Midterm Report

Tianhui Zhao

2025-03-13

github repo link: https://github.com/TeresasaZ/JSC370-Project

Introduction

Urban traffic congestion is a critical issue in large metropolitan areas, impacting commuting efficiency, road safety, and environmental sustainability. Fuel prices are often considered a factor influencing driving behavior, as changes in gasoline costs can affect how frequently individuals use personal vehicles. This study examines how fluctuations in fuel prices influence traffic volume in Toronto. Since driving is a necessity for many individuals, especially in large countries like Canada where distances between destinations can be significant, an important question arises: **Would changes in fuel prices still affect the amount of car usage, given its essential nature?**

For this analysis, two datasets were used:

1. Traffic Volume Data: This dataset was extracted from the Toronto Open Data Portal, providing detailed records of traffic volumes at various locations throughout the city, measured at 15-minute intervals.
2. Fuel Price Data: The fuel price dataset was retrieved from Open Canada, containing monthly fuel price information for major Canadian cities, including Toronto.

The research question is:

How does fuel price affect traffic volume in Toronto?

Methods

Data Acquisition

Traffic volume data was obtained via API extraction from the Toronto Open Data CKAN repository. The API requests returned CSV files containing detailed traffic counts, specific 15 min interval times, specific street locations, longitude and latitude. (2015-2019 traffic data and 2019-2024 traffic data were obtained separately).

https://open.toronto.ca/dataset/traffic-volumes-midblock-vehicle-speed-volume-and-classification-counts/

Fuel price data was sourced from Open Canada as a direct CSV download. The dataset includes date (monthly), fuel price of several major cities including Toronto, and whether the price is untaxed, taxed or total.

https://open.canada.ca/data/en/dataset/3ff1e1de-d665-4398-a12a-e8ce55f887ac/resource/af4537fc-1bf3-4b3a-b093-ebc633c2a21a

Data Cleaning

Filtering Data: Traffic data from 2015–2024 was filtered to include only records from 2018 onward, ensuring alignment with fuel price records, which begin in 2018.

Then, I combined the two traffic volume datasets to get a complete 2018-2024 traffic volume dataset.

Handling Missing Values: Both traffic and pricing datasets were checked for missing values, empty strings, and "NULL" values. There were no missing values, so no need to make changes.

Unique location names and direction from traffic dataset, and unique tax status from pricing dataset were also checked to ensure there were no string placeholders exist. Unique values from these string variables are normal. No placeholders exist.

Data Exploration

Check for import issues: I checked for import issues using dim(), head(), tail(), str() and summary() for both datasets. There were no issues by checking these tables. Among the key variables, time start is of Date type, direction and tax status are of character type, and volume and fuel price are numeric. From summary(), I also made sure that there were no errors in numerical variables for both datasets.

Data Wrangling

Traffic volume dataset only has time-related variable as `time_start` and `time_end`, which are two variables representing each 15-min interval observation. These two times are too specific as they include year, month, day, hour, minute and second. However, the pricing dataset only has a `Date` variable that represents the observation of fuel price per month (which are all on the first day of each month). Due to this situation, I created a `month` variable in each dataset and merged them by month. `month` is extracted from the dates or time from the two datasets.

Although pricing data was observed monthly, observing traffic data in the same way would result in too few data points. So, I created a `actual_date` variable to extract the actual dates (year-month-day) traffic volumes were observed so I can observe daily data.

By examining the number of unique locations per date and other details of the merged dataset using functions like `str()` and `unique()`, I found that each 15-minute interval includes different observation locations. These locations may overlap or vary, and the number of observed locations fluctuates. Additionally, within a single 15-minute interval, a street can be monitored from multiple directions. As a result, the number of observations per time interval varies due to differences in both the observed locations and the recorded directions. Therefore, I chose to analyze the average daily traffic volume. This approach ensures a sufficient number of data points while also accounting for discrepancies in the number of observations per time interval.

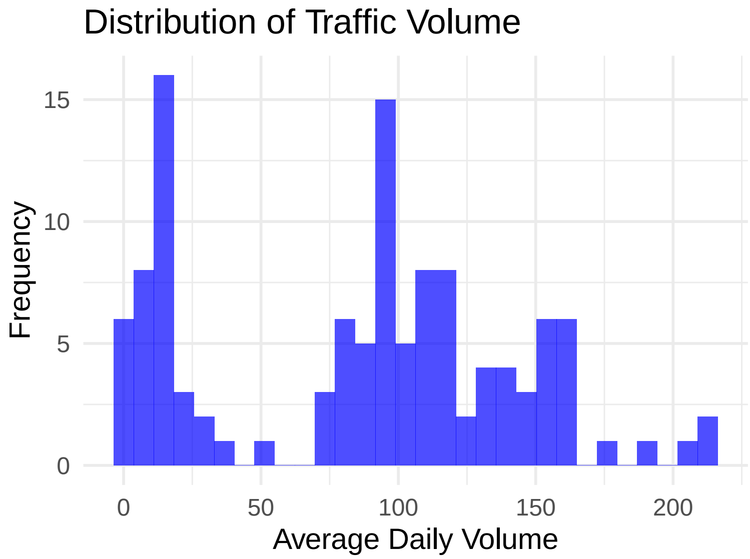
Before analyzing the dataset, I factored the three string variables: location name, direction and tax status.

Lastly, it was unnecessary to retain all three fuel price categories—untaxed, taxed, and total. Therefore, I kept only the Total price, as it provides the most comprehensive representation of fuel costs.

Preliminary Results

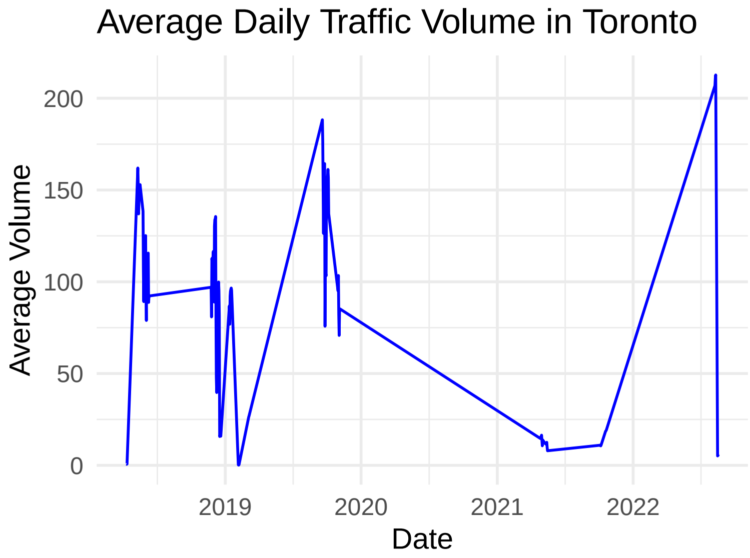
1.

This histogram provides an overview of the spread of average daily traffic volume. The distribution of traffic volume is bimodal, with one peak at around 10 and another peak around 100. Traffic volume observations range from 0 to approcimately 220.



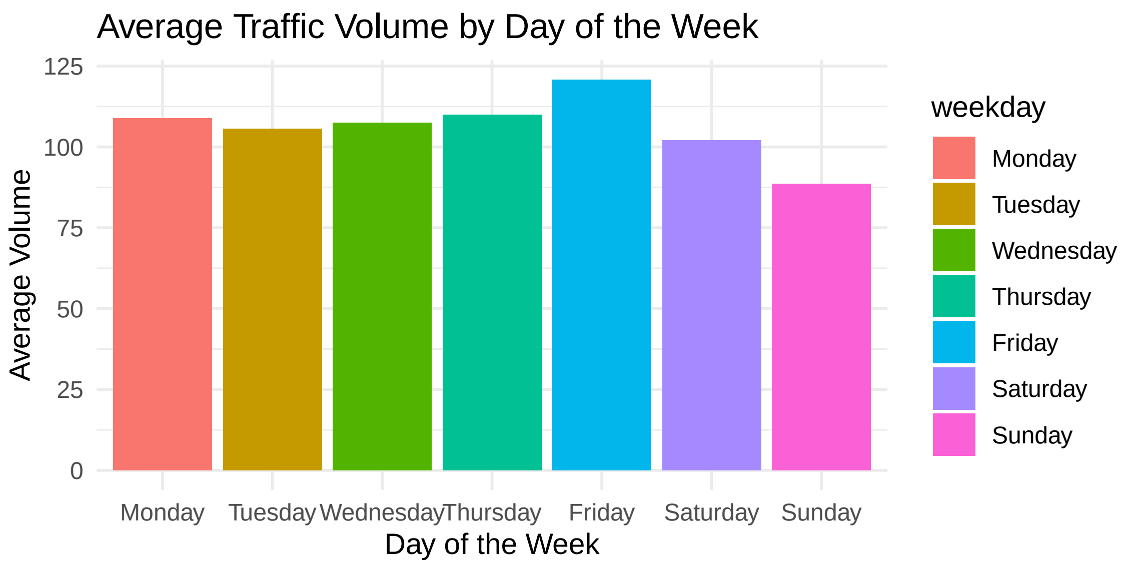
2.

This figure shows how traffic volume in Toronto changed from 2018 to 2024. The graph appears rough due to gaps in data collection, but clear patterns emerge. Before 2019, traffic volume fluctuates with peaks around 150. A sharp decline occurs between 2020 and 2022, likely due to pandemic lockdowns reducing vehicle movement. After 2022, traffic volume rises sharply, suggesting recovery as restrictions ease. Despite missing observations, the trend indicates a pandemic-induced drop followed by a return to pre-pandemic levels.



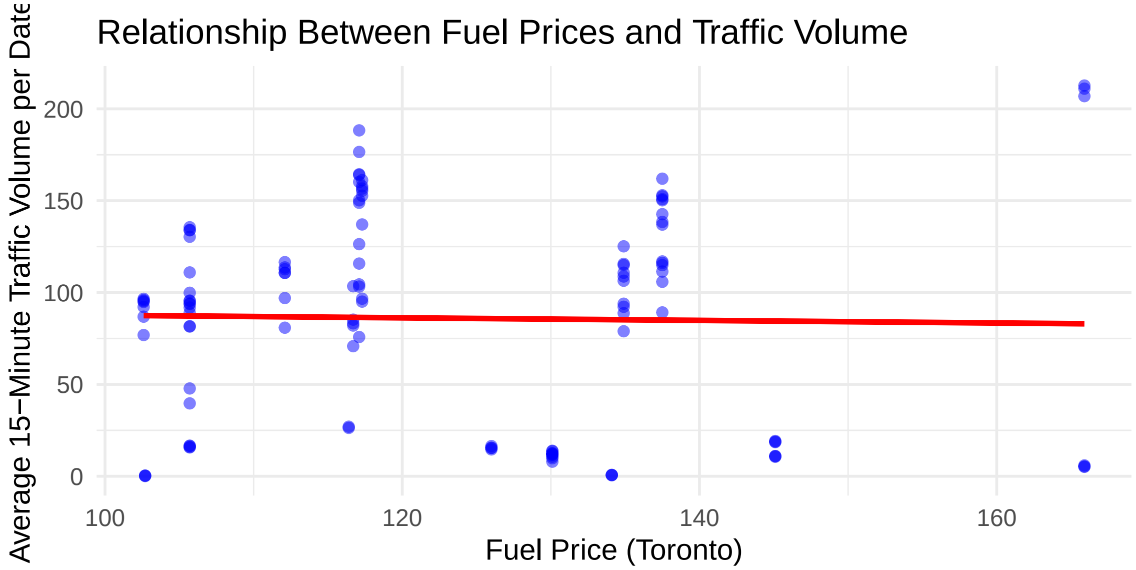
3.

This figure illustrates the average traffic volume for each day of the week, highlighting differences between weekday and weekend traffic. Traffic remains relatively stable from Monday to Friday, peaking on Friday. However, a noticeable decline occurs on weekends, with Sunday experiencing the lowest traffic volume. This suggests reduced commuting activity on weekends compared to weekdays.



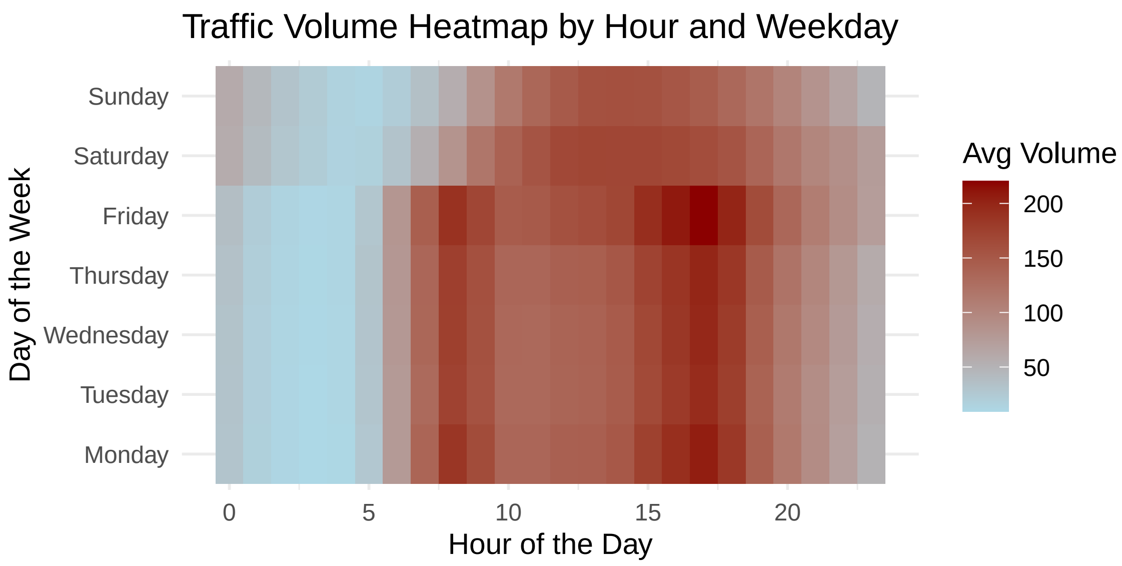
4.

This scatterplot examines the relationship between fuel prices and daily average traffic volume directly, with a fitted trend line. The data points show no clear pattern, especially no potential trend indicating that traffic volume significantly decrease as fuel prices rise. Even at the highest observed fuel price (~160), traffic volume can still peak around 200. The nearly horizontal trend line further suggests a weak or negligible correlation between fuel prices and traffic volume.



5.

This heatmap illustrates traffic volume variations by hour and day of the week. On weekdays, traffic is lowest before 6 AM and after 10 PM, with two peaks around 8 AM and 5 PM, aligning with typical commuting patterns. On weekends, traffic starts increasing later, around 8 AM, and declines after 7 PM. The lower traffic volume during weekday commuting hours on weekends suggests reduced commuting demand.



6.

The correlation value of -0.0200 suggests an extremely weak negative relationship between fuel prices and average traffic volume, implying that traffic volume is unlikely to be significantly affected by fuel price fluctuations. However, when fuel prices are categorized as high or low based on the median and analyzed using a t-test, the p-value of 0.0152 (below the 0.05 threshold) indicates statistically significant evidence that traffic volume differs between the two pricing categories.

Table 1: Summary of Statistical Tests: Correlation and T-test Results

Statistic	Value
Pearson Correlation	-0.0200
T-test p-value	0.0152

Summary

The correlation analysis indicates an extremely weak negative relationship between fuel price and traffic volume. However, when fuel prices are categorized into high and low, a t-test reveals a statistically significant difference in traffic volume between the two pricing groups.

Additionally, two key findings emerge from the exploratory analysis:

1. Weekday-weekend traffic patterns differ significantly, with noticeable peaks during commuting hours on weekdays, while weekend traffic follows a different pattern (as observed in the heatmap and boxplot).
2. Traffic volume dropped sharply during the pandemic, suggesting a strong external influence on travel behavior (as seen in the line graph).

Based on these insights, my next steps are:

1. Fit splines to model the effect of fuel pricing numerically and capture potential non-linear relationships.
2. If splines prove ineffective, categorize fuel prices to better illustrate traffic volume differences.
3. Incorporate weekday-weekend differences as a factor when constructing a predictive model for traffic volume.

4. If feasible, introduce the pandemic as a random event effect in the overall model to account for its impact on traffic trends.