analysis_b

Benjamin Leidig, Monte Thomas, Harmony Pham

2025-04-24

R Markdown

Methods

Although the original dataset includes data on historical hotel occupancy rates, ADR, and daily revenue between April, 2024 and April, 2025, only the data on daily total revenue was retained and used throughout the analysis.

Daily total revenue was plotted over time (represented in days) and the plot was examined for stationary behavior. Visual examination included inspecting whether the variance remained relatively constant over time and if the mean function remained relatively constant over time. Inspecting the mean function also included investigating obvious trends. Alongside visual inspection, more rigorous statistical tests were used to conclude stationarity. An Augmented Dickey-Fuller (ADF) test was conducted, with the null and alternative hypotheses being:

$$H_0 : \text{The time series has a unit root.}$$

$$H_1 : \text{The time series doesn't have a unit root.}$$

Effectively, this means the null hypothesis represents non-stationarity whereas the alternative hypothesis represents stationarity. This test was conducted with a significance level of 0.05. Thus, if the test yielded a p-value below 0.05, it support the conclusion that the series is stationary. Ultimately, however, determining stationarity wasn't definitively concluded using statistical testing alone, although this method provided useful insights and supported the conclusion made by visual inspection.

The autocorrelation function (ACF, $\rho(h)$) was then examined for significance at various lags, h, where statistical significance at the 95% confidence interval is determined approximately by $\pm\frac{2}{\sqrt{n}}$. This inspection was carried through until $h = 100$. To determine the presence of seasonality in the time series, the ACF was also examined for significance and/or patterns at seasonal lags, $h = sk$, where $k \in \mathbb{Z}^+$, k is a seasonal lag, and s is the seasonal component. Significant ACF values at multiple lags of h in intervals of length s were used to determine seasonality and, if seasonal, to determine the seasonal component s.

The partial autocorrelation function (PACF, ϕ_{hh}) was also examined for significance at various lags, h, where statistical significance at the 95% confidence level is determined approximately by $\pm\frac{2}{\sqrt{n}}$, and performed until $h = 100$. Again, the PACF was also inspected at seasonal lags, k, as well.

To determine potential orders for various candidate models, the ACF and PACF were again examined.

It can be shown that for autoregressive models of order p (AR(p)), the ACF "tails out" to insignificance while the PACF cuts off (abrupt insignificance) for lags greater than p ($h > p$). Similarly, for pure seasonal autoregressive models of order P (SAR(P)), the ACF tails out while the PACF cuts off for seasonal lags greater than P ($k > P$).

The inverse is true for moving-average models of order q (MA(q))–that is, the ACF cuts off for lags greater than q while the PACF tails out. Similarly, for pure seasonal moving-average models of order Q (SMA(Q)), the ACF cuts off for seasonal lags greater than Q while the PACF tails off.

However, for autoregressive moving-average models of orders p and q (ARMA(p, q)) and pure seasonal autoregressive moving-average models of orders P and Q (SARMA(P, Q)), both the ACF and PACF tail out (Shumway, Stoffer 94, 148). Accordingly, the behavior of the ACF and PACF of the time series was analyzed and several candidate models of different orders were fitted. Since ARMA and pure SARMA model orders cannot be determined precisely via ACF and PACF plot inspection, the orders were ambiguous, and several candidate models were created with various orders.

Each candidate model was investigated for parameter significance of each parameter estimate at the 95% confidence level. Parameters for each candidate model were estimate using the Gauss-Newton method. Models with insignificant parameter estimates were disregarded. The models kept after the first screening process were then compared using AIC, AICc, and BIC scores, the lowest scores of which are preferred. The candidates were then ranked according to these scores, and the model with the lowest BIC was chosen. To ensure robustness of analysis, a residual analysis was conducted on the chosen model, including a Ljung-Box test with the respective null and alternative hypotheses:

$$H_0 : \text{The residuals don't exhibit autocorrelation.}$$
$$H_1 : \text{The residuals exhibit autocorrelation.}$$

So, p-values above 0.05 are preferred, as an assumption in the modeling is uncorrelated, normally distributed errors with mean zero ($W_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_w^2)$). Alongside this test, the residuals were plotted over time to ensure a zero mean and constant variance. Similarly to how the original time series was investigated, an ADF test was conducted to support the conclusion made on the nature of the residuals in order to conclude zero mean and constant variance, abiding by the aforementioned stipulation of the model assumption.

Five future values were forecasted according to the chosen model using the truncated n-step-ahead forecasting method. The full time series plus the five forecasted future values and associated 80% and 95% prediction intervals were plotted.

Results

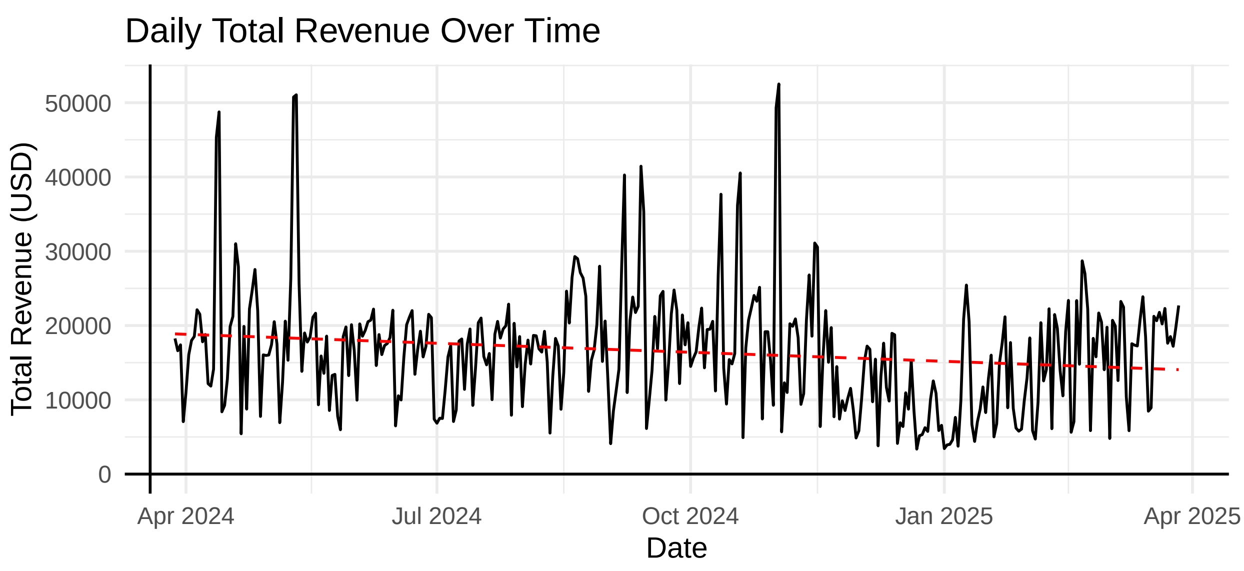
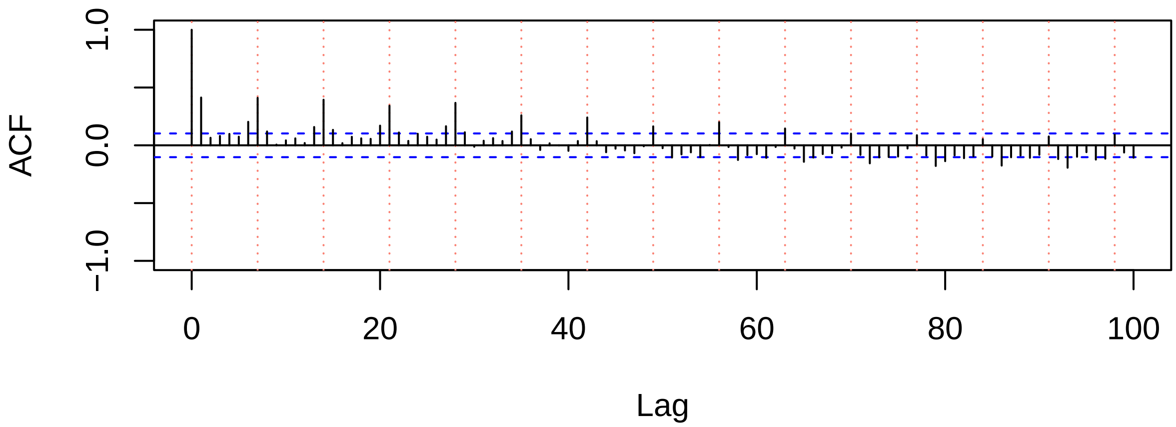


Figure FIGNUMBER demonstrated a relatively stable mean function and non-constant variance. The mean function appeared to increase in the month of October, 2024, and decreased into the months of December, 2024, and January, 2025. The overlaid trend on the time series plot exemplifies the lack of a consistent, non-zero, linear trend in the time series, which suggested that differencing or detrending the data would be unnecessary. The variance function increased drastically in the months of May, September, and November, 2024.

An ADF test yielded a p-value below the 0.05 significance level, which supported the conclusion that this time series was stationary.



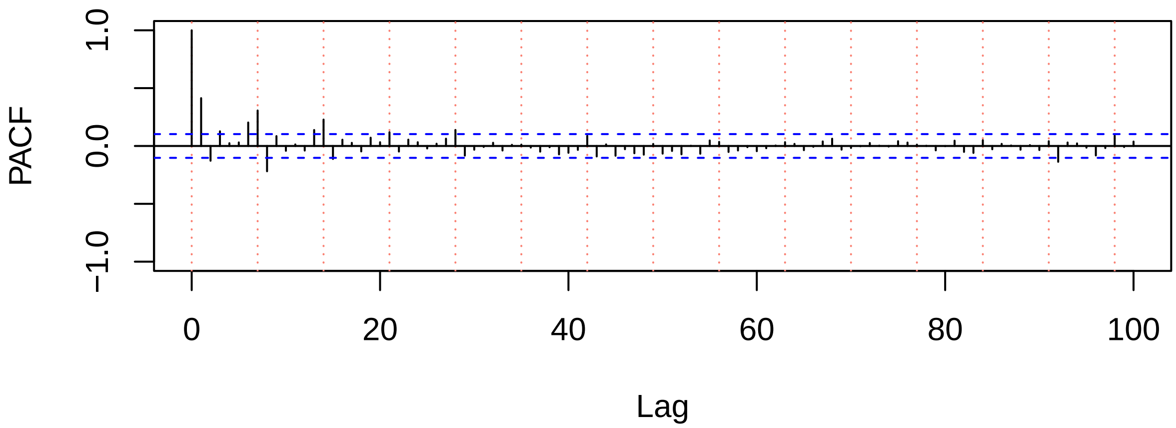


Table TABLE NUMBER exhibited the ACF and PACF of the time series. The various ACF and PACF values for different lags can be interpretable in several ways.

One interpretation was that the ACF spiked at lag $h = 1$, and cut off to 0 for non-seasonal lags afterward. The ACF at seasonal lags tailed out to insignificance. The PACF at non-seasonal lags and seasonal lags tailed out to insignificance. Due to this behavior, the non-seasonal orders can be interpreted as $(\phi = 0, \theta = 1)$, or MA(1) model behavior. The seasonal orders can be interpreted as $(\Phi \geq 1, \Theta \geq 1)_{s=7}$.

A second interpretation was that the ACF at non-seasonal and seasonal lags tailed out to insignificance. The PACF at non-seasonal lags and seasonal lags also tailed out to insignificance. Due to this behavior, the non-seasonal orders can be interpreted as $(\phi \geq 1, \theta \geq 1)$, or ARMA$(\geq 1, \geq 1)$ model behavior. The seasonal orders can be interpreted as $(\Phi \geq 1, \Theta \geq 1)_{s=7}$.

The models of both interpretations were tested with ambiguous orders set to several combinations of 1, 2, and 3 to investigate whether higher orders were necessary in the final model. Models with p, P, and Q orders greater than 2 all had insignificant parameter estimates for the associated order's parameter. No parameter of order 3 or greater from any model was deemed statistically significant.

Accordingly, three models were selected for the final screening phase, each with statistically significant parameter estimates: SARMA$((0, 1) \times (1, 1)_7)$, SARMA$((1, 1) \times (1, 1)_7)$, and SARMA$((1, 2) \times (1, 1)_7)$.

The three aforementioned models were tested using AIC, AICc, and BIC scores in Table 1. SARMA$((1, 2) \times (1, 1)_7)$ obtained the lowest BIC, SARMA$((0, 1) \times (1, 1)_7)$ obtained the second lowest BIC, and SARMA$((1, 1) \times (1, 1)_7)$ obtained the highest BIC. Not only did SARMA$((1, 2) \times (1, 1)_7)$ obtain the

Table 1: Model Performance

Orders	AIC	AICc	BIC
$(1,2)\times(1,1)_7$	20.28101	20.28165	20.35580
$(0,1)\times(1,1)_7$	20.31236	20.31266	20.36578
$(1,1)\times(1,1)_7$	20.30773	20.30818	20.37183

lowest BIC, but it also obtained the lowest AIC and AICc, too. Thus, SARMA$((1,2) \times (1,1)_7)$ was chosen as the final model.

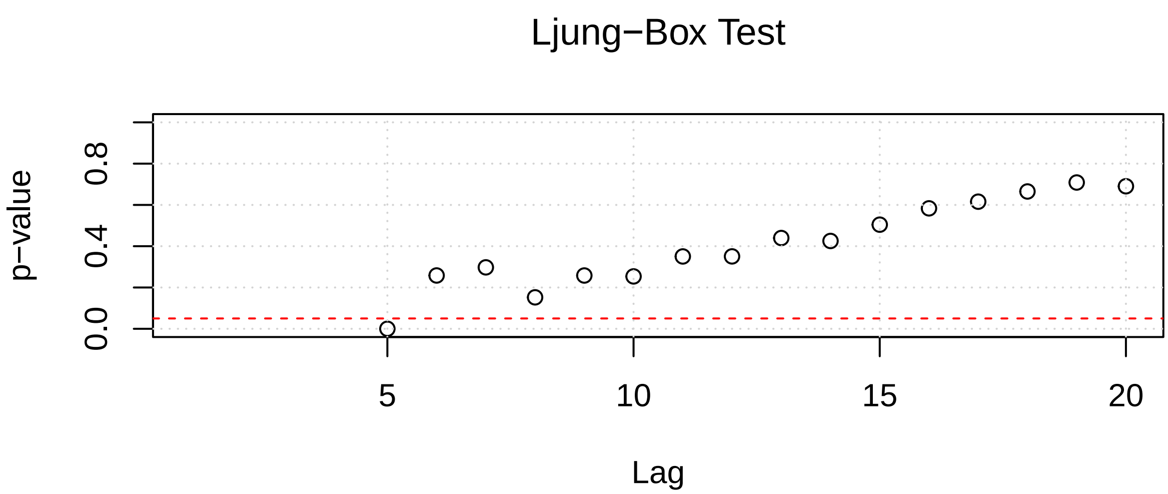
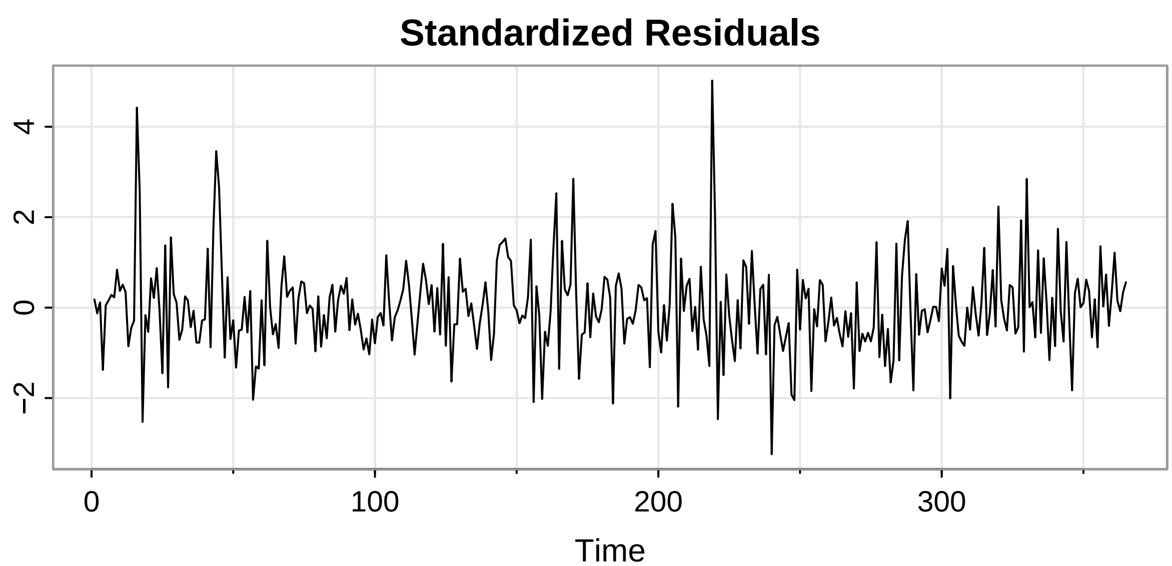


Figure FIGNUMBER showed almost all p-values from the Ljung-Box test to be greater than a significance level of 0.05, failing to reject the null hypothesis that the residuals don't exhibit autocorrelation. So, each observed residual from the model is uncorrelated with every other observed residual.



The standardized residuals of the chosen model appeared to have a constant mean function in Figure FIGNUMBER. However, the variance of the residuals fluctuates mildly, increasing near the start (April, 2024) and middle (October, 2024) of the series. However, an ADF test yielded a p-value below the 0.05 significance level, which suggested the residuals were stationary and thus had constant mean and variance.

Therefore, the residuals of the model were normally distributed with mean zero, which abided by the model assumption $W_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_w^2)$.

Table 2: Model Estimates

Parameter	Estimate	SE	t-value	p-value
$\hat{\phi}$	0.9699	0.0198	48.9615	0
$\hat{\theta}_1$	-0.4487	0.0519	-8.6527	0
$\hat{\theta}_2$	-0.4131	0.0480	-8.6025	0
$\hat{\Phi}$	0.9983	0.0034	289.4971	0
$\hat{\Theta}$	-0.9583	0.0401	-23.8717	0

All of the parameter estimates for the chosen model were significant, as shown in Table 1. Thus, the time series of total daily revenue, $\{X_t\}$, can be most accurately modeled by:

$$X_t = 0.9699X_{t-1} + 0.9983X_{t-7} - 0.4487W_{t-1} - 0.4131W_{t-2} - 0.9583W_{t-7}$$

.

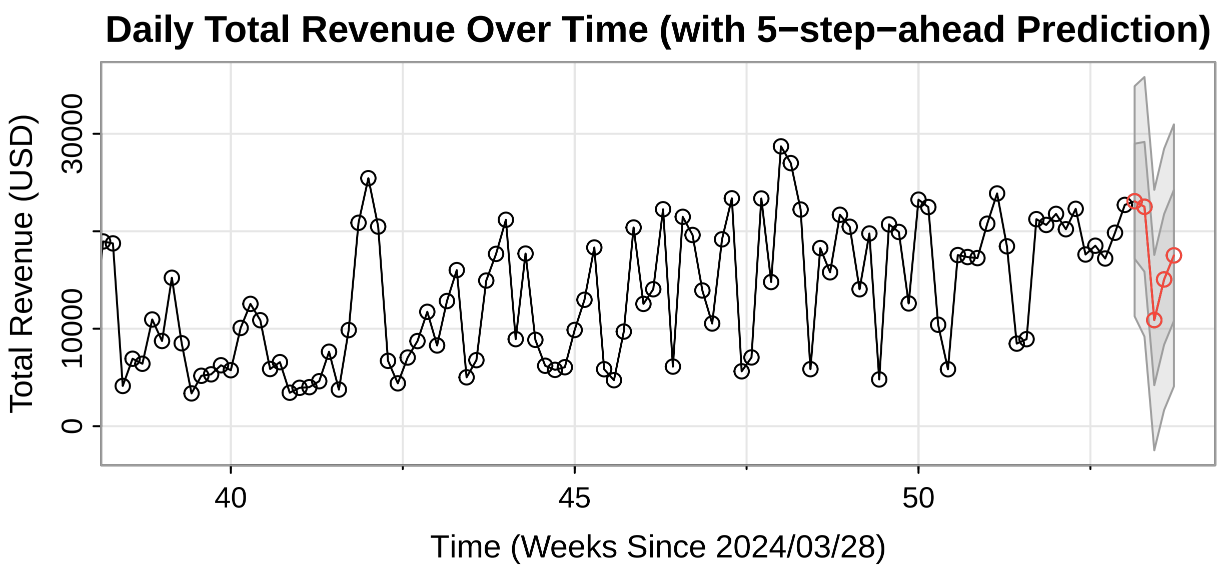


Table 3: Forecasted Values

Date	Forecast	SE
2025/03/28	23079.84	5905.305
2025/03/29	22511.09	6659.038
2025/03/30	10885.84	6681.279
2025/03/31	15057.93	6702.128
2025/04/01	17524.90	6721.676

Figure FIGNUMBER showed the future five forecasted values for $\{X_t\}$, total daily revenue, using the truncated n-step-ahead forecasting method. The dark gray interval represents the 95% prediction interval, and the light gray interval represents the 80% prediction interval. The forecasted values are shown in Table 3.