ETL Project

Shahzad Ansari, Jacob Greenbaum, Ronald Rivas

**Extract:**

We found both our data sets from Kaggle, since Kaggle is well designed we were able to download the files in a clean csv format without having to convert file formats

**Transform:**

All manipulation was done within python before being uploaded into the SQL server, the first data set labeled crime\_df. In crime\_df we extracted the desired columns ‘timestamp’ , ‘ unique key’ , ‘description’ and put them into a new data frame called clean\_crime\_df. We dropped all null values and converted the dates into date time objects. Finally, we renamed the timestamp to be in more line with the rest of the data and called it date. We performed the same operations to the weather\_df, dropping undesirable columns and in the end merging them into one data frame called “merged\_df”.

**Load:**

In PGAdmin we created a new database called “ETL” and a new table called “crimeByweather” with all the necessary columns needed Using SQL Alchemy we created a connection with PostGreSQL via sqlalchemy create\_engine and converted and uploaded the data frame via the to\_sql method.