MTA, at Your Service

By: Jiale (Jerry) Chen, Nossaiba Kheiri, Yihan (Shane) Luo, Sifan (Emily) Tao, and Yan (Felicity) Zhu

1. Introduction

The Metropolitan Transportation Authority (MTA) is the largest transportation network in North America, "serving a population of 15.3 million people across a 5,000-square-mile travel area surrounding New York City, Long Island, southeastern New York State, and Connecticut" (Metropolitan Transportation Authority [MTA], 2023). In New York City, the subway is a lifeline for millions, but its efficiency is hampered by peak-hour congestion. There are many reviews on MTA from Yelp and a trove of hourly ridership data from DATA.NY.GOV; we want to use the data to identify problems with MTA and predict hourly ridership at each station so that we can help MTA better serve the public. After building models that perform well in predicting hourly ridership, we propose dynamic pricing, for which we encourage people to travel off-peak hours by increasing fares during rush hours. Based on the simulation results, it will smooth the traffic and resolve the problem of overcrowdedness during peak hours.

2. Exploratory Data Analysis

First, we visualized MTA Customer Feedback Data and used natural language processing (NLP) to understand people's complaints about MTA. As we can see from Figure 1, MTA received far more complaints than commendations, which makes sense because the dataset is generated from the Customer Relationship Management System and the public is more likely to correspond to MTA about complaints. To understand people's complaints, we used Selenium and Beautiful Soup to extract 123 reviews on MTA from Yelp. People's reactions to MTA are really polarized; there are an equal amount of positive and negative reviews (see Figure 2). As for areas of complaint, we can see that MTA is complained for overcrowded trains, delays, increases in fares, and homeless people on trains (see word clouds in Figure 3).

Second, we performed some financial analysis on MTA. As we can see from Figure 4, MTA has been experiencing a great loss from its operations, and the pandemic aggravated this trend. MTA has been experiencing losses at the gross profit level since 2003. The net incomes are primarily positive, which is attributed to non-operating income. This item consists of multiple sorts of subsidies, including subsidies offered by New York State and Connecticut, tax-related subsidies, Mass Transportation Trust Fund Subsidies, and others. The gap between revenue and gross profits is continuously growing in the last 3 years due to the pandemic. Revenue in 2020 halved that of 2019, and revenue in 2022 was yet to reach the previous level. And the subsidies played a greater role in the last 3 years.

Third, we tried to understand factors that might affect hourly ridership, including time, space, and weather.

- Timewise (see Figure 5): (1) Hour: there are fewer data points from 1 am to 3 am, which makes sense because not many people are traveling during those hours. Additionally, we can see rush hours in NYC are from 7 am to 8 am and from 3 pm to 6 pm. (2) Weekdays: there are fewer data points on Saturday and Sunday, and total hourly ridership on Saturday and Sunday is much less compared with weekdays. (3) There are some seasonal fluctuations in hourly ridership.
- Spacewise (see Figure 6): (1) Brooklyn has more data compared with other boroughs; Brooklyn has the most stations (157), followed by Manhattan (120), Queens (78), Bronx (68), and Staten Island (2). (2) Manhattan has the most traffic. (3) The 7 busiest subway stations are Times Sq-42 St, Grand Central-42 St, 34 St-Herald Sq, 14 St-Union Sq, Fulton St, 34 St-Penn Station, and 59 St-Columbus Circle.
- Timespatial analysis: Manhattan has the most traffic. However, people from Brooklyn and Queens are commuting to Manhattan in the morning rush hours, resulting in higher traffic in some stations.
- Weather (see Figure 7 and Figure 8): (1) there is a distribution for temperature and precipitation; temperature and precipitation have some variations. Temperature and precipitation will be incorporated into our machine learning model. (2) Extreme snowfall and snow depth conditions are rare from 2022 on. (3) While temperature has seasonability, ridership continues to increase and increased ridership is due to returning to normal.

Fourth, we observe that ridership patterns are different from station to station (see Figure 9 and Figure 10). Taking 116 St - Columbia University and Flushing - Main St as an example, Main St Station has the most traffic in the morning while Columbia University has the most traffic in the afternoon. In addition, seasonability for ridership is different.

3. Machine Learning

Hour, Weekdays, and Month are categorical variables, which require dummy coding and dropping first. After dummy coding and combining with weather data, we have 43 columns. Our hourly subway ridership data has 6086184 rows in total. There are 425 stations; on average, each station has 14320 data points. In consideration of the data complexity and variation in stations, it makes more sense to predict hourly ridership for each station.

For each station, we split data with 70% training and 30% testing and fitted linear regression, random forest regressor, gradient boosting regressor, and quadratic regressor to predict hourly ridership for each station. Averaging r^2score on testing data for all stations, we can see that

- Linear regression (even with regularization terms) has the worst performance in predicting hourly ridership, with r^2 of 0.64 on testing data. Linear regression is a parametric model and it assumes a linear relationship between hourly ridership and factors such as hour, weekdays, month, temperature, and precipitation. Our data might not fit these assumptions; as a result, r^2on our testing data is not great.

- However, when we used a quadratic regressor, the performance is much better, with r^2 of 0.77 on testing data.
- Nonparametric models, such as random forest regressor and gradient boosting regression, also have better performance than linear regression, with an r^2 of 0.7 on testing data.
- Support vector regression does not work because our dataset is too big even at the station level.
- Regression based on k-nearest neighbors has really bad performance on the testing data compared with other models and we decided not to use it.

To understand the performance difference in training and testing data, let us take a look at 116 St-Columbia University station. As we can see from Figure 11, random forest regression and gradient boosting regression give the best performance on training data. However, random forest regression and gradient training do not give the best performance on testing data due to overfitting. We have tried to tune parameters for gradient boosting regression and gradient boosting regression, to no avail.

4. Applications

We want to dynamically price fares based on our model prediction, real-time traffic, occurrence of events, etc. First, we use the sigmoid function to constrain our predicted hourly ridership to a range between 0 and 1. Second, we come up with a real-time traffic factor based on data from Waze API and an event factor based on data from Eventbrite API. Third, we come up with an appropriate base fare and multiply it by the real-time traffic factor, the event factor, and the predicted hourly ridership after the sigmoid function, which results in a fare. Lastly, we do not want stations close to each other to have very different fares, which can give rise to some arbitrage opportunities. To avoid exploitation, we also use clustering to adjust the fare.

Computing fares for all predicted hourly ridership, we come up with a fare distribution (see Figure 12). To evaluate the impact of dynamic pricing, we simulated how riders will respond to price changes. As we can see from Figure 13, dynamic pricing can encourage people to travel off-peak hours, which resolves the problem of congestion. As for the revenue, there is no significant difference before and after dynamic pricing. Nevertheless, we can play with the parameters aforementioned in the fare calculation procedure to help MTA make a profit.

5. Limitations and Future Research

First, we did not take heterogeneity into account and assumed that all riders would respond to fare changes in the same way. Going forward, we want to investigate how different riders would respond to fare changes by using Large language models (LLM) so that our simulation results can be more representative.

Second, we can incorporate more features, such as holidays, and explore more complex models. We incorporated ridership in the previous hour and implemented the Long Short-Term Memory(LSTM), which is a recurrent neural network architecture widely used in Deep Learning, but it did not give us better predictions. Thus, we want to use SHAP(Shapley Additive exPlanations) to understand areas of improvement for our models.

Third, we built a website using Streamlit and used it for the presentation. However, due to the capacity limitation, the website is down now and can only be accessed locally by typing ‘streamlit run Home.py’ on the terminal. We hope to deploy our project using React over the winter break.

Reference

Metropolitan Transportation Authority. (2023). https://new.mta.info/about

Figure 1. MTA Complaint and Commendation Count

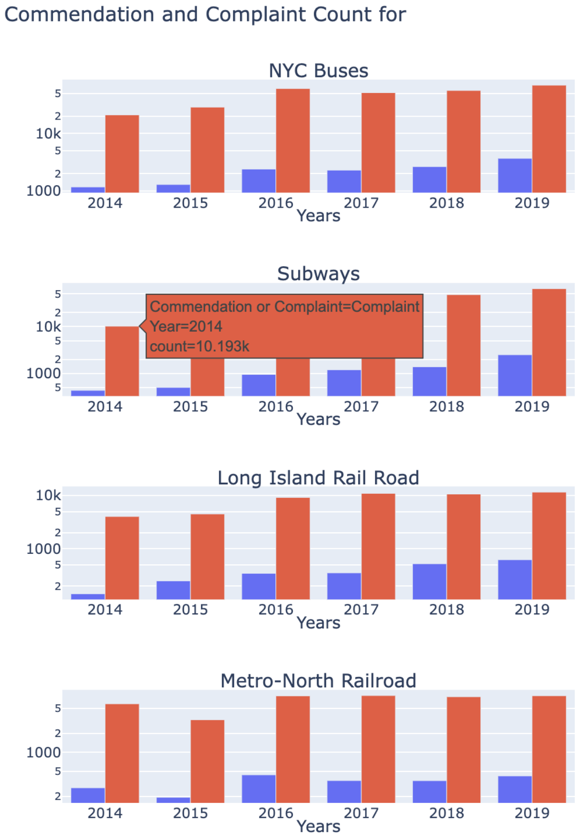


Figure 2. Sentiment Analysis on MTA

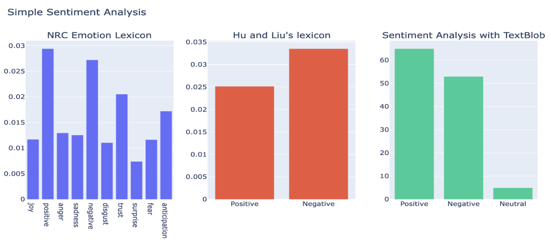


Figure 3. Word Clouds on MTA Reviews

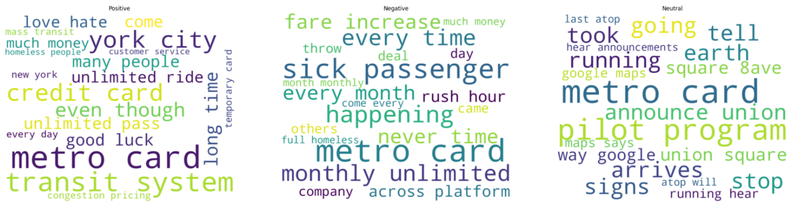
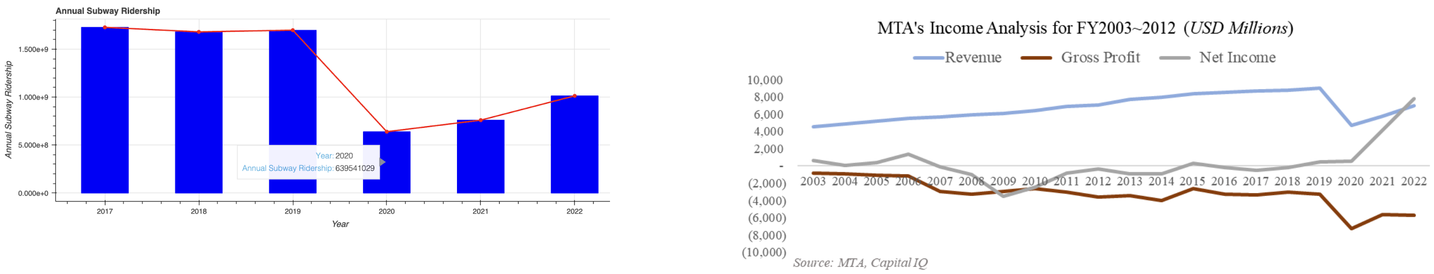


Figure 4. Financial Analysis on MTA



Source: MTA, Capital IQ

Figure 5. Time Analysis

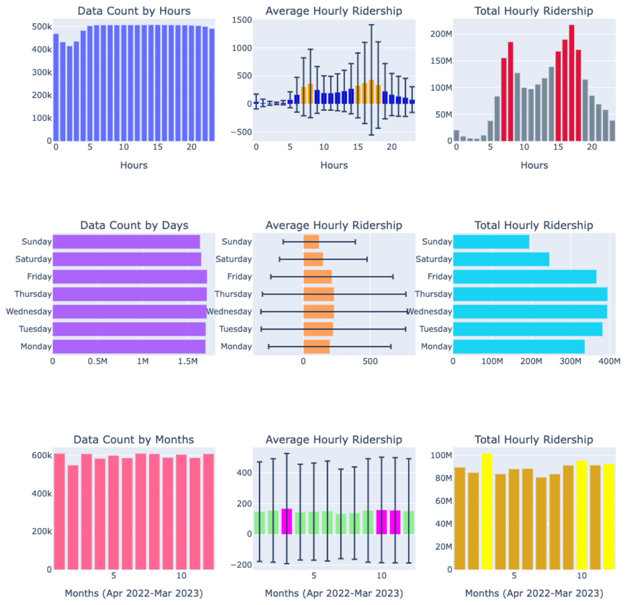


Figure 6. Spatial Analysis

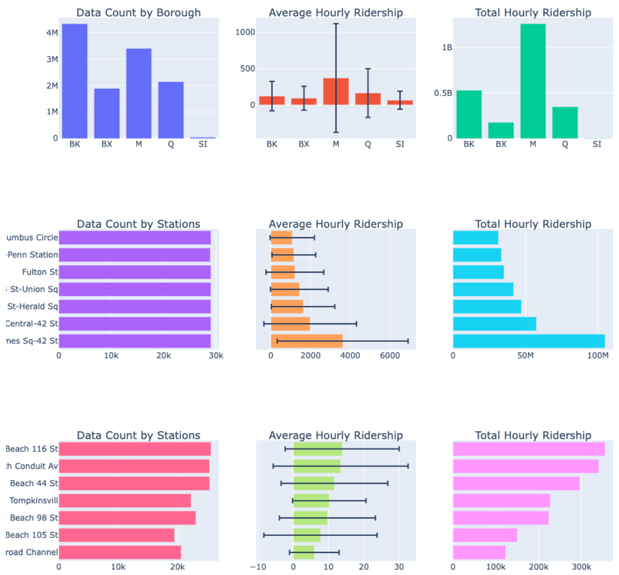


Figure 7. Weather Distribution

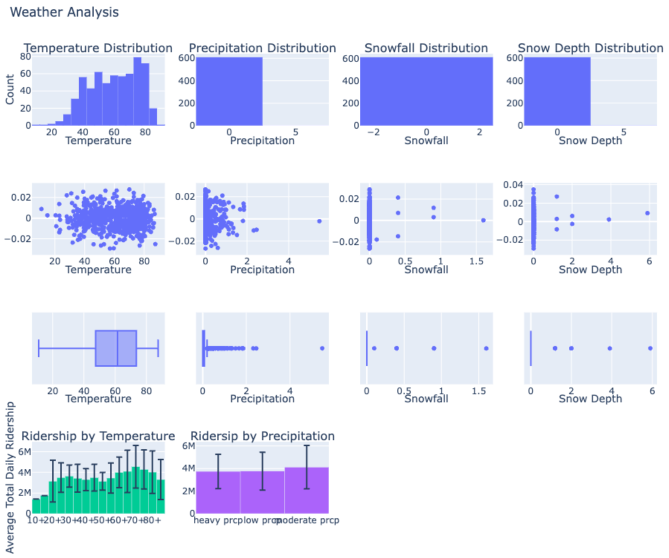


Figure 8. Average Temperature and Ridership

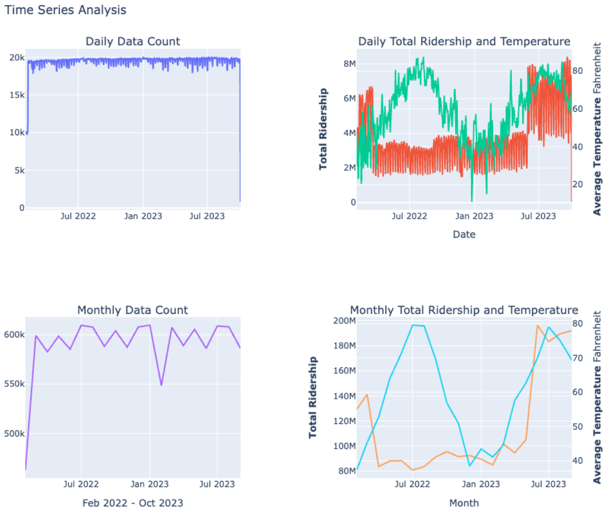


Figure 9. Analysis for Columbia University Station

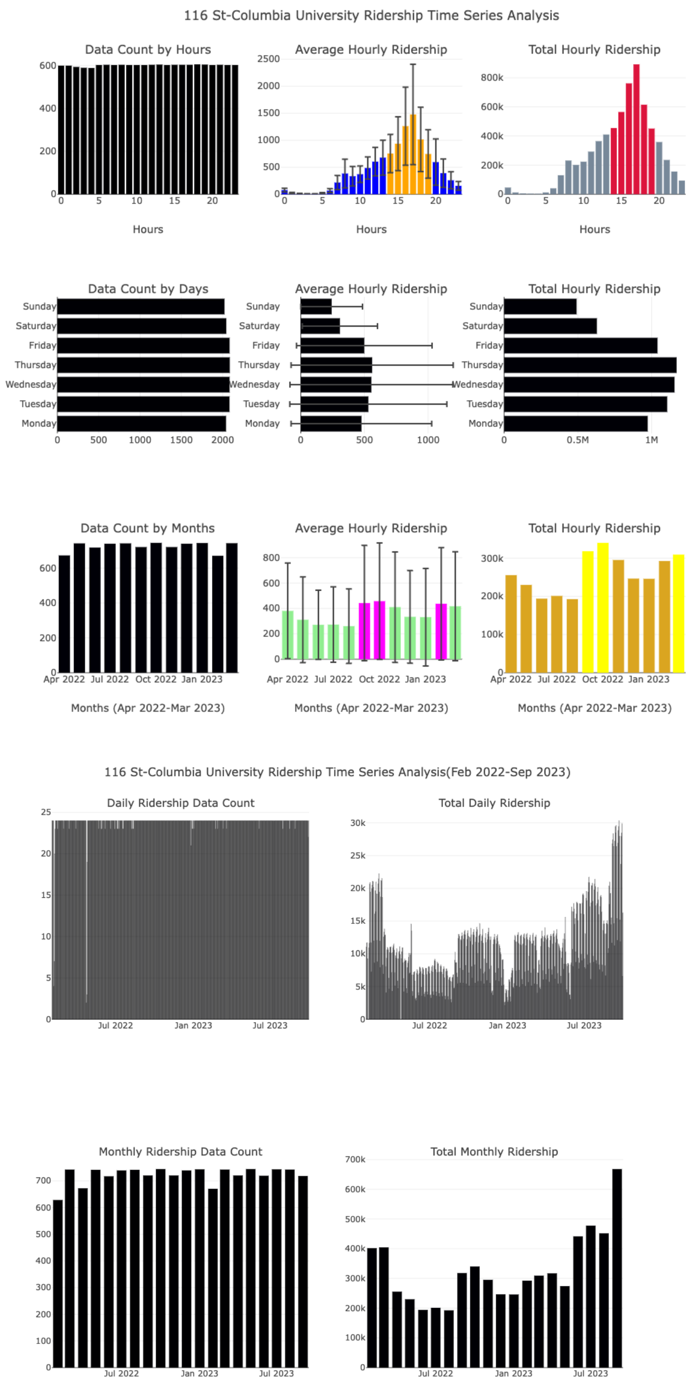


Figure 10. Analysis for Flushing - Main St Station

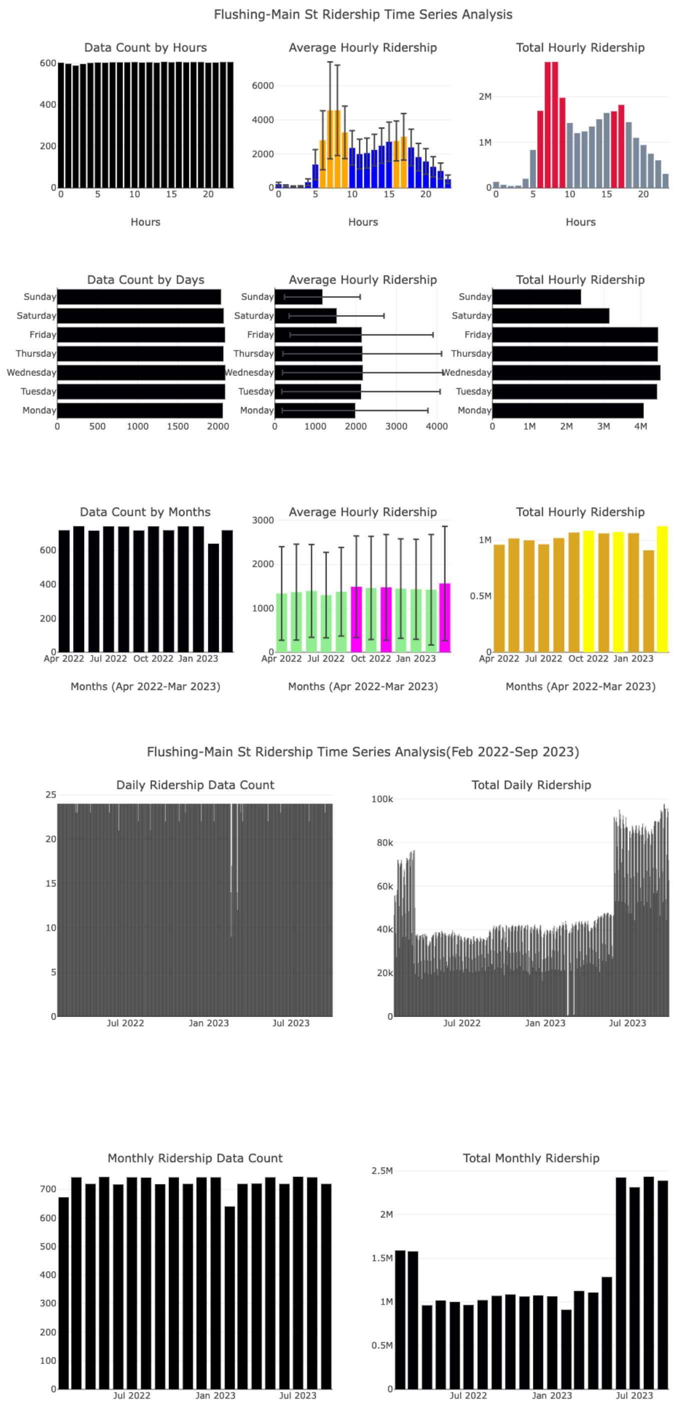
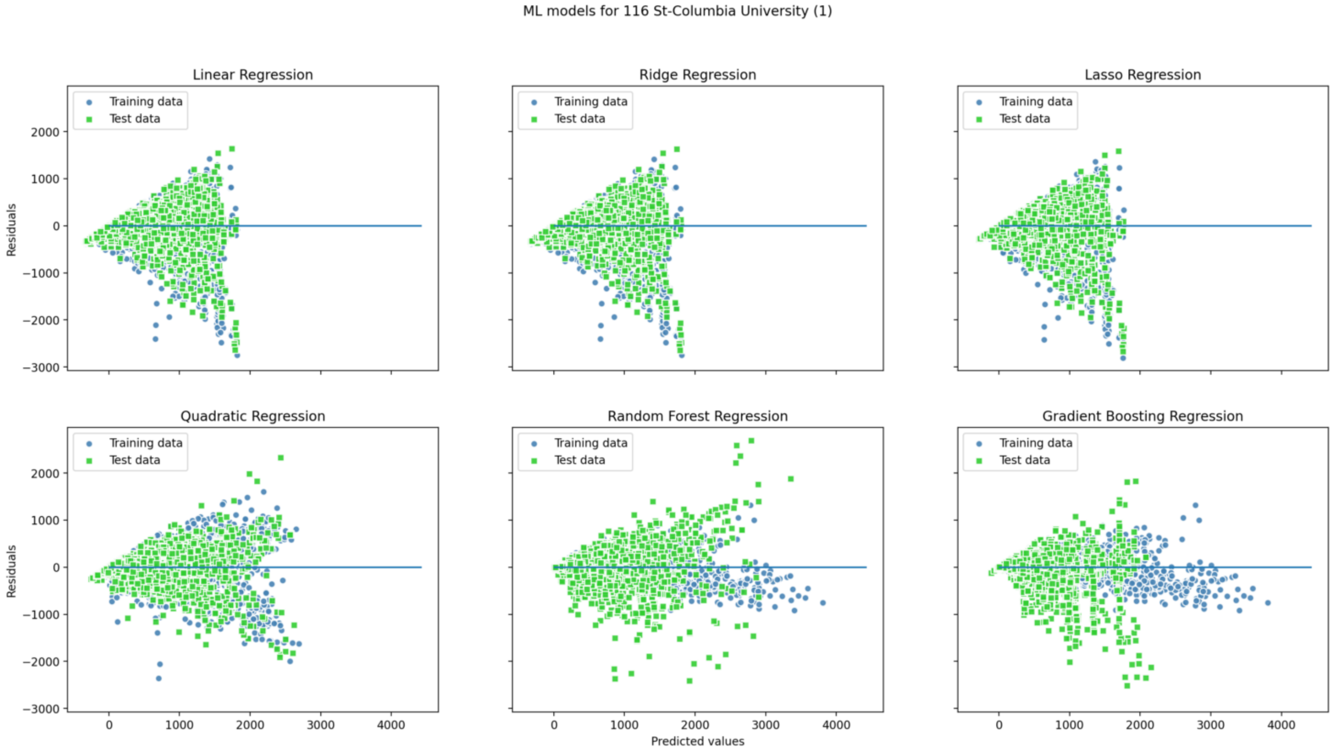
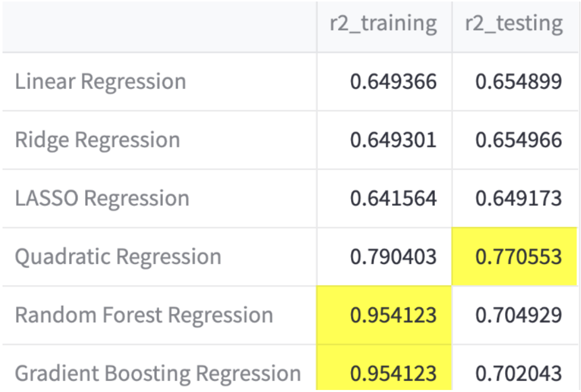
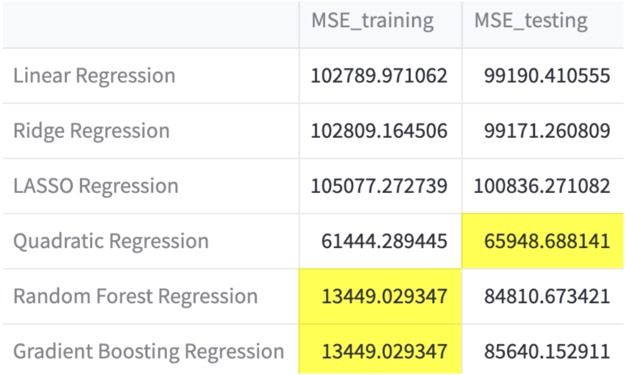


Figure 11. Model performance for 116-St Columbia University





	r2_training	r2_testing
Linear Regression	0.649366	0.654899
Ridge Regression	0.649301	0.654966
LASSO Regression	0.641564	0.649173
Quadratic Regression	0.790403	0.770553
Random Forest Regression	0.954123	0.704929
Gradient Boosting Regression	0.954123	0.702043



	MSE_training	MSE_testing
Linear Regression	102789.971062	99190.410555
Ridge Regression	102809.164506	99171.260809
LASSO Regression	105077.272739	100836.271082
Quadratic Regression	61444.289445	65948.688141
Random Forest Regression	13449.029347	84810.673421
Gradient Boosting Regression	13449.029347	85640.152911

Figure 12. Fare Distribution

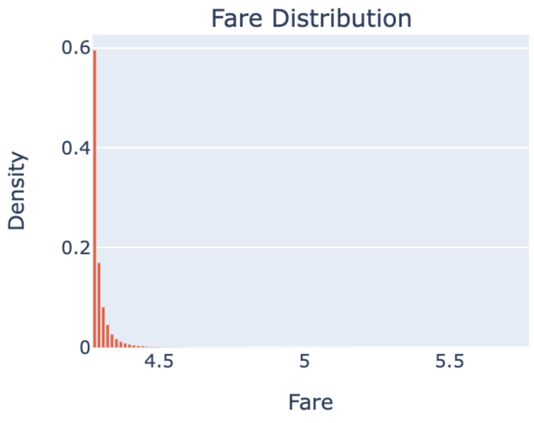
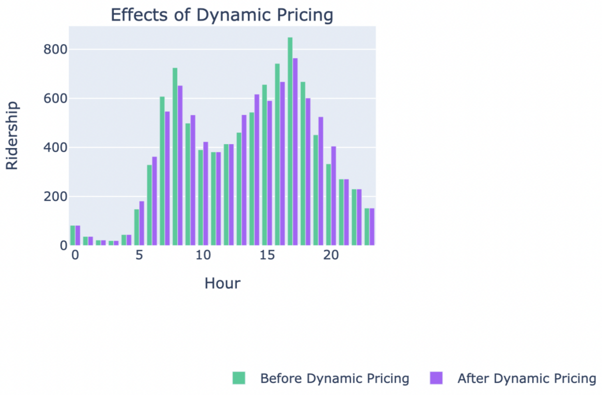


Figure 13. Effects of Dynamic Pricing


Effects of Dynamic Pricing
Ridership
Hour
🟩 Before Dynamic Pricing
🟪 After Dynamic Pricing