MGSC-416-001 Data-driven Models for Operation Analytics

Traffic Prediction and Optimization of Bike-sharing System in New York City

Mingshu Liu: 261044984

Steven Xia: 261039623

Faye Wu: 261043794

Kaibo Zhang: 261110409

Desautels Faculty of Management

McGill University

April 2nd, 2024

1. Introduction

Bike-sharing systems have emerged as a popular and sustainable mode of transportation in urban environments, offering an efficient means of travel for commuters and tourists alike. Understanding the dynamics of bike-sharing demand is crucial for optimizing resource allocation to ensure efficient service delivery and customer satisfaction. Building upon previous research conducted by scholars at the Hong Kong University of Science and Technology (HKUST), this project seeks to delve deeper into the transitional demands for bike-sharing stations in the vibrant landscape of New York City. Leveraging the dataset and preprocessing techniques established by prior researchers, our study aims to develop a simplified yet robust model capable of capturing the nuances of demand variations and customer patterns. Furthermore, this study goes beyond mere prediction by integrating an optimization model to determine the optimal number of bikes to deploy. Focusing on the week of August 16th to August 22nd, we aim to devise a strategy for resource allocation that maximizes efficiency and minimizes operational costs. In essence, this project bridges academic research and practical applications, offering a comprehensive framework for understanding and managing the transitional demands of bike-sharing stations in New York City. By leveraging insights from previous research and employing advanced analytical techniques, we aspire to contribute to the ongoing efforts to enhance the sustainability and accessibility of urban transportation systems.

2. Data and Sample

2.1. Data Overview and Preprocessing

To obtain trip records in New York City, we relied on the same NYC dataset that once served as the primary data for prior research. The dataset comprises two distinct files: one encompasses individual records of 473,621 trips across 325 stations, while the other encompasses meteorological data within the city, both spanning a two-week time horizon. The structures of the two data frames are as follows[1]. As a preprocessing step, we assigned an hour label ranging from 0 to 23 and a day label ranging from 0 to 15 to each row of the rides data frame. Subsequently, the ride data were aggregated based on each day and hour label to obtain the total ride records. In addition, any NaN values in the meteorology data were filled using interpolation methods to ensure data integrity and completeness. Considering the existing values in the 'weather' column, we cleaned and limited the values to 'sunny', 'rainy', and 'haze' to ensure consistency.

2.2. K-mean Clustering

As evident in the previous research, the bike traffic for nearby stations is highly interdependent. When a station is at capacity, individuals naturally seek alternative stations for bike return or retrieval. Similarly, anomalous events such as concerts can significantly disrupt typical bike ride patterns, leading to abrupt surges or declines in demand and return numbers at specific stations without prior indications. Consequently, traffic dynamics at micro-levels are highly volatile. Tackling this issue from the same angle but with a different approach, we aggregated trip records for each station and transformed the table into an adjacency matrix, in which each entry, *aij*, represents the number of trips between any two stations. We standardized the matrix, along with each station's geographical coordinates (longitude and latitude), as inputs for a K-mean clustering algorithm with an optimal number of k=3. By doing so, we adhered to the assumption that stations within each cluster exhibit not only spatial proximity but also analogous patterns in usage behavior and connectivity[2]. This is crucial in subsequent analyses as we will examine demand at the macro scale before extrapolating it back to the individual clusters. This approach ensures that these clusters serve as suitable and representative entities capable of capturing the majority of variations and trends in local demands.

2.3. Probability Computation

The destinations of individual trips vary across different clusters. They are characterized by two possible events: in-cluster/out-cluster rides and within-hour/out-hour rides. Since bike-sharing rides are mostly used for short trips, the probability of a trip lasting more than an hour is negligible. Therefore, we assumed that riders would return all out-of-hour rides within the following hour. To understand these riding habits within each cluster, we based our analysis on theories of conditional probability and Bayes' rules. Initially, we aggregated the data frame by 'start cluster label', 'day', and 'hour' to construct contingency tables. Following this, we computed the probability of whether bikes were returned within the same hour. This involved dividing the number of rides within each group by their respective totals. Using comparable methods, we calculated subsequent conditional probabilities for in/out cluster travels. Then, applying Bayes' rules, we multiplied these conditional probabilities backward through their respective branch[3] to obtain the corresponding intersection probabilities for any two events (e.g. $P(C \cap H) = P(C \mid H) \cdot P(H)$). Later in the analysis, these probabilities were then used as training data for forecasting the trend of the following week.

3. Methodology

3.1. Traffic Prediction Model

After preprocessing efforts, we obtained a foundational input table named 'demand_all[4]', which contains the hourly demand for the entire city. We chose to build the model based on the entire demand function instead of zooming into individual clusters since the potential candidates from the datasets are mostly ubiquitous to the entire city, and the latter may result in overfitting specific local patterns, making the models less generalizable.

3.1.1. Multiple Regression

Using multiple regression as the initial framework to capture the relationship between meteorology data and bike traffic, we first underwent backward selection by dropping predictors that had a p-value above α = 0.25. Examining the four in one plot[5] revealed that the normality assumption was far from satisfied. Consequently, we transformed the y variable using natural logarithms, resulting in a significant improvement in the model's performance. This is not surprising because bike demand inherently remains positive and cannot fall below zero, aligning well with the characteristics of a log-normal distribution. A squared interaction term for hour and temperature was introduced to account for potential non-linear relationships and multicollinearities. As presented in the regression outputs[6], temperature and hour are positively correlated with bike sharing demand, while the negative coefficient in 'rainy' suggests decreases in bike usage during adverse weather. The significance of the squared interaction term implies that the combined effect of temperature and hour has diminishing returns on bike demand, implying a saturation point where further increases in temperature and hour result in diminishing increases in demand. Overall, these findings align with previous research and underscore the importance of meteorological variables in predicting variations in bike demand.

3.1.2 ARIMA Modelling for Residuals

The regression failed to account for almost 34.1% of the variability in the dependent variable. Likewise, a Durbin-Watson statistic of 0.781 indicates potential autocorrelation issues. ARIMA modeling is pursued to account for these unexploited time-dependent variations. The residuals are first extracted and transformed back to standard units using the exponential function to prepare them for analysis. Plotting the residuals reveals high variations and seasonal patterns following a 24-hour cycle[7], which is expected given the hourly nature of the dataset. To address the seasonality, the data is deseasonalized using additive methods. However, even after deseasonalization, strong unexplained seasonality persists, as evidenced by the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots[8], which show spikes changing sign every 3 or 4 lags. Consequently, an *auto_arima* model with orders (2,0,1) and a seasonal period of 4 was fitted to the data to account for this recurring pattern. Its fitted values were then added back with regression outputs to obtain the final prediction. Plotting the predicted values ($\hat{y}$) against the actual observations[9], our model exhibits commendable proficiency in capturing the local minima and cyclical patterns in the provided dataset. Nevertheless, the inherent limitations of ARIMA modeling became apparent when confronted with unexpected peaks.

3.2. Exponential Smoothing for Probability Forecasting

We chose exponential smoothing for predicting bike dynamics because human behavior has a homogenous reliance on history. For instance, it is common for individuals to incorporate bike trips into their daily routines, such as commuting to work in the morning. This habitual behavior is characterized by consistent timing, occurring at the same hour, and repeating across multiple days within a week. Exponential smoothing offers a straightforward yet effective method for capturing trends in historical data and potential seasonal variations. As these dynamics are unique for each cluster and behavioral attributes, a total of 12 ETS models were fitted against the probabilities computed earlier to inform decisions related to bike allocation and resource management. The time series plot[10] demonstrates that the model effectively captured the transitional patterns, accurately depicting the fluctuations in the data over time.

4. Optimizing Initial Bike Deployment

A linear optimization model was used to determine the optimal combination of initial bike allocation across clusters at time 0 on August 16th, 2014, the first day of the week. To simplify our model and ensure consistency in results, we made two following assumptions.

a) Bicycle usage patterns throughout the week strictly follow earlier predictions without any human interference (i.e. bike reallocation during the week done by staff).

b) No out-hour traveling occurs during the last hour on August 22nd (all bikes will be returned to the stations at the end of the week).

Notations for parameters are summarized in this table[11].

Decision Variables

init_bike$_{i}$: denotes the number of initial bikes to deploy at Cluster i.
check in$_{i,h}$: denotes the number of check-ins to Cluster i in hour h.
check_out$_{i,h}$: denotes the number of check-outs at Cluster i in hour h.
active_bikei,h: denotes the number of active bikes at Cluster i in hour h.

Constraints

$$
\begin{aligned}
&1: \mathit{check\_in}_{i,0} = \sum_{j=1}^{J} \mathit{cj\_Ci\_H}_0 * \mathit{check\_out}_{j,0},\ \forall i \neq \forall j \\
&2: \mathit{check\_in}_{i,h} = \sum_{j=1}^{J} \mathit{cj\_Ci\_H}_h * \mathit{check\_out}_{j,h} + \sum_{j=1}^{J} \mathit{cj\_Ci\_Hc}_{h-1} * \mathit{check\_out}_{j,(h-1)},\ \forall i \neq \forall j \\
&3: \mathit{check\_in}_{i,167} = \sum_{j=1}^{J} (\mathit{cj\_Ci\_H}_{167} + \mathit{cj\_Ci\_Hc}_{167}) * \mathit{check\_out}_{j,167} + \sum_{j=1}^{J} \mathit{cj\_Ci\_Hc}_{166} * \mathit{check\_out}_{j,166}
\end{aligned}
\tag{1}
$$

These constraints accurately reflect bikes' inflow based on the outflow patterns observed in the current and preceding hours.

$$
\begin{aligned}
&\mathit{active\_bike}_{i,h} = \mathit{init\_bike}_i + \mathit{check\_in}_{i,h} - \mathit{check\_out}_{i,h},\ \boldsymbol{h=0} \\
&\mathit{active\_bike}_{i,h} = \mathit{active\_bike}_{i,h-1} + \mathit{check\_in}_{i,h} - \mathit{check\_out}_{i,h}\ \forall h > 0
\end{aligned}
\tag{2}
$$

These constraints update active bikes by tracking the number of inflow, outflow, and initial bike deployments.

$$
\begin{aligned}
&\mathit{check\_out}_{i,h} <= \mathit{active\_bike}_{i,h}\ \forall i\ \forall h, \\
&\mathit{check\_out}_{i,h} <= \mathit{demand\_ci}_h\ \forall i\ \forall h,
\end{aligned}
\tag{3}
$$

Ensuring that bikes checked out do not exceed the active bike or the predicted demand for any period.

$$
\mathit{check\_in}_{i,h}, \mathit{check\_out}_{i,h}, \mathit{init\_bike}_i, \mathit{active\_bike}_{i,h} \geq 0 \tag{4}
$$

Nonnegativity constraint.

Objective Function

$$
\mathit{min}(\sum_{i=1}^{I} \sum_{h=0}^{H} (\mathit{demand}_{i,j} - \mathit{check\_out}_{i,h}) + \lambda * \mathit{init\_bike}_i)) \tag{1}
$$

The model's objective seeks to minimize the mismatch between the estimated demand and the number of bikes checked out at each cluster while penalizing attempts to overly increase the number of initial bikes needed, thereby aligning supply with anticipated demand. The optimal $\lambda = 6$ was selected through trials and errors and sensitivity analysis on the magnitude of changes in demand mismatch.

5. Results and Managerial Implications

The outcomes of our optimization model unveil an optimal strategy that captures an aggregated 85.44% of the total demand and assigns 1247, 874, and 535 bicycles to clusters 1, 2, and 3, respectively. This strategic distribution can be explained from two angles. As the week progresses, the dynamics of bike utilization are reflected in the active bike counts. Deployment in the first cluster was notably higher compared to that in Cluster 2 and 3, indicating a trend[12] where trips initiated in these clusters tend to transit toward Cluster 3 rather than staying within each other. Particularly noteworthy is an abrupt surge observed around the beginning of the week, where a considerable number of bikes were circulated to Cluster 2 and 3, resulting in a sudden decrease in available bikes in Cluster 1. Secondly, the preliminary demand for out-cluster travel from Cluster 1 is high, while total demands in Cluster 2 and 3 consistently surpass that in Cluster 1[13]. As a result, bikes directed towards these clusters are less likely to return to Cluster 1. This observation may be attributed to the geographical location of bike stations. Cluster 1 predominantly covers residential areas of the city. Therefore, individuals residing in this area are more likely to have a higher demand for travel to stations in Cluster 2 and 3, where firms and metropolitan areas are located[14].

However, despite efforts to allocate bikes efficiently, the supply matching ratio for Cluster 1 is only 43.64%. This shortfall is particularly evident after a substantial number of bikes were directed toward the other two areas of the city. Furthermore, since surges observed were likely to be a result of morning traffic peaks, these bikes from Cluster 1 remained unused in Cluster 2 and 3 for a considerable amount of time while demand in Cluster 1 began to rise[15] as people started to commute to nearby shoppers or other parts of the neighborhood. Addressing this gap in service provision may necessitate reallocating resources (i.e. human intervention by moving bikes from other clusters) or temporarily increasing the number of bikes in Cluster 1 to better serve users and enhance system efficiency. Alternatively, innovative programs incentivizing users to return bikes to specific areas could also help optimize bike distribution dynamics. These strategies offer potential avenues for improving operational effectiveness and enhancing user satisfaction in the bike-sharing system.

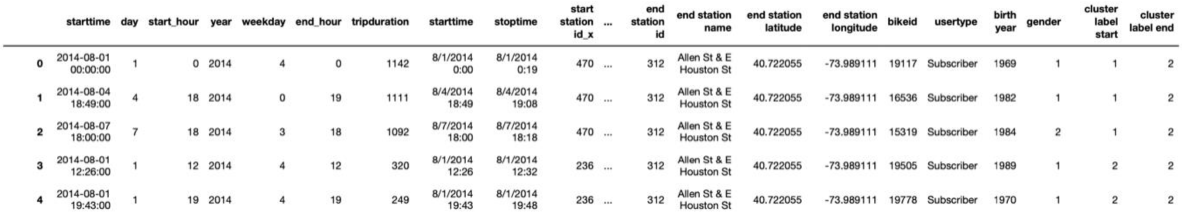
6. Limitations and Future Implementations

For the next steps, it is advisable to delve deeper into customer usage patterns and the predictive accuracy of the prediction model. The current demand prediction model ignored regional variations and could potentially underestimate the influences of meteorology indicators. In addition, other variables not included in the dataset, such as location, could also impact bike traffic. Therefore, consistent data collection on surrounding infrastructures and local building usage (i.e. commercial complexes, residential areas, schools, etc.) can be worthwhile.

7. Appendices

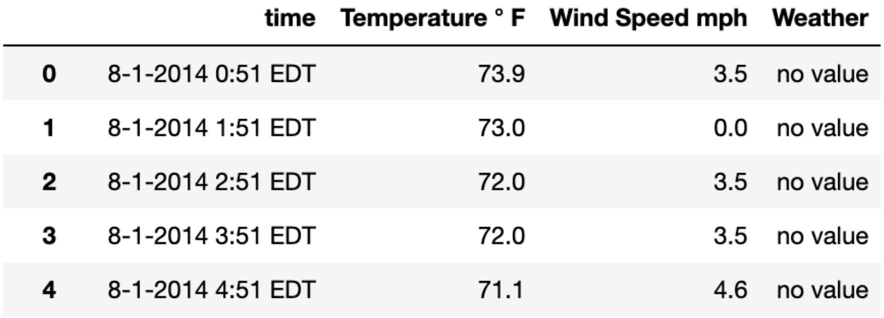
Figure 1: Structures of Data Frames

1.1. NYC Bike Sharing Transaction Data



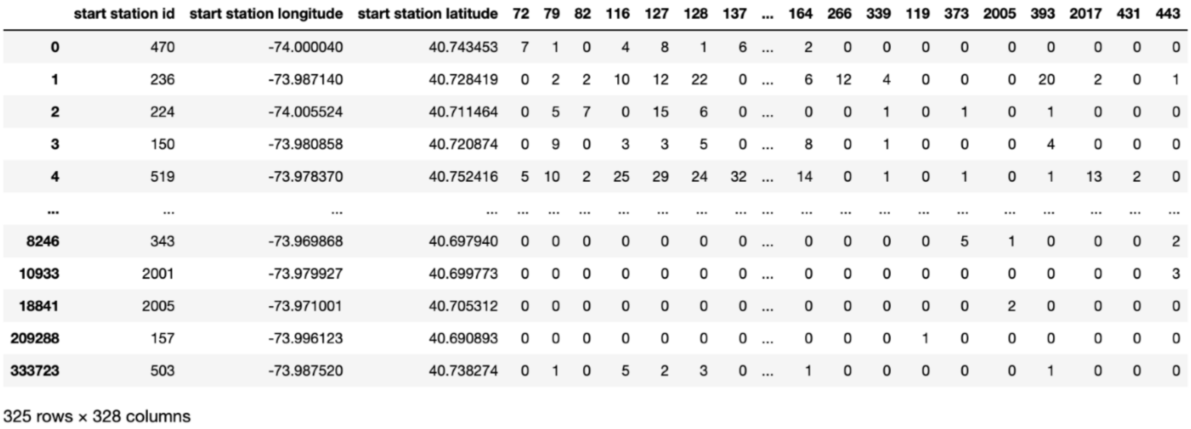
	starttime	day	start_hour	year	weekday	end_hour	tripduration	starttime	stoptime	start station id_x	...	end station id	end station name	end station latitude	end station longitude	bikeid	usertype	birth year	gender	cluster label start	cluster label end
0	2014-08-01 00:00:00	1	0	2014	4	0	1142	8/1/2014 0:00	8/1/2014 0:19	470	...	312	Allen St & E Houston St	40.722055	-73.989111	19117	Subscriber	1969	1	1	2
1	2014-08-04 18:49:00	4	18	2014	0	19	1111	8/4/2014 18:49	8/4/2014 19:08	470	...	312	Allen St & E Houston St	40.722055	-73.989111	16536	Subscriber	1982	1	1	2
2	2014-08-07 18:00:00	7	18	2014	3	18	1092	8/7/2014 18:00	8/7/2014 18:18	470	...	312	Allen St & E Houston St	40.722055	-73.989111	15319	Subscriber	1984	2	1	2
3	2014-08-01 12:26:00	1	12	2014	4	12	320	8/1/2014 12:26	8/1/2014 12:32	236	...	312	Allen St & E Houston St	40.722055	-73.989111	19505	Subscriber	1989	1	2	2
4	2014-08-01 19:43:00	1	19	2014	4	19	249	8/1/2014 19:43	8/1/2014 19:48	236	...	312	Allen St & E Houston St	40.722055	-73.989111	19778	Subscriber	1970	1	2	2

1.2. Meteorology Data



	time	Temperature ° F	Wind Speed mph	Weather
0	8-1-2014 0:51 EDT	73.9	3.5	no value
1	8-1-2014 1:51 EDT	73.0	0.0	no value
2	8-1-2014 2:51 EDT	72.0	3.5	no value
3	8-1-2014 3:51 EDT	72.0	3.5	no value
4	8-1-2014 4:51 EDT	71.1	4.6	no value

Figure 2: Pre-standardized Input for Clustering Model



	start station id	start station longitude	start station latitude	72	79	82	116	127	128	137	...	164	266	339	119	373	2005	393	2017	431	443
0	470	-74.000040	40.743453	7	1	0	4	8	1	6	...	2	0	0	0	0	0	0	0	0	0
1	236	-73.987140	40.728419	0	2	2	10	12	22	0	...	6	12	4	0	0	0	20	2	0	1
2	224	-74.005524	40.711464	0	5	7	0	15	6	0	...	0	0	1	0	1	0	1	0	0	0
3	150	-73.980858	40.720874	0	9	0	3	3	5	0	...	8	0	1	0	0	0	4	0	0	0
4	519	-73.978370	40.752416	5	10	2	25	29	24	32	...	14	0	1	0	1	0	1	13	2	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8246	343	-73.969868	40.697940	0	0	0	0	0	0	0	...	0	0	0	0	5	1	0	0	0	2
10933	2001	-73.979927	40.699773	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	3
18841	2005	-73.971001	40.705312	0	0	0	0	0	0	0	...	0	0	0	0	0	2	0	0	0	0
209288	157	-73.996123	40.690893	0	0	0	0	0	0	0	...	0	0	0	1	0	0	0	0	0	0
333723	503	-73.987520	40.738274	0	1	0	5	2	3	0	...	1	0	0	0	0	0	1	0	0	0

325 rows × 328 columns

Figure 3: Tree Diagram for Calculating Probabilities

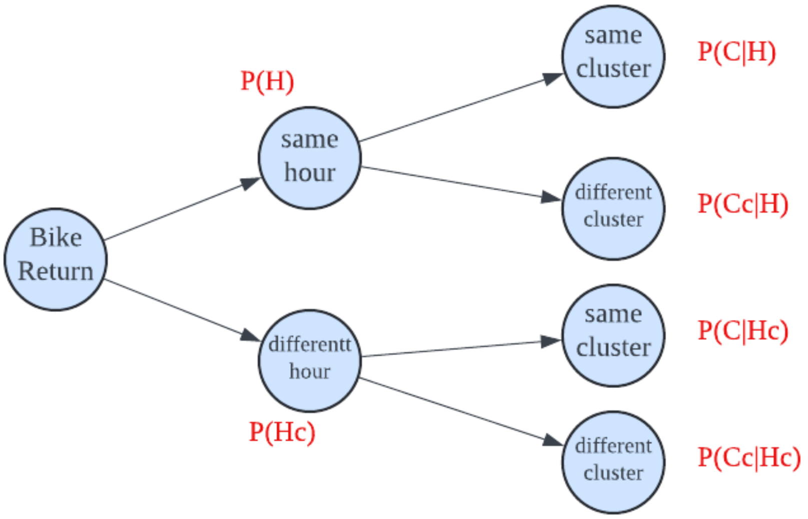
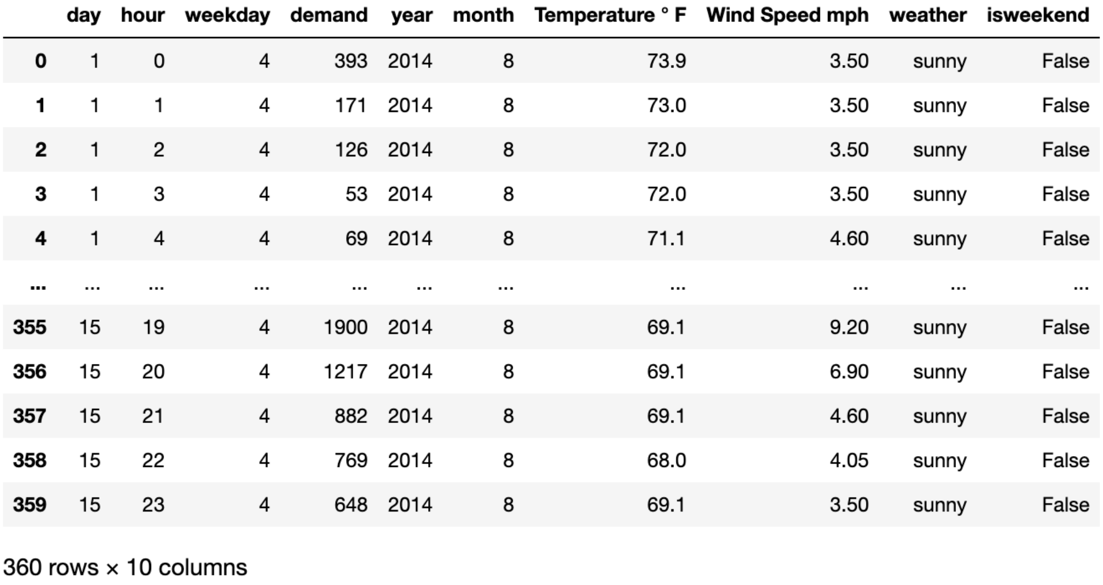


Figure 4: Demand Data Aggregated by Day and Hour



	day	hour	weekday	demand	year	month	Temperature ° F	Wind Speed mph	weather	isweekend
0	1	0	4	393	2014	8	73.9	3.50	sunny	False
1	1	1	4	171	2014	8	73.0	3.50	sunny	False
2	1	2	4	126	2014	8	72.0	3.50	sunny	False
3	1	3	4	53	2014	8	72.0	3.50	sunny	False
4	1	4	4	69	2014	8	71.1	4.60	sunny	False
...	...	...	...	...	...	...	...	...	...	...
355	15	19	4	1900	2014	8	69.1	9.20	sunny	False
356	15	20	4	1217	2014	8	69.1	6.90	sunny	False
357	15	21	4	882	2014	8	69.1	4.60	sunny	False
358	15	22	4	769	2014	8	68.0	4.05	sunny	False
359	15	23	4	648	2014	8	69.1	3.50	sunny	False

360 rows × 10 columns

Figure 5: Four in One Plot from Multiple Regression

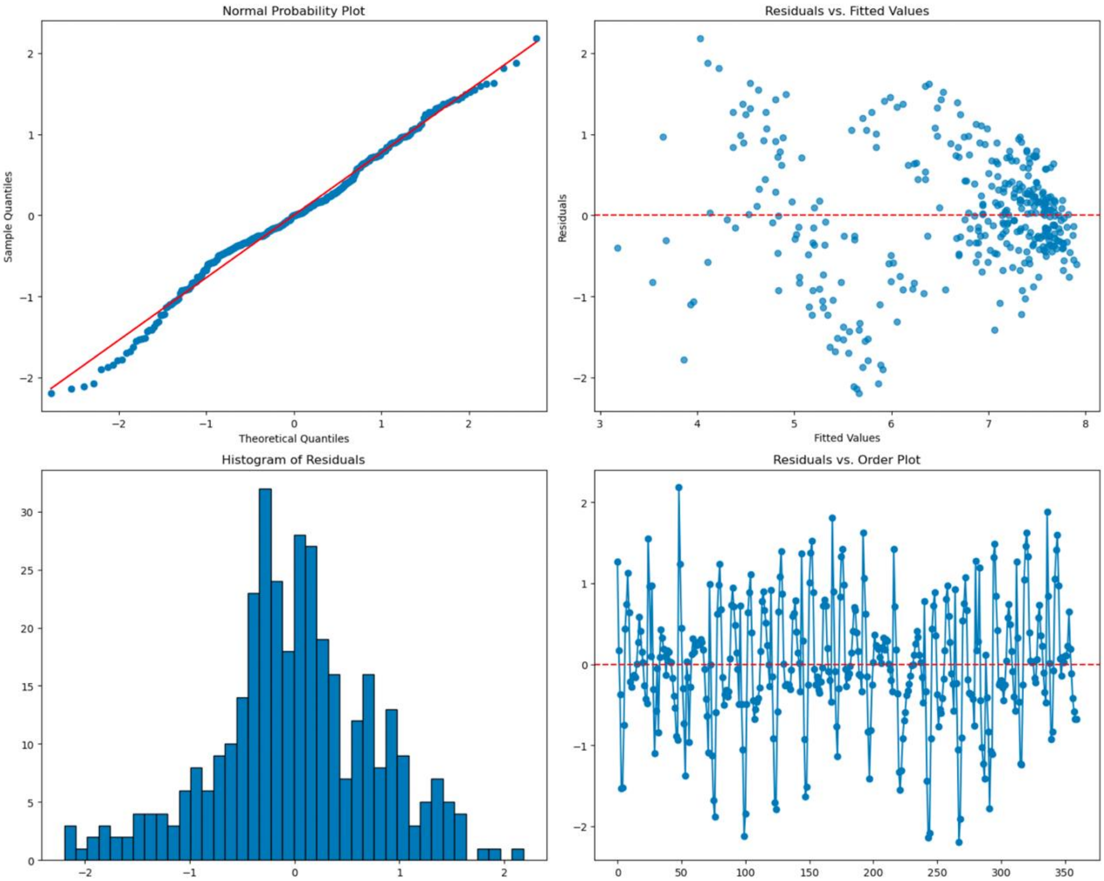
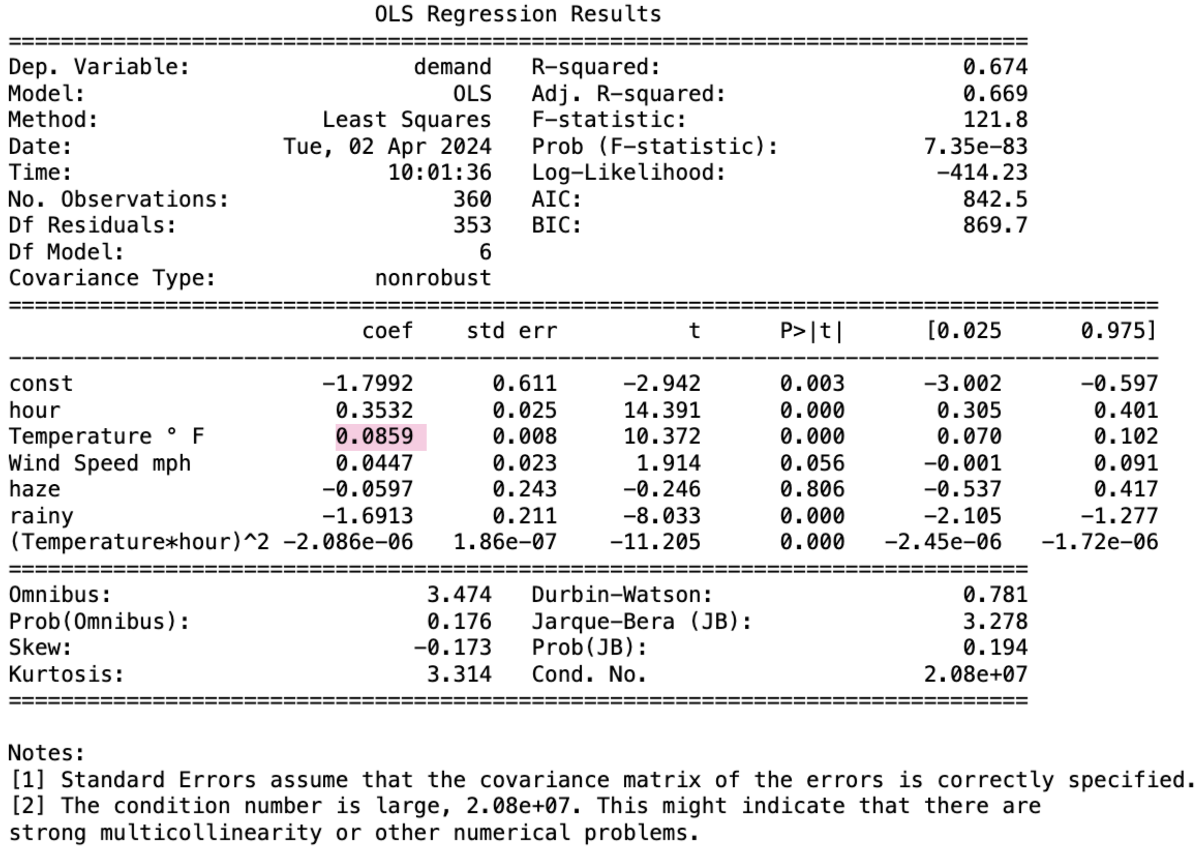


Figure 6： Multiple Regression Results Output



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 demand   R-squared:                       0.674
Model:                            OLS   Adj. R-squared:                  0.669
Method:                 Least Squares   F-statistic:                     121.8
Date:                Tue, 02 Apr 2024   Prob (F-statistic):           7.35e-83
Time:                        10:01:36   Log-Likelihood:                -414.23
No. Observations:                 360   AIC:                             842.5
Df Residuals:                     353   BIC:                             869.7
Df Model:                           6
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
const                   -1.7992      0.611     -2.942      0.003      -3.002      -0.597
hour                     0.3532      0.025     14.391      0.000       0.305       0.401
Temperature ° F          0.0859      0.008     10.372      0.000       0.070       0.102
Wind Speed mph           0.0447      0.023      1.914      0.056      -0.001       0.091
haze                    -0.0597      0.243     -0.246      0.806      -0.537       0.417
rainy                   -1.6913      0.211     -8.033      0.000      -2.105      -1.277
(Temperature*hour)^2 -2.086e-06   1.86e-07    -11.205      0.000   -2.45e-06   -1.72e-06
==============================================================================
Omnibus:                        3.474   Durbin-Watson:                   0.781
Prob(Omnibus):                  0.176   Jarque-Bera (JB):                3.278
Skew:                          -0.173   Prob(JB):                        0.194
Kurtosis:                       3.314   Cond. No.                     2.08e+07
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.08e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 7: Time Series Plot for Original and Desasonalized Residuals

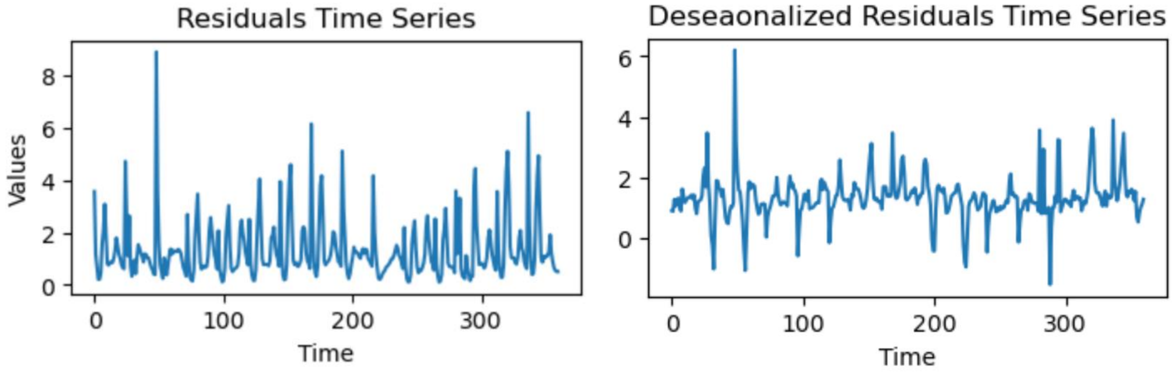


Figure 8: ACF and PACF Plots for Deseaonalized Residuals

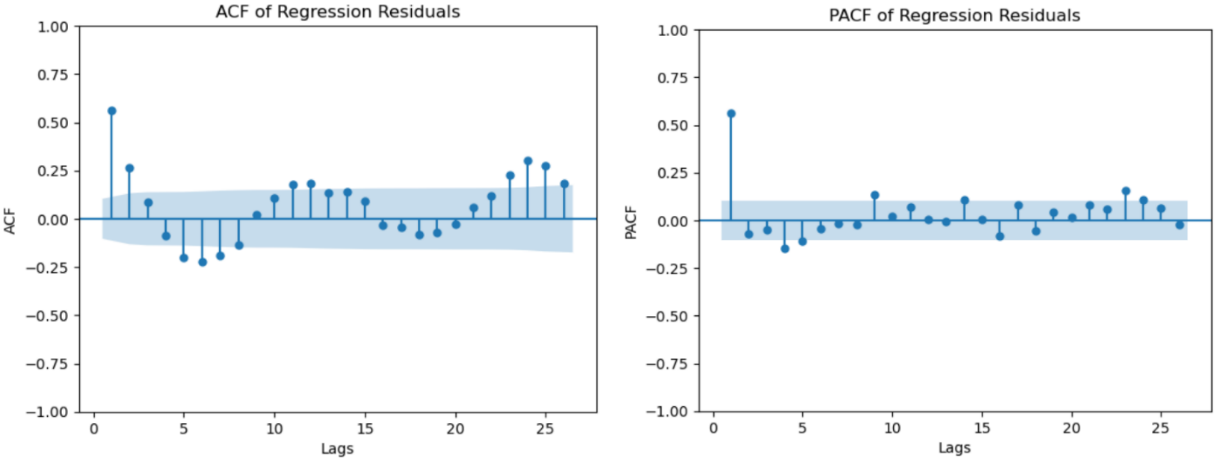


Figure 9: Complete Predictions vs. Actual Demand

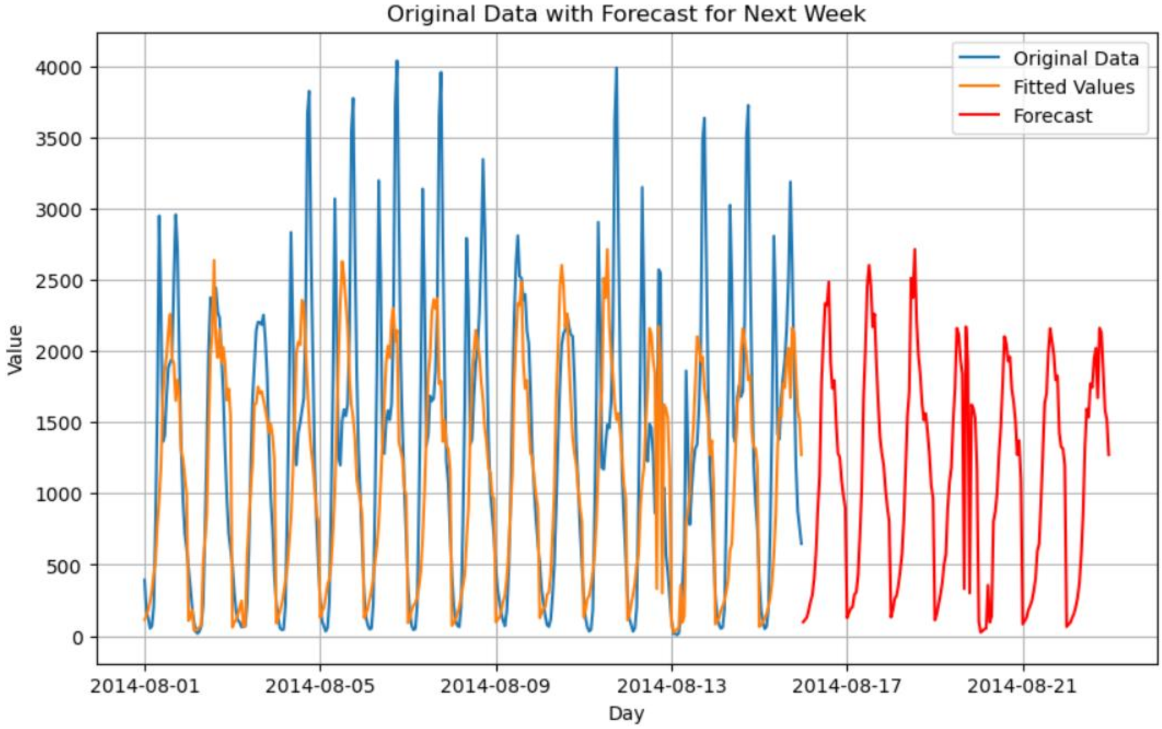


Figure 10: Sample ETS Model Predictions vs. Actual for P(C∩H) in Cluster 2

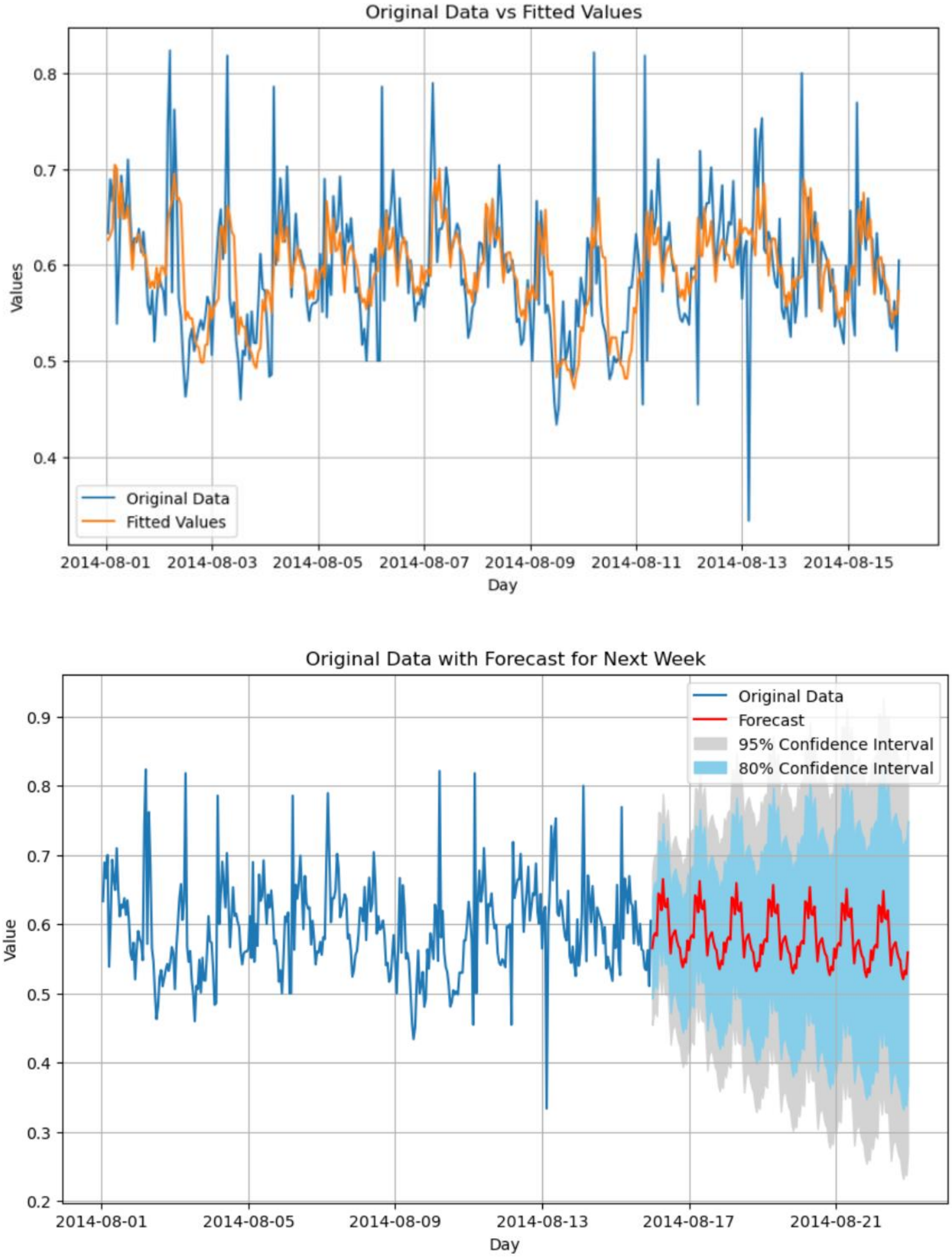


Figure 11: Notation Table

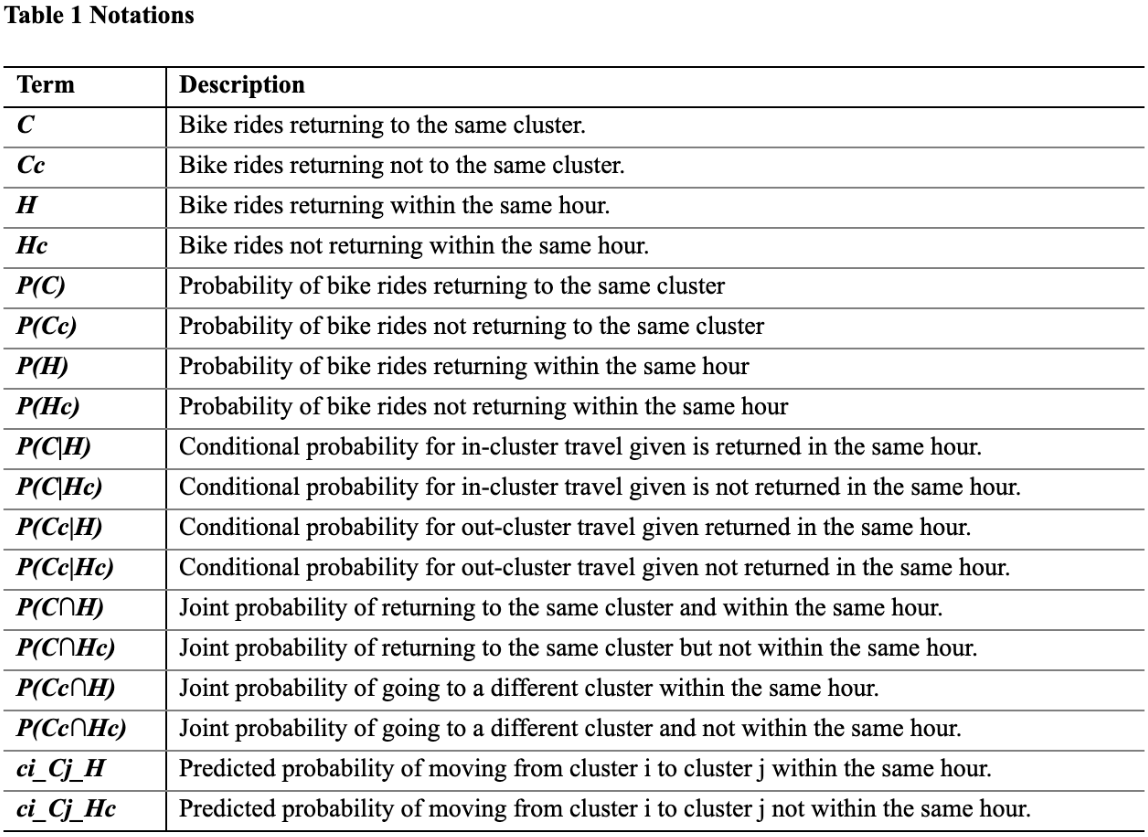


Table 1 Notations

Term	Description
C	Bike rides returning to the same cluster.
Cc	Bike rides returning not to the same cluster.
H	Bike rides returning within the same hour.
Hc	Bike rides not returning within the same hour.
P(C)	Probability of bike rides returning to the same cluster
P(Cc)	Probability of bike rides not returning to the same cluster
P(H)	Probability of bike rides returning within the same hour
P(Hc)	Probability of bike rides not returning within the same hour
P(C|H)	Conditional probability for in-cluster travel given is returned in the same hour.
P(C|Hc)	Conditional probability for in-cluster travel given is not returned in the same hour.
P(Cc|H)	Conditional probability for out-cluster travel given returned in the same hour.
P(Cc|Hc)	Conditional probability for out-cluster travel given not returned in the same hour.
P(C∩H)	Joint probability of returning to the same cluster and within the same hour.
P(C∩Hc)	Joint probability of returning to the same cluster but not within the same hour.
P(Cc∩H)	Joint probability of going to a different cluster within the same hour.
P(Cc∩Hc)	Joint probability of going to a different cluster and not within the same hour.
ci_Cj_H	Predicted probability of moving from cluster i to cluster j within the same hour.
ci_Cj_Hc	Predicted probability of moving from cluster i to cluster j not within the same hour.

Figure 12: Time Series Plot for Active Bikes

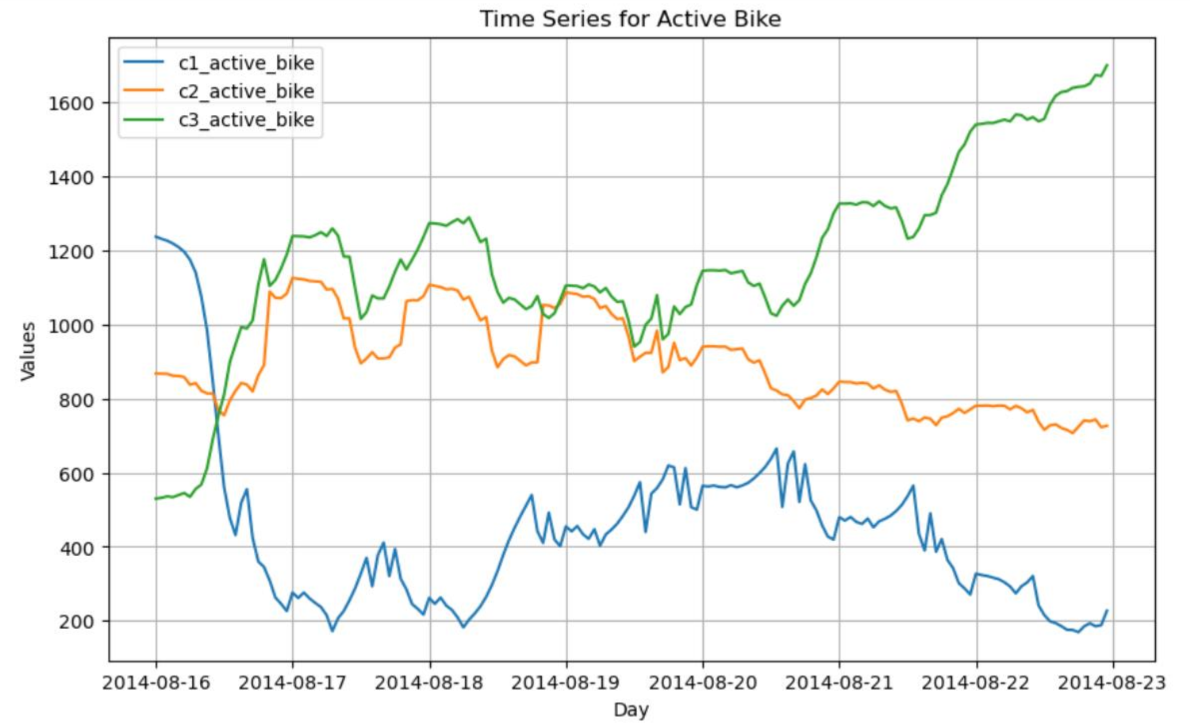


Figure 13: Time Series Plot for Demand in Different Clusters

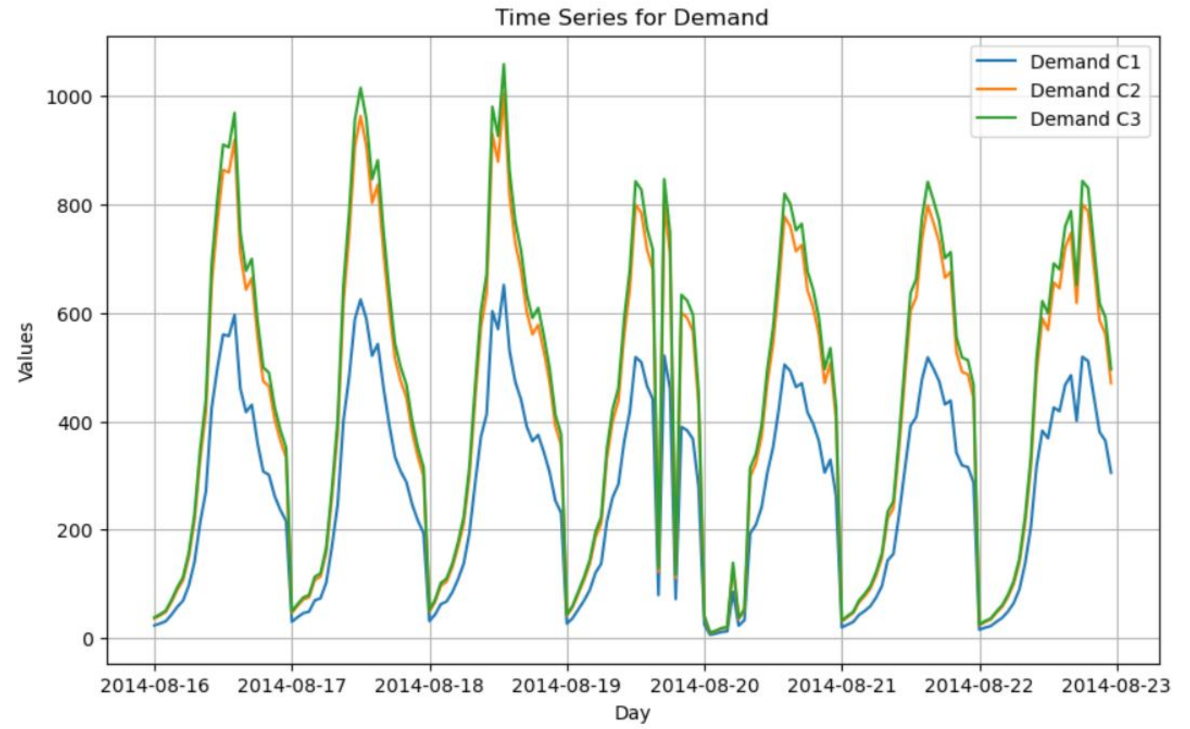


Figure 14: Cluster and Station Distribution on NYC Map

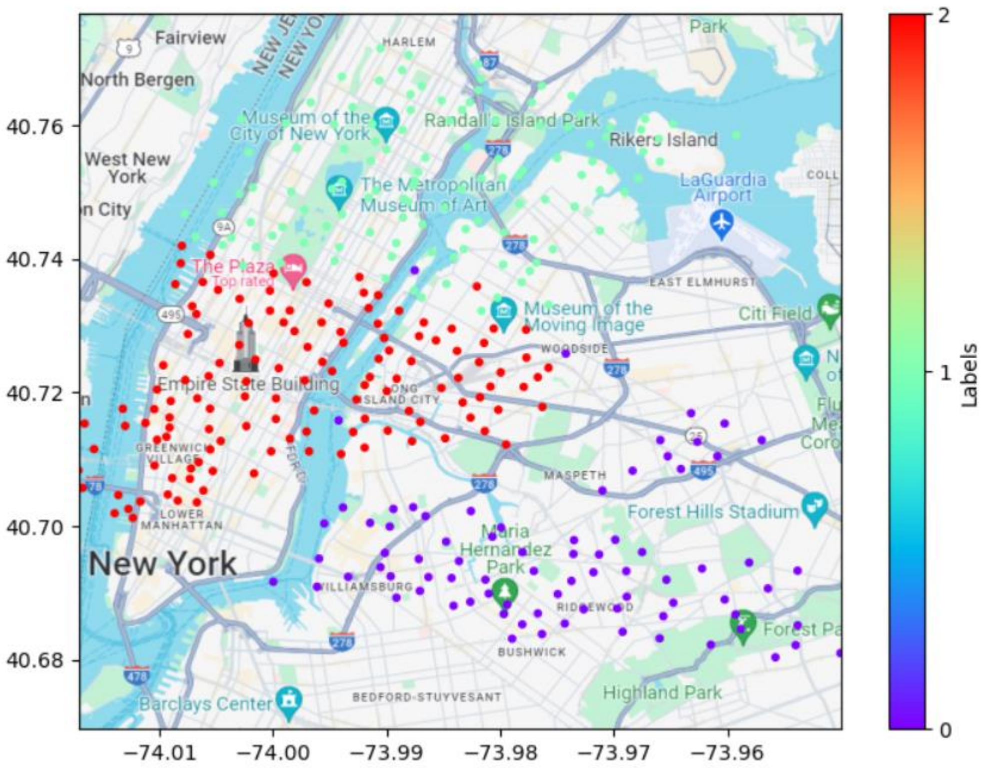
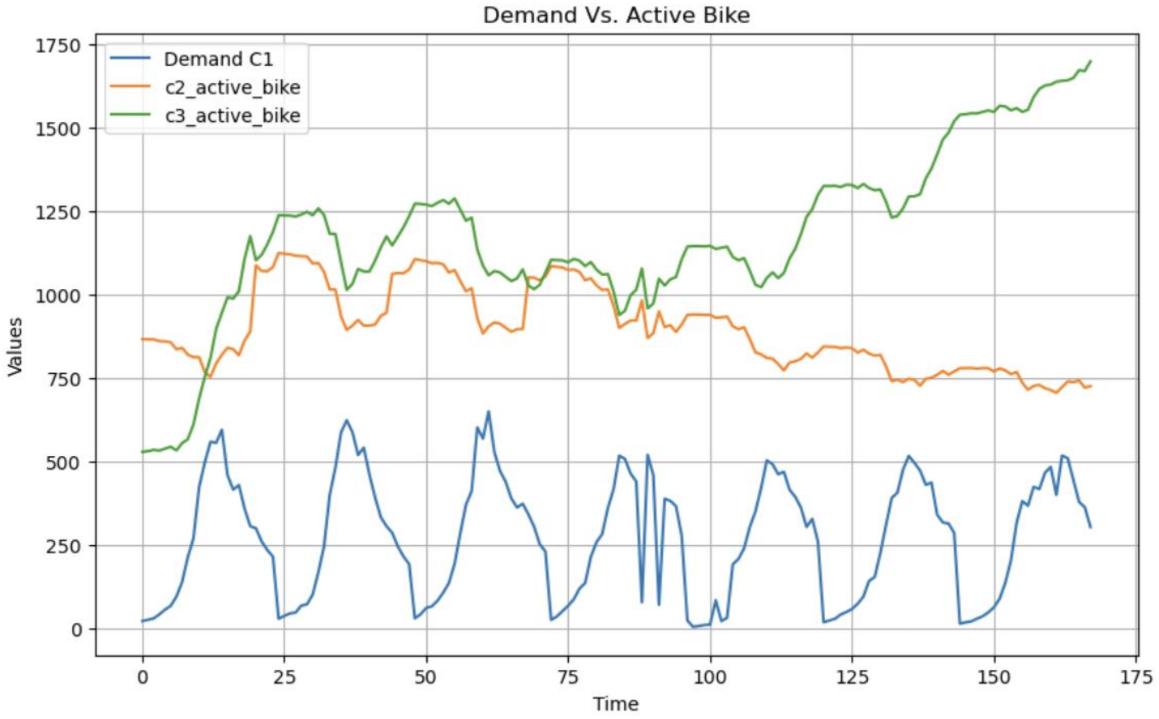


Figure 15: Demand in Cluster 1 vs. Active Bikes in Other Clusters



8. References

Li, Yexin, et al. “Traffic Prediction in a Bike-sharing System.” *ACM Digital Library*, Nov. 2015, https://doi.org/10.1145/2820783.2820837.