Computing for Analytics: Group Project (Team 8)

us Border Crossing

Agraj Agarwal

Anjali Hegde

Juliet Zhang

Keerthi Pullela

Sobhagya Singh Jamwal

Srishti Bisen

**Introduction**

This project aims to draw insights about the US-Mexico and US-Canada border crossings based on the data provided by the Bureau of Transportation Statistics (BTS) ranging from 1996 to 2018. The data is broadly divided into passenger traffic and commercial traffic.

**Hypothesis:** The crossing traffic at the US border, for both Mexico and Canada, has a trend that can be predicted for the future.

In the recent times, issues corresponding to border crossing have been in the spotlight for various political, economic and humanitarian reasons. With the ongoing conflicts and controversies, it would be interesting to analyze this data and see if the results are coherent with our general assumptions/opinions.

The dataset had about 370k rows and 8 columns drilled down to the port of entry level. Our first step was to understand the trends of border crossings over the years (1996-2018). Based on these, we aimed to identify the busiest states for entry into the US along its border with Mexico and Canada. Next, we pinpointed to all the ports of entry in the top 3 busiest states, namely New York, Texas and California, using interactive maps. Post this initial analysis, we ran a regression model to predict the future trends.

An article on Wola.org analyzed a defined dataset to consider nationalities, races, age groups of people etc. to show that the number of people crossing the US-Mexico border has substantially reduced in the recent years. This can be attributed to the stringent immigration policies being put in place by the current administration. Similar articles from other reporting agencies such as CNN, NYT and BBC support this claim. However, illegal border crossings with Mexico also contribute significant numbers while our analysis is based on authorized crossings.

US-Canada border crossings are different and less sensitive when compared to US-Mexico border crossings since most of them are attributed to Tourism and Commerce.

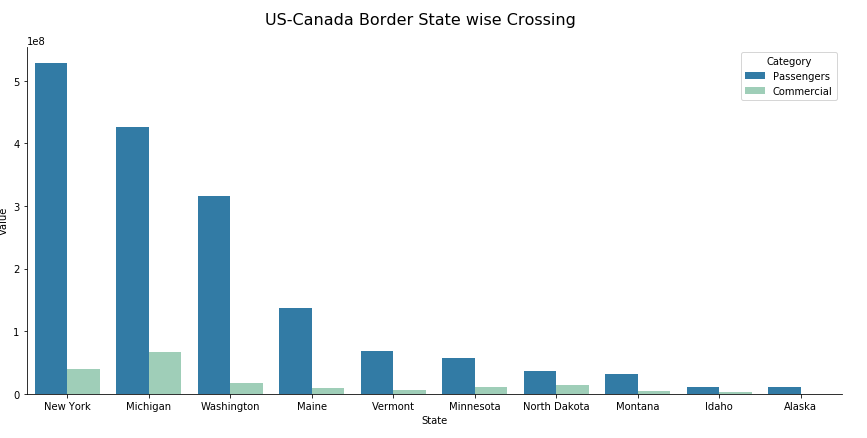
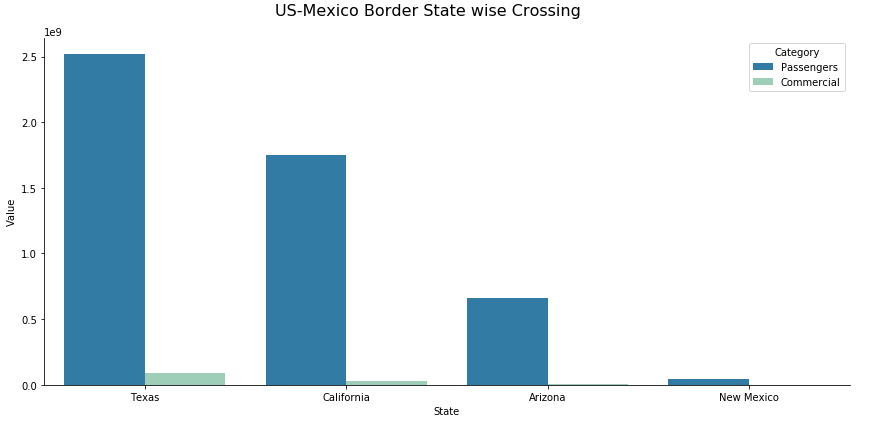
**Computational Setup**

**Read-in:**

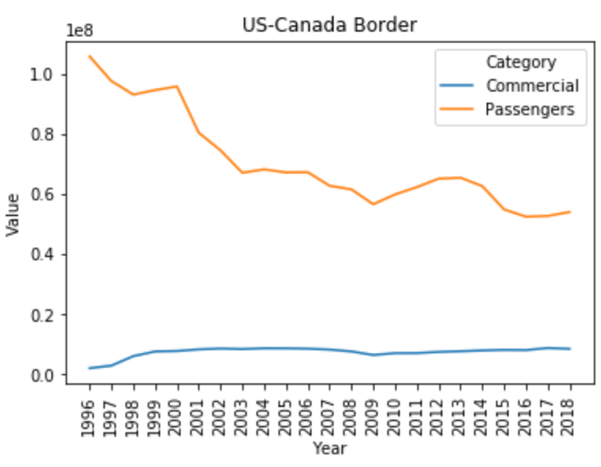
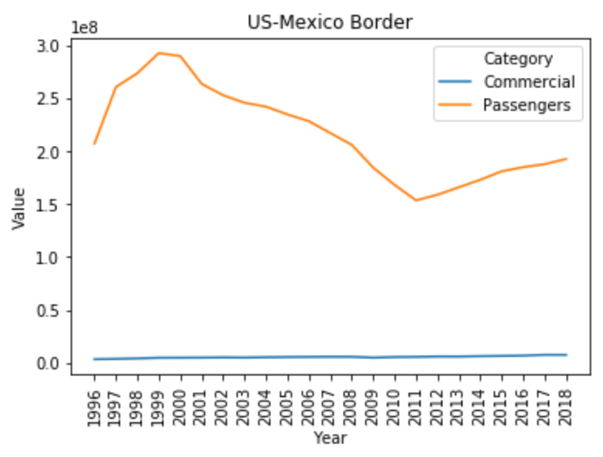
Data has 370k rows and 8 columns, at a state, port of entry and means of transportation level. No null values were found. Redundant means of transportations in the field “Measure” were filtered out from the data. For example, number of “Trucks” does not contribute much to the commercial data when including “Truck containers Full” and “Truck containers empty”. Data for 2019 was incomplete and hence, dropped. Final data was categorized into passengers and commercial for US-Canada and US-Mexico borders.

**Exploratory Analysis:**

We group the data by states to obtain the busiest states over the years for both passenger and commercial traffic. We plotted the border traffic for US borders with Canada and Mexico independently and found out that New York, Michigan and Washington were the busiest states for Canada, whereas Texas and California were the busiest states for Mexico.

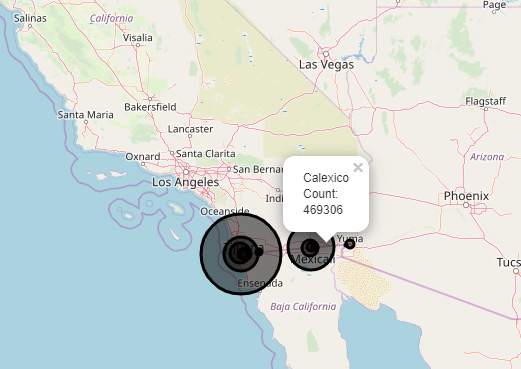
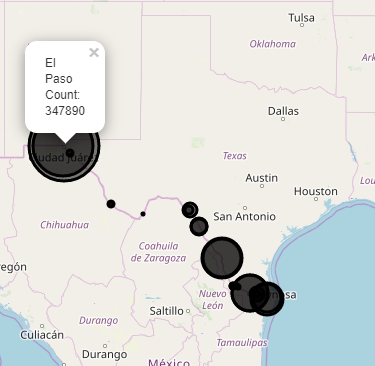
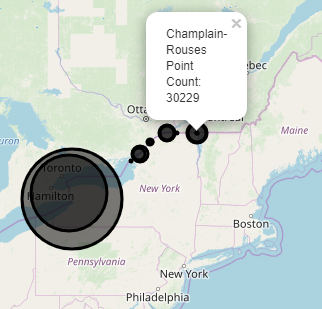
 

Summarizing the data by year, we obtain a time series pattern for the count of passengers/commercial vehicles crossing both US-Canada and US-Mexico borders.

We can see that overall border crossings have decreased over the years. However, the crossings have been extremely unpredictable in the recent years.

Latitude and longitude of each port of entry is one of the crucial features of the data. To make the most of this information, we created interactive maps in folium to visualize the exact location and density of traffic at each port of entry for New York, Texas and California.

New York

Texas

California

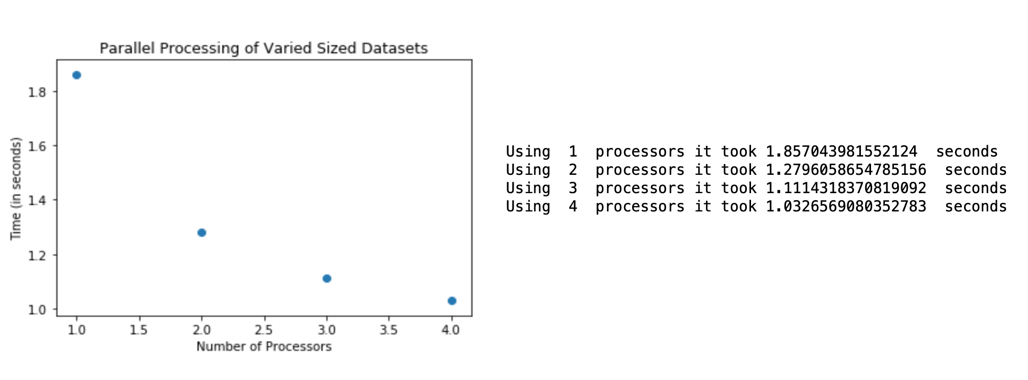
**Regression Model:**

Since the most interesting aspect of the data is passenger traffic and trends of people crossing the border, we have isolated that data to perform a regression analysis using ordinary least squares model with the parameters such as year, state and means of transportation used.

Computational challenges here were the size of the data and time taken for the regression model to run. Random sampling was not an option since there was a possibility that the sample would be skewed or imbalanced based on year. For example, we could not use data between 1996 and 2000 to predict the trends post 2018.

To tackle the data size issue, we decided to strategically sample the data into buckets based on the year. We split the data starting from 2018 and backtracking year wise since the trends post 2018 would be more dependent on the recent years. Using this approach, we came up with 5 subsets ranging from 2018-2013, 2018-2008, 2018-2004, 2018-2000 & 2018-1996.

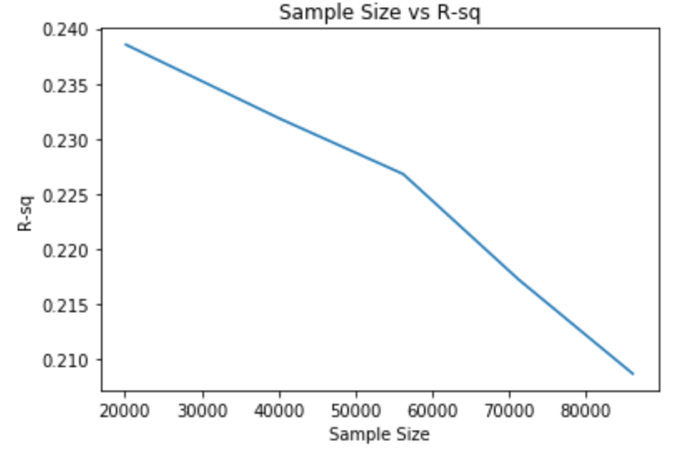
To overcome the runtime for regression over different subsets, we implemented parallel processing. This gave us significant boost in performance which can be seen in the graph below.



Note: Post data processing, our dataset (from 2018 to 1996) had ~90K rows. Thus, the runtimes weren’t significantly high in this case. However, if the dataset were to be scaled up, our solutions to the prior mentioned computational challenges would still be applicable and provide results efficiently.

**Results**

After running our regression model on the strategically sampled buckets, we found that the metric for model accuracy (R-sq) showed a decreasing trend as the data increased (moving backwards from 2018). This is highly unusual since in most regression problems, as the number of observations increase, the models tend to perform better. However, the discrepancy in our model can be fully attributed to the volatile nature of the political situations over the years. Different administrations enforce different immigration policies thereby fluctuating the border crossings drastically. So, when we include a greater number of years, the data becomes more random, thereby negatively impacting the model performance.



**Conclusion**

From our analysis, we conclude that hypothesis stands rejected i.e., we cannot establish a trend for border crossings in the near future. The traffic would be highly dependent on the immigration policies being enforced by the government.

For both the countries (Canada & Mexico), our analysis about border crossings had limited scope. For instance, illegal immigrants contribute significantly to border crossings with Mexico. Similarly, for both Canada and Mexico, crossings via air and sea were not accounted for in our data. Going forward we can include these parameters (and more) to present a more accurate study.

The major computational issues we identified in this study were size of the data and model runtimes. These can be mitigated using the solutions presented in our analysis such as strategic sampling and parallel processing.

**Appendix**

* Kaggle Data Source: <https://www.kaggle.com/akhilv11/border-crossing-entry-data>
* <https://www.wola.org/analysis/us-government-2018-border-data-trump-immigration-asylum-policy/>
* <https://www.cnn.com/2019/10/08/politics/migrants-apprehended-us-mexico-border-fiscal-year-2019/index.html>
* <https://www.bbc.com/news/world-us-canada-46824649>
* <https://python-visualization.github.io/folium/quickstart.html>
* <http://stackoverflow.com/>
* Dr. Yaroslav Rosokha: Codes discussed in class (Parallel Processing, MeasureTime, Regression using OLS)