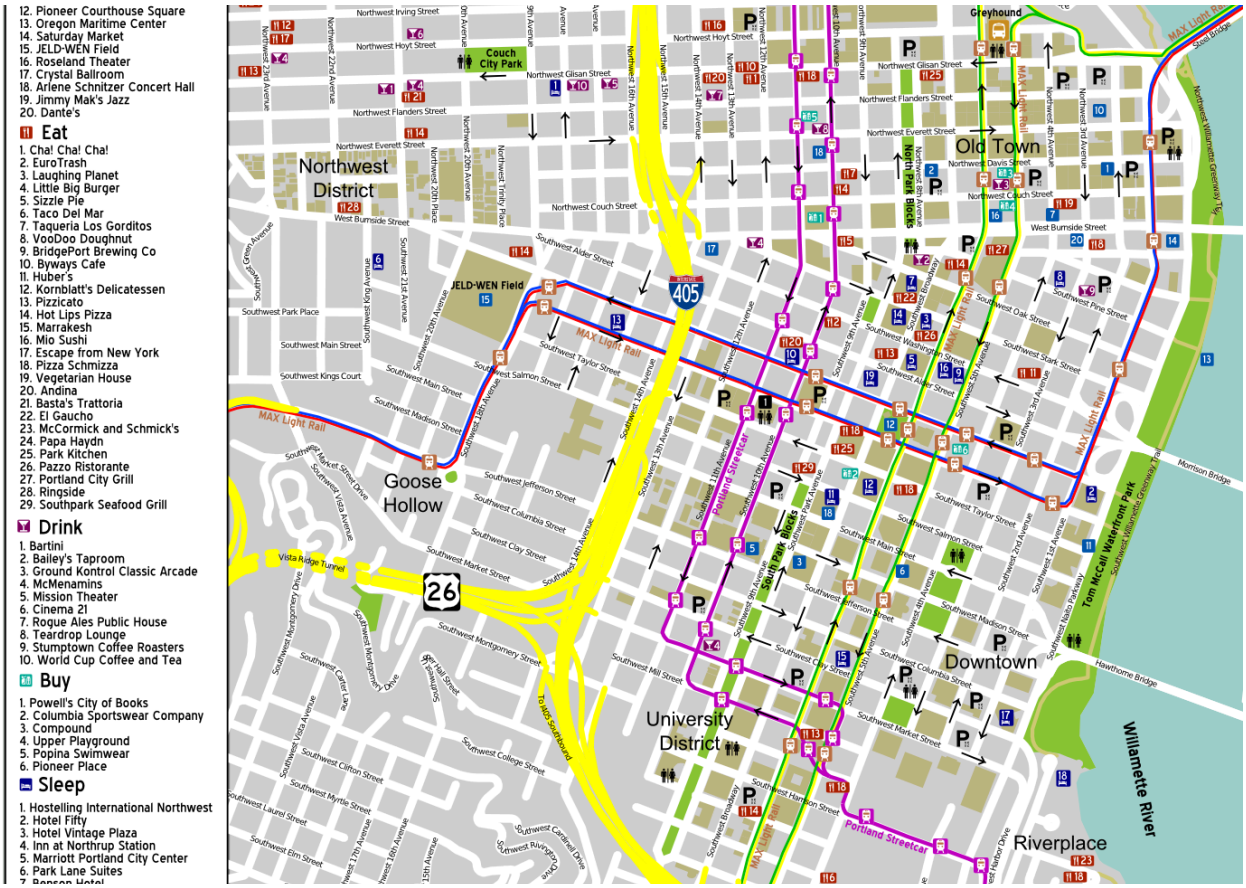
Portland Bus Ridership Project

Modeling Demand Patterns By Analyzing Historical Ridership Data For Public Transit Planning



Sam Sithimolada

Tools used: JMP Pro Software

Techniques: Time Series Analysis, SARIMA Modeling

EXECUTIVE SUMMARY

This report examines monthly bus ridership trends in Portland, Oregon, from 1960 to 1968 to help forecast future demand. The goal was to identify long-term patterns and seasonality to make informed predictions about ridership over the next six months.

The data showed a steady upward trend over time, with noticeable seasonal fluctuations, especially during the same months each year. To prepare the data for forecasting, adjustments were made to remove long-term growth and recurring seasonal effects.

Several forecasting models were compared, and the most accurate one was selected based on how well it fit past data and its ability to make reliable predictions. The final model projected continued growth in bus ridership into 1969, along with expected seasonal ups and downs.

These insights can help city planners, transportation authorities, and policymakers make data-informed decisions about service levels, budgeting, and scheduling for Portland's public transit system.

DATA OVERVIEW

- **Dataset**: Bus Ridership.xlsx
- **Location**: Portland, Oregon
- **Frequency**: Monthly
- **Time Frame**: January 1960 – December 1968
- **Variable**: Average monthly bus riders per 100 people

There are 108 observations in total. The data showed both upward trend and recurring seasonal spikes, indicating that patterns repeat yearly.

INITIAL OBSERVATIONS

Time Series Visualization

The original time series plot revealed a clear upward trend in bus ridership. Starting from lower values in 1960, ridership consistently increased over the years, with

fluctuations that repeated annually *(see Appendix A, Figure A1)*.

Seasonality

Seasonal peaks and dips suggested a strong annual cycle, where ridership was consistently higher or lower in certain months.

TESTING FOR STABILITY OVER TIME

Before building a forecast model, it was important to assess whether the series was stable over time ("stationary"). A standard statistical test (Augmented Dickey-Fuller test) was applied (*see Appendix A, Figures A2-A4)*.

Key Findings:

- The original series was **not stable** due to the presence of a trend and seasonality.
- A transformation (called differencing) was applied to remove the trend.
- A second transformation (seasonal differencing) was applied to remove repeating yearly patterns.
- After these adjustments, the data passed the stationarity test and was ready for modeling.

MODEL DEVELOPMENT

A range of forecasting models was considered, focusing on Seasonal ARIMA (SARIMA) models. These are well-suited for data with both trends and seasonal cycles.

Candidate Models Considered *(see Appendix A, Figure A5)*:

- SARIMA(0,1,0)x$(1,1,1)_{12}$
- SARIMA(0,1,0)x$(2,1,1)_{12}$
- SARIMA(0,1,0)x$(0,1,1)_{12}$

Selection Criteria:

- AIC and SBC: Model selection metrics that reward simplicity and fit
- MAPE: Measures forecast accuracy in percentage terms

Final Model Chosen *(see Appendix A, Figure A6)*:

- **SARIMA(0,1,0)x(1,1,1)$_{12}$**
- Best combination of accuracy and simplicity
- All key parameters were statistically significant

FORECASTING RESULTS

Using the best-fitting model, ridership was forecasted for the first half of 1969 *(see Appendix A, Figure A7)*:

Month	Forecast	95% CI Lower	95% CI Upper
Jan 1969	1,407	1,345	1,469
Feb 1969	1,415	1,327	1,503
Mar 1969	1,377	1,270	1,485
Apr 1969	1,391	1,267	1,515
May 1969	1,364	1,225	1,502
Jun 1969	1,314	1,162	1,466

CONCLUSION

The Portland bus ridership data exhibited both upward trends and recurring seasonal behavior. After stabilizing the data, seasonal ARIMA modeling produced accurate and interpretable forecasts.

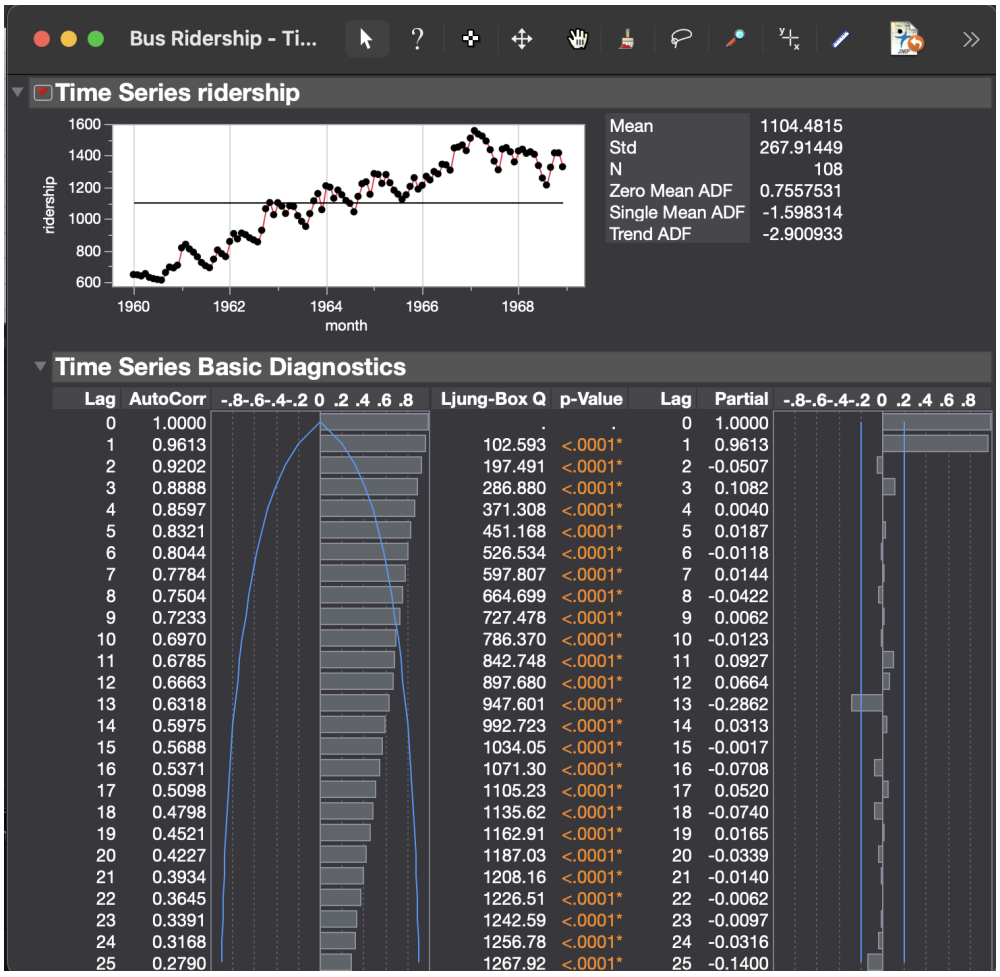
The insights from this analysis can support public transit decisions, such as staffing, fleet

allocation, and budgeting, especially during high-demand months.

APPENDIX A: TIME SERIES DATA MODELING DIAGNOSTICS

A1: Time Series, ACF, PACF Plot of Original Data

From a visual inspection, this data is not stationary because the time series plot shows an increasing trend and therefore the mean is not constant over time. The variance is also not constant as there is more variance towards the recent years, perhaps mid-1966 onward. Also, in the ACF the autocorrelation values gradually decay, indicating that past values are influencing the time series, but their influence diminishes over time which ultimately means that the series' statistical properties are also not constant over time.



Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	0.9613	102.593	<.0001*	1	0.9613
2	0.9202	197.491	<.0001*	2	-0.0507
3	0.8888	286.880	<.0001*	3	0.1082
4	0.8597	371.308	<.0001*	4	0.0040
5	0.8321	451.168	<.0001*	5	0.0187
6	0.8044	526.534	<.0001*	6	-0.0118
7	0.7784	597.807	<.0001*	7	0.0144
8	0.7504	664.699	<.0001*	8	-0.0422
9	0.7233	727.478	<.0001*	9	0.0062
10	0.6970	786.370	<.0001*	10	-0.0123
11	0.6785	842.748	<.0001*	11	0.0927
12	0.6663	897.680	<.0001*	12	0.0664
13	0.6318	947.601	<.0001*	13	-0.2862
14	0.5975	992.723	<.0001*	14	0.0313
15	0.5688	1034.05	<.0001*	15	-0.0017
16	0.5371	1071.30	<.0001*	16	-0.0708
17	0.5098	1105.23	<.0001*	17	0.0520
18	0.4798	1135.62	<.0001*	18	-0.0740
19	0.4521	1162.91	<.0001*	19	0.0165
20	0.4227	1187.03	<.0001*	20	-0.0339
21	0.3934	1208.16	<.0001*	21	-0.0140
22	0.3645	1226.51	<.0001*	22	-0.0062
23	0.3391	1242.59	<.0001*	23	-0.0097
24	0.3168	1256.78	<.0001*	24	-0.0316
25	0.2790	1267.92	<.0001*	25	-0.1400

A2: ADF Test Output (Pre-Differencing)

To be considered stationary, the series has to pass three components of the Augmented Dickey-Fuller test. The Tau statistics of the three tests will have to be more extreme than the critical values in order to reject the null hypothesis and conclude that the series is stationary. With the current Tau statistics, they all fail to reject the null hypothesis because they are no more extreme than their critical value counterparts at a 5% significance level. Because they fail to reject the null hypothesis, the time series is non-stationary.

Hypotheses:

$H_0 : \delta = 0$ $\qquad$ $H_1 : \delta < 0$

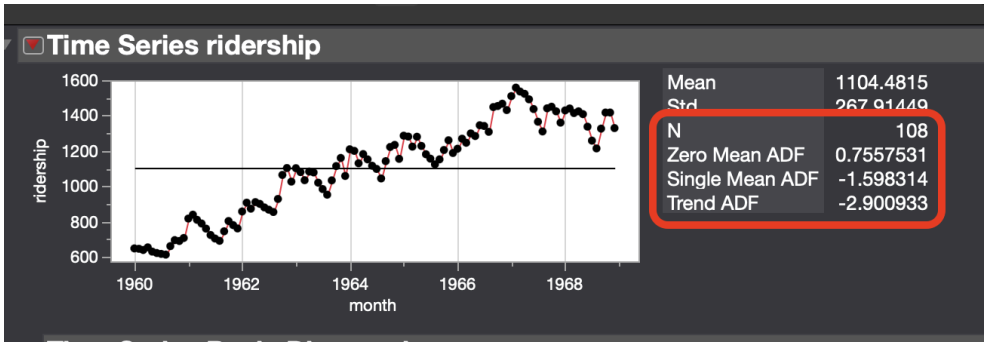
Critical Values:

1% and 5% critical Dickey-Fuller (τ) Values for Unit Root Tests.

Sample Size	t_{nc}^* 1%	t_{nc}^* 5%	t_c^* 1%	t_c^* 5%	t_{ct}^* 1%	t_{ct}^* 5%
25	−2.66	−1.95	−3.75	−3.00	−4.38	−3.60
50	−2.62	−1.95	−3.58	−2.93	−4.15	−3.50
100	−2.60	−1.95	−3.51	−2.89	−4.04	−3.45
250	−2.58	−1.95	−3.46	−2.88	−3.99	−3.43
500	−2.58	−1.95	−3.44	−2.87	−3.98	−3.42
∞	−2.58	−1.95	−3.43	−2.86	−3.96	−3.41

*Subscripts nc, *c*, and ct denote, respectively, that there is no constant, a constant, and a constant and trend term

Tau Statistics:

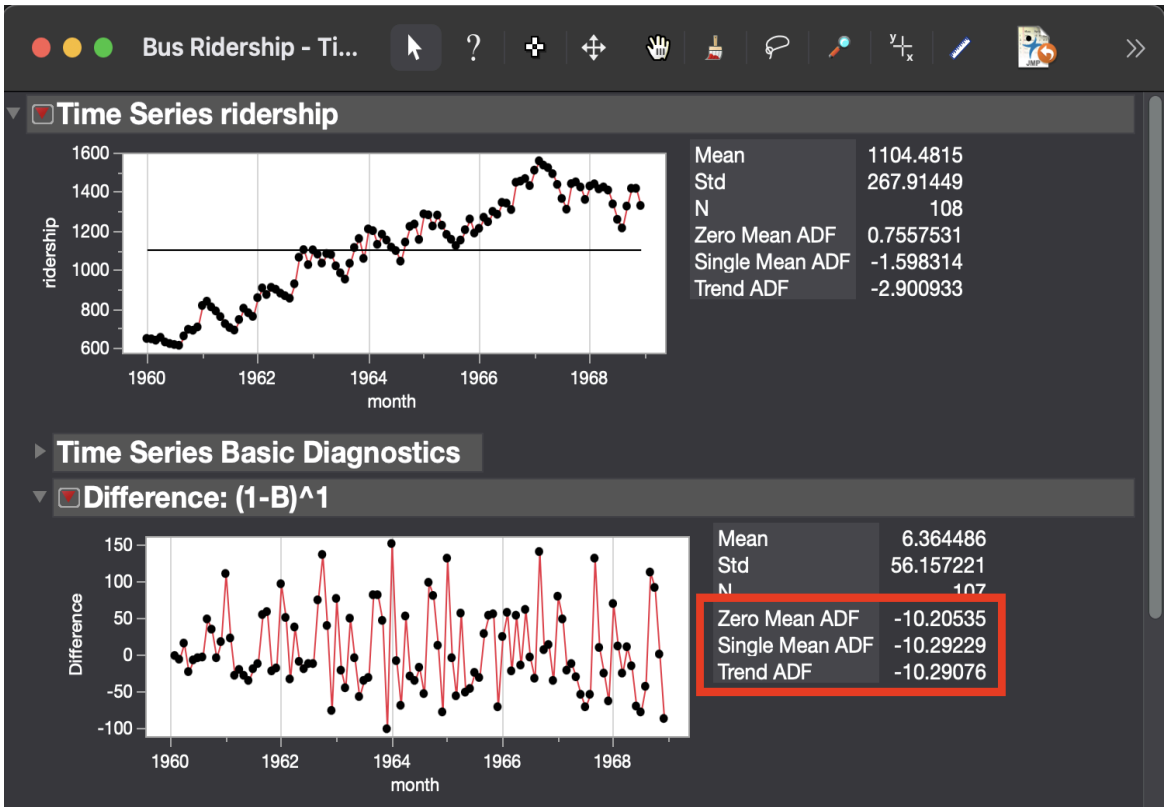


ADF Component	Tau Stat	Critical Value (n = 108)	Tau < CV?	Conclusion

Zero Mean ADF	0.7557	-1.95	No	Fail to reject the null
Single Mean ADF	-1.5983	-2.89	No	Fail to reject the null
Trend ADF	-2.9009	-3.45	No	Fail to reject the null

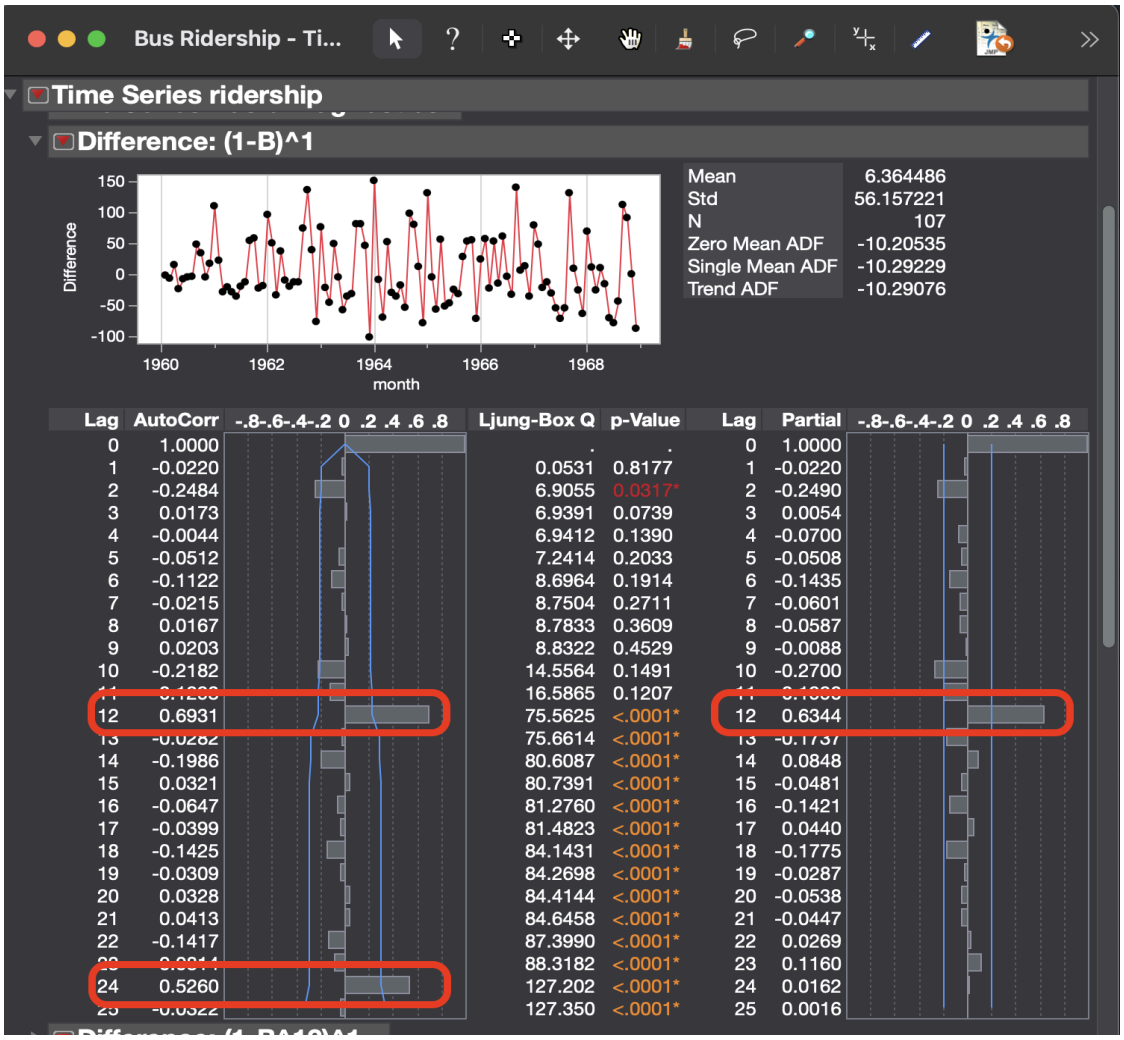
A3: De-trending The Series: Differenced Series Plot

Because we confirmed the series is non-stationary through the Augmented Dickey-Fuller test, we can try to detrend the data by taking the first-order difference or the lag 1 difference, since we identified there was a trend in the data, which undoubtedly affects the series' statistical properties. Taking the first-order difference should help stabilize the series. In the output below, the time series with the first-order difference shows statistically significant Tau statistics, which are more extreme than the critical values for their respective tests. Since they are all statistically significant, we can reject the null hypothesis and conclude the time series is now stationary.



ADF Component	Tau Stat	Critical Value (n = 107)	Tau < CV?	Conclusion
Zero Mean ADF	-10.2054	-1.95	Yes	Reject the null
Single Mean ADF	-10.2923	-2.89	Yes	Reject the null
Trend ADF	-10.2908	-3.45	Yes	Reject the null

Despite being stationary, there is also a seasonal component that needs to be addressed if we want to model this time series. A visual inspection of the correlation function plots reveal statistically significant values at lags of intervals of 12, suggesting an annual seasonal component (every 12 months, since this is monthly data). So, there is a correlation between the same calendar month from across different years.

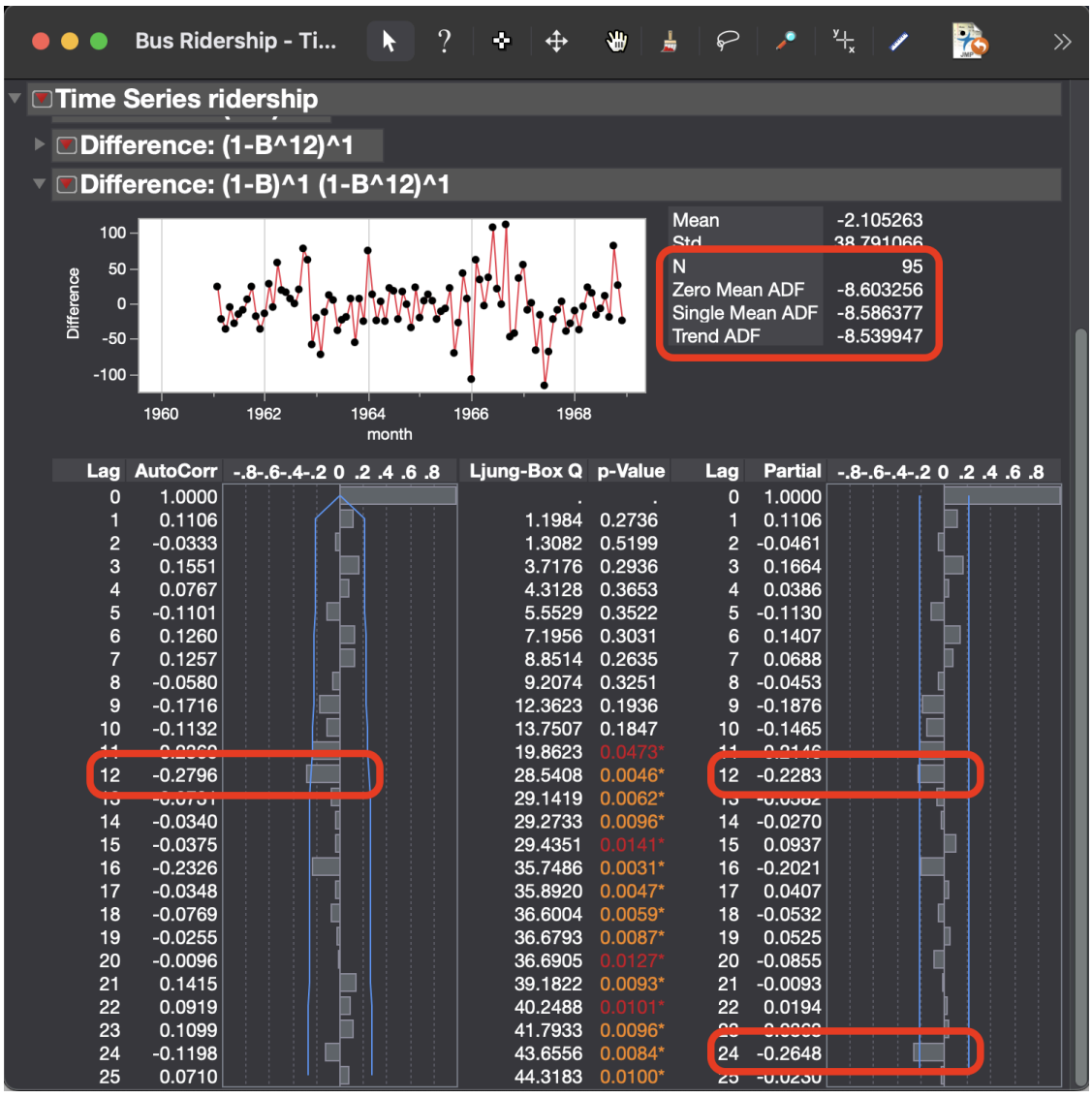


A4: De-seasoning the De-trended Series

From the de-seasoned and de-trended data (d = 1, D =1) we have correlation function plots with a stationary time series since all Tau statistics are more extreme than their critical values and no significant lags before the seasonal lag (lag 12) for the non-seasonal component. However, from the seasonal component (looking at intervals of lag 12) we see lags barely above the thresholds at lag 12 for the ACF and lag 12 and 24 for the PACF, suggesting, at most, a seasonal MA1 and AR2. From this, we can base our model on different parameters.

Non-seasonal aspect: -

Seasonal aspect: MA0, MA1, AR0, AR1, AR2



ADF Component	Tau Stat	Critical Value (n = 95)	Tau < CV?	Conclusion
Zero Mean ADF	-8.6033	-1.95	Yes	Reject the null
Single Mean ADF	-8.5864	-2.89	Yes	Reject the null
Trend ADF	-8.5399	-3.45	Yes	Reject the null

A5: Model Fit and Evaluation

SARIMA (p, d, q)x(P, D, Q)s Model Parameters:
p: 0
d: 1
q: 0
P: [0, 2]
D: 1
Q: [0, 1]
s: 12

These are the top 3 models with low AIC and SBC metrics:

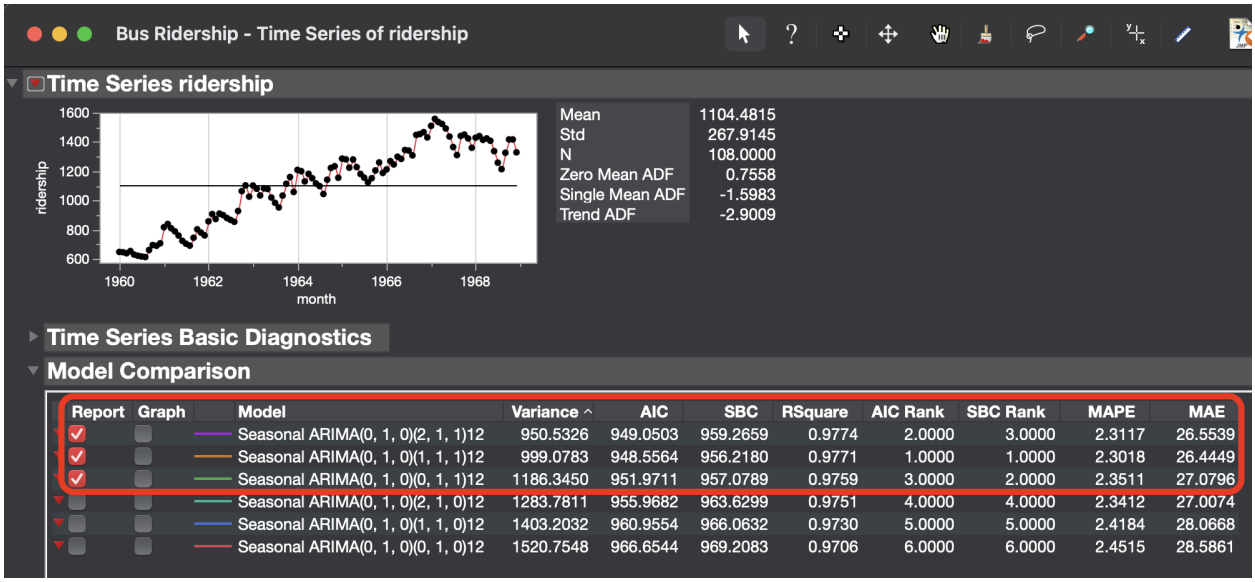
Top 3 Models:

SARIMA(0, 1, 0)x(1, 1, 1)$_{12}$
SARIMA(0, 1, 0)x(2, 1, 1)$_{12}$
SARIMA(0, 1, 0)x(0, 1, 1)$_{12}$

SARIMA(0, 1, 0)x(1, 1, 1)$_{12}$ is clearly the best as it has the lowest AIC (948.5564), SBC (956.2180), and MAPE (2.3018%) metrics. Between the other two SARIMA models, SARIMA(0, 1, 0)x(2, 1, 1)$_{12}$ is the second best because although it interchanges ranks between the AIC and SBC with model SARIMA(0, 1, 0)x(0, 1, 1)$_{12}$, making them comparably similar, SARIMA(0, 1, 0)x(2, 1, 1)$_{12}$ is better between the two because it has a lower MAPE, 2.3117% vs 2.3511%.

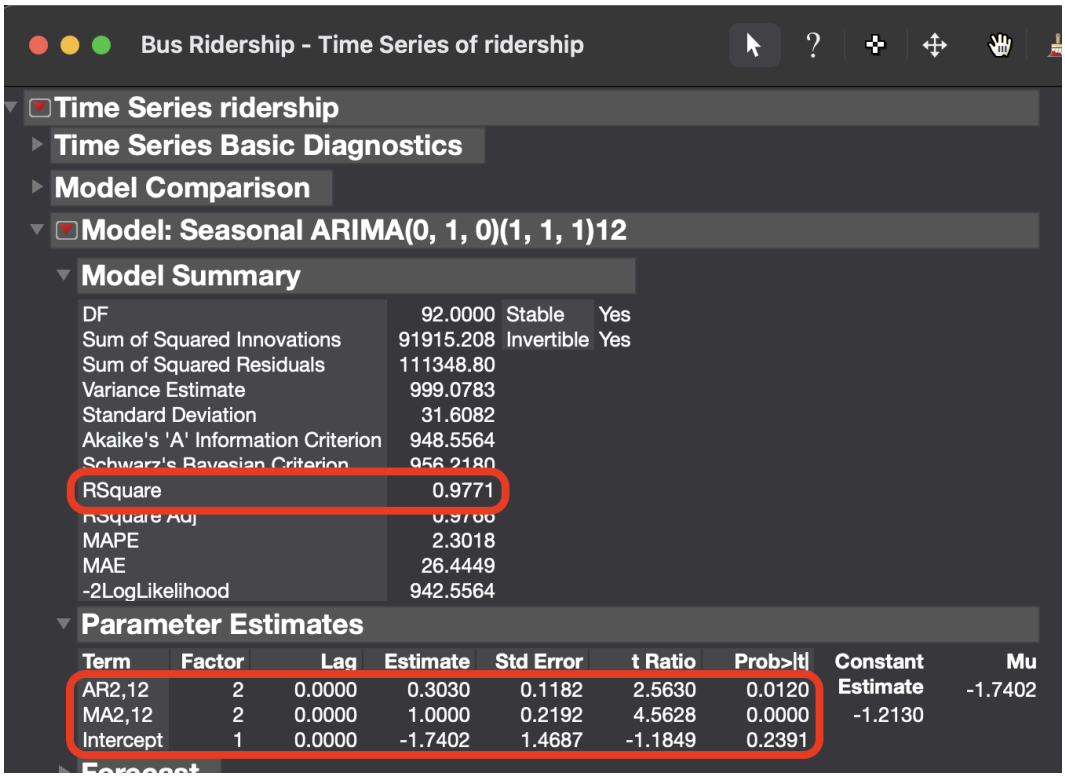
Top 2 Models:

SARIMA(0, 1, 0)x(1, 1, 1)$_{12}$
SARIMA(0, 1, 0)x(2, 1, 1)$_{12}$

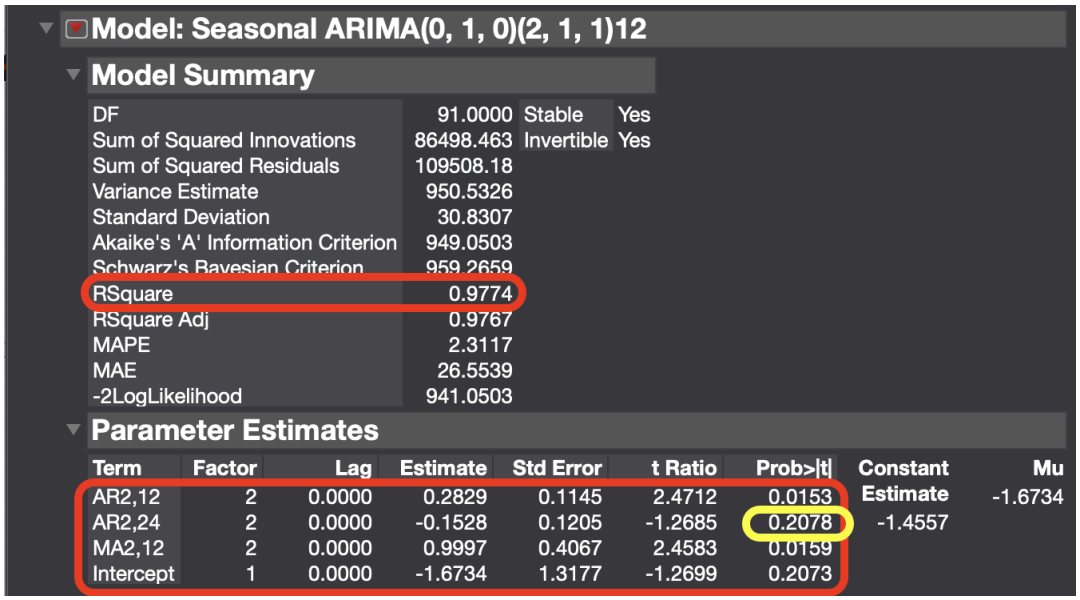


A6: Top Two Model Fit and Parameter Estimates

SARIMA(0, 1, 0)x(1, 1, 1)$_{12}$



SARIMA(0, 1, 0)x(2, 1, 1)$_{12}$



Model: Seasonal ARIMA(0, 1, 0)(2, 1, 1)12

Model Summary

DF	91.0000	Stable	Yes
Sum of Squared Innovations	86498.463	Invertible	Yes
Sum of Squared Residuals	109508.18		
Variance Estimate	950.5326		
Standard Deviation	30.8307		
Akaike's 'A' Information Criterion	949.0503		
Schwarz's Bayesian Criterion	959.2659		
RSquare	0.9774		
RSquare Adj	0.9767		
MAPE	2.3117		
MAE	26.5539		
-2LogLikelihood	941.0503		

Parameter Estimates

Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|	Constant Estimate	Mu
AR2,12	2	0.0000	0.2829	0.1145	2.4712	0.0153		-1.6734
AR2,24	2	0.0000	-0.1528	0.1205	-1.2685	0.2078	-1.4557	
MA2,12	2	0.0000	0.9997	0.4067	2.4583	0.0159		
Intercept	1	0.0000	-1.6734	1.3177	-1.2699	0.2073		

Between the two models, I would select ***SARIMA(0, 1, 0)x(1, 1, 1)$_{12}$*** because although it has similar R^2 scores (0.9771 vs 0.9774), this model has a better AIC, SBC, and MAPE score which are all lower between the two models. Also, model SARIMA(0, 1, 0)x(2, 1, 1)$_{12}$ has a statistically insignificant parameter, AR2, 24, at a 5% significance level–its p-value is 20.78%, whereas, SARIMA(0, 1, 0)x(1, 1, 1)$_{12}$ has all statistically significant parameter estimates aside from the intercept which could be expected as it becomes less irrelevant when we used the first-order difference to remove the trend and lag-12 difference to remove seasonality. Moreover, the insignificant AR2, 24 parameter estimate means that adding the AR2 in the SARIMA model (making it more complex) did not provide any more explanatory power, which can be seen in the R^2 scores (0.9771 vs 0.9774).

Best Model:

SARIMA(0, 1, 0)x(1, 1, 1)$_{12}$

A7: Final Forecast with Confidence Intervals

untitled 15

	Actual ridership	month	Predicted ridership	Std Err Pred ridership	Residual ridership	Upper CL (0.95) ridership	Lower CL (0.95) ridership	
98	1440	1968-02-01	1450.0514743	31.608200571	-10.05147427	1512.002409	1388.1005395	
99	1414	1968-03-01	1403.7243535	31.608200571	10.275646501	1465.6752882	1341.7734188	
100	1424	1968-04-01	1422.0628463	31.608200571	1.9371537168	1484.013781	1360.1119115	
101	1408	1968-05-01	1391.3400272	31.608200571	16.659972805	1453.2909619	1329.3890925	
102	1337	1968-06-01	1367.5600587	31.608200571	-30.5600587	1429.5109934	1305.609124	
103	1258	1968-07-01	1291.4031507	31.608200571	-33.40315068	1353.3540854	1229.4522159	
104	1214	1968-08-01	1214.7888204	31.608200571	-0.788820383	1276.7397551	1152.8378856	
105	1326	1968-09-01	1305.2141916	31.608200571	20.785808404	1367.1651263	1243.2632569	
106	1417	1968-10-01	1360.1256996	31.608200571	56.874300373	1422.0766344	1298.1747649	
107	1417	1968-11-01	1410.8020567	31.608200571	6.197943311	1472.7529914	1348.851122	
108	1329	1968-12-01	1356.0682644	31.608200571	-27.06826441	1418.0191991	1294.1173297	
109	•	1968-12-31	1407.8754194	31.608200571	•	1469.8263542	1345.9244847	
110	•	1969-01-30	1415.4860551	44.70074593	•	1503.0979072	1327.874203	
111	•	1969-03-01	1377.8267417	54.747009325	•	1485.1289082	1270.5245751	
112	•	1969-03-31	1391.8408236	63.216401142	•	1515.7426931	1267.9389541	
113	•	1969-04-30	1364.2002284	70.678085124	•	1502.7267298	1225.6737271	
114	•	1969-05-30	1314.6034867	77.423963087	•	1466.3516659	1162.8553075	
115	•	1969-06-29	1262.2883607	83.627438102	•	1426.1951275	1098.381594	
116	•	1969-07-29	1221.1167422	89.40149186	•	1396.3404464	1045.893038	