STAT-515 FINAL PROJECT

ANALYSIS ON BORDER CROSSING ENTRY DATA

INTRODUCTION

Border crossings are crucial points where people, goods, and ideas cross national boundaries, influencing economies, cultures, and society. The examination of border crossing entrance data gives vital insights into a wide range of phenomena, including immigration patterns, commerce flows, security trends, and more. In this project, I will investigate the US border crossing entry data in order to better understand the patterns and ramifications of cross-border movements.

The importance of analyzing border crossing entrance data cannot be emphasized. These statistics not only provide insight into the volume and characteristics of cross-border migrations, but also serve as a barometer for evaluating the effectiveness of immigration laws, trade agreements, and security measures. By examining trends, patterns, and anomalies in border crossing entry data, I can uncover hidden narratives, identify emerging challenges, and foster evidence-based decision-making.

At the heart of our inquiry lies a deep-seated recognition of the pivotal role that US borders play in shaping the nation's socio-economic landscape, security posture, and immigration dynamics. Understanding the nuances of border crossings at US borders holds profound implications for a myriad of stakeholders, including policymakers, law enforcement agencies, businesses, and communities along the border regions.

RESEARCH QUESTIONS TO BE ANSWERED IN THE PROJECT

- What causes the unequal distribution or several peaks in the number of crossings at U.S. border ports? How do the patterns of different border crossing activities differ?
- How does the geographical placement of ports of entry along the United States' borders correspond with the number of border crossings, and what geographical and socioeconomic factors influence these differences?
- Can the volume and kind of crossings be used to distinguish different clusters based on the underlying patterns in border crossing activity?

DATASET

In this project, I decided to work with the ‘Border_Crossing_Entry_Data’ dataset, obtained from data.gov. I have chosen this dataset since it will effectively address the key research questions relating to border crossings for my project.. This dataset consists of 390,306 rows of data and is of 43.1 MB. The dataset has information regarding the border crossings at various ports of entry along with the US-Canada and US-Mexico borders.

	A	B	C	D	E	F	G	H	I	J	K
1	Port Name	State	Port Code	Border	Date	Measure	Value	Latitude	Longitude	Point	
2	Roma	Texas	2310	US-Mexico Border	Dec-23	Buses	46	26.404	-99.019	POINT (-99.018981 26.403928)	
3	Del Rio	Texas	2302	US-Mexico Border	Dec-23	Trucks	6552	29.327	-100.928	POINT (-100.927612 29.326784)	
4	Willow Creek	Montana	3325	US-Canada Border	Jan-24	Pedestrians	2	49.000	-109.731	POINT (-109.731333 48.999972)	
5	Whitlash	Montana	3321	US-Canada Border	Jan-24	Personal Vehicles	29	48.997	-111.258	POINT (-111.257916 48.99725)	
6	Ysleta	Texas	2401	US-Mexico Border	Jan-24	Personal Vehicle Passengers	521714	31.673	-106.335	POINT (-106.335449846028 31.6731261376859)	
7	Warroad	Minnesota	3423	US-Canada Border	Jan-24	Trucks	837	48.999	-95.377	POINT (-95.376555 48.999)	
8	Wildhorse	Montana	3323	US-Canada Border	Jan-24	Trucks	20	48.999	-110.215	POINT (-110.215083 48.999361)	
9	Wildhorse	Montana	3323	US-Canada Border	Jan-24	Personal Vehicle Passengers	965	48.999	-110.215	POINT (-110.215083 48.999361)	
10	Westhope	North Dakota	3419	US-Canada Border	Jan-24	Truck Containers Loaded	102	49.000	-101.017	POINT (-101.017277 48.999611)	
11	Warroad	Minnesota	3423	US-Canada Border	Jan-24	Truck Containers Loaded	459	48.999	-95.377	POINT (-95.376555 48.999)	
12	Ysleta	Texas	2401	US-Mexico Border	Jan-24	Truck Containers Empty	21355	31.673	-106.335	POINT (-106.335449846028 31.6731261376859)	
13	Ysleta	Texas	2401	US-Mexico Border	Jan-24	Pedestrians	127217	31.673	-106.335	POINT (-106.335449846028 31.6731261376859)	
14	Wildhorse	Montana	3323	US-Canada Border	Jan-24	Personal Vehicles	519	48.999	-110.215	POINT (-110.215083 48.999361)	
15	Wildhorse	Montana	3323	US-Canada Border	Jan-24	Truck Containers Loaded	21	48.999	-110.215	POINT (-110.215083 48.999361)	
16	Westhope	North Dakota	3419	US-Canada Border	Jan-24	Personal Vehicles	339	49.000	-101.017	POINT (-101.017277 48.999611)	
17	Warroad	Minnesota	3423	US-Canada Border	Jan-24	Truck Containers Empty	590	48.999	-95.377	POINT (-95.376555 48.999)	
18	Ysleta	Texas	2401	US-Mexico Border	Jan-24	Truck Containers Loaded	63367	31.673	-106.335	POINT (-106.335449846028 31.6731261376859)	
19	Wildhorse	Montana	3323	US-Canada Border	Jan-24	Truck Containers Empty	10	48.999	-110.215	POINT (-110.215083 48.999361)	
20	Sweetgrass	Montana	3310	US-Canada Border	Jan-24	Truck Containers Empty	1985	48.998	-111.960	POINT (-111.959611 48.998388)	
21	Brownsville	Texas	2301	US-Mexico Border	Jan-24	Rail Containers Empty	7712	25.952	-97.401	POINT (-97.40067 25.95155)	

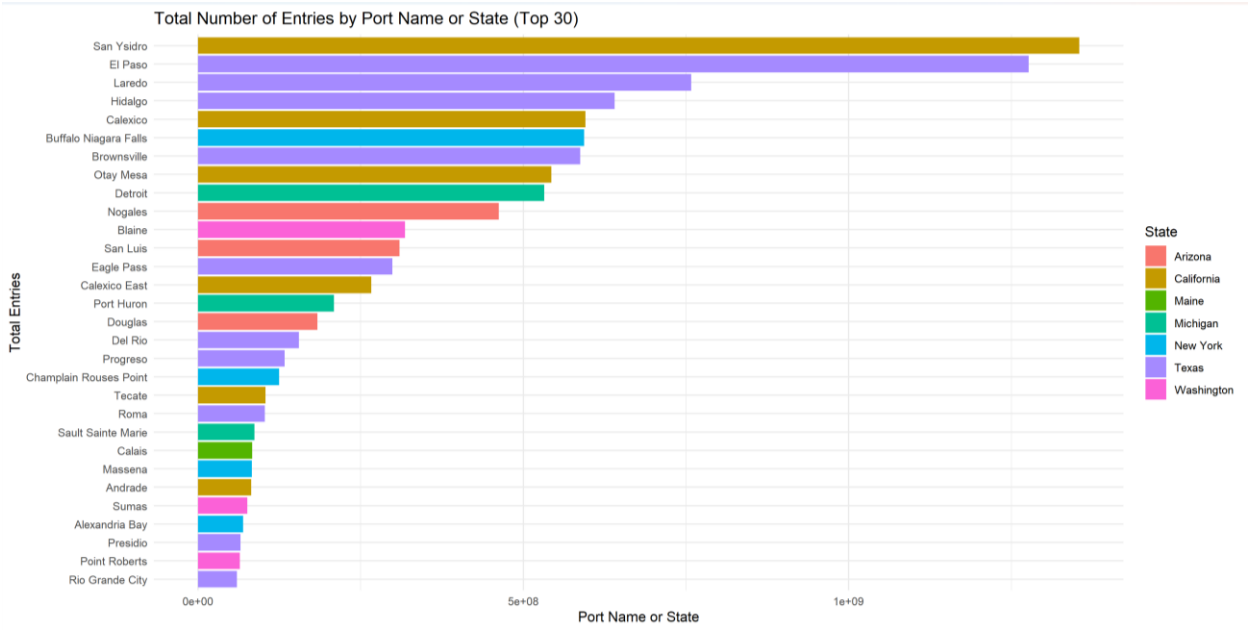
With a total of ten columns, the dataset contains important data that is necessary for our analysis. The details regarding the port are displayed on the top left having the names of each port, the state it is in, the respective port codes for each port, and the border it is located at. We then have the dates at which each mode of transport is used and how often it is crossing the border. Lastly, the latitude and longitude column, and the point column telling us the exact location of a port with the help of longitude and latitude coordinates. All the NOIR data types are present in the dataset, ensuring a diverse range of data visualizations/representations. This dataset contains 10 various variables which has information related to border crossing such as Port Name, State, Port Code, Border, Date, Measure, Value, Latitude, Longitude, and Point.

DATA PRE-PROCESSING

Pre-processing was a necessary step in our border crossing entry dataset preparation to ensure correctness and consistency. This includes eliminating redundant information, taking care of missing numbers, and fixing data inconsistencies. Latitude and longitude values are examples of geospatial data that is normalized for uniformity; normalizing procedures allow for fair comparisons between various border crossing places.

RESULTS AND INTERPRETATION

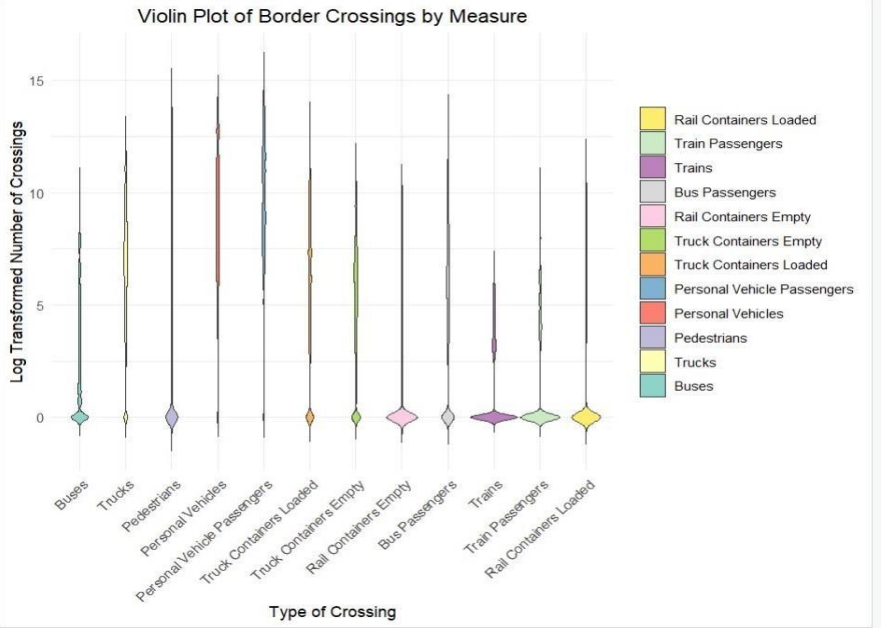
Bar Chart



Based on the total number of entries from the Border Crossing Entry dataset, the top 30 border crossing ports is displayed as a horizontal bar chart. By this visualization, authorities can give

importance to better resource allocation and policy considerations for efficient border control by identifying high- traffic entry points. Every bar represents a port, each port has a colour representation which shows which state it belongs to, and the longer bars represent the high number of entries where the smaller bars signify fewer total entries in that port. The differentiation of state is shown in different colours which can help officials make cross-region comparisons easier.

Violin plot



The above output tells the following details:

- Type of Crossing: A unique violin plot is used to illustrate each type of border crossing that is categorised along the x-axis. These include several kinds of containers and ways to get about, such as rail containers that are loaded and unloaded, cars, buses, and pedestrians.

- Log Transformed Number of Crossings: The log-transformed number of crossings is displayed on the y-axis. In data analysis, log transformation is a frequently employed approach for handling skewed data or bringing all data to the same scale. It can enhance the plot's interpretability and draw attention to patterns.

- Violin Shapes: For each category, the number of crossings is represented by a single "violin." Wider sections of the violin denote a larger frequency of data points in that range since the width of the violin at various places correlates to the density of the data. A violin plot depicts the complete distribution of the data, as contrast to a box plot which offers summary statistics.

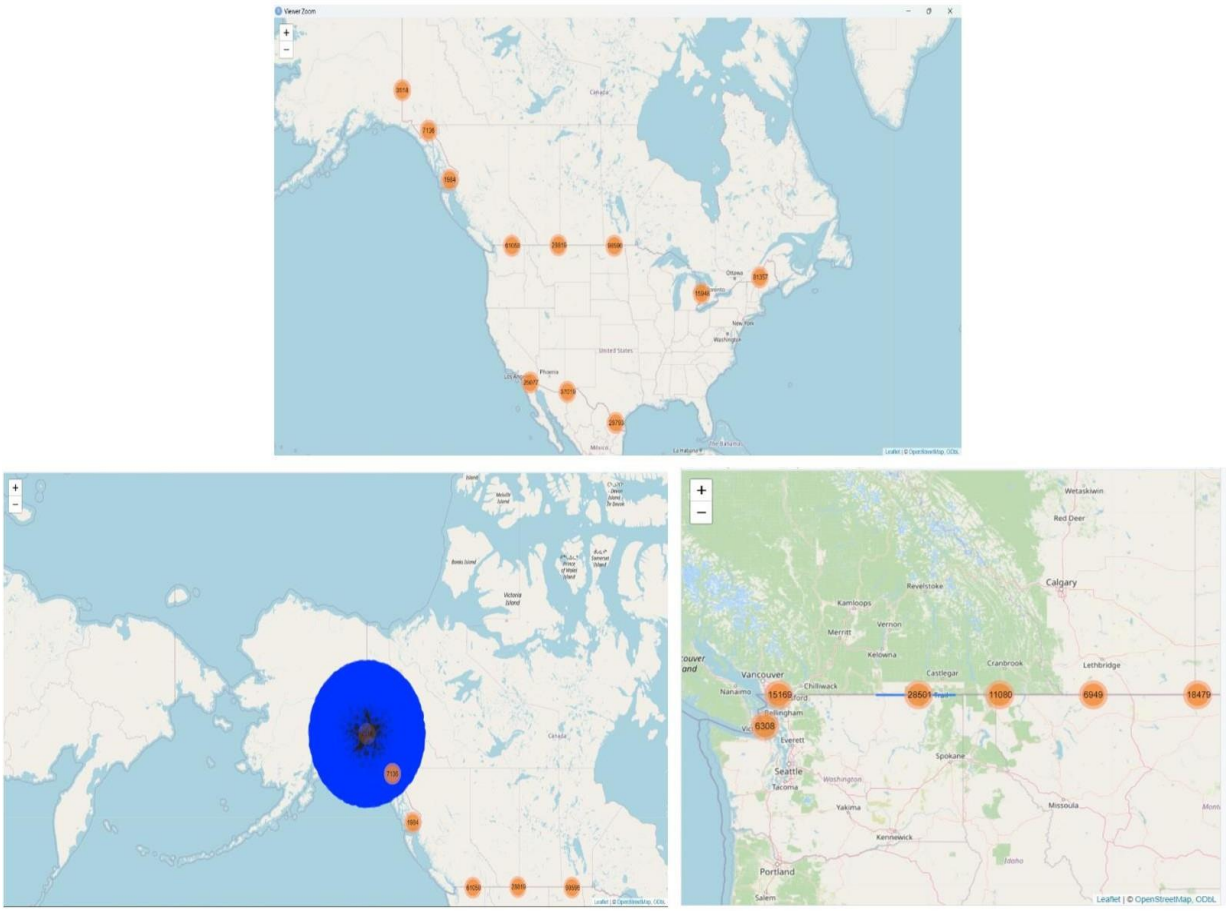
- Internal Box Plot: Usually, there is a tiny box plot inside the violins. The interquartile range, or the middle 50% of data points, is displayed by the box itself, while the thick line inside the box reflects the median of the data. Variability outside of the top and lower quartiles is indicated by the whiskers, which are narrow lines that extend from the box.

- Colour Legend: A colour legend corresponding particular colours to each kind of crossing is located on the right side. This helps to differentiate the many x-axis groupings.
  Data Dispersion and Outliers: The distribution of the data is indicated by the violin plots' thickness. For instance, a very thin violin would imply that the data is closely concentrated around one number, but a wide violin would suggest that there is greater fluctuation. Furthermore, any point that deviates significantly from the violins' body or centre may be regarded as an outlier.

- Utilisation of Logarithmic Scale: The y-axis is equipped with a logarithmic scale, where each tick mark denotes a tenfold increase in the number of crossings. This is especially helpful for presenting data that spans multiple orders of magnitude.

This plot is an effective tool for data analysis since it highlights central tendencies, dispersions, and outliers while offering a thorough understanding of the distribution of data over several categories. When you wish to compare the distributions across several categories or when the data is not normally distributed, you frequently utilise it.

Geospatial Map



The above output is a geospatial output, which is created by using package in called ‘leaflet’ which is used to create an interactive map visualizing the border crossing entry data.

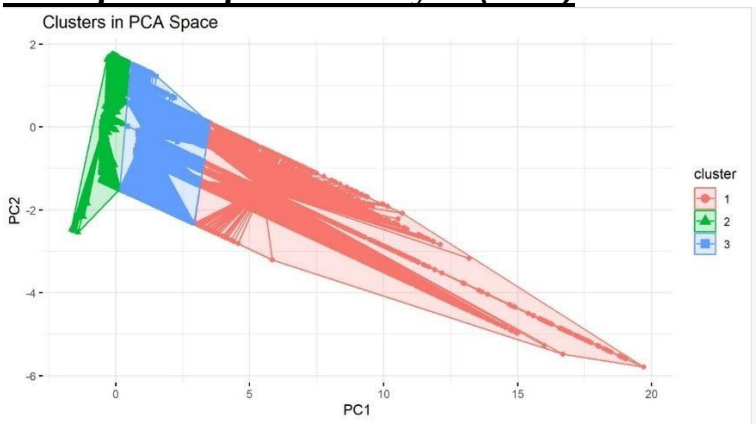
Users will be able to interactively examine the border crossing statistics with this leaflet map. With a given zoom level, the set View method centres the map on the United States. A service like OpenStreetMap can be used to generate base map tiles, and addCircleMarkers can add circle markers for each data point along with a popup window providing further details about each border crossing.

The geographic data dots on the maps each indicate a border crossing place from the dataset. The borders between the United States and Canada and Mexico are highlighted on the map of North America where these points are superimposed.

Cluster Map with Zoom: As indicated by the blue circles with numbers, the map looks to use clustering to put nearby locations together. The numbers indicate how many border crossings or entries there are in each cluster. These clusters would disintegrate into individual marks as one zooms in, as suggested by the "Viewer Zoom" feature. When there are numerous data points near together, this aids in controlling visual clutter.

Individual Data spots: The precise locations of the border crossings are shown on the map with individual orange spots. Each figure could be the overall number of crossings or a specific indicator of interest (such as the quantity of trucks, buses, or pedestrians) that was noted at that location over a specified time frame.

Principle Component Analysis (PCA)



The above visual representation is of clusters in PCA space.

PCA Axes: The first principal component is the horizontal axis, designated as PC1, while the second main component is the vertical axis, designated as PC2. According to PCA, these are the two most important axes of variance in the dataset.

Data Points: Every dot on the plot denotes an observation made using from the dataset and projected onto the PCA space that PC1 and PC2 define.

Three clusters are present, with each one being darkened in a distinct colour (Cluster 1 is red, Cluster 2 is green, and Cluster 3 is blue). After the PCA-transformed data is run through the k-means algorithm, clustering is the outcome.

Cluster 1: It represents "High Personal Vehicle Traffic" and most of the data points show border crossings with plenty of personal automobiles. (Red)

Cluster 2: It represents “High Truck Traffic” and it combines a variety of different measures in different quantities.

Cluster 3: It represents "High Pedestrian Traffic" and depicts crossings that are mostly used by pedestrians. (Blue)

CONCLUSION

Using bar charts, violin plots, and PCA, the study combined various border crossing patterns into illuminating clusters and spatial distributions. Key transit locations were identified using geospatial maps, and complex traffic dynamics were explained using PCA, which provided a framework for smart border management and policy improvement.

REFERENCES

[1] “Department of Transportation - Border Crossing Entry data,” Data.gov, Aug. 24, 2023. https://catalog.data.gov/dataset/border-crossing-entry-data-683ae

Submitted by: Amit Chowdary Kamma