TTC Bus Delay Analysis using Linear Predictive Model of Delay-Causing Incidents

# Introduction

No doubt that buses present an important mean of transportation in metropolitan cities such as city of Toronto, thousands of Torontonians ride them daily to get to their work, schools and many different places. The repeated delays of the buses are equal to enormous waste of time and money and lead to a lot of frustration of the bus riders who facing these delays.

In my capstone I plan to use TTC Bus Delay Data provided by TTC to form a pattern of these delays over past 56 months from Jan 2014 to Aug 2018, study the main incidents causing these delays, correlate the delays with different seasons of the year, different times of the day, road construction works etc…

I will use R-programming language in discovering the data, also use it in building linear regression model that predict the duration of time delay caused by different incidents such as Mechanical, Road Diversions and so on.

Research questions: What are the main incidents causing the delay of TTC buses?

# Literature Review

Key words: Headway, Bus Bunching, EWT (Excessive Waiting Time), PBCs (Performance Based Contracts)

Headway**:** the time between consecutive services, headway of 30 mins is equivalent to a bus comes every half hour

Unreliable bus services can be characterized by:

1. Unequal headways or bus bunching that lead to longer waiting and travel times for bus passengers
2. High passenger loads on routes having unequal arrival times, these irregular arrival times will cause sever crowd on the first bus that arrives after a long headway
3. Increase dwell time at the bus stops to cater for higher passenger movements On and Off the buses

All the above will lead to longer EWT for bus users hence reduce bus service reliability.

While few external factors may contribute to instability of bus services such as Traffic disruptions, Weather conditions and Boarding durations, there are some reasons of service irregularity are within operators’ control, for example, Staff shortages, Mechanical issues etc.…

Prior to improve bus services reliability, it’s necessary create a system of measuring their performance, here I’ll give two examples used worldwide:

* Bus Dispatching System (BDS)

1. Headway deviation: the difference between actual headway and schedule headway
2. Actual headway spatial distribution: depicts the proportion of actual headway from the schedule headway against different stops along the route

* Transport for London (Tfl)

Characterized buses based on the headway durations or number of buses per unit time to:

1. Low frequency buses having 15 minutes headways or more, those buses are assessed based on average excess wait time (EWT) experienced by the commuters
2. High frequency buses having less than 15 minutes headways, they are assessed based on punctuality, means percentage of buses departing On-time according to bus schedules

There are different approaches to implement scientific data-driven methodologies to improve bus service performance:

1. Adjust bus time tables considering observed delays and passenger numbers, first, cluster lines according to their delay patterns to extract lines of most need for time table revisions, then develop a simple approach to modify the time table
2. Achieving reliability through performance-based contracts PBCs, by rewarding operators for achieving a minimum service level (MSL) and for increase in ridership, more revenue for the operators meet social benefits such as reduced waiting time, reduced number of transfers and transfers of riders from car to public transport
3. Introducing headway balancer based on stochastic search, the balances come from introducing dwell time intervals in the schedule to hold the buses in control points in bus stations.

# Dataset

TTC Bus Delay Data

Owner Toronto Transit Commission

Currency August 2018

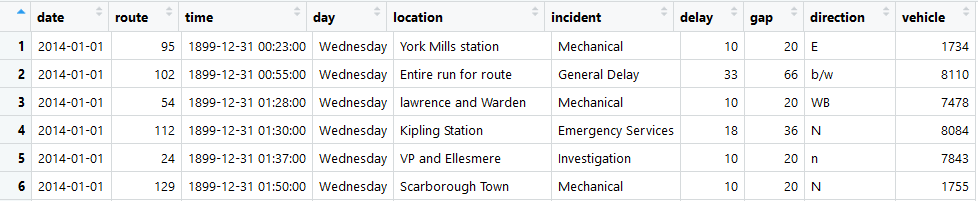
Format xls

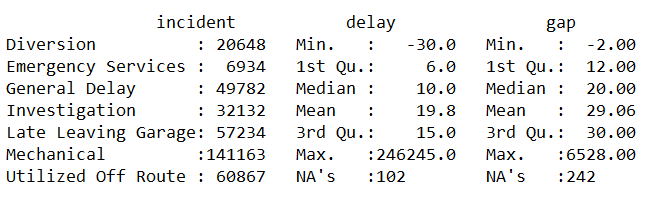
Refresh Rate Monthly

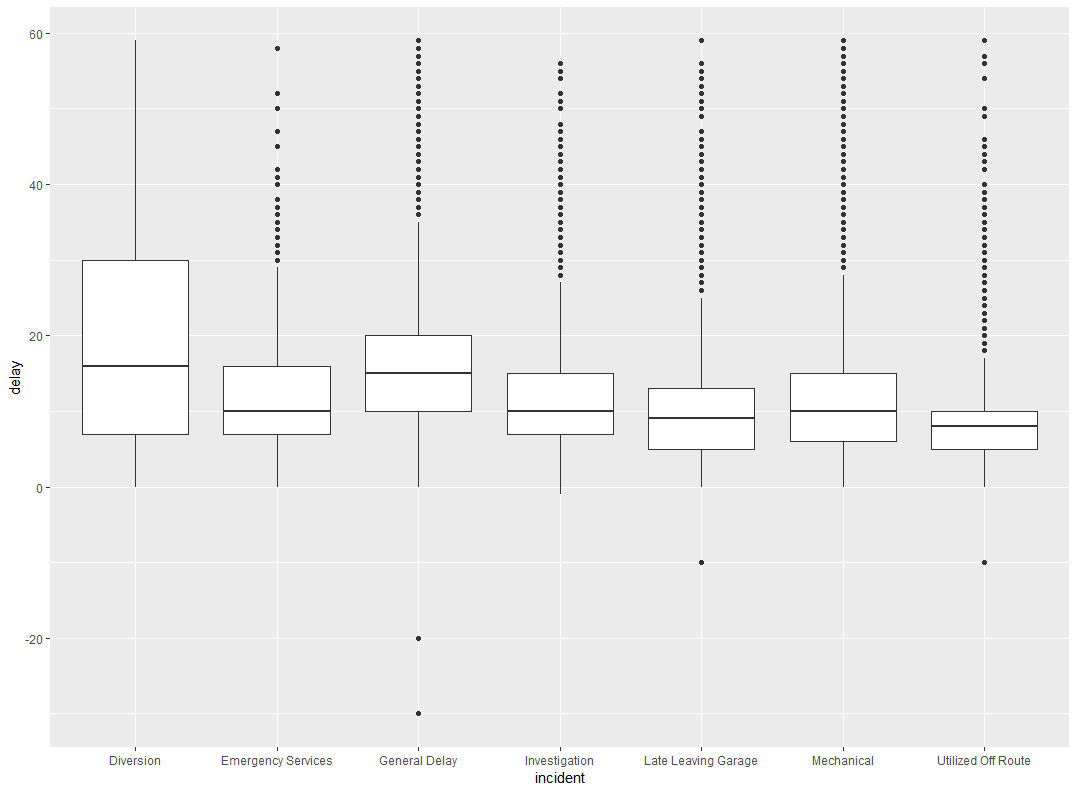
Contact Open Data Team [opendata@toronto.ca](mailto:opendata@toronto.ca)

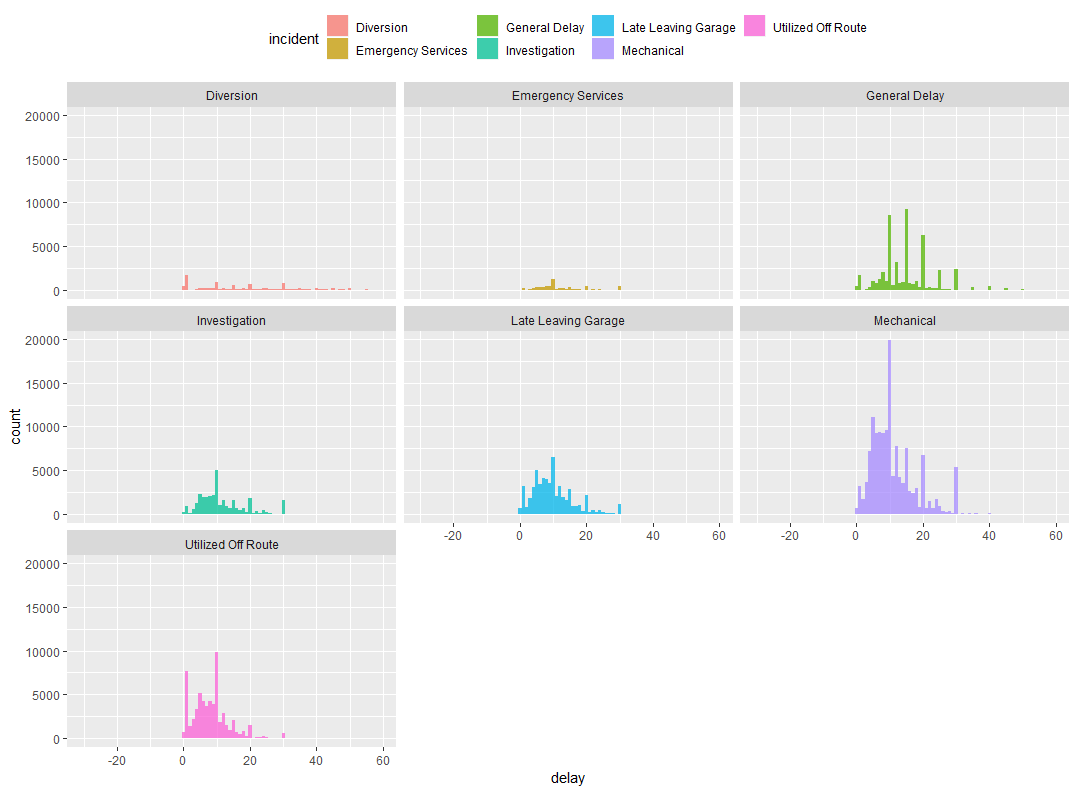
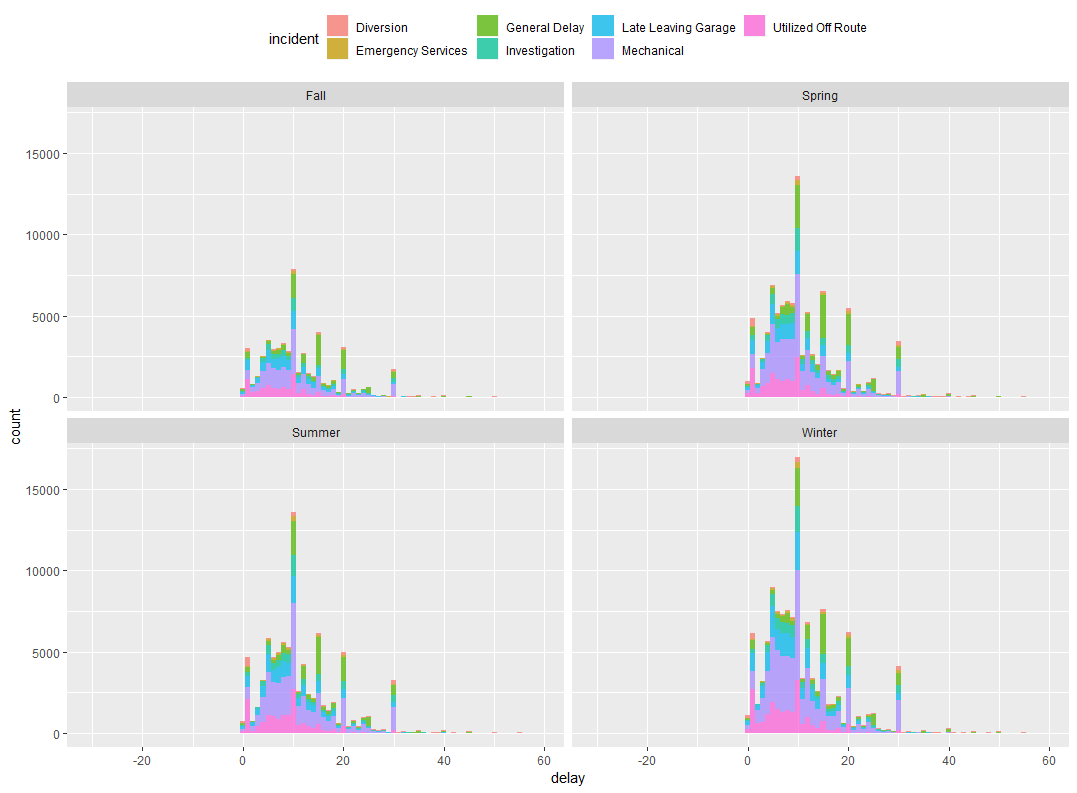
|  |  |  |
| --- | --- | --- |
| **Field Name** | **Description** | **Example** |
| Report Date | The date (YYYY/MM/DD) when the delay-causing incident occurred | 2017/06/20 |
| Route | The number of the bus route | 51 |
| Time | The time (hh:mm:ss AM/PM) when the delay-causing incident occurred | 12:35:00 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 |

Summary of the dataset









# Approach

1. Merge Monthly data into one sheet per year, then combine all years in one big data frame, then save it in CSV format file
2. Read CSV file remove routes have NAs and remove delays more than 60 mins as I assumed the maximum headway is 60 mins, so that the following bus will replace delayed those have delays more than 60 mins
3. Explore dataset using summary statistics and ggplot graphs in order to visualize the distribution of different incidents causing TTC delayed services
4. Divided the dataset into two parts, training part presents 70% random sample of whole dataset and remaining 30% is used for testing the model
5. Create linear regression model using “the time delay” as the response and “incidents” as the predictor attribute
6. Verify the model by predicting the time delay caused by the incidents of the test dataset, then calculate the error which equivalent to the difference between the actual and predicted values and show the errors in a histogram.

Hypotheses:

The main reason of Bus delays is the “Mechanical” incidents, to statistically proof that assumption, I ran z-test in which the population is all delays due to “Mechanical” and “non-Mechanical” reasons and the sample is the delays due to “Mechanical” reasons

Population mean (µo) = 11.45323 mins

Population Standard Deviation (σ) = 7.816287 mins

Mechanical mean (µ) = 11.22805 mins

Size of the sample (n) = 140988

Ho: Mechanical mean (µ) = Population mean (µo)

Ha: Mechanical mean (µ) ≠ Population mean (µo)

Z = (µ – µo)/( σ / √n)

Z = -10.81751

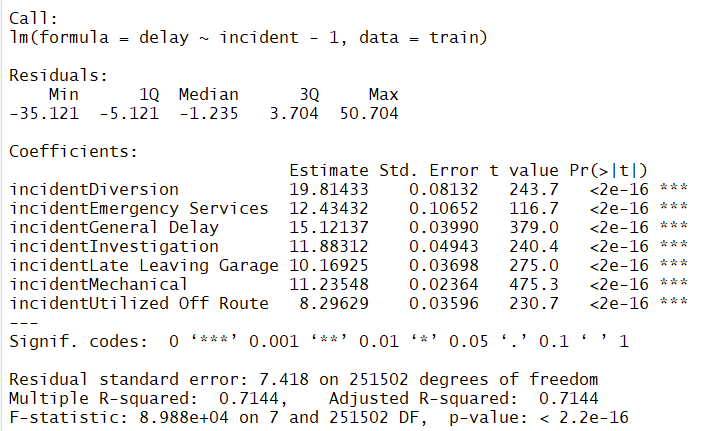
Significance α = .05

From z-table at α/2 = .025 , zc = -1.96

Since calculated z =-10 is located in the rejection region, then there is enough evidence to reject the null hypothesis. There is a difference between the means of “Mechanical” and of all incidents

|  |
| --- |
|  |
| Delay histogram of all incidents |

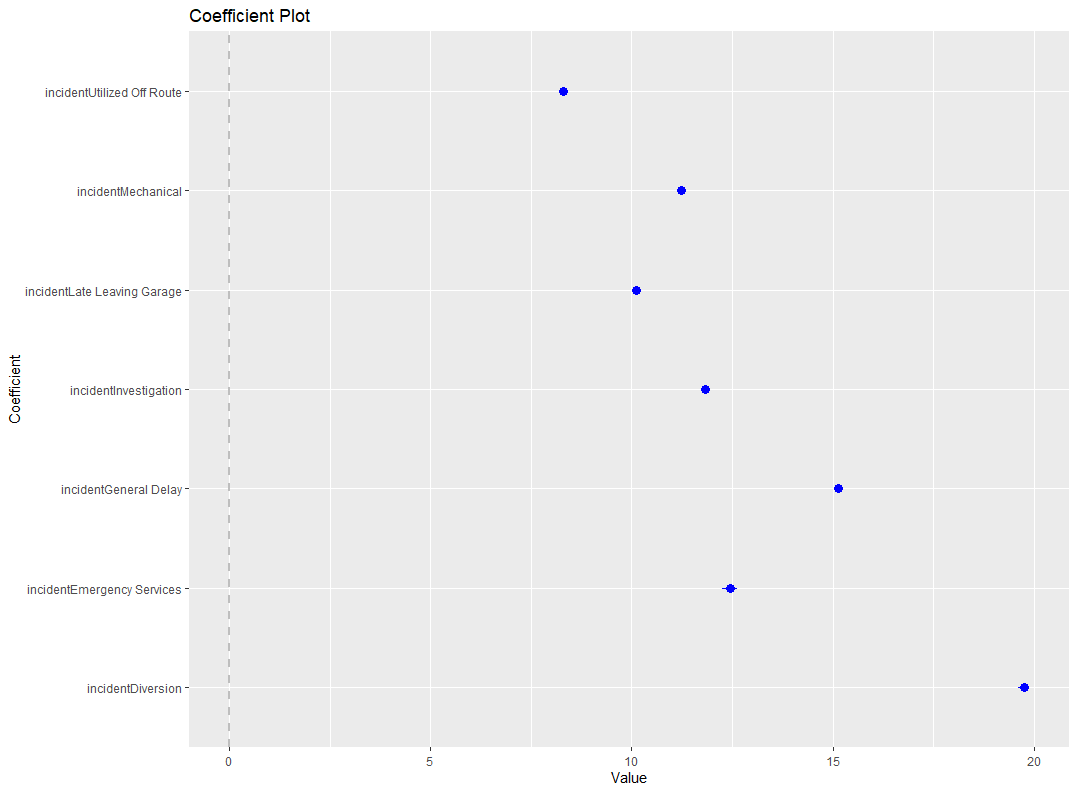
|  |
| --- |
|  |
| Delay histogram of mechanical incidents |



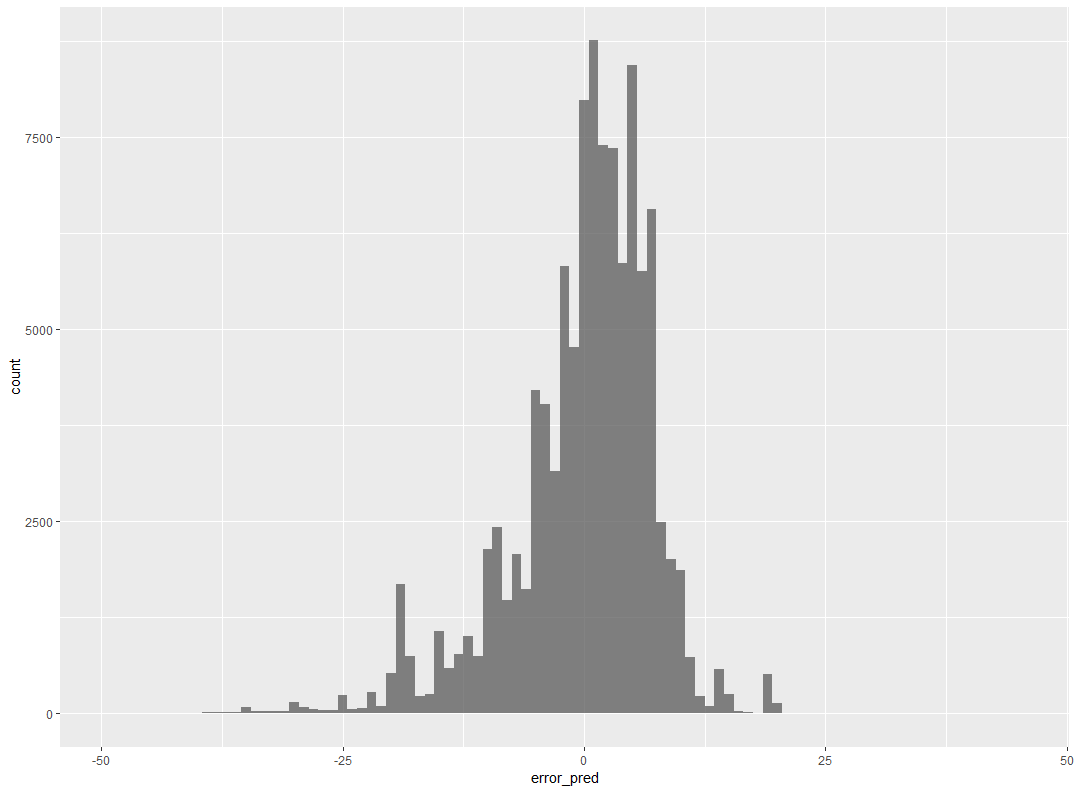
Github link for the code:

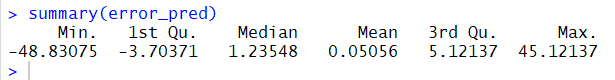
<https://github.com/Amrkenawy/Capstone/blob/master/TTC_Analysis_ver1.1.R>

Using linear Model for Analysis of Variance (ANOVA)



Prediction Error Histogram





# Results

Explain your results here. Consider that you need to communicate your results to executives in an organization. For example:

1. Insert tables and/or charts showing the results
2. Write description of the tables and charts, such that they show the usefulness for an organization
3. Identify the evaluation measures, such as accuracy, precision, recall, etc.

# Conclusions

Give a short summary (one to two paragraphs) of your analysis and conclude the discussion by defining the usefulness of your analysis.