9951666670
INNOMATICS
RESEARCH LABS
DATA SCIENCE | MACHINE LEARNING | ARTIFICIAL INTELLIGENCE
BIG DATA | DIGITAL MARKETING | AWS | DEVOPS
TOSHIBA

INNOMATICS
RESEARCH LABS

WEB SCRAPING

Washing Machine: Product Category, Price Analysis and Insights

Batch No: 154

Prepared by :

Ramesh Isukapalli, Shrikant Telang

Guided By:

Ram sir, Saksham sir

INNOMATICS RESEARCH LABS

CONTENTS:

- Problem Statement
- Web Scraping Definition
- Libraries Used
- Data Frame before cleaning
- Data Cleaning and Manipulation
- Data Frame After cleaning
- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis
- Summary of Insights

Understanding the Problem

Exploring a product from An online E-commerce product catalog

Applying desirable filters to choose best suitable product

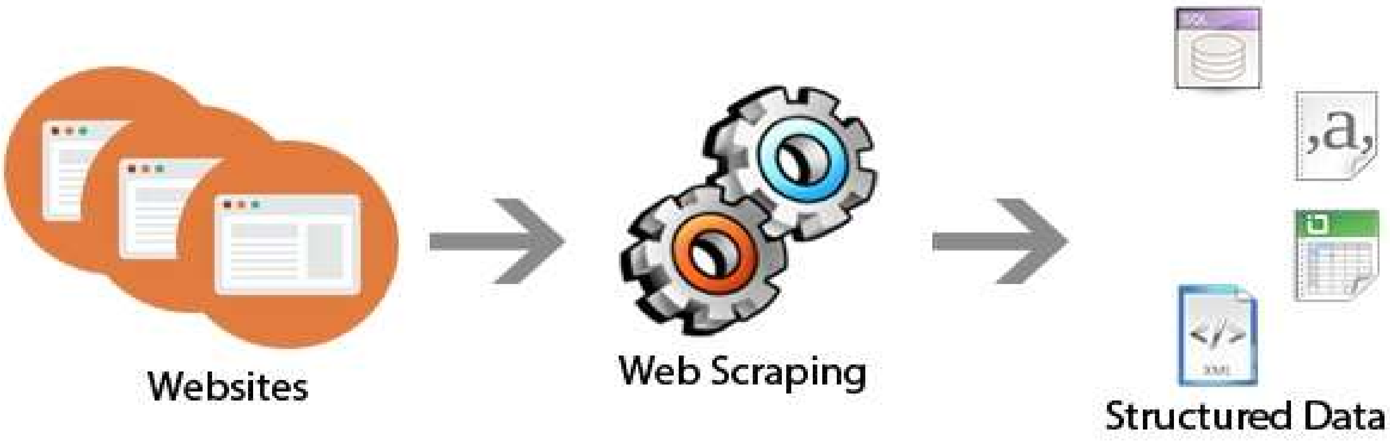
Select the best suitable product

Problem Statement:

Suggest a Washing Machine with best-in-class features out of hundreds of catalog items from an online e-commerce website according to consumer’s budget

INNOMATICS RESEARCH LABS

WHAT IS WEB SCRAPING?



Web scraping is a term used to describe the use of a program or algorithm to extract and process large amounts of unstructured data from the web and exporting into a useful format.

Different Libraries used for Data Scraping:

- Requests Library for Web Scraping
  - This library used for making various types of HTTP requests like **Get, Post** etc., to retrieve contents. it helps to access website HTML contents.

- BeautifulSoup Library for Web Scraping (bs4)
  - This library perhaps the most widely used Python library for web scraping
  - let's say you receive your data in raw HTML, this library will take the said HTML and transform it into a more readable data format that can be easily read and understood.

The combination of Beautiful Soup and Requests is quite common in the Web Scraping.

WEBSITE USED FOR SCRAPING THE DATA:

Flipkart

url:

https://www.flipkart.com/

LIBRARIES USED FOR DATA ANALYSIS

Pandas

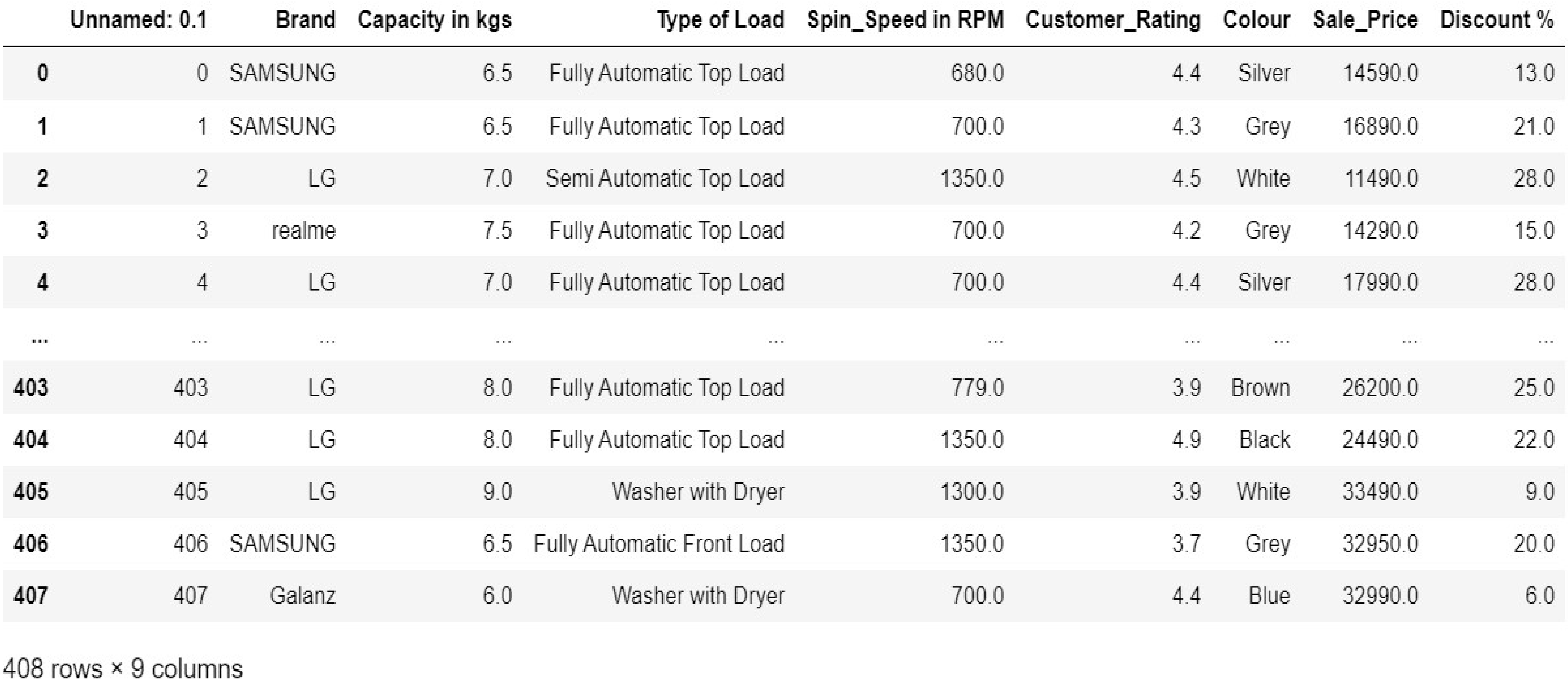
Seaborn

Matplotlib

re(RegEx)

Plotly

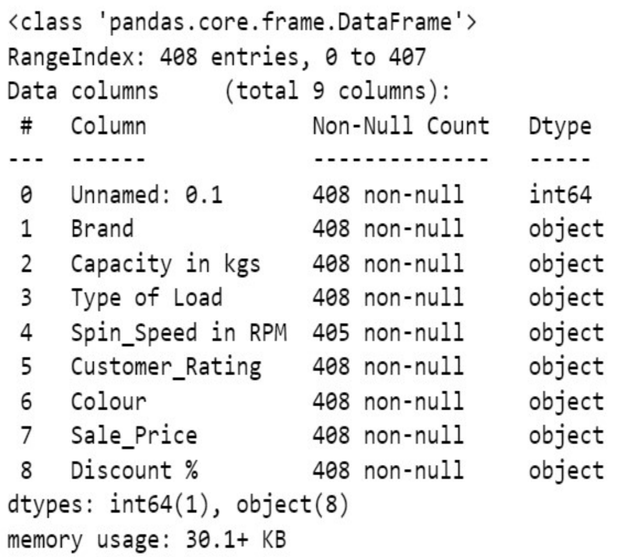
DATA FRAME BEFORE CLEANING :



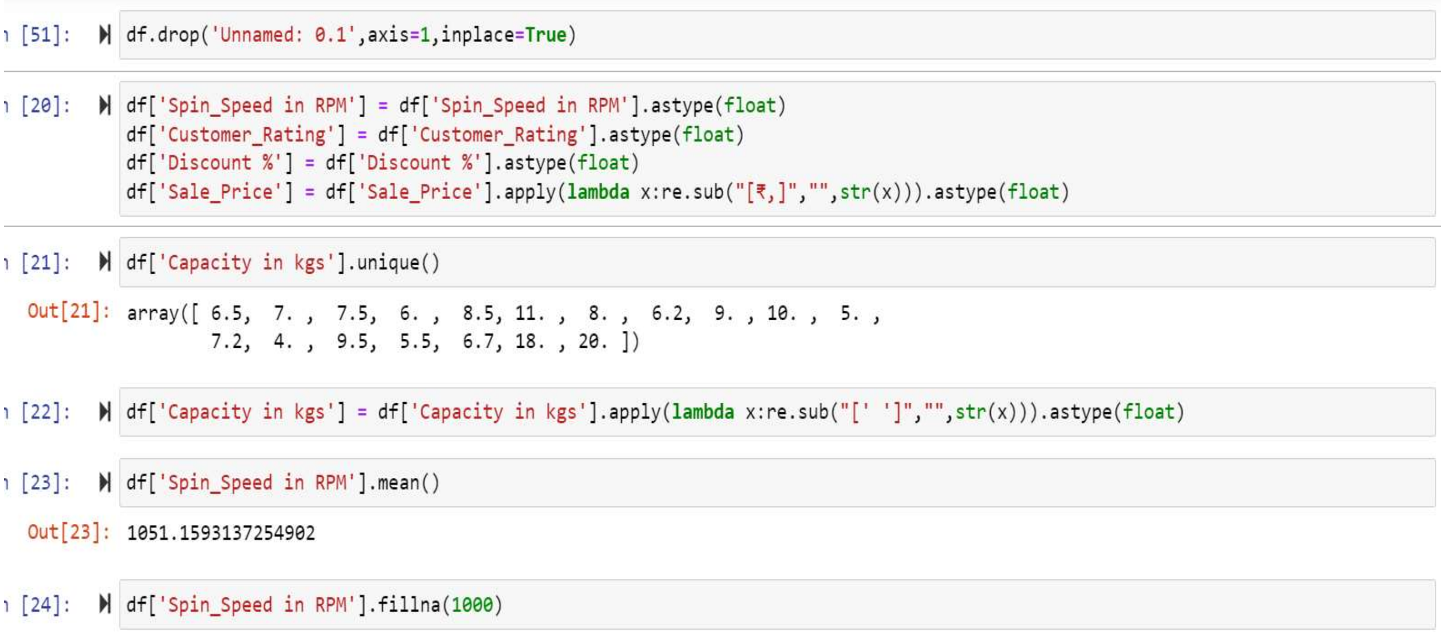
	Unnamed: 0.1	Brand	Capacity in kgs	Type of Load	Spin_Speed in RPM	Customer_Rating	Colour	Sale_Price	Discount %
0	0	SAMSUNG	6.5	Fully Automatic Top Load	680.0	4.4	Silver	14590.0	13.0
1	1	SAMSUNG	6.5	Fully Automatic Top Load	700.0	4.3	Grey	16890.0	21.0
2	2	LG	7.0	Semi Automatic Top Load	1350.0	4.5	White	11490.0	28.0
3	3	realme	7.5	Fully Automatic Top Load	700.0	4.2	Grey	14290.0	15.0
4	4	LG	7.0	Fully Automatic Top Load	700.0	4.4	Silver	17990.0	28.0
...	...	...	...	...	...	...	...	...	...
403	403	LG	8.0	Fully Automatic Top Load	779.0	3.9	Brown	26200.0	25.0
404	404	LG	8.0	Fully Automatic Top Load	1350.0	4.9	Black	24490.0	22.0
405	405	LG	9.0	Washer with Dryer	1300.0	3.9	White	33490.0	9.0
406	406	SAMSUNG	6.5	Fully Automatic Front Load	1350.0	3.7	Grey	32950.0	20.0
407	407	Galanz	6.0	Washer with Dryer	700.0	4.4	Blue	32990.0	6.0

408 rows × 9 columns

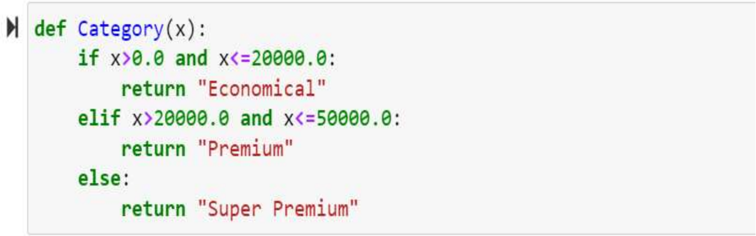
DATA CLEANING AND MANIPULATION



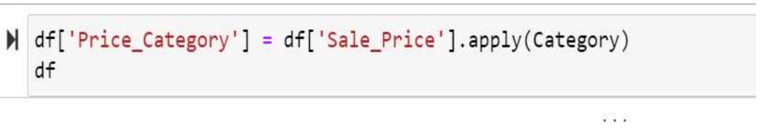
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 408 entries, 0 to 407
Data columns     (total 9 columns):
 #   Column              Non-Null Count    Dtype
---  ------              --------------    -----
 0   Unnamed: 0.1        408 non-null      int64
 1   Brand               408 non-null      object
 2   Capacity in kgs     408 non-null      object
 3   Type of Load        408 non-null      object
 4   Spin_Speed in RPM   405 non-null      object
 5   Customer_Rating     408 non-null      object
 6   Colour              408 non-null      object
 7   Sale_Price          408 non-null      object
 8   Discount %          408 non-null      object
dtypes: int64(1), object(8)
memory usage: 30.1+ KB
```



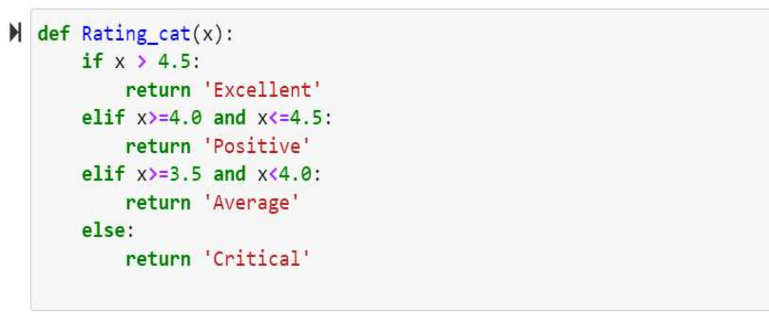
```
[51]:  df.drop('Unnamed: 0.1',axis=1,inplace=True)

[20]:  df['Spin_Speed in RPM'] = df['Spin_Speed in RPM'].astype(float)
       df['Customer_Rating'] = df['Customer_Rating'].astype(float)
       df['Discount %'] = df['Discount %'].astype(float)
       df['Sale_Price'] = df['Sale_Price'].apply(lambda x:re.sub("[₹,]","",str(x))).astype(float)

[21]:  df['Capacity in kgs'].unique()

Out[21]: array([ 6.5,  7. ,  7.5,  6. ,  8.5, 11. ,  8. ,  6.2,  9. , 10. ,  5. ,
                 7.2,  4. ,  9.5,  5.5,  6.7, 18. , 20. ])

[22]:  df['Capacity in kgs'] = df['Capacity in kgs'].apply(lambda x:re.sub("[' ']","",str(x))).astype(float)

[23]:  df['Spin_Speed in RPM'].mean()

Out[23]: 1051.1593137254902

[24]:  df['Spin_Speed in RPM'].fillna(1000)
```



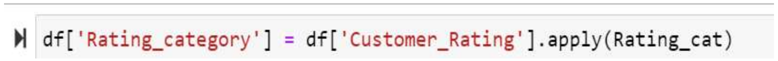
```
def Category(x):
    if x>0.0 and x<=20000.0:
        return "Economical"
    elif x>20000.0 and x<=50000.0:
        return "Premium"
    else:
        return "Super Premium"
```



```
df['Price_Category'] = df['Sale_Price'].apply(Category)
df
```

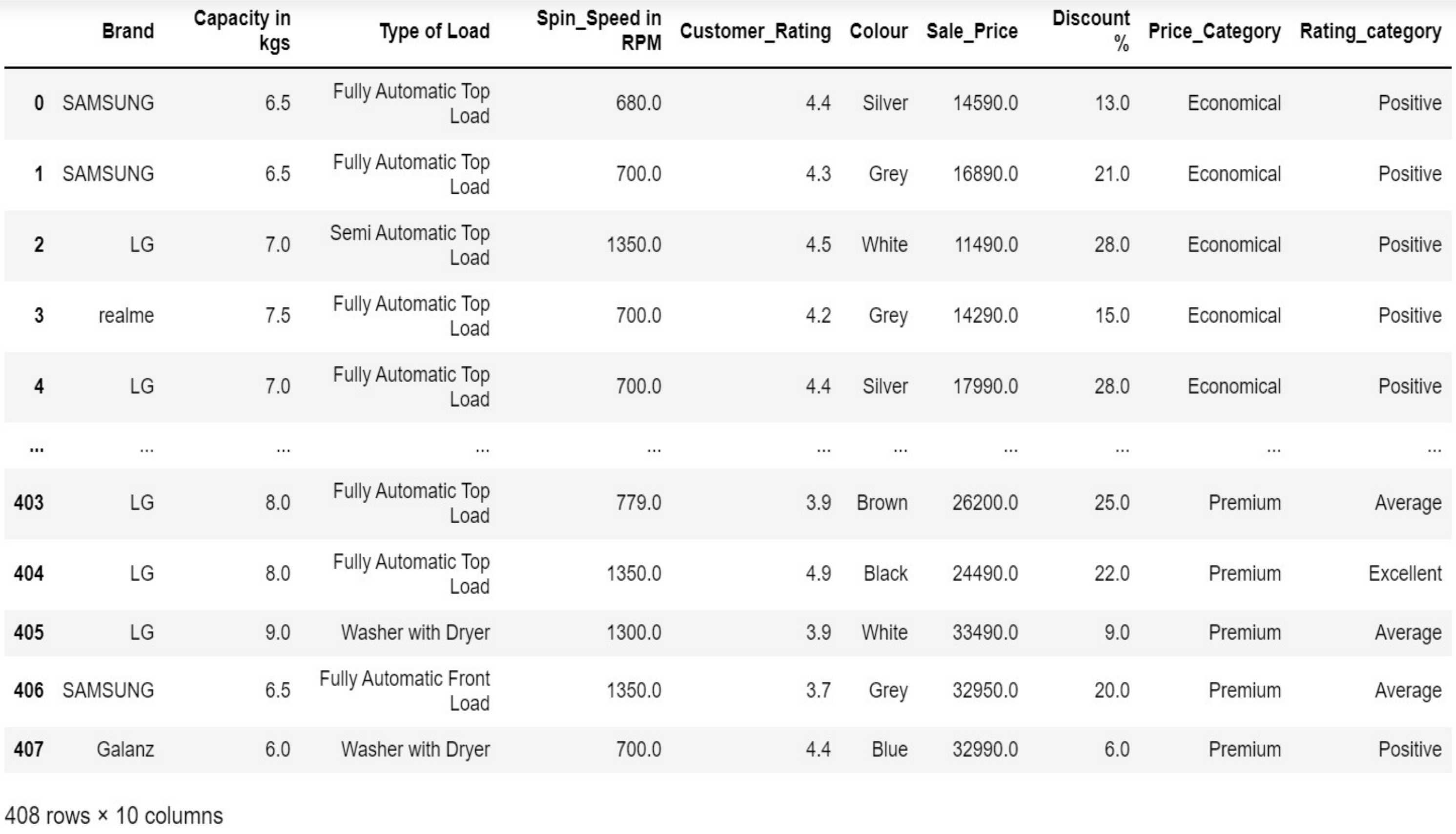


```
def Rating_cat(x):
    if x > 4.5:
        return 'Excellent'
    elif x>=4.0 and x<=4.5:
        return 'Positive'
    elif x>=3.5 and x<4.0:
        return 'Average'
    else:
        return 'Critical'
```



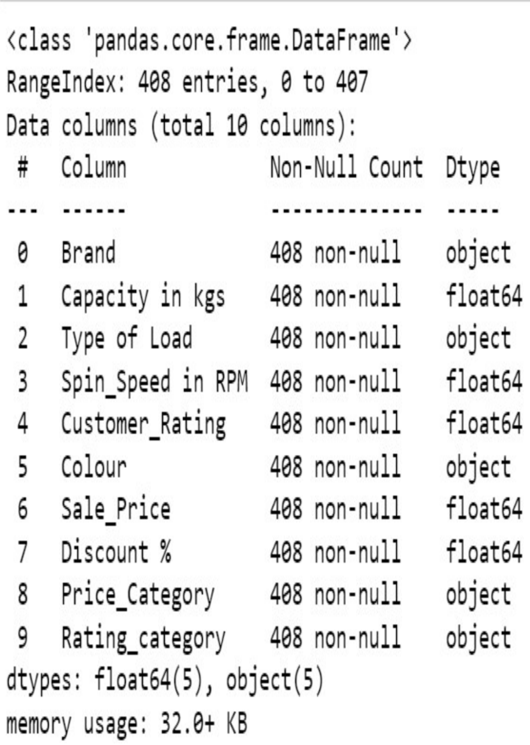
```
df['Rating_category'] = df['Customer_Rating'].apply(Rating_cat)
```

DATA FRAME AFTER CLEANING :



	Brand	Capacity in kgs	Type of Load	Spin_Speed in RPM	Customer_Rating	Colour	Sale_Price	Discount %	Price_Category	Rating_category
0	SAMSUNG	6.5	Fully Automatic Top Load	680.0	4.4	Silver	14590.0	13.0	Economical	Positive
1	SAMSUNG	6.5	Fully Automatic Top Load	700.0	4.3	Grey	16890.0	21.0	Economical	Positive
2	LG	7.0	Semi Automatic Top Load	1350.0	4.5	White	11490.0	28.0	Economical	Positive
3	realme	7.5	Fully Automatic Top Load	700.0	4.2	Grey	14290.0	15.0	Economical	Positive
4	LG	7.0	Fully Automatic Top Load	700.0	4.4	Silver	17990.0	28.0	Economical	Positive
...	...	...	...	...	...	...	...	...	...	...
403	LG	8.0	Fully Automatic Top Load	779.0	3.9	Brown	26200.0	25.0	Premium	Average
404	LG	8.0	Fully Automatic Top Load	1350.0	4.9	Black	24490.0	22.0	Premium	Excellent
405	LG	9.0	Washer with Dryer	1300.0	3.9	White	33490.0	9.0	Premium	Average
406	SAMSUNG	6.5	Fully Automatic Front Load	1350.0	3.7	Grey	32950.0	20.0	Premium	Average
407	Galanz	6.0	Washer with Dryer	700.0	4.4	Blue	32990.0	6.0	Premium	Positive

408 rows × 10 columns



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 408 entries, 0 to 407
Data columns (total 10 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Brand              408 non-null    object 
 1   Capacity in kgs    408 non-null    float64
 2   Type of Load       408 non-null    object 
 3   Spin_Speed in RPM  408 non-null    float64
 4   Customer_Rating    408 non-null    float64
 5   Colour             408 non-null    object 
 6   Sale_Price         408 non-null    float64
 7   Discount %         408 non-null    float64
 8   Price_Category     408 non-null    object 
 9   Rating_category    408 non-null    object 
dtypes: float64(5), object(5)
memory usage: 32.0+ KB
```

UNIVARIATE ANALYSIS:

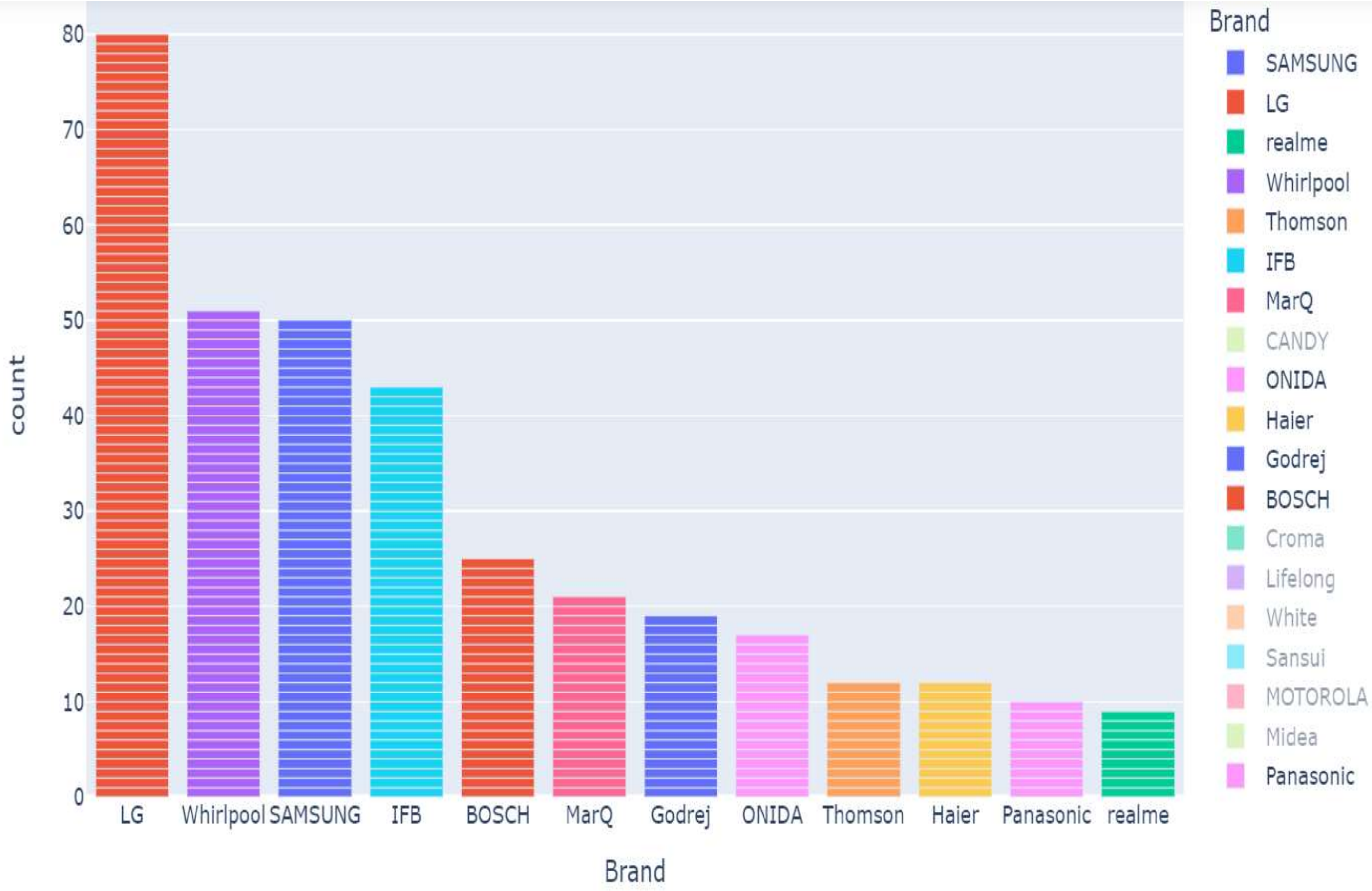
The term univariate analysis refers to the analysis of one variable. You can remember this because the prefix "uni" means "one."

There are three common ways to perform univariate analysis on one variable:

1. **Frequency table** – Describes how often different values occur.
2. **Charts** – Used to visualize the distribution of values.
3. **Summary statistics** – Measures the center and spread of values.

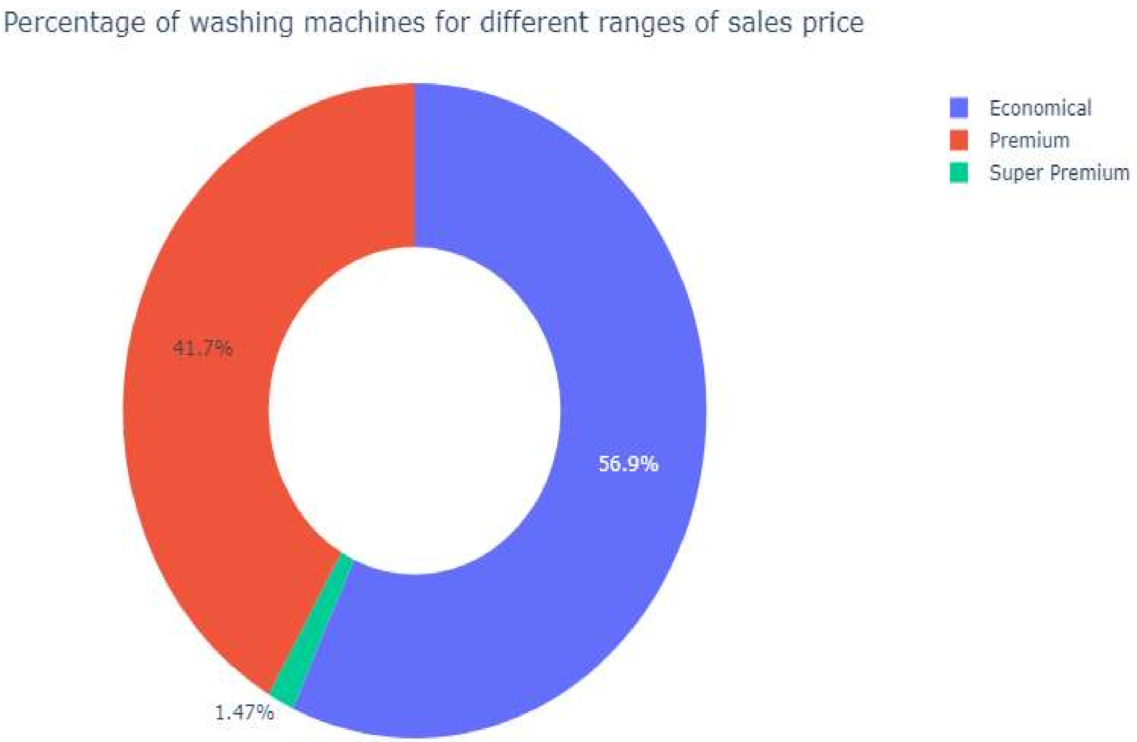
- For Numerical variable: Histogram, Boxplot, violin plot etc.,
- For Categorical Variable: Count plot, Pie chart etc.,

WASHING MACHINE BRANDS LEADING BY INVENTORY



- There are 26 companies selling washing machines on the Market based on our data.
- LG leads the catalog with count 80.
- Whirlpool, SamSung & IFB are trailing behind with count 51, 50 and 42 respectively.

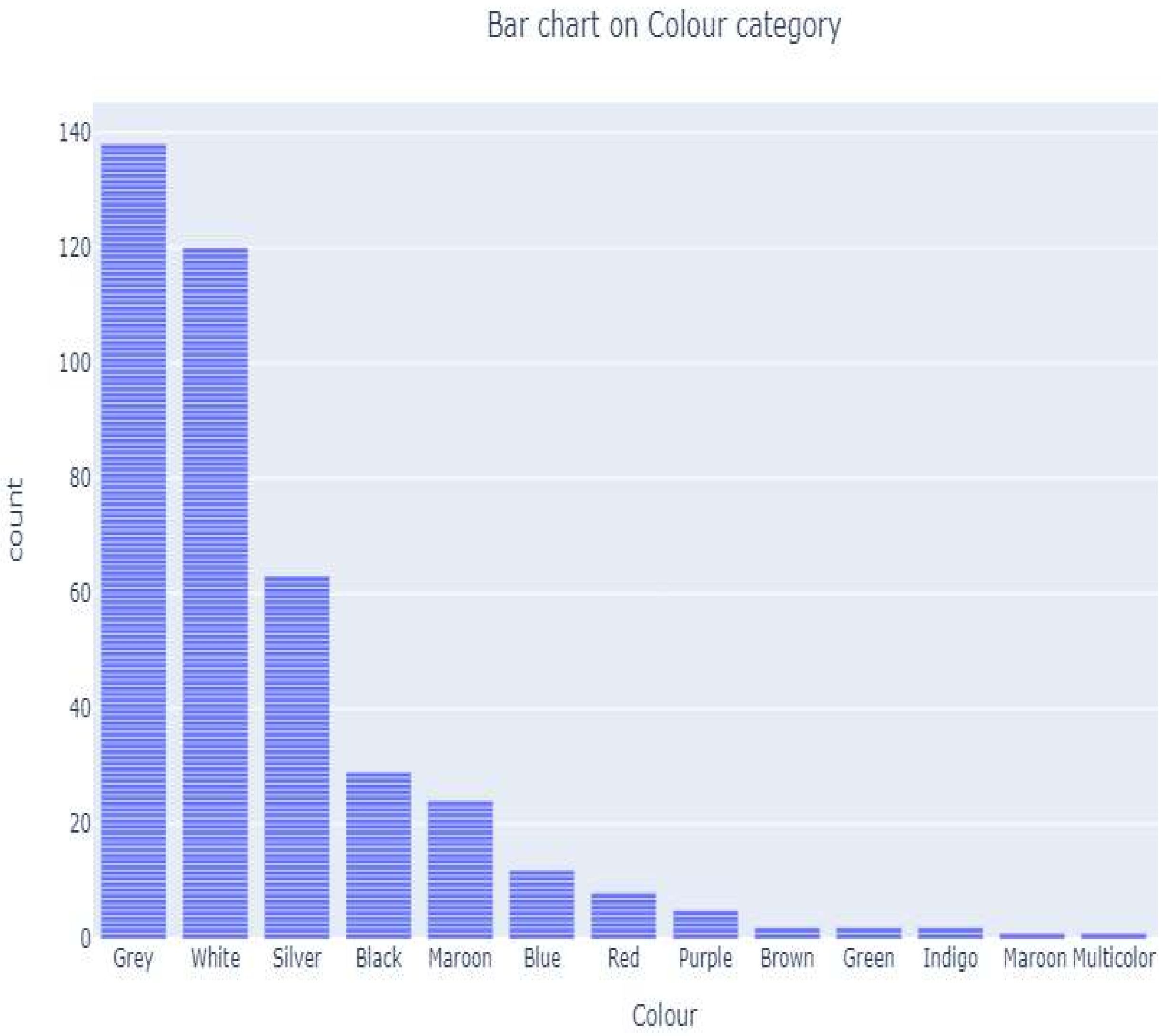
CATEGORY WISE DISTRIBUTION ACROSS DIFFERENT PRICE SEGMENTS



- Out of a total of 408 washing machines, 56.9% machines fall in economy category
- 41.7% washing machines are falling in the premium category.
- super-premium washing machine category occupying just 1.47%.

➢ Count of Washing Machines

❑ Economical	232
❑ Premium	170
❑ Super Premium	6

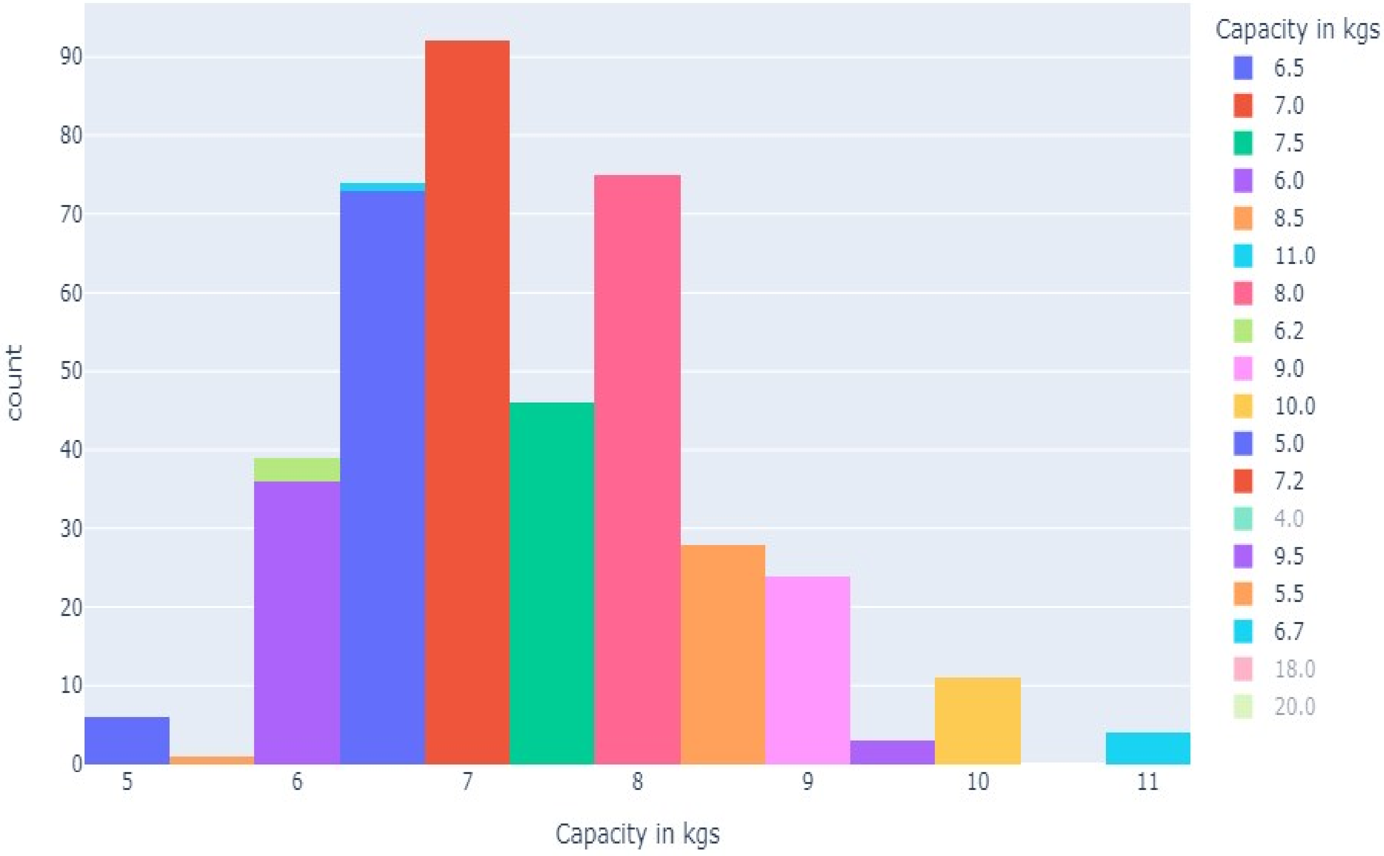


WASHING MACHINES LEADING BY COLOUR IN THE INVENTORY

We can observe from the above chart that three colours Grey(138), White(120) and Silver(63) occuppies the first three places followed by Black(29) and Maroon(24)

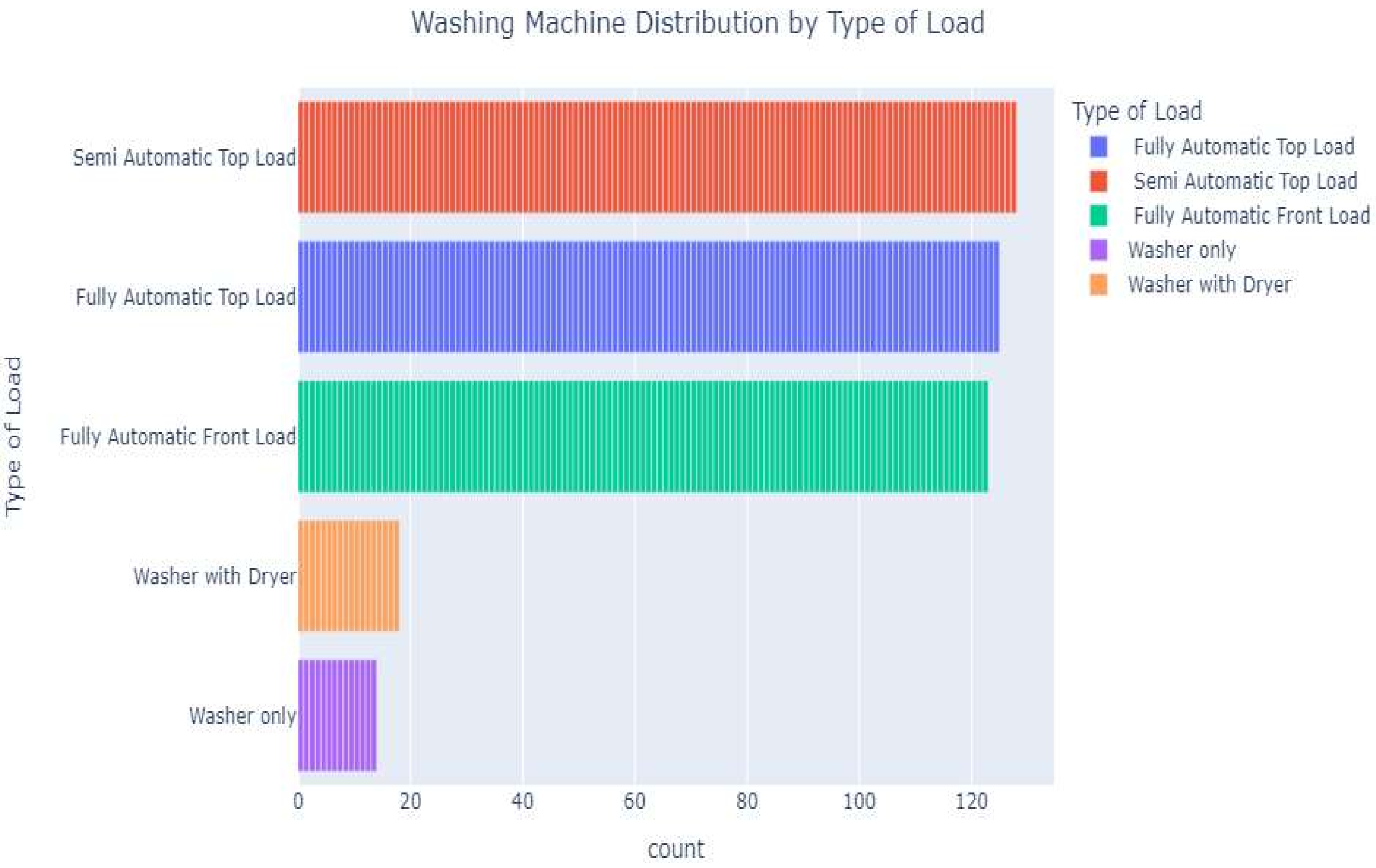
INNOMATICS RESEARCH LABS

WASHING MACHINES WITH DIFFERENT LOAD CAPACITY :



- ❑ Maximum number of Washing machines lies in range of capacity 6kg to 8kg.
- ❑ There are nearly above 90 machines with load capacity 7kg

INNOMATICS
RESEARCH LABS



WASHING MACHINES WITH DIFFERENT TYPE OF LOAD:

- ❑ **3 major categories:**

semi-automatic top load 128
fully automatic top load 125
fully automatic front load 122

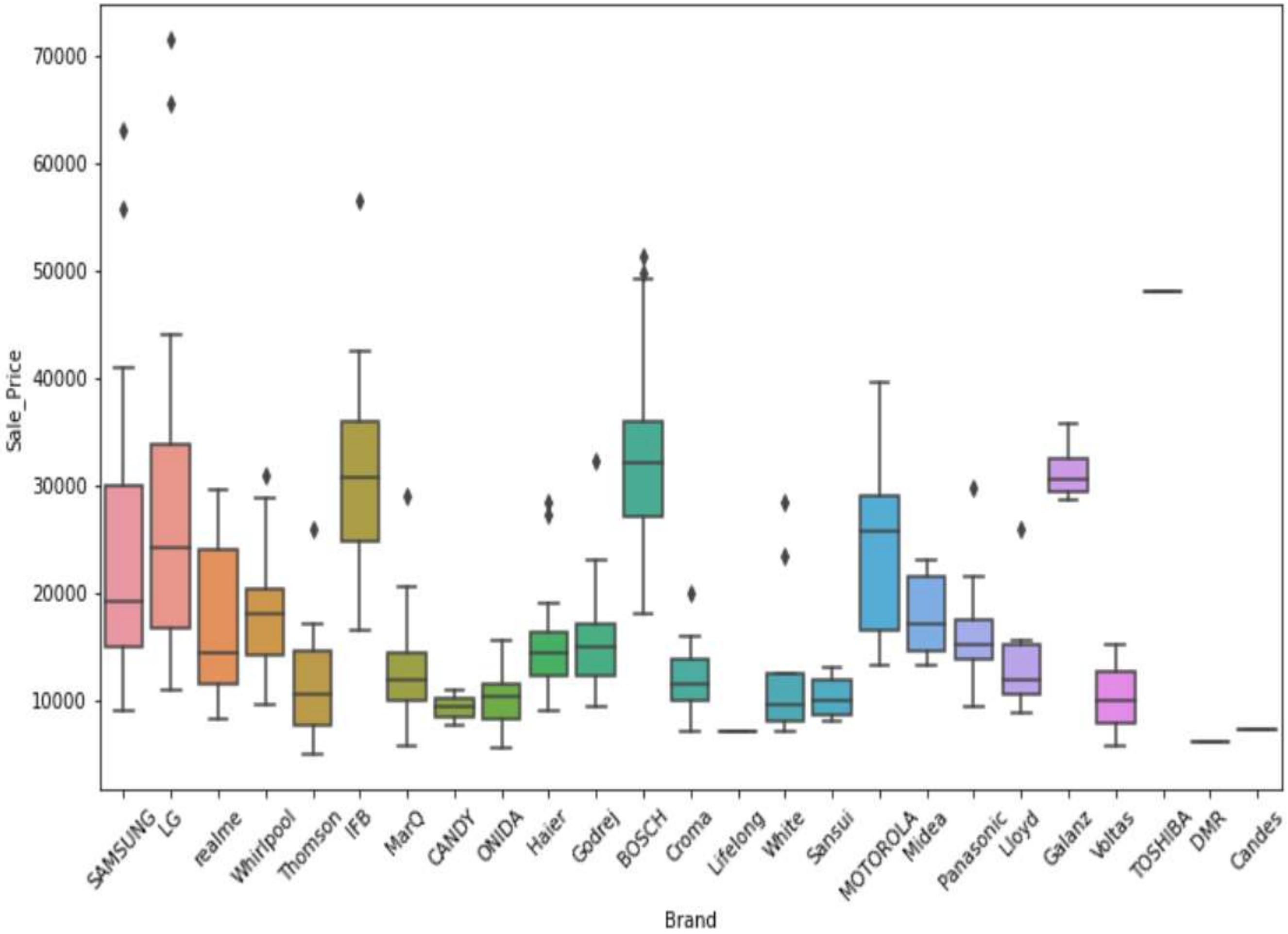
INNOMATICS
RESEARCH LABS

BIVARIATE ANALYSIS

It is performed to find the relationship between each variable in the dataset and the target variable of interest (or) using 2 variables and finding the relationship between them.

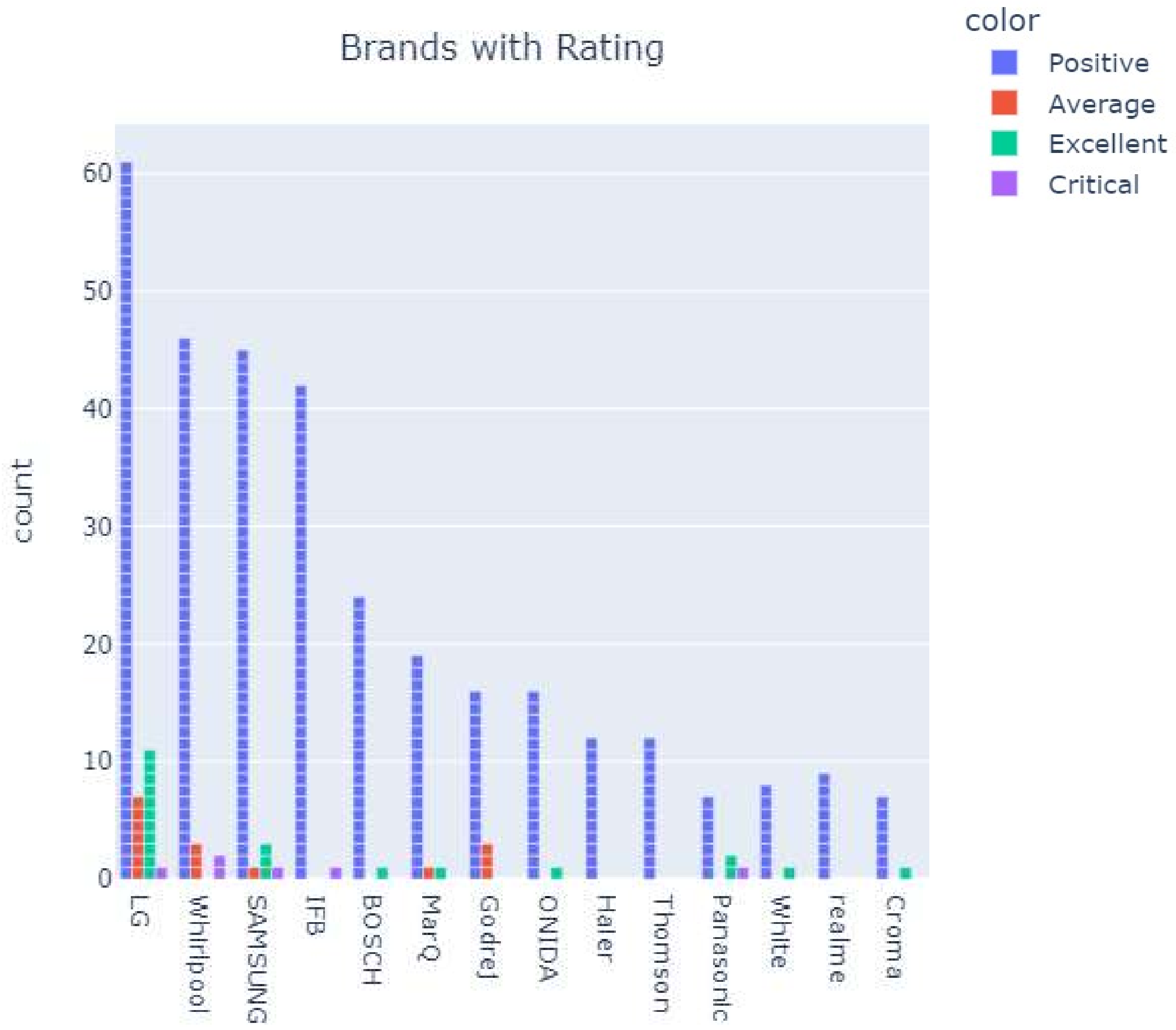
- ➢ Numerical vs Numerical: Scatter plot, Relational Plot, Regression Plot etc.,
- ➢ Numerical vs Categorical: Bar plot, line plot, Box plot, Violin plot etc.,

BOX PLOT BETWEEN BRAND AND SALE PRICE



Range of Sale price of Washing machines per each brand and also some outliers

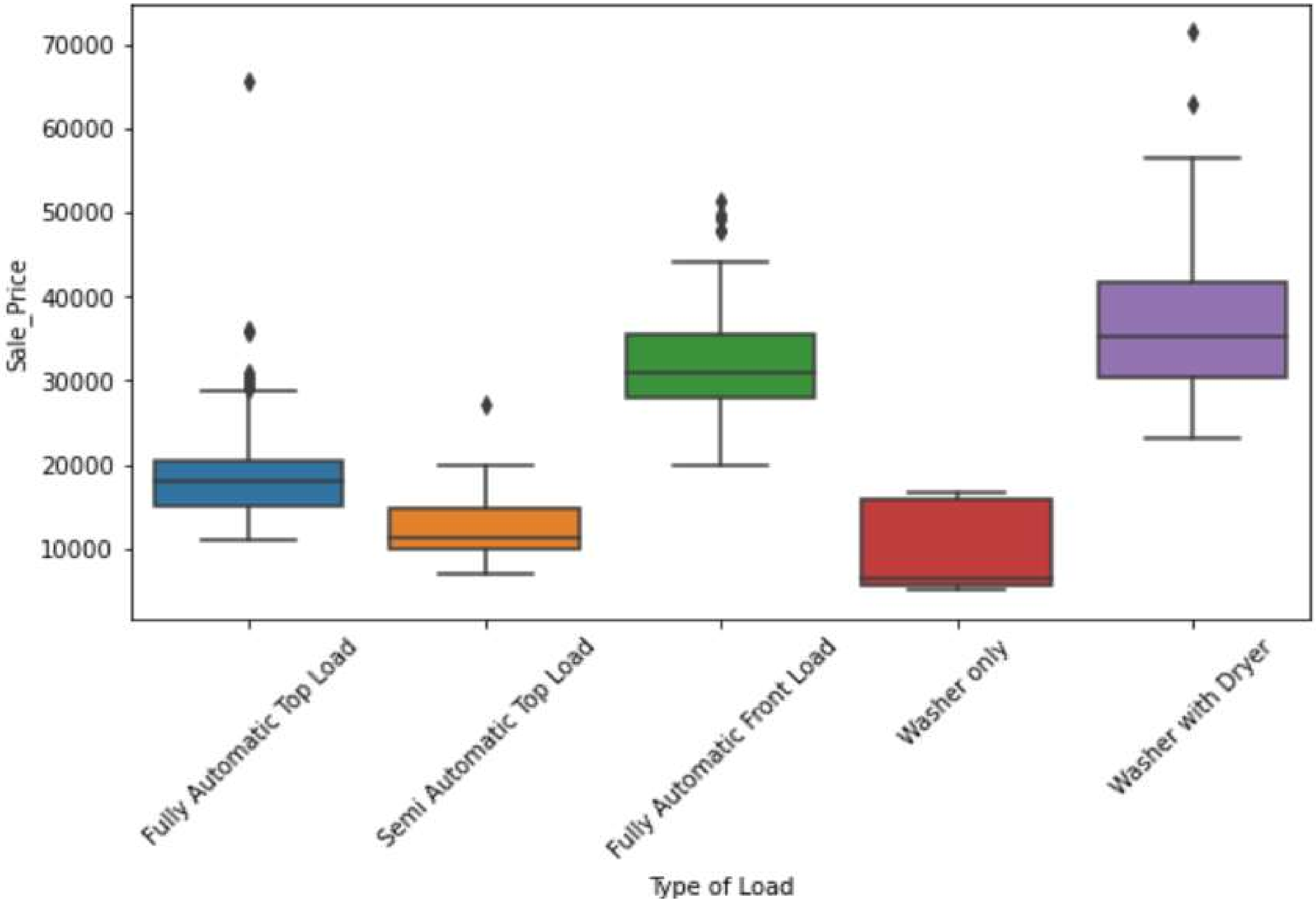
- ✓ SAMSUNG, LG, realme, Whirlpool & MOTOROLA have the max. price range .
- ✓ BOSCH & IFB is having the highest median sale price.
- ✓ Lifelong , Toshiba , DMR and Candes are having less number of products. Thus, less price range
- ✓ SAMSUNG, LG, IFB & BOSCH are having outliers which falls in super premium price range.



OBSERVATIONS:

- ✓ Most of the Washing machines are having Positive rating.
- ✓ LG is having more Excellent Ratings. SAMSUNG & Panasonic are also having some excellent ratings
- ✓ There are some washing machines like LG, Whirlpool, Samsung, Marq and Godrej are having Average ratings

INNOMATICS RESEARCH LABS

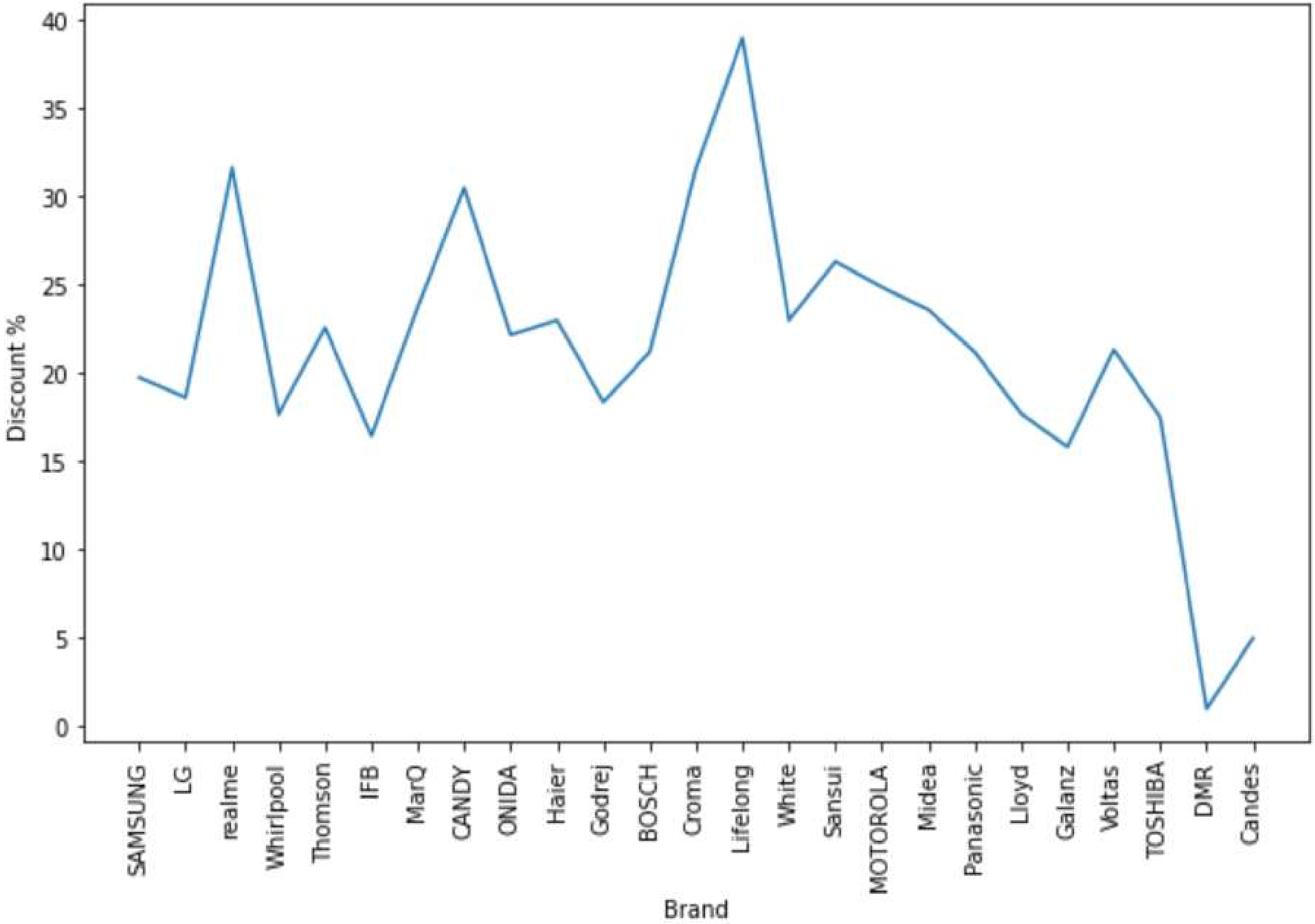


OBSERVATION:

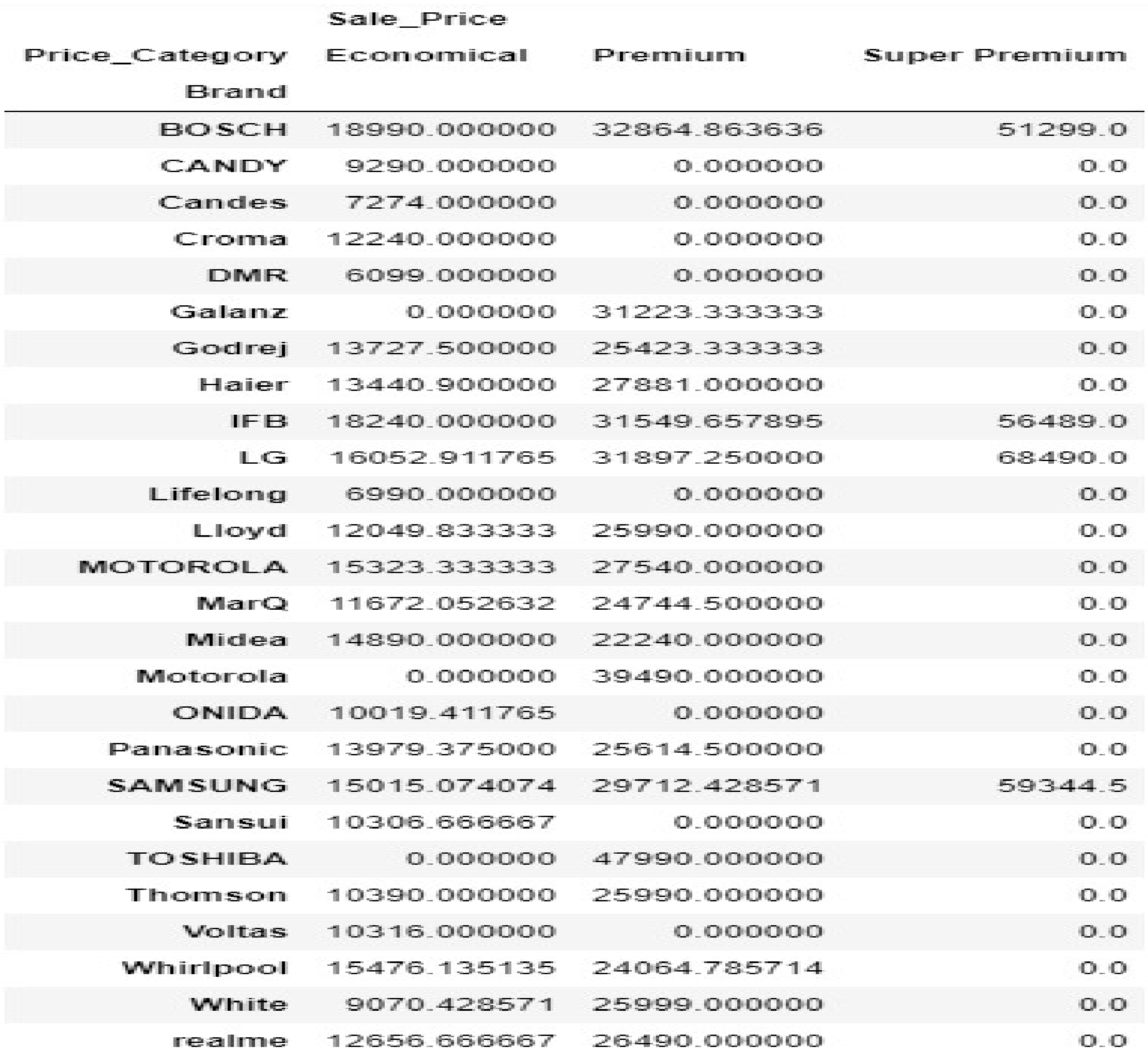
- ✓ Sales price of Washer with dryer is having the highest sale price range ranging from around 25,000 to 71,000
- ✓ Fully automatic front load is the washing machine with sale price range between 20,000 to 50,000
- ✓ Washer only & Semi Automatic Top Load probably having price range under 20,0000
- ✓ Fully Automatic Top Load is having the price range around 11,000 to 35,000 except one machine.

INNOMATICS RESEARCH LABS

LINE PLOT BETWEEN BRAND AND DISCOUNT



- ❑ Brand 'Lifelong' is giving the highest discount around 38%.
- ❑ 'DMR' is giving the lowest discount around 2% only.
- ❑ 'realme' & 'CANDY' are giving the discount around 30%.
- ❑ Rest of the brands are giving nearly around 18% to 25%.

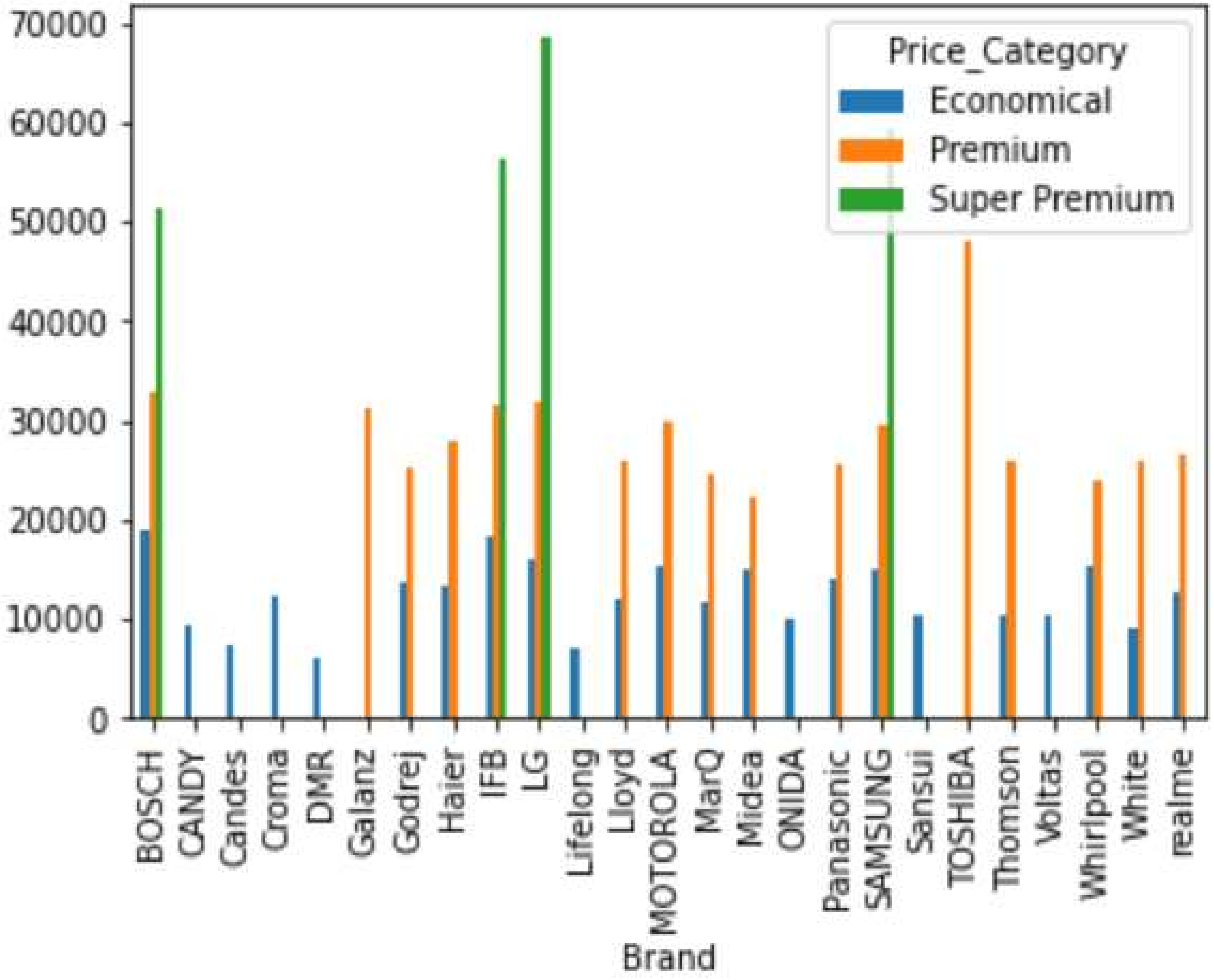


	Sale_Price		
Price_Category	Economical	Premium	Super Premium
Brand			
BOSCH	18990.000000	32864.863636	51299.0
CANDY	9290.000000	0.000000	0.0
Candes	7274.000000	0.000000	0.0
Croma	12240.000000	0.000000	0.0
DMR	6099.000000	0.000000	0.0
Galanz	0.000000	31223.333333	0.0
Godrej	13727.500000	25423.333333	0.0
Haier	13440.900000	27881.000000	0.0
IFB	18240.000000	31549.657895	56489.0
LG	16052.911765	31897.250000	68490.0
Lifelong	6990.000000	0.000000	0.0
Lloyd	12049.833333	25990.000000	0.0
MOTOROLA	15323.333333	27540.000000	0.0
MarQ	11672.052632	24744.500000	0.0
Midea	14890.000000	22240.000000	0.0
Motorola	0.000000	39490.000000	0.0
ONIDA	10019.411765	0.000000	0.0
Panasonic	13979.375000	25614.500000	0.0
SAMSUNG	15015.074074	29712.428571	59344.5
Sansui	10306.666667	0.000000	0.0
TOSHIBA	0.000000	47990.000000	0.0
Thomson	10390.000000	25990.000000	0.0
Voltas	10316.000000	0.000000	0.0
Whirlpool	15476.135135	24064.785714	0.0
White	9070.428571	25999.000000	0.0
realme	12656.666667	26490.000000	0.0

MULTIVARIATE ANALYSIS

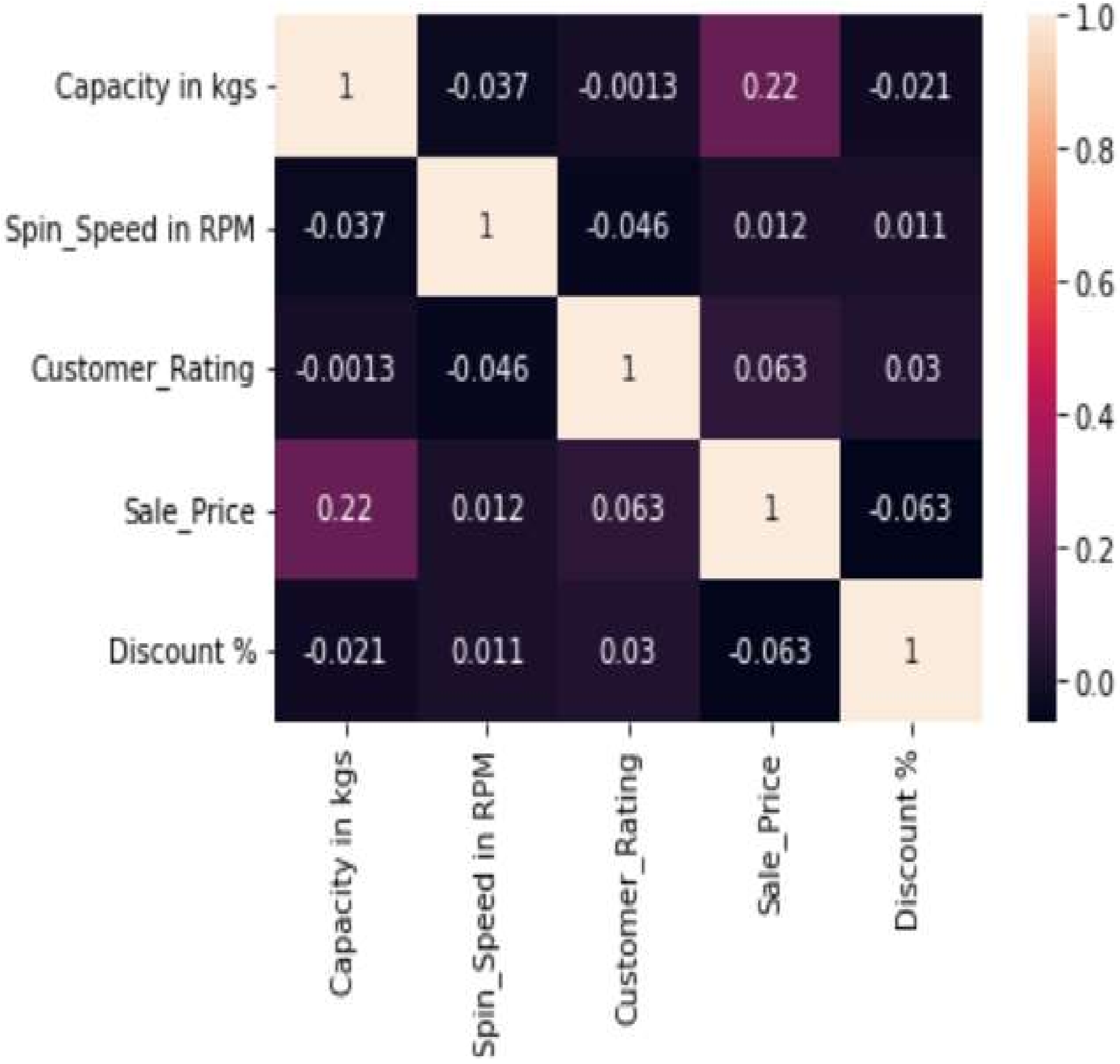
Pivot table between:
1.Avg. Sale_Price
2.Brand
3.Price_Category

INNOMATICS
RESEARCH LABS



BAR PLOT ON THE PIVOT TABLE

- Only four brands BOSCH, IFB, LG & SAMSUNG are having the wasing machines in all three price categories



	Capacity in kgs	Spin_Speed in RPM	Customer_Rating	Sale_Price	Discount %
Capacity in kgs	1	-0.037	-0.0013	0.22	-0.021
Spin_Speed in RPM	-0.037	1	-0.046	0.012	0.011
Customer_Rating	-0.0013	-0.046	1	0.063	0.03
Sale_Price	0.22	0.012	0.063	1	-0.063
Discount %	-0.021	0.011	0.03	-0.063	1

HEAT MAP BETWEEN EACH NUMERICAL VARIABLES

- ✓ Sale price and capacity of washing machine are somewhat positively correlated
- ✓ Other than these two variables, rest of the variables are having almost neutral relation among themselves.

INNOMATICS RESEARCH LABS

SUMMARY OF INSIGHTS:

- Brand wise LG has most number of products in the data followed by Whirlpool, SAMSUNG, IFB & BOSCH.
- We've got the catalog covered by mostly Economical and Premium price categories.
- Major number of products are available in Grey, White, Silver colours.
- Based on Load Capacity, Machines with 6 to 8 kg are in more number.
- Top Load in both semi & Fully Automatic category and Fully Automatic Front Load occupied more portion in our data.
- Most of the brands have positive ratings and LG and SAMSUNG have more excellent ratings.

SUGGESTIONS BASED ON OUR ANALYSIS:

➢ If consumer comes up with certain requirements as mentioned below:

- ❑ Premium Brand with 'Fully Automatic Top Load', having Excellent ratings and Load capacity within the range of 7 to 8 kg
  - ✓ **'LG', 'Panasonic', 'SAMSUNG'** would be the best choice

- ❑ Premium Brand with 'Fully Automatic Front Load', having Excellent ratings and Load capacity within the range of 7 to 8 kg
  - ✓ **'LG' & 'BOSCH'** will meet our requirements.

- ❑ Economical Brand with 'Fully Automatic Front Load', having Positive ratings
  - ✓ **'MarQ' & 'Croma'** would be the only choice

- ❑ Economical Brand with 'Semi Automatic Top Load', having Excellent ratings
  - ✓ **'LG', 'SAMSUNG', 'ONIDA' & 'MarQ'** fits in our specifications

CONCLUSION:

Hence, I can conclude that we have provided our inferences for the problem statement encountered to suggest a washing machine according to the customer's budget from the Flipkart website.

INNOMATICS RESEARCH LABS

THANK YOU

INNOMATICS
RESEARCH LABS