A REPORT

ON

BIG DATA TECHNOLOGIES

BY: TIMOTHY AYODELE OLATUNJI

LIST OF TABLES

LIST OF FIGURES

CONTENTS PAGE

PART I

1.0 INTRODUCTION

This report analysis is divided into parts, the first part contains cleaning and analysis of given dataset while the second part discusses in detail the industry in which this dataset has been drawn from.

In this first part, the given two datasets will be combined into a new dataset, the new merged dataset will be cleaned with Google colab which is a python interface, analyzed using HIVE, MAPREDUCE and machine learning techniques, and be visualized using Python and Power-Bi.

These two datasets, both belonging to the fashion industry, the first dataset named “fashion dataset.csv” is being uploaded to the files section on the google colab interface. For the purpose of reading the dataset and performing other cleaning process, pandas which is a function associated with python is being imported as pd. And the first dataset is being read using the following lines of code. The dataset has 14,329(fourteen thousand, three hundred and twenty-nine) rows and 9(Nine) Columns.

```
import pandas as pd
dtt = pd.read_csv('fashion dataset.csv')
dtt
```

	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...
...	...	...	...	...	...	...	...	...	...
14324	17029604.0	The Chennai Silks Pink & Silver-Toned Floral Z...	3999.0	Pink	The Chennai Silks	NaN	NaN	Design Details Pink and silver-...	{'Better Cotton Initiative': 'Regular', 'Blous...
14325	17600212.0	Kinder Kids Girls Blue & Green Printed Foil Pr...	2050.0	Blue	Kinder Kids	NaN	NaN	Blue and green printed lehenga choli, has foi...	{'Blouse Closure': 'NA', 'Blouse Fabric': 'Cot...
14326	18159266.0	KLOTTHE Women Green & Black Floral Printed Pal...	1659.0	Green	KLOTTHE	NaN	NaN	Green and black woven palazzos ...	{'Body or Garment Size': 'To-Fit Denotes Body ...
14327	18921114.0	InWeave Women Red Printed A-Line Skirt	2399.0	Red	InWeave	NaN	NaN	<p>Red printed A-line skirt, has drawstring cl...	{'Add-Ons': 'NA', 'Body Shape ID': '324,333,42...
14328	19361058.0	BoStreet Women Navy Blue Tapered Fit Trousers	2599.0	Navy Blue	BoStreet	NaN	NaN	Navy blue knitted trousers <li...	{'Add-Ons': 'NA', 'Body Shape ID': '443,333,42...

14329 rows × 9 columns

fig.1.0.0

The dataset consists of columns; p_id which indicate Product ID, Name, Price, Colour, Brand, RatingCount, Avg_Rating, Description and p_attributes.

```
dtt.columns
```

```
Index(['p_id', 'name', 'price', 'colour', 'brand', 'ratingCount', 'avg_rating',
       'description', 'p_attributes'],
      dtype='object')
```

fig 1.0.1

And for the purpose of data cleaning, the dataset has total number of null values as shown below.

```
dtt.isna().sum()
```

```
p_id             18
name             19
price            19
colour           22
brand            24
ratingCount    7748
avg_rating     7748
description      19
p_attributes     19
dtype: int64
```

fig 1.0.2

The second dataset named “fashion brand details” is also being read using the following lines of code and it has 1,020(One thousand and twenty) rows and 2(two) columns as shown below:

```
dtt2 = pd.read_csv('fashion brand details.csv')
dtt2
```

	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
1019	1020	Zola

1020 rows × 2 columns

Fig 1.0.3

The second dataset consist of columns; brand_id and brand_name

```
dtt2.columns
```

```
Index(['brand_id', 'brand_name'], dtype='object')
```

Fig 1.0.4

It also has no null values as shown below;

```
dtt2.isna().sum()
```

```
brand_id      0
brand_name    0
dtype: int64
```

Fig 1.0.5

The first dataset has a duplicate of 59 records while second dataset has no duplicate record

```
dtt.duplicated().sum()
```

```
59
```

fig 1.0.6

```
dtt2.duplicated().sum()
```

```
0
```

fig 1.0.7

Some duplicate record of the first data is shown below:

```
dtt.loc[dtt.duplicated(), :]
```

	p_id	name	price	colour	brand	ratingCount	avg_rating	description	p_attributes
219	19215754.0	BoStreet Women Maroon Flared Skirts	2499.0	Maroon	BoStreet	NaN	NaN	Maroon Flared Skirts, has slip-on closure and ...	{'Add-Ons': 'NA', 'Body Shape ID': '443,324,33...
437	18548534.0	MONTREZ Women Blue Denim Solid Skirts	1499.0	Blue	MONTREZ	36.0	4.194444	<p>Blue solid knitted midi denim A-line skirt...	{'Add-Ons': 'NA', 'Body Shape ID': '443,324,33...
696	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
847	18815858.0	Emprall Women Pink Solid Pleated Mini Skirts	1299.0	Pink	Emprall	NaN	NaN	<p>Pink pleated flared mini skirt, has an elas...	{'Add-Ons': 'NA', 'Body Shape ID': '443,333,42...
3025	18841582.0	Styli Women Black Solid A-Line Skirt	2149.0	Black	Styli	NaN	NaN	<p>Black solid A-line skirt, has a waistband ...	{'Add-Ons': 'NA', 'Body Shape ID': '324,333,42...
3041	18485216.0	Hive91 Women Black Printed A-Line Maxi Skirts	1349.0	Black	Hive91	NaN	NaN	<p>Black printed A-line midi skirt, has an ela...	{'Add-Ons': 'NA', 'Body Shape ID': '324,333,42...
3053	18604646.0	Zink London Women Blue Floral Lace Design Skirts	1999.0	Blue	Zink London	NaN	NaN	<p>Blue floral lace design A-line skirt has a...	{'Add-Ons': 'NA', 'Body Shape ID': '443,324,33...
3168	19235438.0	Oxolloxo Women Navy Blue Printed Knee-Length E...	1599.0	Navy Blue	Oxolloxo	NaN	NaN	<p>Navy blue printed flared knee length skirt,...	{'Add-Ons': 'NA', 'Body Shape ID': '443,324,33...
3394	10604482.0	SASSAFRAS Blue Washed Midi Denim Pure Cotton A...	1999.0	Blue	SASSAFRAS	632.0	4.143987	Blue washed midi denim A-line skirt, has a par...	{'Add-Ons': 'NA', 'Body Shape ID': '443,324,33...
3426	13792358.0	U&F Red & White Printed Accordion Pleat Maxi F...	1699.0	Red	U&F	441.0	4.074830	Red and white printed accordion pleats woven m...	{'Add-Ons': 'NA', 'Body Shape ID': '324,333,42...
3461	13626148.0	SASSAFRAS Burgundy Dobby Weave Maxi Flared Skirt	1699.0	Burgundy	SASSAFRAS	219.0	4.337900	Burgundy dobby weave maxi flared skirt, has an...	{'Add-Ons': 'Comes with a belt', 'Body Shape I...
3936	18879630.0	MANGO Lilac & Green Sequinned Mini Skirt	3690.0	Lavender	MANGO	NaN	NaN	Lilac and green sequinned straight mini skirt,...	{'Add-Ons': 'NA', 'Body Shape ID': '443,424', ...
4032	18977844.0	Popwings Women Lavender-Colored & Black Stripe...	1299.0	Lavender	Popwings	6.0	4.833333	<p>Lavender-colored and black striped pencil k...	{'Add-Ons': 'NA', 'Body Shape ID': '424', 'Bod...

Fig1.0.8

Having seen the two datasets given, it can easily be concluded that it is a large dataset which means it is a big data. It is a big data because it contains the six (6) V's of big data, that is, it has Volume, Variety of columns, Velocity of which the dataset has been collected, Veracity of the dataset, Value and Variability. For this reason, big data techniques will be performed on the two datasets firstly by merging and cleaning before analyzing.

Three research questions will be dealt with using various big data techniques in this report which are:

a) What is the average price of each color of the fashion product, and which has the highest and lowest value?
b) What is the average price of each fashion brands and which fashion product brand having the highest and lowest average price?

These questions will be analyzed, and results will be displayed given these two datasets.

2.0 DATA MERGING AND CLEANING

The two datasets have two common columns which are ‘brand’ and ‘brand_name’, the two datasets are therefore merged at this point to obtain this new dataset.

```
newdata=pd.merge(dtt, dtt2, left_on='brand', right_on='brand_name')
newdata
```

	p_id	name	price	colour	brand	ratingCount	avg_rating	description	p_attributes	brand_id	brand_name
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'...	242	Dupatta Bazaar
1	10711448.0	Dupatta Bazaar Women White Solid Dupatta	599.0	White	Dupatta Bazaar	1531.0	4.536251	White solid dupatta and has a taping borderPol...	{'Border': 'Taping', 'Fabric': 'Poly Chiffon',...	242	Dupatta Bazaar
2	14964708.0	Dupatta Bazaar Orange & Green Dyed Art Silk Ba...	899.0	Orange	Dupatta Bazaar	30.0	4.366667	Orange and green bandhani dyed dupatta has ban...	{'Border': 'Woven Design', 'Fabric': 'Art Silk...	242	Dupatta Bazaar
3	13552234.0	Dupatta Bazaar Black Solid Dupatta	599.0	Black	Dupatta Bazaar	232.0	4.547414	Black solid Dupatta and has a solid borderMate...	{'Border': 'Solid', 'Fabric': 'Poly Chiffon', ...	242	Dupatta Bazaar
4	17663032.0	Dupatta Bazaar Women's Solid Orange Dupatta	599.0	Orange	Dupatta Bazaar	NaN	NaN	Orange solid dupatta Woven design borderPol...	{'Border': 'Woven Design', 'Fabric': 'Poly Chi...	242	Dupatta Bazaar
...	...	...	...	...	...	...	...	...	...	...	...
8093	18512752.0	Mystere Paris Multicoloured Checked Extended S...	999.0	Multi	Mystere Paris	2.0	3.500000	Multicoloured regular peplum top </l...	{'Body Shape ID': '333', 'Body or Garment Size...	605	Mystere Paris
8094	18819296.0	Swasti Women Blue Ethnic Motifs Printed Mirror...	2070.0	Blue	Swasti	NaN	NaN	Colour: blue Ethnic motif...	{'Body Shape ID': '424', 'Body or Garment Size...	861	Swasti
8095	18321444.0	Havida Sarees Brown & Copper-Toned Pure Chiffo...	2399.0	Brown	Havida Sarees	NaN	NaN	Design Details Brown and co...	{'Blouse': 'Blouse Piece', 'Blouse Fabric': 'V...	358	Havida Sarees
8096	18289626.0	Ziva Fashion Women White Floral Yoke Design Ku...	2999.0	White	Ziva Fashion	NaN	NaN	White yoke design Kurta with Trousers with dup...	{'Add-Ons': 'NA', 'Body Shape ID': '333,424', ...	1016	Ziva Fashion
8097	2117164.0	Noi Cream-Coloured & Brown Printed Shawl	1999.0	Cream	Noi	5.0	3.400000	Cream-coloured and brown printed shawl, has fr...	{'Border': 'Printed', 'Fabric': 'Wool', 'Fabri...	637	Noi

8098 rows × 11 columns

Fig 2.0.0

After merging the two datasets, the panda function merged and reduced the total rows to 8,098(eight thousand and ninety-eight) and added the columns together to make it 11(eleven) columns thereby eliminating null rows.

The new dataset has the following sum of null data

```
newdata.isna().sum()
```

```
p_id              0
name              0
price             0
colour            1
brand             0
ratingCount    4139
avg_rating     4139
description       0
p_attributes      0
brand_id          0
brand_name        0
dtype: int64
```

Fig 2.0.1

The new dataset also has 27 duplicated data.

```
newdata.duplicated().sum()
```

```
27
```

Fig 2.0.2

For the purpose of research questions raised, all duplicate data and null colour need to be removed. And ratingCount with avg_rating replaced with 0 as 0 rating is assumed to be same as null rating for the purpose of the research questions.

The last duplicate rows will be removed and keeping the first duplicate rows.

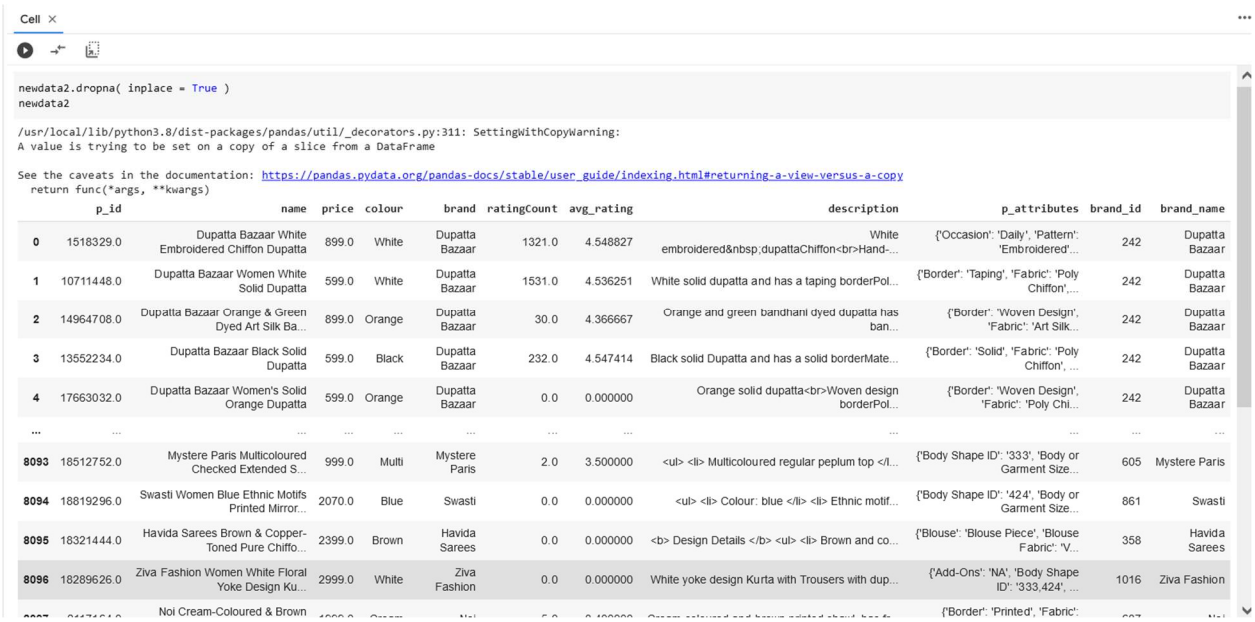
```
newdata2=newdata.drop_duplicates(keep='first')
newdata2
```

	p_id	name	price	colour	brand	ratingCount	avg_rating	description	p_attributes	brand_id	brand_name
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'...	242	Dupatta Bazaar
1	10711448.0	Dupatta Bazaar Women White Solid Dupatta	599.0	White	Dupatta Bazaar	1531.0	4.536251	White solid dupatta and has a taping borderPol...	{'Border': 'Taping', 'Fabric': 'Poly Chiffon',...	242	Dupatta Bazaar
2	14964708.0	Dupatta Bazaar Orange & Green Dyed Art Silk Ba...	899.0	Orange	Dupatta Bazaar	30.0	4.366667	Orange and green bandhani dyed dupatta has ban...	{'Border': 'Woven Design', 'Fabric': 'Art Silk...	242	Dupatta Bazaar
3	13552234.0	Dupatta Bazaar Black Solid Dupatta	599.0	Black	Dupatta Bazaar	232.0	4.547414	Black solid Dupatta and has a solid borderMate...	{'Border': 'Solid', 'Fabric': 'Poly Chiffon', ...	242	Dupatta Bazaar
4	17663032.0	Dupatta Bazaar Women's Solid Orange Dupatta	599.0	Orange	Dupatta Bazaar	NaN	NaN	Orange solid dupatta Woven design borderPol...	{'Border': 'Woven Design', 'Fabric': 'Poly Chi...	242	Dupatta Bazaar
...	...	...	...	...	...	...	...	...	...	...	...
8093	18512752.0	Mystere Paris Multicoloured Checked Extended S...	999.0	Multi	Mystere Paris	2.0	3.500000	Multicoloured regular peplum top </l...	{'Body Shape ID': '333', 'Body or Garment Size...	605	Mystere Paris
8094	18819296.0	Swasti Women Blue Ethnic Motifs Printed Mirror...	2070.0	Blue	Swasti	NaN	NaN	Colour: blue Ethnic motif...	{'Body Shape ID': '424', 'Body or Garment Size...	861	Swasti
8095	18321444.0	Havida Sarees Brown & Copper-Toned Pure Chiffo...	2399.0	Brown	Havida Sarees	NaN	NaN	Design Details Brown and co...	{'Blouse': 'Blouse Piece', 'Blouse Fabric': 'V...	358	Havida Sarees
8096	18289626.0	Ziva Fashion Women White Floral Yoke Design Ku...	2999.0	White	Ziva Fashion	NaN	NaN	White yoke design Kurta with Trousers with dup...	{'Add-Ons': 'NA', 'Body Shape ID': '333,424', ...	1016	Ziva Fashion
8097	2117164.0	Noi Cream-Coloured & Brown Printed Shawl	1999.0	Cream	Noi	5.0	3.400000	Cream-coloured and brown printed shawl, has fr...	{'Border': 'Printed', 'Fabric': 'Wool', 'Fabri...	637	Noi

8071 rows × 11 columns

Fig 2.0.3

Removing the null color row will reduce the row number of the dataset to 8,070.



```
newdata2.dropna( inplace = True )
newdata2
```

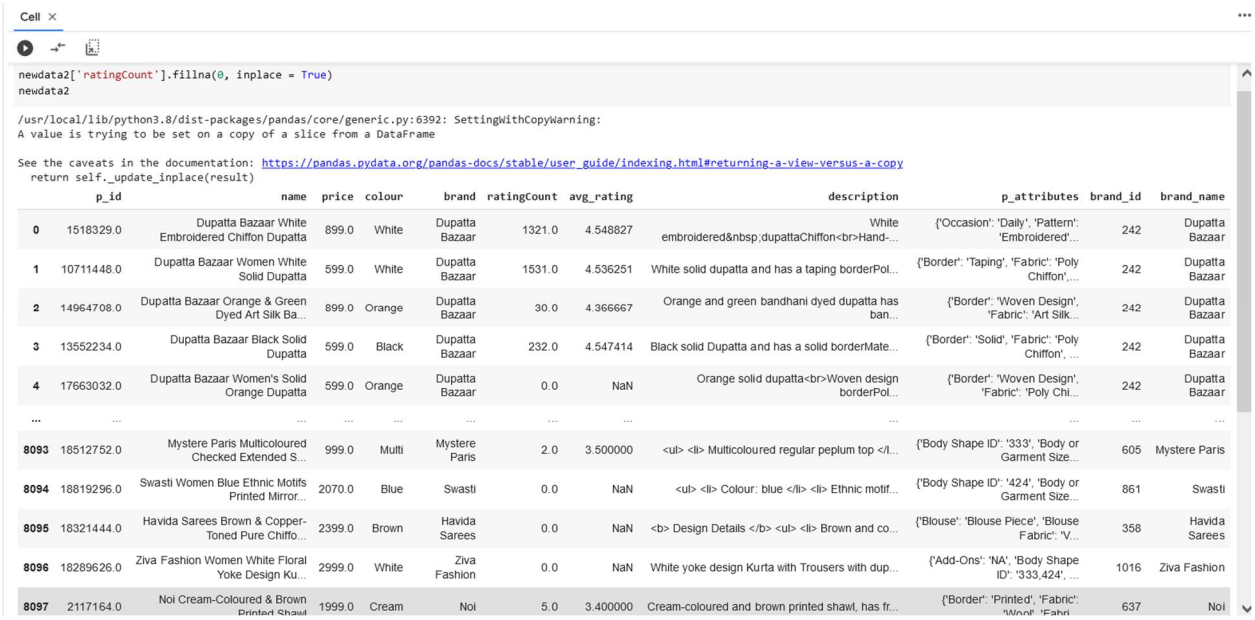
```
/usr/local/lib/python3.8/dist-packages/pandas/util/_decorators.py:311: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  return func(*args, **kwargs)
```

	p_id	name	price	colour	brand	ratingCount	avg_rating	description	p_attributes	brand_id	brand_name
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'...	242	Dupatta Bazaar
1	10711448.0	Dupatta Bazaar Women White Solid Dupatta	599.0	White	Dupatta Bazaar	1531.0	4.536251	White solid dupatta and has a taping borderPol...	{'Border': 'Taping', 'Fabric': 'Poly Chiffon',...	242	Dupatta Bazaar
2	14964708.0	Dupatta Bazaar Orange & Green Dyed Art Silk Ba...	899.0	Orange	Dupatta Bazaar	30.0	4.366667	Orange and green bandhani dyed dupatta has ban...	{'Border': 'Woven Design', 'Fabric': 'Art Silk...	242	Dupatta Bazaar
3	13552234.0	Dupatta Bazaar Black Solid Dupatta	599.0	Black	Dupatta Bazaar	232.0	4.547414	Black solid Dupatta and has a solid borderMate...	{'Border': 'Solid', 'Fabric': 'Poly Chiffon', ...	242	Dupatta Bazaar
4	17663032.0	Dupatta Bazaar Women's Solid Orange Dupatta	599.0	Orange	Dupatta Bazaar	0.0	0.000000	Orange solid dupatta Woven design borderPol...	{'Border': 'Woven Design', 'Fabric': 'Poly Chi...	242	Dupatta Bazaar
...	...	...	...	...	...	...	...	...	...	...	...
8093	18512752.0	Mystere Paris Multicoloured Checked Extended S...	999.0	Multi	Mystere Paris	2.0	3.500000	Multicoloured regular peplum top </l...	{'Body Shape ID': '333', 'Body or Garment Size...	605	Mystere Paris
8094	18819296.0	Swasti Women Blue Ethnic Motifs Printed Mirror...	2070.0	Blue	Swasti	0.0	0.000000	Colour: blue Ethnic motif...	{'Body Shape ID': '424', 'Body or Garment Size...	861	Swasti
8095	18321444.0	Havida Sarees Brown & Copper-Toned Pure Chiffo...	2399.0	Brown	Havida Sarees	0.0	0.000000	Design Details Brown and co...	{'Blouse': 'Blouse Piece', 'Blouse Fabric': 'V...	358	Havida Sarees
8096	18289626.0	Ziva Fashion Women White Floral Yoke Design Ku...	2999.0	White	Ziva Fashion	0.0	0.000000	White yoke design Kurta with Trousers with dup...	{'Add-Ons': 'NA', 'Body Shape ID': '333,424', ...	1016	Ziva Fashion
8097	2117164.0	Noi Cream-Coloured & Brown Printed Shawl	1999.0	Cream	Noi	5.0	3.400000	Cream-coloured and brown printed shawl, has fr...	{'Border': 'Printed', 'Fabric': 'Wool', 'Fabri...	637	Noi

Fig 2.0.4

The null ratingCount is being replaced by 0



```
newdata2['ratingCount'].fillna(0, inplace = True)
newdata2
```

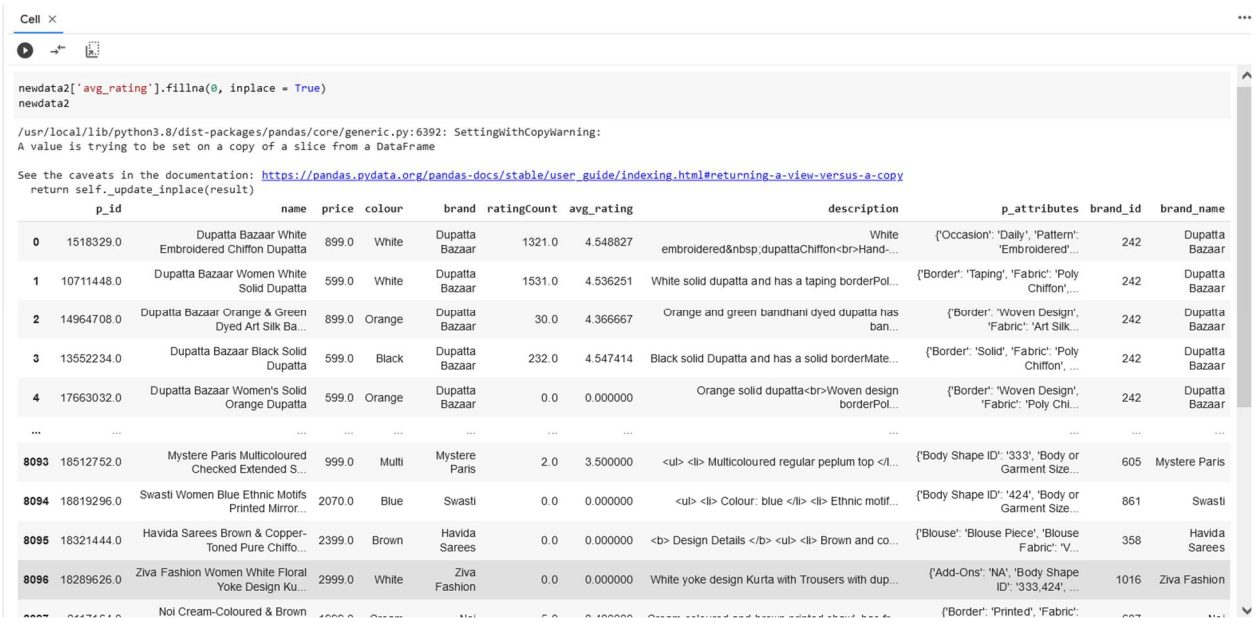
```
/usr/local/lib/python3.8/dist-packages/pandas/core/generic.py:6392: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  return self._update_inplace(result)
```

	p_id	name	price	colour	brand	ratingCount	avg_rating	description	p_attributes	brand_id	brand_name
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'...	242	Dupatta Bazaar
1	10711448.0	Dupatta Bazaar Women White Solid Dupatta	599.0	White	Dupatta Bazaar	1531.0	4.536251	White solid dupatta and has a taping borderPol...	{'Border': 'Taping', 'Fabric': 'Poly Chiffon',...	242	Dupatta Bazaar
2	14964708.0	Dupatta Bazaar Orange & Green Dyed Art Silk Ba...	899.0	Orange	Dupatta Bazaar	30.0	4.366667	Orange and green bandhani dyed dupatta has ban...	{'Border': 'Woven Design', 'Fabric': 'Art Silk...	242	Dupatta Bazaar
3	13552234.0	Dupatta Bazaar Black Solid Dupatta	599.0	Black	Dupatta Bazaar	232.0	4.547414	Black solid Dupatta and has a solid borderMate...	{'Border': 'Solid', 'Fabric': 'Poly Chiffon', ...	242	Dupatta Bazaar
4	17663032.0	Dupatta Bazaar Women's Solid Orange Dupatta	599.0	Orange	Dupatta Bazaar	0.0	NaN	Orange solid dupatta Woven design borderPol...	{'Border': 'Woven Design', 'Fabric': 'Poly Chi...	242	Dupatta Bazaar
...	...	...	...	...	...	...	...	...	...	...	...
8093	18512752.0	Mystere Paris Multicoloured Checked Extended S...	999.0	Multi	Mystere Paris	2.0	3.500000	Multicoloured regular peplum top </l...	{'Body Shape ID': '333', 'Body or Garment Size...	605	Mystere Paris
8094	18819296.0	Swasti Women Blue Ethnic Motifs Printed Mirror...	2070.0	Blue	Swasti	0.0	NaN	Colour: blue Ethnic motif...	{'Body Shape ID': '424', 'Body or Garment Size...	861	Swasti
8095	18321444.0	Havida Sarees Brown & Copper-Toned Pure Chiffo...	2399.0	Brown	Havida Sarees	0.0	NaN	Design Details Brown and co...	{'Blouse': 'Blouse Piece', 'Blouse Fabric': 'V...	358	Havida Sarees
8096	18289626.0	Ziva Fashion Women White Floral Yoke Design Ku...	2999.0	White	Ziva Fashion	0.0	NaN	White yoke design Kurta with Trousers with dup...	{'Add-Ons': 'NA', 'Body Shape ID': '333,424', ...	1016	Ziva Fashion
8097	2117164.0	Noi Cream-Coloured & Brown Printed Shawl	1999.0	Cream	Noi	5.0	3.400000	Cream-coloured and brown printed shawl, has fr...	{'Border': 'Printed', 'Fabric': 'Wool', 'Fabri...	637	Noi

Fig 2.0.5

The null avg_rating is being replaced by 0



```
newdata2['avg_rating'].fillna(0, inplace = True)
newdata2
```

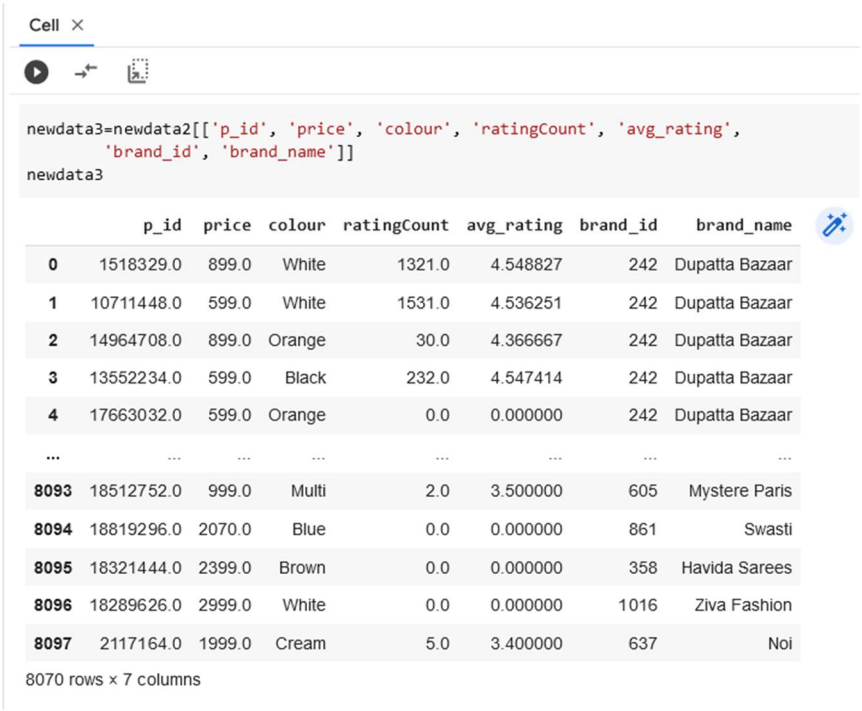
```
/usr/local/lib/python3.8/dist-packages/pandas/core/generic.py:6392: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  return self._update_inplace(result)
```

	p_id	name	price	colour	brand	ratingCount	avg_rating	description	p_attributes	brand_id	brand_name
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'...	242	Dupatta Bazaar
1	10711448.0	Dupatta Bazaar Women White Solid Dupatta	599.0	White	Dupatta Bazaar	1531.0	4.536251	White solid dupatta and has a taping borderPol...	{'Border': 'Taping', 'Fabric': 'Poly Chiffon',...	242	Dupatta Bazaar
2	14964708.0	Dupatta Bazaar Orange & Green Dyed Art Silk Ba...	899.0	Orange	Dupatta Bazaar	30.0	4.366667	Orange and green bandhani dyed dupatta has ban...	{'Border': 'Woven Design', 'Fabric': 'Art Silk...	242	Dupatta Bazaar
3	13552234.0	Dupatta Bazaar Black Solid Dupatta	599.0	Black	Dupatta Bazaar	232.0	4.547414	Black solid Dupatta and has a solid borderMate...	{'Border': 'Solid', 'Fabric': 'Poly Chiffon', ...	242	Dupatta Bazaar
4	17663032.0	Dupatta Bazaar Women's Solid Orange Dupatta	599.0	Orange	Dupatta Bazaar	0.0	0.000000	Orange solid dupatta Woven design borderPol...	{'Border': 'Woven Design', 'Fabric': 'Poly Chi...	242	Dupatta Bazaar
...	...	...	...	...	...	...	...	...	...	...	...
8093	18512752.0	Mystere Paris Multicoloured Checked Extended S...	999.0	Multi	Mystere Paris	2.0	3.500000	Multicoloured regular peplum top </l...	{'Body Shape ID': '333', 'Body or Garment Size...	605	Mystere Paris
8094	18819296.0	Swasti Women Blue Ethnic Motifs Printed Mirror...	2070.0	Blue	Swasti	0.0	0.000000	Colour: blue Ethnic motif...	{'Body Shape ID': '424', 'Body or Garment Size...	861	Swasti
8095	18321444.0	Havida Sarees Brown & Copper-Toned Pure Chiffo...	2399.0	Brown	Havida Sarees	0.0	0.000000	Design Details Brown and co...	{'Blouse': 'Blouse Piece', 'Blouse Fabric': 'V...	358	Havida Sarees
8096	18289626.0	Ziva Fashion Women White Floral Yoke Design Ku...	2999.0	White	Ziva Fashion	0.0	0.000000	White yoke design Kurta with Trousers with dup...	{'Add-Ons': 'NA', 'Body Shape ID': '333,424', ...	1016	Ziva Fashion
8097	2117164.0	Noi Cream-Coloured & Brown	1999.0	Cream	Noi	5.0	3.400000	Cream coloured and brown printed shawl, has fr...	{'Border': 'Printed', 'Fabric':	637	Noi

Fig 2.0.6

And to complete the cleaning process, the name, brand, description and p_attributes column will be removed, and we have this new table.



```
newdata3=newdata2[['p_id', 'price', 'colour', 'ratingCount', 'avg_rating',
       'brand_id', 'brand_name']]
newdata3
```

	p_id	price	colour	ratingCount	avg_rating	brand_id	brand_name
0	1518329.0	899.0	White	1321.0	4.548827	242	Dupatta Bazaar
1	10711448.0	599.0	White	1531.0	4.536251	242	Dupatta Bazaar
2	14964708.0	899.0	Orange	30.0	4.366667	242	Dupatta Bazaar
3	13552234.0	599.0	Black	232.0	4.547414	242	Dupatta Bazaar
4	17663032.0	599.0	Orange	0.0	0.000000	242	Dupatta Bazaar
...	...	...	...	...	...	...	...
8093	18512752.0	999.0	Multi	2.0	3.500000	605	Mystere Paris
8094	18819296.0	2070.0	Blue	0.0	0.000000	861	Swasti
8095	18321444.0	2399.0	Brown	0.0	0.000000	358	Havida Sarees
8096	18289626.0	2999.0	White	0.0	0.000000	1016	Ziva Fashion
8097	2117164.0	1999.0	Cream	5.0	3.400000	637	Noi

8070 rows × 7 columns

Fig 2.0.7 Cleaned dataset to suit research questions.

And finally, the cleaned dataset is now saved and will be used for subsequent analyzes in this report.

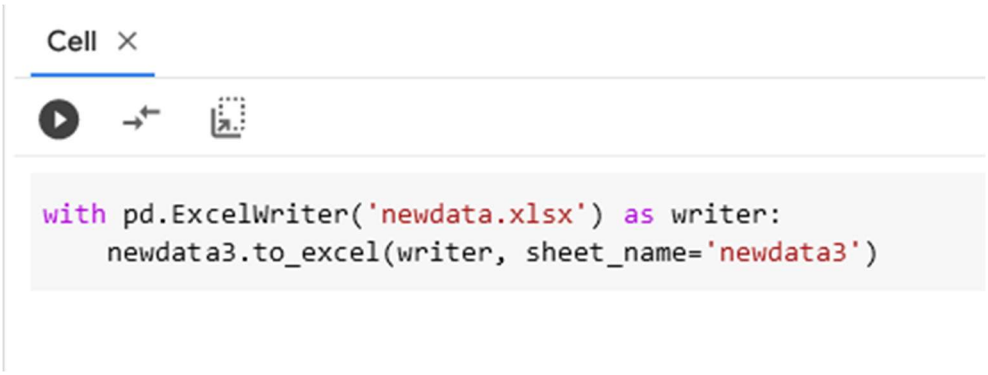

Cell
with pd.ExcelWriter('newdata.xlsx') as writer:
newdata3.to_excel(writer, sheet_name='newdata3')

Fig 2.0.8

3.0 Research Questions and Answers

The new data was loaded on the HIVE using the upload table button, And SQL queries are being used to get answers to research questions.

a) What is the average price of each color of the fashion product, and which has the highest and lowest value?

To answer this question, the new cleaned data will be imported into HIVE and SQL queries will be performed to see the unique colors and average price of each unique color in the cleaned dataset and the color with highest and lowest average price can be deduced.

To obtain distinct colors on the dataset, SQL query will be done on HIVE after uploading the dataset using the upload table.

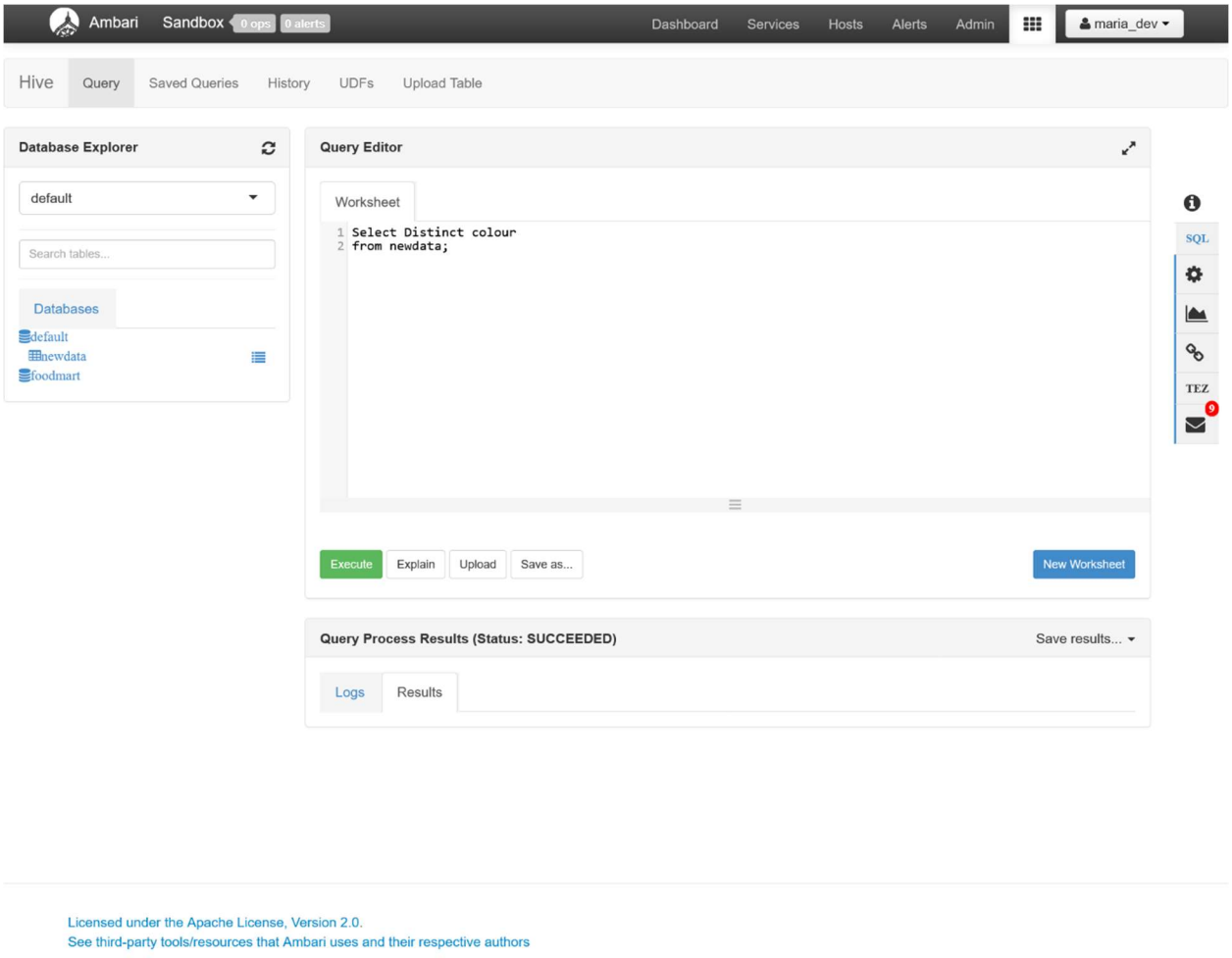


Fig 3.0.0

The query gave 48 unique colours.

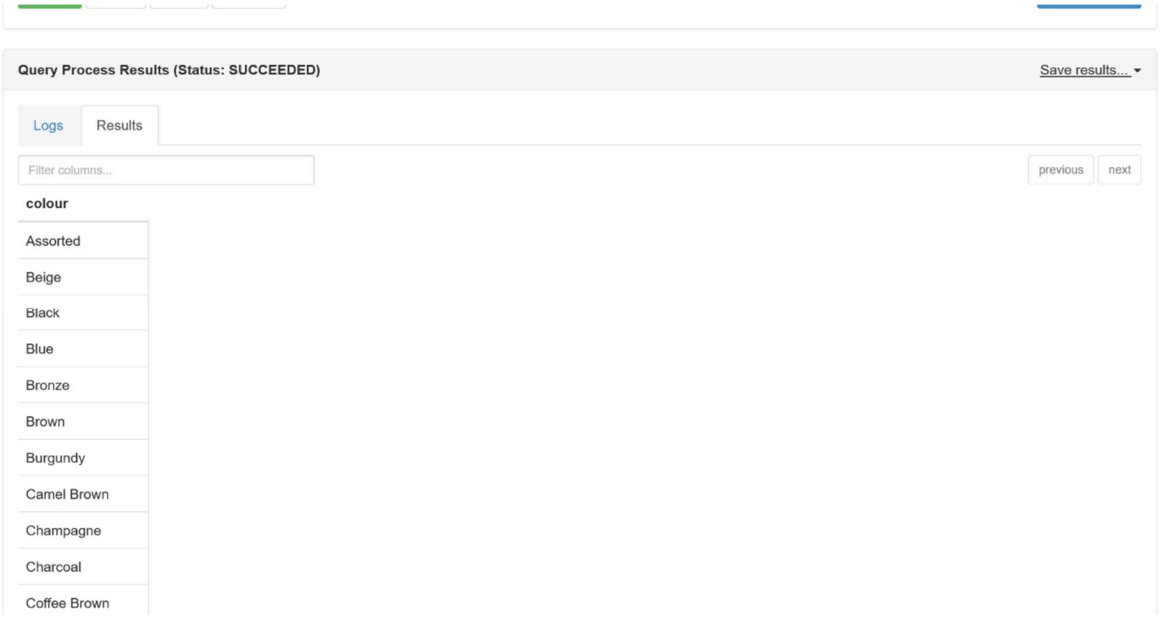


Fig 3.0.1

For color ‘Assorted’, running the SQL query below gives the average price for fashion products with the color.

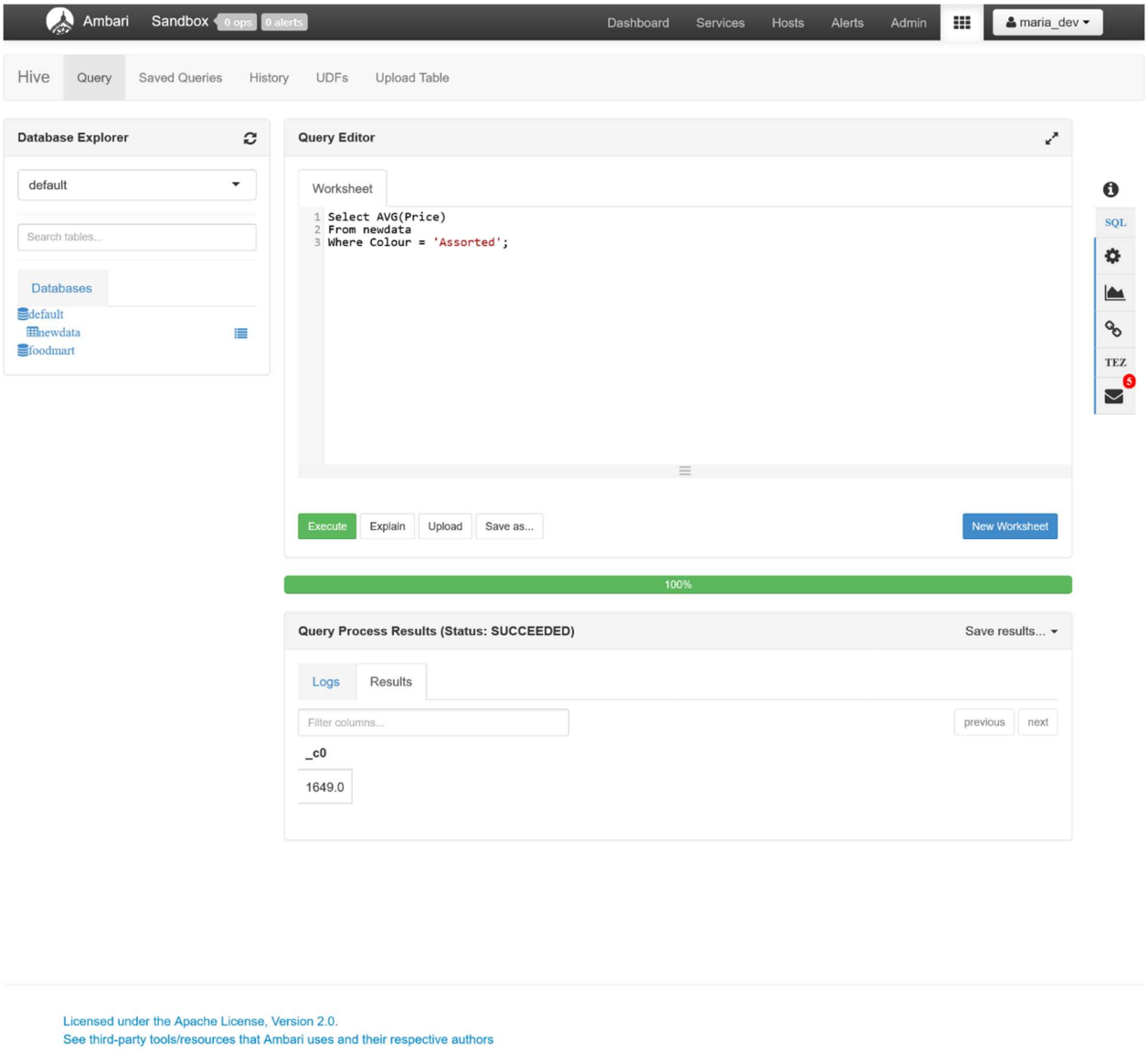


Fig 3.0.2

For color Beige, the SQL query below gives the average price

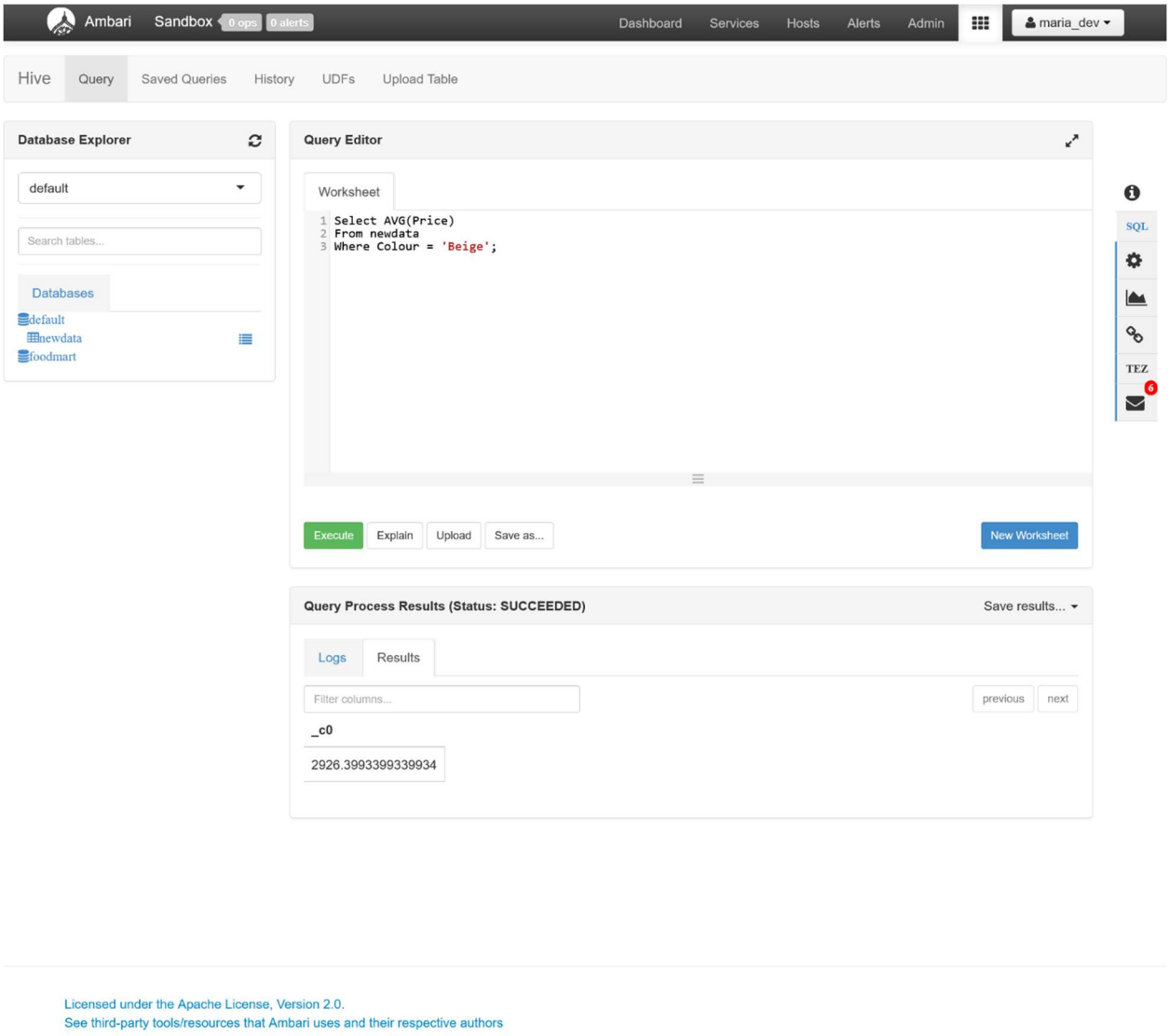


Fig 3.0.3

For color Black, the SQL query below gives the average price

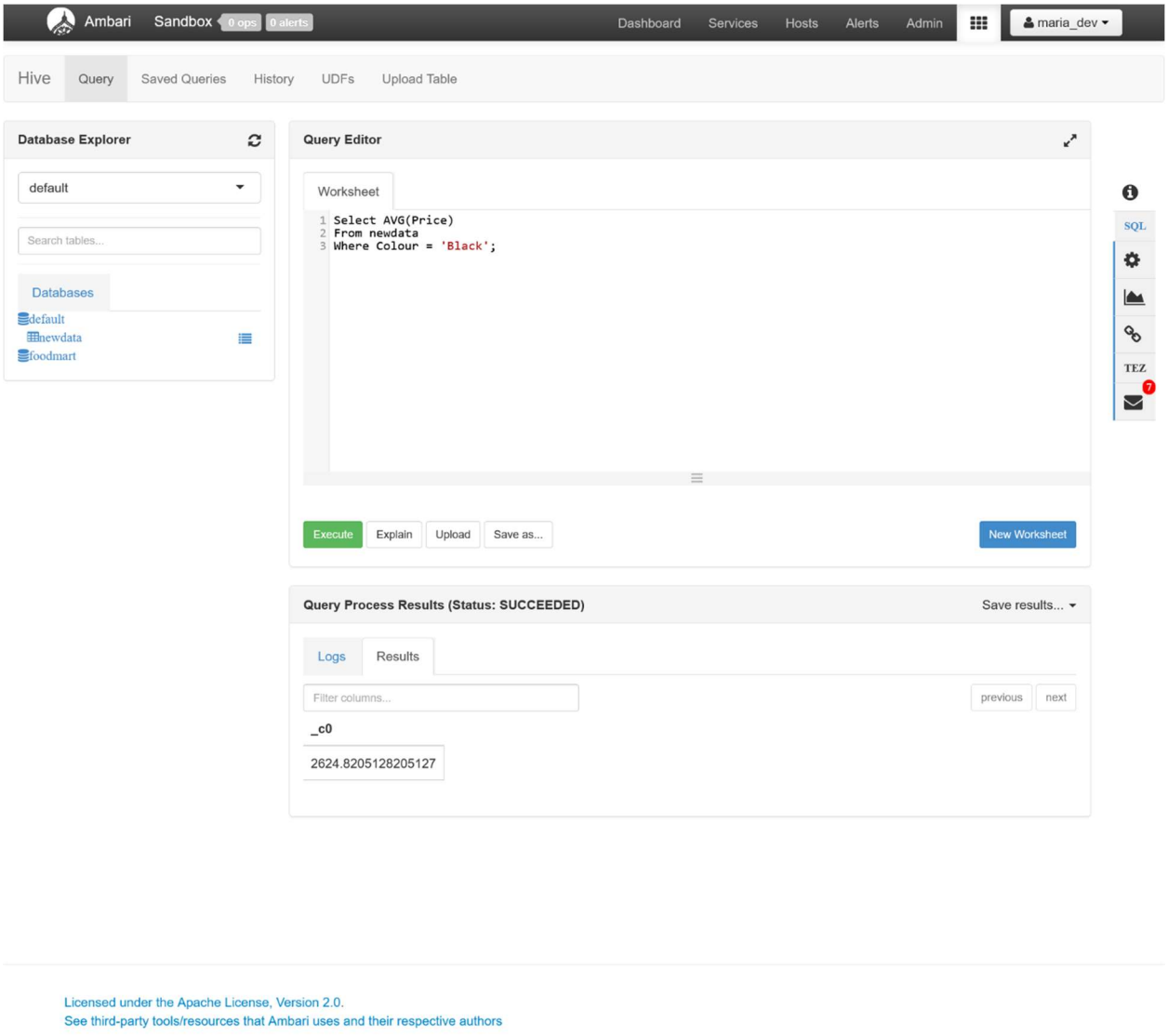


Fig 3.0.4

Changing the color name to respective distinct color, Average prices are being obtained and computed into the table shown below.

The table below shows Color and Average Price for each distinct color.

Table 1.0

Colour	Avg_Price
Assorted	1649.00
Beige	2926.40
Black	2624.82

Blue	2673.14
Bronze	2199.00
Brown	3000.44
Burgundy	3321.46
Camel Brown	2506.14
Champagne	2999.00
Charcoal	2627.68
Coffee Brown	3887.07
Copper	2199.00
Coral	2469.13
Cream	3186.92
Fluorescent Green	1912.67
Fuchsia	2479.12
Gold	2811.49
Green	3054.03
Grey	2992.08
Grey Melange	2403.63
Khaki	2490.50
Lavender	3608.27
Lime Green	2592.97
Magenta	3630.87
Maroon	2908.58
Mauve	3930.49
Multi	2678.61
Mustard	2818.76
Navy Blue	2541.90
Nude	2481.67
Off White	2750.09
Olive	2761.53
Orange	2960.33
Peach	3908.51
Pink	3008.76
Purple	3241.40
Red	2706.03
Rose	2887.25
Rust	2479.14
Sea Green	3827.97
Silver	2511.50
Tan	4067.0
Taupe	3204.56
Teal	4037.44
Turquoise Blue	3549.06
Violet	3841.67

White	2373.73
Yellow	3220.80

According to the table above, it is seen that ‘Assorted’ type of color has the lowest price average as 1,649.00 among other colors for the fashion brand and ‘Tan’ type of color is the highest price average as 4,067.00 among other colors.

It can then be deduced that color of fashion brand has impact on the price of the fashion brand and with the dataset at hand, it is seen that ‘Tan’ color type of fashion products are expensive and ‘Assorted’ color type of fashion products are cheapest.

b) What is the average price of each fashion brands and which fashion product brand having the highest and lowest average price?

To answer this question, the new cleaned data will be imported into HIVE and SQL queries will be performed to see.

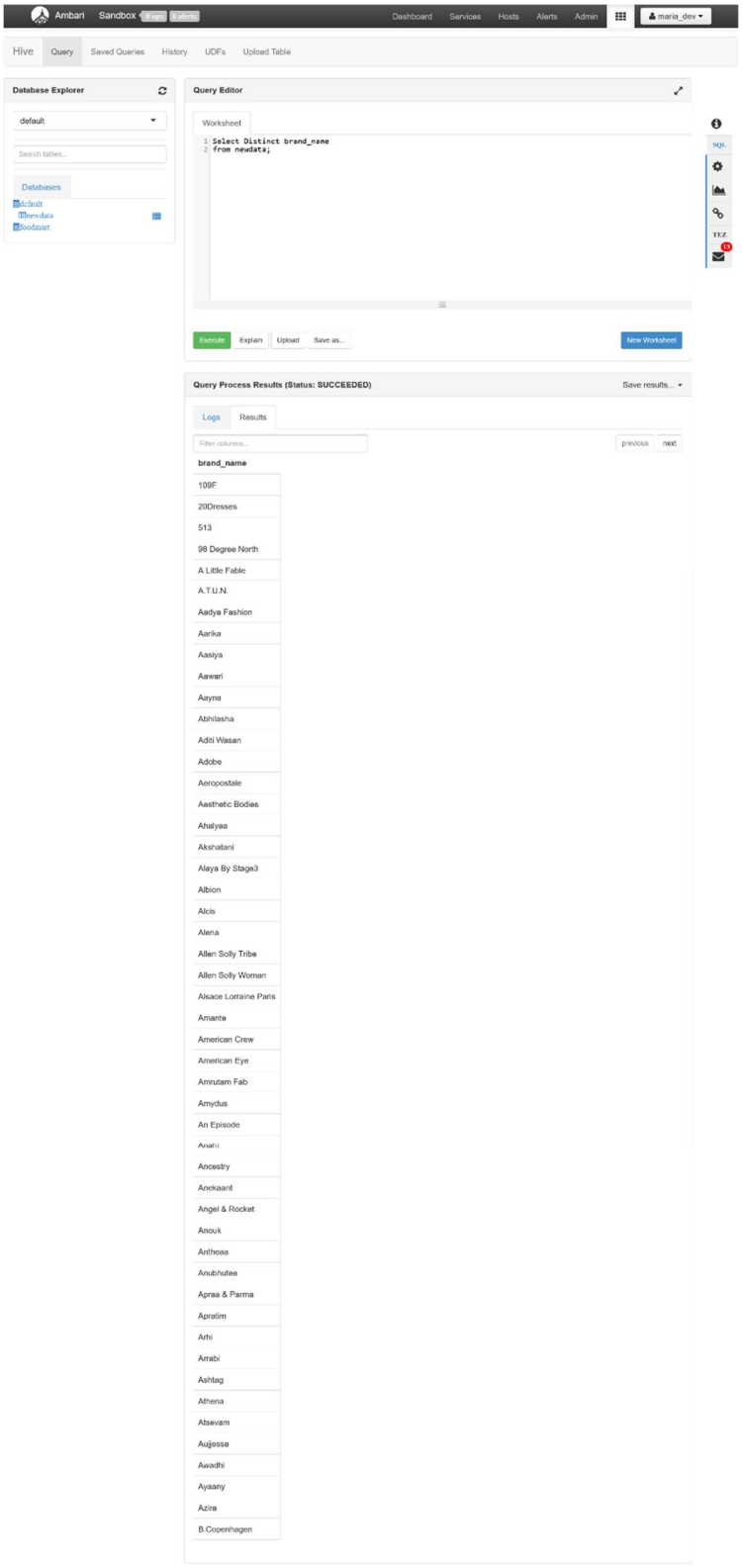


Fig 3.1.0

After running this query, 500 unique brand name was obtained.

To obtain the average price for each unique brand name, the following SQL query will be run

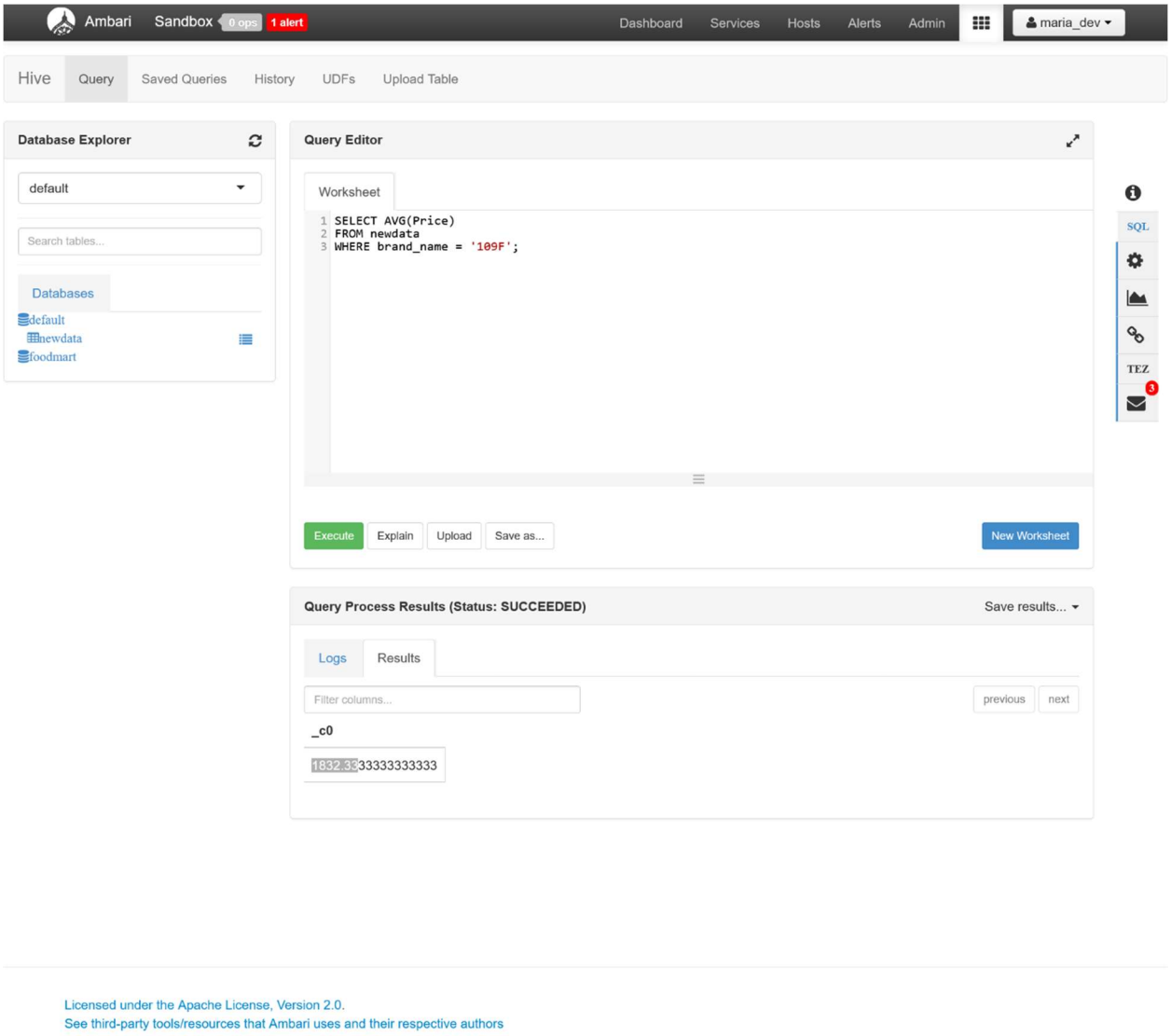


Fig 3.1.1

This query shows fashion brand 1099F has average price of 1,832.33

Running this query and changing the brand name gives the following table on each unique brand name

Table 2.0

brand_name	Average Price
109F	1832.33
20Dresses	2885.70
513	2099.00
98 Degree North	1999.00
A Little Fable	2690.00
A.T.U.N.	1999.00
Aadya Fashion	649.00
Aarika	2160.60
Aasiya	2499.00
Aawari	1045.00
Aayna	2625.20
Abhilasha	499.00
Aditi Wasan	1499.00
Adobe	5124.00
Aeropostale	2559.00
Aesthetic Bodies	2719.50
Ahalyaa	3584.03
Akshatani	1099.00
Alaya By Stage3	2061.19
Albion	836.85
Alcis	2512.33
Alena	1599.00
Allen Solly Tribe	1799.00
Allen Solly Woman	2701.63
Alsace Lorraine Paris	1784.14
Amante	1072.50
American Crew	2999.00
American Eye	1799.00
Amrutam Fab	11749.00
Amydus	2506.69
An Episode	1899.00
Anahi	349.00
Ancestry	3106.67
Anekaant	3340.67
Angel & Rocket	1299.00
Anouk	2572.36
Antheaa	2039.18
Anubhutee	2817.89
Apraa & Parma	1782.50

Apratim	1171.22
Arhi	4081.67
Arrabi	4999.00
Ashtag	2490.00
Athena	2197.56
Atsevam	5999.70
Aujjessa	2519.80
Awadhi	4499.00
Ayaany	1380.25
Azira	1359.00
B.Copenhagen	3935.71
Baby Lakshmi	1096.50
Baby Moo	999.00
Baisacrafts	5032.56
Banarasi Style	3099.00
Bani Women	999.00
Bannos Swagger	2519.00
Bavaria	3205.00
Being Human	2599.00
Belle Fille	2098.90
Belliskey	2599.00
Bene Kleed	1499.00
Berrylush	1871.62
Berrylush Curve	2164.00
Besiva	1859.00
Beverly Hills Polo Club	3799.00
Bewakoof	1515.67
Bhama Couture	2182.55
Biba	3313.21
Bitterlime	1927.57
Bitz	699.00
Blamblack	2540.00
Blissta	4984.64
Blue Giraffe	1449.00
Bossini	1461.50
Bronz	749.00
Calvin Klein Jeans	9165.67
Camey	1099.00
Campana	1999.00
Campus Sutra	2355.06
Candyskin	2099.00
Cantabil	2432.33
Carlton London	1919.00
Cation	2849.00

Cayman	3913.75
Charu	3037.46
Charukriti	3563.51
Chemistry	2247.84
Cherokee	1442.75
Chhabra 555	7164.81
Chidiyaa	1890.00
Chipbeys	2599.00
Chkokko	3458.53
Civil	699.00
Clora Creation	1464.38
Cloth Haus India	3750.00
Clovia	999.00
Club York	2479.00
Code 61	2859.00
Color Cocktail	1699.00
Color Trends	1999.00
Columbia	9499.00
Cot'N Soft	--
Cottinfab	2124.16
Creative Kids	999.00
Crimsoune Club	1911.50
Crozo By Cantabil	1999.00
Cultsport	3658.60
Cutiekins	1516.65
Darzi	1549.00
De Moza	1369.00
Deewa	2265.67
Desi Weavess	2177.20
Disrupt	1149.00
Divena	3724.00
Divine International Trading Co	6219.00
Diwaah	8916.36
Dollar Missy	1558.00
Domyos By Decathlon	2499.00
Dress My Angel	1099.00
Dupatta Bazaar	1281.16
Eavan	1899.00
Ekta Textiles	5365.67
Emprall	1199.00
Enamor	965.67
Enchanted Drapes	1300.00
Espresso	1170.29
Ethnic Junction	2695.10

Ethnic Yard	3992.33
Ethnicity	1539.00
F Loop	875.00
F.R.I.E.N.D.S By Sztori	2099.00
Fab Dadu	4199.00
Fab Viva	3999.00
Fabcartz	5999.00
Fabclub	999.00
Fabindia	2385.88
Fabnest Curve	2950.00
Fabriko	1999.00
Faserz	1278.00
Fashion Basket	3787.00
Fashion Booms	3999.00
Fashionuma	3434.71
Femella	2265.67
Femme Luxe	3424.00
Flambeur	2249.00
Flenzy	1199.00
Fleximaa	899.00
Florence	2221.00
Flying Machine	2228.79
Foreign Culture By Fort Collins	2387.50
Forever New	5314.29
Forevermore	2199.00
Fort Collins	2546.67
Free Authority	1599.00
Frempy	1299.00
Friskers	946.25
Fusion Beats	2074.00
Geroo Jaipur	7159.00
Get Glamr	2899.00
Get Wrapped	658.38
Ginni Arora Label	1449.00
Gipsy	2361.67
Global Desi	2202.08
Globus	1678.41
Go Colors	912.46
Golden Kite	3999.00
Grancy	6674.25
Granthva Fab	3999.00
Grubstaker	1285.00
H&M	1811.50
Hangup	1980.25

Harpa	1737.46
Harvard	1858.53
Havida Sarees	2399.00
Head	1449.00
Heart Up My Sleeves	2000.00
High Star	1743.84
Hive91	1214.33
Hopscotch	2299.00
Hubberholme	1499.00
Huetrap	1359.00
Hypernation	1173.09
I Love She	1899.00
I Saw It First	2974.00
Inddus	5673.71
Indethnic	5912.00
Indian Dobby	1699.00
Indibelle	1656.14
Indietoga	2733.50
Indo Era	4001.90
Instafab Plus	2619.00
Insua	3154.00
Invincible	1199.00
Ira Soleil	3455.00
Iris	3419.00
Ishin	5338.29
Iti	2515.67
Ives	2989.00
J Style	3756.00

4.0 SPARK

In this section, spark code will be run to display in descending order price and average rating of the cleaned dataset

To start with, necessary pyspark tools was installed using the code below

```
!apt-get install openjdk-8-jdk-headless -qq > /dev/null
!wget -q http://archive.apache.org/dist/spark/spark-3.1.1/spark-3.1.1-bin-hadoop3.2.tgz
!tar xf spark-3.1.1-bin-hadoop3.2.tgz
!pip install -q findspark
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
os.environ["SPARK_HOME"] = "/content/spark-3.1.1-bin-hadoop3.2"
import findspark
findspark.init()
from pyspark.sql import SparkSession
spark = SparkSession.builder.master("local[*]").getOrCreate()
spark.conf.set("spark.sql.repl.eagerEval.enabled", True)
```

Fig 4.0.0

The cleaned dataset was then imported and load to display the first 5 rows using the following query

```
df = spark.read.csv('newdata.csv', header=True, sep=",")
df.show(5)
```

```
+---+--------+-----+------+-----------+-----------+--------+--------------+
|_c0|    p_id|price|colour|ratingCount| avg_rating|brand_id|    brand_name|
+---+--------+-----+------+-----------+-----------+--------+--------------+
|  0| 1518329|  899| White|       1321|4.548826646|     242|Dupatta Bazaar|
|  1|10711448|  599| White|       1531|4.536250816|     242|Dupatta Bazaar|
|  2|14964708|  899|Orange|         30|4.366666667|     242|Dupatta Bazaar|
|  3|13552234|  599| Black|        232|4.547413793|     242|Dupatta Bazaar|
|  4|17663032|  599|Orange|          0|          0|     242|Dupatta Bazaar|
+---+--------+-----+------+-----------+-----------+--------+--------------+
only showing top 5 rows
```

Fig 4.0.1

The following query shows price in descending order that is the highest price.

```
df.orderBy('Price', ascending=False).show(truncate=False)
```

```
+----+--------+-----+-------+-----------+-----------+--------+------------------+
|_c0 |p_id    |price|colour |ratingCount|avg_rating |brand_id|brand_name        |
+----+--------+-----+-------+-----------+-----------+--------+------------------+
|1750|17325496|9999 |Green  |0          |0          |583     |Mitera            |
|2428|14602198|9999 |Blue   |16         |4.4375     |509     |Libas             |
|1751|17325490|9999 |Black  |0          |0          |583     |Mitera            |
|5797|15869432|9999 |Teal   |0          |0          |626     |New Balance       |
|5310|18095232|9999 |Green  |0          |0          |969     |Warthy Ent        |
|5311|18095230|9999 |Green  |0          |0          |969     |Warthy Ent        |
|7085|12966308|9999 |Blue   |0          |0          |159     |Calvin Klein Jeans|
|7831|18549594|9999 |Mustard|0          |0          |61      |Amrutam Fab       |
|3965|17515092|9999 |White  |0          |0          |706     |Puma              |
|4556|16215632|9999 |Peach  |14         |3.857142857|408     |Ishin             |
|1787|17325498|9999 |Maroon |0          |0          |583     |Mitera            |
|5306|18095246|9999 |Orange |0          |0          |969     |Warthy Ent        |
|5309|18095234|9999 |Green  |0          |0          |969     |Warthy Ent        |
|1788|17325492|9999 |Green  |0          |0          |583     |Mitera            |
|2815|16376838|9998 |Brown  |0          |0          |832     |Soch              |
|2824|17035692|9998 |Yellow |0          |0          |832     |Soch              |
|2768|17035678|9998 |Purple |0          |0          |832     |Soch              |
|2859|18644402|9998 |Mauve  |0          |0          |832     |Soch              |
|4729|16420664|9995 |Pink   |16         |3.6875     |543     |Mameraa           |
|4647|15795764|9990 |Pink   |0          |0          |173     |Chhabra 555       |
+----+--------+-----+-------+-----------+-----------+--------+------------------+
only showing top 20 rows
```

Fig 4.0.2 First 20 rows of price in descending order

And lastly the query below shows average rating in descending order that is the highest average rating

```
df.orderBy('avg_rating', ascending=False).show(truncate=False)
```

```
+----+--------+-----+--------------+-----------+----------+--------+------------------+
|_c0 |p_id    |price|colour        |ratingCount|avg_rating|brand_id|brand_name        |
+----+--------+-----+--------------+-----------+----------+--------+------------------+
|1648|15889850|3299 |Teal          |6          |5         |868     |Taavi             |
|1659|17664432|1899 |Brown         |7          |5         |868     |Taavi             |
|1528|15823664|1999 |Beige         |6          |5         |353     |H&M               |
|1757|11530694|6248 |Peach         |2          |5         |583     |Mitera            |
|256 |15089126|1599 |Olive         |5          |5         |750     |Roadster          |
|1778|16331704|5999 |Yellow        |9          |5         |583     |Mitera            |
|601 |17535088|5899 |Peach         |2          |5         |389     |Inddus            |
|1849|16291334|6999 |Red           |2          |5         |583     |Mitera            |
|101 |11514822|899  |Blue          |5          |5         |242     |Dupatta Bazaar    |
|1935|15888408|5100 |Blue          |5          |5         |734     |Readiprint Fashions|
|1374|16600180|3299 |Navy Blue     |5          |5         |52      |Allen Solly Woman |
|2212|16630112|1299 |Beige         |7          |5         |776     |Sangria           |
|1393|17548400|1699 |Green         |3          |5         |52      |Allen Solly Woman |
|2228|16844316|1299 |Red           |3          |5         |776     |Sangria           |
|2812|16462232|1298 |Turquoise Blue|5          |5         |832     |Soch              |
|2267|14448166|1999 |Black         |6          |5         |776     |Sangria           |
|251 |15441312|2799 |Brown         |5          |5         |750     |Roadster          |
|2279|17299430|3299 |Burgundy      |4          |5         |776     |Sangria           |
|699 |16200858|1649 |Olive         |8          |5         |903     |Tokyo Talkies     |
|2374|14094034|2499 |Pink          |5          |5         |354     |Hangup            |
+----+--------+-----+--------------+-----------+----------+--------+------------------+
only showing top 20 rows
```

Fig 4.0.3

5.0 MACHINE LEARNING

For the purpose of this report, the machine learning will be done using one of the supervised learning techniques which is the Support Vector Regression (SVR). As support vector helps with both linear and nonlinear regressions.

The machine learning is about two dependent variables which are the brand_id and ratingCount and one independent variable which is Price. So, it imply a model was created to determine price given brand_id and ratingCount.

To start with, necessary tools are being imported into the colab which is the python interface used here. And as seen below

```
import pandas as pd
from sklearn import linear_model
from sklearn import preprocessing
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error, mean_squared_error
import matplotlib.pyplot as plt
```

Fig 5.0.0

The cleaned data is being uploaded and read using panda feature and as seen below

```
newdata = pd.read_csv('newdata.csv')
newdata
```

	Unnamed: 0	p_id	price	colour	ratingCount	avg_rating	brand_id	brand_name
0	0	1518329	899	White	1321	4.548827	242	Dupatta Bazaar
1	1	10711448	599	White	1531	4.536251	242	Dupatta Bazaar
2	2	14964708	899	Orange	30	4.366667	242	Dupatta Bazaar
3	3	13552234	599	Black	232	4.547414	242	Dupatta Bazaar
4	4	17663032	599	Orange	0	0.000000	242	Dupatta Bazaar
...	...	...	...	...	...	...	...	...
8065	8093	18512752	999	Multi	2	3.500000	605	Mystere Paris
8066	8094	18819296	2070	Blue	0	0.000000	861	Swasti
8067	8095	18321444	2399	Brown	0	0.000000	358	Havida Sarees
8068	8096	18289626	2999	White	0	0.000000	1016	Ziva Fashion
8069	8097	2117164	1999	Cream	5	3.400000	637	Noi

8070 rows × 8 columns

Fig 5.0.1

The values for x being the dependent variables was assigned and value for y being independent variable was also assigned below

```
x = newdata[['brand_id', 'ratingCount' ]]
y = newdata['price']
```

Fig 5.0.2

The test size was made to be 0.2, random state = 42 and train size being 0.8 as seen below

```
x_train, x_test,y_train, y_test = train_test_split(x,y, test_size=0.2, random_state=42)
```

Fig 5.0.3

And the x train values as

x_train

	brand_id	ratingCount
2821	832	0
7708	766	0
1393	52	0
812	903	35
2215	776	0
...	...	...
5226	336	0
5390	921	9
860	903	16
7603	879	0
7270	327	0

6456 rows × 2 columns

Fig 5.0.4

And x test values as

x_test

	brand_id	ratingCount
742	903	54
2127	488	0
2107	488	0
2943	968	163
6691	344	0
...	...	...
3196	554	41
6304	208	0
7024	463	0
3473	46	0
4778	77	0

1614 rows × 2 columns

Fig 5.0.5

The SVR function was then imported, and linear kernel was first used for the model as shown below

```
from sklearn.svm import SVR
model= SVR(kernel='linear')
model.fit(x_train, y_train)
```

```
SVR(kernel='linear')
```

Fig 5.0.6

And predict_y values as

```
predict_y = model.predict(x_test)
predict_y
```

```
array([2337.72238901, 2172.12051991, 2172.12051991, ..., 2161.54469422,
       1985.13992162, 1998.25394548])
```

Fig 5.0.7

The mean squared error, mean absolute error and Root mean squared error obtained below

```
from sklearn import metrics
import numpy as np
print("Mean squared error", metrics.mean_squared_error(y_test, predict_y))
print('Mean absolute error', metrics.mean_absolute_error(y_test, predict_y))
print('Root mean squared error', np.sqrt(metrics.mean_squared_error(y_test, predict_y)))
```

```
Mean squared error 7716760.953476627
Mean absolute error 1448.4676099070236
Root mean squared error 2777.905857561884
```

Fig 5.0.8

Changing the kernel to rbf to get the kernel with the best error

```
from sklearn.svm import SVR
model= SVR(kernel='rbf')
model.fit(x_train, y_train)
```

```
SVR()
```

Fig 5.0.9

With predict_y as

```
predict_y = model.predict(x_test)
predict_y
```

```
array([2243.77109898, 2172.5833294 , 2172.5833294 , ..., 2162.01837362,
       2018.41531692, 2023.614834  ])
```

Fig 5.1.0

The mean squared, mean absolute and root mean squared errors for rbf as shown below

```
from sklearn import metrics
import numpy as np
print("Mean squared error", metrics.mean_squared_error(y_test, predict_y))
print('Mean absolute error', metrics.mean_absolute_error(y_test, predict_y))
print('Root mean squared error', np.sqrt(metrics.mean_squared_error(y_test, predict_y)))
```

```
Mean squared error 7729786.007062701
Mean absolute error 1445.5651098940523
Root mean squared error 2780.2492706702938
```

Fig 5.1.1

Changing the kernel to poly as seen below

```
from sklearn.svm import SVR
model= SVR(kernel='poly')
model.fit(x_train, y_train)
```

```
SVR(kernel='poly')
```

Fig 5.1.2

The y_predict as shown below

```
predict_y = model.predict(x_test)
predict_y
```

```
array([2300.81531998, 2134.00248026, 2134.00248026, ..., 2128.56701937,
       2096.79162895, 2096.90673728])
```

Fig 5.1.3

The mean squared, mean absolute and root mean squared errors for poly as shown below

```
from sklearn import metrics
import numpy as np
print("Mean squared error", metrics.mean_squared_error(y_test, predict_y))
print('Mean absolute error', metrics.mean_absolute_error(y_test, predict_y))
print('Root mean squared error', np.sqrt(metrics.mean_squared_error(y_test, predict_y)))
```

```
Mean squared error 7747436.448201017
Mean absolute error 1449.0716482447328
Root mean squared error 2783.421715838442
```

Fig 5.1.4

As it is seen that the kernel linear has the lowest errors and hence the best kernel for the support vector regression analysis model for this dataset.

6.0 DATA VISUALIZATION

The cleaned dataset will be visualized using python seaborn feature and Powerbi. These two applications are great big data visualization tool and help show data analysis in graphical way.

6.1 POWERBI VISUALIZATION

Using Powerbi, the two research questions will be visualized using stacked chart. After launching the Powerbi Desktop. The cleaned dataset is being imported from excel as shown in the figure below

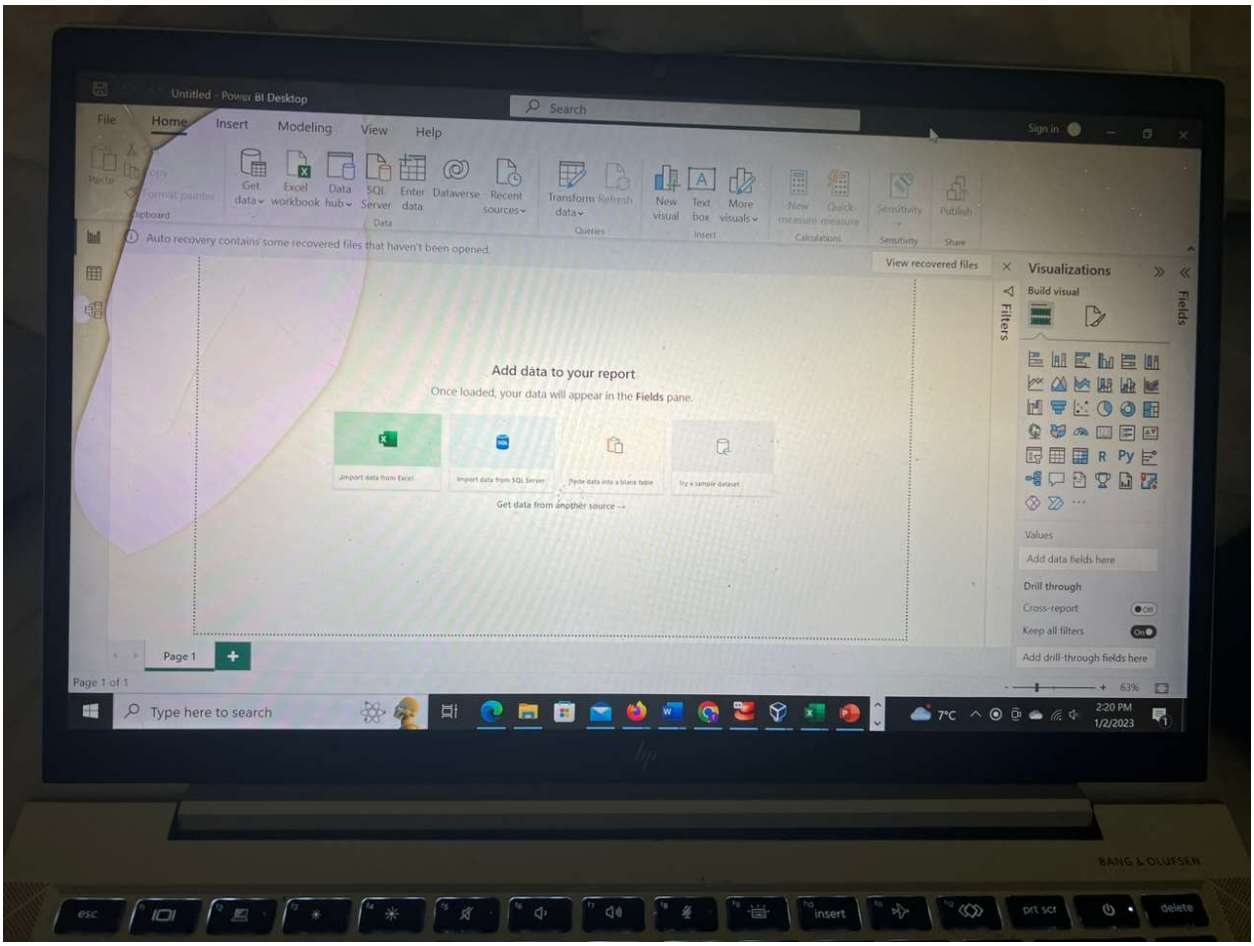


Fig 6.1.0

After importing the dataset, and because it is a cleaned a dataset, the transform data function is not needed, the dataset was loaded into the visualization for modelling.

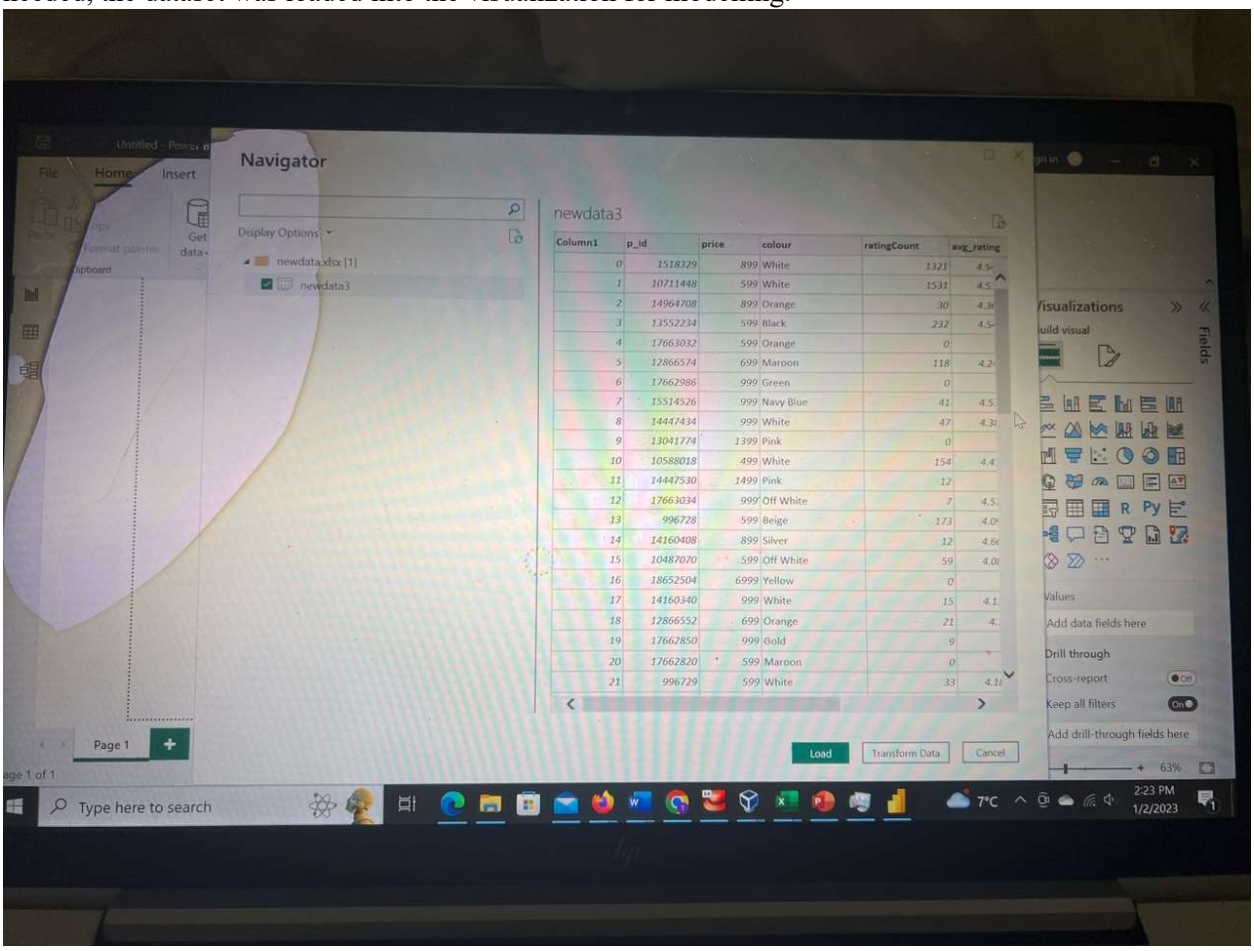


Fig 6.1.1

The data is then modelled and visualized in accordance with the research question

The following figure shows the visualization of first 27 colors with high average price. Which still connote that Tan has the highest average price. Color is being modelled on the y- axis and Average of price is on the x-axis

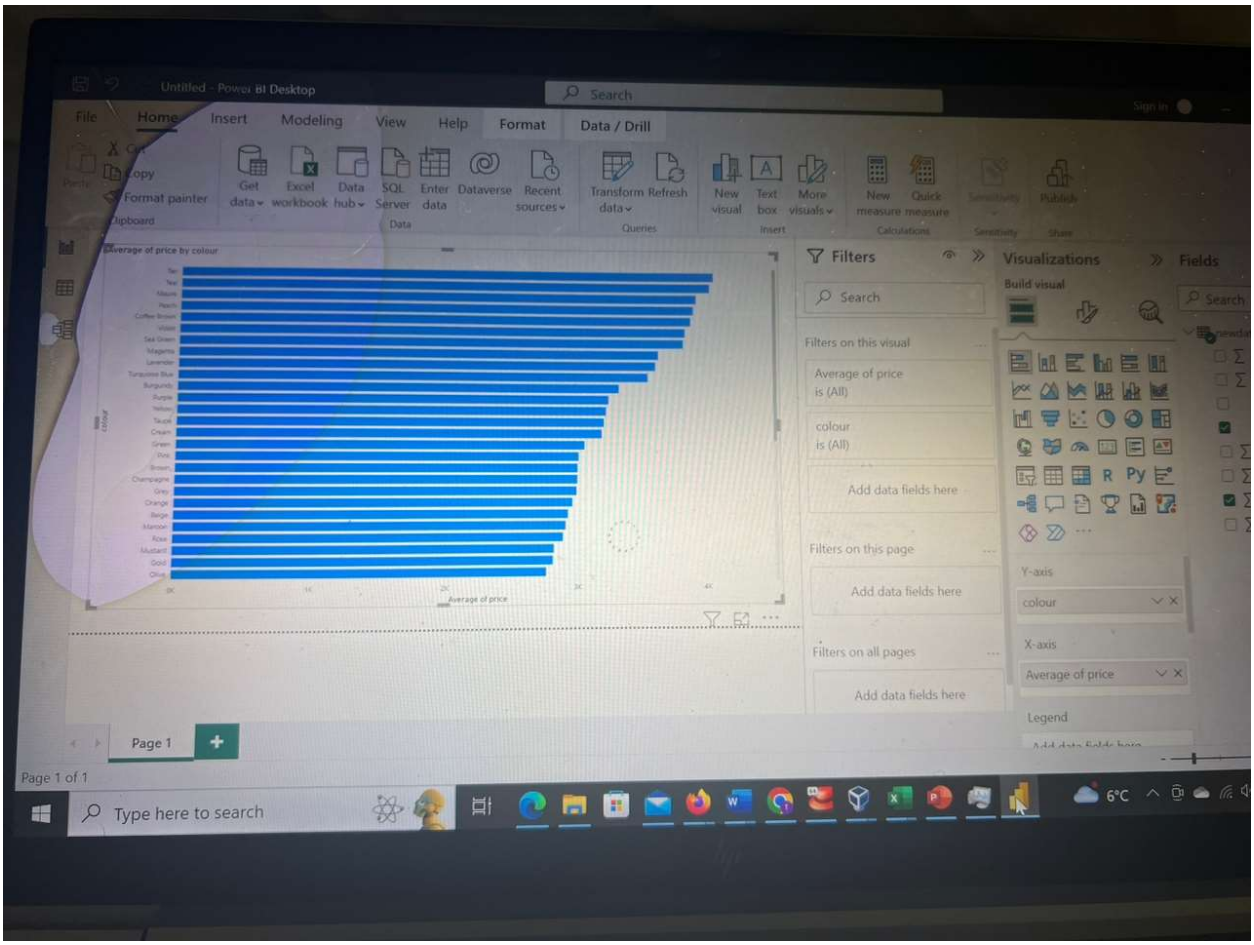


Fig 6.1.2

The following figure shows brand name with the highest average price. With brand name modelled on the y-axis and average price on the x-axis.

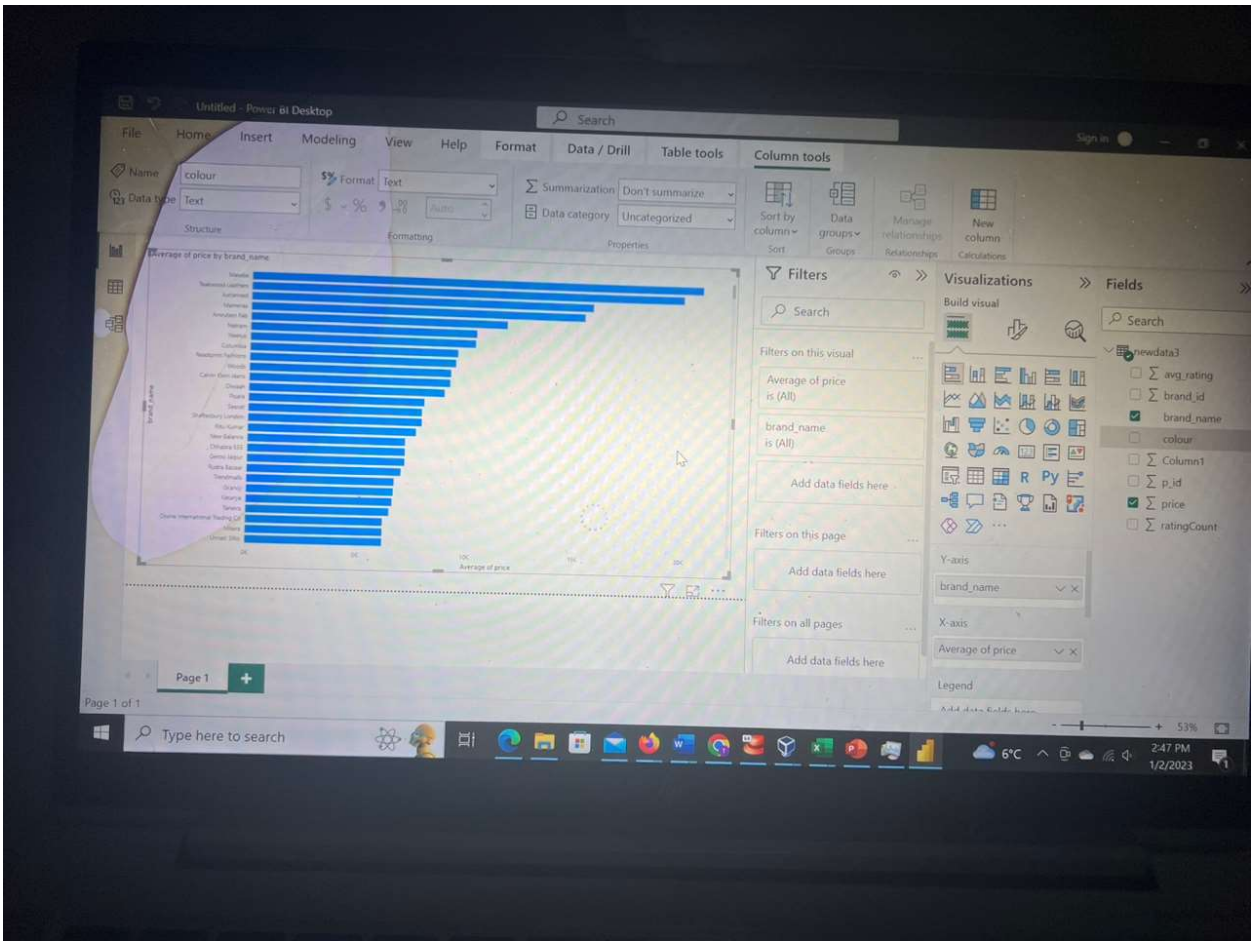


Fig 6.1.3

And that completes the visualization of the dataset with accordance to the two research questions.

6.2 PYTHON VISUALIZATION

The cleaned data set will be visualized in python

This will be done by first importing necessary libraries

```
import pandas as pd
from sklearn import linear_model
from sklearn import preprocessing
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error, mean_squared_error
import matplotlib.pyplot as plt
```

Fig 6.2.1

The cleaned dataset is then being uploaded and read as seen below

```
newdata = pd.read_csv('newdata.csv')
newdata
```

	Unnamed: 0	p_id	price	colour	ratingCount	avg_rating	brand_id	brand_name
0	0	1518329	899	White	1321	4.548827	242	Dupatta Bazaar
1	1	10711448	599	White	1531	4.536251	242	Dupatta Bazaar
2	2	14964708	899	Orange	30	4.366667	242	Dupatta Bazaar
3	3	13552234	599	Black	232	4.547414	242	Dupatta Bazaar
4	4	17663032	599	Orange	0	0.000000	242	Dupatta Bazaar
...	...	...	...	...	...	...	...	...
8065	8093	18512752	999	Multi	2	3.500000	605	Mystere Paris
8066	8094	18819296	2070	Blue	0	0.000000	861	Swasti
8067	8095	18321444	2399	Brown	0	0.000000	358	Havida Sarees
8068	8096	18289626	2999	White	0	0.000000	1016	Ziva Fashion
8069	8097	2117164	1999	Cream	5	3.400000	637	Noi

8070 rows × 8 columns

Fig 6.2.2

The first plot is a bar chart of brand_name against price as seen below

```
f, ax = plt.subplots(figsize=(18,5))
x_axis=newdata.brand_name
y_axis=newdata.price
plt.bar(x_axis, y_axis, label = 'first bar')

ax.legend(fontsize = 14)
plt.show()
```

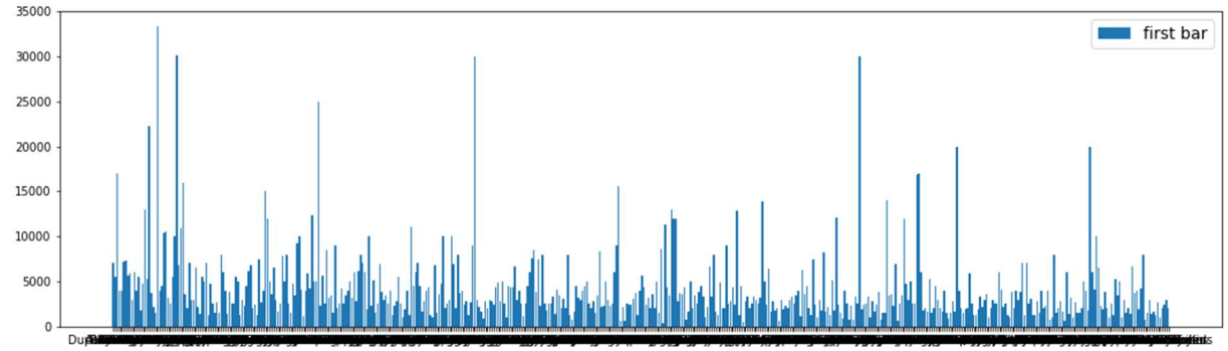


Fig 6.2.3

And the second plot is a bar chart of colour against price as seen below

```
f, ax = plt.subplots(figsize=(18,5))
x_axis=newdata.colour
y_axis=newdata.price
plt.bar(x_axis, y_axis, label = 'Second bar')

ax.legend(fontsize = 14)
plt.show()
```

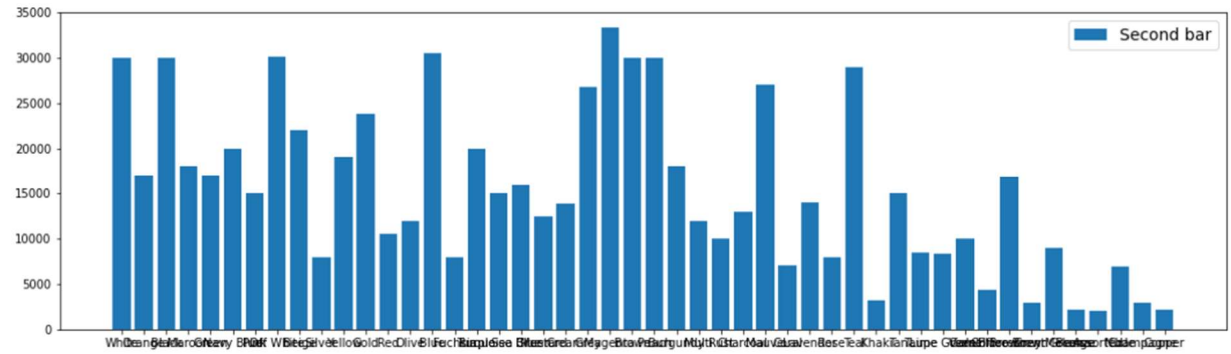


Fig 6.2.4 Bar chart showing color and price

PART II

TOPIC: BIG DATA ISSUES FACING MODERN FASHION INDUSTRY

ABSTRACT

The fashion industry, which involves making of clothing and textiles since inception have been involved in revolutionary trends of industry and presently has keyed into the benefit of 4.0 which involves the use of big data in analyzing demands and supply which is the mainstream for an effective profit for the industry. The fashion industry uses both analytical and visualization approach for data analysis which includes programming languages such as Python, Powerbi, HIVE, Mapreduce and SQL queries. Having keyed into the benefit of big data the industry also subscribed to the issues of big data applications and with pandemic which hit the world in 2020 and pulled the world economy into recession, major industry using big data analytics including fashion industry now want a more profitable trend and to elope the issues facing application of big data in the industry.

KEYWORDS: Big Data, Modern Fashion Industry, Python, Demand and Supply, Consumer, Data Privacy, Data Quality.

INTRODUCTION

The ongoing transformation of societies and economies toward organizational paradigms profoundly influenced by digital technologies is at the center of contemporary debates, engaging researchers and affecting a wide range of fields, from the humanities to science and technology.

As a result, the so-called "Fourth Industrial Revolution" has been described as a model in which new patterns of production and consumption will radically disrupt all major industrial systems and has been identified as a goal for a sustainable future in numerous government programs. While fundamental frameworks characterizing the 4.0 paradigm are codified and readily available, implementation tactics and their consequences for local and sectorial systems remain largely unexplored. From this premise, the essay explores the current state of the art and main developments of the "Fourth Industrial Revolution," perhaps identifying its effects on the textile and garment industries.

The fashion sector has seen tremendous change, notably during the past two decades, when the industry's borders began to expand (Djelic & Ainamo., 1999). Since then, the altering fashion industry dynamics, such as the decline of mass production, the increase in the number of fashion seasons, and the modification of the supply chain's structural characteristics, have compelled retailers to seek low costs and flexibility in design, quality, delivery, and speed to market (Doyle, et al., 2006). In addition to speed to market and design, marketing and capital expenditure have been cited as competitive drivers in the fashion garment sector (Sinha, 2001). (Franks, 2000) proposed' sense and respond' as the most important technique for maintaining a profitable position in a market that is becoming increasingly dynamic and demanding. In this environment, maintaining closer relationships with suppliers and buyers is a crucial defining attribute of rapid reaction and increased flexibility (Wheelright & Clark., 1992).

Until the late 1980s, traditional fashion apparel retailers competed based on their ability to predict customer demand and fashion trends (referred to as ready-to-wear) well in advance of the actual period of consumption (Guercini, 2001). In recent years, however, fashion shops have competed with one another by assuring speed to market by fast delivering the fashion trends shown at fashion shows and on runways. According to (Taplin, 1999), such businesses should be attributed with the adoption of "rapid fashion," which is the result of an unexpected process resulting in a shorter time gap between seasonal design and consumption.

The extremely competitive nature of today's fashion market and the ongoing need to 'refresh' product lines compel many shops to increase the number of 'seasons', or the frequency with which an entire store's stock is replaced. With the emergence of tiny collections of items, fashion merchants are encouraging customers to visit their stores more frequently by promoting the concept of "Here Today, Gone Tomorrow." This signifies a shorter life cycle and larger profit margins from the sale of products that sells quickly, so bypassing the markdown process (Sydney., 2008). In addition, shoppers favor Zara and H&M due to their demand for variety and rapid pleasure accompanied by price savagery (Post., 2009).

Several studies have examined various facets of the buyer-supplier relationship in quick or fast fashion, such as the apparel design process in relation to quick response (Forza & Vinelli., 1996) the role of the supplier in fast moving fashion (Doyle, et al., 2006), buyer behavior (Bruce & Daly., 2006), and financial performance (Hayes & Jones., 2006) . However, there appears to be a void in the literature concerning the emergence of the notion of "fast fashion" in the fashion business from the consumer's perspective. Few studies have focused on the consumer characteristics that drive the changes in the fashion industry (Barnes & Lea-Greenwood., 2006) among the numerous studies on rapid fashion.

The objective of this article is to examine the changes that have occurred in the fashion apparel business over the past two decades and to attempt to comprehend the emergence of fast fashion to its current level. The study focuses on the changes in the fashion garment business that led to the emergence of disposable or fast fashion. A quick literature review assists to systematize and evaluate the available material. This study also aims to connect research skills with market growth potential for fast fashion and recommends new venues for doing research in order to gain a better knowledge of fast fashion as a consumer-driven, as opposed to a supplier-driven, approach.

METHODOLOGY

Several research journals and academic which explicitly discusses the fashion industry and evolutionary trends regarding big data technologies were reviewed and guided in ideas of this report and are well cited and referenced in this report.

RESEARCH QUESTIONS/ ANSWERS

This research will focus on five (5) issues that arises from implementation of big data technologies by the fashion industry

1. Data Quality
   Fashion companies often work with large amounts of data from a variety of sources such as customer transactions, social media and supply chain information. Ensuring quality and accuracy of this data can be challenging. Data quality has six basic rules which are: validity, accuracy, consistency, integrity, timeless and completeness. Just as the data that was worked on in part I of this report, the quality of the data cannot be ascertain with great precision, moreover, problem of data quality has been associated with big data from inception. The quality of data will affect the quality of analysis and data with bad quality will lead to bad results and visualization hence, making the fashion industry run loss and this is what majority of industries are avoiding and making them adopt big data. Bad data can have major commercial ramifications for firms. Poor-quality data is often cited as the root of operational snafus, erroneous analytics and ill-conceived company plans. Data quality issues can result in additional costs when products are supplied to incorrect customer addresses, lost sales opportunities due to inaccurate or incomplete customer records, and penalties for improper financial or regulatory compliance reporting. In 2021, Gartner, a consulting firm, estimated that poor data quality costs businesses an average of $12,9 million annually. IBM's 2016 estimation that the annual cost of data quality issues in the United States was $3.1 trillion is still frequently referenced. Thomas Redman, a data quality consultant, estimated in a 2017 article for the MIT Sloan Management Review that rectifying data inaccuracies and dealing with business difficulties caused by poor data costs organizations 15% to 25% of their yearly sales on average.
   Moreover, a lack of trust in data on the part of corporate leaders and business managers is frequently recognized as one of the most significant barriers to the use of business intelligence (BI) and analytics tools to improve organizational decision-making. All of this necessitates an efficient data quality management plan. (Craig, 2022).

2. Data Security
   With the sensitive personal and financial information that is often stored in fashion company database, data security is a critical concern. Hackers and cyber criminals may target fashion companies in order to access this information. Data security is the discipline of protecting digital information throughout its full lifecycle from unwanted access, corruption, and theft. It incorporates all facets of information security, including the physical security of hardware and storage devices, administrative and access controls, and logical security of software applications. It also covers policies and procedures of the organization
   Robust data security plans, when correctly executed, safeguard an organization's information assets from cybercriminal activity, as well as from insider threats and human mistake, which remain the top causes of data breaches today. Data security entails implementing tools and technology that increase the organization's visibility into the location and usage of its essential data. Ideally, these solutions should be able to protect sensitive files via encryption, data masking, and redaction, as well as automate reporting to facilitate audits and compliance with regulatory standards. Every facet of today's corporate operations and competition is being radically altered by digital transformation.

The sheer volume of data that businesses generate, handle, and store is increasing, necessitating increased data governance. In addition, computing systems are more complex than they used to be, typically encompassing the public cloud, the enterprise data center, and a multitude of edge devices, such as Internet of Things (IoT) sensors, robots, and remote servers. This intricacy increases the attack surface, making it more difficult to monitor and secure.

Concurrently, consumer awareness of the significance of data privacy is growing. Multiple new privacy legislation has recently been enacted, notably Europe's General Data Protection Regulation (GDPR) and the California Consumer Protection Act, in response to rising public demand for data protection efforts (CCPA). These regulations join long-standing data security provisions such as the Health Insurance Portability and Accountability Act (HIPAA), which protects electronic health records, and the Sarbanes-Oxley Act (SOX), which protects shareholders from accounting errors and financial fraud in public companies. With maximum fines in the millions of dollars, there is a strong financial incentive for businesses to assure compliance. The value of data to businesses has never been higher than it is today. The loss of trade secrets or intellectual property (IP) can have a negative effect on future advances and profits. Consequently, trustworthiness is becoming increasingly crucial to consumers, with 75% stating that they will not purchase from businesses that they do not trust to protect their data (IBM, 2022). These contribute largely to the fashion industry investing huge in data security and having impact on their maximum profit.

3. Data Integration

Fashion companies may use variety of software systems and databases to manage different aspects of their business, such as inventory, sales, and customer relationship management. Integrating these systems and ensuring that the data is consistent across them can be a challenge.

The scope and significance of data integration have drastically shifted. Currently, we increase company capabilities by utilizing common SaaS services, while continuing to develop unique applications. With a robust ecosystem of partners ready to exploit an organization's information, the information about a company's services that is exposed to customers is now as vital as the services itself. Today, it is necessary to integrate SaaS, bespoke, and partner apps and the data stored inside them. A company differentiates itself by combining business talents in a novel way today. Numerous businesses, for instance, are evaluating data in-motion and at-rest, utilizing their results to develop business rules, and then applying these rules to react even more quickly to incoming data. This form of innovation typically aims to improve user experiences and corporate operations.

One of the greatest obstacles organizations confront is gaining access to and making sense of the data that represents their operating environment. Every day, organizations collect an increasing amount of data in various formats from a growing number of data sources. Organizations require a method for employees, users, and customers to extract value from data. This implies that enterprises must be able to collect important data from wherever it sits in order to support their reporting and business operations. However, essential data is frequently dispersed across applications, databases, and other data sources housed locally, in the cloud, on IoT devices, or given by third parties. Organizations no longer retain data in a single database; rather, they maintain traditional master and transactional data, in addition to new types of structured and unstructured

data, across numerous sources. For example, a company may have data in a flat file, or it may need to retrieve data from a web service.
The conventional method of data integration is referred to as the physical data integration method. This includes the physical transportation of data from its source system to a staging area, where cleansing, mapping, and transformation occur, before the data is physically moved to a target system, such as a data warehouse or data mart. The alternative method is data virtualization. Using a virtualization layer to connect to physical data stores is entailed by this method. Data virtualization, as opposed to physical data integration, includes the production of virtualized views of the underlying physical environment without the requirement for physical data transportation. Extract Transform and Load (ETL) is a typical data integration technique in which data is physically taken from many source systems, changed into a different format, and loaded into a centralized data storage.

4. Data Privacy
   Data privacy generally refers to an individual's ability to select when, how, and to what degree their personal information is shared or discussed with others. This can include a person's name, location, contact information, and online or offline behavior. Like how one may prefer to exclude specific individuals from a private chat, many Internet users wish to regulate or prohibit the collecting of sorts of personal information. Over the years, as Internet usage has expanded, so has the significance of data privacy. In order to deliver services, websites, applications, and social media platforms must frequently gather and keep users' personal information. Nonetheless, some programs and platforms may exceed user expectations for data gathering and utilization, leaving users with less privacy than they anticipated. Other apps and platforms may not adequately protect the data they collect, which may result in a data breach that breaches user privacy.
   In many nations, privacy is seen as a fundamental human right, and there are laws in place to preserve this right. Data privacy is particularly vital because for users to engage online, they must have faith that their personal information will be managed carefully. Organizations demonstrate to their clients and users that they can be trusted with their personal data by employing data protection practices.
   If personal information is not kept secret or if individuals lack the power to regulate how their information is used, it can be abused in several ways:
   Criminals can utilize personal data to defraud or harass people.
   Without user authorization, entities may sell personal data to advertisers or other third parties, which may result in users receiving unwanted marketing or advertising.
   When a person's activities are followed and monitored, this can impede their capacity to express themselves freely, particularly in countries with authoritarian governments. Any of these effects can be damaging to persons. These outcomes can irreversibly destroy a company's reputation and result in fines, sanctions, and other legal repercussions.
   Fashion companies must ensure that they are complying with laws and regulations related to data privacy, such as the General Data Protection Regulation (GDPR) in the European Union.

5. Data Analysis
   Today's businesses require every possible advantage and competitive edge. Due to constraints such as quickly moving markets, economic uncertainties, fluctuating political

landscapes, picky consumer attitudes, and even worldwide pandemics, firms today operate with narrower error margins. Companies that wish to survive and prosper can increase their chances of success by making intelligent decisions when addressing the question, "What is data analysis?" How do individuals and organizations make these decisions? They gather as much relevant, actionable data as possible and then use it to make more educated judgements. This method is sensible and applicable to both personal and professional life. No one makes significant decisions without first considering the stakes, pros and drawbacks, and potential outcomes. Likewise, no organization that desires success should make judgements based on inaccurate information. Organizations require data and information. Here is where data analysis and data analytics come into play. Today, data is viewed as the 'new oil' on the market, making data comprehension one of the fastest-growing sectors.

Although numerous individuals, organizations, and specialists have varied approaches to data analysis, most of them may be condensed into a single term. Data analysis is the process of cleansing, modifying, and processing raw data in order to obtain useful, usable information that assists businesses in making educated decisions. The technique aids in mitigating the inherent risks associated with decision-making by offering valuable insights and facts, which are frequently displayed as charts, graphics, tables, and graphs. Every time we decide in our daily lives by examining what has occurred in the past or what will occur if we make that decision, we demonstrate a straightforward use of data analysis. This entails assessing the past or the future and planning based on that analysis. Fashion companies may have access to large amounts of data, but they need to be able to analyze and interpret this data in order to make informed business decisions. This can be a challenge, especially for companies that do not have the necessary analytical skills or resources.

CONCLUSION

The fashion industry cannot because of problems associated with big data ignore usage of big data technologies. Big data technologies application in industry has larger benefit and helps boost the financial status of the industry. It also makes the industry produce the right quantity of commodity at the right time and not putting excess price on the final consumer. It also helps fight issues facing modern fashion industry by not making the industry produce excess clothing and also preaches recycling of fashion products thereby limiting the endangered animals where clothing materials are gotten from. The industry will have to invest heavily in tackling the five major issues associated with use of big data technologies and will enjoy the benefit associated with big data.

References

A., U. & E., C., 2017. *Industry 4.0: Managing The Digital Transformation.* London, Springer.

Barnes, L. & Lea-Greenwood., a. G., 2006. Fast fashioning the supply chain: Shaping the research agenda. *Journal of Fashion Marketing and Management,* 10(3), pp. 259-271.

Bruce, G. & Daly., a. L., 2006. Buyer behavior for fast fashion. *Journal of Fashion Marketing and Management,* 10(3), pp. 329-344.

Cachon, G. P. & Swinney, R., 2011. The value of fast fashion: Quick response, enhanced design, and strategic consumer behavior.. *Management science,* 57(4), pp. 778-795.

Christopher, M., Lowson, R. & & Peck, H., 2004. Creating agile supply chains in the fashion industry. *International Journal of Retail & Distribution Management,* 32(8), pp. 367-376.

Craig, S., 2022. *DATA QUALITY.* [Online]
Available at: https://www.techtarget.com/searchdatamanagement/definition/data-quality
[Accessed 05 01 2023].

Djelic, M.-L. & Ainamo., a. A., 1999. The coevolution of new organizational forms in the fashion industry: A historical and comparative study of France, Italy, and the United States. *Organizational Science,* 10(5), pp. 622-637.

Doyle, S., Moore, C. & fashion, a. L. M. 2. S. m. i. f. m., 2006. Supplier management in fast moving fashion retailing. *Journal of Fashion Marketing and Management,* 10(3), pp. 272-281.

Forza, C. & Vinelli., a. A., 1996. An analytical scheme for the change of the apparel design process towards quick response. *International Journal of Clothing,* 8(4), pp. 28-43.

Franks, J., 2000. Supply chain innovation. *Work study,* 49(4), pp. 152-156.

Guercini, S., 2001. Relation between branding and growth of the firm in new quick fashion formulas: Analysis of an Italian case. *Journal of Fashion Marketing and Management,* 5(1), pp. 69-79.

Hayes, S. & Jones., a. N., 2006. Fast fashion: A financial snapshot.. *Journal of Fashion Marketing and Management,* 10(3), pp. 282-300.

IBM, 2022. *Why is data security important?.* [Online]
Available at: https://www.ibm.com/uk-en/topics/data-security
[Accessed 05 01 2023].

K., R. P. & L., &. H., 2016. Reshoring: a strategic renewal of luxury clothing supply chains. *Operations Management Research,* 9(4), pp. 89-101.

Pine, B. & & Gilmore, J., 2007. *Authenticity, What Consumers really want.* Cambridge, Boston., Harvard Business School Press,.

Post., N., 2009. *Fast fashion.* [Online]
Available at: http://trendwatching.com/about/inmedia/articles/2009_fast_fashion.html
[Accessed 05 01 2023].

Seymour, S., 2010. *Functional Aesthetics: Visions in Fashionable Technology.* Vienna, Springer.

Sinha, P., 2001. The machanics of fashion. *Fashion marketing: Contemporary issues,* pp. 165-189.

Sydney., 2008. *Fast fashion is not a trend.* [Online]
Available at: http://www.sydneylovesfashion.com/2008/12/fast-fashion-is-trend.html
[Accessed 05 01 2023].

Taplin, I. 1. C. a. c. i. t. U. a. i. A. s. p., 1999. Continuity and change in the US apparel industry: A statistical profile. *Journal of Fashion Marketing and Management,* 3(4), pp. 360-368.

Wheelright, S. & Clark., a. K., 1992. *Revolutionizing product development: Quantum leaps in speed, efficiency, and quality,* New York: The Free Press.