

DEMAND PREDICTION, STATION SEGMENTATION AND TIME-SERIES FORECASTING FOR CAPITAL BIKESHARE

EXCELLENCE INNOVATIONS
SAMPLE - FOR REFERENCE ONLY

Executive Summary

This research is focused on demand analysis in the context of Capital Bikeshare operations by using predictive analytics and clustering techniques. Trip history data and weather information from May 2020 to August 2024 have been analyzed to understand demand dynamics, compare linear regression modeling with decision tree approaches, cluster the stations based on demand characteristics, and create the forecasting model based on weather and SARIMAX model for 30 days ahead forecasting. It was concluded that bicycle demand is significantly affected by seasonality, weather conditions, and long-term trends. Linear regression showed the best performance in terms of prediction accuracy and three stations' clusters were found to facilitate more efficient management of the stations.

EXCELLENCE INNOVATIONS
SAMPLE - FOR REFERENCE ONLY

Table of Contents

1. Introduction and decision context.....	4
2. Data Preparation and Analytical Design.....	4
2.1 Data Sources and Structure.....	4
2.2 Pre-processing and validation.....	5
3. Exploratory data analysis.....	5
4. Prediction: Regression and Decision Trees.....	7
4.1 Model rationale and implementation.....	7
4.2 Results and interpretation.....	7
5. Clustering and station profiles.....	9
6. Time-series analysis and forecasting.....	10
6.1 Trend, Seasonality and Stationarity.....	10
6.2 SARIMAX Validation and Final Forecast.....	11
7. Recommendations and evaluation.....	14
8. Conclusion.....	15
References.....	16
Appendices.....	18
Appendix A. Full September 2024 forecast.....	18

1. Introduction and decision context

Capital Bikeshare is the public bicycle-sharing system serving metropolitan Washington, DC. Its current network includes more than 5000 bicycles and approximately 600 stations across seven jurisdictions, and bikes can be hired and returned throughout the network. The operator publishes monthly trip-history files for public analysis, including trip times and origin and destination stations. Demand is nevertheless difficult to manage because it varies by location, day, season and weather. Poor anticipation can create empty origin stations, full destination docks and avoidable vehicle rebalancing costs. This research project centres around the analysis of four different aspects. It starts with the analysis of the pattern and relationship associated with the bicycle pickup demand on a day-to-day basis. Next, it assesses the prediction ability of two different models of prediction the clear linear regression model and the nonlinear decision tree to find out which model predicts better daily bicycle pickup demand. This project will also group bicycle stations into meaningful demand clusters through clustering algorithms to allow for different business decisions. It can develop a weather-assisted time series forecast model for bicycle demand prediction for 30 days ahead.

2. Data Preparation and Analytical Design

2.1 Data Sources and Structure

The main analytical file combines Capital Bikeshare activity with daily weather attributes for Washington, DC. Capital Bikeshare provides the underlying trip-history data, while the weather fields are consistent with the variables supplied through the Visual Crossing Timeline Weather API, which supports historical daily data with CSV outputs. The station-day file contains 873318 rows and 12 original columns. It includes numeric pickup and drop-off counts plus temperature, humidity, precipitation, wind speed, cloud cover, visibility and solar energy (Kaggle.com, 2024). Station name and conditions are qualitative fields. Day type, season and simplified weather group were subsequently engineered, thereby providing more than three numeric and three categorical variables for analysis. A second dataset was derived by summing pickups and drop-offs by date. It contains 1584 unique daily records from 1 May 2020 through 31 August 2024, well above the two-year requirement. The date index has no missing values or duplicate dates. Daily pickups averaged 9240.23 (SD 4151.60), ranging from 450 to 20504 and daily drop-offs averaged 9127.91 (SD 4131.03). Both datasets still pertain to the same organisation and process. The unit of analysis was deliberately varied for the purpose of making other kinds of decisions (Biswas et al., 2025). Station-day data retains the regional variation in demand and allows profiling of stations and clusters of stations. Day-level system data eliminates redundant weather information from different stations and provides the appropriate unit for regression at the level

of the whole system. The use of these units prevents the problem of pseudoreplication in Hundreds of similar weather measures per day could make the sample too precise.

2.2 Pre-processing and validation

Dates were transformed to the datetime type and ordered chronologically. Numerical variables were checked, and the notebook reported that all original columns have zero missing values. Duplicate dates were deleted when forming the model tables for daily analysis, and aggregation was done by one value of weather conditions per date because weather conditions are duplicated by stations. Calendar features included month, day of week, weekend, and time trend. Weather conditions were aggregated into operational groups, and station observations were aggregated into one observation per station-date before clustering. As for K-means, feature variables were scaled due to the sensitivity of Euclidean distance to the scale (Ismael, 2025). The number of days when station observations should take place was not less than 1267 out of 1584 days, which excludes relatively new stations with sparse data. Prediction was done with chronological splitting into train-test sets, where train part contains 1267 days and the test part 317 days. Pre-processing was applied only on the train sample. This guarantees maintaining temporal order and minimising information leakage. Model performance metrics included MAE, RMSE and R^2 . The primary selection metric was RMSE, which stays in the same units of daily pickups and penalises large operational errors (Ho et al., 2022).

3. Exploratory data analysis

The demand pattern demonstrates an obvious upward trend and cyclicity throughout every year. There is a 30-day moving average with valleys formed in winter and recovery through spring with higher and higher peaks after 2020. Average monthly demand grows from 4764 rides in January to 11636 rides in July and then declines in autumn. The seasonally aggregated data is as follows summer 11524 rides, autumn 9700 rides, spring 9656 rides, and winter 5407 rides. Pickup and drop-off profiles are almost parallel, suggesting broadly balanced system-level flows even though individual stations may experience local imbalance.

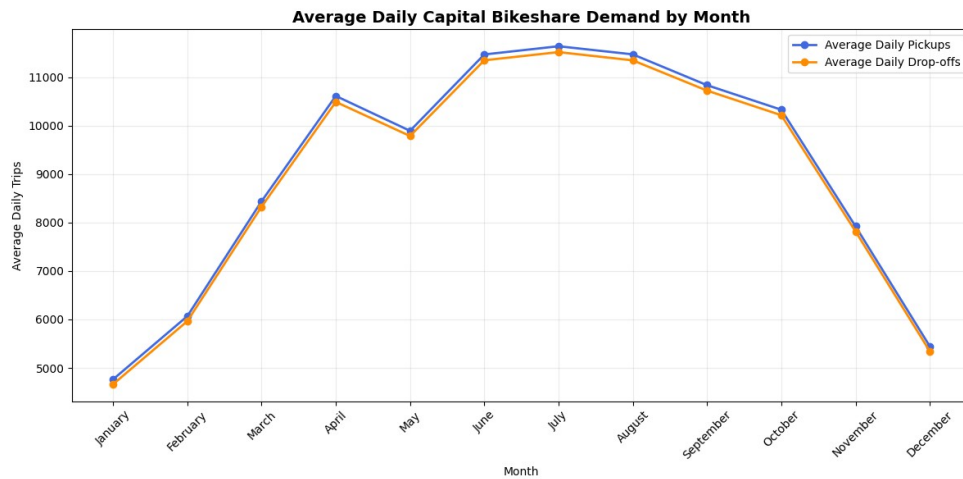


Figure 1: Average Daily Capital Bikeshare Pickups and Drop-Offs by Month

Weather relationships were directionally plausible, and daily pickups correlated positively with temperature ($r = 0.582$), solar energy ($r = 0.565$) and visibility ($r = 0.277$). But daily pickups correlated negatively with precipitation ($r = -0.255$), cloud cover ($r = -0.185$), wind speed ($r = -0.168$) and humidity ($r = -0.130$). These are associations rather than causal effects, and several weather variables are interrelated. For example, humidity correlates with cloud cover and precipitation. However, variance-inflation factors for the seven weather predictors were modest from 1.20 to 2.61. Boxplots additionally showed higher system pickups on weekends and markedly lower winter activity. These findings motivated the inclusion of weather, calendar indicators and a time trend in predictive models, plus weekly seasonality and exogenous weather in time-series forecasting. Station-level counts were far more dispersed than daily system totals: the raw station-day pickup distribution had a mean of 16.76, median of 8 and maximum of 444. Right skew is evident from the median value which is lower than the mean. This is due to the high demand from a few stations with a few days of peak periods. Right skew justified the use of a segmentation strategy that considered not only the average but also variability and maximum demand.

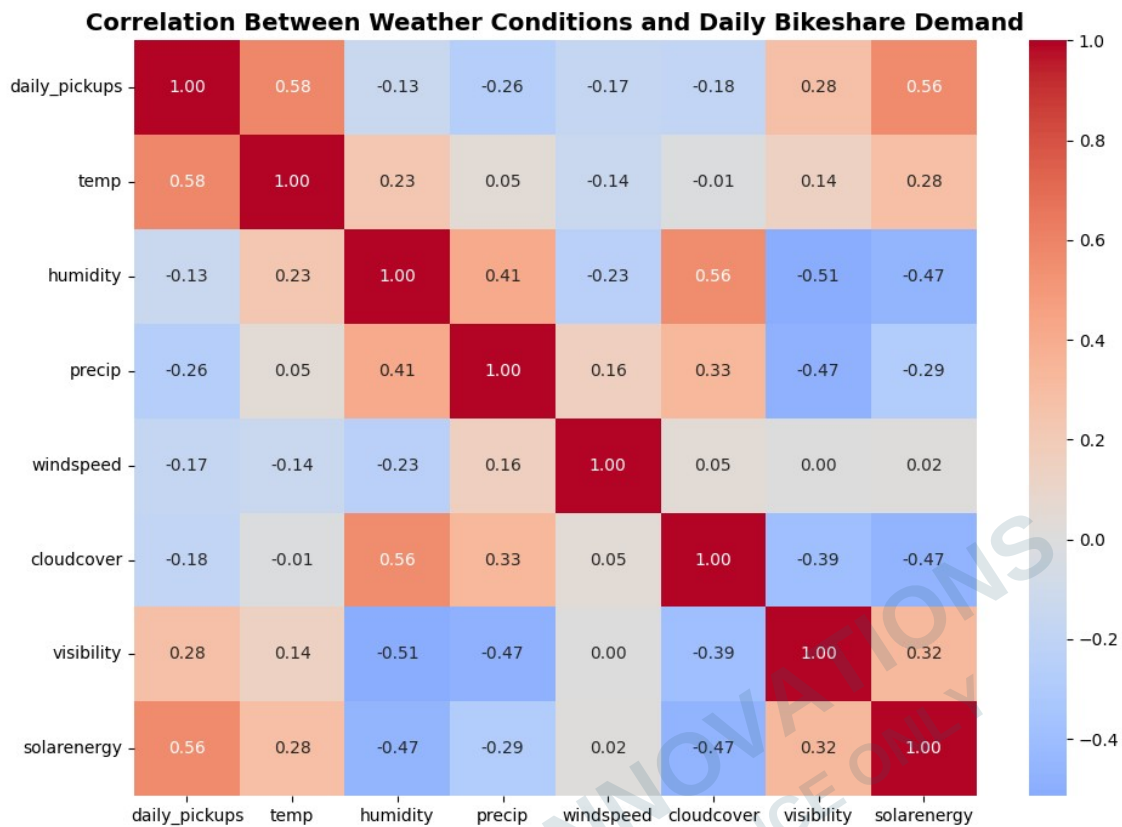


Figure 2: Correlation Matrix for Daily Pickups and Weather Variables

4. Prediction: Regression and Decision Trees

4.1 Model rationale and implementation

The predictive decision goal was to estimate total daily pickups early enough to support system-wide staffing and bicycle rebalancing. Linear regression was selected as an interpretable baseline, while decision-tree regression was used to test whether threshold effects and nonlinear interactions produced better generalisation. Both approaches used exactly the same 1584 daily observations, target and chronological holdout period. Predictors comprised seven weather variables, time trend, month and day of week. Numeric variables were median-imputed and scaled for linear regression categorical variables were imputed. Decision trees used maximum depths of 3, 5 and 8 with a minimum of 20 observations per terminal leaf (Assegie and Elaraby, 2023). Controlling depth reduces overfitting while testing three values satisfies the hyperparameter comparison requirement.

4.2 Results and interpretation

Model	MAE	RMSE	Test R ²	Rank
Linear regression	1467.42	1955.57	0.778	1
Decision tree (depth 8)	2220.08	2690.50	0.580	2

Decision tree (depth 5)	2274.27	2779.13	0.552	3
Decision tree (depth 3)	2755.04	3253.73	0.386	4

Table 1: Regression and Decision Tree Model Performance Comparison

Linear regression performed best that reduce RMSE by 734.93 pickups relative to the strongest tree and explaining 77.8 per cen of variation in unseen daily demand. Within the tree family, depth 8 was best and progressively shallower trees underfit the trend and seasonal structure. The result indicates at daily system level, a stable additive relationship among weather, calendar effects and long-term growth generalised better than the tested tree thresholds (Bobirovich, 2022). Consequently, linear regression is recommended for routine daily-demand prediction with the tree retained as a nonlinear benchmark.

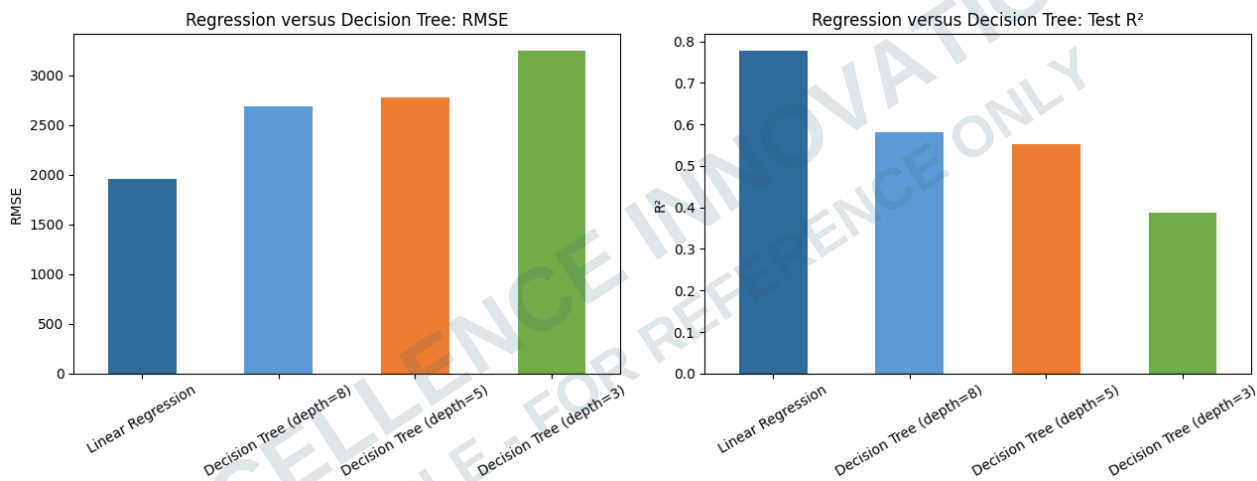


Figure 3: Fair Out-of-Sample Comparison of Linear Regression and Decision-Tree Models

The full-sample explanatory OLS model achieved $R^2=0.854$ and adjusted $R^2=0.851$. Heteroskedasticity-and-autocorrelation-consistent standard errors were used because residual testing indicated serial dependence (Hounyo, 2026). Holding calendar and trend variables constant, temperature was associated with 229.51 additional daily pickups per one-degree increase. Humidity (-37.32), precipitation (-63.00) and wind speed (-51.06) had significant negative coefficients. Visibility ($+122.95$) and solar energy ($+51.97$) were positive. Cloud cover was not significant at 5% after HAC correction ($p=0.0526$). These coefficients should be interpreted as conditional associations, not causal treatment effects. Residual non-normality and autocorrelation also mean uncertainty is better represented by robust errors and separate time-series validation than by conventional OLS assumptions alone (Iheaka, 2025). The full-sample OLS fit estimates conditional relationships and uses all observations for coefficient interpretation, whereas the holdout comparison measures how well each algorithm generalises to a later

period. The higher in-sample R^2 of 0.854 should therefore not be compared directly with the test R^2 of 0.778.

Weather predictor	Coefficient	HAC SE	p-value	Interpretation
Temperature	229.51	20.43	<0.001	Positive
Humidity	−37.32	6.40	<0.001	Negative
Precipitation	−63.00	8.34	<0.001	Negative
Wind speed	−51.06	6.11	<0.001	Negative
Cloud cover	−5.37	2.77	0.0526	Not significant at 5%
Visibility	122.95	45.32	0.0067	Positive
Solar energy	51.97	17.07	0.0024	Positive

Table 3: HAC-Robust Regression Results for Weather Predictors of Daily Pickup Demand

5. Clustering and station profiles

The clustering decision goal was to convert heterogeneous station behaviour into operational tiers for rebalancing, maintenance and capacity review. Three standardised features were used: mean daily pickups, standard deviation of daily pickups and maximum daily pickups. These are the representations of standard loads, volatility, and peak pressures. The k-means clustering algorithm was applied to determine k from 2 to 8 based on inertia and silhouette score. K-means tries to generate compact clusters by minimising the within-cluster sum of squares, while high silhouette scores near one denote well-separated clusters. While k=2 achieved the best silhouette score of 0.6085, k=3 maintained a similar score of 0.5964 along with an elbow drop in inertia.

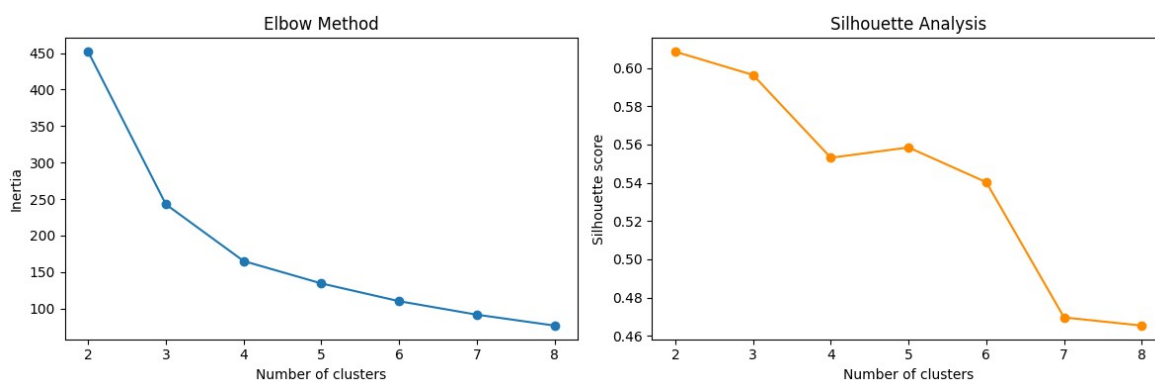


Figure 4: Elbow and Silhouette for Selecting Three Station Clusters

Cluster profile	Stations	Mean daily pickups	Mean variability	Mean maximum	Total pickups
Low demand	231	9.12	5.56	36.68	3200827
Medium demand	131	33.87	17.77	113.35	6949754
High demand	21	66.71	38.41	272.52	2189377

Table 4: K-Means Cluster Profiles of Bicycle Stations Based on Pickup Demand

The 231 low-demand stations warrant relative levels of monitoring as opposed to absolute intensive monitoring of all stations. On the other hand, the 131 medium-demand stations represent the highest aggregated demand and ought to make up the bulk of the routine rebalancing network. However, there are just 21 stations representing the high-demand stations. Notably, the average peak of these stations is almost eight times that of the low-demand stations. Thus, they require constant inventory and fast response during peak times. Cluster classification is therefore descriptive and non-permanent and needs to be reviewed as the network continues to grow or as people change their travelling behaviour (Zamprogno and Esztergár-Kiss, 2024). Consequently, the selection of three clusters represents a management decision instead of an optimal mechanical one. Two clusters could maximise the separation but bring together medium and extreme locations and hence not prioritise stations properly. On the other hand, a three-cluster solution maintains good separation and allows isolation of an important few. Since cluster labels are non-descriptive, centroid statistics were used to give them meanings of low, medium and high-demand respectively.

6. Time-series analysis and forecasting

6.1 Trend, Seasonality and Stationarity

The time series exhibits growth over the long run, a consistent yearly cycle and a significant pronounced weekly cycle. Additive decomposition using period 7 has resulted in splitting up the observed time series into a growing trend, a regular weekly cycle and an irregular component. The growing trend shows recovery and growth of the network, whereas the seasonal component represents the systematic weekly pattern of movement.

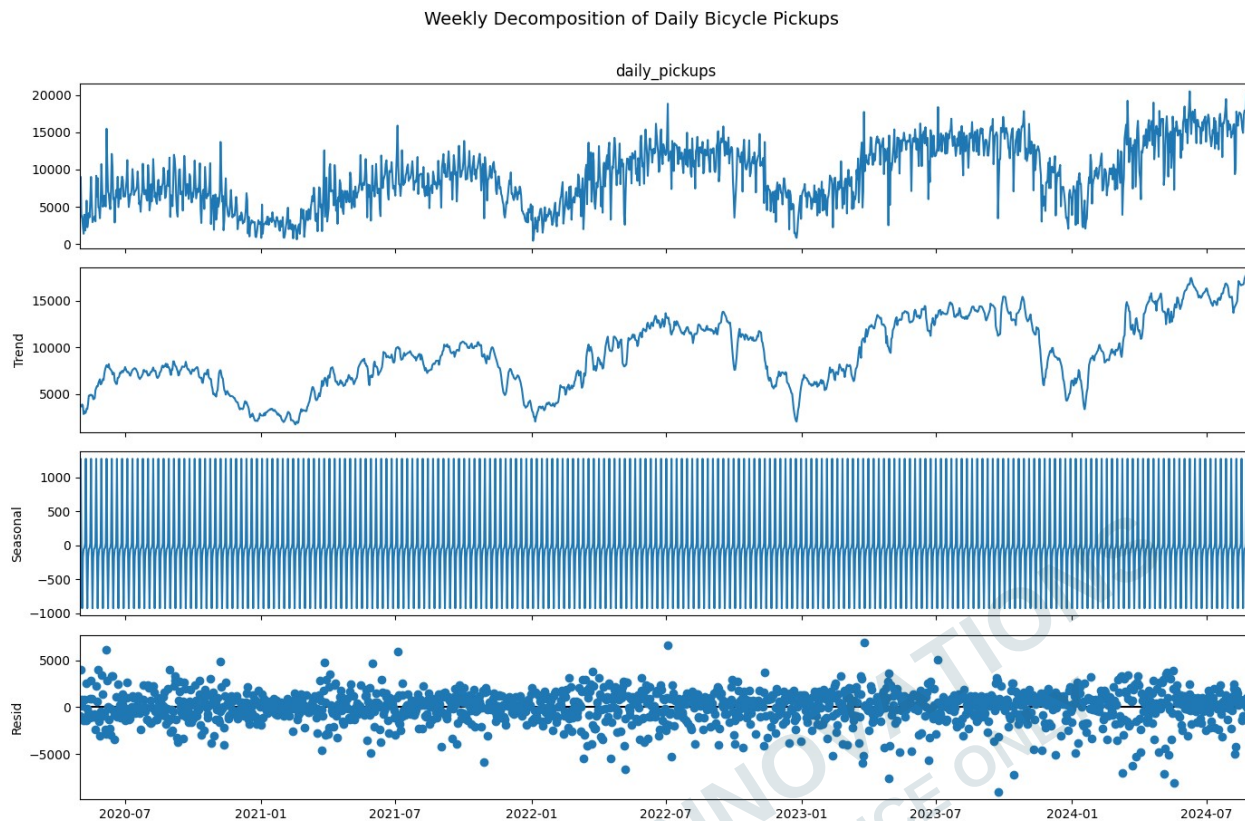


Figure 5: Additive Weekly Decomposition of Daily Pickups

Augmented Dickey-Fuller testing did not reject a unit root for the level series ($p=0.398$) or log level ($p=0.208$). The first-differenced log series was stationary ($ADF = -10.941$, $p<0.001$). This suggested $d=1$ as a statistical candidate, but differencing order was selected on genuine out-of-sample forecasting performance rather than the ADF test alone (Maitra and Politis, 2024).

Series	ADF statistic	p-value	Decision at 5%
Daily pickups	-1.764	0.398	Non-stationary
Log daily pickups	-2.196	0.208	Non-stationary
First-differenced log pickups	-10.941	<0.001	Stationary

Table 5: ADF Test Results for Stationarity of Daily Pickup Series

6.2 SARIMAX Validation and Final Forecast

The selected SARIMAX approach allows for using autoregressive and moving average models with a weekly seasonality and exogenous regressors from the weather dataset (Alharbi and Csala, 2022). The logarithmic transformation of the target variable followed by the inverse transformation with the

smearing correction factor was implemented to minimize the bias. The weather variables were standardized within each training fold.

Model	MAE	RMSE	MAPE	R ²	Mean error
SARIMAX (1,0,1)(1,0,1) ₇	2,191.88	2,792.58	21.10%	0.538	−77.09
SARIMAX (1,1,1)(1,0,1) ₇	2,434.56	3,124.24	24.51%	0.422	854.74

Table 6: Comparison of SARIMAX Model Specifications

The $d = 0$ specification was selected because it achieved lower MAE, RMSE and MAPE, higher R^2 and substantially smaller mean error (Maciejowska *et al.*, 2026). Thus, despite stationarity after first differencing, differencing reduced forecast accuracy in this validation design. The winning model also significantly outperformed last-value and weekly-naïve benchmarks in Diebold-Mariano comparisons ($p=0.028$ and $p=0.001$) (Fawaz *et al.*, 2022). This is strong evidence that weather and dynamic structure add practical predictive value.

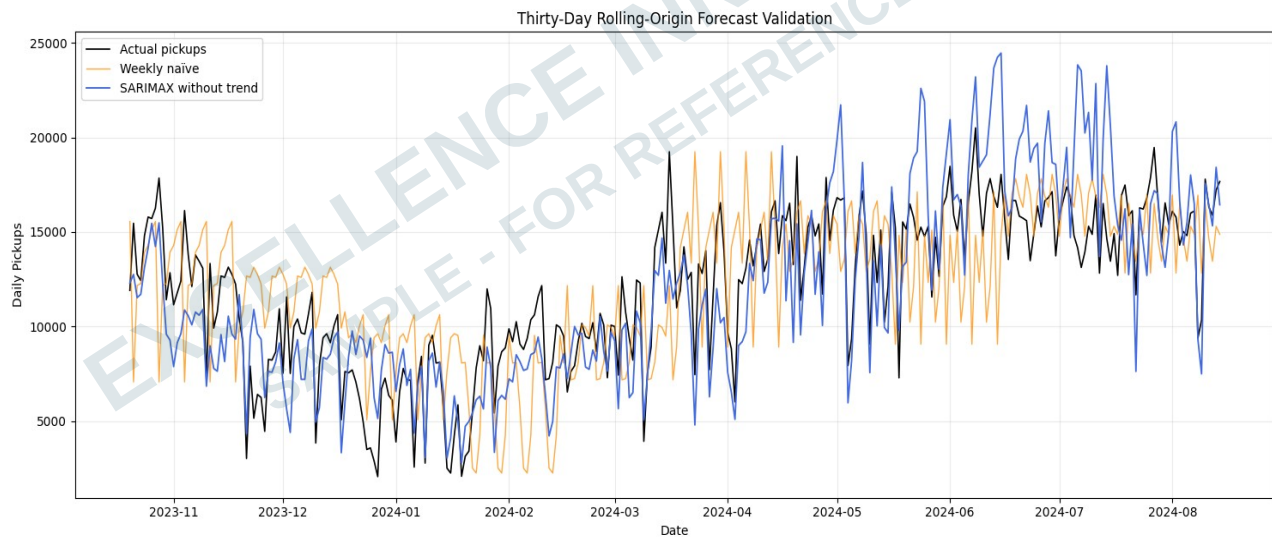


Figure 6: Thirty-Day Rolling-Origin Validation for the Selected SARIMAX model

The residual diagnostic shows deficiencies in addition, as the residual mean is nearly equal to zero (0.0077). However, Ljung-Box test results show statistical significance for lags 7, 14, 21 and 28, which means that there is a serial correlation. Jarque-Bera and ARCH tests were also significant ($p<0.001$) showing non-normality and time-varying variance (Glinskiy *et al.*, 2024). The model is therefore useful for central demand forecasts but its uncertainty should not be treated as perfectly calibrated. Prediction intervals were constructed empirically from rolling-validation log errors to reflect observed out-of-sample dispersion.

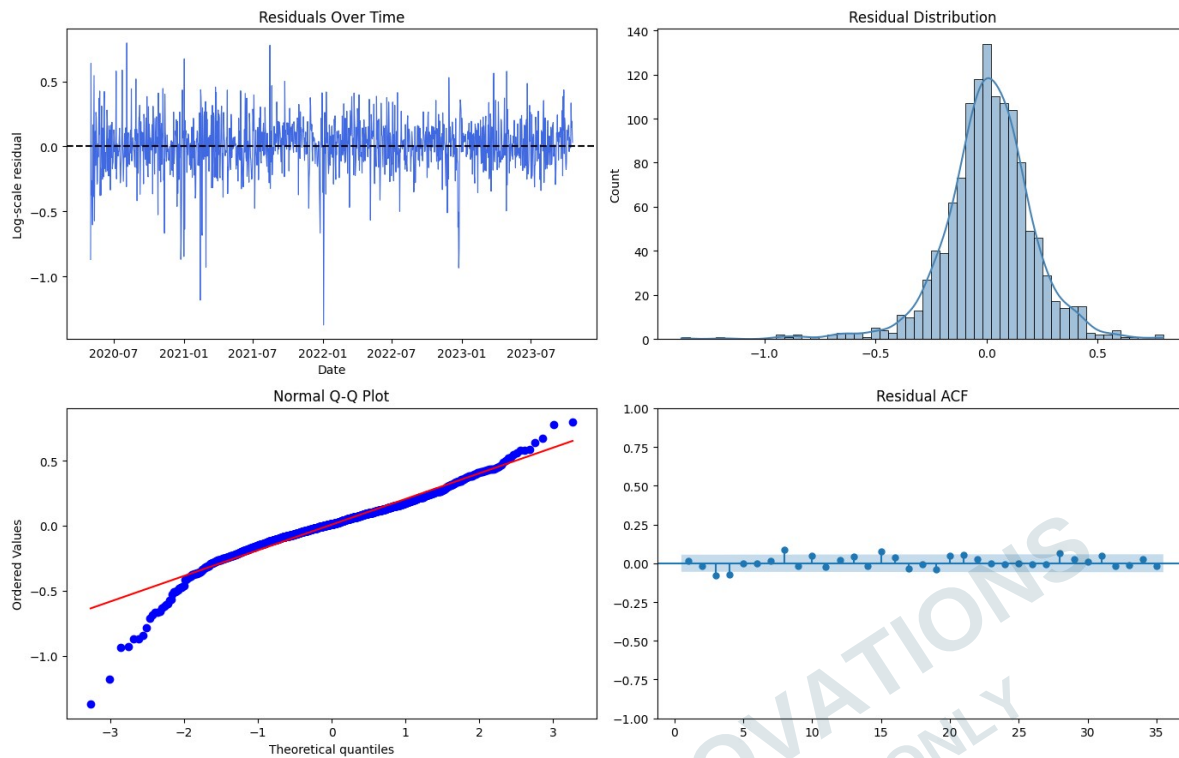


Figure 7: Residual diagnostics for the selected time-series specification.

This model was then estimated for all 1584 days and projected 1 to 30 September 2024 based on historical weather in September. Daily point forecasts range from 8147 to 17147 trips. The mean forecast is around 13086 trips per day, or 392573 trips in total for the month. The intervals are large and 1 September forecast is 14365 with an empirical 95% interval of 8225 to 22102. This again underscores the necessity of planning based on scenarios rather than precise predictions. While the use of climatological weather is justified for illustrating a forecast in the future where real exogenous forecasts are not available, it implies that all outcomes are conditional on average historical September weather. The operator would use current weather forecasts and run the model on a daily basis. It is important to monitor forecast accuracy by horizon, as a model might be accurate only for several days into the future.

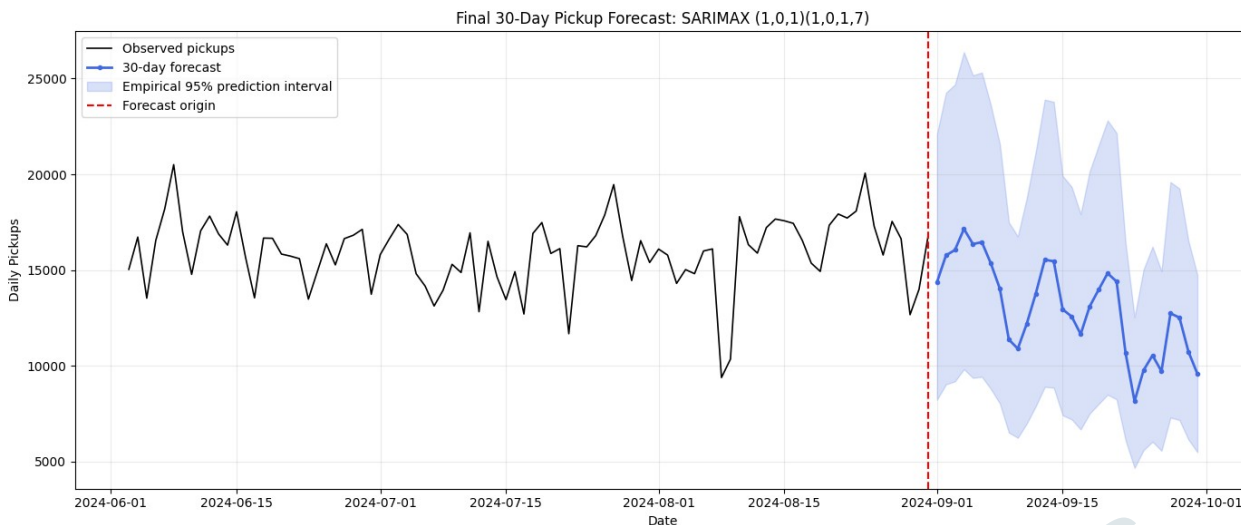


Figure 8: Final 30-Day SARIMAX Forecast with Empirical 95% Prediction Intervals

7. Recommendations and evaluation

Linear regression was chosen as the main model for daily planning because it had the minimum holdout error and was still interpretable (McAlinn and Takanashi, 2025). Refresh the model monthly and compare the RMSE and R-squared with those of the depth-8 tree to have an estimate of performance against the control. Translate weather forecasts into triggers for deploying and balancing. Days characterized by good weather with sunshine and visibility warrant increased bicycle supply, while heavy rainfall and strong winds suggest a conservative approach in deploying the bicycles. Provide differentiated service level by cluster. Observe stations with high demand intraday and pre-position bicycles ahead of the high demand, rebalance stations with medium demand regularly using routes, and send alerts to low-demand stations. For tactical planning for 30 days ahead, use SARIMAX, but provide the forecast interval along with each forecast point. Maintain contingency capacity because the residuals' autocorrelation, non-normality, and Heteroskedasticity mean that strange days are not predictable enough. Create a monthly dashboard that includes error from the model, residual drift, cluster change, station out-of-stock and weather forecast error. If there is an increase in the error or a change in the clusters, then retrain or respecify the model. The usage of business needs to match the analytical scope. In this case, regression and SARIMAX models forecast the picks at the system level, while clustering helps to classify the stations according to their demands regardless of the capacity, geographical location or vehicle travel cost (Zadeh *et al.* 2025). Operational improvement will include station capacity, member type, holidays, public events, subway problems and real-time stock.

8. Conclusion

This project meets the necessary sequence of exploratory data analysis, multi-variate prediction, hyperparameters' comparison, clustering and time series forecasting through two correlated datasets. Demand is seasonal, weather-dependent and growing over time. In a fair chronological test, linear regression beat all decision tree regressors by showing that a good additive model identified the underlying daily patterns correctly. Three cluster stations helped make a sensible differentiation among low, medium and high demand stations. Finally, the rolling origin validation method chose the undifferenced weather-enhanced SARIMAX model even though the ADF test suggested that stationarity occurs when differencing is applied. It shows the necessity of choosing operational models according to their out-of-sample performance. The proposed solution includes predictable daily forecasting, appropriate station management and interval-based time series planning.

EXCELLENCE INNOVATIONS
SAMPLE - FOR REFERENCE ONLY

References

- Alharbi, F.R. and Csala, D., 2022. A seasonal autoregressive integrated moving average with exogenous factors (SARIMAX) forecasting model-based time series approach. *Inventions*, 7(4), p.94. <https://doi.org/10.3390/inventions7040094>
- Assegie, T.A. and Elaraby, A., 2023. Optimal tree depth in decision tree classifiers for predicting heart failure mortality. *Healthcraft frontiers*, 1(1), pp.58-66. <https://doi.org/10.56578/hf010105>
- Biswas, M.E., Hossain, M.D., Huh, E.N. and Sultana, T., 2025. Spatiotemporal data-driven hourly bike-sharing demand prediction using apexboost regression. *International Journal of Data Science and Analytics*, 20(7), pp.6259-6282. <https://doi.org/10.1007/s41060-025-00820-0>
- Bobirovich, M.A., 2022. Regression Based on Decision Tree Algorithm. *Вестник науки и образования*, (4-1 (124)), pp.25-30. <https://cyberleninka.ru/article/n/regression-based-on-decision-tree-algorithm-1>
- Dalmajer, E.S., Nord, C.L. and Astle, D.E., 2022. Statistical power for cluster analysis. *BMC bioinformatics*, 23(1), p.205. <https://doi.org/10.1186/s12859-022-04675-1>
- Fawaz, M., Fayaz, N. and Baseer, M. 2022. Energy Forecasting and Operational Optimization in steel Plants Using Explainable Ensemble Models. Available at SSRN 6883007. <https://papers.ssrn.com/sol3/Delivery.cfm?abstractid=6883007>
- Glinskiy, V., Ismayilova, Y., Khrushchev, S., Logachov, A., Logachova, O., Serga, L., Yambartsev, A. and Zaykov, K., 2024. Modifications to the Jarque–Bera test. *Mathematics*, 12(16), p.2523. <https://doi.org/10.3390/math12162523>
- Ho, G.T.S., Choy, S.K., Tong, P.H. and Tang, V., 2022. A forecasting analytics model for assessing forecast error in e-fulfilment performance. *Industrial Management & Data Systems*, 122(11), pp.2583-2608. <https://doi.org/10.1108/IMDS-01-2022-0056>
- Hounyo, U., 2026. When Does Heteroskedasticity Matter? A Contrast-Specific Theory of Robust Inference. *arXiv preprint arXiv:2607.03331*. <https://arxiv.org/pdf/2607.03331>
- Iheaka, V.C., 2025. Diagnosing and correcting violations of normality and constant variance assumptions in multiple linear regression analysis. *International Journal of Advanced Statistics and Probability*, 12(1), pp.17-27. https://www.researchgate.net/profile/Victor-Iheaka/publication/390065025_Diagnosing_and_correcting_violations_of_normality_and_constant_variance_assumptions_in_multiple_linear_regression_analysis/links/67ddbea53ad6d174c4a4db05/Diagnosing-and-correcting-violations-of-normality-and-constant-variance-assumptions-in-multiple-linear-regression-analysis.pdf

Ismael, O., 2025. The Standard Deviation Score: a novel similarity metric for data analysis. *Journal of Big Data*, 12(1), p.58. <https://doi.org/10.1186/s40537-025-01091-z>

Kaggle.com(2024). Capital bikeshare dataset 2020/05~2024/08. <https://www.kaggle.com/datasets/taweilo/capital-bikeshare-dataset-202005202408>

Maciejowska, K., Lipiecki, A. and Uniejewski, B., 2026. Statistical and economic evaluation of forecasts in electricity markets: Beyond RMSE and MAE. *Energy Conversion and Management*, 356, p.121408. <https://arxiv.org/pdf/2511.13616>

Maitra, S. and Politis, D.N., 2024. Prepivoted augmented dickey-fuller test with bootstrap-assisted lag length selection. *Stats*, 7(4), pp.1226-1243. <https://doi.org/10.3390/stats7040072>

McAlinn, K. and Takanashi, K., 2025. Optimal Hold-Out Size in Cross-Validation. *arXiv preprint arXiv:2511.12698*. <https://arxiv.org/pdf/2511.12698>

Zadeh, H.S., Zhang, L., Fackrell, M. and Anjomshoa, H., 2025. Advanced forecasting of emergency surgical case arrivals: Enhancing operating room performance. *Perioperative Care and Operating Room Management*, 38, p.100451. <https://doi.org/10.1016/j.pcorm.2024.100451>

Zamprogno, M.M. and Esztergár-Kiss, D., 2024. Applying cluster analysis for the investigation of travel behavior and user profiles. *Periodica Polytechnica Transportation Engineering*, 52(4), pp.362-371. <https://doi.org/10.3311/PPtr.25663>

Appendices

Appendix A. Full September 2024 forecast

Date (2024)	Predicted pickups	Lower 95%	Upper 95%
01 Sep	14,365	8,225	22,102
02 Sep	15,766	9,027	24,257
03 Sep	16,046	9,188	24,689
04 Sep	17,147	9,818	26,382
05 Sep	16,360	9,367	25,171
06 Sep	16,453	9,421	25,314
07 Sep	15,340	8,784	23,602
08 Sep	14,025	8,030	21,578
09 Sep	11,370	6,510	17,494
10 Sep	10,892	6,237	16,759
11 Sep	12,200	6,985	18,771
12 Sep	13,769	7,884	21,185
13 Sep	15,534	8,895	23,901
14 Sep	15,459	8,852	23,785
15 Sep	12,943	7,411	19,914
16 Sep	12,571	7,198	19,341
17 Sep	11,649	6,670	17,923
18 Sep	13,096	7,499	20,150
19 Sep	13,971	8,000	21,496
20 Sep	14,827	8,490	22,813
21 Sep	14,401	8,246	22,157
22 Sep	10,663	6,106	16,407
23 Sep	8,147	4,665	12,536

24 Sep	9,761	5,589	15,019
25 Sep	10,542	6,036	16,221
26 Sep	9,709	5,559	14,938
27 Sep	12,735	7,292	19,594
28 Sep	12,516	7,166	19,257
29 Sep	10,738	6,148	16,521
30 Sep	9,578	5,484	14,736

EXCELLENCE INNOVATIONS
SAMPLE - FOR REFERENCE ONLY