Kwanza Tukule Data Analyst Assessment

Moffat Kagiri Ngugi

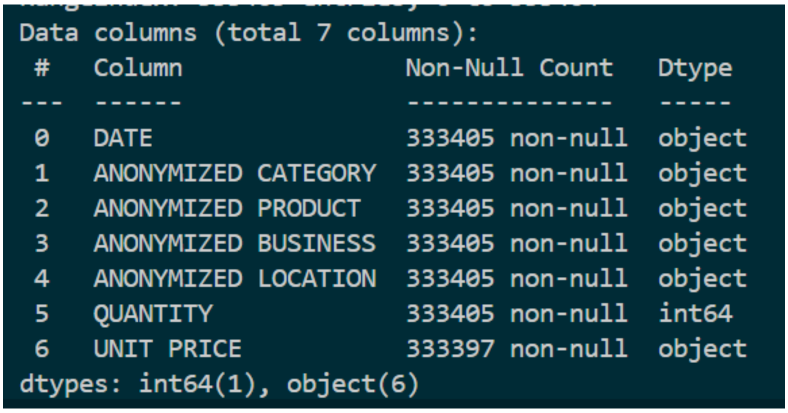
kagirimoffat@outlook.com

Introduction

Sales data is a valuable resource for businesses in any and every industry today. It enables them to determine their strengths and weaknesses and, determine where they need to put more effort to improve their profitability. In this assessment, I analyse the sales data from a model firm whose businesses and products are anonymized. Regardless of this obscurity, the analysis demonstrates impeccable clarity in its findings and the potential value that would arise by implementing the insights.

Data Cleaning and Preparation

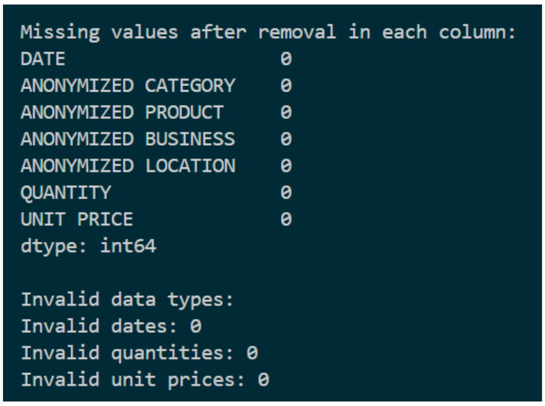
The raw sales data requires a process of cleaning to detect and replace null, missing, or invalid entries. To do this, I used a python script in VS Code to summarize the data and determine whether there were any null values. A preliminary understanding of the raw data was essential in forming a starting point for the analysis. As shown below, the data had a varying count of valid entries, indicating the need for cleaning.



```
Data columns (total 7 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   DATE                  333405 non-null  object
 1   ANONYMIZED CATEGORY   333405 non-null  object
 2   ANONYMIZED PRODUCT    333405 non-null  object
 3   ANONYMIZED BUSINESS   333405 non-null  object
 4   ANONYMIZED LOCATION   333405 non-null  object
 5   QUANTITY              333405 non-null  int64
 6   UNIT PRICE            333397 non-null  object
dtypes: int64(1), object(6)
```

Figure 1 Summary of the raw data

The difference between the counts arose as a result of invalid entries in the unit price column, which I removed and validated that the remaining data was all valid and fit for analysis.

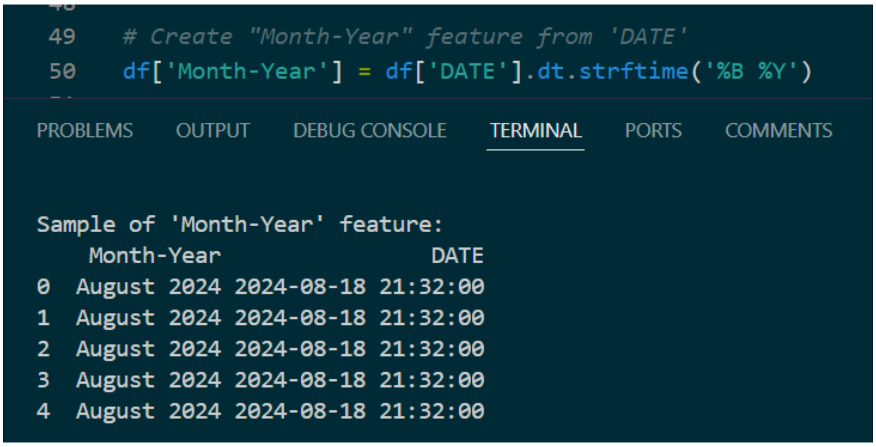


```
Missing values after removal in each column:
DATE                   0
ANONYMIZED CATEGORY    0
ANONYMIZED PRODUCT     0
ANONYMIZED BUSINESS    0
ANONYMIZED LOCATION    0
QUANTITY               0
UNIT PRICE             0
dtype: int64

Invalid data types:
Invalid dates: 0
Invalid quantities: 0
Invalid unit prices: 0
```

Figure 2 Missing or invalid values after removing invalid values in each column

With the data cleaned, I proceeded to add the ‘Month-Year’ column in python as follows:



```
# Create "Month-Year" feature from 'DATE'
df['Month-Year'] = df['DATE'].dt.strftime('%B %Y')
```

```
Sample of 'Month-Year' feature:
    Month-Year                DATE
0  August 2024 2024-08-18 21:32:00
1  August 2024 2024-08-18 21:32:00
2  August 2024 2024-08-18 21:32:00
3  August 2024 2024-08-18 21:32:00
4  August 2024 2024-08-18 21:32:00
```

Figure 3 The 'Month-Year' Feature

The above feature would facilitate ease in understanding the data as it were, and even modelling and analysing the time series as needed. With the data cleaning and necessary transformation complete, I could now proceed to carry out some exploratory analysis to further interrogate the dataset.

Exploratory Data Analysis

To seek out the clear trends in the sales data, I grouped the data by the category as well as the business and investigated the different trends they exhibited. This aggregation showed the varied value of each business and category as it showed the respective figures for each category's turnover for every month. To have a better understanding of the varied levels of success across business and categories, I defined a variable 'VALUE' which is a product of the product price and the quantity of sales.

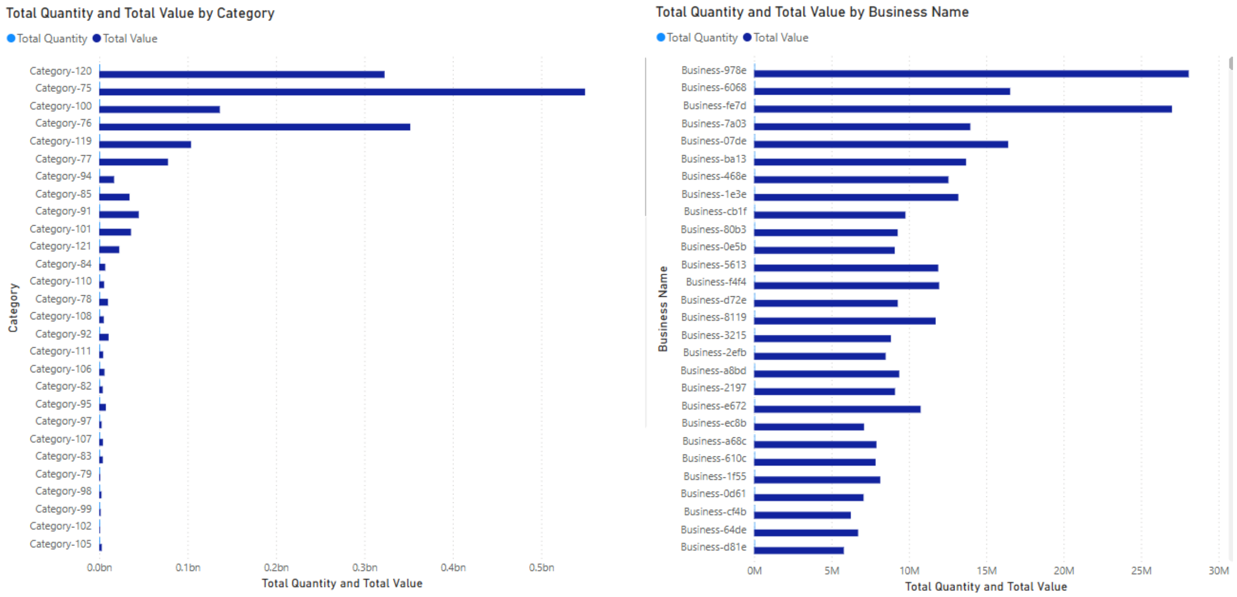


Figure 4 Sales Overview (Quantity & Value by Category and Business)

As the graphs above show, the distribution of value across categories and businesses is highly disproportionate, with less than 20% of the categories and businesses contributing over 60% of the total revenue.

Given the nature of the sales data as a time series, it was also essential to investigate the seasonality of the sales to determine how the sales fluctuated in time. The graph below shows the best illustration of this seasonality.

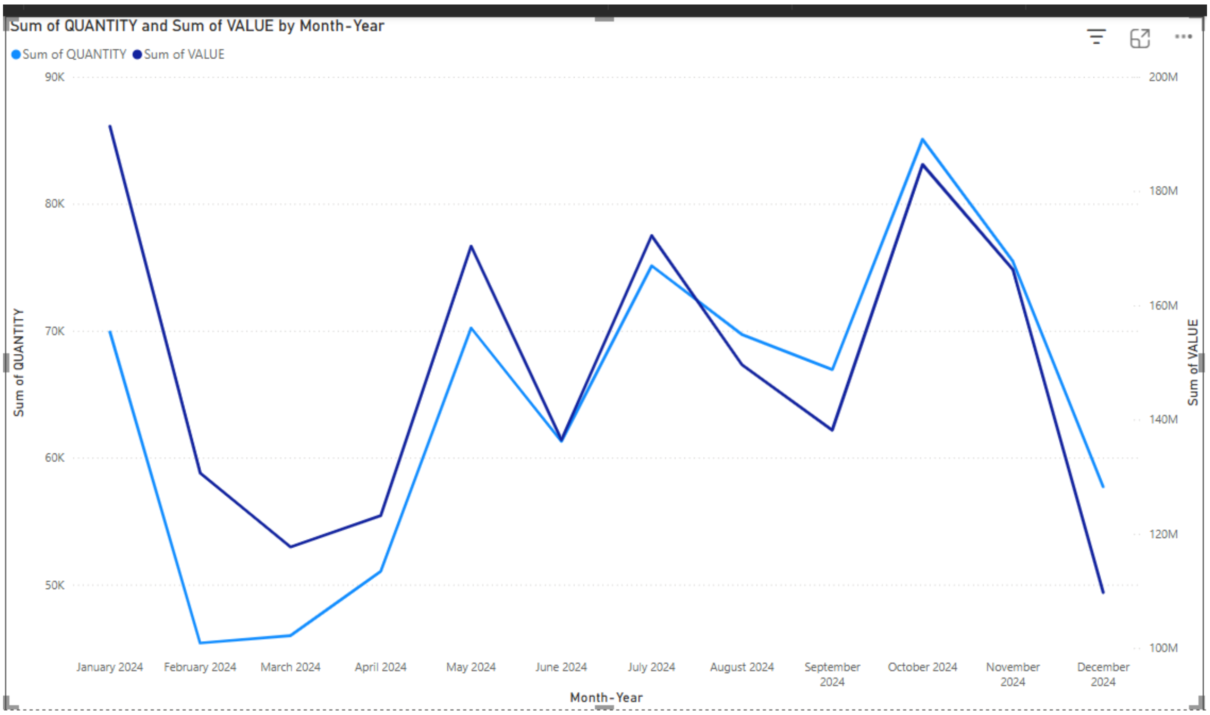
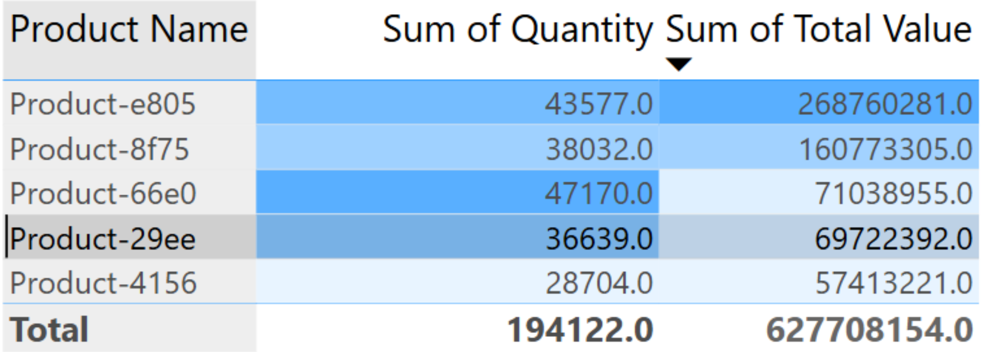


Figure 5 Total Quantity and Value over time

As the data shows, this firm had 4 peak seasons in the year. In these peaks, the firm had sales exceeding the average of the remaining months.

In further exploration, it was necessary to understand the best performing products, categories, and businesses. For this reason, I investigated the products with the highest value in the entire period for which the data was available.



Product Name	Sum of Quantity	Sum of Total Value
Product-e805	43577.0	268760281.0
Product-8f75	38032.0	160773305.0
Product-66e0	47170.0	71038955.0
Product-29ee	36639.0	69722392.0
Product-4156	28704.0	57413221.0
Total	**194122.0**	**627708154.0**

Figure 6 Top 5 Performing Products

As the table above shows, the top 5 products in the total quantities sold were also the ones leading in the total value sold. As such, this firm can be seen making more use of the economics of scale. Where it sells more quantities, it also makes more in revenues. With this understanding came the need to understand just how much different customer segments contributed to the revenue. Having learned enough from the data to make some hypotheses on how different variables were correlated, I concluded the exploratory data analysis and proceeded for a deeper look at the data and the patterns that emerged.

Advanced Analysis

Given the disparities that came up in the exploratory analysis, it was important to review the customer or business profile for the firm. To determine the customer segment that was most profitable to the company, I needed to cluster the data and figure out what factors created or supported the similarities between different customer segments. The first form of clustering was rule-based. With this clustering, I tabulated the total value of the products by their respective segments. The high value segment refers to the businesses that exceeded 75% of the highest value, while the medium value is the cluster that exceeded 25%. As shown below, the high value segment exceeded 90%, which means that the high value segment contributes disproportionately to the total revenue of the firm.

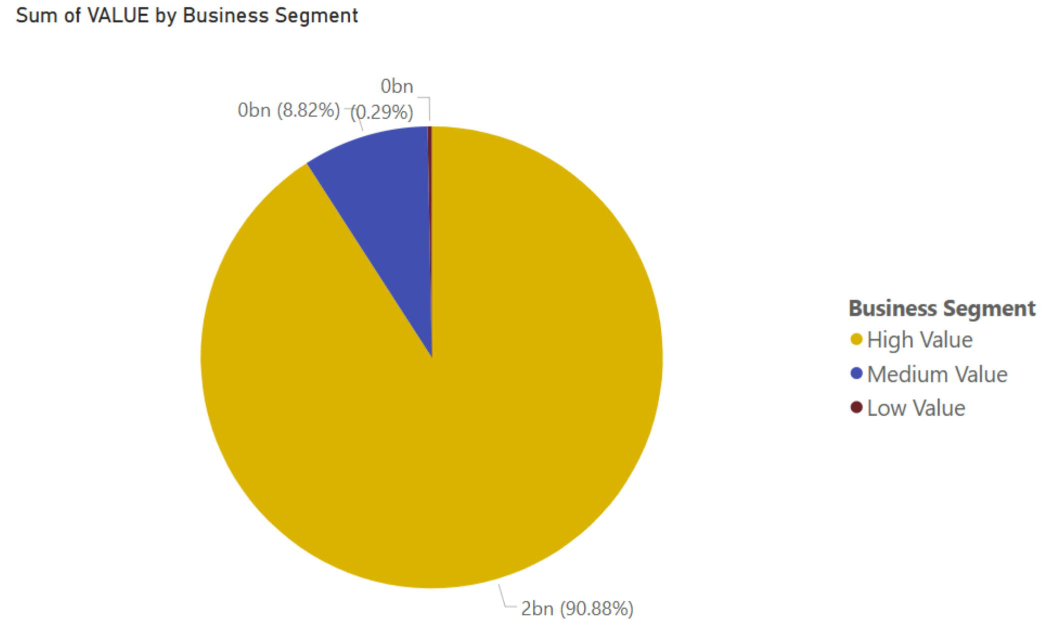
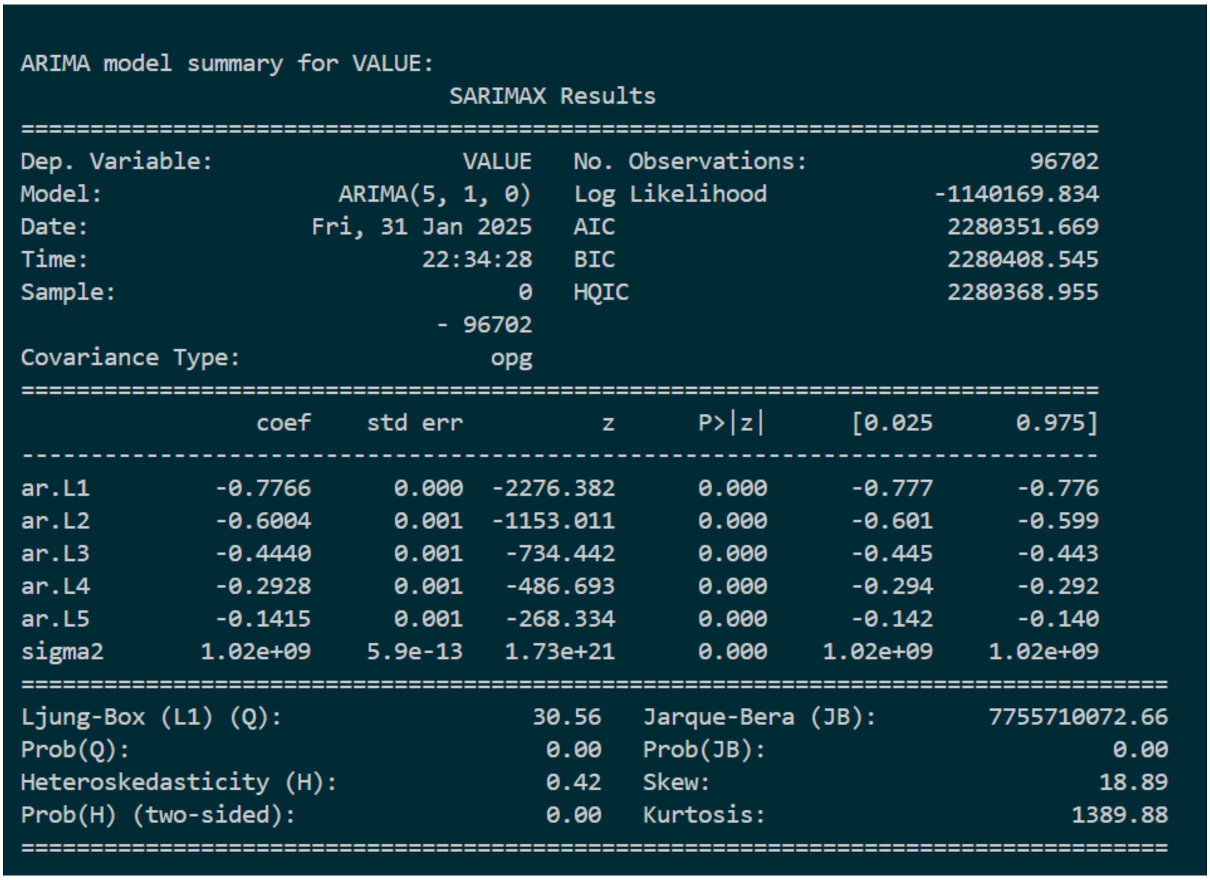


Figure 7 Total value by business segments

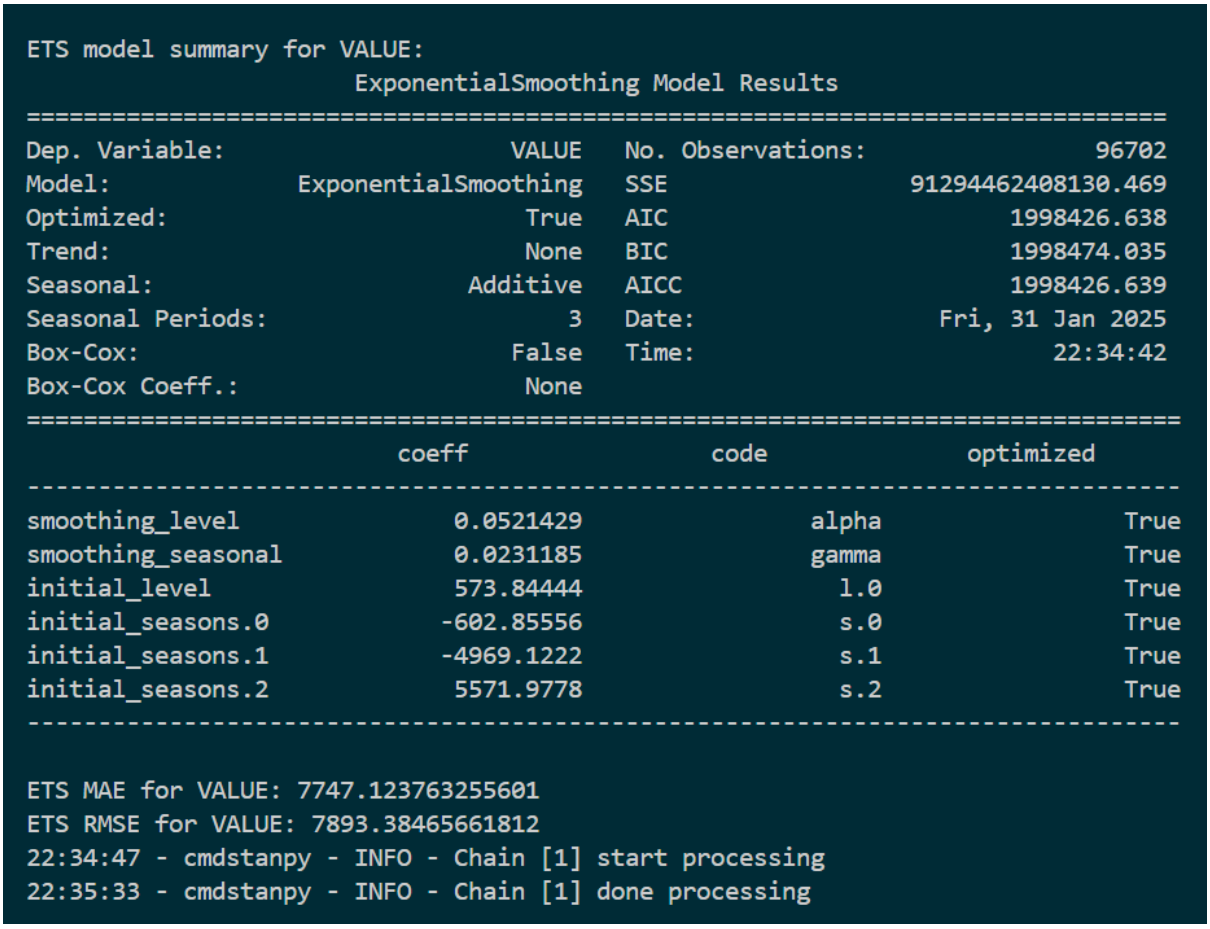
Besides the rule-based clustering method, this analysis employed k-means clustering model to cluster the data points based on their similarities. This facilitated a more flexible and data driven approach to complex patterns without the need for explicitly dictating the rules. The K-means clustering model isolated 3 clusters, just as the rule-based method. With these clusters, I could proceed to conduct a forecast of the sales for the 3 months after the period for which I had the data.

I carried out the forecasting using 3 models: ARIMA, Prophet, and ETS. ARIMA stands for AutoRegressive Integrated Moving Average. It refers to a model backed by a mathematical model that represents time series using its past values. On the other hand, Prophet is an algorithm that was developed by Facebook specifically to predict business time series. Lastly, ETS is useful in making forecasts and recommendations although it lacks as good a fit to test data as ARIMA has. The results of the ARIMA, Prophet, and ETS models fit on the data is shown in the table below.



```
ARIMA model summary for VALUE:
                               SARIMAX Results
==============================================================================
Dep. Variable:                  VALUE   No. Observations:                96702
Model:                 ARIMA(5, 1, 0)   Log Likelihood             -1140169.834
Date:                Fri, 31 Jan 2025   AIC                         2280351.669
Time:                        22:34:28   BIC                         2280408.545
Sample:                             0   HQIC                        2280368.955
                              - 96702
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.7766      0.000  -2276.382      0.000      -0.777      -0.776
ar.L2         -0.6004      0.001  -1153.011      0.000      -0.601      -0.599
ar.L3         -0.4440      0.001   -734.442      0.000      -0.445      -0.443
ar.L4         -0.2928      0.001   -486.693      0.000      -0.294      -0.292
ar.L5         -0.1415      0.001   -268.334      0.000      -0.142      -0.140
sigma2       1.02e+09    5.9e-13   1.73e+21      0.000    1.02e+09    1.02e+09
===================================================================================
Ljung-Box (L1) (Q):                  30.56   Jarque-Bera (JB):         7755710072.66
Prob(Q):                              0.00   Prob(JB):                          0.00
Heteroskedasticity (H):               0.42   Skew:                             18.89
Prob(H) (two-sided):                  0.00   Kurtosis:                       1389.88
===================================================================================
```

Figure 8 ARIMA model summary for VALUE

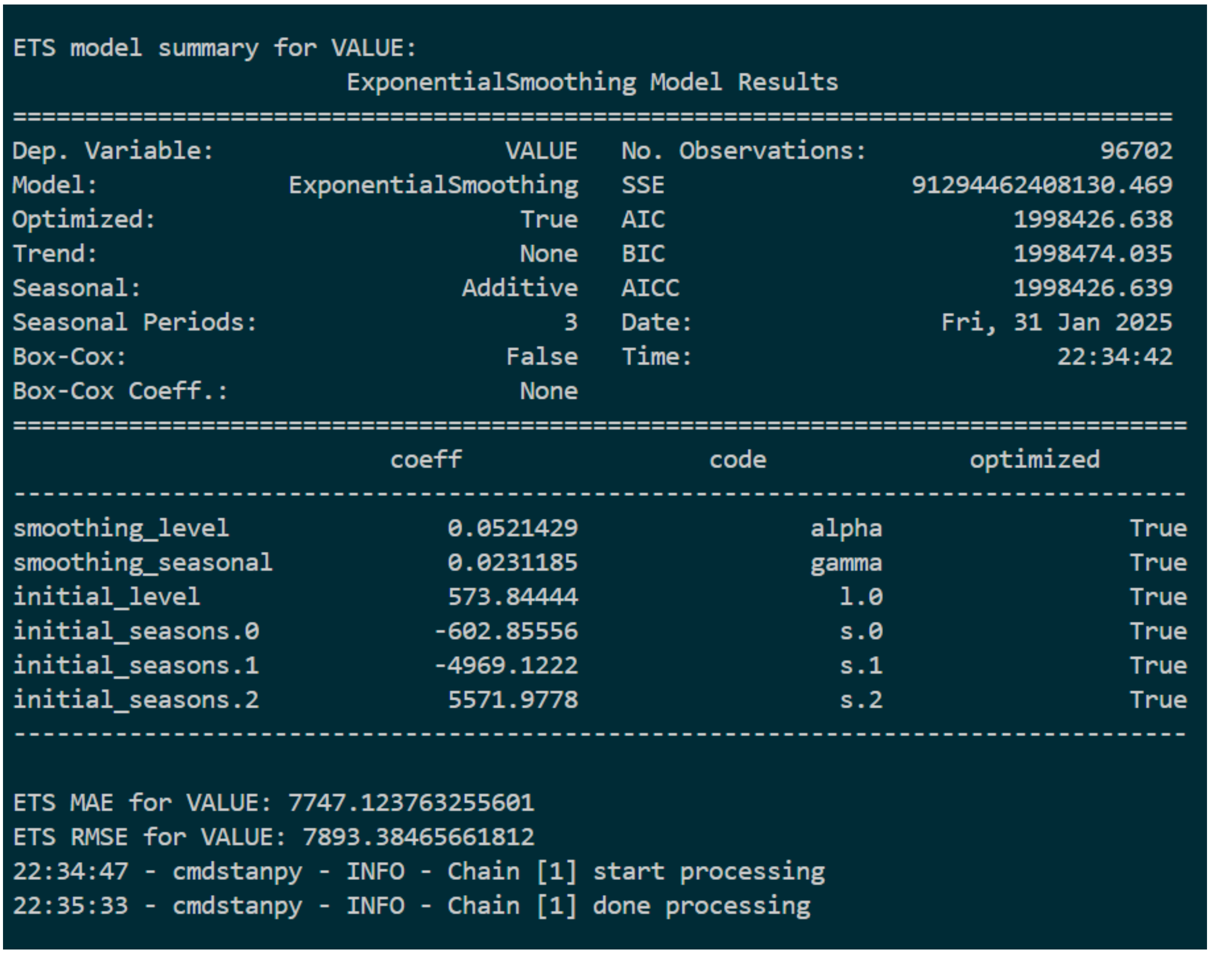


```
ETS model summary for VALUE:
                       ExponentialSmoothing Model Results
================================================================================
Dep. Variable:                     VALUE   No. Observations:                96702
Model:              ExponentialSmoothing   SSE                 91294462408130.469
Optimized:                          True   AIC                        1998426.638
Trend:                              None   BIC                        1998474.035
Seasonal:                       Additive   AICC                       1998426.639
Seasonal Periods:                      3   Date:                 Fri, 31 Jan 2025
Box-Cox:                           False   Time:                         22:34:42
Box-Cox Coeff.:                     None
================================================================================
                          coeff                 code              optimized
---------------------------------------------------------------------------------
smoothing_level               0.0521429                alpha                 True
smoothing_seasonal            0.0231185                gamma                 True
initial_level                 573.84444                  l.0                 True
initial_seasons.0            -602.85556                  s.0                 True
initial_seasons.1            -4969.1222                  s.1                 True
initial_seasons.2             5571.9778                  s.2                 True
---------------------------------------------------------------------------------

ETS MAE for VALUE: 7747.123763255601
ETS RMSE for VALUE: 7893.38465661812
22:34:47 - cmdstanpy - INFO - Chain [1] start processing
22:35:33 - cmdstanpy - INFO - Chain [1] done processing
```

Figure 9 ExponentialSmoothing Model Results

On the other hand, the forecasting from the ETS model is shown below:

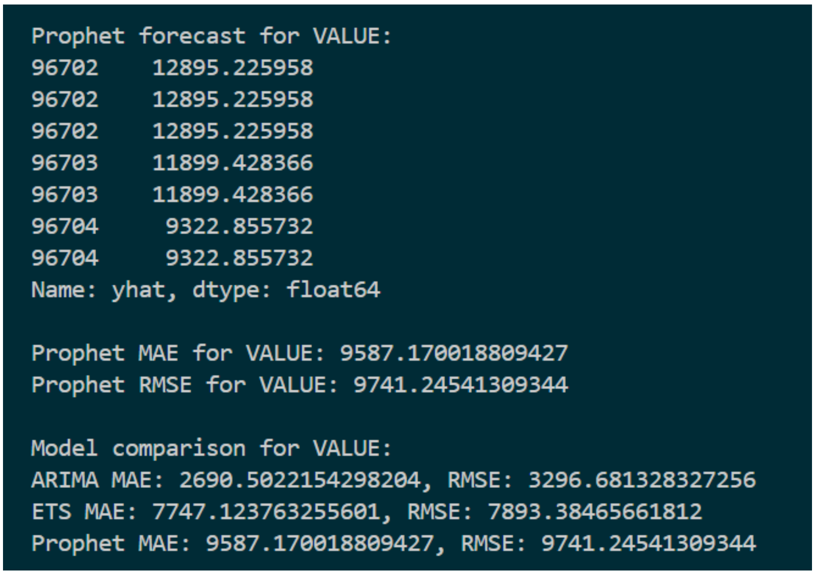


```
ETS model summary for VALUE:
                       ExponentialSmoothing Model Results
================================================================================
Dep. Variable:                    VALUE   No. Observations:                96702
Model:             ExponentialSmoothing   SSE                 91294462408130.469
Optimized:                         True   AIC                        1998426.638
Trend:                             None   BIC                        1998474.035
Seasonal:                      Additive   AICC                       1998426.639
Seasonal Periods:                     3   Date:                 Fri, 31 Jan 2025
Box-Cox:                          False   Time:                         22:34:42
Box-Cox Coeff.:                    None
=================================================================================
                          coeff                 code              optimized
---------------------------------------------------------------------------------
smoothing_level               0.0521429                alpha                True
smoothing_seasonal            0.0231185                gamma                True
initial_level                 573.84444                  l.0                True
initial_seasons.0            -602.85556                  s.0                True
initial_seasons.1            -4969.1222                  s.1                True
initial_seasons.2             5571.9778                  s.2                True
---------------------------------------------------------------------------------

ETS MAE for VALUE: 7747.123763255601
ETS RMSE for VALUE: 7893.38465661812
22:34:47 - cmdstanpy - INFO - Chain [1] start processing
22:35:33 - cmdstanpy - INFO - Chain [1] done processing
```

Figure 10 ETS model summary for VALUE

To determine which model carried out the forecast in the most efficient manner and with the least error, it was just as critical to review their error rates using MAE and RMSE. This would help to determine the most efficient forecasting model. As the error metrics below suggest, ARIMA prevailed over Prophet and ETS. This is because it had the smallest errors.



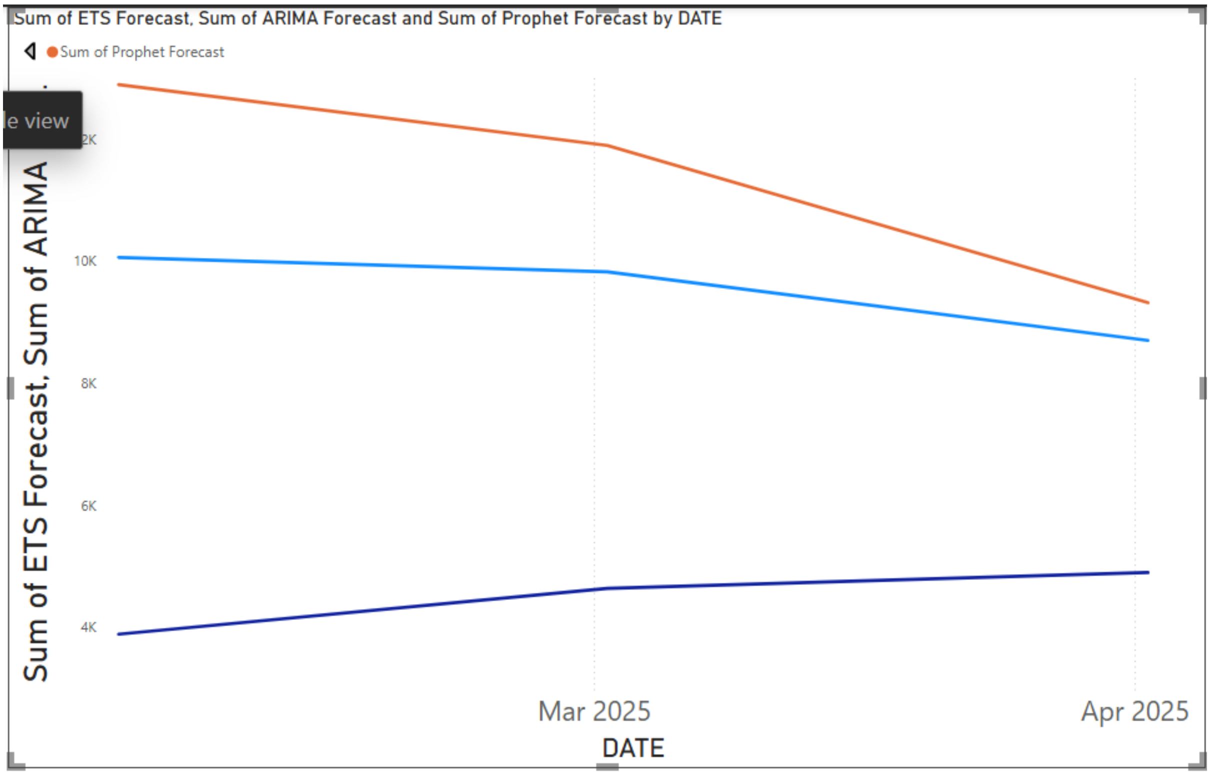
```
Prophet forecast for VALUE:
96702    12895.225958
96702    12895.225958
96702    12895.225958
96703    11899.428366
96703    11899.428366
96704     9322.855732
96704     9322.855732
Name: yhat, dtype: float64

Prophet MAE for VALUE: 9587.170018809427
Prophet RMSE for VALUE: 9741.24541309344

Model comparison for VALUE:
ARIMA MAE: 2690.5022154298204, RMSE: 3296.681328327256
ETS MAE: 7747.123763255601, RMSE: 7893.38465661812
Prophet MAE: 9587.170018809427, RMSE: 9741.24541309344
```

Figure 11 Error Metrics

Having evaluated the models' efficacy and created a thorough understanding of the seasonality and correlation in the data, we have the context necessary to understand the forecasts from the respective models. The graph below shows the forecasts for the next 3 months.



Strategic Insights and Recommendations

1. Focus on High-Value Segments

The high-value segment contributes over 90% of the firm's revenue. The company should prioritize maintaining strong relationships with these customers, offering tailored incentives, and ensuring consistent product availability.

2. Optimizing Category and Business Contributions

Less than 20% of the categories and businesses contribute over 60% of total revenue. The company should allocate more resources to these high-performing categories while assessing underperforming ones for potential improvement or discontinuation.

3. Leverage Seasonal Trends for Inventory and Marketing Strategies

The business experiences four peak seasons annually. This pattern should guide inventory management, marketing campaigns, and pricing strategies to maximize sales during high-demand periods.

4. Enhancing Forecasting Accuracy

ARIMA demonstrated the lowest error rates (MAE & RMSE) in forecasting sales, making it the most reliable model for predicting future trends. The company should integrate ARIMA-based forecasting into decision-making processes for demand planning and financial projections.

5. Product Performance Optimization

The top five products in total quantity sold also led in revenue. Given this, the firm should invest in optimizing production, distribution, and marketing for these key products to capitalize on their success.

6. Refining Customer Segmentation

Both rule-based clustering and K-Means clustering identified three key customer segments. The firm should leverage this insight to develop targeted pricing strategies, loyalty programs, and personalized marketing efforts to maximize engagement and revenue.

Conclusion

This analysis has provided valuable insights into the company's sales trends, key revenue drivers, and customer segmentation. By leveraging data-driven strategies—such as focusing on high-value customers, optimizing product categories, and refining sales forecasts—the firm can enhance profitability and efficiency. With ARIMA proving to be the most effective forecasting model, the company can better anticipate demand and make informed business decisions. Implementing these recommendations will position the company for sustained growth and competitive advantage.