Final Report

1-2. Introduction & Problem Definition

Weather conditions significantly influence public transportation usage, but the extent of this impact remains understudied. In New York City—where millions rely on the subway daily—understanding how factors like temperature, precipitation, and holidays affect ridership can empower transit planners to optimize service and help commuters make informed travel decisions. This project leverages statistical modeling and machine learning with historical data to quantify these relationships and predict subway ridership under varying weather conditions.

3. Literature Survey

As of 2020, the U.S. Census Bureau (Vintage 2023) estimated that New York City and its boroughs have a population of approximately 8.8 million, making NYC the most populous city in the United States. This immense population imposes significant pressure on the city's public transportation infrastructure—particularly on its subway system, the most widely used mode of transit according to MTA (2023). Given these demands, improved methods for forecasting ridership can be highly valuable to transit operators, everyday commuters, researchers, and students interested in external impacts on subway ridership.

Many researchers have studied how weather influences public transportation ridership. Ngo (2024), for example, explores the effects of extreme weather on U.S. transit systems by running separate regressions for different time segments, whereas Arana (2014) gathers and visualizes data to show how temperature, rainfall, and wind correspond with trip volumes. In China, Jiang (2023) employs multiple linear regression and generalized linear fixed-effects models to examine metro ridership in three major cities, and Ding (2018) deconstructs Beijing's subway passenger volume into a deterministic component derived from an ARIMA model and a stochastic volatility component based on a non-linear GARCH family approach. Further explorations include Tang (2021), who applies time-series decomposition (separating seasonal and epoch factors) to predict short-term traffic on Chongqing Rail Transit, and Ding (2016), which uses gradient boosting decision trees for short-term forecasting across three Beijing subway lines. Other notable studies include Brazil (2017), applying regression models to 30 train services on Dublin's DART network, Kim (2020), assessing weather and calendar effects via generalized linear models on 20 months of smart card data, and Vitello (2024), combining mobile crowdsensing data with regression methods to estimate subway station demand. While these works collectively establish a consistent weather-ridership connection, most depend on comparable parametric models, potentially oversimplifying the influences on ridership. Additionally, many focus on more recent Chinese rail systems, which may not translate directly to the historical complexity of older networks like New York City's subway.

Building on these earlier findings, we aim to strengthen predictive accuracy and address limitations in a more nuanced way. Previous research findings indicate that linear regression has proven

useful for revealing certain weather-ridership relationships (e.g., Ngo 2024; Arana 2014). Existing research offers multiple angles on how weather shapes travel mode decisions: Singhal (2014) notes that severe events such as heavy snow may push bicyclists and drivers to switch to transit, whereas Cui (2023) and Ettema (2017) show that wind speed can negatively influence passenger flow and temperature shifts can sometimes increase ridership. Similarly, Lepage (2020) highlights a negative effect of rainfall on usage.

These collected insights directly assist in furthering our team's exploration of NYC's subway system, whose long history and global importance make it an excellent case for assessing and refining the proposed predictive methods.

4. Proposed Method

Building on our literature review, our project will explore the following innovations: predicting ridership by incorporating holiday data & more weather variables; implementation of a multiple linear regression design to capture the relationship between the predictors and ridership; using machine learning to build additional predictive models for predicting ridership; creating interactive visualizations to analyze the ridership and weather interactions; developing a front-end and back-end system for model inferencing to predict daily ridership.

We began the data pre-processing phase focused on cleaning and transforming the data, reducing the large csv dataset of 111 million rows to a 7 GB SQLite database through normalization and creating lookup tables for categorical variables. Next, we explored traditional time series models, such as Holt-Winters Exponential Smoothing and SARIMA, which are useful in identifying some patterns in the data. However, they may fall short in accurately capturing complex patterns influenced by non-linear variables such as weather conditions like temperature and rainfall, and holidays. By incorporating these features, we expect our approach to deliver more accurate forecasts that consider these additional sources of variation.

The dataset for predictive modeling was ingested from various sources which was further processed to obtain the final dataset for modeling. The daily ridership was obtained by aggregating the hourly dataset to daily, which consisted of 1,325 daily ridership from June 2021 to December 2024. We then collected 26 weather features from Open-Mateo API, federal holidays from Azure Open Datasets and NYC public school holidays from and NYC Department of Education website. We merged the datasets, performed Exploratory Data Analysis (EDA) using scatter plots, histograms, and box plots, and conducted multicollinearity analysis (correlation heat map, Variance Inflation Factor – VIF) to clean redundant features such as temperature min/max, apparent temperature min/max, rain_sum, wind_gusts_10m_max, et0_fao_evapotranspiration', etc (Fig 1). Based on the histograms, while most of the data looked normally distributed, features like rain and snowfall were strongly right skewed, and temperature, daylight, ridership had bi-modal peaks, raising corners over the violation of linear assumptions. Based on the scatter plots and box plots, weekend, holiday, precipitation, snowfall, and windspeed displayed promising

correlation with ridership. Furthermore, feature engineering was another key part of the process, where we developed new features such as “temp_2m_range” and categorized weather conditions into three groups based on the World Meteorological Organization (WMO) code (Fig 2, Fig 3). These engineered features, combined with standard data preparation techniques, allowed us to build a more structured and meaningful dataset for our modeling tasks.

The final dataset consists of a range of features, including temperature (mean and range), wind speed, precipitation, weather categories (Fair, Moderate, Severe), holidays, weekends, snowfall, and others. This comprehensive data set enables an in-depth analysis of how various factors influence subway ridership and supports the development of predictive models that account for these effects. The primary research questions guiding this research include: What are the most significant predictors of daily subway ridership? How do weather conditions, such as temperature and precipitation, impact ridership levels? What is the influence of public holidays and weekends on subway usage? Finally, can machine learning models surpass traditional time series methods in terms of predictive accuracy?

Beyond time series methods, we explored a range of machine learning techniques, leveraging both parametric and non-parametric models to analyze the data from multiple perspectives and improve predictive power. To capture the linear relationship between the features and daily ridership, we used multiple linear regression and regularized regression methods–including Ridge, Lasso, and Elastic Net. These models offer interpretability, helping us quantify the impact of individual features and assessing statistical significance. However, they might struggle to fit non-linear and complex real-world patterns, leading to potential underfitting and bias.

On the other hand, non-parametric models such as tree-based approaches (e.g., Decision Tree, Random Forest, Gradient Boosting), dimensionality reduction with Principal Component Analysis (PCA), K-Nearest Neighbors (KNN), and more complex models like Recurrent Neural Networks (RNN) with Long Short-Term Memory (LSTM), were tested to handle feature complexity. These models excel at capturing non-linear patterns and interactions, particularly for variables such as weather and holiday conditions, that may not adhere to strict linear assumptions and improve the prediction performance. Especially, during the training of the RNN-LSTM model, we further feature engineered temporal features from date (month, day, day of week, day of week sin, etc.), which is designed to capture long-range dependencies in the sequential data like daily ridership influenced by historical weather patterns and recurring holidays effects over time. To further enhance the predictive power and optimize the model performance, we performed cross validations over a grid of parameters, such as alpha for regularized regression, learning rates, number of trees, maximum depth for tree-based models, and hidden layers, epochs for RNN-LSTM. Following the performance gain with temporal features in our RNN-LSTM model, we subsequently explored incorporating these temporal features with our previously trained regression and tree-based models.

For the visualization component of the project, we developed an interactive web application that leverages three primary data exploration techniques to create a comprehensive and user-friendly

experience for understanding the data. The frontend of the application is built using React, while a Flask API is employed to integrate our trained models into the user interface. The application is hosted on Heroku, an affordable cloud platform chosen for its ease of deployment and cost-effectiveness.

The web application is organized into six tabs, divided evenly between Project Information and Interactive Visualizations. Within the Project Information section, the Home tab presents an overview of the project, key findings, and team information; the Data tab provides details about the datasets and the preprocessing steps undertaken to prepare it for visualization and modeling; and the Methodology tab outlines the team's approaches, innovations, and conclusions.

The Interactive Visuals section consists of three additional tabs. The D3 tab (Fig 6) features dynamic visualizations that highlight key insights into the relationship between weather conditions and subway ridership. The Tableau tab (Fig 7) contains a complementary dashboard that narrates the story of the data and allows users to explore it further. Finally, the Prediction tab (Fig 8) offers an interactive interface through which users can input feature values into several random forest models to obtain predicted ridership outcomes for various feature combinations.

In the Tableau section, an interactive interface is provided to facilitate in-depth exploration of the ridership dataset. The report's landing page is organized into three tabs: Historical, Weather, and Holiday & Weekend. The Historical tab presents ridership trends through a line chart, offering drill-down capabilities at the hourly, daily, weekly, monthly, and aggregated levels. This visualization is accompanied by two filters, enabling users to examine ridership by year and during holiday periods. The Weather tab provides a comprehensive analysis of the influence of weather variables on ridership. Users can investigate the effects of rainfall, temperature deviation (i.e., the range of temperatures in degrees Fahrenheit), and maximum average temperature on average ridership across a single year or multiple years. These relationships are visualized using scatterplots, bar charts, and line charts. The final tab, Holiday & Weekend, assesses the impact of weekends on average ridership for a selected year. This analysis is presented through a variety of visual formats, including scatterplots, bar charts, line charts, and circle plots. Collectively, these visualizations enable users to explore how temporal, climatic, and calendar-based factors influence average ridership patterns across the MTA subway system.

5. Evaluation

We began with a time series analysis to identify trends and seasonal patterns in the ridership data. The decomposition revealed a subtle upward trend in ridership, supported by linear regression and confirmed by Kendall's tau, which showed a moderate positive correlation between time and ridership. It also identified a strong weekly seasonal pattern (Fig 9), with Sundays consistently having lower ridership than weekdays. Autocorrelation analysis reinforced this with a significant 7-day lag, highlighting the impact of weekly cycles on ridership. We also found that residuals were

skewed, especially during holidays, suggesting anomalies that are challenging to model with standard time series techniques (Fig 10). Despite these issues, the time series analysis offered valuable insights, guiding the development of our machine learning approach.

Within the context of daily ridership-predictive modeling, evaluation will be in the form of measuring the error rate (RMSE) and the proportion of variance explained by our predictive models (R-squared). In terms of experimental design, we divided the data into training and testing sets, using 70% of the data for model training and the remaining 30% for testing using scikit learn library. We standardized the data using the scikit-learn StandardScaler method before training our models. To ensure robust performance and reduce overfitting, we implemented n - fold cross validation (CV) during the model training. For each model, we divided the dataset into n folds, iteratively trained it on $n-1$ folds and validated them on remaining folds for better measure of model generalizability. We trained several machine learning models and compared their test performance based on key metrics like mean squared error (MSE), Root mean squared error (RMSE) and R-squared values. Finally, we picked the model that performed the best in terms of maximizing the R-squared values while minimizing the errors across various weather and holiday conditions (Fig 11). Furthermore, we analyzed the feature importance for the features to understand which factors contribute the most and how they relate to the ridership (Fig 4, Fig 6).

The preliminary exploratory analysis revealed several notable patterns. Specifically, heavy rainfall was associated with a significant 10% decrease in ridership compared to days without precipitation, whereas moderate and light rainfall were linked to smaller declines. Ridership levels were found to be highest on weekdays, particularly from Tuesday to Thursday, while weekend ridership experienced a marked decline, with the most substantial drop occurring on Sundays. These observations were corroborated by initial modeling results, which indicated that weekday variables exerted the most substantial positive influence on daily ridership, whereas severe weather conditions and precipitation levels had a negative impact. These findings offer valuable insights into ridership dynamics and serve to guide and refine subsequent model development.

Among the models evaluated, tree-based algorithms incorporating weather, holiday, and temporal features demonstrated strong performance. The Random Forest model outperformed others, achieving the highest R-squared value of 74.8% and the lowest Root Mean Squared Error (RMSE) of 398,865. This indicates that the model was able to explain 74.8% of the variability in daily New York City subway ridership based on the selected independent variables—an impressive result given the complexity of real-world data and the limited set of weather and holiday-related features. Given that the average daily ridership is approximately 2.96 million, the RMSE represents roughly 13.4% of the mean daily ridership. The optimal Random Forest model was developed using fine-tuned hyperparameters: max_depth set to 20, min_samples_leaf to 2, min_samples_split to 2, and n_estimators to 100. Its performance was rigorously validated through 3-fold cross-validation, ensuring robust and reliable results. A summary of the top-performing models and their corresponding evaluation metrics is provided in Figure 11.

6. Conclusions and Discussion

Our analysis of daily subway ridership highlights several important insights. Based on RMSE and R-squared values, the random forest model—incorporating temporal, weather, and holiday data—emerged as the best-performing model (Fig 11). Notably, the inclusion of both weather and holiday data improved predictive accuracy compared to a purely autoregressive time series model, underscoring the value of exogenous variables in ridership forecasting.

The time series analysis revealed a strong weekly seasonality, with ridership peaking on Tuesdays and Wednesdays and dropping significantly on weekends, particularly Sundays. Major holidays, such as Christmas and Thanksgiving, had a pronounced negative effect on ridership. Weather conditions also played a measurable role: higher temperatures, shorter daylight duration, increased precipitation, higher snow fall, and stronger winds were generally associated with reduced ridership (Fig 4, Fig 5).

While our model demonstrates strong predictive performance, several limitations present opportunities for future research. First, our analysis operates at a daily level, and refining the granularity—such as by hourly ridership or individual station data—could provide more nuanced insights for transit operations. Additionally, while expanding the dataset beyond the 2020–2024 period might improve model performance by providing more training data, the COVID-19 pandemic complicates this approach, as its extreme impact on ridership could distort long-term trends. Future studies could explore segmenting pandemic-era data or applying anomaly detection techniques to mitigate this issue.

Another limitation lies in the variables we incorporated; while weather and holidays proved significant, other factors—such as large events (concerts, sporting games), crime rates, or macroeconomic indicators (e.g., unemployment, gas prices)—could further enhance predictive power. However, collecting such data systematically remains a challenge, particularly for decentralized events. Finally, testing the generalizability of our model across different transit systems or cities with varying weather patterns, cultural norms, or urban layouts could strengthen its broader applicability. Addressing these limitations in future work would not only improve forecasting accuracy but also support more adaptive and data-driven public transit planning.

Our results have practical implications for transit planning and demand forecasting. By accounting for weather, holidays, and weekly trends, agencies can better anticipate fluctuations in ridership and allocate resources efficiently. Furthermore, the success of the random forest model suggests that ensemble machine learning methods are well-suited for transportation demand modeling, especially when nonlinear relationships (e.g., weather effects) are present. Future extensions of this work could refine predictive accuracy by addressing the limitations above, ultimately supporting more resilient and data-driven public transit systems.

All team members have contributed a similar amount of effort

Appendix

Figure 1: Correlation Heat Map

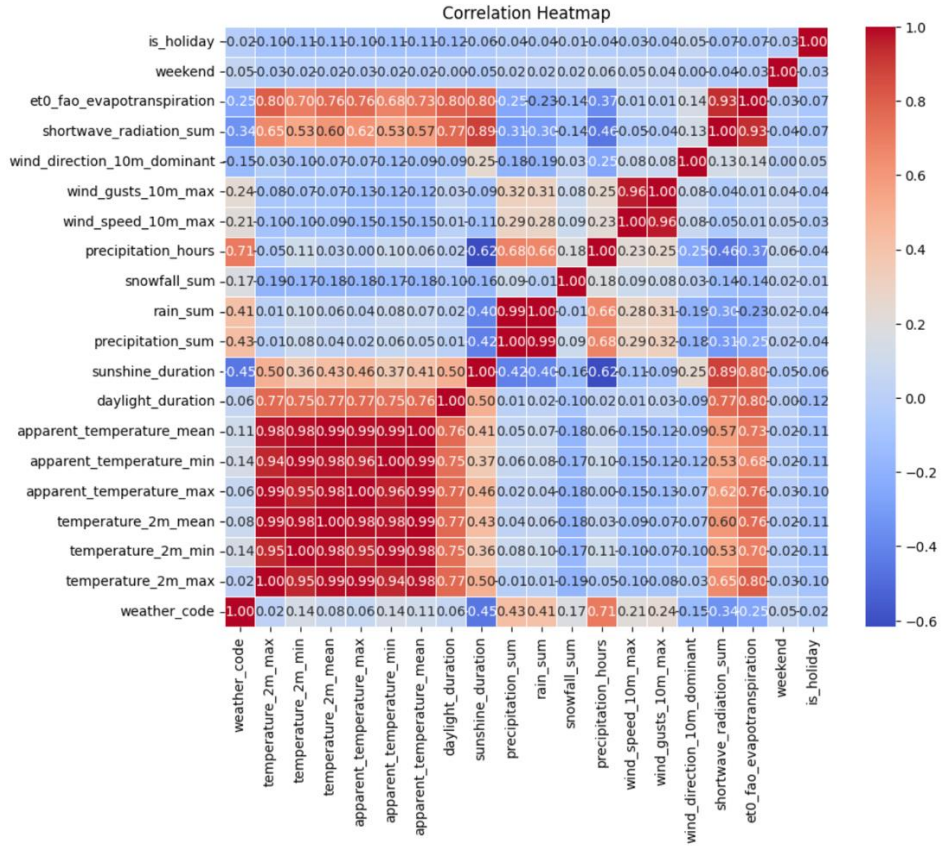


Figure 2: Box Plot for Categorical variables vs Ridership

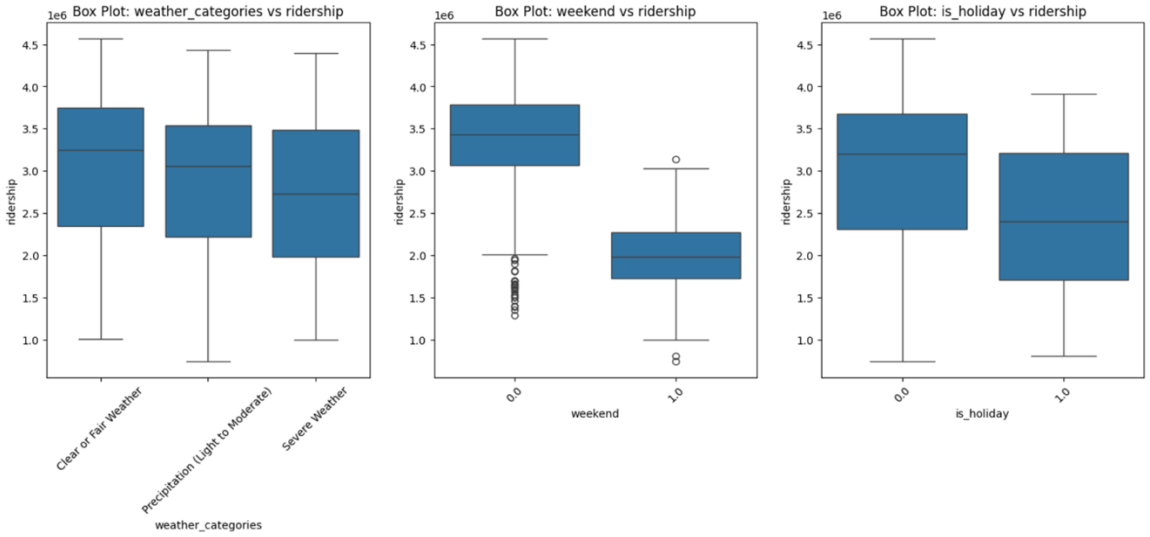


Figure 3: Ridership vs Temperature (Range, Min, Max, Mean) Scatter plot

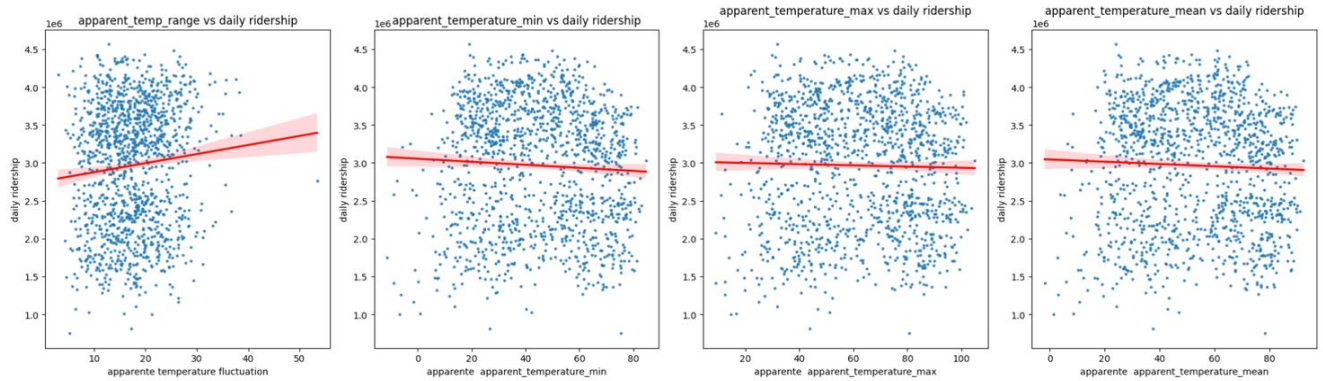


Figure 4: Feature Importance from Random Forest model

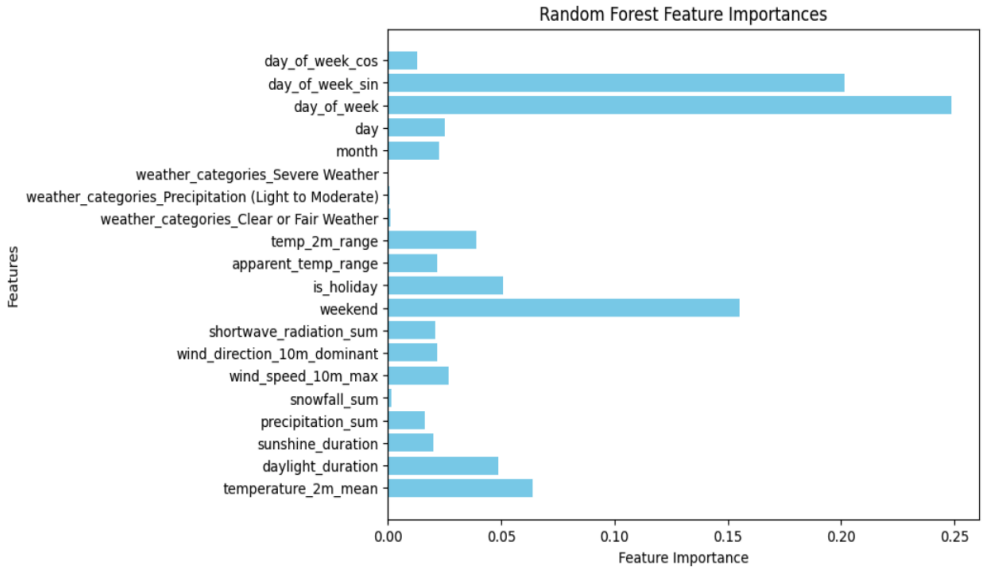


Figure 5: SHAP Features Impact plot for Random Forest model

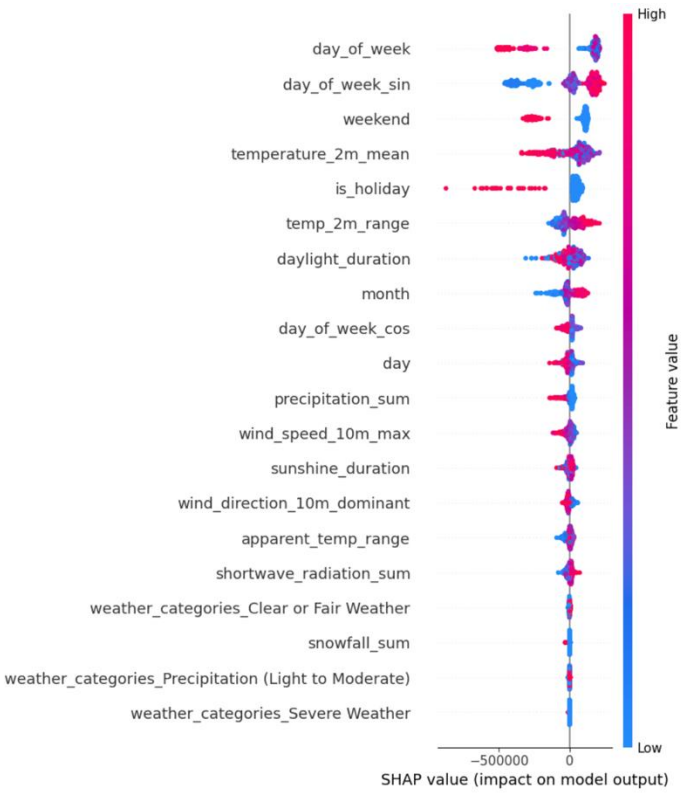


Figure 6: D3

Figure 7: Tableau

Figure 8: Prediction

Figure 9: Decomposition, Seasonality Component Zoomed

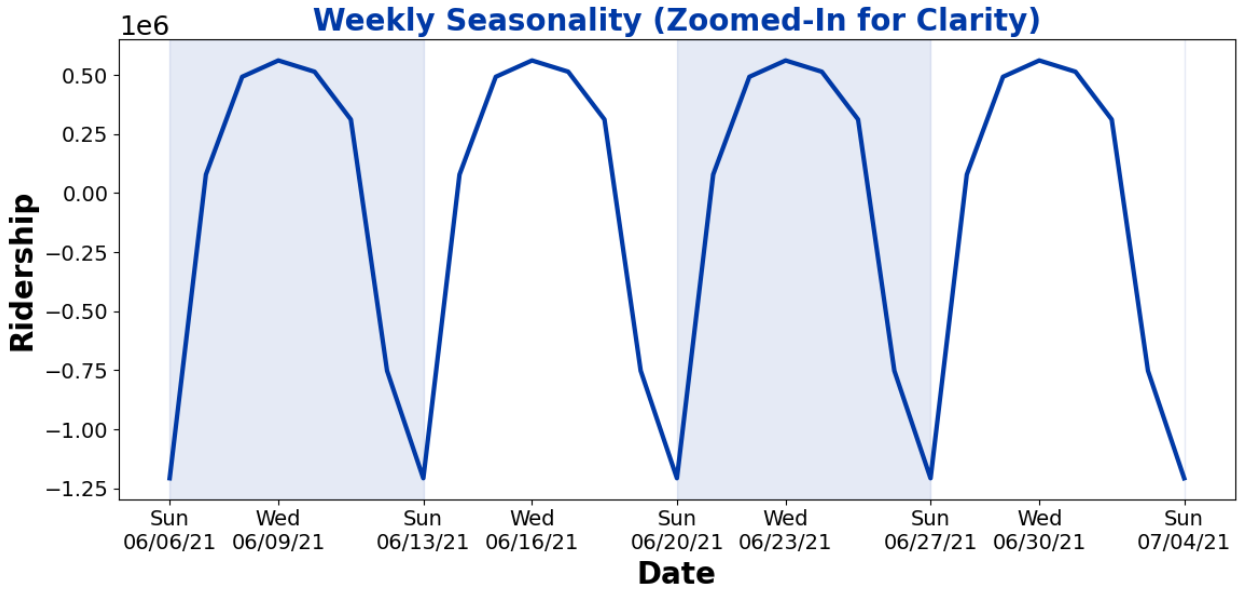
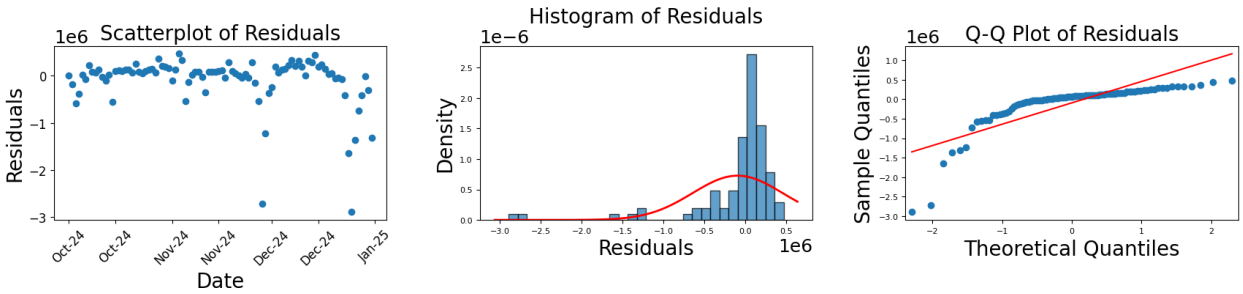


Figure 10: Residuals of Holt-Winters Exponential Smoothing Model



Note the unhealthy residuals showing deviation from theoretical normal distribution in red in both the histogram and Q-Q plot. The largest residuals coincide with major holidays.

Figure 11: Models and evaluation metrics

Model Name	RMSE	R-squared	Best Params
Random Forest + temporal_features	398,866	0.748	{'max_depth': 20, 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 100}
Gradient Boosting + temporal_features	405,769	0.740	{'learning_rate': 0.2, 'max_depth': 4, 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 50}
CATBoost Regressor + temporal_features	409,241	0.735	{'iterations': 500, 'learning_rate': 0.05, 'depth': 6, 'loss_function': 'RMSE', 'verbose': 100}
RNN LSTM	429,261	0.729	{'seq_length': 30, 'hidden_size': 128, 'num_layer': 2, 'output_size': 1, 'batch_size': 32, 'n_epochs': 100}
Decision Tree + temporal_features	431,543	0.705	{'ccp_alpha': 0.01, 'max_depth': 5, 'min_samples_leaf': 4, 'min_samples_split': 10}
Random Forest	443,944	0.688	{'max_depth': 10, 'min_samples_leaf': 6, 'min_samples_split': 2, 'n_estimators': 50}
Gradient Boosting	449,038	0.681	{'learning_rate': 0.1, 'max_depth': 3, 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 50}
CATBoost Regressor	457,475	0.675	{'iterations': 500, 'learning_rate': 0.05, 'depth': 6, 'loss_function': 'RMSE', 'verbose': 100}
XGBoost Regressor	469,653	0.658	{'colsample_bytree': 1.0, 'learning_rate': 0.01, 'max_depth': 4, 'min_child_weight': 10, 'n_estimators': 500, 'subsample': 0.6}
Ridge + temporal_features	482,812	0.631	{'alpha': 10}
ElasticNet + temporal_features	482,848	0.631	{'alpha': 0.1, 'l1_ratio': 0.9}
Lasso + temporal_features	483,417	0.630	{'alpha': 100}
PCA RanfomForest	483,999	0.629	n_components = 9
Decision Tree	488,662	0.622	{'ccp_alpha': 0.01, 'max_depth': 5, 'min_samples_leaf': 2, 'min_samples_split': 10}
PCA Regression	513,923	0.582	n_components = 9
ElasticNet	514,083	0.582	{'alpha': 0.01, 'l1_ratio': 0.5}
Ridge	515,171	0.580	{'alpha': 1}
Linear Regression + temporal_features	515,371	0.580	
Lasso	515,499	0.580	{'alpha': 100}
Linear Regression	515,577	0.580	
Holt-Winters Exponential Smoothing	554,598	n/a	{'trend': None, 'seasonality':'multiplicative'}
SARIMA	617,654	n/a	{'order': (3,1,2), 'seasonal_order': (2,0,2,7)}

References

1. Arana, P., Cabezudo, S., & Peñalba, M. (2014). Influence of weather conditions on transit ridership: A statistical study using data from smartcards. *Transportation Research Part A: Policy and Practice, 59*, 1–12. https://doi.org/10.1016/j.tra.2013.10.019

2. Brazil, W., White, A., Nogal, M., Caulfield, B., O'Connor, A., & Morton, C. (2017). Weather and rail delays: Analysis of metropolitan rail in Dublin. *Journal of Transport Geography, 59*, 69–76. https://doi.org/10.1016/j.jtrangeo.2017.01.008

3. Cui, J., & Liu, Z. (2023). Research on the influence of weather factors on urban rail transit passenger flow. *Transactions on Computer Science and Intelligent Systems Research, 1*, AIEA 2023.

4. Ding, C., Duan, J., Zhang, Y., Wu, X., & Yu, G. (2018). Using an ARIMA-GARCH modeling approach to improve subway short-term ridership forecasting accounting for dynamic volatility. *IEEE Transactions on Intelligent Transportation Systems, 19*(4), 1054–1064. https://doi.org/10.1109/TITS.2017.2711046

5. Ding, C., Wang, D., Ma, X., & Li, H. (2016). Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees. *Sustainability, 8*(11), 1100. https://doi.org/10.3390/su8111100

6. Ettema, D., Friman, M., Olsson, L. E., & Gärling, T. (2017). Season and weather effects on travel-related mood and travel satisfaction. *Frontiers in Psychology, 8*, Article 140. https://doi.org/10.3389/fpsyg.2017.00140

7. Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., & Thépaut, J.-N. (2023). *ERA5 hourly data on single levels from 1940 to present* [Data set]. ECMWF. https://doi.org/10.24381/cds.adbb2d47

8. Jiang, S., & Cai, C. (2023). The impacts of weather conditions on metro ridership: An empirical study from three mega cities in China. *Travel Behaviour and Society, 31*, 166–177. https://doi.org/10.1016/j.tbs.2022.12.003

9. Kim, K. (2020). Effects of weather and calendar events on mode-choice behaviors for public transportation. *Journal of Transportation Engineering, Part A: Systems, 146*(7), 04020054. https://doi.org/10.1061/JTEPBS.0000371

10. Lepage, S., & Morency, C. (2021). Impact of weather, activities, and service disruptions on transportation demand. *Transportation Research Record, 2675*(1), 294–304. https://doi.org/10.1177/0361198120966326

11. Metropolitan Transportation Authority. (2023). *Subway and bus ridership for 2023*. https://www.mta.info/agency/new-york-city-transit/subway-bus-ridership-2023

12. Muñoz Sabater, J. (2019). *ERA5-Land hourly data from 2001 to present* [Data set]. ECMWF. https://doi.org/10.24381/CDS.E2161BAC

13. New York City Department of City Planning. (2023). *Population estimates for New York City and boroughs, vintage 2023*. https://www.nyc.gov/assets/planning/download/pdf/planning-level/nyc-population/population-estimates/current-population-estimates-2023-June2024-release.pdf?r=1

14. Ngo, N. S., & Bashar, S. (2024). The impacts of extreme weather events on U.S. public transit ridership. *Transportation Research Part D: Transport and Environment, 137*, 104504. https://doi.org/10.1016/j.trd.2024.104504

15. Schimanke, S., Ridal, M., Le Moigne, P., Berggren, L., Undén, P., Randriamampianina, R., Andrea, U., Bazile, E., Bertelsen, A., Brousseau, P., Dahlgren, P., Edvinsson, L., El Said, A., Glinton, M., Hopsch, S., Isaksson, L., Mladek, R., Olsson, E., Verrelle, A., & Wang, Z. Q. (2021). *CERRA sub-daily regional reanalysis data for Europe on single levels from 1984 to present* [Data set]. ECMWF. https://doi.org/10.24381/CDS.622A565A

16. Singhal, A., Kamga, C., & Yazici, A. (2014). Impact of weather on urban transit ridership. *Transportation Research Part A: Policy and Practice, 69*, 379–391. https://doi.org/10.1016/j.tra.2014.09.008

18. Tang, J., Zuo, A., Liu, J., et al. (2022). Seasonal decomposition and combination model for short-term forecasting of subway ridership. *International Journal of Machine Learning and Cybernetics, 13*, 145–162. https://doi.org/10.1007/s13042-021-01377-7

19. Vitello, P., Fiandrino, C., Connors, R. D., et al. (2024). TransitCrowd: Estimating subway stations demand with mobile crowdsensing data. *Data Science for Transport, 6*, Article 6. https://doi.org/10.1007/s42421-024-00091-4

20. Zippenfenig, P. (2023). *Open-Meteo.com Weather API* [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.7970649