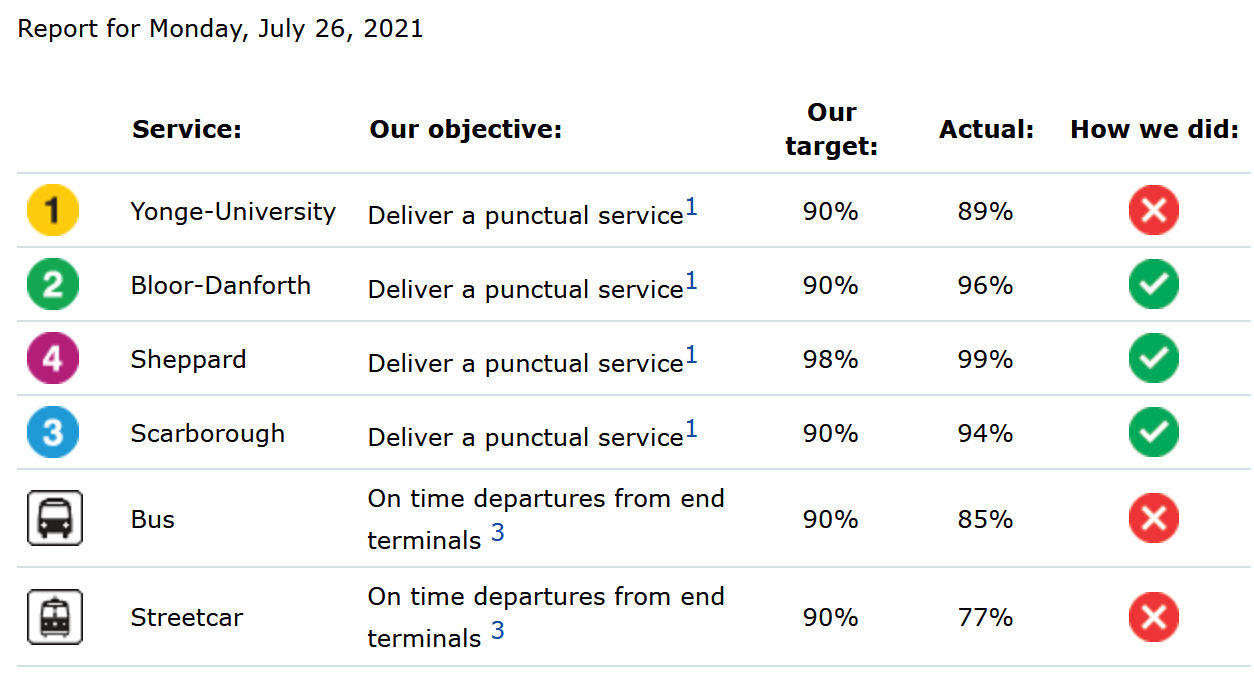
**Predicting Toronto Public Transit Delays**

**Group Number: 2**

**Group Members:** Daniel Cebula,Melissa Hartwick, Aravind Kakarala, McKinleigh Needham, Allan Sales, Athithian Selvadurai

**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 modes of transport missing their goal of 90%+ for on time departures (Toronto Transit Commission, 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 classify and predict transit delays, 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 Toronto Transit Commission (TTC) subway, bus and streetcar delays (from a period of 2014 - 2019) using historical Toronto weather data from the same period.

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 regression machine learning algorithms.

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

**Data Preparation**

Introduction to Dataset

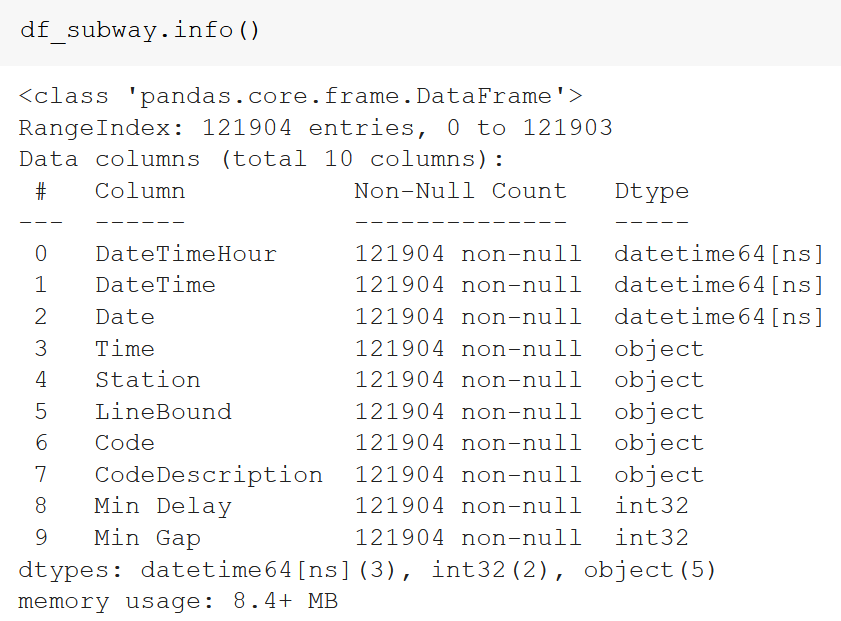
Two datasets were used for our analysis:

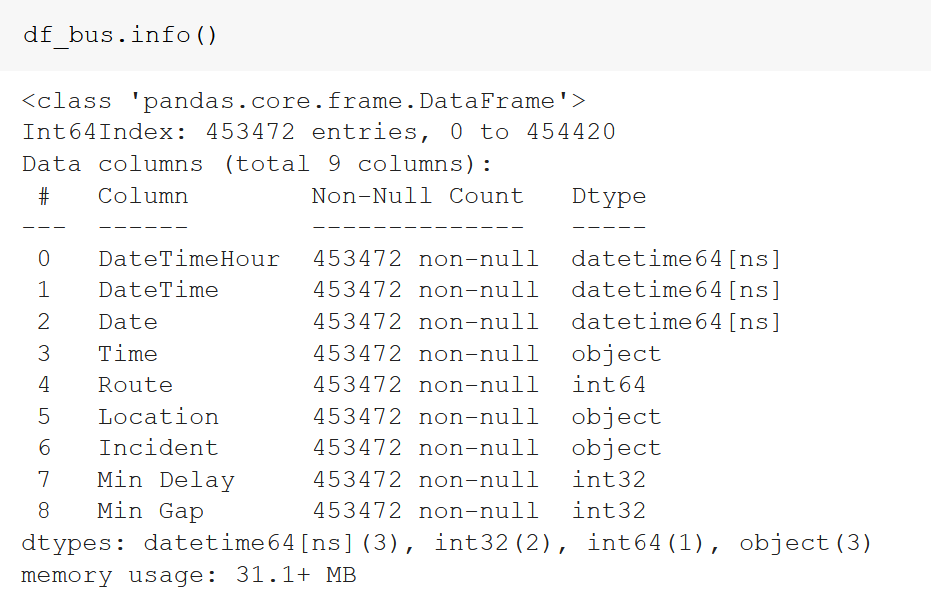
* TTC Delay Data: Subway delay data, Bus delay data, Streetcar delay data
* Toronto Historical Weather Data

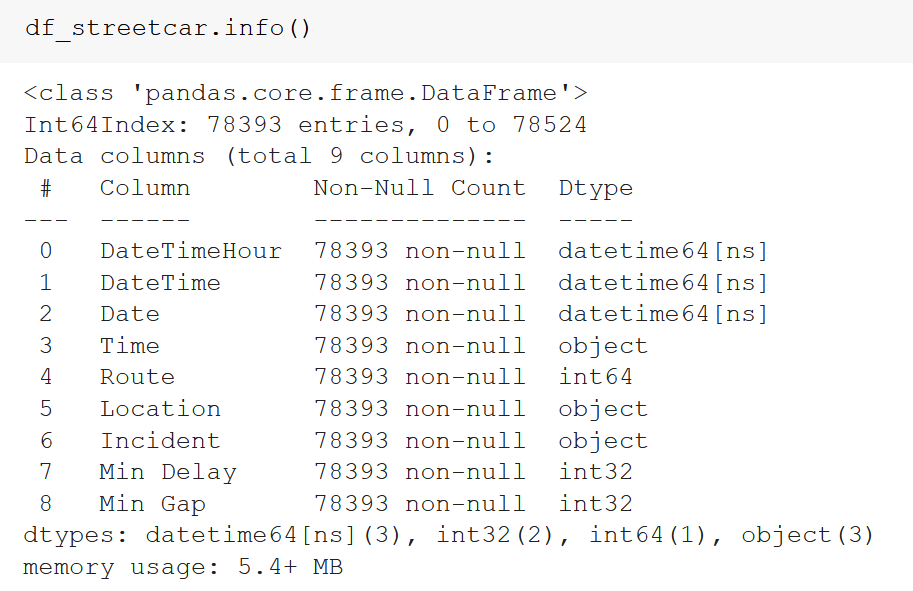
The TTC Delay Data was provided by the Toronto Transit Commission and made publicly available by the City of Toronto’s Open Data Portal through an open license. We retrieved the data for subway, bus, and streetcar delays on 7/11/2021.

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



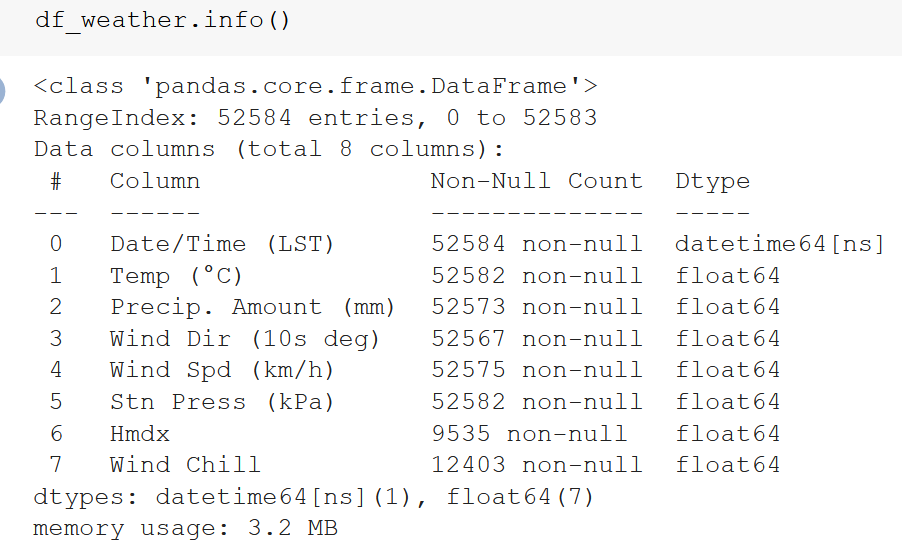




The Toronto Historical Weather Data was provided by the Government of Canada National Climate Services and made publicly available on the Government of Canada website through an open license. We retrieved the data for three weather stations: ‘Toronto City’, ‘Toronto City Centre’, and ‘Toronto International 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 XX: DataFrame Information for Toronto Historical Weather Data**



Data Cleaning & Preparation Process

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 xx,xx,xx. Our process to prepare the dataset for analysis included:

1. We combined each of the three modes of transportation into a single dataset using the concat method. This was an effective approach as each of the three data frames had the same column names.
2. The transportation delay data was combined with the historical weather data. We did this through a join on the ‘DateTimeHour’ and ‘Date/Time’ columns.
3. We added a new column to indicate the day of week for each observation, as well as which days were holidays in case the day of week or holiday day impacted delays, either positively or negatively.
4. Within the historical weather dataset, Temp (Temperature), Precip. Amount (mm), Wind Dir (Wind Direction), Wind Sp (Wind Speed), and Stn Press (Station Pressure), Hmdx (Humidex), and Wind Chill all had varying degrees of null values. We tackled these null values using the following approach: xx
5. TTC delay data has outlier values xx

Data Cleaning & Preparation Challenges

*What challenges did you face?*

[Route, Location and Incident has too many categories]

[Talk about how we wanted to use Toronto Residential House / Condos / Apartments sold house prices (during 2016) to determine if “richer” areas experienced the same frequency and severity of delays and the challenges we ran into, and why we decided not to use it]

After data cleaning and preparation, we had three cleaned datasets, one for each mode of transportation. Each of the three datasets had xx unique columns, with xx being our target variable. The final subway data had xx observations, the bus data had xx observations, and the streetcar data had xx observations.

**Figure xx: Final set of columns within each of the three datasets**

[Insert screenshot]

**Model Design**

*Describe your (two or more) models. How did you choose hyperparameters if they were required? Why did you choose those particular machine learning models?*

[Add in work done on feature generation here]

**Model Evaluation**

*How well did the model perform on the set-aside testing dataset? If you attempted an ensemble model how did it perform vs. the individual models in the ensemble?*

**Conclusion**

*Did you prove/disprove your hypothesis or create a useful model? What did you learn about your dataset? What would you do next to improve your model?*

**References**

[**https://www.ttc.ca/Customer\_Service/Daily\_Customer\_Service\_Report/index.jsp**](https://www.ttc.ca/Customer_Service/Daily_Customer_Service_Report/index.jsp)