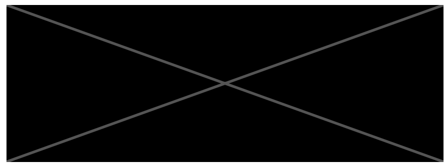
# Is the Yellow Cab Dead? Analyzing Patterns of Taxi Demand in New York using Local Data

STATS 207 Final Project




## Abstract

Accurate prediction of taxi demand is crucial for optimizing city transportation networks, especially in a bustling environment such as New York City. However, this is made challenging because of the complex interplay of factors influencing demand, including commuting patterns, weather conditions, and special events. In this paper, we analyze macro- and micro-level patterns in taxi usage. We find that taxi demand is mostly stationary over time and exhibits periodic behavior on both daily and weekly levels, with deviations for outlier events. We train a SARIMA model to effectively model this consistent structure, incorporating seasonal components for robust short- and long-term predictions. To enhance performance during outlier events, we augment our model with weather and NYC event permit data using a SARIMAX model. By incorporating a linear multiple regressor into our seasonal time series model, our integrated approach improves 1-day forecasting by 4.1% overall and nearly 3x during outlier events such as tropical storms or New Year's celebrations. Relative feature importance analysis via regression coefficients reveals that event-based features are more significant (74% total weight) in predicting demand, likely due to their ability to better model acute fluctuations. Our experiments yield a framework that successfully models inter- and intra-day taxi demand, leveraging diverse data sources to support smarter city planning and better service optimization.


## 1 Introduction

The yellow taxi is as iconic to New York City (NYC) as the Statue of Liberty or the Empire State Building. The vibrant fleet of over 13,500+ cabs is not just a symbol of the city's hustle and bustle but also represents a vital part of its transportation network, shuttling millions of passengers across the five boroughs every year [1]. However, with the rise of ride-share apps such as Uber and Lyft, there are strong incentives to provide more accurate predictions of taxi demand. Improved forecasting allows taxi drivers to dynamically reroute, schedule, and optimize operations, reducing passenger wait times and overall traffic congestion and pollution.

This problem is made especially challenging due to the complex factors involved in taxi demand such as commuting patterns, weather, special events, road work and closures, disruptions in transit services, etc. Furthermore, since the COVID-19 pandemic, conventional traffic patterns have been fundamentally disrupted, especially with the rise of work-from-home. In this project, we aimed to model the key dynamics in NYC taxi demand. Specifically, we answer three primary questions:

1. What are key macro-level and micro-level trends, seasonality, and patterns in taxi usage?
2. What features and models are needed to accurately forecast taxi usage?
3. How can we address outlier moments where taxi usage is different from expected?

We developed a time-series model to understand cyclical demand patterns, and integrated weather and event-data to improve performance during outlier moments. Our experiments yield a framework that successfully models inter- and intra-day taxi demand to help drive smarter city planning and better service optimization.

## 2 Related Work

Previous work has explored different factors that affect taxi trip demand. A 2014 study by Yang et al. fit multiple linear regressions and studied the impact of various demographic factors including population, education, age, income, and employment that affect taxi demand [2]. Interestingly, depending on the time of day, these factors had different degrees of influence on demand. However, the linear models lacked expressivity to model all nuances involved, and a more effective strategy is needed. Faghih et al found success using ARMA models to understand the demand for yellow cabs, subway, Uber, and bicycles throughout Manhattan and outperformed regression-only models by 30% [3]. However, they focused on next-day prediction, ignoring intraday patterns. Safikhani et al found strong performance by incorporating location-based features, using spatio-temporal autoregressive (STAR) models to examine taxi usage patterns across both time and location [4]. A key limitation was the poor performance during outlier events which may cause surges or declines in usage.

Rodrigues et al. took an interesting approach to address this by combining time series data with event text data to train a modified LSTM model with significantly improved forecasting capabilities [5]. Although their results are promising, the analysis was highly specific to taxi demand outside two concert venues that regularly published their events. For a broader city-wide analysis, a more general framework and dataset are needed to incorporate diverse event information. We aim to address the key gaps in the literature, namely providing both intra- and inter-day analysis and incorporating more diverse event data.

## 3 Data

### 3.1 Taxi Trip Data

For the main time-series task, we use data from the NYC Taxi and Limousine Commision (TLC) which contains metadata for each taxi trip taken between 2009 to 2024, including pick-up and drop-off times, start and end locations, trip distances, fares, etc [6]. Given the enormous size of this dataset, for this task, we focus on yellow cab trips in Manhattan in the year 2023 (38.3 million total trips), to be representative of recent trends in taxi demand.

To analyze demand, we preprocessed the dataset by bucketing into 15-minute intervals (to measure intra-day demand) and 1-day intervals (to measure inter-day demand), aggregating the total number of trips within each interval over time. Thus, over 2023, our short-term time series had 35,036 total 15-minute intervals and our long-term time series had 365 1-day intervals.

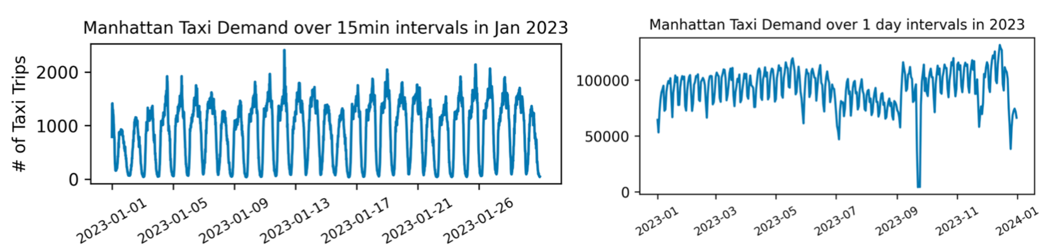



Figure 1: Manhattan taxi demand bucketed into intraday 15-min intervals [left] and interday 1-day intervals [right]. Both exhibit periodic behavior. Note the outlier in September with low demand.

### 3.2 Additional Features

To better model demand on outlier days, we incorporated two more datasets. First is the NYC Permitted Event Information dataset which shows all approved events that will impact streets for at least 5 days for the 1.7 million registered events in 2023 [7]. We preprocessed this dataset by bucketing events into days, and aggregating features such as daily number of event permits, event names, event type, and street closures. Next, we incorporate weather data from the National Oceanic and Atmospheric Administration (NOAA) for the Central Park weather station including average daily temperature, precipitation, snow, windspeed, and atmospheric pressure for every day in 2023 [8]. Due to potential interdependencies, we normalize our data and apply principle component analysis (PCA) to decorrelate the features.

# 4 Methods

## 4.1 Baseline Model

For our baseline model, we trained a seasonal autoregressive integrated moving average (SARIMA) model model [9] directly on the taxi trip data to model the seasonal patterns in taxi demand usage. We first de-trend and normalize the data by fitting a linear model and analyzing residuals. Next, we run sanity and robustness checks, namely the augmented Dickey-Fuller test to verify stationarity [10]. We examine autocorrelation function (ACF) and partial autocorrelation function (PACF) plots to determine parameterizations and verify model assumptions. Then, we fit a SARIMA$(p, d, q) \times (P, D, Q)_s$ model, tuning parameters using a hyperparameter grid search.

$$\Phi(B^s)\phi(B)(1 - B^s)^D(1 - B)^d Y_t = \Theta(B^s)\theta(B)\epsilon_t$$

We chose the model to effectively model observed seasonality patterns. The bounds for the grid search were determined based off ACF/PACF analysis. We evaluate the model on two criterion, Akaike information criterion (AIC) to measure quality of model fit [11], and time series cross-validation to measure forecasting capability [12]. We also use the Ljung-Box statistic to evaluate the final residual autocorrelations.

## 4.2 Augmented Model

To incorporate weather and event data, we trained a SARIMAX model, which fits a SARIMA model with e**X**ogenous features by additionally learning coefficients similar to a multiple regression model [13]. We chose a SARIMAX model, instead of other multivariate time series models, as it was a simple augmentation to our SARIMA baseline and could be parameterized the same, allowing us to directly measure the impact of the additional data. Furthermore, it is computational efficient to train, and the coefficients are easily interpretable. We express our model as SARIMAX$(p, d, q) \times (P, D, Q)_s$

$$\Phi(B^s)\phi(B)(1 - B^s)^D(1 - B)^d Y_t = \Theta(B^s)\theta(B)\epsilon_t + \beta X_t$$

where $\beta$ represents the coefficients from the multiple regression model. We trained the SARIMAX model using the same process as the SARIMA models (detrending, stationarity checks, ACF/PACF analysis, model-fitting hyperparameter grid search).

# 5 Experiments

## 5.1 Intra-Day Demand

We fit a linear model with parameterization $X_t = 0.00042t + 957$, suggesting on average, there are 957 taxi trips every 15 minutes and there is not much fluctuation in this value over time. Analysis of the ACF and detrended data revealed cyclic behavior with a period of 96, corresponding to 1 day of time. It intuitively makes sense since demand patterns for a given time of day are likely correlated with each other.

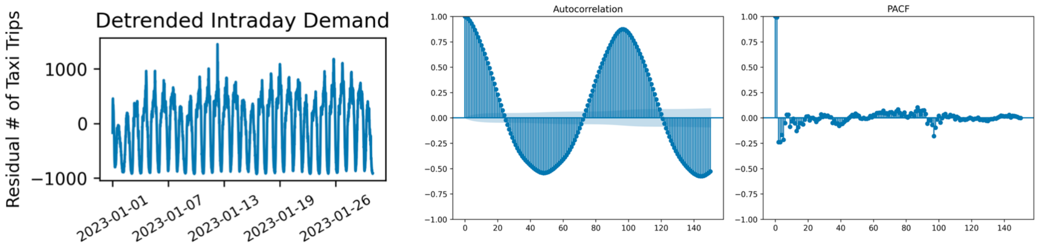



Figure 2: Detrended intraday taxi demand for January 2023. ACF/PACF plots reveal seasonal structure, along with strong AR components

The augmented Dickey-Fuller test on the residuals found strong evidence of stationarity ($p < 0.01$), so combined with verification checks from the ACF/PACF, we proceeded to fit a SARIMA model. We ran a hyperparameter grid search, varying AR/MA/differencing components from 0 to 3, and seasonal AR/MA/differencing components from 0 to 2, trying a total of 128 model parameterizations (five shown below). Note we used a seasonality factor of 96 (1 day), as suggested by the ACF.

| Model | Parameters | AIC | 3hr CV | 1d CV |
|---|---|---|---|---|
| 1 | (1, 1, 1) | 57,444 | 2059 | 182,337 |
| 2 | (3, 0, 3) | 57,214 | 1655 | 65,743 |
| 3 | $(0, 0, 0), (1, 1, 1)_{96}$ | 70,034 | 5060 | 24,761 |
| 4 | $(1, 1, 1), (1, 1, 1)_{96}$ | 55,416 | 1849.1 | 28,677 |
| 5 | $(3, 0, 3), (1, 1, 1)_{96}$ | 49,821 | 1584.3 | 8,092 |

Table 1: Comparison of different SARIMA models with AIC, 3-hr, and 1-day cross-validation mean squared error (CV) metrics. The best performing model is shown in red.

Model 5 performed the best (lowest AIC) with autoregressive and moving average components of order of 3, along with seasonal components. Cross-validation (CV) analysis suggests that seasonal modeling has little impact on seasonal modeling has little impact on 3-hour prediction, with model 5 having mean squared error (MSE) only 4.3% lower than model 2. However, seasonal components have a huge impact on 1 day forecasting, with model 5 having a 88% lower MSE than model 2, which intuitively makes sense given we model our day-level seasonality. Overall, we see a strong seasonal component at a lag of 1 day and a strong ARMA component.

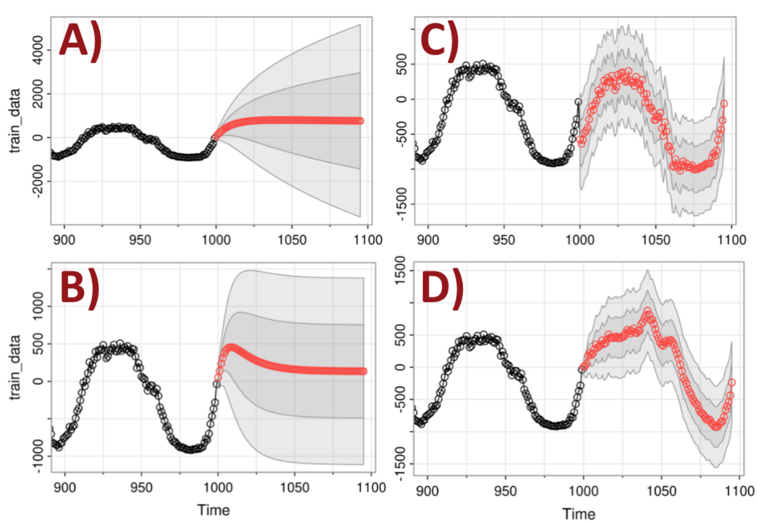



Figure 3: 1-day Short-Term SARIMA forecasts. A) Model 1, B) Model 2, C) Model 3, D) Model 5.

We further analyze the differences across models by looking at 1-day ahead forecasting. Model 1 (A), a simple ARMA model without seasonality, quickly converges to the mean of the dataset and only performs well for very short-term forecasts (few hours into the future). Model 2 (B), a higher order complex ARMA model without seasonality, also converges to the mean, but can produce accurate forecasts for longer than the simple model. By adding seasonality, Model 5 (D), our best performing model with seasonality, is able to do well on long-term forecasts as well by modeling the periodic behavior. In fact, Model 3 (C), with just seasonal components, can get the shape of future observations well, even though it produces poor short-term predictions. This further highlights the importance of the seasonal components.

### 5.2 Inter-Day Demand

We repeat the analyses on the **inter**-day dataset (each data point represents taxi demand aggregated over the day).

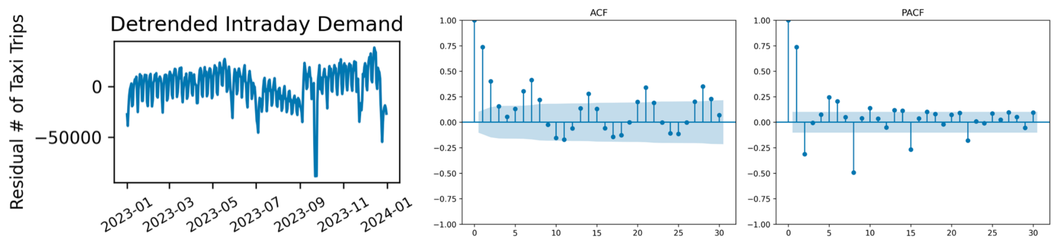



Figure 4: Detrended interday taxi demand for 2023. ACF/PACF plots reveal seasonal structure, along with strong AR components

Analysis of the ACF and detrended data revealed cyclic behavior with a period of 7, corresponding to 1 week of time, which intuitively makes sense since demand patterns for a given day of week are likely correlated with each other. We also see strong AR components. The augmented Dickey-Fuller test on the residuals found strong evidence of stationarity ($p < 0.01$). We ran the same hyperparameter grid search as above, but with a seasonality factor of 7 as suggested by the ACF.

| Model | Parameters | AIC |
|---|---|---|
| 1 | (1, 1, 1) | 7893 |
| 2 | (3, 0, 3) | 7819 |
| 3 | (0, 0, 0), $(1, 1, 1)_7$ | 7909 |
| 4 | (1, 1, 1), $(1, 1, 1)_7$ | 7469 |
| 5 | (3, 0, 3), $(1, 1, 1)_7$ | 7414 |

Table 2: Comparison of different SARIMA models. The best-performing model is shown in red with both high order AR/MA components and seasonality.

Model 5 again performed the best (lowest AIC) with autoregressive and moving average components of order of 3, along with seasonal components. Adding seasonality reduced AIC by 5.2% compared to model 2. It is interesting to note that inter-day analysis, has less variance across the different models than intra-day analysis, likely due to a smaller number of data points and less noisy data due to aggregating over larger buckets. Therefore, SARIMA models show promise in modeling taxi demand in both the short term and long term.

### 5.3 Event and Weather Features

We examined the forecasting performance of the SARIMAX model with weather and event features against the baseline SARIMA model. The SARIMAX model was fit with the same parameterization as model 5, the best-performing SARIMA model for comparison. We evaluate both next day and 5 days ahead performance using time series cross-validation. We focus on inter-day demand because of the limited granularity in our weather and event datasets.

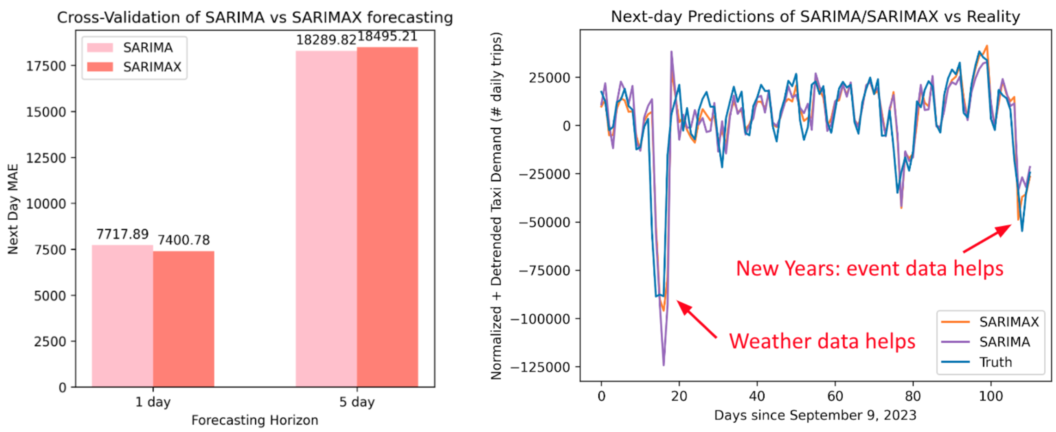



Figure 5: [LEFT] SARIMAX vs SARIMA model in 1-day and 5-day forecasting evaluated with time series cross validation. Event and weather features improve performance in 1-day forecasting, but harm performance in 5-day forecasting. [RIGHT] SARIMA (purple) vs SARIMAX (orange) vs ground truth (blue) predictions on normalized and detrended daily taxi demand data over time. The large downward spike at the start is when Tropical Storm Ophelia hit New York City. The downward spike at the end is the winter holidays.

For 1-day forecasting event and weather features provided a 4.1% reduction in mean absolute error (MAE), suggesting that these features improve overall predictive power. However, for 5-day forecasting, the extra features increased MAE by 1.1%. increase in error suggests that event and weather features provide benefit only for short-term forecasting, which intuitively makes sense since event and weather data is relevant only for the current day.

Next, we compared next-day forecasting results over time for both models. For most of the data, both models perform similarly. However, we see that on black-swan weather events, such as Tropical Storm Ophelia at the end of September, the inclusion of weather features reduces MAE error by almost 4x. Furthermore, we also see that during New Years Eve, when a large number of event permits are approved for celebrations, SARIMAX performs 3.4x better. Therefore, these results suggest that event and weather features can improve model performance during outlier events.

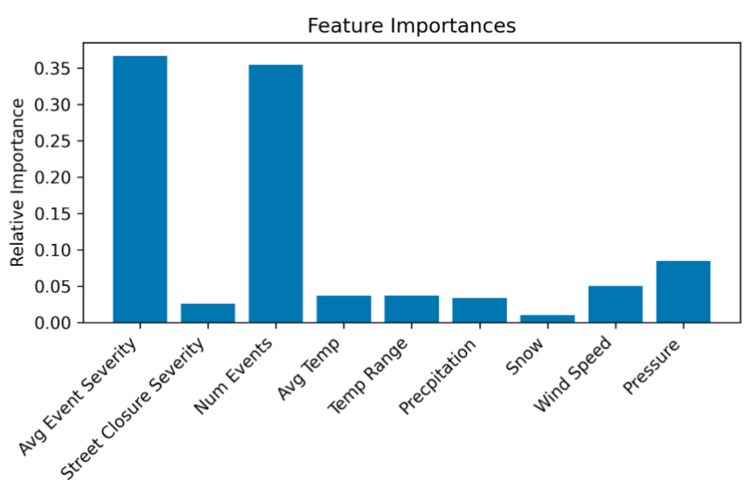



Figure 6: Relative feature importances for exogenous regressors

Lastly, we examined approximate relative feature importances of event and weather features by comparing the normalized magnitudes of regression coefficients. Event-based features have more impact on average, comprising 74% of the exogenous contributions. Interestingly, street closures don't appear to have much impact. Pressure appears to be the most important weather-based feature, but they all contribute roughly equivalently. As a caveat, it is important to note that while regression coefficients are a helpful proxy for approximate feature importance, given that many of these features are correlated with each other, the coefficients may also include interdependencies between the features and may not be representative of the exact contribution of *only* that feature. Regardless, this preliminary analysis is helpful to visualize how exogenous regressors affect the model.

## 6 Conclusion

We were able to develop a modeling approach for New York City taxi demand in Manhattan in 2023. Our analyses show that taxi demand exhibits periodic behavior on both a day-by-day level and week-to-week level and is mostly stationary over time. We saw that this consistent structure can be effectively modeled using SARIMA with higher-order autoregressive and moving average components, along with seasonal components to produce strong short-term and longer-term predictions. However, outlier events, such as tropical storms or New Year's celebrations, can cause large variations. Event and weather-based features improve overall 1-day forecasting by 4.1%, but significantly improve performance (by nearly 3x) on outlier days. Coefficient analysis of SARIMAX regressors revealed event-based features are generally more important than weather-based features, likely because they better model acute fluctuations in taxi demand. Overall, our work underscores the importance of diverse data sources in accurately predicting taxi demand in a complex transportation network.

However, our model is limited in handling outlier events that are not reflected in the weather or NYC event permit dataset. Therefore, we hope to include more diverse datasets to consider more factors that affect demand. We hope to incorporate text features from current events sources such as news headlines, Twitter, or Reddit to better model more nuanced event-based demand. Furthermore, our approach aggregated taxi demand across all of Manhattan, which may not be representative of neighborhood-level variation. Therefore, we hope to conduct further analyses on a smaller geographical scale, incorporating further geospatial features. Regarding the model itself, the underlying multiple regressor in SARIMAX is linear, limiting its expressivity. Therefore, we hope to incorporate more complex multivariate time series models that can more effectively model relationships between these exogenous features such as VARMA, or state-space models such as LSTMs or transformers. We hope that this work will help drive smarter city planning and better service optimization.

## 7 Appendix: Project Codes

Please see all relevant ro ect code at the following Google colab notebook: [redacted]

## References


[1] https://www.nyc.gov/site/tlc/businesses/yellow-cab.page.

[2] Ci Yang and Eric Gonzales. Modeling taxi trip demand by time of day in new york city. In *Transportation Research Record*, 2014.

[3] Sabiheh Faghigh, Arpita Shah, Zehan Wang, Abolfazl Safikhani, and Camille Kamga. Taxi and mobility: Modeling taxi demand using arma and linear regression. *Procedia Computer Science*, 177:186–195, 2020.

[4] Abolfazl Safikhani, Camille Kamga, Sandeep Mudigonda, Sabiheh Sadat Faghih, and Bahman Moghimi. Spatio-temporal modeling of yellow taxi demands in new york city using generalized star models. *International Journal of Forecasting*, 36:1138–1148, 2020.

[5] Filipe Rodrigues, Ioulia Markou, and Francisco Pereira. Combining time-series and textual data for taxi demand prediction in event areas: A deep learning approach. *Information Fusion*, 49:120–129, 2019.

[6] https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

[7] https://data.cityofnewyork.us/city-government/nyc-permitted-event-information-historical/bkfu-528j/about$_{d}ata.$

[8] https://www.ncei.noaa.gov/cdo-web/search?datasetid=ghcnd.

[9] https://en.wikipedia.org/wiki/autoregressive$_{i}ntegrated_{m}oving_{a}verage.$

[10] https://en.wikipedia.org/wiki/dickey

[11] https://en.wikipedia.org/wiki/akaike$_{i}nformation_{c}riterion.$

[12] https://otexts.com/fpp3/tscv.html.

[13] https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.sarimax.html.