

US Borders  
Crossings vs. Captures

[Phil Lee](https://www.linkedin.com/in/escphillee/) | UC Davis Data Analytics | 12/22/2019

# Summary

My data analytics project looked at border crossing data at all US border stations. The goal was to determine the highest volume stations and highest volume measure types. The analysis also looked at illegal immigrant arrest volumes within the US. The goal was to determine the highest volume stations and types. The last piece of analysis asked the question: is there a correlation between the highest volume border crossings and illegal immigrant arrests? I used Excel, SQLite, Jupyter Notebook, Pandas, SQLAlchemy, and matplotlib to complete this work.

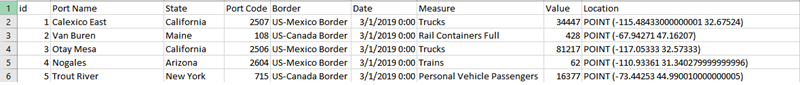
Special thanks to [Akhil Vasvani](https://www.linkedin.com/in/akhil-vasvani/) and the [US Customs and Border Protection](https://www.cbp.gov/) for use of their data. I did not include all the code from my Jypyter Notebook in this report. Here is the link for the [git depository](https://github.com/Phillee58/13-ETL-Project/blob/master/borders_PKL.ipynb), should you wish to reference it.

## Extract & transform

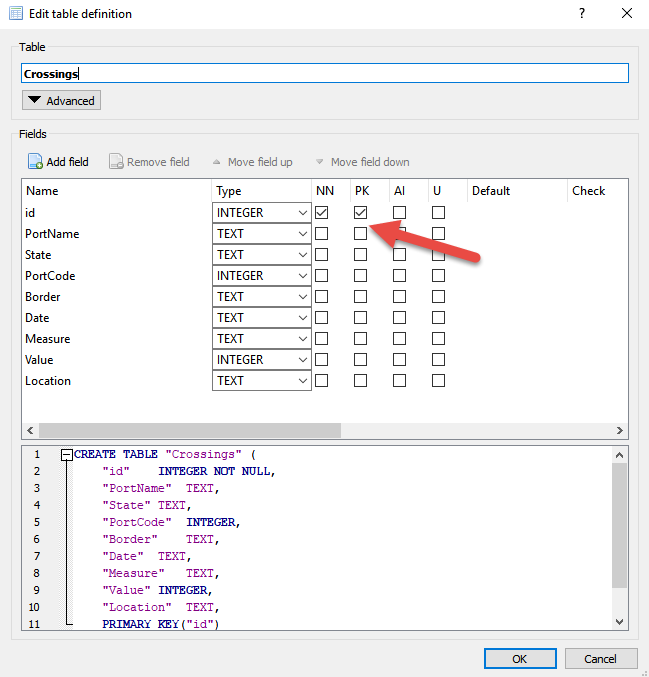
### Data Source 1 – <https://www.kaggle.com/akhilv11/border-crossing-entry-data>



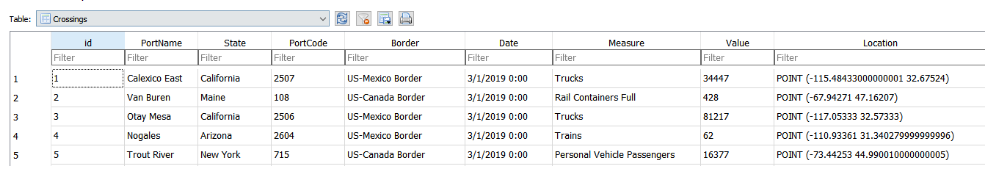
This is a snapshot of the provided Arrest.csv file I fetched from Kaggle.com and Akhil Vasvani. I cleaned out all data with 0 in Measure field. I created the id column and auto numbered 346,733 rows of data.



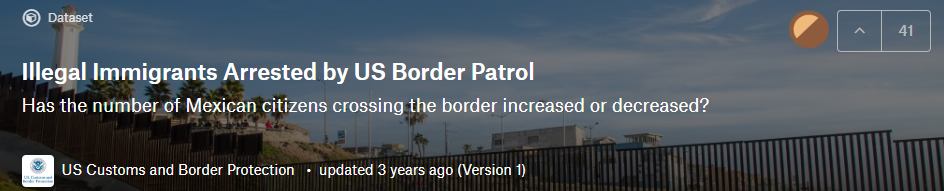
I used SQLite to create the Crossings table and imported the data from the CSV file. I updated the id field checkboxes on for NN and PK, since SQLalchemy AutoMap to run.



Here’s a snapshot of the data...

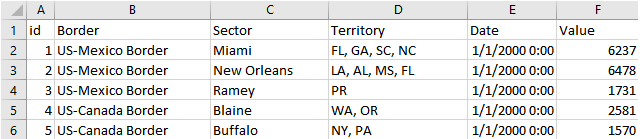


### Data Source 2 – <https://www.kaggle.com/cbp/illegal-immigrants>

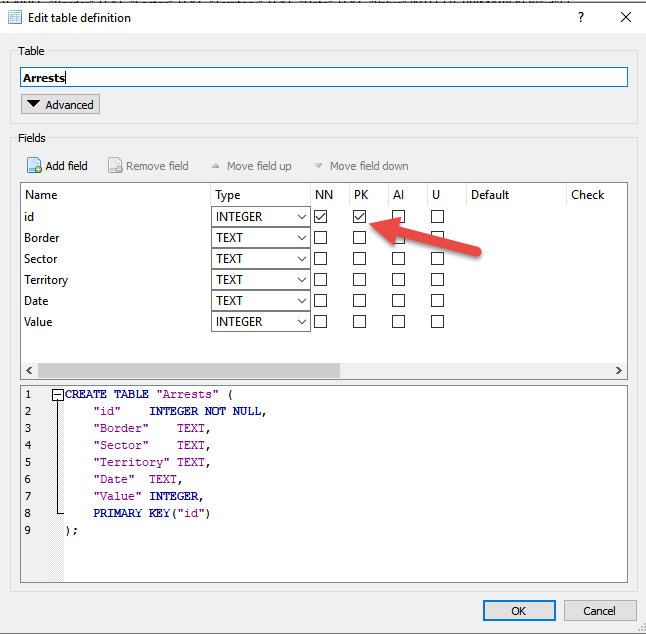


This is a snapshot of the provided Arrest.csv file I fetched from Kaggle.com. I had to clean this data because the provider had placed year totals in rows, so I had to remove them. He also placed the yearly totals in separate columns, so I had to aggregate them into one column and create a date field for it to work. I needed to decide which group in the two tables to use and I settled on using the one from the crossings table and updated it in the arrest table. I created the id column and auto numbered 340 rows of data.

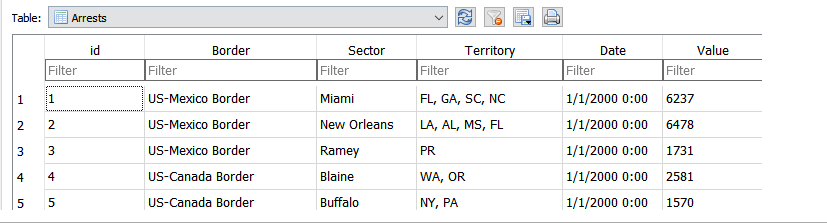
I include the correct [(Legal Definition)](https://www.cato.org/blog/illegal-alien-one-many-correct-legal-terms-illegal-immigrant) of Illegal Immigrants in the link to the left, should you be interested.



I used SQLite to create the Arrests table and imported the data from the CSV file. I updated the id field checkboxes on for NN and PK, since SQLalchemy AutoMap to run.



Here’s a snapshot of the data...



## Load

These are the dependencies I loaded:

*import pandas as pd*

*import datetime as dt*

*import sqlalchemy*

*from sqlalchemy.ext.automap import automap\_base*

*from sqlalchemy.orm import Session*

*from sqlalchemy import create\_engine, func*

*import numpy as np*

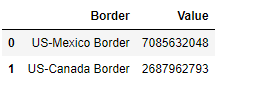
*%matplotlib notebook*

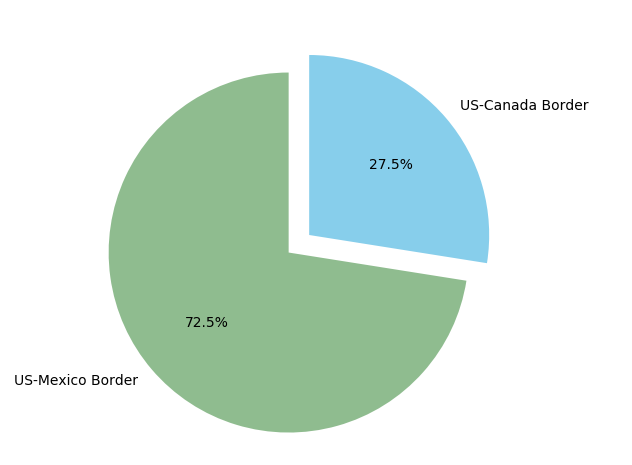
*import matplotlib.pyplot as plt*

## Analysis

#### Crossings:

# Query total crossings by common border

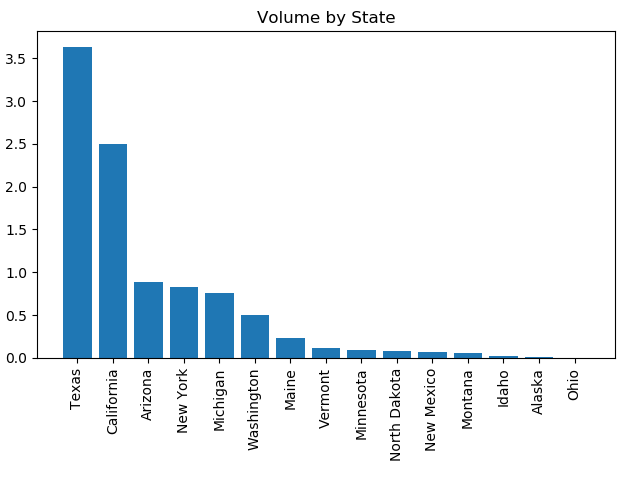




There are almost ten billion crossing per year across both the Mexico and Canada borders. There are roughly three border crossings over the Mexico border for every crossing over the Canadian border.

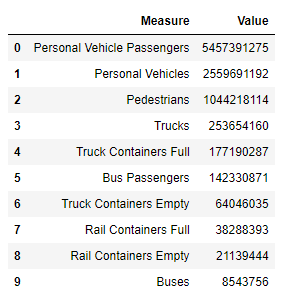
# Query total crossings by state

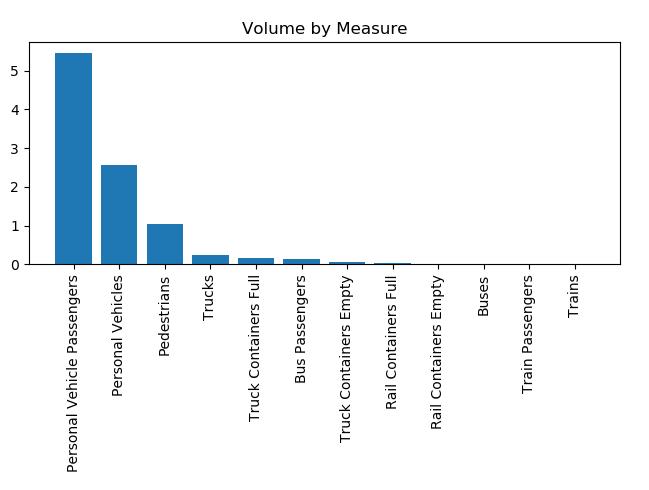




The states of Texas and California make up most crossings per year. Five other states have a small contribution of crossings and the remainder virtually no activity.

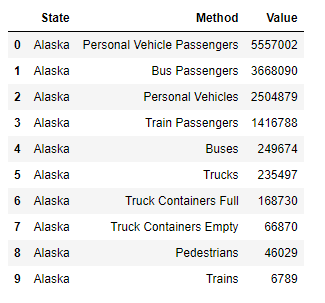
# Query total crossings by measure





Here we are looking at a breakdown of the measure, or type of crossing. Interesting to see that Personal Vehicles, Passengers, and Pedestrians greatly outnumber commercial transportation.

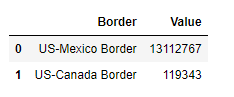
# Query total crossing measures by state

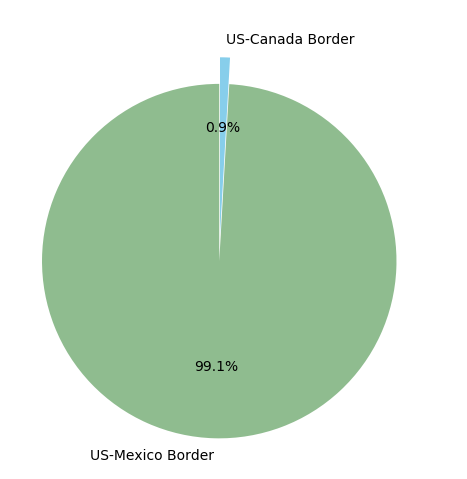


No real finders were found here. However, this is an easy way to zero in crossings for a state.

#### Crossings:

# Query total arrests by common border crossing

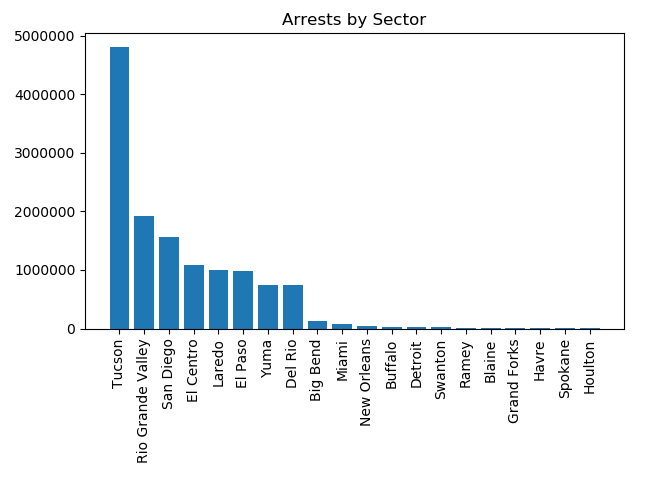




99% of arrests came from the Mexico Border. There is almost no arrest activity on our northern border with Canada.

# Query total arrests by sector crossing





The crossings dataset had stations grouped into sectors. This plot shows the sector data for arrests I found the activity clearly concentrated within California, Arizona, and Texas.

## Conclusion

It is blatantly obvious that the greatest concentration of crossing and arrest activity occurs along our southern border between San Diego and Laredo. This is really no reason to protect our northern border, even though we hear rumblings of this from the news media.

There appears to be a correlation between crossing and arrest activity, however, my data was limited, and I was unable to statistically make a case for this. The dataset provided by the US Customs lacked some key data elements, for example, sector data that would have made this correlation possible.

Overall, I learned some things from this project.

* How to find data from services like Kaggle.com
* How to take a csv file and transform it into a working SQLite file.
* How to use SQLAlchemy to map the files.
* How to group data and create data frames in Juptyer Notebook and Pandas.
* How to create pie and bar charts in matplotlib.