Data Cleaning and Methodology:

The project has 4 Jupyter Notebook files for data cleaning and analysis:

1 – border\_crossing\_analysis – includes data cleaning for first source and analysis

2 – weather\_data\_analysis – includes weather data retrieval from OpenWeatherMap API

3 – weather\_data\_analysis – includes analysis of border crossing traffic based on weather data

4 – statistical\_analysis – statistical analysis to evaluate the significance of average border crossing volume for each year

The original border crossing dataset had data ranging from Dec-1995 to Feb-2019. The original CSV file has 346733 data entries and 8 columns:

1- Port Name

2- State

3- Port Code

4- Border

5- Date

6- Measure

7- Value

8- Location

The first step in our data cleaning process was to filter out data from 2009 to2018. The format for ‘Date’ column was originally string and although the value had year, month, day and time but there was only one date for each month which implies that it was monthly data. We converted the values in this column to ‘datetime’ format as well as retrieved the ‘timestamp’ formatted data in a separate column. The datetime data was split into year, month and date values.

The ‘Location’ column had longitude and latitude values but in one cell. We used string split function to retrieve latitude and longitude values and stored them in separate columns.

Next, we dropped the columns we did not need which were ‘Port Code’ and original columns for ‘Date’ and ‘Location’. Then we filtered the data to our targeted analysis period of Jan-2009 to Dec-2018 which included 144605 data entries. This final retrieved dataset was then exported to a CSV file.

For weather-related analysis, we narrowed our time frame to just 2018 and grouped so we had only one row for each Port Name (116 unique Ports) based on unique timestamp (one timestamp for each month in 2018). The reason for narrowing data for one year was that there was limit on OpenWeatherMap API for historical data retrieval of 2000 entries per day even with subscription. The retrieved data for five different weather variables were added to the dataset which was then imported to another CSV.

Data Analysis:

The data analysis part in Jupyter Notebooks investigates the data for our research questions for this project. The methodology includes filtering, grouping and sorting data in various forms.

For the first few research questions, the data is divided into two overall categories: people and vehicles. This is based on the values in the column ‘Measure’ to differentiate the border crossing traffic for number of people vs. number of vehicles.

There are several charts prepared throughout the analysis which are saved in the Images folder. The syntax to display the charts is then commented out to avoid overwriting the images every single time we re-run the Jupyter Notebbok.

**Analysis**

The second part of the project analyzes border crossing traffic in correlation with weather conditions. The data for weather including Temperature, Humidity, Cloudiness, Wind speed, and Weather condition was retrieved from OpenWeatherMap API for the targeted time frame.

**Introduction**

There are two data sources used for this analysis. The first is a data set from Kaggle about border crossing entry data from the Bureau of Transportation Statistics (BTS) ( <https://www.kaggle.com/datasets/akhilv11/border-crossing-entry-data> ) . It provides information regarding inbound crossings at the U.S.-Canada border and the U.S.- Mexico border between 1996 to 2019. The second is historical weather API from OpenWeatherMap (<https://openweathermap.org/api/history-api-timestamp> ). We decided to choose this data set as there are many socio-political-economic discussions regarding US border activity and we wanted to see if there are any meaningful insights we could gather that can contribute to the discussion. For the purposes of our project, we chose to analyze a 10-year span from 2009 to 2018 as those are the most recent years with information from the entire year.

The data set provides information about the volume of both vehicle traffic and people traffic. It provides the date of each activity, the border port, the states in which the ports are located as well as the vehicle types. The limitations of the data set were that it only had information up to 2019 and the 2019 data only went up to February. Had we had access to information up to possibly 2022, we could have provided some more insight. There is also the issue of the data set only showing inbound border activity to the US nor outbound.

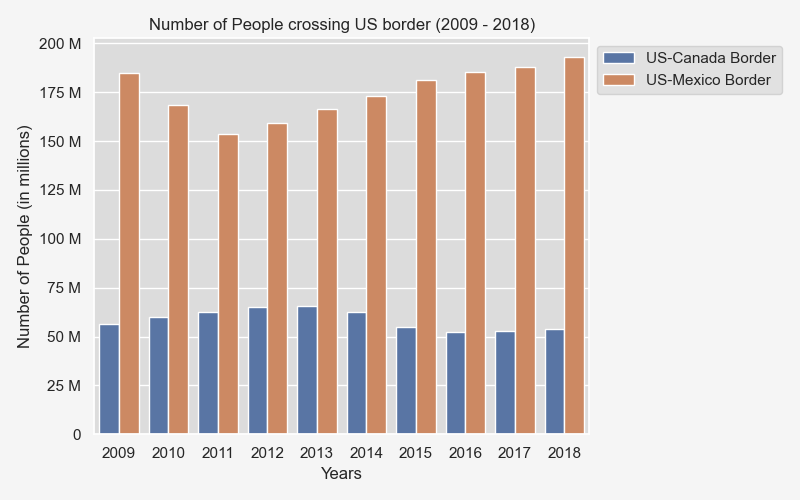
In order for the data to be ready for analysis, there was a data cleaning process that was followed.

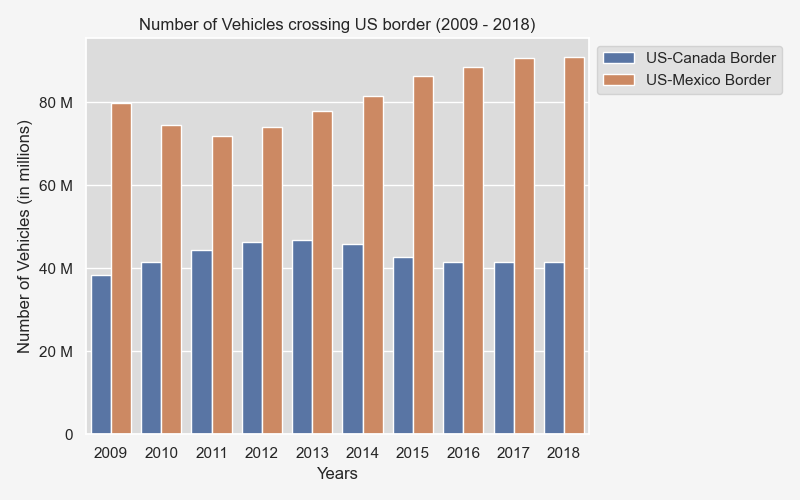
The areas covered in this analysis are the busiest border, the busiest ports at each border, the states with the most border activity, the months with the most activity and if there was any correlation between border activity and the temperature for the most recent year of 2018. In this analysis, busy is defined as the volume of people traffic and vehicle traffic across these two borders. The paragraphs that follow will display each question we posed in our project proposal followed by our findings.

**Findings**

**Question 1 - What’s the busiest border for the US?**

There are two borders for the US: the US-Canada Border and the US-Mexico Border. Two bar graphs were created to determine the answer to this question. One bar graph looked at the people traffic while the second bar graph looked at the vehicle traffic. Next to each of the bar graphs below, there is the table that includes the raw count that makes up the bar graph.

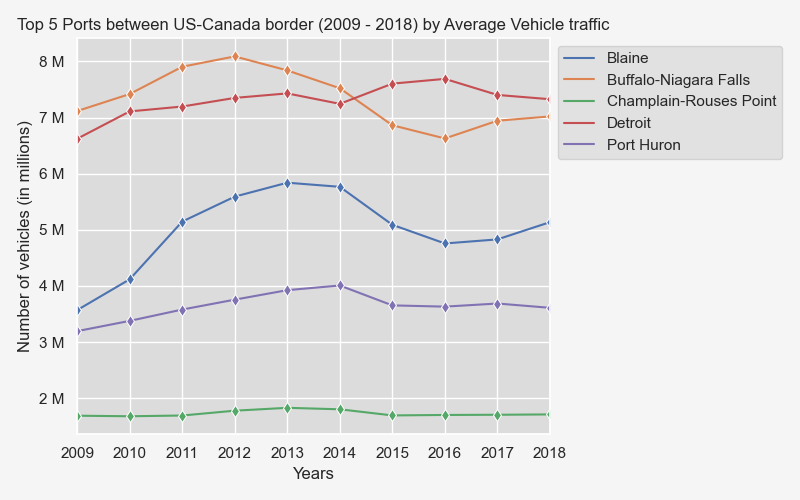
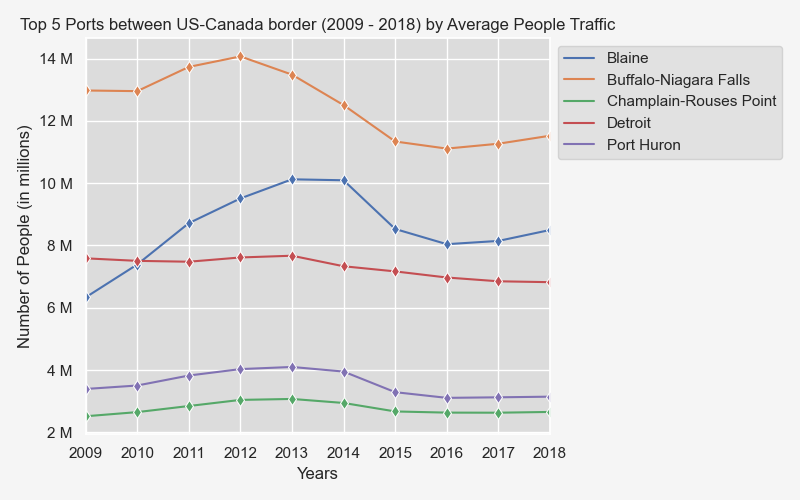




The people traffic for the US-Mexico border is 3 times higher than the US-Canada border overall across the 10 years analyzed i.e. 2009-2018. The US-Mexico border has 1.2 billion more people crossing the border overall. The vehicle traffic for the US-Mexico border is also 2 times higher than the traffic across the US-Canada border. The US-Mexico border has 380 million more vehicles crossing the border overall. For both the people and vehicle traffic, the US-Mexico border has an upward tick of activity from 2012 to 2018, jumping about 23% for its vehicle traffic and jumping about 21% for its people traffic. For both the US-Mexico border and the US-Canada border, the people traffic is much higher in general.

**Question 2 - Which are the top 5 busiest ports between US and Canada?**

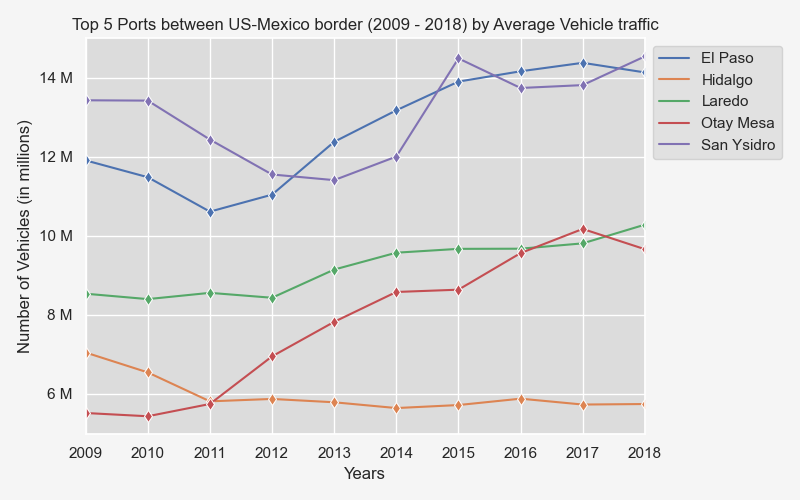
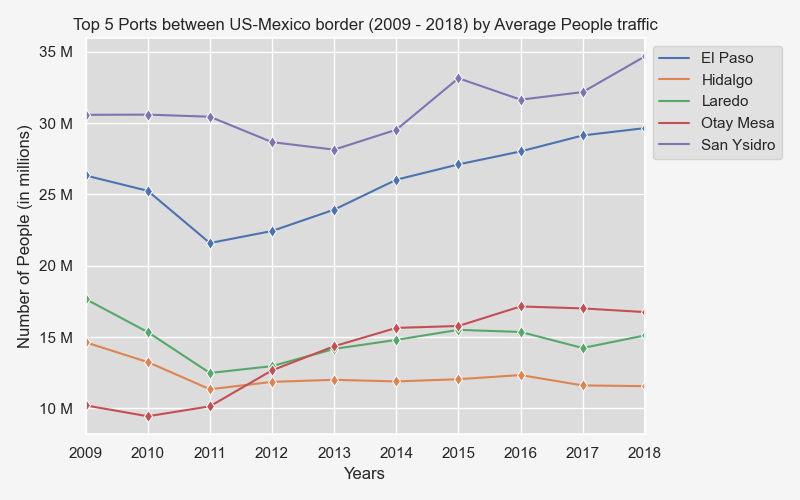
For this question, line graphs were used to look at the people traffic and the vehicle traffic across the US-Canada border.

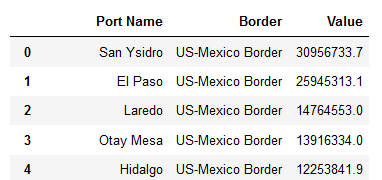
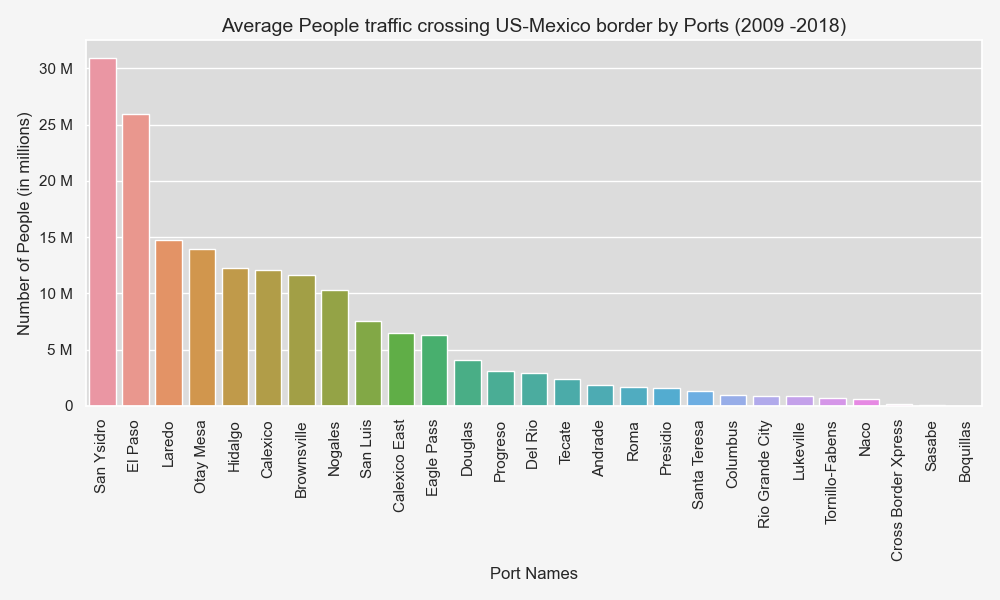


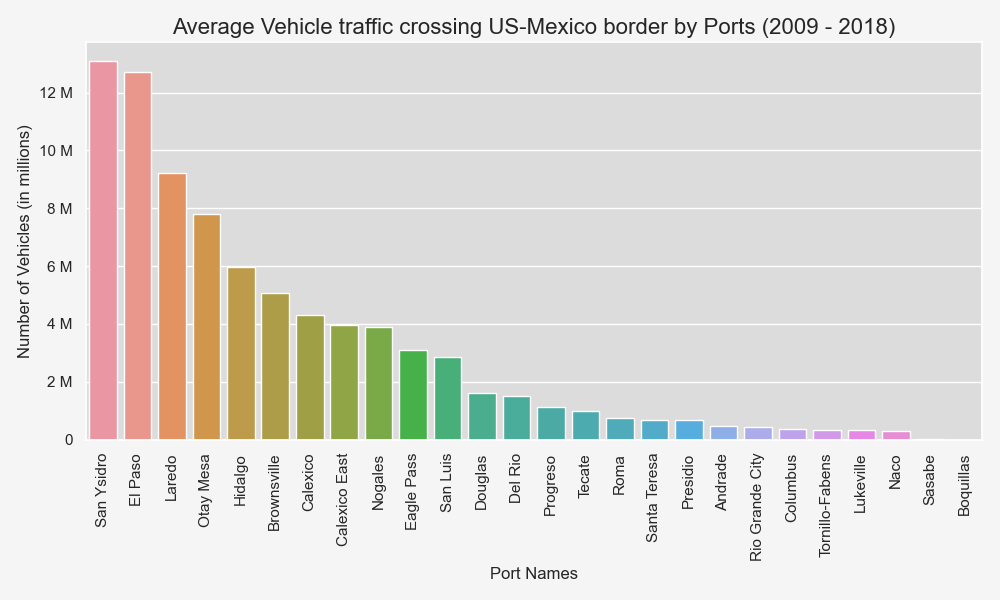
Across the US-Canada border, between 2009 to 2018, the top five busiest ports overall are Buffalo-Niagara Falls, Blaine, Champlain-Rousses Point, Detroit and Port Huron. The top two busiest ports for people traffic are Buffalo-Niagara Falls and Blaine. Both of these ports are close to metropolitan cities and are located towards the end of US-Canada border line. For its vehicle traffic, the top two busiest ports are Buffalo-Niagara Falls and Detroit. There is an increase in people traffic from 2010-2012 in these two ports, then there is a decline and another uptick starting in 2017. The people and vehicle traffic for the Detroit port stays relatively steady across the year compared to the other active ports.

**Question 3: Which are the top 5 busiest ports between US and Mexico?**

Similar to question 2, for this question, line graphs were used to look at the people traffic and the vehicle traffic across the US-Mexico border. However, because there are significantly less ports across the US-Mexico border, bar graphs were also included to provide some additional insight.



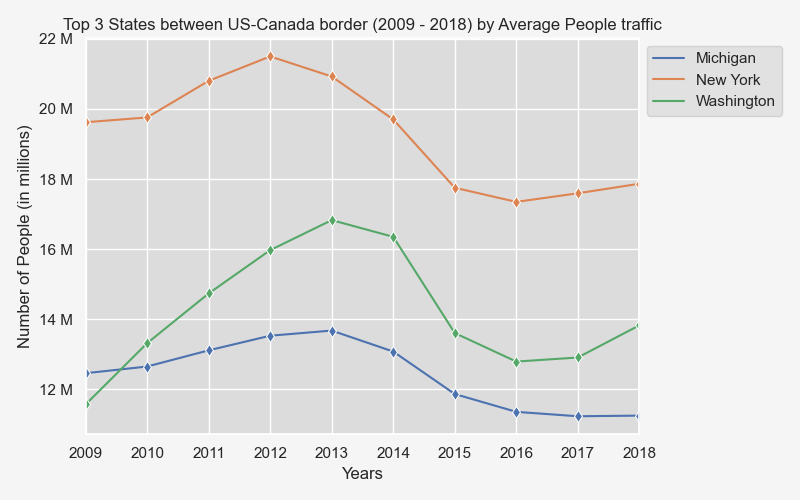
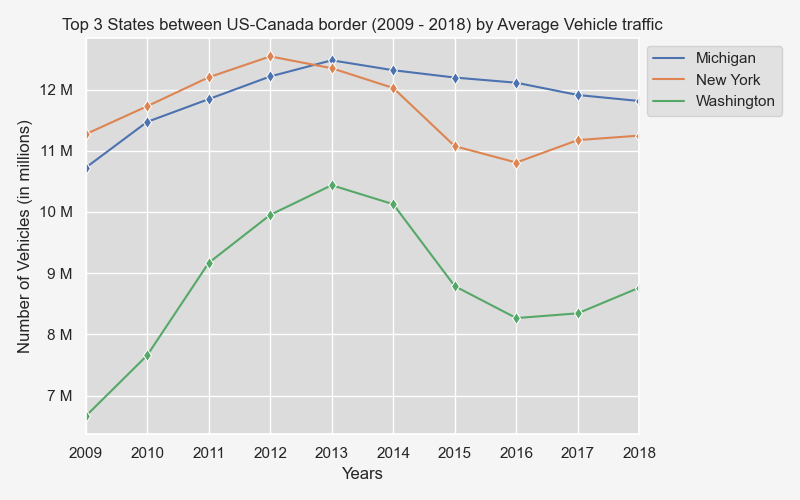


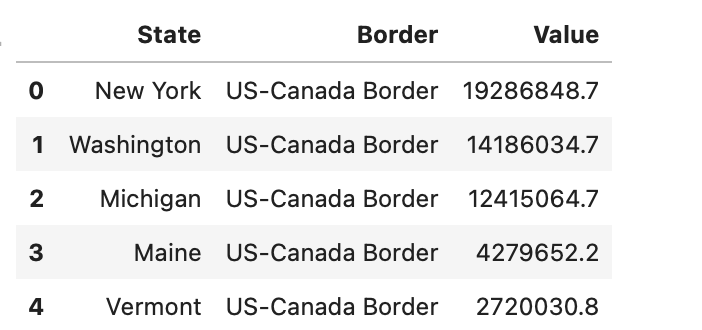
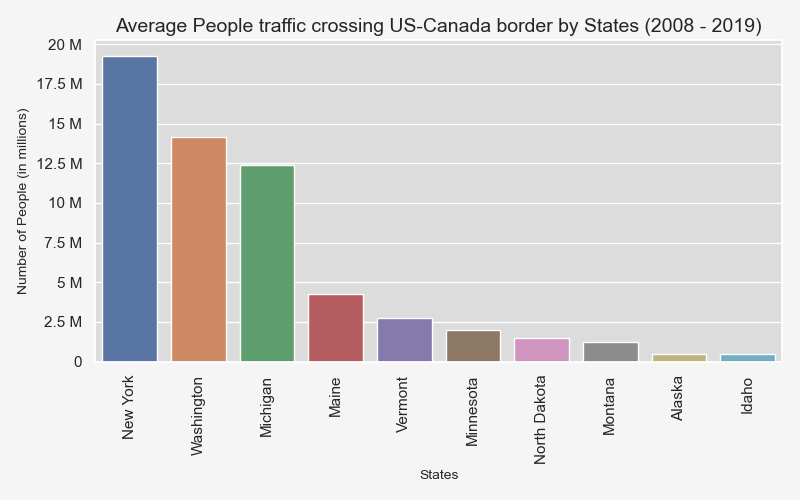


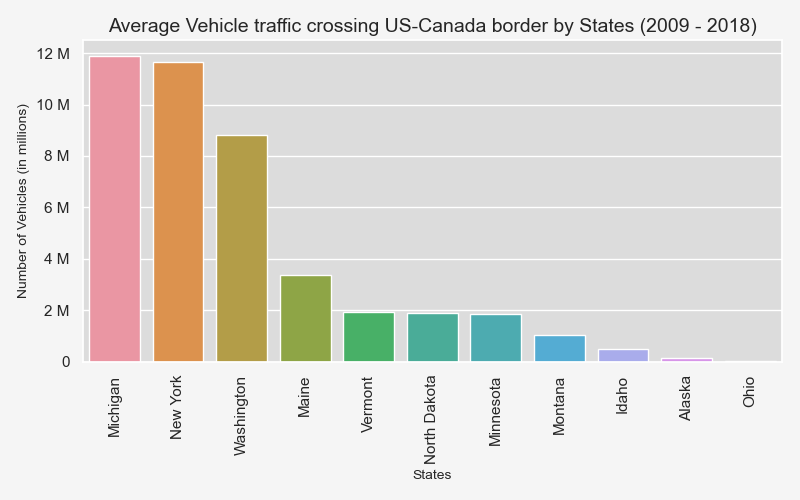
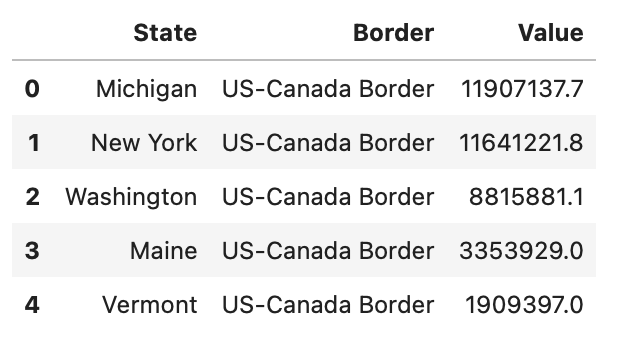
The top 5 busiest ports for both people traffic and vehicle traffic across the US-Mexico border are San Ysidro, El Paso, Laredo, Otay Mesa and Hidalgo, between 2008-2019. The top two busiest ports across both people traffic and vehicle traffic are the San Ysidro and El Paso; these ports are near metropolitan cities. Moreover, across these two busiest ports, there is a decrease from 2009 to end 2012 then a steady increase as of 2013 across both ports. Across the busiest port across the US-Mexico border i.e. San Ysidro, there is on average about 2.5 times people crossing the border than there are vehicles.

**Question 4 - Which are the top three busiest states between US and Canada?**

For this question, both bar graphs and line graphs were used to draw some observations.



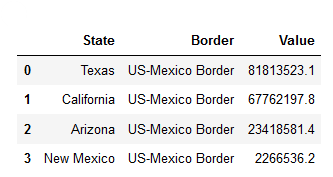
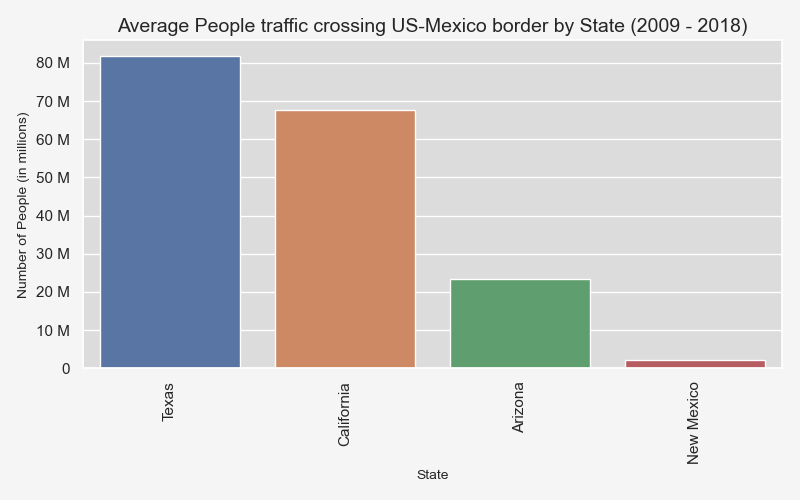
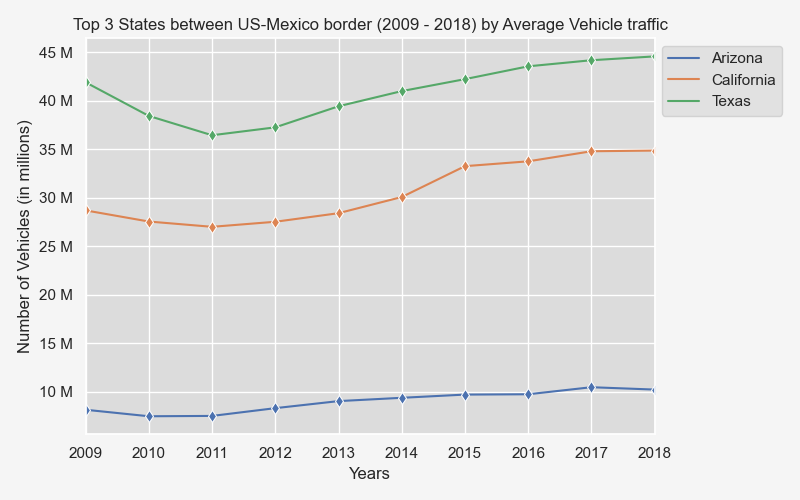
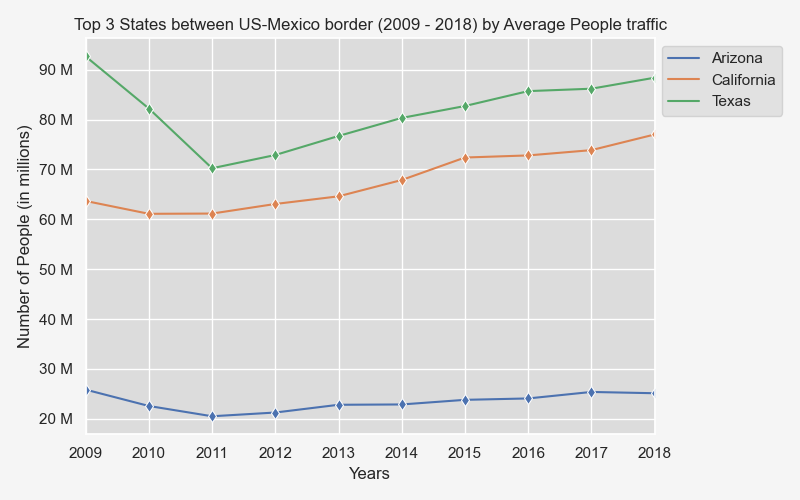


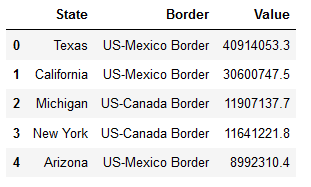
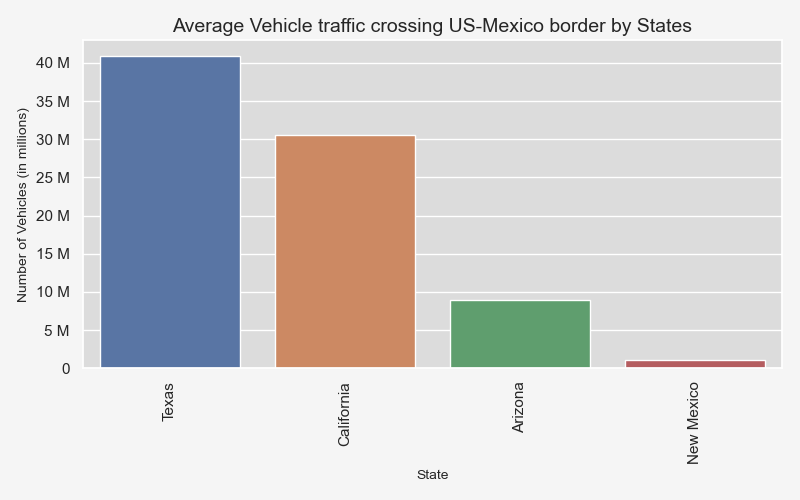
****

Across the US-Canada border, the top three busiest states are Michigan, New York and Washington. Michigan has the most vehicle traffic followed by New York at a close second. New York has the most people traffic followed by Washington. New York has about 1.5 times more people traffic than Washington. New York remains the busiest state overall for both vehicle and people traffic showing a more consistent traffic trend line across the years. In New York, there is a significant difference between the number of people crossing the border versus vehicles crossing the border. The people traffic is about 2 times more. In regards to both vehicle and people traffic, the volumes increased between 2009 and mid-2012 for the top three busiest states and then started to decrease. Of the three, Michigan has remained relatively steady whereas New York and Washington show greater volatility in their decline. Zooming out, the busiest states also coincide with the busiest ports along the US-Canada border i.e. New York and Washington. For the people traffic, New York is on average 7 times busier than Vermont. For vehicle traffic, Michigan is on average on 10 times busier than Vermont.

**Question 5: Which are the top 3 busiest states across US-Mexico border?**

For this question, both bar graphs and line graphs were used to draw some observations.

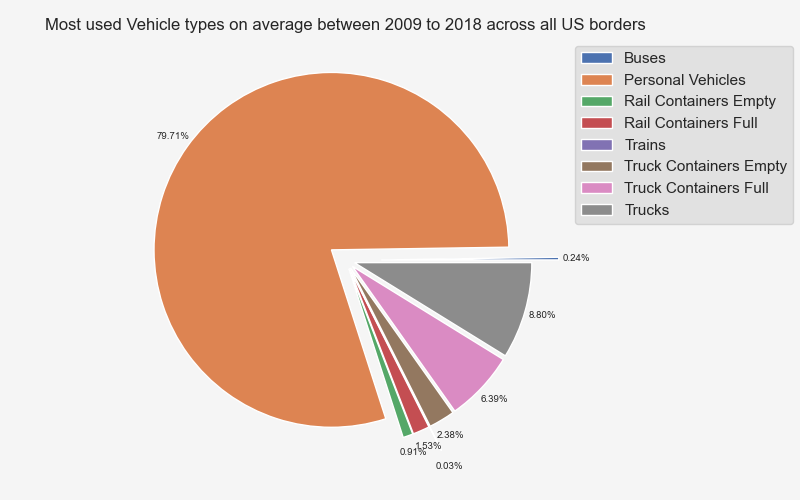


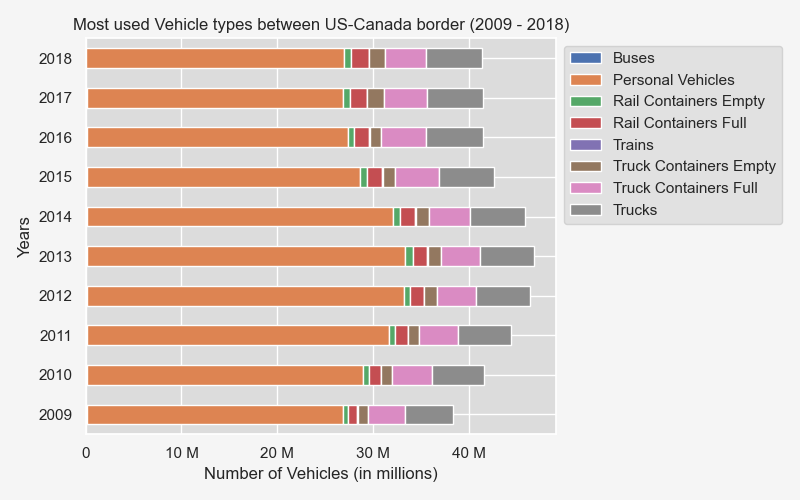
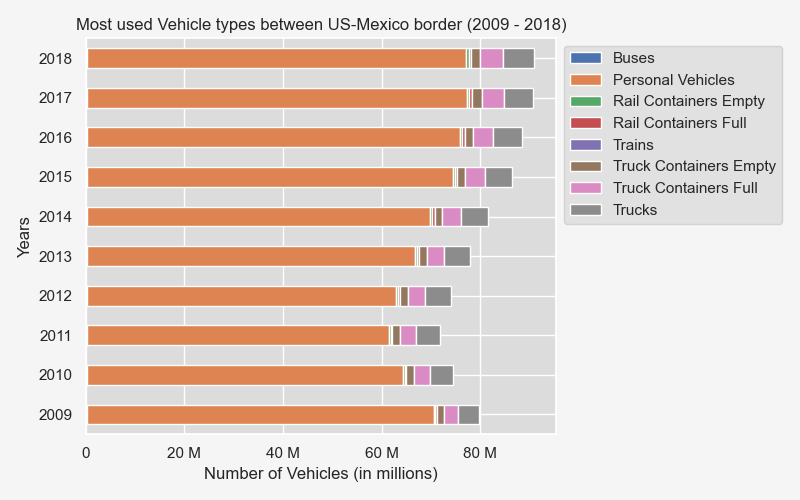


Across the US-Mexico border, the top three busiest states for people traffic are Texas, California and Arizona. The top three busiest states for vehicle traffic are Texas, California and Michigan. Texas has 1.2 times more people traffic than California. Texas remains the busiest state overall for both vehicle and people traffic; there was a decrease in traffic in this state from 2009 to 2011 then an upward increase since 2012. In Texas, there is a significant difference between the number of people crossing the border versus vehicles crossing the border; the people traffic is about 2 times higher. In regards to both vehicle and people traffic, the volumes increased between 2009 and 2011, for Texas and California and then start to trend upward. Of the three, Arizona has remained relatively steady across the years. Zooming out, the busiest states also coincide with the business ports along the US-Mexico i.e. Texas and California. For the people traffic, Texas is 36 times busier than New Mexico. For vehicle traffic, Texas is 5 times busier than New Mexico.

**Question 6: What vehicle types are used the most?**

For this question, we used both a pie chart and horizontal stacked bar graph to make some observations.

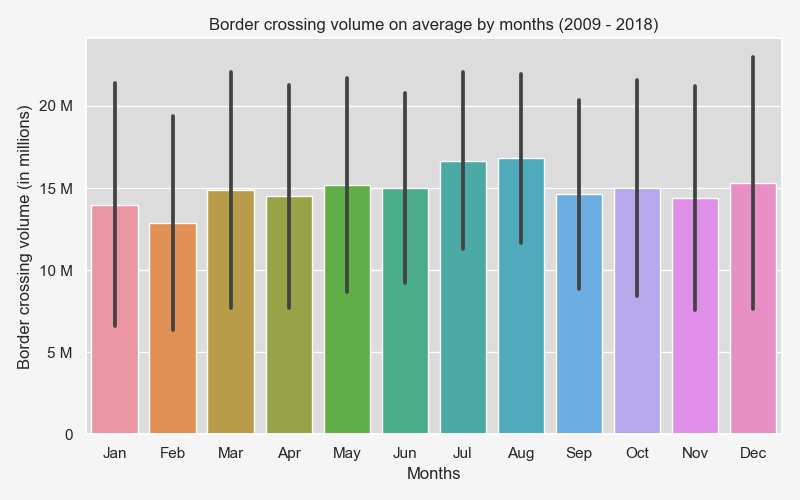
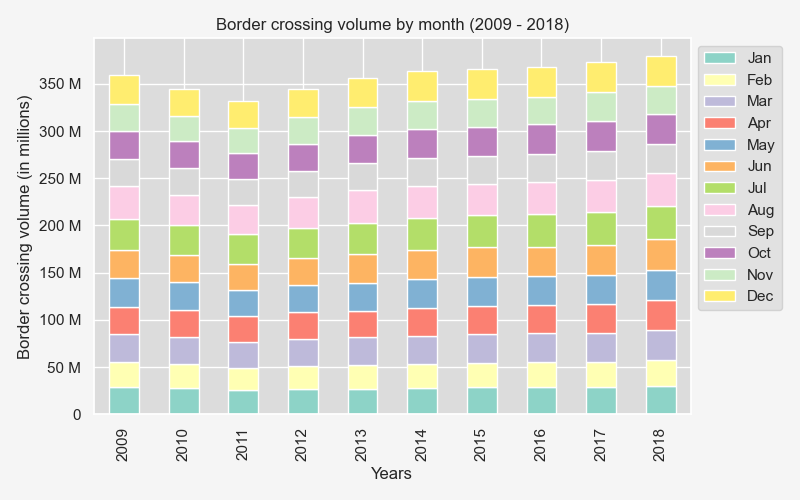


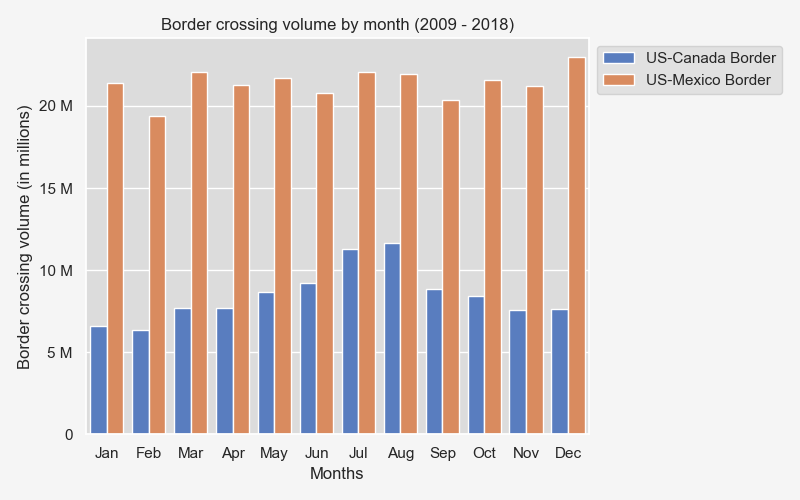
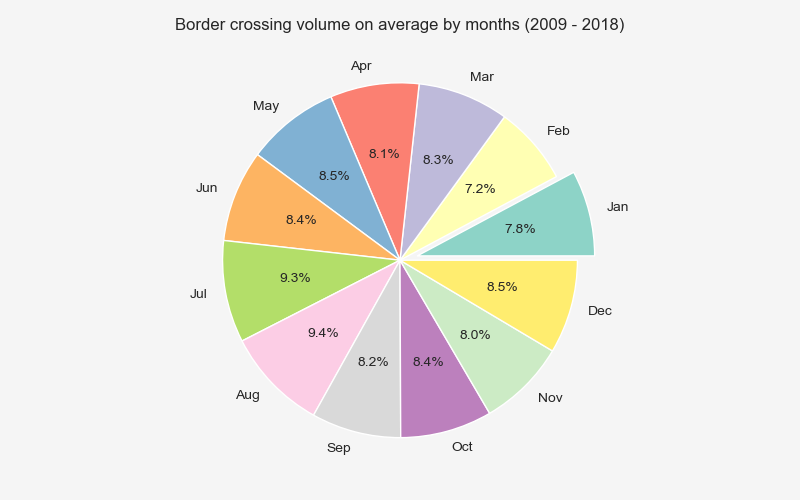


Personal vehicles represent 79.7% of the vehicle volume across the 10 years followed by trucks at 8.8%. For Canada, the use of personal vehicles increased from 2009 to 2013 and then has declined ever since. For Mexico, the use of vehicles declined from 2009 to 2012 and has increased ever since.

**Question 7: Which months are the busiest?**

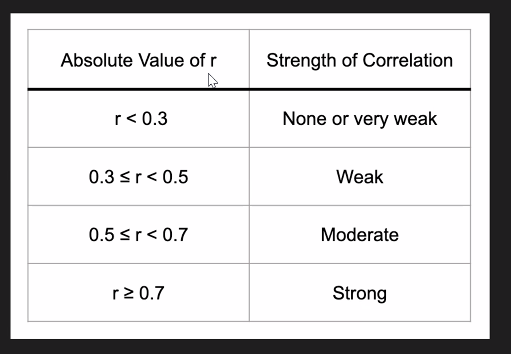
For this question, the following graphs were used: a vertical stacked graph, bar graphs and a pie chart.





Across the two borders, the busiest months across the years are July representing 9.3% of the border activity followed by August representing 9.4% of the border activity. The two months with the least amount of activity across the years are February representing 7.2% of the border activity followed by January representing 7.8% for the border activity. The warmer times of the year garner more traffic because there are less weather concerns and easier mobility. Overall, however, there isn’t one month in particular that has significantly more activity than others.

**Question 8: How does correlation volume correlate between Mexico and Canada over the years?**

****

* **Based on r-value,**

**Conclusion**

There are a few observations that stood out in this analysis. There is consistently more activity across the US-Mexico border than across the US-Canada border. The most border activity along the US-Canada border happens in the New York region and the most border activity along the US-Mexico border happens in the Texas region. The summer months show the most border activities however not by a significant amount.