Technical Report For ETL Project

For this project, the subject of interest is corona virus (COVID19), specific to the New York and Texas population. The data included the number of confirmed cases, the number of deaths and the number of recovered people and was sourced from Kaggle and Data.org. Original files were CSV-formatted. To better analyze the data, we chose to use Pandas, a library associated with Python, to clean as well as modify the data if needed. Our group first loaded each data table into DataFrame and then used a method called info(). This initial step helped us understand the data information and data type given in each file as well as identify any null values in each data set. The data was then narrowed down to the aforementioned subset subject interests; number of confirmed cases, the number of deaths and the number of recovered people limited to New York and Texas. Once all parameters of interest were selected we then cleaned up the data, this included column renaming to avoid postgresql errors and exceptions for capitalized words in the column headers. After finishing with the cleaning process, we employed sqlalchemy as a connection between Python and postgresql (our main database). The reason we chose postgresql was because it is a relational database management system and most importantly our data tables are related to each other. After setting up the connection, our group came to pgAdmin4 to create a database that corresponded to the database declared when we created our connection. There were three tables created in postgresql corresponding to three data tables we already had in jupyter notebook. To make sure our dated loaded into the database successfully, we ran some queries both in jupyter notebook and or in pgAdmin4.