The Pandemic and Traffic: A Look at the Change in the Number of Cars Per Cyclist/Pedestrian Traveling Through Select Toronto Intersections During the COVID-19 Pandemic

Juan Acosta

April 26, 2022 (modified)

Introduction

The COVID-19 pandemic uprooted almost every aspect of our lives, and commuting habits around the world were no exception [1]. Several lockdowns imposed globally, including in the Canadian city of Toronto throughout 2020 and 2021 forced many people to work from home, dramatically decreasing traffic levels of all modes of traffic throughout the city [1], as well as changing the travel mode preferences of Toronto commuters during both lockdown and non-lockdown periods of the pandemic [1].

This change in travel mode preferences put a spotlight on the already-present reality of Toronto roads: more people are preferring a commute by bicycle or walking as opposed to driving a car [2]. Several initiatives by the City of Toronto, such as installing more bike lanes and making intersection enhancements to improve cyclist and pedestrian safety [3], are signs of the changing times. However, these improvements have come at a slow pace [4] and cyclist and pedestrian safety is still a top concern [5].

The aim of this report is to observe if there was an increase in cyclist/pedestrian traffic relative to car traffic as the COVID-19 pandemic progressed, and if there is, then this would be another indication that infrastructure improvements to increase cyclist and pedestrian safety and mobility must be prioritized. More formally, we will seek the answer to the following question: **Did the number of cars per cyclist/pedestrian traveling through intersections in Toronto increase, decrease, or stay the same as the pandemic progressed?**

The following terminology is used throughout the report:

- **e.g.**: Shorthand for "for example".
- **i.e.**: Shorthand for "in other words".
- **etc.**: Shorthand for "and so forth".
- **Independent Variable**: A variable that is manipulated or changed by researchers and whose effects are measured and compared.
- **Dependent Variable**: A variable that depends on the independent variables and their effects.
- **First Quartile**: A numeric value where 25% of the data is below it.
- **Third Quartile**: A numeric value where 75% of the data is below it.
- **Cars Per Cyclist/Pedestrian**: This refers to how many cars travel through an intersection for each cyclist/pedestrian that also travels through the intersection. For example, at a particular intersection, 5 cars travel through for every cyclist/pedestrian that also travels through.

- **Days Since The Pandemic Was Declared**: This refers to the number of days that have passed since the pandemic was declared by the World Health Organization (WHO) on March 11, 2020 [6].

The hypothesis is that there will be a decrease in the number of cars per cyclist/pedestrian traveling through intersections in Toronto as the pandemic progressed. This could be attributed to a change in travel mode preferences as more people took to commuting on their bicycles and walking instead of driving a car, and thus the make-up of traffic traveling through an intersection would change, with a decreased percentage of cars versus an increased percentage of bicycles/pedestrians.

A brief glimpse of what the next sections of this report entail is as follows:

- *Data*: This section will look at the data collection and cleaning processes, an overview of the variables used, numerical summaries, and plots of the data.
- *Methods*: This section will give an overview of the multiple linear regression model used to observe if there is a relationship between the dependent variable (the number of cars per cyclist/pedestrian traveling through an intersection) and various independent variables. There will also be a justification for the use of this model and its significance in relation to the goal of this report.
- *Results*: This section will display the results obtained after running the multiple linear regression model on the traffic data, and an analysis of the results in relation to the goal of this report.
- *Conclusions*: This section will wrap up the report with an interpretation of the results and some final thoughts.
- *Bibliography*: This section holds all the references cited throughout this report.
- *Appendix*: This section displays additional plots and graphs from the *Data* section.

Data

Data Collection Process

The data were collected by the City of Toronto's Transportation Services division and published to the online Toronto Open Data Portal [7]. The data collected were Turning Movement Counts (TMCs). TMCs are movements observed, by each mode of travel, at a specific intersection in Toronto, either through automatic means or by manually counting. These movements include the total volume of traffic moving in a certain direction through an intersection (ex. eastbound, northbound, etc.), including turning traffic. Each TMC generally includes data collected in a series of 15-minute intervals over a span of 8 non-continuous hours throughout a specific day at a specific intersection in Toronto.

Due to the pandemic, the data collected during 2020 and 2021 has been limited to a few intersections across Toronto, with notable absences including no intersections in the downtown core nor in the entire boroughs of Scarborough and North York. In addition, for the intersections that were included in the data collection, there were gaps in the days and times when TMCs were observed and collected (i.e. inconsistent data collection).

Data Summary

The data is organized into a table consisting of several records (i.e. rows) that each correspond to a specific Turning Movement Count (TMC). Each record includes the following variables:

- An ID that identifies the traffic study undertaken to gather the data.
- The date of the count.
- A location ID.
- The name of the intersection (e.g. Islington Avenue at Market Garden Mews).
- Type of the Centreline (location) (since this data consists only of intersections, every value in this column is 2).
- The ID of the Centreline (location) (in the case of this data, the ID of the intersection).
- An identifier for the traffic control signal at the intersection.
- The start time of the 15-minute interval.
- The end time of the 15-minute interval.
- Several columns that correspond to the movements of each mode of travel (cars, trucks, buses, cyclists, pedestrians, and others) through an intersection from each direction (north, south, east and west), including the direction of turning traffic.
- The longitude and latitude coordinates of the intersection.

Using some of the above variables, the data will be aggregated by date and by select travel mode types (cars, cyclists and pedestrians), and some new variables will also be created from existing variables. This will result in a much cleaner and more focused set of data to work with.

The cleaning process makes use of the Tidyverse set of functions in `R` [8], and is as follows:

1. The Tidyverse function `mutate()` was used to aggregate all car movement variables into one car traffic volume variable called `car_traffic_volume`, and aggregate all cyclist and pedestrian movement variables into one combined cyclist and pedestrian traffic volume variable called `bike_ped_traffic_volume`.
2. The Tidyverse function `select()` was used to eliminate the variables that are not needed for the analysis. The variables that were kept are `count_date`, `car_traffic_volume`, and `bike_ped_traffic_volume`.

3. First, the Tidyverse function `group_by()` was used to group the records (rows) by the values of the variable `count_date`. This function, however, grouped the records (rows) by date according to each specific intersection, so there are several records (rows) of the same date, whereas the goal is to have one record (row) per date. Thus, the Tidyverse function `summarise()` was used in the following way: `summarise(across(everything(), sum))`. The arguments inside the function ensure that the values of the variables `car_traffic_volume`, `bike_ped_traffic_volume`, and `total_traffic_volume` across several records (rows) with the same `count_date` value are aggregated into a single record (row) with that same `count_date` value.

4. First, create a new variable named `pandemic_start_date`, and assign to it the Date value 2020-03-11, corresponding to the day that the COVID-19 pandemic was declared by the World Health Organization (WHO). Then, the Tidyverse function `mutate()` was used to create a new variable named `days_since_pandemic_declared` that takes the difference of the value of the variable `count_date` (converted into a Date value) with `pandemic_start_date`, and this difference is converted into a numeric value that gives the number of days since the pandemic began in relation to the date (e.g. March 12, 2020 would be 1 day since the pandemic was declared). Then, the Tidyverse function `filter()` was used to keep only the records (rows) where `days_since_pandemic_declared` is greater than or equal to zero (i.e. only kept the dates during the pandemic).

5. The Tidyverse function `mutate()` was used to create the following variables:

- `cars_per_bike_or_ped`, which was obtained by dividing the value of the variable `car_traffic_volume` by the value of the variable `bike_ped_traffic_volume`, and then round the resulting value to the nearest integer.
- `high_low_car_traffic`, which was obtained by using an if-else statement to assign to this variable either the string `"High Car Traffic"` if the variable `car_traffic_volume` is greater than or equal to `100479`, or the string `"Low Car Traffic"` otherwise.
- `high_low_bike_ped_traffic`, which was obtained by using an if-else statement to assign to this variable either the string `"High Bike + Ped Traffic"` if the variable `bike_ped_traffic_volume` is greater than or equal to `25143`, or the string `"Low Bike + Ped Traffic"` otherwise.

6. First, the summaries of the variables `cars_per_bike_or_ped`, `car_traffic_volume` and `bike_ped_traffic_volume` were taken to reveal the values of the first and third quartiles. Then, the Tidyverse function `filter()` was used to keep only the records (rows) that have values for the variables `cars_per_bike_or_ped`, `car_traffic_volume` and `bike_ped_traffic_volume` that are within a distance of 1.5 times the interquartile range (third quartile minus first quartile) from the third quartile of each variable.

7. Finally, the Tidyverse function `select()` was used to keep the following variables in the following order: `days_since_pandemic_declared`, `car_traffic_volume`, `high_low_car_traffic`, `bike_ped_traffic_volume`, `high_low_bike_ped_traffic`, and `cars_per_bike_or_ped`.

Thus, after the cleaning is done, there are six important variables in this report:

- `days_since_pandemic_declared`: An integer, independent variable that represents a specific date as the number of days since the pandemic was declared by the World Health Organization (WHO) on March 11, 2020. (For example, the date "2020-03-14" occurred 3 days after the pandemic was declared, so the value of the variable would be `3` in this case.)

- `car_traffic_volume`: An integer, independent variable that represents the volume of car traffic on a given day. (For example, a given day may have had a car traffic volume of `250000`.)

- `high_low_car_traffic`: A categorical, independent variable that determines whether a given day had `"High Car Traffic"` or `"Low Car Traffic"` in relation to the threshold of `100479`. (For example, if a given day had a car traffic volume of `250000`, then the variable would take on the value `"High Car Traffic"`.)

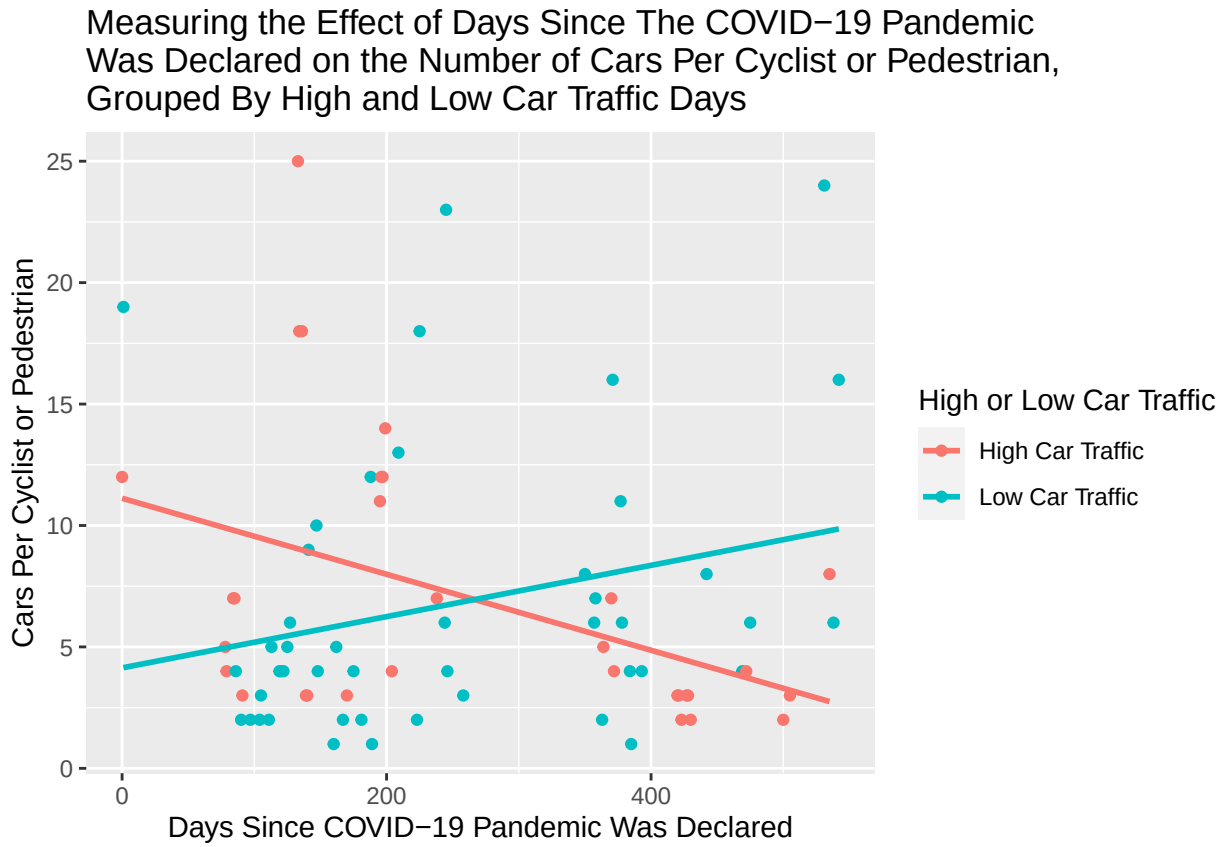
- `bike_ped_traffic_volume`: An integer, independent variable that represents the combined volume of cyclist and pedestrian traffic on a given day. (For example, a given day may have had a combined cyclist and pedestrian traffic volume of `20000`.)
- `high_low_bike_ped_traffic`: A categorical, independent variable that determines whether a given day had `"High Bike + Ped Traffic"` or `"Low Bike + Ped Traffic"` in relation to the threshold of `25143`. (For example, if a given day had a combined cyclist and pedestrian traffic volume of `20000`, then the variable would take on the value `"Low Bike + Ped Traffic"`.)
- `cars_per_bike_or_ped`: An integer, dependent variable that determines how many cars travel through an intersection for each cyclist or pedestrian that travels through an intersection. (For example, if the value of `cars_per_bike_or_ped` is `5`, then 5 cars travel through an intersection for every cyclist or pedestrian that travels through an intersection.)

The following numerical summaries were taken on the variable of interest in this report, `cars_per_bike_or_ped`:

- The mean of `cars_per_bike_or_ped` is 6.844. In other words, the average number of cars traveling through an intersection for each cyclist or pedestrian that travels through an intersection is 6.844.
- The variance of `cars_per_bike_or_ped` is 32.633.
- The standard deviation of `cars_per_bike_or_ped` is 5.713.

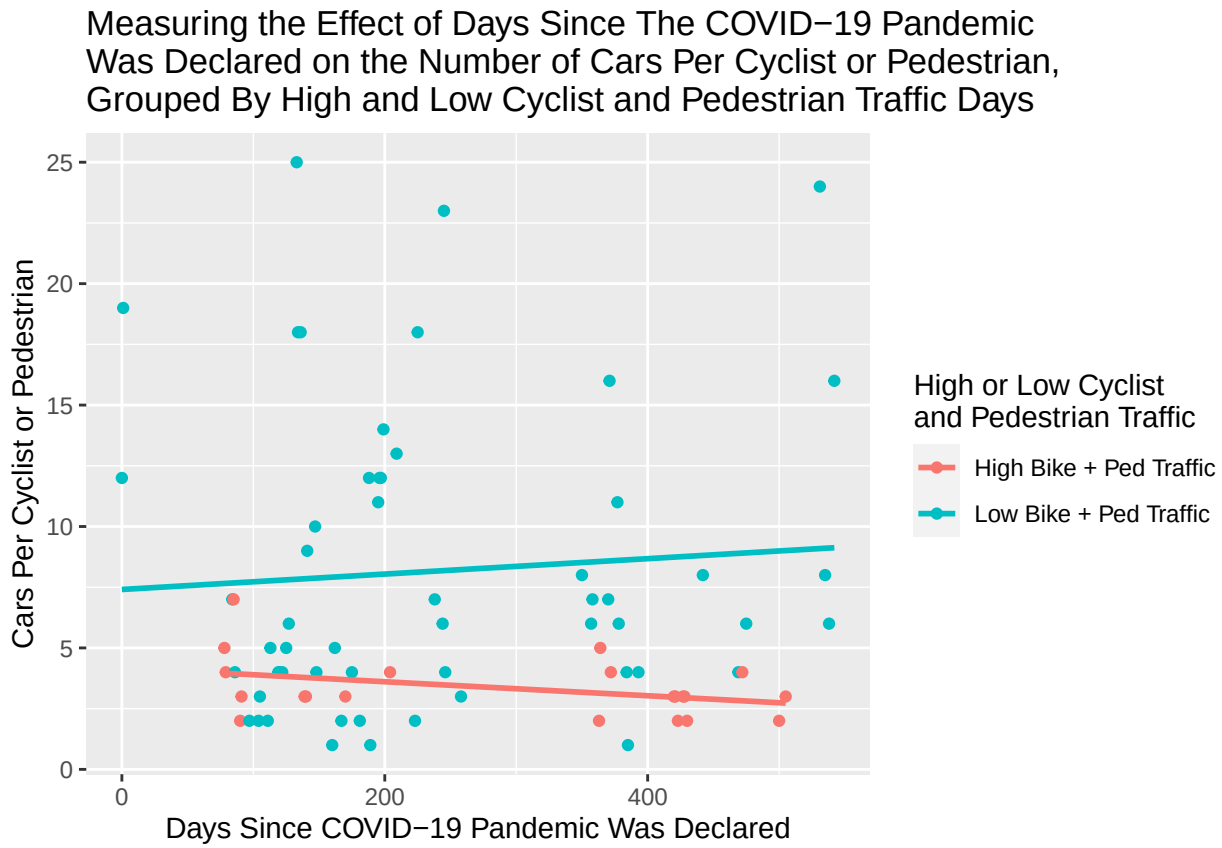
Data Plots

Plot 1



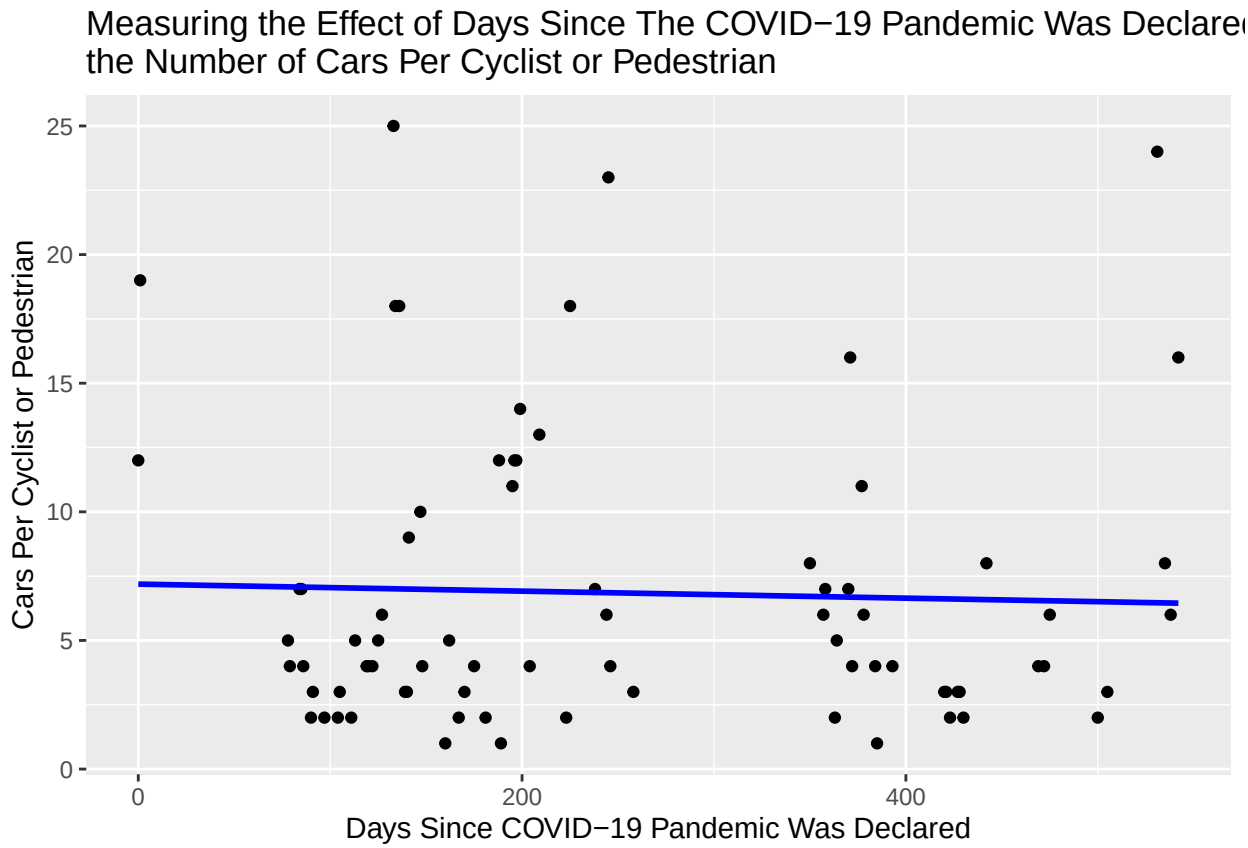
Plot 1 demonstrates that there is a different trend in the number of cars per cyclist or pedestrian as the COVID-19 pandemic progresses depending on whether a day was classified as either a high or low car traffic day. For the group of high car traffic days, there is a notable downward trend, while for the group of low car traffic days, there is a notable upward trend.

Plot 2



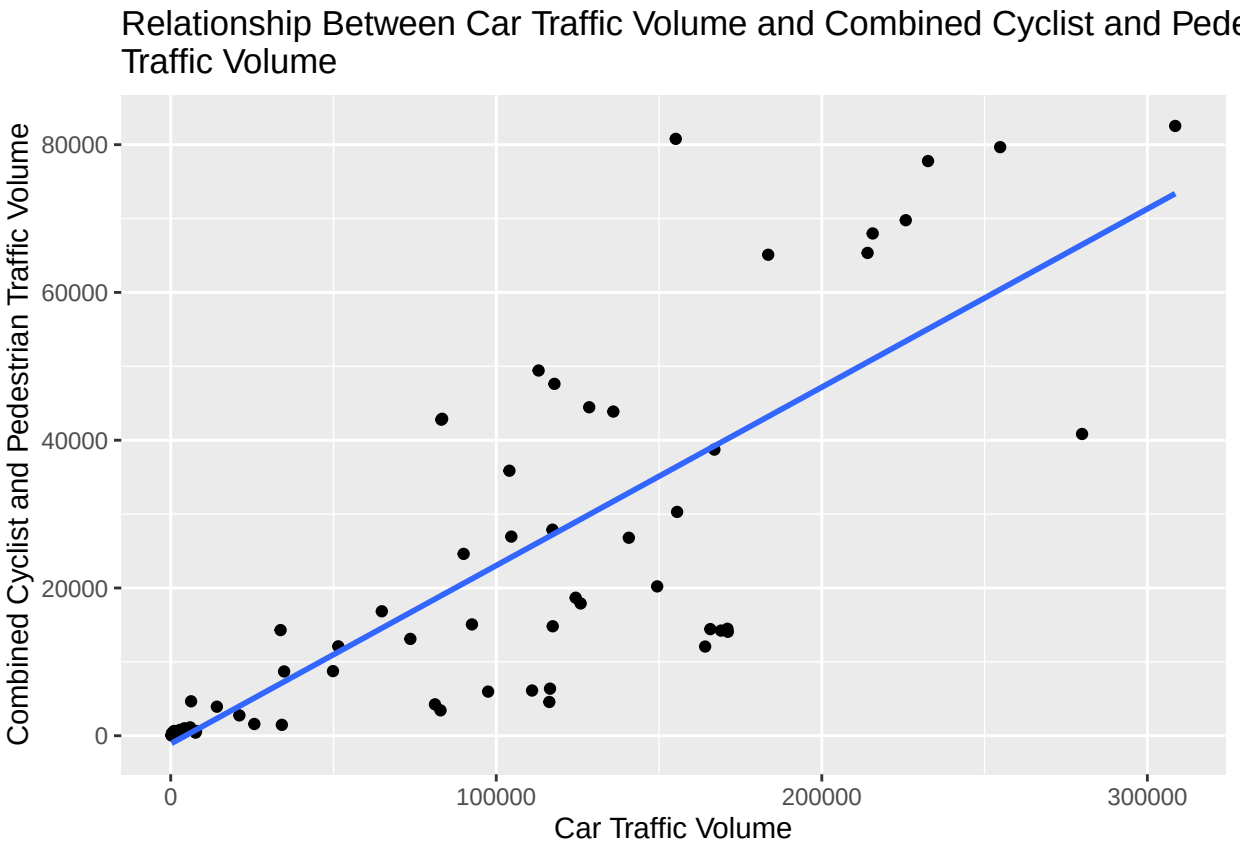
Plot 2 demonstrates that there is a different trend in the number of cars per cyclist or pedestrian as the COVID-19 pandemic progresses depending on whether a day was classified as either a high or low cyclist and pedestrian traffic day. For the group of high cyclist and pedestrian traffic days, there is a weak downward trend, while for the group of low cyclist and pedestrian traffic days, there is a weak upward trend.

Plot 3



Plot 3 demonstrates that there is a weak downward trend in the number of cars per cyclist or pedestrian as the COVID-19 pandemic progresses.

Plot 4



Plot 4 demonstrates that there is a positive linear relationship between car traffic volume and the combined cyclist and pedestrian traffic volume.

Additional plots can be found in the *Appendix* section at the end of this report.

All analysis for this report was programmed using `R version 4.1.1`.

Methods

The methodology that is used in this report is a multiple linear regression model, which is used to predict the effects of several independent variables on a dependent variable. More specifically, the multiple linear regression model used in this report is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon$.

- β_0 represents the intercept.
- β_1 represents the effect of a unit change in the variable `days_since_pandemic_declared` on y.
- β_2 represents the effect of a unit change in the variable `car_traffic_volume` on y.
- β_3 represents the effect of a unit change in the variable `bike_ped_traffic_volume` on y.
- β_4 represents the difference in the value of the variable `high_low_car_traffic` being "Low Car Traffic" versus "High Car Traffic", and the effect of this on y.
- β_5 represents the difference in the value of the variable `high_low_bike_ped_traffic` being "Low Bike + Ped Traffic" versus "High Bike + Ped Traffic", and the effect of this on y.

What this model does is it lets us predict y (the dependent variable) by plugging values into the x_is, and the effect of those values on predicting y is determined by the values of the β_is that are obtained by running the multiple linear regression. If the model is a good fit, then there is confidence that the predicted value of y is a good approximation of the actual value.

This multiple linear regression model is appropriate, because there is an assumption that there is a linear relationship between the predicted y and the independent variables. Visualizing the relationships between y and the x_is reveals that this is indeed the case.

The independent variables included in the model have been chosen using the practical rationale technique. Since we want to observe the effect of days since the COVID-19 pandemic was declared, the car and combined cyclist and pedestrian traffic volumes, whether a day had low or high car traffic, and whether a day had high or low cyclist and pedestrian traffic on the number of cars per cyclist or pedestrian, then the following independent variables will be included in the model: `days_since_pandemic_declared`, `car_traffic_volume`, `high_low_car_traffic`, `bike_ped_traffic_volume`, `high_low_bike_ped_traffic`, and `cars_per_bike_or_ped`.

Note that since `high_low_car_traffic` and `high_low_bike_ped_traffic` are categorical variables, they have been converted into factors for the model. As such, the values "High Car Traffic" and "High Bike + Ped Traffic" are now represented as the number 1, and the values "Low Car Traffic" and "Low Bike + Ped Traffic" are now represented as the number 2.

Results

The variables used in the multiple linear regression model are the following:

- `cars_per_bike_or_ped`: An integer, dependent variable that determines how many cars travel through an intersection for each cyclist or pedestrian that travels through an intersection. (For example, if the value of `cars_per_bike_or_ped` is `5`, then 5 cars travel through an intersection for every cyclist or pedestrian that travels through an intersection.)
- `days_since_pandemic_declared`: An integer, independent variable that represents a specific date as the number of days since the pandemic was declared by the World Health Organization (WHO) on March 11, 2020. (For example, the date "2020-03-14" occurred 3 days after the pandemic was declared, so the value of the variable would be `3` in this case.)
- `car_traffic_volume`: An integer, independent variable that represents the volume of car traffic on a given day. (For example, a given day may have had a car traffic volume of `250000`.)
- `high_low_car_traffic`: A categorical, independent variable that determines whether a given day had `"High Car Traffic"` or `"Low Car Traffic"` in relation to the threshold of `100479`. (For example, if a given day had a car traffic volume of `250000`, then the variable would take on the value `"High Car Traffic"`.)
- `bike_ped_traffic_volume`: An integer, independent variable that represents the combined volume of cyclist and pedestrian traffic on a given day. (For example, a given day may have had a combined cyclist and pedestrian traffic volume of `20000`.)
- `high_low_bike_ped_traffic`: A categorical, independent variable that determines whether a given day had `"High Bike + Ped Traffic"` or `"Low Bike + Ped Traffic"` in relation to the threshold of `25143`. (For example, if a given day had a combined cyclist and pedestrian traffic volume of `20000`, then the variable would take on the value `"Low Bike + Ped Traffic"`.)

Table of Results

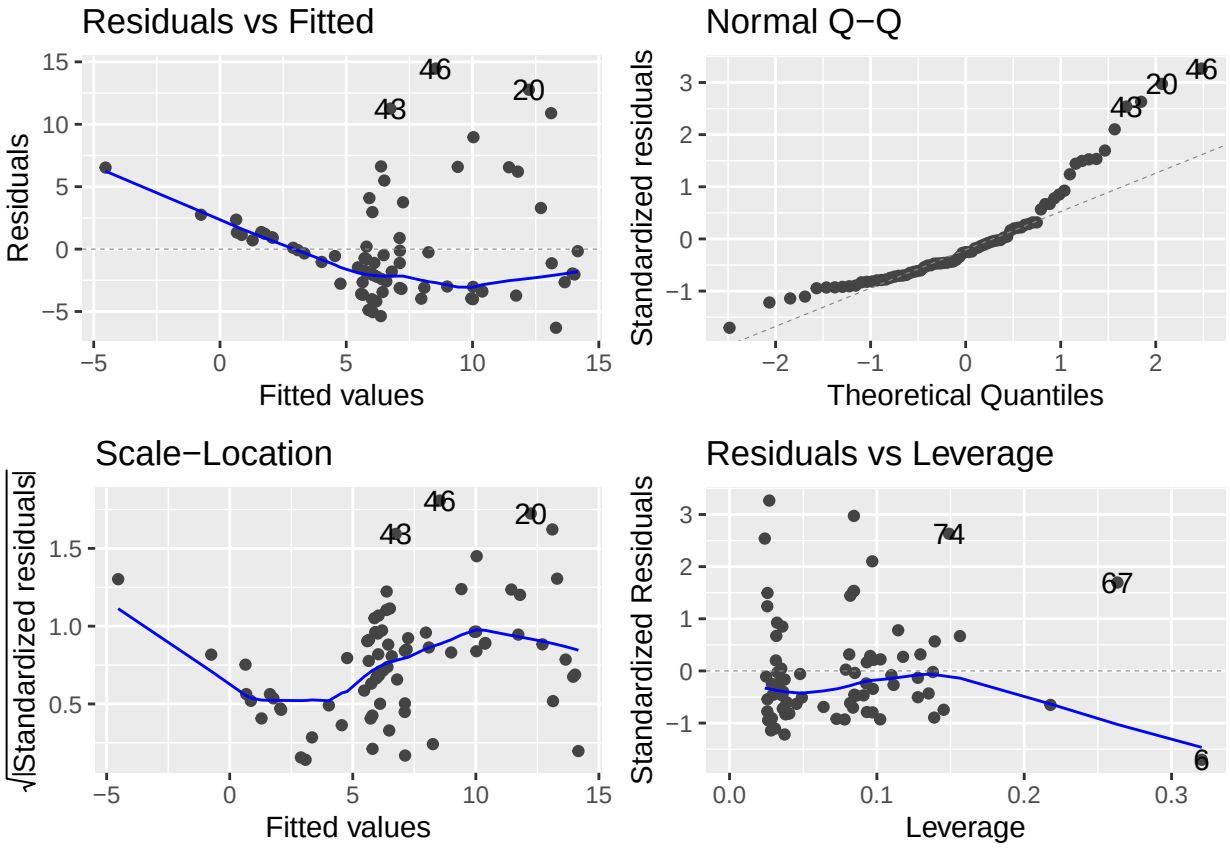
Coefficients	Estimate	Standard Error	t-value	p-value
$\hat{\beta}_0$	2.76079582	2.98003667	0.926	0.35736
$\hat{\beta}_1$	0.00516581	0.00359760	1.436	0.15542
$\hat{\beta}_2$	0.00007401	0.00001751	4.227	0.0000693
$\hat{\beta}_3$	-0.00025988	0.00006487	-4.006	0.00015
$\hat{\beta}_4$	1.00075602	2.14542971	0.466	0.64231
$\hat{\beta}_5$	1.36639639	2.58480158	0.529	0.59871

The estimates in the results table show that four of the $\hat{\beta}_i$s (specifically $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_4$, and $\hat{\beta}_5$) exert a positive effect on the model, while only $\hat{\beta}_3$ exerts a negative effect. Thus, as the values of each independent variable increase, it will be expected that the value for $\hat{y}$, the predicted value for y, will also increase.

For example, say we want to predict the number of cars per cyclist or pedestrian when it's been 600 days since the pandemic was declared, the car traffic volume is 200000, the combined cyclist and pedestrian traffic

volume is 18000, there is high car traffic (represented by the number 1), and there is low cyclist and pedestrian traffic (represented by the number 2). Then $\hat{y} = 2.76079582 + 0.00516581(600) + 0.00007401(200000) - 0.00025988(18000) + 1.00075602(1) + 1.36639639(2) = 19.71799$. Thus, we predict that there will be approximately 19.72 cars per cyclist or pedestrian.

Plots of Results



The above plots check if the model used in this report satisfies the regression assumptions [9].

- The *Residuals vs Fitted* plot checks the linearity assumption of regression [9]. There is not a well-defined pattern in the plot, so linearity of the model can be safely assumed.
- The *Scale-Location* plot checks the homogeneity of variance assumption of regression [9]. The blue line in the plot is not completely horizontal, which suggests that the residuals are not spread out equally [9].
- The *Normal Q-Q* plot checks the normality of residuals assumption of regression [9]. Not all the points follow a straight line, which makes it challenging to assume normality [9].
- The *Constant Leverage: Residuals vs Factor Levels* plot checks for outliers in the regression [9]. In this case, only one point is above 3 standard deviations, which means it is a possible outlier.

Thus, based on the above observations, it is safe to be confident in the results of the model, but further improvements can be made to reinforce all the regression assumptions.

All analysis for this report was programmed using `R version 4.1.1`. The `lm()` function in base `R` was used to derive the estimates of a frequentist multiple linear regression in this section.

Conclusions

The question that was asked at the beginning of this report was the following: **Did the number of cars per cyclist/pedestrian traveling through intersections in Toronto increase, decrease, or stay the same as the pandemic progressed?** In order to answer this question, data from the Toronto Open Data Portal [7] was cleaned, analyzed, plotted, and run through a multiple linear regression model. What is interesting is that while Plot 3 demonstrated a weak decrease in the number of cars per cyclist or pedestrian as the COVID-19 pandemic progressed, the multiple linear regression model predicts that the number of cars per cyclist or pedestrian will increase given the effects of the independent variables. Though this need not be a conflict. One explanation for this contrast can be that the model is predicting that the number of cars per cyclist or pedestrian will continue increasing until the regression line in Plot 3 switches from a negative to a positive slope.

Weaknesses

One weakness of the data is that there was no clearly-defined consistency in the dates that the data was recorded. There were several notable gaps, especially during the early pandemic, which was to be expected due to early public health guidance and restrictions in the city of Toronto. Also, and this was also possibly due to the pandemic, data was collected at only a small selection of intersections in Toronto, which completely left out intersections in the downtown core and the entire boroughs of Scarborough and North York. In order to obtain a more diverse set of data that can more confidently generalize the model to the city of Toronto as a whole, more intersections from all boroughs of Toronto, including the downtown core, must be included in the data collection process.

Next Steps

The next steps to continue the work of this report would be to collect more data from more intersections, and set up a more consistent data collection schedule to avoid large gaps in the data. Also, with more time, a new model could be made to measure the increase in the number of cars per cyclist or pedestrian in the city of Toronto from 1980 to 2021, as this is the data available through the Toronto Open Data Portal [7].

Discussion

In conclusion, the COVID-19 pandemic has ushered in a new era of changing commutes and traffic patterns [1]. In the city of Toronto, the number of cars per cyclist or pedestrian has decreased slightly as the pandemic has progressed. This means that more people are commuting on their bicycle or walking, which also means that the city must invest in infrastructure improvements to increase the safely of all road users. The car is no longer the king of the road, but it can still co-exist in a safer environment shared by all road users, because the road is for everyone.

Bibliography

1. Jackson, H. (2020, June 24). *How will commuting change after coronavirus? Experts weigh in.* Global News. Retrieved October 25, 2021, from https://globalnews.ca/news/7101358/coronavirus-commuting-economic-impacts/.

2. Bradley, S. (2017, December 1). *Major increase in Torontonians biking to work: up to 34% in some neighbourhoods.* Cycle Toronto. Retrieved October 25, 2021, from https://www.cycleto.ca/news/major-increase-torontonians-biking-work-34-some-neighbourhoods.

3. ActiveTO. (2021, September 16). *COVID-19: ActiveTO - Expanding the Cycling Network.* City of Toronto. Retrieved October 25, 2021, from https://www.toronto.ca/explore-enjoy/recreation/covid-19-activeto/covid-19-activeto-expanding-the-cycling-network/#:~:text=The%20Council%2Dapproved%20plan%20is,for%20accelerated%20installation%20in%202020.

4. Koehl, A., & Zaichkowski, R. (2018, January 15). Bike plan implementation in Toronto is too slow. dandyhorse magazine. Retrieved October 25, 2021, from http://dandyhorsemagazine.com/blog/2018/01/15/taking-a-look-at-the-pace-of-the-bike-plan-implementation/.

5. Murphy, J. (2018, July 15). *How Dangerous Are Toronto Streets for the City's Cyclists?* BBC News. Retrieved October 25, 2021, from https://www.bbc.com/news/world-us-canada-44746889.

6. Cucinotta , D., & Vanelli, M. (n.d.). *WHO Declares COVID-19 a Pandemic.* U.S. National Library of Medicine. Retrieved October 25, 2021, from https://pubmed.ncbi.nlm.nih.gov/32191675/.

7. *Traffic Volumes at Intersections for All Modes.* (n.d.). Retrieved October 25, 2021, from https://open.toronto.ca/dataset/traffic-volumes-at-intersections-for-all-modes/.

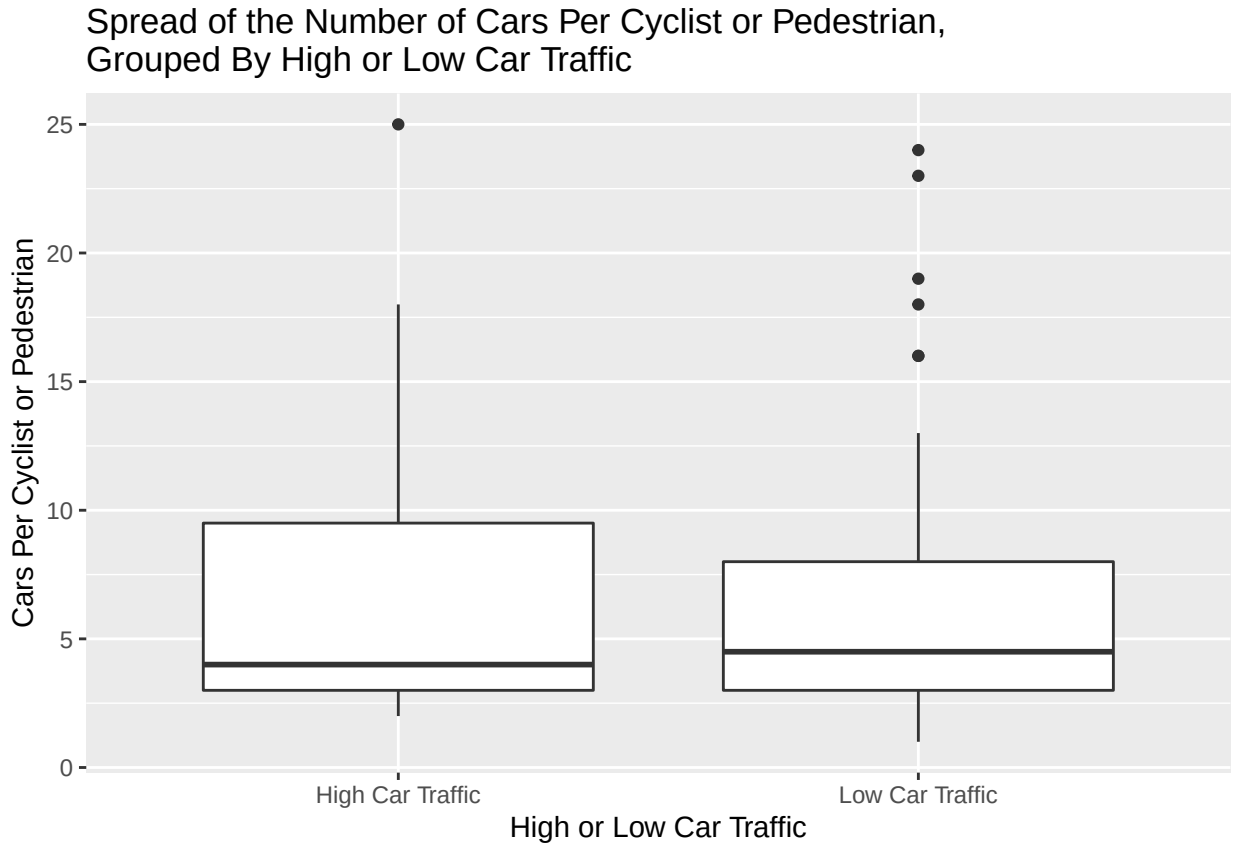
8. *Tidyverse.* (n.d.). Retrieved October 25, 2021, from https://www.tidyverse.org/.

9. Kassambara, Soyan, R., Vividdiagnostics, Eva, Visitor, & Mann, T. (2018, March 11). *Linear regression assumptions and diagnostics in R: Essentials.* STHDA. Retrieved October 25, 2021, from http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/.

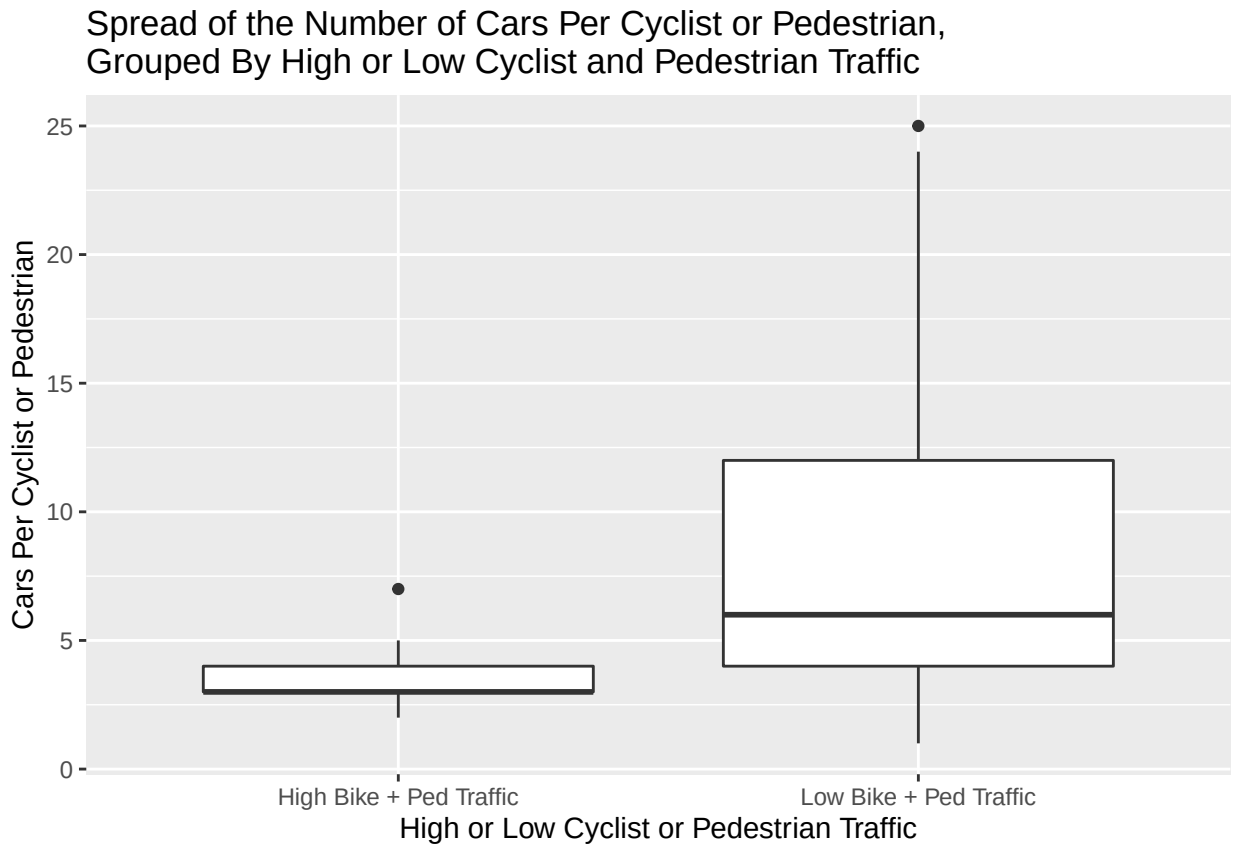
Appendix

Additional Plots

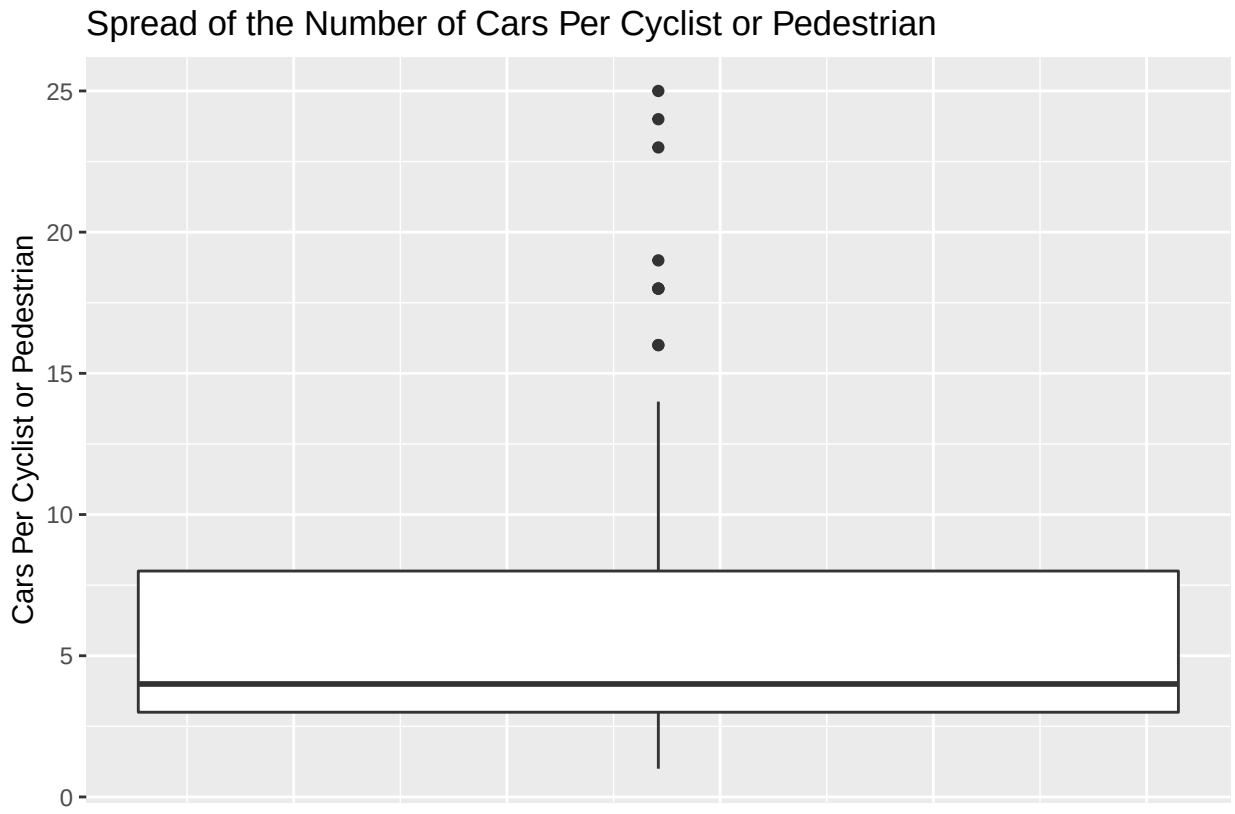
Plot 5



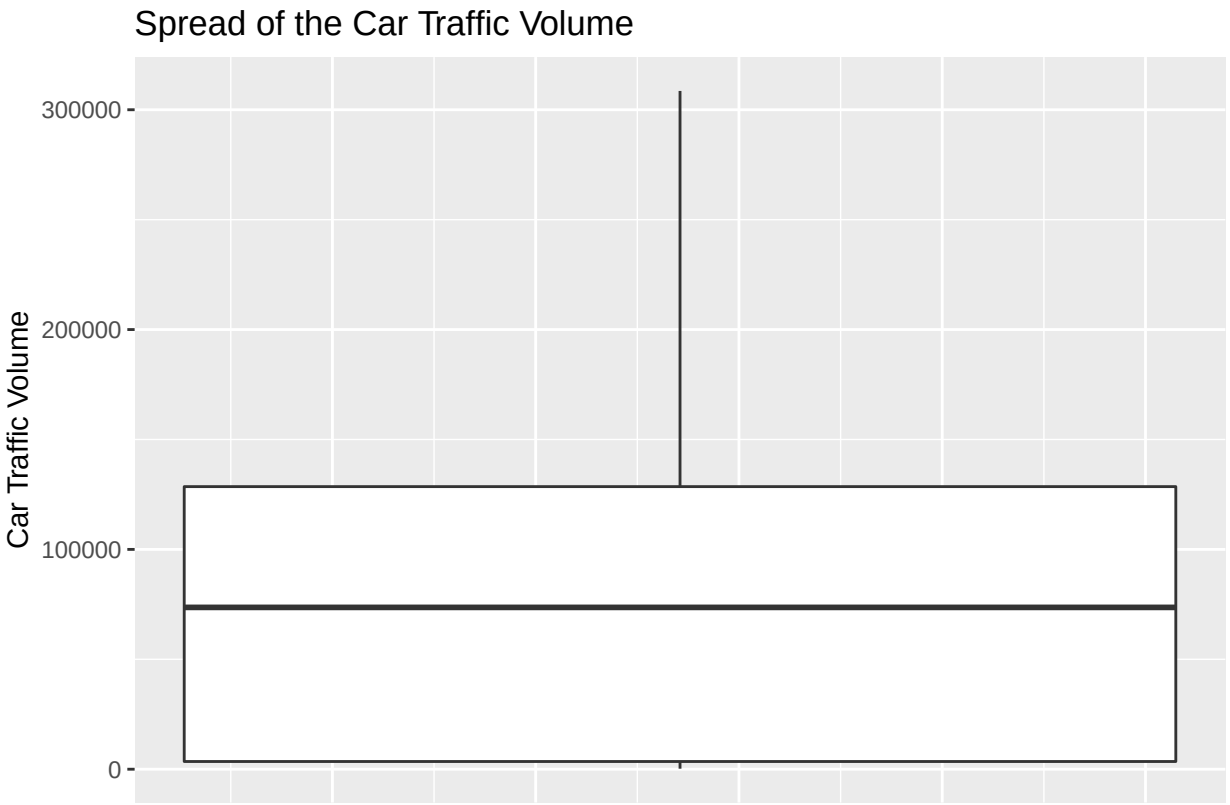
Plot 6


Spread of the Number of Cars Per Cyclist or Pedestrian,
Grouped By High or Low Cyclist and Pedestrian Traffic
Cars Per Cyclist or Pedestrian
25
20
15
10
5
0
High Bike + Ped Traffic
Low Bike + Ped Traffic
High or Low Cyclist or Pedestrian Traffic

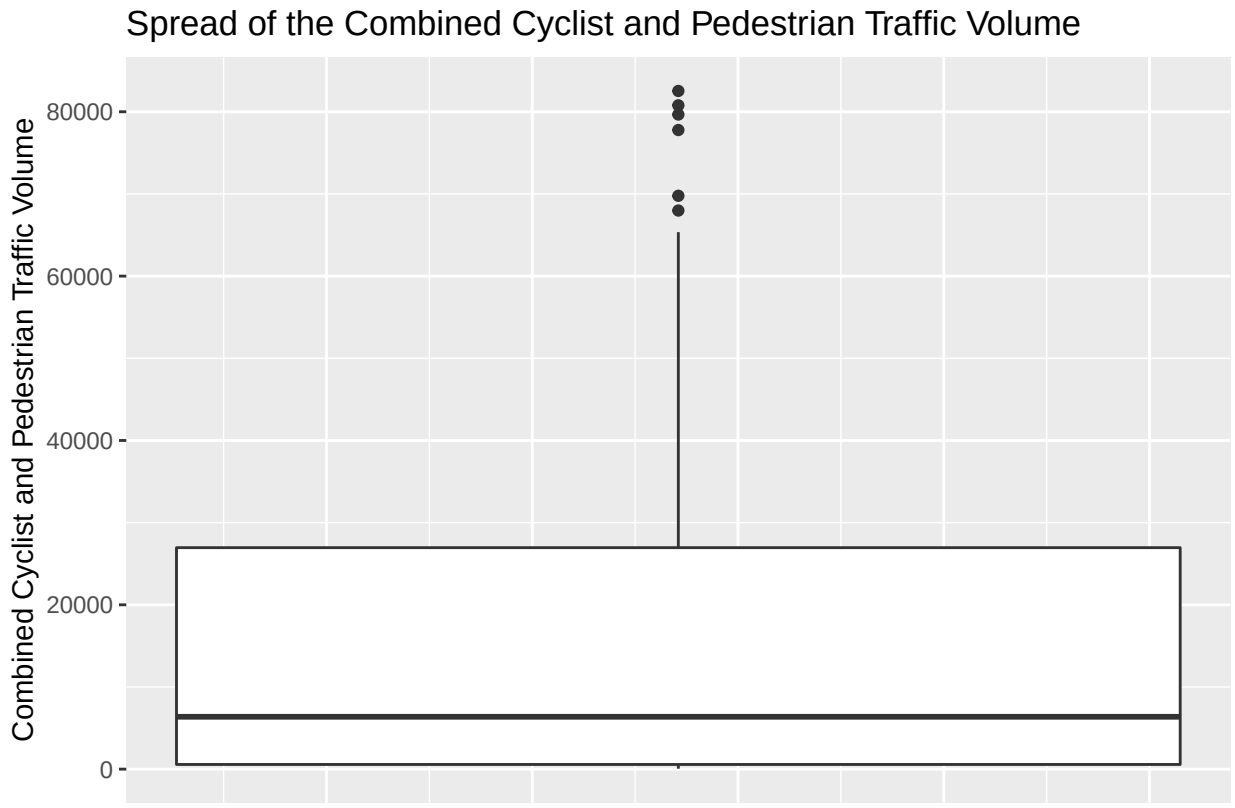
Plot 7


Spread of the Number of Cars Per Cyclist or Pedestrian
Cars Per Cyclist or Pedestrian
25
20
15
10
5
0

Plot 8


Spread of the Car Traffic Volume
Car Traffic Volume
300000
200000
100000
0

Plot 9


Spread of the Combined Cyclist and Pedestrian Traffic Volume
Combined Cyclist and Pedestrian Traffic Volume
80000
60000
40000
20000
0