Final project – milestone 2

Ayush Anand

NORTHEASTERN UNIVERSITY:
COLLEGE OF PROFESSIONAL STUDIES
ALY 6010: PROBABILITY THEORY AND INTRODUCTORY STATISTICS
PROFESSOR XYZ
OCTOBER 17TH, 2024

Report for Milestone 2: Hypothesis Testing and Analysis

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**.

Objective:

The aim of this milestone was to apply inferential statistics and hypothesis testing to analyze border crossing data, focusing on key questions related to the US-Mexico and US-Canada borders. The tests performed included one-sample and two-sample tests, supported by data visualization for comparison and interpretation.

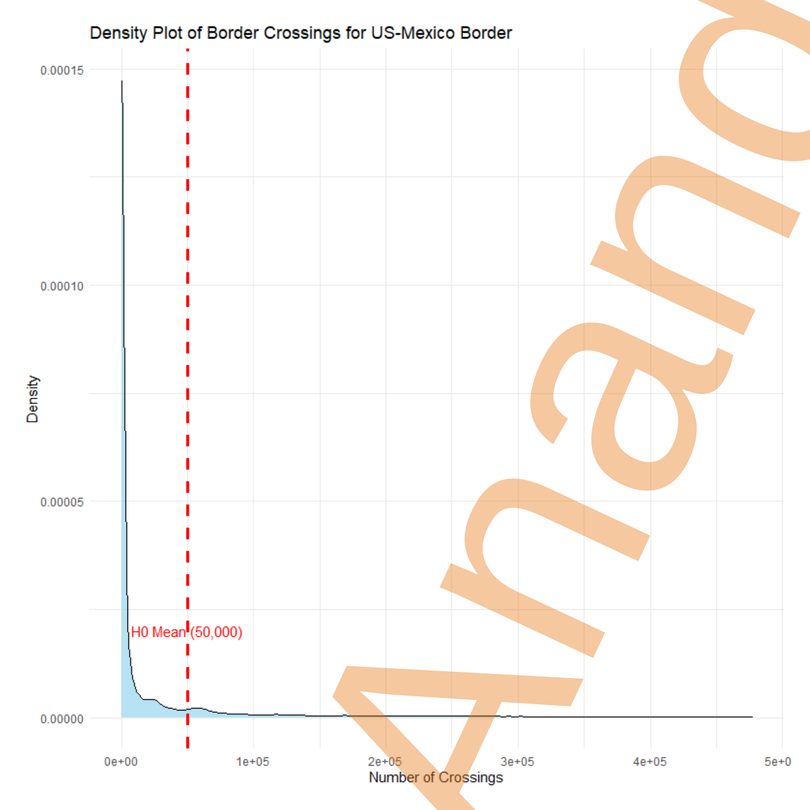
Data Overview:

The dataset used for this analysis contains data on border crossings at various US borders, filtered to focus specifically on the US-Mexico and US-Canada borders. The key variable of interest is the number of crossings, represented as "Value" in the dataset. For the purposes of hypothesis testing, subsets of the data were extracted for each border.

Questions and Hypotheses:

1. One-Sample t-Test (US-Mexico Border Crossings):

- **Question**: Is the average number of border crossings at the US-Mexico border significantly different from 50,000?



Hypothesis Testing Steps:

1. **Define Hypotheses:**
   - **Null Hypothesis (H_0):** The mean number of border crossings at the US-Mexico border is greater than or equal to 50,000. (H_0: $\mu \geq 50{,}000$)
   - **Alternative Hypothesis (H_1):** The mean number of border crossings at the US-Mexico border is less than 50,000. (H_1: $\mu < 50{,}000$)

Test Execution:

- A one-sample t-test was performed on the US-Mexico border crossing data (us_mexico_data$Value). The test compared the sample mean to the hypothesized mean of 50,000 crossings.

Results:

- **t-value:** -54.685
- **Degrees of Freedom (df):** 87,688
- **p-value:** < 2.2e-16
- **Sample Mean:** 34,847.48
- **95% Confidence Interval:** (-∞, 35,303.25)
- The 95% confidence interval suggests that the true mean of crossings lies below 35,303.25.

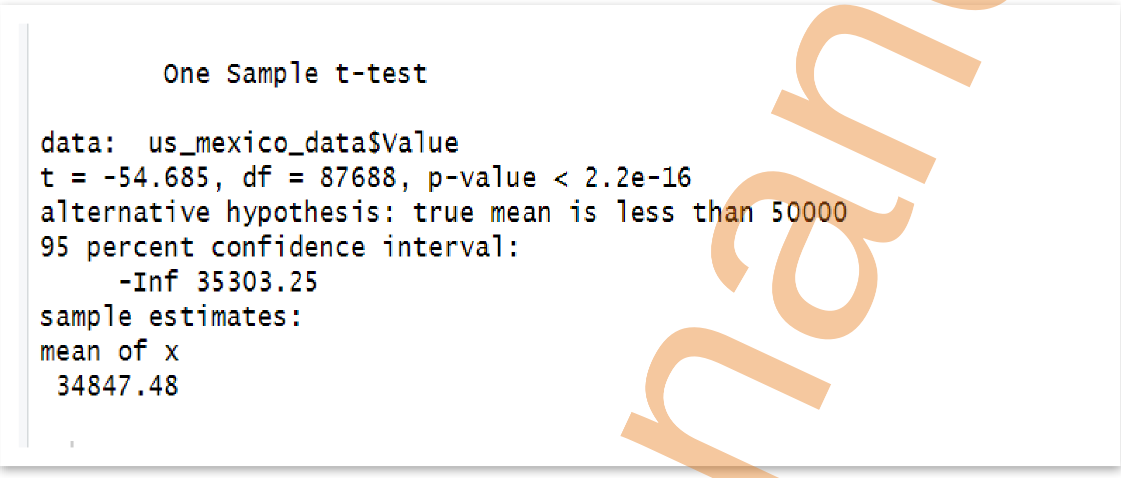
Critical Value and Decision Rule:

- **Critical Value:** 1.644871 (for a one-tailed test with α = 0.05)

- Since the p-value is effectively zero, which is smaller than the significance level (α = 0.05), we reject the null hypothesis. The test statistic is negative (-54.685), clearly showing that the observed mean is significantly lower than the hypothesized mean of 50,000.

Conclusion:

- Based on the test statistic (-54.685) and the p-value (0), we reject the null hypothesis (H_0). This result indicates that the average number of crossings at the US-Mexico border is significantly less than 50,000.

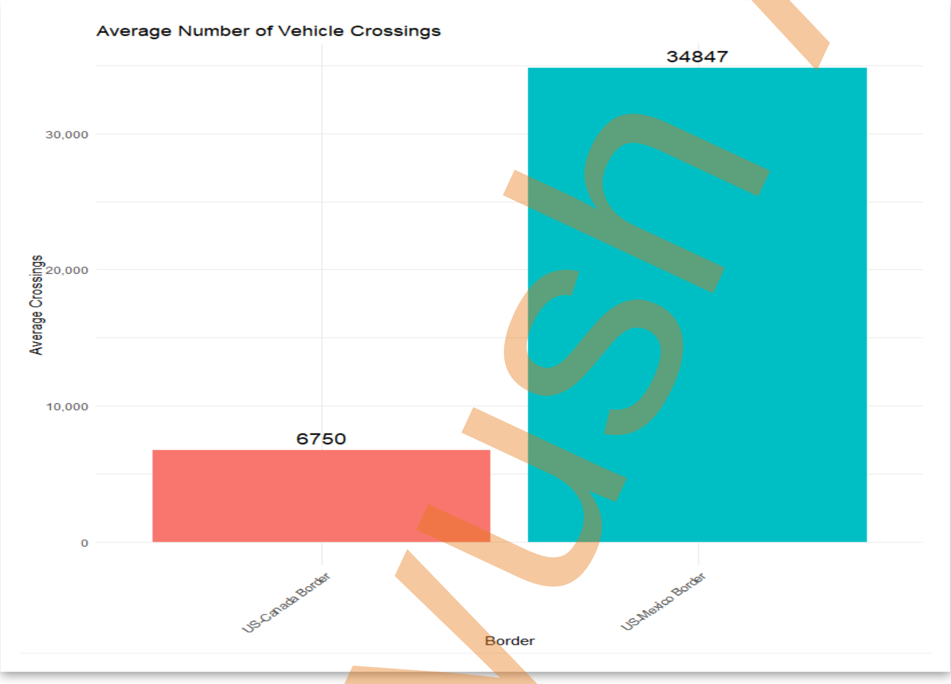


```
        One Sample t-test

data:  us_mexico_data$Value
t = -54.685, df = 87688, p-value < 2.2e-16
alternative hypothesis: true mean is less than 50000
95 percent confidence interval:
     -Inf 35303.25
sample estimates:
mean of x
 34847.48
```

2. Two-Sample Test (Comparison between US-Mexico and US-Canada Borders):

- **Question**: Is there a significant difference in the average number of vehicle crossings between the US-Mexico and US-Canada borders?



Hypotheses:

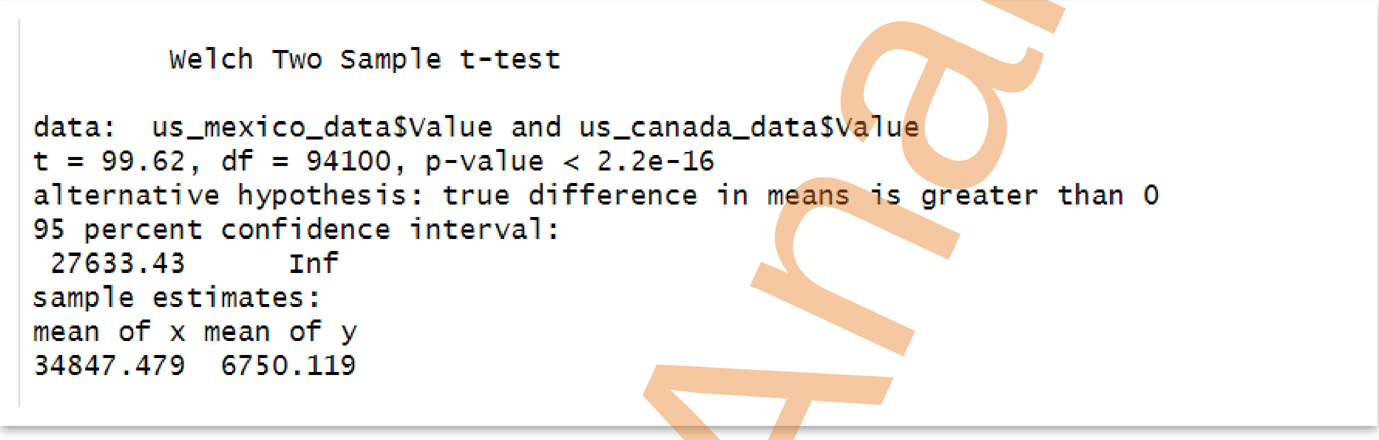
Null Hypothesis (H_0): The average number of vehicle crossings from the US-Mexico border is less than or equal to that of the US-Canada border. (H_0: μ_US-Mexico ≤ μ_US-Canada)

Alternative Hypothesis (H_1): The average number of vehicle crossings from the US-Mexico border is greater than that of the US-Canada border. (H_1: μ_US-Mexico > μ_US-Canada)

Two-Sample Test (US-Mexico vs US-Canada Borders)

- **Test Conducted**: A two-sample t-test was performed to test whether the average number of vehicle crossings at the US-Mexico border is significantly greater than the average number of vehicle crossings at the US-Canada border.
- **Results**:
    - **Test Statistic**: 99.62
    - **P-value**: < 2.2e-16
    - **Critical Value**: 1.64487
    - Based on the p-value and the significance level (alpha = 0.05), we rejected the null hypothesis.

Conclusion: The two-sample t-test results show that there is significant evidence to suggest that the average number of crossings at the US-Mexico border is greater than that of the US-Canada border.



```
        Welch Two Sample t-test

data:  us_mexico_data$Value and us_canada_data$Value
t = 99.62, df = 94100, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 27633.43      Inf
sample estimates:
mean of x mean of y 
34847.479  6750.119 
```

Report Summary

The **US-Mexico border shows significantly lower crossings** compared to the hypothesized value of 50,000 in the one-sample t-test. The comparison between the US-Mexico and US-Canada borders revealed that the **US-Mexico border has significantly higher crossings** than the US-Canada border based on the two-sample t-test.

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

`