Student: Caroline Araujo

Notes on the "Business Challenge: EDA and SQL" Project

1. Dataset Selection

The dataset was sourced from Kaggle, created by Narmelan Tharmalingam and last updated four years ago. I chose this dataset because it provides an extensive overview of the 100 most spoken languages in the world, which is relevant for exploring global language trends.

2. Data Cleaning Process

I used Python for data cleaning. First, I imported necessary libraries such as pandas and numpy. I then read the file and started a preliminary exploration. I noticed that there were spaces in the column names, so I cleaned this up by removing the spaces and sorting the data for better readability.

3. Data Exploration and Visualization

I performed some exploratory data analysis (EDA), visualizing the top 10 languages by the number of total speakers. I used matplotlib and seaborn to create bar charts and correlation matrices to check the relationship between native speakers and total speakers.

4. Business Questions and SQL

I imported the dataset into SQL to answer business-related questions such as:

- What are the languages with the highest number of native speakers?
- What is the geographical distribution of these languages?

5. Python vs SQL Comparison

One interesting aspect of the project was comparing the results obtained from SQL with those generated in Python, observing the strengths of each tool for different types of analysis.

6. Conclusion

This project provided a comprehensive view of how EDA and SQL can be combined to generate business insights effectively, using real-world data.