**ETL PROJECT REPORT – TERRORIZER – Ryan, Nandini, Steve**

**EXTRACT:**

Our data came from three main websites: ourworldindata.org , data.worldbank.org , info.worldbank.org

The info.worldbank.org website had the six WGI (Worldwide Governance Indicators) datasets. These were all in one Excel workbook with a tab for each indicator. These all had to be separated out into one tab each so they could be saved as csv files

The ourworldindata.org website had the main terrorism data and attacks across multiple years. This had to be split into two files due to size issues not being able to be uploaded into github or merged into pandas. These files were also in excel format and had to be converted to CSV

The data.worldbank.org website had the population and GDP information for all our countries across multiple years. Both files came in a CSV format.

All files were put into a resources folder inside github for further ease and access

**TRANSFORM:**

Multiple issues arose when trying to read the csv files (after converted from excel previously).

Regarding the population/GDP files, there was problems with some of the characters in the set, which required some investigation and fixing using a “from chardet import detect” function. The tab delimiter also needed to be used instead of the traditional comma. An outer merge was then used to combine the data frames by country name. Following that was some dropping and renaming of columns. The combined dataframe was then saved as a CSV file in order to be put into the database. Following that was some dropping and renaming of columns. The combined data frame was then saved as a CSV file in order to be put into the database. Column names were renamed to identify if the year was from GDP or Pop and unnecessary columns dropped. The cleaned-up data frame was exported to a new CSV file and inserted into MongoDB.

In regards to the WGI files, after they were converted into 6 different csv files, the files had to be cleaned up in order to be readable through pandas. Blank rows and headers were removed to get to the main headers to be used. Columns were dropped so as only to show the “estimate” values and not the other standard deviation, upper, lower values. A few unnamed columns were given appropriate names. All 6 dataframes were all merged through an outer join on country name across multiple merge efforts. Duplicate columns were then dropped, and a few columns were renamed to better cleanse the dataframe. The final dataframe was then saved as a CSV file to be added to the database.

**LOAD:**

A dataframe was created in MongoDB to add in all the data. Multiple collections were created for storage. The decision to have an unstructured/unrelational database was made after reviewing the different datasets we had. The main terrorism data records was much larger in size (100+ mgb) then the other data. Also, for the terrorism data, each record of attack was on a single row that included a country, year, month, etc. of data. On the flip side, the WGI, Population, GDP data had countries counting down row by row and each year was on top as a column going across. Merging the two would prove difficult and very time consuming, as well as resulting in multiple duplicate values being put into the database

The final check was done in Jupyter Notebook to confirm that the mongo data base was connected and there was data in the database.