Analysis the Delays of TTC Buses in June 2021

Dhanraj Patel

2021-07-18

Contents

Overview

The city of Toronto Prides itself on having one of the best public transportation systems in all of the world. While riding on one of the cities numerous TTC buses one is able to reach every corner of the City and while the city strives to keep their system running on a timely bases, there are situations which cause a delay. Using the data provided by the city of Toronto we will be analyzing the data of all the TTC bus delays that occurred in June 2021.

Data Collection Process

Before delving deeper into the data it is important to understand the process of how the data was collected. Whenever a TTC bus delay occurs, the TTC (Toronto Transit Commission - the public agency in charge of transportation in Toronto) records the event in detail. The TTC makes sure to record the time and day of the delay, the route of the delayed bus, location of delay, the amount of time the bus was delayed for and the cause of the delay. Due to the fact that the TTC is a public agency all the information they record is then passed on to the city of Toronto who then publish it for the public to view. The data is available on https://open.toronto.ca.

Cleaning of data

After retrieving the data set it was important to clean the data so that the results generated would be an accurate reflection of the data. In order to clean the data it was first important to eliminate all the duplicate entries. Within the data set there were instances were the same delay was recorded multiple times. For the purposes of this analysis, since we only want to include one recorded instance of all delays, all the duplicates in the data set were removed.

In addition, within the data set both of the columns of route and direction of the recorded delayed bus contained missing values. As for the direction column about 10 percent the values were missing and for route about 0.5 percent of the values were missing. Due to the fact that the values for route and direction were not used for the numerical or graphical summaries, the missing values for both columns were not removed as not not alter the accuracy of the data that would be caused by removing data.

Lastly, since the delay times were recorded down to the minute, I introduced a new variable called time_interval. With this, I split up the time into 4 sections so incident delays can be analyzed by a factor of time intervals and not by the exact time. This is to avoid overly granular data.

Important Variables

Before looking at the numerical and graphical summaries it is important to understand all the variables in the data set and what they represent.

- Date - The date that the TTC Bus delay occurred
- Route - The route the TTC bus was driving while it was delayed
- Time - The time that the TTC bus was delayed
- Day - The day of the week that the delay occurred
- Location - The intersection location of the TTC bus delay
- Min Delay - The time in minutes of the delay to the schedule for the following bus
- Min Gap - The total time of delay in minutes between the delayed bus and the bus before the delayed bus
- direction - The compass direction of the bus during the delay (e.g North)
- Vehicle - The unique identifier of the delayed bus
- Incident - The description of cause of the delay from the following:
  - Cleaning
  - Collision - TTC
  - Diversion
  - Emergency Services
  - General Delay
  - Held-by
  - Investigation
  - Late Entering Service
  - Late Leaving Garage
  - Management
  - Utilized Off Route
  - Vision
  - Security
  - Road Blocked - NON-TTC Collision
  - Operations-Operator
  - Mechanical
- Time interval - The time interval the delay occurred.
  - Morning: 00:00 to 06:00
  - Afternoon: 00:60 to 12:00
  - Evening: 12:00 to 18:00
  - Night: 18:00 to 24:00

Numerical summeries

With all the data cleaned, we can now move on analyzing important numerical summaries to help us better understand the data set. Within our data set it is important to consider both the location and spread of our data to get a clear idea of how the variables are distributed.

Comparing the time of TTC delays to the days of the week

The following table (labeled Table_1) displays numerical summaries of the delays that occurred in minutes for each day of the week.

(Table_1)

Table 1: The trimmed mean was trimmed by 10 percent

day	Min	Q1	Median	Q3	Max	IQR	Mean	Trimmean_10	Var	SD	Range
Friday	0	9.00	12	20.00	403	11.00	17.37	13.87	764	27.64	403
Monday	0	9.00	12	17.00	875	8.00	15.99	12.96	1399	37.40	875
Saturday	0	8.00	12	20.00	355	12.00	18.91	14.04	1030	32.10	355
Sunday	0	8.00	12	20.00	480	12.00	18.53	13.61	1165	34.13	480
Thursday	0	8.75	12	19.25	807	10.50	18.60	13.43	1582	39.77	807
Tuesday	0	8.00	12	18.00	681	10.00	16.96	12.81	1432	37.84	681
Wednesday	0	9.00	11	17.75	222	8.75	14.59	12.74	234	15.30	222

Gathering information from the table we can see that for each day of the week the minutes of delays are right skewed. From the table we see that for each day the spread of values from q3 to the min value is roughly a tenth of the number of values from max to q3. From this we can surmise that the large majority of delays that occur on the TTC are very short lasting roughly about 10-15 minutes, but there are rare circumstances where there are delays that last much longer. The rare long delays are influencing the mean of the time of the delays and also is the cause of the large amount of variance. Once you trim 10 percent of the mean, the trimmed mean of the delay time for each day of the week decreases. Since the data is right skewed, the best numerical summary would be the median which states the the average delay is around 10 minutes.

Comparing different causes to overall delay time

The following table (labeled Table_2) displays numerical summaries of the delays that occurred in minutes for each cause of delay.

(Table_2)

Table 2: The trimmed mean is trimmed by 10 percent

incident	Count	Min	Q1	Median	Q3	Max	IQR	Mean	Trimmean	Var	SD	Range
Cleaning	1834	0	8.00	11.0	15.00	40	7.00	12.40	11.81	35	5.89	40
Collision - TTC	208	0	8.00	10.0	15.00	60	7.00	11.66	10.65	80	8.93	60
Diversion	29	0	23.00	57.0	136.00	333	113.00	91.41	80.76	7950	89.16	333
Emergency Services	146	0	0.00	10.0	14.00	33	14.00	9.29	8.51	68	8.22	33
General Delay	112	0	14.25	28.5	65.00	875	50.75	76.79	41.90	21981	148.26	875
Held By	17	0	8.00	12.0	18.00	30	10.00	13.12	12.87	78	8.84	30
Investigation	111	0	8.00	12.0	20.50	480	12.50	21.93	13.63	2525	50.25	480
Late Entering Service	3	15	18.00	21.0	22.00	23	4.00	19.67	19.67	17	4.16	8
Late Leaving Garage	1	20	20.00	20.0	20.00	20	0.00	20.00	20.00	NA	NA	0
Management	5	4	7.00	11.0	12.00	30	5.00	12.80	12.80	103	10.13	26
Mechanical	1175	0	8.00	11.0	16.00	43	8.00	13.00	12.23	48	6.93	43
Operations - Operator	714	0	10.00	20.0	24.00	200	14.00	17.96	17.56	120	10.96	200
Road Blocked - NON-TTC Collision	105	0	25.00	49.0	87.00	403	62.00	78.66	61.04	7486	86.52	403
Security	251	0	8.00	10.0	18.00	145	10.00	13.08	11.96	148	12.18	145
Utilized Off Route	30	0	10.25	20.0	23.75	297	13.50	26.23	17.92	2695	51.91	297
Vision	85	0	9.00	10.0	17.00	60	8.00	13.55	12.45	74	8.60	60

While looking at this table of all the incident categories and the minutes of delay they caused, what immediately jumps out is that the incident Late Leaving Garage has a NA value for variance and SD. This is because in all of June 2021, there was only one recorded instance of this cause of delay. With only one recorded data entry, it is not possible to find variance or SD so the table recorded it as NA. When comparing delays between different incident categories, it's important to take into account how many incidents each delay had, the number of incidents can significantly skew the data. The higher to count of the incident category, the more confidence we can have in its numerical summary.

While comparing all the incidents, we notice that general delays and diversions result in the longest delays. They have both the largest mean time and trimmed mean time. In addition it is important to note that those two categories also have the greatest amount of variance, with each of its quartiles having a greater spread than any of the other categories. So while the average of of these 2 delays has a large delay time, the delay times are spread out.

Lastly, aside from a few categories of incidents we can see that most categories have a trimmed mean of roughly 20 minutes with a significant variance. This mirrors the data of the days of the week in the last section where we found that the data was right skewed.

Comparing the times at which the delays occurred

The following table (labeled Table_3) displays numerical summaries of the delays that occurred for each time interval the delay occurred in.

(Table_3)

Table 3: The trimmed mean is trimmed by 10 percent

time_interval	Count	Min	Q1	Median	Q3	Max	IQR	Mean	Trimmean	Var	SD	Range
Afternoon	1736	0	8	11	18	875	10	16.13	12.59	1608	40.10	875
Evening	1876	0	9	12	19	403	10	15.96	13.51	478	21.87	403
Morning	300	0	8	16	30	355	22	25.80	16.76	1788	42.29	355
Night	914	0	8	11	20	419	12	18.01	13.09	951	30.85	419

From the table above we are able to see that identically to the previous two tables, the delayed time for all intervals are right skewed. The majority of the delays only last a few minutes while on rare occasions, there are delays that last much longer. We know this from the fact that in all time intervals, the spread of values from q3 to the min value is significantly smaller than the spread from the max value to q3.

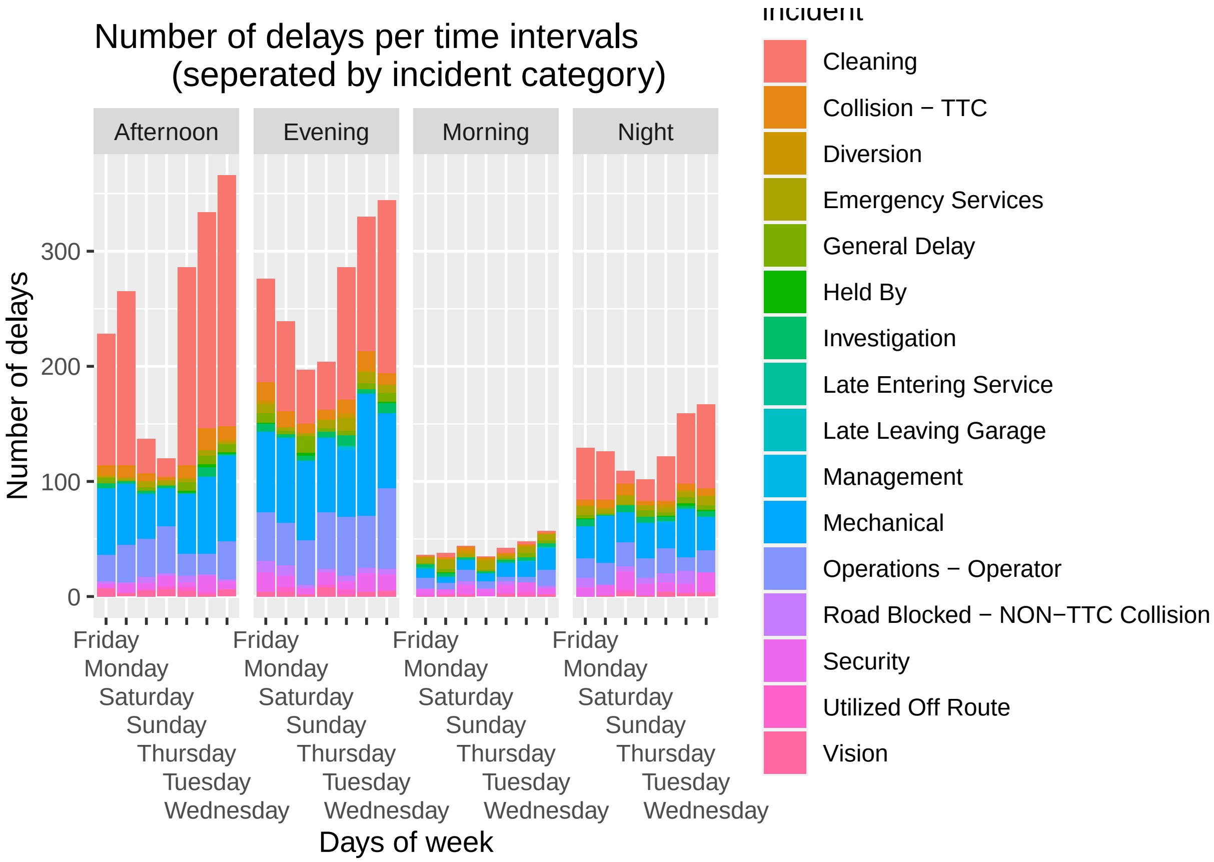
Comparing individual times of day we see that the morning has the largest mean time. Although once you apply a trimmed mean the gap between the mean for the morning and the rest of the interval gets smaller. From this we can infer that the morning time have a large amount of variance in their delay times, this is reinforced as we see that the variance for the morning is greater than all the other intervals. It is also important to note that the number of incidents that occur in the morning are fewer than the rest of the intervals. This can explain the greater variance of the morning interval.

Graphical summaries

Aside from numerical table summaries it is important to see graphical summaries to see trends in the data.

(Graph_1)

Comparing delays between time of day, day of week and category of incident



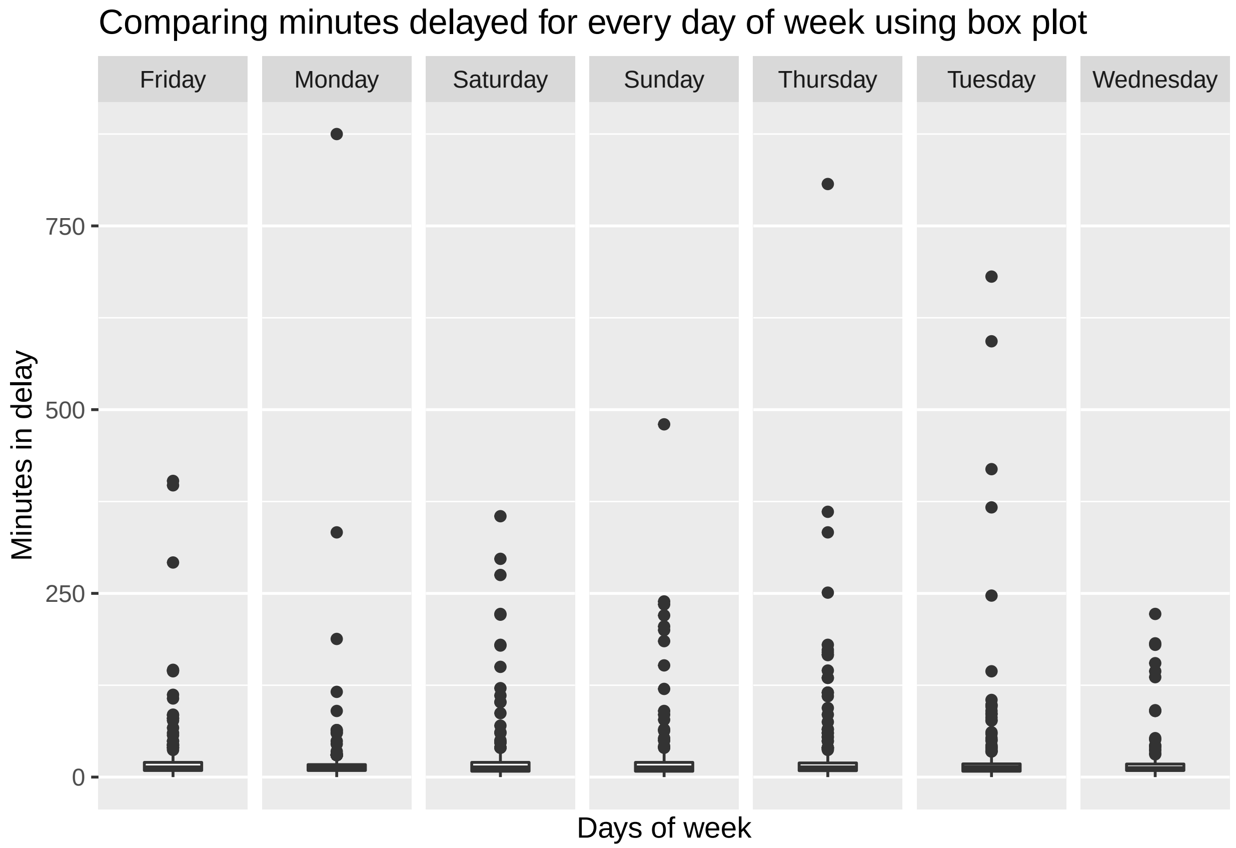
In this graph we can see the number of delays for every day of the week separated by time of day. Within the count of delay incidents we can also the causes of the delays for that day.

Reminder: Morning: 00:00 to 06:00 Afternoon: 06:00 to 12:00 Evening: 12:00 to 18:00 Night: 18:00 to 24:00

At a quick glance of graph_1 we can infer that the afternoon and evening have the highest amount of delays, while the morning has the smallest amount of delays. In addition we can see that in all intervals other than morning, cleaning and mechanical delays are noticeably the largest causes of delays. In contrast, in the morning the causes are more evenly distributed. Also, when comparing the number of delays between the time interval and day of the week we see for all time intervals the two days of the week that have the highest amount of delays are Tuesday and Wednesday. Lastly, for mornings it seems that all days of the week have a roughly even amount of delays, where as for the rest of the time intervals it's is more varied.

(Graph_2)

From graph_2 we are able to see the spread of minutes in delays for every day of the week separated by their quartiles.



From graph_2 we can see that for all days of the week, the data of minutes of delays is right skewed, most of the data is concentrated near the bottom, meaning most of the delays are only a few minutes long. However the inclusion of significant outliers of really lengthy delays skews the data. From this skew it makes it look like the average delay is much higher than it really is and impacts the variance greatly.

Resources

This entire document and all its contents were made by using the programming language R and Rstudio. From pulling the data, to cleaning the data, manipulating the data and creating tables and graphs, was all done using R and Rstudio.