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 as the most relevant for implementing Data Science methods to improve business efficiency. Among them, forecasting hotel booking demand was selected for detailed investigation, as it plays a crucial role in effective planning and decision-making in the travel industry.

To support this use case, a synthetic time series was generated to replicate realistic booking patterns typically observed in hotel operations. Time Series Forecasting techniques were then applied to predict future hotel booking demand.

Four Key Use Cases for the Travel Industry:

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

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

Time Series modelling can be applied to this case to capture trends, seasonal patterns, and fluctuations in the data, allowing the company to predict future demand.

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.

Linear Regression (OLS) can be applied to identify factors that have a significant impact on customer ratings and to quantify the strength and direction of these effects.

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.

Time Series and OLS Regression can be applied to reveal how price depends on hotel characteristics and external factors.

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.

Hypothesis testing (such as t-test or z-test) can be applied to determine whether one design performs significantly better than another.

2. 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.

2.1 Time Series Generation

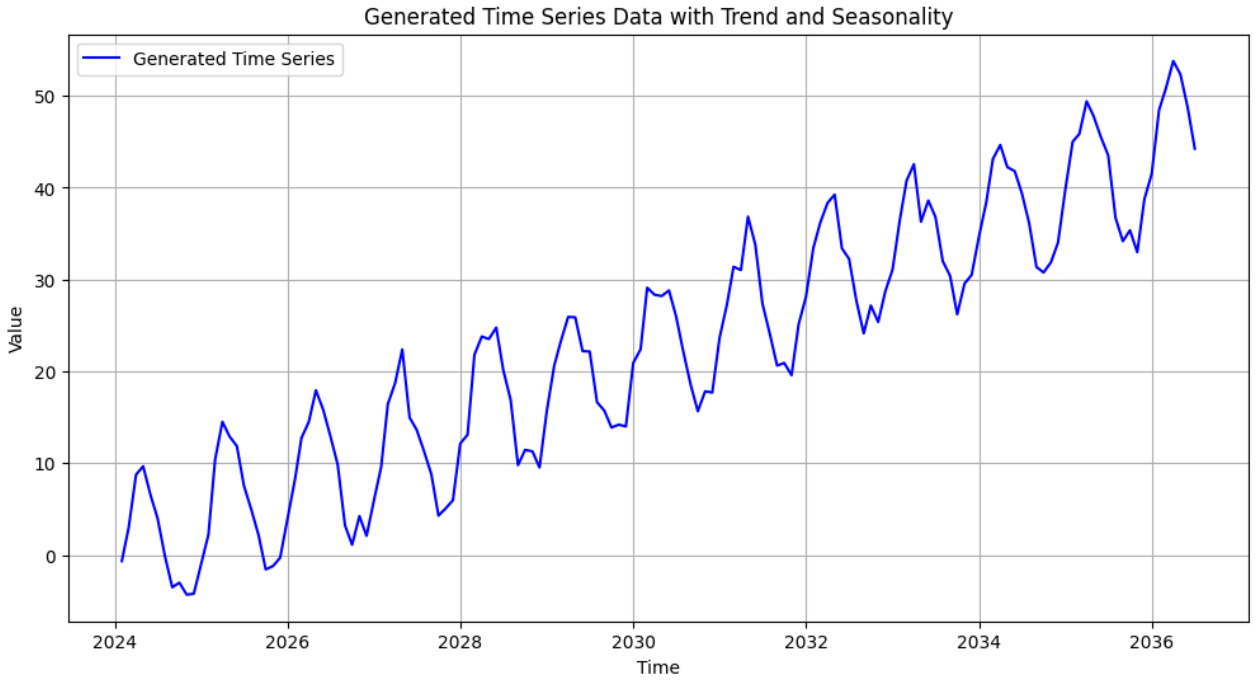
To forecast hotel booking demand, a synthetic time series with a trend and seasonality should be generated. The following dataset reflects realistic booking patterns, where demand increases during peak seasons and decreases during off-peak periods. The generated time series will be then modelled using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

Show code

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.
```

Show code

```
/tmp/ipython-input-2356864722.py:18: FutureWarning: 'M' is deprecated and will be re
  time_index = pd.date_range(start='2024-01-01', periods=n_periods, freq=freq)
```



The generated trend and seasonality alignment chart shows:

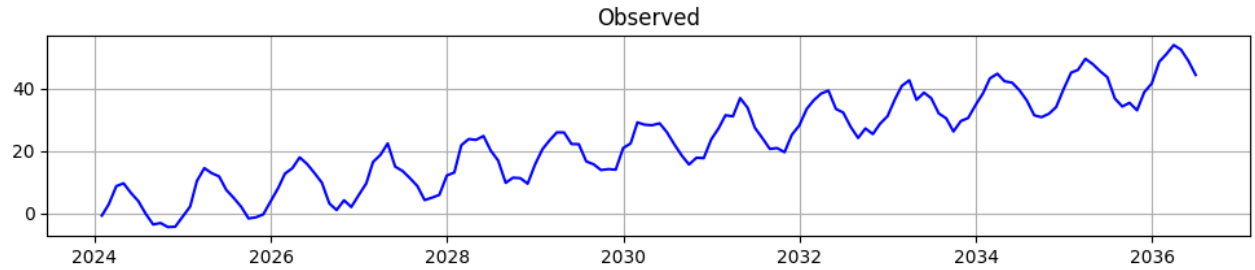
- a clear upward movement (increasing trend);
- regular cyclic up-and-down patterns (seasonality);
- small irregular fluctuations (noise).

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

2.2 Decomposition

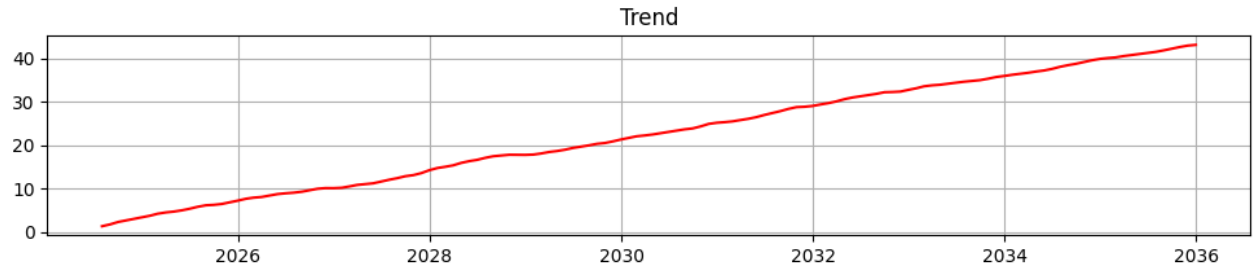
The next step is to decompose the time series into its components to better understanding of the underlying trend, seasonal patterns, and residual noise, which will help determine whether the series is suitable for ARIMA/SARIMA modelling.

Show code



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

Show code



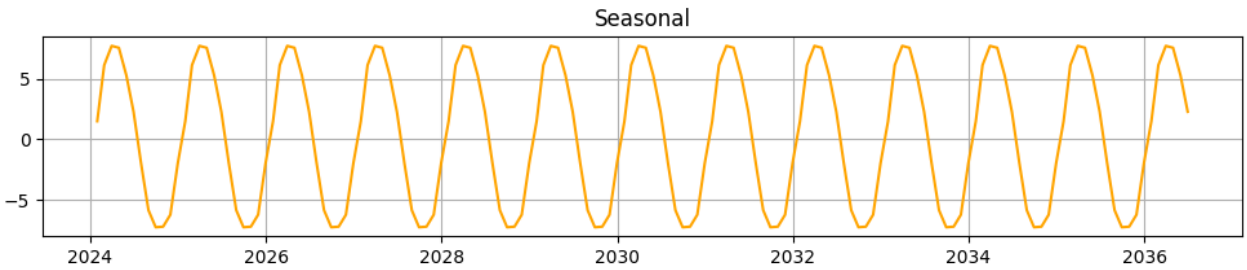
The Trend plot shows that the value of the series increases steadily with each period.

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

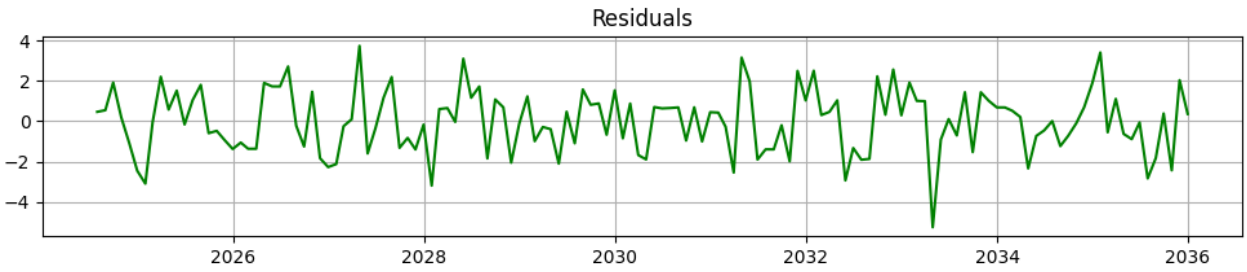


The Seasonal plot shows:

- 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:

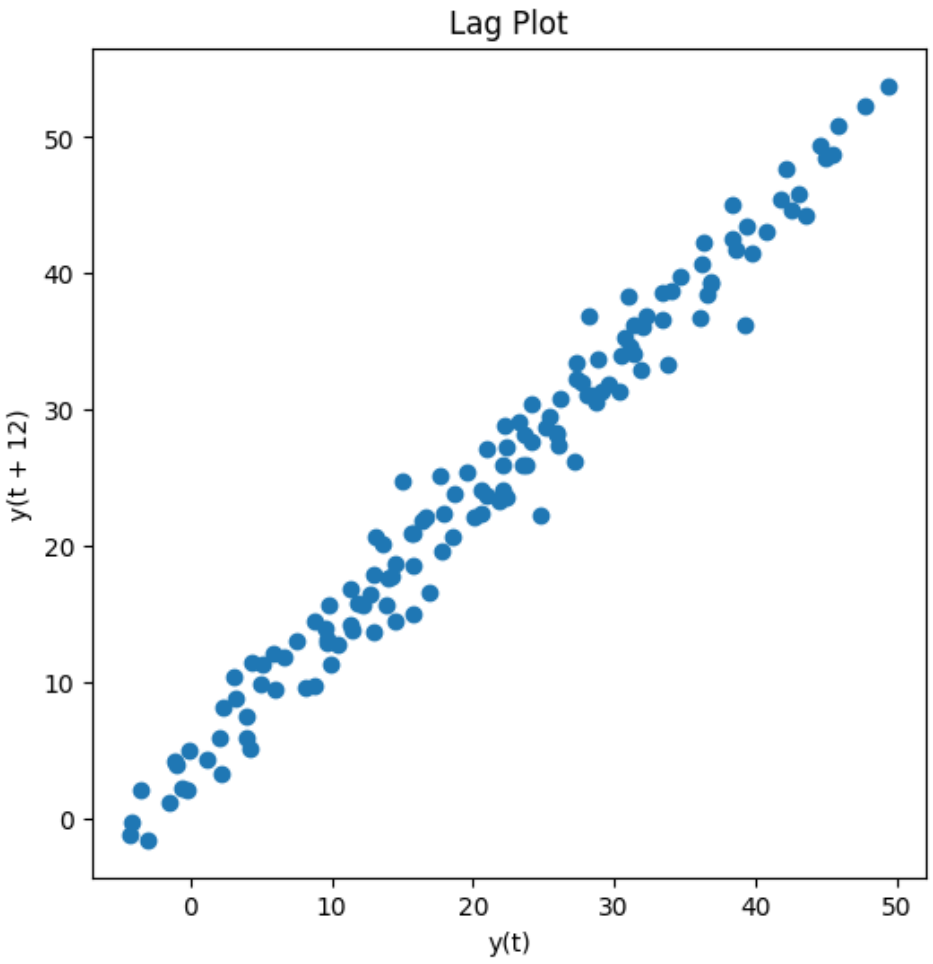
- the trend + seasonality model correctly describes the series;
- there are sufficient data to apply ARIMA/SARIMA;
- the noise component is additive and stationary.

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.

2.3 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/SARIMA 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.

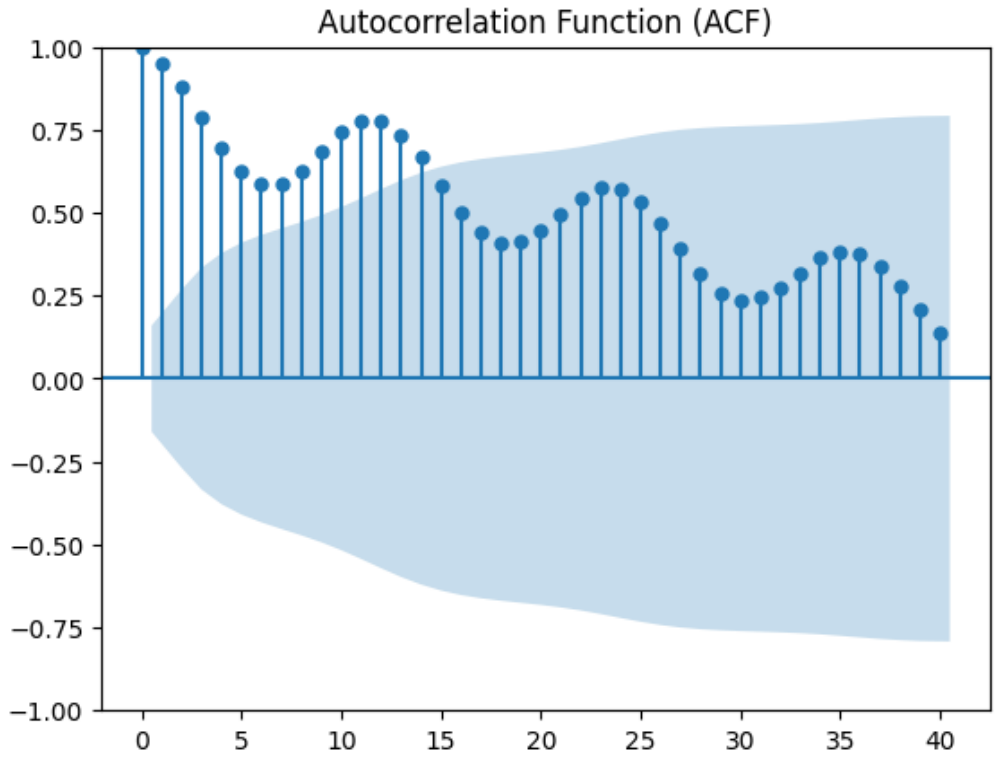
2.4 Autocorrelation Function (ACF) Plot

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>
```



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.

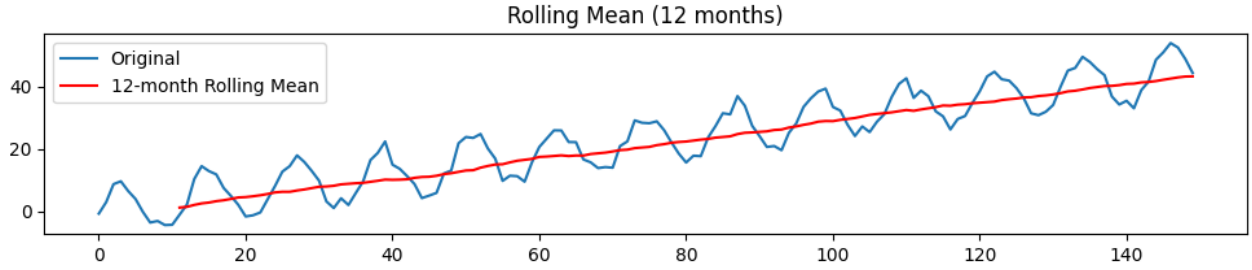
A clear seasonal pattern is visible: a pronounced peak at lag 12 and 24, and a smaller peak near lag 36.

These repeated spikes confirm yearly seasonality in the data.

2.5 Rolling Mean

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

Show code



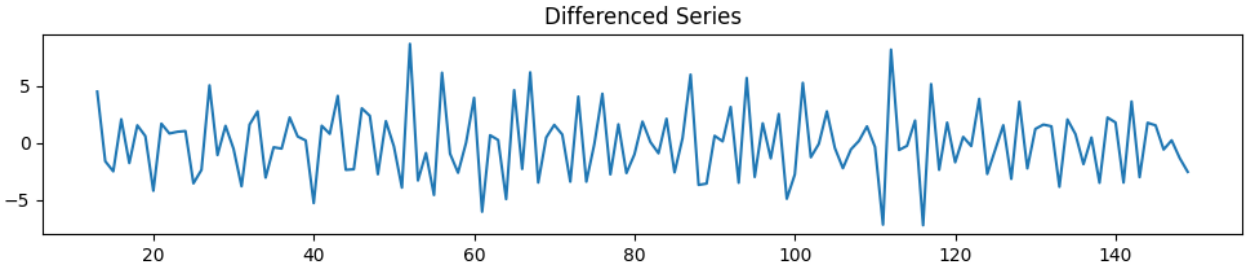
The red rolling-mean curve shows a clear upward trajectory, confirming the presence of a strong positive trend in hotel booking demand.

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.

2.6 Stationarity

Since the generated 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 plot 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.

2.7 Testing for Stationarity

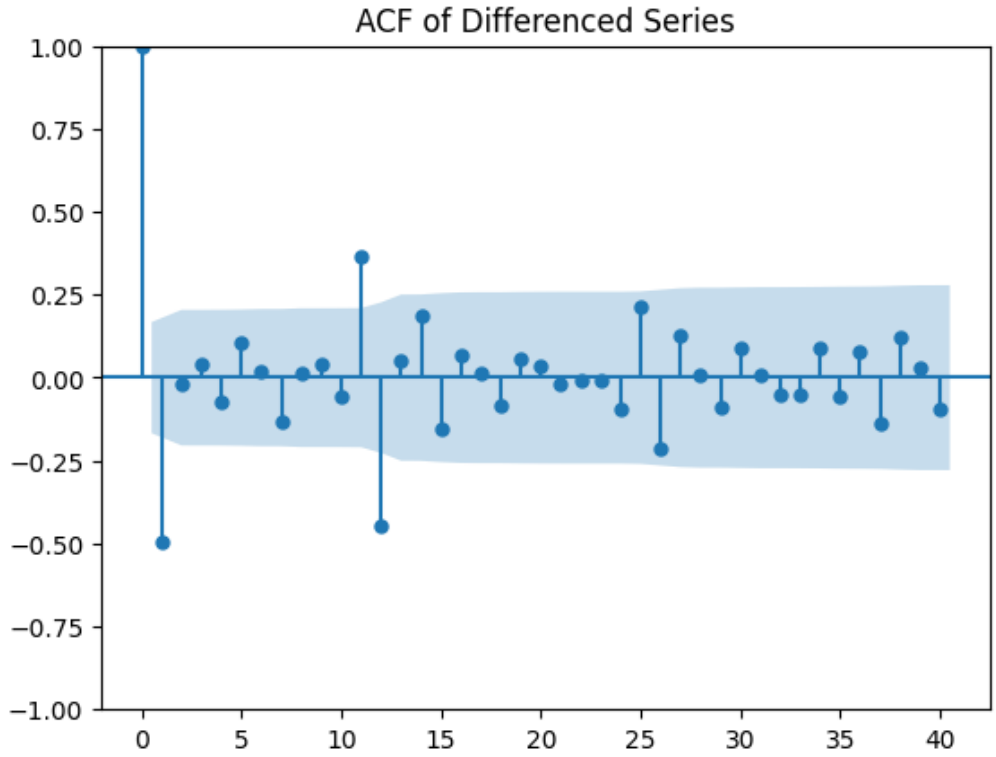
a) Augmented Dickey-Fuller (ADF) Test

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 significance level of 0.05, we reject H_0 and conclude that the series is stationary.

Show code

```
ADF Statistic: -4.034626682686143
p-value: 0.0012396637871963736
Critical values:
   1%: -3.4833462346078936
   5%: -2.8847655969877666
   10%: -2.5791564575459813
<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.

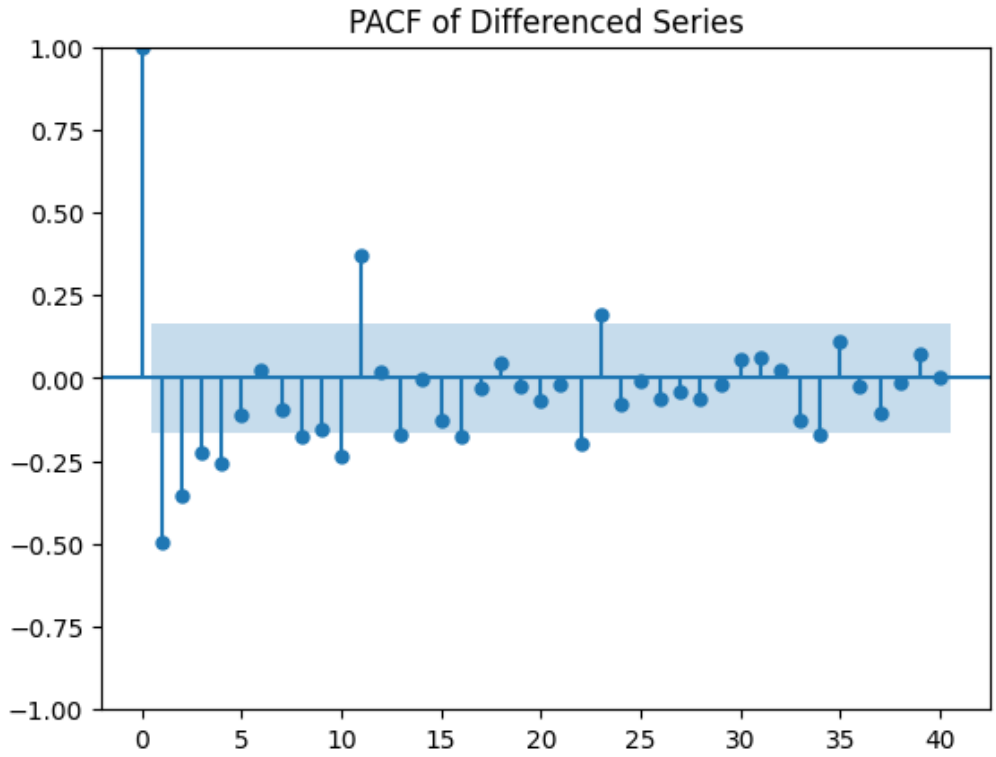
At the same time, **the ACF plot confirms** that the differenced series behaves like white noise and does not contain any remaining trend or seasonal structure.

b) 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.

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

To determine the optimal autoregressive order p, different 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.

Show code

```
Requirement already satisfied: pmdarima in /usr/local/lib/python3.12/dist-packages
Requirement already satisfied: joblib>=0.11 in /usr/local/lib/python3.12/dist-packa
Requirement already satisfied: Cython!=0.29.18,!=0.29.31,>=0.29 in /usr/local/lib/
Requirement already satisfied: numpy>=1.21.6 in /usr/local/lib/python3.12/dist-pac
Requirement already satisfied: pandas>=0.19 in /usr/local/lib/python3.12/dist-packa
Requirement already satisfied: scikit-learn>=0.22 in /usr/local/lib/python3.12/dis
Requirement already satisfied: scipy>=1.13.0 in /usr/local/lib/python3.12/dist-pac
Requirement already satisfied: statsmodels>=0.14.5 in /usr/local/lib/python3.12/di
Requirement already satisfied: urllib3 in /usr/local/lib/python3.12/dist-packages
Requirement already satisfied: setuptools!=50.0.0,>=42 in /usr/local/lib/python3.1
Requirement already satisfied: packaging>=17.1 in /usr/local/lib/python3.12/dist-pa
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.12/
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.12/dist-packa
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.12/dist-pa
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.12/d
Requirement already satisfied: patsy>=0.5.6 in /usr/local/lib/python3.12/dist-packa
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.12/dist-packages
Performing stepwise search to minimize aic
 ARIMA(2,1,2)(1,1,1)[12]             : AIC=inf, Time=2.75 sec
 ARIMA(0,1,0)(0,1,0)[12]             : AIC=691.381, Time=0.05 sec
 ARIMA(1,1,0)(1,1,0)[12]             : AIC=620.338, Time=0.25 sec
 ARIMA(0,1,1)(0,1,1)[12]             : AIC=inf, Time=1.71 sec
 ARIMA(1,1,0)(0,1,0)[12]             : AIC=654.277, Time=0.05 sec
 ARIMA(1,1,0)(2,1,0)[12]             : AIC=604.979, Time=0.48 sec
 ARIMA(1,1,0)(2,1,1)[12]             : AIC=595.220, Time=0.95 sec
 ARIMA(1,1,0)(1,1,1)[12]             : AIC=595.723, Time=0.28 sec
 ARIMA(1,1,0)(2,1,2)[12]             : AIC=593.348, Time=1.61 sec
 ARIMA(1,1,0)(1,1,2)[12]             : AIC=593.912, Time=1.50 sec
 ARIMA(0,1,0)(2,1,2)[12]             : AIC=628.390, Time=1.12 sec
 ARIMA(2,1,0)(2,1,2)[12]             : AIC=582.925, Time=1.79 sec
 ARIMA(2,1,0)(1,1,2)[12]             : AIC=585.320, Time=1.44 sec
 ARIMA(2,1,0)(2,1,1)[12]             : AIC=586.213, Time=1.05 sec
 ARIMA(2,1,0)(1,1,1)[12]             : AIC=588.154, Time=0.46 sec
 ARIMA(3,1,0)(2,1,2)[12]             : AIC=577.940, Time=3.27 sec
 ARIMA(3,1,0)(1,1,2)[12]             : AIC=581.062, Time=2.87 sec
 ARIMA(3,1,0)(2,1,1)[12]             : AIC=582.843, Time=1.08 sec
 ARIMA(3,1,0)(1,1,1)[12]             : AIC=585.476, Time=0.49 sec
 ARIMA(4,1,0)(2,1,2)[12]             : AIC=575.151, Time=2.35 sec
 ARIMA(4,1,0)(1,1,2)[12]             : AIC=578.506, Time=2.42 sec
 ARIMA(4,1,0)(2,1,1)[12]             : AIC=580.295, Time=1.51 sec
 ARIMA(4,1,0)(1,1,1)[12]             : AIC=582.481, Time=0.67 sec
 ARIMA(5,1,0)(2,1,2)[12]             : AIC=576.575, Time=4.43 sec
 ARIMA(4,1,1)(2,1,2)[12]             : AIC=inf, Time=6.08 sec
 ARIMA(3,1,1)(2,1,2)[12]             : AIC=inf, Time=4.63 sec
 ARIMA(5,1,1)(2,1,2)[12]             : AIC=inf, Time=9.11 sec
 ARIMA(4,1,0)(2,1,2)[12] intercept   : AIC=577.116, Time=2.63 sec

Best model:  ARIMA(4,1,0)(2,1,2)[12]
Total fit time: 57.052 seconds
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:
Model:             SARIMAX(4, 1, 0)x(2, 1, [1, 2], 12)   Log Likelihood
Date:                                 Thu, 04 Dec 2025   AIC
Time:                                         21:14:43   BIC
Sample:                                              0   HQIC
                                                   - 150
Covariance Type:                                   opg
------------------------------------------------------------------------------
```

By comparing AIC values across different AR orders, the model with the lowest AIC value was selected as the most appropriate SARIMA specification:

SARIMA(4,1,0)(2,1,2)[12] with AIC = 575.151

This model is the best-performing forecasting model for this time series.

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

Show code

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                 Value   No. Observations:
Model:             SARIMAX(4, 1, 0)x(2, 1, [1, 2], 12)   Log Likelihood
Date:                                 Thu, 04 Dec 2025   AIC
Time:                                         21:17:54   BIC
Sample:                                              0   HQIC
                                                   - 150
Covariance Type:                                   opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.7221      0.127     -5.688      0.000      -0.971      -0.473
ar.L2         -0.5334      0.158     -3.382      0.001      -0.843      -0.224
ar.L3         -0.3133      0.133     -2.361      0.018      -0.573      -0.053
ar.L4         -0.1641      0.111     -1.475      0.140      -0.382       0.054
ar.S.L12      -0.2625      0.280     -0.936      0.349      -0.812       0.287
ar.S.L24      -0.4114      0.115     -3.589      0.000      -0.636      -0.187
ma.S.L12      -0.4686      0.316     -1.482      0.138      -1.089       0.151
ma.S.L24       0.2005      0.266      0.755      0.450      -0.320       0.721
sigma2         3.5653      0.463      7.697      0.000       2.658       4.473
===================================================================================
Ljung-Box (L1) (Q):                   0.03   Jarque-Bera (JB):                 4.71
Prob(Q):                              0.86   Prob(JB):                         0.10
Heteroskedasticity (H):               0.45   Skew:                            -0.25
Prob(H) (two-sided):                  0.02   Kurtosis:                         3.88
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step)
```

The output shows that the best-fitting model is SARIMA(4,1,0)(2,1,2)[12], 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 0.140, meaning it is not statistically significant but does not harm the model.

The Ljung-Box Q-test (p = 0.86) shows no significant autocorrelation remaining in residuals - so the model fits the data reasonably well.

The Jarque-Bera test (p = 0.10) indicates that residuals are normally distributed and the Heteroskedasticity test (p = 0.45) shows variance is stable (no heteroskedasticity)and residuals behave like white noise.

Overall, SARIMA(4,1,0)(2,1,2)[12] 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.

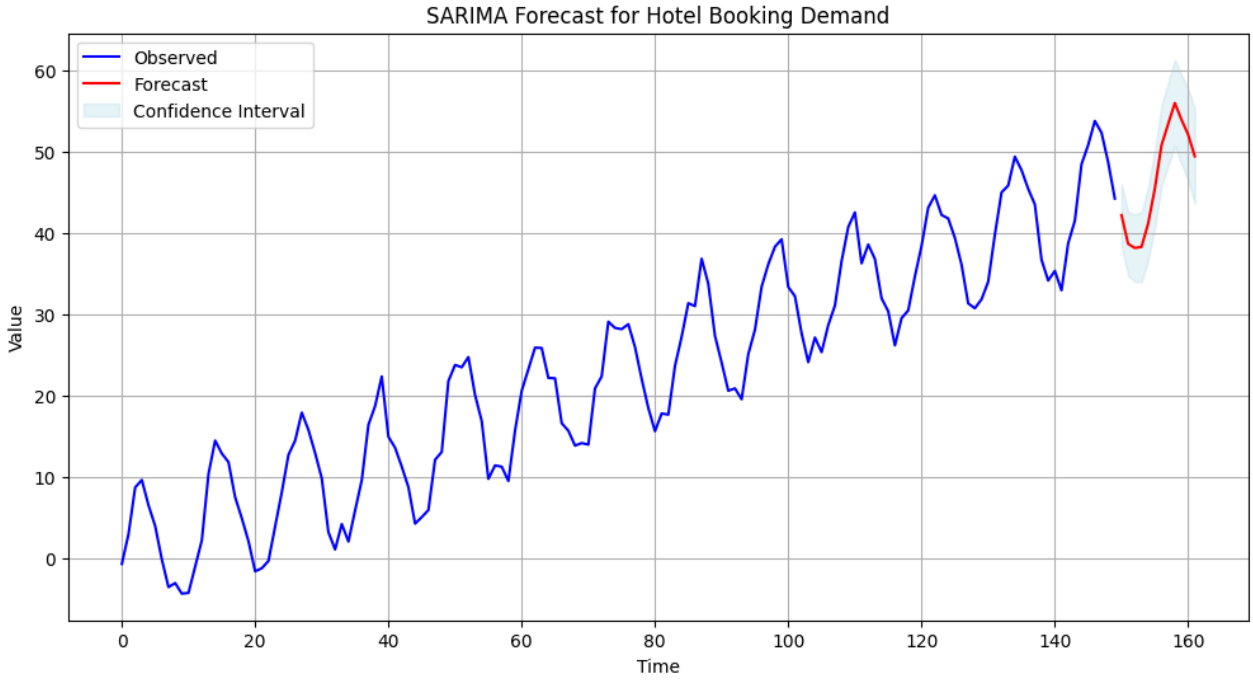
2.10 Forecasting and Interpretation

After fitting the SARIMA(4,1,0)(2,1,2)[12] model, a 12-step forecast can be 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
0
10
20
30
40
50
60
0
20
40
60
80
100
120
140
160

This chart shows how the SARIMA(4,1,0)(2,1,2)[12] 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.
- *Light blue shaded area — 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 light blue 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 uncertainty intervals remain narrow, indicating a stable and reliable model suitable for planning and capacity management.

3. Conclusion

After analysing the operations of the travel conglomerate, four key use cases were identified where Data Science techniques can be effectively implemented to enhance business efficiency. Among these, Time Series modelling was applied in detail to forecast hotel booking demand for the next 12 months.

The forecasting results indicate that V.Ger Travel can expect a continued increase in future hotel bookings, with clearly defined seasonal patterns. Such insights support strategic planning, resource allocation, and revenue optimisation, demonstrating the value of Data Science in improving decision-making across the organisation.