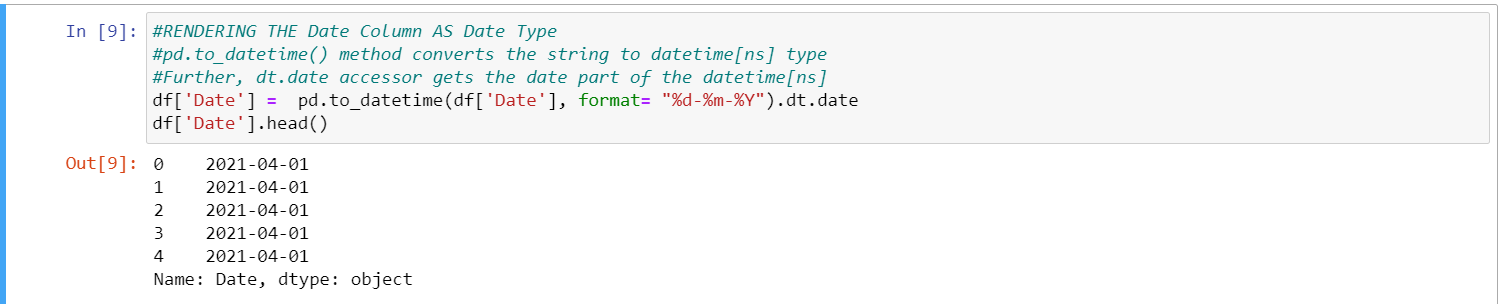
**BASIC STATS – 1**

1. **INFO ABOUT THE DATA SET & INITIAL PREPROCESSING**

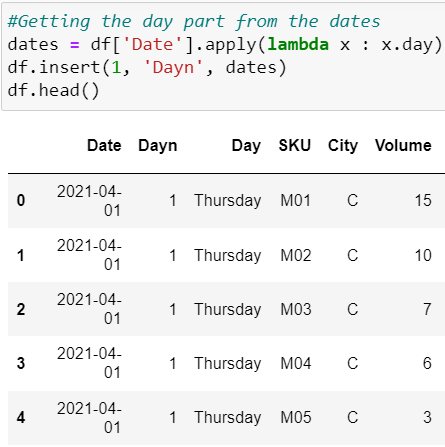
* There are 13 columns and 450 rows in the dataset
* The dataset is the sales data of a SuperMarket. The data is for 30 different products over 15 days
* The columns of the dataset are Date, Dat, SKU, City Volume, BU , Brand, Model, Avg Price, Total Sales Value, Discount Rate (%), Discount Amount, Net Sales Value

**Preprocessing:**

* ‘Date’ column is originally of object type. Convert it to ‘DataTime’ type

****

* Adding a ‘Dayn’ column which holds only the day part of the date e.g: (1, 2, 3,….)

****

1. **DESPCRIPTIVE STATS**

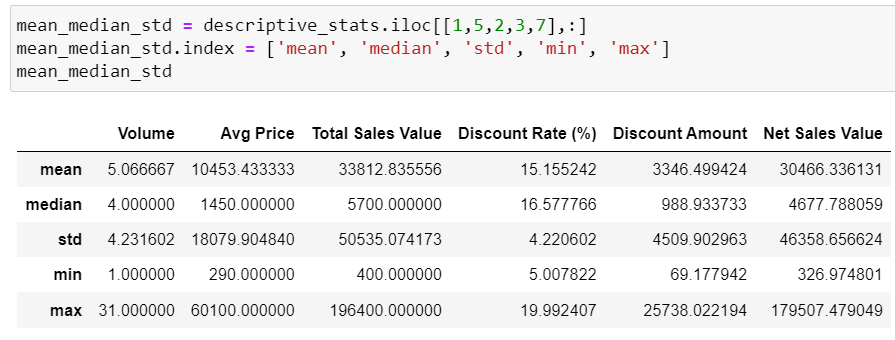
Mean and Median are measures of central tendencies. They give us an idea of where does the center of our data lies. The definition of center differs for mean and median.

**Mean**: The mean is a measure of central tendency that is calculated by summing all the values in the data set and then dividing by the number of values. It gives us the arithmetic average of the data points and is influenced by the magnitude and direction (positive or negative) of the values.

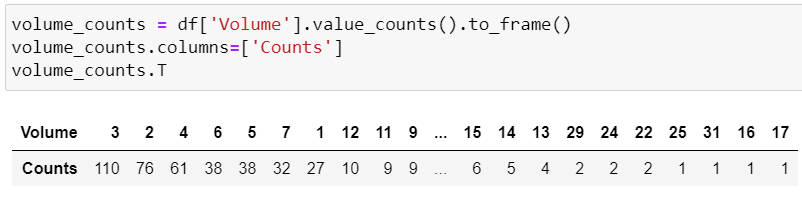
**Median**: The median is a measure of central tendency that is the middle value when the data points are ordered from smallest to largest. If the data set has an odd number of observations, the median is the middle value. If the data set has an even number of observations, the median is the average of the two middle values. The median is determined solely by the order of the values and is not affected by their magnitude or direction.

**STD:** Standard Deviation gives gives the measure of dispersion in the data. It indicates on an average how far the datapoinsts are from the mean value. The disadvantage of STD is that it cannot be compared for two datasets.

* **Mean, Median & STD of the numerical columns**

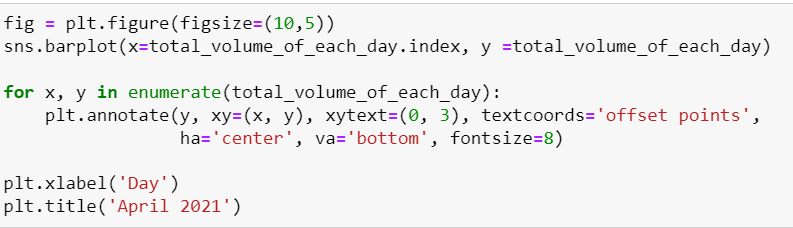


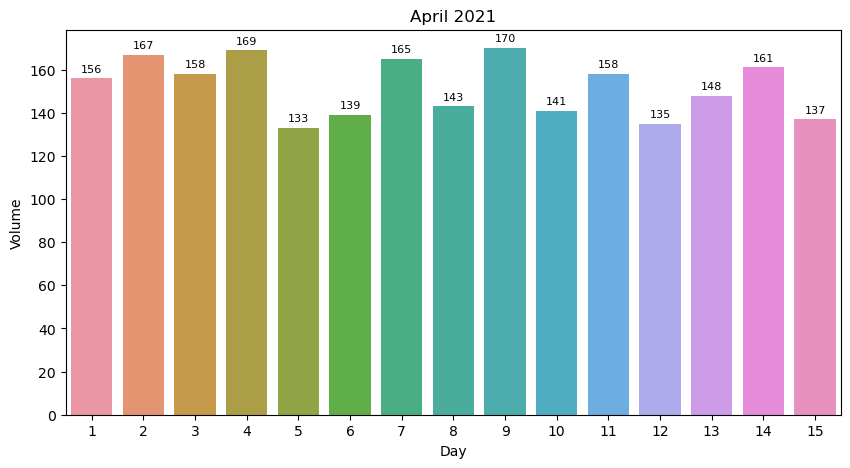
* Except for ‘Discount Rate’ column, Median is Very less compared to the Mean, for rest of the columns
* This indicates that there are a lot of smaller values for these columns compared to bigger values.
* This phenomenon is generally expected in any supermarket, Lower Price Items like (Creams, Shampoo, etc ) are purchased way more in quantity than Higher price items like (mobiles)
* This indicates that the distribution of data is very much skewed at the right side.
* The range (max - min) is very wide, this results in high STD values for each column.
* **Mode of Volume column**



* Mode value of Volume Column is ‘3’ which has the highest frequency of 110
* This explains the difference in mean and median of the volume column.
* Since the frequency of low volumes is very high it results in median lower than the mean.
* Here, Mode<Median<Mean. It is expected for a right skewed distribution.

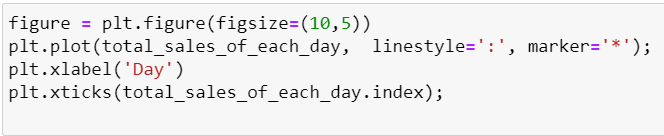
1. **VISUALIZATION**
2. **Bar Plot for Volume of each day**

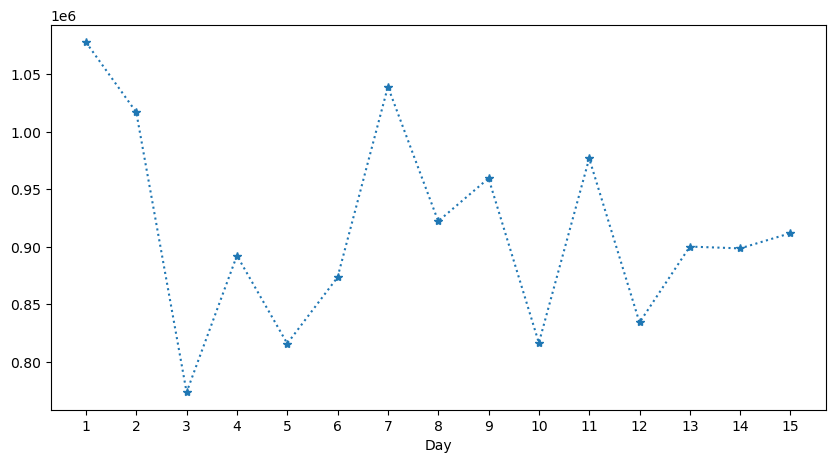


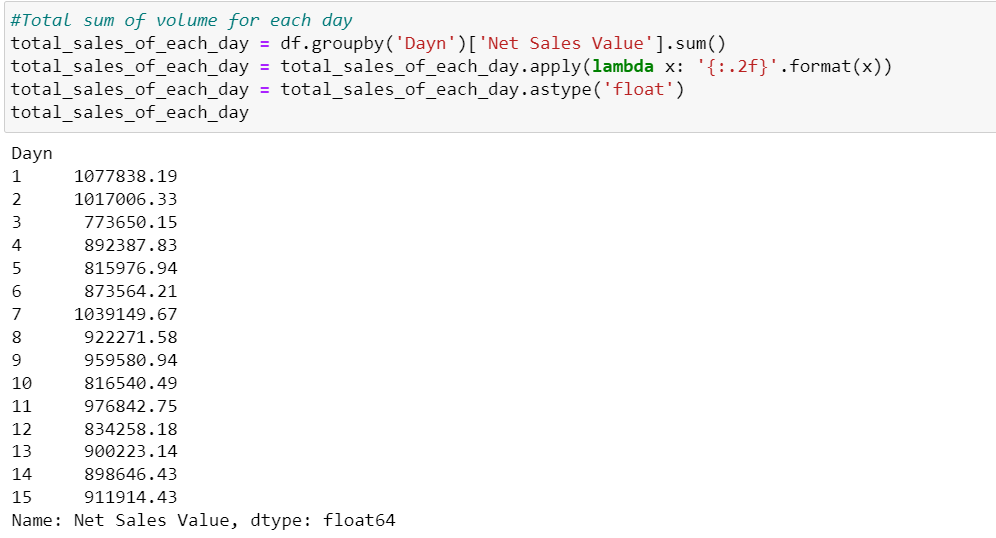


* Median Volume is 156 (Day 1)
* Mean Volume is 152
* The Volume is evenly spread in the range (133-170)
* Highest Volume was on Day 9, Lowest on Day 5

1. **Line graph showing changes in Net Sales Value across 15 days**



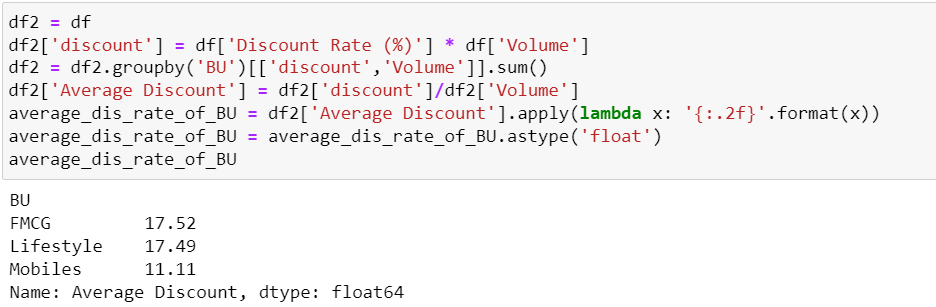


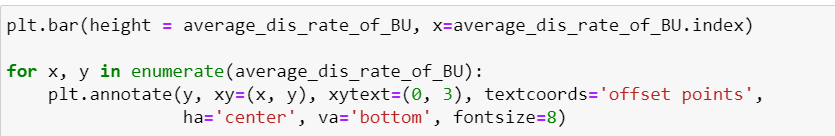


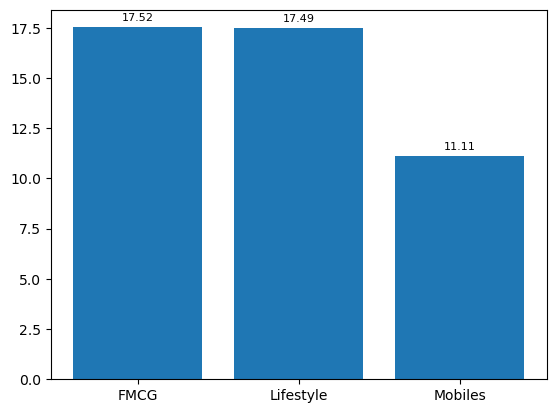
* Highest Sale Value was on Day 1 where as the Volume on Day 1 was the median Volume
* The lowest Sales Value was on Day 3 : 7 Lac. 73 Thousand
* The highest Sales Value was on Day 1 : 10 Lac. 77 Thousand
* Only three days have received above 10 Lac sales. Day 1, 2 & 7
* Median : 9,00,223 Mean : 9,13,990.

1. **Discount Rate (%):**

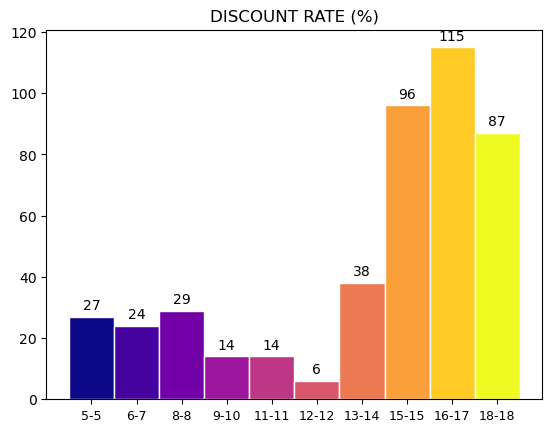
* **Average Discount Rate of each BU**







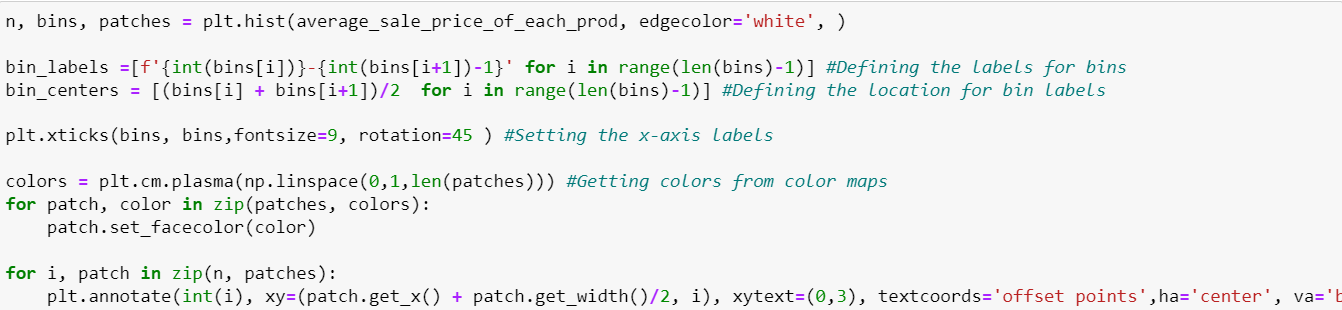
* + FMCG & Lifestyle goods have an equal discount rate of 17.5% as they are mostly lower price items.
  + Mobiles are costlier hence their discount rate is significantly low in comparison: 11%
* **Histogram of Discount Rate:**

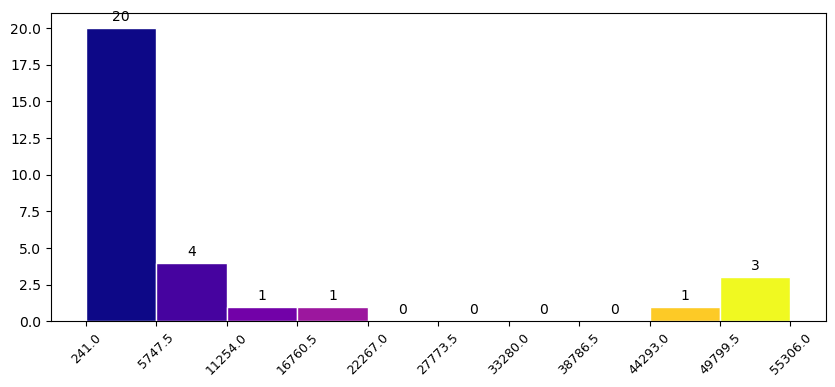


* Histogram of Discount Rate shows highest recurring discounts are in the (16% - 17%) bracket
* An overwhelming majority of discounts are greater than 13%

1. **Average Sales Price of Each Product:**

* Average Sales prices is the average price of a product after deducting the discounts from the price.
* There are 30 different products across all the BUs including FMCG, Lifestyle and Mobiles.
* Average Sales price of all products create the following histogram:

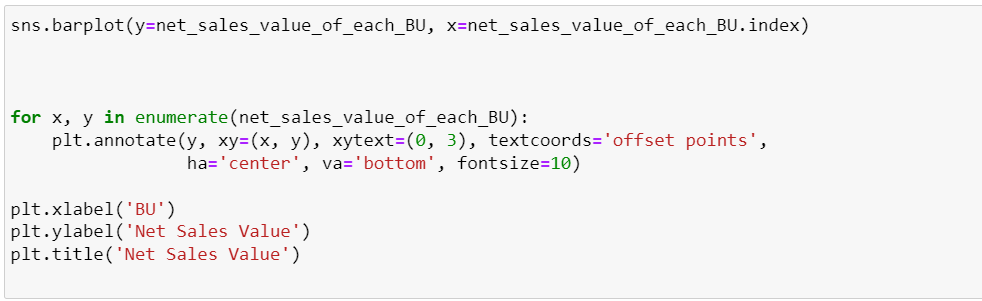


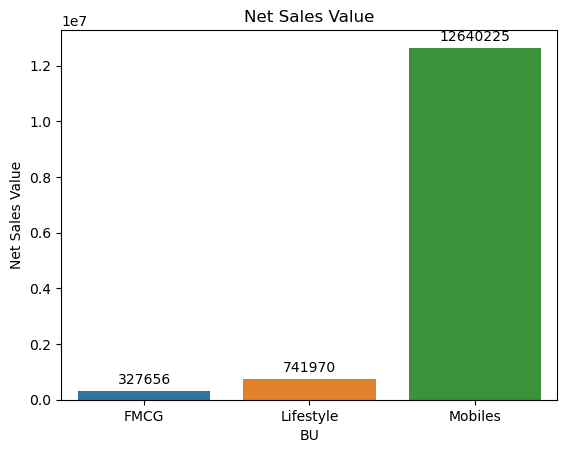


* Here we can see most of the products are below Rs.5,747. These are mostly FMCG & Lifestyle Goods.
* Mobiles are expected to be expensive products. There are 10 different models of mobiles. All these models must be on the higher price bins.
* We can see 4 mobile models have prices greater than Rs. 44,293
* 2 mobiles are between (11k – 22k)
* 4 mobiles are between (5k – 11k)

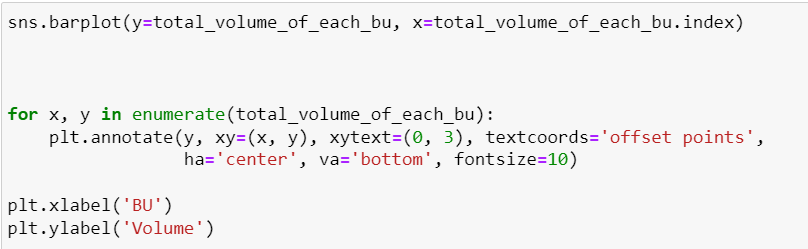
1. **NET SALES VALUE v/s VOLUME**

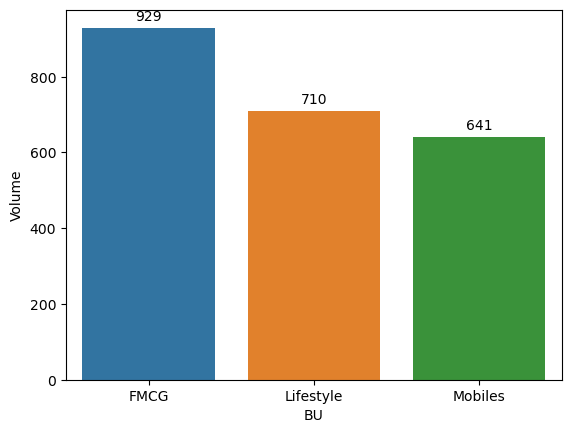
* **Net Sales Value of each BU**





* Mobiles clearly dominate the Net Sales Value of the supermarket.
* Lifestyle sales is more than twice the sales of FMCG
* **Total Volume of each BU:**

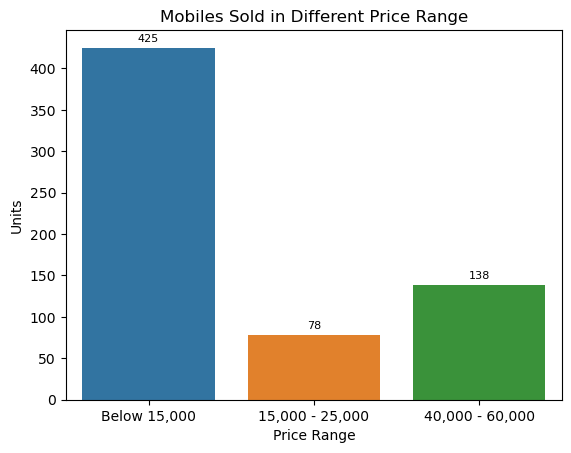




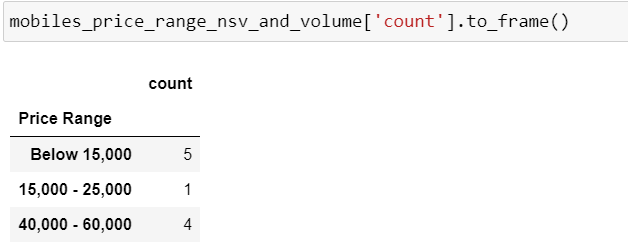
* We can see that the volume of Mobiles and Lifestyle is almost equal and Mobiles is not very less compared to FMCG
* Considering the fact that Lifestyle and FMCG products are much cheaper than Mobiles yet mobiles are very close to them in Volume. This suggests that the supermarket is mainly preferred for Mobiles

1. **MOBILES**

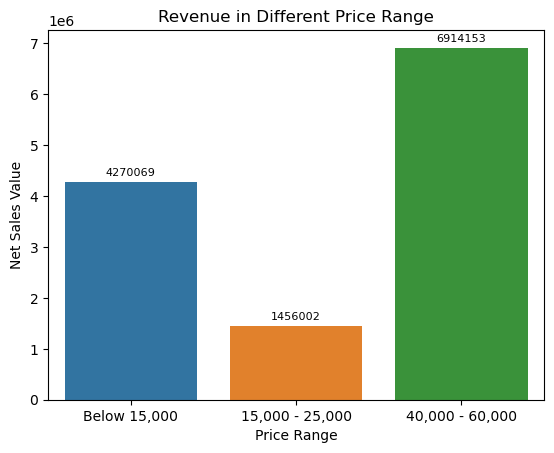
* Mobiles make the most of the revenue of the supermarket.
* There are 10 different models of mobiles with difference prices.
* The price range can be categorized as 'Below 15K , '15K– 25K , '40K – 60K'
* **Units of mobiles sold in difference price range is represented in the below bar graph**



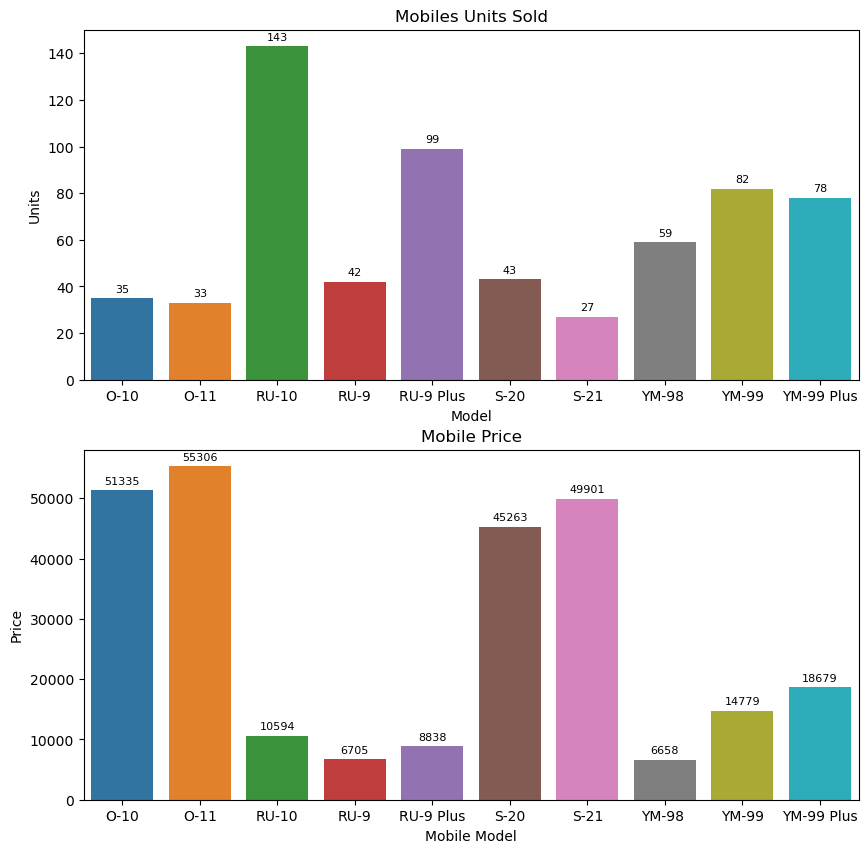
* Mobiles below 15K clearly dominate the sales
* However the super market also sells a very good amount of Upper Price range models (40K – 60K).
* The mid range models are purchased very less.
* This is because there is only one model available in mid range

****

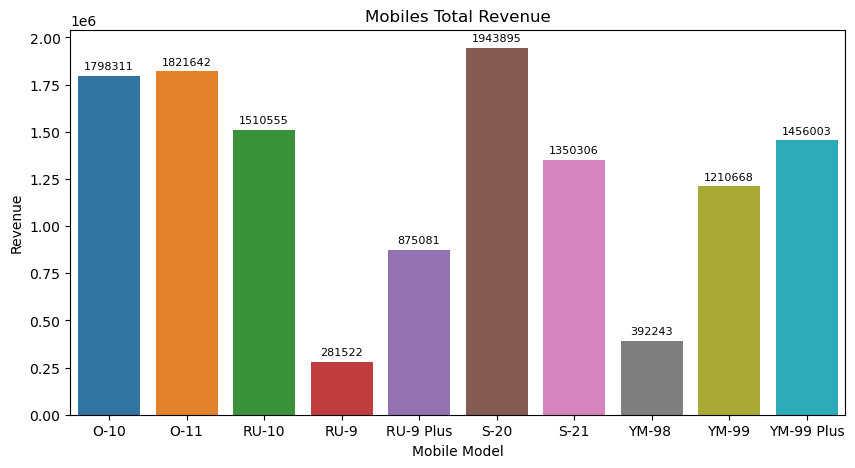
* **Revenue in different price range**



* The units sold of budget models (below 15K) were 3 times the units sold in upper range(40K – 60K)
* But, these many units of upper range were enough to make 1.5 times the revenue of the budget models
* This is because the budget models are very low price models. Out of 5 budget models 3 are below even 10K
* Considering there is only one model in mid range, the revenue of mid range is good.
* **The price of mobiles and their units sold is represented in the below bar graph**



* We can see that the most purchased model is RU-10 which is of price 10.5K
* Top 4 most purchased models are RU-10, RU-9 Plus, YM-99, and YM-99 Plus
* **Revenue Generated by each Model is represented in the below graph**

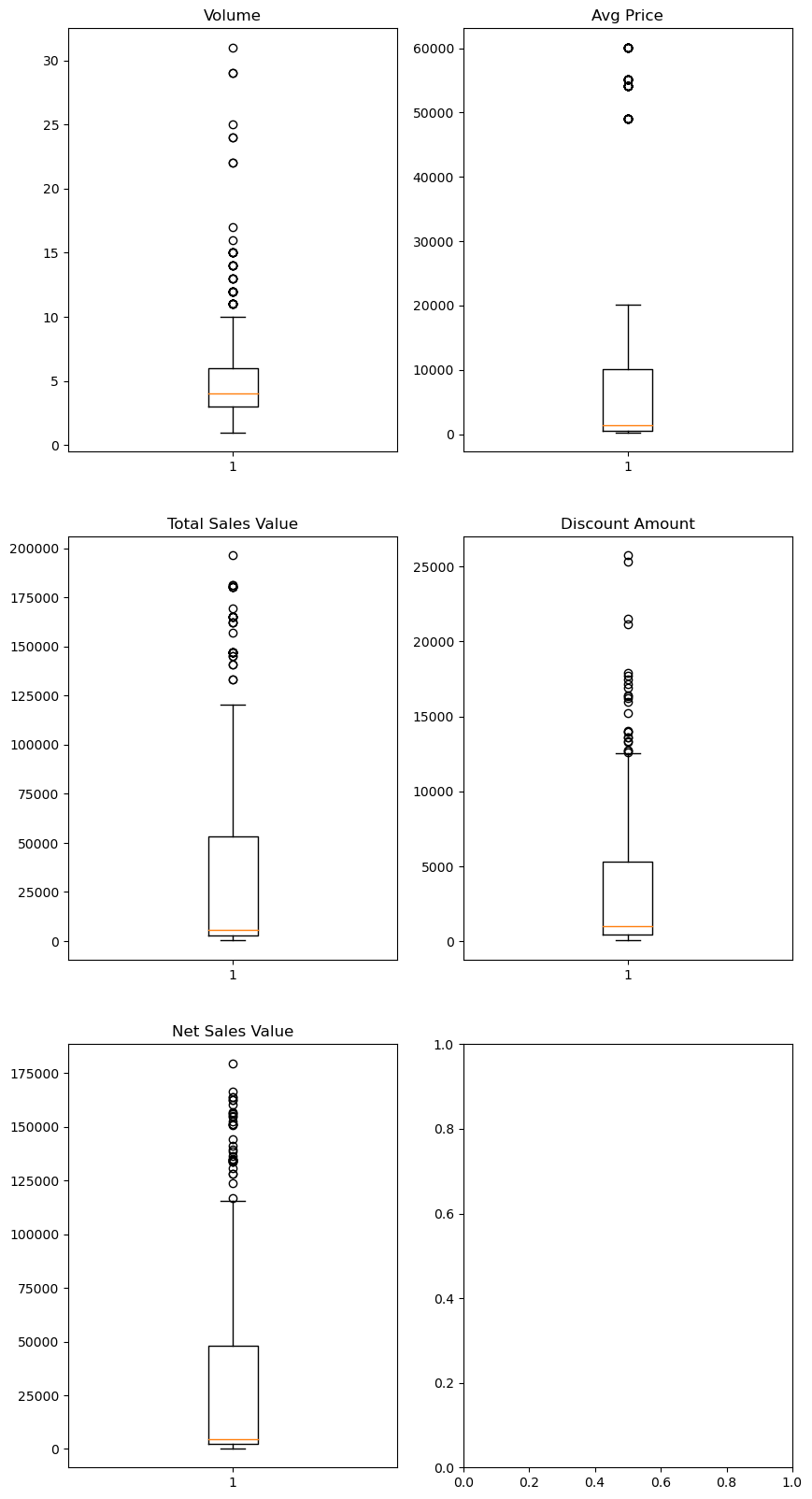


* Most Revenue is generated by S-20 : 19 Lac. 43K
* The top 3 most revenue generating models are of upper range (40K – 60K)

1. **BOX PLOT**

* Box & Whisker plot shows the statistical values such as Min Value, Max Value, Q1, Q2, Q3 and outliers of a numerical distribution
* Outliers of a distribution are defined as the values lying beyond 1.5 times of the IQR from the Q1 on the lower side and Q3 from the upper side
* Box & Whisker plot of numerical columns

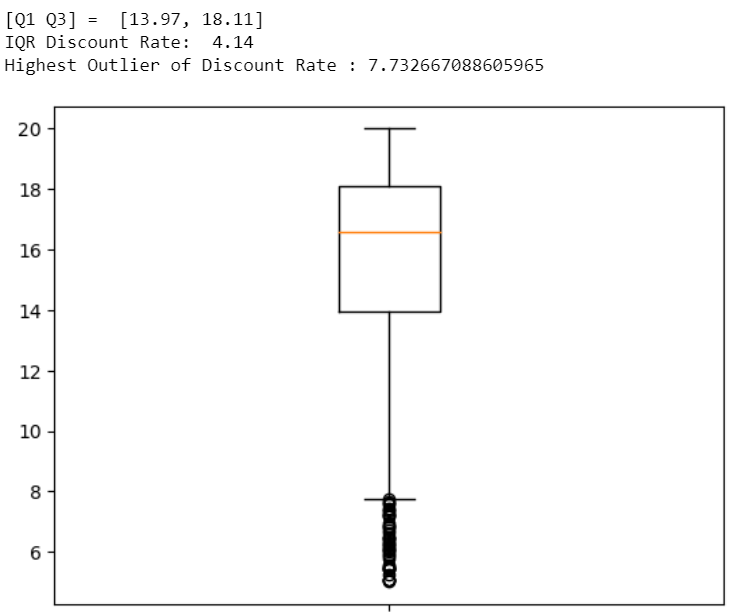




* IQR & Lowest Value of outlers (since there are numerous outliers) for each numerical column



* Box plot of Discount Rate



1. **STANDARDIZATION**

* Standardization is the process of transforming our data into a standard scale, which is of much smaller range. The standardized data still holds the original pattern of variabilty and proportion
* It is also called z-score normalization because the operation that we do on the data values is similar to calculating the z-score for any value in a normal distribution
* To get the new value, we subtract the mean of the data from the original value in the distribution and then divide the obtained value by the standard deviation value of the distribution.
* Its formula can be written as x’ = (x-u)/p

where

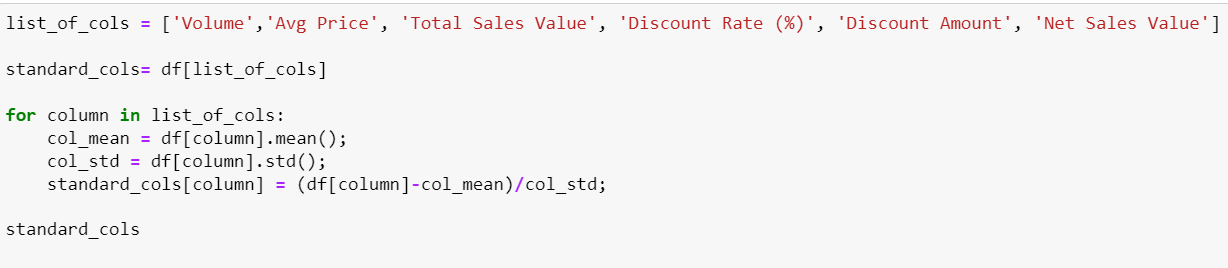
x’: new value

x: original value

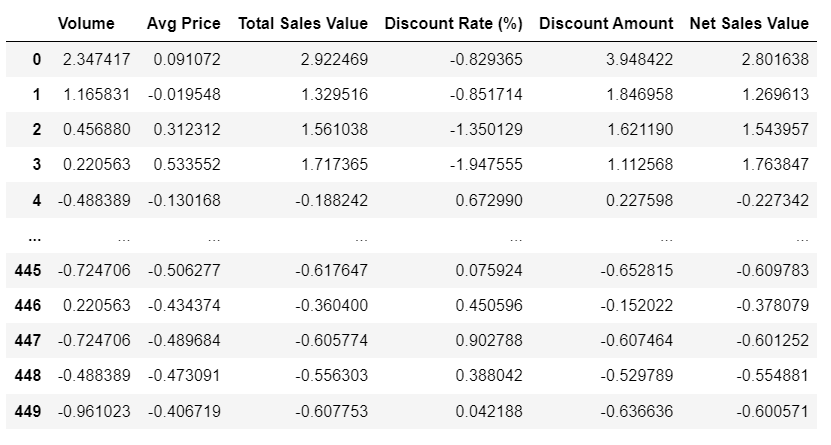
u: mean of data

p: standard deviation of data

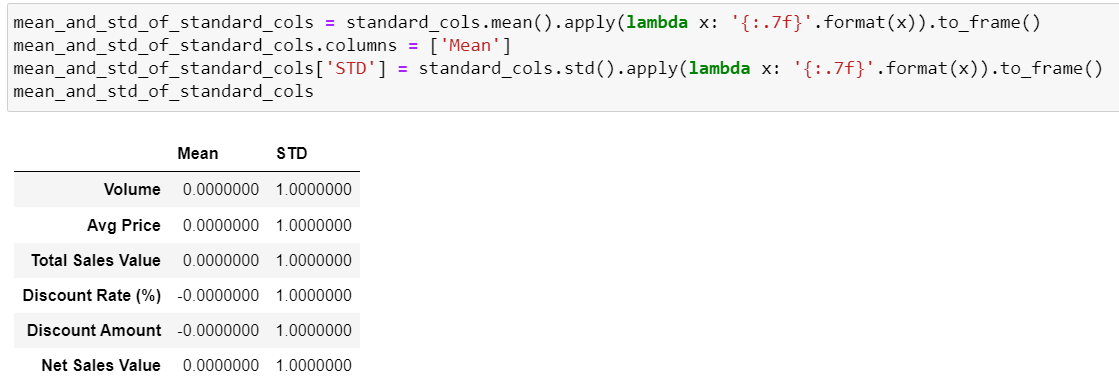
* After standardization the new data will have mean 0 and std 1.
* **Standardizing the numerical columns of the dataset**

****

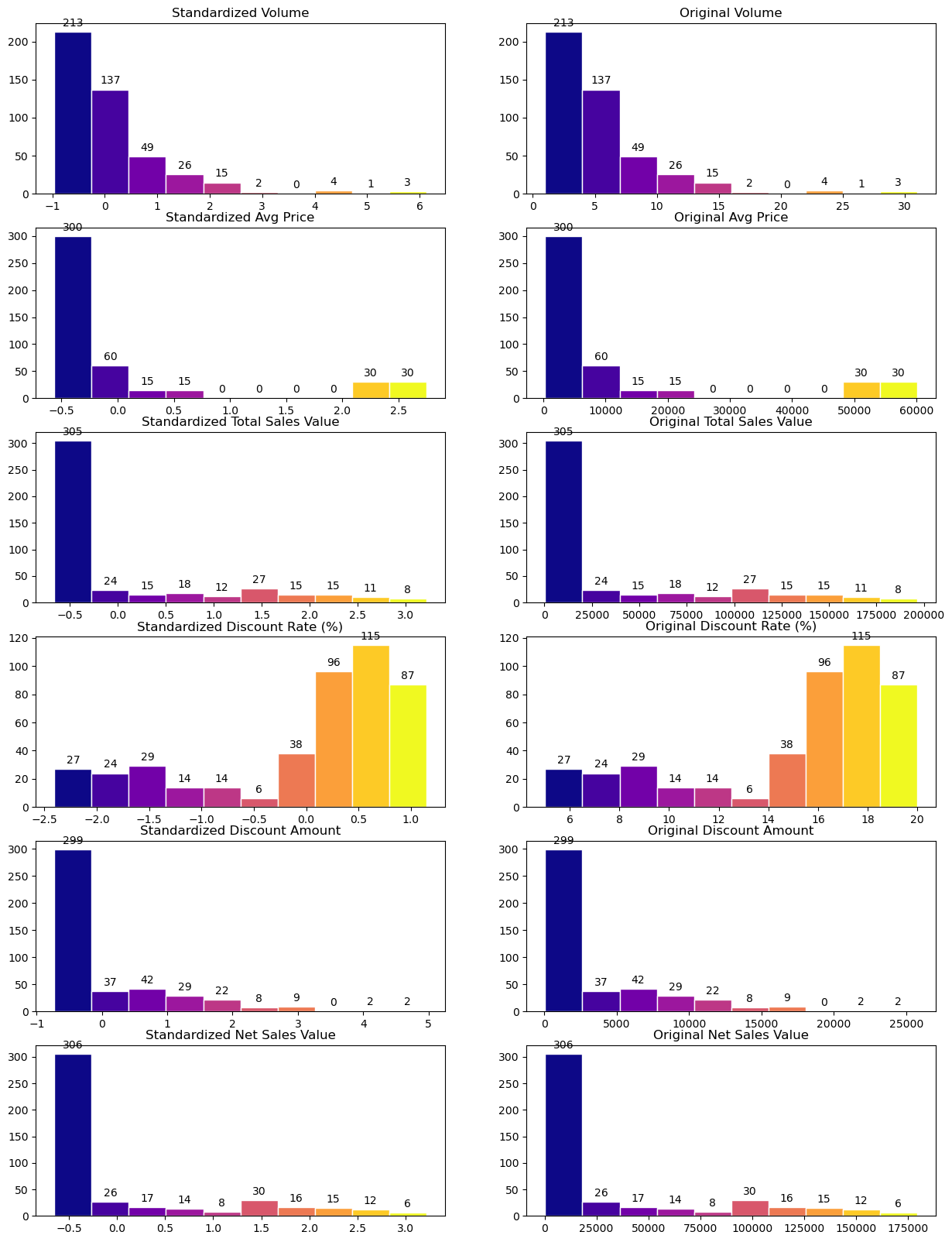
* **Standardized Values**

****

* **Mean & STD of standardized columns**

****

* **Histogram Comparison : Standardized Columns v/s Original Columns**



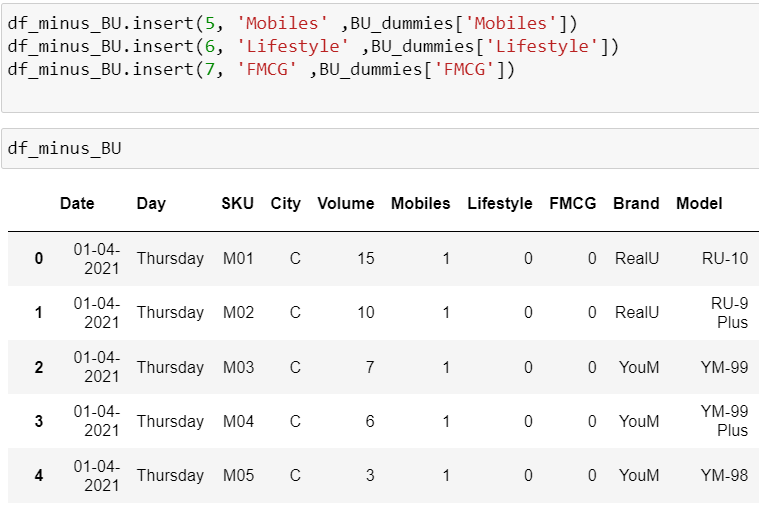
* We can see that the standardization does not changes the distribution of data. It only reduces the scale of values hence it is one of the Scaling Methods

1. **ONE HOT ENCODING**

* All machine learning algorithms work on numerical data. Categorical data must be transformed into numbers.
* Adding dummy variables for each categorical value gives us data into numbers (0,1) in each variable.
* ML algorithms can do computations on this data.
* One -hot encoding can be applied to only one column which is 'BU' as it has just 3 categorical values.
* Applying one-hot encoding to other categorical columns is not appropriate as they have too many values
* **Applying one hot encoding on BU column**



* **Adding the new columns to the original dataset**

****

1. **SUMMARY**

* **Key Findings**
* The Highest Revenue for the Supermarket was on 1st April
* Highest Revenue was 10 Lac. 77 Thousands
* Discount given on Lifestyle & FMCG was on an average arounf 17.5 % while for Mobiles it was around 11%
* 126 Lacs of revenue came from mobile sales whereas 11 Lacs from rest of the goods. Meaning the supermarket’s revenue solely comes from mobile sales
* The above insight suggests that sincethe super market is getting so many mobile customers it can expand its revenue by selling additional accessories for mobiles, adding repair shops, and other communications electronics like PCs and its accessories.
* **Importance of Data Preprocessing**
* Data preprocessing is very important part of Data Analytics because we normally get a data which cannot be given directly to ML algorithms.
* The datasets in their original form do not give any previously unknown information .
* Often the values in the datasets are missing and not the exact type which they should be. Mostly the numerical values are also in object type.
* The datatype must be corrected to the most appropriate type so that we can perform calculations using them for our insights.
* The data may contain some error values. In data there may be certain values which are one of their kind such values are called outliers.
* These error values and outliers must be identifies and treated so that they do not show a different picture than what the majority data actually shows.
* They need to be processed in order to get insights from them.
* The data transformation techniques like standardization and one-hot encoding are very important steps of preprocessing.
* Because the ML algorithms cannot performs operations on string data so they must be converted to numericals.
* Standardized data holds the exact distribution by reducing the scale, this is very important for keeping the equal importance of all features in the dataset for the training and predictions. Otherwise the feature with large scaled data dominates the training and model is trained with amplified importance for that feature and other features account for very less impact during training.