Playing The Waiting Game

An Examination of Wait Times due to TTC Bus Delays during the first half of 2021

Joseph Hotz - 1005953354

October 25, 2021

Introduction

The Toronto Transit Commission (TTC) is the body which manages public transportation in the City of Toronto. Although the TTC's primary focus is on Toronto "proper", the transit system connects to many transit systems such as GO Transit and York Region Transit, which operate beyond Toronto's city limits, as shown on TTC's System Map (TTC, 2021).

The TTC is held in high regard by many who work in the transit industry; in 2017, the TTC was awarded "Transit System of the Year" by the American Public Transportation Association (APTA) (Spurr, 2017), and in 2018, the TTC was rated the 3rd best transit system in North America by Walk Score (Basa, 2018).

Despite these high praises, many Toronto citizens have disputed whether the TTC's service is deserving of such accolades, (Draaisma, 2017), as the TTC's service on weekdays frequently falls prey to issues such as overcrowding, delays, and subway closures, and stretches of the TTC's subway lines are often closed on weekends, with this service replaced by shuttle buses.

Prior to the COVID-19 pandemic, the TTC provided transit to 1.7 million riders per day (TTC, 2021). As one might expect from a global pandemic, TTC ridership fell significantly in 2020. As of September 2021, TTC ridership is steadily increasing from its pandemic lows, but the overall ridership is not expected to reach its pre-pandemic numbers until the year 2023 (Fox, 2021).

In March 2021, the TTC averaged over 1 million passengers boardings per weekday, with over 55% of these boardings occurring on TTC buses (TTC, 2021). The TTC is a major mode of transportation for hundreds of thousands of people per day, and many Torontonians can attest to the frequency of TTC bus delays, both prior to, and during the COVID-19 pandemic, which often increase the time spent waiting for a bus, as well as their overall time on transit.

In this report, we aim to answer the question "*When will the next bus arrive at my stop?*" This is a question which many TTC riders ask themselves daily, and being able to determine factors which affect these arrival times will surely give countless TTC riders some peace of mind.

Using the data posted on the City of Toronto's Open Data Portal, we will examine bus delays on the TTC throughout the first half of 2021, in order to create a linear model which will predict the expected amount of time before a TTC bus arrives at a bus stop when there is a delayed bus on that route.

Our linear model will be based on three factors:

- The time of day when the delay occurred
- Whether the delay occurred on a weekday or a weekend
- The cause of the delay.

In addition, we will use this linear model to determine how these factors affect the expected waiting time for a TTC bus when there is a bus delay. This can help us differentiate between the reliability of TTC service at different times of the day, as well as on weekdays as opposed to weekends, and it can also points us towards determining which incidents have a longer expected waiting time for the next bus to arrive.

We hypothesize that there will be a somewhat negative relationship between the time of day and the waiting time for a bus, because the TTC would likely run buses more frequently during the daytime and peak travelling hours, due to people commuting to school and work, and they would run fewer buses (increasing the gap between successive buses on the same bus route) outside of these hours.

Similarly, we believe that the waiting time when a bus is delayed will be longer on the weekend than on a weekday, because there are fewer people using the TTC to commute to school and/or work on weekends than on weekdays. Furthermore, in their own reports, the TTC primarily measures their ridership statistics in terms of the amount of riders on weekdays. (TTC, 2021) Since the TTC's internal metrics rely on weekday riders, we expect that the quality of TTC service on weekends is not held to the same standard as on weekdays.

Data

Data Collection Process

As mentioned above, the TTC bus delay data which we will examine is the dataset posted on the City of Toronto's Open Data Portal (City of Toronto, 2021).

In particular, we will be examining the data from 2021 which is in the "TTC Bus Delay Data" data (TTC, 2021), which is stored in the "ttc-bus-delay-data-2021" file on the website. This dataset was posted on the website by a representative of the TTC organization, which tells us that the source of the data is a primary source, as the TTC is the body which has published the data about the TTC's bus services.

As the data is recorded and published by the TTC themselves, the largest possible source of error in the data is due to incorrectly-inputted values, but it is unlikely that there was any bias which affected the data which the TTC collected and published. Unfortunately, the Open Data website does not have any further information about the TTC's data collection processes, so we will assume that these the data which we collected is fully representative of the time period from January to June of 2021, although this assumption may not necessarily be true.

The data was last updated on October 21, 2021 (TTC, 2021) to include bus delays which occurred in July and August 2021. However, for some unknown reason (likely a data inputting error), the observations recorded in these months correspond to TTC streetcar delays instead of TTC bus delays as one can see by looking at the vehicle routes listed for "bus" delays in July and August, and comparing these routes to a list of TTC streetcar route numbers. Due to this discrepancy, we will not examine the data for these months, as this streetcar data is unrelated to the bus delays which we chose to investigate, and these are fundamentally different modes of transportation.

Data Summary

The data which we are examining represents TTC bus delays which occurred between January 1, 2021 and June 30, 2021, inclusive. Information about these bus delays was collected, compiled, and posted online to the Open Data Portal by the TTC organization. Each of the observations in the data represents a bus delay which occurred in the timeframe mentioned above.

In addition to providing data for this report on the Open Data Toronto website, the TTC also included a "readme" file with details about the variables in their reported data.

There are 5 variables in the data which was uploaded by the TTC which are important to our analyses. A description of these variables, as given in the TTC's provided "readme" file is displayed below.

Table 1: Descriptions of Important Variables

Variable Name	Description	Example
Time	The time (hh:mm:ss AM/PM) when the delay-causing incident occurred	12:35:00 AM
Day	The name of the day	Monday
Incident	The description of the delay-causing incident	Mechanical
Min Delay	The delay, in minutes, to the schedule for the following bus	10
Min Gap	The total scheduled time, in minutes, from the bus ahead of the following bus	20

Data Cleaning

First, we will look to see if there are any missing and/or improperly-recorded values in the data which we retrieved from the TTC.

Table 2: Missing values for each variable

Column Name	Missing Values
Time	0
Day	0
Incident	0
Min.Delay	0
Min.Gap	0

As we can see above, there are no unrecorded values in our variables of interest, which is excellent. However, as we have already seen with the accidentally-uploaded streetcar data, the data which is collected and posted by the TTC is not immune to human error, so we still need to examine whether there are any values which are not recorded properly, even if they are not necessarily missing.

In addition to examining if there are any improperly recorded values in our data, we will also look at the distribution of the overall data, assuming that everything is recorded properly.

First, we will examine the days of the week when the TTC bus delays occurred, as well as the incidents listed as the causes of these bus delays.

Table 3: TTC bus delays per day of the week

Day	Count
Monday	3732
Tuesday	4134
Wednesday	4281
Thursday	3881
Friday	3810
Saturday	2455
Sunday	2176

As we can see above, the only days which are listed in our data are the seven days of the week, so it would appear that at a minimum, there are no days which were spelled incorrectly, although we cannot be certain that no days were incorrectly inputted altogether.

However, examining this idea with respect to Occam's razor (Andersen-Wood), it is a much simpler (and more reasonable) assumption to just say that there are not any days where the day was spelled correctly but incorrectly inputted, and instead, we will just assume the 'simplest explanation' holds; i.e. the days are all correct.

Table 4: Incidents listed as causes for TTC bus delays

Incident	Count
Cleaning	8953
Collision - TTC	1136
Collision - TTC	1
Diversion	73
e	1
Emergency Services	908
General Delay	521
Held By	65
Investigation	592
Late Entering Service	11
Late Leaving Garage	1
Late Leaving Garage - Mechanical	4
Late Leaving Garage - Operations	4
Late Leaving Garage - Operator	8
Management	9
Mechanical	6037
Operations - Operator	3425
Road Blocked - NON-TTC Collision	638
Security	1455
Utilized Off Route	201
Vision	426

In the list of incidents shown above, there are some anomalous results which are shown as incidents which have caused delays, such as the two seemingly-identical "Collision - TTC" incidents, which are counted separately, as well as the strange incident "e", which provides no information about what actually caused the bus to be delayed.

We will clean these values to ensure that our data are easier to work with, while still representative of the initial data. In particular, we will combine the two "Collision - TTC" incidents into one incident with this name. Additionally, we will change all instances of the incident "Road Blocked - NON-TTC Collision" by replacing this incident with the shorter name "Collision - Non-TTC".

We will also remove the observation with the incident "e" from our data altogether, as it is not clear which delay-causing incident this value refers to.

Lastly, we will reclassify the three separate incidents "Late Leaving Garage - Mechanical", "Late Leaving Garage - Operations", and "Late Leaving Garage - Operator" by combining these all into the "Late Leaving Garage" incident.

Table 5: Incidents listed as causes for TTC bus delays

Incident	Count
Cleaning	8953
Collision - Non-TTC	638
Collision - TTC	1137
Diversion	73

Incident	Count
Emergency Services	908
General Delay	521
Held By	65
Investigation	592
Late Entering Service	11
Late Leaving Garage	17
Management	9
Mechanical	6037
Operations - Operator	3425
Security	1455
Utilized Off Route	201
Vision	426

The table generated above represents the data after "cleaning" the set of incidents by performing the operations which we described above. It is clear from this new table that the cleaned data maintains the same overall information as the 'original' TTC data, but there are no longer prominent issues with the set of delay-causing incidents.

Now that we have looked for (and fixed) improperly recorded values in our data, we can manipulate the data to better represent the variables of interest which we care about.

First, we will convert the recorded time of the bus delays, so that instead of being listed as a time in hours and minutes, they will be counted based on the number of minutes since midnight. To do this, we will define a function which converts a set of characters of the form "HH:MM" to an integer, and apply this function to each individual row to create our new variable using the `mutate()` function.

Next, we will convert the recorded days of the week into weekdays and weekends, where a "weekend" is classified as a day falling on Saturday or Sunday, and a "weekday" is any of the other 5 days of the week.

Table 6: TTC bus delays on Weekdays and Weekends

Type of Day	Count
Weekday	19837
Weekend	4631

Lastly, we will measure the total delay for a TTC bus delay by determining the time until the next bus arrives when a TTC bus is delayed.

As mentioned in the summary of our variables above, "Min Gap" represents the amount of time (in minutes) between the delayed bus and the next bus on its route, according to the TTC's schedule for the route. Similarly, "Min Delay" measures how far behind schedule (in minutes) the delayed bus is, as a result of its delay.

From the perspective of someone at a bus stop waiting for the next bus to arrive, they do not particularly care whether the bus which arrives is the originally-scheduled bus, or if it is the next bus on the schedule, as their priority is simply to get on a bus.

If the scheduled gap is larger than the time which the bus takes to overcome its delay, then the next bus which will arrive at the bus stop for this route will be the delayed bus, but it will be a bit behind schedule, as the bus was delayed before it could reach the stop.

Otherwise, if the bus's delay is larger than the scheduled gap, then the bus which will pick them up at their stop is the next bus on their bus route, so the increased waiting time is dependent on how long it will take for the next bus to arrive, which is given by the "Gap", as long as the following bus stays on schedule.

Thus, for a TTC rider who is waiting for a delayed bus, we can write that Wait $= \min\{\text{Delay}, \text{Gap}\}$, where "Wait" represents the actual time until another bus arrives.

This waiting time will be our new variable which we will use in our analyses, and this variable represents the amount of additional time required for someone to wait for the bus as a result of a delay.

Next, we will determine summaries of the important numerical columns in our data; the minutes since midnight, the delay time, and the gap time. These summaries will be computed using the built-in R functions `min`, `quantile`, `median`, and `mean`.

Table 7: Summaries of our numerical variables of interest

	Minutes After Midnight	Wait Time (Minutes)	Delay (Minutes)	Gap (Minutes)
Minimum	0.000	0.000	0.000	0.000
First Quartile	599.000	9.000	9.000	18.000
Median	811.000	12.000	12.000	24.000
Third Quartile	1029.000	20.000	20.000	37.000
Maximum	1439.000	999.000	999.000	999.000
Mean	805.008	17.552	17.922	30.738

Data Exploration

Next, we will look at some plots of the data which we have collected, so that we can get a better feel for the variables which we are working with, as well as developing a better overall sense of the data which we collected from the TTC, and the distributions of the individual variables which we are using as predictors.

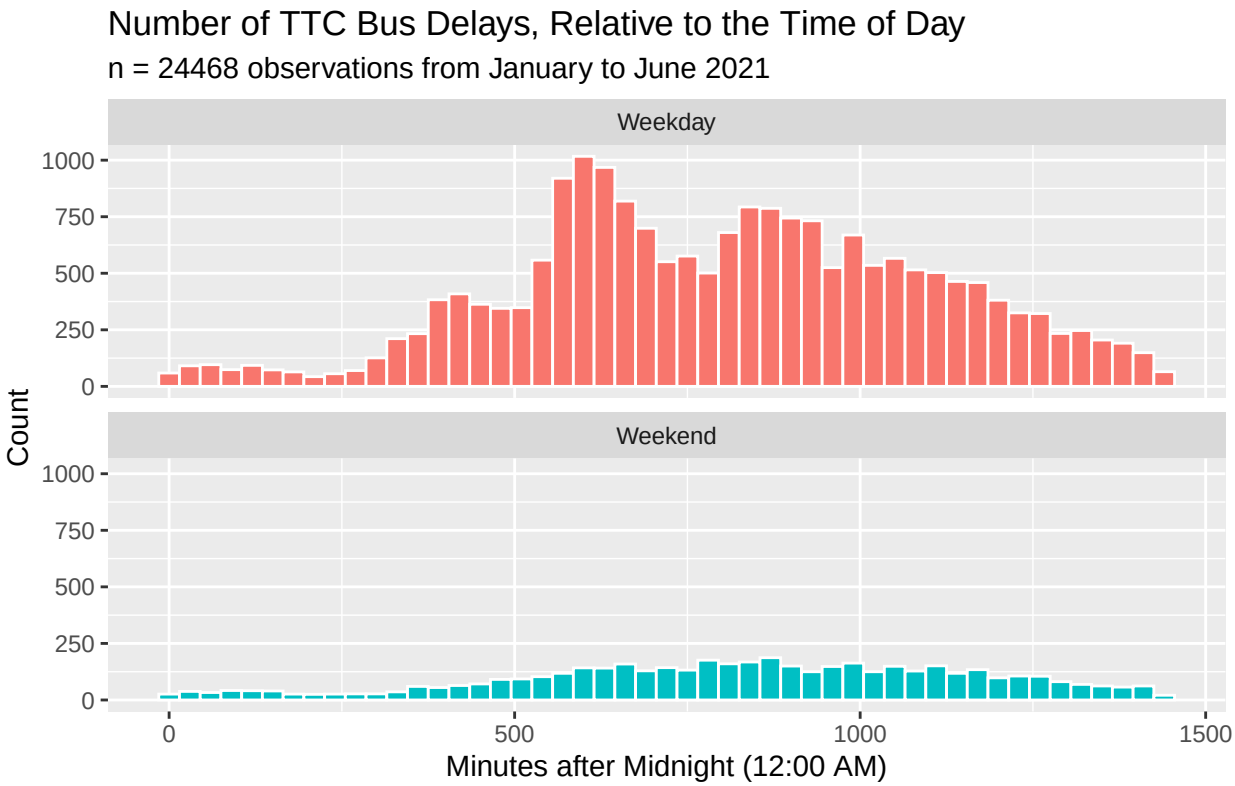


Figure 1: A comparison of the frequency of TTC bus delays at different times of the day, with frequencies computed separately for weekdays and weekends.

The plot above shows the distribution of TTC bus delays with respect to the time of day during weekdays

and weekends. On weekdays, we can see that there are clearly certain times which have more frequent bus delays; these times are between approximately 550 and 1000 minutes after midnight, representing times from about 9:00 AM to 4:45 PM. The distribution of bus delays on weekdays looks slightly bimodal, as there is a bit of a dip, and then it rises again before tapering off. Comparatively, the distribution of bus delays on the weekend is a noticeably flatter distribution; there are still some times of day which have more bus delaying incidents occur, but overall, the distribution of bus delays with respect to time on weekends is noticeably more even throughout the 24-hour span.

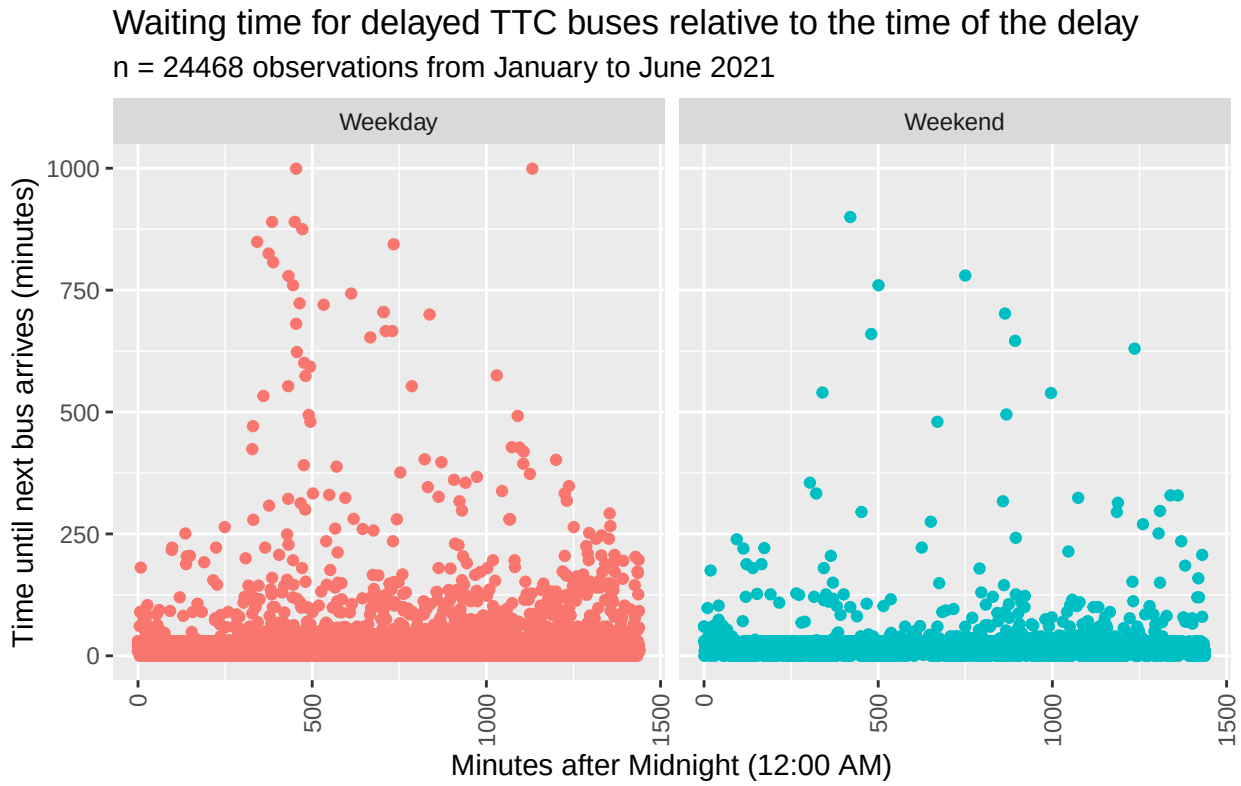


Figure 2: A plot depicting the overall waiting times for delayed TTC buses compared to the time of day of the delay, split based on whether the day of the delay was a weekday or a weekend.

On both weekdays and weekends, we see that for the most part, the majority of overall waiting times for TTC bus delays are quite short, which is shown graphically by the nearly-filled areas near the bottom of the graph.

However, in both instances, there are occasionally buses which are delayed and the amount of time before a new bus comes on that route is quite large. These occasions are depicted in the plots above with the "higher up" points on the plots. We can see that there are more of these "high" points on weekdays than weekends, but since each week has 5 weekdays and 2 weekend days, this is to be expected, as there are simply more opportunities for lengthy bus delays on weekdays by virtue of the amount of weekdays in a week.

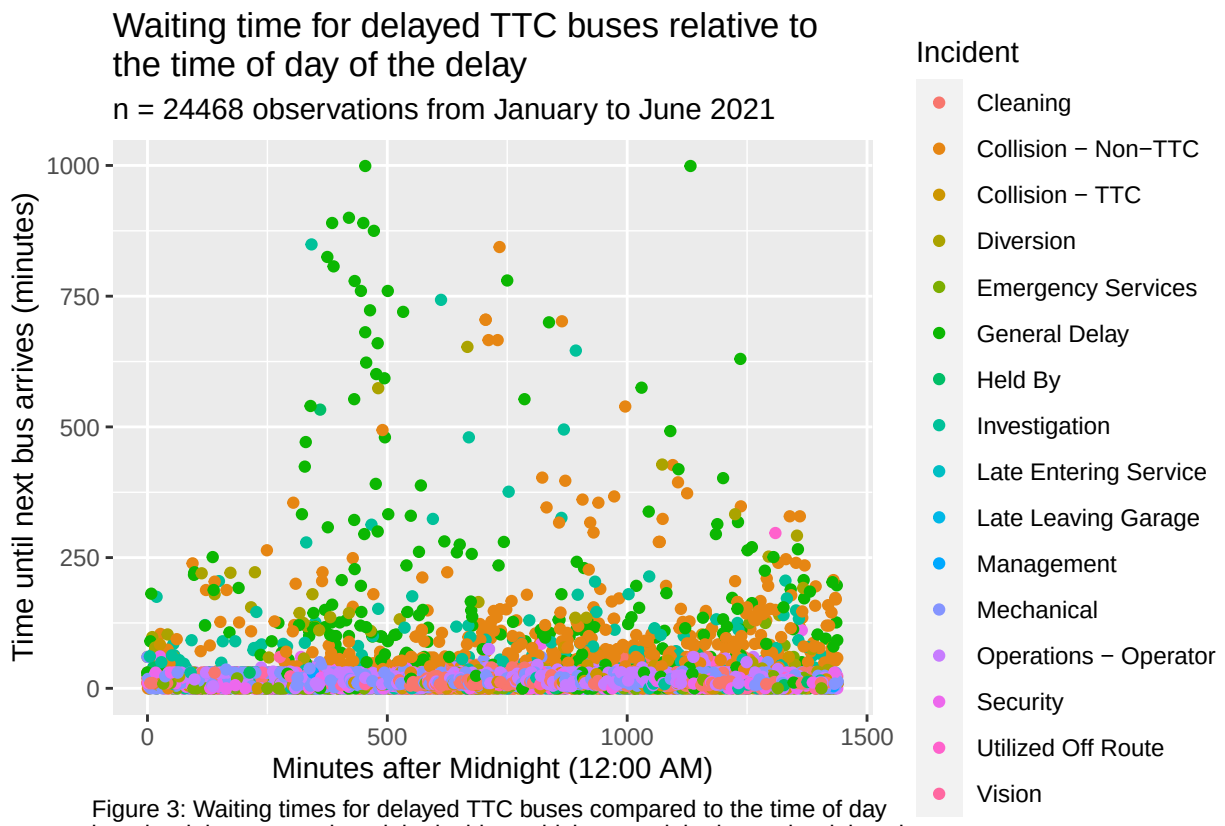


Figure 3: Waiting times for delayed TTC buses compared to the time of day when the delay occurred, and the incident which caused the bus to be delayed.

The plot above depicts the distribution of waiting times for the next bus to arrive (after a delayed bus) on a given TTC route with respect to the incident causing the delay. We can see that many of the observations which had long waiting times fell into one of two particular incident categories; General Delays, and Non-TTC Collisions.

These two categories having particularly long waiting times makes sense, as a General Delay could easily be something which may take a long time to fix, such as a problem with the roads which causes unsafe driving conditions. Similarly, a Non-TTC Collision would also likely be a lengthy ordeal, as generally, when a sever car crash occurs, first responder teams such as the police and ambulances are required to help, and roads may even be shut down, which causes a lot of traffic to be diverted, including TTC bus routes.

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

Other R packages used in the preparation of this report include `opendatatoronto version 0.1.4` (Gelfand, 2020), which was used for finding the data, `openxlsx version 4.2.4` (Schauberger, 2021), which we used to load the data from the Excel documents provided by Open Data Toronto, `car version 3.0.11` (Fox et al., 2021), which was used for certain functions, and `tidyverse version 1.3.1` (Wickham et al., 2021) which was used for managing the data throughout the report.

Additionally, the `knitr version 1.33` (Xie et al., 2021) R package was used for formatting tables which were produced in the report. This report was compiled using `MiKTeX version 21.6`. (Schenk, 2021)

Methods

Explanation

As detailed above, we plan on creating a linear regression model to predict the expected waiting time for the next TTC bus to arrive in terms of our variables of interest; the type of incident which caused the delay,

the day of the week on which the delay occurred, and the time of day at which the delay occurred.

We will be creating a linear regression model based on a frequentist approach. In this statistical approach, we will assume that there are some fixed constants which affect the expected waiting time for a TTC bus when there is a delay, and we will use the data available to us to estimate the values of these fixed constants.

A linear regression model is a function of the form $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$, where y represents the output (dependent variable) which we aim to predict, β_0 is a constant which represents the intercept of the model, i.e. the expected value of the dependent variable when all predictors are equal to 0. An equivalent expression to this model is $\mathbb{E}[y|x_1, x_2, \ldots, x_n] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$, which represents that the mean of our predicted variable is linearly dependent on the predictors $x_1, x_2, \ldots, x_n$, and for fixed values of $x_1, x_2, \ldots, x_n$, the expected value of our dependent variable is given by this formula.

The values $\beta_1, \ldots, \beta_n$ are all constants which represent the coefficients which correspond to $x_1, x_2, \ldots, x_n$, respectively. For $i \in \{1, 2, \ldots, n\}$ the coefficient β_i represents the change in the expected value of the variable y per unit change in the observed value x_i, with all other values $(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ held constant. (Sheather, 2009)

Lastly, the ϵ in the equation above represents a random error term, which allows for randomness in the resulting values. The existence of this random error explains why we can sometimes have the exact same input conditions while observing different results (values of y), as this term helps to encapsulate the effects of inherent randomness as well as external factors which are not random but are not captured in the modelling equation.

Diagnostics

Before creating a linear regression model, we must perform some diagnostic checks to ensure that the necessary assumptions for linear regression modelling are met by our data. (Sheather, 2009) In particular, we will check the four main assumptions for linear regression:

In order to perform such a linear regression, we need to do diagnostic checks on our data to ensure that the underlying assumptions for linear regression are valid. In particular, we have some key assumptions which we always make:

1. The predictors are independent: there is minimal correlation (multicollinearity) between the potential predictor variables
2. The error terms are homoscedastic: the variance of the error terms is constant, and not dependent on the values of our predictors
3. The error terms are all independent of one another: there is minimal autocorrelation between error terms
4. The error terms follow a Normal distribution

In addition to checking whether the data which we have retrieved from Open Data Toronto is suitable for linear regression, we will also examine whether any of the observations in our data are considered to be influential observations which heavily skew our predictions.

Examining Multicollinearity

The three independent (hopefully) variables which we have chosen to use as potential predictors in our linear regression model are the day of the week, the incident which caused the delay, and the time of day when the delay was reported.

As shown in the tables in the Data Cleaning sub-section above, two of these three variables (the day and the incident) are categorical variables; they are not represented by numbers, and each observation belongs to one category in a finite set.

To examine multicollinearity in a dataset which contains both numerical and categorical variables (as our data does), we can use the Generalized Variance Inflation Factor (GVIF). The GVIF is a standardized measure of the correlation between different predictors in a linear model, representing how the correlations

between a chosen predictor variable and the other predictor variables in the data affect the overall accuracy of the linear model's predicted outcomes (Fox and Monette, 1992).

Table 8: Variance Inflation Factors (VIF) of our Predictor Variables

	GVIF	Degrees of Freedom (df)	GVIFˆ(1/(2*df))
Minutes After Midnight	1.007056	1	1.003522
Type of Day	1.020160	1	1.010030
Incident	1.025468	15	1.000839

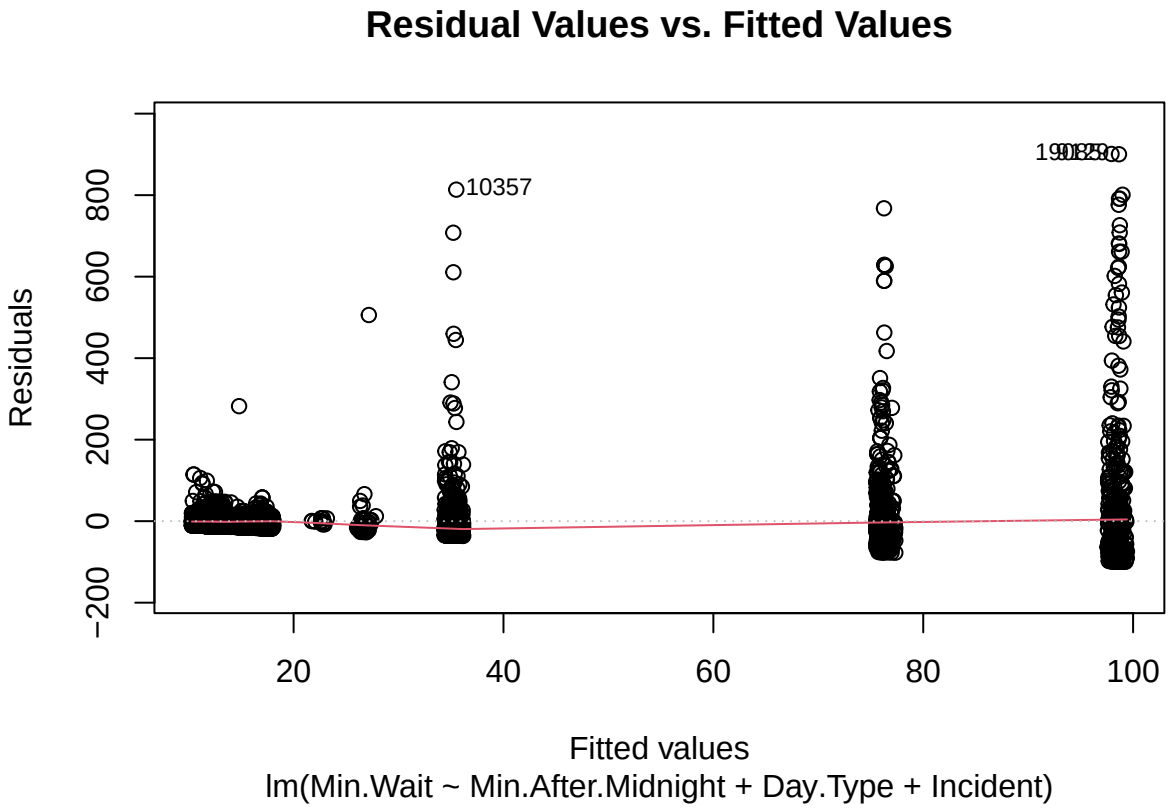
A general rule of thumb for GVIF values is that a predictor is highly correlated with at least one of the other predictors if its value of $\text{GVIF}^{\frac{1}{2\times \text{df}}}$ exceeds 2 (MsGISRocker, 2014).

As we can see in the output above, each of our variables' corresponding values are very close to 1 (the lowest possible value of $\text{GVIF}^{\frac{1}{2\times \text{df}}}$), which tells us that there is minimal correlation between any combinations of our three variables, and thus, the effect of multicollinearity in our data which could affect our linear regression models is negligible.

Examining Homoscedasticity

Next, we will examine the possibility of heteroscedasticity in our error terms, which is a measure of whether or not the standard deviation and variance in our error terms changes with respect to the fitted values. In an ideal linear regression situation, the variance of our error terms is constant with respect to the predictors and predicted values, but this is not always the case.

To examine this assumption, we will create and examine a plot of the residuals and the values 'fitted' by our linear regression model. In linear regression, a residual represents the difference between the value predicted by a model and the true observed value of the outcome variable.



In the plot above, we can see that the red curve (representing the average residual for different vales predicted by our model) is normally quite close to 0, which is a good sign that there is a linear relationship in our data, as the average residuals are quite small throughout this plot, which tells us that our linear regression model is a well-fitted model for our data and the relationship we aimed to examine.

However, we can see that the spread of the standardized residuals grows as the fitted values increase. For observations whose fitted values are below 20, a majority of these observations' residuals are close to 0, but as the value predicted by the model increases, we see that the spread and variation of the residuals is getting larger in size, which tells us that our observations do not satisfy the homoscedasticity assumption, as the variance of our error terms seems to grow as the fitted values increase, which hints at the existence of some underlying relationship between the fitted values and the variance of our error terms, also known as heteroscedasticity.

Examining Independence of Error Terms

Next, we will examine the third assumption for linear regression: the error terms are independent of one another. We saw above that the homoscedasticity assumption for linear regression does not hold water in our dataset, as the variance of these error terms does appear to be correlated with the waiting time predicted by our model. However, this assumption asserts that the error terms are independent of one another, even if they are not necessarily independent of the model's predicted values.

To examine this assumption, we will use the Durbin-Watson Test to determine the amount of autocorrelation and independence between different residuals in our data. The Durbin-Watson Test is a test which can be applied to a statistical model (such as our linear model), and the test will return a value between 0 and 4, where a value of 2 occurs when the residuals are all independent of one another. (Kenton et al., 2021)

Durbin-Watson test statistics which are between 0 and 2 imply that the residuals have negative autocorrelation (if the wait time for a certain delayed TTC bus is short, then the wait time for the next delayed TTC

bus will be long), whereas test statistics between 2 and 4 imply that there is positive autocorrelation (if the wait time for a delayed TTC bus is short, then the wait time for the next bus will also be short) among the residuals in the data. (Kenton et al., 2021)

A general rule of thumb is that a Durbin-Watson test statistic between 1.5 and 2.5 is considered to mean that there is minor autocorrelation, but the autocorrelation present in the model is not considered to be a major problem . (Statology, 2021)

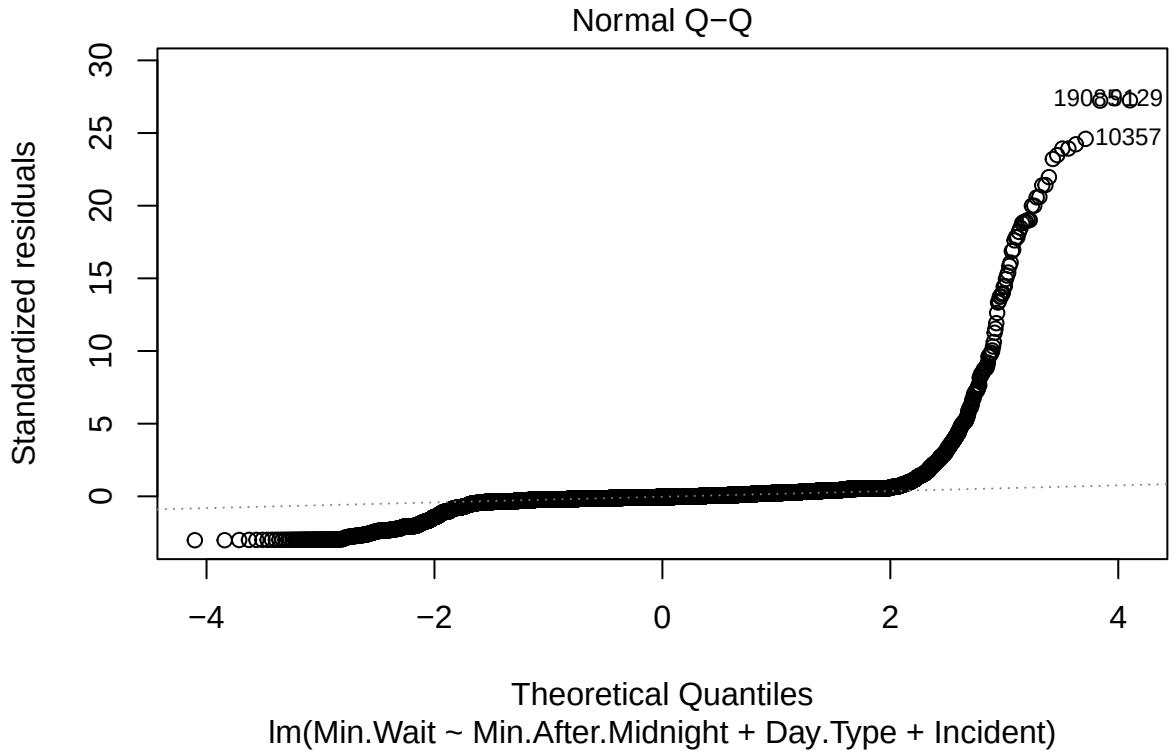
The Durbin-Watson test statistic for our observed data is approximately 1.923, and the p-value for our observed data is 0, which tells us that the probability of significant autocorrelation in our data is negligible, as the p-value for the possibility of autocorrelation is an incredibly small value. Based on the Durbin-Watson test, the error terms in our data are almost entirely independent of one another; there is a slight amount of negative autocorrelation, but this is likely just a statistical quirk in our data.

Examining Normality of Error Terms

Next, we will examine whether the residual error terms (the differences between our model's predicted outcomes and the real observed values of the predicted variable) in our model follow a Normal distribution.

Unlike the first three assumptions, this last assumption is not strictly required in order to create a linear regression model, but it is useful to have as Normally-distributed error terms greatly simplify the computation of the predicted outcomes in terms of the inputs.

To check whether this assumption holds, we will create a Q-Q plot (quantile-quantile plot) to compare the distribution of our linear model's standardized residuals to the distribution of elements which follow a Normal distribution.



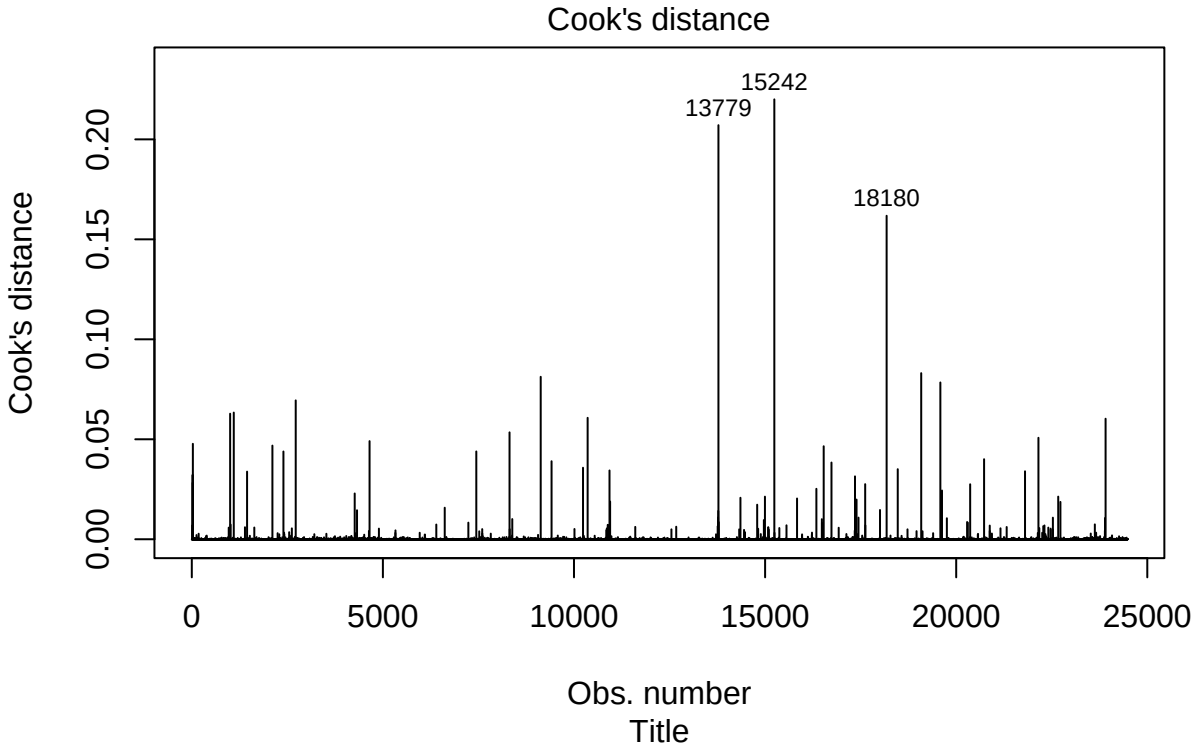
If the error terms were Normally distributed as we had hoped, then the observations in the plot above would be similar to the dotted line shown on the plot above. (Sheather, 2009) For the Theoretical Quantiles between

approximately -2 and 2, the standardized residuals produced by our model fit this dotted line extremely well, but it is clear that the standardized residuals produced by our model deviate from the quantiles of a Normal distribution in these 'outer reaches'.

Detecting Influential Observations

Lastly, we will examine if any of our observations are considered to be influential observations which significantly affect the predictions made by our model.

To determine points which are considered to be excessively influential observations on our data, we will use the Cook's Distance as a measure of influence. The Cook's Distance of an observation is a measure which corresponds to how much removing this observation will change the predictions made by our model (Cook and Weisberg, 1982; Sheather, 2009)



In the plot above, we can see that among our nearly 25000 observations, the largest Cook's Distance values are slightly over 0.20. A common suggested value as the cutoff for determining whether an observation in a dataset is overly influential is the 50th percentile of the F-distribution with p and $n - p$ degrees of freedom (Cook and Weisberg, 1982), where p is the number of predictors in the dataset, and n represents the number of observations in the dataset.

Using the built-in `qf` function to determine the 50th percentile of an F-distribution, we find that the cutoff for an influential observation in our dataset is any observation whose Cook's Distance exceeds 0.965, which is significantly higher than any of the Cook's Distances seen in our data. Thus, there are no significantly influential observations in our data.

Model Selection

Next, we will choose a suitable model for predicting the expected waiting time when a TTC bus delay occurs. We have three potential predictor variables in our data; the number of minutes since midnight, the incident which caused the TTC bus to be delayed, and whether the delay occurred on a weekday or a weekend.

We will create seven distinct linear regression models, by creating a model for every possible choice of our three predictors (or choosing only some of the predictors), and we will then select our 'final' model based on the most suitable of these seven models.

In order to choose which model we will use, we will judge these models based on their adjusted R-squared value. The R-squared value of a regression model is a value between 0 and 1, which represents the ratio of the total variation in the outcome variable to the variance which is explained by the regression model. (Mahendru, 2019)

A problem with the R-squared value is that as more predictors are added to a model, the R-squared value always increases, even if the new variables are not necessary. The adjusted R-squared value of a regression model is similar to the R-squared value, but it introduces a penalty term to help ensure that the adjusted R-squared value only increases if the newly-added predictors improve the overall model. (Sheather, 2009) Thus, using the adjusted R-squared will allow us to compare these models which have different numbers of predictors, instead of the three-predictor model automatically being selected by virtue of having the highest number of predictors.

Table 9: Summaries of our seven linear regression models.

Predictor Variables	Adjusted R-squared
Minutes After Midnight, Type of Day, and Incident	0.20052
Type of Day and Incident	0.20048
Minutes After Midnight and Incident	0.20054
Minutes After Midnight and Type of Day	0.00045
Minutes After Midnight	8.5291e-05
Type of Day	0.00035
Incident	0.2005

In the table above, we can see that the linear regression model which achieved the highest adjusted R-squared value is the model which used the incident and the time of day as predictors, but not whether the TTC delay occurred on a weekday or a weekend.

However, we can see that the adjusted R-squared value of the model which uses the incident alone is only slightly lower than this model's adjusted R-squared value, which informs us that while the time of day is a (somewhat) helpful predictor, the delay-causing incident does a majority of the "heavy lifting" when it comes to making these predictions.

Results

Now that we have selected our model, we can examine the coefficients of the model to determine how changes in the input parameters (time since midnight, and the type of incident) affect the expected waiting time for a delayed TTC bus.

In the table shown above, there are 17 variables, each of which has a different respective coefficient. Of the 17 variables, one of these variables refers to the intercept of our linear regression formula, one variable refers to the time of day when the delay occurred, and the remaining 15 variables are all related to different possible delays which could have occurred and caused the bus to be delayed.

Although there are 15 variables above which refer to different possible delays that could occur, there were actually 16 different delays in our data. The reason for this discrepancy is that when we fit a linear regression

with a categorical variable, one possible category is selected to be a 'default' category, and other categories are compared to this default. For the data which we are examining, this 'default' delay is a delay due to "Cleaning", and so TTC bus delays due to "Cleaning" represent the baseline for delayed buses.

The first variable, "Intercept", refers to the expected waiting time for the next TTC bus to arrive when a delay occurs, if the time of day is 0 minutes past midnight, and the reason for the delay is "Cleaning".

The second coefficient represents the change in the expected waiting time for the next TTC bus to arrive per unit change in Minutes After Midnight, assuming that other factors (the cause of the delay) are held constant. In essence, this coefficient represents that if there is a TTC bus which has experienced some delay, then for each additional minute after midnight when the delay occurred, the expected waiting time for the next bus to arrive decreases by 0.001.

The remaining 15 coefficients all refer to the change in expected waiting time for a delayed TTC bus due to different incidents which could cause the delay. If the delay is one of these 15 variables, then we will add the coefficient corresponding to this variable to the expected delay time. If the delay is not one of the 15 categories with a corresponding coefficient in the table above, then we will not change the value which is given by the other two terms of this linear regression model. Note that a delay can only be caused by exactly one incident, which means that we will only ever add at most one of these 15 coefficients to the value predicted by the intercept and the time of the delay.

For example, suppose that at 8:30 AM (510 minutes after midnight), a TTC bus is delayed, and the reason for the delay is an Investigation. In this case, our linear model predicts that the expected waiting time for the next bus to arrive at this stop is $13.45 + (-0.001)(510) + 22.475 = 35.415$ minutes.

In the equation above, the value 13.45 comes from the intercept term, the value $(-0.001)(510)$ is because 8:30 AM is 510 minutes after midnight, and the coefficient corresponding to this variable is -0.001, and the value 22.475 arises because the cause of the bus's delay was an investigation, which has a corresponding coefficient of 22.475. Adding these values together, we arrive at an estimated waiting time of 35.415 minutes for the next TTC bus to arrive at the stop.

Overall, these coefficients are all reasonably interpretable and make sense within the context of the value which we are trying to predict. Based on these coefficients, our model cannot predict a negative time until the next bus arrives, which matches our logical intuition of buses arriving in the future, and based on the model's coefficients, we see that different types of delays can yield massive changes in the expected waiting time for a bus. This also matches our expectation of bus delays, since a bus which is delayed due to a mechanical failure with the vehicle should logically take longer to reach the next stop than a bus which simply got caught in traffic.

Additionally, we see that there is a very slight negative linear relationship between the time after midnight and the expected waiting time for a delayed bus. This also makes sense (and matches our hypothesis), as we expected the TTC's service to be quicker and more efficient during the daytime, and less frequent at night.

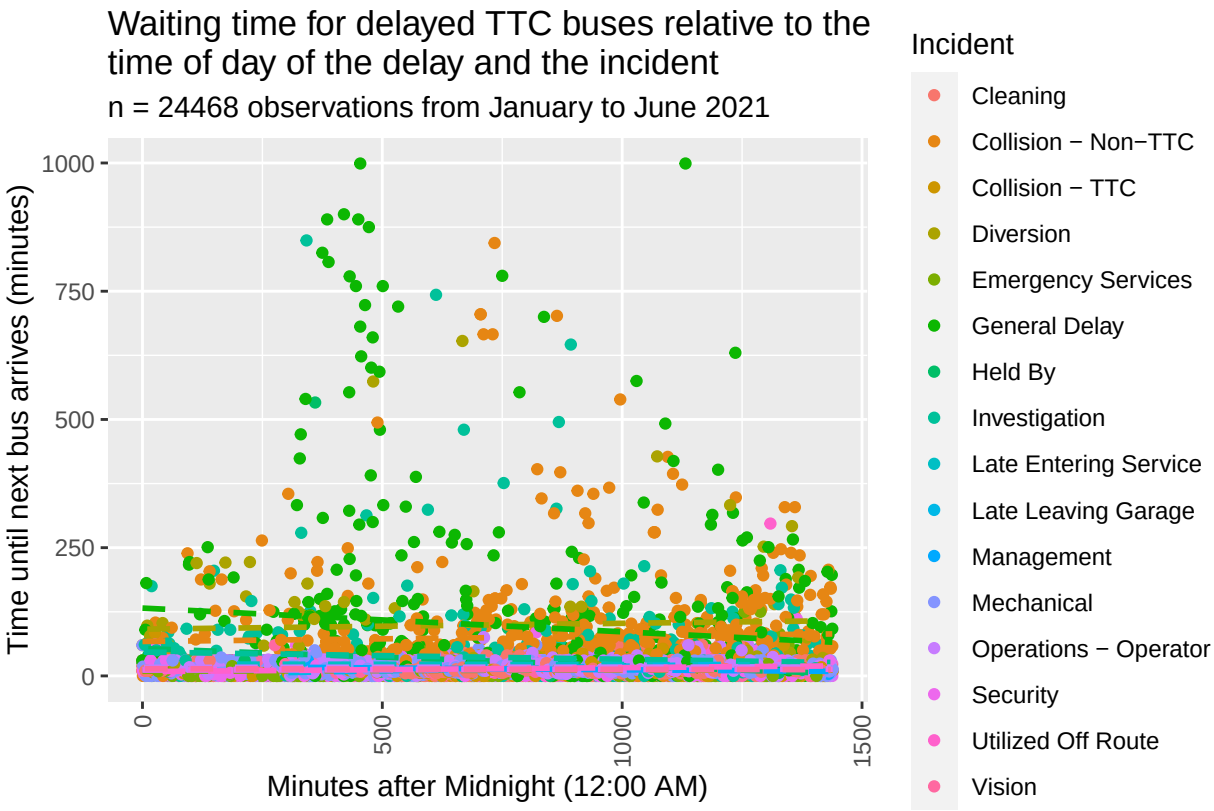


Figure 5: Waiting times for delayed TTC buses compared to the time of day when the delay occurred, and the incident which caused the delay.

The dashed lines represent our linear model's predicted waiting times, given the time of day of the delay,and the incident which occurred.

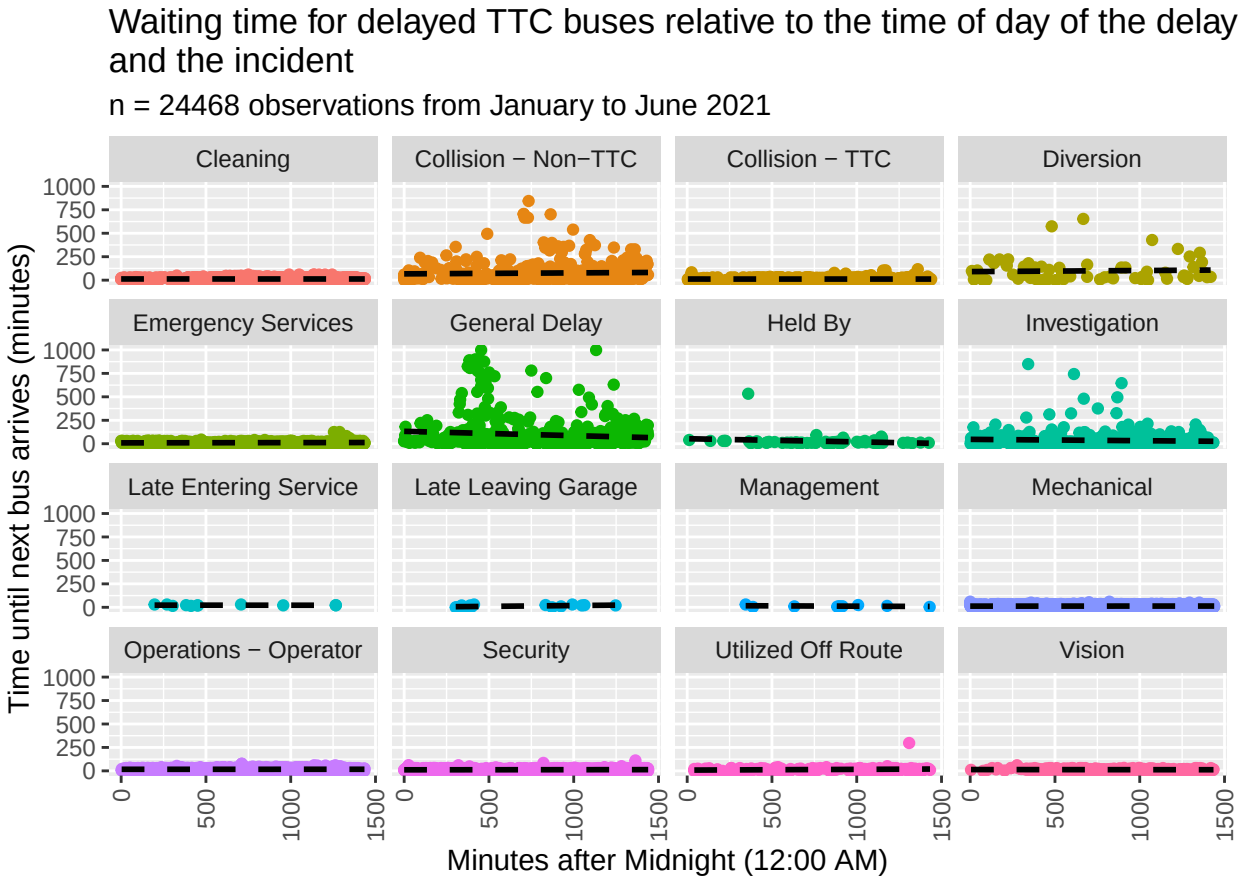


Figure 6: Waiting times for delayed TTC buses compared to the time of the delay, and the incident. The dashed black lines represent our linear model's predicted values given the time and type of delay.

Conclusions

Using the data which we had collected from the TTC via the Open Toronto Data Portal, the main question which we aimed to answer through our research was "*When will the next bus arrive at my stop?*"

In addition to answering this question, we also wanted to determine how different factors such as the day of the week, the incident causing the delay, and the time of day affected these waiting times. We hypothesized that these waiting times would be longer on weekends than on weekdays, and that they would also be longer during nighttime hours, while the delays would be shorter during the day, as many people require the TTC to get around the city.

We chose to fit a linear model to the data, and we tested multiple different combinations of predictors in order to choose the model which had the best predictive ability, in terms of its adjusted R-squared value. The model which we selected had an adjusted R-squared value of just over 0.2, and an R-squared (non-adjusted) value of 0.201, which is not a particularly high adjusted R-squared value for a linear regression model, as this means that nearly 80% of the variance in our observed values is not explained by the model's predictors.

When we examined and compared the different possible linear models which we could construct to predict the wait time for the next bus, we found that whether the day of the delay was a weekday or a weekend was almost entirely inconsequential, and hardly affected the reliability of our linear models' predictions. In fact, the linear model which included whether the delay fell on a weekday or a weekend had a lower adjusted R-squared value than the model which excluded this point from the data! This predictor might not be completely useless, as there likely is some small change in waiting times during weekends and weekdays (which may simply be statistical noise in our data), but including this predictor would make the model

significantly more complicated and the benefit would be a negligible increase in the predictive quality of the model. The expected waiting time for the next bus may not be completely independent of whether the delay occurred on a weekday or a weekend, but the effect of the day of the week is negligible, which goes against our hypothesized result.

We also found that there is a very small negative linear relationship between the time of day (in minutes after midnight) and the waiting time for delayed TTC buses. We included this predictor in our final choice of model because the linear model including this predictor exhibited a slightly higher adjusted R-squared than the linear model which excluded this predictor. As such, the time of day is a somewhat helpful predictor for predicting the expected waiting time for a TTC bus, but it is not as important as the type of incident which caused the delay to occur.

Overall, the linear model which we created has an adjusted R-squared value of 0.2, which is not indicative of this linear model being a good fit for the data which we have collected. This is not incredibly surprising, as the data which we had collected did not fully meet all of the linear regression assumptions, so it is not surprising that our results cannot be predicted accurately by a simple linear model.

Weaknesses

The largest weakness in this report is that the data which we examined does not appear to fit a linear regression model, as the data which we have examined does not meet all of the linear regression assumptions which are necessary for 'safely' performing a linear regression. Furthermore, we know that a linear regression model in terms of our three chosen variables is not sufficient for predicting expected wait times for TTC buses which are delayed, and a different choice of predictive model may have been better suited for the data at hand.

Another obstacle which we faced during this report was a lack of clarity surrounding the data which we collected from the Open Toronto Data Portal. On the Open Toronto website, it says that the data was posted by the Toronto Transit Commission, which means that the data is from a primary source. However, there is not any information on the Open Toronto website, or other websites such as the TTC's website about how the data were collected, measured, and handled. Because of this lack of clarity with regard to how the data was managed, we needed to assume that the data which we collected was representative of all TTC bus delays from January to June 2021, but this is a strong assumption to make, and it would be helpful to have more information about the TTC's standards with respect to their data collection processes.

Next Steps

Based on the linear models which we fitted and examined in this report, we can see that the wait times for a TTC bus cannot be represented well as a linear function in terms of our chosen predictors. Although we know that our data does not necessarily fit a linear model, this does not mean that there is no relationship at all between the expected waiting time at a bus stop and factors such as the time of day when the delay occurred; it simply means that this relationship is likely non-linear. In a future report on TTC bus delay data, it may be interesting to examine the possibility of non-linear relationships between the time of day and the expected waiting time, as we have simply shown that these values are not linearly related to one another, but a different relationship could still exist between these values, and the time of day could potentially be a strong predictor for the expected waiting time.

Some of the results which we have found seem quite promising, for example, different incidents which cause TTC buses to be delayed have vastly different expected waiting times for the arrival of a bus in the future. Another interesting factor which we could examine in future reports on this data (or other similar TTC data) would be to examine whether there are any higher order interaction terms (Sheather, 2009) which could be used for creating a more promising predictive model to predict our outcomes.

Lastly, there are other variables which are not present in the dataset which we retrieved from Open Data Toronto which could be explanatory variables for the expected waiting times. Two possible explanatory variables which we did not have access to are traffic patterns at a certain day/time, and the weather in Toronto at a given day and time. Although these additional predictors may be correlated with some of the

predictors which we have already examined, there is potential for these predictors to help us gain more insight into TTC bus delays, and attempting to include these variables in a future model for these observations may prove effective.

Discussion

Based on our investigation into the TTC's bus delay data, we found that the type of incident which caused the bus to be delayed is a useful predictor in determining how long it will take until the next TTC bus on that route arrives at a stop.

However, some of the other variables which we had investigated, such as whether the delay was on a weekday or a weekend proved to be ineffective as factors which could be used to predict the duration of time before the next bus arrived. This is interesting in its own right, as we had hypothesized that the quality of service on the TTC would be better during the week than on weekends, but this does not appear to be the case.

We also found that there is a very minute linear relationship between the time of day and the expected waiting time for the next bus to arrive, but this linear relationship may actually be due to some statistical noise. However, unlike the weekend/weekday categories, where it is clear that there is not a difference in the expected wait time across these two categories, there still could be a non-linear relationship between the time of day when the delay occurred and the waiting time required to wait for a bus.

Overall, the expected waiting time for a bus to arrive when a bus on that route is delayed does not appear to be a linear function of the predictors which we had examined. The only things which we are able to say with any great certainty are that the TTC appears to be quite consistent in its timing throughout the week, while there is a large amount of variation in terms of this expected waiting time with respect to the different incidents which can cause delays.

Bibliography

- Andersen-Wood, F. (n.d.). Occam's Razor. Retrieved October 25, 2021, from https://conceptually.org/concepts/occams-razor
- APTA 2017 Award. (2017). Retrieved from https://ttc-cdn.azureedge.net/-/media/Project/TTC/DevProto/Documents/Home/About-the-TTC/APTA_Award_2017.pdf
- Basa, E. (2018, January 26). Toronto's Transit System Is Apparently Ranked The 3rd Best In North America. Retrieved from https://www.narcity.com/toronto/torontos-transit-system-is-apparently-ranked-the-3rd-best-in-north-america
- Cook, R. Dennis; Weisberg, Sanford. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall. Retrieved from the University of Minnesota Digital Conservancy, https://hdl.handle.net/11299/37076.
- Draaisma, M. (2017, June 27). 'No one saying it's perfect': TTC named best transit system in North America. Retrieved from https://www.cbc.ca/news/canada/toronto/programs/metromorning/ttc-named-best-transit-system-north-america-1.4178574
- Ebner, J. (2018, November 12). How to use to facet_wrap in ggplot2. Retrieved October 19, 2021, from https://www.sharpsightlabs.com/blog/facet_wrap/
- Fox, C. (2021, September 15). TTC ridership could linger below pre-pandemic levels for at least another two years: Report. Retrieved October 22, 2021, from https://www.cp24.com/news/ttc-ridership-could-linger-below-pre-pandemic-levels-for-at-least-another-two-years-report-1.5586747
- Fox, John, and Georges Monette. "Generalized Collinearity Diagnostics." *Journal of the American Statistical Association*, vol. 87, no. 417, [American Statistical Association, Taylor & Francis, Ltd.], 1992, pp. 178–83, https://doi.org/10.2307/2290467.
- Gibbs, A., Stringer, A., & Caetano, S. (2021). Probability, Statistics, and Data Analysis. Retrieved October 18, 2021, from https://awstringer1.github.io/sta238-book/
- Grolemund, G. (2014, July 16). *Introduction to R Markdown*. Retrieved September 30, 2021, from https://rmarkdown.rstudio.com/articles_intro.html
- Hutt, A. (2021, October 21). *TTC Bus Delay Data*. Retrieved October 15, 2021, from https://open.toronto.ca/dataset/ttc-bus-delay-data/.
- karpfen. (2017, June 30). How to set colour for toc links in R markdown document? Retrieved September 28, 2021, from https://stackoverflow.com/questions/44846350/how-to-set-colour-for-toc-links-in-r-markdown-document
- Kenton, W. (2021, August 4). Durbin Watson Statistic Definition (S. Anderson & A. Courage, Eds.). Retrieved October 24, 2021, from https://www.investopedia.com/terms/d/durbin-watson-statistic.asp
- Khan, S. (2020, September 12). *TTC Delay Data Analysis* [Scholarly project]. In Safi Khan. Retrieved from https://safikhan.ca/
- Leary, R. J. (2021, May). *CEO's Report*. Retrieved from https://transittoronto.ca/archives/.
- Mahendru, K. (2019, June 21). Measuring the Goodness of Fit: R2 versus Adjusted R2. Retrieved October 25, 2021, from https://medium.com/analytics-vidhya/measuring-the-goodness-of-fit-r%C2%B2-versus-adjusted-r%C2%B2-1e8ed0b5784a
- MsGISRocker and Zhao, X. (2020, February 25). Which variance inflation factor should I be using? Retrieved October 23, 2021, from https://stats.stackexchange.com/questions/70679/which-variance-inflation-factor-should-i-be-using-textgvif-or-textgvif
- Sauer, S. (2017, September 08). Different ways to count NAs over multiple columns. Retrieved from https://sebastiansauer.github.io/sum-isna/
- Sheather, S. J. (2009). *A Modern Approach to Regression with R*. New York City, NY: Springer Science Business Media. doi:10.1007/978-0-387-09608-7
- Spurr, B. (2017, June 26). TTC named best public transit agency in North America. Retrieved October 20, 2021, from https://www.thestar.com/news/city_hall/2017/06/26/ttc-named-best-public-transit-agency-in-north-america.html
- Substrings of a Character Vector. (n.d.). Retrieved October 20, 2021, from https://stat.ethz.ch/R-manual/R-devel/library/base/html/substr.html
- Tables. (n.d.). Retrieved October 1, 2021, from https://rmarkdown.rstudio.com/lesson-7.html
- *TTC System Map - October 2021* [PDF]. (2021, October). Toronto: Toronto Transit Commission.

- Zach. (2021, January 21). The Durbin-Watson Test: Definition & Example. Retrieved October 23, 2021, from https://www.statology.org/durbin-watson-test/
- Zach. (2021, March 29). How to Calculate Variance Inflation Factor (VIF) in R. Retrieved October 23, 2021, from https://www.statology.org/variance-inflation-factor-r/