

Time Series Analysis of Weekday Ridership Trends in San Francisco's MUNI System

Aastha Prakash
aprakash@ucsb.edu

March 2025

Abstract

This project analyzes and forecasts weekday ridership trends in San Francisco's MUNI public transportation system using two distinct time series modeling methodologies: SARIMA (Seasonal AutoRegressive Integrated Moving Average) and lagged regression. The aim is to uncover underlying patterns, including trends and seasonality, and develop accurate forecasting models for future ridership.

Public transportation is a cornerstone of urban sustainability and efficiency, and understanding ridership patterns is essential for optimizing resource allocation, improving service quality, and making informed policy decisions. San Francisco's MUNI system faces unique challenges due to its dense urban environment and the influence of external factors such as the tech industry, tourism, and changing work patterns. By applying SARIMA and lagged regression, this study enhances the understanding of ridership behavior, ultimately contributing to more effective planning and resource management.

Introduction

This time series project focuses on analyzing and forecasting weekday ridership trends in San Francisco’s MUNI public transportation system. The primary objective is to identify and understand the underlying patterns in ridership data, including trends, seasonality, and external disruptions, while developing reliable forecasting models for future ridership. Two key methodologies are employed in this study: SARIMA (Seasonal AutoRegressive Integrated Moving Average) and lagged regression.

Public transportation is essential to urban sustainability and efficiency. Understanding ridership trends allows for better resource allocation, improved service quality, and data-driven policy decisions. San Francisco’s MUNI system faces unique challenges, given the city’s dense urban landscape and the influence of external factors such as the booming tech industry, seasonal tourism, and shifts in housing markets. For instance, the rise of remote work in the post-pandemic era has significantly altered commuting patterns, impacting both ridership numbers and transit revenue. Large-scale conferences and tourism peaks further complicate these trends, making it essential to develop accurate forecasting models.

Prior studies on transit ridership trends provide important context for these challenges. Wasserman and Taylor’s *Transit Blues in the Golden State: Regional Transit Ridership Trends in California* highlight that while Greater Los Angeles saw significant transit declines, the Bay Area’s ridership remained relatively stable due to the strong performance of its two largest agencies, MUNI and BART. Similarly, *What’s Behind Recent Transit Ridership Trends in the Bay Area? Volume II: Trends among Major Transit Operators* (Wasserman et al., 2020) emphasizes that MUNI, operating in a high-density, transit-supportive environment, experienced minimal ridership losses between 2014 and 2018, contributing little to statewide declines. These studies suggest that high-density transit corridors play a crucial role in sustaining ridership levels, a factor that remains relevant as cities adapt to post-pandemic commuting patterns.

The need to understand MUNI’s ridership patterns is especially pressing given recent disruptions. Economic downturns, policy changes, and the COVID-19 pandemic have all affected public transit usage. A report from the San Francisco Municipal Transportation Agency (SFMTA, 2022) revealed that MUNI’s recovery has been slower than expected, with weekday ridership remaining below pre-pandemic levels—particularly in downtown areas where office occupancy remains low. Karner et al. (2020) examined transit agency responses during the pandemic, noting how agencies, including SFMTA, adapted by prioritizing essential workers and adjusting services based on real-time ridership data. SFMTA’s equity plan played a key role in ensuring continued service for marginalized communities during this period.

By analyzing fluctuations in ridership, transit agencies can allocate resources more effectively—whether by adjusting the number of buses, trains, or streetcars on different routes, optimizing schedules, or identifying underserved areas. Such data-driven insights enable proactive decision-making regarding fare policies, service routes, and operational efficiency, contributing to a more sustainable, equitable, and responsive transportation network. This, in turn, improves the rider experience and helps ensure public transportation remains a reliable and accessible resource for all.

The insights gained from real-time data adjustments during the pandemic emphasize the importance of predictive models for future planning. This study applies two forecasting methodologies—SARIMA and lagged regression—to analyze time series data and develop forecasts for future ridership. SARIMA is particularly well-suited for capturing seasonality and long-term trends, while lagged regression models are effective in identifying how past ridership influences future patterns.

One notable finding from the data analysis is the significant shift in ridership patterns during the pandemic, particularly the sharp decline in weekday ridership in central business districts and a modest increase in suburban and residential areas. This shift highlights the lasting impact of remote work on commuting behavior, reinforcing the need for transit systems to adapt to these evolving patterns. By leveraging the SARIMA and lagged regression models, MUNI can better anticipate demand fluctuations and tailor its services to meet the changing needs of commuters.

Description of Dataset

Key Findings

- Time Range: Fiscal Year 2016 – Fiscal Year 2025
- Frequency: Monthly
- Values: Average daily weekday boardings
- Unit of Measurement: Number of boardings
- Size: 914 observations
- Data Source: San Francisco Municipal Transportation Agency (SFMTA)
- Web Link: MUNI Ridership Reports

Why This Dataset?

The MUNI transit system plays a vital role in San Francisco’s urban mobility, and understanding ridership trends is crucial for improving transit efficiency and accessibility. This dataset, which tracks weekday ridership, provides essential insights that can guide city planners, policymakers, and transit agencies in making informed decisions about transit operations. Analyzing this data can help optimize resources, improve service, and ensure that transit systems effectively meet public demand.

As a Bay Area resident, I have a personal connection to the MUNI system, which makes this analysis more than just a data exercise. By examining how external factors—such as the COVID-19 pandemic, economic changes, and the rise of remote work—have influenced public transit patterns in San Francisco, the findings hold real-world implications for transit planning and policy.

Data Collection

The dataset is collected using automated passenger counters on buses and traffic checkers on rail vehicles. Data is reported monthly as system-wide average weekday boardings.

Purpose of Study

This project aims to:

- Analyze ridership trends: investigate changes in MUNI ridership over time.
- Identify seasonal and external Factors: explore how seasonal variations and external factors (the COVID-19 pandemic, economic shifts, and remote work) have affected ridership patterns.
- Develop predictive models: Apply time series analysis methods, such as SARIMA and Lagged regression models, to forecast future ridership trends.
- Support transit planning and policy: provide data-driven insights that inform the optimization of MUNI service planning and the allocation of resources.

Methodologies

This project employs two primary time series analysis techniques: the SARIMA model and lagged regression. Both methods aim to model the temporal dependencies in MUNI ridership data and provide forecasts. However, they differ in how they account for seasonality and the influence of past values on future ridership patterns.

SARIMA (p, d, q) x (P, D, Q) Model

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model is a widely used approach for time series forecasting that accounts for both trend and seasonality. The notation SARIMA(p, d, q) x (P, D, Q) represents the following components:

- Autoregressive (AR) (p): models the dependence of the current value on its past values. The order p represents the number of lagged values included in the model.
- Integrated (I) (d): represents the degree of differencing applied to the time series to make it stationary. Differencing involves subtracting the previous observation from the current one. The order d indicates the number of times differencing is performed.
- Moving Average (MA) (q): models the dependence of the current value on past error terms. The order q represents the number of lagged error terms included in the model.
- Seasonal Components (P, D, Q)s: capture the seasonal patterns in the time series. The orders P, D, and Q represent the seasonal AR, I, and MA components, respectively, and s is the seasonal period (e.g., 12 for monthly data with yearly seasonality).

Lagged Regression Model

The Lagged Regression Model is a straightforward approach for forecasting time series data, particularly when the goal is to understand the relationship between current values and past observations. In this model, past values of the dependent variable (in this case, MUNI ridership) are used as independent variables (lagged values) to predict future ridership.

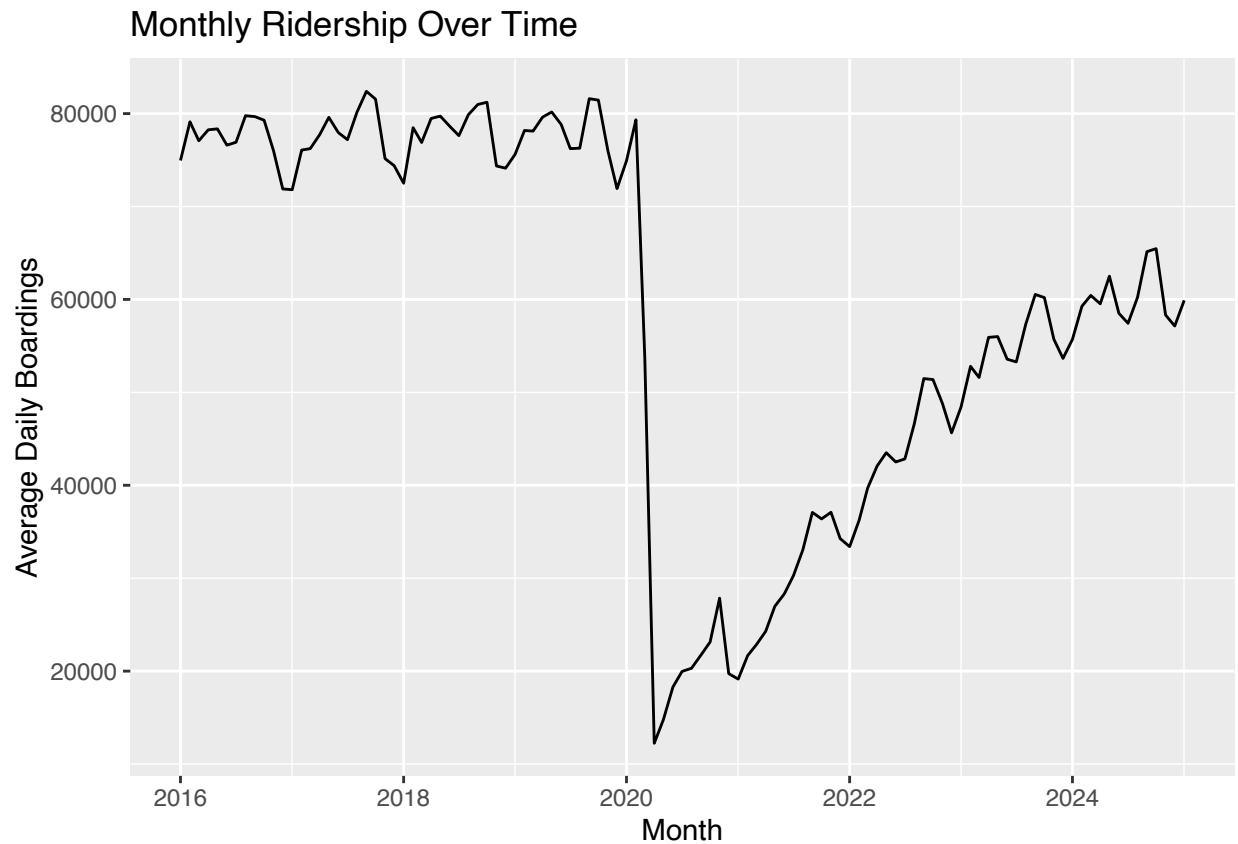
The model works by incorporating lagged variables—previous time points—as predictors for future outcomes. For example, to predict ridership for a given month, we use ridership data from previous months as input features. The model's coefficients for each lag term indicate how much influence past values have on future ridership.

The lagged regression model is useful for capturing the temporal dependencies in the data, particularly when there is an observable pattern of past behavior influencing future outcomes. By fitting a linear regression model with these lagged variables, we can forecast future ridership based on recent trends.

SARIMA and Lagged Regression serve as the core frameworks for analyzing and forecasting MUNI ridership trends, with SARIMA capturing seasonal and trend components, and Lagged Regression focusing on the influence of past ridership values on future predictions.

Results

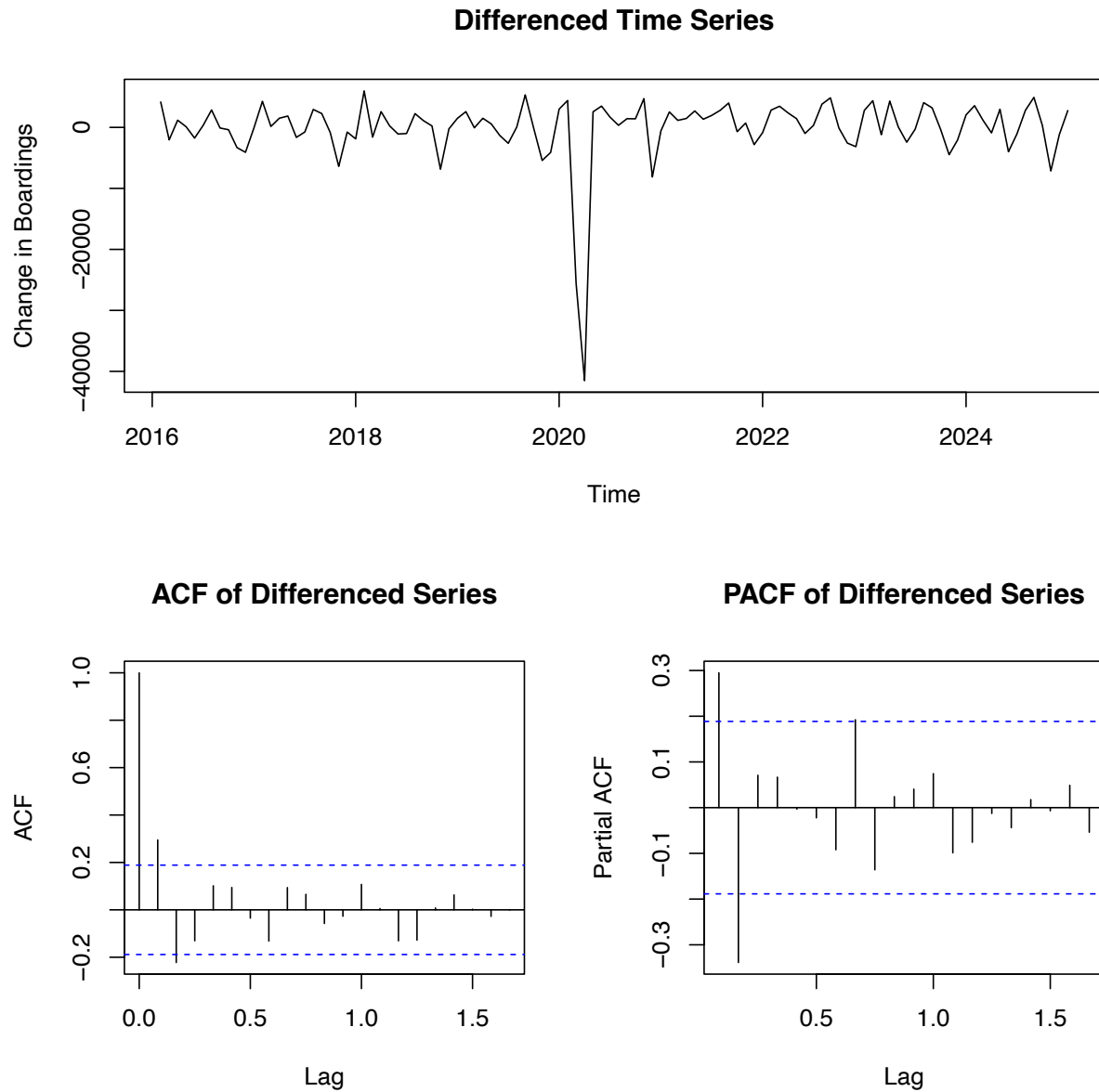
Exploratory Data Analysis



We first examined the monthly ridership trends to identify patterns and anomalies. Figure 1 illustrates the fluctuations in Average Daily Boardings over time, revealing clear seasonality with consistent peaks and declines throughout the year.

However, one notable anomaly is the sharp drop in March 2020, which aligns with the onset of COVID-19 lockdowns and likely explains the significant decrease in ridership. The presence of an overall trend in the data suggests the need for differencing to achieve stationarity.

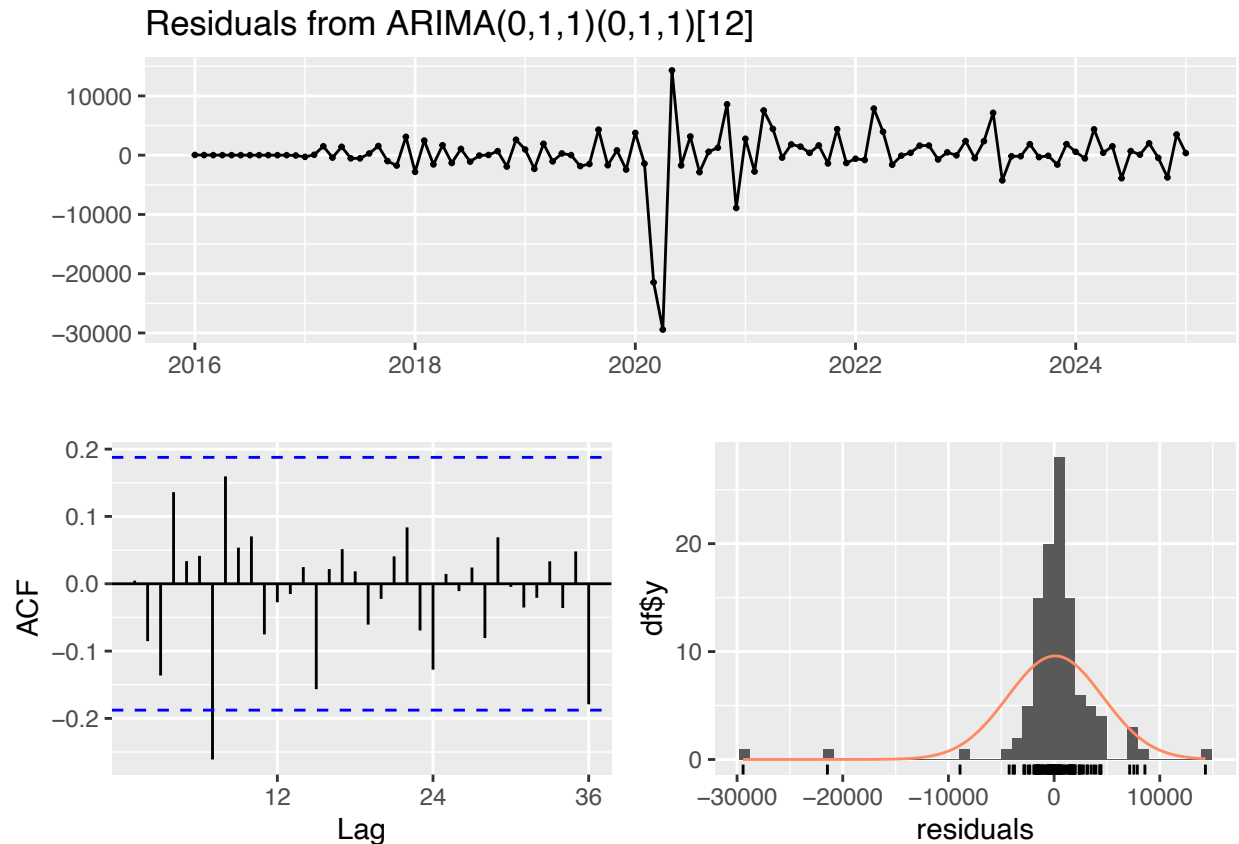
Data Transformation for Stationarity



I applied first differencing to the ridership data to address potential non-stationarity. The resulting differenced time series is shown where the trend has been reduced, and the series appears more stable. To confirm stationarity, I also plotted the ACF and PACF of the differenced series. The ACF plot shows a rapid decay, suggesting a MA (Moving Average) component, while the PACF plot indicates a sharp cutoff, pointing to a potential AR (AutoRegressive) component.

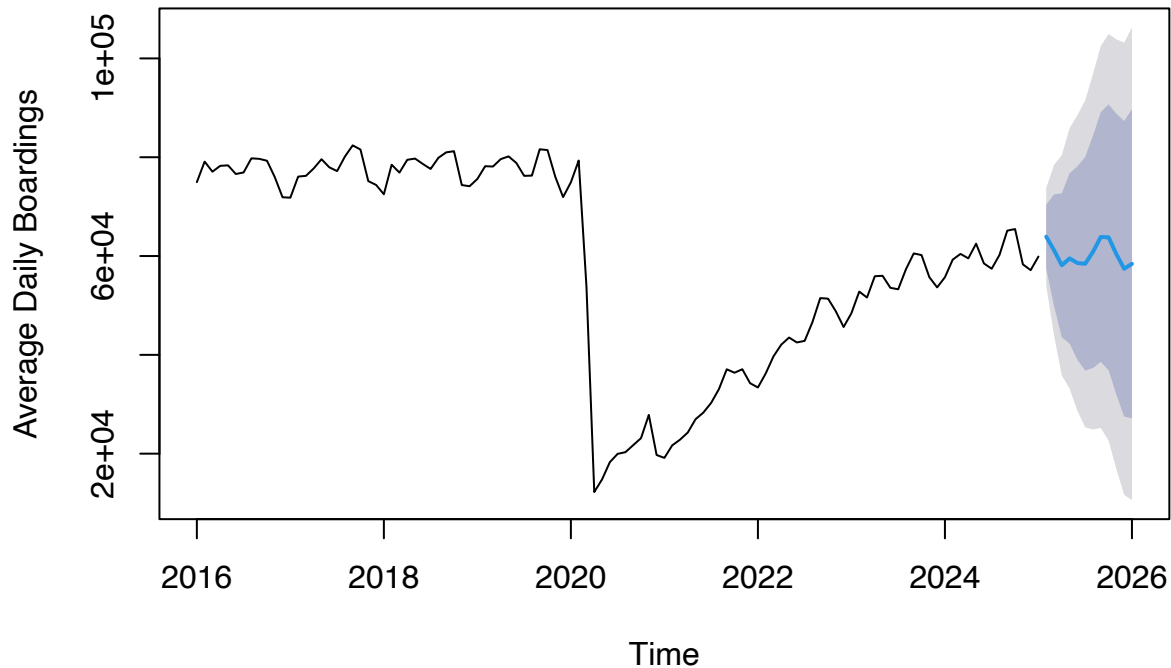
SARIMA Model Analysis

I fitted a Seasonal AutoRegressive Integrated Moving Average (SARIMA) model with the parameters $(0,1,1)(0,1,1)[12]$, which effectively captures both seasonal and non-seasonal dependencies in ridership data.



After fitting the SARIMA(0,1,1)(0,1,1) model, I carefully examined the residuals to ensure the model's adequacy in capturing the underlying patterns within the data. A time series plot revealed mostly random fluctuations, although a notable outlier was observed in 2020, coinciding with the onset of the COVID-19 pandemic and associated lockdowns. The ACF plot indicated some statistical significance at a 12-month lag, but otherwise, the residuals showed no significant autocorrelation, suggesting that the model effectively captured most linear dependencies. Additionally, the histogram indicated a generally normal distribution of residuals, with a slight negative skew potentially related to the outlier in 2020. Overall, these diagnostics suggest that the SARIMA model adequately explains the variance in the observed data, and that it can be used to make informed predictions on the future ridership of the MUNI system.

Forecast with SARIMA



The SARIMA forecast depicted in the graph provides insights into average daily boardings from 2016 to 2026. The historical data, represented by the black line, shows a steady trend until 2020, where there is a sharp decline likely caused by a significant external event, such as the COVID-19 pandemic. Following this drop, the data reflects gradual recovery and stabilization through 2025. The forecast, shown as a blue line, predicts relatively stable average daily boardings from 2025 to 2026, with minimal fluctuations. Surrounding the forecast are confidence intervals shaded in darker and lighter blue, indicating varying levels of uncertainty that increase as the prediction extends into the future. This suggests that while the model anticipates stability, there is some degree of unpredictability in longer-term projections. Overall, the SARIMA model captures historical trends and seasonality effectively while providing a cautious outlook for future boardings.

The model parameters include a non-seasonal moving average (MA) term of 0.4299 and a seasonal MA term of -1.0000. The model's error variance (σ^2) is estimated at 22,847,241, and the log-likelihood value is -962.83, with an AIC (Akaike Information Criterion) of 1931.66.

Training set error metrics suggest that the model has a reasonable fit to the historical data, with a Mean Error (ME) of 97.37, RMSE (Root Mean Squared Error) of 4485.88, and MAE (Mean Absolute Error) of 2216.16. The model's performance is acceptable, given the MAPE (Mean Absolute Percentage Error) of 7.47%, indicating that the model's forecast errors, on average, are within a 7.47% margin of the actual values.

The Ljung-Box test, which tests for autocorrelation in the residuals, returns a p-value of 0.3157, suggesting that there is no significant autocorrelation in the residuals, and the model fits the data well without any significant model mis-specifications.

Table 1: Forecasted Ridership with Confidence Intervals (Rounded)

Month	Forecast	80% CI Lower	80% CI Upper	95% CI Lower	95% CI Upper
Feb 2025	63915	57452	70378	54031	73799
Mar 2025	61187	49917	72458	43951	78424
Apr 2025	58131	43562	72701	35850	80413
May 2025	59522	42274	76771	33143	85902
Jun 2025	58587	39022	78151	28666	88508
Jul 2025	58425	36791	80058	25339	91511
Aug 2025	60858	37337	84380	24885	96831
Sep 2025	63856	38587	89125	25211	102501
Oct 2025	63793	36891	90696	22649	104938
Nov 2025	60377	31934	88820	16877	103877
Dec 2025	57426	27522	87331	11692	103161
Jan 2026	58430	27140	89721	10576	106285

The forecasted ridership values from February 2025 to January 2026 fluctuate between 57,426 and 63,856, with noticeable seasonal variations. For example, February 2025 shows the highest forecast of 63,915, while December 2025 is the lowest at 57,426.

The 95% confidence intervals (CIs) indicate wider uncertainty compared to the 80% CIs, with the 95% CI for February 2025 ranging from 54,031 to 73,799, reflecting greater variability in long-term forecasts. In contrast, months like October 2025 show more reliable forecasts with narrower CIs.

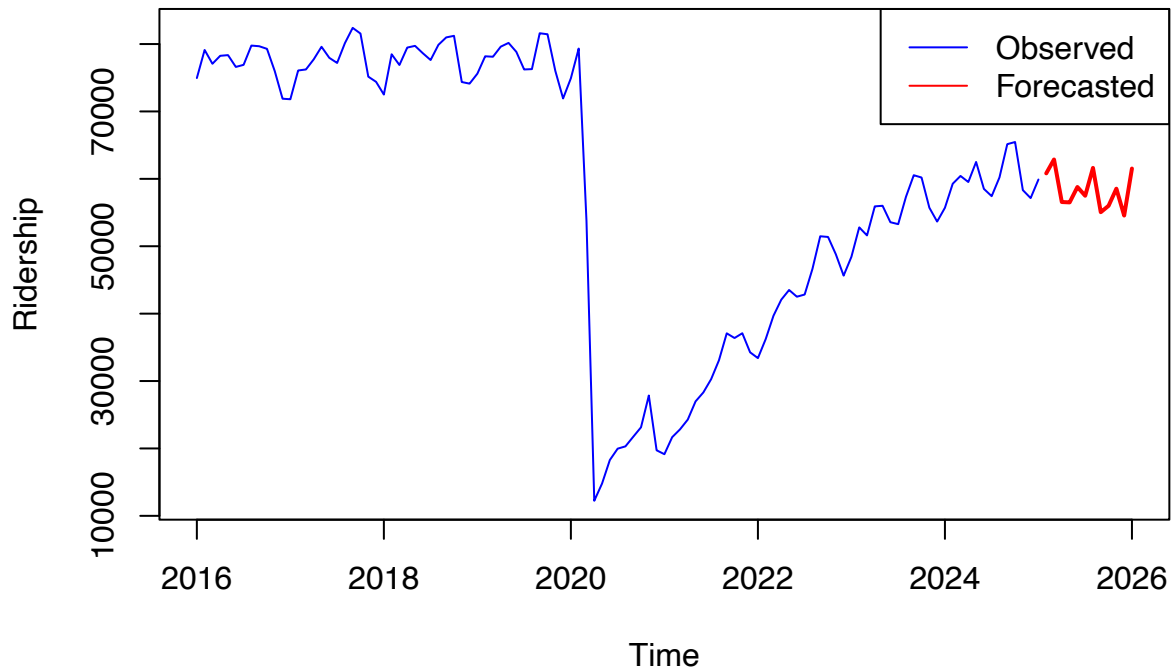
The 80% CIs provide a tighter range, offering a more focused prediction with higher confidence. For instance, February 2025 has an 80% CI between 57,452 and 70,378.

Overall, the forecast suggests stable ridership with moderate fluctuations, but the wide confidence intervals highlight uncertainty, particularly in months with higher variability like March and December 2025.

In conclusion, the SARIMA model provides reliable forecasts for the upcoming months, with good performance on historical data. The forecast intervals offer a reasonable range of uncertainty, which is crucial for planning purposes, especially in predicting the future ridership fluctuations.

Lagged Regression Model Analysis

SF Muni Ridership with Lagged Regression Forecast

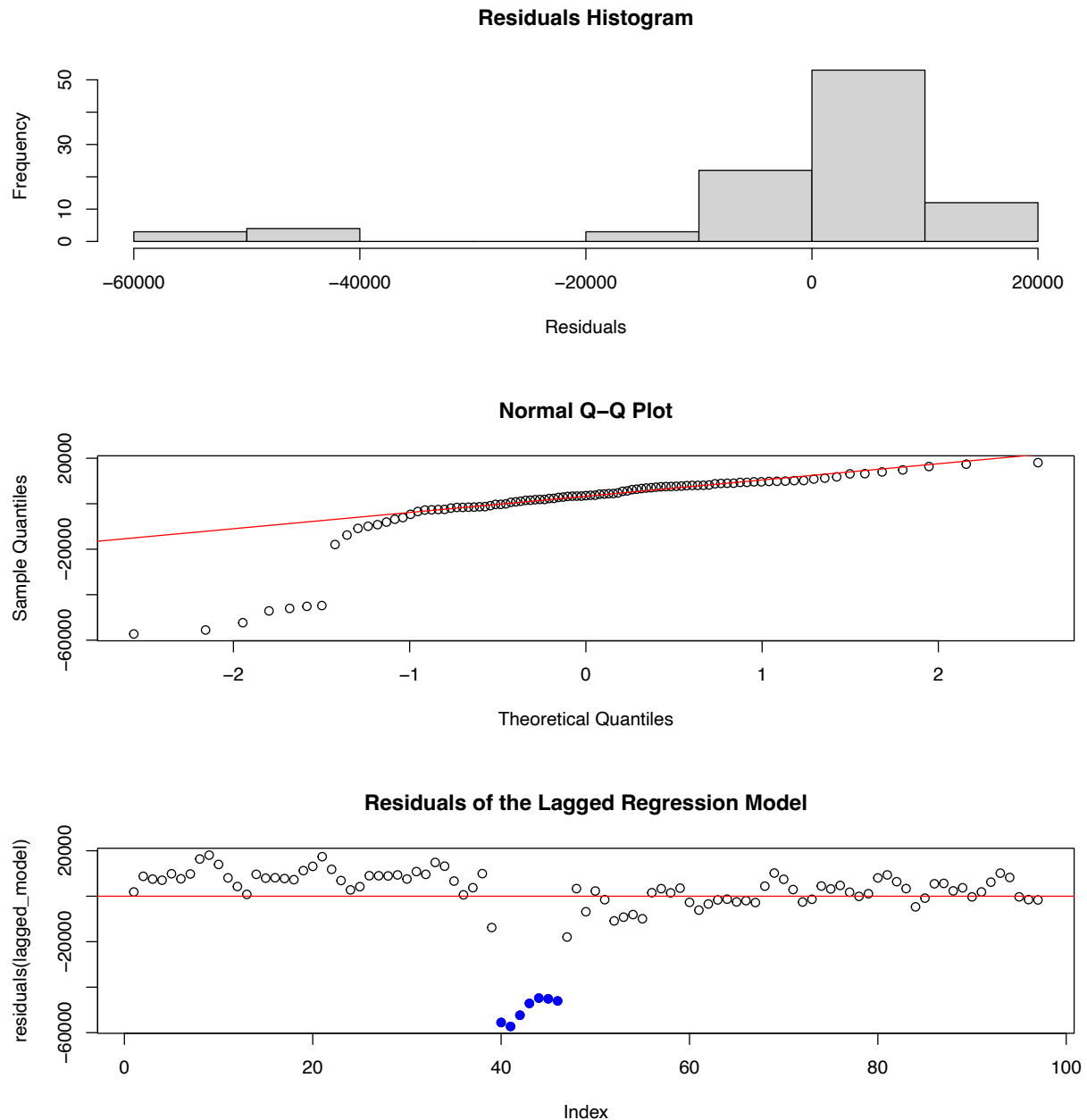


The lagged regression forecast for SF Muni ridership provides an analysis of observed and predicted trends from 2016 to 2026. The blue line represents historical ridership data, showing consistent levels until a sharp decline in 2020, likely due to the COVID-19 pandemic. Following this drop, ridership gradually recovers, approaching pre-pandemic levels by 2025. The red line illustrates the forecasted ridership for 2025 to 2026, indicating a continuation of stable recovery with slight fluctuations. Unlike the SARIMA model, this forecast does not include confidence intervals, making it less explicit about uncertainty. However, it aligns closely with historical trends and suggests steady ridership levels without significant growth or decline. Overall, the lagged regression forecast captures the recovery trajectory effectively and provides a straightforward outlook for future ridership.

The coefficients for the lagged variables suggest that recent trends in ridership (within 1-2 months) have a more substantial effect than older trends. The R-squared value indicates that a significant portion of the variance in current ridership can be explained by past ridership values, with the model providing a reasonable fit.

Residuals analysis showed no major autocorrelation, implying that the model effectively captures the dependencies in the data without missing crucial patterns. However, the model's p-values for some lags indicate potential room for improvement, suggesting the inclusion of additional variables or interaction terms might enhance forecasting accuracy.

Model Diagnostics



The diagnostics for the lagged regression model were conducted to assess the model's validity and performance. The histogram of the residuals reveals a roughly normal distribution, which suggests that the errors of the model are symmetrically distributed around zero, supporting the assumption of normally distributed residuals. Additionally, the Q-Q plot mostly follows the red reference line, indicating that the residuals align with a normal distribution. However, there are some tail outliers present, signaling that the model might not fully capture the extreme values in the data. While the residual plot appears well-behaved overall, there are two notable outliers in blue, which deviate from the rest of the residuals. These outliers could represent unusual events or data points that the model does not adequately address. Despite these outliers, the general pattern of the residuals suggests that the model is a reasonable fit for the data, and the assumptions of normality hold in most instances. However, the presence of these outliers may warrant

further investigation or consideration of alternative modeling approaches to better accommodate such deviations.

Conclusion

In this project, we aimed to develop and analyze forecasting models for ridership data using both Seasonal AutoRegressive Integrated Moving Average (SARIMA) and lagged regression models. The objective was to understand the temporal dependencies in ridership trends and develop models that could reliably predict future ridership. The analysis revealed that the SARIMA model successfully captured both seasonal and non-seasonal dependencies, producing predictions that aligned well with the observed patterns in the data. The forecasted ridership for the upcoming months demonstrated cyclical fluctuations based on historical trends, and the model's confidence intervals provided valuable insights into the range of potential future ridership values. In addition, the SARIMA model proved effective in accounting for the cyclical nature of the data, allowing for robust short-term predictions.

On the other hand, the lagged regression model revealed that past ridership values, particularly at lag-1 and lag-2, were strong predictors of future ridership. This model explained a significant proportion of the variance in the data, underlining the importance of temporal dependencies. Moreover, we examined the residuals for autocorrelation, and the lack of significant patterns suggested that the model adequately captured the underlying structure of the data. The lagged regression model, however, indicated that incorporating additional explanatory variables or considering non-linear terms might enhance prediction accuracy further.

The key takeaway from this analysis is that both models confirmed that historical ridership plays a crucial role in forecasting future trends. The SARIMA model effectively captured seasonal effects, while the lagged regression model focused on direct past relationships. Despite the strong performance of both models, there are areas for further refinement. Future research could explore incorporating exogenous variables such as weather, special events, and holidays, which could have a significant impact on ridership patterns. Adding these external factors to the models could improve accuracy, particularly for non-regular or anomalous ridership trends. Additionally, investigating more advanced techniques such as seasonal decomposition and non-linear models could potentially uncover more subtle relationships that the current models might miss. A more in-depth comparison of time series models, such as ARIMA, SARIMA, and Exponential Smoothing, could provide a better understanding of which model is most suitable for forecasting ridership under different conditions.

In conclusion, this project demonstrated the importance of historical data and temporal dependencies in forecasting ridership. While the models presented here are effective for short-term predictions, future work will focus on enhancing the models by incorporating external data sources, refining the seasonal components, and exploring alternative modeling techniques to improve long-term forecasting accuracy.