SDSS Datathon

EDA REPORT

Student Name
Adelyn Lee
Sara Sanas
Narae Lee
Chay Park

The team’s goal was to create a transit app for TTC commuters to forecast transit time on the TTC based on the most likely delay incidents to occur at a user’s given time. Bus, streetcar, and subway data that was provided all contain the same variables with subway incidents classified as standardized delay codes. All datasets contained some null values for ‘Bound’/’Direction’ columns. It was assumed that these vehicles were not in service and was removed from the datasets. Although some of these entries contained a non-zero minimum delay time, it would not have an impact on the consumer as they are only dependent on running vehicles that TTC users can board. For each dataset, the team also filtered out a set list of delay codes and incidents that were deemed to be random, as they are not an effect of any of the possible features so cannot be a predicted event (eg. SUDP - Disorderly Patron, Emergency Services, Collision - TTC Involved). A weather dataset for 2024 was combined to allow for exploratory data analysis on environmental features, including ‘Mean Temp (°C)’, ‘Total Rain (mm)’, ‘Total Snow (cm)’. To conclude preprocessing, all datasets were combined and another column was created to differentiate between bus, streetcar, and subway delays. All preprocessing was using the Polars library, a Pandas alternative with faster runtimes and freedom with parallel operations.

After the data was cleaned and standardized, EDA was conducted to gain visibility on the correlation between features, possible outliers, and allow the team to make a decision on what predictive models to construct.

A bivariate graph was created to measure the correlation between possible features.

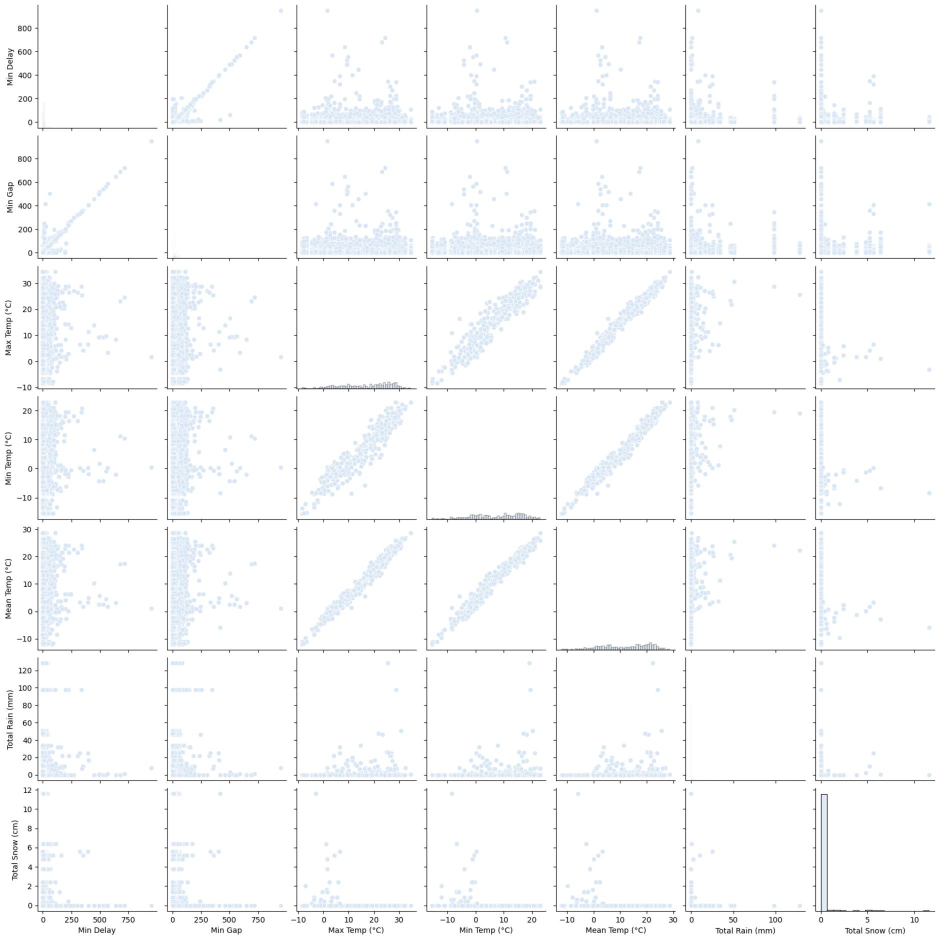
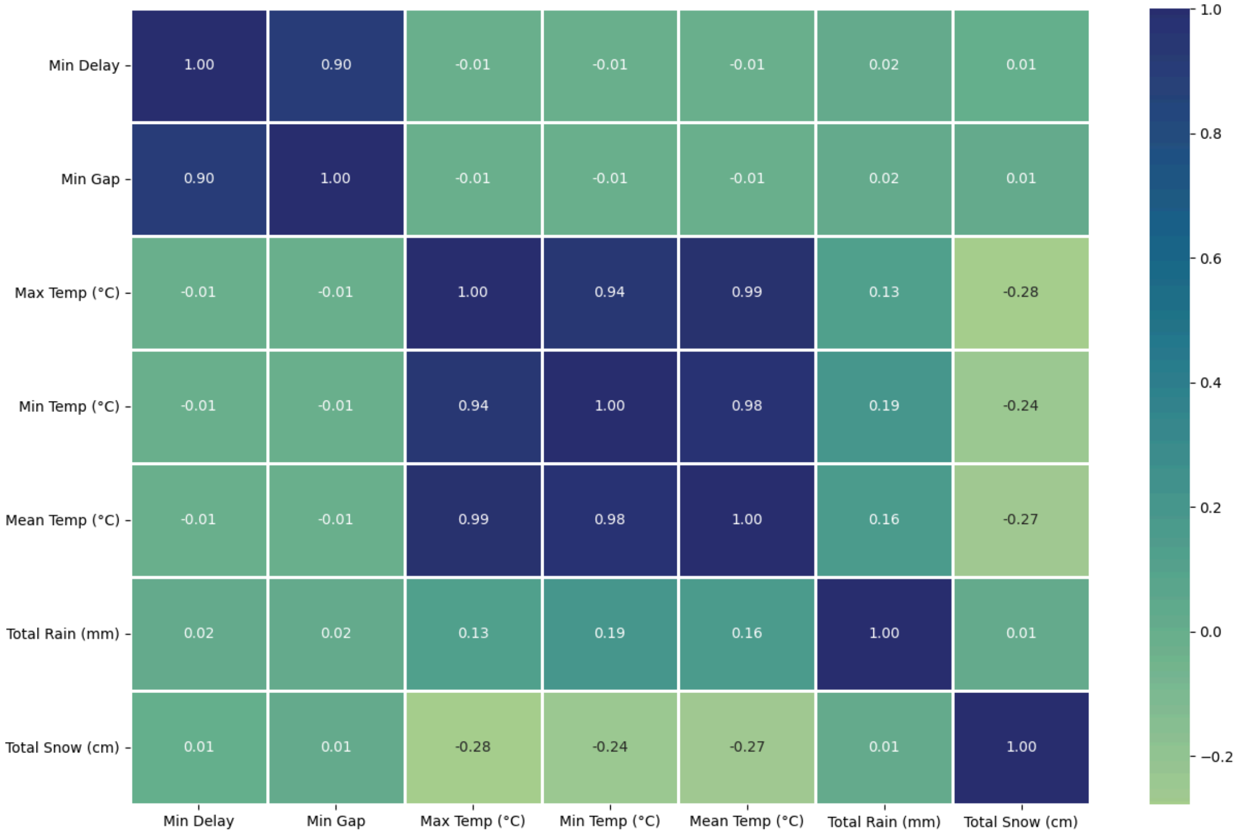


Fig 1. Bivariate graphs of possible features

It was found that there was a very strong correlation between ‘Min Temp’, ‘Mean Temp’, and ‘Max Temp, as well as ‘Min Delay’ and ‘Min Gap’. Therefore, ‘Min Temp’ and ‘Max Temp’ were dropped to avoid redundancies in the features, and a new column was created for the ratio between ‘Min Delay’ and ‘Min Gap’. ‘Min Delay’ is the amount of delay of an individual vehicle from its intended arrival, and ‘Min Gap’ is the amount of time between a vehicle’s arrival and the next vehicle’s arrival at a chosen station/stop on a given route/line. It was surprising to find these features strongly correlated as it seems to measure different components of a transit process, so the team decided to keep both features for the model by creating a new column as the ratio of ‘Min Delay’ to ‘Min Gap’.

A heat map was also used to observe the correlation strength between features.



	Min Delay	Min Gap	Max Temp (°C)	Min Temp (°C)	Mean Temp (°C)	Total Rain (mm)	Total Snow (cm)
Min Delay	1.00	0.90	-0.01	-0.01	-0.01	0.02	0.01
Min Gap	0.90	1.00	-0.01	-0.01	-0.01	0.02	0.01
Max Temp (°C)	-0.01	-0.01	1.00	0.94	0.99	0.13	-0.28
Min Temp (°C)	-0.01	-0.01	0.94	1.00	0.98	0.19	-0.24
Mean Temp (°C)	-0.01	-0.01	0.99	0.98	1.00	0.16	-0.27
Total Rain (mm)	0.02	0.02	0.13	0.19	0.16	1.00	0.01
Total Snow (cm)	0.01	0.01	-0.28	-0.24	-0.27	0.01	1.00

Fig 2. Heat map of possible features

It was found that weather features have a weaker correlation to ‘Min Gap’ and ‘Min Delay’, and it was surprising that ‘Total Rain (mm)’ had a positive correlation with ‘Mean Temp’ whereas ‘Total Snow (cm)’ had a negative correlation. Therefore, it was decided that it was necessary to keep these features in the model.

Finally, a box plot was made between ‘Day’ and ‘Min Delay’ to compare an example of categorical features to numerical features.

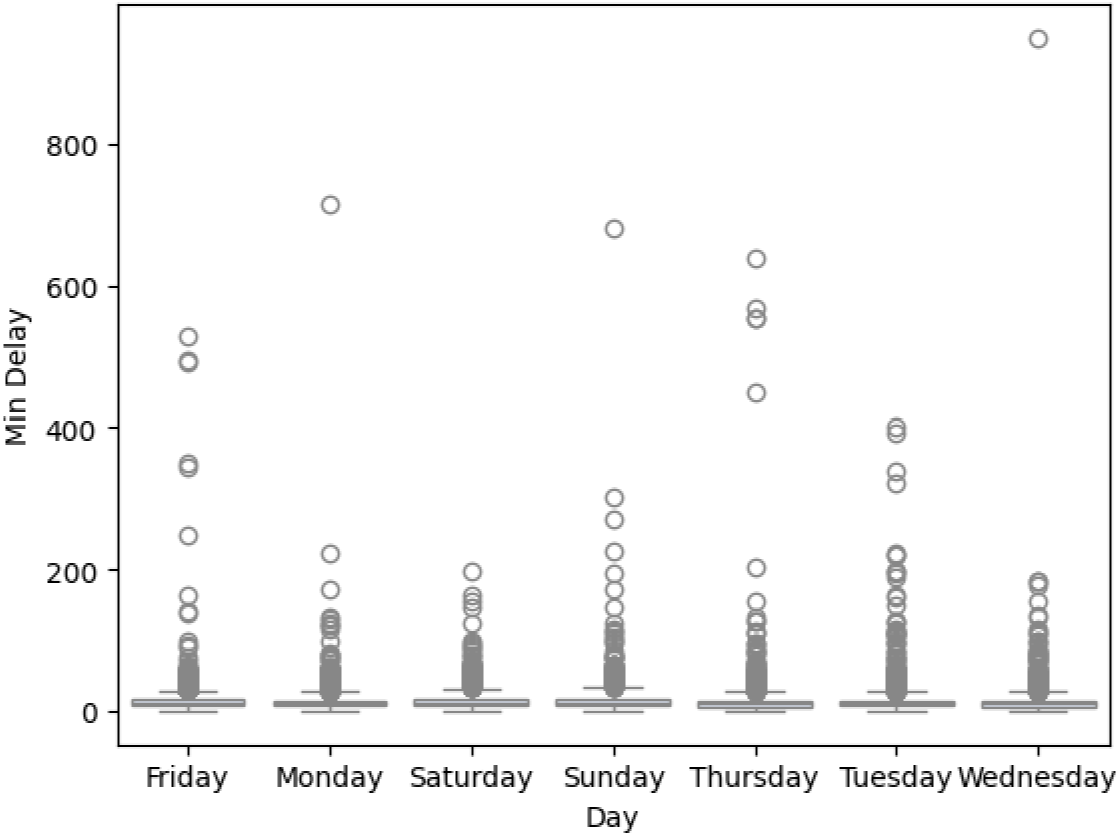


Fig 3. Box plot of ‘Day’ and ‘Min Delay’

It can be seen that most ‘Min Delay’ values are under 200 minutes, and have outliers reaching upward of 900 minutes. The team decided not to exclude these outliers as they are already determined to be predictable issues (random incidents were filtered out in preprocessing), and understanding the amount of noise in the data allowed us to choose what predictive models to build.

The team decided to construct a Multinomial Linear Regression model as it is best suited for classification methods that require predicting multiple possible outputs, as well as a Random Forest model because it is resistant to outliers, apparent in Fig 3.