Data Analysis on the 2019 Stack Overflow Survey of Software Professionals

Kevin Nguyen

September 2023

OUTLINE

EXECUTIVE SUMMARY

- This report analyzed the data from a 2019 Stack Overflow survey of software professionals.
- Respondents were asked a series of questions with relevance to personal background, current career, future aspirations, and more.
- Data analysis process included:
  - Data mining
  - Data wrangling and normalization
  - Data analysis and visualization
- The findings in this report are not only representative of the software professionals community, but also predictive of what skills and traits are important to acquire for career purposes.

INTRODUCTION

- This report is about the data gathered from the 2019 Stack Overflow survey of software professionals.
- The data is mined and wrangled for data analysis and the relationships between some of the variables of the survey were compared for potential correlation.
- The findings of the reports are relevant to aspiring and current software professionals because it could be predictive of what the current job market is like and what skills are necessary or important.
- The report explores the data to answer initial questions proposed such as:
  - What current technologies are relevant to software professionals?
  - Which technologies could be of importance to learn for the future or to meet the needs of the job market?
  - What are the demographic qualities of the respondents?

METHODOLOGY

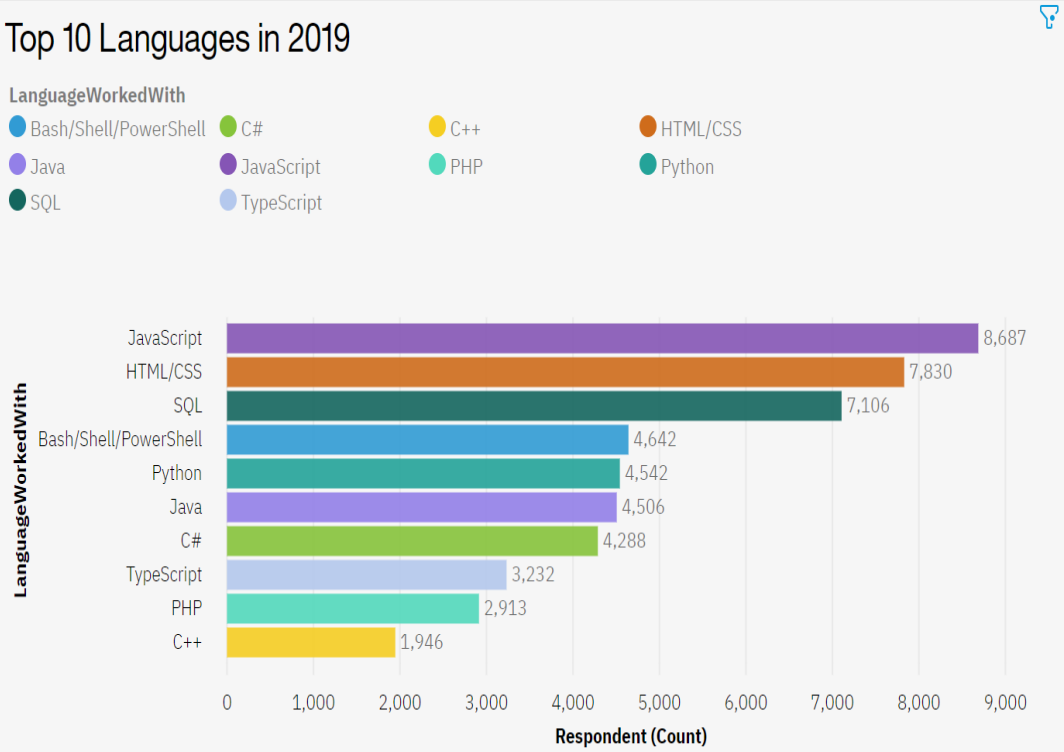
- The data comes from the following source: https://stackoverflow.blog under a ODbL: Open Database License.
- Survey dataset exploration using the Python software library, Pandas, to examine the dimensions of the dataset and the datatypes of each variable for the purpose of using statistical methods.
- Data wrangling to identify and remove duplicate and missing values in the dataset. Data normalization was performed for the sake of comparison among different variables.
- Exploratory data analysis using Python libraries such as Matplotlib and Seaborn to identify the distribution of the data, if there are any outliers, and if any correlation exists between features of the dataset.
- Data visualization using SQLite, Matplotlib, and IBM Cognos Analytics to visualize the distribution, composition, and comparison of data.

RESULTS

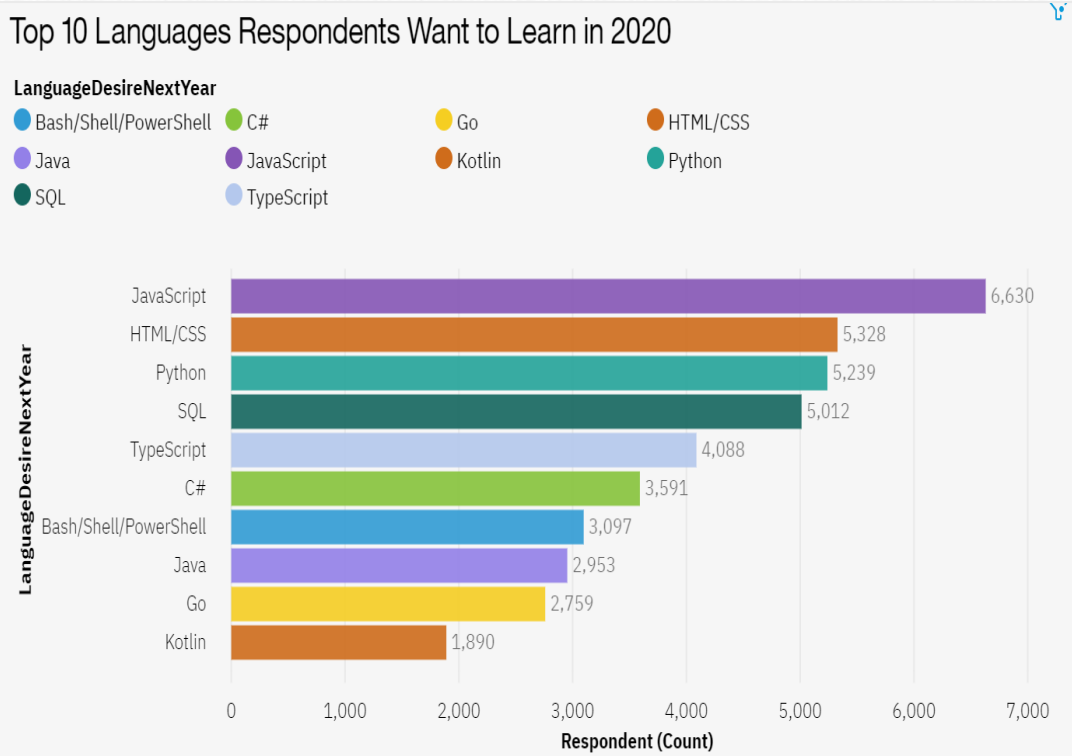
- The following slides show visualizations along with findings and implications for the dataset.
- A link to the read-only Cognos dashboard is available on slide 11.

PROGRAMMING LANGUAGE TRENDS

2019



2020



PROGRAMMING LANGUAGE TRENDS - FINDINGS & IMPLICATIONS

Findings

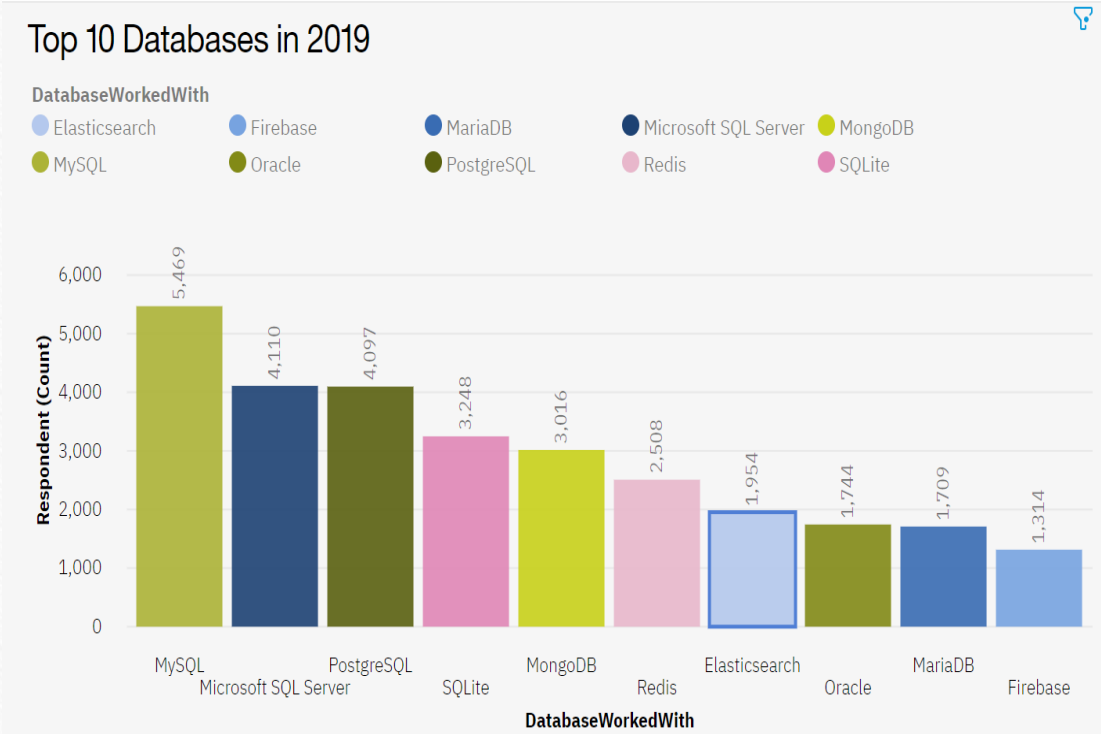
- JavaScript is the top programming language for both bar charts.
- As far as change in the top 3, Python overtook SQL in 2020 for the 3rd rank spot.
- Go and Kotlin overtook PHP and C++ to finish out the top 10 for 2020.

Implications

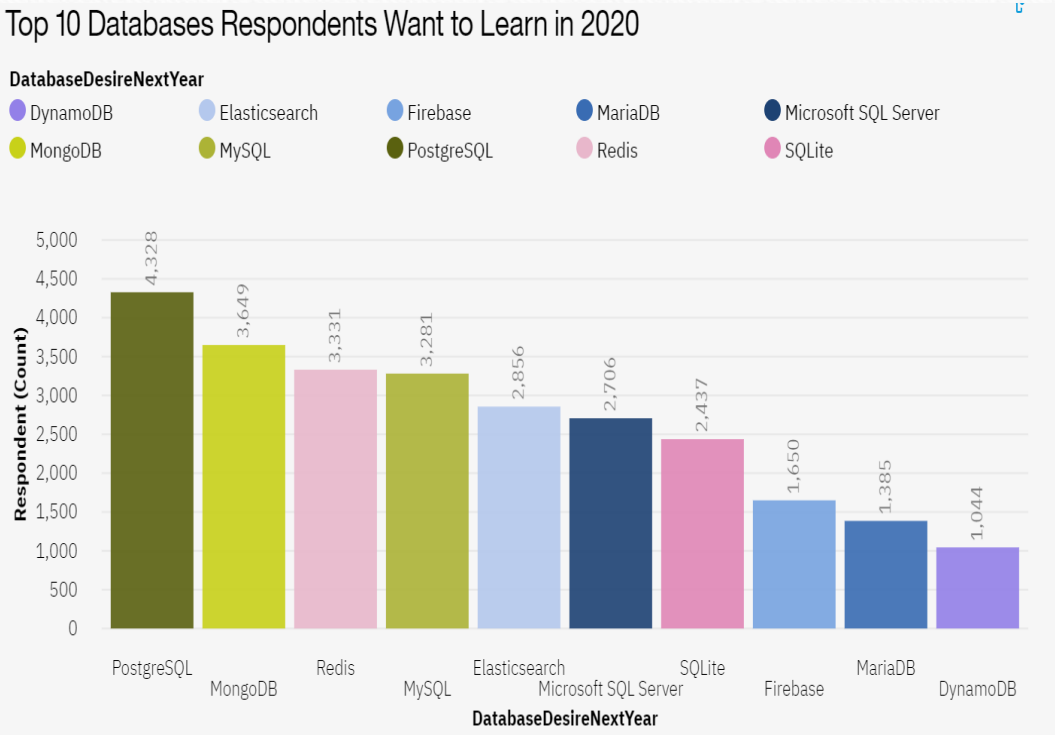
- JavaScript continues to be a programming language relevant to learn for the workforce among software professionals.
- Python is gaining popularity, more so than SQL.
- Go and Kotlin has become more relevant in the software space.

DATABASE TRENDS

Current Year



Next Year



DATABASE TRENDS - FINDINGS & IMPLICATIONS

Findings

- PostgreSQL overtook MySQL for the number one rank in 2020.
- Oracle fell out of the top 10, and DynamoDB made the top 10 for 2020.
- Outside of the change aforementioned, the rest of the top 10 databases from 2019 remain in the top 10 for 2020.

Implications

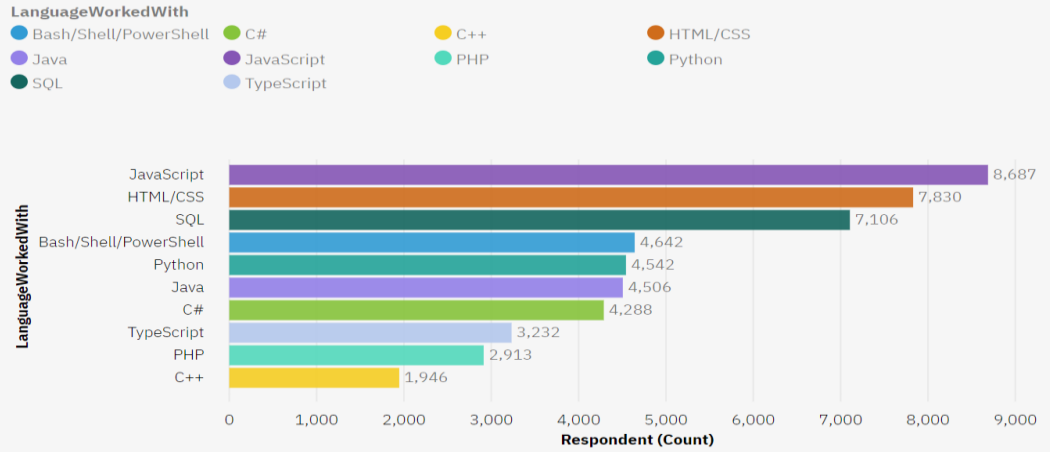
- PostgreSQL has become an extremely valuable language to know in the software space.
- Oracle regressed in popularity while DynamoDB increased in popularity.
- A majority of the top 10 databases from 2019 remain valuable languages to know for the software workspace.

DASHBOARD

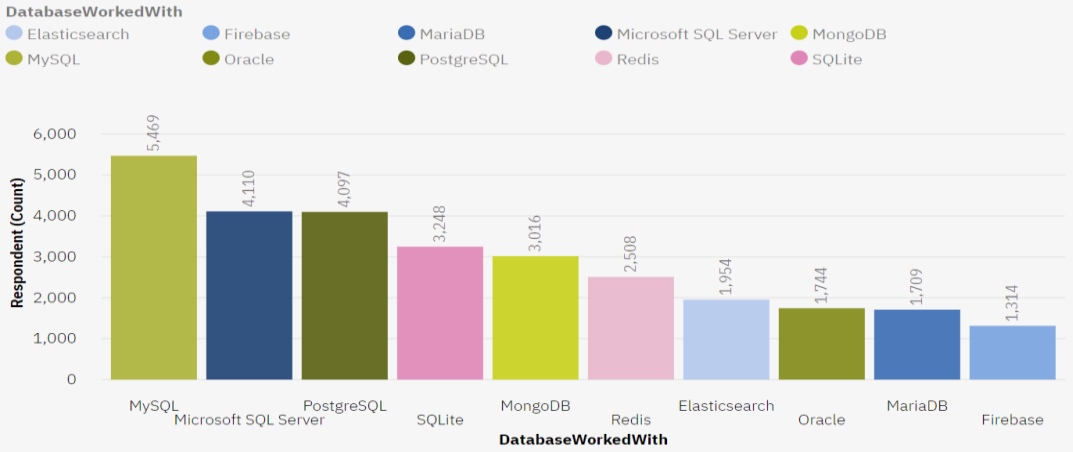
The permanent link of the read-only view of the Cognos dashboard: https://eu-de.dataplatform.cloud.ibm.com/dashboards

DASHBOARD TAB 1

Top 10 Languages in 2019

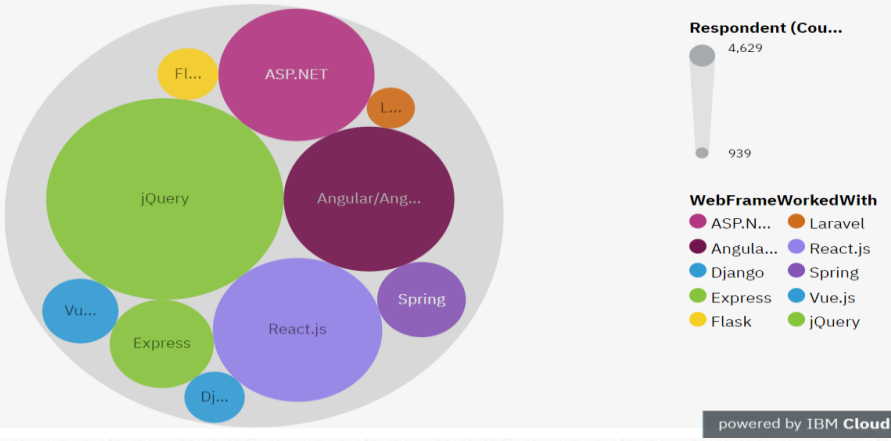


Top 10 Databases in 2019

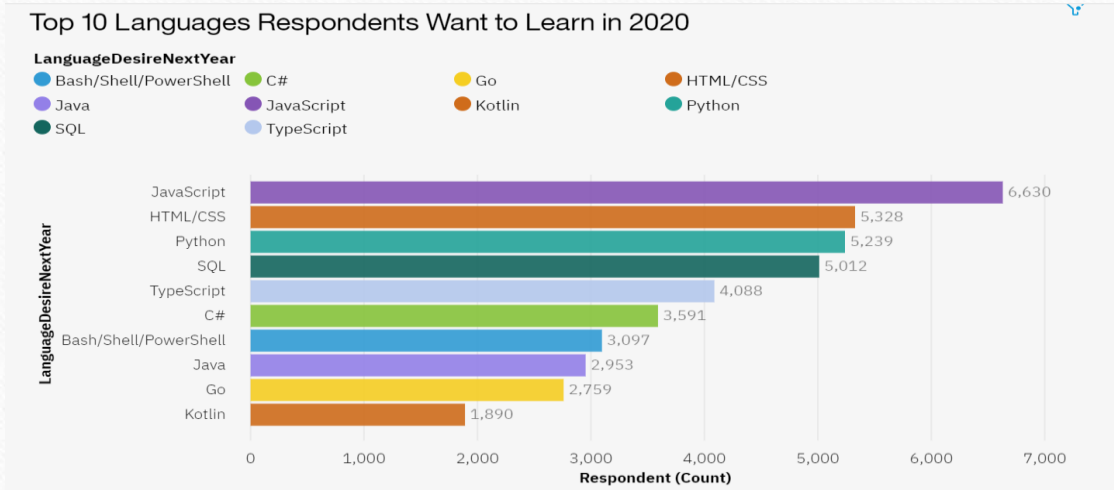


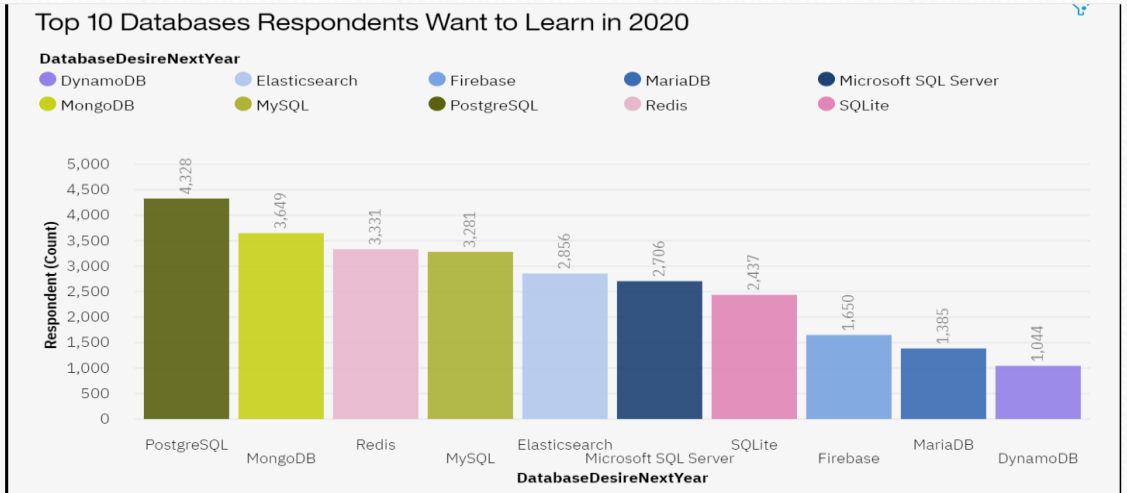
Platforms Utilized in 2019

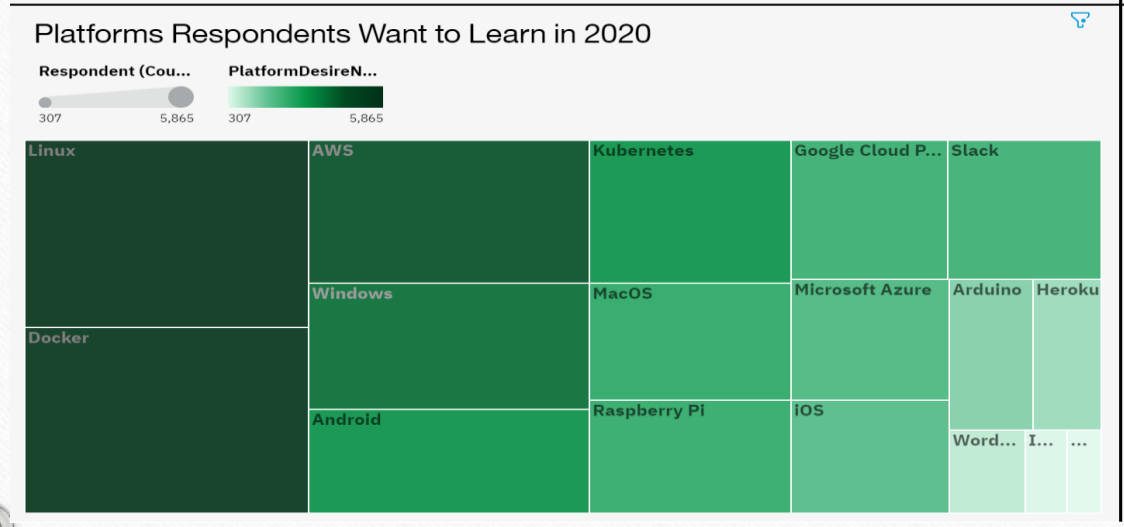
Top 10 Web Frames in 2019

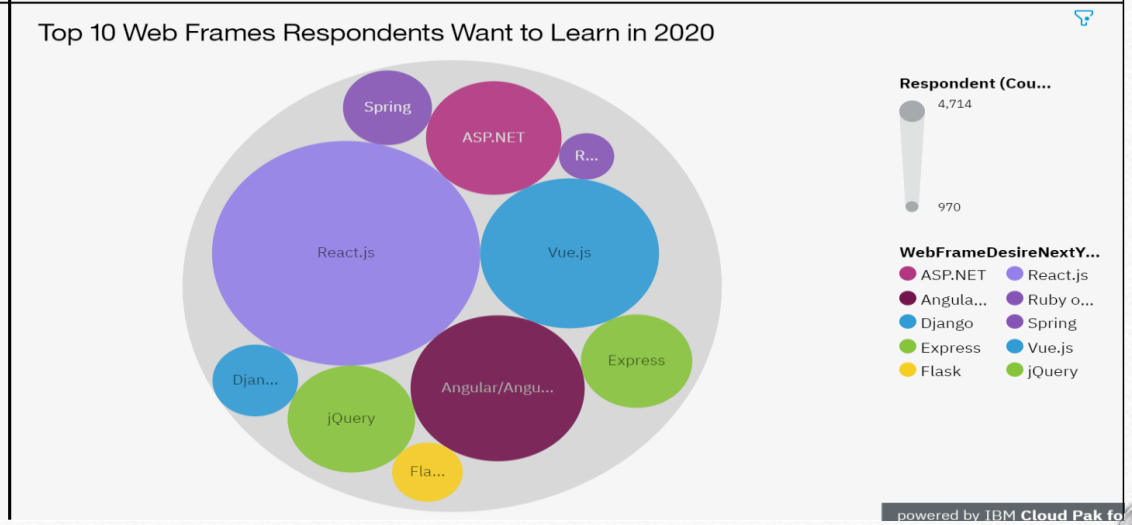


DASHBOARD TAB 2

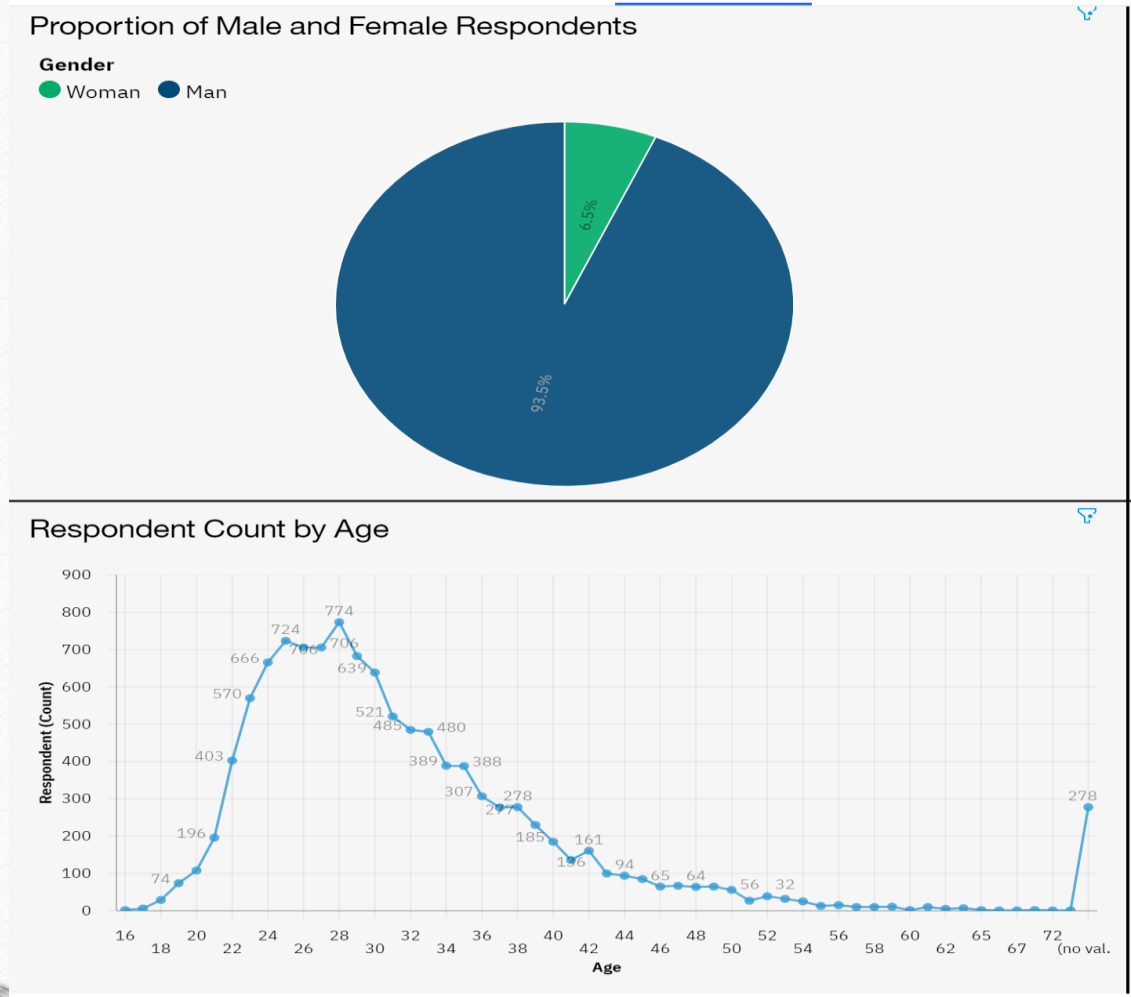








DASHBOARD TAB 3



DISCUSSION

- The results provide statistical evidence for the relationships among different variables.
- Correlation, or the lack thereof, can also be drawn from the visualizations.

OVERALL FINDINGS & IMPLICATIONS

Findings

- For the most part, the top 10 programming languages, databases, platforms, and web frames remain unchanged from 2019 to 2020.
- 93.5% of respondents are men, 6.5% are women.
- The United States accounts for the most number of respondents.
- Majority of respondents are ages 20-40.
- Majority of men and women have a bachelor's degree.

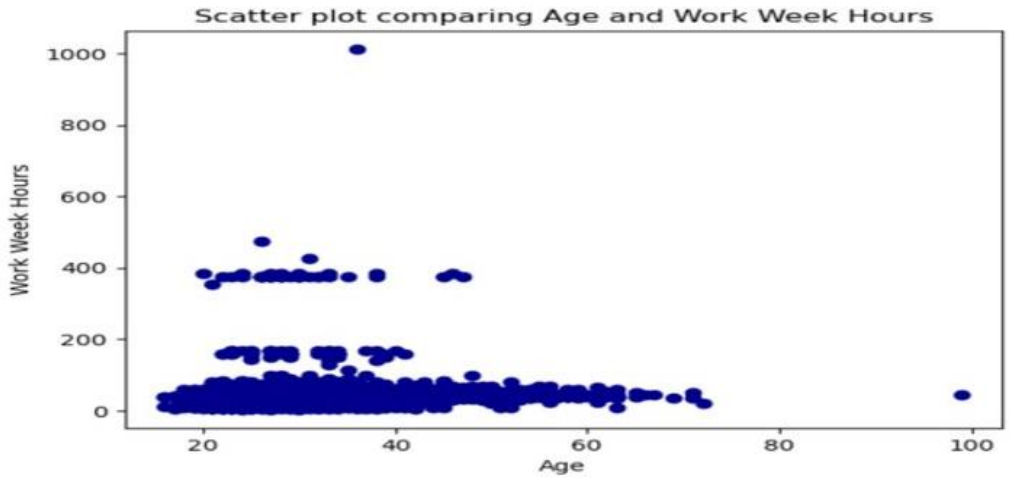
Implications

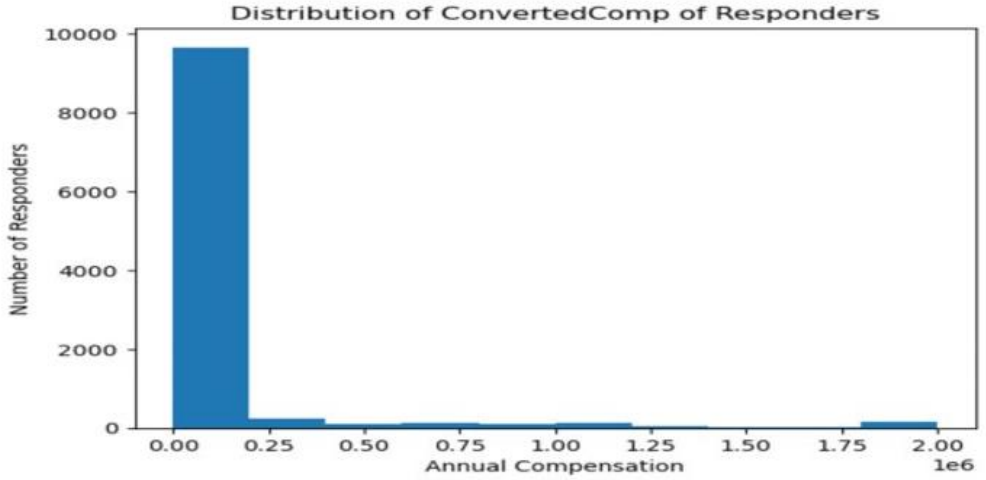
- The technologies relevant in 2019 are still relevant in 2020 and are worth learning for a profession in software.
- Software/tech careers are dominated by men.
- The concentration of respondents come from the USA.
- The age range could indicate that the median age of users of Stack Overflow are of ages 20-40 or that the majority of software professionals fall in this range.
- A bachelor's degree is the minimum educational requirement or preference for a majority of software roles.

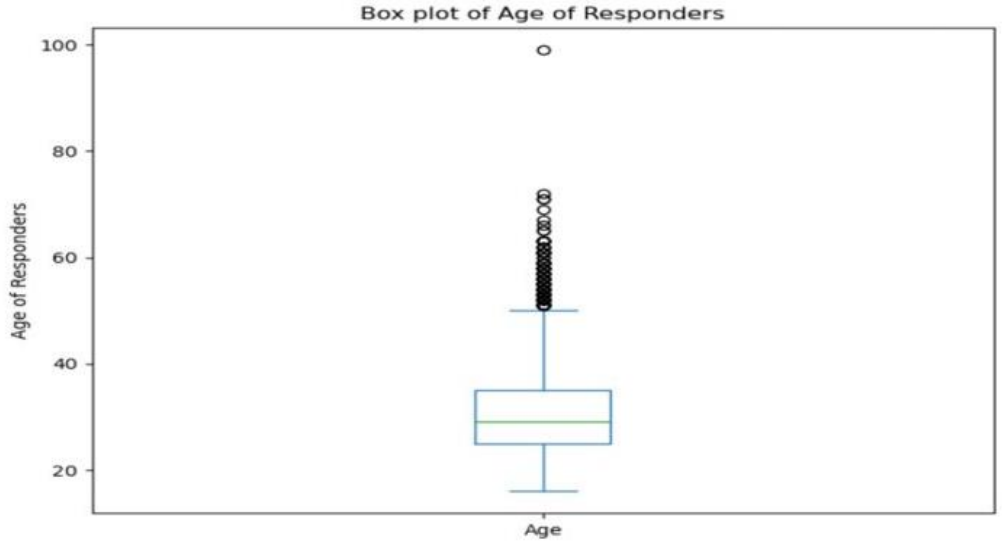
CONCLUSION

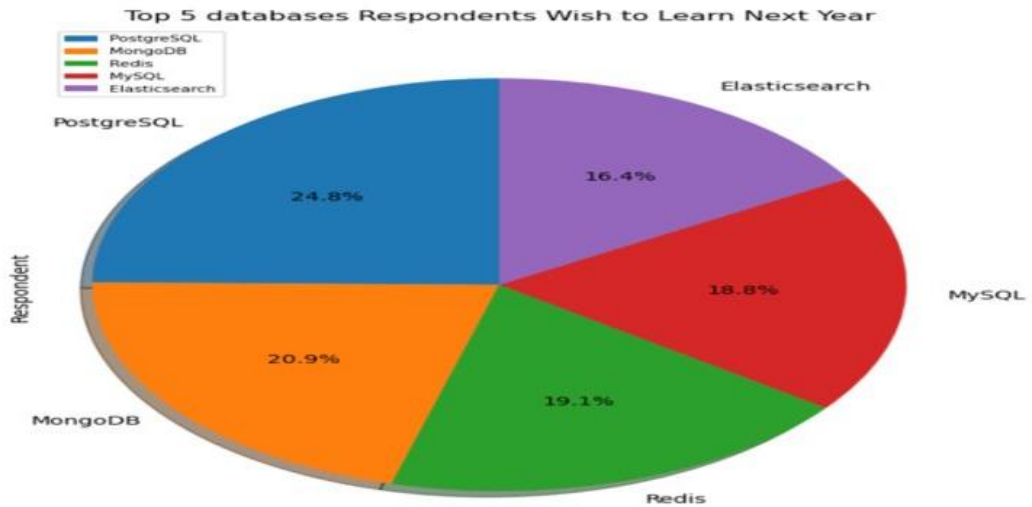
- Majority of the technologies remain relevant in the software workspace from year to year.
- Learning the technologies in the top 10 can be really valuable for aspiring software professionals to land a career or even for current professionals to move up in status.
- Software/tech roles remain a male-dominated space and success in landing such roles seemingly require, at minimum, a bachelor's degree.

APPENDIX

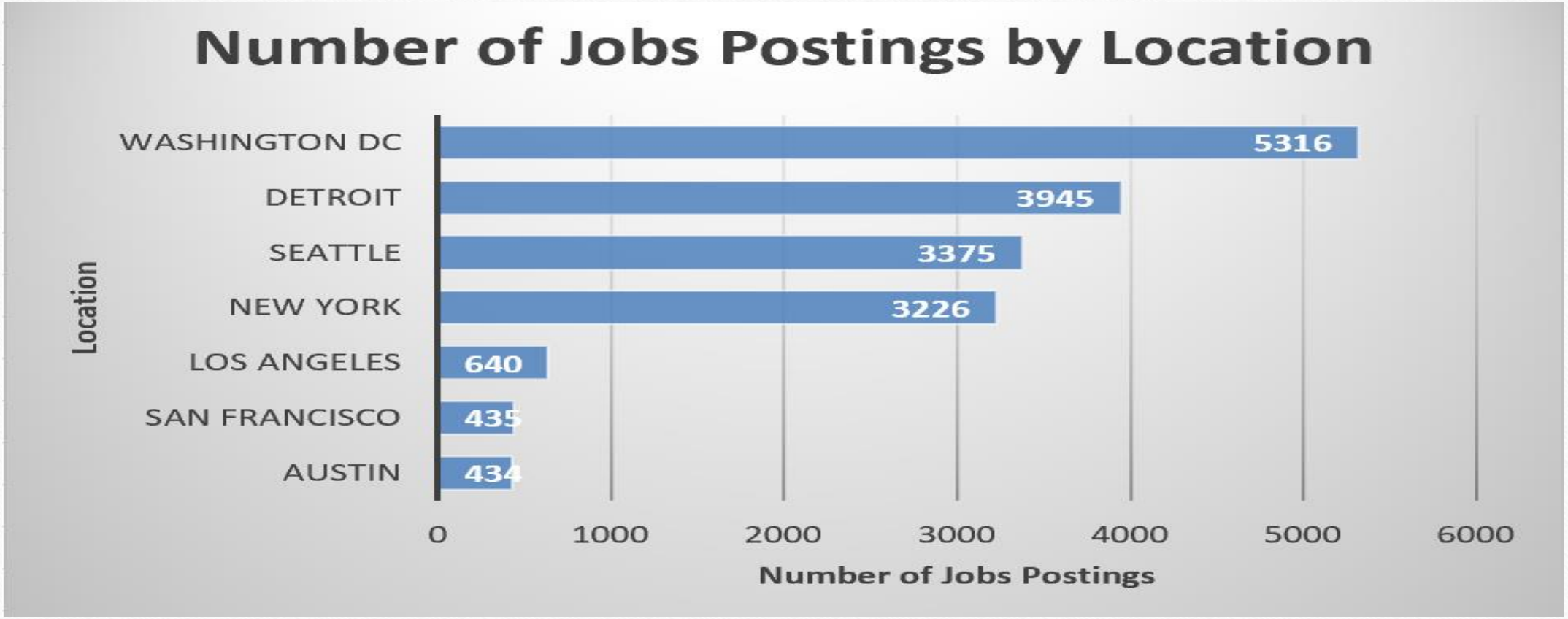

Scatter plot comparing Age and Work Week Hours
Work Week Hours
Age
0
200
400
600
800
1000
20
40
60
80
100


Distribution of ConvertedComp of Responders
Number of Responders
Annual Compensation
1e6
0
2000
4000
6000
8000
10000
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00


Box plot of Age of Responders
Age of Responders
Age
20
40
60
80
100


Top 5 databases Respondents Wish to Learn Next Year
PostgreSQL
MongoDB
Redis
MySQL
Elasticsearch
Respondent
24.8%
16.4%
18.8%
20.9%
19.1%

JOB POSTINGS



POPULAR LANGUAGES