15.0 Capstone 2 Milestone Report

Problem Statement

As the city of Toronto is growing, so is the population which will affect how people get around the city. One of the best ways to get around the city is the public transportation system, especially the subways. Unfortunately, subway delays are common and can ruin anyone's day. The goal of this project would be to forecast the pattern of delays over the last 5 years and hopefully come up with some insight.

The client in this scenario would be the local government and Toronto Transit Commission. It would be in their best interest to identify when to address inefficiencies in the public transportation system so that there are less people driving their own cars. This could lead to overcrowded highways and congestion on roads which could also lead to accidents.

Dataset

For this analysis, I will be obtaining data from Toronto Open Data (data collected by the city of Toronto) which will be in the form of multiple .csv files spanning over 5 years. I will use time series analysis techniques to forecast trends in the data for delays so hopefully, there can be actions taken to mitigate it.

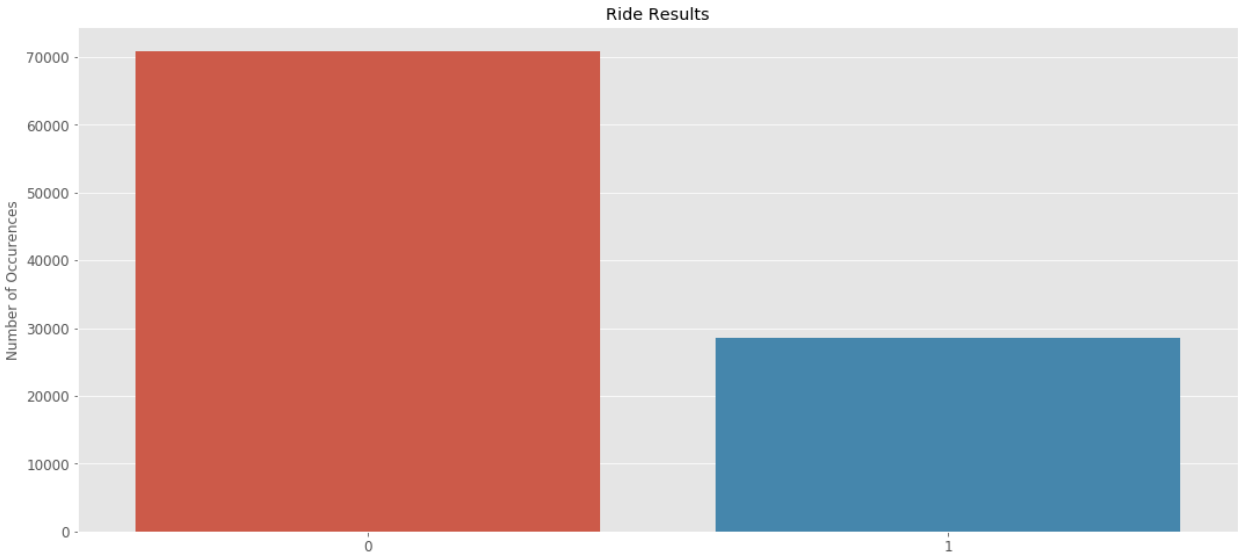
During the process of reading in the data, it was found that there was very little missing data but there were inconsistencies with several of the values of features. In the case of missing data, it was viable to drop the rows due to missing data in so many other features. The problem of inconsistent naming convention was for categorical features that introduced many values which would result in too many features in the encoding process. This was solved by combining values that are essentially identical and dropping values that don't make any sense.

For the time series analysis, the only features that were important are the timestamp and the engineered feature of whether a delay occurred or not.

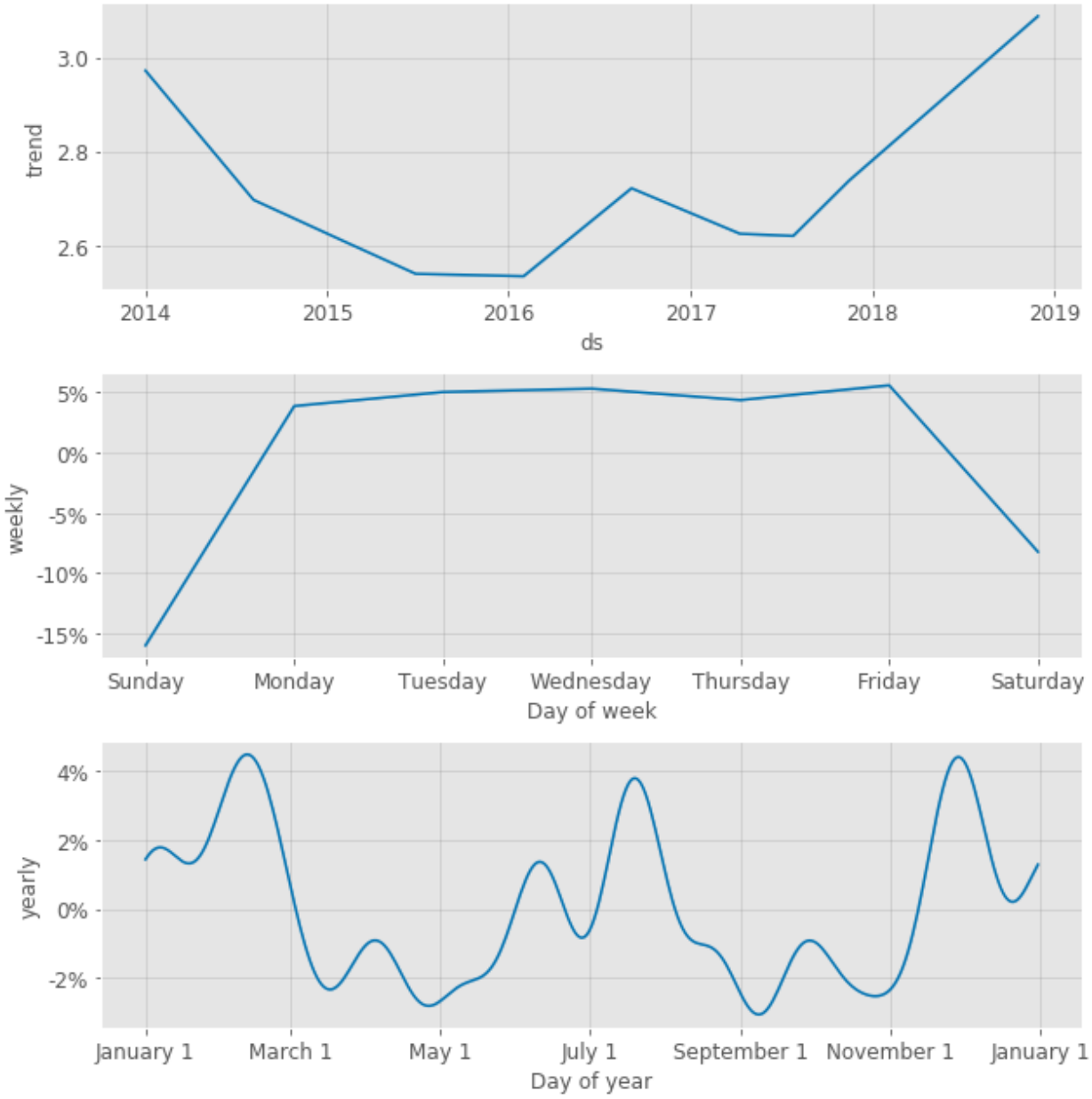
Initial findings from exploratory analysis

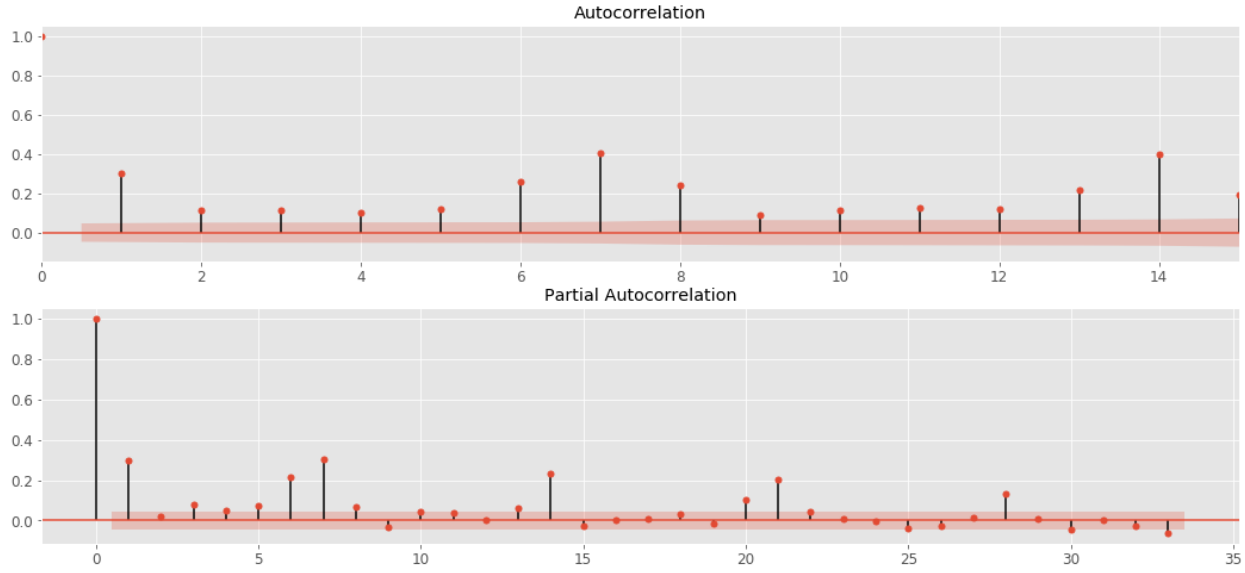
Five years of data is a lot to plot out, especially when observations are minutes apart. What I decided to do was aggregate the data on a daily basis and summed up the amount of delays that occurred on any particular day. After plotting the new data set out it looks like there was an event that happened which caused the number of delays to increase ever since 2014.

When the ratio of delays and non-delays are observed below, we can see that there are a significant amount of delays. The good news is that the data is relatively balanced so we can keep it as is.



The data was then broken down to its components of yearly and weekly seasonality as well as any trends that exist as seen in the figure below. The weekly seasonality makes a lot of sense since it peaks on weekdays and dips on weekends which makes sense according to the average work schedule. There also seems to be a pattern in the yearly seasonality of peaking during the winter and summer and dipping during the spring and fall seasons.


trend
3.0
2.8
2.6
2014
2015
2016
2017
2018
2019
ds
weekly
5%
0%
-5%
-10%
-15%
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Day of week
yearly
4%
2%
0%
-2%
January 1
March 1
May 1
July 1
September 1
November 1
January 1
Day of year



Autocorrelation exists for 1 and intervals of 7 which is also the case for the partial autocorrelation. This matches our understanding of the work week.