneethan

05/24/2023

OUTLINE

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

EXECUTIVE SUMMARY

- Data analysis of a subset of data from the source **2019 Stack Overflow Developer Survey** to identify the trends for emerging technologies.
- The process involves following key steps:
  - Data collection - from Jobs API and web scraping
  - Data wrangling - clean the data for further analysis
  - Exploratory Data Analysis – analyze the cleaned dataset to identify data distribution and correlation
  - Data Visualization – visualize results of data analysis using charts, plots and graphs
  - Dashboard creation – visualize the key points of the survey dataset
- The key results of the analysis include the following:
  - Current Technology Usage - top languages, databases, platforms and web frames at the time of data collection
  - Future Technology Trend - most preferable technologies in the future
  - Demographics - distribution of respondents by gender, countries, age and education

INTRODUCTION

Objective:

To perform data analysis of the Developer Survey dataset, identify trends in emerging software skills and technologies, and finally present the results of the analysis.

Purpose:

- To identify the top programming languages, database skills in demand
- To determine the popular IDE's
- To project the software technologies/skills requirements for future
- To get an insight of statistical data relating to the human resources in the IT industry

Audience:

- Global IT and business consulting services
- Current and aspiring IT professionals

METHODOLOGY

Data Collection:

- Sources:
  1. Subset of data from the source 2019 Stack Overflow Developer Survey - using Python's Pandas library
  2. GitHub Jobs API – using Python requests library
  3. Popular Programming Languages - using Web scraping

Data wrangling:

- Performed to clean the data from duplicates and missing values.
- Normalization - making information easier to group and analyze
- Tool – Pandas library

Exploratory data analysis:

- To identify distribution of data, remove outliers and find the correlation between features in the dataset.
- Tool – Pandas, NumPy libraries

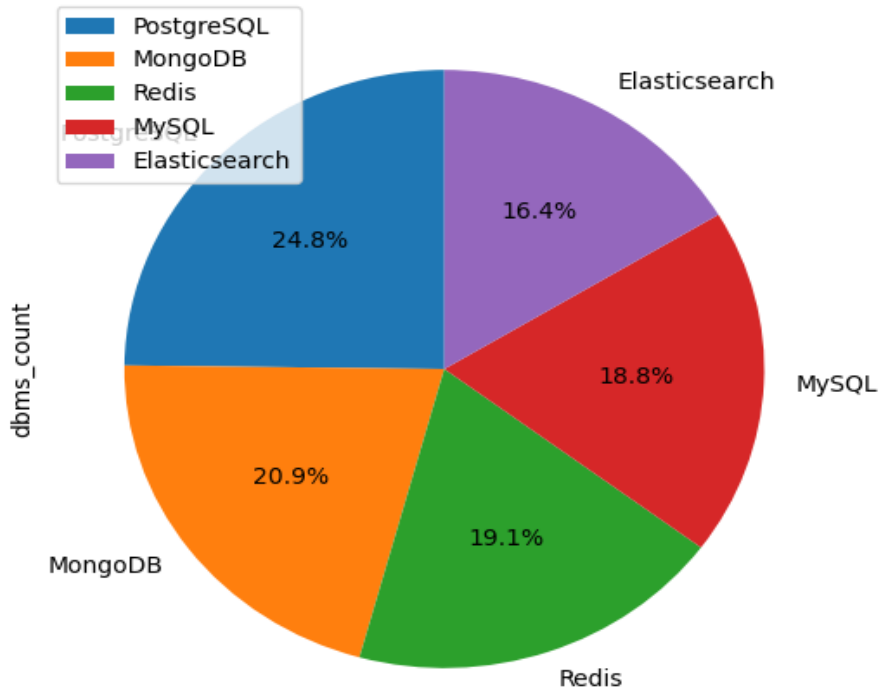
Data visualization:

- Charts, plots and graphs to visualize the distribution, composition, comparison of data and relationships between data.
- Tool - SQLite, Pandas, NumPy, matplotlib libraries and SQL
- The results and key points of the analysis are visualized into appropriate chart types, thus assembling a dashboard. Tool - IBM Cognos Analytics.

RESULTS

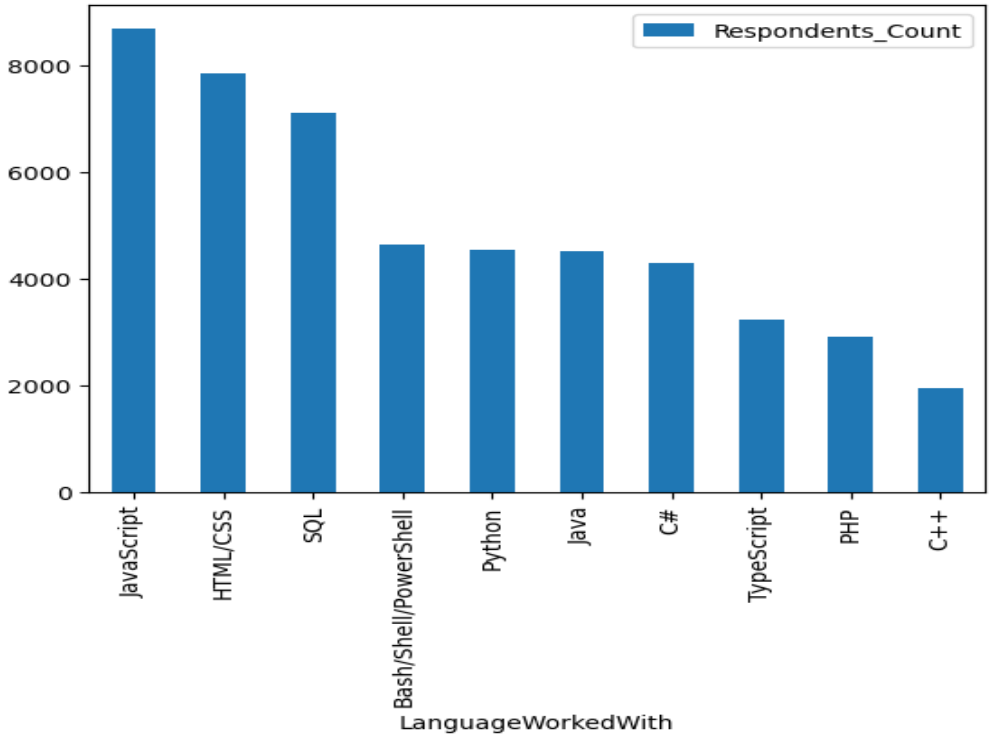
- Total no of rows after outliers removal = 10,519
- Statistical analysis of total annual compensation(in USD) gives:
  1. Median = $52,704
  2. Mean = $59,883
- Age shows a positive correlation with all other numerical columns except for CodeRevHrs (hours per week spent on code review)
- The most desired technologies in the future:
  1. Database – PostgreSQL
  2. Language – JavaScript
  3. Platform – Linux
- Gender distribution – Male respondents(93.7%) are more predominant than female respondents(6.3%)

Top 5 databases that respondents wish to learn next year

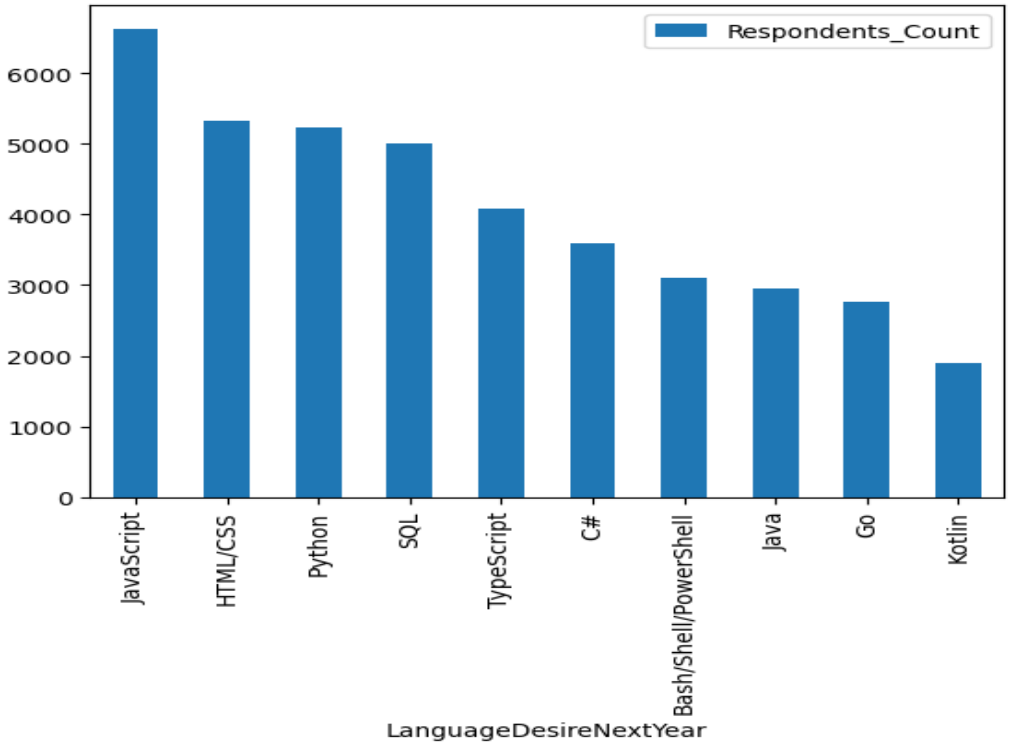


PROGRAMMING LANGUAGE TRENDS

Top 10 programming languages for the current year



Top 10 programming languages for the next year



PROGRAMMING LANGUAGE TRENDS - FINDINGS & IMPLICATIONS

Findings:

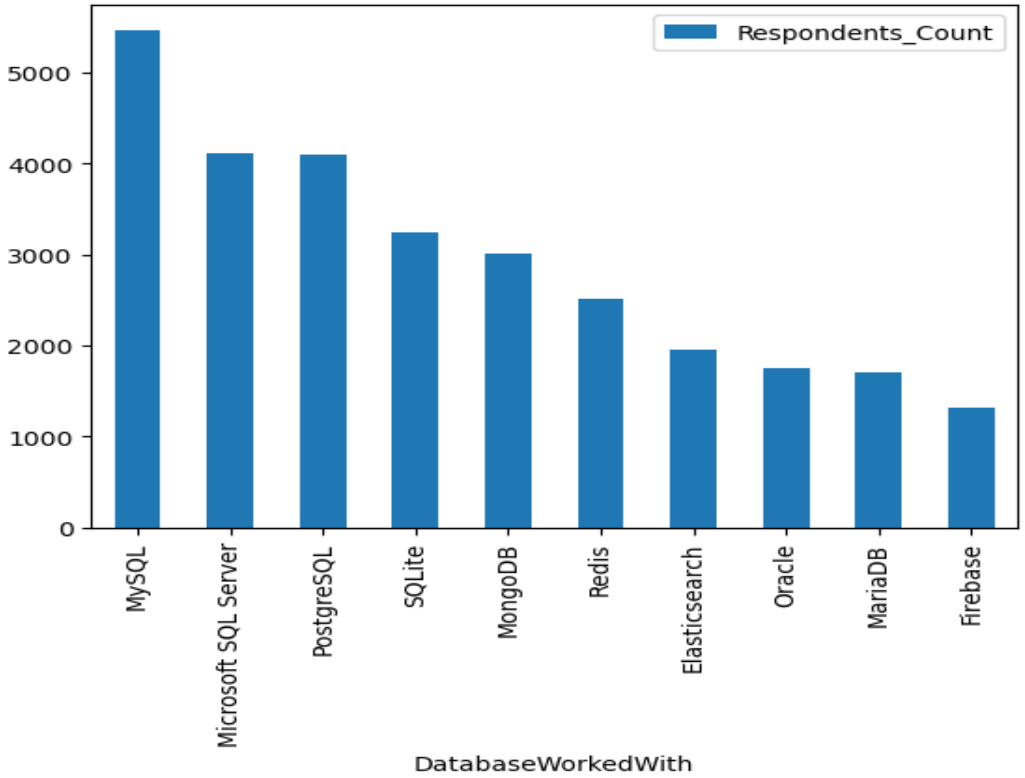
- JavaScript and HTML/CSS are the top two preferred languages, both in the current and next year
- Python and TypeScript are becoming more popular
- Demand for SQL is likely to remain the same in the future as in the present
- Popularity for Bash/Shell/PowerShell is decreasing

Implications:

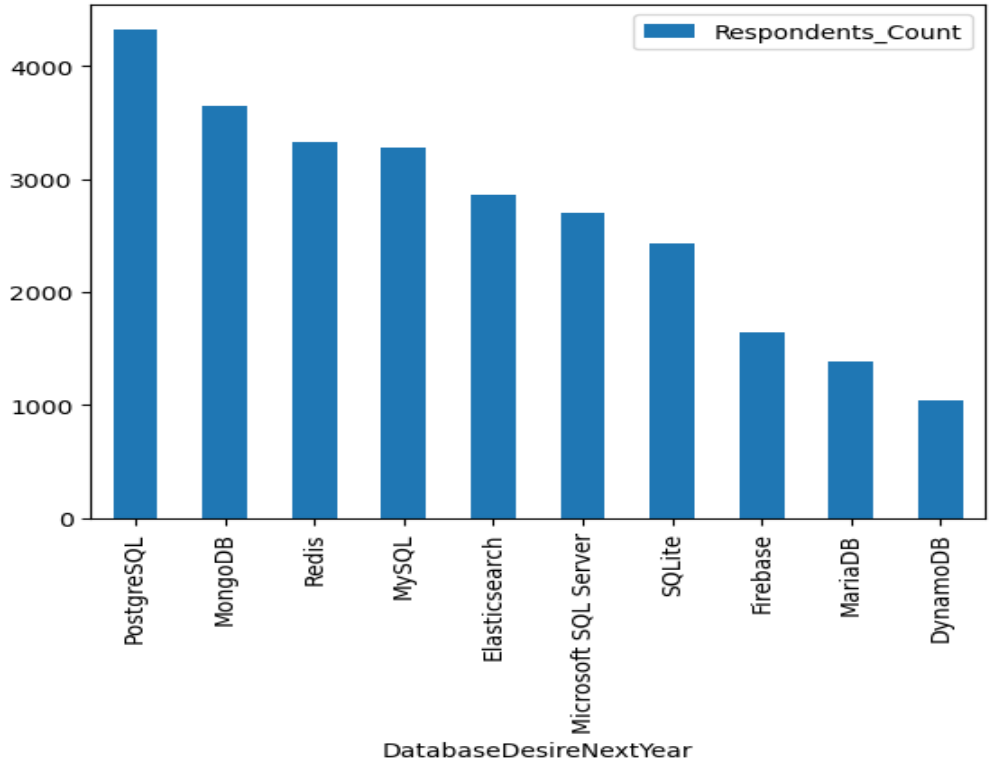
- Web development and web designing are the most dominating areas
- Python's increasing preference indicates data science, automation, AI, and machine learning fields will remain in high demand
- SQL is going to be the most preferred language used in relational database communications

DATABASE TRENDS

Top 10 databases for the current year



Top 10 databases for the next year



DATABASE TRENDS - FINDINGS & IMPLICATIONS

Findings:

- MySQL is the top database used currently, but PostgreSQL will likely be in highest demand in the future
- MongoDB and Redis are gaining popularity
- Microsoft SQL Server stands as the second-highest-used database currently but may grow less in demand in the future
- Increasing interest in Elasticsearch can be seen

Implications:

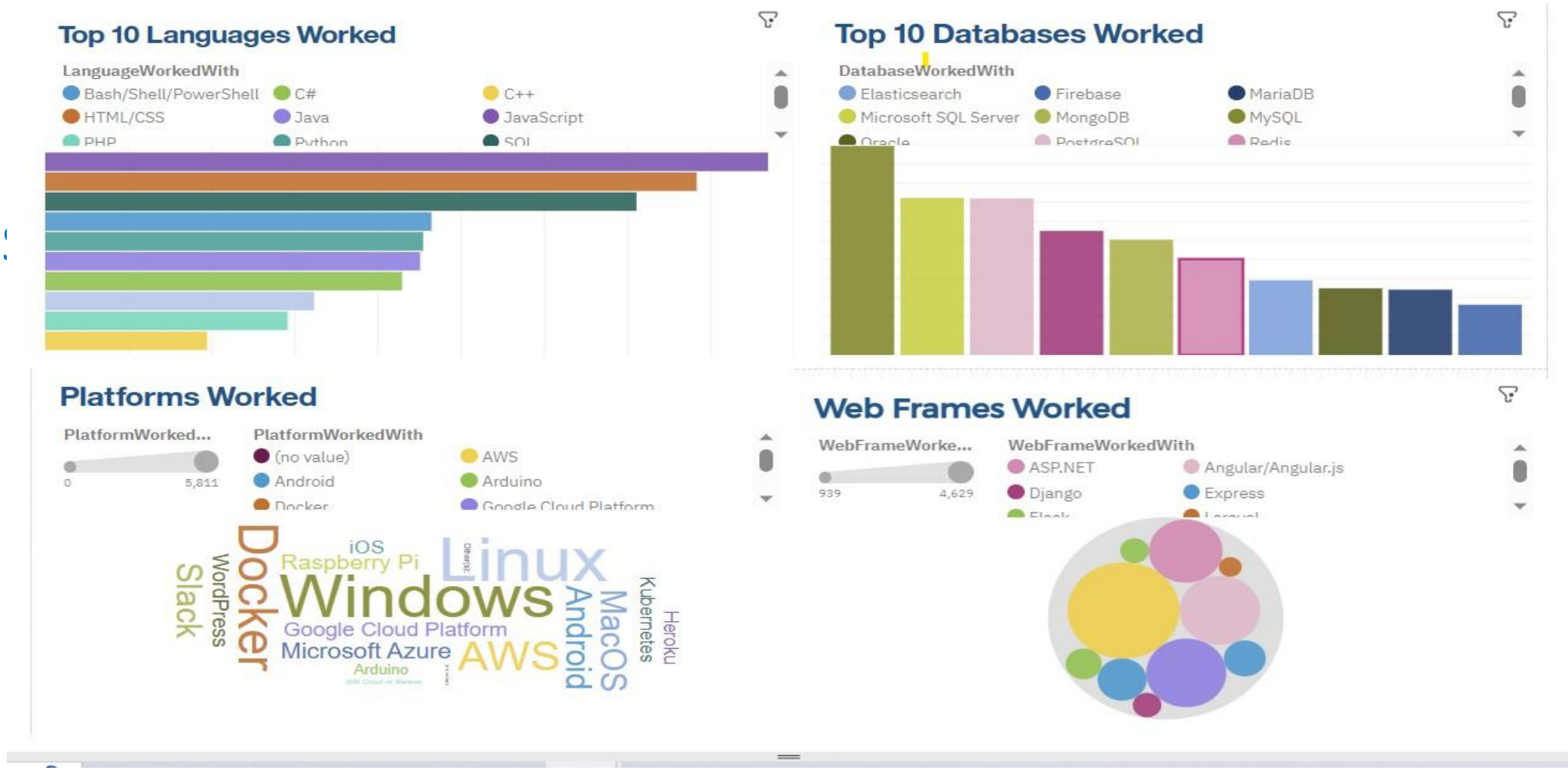
- Increasing demand for PostgreSQL, MySQL indicates a higher preference for open-source relational database programs
- NoSQL database programs are to be widely used in the future to store all data types, including structured and unstructured
- As open-source RDBMS gain popularity, proprietary database systems (like Microsoft SQL Server) might become less used.
- Programmers are likely to develop an interest in distributed, open-source search and analytics engines with the usage of Elasticsearch

DASHBOARD

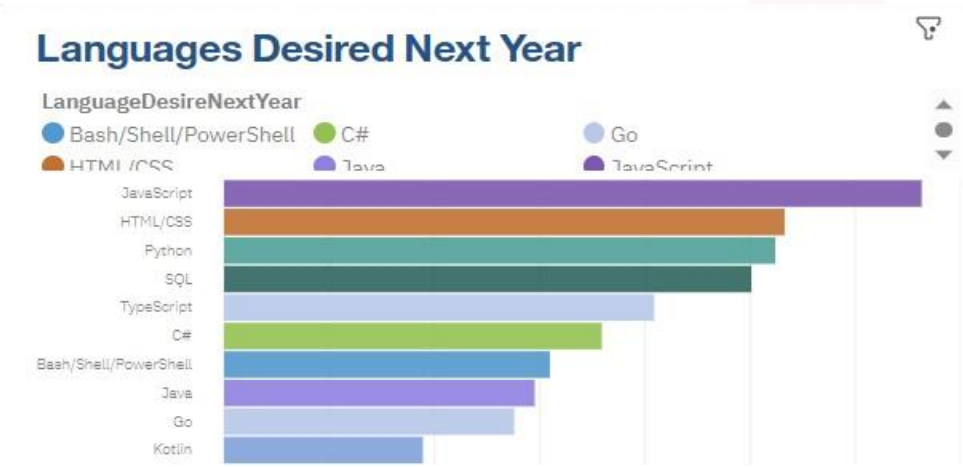
The below link for the Cognos dashboard highlights current and future technologies in demand and demographic distribution

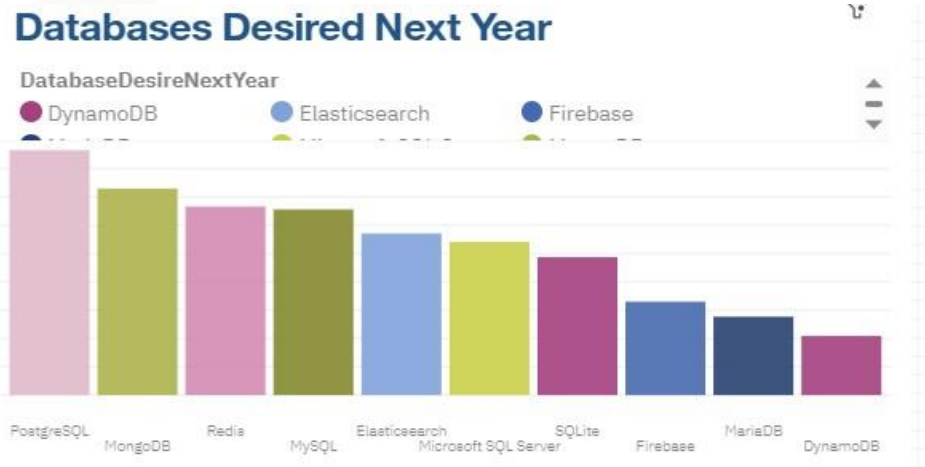
Link : https://dataplatform.cloud.ibm.com/dashboards/4f3e285f-035e-4099-938d-5cb8ea5240e9/view/7b3ea10465f563d473fdc0e4079b2d557b662d0eb4bbd10589827b495d627997f03f17c7c82e495cdb16026bf5e81a5a9d

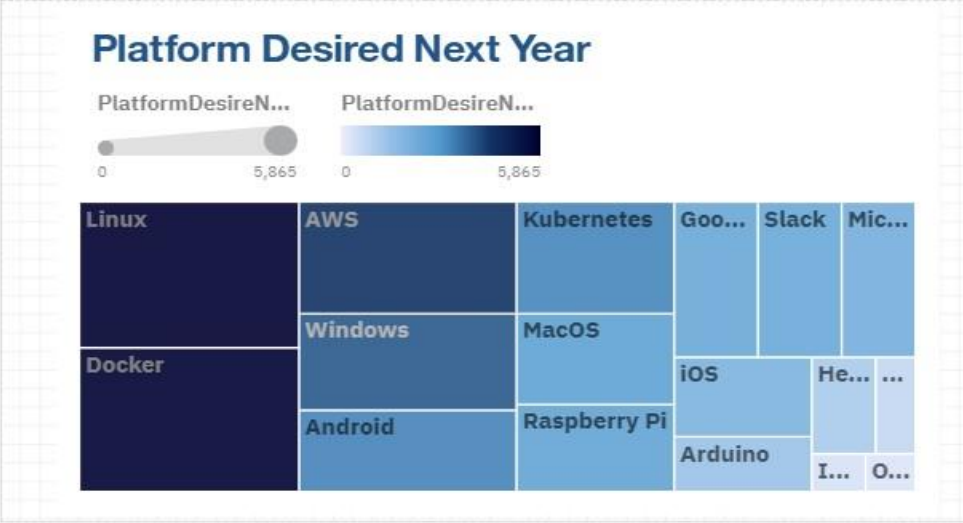
DASHBOARD TAB 1



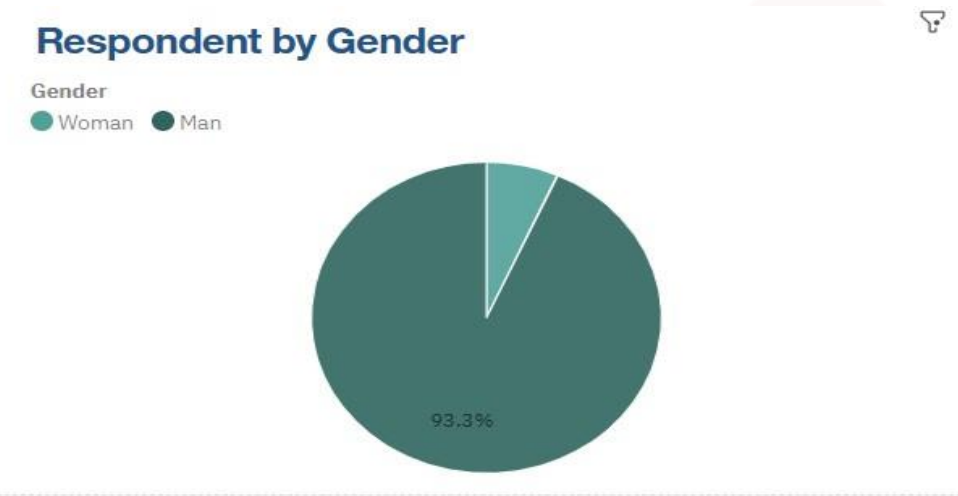
DASHBOARD TAB 2

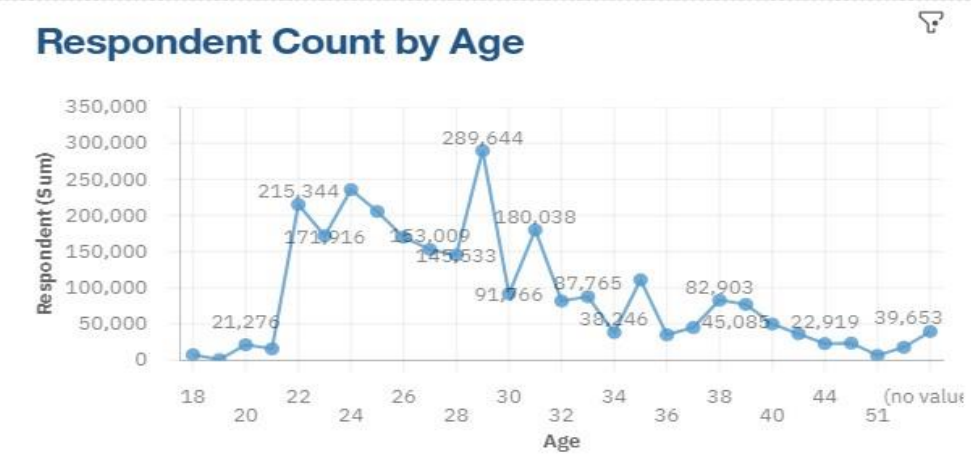


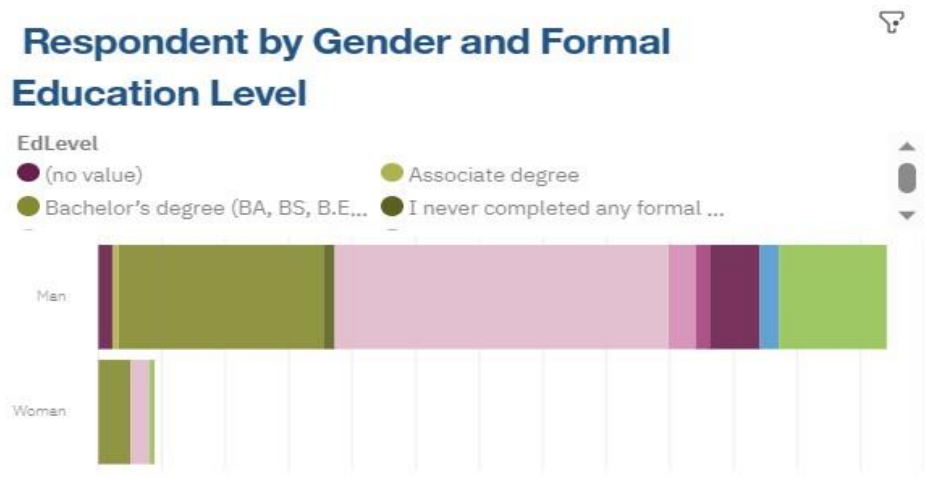




DASHBOARD TAB 3







DISCUSSION

Findings from analyzing the developer survey data can be taken as guidance in the following aspects:

- Top technologies that are used most frequently in the present and those that are going to be in high demand in the future.
- The results of the analysis help prospective developers, programmers, and data professionals gear their careers towards emerging technologies.
- Employers and consulting services in the IT industry can identify current and future skill requirements.
- Educators and aspiring students can focus on teaching and learning these in-demand skills.
- Bridging the gaps in demographics with respect to gender, age, location, and education.

OVERALL FINDINGS & IMPLICATIONS

Findings:

- JavaScript and HTML/CSS are the most dominant languages currently and in the future.
- Python, SQL, and Typescript are predicted to gain more popularity in the future, while usage of Bash, Shell, and PowerShell will decline.
- MySQL, Redis, and MongoDB databases are expected to be in high demand. PostgreSQL will be the most widely used database in the future.
- Linux and Docker will be widely used platforms next year. React.js, Vue.js, and Angular.js are going to be the top three preferred web frames.
- A huge gender gap is seen, with male workers being the dominant group.
- A high concentration of respondents are from the United States.
- Median age of respondents is 29 years.
- The majority of the respondents hold either a bachelor's or master's degree.

Implications:

- Web development and web designing are the top preferred areas now and in future. Prospective developers will highly benefit from learning JavaScript, HTML/CSS and Typescript.
- With the growing need to handle huge amounts of data and the emergence of AI, data professionals should consider enhancing their skills in NoSQL database programs in addition to Python, SQL, and PostgreSQL.
- Prospective employers and business services should consider bridging the gender gap by hiring more female workers.
- IT businesses and consulting services should keep pace with emerging technologies to remain competitive.
- Technology should be spread among lagging countries in need.
- Education organizations and online learning platforms in IT fields should gear their curriculum towards high-demand technologies.

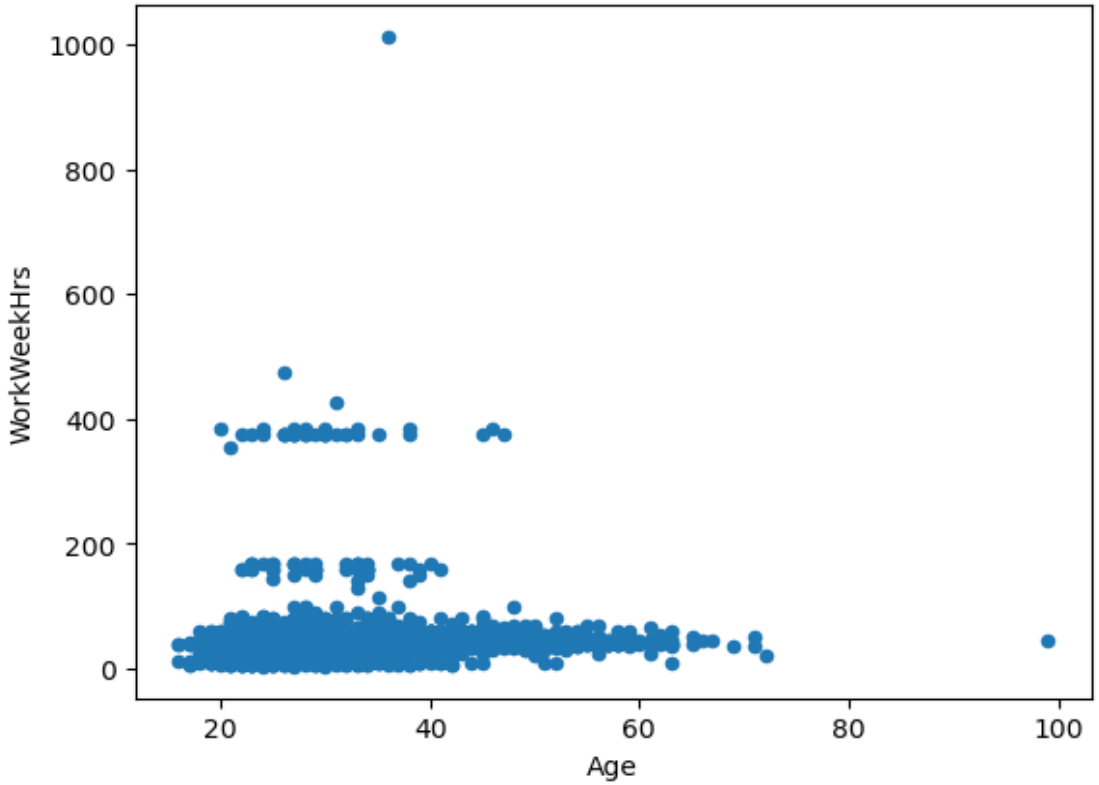
CONCLUSION

Thus, by analyzing a subset of the developer survey dataset, the following insights were drawn:

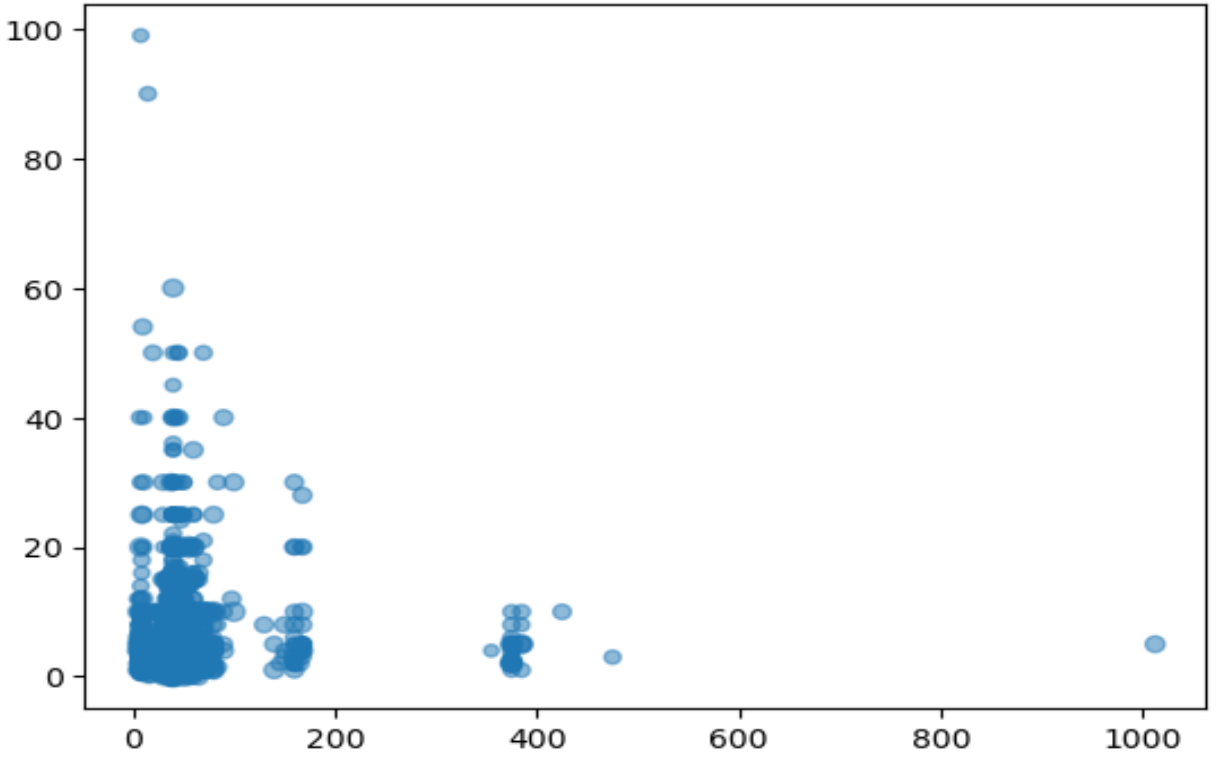
- Top trending technologies preferred by developers in the current and future
- Demographical distributions among IT professionals
- These insights will help business services, IT professionals, and educators identify and meet future skill requirements. They need to keep pace with emerging technologies to remain competitive. Employers can work towards balancing the gaps inferred from demographic data.

APPENDIX A

Scatter plot of Age and WorkWeekHrs(hours working per week)

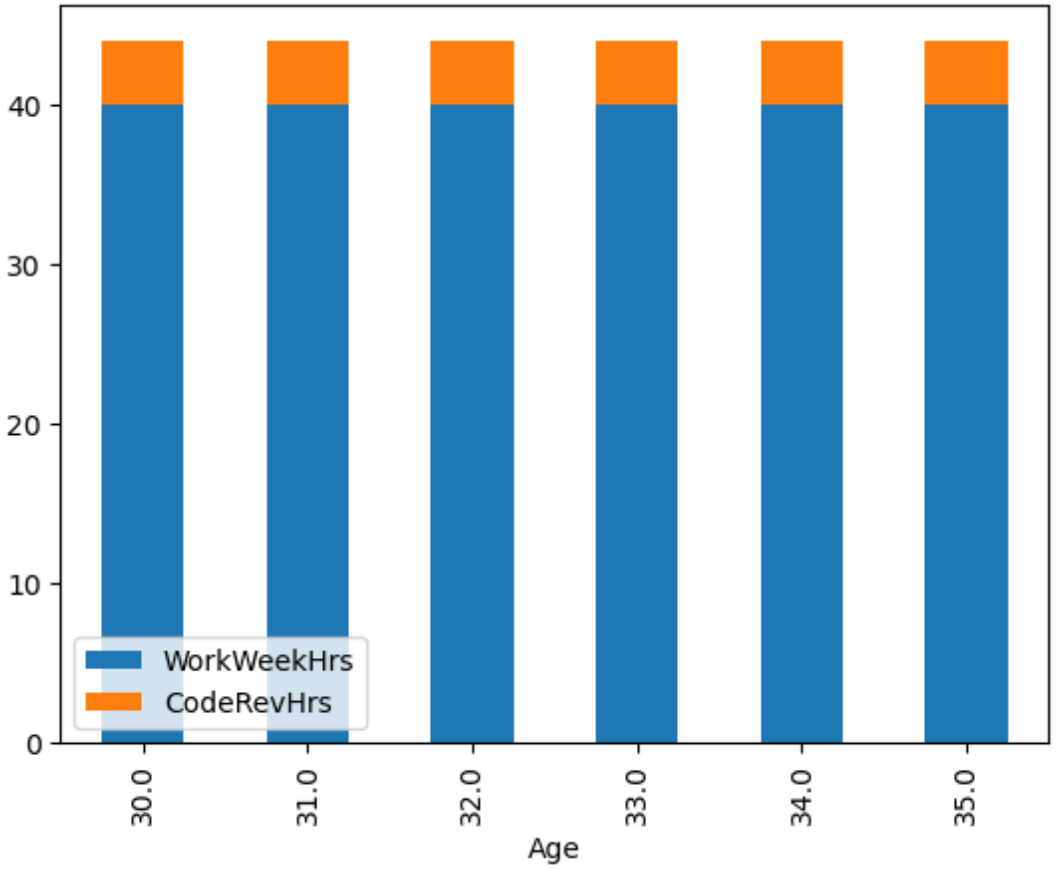


Bubble plot of Age and WorkWeekHrs

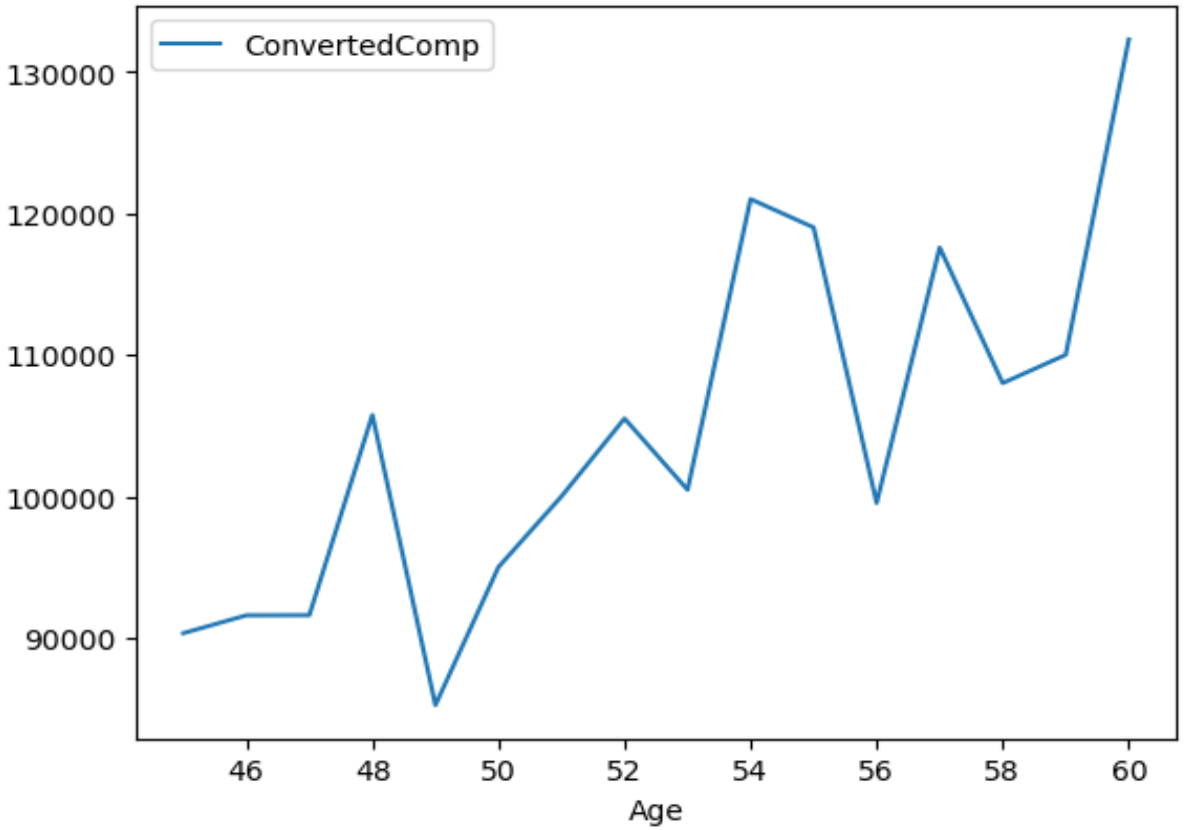


APPENDIX B

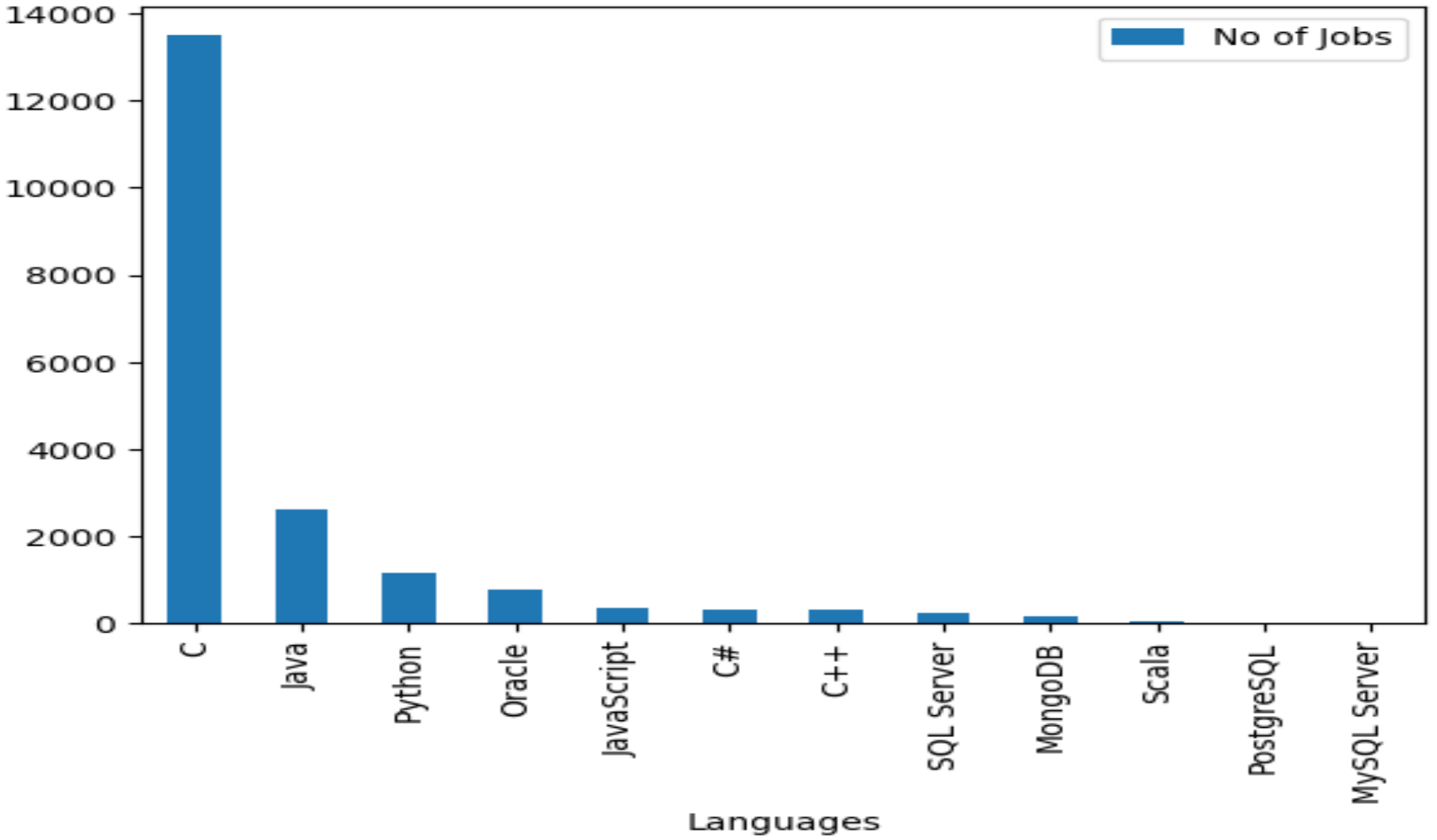
Stacked chart of median WorkWeekHrs(Working hours per week) and CodeRevHrs(hours per week spent on code review)



Median total annual compensation for all ages from 45 to 60



JOB POSTINGS



POPULAR LANGUAGES