Business forecast report for the year 2022-23

By

Krishna Rao Ramesh

Table of Contents

1. Problem statement →

To provide detailed insight on given time series dataset through the use of different data analysis techniques and build a forecasting model to project values for the time period *01-09-2022 to 01-08-2023*. This analysis will include:

- EDA with the help of summary statistics and visualizations.
- Cleaning data by fixing outliers, missing values etc.
- Detection of Trend, Cyclicity and Seasonality.
- Forecasting for target time period.
- Deriving inferences from above analysis.

The Time series dataset has three variables;

a. Date (in dd-mm-yyyy)
b. Value (No unit)
c. Series_id (Constant feature)

2. Solution Statement →

To arrive at accurate forecast values, the following set of actions will be undertaken.

- ✓ Calculating *summary statistics* (*Mean, median, Standard deviation etc*) and *plotting visualizations* (*line graphs, boxplots etc*) on given time series to identify *trends, outliers* and *seasonality* patterns.
- ✓ Perform data cleaning operations for filling missing values and replacing outliers by common estimation methods such as *Linear forecast* and *Measure of Interquartile range (IQR).*
- ✓ Creating *Dummy variables* and carrying out *feature engineering* to discern seasonality more accurately.
- ✓ Building different time series models to identify the *best-fit model* for given dataset. Evaluating performance of each model through the use of metrics such as *Mean absolute deviation (MAD), Mean absolute percentage error (MAPE), Mean squared error (MSE)* and *Bias*.
- ✓ Providing insights and pain points from analysis. Offering *recommendations* to organization based on insight inferred.

3. Exploratory data analysis and Data pre-processing →

Data cleaning

a. Since most values are around the range of 0 to 100, an educated guess is made to identify the unit for the values; the data could be reflective of the **general flow of stock price value across time**.

b. Considering the values are stock prices, we are *removing the negative sign for stock values* (instance of typo) since they cannot be negative.

c. To fill in missing values, I am taking the help of different forecast functions in excel. Since the **data has a linear upward trend and seasonality** (identified by *Forecast.ets.seasonality*), different columns were created to represent different methodologies to replace missing values;

   → **Forecast.linear:** Used in this business context since data has clear trend. Predicts future values along a trendline based on existing values.

   → **Forecast.ets.seasonality:** Ascertains at which intervals data exhibits seasonality or recurring pattern (8 for our dataset).

   → **Forecast.ets:** A forecasting function that takes seasonality (8 for our dataset) into account while computing values, with the help of exponential smoothing.

After calculating and comparing results from above interpolation/ forecasting techniques, **values from forecast.ets** was chosen since the function focuses on estimating values with seasonality period as a parameter.

d. Since the datapoints have trend and seasonality, using ***IQR on actual data could lead to incorrect classification of outliers***. IQR does not take these temporal variation or changes into account, which might result in detection of an outlier that is otherwise a peak seasonality or cyclical datapoint.

e. Seasonal decomposition can help in addressing this issue. This method decomposes data into trend, seasonal and residual values. In our dataset, a *centered moving average for 8-month period* would help to remove the trend exhibited by data and a *centered seasonal component* would isolate the seasonality pattern. Subtracting the sum of these two components from the actual values would give us the residual i.e. the values that are random variations not captured by trend and seasonal components. Applying IQR on the residual helps us to identify the values or "noise" that are deviating significantly from calculated upper and lower bounds. **No outliers were detected based on above set of calculations.**

Note - A centered calculative approach is used for both moving average and seasonal component since this would include all values while computing decomposition. A non-centered approach would neglect values present towards the edges of dataset.

Other observations

- ✓ A skewness of 0.34 suggest that data is **slightly skewed to the right**. A kurtosis of -0.74 indicates **low kurtosis**, meaning that distribution is spread apart with few extreme values (outliers).

Summary

- ✓ "*Values*" column denoting stock price value.
- ✓ Removal of negative sign for stock price values (typo).
- ✓ **Forecast.ets.seasonality** function used to determine seasonality period. A result of 8 indicates seasonality occurs 8 months once.
- ✓ **Forecast.ets** function used to fill missing values. This function calculates values by considering seasonality as a parameter.
- ✓ IQR on actual data would lead to incorrect outlier detection. Rather, applying IQR on residual data after seasonal decomposition would help in sidelining "abnormal residuals" that differ drastically from other residuals.
- ✓ Dataset is slightly skewed to the right (0.34) and possesses low kurtosis (-0.74).

Visualizations

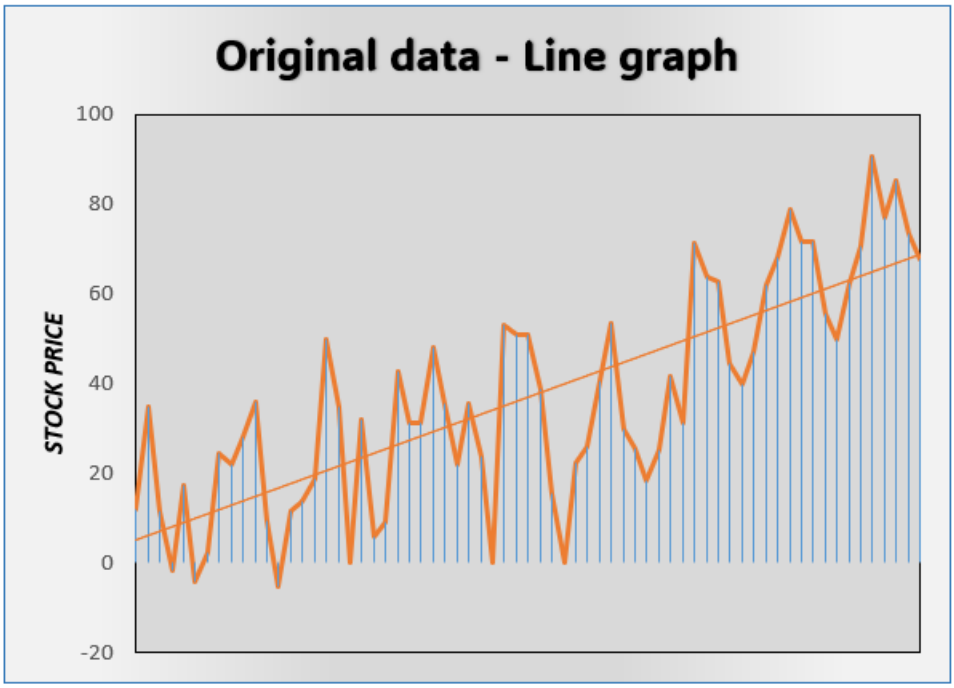


Fig 1. Flow of data over time (Trend) and presence of negative values in original data.

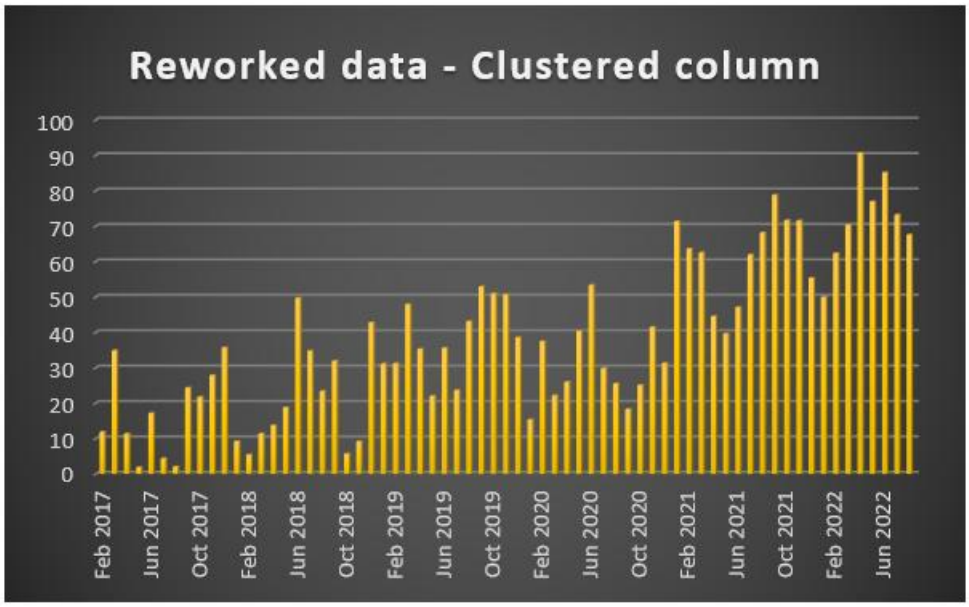


Fig 2. Distribution of data over a 5-year period after data cleaning.

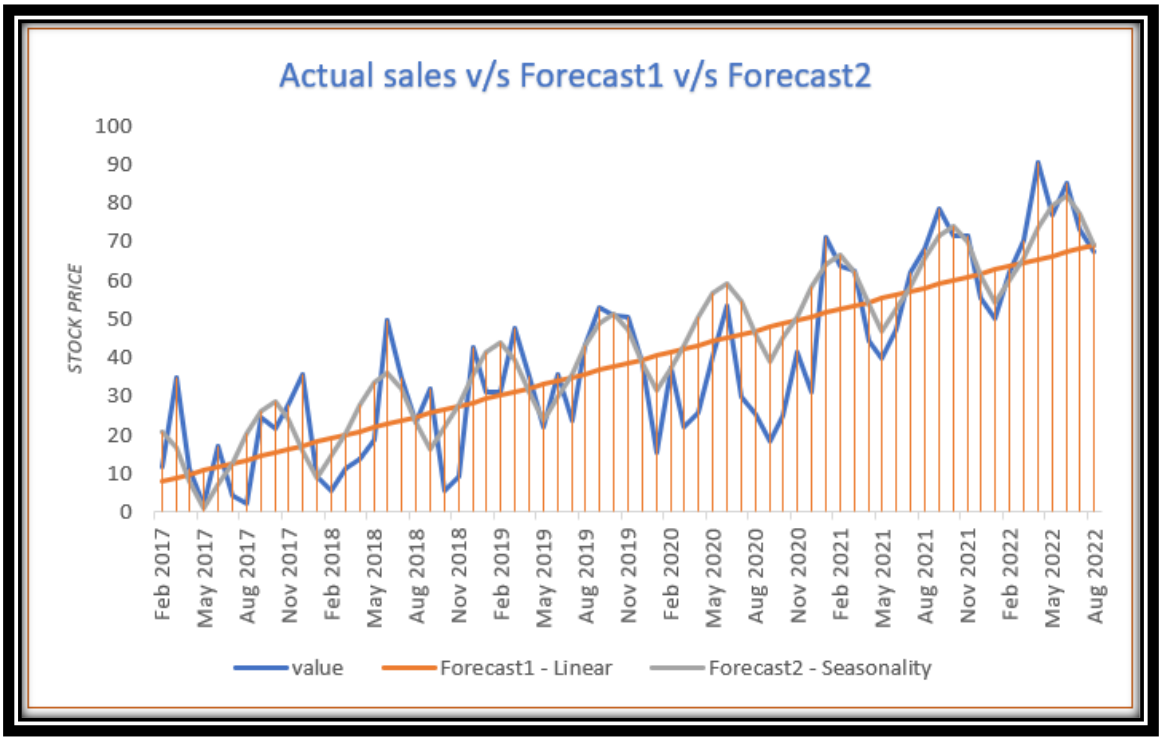


Fig 3. A comparison between actual values, values from linear forecast and seasonal forecast functions.

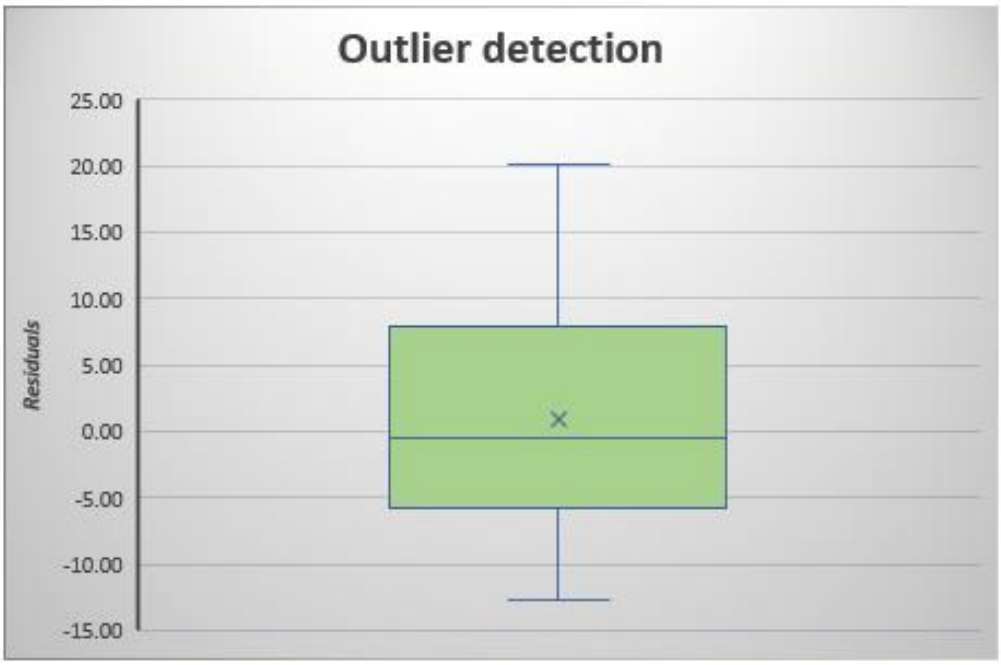


Fig 4. Boxplot to observe presence of outliers (residual values)

4. Modelling and evaluation →

Since our dataset has clear trend and seasonality, I am going to build two different time series algorithms that extract the same:

a) Linear Trend + Seasonal Model

b) Linear Trend + Seasonal Model with Dummy Variables

Note: a model like moving average (MA) or exponential smoothing would not be able to capture seasonality and trend components present in time series.

a) Linear Trend + Seasonal Model

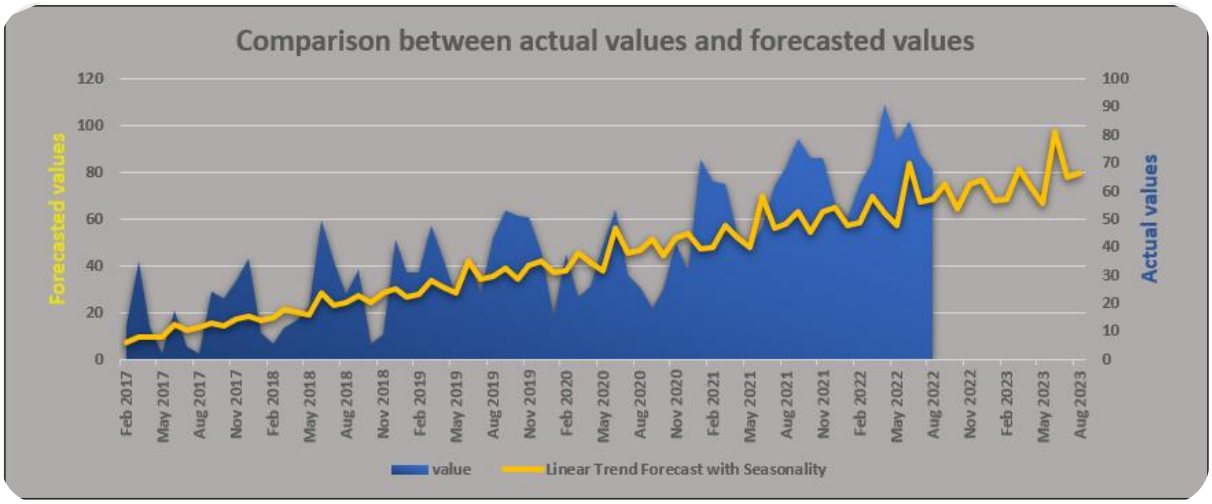
This model is used when **data has seasonality present at certain time periods along with trend**. Each month is imparted with a **seasonal index**, which would capture seasonality that is dependent on the months of a year.

Upon forecasting, the model produced the following numbers after evaluation.

MAD	**11.04**
MAPE	**59.00%**
MSE	**179.67**
Bias	**-0.02**

- ✓ MAD indicates the **forecasted values deviate** from the observed values by 11.04 units on average.
- ✓ MAPE indicates that forecasts are off by 59% on average.
- ✓ High MSE of 179.67 indicates the **presence of large deviations** in our calculations.
- ✓ Bias of -0.02 indicates that, on average, forecasts are slightly above actual values.

Fig 5. The graph explicitly visualizes the variation in forecast and observed values



Sep 2022	74.93
Oct 2022	64.47
Nov 2022	74.85
Dec 2022	76.91
Jan 2023	67.68
Feb 2023	68.38
Mar 2023	81.41
Apr 2023	73.27
May 2023	66.80
Jun 2023	97.55
Jul 2023	78.04
Aug 2023	79.67

Forecasted values for Sep 2022 to Aug 2023

b) Linear Trend + Seasonal Model with Dummy Variables

Building a time series model with dummy variables helps us to record seasonality along with other categorical effects associated with specific time periods.

Since our data has seasonality of 8, I **create 7 dummy variables** to avoid multicollinearity. After assigning dummy variables to each time period, we take up the intercept coefficient and add it to the corresponding product of dummy variable and its coefficients. This calculation gives us the forecasted value based on an 8-month seasonality period.

Note: Coefficients are calculated using regression techniques.

MAD	**7.99**
MAPE	**42.46%**
MSE	**106.82**
Bias	**0.00**

- ✓ Although the evaluation metrics show some improvement over previous model, a high MAD score of 7.99 indicates that there is significant difference between actual and forecasted values. However, a bias of 0 denotes that the **model does not consistently over-forecast or under-forecast**, but rather deviates randomly.

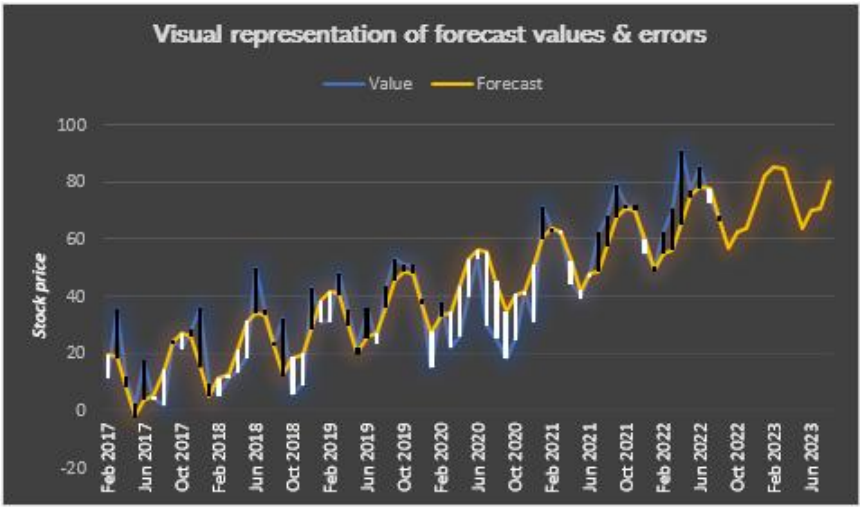


Fig 6. The graph captures the variations in forecasting (black lines indicate under-forecasting and white lines over-forecasting)

Sep 2022	56.51
Oct 2022	62.45
Nov 2022	63.55
Dec 2022	72.70
Jan 2023	82.30
Feb 2023	85.45
Mar 2023	84.77
Apr 2023	74.11
May 2023	63.84
Jun 2023	69.78
Jul 2023	70.87
Aug 2023	80.03

Forecasted values for Sep 2022 to Aug 2023

5. Insights and conclusions →

- A seasonality period of **approximately 8 months** is present in the dataset.

- A maximum **stock price of 90.71** was observed in April 2022, which occurs slightly earlier to designated 8-month seasonality period since previous peak was

observed in Oct 2021. This suggests **random variations in data** and these fluctuations are not captured by model.

- Similarly, many other instances of data variations in observed values are present that stray away from the 8-month seasonality pattern. This **indifference in data distribution** poses a challenge to correctly forecast future values as **model cannot comprehend external factors apart from seasonality**.

- Overall, the stock value of business is steadily growing upwards, hitting peak stock prices approximately 8 months once. The value surge could be **influenced by a number of aspects** such as sales cycles, new product development etc.

- To accurately forecast values, a **more sophisticated model** that takes multiple factors into account (at once) should be built and should be chosen over algorithms used above.