PROJECT 4

REPORT

GROUP 5:

ASHAY AGLAWE

UDAY KUMBHAR

SHRINIVAS SANGLIKAR

INTRODUCTION

H&M Hennes & Mauritz AB (H&M) is a Swedish international apparel retailer based in Stockholm. Fast-fashion clothes for men, women, youth, and children is its main focus. H&M employs 126,000 full-time employees in 74 countries as of November 2019, with over 5,000 locations under several company names. H&M is the world's second-largest clothing store, after Spain's Inditex (parent company of Zara).

BACKGROUND KNOWLEDGE

Erling Persson founded the company in 1947, when he opened his first shop in Västers, Sweden. Hennes (Swedish for "hers") was a shop that only sold women's clothing. In 1964, a store opened in Norway.

Persson purchased the hunting apparel retailer Mauritz Widforss in Stockholm in 1968, which resulted in the addition of a menswear collection to the product line and the name change to Hennes & Mauritz.

In 1974, the company was listed on the Stockholm Stock Exchange. In 1976, the first store outside of Scandinavia opened in London.

The opening of its first U.S store on 31 March 2000 on Fifth Avenue in New York City marked the start of its expansion outside of Europe.

METHODOLOGY

Data Exploration:

We started by exploring the dataset. Since we had 3 data files namely articles, customers & transactions, we decided to investigate each of them individually to begin with.

Checking the shape and some statistics about the data formed the initial part of the process. Then we looked at the different features/attributes in the dataset and if any of them had missing values.

We then proceeded onto visualizations for better understanding of the data. The dataset was quite large with size of over 3.5GB, with the trasactions data having over 31M samples.

```
articles.shape
```

```
(105542, 25)
```

The articles data has 105542 samples with 25 attributes.

Summary of the articles data:

```
articles.describe().round(2)
```

	article_id	product_code	product_type_no	graphical_appearance_no	colour_group_code	perceived_colour_value_id	perceived_colour_master_id
count	1.055420e+05	105542.00	105542.00	105542.00	105542.00	105542.00	105542.00
mean	6.984246e+08	698424.56	234.86	1009515.08	32.23	3.21	7.81
std	1.284624e+08	128462.38	75.05	22413.59	28.09	1.56	5.38
min	1.087750e+08	108775.00	-1.00	-1.00	-1.00	-1.00	-1.00
25%	6.169925e+08	616992.50	252.00	1010008.00	9.00	2.00	4.00
50%	7.022130e+08	702213.00	259.00	1010016.00	14.00	4.00	5.00
75%	7.967030e+08	796703.00	272.00	1010016.00	52.00	4.00	11.00
max	9.594610e+08	959461.00	762.00	1010029.00	93.00	7.00	20.00

To check if the data has any missing values:

```
articles.isnull().sum()
```

```
article_id                      0
product_code                    0
prod_name                       0
product_type_no                 0
product_type_name               0
product_group_name              0
graphical_appearance_no         0
graphical_appearance_name       0
colour_group_code               0
colour_group_name               0
perceived_colour_value_id       0
perceived_colour_value_name     0
perceived_colour_master_id      0
perceived_colour_master_name    0
department_no                   0
department_name                 0
index_code                      0
index_name                      0
index_group_no                  0
index_group_name                0
section_no                      0
section_name                    0
garment_group_no                0
garment_group_name              0
detail_desc                   416
dtype: int64
```

Some of the product names and their counts:

```
articles['prod_name'].value_counts()
```

```
Dragonfly dress               98
Mike tee                      72
Wow printed tee 6.99          70
1pk Fun                       55
TP Paddington Sweater         54
                              ..
W MARCIE DRESS CNY             1
W NAPOLI SKIRT CNY             1
BEANIE JERSEY FLEECED LINED    1
H-string multicolour           1
Lounge dress                   1
Name: prod_name, Length: 45875, dtype: int64
```

```
customers.shape
```

```
(1371980, 7)
```

The customer data has 1371980 samples with 7 attributes.

Summary of the customer data:

```
customers.describe().round(2)
```

	FN	Active	age
count	476930.0	464404.0	1356119.00
mean	1.0	1.0	36.39
std	0.0	0.0	14.31
min	1.0	1.0	16.00
25%	1.0	1.0	24.00
50%	1.0	1.0	32.00
75%	1.0	1.0	49.00
max	1.0	1.0	99.00

```
customers.isnull().sum()
```

```
customer_id                  0
FN                      895050
Active                  907576
club_member_status        6062
fashion_news_frequency   16009
age                      15861
postal_code                  0
dtype: int64
```

The customer data had lot of missing values as can be seen from the above picture.

```
transactions.shape
```

```
(31788324, 5)
```

The transaction data was the largest with over 31M samples & 5 attributes.

```
transactions.describe().round(2)
```

	article_id	price	sales_channel_id
count	3.178832e+07	31788324.00	31788324.00
mean	6.962272e+08	0.03	1.70
std	1.334480e+08	0.02	0.46
min	1.087750e+08	0.00	1.00
25%	6.328030e+08	0.02	1.00
50%	7.145820e+08	0.03	2.00
75%	7.865240e+08	0.03	2.00
max	9.562170e+08	0.59	2.00

```
transactions.isnull().sum()
```

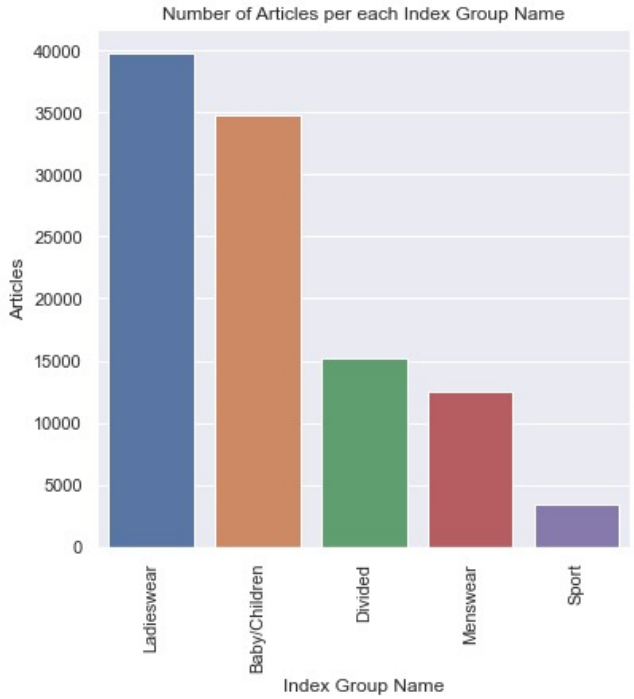
```
t_dat               0
customer_id         0
article_id          0
price               0
sales_channel_id    0
dtype: int64
```

This data did not contain any missing values.

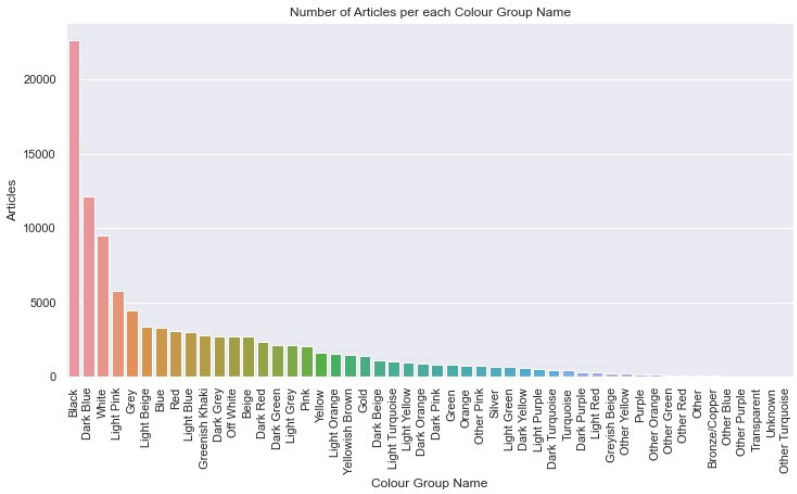
Visualizations:

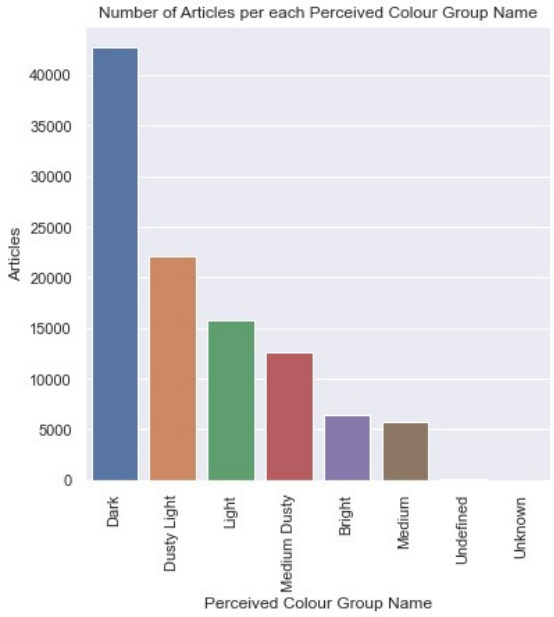
Some of the visualizations we looked at to understand the data:

Number of articles per Group Name:

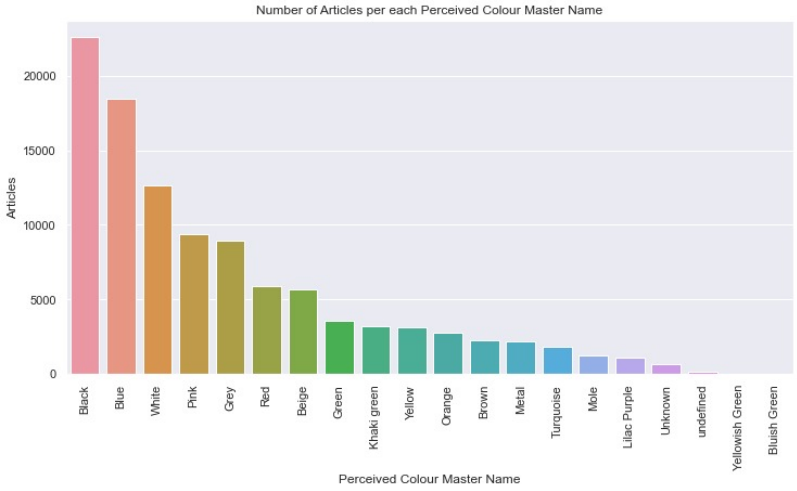


Number of articles per Group Color:

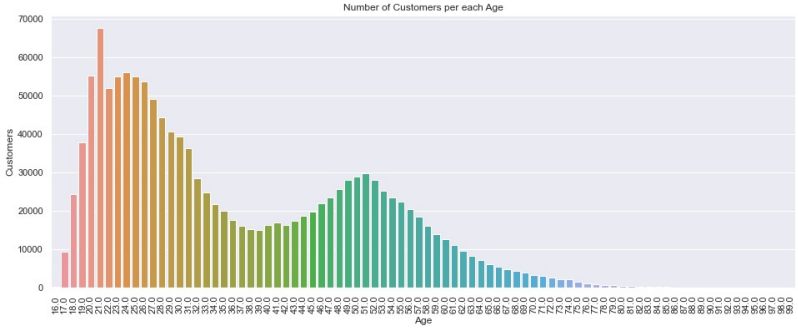




As we can see that Dark & Dusty light colors are popular among customers of H&M.

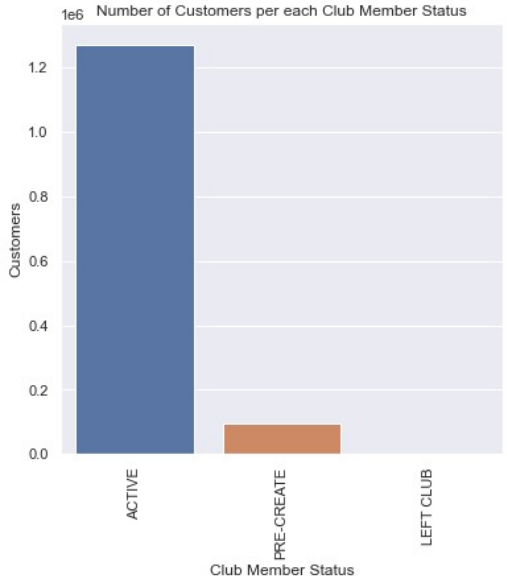


Distribution of Customers as per Age:



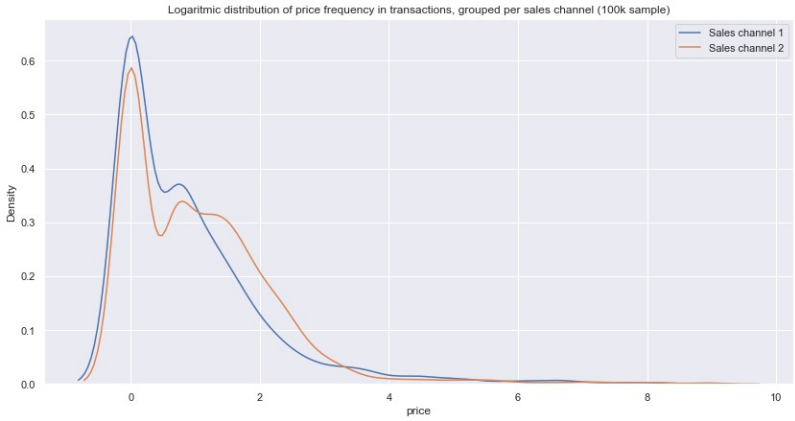
From the above age distribution we can see that people between 18-30 years of age shop in large numbers at H&M which is as expected.

Number of customers per Membership Status:



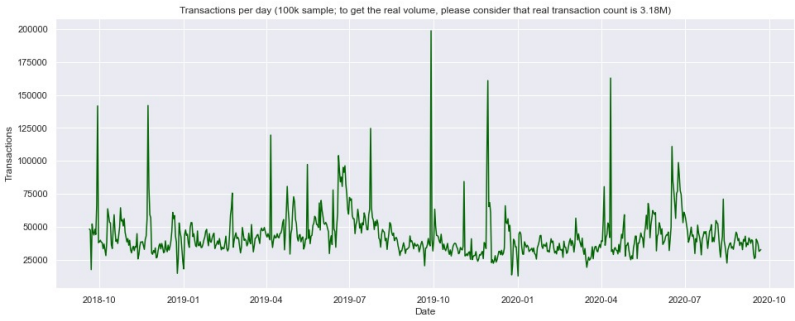
As the number of customers with active membership is high, this might suggest that people prefer maintaining membership status as it may have loyalty benefits such as reward points and exclusive offers while shopping at H&M.

Price Frequency per Sales Channel:

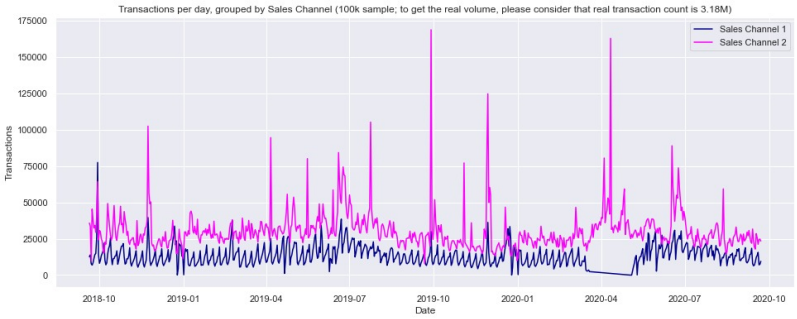


Here Sales channel 1 & 2 indicates offline mode and online mode.

Transactions per day (100K Samples):

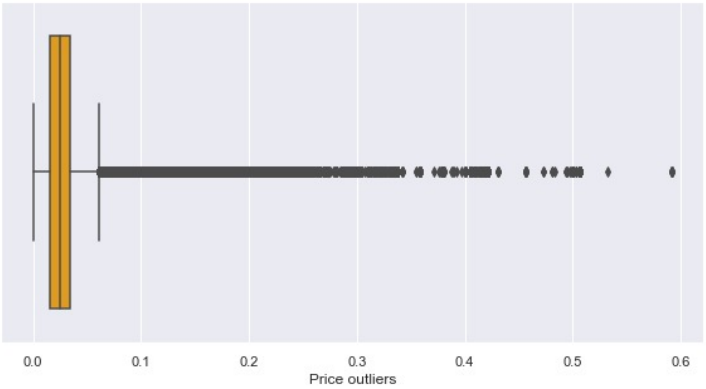


Transactions per day (100K Samples) by sales channel:

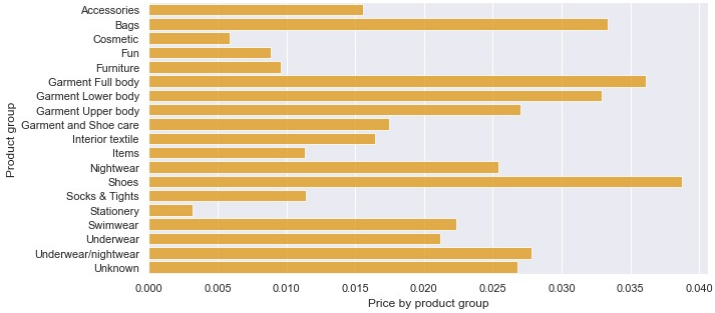


In this plot we see the transactions per day via the 2 channels. During March-May 2020 we see a flat line which is because of the stores being closed due to lockdowns caused by COVID-19.

Price Outliers:



Price by Product Group:



CUSTOMER SEGMENTATION

Customer segmentation is the process of dividing consumers into groups based on shared criteria. It is done for the following reasons:

- Target clients with the right marketing message and personalize their offers to a certain demographic. This not only helps businesses increase revenue, but it also helps them better maintain customer relationships and gain a deeper understanding of the customers.
- Identify ways to improve products or service opportunities.
- Test pricing options.
- Focus on the most profitable customers.

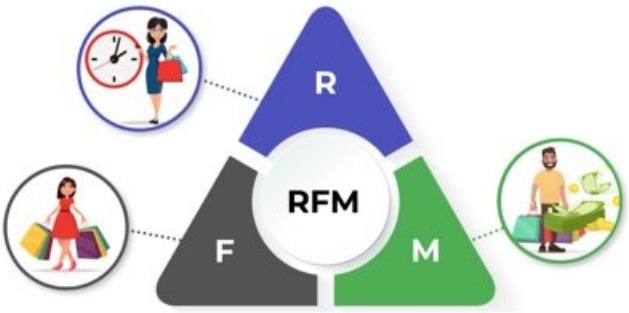
RFM Analysis

RFM stands for Recency, Frequency, Monetary Value.

- Recency: Days since last purchase/order of the client
- Frequency: Total number of purchases the customer made by the customer
- Monetary Value: Total money the customer spent per order

RFM is based on a basic concept:

1. Customers who have recently purchased from you are more likely to purchase from you again than customers who haven't purchased from you in a long time.
2. Customers who spend more money are more likely to purchase again than those who spend less money.



Segments:

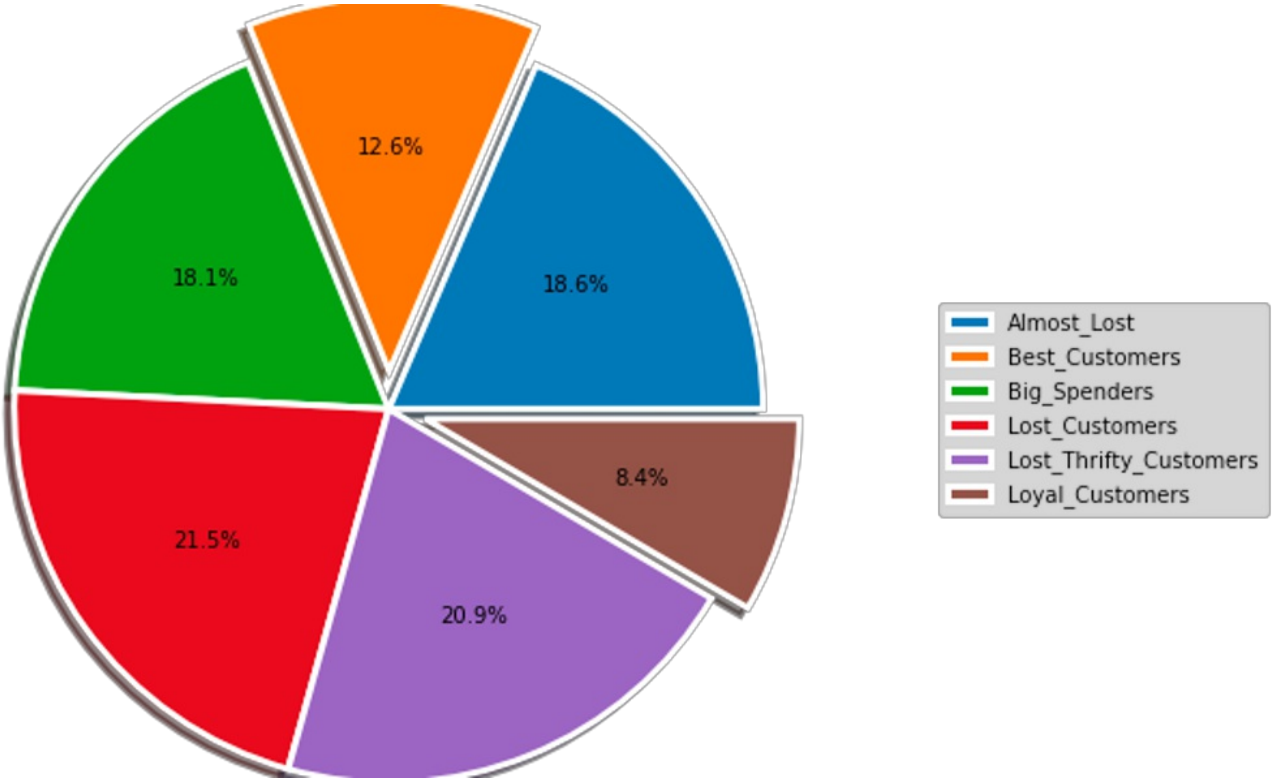
With the help of these RFM values we obtained RFM scores ranging from 1 to 4. (1 being lowest and 4 being highest).

After allocating those scores, we formed the segments such as, Best Customers, Loyal Customers, Big Spenders, etc.

The conditions were as follows:

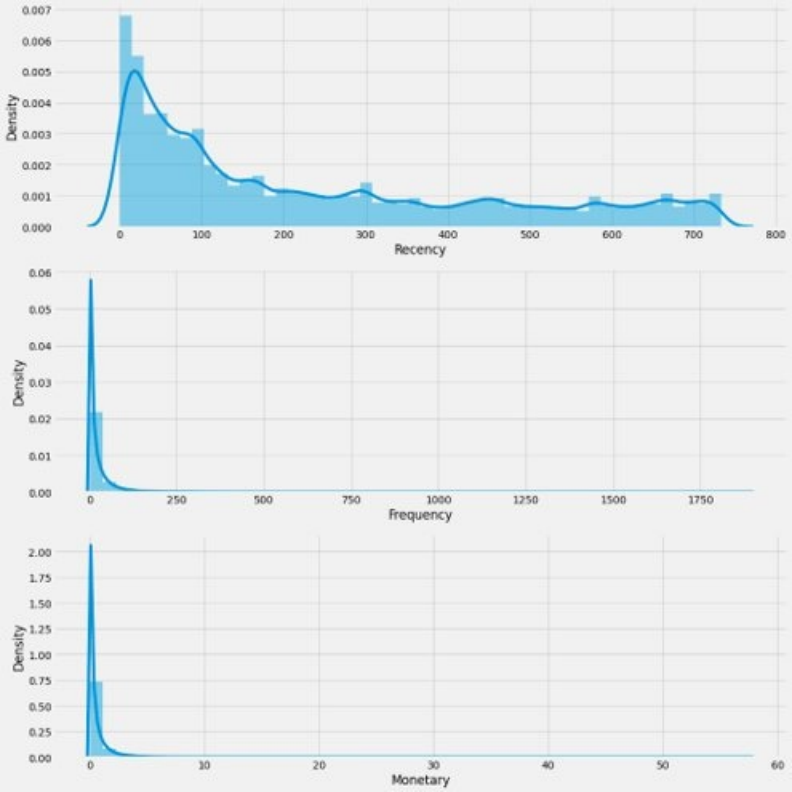
1. If RFM_Scores == '444', then "1-Best Customers"
2. If RFM_Scores == 'X4X', then "2-Loyal Customers"

3. If RFM_Scores == 'XX4', then "3-Big Spenders"
4. If RFM_Scores == '244', then "4-Almost Lost"
5. If RFM_Scores == '144', then "5-Lost Customers"
6. If RFM_Scores == '111', then "6-Lost Thrifty Customers"



The above pie chart illustrates that 20% of the loyal customers generate 80% of the revenue for H&M.

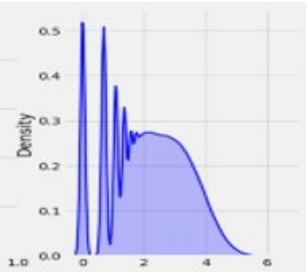
We looked at the data distribution of RFM and it was as follows:

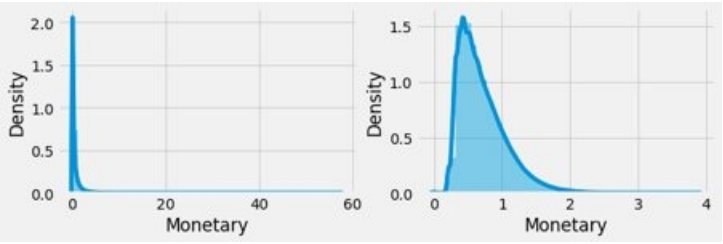


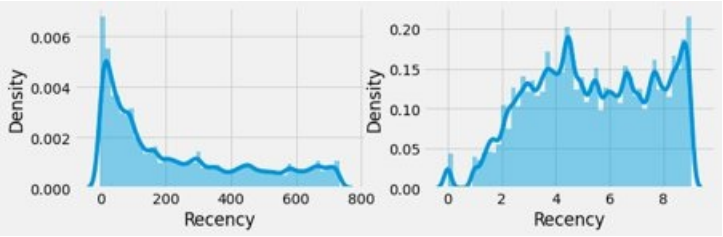
Data is mostly skewed for all three KPI's.

We used Box Cox transformation and Cube root transformation so that data will look like normal distribution.

<u>After transformation:</u>





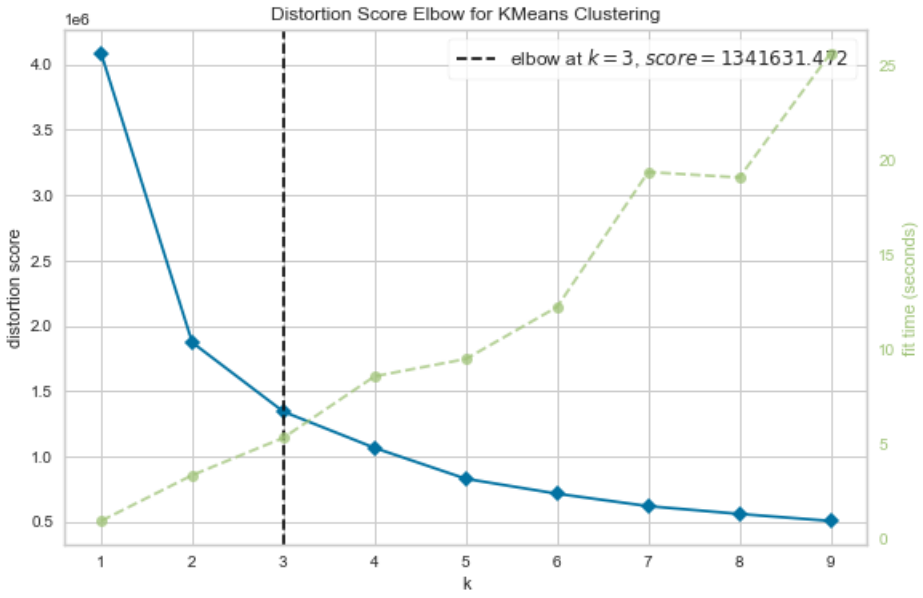


MODEL SELECTION

As our task involved clustering, we chose K-means as our model for training. K-means is a clustering method that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.

MODEL TUNING

Since the dataset is large, we used the elbow method to determine the k value required for training the k-means model.



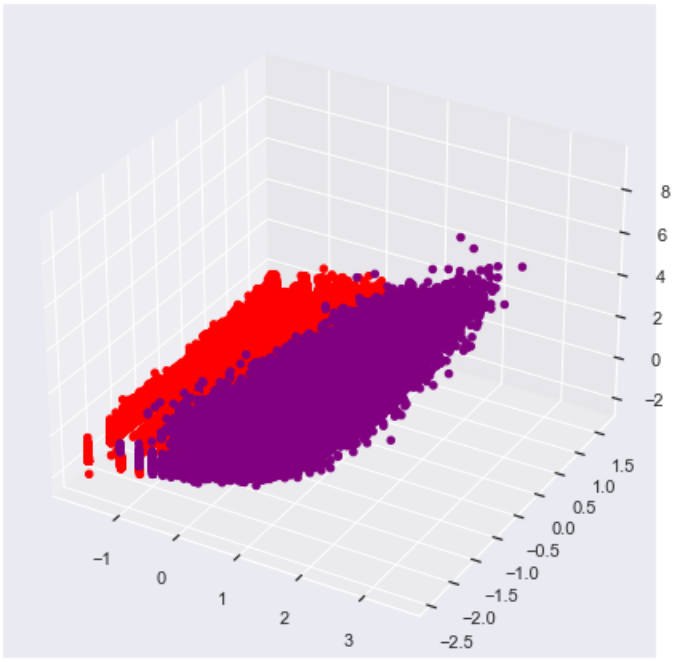
The bend is obtained at k = 2 but the elbow visualizer in Python selects k = 3.

MODEL EVALUATION

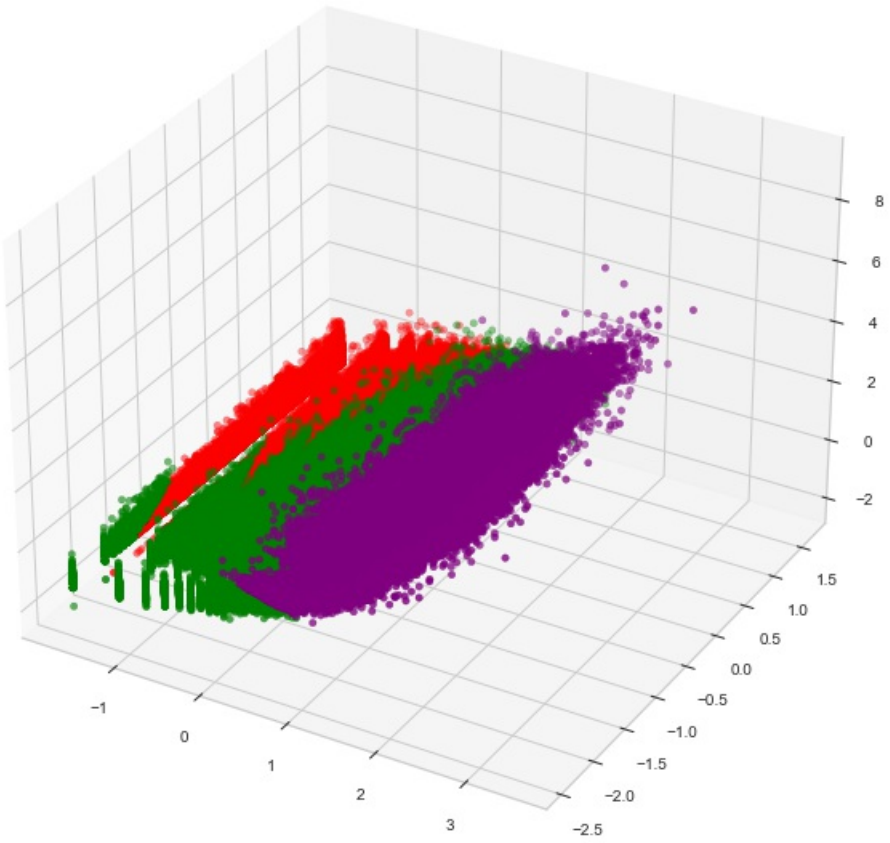
From the picture depicting elbow method, we can see that the bend is obtained at k = 2 but when we used the elbow visualizer, it chose k = 3. Hence we decided to evaluate the model for both values of k.

- We used Silhouette Coefficient to determine the overall quality of cluster.
- For K = 2, Silhouette score was around 0.45 and for K = 3, Silhouette score was around 0.4.
- These scores signify that the clusters are nearly overlapping.

Clusters for K = 2



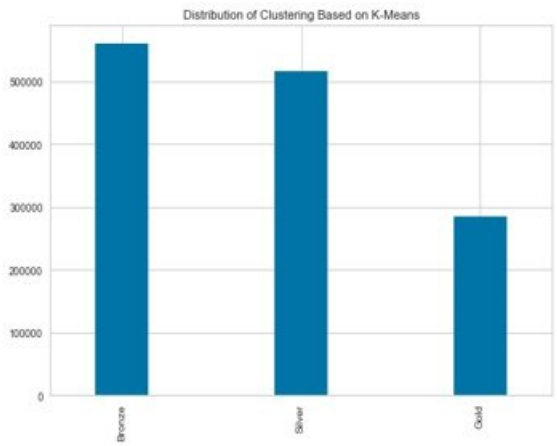
Clusters for K = 3



Analyzing Segments and Clusters

As the clusters were nearly overlapping, we created 3 packages (Gold, Silver and Bronze) and distributed the RFM segments accordingly.

RFM_Points_Segments Kmeans_Label	Best_Customers	Loyal_Customers	Big_Spenders	Almost_Lost	Lost_Customers	Lost_Thrifty_Customers
Gold	169540	89902	25482	5	0	0
Silver	1502	25210	220932	221997	46976	0
Bronze	0	0	0	30898	245425	284412



The best & the loyal customers or Gold package customers should be the most prioritized customers, most of

the resources should be spent on these them because they can be early adopters for new products and their feedback regarding those products could be obtained. Furthermore most expensive products should be marketed to the Gold package customers as they are more willing to spend their credit.

Customers in the bronze package who are lost or almost lost can be re-targeted with email sequence offering them with incentives. Also H&M should not waste too much time in reacquiring them.

CONCLUSIONS

We provided H&M with a better understanding of its customers, sales, and products. With this analysis, H&M can create new marketing strategies for their loyal customers and provide them with a superior shopping experience in the future.

For this analysis, it would be interesting to experiment with alternative clustering models such as DBSCAN and OPTICS to see if we get better results than K-means.

CHALLENGES

- In RFM analysis, coding for multiple combinations of RFM values becomes a daunting task.
- Optimizing the Silhoutte coefficient for such a large dataset with few dimensions was a tough task.
- Training the K-means model on ace cluster took about 9 hours to train due to large number of samples in the dataset.
- Since the transaction dataset was very large (>31M samples), merging it with the other two files was challenging as we faced issues with RAM allocation.
- The dataset that we had was largely inconsistent as the articles and transactions data did not have much missing values whereas the customer data had lots of missing values.
- The RFM distribution data was highly right skewed so we had to apply transformations to convert it into normal distribution.

STORY OF THE GROUP

We initiated the case study by downloading the dataset from Kaggle. As we had 3 files part of the dataset, Ashay worked on exploring the data of each individual file by checking the features, summary & missing values in the attributes(if any).

After this Shrinivas did the visualizations on the dataset to get a better understanding of the data through various plots and charts. Once we got a basic understanding of the same, Uday proceeded to find the RFM values which would help us in training the K-means clustering model.

Then we used the elbow method to determine k value for training the clustering model. On obtaining the k values, we found the Silhouette Coefficient and plotted the clusters to complete the final analysis.