Predicting New York City Public Transportation Ridership

Marc Hughes
The Pennsylvania State University
State College, PA
mxh917@psu.edu

Tyler Otero
The Pennsylvania State University
State College, PA
tdo5076@psu.edu

Brady Miller
The Pennsylvania State University
State College, PA
bjm6583@psu.edu

Alvaro Tapia
The Pennsylvania State University
State College, PA
abt5572@psu.edu

Advait Ashtikar
The Pennsylvania State University
State College, PA
aaa6940@psu.edu

Brian MacCurtin
The Pennsylvania State University
State College, PA
bjm6594@psu.edu

Abstract

New York City is renowned for being one of the most densely populated and busiest cities in the United States, with hundreds of thousands of individuals commuting from one location to another on a daily basis for various reasons. However, various factors, such as weather conditions and day of the week, can significantly impact the demand for daily transportation modes, including buses, trains, and subways. Despite the availability of various weather forecasting models, the general public remains largely unaware of the potential impact of these conditions on public transportation ridership demand. Therefore, we aim to investigate and evaluate any potential circumstances related to weather conditions and day of the week that may affect public transportation ridership demand in New York City.

1. Introduction

New York City is one of the most densely populated cities in the United States, with a population density of over 28,211 persons per square mile ([1]). This significant population density within the urban center has a significant impact on the mobility of its residents, who are faced with challenges such as affordability and traffic congestion that make it increasingly difficult to navigate the city. As a result, many people opt for the use of public transportation due to its accessibility and extensive coverage, with over 3.6 million riders daily utilizing the buses and subways as their main methods of transportation ([2]).

Given the importance of transportation in modern urban life, it is critical to understand the factors that impact public transportation demand, including weather conditions, which can have a significant influence. In this sense, understanding the relationship between weather conditions and transportation demand is essential, as it can have a substantial impact on the commuting experience of many individuals. This gains a greater importance as with this, transportation agencies can better allocate their resources to meet demand during different weather conditions, such as snow or rain, and individuals can plan their day accordingly. This is the main motivation for the development of the project, providing better valuable insights and information to transportation agencies, businesses, and individuals alike, enabling them to better prepare for and navigate the transportation system during different weather conditions.

Therefore, this project aims to train and evaluate the bus ridership and subway ridership of NYC, with a focus on forecasting the demand for both methods of transportation regarding the impact of different factors. Additionally, we will analyze the correlations and accuracy of ridership and weather parameters to provide insight into the relationship between the two variables.

2. Source of the Data

For the purpose of this investigation, we utilized two primary data sources: Visual Crossing, a weather data and analysis tool developed by data scientists and weather professionals, and the Metropolitan Transportation Authority (MTA), the largest transportation network in North America, serving over 15.3 million individuals in the New York, Long Island, and Connecticut area.

The weather data utilized in this study was obtained through Visual Crossing's APIs and Enterprise Tools, which are available to participants globally. The transportation data, on the other hand, was acquired directly from the MTA.

It is worth noting that the data collected for this study was raw meaning that no prior analysis or manipulation had been conducted on the data prior to our investigation.

3. Description of Data

Our data collection process involved multiple sources, one of which was the Metropolitan Transportation Authority (MTA), a leading transportation agency. The MTA's Open Data Catalog provided us with a comprehensive dataset comprising numerous records on transportation. This encompassed a wide range of metrics such as daily ridership, total customers, stops covered, and average duration of trips. For this project, we focused on the average daily ridership figures for buses and subways, which were critical to addressing our research problem.

Additionally, we sourced weather data from Visual Crossing, a weather data and analysis tool developed by experienced data scientists and weather professionals. With almost a century's worth of historical weather data for New York City, Visual Crossing provided us with reliable weather data from the post-COVID era. This data was crucial in creating a predictive model that offered valuable insights into the relationship between weather conditions and daily ridership.

By combining the average daily ridership data with the "Temp" and “Snow Depth” weather data obtained from Visual Crossing, we were able to draw insightful conclusions and generate actionable reports that could aid transportation offices and everyday New York City residents.

4. EDA and Results Interpretation

In the initial stage of our project, it was imperative to conduct a thorough examination and assessment of our data, with a focus on identifying any possible correlations or similarities between the ridership and weather data. To achieve this, we began by analyzing the demand for public transportation on different days of the week, to determine the day with the highest ridership.

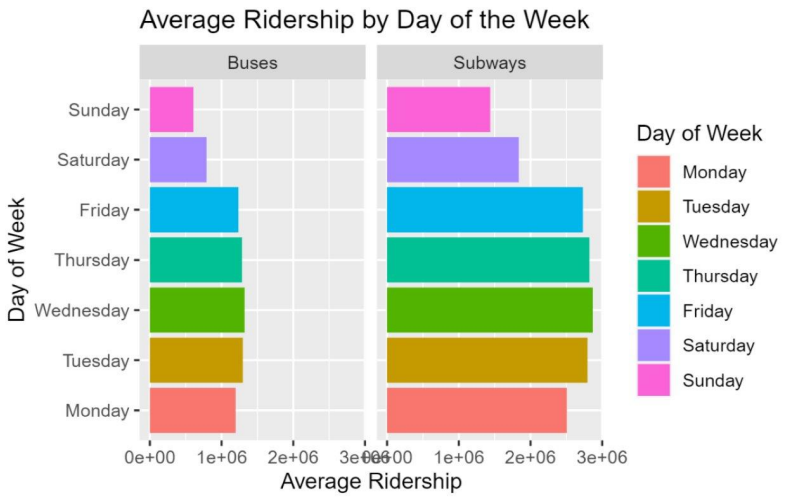

Figure 1: Illustration of Average Ridership vs. Day of the Week between Buses and Subways

Based on Figure 1, it is possible to visualize that Wednesday had the highest average ridership, with Thursday and Friday following closely behind with approximated values of 1.3 million and 1.2 million riders respectively. It can be noted that there was a significant drop in ridership over the weekend. Additionally, we discovered that the average ridership of the subway was nearly three times greater than that of the buses with approximately 2.9 million riders on Wednesday. This finding is particularly relevant as it highlights the sensitivity of the subway system to changes, given its larger scale of data.

Following the initial analysis, it was imperative to explore potential correlations between the chosen modes of public transportation and specific weather parameters. To achieve this, a bar plot was constructed, clearly classifying the weather values and transportation types.

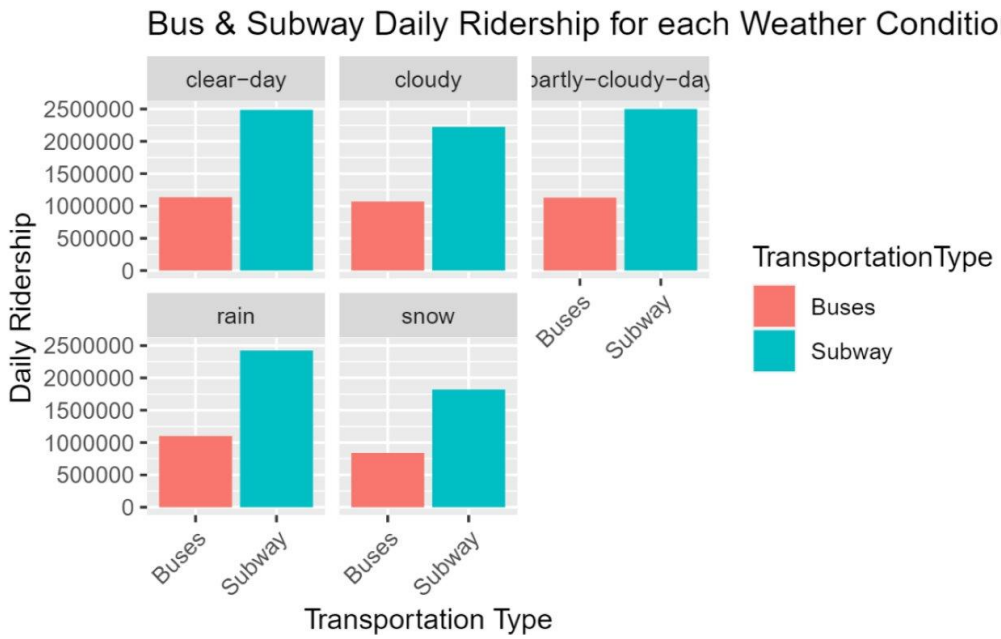

Figure 2: Visualization of Transportation Type vs. Daily Ridership between selected weather parameters

This analysis of public transportation data reveals a notable difference in ridership between buses and subways. In particular, subway ridership exceeds that of buses by a factor of two and a half, indicating a significant disparity between the two modes of transportation. However, on days with clear weather conditions, there is a marked increase in the number of commuters utilizing both bus and subway services. Conversely, snowy weather leads to a significant reduction in subway ridership, while bus ridership remains relatively stable across all weather conditions, assuming other factors remain constant.

To continue with the investigation, it was highly important to also analyze the temperature for each day of the week since this could be impactful on ridership numbers.

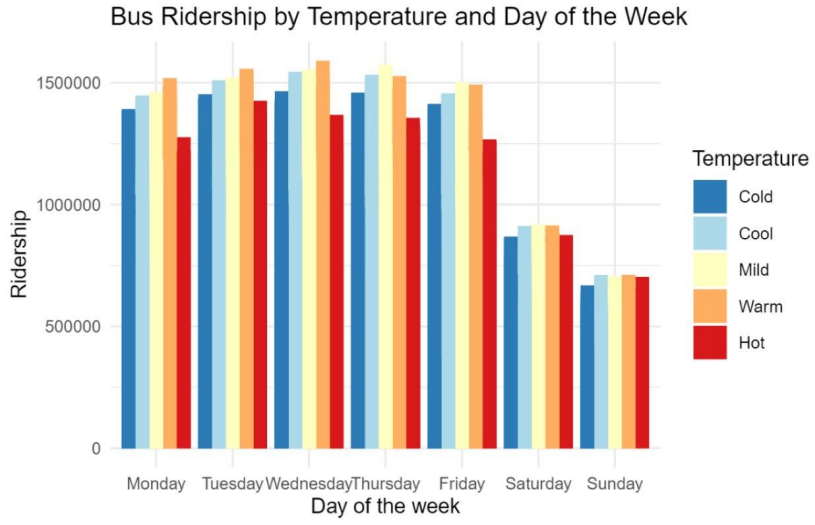

Figure 3: Visualization of Bus Ridership vs the Day of the Week divided up by temperature ranges

We learned something very interesting seeing the correlation of “Temperature” and daily “Ridership” for buses. Buses tend to be highly affected by both cold and hot temperatures. As the temperature nears the extremes of both hot and cold, bus ridership begins to significantly drop.

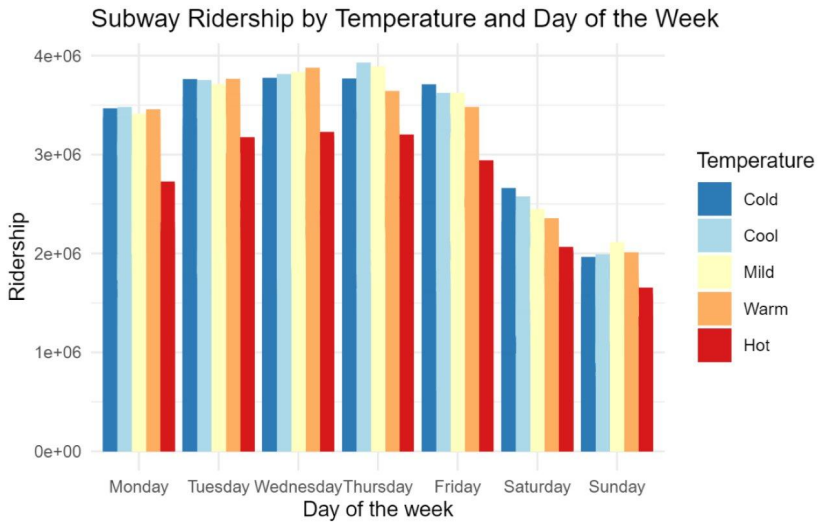

Figure 4: Visualization of Subway Ridership vs the Day of the Week divided up by temperature ranges

When doing our analysis on the subway ridership, we learned that there was a negative correlation with hot temperatures and average daily ridership. Ridership dramatically drops as the temperature increases. This is most likely due to a drop in ridership during the hottest months of the year.

5. Referencing Similar Investigations

After conducting a thorough investigation into potential similar projects or studies related to our own, we found one comparable work undertaken by Cornell Tech University [4]. This inquiry examined the impact of weather and special events on public transportation ridership demand in New York City, with a focus solely on subway ridership, whereas our research examines both subway and bus ridership. The Metropolitan Transportation Authority (MTA) dataset for public transportation ridership in NYC is a valuable resource utilized by both projects. However, there are differences in the weather datasets used, as the Cornell study acquired data from the National Center for Environmental Information [5], while we employed Visual Crossing, a weather data and analysis tool developed by data scientists and weather professionals, which offers a more extensive range of parameters and features related to weather.

While both datasets are legitimate and valid for this study, our preference for Visual Crossing was based on its greater comprehensiveness. Furthermore, our research objectives differ from those of the Cornell study. Whereas their focus was on reducing congestion in high-traffic stations and minimizing the risk of danger to the public, our study seeks to forecast ridership demand, allowing individuals to plan their daily travel according to the impact of weather, events, and holidays on public transportation usage. Additionally, we aim to assist transportation agencies in better allocating their resources to meet ridership demand during different weather conditions. Thus, while both projects use some similar datasets and analyze common factors, there were distinct approaches that were taken to provide value to New York City.

6. Model Creation

We first decide to fit two full models with all variables in order to predict bus and subway ridership. From both of the full model summaries, there were some variables that were significant, but not all. We realized that there could be multicollinearity so we wanted to explore the relationships between variables further.

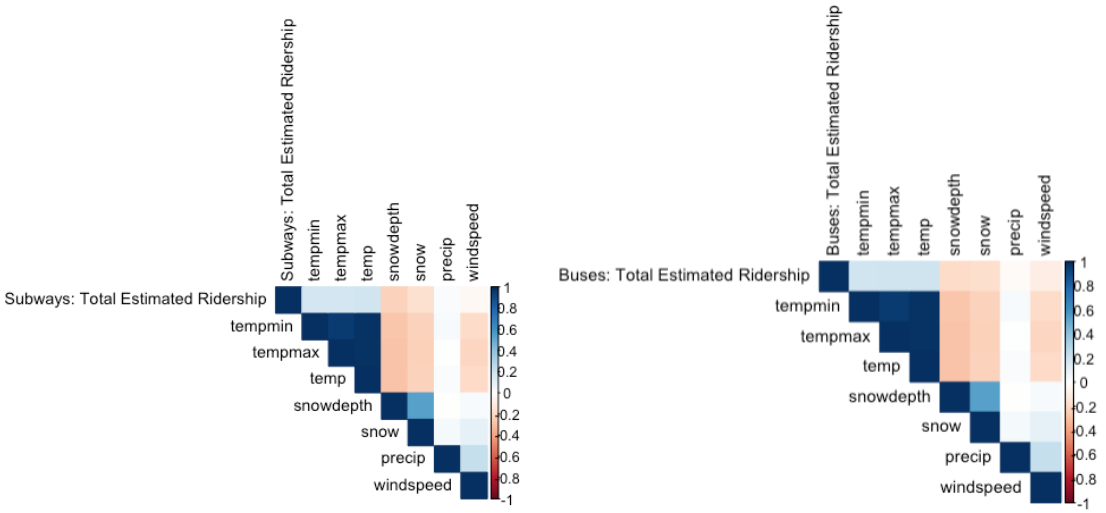

Figure 5: Correlation Matrices for Buses and Subway Variables

We looked at the full model correlation matrix for all of the numerical variables to look for potential multicollinearity for both models. We saw that both models gave very similar results. Variables like ‘tempmin’, ‘tempmax’, are naturally heavily correlated with the ‘temp’ variable because all three of these variables are giving us information on temperature for the day. These variables all had VIF values of greater than 90. Also, the variable ‘snow’, which is measured daily, was correlated with ‘snowdepth’, which is the measurement of the total snow on the ground. These both had extremely high VIF values. Besides that, none of the other variables seemed heavily correlated with any other variables

Due to a lot of the variables giving us very similar information, we decided that it would make sense to include just one temperature related variable and just one snow related variable. We decided to use the ‘temp’ variable over the ‘tempmin’ and ‘tempmax’ variables because we concluded that ‘temp’ variable was just a generalization of the other two. We also thought that ‘snowdepth’ would be better than just the ‘snow’ variable because ‘snowdepth’ affects people's ability to use public transportation more.

With this in mind, we used the remaining variables in our model selection process. We decided for both our subway and bus data sets to use the backwards selection technique to create the best possible subset of predictors. Using backwards selection, a larger percentage of the variables became significant. Although a few variables aren't significant, they are barely insignificant and we didn’t want to remove them because we thought that the backwards model gave us the best possible subset of predictors we could use. In both models, the variables that were used were day of the week, temperature, and precipitation. The buses model also included snow depth as a covariate.

We then checked again for the presence of multicollinearity in the backwards models. From the table, we can see that none of the variables have a GVIF above 2. We use GVIF since we have categorical predictors in our models. GVIF is just the square root of the VIF, so they behave similarly. We get GVIF values well below 2, which is a sign that there is no presence of multicollinearity in our models.

Buses

	GVIF
Day of Week	1.010600
Buses: % of Comparable Pre-Pandemic Day	1.179453
temp	1.134486
precip	1.014720
snowdepth	1.172811

Subways

	GVIF
Day of Week	1.102527
Subways: % of Comparable Pre-Pandemic Day	1.208726
temp	1.118419
snowdepth	1.147487

Figure 6: GVIF Values Table for Buses and Subways Backwards Model

From here, we wanted to begin creating models to predict public transportation ridership using the insights gained above. Due to our data being time series data, k-fold cross validation would have been an improper way of validating our data. As a result, walk-forward validation was used in order to properly capture our various models’ accuracy. We used this cross validation method by sectioning our data into five time periods of increasing size, training and testing on each period, and then calculating the accuracy. When we performed the validation method on our time series data, we used it alongside support vector machine (SVM) and linear regression (LM)

models with the ‘train()’ function. After finding the root-mean-square error (RMSE) for each specified time period within walk-forward validation, we found the mean of those RMSE values to get our overall accuracy values pictured below.

Model <chr>	RMSE <dbl>
Bus SVM	122482.8
Sub SVM	287872.4
Bus LM	49402.5
Sub LM	120341.6

Figure 7: Table of RMSE Values

While these RMSE values may seem quite high, we believe this is due to the context of our problem. The ridership value is in the hundreds of thousands and even millions, so a difference in predicted ridership compared to actual ridership of about 50,000 isn't necessarily a huge difference or 'issue' in the context of this problem. These seemingly large differences along with some outliers, could cause the root mean square error to be skewed and very large, but many of the predictions may not be super far off in terms of being in a certain range from the actual ridership. Based on all this, we decided that despite having a non-linear set of data for the response variable, we would use the linear model method to make predictions.

For each data set, we split the data sequentially into training, and testing, using the training data to make predictions on the test data. Below is the graph for bus ridership predictions, with the actual ridership in red and the predictions in blue

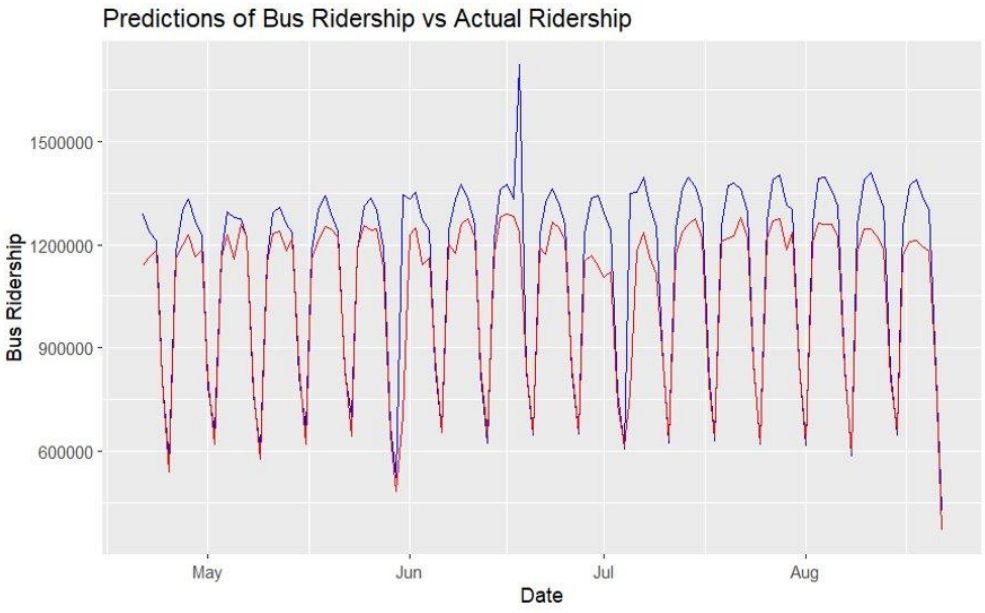

Figure 8: Predictions of Bus Ridership vs. Actual Ridership

As you can see it does a good job following the overall trend of the actual ridership, but it does over predict the ridership at the top of the curves. There also seems to be a massive outlier occurring in mid-June, which we don't currently have an explanation for. Another thing we noticed from the graph is that the graph lines dip a few times every month, in a cyclical pattern, which we attributed to the weekends having less bus ridership, probably due to the weekdays requiring people to commute to work.

The subway data set was split in the same fashion as the bus data set, and the graph below shows the predictions of subway ridership in blue compared to the actual ridership in red.

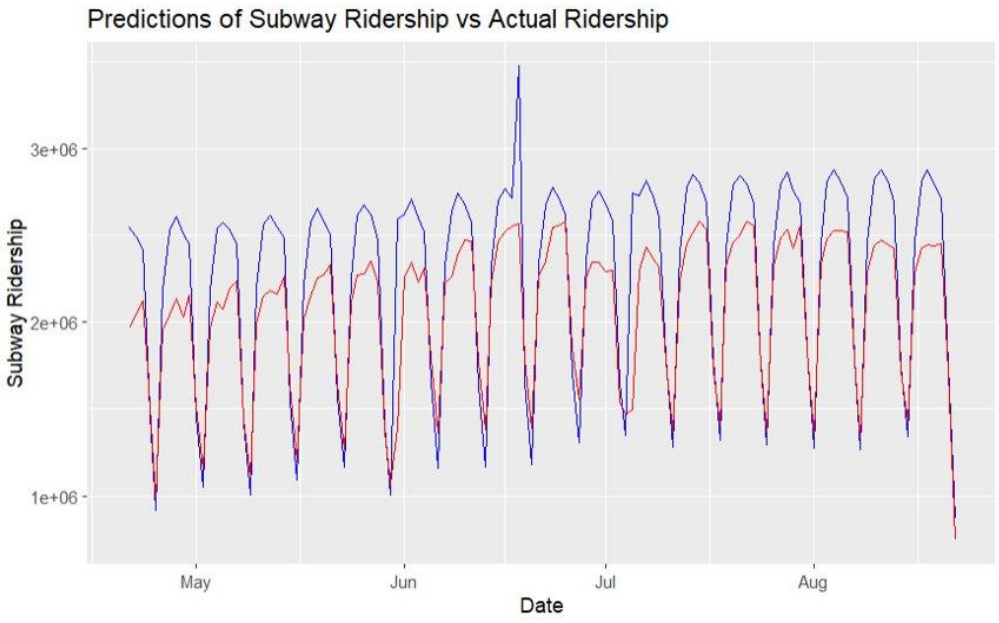

Figure 9: Predictions of Subway Ridership vs. Actual Ridership

Similar to the bus graph, it does a good job following the overall trend of the actual ridership. However, it is worse than the bus data was in over predicting the ridership at the top of those curves, and it also has that one outlier point sticking out in the middle. This model also seems to under-predict the ridership at the bottom of the curves as you can see the actual values are a bit above the predicted values at each point along the bottom.

7. Conclusion

Based on the results obtained from the models and plots created for this project, it can be concluded that we have developed a model to help predict the demand for bus and subway ridership in New York City. The project has effectively met all the expectations and goals previously established.

The outcomes of this project provide significant value by offering essential insights to residents of NYC, businesses, transportation systems, tourists, and city officials regarding the projected level of activity for the public transportation systems on a given day based on the effect of the weather and day of the week. Overall, the results of this project have demonstrated the potential to significantly enhance the efficiency and effectiveness of the public transportation systems in New York City.

8. Future Investigations

For the future development of this project, there are several areas that could be improved upon. Firstly, it would be beneficial to analyze how specific types of natural disasters, such as tornadoes or winter storms, could impact the weather. Currently, our predictive capabilities may not be entirely accurate in the face of such events, as we lack the necessary data and these occurrences often happen without warning. By addressing this gap, we could enhance the accuracy of our predictions.

Additionally, the project could be refined by tracking the hour to hour ridership on subways to more precisely predict ridership. As shown above, our models regarding subway ridership did not perform as well as our models predicting bus ridership. This is largely due to the nature of weather's inherent interaction with buses, which are outside, versus subways, which are underground. Utilizing the hourly ridership of specific subway routes would undoubtedly help bridge this gap in accuracy.

Finally, while we are already providing valuable assistance and information to transportation companies and individuals by predicting rider demand on specific days based on weather conditions, future work could involve real-time updates or data in the event of unexpected incidents. Being able to notify users of any changes before they take action could further enhance the usefulness and reliability of the research.

9. References

[1] Maciag, M. (2021, June 2). *Population density for U.S. cities statistics*. Governing. Retrieved April 20, 2023, from https://www.governing.com/archive/population-density-land-area-cities-map.html

[2] MTA. (n.d.). *Subway and bus ridership for 2021*. MTA. Retrieved April 20, 2023, from https://new.mta.info/agency/new-york-city-transit/subway-bus-ridership-2021

[3] Visual, C. (n.d.). *Historical weather data for New York City,USA*. Historical weather data for New York City,USA | Visual Crossing. Retrieved April 20, 2023, from https://www.visualcrossing.com/weather-history/New%20York%20City,USA/us/2022-01-01/2023-04-09

[4] Rasool, R. (2019, May 13). *Analyzing modes of transportation in New York City*. Medium. Retrieved April 24, 2023, from https://towardsdatascience.com/analyzing-modes-of-transportation-in-new-york-city-dfb4a1726ac4

[5] National Centers for Environmental Information (NCEI). (n.d.). *Climate Data Online: Dataset Discovery*. Datasets | Climate Data Online (CDO) | National Climatic Data Center (NCDC). Retrieved April 24, 2023, from https://www.ncdc.noaa.gov/cdo-web/datasets#GHCND