THE IMPACT OF NATURAL DISASTERS ON BORDER CROSSINGS IN THE US

By

Sai Harshitha Dalli, BS Data Science

A thesis submitted to the Graduate Committee of

Ramapo College of New Jersey in partial fulfillment

of the requirements for the degree of

Master of Science in Data Science

Spring, 2025

Committee Members:

Osei Tweneboah, Advisor

Amanda Beecher, Reader

Kaitlin Sidorsky, Reader

Dedication

I would like to dedicate this paper to my family, who have been my pillar throughout my academic journey. I show my gratitude toward my parents for their continuous backing and dedication, along with teaching me the importance of education through their example. I would like to thank my sister for always being there for me and cheering me on to keep working hard, even in the most difficult moments.

Acknowledgments

I extend my deepest gratitude to my advisor, Dr. Osei Tweneboah, for his invaluable guidance, encouragement, and support throughout my research journey. I would also like to sincerely thank my readers, Professor Amanda Beecher and Professor Kaitlin Sidorsky, for their time, thoughtful feedback, and commitment to helping me strengthen my work. Their expertise and insights greatly contributed to the quality of this research.

Table of Contents

List of Tables

List of Figures

Abstract

The U.S.-Mexico and U.S.-Canada borders are vital to the trade and tourism sector of the country, which influences the economy, but the borders are susceptible to natural disasters. The research explores natural disasters as a factor for border crossing values by utilizing two datasets: U.S. border crossing entry data and FEMA disaster declarations data. The goal is to analyze patterns and use modeling to forecast the border volumes and predict volumes based on disasters to help find the impact of the two.

Exploratory Data Analysis (EDA), K-Means and K-Prototypes clustering, and SARIMA and ARIMA forecast models were applied to the border crossing data set to identify the important factors influencing the border volumes. The two data sets were joined by state, year, and month, for which statistical tests such as the Welch's test were used to test the difference between volumes one month before and one month after a specific disaster type had occurred. A Generalized Linear model (GLM) with Poisson distribution, and a Negative Binomial model were fit for both the U.S.-Mexico and U.S.-Canada borders after checking for dispersion statistics for predicting border volumes based on disaster count and other border predictors.

The clustering model showed traffic patterns where there were more truck traffic crossings in the southern ports, and personal vehicle traffic was higher in the northern ports. The forecasting models captured seasonal trends, showing future volumes. The disaster periods show higher differences in traffic volumes one month before and after a disaster occurrence, though it is not statistically significant by Welch's test. The negative binomial models suggested that disaster declarations were not strong predictors of crossing values, though they were slightly positive. Variables such as personal vehicle passengers and personal vehicles are strong

predictors for both models. Pedestrian crossings are more impactful in predicting the number of crossings in the Mexico border, while rail containers were more influential in predicting the number of crossings in the Canada border.

These findings call for additional research in the future about the border-specific, state-specific, or disaster-specific analysis to dig deeper into the analysis and help build better disaster response policies. Since the data that is used is considered as count data, which is data aggregated by month or year, different regression models such as Generalized Linear Models were used in the research. In the future, more count analysis and models could be explored, and can be compared to the Generalized Linear Model to compare and find the best fit predictive model.

Chapter 1

Introduction

The U.S.-Mexico and U.S.-Canada are two important international land borders that ease trade and tourism, which play a crucial role in the U.S. economy. The connected nations enable the straightforward movement of people along with goods across international boundaries. The borders enable retail businesses, agricultural producers, and manufacturing operations to handle international import and export activities. They also contribute to the migration and immigration of people into the country, which increases the country's economy. These borders ease importing, exporting goods, as well as entering and exiting the country. Multiple factors could affect the number of border crossings, such as weather in both countries, wait times at the ports, types of security, resources at the ports, immigration, and trade policies.

One important factor to study is natural disasters. The occurrence of natural disasters has increased in recent years. Disasters such as hurricanes, wildfires, floods, and pandemics have happened more often and more forcefully over the last few decades, and that has instilled more fear of what their consequences for human populations and economic infrastructures would be. Disasters are causes of breakdown of transport networks, damage to vital infrastructure, and disruption of both commercial and tourism activities. For example, if a disaster damages highway or border operations, it can slow shipments, reduce passenger movement, stop supply chains, and eventually lead to financial and larger economic problems. Since many people cross the border daily, any interruption from natural disasters will have negative economic consequences. Therefore, studying how disasters impact border crossing volumes is one of the most important aspects in developing well-informed policies, developing strong emergency measures, and increasing the ability of cross-border operations to withstand disruption.

This research investigates the combination of both border crossings and natural disasters, which are both crucial to study, and identifies any ties or patterns that can be observed by combining the topics. The research shows whether disasters impact the border crossings into the U.S. volumes. This study utilizes Python libraries, data analysis, visualization, and machine learning to analyze, visualize, forecast, and predict border crossings by adding disaster as a predictor. This could show us significant factors that could impact the number of border crossings at both the borders separately.

The following chapters in the study are organized as follows: Chapter 1 introduces natural disasters and cross-border issues in the U.S and tells the research purpose. Chapter 2 is the background and literature review covering existing research about trade, tourism, migration, and natural disasters. Moreover, it shows how my research fills in the gaps between the previous studies and this research. Chapter 3, the methodology, explains different types of exploratory data analysis (EDA), models, model evaluations, and visualizations used in the research. It simplifies all the techniques and tools that were used for a better understanding of the results. Chapter 4 is the results and discussion section that shows the models such as K-Means and K-Prototype clustering, SARIMA and ARIMA forecasting, generalized linear model with Poisson distribution, and negative binomial regression. It also discusses statistical tests such as the Kruskal-Wallis test for seasonality, the Welch's test, analysis of variance (ANOVA), and dispersion statistics. Chapter 5 is the conclusion, which discusses how the results relate to the problem and how they could be used for future research.

Chapter 2

Background

Section 2.1: Trade in the United States

Border crossings represent a vital area for study in the United States because business activity depends heavily on them. A nation's economic growth heavily depends on trade activities, and the U.S maintains its status as the world's 2nd largest trading nation, alongside more than 200 established trade partners. The United States is the world's leading importer of goods and services with a total annual sum of $3.2 trillion in 2022, which represents an increase of 14.6% ($413.7 billion) for trade of goods and a 23.7% rise ($130.3 billion) from 2021 for services trade. China is one of the important import markets for U.S. goods, alongside Mexico, Canada, Japan and Germany. The United Kingdom along with Germany, Canada, Japan, and Mexico are the top import origins for services (Office of the United States Trade Representative, 2025).

Mexico was the top overall trade partner with 15% of total trade and 14.4% of imports, followed by Canada, which had 13.9% of total trade and 12.1% of total imports in February 2025 (Bureau, 2025). "Mexico supplied roughly one third of U.S. horticultural product imports–including fruit, vegetables, and alcoholic beverages. Canada is a large supplier of processed food products (baked goods), meat, vegetable oils, and vegetables" (Kaufman, 2025). Canada was also the largest supplier of U.S energy imports such as natural gas, electricity, and crude oil. This shows how important these countries are for trading and how influential they are in the economy of the U.S.

Since Mexico and Canada share borders with the United States, it is paramount to study the trade that happens across their borders and view the effects of sharing a border on trade. The United States-Mexico-Canada Agreement (USMCA) is a trade agreement which replaced the 1994 North American Free Trade Agreement (NAFTA), which came into effect on July 1, 2020. NAFTA created a free-trade area for U.S.-Canada and Mexico with no tariffs over time. The USMCA applies additional digital trade provisions, plus environmental standards and labor regulations, and agricultural guidelines to NAFTA. The agreement functions as a major pact because it creates trade regulations for U.S.-Canada and U.S.-Mexico that affect supply chains and market accessibility for American businesses and consumers.

Under the existing free-trade agreement, President Donald Trump of the U.S. imposed a 25% tariff on Canadian and Mexican imports through the International Emergency Economic Powers Act (IEEPA) in February 2025. The imposed tariff would raise product prices, which might affect U.S. industries dependent on exports and potentially cause disruptions in supply networks and cost increases when producing goods. The amount of border crossings toward the U.S. may be changed because trade operations primarily occur at borders. For example, if there is less trading that happens due to the tariffs, there could be fewer border crossings taking place.

Section 2.2: Travel and Tourism in the United States

The travel and tourism industry is a powerful market that contributes to a country's economy. It has a leading market in the U.S, being the world's largest industry and contributing $2.36 trillion to the national economy in 2023, according to "The World Travel & Tourism

Council (WTTC)”. This sector not only helps in supporting numerous jobs in the world but also contributes to the increasing Gross Domestic Product (GDP) of the respective countries. Travel reduced in the COVID-19 pandemic due to restrictions on travelling and lockdowns. The number of travelers who come to the U.S, has not returned to the pre-pandemic levels yet, but the incoming traveling and tourism crossings are likely to go up and get to the pre-pandemic level by the end of 2025.

Both border crossing volumes, travel, and tourism affect one another. For instance, more travel and tourism into the U.S could lead to higher numbers of border crossing volumes. Consequently, higher volumes of border crossings could lead to higher wait times at the ports of entry, which could demotivate tourists who want to travel to the U.S. The strict U.S border security leads to people planning on banning travel to the U.S this year, especially in Canada. An article named “Trump’s bluster and border arrests cool Canadians’ appetite for U.S. tourism” reveals that there has been a more than 22% drop in foreign travel to the U.S between March 2024 and March 2025 (News, 2025). That is a significant drop in numbers and shows that travel difficulties can lead to fewer border crossings, which shows how interconnected border crossing volumes and tourism are.

Section 2.3: Existing Research on Tourism and Border Crossings

The research paper “US-Mexico border tourism and day trips: An aberration in globalization?” by John Berdell and Animesh Ghoshal studies the post 9/11 US border policy changes on short-term tourism and day trips across the U.S.-Mexico border and how they affected the border crossings. The study uses both Ordinary Least Squares (OLS) regression and Autoregressive Integrated Moving Average (ARIMA) time series models to look at the differences before and after 9/11. The results showed a 24% decrease in day trips after implementing stricter border security measures in 2008, with the decrease deepening to 60% by 2012. John Berdell and Animesh Ghoshal conclude that although they used aggregate data, there is not strong evidence to suggest a causal link between border security policy changes and reduced day trips and short trips (Berdell & Ghoshal, 2015). Similarly, this research also uses OLS, ARIMA, and SARIMA models for prediction on future border crossings with consideration on how these issues are affected by the existence of disasters.

Section 2.4: Natural Disasters in the U.S

There has been an increase in the number of natural disasters in the U.S over the past few years, which could be due to the increase in global warming and climate change. These disasters disrupt lots of lives, cause damage to infrastructure and properties. The article “Is climate change increasing the risk of disasters?” by Daniel Vernick examines climate change and its effects on natural disasters such as floods, hurricanes, tornadoes, and wildfires. It is beneficial to understand how each type of disaster affects infrastructure and human movement, as this can directly influence cross-border travel and trade disruptions.

Oftentimes, natural disasters affect communities disproportionately, and some communities might be affected more than others. Historically, minorities or people of low socioeconomic status tend to be more affected in disasters because of the systemic disadvantages through which they lack access to critical items such as stable housing and transport. For instance, discriminatory housing practices have led many low-income and minority populations to end up living in areas that are prone to natural hazards such as flood zones. In cases when major floods or hurricanes take place, these communities are hit hardest. Internal migration might also expand as more people migrate to safer places. The same will apply to people from neighboring countries such as Mexico, or Canada who can run to the United States for refuge in case natural disaster affects their country and it is severely hit. This is applicable to this study since it outlines the way how natural disasters can affect patterns of migration and therefore lead to variations in border crossing volumes, especially when it comes to humanitarian purposes or displacement.

Section 2.5: Natural Disasters and Migration

The research study "Natural Disasters, Economic Development, and Humanitarian Aid" by David Stromberg examines global natural disaster distributions as well as their sources and consequences through data analysis of EM-DAT database records spanning from 1980 to 2004. Between 1980 to 2004, natural disasters brought death to 1.5 million people while creating disturbance for 4.7 billion individuals. Reported disaster numbers have risen by a 5% annual increase, yet mortality rates have shown continuity because improved warning systems, together with enhanced infrastructure, have decreased per-person disaster susceptibility. The research explores the reason why high-income nations survive natural disasters with lower mortality rates than low-income countries, even though their exposure to potential risks remains consistent.

Research established that economic development, government effectiveness, and democratic institutions, as well as equal income distribution, result in reduced vulnerability. During the period from 1980 to 2004, high-income countries experienced similar disaster events as low-income countries, yet low-income countries recorded significantly higher death tolls, even though populations at risk were comparable. Statistical data shows that low-income nations faced disaster fatalities that amounted to four times the rate of those in other countries. The research evaluates disaster fatalities by using regression analysis while it incorporates three controlled factors, including disaster severity levels and population numbers, and risk zones. By analyzing the variables of income and government effectiveness correlated with politics and income equality, it demonstrates a clear negative relationship with disaster fatalities. The author acknowledges that complete control of all variables is not possible so the results should be reviewed carefully. Research indicates that both economic development teamed with strong governance systems work as leading factors to reduce the fatalities from natural disasters.

Historically, natural disasters played a significant role in migration patterns from ancient civilizations to modern times. The research paper “Natural Disasters and Migration”, by Ariel Belasen and Solomon W. Polache, explores natural disasters and their impact on migration by looking at historical examples and applying models such as Generalized Difference-in-Difference (GDD). The study reviews 52 studies that support the migration because of disasters. The event study uses the GDD model and analyzes the impact of hurricanes on Florida counties. The model suggests population growth rates in disaster-hit counties drop by 75% compared to the non-disaster counties. It also says that individuals in urban areas are more likely to migrate than those in rural areas. This study also shows clear evidence that natural disasters can prompt migration when there is an opportunity to improve the quality of life.

There has already been research done that uses machine learning to forecast and predict border crossings based on just the factors that relate to border activity. Additionally, some studies examine the effects of weather on traffic on roads, but there has not been any study that was conducted particularly on the natural disasters being a factor to predict volumes. This research fills in the gap by exploring machine learning models, such as forecasting and predictive analysis, to see whether disaster impacts the volumes of border traffic or not.

Chapter 3

Methodology

The research uses two data sets. The primary data set is the Border Crossing Entry data, which is from the U.S. Department of Transportation's Bureau of Transportation Statistics (BTS). The data was exported as CSV and has 394,866 observations and 10 variables that contain historical data on incoming vehicle, container, passenger, and pedestrian counts in the U.S.-Mexico and U.S.-Canada land border ports. This data set accounts for the years 1996 to 2023 and has variables, Port Name, State, Port Code, Border, Latitude, Longitude, and Point, which are the geographic variables. There is a Date variable that has the year and month of the border entry. The variable Measure is the vehicles, containers, and pedestrians entering the U.S. from the ports. Value is the number of each Measure that was entered that month in a given year. For example, in December 2024 (Date), there were 20000 (Value) Trucks (Measure) that entered from El Paso (Port Name), which is in Texas (State) at the U.S.-Mexico border (Border). Each month of the year, there will be a value for each type of measure in a certain port. So, taking the El Paso example from before, we can expect there to be a Value for each of the measures in December 2024 in El Paso.

State	Port	Border	Date	Measure	Value
Texas	El Paso	U.S-Mexico	Dec 2024	Trucks	20000
Texas	El paso	U.S.-Mexico	Dec 2024	Bus	67400
Texas	El Paso	U.S-Mexico	Dec 2024	Personal Vehicles	70800

Table 3.1: *Example of Border data set*

The second data set that was used is the US Natural Disaster Declarations data set from Kaggle. This dataset includes information on natural disasters that were declared from 1953 to 2023. It is a summary dataset of all federally declared disasters that came from the FEMA website. There are 64,092 observations and 23 variables. The variables include state, designated area, place code, and FIPS, which is a five-digit county code, that are geographic data. FEMA declaration string, the disaster number, declaration type, declaration date, fiscal year that the disaster was declared, incident type, and declaration title are some variables that are related to the declaration of the disaster. Individuals and Household Program, Individual Assistance Program, Public Assistance Program, and Hazard Mitigation Program are binary flags indicating whether the programs were declared for the disaster. There are variables such as state, designated area, declaration date, and incident type that are key variables that will be used throughout the research.

State	Designated Area	Declaration Date	Incident Type
TX	Statewide	2005-09-21	Hurricane
UT	Salt Lake (County)	2020-07-09	Earthquake

Table 3.2: *Example of Disaster Declaration data set*

Section 3.1: Data Cleaning

The border crossing entry data set is exported as a CSV and loaded as a data frame using the pandas library. The border data is then looked at for missing data, inconsistencies, and duplicate values. The dataset had no missing values or duplicate values. The natural disasters data set was also loaded as a pandas data frame. The information about the data set was viewed, such as missing values, data types of variables, and outliers. There were 46,339 missing values in the variables like last_ia_filling_date, 15,180 missing values in disaster_closeout_date, and 8,410 in incident_end_date. These variables were dropped since they were variables that were not relevant to my research. Along with them, fema_declaration_string, fips, place_code, last_refresh, hash, declaration_request_number, and id were dropped since they don't contain any important information that could be of use to my research. The declaration date and the incident start date are two dates that were provided, the declaration date will be used for my analysis since both of them are the same value. The declaration date is of object data type, which is then changed to datetime data type as it would be needed for analysis and modeling purposes.

Since both datasets will be joined, the values in the variables that will be joined should be consistent. The two data sets can be joined with each other on similar variables; one similar variable that is in both datasets is State. In the disaster data, the states are abbreviated as "TX", whereas in the border data, the states are in full forms like "Texas". The states were mapped using the pandas map function for state abbreviations to make their full names. Another variable that needs to be consistent for joining is the Date variable. The border data has a Date which is in Month and year format like "Dec 2023", and the declaration date in disaster data is in Year, Month, and Day format like "2023-12-15". Since the data sets are at different levels of granularity, they need to be made at the same level to join. The date variable in border data is an object data type, which is changed to a datetime type using the pandas to_datetime function. Then, the date of both data sets is split up into two separate columns, one for the Year and one for the Month. In the border data, the Year and Month columns are created, Year has the year like "2023", and Month has the month like "12". The declaration date is also split into Year and Month columns, just like the border data set.

Section 3.2: Exploratory Data Analysis (EDA)

It is important to clearly understand the variables in both datasets before joining and modeling, which is why EDA is performed. First, the two datasets are explored separately, starting with the border data. Distribution of states and ports is done using geographic mapping and heatmaps. To visually see the number of ports in each state/border, a Census Bureau cartographic boundary shapefile is downloaded from the U.S. Census Bureau website and read using the geopandas library. Then, the Mexico and Canada borders are filtered and plotted separately, with states plotted in light grey, the ports in Mexico border plotted in blue, and the ports in Canada border plotted in green.

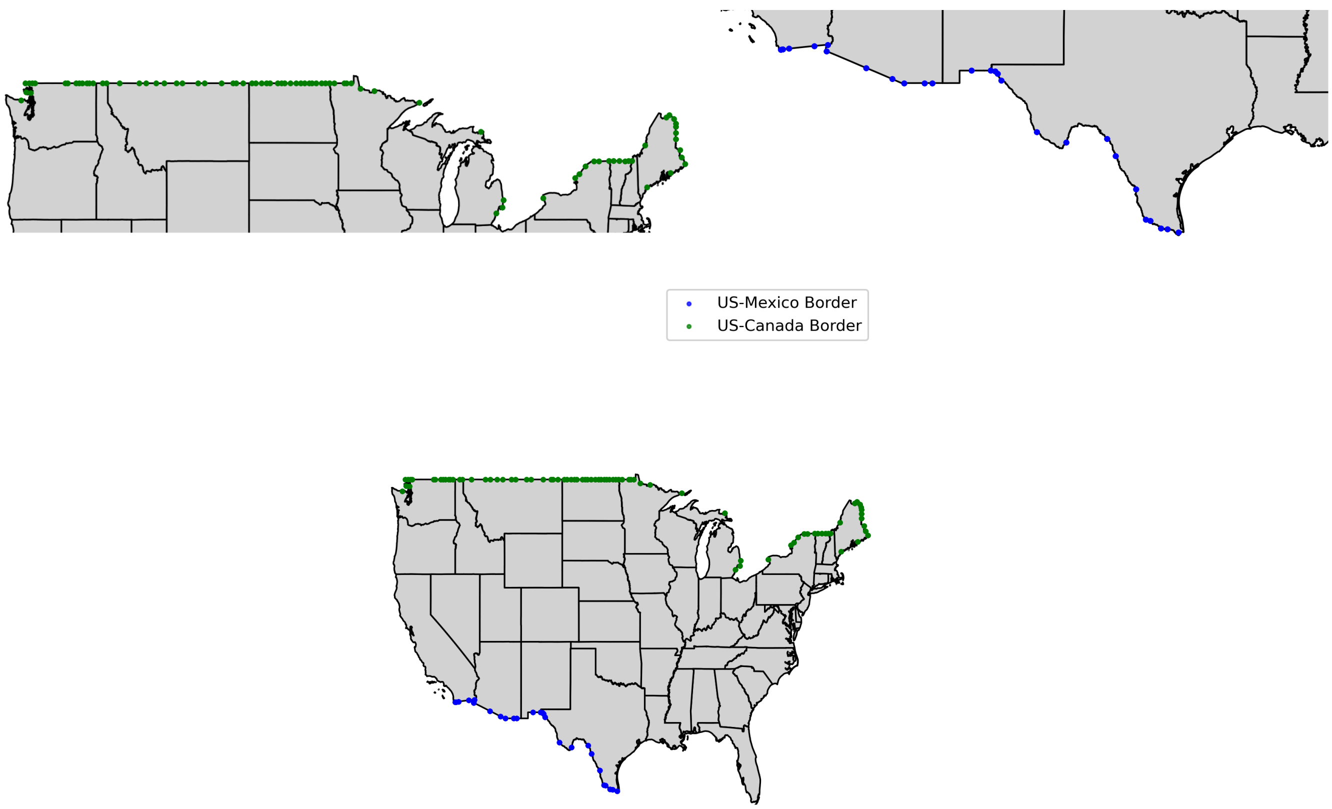

Figure 3.1a: *Distribution of Ports in U.S.-Canada(green) and U.S.-Mexico(blue)*

Figure 3.1b*: Zoomed view of ports at U.S.-Canada (Left) and U.S.-Mexico Border (Right)*

The green and blue dots in *Figure 3.1a* represent the number of ports in each state at their respective border. This shows that the U.S.-Canada border has more ports that span from the west coast to the east, covering almost all the states on the border. There are ports in Alaska and some ports in the south of Maine that are connected to international waters. These ports are maritime ports of entries and customs inspection ports for international vessels and they are mainly utilized by ferries or cargo ships that come from Canada. The U.S.-Mexico border has fewer ports compared to the Canada border. *Figures 3.1b* and *3.1c* show the zoomed view of the maps at the U.S.-Canada and U.S.-Mexico border. U.S.-Canada has 120 unique ports and U.S.-Mexico has 26 unique ports.

Variable	**Sum**
Total Border Crossings (Value)	11,383,189,161

Table 3.3*: The sum of Total Border Crossings (Includes both people and transports)*

Table 3.3 shows that there are a total of 11,383,189,161 border crossings that happened throughout 1995 to 2024, which is about 30 years. Next, a choropleth map of U.S border crossings aggregated by state is generated. This also uses the Census Bureau cartographic boundary shapefile (*Cartographic Boundary Files - Shapefile*, 2021) and maps the total volume of border crossings for each state at the borders. Using the geopandas library, the shapefile is loaded and merged with the border data for the state. The Matplotlib library is used to visualize the map, and the map is then colored in a way that the dark colors represent higher border traffic and the light colors represent lower border traffic in those states.

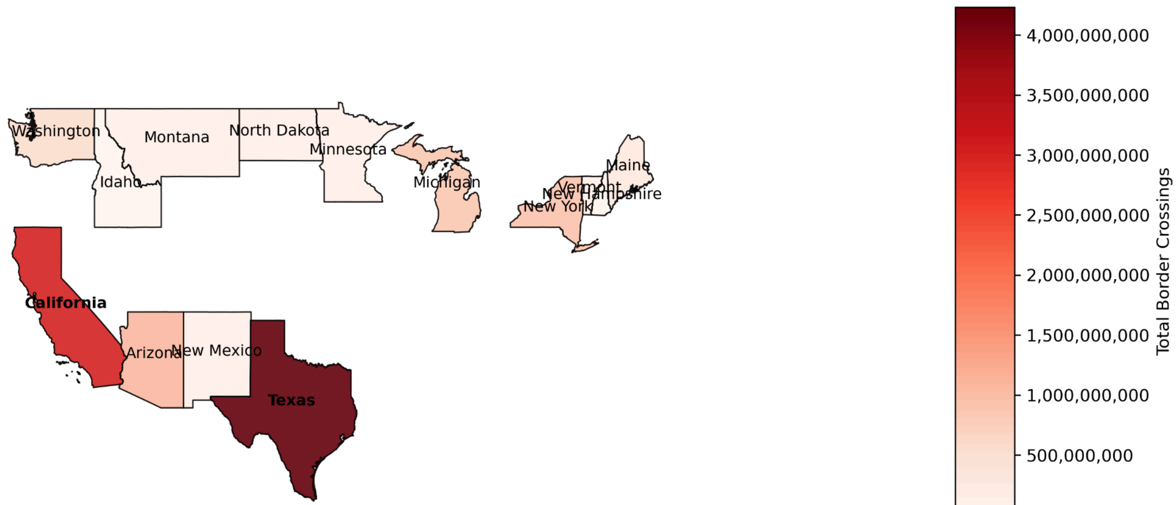

***Figure 3.2**: Choropleth Map of the Total Border Crossings by State*

In *Figure 3.2*, Texas is the darkest colored state on the map, followed by California. This suggests that Texas is the busiest border state with 4 billion total crossings, and California also experiences high traffic, with 3 billion crossings. In comparison, northern border states see fewer crossings since they are lighter in color. These states appear almost white, which is on the lower end of the scale, the crossings have between 0 and 5 million. This could suggest that there is less border crossing activity in those states either because of less trade and tourism happening there or because there are stricter border regulations which makes people prefer the Mexico border.

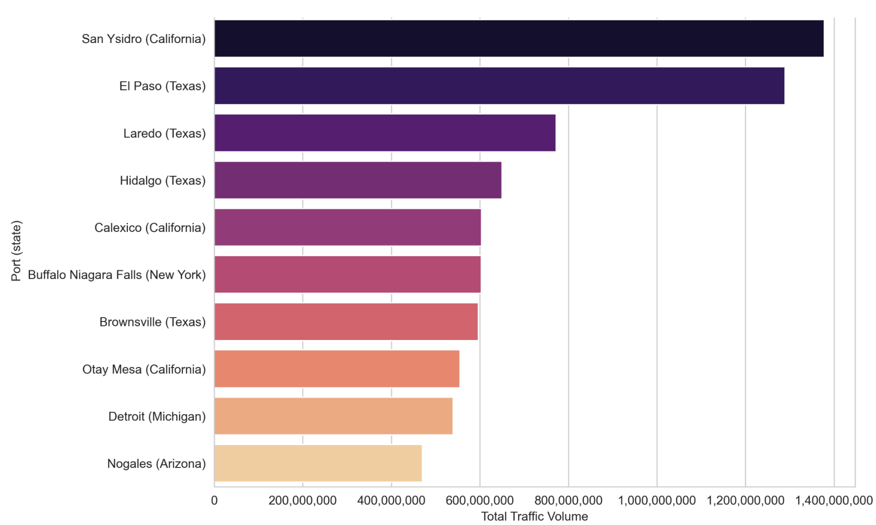

***Figure 3.3**: Bar chart of the Top 10 Border Ports*

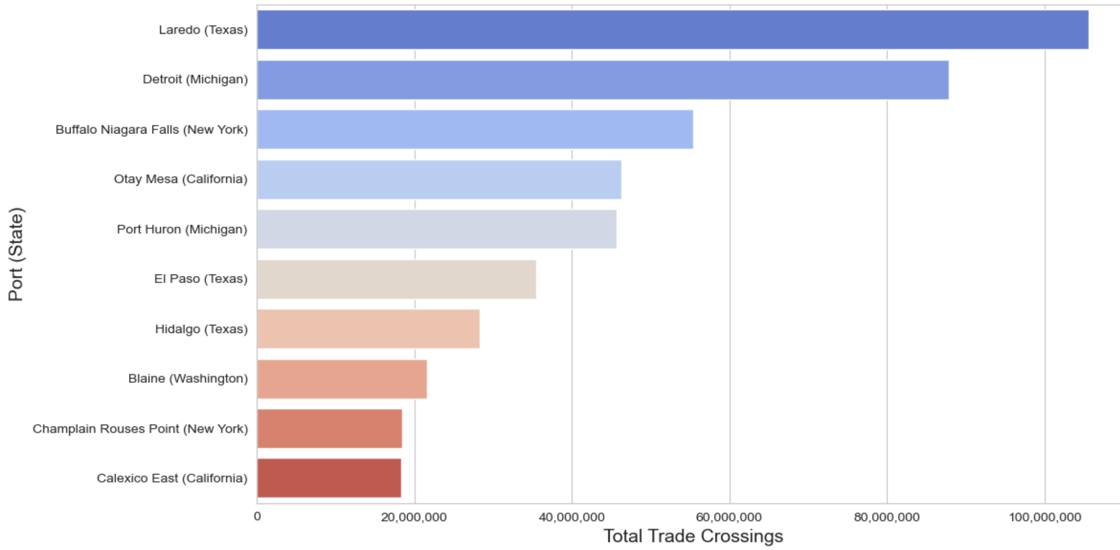

***Figure 3.4**: Bar chart of the Top 10 Border Ports with Truck Crossings*

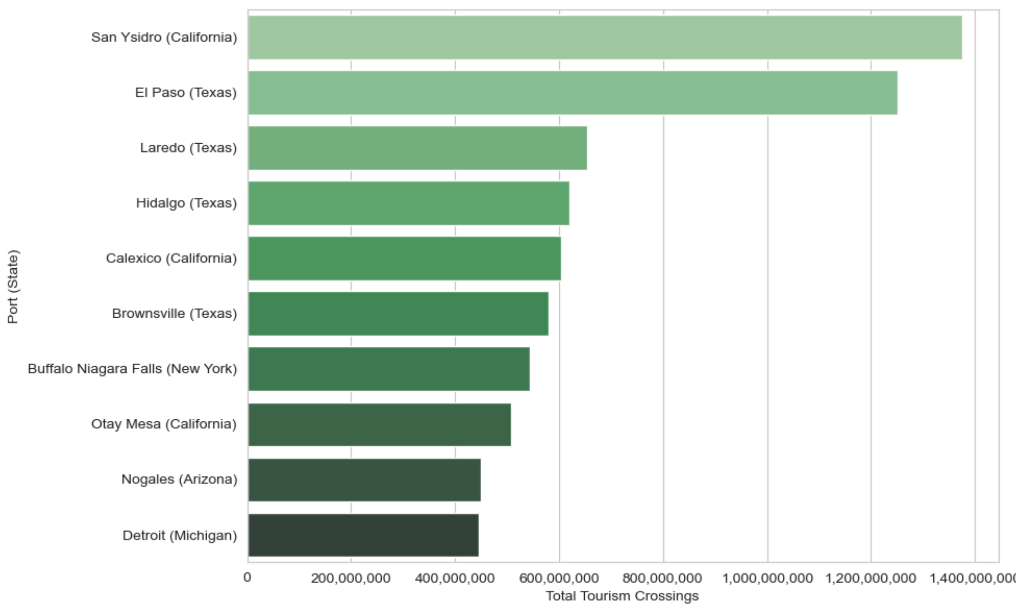

***Figure 3.5**: Bar chart of the Top 10 Personal Vehicle and Pedestrian Crossings*

The top 10 border ports with the most border crossings are visualized using a bar chart by grouping port name, state, and traffic value. *Figure 3.3* shows that the San Ysidro port in California has the highest number of border crossings of around one billion three hundred fifty million. Followed by the El Paso, Laredo, and Hidalgo ports in Texas. From *Figure 3.2*, it was shown that Texas had the highest crossings, which makes sense because four out of the top ten ports are in Texas. Notably, northern ports such as Niagara Falls and Detroit also appear in the top ten. These locations are special cases as they not only can be entry points to the other country, but also one can use them as shortcuts for domestic U.S.-to-U.S. travel. For instance, if someone were to drive from upstate New York to Michigan, instead of taking the long U.S. only route, they could cross into Canada at Niagara Falls and drive through Ontario and reenter at the U.S. port at Detroit. So even if the trip started and ended in the U.S, once someone enters and exits Canada, it's logged as an international crossing. This is important to the research because it demonstrates that not every crossing is an international visit and there are certain travelling shortcuts which are good to know for interpretations of traffic data and modeling its effects.

Then, a port analysis with the measures for trucks and truck crossings is created in *Figure 3.4,* along with measures such as pedestrians and personal vehicles in *Figure 3.5*. This splits the ports by trade and tourism since trade happens mostly using trucks and tourism takes place when it is either by personal vehicles or walking through the borders (pedestrians). For *Figure 3.4* which are the trade ports consisting of truck crossings, Laredo (TX), and Detroit (MI) ports have the highest crossing activity. Other ports such as Buffalo Niagara Falls (NY), Port Huron (MI), and Otay Mesa (CA) also have high activity. For *Figure 3.5,* San Ysidro (CA) is the busiest port in tourism crossings. Texas and California ports are also prominent. There are only a couple of ports in the North, like Detroit, Buffalo Niagara Falls. There are major trade-focused ports both in northern and southern regions, and ports that are primarily for tourists lie in the southern regions, which shows more cultural and personal connections with Mexico. Ports such as Laredo, Otay Mesa, and El Paso handle large amounts of trade and tourism, which makes them indispensable and in need of good infrastructure and staffing. Texas, having the highest number of ports with total crossings, is notable and highlights its central role in U.S.-Mexico trade and tourism.

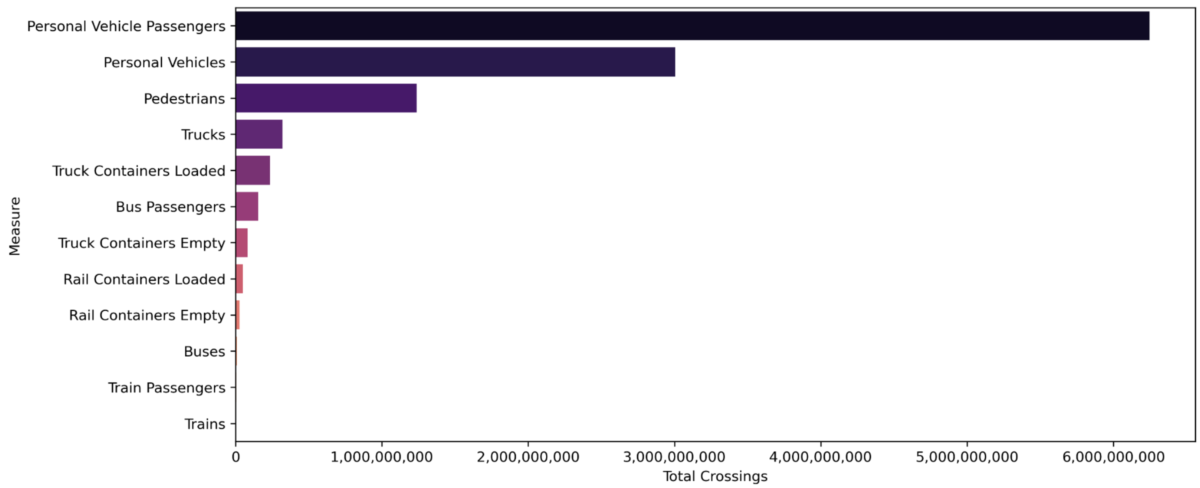

Figure 3.6: *Bar chart of Total Border Crossings by Measure*

A graph with the number of total crossings is visualized for each of the measures. The measures are the types of vehicles that come into the United States through the borders. In *Figure 3.6*, each bar represents the total count of crossings across all dates and ports for that vehicle type. Personal Vehicle passengers, which count each of the passengers in a personal vehicle, have the highest volume, with over 6 billion crossings; this shows that most people prefer to cross the borders in their vehicles. Trains and buses account for a very small part of the border crossings.

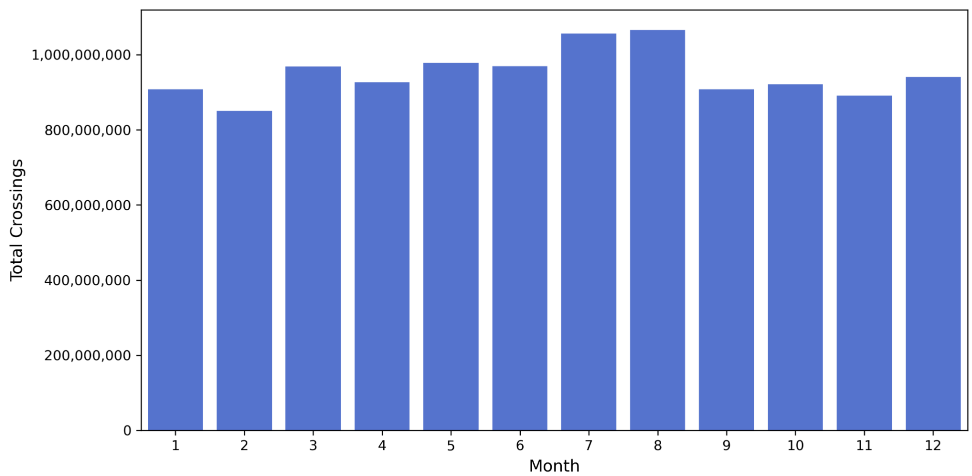

Figure 3.7: *Total Border Crossings by Month*

Figure 3.7, groups by month and sum of the border values, and plots a bar chart of total border crossings by month, is plotted with the seaborn library. The graph shows that the highest border crossings took place in July and August. This could be due to the weather conditions, summer months could see a high in border crossings because of people traveling for vacation. The lowest crossings took place in February. The winter months see a low because of the cold weather, or bad road conditions.

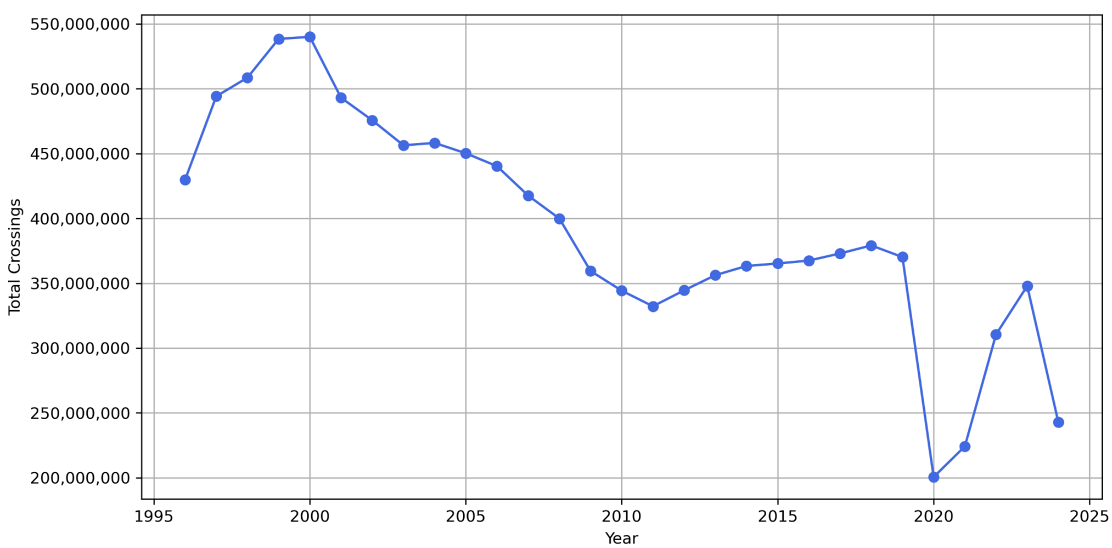

Figure 3.8*: Total Border Crossings Over Years (1995-2024)*

The line graph of the border crossings *(Figure 3.8)* is made by grouping the years and the sum of the border crossing values. The total border crossing analysis reveals that 1998-1999 registered the greatest number of movements at about 540 million, whereas 2020 demonstrated the fewest border crossings, reaching 200 million. Travel restrictions and the quarantine measures combined with COVID-19 caused a significant drop in border crossings in 2020. From the early 2000s until now, there has been a steady decline in border crossings, which this graph displays.

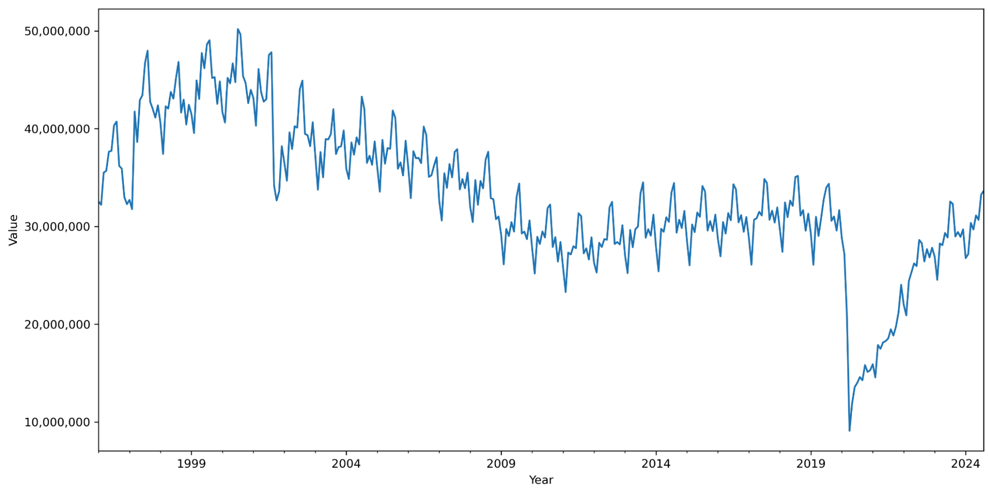

Figure 3.9*: Monthly Border Crossings over Time*

Border data over time is graphed by doing a time series analysis on border data and value, which represents the count of measures and visualizes it in a monthly line chart. The data is aggregated by month and summed by value. This is important to do to see seasonality or any long-term trends. The graph in *Figure 3.9* shows a rising trend until around 2001-2002 and then declines and has a stable trend after 2002. There are noticeable spikes each year, which suggests seasonality. There is a sharp drop in 2020 due to the border closures or travel restrictions during the pandemic, and crossings gradually recover after 2020 through 2023.

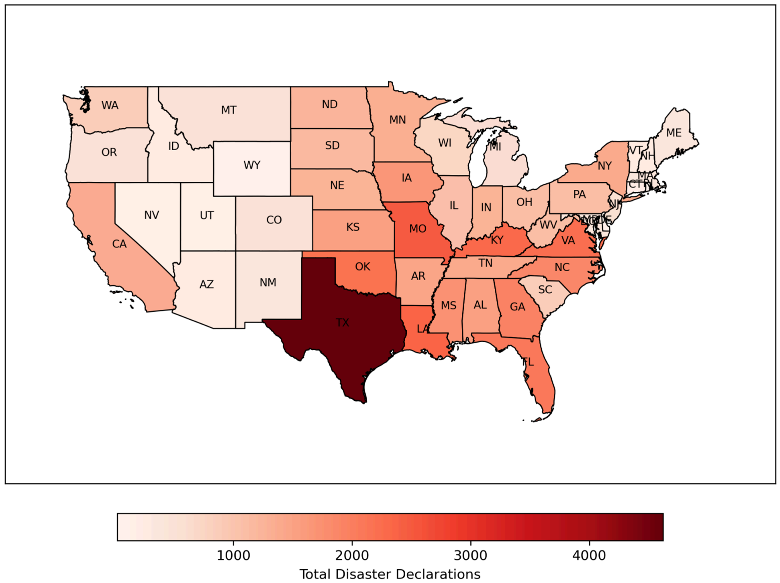

***Figure 3.10**: Choropleth Map of the Total Disaster Declarations by State*

Now, EDA is performed on the disaster dataset to learn more about the important variables. The distribution of disasters by state is examined for the disaster dataset. The disaster data is grouped by state, and a count is taken for the number of disaster declarations. The same Census Bureau cartographic boundary shapefile (*Cartographic Boundary Files- Shapefile*, 2021) is used to merge with the disaster data on state level. Then, a choropleth map is used to visualize the disasters that were declared in each state, the dark colors representing high disaster declarations and the light colors representing low disaster declarations.

Texas stands as the most disaster-declared state in *Figure 3.10* because it appears in the darkest color shade. Texas maintains its position as the second largest state regarding size while being the main target of hurricanes, floods, and tornadoes. Southern states like Florida, Louisiana, Georgia, and North Carolina also show high disaster counts. Midwest states like Kentucky, Missouri, and Arizona show moderate counts, and northwestern states are lighter, which could mean fewer disasters. This could be due to factors such as the population sizes, the geography of the states, and the state policies in reporting a disaster.

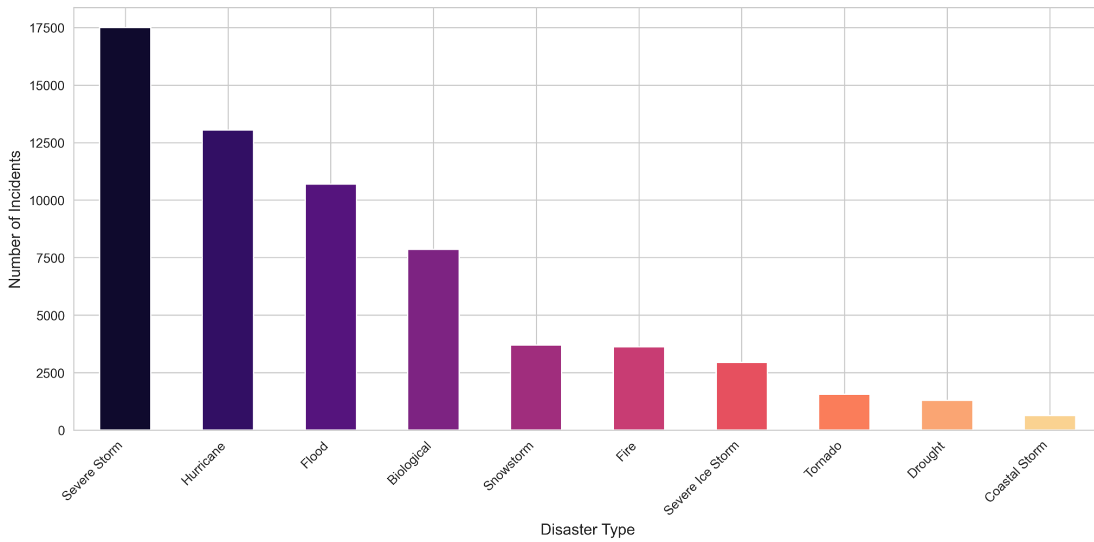

Figure 3.11: *Bar chart of the Top 10 Disaster Types*

The top 10 FEMA-declared disasters by disaster type were generated using the seaborn library. In *Figure 3.11*, severe storms are the most common declared disaster, followed by hurricanes. This could also be due to not only the number of times they occur but also the wide geographic spread and financial damage that they cause resulting in more declarations on federal levels. Contrary, Coastal Storms and Droughts have the least number of declarations as they are confined geographically or take a long time to form thus resulting in fewer emergency responses. There is also an incident type categorized as biological, which is the COVID-19 pandemic and the declarations. There have been around 7500 incidents of COVID-19, but this could be because the data is at a county level, and the incident could have been declared in many counties.

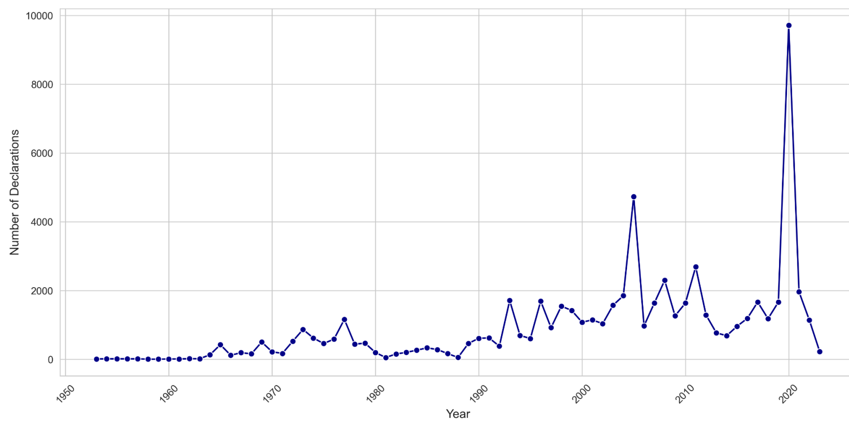

***Figure 3.12**: Time-Series Graph of Disaster Declarations by Year (1950-2023)*

A time series graph is created for the disaster declarations over the years by extracting the years from the variable declaration date and counting the declarations per year. The graph in *Figure 3.12* suggests that the highest number of disasters was declared in 2020, and declarations of disasters were low before 1992, with only a few a year. The declarations beginning to grow from the 1990s could be due to the rise in technology and more ways to declare disasters. Before the 1990s, it might have been harder to report the disaster, especially if they happened in only a few places. The rise in declarations in 2020 is because of COVID-19 declarations that were made in many counties and states, since it was a global pandemic.

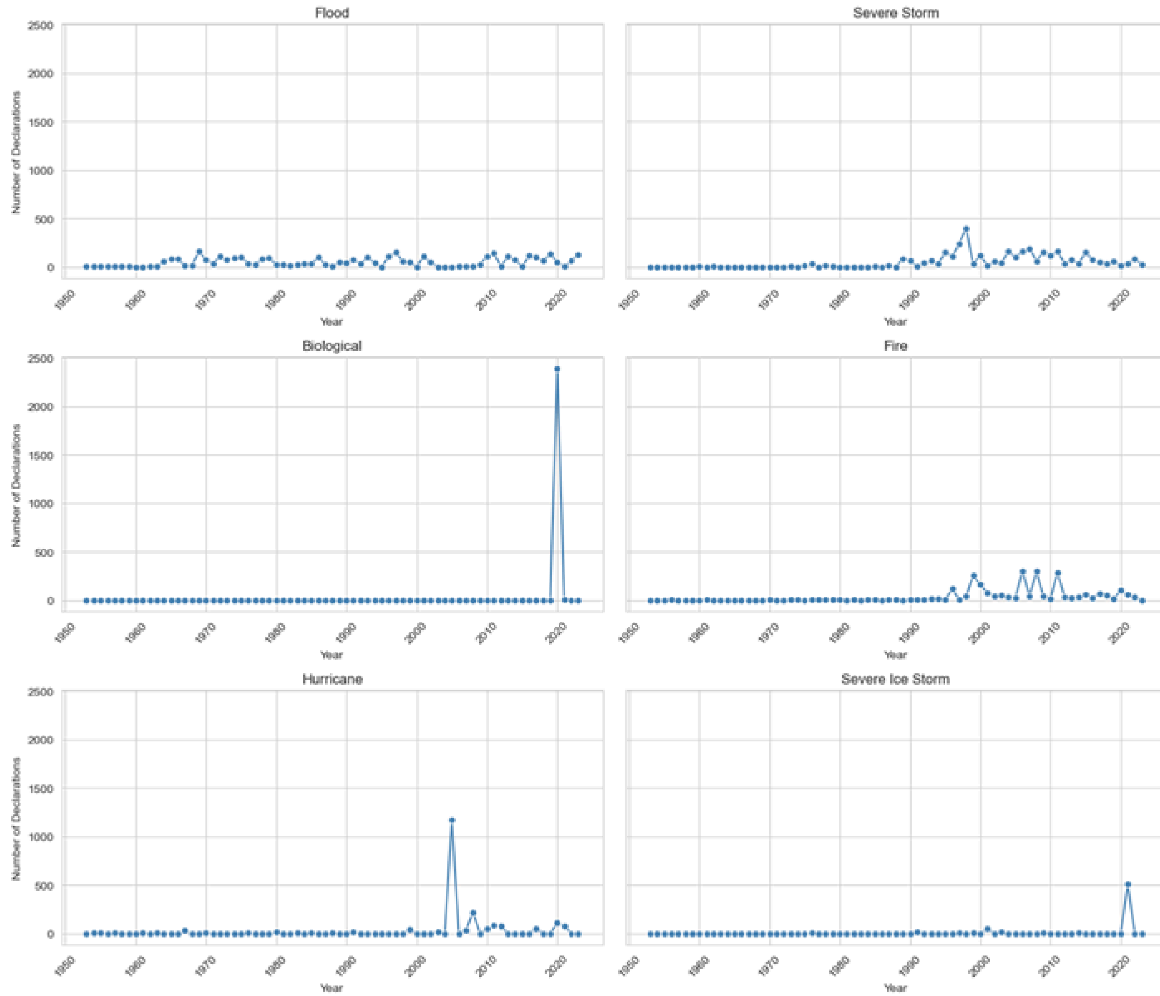

***Figure 3.13**: Disaster Declarations by Year for Top 6 Incident Types*

To look more into the disaster declaration by year, the top 6 disaster types are seen over time using the count of disaster declarations per year in the U.S, grouped by the disaster type. The graphs from *Figure 3.13* show some important trends that were not visible in *Figure 3.12*, it gives a clear view into the peaks and the drops. The flood graph (Top Left) shows a degree of consistency regarding low levels. There are several peaks around the 1990s as well as the early 2000s, in addition to another slight increase throughout 2015-2020. The severe storm graph (Top Right) shows a distinct increase in storms around 1990 with peaks in the early 2000s and again 2010-2015, along with a slight decrease after 2015. The peak within Figure 11 during 2005 could be the reason as well. Values are shown in the biological graph (Middle Left) only for 2020 and 2021, plus there is a sharp drop from 2020 to 2021, reflecting the declarations related to COVID-19 in 2020. The Fire (Middle Right) graph has very low counts until the 1990s and more consistent declarations from the 2000s, likely because of the increasing wildfires. Consistent declarations are shown on the hurricane graph (Bottom Left) along with a sharp spike during 2005, plausibly corresponding with Hurricane Katrina. The severe ice storms graph (Bottom Right) shows a greatly low frequency throughout the years, with some activity occurring in 2000 and a peak in 2005.

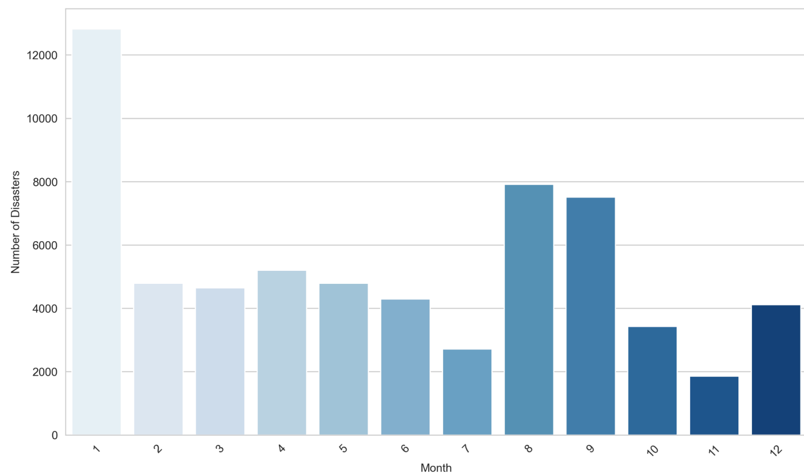

Figure 3.14*: Number of Disasters Declared by Month*

Disaster declaration by month is looked at by grouping and counting months from the disaster declaration and the disaster number. The bar graph in *Figure 3.14* implies that January has the most disaster declarations, while November has the lowest disaster declarations. January might have the highest declarations since it is the peak winter, which often has snowstorms and flooding. Some declarations that happen late in the year might also be declared or processed when the calendar year begins. November could have the lowest declarations since it is a transition period between hurricane season and winter storms.

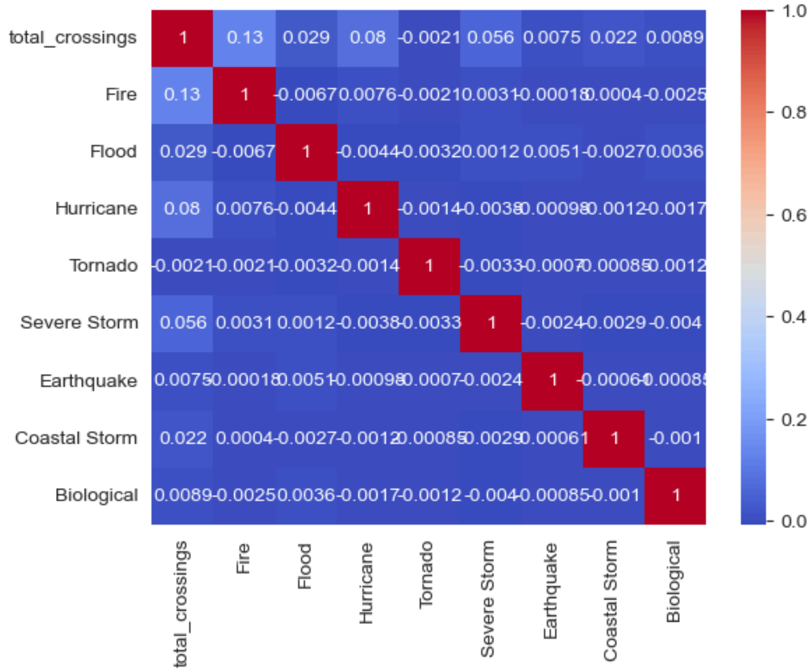

Figure 3.15: *Correlation Between Disasters and Border Crossings*

A correlation heatmap was created to see the correlation between the disaster types and total border crossings. In *Figure 3.15*, only the top row is meant to be looked at for the correlation. There is a slight significant correlation between total crossings and the chosen disaster types. Despite this, the data shows a faint negative pattern between tornado occurrences and the total number of crossings at -0.0021. The second most common disaster type, called fire, exhibits a lighter blue coloration and a correlation value of 0.13, indicating a moderate link between its occurrences and others.

State	Year	Month	Total disasters
Texas	2005	9	509
Texas	2005	8	254
Texas	2005	11	254
Texas	1999	8	244
Texas	2008	3	239

Table 3.4: *Top 5 Highest Total Crossings*

Then, a sum across the disaster types by the disaster occurrences was taken and put into a new variable and sorted to find the worst month in terms of disasters and the highest disasters. The period of September 2005 in Texas marks both the worst year alongside the most disaster events with 509 occurrences of total disasters as recorded through *Table 3.4*. During August and September of 2005, Hurricane Katrina, along with Hurricane Rita, struck the region. There might have been a lot of declarations in 2005 since FEMA probably declared many individual disasters across many counties. There is one disaster declaration for every county that qualifies for federal aid, and Texas has the highest number of counties.

Section 3.3: Comparative Analysis of U.S.-Mexico and U.S.-Canada Border

An analysis of the differences between the U.S.-Canada and U.S.-Mexico is performed, where we split the border data by both borders. Both borders are different geographically, and could be used to cross for different purposes. For instance, one border could be prone to tourism, and the other could have more goods being imported. Analyzing both the borders separately could be helpful to implement border-specific or state-specific policies for better disruption preparedness.

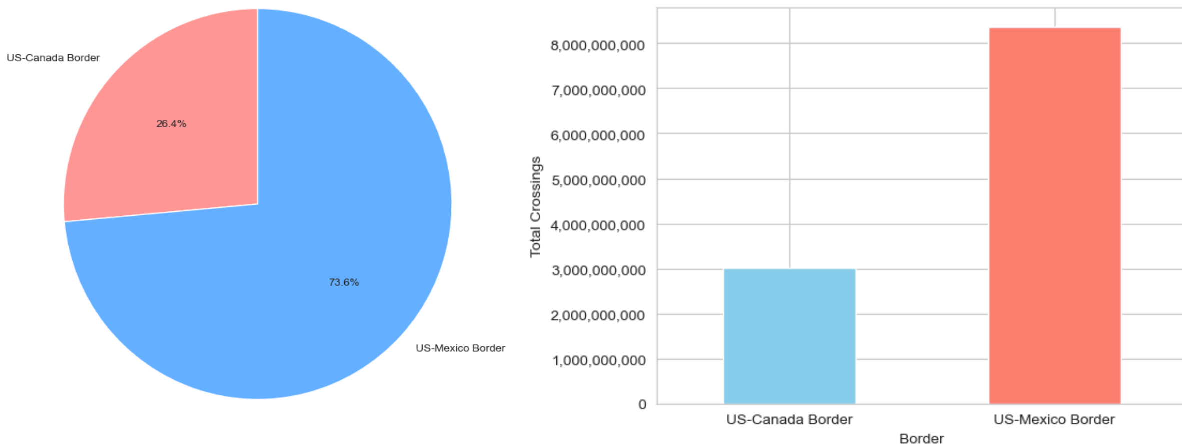

Figure 3.16: *Volumes by Border: Pie Chart (Left) and Bar Chart (Right)*

The analysis of border distribution between the U.S.-Canada and U.S.-Mexico uses a pie chart together with a bar chart, which presents borders and traffic volume data in *Figure 3.16*. The U.S.-Mexico border sector represents 73.6% of the overall port traffic, whereas the U.S.-Canada border manages 26.4% of the total movement in *Figure 3.16*. Total border crossing statistics reveal that the U.S.-Canada border experiences 3 billion crossings, but the U.S.-Mexico border observes above 8 billion crossings. This could be because the U.S and Mexico are major trading partners, and a large portion of trade happens over land. This could also be because the U.S.-Mexico border passes through the state of Texas, with popular ports that are highly busy.

It is important to note that both borders have different types of traffic. According to the article "*What the data says about immigrants in the U.S.*", U.S.-Mexico border has more immigrants that come in compared to the U.S.-Canada border, since people seek asylum in the U.S. from Mexico (Moslimani & Passel, 2024). And, the article "*The number of migrants crossing the US-Mexico border is likely to keep growing.*" explains why there is an increase in the Mexican border crossings and because of the migrants who are likely to seek asylum in the

U.S. It says that “In July 2022, for example, CBP figures indicate 4,000 Mexican family encounters at the border. A year later, that number had more than quadrupled, reaching nearly 22,000” (Shoichet, 2023). Although the data set that is used for research does not have data on immigration and migration, this still shows how migrants and immigrants could account for many of the crossings that come into the U.S.

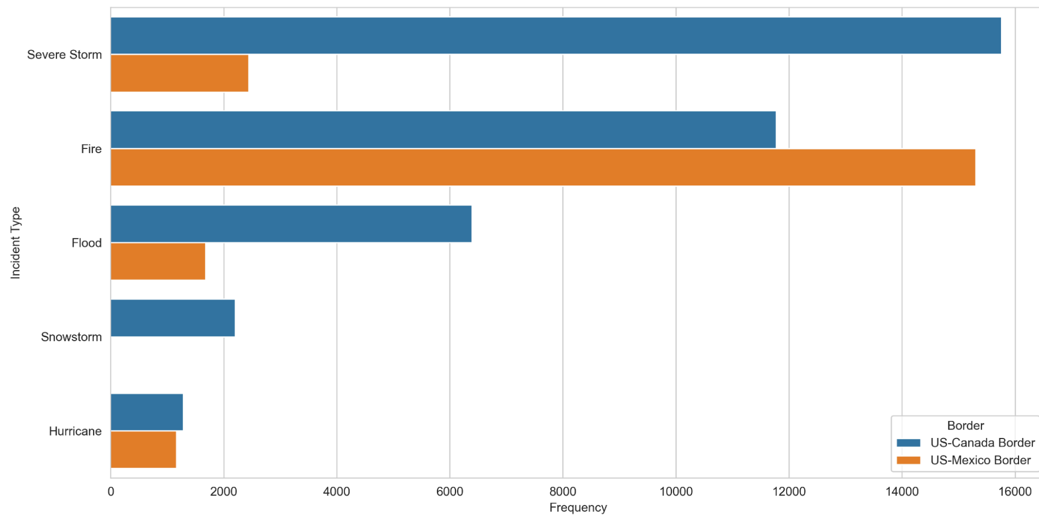

***Figure 3.17**: Comparison of Incident Types between Borders*

A group bar graph is created to compare the top 5 common disaster types (Severe Storms, Fire, Flood, Snowstorm, and Hurricane) between the two borders. *Figure 3.17* suggests that severe storms have the highest declarations in the U.S.-Canada border, with a count of 16,000 occurrences. This is less frequent along the U.S.-Mexico border, with under 5,000 occurrences. And fires have the highest declarations for the U.S.-Mexico border, where it has 15,000 incidents. It is also frequent in the U.S.-Canada border with over 11,000 occurrences. This means that fire is an issue for both borders, probably because of the wildfires. Floods are a U.S.-Canada border issue, with over 6,000 incidents and significantly fewer in the U.S.-Mexico border, with under 2,000 incidents. Snowstorms are more common in the U.S.-Canada border since the colder climate, and there are no snowstorms in the U.S.-Mexico border because of the warm climate. Hurricane is balanced between both borders, but is still low in total volume.

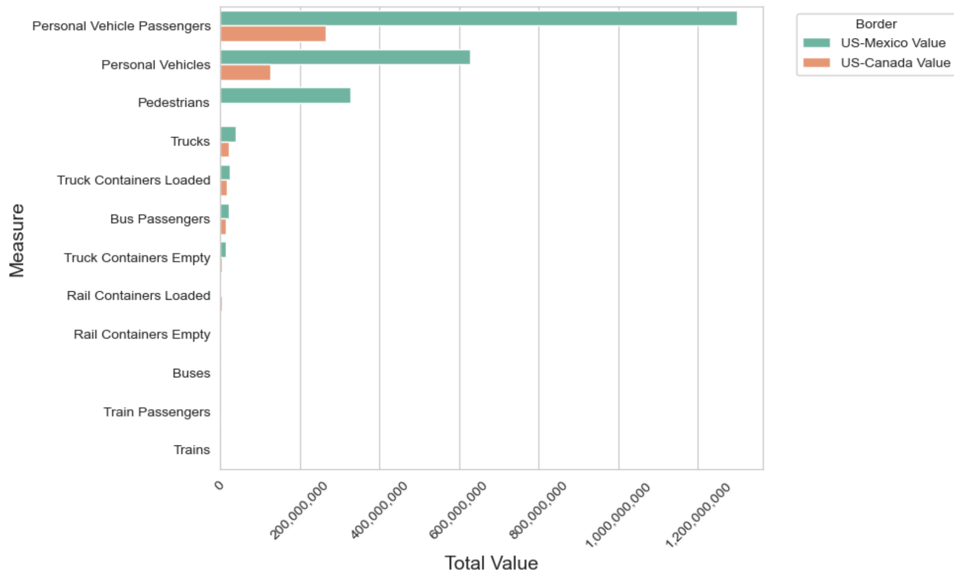

Figure 3.18*: Comparison of Measures By Border*

To compare the vehicle types (Measures) between the U.S.-Mexico and U.S.-Canada border, both borders are grouped by measure and value. Then, a horizontal bar plot is created using the seaborn library to visually compare and analyze the trends. In *Figure 3.18*, personal vehicle passengers have a lot of activity at the U.S.-Mexico border with over 1.2 billion crossings, and have much smaller crossing volumes at the U.S.-Canada border. Pedestrians have high crossings in the U.S.-Mexico border, and not a lot of pedestrians in the U.S.-Canada border, likely due to the climate there. Trucks, containers, and rail traffic exist at both borders but in smaller numbers compared to the personal travels.

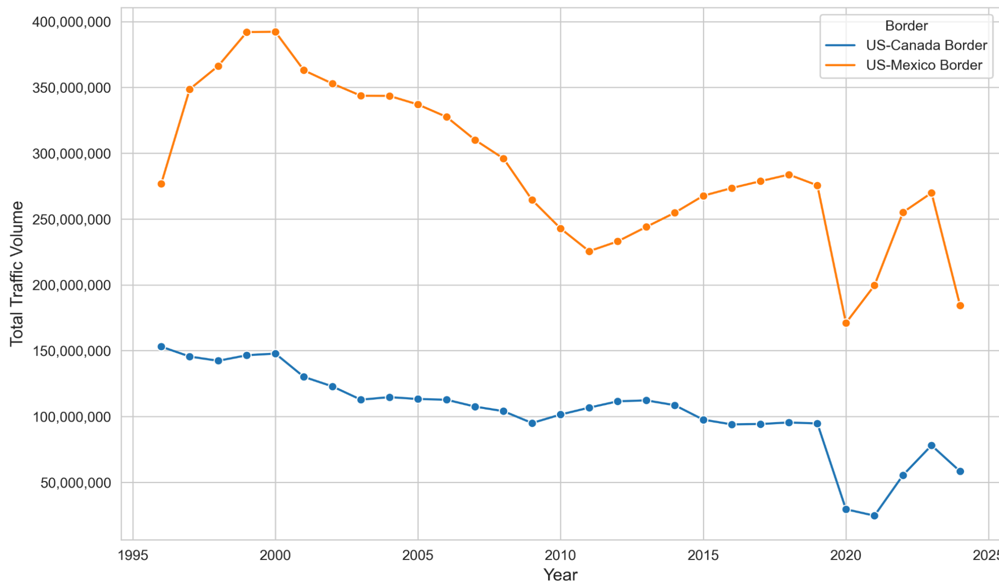

Figure 3.19: *Yearly Border Traffic Volume: U.S.-Mexico vs U.S.-Canada*

A yearly line graph comparing total traffic volumes in the U.S.-Mexico and U.S.-Canada borders is created by grouping the year and border and summing the total traffic volume. The graph in *Figure 3.19* has one line for U.S.-Mexico border traffic over time (Orange) and one line for U.S.-Canada border traffic over time, and it tracks total traffic volume per year from all the measures combined. The U.S.-Mexico Border saw traffic volumes that started at over 250 million crossings, as well as rose steadily to approximately 390 million by the year 2000. Volumes declined consistently across 2011 after such a peak. This could be because of the stricter border security regulations after 9/11 as discussed in Chapter 2. A meaningful decrease to about 170 million happened in 2020. Approximately 150 million crossings marked the start of the U.S.-Canada Border, in comparison, gradually declining to 90 million before a sharp fall to about 30 million in 2020. Overall, this graph also shows that the U.S.-Canada border has fewer crossings compared to the U.S.-Mexico border, though both borders dropped in volume in 2020, likely because of the COVID-19 pandemic.

Section 3.4: Clustering Analysis

Clustering techniques such as K-means and K-Prototypes are used to find any interesting groups or patterns between the datasets. K-means clustering is used to cluster the border data, and it takes in numerical features such as value, which is the border traffic volume, longitude, and latitude, which are the geographical metrics. These three features are selected to use for clustering similar ports together. The features are standardized, which makes the mean 0 and the standard deviation 1 for the features since the features are on different numeric scales. A K-Means model is fit to 4 clusters, which predicts a cluster label (0,1,2,3) for each port. The clusters are visualized using a scatter plot in *Figure 4.1,* where each dot represents a port colored according to the cluster it belongs to. This visual shows how the ports are grouped based on their location and traffic volume. This could show potential outliers or unique ports that need specific attention in terms of research.

A K-Prototype clustering is performed on the K-means clustering features, which essentially adds a categorical feature to the model. A K-Prototype acts as an extension of the K-means clustering, which uses both categorical and numerical features. This algorithm puts numerical data points and assigns them to the nearest Euclidean distance, and puts the categorical data points to the nearest mode. Measure variable, which is the vehicle type, is added to the K-Means algorithm from before. The variable is one-hot encoded since it's a categorical feature, and 4 clusters are fitted and plotted on a scatter plot in *Figure 4.2a*. This clusters on how much traffic there is, the location, and also the type of crossings (measure). Ports that are close together but have different types of vehicles could be placed into different clusters which is shown in *Figure 4.2a.* In *Figure 4.2b,* a bar plot is created in which each bar represents a cluster and it displays the measure that is dominant in each cluster.

Section 3.5: Forecasting Models

To forecast the border crossing values for the future years, models such as Autoregressive Integrated Moving Average (ARIMA), and Seasonal Autoregressive Moving Average (SARIMA) are used. This model is an univariate one since it uses one variable as input to make predictions. In these models we use only the border crossings (Value) which are the historical values of that variable to predict future values and no other explanatory variables. These forecasting models are used because in *Section 2.3,* the study *"US-Mexico border tourism and day trips: An aberration in globalization*" uses ARIMA model to look at the forecast of short trips after 9/11. Similarly, this research uses ARIMA and SARIMA models to forecast border crossings since it has already been done on similar research.

Figure 3.9 shows the monthly crossings over the years, and there were peaks at certain points every year, suggesting seasonality. Seasonality occurs when there are repeating patterns within a time series graph at the same time frame each year. To further prove seasonality, a Kruskal-Wallis test was performed on the border data, which tests for seasonality of the time series. Kruskal-Wallis test takes groups that are the different time intervals within the series and compares them across all intervals. "The null hypothesis states that all months (or quarters, respectively) have the same mean. When this hypothesis is rejected, it is assumed that time series values differ significantly between periods" (*Kruskal-Wallis Test*). This test is done by taking the data column, which includes the timestamp of crossing data and the value column, and resampling is done by summing all values for each calendar month.

- **Null Hypothesis (H0):** The months have the same mean of border crossing volumes.
- **Alternative Hypothesis (H1):** At least one month has a different mean.

The test checks whether monthly border crossing totals differ significantly, which indicates seasonality. If the p-value is less than 0.05, which is our significance level, then the null hypothesis is rejected and there is significant seasonality. However, if the p-value is greater than or equal to 0.05, we fail to reject the null hypothesis, which means that there is no statistically significant seasonality.

Identifying seasonality plays an important role in selecting a suitable forecasting model. If seasonality exists for the border crossing data, then a Seasonal Autoregressive Integrated Moving Average (SARIMA) can be used to forecast the border crossings. SARIMA takes care of handling time series with seasonality and is used to forecast various kinds of data. There are four main components to it: the seasonal component (S), the autoregressive component(AR), the integrated component (I), and the moving average component (MA). To fit a SARIMA model, initializing an order and a seasonal order is crucial, this order consists of *(p,d,q)* and *(p,d,q,s)* (Noble, 2024).

The order for the SARIMA model used is *(1, 1, 2)*, and the seasonal order used is *(1,1,1,12)* for the total crossings which uses values from both the borders. The *p* in the order represents the number of non-seasonal autoregressive (AR) terms, meaning it uses the previous month's value to predict the current. The *d* in the order represents the non-seasonal differencing, which means that it helps remove any trends and makes the data stationary. The *q* in the order represents the forecast error term from the previous two time steps that improves prediction. The *p*,*d*, and *q* in the seasonal order essentially do the same thing, where they use values from the same month in the previous year to predict the current value, remove any repeating yearly seasonal trends, and use the forecast error from the previous year. There is an additional parameter in the seasonal order, which is the s, the s is the seasonal period, which is defined as 12, which represents the monthly data with yearly seasonality.

Order	Seasonal Order	AIC	BIC
(1,1,1)	(1,1,0,12)	10017.64	10032.69
(1,1,1)	(0,1,1,12)	9971.18	9986.22
(1,1,1)	(1,1,1,12)	9970.71	9989.50
(0,1,1)	(1,1,0,12)	10048.41	10059.71
(0,1,1)	(0,1,1,12)	9970.14	9981.42
(0,1,1)	(1,1,1,12)	9970.23	9985.26
(1,1,0)	(1,1,0,12)	10022.16	10033.45
(1,1,0)	(0,1,1,12)	10005.54	10016.82
(1,1,0)	(1,1,1,12)	10005.68	10020.72
(1,1,2)	(1,1,0,12)	10018.62	10037.43
(1,1,2)	(0,1,1,12)	9940.19	9958.97
(1,1,2)	(1,1,1,12)	9939.78	9962.32

Table 3.5: *Order and Seasonal Order Combination AIC and BIC for total crossings*

The parameters *(1,1,2)* and *(1,1,1,12)* for SARIMA were selected by grid search which is a hyperparameter tuning technique that is used to search all possible combinations of parameters. A for loop was created to check the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) score of the different combinations. AIC and BIC are model selection variables that examine the fit of the model, it is best to have a low AIC and BIC since it indicates a better fit. This combination had an AIC of 9940.19 and BIC of 9958.97 which is the lowest among the other combinations (*Table 3.5*).

Order	Seasonal Order	AIC	BIC
(1,1,1)	(1,1,0,12)	9413.18	9428.23
(1,1,1)	(0,1,1,12)	9360.92	9375.95
(1,1,1)	(1,1,1,12)	9363.14	9381.93
(0,1,1)	(1,1,0,12)	9439.93	9451.23
(0,1,1)	(0,1,1,12)	9359.54	9370.81
(0,1,1)	(1,1,1,12)	9361.14	9376.17
(1,1,0)	(1,1,0,12)	9411.96	9423.25
(1,1,0)	(0,1,1,12)	9389.48	9400.77
(1,1,0)	(1,1,1,12)	9391.10	9406.15
(1,1,2)	(1,1,0,12)	9415.23	9434.04
(1,1,2)	(0,1,1,12)	9332.60	9351.38
(1,1,2)	(1,1,1,12)	9333.97	9356.50

Table 3.6: *Order and Seasonal Order Combination AIC and BIC for U.S.-Canada border*

For further exploration of the border crossing forecast model, another model is fit where the border crossings were split into two borders and fit as different models, U.S.-Canada and

U.S.-Mexico. First, the data is filtered by border, and then for each border, the data is pivoted and grouped by month. A Kruskal-Wallis test is performed for both borders separately to see if there is seasonality present in both. The U.S.-Canada border has seasonality and U.S.-Mexico border does not have any seasonality according to the test. A SARIMA is fit for the U.S.-Canada border and an ARIMA is fit for the U.S.-Mexico border. The parameters *p,d,* and *q* for ARIMA and *p,d,q,s* for SARIMA were obtained by testing out a combination of order and seasonal order by Grid Search for the best parameters again for each model separately. A SARIMA and ARIMA model is fit after finding the best parameters with AIC of 9332.69 and BIC of 9351.38 (*Table 3.6*).

The parameters *Order=(1, 1, 2)*, *Seasonal=(0, 1, 1, 12)* were used for the SARIMA model because it was the combination that had the lowest AIC and BIC score. This means that for the order the model uses previous months values to predict the current values, the differencing of 1 removes any long term growth trend, and MA uses the last 2 months of prediction errors. For the seasonal order, the model doesn't use the same month last year directly to predict future values; it uses differencing of 1 to remove yearly seasonality, and uses the seasonal moving average error from the same month last year, and the 12 indicates monthly data with a 12-month seasonal cycle. The ARIMA parameters are *(2, 1, 3)*, since it had the lowest AIC score of 10736.11 after doing Grid Search. This means the number of crossings this month depends on the previous 2 months, the model removes trends, and uses errors from the last 3 months of forecast.

Section 3.6: Natural Disasters' impact on border crossings

An analysis of the disaster months and non-disaster months is done to check how and in what ways natural disasters impact border crossings. To do that, a new dataset is created by grouping state, year, and month, and aggregating it by the border crossing value and disaster counts. Disaster count is a column that was created to show how many disasters took place in a given year, month, and state. A disaster occurrence variable is created and set to 1 if there is more than one disaster in that month and year, and 0 otherwise. Then the border crossings and disaster flags are combined by state, year, and month. In the first row in *Table 3.7*, there is No Disaster (0) in Alaska in January 1996 since no disaster has occurred that month (Disaster Count = 0). But, in the state of Alaska in February 1996, there were 12 disasters that took place which is why the Disaster Occurred is 1. A bar plot (*Figure 4.5*) is generated for the total crossings, where one bar represents no disaster months (0) and the other represents Disaster months (1).

State	Year	Month	Total crossings	Disaster Count	Disaster occurred
Alaska	1996	1	7705	0	0
Alaska	1996	2	7585	12	1

Table 3.7: *Example of data after adding Disaster Occurrence*

An independent U test is performed to test if the average number of total crossings is different when there is a disaster compared to when there is no disaster. This test compares the means of two different groups and sees if they are significantly different. The first group is the total crossings when a disaster occurred, and the second group is the total crossings when no disaster occurred. A Mann Whitney U test which tests two independent groups to determine if one tends to have higher values than the other is performed. The p-value is calculated to measure how big the difference between groups is and whether the difference is statistically significant. The p-value is less than 0.05, which means that there's a statistically significant difference between the two groups.

A pre-disaster and post-disaster analysis is performed to check the impact of a disaster type on border crossings. The disasters that were checked are Fire, Severe Storm, Flood, Biological, Earthquake, and Tsunami. The Year and Month column is converted into a full date column and set to the first day of each month. Dates are then found for the months in which the specific disaster type has occurred. It identifies dates when the disaster occurred and labels the surrounding months as either “Pre-Disaster” or “Post-Disaster.” It extracts border crossing volumes for these periods and conducts an independent t-test to statistically compare average crossings before and after the disaster. Bar charts are plotted in *Figure 4.6a-f*, which show the change in the average crossing between Pre-Disaster and Post-Disaster.

Section 3.7: Predicting Border Values

Since the Value in the border data is a numerical count of the border crossings by month, it is considered count data. Thus, count data models such as Generalized Linear Model (GLM) with a Poisson distribution or Negative Binomial distribution would be optimal to use instead of regression models such as linear or logistic regression. A GLM acts as an extension of linear regression, and it works on discrete data such as the border data with crossings aggregated by month. U.S.-Canada and U.S.-Mexico borders are different politically, geographically, and have different seasonalities, they are modeled separately. First, data is filtered by the borders U.S.-Mexico and U.S.-Canada. Then, both datasets are randomly split into a training set, which has 80% of the data, and a testing set with 20% of the data. A Generalized Linear Model (GLM) with a Poisson distribution is fit for both the borders, with Value (border crossings) being the target variable and measure, month, and disaster count as predictors. Disaster count is chosen as a predictor as it shows either the increase or decrease in the crossings based on the number of disasters that take place. Then, the dispersion statistic is calculated, which shows the variability of the model.

Dispersion statistic *(Equation 3.3)* is the Pearson chi-square *(χ^2)* from the model divided by the residual degrees of freedom, which is obtained from the model summary. The Pearson chi-square statistic *(χ^2)* is how the observed data *(O_i)* fits the expected data *(E_i)* as shown in *Equation 3.1*. Residual Degrees of Freedom (df residuals) as shown in *Equation 3.2* is the number of data points *(η)* minus the number of estimated parameters *(k),* which includes the intercept. If the dispersion statistic is around 1, there is no overdispersion, and a Poisson model is appropriate. If the dispersion is greater than 1, it indicates a strong overdispersion, which suggests that a Negative Binomial model should be fit instead. A Negative Binomial Model is also fit with the same predictors as the Poisson model, where the value is being predicted for both borders using measure, month, and disaster count to predict. The U.S.-Canada data is split into training and testing sets. A model summary showing the coefficients for each predictor, *p*-value, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) for the test set is calculated to look at the model's performance in *Table 4.6*.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \qquad \text{(Equation 3.1)}$$

$$df\ residuals = \eta - \kappa \qquad \text{(Equation 3.2)}$$

$$Dispersion\ Statistic\ = \frac{\chi^2}{df\ residuals} \qquad (Equation\ 3.3)$$

Chapter 4

Analysis and Discussion

This chapter focuses on the results and discusses how the models and graphs that we created are relevant in our research and how they could help answer a real-world problem. It analyzes border crossings data with clustering, time series, forecasting, and disaster impact assessments. K-Means and K-Prototypes clustering models are used to explore groups and trends in the border data by geographic location and value of the border crossings. Next, forecasting models such as SARIMA and ARIMA models are analyzed to see future trends from 2024 through 2028, looking at seasonal and non-seasonal patterns, which are different for each border. Natural disasters and border crossings are explored together with each other by doing t-tests and ANOVA, showing patterns in how disaster events impact border traffic. These analyses exhibit insights into geographic and external factors that influence border crossing traffic.

Section 4.1: Clustering

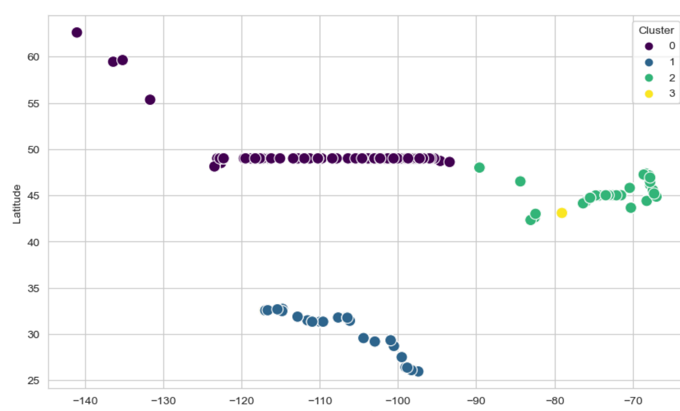

Figure 4.1: *K-means Port Clusters Based on Traffic Volume and Location*

The scatter plot in *Figure 4.1* shows the results of K-means clustering that is applied to border data, where ports are grouped based on the latitude, longitude, and traffic volume. Four clusters are used after looking at the elbow method which is a technique used to find the ideal number of clusters. The ports in cluster 0 (dark purple) are located mostly in the U.S.-Canada border and are in northern U.S. states, and a few in Alaska. Cluster 1 (blue) ports are in the southern U.S states, like Texas, Arizona, and California, and close to the U.S.-Mexico border. Cluster 2 (green) ports are in eastern U.S.-Canada border crossings like New York, Vermont, and Michigan. There is only one port in cluster 3 (yellow), which is most likely an outlier, but it could also be a unique port with a big difference in the volumes or a difference in location. The unique port in cluster 3 is Buffalo Niagara Falls, this could be because as previously mentioned in *Section 3.2,* this is a port that is used as a shortcut for U.S.-to-U.S. travel. This means that there could be more crossings that take place in this port compared to the ports that are around it.

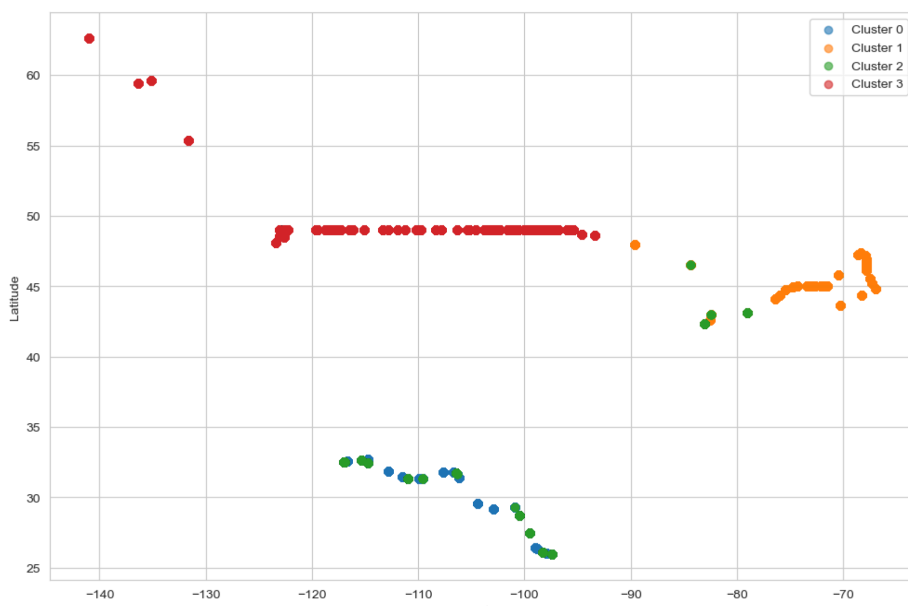

Figure 4.2a: *K-Prototypes Border Crossing Clusters by Measure*

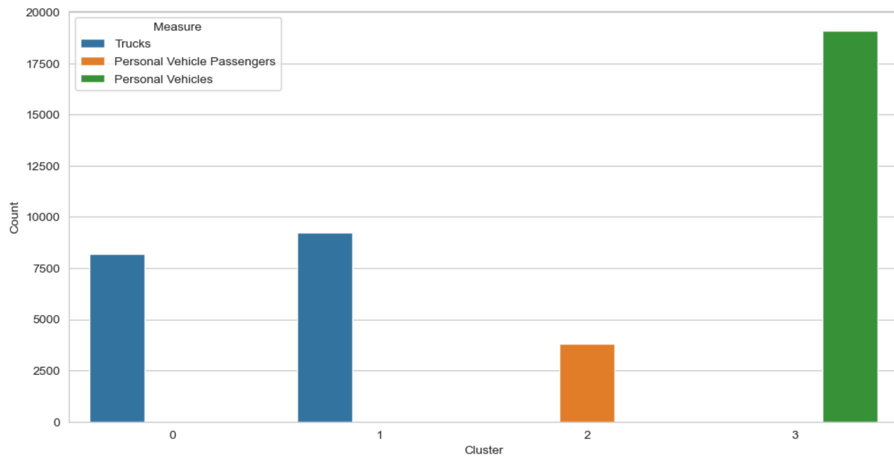

Figure 4.2b: *Highest Measure Type for Each Cluster*

Figure 4.2a shows the scatterplot of the border crossing ports across the U.S. grouped into 4 clusters based on their location. Again, the elbow method was used to find the best number of clusters which is 4. In this map, each point corresponds to a port that gets its color based on its cluster identifier. High truck crossings at the southeastern U.S.-Mexico border and eastern U.S.-Canada border are highlighted in *Figure 4.2b* since these are the main centers of the cluster 0 (blue) and cluster 1 (orange). The green-colored Cluster 2 stretches across the United States from border points across the American and Canadian border. Personal Vehicle Passengers control these ports strongly, which indicates these ports serve primarily urban areas where individuals travel independently instead of transporting cargo. The red cluster identifies personal vehicles as the dominant modality for border travel between the United States and Canada based on its location in the U.S.-Canadian border. Cluster 3 presents the maximum crossing volume, indicating there is personal car traffic between the U.S. and Canada borders. This indicates that the North West ports are mostly used for commuting and tourism compared to the South ports, which are truck-heavy, indicating more trade occurring. This model helps address the geographical and functional differences between the border regions.

Section 4.2: Forecasting Results

The Kruskal-Wallis test for seasonality to prove that the border data has seasonal patterns in *Table 4.1* shows that the p-value for both the overall border crossing time series is 0.0004, which is less than the significance level of 0.05. This rejects the null hypothesis that the group medians are equal between periods, which indicates that there is significant seasonality in border crossing values. *Table 4.1* shows that the p-value for the U.S.-Mexico border is 0.7617, which is higher than the significance level of 0.05. The data indicate that no consistent seasonal patterns exist at the U.S.-Mexico border, and the examined monthly variations are statistically insignificant. The recorded p-value for U.S.-Canada amounts to 1.46×10^{-63} demonstrating strong seasonality since it exceeds both 0.05 and 0. New data provides an even stronger indication of seasonality. Statistics show that seasonality exists and monthly differences hold statistical significance.

Border Name	P-Value
U.S.-Mexico and U.S.-Canada	0.0004
U.S.-Mexico	0.7617
U.S.-Canada	1.46×10^{-63}

Table 4.1: *Kruskal-Wallis Test Results for Border Crossing Seasonality*

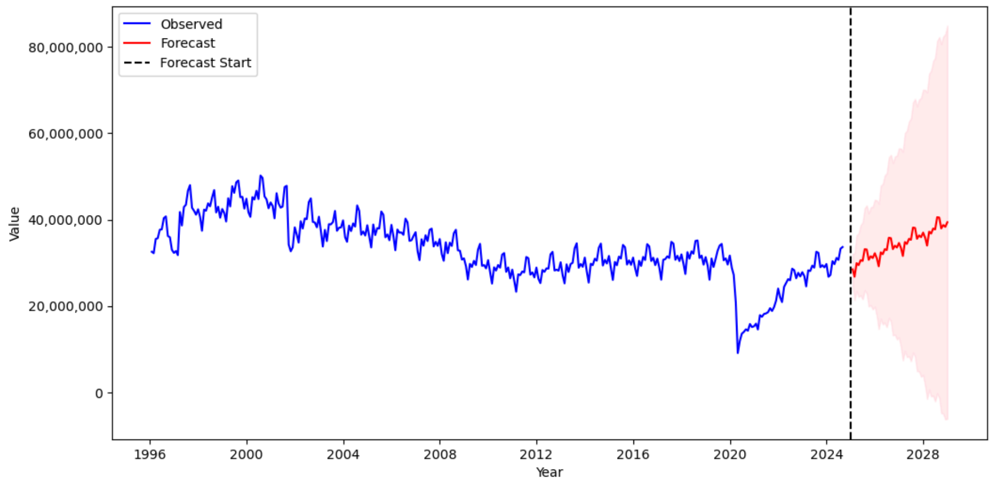

***Figure 4.3**: SARIMA Forecast of the border crossings (2025-2028)*

Since it has been proven that seasonality exists, we created a SARIMA model in *Figure 4.3*, which shows a forecast of border crossings for the years 2024-2028. The parameters that were used for this model are (1,1,2) for order and (1,1,1,12) for seasonal order. This was chosen since the AIC for this was 9939.78 and the BIC was 9962.32, which were the lowest of all combinations. The blue line in the graph shows the observed trend in the crossings, which is basically what was looked at in *Figure 3.9*. This red line in the graph represents the forecast which shows that there will be a slight increase from now and four years where the crossings might go as high as 39,000,000. The shaded red area is the 95% confidence interval, which reflects the increasing uncertainty as the model looks further ahead. The prediction interval suggests that the future values could range higher or lower as well, depending on some external factors.

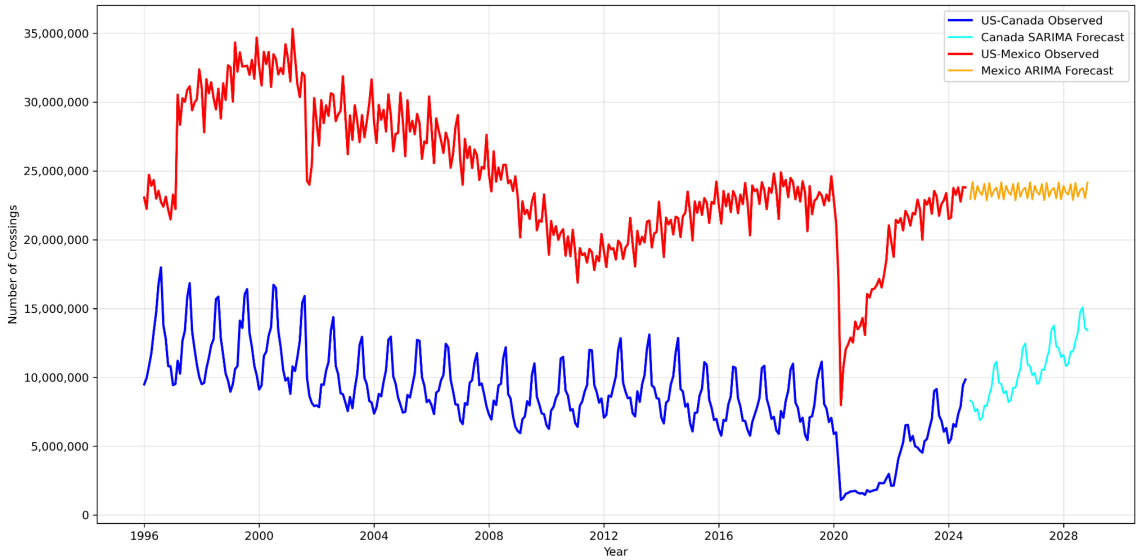

Figure 4.4: *Border Crossings: Canada (SARIMA) vs Mexico(ARIMA)*

As a way to explore the border crossings separately by border, a forecast model is created where each border is fit separately. To determine the best model to use, a Kruskal-Wallis test is performed to test seasonality of both borders. Since there is seasonality in the U.S.-Canada border, a SARIMA model is fit for that border. The best parameters after the Grid Search for the SARIMA model are (1, 1, 2) for order and (0, 1, 1, 12) for seasonal order, since it had an AIC of 9332.60 and BIC of 9351.38, which was the lowest of all the combinations. In *Figure 4.4*, the dark blue line represents the observed border crossings, and the cyan (light blue) line represents the U.S.-Canada border crossings forecast after 2024 to 2028. The dark blue line shows seasonality, where peaks and drops happen every year. There also seems to be a big drop in crossings in 2020, likely due to the impact of COVID-19. The SARIMA forecast (cyan) shows that the seasonality will continue to rise and dip in the coming years, with summer crossings being high and winter being low. It also shows that the average level of crossings will be rising gradually year after year.

On the contrary, the U.S.-Mexico border does not have a seasonal trend, so an ARIMA model is used. The parameters that were used for the model were the order (2,1,3) after performing grid search since they had the lowest AIC of 10736.11. In *Figure 4.4*, the red line represents the observed border crossings for the U.S.-Mexico border, and the orange line represents the forecast of the border crossings. The red line shows no clear seasonality since it is mostly flat with big level shifts. The orange line (forecast) also shows no seasonality or strong trends where there is a clear increase or decrease in the crossings in the years to come. The model does not show any extreme highs or lows and is steady.

The border crossings in the dataset take into account all the border crossings that happen coming into the U.S. This includes trade, day trips, short-term trips, immigrants who want to stay in the U.S, and migrants who seek asylum. Although the forecasting models show that there is a slight increase in border crossings for the future years 2024-2028, it is important to note that external factors such as political tensions or economic crises could affect these. For instance, after President Donald Trump was elected in 2025, border security became stricter and which directly affects immigration. This could also discourage people from traveling to the U.S and could decrease the number of crossings significantly. Likewise, the 25% tariff that was imposed on Canadian and Mexican imports could also bring down the number of border crossings in the future. These exogenous events could change the forecasting model.

Section 4.3: Natural Disasters Occurrence Analysis Results

This section examines the differences between the border crossings with no disaster and disaster months. Looking at the total border crossings by disaster occurrences might falsely conclude that the disasters are related to lower crossings. However, that could just be because there are fewer disaster months compared to the months with disasters. But looking at the average crossings by state-month would normalize the data and help make a better comparison between the groups. Statistical tests like the Mann-Whittney U studies the distribution of the two groups and not the total which makes it unbiased.

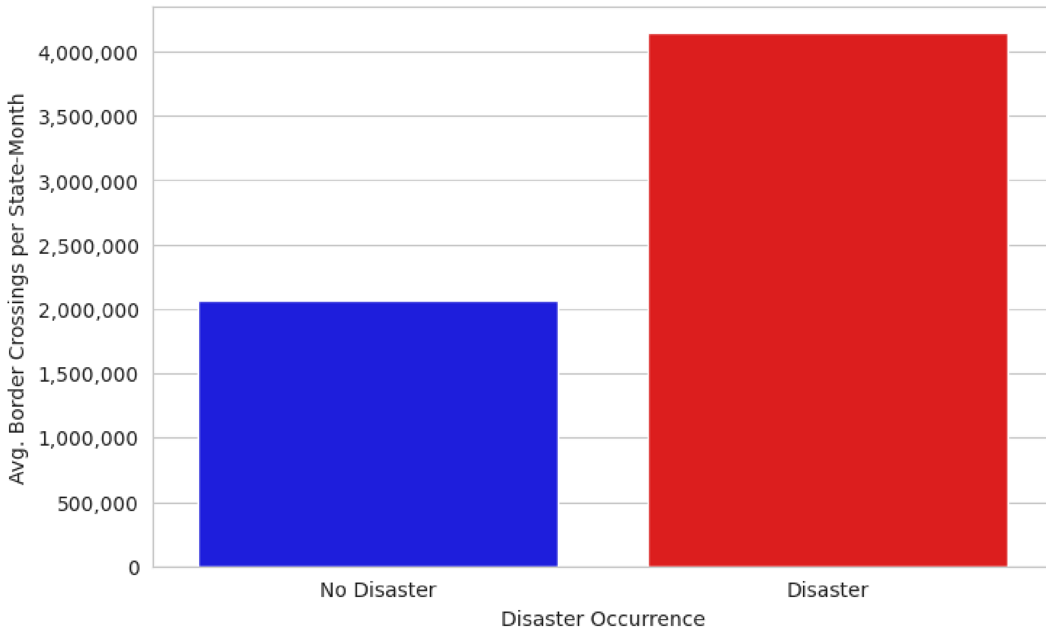

Figure 4.5*: Average Border Crossings per State-Month: Disaster vs. No Disaster Months*

Figure 4.5 shows that the months with disaster occurrences (1) have 4,000,000 average border crossings and months with no disaster occurrences (0) have 2,000,000 on average. On average, border crossings were higher in months where a disaster occurred than in months with no disaster. This may be attributed to people having to move supplies into affected areas using more trucks, and more deliveries into those states might be occurring. Unexpectedly, there were more average crossings during the disaster months than there were when there were not. A potential explanation could be that there is increased travel for emergency response and aid delivery. Although, border crossings are higher during disaster months doesn't mean that the disasters caused that.

The U test examining the average number of crossings between disaster and non-disaster conditions produces a p-value of 0.000. Results from the U test indicate that crossing frequencies differ substantially between disaster and non-disaster periods because the calculated p-value falls below 0.05. The average number of crossings becomes greater during disaster situations than during normal times. The test confirms that the average crossings during disaster months and no-disaster months are statistically different and not due to chance. A further investigation into these findings is necessary to better understand the origins of this phenomenon.

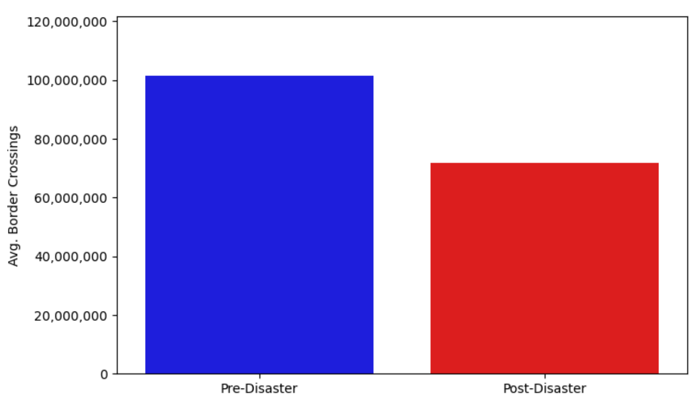

***Figure 4.6a**: Fire Impact*

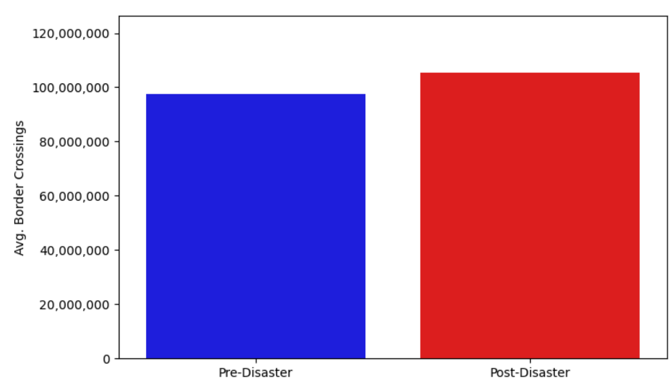

***Figure 4.6b:** Severe Storm Impact*

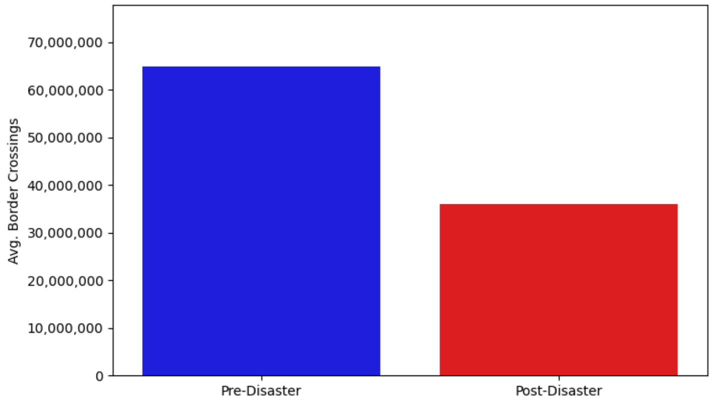

Figure 4.6c: *Flood Impact*

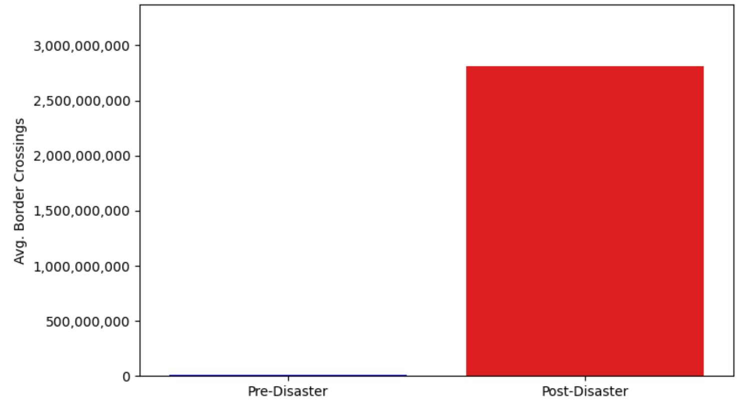

Figure 4.6d: *Biological Impact*

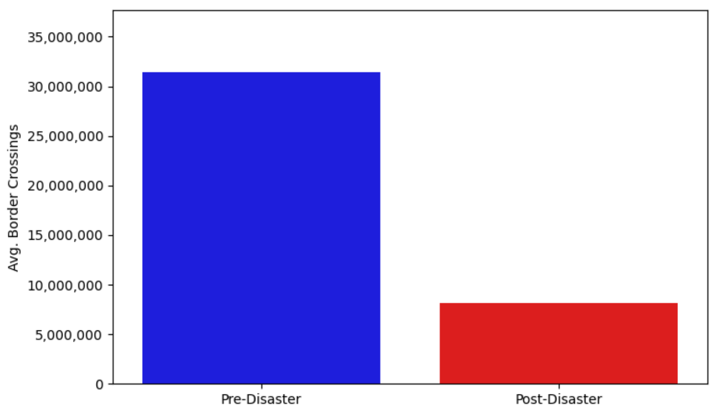

Figure 4.6e*: Earthquake Impact*

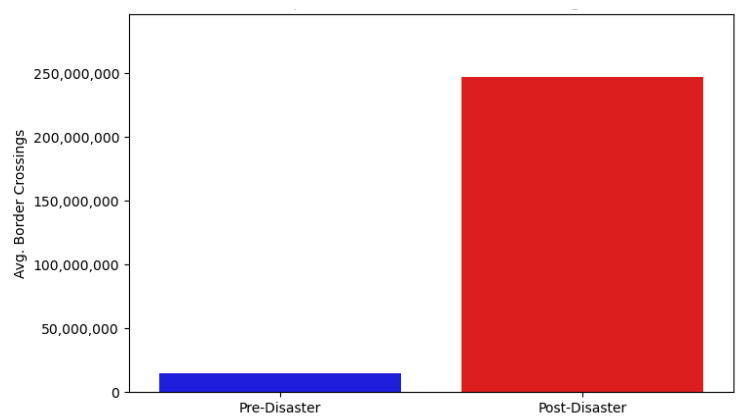

Figure 4.6f: *Tsunami Impact*

Disaster Type	Average Pre-Disaster Crossings	Average Post-Disaster Crossings	P-Value
Fire	101,328,401.95	71,807,890.52	0.4576
Severe Storm	97,420,208.73	105,256,862.73	0.8598
Flood	64,828,061.42	36,024,821.42	0.3185
Biological	9,651,296.00	2,807,900,912.00	0.2710
Earthquake	31,381,276.65	8,118,520.19	0.1546
Tsunami	15,108,559.67	246,660,166.00	0.5160

Table 4.2: *T-test Results on the Impact of Disaster Types on Crossings*

Although the bar graphs in *Figure 4.6a-4.6f* show that there is a high difference between the variables, the p-values from independent t-test in *Table 4.2* suggest that none of the disaster types has a statistical difference. Biological disasters also have a huge jump from 9 million to 2.8 billion crossings and a p-value of 0.2710. This is likely because of COVID-19 and the fact that the crossings went down quickly in just a few months. It is interesting to find out that the disasters are not statistically significant, this may be attributed to high variability, which could lead to noise in the data. Also, there is a small sample size since there are very few months of pre-disaster and post-disaster data, and that makes it hard for the t-test to be confident that the difference is real.

Section 4.4: Count Models for Border Crossings

Count models such as Generalized Linear Models with Poisson and negative binomial distribution are fit to predict the border crossing values for both the U.S.-Mexico and U.S.-Canada borders. The predictors that will be used are measures, month, and disaster counts. This is considered a multivariate model since we use two or more predictors to predict the outcome. A comparison between the model performances for both the distributions is also examined.

Border	Degrees of freedom Residual	Pearson Chi-Square	Test RMSE	Test MAE
U.S.-Mexico	74437	1.10×10^{10}	236784.17	85220.19
U.S.-Canada	241425	2.13×10^{10}	59938.19	13660.15

Table 4.3: *GLM with Poisson Distribution Performance*

A Generalized Linear Model (GLM) with Poisson distribution for the U.S.-Mexico is fit on both training and testing data, and the model performance is in *Table 4.3*. From the table, the degrees of freedom residuals and Pearson chi-square are taken and used for calculating the dispersion statistic, which shows the variability of the model. The Dispersion Statistic from *Equation 4.1* suggests overdispersion since it is way greater than 1. The test Mean Absolute Error (MAE) is 85220, which means that the model is wrong by 85,220 crossings for every prediction. The Root Mean Squared Error (RMSE) for the test is 236784.17 crossings, which means that the model is wrong by as much as 236,784 crossings when errors are large. These are really large errors, which make the GLM model with poisson distribution not reliable for predicting border crossings for the U.S.-Mexico border.

$$Dispersion\ Statistic\ for\ U.S. -\ Mexico\ = \frac{1.10\times10^{10}}{74437}\ = 147775.97 \qquad \textit{(Equation 4.1)}$$

Next, the model fits U.S.-Canada data on training and testing data as well, and the GLM model performance is given in *Table 4.3*. The dispersion statistic is again calculated with the Pearson Chi-Square and the degrees of freedom residuals. *Equation 4.2* shows that the dispersion statistic is about 88226, which is again greater than 1; this suggests overdispersion for this model. The test MAE suggests that the average difference between the predicted and actual is 13,660 crossings. The test RMSE indicates that the model predictions are around 59,938.19 crossings off when there are large errors. Compared to the U.S.-Mexico border model, this performs better in terms of fewer errors, but that could also be because the U.S.-Mexico crossings are higher than the U.S.-Canada crossings. This model is also not appropriate to use for prediction since it has high errors.

$$Dispersion\ Statistic\ for\ U.S. -\ Canada = \frac{2.13\times10^{10}}{241425}\ = 88226.16 \qquad \textit{(Equation 4.2)}$$

Generalized Linear models with a negative binomial distribution are better to use for predicting the border crossings because of the overdispersion from the Poisson distribution models. First, a GLM model with a negative binomial distribution is fit for the U.S.-Mexico border with training and testing data. The dispersion statistic is calculated in *Equation 4.3* and is 4.77, which is still greater than 1, but it is not as high as the Poisson model. The test RMSE measures the average error between predicted and actual values, and this model has errors of 239,697 crossings. The test MAE for Mexico is 85,064, which suggests that on average, the model's predictions differ from the actual number of crossings by 85,000. The U.S.-Canada border has a dispersion statistic of 13.58 from *Equation 4.4.* This is again greater than 1 but better than the Poisson distribution. The test RMSE is 60,143 crossings, which is the average difference between predicted values and the actual values. The test MAE is 13,698, which is the number of crossings that the model's predictions differ from the actual values. The MAE and RMSE for the models are high, which indicates that the model still needs to be improved but also means that it has some predictive power and could detect trends.

Border	Degrees of freedom Residual	Pearson Chi-Square	Test RMSE	Test MAE
U.S.-Mexico	74428	5.16×10^5	239697.71	85064.69
U.S.-Canada	241416	3.28×10^6	60143	13698.53

Table 4.4: *Generalized Linear Model with Negative Binomial Distribution*

$$Dispersion\ Statistic\ for\ U.S - Mexico = \frac{5.16\times10^5}{74428} = 6.93 \qquad \textit{(Equation 4.3)}$$

$$Dispersion\ Statistic\ for\ U.S. - Canada = \frac{3.28\times10^6}{241416} = 13.58 \qquad \textit{(Equation 4.4)}$$

The model summary shows the coefficients for the predictors, which could tell us how much each predictor impacts the border crossings. The coefficients are given on a logarithmic scale since a log link was used for fitting the model. The coefficients are turned into exponential form to be able to see the impacts. *Table 4.4* shows the exponential coefficients for each border, which helps identify the impacts of each predictor.

Predictors	U.S.-Mexico	U.S.-Canada
Intercept	10416.914992	2695.507147
Measure (Buses)	0.079270	0.042319
Measure (Pedestrians)	13.856199	0.170173
Measure (Personal Vehicle Passengers)	51.661544	17.813899
Measure (Personal Vehicles)	24.912511	8.635397
Measure (Rail Containers Empty)	0.170705	0.224955
Measure (Rail Containers Loaded)	0.156196	0.543849
Measure (Train Passengers)	0.005638	0.086145
Measure (Trains)	0.003715	0.010961
Measure (Truck Containers Empty)	0.626554	0.354335
Measure (Truck Containers Loaded)	1.237883	1.532887
Measure (Trucks)	1.790255	1.879825
Month (February)	0.871634	0.948565
Month (March)	0.979282	1.111123
Month (April)	0.937948	1.117035
Month (May)	0.969380	1.285189
Month (June)	0.974924	1.381606
Month (July)	0.965032	1.527377
Month (August)	1.008855	1.630094
Month (September)	0.874226	1.321656
Month (October)	1.001788	1.252358
Month (November)	0.919847	1.066572
Month (December)	0.959948	1.035687

disaster_count	1.001264	1.002008

Table 4.5: *Coefficients for Negative Binomial distribution model*

For the U.S.-Mexico border, the intercept of 10,416.91 is the baseline for the number of crossings when predictors are January and bus passengers. From the modes of transportation, personal vehicles and personal vehicle passengers had coefficients of 51.66 and 24.91, which suggests that they are strong predictors of border crossings and are associated with higher volumes of crossings. Pedestrians had a coefficient of 13.856 which is also a strong positive predictor. Trucks, truck containers loaded, and truck containers empty were also positive predictors, but were less significant in predicting the crossings compared to personal vehicles. Trains and Train Passengers had very little impact, suggesting that they don't contribute much to predicting the crossings at the Mexico border. All month, predictors were almost equal to 1, and that means that they do not play a great role in the prediction. The effect on border crossings is smaller for February and September (both with an effect of 0.87), and a higher influence is witnessed for August and October. However, the effect of the season on predicting crossings is not significant. The disaster count coefficient 1.0013 implies that every federally declared disaster results in a small increment of border crossings by 0.13%. This influence is small, but it strives in a positive direction all the time.

For the U.S.-Canada border model, the intercept is 2,695.51, which is lower and reflects fewer baseline crossings compared to Mexico. Similar to Mexico, personal vehicle passengers and personal vehicles are the strong positive predictors. Trucks and truck containers loaded have a slightly stronger influence on border crossings compared to Mexico. However, pedestrians have a smaller impact since it is harder to walk into the U.S from Canada than in Mexico.

Loaded rail containers (0.54) are more important than in Mexico, suggesting that rail crossings are more important for the movements of goods between the U.S. and Canada. In Canada, month coefficients indicate seasonality with an incremental increase from 1.11 in March to 1.63 in August. This shows that crossings are higher in the summer months, possibly because of tourism and seasonal trade. This was already discussed in *Section 4.2* when looking at the SARIMA model for the Canada border. The disaster count coefficient (1.0020) implies that there is a 0.2% increase in crossings for every additional disaster. This is slightly greater than the Mexico value and also shows a subtle impact, and not a significant one.

Chapter 5

Conclusions

This research studies the border crossings data from both the U.S.-Mexico and U.S.-Canada borders through different techniques like EDA, clustering, forecasting, and predicting border crossings. The research uncovers various insights into the factors that influence border crossing patterns and trends that show implications for future research and policy making.

The clustering analysis showed clear geographical patterns where there is a differentiation between the U.S.-Mexico and U.S.-Canada border. The K-Means and K-Prototypes models disclose the differences, such as high volume of truck crossings in the ports in the South of the U.S and the high traffic flow in the North of the U.S ports because of Personal Vehicles and commuter traffic. This is important to note since it could be helpful for future planning and allocating specific resources through specific ports based on the traffic type and location.

For the forecasting models, the U.S.-Canada border had a seasonal and more stable trend, whereas the U.S.-Mexico border had non-seasonal trends. The forecasting could help with border agencies being prepared for the rushes and drops at the border in the future by being prepared for more staffing or resource allocation during the peaks and drops as well. However, it is important to understand that political and economic factors could change the forecast of these crossings. Border crossings can be unpredictable in the future since there are many uncertainties about the events that could occur which can change the predictions from the forecast. Having more data on the previous political affairs, policy changes, or economic factors could help predict how future crossings would be affected by different types of factors. The analysis of natural disasters suggests that the disaster periods correlate with more border crossings, since there is an increased movement of emergency resources coming in through the ports and having higher aid deliveries depending on how major a disaster is. Border preparedness for emergencies demonstrates its critical role because infrastructure requires readiness to manage traffic surges during disasters. This could help policymakers improve management in the disaster sector and have strategies to respond at the border during and after the disaster occurrence.

The implementation of a Negative Binomial model for both U.S.-Mexico and U.S.-Canada borders became necessary after detecting overdispersion along with suspected model overfitting in the Poisson model, leading to improved predictive power. For both the U.S.-Mexico and the U.S.–Canada border crossings, important factors of border crossings relate to movements involving vehicles, mainly Personal Vehicle Passengers and Trucks. This pattern demonstrates constant and positive correlation with volumes of traffic, which emphasises the important role of vehicles in cross-border movements. However, the nature of border crossings is very different from the U.S.-Mexico and U.S.-Canada borders: U.S.-Mexico border is significantly defined in terms of pedestrian crossings because urban border areas tend to draw huge volumes of pedestrians as seen in the negative binomial model. In contrast, the dominance of freight rail in northern cross-border trade is highlighted at the U.S.-Canada border by the substantial effect of Rail Containers Loaded. Since disaster count serves to have very little impact in both models, it indicates that disaster occurrences have little positive impact on border crossings. This reveals the necessity for unique policies and methods in funding infrastructure and disaster management at each border.

Section 5.1: Future work

More attention will be given in upcoming years to enhance disaster classification methods by adding biological threats such as pandemic diseases. Knowledge about disaster-specific effects on border conditions could enable better planning of disaster preparation strategies by type. Additionally, the research needs to include additional important variables, which should include daily weather reports. Both weather conditions affecting border crossing person numbers and port entry lockdowns occur because of adverse meteorological conditions, including extreme heat, snowstorms, and hurricanes.

The research should include data relating to policy changes as an element that influences border crossing activities. The measures implemented related to immigration policy and trade activities, and border security procedures have the potential to significantly impact border crossing activities. Border policies that raise tariffs or heighten inspections at ports provide either positive or negative incentives to people, along with goods to cross borders. The development of forecasting models to predict border volume changes from different immigration policies and trade adjustments should become possible. Future policy modifications could find clarification about their effects on border crossings through this method.

The data that was used is count data which is not continuous because it is aggregated crossing by month. In the future, it would be beneficial for the data to be continuous, which means that the data shouldn't be aggregated like border data is, since we would be able to use more predictive models. Count models, such as generalized linear models with Poisson distribution and negative binomial distribution, were fit, but the model did not perform that well on the test set. In the future, it would be better if there were models that fit better when there is overdispersion in the model. Models such as Zero Inflated models, Conway-Maxwell-Poisson model, and Quasi-Poisson regression model, could be used to predict the border crossing volumes as well. These models were not fit for reasons beside time constraints. First, the nature of the data and the characteristics available did not particularly indicate a high proportion of excess zeros, which constitutes a major reason for fitting zero-inflated models. Second, the Conway-Maxwell-Poisson is computationally intensive and less available in standard base Python packages, and therefore harder to use and explain. Third, while the Quasi-Poisson model can handle overdispersion, it does not offer a full likelihood structure, thereby limiting use for some prediction and model comparison goals. Therefore, generalized linear models with Poisson and negative binomial distributions were prioritized due to their interpretability, and application for overdispersed count data. It would be helpful to also have classifications for each crossing into the U.S such as “trade”, “tourism”, “Immigrant”, and “Asylee”. It would be good to look at specific types of crossings and forecast with that information for more accuracy.

References

Belasen, A., & Polachek, S. W. (2011). *Natural Disasters and Migration*.

Berdell, John & Ghoshal, Animesh. (2015). US-Mexico border tourism and day trips: An aberration in globalization?. *Latin American Economic Review.* 24. 10.1007/s40503-015-0023-9.

Bhattacharjee, A. (2023, September 10). *Kruskal-Wallis Test: A Powerful Tool for Detecting Seasonality in Time Series Data using Python. Medium.* https://medium.com/@avijit.bhattacharjee1996/kruskal-wallis-test-a-powerful-tool-for-detecting-seasonality-in-time-series-data-using-python-d827ef23d29e

Bureau, U. C. (2025). *International Trade*. Www.census.gov. https://www.census.gov/foreign-trade/statistics/highlights/topcm.html#total

Cartographic Boundary Files - Shapefile. (2021). The United States Census Bureau. https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html

dremio. (n.d.). *Grid Search | Dremio.* Www.dremio.com. https://www.dremio.com/wiki/grid-search/

Facts About Texans and Their County Governments. (2020). Texas Association of Counties. https://www.county.org/resources/resource-library/news-and-information/10-facts-about-texas-counties

GeeksforGeeks. (2023, November 17). *SARIMA (Seasonal Autoregressive Integrated Moving Average). GeeksforGeeks.*

https://www.geeksforgeeks.org/sarima-seasonal-autoregressive-integrated-moving-average/

Heads or Tails. (2016). *US Natural Disaster Declarations.* Kaggle.com. https://www.kaggle.com/datasets/headsortails/us-natural-disaster-declarations

Kaufman, J. (2025, April 1). *Canada and the European Union are the two largest suppliers of U.S. agricultural imports, followed by Mexico* | Economic Research Service. Usda.gov. https://www.ers.usda.gov/data-products/chart-gallery/chart-detail?chartId=58394

Kruskal-Wallis test. Jdemetradocumentation.github.io. https://jdemetradocumentation.github.io/JDemetra-documentation/pages/theory/Tests_KW.html

Kitamura, K. H. (2025, March 17). U.S.-Canada Trade Relations. Congress.gov. https://www.congress.gov/crs-product/IF12595

Matheson, P. (2022, November 15). *The growing importance of cross-border securities holdings for the US economy.* Sifma. https://www.sifma.org/resources/news/blog/the-growing-importance-of-cross-border-securities-holdings-for-the-u-s-economy/

Maxwell, O., Mayowa, B., Chinedu, I., & Peace, A. (2018). *Modelling Count Data; A Generalized Linear Model Framework.* Scientific & Academic Publishing. http://article.sapub.org/10.5923.j.ajms.20180806.03.html#:%7E:text=In%20this%20section%20we%20shall,the%20Negative%20Binomial%20Regression%20Model

Moslimani, M., & Passel, J. S. (2024, September 27). *What the Data Says about Immigrants in the U.S.* Pew Research Center; Pew Research Center. https://www.pewresearch.org/short-reads/2024/09/27/key-findings-about-us-immigrants/

News, P. (2025, April 25). *Trump's bluster and border arrests cool Canadians' appetite for U.S. tourism*. PBS News. https://www.pbs.org/newshour/world/trumps-bluster-and-border-arrests-cool-canadians-appetite-for-u-s-tourism

Noble, J. (2024, May 24). *ARIMA models*. Ibm.com. https://www.ibm.com/think/topics/arima-model

Office of the United States Trade Representative. (2025). Countries & Regions | United States Trade Representative. Ustr.gov; Office of the United States Trade Representative. https://ustr.gov/countries-regions

Python Machine Learning - Grid Search. Www.w3schools.com. https://www.w3schools.com/python/python_ml_grid_search.asp

Reddy, Y. (2023, April 10). *K-means, kmodes, and k-prototype.* Medium. https://medium.com/@reddyyashu20/k-means-kmodes-and-k-prototype-76537d84a669

Ritchie, H., & Rosado, P. (2024). *Is the number of natural disasters increasing?* Our World in Data. https://doi.org/10.1080/17477891.2023.2239807)

Shoichet, C. E. (2023, September 21). *The number of migrants crossing the US-Mexico border is likely to keep growing. Here are 3 reasons why*. CNN. https://www.cnn.com/2023/09/21/us/us-mexico-border-migrant-crossings-explainer-cec/index.html

Singh, G. (2025, March 26). *A simple introduction to ANOVA.* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2018/01/anova-analysis-of-variance/

Strömberg, D. (2007). Natural Disasters, Economic Development, and Humanitarian Aid. *The Journal of Economic Perspectives*, *21*(3), 199–222. http://www.jstor.org/stable/30033741

U.S. Department of Homeland Security, Customs and Border Protection. (2017, March 13). *Border Crossing Entry Data | Tyler Data & Insights.* Data.bts.gov. https://data.bts.gov/Research-and-Statistics/Border-Crossing-Entry-Data/keg4-3bc2/about_data

U.S. Remains the World's Most Powerful Travel & Tourism Market. (2024, September 4). Wttc.org; World Travel & Tourism Council. https://wttc.org/news/us-remains-the-worlds-most-powerful-travel-and-tourism-market

USA Facts. (2024, September 6). *Is the number of major natural disasters increasing?* USAFacts. https://usafacts.org/articles/are-the-number-of-major-natural-disasters-increasing/

Vernick, D. (2025, January 14). *Is Climate Change Increasing the Risk of Disasters?* World Wildlife Fund. https://www.worldwildlife.org/stories/is-climate-change-increasing-the-risk-of-disasters

Vaughn, V. (2025, March 3). *U.S. TARIFFS ON CANADA AND MEXICO TAKE EFFECT: WHAT IT MEANS FOR CROSS-BORDER TRADE AND DE MINIMIS SHIPMENTS*. Green Worldwide Shipping. https://www.greenworldwide.com/u-s-tariffs-on-canada-and-mexico-take-effect-what-it-means-for-cross-border-trade-and-de-minimis-shipments/

Zahniser, S. (2025, April 1). USMCA, *Canada, & Mexico - Mexico: Policy | Economic Research Service*. Usda.gov. https://www.ers.usda.gov/topics/international-markets-us-trade/countries-regions/usmca-canada-mexico/mexico-policy