**ETL PROJECT**

This is a project by Christy Dafonte, Fred Castrow, Mustafa Ancin. In this project we aimed to chase a relation between the border crossings and the incarceration rates in the border states of the USA. We will be looking into the states up north on the border with Canada like Maine, Vermont, New York, Idaho, Michigan, etc. We will also look down to Texas, Arizona, California, etc. From these states, the major cities that has access ports through the border will be taken into consideration.

**Extract:**

In order to extract the required datasets we dived into some websites and had our csv files from [www.kaggle.com](http://www.kaggle.com) , <https://data.world/> mainly.

The “Crime and Incarceration Rates” dataset had all the states so we pulled out the border states only. The data provided info between the years 2001-2016. This led us to go through these years with our second data which is “Border Crossing Entry”.

“Border Crossing Entry” is a huge dataset with a vast information sets such as type of entry (by car, walking, by train etc.), types of vehicles cars, trucks, trains, ship, buses, etc.

**Transform:**

We had to do some cleaning and get rid of the unnecessary data such as location which was given on a longitude and latitude basis. Since there are more than one ports of entry in a state, we had to sum the values of each ports to be able get a number on the state basis.

Then we had to transform the number of people to an understandable value from “Value’ to “Crossing Border Count”.

The dates were given in date and time in our dataset where we had to transform them into day month year format.

“The Prison Custody” dataset contained all the states, from that set we had to extract the border states.

Once this was done, we groupby’ed the states and the dates (years).

On the Crime and Incarceration Rates side we cleaned the .csv file, got rid of the N/A values and blank cells. Later we scraped the years and took out the states that are not on the north and south borders. Once these were done, the two files were ready to be merged and groupby’ed.

**Load:**

At this point to make sure we have the successful merge and flawless data output, we created a sql database called “ETL\_Project”. We used SqlAlchemy to create an engine that would connect the dataset to the database. Then we used python codes to append the data frame to the database. Then we ran a Sql Query to confirm the data was created in the database.