INT 353

EDA FINAL PROJECT

LOVELY PROFESSIONAL UNIVERSITY, PHAGWARA

BTech CSE specialization with ML and AI (UPGRAD)

Dataset- Myntra Fashion Clothing

Submitted to – Abhijeet Dutta

Name -Aman Gautam

Roll no. – RK20RUA20

Registration Number-12016284

Section – K20RU

Overview –

Myntra is a one stop shop for all your fashion and lifestyle needs. Being India's largest e-commerce store for fashion and lifestyle products, Myntra aims at providing a hassle free and enjoyable shopping experience to shoppers across the country with the widest range of brands and products on its portal. The brand is making a conscious effort to bring the power of fashion to shoppers with an array of the latest and trendiest products available in the country.

Headquarters – Bengaluru

Founder - *Mukesh Bansal, Ashutosh Lawania, Vineet Saxena founded Myntra in 2007.*

Website -*https://www.myntra.com*

About Project –

In this dataset I analysis the trends of different products according to reviews and ratings according to different categories and brands. I will also remove the different outliers and null values which affect the data.

*This dataset has **526564 rows** and **13columns.***

Techniques used –

- *Exploratory Data Analysis*
- *Data Cleaning*
- *Handling Outliers*
- *Univariate Analysis*
- *Bivariate Analysis*
- *Multivariate Analysis*
- *Hypothesis Testing*

Column description –

The columns in the dataset are described as follows:

1.URL of the Product – It has the URL of all product list present in dataset.

2.Productid – It has the id of all product present in dataset and each value has its own unique id so that that no two products have similar id.

3.Brand Name- it contains all top brand names whose products are sell by Myntra online.

4.Category – It contains all categories of clothes whether its western, Indian and other categories.

5.Individual category – It contains all individual details like whether its shirt, pants or any other women wear.

6.Category by gender – It tells whether the products are for women or men.

7.Description – It tells a detailed description of the product with information about the clothes so that user can think about it before it.

8.Discount price – It shows the product price after the discount amount is reduced from the main price.

9.Original price – It shows the original price without reducing the discount price from data.

10.Discount – The column shows the discount percent of all the products that how much discount is present in data.

11.Size- It shows how many different sizes of a product is available in store.

12.Ratings – it shows the ratings of a product given by users to the product.

13.Reviews - It shows the total reviews shown by different buyers after buying the product from Myntra

IMPORTING PYTHON LIBRARIES-

In [6]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
import math
warnings.filterwarnings("ignore")
```

Checking the top 5 rows of dataset

In [9]:
```
data.head()
```

Out[9]:

URL	Product_id	BrandName	Category	Individual_category	category_by_Gender	Description	DiscountPrice (in Rs)	OriginalPric (in R
:ps://www.myntra.com/jeans/roadster/roadster...	2296012	Roadster	Bottom Wear	jeans	Men	roadster men navy blue slim fit mid rise clean...	824.0	1499
ps://www.myntra.com/track-pants/locomotive/...	13780156	LOCOMOTIVE	Bottom Wear	track-pants	Men	locomotive men black white solid slim fit tra...	517.0	1149
tps://www.myntra.com/shirts/roadster/roadste...	11895958	Roadster	Topwear	shirts	Men	roadster men navy white black geometric print...	629.0	1399
//www.myntra.com/shapewear/zivame/zivame...	4335679	Zivame	Lingerie & Sleep Wear	shapewear	Women	zivame women black saree shapewear zi3023core0...	893.0	1295
ttps://www.myntra.com/tshirts/roadster/roadst...	11690882	Roadster	Western	tshirts	Women	roadster women white solid v neck pure cotton ...	NaN	599

DATA CLEANING-

First, we set the columns in proper manner as we can see the URL column is our first column so we can shift it to last so our data will look proper sequence of columns.

```
In [10]: new_order = [1,2,3,4,5,6,7,8,9,10,11,12,0]
         data=data[data.columns[new_order]]
```

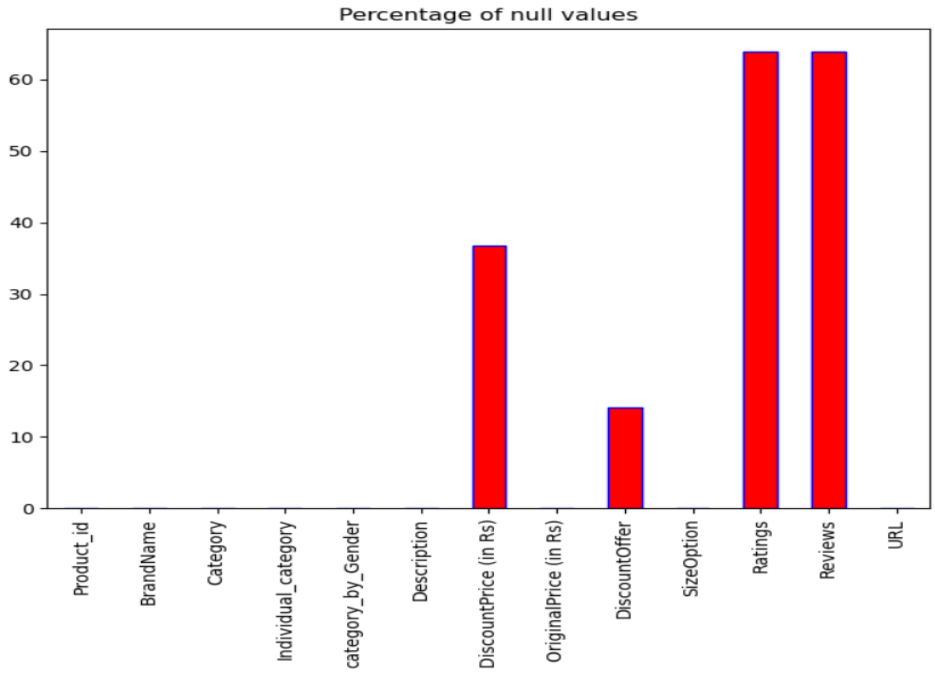
```
In [11]: data.head()
```

Out[11]:

	Product_id	BrandName	Category	Individual_category	category_by_Gender	Description	DiscountPrice (in Rs)	OriginalPrice (in Rs)	DiscountOffer	SizeOption	Ratings
0	2296012	Roadster	Bottom Wear	jeans	Men	roadster men navy blue slim fit mid rise clean...	824.0	1499.0	45% OFF	28, 30, 32, 34, 36	3.9
1	13780156	LOCOMOTIVE	Bottom Wear	track-pants	Men	locomotive men black white solid slim fit tra...	517.0	1149.0	55% OFF	S, M, L, XL	4.0
2	11895958	Roadster	Topwear	shirts	Men	roadster men navy white black geometric print...	629.0	1399.0	55% OFF	38, 40, 42, 44, 46, 48	4.3
3	4335679	Zivame	Lingerie & Sleep Wear	shapewear	Women	zivame women black saree shapewear zi3023core0...	893.0	1295.0	31% OFF	S, M, L, XL, XXL	4.2
4	11690882	Roadster	Western	tshirts	Women	roadster women white solid v neck pure cotton ...	NaN	599.0	35% OFF	XS, S, M, L, XL	4.2

Checking Null values of dataset-

```
In [14]: data.isnull().sum()

Out[14]: Product_id                   0
         BrandName                    0
         Category                     0
         Individual_category          0
         category_by_Gender           0
         Description                  0
         DiscountPrice (in Rs)   193158
         OriginalPrice (in Rs)        0
         DiscountOffer            74306
         SizeOption                   0
         Ratings                 336152
         Reviews                 336152
         URL                          0
         dtype: int64
```

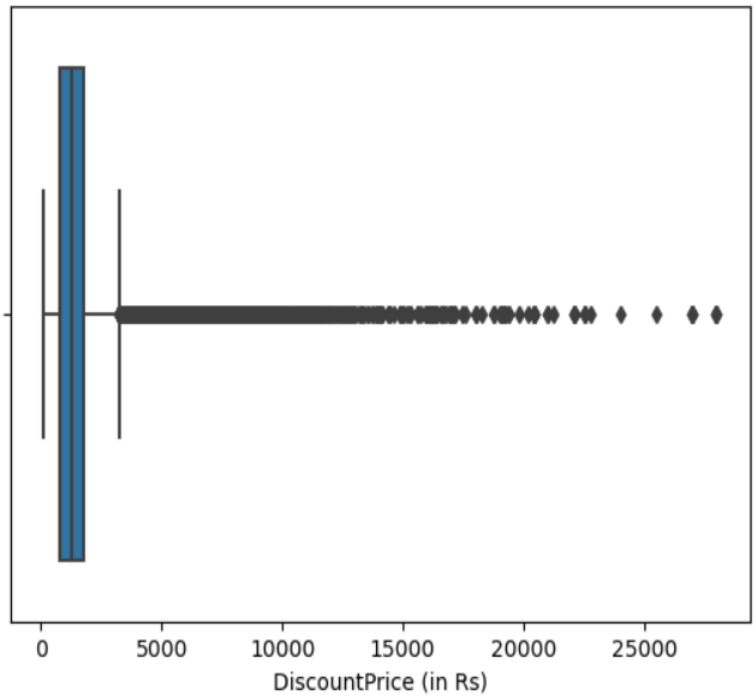


As we can see there are 3 columns with Null Values –

- Discount Price
- Discount Offer
- Ratings
- Reviews

1. ***Handling Outliers on Discount Price-***

Boxplot for discount Price



As we can see there maybe outliers present after 20000 so we have to check using IQR method.

```
In [40]: data["DiscountPrice (in Rs)"].quantile([0.5,0.7,0.9,0.95,0.99,1])

Out[40]: 0.50     1299.000000
         0.70     1790.085106
         0.90     1849.000000
         0.95     2432.000000
         0.99     4453.900000
         1.00    27996.000000
         Name: DiscountPrice (in Rs), dtype: float64
```

Conclusion –

As we can see there is an outlier but they will not affect data so can ignore them as there is high discount on higher cost clothes.

2. ***Handling Null values on Discount Price-***

We can fill the null rows of this column by dividing the column on the basis of Individual category and taking its mean.
After that we can fill the values as mean discount price of that individual category in null rows.

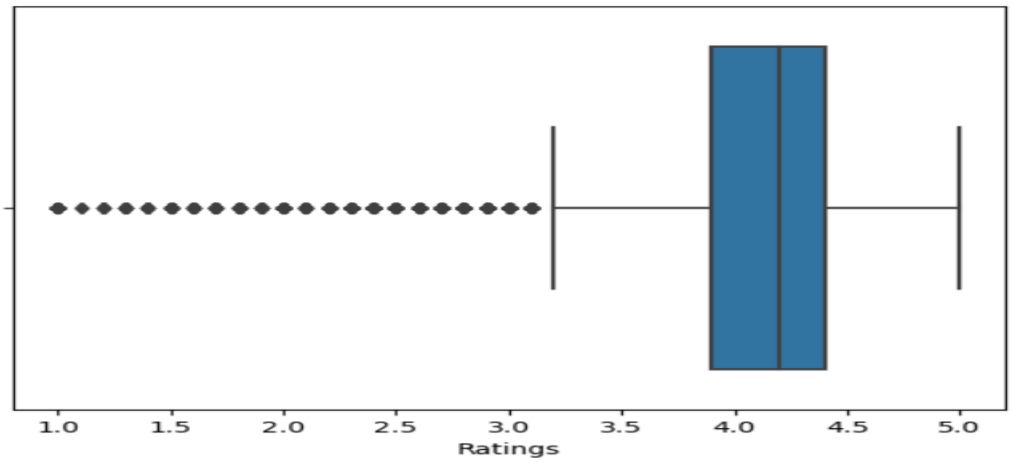
```
1]: (in Rs)'] = data['DiscountPrice (in Rs)'].fillna(data.groupby('Individual_category')['DiscountPrice (in Rs)'].transform('mean'))
```

We successfully remove null value –

```
]: data["DiscountPrice (in Rs)"].isnull().sum()
]: 0
```

3. ***Handling outliers on Ratings –***

Boxplot for Ratings



As we can see there is continuous decrease of ratings so there is very less chances of these beings as outliers.

4. ***Handling Null Values on Ratings –***

Applying same method as used in discount price to fill null values.

- Filling null rows as mean of Ratings using group by method on individual category column –

```
In [35]: data['Ratings'] = data['Ratings'].fillna(data.groupby('Individual_category')['Ratings'].transform('mean'))
```

5. ***Handling Null values on Discount Offer –***

Data in this column is present in object which contain int and String values.

- Removing the String values from column –

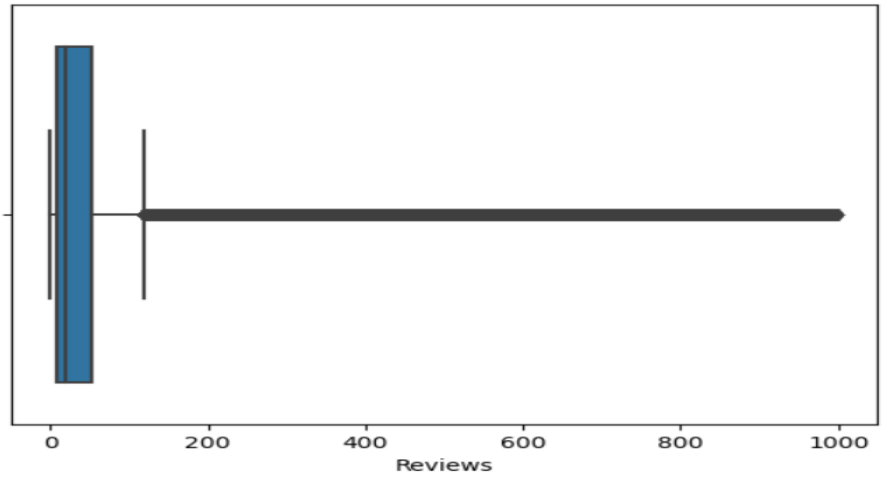
```
In [49]: #data["DiscountOffer)"]=data["DiscountOffer"].str.replace(" OFF","% OFF")
         data["DiscountOffer"]=data["DiscountOffer"].str.replace('[OFF,"OFF",Hurry*,' ',"%","Rs. "]', '')

In [50]: data["DiscountOffer"].value_counts()

Out[50]: 50      53208
         60      40518
         40      27902
         20      25595
         55      25176
                 ...
         1126        1
         1153        1
         1103        1
         1114        1
         283         1
         Name: DiscountOffer, Length: 1347, dtype: int64
```

- Filling null values as 0 because there is no discount offer on these products

```
In [51]: data['DiscountOffer'] = data['DiscountOffer'].fillna(0)
```

6. ***Handling outliers on Reviews –***

Boxplot for Reviews



The data is continuously increasing so there are no outliers present in the column.

7. ***Handling Null values on Reviews –***

We can use backward fill or forward fill here because the data is similar in many adjacent rows as we go down in column –

	Reviews	
	999.0	https
	999.0	https
	999.0	http
	999.0	https://w
	999.0	http
	999.0	https:/
	998.0	https
	998.0	https:/

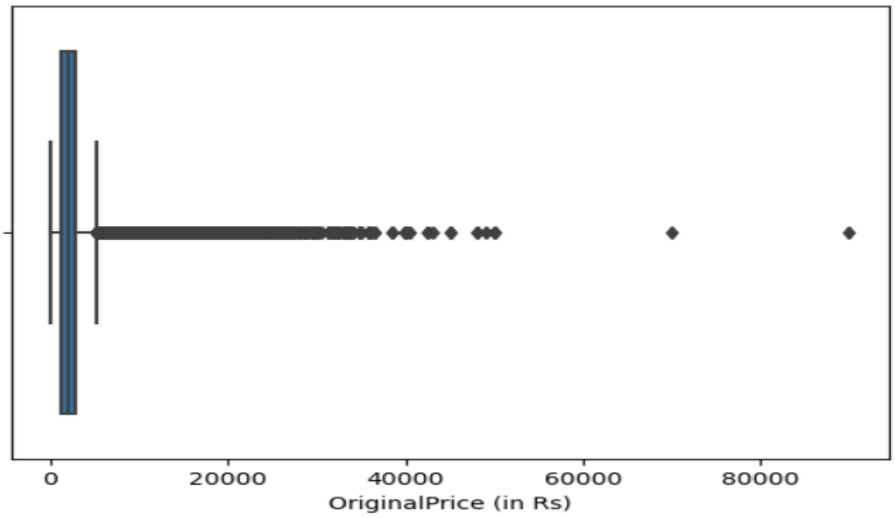
- Using forward fill to fill the null values similar to front column –

```
In [57]:
data['Reviews'] = data['Reviews'].fillna(method="ffill")
#data['Reviews'] = data['Reviews'].interpolate()
```

- Using interpolate method which is used to guess is null values give the same result as forward fill.

8. ***Handling outliers on Original Price –***

Boxplot for Original Price



So, there are definitely some outliers present in this column

- Checking outliers –

```
In [62]: data["OriginalPrice (in Rs)"].quantile([0.5,0.7,0.9,0.95,0.96,0.97,0.99])
```

```
Out[62]: 0.50    1999.0
         0.70    2599.0
         0.90    4399.0
         0.95    5799.0
         0.96    6000.0
         0.97    6799.0
         0.99    9299.0
         Name: OriginalPrice (in Rs), dtype: float64
```

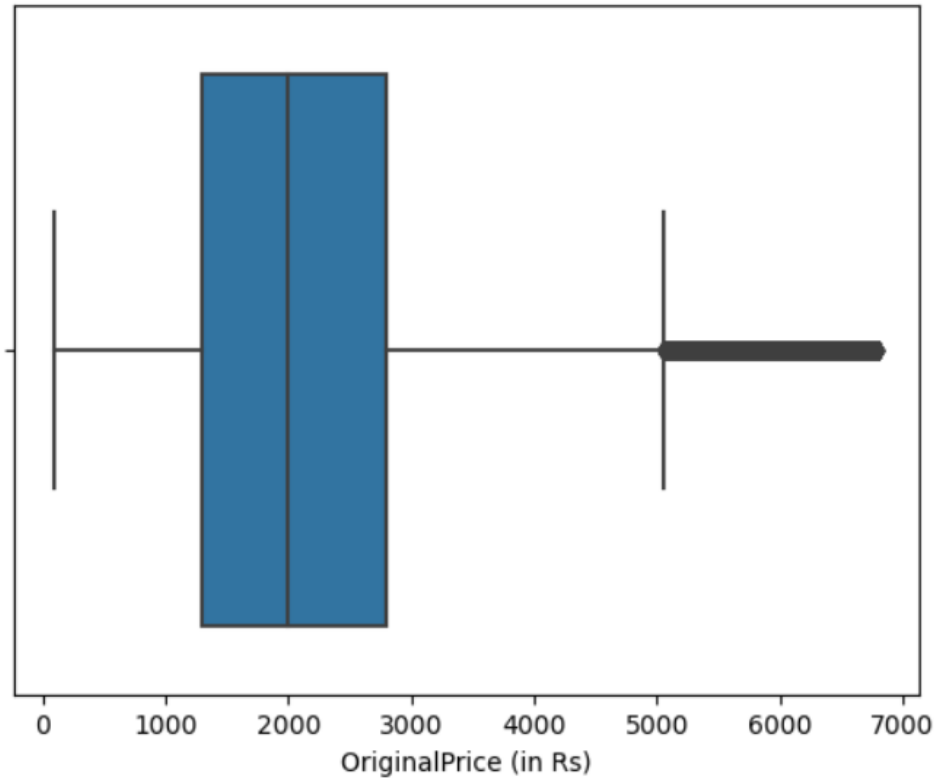
So, by using quantile we can see there is outliers because of sudden increase in price after 97%

- Removing outliers after 97% -

```
In [63]: min_thresold=data["OriginalPrice (in Rs)"].quantile(0.97)
```

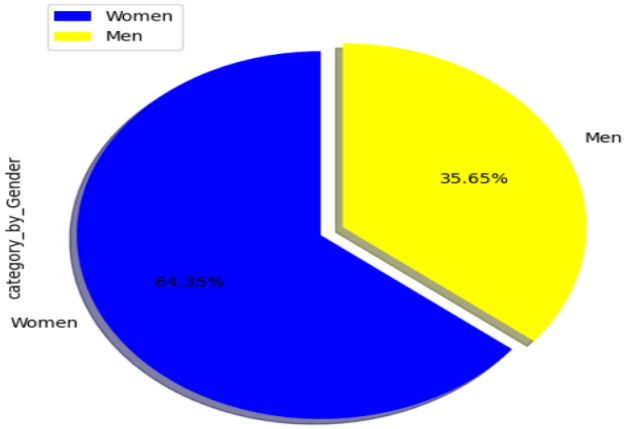
```
In [65]: data=data[data["OriginalPrice (in Rs)"]<min_thresold]
```

- *Boxplot after removing Outliers from Original Price –*



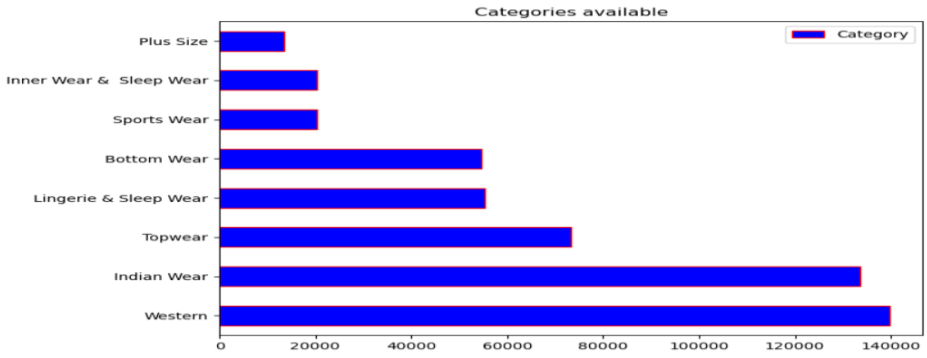
Univariate Analysis –

- ***Do we have more Male or Female clothes categories –***



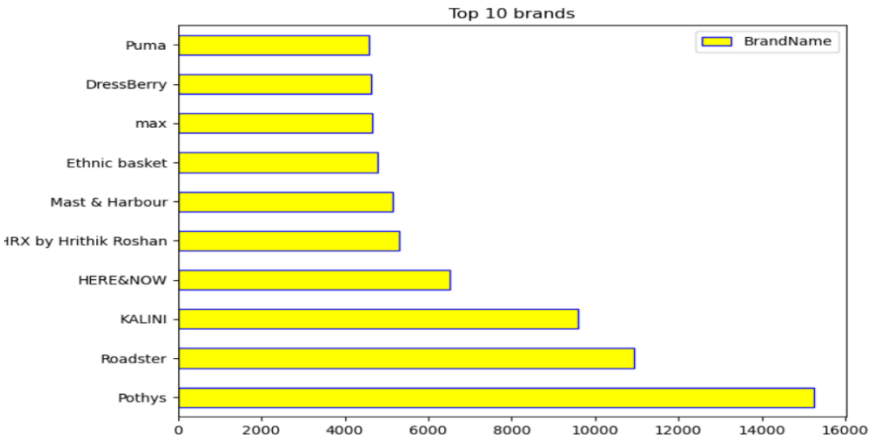
Inference – *As we can conclude woman clothes category is more available at Myntra as compare to men clothes category.*

- ***Which category is mostly available on Myntra platform?***



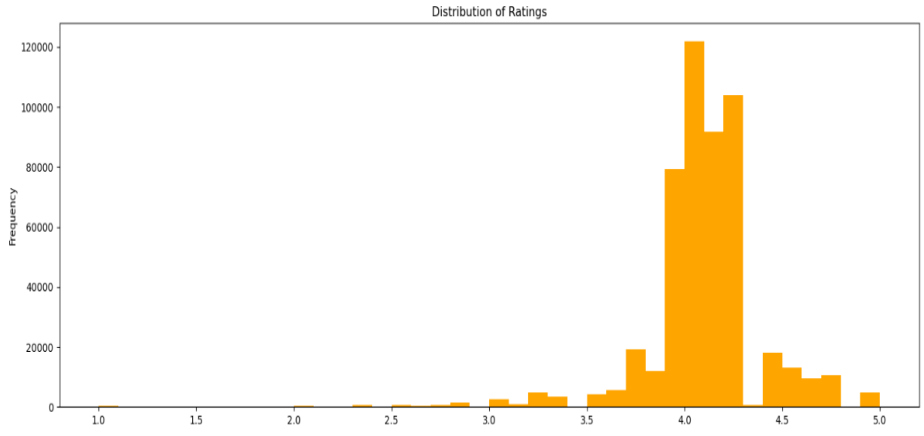
Conclusion- *Western clothes is mostly available in Myntra and then Indian wear is on second place.*

- ***What types of brand product Myntra mostly Provide?***



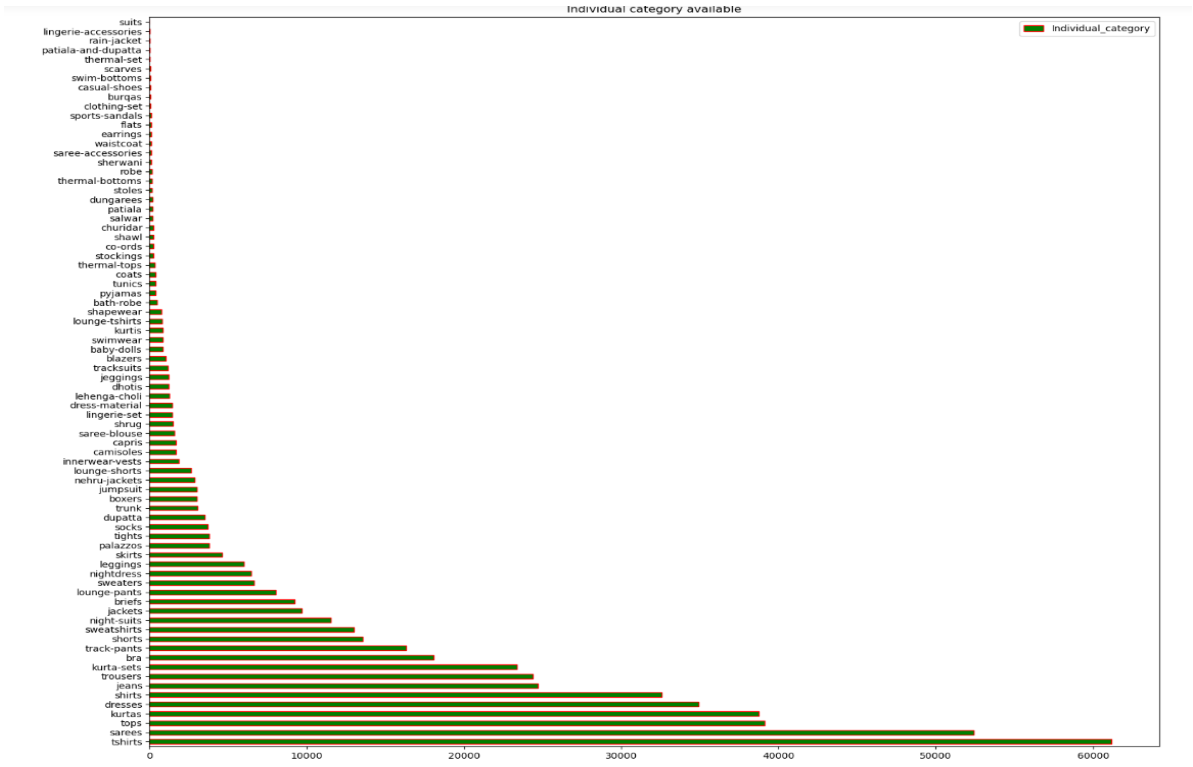
Conclusion – *So, Pothys is the brand whose products are mostly available.*

- ***Where the maximum distribution of Ratings lies?***



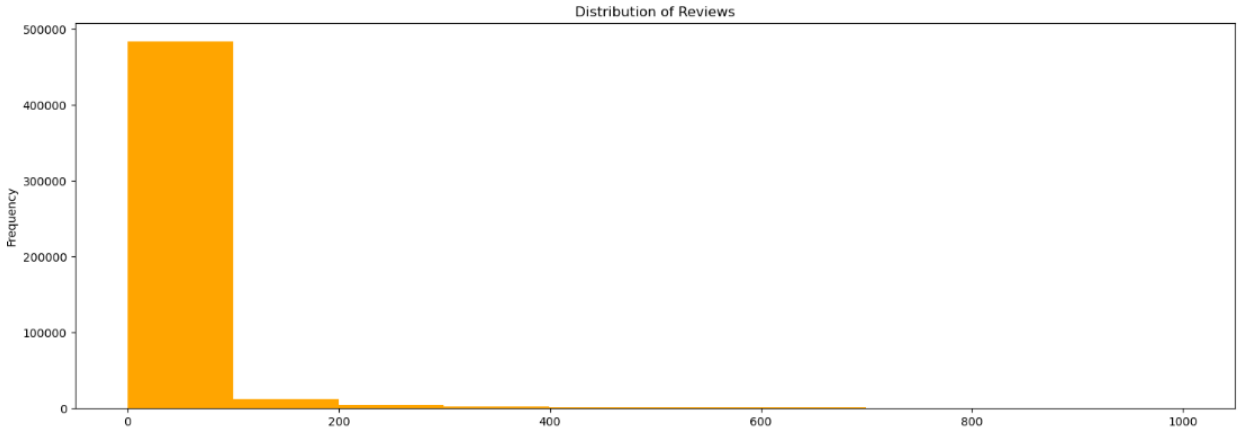
Conclusion *– The maximum rating distribution lies between 3.5 to 4.5 for all the categories*

- ***What type of individual cloth category is highly available for customers?***



Conclusion –> *T-shirts and sarees is mostly there for customers*

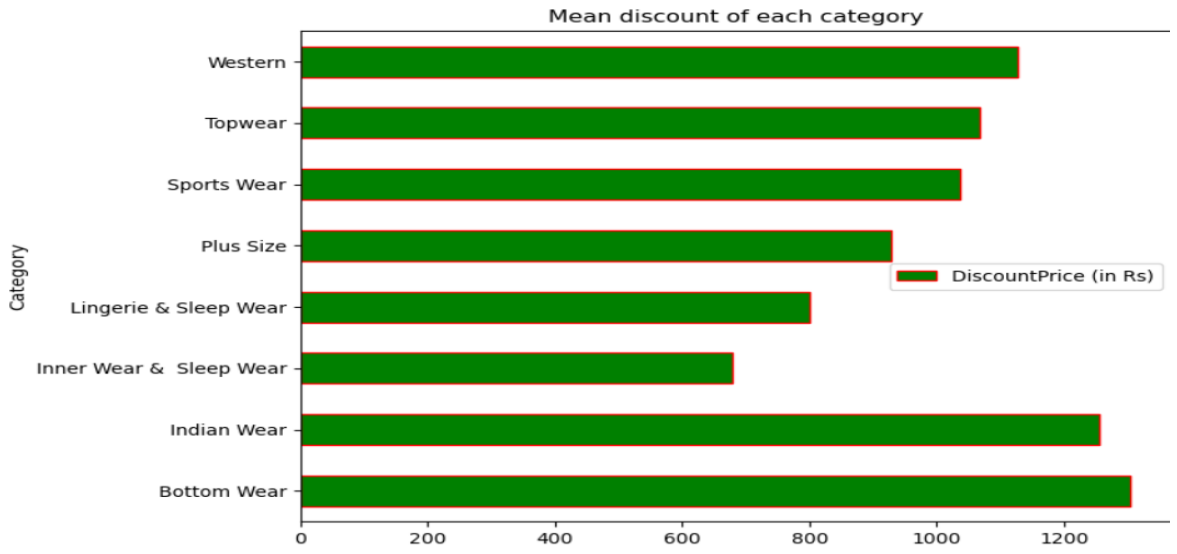
- ***In which range the most of the reviews given to products?***



Conclusion *– Range of reviews is greatly lies between 0-200.*

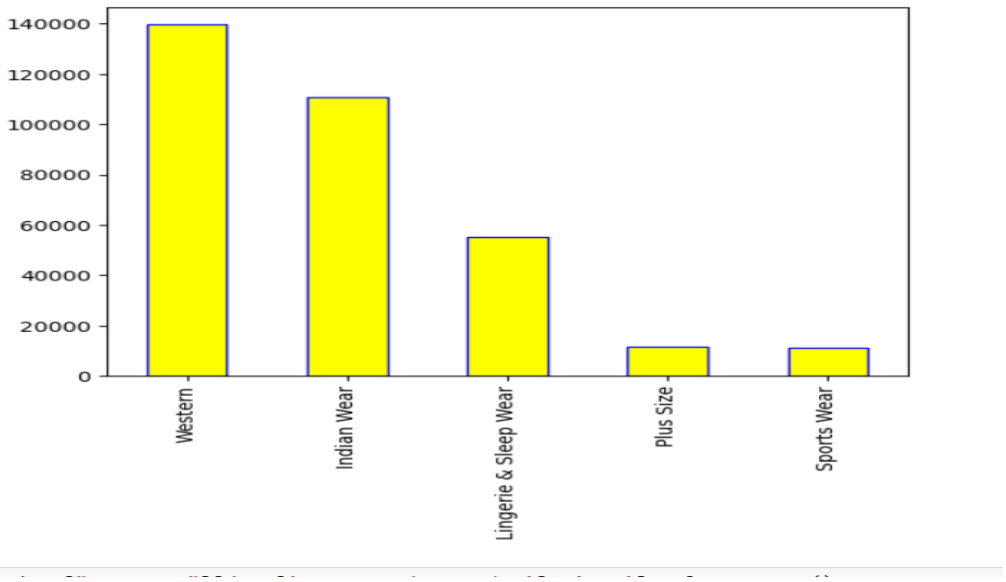
Bivariate Analysis –

- ***What is the average discount on category of products?***



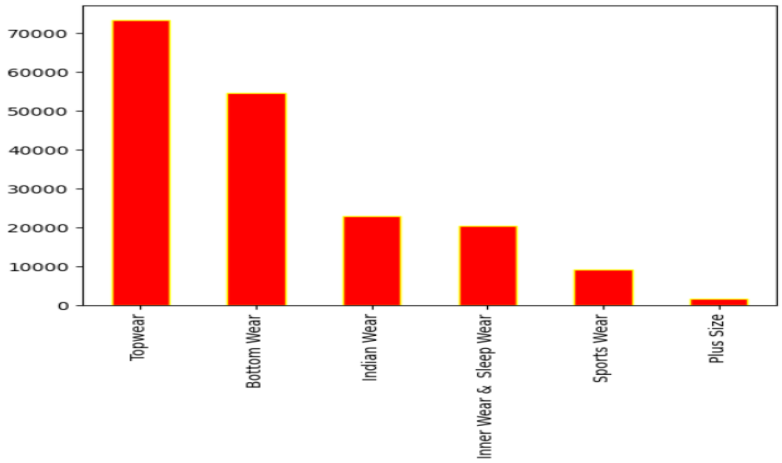
Conclude *– Myntra provide maximum discount on products of Bottom wear then Indian wear is on second.*

- ***What Cloth category is highly in demand for women?***



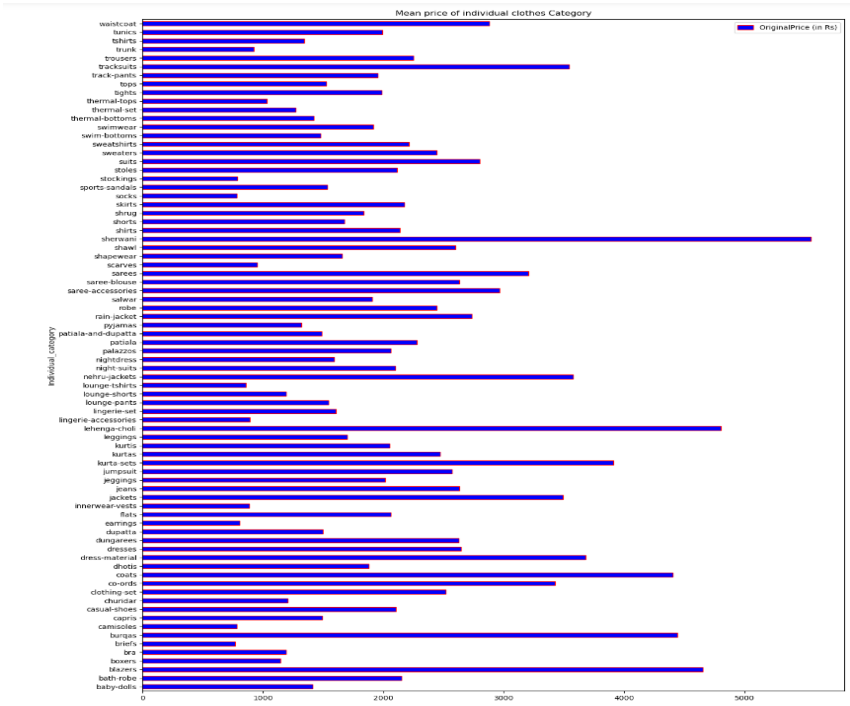
Conclusion *– Western is highly available for women*

- ***What Cloth category is highly in demand for men?***



Conclusion *– Top wear is mostly available for men.*

- ***Which individual Category Product as high Mean Price?***

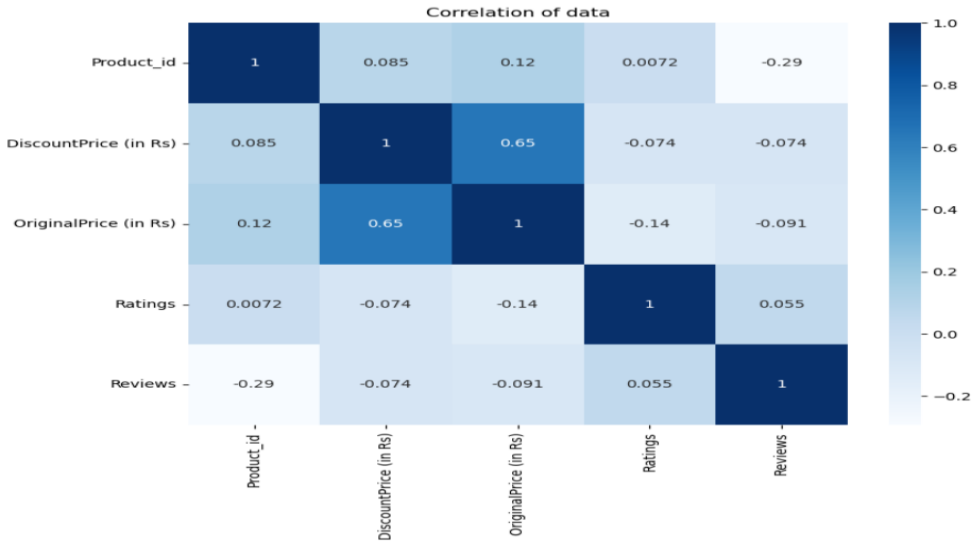


Conclusion-

Sherwani is expensive cloth on Myntra

Multivariate Analysis –

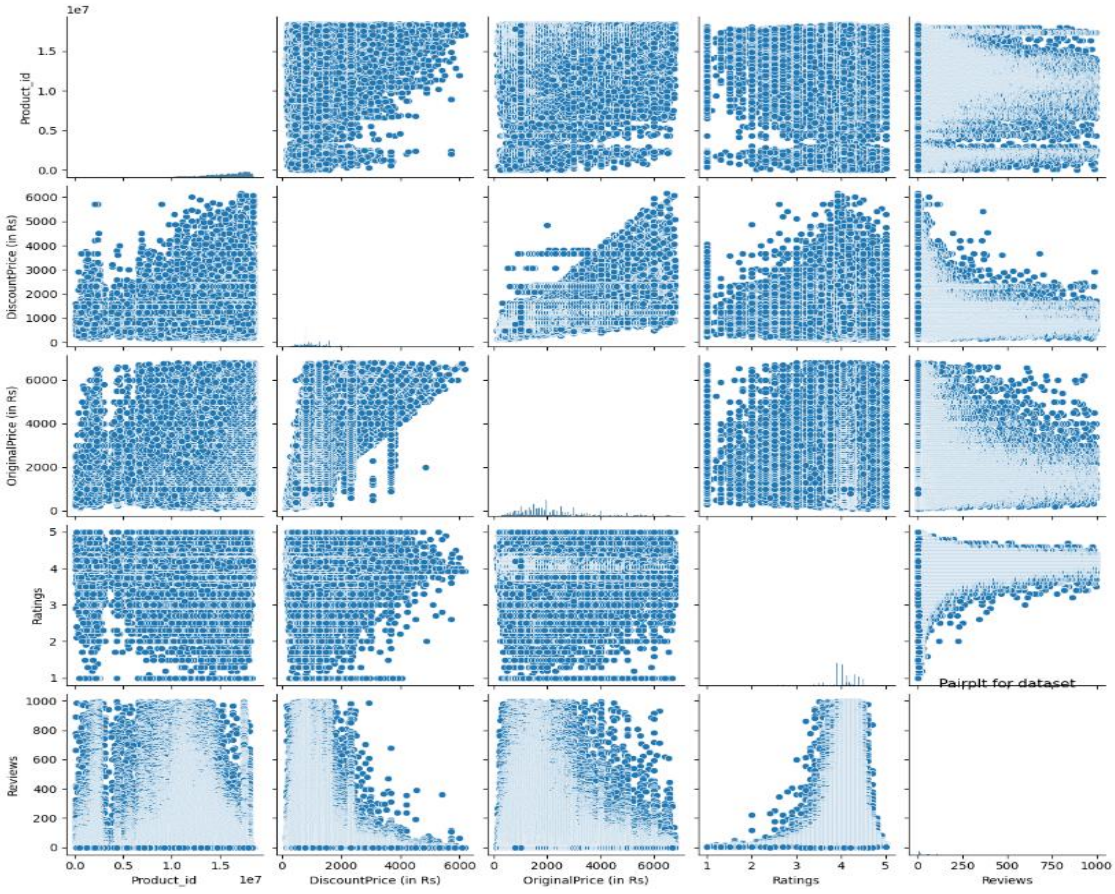
1. *Using correlation heatmap for Multi variate Analysis*



Inference-

- *As we can see the value of original Price and Discount Price is correlated which shows that the product with High price has high discount too.*
- *Ratings have negative .14 relation with original Price which shows that ratings are low for high price products. One reason for this can be high price from which buyers cannot be satisfied.*

1. *Using Pair Plot for Multi Variate Analysis –*



Inference – *Using Pair plot we found the most relation of where the most of data lies when compare with other columns.*

For example, we can see higher the reviews for high ratings.

Hypothesis Testing –

Let's apply hypothesis T-test on column ratings –

- *Finding mean of Ratings –*

```
In [139]: # importing the library for t-test
          #let alpha value be 0.05 or 5%
          rating_mean=np.mean(data["Ratings"])
          rating_mean

Out[139]: 4.082845154868895
```

- *Now importing Library for T-test and taking a random sample of 10000 rows –*

```
In [140]: # creating random sample from the column rating
          from scipy.stats import ttest_1samp
          sample_size=10000
          rating_sample=np.random.choice(data["Ratings"],sample_size)
```

- *Let significance value be 5% and now Applying T-test to find the P-value for the random sample –*

```
In [106]: # finding p-value for sample
          ttest,p_value=ttest_1samp(rating_sample,4)

In [107]: print(p_value)

          1.0651423710497649e-169
```

- Let's check if null hypothesis rejected or not –

```
In [108]: if(p_value < 0.05):
              print("We are rejecting the null hypothesis")
          else:
              print("we are accepting the null hypothesis")

          We are rejecting the null hypothesis
```

Conclusion –

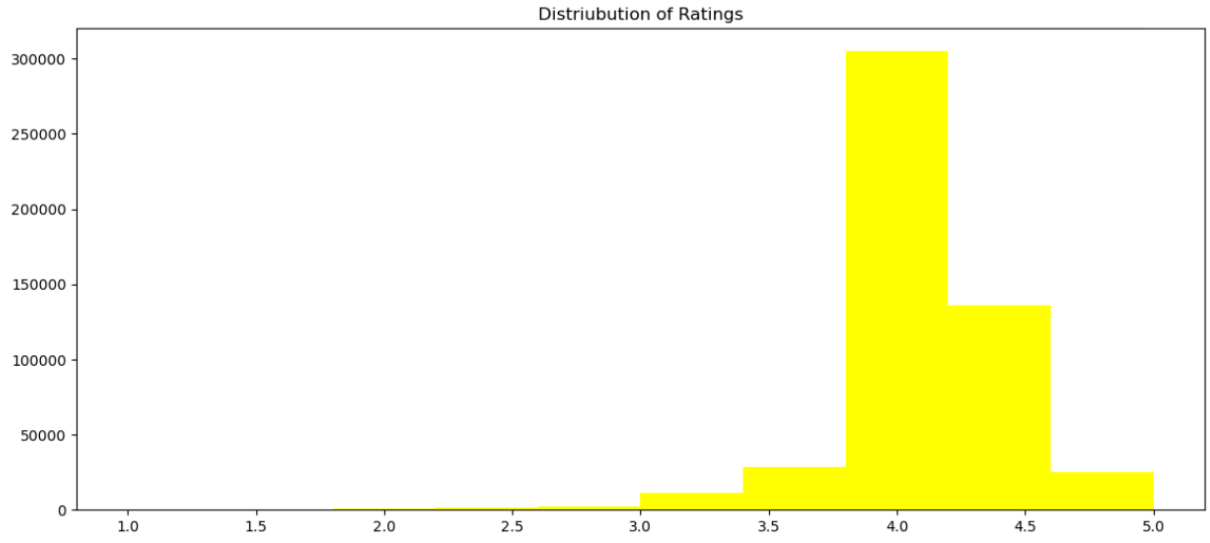
As we can see the null hypothesis is rejected

So, reason for null Hypothesis rejection is because the data is not in Normal Distribution.

- *From Shapiro library we can find if data is in normal Distribution or not –*

```
In [110]: from scipy.stats import shapiro
          dataToset =data["Ratings"]
          stat, p=shapiro(dataToset)
          print("stat=%.2f , p=%.30f" % (stat, p))
          if p>0.05:
              print("Normal Distribution")
          else:
              print("Not a normal distribution")

          stat=0.81 , p=0.000000000000000000000000000000
          Not a normal distribution
```

- *Plotting Histogram for data Distribution of Ratings –*



Conclusion –

As we can see from using Shapiro and histogram null hypothesis is rejected.

Bibliography –

- https://www.upgrad.com/data-science-course/
- https://www.geeksforgeeks.org/
- https://github.com/aman9650/EDA_Poject
- https://www.kaggle.com/datasets/shivamb/fashion-clothing-products-catalog


Thank You!