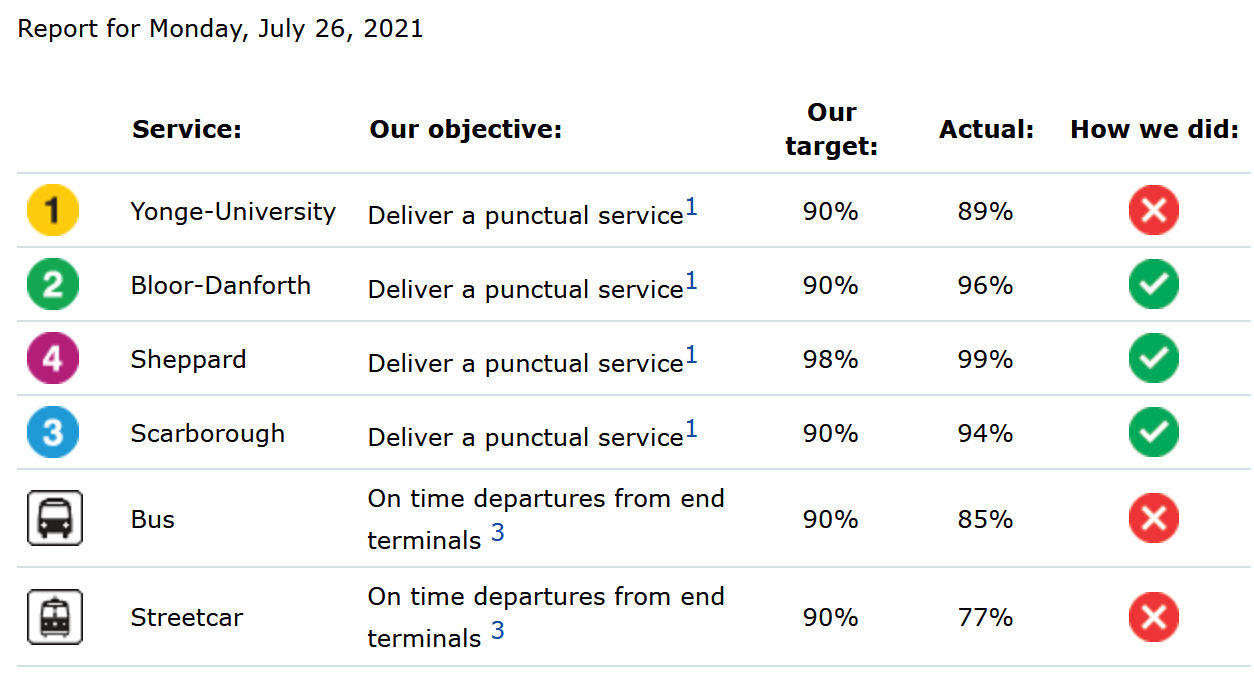
**Predicting Toronto Public Transit Delays**

**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**

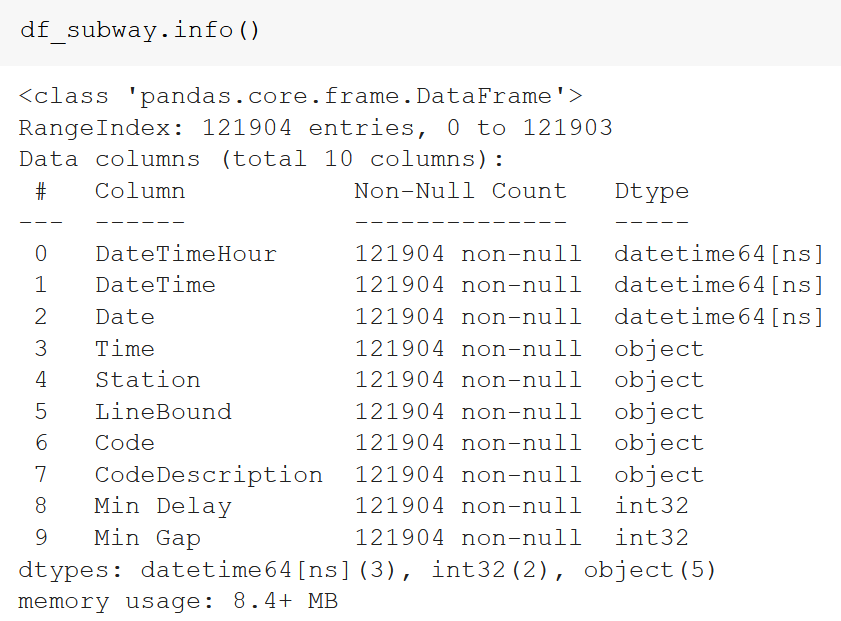
Introduction to Dataset

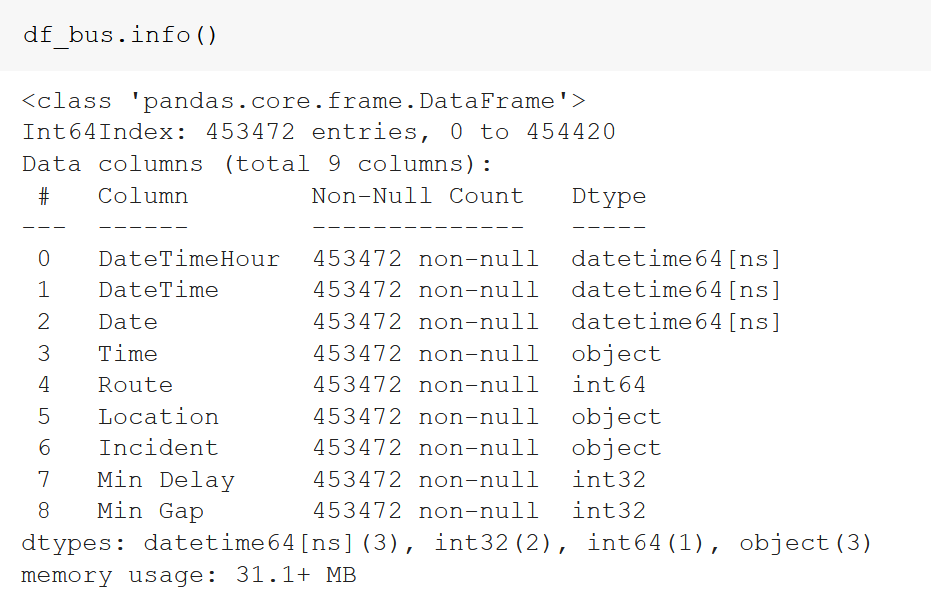
Two datasets were used for our analysis:

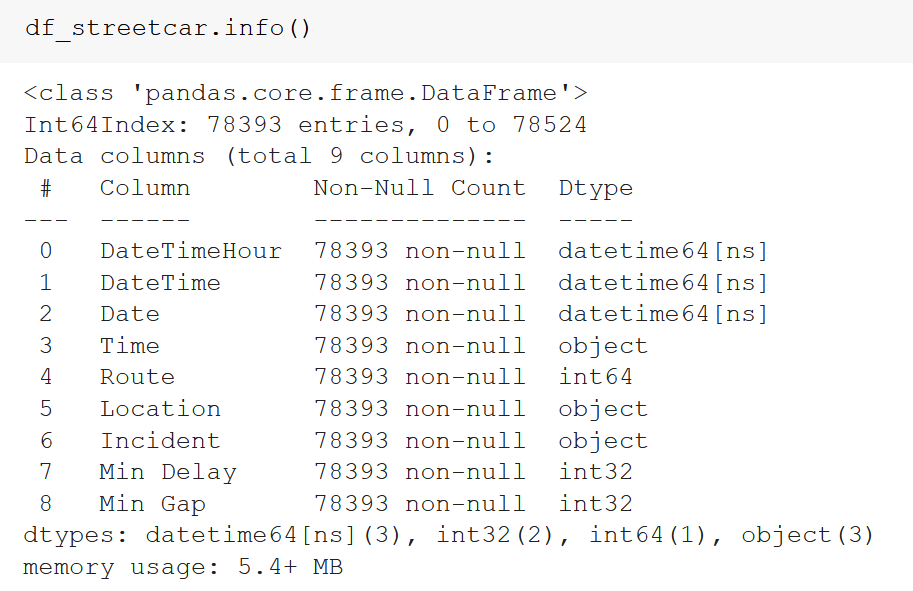
* TTC Delay Data
  + Subway delay data
  + Bus delay data
  + Streetcar delay data
* Toronto Historical Weather Data

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 through a series of HTTP requests 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. In total, the raw dataset included [insert Number of observations and variables].



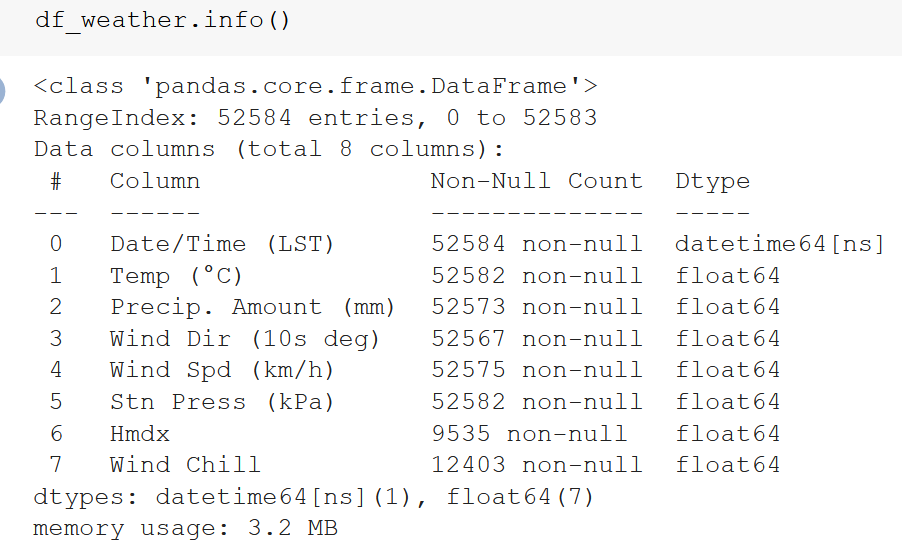




Data dictionary can be found here: [insert link]

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 also through a series of HTTP requests on 7/11/2021.

This raw dataset included [insert Number of observations and variables].



Data Cleaning & Preparation Process

*How good was the data quality? What did you need to do to procure it? What tools or code did you need to use to prepare it for analysis?*

The data was downloaded as a comma-separated values (csv) file, and imported into a pandas dataframe in Jupyter Notebook for data cleaning and preparation.

The variables outlined in the dataset were a mix of numerical and categorical.

[Talk about adding in the holidays, adding in date values, removing nulls,

Data Cleaning & Preparation Challenges

*What challenges did you face?*

[Talk about the codes in the delay data]

[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]

**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?*

**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)