Ali Ali
Diego Munoz
Dr. Jae Kim
ISYE480 - Data Analytics
December 22, 2019

Final Project

Introduction

The United States Department of Transportation (UDOT), has recently hired me (Diego) and Ali Ali, to provide an analysis of the number of monthly border entries across the southern and northern border of the lower 48 states. Where one of our tasks is to visualize the data so that we have a better understanding of the border crossings at both the U.S.-Canada and U.S.-Mexico border. In addition, we will also create predictive models that will accurately predict the entries by modality(i.e. personal vehicles, trucks) given any port of entry name.

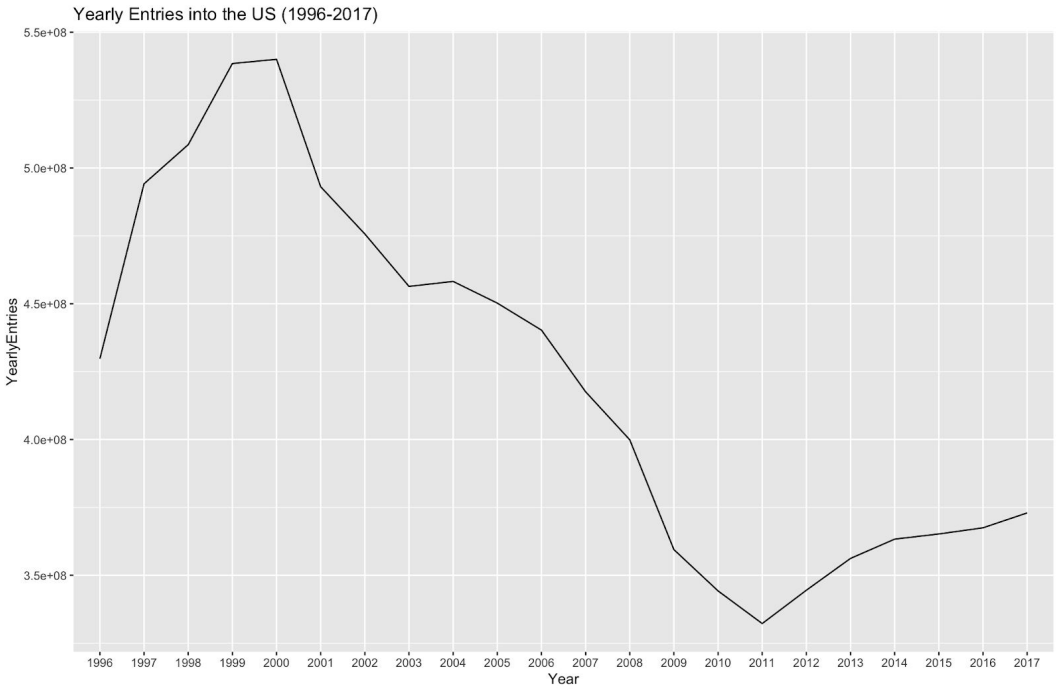
In order to understand the trends of traffic across the northern and southern borders. We first created various plots to visually represent the data. Showing the number of entries by location and modality at both the U.S-Canada and U.S.-Mexico border over a 20+ year period. We have taken steps to find the significant factors that affect the rate of pedestrian traffic. Before any model is created, a foundational grasp of the data must be had in order for the models, that were to be created later, to make sense.

Task 1

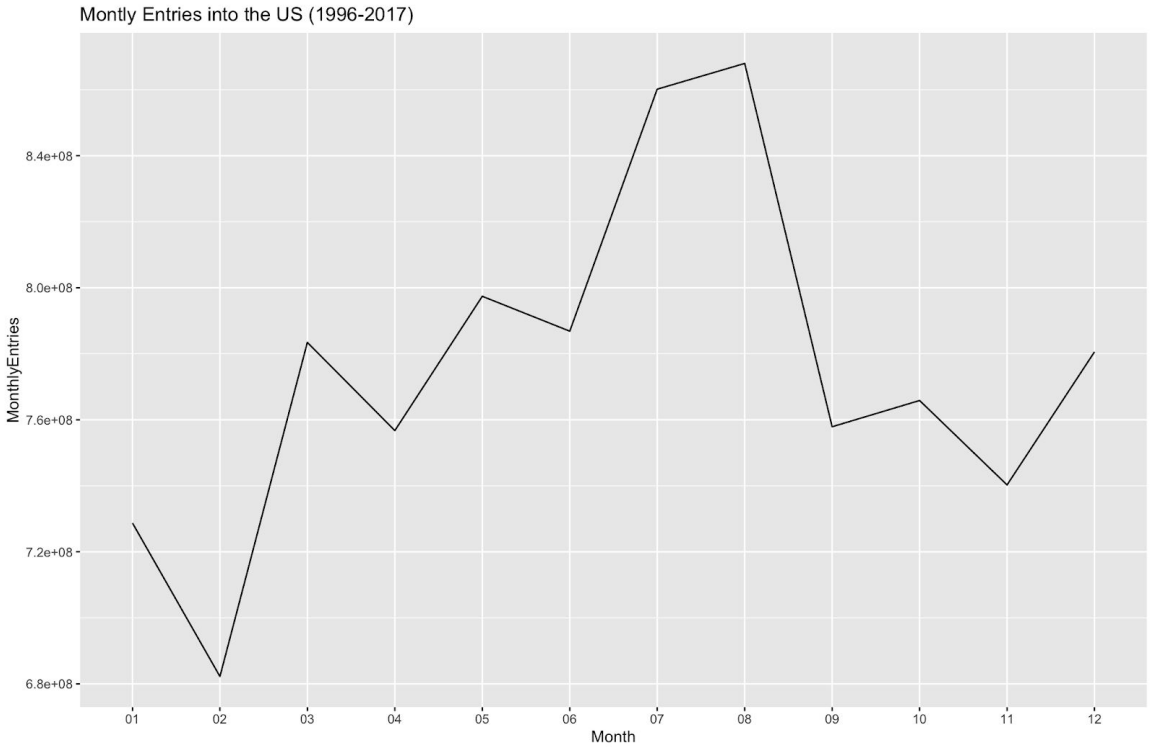
Data Visualization

In the first part of this analysis, we were asked to analyze the data by looking at trends and patterns that have occurred from 1996 to 2018 in terms of entries into the US. We were looking at factors such as year, time of year, the state of entry, the measure used for entry, as well as border of entry. From those, we can see what have been significant factors in the number of entries throughout the years from neighboring countries.

For our visualizations, we decided to remove 2018 data because we only had data up until March, which skewed the overall dataset. The first step of our analysis was to simply look at how the sum of entries changes from year to year.

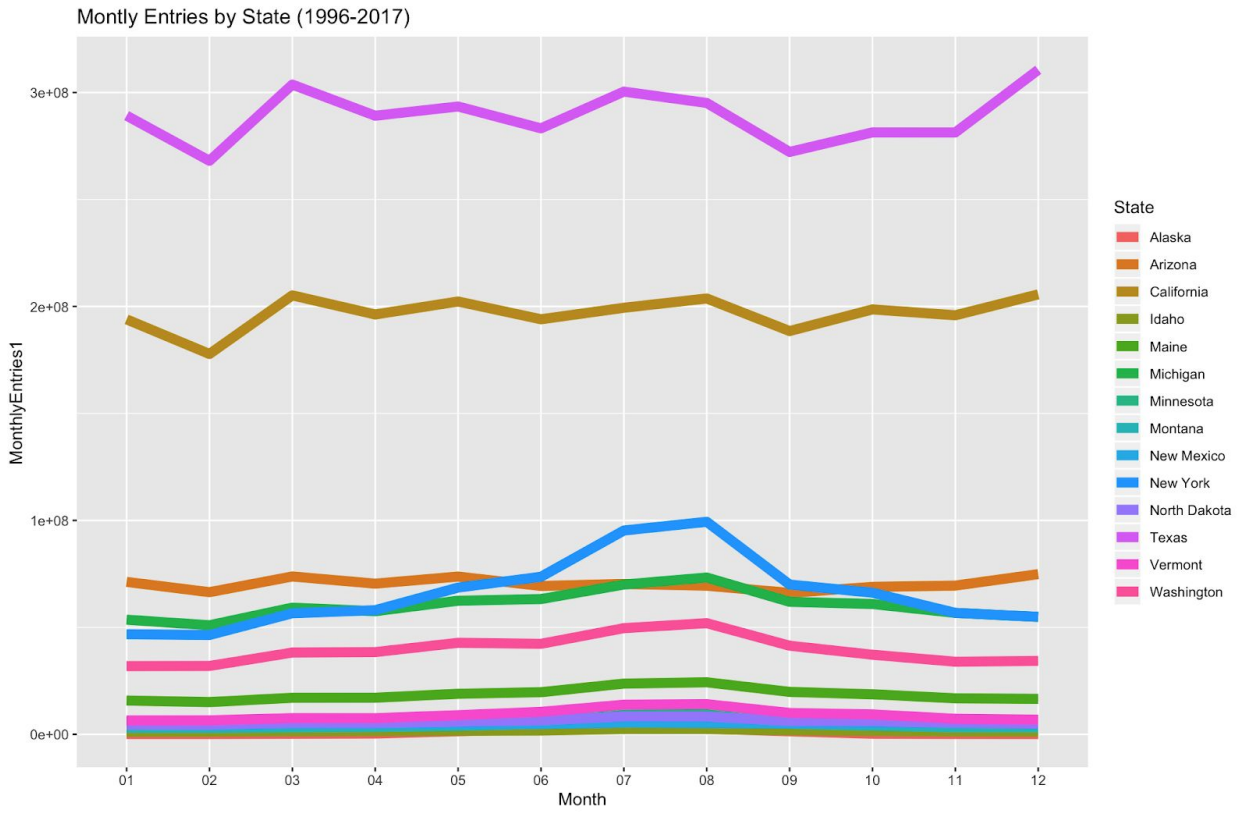


The plot above shows us that the number of entries has gone down significantly from around 2000, even though it has seemed to increase again starting in 2011 again, but not to the extent it was from 1996 to 1999.



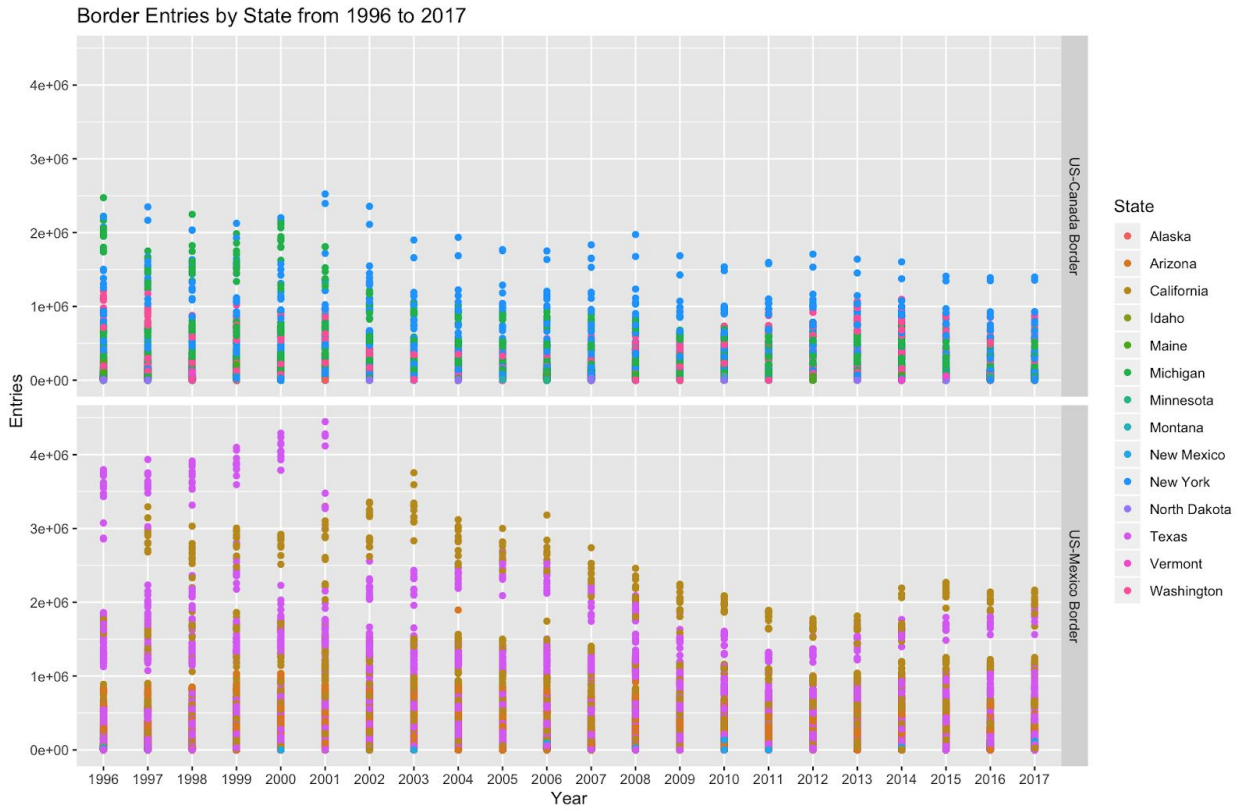
Another simply analysis was to show the trends in different times of the year, and the plot above shows us that the Summer leads to the highest number of entries. So, from a predictive standpoint, we might be able to conclude that June, July, and August will lead to a greater number of entries in the upcoming years. That makes perfect sense as it gives people more time to turn around and adapt to new surroundings.

Next, we went a little more in depth. We looked at which states had the highest number of entries, and how their numbers changed over the course of the year.



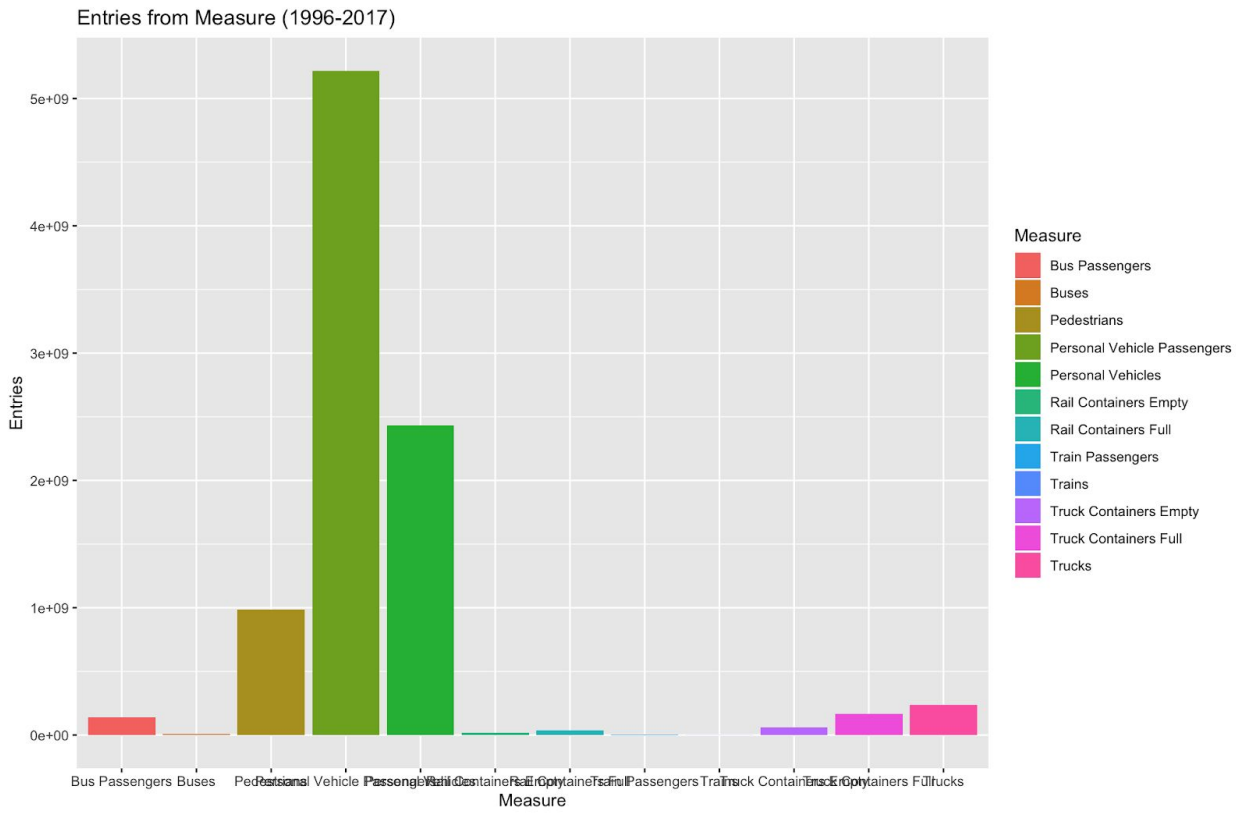
From the plot above, Texas and California clearly have the highest number of entries, and the time of the year does not change that. A lot of states such as New Mexico, Vermont, North Dakota, etc. have a very small amount of entries throughout the year and don't change much. The one state that does change is New York (represented by the blue line the middle), as we can we can see that it is affected by seasonality. Its number of entries increases in the summer more than any other state.

The next step was taking a look at the difference of entries between the two borders, the US-Canada border and the US-Mexico border. The plot below compares the two borders by State between 1996 and 2017.

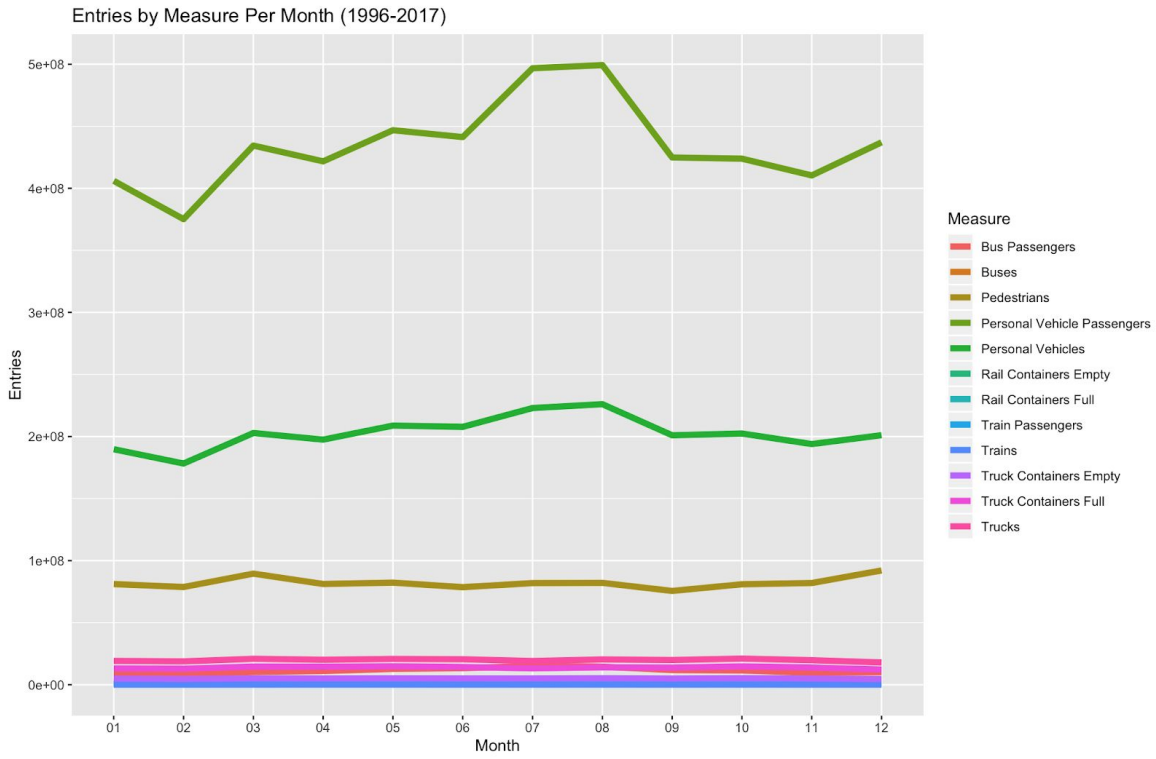


The plot above clearly shows us that the US-Mexico border has a greater number of countries caused by the presence of states such as Texas and California. One interesting observation is that the number one entries throughout the 21 years has decreased more at the US-Mexico border than the US-Canada border.

Then, we attempted to analyze the measure of entry, which would tell us how most people enter the US. This was done through pure value and monthly trends.

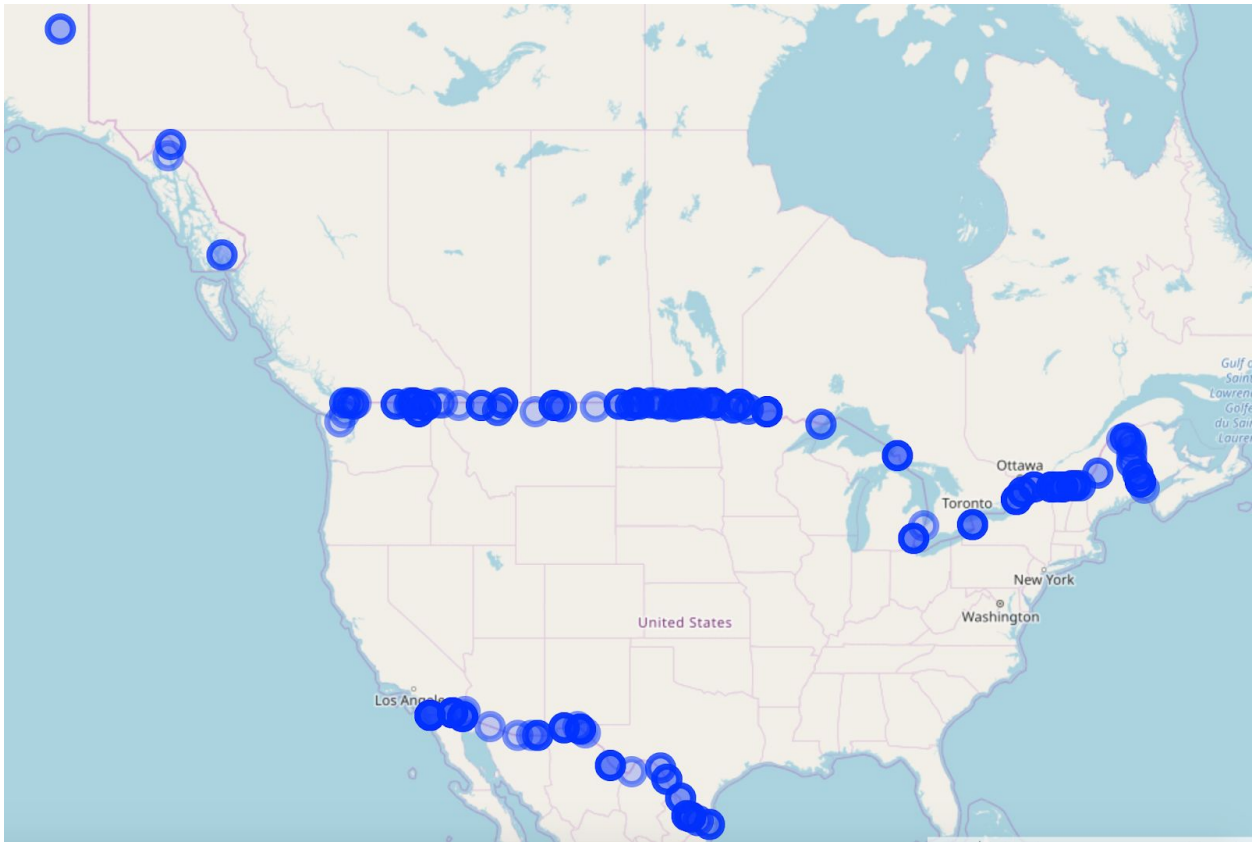


The bar chart tells us that Personal Vehicle Passengers, Personal Vehicles, and Pedestrians are the measures used for most of the entries into the US by a significant margin.



And the line charts suggests that the aforementioned measures stay the most used method of entry no matter the time of the year.

Lastly, we took the first 250 data points of the dataset and placed them on to a map of the US. The plot only includes 250 points because having more would have overpopulated the map.



The plot adds markers to visually show the exact location of each entry point according to its latitude and longitude. The higher the value, the darker the circle.

We used the visualizations to create models to predict values for upcoming dates and areas in Task 2.

Task 2

Modeling

Our goal for this case study was to accurately predict the number of entries for future time periods depending on the area and the measure of entry. In order to help us create better models, we decided to break the dataset into 8 subsets. A summary of the 8 subsets is given in the table below.

Subset	Border	Measure
DF1	US-Mexico	Personal Vehicles
DF2	US-Mexico	Pedestrians
DF3	US-Mexico	All Other
DF4	US-Canada	Personal Vehicles
DF5	US-Canada	Pedestrians
DF6	US-Canada	All Other
DF7	US-Mexico	Personal Vehicle Passengers
DF8	US-Canada	Personal Vehicle Passengers

For all subsets using the US-Mexico Border, we included the monthly average Peso to Dollar exchange rate from January 1996 to March 2018.

For all subsets using the US-Canada Border, we included the monthly average Unemployment Rate in Canada also from January 1996 to March 2018.

For all subsets, we created a linear regression model to find the significant factors and predict the number of entries (the Value column in the dataset). Each one had a training set within each subset, which included 70% of the subset split up using the Value column, with the other 30% comprising the testing set. All linear models were created used using the training sets of each respective subset, then applied to each testing set to find the error associated with our predictions.

Before making the models, we decided not to take into account some factors. We did not include Port Code because we included Port Name, which would be redundant, as well as opting for Latitude and Longitude instead of Location, for the same reasons.

A summary of each model is included in the table below.

Subset	Significant Factors in Final Model	R-Squared (%)	RMSE
DF1	- Port Name - USxMX Exchange Rate - Date - Longitude	92.32	575.69
DF2	- Port Name - Date - Longitude	87.94	2115.27
DF3	- Port Name - USxMX Exchange Rate	35.48	2932.52
DF4	- Port Name - Canada Unemployment Rate - Date - Longitude - Latitude	91.26	486.22
DF5	- Port Name - Date	52.83	81.08
DF6	- Port Name - Canada Unemployment Rate - Date - Longitude - Latitude	38.09	981.45
DF7	- Port Name - USxMX Exchange Rate - Date - Longitude - Latitude	84.65	1214.57
DF8	- Port Name - Canada Unemployment Rate - Data - Longitude - Latitude	84.71	74.56

Predictions

These are the final models that we are recommending for use. These models were used to predict values for the evaluation set. This was done using a for loop that reads the evaluation set, and checks the conditions of each row. The two conditions applied in this case were the Border and Measure, because those were the conditions of our subsets. This ensures that each individual row of the dataset is predicted using the correct model from the 8 that were initially created for each set of specific characteristics. The predictions were generated, and were placed in a new column in the evaluation dataset.