Table of Contents

0. Instructions to Run the Code:

As part of the submission, five .py files were submitted. These files contain the code for task 1, and each of the clustering and classification techniques. Due to the difference in the Python environment used by the group members, running the code for some of the .py files requires different requisites:

a) **task1.py:** This file contains the code for task 1 and requires the nine .csv files provided in the assignment to be in the same folder to be able to run the code effectively without errors. Hence, they were included in the submission to facilitate the running of the code.

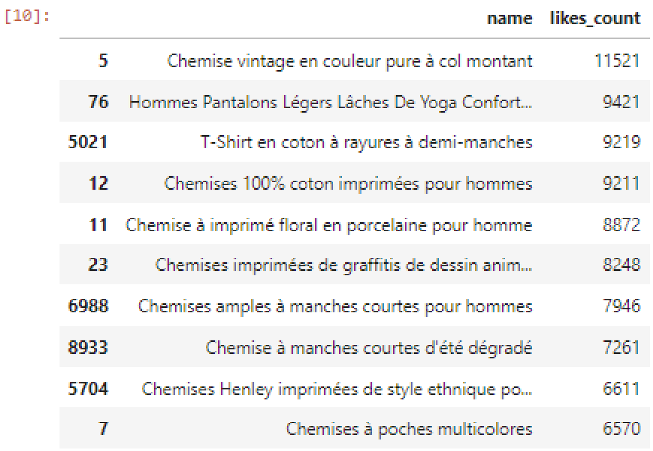
b) **clustering_kmeans.py, DBSCAN Clustering.py, and T3RFC.py:** These files contain the code for the clustering techniques mentioned in task 2 and one of the classification techniques mentioned in task 3. The Python environment used in the creation of those files was Google Colab (Google Colab 2024) and to avoid uploading the file every time the code is run, the final dataset file, Updated_CSV.csv, that was created by task 1 was uploaded and accessed through Google Drive (Google Drive 2024). That file can be accessed here.

c) **knn.py:** This file contains the code for one of the classification techniques mentioned in task 3 and it requires the Updated_CSV.csv file produced by task 1 to be in the same folder to avoid any errors when running the code. Hence, we have also included the Updated_CSV.csv file in our submission.

1. Problem Analysis:

We are given snapshots of product records stored by an e-commerce website, NewChic. The snapshots are categorised; they relate to information that falls into one of nine overall categories. In order to conduct a thorough analysis of these records we decided to utilise all nine provided categories in our analysis. We were tasked with finding the top 10 products from all the selected categories, as well as the best category among the selected categories. Hence, we first sorted the records of each categorical .csv file in the descending order of the value of a particular variable, likes_count. By utilising this technique we were able to find out the top 10 products from each category and hence the top 10 products overall. The following figures are the records of the top 10 products from each category (refer to Figures 1 - 10):

1. Men Category:

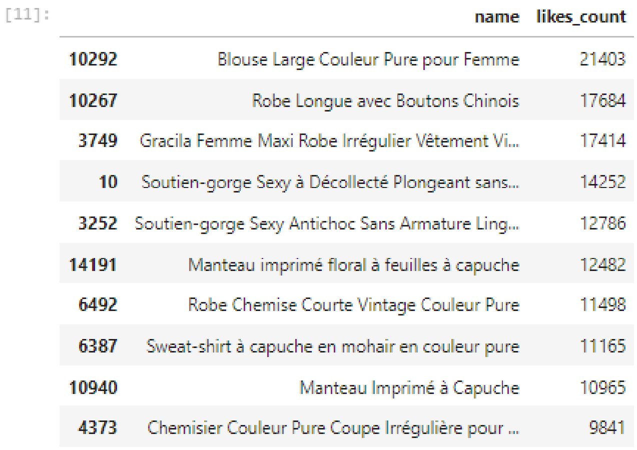


[10]:

	name	likes_count
5	Chemise vintage en couleur pure à col montant	11521
76	Hommes Pantalons Légers Lâches De Yoga Confort...	9421
5021	T-Shirt en coton à rayures à demi-manches	9219
12	Chemises 100% coton imprimées pour hommes	9211
11	Chemise à imprimé floral en porcelaine pour homme	8872
23	Chemises imprimées de graffitis de dessin anim...	8248
6988	Chemises amples à manches courtes pour hommes	7946
8933	Chemise à manches courtes d'été dégradé	7261
5704	Chemises Henley imprimées de style ethnique po...	6611
7	Chemises à poches multicolores	6570

Figure 1: Top 10 Products in the Men Category Based on the Highest “likes_count”

2. Women Category:

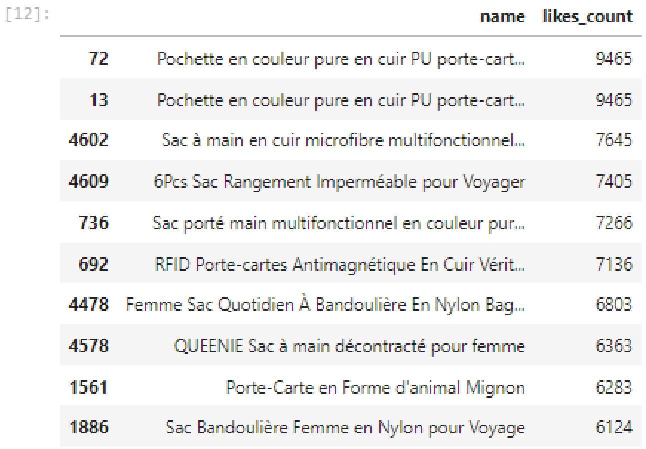


[11]:

	name	likes_count
10292	Blouse Large Couleur Pure pour Femme	21403
10267	Robe Longue avec Boutons Chinois	17684
3749	Gracila Femme Maxi Robe Irrégulier Vêtement Vi...	17414
10	Soutien-gorge Sexy à Décollecté Plongeant sans...	14252
3252	Soutien-gorge Sexy Antichoc Sans Armature Ling...	12786
14191	Manteau imprimé floral à feuilles à capuche	12482
6492	Robe Chemise Courte Vintage Couleur Pure	11498
6387	Sweat-shirt à capuche en mohair en couleur pure	11165
10940	Manteau Imprimé à Capuche	10965
4373	Chemisier Couleur Pure Coupe Irrégulière pour ...	9841

Figure 2: Top 10 Products in the Women Category Based on the Highest “likes_count”

3. Bags Category:

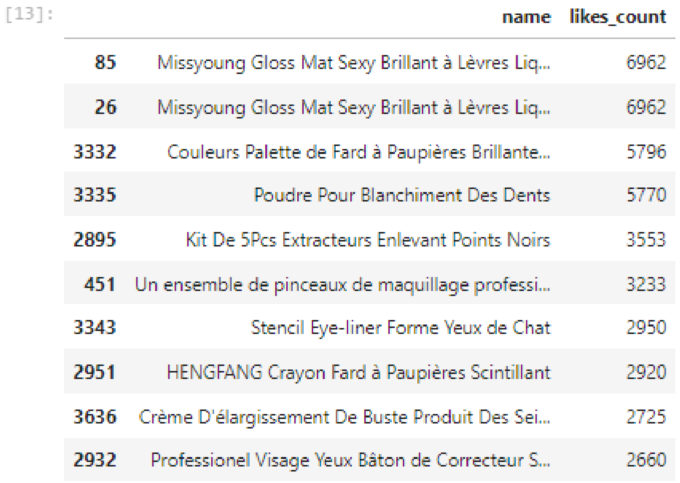


[12]:

	name	likes_count
72	Pochette en couleur pure en cuir PU porte-cart...	9465
13	Pochette en couleur pure en cuir PU porte-cart...	9465
4602	Sac à main en cuir microfibre multifonctionnel...	7645
4609	6Pcs Sac Rangement Imperméable pour Voyager	7405
736	Sac porté main multifonctionnel en couleur pur...	7266
692	RFID Porte-cartes Antimagnétique En Cuir Vérit...	7136
4478	Femme Sac Quotidien À Bandoulière En Nylon Bag...	6803
4578	QUEENIE Sac à main décontracté pour femme	6363
1561	Porte-Carte en Forme d'animal Mignon	6283
1886	Sac Bandoulière Femme en Nylon pour Voyage	6124

Figure 3: Top 10 Products in the Bags Category Based on the Highest "likes_count"

4. Beauty Category:

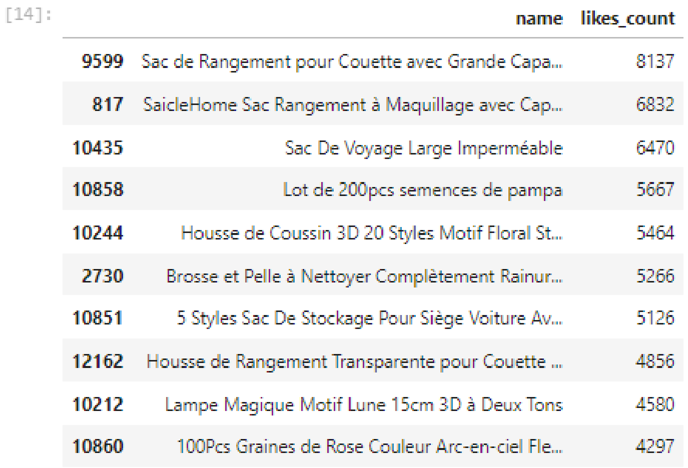


[13]:

	name	likes_count
85	Missyoung Gloss Mat Sexy Brillant à Lèvres Liq...	6962
26	Missyoung Gloss Mat Sexy Brillant à Lèvres Liq...	6962
3332	Couleurs Palette de Fard à Paupières Brillante...	5796
3335	Poudre Pour Blanchiment Des Dents	5770
2895	Kit De 5Pcs Extracteurs Enlevant Points Noirs	3553
451	Un ensemble de pinceaux de maquillage professi...	3233
3343	Stencil Eye-liner Forme Yeux de Chat	2950
2951	HENGFANG Crayon Fard à Paupières Scintillant	2920
3636	Crème D'élargissement De Buste Produit Des Sei...	2725
2932	Professionel Visage Yeux Bâton de Correcteur S...	2660

Figure 4: Top 10 Products in the Beauty Category Based on the Highest "likes_count"

5. House Category:

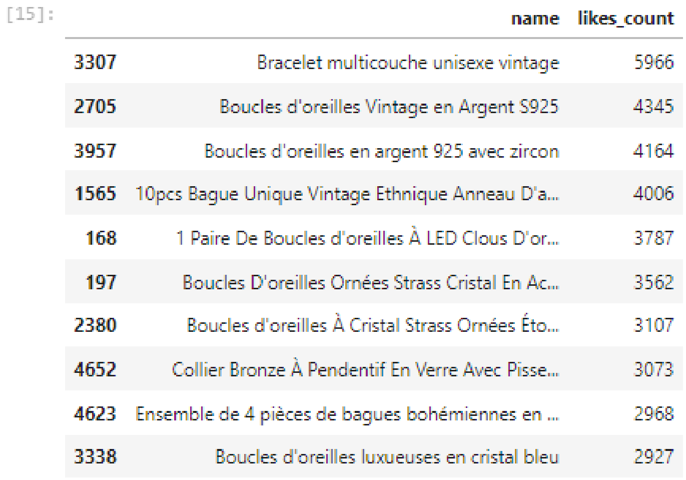


[14]:

	name	likes_count
9599	Sac de Rangement pour Couette avec Grande Capa...	8137
817	SaicleHome Sac Rangement à Maquillage avec Cap...	6832
10435	Sac De Voyage Large Imperméable	6470
10858	Lot de 200pcs semences de pampa	5667
10244	Housse de Coussin 3D 20 Styles Motif Floral St...	5464
2730	Brosse et Pelle à Nettoyer Complètement Rainur...	5266
10851	5 Styles Sac De Stockage Pour Siège Voiture Av...	5126
12162	Housse de Rangement Transparente pour Couette ...	4856
10212	Lampe Magique Motif Lune 15cm 3D à Deux Tons	4580
10860	100Pcs Graines de Rose Couleur Arc-en-ciel Fle...	4297

Figure 5: Top 10 Products in the House Category Based on the Highest "likes_count"

6. Jewellery Category:

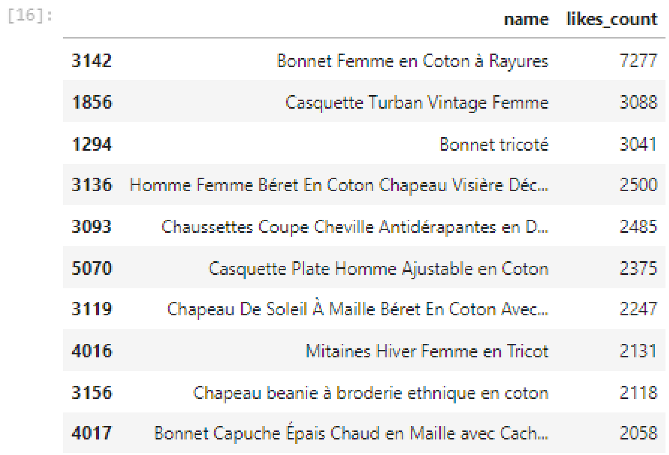


[15]:

	name	likes_count
3307	Bracelet multicouche unisexe vintage	5966
2705	Boucles d'oreilles Vintage en Argent S925	4345
3957	Boucles d'oreilles en argent 925 avec zircon	4164
1565	10pcs Bague Unique Vintage Ethnique Anneau D'a...	4006
168	1 Paire De Boucles d'oreilles À LED Clous D'or...	3787
197	Boucles D'oreilles Ornées Strass Cristal En Ac...	3562
2380	Boucles d'oreilles À Cristal Strass Ornées Éto...	3107
4652	Collier Bronze À Pendentif En Verre Avec Pisse...	3073
4623	Ensemble de 4 pièces de bagues bohémiennes en ...	2968
3338	Boucles d'oreilles luxueuses en cristal bleu	2927

Figure 6: Top 10 Products in the Jewellery Category Based on the Highest “likes_count”

7. Accessories Category:

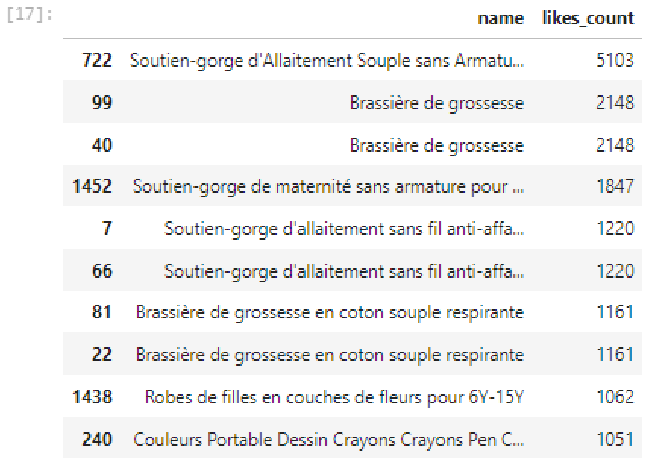


[16]:

	name	likes_count
3142	Bonnet Femme en Coton à Rayures	7277
1856	Casquette Turban Vintage Femme	3088
1294	Bonnet tricoté	3041
3136	Homme Femme Béret En Coton Chapeau Visière Déc...	2500
3093	Chaussettes Coupe Cheville Antidérapantes en D...	2485
5070	Casquette Plate Homme Ajustable en Coton	2375
3119	Chapeau De Soleil À Maille Béret En Coton Avec...	2247
4016	Mitaines Hiver Femme en Tricot	2131
3156	Chapeau beanie à broderie ethnique en coton	2118
4017	Bonnet Capuche Épais Chaud en Maille avec Cach...	2058

Figure 7: Top 10 Products in the Accessories Category Based on the Highest “likes_count”

8. Kids Category:

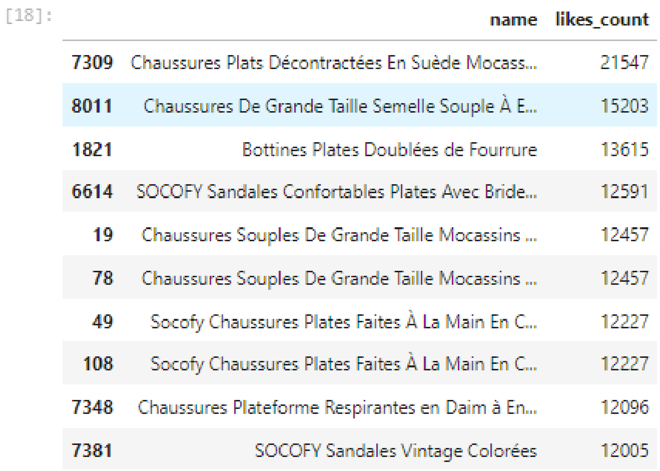


[17]:

	name	likes_count
722	Soutien-gorge d'Allaitement Souple sans Armatu...	5103
99	Brassière de grossesse	2148
40	Brassière de grossesse	2148
1452	Soutien-gorge de maternité sans armature pour ...	1847
7	Soutien-gorge d'allaitement sans fil anti-affa...	1220
66	Soutien-gorge d'allaitement sans fil anti-affa...	1220
81	Brassière de grossesse en coton souple respirante	1161
22	Brassière de grossesse en coton souple respirante	1161
1438	Robes de filles en couches de fleurs pour 6Y-15Y	1062
240	Couleurs Portable Dessin Crayons Crayons Pen C...	1051

Figure 8: Top 10 Products in the Kids Category Based on the Highest “likes_count”

9. Shoes Category:

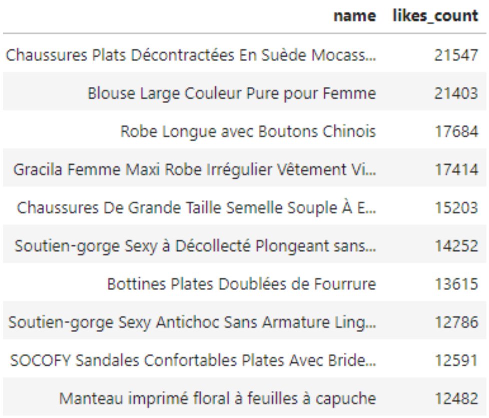


[18]:

	name	likes_count
7309	Chaussures Plats Décontractées En Suède Mocass...	21547
8011	Chaussures De Grande Taille Semelle Souple À E...	15203
1821	Bottines Plates Doublées de Fourrure	13615
6614	SOCOFY Sandales Confortables Plates Avec Bride...	12591
19	Chaussures Souples De Grande Taille Mocassins ...	12457
78	Chaussures Souples De Grande Taille Mocassins ...	12457
49	Socofy Chaussures Plates Faites À La Main En C...	12227
108	Socofy Chaussures Plates Faites À La Main En C...	12227
7348	Chaussures Plateforme Respirantes en Daim à En...	12096
7381	SOCOFY Sandales Vintage Colorées	12005

Figure 9: Top 10 Products in the Shoes Category Based on the Highest “likes_count”

10. Top 10 Products Overall:



name	likes_count
Chaussures Plats Décontractées En Suède Mocass...	21547
Blouse Large Couleur Pure pour Femme	21403
Robe Longue avec Boutons Chinois	17684
Gracila Femme Maxi Robe Irrégulier Vêtement Vi...	17414
Chaussures De Grande Taille Semelle Souple À E...	15203
Soutien-gorge Sexy à Décollecté Plongeant sans...	14252
Bottines Plates Doublées de Fourrure	13615
Soutien-gorge Sexy Antichoc Sans Armature Ling...	12786
SOCOFY Sandales Confortables Plates Avec Bride...	12591
Manteau imprimé floral à feuilles à capuche	12482

Figure 10: Top 10 Products Overall Based on the Highest “likes_count”

2. Data Preprocessing:

After identifying the top 10 products from each category and the top 10 products overall, we proceeded to perform preprocessing, which is a necessary step to be able to perform clustering & classification on this data and also for identifying the best category.

Firstly a combined dataset was created by concatenating all the provided .csv files. The next few steps ensure that the data is processed and ready for modelling. They are explained briefly as follows:

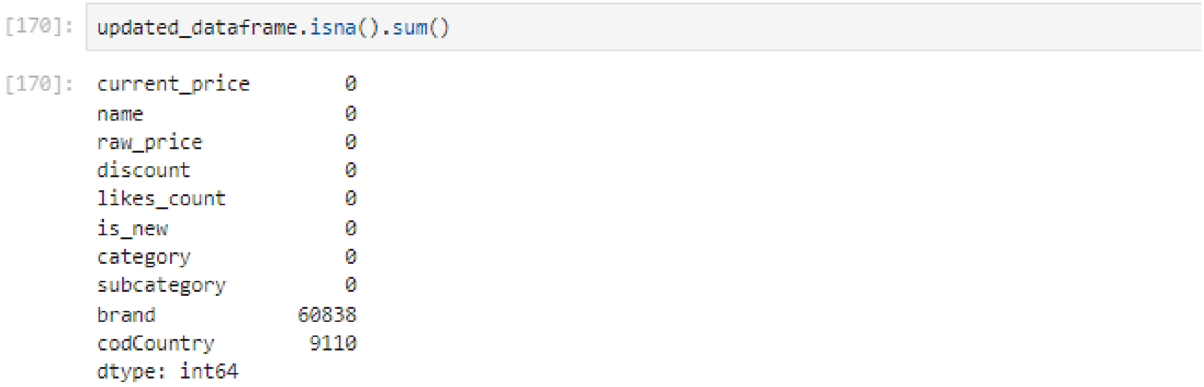
2.1 Preprocessing steps

1. **Selecting relevant features “columns” -** several features were discarded from the combined dataset as the information they contained was not relevant to the analysis. This includes all the url, color, thumbnail, and image columns, as well as the currency, id, and model columns as the data is more relevant to the company but will not be beneficial in finding the top 10 products and best product category. The removal of these columns will aid in the generation of optimised models for the analysed phenomena (Gupta 2024). The list below shows the columns that have been removed in this step:

    a) brand_url
    b) codCountry
    c) variation_0_color
    d) variation_1_color
    e) variation_0_thumbnail
    f) variation_0_image
    g) variation_1_thumbnail
    h) variation_1_image
    i) image_url
    j) url
    k) currency
    l) id
    m) model

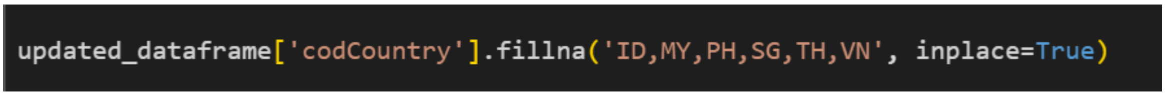
2. **Missing values -** an important step in machine learning is to handle any missing values as this can cause the implemented machine learning algorithms to provide skewed results or lose accuracy if the missing data is not handled appropriately (Olmez 2024). Hence, it is important to either carefully complete the missing data or eliminate them (Olmez 2024). Upon investigating our nine datasets, two columns were found to have missing data, ‘brand’ and ‘codCountry’ as shown in Figure 11. The column ‘brand’ was

dropped because it was revealed that it had over 60k missing values which is a huge portion of the data. 'codCountry', on the other hand, had 9k missing records, which were filled using the fillna() method present in Pandas library with the most used record "ID,MY,PH,SG,TH,VN" as shown in Figure 12.



```
[170]: updated_dataframe.isna().sum()

[170]: current_price        0
       name                 0
       raw_price            0
       discount             0
       likes_count          0
       is_new               0
       category             0
       subcategory          0
       brand            60838
       codCountry        9110
       dtype: int64
```

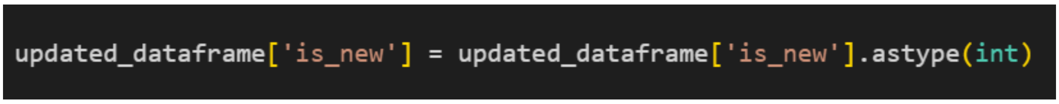
Figure 11: Missing Values in Each of the Chosen Columns



```
updated_dataframe['codCountry'].fillna('ID,MY,PH,SG,TH,VN', inplace=True)
```

Figure 12: Code for Replacing Missing Values in 'codCountry'

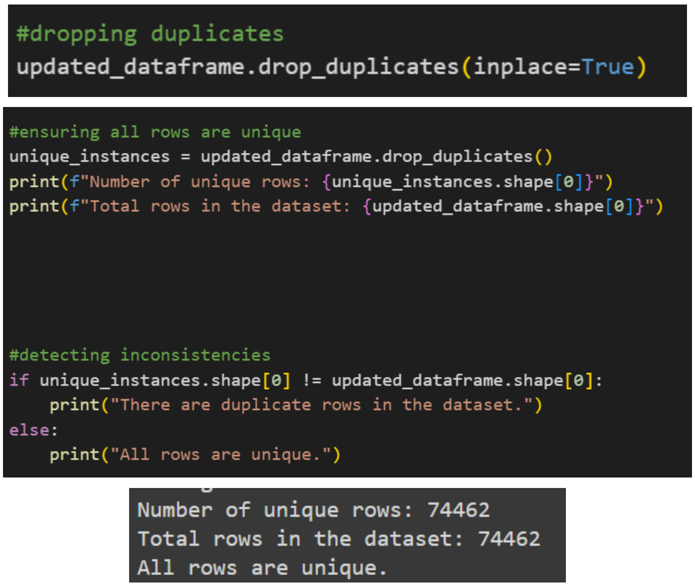
3. **Data Types:** feature "is_new" initially had values of Boolean data type, this was modified such that the feature would have a value of 0 for every 'False' record and a value of 1 for every 'True' record. This allows for easier interpretability (Gupta 2024). The code for this is shown in the figure below (refer to Figure 13):



```
updated_dataframe['is_new'] = updated_dataframe['is_new'].astype(int)
```

Figure 13: Code for the Conversion of 'is_new"s Boolean Values into Numerical Values

4. **Dropping duplicates & checking inconsistencies in data:** duplicates in a dataset can result in slow processing (Durgapal 2023). Hence, removing these is crucial as part of the data pre-processing (Durgapal 2023). In this dataset, the following code was used to remove any duplicates and then the data was checked against the total number of rows in the dataset to ensure that each record (row) was unique (since each row represents a different product) as shown in the code below (refer to Figure 14):



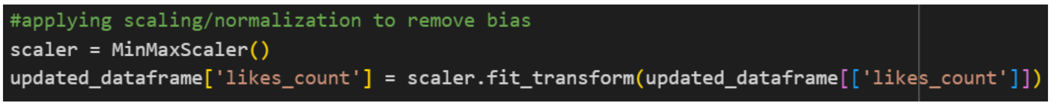
```
#dropping duplicates
updated_dataframe.drop_duplicates(inplace=True)
```

```
#ensuring all rows are unique
unique_instances = updated_dataframe.drop_duplicates()
print(f"Number of unique rows: {unique_instances.shape[0]}")
print(f"Total rows in the dataset: {updated_dataframe.shape[0]}")




#detecting inconsistencies
if unique_instances.shape[0] != updated_dataframe.shape[0]:
    print("There are duplicate rows in the dataset.")
else:
    print("All rows are unique.")
```

```
Number of unique rows: 74462
Total rows in the dataset: 74462
All rows are unique.
```

Figure 14: Code for Removing Duplicates and Ensuring that All Rows are Unique

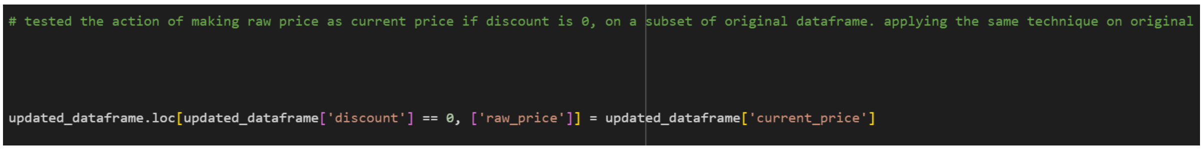
5. **Scaling:** it was observed that the likes_count feature has high range values, with the max being 21,547. Hence, it was decided to perform a scaling operation on this feature. Scaling is performed in order to remove any potential bias or skewness (Shivanipickl 2023). For this dataset and feature, we imported and used MinMaxScaler from the scit-kit learn preprocessing module. The code for this is shown below in Figure 15:



```
#applying scaling/normalization to remove bias
scaler = MinMaxScaler()
updated_dataframe['likes_count'] = scaler.fit_transform(updated_dataframe[['likes_count']])
```

Figure 15: Code for Scaling likes_count

6. **Ensuring the correctness of data:** 'raw_price' is the product's price before any discounts were made while 'current_price' is the price of the product after a discount. Hence, to ensure that there are no misleading data that can affect the results of the models negatively (Acharya 2024), it is important to ensure that each product that has 0 discount has the same exact 'raw_price' and 'current_price' as shown in Figure 16:

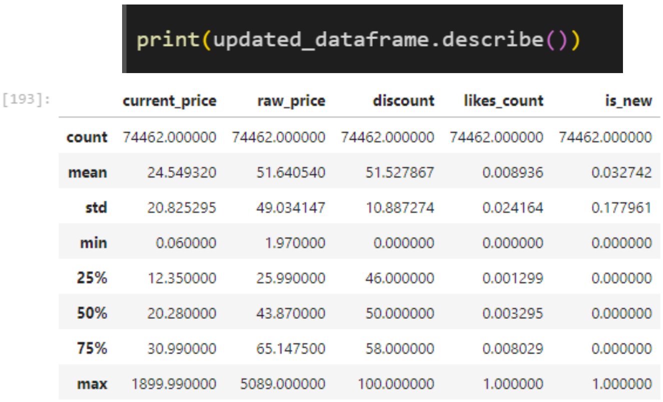


```
# tested the action of making raw price as current price if discount is 0, on a subset of original dataframe. applying the same technique on original


updated_dataframe.loc[updated_dataframe['discount'] == 0, ['raw_price']] = updated_dataframe['current_price']
```

Figure 16: Code for Ensuring the Correctness of the Data

7. **Statistical summary:** After applying the aforementioned preprocessing steps, a statistical summary was created by using the describe() method in the

Pandas library on the numerical data. This provides the count, the mean, the standard deviation, the minimum value, the 25th percentile, the median, the 75th percentile, and the maximum value allowing us to have an insight into the dataset (Venkataramanan 2021). This helps ensure that there are no missing values and that the scaling of likes_count was successful as well as the conversion of the 'is_new' values from boolean to numerical. The code and result of this is shown in Figure 17:



```
print(updated_dataframe.describe())
```

[193]:

	current_price	raw_price	discount	likes_count	is_new
count	74462.000000	74462.000000	74462.000000	74462.000000	74462.000000
mean	24.549320	51.640540	51.527867	0.008936	0.032742
std	20.825295	49.034147	10.887274	0.024164	0.177961
min	0.060000	1.970000	0.000000	0.000000	0.000000
25%	12.350000	25.990000	46.000000	0.001299	0.000000
50%	20.280000	43.870000	50.000000	0.003295	0.000000
75%	30.990000	65.147500	58.000000	0.008029	0.000000
max	1899.990000	5089.000000	100.000000	1.000000	1.000000

Figure 17: Code and Result of the Statistical Summary

2.2 Visuals

In order to draw some insights from the data before beginning modelling, we created a few visuals and charts. This helped us in getting familiar with the data and the relationships between features and categories. Most of these visuals, apart from the correlation matrix, were created using Microsoft Power BI (Microsoft 2024).

Figure 18 illustrates the count of products by category, where each category is represented by a box. The larger the box the higher the count. The 'women' category has the highest count with 14k products (refer to Figure 18). This is followed by the 'house' category which has 11k products (refer to Figure 18). The 'shoes' and 'men' categories both have a total of 9k products each (refer to Figure 18). The 'accessories', 'bags', and 'jewellery' categories are in the fourth, fifth, and sixth places respectively with 6k, 5k, and 4k products respectively (refer to Figure 18). Lastly, the 'beauty' and 'kids' categories each have 3k products (refer to Figure 18). With this data, we can deduce that 30% of the products sold by NewChic are categorized as "Women" while only about 13% are categorized as either 'kids' or 'beauty'.

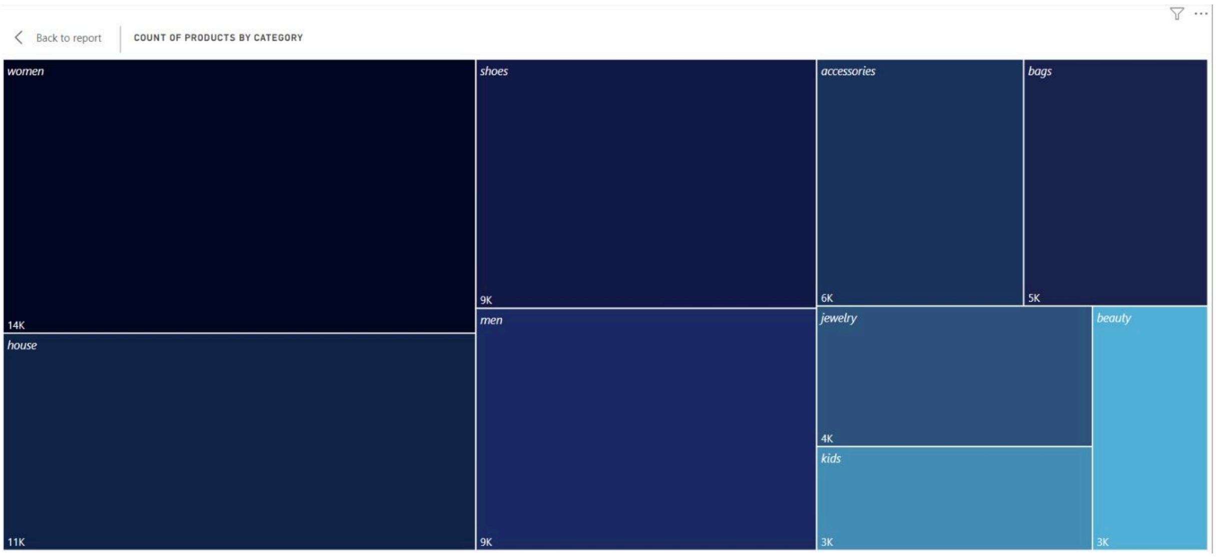


Figure 18: Count of Products by Category

The following figure indicates the average 'current_price' for each category (refer to Figure 19). 'shoes' is clearly the category with the highest average 'current_price', while 'accessories' has the lowest average 'current_price' (refer to Figure 19). This indicates that 'shoes' may have fewer discounts or are generally more expensive compared to the rest of the categories. Some categories, such as 'beauty' and 'women' as well as 'kids' and 'house', have similar average current_price (refer to Figure 19), which is an interesting observation as each of those pairs has a category with one of the highest products as well as one with the lowest products as was shown in Figure 18. This indicates that the number of products a category has does not influence its average 'current_price'.

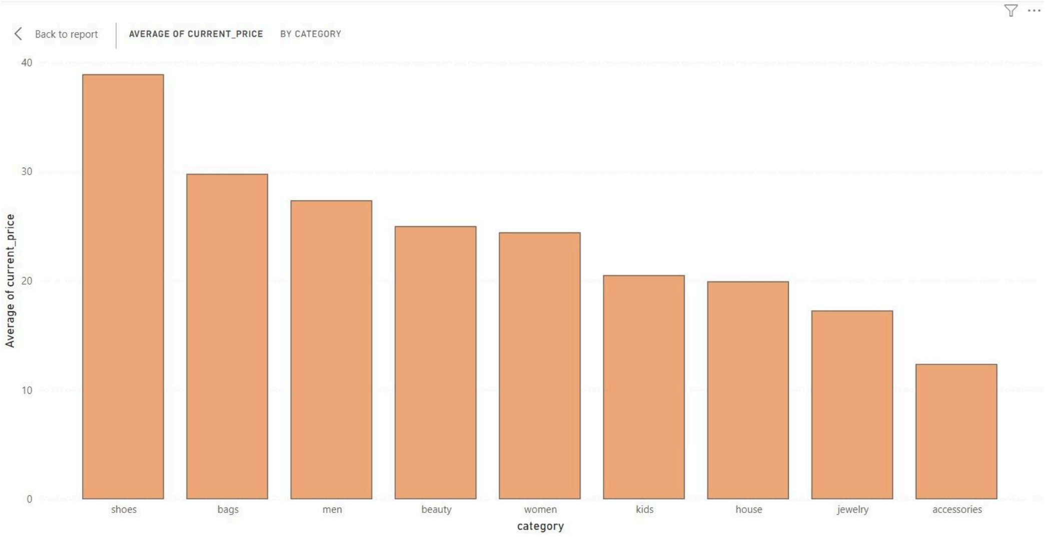


Figure 19: Average 'current_price' vs Category

The total likes_count for each category was computed in order to visualize the most popular category. The results of this were almost similar to that of Figure 18 except that ‘shoes’, which has the highest average ‘current_price’ as shown in Figure 19, is the second highest liked category instead of being in the third place as it was in Figure 18. ‘accessories’ was also knocked down to seventh place in Figure 20 compared to being the fifth in Figure 18 even though it has the lowest average ‘current_price’ as illustrated in Figure 19. The bar chart in Figure 20 indicates that the total likes_count for each category is more widespread compared to the results shown in Figures 18 and 19 with the ‘women’ category having the highest total likes_count which coupled with the data from Figure 19 indicates that NewChic is more focused on ‘women’ products.

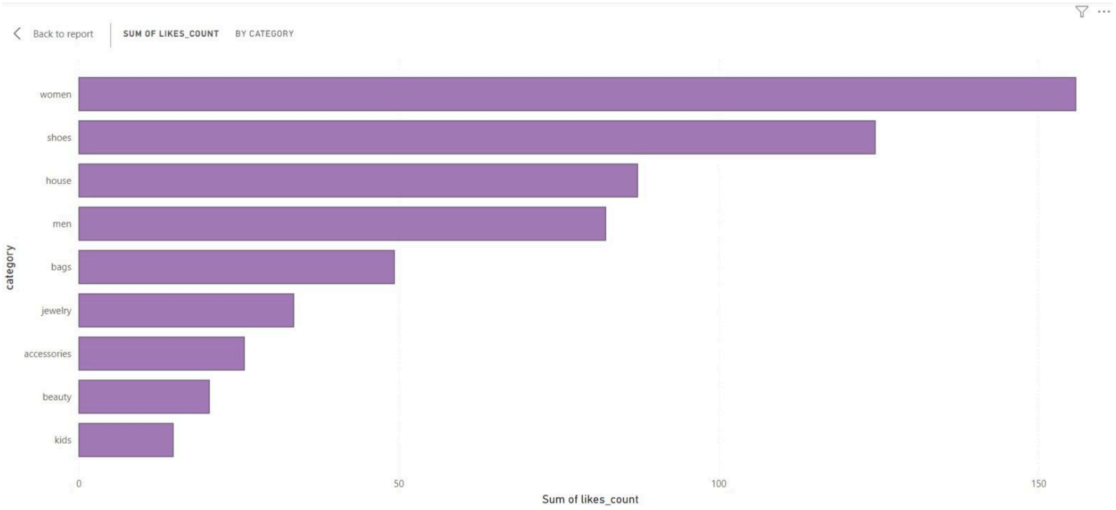


Figure 20: Total likes_count by Category

Figure 21 illustrates the average ‘discount’ of each category. The ‘kids’ category has the highest average ‘discount’, while the ‘men’ category has the lowest average ‘discount’. Several pairs of categories have similar average discounts which indicates that NewChic implements similar strategies for them. This, however, does not apply to the kids, women, house, and men categories that have different average discounts indicating that different strategies are applied to them. An interesting observation is that although the kids category has the highest average discounts it still has the least total likes_count as was illustrated by Figures 20 and 21. This might indicate that the company has either newly introduced the kids category or that it is trying to increase its likes_count and hence its sales in this sector by providing the customers with more discounts as an incentive to attract new budget-conscious customers (Active Campaign 2024).

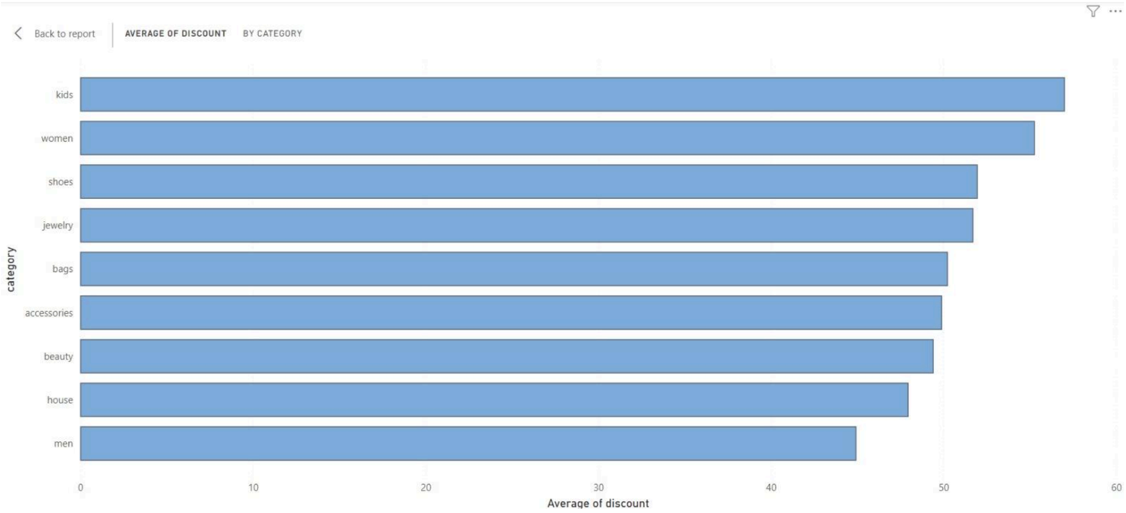
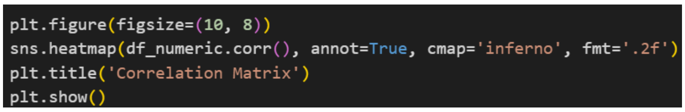


Figure 21: Average 'discount' by Category

In order to further understand the relationship between the numerical attributes, a confusion matrix was created (refer to Figure 22). In a correlation matrix, the closer the relationship value of the attributes to 1, the stronger their relationship (Wagavkar 2023). However, the closer the relationship value of the attributes to -1, the weaker their relationship (Wagavkar 2023). Relationship values closer to '0' indicate that the attributes have neutral relationships (Wagavkar 2023). From this information and the results provided in Figure 22, we can deduce that almost all numerical attributes have neutral relationships except 'raw_price' and 'current_price' which have a stronger relationship with each other. This aligns with the fact that the 'current_price' = 'raw_price' - 'discount'.



```
plt.figure(figsize=(10, 8))
sns.heatmap(df_numeric.corr(), annot=True, cmap='inferno', fmt='.2f')
plt.title('Correlation Matrix')
plt.show()
```

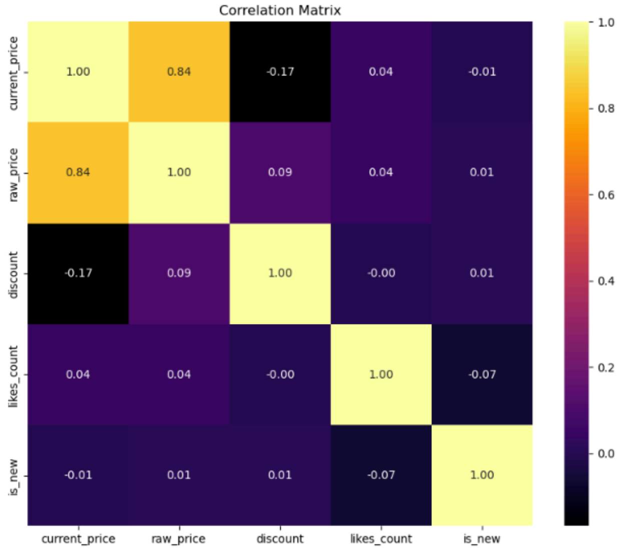


Figure 22: Correlation Matrix

3. Clustering:

3.1. Clustering Algorithm 1 - K-Means:

In supervised learning, every dataset has a label or target value that represents the outcome that the model should predict. On the other hand, in unsupervised learning, we don't have these labels or targets. We only have the data to visualise and analyse, with no clear label or target on what outcome we're looking for. One very popular approach to an unsupervised data frame is clustering, where we group similar data points to discover patterns or structures within the data (S, 2022). Clustering helps us explore the data and find groups within the data where each cluster represents data points that are similar (Dabbura, 2022).

One of the popular methods of unsupervised clustering is the K-Means algorithm which is widely used in data mining and the field of data analysis (Li & Wu 2012). This algorithm is very effective for grouping large datasets where manual grouping would be tedious and impractical (Jeffares 2021).

3.1.1 Model Planning

The goal is to analyze a dataset of products of different categories (e-commerce website, NewChic) by applying the K-Means clustering algorithm where we will try to group similar products based on important features such as the number of likes for each product (likes_count), percentage of discount given to a product (discount), and the current price of the product (current_price). By doing so we expect to have multiple clusters where the data points within each group/cluster are similar and dissimilar to those in other groups. Our goal is to find out the top 10 products from all nine given categories, a total of 74462 products, and finally, find out the best category. We followed the following steps to build our model.

Data Collection and Preprocessing

Before we started building the K-Means model several preprocessing steps were applied to the dataset such as checking and removing null values, standardisation of data, and dropping non-numerical values. A detailed overview is described in the model-building phase.

Choosing the Value of K

One of the most important steps before applying the K-Means clustering algorithm is determining the number of clusters also known as the value of K (S 2022). We must execute the K-Means clustering algorithm for different values of K and analyse the results to find the optimal number of clusters required for the data. The value of K determines how well the K-Means algorithm performs as it is essential to achieve good results.

Perform K-Means Clustering

After determining the ideal number of clusters using the Elbow method, we implement the K-Means clustering algorithm.

- Initialise the Cluster Centres: First, the process starts with k initial cluster centres, called centroids where these values are initially selected randomly from the data points. (KMEANS n.d.)
- Assign Data Points: Secondly, each data point is assigned to the closest cluster by calculating the sum of the squared distance. From each data point distance to all clusters is measured to find the closest cluster and therefore assign the data point to a centroid (Dabbura 2022).
- Update Cluster Centres: Once all data points are assigned, the cluster centroids are updated by taking an average (mean) of all of the data points that are associated with the respective cluster. (Jeffares 2021)
- Iterate: Finally, above mentioned assigning data points and updating cluster centres steps are iterated until the centres no longer update or we can stop the iteration by setting a pre-defined number of iterations (Dabbura 2022).

Analyse the Clusters

In this step, we analyse each cluster and its associated data to determine the similarity or pattern to identify what common features grouped them and what features differentiated them from other groups.

3.1.2 Model Building

Once we determine the model planning and steps to follow, we can start building the model. It consists of several steps namely data collection, pre-processing, finding the optimal value of K, applying K-Mean clustering, and analysis of clusters.

Data Collection:

We have the pre-processed dataset where several data cleaning and preprocessing tasks have been performed such as keeping relevant columns of data and dropping irrelevant columns such as image URL, image, and thumbnail. The missing values were dropped and Boolean values were converted to numeric values. To ensure data integrity, duplicate values were removed as well and high-range values were scaled. This pre-processed data is accessed via Google Drive and loaded using the ‘pandas’ library. The file is named ‘Updated_CSV.csv’.

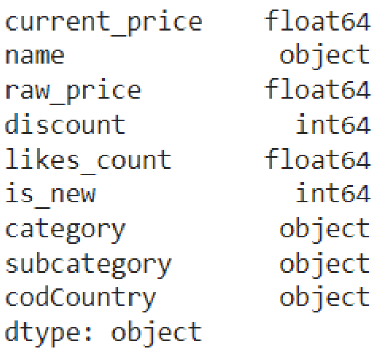
Pre-processing:

To tailor the data for the K-Mean clustering algorithm we need to compute additional pre-processing steps to ensure that the dataset is optimised and scaled properly.

- **Handling Missing Values:** We start by checking for any null or missing values in the dataset to ensure the integrity of the data because null values may hinder the clustering process. Fortunately, there were no missing values.

- **Normalisation/Scaling:** Upon visualising the dataset we can see only the likes_count column is scaled because of high value numbers. To standardise the data and to ensure that all features contributed equally to the clustering, columns like raw_price, discount, and current_price in the dataset were also scaled using MinMaxScaler which is imported from sklearn.preprocessing library. This step converted the data to a range of 0 to 1, making it suitable for clustering.

- **Feature Selection:** In this step, we only kept the numeric value features and dropped object type ones.

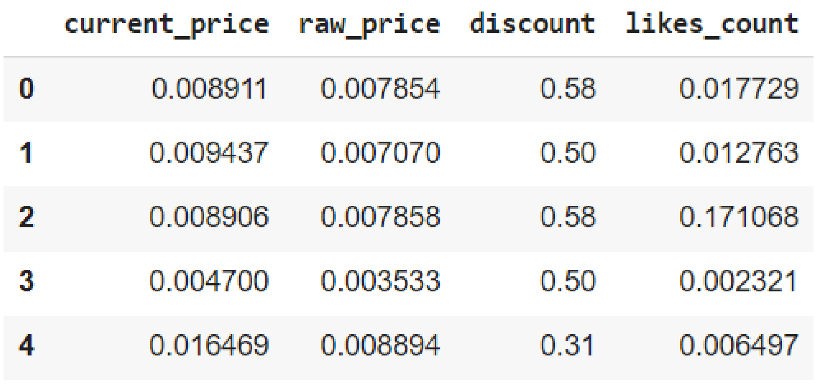


```
current_price    float64
name              object
raw_price        float64
discount           int64
likes_count      float64
is_new             int64
category          object
subcategory       object
codCountry        object
dtype: object
```

Figure 23: Data types of Dataset

The columns that are being removed are ‘name’, ‘is_new’, ‘category’, ‘subcategory’, and ‘codcountry’. The purpose of dropping these columns is necessary because they are not needed as we are only focusing on numerical data for clustering.

After completing further preprocessing we have the following dataset-



	current_price	raw_price	discount	likes_count
0	0.008911	0.007854	0.58	0.017729
1	0.009437	0.007070	0.50	0.012763
2	0.008906	0.007858	0.58	0.171068
3	0.004700	0.003533	0.50	0.002321
4	0.016469	0.008894	0.31	0.006497

Figure 24: Pre-processed Dataset

Determining the Optimal Number of Clusters

Determination of an optimal number of clusters is crucial before performing K-Means clustering.

- **The Elbow Method:** By using this approach, we can understand how many clusters the data will be organised making it relevant and well-defined. First, we defined a range of K (1 to 9) of potential cluster numbers. For each value of K, we applied the K-Means clustering algorithm to the dataset defining clustering based on three key features: likes_count, discount, and current_price. This method involves calculating the sum of squared distances between each data point and its nearest cluster centre against the number of clusters. To find out the relationship between the number of clusters and the corresponding sum of squared distance values, we plotted an "Elbow Graph."

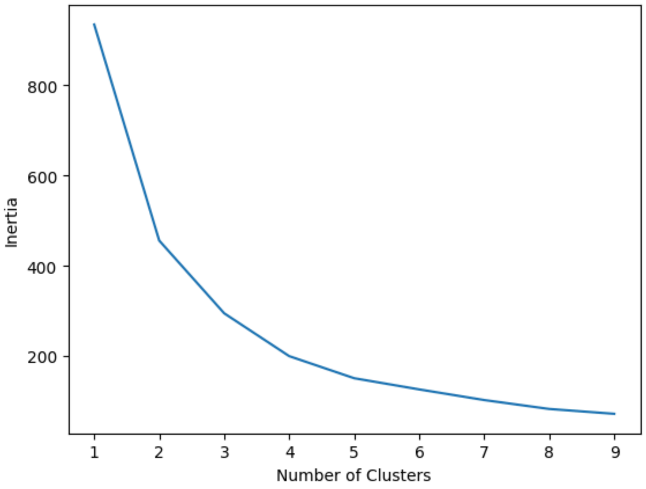


Figure 25: The Elbow Graph

The x-axis of the graph represents the number of clusters, while the y-axis represents the inertia values. The point where the line begins to drop and takes the shape of an elbow, suggests the optimal number of clusters (Lee 2024). Here, the graph suggests an optimal number of cluster values of 3.

K-Means Clustering

Based on the results from the Elbow method we determined the optimal number of clusters which helped us avoid the trial and error of having too low or too many clusters.

The purpose of this step is to apply the KMeans clustering algorithm to our dataset, specifically focusing on the likes_count, discount, and current_price features. Based on our earlier analysis using the Elbow Method, we have determined that three clusters (k=3) are likely optimal for our data. We initialized the KMeans algorithm with n_clusters=3, indicating that we want to partition our data into three distinct clusters.

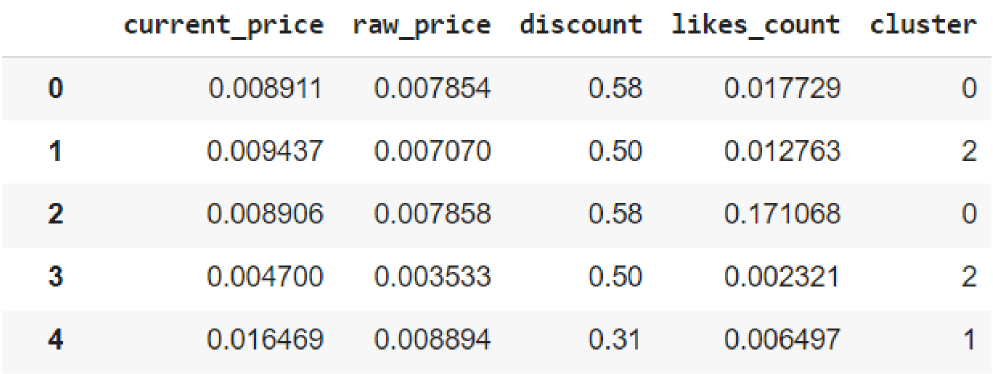
- **KMeans Clustering:** First, we initialized the K-Means algorithm with n_clusters=3.

```
# Using KMeans Cluster
km = KMeans(n_clusters=3)
y_predicted = km.fit_predict(dataset[['likes_count', 'discount', 'current_price']])
dataset['cluster']=y_predicted
```

Figure 26: Code Snippet for K-Means

The KMeans algorithm was applied to our dataset targeting the defined features likes_count, discount, and current_price columns. According to their significance, these features were selected. The number of likes counts of a product is a big indicator of whether the product is liked by the customers or not and its popularity. Similarly, the amount of discount and price drives the sale and the product. Lastly, the model assigned each product to one of the three clusters [0, 1 or 2] based on the features likes_count, discount, and current_price.

- **Cluster Label:** Finally, a new column named cluster was added to the dataset ‘dataset’ which indicates which group the product belongs. Here is the dataset after adding the cluster column. Each product belongs to either 0, 1, or 2 clusters based on feature characteristics.



	current_price	raw_price	discount	likes_count	cluster
0	0.008911	0.007854	0.58	0.017729	0
1	0.009437	0.007070	0.50	0.012763	2
2	0.008906	0.007858	0.58	0.171068	0
3	0.004700	0.003533	0.50	0.002321	2
4	0.016469	0.008894	0.31	0.006497	1

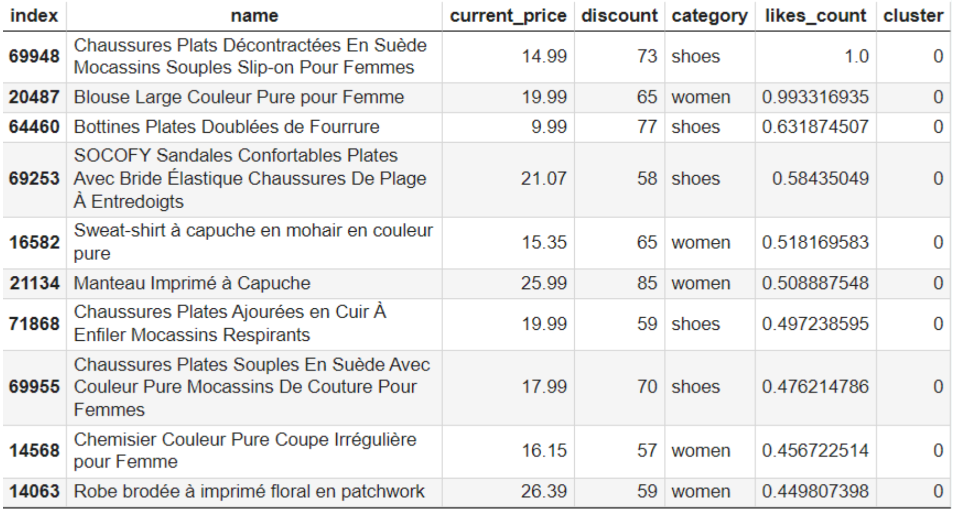
Figure 27: Dataset After Adding the Cluster Column

Visualising the Clusters

Now, we can start analysing the data to visualise which products belong to which cluster, the possible logical reason for that and find a better understanding of the correlations between the three important variables—likes_count, discount, and current_price.

Dataset of Each Cluster Analysis

Cluster 0: The table displays a subset of products presented in cluster 0 sorted by highest likes count.

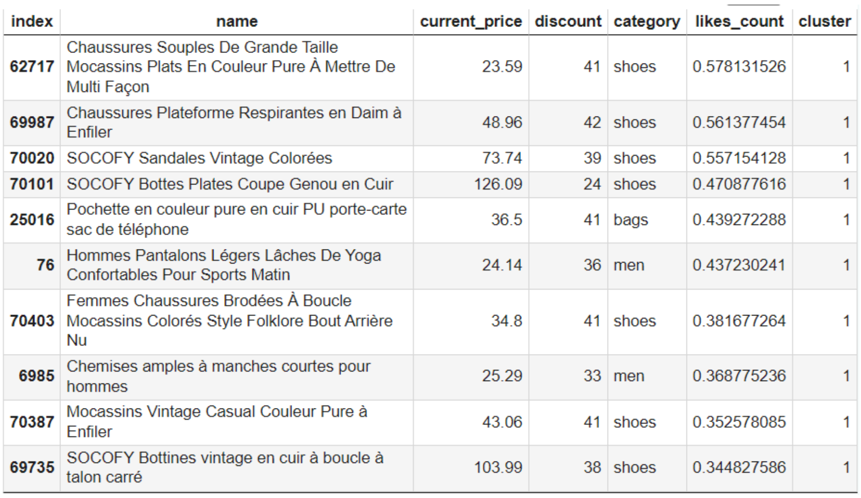


index	name	current_price	discount	category	likes_count	cluster
69948	Chaussures Plats Décontractées En Suède Mocassins Souples Slip-on Pour Femmes	14.99	73	shoes	1.0	0
20487	Blouse Large Couleur Pure pour Femme	19.99	65	women	0.993316935	0
64460	Bottines Plates Doublées de Fourrure	9.99	77	shoes	0.631874507	0
69253	SOCOFY Sandales Confortables Plates Avec Bride Élastique Chaussures De Plage À Entredoigts	21.07	58	shoes	0.58435049	0
16582	Sweat-shirt à capuche en mohair en couleur pure	15.35	65	women	0.518169583	0
21134	Manteau Imprimé à Capuche	25.99	85	women	0.508887548	0
71868	Chaussures Plates Ajourées en Cuir À Enfiler Mocassins Respirants	19.99	59	shoes	0.497238595	0
69955	Chaussures Plates Souples En Suède Avec Couleur Pure Mocassins De Couture Pour Femmes	17.99	70	shoes	0.476214786	0
14568	Chemisier Couleur Pure Coupe Irrégulière pour Femme	16.15	57	women	0.456722514	0
14063	Robe brodée à imprimé floral en patchwork	26.39	59	women	0.449807398	0

Figure 28: Data Table for Cluster 0

In this table, we can see the current price ranges from 9.99 to 26.39 neither too low nor very high, and discounts range from 57% to 85%, which indicates cluster 0 might be associated with products that generally have high discount rates or items on clearance. The effect of higher discounts can be seen in the like count as most values above 0.4 suggest that products in this cluster are well-liked. On the other hand, the highest discounts do not necessarily result in the highest likes count. For example, the highest-liked item is associated with a 73% discount which is not the highest discount in the cluster. By looking at the price we can assume these products target budget-conscious consumers, considering low prices and significant discounts.

Cluster 1: The table shows a subset of products presented in cluster 1 sorted by highest likes count.

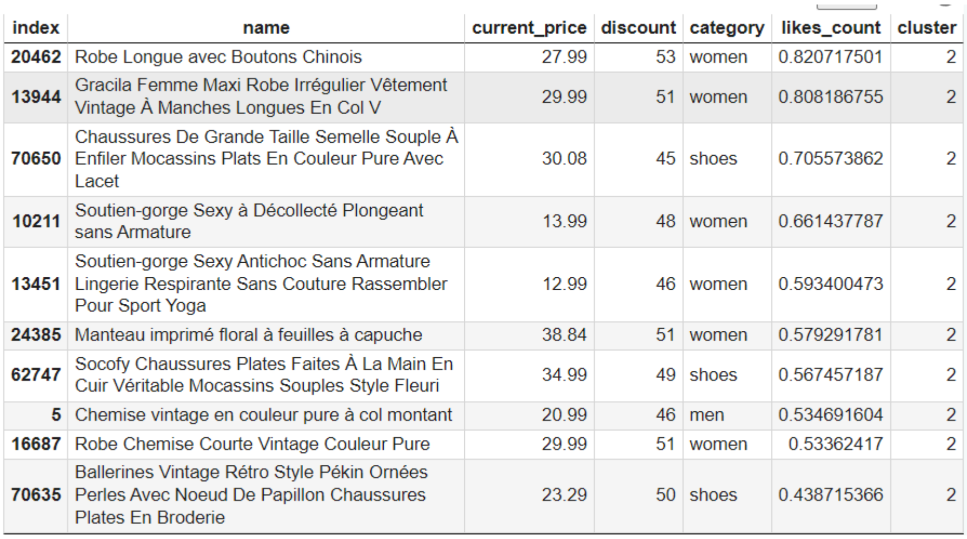


index	name	current_price	discount	category	likes_count	cluster
62717	Chaussures Souples De Grande Taille Mocassins Plats En Couleur Pure À Mettre De Multi Façon	23.59	41	shoes	0.578131526	1
69987	Chaussures Plateforme Respirantes en Daim à Enfiler	48.96	42	shoes	0.561377454	1
70020	SOCOFY Sandales Vintage Colorées	73.74	39	shoes	0.557154128	1
70101	SOCOFY Bottes Plates Coupe Genou en Cuir	126.09	24	shoes	0.470877616	1
25016	Pochette en couleur pure en cuir PU porte-carte sac de téléphone	36.5	41	bags	0.439272288	1
76	Hommes Pantalons Légers Lâches De Yoga Confortables Pour Sports Matin	24.14	36	men	0.437230241	1
70403	Femmes Chaussures Brodées À Boucle Mocassins Colorés Style Folklore Bout Arrière Nu	34.8	41	shoes	0.381677264	1
6985	Chemises amples à manches courtes pour hommes	25.29	33	men	0.368775236	1
70387	Mocassins Vintage Casual Couleur Pure à Enfiler	43.06	41	shoes	0.352578085	1
69735	SOCOFY Bottines vintage en cuir à boucle à talon carré	103.99	38	shoes	0.344827586	1

Figure 29: Data Table for Cluster 1

The price in this cluster ranges widely from 23.59 to 126.09 units, with most of the products priced higher than those in Cluster 0. On the other hand, the discount rates in this cluster are generally lower, ranging from 24% to 42%. This is a significantly narrower range compared to Cluster 0, indicating high-value items with fewer discounts. Similarly, the likes count in this cluster is slightly lower on average which also indicates these products may be premium products with specific target audiences.

Cluster 2: The table depicts a subset of products presented in cluster 2 sorted by highest likes count.



index	name	current_price	discount	category	likes_count	cluster
20462	Robe Longue avec Boutons Chinois	27.99	53	women	0.820717501	2
13944	Gracila Femme Maxi Robe Irrégulier Vêtement Vintage À Manches Longues En Col V	29.99	51	women	0.808186755	2
70650	Chaussures De Grande Taille Semelle Souple À Enfiler Mocassins Plats En Couleur Pure Avec Lacet	30.08	45	shoes	0.705573862	2
10211	Soutien-gorge Sexy à Décolleté Plongeant sans Armature	13.99	48	women	0.661437787	2
13451	Soutien-gorge Sexy Antichoc Sans Armature Lingerie Respirante Sans Couture Rassembler Pour Sport Yoga	12.99	46	women	0.593400473	2
24385	Manteau imprimé floral à feuilles à capuche	38.84	51	women	0.579291781	2
62747	Socofy Chaussures Plates Faites À La Main En Cuir Véritable Mocassins Souples Style Fleuri	34.99	49	shoes	0.567457187	2
5	Chemise vintage en couleur pure à col montant	20.99	46	men	0.534691604	2
16687	Robe Chemise Courte Vintage Couleur Pure	29.99	51	women	0.53362417	2
70635	Ballerines Vintage Rétro Style Pékin Ornées Perles Avec Noeud De Papillon Chaussures Plates En Broderie	23.29	50	shoes	0.438715366	2

Figure 30: Data Table for Cluster 2

In this table, most items are priced between 12.99 and 38.84, this moderate price range falls in between cluster 0 and cluster 1. Unlike other clusters, there is a correlation between higher discounts and likes count where higher discounts drive higher customer engagement. Most of the products have a discount range of 45-53% with higher likes count indicating these might be popular trendy products.

3.1.3 Result Communication

Scatter Plot: Scatter plots are a simple yet powerful way to show the relationship between two numerical variables. Scatter plots can help us identify patterns, similarities, and correlations in the data (Pathak 2024). To use scatter plots we imported the library from using matplotlib. The purpose of this analysis is to visually examine the clusters generated by the KMeans algorithm by plotting relationships between key variables like count, discount, and current price.

Scatter Plot: Discount vs. Likes Count

- **Visualization:**
    - Cluster 0: Green Dots
    - Cluster 1: Red Dots
    - Cluster 2: Black Dots

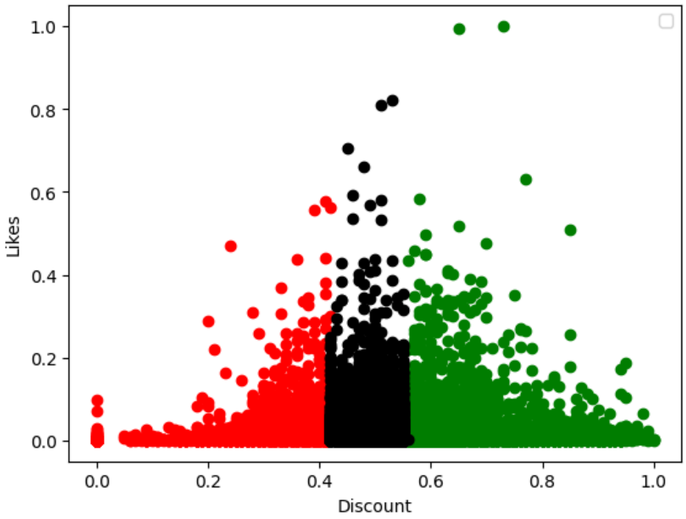


Figure 31: Scatter Plot - discount vs likes_count

- **Distribution Across Clusters**
    - *Cluster 0 (Green):* Upon visualising the scatter plot, we can see it is scattered on the higher end of the discount axis, having more than 56% and above discounts. The Likes count in this cluster spread across the y-axis, indicates a wider variation in popularity. Products in this cluster may receive strong engagement, possibly due to attractive discounts however the downward trend in likes count with an increase of discount suggests otherwise.
    - *Cluster 1 (Red):* Having between 0% and around 42% discount, this cluster concentrated on the lower end of the discount axis. The Likes in this cluster are generally low following lower discounts but show an upward trend with gradual increases in discounts.

- *Cluster 2 (Black):* On the x-axis, it occupies the 42% to 55% discount range. The likes in this cluster are more diverse and spread out compared to the other two clusters.

- **Overall Observation**

  There's a noticeable rise in likes around the 42% to 55% discount range (Cluster 2). This might suggest that these are a bestselling range of products as higher discounts in this range are more effective at driving engagement compared to very low or very high discounts.

  In cluster 0 (Green), there’s more diversity in likes. Some products with high discounts receive a lot of likes, while others having even higher discounts don't, indicating that high discounts alone might not guarantee likes.

  Cluster 1 (Red) shows a straightforward relation of products with low discounts with fewer likes, which may indicate that smaller discounts are less effective at generating interest.

Scatter Plot: Likes Count vs. Current Price

- **Visualisation:**
  - Cluster 0: Green Dots
  - Cluster 1: Red Dots
  - Cluster 2: Black Dots

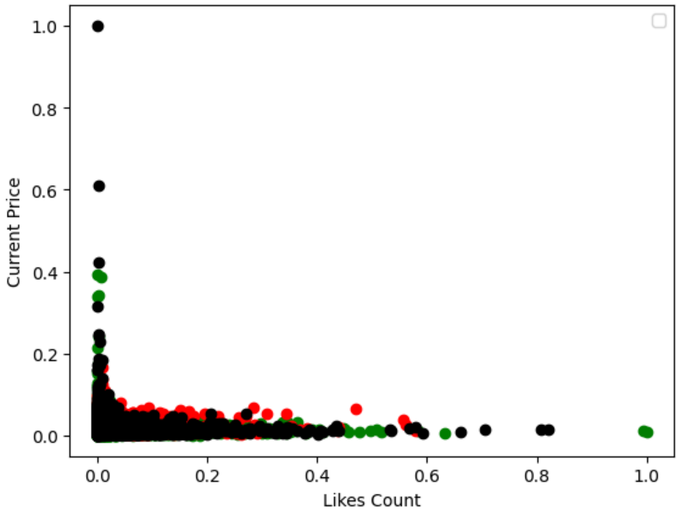


Figure 32: Scatter Plot - likes_count vs current_price

The scatter plot depicts the relationship between Likes Count (x-axis) and Current Price (y-axis). Here are some key observations:

Most of the items have low price range and lower likes counts which indicates a lot of lower-priced products are not liked well by customers.

Scatter Plot: Discount vs. Current Price

- **Visualisation:**
    - Cluster 0: Green Dots
    - Cluster 1: Red Dots
    - Cluster 2: Black Dots

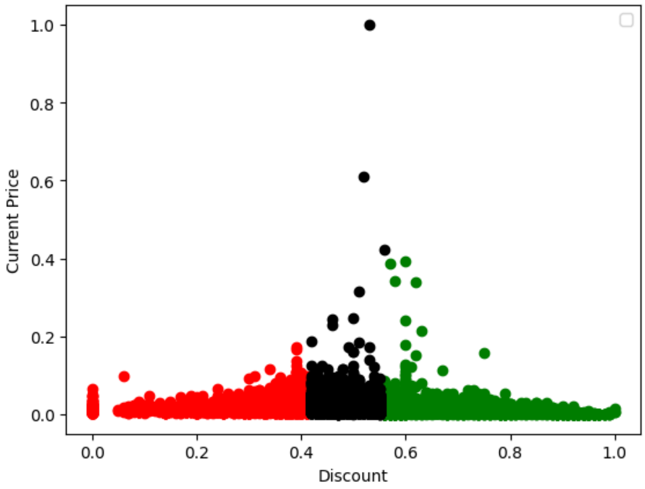


Figure 33: Scatter Plot - discount vs current_price

The scatter plot shows the relationship between Discount (horizontal axis) and Current Price (vertical axis). By analysing the data, we can say most of the items belong to the lower price group, even the high-priced items are highly discounted which resulted in lower prices for those products as well.

Based on the visuals of scatter plots, we can draw the conclusion that features Discount and Likes count produced the most diverse data range. To further analyse the relation and diversity of discount and likes count we can draw a box chart to better understand.

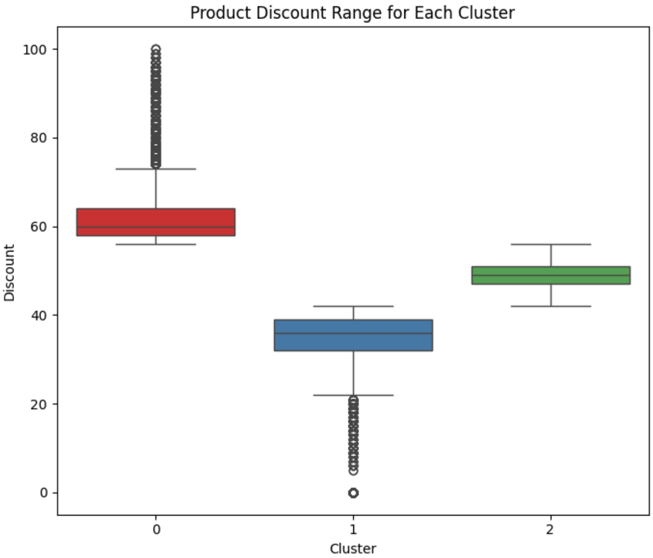


Figure 34: Product Discount Range for Each Cluster

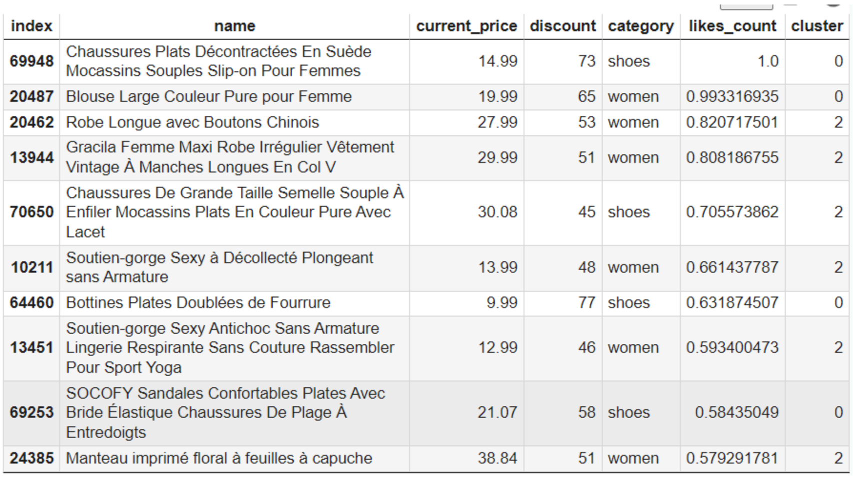
It displays the distribution of the discount percentage for all the products within three different clusters. We can visualise how different ranges of discounted products are grouped in respective clusters. For example, all the products ranging from 0-42% discounts belong to cluster 1.

Bar charts:

To evaluate the results, we can draw bar charts to depict the patterns of different clusters based on categories, like counts and discounts.

Cluster distribution for Top 10 products across all categories

This table represents the dataset for the most liked 10 products from all categories combined.



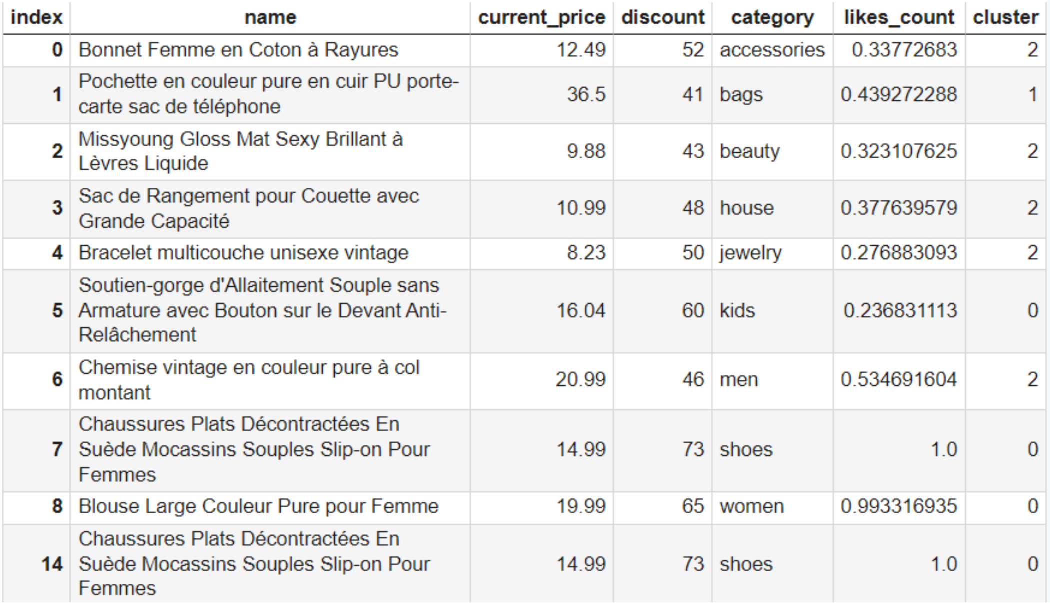
index	name	current_price	discount	category	likes_count	cluster
69948	Chaussures Plats Décontractées En Suède Mocassins Souples Slip-on Pour Femmes	14.99	73	shoes	1.0	0
20487	Blouse Large Couleur Pure pour Femme	19.99	65	women	0.993316935	0
20462	Robe Longue avec Boutons Chinois	27.99	53	women	0.820717501	2
13944	Gracila Femme Maxi Robe Irrégulier Vêtement Vintage À Manches Longues En Col V	29.99	51	women	0.808186755	2
70650	Chaussures De Grande Taille Semelle Souple À Enfiler Mocassins Plats En Couleur Pure Avec Lacet	30.08	45	shoes	0.705573862	2
10211	Soutien-gorge Sexy à Décollecté Plongeant sans Armature	13.99	48	women	0.661437787	2
64460	Bottines Plates Doublées de Fourrure	9.99	77	shoes	0.631874507	0
13451	Soutien-gorge Sexy Antichoc Sans Armature Lingerie Respirante Sans Couture Rassembler Pour Sport Yoga	12.99	46	women	0.593400473	2
69253	SOCOFY Sandales Confortables Plates Avec Bride Élastique Chaussures De Plage À Entredoigts	21.07	58	shoes	0.58435049	0
24385	Manteau imprimé floral à feuilles à capuche	38.84	51	women	0.579291781	2

Figure 35: Top 10 Products from All Categories

From this data, we can predict from the table that the most liked products from all categories have higher discounts (42% and above). That is why all these products belong to clusters 0 and 2.

Cluster distribution for Top 10 products (at least 1 product from each category)

This table represents the dataset for the most liked 10 products where at least 1 product is present from every unique category.



index	name	current_price	discount	category	likes_count	cluster
0	Bonnet Femme en Coton à Rayures	12.49	52	accessories	0.33772683	2
1	Pochette en couleur pure en cuir PU porte-carte sac de téléphone	36.5	41	bags	0.439272288	1
2	Missyoung Gloss Mat Sexy Brillant à Lèvres Liquide	9.88	43	beauty	0.323107625	2
3	Sac de Rangement pour Couette avec Grande Capacité	10.99	48	house	0.377639579	2
4	Bracelet multicouche unisexe vintage	8.23	50	jewelry	0.276883093	2
5	Soutien-gorge d'Allaitement Souple sans Armature avec Bouton sur le Devant Anti-Relâchement	16.04	60	kids	0.236831113	0
6	Chemise vintage en couleur pure à col montant	20.99	46	men	0.534691604	2
7	Chaussures Plats Décontractées En Suède Mocassins Souples Slip-on Pour Femmes	14.99	73	shoes	1.0	0
8	Blouse Large Couleur Pure pour Femme	19.99	65	women	0.993316935	0
14	Chaussures Plats Décontractées En Suède Mocassins Souples Slip-on Pour Femmes	14.99	73	shoes	1.0	0

Figure 36: Top 10 Products (At Least 1 Product from Each Category)

Instead of picking the 10 best products from all categories, if we take at least 1 product from each category, this is the tabular representation where the majority of the items belong to clusters 0 and 2 justifying our claim that higher discount items tend to have higher engagement from customers.

Cluster distribution for the Top 10 products from each category

This bar chart depicts the top 10 products from each category and their distribution across the clusters. Here we have taken 10 products from each category, totalling 90 products from 9 categories and we can visualise that almost 85% of top products belong to clusters 0 and 2 combined.

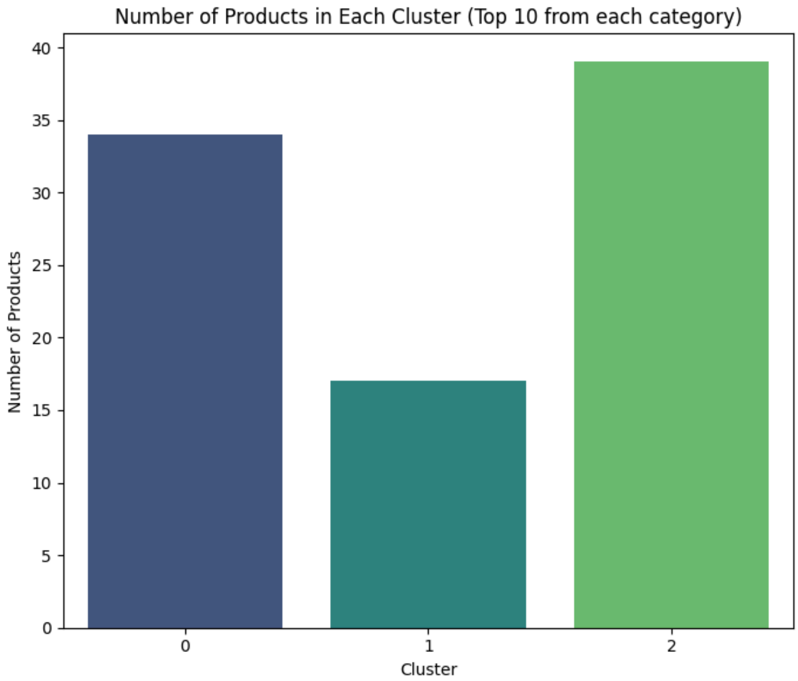


Figure 37: Top 10 Products from Each Category and Distribution in Clusters

Cluster distribution for the Top 100 products from all categories

Similarly, if we compute another bar chart that shows the distribution of clusters for the top 100 products from all categories, we can get similar results where 88% of products belong to clusters 0 and 2.

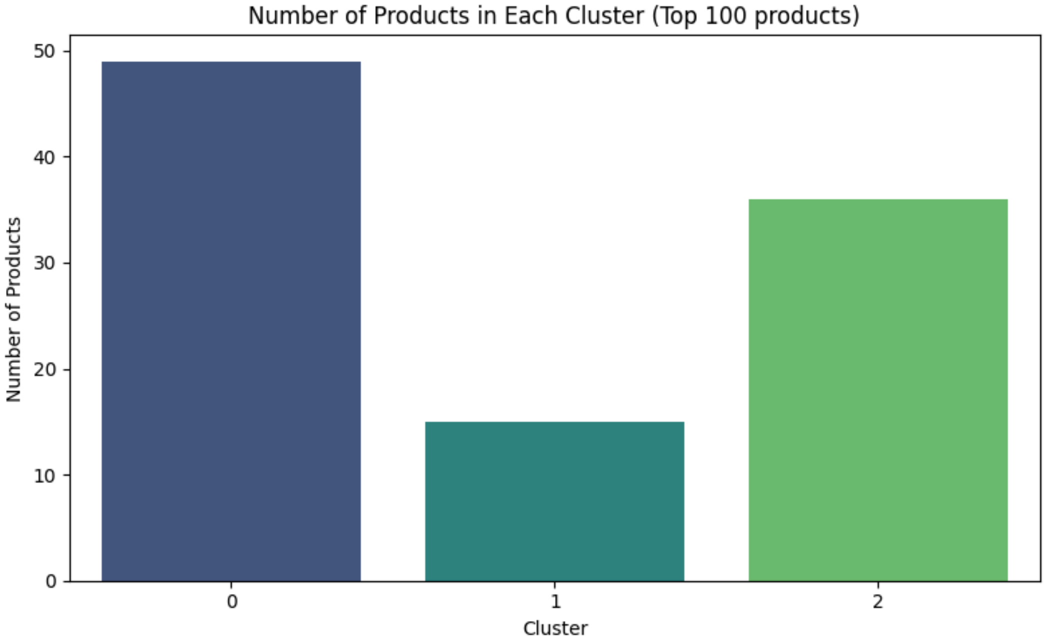


Figure 38: Cluster Distribution for the Top 100 Products from All Categories

Top 100 products category and cluster distribution

We can also plot a bar chart to display the top 100 products and their respective categories and clusters.

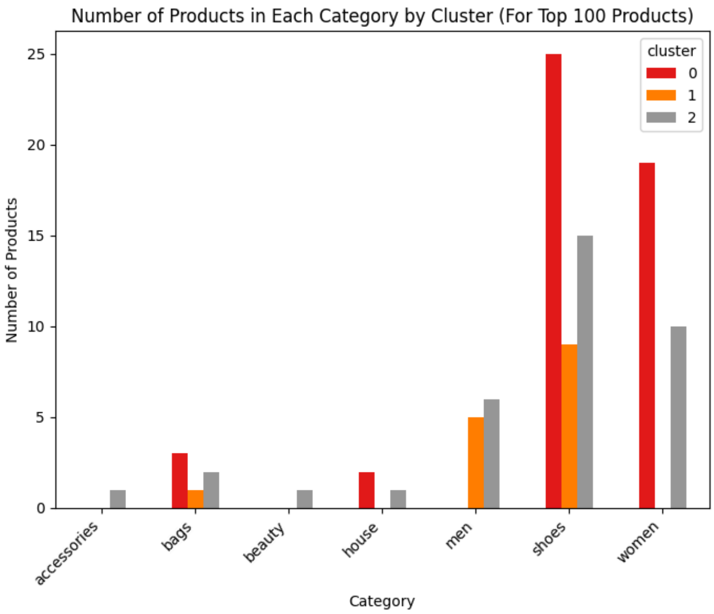


Figure 39: Top 100 Products Category and Cluster Distribution

In this bar chart, we can see two categories' women and shoes have the most liked products indicating shoes as the potential best category,

Silhouette Score:

Finally, to evaluate the quality of the clustering, the Silhouette Score was calculated for each category to determine how well the KMeans algorithm performed. Here are the scores-

Category	K-Means Silhouette Score
Men	0.33
Women	0.31
Bags	0.19
Kids	0.27
Beauty	0.16
House	0.21
Jewellery	0.29

Accessories	0.47
Shoes	0.15

Table 1: Silhouette Score By Category

The range of silhouette scores is from -1 to +1 where a positive number closer to 1 means the clusters are separated well whereas a negative value indicates there is overlapping in data points and clustering did not work well. In our K-Means model, we obtained all positive values having the highest 0.47 and the lowest 0.15. As a result, our model is an average-performing one.

In conclusion, the KMeans clustering algorithm was able to effectively group the products into three clusters, revealing distinct and diverse patterns in products based on likes, pricing, and discount strategies. The visualisations provided clear insights into how different products belong to different clusters based on their features. This analysis can provide valuable insight for e-commerce businesses to understand customer reactions based on different strategies.

3.2. Clustering Algorithm 2 - DBSCAN:

3.2.1 Model Planning:

Overview

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm that identifies clusters by detecting areas of high data point density, effectively distinguishing these clusters from outliers (Ester et al., 1996). Unlike centroid-based approaches like K-Means, DBSCAN doesn’t require specifying the number of clusters in advance, making it ideal for exploratory data analysis (Ester et al., 1996).

- Though there are several features available, including `current_price`, `raw_price`, `is_new`, `category`, and ‘subcategory’, `discount` is chosen as the X-axis because it directly influences consumer purchasing decisions, while likes_count is selected as the Y-axis as it indicates customer engagement and product popularity. These two features together provide a complementary perspective on pricing strategy and consumer behaviour, making them ideal for clustering analysis to uncover patterns in product performance.

Key Parameters

- **Min_samples**: Specifies the minimum number of points required to form a dense region, which is considered a cluster. Points meeting this criterion are labelled as core points (Scikit-learn, 2017).
- **Epsilon (ε)**: Defines the maximum distance between two points for them to be considered as part of each other's neighborhood, effectively setting the radius around a core point (Scikit-learn, 2017).

DBSCAN Concepts

- **Core Points**: Points that have at least the number of Min_samples within their Epsilon distance, forming the backbone of clusters (Scikit-learn, 2017).
- **Border Points**: Points that fall within the Epsilon range of a core point but don't have enough neighbouring points to be considered core points themselves (Scikit-learn, 2017).
- **Noise Points**: Points that are neither core points nor border points and are treated as outliers, excluded from any clusters (Scikit-learn, 2017).

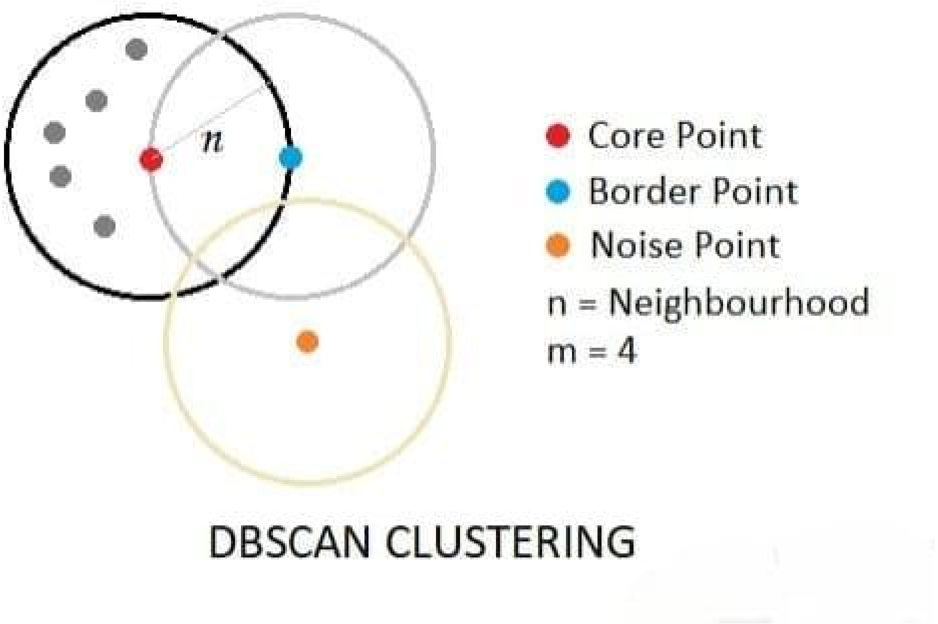


Figure 40: DBSCAN Clustering (Scikit-learn, 2017)

Advantages of DBSCAN

- **Handling Noise and Outliers**: DBSCAN is proficient at managing noise and outliers by focusing on clusters as dense regions separated by less dense areas (Scikit-learn, 2017).
- **Cluster Shape Flexibility**: DBSCAN is capable of identifying clusters of various shapes and sizes, unlike methods that assume spherical or similarly sized clusters (Scikit-learn, 2017).

Practical Implementation

- **Data Preprocessing**: Data should be normalized to ensure all features contribute equally to the distance calculations (Scikit-learn, 2017).
- **Parameter Tuning**: Choosing appropriate values for Min_samples and Epsilon is critical for effective clustering. The k-distance graph can assist in determining a suitable Epsilon value (Scikit-learn, 2017).

3.2.2 Model Building:

Data Preparation

- **Filtering and Feature Selection**: The dataset was filtered to focus on relevant product categories. The selected features for clustering included current_price, discount, and likes_count for the same reasons stated in Section 3.1.2.
- **Normalisation**: The features were normalised to ensure that each one contributed equally to the clustering process (Han, 2000).

Model Training and Tuning

- **Parameter Tuning**: Various combinations of eps (ε) and Min_samples were tested to determine the optimal parameters for DBSCAN. The silhouette coefficient was used to measure the quality of clustering, with higher scores indicating better cluster definition.

- **DBSCAN Clustering**: The DBSCAN algorithm was applied using the optimal parameters identified during the tuning process, clustering data points based on their density (Rousseeuw, 1987).

Figures:

- **Figure 1**: **Data Proximity Plot** - Illustrates the distance to each point's k-nearest neighbours, aiding in the selection of an appropriate epsilon value (Wickman, 2016).

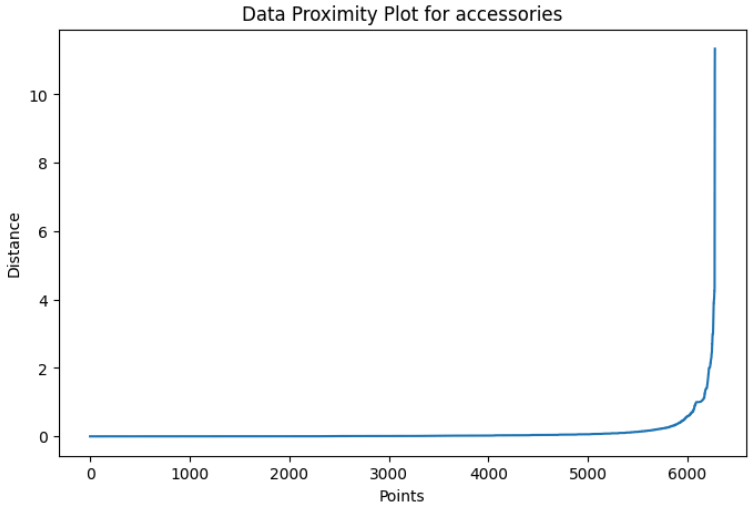


Figure 41: Data Proximity Plot

- **Figure 2**: **Clustering Results** - Visual representation of the clusters formed by DBSCAN, highlighting how the algorithm groups dense data regions while classifying sparse regions as noise.

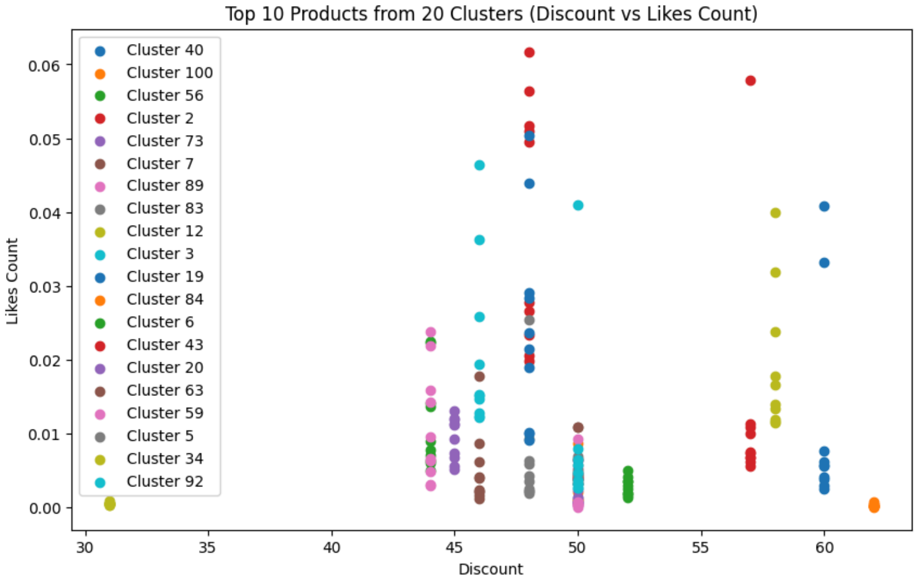


Figure 42: Clustering Result

Model Evaluation

- **Silhouette Coefficient**: Used to evaluate the effectiveness of the clustering. A high silhouette score indicates well-separated and well-defined clusters (Wickman, 2016).
- **Visualization of Clusters**: Scatter plots and bar plots were created to visualize the top products within each cluster as well as the overall top products (Wickman, 2016).

3.2.3 Result Communication:

Clustering Results Overview

DBSCAN proved to be a robust unsupervised clustering method, capable of identifying clusters without predefining their number. The implementation involved systematically testing various combinations of epsilon and Min_samples values to find the best clustering parameters. The algorithm was applied to NewChic’s dataset, covering categories like Accessories, Bags, Kids, Beauty, Jewelry, Women, Men, House, and Shoes. The results showed that DBSCAN effectively identified well-defined clusters, as evidenced by high silhouette scores. This density-based approach ensured that data points within clusters were closely related to their core points, confirming the algorithm's effectiveness in partitioning the data meaningfully (Jain et al., 1999).

Top Products by Category

For each analyzed category, the top 10 products were identified based on the clustering results and ranked by likes_count. The key findings are summarized below:

- **Accessories Category:**
  - *Silhouette score:* -0.1161
  - *Epsilon value:* 0.151
  - *Min Sample value:* 8

  For the Accessories dataset, the best DBSCAN parameters were determined by tuning the epsilon value to 0.151 and setting the min_samples value to 8.

Despite a negative silhouette score of -0.1161, the clustering revealed key insights into product popularity. High-engagement items like 'Beanie Hat' dominated the top products, followed closely by 'Chaussettes & Collants,' 'Baseball Caps,' and 'Straw Hats.' These products showed a strong balance between discounts and user engagement, reflecting their appeal within the Accessories category.

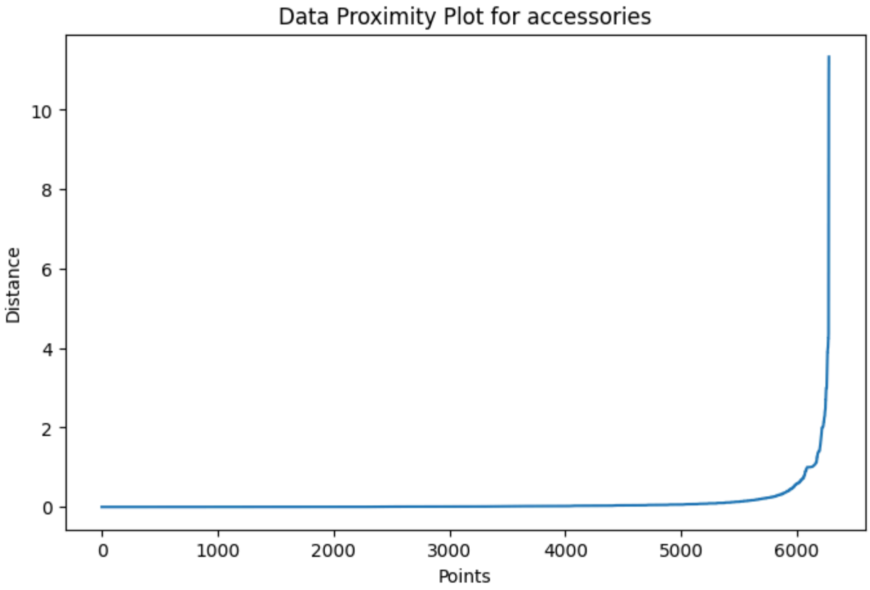


Figure 43: Data Proximity Plot for Accessories

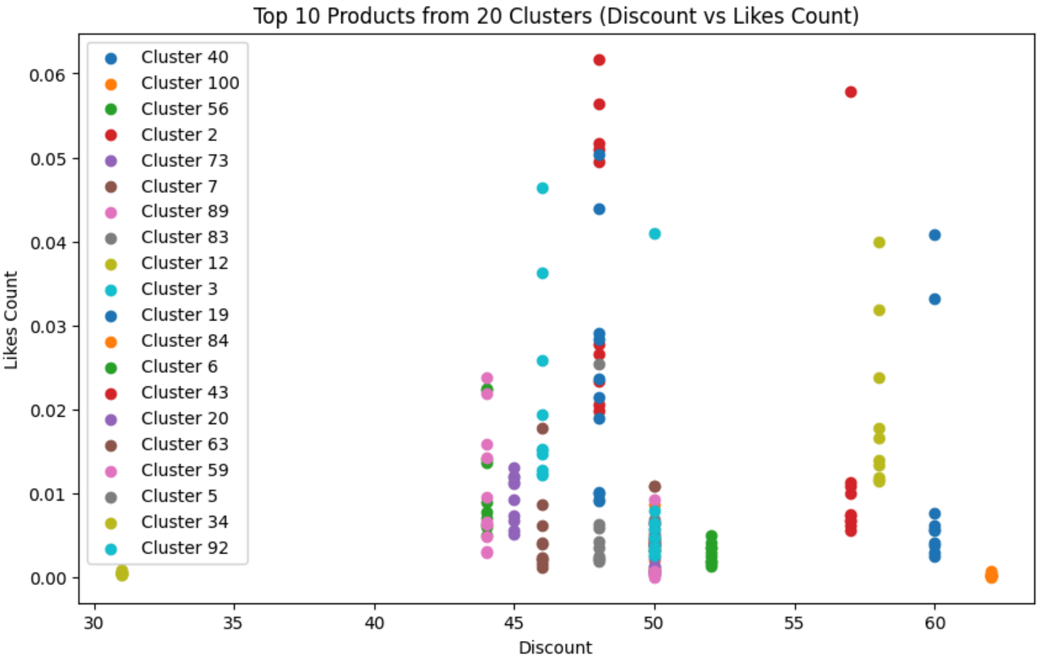


Figure 44: Top 10 Products from 20 Clusters for Accessories

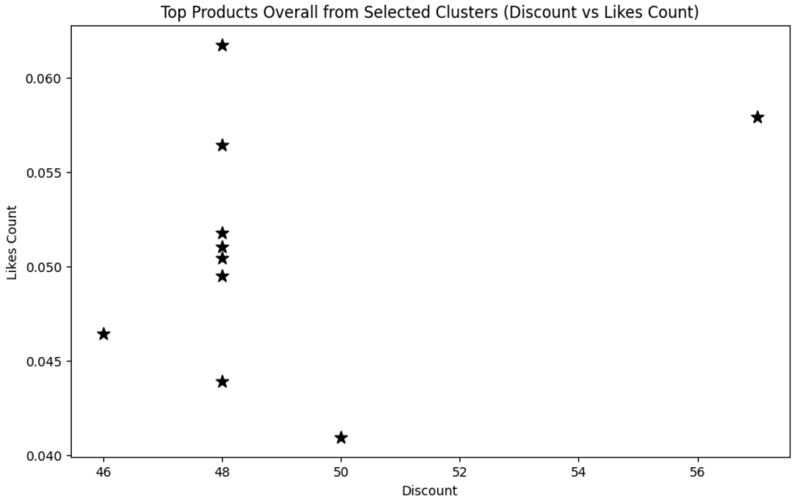


Figure 45: Top Products Overall from Selected Clusters for Accessories

- **Bags Category:**
  - *Silhouette score:* -0.2913
  - *Epsilon value:* 0.788
  - *Min Sample value:* 11

In the Bags category, the DBSCAN algorithm was tuned to an epsilon value of 0.788 and a min_samples value of 11. The silhouette score of -0.2913 suggested some challenges in cluster separation. However, the analysis highlighted top products such as 'Sac bandoulière' and 'Sacs à main,' which exhibited high likes counts and significant discounts. Items like 'Sacs de voyage' and 'Portefeuilles' also stood out, indicating their strong market presence within the category.

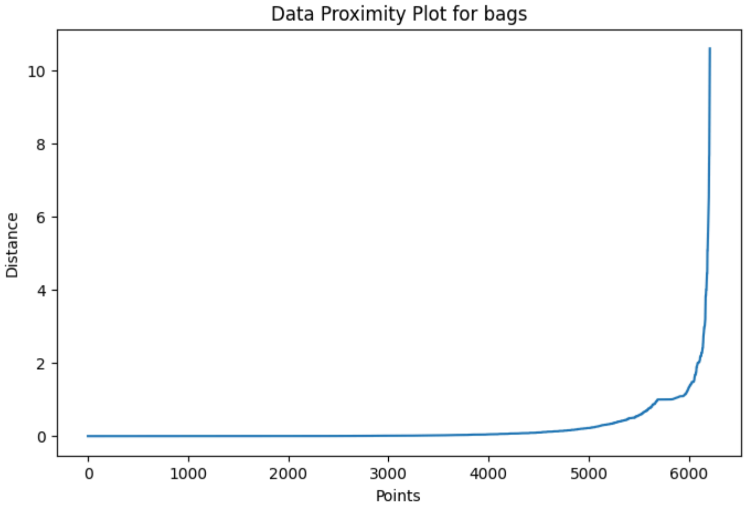


Figure 46: Data Proximity Plot for Bags

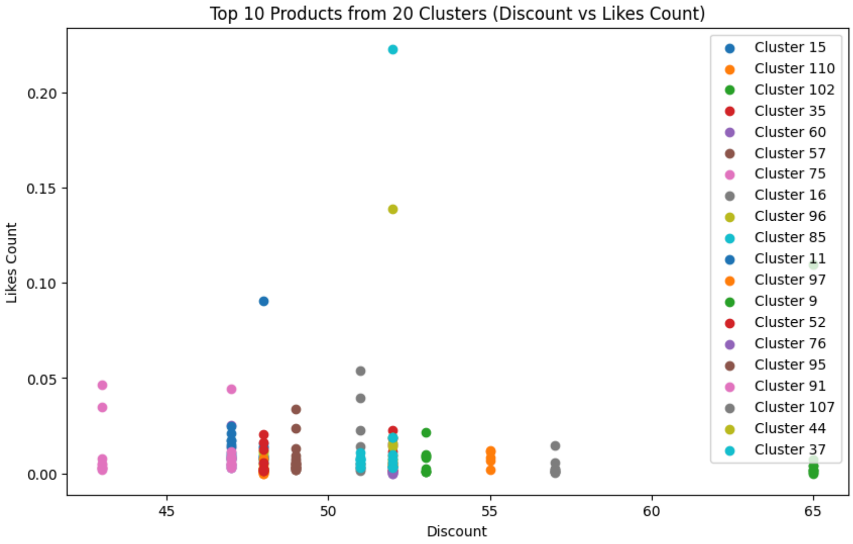


Figure 47: Top 10 Products from 20 Clusters for Bags

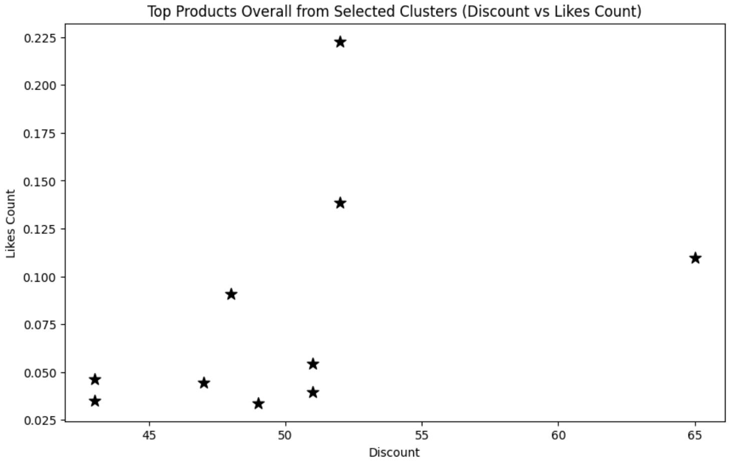


Figure 48: Top Products Overall from Selected Clusters for Bags

- **Kids Category:**

  - *Silhouette score:* -0.0377
  - *Epsilon value:* 0.151
  - *Min Sample value:* 8

  The clustering for the Kids category was conducted using an epsilon value of 0.151 and a min_samples value of 8. The resulting silhouette score of -0.0377 indicated close proximity of clusters. Despite this, the top products identified were 'Brassières de grossesse' and 'Pantalons & Jupes,' which had high engagement. Other notable items included 'Sous-vêtements de maintien de grossesse' and 'Costume & Jupe-culotte,' which also attracted significant attention from users.

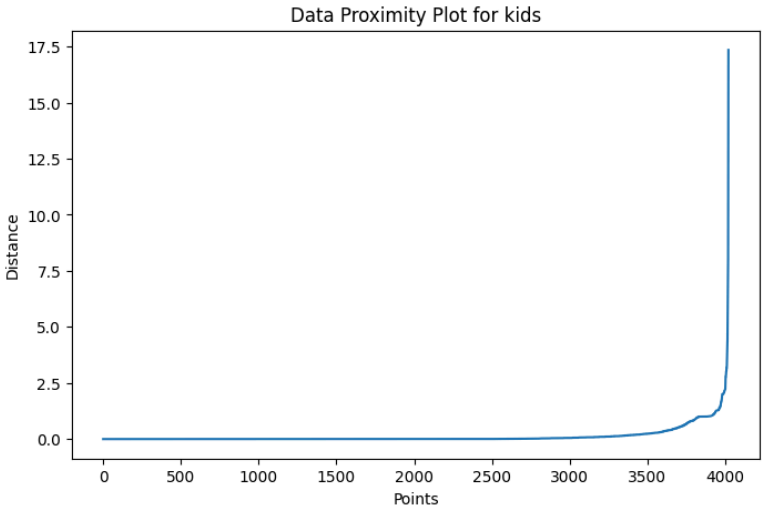


Figure 49: Data Proximity Plot for Kids

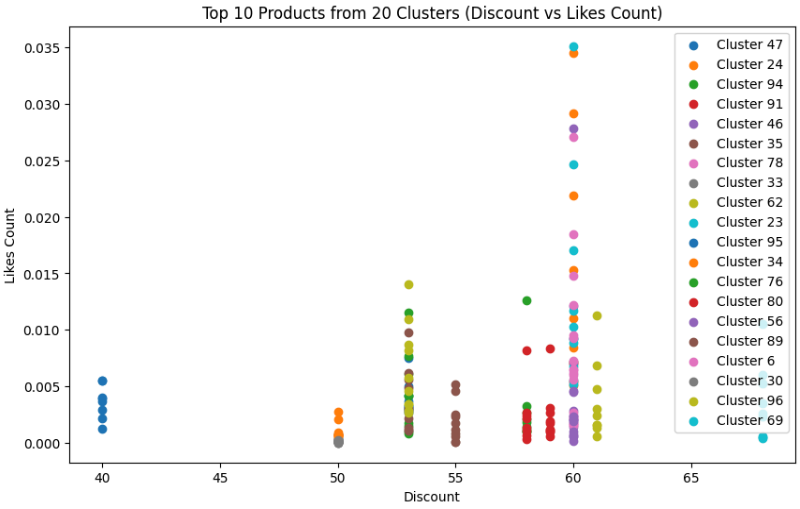


Figure 50: Top 10 Products from 20 Clusters for Kids

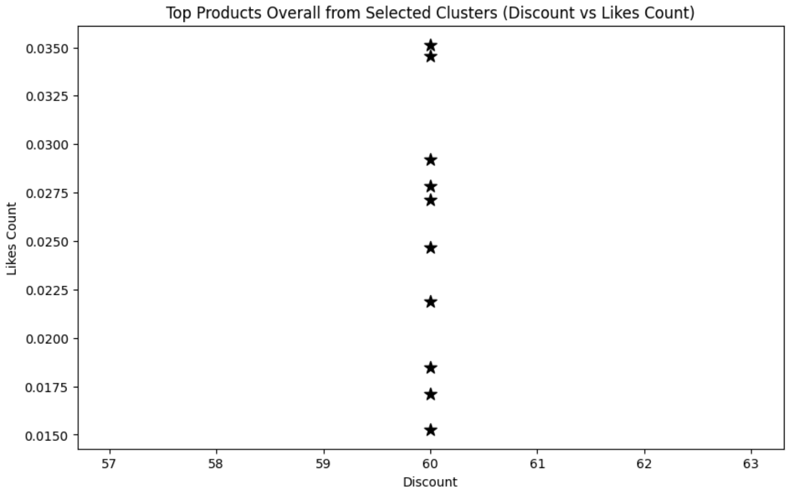


Figure 51: Top Products Overall from Selected Clusters for Kids

- **Beauty Category:**

  - *Silhouette score:* -0.2113
  - *Epsilon value:* 0.576
  - *Min Sample value:* 5

In the Beauty category, DBSCAN was optimized with an epsilon value of 0.576 and a min_samples value of 5. The silhouette score of -0.2113 suggested moderate cluster definition. The clustering process highlighted popular items like 'Gloss à lèvres' and 'Eyeliner,' which received substantial likes. Other key products included 'Soins des pieds,' 'Correcteur,' and 'Rouge à lèvres,' each demonstrating a blend of appealing discounts and high user engagement.

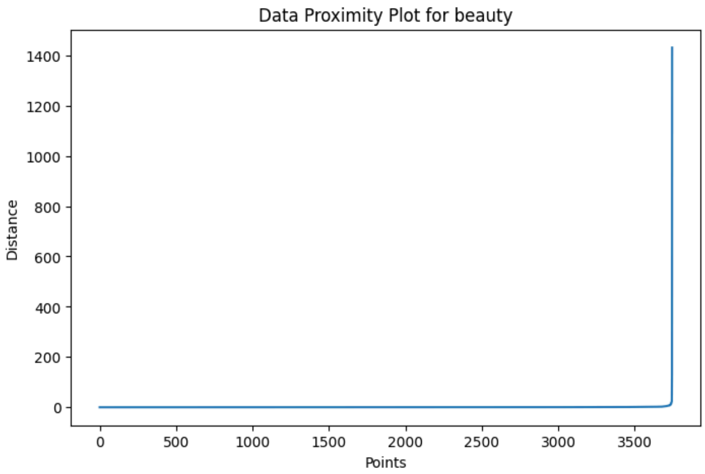


Figure 52: Data Proximity Plot for Beauty

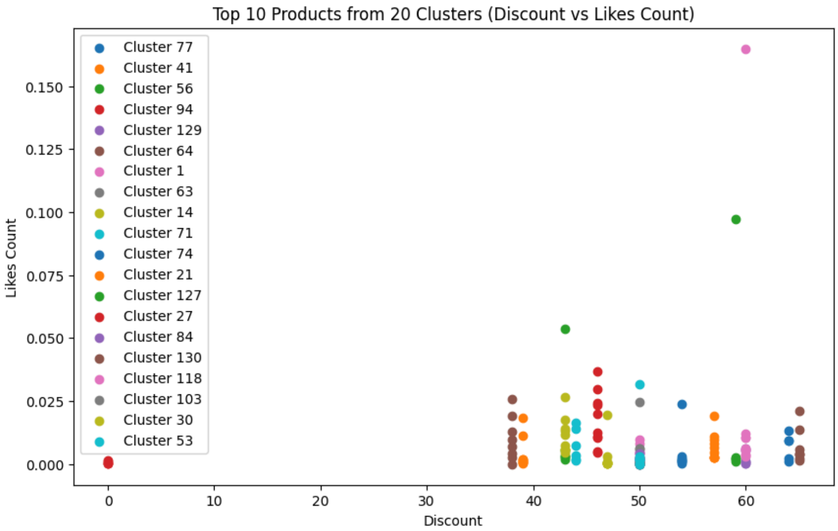


Figure 53: Top 10 Products from 20 Clusters for Beauty

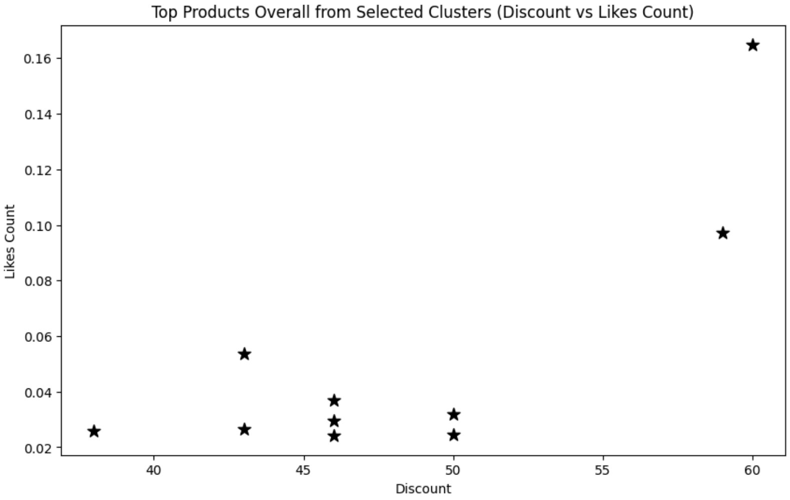


Figure 54: Top Products Overall from Selected Clusters for Beauty

- **Jewellery Category:**

  - *Silhouette score:* -0.0181
  - *Epsilon value:* 0.505
  - *Min Sample value:* 5

For the Jewelry category, the best parameters included an epsilon value of 0.505 and a min_samples value of 5. The silhouette score of -0.0181 indicated very close clustering, with top products like 'Bracelets' and 'Boucles d'oreilles' leading in popularity. Other prominent items in the clusters were 'Colliers' and 'Montres pour femme,' which were well-received, showcasing their appeal within the Jewelry category.

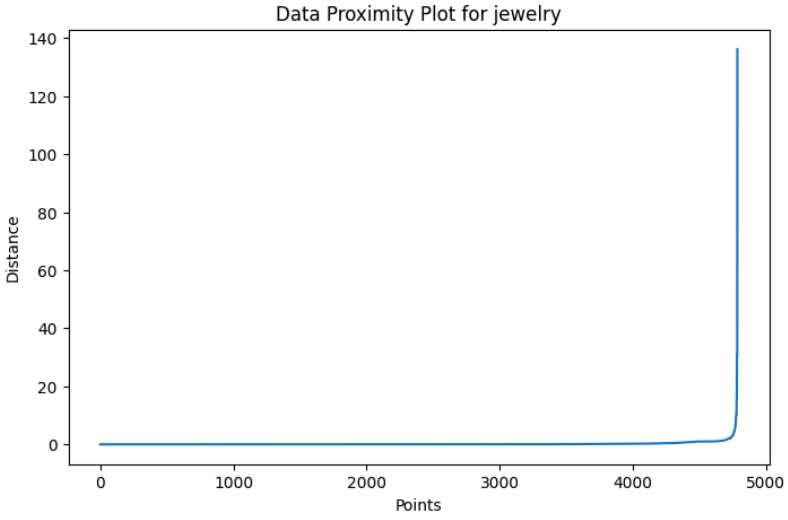


Figure 55: Data Proximity Plot for Jewellery

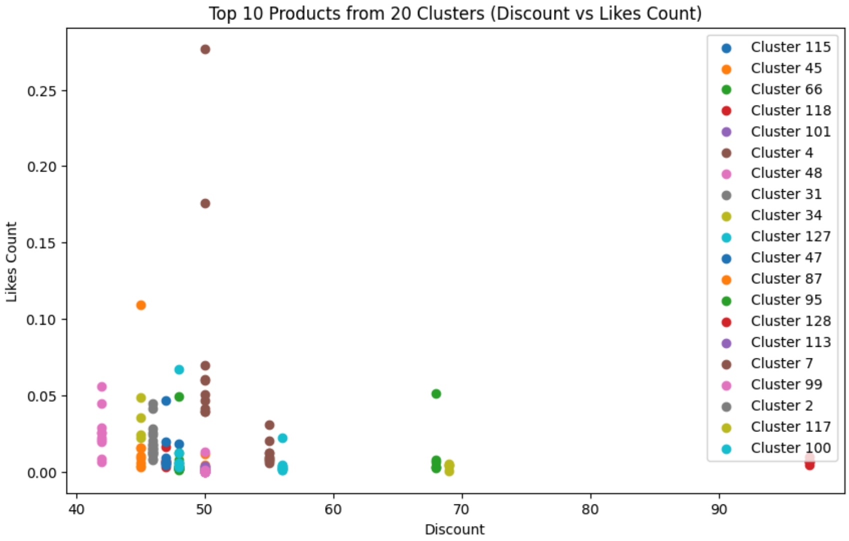


Figure 56: Top 10 Products from 20 Clusters for Jewellery

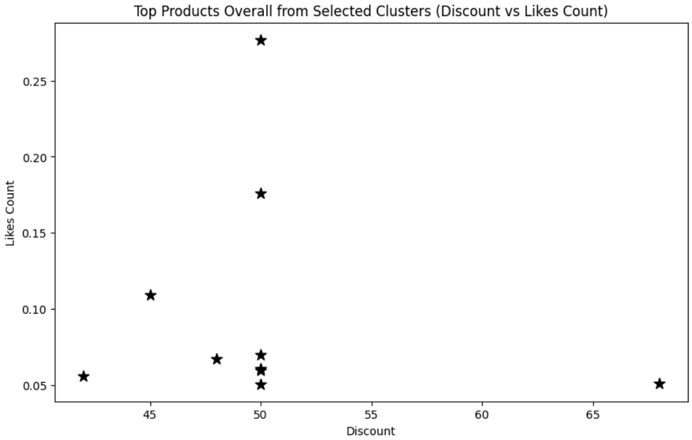


Figure 57: Top Products Overall from Selected Clusters for Jewellery

- **Women Category:**
  - *Silhouette score:* -0.3197
  - *Epsilon value:* 0.788
  - *Min Sample value:* 11

In the Women category, the optimal DBSCAN parameters were an epsilon value of 0.788 and a min_samples value of 11. The silhouette score of -0.3197 reflected some challenges in cluster clarity. Nonetheless, products like 'Soutiens-gorge' and 'Vestes & Gilets' were identified as top items, attracting a high number of likes. Other notable products included 'Chemises,' 'Blouses & Chemises,' and 'Robes imprimées,' each demonstrating strong consumer interest.

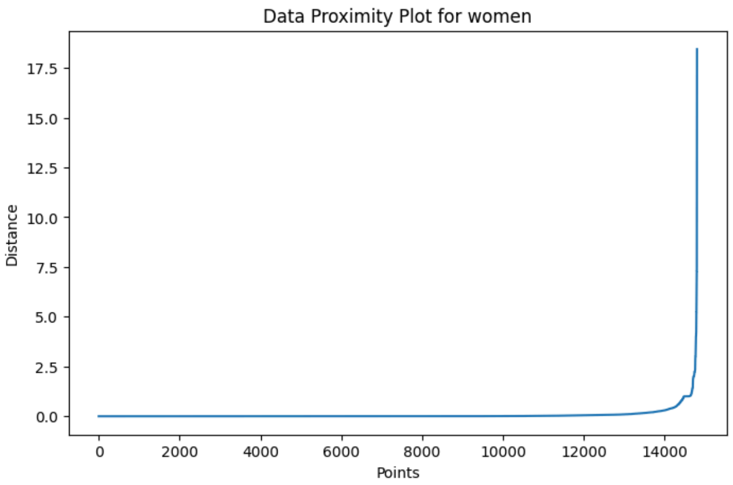


Figure 58: Data Proximity Plot for Women

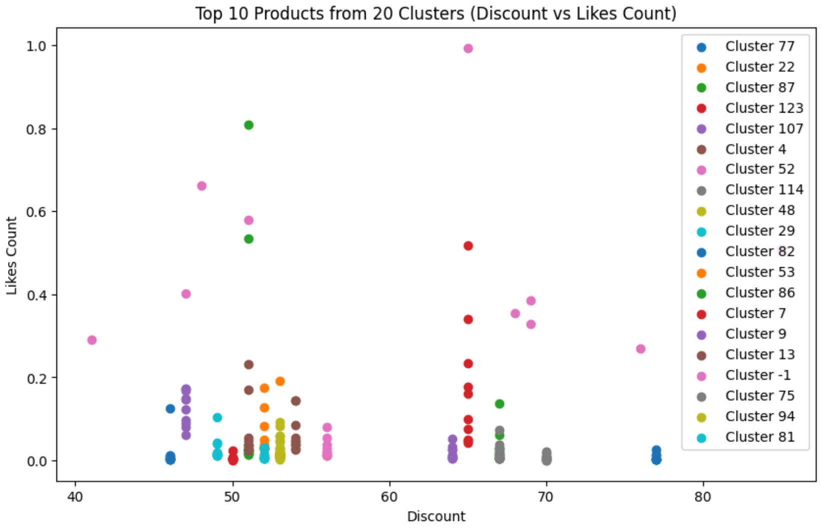


Figure 59: Top 10 Products from 20 Clusters for Women

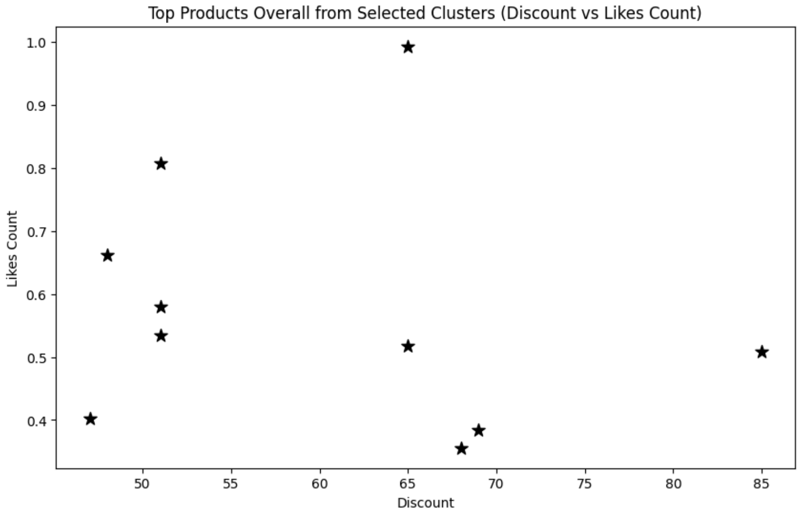


Figure 60: Top Products Overall from Selected Clusters for Women

- **Men Category:**

  - *Silhouette score:* 0.0656
  - *Epsilon value:* 0.505
  - *Min Sample value:* 14

For the Men category, DBSCAN was fine-tuned with an epsilon value of 0.505 and a min_samples value of 14. The silhouette score of 0.0656, although modest, indicated some level of cluster distinction. The analysis highlighted popular items such as 'Shirts' and 'Henley Shirts,' which garnered significant likes. Other top products included 'Pantalons,' 'Hoodies,' and 'Boxers,' each reflecting strong user preference and engagement.

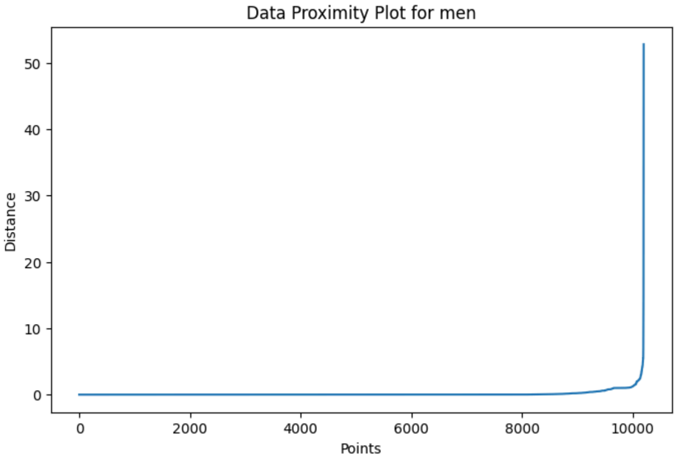


Figure 61: Data Proximity Plot for Men

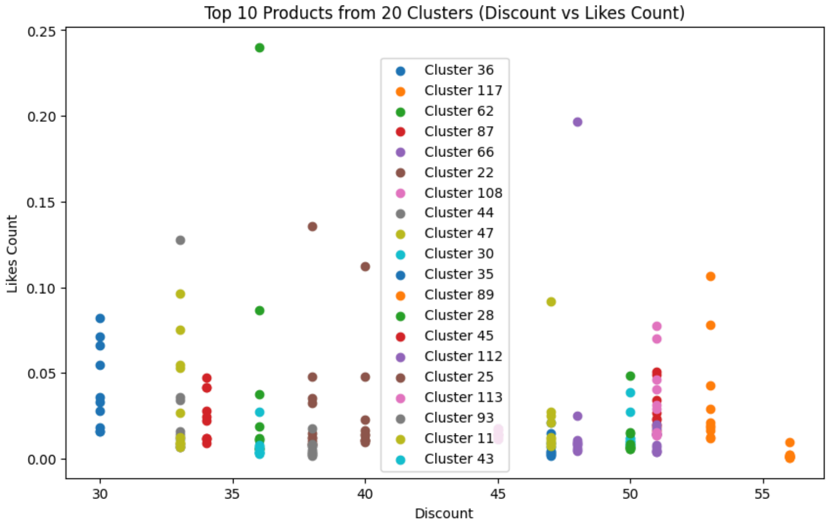


Figure 62: Top 10 Products from 20 Clusters for Men

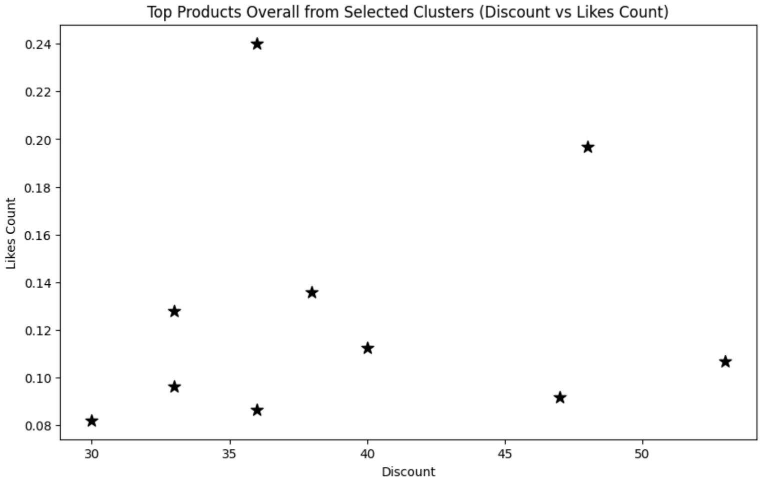


Figure 63: Top Products Overall from Selected Clusters for Men

- **House Category:**

  - *Silhouette score:* -0.1475
  - *Epsilon value:* 1.0
  - *Min Sample value:* 14

In the House category, the DBSCAN algorithm was applied with an epsilon value of 1.0 and a min_samples value of 14. The silhouette score of -0.1475 suggested challenges in cluster separation. Nevertheless, the clustering identified popular products like 'Flowers' and 'Étuis & Crochets' as top items, with others like 'Grasses,' 'Flatware & Utensil Storage,' and 'Gloves' also ranking high in user engagement.

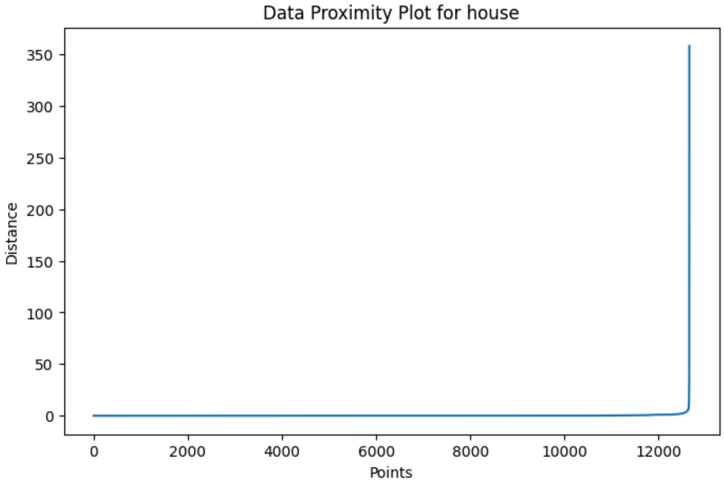


Figure 64: Data Proximity Plot for House

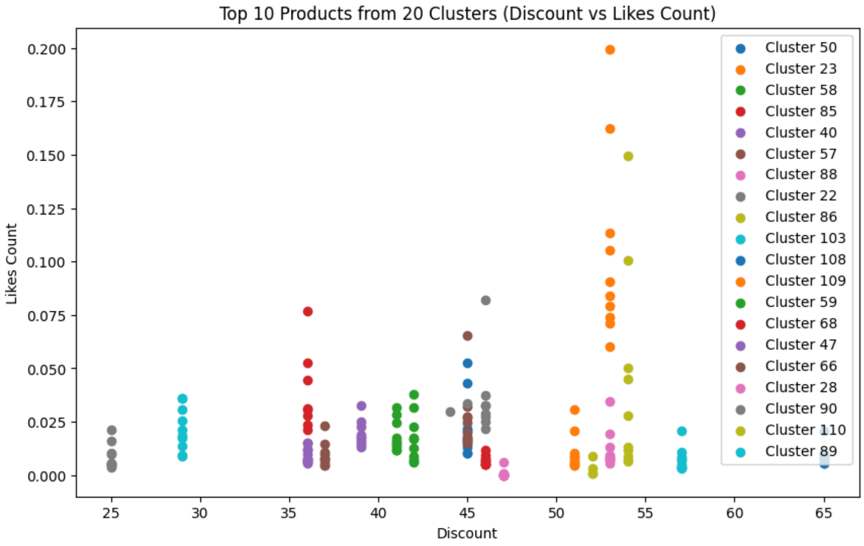


Figure 65: Top 10 Products from 20 Clusters for House

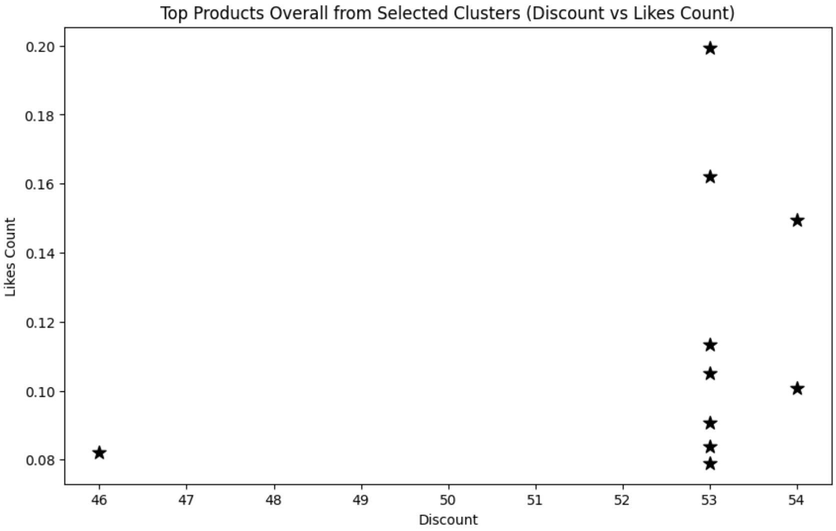


Figure 66: Top Products Overall from Selected Clusters for House

- **Shoes Category:**
  - *Silhouette score:* -0.4035
  - *Epsilon value:* 0.364
  - *Min Sample value:* 14

For the Shoes category, the best DBSCAN parameters were found with an epsilon value of 0.364 and a min_samples value of 14. Despite a silhouette score of -0.4035 indicating significant overlap among clusters, the analysis revealed that 'Derbies & Mocassins' was the most popular subcategory, followed by 'Bottes & Bottines' and 'Sandales & Mules.' These items showed a strong balance between discount rates and likes count, highlighting their appeal in the Shoes category.

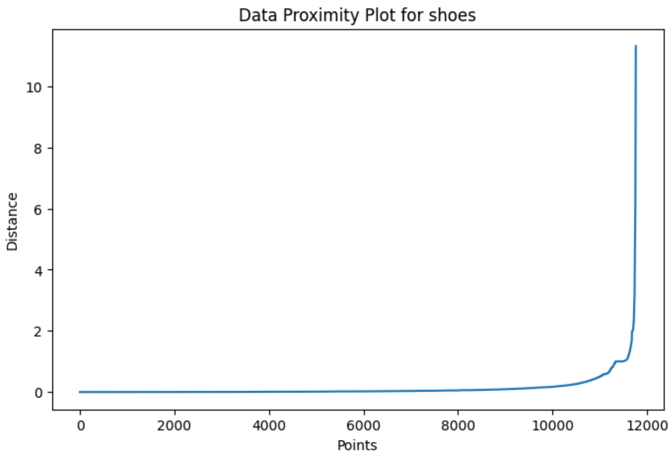


Figure 67: Data Proximity Plot for Shoes

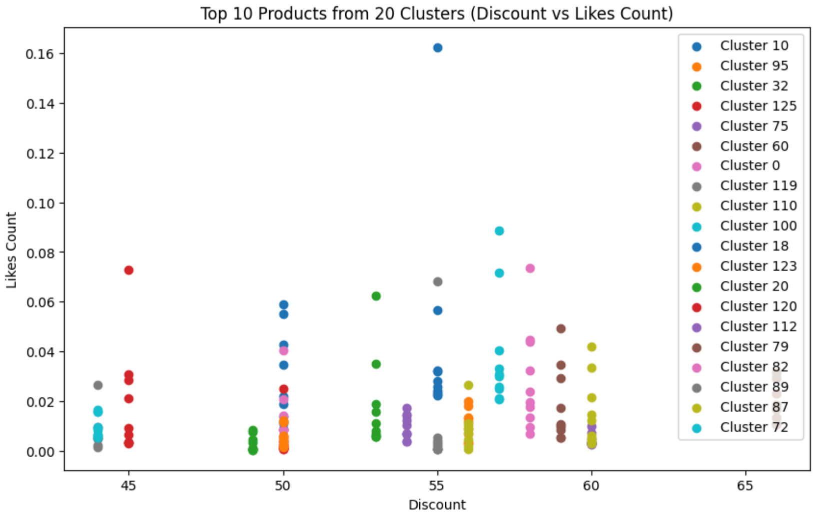


Figure 68: Top 10 Products from 20 Clusters for Shoes

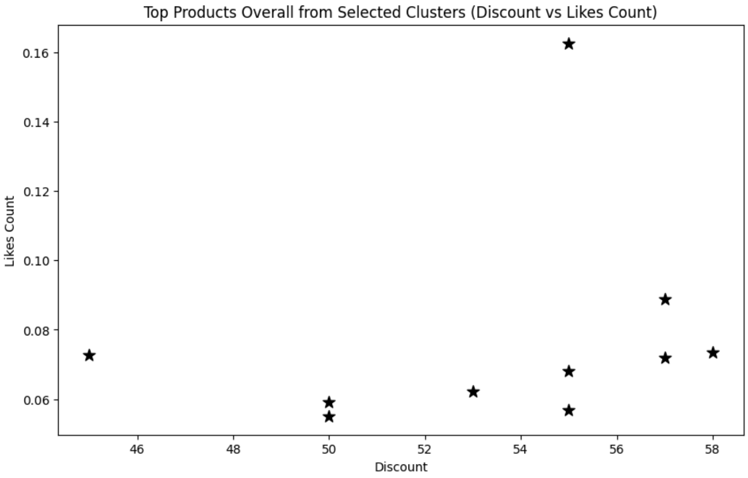


Figure 69: Top Products Overall from Selected Clusters for Shoes

Top 10 Products Overall Across All Categories

After analyzing the top 10 products from each selected cluster, the top 10 products overall across all categories were identified, revealing a strong dominance of items from the Women and Men categories. Leading the list is the 'Blouses & Chemises' subcategory from the Women category, with an impressive 65% discount and a standout likes count of 0.993. Other top performers include 'Soutiens-gorge,' 'Brassières de sport,' and 'Vestes & Gilets,' all from the Women category, reflecting high consumer interest through substantial likes counts and attractive discount rates. The Men category also showed strong contenders, with 'Henley Shirts' and 'Pantalons' achieving notable likes counts and competitive discounts. 'Henley Shirts' were particularly popular, appearing twice in the top 10, underscoring their appeal. These products exemplify a balance between appealing discounts and high user engagement, making them stand out in their respective categories.

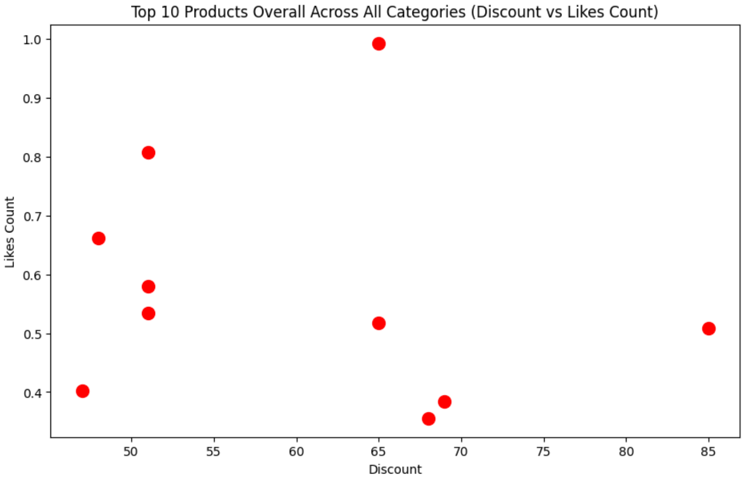


Figure 70: Top 10 Products Overall Across All Categories

Category Distribution

The distribution of these top products across categories was visualized to understand clustering patterns and any potential biases, highlighting the model's effectiveness in pinpointing prominent products within each category.

Visualizing Top Products

To provide a detailed view of the most popular products within specific categories, scatter plots and bar plots were generated. These visualizations are crucial for understanding user preferences and can significantly inform marketing strategies and inventory decisions.

Conclusion

The DBSCAN algorithm successfully segmented the product data into distinct clusters without needing a predefined number of clusters, demonstrating its capability to handle noise and uncover meaningful patterns. The analysis produced well-defined clusters, pinpointing significant products within each category based on user engagement and discount levels. The identification of the top 10 products overall across all categories offered valuable insights into the most popular items, emphasizing the prominence of certain subcategories. While the outcomes are promising, further enhancements, such as exploring additional features or utilizing advanced algorithms, could improve clustering accuracy and provide deeper insights.

3.3. Comparison:

Best Clustering Algorithm: K-Means

When comparing the clustering performance between **K-Means** and **DBSCAN**, **K-Means** emerges as the superior algorithm. This conclusion is based on an analysis of the silhouette scores across various product categories (Rousseeuw, 1987).

Table 2 demonstrates the Silhouette score for each category for both clustering techniques, whereas Figures 71 and 72 illustrate those scores in the form of a bar chart. The displayed scores in the table and the figures are explained in more detail below.

Category	K-Means Silhouette Score	DBSCAN Silhouette Score
Men	0.33	0.0656
Women	0.31	-0.3197
Bags	0.19	-0.2913
Kids	0.27	-0.0377
Beauty	0.16	-0.2113

House	0.21	-0.1475
Jewellery	0.29	-0.0181
Accessories	0.47	-0.1161
Shoes	0.15	-0.4035

Table 2: Silhouette Score Comparison Between the Two Clustering Techniques

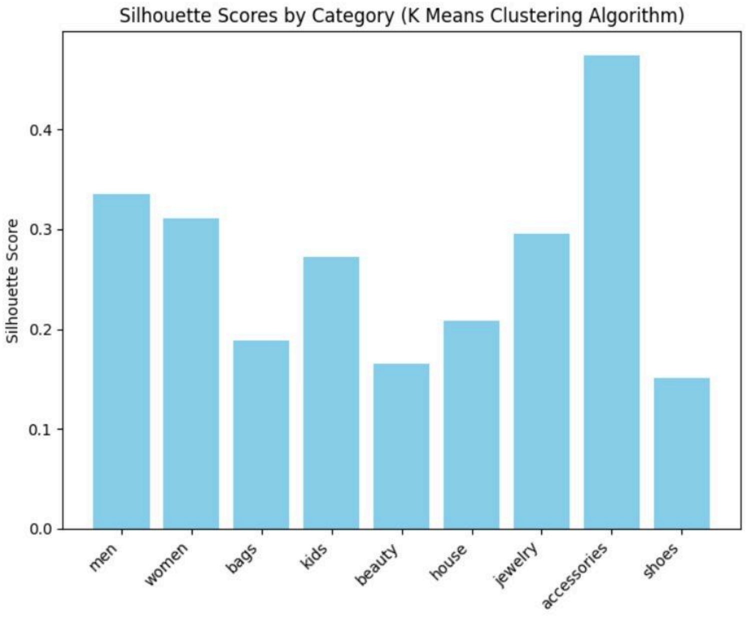


Figure 71: Silhouette Scores by Category for K-Means Clustering

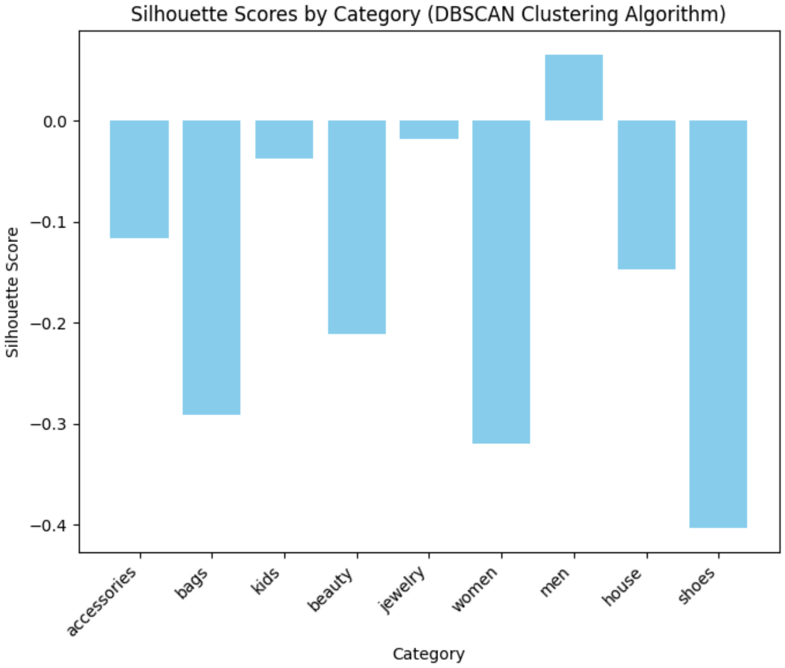


Figure 72: Silhouette Scores by Category for DBSCANS Clustering

K-Means Clustering

- **Silhouette Scores**: The silhouette scores for K-Means range from approximately 0.1 to 0.45 with higher scores generally indicating better-defined clusters.

- **Categories with High Scores**:
  - **Accessories** and **Shoes** stand out with the highest silhouette scores, with Accessories scoring above 0.4.
  - **Men** and **Women** categories also show relatively high silhouette scores, around 0.3 (MacQueen, 1967).

- **Categories with Lower Scores**:
  - **Bags** and **Beauty** have the lowest scores, ranging from 0.1 to 0.2, indicating less well-defined clusters compared to the other categories.

DBSCAN Clustering

- **Silhouette Scores**: The silhouette scores for DBSCAN mostly range from negative to zero, indicating poor clustering performance or overlapping clusters.

- **Categories with Positive Scores**:
  - **Men** and **House** categories have slightly positive scores, just above 0.0, suggesting some clustering quality but not as robust as K-Means (Ester et al., 1996).

- **Categories with Negative Scores**:
  - Most categories, such as **Bags**, **Kids**, **Jewelry**, and **Shoes**, have negative scores. The **Shoes** category, in particular, has the lowest score around -0.4, reflecting poorly defined clusters.

Comparison

- **Overall Clustering Quality**:
  - **K-Means** demonstrates superior clustering quality, with most categories showing positive silhouette scores, especially Accessories

and Shoes. This indicates that K-Means successfully forms distinct and well-defined clusters.

  - **DBSCAN** performs worse overall, with many categories having negative silhouette scores, indicating poorly defined clusters or a significant amount of noise in the data. (Xu and Wunsch, 2005)

- **Consistency Across Categories**:
  - Categories like **Men** perform relatively well under both algorithms, but with much better results in K-Means.
  - **Shoes** and **Bags** perform poorly in DBSCAN but show reasonable clustering quality in K-Means.

Reasons for K-Means Being the Best

1. **Positive Silhouette Scores**:
   - The majority of categories in the K-Means clustering graph have positive silhouette scores, ranging from 0.1 to 0.45. These positive scores indicate well-separated clusters where data points are grouped appropriately within their clusters.
2. **High Clustering Quality**:
   - Categories like **Accessories** (above 0.4) and **Shoes** (around 0.45) exhibit high silhouette scores in K-Means, reflecting well-defined and distinct clusters.
3. **Consistency**:
   - K-Means demonstrates consistent clustering quality across different categories, with few instances of very low or negative silhouette scores. This consistency underscores K-Means' ability to reliably identify meaningful clusters.

Expanded Comparison Between K-Means and DBSCAN

1) **Silhouette Scores:**
    - *K-Means:*
        - Scores range from around 0.1 to 0.45, with positive values indicating well-formed clusters. The algorithm performs well across most categories, with **Accessories** and **Shoes** showing the highest silhouette scores, indicating strong intra-cluster cohesion and inter-cluster separation.
    - *DBSCAN:*
        - The silhouette scores mostly range from negative to zero, with many categories having negative scores. **Shoes** (-0.4) and **Bags** (around -0.3) are particularly poorly clustered. Negative scores suggest that the clusters may be poorly defined, with data points closer to points in other clusters than to points within their own cluster (Xu and Wunsch, 2005).

2) **Clustering Behavior:**
    - *K-Means:*
        - Assumes clusters are spherical and evenly sized, which works well when clusters are relatively uniform and well-separated, as seen in the dataset. K-Means excels with datasets that have a clear separation between clusters, resulting in better-defined clusters and higher silhouette scores.
    - *DBSCAN:*
        - Designed to find arbitrarily shaped clusters and effectively identify noise (points that do not belong to any cluster). However, DBSCAN struggles with datasets where clusters overlap or are not well-defined, as indicated by the negative silhouette scores. DBSCAN is also more sensitive to its parameters (e.g., epsilon and minimum points), which might not be well-tuned for this particular dataset (Rodriguez et al., 2019).

3) **Sensitivity to Noise and Outliers:**

- *K-Means:*
    - ○ Less robust to noise and outliers, as it assigns every point to a cluster, potentially resulting in misleading results if there is significant noise. Despite this, K-Means managed to achieve better clustering in this analysis, indicating that the dataset might not have excessive noise or that K-Means effectively grouped relevant points (Rodriguez et al., 2019).
- *DBSCAN:*
    - ○ More robust to noise and outliers, as it can identify and ignore them, labelling them as noise instead of forcing them into a cluster. In this case, the high number of negative silhouette scores suggests that DBSCAN may have identified too many points as noise or failed to properly separate the clusters due to parameter sensitivity (Xu and Wunsch, 2005).

4) **Use Case Suitability:**

- *K-Means:*
    - ○ Best suited for datasets with well-defined, convex clusters, as seen with categories like **Accessories** and **Shoes**. K-Means provides clear, meaningful clusters, making it the preferred algorithm for this dataset (Xu and Wunsch, 2005).
- *DBSCAN:*
    - ○ Better suited for datasets with varying cluster shapes, uneven distribution, or significant noise. However, in this case, DBSCAN did not perform as well, likely due to parameter sensitivity and the nature of the dataset not being ideal for DBSCAN.

Conclusion:

K-Means is the recommended clustering method for this dataset due to its ability to form distinct and well-separated clusters, as indicated by higher silhouette scores across most categories. DBSCAN, while useful in specific scenarios, did not perform well with this dataset, likely due to its sensitivity to parameters and the nature of the data, leading to poorly defined clusters with negative silhouette scores. Therefore, K-Means is the preferred clustering algorithm in this comparison (Rodriguez et al., 2019).

4. Classification:

4.1. Classification Algorithm 1 - Random Forest Classifier:

4.1.1 Model Planning

- **Objective:** The primary objective of this analysis is to classify products into predefined categories based on features such as discount, likes_count, current_price, and raw_price. The categories of interest include accessories, bags, beauty, jewellery, kids, men, women, house, and shoes. The goal is to build a predictive model using the Random Forest algorithm that can accurately assign a category to each product based on these features (Breiman, 2001).
- **Data Understanding:** The dataset used for this analysis comprises various product attributes, with the category being the target variable. The key features identified for model building are:
  - **discount**: Percentage discount on the product.
  - **likes_count**: Number of likes or user engagement with the product.
  - **current_price**: The current selling price of the product.
  - **raw_price**: The original price of the product before any discounts (Liaw and Wiener, 2011).

The selection of the columns discount, likes_count, current_price, and raw_price was driven by their relevance to understanding customer behaviour and product performance in the context of an e-commerce setting. Here's the reasoning behind selecting each of these features:

1. Discount:

- **Why Selected:** Discounts are a major driving factor in customer decision-making when it comes to online shopping. Offering a product at a lower price than its original cost often attracts customers and increases the likelihood of engagement and purchase.

- **Significance:** It is crucial to understand how price reductions affect product popularity across different categories. This feature helps in identifying which categories or products benefit most from discounts.

2. Likes Count:

- **Why Selected:** Likes count measures user engagement and interest in a product, serving as a proxy for popularity. Higher engagement often correlates with customer satisfaction or at least interest, which can translate into purchases.
- **Significance:** Including this feature helps the model capture the social proof element, which is important in e-commerce platforms where products with more likes are more likely to be noticed by other customers.

3. Current Price:

- **Why Selected:** The current price is one of the most important factors influencing whether a customer will buy a product. It directly impacts sales conversion and consumer choices.
- **Significance:** Analyzing the current price allows the model to understand how pricing strategies (in conjunction with discounts) influence customer behaviour and help differentiate products within the same category.

4. Raw Price:

- **Why Selected:** The raw price provides context for the discount and the pricing history of a product. It indicates the original valuation and helps to assess the perceived value customers might derive from the discount offered.
- **Significance:** Knowing the raw price helps in understanding the magnitude of the discount being offered and allows for better insights into how much price reduction contributes to driving engagement and sales.

Overall Justification:

These features are closely tied to customer behaviour on e-commerce platforms. They offer a balance of price-related data (e.g., current_price, raw_price, discount) and behavioural data (e.g., likes_count). Together, they provide a comprehensive view of both the product's financial attributes and how customers interact with it, which is essential for building a predictive model that can classify products effectively across different categories.

Preprocessing Strategy:

- Filter the dataset to include only relevant categories.
- Encode the target variable (category) into numerical labels for compatibility with machine learning algorithms.
- Split the dataset into training and testing sets to evaluate the model's performance (James et al., 2013).

Model Selection:

The Random Forest Classifier was selected due to its robustness in handling large datasets and its ability to model complex interactions between features. It also provides feature importance metrics, which can be valuable in understanding the influence of each feature on the prediction (Breiman, 2001).

4.1.2 Model Building

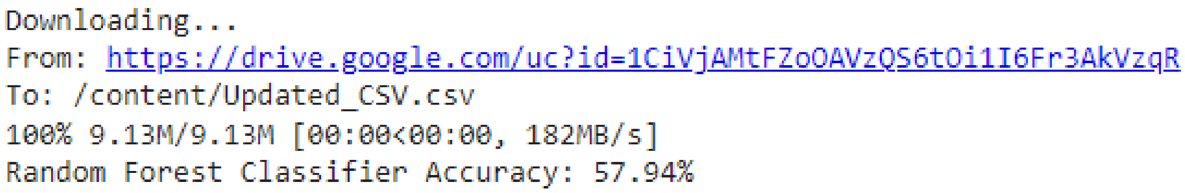
Data Preparation:

- The dataset was first filtered to include only the selected categories.
- The target variable category was encoded into numerical values using Label Encoding, where each category was assigned a unique integer.
- The feature set was defined to include discount, likes_count, current_price, and raw_price (Pedregosa et al., 2011).
- The dataset was then split into training (70%) and testing (30%) sets using the train_test_split function from the sklearn.model_selection module (Pedregosa et al., 2011).

Model Training:

- A Random Forest Classifier with 100 trees (estimators) was instantiated. The model was configured with a random seed of 42 to ensure reproducibility.
- The model was trained on the training data (X_train, y_train) using the fit method, where it learned to map the feature values to the corresponding product categories (Breiman, 2001).

Model Evaluation:



```
Downloading...
From: https://drive.google.com/uc?id=1CiVjAMtFZoOAVzQS6tOi1I6Fr3AkVzqR
To: /content/Updated_CSV.csv
100% 9.13M/9.13M [00:00<00:00, 182MB/s]
Random Forest Classifier Accuracy: 57.94%
```

Figure 73: Accuracy of Random Forest

- The model's performance was evaluated using the test set (X_test, y_test). Predictions were made on the test set using the predict method.
- The model achieved an **accuracy of 57.94%**. While this accuracy suggests that the model captures some of the underlying patterns in the data, there is potential for improvement (James et al., 2013).

Feature Importance:

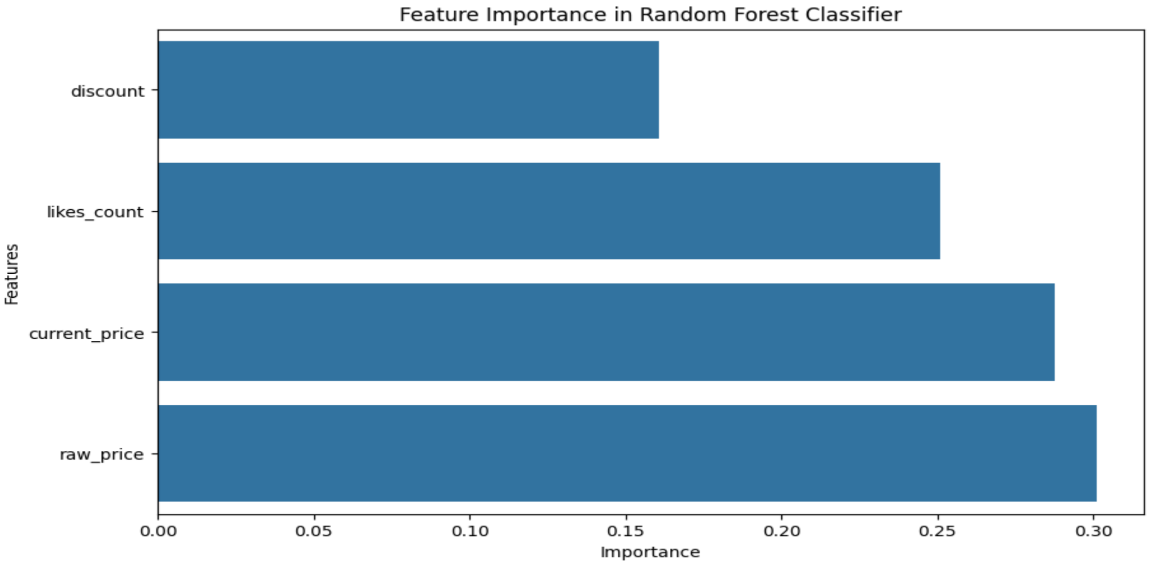


Figure 74: Feature Importance in Random Forest

- The Random Forest model provides a measure of the importance of each feature in making predictions. The feature importance scores were as follows:
  - **likes_count**: Highest importance, indicating that user engagement is a critical factor in determining the product category.
  - **current_price**: Also highly important, suggesting that the pricing of products is a key differentiator among categories.
  - **discount** and **raw_price**: While less important than the other two features, these still contribute to the model's decision-making process (Liaw and Wiener, 2011).
- A bar plot of feature importances was generated, visually depicting the contribution of each feature to the model.

Confusion Matrix:

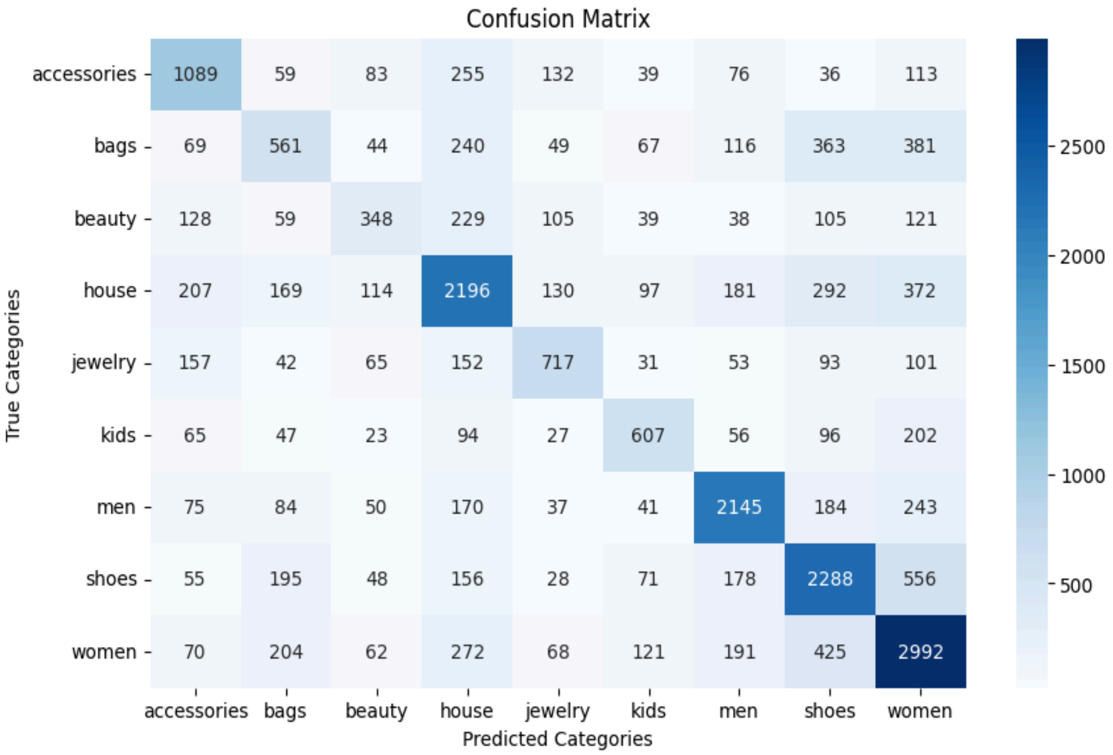
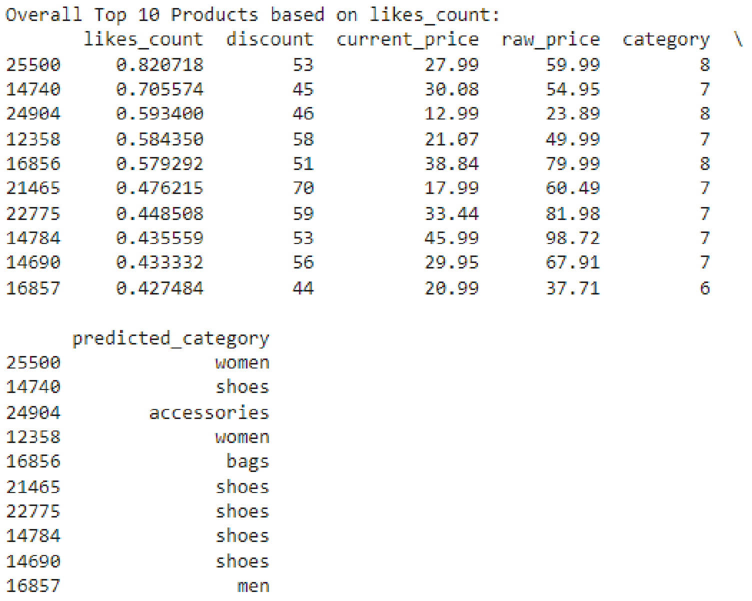


Figure 75: Confusion Matrix of Random Forest

- A confusion matrix was plotted to provide a detailed breakdown of the model's performance across different categories. This matrix shows the true versus predicted categories, highlighting where the model tends to make errors (Pedregosa et al., 2011).

4.1.3 Result Communication

Overall Top 10 Products by all Categories:



```
Overall Top 10 Products based on likes_count:
       likes_count  discount  current_price  raw_price  category  \
25500     0.820718        53          27.99      59.99         8
14740     0.705574        45          30.08      54.95         7
24904     0.593400        46          12.99      23.89         8
12358     0.584350        58          21.07      49.99         7
16856     0.579292        51          38.84      79.99         8
21465     0.476215        70          17.99      60.49         7
22775     0.448508        59          33.44      81.98         7
14784     0.435559        53          45.99      98.72         7
14690     0.433332        56          29.95      67.91         7
16857     0.427484        44          20.99      37.71         6

      predicted_category
25500              women
14740              shoes
24904        accessories
12358              women
16856               bags
21465              shoes
22775              shoes
14784              shoes
14690              shoes
16857                men
```

Figure 76: Random Forest - Overall Top 10 Products based on likes_count

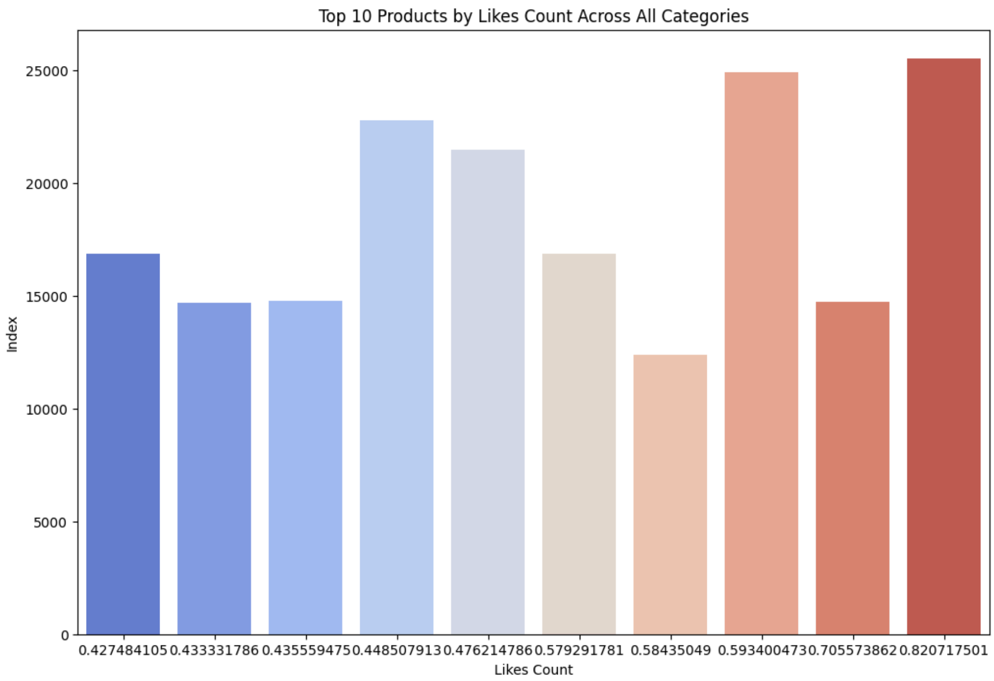


Figure 77: Random Forest - Top 10 Products by likes_count Across All Categories

"Top 10 Products by Likes Count Across All Categories." The x-axis represents the likes count (appearing as floating-point numbers), while the y-axis shows the index or ranking of the products based on their likes count.

Analysis:

1. **Top-Ranked Products:**
    - The product with the highest likes count (0.820717501) belongs to the "women" category.
    - The second-highest likes count (0.705573862) corresponds to a product in the "shoes" category.
    - A product in the "accessories" category also appears within the top 10, though with a relatively lower likes count (0.593400473).
2. **Categories:**
    - Women: This category seems to dominate, with the highest likes count and another product also making it into the top 10.
    - Shoes: This category is well-represented, appearing multiple times across different rankings.
    - Accessories, Bags, and Men: Each of these categories has at least one product within the top 10, indicating their popularity but to a lesser extent.
3. **Discount and Pricing:**
    - Products in the top 10 have a varied discount range from 44% to 70%, suggesting that substantial discounts could be a factor in the number of likes.
    - Current prices for these products range from $12.99 to $45.99, indicating a broad spectrum of affordability among popular items.

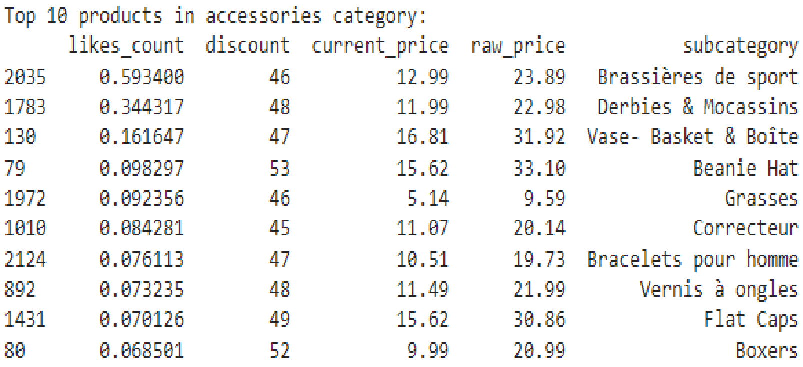
Key Insights:

- Popularity and Pricing: There's a potential correlation between higher discounts and likes count, as the top products all feature significant discounts. This might imply that consumers are attracted to deals, leading to higher engagement.
- Category Preferences: Categories like "women" and "shoes" seem to be particularly popular, as evidenced by the number of entries in the top 10.
- Likes Distribution: The distribution of likes shows a clear preference towards certain products, possibly driven by a combination of factors like price, discount, and category.

This analysis can contribute to understanding consumer behavior and preferences, which can be useful for marketing strategies or inventory planning in related sectors.

Top Products by Category:

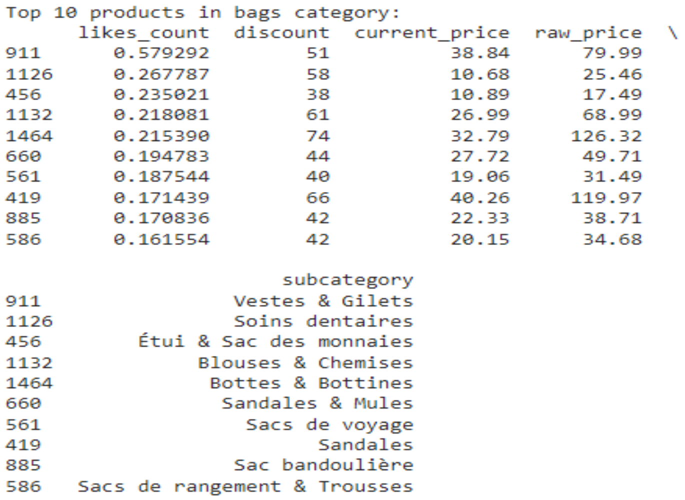
- For each of the nine categories, the top 10 products were identified based on the predicted categories and sorted by likes_count. This approach provides insights into which products within each category are most popular, as measured by user engagement. Below is a summary of findings:
  - **Accessories**: Products with moderate discounts and likes were identified as top products, indicating a balanced appeal between affordability and engagement.



```
Top 10 products in accessories category:
      likes_count  discount  current_price  raw_price           subcategory
2035     0.593400        46          12.99      23.89    Brassières de sport
1783     0.344317        48          11.99      22.98    Derbies & Mocassins
130      0.161647        47          16.81      31.92  Vase- Basket & Boîte
79       0.098297        53          15.62      33.10            Beanie Hat
1972     0.092356        46           5.14       9.59               Grasses
1010     0.084281        45          11.07      20.14            Correcteur
2124     0.076113        47          10.51      19.73   Bracelets pour homme
892      0.073235        48          11.49      21.99        Vernis à ongles
1431     0.070126        49          15.62      30.86             Flat Caps
80       0.068501        52           9.99      20.99                Boxers
```

Figure 78: Random Forest - Top 10 Products in Accessories

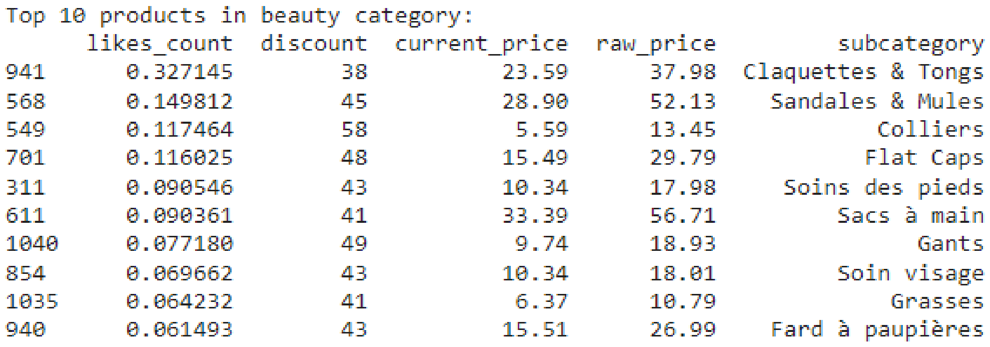
  - **Bags**: The top products had significant variability in likes_count and discount percentages, suggesting that both luxury and budget-friendly options are popular.



```
Top 10 products in bags category:
      likes_count  discount  current_price  raw_price  \
911      0.579292        51          38.84      79.99
1126     0.267787        58          10.68      25.46
456      0.235021        38          10.89      17.49
1132     0.218081        61          26.99      68.99
1464     0.215390        74          32.79     126.32
660      0.194783        44          27.72      49.71
561      0.187544        40          19.06      31.49
419      0.171439        66          40.26     119.97
885      0.170836        42          22.33      38.71
586      0.161554        42          20.15      34.68

                       subcategory
911               Vestes & Gilets
1126             Soins dentaires
456       Étui & Sac des monnaies
1132           Blouses & Chemises
1464            Bottes & Bottines
660             Sandales & Mules
561               Sacs de voyage
419                     Sandales
885              Sac bandoulière
586   Sacs de rangement & Trousses
```

Figure 79: Random Forest - Top 10 Products in Bags

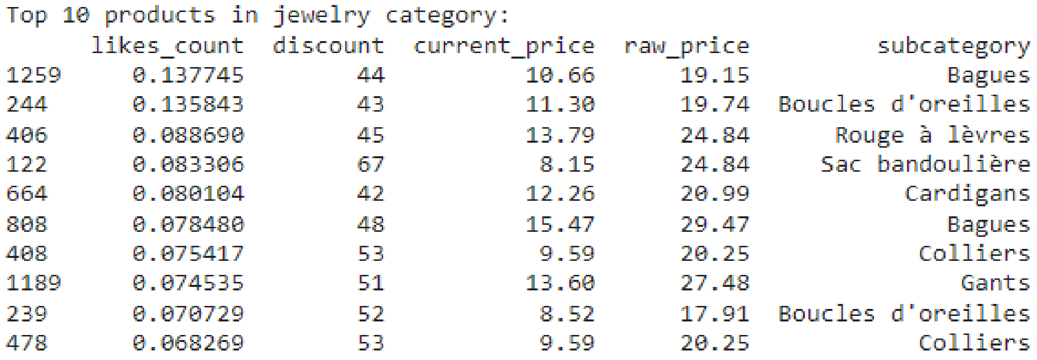
- **Beauty**: The category featured products with moderate engagement, often with discounts around 40-50%.



```
Top 10 products in beauty category:
      likes_count  discount  current_price  raw_price          subcategory
941      0.327145        38          23.59      37.98  Claquettes & Tongs
568      0.149812        45          28.90      52.13    Sandales & Mules
549      0.117464        58           5.59      13.45             Colliers
701      0.116025        48          15.49      29.79            Flat Caps
311      0.090546        43          10.34      17.98     Soins des pieds
611      0.090361        41          33.39      56.71          Sacs à main
1040     0.077180        49           9.74      18.93                Gants
854      0.069662        43          10.34      18.01          Soin visage
1035     0.064232        41           6.37      10.79              Grasses
940      0.061493        43          15.51      26.99     Fard à paupières
```

Figure 80: Random Forest - Top 10 Products in Beauty

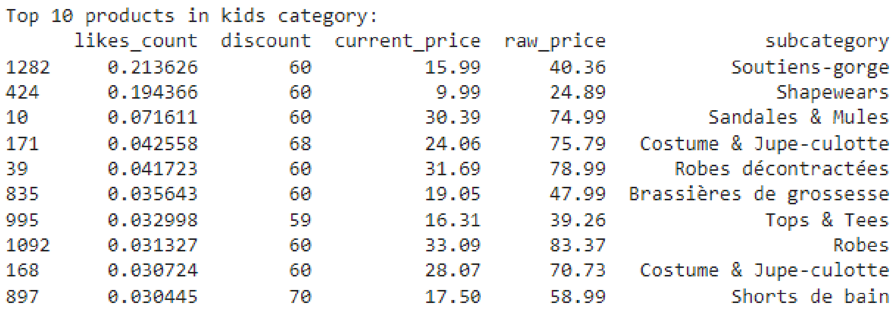
- **Jewellery**: Lower likes_count values in this category indicate that these products may have niche appeal or that the model struggles with accurate classification here.



```
Top 10 products in jewelry category:
      likes_count  discount  current_price  raw_price          subcategory
1259     0.137745        44          10.66      19.15               Bagues
244      0.135843        43          11.30      19.74    Boucles d'oreilles
406      0.088690        45          13.79      24.84         Rouge à lèvres
122      0.083306        67           8.15      24.84       Sac bandoulière
664      0.080104        42          12.26      20.99            Cardigans
808      0.078480        48          15.47      29.47               Bagues
408      0.075417        53           9.59      20.25             Colliers
1189     0.074535        51          13.60      27.48                Gants
239      0.070729        52           8.52      17.91    Boucles d'oreilles
478      0.068269        53           9.59      20.25             Colliers
```

Figure 81: Random Forest - Top 10 Products in Jewellery

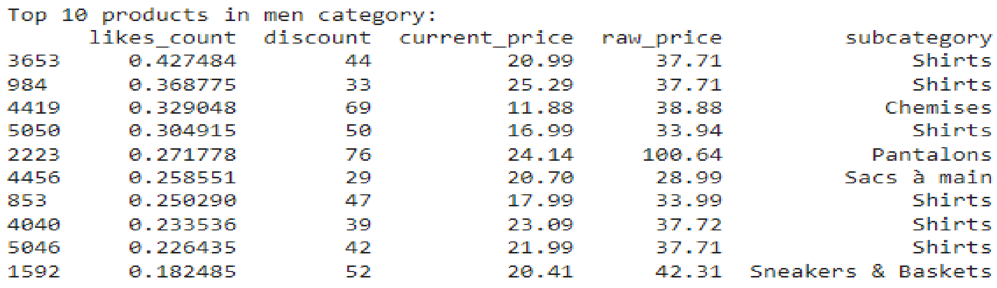
- **Kids**: Products in this category had moderate to low likes_count, with consistent high discounts.



```
Top 10 products in kids category:
      likes_count  discount  current_price  raw_price               subcategory
1282     0.213626        60          15.99      40.36            Soutiens-gorge
424      0.194366        60           9.99      24.89                Shapewears
10       0.071611        60          30.39      74.99          Sandales & Mules
171      0.042558        68          24.06      75.79   Costume & Jupe-culotte
39       0.041723        60          31.69      78.99        Robes décontractées
835      0.035643        60          19.05      47.99   Brassières de grossesse
995      0.032998        59          16.31      39.26               Tops & Tees
1092     0.031327        60          33.09      83.37                     Robes
168      0.030724        60          28.07      70.73   Costume & Jupe-culotte
897      0.030445        70          17.50      58.99           Shorts de bain
```

Figure 82: Random Forest - Top 10 Products in Kids

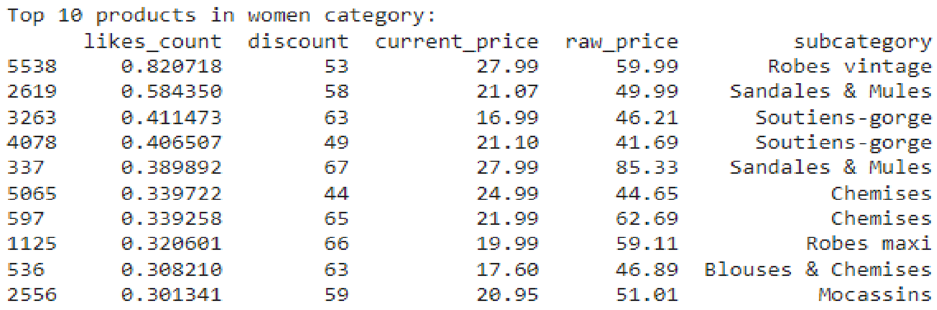
- **Men**: The top products exhibited a wide range of likes_count, with a noticeable influence of discount percentages.



```
Top 10 products in men category:
      likes_count  discount  current_price  raw_price         subcategory
3653     0.427484        44          20.99      37.71              Shirts
984      0.368775        33          25.29      37.71              Shirts
4419     0.329048        69          11.88      38.88            Chemises
5050     0.304915        50          16.99      33.94              Shirts
2223     0.271778        76          24.14     100.64           Pantalons
4456     0.258551        29          20.70      28.99         Sacs à main
853      0.250290        47          17.99      33.99              Shirts
4040     0.233536        39          23.09      37.72              Shirts
5046     0.226435        42          21.99      37.71              Shirts
1592     0.182485        52          20.41      42.31  Sneakers & Baskets
```

Figure 83: Random Forest - Top 10 Products in Men

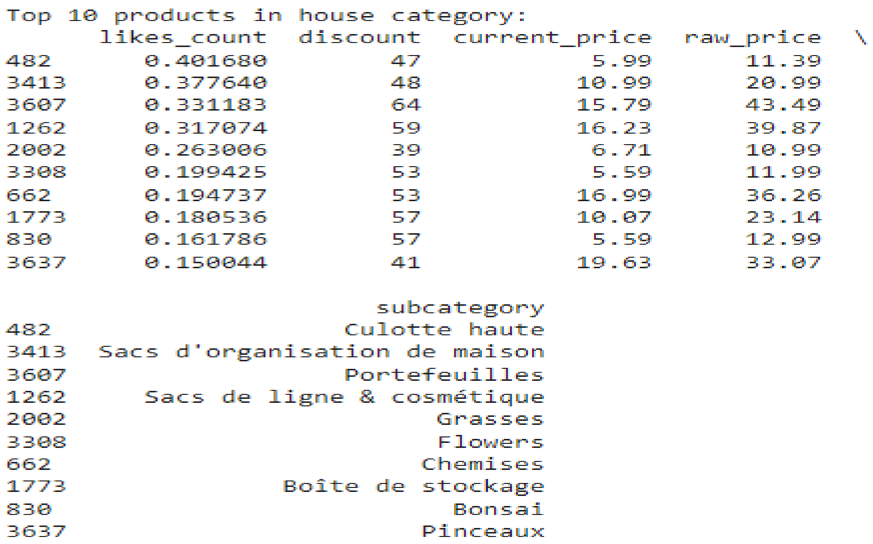
- **Women**: This category had the highest engagement, with top products receiving the most likes and showing substantial discounts.



```
Top 10 products in women category:
      likes_count  discount  current_price  raw_price         subcategory
5538     0.820718        53          27.99      59.99       Robes vintage
2619     0.584350        58          21.07      49.99     Sandales & Mules
3263     0.411473        63          16.99      46.21      Soutiens-gorge
4078     0.406507        49          21.10      41.69      Soutiens-gorge
337      0.389892        67          27.99      85.33     Sandales & Mules
5065     0.339722        44          24.99      44.65            Chemises
597      0.339258        65          21.99      62.69            Chemises
1125     0.320601        66          19.99      59.11          Robes maxi
536      0.308210        63          17.60      46.89   Blouses & Chemises
2556     0.301341        59          20.95      51.01           Mocassins
```

Figure 84: Random Forest - Top 10 Products in Women

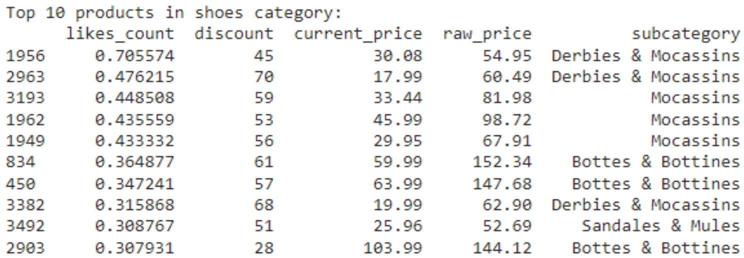
- **House**: Products were generally affordable with moderate user engagement.



```
Top 10 products in house category:
      likes_count  discount  current_price  raw_price  \
482      0.401680        47           5.99      11.39
3413     0.377640        48          10.99      20.99
3607     0.331183        64          15.79      43.49
1262     0.317074        59          16.23      39.87
2002     0.263006        39           6.71      10.99
3308     0.199425        53           5.59      11.99
662      0.194737        53          16.99      36.26
1773     0.180536        57          10.07      23.14
830      0.161786        57           5.59      12.99
3637     0.150044        41          19.63      33.07

                         subcategory
482                   Culotte haute
3413  Sacs d'organisation de maison
3607                  Portefeuilles
1262     Sacs de ligne & cosmétique
2002                        Grasses
3308                        Flowers
662                        Chemises
1773             Boîte de stockage
830                          Bonsai
3637                       Pinceaux
```

Figure 85: Random Forest - Top 10 Products in House

- **Shoes**: High engagement was observed, especially for products with deep discounts (Breiman, 2001) (James et al., 2013).



```
Top 10 products in shoes category:
      likes_count  discount  current_price  raw_price          subcategory
1956     0.705574        45          30.08      54.95  Derbies & Mocassins
2963     0.476215        70          17.99      60.49  Derbies & Mocassins
3193     0.448508        59          33.44      81.98            Mocassins
1962     0.435559        53          45.99      98.72            Mocassins
1949     0.433332        56          29.95      67.91            Mocassins
834      0.364877        61          59.99     152.34    Bottes & Bottines
450      0.347241        57          63.99     147.68    Bottes & Bottines
3382     0.315868        68          19.99      62.90  Derbies & Mocassins
3492     0.308767        51          25.96      52.69     Sandales & Mules
2903     0.307931        28         103.99     144.12    Bottes & Bottines
```

Figure 86: Random Forest - Top 10 Products in Shoes

Category Distribution:

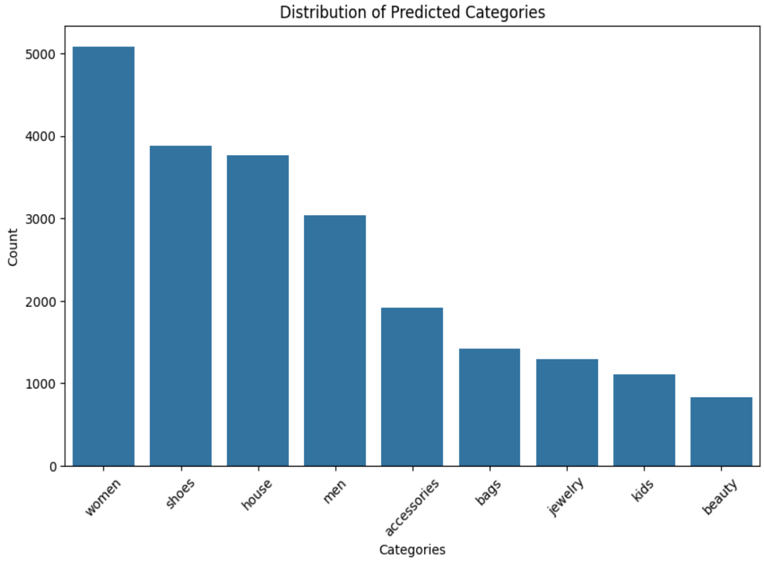
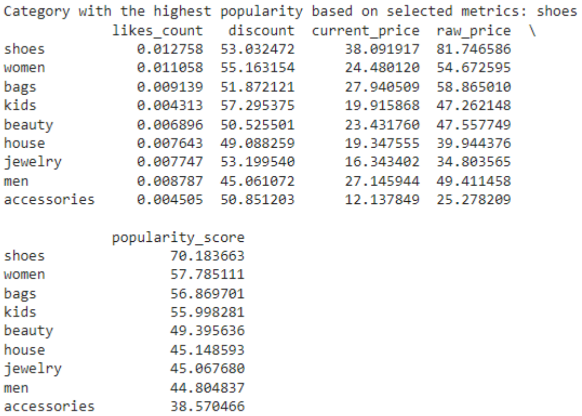


Figure 87: Random Forest - Distribution of Predicted Categories

- The distribution of predicted categories was visualized to assess the model's tendency to favour certain categories over others. This analysis helps in understanding potential biases in the model (Liaw and Wiener, 2011).

Popularity Score by Category:



```
Category with the highest popularity based on selected metrics: shoes
             likes_count   discount  current_price  raw_price  \
shoes           0.012758  53.032472      38.091917  81.746586
women           0.011058  55.163154      24.480120  54.672595
bags            0.009139  51.872121      27.940509  58.865010
kids            0.004313  57.295375      19.915868  47.262148
beauty          0.006896  50.525501      23.431760  47.557749
house           0.007643  49.088259      19.347555  39.944376
jewelry         0.007747  53.199540      16.343402  34.803565
men             0.008787  45.061072      27.145944  49.411458
accessories     0.004505  50.851203      12.137849  25.278209

             popularity_score
shoes               70.183663
women               57.785111
bags                56.869701
kids                55.998281
beauty              49.395636
house               45.148593
jewelry             45.067680
men                 44.804837
accessories         38.570466
```

Figure 88: Random Forest - Most Popular Category - Shoes

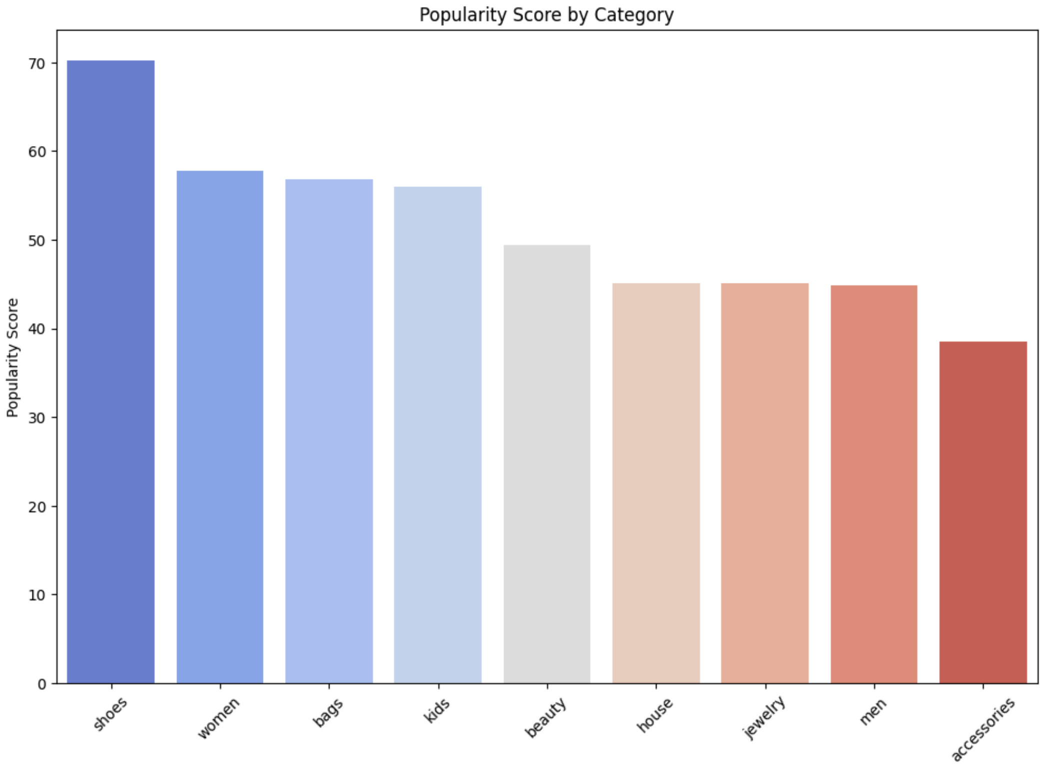


Figure 89: Random Forest - Popularity Score by Category

"Popularity Score by Category" shows the popularity scores of different categories based on selected metrics. The x-axis represents various categories, while the y-axis shows their corresponding popularity scores.

Analysis:

1. **Category Rankings**:
   - **Shoes**: This category has the highest popularity score of approximately 70, making it the most popular category based on the selected metrics.
   - **Women and Bags**: These categories also rank high in popularity, with scores around 57 and 56, respectively.
   - **Accessories**: This category has the lowest popularity score at around 38, suggesting it’s the least popular among the categories analyzed.
2. **Likes Count and Discount**:
   - The category **Shoes** not only has the highest popularity score but also maintains a relatively high likes count (0.012758) and discount rate (53%). This suggests that the combination of a good discount and moderate pricing could be driving its popularity.

    - **The women** category also has a significant likes count (0.011058) and the highest discount rate (55.16%) among the top categories, further contributing to its popularity.
3. **Pricing Insights**:
    - **Current Prices**: Categories like **Shoes** and **Women** have higher current prices ($38.09 and $24.48, respectively) compared to other categories, yet they still manage to maintain high popularity scores. This might indicate a consumer preference for these categories even at higher price points.
    - **Raw Prices**: The raw prices (i.e., original prices before discount) show that categories like **Shoes** and **Bags** were initially more expensive, but the discounts seem to have played a significant role in boosting their popularity.
4. **Popularity Score**:
    - The popularity score is a composite metric, likely influenced by factors like likes count, discounts, and prices. Categories like **Shoes** dominate with the highest score, indicating that consumers are particularly drawn to this category, possibly due to a combination of substantial discounts and moderate pricing.
    - **Accessories**, despite being a common category, ranks the lowest, possibly due to a combination of lower likes count and less favorable discount or pricing strategies.

Key Insights:

- **Category Dominance**: The **Shoes** category clearly stands out in terms of popularity, followed by **Women** and **Bags**. These categories seem to be driving the most consumer interest, likely due to a combination of attractive discounts and reasonable prices.
- **Consumer Behavior**: The analysis suggests that consumers might prioritize categories with better discounts, even if the current price is relatively higher, as seen in the case of **Shoes** and **Women**.
- **Marketing Strategy**: Focusing marketing efforts on categories like **Shoes**, which already have high popularity scores, could yield even better results. On

the other hand, categories like **Accessories** might require strategic adjustments, such as better discounts or pricing, to improve their popularity.

This analysis can be useful for refining product offerings, pricing strategies, and marketing campaigns to align with consumer preferences and maximize engagement (Pedregosa et al., 2011).

Conclusion:

The Random Forest Classifier provided moderate accuracy in predicting product categories, with likes_count and current_price emerging as the most influential features. While the model shows promise, there is room for further refinement, particularly in improving accuracy across categories with lower engagement, such as jewelry. Future work could involve exploring additional features, applying more advanced algorithms, or fine-tuning the existing model to enhance performance (Breiman, 2001) (James et al., 2013).

4.2. Classification Algorithm 2: KNN

This section details the results obtained using the K-Nearest Neighbors (KNN) algorithm. The objective, data understanding and preprocessing strategy remain similar to Section 4.1. The model's performance was evaluated, and hyperparameter tuning was performed to identify the optimal number of neighbours (k).

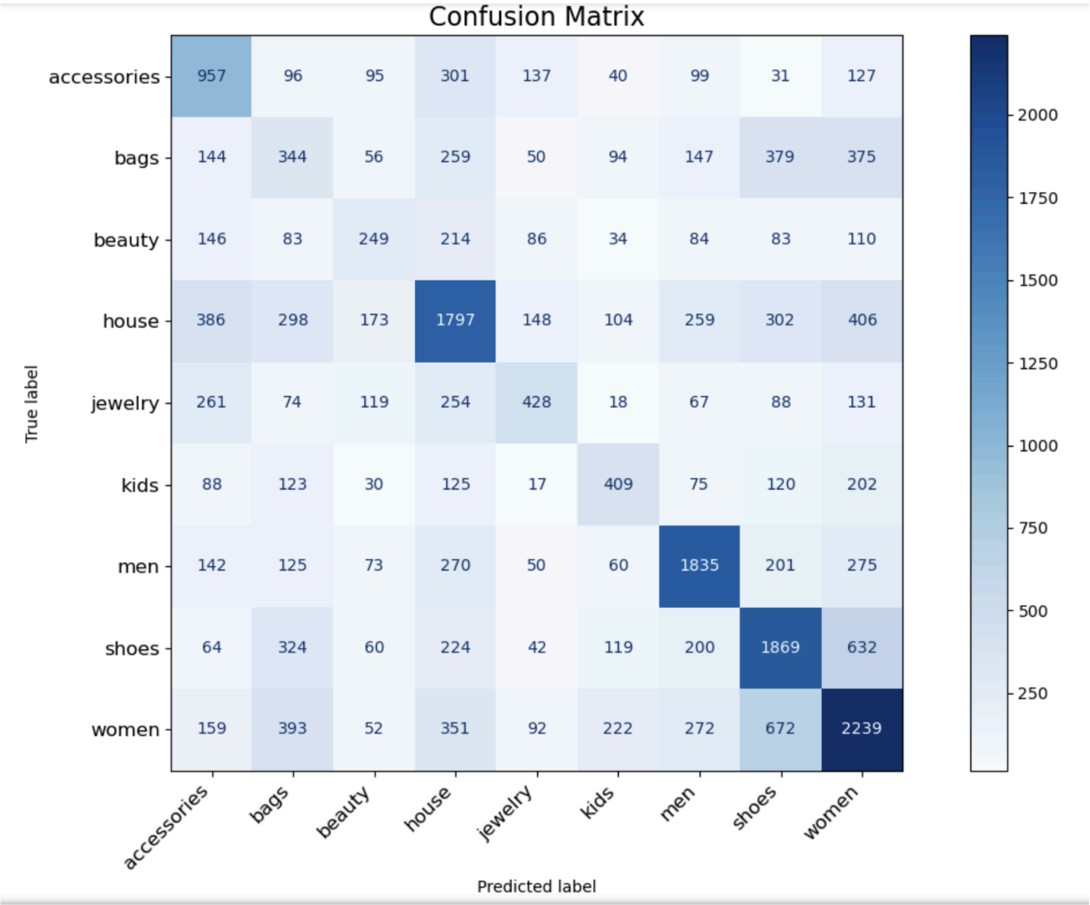
4.2.1. Data Preparation

The dataset was preprocessed to include all 9 categories. Four features discount, likes_count, current_price, and raw_price were selected for the model. The target variable category was encoded using LabelEncoder to convert the categorical values into numeric labels. The data was split into training and testing sets with a 70-30 ratio. The features were standardised using StandardScaler to improve the performance of the KNN algorithm, which relies on distance metrics.

4.2.2. Model Training and Evaluation

The following outcomes were observed after training and evaluation:

- Accuracy of the KNN Model: The model achieved an accuracy of **45.33%** on the test set.
- Confusion Matrix: To visualise the model's performance across different categories, a confusion matrix helped to identify where the model was making the most errors and where it was performing well.
- Best Performing Category: The category that the model predicted most accurately was identified as **men**. This was determined by analysing the diagonal of the confusion matrix, which represents the correct predictions for each category. The data for the men category was predicted with an accuracy of 60.54%.



	accessories	bags	beauty	house	jewelry	kids	men	shoes	women
accessories	957	96	95	301	137	40	99	31	127
bags	144	344	56	259	50	94	147	379	375
beauty	146	83	249	214	86	34	84	83	110
house	386	298	173	1797	148	104	259	302	406
jewelry	261	74	119	254	428	18	67	88	131
kids	88	123	30	125	17	409	75	120	202
men	142	125	73	270	50	60	1835	201	275
shoes	64	324	60	224	42	119	200	1869	632
women	159	393	52	351	92	222	272	672	2239

Figure 90: KNN - Confusion Matrix

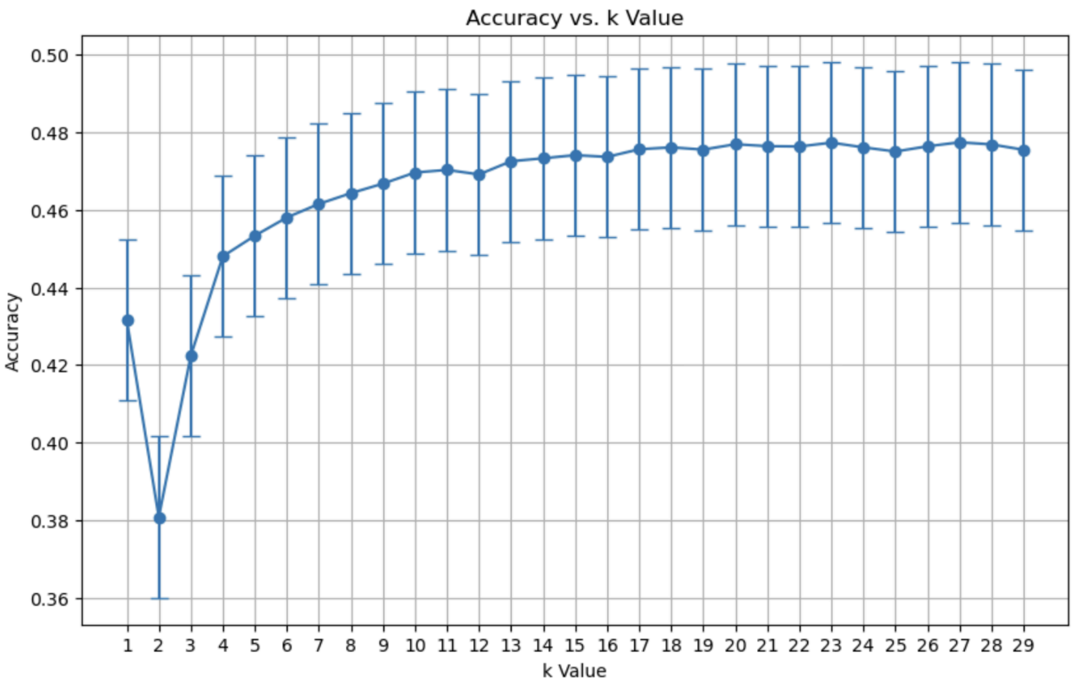


Figure 91: KNN - Accuracy vs k Value

4.2.3. Hyperparameter Tuning with GridSearchCV

To optimize the KNN model, the number of neighbours (k) was fine-tuned using GridSearchCV. The grid search was performed over a range of k values, and 10-fold cross-validation was used to ensure better results. The following key findings were observed:

- Best K Value: The optimal number of neighbours was found to be **26**. This value provided the best balance between bias and variance, resulting in the highest accuracy.
- Best Accuracy from GridSearchCV: The highest accuracy achieved during the grid search was **47.84%**. This indicates a slight improvement over the manually selected k value of 5.
- Comparison of Accuracies: The relationship between k and accuracy was plotted to compare the manual selection process with the grid search results. The grid search identified a more optimal k value, leading to better model performance.

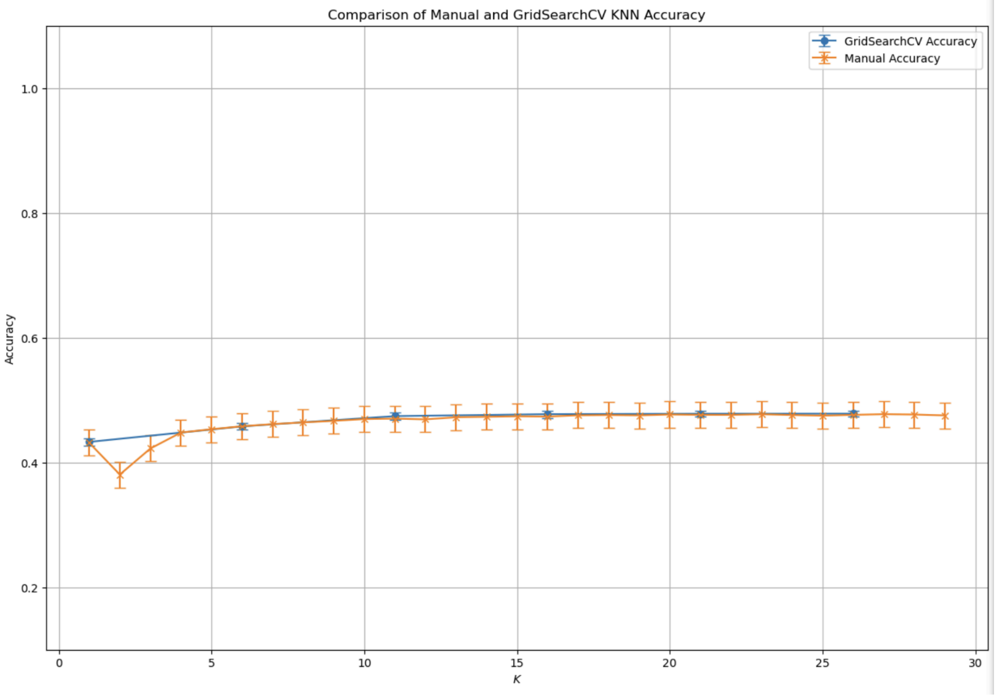
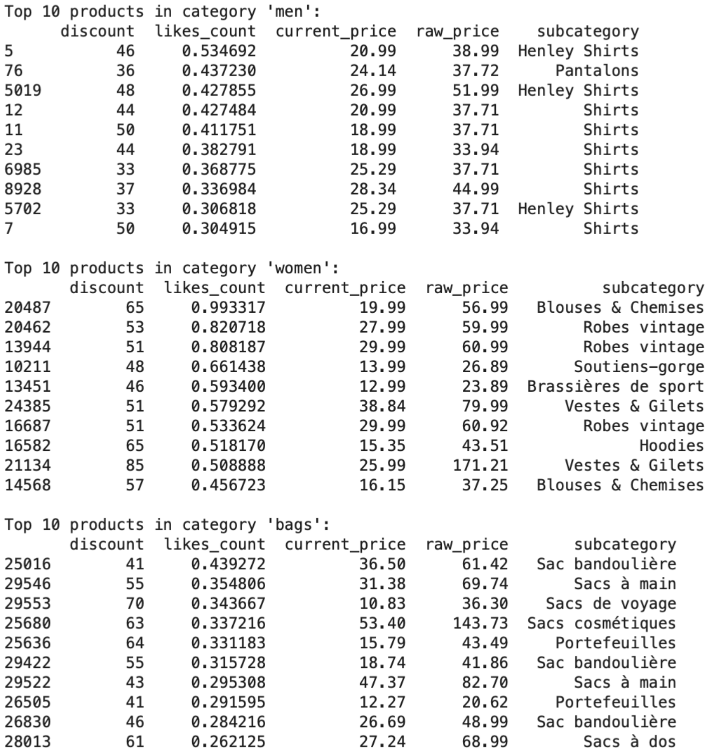


Figure 92: KNN - Comparison of Manual and GridSearchCV KNN Accuracy

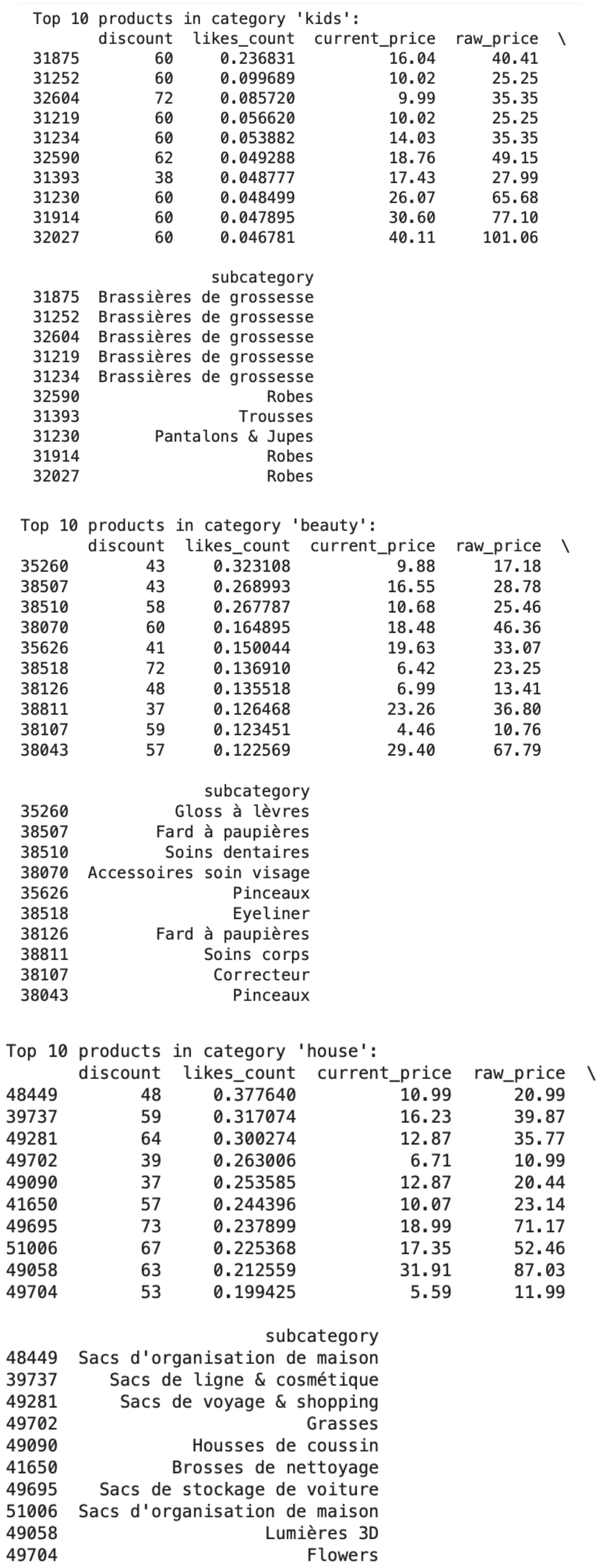
4.2.4. Top 10 Products (in each category and overall)

The top 10 products for each category and overall were identified based on the likes_count column.



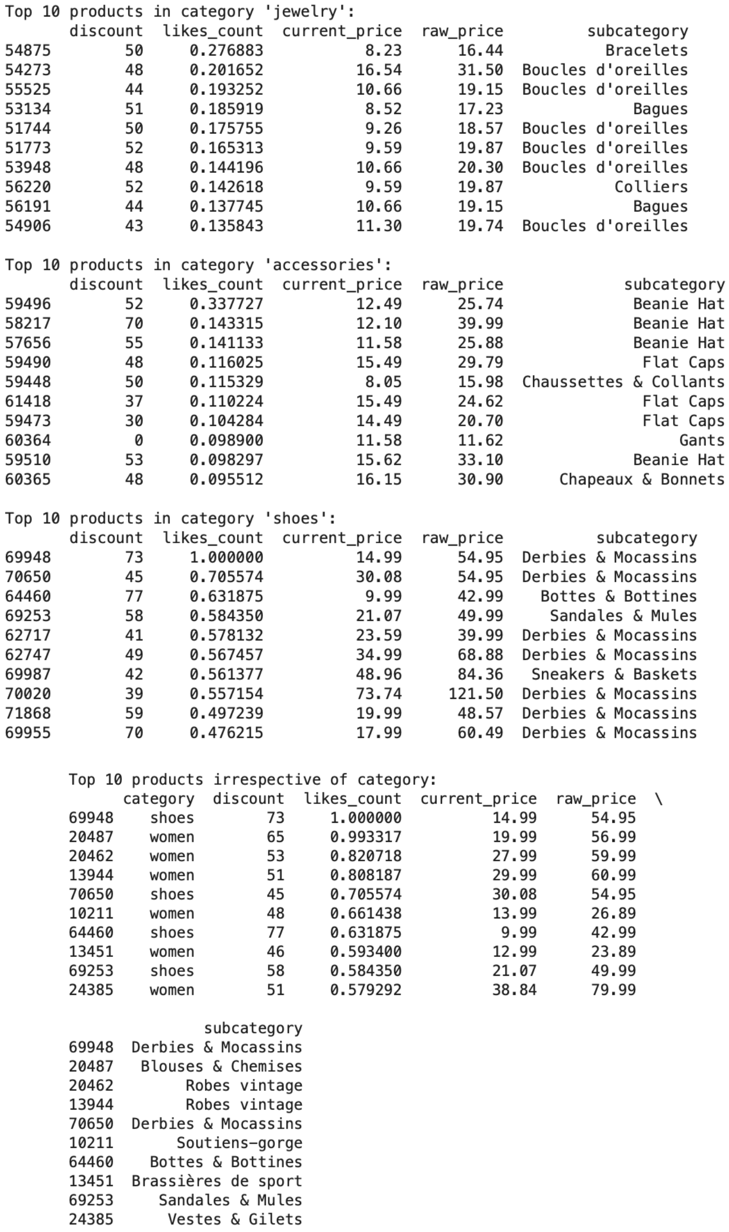
```
Top 10 products in category 'men':
      discount  likes_count  current_price  raw_price    subcategory
5           46     0.534692          20.99      38.99  Henley Shirts
76          36     0.437230          24.14      37.72      Pantalons
5019        48     0.427855          26.99      51.99  Henley Shirts
12          44     0.427484          20.99      37.71         Shirts
11          50     0.411751          18.99      37.71         Shirts
23          44     0.382791          18.99      33.94         Shirts
6985        33     0.368775          25.29      37.71         Shirts
8928        37     0.336984          28.34      44.99         Shirts
5702        33     0.306818          25.29      37.71  Henley Shirts
7           50     0.304915          16.99      33.94         Shirts

Top 10 products in category 'women':
       discount  likes_count  current_price  raw_price          subcategory
20487        65     0.993317          19.99      56.99   Blouses & Chemises
20462        53     0.820718          27.99      59.99        Robes vintage
13944        51     0.808187          29.99      60.99        Robes vintage
10211        48     0.661438          13.99      26.89       Soutiens-gorge
13451        46     0.593400          12.99      23.89  Brassières de sport
24385        51     0.579292          38.84      79.99      Vestes & Gilets
16687        51     0.533624          29.99      60.92        Robes vintage
16582        65     0.518170          15.35      43.51              Hoodies
21134        85     0.508888          25.99     171.21      Vestes & Gilets
14568        57     0.456723          16.15      37.25   Blouses & Chemises

Top 10 products in category 'bags':
       discount  likes_count  current_price  raw_price       subcategory
25016        41     0.439272          36.50      61.42   Sac bandoulière
29546        55     0.354806          31.38      69.74       Sacs à main
29553        70     0.343667          10.83      36.30    Sacs de voyage
25680        63     0.337216          53.40     143.73  Sacs cosmétiques
25636        64     0.331183          15.79      43.49     Portefeuilles
29422        55     0.315728          18.74      41.86   Sac bandoulière
29522        43     0.295308          47.37      82.70       Sacs à main
26505        41     0.291595          12.27      20.62     Portefeuilles
26830        46     0.284216          26.69      48.99   Sac bandoulière
28013        61     0.262125          27.24      68.99        Sacs à dos
```

Figure 93: KNN - Top 10 Products for each Product Part 1



```
Top 10 products in category 'kids':
       discount  likes_count  current_price  raw_price  \
31875        60     0.236831          16.04      40.41
31252        60     0.099689          10.02      25.25
32604        72     0.085720           9.99      35.35
31219        60     0.056620          10.02      25.25
31234        60     0.053882          14.03      35.35
32590        62     0.049288          18.76      49.15
31393        38     0.048777          17.43      27.99
31230        60     0.048499          26.07      65.68
31914        60     0.047895          30.60      77.10
32027        60     0.046781          40.11     101.06

                   subcategory
31875  Brassières de grossesse
31252  Brassières de grossesse
32604  Brassières de grossesse
31219  Brassières de grossesse
31234  Brassières de grossesse
32590                    Robes
31393                 Trousses
31230        Pantalons & Jupes
31914                    Robes
32027                    Robes

Top 10 products in category 'beauty':
       discount  likes_count  current_price  raw_price  \
35260        43     0.323108           9.88      17.18
38507        43     0.268993          16.55      28.78
38510        58     0.267787          10.68      25.46
38070        60     0.164895          18.48      46.36
35626        41     0.150044          19.63      33.07
38518        72     0.136910           6.42      23.25
38126        48     0.135518           6.99      13.41
38811        37     0.126468          23.26      36.80
38107        59     0.123451           4.46      10.76
38043        57     0.122569          29.40      67.79

                   subcategory
35260          Gloss à lèvres
38507        Fard à paupières
38510         Soins dentaires
38070  Accessoires soin visage
35626                Pinceaux
38518                Eyeliner
38126        Fard à paupières
38811             Soins corps
38107              Correcteur
38043                Pinceaux

Top 10 products in category 'house':
       discount  likes_count  current_price  raw_price  \
48449        48     0.377640          10.99      20.99
39737        59     0.317074          16.23      39.87
49281        64     0.300274          12.87      35.77
49702        39     0.263006           6.71      10.99
49090        37     0.253585          12.87      20.44
41650        57     0.244396          10.07      23.14
49695        73     0.237899          18.99      71.17
51006        67     0.225368          17.35      52.46
49058        63     0.212559          31.91      87.03
49704        53     0.199425           5.59      11.99

                       subcategory
48449  Sacs d'organisation de maison
39737     Sacs de ligne & cosmétique
49281      Sacs de voyage & shopping
49702                        Grasses
49090             Housses de coussin
41650           Brosses de nettoyage
49695    Sacs de stockage de voiture
51006  Sacs d'organisation de maison
49058                    Lumières 3D
49704                        Flowers
```

Figure 94: KNN - Top 10 Products for each Product Part 2



```
Top 10 products in category 'jewelry':
       discount  likes_count  current_price  raw_price         subcategory
54875        50     0.276883           8.23      16.44           Bracelets
54273        48     0.201652          16.54      31.50  Boucles d'oreilles
55525        44     0.193252          10.66      19.15  Boucles d'oreilles
53134        51     0.185919           8.52      17.23              Bagues
51744        50     0.175755           9.26      18.57  Boucles d'oreilles
51773        52     0.165313           9.59      19.87  Boucles d'oreilles
53948        48     0.144196          10.66      20.30  Boucles d'oreilles
56220        52     0.142618           9.59      19.87            Colliers
56191        44     0.137745          10.66      19.15              Bagues
54906        43     0.135843          11.30      19.74  Boucles d'oreilles

Top 10 products in category 'accessories':
       discount  likes_count  current_price  raw_price             subcategory
59496        52     0.337727          12.49      25.74              Beanie Hat
58217        70     0.143315          12.10      39.99              Beanie Hat
57656        55     0.141133          11.58      25.88              Beanie Hat
59490        48     0.116025          15.49      29.79               Flat Caps
59448        50     0.115329           8.05      15.98  Chaussettes & Collants
61418        37     0.110224          15.49      24.62               Flat Caps
59473        30     0.104284          14.49      20.70               Flat Caps
60364         0     0.098900          11.58      11.62                   Gants
59510        53     0.098297          15.62      33.10              Beanie Hat
60365        48     0.095512          16.15      30.90     Chapeaux & Bonnets

Top 10 products in category 'shoes':
       discount  likes_count  current_price  raw_price          subcategory
69948        73     1.000000          14.99      54.95  Derbies & Mocassins
70650        45     0.705574          30.08      54.95  Derbies & Mocassins
64460        77     0.631875           9.99      42.99    Bottes & Bottines
69253        58     0.584350          21.07      49.99     Sandales & Mules
62717        41     0.578132          23.59      39.99  Derbies & Mocassins
62747        49     0.567457          34.99      68.88  Derbies & Mocassins
69987        42     0.561377          48.96      84.36   Sneakers & Baskets
70020        39     0.557154          73.74     121.50  Derbies & Mocassins
71868        59     0.497239          19.99      48.57  Derbies & Mocassins
69955        70     0.476215          17.99      60.49  Derbies & Mocassins

Top 10 products irrespective of category:
      category  discount  likes_count  current_price  raw_price  \
69948    shoes        73     1.000000          14.99      54.95
20487    women        65     0.993317          19.99      56.99
20462    women        53     0.820718          27.99      59.99
13944    women        51     0.808187          29.99      60.99
70650    shoes        45     0.705574          30.08      54.95
10211    women        48     0.661438          13.99      26.89
64460    shoes        77     0.631875           9.99      42.99
13451    women        46     0.593400          12.99      23.89
69253    shoes        58     0.584350          21.07      49.99
24385    women        51     0.579292          38.84      79.99

               subcategory
69948  Derbies & Mocassins
20487   Blouses & Chemises
20462        Robes vintage
13944        Robes vintage
70650  Derbies & Mocassins
10211       Soutiens-gorge
64460    Bottes & Bottines
13451  Brassières de sport
69253     Sandales & Mules
24385      Vestes & Gilets
```

Figure 95: KNN - Top 10 Products for each Product Part 3

4.2.5. Conclusion

The KNN classifier provided reasonable accuracy in classifying products into categories, with the men category being the best-performing according to the model. While the initial accuracy was moderate, the use of GridSearchCV improved the model's performance, demonstrating the importance of hyperparameter tuning in machine-learning tasks. Additionally, the identification of top products within each category adds practical value, enabling better decision-making for product management on the e-commerce platform.

This process highlights the effectiveness of combining machine learning algorithms with systematic hyperparameter tuning to achieve optimal performance in classification tasks. Further improvements could be made by exploring other classifiers or by engineering additional features to enhance model accuracy.

4.3. Comparison: Random Forest vs. K-Nearest Neighbors (KNN) Classifiers

Model Overview:

In this comparison, we examine the performance and characteristics of two classification algorithms—Random Forest (RF) and K-Nearest Neighbors (KNN)—used to classify products into categories such as accessories, bags, beauty, jewellery, kids, men, women, house, and shoes. Both models were trained on the same dataset, with identical preprocessing steps, but they utilized different methodologies for learning and prediction. The comparison covers multiple aspects, including accuracy, feature importance, model complexity, and potential areas for improvement.

4.3.1 Accuracy and Performance:

Table 3 shows the accuracies of both classification models

Random Forest	KNN
57.94%	47.84

Table 3: Accuracies of Both Classification Techniques

- **Random Forest Classifier:**
    - **Accuracy:** The Random Forest model achieved an accuracy of **57.94%**, which was significantly higher than KNN's accuracy. This suggests that Random Forest could better capture the relationships between the features and categories.
    - **Strengths:** The Random Forest algorithm is particularly effective at handling complex datasets with nonlinear relationships and interactions among features. It benefits from ensemble learning, which aggregates

the predictions of multiple decision trees to reduce overfitting and improve generalization to unseen data (Breiman, 2001).

    - **Weaknesses:** Despite its higher accuracy, the Random Forest model did not perform equally well across all categories. For example, it struggled to accurately classify categories like jewelry, which have fewer likes or lower engagement levels.

- **K-Nearest Neighbors Classifier:**
    - **Accuracy:** The KNN classifier achieved an initial accuracy of **45.33%**, which improved slightly to **47.84%** after hyperparameter tuning using GridSearchCV. The accuracy remained lower than Random Forest, highlighting KNN's limitations in this context.
    - **Strengths:** KNN is a simple, intuitive model that performs well in cases where data is evenly distributed. Its performance can be improved by feature scaling and careful selection of hyperparameters, as shown by the improvement from GridSearchCV.
    - **Weaknesses:** KNN relies on distance metrics, making it sensitive to outliers and the curse of dimensionality. It struggles when there are many features, as the differences between nearby and far-away points become less distinct. Additionally, the KNN model showed a tendency to perform better in categories like men but was less effective across more complex categories like jewellery or shoes.

4.3.2 Model Interpretability and Feature Importance:

- **Random Forest:**
    - **Feature Importance:** One of the key advantages of Random Forest is its ability to provide insights into feature importance. In this analysis, **likes_count** emerged as the most influential feature, followed by **current_price**. These metrics indicate that user engagement and pricing were critical factors in determining product categories. Random Forest's built-in feature importance metrics help to explain which features drive predictions, offering valuable interpretability to the model (Liaw and Wiener, 2011).

    - **Confusion Matrix and Errors:** The confusion matrix for Random Forest revealed misclassification patterns, which were more pronounced in categories with lower engagement, such as jewellery. This misclassification is typical in ensemble models that rely heavily on dominant features (Pedregosa et al., 2011).
- **K-Nearest Neighbors:**
    - **Lack of Feature Importance:** Unlike Random Forest, KNN does not provide feature importance measures. This limits the interpretability of the model, as it is not possible to determine which features contribute the most to its predictions. The model operates based solely on the proximity of data points in feature space, without an internal mechanism for weighting the importance of those features (James et al., 2013).
    - **Confusion Matrix and Errors:** Similar to Random Forest, the KNN confusion matrix revealed that the model was more accurate for some categories (like men) but struggled with others, especially those with low likes_count or complex pricing structures. This suggests that KNN is less adaptable to the intricacies of the dataset compared to Random Forest (Liaw and Wiener, 2011).

4.3.3 Computational Complexity and Scalability:

- **Random Forest:**
    - **Training Time and Scalability:** Random Forest is computationally intensive, particularly when the number of trees (estimators) is large. However, once trained, the model can make predictions efficiently. Random Forest can also handle large datasets better than KNN, as it builds multiple trees in parallel and can scale more effectively with data size (Breiman, 2001).
    - **Hyperparameter Tuning:** The tuning process for Random Forest involves adjusting parameters like the number of trees and maximum depth of the trees. While tuning these parameters can be time-consuming, it is generally straightforward and can lead to significant improvements in performance.

- **K-Nearest Neighbors:**
    - **Training Time and Scalability:** KNN is relatively simple to train, but it suffers from high computational costs during prediction. Since the algorithm must compute the distance between the test point and every point in the training set, the prediction time increases significantly with larger datasets. This makes KNN less scalable than Random Forest, especially for datasets with many observations or features (James et al., 2013).
    - **Hyperparameter Tuning:** Hyperparameter tuning for KNN primarily involves selecting the optimal number of neighbors (k). The GridSearchCV approach improved the model's accuracy slightly, but even the tuned KNN could not outperform the Random Forest model. Moreover, the tuning process is computationally expensive due to the repeated distance calculations.

4.3.4 Application and Use Case:

- **Random Forest:**
    - **Best Use Case:** Random Forest is best suited for applications where accuracy and robustness are critical, and interpretability is needed. It is highly effective in scenarios involving complex interactions between features and can provide actionable insights through feature importance. For e-commerce platforms, Random Forest would be ideal for more intricate tasks, such as identifying key drivers of customer engagement or optimizing pricing strategies across various product categories (Liaw and Wiener, 2011).
- **K-Nearest Neighbors:**
    - **Best Use Case:** KNN is most appropriate for smaller, simpler datasets where interpretability is less important, and computational resources are limited. KNN can be a good choice for quick, baseline classification tasks, but it is less suitable for large-scale applications or cases where model complexity needs to be accounted for, as seen in the product categorization task in this analysis (James et al., 2013).

4.3.5 Conclusion and Future Directions:

In summary, Random Forest emerged as the better-performing algorithm in this analysis, with higher accuracy, better interpretability through feature importance, and greater robustness to the complexities of the dataset. The KNN classifier, while simpler, struggled with the dataset's intricacies and was less effective in predicting product categories. Future work could focus on further optimizing the Random Forest model by experimenting with different ensemble techniques, such as Gradient Boosting or XGBoost, or incorporating additional features to improve accuracy across all categories.

5. Result Discussion:

5.1. Findings

- **Are the clusters well separated from each other?**
  - Clusters generated by K-means and DBSCAN vary significantly.
  - K-means clustering demonstrates better cluster separation with silhouette scores ranging approximately between 0.1 to 0.45 (refer to Figure 71).
  - When considering the DBSCAN clustering, it showed poor cluster separation. The silhouette scores for DBSCAN were mostly negative or close to zero showing an overlap between the clusters (refer to Figure 72).
  - This demonstrates that due to the complexity of the data, DBSCAN could not differentiate between the clusters.
  - Overall, the analysis shows that clusters that were generated by the K-means clustering were well separated, while those generated by the DBSCAN clustering showed significant overlap.

- **Did the classifiers well separate products from each other into different classes?**
  - The Random Forest and the KNN models both achieved 57.94% and 45.33% accuracies respectively.
  - Despite the low accuracy of the Random Forest model, it performed reasonably well as according to its confusion matrix (refer to Figure 75), for each category the highest prediction types were those of the True Positives which indicates that to some extent the Random Forest model was successful in separating the products into the correct classes. However, there was still some confusion and misclassification of some products (refer to Figure 75). For instance, the “women” category has the highest number of True Positives, 2992 (refer to Figure 75), which indicates that this was the category that the model was most accurate in. However, this might purely be because the “women” category has

the highest number of products, 14k, as indicated in Figure 18. Despite the high number of True Positives for this category, there were still some significant misclassifications into categories like “shoes”, “house”, and “men” (refer to Figure 75) which is interesting as these were the second, third, and fourth categories that had the next highest number of products as indicated by Figure 18. Similar confusions happened for the “house” (confused with “jewellery” and “men”), “shoes” (confused with “women” and “men”), and “accessories” (confused with “bags” and “beauty”) categories (refer to Figure 75). As a result, it can be deduced that the model is struggling a little bit to distinguish between these particular product categories compared to others which might be because of some similar features between these classes.

- The KNN model also had a low accuracy but it was much lower than that of the Random Forest model. This indicates that it had an even harder time separating the products into different classes effectively. According to its confusion matrix (refer to Figure 90), The KNN model achieved for most of the categories the highest prediction types were those of the True Positives, except for the “bags” category. Like the Random Forest Model, the KNN model achieved the highest True Positive predictions in the “women” category (refer to Figure 90) which as was indicated before might just be because of the high number of products in the “women” category (refer to Figure 18). However, it still has quite a significant amount of misclassifications, especially in the “shoes”, “bags” and “house” categories (refer to Figure 90). Similar misclassifications were with the “accessories” (confused with “house”, bags”, and “beauty”), “bags” (confused with “shoes”, “women”, and “house”), “house” (confused with “accessories”, “shoes” and “jewellery”), and “women” (confused with “shoes”, “bags”, and “house”) categories (refer to Figure 90).
- Overall, the Random Forest Model was able to a certain extent to correctly classify the products into the appropriate categories. The KNN model, on the other hand, was unable to classify the products into the correct categories.

- **Do any of the clusters/classes have only a few points?**
    - Most classes in both classification models had the highest number of True Positives, except the “bags” category in the KNN model which had 379 and 375 products incorrectly misclassified as “shoes” and “women” respectively compared to the 344 True Positive products that were correctly classified as “bags” (refer to Figure 90). Hence, it is safe to say that no classes had only a few points.
    - None of the clusters in DBSCAN and K-means have only a few points and the data points were all well distributed among the different clusters of the models. This indicates that the clusters were well-defined and that allowed us to study the clusters better and identify patterns because of the clear segmentation between them.

- **Are there meaningful and non-meaningful clusters/classes to the analytics problems questioned in Task 1?**
    - To answer this question, it is better to look at the best classification model and the best clustering model.
    - **Random Forest:**
        - *Meaningful classes:* Based on the popularity score of each category, we have analysed that the “shoes” and “women” categories are the most meaningful classes based on their rankings, likes_count and discounts, pricing insights, and popularity score as was discussed in Section 4.1.3. These two classes will help us choose the best category and the top 10 products that mostly fall underneath them according to Figure 76.
        - *Non-meaningful classes:* Based on the top 10 products stated in Figure 76, the non-meaningful classes would be the classes that do not have the top 10 products, including the “beauty”, “house”, “jewellery”, and “kids” classes.

- **K-Means:**

  - *Meaningful Clusters:* In K-Means clustering, it is important to evaluate the meaningful clusters because only meaningful, logical, interpretable data that has real-world meaning and relevance can be interpreted into valuable insight and better decision-making. A characteristic of the meaningful cluster is, that similar data points are likely to form within a cluster and each cluster is well-separated from the other and has distinct characteristics that separate one cluster from another (Sachinsoni 2023). Here is an example of a meaningful cluster where each cluster data point is separated from data points of other clusters, and it produces meaningful insights about the relation of how the amount of discount can attain customer attention resulting in more sales and popularity for a product (Sachinsoni 2023).

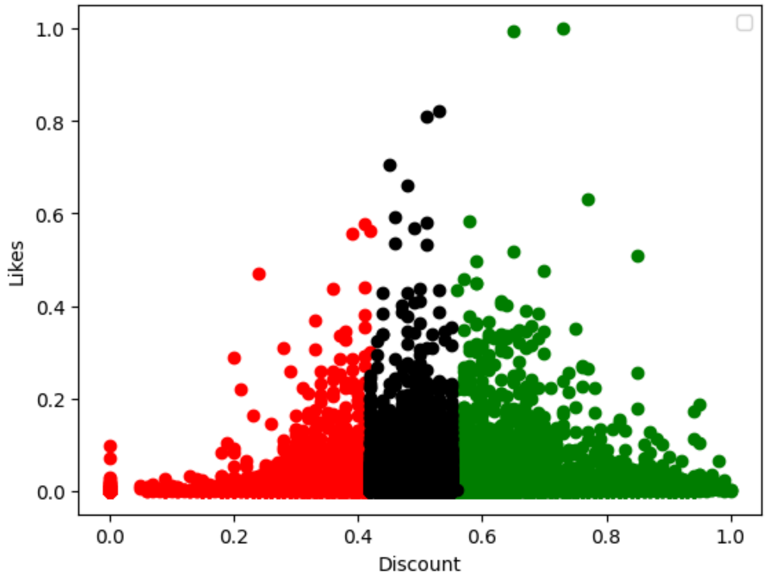


Figure 96: Meaningful Cluster (discount vs likes_count)

Another example from this scatter plot can be derived, that is, the black data points indicate high discounts and high engagement (likes) for the group of products which suggests this cluster represents popular best-selling items, which could be useful for business strategies.

- *Non-meaningful clusters:* A non-meaningful cluster groups data points that are non-related to each other, don't interpret meaningful data and may be misleading. A common characteristic of non-meaningful clusters is that data points of clusters overlap with each other and are not easily separable (Sachinsoni 2023).

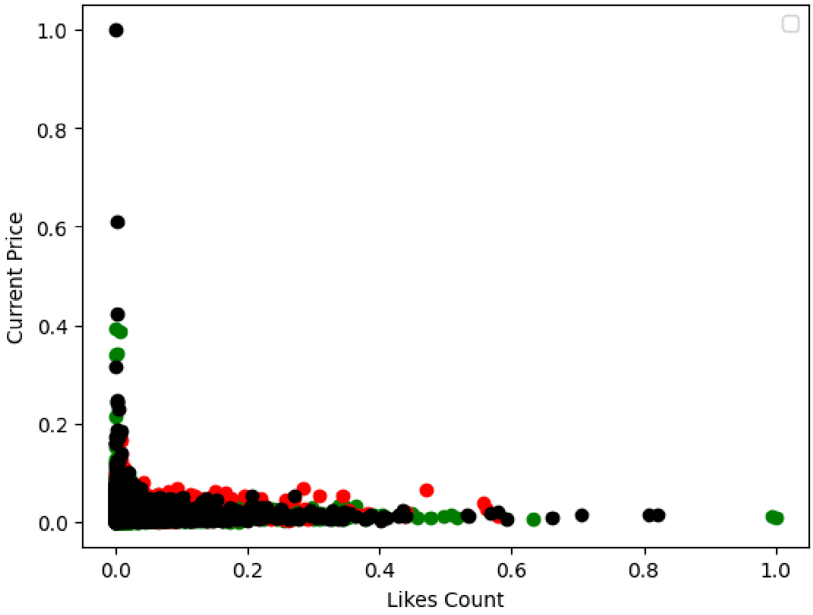


Figure 97: Non-meaningful Clusters (likes_count vs current_price)

Here, in the above-mentioned scatter plot, we can see that points of different clusters are overlapping and there is not much diversity. This type of clustering in data distribution is non-meaningful.

- **What are the advantages, shortages for clustering and classification algorithms in this analytics case? Which one provides results of greater value?**

  - There are distinct advantages for clustering and classification algorithms in this analytics case and also there are certain limitations.
  - Clustering algorithms are experts in finding inherent groupings within the data without any predefined labels (Ester et al., 1996). Here we have used two types of clustering; K-means and DBSCAN. Comparing

K-means and DBSCAN; K-means were successful in identifying the clusters with better cluster separation.

  - Clustering has its own disadvantages also. The challenges occur in selecting the optimal number of clusters for K-mean and appropriate parameters for DBSCAN. The silhouette scores, especially in DBSCAN clearly show the difficulty in clearly distinguishing between different clusters.
  - Classification algorithms focus on predicting data on known features. The classification algorithms used here are Random Forest and KNN. Random Forest showed higher accuracy, highlighting the importance of likes_count and current_price (Breiman, 2001; Liaw & Wiener, 2011).
  - The shortcomings or disadvantages of the classification algorithm are that it can produce inaccurate results if it is implemented incorrectly (Varun Samarth, 2023).
  - Specific analytical goals determine the algorithm of greater value. If the aim of the analysis is to divide the products based on their similarities and to find hidden patterns, clustering provides valuable insights.
  - If the aim of the analysis is to categorise the products for applications like inventory management then classification provides greater value.

- **Are the examined algorithms suitable for Big Data analytics? and why in your opinion?**

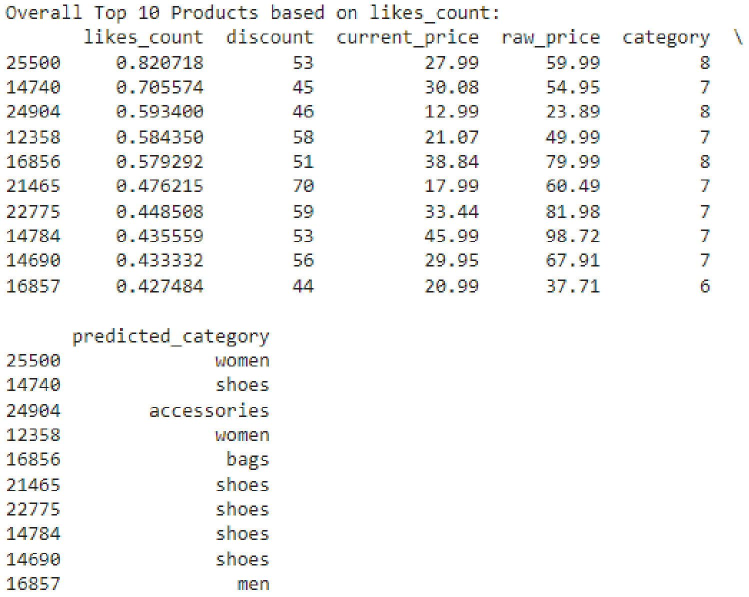
  - The algorithms that are analysed can be useful for Big Data analytics.
  - K-means clustering- this one is a pretty basic algorithm which could be easily understandable and was well for getting started. It is scalable so that it can deal with big data. In contrast, DBSCAN is used to find complex structures in Big Data approaches.
  - Since tree construction is easily parallelizable, random forest classification works very well on large datasets. This gives insights to identify which features play a significant role in the classification. In contrast is KNN which has the advantage of being easy to understand and implement. However, KNN is computationally quite heavy for large datasets.

  - In summary, K-means and random forest are more scalable and performant solutions for Big Data analytics in comparison with KNN or DBSCAN.
  - The algorithms are suited to different types of data and results
  - Sometimes there could be a need for the hybrid methodology to combine and use the best part of both these algorithms.

- **Will data preprocess affect clustering and classification results? and why in your opinion?**
  - Yes, it will affect both results. This is because there have been several missing data and duplicates in the original dataset provided by the assignment and not handling this data would have negatively affected the results of the models and led to inaccurate cluster formations.
  - Additionally, not scaling the likes_count could have also led to clusters being based on likes_count, which would have had the highest scale, while ignoring the ones with the smaller scales which can affect the identification of core points in DBSCAN and hence the formation of the appropriate clusters. The classification techniques' results might have also been negatively affected by the large scale of the likes_count.
  - Additional data preprocessing methods should have been implemented as well to handle the noisy and outlier data that negatively affects the classification and clustering techniques' results which will help in achieving better results.

5.2. Top 10 Products

In Sections 3.3 and 4.3, it was decided that the K-Means and Random Forest models were the best because of their results. In order to choose the top 10 products, it has been decided to choose the products reported by the best model out of these two. According to Figure 75, the Random Forest model has strong performance in classifying the products of the “women”, “shoes”, and “men” categories. However, it does not perform well in classifying the “beauty” and “kids” products (refer to Figure 75). K-Means, on the other hand, has high silhouette scores for the “accessories” and “men” categories (refer to Table 2). However, its performance is still generally weaker than that of the Random Forest model. Hence, we will be reporting the top 10 products generated by the Random Forest Model (refer to Figure 98).

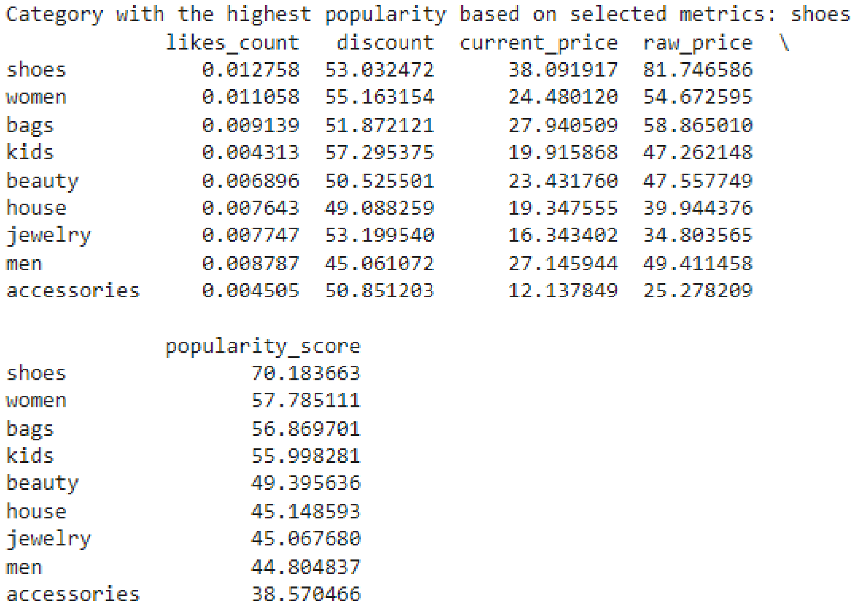


```
Overall Top 10 Products based on likes_count:
      likes_count  discount  current_price  raw_price  category  \
25500    0.820718        53          27.99      59.99         8
14740    0.705574        45          30.08      54.95         7
24904    0.593400        46          12.99      23.89         8
12358    0.584350        58          21.07      49.99         7
16856    0.579292        51          38.84      79.99         8
21465    0.476215        70          17.99      60.49         7
22775    0.448508        59          33.44      81.98         7
14784    0.435559        53          45.99      98.72         7
14690    0.433332        56          29.95      67.91         7
16857    0.427484        44          20.99      37.71         6

     predicted_category
25500             women
14740             shoes
24904       accessories
12358             women
16856              bags
21465             shoes
22775             shoes
14784             shoes
14690             shoes
16857               men
```

Figure 98: Overall Top 10 Products based on likes_count

5.3. Best Category

Using the results generated by the Random Forest model for the same reasons stated in Section 5.2, the best category is the “shoes” category as shown in Figure 99.



```
Category with the highest popularity based on selected metrics: shoes
             likes_count   discount  current_price  raw_price  \
shoes           0.012758  53.032472      38.091917  81.746586
women           0.011058  55.163154      24.480120  54.672595
bags            0.009139  51.872121      27.940509  58.865010
kids            0.004313  57.295375      19.915868  47.262148
beauty          0.006896  50.525501      23.431760  47.557749
house           0.007643  49.088259      19.347555  39.944376
jewelry         0.007747  53.199540      16.343402  34.803565
men             0.008787  45.061072      27.145944  49.411458
accessories     0.004505  50.851203      12.137849  25.278209

             popularity_score
shoes               70.183663
women               57.785111
bags                56.869701
kids                55.998281
beauty              49.395636
house               45.148593
jewelry             45.067680
men                 44.804837
accessories         38.570466
```

Figure 99: Category with The Highest Popularity

5.4. Suggestions

1. Focus on the models' high-performing categories:
    a. Invest in marketing campaigns, promotions and personalised recommendations for Women, Shoes and Men. These categories are likely to represent strong, established segments within the company's customer base.
2. Improve Product Variation for Weak Categories:
    a. Review and refine the product offerings in the Beauty and Bags category since as per both classification and clustering these two categories show poor performance.
    b. Create product bundles that combine weak-performing items with stronger items. This might attract customers to the weaker-performing items.
3. Optimise Pricing and Promotions Based on Customer Segments:
    a. Use the clustering insights to design the pricing strategies and promotions such as offering discounts on accessories when customers purchase items from Women or Men categories. This can potentially increase sales conversion rates
4. Expand Successful Categories into New Markets or Product Lines:
    a. As the Women and Shoes categories are strong performers, conduct market research to identify the potential new markets or product extensions for these categories. This will enable better business decisions.
5. Customer Feedback and Sentiment Analysis:
    a. Implement customer surveys or sentiments analysis in reviews and social media to gather insights. By using these feedback the company can make informed decisions about product improvements, customer service and marketing strategies for products in weaker categories like Beauty and Bags.

By focusing on the strengths and addressing the weaknesses identified in the analysis, NewChic can improve its product offerings, optimise marketing strategies and increase sales. Customer feedback and market insights will enable the company to stay competitive and grow in an efficient manner.

6. References:

Acharya, A., 2024, ‘How Poor Data is Killing Your Models and How to Fix It’, *Encord*, viewed 20 August 2024, <https://encord.com/blog/improve-ai-models-data-quality/#:~:text=Poor%20data%20quality%20can%20drastically,leading%20to%20harmful%20treatment%20recommendations>.

Active Campaign, ‘*Sales Discounts in Business - Definition & Benefits*’, iewed 20 August 2024, < https://www.activecampaign.com/glossary/sales-discount#:~:text=By%20offering%20discounts%20on%20certain,or%20looking%20for%20a%20bargain>.

Breiman, L, 2001, *Random forests Machine Learning*, Statistics Department University of California Berkeley.

Celebi, ME, Kingravi, HA & Vela, PA 2013, “A comparative study of efficient initialization methods for the k-means clustering algorithm,” *Expert Systems With Applications*, 40(1):200–210, https://www.sciencedirect.com/science/article/abs/pii/S0957417412008767?via%3Dihub.

Dabbura, I 2022, *K-Means Clustering: algorithm, applications, evaluation methods, and drawbacks*, *Medium*, https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a.

Ester, M., Kriegel, H.-P., Sander, J. & Xu, X., 1996, ‘A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise’, *Institute for Computer Science, University of Munic*, AAAI Press, p. 226 - 231.

Durgapal, A., 2023, ‘Data Preprocessing — Handling Duplicate Values and Outliers in a dataset’, *Medium*, viewed 18 August 2024, <https://medium.com/@ayushmandurgapal/handling-duplicate-values-and-outliers-in-a-dataset-b00ce130818e>.

Google Colab, viewed 20 August 2024, <https://colab.google/>.

Google Drive, viewed 20 August 2024, <https://drive.google.com/drive/home>.

Gupta, A., 2024, ‘Feature Selection Techniques in Machine Learning’, *Analytics Vidhya*, viewed 18 August 2024, <https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/>.

Han, J, 2000, ‘Data Mining: Concepts and Techniques’, *University of Illinois at Urbana-Champaign*, 3rd edn.

Jain, A.K., Murty, M.N. and Flynn, P.J., 1999, ‘Data clustering: a review’, *ACM Computing Surveys*, vol.31, iss.3, pp.264–323

James, G, Witten, D, Hastie, T & Tibshirani, R 2013, *An Introduction to Statistical Learning with Applications in R*, 2nd edn.

Jeffares, A 2021, *K-Means: A complete introduction - towards data science*, *Medium*, https://towardsdatascience.com/k-means-a-complete-introduction-1702af9cd8c. *KMEANS*, https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

Lee, S 2024, *Implementing K-Means clustering with Python for data analysis*, *Medium*, https://medium.com/@sunqlee1004/implementing-k-means-clustering-with-python-for-data-analysis-43c41f8390d1.

Li, Y & Wu, H 2012, "A clustering method based on K-Means algorithm," *Physics Procedia*, 251104–1109, https://www.sciencedirect.com/science/article/pii/S1875389212006220.

Liaw, A & Wiener M 2011, 'Classification and Regression by randomForest', *Forest*, vol. 23.

MacQueen, J., 1967, 'Some methods for classification and analysis of multivariate observations', vol.1, pp.281–298.

Microsoft, 2024, '*Power BI - Data Visualization | Microsoft Power Platform'*, viewed 20 August 2024, < https://www.microsoft.com/en-us/power-platform/products/power-bi>.

Olmez, R. 2024, 'Handling Missing Values in Data Science / Machine Learning Projects: Strategies and Practice', *Medium*, viewed 17 August 2024, <https://medium.com/@ramazanolmeez/handling-missing-values-in-data-science-machine-learning-projects-strategies-and-practice-42d7376ca94a#:~:text=1.-,Introduction,values%20or%20carefully%20complete%20them>.

Pathak, H 2024, *Scatter Plot Visualization in Python using matplotlib*, https://www.analyticsvidhya.com/blog/2024/02/scatter-plot-visualization-in-python-using-matplotlib/.

Pedregosa, F, Varoquaux, G, Gramfort, A, Michel, V, Thirion, B, Grisel, O, Blondel, M, Prettenhofer, P, Weiss, R, Dubourg, V, Vanderplas, J, Passos, A, Cournapeau, D, Brucher, M, Perrot, M & Duchesnay, E 2011, 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830.

Qi, J, Yu, Y, Wang, L, Liu, J & Wang, Y 2017, "An effective and efficient hierarchical K-means clustering algorithm," *International Journal of Distributed Sensor Networks*, 13(8):155014771772862, https://doi.org/10.1177/1550147717728627.

Rodriguez, M.Z., Comin, C.H., Casanova, D., Bruno, O.M., Amancio, D.R., Costa, L. da F. and Rodrigues, F.A., 2019, 'Clustering algorithms: A comparative approach', *PLOS ONE*, vol.14, iss.1.

Rousseeuw, P.J., 1987, ‘Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis’ *Journal of Computational and Applied Mathematics*, vol.20, pp.53–65.

S, A 2022, *K-Means clustering is used to identify and infer intrinsic groups within an unlabeled dataset. It is based on centroid-based clustering | Medium*, *Medium*, https://amansinganamala.medium.com/k-means-clustering-algorithm-1094514fce10.

Sachinsoni 2023, *The Art and Science of K-means Clustering: A Practical guide*, *Medium*, https://medium.com/@sachinsoni600517/the-art-and-science-of-k-means-clustering-a-practical-guide-e71b11638867.

Scatter plot — Matplotlib 3.9.2 documentation, https://matplotlib.org/stable/gallery/shapes_and_collections/scatter.html.

Scikit-learn, 2017, ‘DBSCAN — scikit-learn 1.5.1 documentation, viewed 20 August 2024, <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html>.

Shivanipickl, 2023, ‘What is Feature Scaling and Why Does Machine Learning Need It?’, *Medium*, viewed 20 August 2024, <https://medium.com/@shivanipickl/what-is-feature-scaling-and-why-does-machine-learning-need-it-104eedebb1c9#:~:text=Skewed%20data%20and%20outliers%20can,makes%20the%20model%20more%20robust>.

silhouette_score, https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html.

Varun Samarth (2023). Professional Certificate in Data Engineering with Microsoft Azure.[online] Emeritus India. Available at: https://emeritus.org/in/learn/data-science-classification-analysis/#disadvantages-of-classification-analysis [Accessed 23 Aug. 2024].

W3Schools.com, https://www.w3schools.com/python/python_ml_k-means.asp.

Wagavkar, S., 2023, ‘Introduction to The Correlation Matrix | Built In’, *Builtin*, viewed 18 August 2024, <https://builtin.com/data-science/correlation-matrix>.
Wickham, H., 2016, ‘Data Analysis’, pp.189–20.
Xu, R. & WunschII, D., 2005, ‘Survey of Clustering Algorithms’, *IEEE Transactions on Neural Networks*, vol.16, iss.3, pp.645–678