Enhancing the Business Efficiency of V.Ger Travel Through Data Science

1. Introduction

The travel conglomerate was analysed, and four key use cases were identified where Data Science methods can significantly improve operational efficiency. Time Series Forecasting using SARIMA was applied to predict hotel booking demand, while OLS Regression was used to model customer satisfaction.

Use Cases for the Travel Industry:

1. Forecasting Demand for hotel, flight, and car bookings

Forecasting demand for various services helps allocate staff correctly, manage hotel occupancy, plan procurement and maintenance, and prevent shortages or oversupply. This forms the foundation of dynamic pricing.

2. Analysing Customer Satisfaction

Customer satisfaction analysis identifies the factors that influence guests' ratings, such as price, room comfort, reception service, staff responsiveness, and food quality.

This helps businesses:

- improve service where it matters most;
- enhance ratings on platforms like Booking.com and TripAdvisor;
- reduce the number of complaints;
- increase repeat bookings.

It is highly valuable for quality management in hospitality services.

3. Price Modelling

Price modelling evaluates how the price per night depends on factors such as seasonality, hotel rating, room type, distance to the beach, and customer reviews. Understanding these relationships allows businesses to optimise pricing strategies and maximise revenue.

4. Comparing user responses to website designs (A/B Testing)

A/B testing of website layouts and advertising materials allows companies to compare different page designs, promotional banners, text, call-to-action buttons, and hotel search structures.

This method:

- increases website conversion rates;
- boosts sales without raising the advertising budget;
- helps select the most effective design based on data.

2. SARIMA Time Series Forecasting for Hotel Booking Demand

Demand forecasting is essential in the travel industry, as most operational processes depend on seasonality, long-term trends, and advance planning.

To forecast hotel booking demand, a synthetic time series with a clear trend and seasonality was generated. This dataset reflects realistic booking patterns, where demand increases during peak seasons and decreases during off-peak periods. The generated time series was then modelled using an ARIMA-based forecasting approach to predict future hotel bookings.

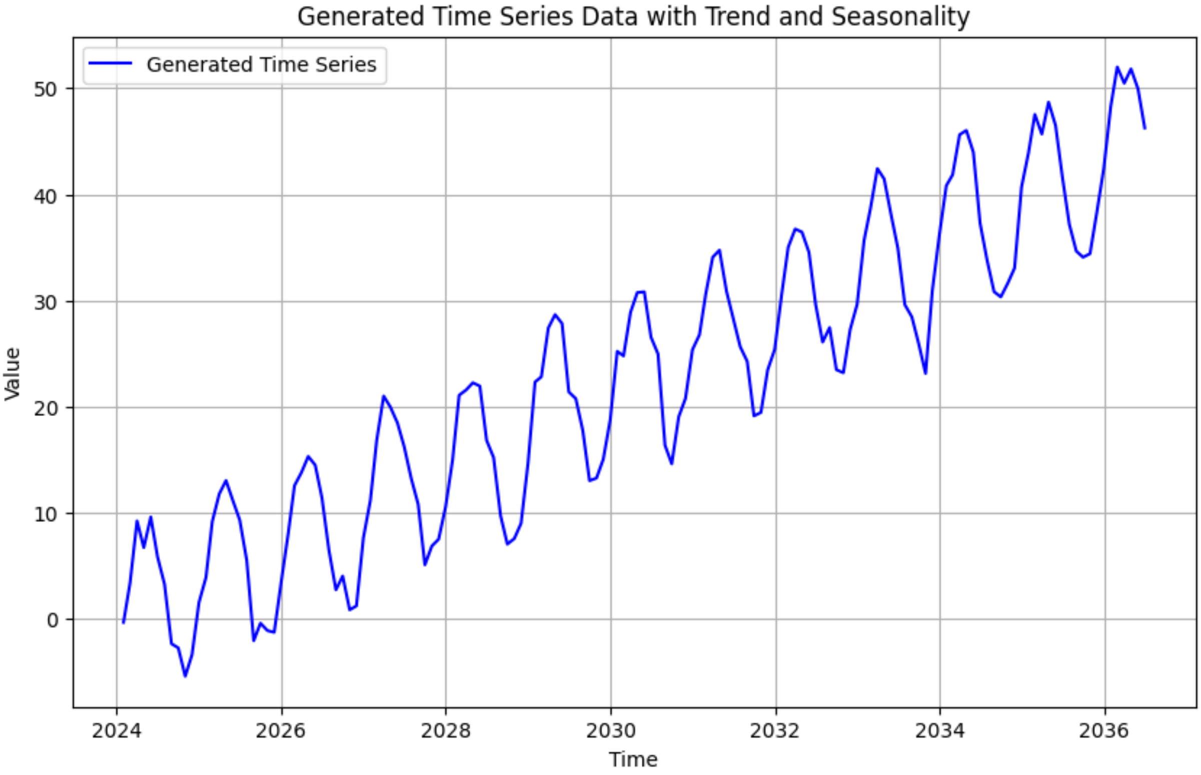
Show code

```
Drive already mounted at /content/drive; to attempt to forcibly remo
```

```
# Decompose the Time Series
t_s = generate_time_series(n_periods=150, frequency='months', trend

# decompose_time_series(t_s)
```

```
/tmp/ipython-input-4049733946.py:17: FutureWarning: 'M' is deprecate
  time_index = pd.date_range(start='2024-01-01', periods=n_periods,
```



The generated trend and seasonality alignment chart shows:

- a clear upward movement (increasing trend);
- regular cyclic up-and-down patterns (seasonality);
- small irregular fluctuations (noise).

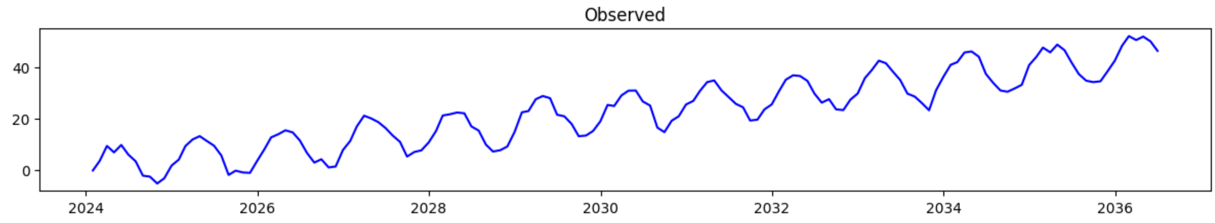
This indicates that the time series is predictable and follows an additive structure.

Decomposition

The next step is to decompose the time series into its components in order to better understand the underlying trend, seasonal patterns, and residual noise, which will help determine whether the series is suitable for SARIMA modelling.

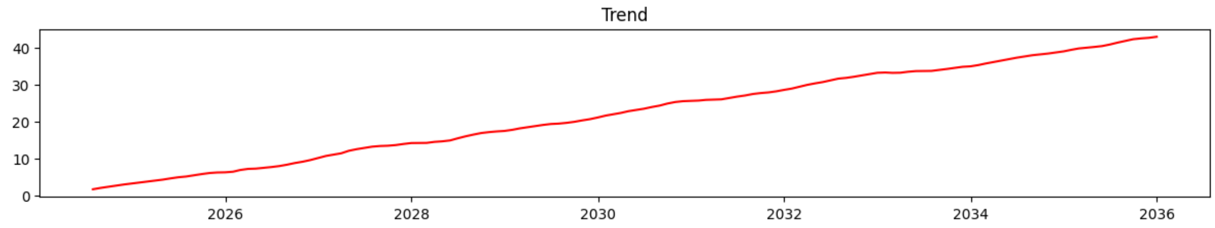
Show code

```
decompose_time_series(t_s, frequency='M')
```



The Observed plot represents the original time series used in the decomposition process.

Show code



The time series exhibits a pronounced linear upward trend.

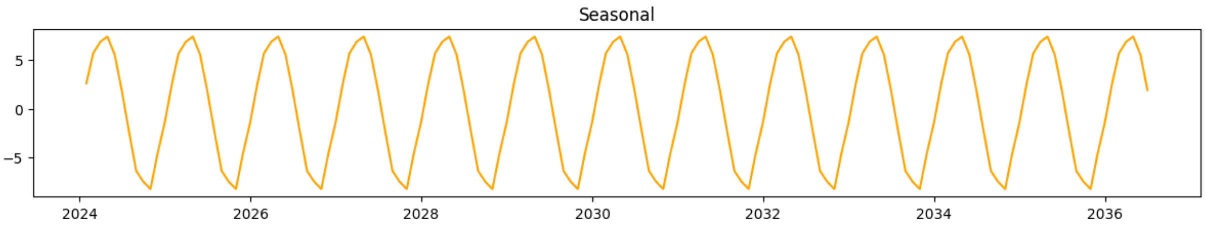
The Trend component shows that the value of the series increases steadily with each period.

This means:

- bookings (or the value we are modeling) are increasing over time;
- the trend is stable, without sharp jumps;
- the trend is nearly linear, which simplifies ARIMA/SARIMA modeling.

Overall, this indicates that the number of bookings is rising year after year — possibly due to growing demand, expansion of the hotel network, or general growth in the tourism industry.

Show code

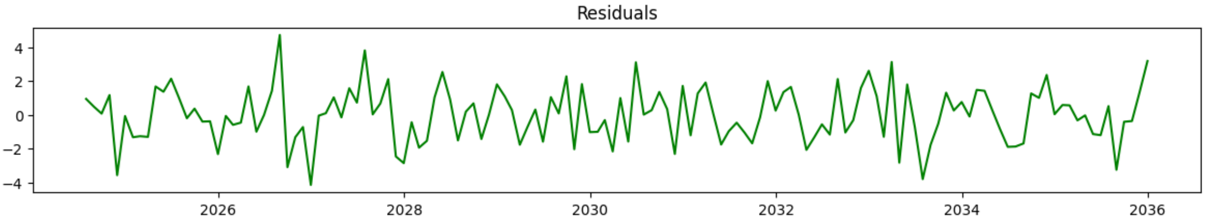


The Seasonal chart shows:

- clear, repeating cycles of the same shape;
- seasonality is stable in amplitude and frequency;
- peaks are repeated every period (in our case, monthly seasonality).

That means that in the summer (or certain months), bookings increase, while in the winter, they decline. This is typical for the tourism industry.

Show code



The residuals graph shows that the remaining variation in the series is random and does not exhibit any visible structure.

In the residuals plot: there is no trend; no seasonality; the fluctuations are irregular and roughly uniformly distributed.

This indicates that:

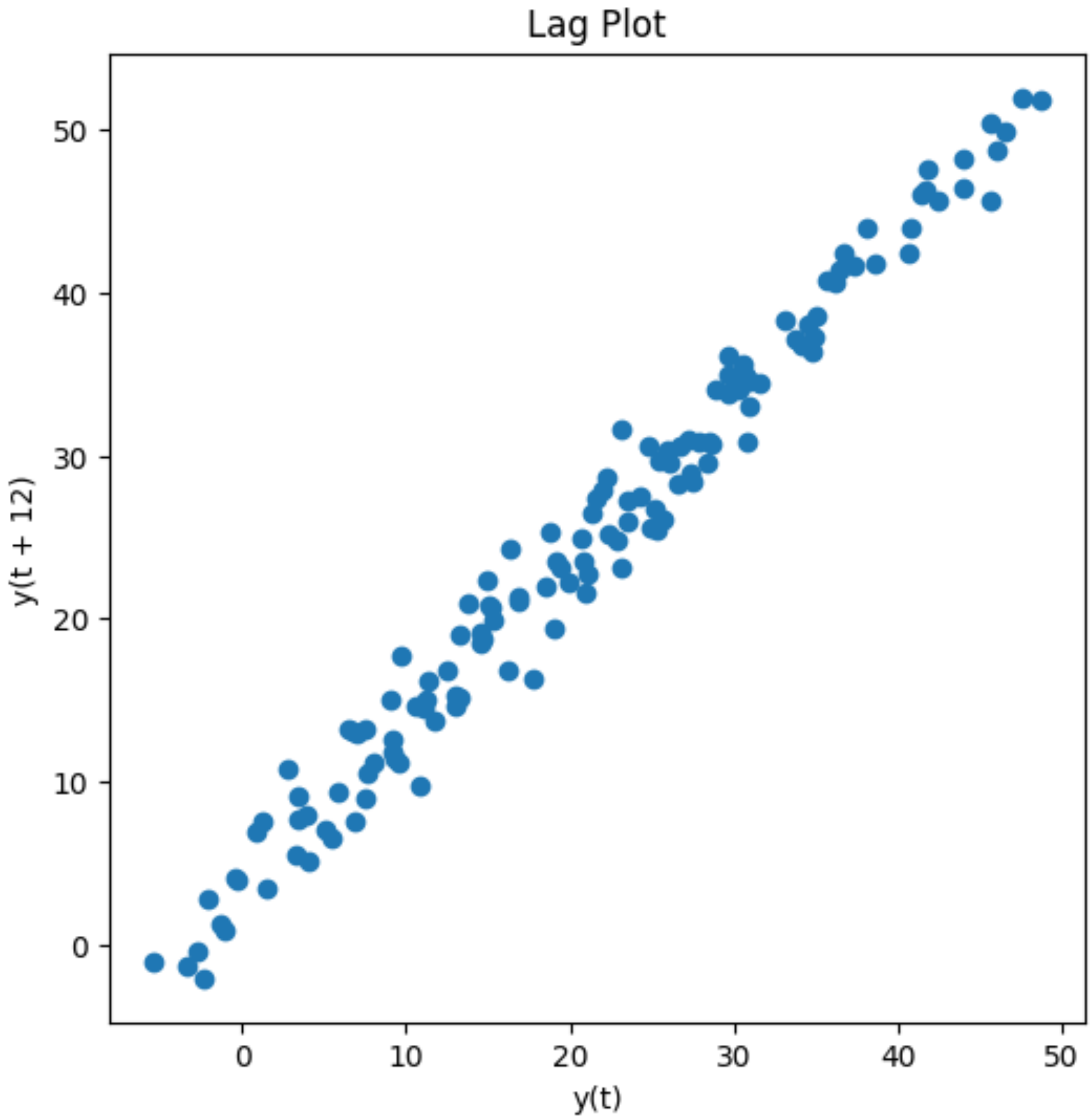
- the trend + seasonality model correctly describes the series;
- there are sufficient data to apply ARIMA/SARIMA;
- the noise component is additive and stationary.

ARIMA/SARIMA models will perform well because the systematic part of the series is explained by deterministic components (trend and seasonality), while the residuals behave like random noise without patterns.

Lag Plot

A lag plot helps to check whether the current value of a time series is related to its past values (lags)— an important requirement for ARIMA modelling.

Show code



The lag plot with a 12-month lag shows a clear, strong line and indicates that the value of the time series one year ahead is highly correlated with its value in the same month of the previous year, confirming consistent annual cycles in hotel booking demand.

This pattern suggests that:

- the series exhibits strong yearly seasonality;
- observations 12 months apart follow very similar dynamics;
- the series is highly predictable at seasonal intervals;
- the data structure is suitable for SARIMA, which requires seasonal autocorrelation.

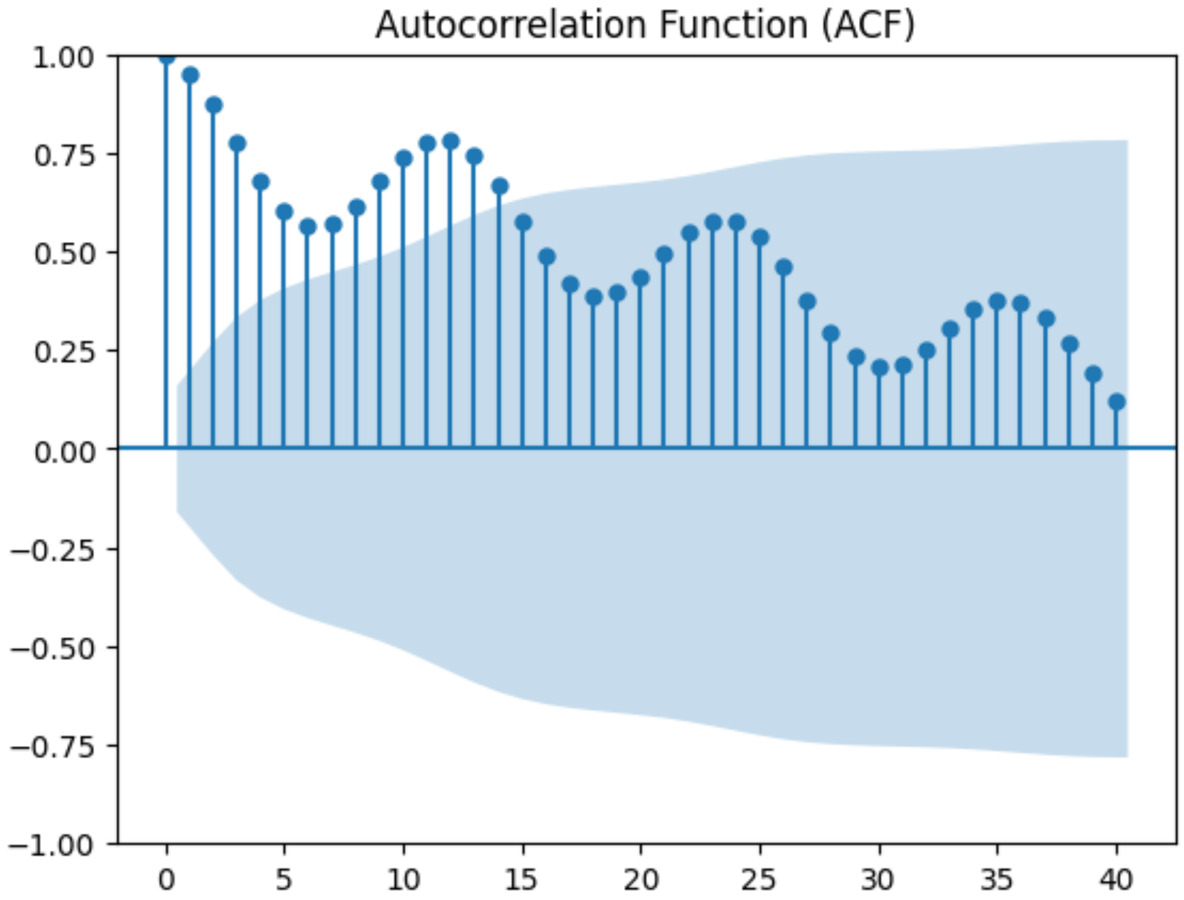
ACF Plot

The ACF plot shows how a time series is related to its past values.

It helps identify short-term dependencies, the presence of autocorrelation, and seasonal patterns — key elements for selecting appropriate ARIMA/SARIMA model parameters.

Show code

```
<Figure size 1000x600 with 0 Axes>
```


Autocorrelation Function (ACF)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
5
10
15
20
25
30
35
40

The ACF plot shows strong positive autocorrelation at the first several lags, indicating that recent values of the series are strongly dependent on previous values. The autocorrelation decreases gradually, which is typical for a time series containing a trend component.

A clear seasonal pattern is visible: a pronounced peak at lag 12 and 24, and a smaller peak near lag 36.

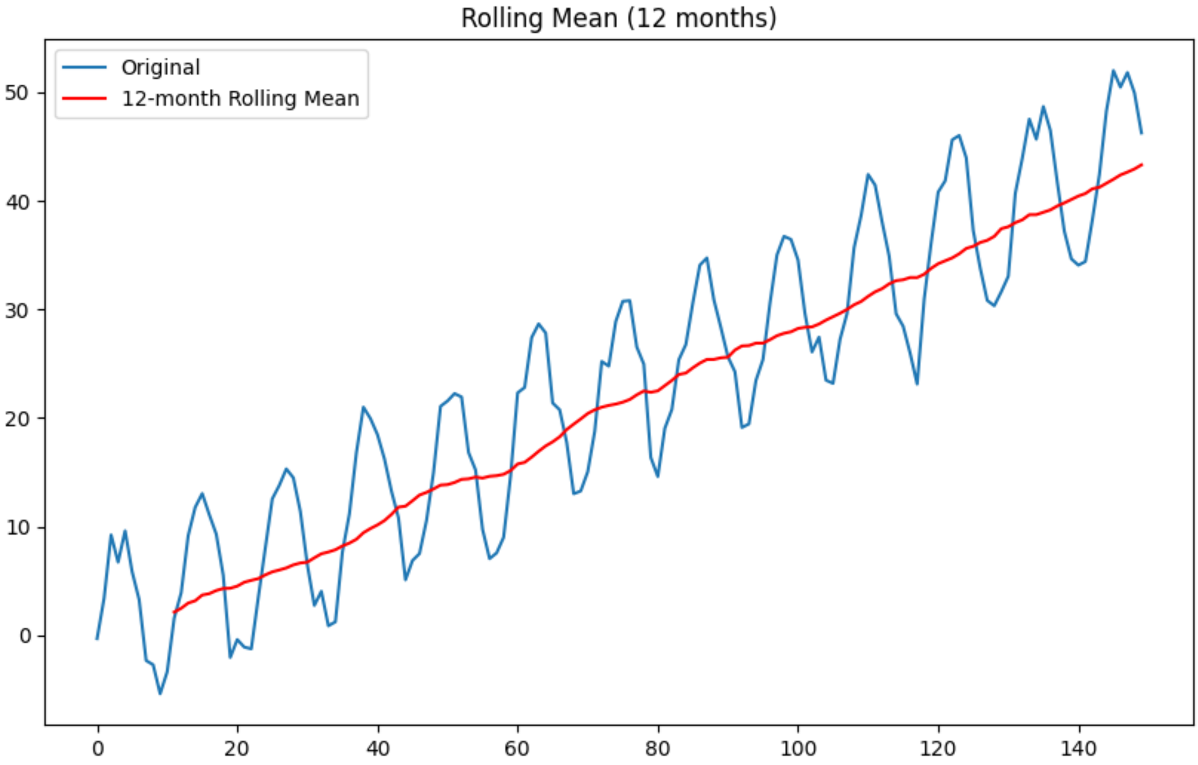
These repeated spikes confirm yearly seasonality in the data.

The slow decay of autocorrelation combined with recurring 12-month peaks is characteristic of a trending, strongly seasonal time series — exactly the type of data for which SARIMA performs well.

Rolling Mean

The rolling mean smooths out short-term fluctuations and highlights the long-term trend in the data.

Show code


Rolling Mean (12 months)
Original
12-month Rolling Mean
50
40
30
20
10
0
0
20
40
60
80
100
120
140

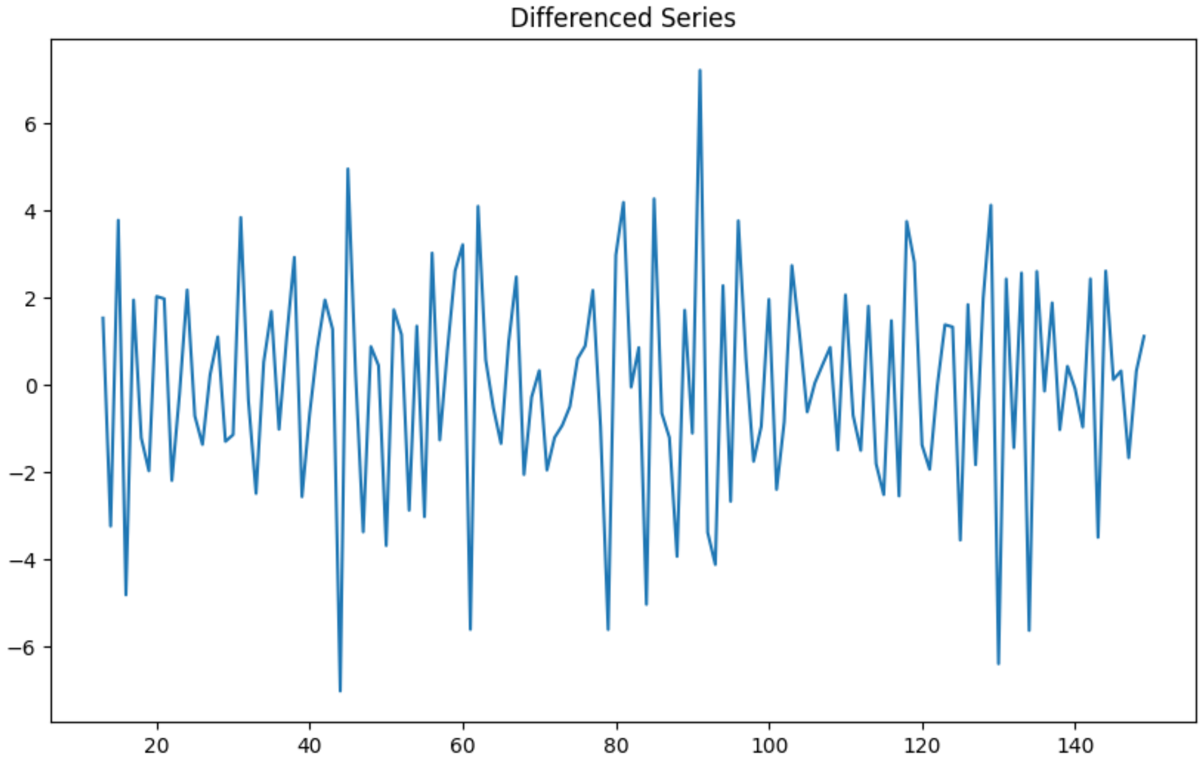
The red rolling-mean curve shows a clear upward trajectory, confirming the presence of a strong positive trend in hotel booking demand.

Overall, the rolling mean confirms that the time series is non-stationary and contains a strong trend component, which justifies the need for differencing when applying ARIMA/SARIMA models.

Stationarity

Since the original time series is non-stationary, it must be transformed by applying differencing before fitting forecasting models. Differencing removes the trend and seasonal components, producing a stationary series suitable for ARIMA/SARIMA modelling.

Show code



The differenced series represents the result of applying both first non-seasonal differencing and seasonal differencing with a 12-month lag. This transformation removes the trend and the yearly seasonality, leaving a series that fluctuates around a constant mean around zero.

These characteristics indicate that the series has become stationary, meaning its statistical properties (mean, variance, autocorrelation) no longer change over time.

Testing for Stationarity

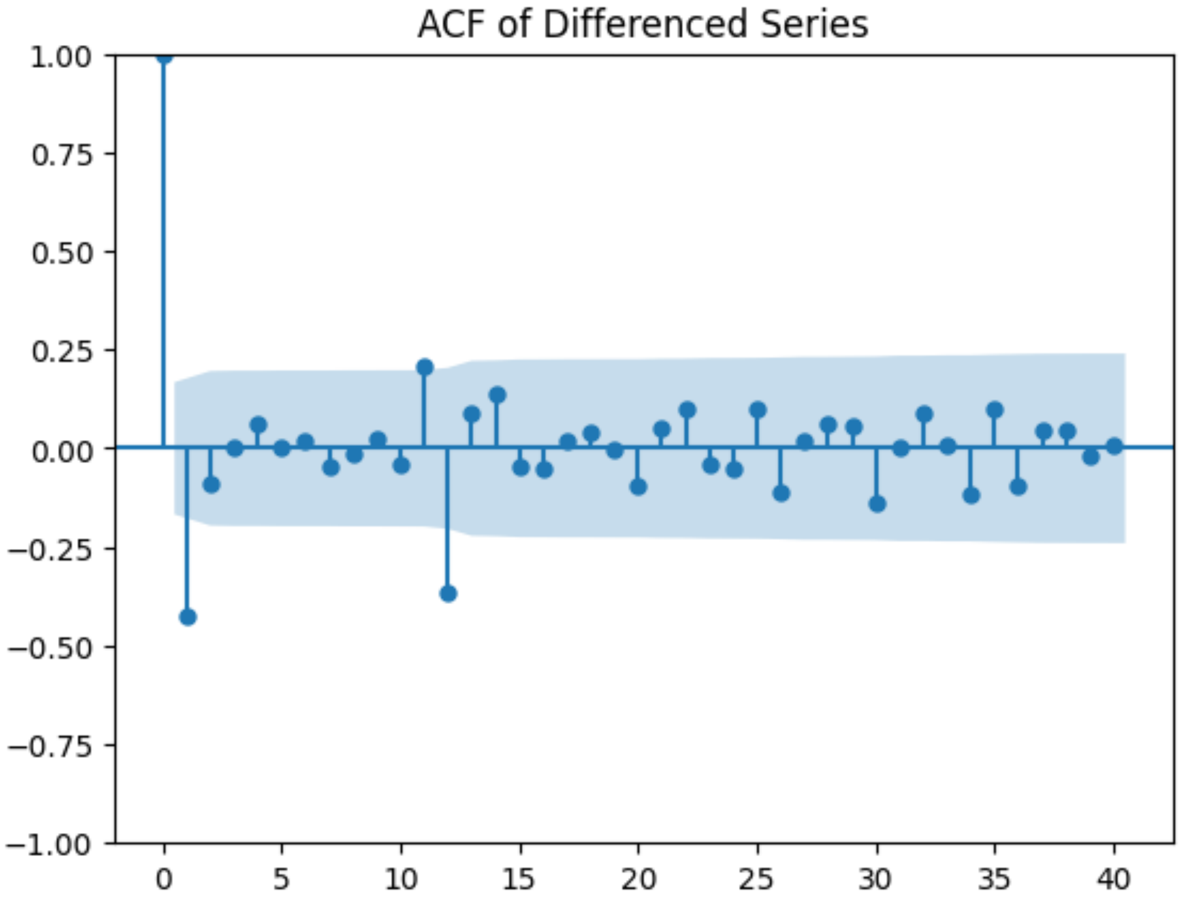
1. Augmented Dickey-Fuller (ADF) Test

The Augmented Dickey–Fuller test evaluates whether a time series contains a unit root.

The null hypothesis (H_0) assumes that the series is non-stationary, meaning its statistical properties change over time. If the p-value is below the chosen significance level (typically 0.05), we reject H_0 and conclude that the series is stationary.

Show code

```
ADF Statistic: -5.010264943055501
p-value: 2.117808902402306e-05
Critical values:
   1%: -3.484219653271961
   5%: -2.885145235641637
   10%: -2.579359138917794
<Figure size 1000x600 with 0 Axes>
```



The p-value is about 0.00002, which is far below 0.05. That means that we reject the null hypothesis of non-stationarity and infer that the differenced time series is stationary and suitable for ARIMA/SARIMA modelling.

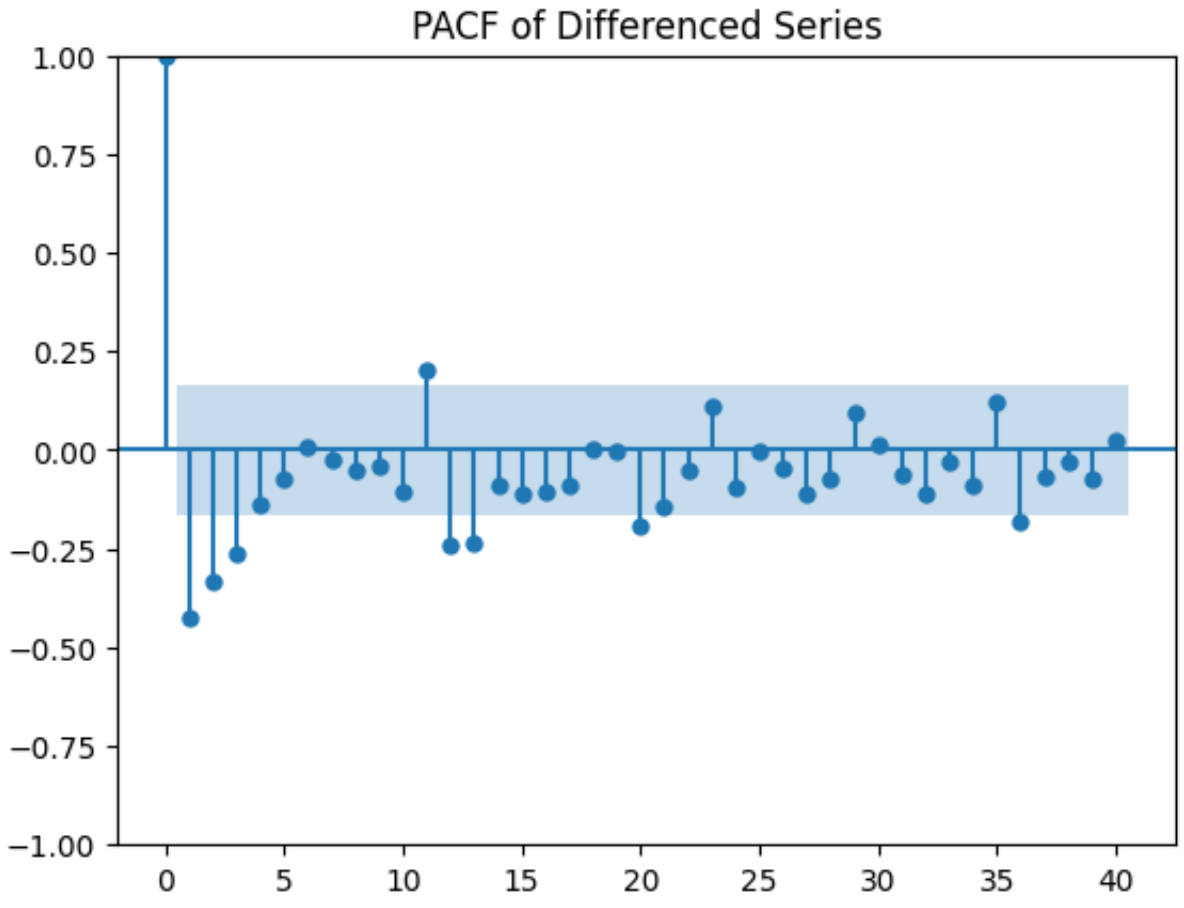
At the same time, the ACF plot confirms that the differenced series behaves like white noise and does not contain any remaining trend or seasonal structure.

2. Partial Autocorrelation Function (PACF) of Differenced Series

The PACF plot shows how many lagged values (previous time points) have a direct influence on the current observation. This information is used to determine the appropriate autoregressive order AR(p) for the ARIMA or SARIMA model.

Show code

```
<Figure size 1000x600 with 0 Axes>
```



The PACF plot displays a dominant spike at lag 1, after which all partial autocorrelations drop to near zero. This indicates an autoregressive structure of order AR(1) and supports choosing p = 1 for ARIMA/SARIMA modelling.

Choose AR Order (p) Using Akaike Information Criterion (AIC)

To determine the optimal autoregressive order p, several ARIMA (p,d,0) models were fitted and compared using the Akaike Information Criterion (AIC). AIC evaluates model quality based on both goodness of fit and model complexity, where lower AIC values indicate a better model.

By comparing AIC values across different AR orders, the model with the lowest AIC is selected as the most appropriate ARIMA specification.

Show code

```
 ARIMA(0,1,0)(0,1,0)[12]             : AIC=640.138, Time=0.03 sec
 ARIMA(1,1,0)(1,1,0)[12]             : AIC=595.393, Time=0.10 sec
 ARIMA(0,1,1)(0,1,1)[12]             : AIC=inf, Time=0.67 sec
 ARIMA(1,1,0)(0,1,0)[12]             : AIC=614.606, Time=0.04 sec
 ARIMA(1,1,0)(2,1,0)[12]             : AIC=589.958, Time=0.25 sec
 ARIMA(1,1,0)(2,1,1)[12]             : AIC=583.938, Time=0.95 sec
 ARIMA(1,1,0)(1,1,1)[12]             : AIC=581.981, Time=0.36 sec
 ARIMA(1,1,0)(0,1,1)[12]             : AIC=581.407, Time=0.18 sec
 ARIMA(1,1,0)(0,1,2)[12]             : AIC=581.980, Time=0.59 sec
 ARIMA(1,1,0)(1,1,2)[12]             : AIC=583.961, Time=1.40 sec
 ARIMA(0,1,0)(0,1,1)[12]             : AIC=605.381, Time=0.13 sec
 ARIMA(2,1,0)(0,1,1)[12]             : AIC=569.244, Time=0.24 sec
 ARIMA(2,1,0)(0,1,0)[12]             : AIC=600.803, Time=0.06 sec
 ARIMA(2,1,0)(1,1,1)[12]             : AIC=570.360, Time=0.38 sec
 ARIMA(2,1,0)(0,1,2)[12]             : AIC=570.258, Time=0.69 sec
 ARIMA(2,1,0)(1,1,0)[12]             : AIC=582.379, Time=0.13 sec
 ARIMA(2,1,0)(1,1,2)[12]             : AIC=572.244, Time=1.50 sec
 ARIMA(3,1,0)(0,1,1)[12]             : AIC=564.638, Time=0.44 sec
 ARIMA(3,1,0)(0,1,0)[12]             : AIC=593.154, Time=0.14 sec
 ARIMA(3,1,0)(1,1,1)[12]             : AIC=566.135, Time=1.24 sec
 ARIMA(3,1,0)(0,1,2)[12]             : AIC=566.102, Time=1.77 sec
 ARIMA(3,1,0)(1,1,0)[12]             : AIC=574.457, Time=0.57 sec
 ARIMA(3,1,0)(1,1,2)[12]             : AIC=568.101, Time=2.36 sec
 ARIMA(4,1,0)(0,1,1)[12]             : AIC=564.548, Time=0.32 sec
 ARIMA(4,1,0)(0,1,0)[12]             : AIC=592.576, Time=0.09 sec
 ARIMA(4,1,0)(1,1,1)[12]             : AIC=566.289, Time=0.70 sec
 ARIMA(4,1,0)(0,1,2)[12]             : AIC=566.269, Time=0.84 sec
 ARIMA(4,1,0)(1,1,0)[12]             : AIC=573.165, Time=0.26 sec
 ARIMA(4,1,0)(1,1,2)[12]             : AIC=568.265, Time=2.36 sec
 ARIMA(5,1,0)(0,1,1)[12]             : AIC=565.702, Time=0.39 sec
 ARIMA(4,1,1)(0,1,1)[12]             : AIC=inf, Time=1.73 sec
 ARIMA(3,1,1)(0,1,1)[12]             : AIC=inf, Time=2.12 sec
 ARIMA(5,1,1)(0,1,1)[12]             : AIC=inf, Time=5.17 sec
 ARIMA(4,1,0)(0,1,1)[12] intercept   : AIC=566.540, Time=0.44 sec

Best model:  ARIMA(4,1,0)(0,1,1)[12]
Total fit time: 30.893 seconds
                                      SARIMAX Results
=========================================================================================
Dep. Variable:                                    y   No. Observat
Model:             SARIMAX(4, 1, 0)x(0, 1, [1], 12)   Log Likeliho
Date:                              Mon, 01 Dec 2025   AIC
Time:                                      23:25:22   BIC
Sample:                                           0   HQIC
                                              - 150
Covariance Type:                                opg
=========================================================================================
                 coef    std err          z      P>|z|      [0.025
-----------------------------------------------------------------------------------------
ar.L1         -0.6590      0.081     -8.089      0.000      -0.819
ar.L2         -0.5015      0.097     -5.185      0.000      -0.691
ar.L3         -0.3028      0.096     -3.139      0.002      -0.492
ar.L4         -0.1262      0.093     -1.363      0.173      -0.308
ma.S.L12      -0.5395      0.092     -5.853      0.000      -0.720
sigma2         3.1916      0.392      8.147      0.000       2.424
=========================================================================================
Ljung-Box (L1) (Q):                   0.03   Jarque-Bera (JB):
Prob(Q):                              0.87   Prob(JB):
```

Based on the stepwise AIC comparison, the model with the lowest AIC value is:

SARIMA(4,1,0)

This model achieves an AIC of 564.55, which is substantially lower than all alternative ARIMA and SARIMA candidates tested.

Therefore, **SARIMA(4,1,0)(0,1,1)[12]** is the best-performing forecasting model for this time series.

Fitting the SARIMA (4,1,0) and Review the Summary

Show code

```
                                     SARIMAX Results
=============================================================================================
Dep. Variable:                                   Value   No. Observatio
Model:             SARIMAX(4, 1, 0)x(0, 1, [1], 12)   Log Likelihood
Date:                             Mon, 01 Dec 2025   AIC
Time:                                     23:47:47   BIC
Sample:                                          0   HQIC
                                             - 150
Covariance Type:                               opg
=============================================================================================
                 coef    std err          z      P>|z|      [0.025
---------------------------------------------------------------------------------------------
ar.L1         -0.6347      0.085     -7.504      0.000      -0.800
ar.L2         -0.5095      0.100     -5.107      0.000      -0.705
ar.L3         -0.3159      0.099     -3.205      0.001      -0.509
ar.L4         -0.1340      0.095     -1.404      0.160      -0.321
ma.S.L12      -0.5462      0.093     -5.868      0.000      -0.729
sigma2         3.2185      0.402      8.004      0.000       2.430
=============================================================================================
Ljung-Box (L1) (Q):                  0.04   Jarque-Bera (JB):
Prob(Q):                             0.84   Prob(JB):
Heteroskedasticity (H):              1.00   Skew:
Prob(H) (two-sided):                 0.99   Kurtosis:
=============================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradient
```

The output shows that the best-fitting model is SARIMA(4,1,0), meaning the series was differenced once (d = 1) and includes 4 autoregressive lags (p = 4).

Most AR coefficients (AR(1), AR(2), and AR(3)) are statistically significant (p-values < 0.05), indicating that recent observations meaningfully predict the next value. AR(4) has a p-value of 0.160, meaning it is not statistically significant but does not harm the model.

The Ljung-Box Q-test (p = 0.84) shows no significant autocorrelation remaining in residuals - so the model fits the data reasonably well.

The Jarque-Bera test (p = 0.93) indicates that residuals are normally distributed and the Heteroskedasticity test (p = 0.99) shows variance is stable (no heteroskedasticity)and residuals behave like white noise.

Overall, SARIMA(4,1,0) is a strong, well-fitting model that successfully captures both the upward trend and the yearly seasonality in the time series. All diagnostics confirm that the model is appropriate and ready for forecasting.

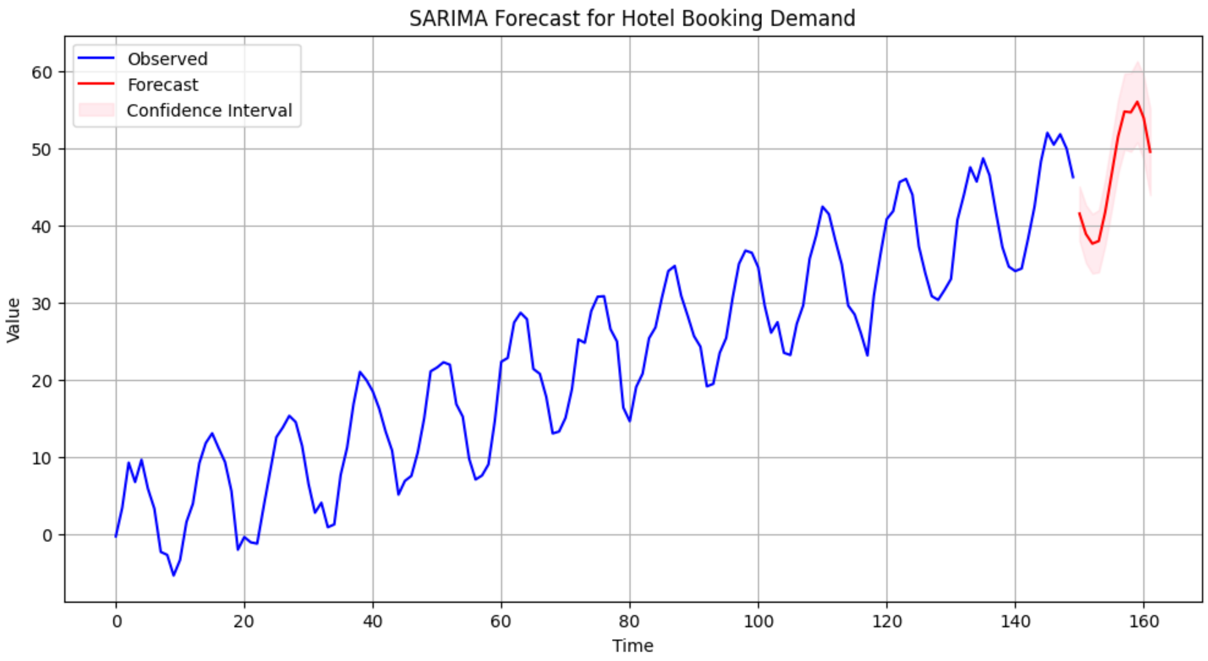
Forecasting and Interpretation

After fitting the SARIMA(4,1,0) model, a 12-step forecast was generated to predict future hotel booking demand.

The forecast includes:

- point predictions (expected future values);
- 90% and 95% confidence intervals, which reflect uncertainty in the forecast;
- future seasonal fluctuations, consistent with the historical yearly pattern.

Show code


SARIMA Forecast for Hotel Booking Demand
Observed
Forecast
Confidence Interval
Value
Time
60
50
40
30
20
10
0
0
20
40
60
80
100
120
140
160

This chart shows how the SARIMA(4,1,0) model predicts hotel booking demand for the next 12 months.

- Blue line — Observed historical data (generated time series). It shows a strong upward trend and clear yearly seasonality. Peaks happen every 12 steps (months).
- Red line — Forecast - predictions for the next 12 months. The red curve continues: the upward growth trend; the seasonal shape; the same pattern of ups and downs.
- Pink shaded area — Confidence Interval (uncertainty range) - where the future values are expected to fall with 95% probability.

Forecasting predicts the following:

1. **Demand will continue to increase**. The red curve rises higher than the last blue values, so the model predicts continued long-term growth in bookings.
2. **Seasonality stays stable and predictable**. The red line forms the same wave-like pattern : seasonal highs and lows remain consistent from year to year.
3. Forecast uncertainty is low. The pink band is narrow, meaning the model is confident and residuals behaved like white noise.
4. The forecast follows the exact structure of the generated time series The SARIMA model captured trend, yearly seasonality and noise. The predictions look realistic and smooth with no strange jumps.

To sum up, the SARIMA model forecasts a continued upward trend in hotel booking demand with well-defined seasonal peaks. The confidence intervals remain narrow, indicating a stable and reliable model suitable for planning and capacity management.