e-COMMERCE SALES MODELING & FORECASTING

Interim Project report submitted in fulfillment of the requirement for

POST-GRADUATE DIPLOMA IN STATISTICAL METHODS AND ANALYTICS

Submitted by

SANDIP MAJI : DSTC-022
MOUSUMI MAITY : DSTC-015
SANATAN PAUL : DSTC-021

(june, 2022)

INDIAN STATISTICAL INSTITUTE

Indian Statistical Institute (ISI), Chennai Centre
110, Nelson Manickam Road
Aminjikarai
Chennai, 600 029.Email : *head[at]isichennai.res.in*
Phone : *+91 44 23740612 , +91 44 23740256*

Date: 20/06/2022

CERTIFICATE

This is to certify that Mr. Sandip Maji, Mr. Sanatan Paul and Ms. Mousumi Maity has done the project under my supervision and guidance (from February 20th to June 20th). This is an original project report based on work carried out by them in fulfillment of the requirement for the Post-Graduate Diploma in Statistical Methods and Analytics programme of the Indian Statistical Institute, Chennai Centre

[Signature]

Dr. G. Ravindran

ACKNOWLEDGEMENT

First we would like to thank Indian Statistical Institute, Chennai center for providing the Infrastructure and opportunity. We would also like to express our sincere gratitude, gratefulness and indebtedness to reverend teacher and supervisor Dr. G. Ravindran, Indian Statistical Institute (Chennai Centre), for his constant encouragement, without which it would not have been possible for us to complete this project work.

[Signature(s)]

Sandip Maji (DSTC-022)

Sanatan Paul (DSTC-021)

Mousumi Maity (DSTC-015)

CONTENTS

INTRODUCTION

It is projected that eCommerce will account for more than $6.5 trillion in sales by 2023, which is 22% of retail sales, globally. It's not just retail eCommerce sales that are soaring, as the global B2B Ecommerce market reached $5.7 trillion in 2019.In INDIA also e-commerce market is expected to reach US$ 111 billion by 2024 and US$ 200 billion by 2026.

In order to enhance the logistics service experience of customers in the e-commerce industry chain, supply chain collaboration requires that commodities are stocked in advance in local warehouses of various markets around the world, which can effectively reduce logistics time. Sales forecasting is even more vital for supply chain management in e-commerce with a huge amount of transaction data generated every minute.
Therefore, algorithms and technologies of big data analysis are widely applied to predict sales of e-commerce commodities, which provide the data basis for the supply chain management.

Here we are using different methods to find the behaviour of variables involved and their interdependencies .At first we are using some classical techniques and then some advanced machine learning techniques to find some important factors which affects the e-commerce activity and the profit as a whole ,then we are using time series techniques and some machine learning techniques to model and forecast sales .And come to a conclusion with highest accuracy .This will help business entities to enhance logistics service experience of customers and effective warehouse planning.

DATA DESCRIPTION

Dataset and its descriptions:

- **Source:**

Kaggle allows users to find and publish data sets, explore and build models in a web based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges. We took our dataset from kaggle data dictionary.

Dataset link: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

- **About Data:**

This is a Brazilian ecommerce public dataset of orders made at Olist Store .The dataset has information of 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. Its features allows viewing an order from multiple dimensions: from order status, price, payment and freight performance to customer location, product attributes and finally reviews written by customers. We also released a geolocation dataset that relates Brazilian zip codes to latitude/longitude coordinates.

There are nine separate data set. The data sets are Customer dataset, Geolocation dataset, order item dataset, Order payment dataset, Order review dataset, Orders dataset, Product dataset, Sellers dataset, Product category name translation dataset.

- **Data:**

Customer dataset:

```
Shape of the dataframe is : (99441, 5)
Data columns (total 5 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   customer_id               99441 non-null  object
 1   customer_unique_id        99441 non-null  object
 2   customer_zip_code_prefix  99441 non-null  int64
 3   customer_city             99441 non-null  object
 4   customer_state            99441 non-null  object
```

Seller dataset:

```
Shape of the dataframe is : (3095, 4)
**********************************************************
Data columns (total 4 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   seller_id               3095 non-null   object
 1   seller_zip_code_prefix  3095 non-null   int64
 2   seller_city             3095 non-null   object
 3   seller_state            3095 non-null   object
```

Geolocation dataset:

```
Shape of the dataframe is : (1000163, 5)
************************************************************
Data columns (total 5 columns):
 #   Column                        Non-Null Count    Dtype
---  ------                        --------------    -----
 0   geolocation_zip_code_prefix   1000163 non-null  int64
 1   geolocation_lat               1000163 non-null  float64
 2   geolocation_lng               1000163 non-null  float64
 3   geolocation_city              1000163 non-null  object
 4   geolocation_state             1000163 non-null  object
```

Product dataset:

```
Shape of the dataframe is : (32951, 9)
************************************************************
Data columns (total 9 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   product_id                  32951 non-null  object
 1   product_category_name       32341 non-null  object
 2   product_name_lenght         32341 non-null  float64
 3   product_description_lenght  32341 non-null  float64
 4   product_photos_qty          32341 non-null  float64
 5   product_weight_g            32949 non-null  float64
 6   product_length_cm           32949 non-null  float64
 7   product_height_cm           32949 non-null  float64
 8   product_width_cm            32949 non-null  float64
```

category name translation dataset:

```
Shape of the dataframe is : (71, 2)
************************************************************
Data columns (total 2 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   product_category_name          71 non-null     object
 1   product_category_name_english  71 non-null     object
```

Orders dataset:

```
Shape of the dataframe is : (99441, 8)
************************************************************
Data columns (total 8 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   order_id                       99441 non-null  object
 1   customer_id                    99441 non-null  object
 2   order_status                   99441 non-null  object
 3   order_purchase_timestamp       99441 non-null  object
 4   order_approved_at              99281 non-null  object
```

```
 5   order_delivered_carrier_date   97658 non-null  object
 6   order_delivered_customer_date  96476 non-null  object
 7   order_estimated_delivery_date  99441 non-null  object
```

Order items dataset:

```
Shape of the dataframe is : (112650, 7)
************************************************************
Data columns (total 7 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   order_id             112650 non-null  object
 1   order_item_id        112650 non-null  int64
 2   product_id           112650 non-null  object
 3   seller_id            112650 non-null  object
 4   shipping_limit_date  112650 non-null  object
 5   price                112650 non-null  float64
 6   freight_value        112650 non-null  float64
```

Order payments dataset :

```
Shape of the dataframe is : (103886, 5)
************************************************************
Data columns (total 5 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   order_id              103886 non-null  object
 1   payment_sequential    103886 non-null  int64
 2   payment_type          103886 non-null  object
 3   payment_installments  103886 non-null  int64
 4   payment_value         103886 non-null  float64
```

Order reviews dataset :

```
Shape of the dataframe is : (99224, 7)
************************************************************
Data columns (total 7 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   review_id                99224 non-null  object
 1   order_id                 99224 non-null  object
 2   review_score             99224 non-null  int64
 3   review_comment_title     11568 non-null  object
 4   review_comment_message   40977 non-null  object
 5   review_creation_date     99224 non-null  object
 6   review_answer_timestamp  99224 non-null  object
```

METHODOLOGY

1. Objective:

- Understanding the behavior of data eventually getting a sense of underlying process.
- Assessing the impact of different product attributes like- price, product photos, product weight , distance between buyer and seller ,customer-city, states, distance between buyer and seller, and some after sales attributes like review score ,payment type , payment sequential etc. on sales And their interdependent relations.
- Overall sales forecasting.
- Product wise sales forecasting.

2. Missing value treatment:

The real-world data often has a lot of missing values. The cause of missing values can be data corruption or failure to record data. The handling of missing data is very important during the preprocessing of the dataset as many machine learning algorithms do not support missing values.

The missing techniques are

- Dropping records with at least one missing value
- Drop columns that are least significant and has majority of missing value

3. Exploratory data analysis :

Exploratory Data Analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. Following are the different steps involved in EDA :

- Data Collection
- Data Cleaning
- Data Preprocessing
- Data Visualization

4. Modeling The Data

- **SEASONAL DECOMPOSE OF DATA**

There are 2 types of seasonal decomposition which we are using here are additive decomposition and multiplicative decomposition

- **AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA)**

ARIMA stands for autoregressive integrated moving average, which deals with time series data by combination of three methods auto regression, differencing and moving average. Advantage of ARIMA

is it can work with nonstationary data. Stationary data means which is constant throughout the period, where non-stationary data is seasonal i.e. data changes consistently through various time periods. ARIMA (p, d, q) can be represented as

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots\ldots + \varphi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-1} + \ldots\ldots + \theta_q \varepsilon_{t-q}$$

Where y_t is time series values at time t and φ and θ are constant also p and q are auto regressive and lagged error terms respectively.

- **REGRESSION TREE**

A regression tree refers to an algorithm where the target variable is and the algorithm is used to predict its value. In a regression tree, a regression model is fit to the target variable using each of the independent variables. After this, the data is split at several points for each independent variable. At each such point, the error between the predicted values and actual values is squared to get "A Sum of Squared Errors" (SSE). The SSE is compared across the variables and the variable or point which has the lowest SSE is chosen as the split point. This process is continued recursively.

As an example of a regression type problem, you may want to predict the selling prices of a residential house, which is a continuous dependent variable. This will depend on both continuous factors like square footage as well as categorical factors like the style of home, area in which the property is located, and so on.

- **RANDOM FOREST**

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result

- **XGBOOST**

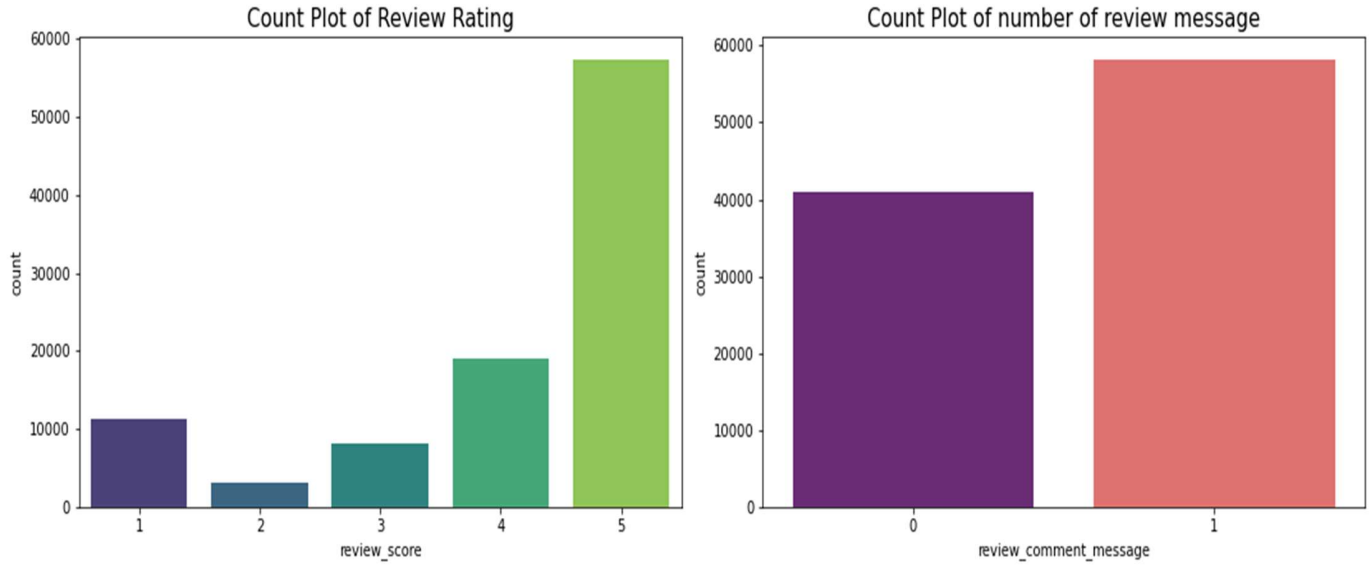
XGBoost is an abbreviation "Extreme Gradient Boosting" proposed by Friedman. In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results.

MISSING VALUE TREATMENT

- All duplicate rows are first deleted
- Missing values present in product_data are replaced by mean
- While data preparation data from order_data is extracted so null values have no impact so left as it is.

EXPLORATORY DATA ANALYSIS

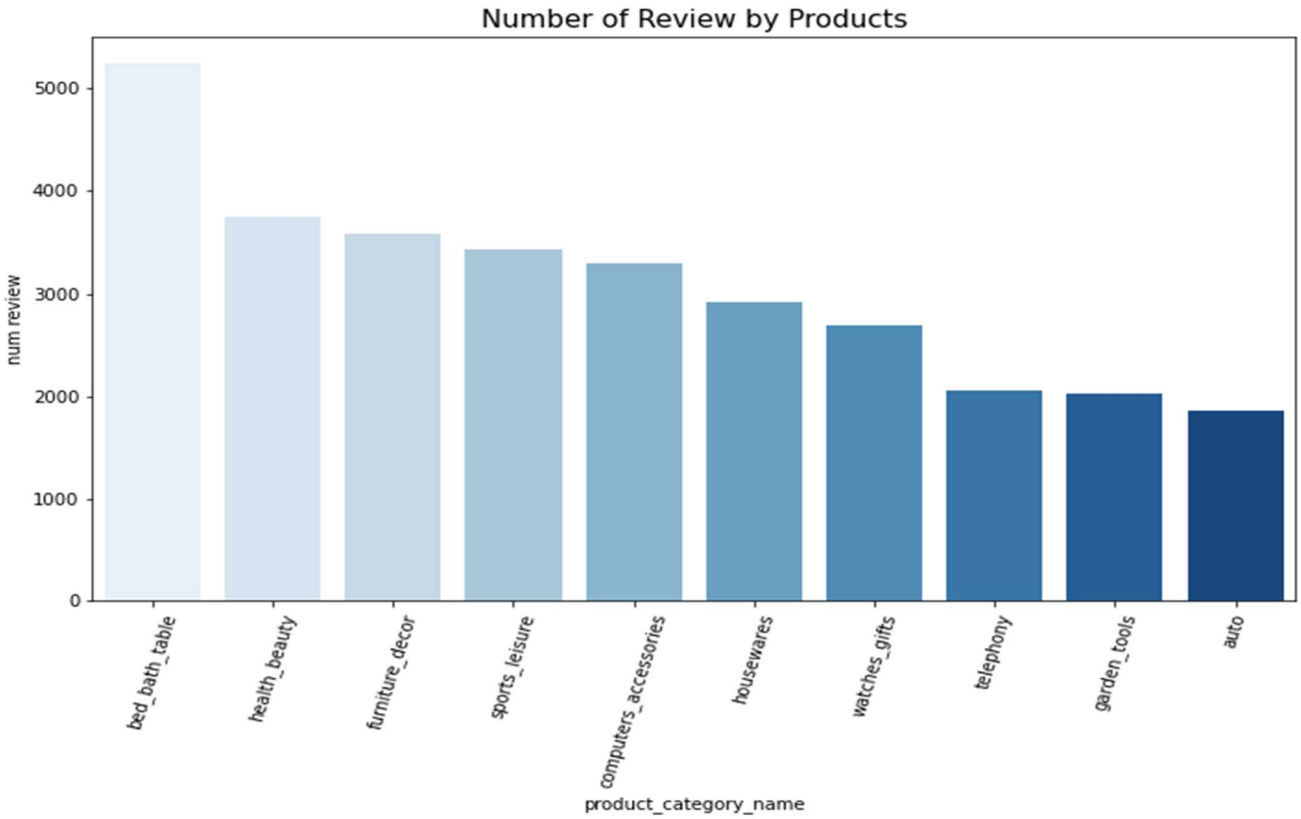
ORDER REVIEW DATA ANALYSIS :

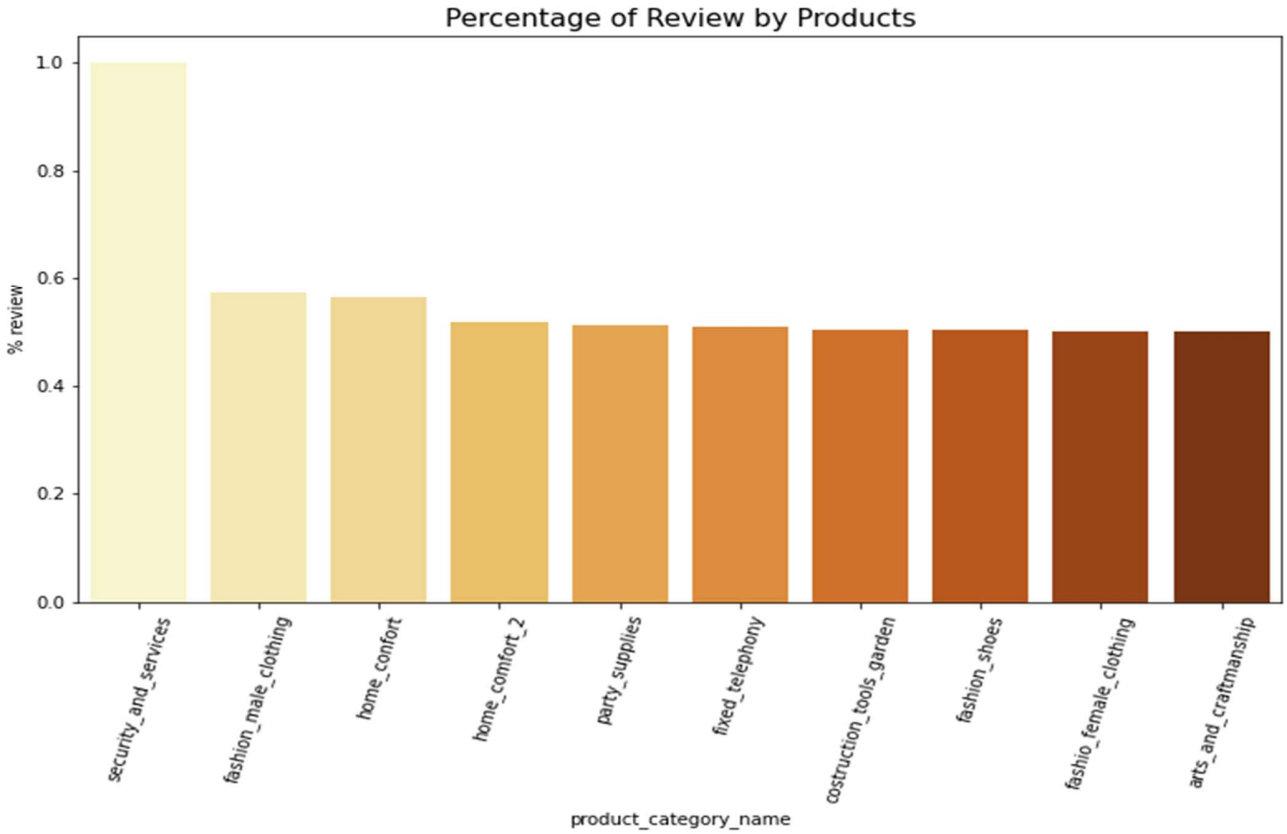


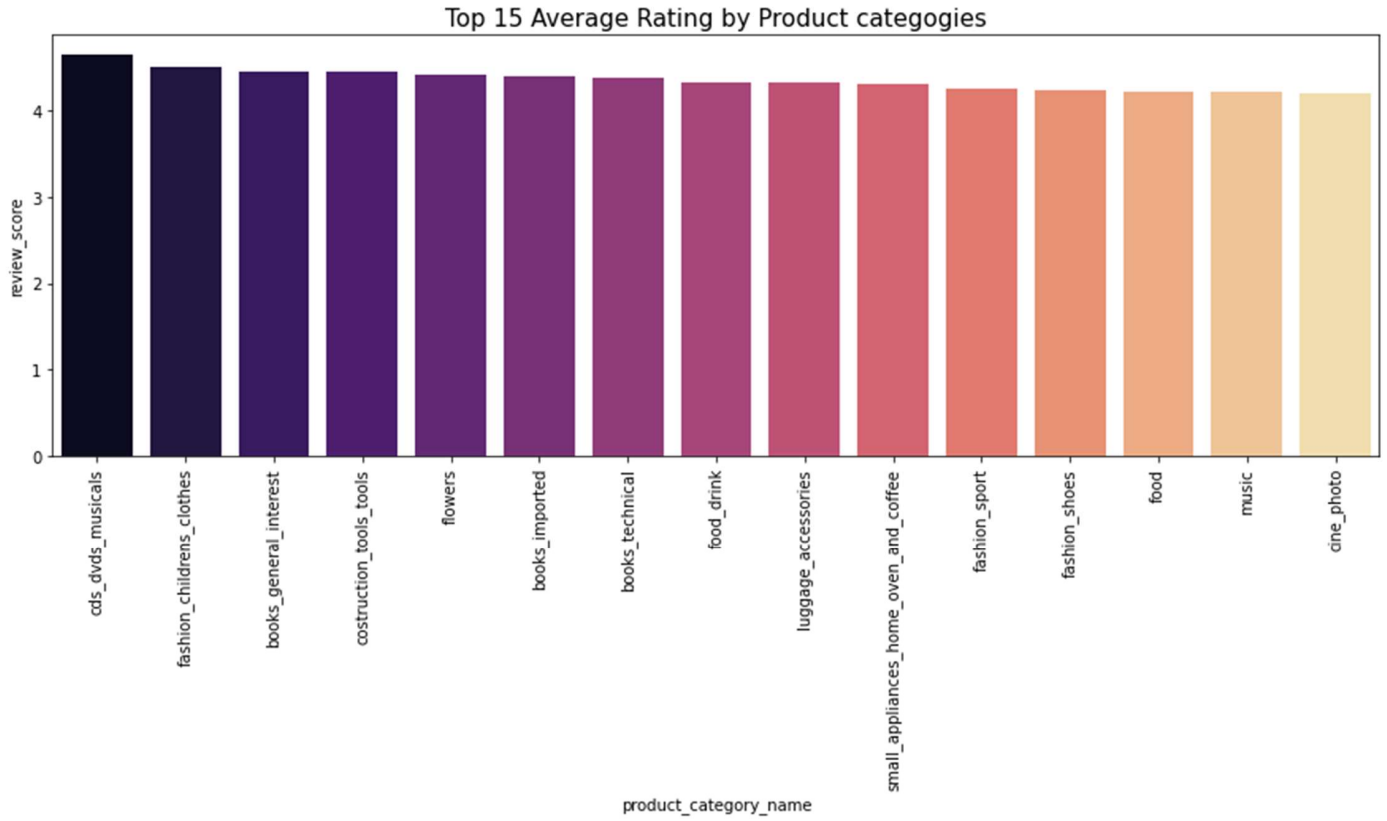
Interpretation:

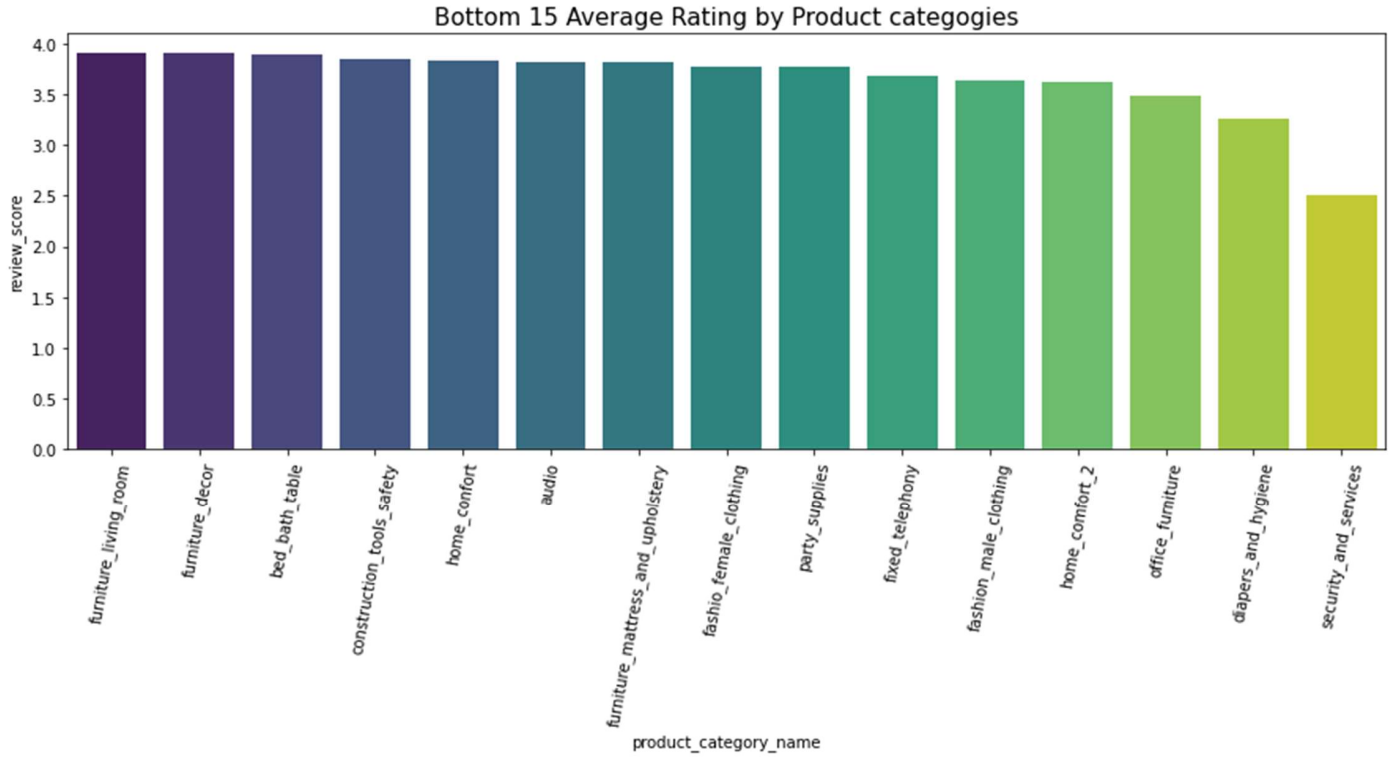
- In general, all the orders have high review ratings, with over 77% of high ratings (4,5)
- Over half of customer don't leave comments (around 58.7%)

- **Average Rating and Number of Review by Product categories**





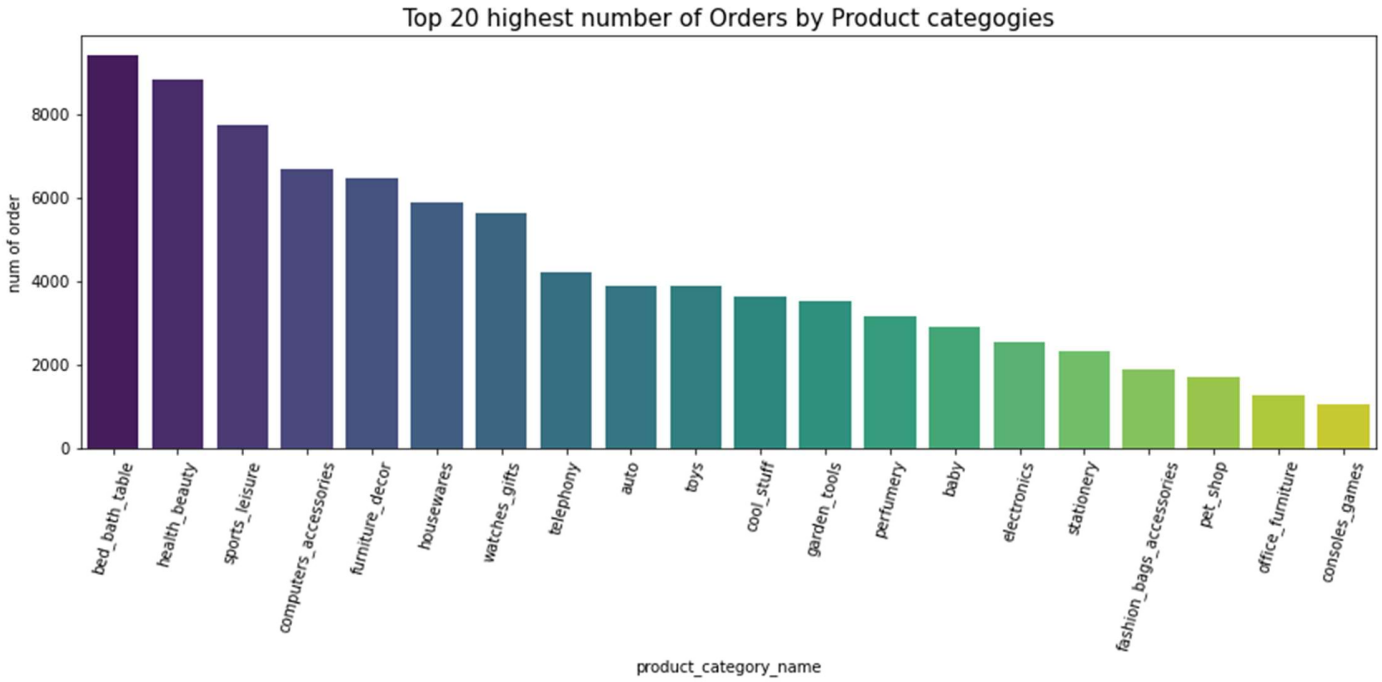


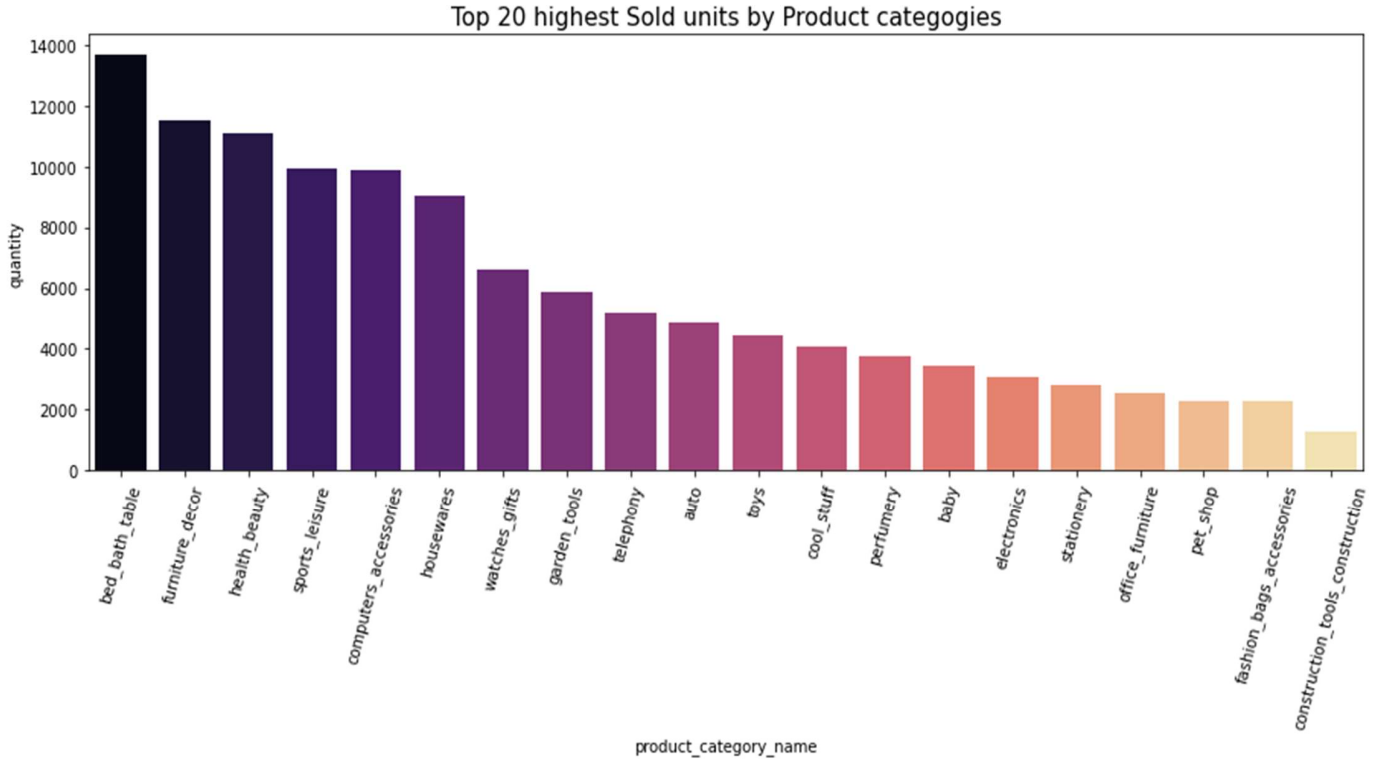


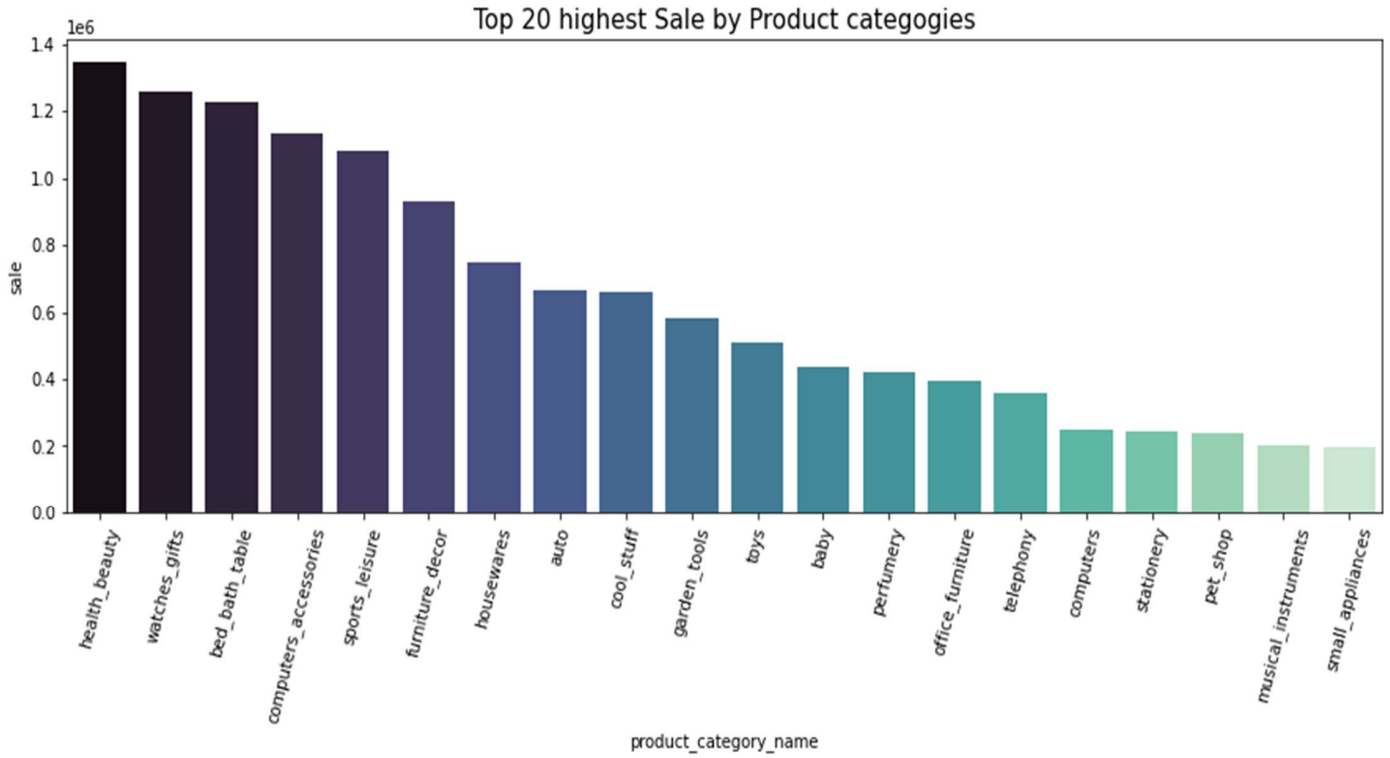
Interpretation:

- Overall, all product Categories have high and consistent average rating scores, with low rating is just roughly under 4/5
- The list of top 15 and bottom 15 product categories are shown above

- **Number of orders, Quantity of sold units and Sale by Product categories**







Interpretation:

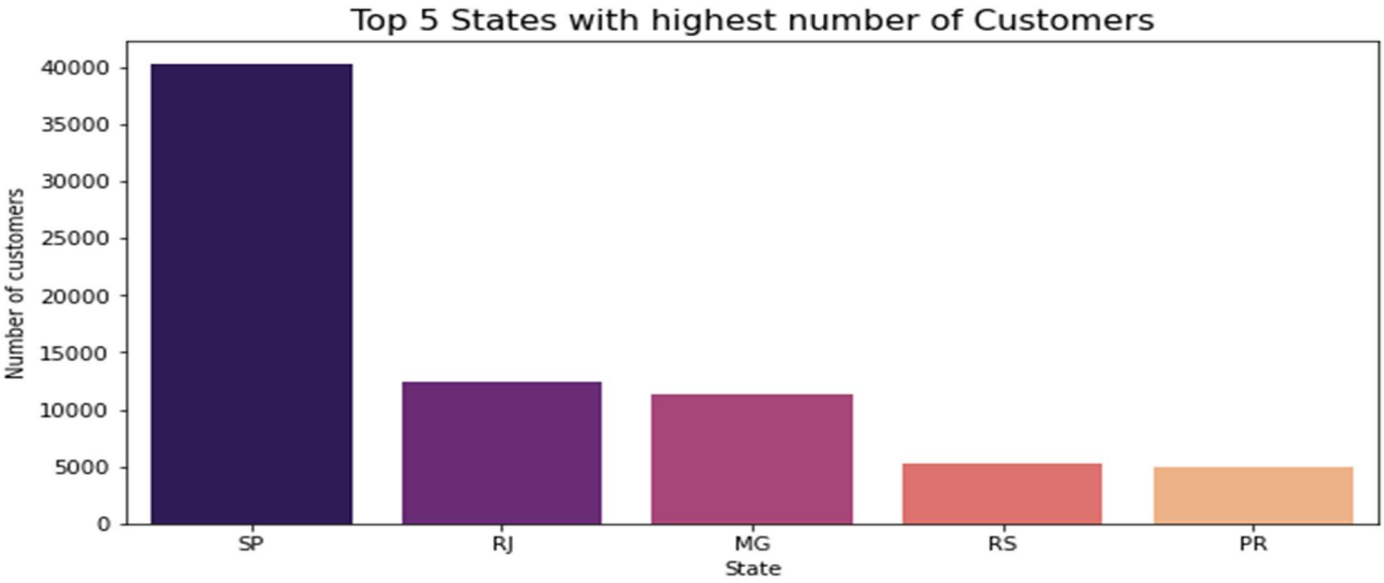
- Visually, the num order, sale and sold quantity amongst product categories range a lot
- Top 20 highest product categories in term of num orders, sold units and sale are plotted above

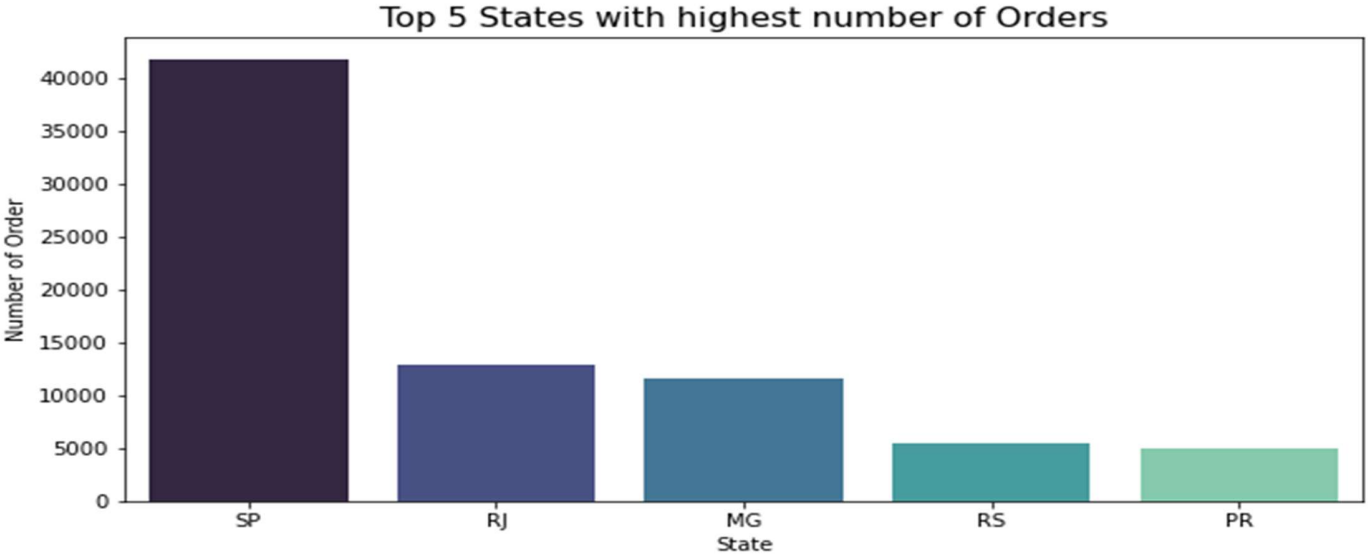
The Target Products should be

- 'bed_bath_table',
- 'garden_tools',
- 'watches_gifts',
- 'pet_shop',
- 'stationery',
- 'small_appliances',
- 'consoles_games',
- 'cool_stuff',
- 'auto',
- 'electronics',
- 'musical_instruments',
- 'home_construction',
- 'perfumery',
- 'furniture_living_room',
- 'baby',

- 'housewares',
- 'luggage_accessories',
- 'sports_leisure',
- `'office_furniture',`
- `'fashion_bags_accessories'`

- **Which States have highest number of Customers, Orders?**

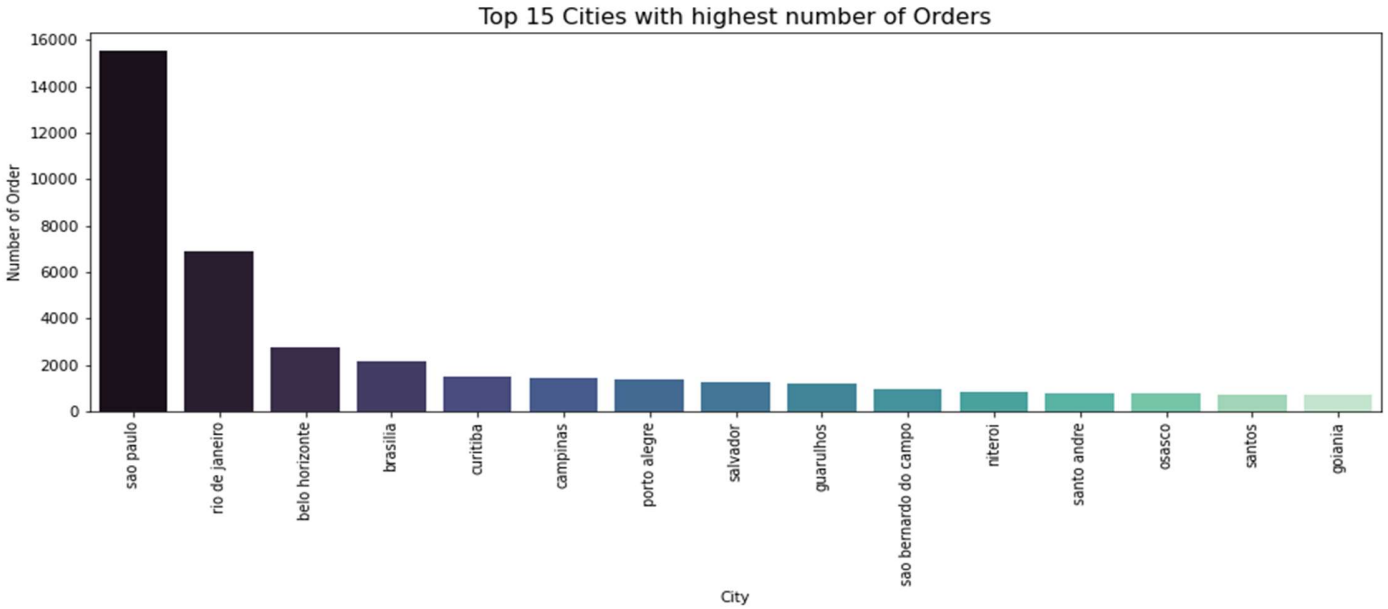




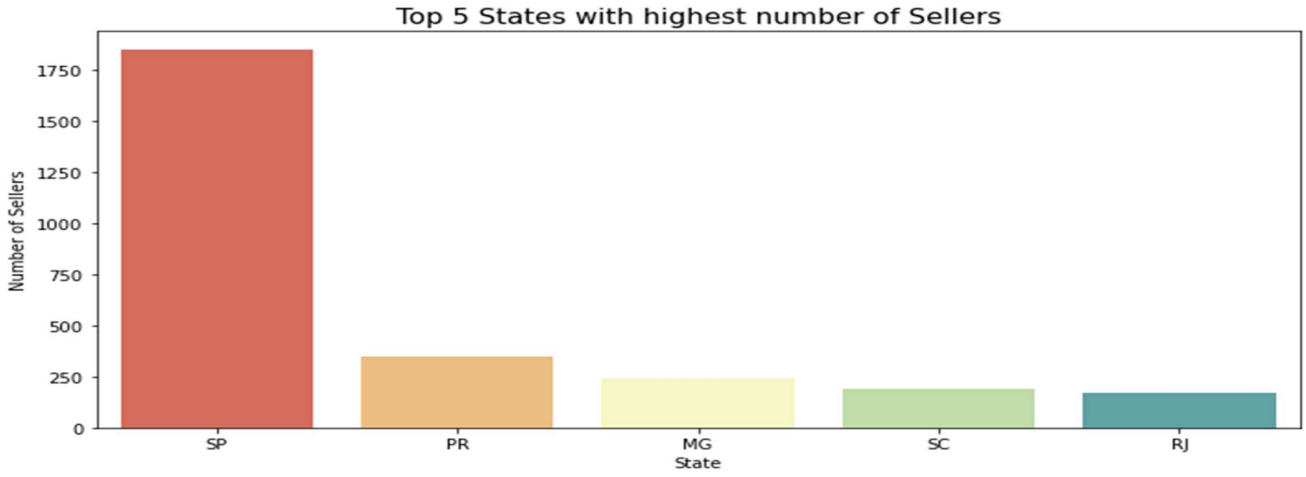
Interpretation:

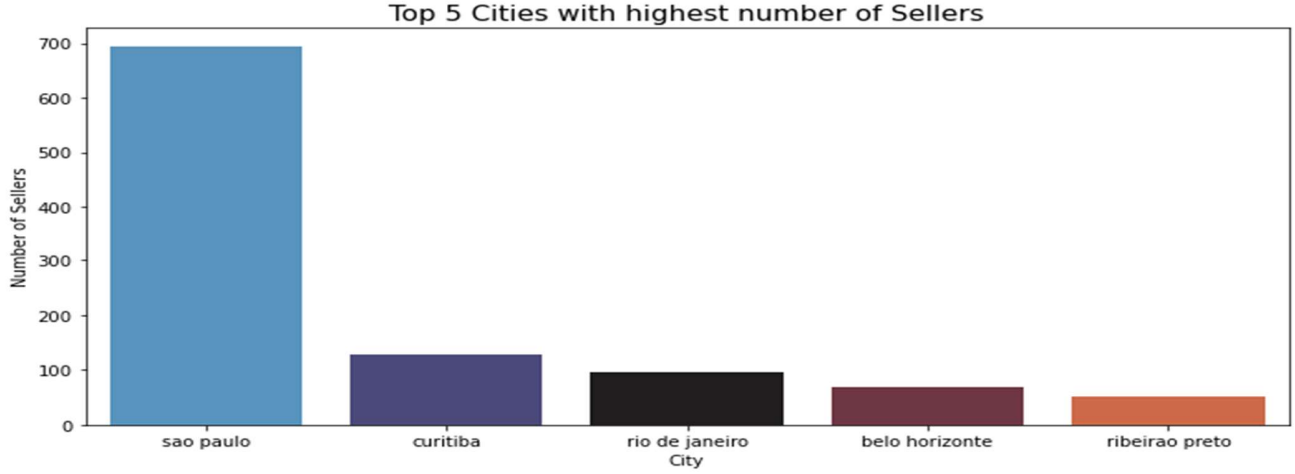
- Figures of top state in term of number of customers and orders are roughly similar
- Most of customers and orders are dominantly came from SP(Sao Paulo) State

- **Which Cities have highest number of Orders ?**



- **Which States, Cities have highest number of Sellers ?**

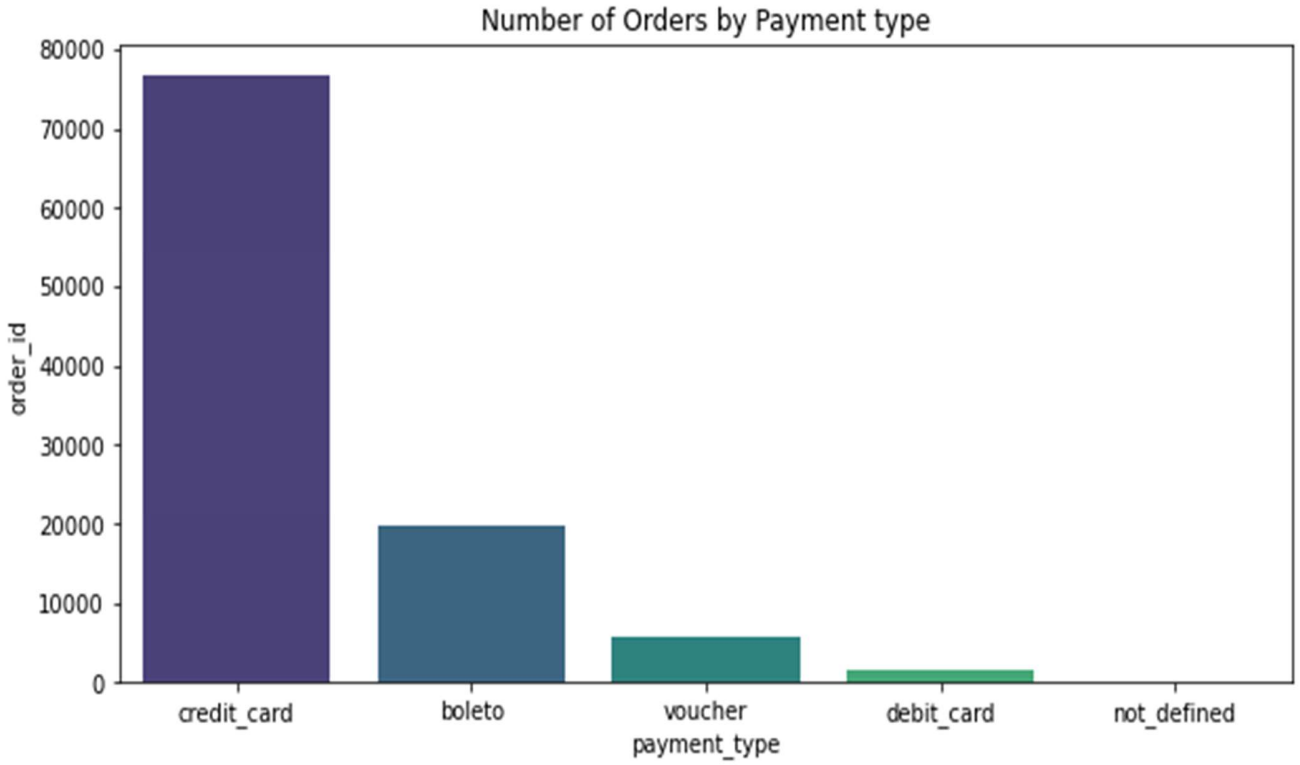




Interpretation:

- In general, top state by number of sellers coincide with top state by customers amounts (except RS for customers and SC for sellers)
- Similarly for top cities in term of customers and sellers (except Brasilia for customers and Ribeirao Preto for sellers)

- **About Payment types ?**



Interpretation:

- We see that most of the payment done by credit_card.
- Next most payment method is boleto.

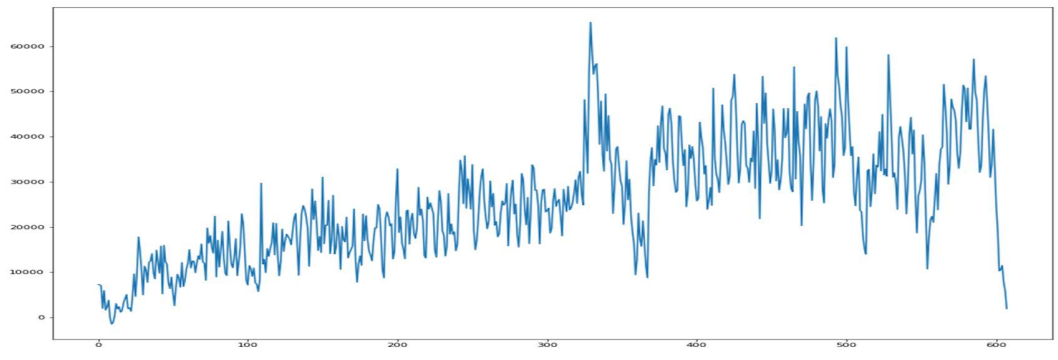
DATA PREPARATION BEFORE MODELING THE DATA:

- To get the product category name in English we have merged product_data and category_name_trn on product category name
- Then order_item_data and order_payment_data is merged on order_id
- New data is merged with order_data on order_id
- Next new data is merged with product_data on product_id
- Then merged with seller_data on seller_id
- Then mereged with customer_data on customer_id
- Then datetime is converted into datetime fromat from object format
- Missing values in product attributes having numerical variables are replaced by mean and for categorical variable a category called null is created
- A new table is created by extracting product attributes and payment attributes from earlier table for further analysis

DETAILS OF PREPARED DATA:

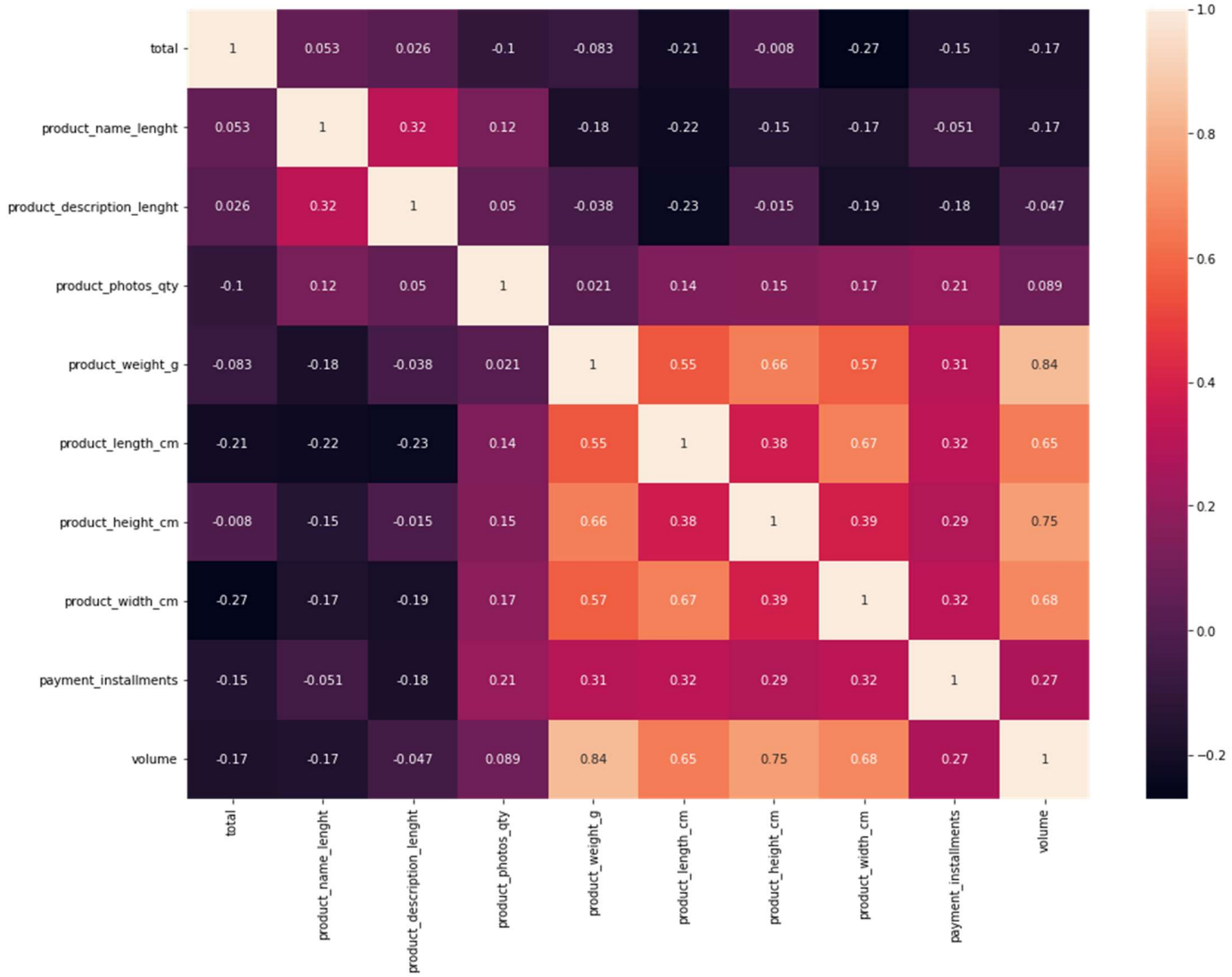
```
Shape of the dataframe is : (611, 11)
************************************************************
Data columns (total 11 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   order_purchase_timestamp     611 non-null    object
 1   total                        611 non-null    float64
 2   product_name_lenght          611 non-null    float64
 3   product_description_lenght   611 non-null    float64
 4   product_photos_qty           611 non-null    int32
 5   product_weight_g             611 non-null    float64
 6   product_length_cm            611 non-null    float64
 7   product_height_cm            611 non-null    float64
 8   product_width_cm             611 non-null    float64
 9   payment_installments         611 non-null    int32
 10  volume                       611 non-null    float64
```

This data represents daily overall sales.



ANALYSIS & RESULTS

CORRELATION STUDY:

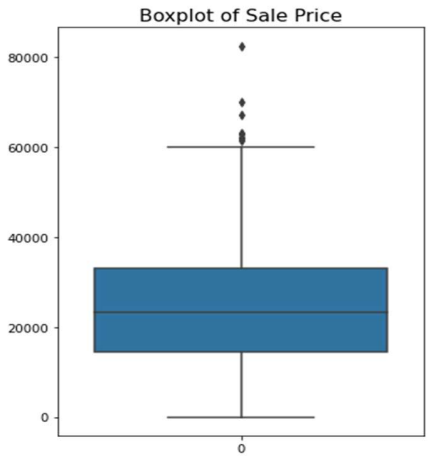


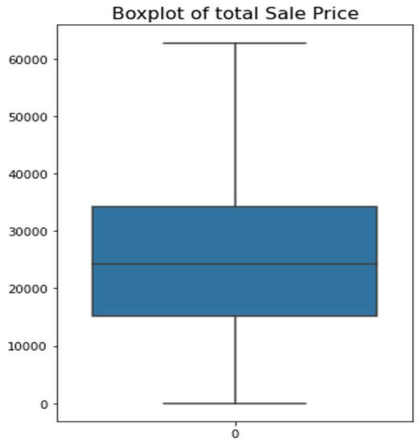
Interpretation:

- With total sale product weight has very low correlation so we are dropping that and column
- Volume has high correlation with product length and product width so we are dropping volume and considering the other two for our analysis.

Overall sale modeling :

To check whether any outlier present in the sale data .boxplot of the data is shown below





Interpretation:

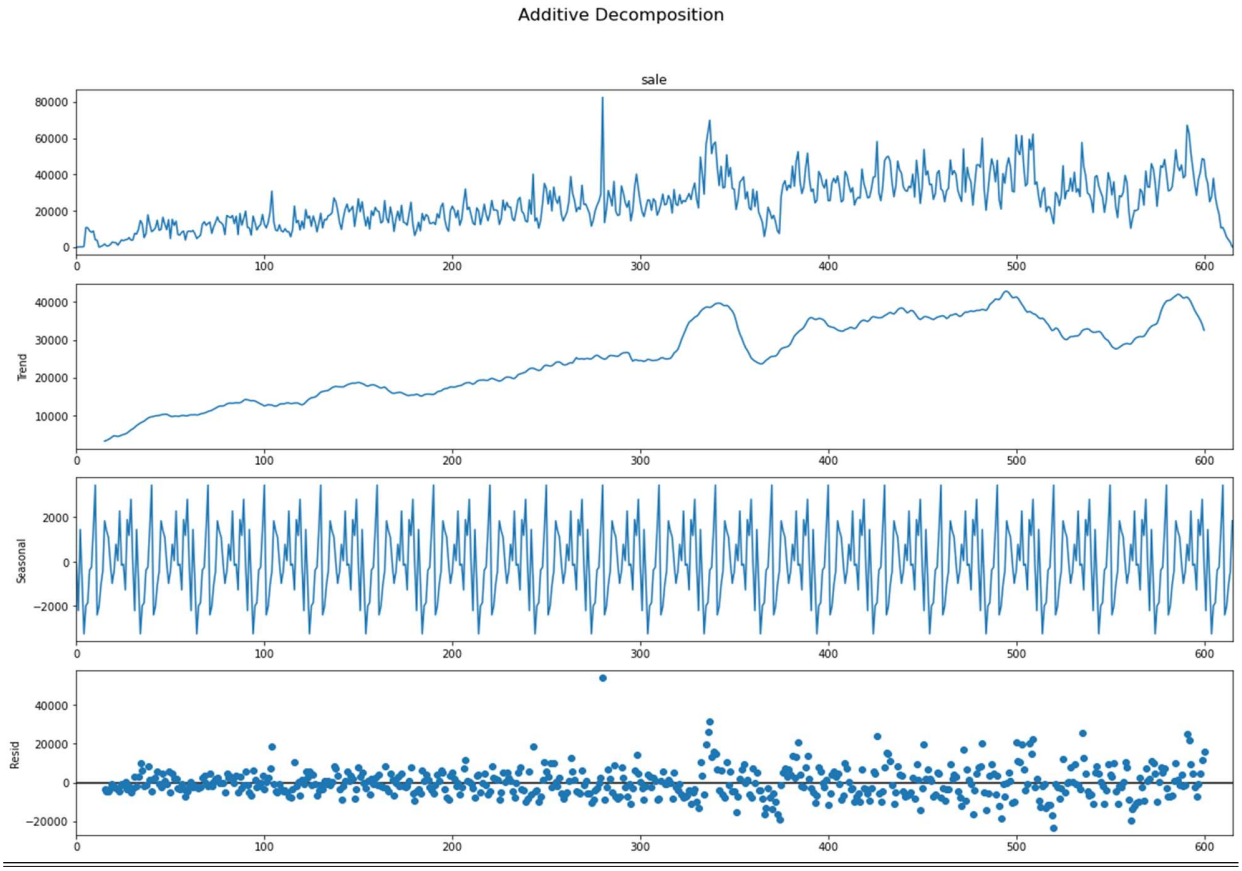
Plot clearly shows there are some outlier presents in the data. So, outlier treatment is done and we have got outlier free data.Here outlier is replaced by the mean of before and after data of outlier.

SEASONAL DECOMPOSITION :

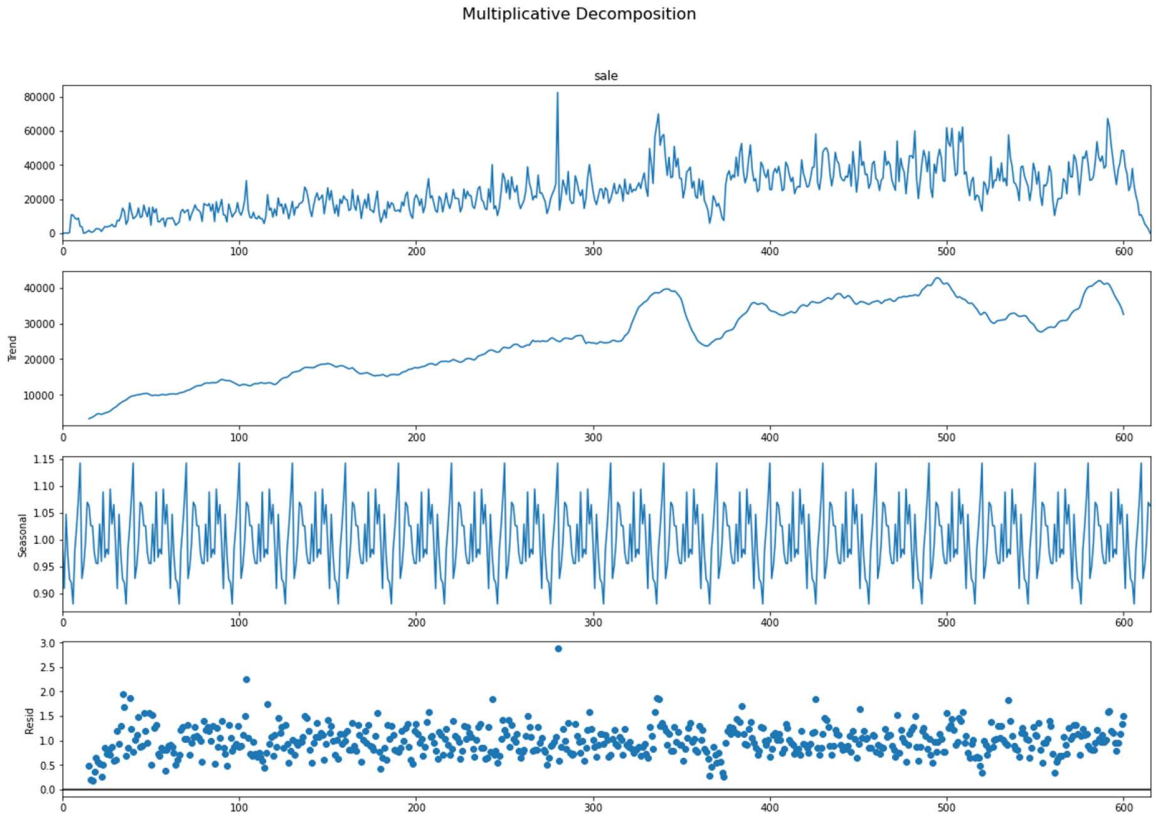
After performing seasonality test using 2 types of models – multiplicative and additive

We are plotting the result

- **Additive model**



- **Multiplicative model**



Interpretation:

Here in additive model residual increases over time so additive model is not suitable to use so multiplicative model is used to determine seasonality.

Deseasonalised series

Using multiplicative model seasonal component is subtracted from overall sale component and deseasonalised series is obtained.

Train Test Split:

To verify the model performance first we divided the data into 90% and 10% as training set and validation set. Depending on the performance on validation set we will decide the model. While creating random forest, regression tree, support vector regression, xgboost we are creating 3 columns containing previous predicted values which we are treating as variable to forecast using these models

Performance measure of different model:

We are using two methods to measure the performance of the model

1. RMSE: RMSE stands for root mean square error. And the equation of this is

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

2. MAPE: MAPE stands for mean absolute percentage error using mape we can compare among different models

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100$$

MODELING THE DATA

Model Building (ARIMA)

Stationarity

A stochastic process is said to be strictly stationary if the joint distribution of $x(t_1), \ldots, x(t_n)$ is same as the joint distribution of $x(t_1 + \tau), \ldots, x(t_n + \tau)$ for all $t_1, t_2, \ldots, t_n, \tau$

Shifting the time origin by an amount .. Has no effect on joint distributions.

We consider second order stationarity (less restricted way) if first order and second order moments are finite and does not depend on time t and autocorrelation depends only on lag τ

$$E[x_t] = \mu$$
$$\text{var}(x_t) = \sigma^2$$
$$y(\tau) = \text{cov}(x_t, x_{t+\tau})$$

Test for stationarity: ADF test

For testing whether the time series is stationary or not we can use Augmented Dickey-Fuller or KPSS test which will be discussed further

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t,$$

$$H_0: \gamma = 0 \qquad H_1: \gamma < 0$$

Null hypothesis : Series is Non-stationary

Alternative hypothesis : Series is Stationary

Results of Dickey-Fuller Test after performing first differencing on deseasonalised series:

- Test Statistic -2.909227
- p-value 0.044285
- #Lags Used 14.000000
- Number of Observations Used 601.000000
- Critical Value (1%) -3.441278
- Critical Value (5%) -2.866361
- Critical Value (10%) -2.569338

ARIMA

$$y_t = \phi_o + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$$

where y_t is following ARMA(p,q) process if y_t is stationary

If process y_t is non-stationary, then we take lagged difference

After ith order lagged difference, the series obtained is stationary, it is said to follow ARIMA(p,I,q) where 'I' is called order of integration

- Order of integration can be found by taking differences and performing test for stationarity
- '*p*' is called order of AR (Auto-Regressive) and '*q*' is called order of MA
- p, q can be found out ACF,PACF and selecting model with minimum AIC

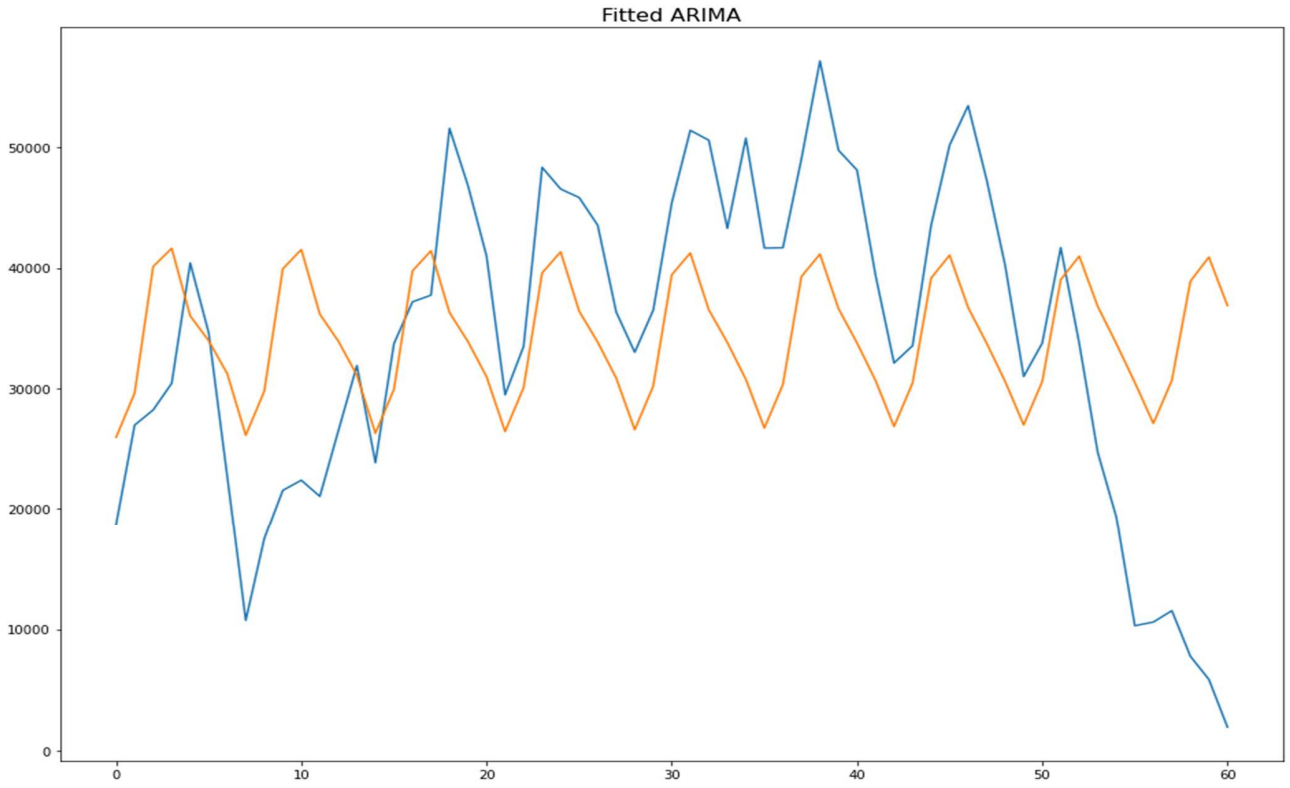
After fitting **ARIMA model** we have found that the model is ARIMA(4,1,5)

SARIMAX Results

Dep. Variable:	y	**No. Observations:**	547
Model:	SARIMAX(4, 1, 5)	**Log Likelihood**	-5477.122
Date:	Thu, 23 Jun 2022	**AIC**	10974.244
Time:	03:48:21	**BIC**	11017.270
Sample:	0	**HQIC**	10991.063

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.8049	0.017	47.860	0.000	0.772	0.838
ar.L2	-1.4308	0.023	-63.220	0.000	-1.475	-1.386
ar.L3	0.7826	0.022	35.564	0.000	0.739	0.826
ar.L4	-0.9792	0.015	-64.834	0.000	-1.009	-0.950
ma.L1	-1.3625	0.043	-31.668	0.000	-1.447	-1.278

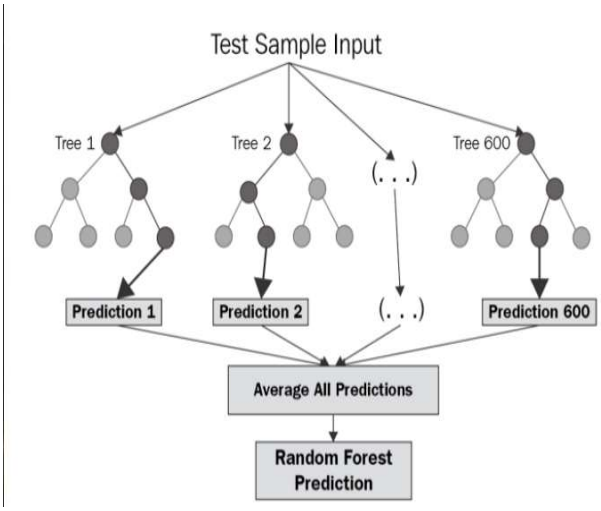
ma.L2	1.7888	0.060	29.908	0.000	1.672	1.906
ma.L3	-1.4907	0.078	-18.997	0.000	-1.645	-1.337
ma.L4	1.3306	0.060	22.275	0.000	1.214	1.448
ma.L5	-0.5246	0.042	-12.580	0.000	-0.606	-0.443
sigma2	3.635e+07	2.83e-10	1.28e+17	0.000	3.64e+07	3.64e+07



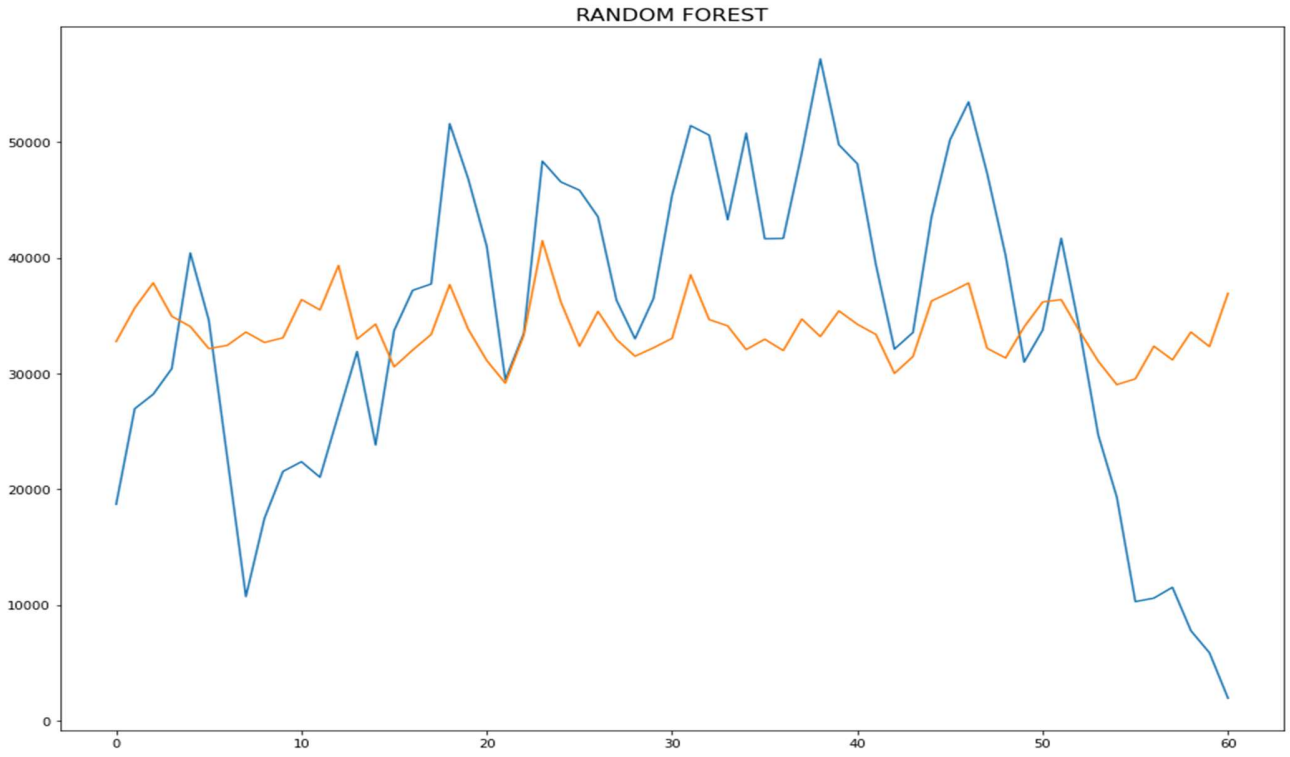
Model performance: rmse = 13024.2736235761
mape is = 80.00905854894125

Interpretation:
The model is not performing good on validation set

- RANDOM FOREST REGRESSION:
  Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.



Random Forest applying on the model



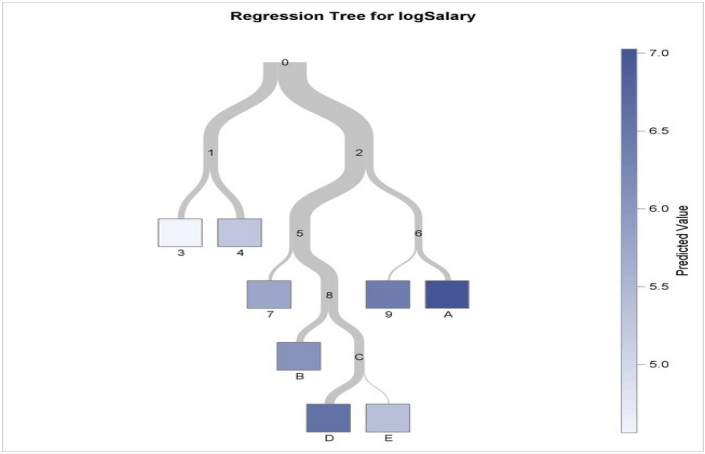

```
Model performance: rmse = 12948.245190958638
mape is =  78.10149522682453
```

<u>Interpretation:</u> This model is also not working good on validation set

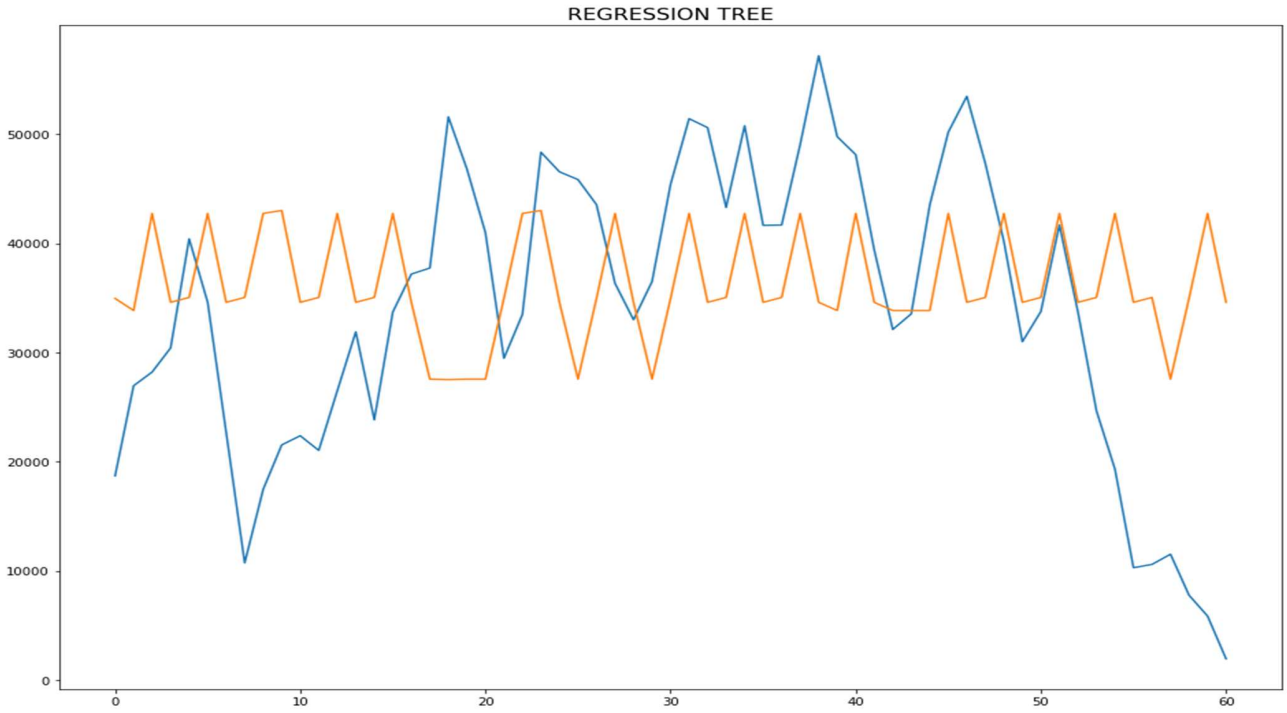
- REGRESSION TREE:

A regression tree is a type of decision tree.It uses sum of squares and regression analysis to predict values of the target field. The predictions are based on combinations of values in the input fields.A regression tree calculates a predicted mean value for each node in the tree. This type of tree is generated when the target field is continuous

Visualization of regression tree:



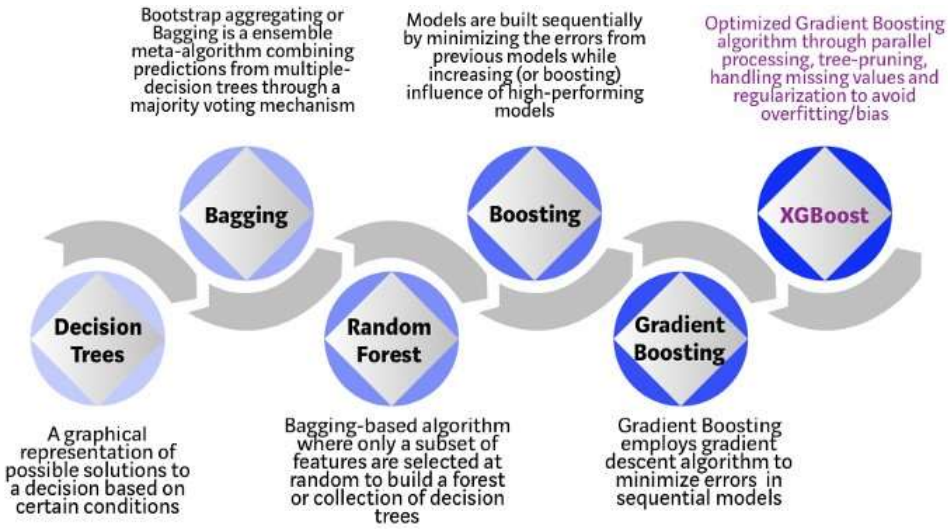
Regression tree applying on overall sales:

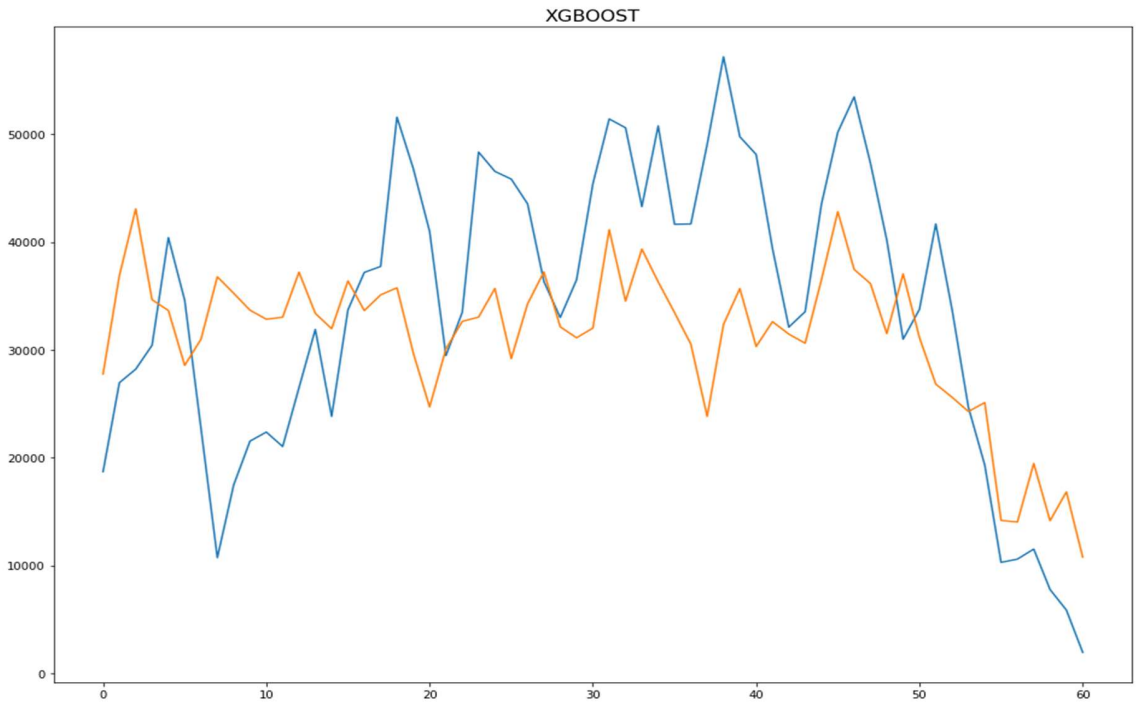



```
Model performance: rmse = 14499.882697563018
mape is =  84.71271415142084
```

INTERPRETATION: The result shows that this model is not suitable to use

- XGBOOST:
  XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now. Please see the chart below for the evolution of tree-based algorithms over the years.






```
Model performance: rmse = 11284.767918070555
mape is =  31.3110130025276
```

INTERPRETATION : This model is better than previous all for this data as for 70% cases the model is forecasting correctly.

CONCLUSION

From the data we have found some insights which can help the company to take important business decisions.These insights are :

- More than 50% people who shops don’t give any feedback so to improve business company can ask customers to rate and give some feedback about the product.

- Only Sao Paolo contributes maximum sale of the company. So company can maintain warehouse operational to get maximum profit. Also in the states with least sale company has opportunity for marketing and increase the sale.

By performing analysis we have found that the data is highly volatile so it was very difficult to achieve the accurate model.Generally in e-commerce industry promotional events, discount sales and festival sales happen during the year,so the data becomes highly volatile.However we have achieved 70% accuracy in creating the model.

- An Application of a Three-Stage XGBoost-Based Model to Sales Forecasting of a Cross-Border E-Commerce Enterprise by **Shouwen Ji,[1] Xiaojing Wang ,[1] Wenpeng Zhao,[2] and Dong Guo[3].Research article** published in Hindawi Mathematical Problems in Engineering, Volume 2019, Article ID 8503252, 15 pages https://doi.org/10.1155/2019/8503252
- Kuhn, M. , & Johnson, K. (2020). Feature Engineering and Selection.New York, NY, USA: Chapman and Hall/CRC
- Suykens, Johan AK, and Joos Vandewalle. "Least squares support vector machine classifiers." *Neural processing letters* 9, no. 3 (1999): 293-300.
- Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (1984). Classification And Regression Trees (1st ed.). Routledge. https://doi.org/10.1201/9781315139470