# MGOD30: Midterm Case Study

# TTC Delays

Names: Dann Sioson
Student Number: 1001346377
Instructor: Professor Cire
Date: March 4, 2018

# Introduction

Toronto, being home to nearly 2.8 million Canadians, is considered to be one of the many notable cities across Canada. Aside from it's many attractions that provides the population growth rate of approximately 6.2% since 2011 [1], the main long term advantage that Toronto has is its GDP representation. With Toronto representing of approximately 10% of Canada's GDP[2], it can be easily seen that many jobs are needed for Toronto to "operate".

For an individualized person to be living in Toronto, "Torontonians" often face the issue of choosing their housing options or commuting to get through their day to day activities; this abstractly comes down to the matter of convenience. Specifically within commuting, Torontonians have one of three options to get to navigate across Toronto per day to get to their distant destinations:

1. Appropriately finance a personal vehicle
2. Use services in the private passenger transportation industry (i.e. Uber/Taxis)
3. Use the TTC (Toronto Transit Commission)

Depending on a person's location and their destination, one may seem more appropriate than the other. For this report, we will be focusing on the TTC (Toronto Transit Commission) and its services towards Torontonians. Specifically, this report analyzes delays within the TTC's subway system, providing managerial intuitions/recommendations and answering delay scenarios based on data given. Three main files are given: "Subway_Delays_Metadata.xlsx", "Subway_SRT_LogCodes.xlsx", and "Subway_STR_Logs_January_2018.xlsx". Complementary to this report will have a .ipynb file to demonstrate how the data was used to create conclusions through python code, as well as a .csv file of the January logs.

Given these files, let it be known that the data is solely focused on the month of January and is not an accurate representation of the TTC's subway operations throughout the year (since there may exist seasonality factors). Likewise, January is quite an odd month to report considering the change from 2017 to 2018. An example of such includes the TTC's initiative to offer free rides on New Years Eve to combat DUI[3]. However, the sole focus of this report is to analyze the TTC's subway performance throughout January. All technical and low level detail can be found on the ipynb file.

[1] found http://nationalpost.com/news/toronto/canada-census-2016-toronto-growth-well-above-the-already-high-national-average

[2] found http://www.tfsa.ca/toronto-advantage/

[3] found https://www.thestar.com/news/gta/2017/12/29/ttc-offers-free-rides-on-new-years-eve.html

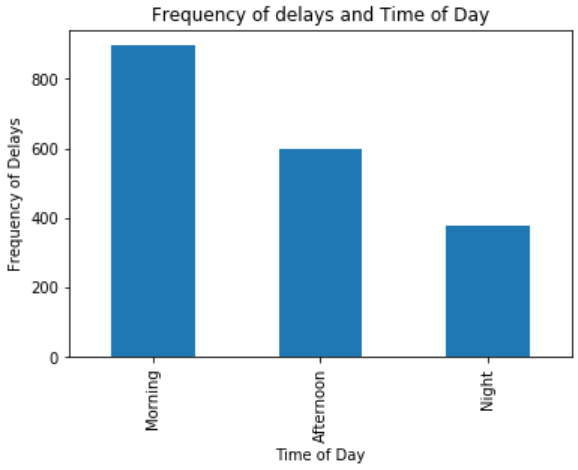
# Managerial Questions

1. **During which periods of the day do the delays occur more often?**

   First, in a broad scope, we have divided the data into three parts, depending on the hour of the day:

| Category of time | Hour of Day |
|---|---|
| Morning | 12:00am to 12:00pm |
| Afternoon | 12:00pm to 7:00pm |
| Night | 7:00pm to 12:00am |

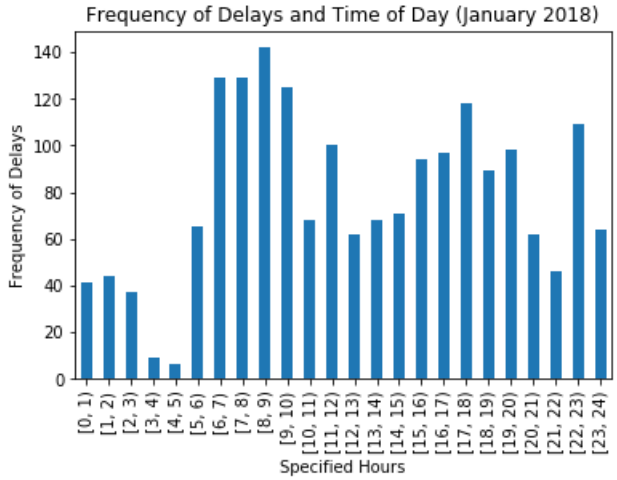
   Given this criteria, where the question is interpreted as how frequent delays occur in a specified period of day, we are able to produce the following graph:




   What we can infer from this graph is that that mornings tend to have the highest frequency of delays in January. Also, as the day progresses, we can see that the frequency of delays decreases. To provide some inference, this piece of knowledge is convenient for a typical TTC customer that rides the subway. This is because riders may not expect when their commute may be delayed on a day to day basis. For further inference, we can see that riders must make adjustments in the morning to get to their destination.

   In a managerial standpoint, what would be ideal is to have the frequency of delays uniformly distributed between morning, afternoon, and night. Since there is a significant amount of riders in the morning compared to the evening, it's much more inconvenient for riders to be riding the subway in the morning as they may be late to get to their destination due to such delays.

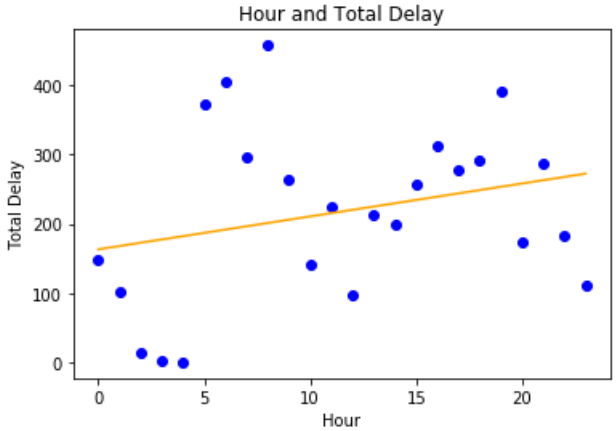
   To solve this, what can be proposed is having reinforced maintenance practices at night so that there is significantly less delays in the morning. In this case, what we have to do is to look at what hours of the day consist of low frequency of delays (at night), and then provide reinforced maintenance on those hours. We have the following graph below that displays the frequency of the delays at a specified hour:




Since "night" consists of 5 hours (7:00pm to 12:00am or 19:00 to 24:00), it may make some difference to provide maintenance during the hours that have the lowest frequency of delays (in this case, from 8:00pm to 10:00pm). However, what is found to be interesting is that the large majority of the delays occur from 6:00am to 10:00am, which does make some sense since subway services typically start operating at 6:00am. However, let it be known that if there were more reinforced maintenance practices from 12:00am to 6:00am, then the frequency of delays in the morning may decrease significantly.

2. **Is there any correlation between the time of the day and the total delay time?**

In the broad, general picture, we are able to find the following visual representation between time of day and total delay time, along with the Pearson correlation coefficient for all days in January:




Pearson Correlation coefficient:

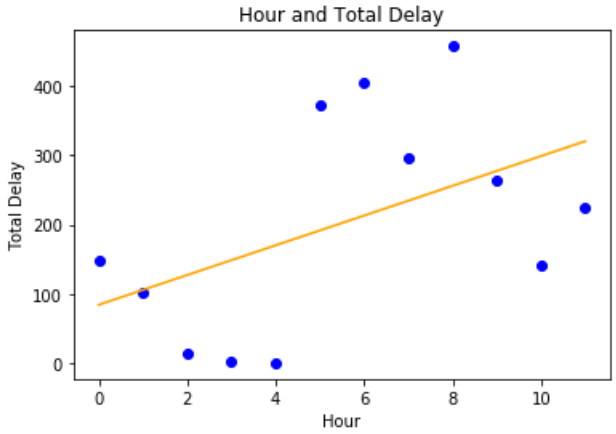
| | Hour of Day | Min Delay |
|---|---|---|
| **Hour of Day** | 1.000000 | 0.265789 |
| **Min Delay** | 0.265789 | 1.000000 |

Given the Pearson Correlation Coefficient (at 0.27), there is a weak positive correlation between the time of day and total delay time. What was found interesting was that the data looks like it "bounces" around in a sinusoidal-like function. This raises our suspicion that it may be best to divide the data into three parts like what was done in question 1:

| Category of time | Hour of Day |
|---|---|
| Morning | 12:00am to 12:00pm |
| Afternoon | 12:00pm to 7:00pm |
| Night | 7:00pm to 12:00am |

Here is the graphical representation of the data for mornings, along with the Pearson correlation coefficient:

```
There are 895 rows used
```



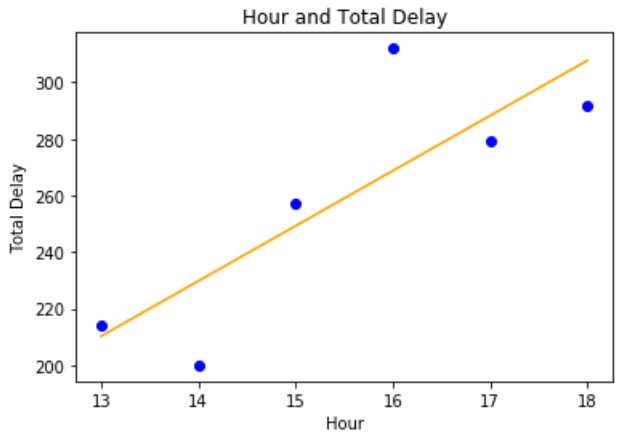

```
The correlation coefficient in mornings:
```

| | Hour of Day | Min Delay |
|---|---|---|
| **Hour of Day** | 1.000000 | 0.484116 |
| **Min Delay** | 0.484116 | 1.000000 |

We see that the Pearson correlation coefficient is raised a bit more in comparison to all hours of the day (from 0.27 to 0.48), making a stronger correlation.

Here is the graphical representation of the data in the afternoon, along with the Pearson correlation coefficient:

There are 537 rows used



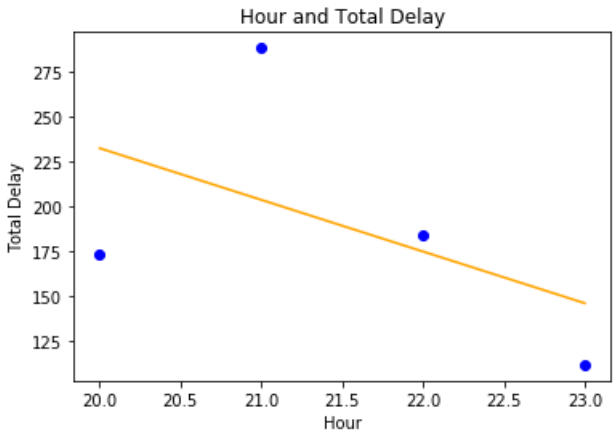

The correlation coefficient in afternoons:

| | Hour of Day | Min Delay |
|---|---|---|
| **Hour of Day** | 1.000000 | 0.823086 |
| **Min Delay** | 0.823086 | 1.000000 |

We see that the Pearson correlation coefficient is at 0.82, which is considered to be a strong positive correlation between hour of the day(in the afternoon) and the total delays.

Here is the graphical representation of the data at night, along with the Pearson correlation coefficient:

There are 281 rows used




The correlation coefficient at night:

| | Hour of Day | Min Delay |
|---|---|---|
| **Hour of Day** | 1.000000 | -0.507179 |
| **Min Delay** | -0.507179 | 1.000000 |

We see that the Pearson correlation coefficient is at -0.51, which is considered to be a strongly negative correlation between the hour of the day (at night) and the total delays.

Breaking the data into parts, we can see that overall, the time of day correlates to the total delay time. To provide a managerial inference of the data shown, we can determine another measure of where maintenance is needed as the total time in delay significantly outweighs frequency of delays for a typical commute.

Again, it's ideal to have uniform total delay time when subways are operating. In this case, in the morning and reiterating intuitions from question 1, reinforced maintenance practices should be conducted when the subways are not operating (1:30am to 6:00am). In the afternoon, more maintenance should be conducted prior to rush hours / "hot hours" (before 4:00pm) as the congestion of the commute may have an effect on the delay time. Evenings are somewhat ideal considering that there would be less people and delay time as the subways stop operating. However, we do not negate the fact that reinforced maintenance practices are needed when subways are about to stop operating so that mornings can have less of a delay time.

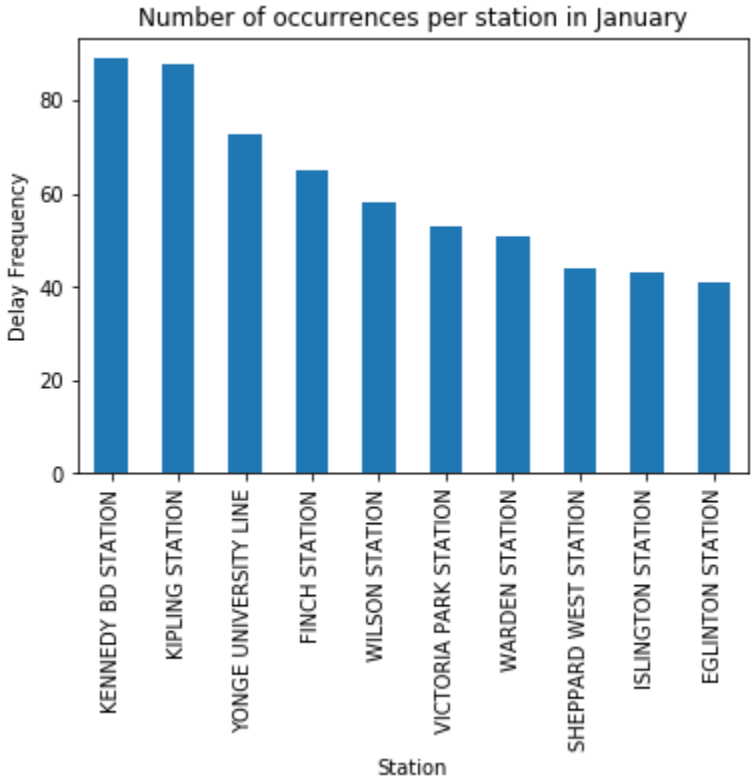
3. **What are the 10 stations that suffer the most delays? Consider both the total number of delay occurrences and the length of the delays as criteria**

In this case, we will initially consider two pieces of the criteria, one being the frequency of delays that occur in the respective station, and the other being the total delay (in minutes) of the specific station for all of January.

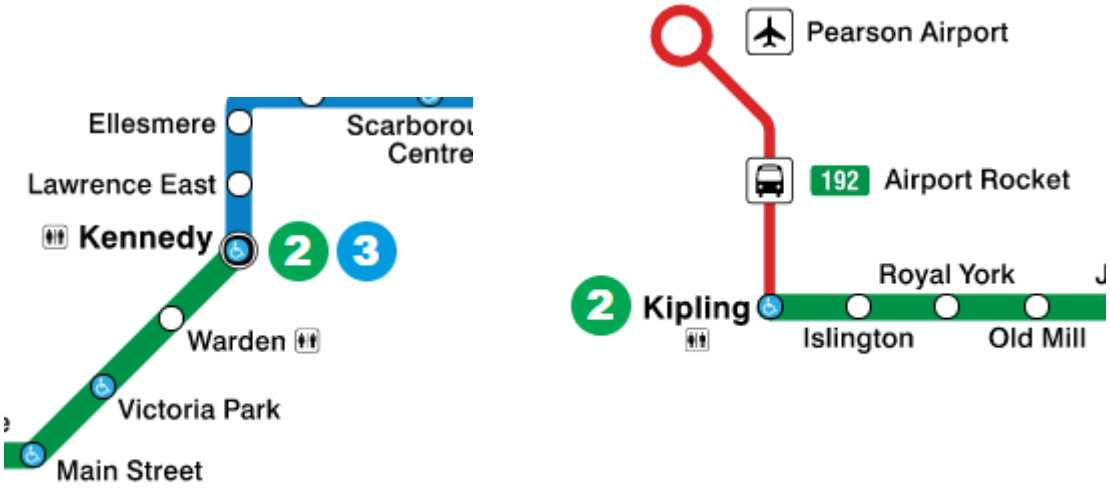
First, we will consider the top 10 stations that have the highest frequencies with the table shown below:

| Station | Delay Frequency |
|---|---|
| KENNEDY BD STATION | 89 |
| KIPLING STATION | 88 |
| YONGE UNIVERSITY LINE | 73 |
| FINCH STATION | 65 |
| WILSON STATION | 58 |
| VICTORIA PARK STATION | 53 |
| WARDEN STATION | 51 |
| SHEPPARD WEST STATION | 44 |
| ISLINGTON STATION | 43 |
| EGLINTON STATION | 41 |

Upon further investigation, "BD" means "Bloor-Danforth" Line. This is stated because simply labelling "Kennedy Station" would be quite confusing considering that Kennedy Station is connected to both the Bloor Danforth line and the SRT. With the data above, we can graphically see the delay frequencies:



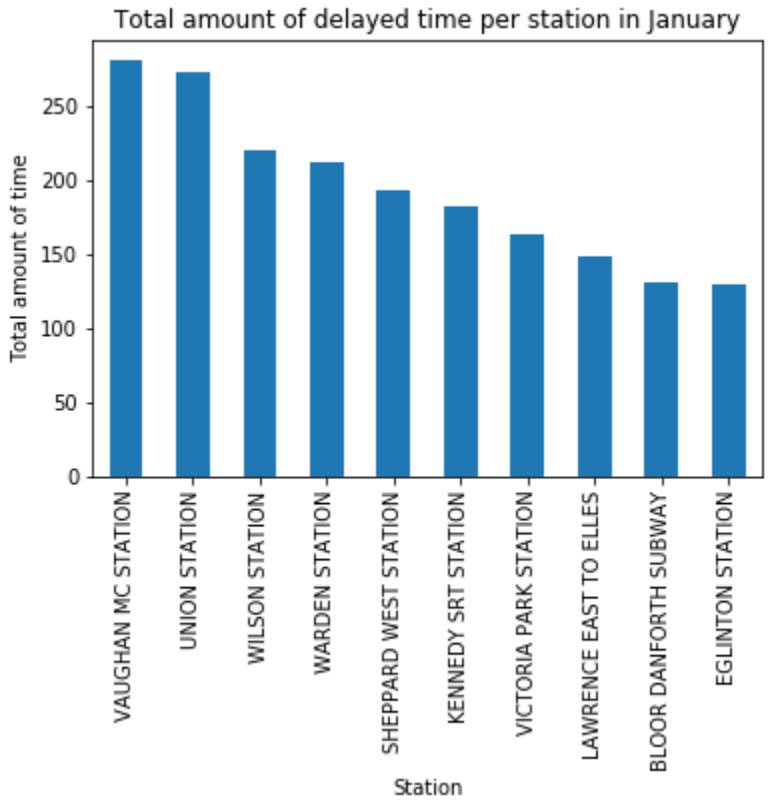

Given the graph above, the number of delays allocated to a specified station is quite interesting considering that the top two stations are on the opposite ends (and start) of the Bloor Danforth line:



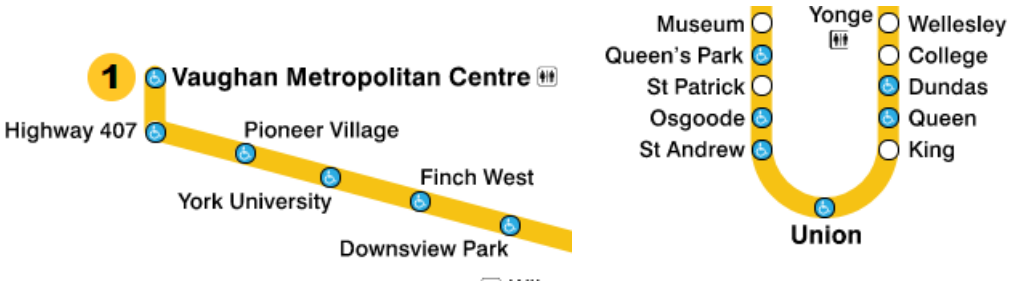

We will now consider the top 10 stations that have the highest delay time with the table shown below:

| Station | Min Delay |
|---|---|
| VAUGHAN MC STATION | 280 |
| UNION STATION | 273 |
| WILSON STATION | 220 |
| WARDEN STATION | 212 |
| SHEPPARD WEST STATION | 193 |
| KENNEDY SRT STATION | 182 |
| VICTORIA PARK STATION | 163 |
| LAWRENCE EAST TO ELLES | 148 |
| BLOOR DANFORTH SUBWAY | 131 |
| EGLINTON STATION | 129 |

And below is the graphic representation of the table above:



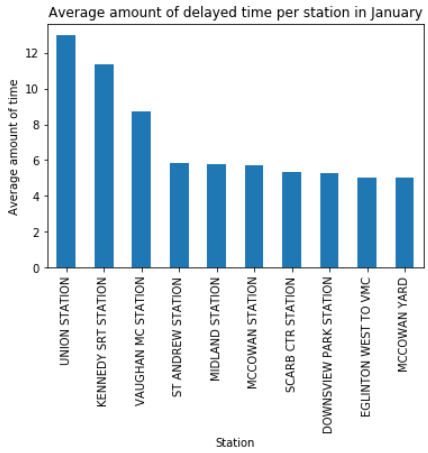

The graph above is also interesting as well. In regards to the top two stations that have the highest amount of delayed time, they are located at the most northern station and southern stations on the TTC map:




Reiterating the question, the criteria above and as described in the question does not signify what station is considered to be the "worst" station in terms of delays. In instances like these, information may be heavily skewed (i.e. many "delays" occurred, but they might not have lasted a minute OR maybe there was one unlucky station that had many delays in tandem). To determine the station that suffered the most, we consider the average time per delay of each station. The top 10 stations that have the highest average time per delay are shown below:

| Station | Min Delay | Delay Frequency | Average time per delay |
|---|---|---|---|
| UNION STATION | 273 | 21 | 13.000000 |
| KENNEDY SRT STATION | 182 | 16 | 11.375000 |
| VAUGHAN MC STATION | 280 | 32 | 8.750000 |
| ST ANDREW STATION | 93 | 16 | 5.812500 |
| MIDLAND STATION | 23 | 4 | 5.750000 |
| MCCOWAN STATION | 114 | 20 | 5.700000 |
| SCARB CTR STATION | 64 | 12 | 5.333333 |
| DOWNSVIEW PARK STATION | 37 | 7 | 5.285714 |
| EGLINTON WEST TO VMC | 5 | 1 | 5.000000 |
| MCCOWAN YARD | 5 | 1 | 5.000000 |

The graphical representation of the data is below:



With the instances above, we can conclude that geographical location does play a large role in terms of delays. Specifically, Union station appears as though it suffered the most, while honorably all of Kennedy station suffered as a runner up. In the third place, we have Vaughn Station, which located all the way north.

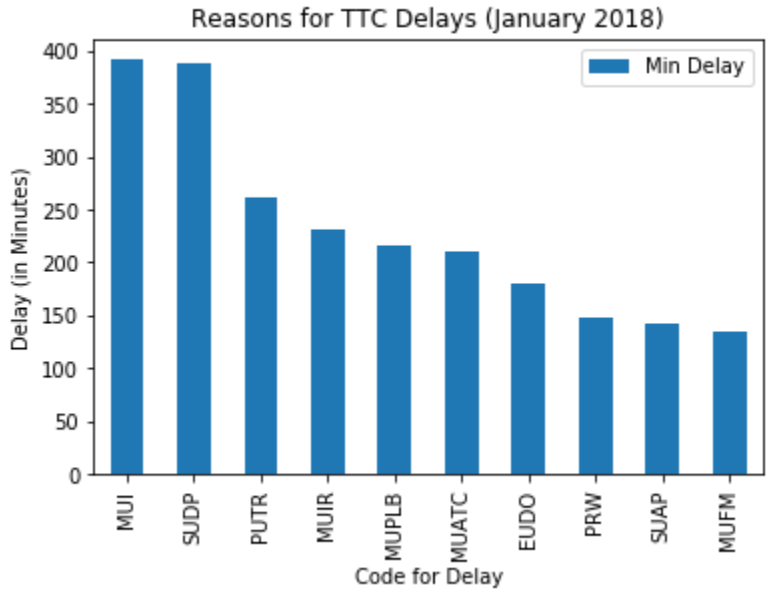
We can conclude that stations that have either an intersection or is in an extreme geographical location (i.e. all the way West/North etc) may need more maintenance and attention than a station that is located central of Toronto. By focusing on these stations, it would significantly reduce the amount of delays and the time of delays.

## 4. What are the top 10 reasons for delays (in terms of total time)?

Below are the top 10 codes that caused the most delay along with its data in January:

| Code | Description of Code | Total delay in January |
|---|---|---|
| MUI | Injured or ill Customer (on train) - Transported | 392 |
| SUDP | Disorderly Patron | 389 |
| PUTR | Rail Related Problem | 262 |
| MUIR | Injured or ill Customer (on train) - Medical Aid Refused | 231 |
| MUPLB | Fire/Smoke Plan B - Source TTC | 216 |
| MUATC | ATC Project | 211 |
| EUDO | Door Problems - Faulty Equipment | 179 |
| PRW | Rail Defect/Festenings/Power Rail | 148 |
| SUAP | Assault / Patron Involved | 142 |
| MUFM | Force Majeure | 135 |

Here is a graphical perspective of the data above:



We see that the top two reasons are significantly higher causes of the delay time compared to the rest of the top 10. With these top two reasons, it looks like it involves people outside of the TTC. For clarification, a "Disorderly Patron" is a person that is described to behave in a disorderly manner, generating risk to themselves or even others. This is a fairly difficult problem to tackle because these problems aren't entirely in the TTC's control. However, what can be done is to regulate stricter TTC by-laws on the subway (or even in general) so that there is no tolerance of disruption involving a patron (this, of course would need further research than the data given). Another suggestion would be to employ TTC moderators to ride the subway so that hey can handle specific social situations (i.e. provide medical need or enforce existing laws). Another suggestion would be to have an anti-bystander program so that people stand up to disorderly patrons, and are rewarded for it. This is an idea that can be abused, but then again, nearly all of the TTC subways have cameras for a specified vehicle, thus people can provide proof for this program given with the TTC's archive.

Though these are a few ideas, it may be a great starting point to cut down delays.

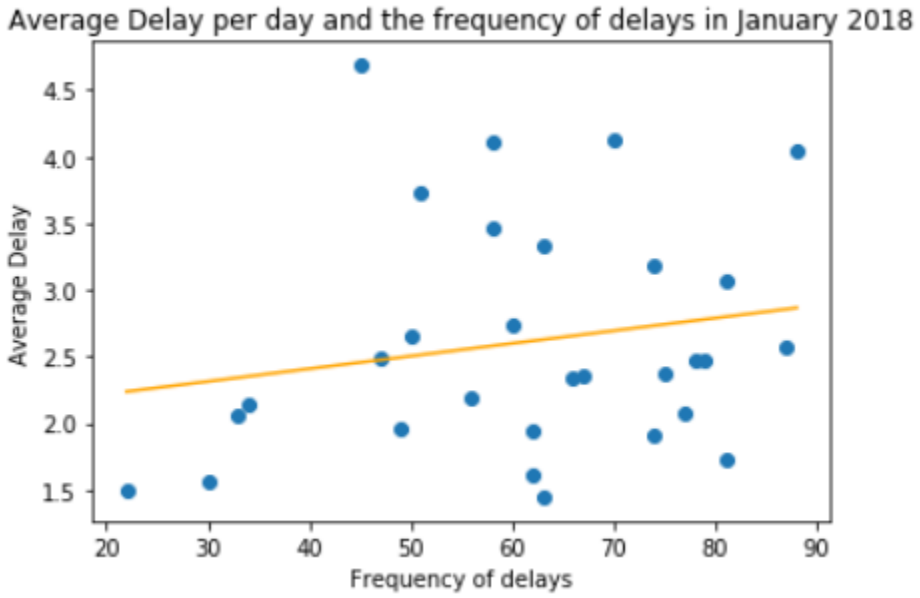
5. **How many times per day did a delay occur in January? What is the average delay time per day?**

On an average day there are approximately 60 delays, lasting approximately 2.8 minutes (this can be found in the ipynb file). Not much can be inferred with these piece of knowledge.

However, what can be taken into initiative given how the question was worded, is to see if there is a correlation between the average delay time per day and the frequency of delays. Which raises the question: "If there are more frequent delays, does that mean that the average delay time decreases?". If this statement were to be true, then it may be more convenient to an average TTC subway rider to know this. It may be more preferable for a subway rider to know that their commute will have consistent delays, which they can schedule around, compared to long unexpected delays which a rider can't schedule around.

This would also fall under the question "If there were more maintenance hours placed in the TTC subway system during the day to fix or control these issues, would this cause more of a convenience to customers?". Theorizing with that question, it would presume that if there was constant maintenance everywhere in the TTC subway system, then it means that there doesn't exist any delays.

To test this, we have drawn a scatter plot below:




Below is a Pearson correlation coefficient matrix. In this case, we are focusing on the entry between the frequency of delays and "Min delay" or in this case, the average delay per day:

| | Min Delay | Min Gap | Vehicle | Frequency of delays |
|---|---|---|---|---|
| **Min Delay** | 1.000000 | 0.928551 | 0.071239 | 0.188295 |
| **Min Gap** | 0.928551 | 1.000000 | 0.053520 | 0.010898 |
| **Vehicle** | 0.071239 | 0.053520 | 1.000000 | 0.080066 |
| **Frequency of delays** | 0.188295 | 0.010898 | 0.080066 | 1.000000 |

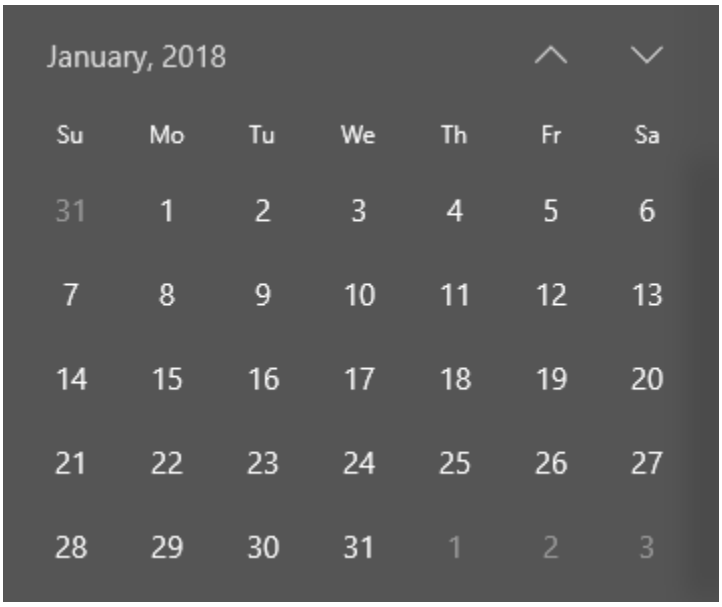
Sadly, we see that there is a positively (weak) correlation between the average delay per day and the frequency of delays. Meaning that, the higher the frequency, the higher the average delay time. However, let it be known that we only have a limited of set of data (because we are only doing the dates in January), and more historical data may be needed to provide an accurate conclusion. Future considerations of hypothesizing the relationship between the frequency of delays and average delay time should be pursued.

6. **Is there any correlation between day of the week and the total number of delays?**

With the data given, it was rather difficult to categorize the days of the week (i.e. Monday, Tuesday etc.) to a numerical value. In this case, what was done is that the days of the week have a numerical representation, where Monday is 1, Tuesday is 2, and so forth.

In regards to the original question asked, categorizing by the total number of delays of each day of the week is not a great measure for the month of January as the data may be skewed. In the month of

January; Monday, Tuesday, and Wednesday have three extra days:

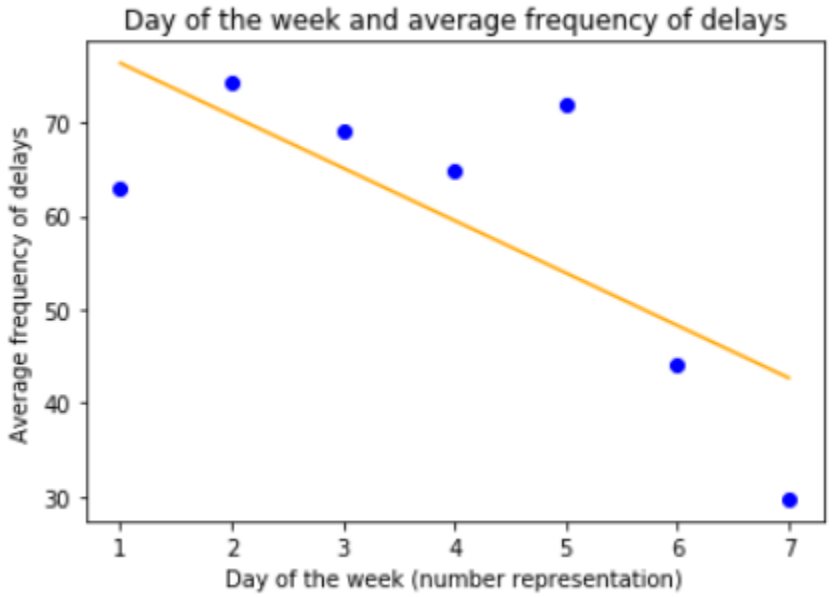



In this case, a better measure is to find the average frequency of the delays associated with the day of the week. Given this criteria, the table is shown below:

Our data looks like:

| Day | Frequency of delays | Days (number representation) | Frequency of days in January | Average frequency of delays |
|---|---|---|---|---|
| **Monday** | 315 | 1 | 5 | 63.00 |
| **Tuesday** | 371 | 2 | 5 | 74.20 |
| **Wednesday** | 345 | 3 | 5 | 69.00 |
| **Thursday** | 259 | 4 | 4 | 64.75 |
| **Friday** | 288 | 5 | 4 | 72.00 |
| **Saturday** | 176 | 6 | 4 | 44.00 |
| **Sunday** | 119 | 7 | 4 | 29.75 |

From this data, we can provide a scatter plot graph:




Below is the Pearson correlation coefficient matrix. In this case, we are focused on the intercept between "Days (number representation)" and "Average frequency of delays":

| | Frequency of delays | Days (number representation) | Frequency of days in January | Average frequency of delays |
|---|---|---|---|---|
| **Frequency of delays** | 1.000000 | -0.875683 | 0.780590 | 0.940085 |
| **Days (number representation)** | -0.875683 | 1.000000 | -0.866025 | -0.736117 |
| **Frequency of days in January** | 0.780590 | -0.866025 | 1.000000 | 0.522762 |
| **Average frequency of delays** | 0.940085 | -0.736117 | 0.522762 | 1.000000 |

In this case, the average frequency of delays is strongly and negatively correlated with the day of the week, listing its Pearson's correlation coefficient at -0.73. To translate, as the week progresses, there are less occurrences of delays.

From a managerial standpoint, we can conclude that heavy/long-term maintenance should occur during the weekend as Saturdays and Sundays have the lowest average amount of delays. Intuitively, it is believed that providing maintenance during those days may reduce the amount of technical delays during the weekday.

The graph also suggests that congestion may have an effect on the frequency of delays as the most popular days to ride the TTC are the weekdays. This is because Torontonians need to get to their day-to-day destinations for a living (i.e. work/school), using the TTC as their main transportation. This does relate back to question 4, where the top delays occur when there's more people using TTC services. Thus, this graph reinforces the intuitions/ideas that was discussed in question 4.

7. **Is there any correlation between day of the week and average delay time? How about time of the day and minimum gaps?**

This question will divided into two parts. The first part being the correlation between day of the week and average delay time, and the second part being the correlation between time of day and gaps between trains.

(a) Day of the week and average delay time

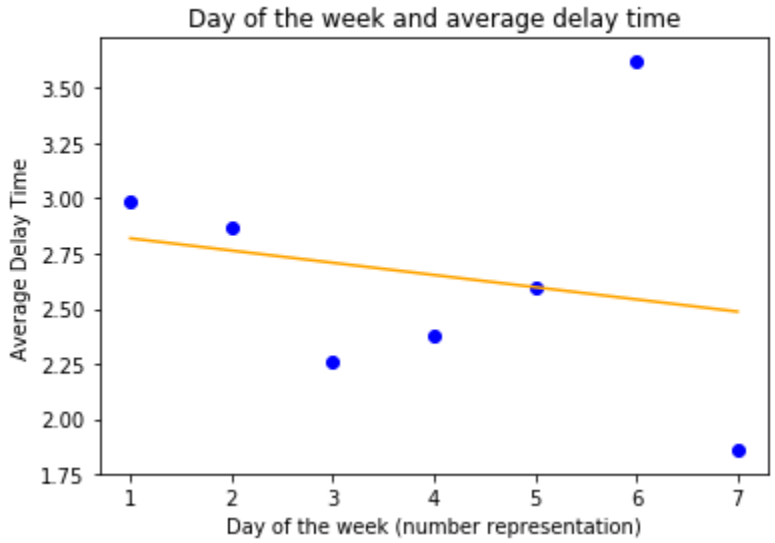
Similar to question 6, every day of the week had an associated number representation with it (i.e. Monday is 1, Tuesday is 2 and so forth). In this case, since we are talking about averages (and not frequencies), there isn't a need to insert the amount of days there exists a particular day of the week in January as a measure.

Data needed for day of the week and delay time:

| Day | Min Delay | Min Gap | Vehicle | Days (number representation) |
|---|---|---|---|---|
| **Friday** | 2.597222 | 3.829861 | 4105.534722 | 5 |
| **Monday** | 2.987302 | 4.139683 | 4044.476190 | 1 |
| **Saturday** | 3.622378 | 5.538462 | 4175.741259 | 6 |
| **Sunday** | 1.857143 | 3.134454 | 4028.193277 | 7 |
| **Thursday** | 2.374517 | 3.374517 | 3907.888031 | 4 |
| **Tuesday** | 2.870620 | 3.897574 | 4066.420485 | 2 |
| **Wednesday** | 2.257971 | 3.321739 | 4167.640580 | 3 |

With the data above, we can create a scatter plot graph.

Visual representation of data:




From the data, we can produce Pearson's correlation coefficient matrix. In this case, we are focused on the columns "Days (number representation)" and "Min Delay" (which is the average delay time)

Correlation coefficient matrix:

| | Min Delay | Min Gap | Vehicle | Days (number representation) |
|---|---|---|---|---|
| **Min Delay** | 1.000000 | 0.953175 | 0.414956 | -0.208448 |
| **Min Gap** | 0.953175 | 1.000000 | 0.504381 | 0.073684 |
| **Vehicle** | 0.414956 | 0.504381 | 1.000000 | 0.090413 |
| **Days (number representation)** | -0.208448 | 0.073684 | 0.090413 | 1.000000 |

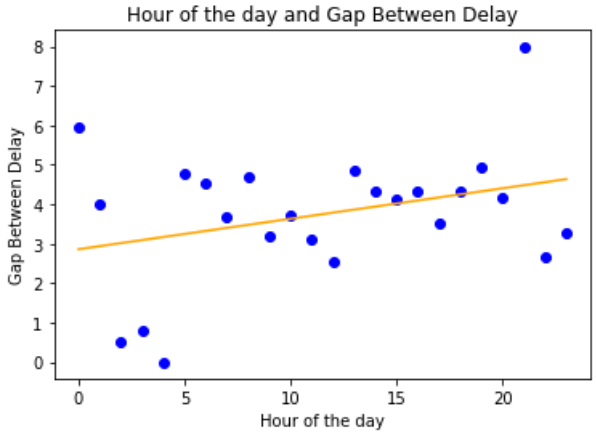
From the intersection of "Days (number representation)" and "Min Delay" in the matrix, we see that the Pearson correlation coefficient between the two is weakly and negatively correlated, having a correlation coefficient of approximately -0.21.

One similar managerial intuition appears here as it did in question 6. We see that Sunday had the lowest amount of frequency and it had the lowest amount of average delay time here. This reinforces the answer that Sunday should definitely be a day for a lot of maintenance. Unexpectedly though, in question 6, we stated that Saturday is a day where there aren't frequent delays. But this graph tells us that there are longer delays on Saturdays as it is the day with the highest average delay time. Since the line of best fit is still at a negative slope, we can also reiterate some conclusions from question 6 (i.e. congestion plays a huge factor).

(b) Time of day and gaps between trains

In this question, we are able to use some data from question 1 in regards to the hour of the day and all other factors. For this question specifically, we will be looking at the correlation between the hour of the day, and the gaps between trains. Intermediate steps to produce the scatter plot can be found in the .ipynb file.

Scatter plot representation:



Correlation coefficient matrix:

| | Min Delay | Min Gap | Vehicle | Hour of the day |
|---|---|---|---|---|
| **Min Delay** | 1.000000 | 0.919572 | 0.543298 | 0.271751 |
| **Min Gap** | 0.919572 | 1.000000 | 0.672003 | 0.321552 |
| **Vehicle** | 0.543298 | 0.672003 | 1.000000 | 0.040045 |
| **Hour of the day** | 0.271751 | 0.321552 | 0.040045 | 1.000000 |

Since the data corresponds to the hour of the day, we see that the data "bounces" in the scatter plot as it depends on the time of day (i.e. splicing the data into buckets of "morning", "afternoon" and, "evening" would produce a strong Pearson correlation coefficient). Relatively, this is expected because the correlation between "delays" and "gaps" is strongly and positively correlated, with a correlation coefficient at 0.92. By this piece, we can reference similar intuitions like we did in question 2 (i.e. reinforced maintenance should occur prior to subway operating hours and rush hour).

Focusing on the Pearson coefficient correlation matrix, we see that the overall correlation between the hour of the day and gap between delays is positively moderate, with a correlation coefficient of 0.32.

8. **How has the number of delays progressed per week during January? That is, did he number of delay increase, decrease or remain stable during this month as weeks progress?**

Our data would have been skewed if we tackle this issue head on. This is because of an instance like question 6, where there are going to be 5 weeks in January, but one out of the five weeks is going to contain three days instead of a typical seven days. To combat this, a better analysis is to look at the progression of each individual day of the week throughout the month, and the frequency of delays that occurred for that week.

Our data will be split into 5 different weeks:

| Week | Date |
|---|---|
| 1 | January 1-7 |
| 2 | January 8-14 |
| 3 | January 15-21 |
| 4 | January 22-28 |
| 5 | January 29-31 |

However, we represent the data for a specified day of the week (i.e. Week progression for Monday, Week progression for Tuesday etc.). In this case, we have been able to determine which delay goes to what week it occurred. Sample data is shown below:

Data that we will be using:

| Date | Time | Day | Station | Code | Min Delay | Min Gap | Bound | Line | Vehicle | Date (number representation) | Week |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **2018-01-01** | 00:29 | Monday | SHEPPARD WEST STATION | MUATC | 10 | 15 | N | YU | 5986 | 1 | 1 |
| **2018-01-01** | 01:07 | Monday | DUNDAS STATION | MUNCA | 0 | 0 | NaN | YU | 0 | 1 | 1 |
| **2018-01-01** | 01:22 | Monday | MUSEUM STATION | MUSC | 0 | 0 | N | YU | 5751 | 1 | 1 |
| **2018-01-01** | 01:28 | Monday | BAY LOWER | EUOE | 0 | 0 | NaN | BD | 5222 | 1 | 1 |
| **2018-01-01** | 01:39 | Monday | MUSEUM STATION | MUO | 6 | 11 | S | YU | 5781 | 1 | 1 |

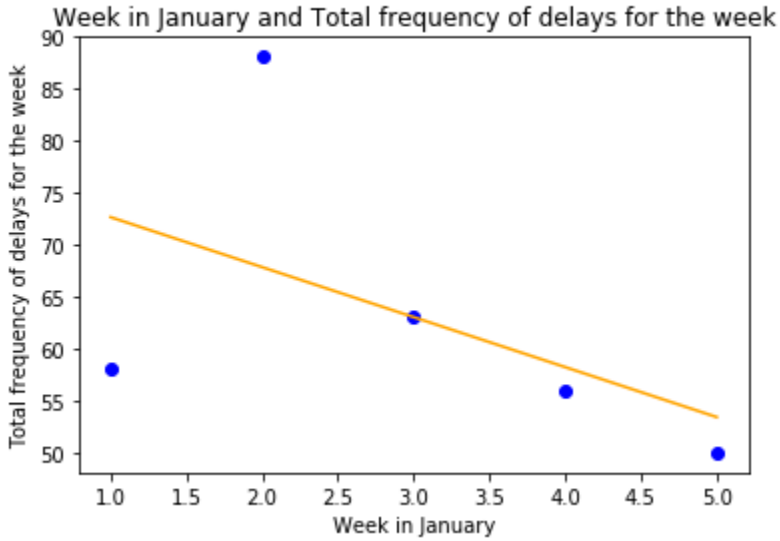
To give an introductory picture, all days of the week have the following in the ipynb file: Week progression chart for the day of, Scatter plot representation for the day of, and the Pearson Correlation Coefficient Matrix. However, to concisely insert our intuitions (and to save document space), we will be using the scatter plot to determine the progression of a specified day, so that we can analyze our data visually. What is going to be shown is Monday's data (for an introductory picture), but all other pieces can be found in the .ipynb file.

On Monday we have the following data:

Week Progression of Monday:

| Week | Frequency of Delays |
|---|---|
| **1** | 58 |
| **2** | 88 |
| **3** | 63 |
| **4** | 56 |
| **5** | 50 |

Scatter plot representation of Mondays:

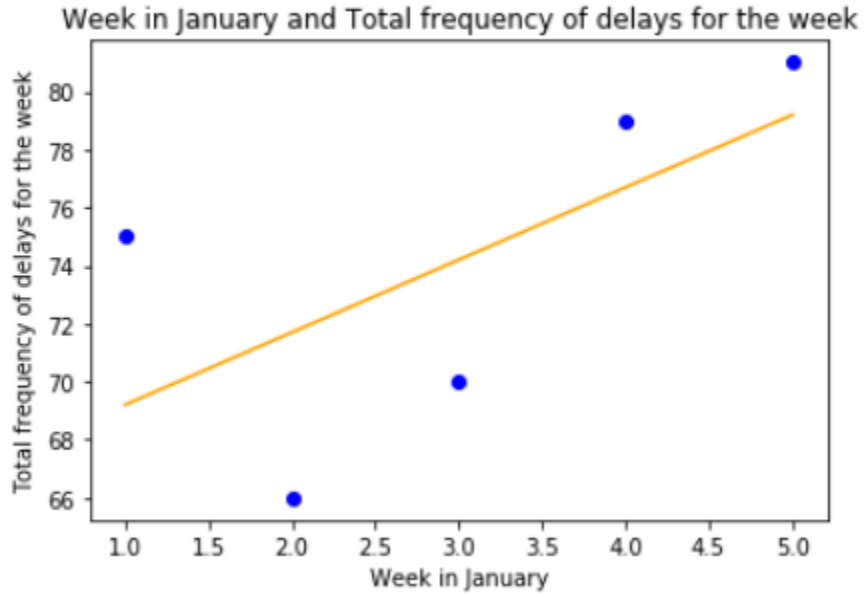



Correlation Coefficent Matrix:

| | Frequency of Delays | Week in January |
|---|---|---|
| **Frequency of Delays** | 1.000000 | -0.515207 |
| **Week in January** | -0.515207 | 1.000000 |

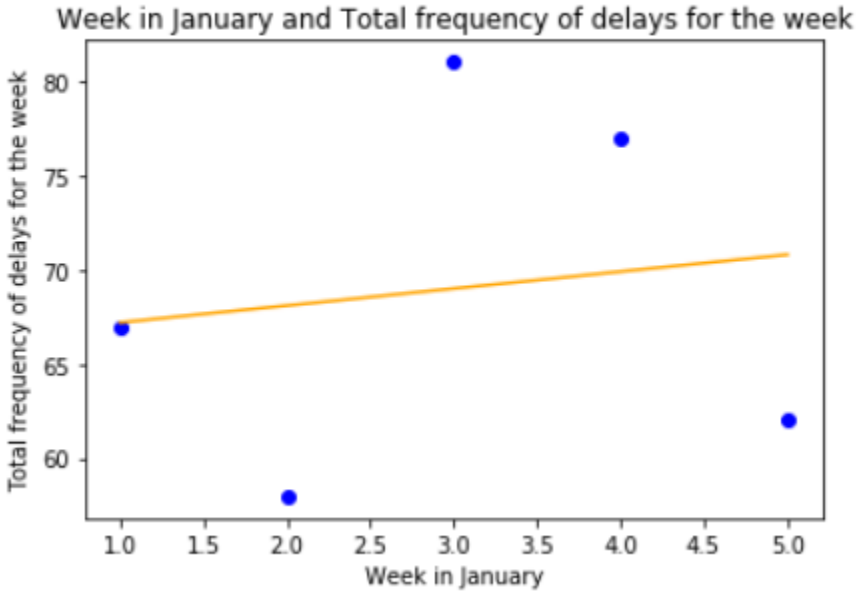
Given the Pearson correlation coefficient matrix, where the correlation coefficient between the week in January and the frequency of delays is -0.52, we can see that the number of delays on Monday decreases as time progresses due to the negatively strong correlation.

Scatter plot representation of Tuesdays:




Given the Pearson correlation coefficient matrix, where the correlation coefficient between the week in January and the frequency of delays is 0.63 (found in the ipynb file), we can see that the number of delays in Tuesday increases as time progresses due to the positively strong correlation.

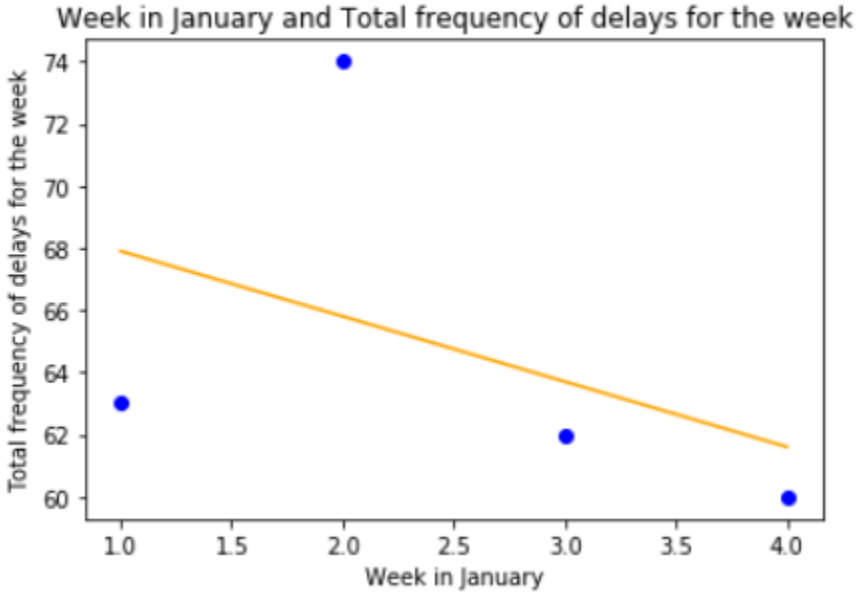
Scatter plot representation of Wednesdays:




Given the Pearson correlation coefficient matrix, where the correlation coefficient between the week in

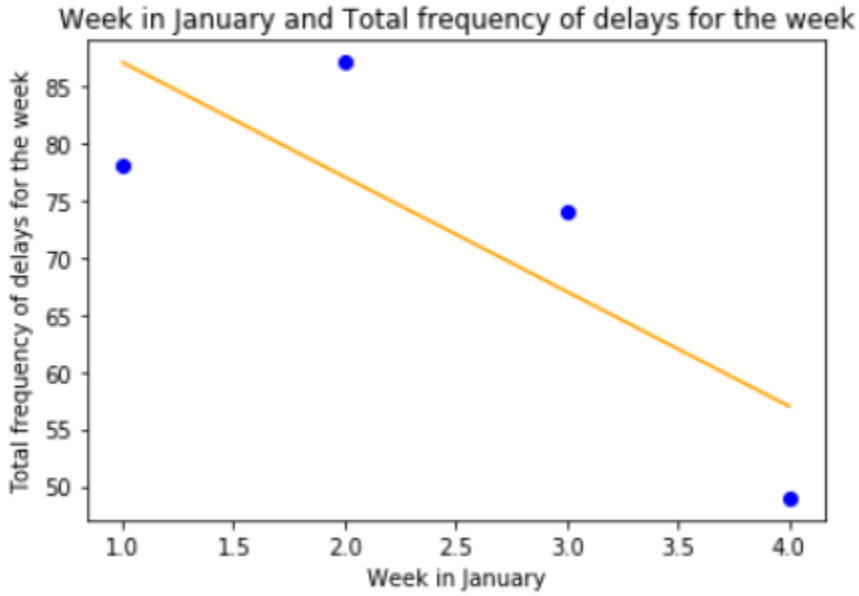
January and the frequency of delays is 0.15 (found in the ipynb file), we can see that the number of delays on Wednesdays stays moderate as time progresses due to the positively weak correlation.

```
Scatter plot representation of Thursdays:
```



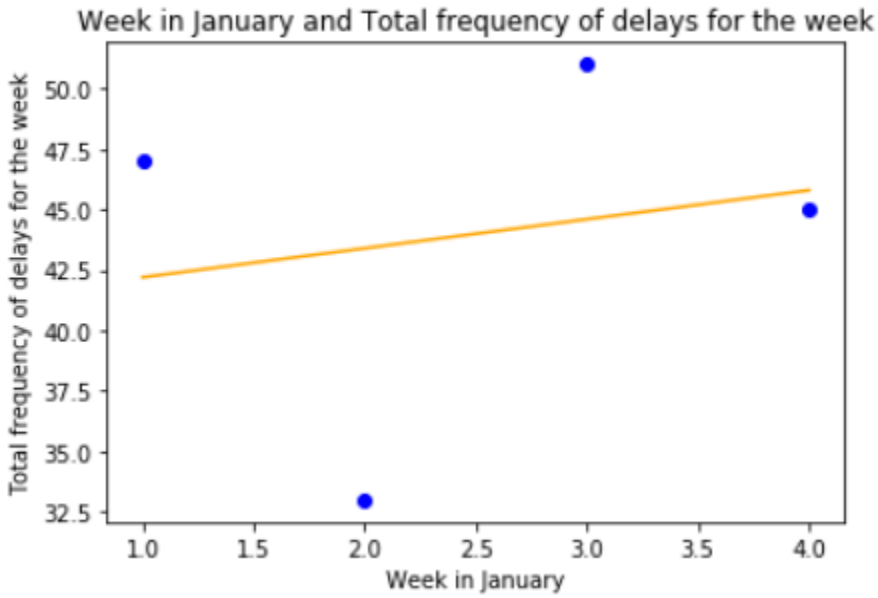

Given the Pearson correlation coefficient matrix, where the correlation coefficient between the week in January and the frequency of delays is -0.43 (found in the ipynb file), we can see that the number of delays on Thursdays decreases as time progresses due to the negatively moderate correlation.

```
Scatter plot representation of Fridays:
```



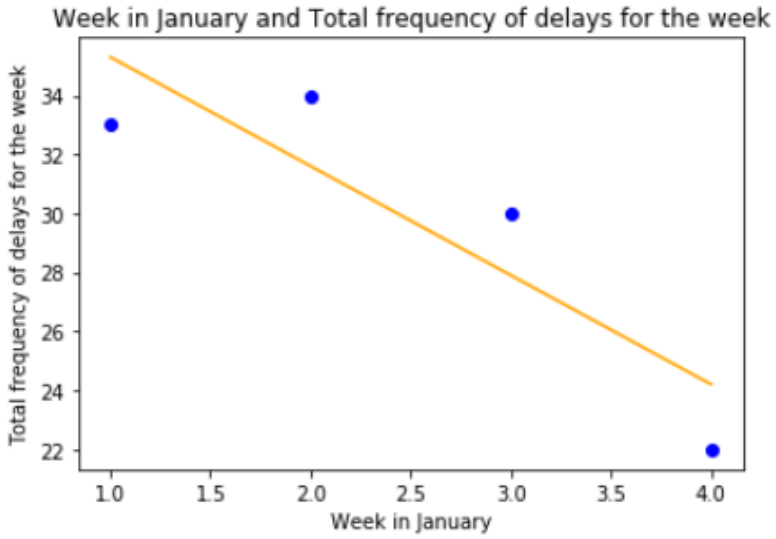
Given the Pearson correlation coefficient matrix, where the correlation coefficient between the week in January and the frequency of delays is -0.79 (found in the ipynb file), we can see that the number of delays on Fridays decreases as time progresses due to the negative strong correlation.

Scatter plot representation of Saturdays:




Given the Pearson correlation coefficient matrix, where the correlation coefficient between the week in January and the frequency of delays is 0.2 (found in the ipynb file), we can see that the number of delays on Saturdays stays moderate as time progresses due to the positively weak correlation.

Scatter plot representation of Sundays:



Given the Pearson correlation coefficient matrix, where the correlation coefficient between the week in January and the frequency of delays is -0.88 (found in the ipynb file), we can see that the number of delays on Sundays decreases as time progresses due to the negatively strong correlation.

In this case, since different days have different results in terms of their week progression, it is difficult to say whether all of January's frequencies of delays increased/decreased/stayed moderate. Instead, what can be more meaningful is the specific day of the week:

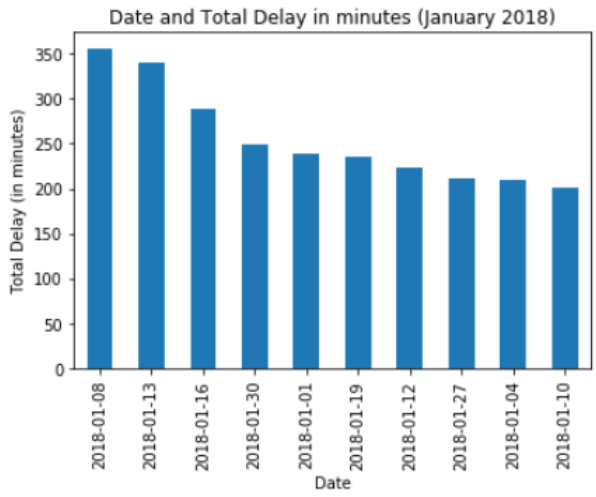
| Day | Week Progression |
|---|---|
| Monday | Decreases |
| Tuesday | Increases |
| Wednesday | Stable |
| Thursday | Decreases |
| Friday | Decreases |
| Saturday | Stable |
| Sunday | Decreases |

Obliviously, if we were to use the "mode" as a measure, then it seems as though the number of delays decreased as time progressed. In a managerial point of view, it's best to first research benchmarks compared to previous years as to how the staff performed. If the staff has done well (i.e. exceeds historic averages), then it's highly suggested to reward the staff that reduced the number of delays so that they can further continue their efforts of reducing the amount of delays.

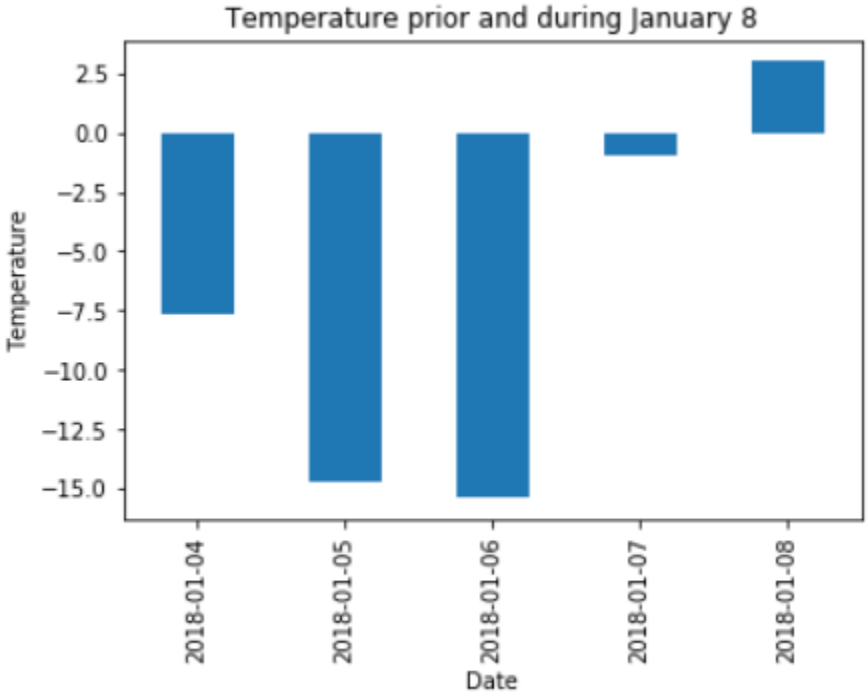
9. **What was the worst day in January in terms of total delay time? What reason would you attribute to this phenomenon? Answer the same question considering the best day in terms of total delay time**

After conducting the intermediate steps to find the absolute "worst" day for delays we are able to find the top 10 worst days of all time. Shown below is a visual representation of the worst days in terms of total delay time

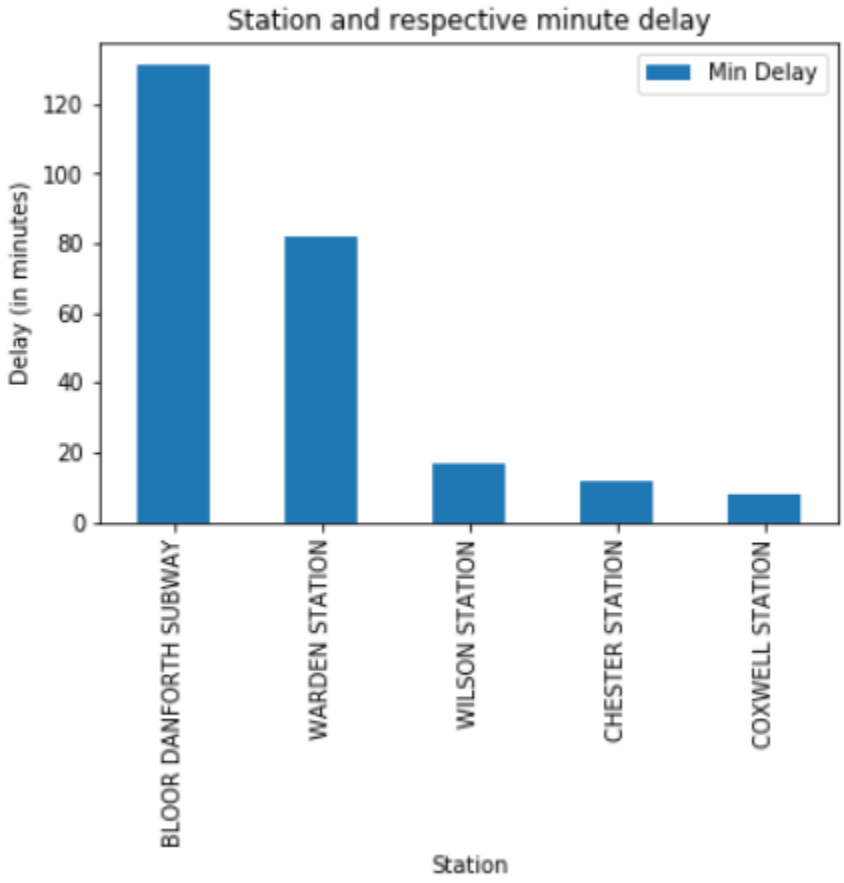
```
A visual representation:
```



Thus, we can conclude that the absolute worst day in terms of delays in January occurred on January 8. One retrievable piece of data that may have attributed to January 8 being the absolute worst day was the weather. Thanks to the weather network, a csv file was made to demonstrate the weather on and prior to January 8. Thus a visual representation is shown below:




Since there are rising temperatures for as low as -14 degrees Celsius, it can be theorized that there may have been a lot of ice/slush on that day which may have attributed to injuries on subway platforms (especially those that are outside). However, further investigation has been conducted on January 8. Visually, we are able to see the top 5 "stations" that have been affected by one sole delay:




TTC in this case lists the whole Bloor Danforth line as a major delay. Investigating further, we can look at the raw data that attributed to how the Bloor Danforth line got listed as a top sole delay:

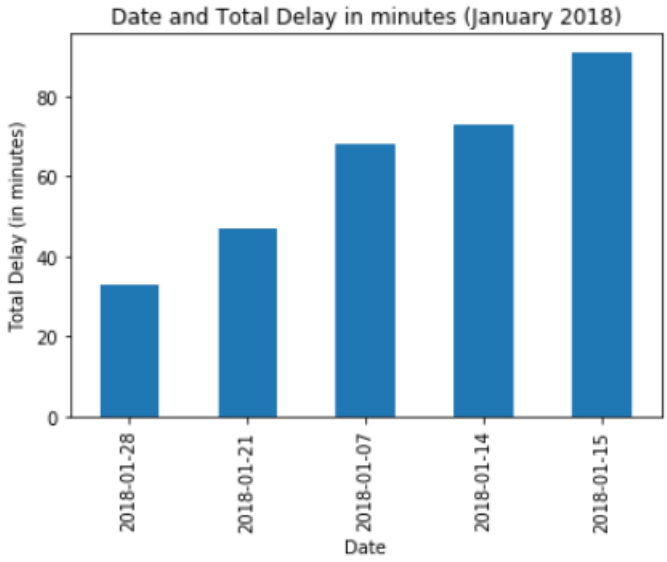
```
Top 5 instances of the delays:
```

| Date | Time | Day | Station | Code | Min Delay | Min Gap | Bound | Line | Vehicle |
|---|---|---|---|---|---|---|---|---|---|
| **2018-01-08** | 05:45 | Monday | BLOOR DANFORTH SUBWAY | MUFM | 131 | 134 | E | BD | 5212 |
| **2018-01-08** | 15:28 | Monday | WARDEN STATION | PUSSW | 82 | 87 | E | BD | 5006 |
| **2018-01-08** | 05:36 | Monday | WILSON STATION | EUTL | 17 | 0 | S | YU | 5411 |
| **2018-01-08** | 22:11 | Monday | CHESTER STATION | MUDD | 12 | 16 | E | BD | 5085 |
| **2018-01-08** | 13:54 | Monday | COXWELL STATION | TUNOA | 8 | 11 | W | BD | 5276 |

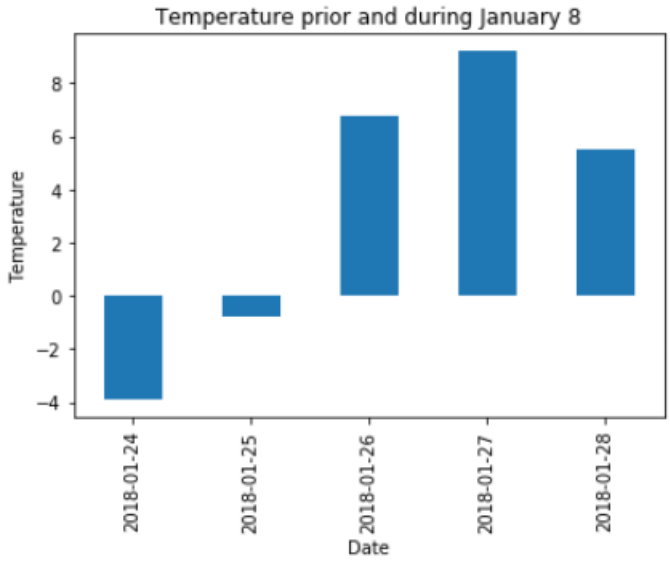
Investigating even more further, the code "MUFM" means "Force Majeure" where Wikipedia describes those key words as an "chance/occurrence/unavoidable accident". The real question that is asked "Well what exactly is this phenomenon?". Looking through CityNews' archives on January 8, it appears that power from Hydro One has been cut in the morning. As a result, stations from Broadview to Woodbine

were not capable of operating, while at the same time, making trains inoperable on the Danforth and Greenwood yard.

In comparison, the graph below shows the "best days" in terms of total delay time:




Thus, we can conclude that the absolute best day for the month of January in terms of total delay time occurred on January 28, 2018. Playing with what was mentioned earlier in terms of weather, we can theorize that weather may have attributed to January 28 being the best day perhaps even having warmer conditions:

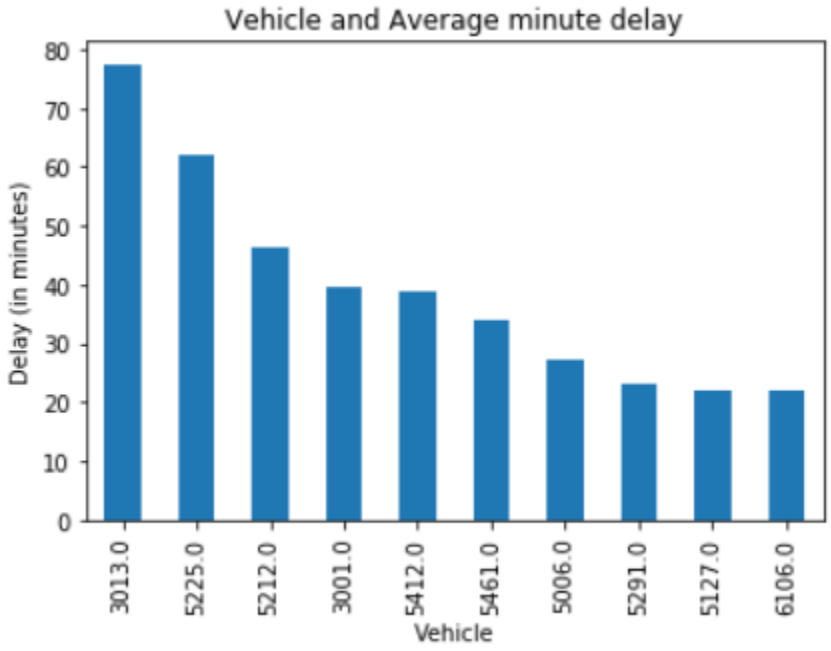



As an implication, what can be assumed is that a lot of ice/snow has been melted due to the rise of warmer, safer conditions. Thus, causing less delays. Also, in reference back to Question 7 we have seen that the best days of the week occur on Sundays, to which January 28 falls under.

10. **What are the top 10 vehicles associated with longer delays? That is, is there any link between the vehicles and delays? What reason could you attribute to this longer delay time?**

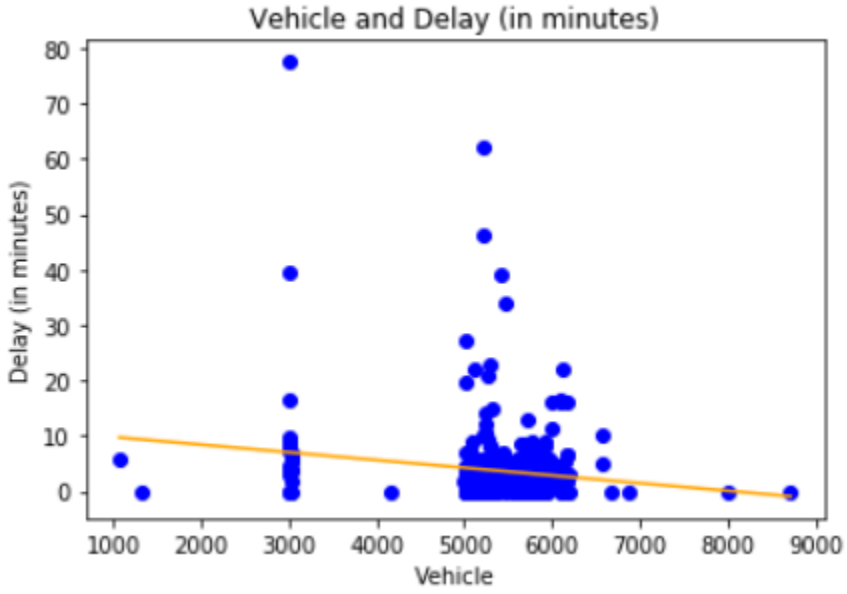
Initializing the data for this specific question was fairly difficult to start, since a lot of vehicle numbers were listed as 0. What was interestingly found was that the SRT trains are immensely different from the trains from all other lines due to its railway system (even in the metadata given, the SRT trains have a different set of codes!). In this case, we picked up on the fact that any train that is running on the SRT system is in the 3000 series. Likewise, We also picked up on the fact that any train that can run the Bloor-Danforth Line is in the 5000 series. Proper adjustments have been made to fit our data.

In regards to the first question asked, the top 10 vehicles with the highest average minute delay are visually shown below. We use averages because it best assumes performance of each vehicle:



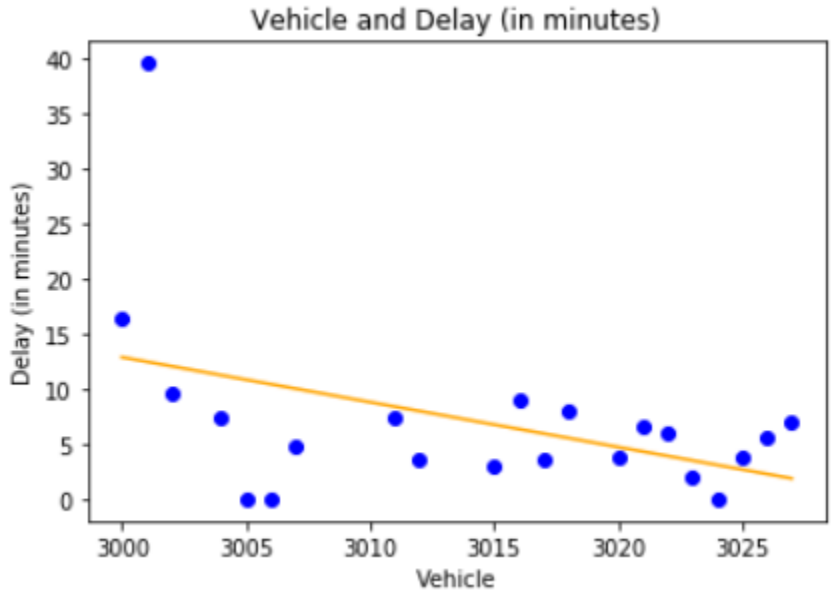

In this case we see the top two trains, one running on the SRT and the other being in the Bloor Danforth Line. This does raise a question, "does vehicle number correlate to the average delay?"

In this case, we have drawn an overall scatter plot of all data of a specified vehicle number



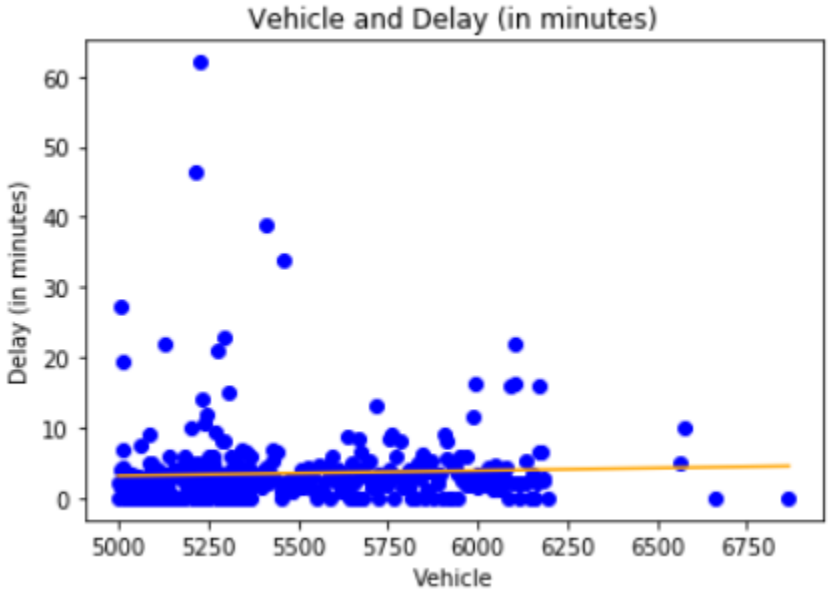

As expected, it's weakly correlated (with a Pearson correlation coefficient at -0.15). One thing to mention though, is the vertical stretches on the 3000 series and the 5000 to 7000 series. In this case, it's best to splice our data up into two parts, one being all SRT vehicles and the "other vehicles". In this case, the "other" vehicles will be consisting of vehicles in the 5000 to 7000 series since data aside from the 3000-5000-6000 series appears to be stray outliers.

Below is the SRT scatter plot




The Pearson correlation coefficient matrix indicates that the correlation coefficient between vehicle and average delay of that vehicle is -0.44, which is a moderate correlation coefficient. By a managerial

standpoint this means that the higher the vehicle number, the less there are delays. This does makes sense if the vehicle number represents the acquitted train. For further inference we can make a claim that newer vehicles have less maintenance needs compared to older vehicles whose value has depreciated.




The Pearson correlation coefficient matrix indicates that the correlation coefficient between vehicle and average delay of that vehicle is approximately 0.05, which shows no correlation. Though, something that looks interesting is that the line of best fit goes upward. Recall that trains in the 5000 series represent a large amount of trains on the Bloor-Danforth line, but the positive slope may be attributed to the vehicles in the Yonge University line. This is fairly interesting considering that the Yonge University line is the oldest line in the Toronto subway system. To which we can somewhat come to a conclusion that trains on the Yonge University line are being fairly robust. With the upward trend, we can somewhat come to an agreement that newer models on the Yonge University line are being depreciated quicker due to the path that the Yonge University line faces, and these models need more maintenance fixtures (i.e. some stations are outdoors compared to the Bloor Danforth line where the Bloor Danforth line is almost entirely underground).

# Bibliography

Aguilar, B. (2017, December). TTC offers free rides on New Year's Eve. Retrieved March 3, 2018 from https://www.thestar.com/news/gta/2017/12/29/ttc-offers-free-rides-on-new-years-eve.html

Canada Census 2016: Toronto growth well above the already high national average. (2017, February). Retrieved March 1, 2018 from http://nationalpost.com/news/toronto/canada-census-2016-toronto-growth-well-above

Force Majeure. (n.d.). Retrieved March 3, 2018 from https://en.wikipedia.org/wiki/Force_majeure

Interactive Map (n.d.). Retrieved March 4, 2018 from https://www.ttc.ca/Subway/interactive_map/interactive_map.jsp#

Jan, 2018. (2018, February). Retrieved March 1, 2018 from https://www.theweathernetwork.com/ca/monthly/ontario/toronto?year=2018&month=1&dispt=calendar-container-monthly

News Staff. (2018, January). Power restored in east end, TTC subway service resumes. Retrieved March 3, 2018 from http://toronto.citynews.ca/2018/01/08/large-power-outage/

THE SCARBOROUGH RAPID TRANSIT LINE. (2017, October). Retrieved March 4, 2018 from https://transit.toronto.on.ca/subway/5107.shtml

Toronto – Canada's business and financial capital. (n.d.). Retrieved March 1, 2018 from http://www.tfsa.ca/toronto-advantage/

Service Information. (n.d.). Retrieved March 4, 2018 from https://www.ttc.ca/Riding_the_TTC/Frequently_Asked_Questions/Service_Information.jsp