# A Statistical Modelling of the TTC Subway Delays Times

Assignment 2

Chun Ki Yip - 1004532589

## Introduction

The Toronto Transit Commission (TTC) is public transit system of the city of Toronto. It was named the best transit system in North America in 2017 and it is the third busiest transit system in North America only after the transit system of New York City and Nex Mexico City [10]. It might be a surprise to many Torontonians that TTC is considered one of the best transit system in North America as it is often plagued with delays and temporary closures. The constant delays and closure of the subway impacts the livelihood and causes problems for hundred of thousands of people every year.

The goal of this report is to use the most recent TTC Subway Delay Log from September 2021 from Toronto Open Data to answer the question **"Is it possible to predict the average delay time of the TTC using number of delay occurrences and the day of the week?"**. This analysis is important because it could potentially give people a better idea on how much time they are expected to delay so they could plan ahead and minimize the impact of the delays.

A preliminary hypothesis would be that it is possible to predict the average delay time of the day with the number of occurrences of delays and the day of the week because firstly, delay time should be heavily correlated with the occurrences of delay because delays stack up especially in the subway as there is only one track for each direction, also if more delay happens then there would be more people waiting in the station and causing more traffic and furthering the delay. Secondly, the time of the week should also influence the average delay time because there are more users of the TTC during weekdays as opposed to weekends as most people have to commute to work and school only during weekdays, so it would lengthen the average delay times caused by station traffic, but there could also be a case where the weekends would average a higher delay time because the TTC usually chooses the weekend for maintenance as it has less usage, which may increases the average delay times potentially.

Some important terminology that are used later in this report are: linear regression, which is statistical model that is used to model and analyze the correlation between numerical data with its independent variables [7]. Parameters are the unknown numbers that statisticians estimate which makes up a statistical model and ties the independent variables with the dependent variables. Lastly, p-value is a threshold which statisticians use to determine whether the result of the analysis is plausible or not.

## Data

### Data Collection Process

The TTC Subway Delay Data comes directly from the TTC, it is described that the TTC relies on staffs to manually record and submit reports of each delay in a Service Efficiency Study on the Toronto Transit Commission. The final data available to the public on Toronto Open Data includes information such as the data, time, day of the week, station, a code describing the cause of delay, minutes of delay, the minute gap between the delayed subway and the previous subway, the direction, line, and vehicle number of each delay [8].

The following will describe the cleaning process of the raw data.

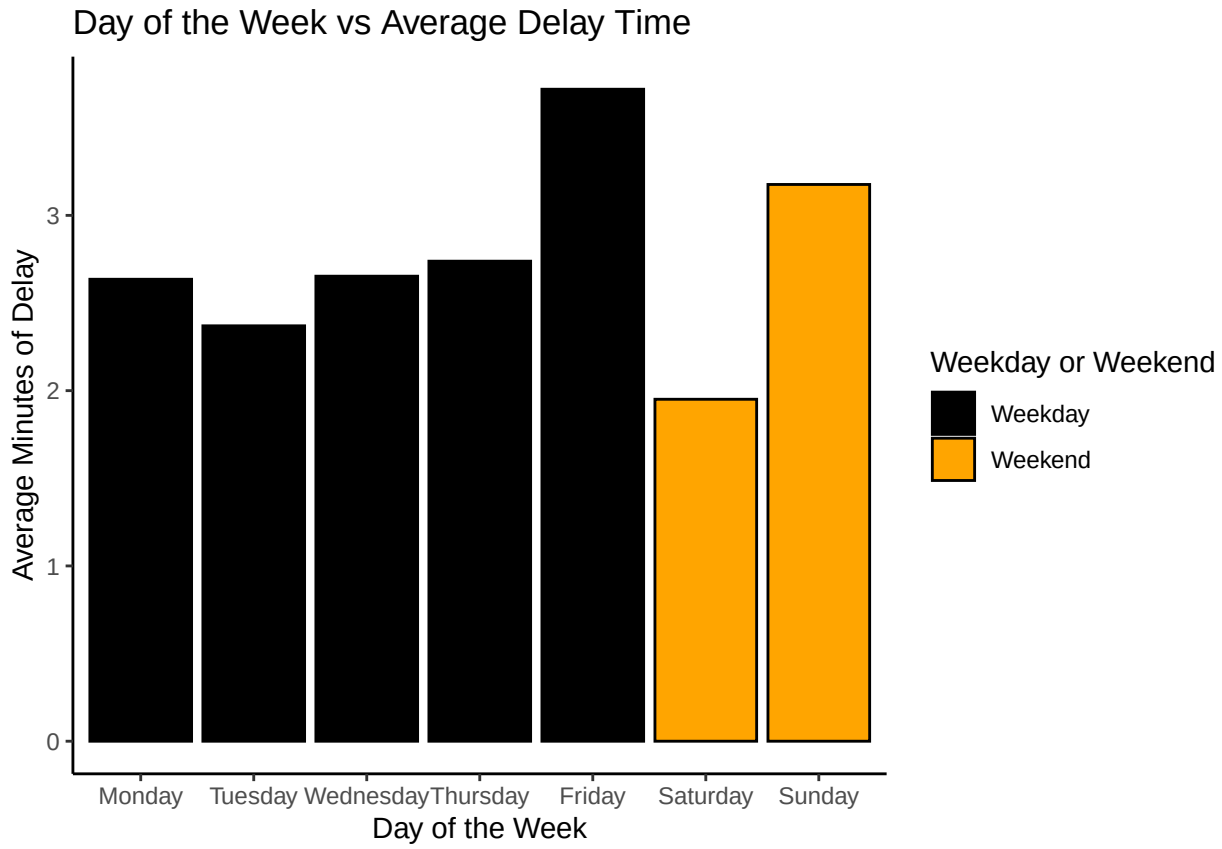
1. Data was pulled directly from Toronto Open Data using the developer package provided [9].
2. The most recent data from the month of September 2021 was chosen as the target data [6].
3. An extra column was created using the mutate function to denote whether or not the day of the week is a weekday or a weekend using 0 and 1, with 0 representing weekday and 1 representing weekend.
4. Each observation of delay of the same day was then grouped together using the group by function and the average delay time in minutes would be calculated using the summarized and mean functions to as the second variable of the grouped data. Meanwhile the binary categorical variable of whether if it's during the weekend or not, would be the third variable of this data calculated by the mean of the 1 and 0 of each occurrence of the delay of the day, this calculation is trivial because all the delays are already grouped such that they are the same day of the week so it would either be all 0s or all 1s which would result in the average or 0 and 1 respectively too. This intermediate clean data would then be named data3 for later use.
5. Going back to the original raw data, each observation of delay of the same day would be grouped together again, but this time the second variable that is summarized is the total number of delay on the given day, this would be counted using the tally function. This second intermediate cleaned data would be named data2 for later use.
6. The two intermediate cleaned data would then be merged using the merge function to create a data set with 4 variables, the date, number of delays, whether it's on the weekend or not, and the average minute of delay.
7. Then the another extra column would be added using the mutate function to translate the 0 and 1 into a more comprehensible form of "No" or "Yes" denoting whether it is during the weekend or not.
8. Finally, the important variable which are interested namely the revised binary vategorical variable of whether or no it is during the weekend, the number of occurence, and the average minutes of delay would be picked out using the select function.

Some foreseeable limitation of this process is that an important variable of the cause of delay was eliminated from the start due to it's complex nature of having more than 200 different categories, this may severely impact the reliability of the model as it may be an important variable to predicting the delay time. Another drawback is that grouping the categorical of the day of the week into binary options is less optimal because it makes the data less descriptive and it may hinder the accuracy of the model too. Another drawback is that by choosing the total number of delays as an independent variable for the dependent variable of average minutes of delay, implies that the passenger would have knowledge of the total number of accidents that would occur during the rest of the day, but realistically, the passenger would only be able to know a very rough estimate of the number of delays by counting the delay announcements on the TTC website up to when they are getting on the subway. This would make the model less useful in terms of daily uses but it could still be accurate to a certain degree by using number of delays up to the current time if it is later discovered that the number of occurrences is in fact a significant variable.

## Data Summary

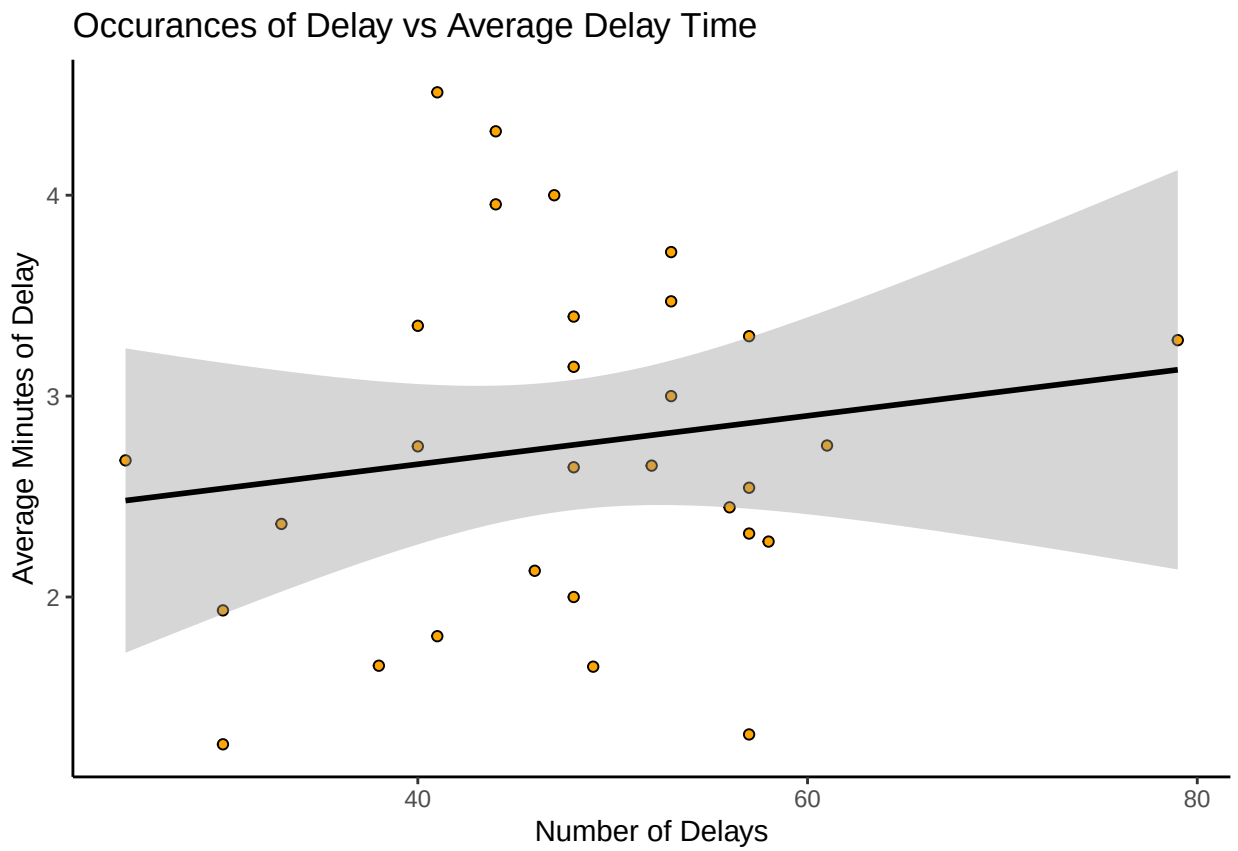
As described above, the 3 important variables chosen for this analysis are the binary categorical variable of whether it is the weekend or not, the total number of delays of the day, and the average minutes of delay of the day, with the first two variables being the independent variable and the third being the dependent variable. The following section would include some summary statistics of the variables and graphs that would help the readers understand the data more.

The average delay by minutes on a weekday is 2.8231308 and the average delay minutes on the weekend is 2.5634575. Below would show a more detailed graph of the average delayed times on each given day of the week.




This bar graph shows the average delay time on each day of the week and color codes the weekdays from the weekends, from the graph, it could be seen that Sunday of the weekend actually has the second highest average delay but Saturday has the lowest average delay making the average time of the weekend lower than the weekdays. It could also be seen that Friday has the highest average delay while the other four weekdays have simililar delay times, so it could be said that weekday delay times are relatively more consistent than the weekend delay times. It is also important to note that the result of this data matches the preliminary hypothesis that weekdays should have higher average delay time.

The minimum number of delay on a day is 25 in the month of September of 2021, and the maximum number of delay on a day is 79 times in the month of September of 2021, and the average number of delays in the month of September 2021 is 47.7666667. The scatter plot below graphs the number of delays of the given day against the average time of delay.




This graph helps readers visualize the positive correlation between the number of delays and the average time of delay. The black line shows a simple linear regression line between the two variables which shows a positive correlation agreeing with the preliminary hypothesis, and the grey shadeing shows the error of the regression line with the actual data.

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

## Methods

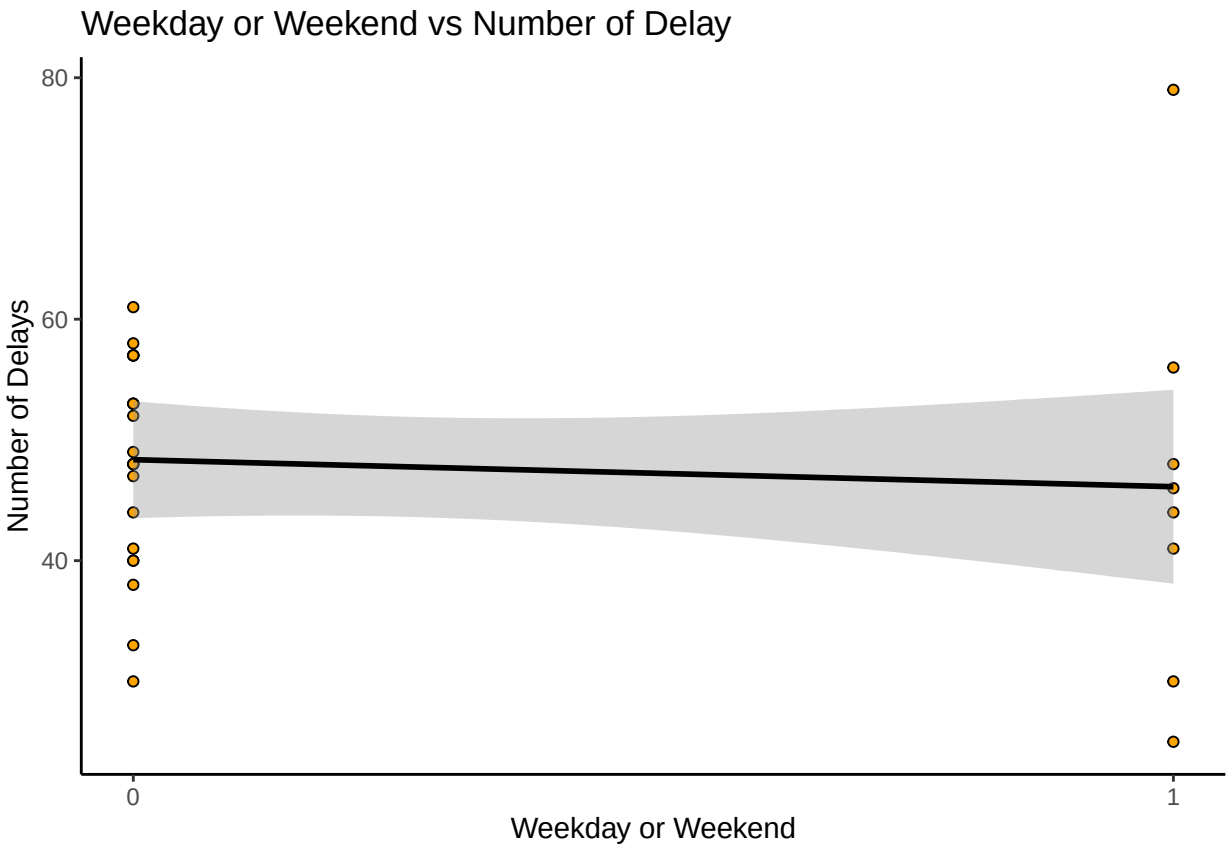
The methodology used to model the data is a multiple linear regression. A multiple linear regression uses two or more explanatory variable to predict a numerical response variable, in this case there would be two explanatory variable which are the number of occurrences of delay and the binary categorical variable of whether if the day is on the weekend. The mathematical formula of a multiple linear regression with two explanatory variables is shown as below.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_{2i} + \epsilon_i$$

Where $y_i$ is the dependent variable, $\beta_0$ is the y-intercept, $\beta_1$ is the parameter or coefficient of the first explanatory variable, $\beta_2$ is the parameter or coefficient of the second explanatory variable, $x_i$ and $x_{2i}$ represents the two explanatory variables, and $\epsilon_i$ represents the error of the model.

This statistic model predicts the parameters that best fit the 3 variables in this and the data to create a complete model. This model is appropriate because the dependent variable is a numeric value so a linear regression is most fitting, while having more than one explanatory variable makes using a multiple linear regression more fitting than a simple linear regression.

An assumption is made that the two explanatory variables are independent of each other, which would be checked using a simple linear regression of the two, with the weekend coded into dummy variables of 0 and 1 with 0 representing a weekday and 1 representing the weekend.




We can see that the regression line is fairly straight meaning that the two variables' correlation should be close to 0, and the p-value for this linear regression is 0.629 which is much higher than the statistical significance of 0.05 meaning it has a 62.9 percent chance that this correlation is a coincidence and therefore the assumption that the two independent variable are independent of each other is true.

## Results

Since it is checked that the two independent variables are not correlated, it satisfies the assumption to be variables of a multiple linear regression. Namely, the first independent variable would be the binary categorical variable of whether if it's weekend or not represented as $\hat{\beta}_1$ and the second independent variable would be the number of delays represented by $\hat{\beta}_2$.

Below are the results

| | |
|---|---|
| $\hat{\beta}_0$ | 2.268 |
| $\hat{\beta}_1$ | -0.20113 |
| $\hat{\beta}_2$ | 0.01130 |

This table shows the predicted parameters of the model, with $\hat{\beta}_0$ being the constant term being 2.268, the $\hat{\beta}_1$ being the coefficient of the weekend variable, and $\hat{\beta}_2$ being the coefficient of the number of delay variable.

An example of how this model works would be to predict the expected average delay time on a weekday where 40 delays occurred would be 2.72 minutes, the formula would be as follow.

$$2.268 + -0.20113(0) + 0.01130(40) = 2.72$$

All analysis for this report was programmed using `R version 4.1.1`. I used the `glm()` function in base `R` to derive the estimates of a frquentist logistic regression in this section [4].

# Conclusion

Although a model is successfully generated, the p-values of both independent variables and the multiple linear regression are all very high (0.58459, 0.45609, 0.623 respectively), meaning there is not enough evidence that the two independent variables nor the multiple linear regression is effective in predicting the dependent variable. Therefore, to answer the research question of **"Is it possible to predict the average delay time of the TTC using number of delay occurrences and the day of the week?"**, the answer would be no, and we would have to reject the hypothesis of which the two variables are able to model the average delay time of the TTC subway.

Some limitations that was speculated before was that grouping the categorical data of the day of the week into a binary category makes it way less descriptive and therefore limiting the accuracy of the model, another limitation that was speculated before was that eliminating the code that describes the reasons of delay was also a bad decision because it is a very descriptive quantitative data with more than 200 categories which may as well be the strongest predictor for predicting average time delay. In the future when time is not a limitation, a model could be made using more indepent variables such as time, and the delay codes, and also using better model selection methods to select variables, as selection method wasn't used in this report due to the choice of only using two variables and it is the minimum requirement to use at least two variables in this specific report.

# Bibliography


1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: October 12, 2021)

2. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.

3. Allaire, J.J., et. al. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: October 12, 2021)

4. Peter Dalgaard. (2008) *Introductory Statistics with R, 2nd edition.*

5. Wickham, H., et. al. *Create Elegant Data Visualisations Using the Grammar of Graphics ggplot2.* from https://ggplot2.tidyverse.org/. (Last Accessed: October 26, 2021)

6. Wickham, H., et. al. *Tidyverse.* from https://tidyverse.org/. (Last Accessed: October 26, 2021)

7. Gibbs , A., & Stringer, A. (2021, January 20). *Probability, statistics, and data analysis - GitHub Pages. GitHub.*, from https://awstringer1.github.io/sta238-book/. (Last Accessed: October 26, 2021)

8. Hayes, A. (2021, September 7). *Multiple linear regression (MLR) definition.* Investopedia. from https://www.investopedia.com/terms/m/mlr.asp. (Last Accessed: October 26, 2021)

9. Toronto Open Data . *City of Toronto Open Data.* Retrieved October 26, 2021, from https://open.toronto.ca/. (Last Accessed: October 26, 2021)

10. Toronto Transit Commission. *TTC.ca.* from http://www.ttc.ca/. (Last Accessed: October 26, 2021)

11. Gelfand, S. *Access the City of Toronto open data portal.* . from https://sharlagelfand.github.io/opendatatoronto/. (Last Accessed: October 26, 2021)

12. Toronto.ca. (2011). *Service Efficiency Study on the Toronto Transit Commission.* toronto.ca. , from https://www.toronto.ca/wp-content/uploads/2017/08/9711-ttc_ses.pdf. (Last Accessed: October 26, 2021)