IBM Data Analyst Capstone Project

Josey Tan

23 May 2021

OUTLINE

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

EXECUTIVE SUMMARY

- This report is about the Stack Overflow Developer Survey 2019, Github job posting and IBM website for current trends, job vacancies and annual salary based on popular languages.

Popular Programming Language	
Current Trend	**Next Year**
Javascript	Javascript
HTML/CSS	Python
SQL	HTML

Popular Database Language	
Current Trend	**Next Year**
MySQL	PostgreSQL
PostgreSQL	MongoDB
MSFT SQL Server	Redia

- The top 3 job vacacies based on programming languages are C, Javascript and Python.
- The 3 best paying languages with the highest average annual salary are Swift, Python and C++.
- Based on findings, web pages interactive and Python are the most desirable and well payed language for the seeable future.

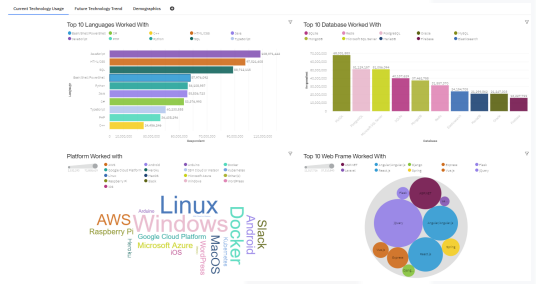
INTRODUCTION

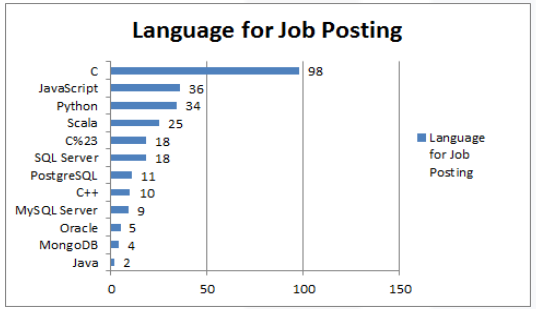
- This report is about the survey of respondent who worked as a developer as profession or those that are not primarily a developer, but I write code sometimes as part of my work.
- The target audience for this report is for professional, companies and recruiter who wants to know the current trend in their field of work.
- Able to gain insight on demand of languages for recent job posting and average annual salary range
- Able to gain insight on demographics of respodent
  - Gender in the workforce
  - Education level

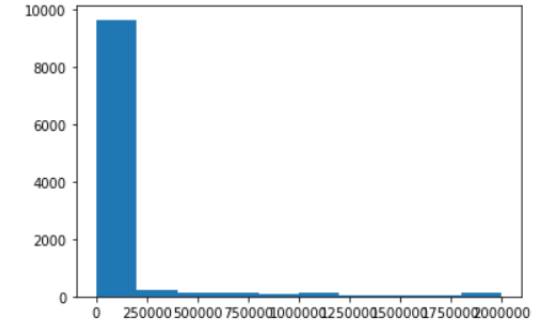
METHODOLOGY

- The data source for the online survey is from Stack Overflow Developer Survey 2019 data where the responses are compiled into csv files.
- The data source for job posting is from GitHub Jobs API where API get request call was used to obtain the result.
- The data source for average annual salary for the languages is from IBM website where web scrapping with get request was used to obtain result
- EDA and data visualization was done on mulitple platform:
  - IMB Cognos Dashboard
  - Jupyter Notebook
  - Excel

RESULTS



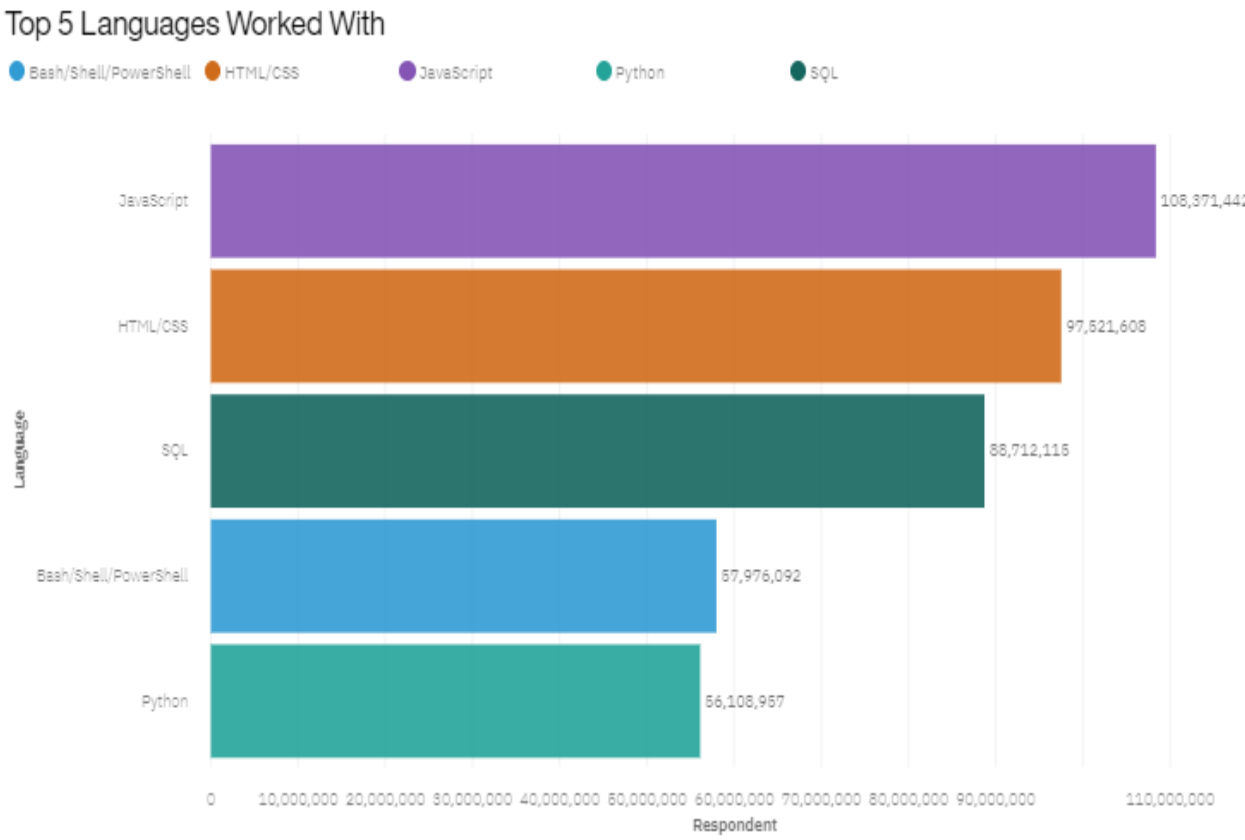




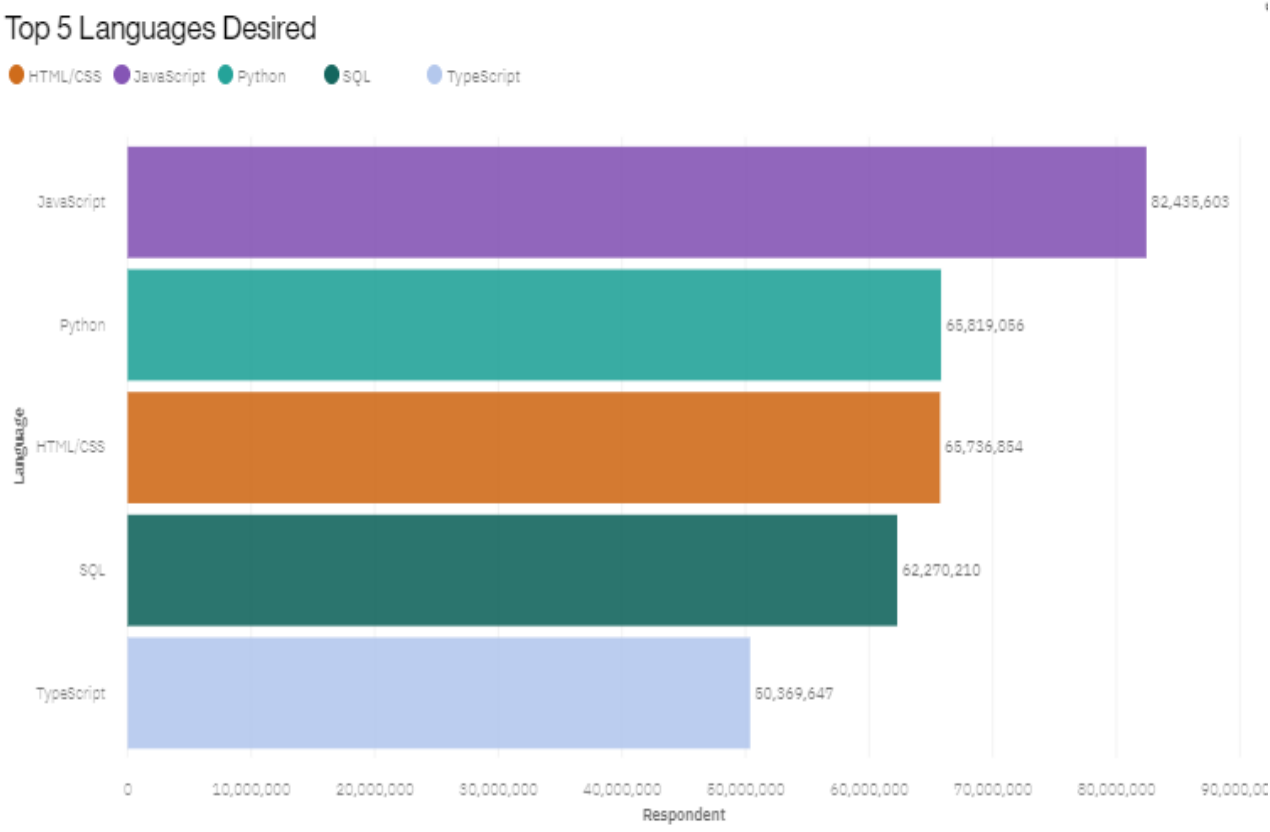
- From the Cognos Dashboard, we visualized the current and desired trend from professional in their field of work.
- From the Excel, we visualized the number of current job posting on Github and average annual salary range by popular languages.
- From the JupyterNotebook, exploratory data analysis was done to find out the characteristics of the respondents in the survey and results can be visualized too.

PROGRAMMING LANGUAGE TRENDS

Current Year



Next Year



PROGRAMMING LANGUAGE TRENDS - FINDINGS & IMPLICATIONS

Findings

Current Year

1. Javascript
2. HTML/CSS
3. SQL

Next Year

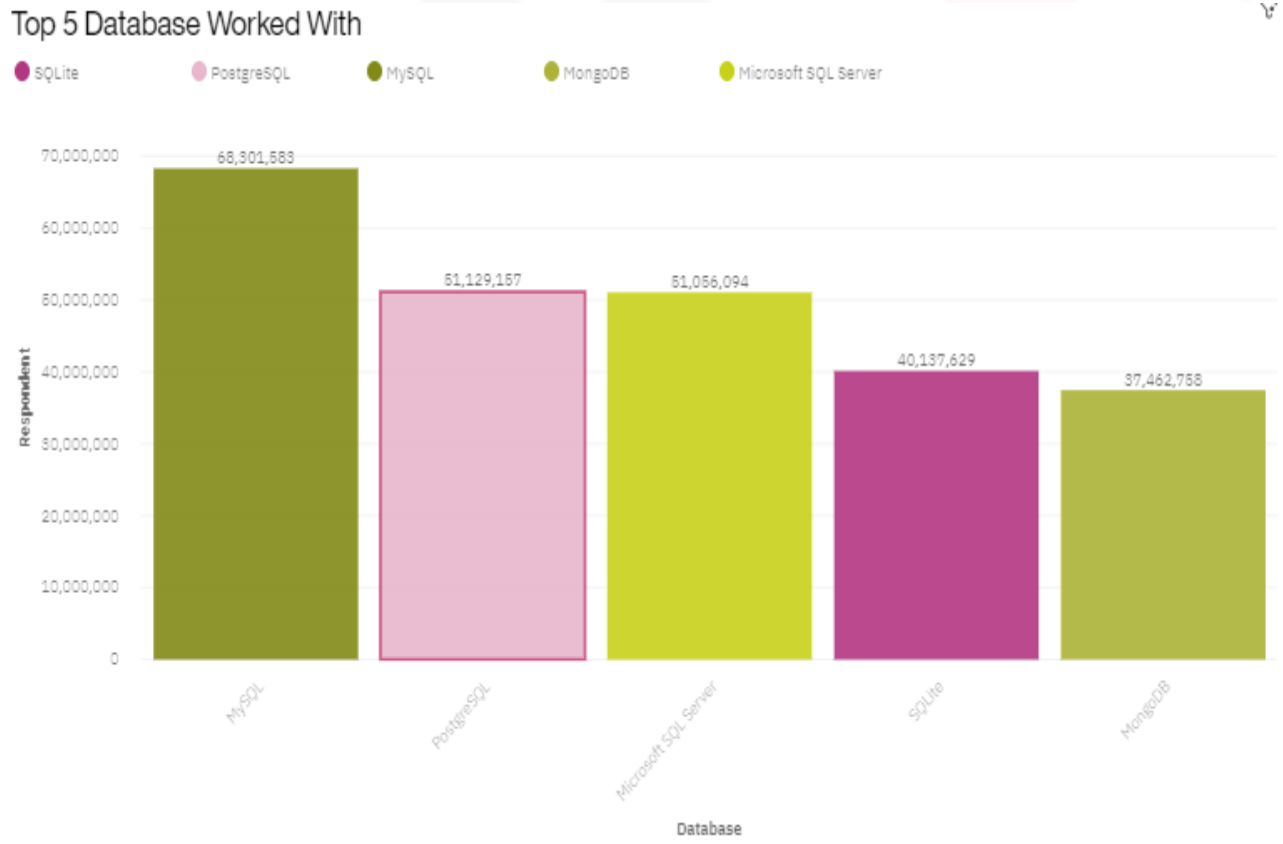
1. Javascript
2. Python
3. HTML/CSS

Implications

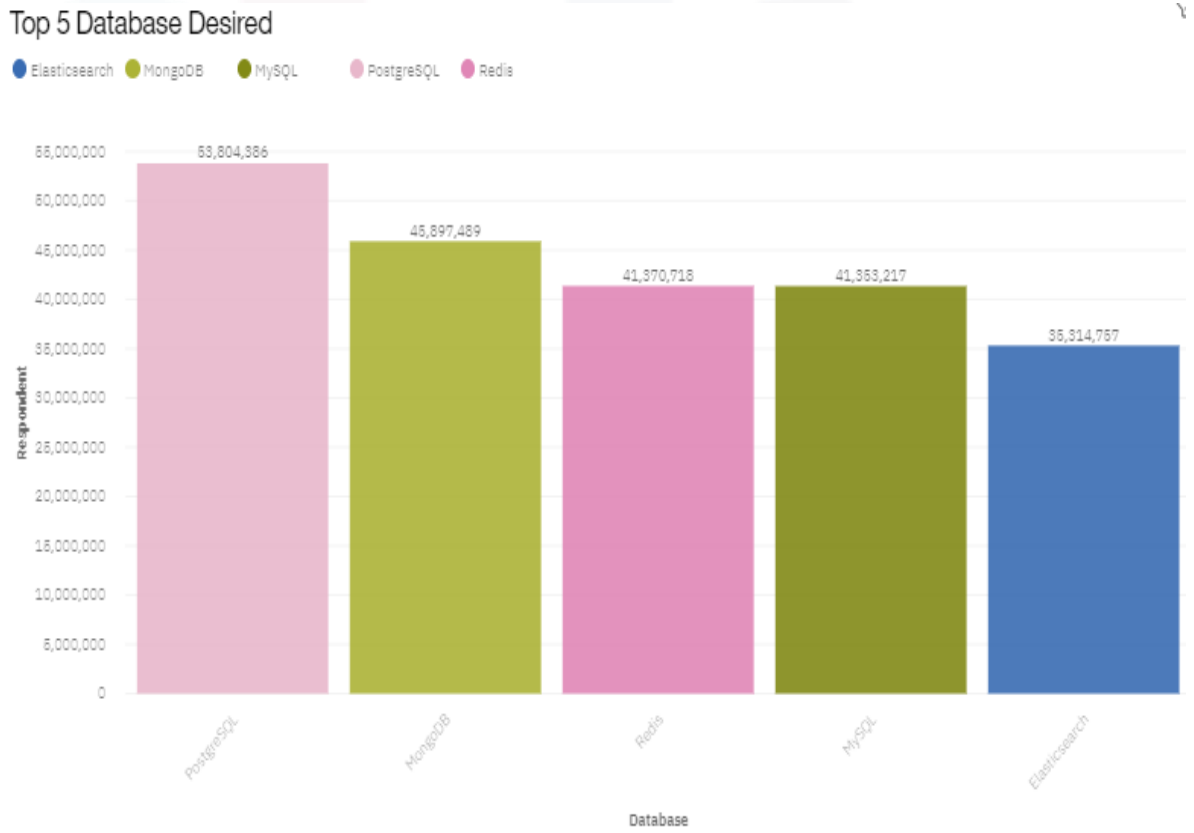
- Web pages interactive languages are the most important.
- Javascript and HTML/CSS are still the most desired language currently and in the furture.
- Python is the upcoming trend for programming languages

DATABASE TRENDS

Current Year



Next Year



DATABASE TRENDS - FINDINGS & IMPLICATIONS

Findings

Current Year

1. MySQL
2. PostgreSQL
3. Microsoft SQL Server

Next Year

1. PostgreSQL
2. MongoDB
3. Redis

Implications

- SQL remains as the most desired query language and database
- However, SQL databases face stiff competition from upcoming database like MongoDB and Redis.

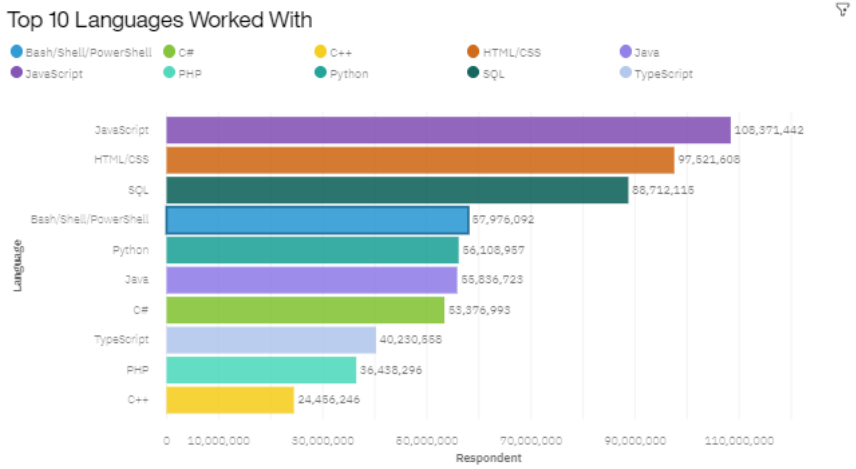
DASHBOARD

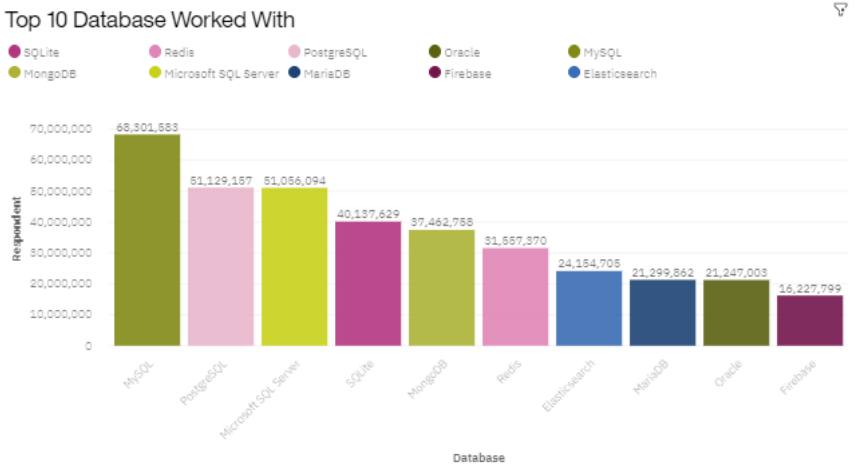
The permanent link of the read-only view of the Cognos dashboard:

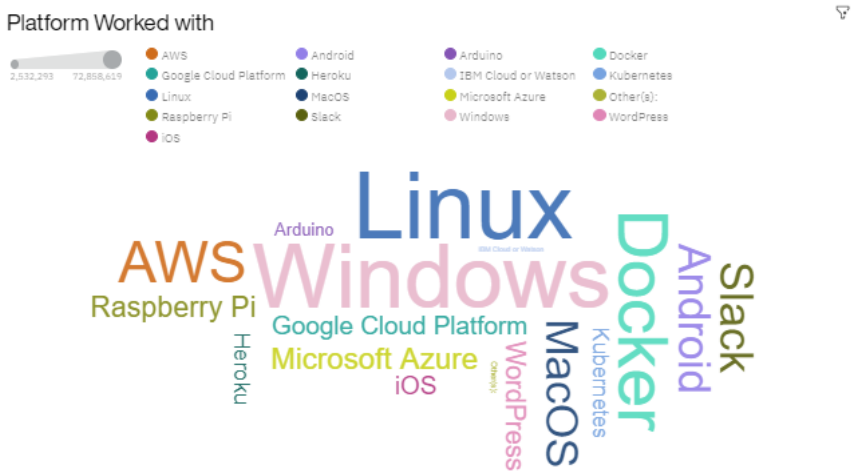
https://dataplatform.cloud.ibm.com/dashboards/be9b9bcd-34b4-4614-9a2e-63173e2f7d8c/view/4c34f61b38b531e077c3d4e407cd7e057b30225ee7bb825088867b495c617697a93d1296c82848598c150162a2ed150d9a

DASHBOARD TAB 1

Current Technology Usage | Future Technology Trend | Demographics





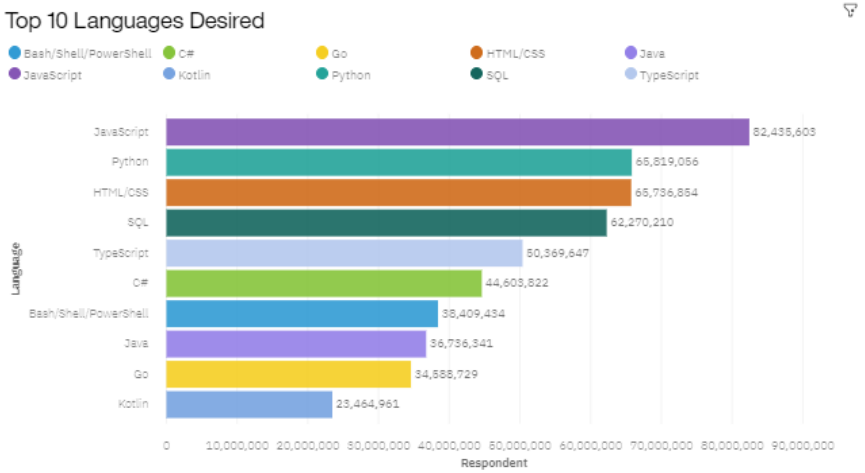


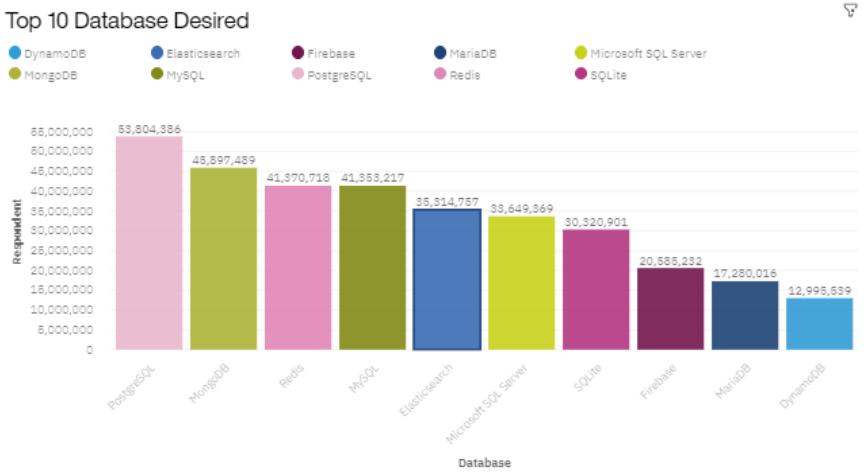
IBM Developer

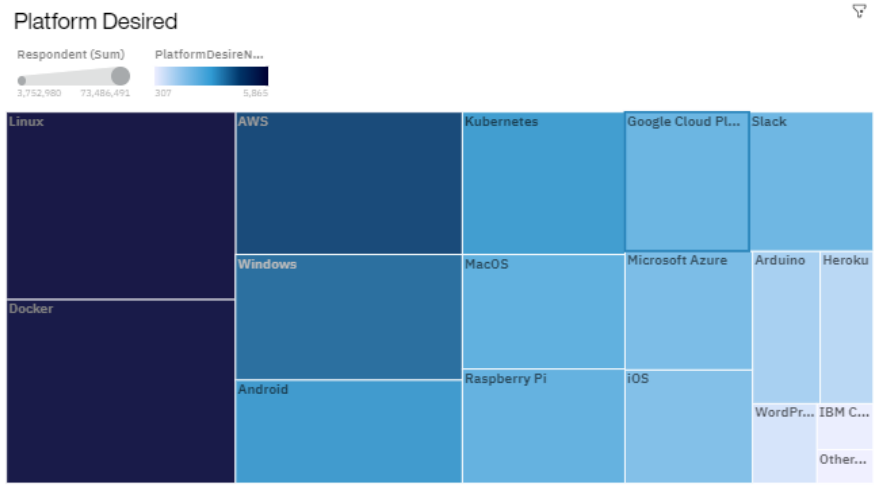
SKILLS NETWORK

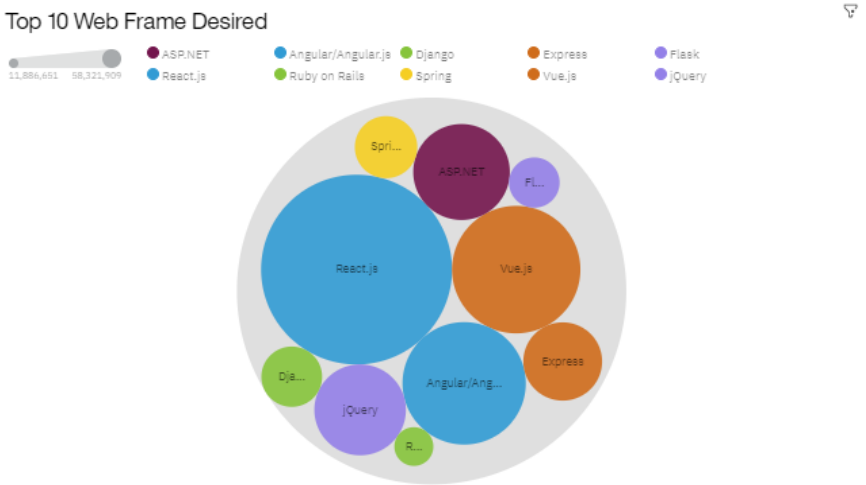
DASHBOARD TAB 2

Current Technology Usage | Future Technology Trend | Demographics





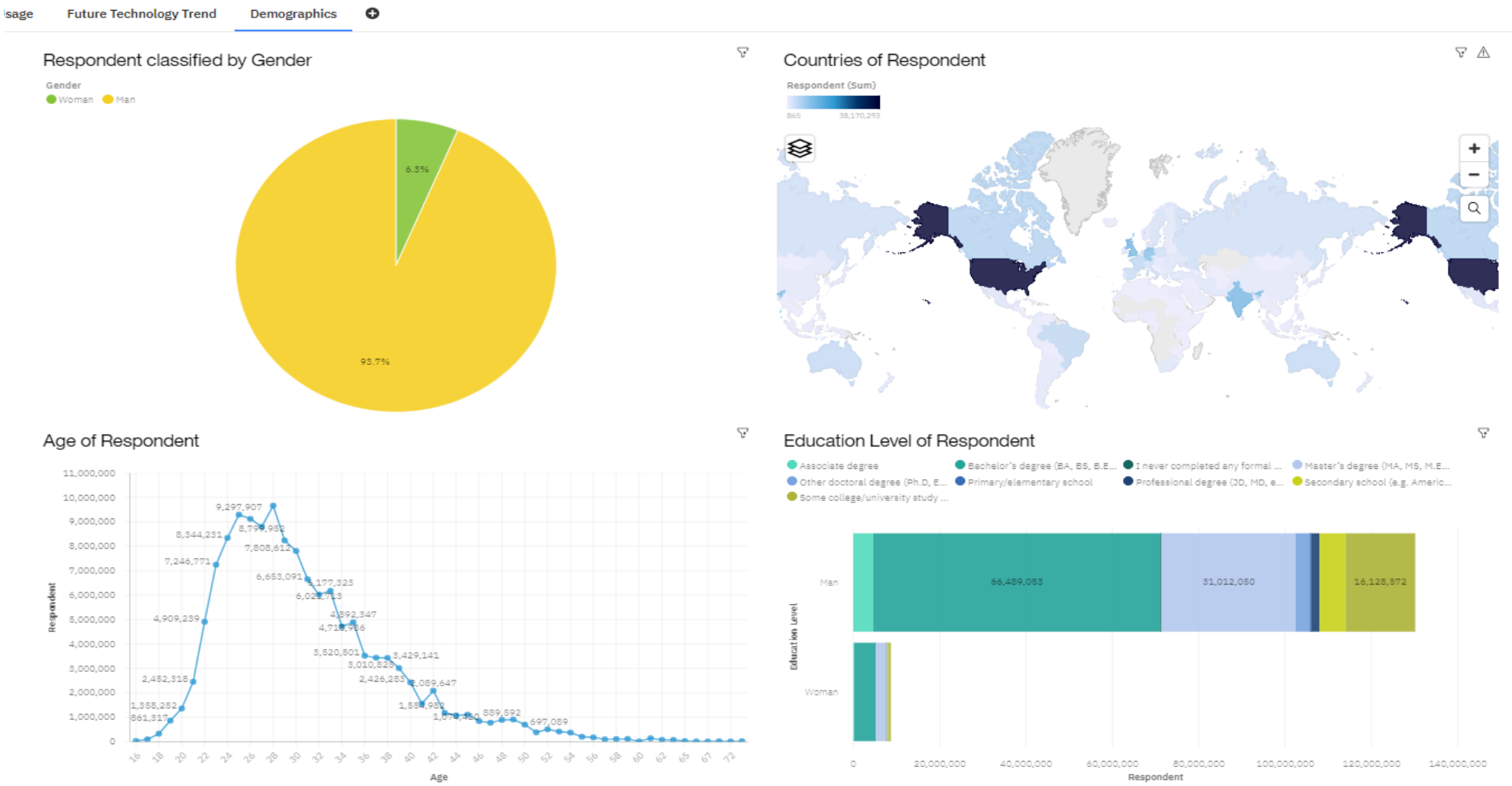




IBM Developer

SKILLS NETWORK

DASHBOARD TAB 3



IBM Developer

SKILLS NETWORK

DISCUSSION

- The majority of respondents are developer by professional.
- The mean age of respondents is 30 years old and the range of most respondents is between 20 – 40 years old
- A total of 11398 survey respondents were recorded.
- This field of work is dominated by male over female with a ratio of 93.7% to 6.3% respectively.
- Most professional have at least a degree for educational level.
- For the age range of 25 - 30 year old, the compensation has a positive correlation with age.
- The normalised annual salary of respondents is USD 103,000.

OVERALL FINDINGS & IMPLICATIONS

Findings

- Web page interactive languages are still in demand while Python is the trend
- SQL database are still in demand
- Linux remains as the preferred platform
- React.js is the future trend for web frame

Implications

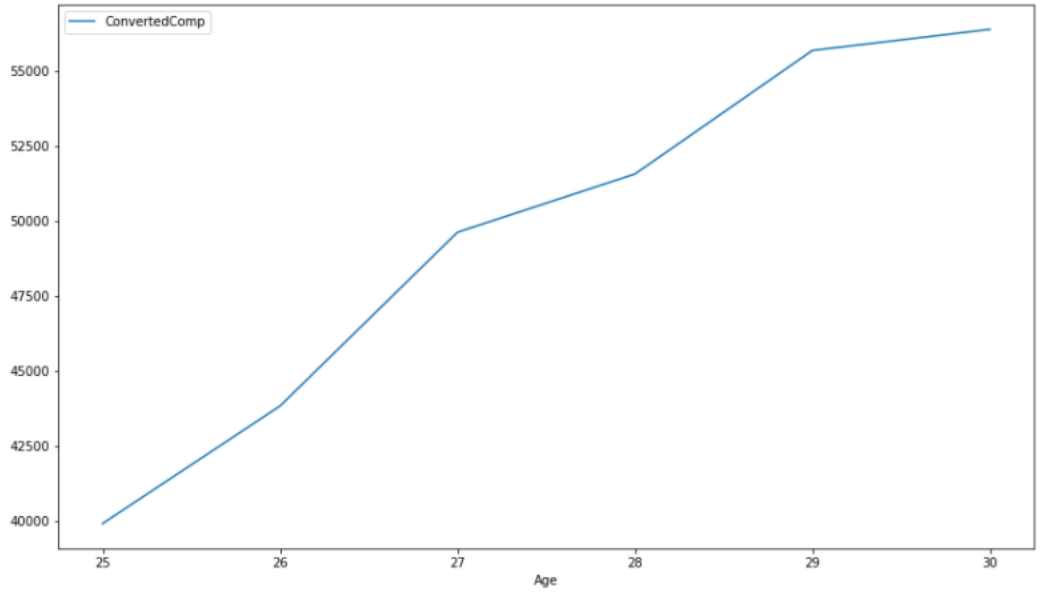
- Web page interactive languages are still in demand due to digitalization.
- Professional are slowly picking up Python for programming language
- Professional are migrating from Window platform and jQuery web frame

CONCLUSION

- Webpage interactive languages like Javascript and HTML/CSS are still in high demand and are relatively well paid.
- Python is the upcoming programming language with high paying salary and plenty of job opportunity
- Professional working on platform and web frame are constantly changing
- This field of work need more gender diversity

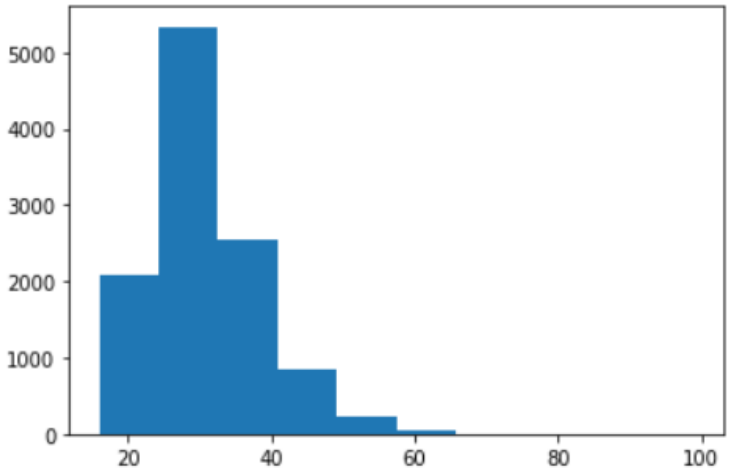
APPENDIX

```
# your code goes here
QUERY = """
SELECT ConvertedComp, Age
FROM master
WHERE Age BETWEEN 25 AND 30
"""
# the read_sql_query runs the sql query and returns the data as a dataframe
df = pd.read_sql_query(QUERY,conn)
df_age = df.groupby('Age').median()
df_age.plot(kind='line', figsize=(14, 8))
plt.show()
```



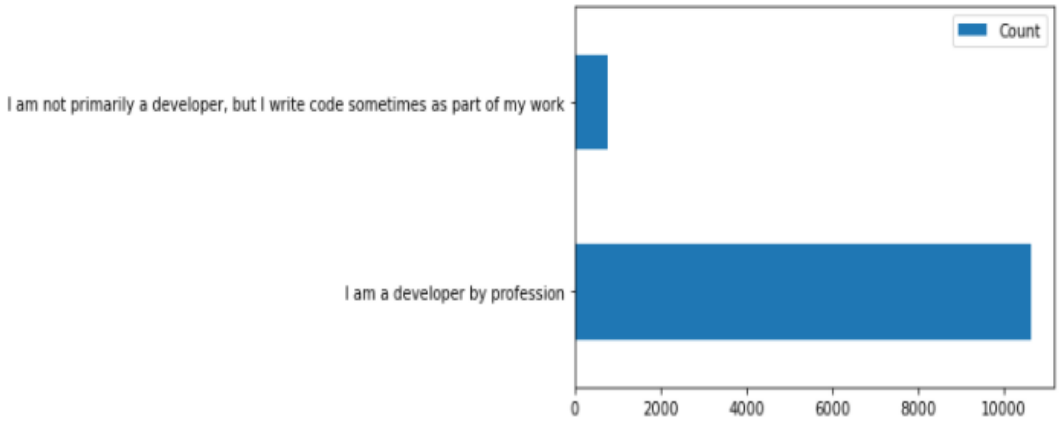

```
: 1 # your code goes here
  2 plt.hist(x = df['Age'])
```

```
: (array([2.094e+03, 5.337e+03, 2.557e+03, 8.420e+02, 2.250e+02, 4.900e+01,
        6.000e+00, 0.000e+00, 0.000e+00, 1.000e+00]),
 array([16. , 24.3, 32.6, 40.9, 49.2, 57.5, 65.8, 74.1, 82.4, 90.7, 99. ]),
 <a list of 10 Patch objects>)
```



APPENDIX

```
QUERY = """
SELECT count(MainBranch) as Count, MainBranch
FROM master
GROUP BY MainBranch
"""
# the read_sql_query runs the sql query and returns the data as a dataframe
df = pd.read_sql_query(QUERY,conn)
import numpy as np
bars = (df['MainBranch'][0],df['MainBranch'][1])
y_pos = np.arange(len(bars))
df.plot(kind='barh')
plt.yticks(y_pos, bars)
plt.show()
```



Double click to see the **Hint**.

```
# your code goes here
def normalise(freq,total):
    result = 0
    if freq == 'Yearly':
        result = total
    elif freq == 'Monthly':
        result = total * 12
    elif freq == 'Weekly':
        result = total *52
    return result
```

```
# freq = df['CompFreq'].tolist()
# total = df['CompTotal'].tolist()
annual = [normalise(freq, total) for freq, total in zip(df['CompFreq'], df['CompTotal'])]
df['NormalizedAnnualCompensation'] = annual
```

```
df.shape
```

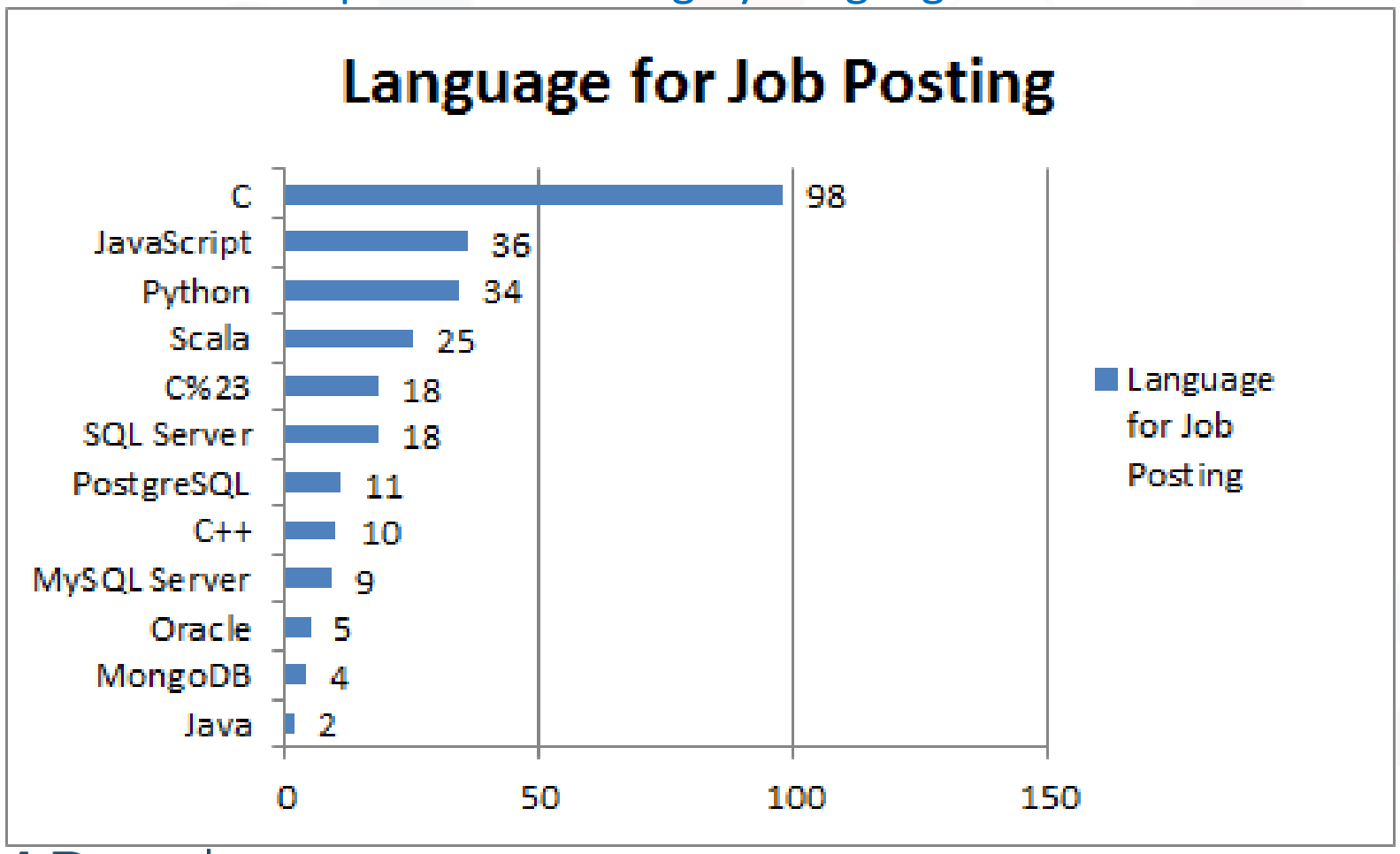
```
(11398, 86)
```

```
df['NormalizedAnnualCompensation'].median()
```

```
103000.0
```

GITHUB JOB POSTINGS

Bar Chart of Popular Job Posting by Languages



POPULAR LANGUAGES

Bar Chart of Average Annual Salary by Popular Languages