**ETL Project Report Group 6**

**Extract:**

We found our COVID dataset from the Data World website and its original file was formatted in CSV. This dataset included the total cases, total deaths, new cases, new deaths from 2020 to 2021. Furthermore, we found our World Happiness Report on the Kaggle website and its original was also formatted in CSV. This dataset tracked global happiness from various countries around the world for the past year and included metrics such as social support, healthy life expectancy, perceived freedom to make life choices, and perceptions of corruption. The responses were collected from the Gallup World Poll.

**Transform:**

Our transformation included cleaning and analyze the data from these four csv files that we downloaded our data sets: full\_data.csv, population\_by\_country\_2020.csv, world-happiness-report-2021.csv, and world-happiness-report.csv. We also cleaned any irrelevant columns, dropped duplicates, and formed the DataFrames (df). The technologies used were Jupyter Notebook (Python 3) and PostgresSQL database. The first step in our transformation was to utilize the function that read the csv files into the df. The second step was to merge the df on the covid\_happy.ipynb file. We then cleaned the merged df by dropping NaN rows which were rows that had empty data. Afterwards, we cleaned the full\_data.csv file and selected five columns which were year, location, new\_cases, new\_deaths and total\_cases. Moreover, we cleaned the world-happiness-report-2021.csv and selected the year, country, life\_expentency, and life\_freedom columns. Finally, we exported these dfs into SQL tables, and we developed one final table to show all the data together.

**Load:**

We faced a few challenges along the way form cleaning the data and exporting it into SQL. An obstacle that we initially encountered was figuring out how to find the latest date for each country since we released the total COVID cases were being added on as the dates passed. Our objective was to figure out the latest date for each location which would have the final total of the COVID cases that were looking for. The final database was loaded into SQL named covid\_happy. We queried the table called full\_21 which housed the year, country, total cases, and total deaths from COVID which is what we created in our transformation stage in jupyter notebook. Our other queried table from the juptyer notebook is called world\_21 which includes the year and countries along with the global happiness report columns of life expectancy, life freedom, and perceived corruption. We then joined the tables into one final table through the countries because we wanted to see how the number of cases and deaths from COVID throughout 2020 and 2021 affected the happiness report within each country, including the life expectancy of the populations and the perceived freedom to make choices from the polcies or legislations that were put in place due to the pandemic.