Analysis of Recommendation Systems

Author: Mehrdad Baradaran

Abstract:

This project focuses on enhancing the recommendation capabilities of Amazon by implementing two distinct methods: content-based methods and collaborative filtering methods. The goal is to design and implement at least one model for each method, utilizing these approaches to provide personalized and accurate suggestions to users. The dataset used for this task is related to over 2 million customer reviews and ratings of Beauty-related products on Amazon.

Amazon Ratings (Beauty Products) - Content-Based Methods

Introduction:

Amazon's recommendation system is a critical component of its success, leveraging customer reviews and purchase history. In this project, we delve into the Content-Based Methods, focusing on creating personalized shopping guides for Beauty Products on Amazon. The dataset in question comprises over 2 million customer reviews and ratings, offering a comprehensive view of user preferences and product details.

Dataset Description:

The dataset contains essential information such as unique UserId, product ASIN (Amazon's unique product identification code), Ratings ranging from 1-5 based on customer satisfaction, and the Timestamp of the rating. The rich dataset spans from May 1996 to July 2014, providing a historical context for the analysis. This dataset serves as the foundation for implementing content-based recommendation methods.

Content-Based Methods Approach:

Dataset Features:

- UserId
- ProductId (ASIN)
- Rating
- Timestamp Content-Based Methods involve creating personalized recommendations by analyzing user and product profiles. Features such as product descriptions, categories, and timestamps will be crucial in forming these profiles. The preprocessing steps include feature engineering and selection to enhance the model's ability to provide tailored suggestions to users.

Methodology:

- Feature Engineering: Utilizing product descriptions and categories to create informative features for content analysis.
- User and Product Profiles: Forming detailed profiles for users and products based on the engineered features.
- Model Implementation: Designing and implementing a content-based recommendation model to provide personalized suggestions.
- Evaluation: Assessing the model's effectiveness in handling the extensive product range on Amazon.

Innovation and Exploration: Exploration and testing of advanced algorithms will be conducted to uncover innovative insights, potentially leading to improved recommendation accuracy.

BigBasket Entire Product List - Collaborative Filtering Methods

Introduction:

BigBasket, India's largest online grocery supermarket, has become a staple for many consumers. In this project, we aim to enhance the recommendation capabilities of BigBasket by implementing Collaborative Filtering Methods. As the e-commerce industry continues to thrive, understanding user preferences becomes paramount, and collaborative filtering offers a team-based approach to achieving this goal.

Dataset Description:

The dataset comprises 10 attributes, including product titles, categories, brands, sale prices, market prices, and ratings. These attributes provide valuable insights into user preferences and product characteristics. The creation of a "discount" feature during feature engineering enhances the dataset's richness for a more comprehensive analysis.

Collaborative Filtering Methods Approach:

Dataset Features:

- Index
- Product
- Category
- Sub_category
- Brand
- Sale_price
- Market_price
- Type
- Rating
- Description Collaborative Filtering Methods consider the preferences of users with similar tastes, generating suggestions based on observed patterns in user behavior. The "discount" feature adds an additional layer for understanding consumer choices. Feature engineering and collaborative filtering techniques will be employed to provide accurate recommendations.

Methodology:

- Feature Engineering: Creating the "discount" feature and exploring its impact on user preferences.
- User Preferences: Analyzing user behavior and preferences based on ratings and product attributes.
- Collaborative Filtering: Implementing a collaborative filtering model to generate suggestions based on similar user tastes.
- Model Evaluation: Assessing the model's adaptability to the dynamic online grocery store with a diverse product range.

Innovation and Exploration: The project encourages exploration and testing of advanced collaborative filtering algorithms for potential improvements in recommendation accuracy.

Insights from Amazon Dataset

Number of Records and Columns:

- The dataset contains 2,023,070 records and 5 columns.

Column Information:

- UserId: Unique identification for each customer (1,210,271 unique values).
- ProductId: Unique identification code for each product (249,274 unique values).
- Rating: Ratings ranging from 1 to 5.
- Timestamp: Time at which the rating was recorded (4,231 unique values).
- user_id: Additional user identification (1,210,271 unique values).

Missing Values:

- There are no missing values in any of the columns.

Data Types:

- UserId and ProductId are of object type, Rating is of float64 type, and Timestamp and user_id are of int64 type.

Summary Statistics:

- Rating: Mean rating is approximately 4.15, with a standard deviation of 1.31. Ratings range from 1 to 5.
- Timestamp: Mean timestamp corresponds to approximately 1360 seconds since UNIX epoch. The dataset spans a considerable timeframe.
- user_id: Mean user_id is around 503,609, with a standard deviation of 353,574.

Correlation Matrix:

- Rating and Timestamp: Negligible correlation (0.00033).
- Rating and user_id: Weak negative correlation (-0.03).
- Timestamp and user_id: Moderate positive correlation (0.26).

Number of Duplicated Rows:

- There are no duplicated rows in the dataset.

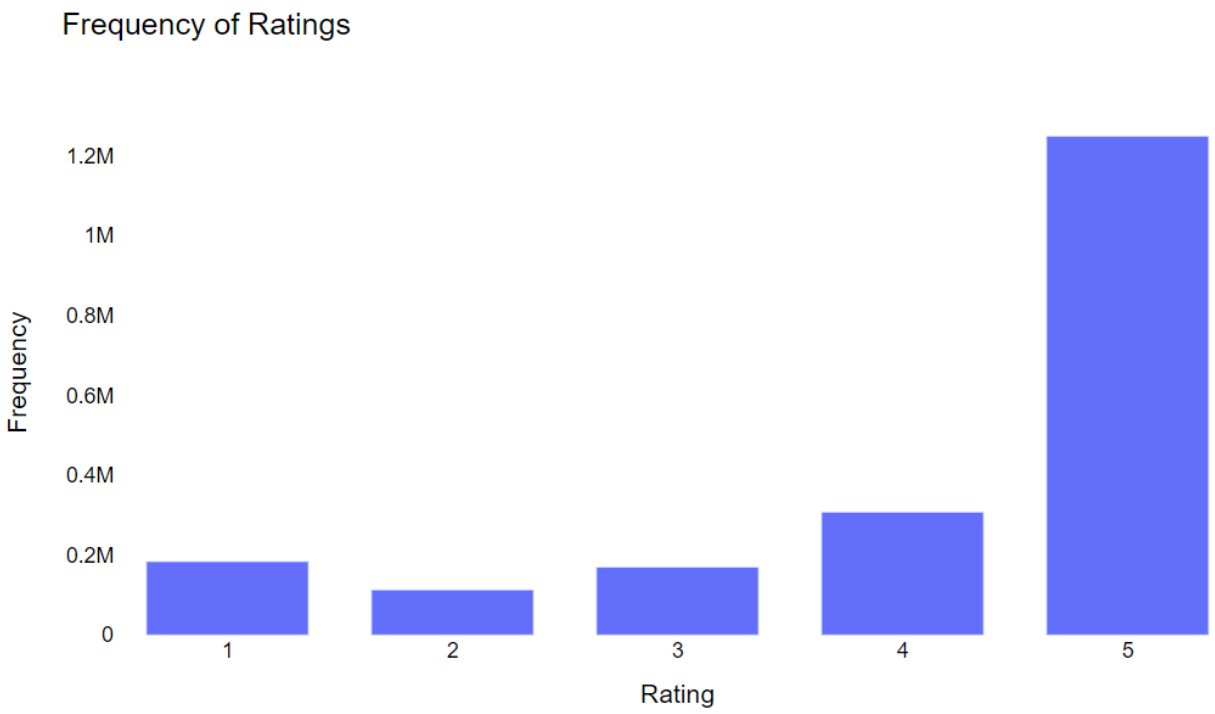
Interpretation

- Ratings Distribution: Ratings are generally positive, with a mean rating close to 4.15. Most ratings fall within the 4 to 5 range.
- Time Span: The dataset covers a substantial time span, indicated by the wide range of timestamps.
- User Identification: The user_id and UserId columns seem to represent similar information, but user_id has a more consistent numeric format. Further investigation is needed to understand the relationship between the two.
- Correlation Analysis: The weak correlation between Rating and user_id suggests that user-specific characteristics may not heavily influence the ratings. The positive correlation between Timestamp and user_id indicates a potential temporal pattern in user activities.
- Data Integrity: No missing values or duplicates were found, indicating a well-maintained dataset.

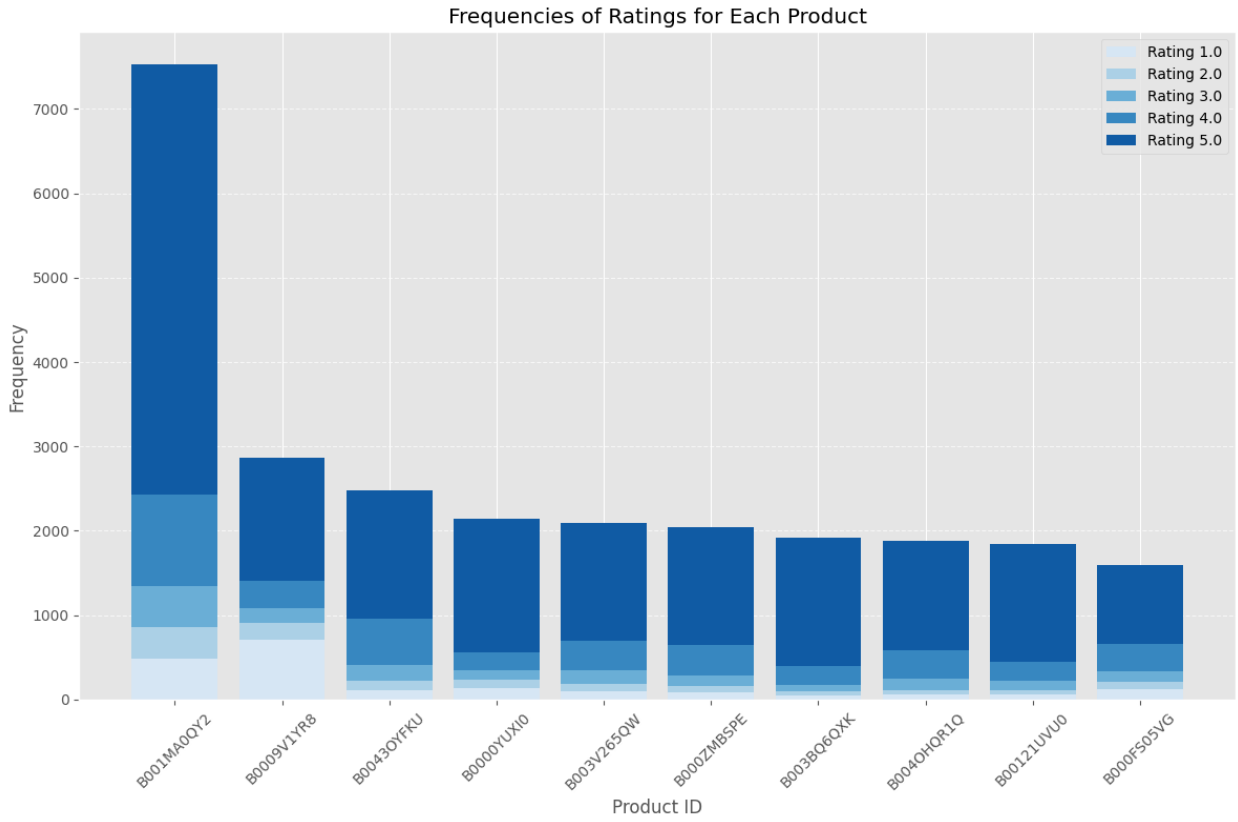
	feature	unique_count	unique_values	data_type
UserId	UserId	1210271	[A39HTATAQ9V7YF, A3JM6GV9MNOF9X, A1Z513UWSAAO0...	object
ProductId	ProductId	249274	[0205616461, 0558925278, 0733001998, 073710447...	object
Rating	Rating	5	[5.0, 3.0, 4.0, 1.0, 2.0]	float64
Timestamp	Timestamp	4231	[1369699200, 1355443200, 1404691200, 138257280...	int64
user_id	user_id	1210271	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,...	int64

The provided dataset contains information about various products, including their unique identifiers ('ProductId') and associated ratings ('Rating'). To gain insights into the overall satisfaction level for each product, an analysis was conducted to calculate the average rating.

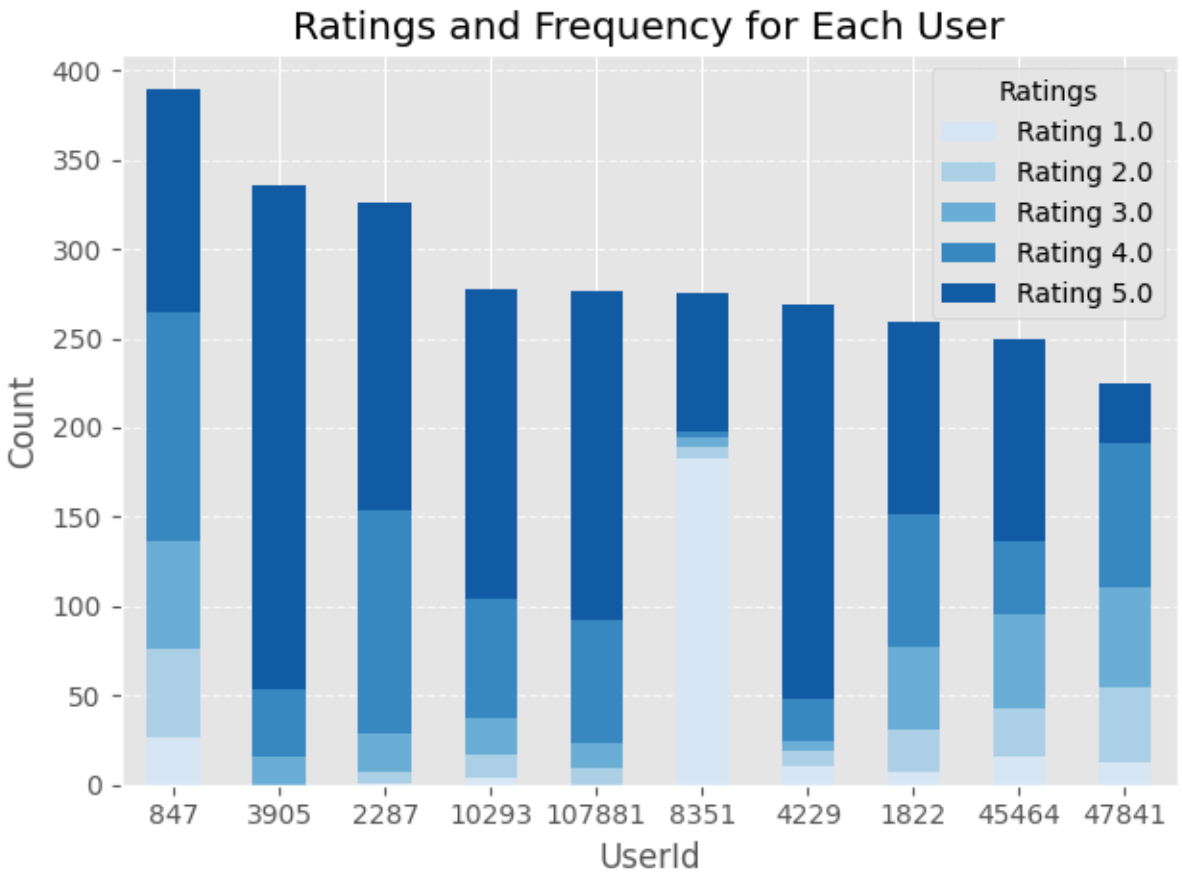
The resulting table presents the 'ProductId' alongside its corresponding average rating, summarizing the collective feedback received. The table showcases the diversity of products within the dataset, ranging from those consistently receiving high ratings to those with more mixed reviews.


Frequency of Ratings
1.2M
1M
0.8M
0.6M
0.4M
0.2M
0
Frequency
1
2
3
4
5
Rating


Average Rating Distribution
100k
80k
60k
40k
20k
0
Count
1
1.5
2
2.5
3
3.5
4
4.5
5
Rating



user_id	index	Rating 1.0	Rating 2.0	Rating 3.0	Rating 4.0	Rating 5.0	Rating Frequency
847	847	27	49	60	129	124	389
3905	3905	0	0	16	38	282	336
2287	2287	1	6	22	125	172	326
10293	10293	4	13	20	67	174	278
107881	107881	0	9	14	69	184	276
8351	8351	183	6	6	3	77	275
4229	4229	10	9	6	23	221	269
1822	1822	7	24	46	75	107	259
45464	45464	16	27	52	41	113	249
47841	47841	13	42	56	80	34	225


Ratings and Frequency for Each User
Ratings
Rating 1.0
Rating 2.0
Rating 3.0
Rating 4.0
Rating 5.0
Count
400
350
300
250
200
150
100
50
0
847
3905
2287
10293
107881
8351
4229
1822
45464
47841
UserId

Code Overview:

1. **Data Preparation:**
    - Limited the dataset to 30,000 rows and sorted by user_id.
    - Created a user-item matrix from the Amazon ratings dataset.
    - Applied TruncatedSVD for dimensionality reduction.
2. **Recommendation Functions:**
    - **recommend_products**: Recommends products for a user using collaborative filtering based on SVD.
    - **generate_recommendations**: Generates recommendations for a user using item-item similarity.
3. **SVD Implementation:**
    - Used TruncatedSVD for matrix factorization, reducing the user-item matrix to 100 dimensions.
4. **Item-Item Similarity:**
    - Calculated the item-item similarity matrix (cosine similarity) based on the user-item matrix.
5. **Recommendation Generation:**
    - Generated recommendations for a user using both SVD and item-item similarity.
6. **Example Usage:**
    - Demonstrated how to recommend products for a user, considering the top N recommendations.

Example :

```
Recommendations for User 3905: Index(['B000052YOL', 'B000055Z3C', 'B000052YM7'], dtype='object', name='ProductId')
```

BigBasket Entire Product List - Collaborative Filtering Methods

Checking the shape of the data i.e. the number of rows and columns present in the data.

Number of data points : 27555
Number of features/variables: 10

```
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   index         27555 non-null  int64
 1   product       27554 non-null  object
 2   category      27555 non-null  object
 3   sub_category  27555 non-null  object
 4   brand         27554 non-null  object
 5   sale_price    27555 non-null  float64
 6   market_price  27555 non-null  float64
 7   type          27555 non-null  object
 8   rating        18929 non-null  float64
 9   description   27440 non-null  object
```

Of these 10 features, we will be using only 6 features for our Model.

category - Category into which product has been classified-11 unique categories sub_category - Subcategory into which product has been kept-90 unique sub_categories brand - Brand of the product-2314 unique brands type- Type into which product falls-426 unique types sale_price - Price at which product is being sold on the site description- Description of the product (in detail) In addition to the above 6 features we will use newly built feature discount on product(i.e, market price-sell price/market price) , to rank the order of Recommended products

After removing missing 'product','brand','description' datapoints reduced from 27555 to 27439 (99.57 % data retained)

Split data into train and test with test size of 0.2 and do Exploratory analysis on Train data features After Split

Train data :21951 data points

Test data : 5488 data points

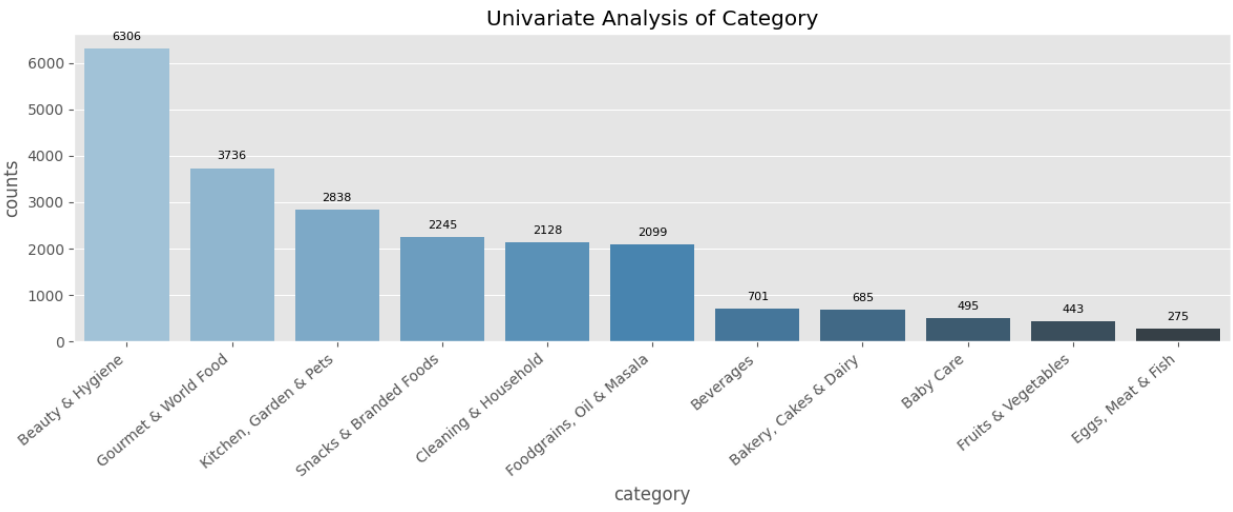
Out of 21951 train data, there are 19301 Unique values for product and top most occuring product is 'Turmeric Powder/Arisina Pudi' with 21 times

```
count                            21951
unique                           19301
top       Turmeric Powder/Arisina Pudi
freq                                21
```

	category	counts
0	Beauty & Hygiene	6306
1	Gourmet & World Food	3736
2	Kitchen, Garden & Pets	2838
3	Snacks & Branded Foods	2245
4	Cleaning & Household	2128
5	Foodgrains, Oil & Masala	2099
6	Beverages	701
7	Bakery, Cakes & Dairy	685
8	Baby Care	495
9	Fruits & Vegetables	443
10	Eggs, Meat & Fish	275

	category	%	cum_%
0	Beauty & Hygiene	0.287276	Beauty & Hygiene
1	Gourmet & World Food	0.170197	Beauty & HygieneGourmet & World Food
2	Kitchen, Garden & Pets	0.129288	Beauty & HygieneGourmet & World FoodKitchen, G...
3	Snacks & Branded Foods	0.102273	Beauty & HygieneGourmet & World FoodKitchen, G...
4	Cleaning & Household	0.096943	Beauty & HygieneGourmet & World FoodKitchen, G...
5	Foodgrains, Oil & Masala	0.095622	Beauty & HygieneGourmet & World FoodKitchen, G...
6	Beverages	0.031935	Beauty & HygieneGourmet & World FoodKitchen, G...
7	Bakery, Cakes & Dairy	0.031206	Beauty & HygieneGourmet & World FoodKitchen, G...
8	Baby Care	0.022550	Beauty & HygieneGourmet & World FoodKitchen, G...
9	Fruits & Vegetables	0.020181	Beauty & HygieneGourmet & World FoodKitchen, G...
10	Eggs, Meat & Fish	0.012528	Beauty & HygieneGourmet & World FoodKitchen, G...

- 28.72% of products fall in category of Beauty & Hygenie
- Gourmet & World Food occupies 17.01 % of total products
- followed by Kitchen, Garden & Pets which has a 12.92% share
- Snacks & Branded Foods has 10.2% share
- Rest of the categories have less than 1 percent share each in total products list



	sub_category	counts
0	Skin Care	1814
1	Health & Medicine	927
2	Hair Care	830
3	Fragrances & Deos	822
4	Storage & Accessories	821
5	Bath & Hand Wash	779
6	Masalas & Spices	700
7	Crockery & Cutlery	694
8	Snacks, Dry Fruits, Nuts	677
9	Men's Grooming	648

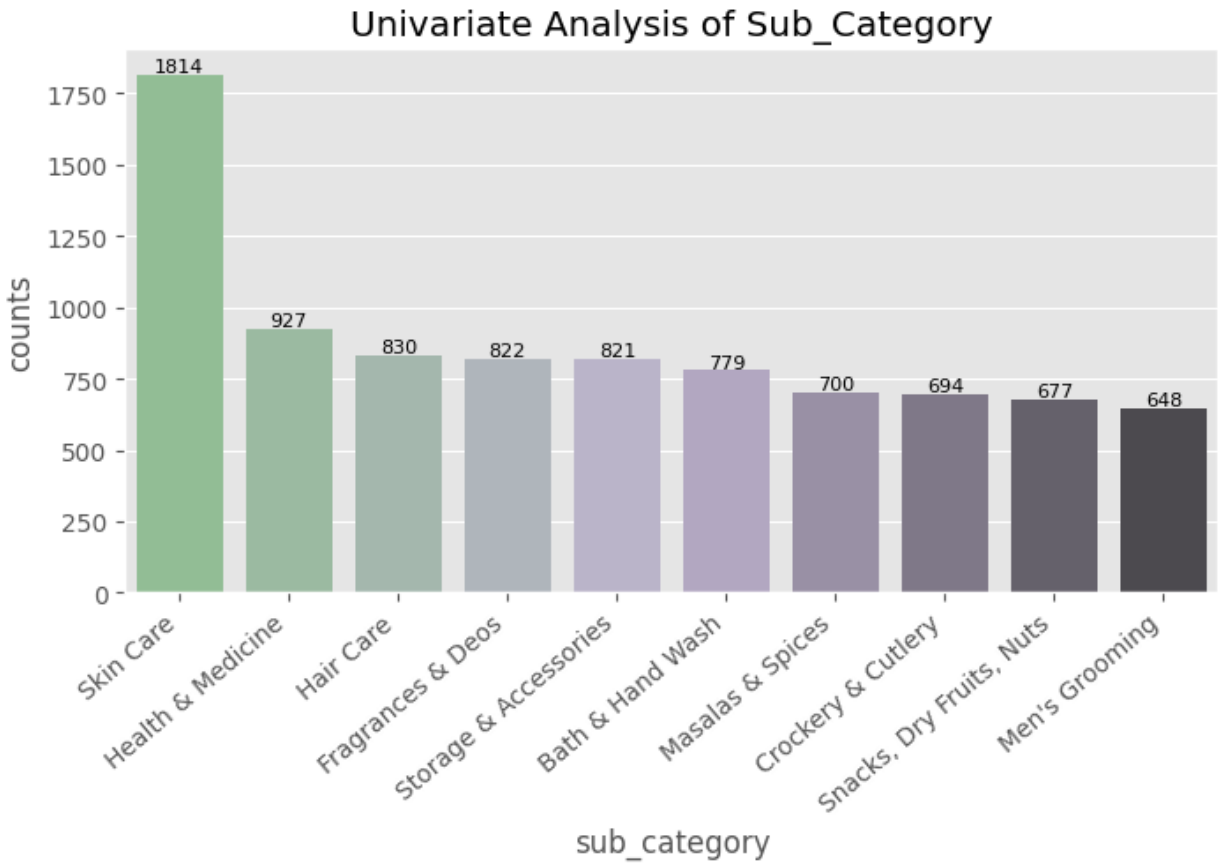
Sub-Category:

There are 90 unique sub-categories and Skin care is top occurring with 1814 times

- The top four major sub_categories are as follows
- Skin Care- 8.26%
- Health & Medicine- 4.22%
- Hair Care-3.78%
- Fragrances & Deos — 3.74%

Type:¶

- There are 423 unique types of product
- Face care is the top type of product with 1181 counts- 5.38%

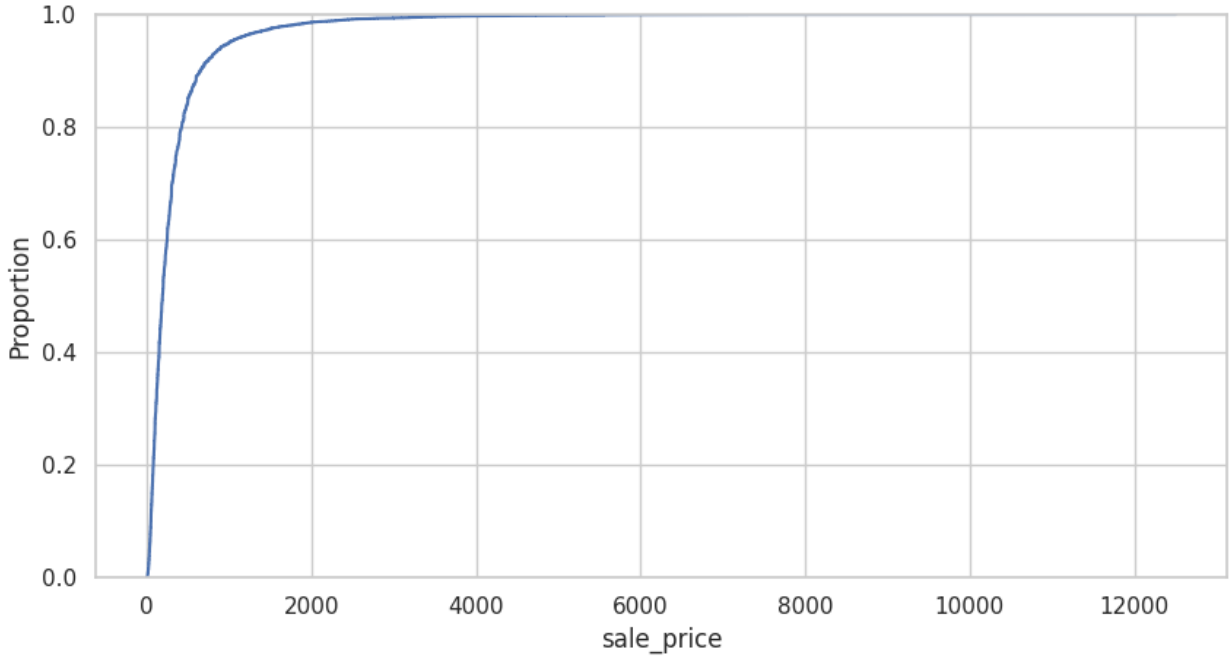



```
brand
Fresho              2.305134
bb Royal            1.954353
BB Home             1.594460
DP                  0.883787
Fresho Signature    0.605895
bb Combo            0.592228
Amul                0.546672
GoodDiet            0.537561
Dabur               0.510227
INATUR              0.505672
```

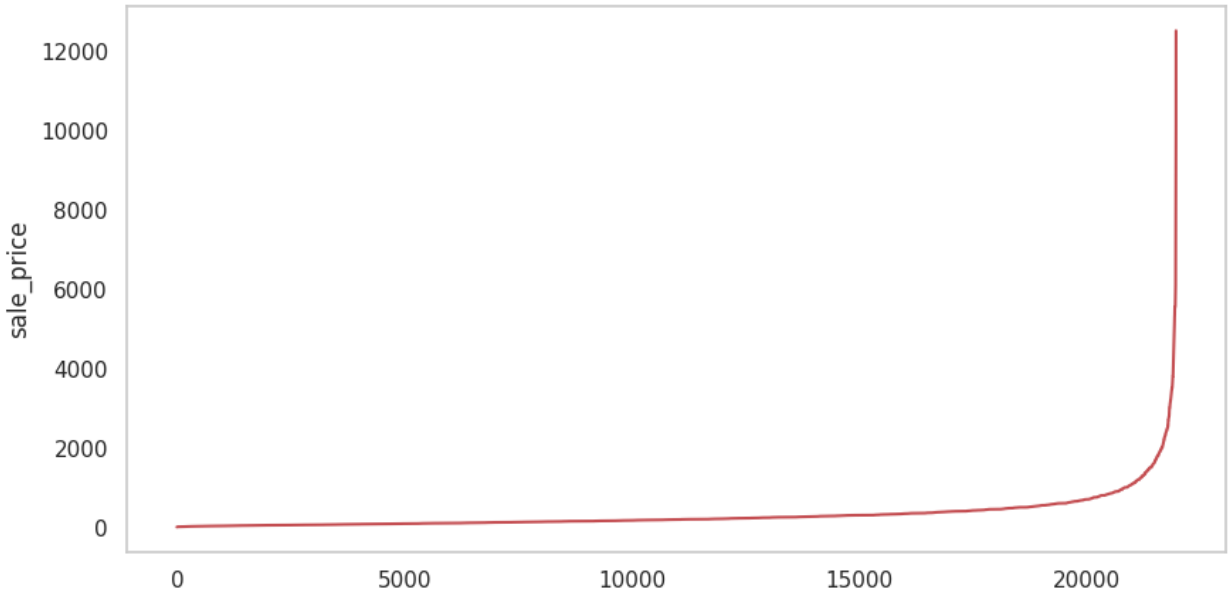
- There are 2195 unique brands
- The brand which has highest number of products is Fresho - 506 counts - 2.3%

type	
Face Care	5.380165
Ayurveda	2.022687
Men's Deodorants	1.854130
Shampoo & Conditioner	1.658239
Containers Sets	1.562571
Glassware	1.489682
Bathing Bars & Soaps	1.403125
Blended Masalas	1.394014
Gourmet Tea & Tea Bags	1.257346
Body Care	1.248235

- There are 423 unique types of product
- Face care is the top type of product with 1181 counts- 5.38%



99% of products sale price is below 2000 in train data



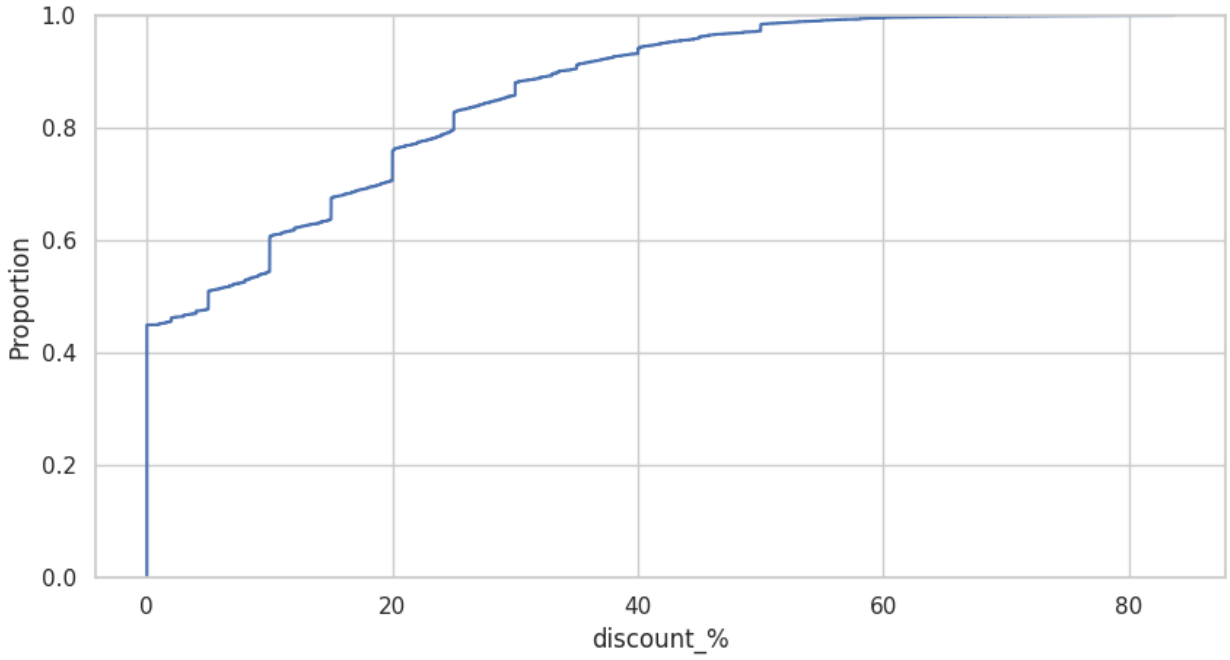
- From the above line plot there is a steep change at sale_price 2000(Inflection point)
- So will remove products whose sale_price is above 2000

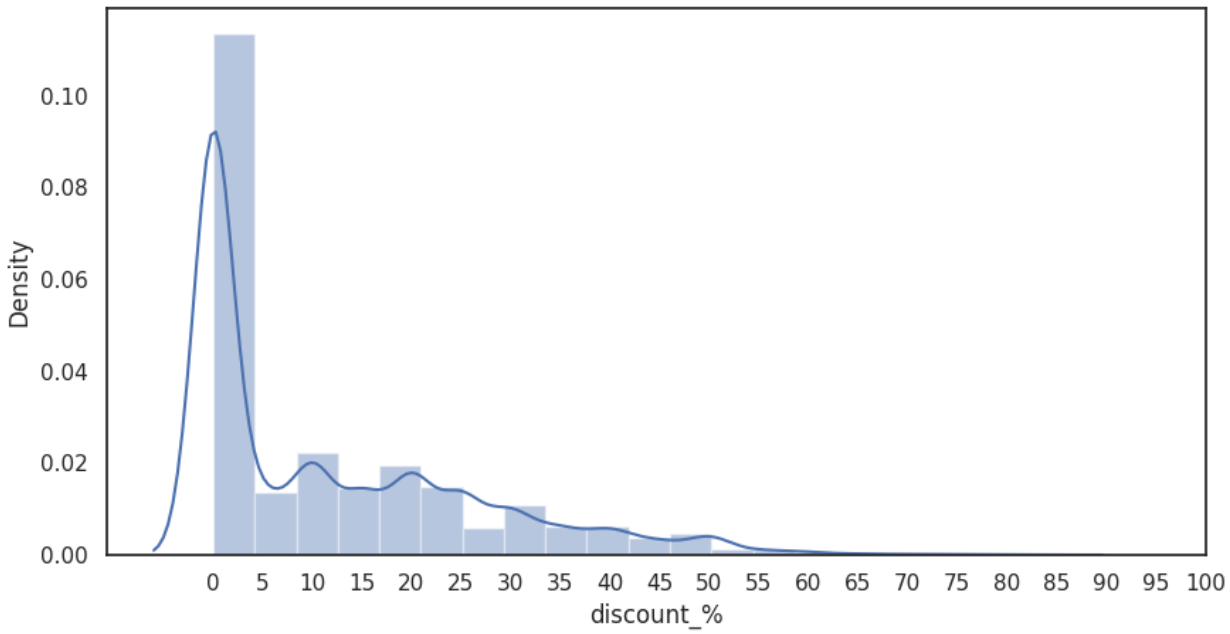
Datapoints reduced to 21648 (99.98% of original data)in train data from 21951 after removing products with sale price greater than 2000

Text preprocessing done on categorical features ; category,sub_category,brand ,type

Same way done for sub_category,type,brand columns

For Description text preprocessing is also done . nltk stop words are used to remove stop words in description





- Approx 46% of products have No discount(zero discount)
- Maximum discount for a product is 60% which is only for a small potion of total products(i.e,~1%)
- ~ 30% of products have discount from 1 to 20%
- ~ 15% of products have discount in range of 21-40% of market_price

Basic Feature Extraction and its importance:

- Discount_% feature is built by using sale_price and market_price from the given data for model building (i.e, market price-sell price/market price)

By this we can have discount range of products in finding similar products of the queried one

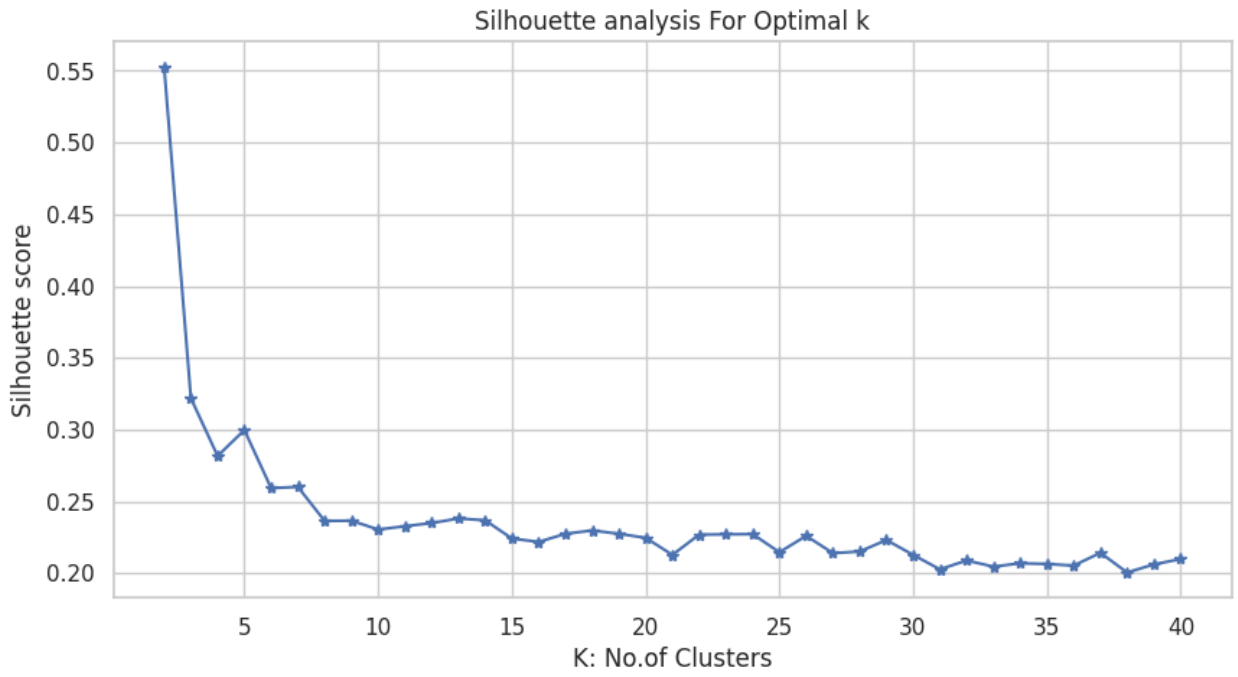
Advanced Feature Extraction and it's Importance:

As part of Advanced Feature Extraction , we will do clustering for the train data ;add cluster label as a feature. Also save the cluster means as pickle file to assign a query point/unseen point to the nearest cluster mean and label it for that cluster

For clustering we will be using Scaled sale_price (standardized using min-max scaler), discount_% and sentiment scores obtained from preprocessed description text.

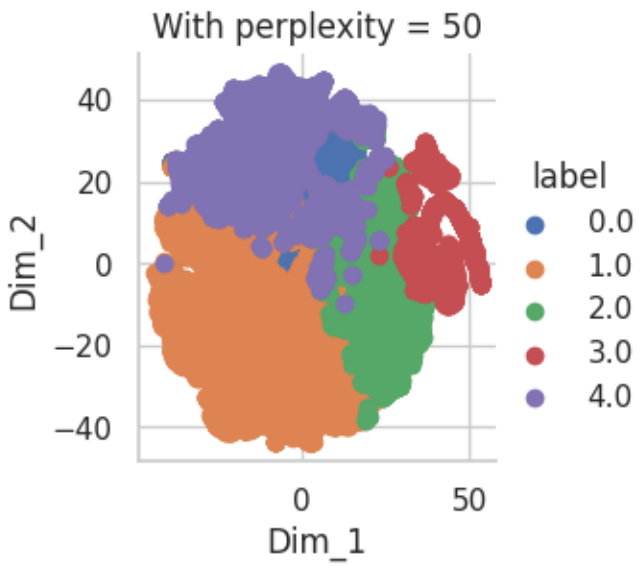
We compute sentiment scores for description text which we will use in clustering analysis

For a good clustering silhouette score should be as close as to 1, but in our case the K with maximum silhouette score is 2, for optimum number of clusters will choose 5 as best_k by elbow method



We cluster our data into 5 groups as per silhouette score and elbow method and label them accordingly

T-SNE visualization of cluster analysis for better understanding



we vectorize preprocessed description text with tfidf weighted W2V

tfidf_W2V vectorization of preprocessed description feature

We use TF-IDF weighted Word2Vec representation of description text for featurization of text data to numerical data as ML models take numerical inputs

We use GloVe: Global Vectors (300 dimensions) for Word Representation for our featurization The link for downloading pretrained glove word embeddings is: https://nlp.stanford.edu/projects/glove/

The TF-IDF weighted Word2Vec of all preprocessed description text of train data are stored in a list for using it in computing item-item cosine similarity

Label encoding categorical features 'category', 'sub_category', 'brand', 'type'

While Label encoding categorical columns ,fit is done on train data and for any unseen label in test data ,will be labelled as (n+1)th ,for n unique values in train data. Encoders returned from the functioned are stored for encoding test data/query data points

we build a matrix from train data to compute cosine similarity

Model Building : Item- Item based Collaborative Filtered recommender system

- will give attributes of query product
- query product should have category,sub_category,brand,type,description,sale_price,market_price
- using the above features will do encoding,vectorizing ,scale sale price to train data,compute discount_%, calculate sentiment scores, and assign to nearest cluster

The text preprocessing of categorical features (category,sub_category,brand,type) and text of description of query product is done in same way as for train data by defining functions for same

The query data point is assigned to nearest(euclidean distance) one among the five cluster means which are obtained by using train data for optimal K in Advance Feature Engineering part

Now after the text preprocessing, building advance features, of query point we compute cosine similarity of the query product with other products in train data and recommend the similar products with highest cosine similarity

We define get_similar_products function for:

The function takes the query point and does the following steps

- Checks whether the query point has all the needed columns , if not will give a warning message and function exits
- After checking for any missing columns , it does preprocessing for categorical columns(category,sub_category,brand,type) and then encode with transform using label encoders which are fit on train data
- Next check is done on sale_price whether the sale price is in +-15% of maximum sale price of the same brand in train data , if not function exits with a message -salepricecheck function
- Preprocessing of description text of query point using preprocess_description function
- discount_% is computed after sale_price check
- Get sentiment scores of preprocessed query's description and scaling sale price with min and max of train data
- Assigning cluster label using cluster means from train data
- Now stacking all vectors and features of query product and compute cosine similarity with products in train data
- Returning top n similar product based on cosine similarity

checking for some random query products

```
query
```

	product	category	sub_category	brand	type	description	sale_price	market_price
0	Kantan Watermelon Slice	Gourmet & World Food	Chocolates & Biscuits	Fini	Marshmallow, Candy, Jelly	Fini Fizzy Watermelon Slices containing fizzy ...	110.0	110.0

The searched/Queried product is:
Kantan Watermelon Slice

Top 10 Similar products for "**Kantan Watermelon Slice**" are:
==========

8506 : Assam Black Tea
Cosine Similarity with queried product is : 0.999965
Discount %: 0.0

13075 : Mandarin Citrus Handmade Soap
Cosine Similarity with queried product is : 0.999964
Discount %: 0.0

16625 : Insta Fair & Glow Fairness Cream
Cosine Similarity with queried product is : 0.999964
Discount %: 0.0

5774 : Artisan Bread Flour
Cosine Similarity with queried product is : 0.999964
Discount %: 0.0

5895 : Whitening Smooth Skin Women Deodorant For 48h Protection
Cosine Similarity with queried product is : 0.999961
Discount %: 0.0

18589 : Bio Farm Organic Fertiliser
Cosine Similarity with queried product is : 0.999961
Discount %: 0.0

21486 : Almond & Rose Soap
Cosine Similarity with queried product is : 0.999961
Discount %: 0.0
---

970 : Protein Packed Nachos - Peri Peri Masala
Cosine Similarity with queried product is : 0.999958

Discount %: 0.0

10965 : Total Effects Whip - UV SPF 30
Cosine Similarity with queried product is : 0.999958
Discount %: 0.0

21605 : Fields of Gold - Organic Flaxseeds
Cosine Similarity with queried product is : 0.999958
Discount %: 0.0

Similar products for: "Kantan Watermelon Slice"

	Similar products
0	Assam Black Tea
1	Mandarin Citrus Handmade Soap
2	Insta Fair & Glow Fairness Cream
3	Artisan Bread Flour
4	Whitening Smooth Skin Women Deodorant For 48h Protection
5	Bio Farm Organic Fertiliser
6	Almond & Rose Soap
7	Protein Packed Nachos - Peri Peri Masala
8	Total Effects Whip - UV SPF 30
9	Fields of Gold - Organic Flaxseeds

The similar products shown have cosine similarity of 0.99 range