# Prediction on taxi hourly pickup count using S-ARIMA and ST-ARIMA

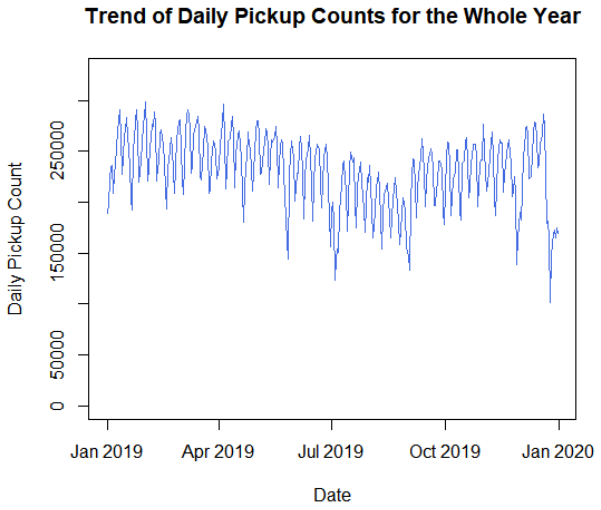
## 1. INTRODUCTION

### 1.1 Background

Understanding taxi demand is crucial for optimizing transportation resources, particularly in cities like New York, a popular tourist destination and business hub in the US. This study aims to investigate the effectiveness of S-ARIMA and STARIMA models in predicting hourly taxi pickup counts in a target taxi zone. By accurately forecasting pickup counts in various taxi zones, drivers can strategize their routes and optimize resource allocation based on anticipated demand.

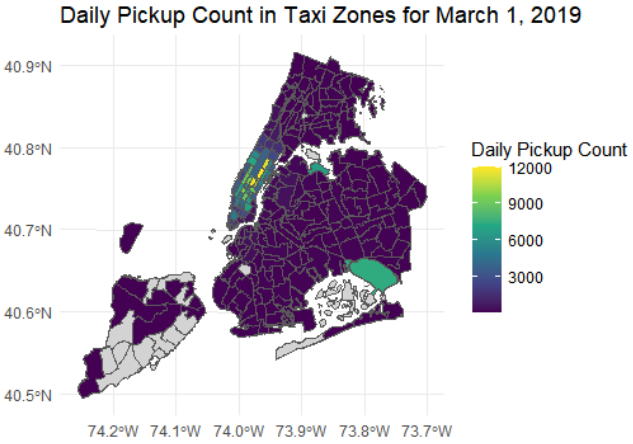
### 1.2 Data

The dataset used in this study was obtained from Kaggle and originates from the NYC Taxi & Limousine Commission. It encompasses taxi trip records spanning from January 2019 to June 2020, featuring information such as pickup and dropoff locations, timestamps, fare amounts, and passenger counts.




For simplicity, the analysis focuses on a single month, with the first 25 days designated for training data and the remaining days for testing. March was chosen due to the absence of national holidays or festivals during weekdays and the lack of discernible trends in daily pickup counts across consecutive months in 2019.

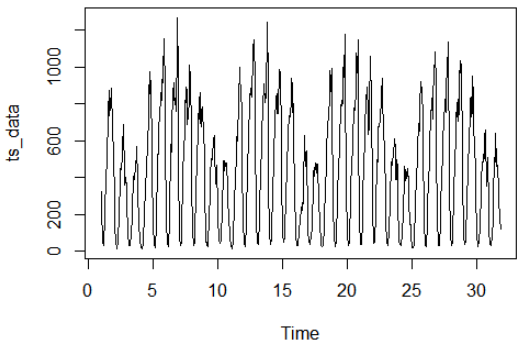
Besides, as observed in the daily pickup count map on March 1st, Manhattan emerges as the focal point of taxi pickups. Therefore, this study focuses exclusively on this borough, allowing for a more targeted analysis of taxi demand patterns within Manhattan.



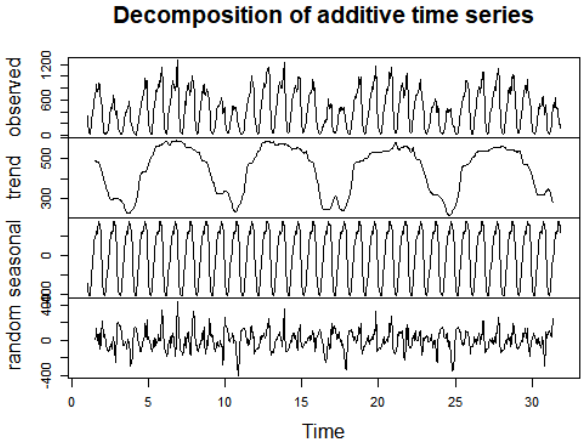

## 2. EXPLORATORY DATA ANALYSIS

### 2.1 Temporal Analysis

The univariate time series of hourly pickup counts in March reveals distinct patterns. There is evident weekly seasonality, with fluctuations observed over each week. Apart from that, the mean and variance seem to remain stable.




Further decomposition of the time series confirms a recurring trend within each week. A cyclical pattern can be observed with daily fluctuations. These observations indicate both short-term and long-term temporal dynamics within March. Therefore, differencing might be necessary to stabilize the series.




To verify stationarity, an Augmented Dickey-Fuller (ADF) test was conducted. The p-value returned was nonsignificant, indicating that the hourly pickup counts within this month can be considered stationary.

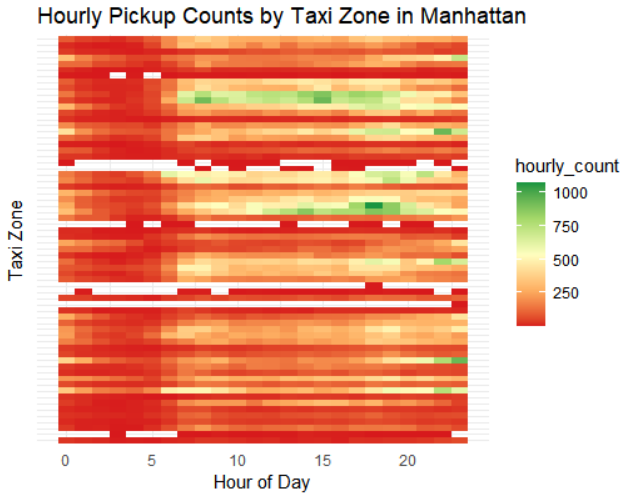
```
In adf.test(ts_data_161) : p-value smal
ler than printed p-value
> print(adf_result)

        Augmented Dickey-Fuller Test

data:  ts_data_161
Dickey-Fuller = -10.228, Lag order = 9,
 p-value = 0.01
alternative hypothesis: stationary
```

## 2.2 Spatial Analysis

I conducted two heatmap visualizations to explore the spatial characteristics of the dataset. The initial heatmap depicted the hourly pickup counts specifically for the first week of March, while the subsequent heatmap extended the analysis to cover the entire month.

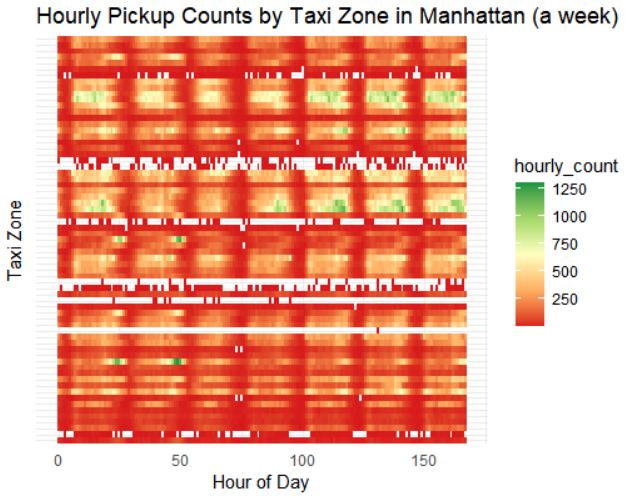
March 2019 commenced on a Friday, and the heatmap illustrations show some notable spatial patterns. Certain zones, as in green, exhibited significantly higher hourly pickup counts compared to others. Situated in the central region of Manhattan, these zones serve as focal points of taxi activity. Conversely, zones located in the suburbs or remote islands recorded minimal pickup counts, depicted as white areas with null values during those times.

Analyzing the daily heatmap, distinct patterns emerged between different types of zones. Some zones, including the Upper East Side North and South, and Midtown Center and East, experienced heightened pickup activity during daytime hours, particularly from 7 AM to 7 PM. These zones, situated in the heart of Manhattan, are recognized as bustling hubs. Conversely, zones like Lower East Side and Lincoln




Square East exhibited increased pickup counts during nighttime hours, spanning from 8 PM to early morning.

Examining the weekly heatmap, it became evident that the observed patterns aligned with specific temporal trends. The first feature, characterized by heightened activity during weekdays, likely reflects commuter behavior. Conversely, the second feature, observed during Friday and Saturday nights, likely corresponds to weekend leisure and entertainment activities.




Based on these insights, Zone 161 (Midtown Center) was selected as the focal point for further analysis. This decision was influenced by its status as one of the busiest taxi zones, as evidenced by the heatmap, and its characteristic representation of the weekday commuting behavior observed in feature one.

# 3. METHODOLOGY

## 3.1 Model

I used ARIMA and STARIMA to model the hourly pickup counts in Zone 161 in March.

ARIMA is a statistical model that combines Auto-Regressive (AR) and Moving Average (MA) and integrates the two. AR captures the autocorrelation while MA captures the error. Together they can make predictions purely based on past temporal data without other assistant features. However, to deal with seasonality in the dataset, I used S-ARIMA, which differences the time series to remove the seasonality and better model the predictions. As seen from the previous explanatory data analysis, S-ARIMA must deal with the daily seasonality in the pickup counts data.

STARIMA is an extended version of ARIMA to deal with spatio-temporal data. It introduces spatiotemporal autocorrelation to the auto-regressive part of ARIMA, instead of treating the modeling target as an isolated object. it is believed to have better performance where spatial features have an influence.

## 3.2 Modelling Process

### *3.2.1 Preprocessing*

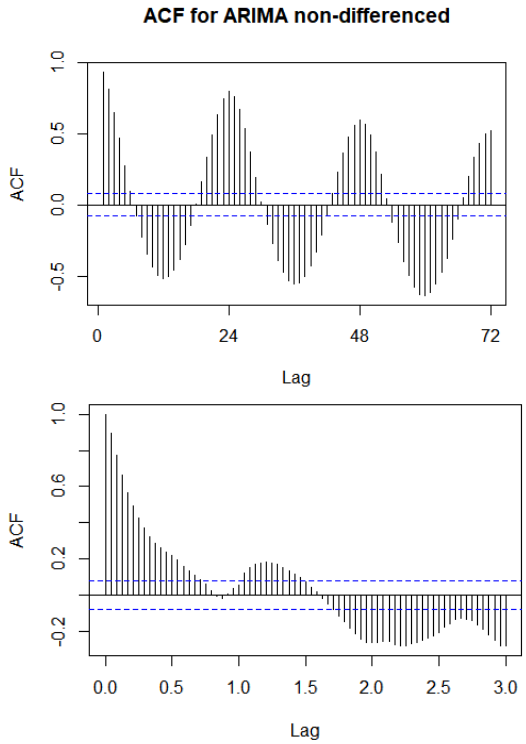
Hourly counts in March: the original dataset was first aggregated into hourly pickup counts, grouped by taxi zone and date. The date data, together with pickup hours in a day, was combined and converted back to the DateTime column. The aggregated dataset was then filtered to include data only in March and NA values were replaced by zeros.

This filtered dataset was used to create a data matrix, consisting of complete data in all taxi zones in Manhattan, and was also filtered to include data only in Zone 161. The double-filtered dataset was then split into two sets of data, one training set from 2019-03-01 00:00:00 to 2019-03-26 23:00:00, and a test set from 2019-03-27 00:00:00 to 2019-03-31 23:00:00. They were then converted into two time series for fitting S-ARIMA model.

For ST-ARIMA, isolated taxi zones were filtered out as the spatial weight matrix function cannot compute weighting for isolated objects, while considering they are distant from the target zone 161 and in more than one spatial order, omitting them does not have much impact on the result. Same as in S-ARIMA, the data matrix was split into train-test sets according to the time range, where the first 623 rows of data (corresponding to the period from March 1 to 26) were selected for training and the rest of the rows for testing.

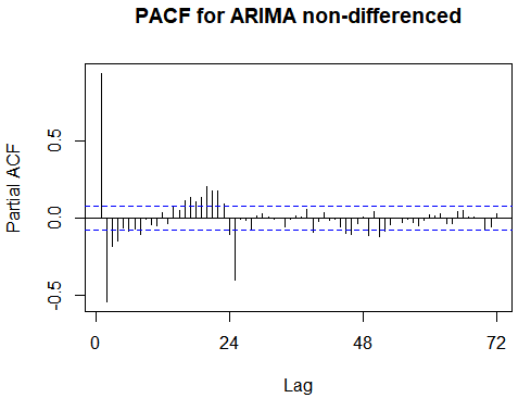
### *3.2.2 S-ARIMA (Self-Selecting)*

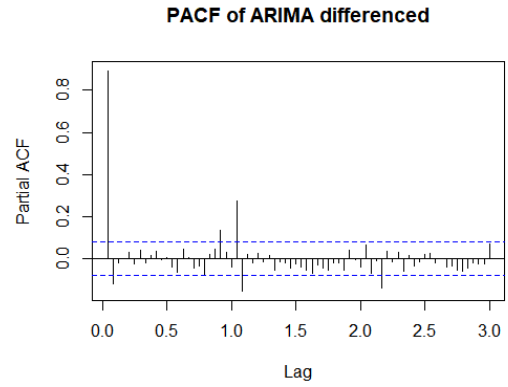
in the S-ARIMA modeling process, I followed the Box-Jenkins Method, which uses ACF and PACF plots to determine the p, q, and d parameters. As shown in the ACF non-differenced plot, there is strong seasonal change with a lag of 24. Therefore, I then differenced the original time series by a lag of 24 to remove the seasonality.



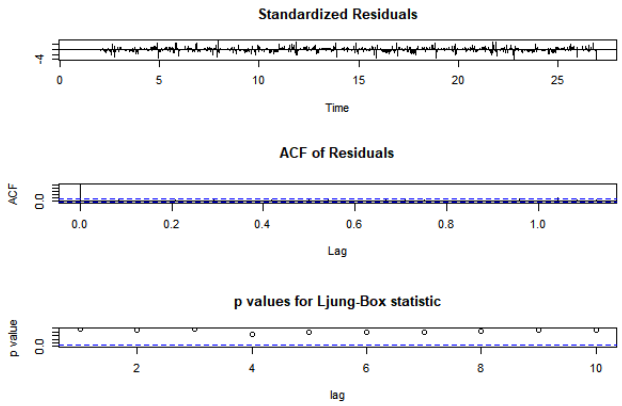

As shown in the differenced ACF, it now loosely follows an exponential trend and decays to zero, which may suggest an AR model, while in the non-differenced ACF plot, it alternates positive and negative. I then used both non-differenced differenced PACF plots to identify the order of AR models. Six candidate models were then fitted and tested to choose the best-performing one. Two metrics AIC and log likelihood were used to evaluate the model fit, as shown in the table. the fourth model with parameters (2,0,0)(1,1,2)[24] won and also passed the diagnostic checking as no autocorrelation was shown in the ACF of residuals

(other than lag 0) and insignificant p values in the Ljung-Box test.




| Model | Parameters | AIC | Log_Likelihood |
|---|---|---|---|
| fit1 | (3,0,0)(0,1,2)[24] | 6,860.708 | -3,424.354 |
| fit2 | (4,0,1)(0,1,2)[24] | 6,863.823 | -3,423.911 |
| fit3 | (4,0,1)(1,1,2)[24] | 6,854.005 | -3,418.002 |
| fit4 | (2,0,0)(1,1,2)[24] | 6,848.935 | -3,418.467 |
| fit5 | (3,0,0)(1,1,2)[24] | 6,850.769 | -3,418.384 |
| fit6 | (2,0,0)(2,1,0)[24] | 6,949.055 | -3,469.528 |

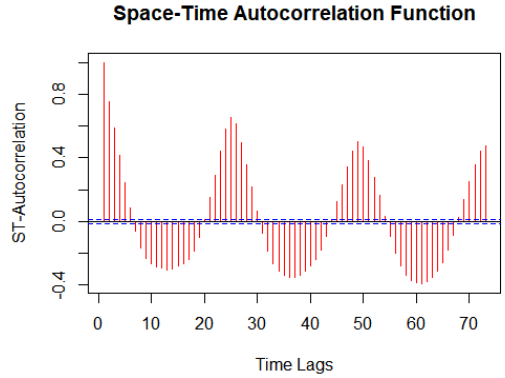


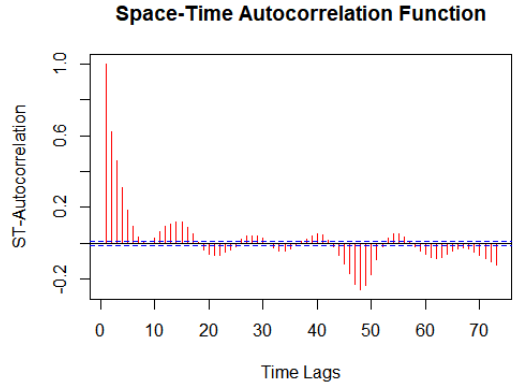

### *3.2.3 S-ARIMA (Auto-Selecting)*

I also fit the model using auto S-ARIMA, which automatically chooses the optimal parameters. It chose a combination of ARIMA(2,0,0)(0,1,1)[24] and also passed the diagnostic test. We will later compare its performance with the other two models in the result section.

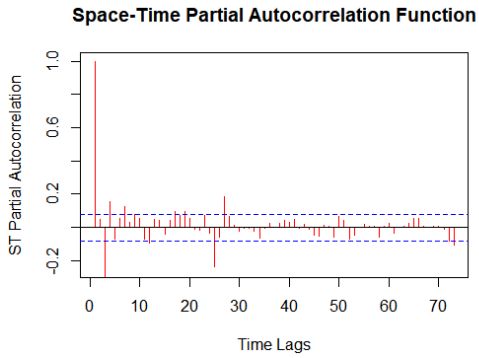
### *3.2.4 ST-ARIMA*

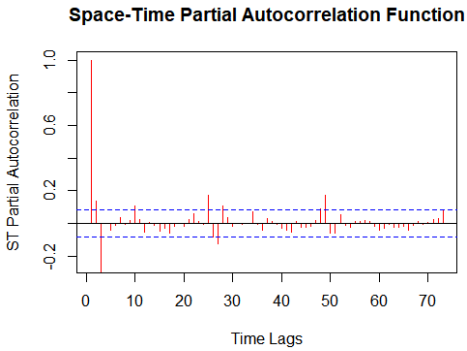
Similar to S-ARIMA, fitting an ST-ARIMA model can implement the Box-Jenkins Method. What is different is calculating the spatial weight matrix, which had been done in the preprocessing section, and incorporating the matrix into both parameter identification and model fitting.




Compared with the one in S-ARIMA, the ACF plot also shows a similar seasonal pattern. A temporal lag of 24 was applied to remove the autocorrelation, with a spatial order of one applied in both non-differenced and differenced series.






A combination of parameters STARIMA(3,1,3)24 was chosen to fit the model. The box test was also passed with a nonsignificant p-value.

```
> Box.test(fit.starima$RES[,6],lag=1,
  type="Ljung")

        Box-Ljung test

data:  fit.starima$RES[, 6]
X-squared = 73.723, df = 1, p-value <
  2.2e-16
```

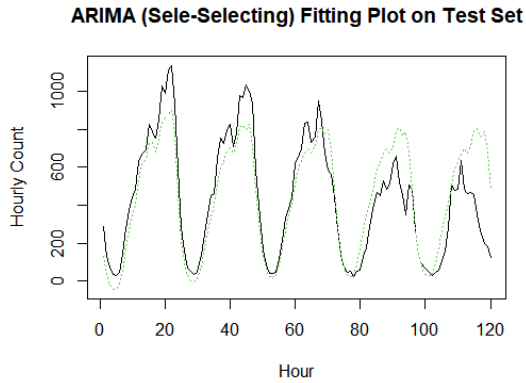
## 4. RESULT

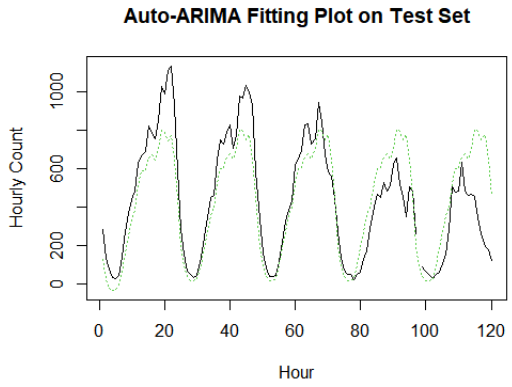
The table below displays how well the three models predicted hourly taxi pickups using data from the last five days of March. We calculated the Normalized RMSE and MAE to measure their performance. The ST-ARIMA model showed the best results, with an RMSE and MAE of 0.057 and 0.040 respectively. In comparison, the other two models had errors nearly three times as high.

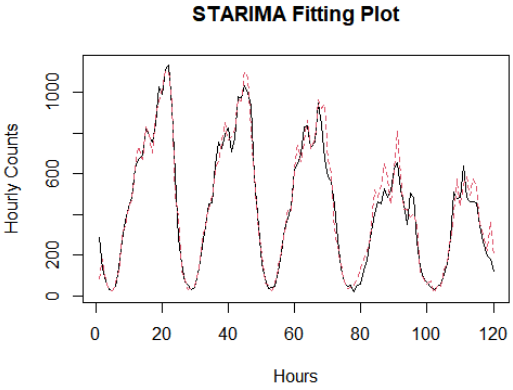
| Model | RMSE | MAE |
|---|---|---|
| S-ARIMA | 0.1416996 | 0.1050976 |
| S-ARIMA.auto | 0.1485491 | 0.1134344 |
| ST-ARIMA | 0.0572794 | 0.0404205 |

The fitting plot also illustrates the prediction results. Initially, the S-ARIMA model performed reasonably well, but it struggled to capture long-term trends in later days as the number of pickups declined. Additionally, it failed to predict the peak values accurately, especially on the first day. Comparing the two S-ARIMA models, the one with manually selected parameters performed slightly better, consistently reporting higher peak values.

On the other hand, STARIMA successfully captured seasonal variations throughout the test period, closely matching the actual data on the first day. However, its accuracy decreased over time, with more significant deviations observed in the last two days.

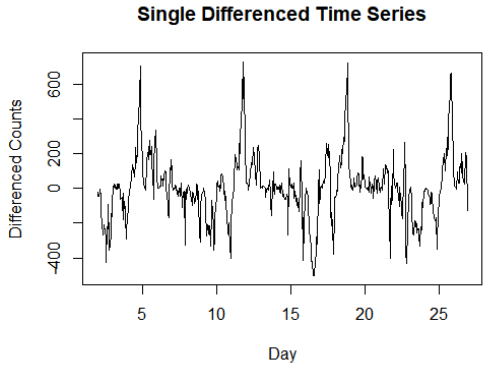



## 5. DISCUSSION & CONCLUSION

In the modelling process of nominating parameters, after undergoing a single differencing




process, the time series plot still indicates some

seasonality persists within the data. Considering a second differencing could be worthwhile to experiment with to remove the weekly cycle.

In comparing the results of the models used in predicting hourly taxi pickups, variations in their performance were observed. The ST-ARIMA model outperformed the other two models, displaying lower errors in terms of Normalized RMSE and MAE. This superior performance can be attributed to the ST-ARIMA model's ability to incorporate both temporal and spatial autocorrelation, which is crucial for accurately capturing the complex patterns present in the data. However, the STARIMA may struggle with computational complexity and increased running time, as I observed in computing PACF and model fitting process.

Due to the limitation of time and computational resources, this study focuses on a single taxi zone and a certain Month. the performance of S-ARIMA and ST-ARIMA is not tested in the remaining study area. More work could be done to explore the variation of prediction accuracy within the whole borough.

Besides, incorporating additional variables such as weather conditions or events could be implemented to enhance the performance of the models. More advanced machine learning techniques such as long short-term memory may offer improved accuracy and flexibility by introducing other feature variables.

## 6. REFERENCES


Commission, N. T. L. (2024). *TLC Trip Record Data*. https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

James, H., & Tao, C. (2024). *Spatio-temporal Analytics in R*. https://moodle.ucl.ac.uk/pluginfile.php/7313499/mod_resource/content/24/_book/index.html

SRIPATHI, M. (2020). *Newyork Taxi Trip Data*. https://www.kaggle.com/datasets/microize/newyork-yellow-taxi-trip-data-2020-2019

Tinsae. (2019). *Yellow-Taxi-Demand-Prediction*. https://github.com/Tinsae/Yellow-Taxi-Demand-Prediction?tab=readme-ov-file

Xinyi025. (2023). *ARIMA & ST-ARIMA.Rmd*. https://github.com/Xinyi025/0042STDM/blob/main/ARIMA%20%26%20ST-ARIMA.Rmd