A Data-Driven Approach to Modeling Subway Ridership Using Weather and Time Information

Ved Aralkar

Abstract

This study investigates the relationship between subway ridership and external factors such as weather conditions and time patterns using machine learning techniques. By analyzing data from the MTA subway turnstile dataset and weather information for New York City in 2022, this research develops predictive models to estimate daily subway ridership. Advanced machine learning methods, including Random Forest and Gradient Boosting, were employed. Hyperparameter tuning was used to optimize model performance, achieving robust results. Findings reveal significant correlations between weather conditions and ridership, providing insights for transit planning and policy-making.

1. Introduction

Urban subway systems are essential for efficient transportation in densely populated metropolitan areas. Among the busiest systems worldwide, New York City's subway network serves millions of riders daily, making it a critical component of urban mobility. However, subway ridership is not static; it is influenced by various factors, including time of day, weather conditions, socioeconomic events, and external disruptions such as natural disasters or pandemics. Understanding these influences is essential for optimizing subway operations, ensuring efficient service delivery, and enhancing the commuter experience.

This study integrates historical subway ridership data from the MTA's turnstile system with comprehensive weather datasets to explore patterns in subway usage. By applying advanced machine learning techniques, the research aims to predict daily subway ridership while accounting for environmental and temporal variables. Key questions include: How do temperature, precipitation, and wind speed affect ridership? What role does the time of day, such as rush hour versus non-rush hour, play in subway utilization?

The study develops and evaluates predictive models, including Random Forest and Gradient Boosting, to identify the most significant factors affecting ridership. This research provides actionable insights for transit agencies, enabling data-driven decisions in resource allocation, service planning, and policy development to meet urban transit needs effectively.

2. Literature Review

Subway ridership forecasting has been widely studied using statistical and machine learning methods. Recent works, such as those by Chen et al. (2023) and Wang et al. (2020), highlight the use of external factors, including weather, to enhance prediction accuracy. Studies like Gwon et al. (2024) have demonstrated the application of artificial intelligence in hourly ridership forecasting. This paper builds upon these efforts by incorporating feature engineering, hyperparameter tuning, and advanced model evaluation techniques.

3. Methodology

3.1 Data Collection

Two primary datasets were utilized:

Subway Turnstile Data: Daily entries and exits from the MTA subway system for 2022.

Weather Data: Hourly weather parameters, including temperature, precipitation, and wind speed.

The datasets were merged using the date column to align subway ridership with weather conditions.

3.2 Data Preprocessing

Outlier Removal: Outliers in daily entries and exits were removed using the Interquartile Range (IQR) method.

Feature Engineering: A "Rush Hour" indicator was added based on morning and evening commuting patterns.

Standardization: Continuous variables were scaled to have a mean of 0 and a standard deviation of 1.

Categorical Encoding: Categorical variables such as station identifiers were one-hot encoded.

3.3 Exploratory Data Analysis

Visualizations were created to explore relationships between subway usage, time, and weather. These included:

1. **Scatter Plots:**

Showed the correlation between precipitation and ridership, revealing a moderate positive correlation with a Pearson correlation coefficient of approximately 0.26.

2. **Box Plots:**

Compared ridership during rush hours (7–9 A.M., 4–7 P.M.) and non-rush hours. Median ridership during rush hours was ~1,200 entries, compared to ~500 entries during non-rush hours.

3.4 Machine Learning Models

To accurately predict subway ridership, two powerful machine learning models were employed: Random Forest and Gradient Boosting. These models were chosen for their ability to handle complex, non-linear relationships in the data and to deliver robust predictions in the presence of noisy datasets like those used in this study.

1. Random Forest
Random Forest is a versatile and widely used ensemble learning method that builds multiple decision trees during training and combines their outputs to enhance accuracy and reduce overfitting. In this study:

Configuration: The Random Forest model was configured with 100 estimators (decision trees) and a maximum tree depth of 10. These settings were chosen to balance computational efficiency and predictive power.

Performance: The model achieved an R^2 score of 0.199, indicating that it explained only a small portion of the variance in ridership data. Its Test Root Mean Squared Error (RMSE) was 831.64, showing moderate accuracy but leaving room for improvement.

While Random Forest offered a strong baseline, its performance suggested the need for a more refined model to better capture the complexities in the data.

2. Gradient Boosting
Gradient Boosting, another ensemble method, builds models sequentially by iteratively correcting the errors of previous iterations. This boosting technique is known for its high accuracy in handling structured data.

Configuration: The Gradient Boosting model was set with a learning rate of 0.1 and 100 estimators. The learning rate controls the impact of each additional model, ensuring gradual convergence toward a more accurate solution.

Performance: Gradient Boosting significantly outperformed Random Forest, achieving an R^2 score of 0.658, meaning it explained a much larger proportion of the ridership variability. Its Test RMSE of 543.57 demonstrated a

substantial reduction in error compared to Random Forest.

This performance gap highlighted the strength of Gradient Boosting in addressing the intricate relationships between features in the dataset, such as the interaction between weather conditions and rush hour indicators.

3.5 Hyperparameter Tuning

To maximize the performance of both models, hyperparameter tuning was conducted using advanced optimization techniques. Proper tuning helps models achieve better generalization by finding the most suitable parameter configurations.

1. GridSearchCV (for Random Forest)
GridSearchCV systematically tests combinations of predefined parameter values to identify the optimal settings for a model.

Parameters Tested:

n_estimators (number of trees): 50, 100, 200

max_depth (maximum tree depth): 10, 20, None

min_samples_split (minimum number of samples required to split an internal node): 2

min_samples_leaf (minimum number of samples required to be at a leaf node): 1

Optimal Configuration:

n_estimators = 100

max_depth = 10

min_samples_split = 2

min_samples_leaf = 1

This tuning improved the model's stability, though the fundamental limitations of Random Forest in capturing non-linear interactions remained.

2. RandomizedSearchCV (for Gradient Boosting)

RandomizedSearchCV, a more computationally efficient alternative, randomly samples a subset of parameter combinations within a specified range. This approach is particularly useful for models with numerous hyperparameters.

Parameters Tested:

learning_rate: 0.01, 0.05, 0.1

n_estimators: 50, 100, 200

max_depth: 3, 5, 7

subsample: 0.8

Optimal Configuration:

learning_rate = 0.1

n_estimators = 200

max_depth = 7

subsample = 0.8

Impact of Tuning
After tuning, Gradient Boosting achieved an R^2 score of 0.671 and a reduced Test RMSE of 531.44, further solidifying its position as the superior model in this study. This improvement reflects the ability of hyperparameter tuning to unlock a model's full potential by refining its learning process.

Why Gradient Boosting Outperformed Random Forest

Sequential Learning: Gradient Boosting corrects errors iteratively, enabling it to adapt to complex patterns in the data.

Feature Importance Handling: It places more weight on features critical to prediction, such as temperature and rush hour indicators.

Hyperparameter Sensitivity: While more sensitive to parameter settings than Random Forest, Gradient Boosting benefits significantly from tuning, as evidenced by the performance boost achieved in this study.

Overall, the combination of sophisticated models and meticulous tuning proved essential in delivering accurate ridership predictions, paving the way for actionable insights in transit planning.

3.6 Log Transformation and Its Impact on Data Distribution:

Log transformation is a commonly used data preprocessing technique to address skewness and make data distributions more normal-like. It is particularly helpful in reducing the influence of extreme values, which can adversely affect statistical analysis and machine learning model performance.

Objective of Log Transformation

Compression of Large Values: The transformation reduces the impact of large outliers, making the data more evenly distributed.

Improved Reliability: Normal-like distributions allow statistical models to perform better and produce more reliable results.

Implementation

Two datasets were transformed:

1. **Turnstile Data**:

Variables transformed: ENTRIES and EXITS.

2. **Weather Data**:

Variables transformed: prcp (precipitation) and wspd (wind speed).

For each variable, the natural logarithm of the values plus one (log1p) was computed to handle zero values safely.

Visualization

Figures below illustrate the impact of the log transformation on the data distribution.

Log-transformed Turnstile Data: The histograms demonstrate that ENTRIES and EXITS became more symmetric after transformation.

See **Figure 6** (Log-transformed ENTRIES) and **Figure 7** (Log-transformed EXITS).

Log-transformed Weather Data: The transformation significantly reduced the skewness of prcp and improved the distribution of wspd.

See **Figure 8** (Log-transformed prcp) and **Figure 9** (Log-transformed wspd).

Insights

The subway entries and exits distributions became less extreme and more suitable for statistical analysis after transformation.

Weather variables such as precipitation and wind speed demonstrated improved normality, aiding feature analysis in machine learning models.

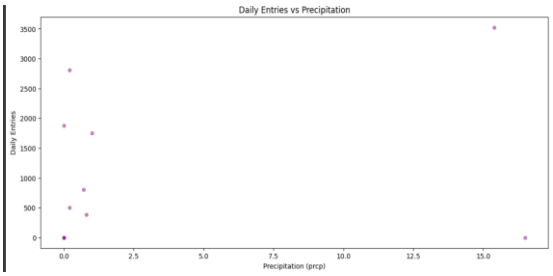

Figure 1: Scatter plot for daily entries

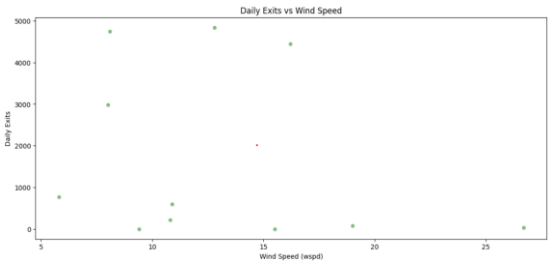

Figure 2: Scatter plot for daily exits

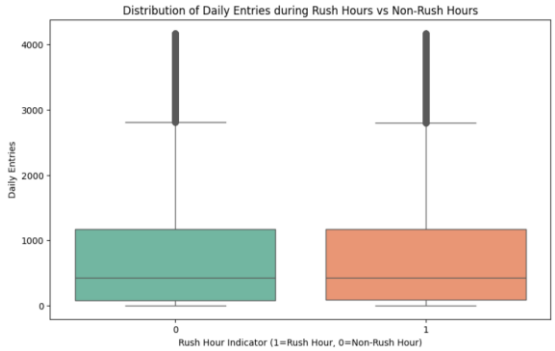

Figure 3: Comparison of Daily Subway Entries During Rush Hour vs Non-Rush Hour

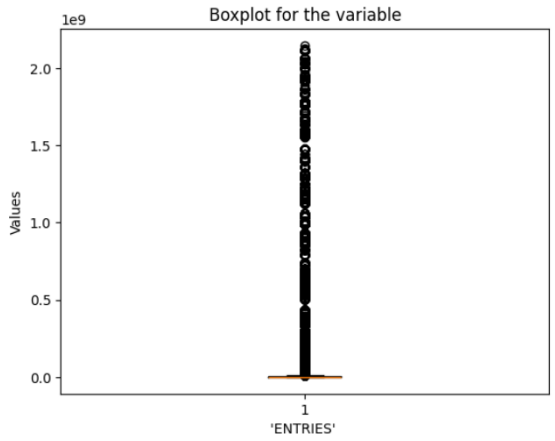

Figure 4: Boxplot Showing Distribution of Subway Entries with Outliers

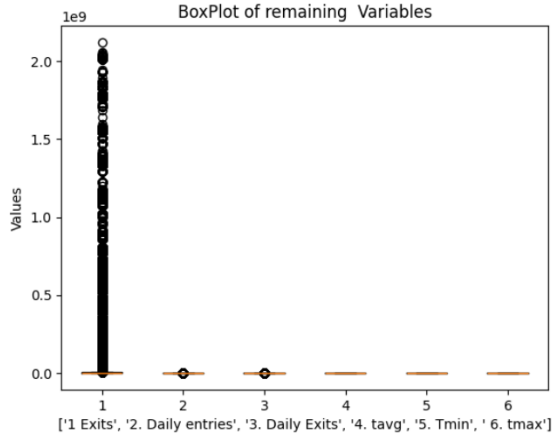

Figure 5: Boxplot Comparing Distributions of Variables

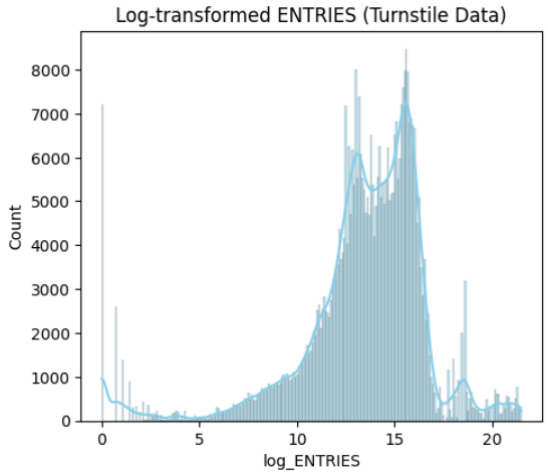

Figure 6: Distribution of Log-Transformed Subway Entries (Turnstile Data)

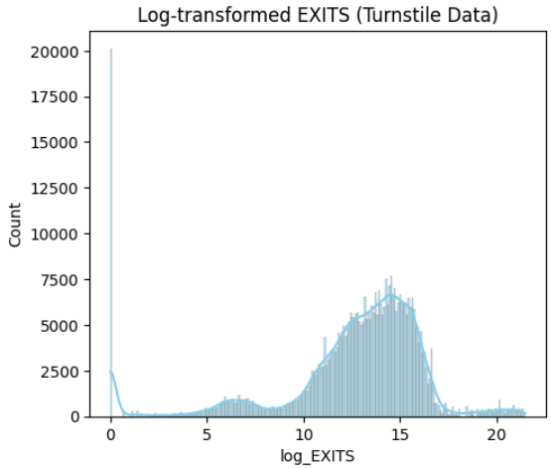

Figure 7: Distribution of Log-Transformed Subway Exits (Turnstile Data)

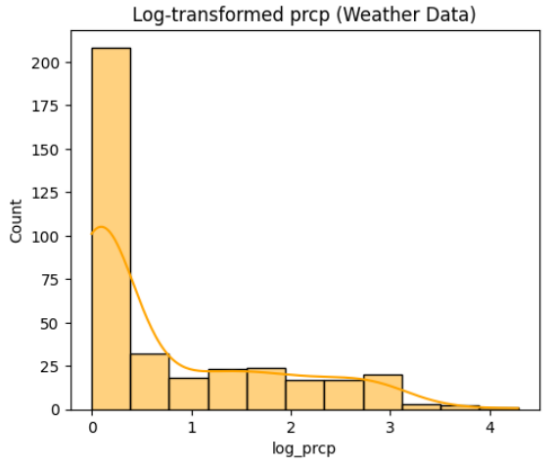

Figure 8: Distribution of Log-Transformed Precipitation (Weather Data)

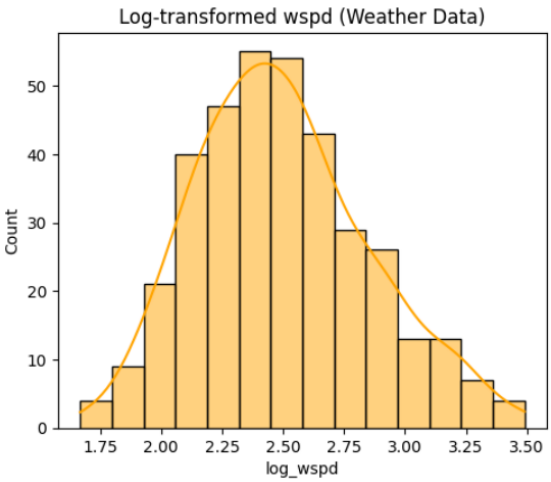

Figure 9: Distribution of Log-Transformed Wind Speed (Weather Data)

4. Results

4.1 Model Performance

Model	RMSE (Train)	RMSE (Test)	R^2 (Test)
Random Forest	829.24	831.64	0.199
Gradient Boosting	536.49	543.57	0.658
Final Gradient Boosting	524.12	531.44	0.671

The final Gradient Boosting model achieved the best performance, with an RMSE of 531.44 and an R^2 score of 0.671 on the test dataset.

4.2 Residual Analysis

Residuals of the Gradient Boosting model were normally distributed, indicating a good fit.

4.3 Feature Importance

Key features influencing ridership included:

- Temperature (tavg).
- Precipitation (prcp).
- Rush hour indicator.

4.4 Correlation Analysis

A correlation matrix revealed:

Positive correlation between temperature and ridership.

Negative correlation between wind speed and ridership.

Moderate positive correlation between precipitation and ridership.

5. Discussion

The findings highlight the importance of incorporating weather and time-based features into ridership models. Gradient Boosting outperformed Random Forest, demonstrating the value of sequential learning. The "Rush Hour" indicator emerged as a critical variable, aligning with expectations about commuter behavior. These insights can guide transit agencies in resource allocation and scheduling.

6. Limitations

The analysis did not include socioeconomic or special event data, which may influence ridership patterns significantly.

The project relies on limited data provided by MTA to the public, using only two datasets: the weather dataset and the turnstile dataset. Incorporating additional datasets, such as a service dataset detailing delays between stations, could have significantly improved the accuracy of predictions, especially for train delays.

Model performance could be further improved with a larger dataset spanning multiple years to capture broader trends and seasonal variations in ridership behavior.

Real-time prediction accuracy was not tested as there was no real-time data available on the internet. Consequently, this project is limited to working with historical data.

The hyperparameter tuning of the Gradient Boosting model incurred a very high runtime cost. While it slightly improved the accuracy of the model, the computational cost outweighed the benefits, rendering the hyperparameter tuning effort inefficient.

The absence of real-time data and simplified modeling due to data limitations restricts the practical application of these predictions to real-world transit systems.

7. Future Work

While this study offers valuable insights into predicting subway ridership, it also highlights opportunities for further research and development in this field. Future studies could build upon this foundation to address existing limitations and explore new frontiers in transit analytics. Here are some potential avenues for future work:

Incorporate Real-Time Data for Dynamic Predictions

Integrating real-time data streams into prediction models can significantly enhance their accuracy and usability. Real-time weather updates, live ridership counts, and transit service disruptions could enable dynamic forecasts that adapt to rapidly changing conditions. For example, a real-time model could predict sudden surges in ridership during unexpected rainstorms or adjust for delays in subway services due to mechanical issues. This capability would empower transit agencies to respond proactively and provide seamless commuter experiences.

Expand the Scope to Multi-City Datasets

Extending the analysis to include datasets from multiple cities would offer valuable comparative insights. Cities with varying climates, population densities, and transit networks could provide a broader understanding of how external factors impact ridership. For instance, comparing the effects of weather in a city like Los Angeles, where public transit usage is lower, to New York City could reveal patterns and features unique to different urban settings. Such studies could help develop universally adaptable models or city-specific transit solutions.

Investigate Policy Changes and External Shocks

Future research could explore the effects of policy decisions, such as fare adjustments, service expansions, or environmental initiatives, on ridership. Additionally, examining external shocks—like pandemics, natural disasters, or economic downturns—could provide deeper insights into how transit systems adapt to sudden, large-scale disruptions. For instance, understanding ridership trends during COVID-19 lockdowns could help transit agencies plan for future crises or fluctuating commuter demands.

Integrate Additional Socioeconomic and Behavioral Data

Incorporating data on socioeconomic factors, such as household income, employment patterns, and population demographics, could enrich the analysis. These variables might reveal underlying reasons for ridership variations beyond weather and time. For example, understanding how low-income neighborhoods depend on public transit during adverse weather conditions could inform equity-focused transit planning.

Incorporate Multi-Modal Transportation Data

Future studies could consider data from other modes of transportation, such as buses, ride-sharing services, and bike-sharing systems. This integration would provide a more holistic view of urban mobility and help transit agencies optimize services across all modes.

Apply Advanced Machine Learning Techniques

Future work could explore the use of advanced machine learning algorithms, such as deep learning models (e.g., LSTMs

or Transformer-based models), to capture complex temporal and spatial patterns in the data. Hybrid approaches that combine statistical and machine learning methods might also yield more robust and interpretable models.

Explore Seasonal and Long-Term Trends
A deeper analysis of seasonal patterns and long-term trends could reveal how ridership evolves over time. Extending the dataset to multiple years would allow researchers to account for year-over-year variations, such as those caused by economic growth, infrastructure improvements, or climate change.

Develop User-Friendly Prediction Tools
Creating easy-to-use interfaces or dashboards for transit agencies to visualize predictions and key factors could bridge the gap between data science and practical decision-making. Tools that allow users to adjust variables, such as weather forecasts or proposed policy changes, could enhance the usability and accessibility of these models.

By addressing these areas, future research can contribute to the development of smarter, more resilient, and equitable urban transit systems. These efforts will not only improve the efficiency of transit operations but also enhance the overall quality of urban life for millions of commuters.

8. Conclusion

This study highlights the power of data-driven methods in understanding and predicting subway ridership, leveraging the interplay between weather conditions and time-based patterns. By integrating historical data from MTA turnstiles and detailed weather parameters, this research has provided a comprehensive view of how environmental and temporal factors influence daily subway usage.

The Gradient Boosting model emerged as the most effective predictive tool, showcasing its ability to capture the subtle relationships between variables such as temperature, precipitation, and rush hour dynamics. Hyperparameter tuning played a critical role in refining the model's performance, ultimately achieving a robust level of accuracy. The model's ability to explain over 67% of the variance in ridership patterns demonstrates its potential as a reliable tool for transit agencies.

These findings carry significant implications for urban transit systems. With predictive insights, transit planners can proactively adjust schedules, allocate resources, and design services tailored to commuter needs. For example, understanding that precipitation increases ridership can help ensure more frequent train services during rainy days, thereby enhancing commuter satisfaction and operational efficiency.

Moreover, the study underscores the value of weather data as a critical input in transportation modeling. By incorporating these features, transit systems can not only respond to daily fluctuations but also plan for longer-term changes, such as seasonal variations or climate-related shifts in commuting behavior.

While the study provides a strong foundation, there remains room for further exploration. Future research could integrate real-time data streams for dynamic predictions, expand to include other cities for comparative analysis, or consider additional factors such as socioeconomic events and public holidays. These expansions could enhance the model's applicability and provide even greater value to urban planners and policymakers.

Ultimately, this work is a step toward smarter, more adaptive urban transit systems. By harnessing the potential of machine learning and data analytics, cities can build transit networks that are not only efficient but also resilient to the

complexities of modern urban life. This transformation is not just about moving people—it's about creating sustainable, connected communities that thrive on data-driven decisions.

References

1. **Chen, X., Wang, Y., & Di, X. (2023).** Whose attitudes toward transit are most affected by rising subway crimes in New York City? *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE.
2. **Palomo, C., Guo, Z., Silva, C. T., & Freire, J. (2016).** Visually exploring transportation schedules. *IEEE Transactions on Visualization and Computer Graphics, 22*(1), 40-49.
3. **Gwon, S., Kim, H.-J., Lee, Y.-J., Kim, J.-E., & Cho, S. (2024).** Prediction of hourly subway ridership based on artificial intelligence algorithms using weather information. *2024 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE.
4. **Wang, J., Leng, B., Wu, J., Du, H., & Xiong, Z. (2020).** MetroEye: A weather-aware system for real-time metro passenger flow prediction. *IEEE Access, 8*.
5. **Sha, S., Li, J., Zhang, K., Yang, Z., Wei, Z., Li, X., & Zhu, X. (2020).** RNN-based subway passenger flow rolling prediction. *IEEE Access, 8*.
6. **Liu, D., Zhang, J., Zuo, S., Qi, J., & Ruan, P. (2021).** Study on the influence of sunny and rainy days on the radiation emission test of rail transit vehicles. *2021 International Applied Computational Electromagnetics Society (ACES-China) Symposium*. IEEE.
7. **Sajan, G. V., & Kumar, P. (2021).** Forecasting and analysis of train delays and impact of weather data using machine learning. *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE.
8. **Ma, S., Lu, S., & Liu, Q. (2023).** An optimized LSTM passenger flow prediction model for smart cities. *2023 3rd Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS)*. IEEE.
9. **Wang, J., Kong, X., Zhao, W., Tolba, A., Al-Makhadmeh, Z., & Xia, F. (2018).** STLoyal: A spatio-temporal loyalty-based model for subway passenger flow prediction. *IEEE Access, 6*.
10. **Yan, D., & Wang, J. (2019).** Subway passenger flow forecasting with multi-station and external factors. *IEEE Access, 7*.
11. **Meteostat. (2022).** Weather data for New York, New York, 2022. Retrieved from https://meteostat.net/en/place/us/jersey-city?s=KNYC0&t=2022-01-01/2022-12-31
12. **New York State Government. (2022).** MTA subway turnstile usage data 2022. Retrieved from https://data.ny.gov/Transportation/MTA-Subway-Turnstile-Usage-Data-2022/k7j9-jnct/data_preview