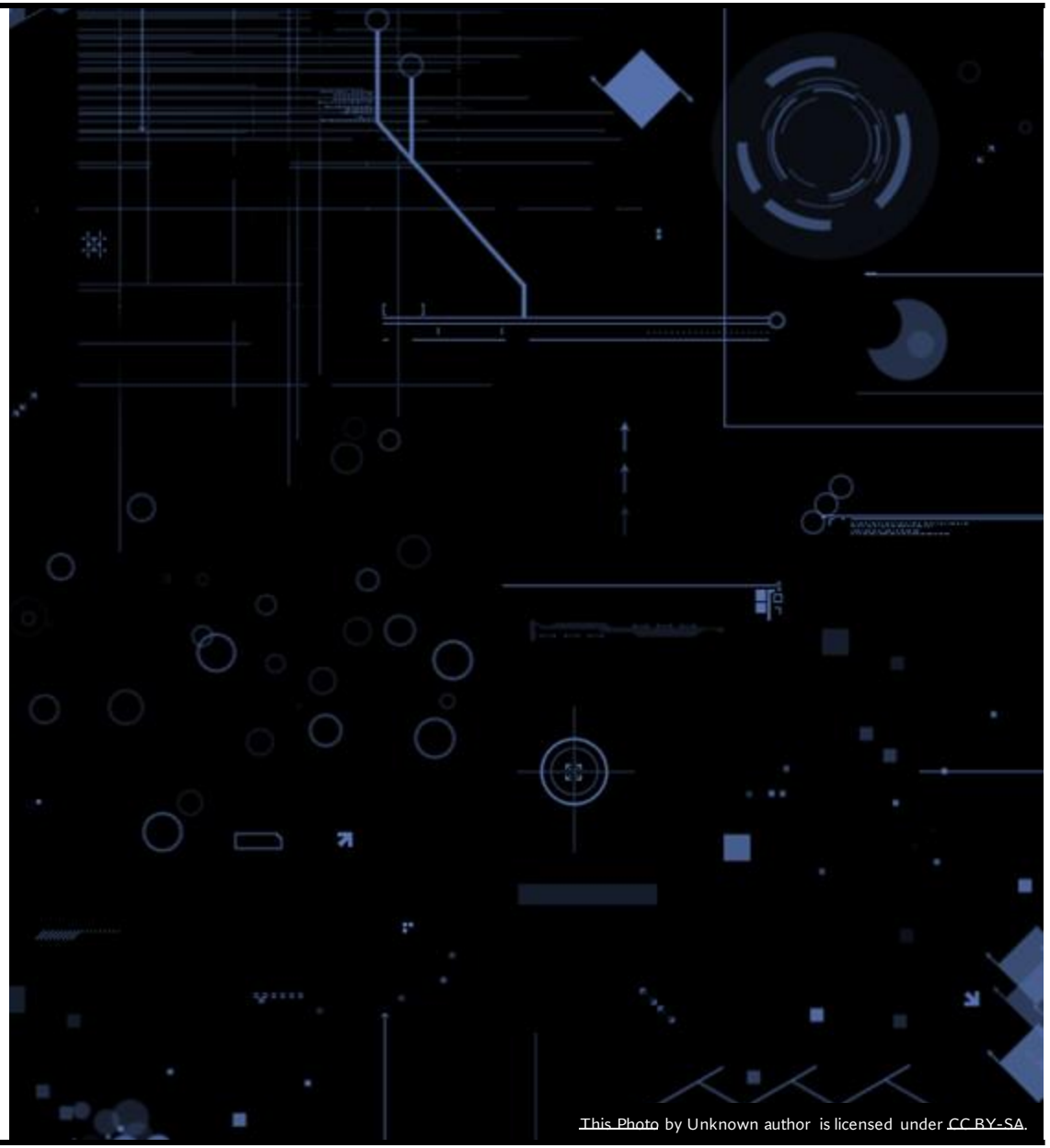
Stack Overflow Developer Survey Data Analysis

Shirish Senthil Kumar

05/01/2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion



Executive Summary

- The necessary abilities for IT and business consulting are always changing and evolving.
- To keep up with evolving technology and to stay competitive, it's critical to understand future skill requirements and trends.
- This presentation will show current and future trends in Programming Languages, Databases, Platforms and Webframes.
- The presentation will also aid in recognizing key skill shortages and the analysis will present data that can be used to make more informed data-driven budgetary decisions.

Introduction

Objective: To conduct a comprehensive analysis for stakeholders and business decision makers within the global IT and business consulting services firm based on StackOverflows Survey Data.

Desirable Outcomes:

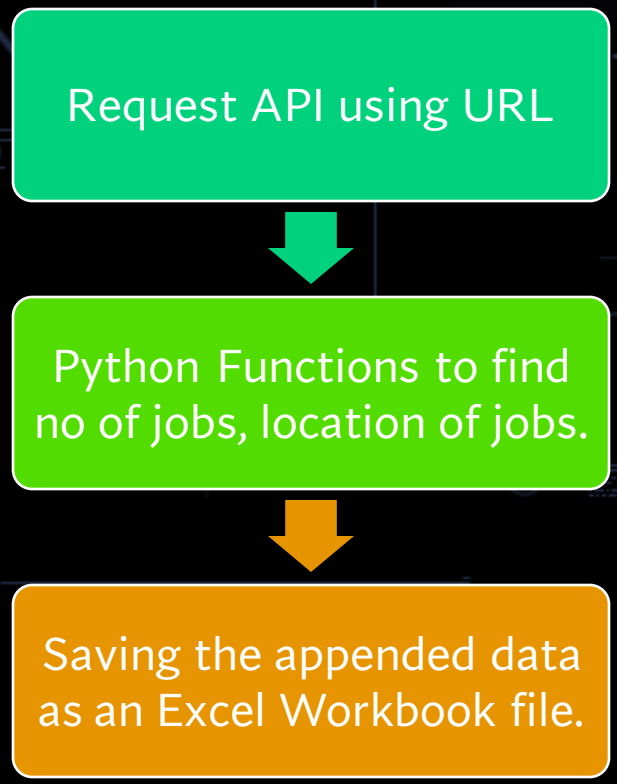
- Show strong correlation in key skill requirements data using SQL and Python.
- Dashboard analytics to show key charts and maps.

Methodology

Data Collection - API

- The data collection is carried through an API that connects to the survey data.
  URL: http://127.0.0.1:5000/data
- Payload Key-Value parameterization is used to access the key data.
- Several functions is called to verify the results.

Source Code:https://github.com/Shirish026/IBM-Data-Analytics-Capstone-Project



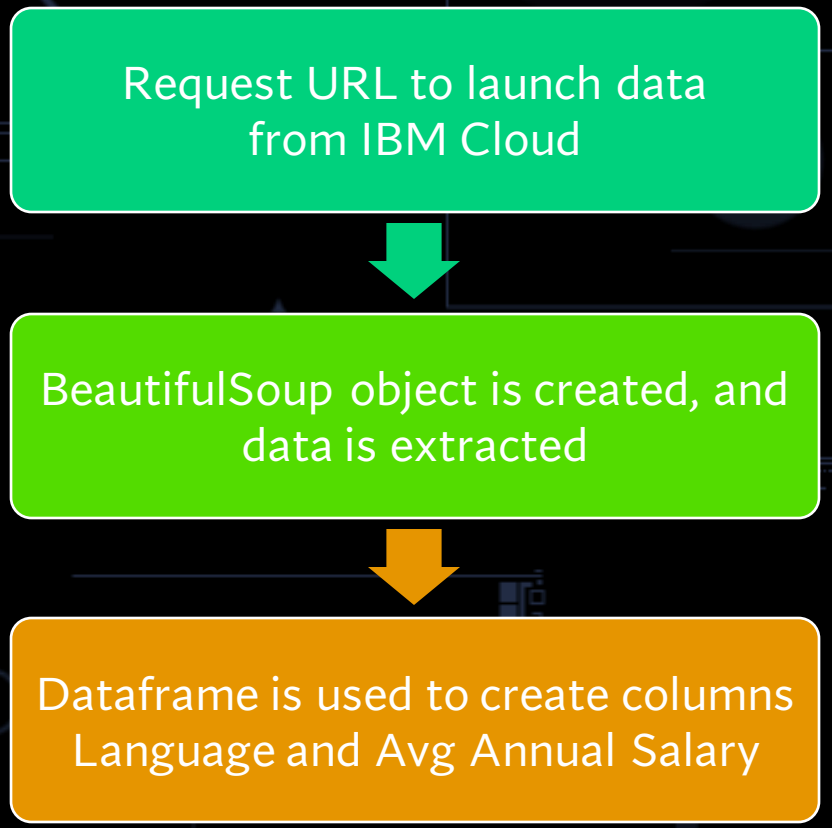
Web Scrapping

- Web Scrapping is carried through BeautifulSoup liibrary on the following URL

  URL: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/labs/datasets/Programming_Languages.html

- Using the BeautifulSoup Object, data extraction is conducted.

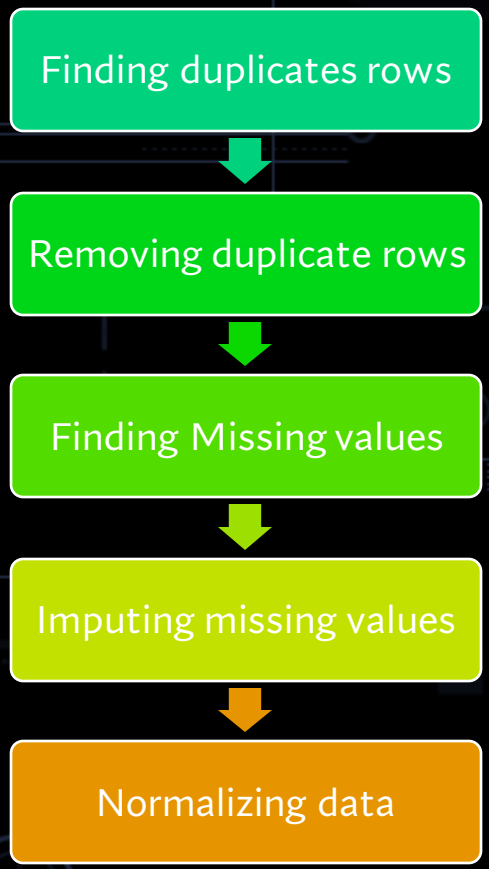
Source Code:https://github.com/Shirish026/IBM-Data-Analytics-Capstone-Project



Data Wrangling

- Data wrangling is the process of removing errors and combining complex data sets to make them more accessible and easier to analyze.
- A function to calculate the annual compensation.

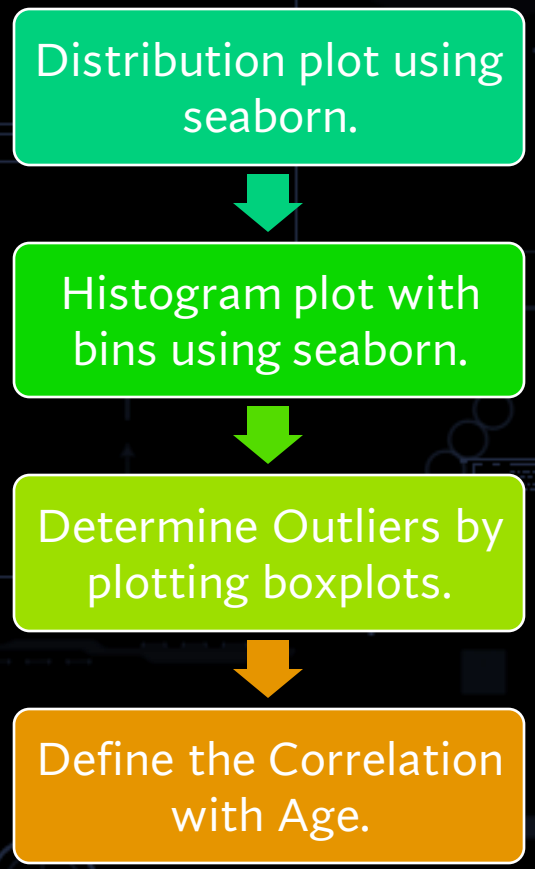
Source Code:https://github.com/Shirish026/IBM-Data-Analytics-Capstone-Project/blob/main/M2DataWrangling-lab.ipynb



Exploratory Data Analysis

- Exploratory Data Analysis refers to the critical process of performing initial investigations on data to discover patterns and trends.
- Seaborn and Matplotlib to plot the graphs and charts to better understand the data.

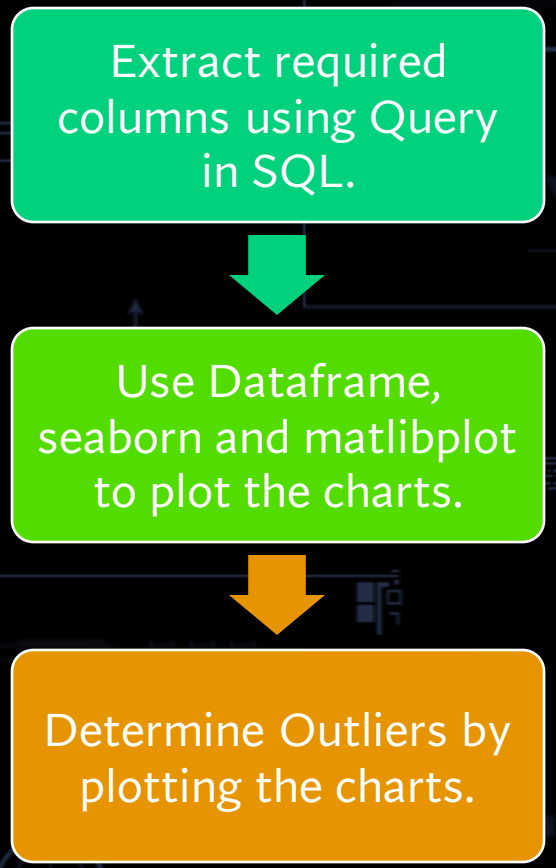
Source Code: https://github.com/Shirish026/IBM-Data-Analytics-Capstone-Project/blob/main/M3ExploratoryDataAnalysis-lab.ipynb



Data Visualization

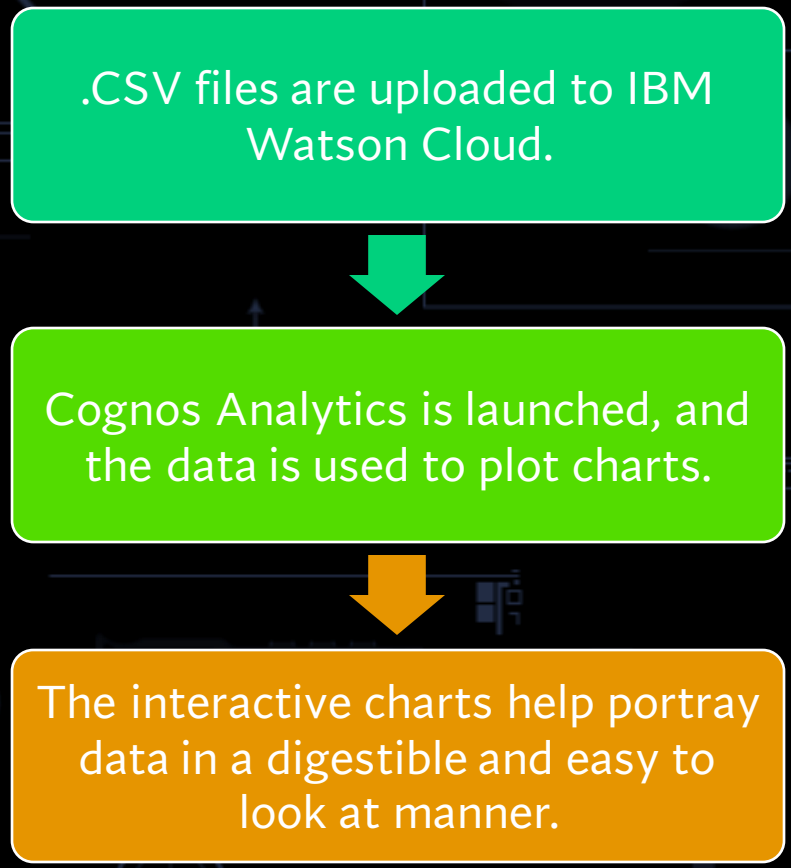
- Using Seaborn and Matplotlib to plot the graphs to better understand the data.
- Graphs plotted include Box, Histogram, Scatter plots, Bubble Plot, Pie Charts, Stacked Chart, Line Chart, Bar Chart.

Source Code:https://github.com/Shirish026/IBM-Data-Analytics-Capstone-Project/blob/main/M4DataVisualization-lab.ipynb



Dashboard on Cognos Analytics

- .CSV files of survey data are uploaded to IBM Watson and analyzed.
- Three separate dashboards were set up, primarily: current trend, desired future trends and the demographic associated with these key IT skill roles.
- Numerous Interactive Charts are plotted that can be filtered, sorted on the go.
- Source Code:https://github.com/Shirish026/IBM-Data-Analytics-Capstone-Project



Results

Data Collection using API

Screenshot shows the use of API URL to extract data with the use of function "get_number_of_jobs_Loc_list and payload parameters.

```
locations=["Los Angeles", "New York", "San Francisco", "Washington DC", "Seattle"]

def get_number_of_jobs_Loc_list(locations):
    number_of_jobs_list = []
    for location in locations:
        payload={"Location":location}
        response=requests.get(api_url, params=payload)
        if response.ok:
            data=response.json()
            number_of_jobs = len(data)
            number_of_jobs_list.append({location: number_of_jobs})

    return number_of_jobs_list
```

```
#your code goes here
get_number_of_jobs_Loc_list(locations)
```

```
[{'Los Angeles': 640},
 {'New York': 3226},
 {'San Francisco': 435},
 {'Washington DC': 5316},
 {'Seattle': 3375}]
```

Web Scrapping

Screenshot shows the use of BeautifulSoup to extract data from the html link to create two rows of "Languages" and "Average Annual Salary".

```
Soup = BeautifulSoup(data,'html')
```

Scrape the `Language name` and `annual average salary`.

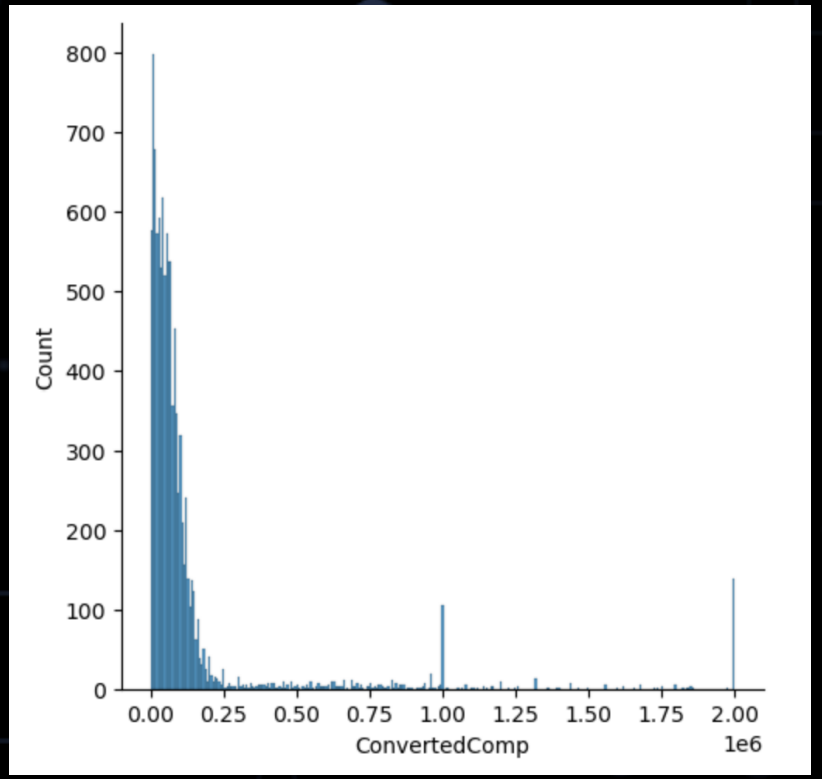
```
table = Soup.find("table")
language_name = []
Average_Salary = []
for row in table.find_all("tr"):
    language_name.append(row.find_all("td")[1].getText())
    Average_Salary.append(row.find_all("td")[3].getText())

df = pd.DataFrame(columns=[language_name,Average_Salary]).T.reset_index()
df = df.rename(columns={"level_0": "Language", "level_1": "Average Annual Salary"})[1:]
```

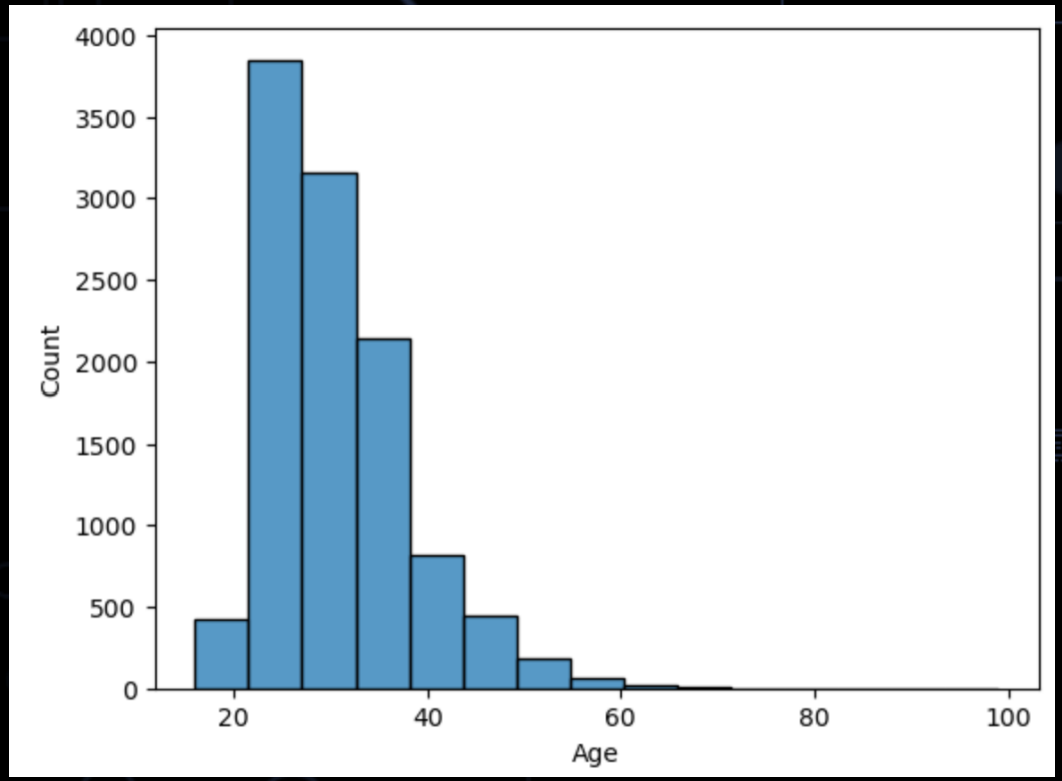
	Language	Average Annual Salary
1	Python	$114,383
2	Java	$101,013
3	R	$92,037
4	Javascript	$110,981
5	Swift	$130,801
6	C++	$113,865
7	C#	$88,726
8	PHP	$84,727
9	SQL	$84,793
10	Go	$94,082

Exploratory Data Analysis

Screenshots shows the use of plots to portray data that can be easily interpreted and analyzed.



Distribution Plot

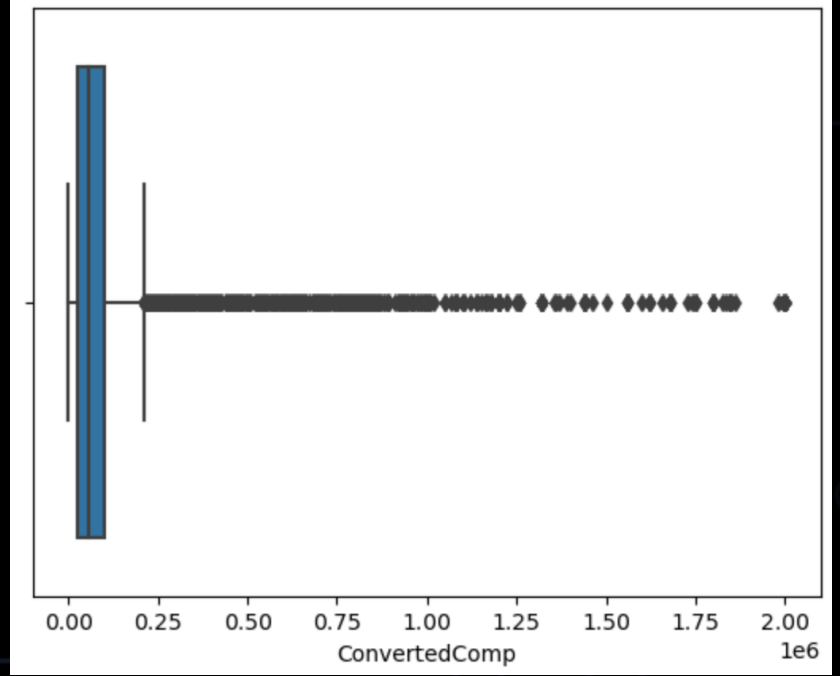


Histogram Plot

Exploratory Data Analysis

Screenshot shows the use of Boxplots and inter quantile ranges to show outliers in the data.

```
sns.boxplot(data=df, x="ConvertedComp");
```

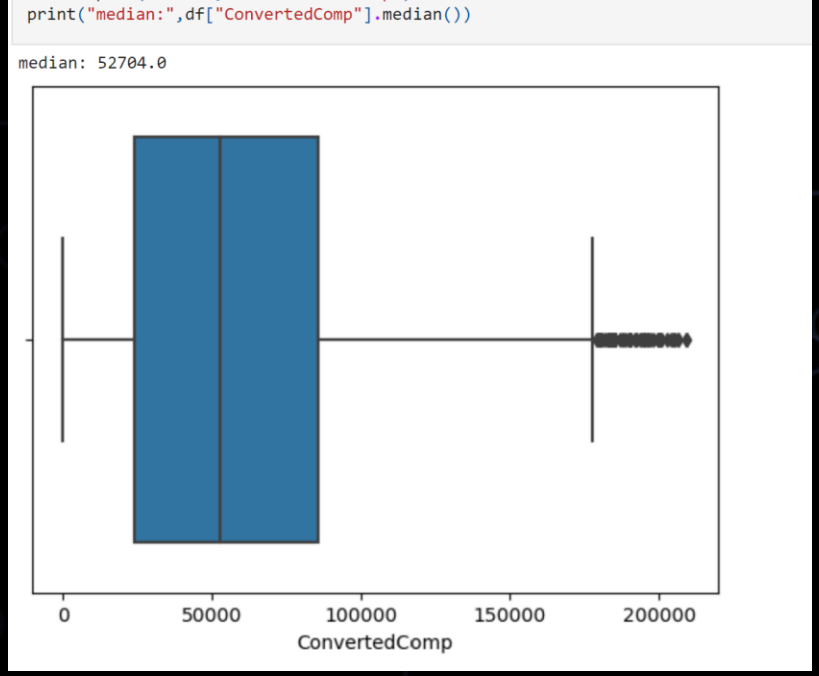


Box Plot with Outliers

```
df["ConvertedComp"] = df["ConvertedComp"].drop(index=outliers.index).to_frame()

sns.boxplot(data=df, x="ConvertedComp")
print("median:",df["ConvertedComp"].median())
```

```
median: 52704.0
```



Box Plot without Outliers

Exploratory Data Analysis

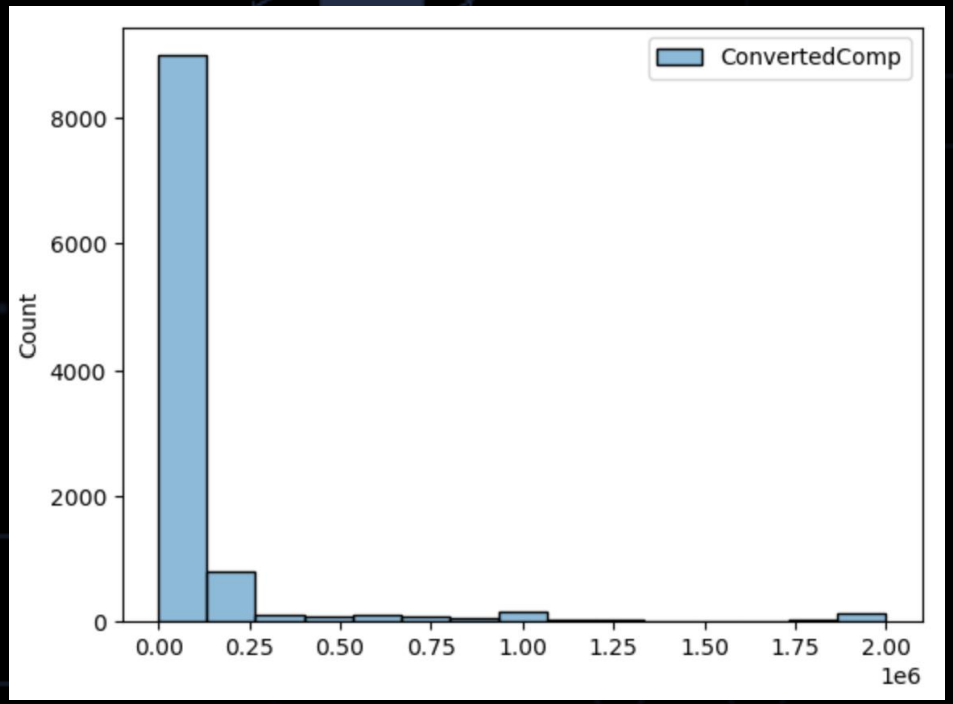
- Screenshot shows the correlation of the variable 'Age' with other variables.
- It also indicates the positive, negative and less to no correlation with the variable Age.

```
df.corr()['Age']
```

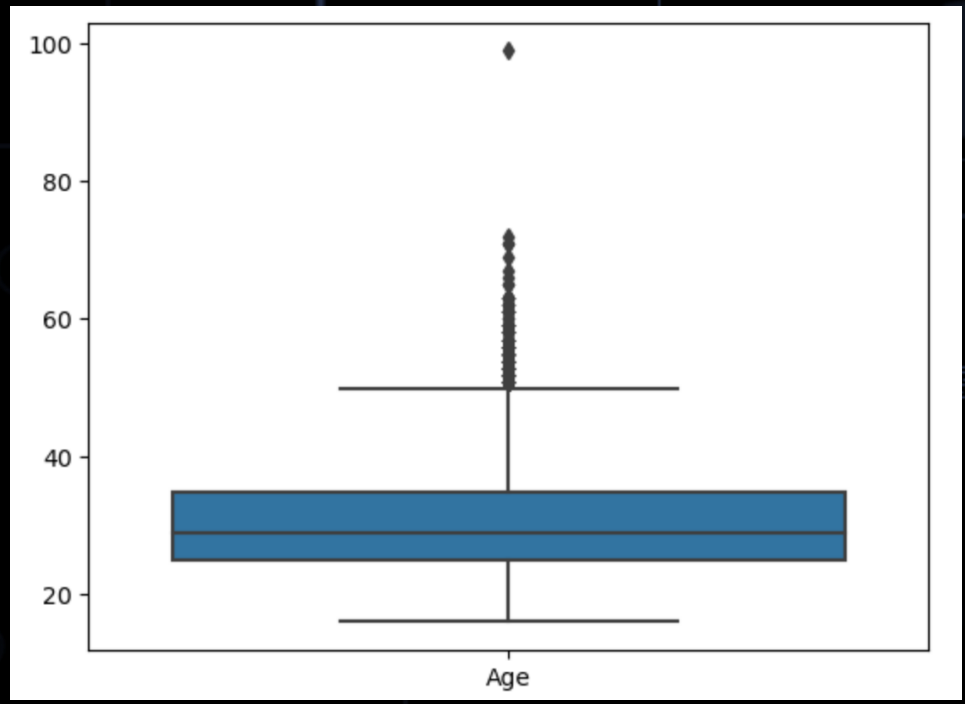
```
Respondent        0.004041
CompTotal         0.006970
ConvertedComp     0.401821
WorkWeekHrs       0.036518
CodeRevHrs       -0.020469
Age               1.000000
Name: Age, dtype: float64
```

Data Visualization with Charts

Data Visualization as a data analyst deals with numerous charts and data sets.



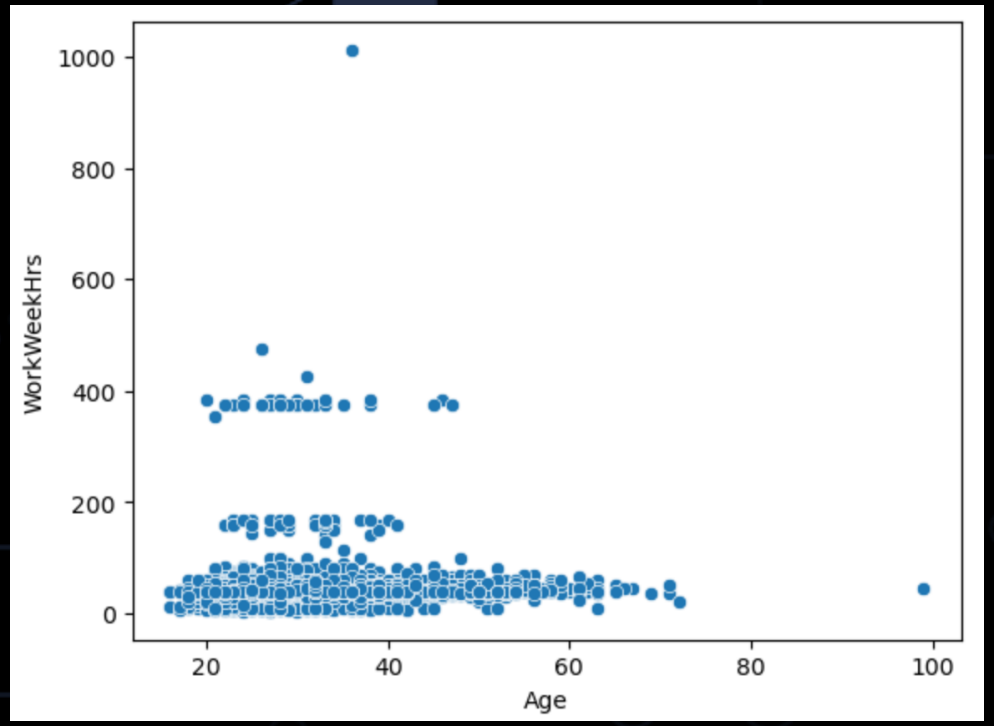
Histogram Plots



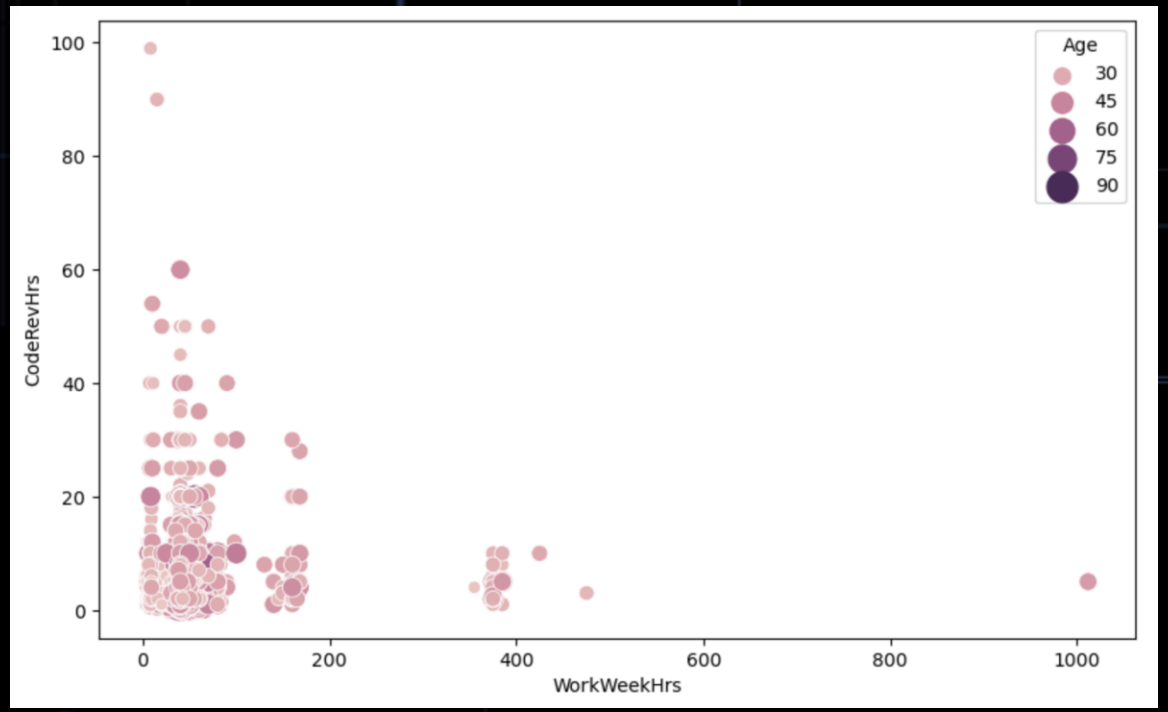
Box Plots

Data Visualization with Charts

Data Visualization as a data analyst deals with numerous charts and data sets.



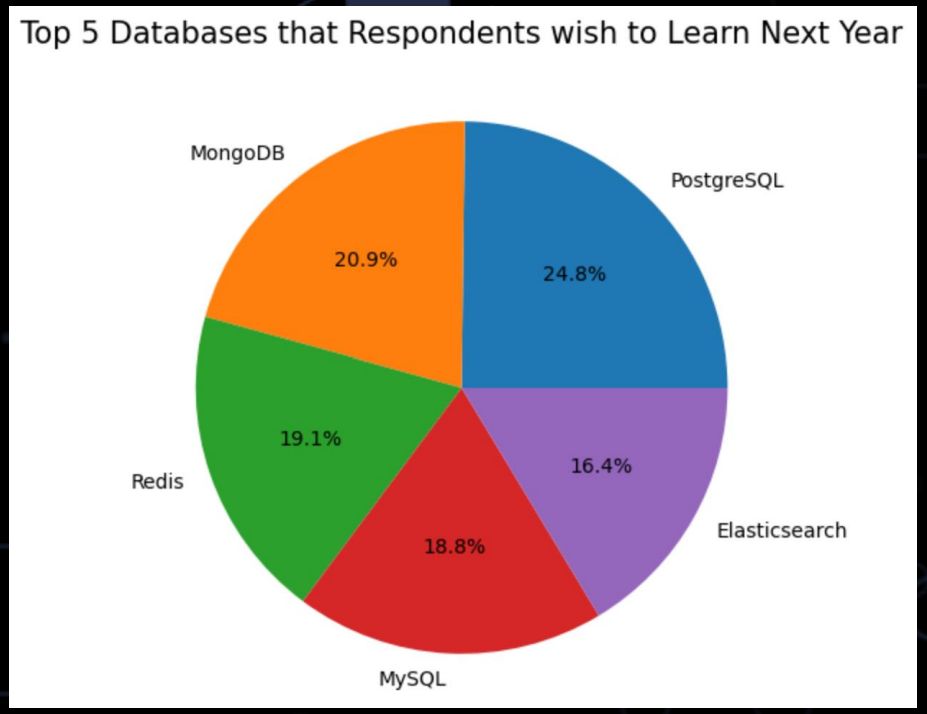
Scatter Plots



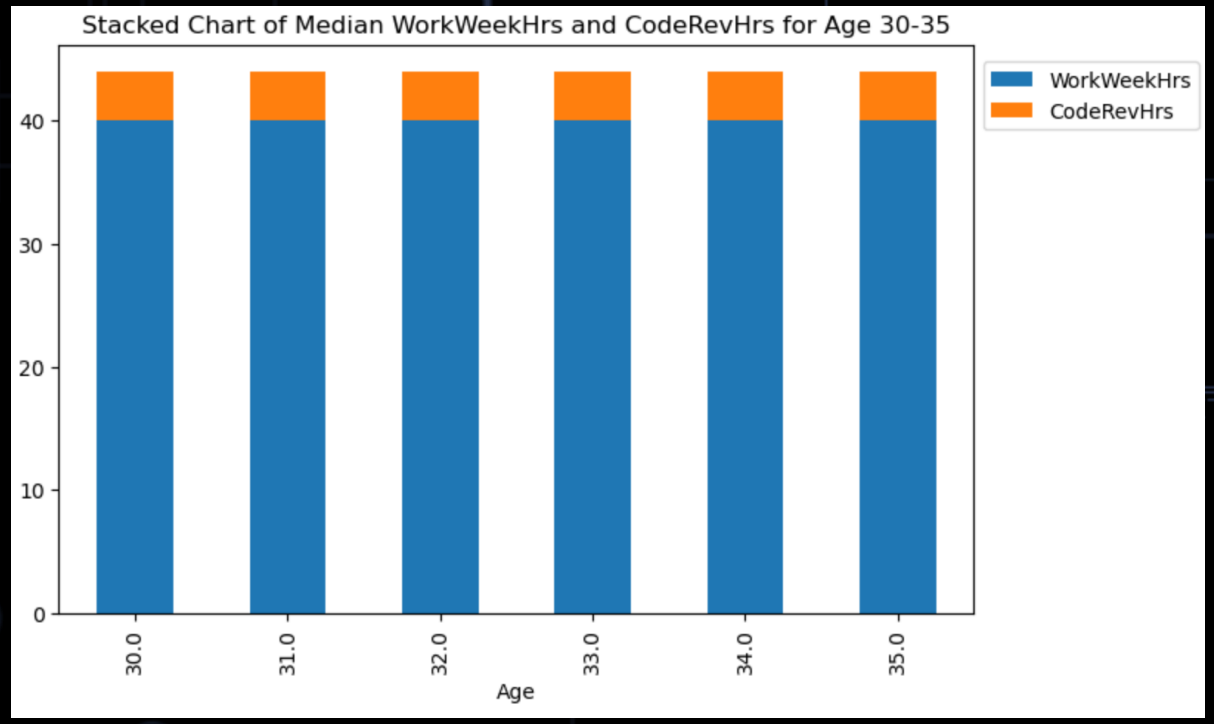
Bubble Plots

Data Visualization with Charts

Data Visualization as a data analyst deals with numerous charts and data sets.



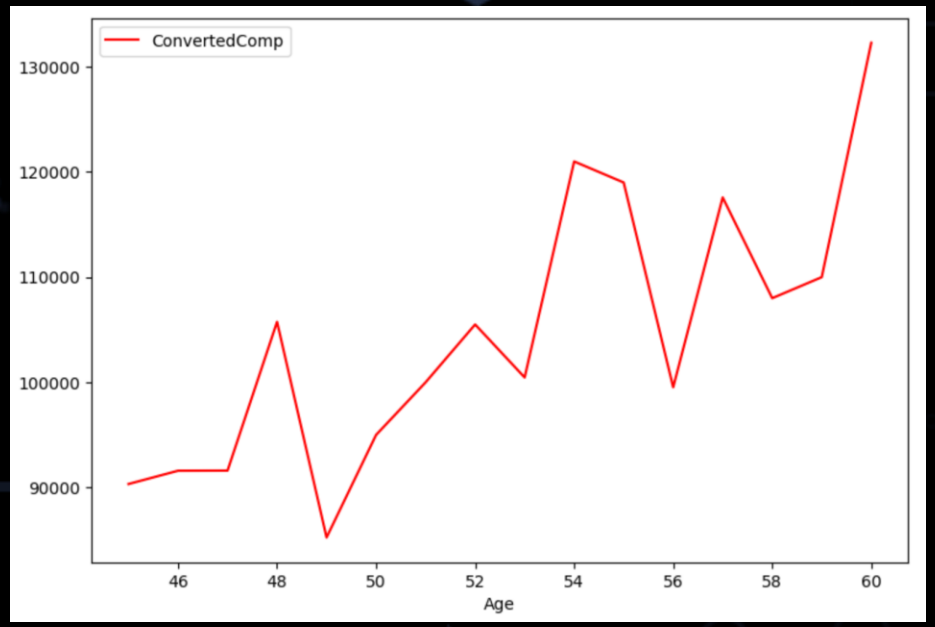
Pie Chart



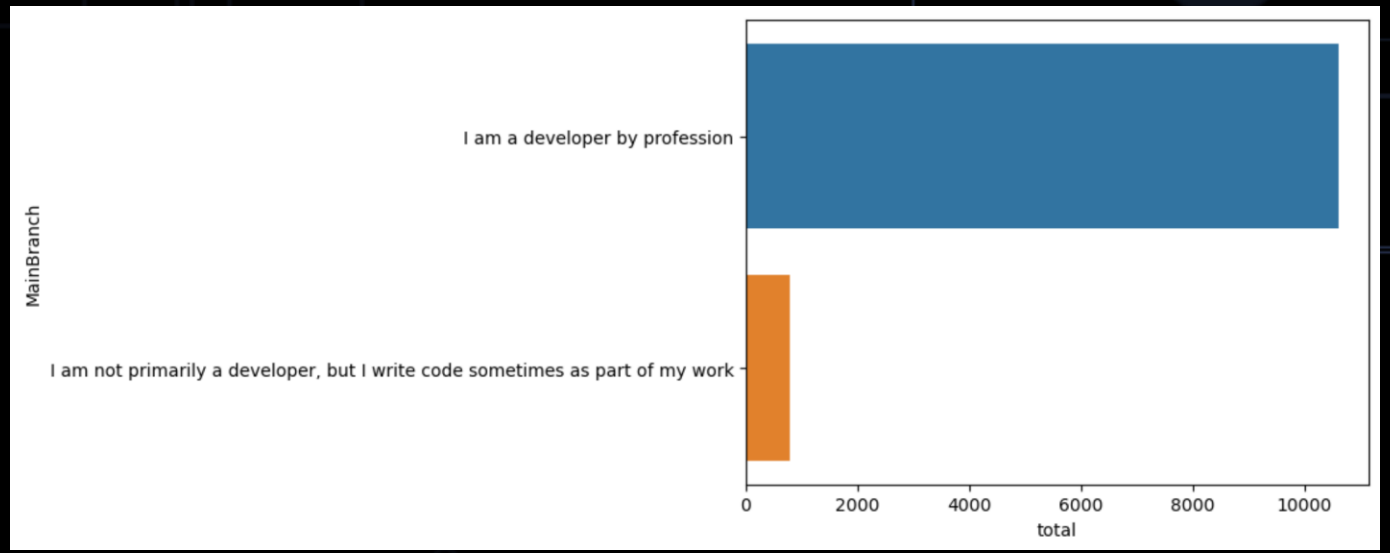
Stacked Charts

Data Visualization with Charts

Data Visualization as a data analyst deals with numerous charts and data sets.



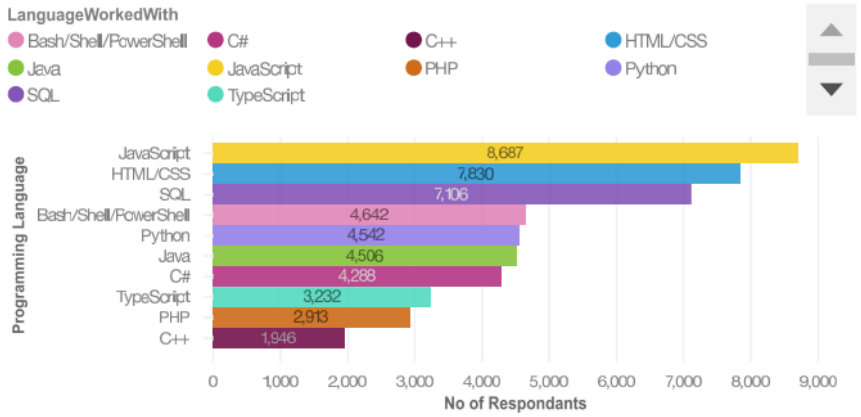
Line Plot



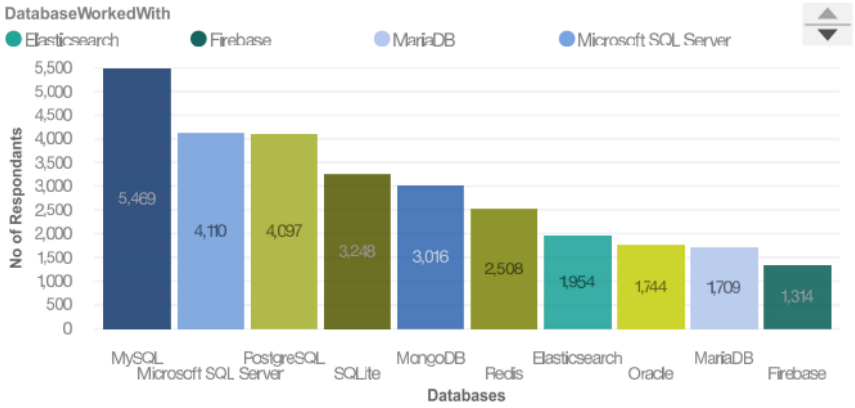
Bar Plot

Dashboard Using IBM Cognos Analytics

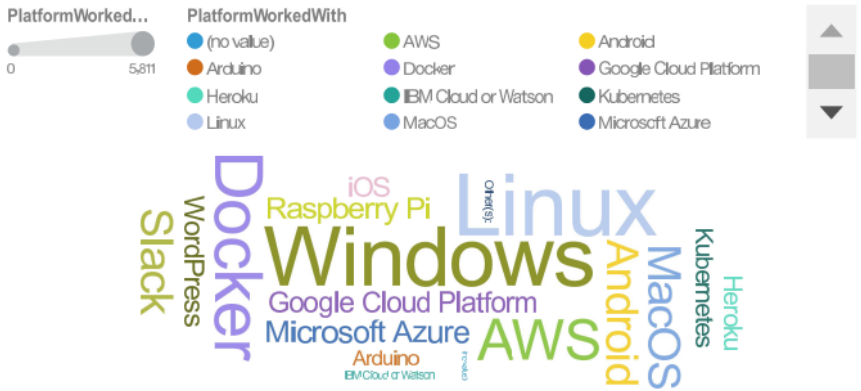
Top 10 Languages Used



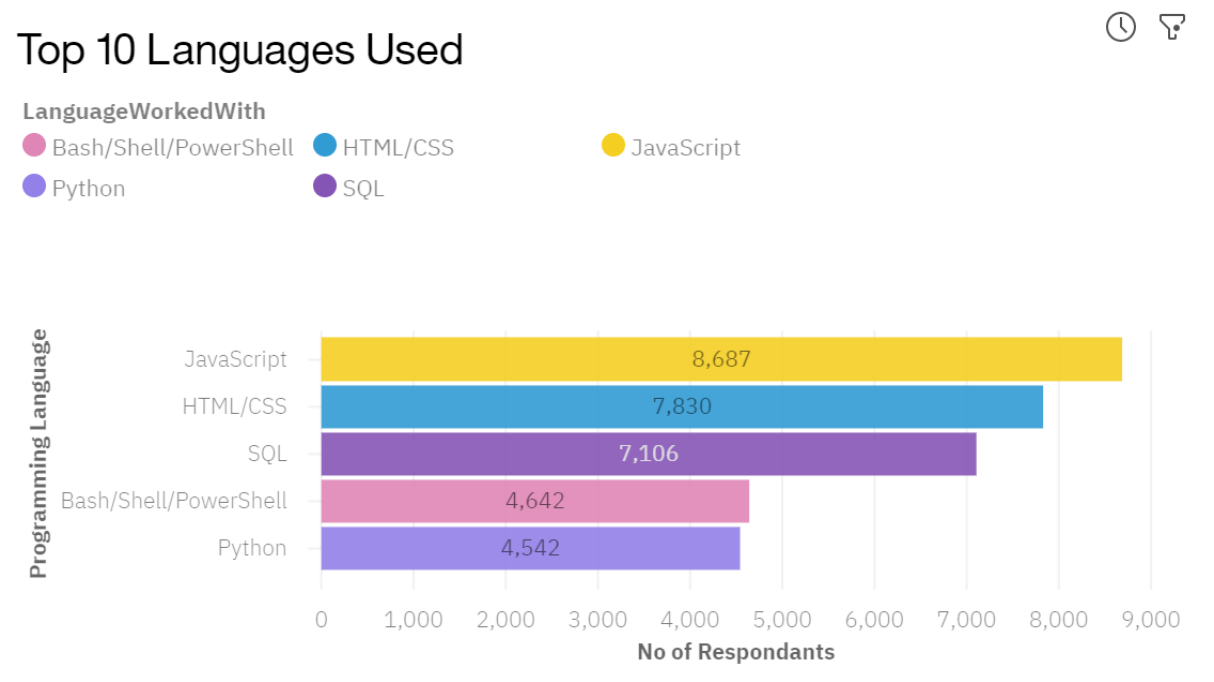
Top 10 Database Used

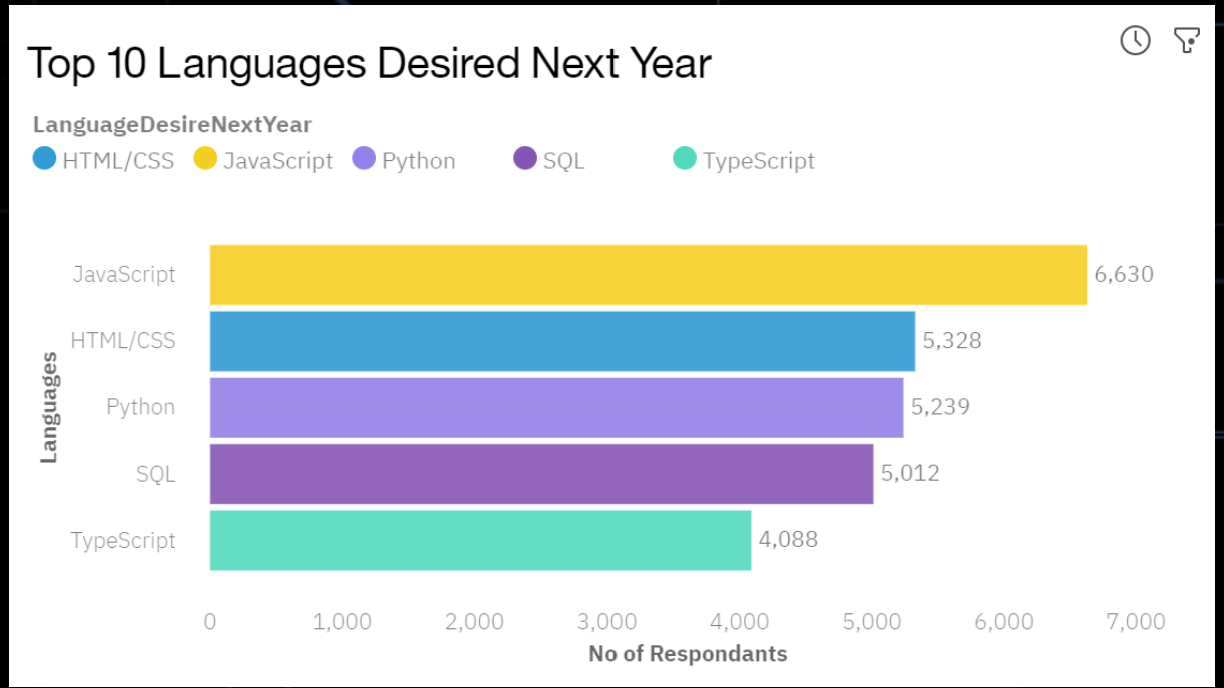


Platform worked with



Top 5 Programming Languages Trends

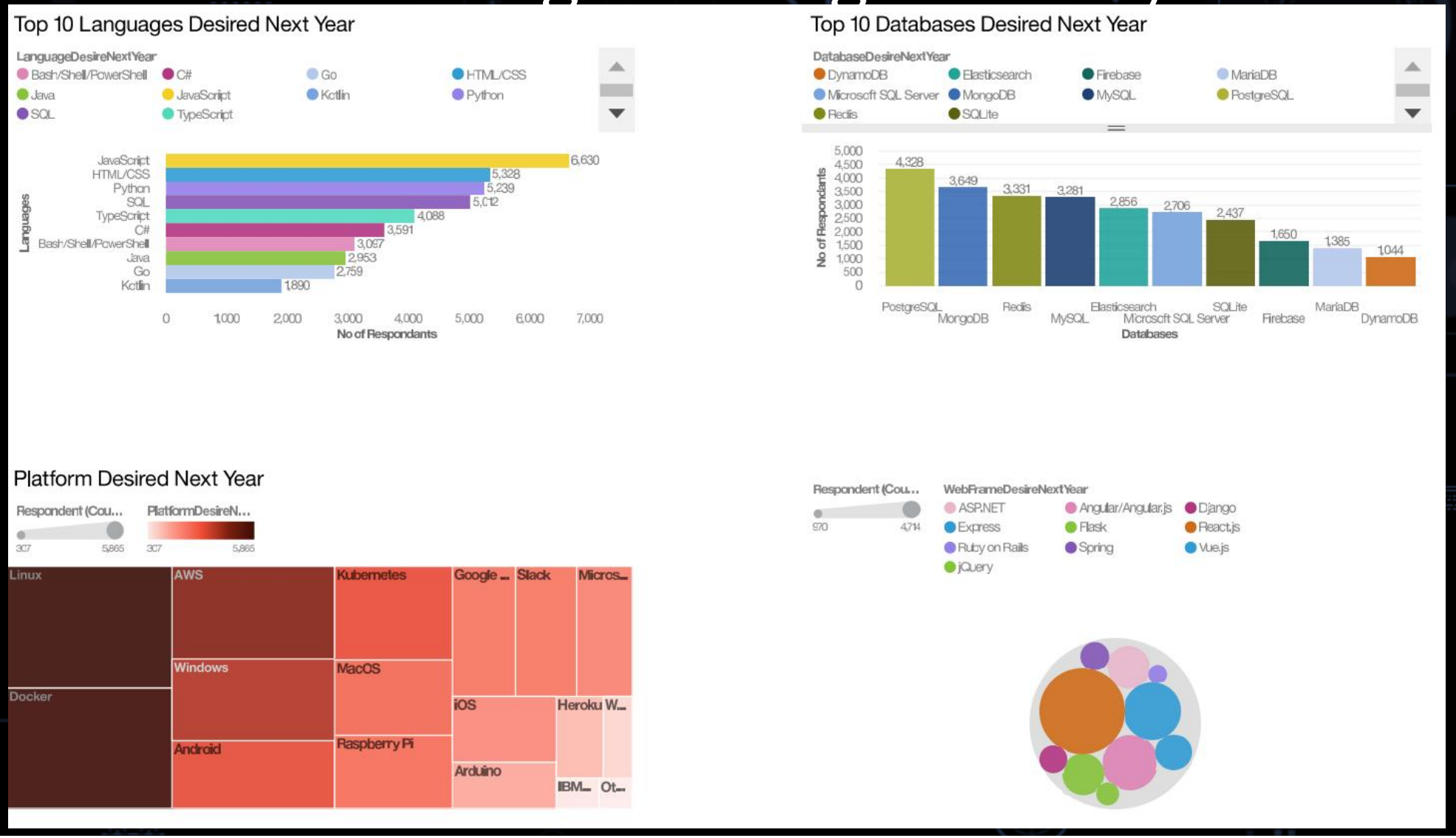




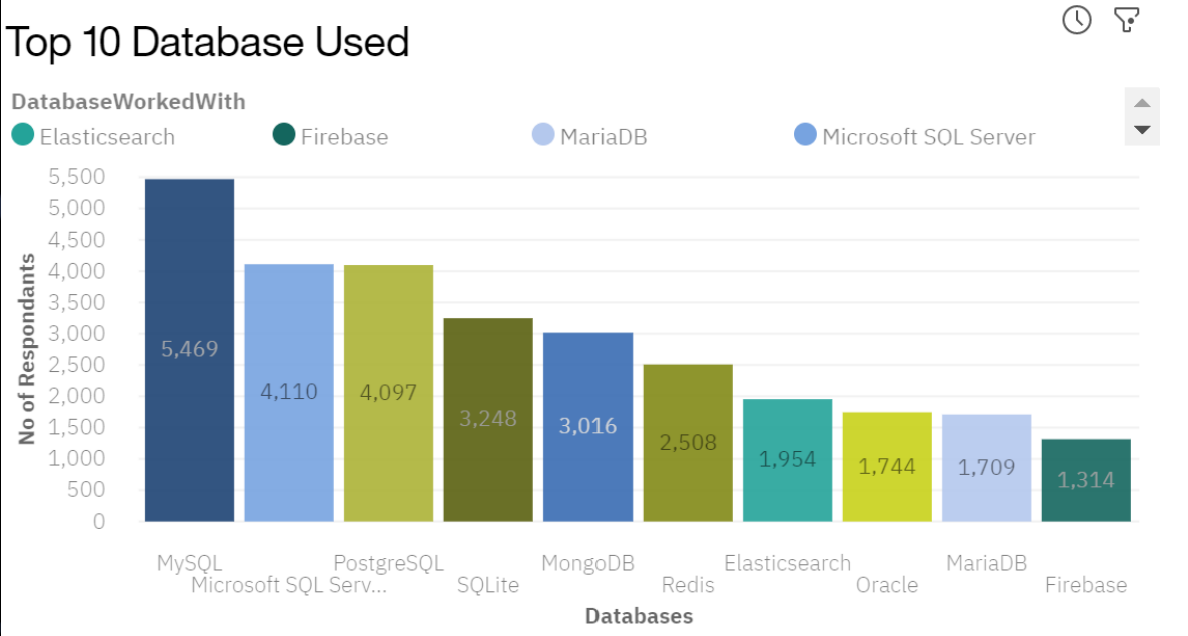
Key Inferences - Programming Languages

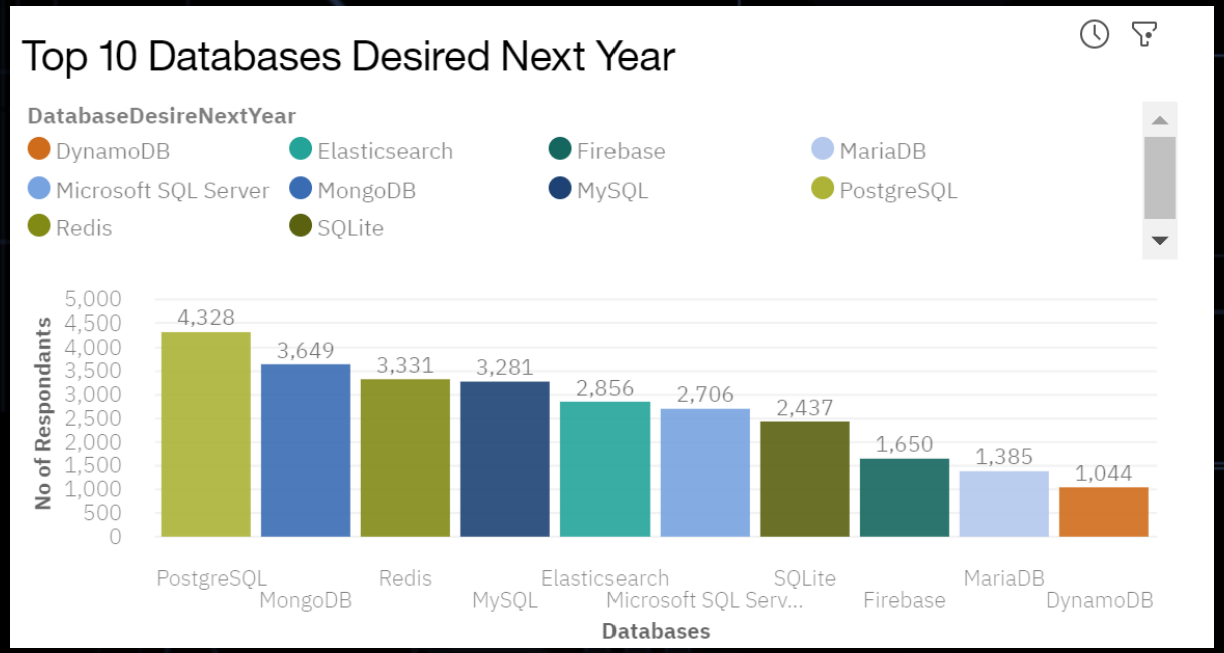
Inference	Implications
1. JavaScript and HTML/CSS stay top 2 used programming languages and will continue to be top 2 next year.	1. Employment with JavaScript and HTML/CSS will remain in demand next year.
2. The use of Python is expected to increase next year from current year.	2. Python developers' employment demand will surge next year.
3. Interest to use SQL/Powershell SQL in the next year has decreased.	3. SQL developers' employment will reduce next year

Dashboard Using IBM Cognos Analytics



Top 10 Databases Trends



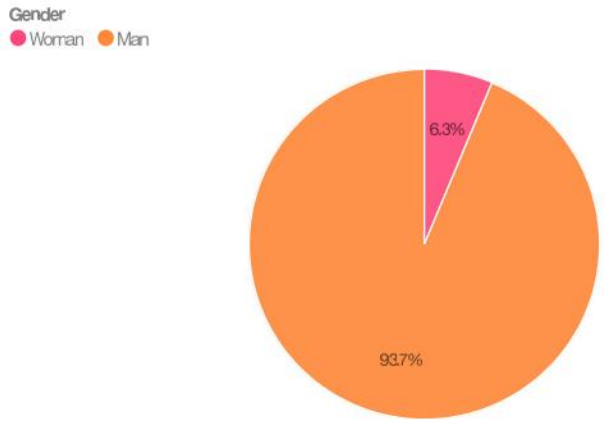


Key Inferences - Databases

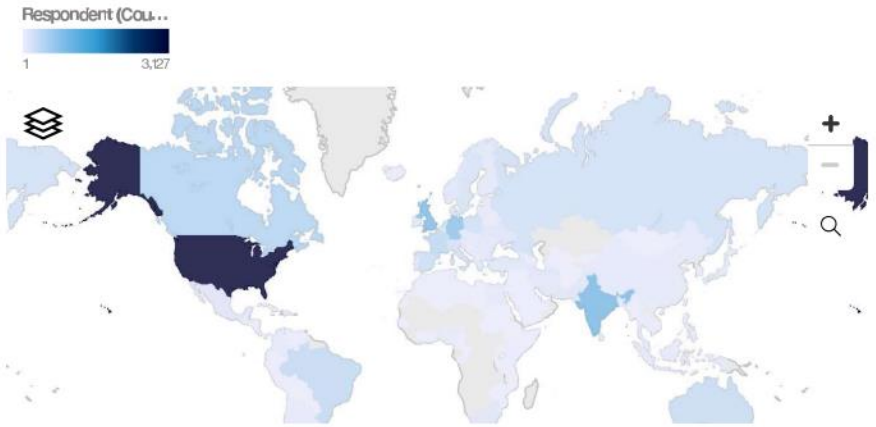
Inference	Implications
1. Desire to use MySQL, Microsoft SQL server and SQLlite has decreased in the next year.	1. Less employment in MySQL, Microsoft SQL server and SQLlite as compared to previous year.
2. The desire of PostgreSQL and MongoDB is expected to increase next year.	2. Employment in PostgreSQL and MongoDB is likely to increase next year
3. Interest to Redis and Elasticsearch in the next year has increased.	• Employment in Redis and Elasticsearch is set to increase.

Dashboard Using IBM Cognos Analytics

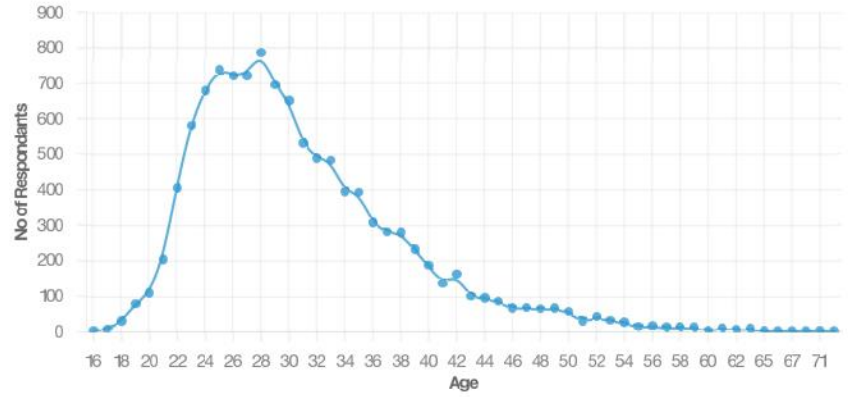
Respondent Classified by Gender



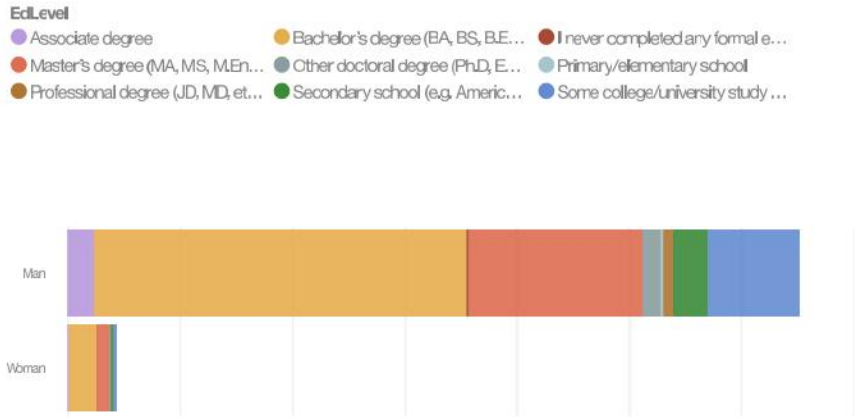
Respondent Count by Countries



Respondent Count by Age



Respondent by Gender and Education Level.



Key Inferences - Demographics

Inference	Implications
1. 93.6% respondents in the survey are men.	1. The tech industry seems to be male dominated.
2. Most respondents are aged 20-30 with maximum education at a bachelor's level	2. The bulk of the industry's employees are bachelor's degree, less experienced.
3. Most respondents are from North America (Canada and United States).	• Most tech respondents are situated in these countries.

Conclusions

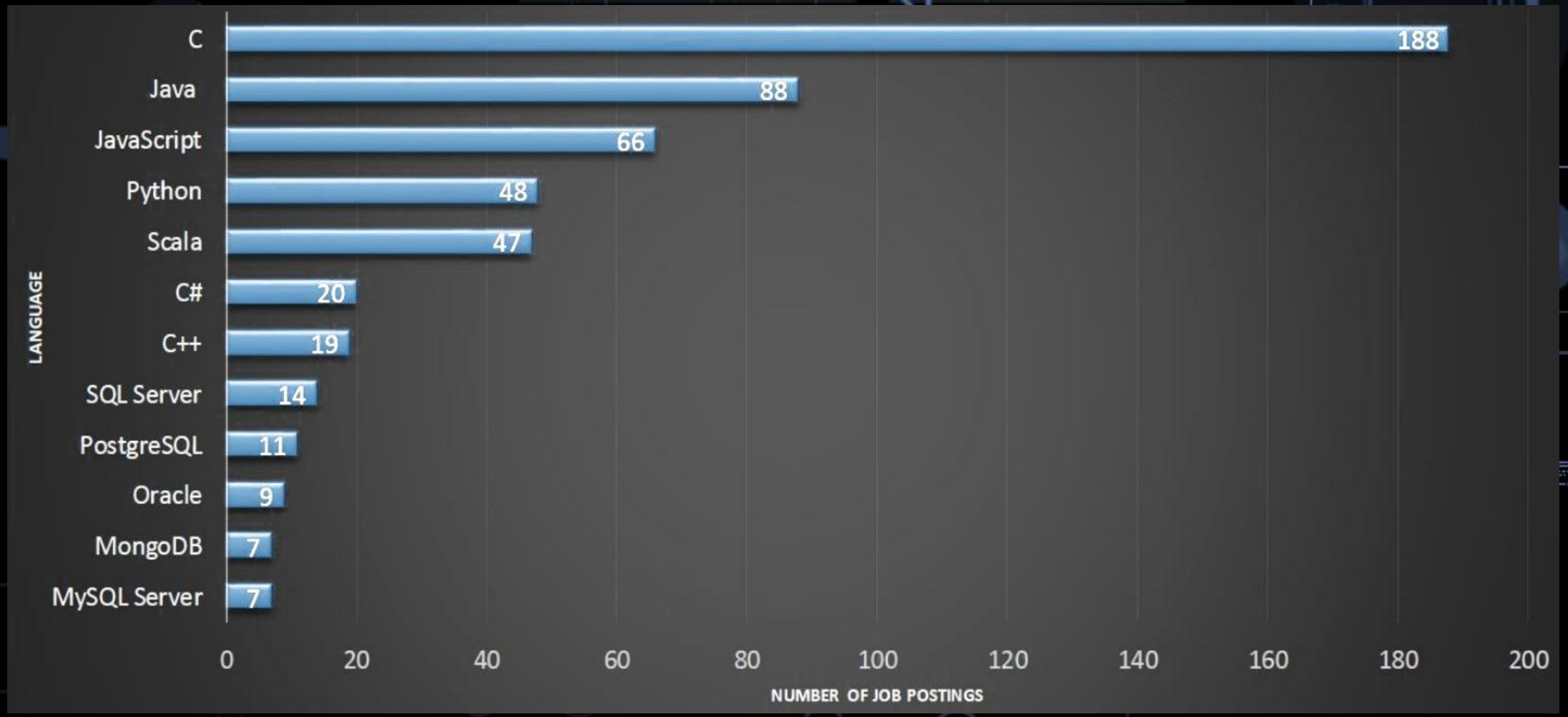
1. Certain Programming Languages like Python and Databases like MongoDB have a rise in popularity amongst developers.

2. Companies can hire employees according to the desired next year market trends and company internal architecture.

3. Upskilling can be done to keep employees up to date with the latest trend.

4. More women can be employed into the workforce to promote an inclusive work environment.

Appendix




Thank you