University of Bolton

Master of Science in Data Analytics and Technologies-extended

Big Data Technologies (DAT7015)

“Fashion Dataset Analysis”

Submitted to: Professor Dr Anchal Garg

Submitted by: Bandana Adhikari (2118001)

Abstract

In today's world enormous amount of data are generated from every sector like fashion, health, social media every second (terabyte to petabytes) and the data types are becoming more complex. The concept of big data is to analyze voluminous data to extract valuable information for better decision making. The way designer produces and promote their items, fashion industry is also facing lots of changes which results in big data. Analysis of these data can help company to find out customer demands, trends so that they can produce right product at right time and in right quantities. Analyzing and managing those big data is quite challenging. Last few decades different work has been carried out in Big Data. To solve this problem Hadoop MapReduce Framework can be used for big data analytics. In this paper, this project deals with analysis of fashion datasets using Hadoop MapReduce framework. As there were two datasets to analyze from, so the data needs to be merged and cleaned before using it. so, SQL (Structure Query language) language is introduced for data manipulation and merge. Python program is carried out for data preprocessing stage and machine learning. According to the data provided the clustering of data is optimal solution. So, to from the clusters among similar data clustering algorithm I.e.; K-mean clustering algorithm is used. Here, Hive is used for data analysis (using query method similar to structure query language) and PowerBI is a tool which has been used to represent the data in visual form for better understanding which will enhance the decision-making process.

Keywords: Machine Learning, PowerBI, Hive, KMean, Dataframe, SQL, Python, MapReduce

Table of Contents

List of Figures

List of abbreviations

SQL	Structured Query Language
PowerBI	Power Business Intelligence
df	Dataframe
pd	pandas
np	NumPy

1 Introduction

Social media and e-commerce have increased the fashion industry. Over last the 20 years fashion industry has significantly evolved since then it started to expand. Until 1980s the retailers used to forecast demand and trend with their own capability called as ready-to wear (Silva et al., 2019). Trends and massive production of clothing has collected the user's perspective, brand recognition and huge amount of data. In today's world fashion industry have made a remarkable growth it is generating immense about of data which is called big data which can be in structured and unstructured data such as word, image. Since recent year fashion studies have received attention form machine learning, computer vision and from different multimedia community. The data which is generated from fashion portrays the features of big data. Big data consists of 4V's- Velocity, Veracity, Volume, Variety. With the change in dynamics of fashion industry retailers are forced to decrease the product price, design the products according to customers choice so that they can place a profitable place in the market. Customer demands are changing frequently this is why Data analysis in the fashion industry gained a significant importance so that companies can analyze customer behavior, trend forecast, customer preference for the products and so on. So, without the analysis of these big data companies face failure and lose a lot of money due to this changing trends. There is development in many technologies which fulfill the customer needs and satisfy their needs (Bhardwaj & Fairhurst, 2010). In past, historical sales was insufficient due to the increase in customer demands and high personalized shopping behavior (Silva et al., 2019). Like a famous brand Zara mines the big data to analyze and understand the customer behavior and meet their expectation on daily basis. Fast fashion trend was only associated with the H&M and Zara which were able to produce 10,000-15000 new items every year. But now these companies are able to produce 50,000-100,000 new items every year form efficient Big Data Analysis. The analysis and process of these data provides a valuable information to an industry which helps to meet the needs of each customer. Increase in dependency of technology has generated the huge amount of data. Store, analyze, share, visualize and process of these data with normal data base and traditional software method is not possible. Nowadays, many tools like Hadoop, PowerBI, Tableau, Spark, Hive are available for the processing of these data. That is why, Apache Hadoop and MapReduce is used to store and compute the complex data in this project. Apache Hadoop is a popular software for the distributed storage and processing of data at very high speed. The goal of fashion data understanding is to explore clothing attributes like color, brand names to support advance fashion applications. In the proposed work, there is two datasets of fashion are considered for the analysis.

2 Methodology

2.1 Business Understanding

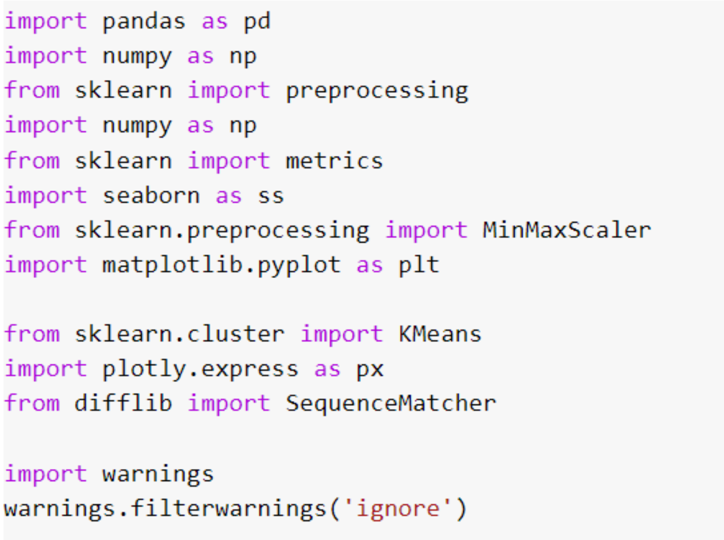
Whenever people think of fashion, they think it as the world of glamour and luxurious brands, but in reality, it is a collection of different sectors which it relies upon to function properly.

Fashion is formed by the combination of textile design and production, fashion designing, fashion shows, media and marketing, retailing and merchandising. The fashion industry sustains if they are able to explore and understand the fashion trends, understand the consumption of product and able to recognize good staffs and manage the resources. Last couple of year has not been good for fashion industry due to covid outbreak but coming to this year the graph seems to increase.

2.2 Data Understanding

Data understanding is the process to gain insights about the data which will further help in data analysis process. So, to have knowledge of data is very important in every sectors. After having business understanding we need to understand the quality of data, what is there in the data, where to find the data, what tools to use to extract the data and what data do you have. In this project there are two different datasets about the fashion industry where one dataset one is fashion brand details and other one is fashion datasets.

2.2.1 Importing Libraries



```
import pandas as pd
import numpy as np
from sklearn import preprocessing
import numpy as np
from sklearn import metrics
import seaborn as ss
from sklearn.preprocessing import MinMaxScaler
import matplotlib.pyplot as plt

from sklearn.cluster import KMeans
import plotly.express as px
from difflib import SequenceMatcher

import warnings
warnings.filterwarnings('ignore')
```

Figure 2.1: Library used for Data analysis

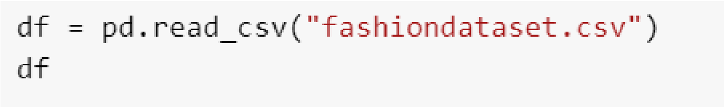
Pandas library is used to work with dataset which has functions to analyze, explore, manipulate, clean, arrange the data.

NumPy library is used to work with arrays

Seaborn and matplotlib is used for visualization (to plot graphs)

Scikit-learn (sklearn) is used for machine learn which provides efficient tool for statistical model or machine learning.

2.2.2 Loading csv file to python

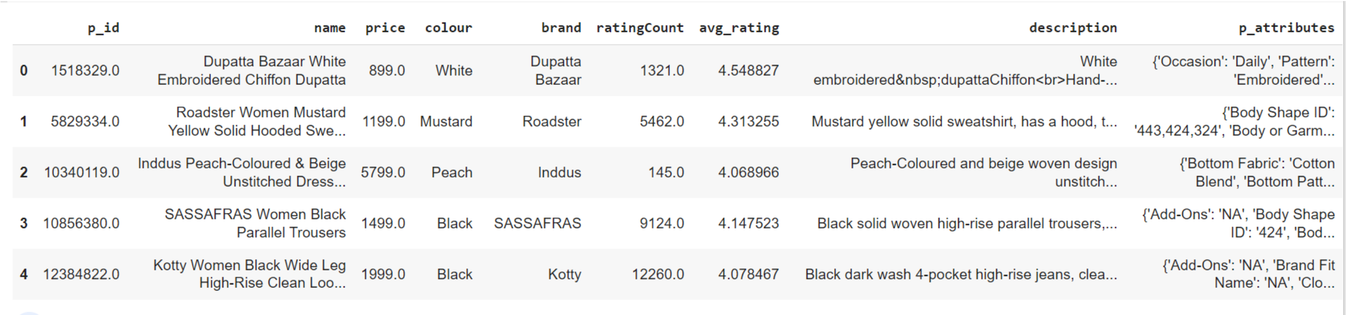


```
df = pd.read_csv("fashiondataset.csv")
df
```

Figure 2.2: Loading the data into df dataframe

The dataset is store in df data frame. Dataframe is the most common data structure used for data analytics process because they are more flexible for the storing and working of data, where data is stored into 2-dimensional table which is row and column like spreadsheet as shown in figure 3

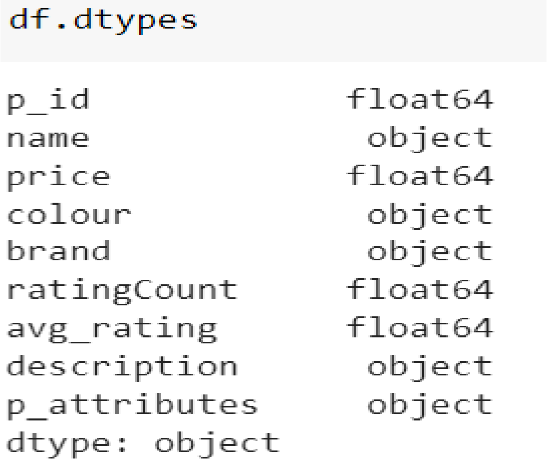
First dataset is imported using the above figure command.



	p_id	name	price	colour	brand	ratingCount	avg_rating	description	p_attributes
0	1518329.0	Dupatta Bazaar White Embroidered Chiffon Dupatta	899.0	White	Dupatta Bazaar	1321.0	4.548827	White embroidered dupattaChiffon Hand-...	{'Occasion': 'Daily', 'Pattern': 'Embroidered'...
1	5829334.0	Roadster Women Mustard Yellow Solid Hooded Swe...	1199.0	Mustard	Roadster	5462.0	4.313255	Mustard yellow solid sweatshirt, has a hood, t...	{'Body Shape ID': '443,424,324', 'Body or Garm...
2	10340119.0	Inddus Peach-Coloured & Beige Unstitched Dress...	5799.0	Peach	Inddus	145.0	4.068966	Peach-Coloured and beige woven design unstitch...	{'Bottom Fabric': 'Cotton Blend', 'Bottom Patt...
3	10856380.0	SASSAFRAS Women Black Parallel Trousers	1499.0	Black	SASSAFRAS	9124.0	4.147523	Black solid woven high-rise parallel trousers,...	{'Add-Ons': 'NA', 'Body Shape ID': '424', 'Bod...
4	12384822.0	Kotty Women Black Wide Leg High-Rise Clean Loo...	1999.0	Black	Kotty	12260.0	4.078467	Black dark wash 4-pocket high-rise jeans, clea...	{'Add-Ons': 'NA', 'Brand Fit Name': 'NA', 'Clo...

Figure 2.3: Display of Data of df dataframe

2.2.3 Datatypes Check and Description of Datatypes



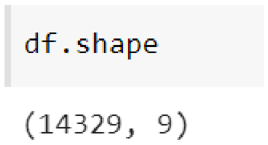
```
df.dtypes
```

```
p_id              float64
name               object
price             float64
colour             object
brand              object
ratingCount       float64
avg_rating        float64
description        object
p_attributes       object
dtype: object
```

Figure 2.4: Display of data types in df dataframe

Data types are the classification of data item which represent the value and tells what kind of operation can be performed in particular data. There are different types of data types like Numeric, String, Boolean and so.

In this dataset there are only two kinds of data types which are float which is numeric data types and string.



```
df.shape
```

```
(14329, 9)
```

Figure 2.5: shape of data frame df

As shown in figure 5 there are 14329 values and 9 different attributes

First dataset name fashion contains 14329 values with nine different columns which are: -

P_id: product id

Name: object

price: float64

color: object

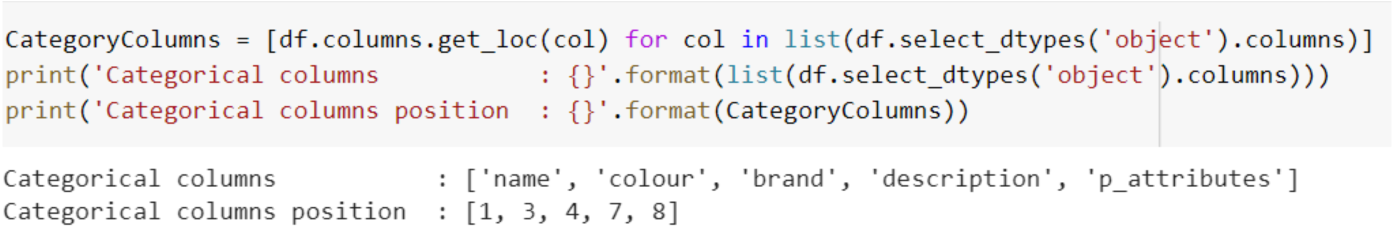
brand: object

ratingCount: float64

description: object

p_attributes: object(product attributes)

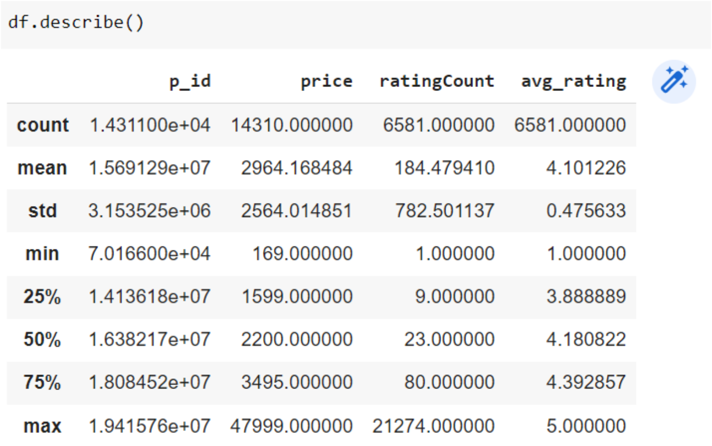
In this dataset there are five categorical column which are name, color, brand, description, p_attributes.



```
CategoryColumns = [df.columns.get_loc(col) for col in list(df.select_dtypes('object').columns)]
print('Categorical columns          : {}'.format(list(df.select_dtypes('object').columns)))
print('Categorical columns position  : {}'.format(CategoryColumns))
```

```
Categorical columns           : ['name', 'colour', 'brand', 'description', 'p_attributes']
Categorical columns position  : [1, 3, 4, 7, 8]
```

Figure 2.6: Shows the which and what are the categorical data present in dataframe df

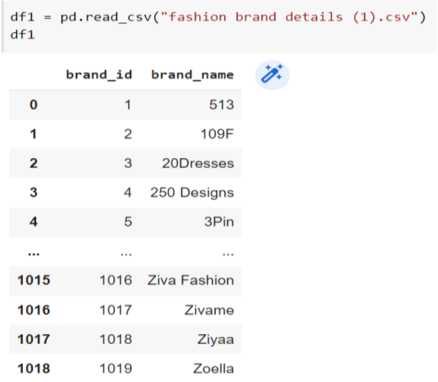


```
df.describe()
```

	p_id	price	ratingCount	avg_rating
count	1.431100e+04	14310.000000	6581.000000	6581.000000
mean	1.569129e+07	2964.168484	184.479410	4.101226
std	3.153525e+06	2564.014851	782.501137	0.475633
min	7.016600e+04	169.000000	1.000000	1.000000
25%	1.413618e+07	1599.000000	9.000000	3.888889
50%	1.638217e+07	2200.000000	23.000000	4.180822
75%	1.808452e+07	3495.000000	80.000000	4.392857
max	1.941576e+07	47999.000000	21274.000000	5.000000

Figure 2.7: Description about the data frame

The above figure shows the mean, count, standard deviation, maximum and minimum values, calculated of all the numeric data.

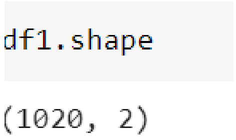


```
df1 = pd.read_csv("fashion brand details (1).csv")
df1
```

	brand_id	brand_name
0	1	513
1	2	109F
2	3	20Dresses
3	4	250 Designs
4	5	3Pin
...	...	...
1015	1016	Ziva Fashion
1016	1017	Zivame
1017	1018	Ziyaa
1018	1019	Zoella

Figure 2.8: Second dataset loaded and store in df1

Second dataset is about fashion brand details is stored in data frame df1.

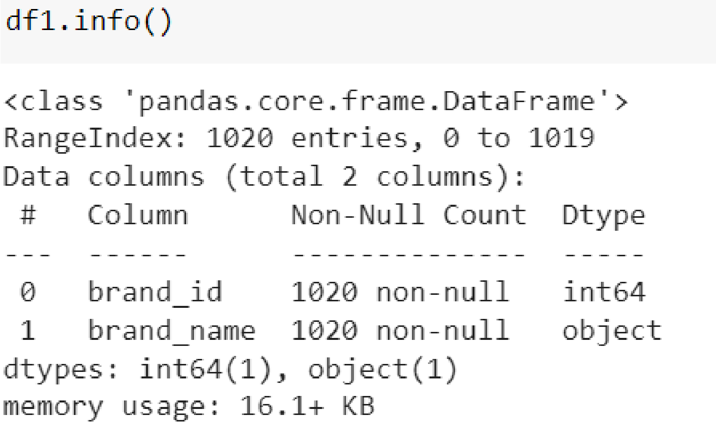


```
df1.shape
```

```
(1020, 2)
```

Figure2.9: Shape of dataframe df1

This dataset contains 1020 values and 2 columns which are:



```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1020 entries, 0 to 1019
Data columns (total 2 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   brand_id    1020 non-null   int64 
 1   brand_name  1020 non-null   object
dtypes: int64(1), object(1)
memory usage: 16.1+ KB
```

Figure 2.10: Information about dataframe df1

Figure 2.10 shows that in dataframe df1 there are two attributes brand_id, brand_name anda data type is integer and string.

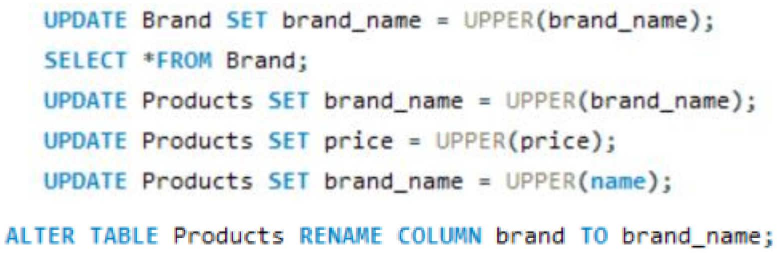
2.3 Data Preparation

Data is collected from multiple data sources which can be structure or unstructured, noisy, unformatted, we should perform Data preparation as it is the first step in data analytics projects. This step is carried out to ensure that there are no missing values, outliers, noisy data so that the data is readable and accurate for further analysis. Analyzing those data which is not prepared carefully can lead to mislead solutions. So good data preparation led to effective data analysis which limits inaccuracies and errors during the data processing phase.

In this project there are two datasets which needs to be merged together and form a new dataset. So, for the merging process Microsoft workbench software is used where Structured Query Language is implemented.

Another step is changing all the string values to upper case in both datasets to eradicate the duplicate values. Before the merging step there is two attributes in the dataset with same values but with different attribute name. So, to remove duplication rename of the attribute is done.

2.3.1 Data Manipulation

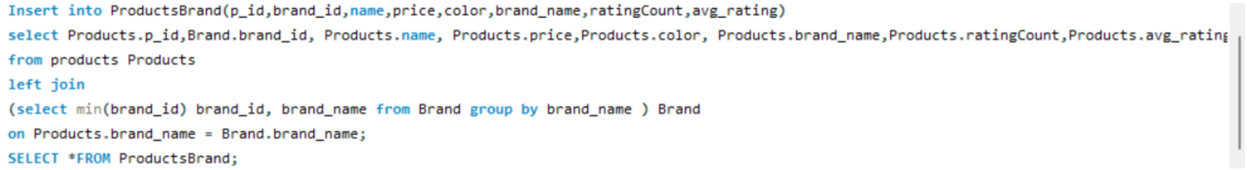


```sql
UPDATE Brand SET brand_name = UPPER(brand_name);
SELECT *FROM Brand;
UPDATE Products SET brand_name = UPPER(brand_name);
UPDATE Products SET price = UPPER(price);
UPDATE Products SET brand_name = UPPER(name);

ALTER TABLE Products RENAME COLUMN brand TO brand_name;
```

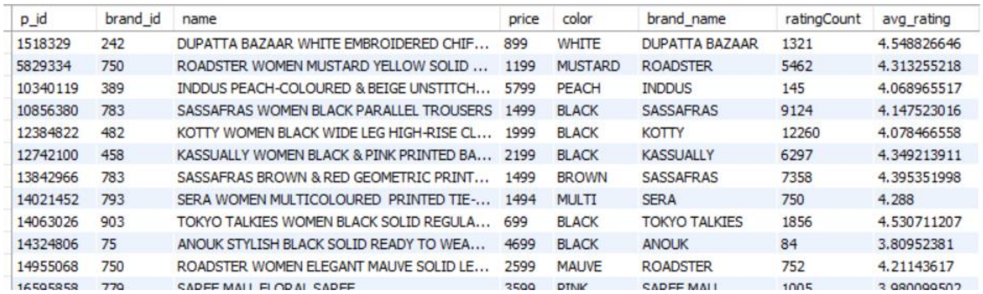
Figure 2.11: Changing into uppercase and renaming the attribute

As last two attributes which are p_attributes and description are in Json format so those columns are removed before the merging step.

2.3.2 Merging the dataset in SQL



```sql
Insert into ProductsBrand(p_id,brand_id,name,price,color,brand_name,ratingCount,avg_rating)
select Products.p_id,Brand.brand_id, Products.name, Products.price,Products.color, Products.brand_name,Products.ratingCount,Products.avg_rating
from products Products
left join
(select min(brand_id) brand_id, brand_name from Brand group by brand_name ) Brand
on Products.brand_name = Brand.brand_name;
SELECT *FROM ProductsBrand;
```

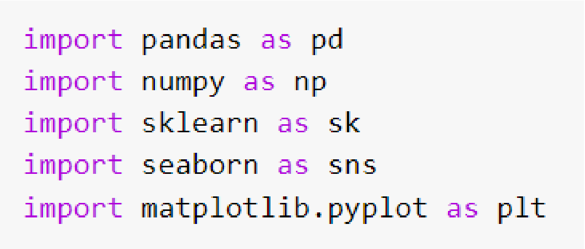


p_id	brand_id	name	price	color	brand_name	ratingCount	avg_rating
1518329	242	DUPATTA BAZAAR WHITE EMBROIDERED CHIF...	899	WHITE	DUPATTA BAZAAR	1321	4.548826646
5829334	750	ROADSTER WOMEN MUSTARD YELLOW SOLID ...	1199	MUSTARD	ROADSTER	5462	4.313255218
10340119	389	INDDUS PEACH-COLOURED & BEIGE UNSTITCH...	5799	PEACH	INDDUS	145	4.068965517
10856380	783	SASSAFRAS WOMEN BLACK PARALLEL TROUSERS	1499	BLACK	SASSAFRAS	9124	4.147523016
12384822	482	KOTTY WOMEN BLACK WIDE LEG HIGH-RISE CL...	1999	BLACK	KOTTY	12260	4.078466558
12742100	458	KASSUALLY WOMEN BLACK & PINK PRINTED BA...	2199	BLACK	KASSUALLY	6297	4.349213911
13842966	783	SASSAFRAS BROWN & RED GEOMETRIC PRINT...	1499	BROWN	SASSAFRAS	7358	4.395351998
14021452	793	SERA WOMEN MULTICOLOURED PRINTED TIE-...	1494	MULTI	SERA	750	4.288
14063026	903	TOKYO TALKIES WOMEN BLACK SOLID REGULA...	699	BLACK	TOKYO TALKIES	1856	4.530711207
14324806	75	ANOUK STYLISH BLACK SOLID READY TO WEA...	4699	BLACK	ANOUK	84	3.80952381
14955068	750	ROADSTER WOMEN ELEGANT MAUVE SOLID LE...	2599	MAUVE	ROADSTER	752	4.21143617
16595858	779	SAREE MALL FLORAL SAREE	3599	PINK	SAREE MALL	1005	3.980099502

Figure 2.12: Merging the two datasets and forming new dataset (ProductsBrand)

2.3.3 Loading Library for Data Preparation

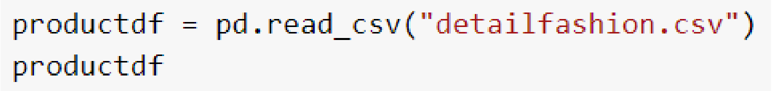
These are the libraries used for the data preparation of the data.



```python
import pandas as pd
import numpy as np
import sklearn as sk
import seaborn as sns
import matplotlib.pyplot as plt
```

Figure 2.13: Libraries that are used for Data Preprocessing

Loading data



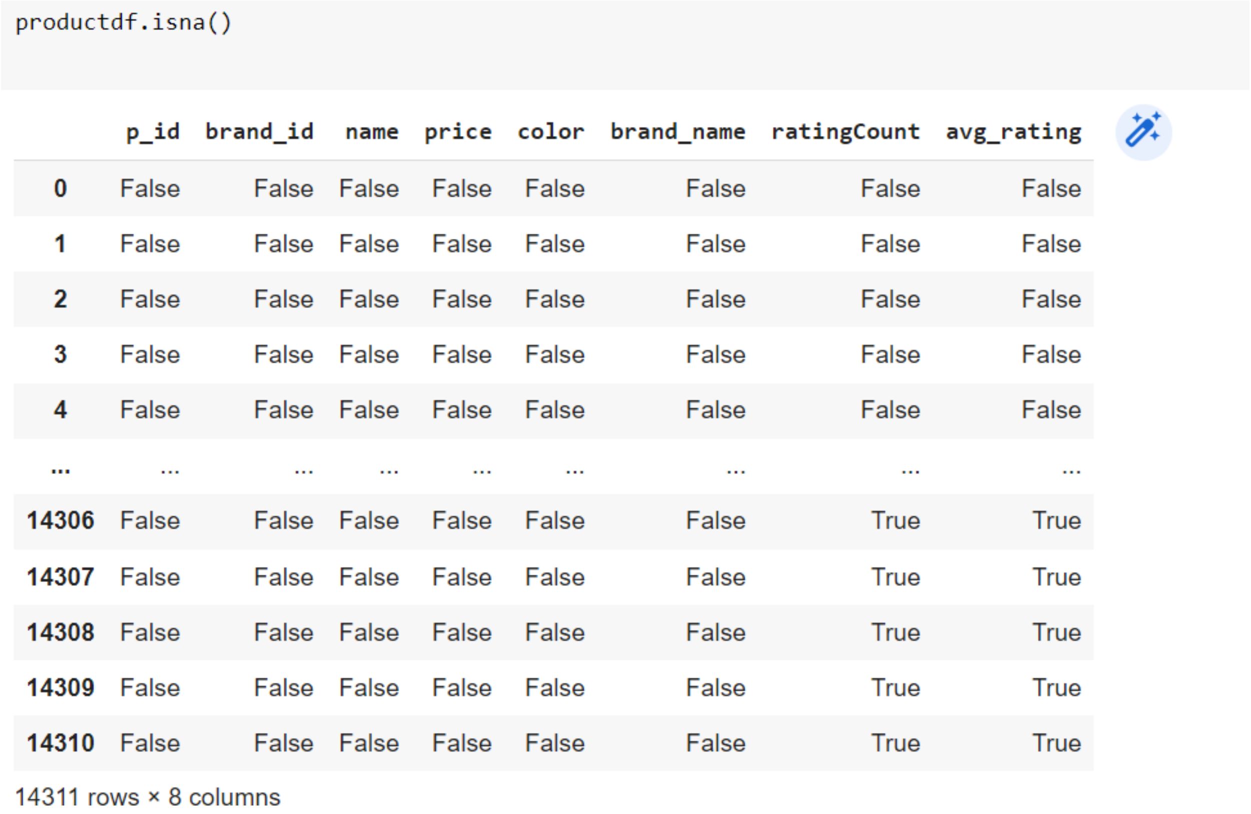
```
productdf = pd.read_csv("detailfashion.csv")
productdf
```

Figure 2.14: Loading the new dataset into new dataframe productdf

As a part of data preprocessing Data Cleaning steps is performed. It is an important early step in data analytics process which remove incorrect, duplicate, redundant entries, incorrect format data and so on. This step helps to improve the data quality and also better business decision. Data cleaning process varies according to the dataset and organizational needs. However, following are the steps that are carried out for this project.

2.3.4 Missing Values Detection

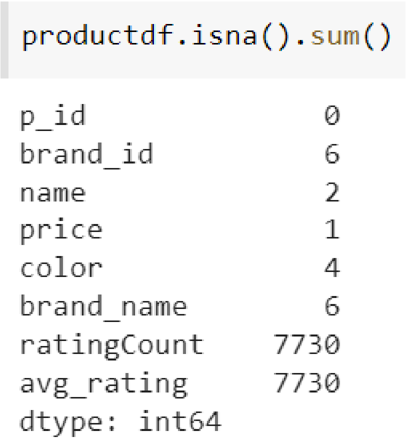
Missing value detection is one of the step-in data cleanings as those values are error because they do not represent the true value. So, we need to consider missing value because it will help to find the type of missing value and what to do for that.



```
productdf.isna()
```

	p_id	brand_id	name	price	color	brand_name	ratingCount	avg_rating
0	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...
14306	False	False	False	False	False	False	True	True
14307	False	False	False	False	False	False	True	True
14308	False	False	False	False	False	False	True	True
14309	False	False	False	False	False	False	True	True
14310	False	False	False	False	False	False	True	True

14311 rows × 8 columns

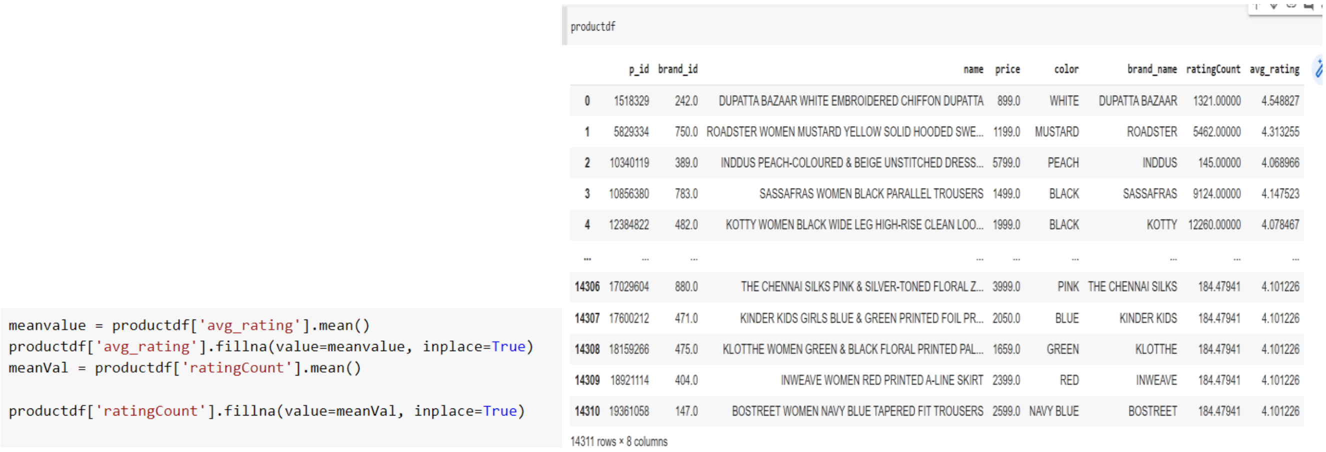


```
productdf.isna().sum()
```

```
p_id              0
brand_id          6
name              2
price             1
color             4
brand_name        6
ratingCount    7730
avg_rating     7730
dtype: int64
```

Figure 2.15 Null value check in the dataframe productdf

It shows that there are 7730 missing values in ratingCount and avg_rating which is huge amount of missing data so for the further processing NAN is replaced with mean value

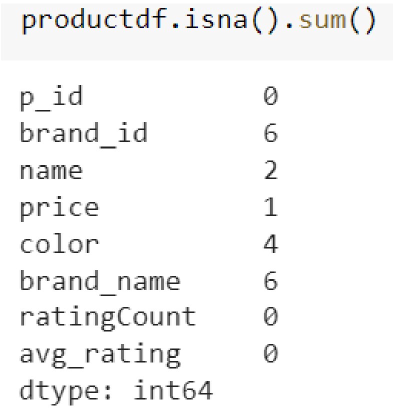


```
meanvalue = productdf['avg_rating'].mean()
productdf['avg_rating'].fillna(value=meanvalue, inplace=True)
meanVal = productdf['ratingCount'].mean()

productdf['ratingCount'].fillna(value=meanVal, inplace=True)
```

```
productdf
```

	p_id	brand_id	name	price	color	brand_name	ratingCount	avg_rating
0	1518329	242.0	DUPATTA BAZAAR WHITE EMBROIDERED CHIFFON DUPATTA	899.0	WHITE	DUPATTA BAZAAR	1321.00000	4.548827
1	5829334	750.0	ROADSTER WOMEN MUSTARD YELLOW SOLID HOODED SWE...	1199.0	MUSTARD	ROADSTER	5462.00000	4.313255
2	10340119	389.0	INDDUS PEACH-COLOURED & BEIGE UNSTITCHED DRESS...	5799.0	PEACH	INDDUS	145.00000	4.068966
3	10856380	783.0	SASSAFRAS WOMEN BLACK PARALLEL TROUSERS	1499.0	BLACK	SASSAFRAS	9124.00000	4.147523
4	12384822	482.0	KOTTY WOMEN BLACK WIDE LEG HIGH-RISE CLEAN LOO...	1999.0	BLACK	KOTTY	12260.00000	4.078467
...	...	...	...	...	...	...	...	...
14306	17029604	880.0	THE CHENNAI SILKS PINK & SILVER-TONED FLORAL Z...	3999.0	PINK	THE CHENNAI SILKS	184.47941	4.101226
14307	17600212	471.0	KINDER KIDS GIRLS BLUE & GREEN PRINTED FOIL PR...	2050.0	BLUE	KINDER KIDS	184.47941	4.101226
14308	18159266	475.0	KLOTTHE WOMEN GREEN & BLACK FLORAL PRINTED PAL...	1659.0	GREEN	KLOTTHE	184.47941	4.101226
14309	18921114	404.0	INWEAVE WOMEN RED PRINTED A-LINE SKIRT	2399.0	RED	INWEAVE	184.47941	4.101226
14310	19361058	147.0	BOSTREET WOMEN NAVY BLUE TAPERED FIT TROUSERS	2599.0	NAVY BLUE	BOSTREET	184.47941	4.101226

14311 rows × 8 columns

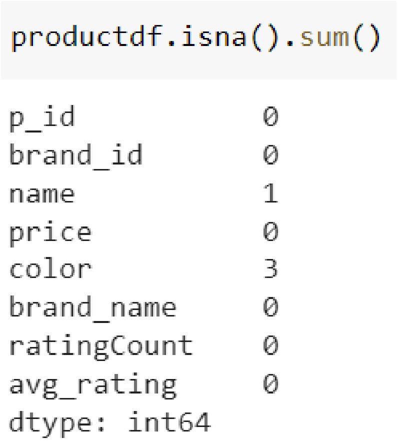


```
productdf.isna().sum()
```

```
p_id           0
brand_id       6
name           2
price          1
color          4
brand_name     6
ratingCount    0
avg_rating     0
dtype: int64
```

Figure 2.16 Replacing null value by mean value and checking the null value

It shows that there is 6 missing brand_id as brand it is always unique so with the help of price, color, brand_name it is not possible to find the value for it. So, I drop the product with brand_id NAN.



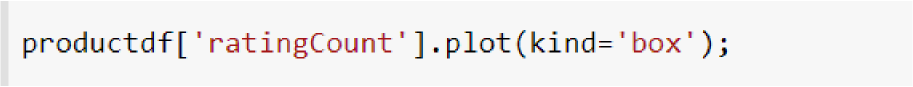
```
productdf.isna().sum()
```

```
p_id          0
brand_id      0
name          1
price         0
color         3
brand_name    0
ratingCount   0
avg_rating    0
dtype: int64
```

Figure 2.17 Checking whether null value is replaced or not

Now it shows there is only one name missing and 3 color which is not big amount of data missing. Now the data is ready for further processing.

2.3.5 Outlier detection



```
productdf['ratingCount'].plot(kind='box');
```

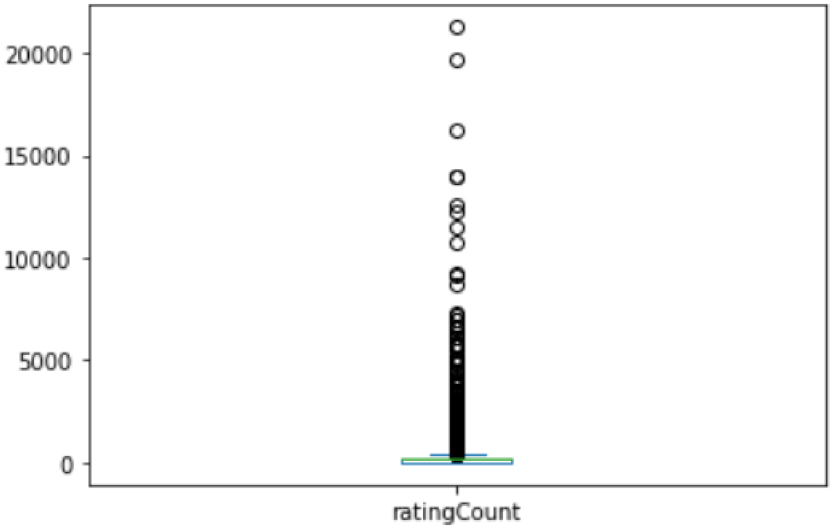
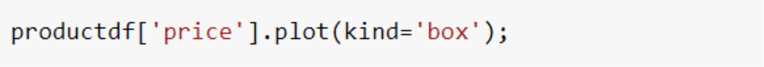


Figure 2.18 outlier check of rating Count



```
productdf['price'].plot(kind='box');
```

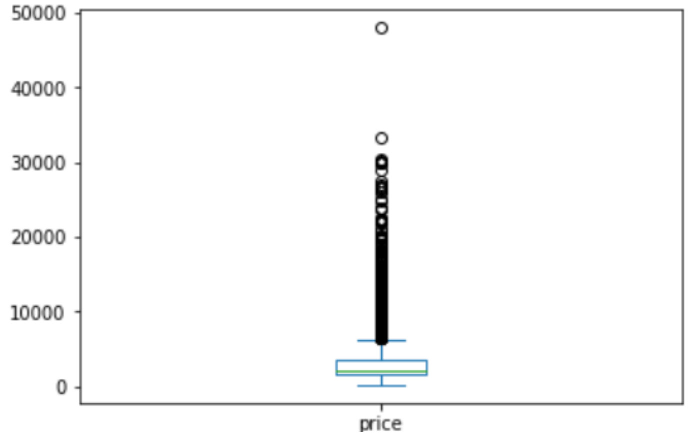
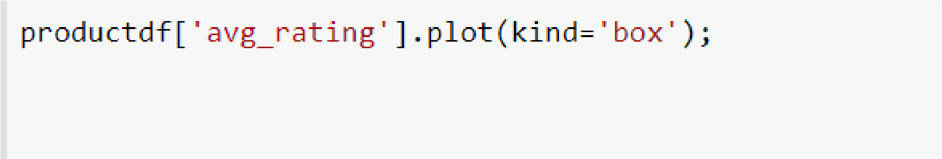


Figure 2.19 outlier check of price



```
productdf['avg_rating'].plot(kind='box');
```

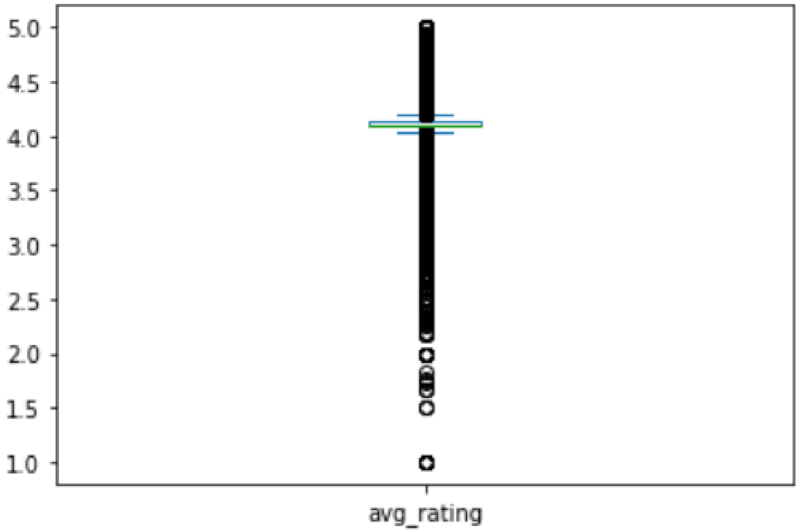
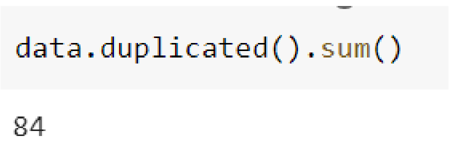


Figure 2.20 outlier check of avg_rating

As in above three figure it shows that there are outliers in avg_rating, ratingCount and price. The data used in this project is about fashion so people can rate the dress or brand according to their choice so removing or replacing those values with minimum value, maximum value or with quartile value will not be effective. Also, about the price the brand can be either most expensive or cheap in accordance to the brand color, name. So, in here there is no necessary to remove the outliers

2.3.6 Duplicate data

Identifying and removing duplicate data are one of the major parts in Data Cleaning.



```
data.duplicated().sum()
```

```
84
```

Figure 2.21 duplicate data check

Sum() method is used to count the number of true.

The above figure shows that there are 84 duplicate brand data in total so I have removed those data from the dataset.

2.3.7 Export data as csv from python

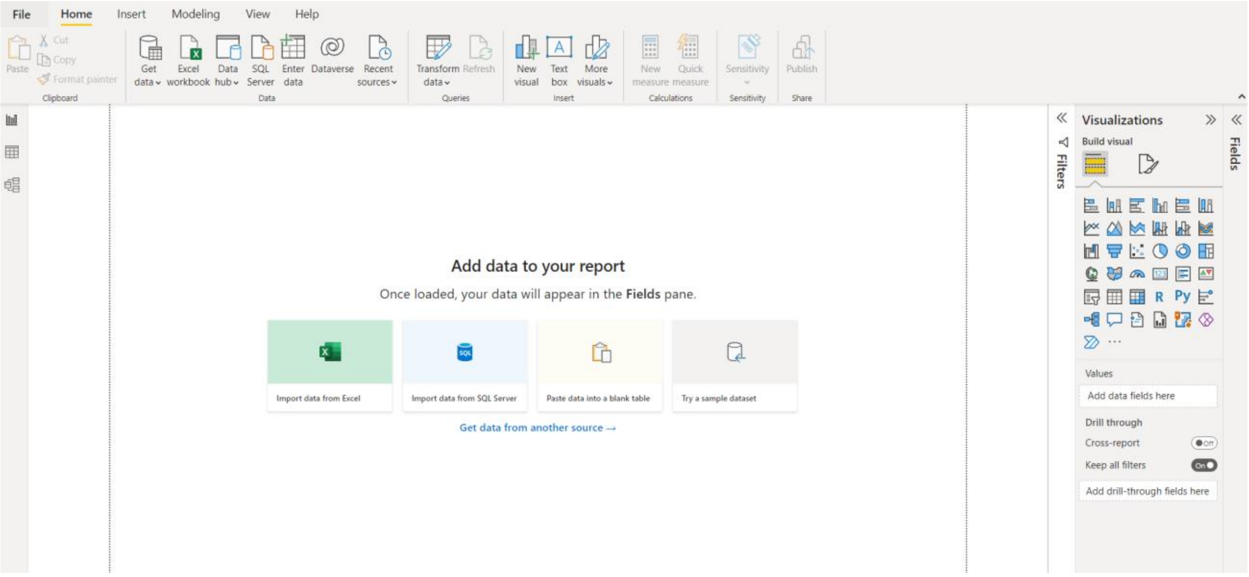
```
data.to_csv('branddetail.csv',index=False)
```

After this preprocessing step data is further used in Power Bi for data analysis and to get the insights from those data.

3 Data Visualization using Power BI

Microsoft Power BI is a Business Intelligence and interactive visualization software which was developed by Microsoft. It has developed most of the powerful business analytics tools within short period of time. This can be used by non-technical business users and study the data by plotting graphs, pie charts, graphs and so one. This helps to turn a dataset into interactive and visually immersive sights. Power Bi helps to connect disparate datasets.

The figure shown below is the dashboard of Powerbi where we import the data and different shapes are used to visualize those data. Also, different attributes, tag are present there to make it more readable and user interactive.



Add data to your report

Once loaded, your data will appear in the **Fields** pane.

Figure 3.1: Dataset loaded into PowerBi

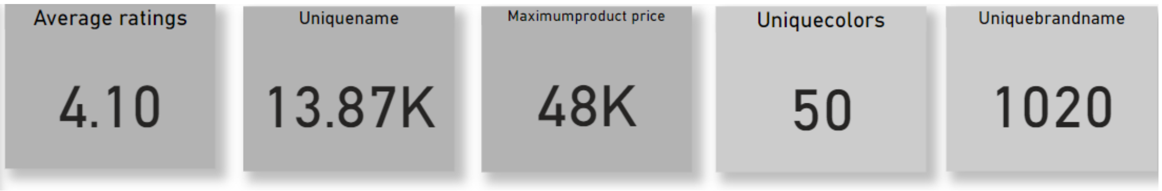


Figure 3.2: Displayed of avg_rating, uniquename, maximumproductprice,uniquecolors and uniquebrandname using card shape

The figure above shows the average of ag_rating, count of unique name, maximum price of a brand, total number of unique colors, and unique brand name which are the attributes in the datasets. These results are displayed using card.

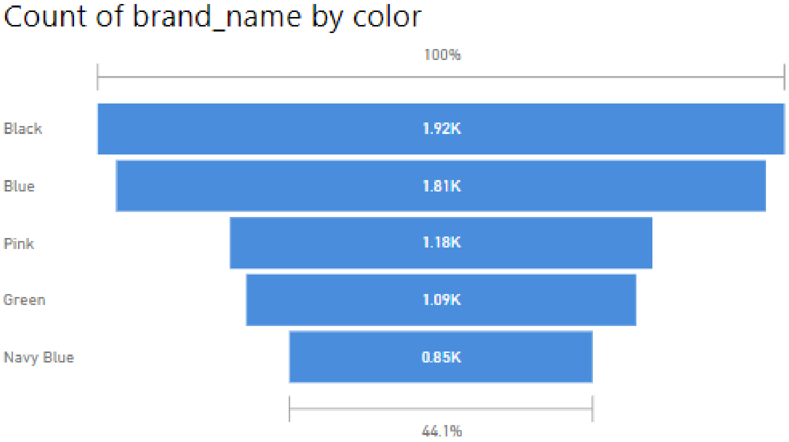


Figure 3.3: Counting the number of brand name

How many color does a brand name have is displayed as show in figure6.3 above by using funnel diagram. It showed the top 5 color and count the brand name in each color.

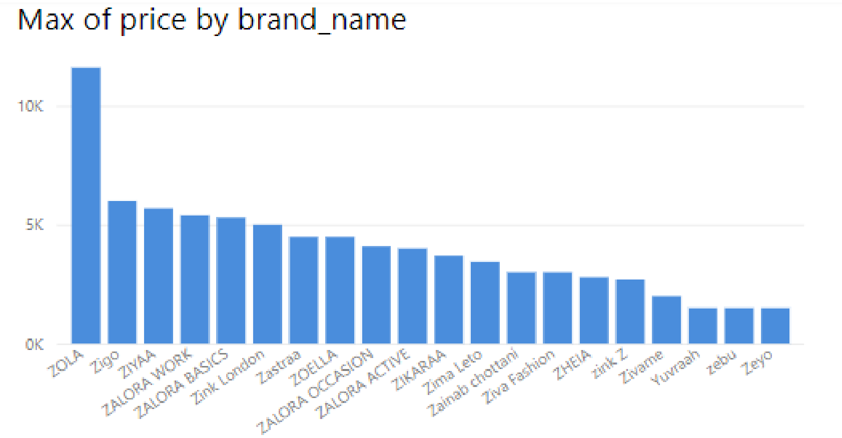


Figure 3.4: Display of expensive brand name

Figure 6.4: Top 20 brand name is displayed according to there maximum price which is shown in above figure. It is represented using stacked column chart.

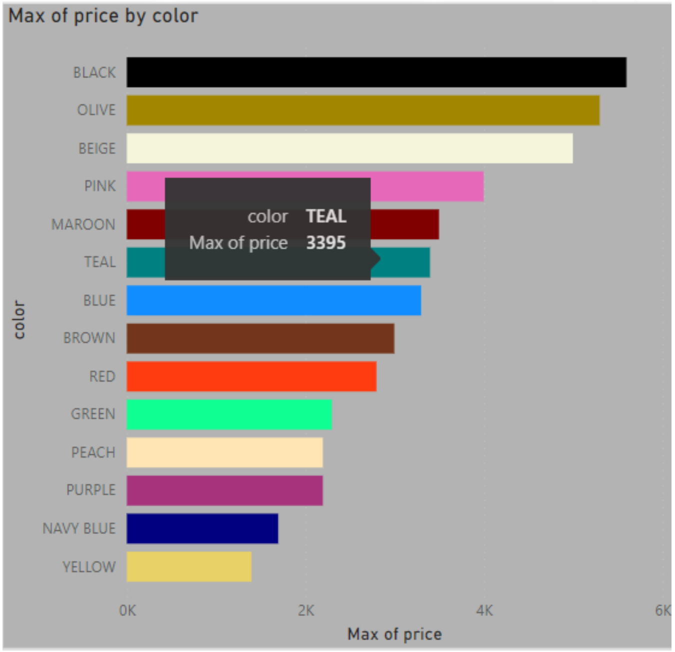


Figure:3.5: Displayed the maximum price based on color

From the above figure it can be analyzed that color can affect the price of a product. It can be seen that in compare to yellow black product is more expensive. The above diagram is visualized using stacked bar chart.

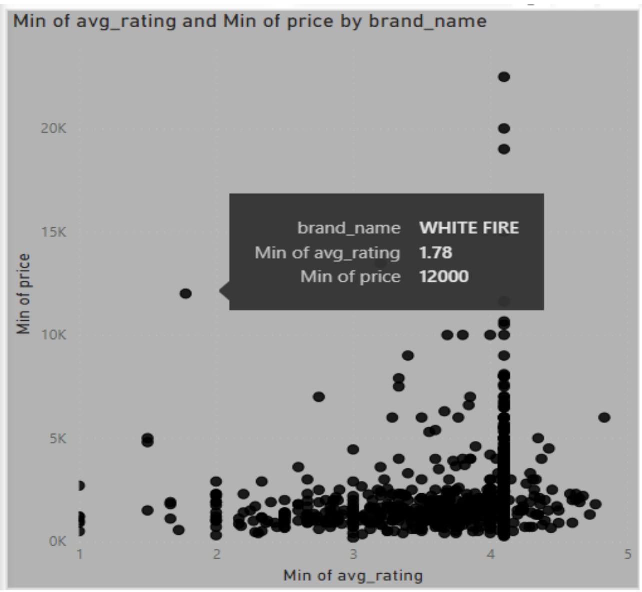


Figure 3.6: Shows the brand with low price and averagerating

A scatter chart is used to visualize the minimum price and avg_rating by brand name where each markers shows the brand name with low average rating and price as shown in above figure. You can analyze that by hover on the marker as it can be observed that brand name is white fire with 1.78 avg rating and 12000 is the minimum price

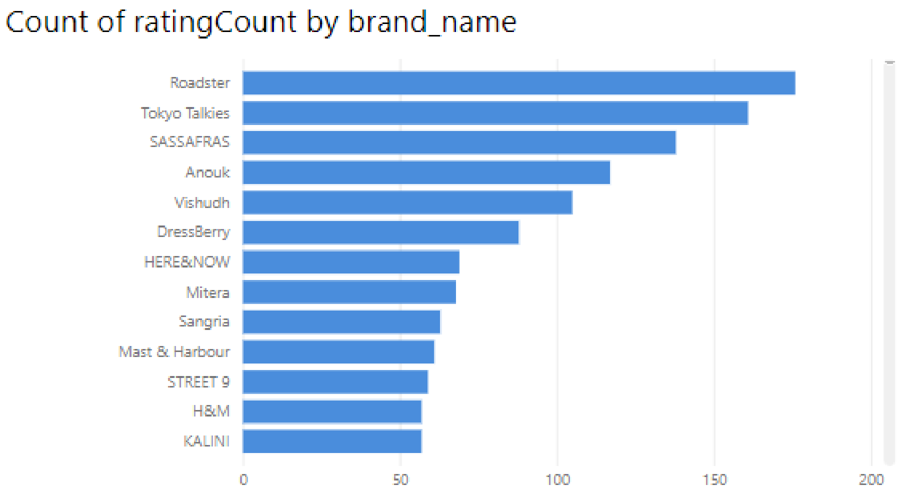
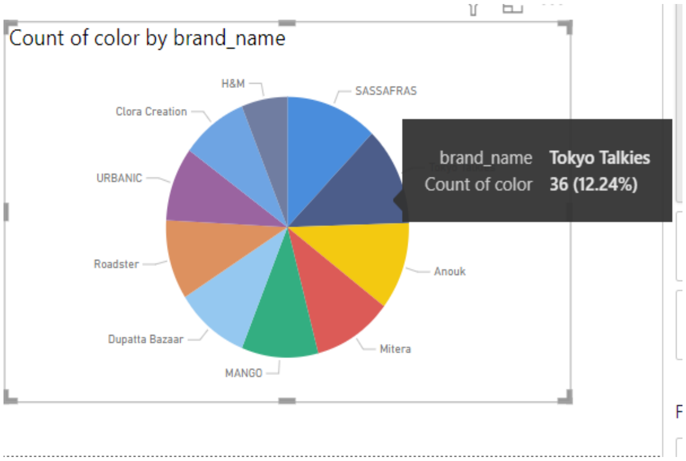


Figure 3.7: Number of same ratingcount based on brand_name

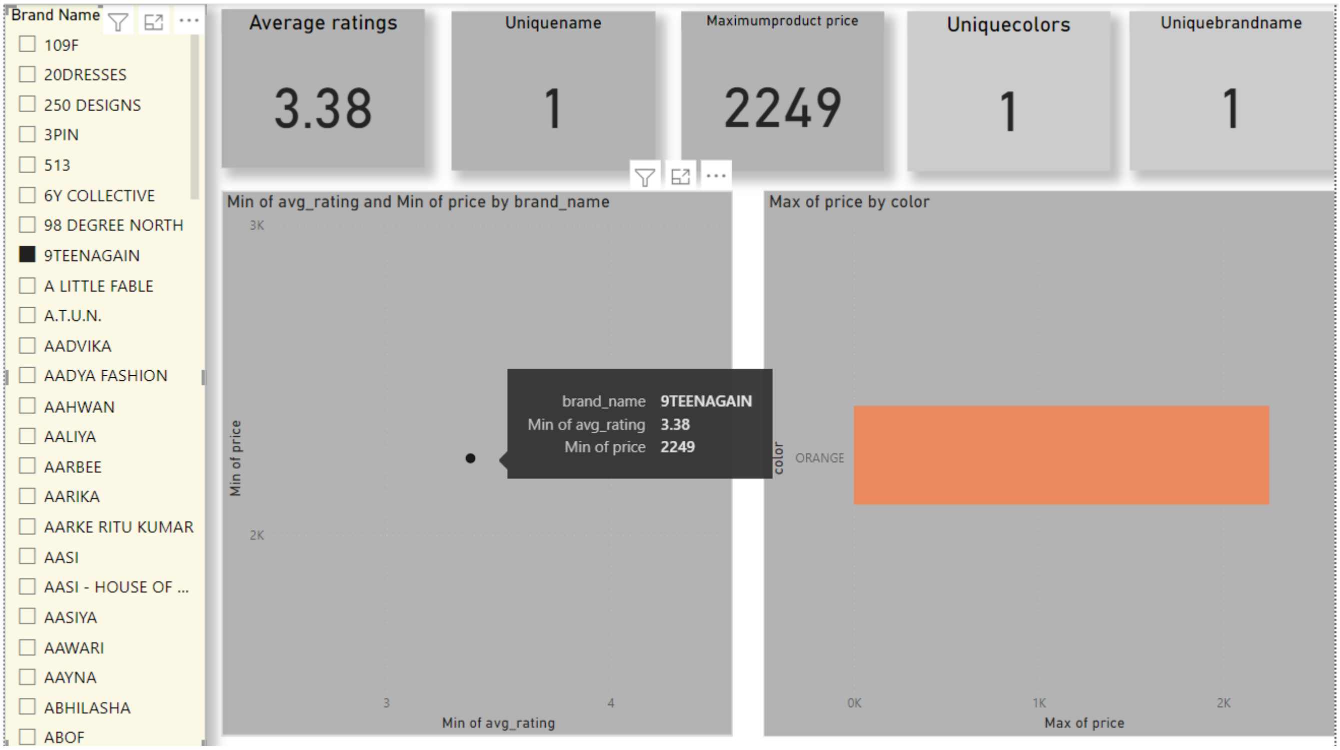
The figure above shows the rating count by brand name. By analyzing the figure, we can say that roadster brand name has more rating count in compare to Anouk, Kalimi, H &M and so on.



brand_name Tokyo Talkies
Count of color 36 (12.24%)

Figure 3.8:Brand name which used the same number of color is displayed

The figure above counts the number of color present in a brand like a brand name Tokyo Talkies have 36 blue color


Brand Name

- ☐ 109F
- ☐ 20DRESSES
- ☐ 250 DESIGNS
- ☐ 3PIN
- ☐ 513
- ☐ 6Y COLLECTIVE
- ☐ 98 DEGREE NORTH
- ■ 9TEENAGAIN
- ☐ A LITTLE FABLE
- ☐ A.T.U.N.
- ☐ AADVIKA
- ☐ AADYA FASHION
- ☐ AAHWAN
- ☐ AALIYA
- ☐ AARBEE
- ☐ AARIKA
- ☐ AARKE RITU KUMAR
- ☐ AASI
- ☐ AASI - HOUSE OF ...
- ☐ AASIYA
- ☐ AAWARI
- ☐ AAYNA
- ☐ ABHILASHA
- ☐ ABOF

Average ratings 3.38

Uniquename 1

Maximumproduct price 2249

Uniquecolors 1

Uniquebrandname 1

brand_name 9TEENAGAIN
Min of avg_rating 3.38
Min of price 2249

Figure 3.9 Working mechanism of different shape together and how they are correlated

Figure above is the visualization of brand_name with different other attributes. In powerbi we can use slider and other different chart together to analyze the report. Just like the figure shown you can see that when 9TEENAGAIN is clicked, minimum average rating is 3.38 and price is 2249. And the brand with orange color is most expensive. Like this way you can analyze and visualize your data so that it will be more understandable.

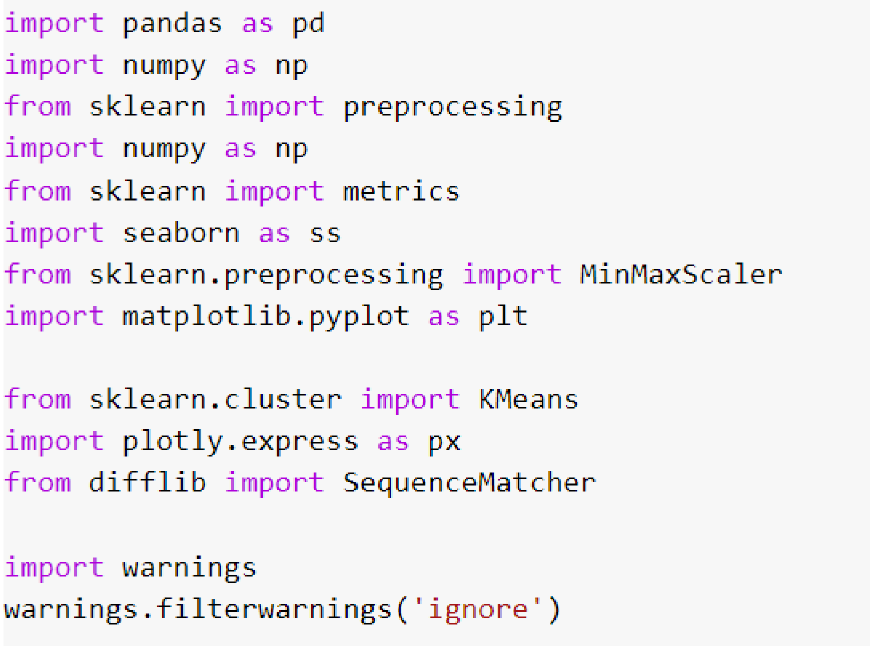
4 Data Modeling

4.1 Machine Learning

Machine learning is the process of data analysis which automates analytical model building. It is a part of artificial intelligence where system can learn from data, make decision, can identify patterns with the minimal intervention from human being. It helps to analyze big data which makes data scientist task easy which provides high value predictions and smart actions that helps in better decision without human intervention. Machine learning is important for data analysis because of its iterative aspect because every time model is exposed to new data which ML can adopt easily. Python, JavaScript, R, Java, C++ are some of the programming languages which can be used for Machine Learning.

In this project Google Colab is used which allows anyone to write and execute the python code through the browser.

4.2 Loading Libraries

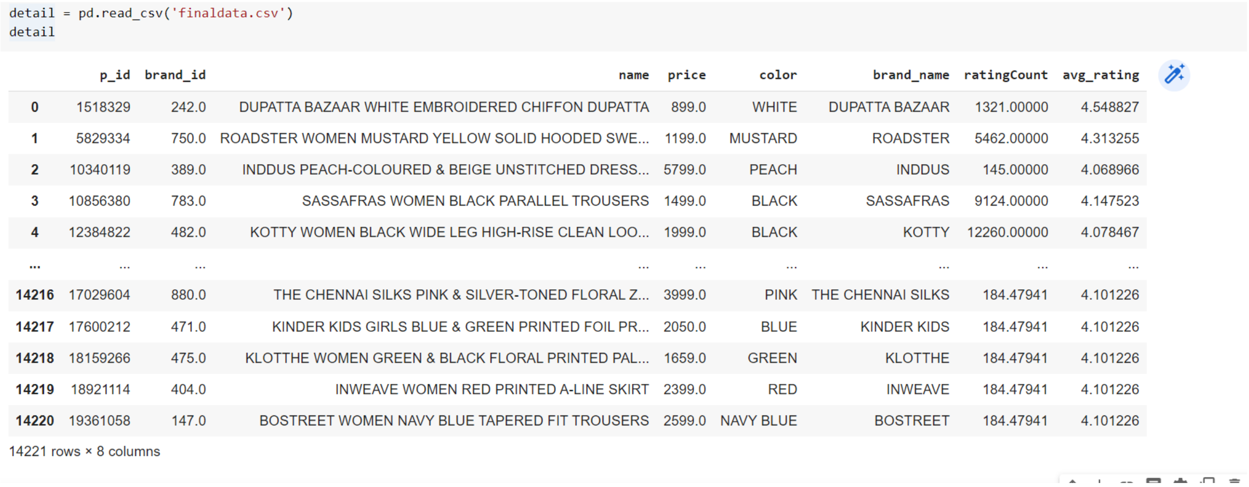


```python
import pandas as pd
import numpy as np
from sklearn import preprocessing
import numpy as np
from sklearn import metrics
import seaborn as ss
from sklearn.preprocessing import MinMaxScaler
import matplotlib.pyplot as plt

from sklearn.cluster import KMeans
import plotly.express as px
from difflib import SequenceMatcher

import warnings
warnings.filterwarnings('ignore')
```

Figure 4.1: Libraries that are used for Machine Learning implementation

Minmax Scaler is imported for Normalization process

4.3 Loading data



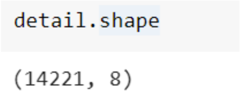
```
detail = pd.read_csv('finaldata.csv')
detail
```

	p_id	brand_id	name	price	color	brand_name	ratingCount	avg_rating
0	1518329	242.0	DUPATTA BAZAAR WHITE EMBROIDERED CHIFFON DUPATTA	899.0	WHITE	DUPATTA BAZAAR	1321.00000	4.548827
1	5829334	750.0	ROADSTER WOMEN MUSTARD YELLOW SOLID HOODED SWE...	1199.0	MUSTARD	ROADSTER	5462.00000	4.313255
2	10340119	389.0	INDDUS PEACH-COLOURED & BEIGE UNSTITCHED DRESS...	5799.0	PEACH	INDDUS	145.00000	4.068966
3	10856380	783.0	SASSAFRAS WOMEN BLACK PARALLEL TROUSERS	1499.0	BLACK	SASSAFRAS	9124.00000	4.147523
4	12384822	482.0	KOTTY WOMEN BLACK WIDE LEG HIGH-RISE CLEAN LOO...	1999.0	BLACK	KOTTY	12260.00000	4.078467
...	...	...	...	...	...	...	...	...
14216	17029604	880.0	THE CHENNAI SILKS PINK & SILVER-TONED FLORAL Z...	3999.0	PINK	THE CHENNAI SILKS	184.47941	4.101226
14217	17600212	471.0	KINDER KIDS GIRLS BLUE & GREEN PRINTED FOIL PR...	2050.0	BLUE	KINDER KIDS	184.47941	4.101226
14218	18159266	475.0	KLOTTHE WOMEN GREEN & BLACK FLORAL PRINTED PAL...	1659.0	GREEN	KLOTTHE	184.47941	4.101226
14219	18921114	404.0	INWEAVE WOMEN RED PRINTED A-LINE SKIRT	2399.0	RED	INWEAVE	184.47941	4.101226
14220	19361058	147.0	BOSTREET WOMEN NAVY BLUE TAPERED FIT TROUSERS	2599.0	NAVY BLUE	BOSTREET	184.47941	4.101226

14221 rows × 8 columns

Figure 4.2: Loading the preprocessed data in data frame detail

The preprocessed data is loaded to use machine learning algorithm in it.
shape of Dataset

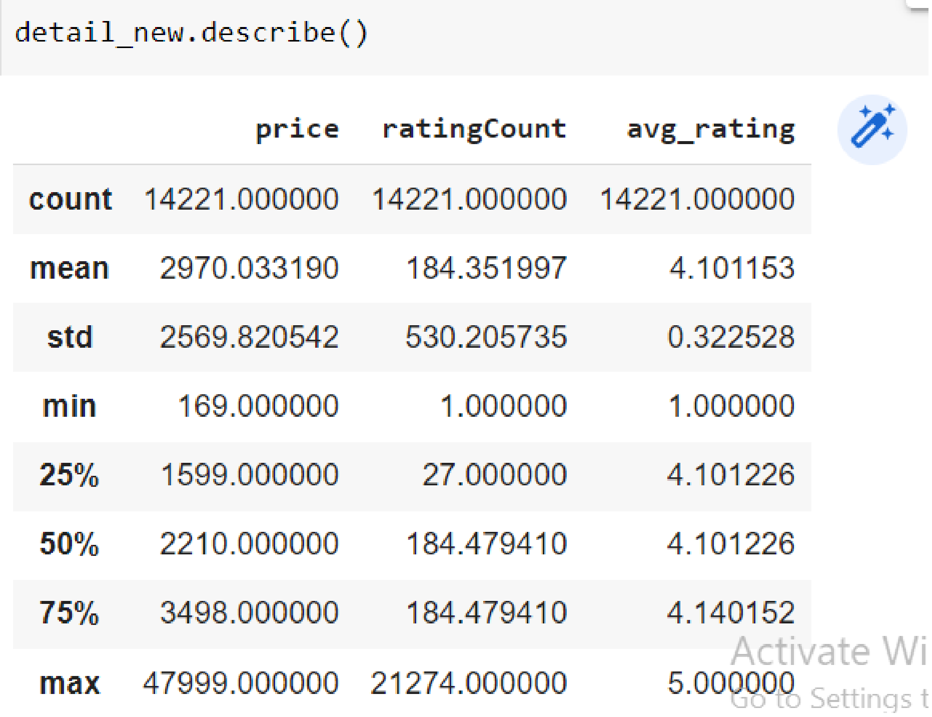


```
detail.shape
```

```
(14221, 8)
```

Figure 4.3: Shape of the preprocessed data

4.4 Descriptive Analysis



```
detail_new.describe()
```

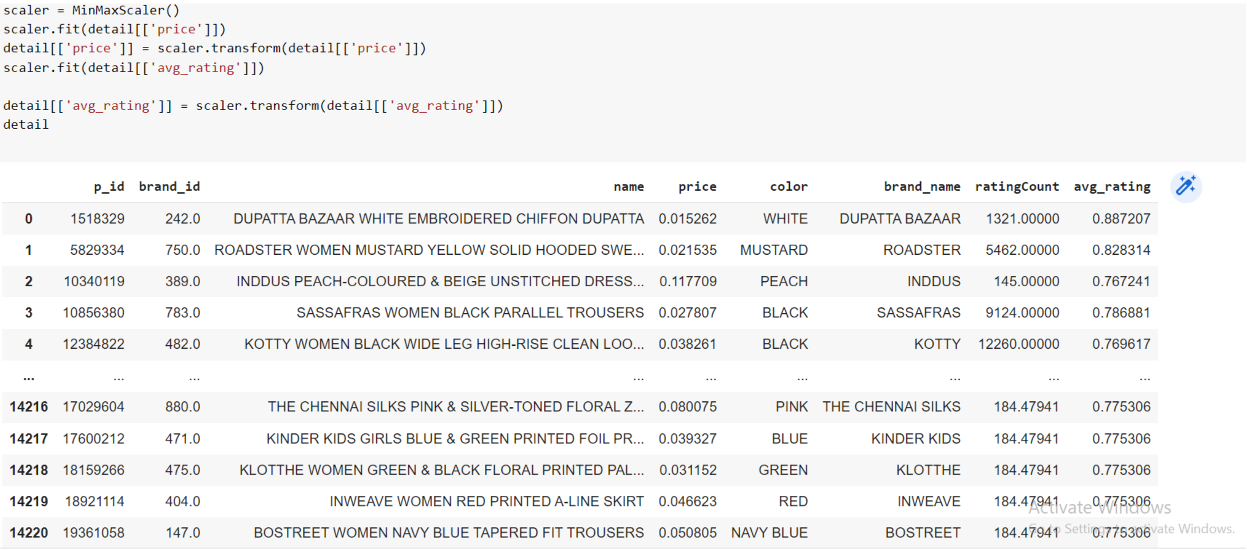
	price	ratingCount	avg_rating
count	14221.000000	14221.000000	14221.000000
mean	2970.033190	184.351997	4.101153
std	2569.820542	530.205735	0.322528
min	169.000000	1.000000	1.000000
25%	1599.000000	27.000000	4.101226
50%	2210.000000	184.479410	4.101226
75%	3498.000000	184.479410	4.140152
max	47999.000000	21274.000000	5.000000

Figure 4.4: Description about the loaded data

It shows that mean value of price is 2970.033190, of ratingCount is 184.351997, and average rating is 4.101153. Maximum value for price is 47999, maximum average rating is 5

4.5 Normalization

Normalization is carried out in this step because as you can see in figure above price and avg_rating have high difference in their values price is 10000 and the rating is 1, 2 and so on. So we need to carry out normalization to transform features so that it become on similar scale. This improves training stability and performance of the model.



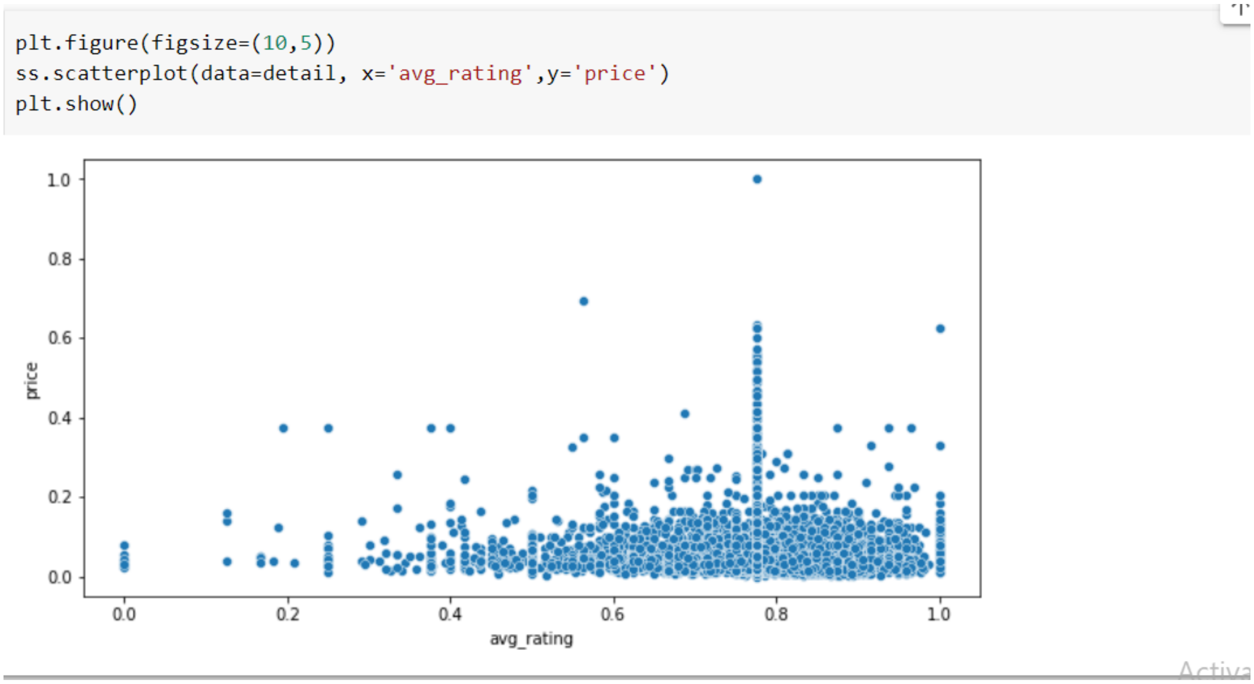
```
scaler = MinMaxScaler()
scaler.fit(detail[['price']])
detail[['price']] = scaler.transform(detail[['price']])
scaler.fit(detail[['avg_rating']])

detail[['avg_rating']] = scaler.transform(detail[['avg_rating']])
detail
```

	p_id	brand_id	name	price	color	brand_name	ratingCount	avg_rating
0	1518329	242.0	DUPATTA BAZAAR WHITE EMBROIDERED CHIFFON DUPATTA	0.015262	WHITE	DUPATTA BAZAAR	1321.00000	0.887207
1	5829334	750.0	ROADSTER WOMEN MUSTARD YELLOW SOLID HOODED SWE...	0.021535	MUSTARD	ROADSTER	5462.00000	0.828314
2	10340119	389.0	INDDUS PEACH-COLOURED & BEIGE UNSTITCHED DRESS...	0.117709	PEACH	INDDUS	145.00000	0.767241
3	10856380	783.0	SASSAFRAS WOMEN BLACK PARALLEL TROUSERS	0.027807	BLACK	SASSAFRAS	9124.00000	0.786881
4	12384822	482.0	KOTTY WOMEN BLACK WIDE LEG HIGH-RISE CLEAN LOO...	0.038261	BLACK	KOTTY	12260.00000	0.769617
...	...	...	...	...	...	...	...	...
14216	17029604	880.0	THE CHENNAI SILKS PINK & SILVER-TONED FLORAL Z...	0.080075	PINK	THE CHENNAI SILKS	184.47941	0.775306
14217	17600212	471.0	KINDER KIDS GIRLS BLUE & GREEN PRINTED FOIL PR...	0.039327	BLUE	KINDER KIDS	184.47941	0.775306
14218	18159266	475.0	KLOTTHE WOMEN GREEN & BLACK FLORAL PRINTED PAL...	0.031152	GREEN	KLOTTHE	184.47941	0.775306
14219	18921114	404.0	INWEAVE WOMEN RED PRINTED A-LINE SKIRT	0.046623	RED	INWEAVE	184.47941	0.775306
14220	19361058	147.0	BOSTREET WOMEN NAVY BLUE TAPERED FIT TROUSERS	0.050805	NAVY BLUE	BOSTREET	184.47941	0.775306

Figure 4.5: Normalization of avg_rating and price before using K mean algorithm

4.6 Data Visualization

For Data Visualization in x axis avg_rating and for y axis price is allocated.



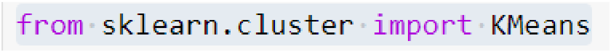
```
plt.figure(figsize=(10,5))
ss.scatterplot(data=detail, x='avg_rating',y='price')
plt.show()
```

Figure 4.6: Plot of avg_rating and price

4.7 Implementation of KMean Algorithm

K mean Algorithm is iterative algorithm or unsupervised learning which is used to solve the classification problems. It segregates the unlabeled data into various group (k subgroups) known as clusters in accordance to the presence of similar features or patterns and mean distance from centroid of formed clusters.

Clustering is the process of dividing the data points into number of same or similar groups. Basically, it is the collection of objects on the basis of its similarity and dissimilarity in unlabeled data (Pham et al., 2005).



```
from sklearn.cluster import KMeans
```

Figure 4.6 Library which is used for K Mean Algorithm

This is the library which needs to be imported to execute K mean algorithm.

4.7.1 Features Extract

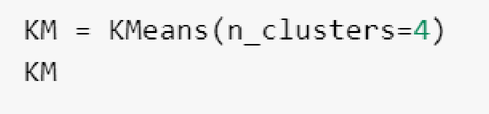
After that I have taken avg_rating and price as the main attribute for further processing

```
f_extract = detail[['avg_rating','price']]
```

Figure 4.7: Attributes which before fit into model

4.7.2 Model

Model in machine learning is a file which is trained to find certain patterns or to make decision from unseen dataset.



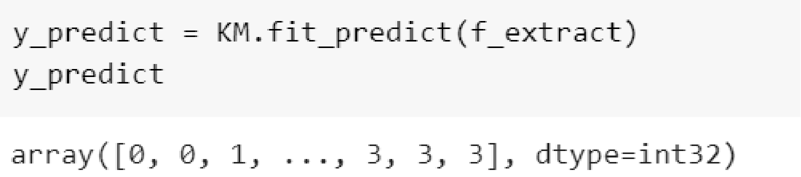
```
KM = KMeans(n_clusters=4)
KM
```

Figure 4.8: Cluster formation

I have assigned k value 4 as an assumption for the number of clusters to be formed. Cluster is defined as the collection of data with different or similar characteristics

4.7.3 Fit to Train

Next step is to fit the avg_rating and price features into the model where the data is store after the process is storage in y_predict where fit_predict() is the function used.

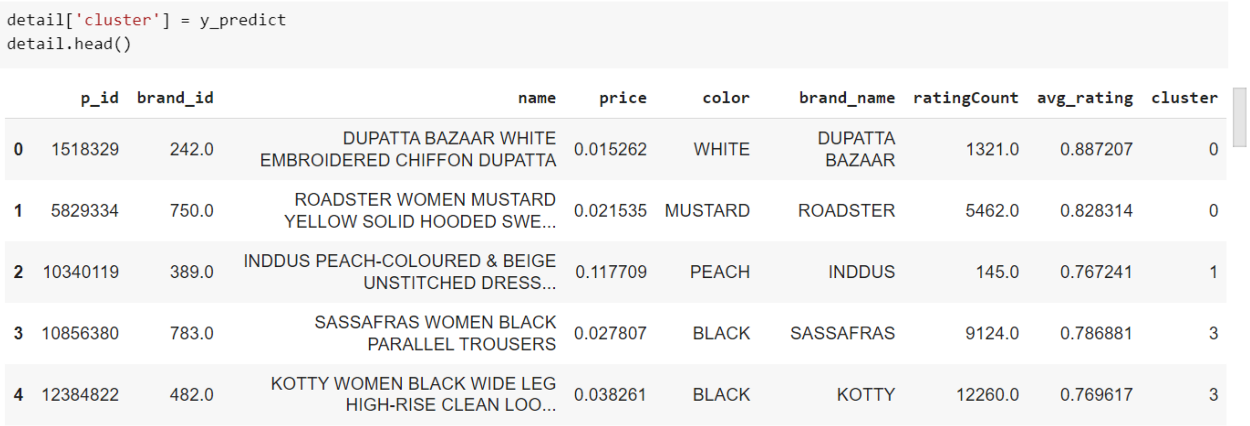


```
y_predict = KM.fit_predict(f_extract)
y_predict
```

```
array([0, 0, 1, ..., 3, 3, 3], dtype=int32)
```

Figure 4.9: Fit and train data and result is stored in y_predict

4.7.4 Cluster



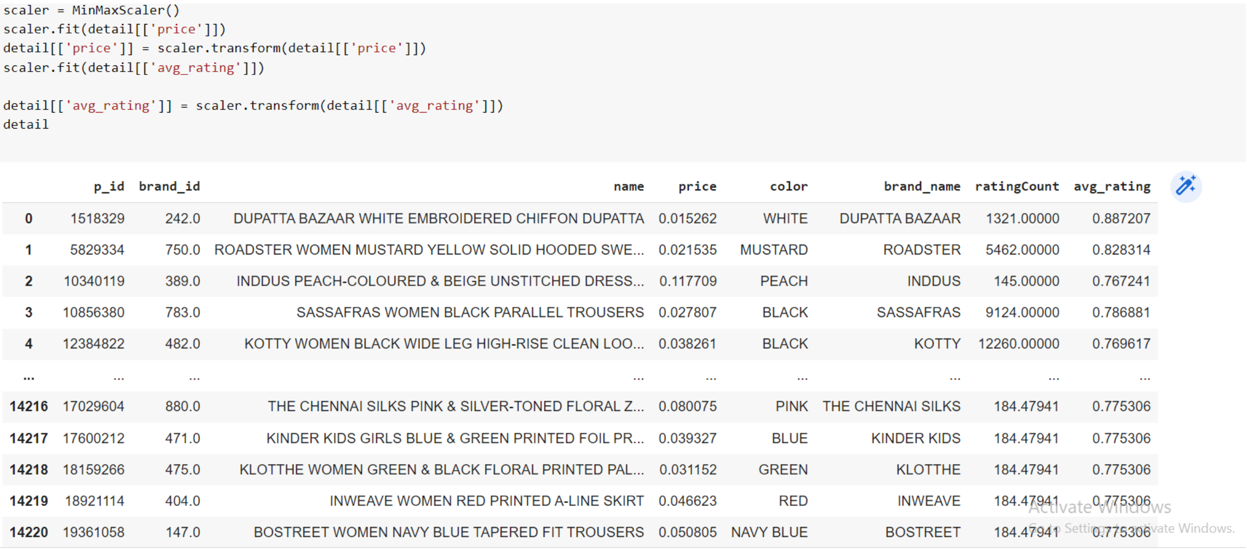
```
detail['cluster'] = y_predict
detail.head()
```

	p_id	brand_id	name	price	color	brand_name	ratingCount	avg_rating	cluster
0	1518329	242.0	DUPATTA BAZAAR WHITE EMBROIDERED CHIFFON DUPATTA	0.015262	WHITE	DUPATTA BAZAAR	1321.0	0.887207	0
1	5829334	750.0	ROADSTER WOMEN MUSTARD YELLOW SOLID HOODED SWE...	0.021535	MUSTARD	ROADSTER	5462.0	0.828314	0
2	10340119	389.0	INDDUS PEACH-COLOURED & BEIGE UNSTITCHED DRESS...	0.117709	PEACH	INDDUS	145.0	0.767241	1
3	10856380	783.0	SASSAFRAS WOMEN BLACK PARALLEL TROUSERS	0.027807	BLACK	SASSAFRAS	9124.0	0.786881	3
4	12384822	482.0	KOTTY WOMEN BLACK WIDE LEG HIGH-RISE CLEAN LOO...	0.038261	BLACK	KOTTY	12260.0	0.769617	3

Figure 4.10: Result displayed about cluster formation

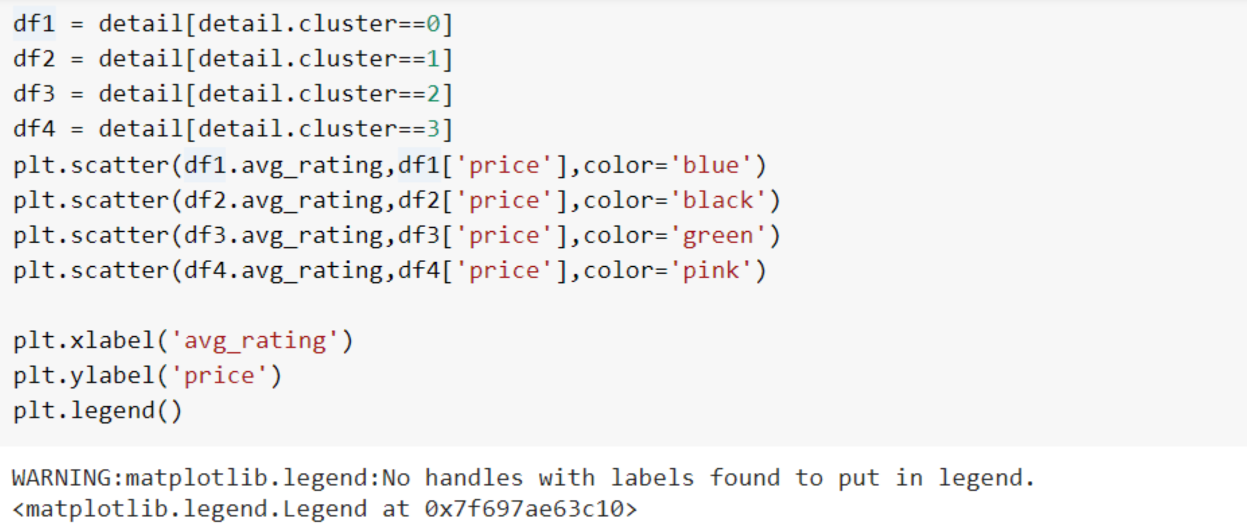
From above you can analyze that avg_rating 0.887202, 0.828314 have similar feature so they are cluster 0 And 4.068966 falls under cluster 3. Meanwhile price of 0.015262,0.021532 falls under 0 cluster.

4.7.5 Visualization Of clusters



```
scaler = MinMaxScaler()
scaler.fit(detail[['price']])
detail[['price']] = scaler.transform(detail[['price']])
scaler.fit(detail[['avg_rating']])

detail[['avg_rating']] = scaler.transform(detail[['avg_rating']])
detail
```

	p_id	brand_id	name	price	color	brand_name	ratingCount	avg_rating
0	1518329	242.0	DUPATTA BAZAAR WHITE EMBROIDERED CHIFFON DUPATTA	0.015262	WHITE	DUPATTA BAZAAR	1321.00000	0.887207
1	5829334	750.0	ROADSTER WOMEN MUSTARD YELLOW SOLID HOODED SWE...	0.021535	MUSTARD	ROADSTER	5462.00000	0.828314
2	10340119	389.0	INDDUS PEACH-COLOURED & BEIGE UNSTITCHED DRESS...	0.117709	PEACH	INDDUS	145.00000	0.767241
3	10856380	783.0	SASSAFRAS WOMEN BLACK PARALLEL TROUSERS	0.027807	BLACK	SASSAFRAS	9124.00000	0.786881
4	12384822	482.0	KOTTY WOMEN BLACK WIDE LEG HIGH-RISE CLEAN LOO...	0.038261	BLACK	KOTTY	12260.00000	0.769617
...	...	...	...	...	...	...	...	...
14216	17029604	880.0	THE CHENNAI SILKS PINK & SILVER-TONED FLORAL Z...	0.080075	PINK	THE CHENNAI SILKS	184.47941	0.775306
14217	17600212	471.0	KINDER KIDS GIRLS BLUE & GREEN PRINTED FOIL PR...	0.039327	BLUE	KINDER KIDS	184.47941	0.775306
14218	18159266	475.0	KLOTTHE WOMEN GREEN & BLACK FLORAL PRINTED PAL...	0.031152	GREEN	KLOTTHE	184.47941	0.775306
14219	18921114	404.0	INWEAVE WOMEN RED PRINTED A-LINE SKIRT	0.046623	RED	INWEAVE	184.47941	0.775306
14220	19361058	147.0	BOSTREET WOMEN NAVY BLUE TAPERED FIT TROUSERS	0.050805	NAVY BLUE	BOSTREET	184.47941	0.775306



```
df1 = detail[detail.cluster==0]
df2 = detail[detail.cluster==1]
df3 = detail[detail.cluster==2]
df4 = detail[detail.cluster==3]
plt.scatter(df1.avg_rating,df1['price'],color='blue')
plt.scatter(df2.avg_rating,df2['price'],color='black')
plt.scatter(df3.avg_rating,df3['price'],color='green')
plt.scatter(df4.avg_rating,df4['price'],color='pink')

plt.xlabel('avg_rating')
plt.ylabel('price')
plt.legend()
```

```
WARNING:matplotlib.legend:No handles with labels found to put in legend.
<matplotlib.legend.Legend at 0x7f697ae63c10>
```

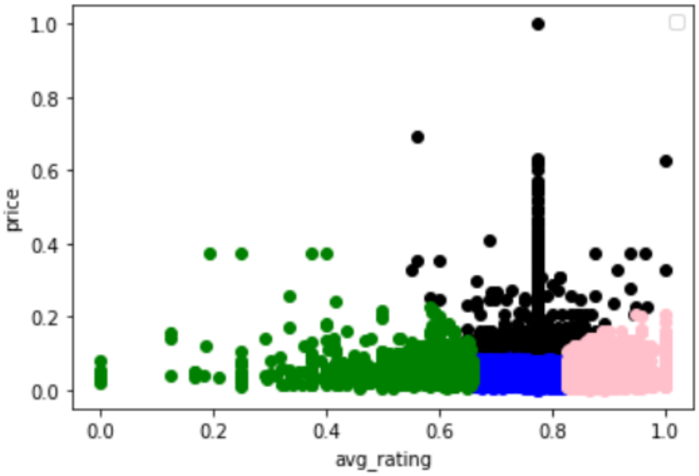
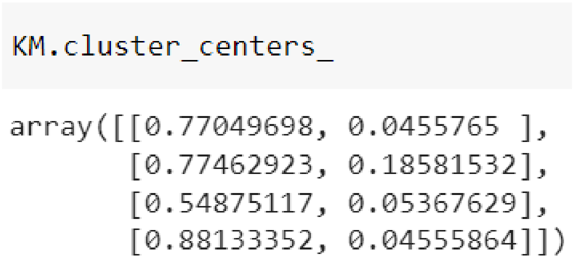


Figure: 4.11: Visualization after cluster formation using scatter plot

4.7.6 Cluster Centers

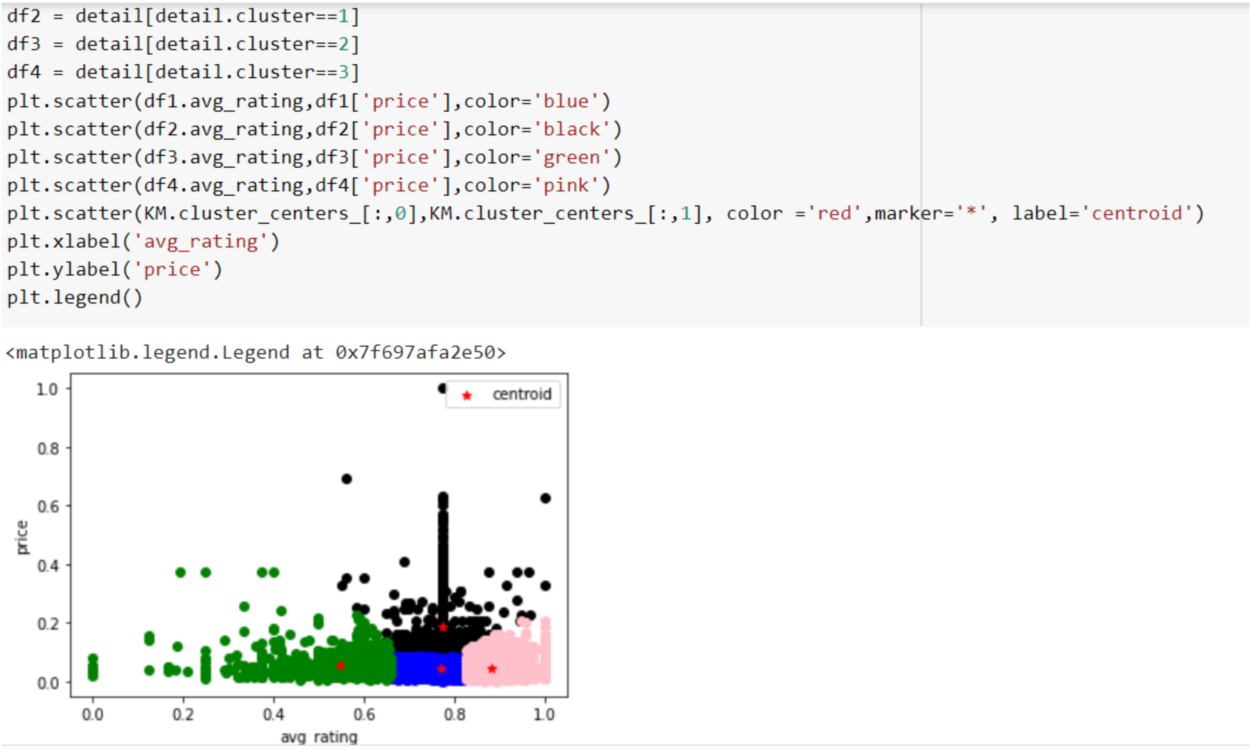
Cluster center is the mean of all the points present to that cluster. Each cluster is represented by its center which is called centroid (the mean)



```
KM.cluster_centers_
```

```
array([[0.77049698, 0.0455765 ],
       [0.77462923, 0.18581532],
       [0.54875117, 0.05367629],
       [0.88133352, 0.04555864]])
```

Visualization of center



```
df2 = detail[detail.cluster==1]
df3 = detail[detail.cluster==2]
df4 = detail[detail.cluster==3]
plt.scatter(df1.avg_rating,df1['price'],color='blue')
plt.scatter(df2.avg_rating,df2['price'],color='black')
plt.scatter(df3.avg_rating,df3['price'],color='green')
plt.scatter(df4.avg_rating,df4['price'],color='pink')
plt.scatter(KM.cluster_centers_[:,0],KM.cluster_centers_[:,1], color ='red',marker='*', label='centroid')
plt.xlabel('avg_rating')
plt.ylabel('price')
plt.legend()
```

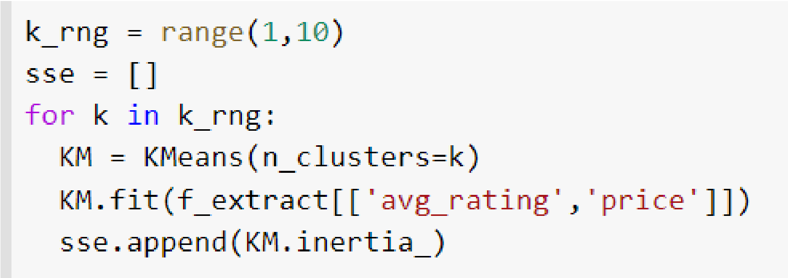
```
<matplotlib.legend.Legend at 0x7f697afa2e50>
```

Figure 4.12: Calculation of cluster centers and visualization using scatter plot

As you can see from above figure centroid is represented by red start symbol which is in the middle of each cluster respectively.

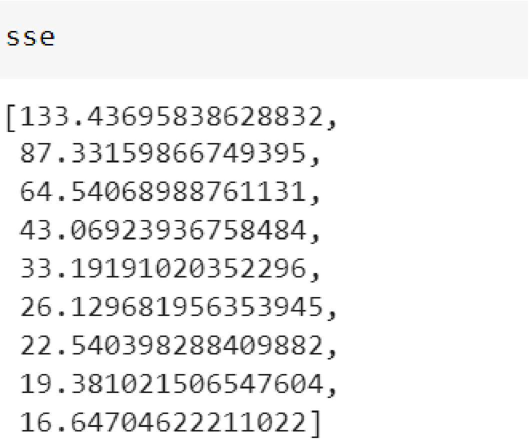
4.7.7 SSE check

Sum of the squared error sum of squared Euclidean distance of each point to its closest centroid. Inertia plot is used to represent the Number of clusters in X axis and SSE values in Y axis



```
k_rng = range(1,10)
sse = []
for k in k_rng:
  KM = KMeans(n_clusters=k)
  KM.fit(f_extract[['avg_rating','price']])
  sse.append(KM.inertia_)
```

SSE values



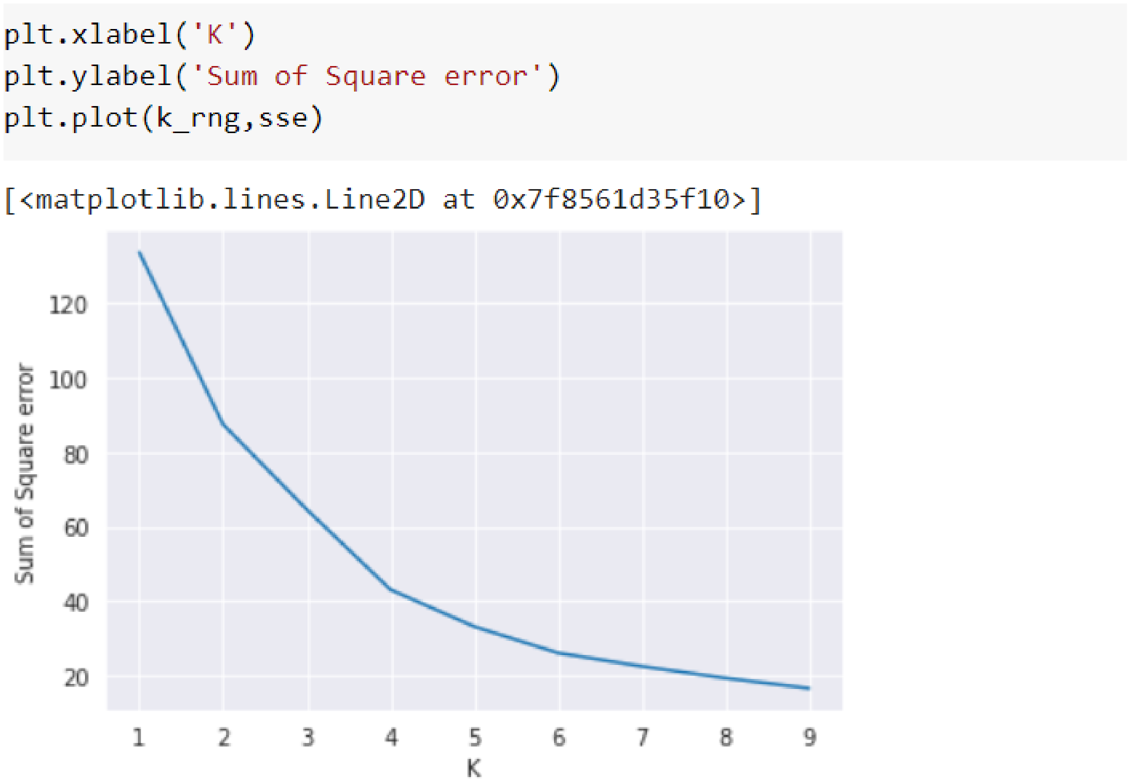
```
sse
```

```
[133.43695838628832,
 87.33159866749395,
 64.54068988761131,
 43.06923936758484,
 33.19191020352296,
 26.129681956353945,
 22.540398288409882,
 19.381021506547604,
 16.64704622211022]
```

Figure 4.13: Calculation of sum of squared error and displayed result

4.7.8 Elbow method

Elbow method is the graphical representation to find the optimal k during k mean clustering. It is used to find the sum of square between centroid of cluster and cluster. The elbow graph represents the Sum of square error on y axis in correspond to values of K on x axis.



```
plt.xlabel('K')
plt.ylabel('Sum of Square error')
plt.plot(k_rng,sse)
```

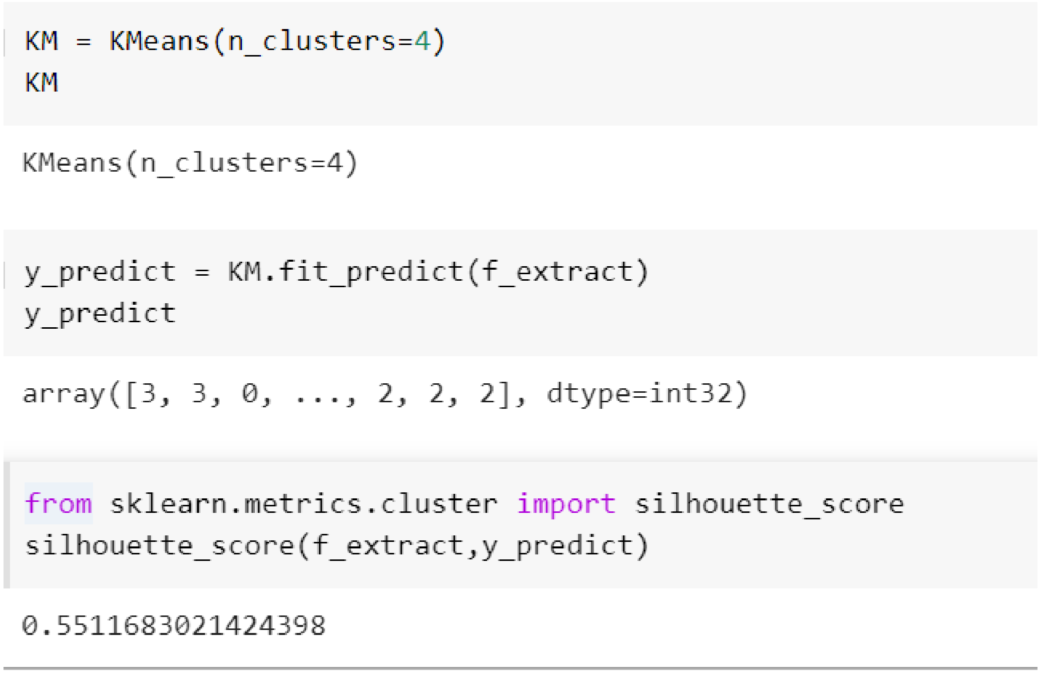
```
[<matplotlib.lines.Line2D at 0x7f8561d35f10>]
```

Figure 4.14: Display the sum of square error and k

4.7.9 Evaluation Of Elbow Point

The above figure shows the elbow shape is in 2 and 4 so to find the accurate cluster number I used Silhouette Score. It is very useful to find the accurate k value while using K mean algorithm which is used when the elbow it does not show the exact Elbow point. The value of silhouette score defines how similar are those objects in their own formed cluster where value of it range from -1to1. Very high value indicates each point is well matched to their own cluster and are poorly mismatch with other clusters. The value -1 of silhouette score means the value is assigned to wrong cluster whereas 0 indicates the sample is very close to two neighboring clusters. 0.6 is considered to be good clustering solution according to silhouette score.

4.7.10 Optimal Cluster



```
KM = KMeans(n_clusters=4)
KM
```

```
KMeans(n_clusters=4)
```

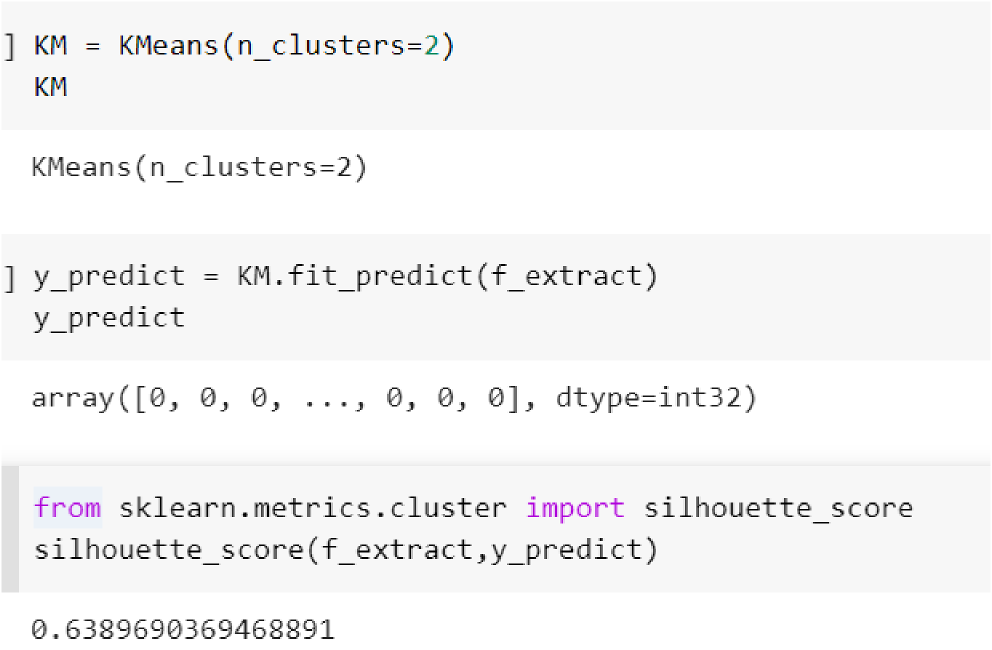
```
y_predict = KM.fit_predict(f_extract)
y_predict
```

```
array([3, 3, 0, ..., 2, 2, 2], dtype=int32)
```

```
from sklearn.metrics.cluster import silhouette_score
silhouette_score(f_extract,y_predict)
```

```
0.5511683021424398
```

Figure 4.15: Calculation of silhouette score with cluster 4 displayed



```
] KM = KMeans(n_clusters=2)
  KM
```

```
KMeans(n_clusters=2)
```

```
] y_predict = KM.fit_predict(f_extract)
  y_predict
```

```
array([0, 0, 0, ..., 0, 0, 0], dtype=int32)
```

```
from sklearn.metrics.cluster import silhouette_score
silhouette_score(f_extract,y_predict)
```

```
0.6389690369468891
```

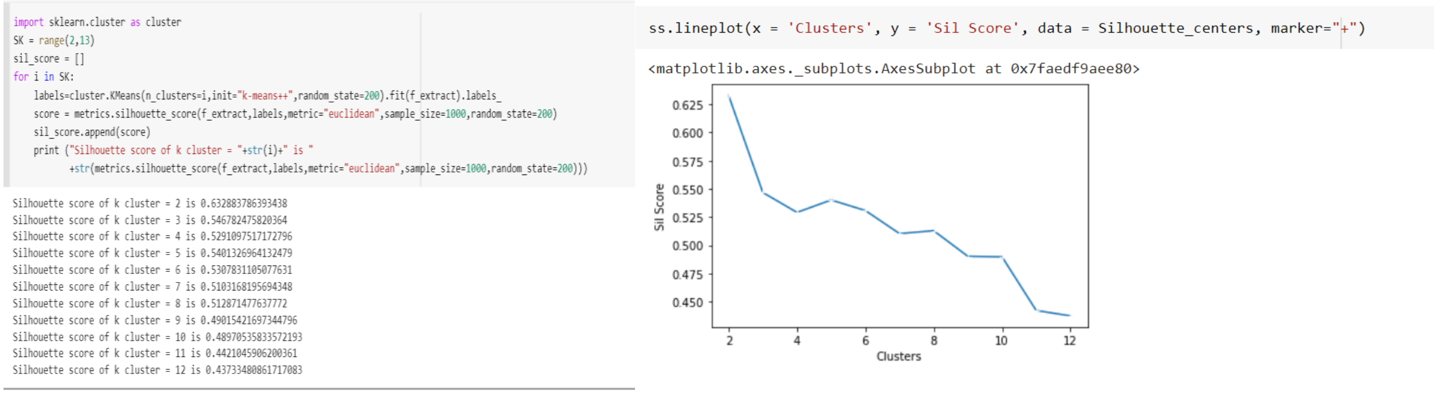
Figure 4.16: Calculation of silhouette score with cluster 2 displayed

4.7.11 Evaluation of cluster

The figure above shows that when the cluster number is 4 silhouette score is 0.55. But according to it the 0.6 it is considered as a good clustering solution.

In figure above the elbow point is in 2 or 4 so I allocated cluster value to 2.

As when the cluster value is 2 the silhouette score is 0.638 which is considered good. So, its concluded that the optimal k value is 2.



```
import sklearn.cluster as cluster
SK = range(2,13)
sil_score = []
for i in SK:
    labels=cluster.KMeans(n_clusters=i,init="k-means++",random_state=200).fit(f_extract).labels_
    score = metrics.silhouette_score(f_extract,labels,metric="euclidean",sample_size=1000,random_state=200)
    sil_score.append(score)
    print ("Silhouette score of k cluster = "+str(i)+" is "
          +str(metrics.silhouette_score(f_extract,labels,metric="euclidean",sample_size=1000,random_state=200)))
```

```
Silhouette score of k cluster = 2 is 0.632883786393438
Silhouette score of k cluster = 3 is 0.546782475820364
Silhouette score of k cluster = 4 is 0.5291097517172796
Silhouette score of k cluster = 5 is 0.5401326964132479
Silhouette score of k cluster = 6 is 0.5307831105077631
Silhouette score of k cluster = 7 is 0.5103168195694348
Silhouette score of k cluster = 8 is 0.512871477637772
Silhouette score of k cluster = 9 is 0.49015421697344796
Silhouette score of k cluster = 10 is 0.48970535833572193
Silhouette score of k cluster = 11 is 0.4421045906200361
Silhouette score of k cluster = 12 is 0.43733480861717083
```

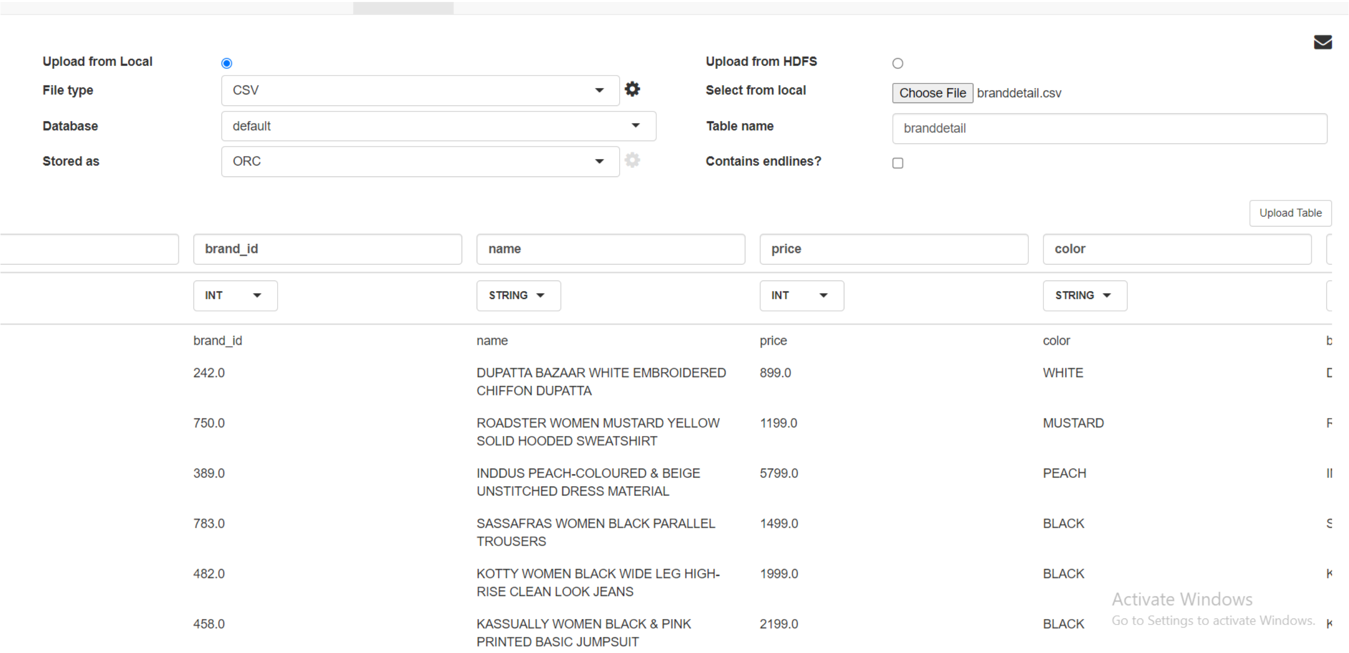
```
ss.lineplot(x = 'Clusters', y = 'Sil Score', data = Silhouette_centers, marker="+")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7faedf9aee80>
```

Figure 4.17: Calculation and Visualization of silhouette score.

5 Hive

A new dataset is created after the data cleaning and preprocessing. Then Hive is used for data queries and analysis. Hive is an open-source framework, built on the top of Apache Hadoop which is used efficiently to store and process the data. Hadoop is the popular software framework which is designed to store and process big data sets. As, result Hive is closely integrated with Hadoop which helps to work effectively and quicky on petabytes of data. In hive, data files are stored directly in Apache Hadoop distributed system or in other storage systems like Apache HBase. It enables Structured Query Language (SQL) developer to write or implement Hive Query Language for reading, writing and managing large data files. The uniqueness of hive that it allows the users to query large amount of data by leveraging MapReduce with the help of SQL like interface. Software used here is Hortonworks Sandbox HDP 2.6.5, which is open-source framework used for the processing and provide distributed storage data sets from multiple sources.



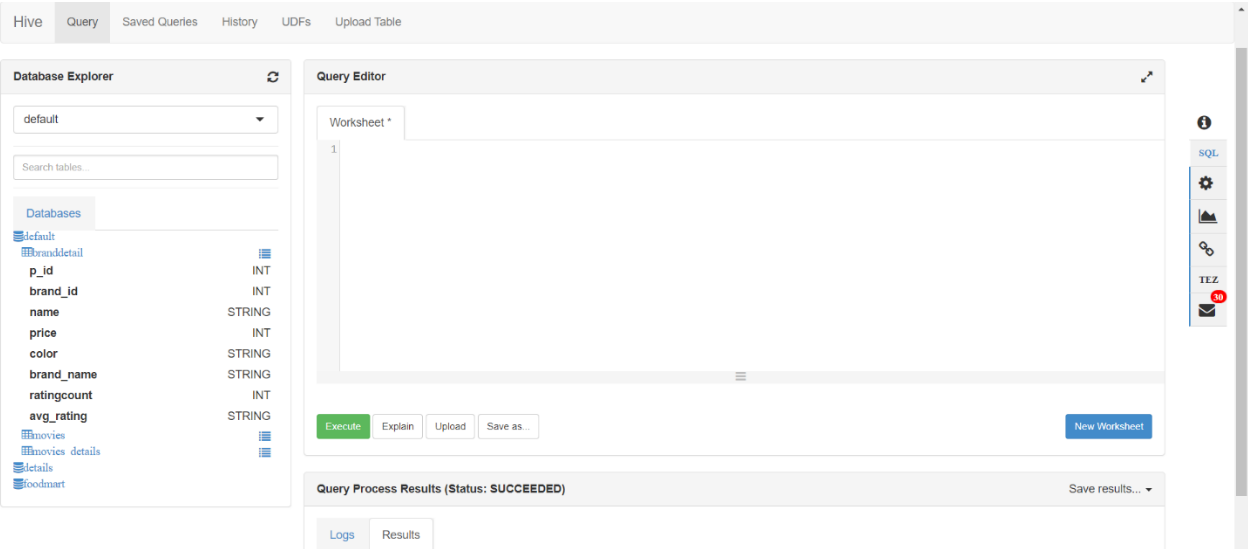
Upload from Local ●
File type: CSV
Database: default
Stored as: ORC

Upload from HDFS ○
Select from local: Choose File branddetail.csv
Table name: branddetail
Contains endlines? □

brand_id	name	price	color	b
INT	STRING	INT	STRING	
brand_id	name	price	color	b
242.0	DUPATTA BAZAAR WHITE EMBROIDERED CHIFFON DUPATTA	899.0	WHITE	D
750.0	ROADSTER WOMEN MUSTARD YELLOW SOLID HOODED SWEATSHIRT	1199.0	MUSTARD	F
389.0	INDDUS PEACH-COLOURED & BEIGE UNSTITCHED DRESS MATERIAL	5799.0	PEACH	I
783.0	SASSAFRAS WOMEN BLACK PARALLEL TROUSERS	1499.0	BLACK	S
482.0	KOTTY WOMEN BLACK WIDE LEG HIGH-RISE CLEAN LOOK JEANS	1999.0	BLACK	K
458.0	KASSUALLY WOMEN BLACK & PINK PRINTED BASIC JUMPSUIT	2199.0	BLACK	K

Figure 5.1 Loading data into Hive

This figure shows that data is storage in default database with table name branddetail. The default columns name is changed with the dataset column name. After that we can do further analysis of data by writing SQL, structured queries in query editor as shown in figure below.



default

default
branddetail

p_id	INT
brand_id	INT
name	STRING
price	INT
color	STRING
brand_name	STRING
ratingcount	INT
avg_rating	STRING

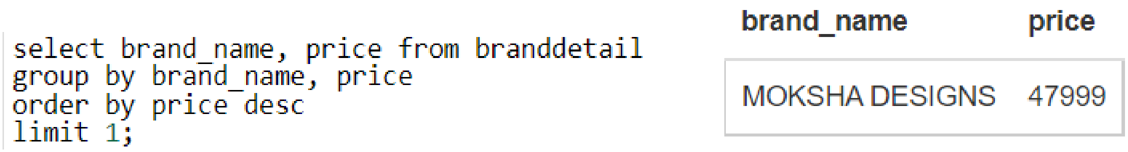
movies
movies_details
details
foodmart

Query Process Results (Status: SUCCEEDED)

Figure 5.2: Display of query editor

5.1 Hive Queries for the datasets

5.1.1 Which brand name is expensive and cheap?

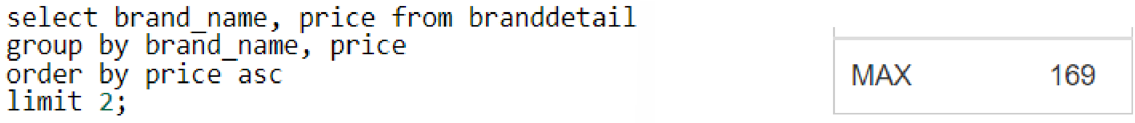


```
select brand_name, price from branddetail
group by brand_name, price
order by price desc
limit 1;
```

brand_name	price
MOKSHA DESIGNS	47999

Figure 5.3 Display of query and result for most expensive brand name

The above figure shows that the most expensive brand is Moksha Designs with price 47999.



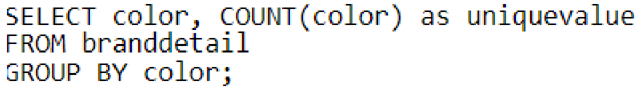
```
select brand_name, price from branddetail
group by brand_name, price
order by price asc
limit 2;
```

MAX	169

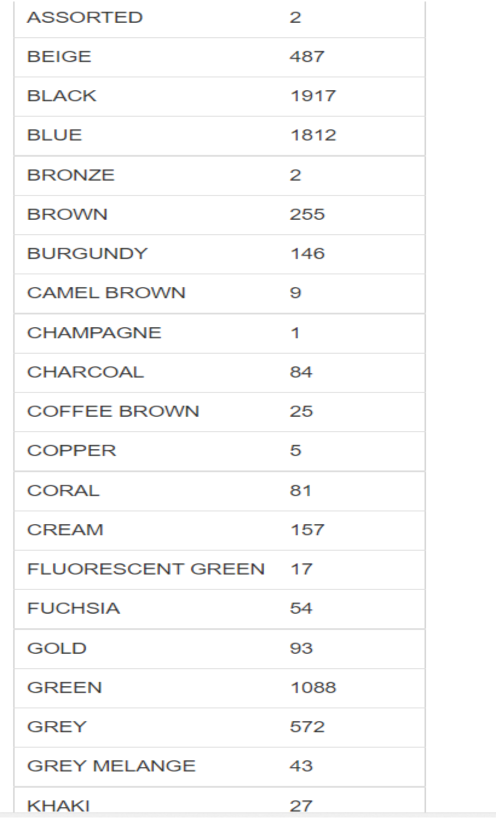
This shows that the cheapest brand is Max with price 169.

Figure 5.4 Display of query and result for most cheap brand name

5.1.2 How many times the same color is used?



```
SELECT color, COUNT(color) as uniquevalue
FROM branddetail
GROUP BY color;
```

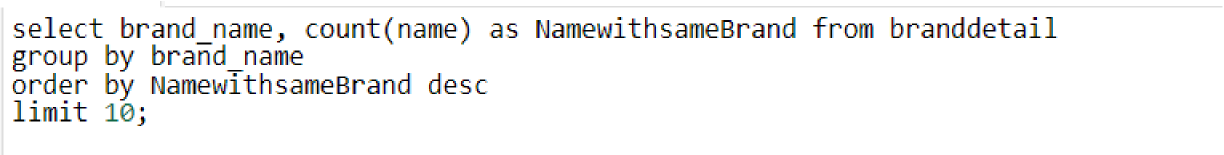


ASSORTED	2
BEIGE	487
BLACK	1917
BLUE	1812
BRONZE	2
BROWN	255
BURGUNDY	146
CAMEL BROWN	9
CHAMPAGNE	1
CHARCOAL	84
COFFEE BROWN	25
COPPER	5
CORAL	81
CREAM	157
FLUORESCENT GREEN	17
FUCHSIA	54
GOLD	93
GREEN	1088
GREY	572
GREY MELANGE	43
KHAKI	27

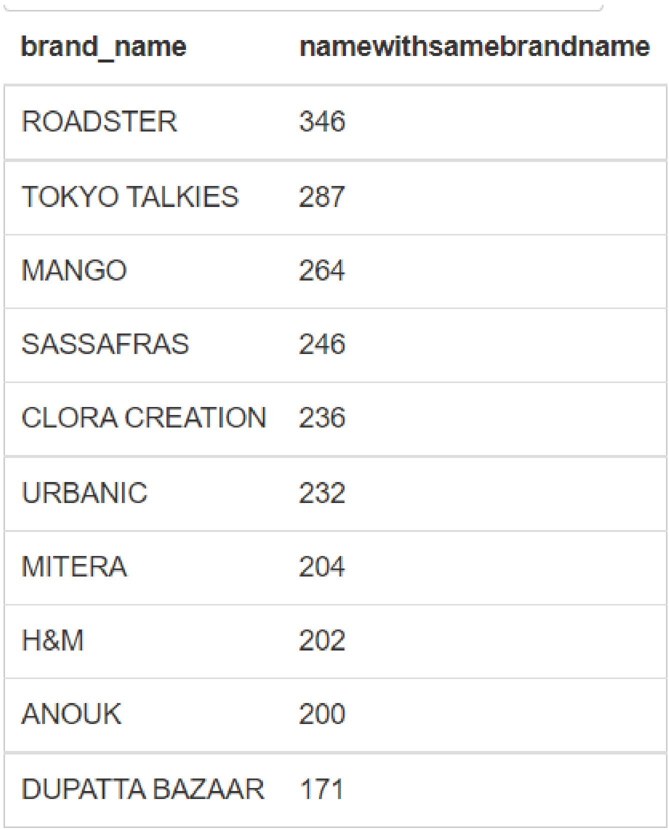
Figure 5.5: query and result for most used color

This shows that Black, Blue, Green is the most repeated color whereas Champagne, Assorted, Bronze is the least repeated color in compare to other color.

5.1.3 How many products are there with same brand name?



```
select brand_name, count(name) as NamewithsameBrand from branddetail
group by brand_name
order by NamewithsameBrand desc
limit 10;
```

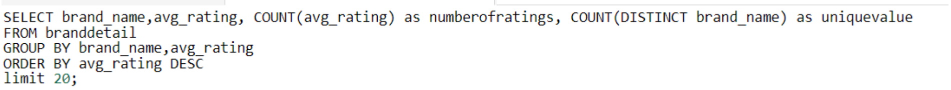


brand_name	namewithsamebrandname
ROADSTER	346
TOKYO TALKIES	287
MANGO	264
SASSAFRAS	246
CLORA CREATION	236
URBANIC	232
MITERA	204
H&M	202
ANOUK	200
DUPATTA BAZAAR	171

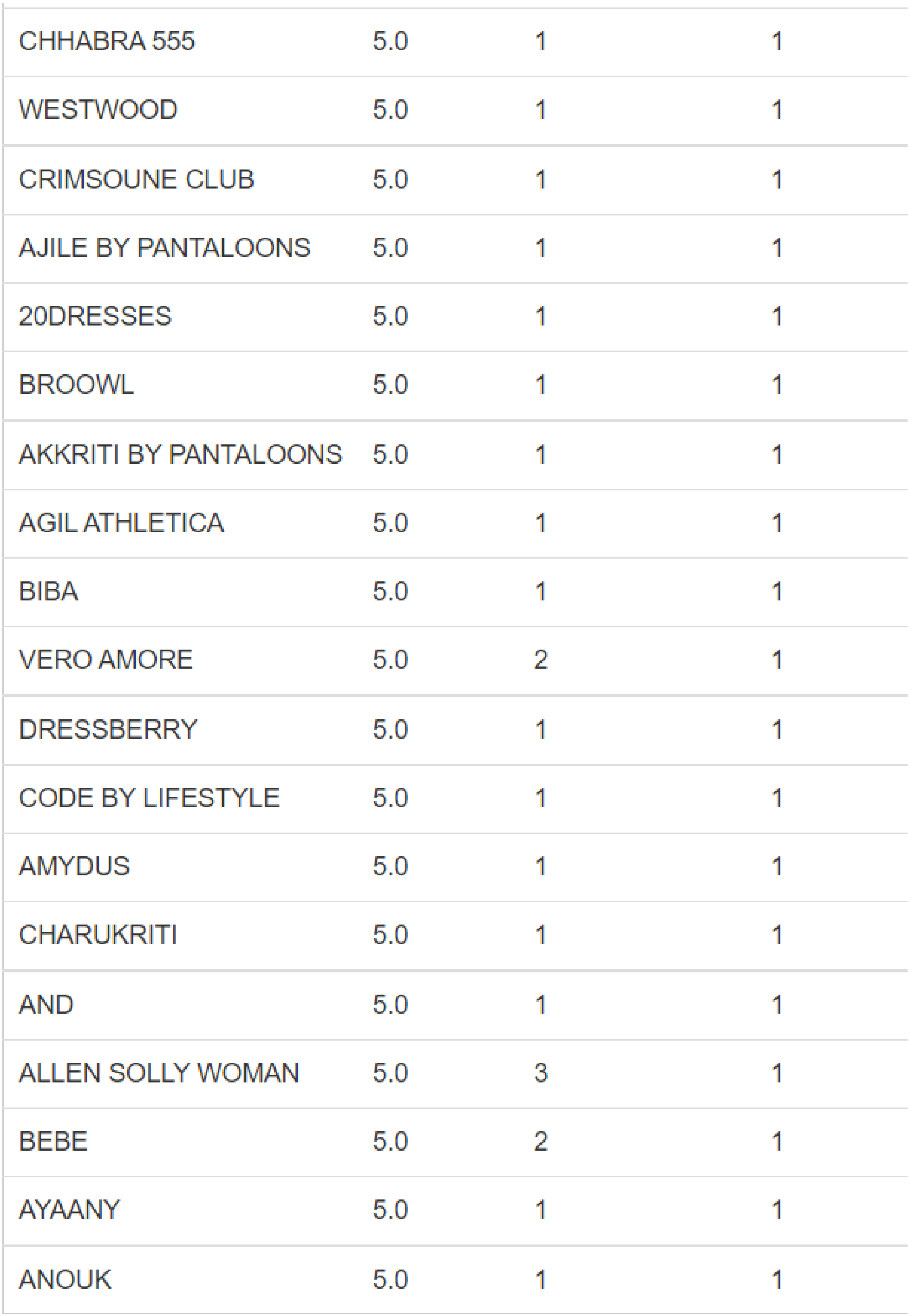
Figure 5.6: Query and result for the name with same brand name

By looking at the figure above we can analyze that there are 346 names with same brand name roadster and 171 name with dupatta bazzar brand name.

5.1.4 How many times a brand is rated with 5?



```
SELECT brand_name,avg_rating, COUNT(avg_rating) as numberofratings, COUNT(DISTINCT brand_name) as uniquevalue
FROM branddetail
GROUP BY brand_name,avg_rating
ORDER BY avg_rating DESC
limit 20;
```

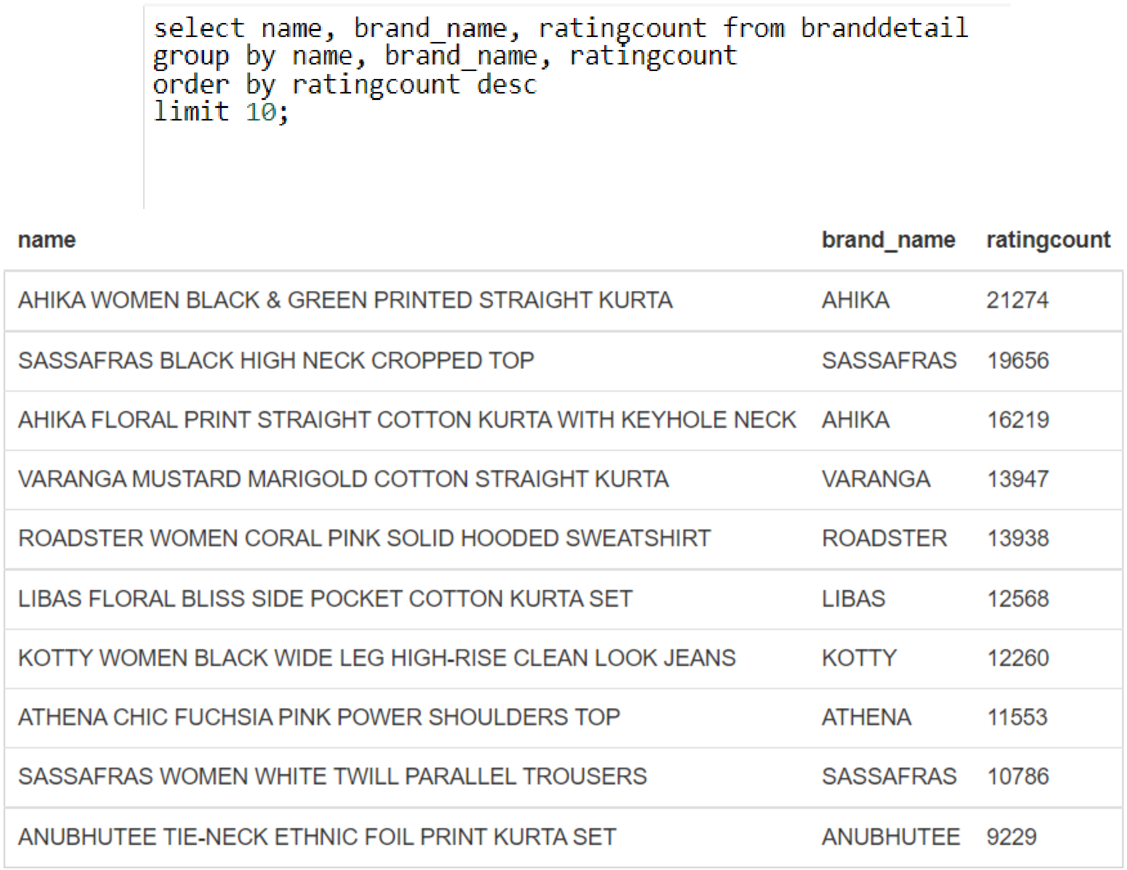


CHHABRA 555	5.0	1	1
WESTWOOD	5.0	1	1
CRIMSOUNE CLUB	5.0	1	1
AJILE BY PANTALOONS	5.0	1	1
20DRESSES	5.0	1	1
BROOWL	5.0	1	1
AKKRITI BY PANTALOONS	5.0	1	1
AGIL ATHLETICA	5.0	1	1
BIBA	5.0	1	1
VERO AMORE	5.0	2	1
DRESSBERRY	5.0	1	1
CODE BY LIFESTYLE	5.0	1	1
AMYDUS	5.0	1	1
CHARUKRITI	5.0	1	1
AND	5.0	1	1
ALLEN SOLLY WOMAN	5.0	3	1
BEBE	5.0	2	1
AYAANY	5.0	1	1
ANOUK	5.0	1	1

Figure 5.7: Query and Result with most rated brand

The figure above shows that ALLEN SOLLY WOMAN is rated 3 time with 5. Likewise ANOUK, AYAANY is rated only once.

5.1.5 Which brand, name have the highest rating Count?



```
select name, brand_name, ratingcount from branddetail
group by name, brand_name, ratingcount
order by ratingcount desc
limit 10;
```

name	brand_name	ratingcount
AHIKA WOMEN BLACK & GREEN PRINTED STRAIGHT KURTA	AHIKA	21274
SASSAFRAS BLACK HIGH NECK CROPPED TOP	SASSAFRAS	19656
AHIKA FLORAL PRINT STRAIGHT COTTON KURTA WITH KEYHOLE NECK	AHIKA	16219
VARANGA MUSTARD MARIGOLD COTTON STRAIGHT KURTA	VARANGA	13947
ROADSTER WOMEN CORAL PINK SOLID HOODED SWEATSHIRT	ROADSTER	13938
LIBAS FLORAL BLISS SIDE POCKET COTTON KURTA SET	LIBAS	12568
KOTTY WOMEN BLACK WIDE LEG HIGH-RISE CLEAN LOOK JEANS	KOTTY	12260
ATHENA CHIC FUCHSIA PINK POWER SHOULDERS TOP	ATHENA	11553
SASSAFRAS WOMEN WHITE TWILL PARALLEL TROUSERS	SASSAFRAS	10786
ANUBHUTEE TIE-NECK ETHNIC FOIL PRINT KURTA SET	ANUBHUTEE	9229

Figure 5.8: Query and Result for name, brand name with highest rating count

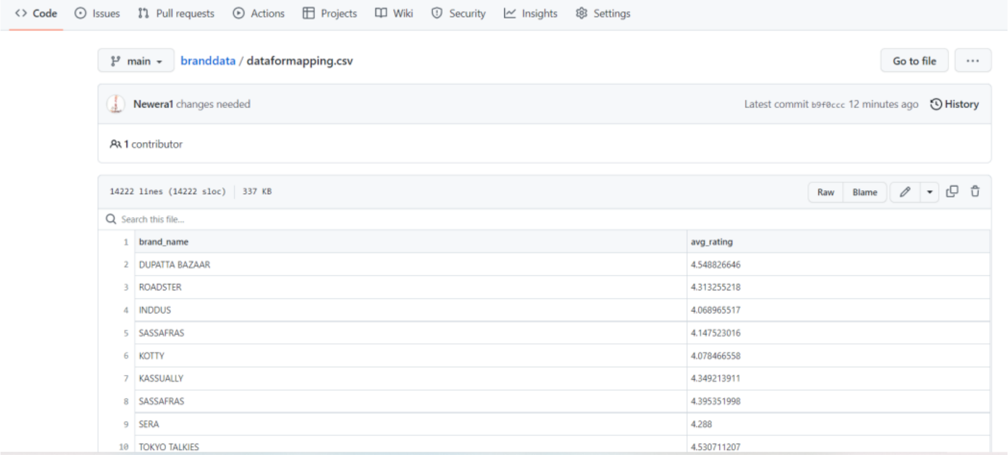
By looking at the above figure you can analyze the brand name with highest rating count which is AHIKA. In compare to that brand ANUBHUTEE brand name have low count which is 9229.

6 Map Reduce

Map Reduce is a built-in model in Apache Hadoop which is a software framework which process on multi-terabyte datasets. It is the processing unit of Hadoop using which data store in Hadoop can be processed. Execution steps of Map reduce contains of three steps which are map stage, shuffle stage and reduce stage. It is a specialization of split-apply-combine strategy used for analysis of data. Basically, map job is to break the tuples into key and value pairs.

Shuffle job is to redistribute the data according to the output key produce by map. And reducer job is to take the output of map as input and combine those tuples of data into further smaller tuples set. The job of reducer is always performed after mapper job as the sequence of the name MapReduce. The output and input are always store in file system.

Map reduce is written in Java but its capable of running different language like Python, Ruby, C++. In this project python programming language is used for map reduce along with MRjob Package. Then data set used is dataformapping.csv where there is two attributes brand_name and avg_rating. To use the dataset first the data is store in git hub repository.



brand_name	avg_rating
DUPATTA BAZAAR	4.548826646
ROADSTER	4.313255218
INDDUS	4.068965517
SASSAFRAS	4.147523016
KOTTY	4.078466558
KASSUALLY	4.349213911
SASSAFRAS	4.395351998
SERA	4.288
TOKYO TALKIES	4.530711207

Figure 6.1: Data uploaded in GitHub

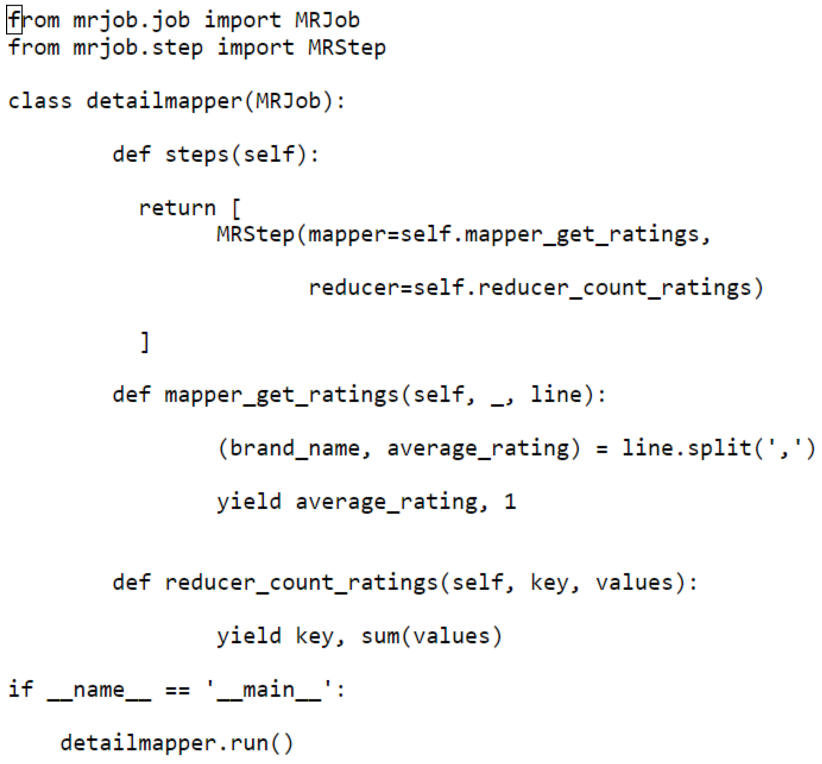
Then the data is fetch using:

```
wget https://raw.githubusercontent.com/Newera1/branddata/main/dataformapping.csv
```

wget is a command which is used as a convenient solution to download multiple files recursively from HTTP or FTP.

After that we can write code in using nano detailmapper.py (python file).

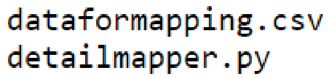
Nano is modeless editor where we can write and edit our code or text after immediately opening where CTRL+O is to save file in it



```
from mrjob.job import MRJob
from mrjob.step import MRStep

class detailmapper(MRJob):

        def steps(self):

          return [
                MRStep(mapper=self.mapper_get_ratings,

                       reducer=self.reducer_count_ratings)

          ]

        def mapper_get_ratings(self, _, line):

                (brand_name, average_rating) = line.split(',')

                yield average_rating, 1


        def reducer_count_ratings(self, key, values):

                yield key, sum(values)

if __name__ == '__main__':

    detailmapper.run()
```

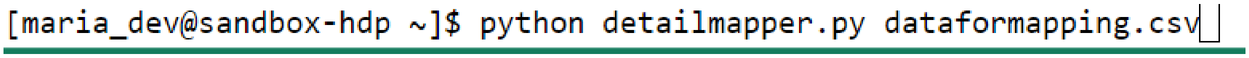
Figure 6.2: Mapreduce code

Verify whether the data is present in the local directory or not using:
ls



```
dataformapping.csv
detailmapper.py
```

Figure 6.3: check whether csv file or python file is there or not
As we can see data is present there
or
Hadoop fs -ls
MRJob is a python library which helps to write Map reduce code in python also due to which it is possible to write mapper and reducer function in single code. This helps developer to test MapReduce Python code locally which is written with mrjob. Here we are having a detailmapper.py a python file which includes mapper function, reducer function and definition. Step function is used to define both mapper and reducer function. Yield keyword in python is similar to return statement which is used to return values in Python.
We need to run python script with mrjob so that we can run detailmapper.py locally or with Hadoop.

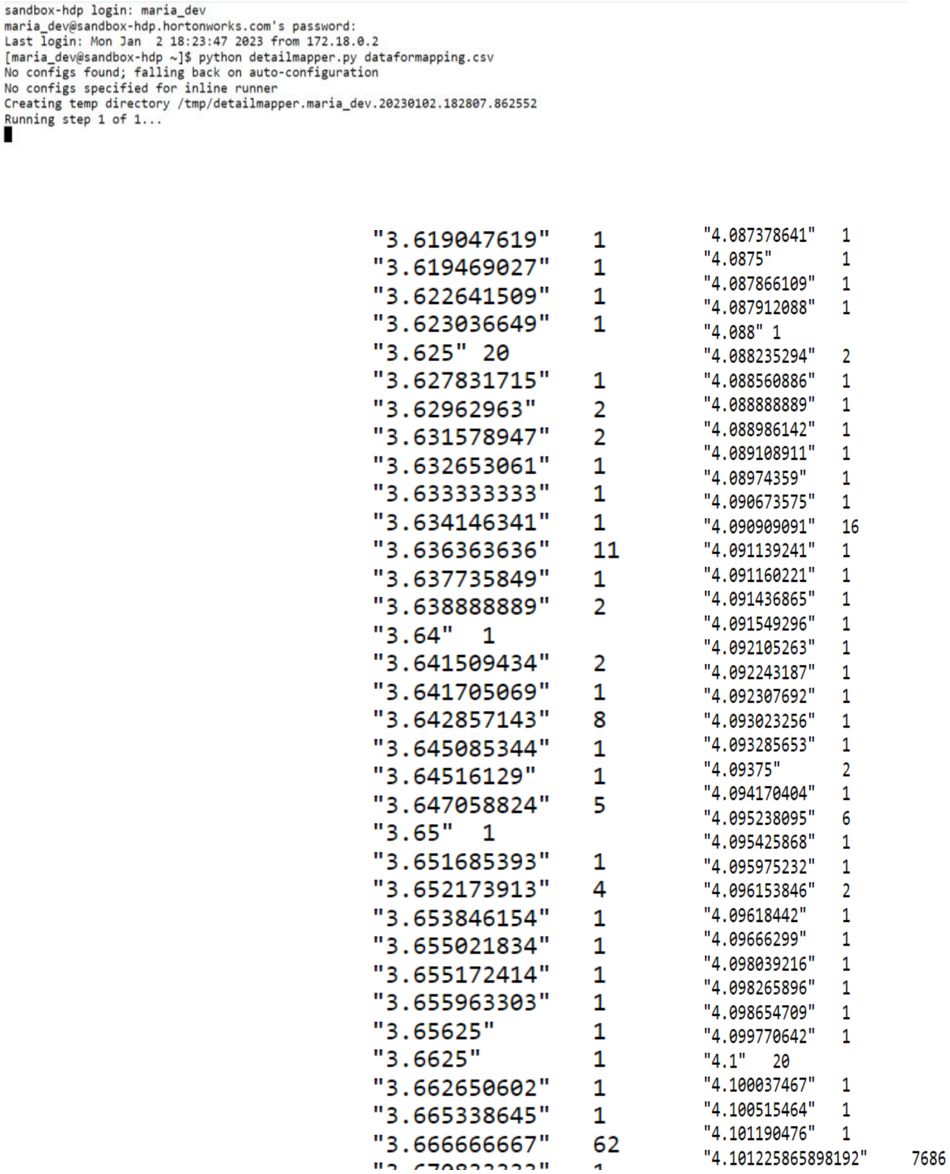


```
[maria_dev@sandbox-hdp ~]$ python detailmapper.py dataformapping.csv
```

Figure 4.4: Code to run python file and csv file together

To run locally this code needs to be executed where detailmapper.py is python file and dataformapping.csv is dataset.

Result



```
sandbox-hdp login: maria_dev
maria_dev@sandbox-hdp.hortonworks.com's password:
Last login: Mon Jan  2 18:23:47 2023 from 172.18.0.2
[maria_dev@sandbox-hdp ~]$ python detailmapper.py dataformapping.csv
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/detailmapper.maria_dev.20230102.182807.862552
Running step 1 of 1...
█
```

```
"3.619047619"   1
"3.619469027"   1
"3.622641509"   1
"3.623036649"   1
"3.625" 20
"3.627831715"   1
"3.62962963"    2
"3.631578947"   2
"3.632653061"   1
"3.633333333"   1
"3.634146341"   1
"3.636363636"   11
"3.637735849"   1
"3.638888889"   2
"3.64"  1
"3.641509434"   2
"3.641705069"   1
"3.642857143"   8
"3.645085344"   1
"3.64516129"    1
"3.647058824"   5
"3.65"  1
"3.651685393"   1
"3.652173913"   4
"3.653846154"   1
"3.655021834"   1
"3.655172414"   1
"3.655963303"   1
"3.65625"       1
"3.6625"        1
"3.662650602"   1
"3.665338645"   1
"3.666666667"   62
"3.670833333"   1
```

```
"4.087378641"   1
"4.0875"        1
"4.087866109"   1
"4.087912088"   1
"4.088" 1
"4.088235294"   2
"4.088560886"   1
"4.088888889"   1
"4.088986142"   1
"4.089108911"   1
"4.08974359"    1
"4.090673575"   1
"4.090909091"   16
"4.091139241"   1
"4.091160221"   1
"4.091436865"   1
"4.091549296"   1
"4.092105263"   1
"4.092243187"   1
"4.092307692"   1
"4.093023256"   1
"4.093285653"   1
"4.09375"       2
"4.094170404"   1
"4.095238095"   6
"4.095425868"   1
"4.095975232"   1
"4.096153846"   2
"4.09618442"    1
"4.09666299"    1
"4.098039216"   1
"4.098265896"   1
"4.098654709"   1
"4.099770642"   1
"4.1"   20
"4.100037467"   1
"4.100515464"   1
"4.101190476"   1
"4.10122586589192"      7686
```

Figure 6.5: Implementation of Mapreduce code and result displayed

As shown in figure 4.5, MapReduce breaks the input data into fragments and distributed those fargments into different machines. The input fragment consist pair of key and value. After that reducer will copy the sorted out from each Mapper(intermediate key_value pair). The shuffle and sort work simultaneously and the data generated by reducer is the final step as shown in above figure. Now the result above shows how many times an average rating value is repeated in whole dataset. The data is sorted in ascending order. For example, 3.666666667 is repeated for 62 times. Some of the value is repeated only once. This way map reduce can be applied in big datasets.

7 Issues and Solution in use of Big Data Analytics in the fashion retail industry

Big data in Fashion

Big data refers to the structured or unstructured, large and growing volume of dataset which an organization collects to analyze those data so that a useful information is formed. But this is not possible by using traditional data processing software. As, it is the era of fast fashion it generates and creates data in various form like words, images rapidly and in with different trends. Since last few decades, big data have gained a significant importance in fashion industry as those data generated portrays all the features of big data. In fashion industry discovering and developing trend is a lifeline. Nowadays the demands of customer are changing frequently like they want personalized or customized garments, color, fit, pattern. Due to which fashion industry face loss in the business due to excessive stock. For this problem to be solve big data analysis should be done (Silva et al., 2019). Big data analytics is the key process in preventing fraudulent activities and in better decision making in any organization. Basically, it is used by any organization or business to have access to right data at right moment so that it can improve every aspect of business which is from production phase to marketing phase. Popular brands like Zara analyze big data to understand customer demands and then translate it into tangible design so that it meets the consumer expectations and demands. Top shop is using Big Data form different blogs about fashion and social media to determine the evolving trends by using predictive analytics. Following are some issue and solution in use of Big Data Analytics in fashion retail industry: -

Mass Customization

The main thing a person want from fashion industry is customization (Silva et al., 2019). There are different technologies which are used in fashion industry for creating different new ways to satisfy customer changing needs and demands. As industry should shift from mass production to mass customization which is the main issue of fashion industry. If the level of customization increases complexity increases. Due to the lack of professional design knowledge customer are unaware of what they need like a customer can like different design, color, print at the same time. Because of which the product is not formed according to their requirement, hence customer will be dissatisfied. What will the retailer do if the product is returned it becomes less likely to be sold out because it is a unique product.

Solution

According to research 50 percent of customer showed the interest in buying customized product and willing to pay 20 % more for the customized product. For this customer need to understand and have knowledge about product specification. Like each customer can have different demand for products so there should be good relationship between manufacturer and consumers like customer can directly talk with the manufacturer about their personal requirement. Companies should involve customer for the product configuration so that marketing, manufacturing, distribution, sales all will have the knowledge about the customer requirement. Different policies should be formulated about whether the customized products can be returned or not.

Social and Environmental impact

Over last 20 years Fast fashion have gained massive popularity due to the affordable clothing and constantly changing trends. Many people love to shop daily because they feel excited, addicted, enjoy but the clothes are no necessary. Just like some like to eat new food, meeting fiend or visiting new places. Fast fashion trend in fashion industry promotes the excessive consumerism, throwaway culture as many customers buying decision is based on their emotions (Bhardwaj & Fairhurst, 2010). Consumer only wear these types of clothes in average of 7 times as they don't keep these clothes for long period of time. Due to which it directly effects the environment to make a garment we need to extract from raw material, manufacture it as well as distribute. So, frequently producing mass number of garments and not using it properly it's to pollution. About 10 % of global carbon is emitted from fashion industry The demand for fast fashion id still growing which exploits local communities in sweatshop for the production of cheap garments. Many young girls face these terrible conditions and abuse also they are forced to abort the pregnancy for the continuation in work. About 80 % of garment worker are women between the age of 18 to 35. Fast Fashion takes places in overseas country with poor laws and human protection where workers work long hours and are provided with very low income. In Bangladesh 2013, Rana Plaza garment factory collapse due to the catastrophic events due to its unsafe work conditions. Poorly paid and dangerous working environment is very common in fast fashion. Women are the main victim because of their gender.

Solution

Using 3D printing for the testing purpose to check the quality, design and fit before its production.

Need to use less harmful material for environment like sustainable materials

Encourage customer to rent or share their clothing after one or two use instead of throwing it.

Building cyber resilience and Rising Distrust

Due to the evolution of fashion industry big data in fashion have reached to a high due to which there is increase in cyber-attacks. This is fourth most target industry, on 2019 and 2020 there was a data comprising events in this industry. The loss in trust will cost millions amount in business. Also, different regulatory Act can huge fines for the targeted business. About 60% of fashion have been classified as misleading and unsubstantiated by Changing Markets Foundation. Making data available to retailers can lead to the collection of personal information like name, addresses, email addresses. In addition, different technical information like Facebook, twitter information can be collected also retailers can collect these data from other sources too. Majority of customer want more transparency form the fashion brand and expect them to take all the responsibility for pollution. There is data breaches at number of online fashion companies which left out customer to think before they share their information with brand and retailers. In fashion industry if you don’t have trust, you would not be able to win customer over time.

Solution

Brand should allocate the greater portion of their budgets for cybersecurity, As the threat is constantly evolving brand should monitor the cyber risk and have the idea about have data is handle from use to the disposal of data. A cyber resilience strategy should be mentioned by every retailer. They should understand that data are very valuable assets. As it includes everything about customer like location, age, gender to data about process used for production, about sales as well.

To rebuild the trust brand should include Creative Integrity, Data Protection, Authenticity, transparency. Information like material used in manufacturing process and supplier working condition can be shared through Blockchain-enable product passport. Standardization can help to find out the social and environmental impact throughout the product life cycle.

8 Conclusion

Big data analytics plays an important role for determining business strategies which simultaneously help in better decision making. In this paper, a methodology for big data analysis with machine learning, hive, map reduce framework, powerbi, SQL is used. In my project there is 14,221 datasets to be analyzed so hive is used to process those data. There are two datasets so to merge SQL is used. In order to analyze the attributes like price brand name, rating count, average rating HiveQL is used as it is faster than SQL. In my dataset there is both categorical data and numeric data so while using K mean clustering, I have only used numeric data which is price and average rating. These categorical data can be used in future for further processing. Powerbi is a visualization tools which help to analyze the data forming different shapes like funnel, bar, pie charts, cards, slider, scatterplot which I have used in my project. As in this paper voluminous data is present so Map reduce framework is used to process those data in short period of time by dividing the data into chunks, store and process it So average rating attribute is used to count how many times same rating is giving for the brand in this paper.

9 References

Ashwitha, T.A., Rodrigues, A.P. and Chiplunkar, N.N. (2017) “Movie dataset analysis using Hadoop-Hive,” 2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS) [Preprint]. Available at: https://doi.org/10.1109/csitss.2017.8447828.

Bante, P.M. and Rajeswari, K. (2017) “Big Data Analytics using Hadoop map reduce framework and data migration process,” 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA) [Preprint]. Available at: https://doi.org/10.1109/iccubea.2017.8463824.

Bhardwaj, V. and Fairhurst, A. (2010) “Fast fashion: Response to changes in the fashion industry,” The International Review of Retail, Distribution and Consumer Research, 20(1), pp. 165–173. Available at: https://doi.org/10.1080/09593960903498300.

Big Data and fashion: How's it changing the industry? (2022) Crayon Data. Available at: https://www.crayondata.com/big-data-fashion-changing-industry/ (Accessed: January 7, 2023).

Big Data in fashion industry - researchgate (no date). Available at: https://www.researchgate.net/publication/320951232_Big_data_in_fashion_industry (Accessed: January 7, 2023).

Manwal, M. and Gupta, A. (2017) “Big Data and Hadoop — a technological survey,” 2017 International Conference on Emerging Trends in Computing and Communication Technologies (ICETCCT) [Preprint]. Available at: https://doi.org/10.1109/icetcct.2017.8280345.

Merla, P.R. and Liang, Y. (2017) “Data analysis using Hadoop MapReduce Environment,” 2017 IEEE International Conference on Big Data (Big Data) [Preprint]. Available at: https://doi.org/10.1109/bigdata.2017.8258541.

Pham, D.T., Dimov, S.S. and Nguyen, C.D. (2005) “Selection of k in k-means clustering,” Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 219(1), pp. 103–119. Available at: https://doi.org/10.1243/095440605x8298.

Ramya, A.V. and Sivasankar, E. (2014) “Distributed pattern matching and document analysis in big data using Hadoop mapreduce model,” 2014 International Conference on Parallel, Distributed and Grid Computing [Preprint]. Available at: https://doi.org/10.1109/pdgc.2014.7030762.

Saravana, M.K. and Harish, K. (2017) “A case study on analyzing uber datasets using Hadoop framework,” 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS) [Preprint]. Available at: https://doi.org/10.1109/icecds.2017.8389665.

Schultz, J., Vierya, J. and Lu, E. (2012) “Abstract: Analyzing patterns in large-scale graphs using mapreduce in Hadoop,” 2012 SC Companion: High Performance Computing, Networking Storage and Analysis [Preprint]. Available at: https://doi.org/10.1109/sc.companion.2012.257.

Shaikh, F. et al. (2018) “YouTube data analysis using mapreduce on Hadoop,” 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information &amp; Communication Technology (RTEICT) [Preprint]. Available at: https://doi.org/10.1109/rteict42901.2018.9012635.

Silva, E.S., Hassani, H. and Madsen, D.Ø. (2019) Big Data In Fashion: Transforming the retail sector, Journal of Business Strategy. Emerald Publishing Limited. Available at: https://www.emerald.com/insight/content/doi/10.1108/JBS-04-2019-0062/full/html (Accessed: January 7, 2023).

Singh, K. and Kaur, R. (2014) “Hadoop: Addressing challenges of Big Data,” 2014 IEEE International Advance Computing Conference (IACC) [Preprint]. Available at: https://doi.org/10.1109/iadcc.2014.6779407.

Varela, B., Bernardino, J. and Pedrosa, I. (2020) “Twitter sensitivity analysis in a higher school using power Bi,” 2020 15th Iberian Conference on Information Systems and Technologies (CISTI) [Preprint]. Available at: https://doi.org/10.23919/cisti49556.2020.9140979.