ETL Project Report

Group #2

Katherine Morriss, Cindy Pappas, Ergin Bostanci, & Matt Mullinazi

For the ETL Project, two CSV files were chosen. The first dataset reflected data for crimes in Los Angeles and was downloaded from Kaggle. The second dataset reflected data for crimes in Kansas City and was downloaded from Open Data KC. The data was not collected in the exact same manner in each city, so cleanup was needed to keep the tables consistent with each other. The idea was to collect this data as if a group were wanting to be able to compare crime between the two cities.

Jupyter notebooks was used to read in the CSV files as pandas data frames. The cleanup involved changing column names since they had similar data but different labeling for each source. We dropped duplicate entries based on the incident ID column and created a column that reflected which crime data belonged to each individual city. MySQL was used to create queries for the two different data files.

To load the data, we created a connection with the database by using the local host and then the create engine function. After confirming the two tables in MySQL, we loaded the data frames by using the “.to\_sql” function. We ran into a problem here. The connection would cause an error and drop our primary key. We did some troubleshooting and found the reason for the error was because we had the incident\_id set as our index. Once we changed this, the connection filled through properly.