**Order of Notebook to be Ran:**

1. **Exploratory Data Analysis (EDA)**
2. **Preprocessing**

**¾) Regression/Time Series Model**

**Problem Statement:**

The dataset analyzed involves information on turnstile entry data of stations within New York City. The models built out of the dataset will analyze the biggest weights or driving factors of New York City Subway Ridership. This will be geared towards the operations planning group of the MTA which is responsible for making service decisions.

This is an important dataset because the models that are generated from this dataset can determine the direction of the ridership of the subway line and whether there should be more service increased on the line or more stairs constructed into the station to allow more space. It can also provide cost savings for the agency because some stations may have declining ridership.

For example, if season has the biggest weight on the model, service may be adjusted by time to allow most efficiency of service.

**Exploratory Data Analysis for the Ridership Dataset:**

We mainly conduct EDA on our dataset using an EDA function which inputs our data and conducts several EDA commands and outputs a new dataset.

Merging New Data:

Another potentially important piece of information that may help the analysis of ridership is whether the station is accessible or not. The subway system was built in 1904 which was before the passing of the Americans with Disabilities Act of 1990. Only a quarter of the 472 stations are ADA accessible with elevators. Elevators may have an influence on whether someone like an elderly or disabled person would take the subway over other transportation options like the bus. The accessible column in our dataset indicates whether a station is accessible (1) or not (0). Another potentially important piece of data that could impact the analysis is the population of the boroughs. However, the population of the boroughs could be represented by the borough column itself. Therefore, it was not necessary to implement that data.

For the time series data analysis, we combine the ridership numbers for the years of 2018 and 2019 to analyze if there is a relationship between seasonality and ridership.

NULL Values:

The dataset initially had null values. Most the missing data values were not missing at random. The systems including the PATH, Roosevelt Tram, and JFK AirTrain all have turnstiles collecting number of entries into the system. These null values were dropped due to the systems being outside of our jurisdiction focus which is the New York City Subway. The remaining null values involve 46 rows where the turnstile entry data was missing completely at random. These rows have also been dropped. It is .02% of our entire dataset. The possible reasons to why the data is missing could be due to closure of the station for the day due to an event, construction, crime scene, etc. In addition, the dataset only features boroughs of NYC subway jurisdiction which does not include Staten Island.

One Hot Encoding:

The boroughs of New York City had to be one hot encoded of {'Q': 1, 'Bk': 2, 'M': 3, 'Bx': 4, 'SI': 5}). All lines had to use pd\_get\_dummies to one hot encode since there are too many different lines. The dates were converted to one hot encoded season of (1)- Winter, (2)- Spring, (3)- Summer, and (4)- Fall for the purposes of a facilitating the development of regression models.

**Visualizations: Some features are represented through graphs.**

* Heatmap

Based off the correlation heatmap, we notice that entries have the highest correlation with variables including complex\_id, accessibility, borough, the Flushing Line and Lexington Ave lines. In terms of subway lines this could mean that the Flushing Line and Lexington Ave lines have the highest ridership in the entire system. However, we notice that all other lines are very correlated with each other. Therefore, models including RandomForests and GradientBoosting models tend to work better with datasets like this.

* Histograms

Based off the histogram representation of the ridership, we notice that entry data is skewed. We should take the log transform to reduce the skewedness or by using Synthetic Minority Oversampling Technique.

* Pie Graphs

When using a pie graph to represent the relationship between ridership and season, we see that the Spring and Summer both have the highest ridership. There may be a seasonal relationship as ridership increases along with warmer weather.

When using a pie graph to represent the relationship between ridership and accessibility, we notice that accessible stations overall have less than half ridership numbers than non-accessible stations. This is not surprising as only a quarter of the subway stations are accessible.

* Scatterplots

When using a scatter plot to represent the relationship between the station (complex\_id), we see that the ridership tends to increase as the complex\_id gets bigger.

* Time Series Visualization

When viewing the trends of the time series plot of ridership, we see that ridership dips when the new year comes around. This indicates a potential season relationship in ridership.

**Preprocessing and Model Selection**

Predictions:

* We want to use entry data as our predictor variable and not exit data. Hundreds of riders prefer to use the gate to exit a crowded station for speedier commutes. Using exit data would lower the predictive power of the model.
* Powertransforming and train-test splitting

Our data is skewed, therefore we want to perform a method to our datasets to reduce this. After powertransforming our regression dataset, we train test split our data to see how well our data generalizes to future data.

**Modeling and Results:**

There were two types of models generated to represent this data including Regression and Time Series.

Regression Modeling:

The predictor variables include the accessibility, season, borough, complex\_id, and line. Regression models that were generated include linear regression, Random Forests, and Gradient Boosting Models. These models are the most widely used for regression models and tend to result in better scores. The metrics that are primarily involved to evaluate regression modeling are R squared scores and root mean squared errors.

* The Linear Regression Model: Produces the R squared scores of around ..47 for both the training and testing data. The root mean squared error is around .73 for both the training and testing data.
* RandomForests: Produces the best R squared scores of around .76 for both the training and testing data. The root mean squared error are the lowest being around .48 and .49 for both the training and testing data respectively.
* Gradient Boosting: Produces an R squared score of around .60 for both the testing and training data. The root mean squared error for both the testing and training data being .63
* AdaBoost: The Adaboost results produce the lowest scores of R squared being around .32 for both the training and testing sets and .83 RMSE for both the training and testing sets.

GridSearch was near to impossible to interpret due to the code taking multiple hours to run without producing metrics.

Time Series Modeling:

The sole purpose of the time series model is to determine whether there is a relationship between seasonality and number of turnstile entries. The possible reasoning is due to subway ridership potentially having a spike in the summer due to the increased number of tourists or people preferring to drive during the winter. In addition, the dataset involves dates which often tends to lead to time series models.

Linear Time Series Model: The model proved to be unreliable in predictive power. The testing and training R squared scores hovered around .18 and the RMSE scores were significantly high in the millions.

**Conclusion and Recommendations:**

The regression model proved to be more reliable than the time series model. One possible factor contributing to this result could be that the data for the time series model spanned only for two years. A longer time span could potentially increase the predictive power of the model.