PROJECT OVERVIEW

I. INTRODUCTION

In this project for the capstone framework of our master's program in data science, DTSC-961, we propose to work on some statistics concerning the COVID-19 pandemic. To begin with, we intend to limit our analysis to a particular period. This is why we will analyze data from January 22, 2020, to July 27, 2020, on the COVID-19 pandemic. Our data are from kaggles.com, which is affiliated with the World Health Organization (WHO), which is its acronym.

II. II. THE PLAN OF OUR ANALYSIS.

1. Cleaning, manipulating, and analyzing data

We started by doing a deep clean of each data table, i.e. 8 in our database. To accomplish this task, we used a Jupiter notebook with Python to clean the data, handle datatype conflicts in certain columns, and manage missing data. We used different notions of Python programming in order to accomplish this task. Once each data table had been cleaned, we saved the modified table as a new csv file.

2. 2. Database

In order to manipulate our data more advancedly, we used PostgreSQL with PGAdmin4. We created three sql files, the first 'Databases_import.sql' created each table and placed the appropriate constraints on the tables. The first of these files was 'Databases_import.sql'. Secondly, we developed 'import_datasets.sql', which enables us to import data from our .csv files into our PGAdmin tables. Finally yet importantly, we created the file 'Database_queries.sql' that enabled us to query the databases and carry out sufficient analysis to obtain answers to our business questions, as described in our project proposal.

3. 3. Visualization

We illustrated our business questions with Tableau and developed a storytelling approach to argue our results with simple, easy-to-understand graphics.

III. DIFFICULTIES

I want to emphasize that we encountered some difficulties during our project. Indeed, along the way, I always checked the numerical results with different official sites on Google. However, putting the tables together using the JOIN method resulted in the data becoming meaningless and the numbers being significantly skewed. As we stated above, we wanted our results to reflect reality within the limits of reality. Ultimately, we did not need to put all eight tables together to answer our business questions.

IV. IV. CONCLUSION

Our DTSC-961 capstone project offered insights into the COVID-19 pandemic via meticulous data analysis from January 22 to July 27, 2020, using Kaggle data affiliated with WHO. We conducted thorough data cleaning and analysis in Python, leveraging PostgreSQL for advanced manipulation. Despite challenges, we ensured data integrity. Through Tableau, we presented our findings effectively. Our project highlights the significance of robust data analysis in addressing crises like COVID-19, contributing to informed decision-making and future research in data science.