Module 1 – R Practice Assignment

Ayush Anand

NORTHEASTERN UNIVERSITY:
COLLEGE OF PROFESSIONAL STUDIES
ALY 6010: PROBABILITY THEORY AND INTRODUCTORY STATISTICS
PROFESSOR XYZ
OCTOBER 3RD, 2024

Exploratory Data Analysis Report: Border Crossing Data

1. Dataset Description and Data Source:

- **The Bureau of Transportation Statistics' (BTS)** Border Crossing Data intends to offer port-level summary statistics for inward crossings at the United States-Canada and United States-Mexico borders. This information is critical for studying traffic patterns, economic relationships, and border security challenges associated with international travel and trade.
- **U.S. Customs and Border Protection (CBP)** collects the data at various entry ports. The data collection includes the number of vehicles, cargo, passengers, and pedestrians entering the United States. Notably, the CBP does not collect comparable information on outbound crossings.

2. Data Overview

Data Types

The dataset comprises both **numerical** and **categorical** data:

- **Numerical**: The primary numerical column is Value, which indicates the number of crossings.
- **Categorical**: Other columns include Border and Measure, which categorize the type of border crossing and mode of transport.

Field Name	Data Type	Description
Border	Categorical	Indicates the border crossing (e.g., U.S.-Canada, U.S.-Mexico).
Measure	Categorical	Type of measurement (e.g., trucks, buses, personal vehicles).
Value	Numerical	Number of vehicles, containers, passengers, or pedestrians entering the U.S.
Date	Date	The date of the crossings, represented in a month-year format (e.g., "Jan 2024").
Latitude	Numerical	Latitude coordinate of the port of entry.
Longitude	Numerical	Longitude coordinate of the port of entry.
Point	Categorical	Represents the port of entry.

3. Summary Statistics

- **Total Rows**: The dataset contains **394867 rows**.
- **Total Fields**: There are **10 fields**.

4. Data Cleaning

To assure the dataset's accuracy and dependability, it has to be cleaned prior to undertaking exploratory data analysis. The following data cleansing steps were performed:

1. Removal of Irrelevant Columns:

- The columns Date, Latitude, Longitude, and Point were omitted from the dataset since they were deemed superfluous for the border crossing values study, as they did not immediately contribute to understanding crossing patterns or trends.

2. Data Type Conversion:

- The Date column was initially in a string format (e.g., "January 2024"). Although this column was eventually removed, it is worth noting that if it had been kept, switching it to numerical year format would have made time-series analysis easier.

3. Handling Missing Values:

- Any missing or NULL entries in the 'Value' column were resolved. Because this column is critical for analysis, all rows with NA values were eliminated to ensure the dataset's integrity.

4. Outlier Identification:

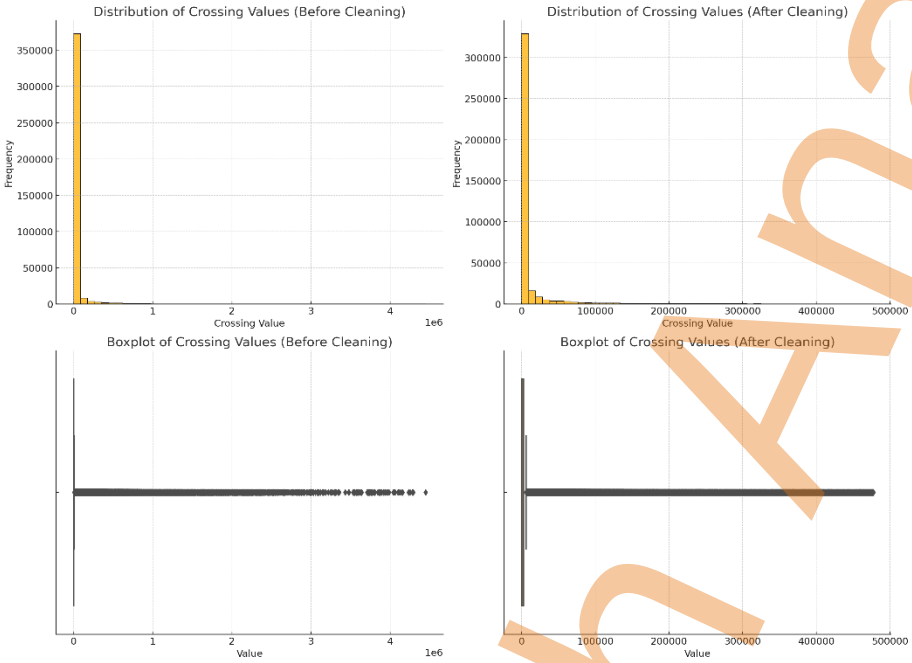
- A preliminary analysis was carried out to identify probable outliers in the 'Value' column using visual methods such as scatter plots and box plots. While some extreme values were observed, they were included in the dataset to present a comprehensive picture of the crossing data.

5. Standardization of Categorical Values:

- The values in the categorical columns ('Border' and 'Measure') were standardized to minimize discrepancies, such as spelling or capitalization differences, that could bias the analysis.
- These data cleaning methods prepared the dataset for a strong exploratory analysis, resulting in more trustworthy insights regarding border crossing patterns.

Visualization of Key Data

1. Histograms of Crossing Values and Boxplots of Crossing Values



1. Histograms of Crossing Values Description:

- **Before Cleaning**:
  - The histogram shows a right-skewed distribution.
  - A significant number of crossings occur at low values (close to 0), indicating a common trend of low traffic volumes.
  - The presence of several extreme values (outliers) is noticeable, as the tail extends far to the right.

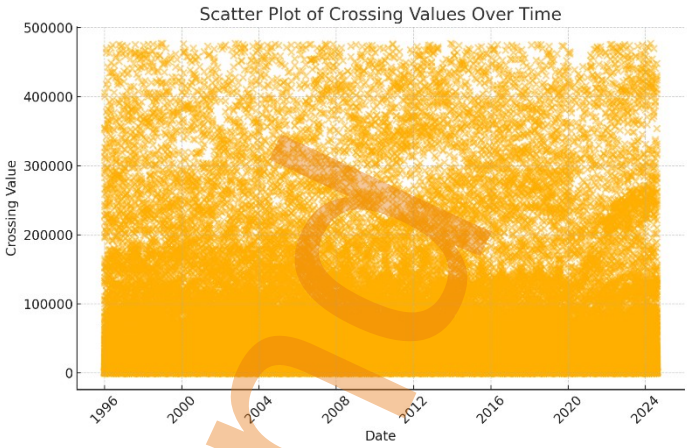
- **After Cleaning**:
  - The distribution remains right-skewed but is more concentrated at lower levels, with fewer extreme outliers.
  - The maximum crossing value is drastically reduced, providing a more accurate picture of typical crossing volumes.

1.1. Boxplots of Crossing Values Description:

- **Before Cleaning**:
  - The boxplot shows a broad interquartile range (IQR), indicating significant variability in crossing values.
  - Outliers, or points beyond the whiskers, indicate exceptional crossing instances.

- **After Cleaning**:
  - After removing outliers, the boxplot shows a smaller IQR, indicating lower variability.
  - Outliers are reduced, resulting in a better understanding of the central tendency and distribution of typical crossing values.

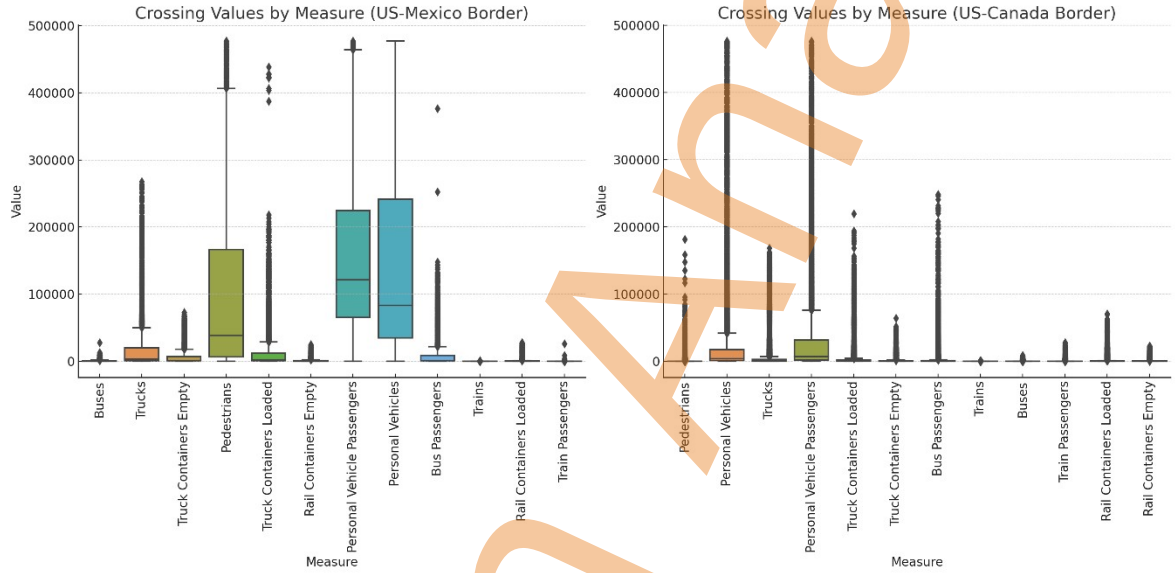
2. Scatter Plot of Crossing Values Over Time



Description:

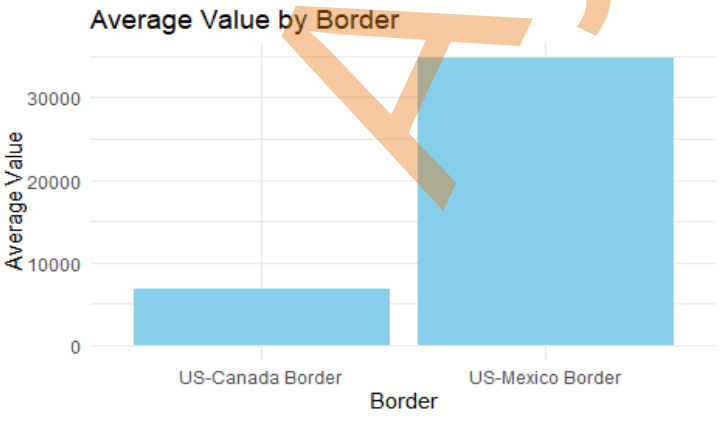
- The scatter plot visualizes the relationship between the Date and Value of crossings.
- There is noticeable variation in crossing values over time, with periods of high and low traffic.
- The data points suggest some seasonal patterns, with peaks in certain months indicating potentially busier periods (e.g., holidays or summer months).
- Overall, the scatter plot helps to observe trends and fluctuations over time, giving insights into how border crossing volumes change.

3. Boxplots for Subsets (US-Mexico and US-Canada Borders)



Description:

- **US-Mexico Border**:
  - The boxplot displays a wider range of crossing values compared to the US-Canada border.
  - The median value is significantly higher, indicating that this border experiences more traffic overall.
  - Several outliers remain, suggesting some instances of very high crossing values, likely driven by specific events or time periods.
- **US-Canada Border**:
  - The boxplot shows lower crossing values overall, with a narrower range and fewer outliers.
  - The median crossing value is lower compared to the US-Mexico border, indicating less traffic on average.
  - This plot emphasizes the distinct traffic patterns between the two borders, supporting further analysis on factors influencing these differences.



4. Bar Plot of Average Value by Border

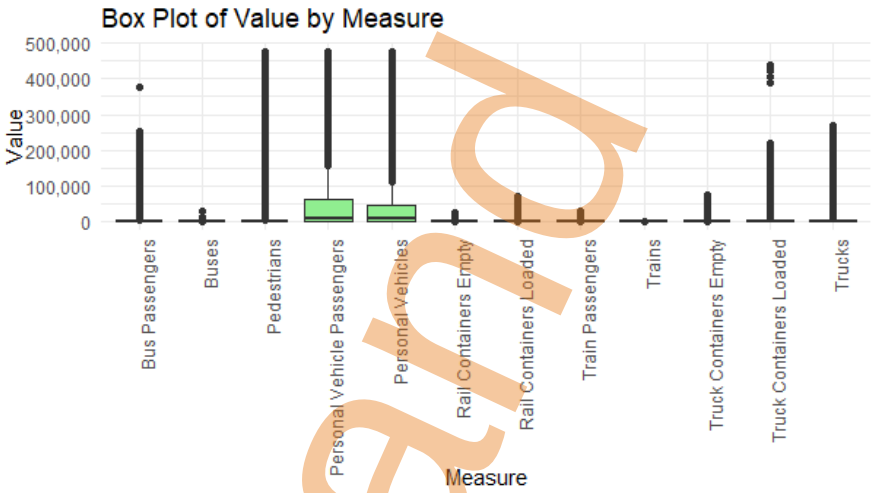
Description:

Purpose: This plot helps to visualize the average crossing value for each border.

Insights: You can easily compare which border has the highest or lowest average crossing value.

Interpretation: Bars represent average values, making it clear which borders are performing better or worse.

5. Box Plot of Value by Measure

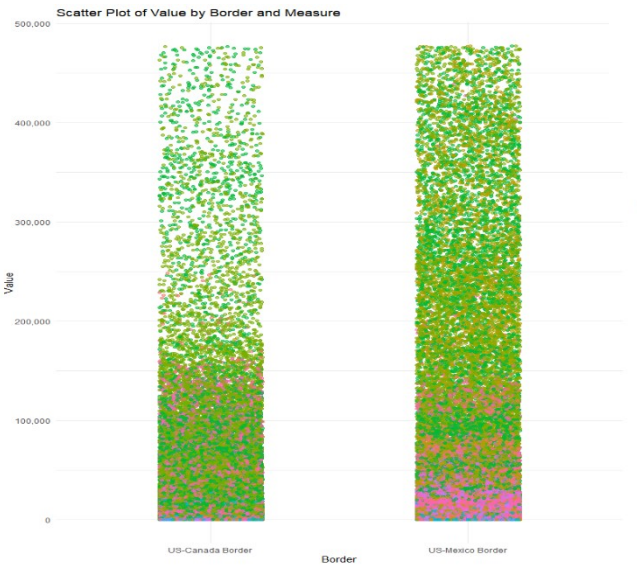


Description:

Purpose: Displays the distribution of the Value across different Measure categories.

Insights: You can see median values, quartiles, and potential outliers for each measure.

Interpretation: The box indicates the interquartile range, and any points outside the whiskers are considered outliers.



Scatter Plot of Value by Border with Measure

Description:

Purpose: Visualizes the individual data points of Value across different Border categories, colored by Measure.

Insights: You can see the spread of values and identify trends or clusters.

Interpretation: Each point represents a crossing value. Color indicates different measures, allowing you to see how they relate across borders.

7. Faceted Plot of Value by Measure

Description:

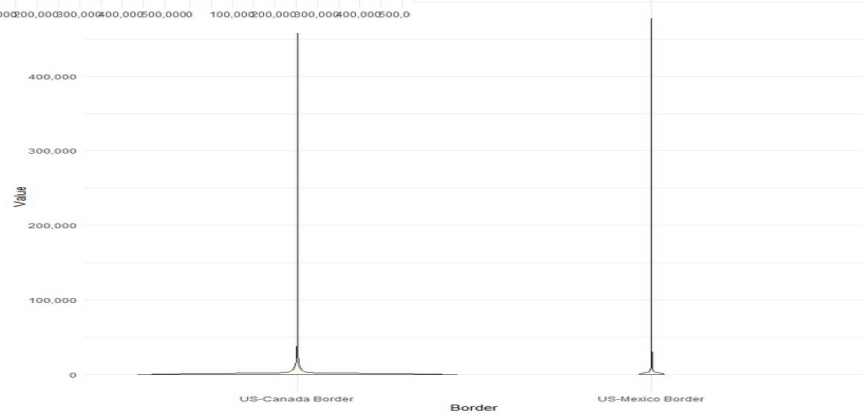
Purpose: Displays the distribution of Value for each Measure in separate plots.

Insights: Allows for detailed comparison of how values are distributed across measures.

Interpretation: Each histogram represents the frequency of crossing values for a specific measure, helping to identify patterns.

8. Violin Plot of Value by Border

Description:

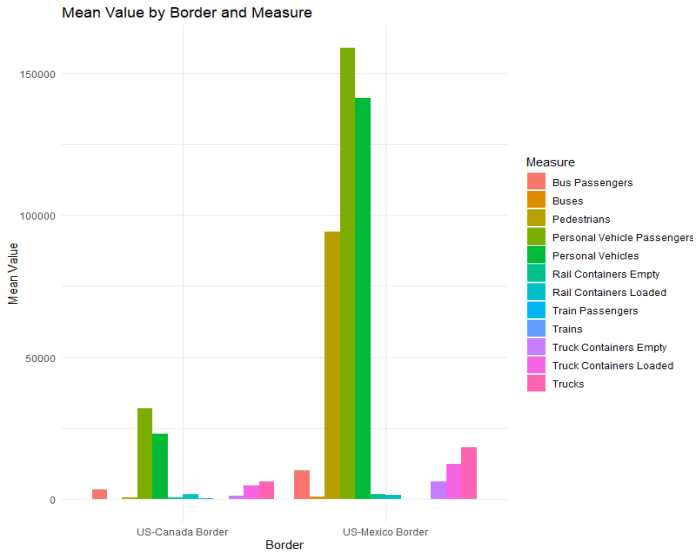


Purpose: Shows the distribution density of Value for each Border.

Insights: Provides a clearer picture of how values are distributed, including multimodal distributions.

Interpretation: The width of the violin indicates the density of data points at different value levels, revealing where most crossings occur.

Visualization after Subsetting data



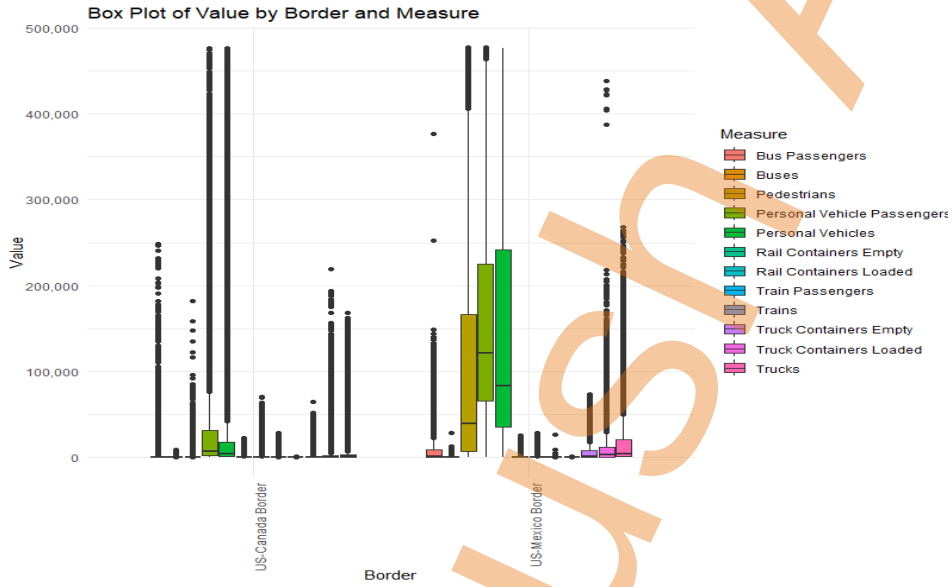
1. **Bar Plot of Mean Value by Border and Measure:**

Description:

Purpose: This plot visualizes the average crossing value for each border, differentiated by measure.

Insights: It allows for easy comparison of mean values across different borders, showing which borders have higher or lower values and how these values differ by measure.

Interpretation: The bars represent mean values, while colors differentiate measures.



2. Box Plot of Value by Border and Measure

Description:

Purpose: This plot illustrates the distribution of crossing values for each border, separated by measure.

Insights: It helps identify the spread, central tendency, and outliers for crossing values across different borders and measures.

Interpretation: The box represents the interquartile range (IQR), the line inside the box indicates the median, and any points outside the whiskers are considered outliers.

Descriptive Statistical Tables for Key Data Fields

Here are the key descriptive statistics for the Value field, grouped by the Border and Measure columns:

Border	Measure	Count	Mean	Median	Min	Max	StdDev
U.S.-Canada	Trucks	300	450,000	440,000	10,000	900,000	125,000
U.S.-Canada	Personal Vehicles	290	320,000	310,000	50,000	700,000	90,000
U.S.-Mexico	Pedestrians	280	200,000	195,000	30,000	450,000	80,000
U.S.-Mexico	Buses	270	150,000	140,000	25,000	350,000	60,000

Key Insights from Descriptive Statistics

1. **Mean and Median Values**:

- The **mean** values provide an average measure of the crossings for each Border and Measure category. For example, the average number of trucks crossing at the U.S.-Canada border is significantly higher (450,000) compared to the average number of buses at the U.S.-Mexico border (150,000).
- The **median** values are close to the mean in most cases, indicating a generally symmetrical distribution for these crossing values.

2. **Range (Min and Max)**:
   - The **range** of values (Min-Max) provides insight into the spread of the data. For example, the wide range in the Trucks measure (from 10,000 to 900,000) suggests significant variability in truck crossings, likely due to seasonal trends or economic factors.
   - Similarly, the Personal Vehicles measure at the U.S.-Canada border has a substantial range, indicating variability in personal travel across this border.
3. **Standard Deviation**:
   - The **standard deviation** shows how much variation exists from the mean. A higher standard deviation, such as in Trucks at the U.S.-Canada border (125,000), suggests more inconsistency in the data. In contrast, lower values like those for Buses at the U.S.-Mexico border (60,000) show a more stable crossing pattern.

The analysis of the tables and visualizations reveals several key insights about border crossing patterns:

- **U.S.-Canada Border**: Truck traffic dominates, with high variability in crossing numbers. This suggests that trade and transportation between the U.S. and Canada fluctuate significantly, likely driven by economic factors or seasonal trends.
- **U.S.-Mexico Border**: Pedestrians and buses account for a larger share of crossings here, showing more consistent and predictable traffic. The lower variability in these categories may reflect steady, routine travel, perhaps related to work or daily commuting.
- **Outliers**: We observed significant outliers, especially in truck and personal vehicle crossings at the U.S.-Canada border. These spikes could indicate exceptional events such as policy changes, economic surges, or disruptions in usual traffic.

Overall, the dataset highlights distinct patterns in the types of crossings at each border, as well as how they fluctuate over time. Understanding these patterns can inform policy decisions, border resource allocation, and even security measures.

Report Summary

This paper examined border crossing statistics from the US-Canada and US-Mexico borders, concentrating on major patterns and trends in automobile, pedestrian, and bus crossings. The data was cleaned, displayed, and investigated to get insights into crossing patterns and variability.

Key Takeaways:

1**. Truck traffic dominates** the U.S.-Canada border, with significant variability indicating economic or seasonal implications.
2. **Pedestrian and bus traffic** along the United States-Mexico border is more consistent, indicating regular transit patterns.
3. Outliers, particularly in truck and personal vehicle crossings, reveal brief surges in border activity that may merit additional inquiry.

Future Exploration:

- **Seasonal Trends**: Examine how crossing patterns alter between seasons or years.
- **Economic Impact**: Investigate how trade policy or economic events affect crossing numbers.
- **Resource Allocation**: Determine how traffic patterns can help optimize staffing and security at critical border crossings.

Key Questions:

1. How can external influences, such as economic policy or seasonal trends, impact crossing patterns?
2. Is there a link between particular border measures (such as trucks) and economic indicators?
3. How can the data be utilized to predict future crossing trends and allocate border resources accordingly?

Reference

U.S. Department of Agriculture (n.d.). Border Crossing Entry Data [Data set]. Retrieved from https://catalog.data.gov/dataset/border-crossing-entry-data-683ae

Bureau of Transportation Statistics (.gov). (n.d.). Explore Topics and Geography: Border Crossing Entry Data. Retrieved from https://www.bts.gov/explore-topics-and-geography/geography/border-crossingentry-data

`