CS432/532: Final Project Report

Project Title: Data Analysis on Stack Overflow Survey 2021 dataset.
Team Member(s): Kale Ritika, Patel Poojan.

I. Problem

We performed data analysis on Stack Overflow Survey 2021 dataset. We used the official dataset from the website.

The URL: https://info.stackoverflowsolutions.com/rs/719-EMH-566/images/stack-overflow-developer-survey-2021.zip

Or https://insights.stackoverflow.com/survey/2021

We have performed 4 data analysis on the dataset which are as follows:

a. Analyzing average hourly wages for developers in country.

b. Analyzing most popular programming languages among developers.

c. Analyzing most popular databases among developers.

d. Analyzing the percentage of respondents from each country in the survey.

II. Software Design and Implementation

A. Software Design and NoSQL-Database and Tools Used:

We used Python for Programming language and for database connection we used MongoDB. We did the code on Visual Studio code. We used matplotlib to represent it in graph form.

B. Parts that you have implemented

We have implemented the four analysis using aggregation pipeline framework for data processing and transformation in MongoDB. We both did 2-2 analysis each. We used NoSQL query and different aggregation like match, project, group, unwind, sort, limit for the attributes. We have connected to MongoDB using pymongo library. And read the data from csv file using pandas library.

III. Project Outcome

After importing the datset in MongoDB we created aggregation pipeline using pymongo library(to connect the NoSQL Database with Python. And then performed the analysis on the data.
The Analysis are explained below.

Analysis 1:
In analysis 1 we have performed analysis on the LanguageHaveWorkedWith attribute in the Collection. The first stage filters out the null values. The second stage performs unwind to separate each element. In further stage calculates the count for each group and then it is sorted in descending order based on the count. And we used matplotlib to represent it in the graph form. The analysis shows the top 10 languages which are popular.

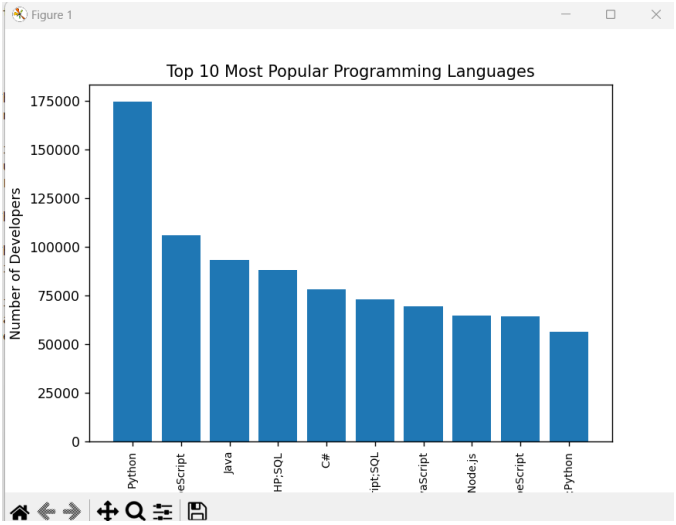

Fig a: Analysis 1

Analysis 2:
In analysis 2 we have displayed the average hourly wages for developers in the Carribean country. It shows the top 4 countries in Carribean which has highest wage. It is represented in a pie chart graph .

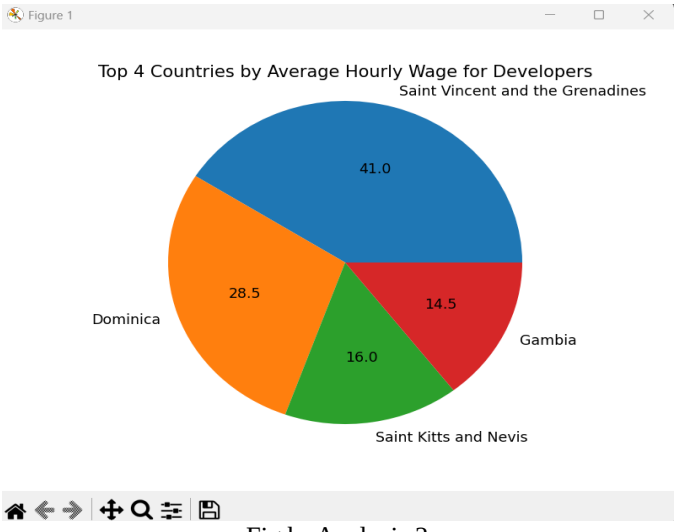

Fig b: Analysis 2

Analysis 3:

In analysis 3 we have derived the top most databases used by individuals in a group . The graph below displays that. (Fig c)

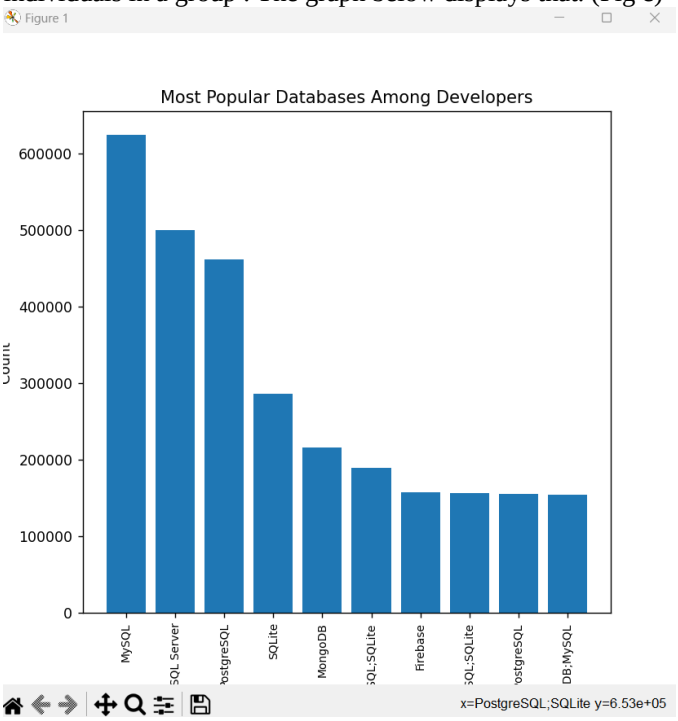

Fig c: Analysis 3

Analysis 4:

In analysis 4 we have calculated the percentage of respondents from each country in the survey. The graph below shows that United States of America has highest respondent who participated in the Survey followed by India.

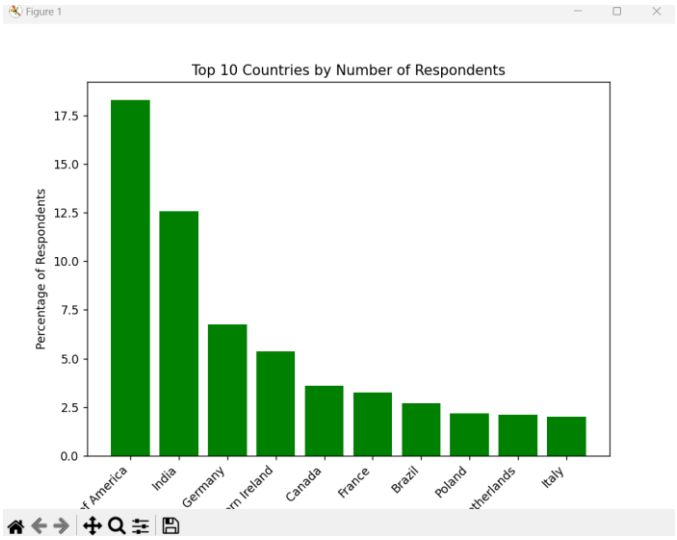

Fig d: Analysis 4

REFERENCES

[1] https://insights.stackoverflow.com/survey/2021

[2] https://numpy.org/doc/stable/

[3] https://www.geeksforgeeks.org/bar-plot-in-matplotlib/

[4] https://matplotlib.org/stable/tutorials/index

[5] https://www.geeksforgeeks.org/plot-a-pie-chart-in-python-using-matplotlib/

[6] We also reffered on google for how to connect to database and how to use aggregate pipeline.