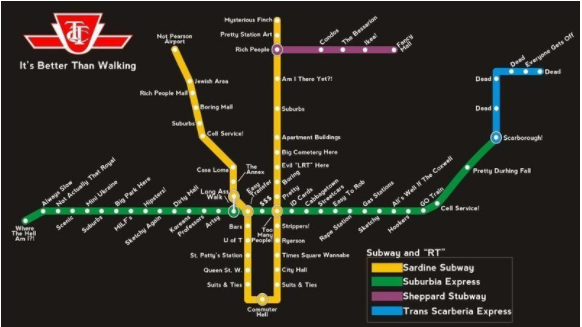
## horizontal line



Report: Predicting Toronto Public Transit Delays

**DS4 - Group 2**

**Group Members:**

1) Daniel Cebula

2) Melissa Hartwick

3) Allan Sales

4) McKinleigh Needham

5) Aravind Kakarala

6) Athithian Selvadurai

[**Introduction**](#_au51mny0sx6) **3**

[**Objectives**](#_3at9u9s4e0vp) **4**

[**Data Preparation**](#_4p7xi5bvhxdr) **4**

[Introduction to Dataset](#_r3ulv59naim9) 4

[Data Cleaning & Preparation](#_5k6z1ori9u1x) 6

[Data Cleaning & Preparation Challenges](#_622gk9amm395) 8

[**Model Design**](#_yyrhu7ml5bea) **9**

[**Model Evaluation**](#_vvd7jrcdppmu) **9**

[**Conclusion**](#_laa0sbfqrf8) **10**

[**References**](#_zhl8rzbn366s) **10**

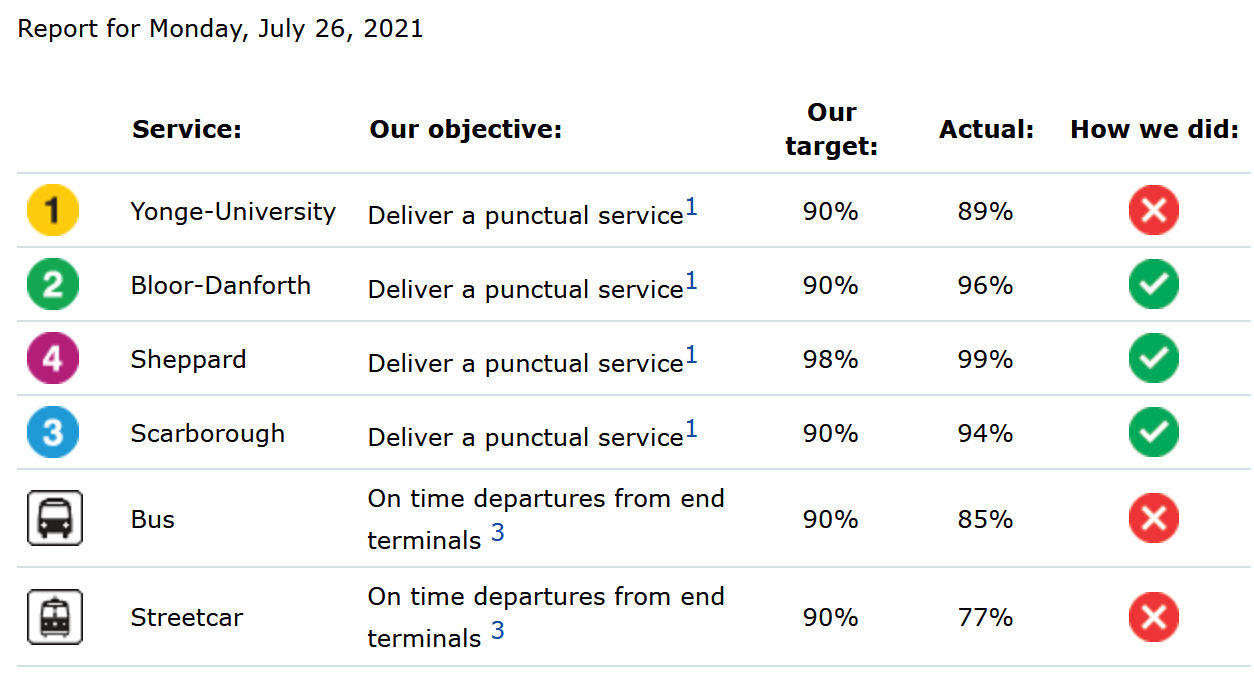
[TTC Daily Customer Service Report:](#_oyq64068rrn9) 10

[Open Data Toronto TTC Datasets:](#_kow7xctf3srk) 10

[Canadian Historical Weather](#_13ygn4i3j19c) 10

# Introduction

During the COVID-19 pandemic, a welcome change felt by non-essential workers around the world was not needing to commute into the office, and during this commute, deal with public transit delays. In Toronto, the Toronto Transit Commission (TTC)’s daily scorecard for service levels for 7/26/2021, showed many service routes missing their goal of 90%+ for on time departures (Toronto Transit Commission, 2021), an indication that Toronto’s transportation system suffers from disruptions and delays.



***Figure 1:*** *TTC’s Daily Scorecard for Service Levels (26th July 2021)*

As Toronto and many cities around the world start to re-open, commuting will soon re-emerge as part of our daily routines, and the “commuting conundrum” of taking public transit vs. driving a car will become a question for some commuters. Do we take public transit and deal with potential overcrowding and delays? Or, do we take a car and deal with potential traffic congestion?

As part of this re-opening, we felt it was timely to review TTC data to determine if we could predict length of transit delays across the three main modes of transportation: subway, bus, and streetcar, in hopes of being more prepared for our organizations’ return to work plans, and provide one more data point to help us make the right decision for our commute.

# Objectives

The objective of our analysis is to predict the length of Toronto Transit Commission (TTC) subway, bus and streetcar delays for the period 2014 - 2019. Our hypothesis is that the publicly available TTC delay data, when combined with weather data, is sufficient enough to predict the length of a delay.

The data will undergo Dimensionality Reduction through Principal Component Analysis (PCA) and other methods and will be fed through a pipeline into a variety of classifier machine learning algorithms.

The best performing machine learning algorithms will form an ensemble to better predict length of subway, bus, and streetcar delays.

# Data Preparation

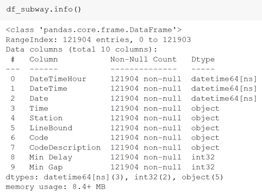
## Introduction to Dataset

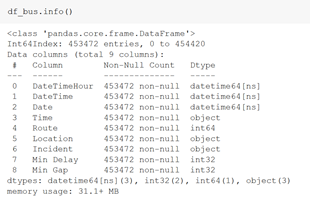
Two datasets were used for our analysis:

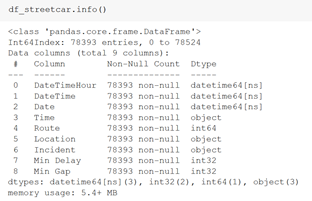
* Toronto Transit Commission (TTC) Delay Data:
  + Subway delay data
  + Bus delay data
  + Streetcar delay data
* Toronto Historical Weather Data

The TTC Delay Data was provided by the TTC and made publicly available by the City of Toronto’s Open Data Portal through an open license. We retrieved the data for subway, bus, and streetcar delays on 7/11/2021.

The data spans from January 1, 2014 - December 31, 2019. We chose to use this timeframe to negate any effects of COVID-19 on the TTC system. We were initially intrigued by this dataset as a candidate for our analysis due to its comprehensiveness in terms of number of observations and variables. After an initial clean-up to remove observations where the target variable (delay) was null, the initial dataset included 121,904 observations for subway, 453,472 observations for bus, 78,393 observations for streetcar, and 13 unique variables.



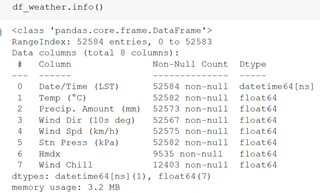




***Figure 2:*** *DataFrame Information for TTC Delay Data by Mode of Transportation*

The Toronto Historical Weather Data was provided by the Government of Canada National Climate Services and made publicly available on the Government of Canada website through an open license. We retrieved the data for three weather stations: ‘Toronto City’, ‘Toronto City Centre’, and ‘Toronto International Airport’ for the same time period as noted above for the TTC Delay Data, and was retrieved on 7/11/2021.

This raw dataset included 54,584 observations and 8 unique variables.



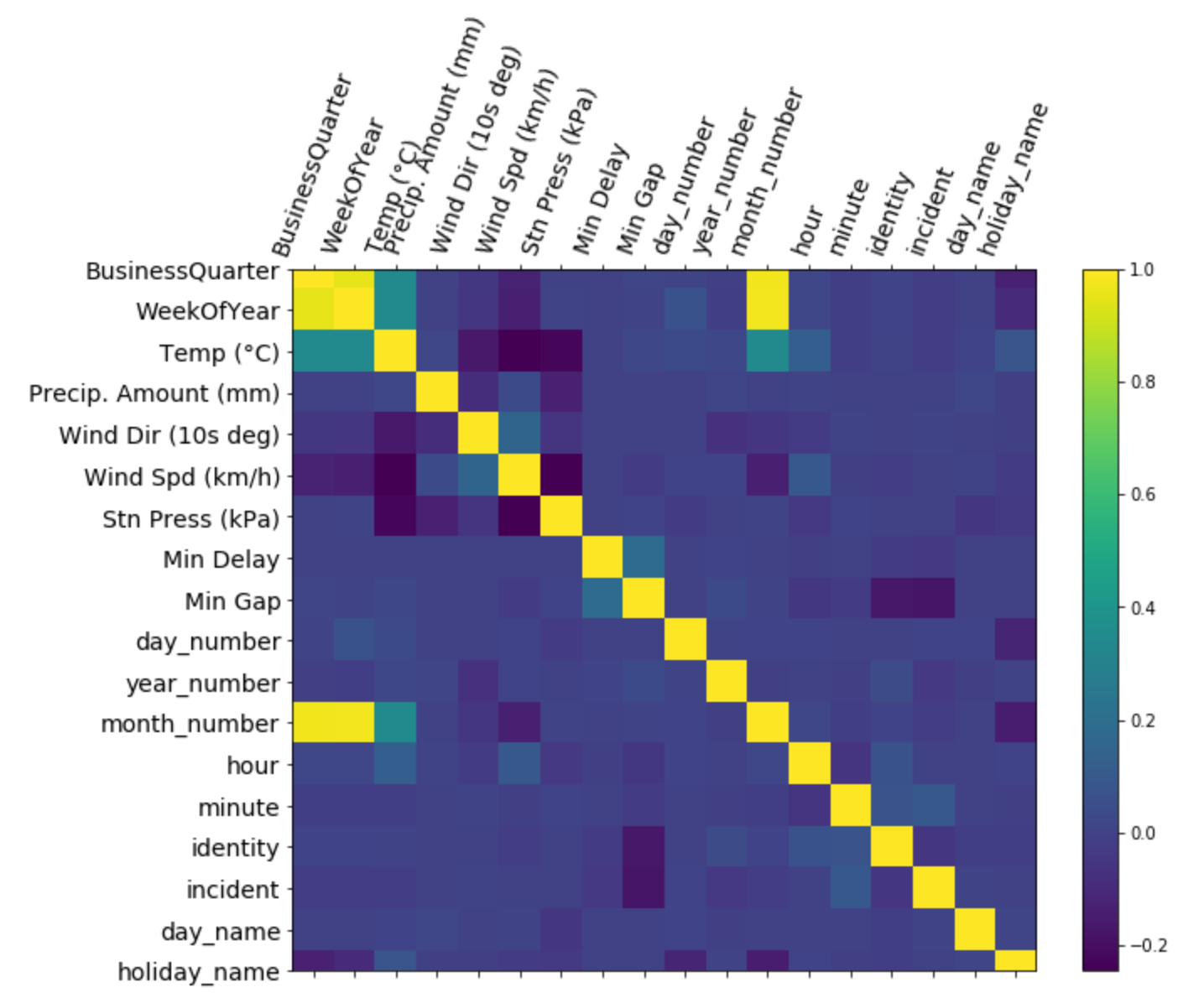
***Figure 3:*** *DataFrame Information for Toronto Historical Weather Data*

## Data Cleaning & Preparation

The data was retrieved through a series of HTTP requests and downloaded as a comma-separated values (csv) file. The data was then imported into a pandas dataframe in Jupyter Notebook for data cleaning and preparation.

Overall the data quality was sufficient in terms of its completeness (minimal null values), accuracy (data appears to be valid and the ranges for numerical values make sense), and timely (historical data available down to the minute). Our process to prepare the dataset for analysis included:

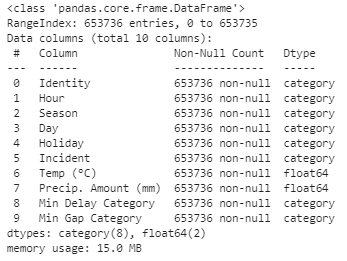
1. Combined each of the three modes of transportation into a single dataset using the concat method. We renamed columns in the subway dataset to match columns names in streetcar and bus, as well as added in an ‘Identity’ column to indicate which mode of transportation a delay was for.
2. The transportation delay data was combined with the historical weather data. This was accomplished through a join on the ‘DateTimeHour’ and ‘Date/Time’ columns.
3. Reduced the copious amount of date and time information to help with model performance and run-time. In its place two columns were created:
   1. ‘Day’ column to combine the date information and to indicate if the observation was a ‘weekday’ or ‘weekend’.
   2. ‘Hour’ column to indicate ‘Morning’, ‘Afternoon/Evening’, or ‘Night’.
   3. ‘Season’ column to indicate the season as ‘Winter’, ‘Spring’, ‘Summer’, ‘Fall’.
4. Added a new column to indicate which days were holidays in case the holiday day impacted delays, either positively or negatively.
5. Within the historical weather dataset, we decided to keep the two variables with the most predictive power: ‘Temp’ (Temperature) and ‘Precip. Amount (mm)’. Both variables had minimal missing information (<1% of observations). Missing values were filled by taking the averages for each column.
6. Removed ‘Location’ data as Latitude / Longitude information was available.
7. ‘Route’ data was also removed as we found it to be very unique to be used for Machine Learning.
8. Looking at the ‘Incident’ data (causes for delay), there were 156 unique categories to start with. The incidents with less frequency of occurrence (<200 occurrences) were grouped into the ‘other’ category. Incident types that were similar were also grouped together (e.g. Assigned observations to ‘Door Problems’ if the word ‘Door’ was in the text string. This resulted in a reduction of Incident categories to 45 types.
9. Based on the correlation matrix (shown below) we saw that ‘Min Gap’ and ‘Min Delay’ (i.e. minutes of a delay) were highly correlated with each other. Both variables are to indicate when a delay has occurred and the length of that delay. ‘Min Gap’ will be used as the target variable and ‘Min Delay’ will be excluded from the train-test sets.

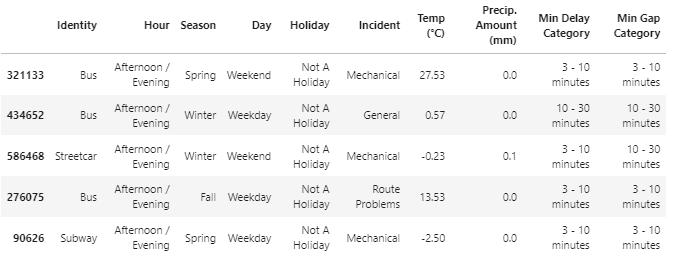


***Figure 4*:** *Correlation matrix for variables in combined dataset*

1. For our target variable ‘Min Gap’ (i.e. gap in minutes between scheduled time and actual time) we knew trying to predict the delay down to the exact minute would be quite challenging. The data closely followed a logarithmic scale and therefore we bucketed the data within this column into 6 buckets: 0 - 3 minutes, 3 - 10 minutes, 10 - 30 minutes, 30 - 60 minutes, 60 - 180 minutes, 180+ minutes.

After data cleaning and preparation, we had one final data set that had 653,736 observations and 8 columns with ‘Min Gap’ to be used as the target variable for predictions.

****

****

***Figure 5*:** *Final dataset to be used for model development and evaluation*

## Data Cleaning & Preparation Challenges

The key challenge we had with our dataset was that even with adding in the weather data and indication for holidays, there was still poor correlation with our target variable ‘Min Gap’. Based on this information, we knew that it would be challenging to develop a model that would produce a high level of accuracy.

# Model Design

We tested the following three models to determine the model which would be most effective at predicting the ‘Min Gap’:

1. Random Forest Classifier
2. Gradient Boosted Classifier
3. Support Vector Classifier

These three models were chosen due to their ease of use and versatility in Classification Machine Learning.

Of our features, we categorized them into three specific groups. Min Gap was set as the dependent variable. We then created a pipeline to transform our variables and combine them together. Our categorical variables, which we used One Hot Encoding to assign values for machine learning, were the following:

1. Identity (Subway, Streetcar or Bus)
2. Hour (Morning, Afternoon / Evening or Night)
3. Season (Winter, Spring, Summer or Fall)
4. Day (Weekday or Weekend)
5. Holiday (Holiday or Not a Holiday)
6. Incident (General, Mechanical, Route Problems or Investigation / Emergency)

The continuous variables were Temp (°C) and Precip. Amount (mm) which were transformed using a MinMax Scaler. Principal Component Analysis was then used to reduce dimensions while preserving 95% of variance. Finally, we combined categorical and continuous variables via a Column Transformer.

The data was then split into a training set and testing split using StratifiedTrainTestSplit ensuring the proportion of Minute Gap Classes in ‘Min Gap’ (our target variable) were distributed equally between the training and testing set.

We then set-up pipelines to fit RandomForestClassifier, GradientBoostedClassifier and LinearSuportVectorClassfier. GridSearchCV was then done on selected hyperparameters to find the highest accuracy. The following hyperparameters were chosen for the models, as these yielded optimal accuracy while ensuring an acceptable run-time:

1. Random Forest Classifier (n\_estimators=10)
2. Gradient Boosted Classifier (n\_estimators=10)
3. Support Vector Classifier (tol=0.1)

Once these models were developed, we also ran an ensemble model in an attempt to reduce the generalization error of the predictions from each individual model. The VotingClassifier ensemble was built using these three classifiers using a hard voting scheme.

# Model Evaluation

Of the models used, their accuracy in predicting ‘Min Gap’ was as follows:

| **Model Name** | **Accuracy Score** |
| --- | --- |
| Random Forest Classifier | 46.6% |
| Gradient Boosted Classifier | 59.7% |
| Support Vector Classifier | 60.2% |
| Ensemble | 59.8% |

As we can see from the above table, we achieved around a 60% accuracy score for our models when looking to predict the length of delay, regardless of when we used a single model or an ensemble. We hypothesized that this level of accuracy might be as high as we would get as predicting length of delays, whether it be subways, streetcars, buses, trains, airplanes, or any other mode of transportation is notoriously difficult due to the sheer number of factors that can contribute to a delay which are hard to track and document (e.g. is the right tool on hand, what is the knowledge level of the mechanic, etc.).

# Conclusion

In conclusion, we were not able to prove our hypothesis that publicly available TTC delay data, when combined with weather data, is sufficient enough to predict the length of a delay, resulting in only a 60% model accuracy, even when using an ensemble.

As mentioned in the model evaluation section, upon further research, we realized that predicting delays is actually quite difficult due to the additional factors that are not typically present in a dataset (e.g. knowledge level of the mechanic).

For future iterations of the model, we would recommend keeping the mode of transportation data separate and looking to create a prediction model of each of the three modes of transportation (i.e. subway, streetcar, and bus). We hypothesize that by combining the modes of transportation, we assume that the nature of these delays are somewhat similar which isn’t necessarily the case.

To further improve the model, we would also recommend looking at housing price data to see if “richer” areas saw shorter delays, as well as looking at population density to determine if “busier” areas along each route see shorter/longer delays.

# 

# References

Toronto Transit Commission. (2021, July 26). *TTC Daily Customer Service Report*. https://www.ttc.ca/Customer\_Service/Daily\_Customer\_Service\_Report/index.jsp

City of Toronto. (2021, July 12). *TTC Subway Delay Data*. Open Data Toronto. <https://open.toronto.ca/dataset/ttc-subway-delay-data/>

City of Toronto. (2021, July 12). *TTC Bus Delay Data*. Open Data Toronto. <https://open.toronto.ca/dataset/ttc-bus-delay-data/>

City of Toronto. (2021, July 12). *TTC Streetcar Delay Data*. Open Data Toronto. <https://open.toronto.ca/dataset/ttc-streetcar-delay-data/>

Government of Canada. (2021, June 1). *Historical Data - Climate - Environment and Climate Change Canada*. Past Weather and Climate. https://climate.weather.gc.ca/historical\_data/search\_historic\_data\_e.html

**Group 2 GitHub Repository:**

Cebula, A. (2021, August 1). *Data Science ML Group 2*. GitHub Repository. https://github.com/cebulada/Data\_Science\_ML-Group\_2