School of Continuing Studies
Foundations of Data Science

INSIGHTS INTO TTC BUS DELAY DATA
(GROUP 20)

Group Project

April 1, 2019

Authors:

Sakshi Sharma, Aaron Fernandes, Christopher Dennis, Sadaf Sadeghian, and Rahim Jiwa

TABLE OF CONTENTS

1.0 Introduction

Transit delays in Toronto cost the city almost a month of lost service every year[4]. Much of time passengers could find alternate routes and/or means of transit if they could be made aware of these delays in advance. However, by definition delays are unexpected and passengers will not realize that a delay has occurred until they are en route and the delay actually happens. But what if delays although unexpected are not in fact random? What if delays occur in patterns or during certain weather events? If this is the case it should be possible to use historical data analysis to allow city planners to direct resources more efficiently. As well it should also be possible to inform travelers when delays are most likely to occur.

1.1 Scope

We will be looking at historical data from the period of 2014 and ending in the current year 2019. We are restricting our analysis to TTC bus data, this data is easily available and different routes can be compared to each other more easily. We have restricted ourselves to the last 5 years because a) The data more available b) Changes in routes and bus models make predictions less accurate the further into the past we look.

For this project we are using the city of Toronto's open data catalog of TTC Bus delay data[2]. In particular we are looking at the data from the years 2014-2019. In addition we used data from Weatherstats.ca[3] to correlate weather events with TTC Bus delays.

We chose to use the city of Toronto's open data project for the TTC delay data because of its availability and the fact that it is the authoritative source for TTC data.

Weatherstats.ca was used because it is a reliable source of weather data for Canada and the data it produces is rich and easy to parse. One fortunate consequence of picking both of these data sources is that almost no cleaning need be performed, the records are extremely complete. The TTC data is formatted according to open data standards. In the case of Weatherstats.ca the dataset is complete by its nature.

Even parsing out the data turns out to be fairly simple given that both datasets are formatted in a similar way. This enhances our ability to make accurate predictions about future outcomes.

We used a number of different methods to obtain our results. As an initial first attempt we can use visualization to compare our data sets month by month to see if an obvious outlier months appear. As well we can examine the types of diversions which are most likely to occur and what routes are most likely to experience delays.

As an additional analysis we will attempt to use the "K Nearest Neighbor" algorithm attempt to proactively identify a delay. We will use the data from the years 2014 to 2017 to train our model and then test it against data from 2018 to 2019.

This could be an invaluable tool to help warn passengers of potential incidents before they even occur. If certain conditions point to a high likelihood that a delay will occur we can use this information to alert passengers of possible incidents.

1.2 Objectives

We hope to use data science to determine the following:

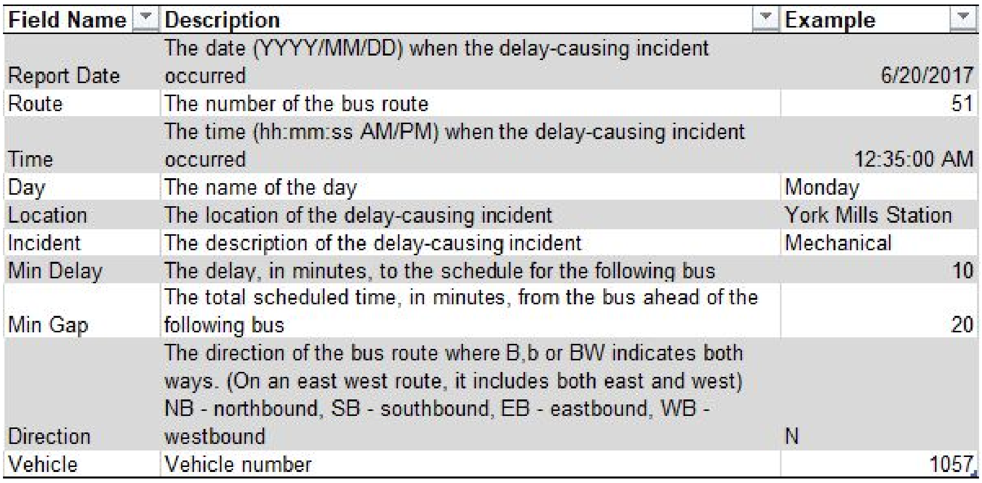
1. What were the leading causes that lead to delays in TTC service?
2. When incidents are most likely to occur (month/day/season)?
3. Which routes have the most frequent delays?
4. What weather conditions lead to delays most frequently?
5. Can we predict future delays from the known dataset.

We hope that by analyzing TTC bus delay data we can develop a model of when and how buses are most likely to break down. By doing so we can predict when future breakdowns are most likely to occur given the appropriate conditions.

2.0 Data Preparation

2.1 Data Quality

The fields available in the data set are summarized in Figure 2-1. The data was available in the city of Toronto's open data project. The data was organized as multiple excel sheets (by month) in multiple files (one per year). The data was available from Jan 2014 to Feb 2019.



Field Name	Description	Example
Report Date	The date (YYYY/MM/DD) when the delay-causing incident occurred	6/20/2017
Route	The number of the bus route	51
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
Location	The location of the delay-causing incident	York Mills Station
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
Direction	The direction of the bus route where B,b or BW indicates both ways. (On an east west route, it includes both east and west) NB - northbound, SB - southbound, EB - eastbound, WB - westbound	N
Vehicle	Vehicle number	1057

Figure 2-1 - Summary of Fields in the data set (TTC Bus Delay)

As a first step, the data was imported in a single dataframe from multiple excel files and associated sheets. This data was then exported as a csv. The primary reason was to increase the execution speed of the code. The excel import into dataframe was slower when compared to the importing of a csv file. This file was used for analysis done in subsequent sections.

The influence of weather on bus delays is considered herein. The weather data was obtained from the Federal Government Climate website[5] in CSV form and was fairly comprehensive.

The weather data was available on a daily basis from 2014-2019, and so was read into a dataframe. A subset of the weather data was generated, as a select few columns contained the data considered valuable for analysis. Data which was omitted included day numbers within a year and all wind data as wind was assumed to not influence bus travel. “Snow on Ground” was also omitted as snow itself was included and the snow data was much more complete. The data which was included related to temperatures and precipitations.

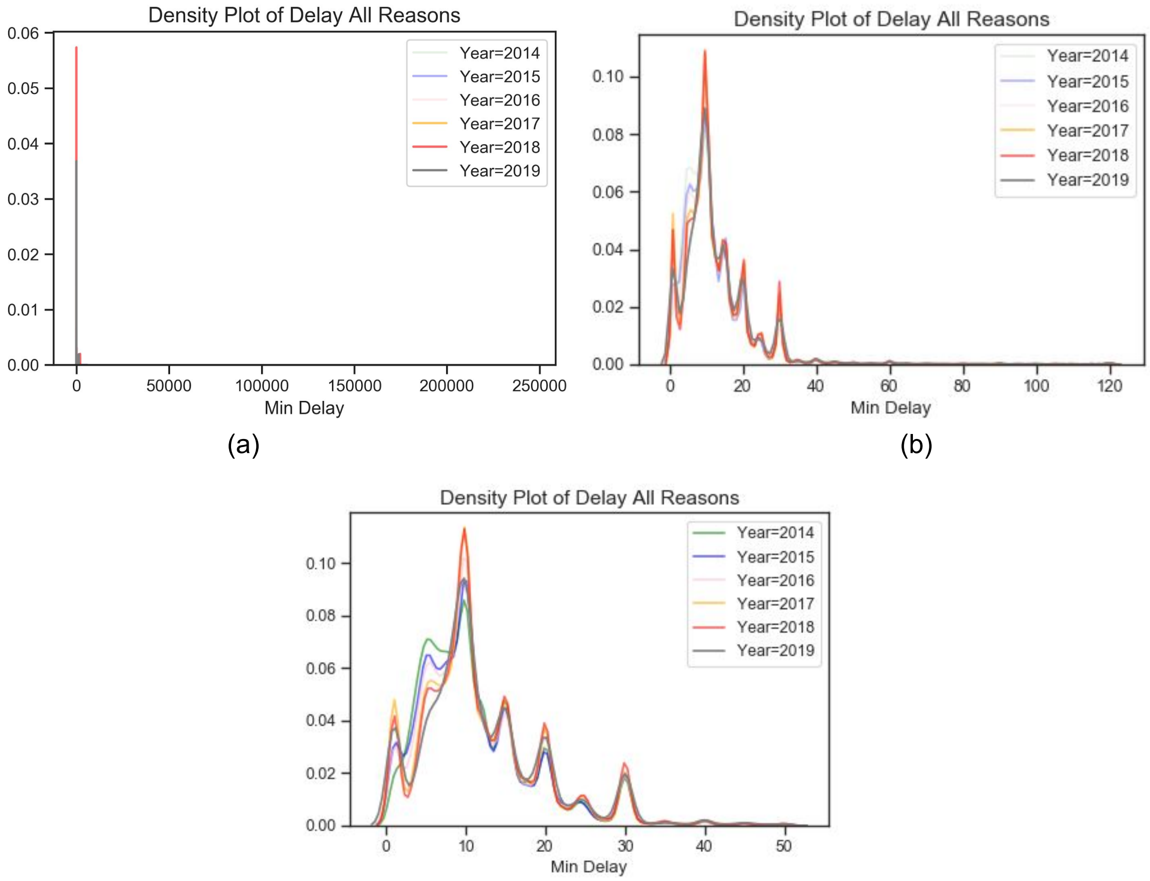
Bus data was taken from the main bus data files, namely the average number of incidents for each day in the time range and also the average delay time for all of the incidents on each of those days. This bus data was merged into the original weather data to create the main dataframe for analysis.

2.2 Challenges

One of the challenges that we faced was a distribution plot of delay data had a significant **positive** skew i.e. long tail is on the positive side of the peak. The mean is on the right of the peak value. This is illustrated in Figure 2-2(a) and posed a challenge for a meaningful analysis.

A filter was therefore applied on this data to focus on the data that had majority of the data. A filter of 120 minutes on the delay still had significant tail on the right with not many data points (Figure 2-2(b)). In addition, this filter resulted in an exclusion of just 2.5% data points.

The filter was further narrowed to 50 minutes to address the skew. A distribution plot of the delay measured variable is shown in Figure 2-2(c). The distribution plot indicates a multi-modal data set. In addition, this filter resulted in an exclusion of just 3.7% data points. Therefore, this filter was considered appropriate for further analysis. The filtered dataset suggests an average delay of 11.6 minutes across TTC service since 2014.



(a)

(b)

(c)

Figure 2-2- (a)- Density plot of all data from 2015 to 2019, (b) Density plot of all data from 2014 to 2019 (Filtered Data Set, Min Delay < 120). (c) Density plot of all data from 2014 to 2019 (Filtered Data Set, Min Delay < 50)

2.3 Feature Creation

2.3.1 Session Category

The bus delay is observed during different sessions throughout the day. The time column in the dataset has been used to create this feature. These sessions are defined as follows:

- Morning: 6:00 am to 12:00 pm
- Afternoon: 12:00 pm to 16:00 pm
- Evening: 16:00 pm to 20:00 pm
- Night: 20:00 pm to 12:00am
- Midnight: 12:00 am to 6:00 am

2.3.2 Route Categories

The route column in the dataset has been used to create this feature. The following categories were created from the route numbers specified in the TTC website[1]: Regular Bus Routes, Express Bus Routes, Downtown Express Bus Routes, Night Bus Routes, Subway Routes, Community Routes, and Other (Uncategorized Routes).

2.3.3 Season Categories

The month data in the reported date is further used to create another feature which infers the data into following season- winter, summer, spring and fall.

3.0 Analysis

3.1 Categorical Analysis

This section discusses the reasons of failure as reported by TTC during the years from 2014 to 2019. The data identifies the following primary reasons of delay - Mechanical, Late Leaving Garage, Utilized off-route, Diversion, Emergency Services, Investigation and General Delay.

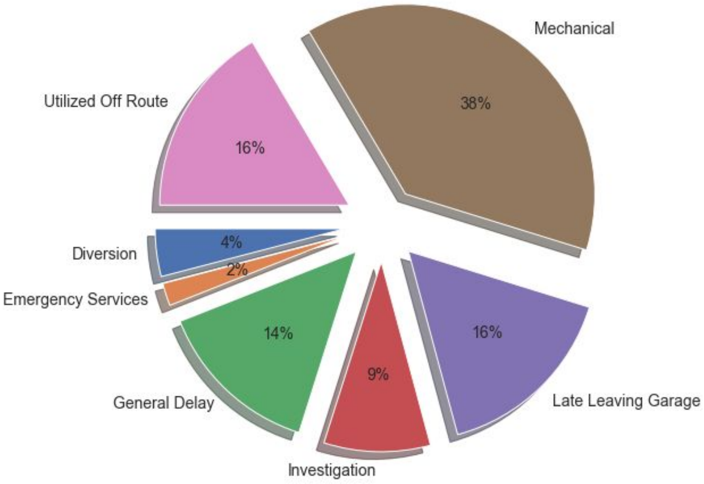


Figure 3-1-1- Pie Plot Illustrating Reasons of TTC Delay

Figure 3-1-1 illustrates the primary reasons for delay in the TTC service. The main cause of delays are “mechanical”, followed by “vehicle utilized off-route” and “late leaving garage”. Mechanical is understandably the largest reason of delays in TTC system. “Utilized off-route” can be considered as TTC trying to increase efficiencies in the system by using vehicles either where the demand is higher than the current route or in scenarios where the demand is lower on the current route. An interesting cause of delay is listed as vehicles leaving late from the garage. These above reasons are completely within TTC's control.

Other reasons that cause delays and can be considered outside of TTC's control. These are as follows -

1. Diversion - most of the diversions in the TTC service stems from either accidents or construction events etc. leading to delays in the service
2. Emergency Services - These delays are due to the passengers using the TTC service
3. Investigations - These are gain most likely from the service being used by the passengers.
4. General Delay - It is assumed that these delays stem out from delays due to general increase of traffic or road delays.

Figure 3-2-2 shows the statistical details associated with all the possible incidents resulting in TTC delays. Note that diversion is taken out of the analysis for this section as it is considered completely out of TTC's control. Table 3-1-1 quantifies the mean delays due to all the incidents.

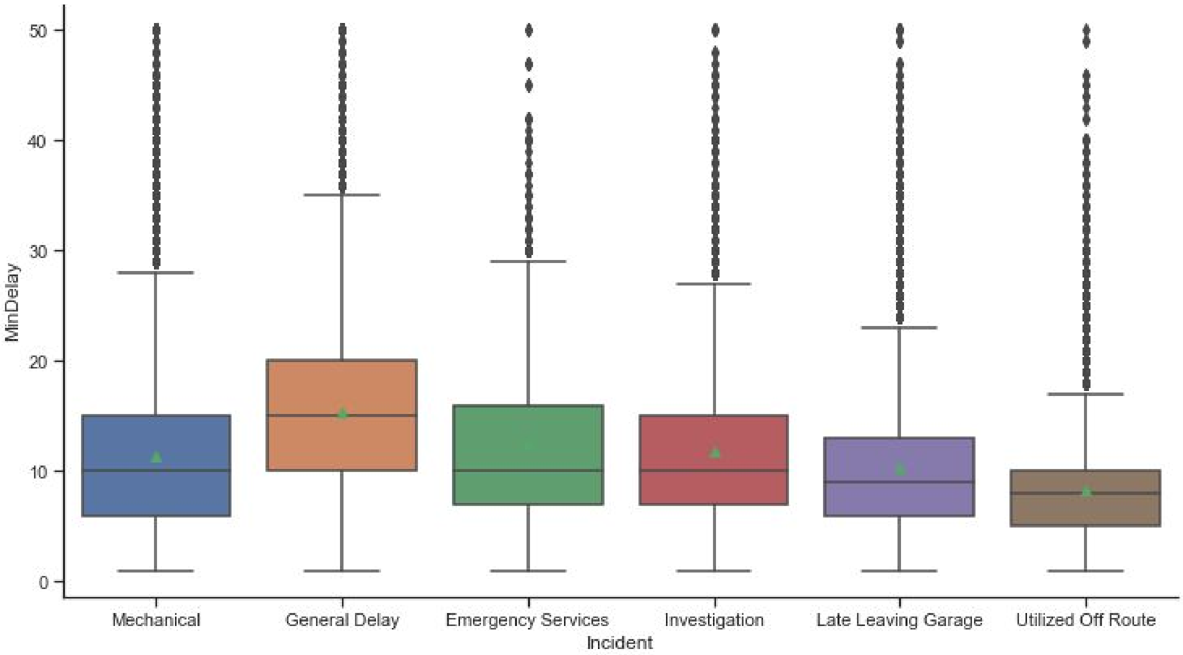


Figure 3-1-2 - Box Plot Indicating Delays Associated with Various Incidents

This analysis will further focus on the issues that are in TTC's control i.e. mechanical and late leaving garage. The other categories are not considered.

Table 3-1-1 - Mean Time of Delay resulting from Various Incidents from 2014-2019

Incident	Mean (Delay), Minutes
Emergency Services	12.6
General Delay	15.4
Investigation	11.9
Late Leaving Garage	10.4
Mechanical	11.3
Utilized Off Route	8.3

3.1.1 Mechanical Delays

Figure 3-1-3 shows illustrates that the even though the number of incidents have decreased through the years, the delays caused due to these incidents have increased steadily over time. This suggests that the mechanical delays are longer and signs indicate to aging fleet.

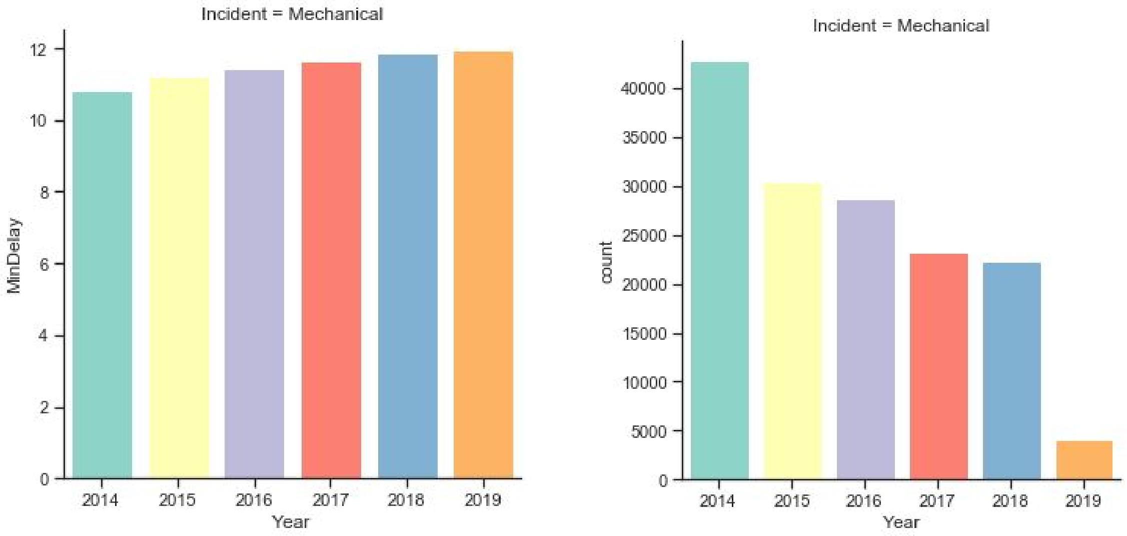


Figure 3-1-3 - (a) Bar Plot of Delays due to Mechanical Problems (b) Bar Plot showing number of incidents per year (2019 until February 2019).

Further analysis on mechanical category associated with different routes is illustrated in Figure 3-1-4 The following observations are made from the data:

- Although the number of incidents in the regular route has decreased, the average delay in service has increased over the past 5 years. The number of incidents have decreased to less than 50% of incidents in 2014.
- The minimum delay due to a mechanical reason has also increased significantly over the past 5 years for the night bus service of the TTC.
- Even though the number of incidents due to mechanical failure in the subway are negligible, the average delay per year has consistently increased.
- TTC has however been able to reduce the average delays in the express bus routes.

The mechanical delays are considered for the different route categories and the times during the day is illustrated in Figure 3-1-5. The following observations are made from the data:

- The average delays due to mechanical issues are longer during the midnight and night than at other times during operations of TTC service. The average minimum delay at these times is close to 14 minutes compared to an average of around 10 minutes at other times.
- The delay in the morning time due to mechanical failure is increasing although the number of incidents have decreased significantly.
- Most of the delays due to mechanical problems occur during the morning followed by afternoon and then the evening periods. Even though the number of incidents are lower in the evening, the average delay is still around 10 minutes when a TTC service is delayed due to any mechanical reasons.
- TTC in general has been able to reduce the overall number of delays in the service due to mechanical reason over the past 5 years.

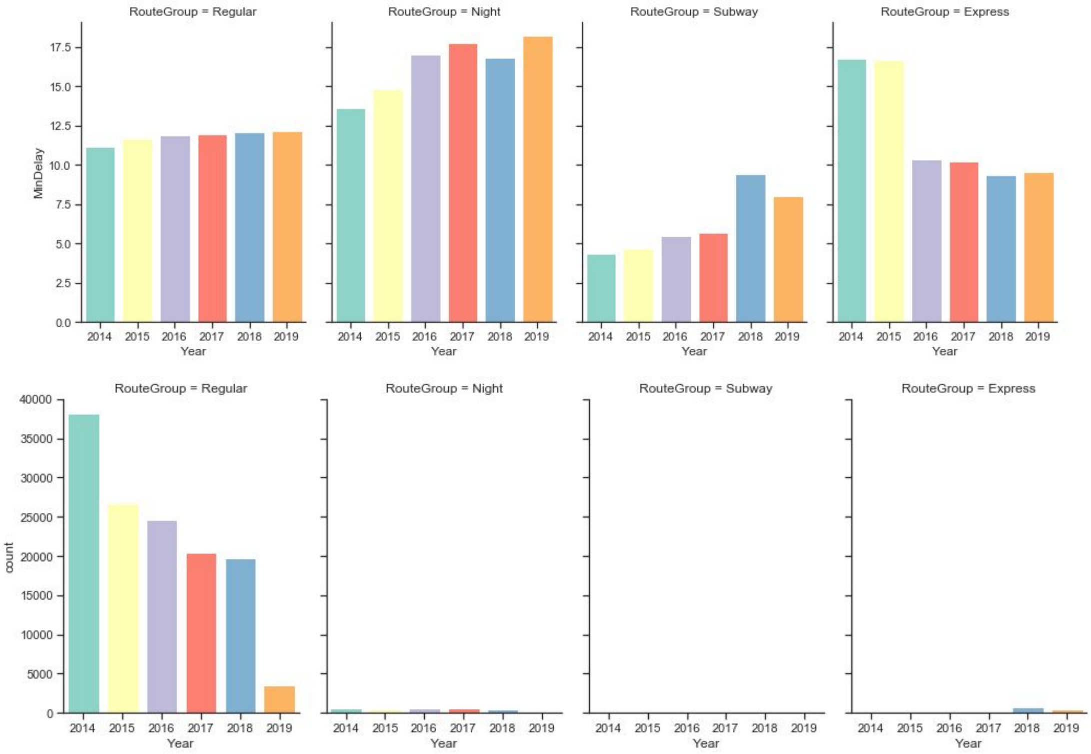


Figure 3-1-4 - (Top) Bar Plot of Delays due to Mechanical Problems on different Routes, (Bottom) Bar Plot of Number of Delays on different Routes (left to right - Regular Buses, Night Buses, Subway, Express)

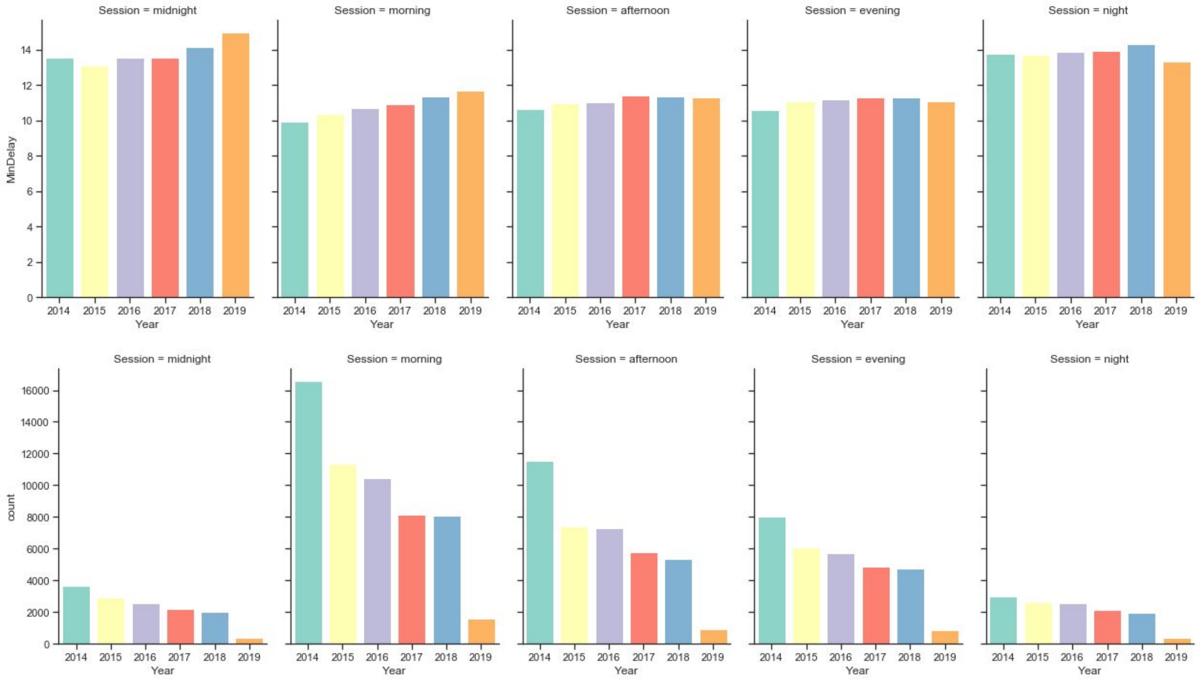


Figure 3-1-5 - (Top) Bar Plot of Delays due to TTC Vehicles Leaving Late from Garage at different times, (Bottom) Bar Plot of Number of Delays at different times (left to right - Midnight, Morning, Afternoon, Evening and Night)

3.1.2 Late Leaving Garage

The delays due to the TTC vehicle leaving garage late have increased over time and the number of incidents have more or less remained consistent. This is illustrated in Figure 3-1-6.

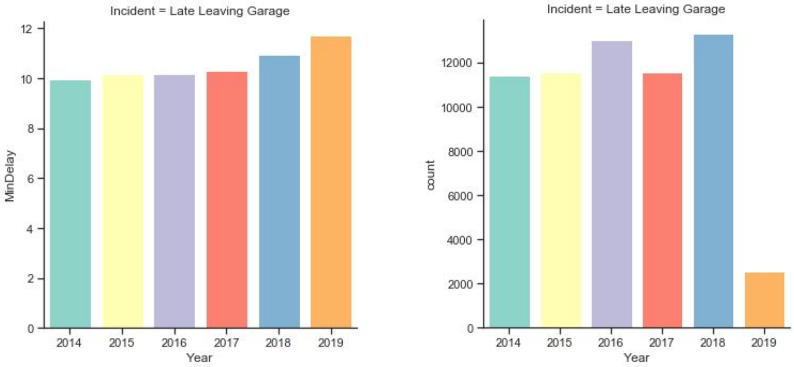


Figure 3-1-6 - (a) Bar Plot of Delays due to leaving garage late (b) Bar Plot showing number of incidents per year (2019 until February 2019).

Figure 3-1-7 suggests that most of the TTC vehicles are delayed during the night service and that TTC has been able to reduce the delays in the express routes due to vehicles leaving late from garage. Most of the delays are associated with buses on regular routes.

Figure 3-1-8 again confirms the observations from Figure 3-1-7 indicating that the delays are longer in the night. Interesting trend to note from this graph is that the morning rush hour has more delays due to vehicles leaving late rather than in the evening rush hour traffic.

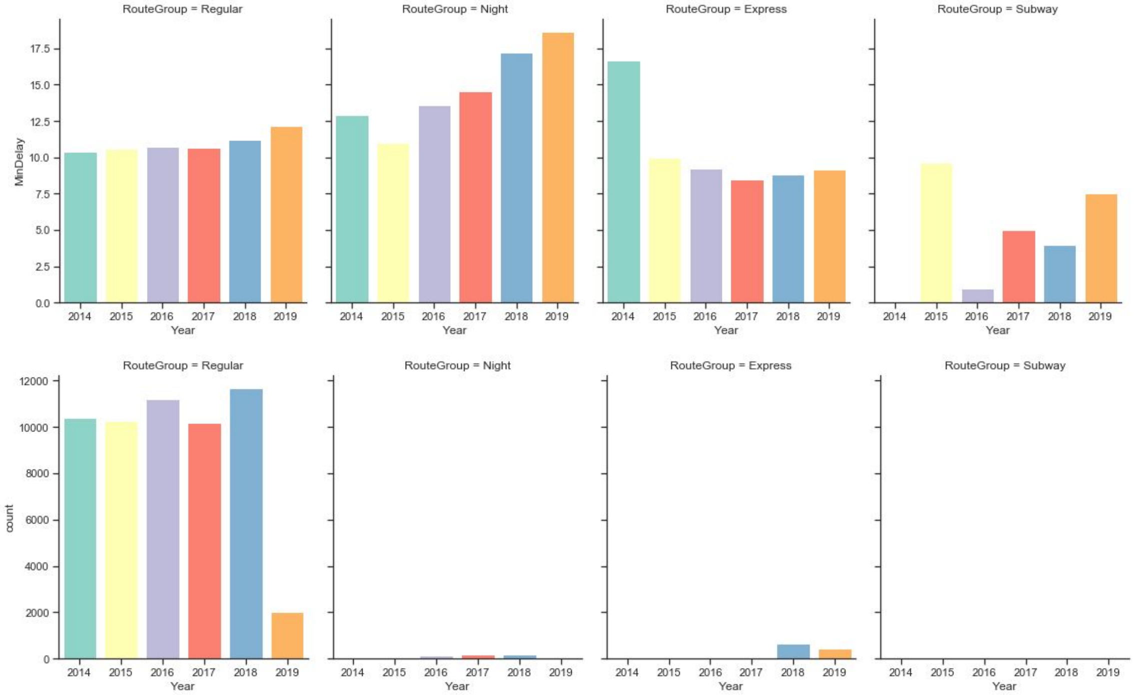


Figure 3-1-7 - (Top) Bar Plot of Delays due to TTC Vehicles Leaving Late from Garage on different Routes, (Bottom) Bar Plot of Number of Delays on different Routes (left to right - Regular Buses, Night Buses, Subway, Express

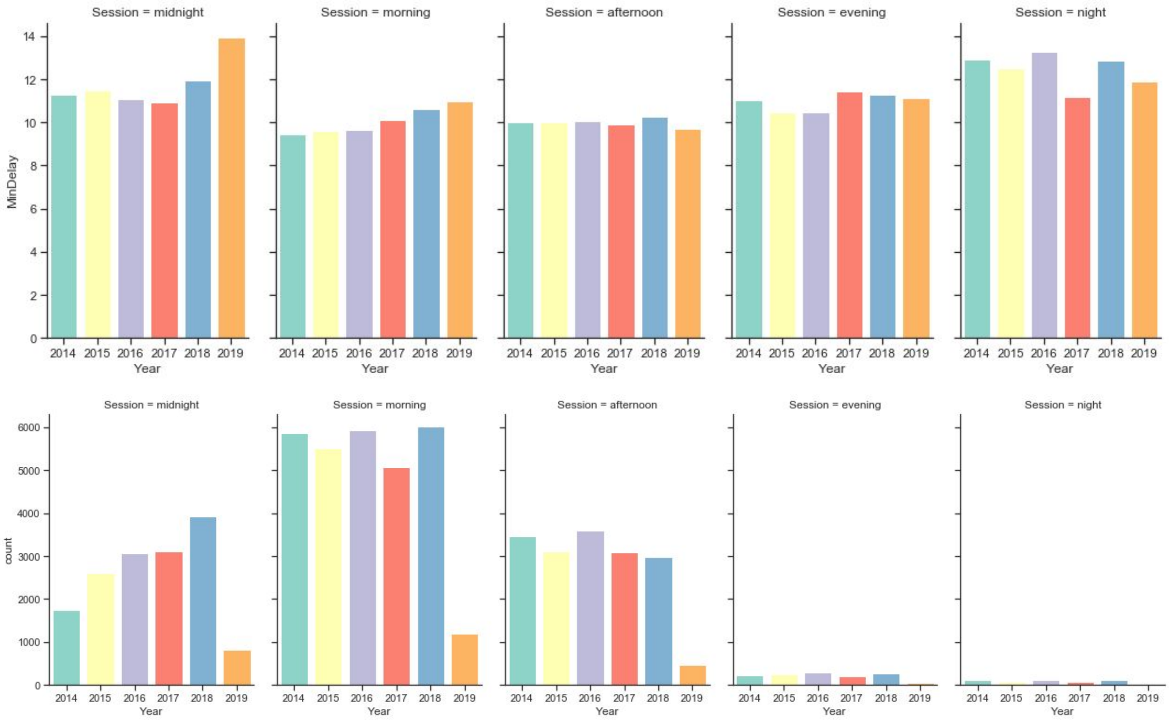


Figure 3-1-8 - (Top) Bar Plot of Delays due to TTC Vehicle Leaving Late from Garage at different times, (Bottom) Bar Plot of Number of Delays at different times (left to right - Midnight, Morning, Afternoon, Evening and Night)

3.2 Date and Calendar Analysis

This section discusses the date and calendar effects on the bus delays in the city of Toronto from 2014 to 2019. This analysis is presented in two parts: days and different sessions of the dat, as well as the months and different seasons.

3.2.1 Days and Sessions

Figure 3-2-1 illustrates the bus delays on different days from 2014 to 2019. From this figure, it is safe to conclude that Saturdays and Sundays have higher delays recorded for the TTC buses of City of Toronto. Sundays have highest delays among the days of the week.

Figure 3-2-2 illustrates the amount of delays in the buses in different sessions of the day, such as morning, afternoon, evening, night, and midnight. Based on this analysis, bus delays from nights are mostly higher than bus delays in any other parts of the day, except for midnights of 2019.

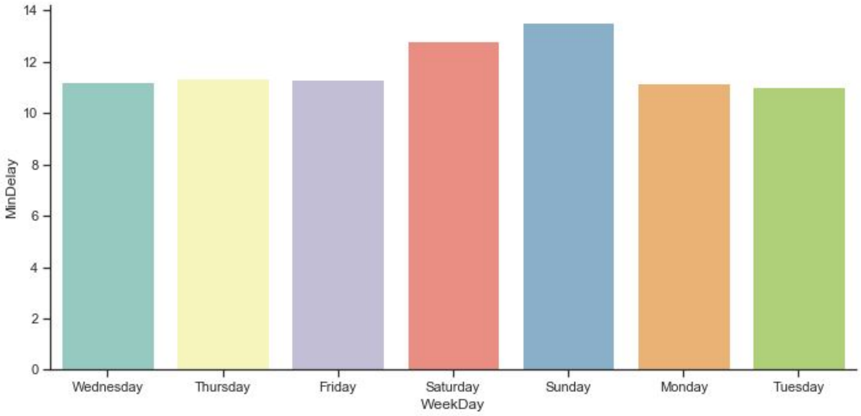


Figure 3-2-1- Bus delays on different days compared on the whole data from 2014 to 2019

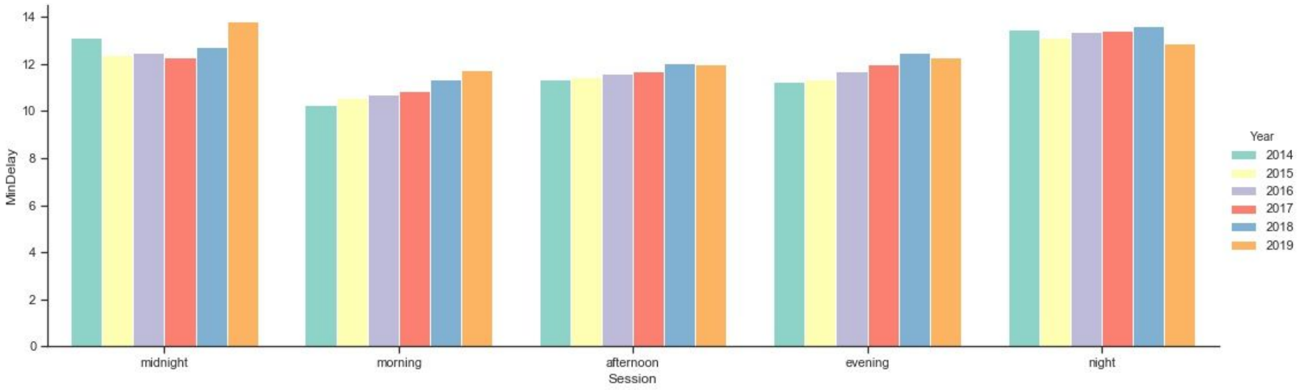


Figure 3-2-2- Bus delay on different sessions of the day (Morning, Afternoon, Evening, Night, and Midnight) from 2014 to 2019

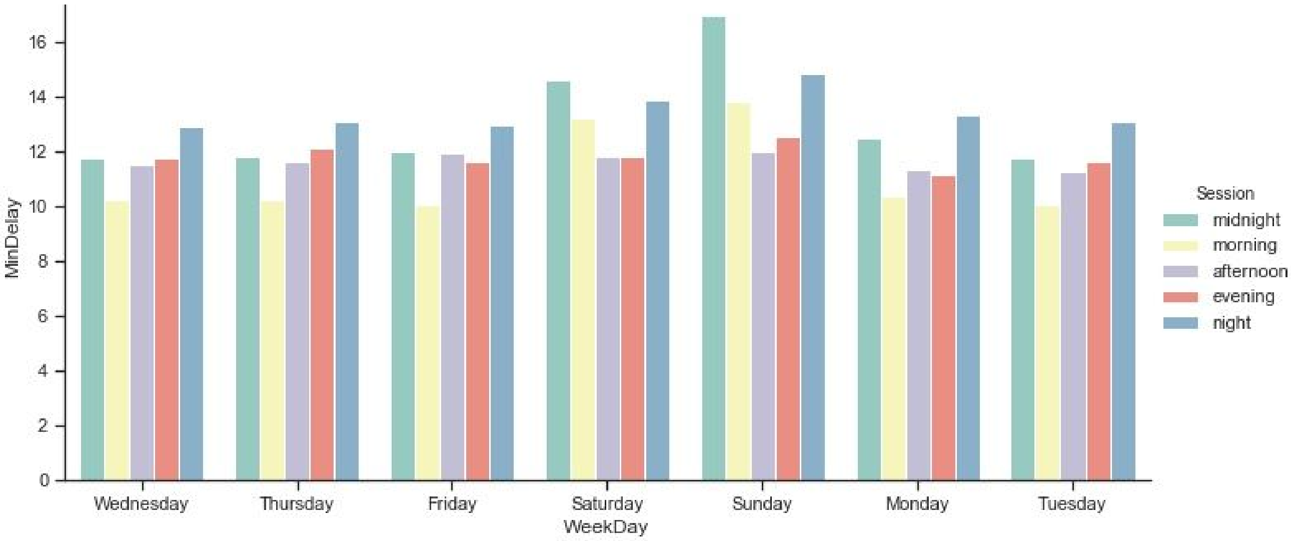


Figure 3-2-3- Bus delays on different sessions of different days of the week

Figure 3-2-3 presents the delays of busses in the city of Toronto through different times of the day, and on different days of the week. This visualizations illustrates that nights and midnight have higher bus delays throughout days. Among different days of the week, Saturdays and Sundays experience highest delays on midnights, while the nights of any other day have the highest delays. It is also obvious that Sunday midnights experience the most delays.

3.2.2 Months and Seasons

Bus delay is also analysed based on different months and different seasons. The seasons are as follows:

- Winter: December, January, February
- Spring: March, April, May
- Summer: June, July, August
- Fall: September, October, November

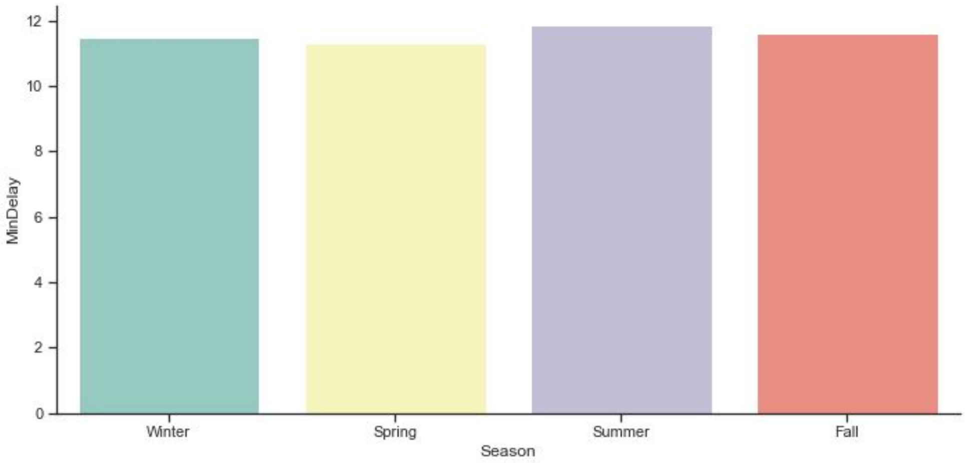


Figure 3-2-4- Bus delays on different Seasons during the whole data from 2014 to 2019

Figure 3-2-4 illustrates bus delays on different seasons. From this figure it could be concluded that summer in total has had higher delays recorded.

Figure 3-2-5 presents the bus delay on different seasons from 2014 to 2019. This illustration shows that bus delays in the year of 2018 has been higher than the rest of the years in every season, except for Winter of 2019. Also, Summer shows higher delays compared to other seasons in every year, except for Winter 2017 that seems slightly higher than Summer of 2017.

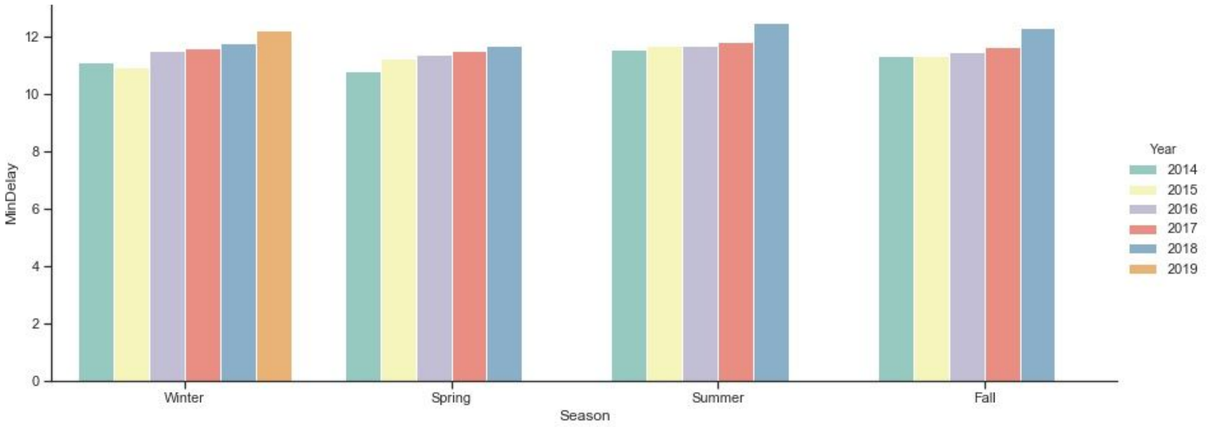


Figure 3-2-5- Bus delay based on different seasons from 2014 to 2019

3.2.3 Combined Calendar Analysis

Finally, bus delay has been observed on different sessions of the day throughout different months and seasons.

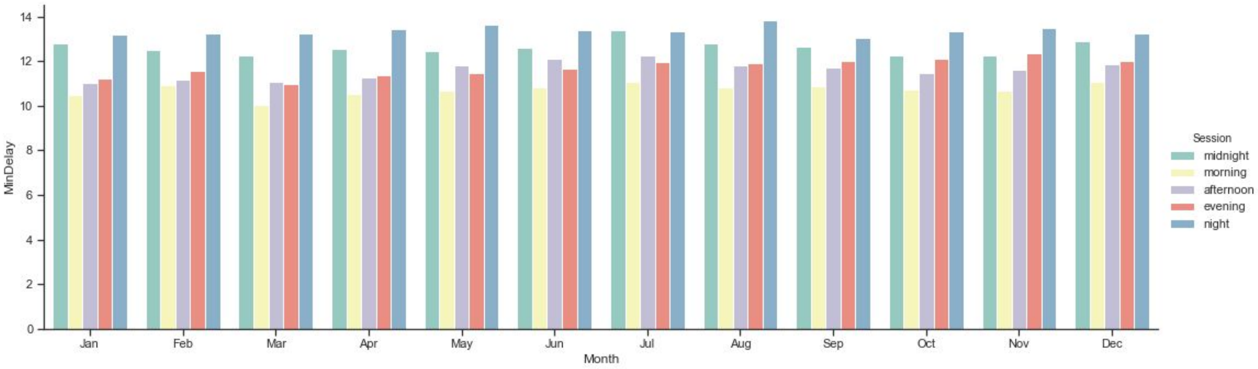


Figure 3-2-6- Bus delay on different sessions of the day on different months from 2014 to 2019

Figure 3-2-6 illustrates the bus delay, on different session of the day and on different months. This figure, also confirms the previous observation that midnights and nights have higher bus delays recorded. The nights of the month August have higher bus delays recorded.

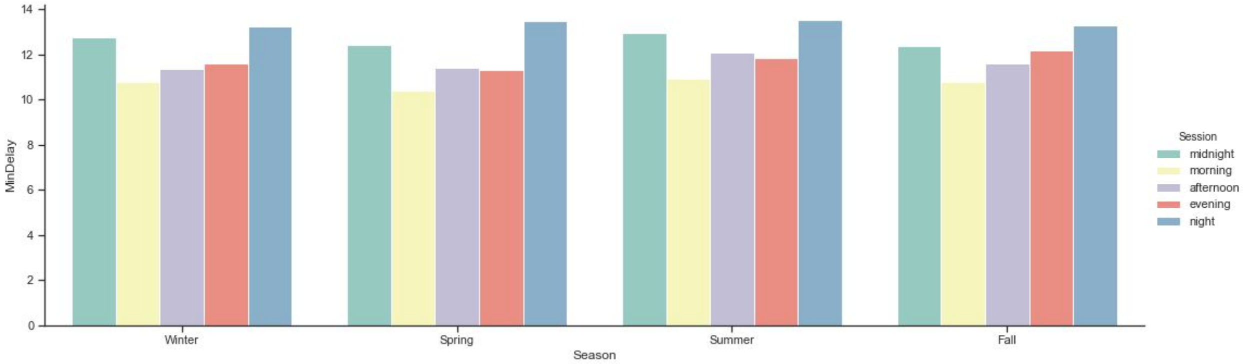


Figure 3-2-7- Bus delays on different sessions of the day on different seasons from 2014 to 2019

Figure 3-2-7, presenting delays based on different seasons and different sessions of the day, also illustrates the higher delays recorded on nights and midnights. Mornings have lowest delays recorded in every season, and the mornings of Spring have lowest delay recorded among the seasons.

3.3 Weather

3.3.1 Temperature Effects

The maximum and minimum daily temperatures were investigated for trends relating to bus delays. Looking at average number of incidents per day, it becomes clear that on the coldest of days, there are the largest number of incidents for buses. This certainly suggests that temperature may play a role in the number of delays caused. The duration of delays paints a different picture though, as no strong trend exists. If anything, there is a slight increase in duration of delay at higher temperatures.

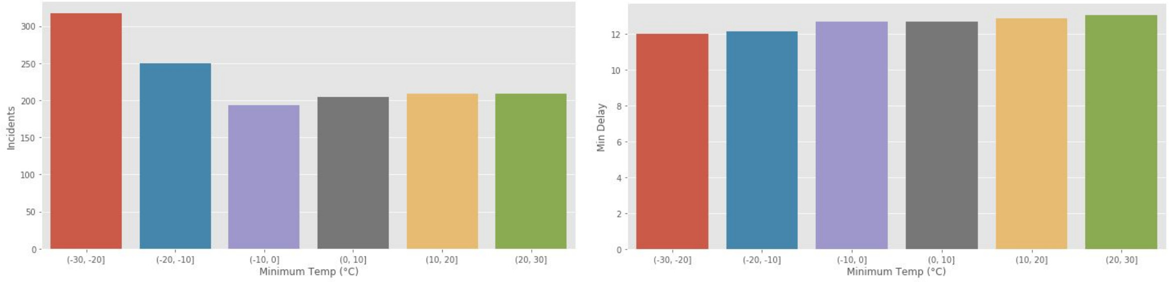


Figure 3-3-1 - Number of incidents and Min delay as a function of Minimum Daily Temperature

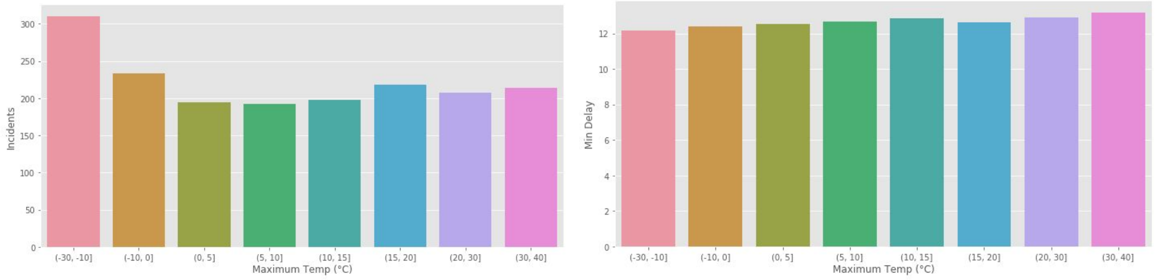


Figure 3-3-2 - Number of incidents and Min delay as a function of Maximum Daily Temperature

3.3.2 Precipitation

Rain fall shows some correlation with delays, but not a strong one. There is some suggestion that with more rain there are more incidents, but the duration of delays is not affected by rainfall.

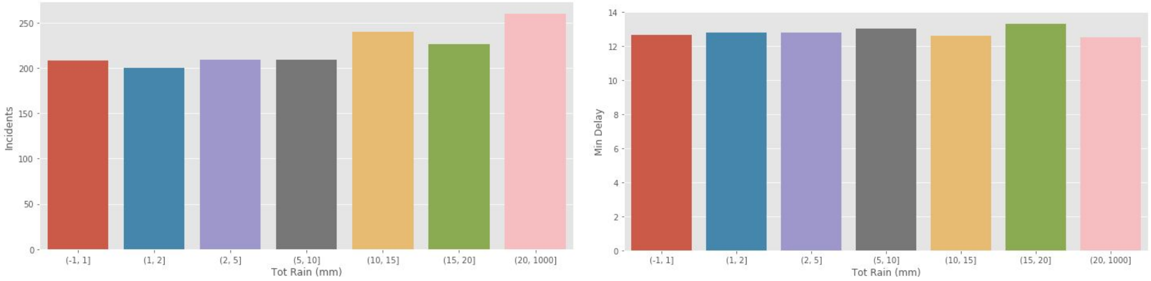


Figure 3-3-3 - Number of incidents and Min delay as a function of Daily Rainfall

Snow has a more pronounced impact than rain. With increased snowfall, there are quite a few more delay incidents and they last a longer time.

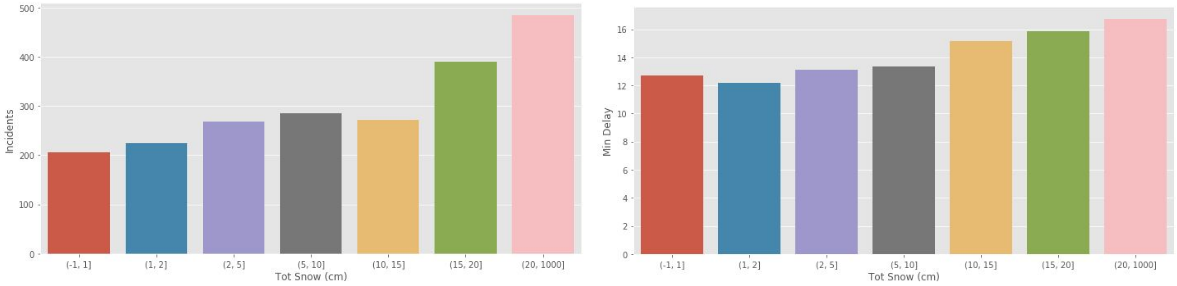


Figure 3-3-4 - Number of incidents and Min delay as a function of Daily Snowfall

3.3.3 Temperature and Reason for Delay

In drawing conclusions around this data, it helps to identify the deeper reason for the incidents. One area where this is strong is temperature. The cold temperatures could cause extra traffic, create ice that necessitates diversions or influence the engines of buses. To address this, the data was re-investigated and broken down by incident type. Normalizing, it becomes clear that mechanical incidents are a larger fraction of total incidents at low temperatures and decrease by percentage at higher temperatures.

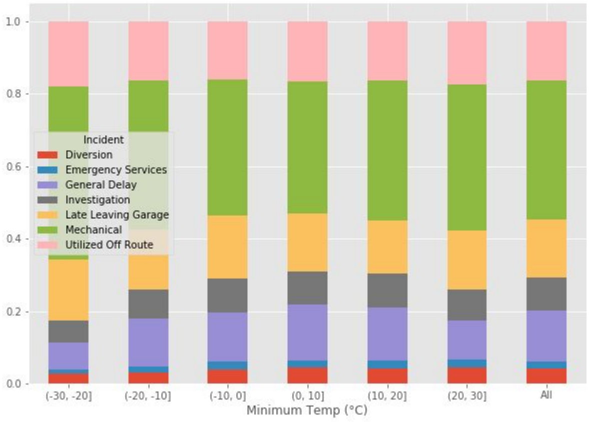


Figure 3-3-5 - Breakdown of incidents with different Minimum Daily Temperature

4.0 KNN model - Predicting Delays

Given the historical nature of this data, one valuable question is how can we use it going forward. Is it possible to create a model that can predict what the delay would be given the independent variables? Being able to predict this would have practical application and would help TTC better serve their customers.

The model that was used in order to come up with predictions was the K Nearest Neighbors model. Once the data was cleaned, the input variables that were used were: Route, Day, Incident and, Hour (an engineered variable to represent the hour of the day when the delay occurred). The target variable was Min Delay. The approach for this analysis was to take the cleaned data, split it into 2 groups: one for the data between 2014-2017 and the other for the data between 2018 and 2019. The 2014-2017 dataset was used to generate the training dataset and the 2018/2019 dataset was used to generate the testing data. The reason for this is that it would not make sense to have instances of more current data being used to predict older data. The split helps to account for this and the time series nature of the data. Once the training and testing datasets were generated, the data was fit to the KNN model and predictions were generated.

The results indicated an accuracy score of 24.1%, which indicates that the previous delay data would not enable us to predict the length of delay based on the: Route, Hour, Day, and Incident. However, it should be noted that this expects an exact match and does not help with delay predictions that are relatively close to the actual delay. In order to build on this further, the accuracy was recalculated to allow a buffer of 15%. If the bus was within 15% of the prediction, it was considered to be fairly accurate. With this implemented, the accuracy score jumped to 34.6%. While being quite an improvement, it is still not good enough to be used to predict the delay in minutes.

Moving forward, in order to predict the TTC delays, the model would need to be updated. The ways to do this would be either: to explore other ML algorithms, to refine the number of neighbours to look at, and to re-evaluate which variables are being used for the prediction.

5.0 Conclusions

The primary cause of TTC service delay is Mechanical (delay of 11.3 minutes) which accounts for approximately 40% of the total number of delay incidents. The number of such failures are reducing year over year but these incidents are causing longer delays.

Another big reason for delay in TTC service is the fact that the bus or subway is delayed in leaving the garage itself. This is worse for night service and morning service accounting for largest proportion of delays. The average delay due to such incidents is approximately 10.5 minutes. One way to improve the service is for TTC to aim for punctuality and try to leave the garage on-time. TTC can achieve this by incentivizing operators leaving the garage on time.

General delay causes longest delays (15 minutes) on routes, however more information is necessary to draw conclusions from these incidents (14% of total incidents). TTC should better categorize this field that would allow for further root-cause analysis.

Based on the data observed and studied, among the days, Saturdays and Sundays have higher bus delays. Also, year of 2018 had highest delays among the years, except for January and February of 2019. Comparing the different sessions of the day, it could also be concluded that nights and midnights would have higher bus delays. Summers also have higher bus delays compared to other seasons. Therefore it could be assumed that the nights and midnights of Saturday/Sundays of Summers would have higher delays**.**

The weather has been shown to influence delays as well. Increased snow creates a pronounced increase in the number and length of delays, while rain has a small effect on the number of incidents, it did not seem to influence the duration of delay. Cold temperatures increase the number of delays, and the most extreme cold causes more mechanical delays in particular.

When looking at predicting the delays, it was found that the historical data could not be used to accurately predict future delays. The accuracy score of the K Nearest Neighbors model was 24.1% and even when a buffer was added, it was found that 34.6% of predictions were accurate within 15% of the actual target. This indicates that further investigation is required in order to develop a model with greater predictive ability. The next steps would be to investigate how accuracy changes when refining relevant inputs and considering alternative models or algorithms.

6.0 Appendix

6.1 Citations

1. Buses." *TTC Buses*, Accessed: March 20, 2019, www.ttc.ca/Routes/Buses.jsp.
2. City of Toronto. "Transportation - Data Catalogue." *City of Toronto*, 19 Jan. 2018, www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/transportation/#bb967f18-8d90-defc-2946-db3543648bd6.
3. "Weather Dashboard for Toronto." *Amateur Weather Statistics for Toronto, Ontario*, toronto.weatherstats.ca/.
4. D'Amore, Rachael. "TTC Delays in 2016 Amount to 26 Days of Lost Service." *Toronto*, CTV New, 17 Feb. 2017, https://toronto.ctvnews.ca/ttc-delays-in-2016-amount-to-26-days-of-service-1.3289190
5. Climate Change Canada. "Daily Data Report for March 2019." *Climate*, 20 Feb. 2019, climate.weather.gc.ca/climate_data/daily_data_e.html?StationID=51459.