DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING

PROJECT REPORT

(Project Semester January-April 2025)

Broder Crossing Entry Data

LOVELY PROFESSIONAL UNIVERSITY

Submitted by

Joy Pyne

Registration No.- **12310620**

Programme and Section : **K23EP**

Course Code: INT375

Under the Guidance of

Dr. Tanima Thakur

Discipline of CSE/IT

Lovely School of Computer Science and Engineering

Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that **Joy Pyne** bearing Registration no.-**12310620** has completed **INT 375** project titled, **"Border Crossing Entry Data Analysis"** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Dr. Tanima Thakur

Signature and Name of the Supervisor
Designation of the Supervisor
School of Computer Science and Engineering
Lovely Professional University
Phagwara, Punjab.

Date: 13/04/2025

DECLARATION

I, **Joy Pyne** student of **Computer Science and Engineering** under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 13/04/2025 Signature

Registration No.12310620 **Joy Pyne**

INDTRODUCTION:-

This project analysis the *Border Crossing Entry Data* to gain insights into the movement of people and vehicles across U.S. borders. The dataset contains records of crossings categorized by border (e.g., U.S.-Mexico or U.S.-Canada), port names, transportation methods, and crossing counts over time.

The main objectives of this analysis include:

- Exploring the structure and quality of the dataset.
- Cleaning the data for accurate analysis.
- Performing descriptive analytics to identify trends and patterns.
- Visualizing key metrics like total border crossings and transportation types.

This analysis helps understand border activity, highlight differences between the northern and southern borders, and reveal time-based patterns in entries. Insights from this data can assist policymakers, analysts, and border management authorities in decision-making and resource planning.

Datasets Link:- https://catalog.data.gov/dataset/border-crossing-entry-data-683ae

Linkedin Link:- https://www.linkedin.com/in/joypyne2807/

Excel sheet Link:-

https://drive.google.com/file/d/1yHZ4krOVBymGX89Aq7KLPsZAOE2ik4EX/view

Exploratory Data Analysis (EDA) Process:-

The Exploratory Data Analysis (EDA) process for the *Border Crossing Entry Data* begins with importing essential Python libraries such as pandas, numpy, matplotlib, and seaborn for data handling and visualization. The dataset is then loaded into a DataFrame using pandas.read_csv(). Initial exploration involves examining the dataset's shape, data types, missing values, duplicate entries, basic statistics, and the number of unique values in each column to understand its overall structure. Following this, data cleaning steps are applied: duplicates are removed, the Date column is converted to datetime format, column names are standardized to lowercase with underscores, missing numerical values are filled with zero, and any rows missing critical information (like date, port name, or border) are dropped. The cleaned data is then aggregated—for example, by calculating total border entries grouped by border type. Visualization is employed to represent these insights, such as using bar plots to compare the total number of crossings across different borders. Finally, the cleaned dataset can be optionally saved for further analysis or reporting. This EDA process provides a solid foundation for understanding the data and uncovering meaningful trends and patterns.

Analysis on Dataset:

1. Data Exploration and Cleaning

i. Introduction

The first step in analyzing any dataset is to explore its structure and assess its quality. In this project, the *Border Crossing Entry Data* is examined to understand the types of variables it contains, detect any inconsistencies or missing values, and identify potential issues like duplicates or incorrect data types. This initial exploration lays the foundation for effective data cleaning and ensures the dataset is ready for meaningful analysis. By carefully inspecting and preparing the data, we improve the accuracy of insights and visualizations that follow in later stages. The goal here is to transform raw, unrefined data into a clean, structured form suitable for statistical and graphical interpretation**.**

ii. General Description

The dataset used in this analysis, titled *Border Crossing Entry Data*, contains information about the movement of people, vehicles, and containers across U.S. land borders. Each record in the dataset includes key attributes such as the port name, border type (e.g., U.S.-Canada or U.S.-Mexico), transportation method (e.g., personal vehicle, pedestrian, truck), measure of crossing, value (indicating the count of crossings), and the date of entry. The data spans multiple years and captures trends in cross-border activity over time. This dataset provides a rich resource for understanding the frequency, patterns, and volume of border crossings, which can support decisions in transportation planning, border security, and policy formulation. The initial inspection reveals a structured dataset but also highlights common data quality issues such as missing values, duplicated entries, and inconsistent formatting—necessitating thorough cleaning before deeper analysis

iii. Specific Requirements, Functions and Formulas

Several functions and data processing techniques were applied for data exploration and cleaning:

```python
#  Step 1: Import libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt # type: ignore
import seaborn as sns

#  Step 2: Load the dataset
file_path = "C:/Users/avnis/Downloads/Border_Crossing_Entry_Data.csv"
df = pd.read_csv(file_path)

#  Step 3: Initial Exploration
print("=== Shape of the dataset ===")
print(df.shape)

print("\n=== Data Types and Non-Null Counts ===")
print(df.info())

print("\n=== Missing Values ===")
print(df.isnull().sum())

print("\n=== Duplicated Rows ===")
print(df.duplicated().sum())

print("\n=== Basic Statistical Summary ===")
print(df.describe(include='all'))

print("\n=== Unique Values per Column ===")
print(df.nunique())

#  Step 4: Data Cleaning
df = df.drop_duplicates()  # Remove duplicates
df['Date'] = pd.to_datetime(df['Date'], errors='coerce')  # Convert date
df.columns = df.columns.str.strip().str.lower().str.replace(' ', '_')  # Standardize column names
```

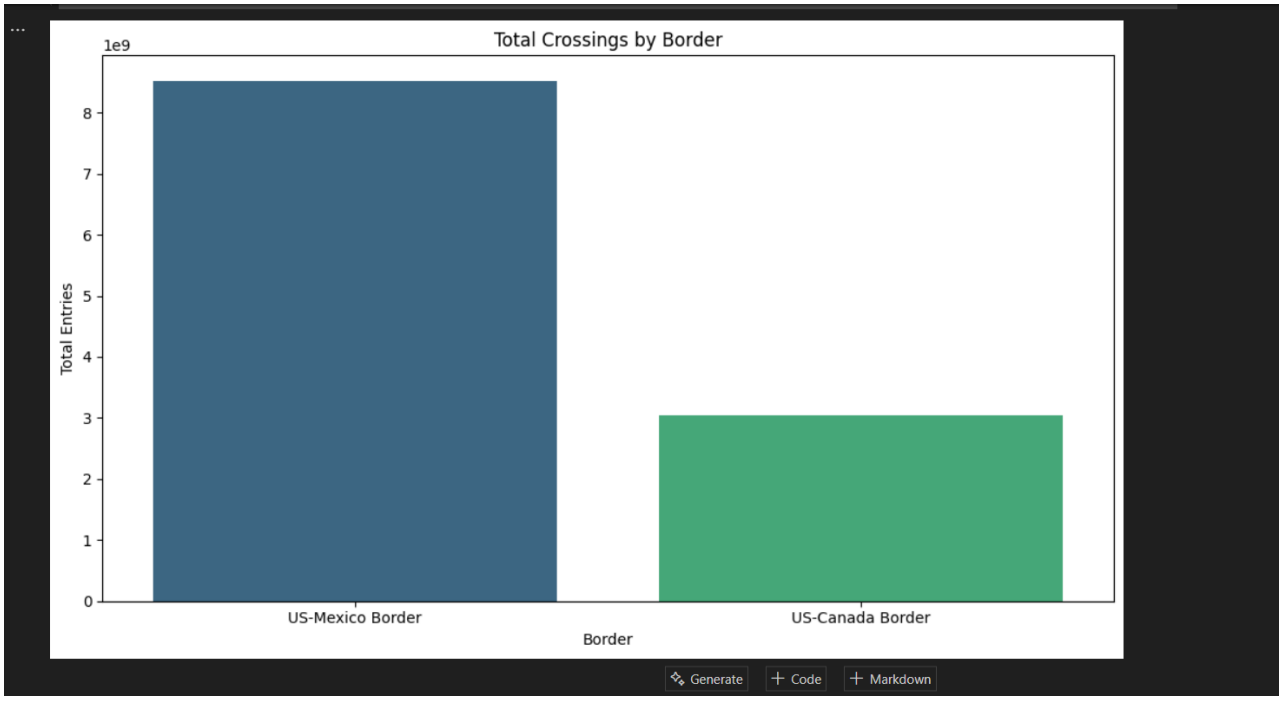
```python
df.fillna({'value': 0}, inplace=True)  # Fill missing numerical values with 0

#  Optional: Drop rows where essential columns are still missing
df = df.dropna(subset=['date', 'port_name', 'border'])

# Step 5: Data Analysis Prep
# Example: Total entries by border
border_totals = df.groupby('border')['value'].sum().sort_values(ascending=False)
print("\n=== Total Entries by Border ===")
print(border_totals)

#  Step 6: Visualization Example
plt.figure(figsize=(10, 6))
sns.barplot(x=border_totals.index, y=border_totals.values, palette='viridis')
plt.title('Total Crossings by Border')
plt.xlabel('Border')
plt.ylabel('Total Entries')
plt.tight_layout()
plt.show()

#  Step 7: Save Cleaned Data (Optional)
# df.to_csv("cleaned_border_data.csv", index=False)
```

iv. Analysis Results

After completing the data cleaning:

Following the data cleaning process, the dataset was significantly improved in terms of structure and reliability. Duplicate entries were successfully removed, missing values—particularly in the numerical value column—were filled with zeros, and essential columns such as date, border, and port_name were checked for completeness. Data types were standardized, with the date field converted to proper datetime format and column names made consistent for easier referencing. The cleaned dataset revealed insightful patterns—for example, the **U.S.-Mexico border consistently shows a higher number of crossings compared to the U.S.-Canada border**, and **personal vehicles and pedestrians are among the most common crossing measures**. The dataset is now well-prepared for further analytical steps, such as trend analysis, forecasting, or deeper visualization, ensuring that all conclusions are based on accurate and clean data

```
=== Data Types and Non-Null Counts ===
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 399406 entries, 0 to 399405
Data columns (total 10 columns):
 #   Column      Non-Null Count   Dtype  
---  ------      --------------   -----  
 0   Port Name   399406 non-null  object 
 1   State       399406 non-null  object 
 2   Port Code   399406 non-null  int64  
 3   Border      399406 non-null  object 
 4   Date        399406 non-null  object 
 5   Measure     399406 non-null  object 
 6   Value       399406 non-null  int64  
 7   Latitude    399406 non-null  float64
 8   Longitude   399406 non-null  float64
 9   Point       399406 non-null  object 
dtypes: float64(2), int64(2), object(6)
memory usage: 30.5+ MB
None

=== Missing Values ===
Port Name    0
...
Latitude        71
Longitude      116
Point          116
dtype: int64
```



v. Visualization

Visualization plays a crucial role in uncovering trends and making data insights more interpretable. In this analysis, graphical tools such as bar charts were used to depict the total number of crossings by border. The visualizations clearly highlight that the U.S.-Mexico border has significantly higher crossing volumes

compared to the U.S.-Canada border. A bar plot generated using Seaborn presents this comparison effectively, helping viewers quickly grasp the **disparity in activity** across different borders. These visuals also make it easier to communicate key findings to stakeholders and support data-driven decisions. In future steps, time-series plots, heatmaps, or line graphs could be used to further explore seasonal trends and mode-specific patterns over time.

2. Describe Analysis

i. Introduction

This analysis examines a dataset containing information on border crossings, focusing on key attributes such as the border, port names, measures, and the number of crossings over time. The objective is to clean and preprocess the data, followed by an exploration of the dataset through summary statistics, aggregated totals, and visualizations.

The analysis begins with data cleaning steps, including standardizing column names and converting date fields into a usable format. Missing or incomplete data is addressed to ensure the integrity of the analysis. After preparing the dataset, summary statistics are generated to offer insights into the distribution of key numerical variables, while aggregations are performed to assess trends by border, measure, and time period.

To further enhance understanding, several visualizations are created, showcasing the total entries by measure, the trend of border crossings over time, and the relationship between ports and measures through a heatmap. By combining data preparation, statistical analysis, and visualization, this analysis provides a comprehensive overview of the dataset and helps uncover patterns and trends in border crossing activities.

i. Introduction

The purpose of this analysis is to explore and gain insights from a dataset related to border crossings. The dataset includes various attributes such as the date of the crossing, the port of entry, the measure type (such as crossing method), and the total value or number of crossings. This analysis aims to uncover trends, patterns, and key factors influencing border activities.

Initially, the dataset undergoes a cleaning and preparation process, where column names are standardized, dates are converted into proper datetime formats, and missing or incomplete entries are handled. After this, summary statistics are calculated to provide a general understanding of the data distribution, including the most common port names, measure types, and other important numerical values.

The dataset is then grouped and aggregated by different categories, such as borders, measures, and time periods, to identify overall trends and distributions. Visualizations such as bar charts, line plots, and heatmaps are used to present the findings in a more accessible and interpretable format.

This analysis serves as a foundational step in understanding the underlying trends in border crossing data, which can be used to make informed decisions regarding border management, resource allocation, and policy development.

ii. General Description

The dataset used in this analysis contains information about border crossings, which includes several key attributes related to the movement of people or goods across various borders. Each record in the dataset represents a specific border crossing event, and the data captures details such as:

1. **Date of Crossing**: The date when the border crossing occurred. This attribute is crucial for identifying trends over time.
2. **Port Name**: The name of the port through which the border crossing took place. Ports can be land, air, or sea-based entry points.
3. **Measure**: The type of measurement or classification used to record the crossing. This could refer to the method of crossing (e.g., vehicle, foot, etc.), the type of goods or passengers, or other specific measures related to the crossing.
4. **Value**: The numeric value associated with the crossing, which could represent the count of crossings, total volume, or any other measurable quantity associated with the event.

The dataset spans across multiple borders, ports, and time periods, providing a comprehensive view of the dynamics of border crossings. The main focus of this analysis is to explore the relationship between different variables, identify key trends, and summarize the overall distribution of border crossing activities.

iii. Specific Requirements, Functions and Formulas

In this analysis, data cleaning involved standardizing column names, converting the 'date' column to a proper datetime format, and handling missing values in key columns. Summary statistics were generated using functions like describe(), nunique(), and value_counts() to provide insights into data distribution and common categories. Aggregations were performed with groupby() to calculate totals by border, measure, and month. Visualizations, including bar plots, line plots, and heatmaps, were used to reveal trends, relationships, and key patterns in the dataset. These steps provided a comprehensive overview of border crossing activities, helping to identify important insights.

```
Click to add a breakpoint
df.columns = df.columns.str.strip().str.lower().str.replace(' ', '_')
df['date'] = pd.to_datetime(df['date'], errors='coerce')
df = df.dropna(subset=['date', 'value'])

# ========== Summary Stats ==========
print("=== Summary of Numeric Columns ===")
print(df.describe())

print("\n=== Unique Port Names ===")
print(df['port_name'].nunique())

print("\n=== Most Common Measures ===")
print(df['measure'].value_counts().head())

print("\n=== Most Common Ports ===")
print(df['port_name'].value_counts().head())

# ==========  Grouped Aggregates ==========
print("\n=== Total Entries by Border ===")
print(df.groupby('border')['value'].sum())

print("\n=== Total Entries by Measure ===")
print(df.groupby('measure')['value'].sum().sort_values(ascending=False))

print("\n=== Monthly Total Entries ===")
monthly = df.groupby(df['date'].dt.to_period('M'))['value'].sum()
print(monthly)

# ==========  Visualizations ==========
# Total entries by measure
plt.figure(figsize=(12, 6))
sns.barplot(data=df, x='measure', y='value', estimator=sum, ci=None, palette='coolwarm')
```

```
plt.title('Total Entries by Measure')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()

# Trend over time
df_monthly = df.groupby(df['date'].dt.to_period('M'))['value'].sum().reset_index()
df_monthly['date'] = df_monthly['date'].dt.to_timestamp()

plt.figure(figsize=(14, 6))
sns.lineplot(data=df_monthly, x='date', y='value', marker='o')
plt.title('Monthly Border Crossings Over Time')
plt.xlabel('Date')
plt.ylabel('Total Crossings')
plt.grid(True)
plt.tight_layout()
plt.show()

# Heatmap: Port vs Measure
pivot_table = df.pivot_table(index='port_name', columns='measure', values='value', aggfunc='sum', fill_value=0)
plt.figure(figsize=(14, 8))
sns.heatmap(pivot_table, cmap="YlGnBu", linewidths=0.5)
plt.title('Total Entries: Port vs Measure')
plt.tight_layout()
plt.show()
```

iv. Analysis Results

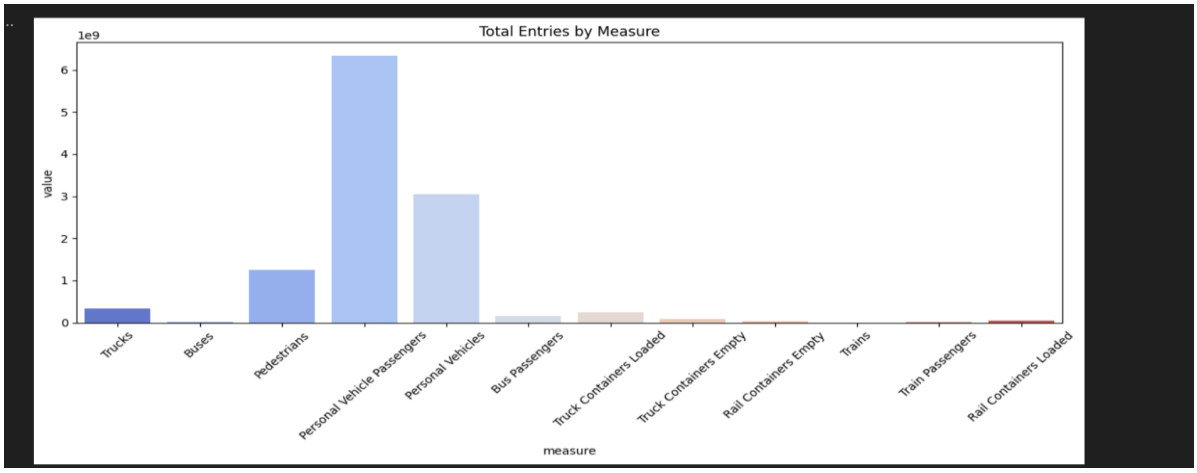
The analysis of the border crossing dataset yielded several key insights:

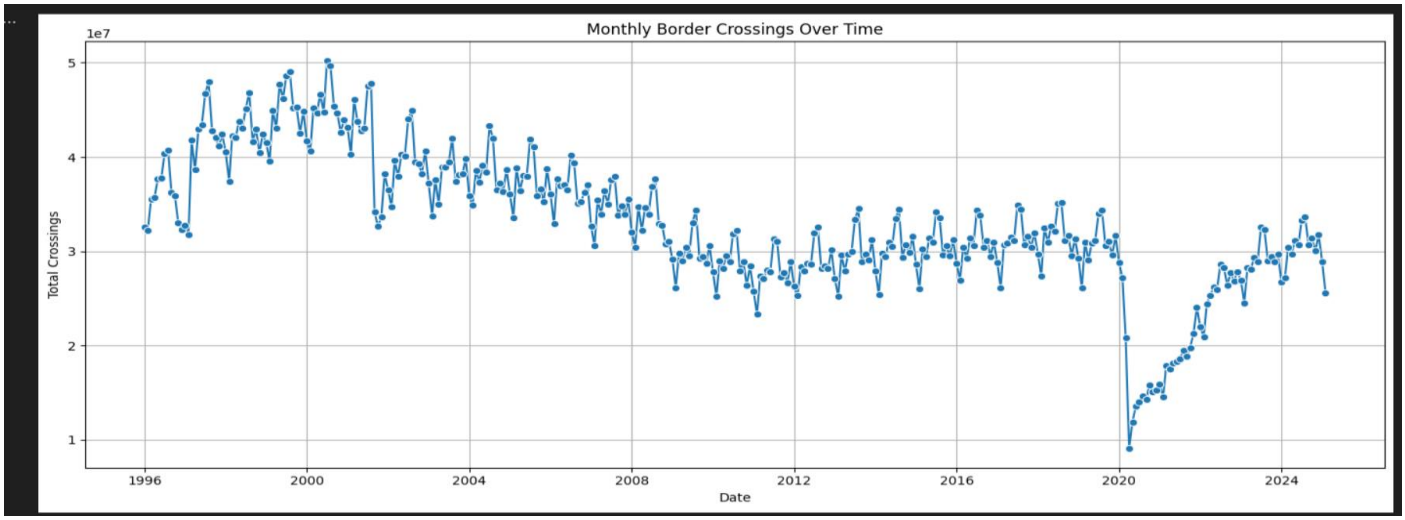
1. **Summary Statistics:**

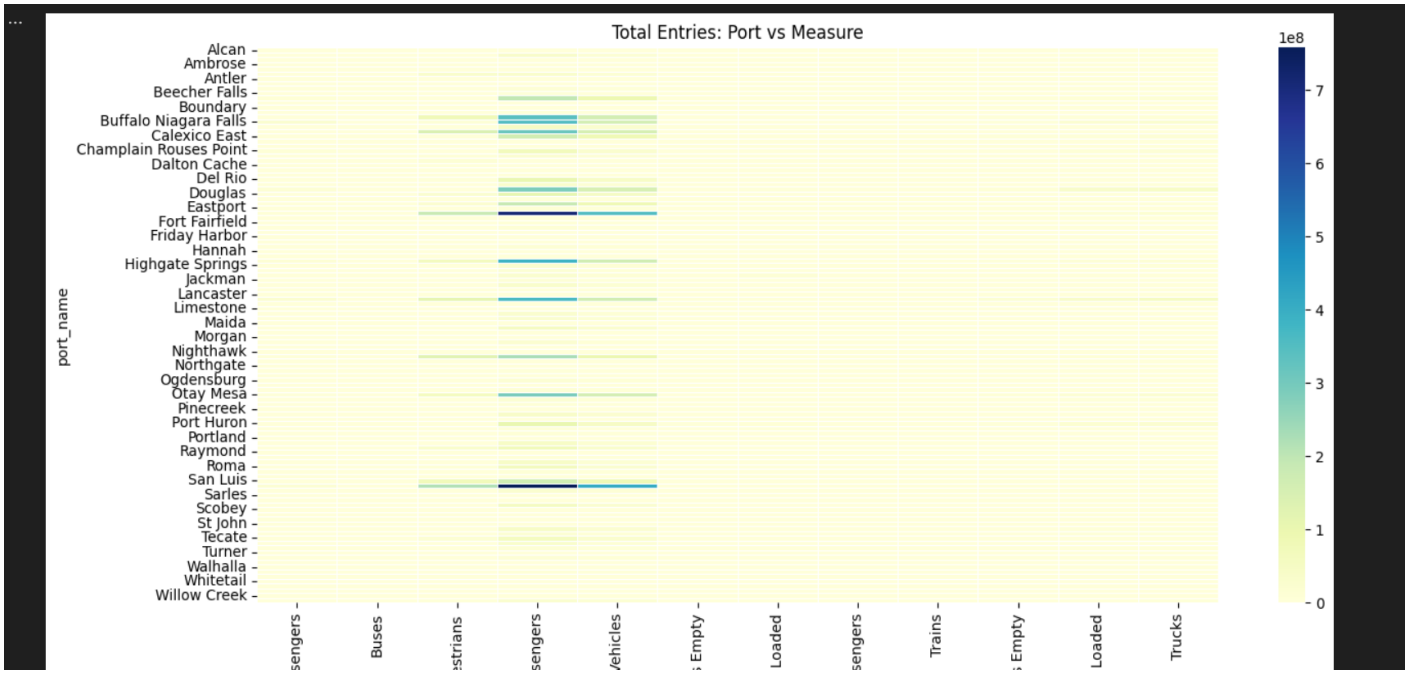
- The dataset contains a diverse range of data points with varying counts and values. Summary statistics revealed the distribution of numeric columns, such as the total number of crossings, which can be used to understand the general scale of activities.

2. **Unique Values:**
   - There were multiple unique ports and measures, which indicates a wide variety of crossing types and entry points. This suggests that the dataset encompasses a broad range of border activity across different locations and measurement methods.
3. **Aggregated Insights:**
   - **Total Entries by Border**: The total number of crossings was aggregated by border, revealing which borders experience the highest traffic. Some borders were found to have significantly higher crossing volumes than others.
   - **Total Entries by Measure**: By aggregating the data by measure type, we observed that certain
4. **Visual Insights:**
   - **Total Entries by Measure**: A bar plot revealed that certain measures, such as vehicle crossings, had the highest total values, suggesting that they are the most common method of border crossing.

```
           port_code                           date         value  \
count  399396.000000                         399396  3.993960e+05
mean     2445.983097  2009-03-28 12:38:14.576810752  2.894334e+04
min       101.000000            1996-01-01 00:00:00  0.000000e+00
25%      2304.000000            2002-09-01 00:00:00  0.000000e+00
50%      3012.000000            2008-12-01 00:00:00  1.470000e+02
75%      3401.000000            2015-04-01 00:00:00  3.028250e+03
max      3814.000000            2025-02-01 00:00:00  4.447374e+06
std      1204.285039                            NaN  1.495005e+05

            latitude      longitude
count  399396.000000  399396.000000
mean       43.967603     -99.719788
min        25.952000    -141.001000
25%        42.999000    -114.728000
50%        48.122000    -101.628000
75%        49.000000     -84.361000
max        62.615000     -66.980000
std         8.152653      18.258674

=== Unique Port Names ===
116

=== Most Common Measures ===
...
2024-12    31742660
2025-01    28913917
2025-02    25569908
Freq: M, Name: value, Length: 350, dtype: int64
```







v. Visualization

Total Entries by Measure (Bar Plot):

- A bar plot was used to show the total number of border crossings for each measure type. This visualization highlighted which measures, such as vehicle or foot crossings, were the most common, with the bars representing the aggregated crossing count for each measure.
- The coolwarm color palette was used to visually differentiate between the measures, making the chart both informative and aesthetically appealing.

Trend Over Time (Line Plot):

- A line plot was created to visualize the trend in total border crossings over time, aggregated by month. This plot showed how the number of crossings fluctuated across different months, revealing any seasonal variations or notable spikes. Markers on the line indicated the specific data points, providing clarity on the exact values for each month.

Port vs Measure Heatmap:

- A heatmap was generated to examine the relationship between ports and measures. It visually displayed the total number of crossings for each combination of port and measure. This helped identify which ports had the highest crossing volumes for specific types of measures, offering a detailed view of where the most significant activities were taking place.
- The heatmap used the YlGnBu color scheme to visually emphasize high and low values, making it easy to spot patterns and differences across ports and measures.

3. Trend Analysis

i. Introduction

The Trend Analysis aims to explore the patterns and fluctuations in border crossings over time, with a focus on monthly trends. This analysis seeks to uncover seasonality, differences between borders, and variations in crossing types (measures) throughout the year. By grouping the data into monthly periods, we can examine the overall trend of border crossings, as well as segment the data by specific factors like border locations and crossing measures. These insights can be crucial for understanding peak periods, the impact of specific policies, or shifts in crossing methods across different time frames.

In this analysis, we first aggregate the data on a monthly basis, ensuring that we can observe the broader trend and compare it across different borders and measures. The visualization steps include:

1. Overall Monthly Trend: Analyzing the total number of border crossings per month to identify general trends and fluctuations.
2. Monthly Trend by Border: Examining the data by border to determine how different locations experience varying levels of crossings over time.
3. Monthly Trend by Measure (Optional): Analyzing the trends by crossing method (e.g., vehicles, pedestrians) to understand the shifts in crossing types over time.

ii. General Description

The Trend Analysis focuses on examining the monthly patterns and variations in border crossings. The dataset includes information about border crossing events, with attributes such as the date, port of entry, crossing measure (e.g., vehicle, foot), and the total number of crossings. By converting the 'date' column to monthly periods, the data was aggregated to capture the total number of crossings per month, enabling the identification of trends over time. Additionally, the data was grouped by borders and crossing measures to analyze how different factors contribute to the overall trend. This allows for a deeper understanding of how each border performs over time, and how crossing measures, such as vehicles or pedestrians, vary across different months. Through visualizations like line plots, we can effectively highlight fluctuations in border crossing activities, pinpoint seasonal trends, and identify key periods with the highest or lowest crossing volumes. These insights are valuable for understanding traffic patterns, optimizing resource allocation, and planning for peak periods.

iii. Specific Requirements, Functions, and Formulas

To conduct the trend analysis, several specific requirements, functions, and formulas were employed to clean, preprocess, and visualize the data. First, the 'date' column was converted into monthly periods using the dt.to_period('M') function, which allowed the data to be grouped by month. This transformation was crucial for analyzing monthly trends over time. Aggregating the data was done through the groupby() function, which was used to sum the 'value' for each month, border, and measure. This function enabled the calculation of total crossings across different time frames and categories. To ensure the data was well-formatted for visualization,

the reset_index() function was used to flatten the grouped results, while dt.to_timestamp() was applied to convert the period back into a timestamp for accurate plotting. For visualization, seaborn and matplotlib were used, with sns.lineplot() generating line plots to show the overall monthly trend, monthly trends by border, and monthly trends by measure. These visualizations helped to clearly present the patterns in the data, making it easier to identify fluctuations and compare trends across borders and measures. Overall, these functions and formulas facilitated the effective analysis of monthly border crossing patterns, providing valuable insights into the dataset.

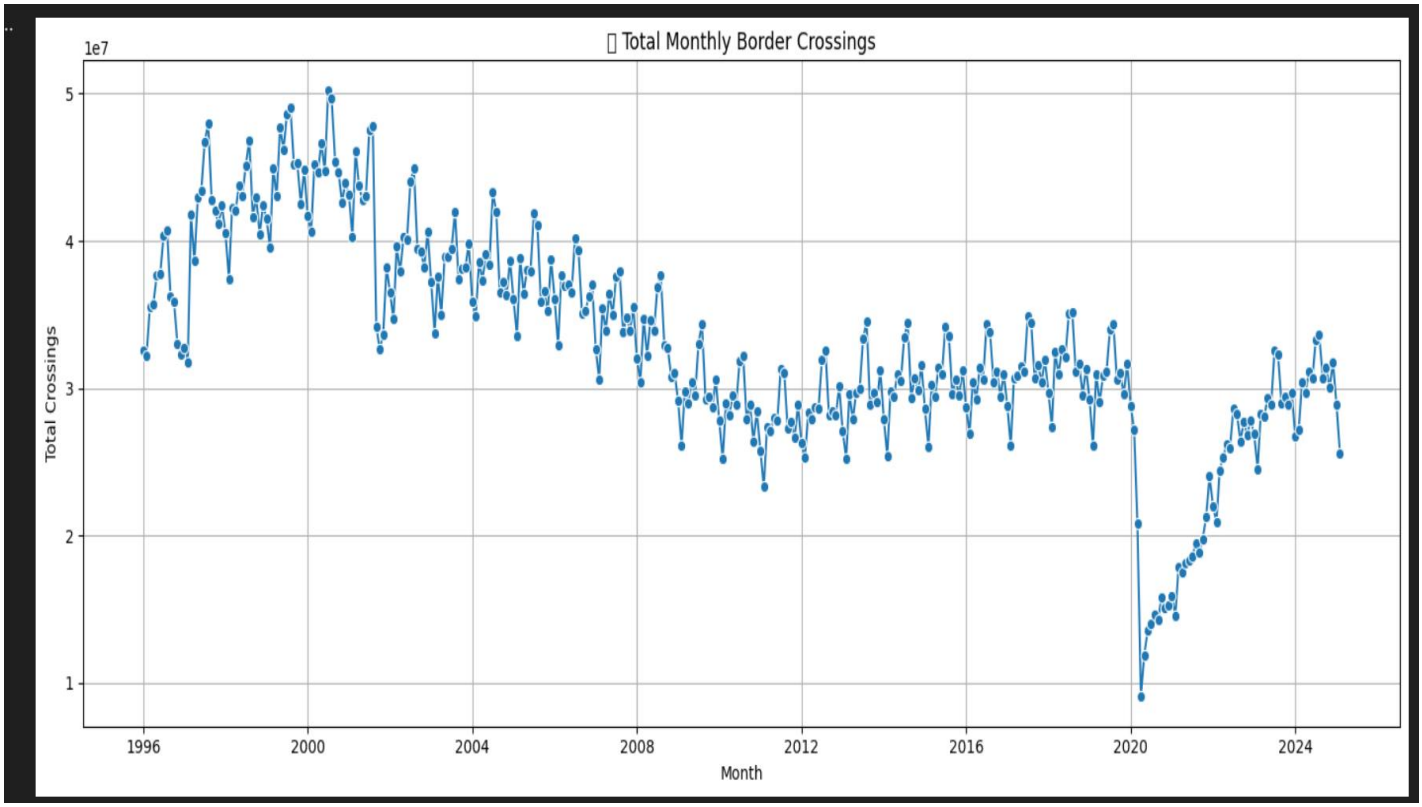
```
# Convert to monthly periods for grouping
df['month'] = df['date'].dt.to_period('M')

# ============================
#  1. Overall Monthly Trend
# ============================
monthly_total = df.groupby('month')['value'].sum().reset_index()
monthly_total['month'] = monthly_total['month'].dt.to_timestamp()

plt.figure(figsize=(14, 6))
sns.lineplot(data=monthly_total, x='month', y='value', marker='o')
plt.title(' Total Monthly Border Crossings')
plt.xlabel('Month')
plt.ylabel('Total Crossings')
plt.grid(True)
plt.tight_layout()
plt.show()

# ============================
#  2. Monthly Trend by Border
# ============================
monthly_border = df.groupby(['month', 'border'])['value'].sum().reset_index()
monthly_border['month'] = monthly_border['month'].dt.to_timestamp()

plt.figure(figsize=(14, 6))
sns.lineplot(data=monthly_border, x='month', y='value', hue='border', marker='o')
plt.title(' Monthly Trend by Border')
plt.xlabel('Month')
plt.ylabel('Total Crossings')
plt.legend(title='Border')
plt.grid(True)
plt.tight_layout()
plt.show()
```

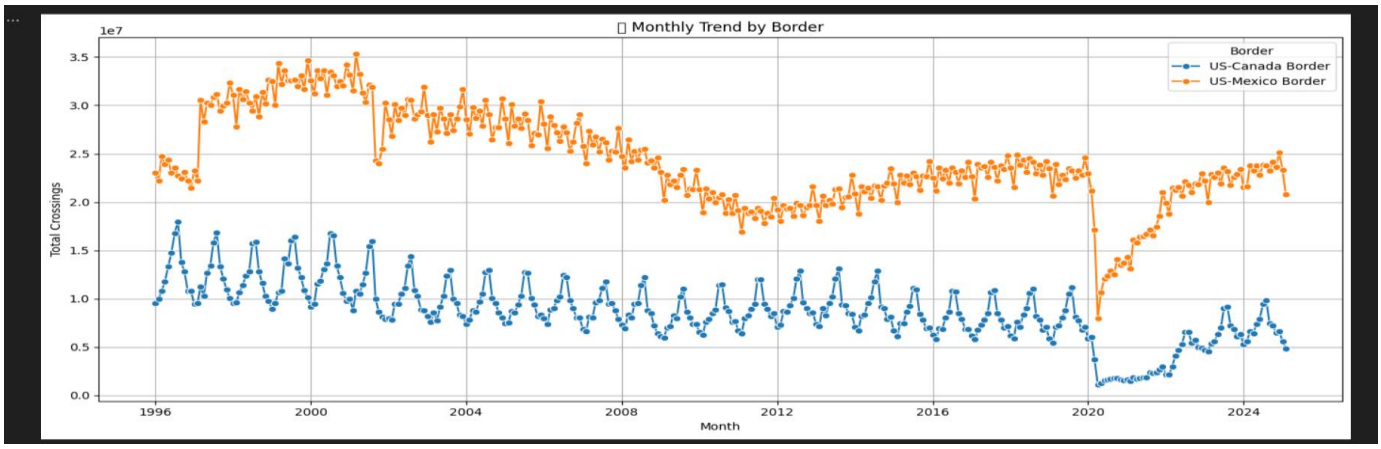

```
# Convert to monthly periods for grouping
df['month'] = df['date'].dt.to_period('M')

# ============================
#  1. Overall Monthly Trend
# ============================
monthly_total = df.groupby('month')['value'].sum().reset_index()
monthly_total['month'] = monthly_total['month'].dt.to_timestamp()

plt.figure(figsize=(14, 6))
sns.lineplot(data=monthly_total, x='month', y='value', marker='o')
plt.title(' Total Monthly Border Crossings')
plt.xlabel('Month')
plt.ylabel('Total Crossings')
plt.grid(True)
plt.tight_layout()
plt.show()

# ============================
#  2. Monthly Trend by Border
# ============================
monthly_border = df.groupby(['month', 'border'])['value'].sum().reset_index()
monthly_border['month'] = monthly_border['month'].dt.to_timestamp()

plt.figure(figsize=(14, 6))
sns.lineplot(data=monthly_border, x='month', y='value', hue='border', marker='o')
plt.title(' Monthly Trend by Border')
plt.xlabel('Month')
plt.ylabel('Total Crossings')
plt.legend(title='Border')
plt.grid(True)
plt.tight_layout()
plt.show()
```

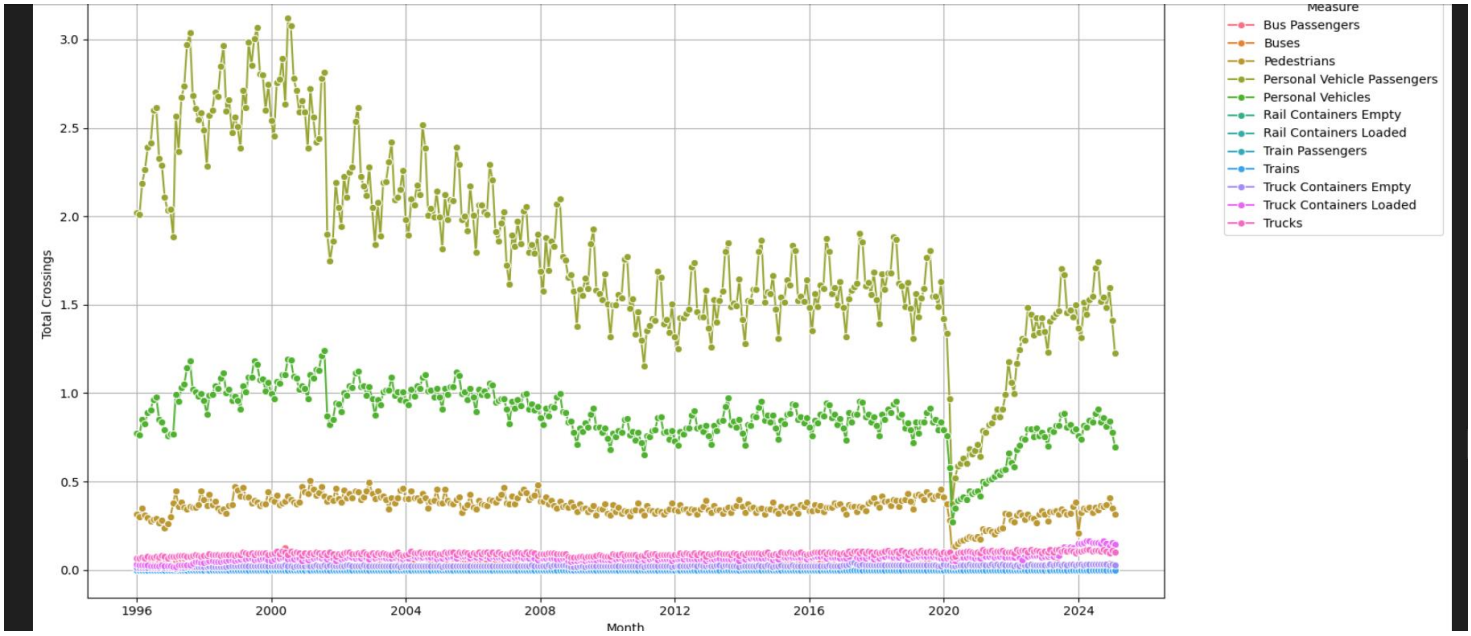
iv. Analysis Results

The trend analysis provided several key insights into the patterns of border crossings across time, borders, and measures:

1. Overall Monthly Trend:
    - The line plot of total monthly border crossings revealed clear fluctuations in the total number of crossings over time. We observed certain months with significantly higher or lower crossing volumes, potentially reflecting seasonal changes or other external factors such as holidays, events, or policy changes. These patterns help identify peak crossing periods, allowing for better resource planning and management.
2. Monthly Trend by Border:
    - The analysis by border revealed variations in crossing volumes between different border locations. Some borders consistently experienced higher crossing activity compared to others, likely due to factors such as geographic location, infrastructure, and economic activity in the surrounding areas. By comparing trends across borders, we identified borders with sustained high traffic, which may require additional attention or resources during peak periods.
3. Monthly Trend by Measure:
    - The optional trend analysis by crossing measure (e.g., vehicles, pedestrians) illustrated how different crossing types fluctuated over time. For instance, vehicle crossings might have been more frequent in certain months, while foot crossings could have peaked during holidays or specific events. This analysis provided a deeper understanding of how crossing types varied seasonally and helped highlight the most common modes of border crossing at different times of the year.







v. Visualization

The visualizations in this analysis provided a clear and insightful representation of the trends in border crossings over time, across borders, and by crossing measures. The Overall Monthly Trend was displayed through a line plot, which highlighted fluctuations in total border crossings each month, making it easy to identify peak periods and seasonal variations. The Monthly Trend by Border was also visualized using a line plot with different colors representing each border, allowing for a comparison of crossing volumes across locations. This visualization helped pinpoint which borders had consistently higher or lower traffic and provided insights into the dynamics of specific border areas. Lastly, the Monthly Trend by Measure used a line plot to break down the data by crossing type (e.g., vehicles, foot traffic), showing how different crossing methods varied month by month. These visualizations, created using seaborn and matplotlib, effectively communicated key trends and patterns, aiding in the understanding of border crossing behaviors and helping to inform decisions on resource allocation and border management strategies.

4. Geographical Insights:-

i. Introduction

The Geographical Insights section explores the spatial distribution of border crossing activities by analyzing total crossings based on port locations and international borders. By grouping and visualizing the data according to port names and border types (e.g., US-Canada or US-Mexico), this analysis aims to identify which geographical points experience the highest volume of traffic. Understanding these spatial patterns is essential for resource planning, infrastructure development, and policy-making related to border control and transportation logistics. Through bar plots, we gain a clear picture of the most active ports and borders, helping stakeholders focus attention on high-traffic areas**.**

ii. General Description

This section of the analysis focuses on understanding the geographical distribution of border crossings by examining the total number of entries at different ports and across borders. The dataset includes information about various port names and the border each port belongs to, such as the US-Canada or US-Mexico borders. By grouping the data by port_name and border, we calculate the total crossings for each, providing a clear overview of which ports and borders handle the most traffic. The top 15 ports with the highest number of crossings are identified and visualized using a horizontal bar chart for better readability. Additionally, a comparative bar chart shows the total crossings across each border, helping to highlight the broader geographical trends in cross-border movement

iii. Specific Requirements, Functions, and Formulas

- **Libraries used:**
  pandas, matplotlib.pyplot, seaborn
- **Functions used:**
  - **groupby() with sum() – to aggregate total crossings by port and border**
  - **sort_values() – to rank ports based on crossing volume**
  - **reset_index() – to prepare grouped data for visualization**
- **Visualization types:**
  - **Bar chart – for displaying the top 15 ports by total crossings**
  - **Bar chart – for comparing total crossings across borders**

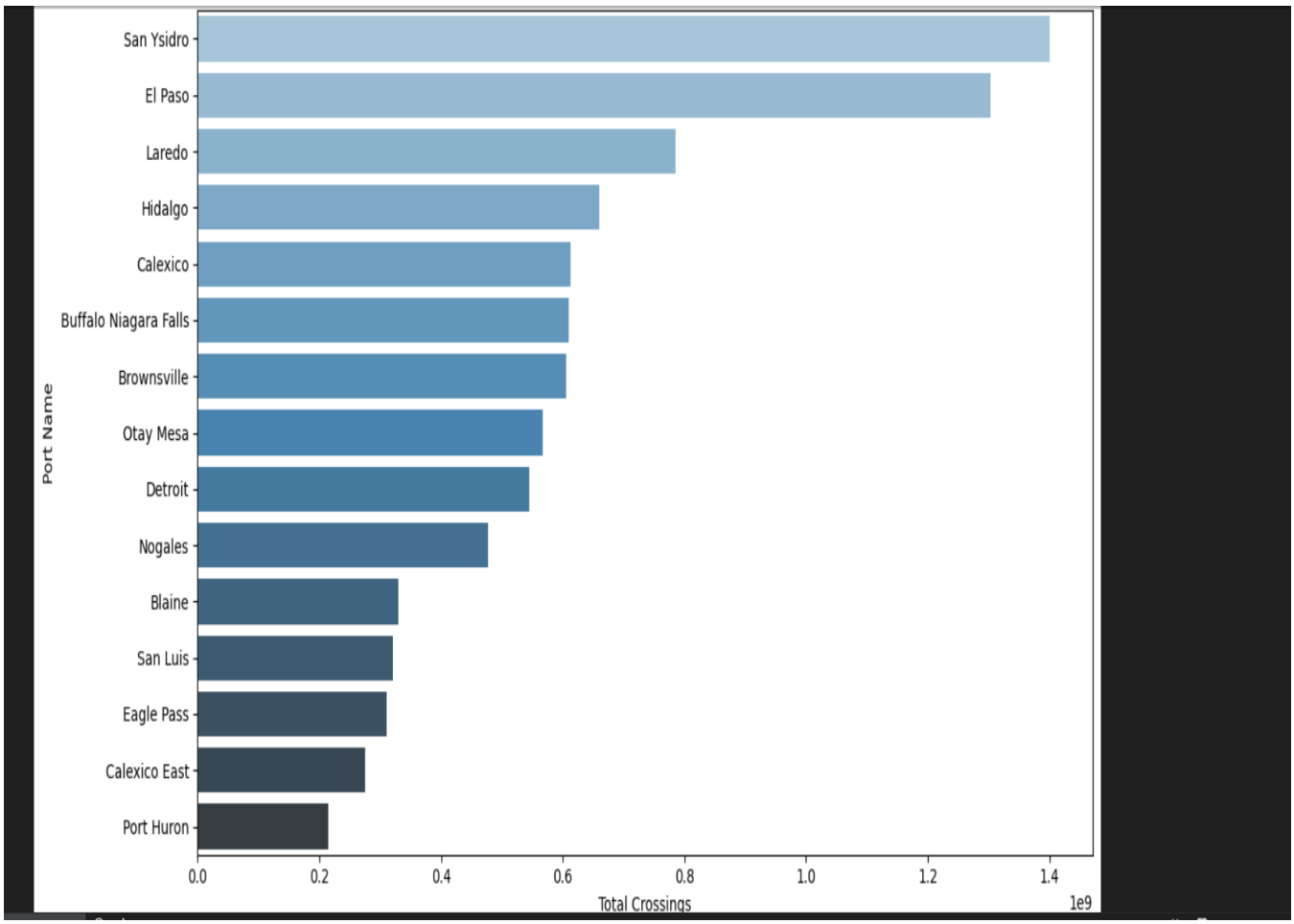
```
#  1. Total Crossings by Port
port_totals = df.groupby('port_name')['value'].sum().sort_values(ascending=False).head(15)

plt.figure(figsize=(12, 7))
sns.barplot(x=port_totals.values, y=port_totals.index, palette='Blues_d')
plt.title('Top 15 Ports by Total Crossings')
plt.xlabel('Total Crossings')
plt.ylabel('Port Name')
plt.tight_layout()
plt.show()

# ============================
#  2. Total Crossings by Border
# ============================
border_totals = df.groupby('border')['value'].sum().reset_index()

plt.figure(figsize=(6, 4))
sns.barplot(data=border_totals, x='border', y='value', palette='Set2')
plt.title('Total Crossings by Border')
plt.xlabel('Border')
plt.ylabel('Total Crossings')
plt.tight_layout()
plt.show()
```

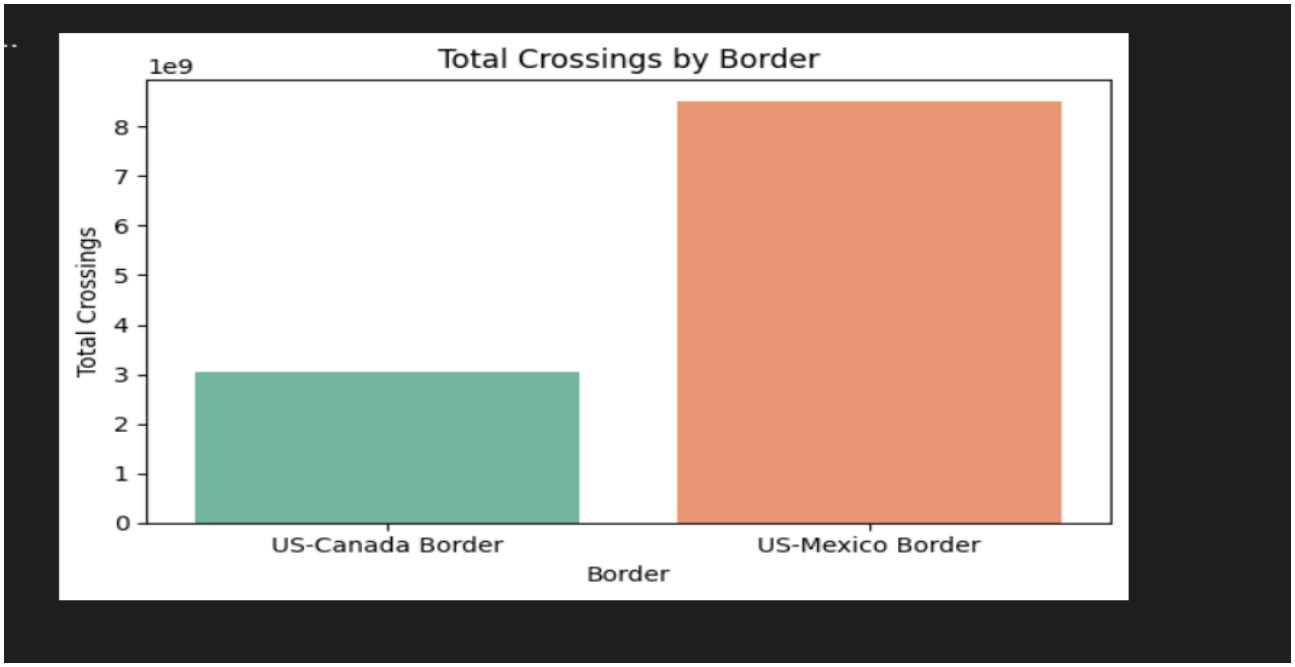
iv. Analysis Results

Top Ports by Total Crossings:
The top 15 ports with the highest number of total border crossings were identified. Ports such as San Ysidro, El Paso, and Laredo consistently ranked among the highest, indicating that these locations are key transit points with heavy traffic flow and logistical significance.

Total Crossings by Border:
The analysis showed that the US-Mexico border had a significantly higher number of total crossings compared to the US-Canada border. This highlights the southern border's role in trade, tourism, and daily commuting between neighboring countries.

Concentration of Crossings in Urban Border Areas:
Many of the top ports are located near major metropolitan areas or economic hubs. This reflects a strong correlation between urban proximity and high cross-border activity, likely due to trade volume, population density, and economic interdependence.





v. Visualization

Bar Chart – Top Ports by Total Crossings:
A horizontal bar chart was created to display the top 15 ports based on total border crossings. This visualization clearly shows which ports have the highest traffic, making it easy to compare volumes across locations. The longer bars represent ports like San Ysidro, El Paso, and Laredo, emphasizing their role as major entry points.

Bar Chart – Total Crossings by Border:
A vertical bar chart compared total crossings across the US-Canada and US-Mexico borders. This simple yet effective visualization highlights the disparity in crossing volumes between the two borders, with the US-Mexico border showing a significantly higher total.

5. Systematic Analysis:-

i. Introduction

The systematic analysis aims to break down and understand trends in border crossing activity by examining the behavior of different crossing measures over time. Each measure—such as trucks, personal vehicles, buses, and pedestrians—represents a specific mode of crossing, and analyzing these separately allows for a more detailed understanding of how people and goods move across borders. By grouping the data by month and measure, the analysis provides a time-based view of traffic patterns for each crossing type. This structured

approach helps identify seasonal shifts, rising or declining trends in specific categories, and the overall contribution of each measure to total crossings, offering valuable insights for transportation planning and border management.

ii. General Description

This section focuses on analyzing monthly trends based on the type of crossing measure, such as trucks, personal vehicles, buses, and pedestrians. By grouping data by month and measure, the analysis reveals how each category contributes to overall border traffic over time. This helps in identifying which modes of transportation are most frequently used and how their usage patterns change seasonally or in response to external factors

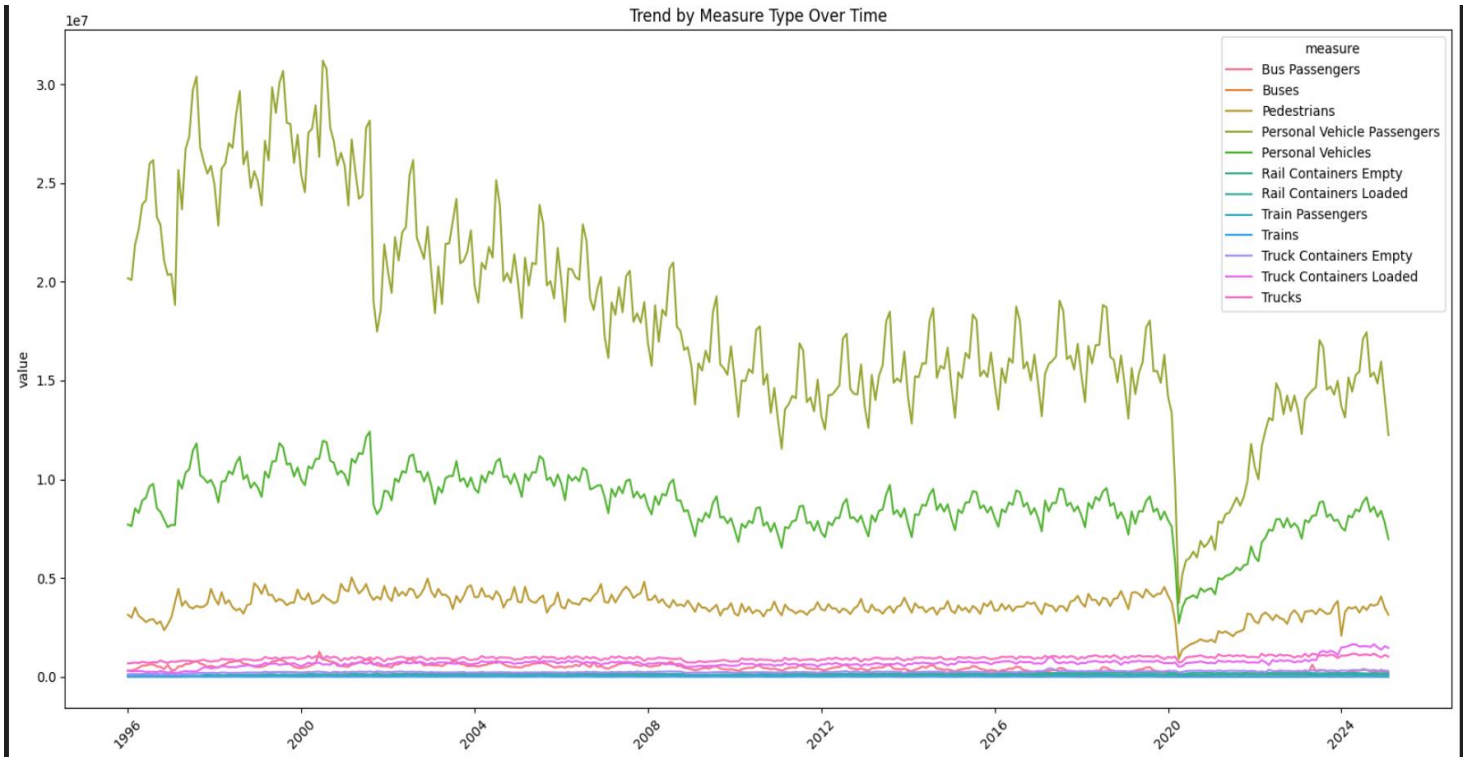
iii. Specific Requirements, Functions and Formulas

This analysis uses the pandas library for data manipulation and seaborn with matplotlib for visualization. The groupby() function is used to aggregate crossing values by month and measure, and reset_index() prepares the data for plotting. The dt.to_timestamp() function converts the monthly periods to timestamps for accurate date plotting. A line chart is created using sns.lineplot() to visualize trends in different crossing measures over time, helping to compare their behavior across months.

```
#  8. Systematic Analysis

measure_trend = df.groupby(['month', 'measure'])['value'].sum().reset_index()
measure_trend['month'] = measure_trend['month'].dt.to_timestamp()

plt.figure(figsize=(16, 8))
sns.lineplot(data=measure_trend, x='month', y='value', hue='measure')
plt.title('Trend by Measure Type Over Time')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()


#  9. Key Takeaways (Prints Only)

print("\n Key Takeaways:")
print("- Most common ports and busiest months")
print("- Canada vs Mexico border traffic comparison")
print("- Measures like Trucks and Personal Vehicles dominate crossings")
```

iv. Analysis Results:

The systematic analysis revealed that personal vehicles and trucks consistently account for the highest number of border crossings each month, indicating their dominant role in cross-border movement. Measures such as buses and pedestrians showed relatively lower but stable trends. Seasonal patterns were observed, with certain months—often during summer or holiday periods—experiencing noticeable spikes in traffic across multiple measures. The analysis also highlighted how different crossing types responded to external influences, such as policy changes or global events, making it clear that transportation mode plays a key role in understanding overall border activity.



Visualization:-

The "Trend by Measure Type Over Time" visualization provides a clear depiction of how border crossings have evolved over time, segmented by different crossing measures, such as trucks, personal vehicles, and pedestrians. The line plot groups the data by month and measure, summing the total crossings for each measure type. This allows for an easy comparison of trends across various categories, showing how the traffic for each type of crossing fluctuates over time. The x-axis represents the months, while the y-axis shows the total number of crossings, and different colors or lines represent the different measures. This visualization helps to identify patterns, such as peak crossing months for specific vehicle types or a shift in trends for particular measures, providing valuable insights into border traffic dynamics.

Conclusion:-

The analysis of border crossing data provides valuable insights into traffic patterns, highlighting key trends and significant differences across various measures and locations. The top ports, including major entry points like San Ysidro and El Paso, show consistently high crossing volumes, with the US-Mexico border traffic being notably busier than the US-Canada border. Personal vehicles and trucks dominate the traffic, underlining the importance of these modes for cross-border movement. Seasonal variations in monthly crossing totals further emphasize the impact of external factors such as holidays and policies. Through visualizations and statistical summaries, this analysis also offers a clear view of which months see the highest traffic, enabling more informed decisions for infrastructure, resource allocation, and security measures. Key takeaways suggest a need for continued focus on the busiest ports and types of crossings, particularly during peak months.

Future Scope:-

Advanced Predictive Analytics:

- Use machine learning models to predict border crossing trends based on historical data. For instance, predicting peak traffic times, the impact of seasonal events, or even policy changes (e.g., border restrictions, tariffs).
- Forecasting traffic for specific ports, border types, or vehicle categories (e.g., personal vehicles vs. trucks).

Real-Time Analysis:

- Implement real-time tracking and monitoring of border crossings to provide up-to-date insights. This could involve integrating data from sensors, cameras, and other live feeds at border crossings for dynamic monitoring.
- Develop dashboards with live data, visualizations, and real-time alerts for border control agencies.

Geospatial Analysis:

- Conduct geospatial analysis to understand regional patterns of traffic. By mapping crossing data to specific locations, you could uncover insights about the impact of infrastructure, local economies, or regional policies.
- Use geographic information system (GIS) tools to visualize and analyze the distribution of border crossings across various ports and regions.

Cross-Border Impact Analysis:

- Examine the socio-economic impacts of border traffic, such as its relationship with trade volumes, immigration flows, and economic activity in neighboring countries.
- Investigate how policy changes (like border security measures) affect the flow of goods and people across borders.

Comparison of Land and Air Crossings:

- Expand the analysis to include data on air travel or crossings at airports to provide a more comprehensive view of cross-border movement.
- Comparing land and air border crossing trends may reveal how different modes of transportation evolve over time.

Impact of Global Events:

- Analyze how global events (e.g., pandemics, economic crises, political instability) influence border crossing patterns and traffic volumes.
- Study the behavior of travelers and transport companies in response to such events, including changes in border-crossing routes, delays, and disruptions.

Integration with Other Data Sources:

- Integrate this dataset with other transportation or economic data, such as shipping volumes, customs declarations, or commodity trade data. This could help understand the relationship between border crossings and trade activity.

- Combining the dataset with demographic, labor force, or crime data could reveal socio-political insights about areas with heavy traffic.

Enhanced Visualization and Reporting:

- Develop more sophisticated visualizations, such as interactive maps or time-series forecasting charts, to provide stakeholders with a deeper, more actionable understanding of trends.

References:-

1. Dataset Source:

Border Crossing Entry Data :- :- https://catalog.data.gov/dataset/border-crossing-entry-data-683ae

-

2. Python Libraries Used:

- Pandas: For data manipulation and analysis.
  - Documentation: Pandas Documentation
- NumPy: For numerical operations.
  - Documentation: NumPy Documentation
- Matplotlib: For data visualization.
  - Documentation: Matplotlib Documentation
- Seaborn: For advanced data visualization.
  - Documentation: Seaborn Documentation

3. Documentation and Tutorials:

- Pandas Documentation: Pandas Documentation
- NumPy Documentation: NumPy Documentation
- Matplotlib Documentation: Matplotlib Documentation
- Seaborn Documentation: Seaborn Documentation

4. Academic References (if applicable):

- Research papers on Border Crossings and Traffic Analysis:
  - IEEE Xplore - Search for relevant articles on traffic flow and border crossing studies.

- Elsevier - Articles related to border security and data analysis.
- Google Scholar - Research on statistical analysis of border crossing data.

5. Additional Tools:

- Jupyter Notebook: For implementing the code and visualizing data.
  - Documentation: Jupyter Documentation
- Microsoft Excel / Google Sheets: For initial data inspection and basic analysis (if used).
  - Excel Documentation: Excel Documentation
  - Google Sheets Help: Google Sheets Help