Statistical Analysis of U.S. Border Crossings Data
IE 6200 - Engineering Probability & Statistics

Project Report

Introduction

The border crossing ports are ports of entry for land modes along the US-Canada and US-Mexico borders and the Bureau of Transportation Statistics (BTS) is responsible for maintaining the summary statistics for these inbound crossings. The BTS was created in 1992 and being part of the United States Department of Transportation (DOT), analyzes, compiles, collects information on intermodal transportations in all areas, and improves the quality and effectiveness of the DOT's statistical programs through research, development of guidelines and promotion of improvements in data acquisitions and usage.

Objective

We intend to analyze the trends and present meaningful statistical insights about how the number of vehicles, passengers, containers, pedestrians entering the United States via the US-Canada Border and US-Mexico Border have changed over the past two decades.

Data Description

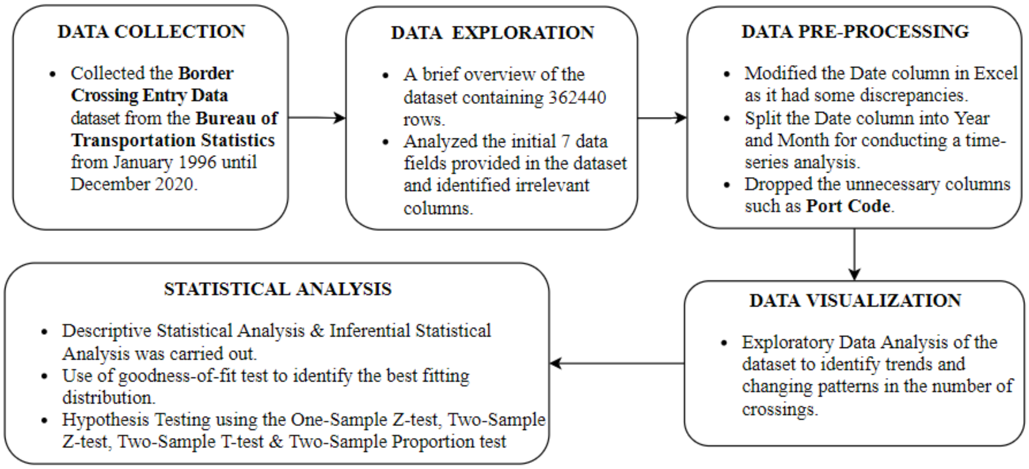
The primary source of data used for our analysis has been taken from The Bureau of Transportation Statistics, an official US government organization that is responsible for managing data pertaining to the various forms of transit in the United States. The dataset we have selected is the Border Crossing Entry Dataset, which contains port level data about inbound crossings into the United States from the two adjacent countries, namely Canada and Mexico from January 1996 to December 2020. It comprises entries of passengers and different modes of transportation such as personal vehicles, trucks, buses, containers, trains. Our dataset contains the following variables :

1. Port Name: The port of entry into the US
2. State: The state in which the port is stationed
3. Port Code: Code to identify the port
4. Border: US-Canada Border and US-Mexico Border
5. Date: Date of entry into the United States(custom split into Month & Year)
6. Measure: Mode of transportation such as trucks, buses, rail, personal vehicles,etc
7. Value: Total count of inbound entries of a particular measure for a specific port

Data Processing

After collecting the dataset from The Bureau of Transportation Statistics, we examined the dataset to remove extraneous columns and the flow of implementation has been explained in

Fig.1.The format of the Date column in the dataset was erroneous, and hence was modified using Excel and then loaded in R. The Date column was further split into Month & Year columns to identify change in statistics of crossings with respect to time. Thereafter, exploratory data analysis was conducted followed by Statistical Analysis. The Null Hypothesis & Alternate Hypothesis were formulated for drawing inferences based on the one-sample and two-sample Z-test, two-sample t-test and two-sample proportion test.



DATA COLLECTION

- Collected the **Border Crossing Entry Data** dataset from the **Bureau of Transportation Statistics** from January 1996 until December 2020.

DATA EXPLORATION

- A brief overview of the dataset containing 362440 rows.
- Analyzed the initial 7 data fields provided in the dataset and identified irrelevant columns.

DATA PRE-PROCESSING

- Modified the Date column in Excel as it had some discrepancies.
- Split the Date column into Year and Month for conducting a time-series analysis.
- Dropped the unnecessary columns such as **Port Code**.

DATA VISUALIZATION

- Exploratory Data Analysis of the dataset to identify trends and changing patterns in the number of crossings.

STATISTICAL ANALYSIS

- Descriptive Statistical Analysis & Inferential Statistical Analysis was carried out.
- Use of goodness-of-fit test to identify the best fitting distribution.
- Hypothesis Testing using the One-Sample Z-test, Two-Sample Z-test, Two-Sample T-test & Two-Sample Proportion test

Fig.1. : Data Processing Steps performed on the Border Crossings Entry Dataset

1. Descriptive Statistical Analysis

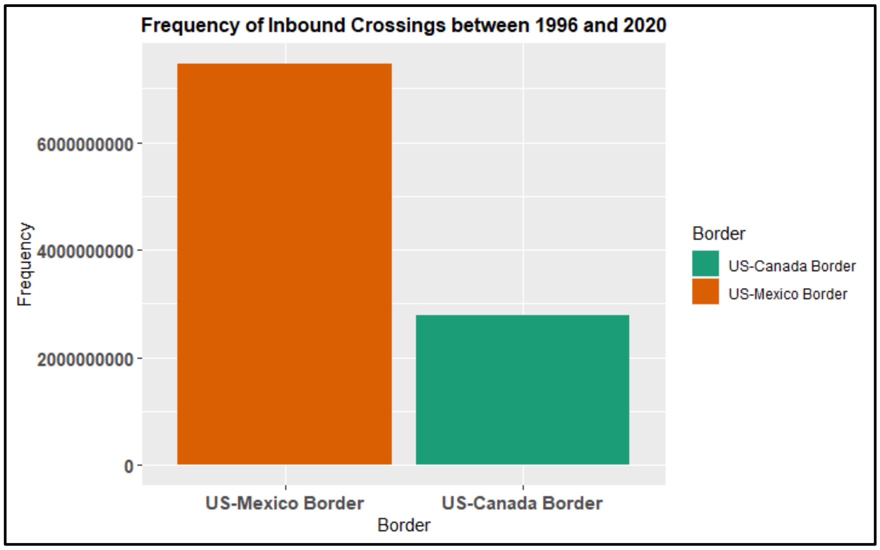


Fig.2. : Frequency of Inbound Crossings between 1996 & 2020

Fig.2. depicts the comparison between the total number of border entry crossings from 1996 to 2020 between the US-Mexico Border and the US-Canada Border. From this visualization, we infer that of the total entry crossings into the US, the US-Mexico Border has recorded around 7.5 Billion inbound crossings since January 1996, while the US- Canada Border has recorded around 2.8 Billion inbound crossings since January 1996, thereby indicating that the Us-Mexico Border is approximately 2.5 times more busier than the US-Canada Border.

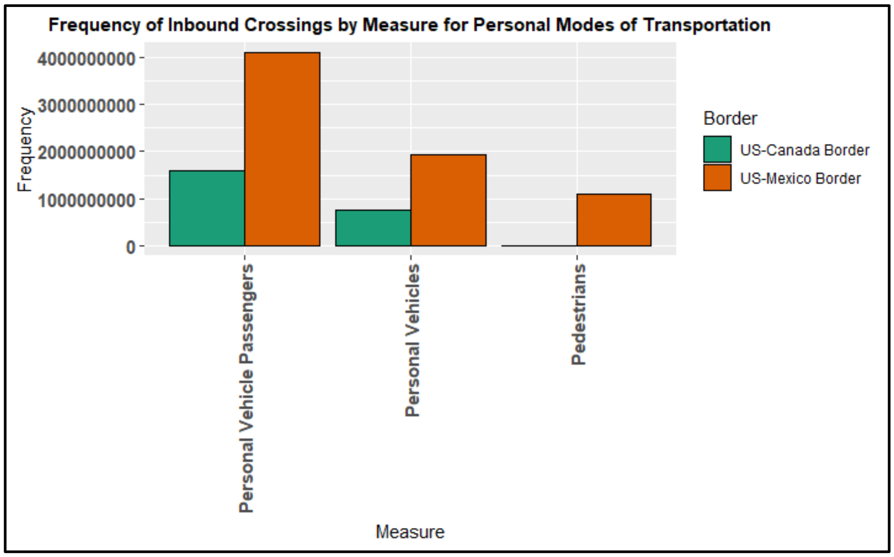


Fig.3. : Frequency of Inbound Crossings of Personal Modes of Transportation

From Fig.3., we observe how non-business related people are entering the US, via personal vehicles or on foot (pedestrians). The data again reaffirms the fact that the US-Mexico Border handles more traffic than the US-Canada Border. It is also visible that the US-Canada border barely has any significant pedestrian entries when compared to the US-Mexico border.

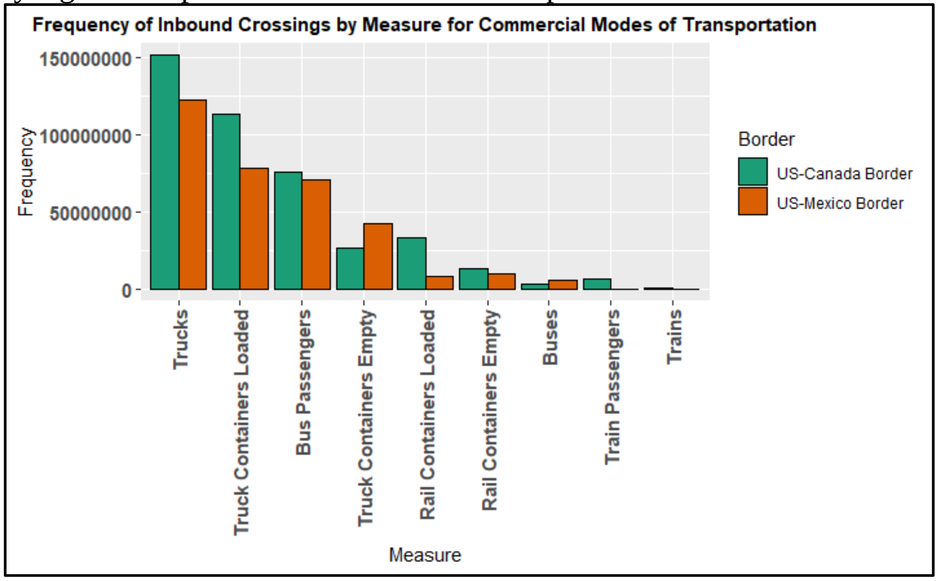


Fig.4. : Frequency of Inbound Crossings by Measure for Commercial Vehicles

The visualization depicted in Fig.4., indicates all the types of Commercial Vehicles that have been seen to enter the US at the two borders. It is visible that trucks and loaded truck containers have the most frequency followed by bus passengers and loaded rail containers. We can infer from Fig.2 & Fig.3., that business related crossings are a lot more in number than personal related crossings.

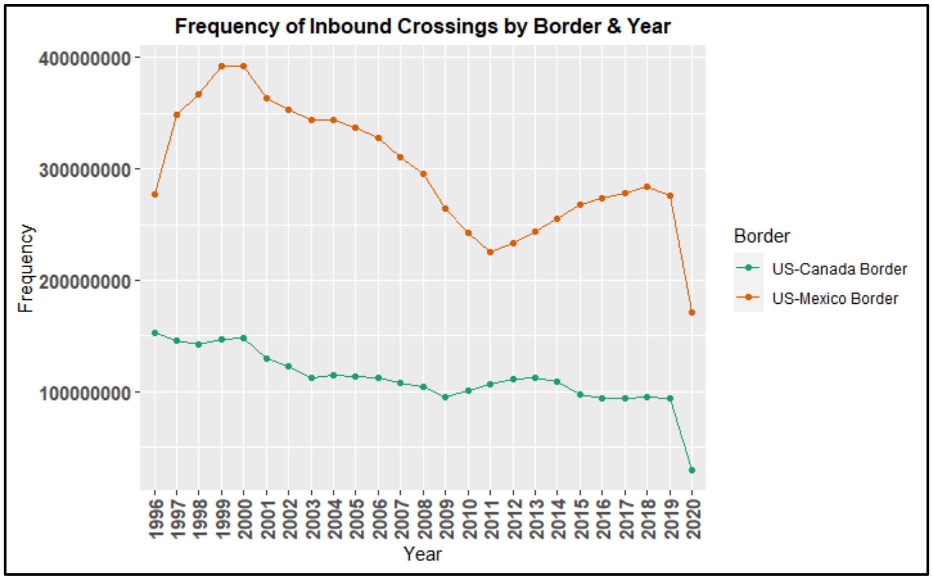


Fig.5. : Frequency of Inbound Crossings by Border & Year

The visual representation in Fig.5.,demonstrates an year-on-year comparison of the varying trends in number of inbound crossings at the US-Canada Border & the US-Mexico Border. The US-Mexico border shows a steep decline trend in crossings from 2000 until 2011, and then a sharp rise in inbound crossings from 2012 until 2019. On the other hand, the US- Canada Border shows an inconsistent pattern in the number of inbound crossings. However, both indicate a significant dip in crossings in 2020, as expected due to the COVID-19 pandemic.

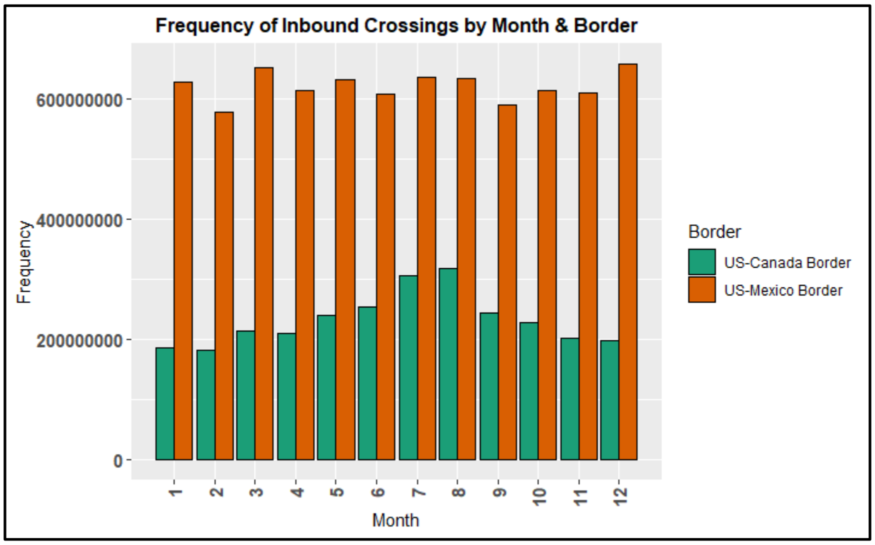


Fig.6. : Frequency of Inbound Crossings by Border & Month

Fig.6. indicates the change in pattern of the frequency of inbound crossings at both the Borders from January until December over the years. It is interesting to see that the US-Canada Border is mostly busy during July & August, while the US-Mexico Border seems the busiest during March & December.

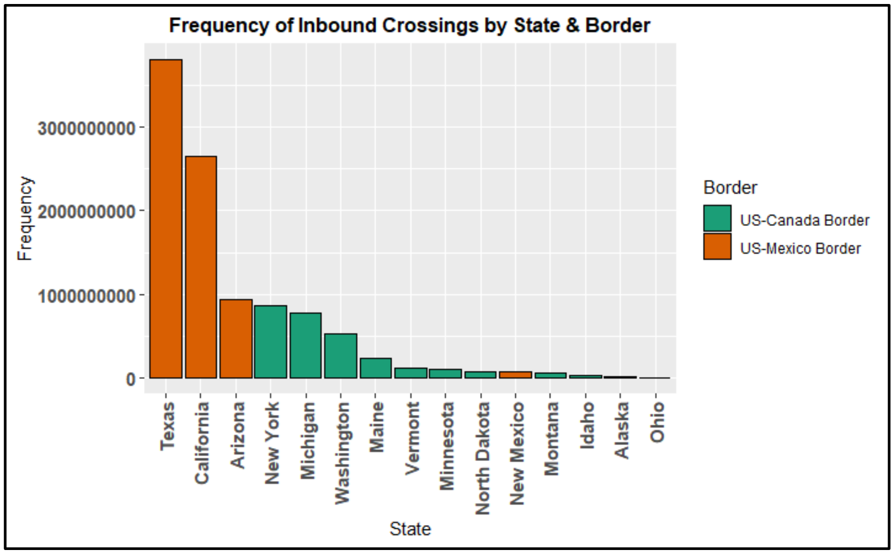


Fig.7. : Frequency of Inbound Crossings by State & Border

From Fig.7, we can infer that the top 3 busiest states in terms of inbound border crossings at the US-Mexico Border are Texas, California and Arizona; followed by the top 3 busiest states at the US-Canada Border being New York, Michigan & Washington.

2. Inferential Statistical Analysis

2.1 Conditional Probability

a. We want to find the probability that a truck entering the US will go to El Paso port given that it is crossing the US-Mexico border.

A = event that a truck has entered visa the US-Mexico Border
B = event that a truck goes to El Paso

The conditional probability of the subset B given the event A is defined as

$$\boxed{(B|A) = P(A\cap B)\ /P(A)}$$

Therefore, we found that the probability of a truck entering the US and going to the port of ***El Paso***, given that it is crossing the US-Mexico border is **0.06331491.** In other words, there is approximately a **6.3%** chance of a truck entering the US and going to the port of ***El Paso***, given that it is crossing the US-Mexico border.

b. We want to find the probability that a person has entered the US for a non-business trip via the US-Canada Border, given that the person is a Pedestrian.

A = event that a person has entered the US on a non-business trip(Bus, Pedestrian, etc.)
B = event that a person is a Pedestrian amongst all non-business persons

The conditional probability of the subset B given the event A is defined as

$$\boxed{(B|A) = P(A\cap B)\ /P(A)}$$

Therefore, we found that the probability that a person has entered the US for a non-business trip via the US-Canada Border, given that the person is a Pedestrian is **0.01218567.** In other words, there is approximately a **1.2%** chance that a person has entered the US for a non-business trip via the US-Canada Border, given that the person is a Pedestrian.

2.2 Fitting Distribution

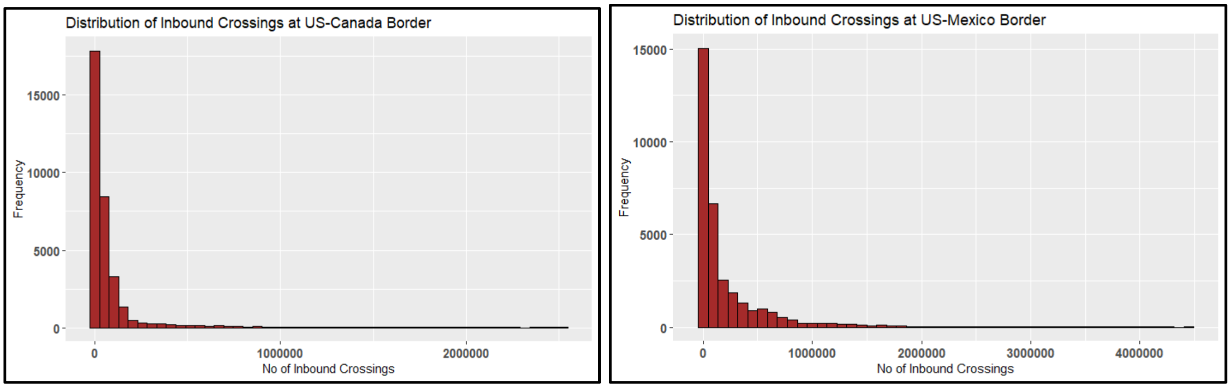
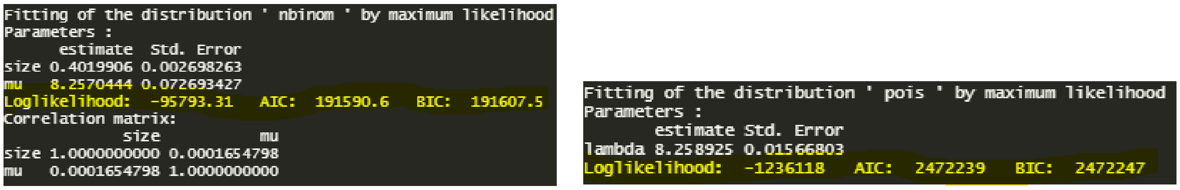


Fig.9.: Right-skewed distribution of Inbound Crossings at both Borders

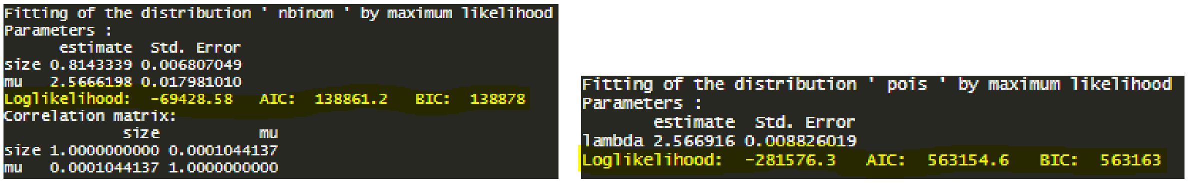
From Fig.9., we can see that the number of inbound crossings at both the Borders is a discrete variable, and also is not normally distributed, i.e., the distribution is skewed in nature. The histogram in both the figures shows the tail of the distribution is stretched towards the right, thus indicating a right-skewed distribution.



```
Fitting of the distribution ' nbinom ' by maximum likelihood
Parameters :
       estimate  Std. Error
size 0.4019906 0.002698263
mu   8.2570444 0.072693427
Loglikelihood:  -95793.31   AIC:  191590.6   BIC:  191607.5
Correlation matrix:
                size           mu
size 1.0000000000 0.0001654798
mu   0.0001654798 1.0000000000
```

```
Fitting of the distribution ' pois ' by maximum likelihood
Parameters :
         estimate Std. Error
lambda 8.258925 0.01566803
Loglikelihood:  -1236118   AIC:  2472239   BIC:  2472247
```

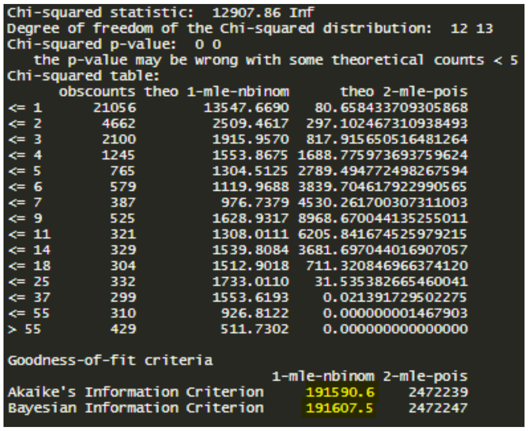
Fig.10.a.: Fitting Negative Binomial Distribution & Poisson Distribution for US-Canada Border



```
Fitting of the distribution ' nbinom ' by maximum likelihood
Parameters :
       estimate  Std. Error
size 0.8143339 0.006807049
mu   2.5666198 0.017981010
Loglikelihood:  -69428.58   AIC:  138861.2   BIC:  138878
Correlation matrix:
                size           mu
size 1.0000000000 0.0001044137
mu   0.0001044137 1.0000000000
```

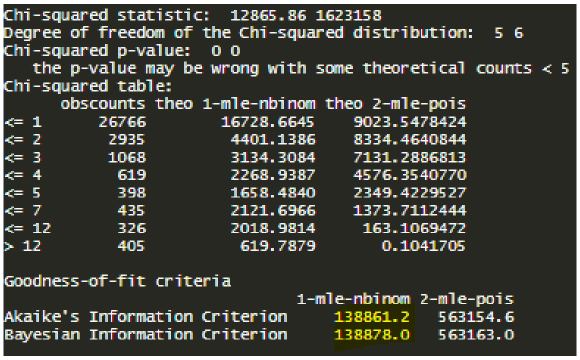
```
Fitting of the distribution ' pois ' by maximum likelihood
Parameters :
         estimate  Std. Error
lambda 2.566916 0.008826019
Loglikelihood:  -281576.3   AIC:  563154.6   BIC:  563163
```

Fig.10.b.: Fitting Negative Binomial Distribution & Poisson Distribution for US-Mexico Border



```
Chi-squared statistic:  12907.86 Inf
Degree of freedom of the Chi-squared distribution:  12 13
Chi-squared p-value:  0 0
   the p-value may be wrong with some theoretical counts < 5
Chi-squared table:
      obscounts theo 1-mle-nbinom      theo 2-mle-pois
<= 1      21056        13547.6690   80.658433709305868
<= 2       4662         2509.4617  297.102467310938493
<= 3       2100         1915.9570  817.915650516481264
<= 4       1245         1553.8675 1688.775973693759624
<= 5        765         1304.5125 2789.494772498267594
<= 6        579         1119.9688 3839.704617922990565
<= 7        387          976.7379 4530.261700307311003
<= 9        525         1628.9317 8968.670044135255011
<= 11       321         1308.0111 6205.841674525979215
<= 14       329         1539.8084 3681.697044016907057
<= 18       304         1512.9018  711.320846966374120
<= 25       332         1733.0110   31.535382665460041
<= 37       299         1553.6193    0.021391729502275
<= 55       310          926.8122    0.000000001467903
> 55        429          511.7302    0.000000000000000

Goodness-of-fit criteria
                               1-mle-nbinom 2-mle-pois
Akaike's Information Criterion     191590.6    2472239
Bayesian Information Criterion     191607.5    2472247
```



```
Chi-squared statistic:  12865.86 1623158
Degree of freedom of the Chi-squared distribution:  5 6
Chi-squared p-value:  0 0
   the p-value may be wrong with some theoretical counts < 5
Chi-squared table:
      obscounts theo 1-mle-nbinom theo 2-mle-pois
<= 1      26766        16728.6645    9023.5478424
<= 2       2935         4401.1386    8334.4640844
<= 3       1068         3134.3084    7131.2886813
<= 4        619         2268.9387    4576.3540770
<= 5        398         1658.4840    2349.4229527
<= 7        435         2121.6966    1373.7112444
<= 12       326         2018.9814     163.1069472
> 12        405          619.7879       0.1041705

Goodness-of-fit criteria
                               1-mle-nbinom 2-mle-pois
Akaike's Information Criterion     138861.2   563154.6
Bayesian Information Criterion     138878.0   563163.0
```

Fig.10.c.: Using goodness-of-fit test to evaluate fitness of Binomial Distribution & Poisson Distribution

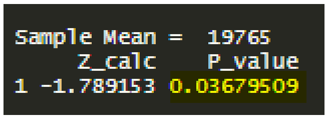
From Fig.10.a.,10.b., and 10.c.,the lower AIC and BIC values and higher Log Likelihood of Negative Binomial Distribution compared to these parameters obtained via Poisson Distribution indicate that it is evident that Negative Binomial Distribution is the best fit for our data.

2.3 Hypothesis Testing

a. **One sample Z-test** : Assuming the entire dataset is the population, we want to test if the mean value of inbound crossings of our sample data is equivalent to the population mean value of inbound crossings.
Calculating the Population Mean (µ) = **28302.55**
X ≡ R.V. of number of inbound crossings in the U.S.

Null Hypothesis -> H0: µ = 28302.55
Alternate Hypothesis -> H1: µ ≠ 28302.55

Conclusion:



```
Sample Mean =  19765
      Z_calc    P_value
1 -1.789153 0.03679509
```

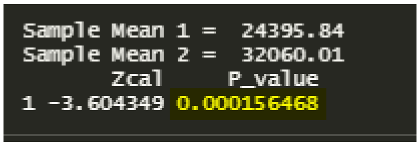
Fig.11.: Results obtained from One-Sample Z-test

From Fig.11, we can see that as **P_value = 0.03679509 < 0.05,** we reject the null hypothesis and thus, conclude that there is a significant difference between the sample mean value of inbound crossings and population mean value of inbound crossings.

b. **Two sample Z-test** : Segregating the dataset into two different populations , we want to test if the mean value of inbound crossings of two different sets of samples taken from two different populations is the same or not.
X1 ≡ R.V. of inbound crossings in the U.S. from first sample
X2 ≡ R.V. of inbound crossings in the U.S. from second sample

Null Hypothesis -> $H0: \mu1 - \mu2 = 0$
Alternate Hypothesis -> $H1: \mu1 - \mu2 \neq 0$

Conclusion:



```
Sample Mean 1 =  24395.84
Sample Mean 2 =  32060.01
        Zcal     P_value
1 -3.604349 0.000156468
```

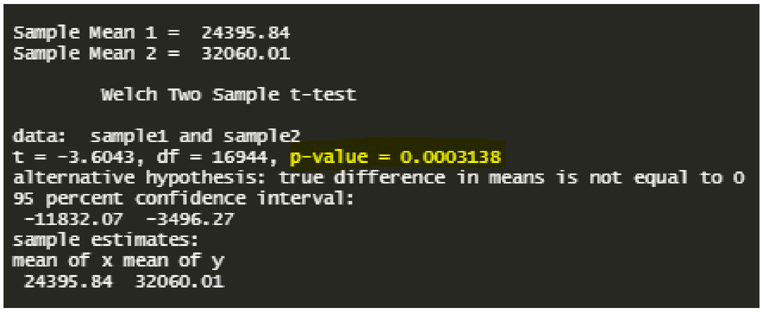
Fig.12.: Results obtained from Two-Sample Z-test

From Fig.12., as **P_value = 0.000156468 < 0.05,** we reject the null hypothesis and thus conclude there is a significant difference between the sample mean value of inbound crossings of the two samples.

c. **Two sample t-test :** We want to test the significance between means of two samples means if populations are independent and the variances are unknown, i.e., to test the equality of the mean value of inbound crossings of two random samples generated from the border crossing data.
X1 ≡ R.V. of inbound crossings in the U.S. from first sample
X2 ≡ R.V. of inbound crossings in the U.S. from second sample

Null Hypothesis -> $H0: \mu1 - \mu2 = 0$
Alternate Hypothesis -> $H1: \mu1 - \mu2 \neq 0$

Conclusion:



```
Sample Mean 1 =  24395.84
Sample Mean 2 =  32060.01

        Welch Two Sample t-test

data:  sample1 and sample2
t = -3.6043, df = 16944, p-value = 0.0003138
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11832.07  -3496.27
sample estimates:
mean of x mean of y
 24395.84  32060.01
```

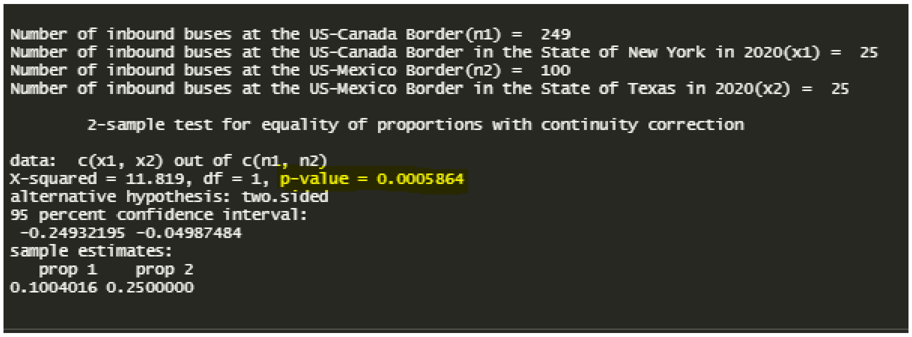
Fig.13.: Results obtained from Two-Sample t-test

From Fig.13., as **P_value = 0.0003138 < 0.05,** we reject the null hypothesis and conclude there is a significant difference between the sample mean value of inbound crossings of the two samples.

d. **Two-sample proportion test** : Filtering out the dataset to test the equality of proportions of inbound bus crossings in the State of New York at the US-Canada Border & in the State of Texas at the US-Mexico Border in 2020.
X1 ≡ R.V. of number of inbound bus crossings at US-Canada Border
X2 ≡ R.V. of number of inbound bus crossings at US-Mexico Border

Null Hypothesis -> $H0: p1 - p2 = 0$
Alternate Hypothesis -> $H1: p1 - p2 \neq 0$

Conclusion:



```
Number of inbound buses at the US-Canada Border(n1) =  249
Number of inbound buses at the US-Canada Border in the State of New York in 2020(x1) =  25
Number of inbound buses at the US-Mexico Border(n2) =  100
Number of inbound buses at the US-Mexico Border in the State of Texas in 2020(x2) =  25

        2-sample test for equality of proportions with continuity correction

data:  c(x1, x2) out of c(n1, n2)
X-squared = 11.819, df = 1, p-value = 0.0005864
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.24932195 -0.04987484
sample estimates:
   prop 1    prop 2
0.1004016 0.2500000
```

Fig.14.: Results obtained from the Two-sample proportion test

From Fig. 14, as **P_value = 0.0005864 < 0.05**, we reject the null hypothesis and thus conclude there is a significant difference between the proportions of inbound bus crossings in New York at US-Canada Border & in Texas at US-Mexico Border in 2020.

Limitations

The dataset consists of entries that are disjoint in nature. For example, a truck that has entered the US cannot make a border entry at two ports at the same time or a pedestrian entering the US cannot enter via both the borders at the same time. Hence, due to the lack of overlap between two given variables(columns), it limits our ability to calculate joint probability distributions.

Conclusion

From the studies, we observe that the US-Mexico border handles significantly more traffic than the US-Canada border. It is observed that the most frequent measure to cross the border was trucks. A trend can also be observed that there has been a steady decline in the number of crossings for the US-Canada border. The results inferred from the analysis show that the most number of crossings occur at the borders of the state Texas, and California respectively. Detailed statistical inferences have been drawn in the inferential statistics section. The report displays the unique characteristics of the two borders.

References

[1] https://data.bts.gov/Research-and-Statistics/Border-Crossing-Entry-Data/keg4-3bc2