**Extract Transform Load**

**Report**

***Extract***

We found two csv files on Kaggle.com after researching through both Kaggle and Data.gov. For our topic we chose crime because we were able to use the DR Number as our primary key (column). After looking through the datasets for a while and trying to find data from different sources, we had to settle for two csv files that both included the DR Number. The first csv file is traffic data from Los Angeles and the second csv file is crime data from Los Angeles. We downloaded and saved both files onto our local directories and prepared for transformation.

***Transform***

For the transformation stage of the project, we decided to use Pandas to clean up the data. We imported Pandas onto jupyter notebook to read the csv files and create data frames out of them. When looking at both the data frames we were able to establish that some of the columns needed to be dropped in order for the data to make more sense. We dropped the unnecessary columns and concatenated the two data frames together by row as they both now have the same column names as well. Once we aggregated and transformed the data, we exported the data frames back out to csv files and saved them in our ETL folder for loading.

***Load***

To start the loading part of the project we decided to utilize postgresql, pgAdmin 4. In order to import the cleaned-up csv files onto postgres, we had to create respective tables and columns for both the csv files. The import would fail giving us various errors otherwise. We had to clean up the data in pandas, word pad, and excel in different ways to get rid of some of the errors, i.e. “extra data after last expected column”. Once the csv files were successfully imported, we established that the ETL process was complete and we could analyze the data and compare the two data sets.

***Report***

We uploaded all of our finalized files to github and are finalizing our report.