SHIVAN PRAMLALL

EDUV4945866@VOSSIE.NET

ITSFA

PROJECT 1

2024

**Table of Contents**

[**QUESTION 1** 3](#_Toc171719399)

[a. Perform any data cleaning and transformation steps needed. 3](#_Toc171719400)

[b. What insights can you gather based on the total sales per state? 3](#_Toc171719401)

[c. Choose two other meaningful columns to plot based on the sales data. Justify your choice of columns and explain what insights you can derive from the plots. 4](#_Toc171719402)

[d. Plot the total sales data per month, as well as the total sales per month for each unique category (furniture supplies, office supplies, technology) in the Category column. What insights, trends and patterns can be derived from this plot? 6](#_Toc171719403)

[**QUESTION 2** 8](#_Toc171719404)

[a. Plot the monthly decomposition sales for trends and seasonality, and comment on the results. 8](#_Toc171719405)

[b. Generate the ACF and PACF plots and comment on the results. 8](#_Toc171719406)

[c. Check for stationarity using the Dickey-Fuller test and comment on the results. 9](#_Toc171719407)

[**QUESTION 3** 10](#_Toc171719408)

[a. Model the Holt-Winters Exponential Smoothing with the Order date and sales. Use the TotalSalesPerMonth table to model it per month 10](#_Toc171719409)

[b. Assess the Model fit/performance above and comment on the results. 10](#_Toc171719410)

[c. Forecast and plot the sales for the next 12 months. Does the results make sense? 11](#_Toc171719411)

[**QUESTION 4** 12](#_Toc171719412)

[Fitting and assessing AR, MA, ARIMA and SARIMA models 12](#_Toc171719413)

[**REFERENCES** 15](#_Toc171719414)

# **QUESTION 1**

a. Perform any data cleaning and transformation steps needed.

Postal code has 11 null values. All the postal code for Burlington city East are null. These values can’t be replaced since there aren't any values where Burlington city East are not null. So, I dropped these null values. Also, column named “Unnamed: 0” was dropped as it is an unnecessary column.

b. What insights can you gather based on the total sales per state?

A screenshot of a cell phone

Description automatically generated

1. The data contains 48 states from USA.
2. The average sales in USA is $46929.32
3. 75% of the states have total sales are <= 48343.93
4. 25% of the states have total sales are > 48343.93

A table with numbers and a state

Description automatically generated

Shown above are the top 5 states with highest total sales.

A table with numbers and a state

Description automatically generated

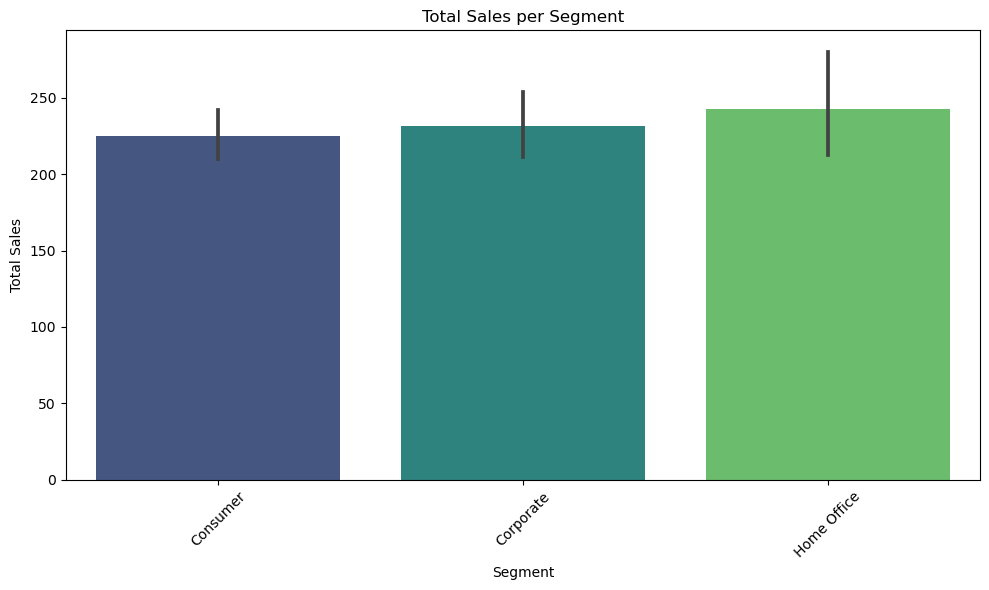
Shown above are the top 5 states with lowest total sales.

c. Choose two other meaningful columns to plot based on the sales data. Justify your choice of columns and explain what insights you can derive from the plots.

A graph of sales per region

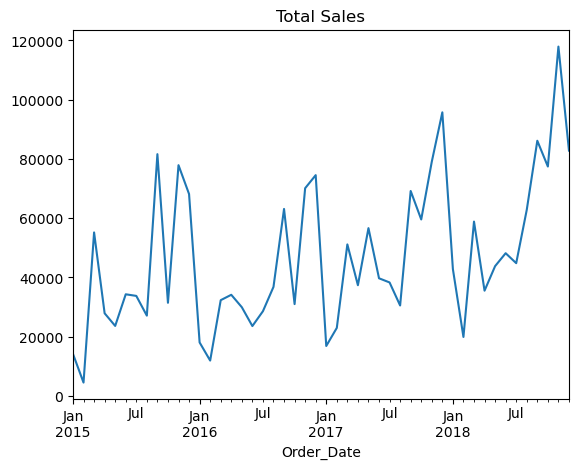
Description automatically generated

**From the plot above, we can see that from all regions in the US, South and East regions have the highest sales.**



**From the plot above, we can see that from all the different product segments. Home Office has the highest sales followed by Corporate products and them Consumer products.**

## d. Plot the total sales data per month, as well as the total sales per month for each unique category (furniture supplies, office supplies, technology) in the Category column. What insights, trends and patterns can be derived from this plot?



**Shown above is the total sales per month. We can see that there is an increasing trend. Also there is some pattern/seasonality present.**

A graph of sales

Description automatically generated

**Furniture sales are most in Decmeber. (Maybe due to Christmas deals)**

**Office Supplies sales are most in August.**

**Technology product sales are most in November. (Maybe due to Cyber monday deals)**

# **QUESTION 2**

1. Plot the monthly decomposition sales for trends and seasonality, and comment on the results.

A graph of blue lines

Description automatically generated with medium confidence

1. Generate the ACF and PACF plots and comment on the results.

A graph with blue lines and dots

Description automatically generated

A graph with blue dots and lines

Description automatically generated

***From the ACF and PACF plot, we can see that there is a correlation at lag = 1 and lag = 12***

1. Check for stationarity using the Dickey-Fuller test and comment on the results.

***Performing the Dickey-Fuller test for stationarity***

H0: The time series is non-stationary.

H1: The time series is stationary.

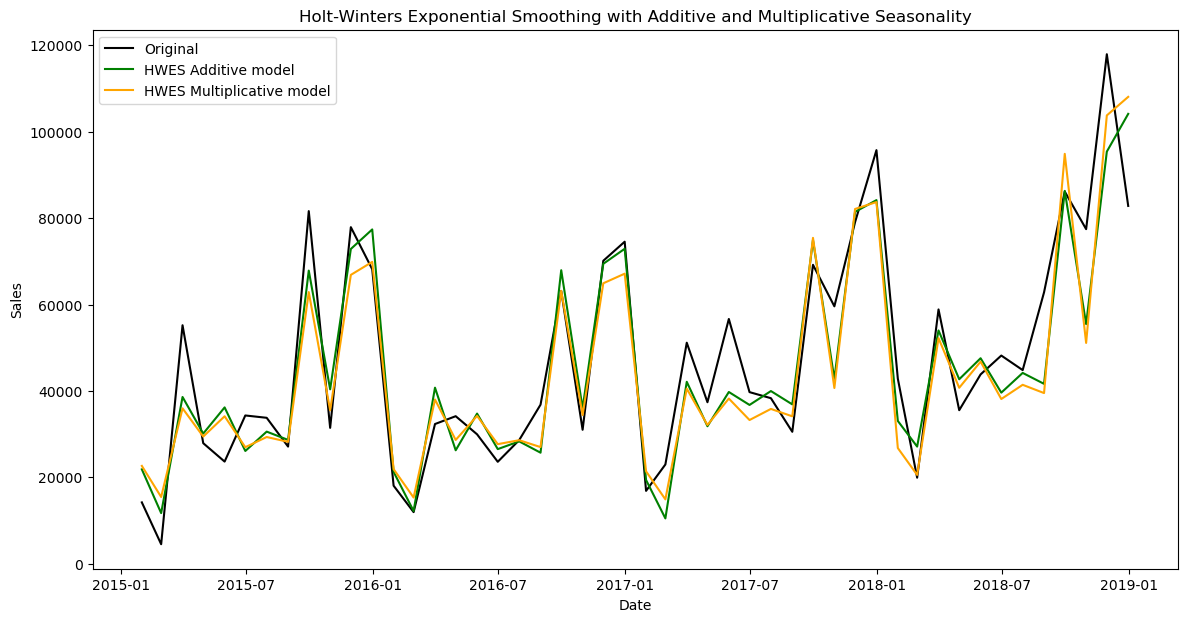
Test statistics = -4.425

P-Value = 0.0003

Since the p-value is less than .05, we reject the null hypothesis. This means the time series is stationary.

# **QUESTION 3**

a. Model the Holt-Winters Exponential Smoothing with the Order date and sales. Use the TotalSalesPerMonth table to model it per month



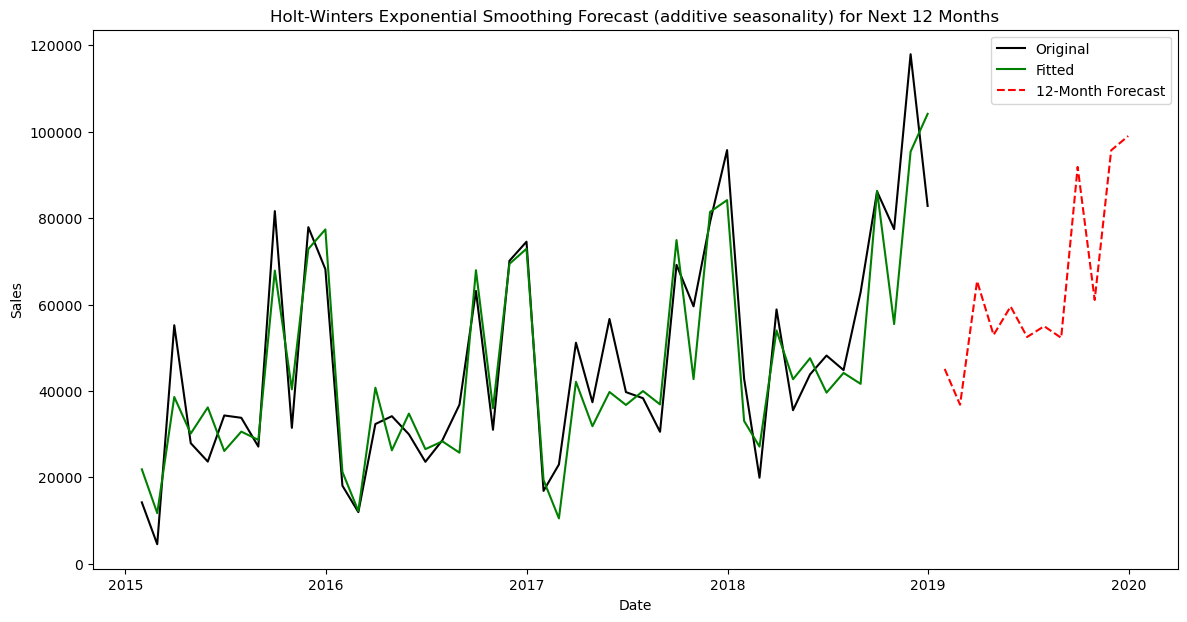
## b. Assess the Model fit/performance above and comment on the results.

|  |  |  |  |
| --- | --- | --- | --- |
| **HWES Model**  **with:** | **Mean Squared Error (MSE):** | **Root Mean Squared Error (RMSE):** | **Mean Absolute Error (MAE):** |
| Additive Seasonality | 96479894.12 | 9822.42 | 7714.47 |
| Multiplicative Seasonality | 111120101.32 | 10541.35 | 8210.46 |

**Conclusion:**

Based on these performance metrics, the HWES model with additive seasonality model generally outperforms the multiplicative seasonality model in terms of prediction accuracy. It demonstrates lower errors across all metrics (MSE, RMSE, MAE), suggesting that it captures the underlying patterns and seasonality more effectively than the HWES model with multiplicative seasonality.

c. Forecast and plot the sales for the next 12 months. Does the results make sense?



The results do make sense as the forecasted values do possess trend and seasonality. Also, we can see the similar patterns being forecasted.

# **QUESTION 4**

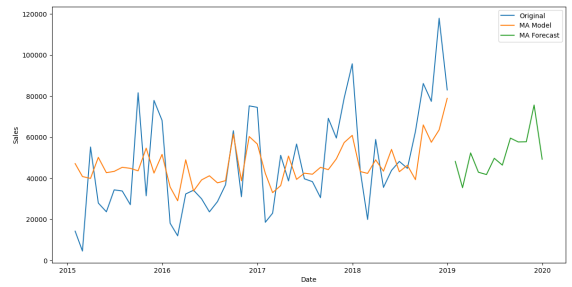
## Fitting and assessing AR, MA, ARIMA and SARIMA models

**AutoRegression (AR) Model**

A graph of blue and orange lines

Description automatically generated

**Moving Average (MA) Model**



4.1.b. the AR model seems to fit the data better than the MA model.

**ARIMA Model**

A graph of blue and orange lines

Description automatically generated

4.2.b. The ARIMA model fits makes sense as it captures the trend.

**SARIMA Model**

A graph with blue and orange lines

Description automatically generated

4.3.b. The SARIMA seems to fit the data the best as it captures the trend and seasonal components .

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Mean Squared Error (MSE):** | **Root Mean Squared Error (RMSE):** | **Mean Absolute Error (MAE** |
| AR | 319595396.65 | 17877.23 | 14451.01 |
| MA | 387595628.06 | 19687.45 | 16231.75 |
| ARIMA | 306360185.14 | 17503.15 | 13556.22 |
| SARIMA | 304892173.94 | 17461.16 | 12761.9 |

**Conclusion:**

Based on these performance metrics, SARIMA appears to be the best performing model among the ones tested for your dataset. It demonstrates lower errors across all metrics (MSE, RMSE, MAE), suggesting that it captures the underlying patterns and seasonality in your data more effectively than AR, MA, and ARIMA models.

# **REFERENCES**

* https://www.statology.org/dickey-fuller-test-python/
* https://www.youtube.com/watch?v=y8opUEd05Dg&t=7s
* https://www.youtube.com/watch?v=YBelhLOohpU&t=517s
* https://youtube.com/watch?v=1opjnegd\_hA
* <https://medium.com/analytics-vidhya/python-code-on-holt-winters-forecasting-3843808a9873>
* <https://spureconomics.com/interpreting-acf-and-pacf-plots/>
* <https://www.kaggle.com/code/carlmcbrideellis/time-series-a-simple-moving-average-ma-model>
* <https://buse-koseoglu13.medium.com/guide-to-time-series-analysis-with-python-3-autoregressive-process-3f16857d3e8a>
* <https://www.geeksforgeeks.org/autoregressive-ar-model-for-time-series-forecasting/>
* <https://developer.ibm.com/tutorials/awb-creating-autoregressive-models-in-python/>
* <https://www.youtube.com/watch?v=xDVOwL1fAZI>
* <https://spotintelligence.com/2023/10/25/autoregressive-ar-models/>
* <https://www.geeksforgeeks.org/time-series-decomposition-techniques/>
* <https://subashpalvel.medium.com/time-series-forecasting-with-seasonal-decomposition-b86897a4b68b>
* <https://www.kaggle.com/code/iamleonie/time-series-interpreting-acf-and-pacf>
* <https://itsudit.medium.com/deciphering-acf-and-pacf-plots-a-guide-to-time-series-forecasting-3323948935fb>
* https://www.youtube.com/watch?v=mQ352TiPvBI&t=252s