**E-COMMERCE SALES PREDICTION​**

**​**

MACHINE LEARNING – 18CSC383​

Ms. PRAJNA DORA​

|  |  |
| --- | --- |
| ROLL NUMBER | NAME |
| AASHIKA B S | CB.SC.I5DAS19003 |
| DHANUSHRI M | CB.SC.I5DAS19015 |
| KEERTHANA PRIYA D | CB.SC.I5DAS19018 |
| SHIVADA K P | CB.SC.I5DAS19028 |

**DESCRIPTION:**

Ecommerce is a platform where people buy and sell products conveniently from anyplace or anytime. So, in order to make this experience even better and to improve the profit of the company machine learning approach is taken. E-commerce sales prediction project focuses on predicting the profit

**MOTIVATION:**

People don’t just shop from home; they are making purchases anywhere they have Wi-Fi or phone service. Sixty-two percent of smartphone users have made a purchase online using their mobile device in the last six months. Selling via online increases reach of products.

By incorporating machine learning in e-commerce sector, meaningful insights can be drawn by the business stakeholders and they can make necessary changes to improve their business.

* Determine incremental impacts of new initiatives
* Project future budgets
* Reduced spoilage and fresher, more appealing products through more accurate stock allocation

It will surely create impact on lots of manufactures who uses sales and each and every manufacturer in the world to optimise their inventory and they can get reasonable estimate for their production.

**CHALLENGES:**

Dataset of the e-commerce is too vast to analyse and draw insights from it. Start-up cannot predict the profit due to lack of knowledge. Without proper sales forecasting, many business decisions are based on unreliable estimates or instinct – which leads to many inefficiencies and missed opportunities.

**DESCRIPTION OF DATASET:**

This project uses a dataset from an e-commerce website. The dataset has information of 4117 orders from 2011 to 2014 made at multiple marketplaces in the world. Its features allow viewing an order from multiple dimensions**.**

The dataset is in excel format which contains 2 sheets.

* **LIST OF ORDERS**

Table

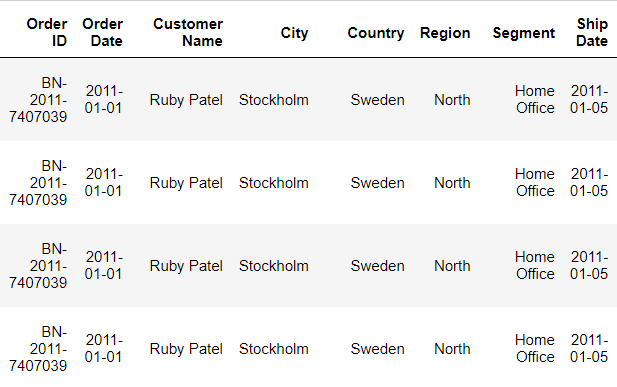
Description automatically generated

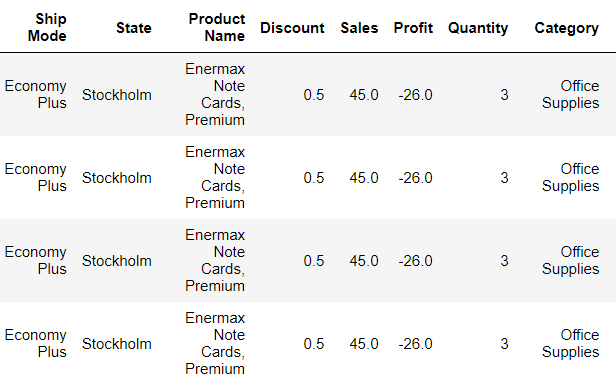
* **ORDER BREAKDOWN**

Table

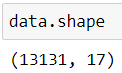
Description automatically generated

**DATASET AFETR MERGING THE SHEETS:**

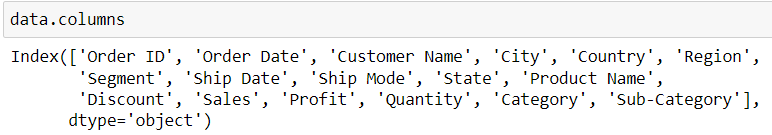
****

****

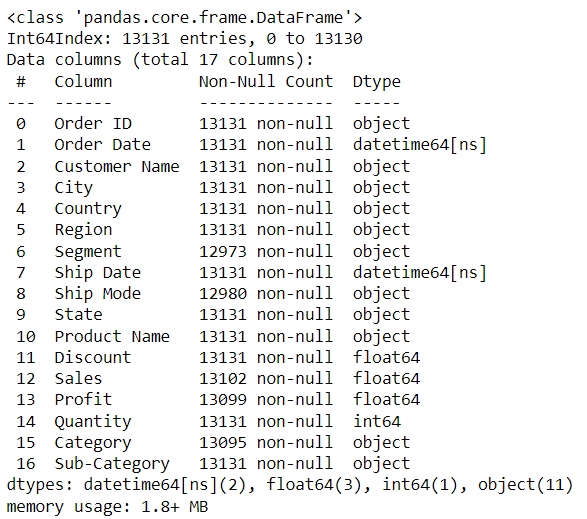
**SHAPE OF DATASET:**

****

**ATTRIBUTES IN DATASET:**

****

**INFORMATION OF DATASET:**

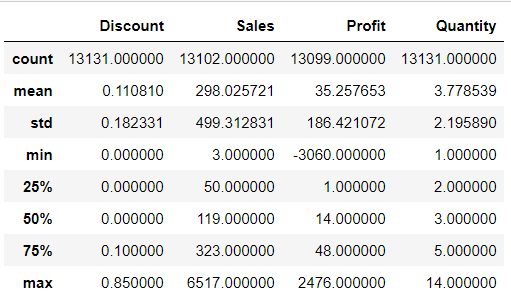
****

**NUMERICAL COLUMNS:**

A picture containing chart

Description automatically generated

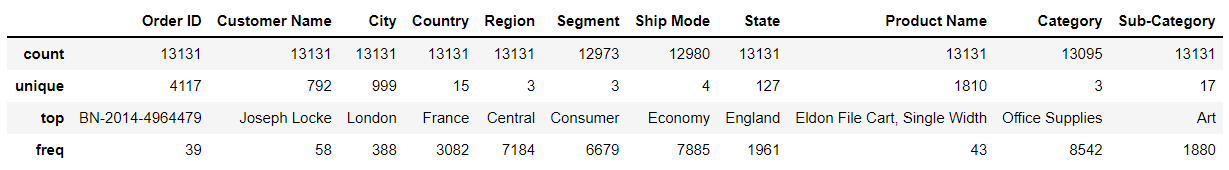
**DESCRIPTION OF NUMERICAL ATTRIBUTES:**



**CATEGORICAL COLUMNS:**

****

**DESCRIPTION OF CATEGORICAL ATTRIBUTES:**

****

**INDEPENDENT ATTRIBUTES:**

All attributes except Profit attribute.

**DEPENDENT ATTRIBUTES:**



**LINK TO DATA:**

[E-COMMERCE-DATASET](https://github.com/Keerthanadevaraj11/E-COMMERCE-DATASET)

**TO PRDEICT PROFIT:**

**KIND OF ATTRIBUTES**:

**DEPENDENT ATTRIBUTE**: Profit

**INDEPENDENT ATTRIBUTE:** All attributes except Profit attribute.

**NATURE OF OUTPUT**:

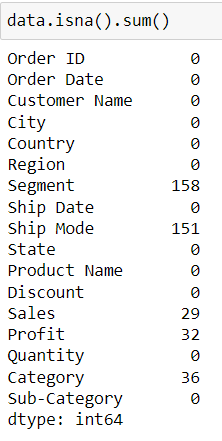
**REGRESSION MODEL:** Profit – Numerical values

**ALGORITHMS USED:**

Linear Regression, Decision Tree, Random Forest, KNN

**DATA PRE-PROCESSING:**

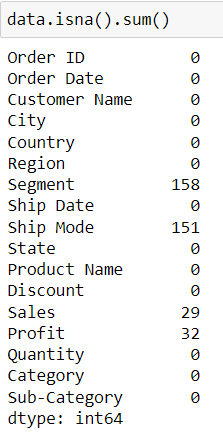
**HANDLING NULL VALUES:**



**HANDLING NULL VALUES IN CATEGORY ATTRIBUTE:**

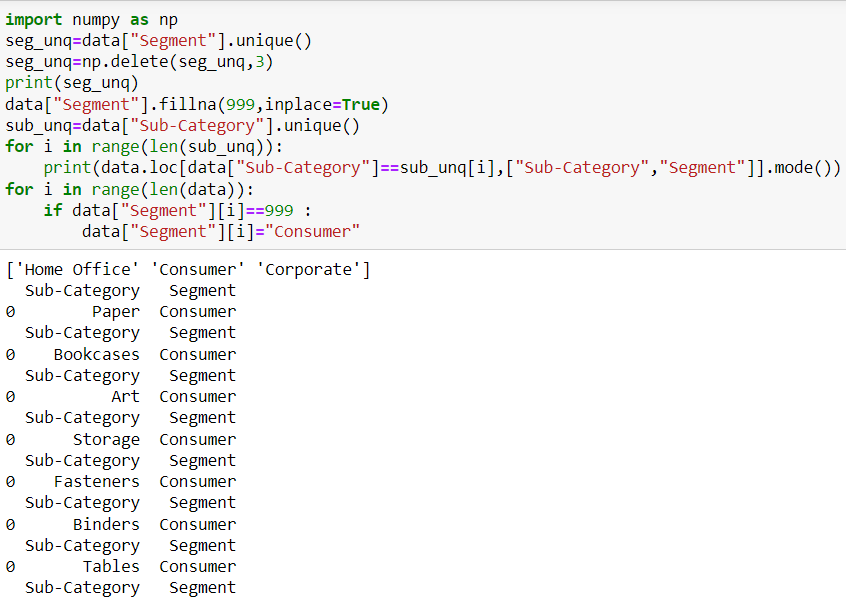
Null values in Category attribute are handled by considering their sub-category. The mode of each sub-category’s category is used to replace null values.

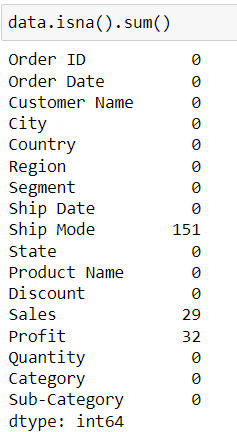




**HANDLING NULL VALUES IN SEGMENT ATTRIBUTE:**

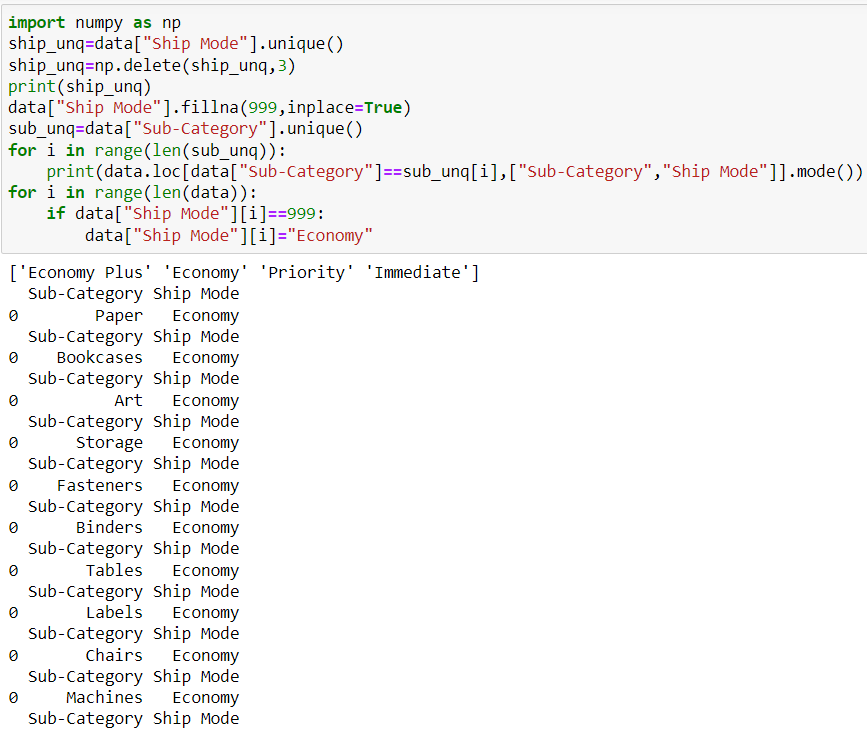
Null values in Segment attribute are handled by considering their sub-category. The mode of each sub-category’s segment is used to replace null values.

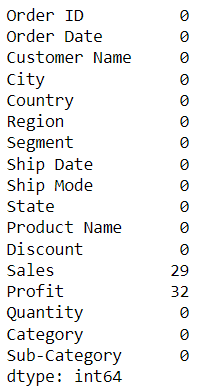




**HANDLING NULL VALUES IN SHIPPING MODE ATTRIBUTE:**

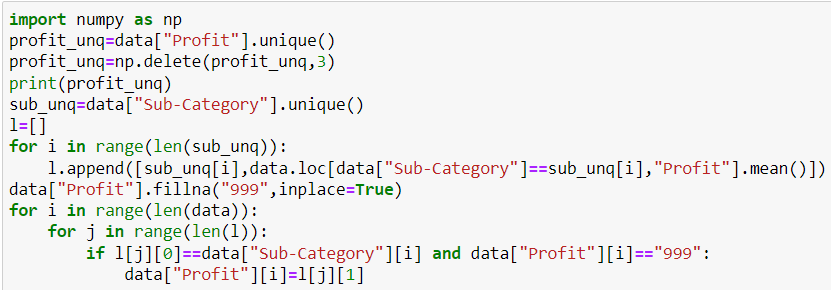
Null values in Shipping mode attribute are handled by considering their sub-category. The mode of each sub-category’s Shipping mode is used to replace null values.

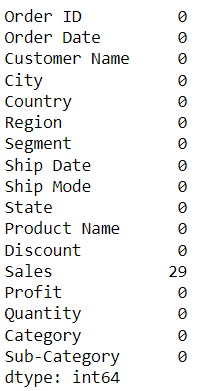




**HANDLING NULL VALUES IN PROFIT ATTRIBUTE:**

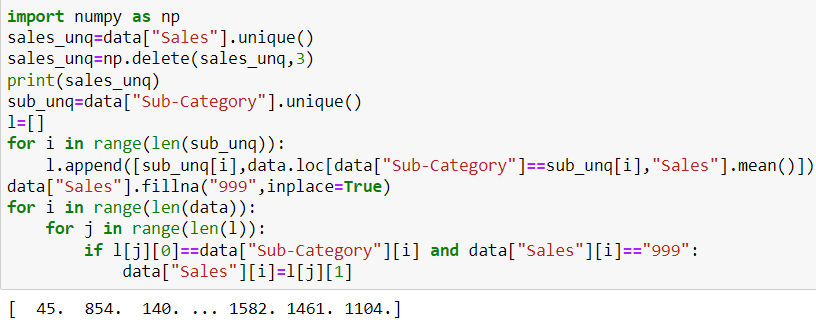
Null values in Profit attribute are handled by taking mean of profit of the sub-category they belong to.

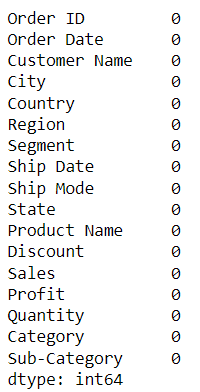




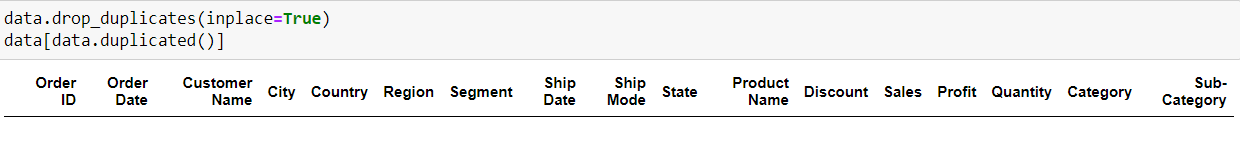
**HANDLING NULL VALUES IN SALES ATTRIBUTE:**

Null values in Sales attribute are handled by taking mean of profit of the sub-category they belong to.





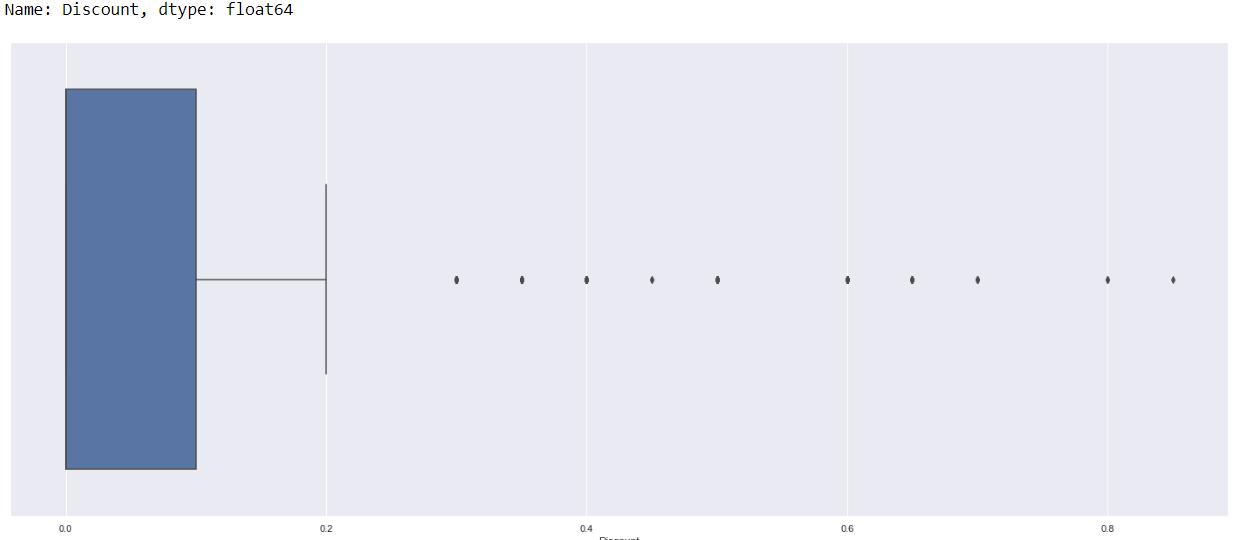
**HANDLING DUPLICATE VALUES:**

****

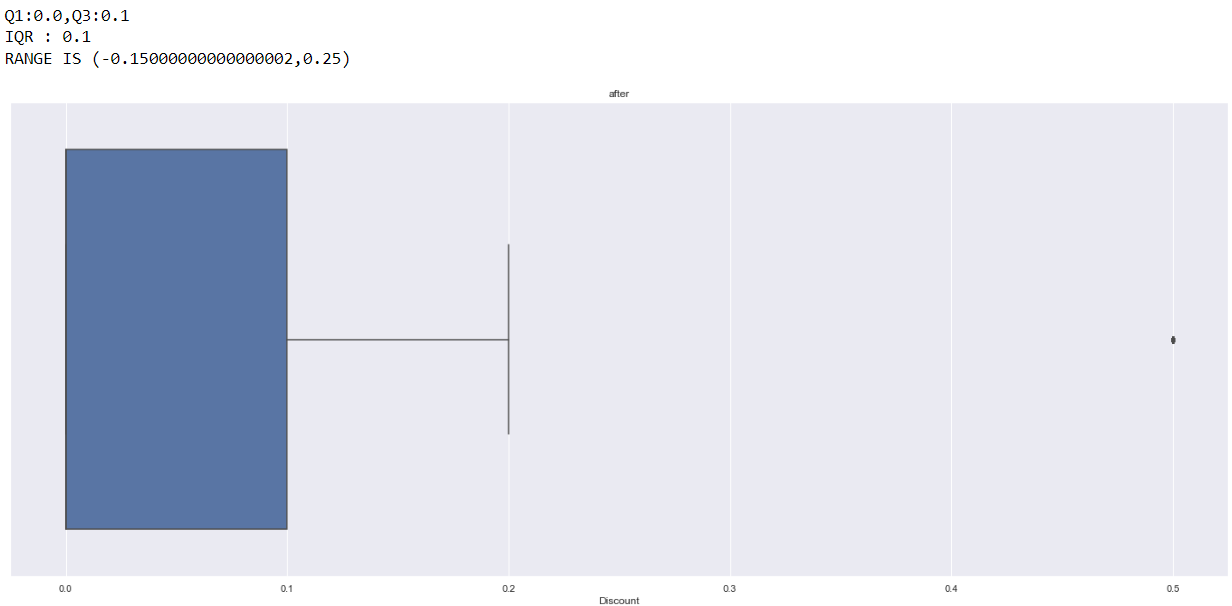
**OUTLIERS:**

Removing outliers from numerical attributes.

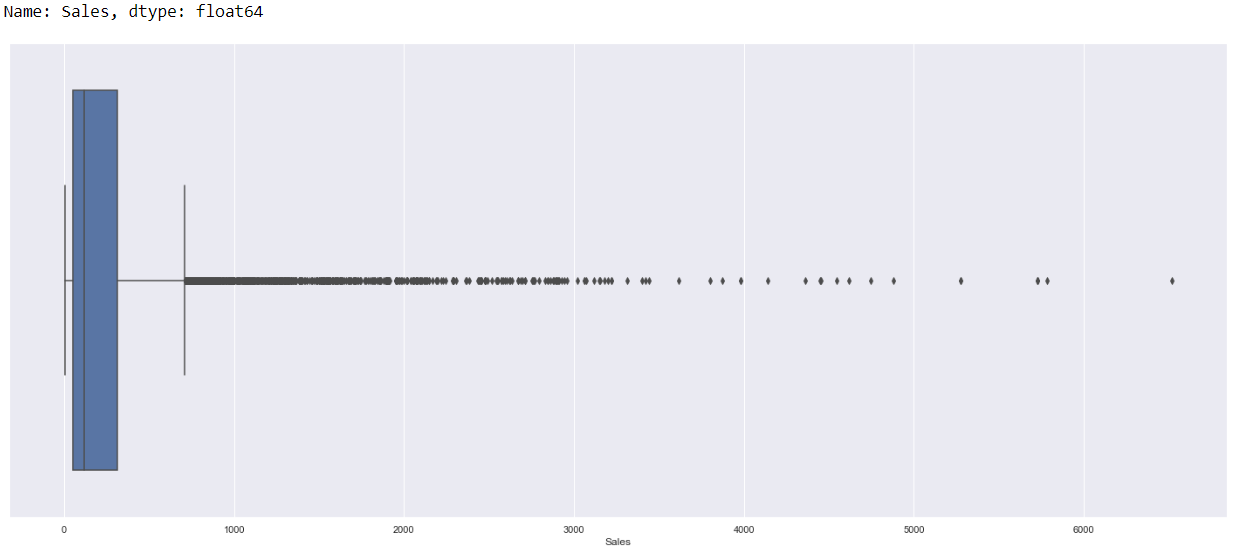
**BEFORE OUTLIER TREATMENT IN DISCOUNT ATTRIBUTE:**



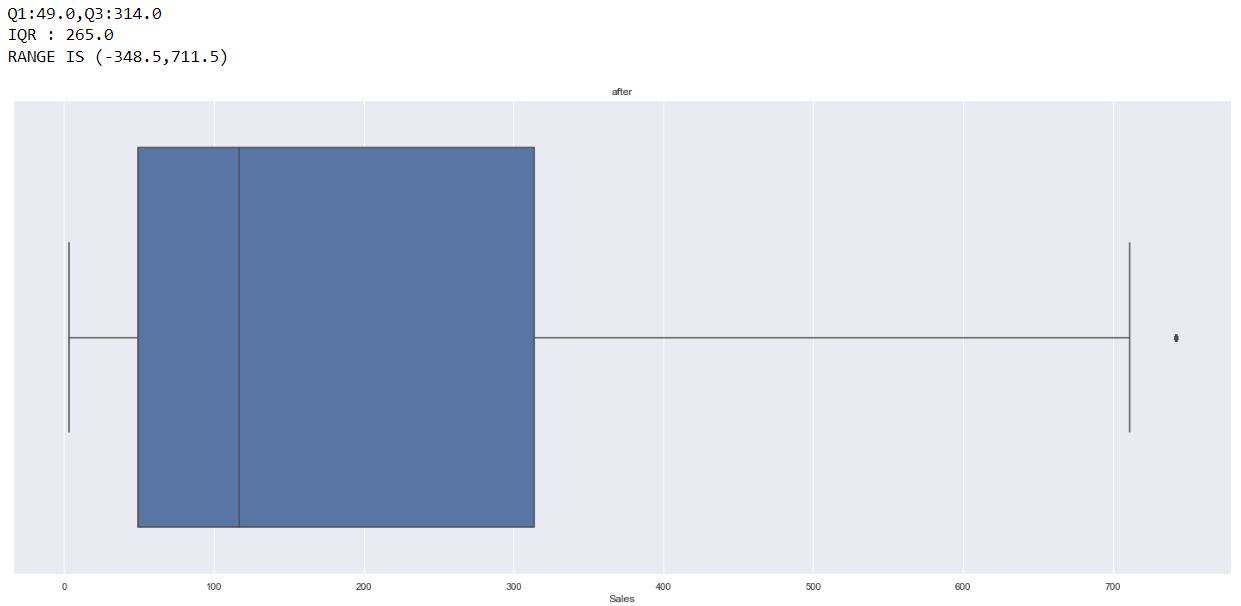
**AFTER OUTLIER TREATMENT IN DISCOUNT ATTRIBUTE:**



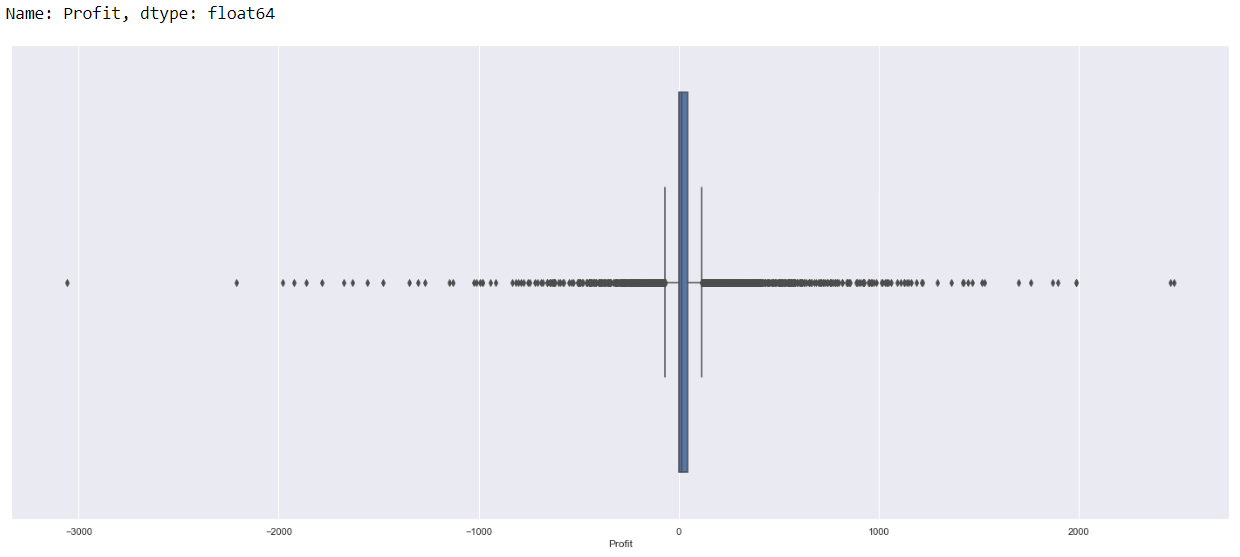
**BEFORE OUTLIER TREATMENT IN SALES ATTRIBUTE:**



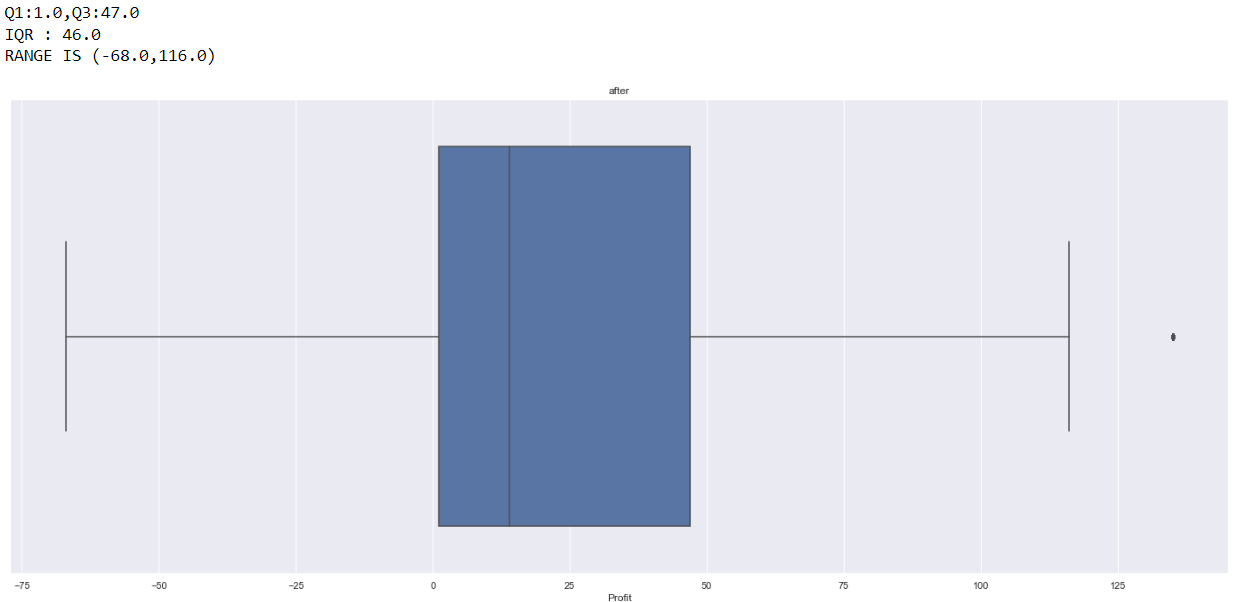
**AFTER OUTLIER TREATMENT IN SALES ATTRIBUTE:**



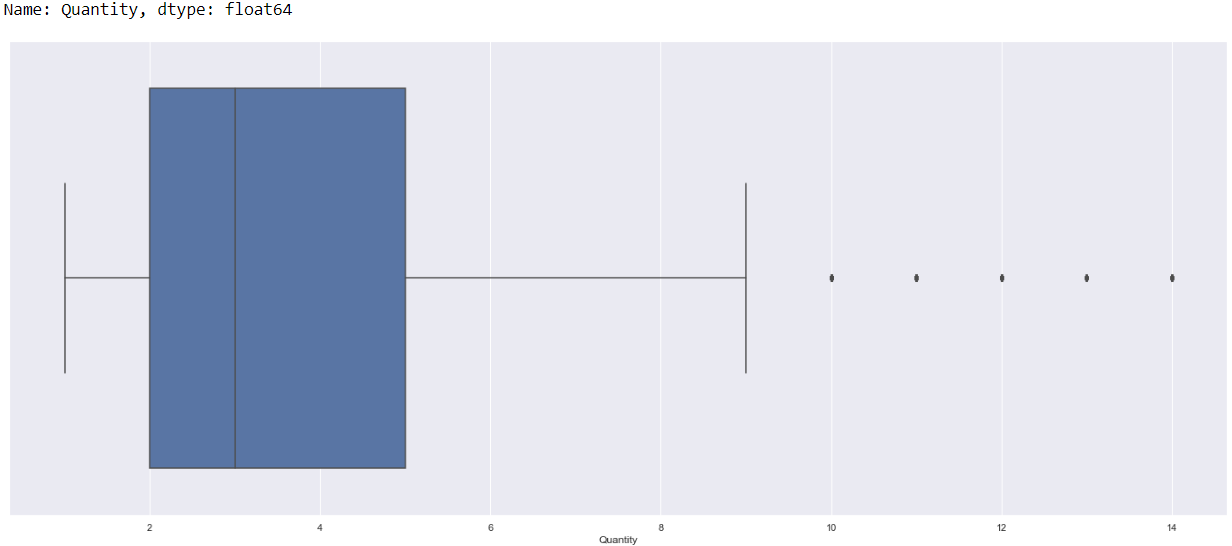
**BEFORE OUTLIER TREATMENT IN PROFIT ATTRIBUTE:**

****

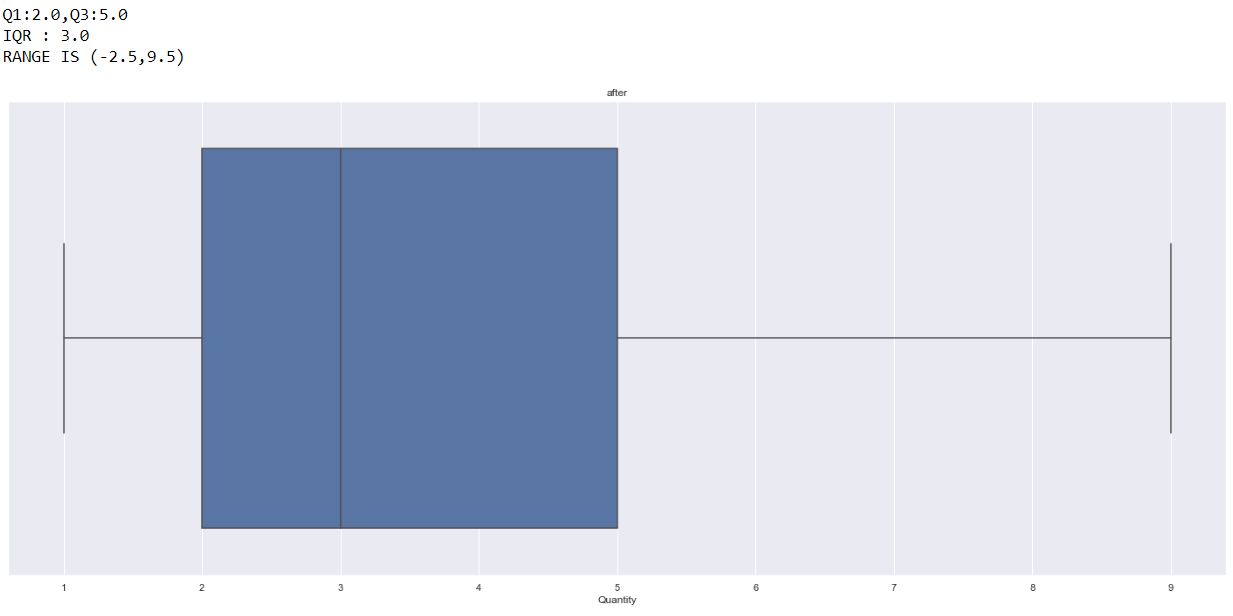
**AFTER OUTLIER TREATMENT IN PROFIT ATTRIBUTE:**



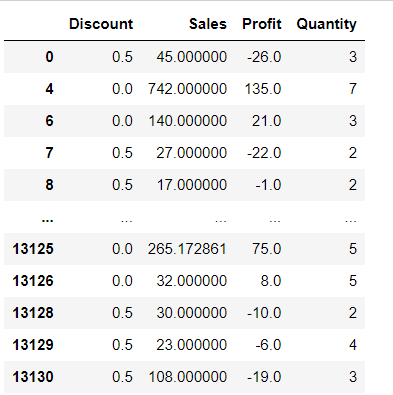
**BEFORE OUTLIER TREATMENT IN QUANTITY ATTRIBUTE:**



**AFTER OUTLIER TREATMENT IN QUANTITY ATTRIBUTE:**

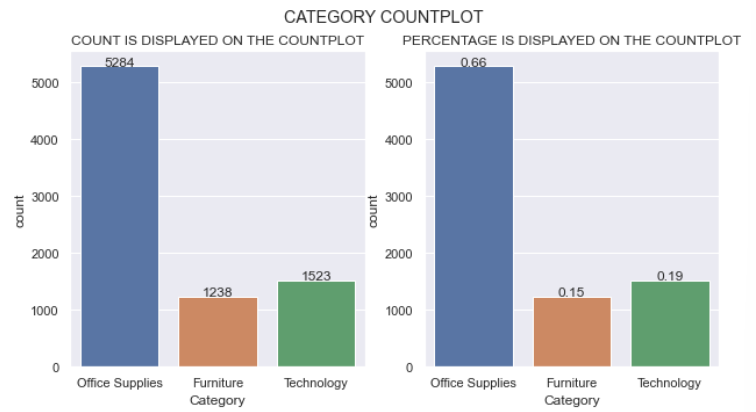


**NORMALIZATION OF NUMERICAL ATTRIBUTES:**

****

**EXPLORATORY DATA ANALYSIS:**

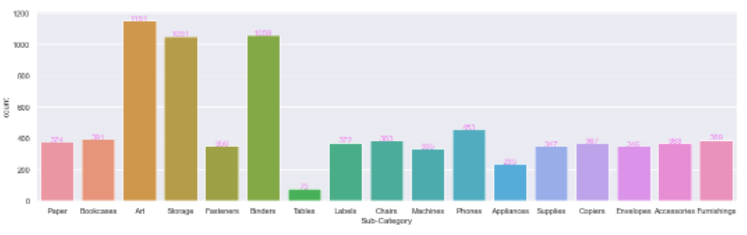
**SALES PERCENTAGE FOR 3 DIFFERENT CATEGORIES:**

****

**INFERENCE:**

Office supplies (0.65) > Technology (0.19) >Furniture (0.16). Therefore, Office supplies have sold more than furniture and technology.

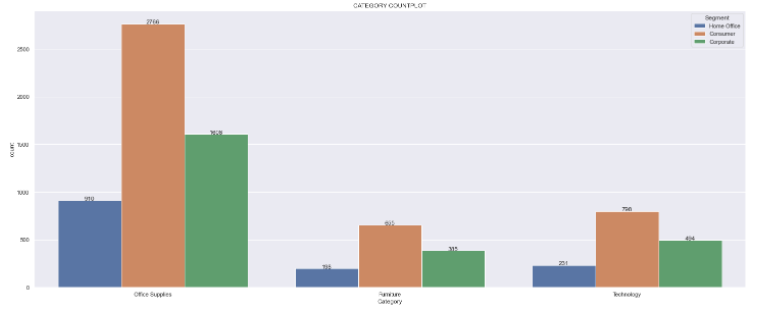
**SALES PERCENTAGE FOR DIFFERENT SUB CATEGORIES:**

****

**INFERENCE:**

Art sub-category has higher sales and takes sub-category has lower sales.

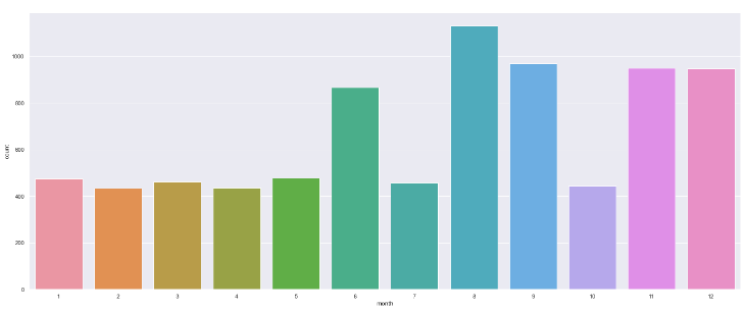
**SALES PERCENTAGE FOR 3 DIFFERENT CATEGORIES ACCORDING TO THE SEGMENT OF CUSTOMERS:**

****

**INFERENCE:**

Consumer segment have purchased more products than others in all category and home office segment have purchased less products than others in all category.

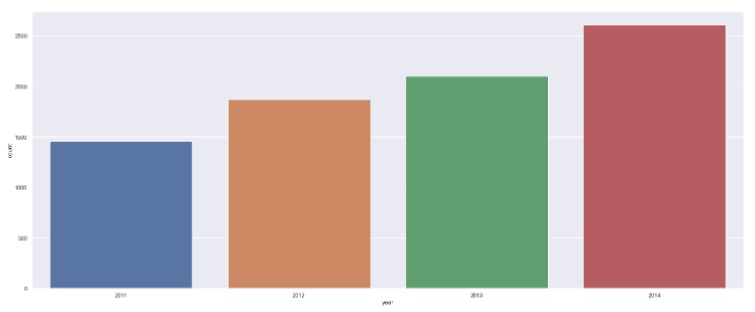
**MONTHLY AVERAGE PROFIT FOR ALL YEARS:**

****

**INFERENCES:**

In the month of August, sales were higher and the second is September and in February the sale is down.

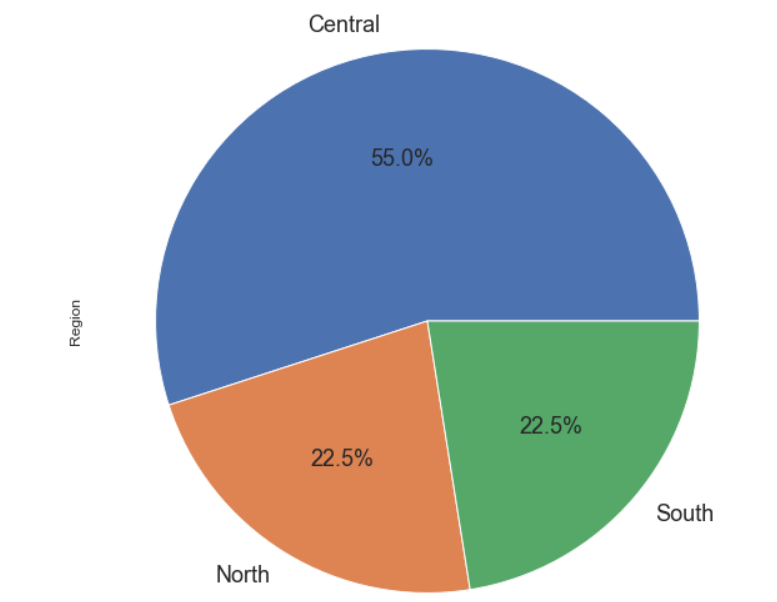
**YEARWISE PROFIT:**



**INFERENCES:**

As the year passes, sales are gradually increasing.

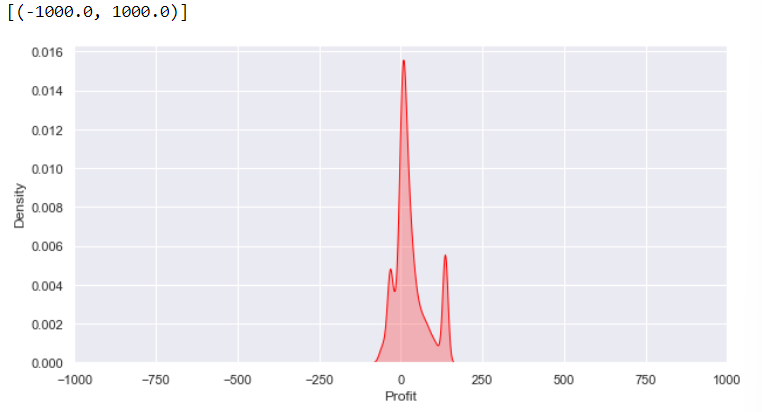
**SALES OF PRODUCTS FOR DIFFERENT REGIONS:**

****

**INFERENCE:**

Many people from Central region have purchased more products than other regions and South and North region people have purchased only less products.

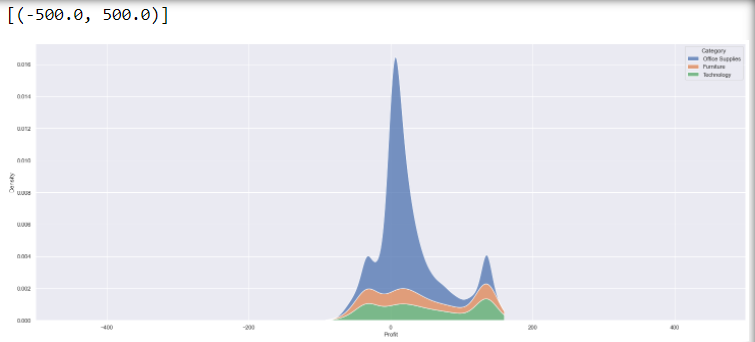
**DISTRIBUTION OF PROFIT:**

****

**INFERENCE:**

Distribution of profit is very much higher than distribution of loss.

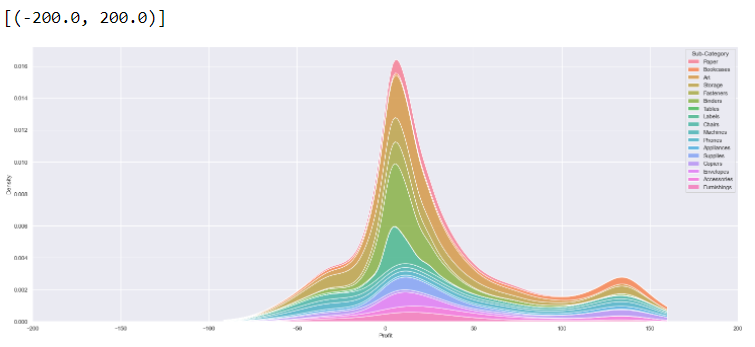
**PROFIT DISTRIBUTION DEPENDING ON CATEGORY:**



**INFERENCES:**

Probability density for profit and categorized by category attribute. From the above, we can see that office supplies have the highest profit.

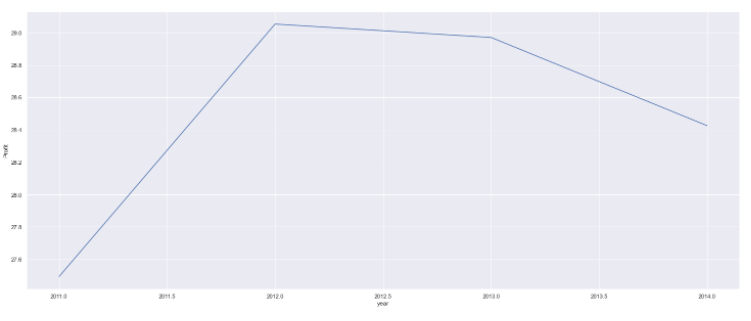
**PROFIT DISTRIBUTION DEPENDING ON SUB-CATEGORY:**

****

**INFERENCES:**

Probability density of profit categorized by sub-category attribute where labels sub category has the highest profit.

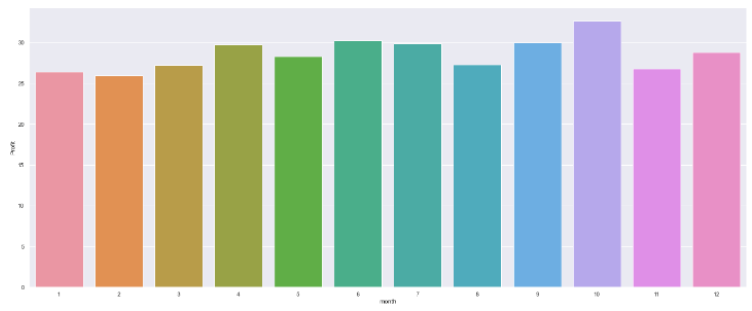
**DISTRIBUTION OF PROFIT YEARWISE:**

****

**INFERENCE:**

The profit is highest for the year 2012 and it is lowest for the year 2011.

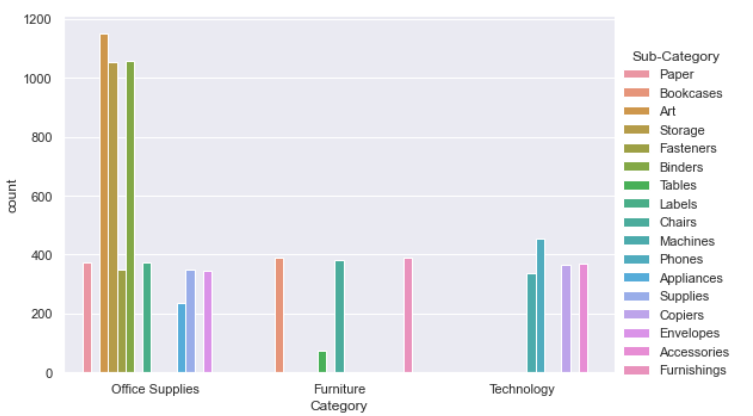
**DISTRIBUTION OF PROFIT MONTH WISE**:



**INFERENCE:**

The profit is higher at the month of October and lowest at the month of February.

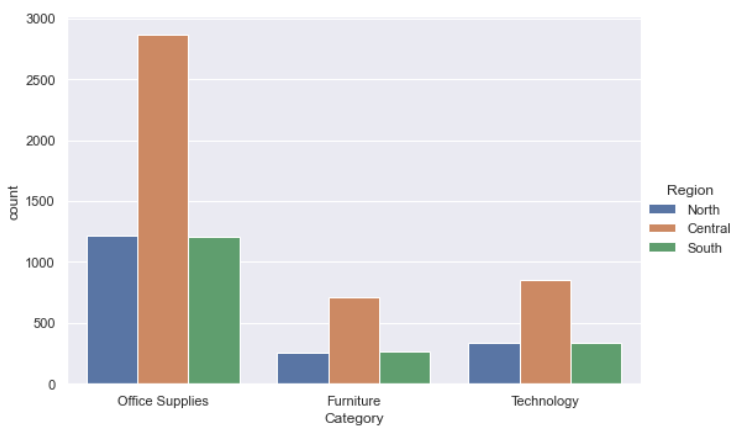
**SALES OF CATEGORY DEPENDING ON SUB- CATEGORY:**

****

**INFERENCE:**

Sub Categories of Office supplies are purchased more and Sub Categories of Furniture and Technology are purchased in lower rate

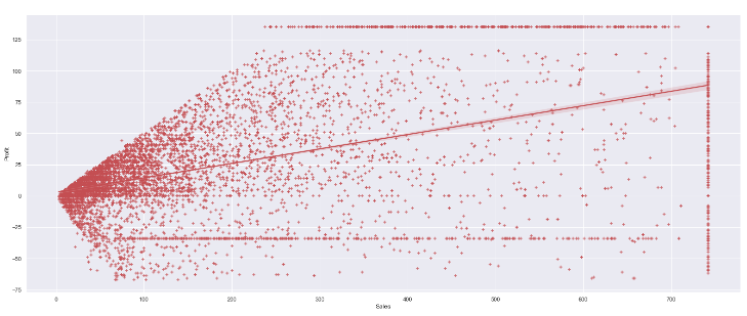
**SALES OF CATEGORY DEPENDING ON REGION:**

****

**INFERENCE:**

Office supplies are purchased more than furniture and technology category. Central region has more sales in all categories.

**REGRESSION PLOT FOR SALES AND PROFIT:**

****

**INFERENCES:**

The best fit line for the attributes sales and profit.

**SALES OF DIFFERENT CATEGORIES AND SEGMENTS DEPENDING ON SUB-CATEGORY:**

****

**INFERENCE:**

Consumer segment has purchased more in all categories than the other segments. In each segment, Office Supplies are purchased by many customers.

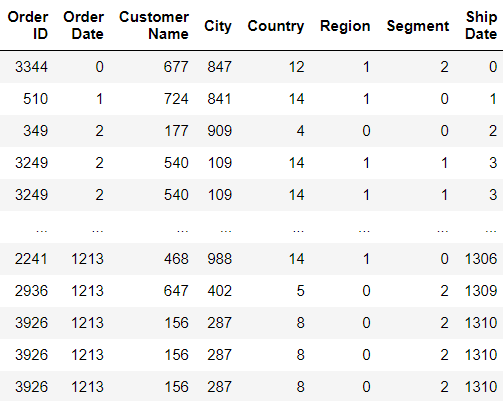
**CORRELATION MATRIX:**

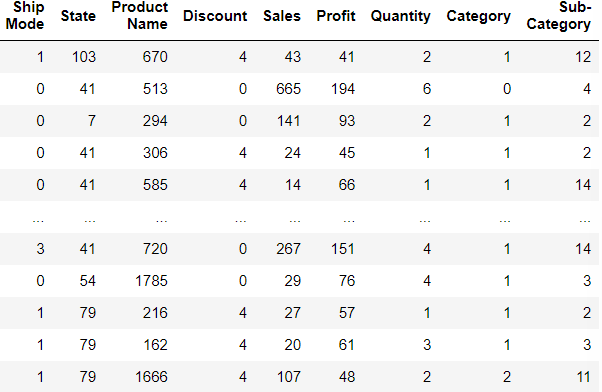


**INFERENCES:**

It shows the correlation between each and every numerical attribute of the dataset. Profit and sales have the second highest correlation.

**ENCODING OF CATEGORICAL VARIABLES:**

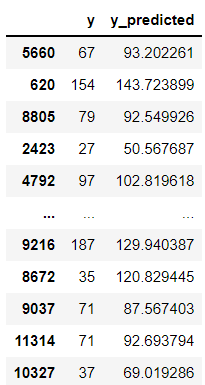
****

****

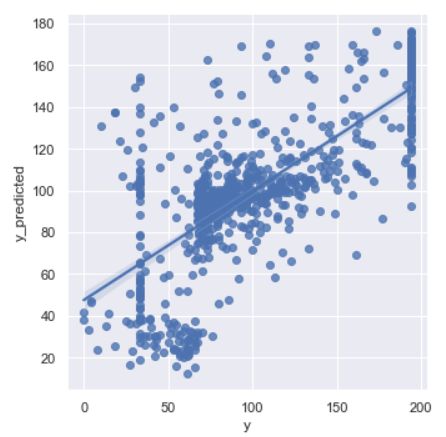
**LINEAR REGRESSION WITHOUT PCA AND BAGGING:**

****

**ACTUAL VALUE AND PREDICTED VALUE:**

****

**PLOT TO REPRESENT DIFFERENCE BETWEEN ACTUAL VALUE AND PREDICTED VALUE:**

****

**LINEAR REGRESSION WITH PCA AND BAGGING:**

****

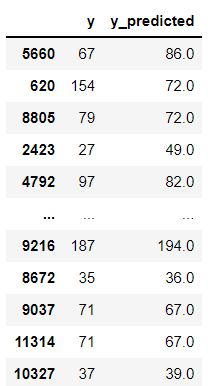
**HYPERPARAMETER TUNING:**

****

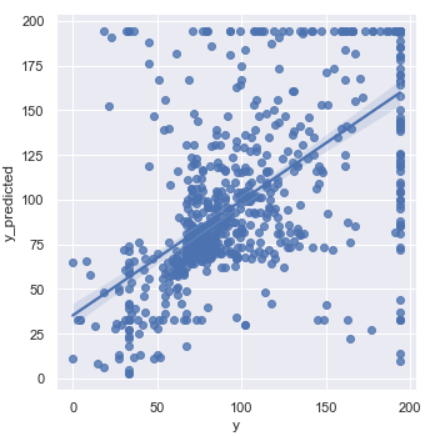
**DECISION TREE WITHOUT PCA AND BAGGING:**

****

**ACTUAL VALUE AND PREDICTED VALUE:**



**PLOT TO REPRESENT DIFFERENCE BETWEEN ACTUAL VALUE AND PREDICTED VALUE:**



**DECISION TREE WITH PCA AND BAGGING:**

****

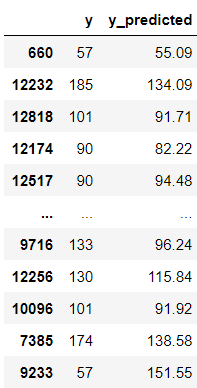
**HYPERPARAMETER TUNING:**

****

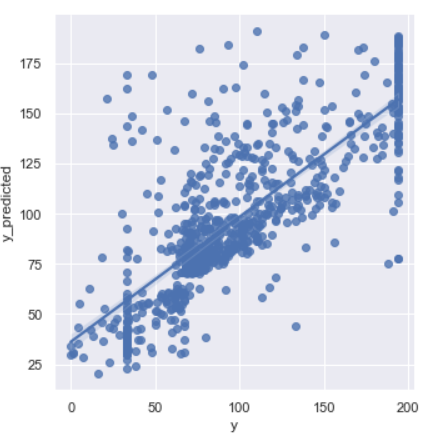
**RANDOM FOREST WITHOUT PCA AND BAGGING:**

****

**ACTUAL VALUE AND PREDICTED VALUE:**

****

**PLOT TO REPRESENT DIFFERENCE BETWEEN ACTUAL VALUE AND PREDICTED VALUE:**

****

**RANDOM FOREST WITH PCA AND WITHOUT BAGGING:**

****

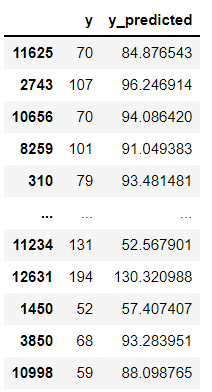
**HYPERTUNING:**

****

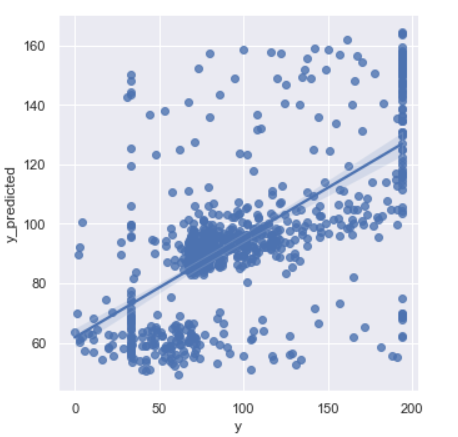
**KNN WITHOUT PCA AND BAGGING:**

****

**ACTUAL VALUE AND PREDICTED VALUE:**

****

**PLOT TO REPRESENT DIFFERENCE BETWEEN ACTUAL VALUE AND PREDICTED VALUE:**

****

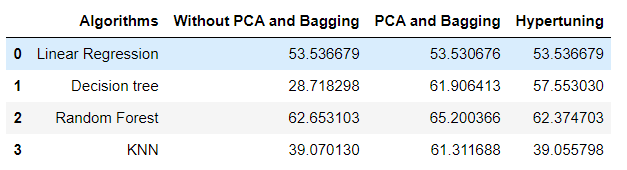
**KNN WITH PCA AND BAGGING:**

****

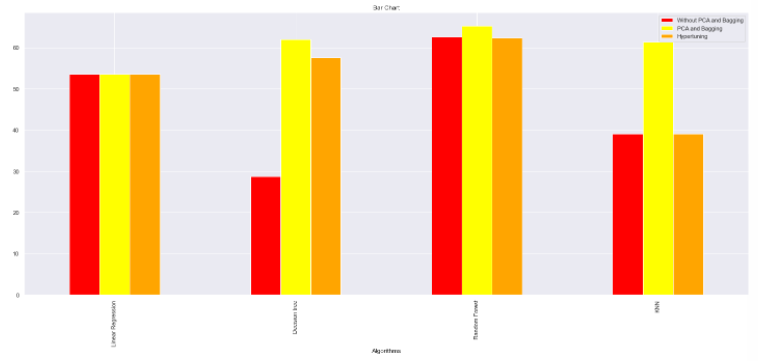
**HYPERPARAMETER TUNING:**

****

**ACCURACY OF EACH MODEL:**

****

**VISUALIZATION OF EACH MODEL ACCURACY:**

****

**BOOSTING:**

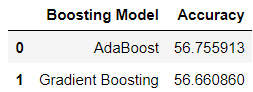
**ADABOOST:**

****

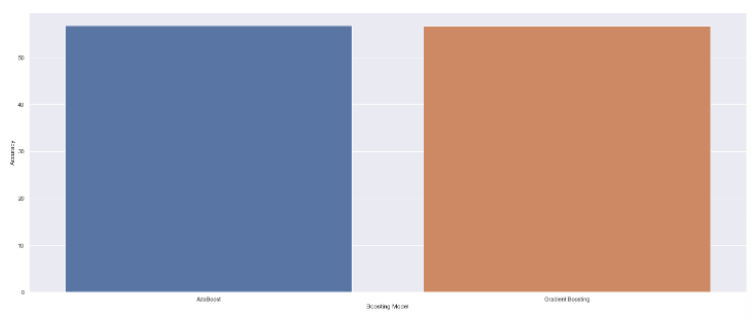
**GRADIENT BOOSTING REGRESSOR:**

****

**ACCURACY OF EACH BOOSTING MODEL:**

****

**VISUALIZATION OF EACH BOOSTING MODEL:**

****

**CONCLUSION:**

It is seen that Random Forest gives more accuracy in predicting e-commerce profit than Decision Tree, Linear Regression, KNN. Considering the criteria’s:

* Without PCA and Bagging
* With PCA and Bagging
* Hyperparameter tuning

In all these three criteria’s Random Forest gives the maximum accuracy.