IBM Capstone Project

A Stack Overflow Study of the Relationship Between Current Use and Future Use of Computer Technology and the Demographic Trends of the Respondents.

Angelina Dye, Analyst
Global IT & Business Services Firm
12/7/2024

OUTLINE

EXECUTIVE SUMMARY/ABSTRACT

This study is a categorical analysis of a subset of 2019 survey from the website Stack Overflow. The relationship between current and future technology use and the demographic trends were examined.

Two types of categorical data: Technology and Demographics were analyzed:

- Technology had subcategories: Language, Database, Platform, & WebFrame. These were each examined for their current and future technology use. After administering Chi-Squared Independence Tests, significant differences in relationships were found in Language, Database, and WebFrames.
- Demographics had subcategories: Age, Gender, Country, & Education Level. The most frequent Respondent was Male, in his early 20s, living in the United States, and had attained a Bachelor Degree or higher. The most significant differences found by Pivot Tables were in Gender, Country, and Education Level.

Top 10 rankings Dashboards were included. Trends found in the technologies and demographics are congruent with past research studies done by others. Further analysis with a larger dataset and longer time span and a, statistical analysis converted categorical data to numerical would be beneficial to follow this study for improved validity and replication.

INTRODUCTION

Purpose To analyze a subset of survey data for current technology use and future technologies Respondents desired to use the next year in 2020, along with the demographic trends of the Respondents.

Historical: Technology

Compiled past research from 4 popular online indices (Appendix C):

- IEEE Spectrum Index ▪ PYPL Index ▪ RedMonk Index ▪ TIOBE Index

2019 Rankings	Language	n/4
1st Place	JavaScript	2/4
	Python	1/4
	Java	1/4
2nd Place	Java	3/4

Figure 1: 2019 Index Rankings for Languages

2020 Rankings	Language	n/4
1st Place	Python	2/4
	JavaScript	1/4
	Java	1/4
2nd Place	Java	2/4

Figure 2: 2020 Index Rankings for Languages

INTRODUCTION

Historical: Demographics

1) One study examined the role of gender diversity in technology companies, as well as in academia (Botella, 2019, p. 30); They found that female STEM (Science, Technology, Engineering, Math) students had been decreasing for 20 years up until then, and professional females resigning from Technological careers.

2) Another study found that "Younger adults generally score higher on digital knowledge questions" (Vogels & Anderson, 2019).

Figure 3: Hypothetical Demographics

Hypothetical Demographics	
Age:	Low 20s
Gender:	Male
Country:	United States
Education Level:	>= Bachelor's
Programming Language Use:	JavaScript Python

INTRODUCTION

Hypothesis

Current & Future Technology Use

(Alternative Hypothesis)
If there is a relationship difference between the current and future Respondents' use of technology, then their survey answers for both 2019 and 2020 will be different, and the Null hypothesis will be rejected.

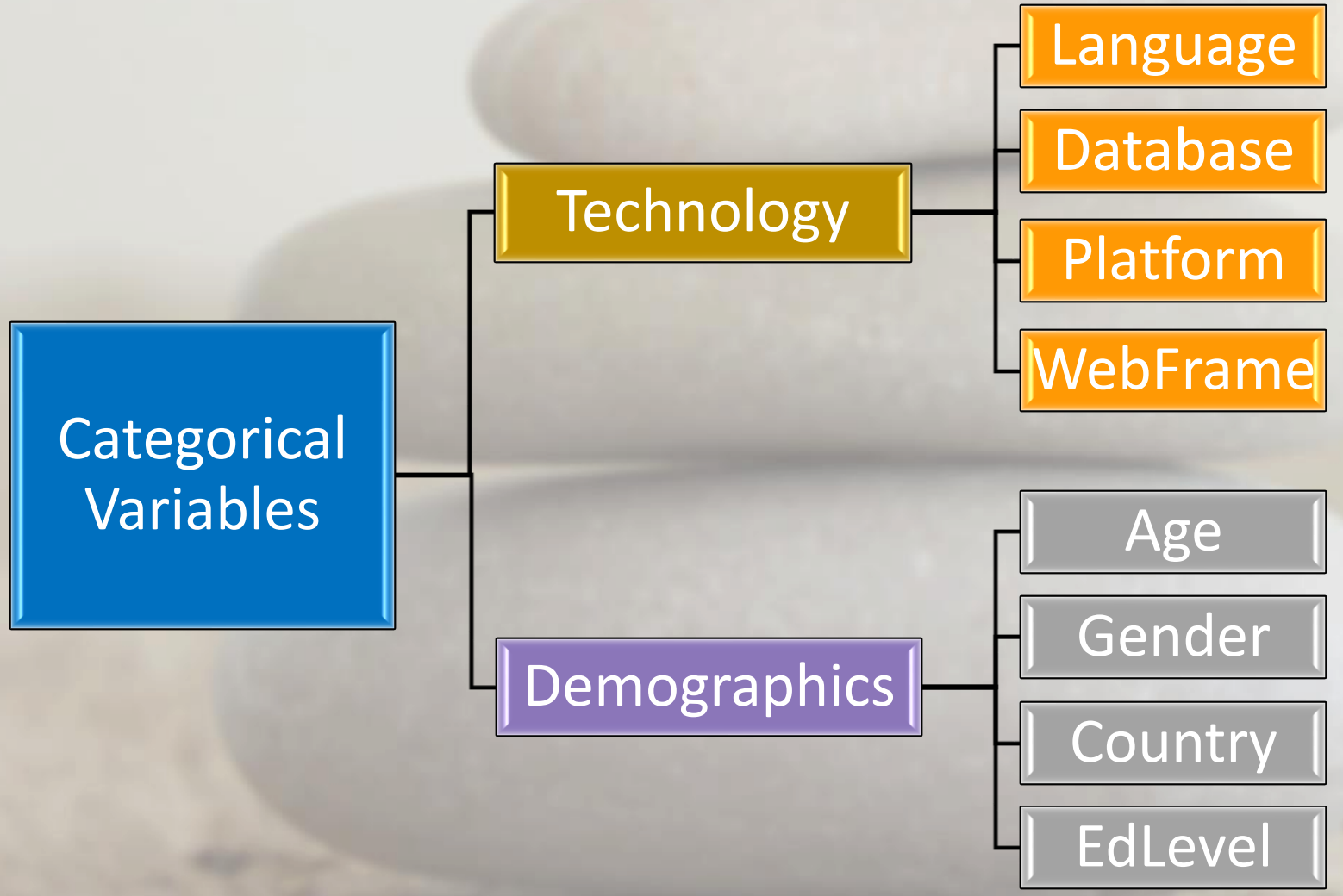
If there is no relationship difference between the current and future Respondents' use of technology, then the survey answers for both 2019 and 2020 will be the same, and the Null hypothesis will be affirmed.

Congruent with the past technology research, it is expected that JavaScript will be the most frequently used programming language.

Respondent Demographics

From the demographic reading, it is also expected that the Respondents answering the survey will most frequently be male and in early adulthood.

METHODOLOGY: Categorical Variable Organization



Future Use (2020)
Independent Variable

Current Use (2019)
Dependent Variable

Figure 4: Categorical Variable Organization

METHODOLOGY

Dataset: A Randomized Subset of Stack Overflow Developer Survey 2019

Provided by IBM/Coursera as a .csv file for this project.

https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/LargeData/m2_survey_data.csv

Methods & Tools:

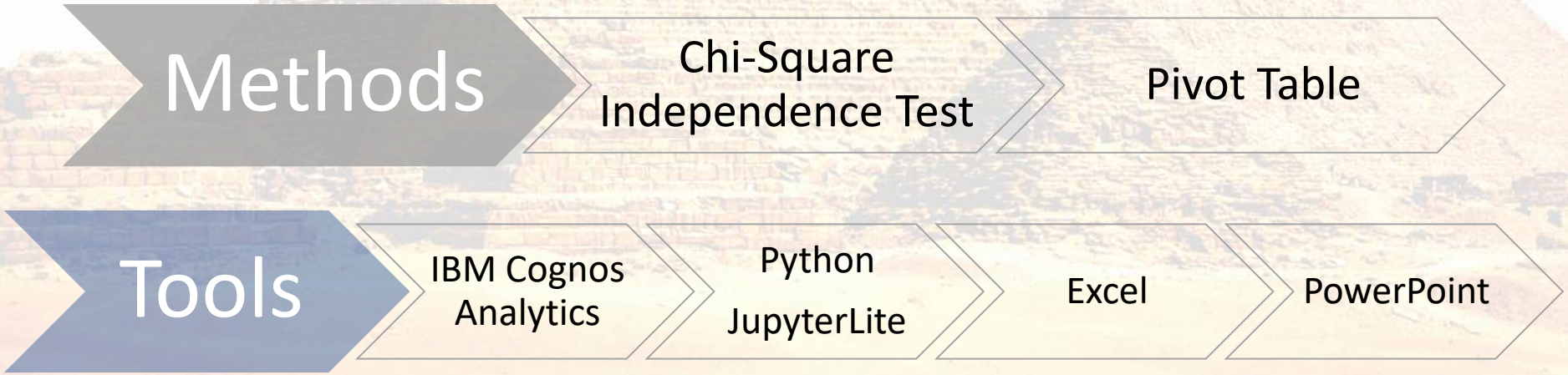


Figure 6: Methods and Tools

DATA: 11552 Rows/before cleaning, 9703/after cleaning; & 85 Columns

Duplicates

- n=154; Removed from dataset.

Missing Values

- n=32; In the column: “WorkLoc”

Imputed

- n=32; Replace in “WorkLoc” ; ∴ Most frequent (majority data): ‘Office’

Normalization

- Created new column: “NormalizedAnnualCompensation”
- Contains column “Annual Compensation” irrespective of the “CompFreqare currency.

Distribution

- Column “ConvertedComp” converted to annual USD (exchange rate 2019/02/01)

Outliers

- In “ConvertedComp”, using Boxplot, found and removed 879 Outliers.

Figure 5: Data Cleaning

RESULTS : Technology By Current & Future Use

Chi-Squared Analysis Technologies: Current Usage/Future Use

Technology	Chi-Squared Statistic	Degrees of Freedom	P-Value	Significant
LWW | LDNY	507814.8433	472098	4.525E-282	Significant
DWW | DDNY	1377.375	1292	0.048766676	Significant
PWW | PDNY	1850	1764	0.075475223	Not Significant
WWW | WDNY	1556.25	1406	0.002972611	Significant

Legend				
LWW | LDNY :	LanguageWorkedWith/ LanguageDesireNextYear		DWW | DDNY :	'DatabaseWorkedWith' /'DatabaseDesireNextYear'
PWW| PDNY :	PlatformWorkedWith/ PlatformDesireNextYear		WWW | WDNY :	'WebFrameWorkedWith' /'WebFrameDesireNextYear'

Figure 7: Chi-Squared Analysis Table

RESULTS : Technologies By Current & Future Use

Chi-Squared Analysis Results:

1) **Languages, Databases, and WebFrames:**

   - Had a statistically significant difference ($p<0.05$) between Current Usage and Future Use, with the strongest difference in Languages ($p<4.5E-282$).

2) **Platforms**:

   - There was no significant difference between current and future Platforms ($p=-0.08$).

RESULTS : Technologies By Current & Future Use

Chi-Square Languages Current & Future Heatmap

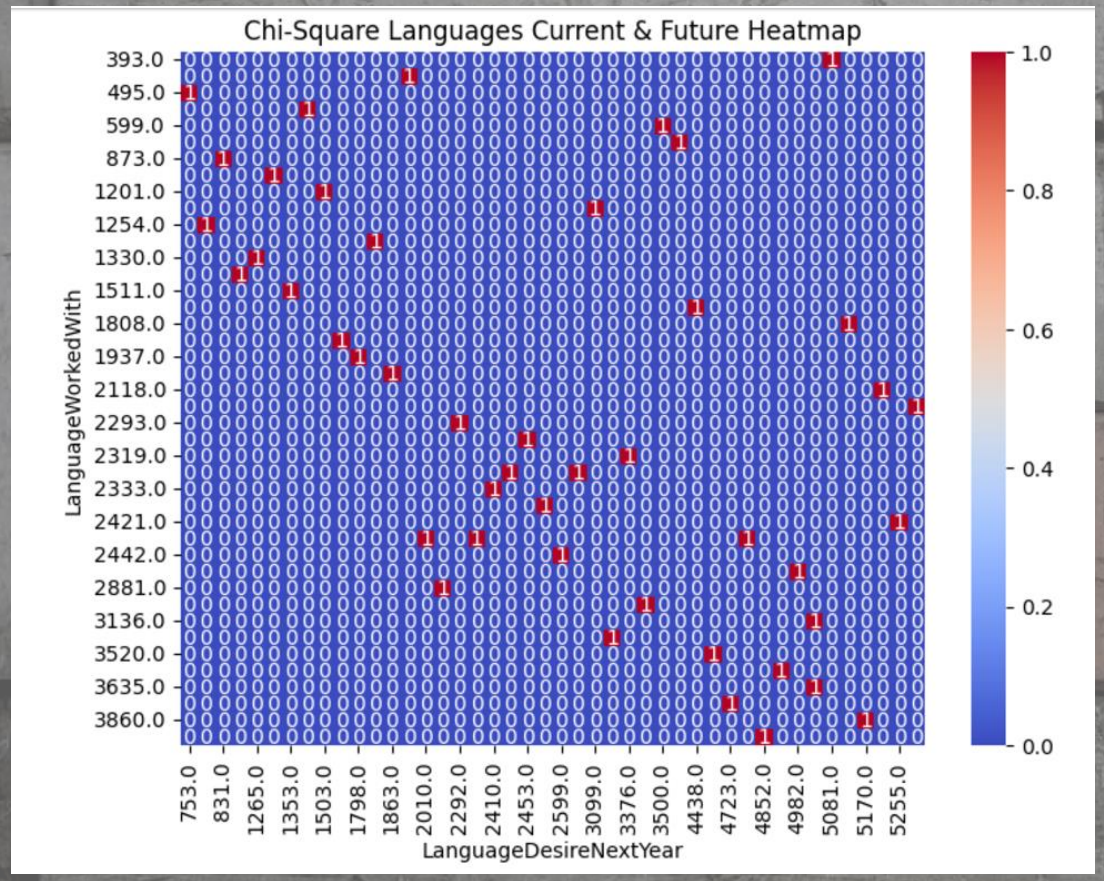


Figure 8: Chi-Square Languages Heatmap

Chi-Square Databases Current & Future Heatmap

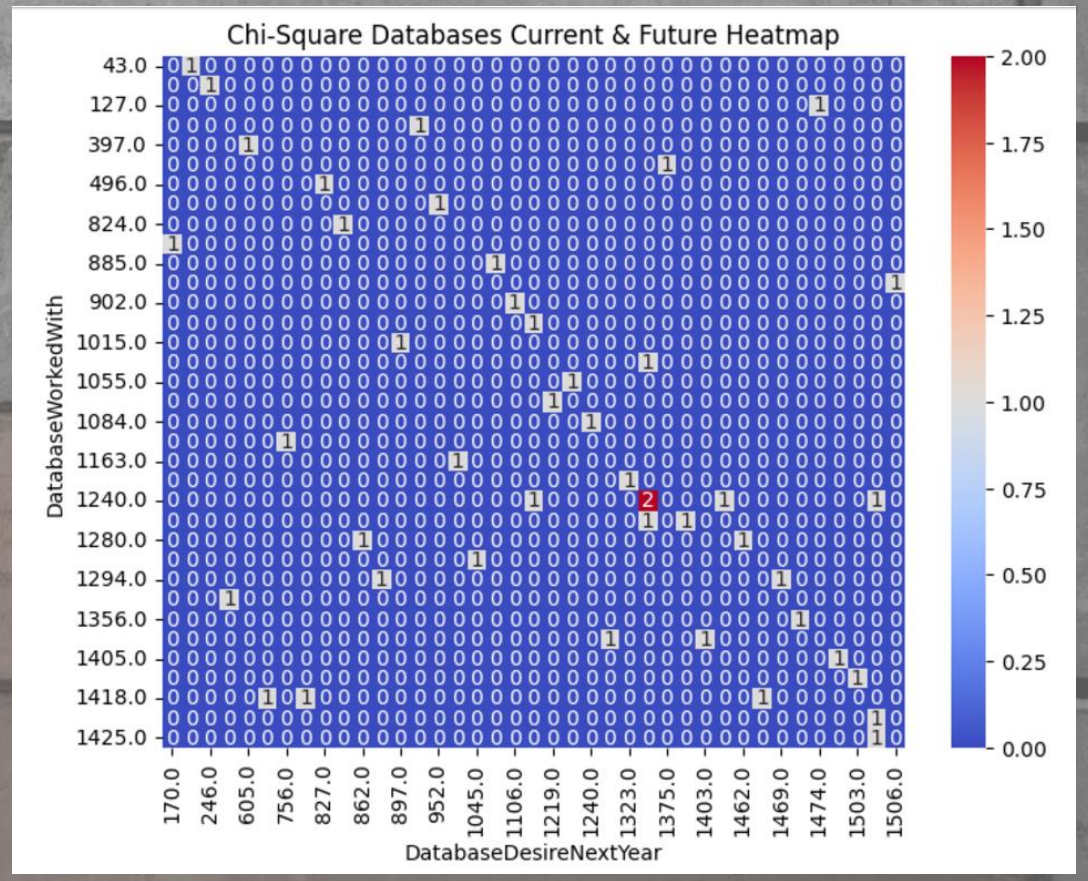


Figure 9: Chi-Square Databases Heatmap

RESULTS : Technology By Current & Future Use

Chi-Square Platforms Current & Future Heatmap

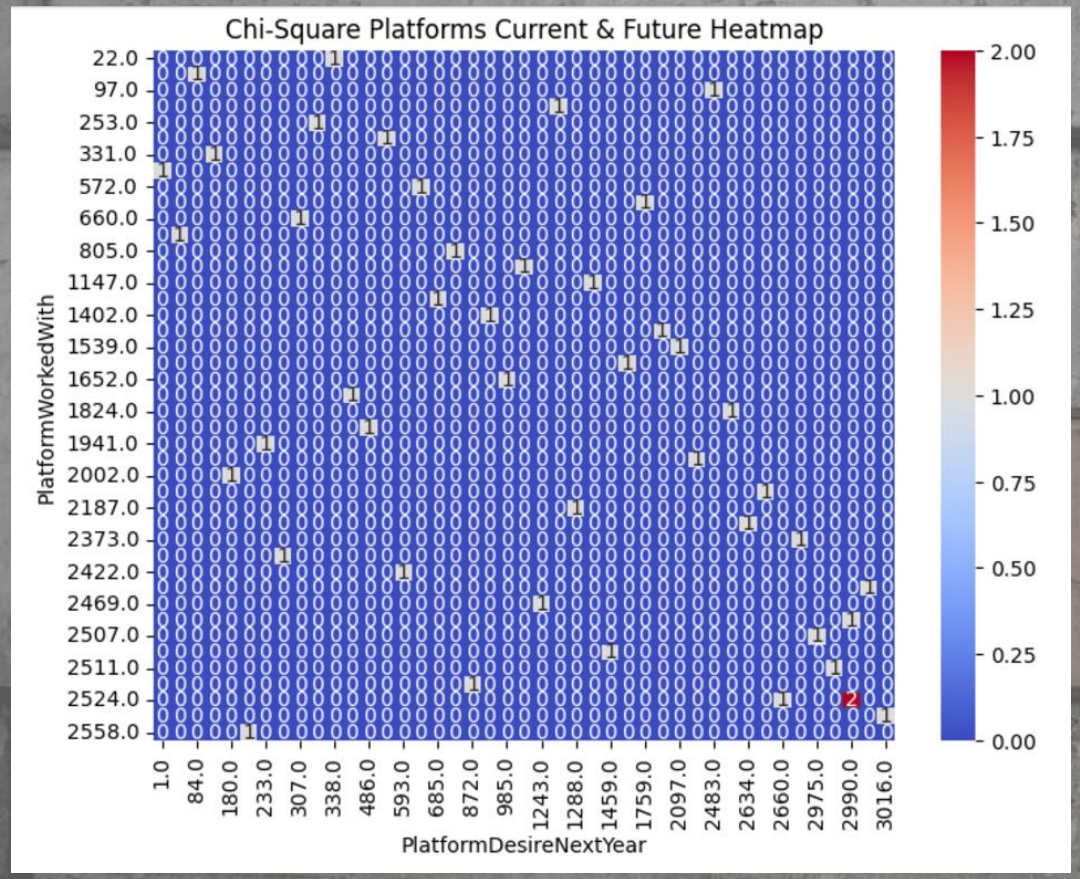


Figure 10: Chi-Square Platforms Heatmap

Chi-Square WebFrames Current & Future Heatmap

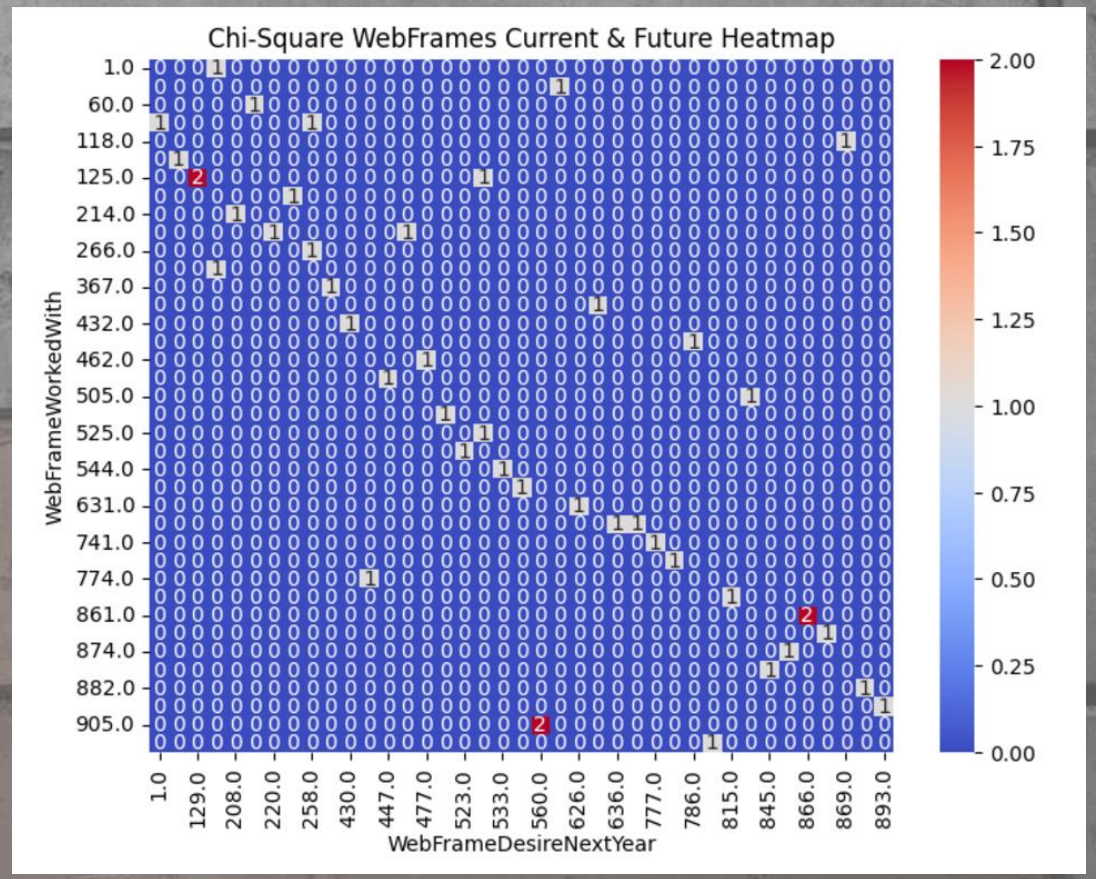


Figure 11: Chi-Square WebFrames Heatmap

RESULTS : Demographics of Respondents

Figure 12: Country & EdLevel Pivot Table

Country	Counts	Percentage	Difference
United States	2598	27%	20%
India	714	7%	
United Kingdom	657	7%	
Germany	622	6%	
Canada	360	4%	

EdLevel	Counts	Percentage	Difference
Bachelor's degree (BA, BS, B.Eng., etc.)	4922	51%	27%
Master's degree (MA, MS, M.Eng., MBA, etc.)	2310	24%	
Some college/university study without earning a degree	1195	12%	
Secondary school (e.g. American high school, German Realschule or Gymnasium, etc.)	455	5%	
Associate degree	328	3%	

Total Number of Respondents:	9703

Top 5 Counts and Percentages for 2019:

Country:

- The United States had the most Respondents (27%), compared to the second most popular country India (7%), with a big 20% difference.

Education Level (EdLevel):

- Half of the Respondents (51%) noted that their highest education level was a Bachelor's degree, while the second, with a quarter of Respondents, had attained a Master's degree(24%).

RESULTS : Demographics of Respondents

Figure 13: Age & Gender Pivot Table

Age	Counts	Percentage
28	683	7%
25	648	7%
26	636	7%
27	604	6%
24	591	6%

Gender	Counts	Percentage
Man	8895	92%
Woman	642	7%
0 (No answer)	65	1%
Non-binary, genderqueer, or gender non-conforming	52	1%
Man;Non-binary, genderqueer, or gender non-conforming	26	0%

Top 5 Counts and Percentages for 2019:

Age:

- There was no significant difference in the answers regarding Respondents' Age (Top: 28, 7%), with the top 5 ranging between 24:28 years old.

Gender:

- The most significant difference was in Respondents' Gender, with 92% of Respondents identifying as a Man, with Women at trailing far behind at 7%.

PROGRAMMING LANGUAGE TRENDS

Current Use: 2019

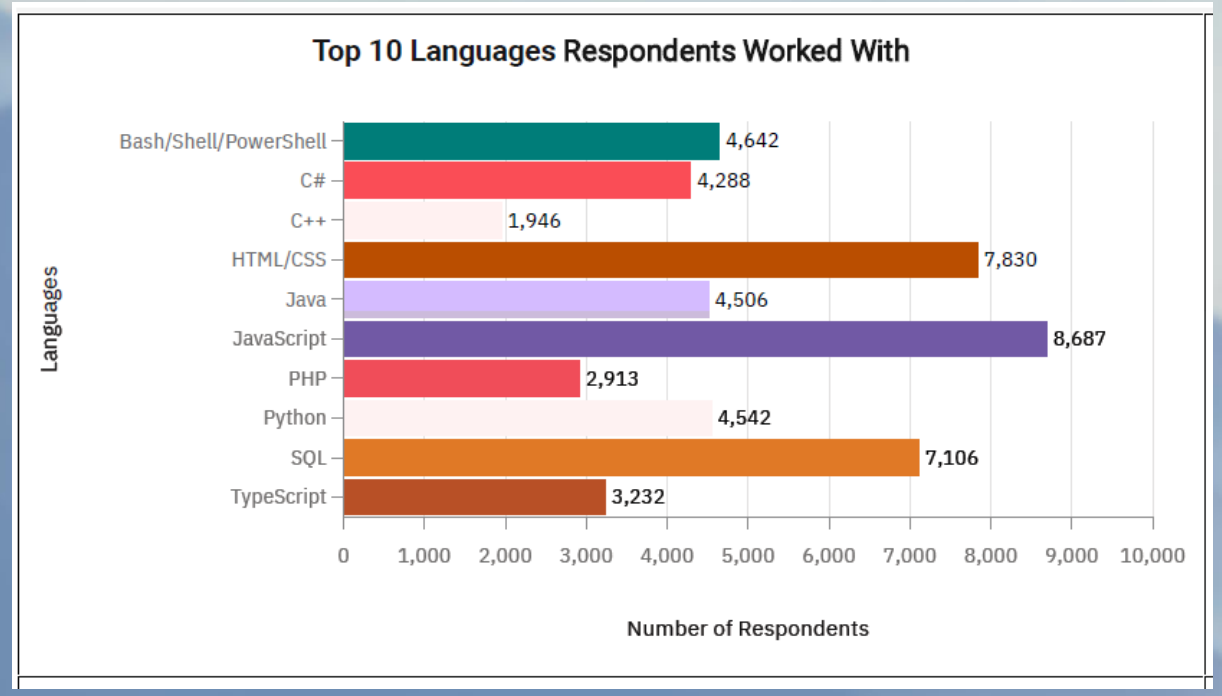


Figure 14: Top 10 Languages Respondents Worked With

Future Use: 2020

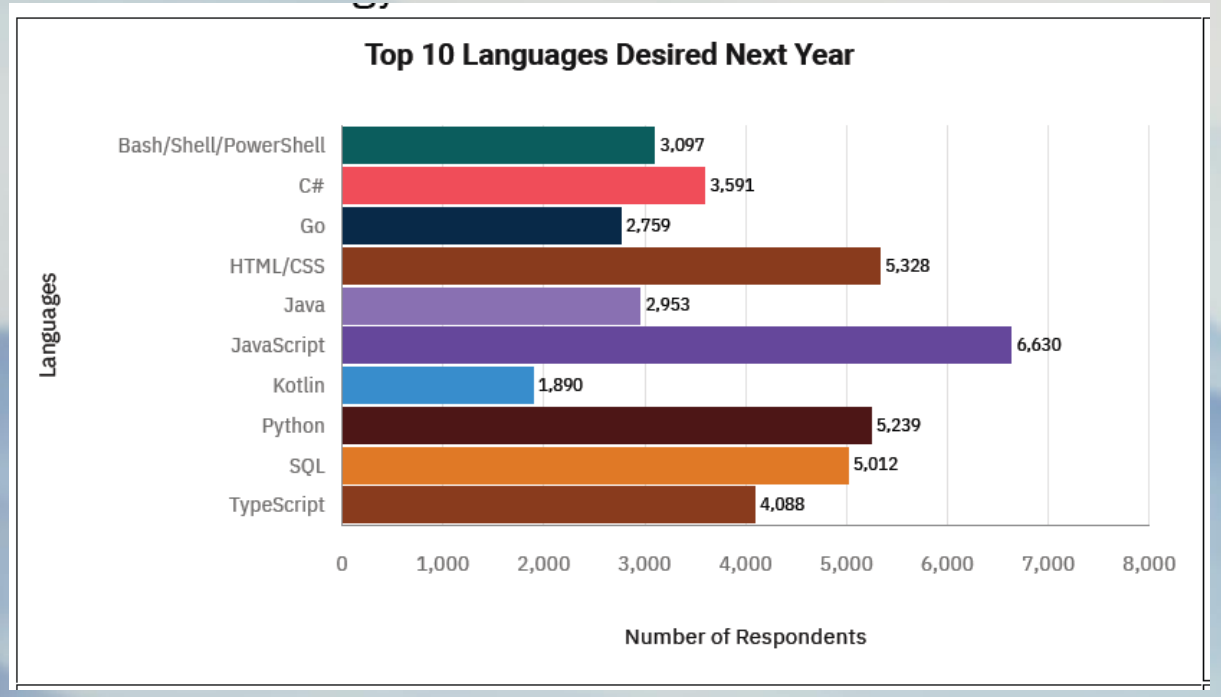


Figure 15: Top 10 Languages Desired Next Year

PROGRAMMING LANGUAGE TRENDS: FINDINGS & IMPLICATIONS

Findings

Top Languages Currently Used in 2019:

1. JavaScript
2. HTML/CSS
3. SQL

Top Languages Desired for 2020:

1. JavaScript
2. HTML/CSS
3. Python

2019: Top Popular Languages By Salary

1. Swift
2. C++
3. Python
4. JavaScript

Implications

- JavaScript, tops both years, also had the 3rd in salary.
- HTML/CSS was popular as 2nd in both years.
- SQL, 3rd in 2019, was overtaken by Python in 2020.
- Swift earned the top salary with C++ as 2nd place.
  - Notes: Neither of these are in the Top 10, and for 2020 Top 10, C++ came in last place.
    - A Respondent could possibly earn a higher salary if the program language had a lower usage.
    - Python may have been rising in popularity in 2020 because it had the 4th highest salary, or the reverse could be true. (Appendix A)

DATABASE TRENDS

Current Use: 2019

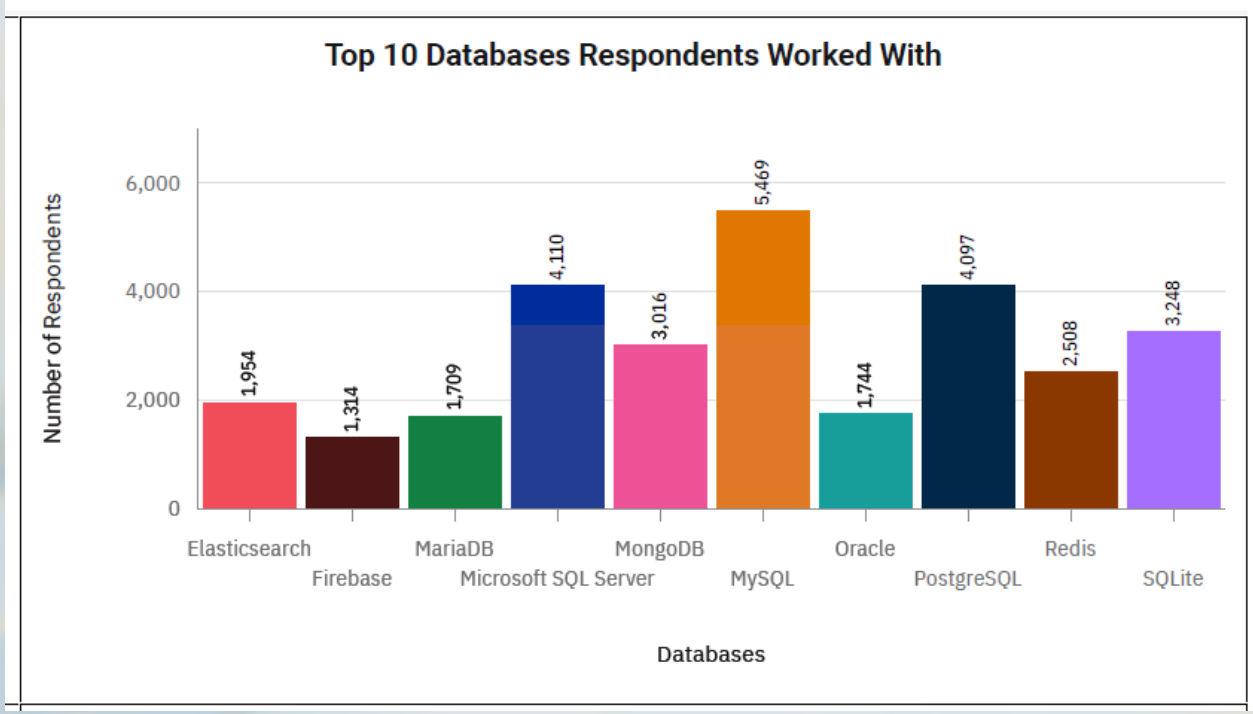


Figure 16: Top 10 Databases Respondents Worked With

Future Use: 2020

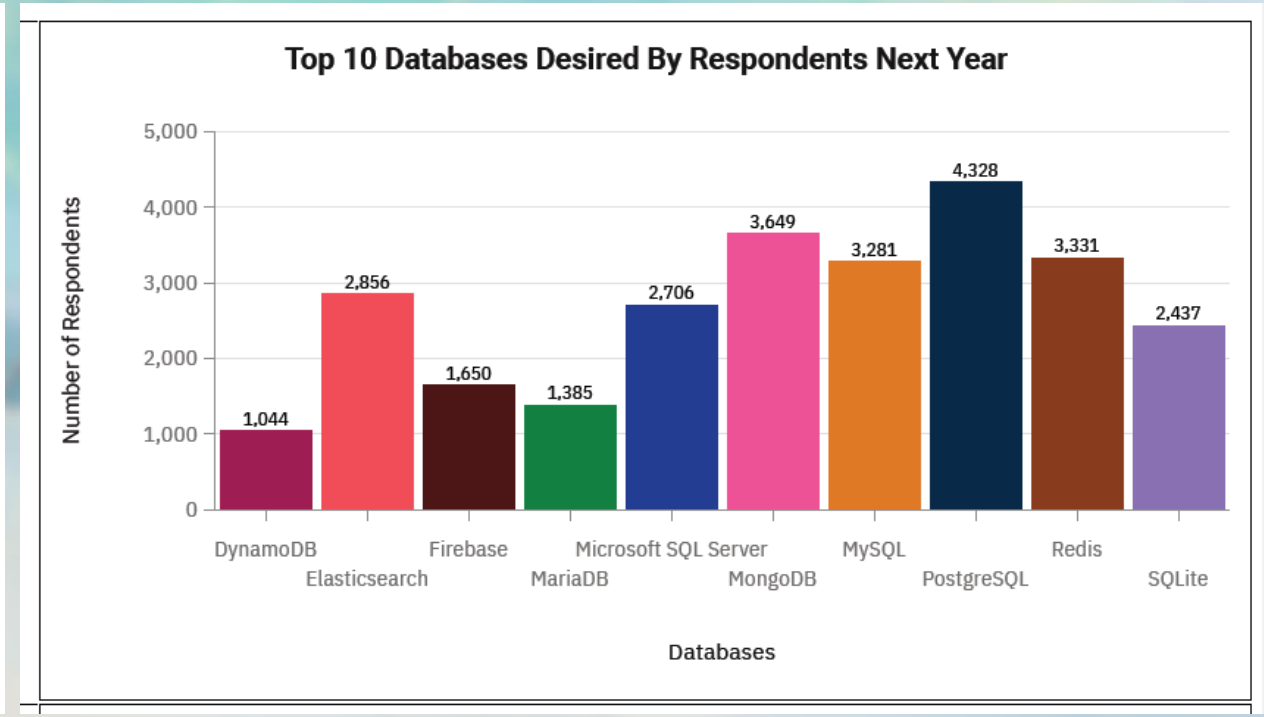


Figure 17: Top 10 Databases Desired By Respondents Next Year

DATABASE TRENDS:
FINDINGS & IMPLICATIONS

Findings

Top Languages Currently Used in 2019:

1. MySQL
2. Microsoft SQL Server
3. PostgreSQL

Top Languages Desired for 2020:

1.PostgreSQL

2.MongoDB

3.Redis

Implications

- In 2019, all Top 3 Databases were relational database management systems (RDBMS), using Structured Query Language (SQL).
- PostgreSQL, RDBMS 3rd for 2019, rose high in popularity to the Top Database for 2020.
- They also desired 2 different NoSQL Databases (MongoDB 2nd, Redis 3rd) for 2020.
- While RDBMSs were popular with 3 Top spots in 2019, the trend could possibly be leaning toward using more flexible, open-sourced, document-centered Databases, as in 2020.

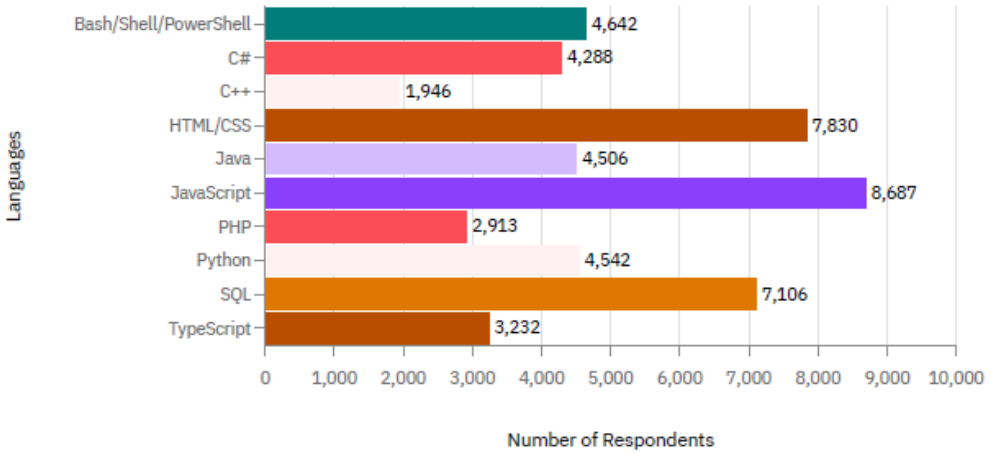
DASHBOARD: GitHub Link

https://github.com/adye7/IBM-Capstone-Project-Dye/blob/7211a0261875bfaf6e5e4f3ea910f1deac78f0f7/IBM%20Capstone%20Project-Dye.pdf

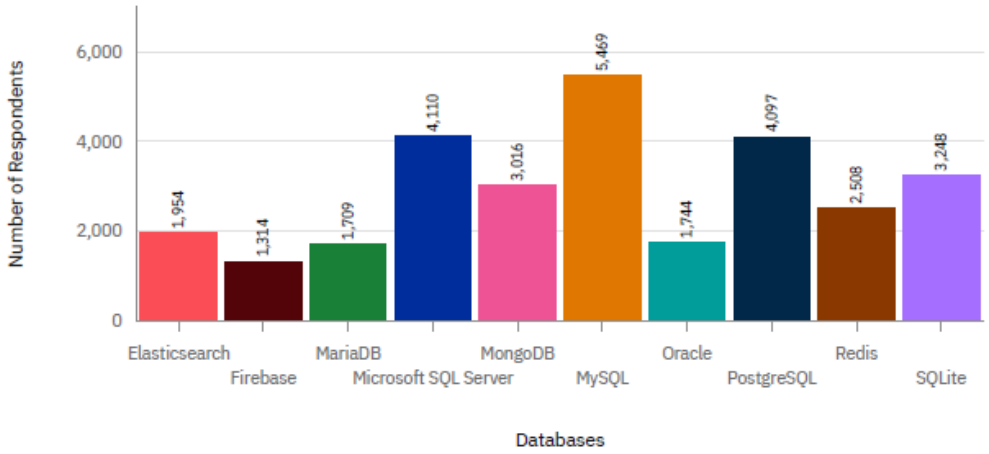
DASHBOARD 1: Current Technology Usage

Current Technology Usage

Top 10 Languages Respondents Worked With

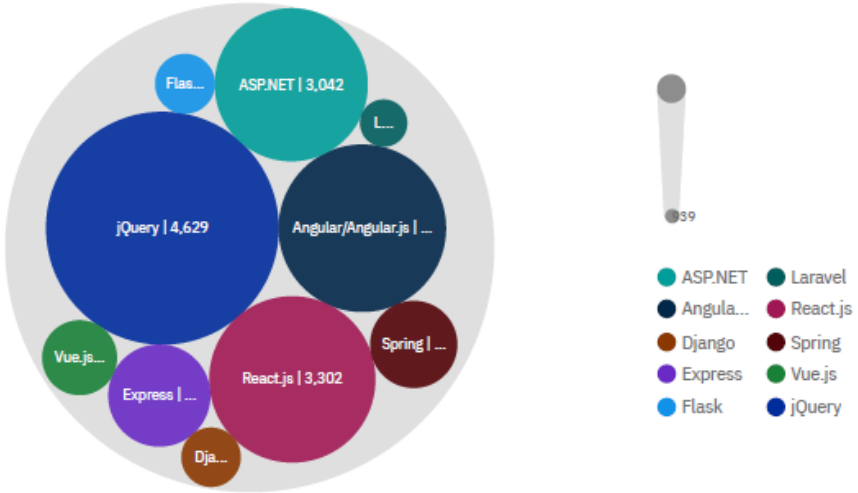


Top 10 Databases Respondents Worked With



Platform Respondents Worked With

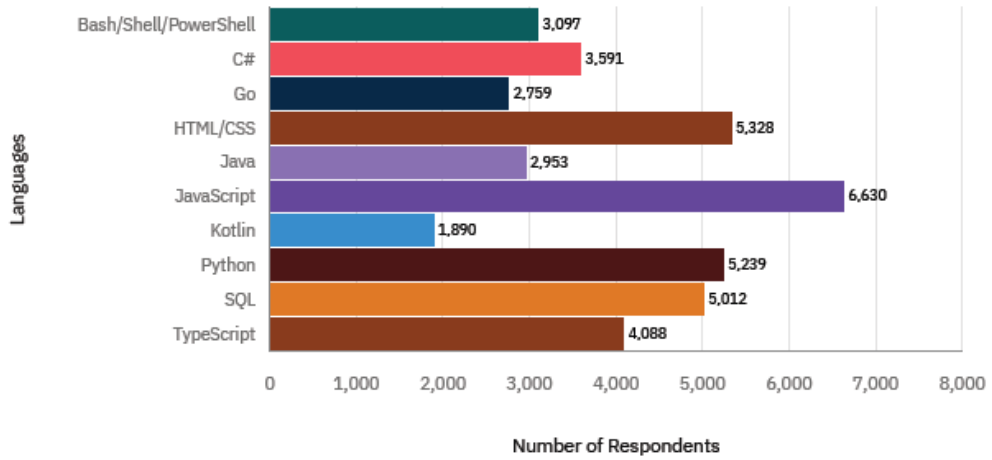
Top 10 Web Frameworks Respondents Worked With



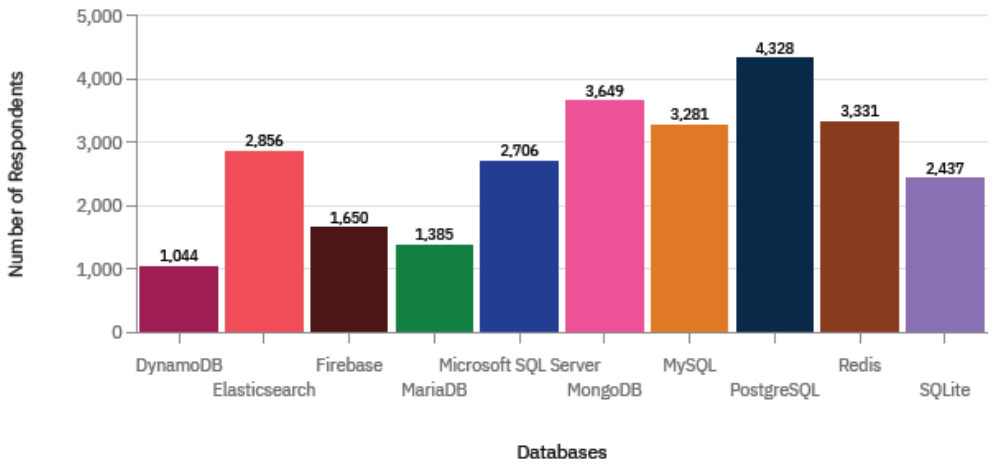
DASHBOARD 2: Future Technology Trends

Future Technology Trend

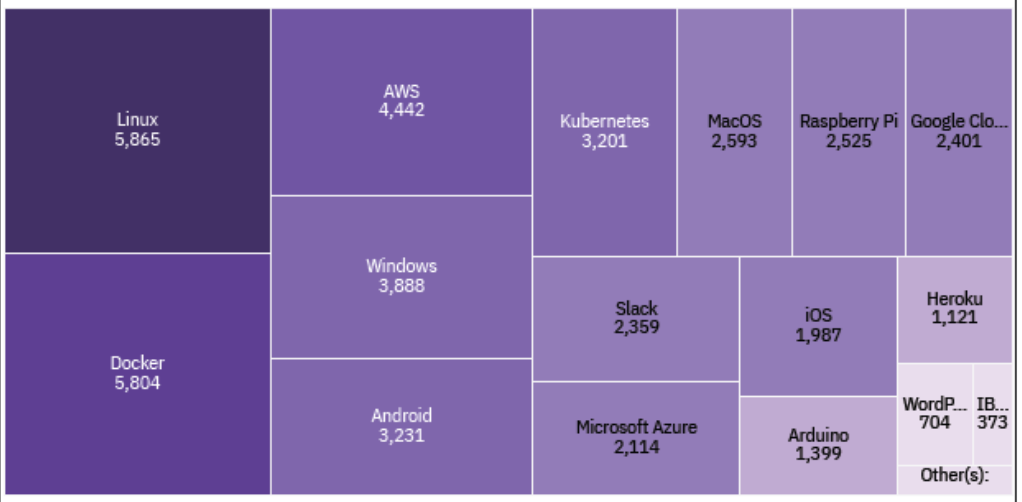
Top 10 Languages Desired Next Year



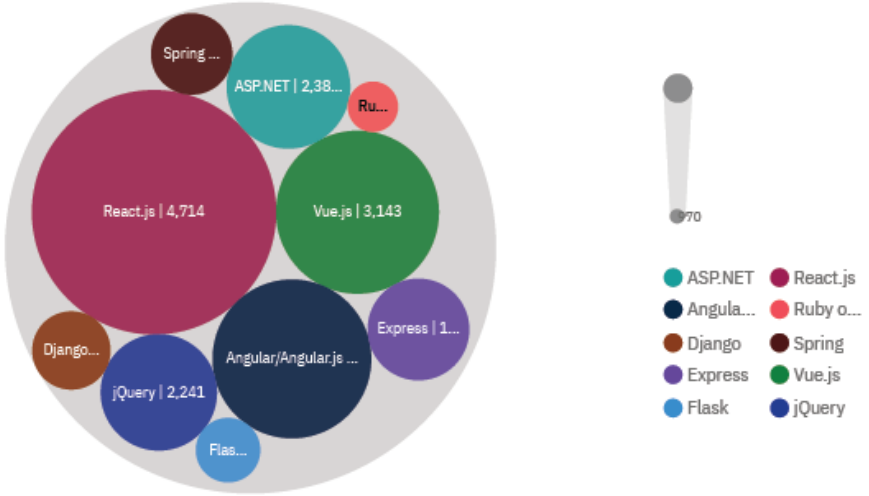
Top 10 Databases Desired By Respondents Next Year



Platforms Respondents Desired Next Year



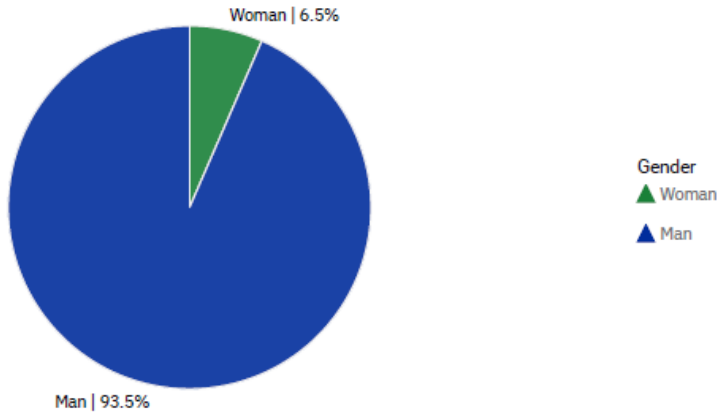
Top 10 WebFrames Desired By Respondents Next Year



DASHBOARD 3: Respondent Demographic Trends

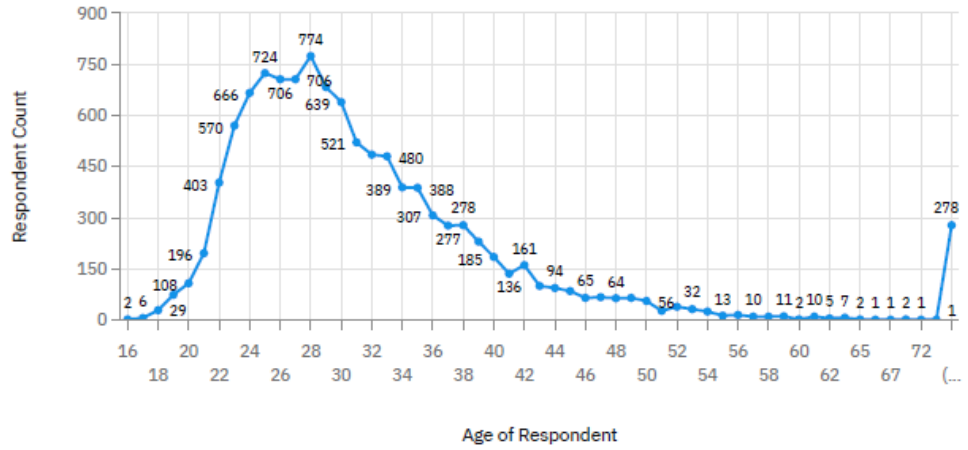
Demographics

Respondent Classified by Gender

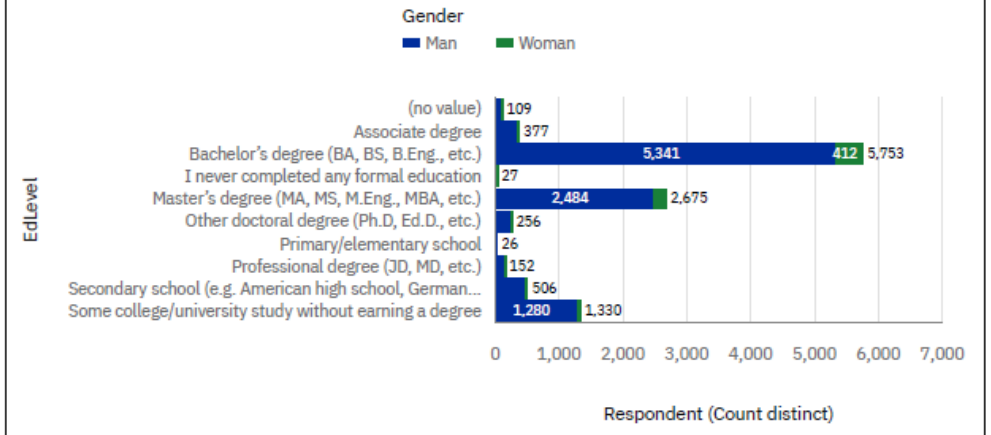


Respondent Count for Countries

Respondent Count By Age



Respondent Count By Gender, Classified By Formal Education Level



OVERALL FINDINGS: Technologies

Languages:

*Significant relationship Current/Future Use

- JavaScript and HTML/CSS were 1st and 2nd choice for 2019 and 2020. For 3rd place SQL replaced Python for 2020.

Databases:

*Significant relationship Current/Future Use

- 2019 had the Top 3 choices taken by RDBMSs, led by MySQL. 2020 PostgreSQL went from 3rd to 1st, but the next two choices were NoSQL Databases, MongoDB & Redis.

Platforms:

No Significance between Current/Future Use

- In 2019, Windows led the platforms, but for 2020, Linux and Docker went from 2nd/3rd to 1st/2nd while Windows dropped to 4th.

WebFrames:

*Significant relationship Current/Future Use

- React.js went 2nd to 1st and jQuery went from 1st to 5th.

OVERALL FINDINGS: Demographics

Age:

- There was no significant difference between Respondents, with the Top 5 ranging between 24:28 years old.

Gender:

- This was the most significant finding, with 92% of Respondents identifying as a Man, with Women at trailing far behind at 7%. The Pie Chart on Dashboard 3 is a striking illustration of this.

Country:

- The United States had by far the most Respondents at 27%

Education Level (EdLevel):

- Most of the Respondents had either a Bachelor's Degree (51%), or a Master's Degree (24%).

OVERALL IMPLICATIONS : Technology

Languages

For 3rd , it appears that SQL may have lost some of the Respondents to Python, which may account for their switch for 3rd place in 2020.

Databases

While the SQL RDBMS were popular in 2019 and held the Top 3, in 2020 the 2nd/3rd choices were NoSQL, Open-Sourced, Document-Centered Databases. This could possibly be seen as a future trend, but further data and analysis is needed.

Technology

Platforms

In 2019, Windows was 1st, but for 2020; but in 2020, Linux and Docker went from 2nd/3rd to 1st/2nd. Windows dropped to 4th. Further research is needed to see any relationship or long-term trends from this.

WebFrames

React.js went 2nd to 1st, and jQuery was downgraded from 1st to 5th. Further research could also show future trends.

DISCUSSION

This study analyzed a 2019 subset of survey data (n = 11552) from Stack Overflow, a survey website for computer programmers. Two main types of categorical data were addressed: Technologies and Demographics.

The Technologies category was administered Chi-Squared Independence Tests and found significant relationships between current and future technology use in three of the four subcategories:

- Languages ($p=4.52E-282$), Databases ($p=0.049$), and WebFrames ($p=0.003$)
  - These significant results validate this study's alternate hypothesis that the current and future technologies have a significant relationship difference.
- There was no significant relationship in the fourth, Platforms ($p=0.075$).
  - This affirmed the Null hypothesis of no significant relationship, but if there were a larger dataset analyzed, the small ($p=0.075$ difference would might possibly be within the significant range.

DISCUSSION

Language

- JavaScript, the for 2019/2020, was consistent with the past research introduced and with the study's alternate hypothesis.
- Also, HTML/CSS, at the top 2 spots for each year, was the same as the indices. SQL (2019) 3rd but then switched to Python (2020).

Database

- MySQL was the Top Database (2019), and the other two were RDBMSs as well.
- PostgreSQL held the Top spot (2020)
- The 2nd and 3rd were instead NoSQL databases: MongoDB and Redis.
- This could hint at a moving trend toward away from SQL databases to the more document led, open-sourced NoSQL databases.

Demographics

- The most typical Respondent was Male, 28 years old, from the United States, and had a Bachelor Degree. This confirmed the male hypothesis but was slightly older than expected.
- The most observable differences were 1st/2nd: Male (92%)/Female (7%); United States (27%)/India (7%); Education Level (EdLevel) Bachelor's Degree (51%)/Master's Degree (24%).
- The highest Age ranged from 24:28 years old, with 28 as the highest. This was a bit older than the past literature and the hypothesis surmised at younger 20s.

LIMITATIONS:

1) Small data sample (n=9703)

2) One more factor that was not alluded to in this study was the world events between 2019-2020. This was when the globe had the Coronavirus pandemic, and many people were at home and were using computers more so than before in general. Jaumotte et. al., had one analysis of this that found a 6% increase in more advanced countries (2023). This may have come to play in many research factors enough to possibly skew the data and findings. This will need to be taken into consideration when others study from this time period and further research is needed on this.

3) Last problem is that repeated, this study would have taken more time to use counts of the categorical data to run some more conclusive statistical tests on the data, but this study had a limited time to complete. However, the results are still interesting and would be an aid to anyone looking for this information.

CONCLUSION

- Three types of technology were found to have a significant difference in their relationship between current and future use in the survey data.
- The respondents were most frequently male, 28 years of age, from the USA, and had a Bachelor's degree.
- JavaScript (1st) and HTML/CSS (2nd) were found, mirroring the indices previous studies of technology use.
- Also, Age was found to be on the lower end of adulthood (24:28), which is a bit above the previous research study introduced regarding Age and technology.
- Gender was by far the highest significance result, even stronger than the previous study addressed. These results are very interesting and would be a helpful snapshot of anyone interested in the subject matter or if they are planning on studying the survey data as well.

REFERENCES

Botella, C., Rueda, S., López-Iñesta, E., & Marzal, P. (2019). Gender diversity in STEM disciplines: A multiple factor problem. *Entropy*, *21*(1), p. 30.

Cass, S. (2019). The Top Programming Languages (2019). Python remains the big kahuna, but specialist languages hold their own. *IEEE Spectrum*, Accessed on 12/05/2024: https://spectrum.ieee.org/the-top-programming-languages-2019.

Cass, S. (2020). Top Programming Languages (2020). Python rules the roost, but Cobol gets a pandemic bump. *IEEE Spectrum,* Accessed on 12/05/2024: https://spectrum.ieee.org/top-programming-language-2020.

Jansen, Paul. (2024), Very Long Term History. *TIOBE Index for November 2024*. Accessed on 12/05/2024: https://www.tiobe.com/tiobe-index.

Jaumotte, F., Oikonomou, M., Pizzinelli, C., & Tavares, M. M. (2023). How Pandemic Accelerated Digital Transformation in Advanced Economies. *International Monetary Fund*, 21.

O'Grady, S. (2019, March 20). The RedMonk Programming Language Rankings: January 2019. *RedMonk*. Accessed on 12/05/2024: https://redmonk.com/sogrady/2019/03/20/language-rankings-1-19/.

O'Grady, S. (2020, February 28). The RedMonk Programming Language Rankings: January 2020. *RedMonk*. Accessed on 12/05/2024: https://redmonk.com/sogrady/2020/02/28/language-rankings-1-20/.

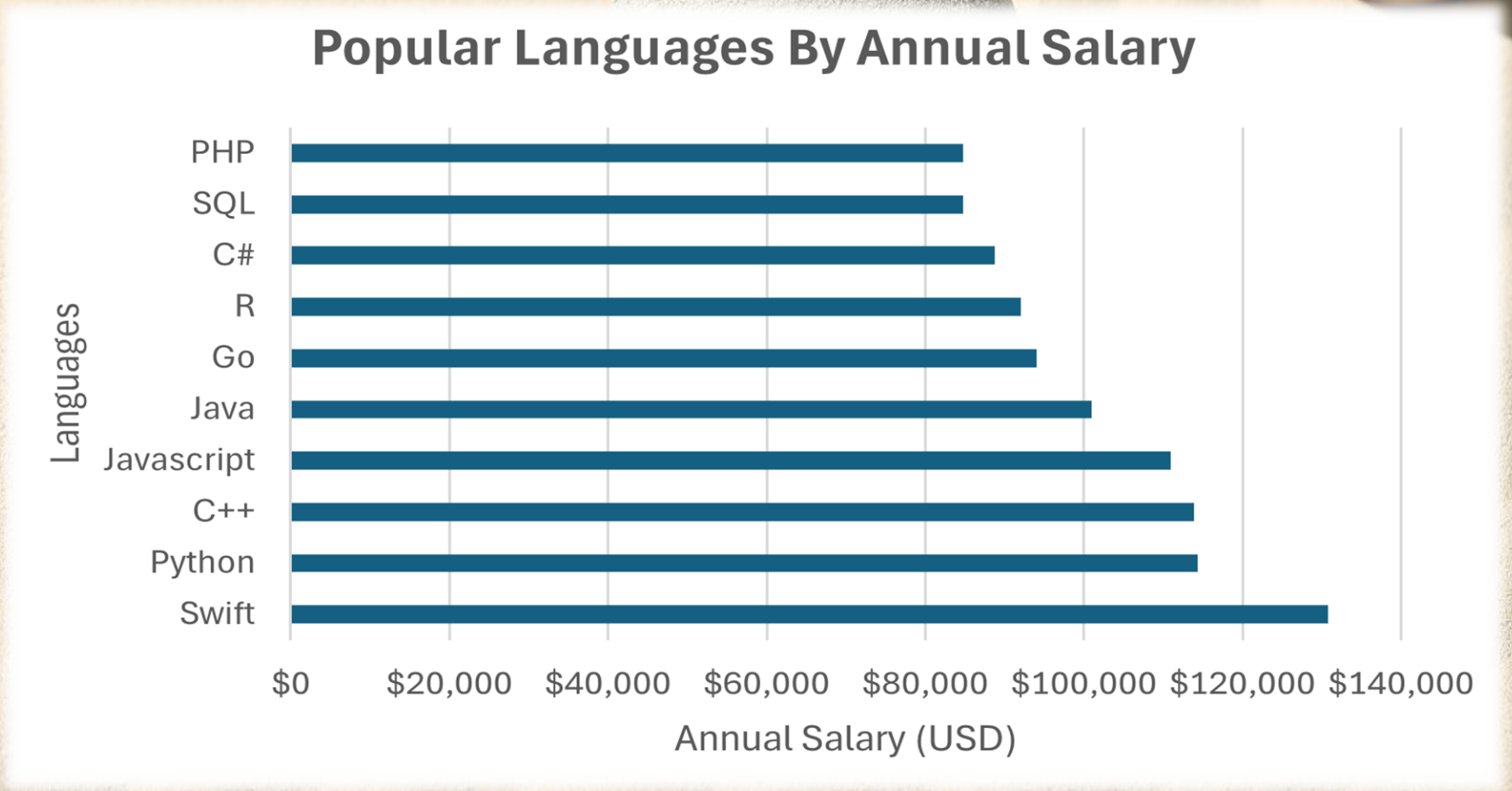
PYPL PopularitY of Programming Language Index. (2019). *PYPL*. Accessed from: https://pypl.github.io/PYPL.html on 12/10/2024.

PYPL PopularitY of Programming Language Index. (2020). *PYPL*. Accessed from: https://pypl.github.io/PYPL.html on 12/10/2024.

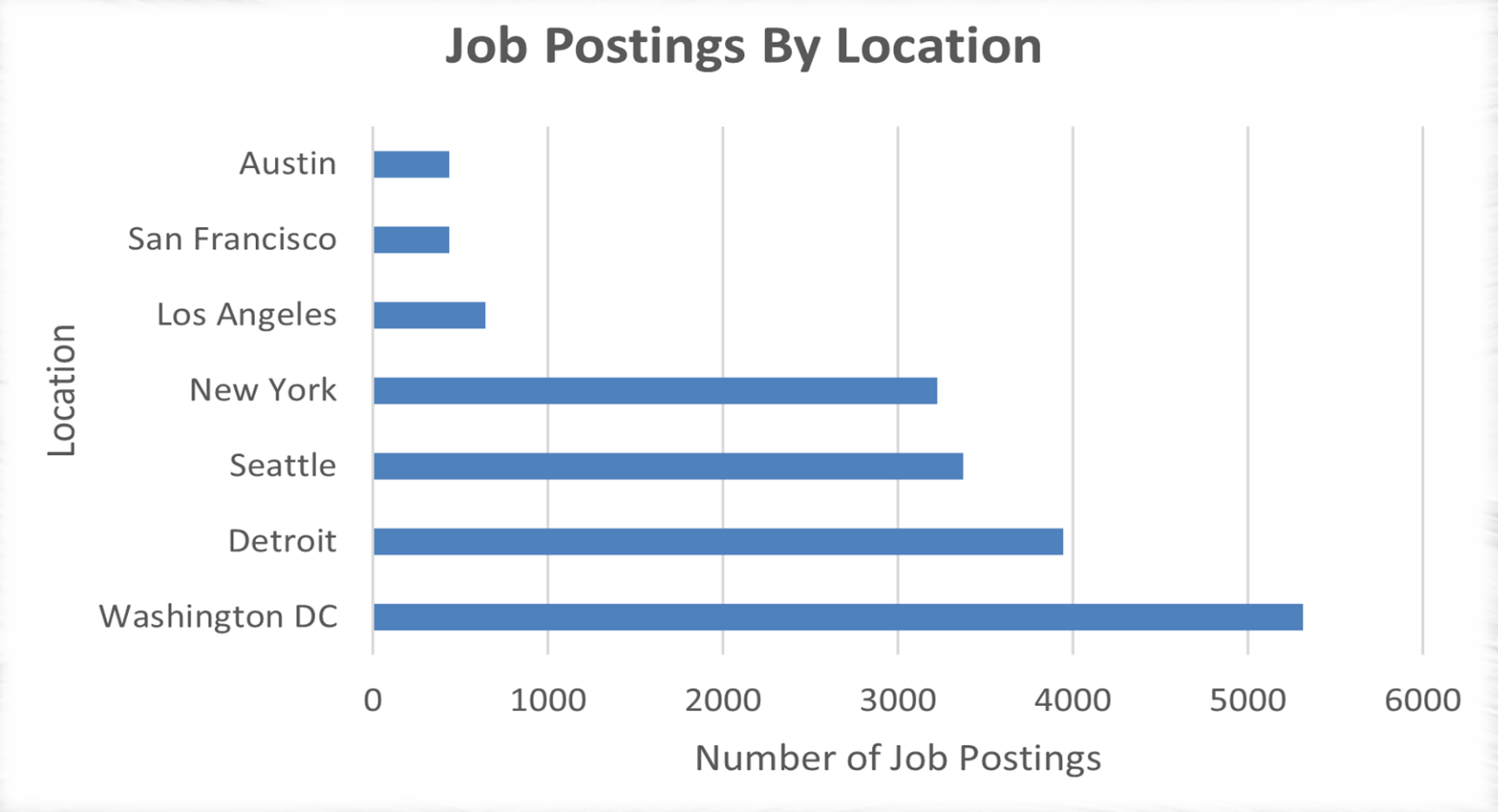
Vogels, E. A., & Anderson, M. (2019, October 9). Americans and digital knowledge in 2019. *Pew Research Center*. Accessed on 12/05/2024: https://www.pewresearch.org/internet/2019/10/09/americans-and-digital-knowledge/.

Appendix

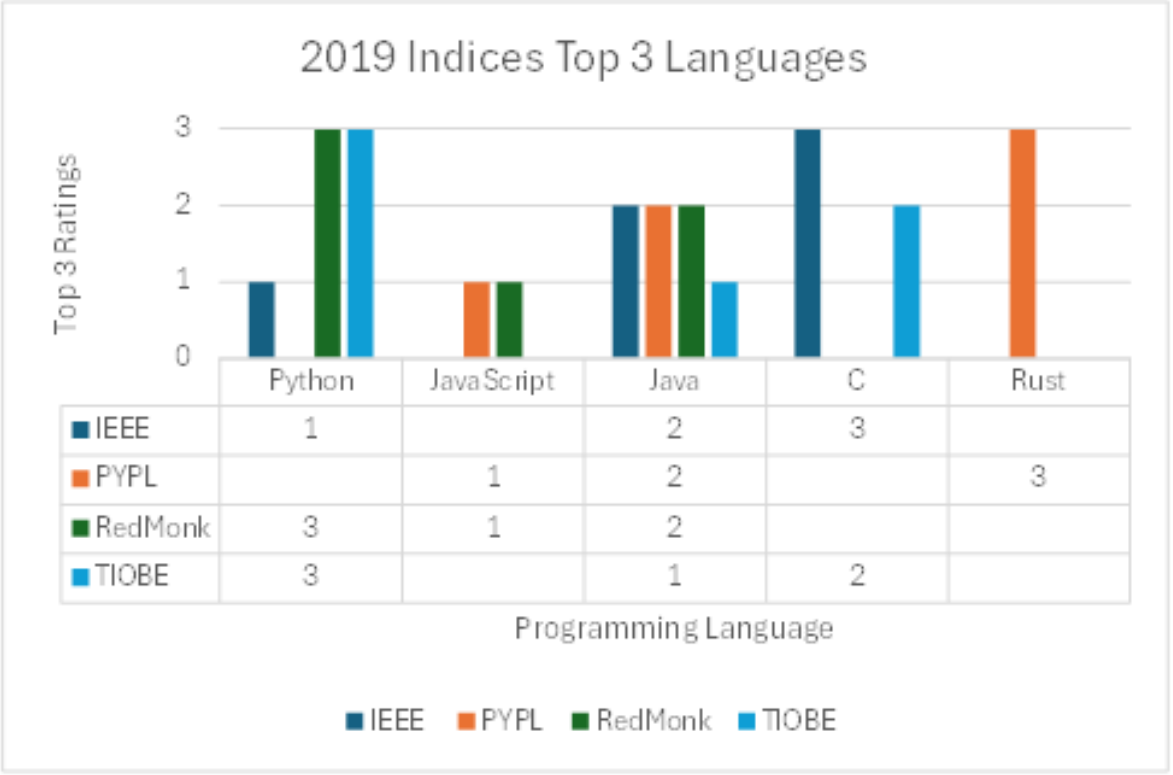
Appendix A: Popular Languages By Annual Salary

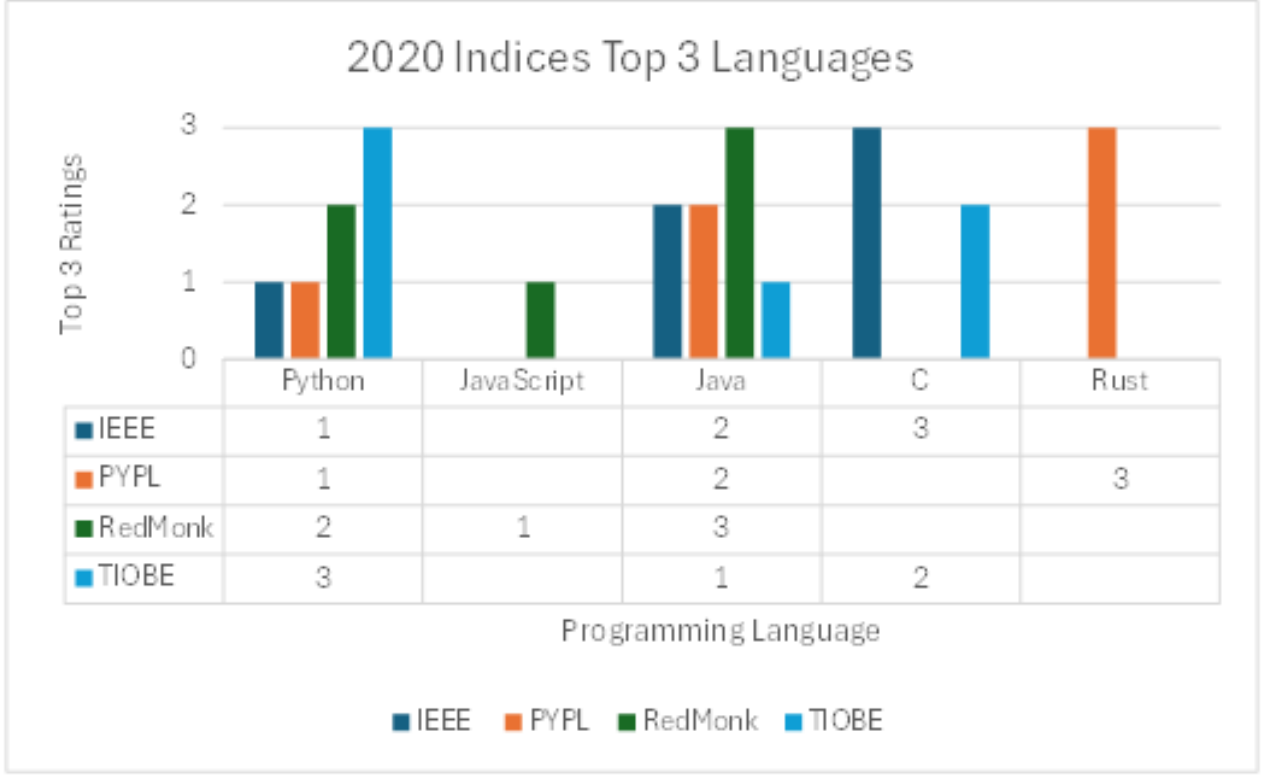


Appendix B: Job Postings By Location

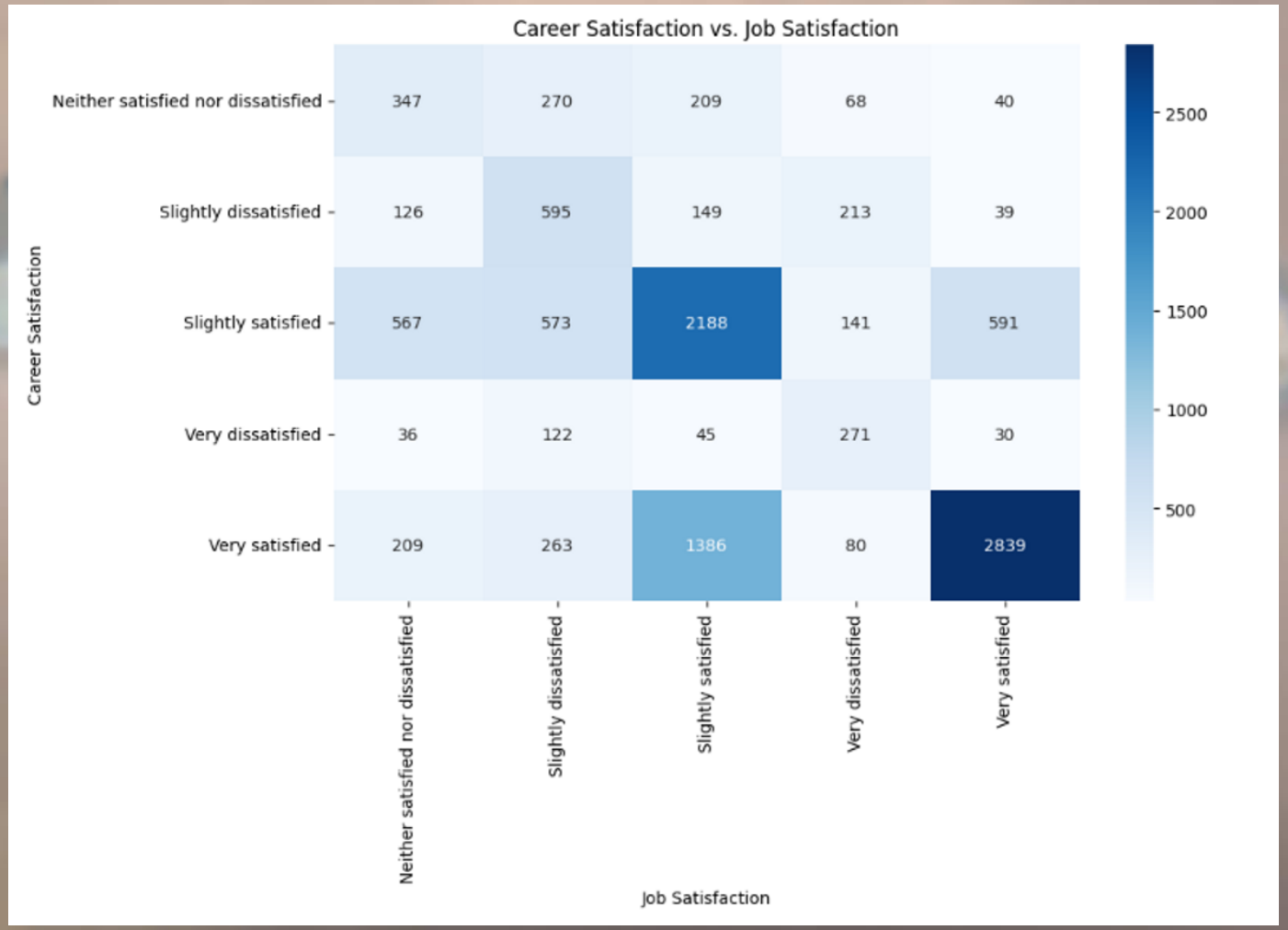


Appendix C: 4 Online Indices Rankings

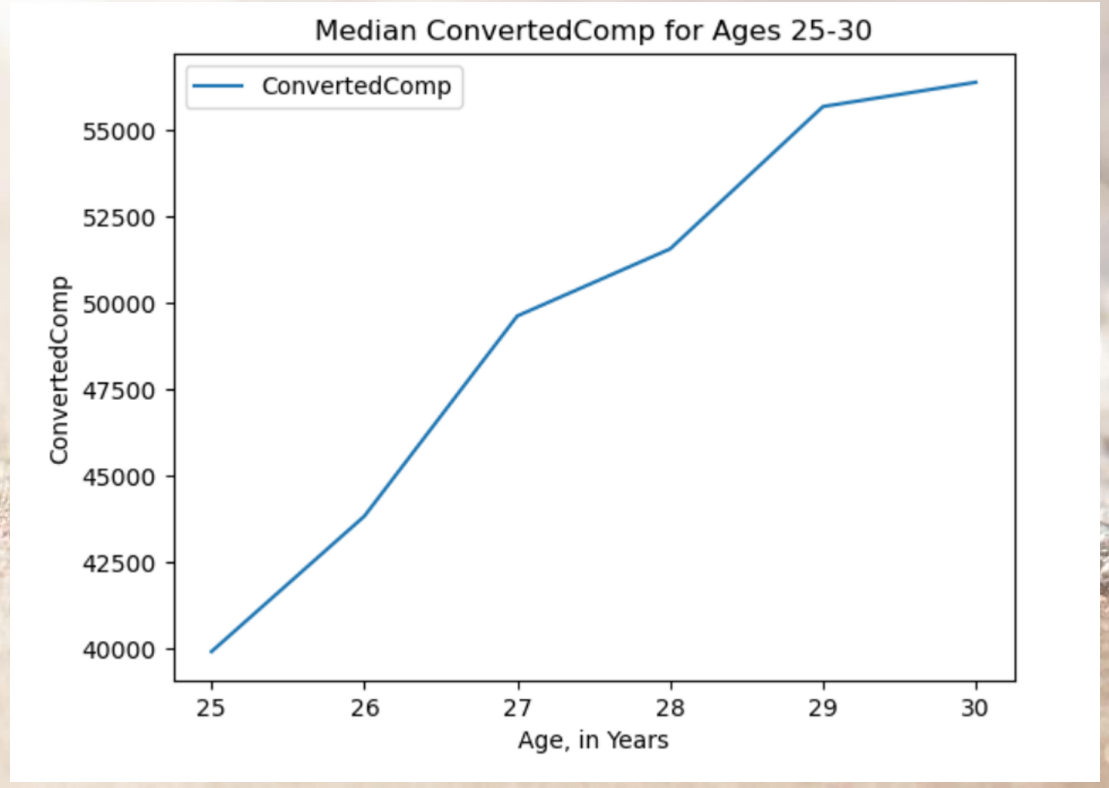


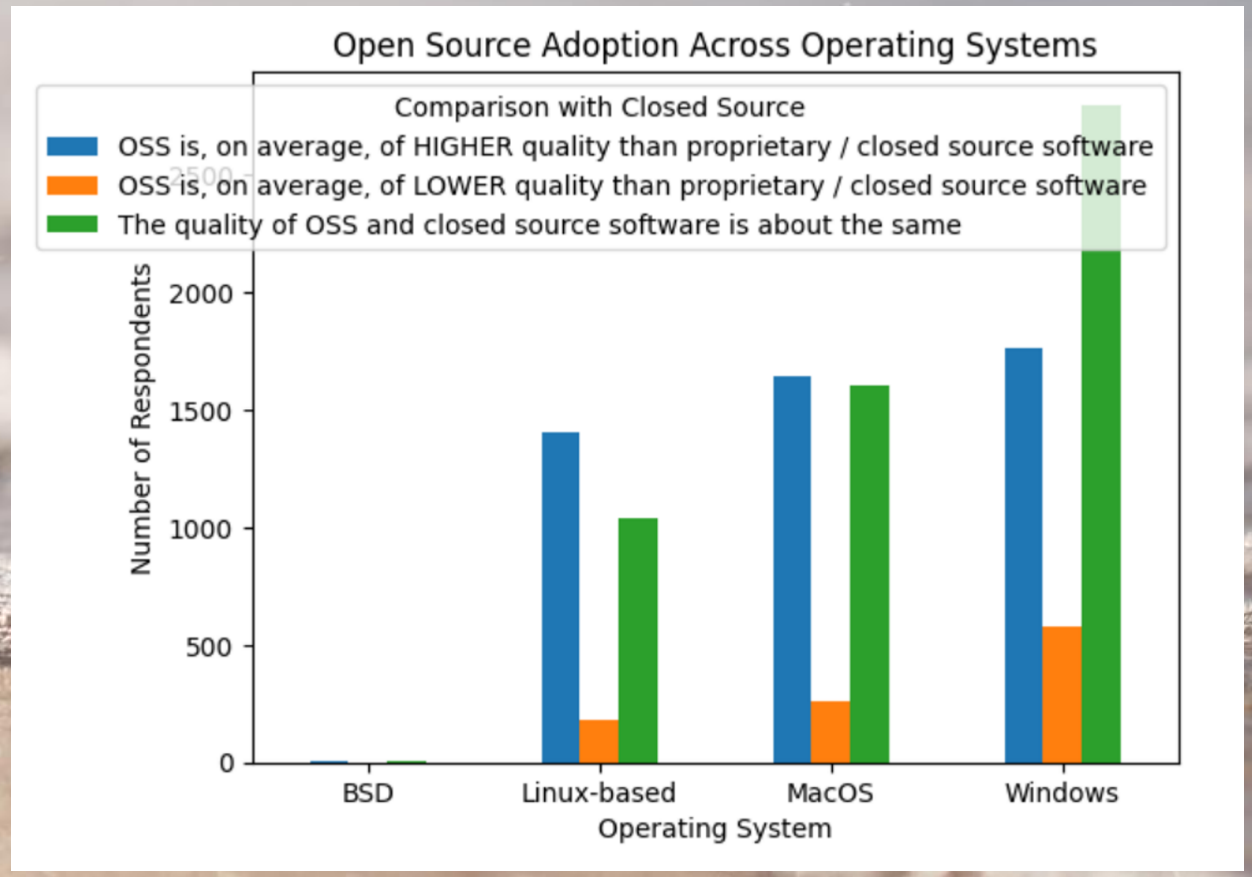


Appendix D: Career Satisfaction vs. Job Satisfaction

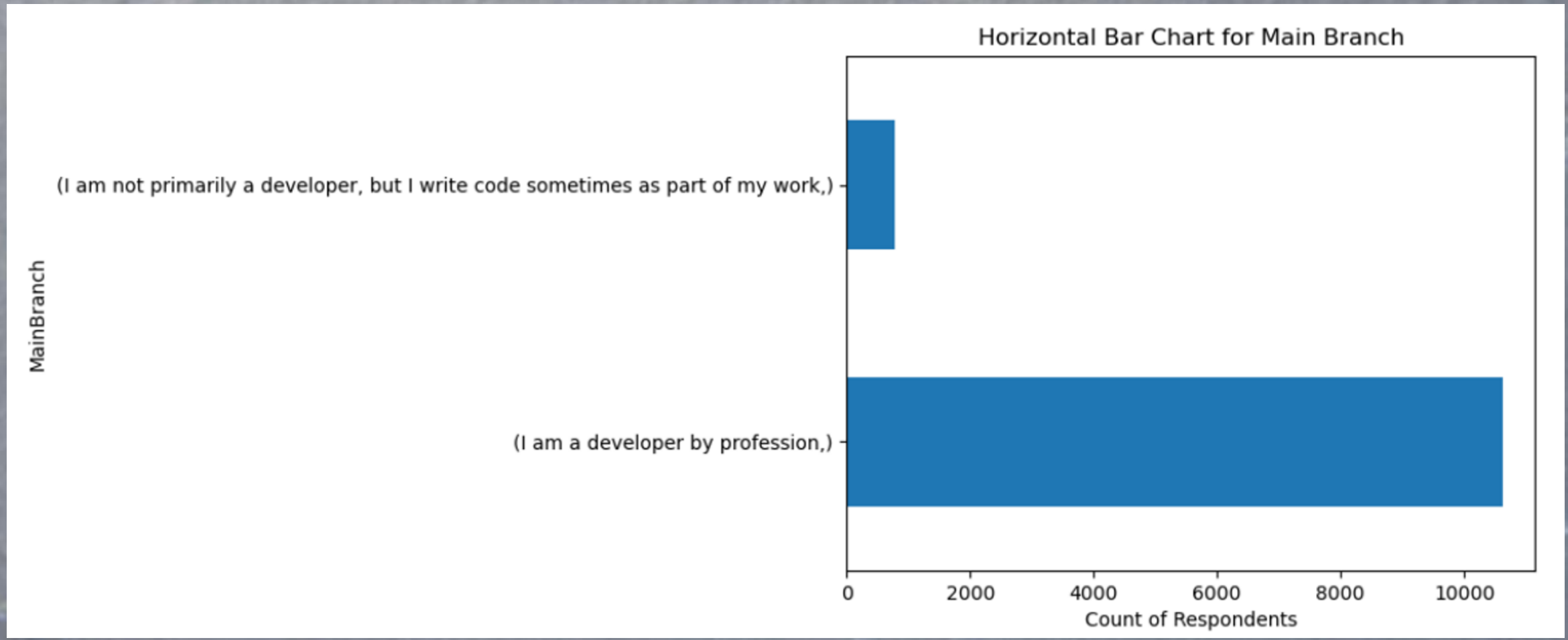


Appendix E: Median ConvertedComp for Ages 25-30/ Open Source Adoption Across Operating Systems





Appendix F: Horizontal Bar Chart for Main Branch



Appendix G: Scatter Plot on Relationship Between Age & WorkWkHrs/ Median WorkWeek & CodeRev Hours in 30~35 Age Group