Brief Introduction

The ETL-project was completed by three team-members. Micheal Pearson, Moumita Ghanti, and Ghaith Ramahi. We decided to Extract, Transform, and Load retrieved data we managed to obtain from respected sources like CA.gov Open Data Portal, as well as Kaggle Data Set.

Firstly, we managed to obtain all the data and analyze it in regards to the state of California. Moumita performed an Extract, Transform, and Load data concerning crime rates and gross domestic product by county in the state of California. Meanwhile, Micheal performed an Extract, Transform, and Load data in regards to the earthquake data in the state of California. Lastly, Ghaith also performed performed an Extract, Transform, and Load data concerning traffic jams in the state of California.

Extraction, Transform, & Load Process:

1. Ghaith Ramahi, relied on the CA.gov open data portal to extract, load, read, and clean the csv file in Pandas and Jupyter Notebook. The data was cleaned as well as particular columns were pulled and extracted into a final data frame. All three csv files representing the years 2015, 2016, and 2017, of traffic data, were all combined, cleaned and placed in one data frame. The Transformation of particular sets was then used to finalize this step.Thereafter, an engine was created and established a connection to pgAdmin SQL. Furthermore, the table was populated the table with data obtained from the Data Frame.

2. Moumita Ghanti, relied on the CA.gov open data portal to extract, load, read, and clean the csv file in Pandas and Jupyter Notebook. The file was obtained from CA.gov concerning the crime data from 2000 to 2013. Thereafter, it was loaded in the Jupyter notebook, the certain columns were extracted from the original file and then transformed into data reflecting the state of California. The data was cleaned and formatted accordingly. Thereafter, certain mathematical calculations were applied to one of the columns to calculate the exact rate for various crimes as well total crime rate per year. Thereafter, an engine was created and established a connection to pgAdmin SQL. Later, the table was populated.
   Similarly, the Gross Domestic Product data(from 1995 to 2016) was retrieved from the CA.gov website. A similar process was used as the previous and then cleaned as there was similar values with multi label descriptions.Columns splitting is done of the coordinate level to retrieve latitude and longitude data separately. Then data was transformed. Thereafter, the Data was loaded.

3. Micheal pearson, relied on USGS-api, to try and retrieve the California Earthquake data. However, the api response came with a lot of dirty data, which was not the best source for cleaning it. Hence, the time frame, it was not possible to perform all the cleaning practices and further finalize the extracted clean data. Therefore, Kaggle was later on utilized to perform the extraction process. For the transformation process, certain columns were used and later on placed in a data frame. The Google

Developer api was practiced to try and clean the data, but ended up not working. The rows were narrowed down to those whose coordinates were within the range of California’s latitude and longitude. Thereafter, an engine was created and established a connection to pgAdmin SQL.

Sources used:

1) https://data.ca.gov/dataset/annual-average-daily-traffic-volumes
2) https://data.ca.gov/dataset?q=traffic
3) https://data.ca.gov/dataset/violent-crime-rate/resource/91e7c556-54cc-4848-8811-500137d5ede2
4) https://www.kaggle.com/usgs/earthquake-database
5) https://earthquake.usgs.gov/fdsnws/event/1/#parameters
6) https://data.ca.gov/dataset/b-7-adjusted-gross-income-by-county