Analyzing Model Performance of the online selling from India: Case study on Amazon Sales Report 2022

Executive Summary

This project builds upon Capstone Two, with the primary goal of assessing the dataset using time series analysis. While the data source, total observations, and variables remain the same, there are notable updates: the inclusion of the “Date” variable, which was omitted in the previous study, and the introduction of a new variable, “Sales,” derived from “Qty” and “Amount.” It is important to note that the results from the previous study will not be the focus of this analysis; they are only used to compare improvements against the current time series model. The last transformed dataset from the previous study was utilized to incorporate the “Date” and “Sales” variables. We began with exploratory data analysis (EDA) on both the original and transformed datasets. Following this, we performed data transformation and visualization to identify the optimal parameters (p, d, and q) for the ARIMA model, using AIC as the error criterion. These parameters were determined using the Auto-ARIMA model. The dataset, sourced from Kaggle and titled 'Amazon Sales Report,' pertains to online sales from India. The primary objective of this study is to perform time series analysis using ARIMA with the “Sales” variable, applying the best parameters [5,1,0] provided by the Auto-ARIMA model.

Introduction

The dataset used in this study covers online sales from April to July. Although it is labeled as the Amazon online selling report for 2022, the limited timeframe of the data appears disproportionate relative to India’s large population. This discrepancy raises questions about potential factors such as limited internet availability, competition from other online retailers, or the nascent stage of e-commerce in India. Despite these challenges, internet penetration in India has been increasing since 2022, with usage exceeding 60%. Given India's population of over 1.4 billion, the market shows significant promise for online sales. Although India’s internet penetration is lower compared to developed nations, its growing online user base, projected to expand further by 2024, highlights its potential as a lucrative market. Therefore, this study aims to analyze the "Sales" data using an ARIMA model to forecast sales trends. The objective is to evaluate sales patterns with the optimal parameters [5,1,0] to inform decision-making for product development and market strategy.

Data Collection, Preparation , and Visualization

The dataset used in this study covers only online sales from April to July. Although the dataset is labeled as the Amazon online selling report for 2022, the limited time frame raises questions about its representativeness relative to India’s large population. This limitation could be due to various factors, such as internet availability, competition from other online retailers, or the nascent stage of e-commerce in India. However, with internet penetration rising since 2022 and usage exceeding 60%, India holds significant promise for online sales, bolstered by its population of over 1.4 billion. Despite lower internet penetration compared to developed nations, India’s growing online user base suggests substantial market potential. Thus, this study aims to

analyze the sales data using an ARIMA model, forecasting the “Sales” variable. The objective is to examine sales trends with the optimal parameters [5,1,0] to inform decision-making for product development and market strategy.

- Data Collection

The data utilized in this study were sourced from Kaggle and are described as follows: 3 categorical variables listed in figure1, 4 numerical variables, and 3 float variables.

Figure 1 Table of data types

```
data12.dtypes
```

```
index                 int64
Status               object
FulfilmentB           int64
Category             object
Size                 object
Qty                   int64
Amount               float64
ship-postal-code     float64
Sales                 int64
dtype: object
```

- Data Preparation

This study focuses exclusively on data manipulation techniques relevant to time series analysis. The data manipulation methods from the previous capstone are not covered here, as they are detailed in that earlier work. To clarify, this time series analysis uses the same dataset as the previous capstone, with the addition of the "Date" variable, converted to the "datetime" format, and the newly created "Sales" variable, derived from "Qty" and "Amount." In other words, the dataset for this analysis is the final version from the previous study, enhanced with the "Date" and "Sales" variables. It is important to note that, unlike previous models, dummy variables have

been applied to the three categorical variables to make the dataset more suitable for regression analysis, although regression is not the focus of this study. Following data transformation and parameter selection using Auto-ARIMA, the best ARIMA model identified had the parameters [5,1,0] for p, d, and q, respectively.

- Data Visualization

Figure 2 Dataset for the TS analysis

```
data12.head()
```

	index	Date	Status	FulfilmentB	Category	Size	Qty	Amount	ship-postal-code	Sales
89698	1	4/30/2022	Shipped - Delivered to Buyer	1	kurta	3XL	1	406.0	560085.0	406
89699	26	4/30/2022	Shipped - Delivered to Buyer	1	kurta	3XL	1	299.0	495001.0	299
89701	85	4/30/2022	Shipped - Delivered to Buyer	1	Bottom	3XL	1	377.0	700078.0	377
89702	92	4/30/2022	Shipped - Delivered to Buyer	1	kurta	3XL	1	725.0	600028.0	725
89703	129	4/30/2022	Shipped - Delivered to Buyer	1	kurta	3XL	1	579.0	560055.0	579

```
data12.head()
```

Date	index	Status	FulfilmentB	Category	Size	Qty	Amount	ship-postal-code	Sales
2022-04-30	1	Shipped - Delivered to Buyer	1	kurta	3XL	1	406.0	560085.0	406
2022-04-30	26	Shipped - Delivered to Buyer	1	kurta	3XL	1	299.0	495001.0	299
2022-04-30	85	Shipped - Delivered to Buyer	1	Bottom	3XL	1	377.0	700078.0	377
2022-04-30	92	Shipped - Delivered to Buyer	1	kurta	3XL	1	725.0	600028.0	725
2022-04-30	129	Shipped - Delivered to Buyer	1	kurta	3XL	1	579.0	560055.0	579

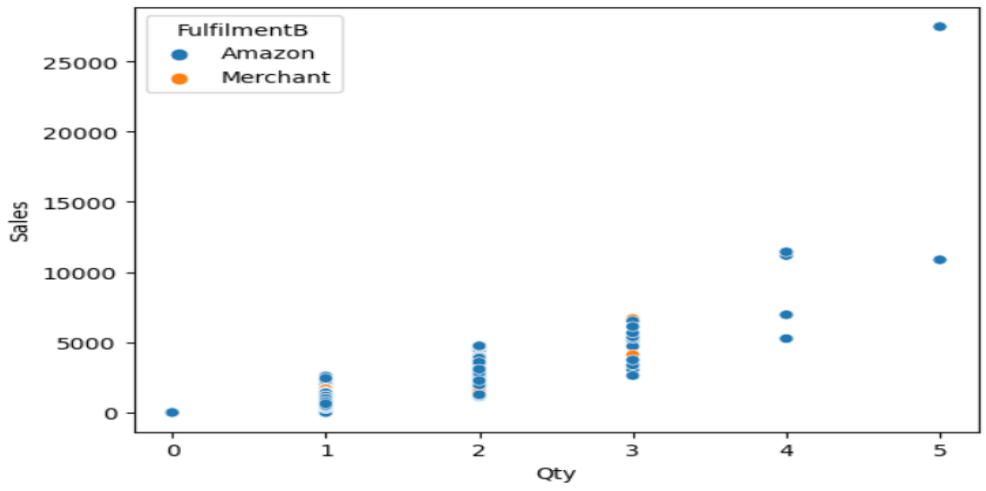
Figure 3 Categorical columns and their associated levels

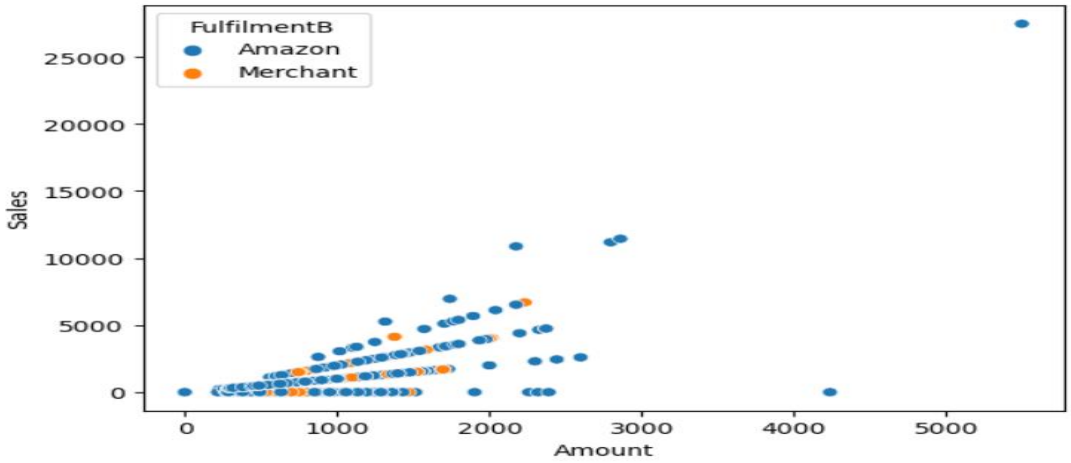
:

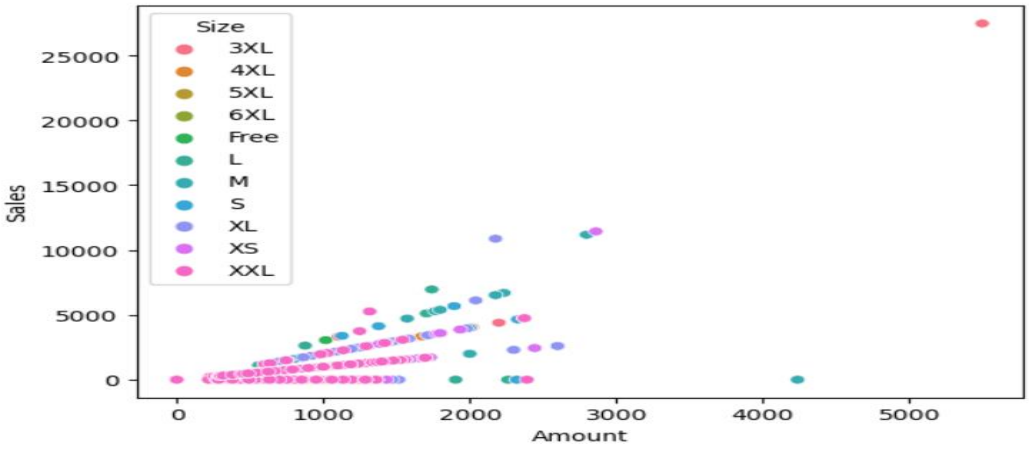
	VarName	LevelsCount
0	Date	91
1	Status	11
2	FulfilmentB	2
3	Category	8
4	Size	11

Figure 4 Descriptive Statistics for the Numerical variables

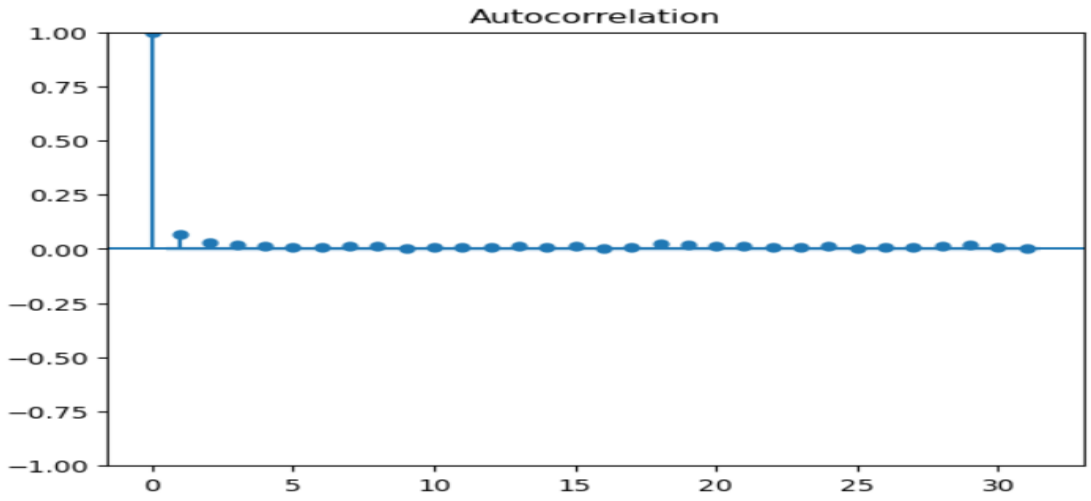
	index	Qty	Amount	ship-postal-code	Sales
count	37528.000000	37528.000000	37528.000000	37528.000000	37528.000000
mean	60932.856401	0.867406	646.523191	463355.001359	567.938446
std	36853.971158	0.354206	279.995058	194518.859123	409.562168
min	0.000000	0.000000	0.000000	110001.000000	0.000000
25%	27192.750000	1.000000	458.000000	380001.000000	368.000000
50%	63448.500000	1.000000	629.000000	500019.000000	568.000000
75%	91786.250000	1.000000	771.000000	600042.000000	759.000000
max	128891.000000	5.000000	5495.000000	989898.000000	27475.000000




FulfilmentB
Amazon
Merchant
Sales
Amount
0
5000
10000
15000
20000
25000
0
1000
2000
3000
4000
5000

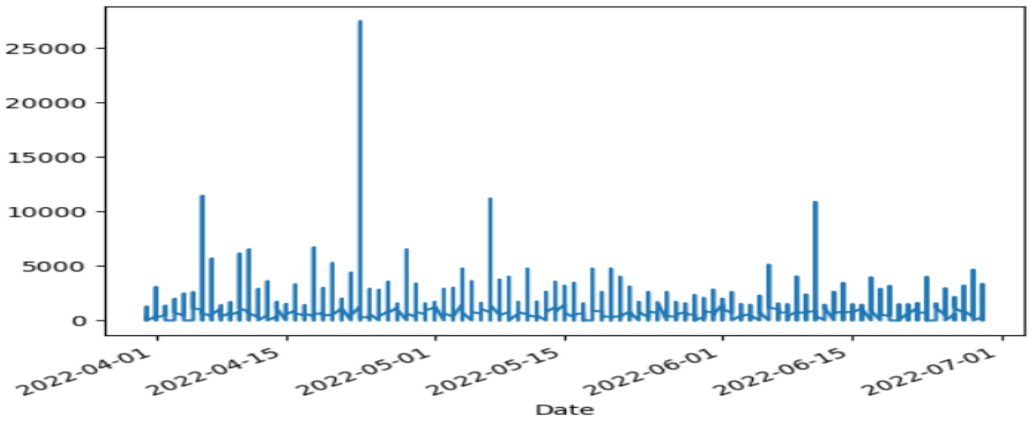

Size
3XL
4XL
5XL
6XL
Free
L
M
S
XL
XS
XXL
Sales
Amount
0
5000
10000
15000
20000
25000
0
1000
2000
3000
4000
5000

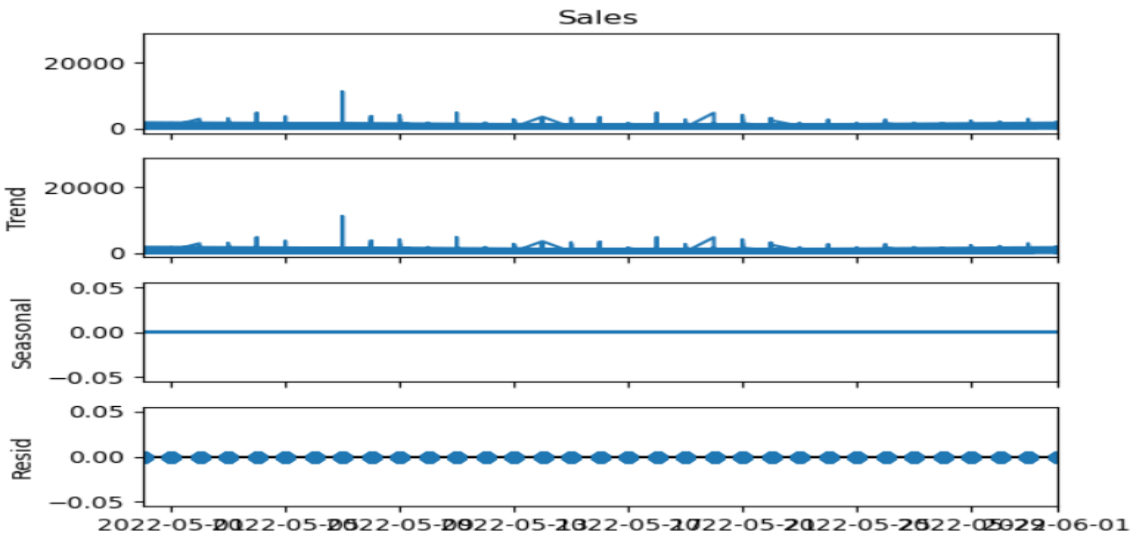
```
from statsmodels.graphics.tsaplots import plot_acf
plot_acf(data12.Sales, lags=31)
pyplot.show()
```




```
data12['Sales'].plot()
```

```
<Axes: xlabel='Date'>
```





Methodology

To conduct the analysis, we began with data visualization to select the values for p, d, and q, using lag plots. These plots revealed that the correlation could be either positive or negative, leading us to perform an autocorrelation test. This test confirmed that the overall data trend is negatively correlated. Since our goal was to analyze the Sales trend over time, we proceeded with the ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) tests for the "Sales" variable. The ACF plot confirmed a negative correlation trend for "Sales," indicating that the data is not suitable for direct autoregression without transformation. The PACF showed considerable changes in autocorrelation, with a notable dominance of positive data points in the first differencing. To determine the best order for the ARIMA model, we used the ADF (Augmented Dickey-Fuller) test to evaluate the p-values. However, the p-values did not reveal a significant difference between the differentiations, providing insufficient evidence to determine the optimal values for p, d, and q. As a result, we employed Auto-ARIMA to find the optimal parameters using the Akaike Information Criterion (AIC) as the error metric. Based on the Auto-

ARIMA results, our final ARIMA model was defined as ARIMA(p=5, d=1, q=0), as shown in the screenshot below:

```
Performing stepwise search to minimize aic
 ARIMA(1,1,1)(0,0,0)[0] intercept   : AIC=inf, Time=24.70 sec
 ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=581335.814, Time=0.83 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=571579.989, Time=2.41 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=inf, Time=20.10 sec
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=581333.814, Time=0.61 sec
 ARIMA(2,1,0)(0,0,0)[0] intercept   : AIC=567541.236, Time=3.27 sec
 ARIMA(3,1,0)(0,0,0)[0] intercept   : AIC=565356.248, Time=4.14 sec
 ARIMA(4,1,0)(0,0,0)[0] intercept   : AIC=564001.071, Time=5.34 sec
 ARIMA(5,1,0)(0,0,0)[0] intercept   : AIC=563055.472, Time=6.61 sec
 ARIMA(5,1,1)(0,0,0)[0] intercept   : AIC=inf, Time=80.93 sec
 ARIMA(4,1,1)(0,0,0)[0] intercept   : AIC=inf, Time=92.22 sec
 ARIMA(5,1,0)(0,0,0)[0]             : AIC=563053.472, Time=2.25 sec
 ARIMA(4,1,0)(0,0,0)[0]             : AIC=563999.071, Time=2.24 sec
 ARIMA(5,1,1)(0,0,0)[0]             : AIC=inf, Time=43.20 sec
 ARIMA(4,1,1)(0,0,0)[0]             : AIC=inf, Time=22.69 sec

Best model:  ARIMA(5,1,0)(0,0,0)[0]
Total fit time: 311.649 seconds
```

Additionally, to perform the analysis, we first applied the decomposition method using an additive model to visualize the trend, seasonality, and residuals of the data. We then transformed the data to achieve stationarity. For model evaluation, we split the dataset into 70% training data and 30% test data. Then forecasting for April 23rd, and the entire month of June.

Modeling

The modeling involved applying the ARIMA model to the dependent variable "Sales" using the parameters [5,1,0], as determined by the Auto-ARIMA test. Mathematically, the model is described as follows:

: arima_model=sm.tsa. ARIMA(data12['Sales'], order = (5,1,0))

model55 = arima_model44.fit()

Evaluation and Conclusion

By evaluating the model using a 70% training set and a 30% test set, the results were consistent with those provided by the ARIMA model, as both followed the same trend. For example, the autoregressive terms in both models had p-values below the significance level of 0.01. This indicates that the autoregressive coefficients are statistically significant in both models, justifying their inclusion.

In all, the time series analysis provided valuable insights into online selling in India, revealing that sales are concentrated during specific periods. While the previous four models (from Capstone 2) did not conclusively favor one over another, the time series analysis highlighted limitations affecting forecasting performance.

All lags in the model (L1, L2, L3, L4, L5) were statistically significant, with p-values below the 0.01 threshold. The Ljung-Box test statistic at L1 showed a Prob(Q) of 19.37 with a p-value of 0.00, allowing us to reject the null hypothesis that errors are white noise, as the p-value is less than 0.01.

Further analysis revealed heteroscedasticity, indicated by a test statistic of 0.79 and a p-value of 0.00, which is below the 0.01 threshold. The normality test using the Jarque-Bera statistic showed a value of 270,768,797.68 with a p-value of 0.00, leading us to reject the null hypothesis of normality. The data exhibited a slight positive skew and large kurtosis. For fitting the model, Auto-ARIMA was used to determine the optimal parameters, with the Akaike Information Criterion (AIC) guiding the selection. The best ARIMA model identified was ARIMA(5,1,0)(0,0,0)[0].

Our ARIMA model successfully forecasted the peak sales month and the sales figures for June, despite data limitations. Given the sparse data covering only April to July, it appears that e-commerce in India may still be in its early stages or not fully integrated into the culture. The observed variability in sales across different months suggests that understanding local cultural dynamics is crucial for investment. Increasing inventory during this period may be essential for capitalizing on potential sales opportunities and enhancing business profitability.