



2023 9th Annual Data Science Hackathon

Make  Great Again

Amanda Lin

Jung A Lim

Joy Yoon Soo Kim

Contents

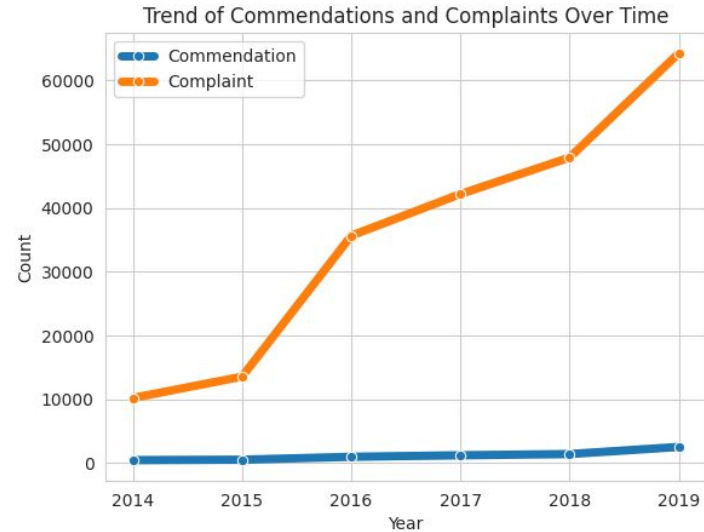
- I. Introduction
- II. Data Exploration
- III. Modeling
- IV. Conclusion
- V. Future Studies
- VI. Appendix - Datasets



I. Introduction

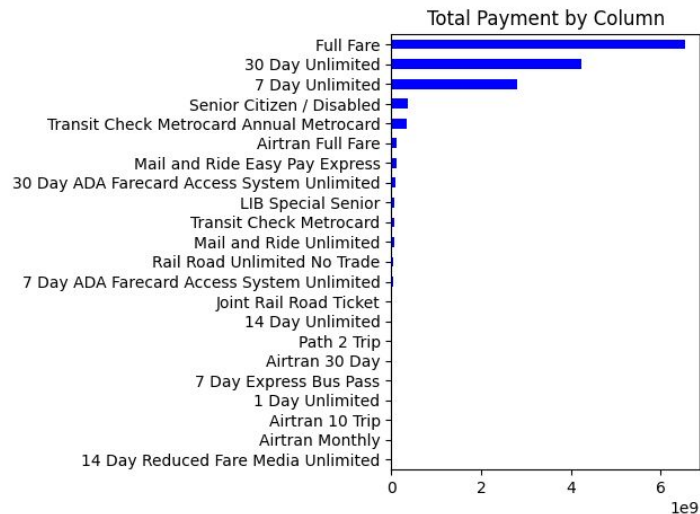
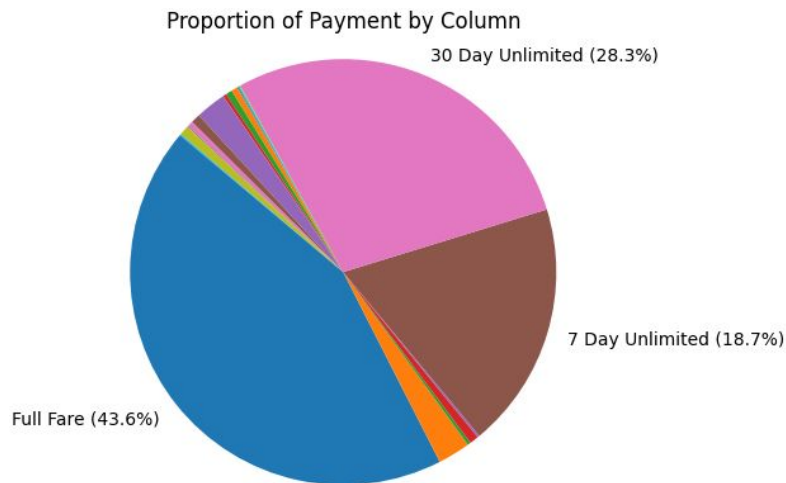


<https://youtube.com/shorts/gQc0yYHR988?si=pca7PwCkpuD5Ebsh>



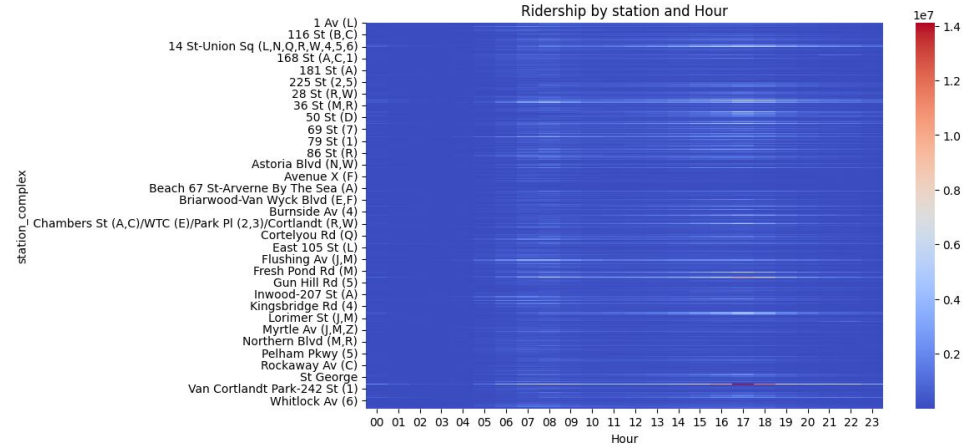
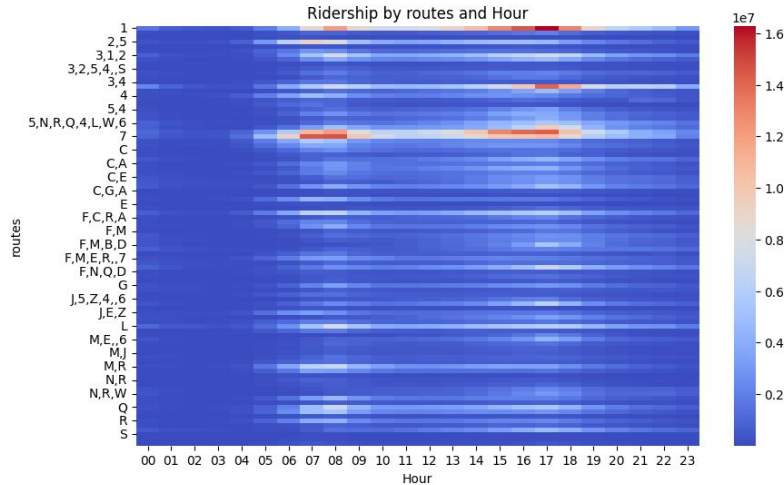
- While NYC MTA has served New York citizens for an extended period, data from the MTA's database reveals a consistent increase in customer complaints over the years.

II. Data Exploration: Ridership by type of payment method



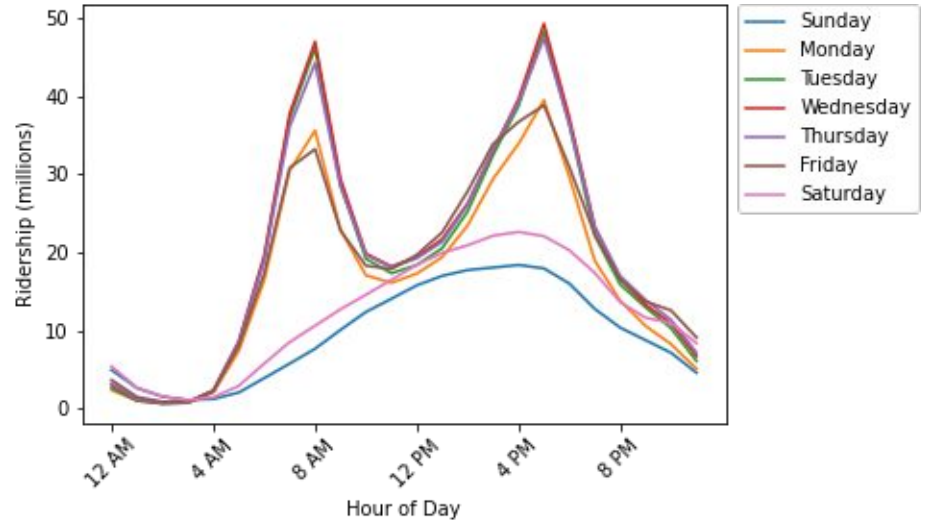
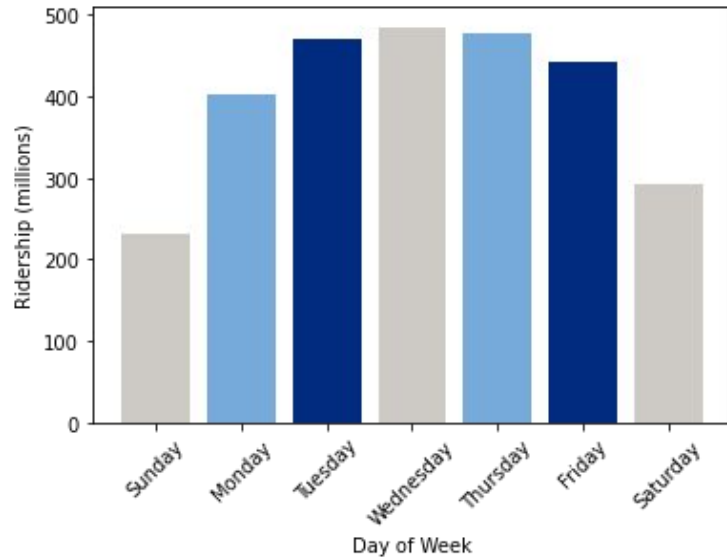
Approximately 90% of all payments fall into three main categories : “Full-fare”, “30 Day Unlimited”, and “7 Day Unlimited”.

II. Data Exploration: Ridership per hour at each stop



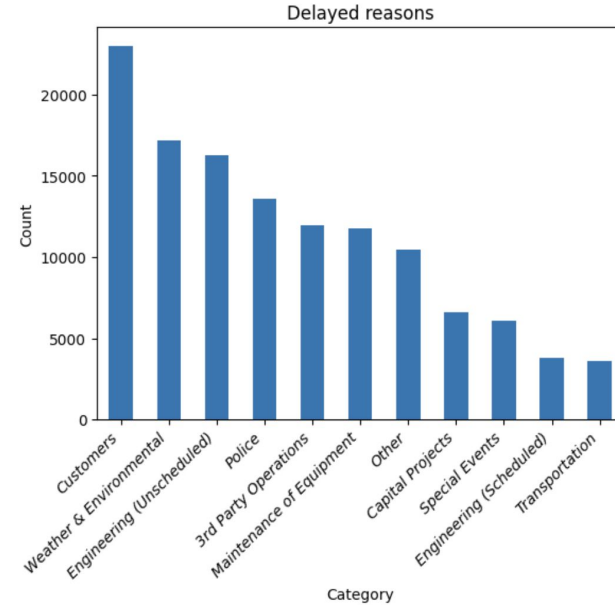
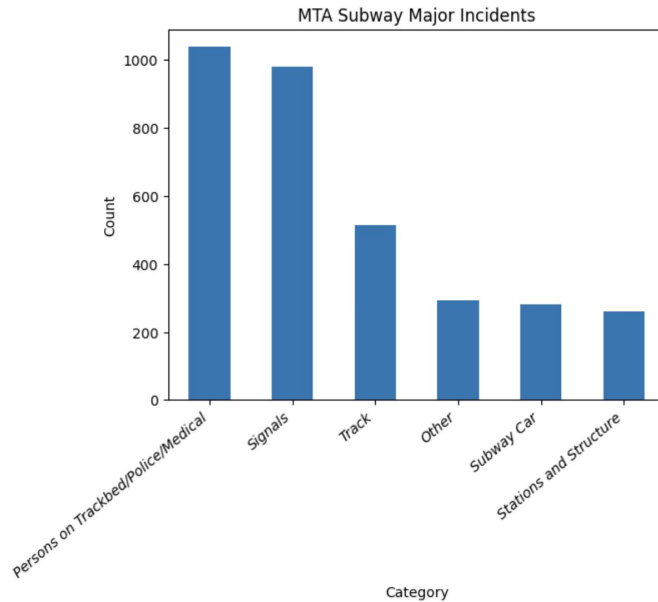
- This heatmap with hours and routes clearly shows that there's a peak hour in most of the lines. Based on this heatmap, we did more analysis for peak hour and peak day
- We found that some of the stations have more riders than the others, so we will use the clustering for the predictive model

II. Data Exploration : Number of Rideships for each day of week and hour of day



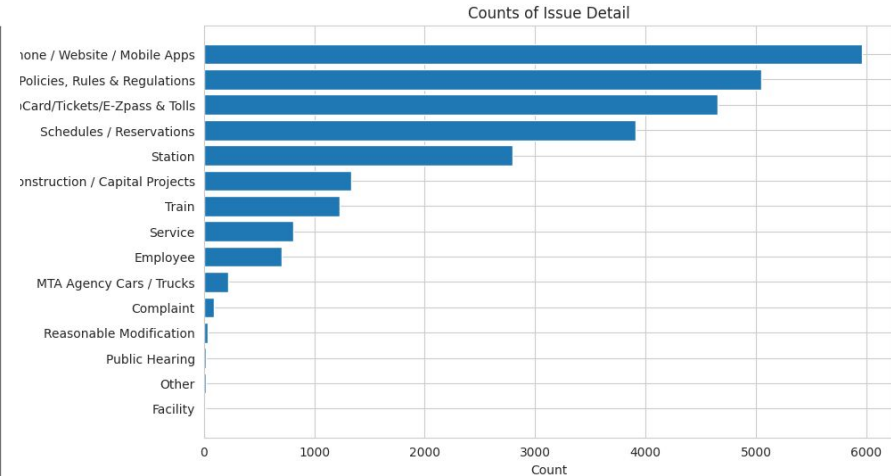
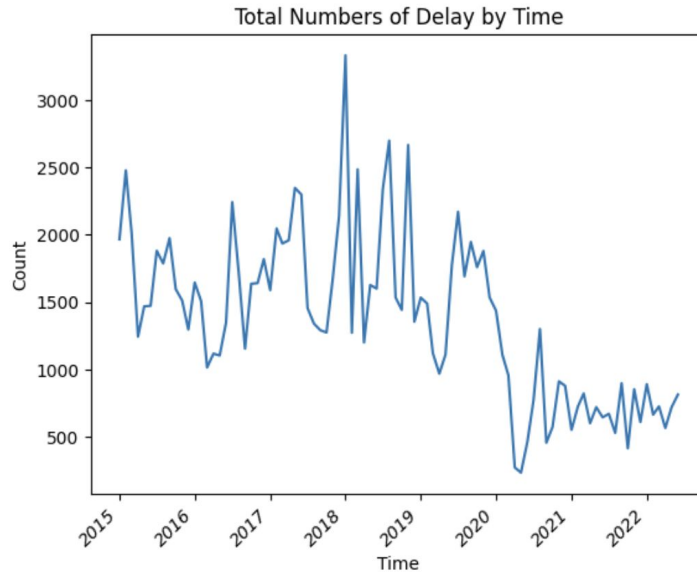
- Wednesday saw the highest riderships among the days of the week.
- Additionally, peak ridership occurred at 8 AM and 5 PM.

II. Data Exploration : Major Incidents and Delayed reasons in Subway



- Top three incidents occurred in the MTA Subway System are “Persons on Trackbed / Police / Medical”, “Signals”, and “Track”.
- The leading causes of delays in the MTA Subway are “Customers”, “Weather & Environmental”, and “Engineering (Unscheduled)”.

II. Data Exploration : Numbers of Delay and Counts of Issue

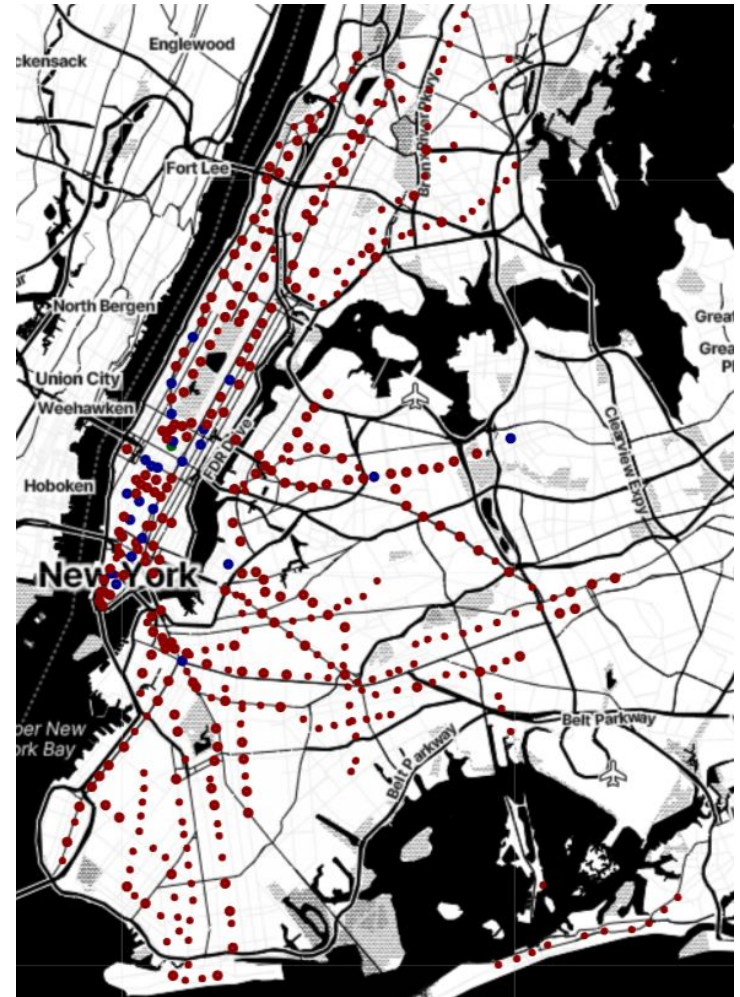


- The total number of delays varies each year, with a significant reduction observed in 2020. Subsequently, there has been a continuous decrease in the in delays since then.
- Based on customer complaints and the primary causes of delays, we recommend enhancing the air conditioning system during peak days and hours. Implementing energy-efficient AC usage or using it selectively can also contribute to environmental sustainability. Furthermore, employing time-dependent AC usage can enhance passenger comfort.

II. Data Exploration - Clustering

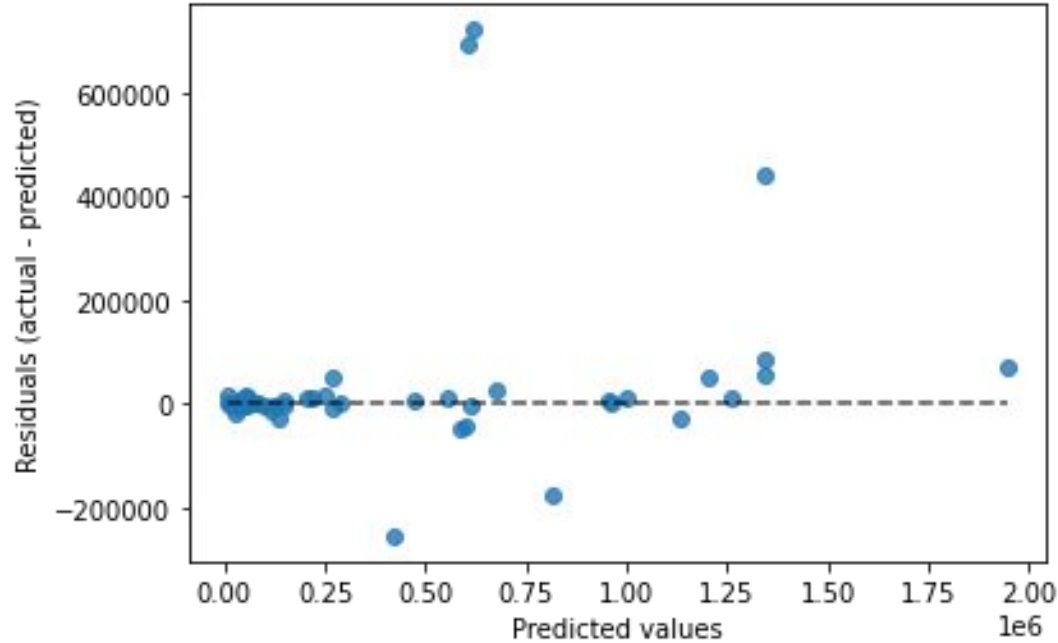
- Used K-means clustering by ridership.
- Time Square is its own cluster.

Cluster 1
Cluster 2
Cluster 3



III. Modeling

Decision Tree Regressor

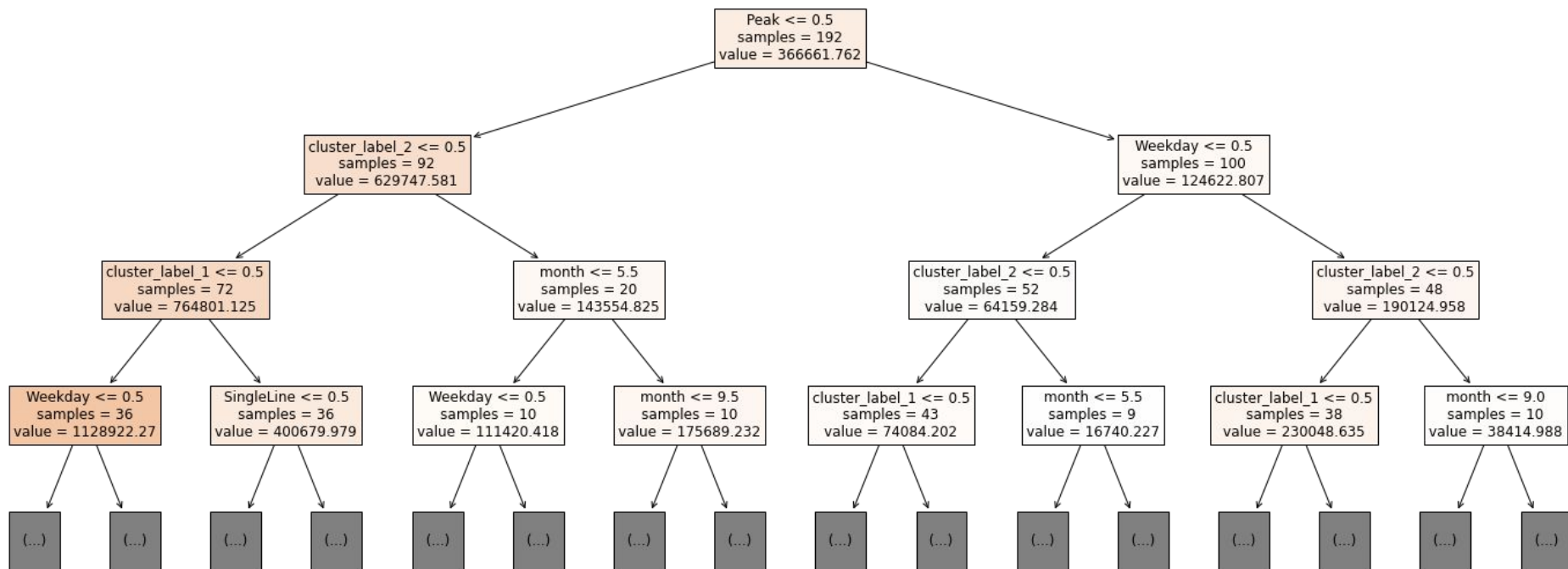


$R^2 = 0.905003$

Adj $R^2 = 0.902556$

III. Modeling

Decision Tree Regressor



IV. Conclusion

Based on our data analysis:

- The MTA can use ML techniques to predict surges of passengers
- AC and additional train cars can be used during such surges
- Times Square should be targeted for improvements because the most passengers transit through there
- This will reduce customer complaints

V. Future Studies

- Compare express vs local lines to determine if any stops should be made express stops or vice versa
- Create mathematical model to determine which stations should be prioritized for improvements such as AC, wifi, etc.
- Compare OMNY vs metrocard riders in different stations
- Listen to customers to improve MTA

VI. Appendix - Datasets

- MTA Hourly Ridership - Beginning February 2022
- MTA Fare Card History - Beginning 2010
- MTA Subway Major Incidents - Beginning 2020
- MTA Systemwide Delays - Beginning 2015
- MTA Complaints - Beginning 2014



2023 9th Annual Data Science Hackathon

Make  Great Again
Thank you.

Amanda Lin
Jung A Lim
Yoon Soo Kim