Final Presentation for IBM Data Analyst Capstone Project

Guan Pengfei

Mar 16, 2023

IBM Developer

SKILLS NETWORK

OUTLINE

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

EXECUTIVE SUMMARY

- Data Collection:
  - Using APIs
  - Using web scraping
  - Exploring data set
- Data wrangling:
  - Finding and removing duplicates
  - Finding and imputing missing values
  - Normalizing data
- Exploratory Data Analysis (EDA)
  - Analyzing data distribution
  - Handling outliers
  - Finding correlation
- Data Visualization
  - Visualizing the distribution of data
  - Visualizing the relationship between two features
  - Visualizing the composition of data
  - Visualizing comparison of data
- Building dashboard
  - Create a dashboard using IBM Cognos Analytics

INTRODUCTION

- Data Collection
  - Using APIs to collect jobs data (number of jobs, technologies used for jobs, job locations, etc.) and storing the collected data in an excel spreadsheet
  - Using web scraping to retrieve data from a website and saving the data into a CSV file
  - Exploring a data set by loading the dataset, finding the number of rows and columns, and identifying the data types of each column

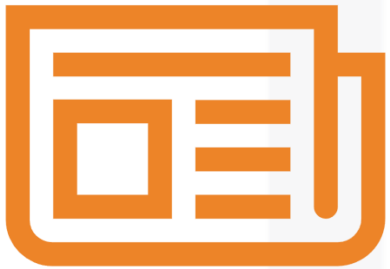
- Data wrangling
  - Finding the number and location of duplicated rows in a dataset and removing them
  - Finding the missing values and imputing (replacing) them with other meaningful values
  - Normalizing relevant columns in a dataset and making them easy for analysis

- Exploratory Data Analysis (EDA)
  - Determining how data is distributed
  - Finding the outliers in a dataset and removing them
  - Finding a correlation between features in a dataset

- Data Visualization
  - Visualizing the distribution of data in graphs (e.g., histogram, boxplot)
  - Visualizing the relationship between two features in graphs (e.g., scatter plot, bubble plot)
  - Visualizing the composition of data in graphs (e.g., pie chart, stacked chart)
  - Visualizing comparison of data in graphs (e.g., line chart, bar chart)

- Building dashboard
  - Create a dashboard using IBM Cognos Analytics, refer to slides 11 to 14

METHODOLOGY



- Data Collection
  - Using APIs to collect jobs data and storing the data in an excel spreadsheet would use the Python libraries: request, pandas, and openpyxl
  - Using web scraping to retrieve data from a website would use the Python libraries: request, pandas, and BeautifulSoup
  - Exploring a data set by loading the dataset, finding the number of rows and columns, and identifying the data types of each column. The process would require Python library pandas methods (read_csv(), head(), shape, dtypes)
- Data wrangling
  - Finding and removing duplicated rows would need Python library pandas methods (duplicated(), drop_duplicates())
  - Finding and imputing (replacing) missing values would use library pandas methods (count(), value_counts(), idxmax(), fillna())
  - Normalizing relevant columns would see the columns case by case
- Exploratory Data Analysis (EDA)
  - Determining how data is distributed would require graphs like histogram plot
  - Finding the outliers in a dataset would need to calculate Q1, Q3, and Inter Quartile Range (IQR)
  - Finding a correlation between features in a dataset would need the Python library pandas method (.corr())
- Data Visualization
  - Using SQL knowledge
  - Using Python libraries (matplotlib.pyplot, pandas) to plot graphs and visualize distribution, relationship, composition, and comparison of data in a dataset
- Building dashboard
  - Using IBM Cognos Analytics to make various charts and assemble a dashboard, refer to slides 11 to 14

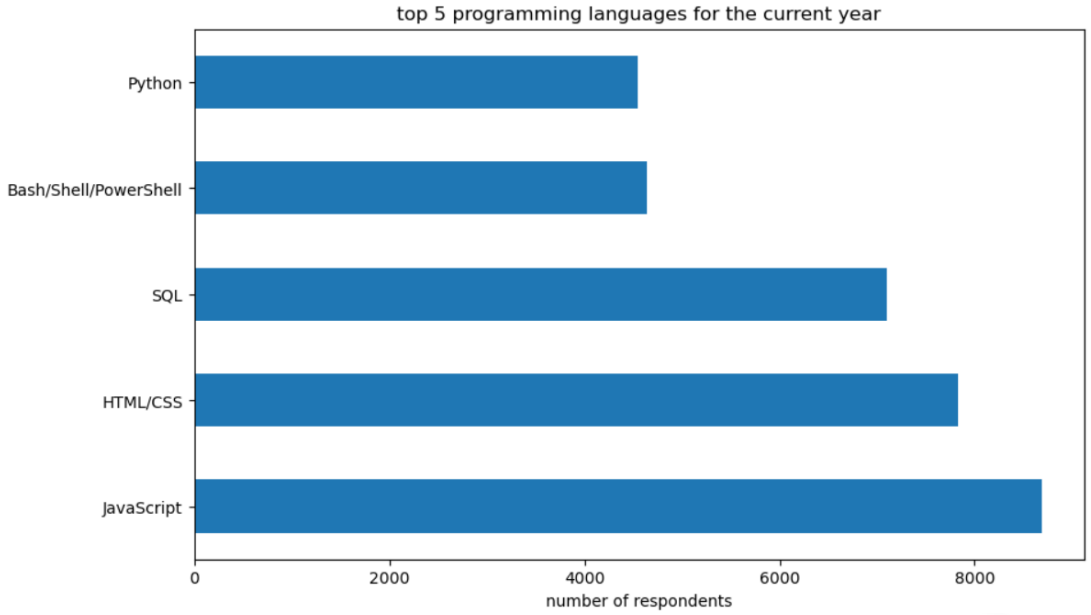
RESULTS

Please refer to the charts (slides 7 to 10) and dashboard (slides 11 to 14)

PROGRAMMING LANGUAGE TRENDS

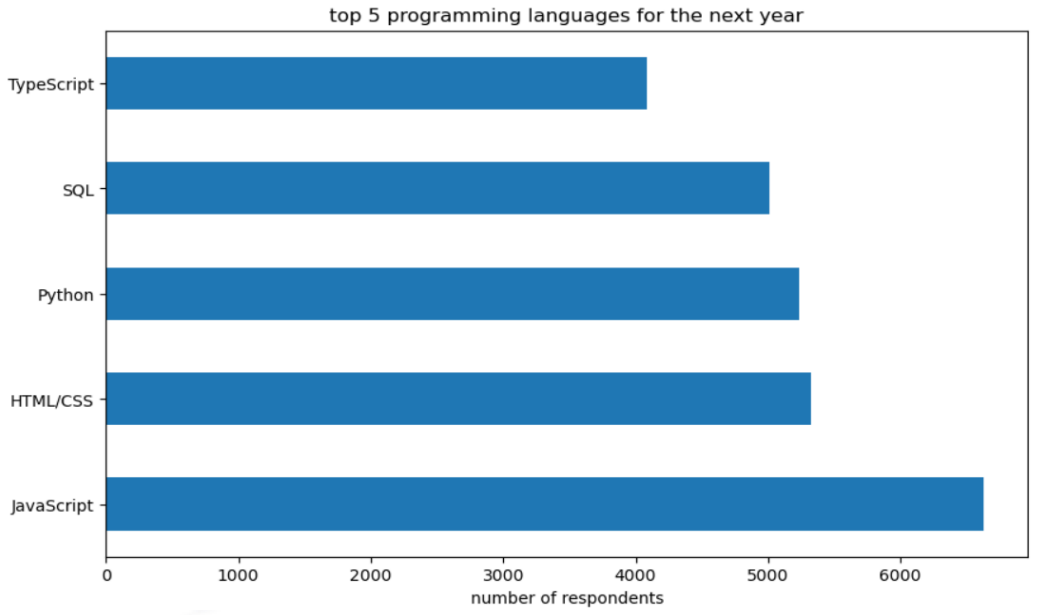
Current Year

Bar chart of top 5 programming languages for the current year:



Next Year

Bar chart of top 5 programming languages for the next year:



PROGRAMMING LANGUAGE TRENDS - FINDINGS & IMPLICATIONS

Findings:

- Finding 1: JavaScript and HTML/CSS are the top 2 languages in both current and next year
- Finding 2: Python improves from 5th in the current year to 3rd in the next year

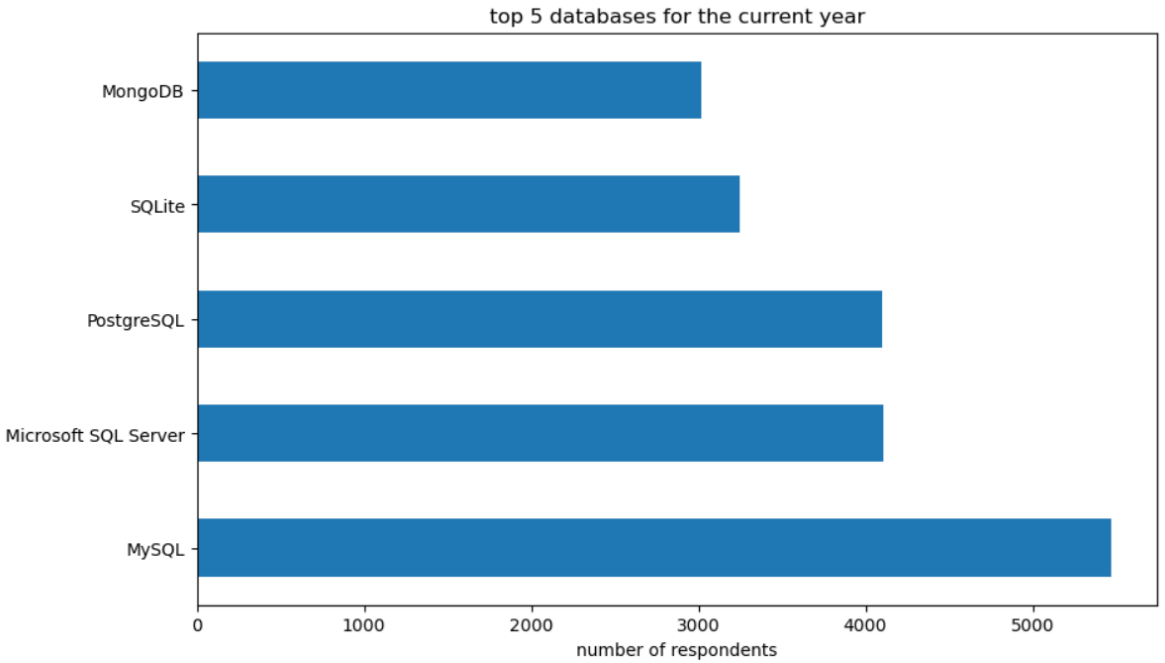
Implications:

- Implication 1: JavaScript and HTML/CSS are the most popular languages in these two years
- Implication 2: More people would like to learn or use Python

DATABASE TRENDS

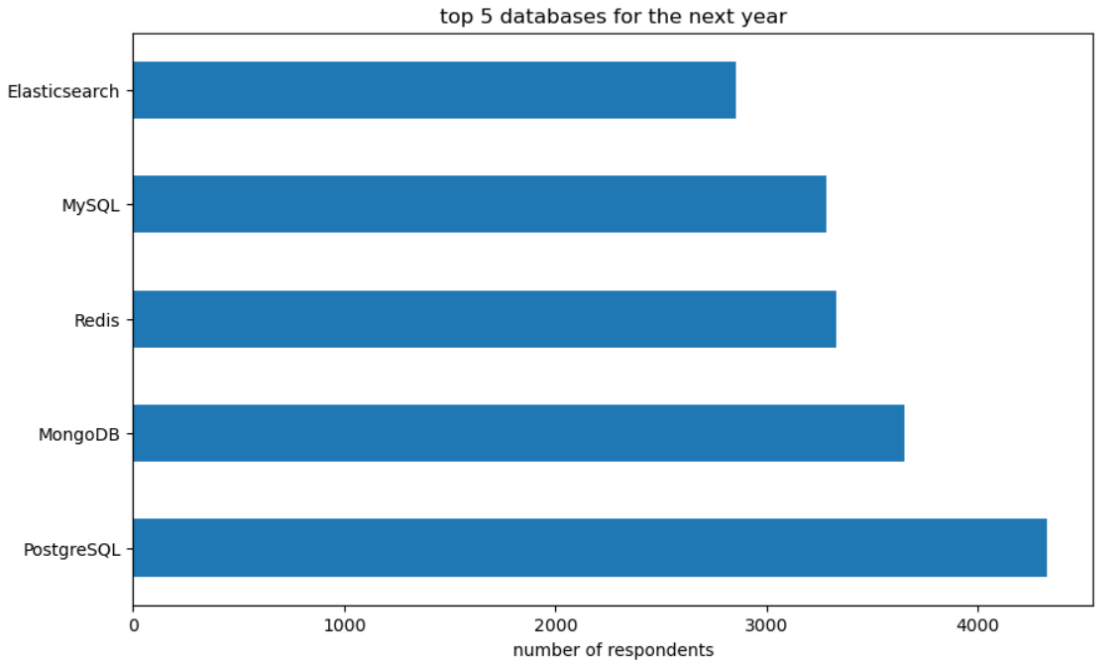
Current Year

Bar chart of top 5 databases for the current year:



Next Year

Bar chart of top 5 databases for the next year:



DATABASE TRENDS - FINDINGS & IMPLICATIONS

Findings:

- Finding 1: SQLite and Microsoft SQL server appear in the current year, but are replaced by Elasticsearch and Redis in the next year
- Finding 2: PostgreSQL improves form 3rd in the current to 1st in the next year

Implications1:

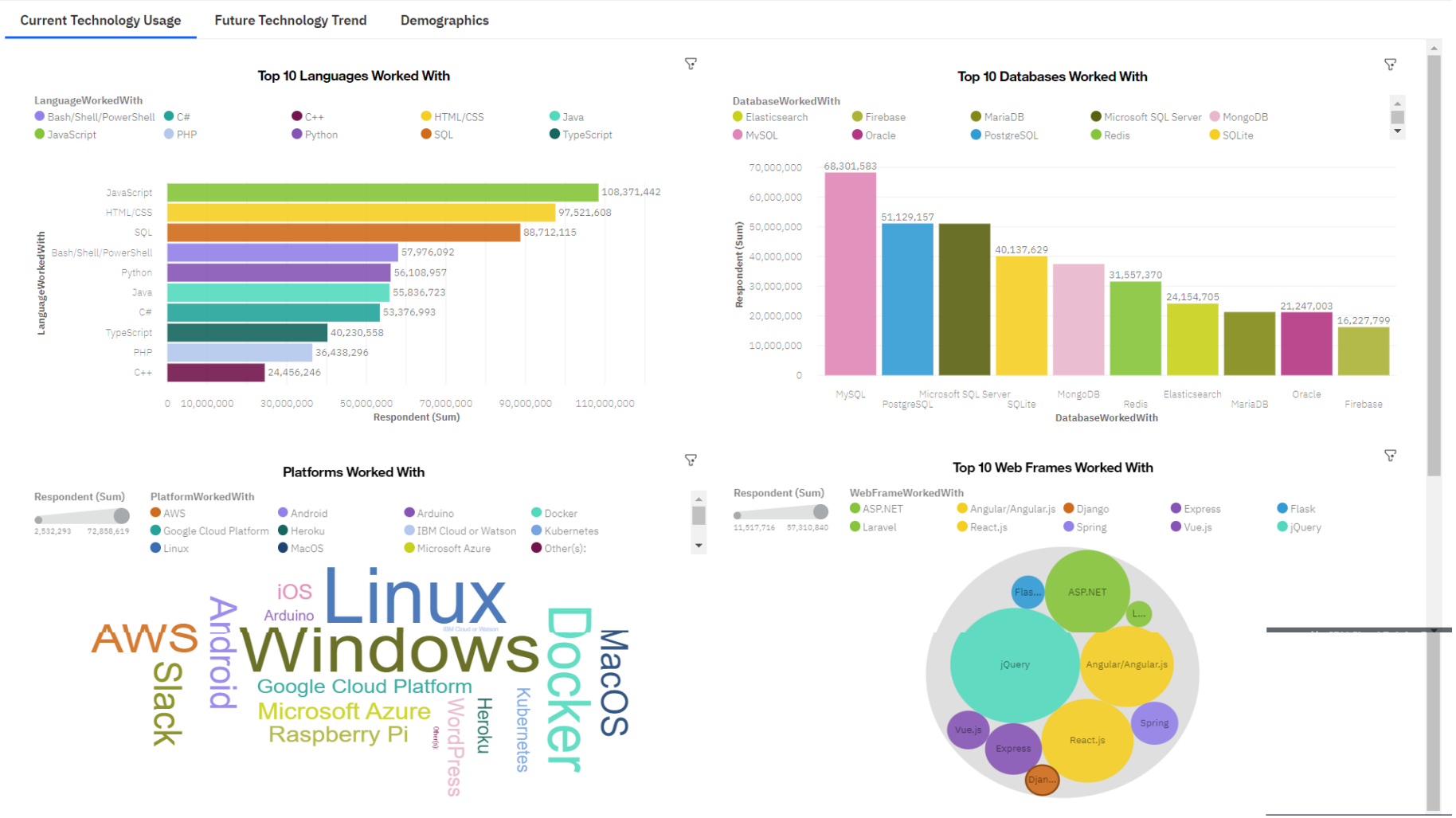
- Implication 1: The database market has high competition
- Implication 2: PostgreSQL becomes more accepted in these two years

DASHBOARD

The permanent link of the read-only view of the Cognos dashboard: https://jp-tok.dataplatform.cloud.ibm.com/dashboards/c8470fe4-f8bb-410a-9860-1e069448c565/view/4406c909648d3dfe67c5bde4079d2a027d35245dbbbbd55284d47b4909637497a83c41c5c82b4c5fd210076bf7bf115b9f

DASHBOARD TAB 1

Screenshot of dashboard tab 1 (Current Technology Usage):

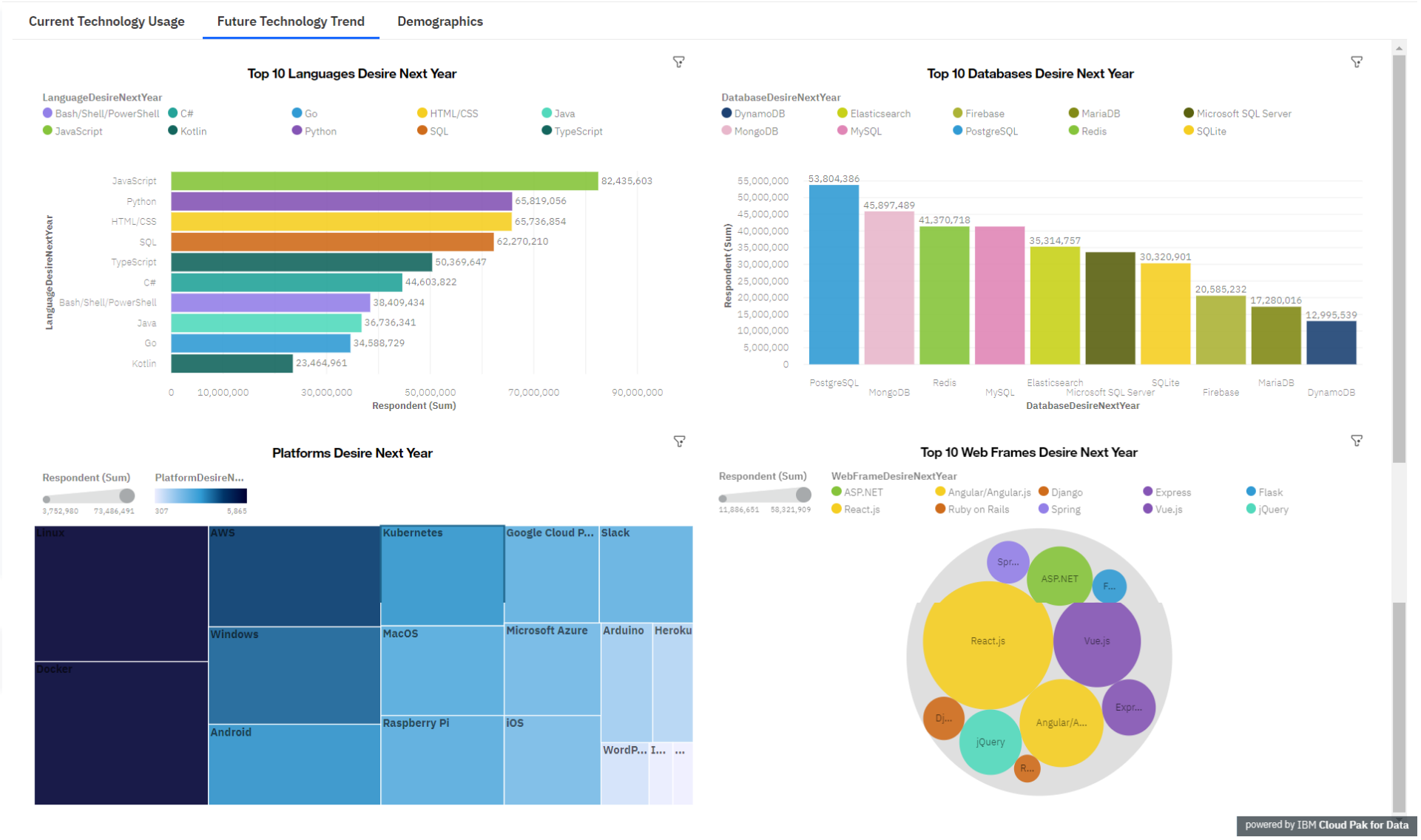


IBM Developer

SKILLS NETWORK

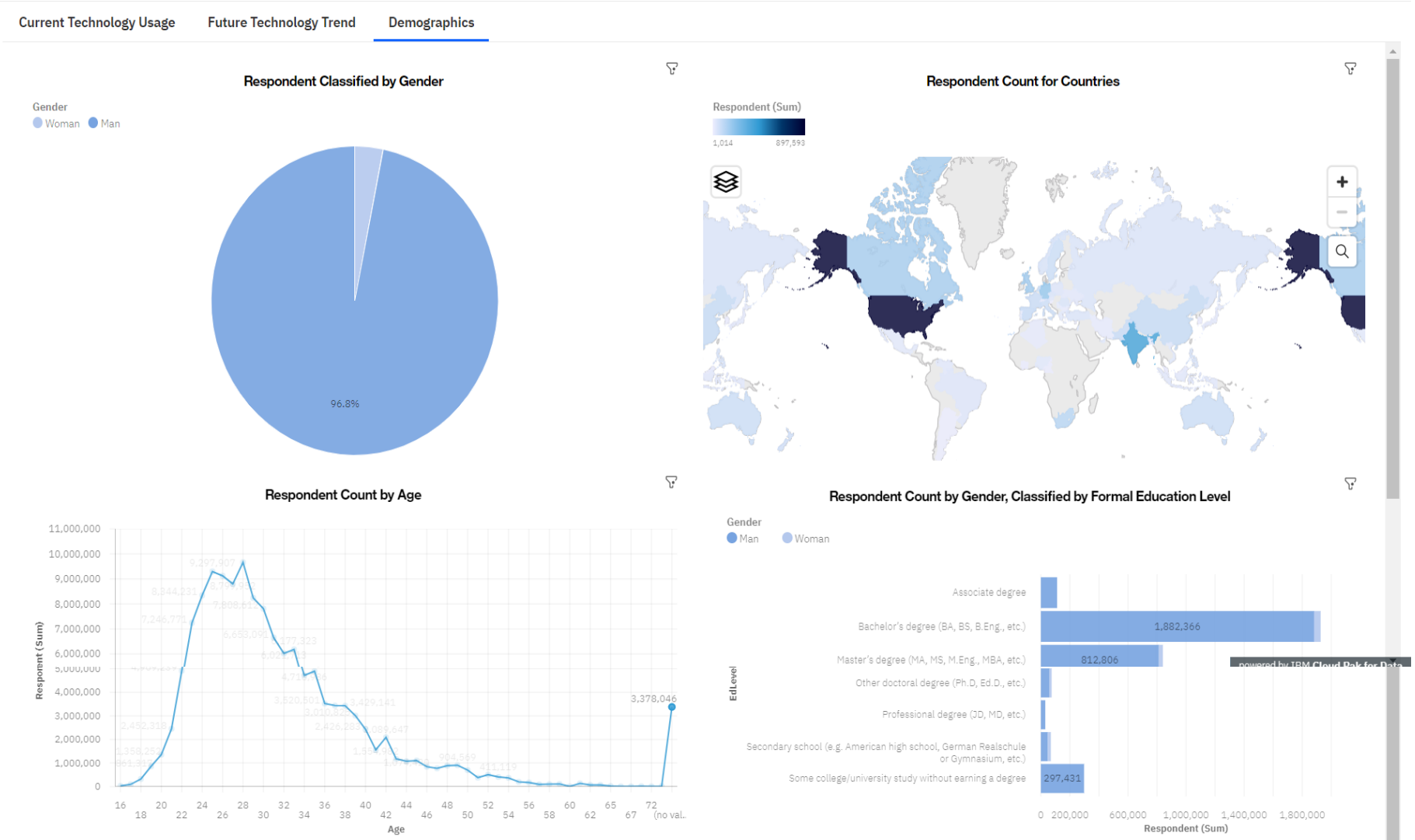
DASHBOARD TAB 2

Screenshot of dashboard tab 2 (Future Technology Trend):



DASHBOARD TAB 3

Screenshot of dashboard tab 3 (Demographics):



IBM Developer

SKILLS NETWORK

DISCUSSION

- Please refer to slide 16

OVERALL FINDINGS & IMPLICATIONS

Findings:

- Finding 1: Most respondents (96.8%) are male
- Finding 2: Most respondents are aged between 20 and 40
- Finding 3: Most respondents (1,882,366) have Bachelor's degree (BA, BS, B.Eng., etc.) no matter male or female

Implications:

- Implication 1: Male is the major workforce in computer science area
- Implication 2: People aged between 20 and 40 would more like to participate survey
- Implication 3: Bachelor's degree (BA, BS, B.Eng., etc.) is the most common education level for people working in computer science area

CONCLUSION

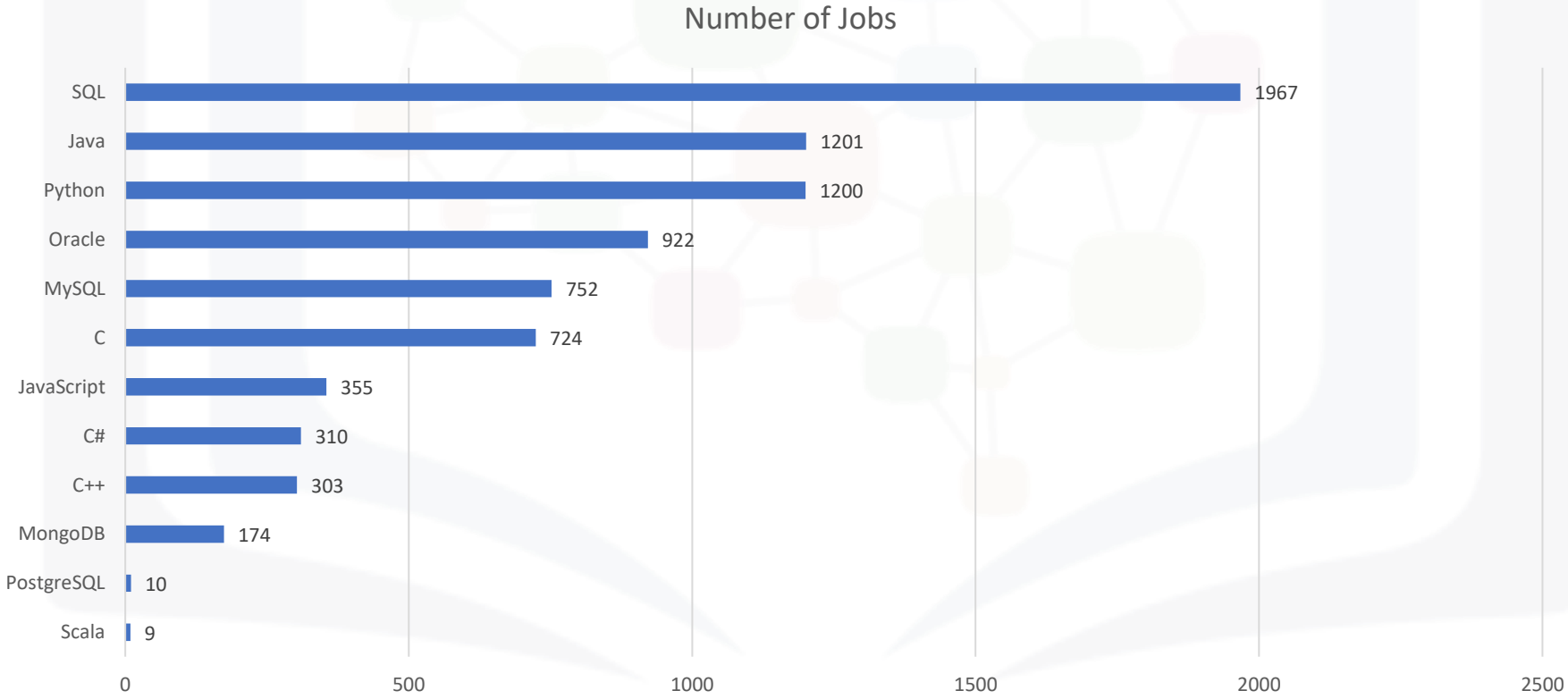
- Python language becomes more and more popular, and more people would like to learn or use it
- Male is the major gender in computer science area
- People aged between 20 and 40 would be the major workforce in computer science

APPENDIX

- Include any relevant additional charts, or tables that you may have created during the analysis phase.
- Please refer to slides 19 to 20

JOB POSTINGS

In Module 1 you have collected the job posting data using Job API in a file named “job-postings.xlsx”. Present that data using a bar chart here. Order the bar chart in the descending order of the number of job postings.



POPULAR LANGUAGES

In Module 1 you have collected the job postings data using web scraping in a file named “popular-languages.csv”. Present that data using a bar chart here. Order the bar chart in the descending order of salary.