Fashion Trend Analysis & Recommendation System

A project paper submitted in partial fulfillment of the requirements for the Degree of Master's of Science (Engg.) in Information & Communication Technology

Submitted by-

ID No: 22209047

Registration No: 22209047

Session: 2021-22

Department of Information & Communication Technology

Comilla University

Cumilla, Bangladesh

Submission: February, 2025

Fashion Trend Analysis & Recommendation System

A project paper submitted in partial fulfillment of the requirements for the Degree of Master's of Science (Engg.) in Information & Communication Technology

Submitted by-

ID No: 22209047

Registration No: 22209047

Session: 2021-22

Department of Information & Communication Technology

Comilla University

Cumilla, Bangladesh

Submission: February, 2025

Abstract-

In order to ease the issue of information overload, which has become a potential concern for many Internet users, it is necessary to filter, prioritize, and efficiently distribute pertinent information on the Internet, where the quantity of options is overwhelming. One of the newest technological developments, big data has the potential to drastically alter how companies analyze and turn consumer behavior into insightful knowledge. Decision trees are also effective tools for data analysis. In order to give users individualized content and services, recommender systems scan through vast amounts of dynamically created data. Fashion recommendation systems are advanced technological solutions designed to provide personalized clothing suggestions by leveraging artificial intelligence, machine learning, and deep learning techniques. This advanced system aims to simplify fashion decision-making by offering intelligent, context-aware recommendations tailored to individual user preferences. The primary objective is to help users find the most suitable and complementary clothing items based on their existing wardrobe, personal preferences, and specific occasion requirements.

List of Chapters

List of Figures

List of Tables

List of Chart & Graph

Introduction

The fashion industry is experiencing a transformative phase driven by data science and machine learning technologies. This project presents a comprehensive Fashion Trend Analysis and Recommendation System designed to revolutionize how fashion retailers understand customer preferences and optimize product recommendations. Finding all the hidden information in the vast amount of data gathered from a wide range of sources is currently a brand's largest difficulty. Here, the Big Data Analytics enters the scene. Customer analytics, often known as customer behavior analysis, is one of them [1]. By enabling businesses to forecast consumer behavior, customer analytics helps transform large data into significant value. This improves sales, market optimization, inventory planning, fraud detection, and many other applications. [2]

An issue of information overload may arise as a result of a clothes and accessory brand's rapid expansion and the volume of customers. The need for recommendation systems has never been higher as a result of this. Recommendation systems' job is to forecast users' potential future interests and likes based on user data and preferences. Web entrepreneurs at the vanguard of the information revolution recognized the enormous potential of recommender system research, which lies at the intersection of science and socioeconomic life. Although recommendation was initially a discipline dominated by computer scientists, it now attracts interest from mathematicians, physicists, and psychologists as well as requires contributions from a variety of fields [3]. In order to address the issue of information overload, recommendation systems filter important information fragments from vast amounts of dynamically generated data based on user preferences, interests, or observed behavior about an item [4]. Based on a person's profile, a recommender system can determine whether or not a specific user would prefer an item. Both service providers and users benefit from recommender systems [5] [6]. They lower the transaction costs associated with locating and choosing products in an online marketplace [7] [8] is becoming increasingly

significant for modern consumers due to the seemingly endless options of constantly evolving styles, which has piqued the interest of the internet retail sector [9].

Here, in our project we have used the datasets of H&M Brand. They have some precised datasets of their product articles, customers' details, transaction details and a folder consisting of subfolders with images in them. Since, there are many different strategies that can be used for the recommendation system itself, we have used K-means clustering, an unsupervised learning. We have divided the customers into few clusters according to some of the significant features from the datasets.

Literature Review

Customer surveys and statistical analysis of historical data have historically been the two main methods used in customer behavior or trend analysis. By applying mathematical methods to historical statistics like as policy details, demographics, and claims history, statistical analysis enables merchants to spot general trends and patterns within their clientele. Correlations between variables, such as age groups and risk profiles or geographic regions and claim frequency, can be found using this method. However, in the current environment, which is marked by an explosion of data and changing client interactions, these approaches have serious drawbacks.

The intrinsic rigidity of conventional statistical models is another drawback [10]. Although these models are good at seeing general patterns, they have trouble capturing the subtleties and complexity of contemporary consumer behavior. Customer connections are dynamic and impacted by a wide range of changing requirements, risk perceptions, and digital touchpoints; they are no longer limited to static demographics. These complex dynamics may be missed by traditional statistical models, which frequently rely on predetermined assumptions, providing an insufficient picture of the clientele.

The other cornerstone of conventional customer behavior analysis is customer surveys, which provide direct input from consumers on their preferences for products, risk perceptions, and levels of satisfaction [11]. However, bias and sample size constraints can affect surveys. It can be difficult to reach a representative sample of the whole customer base, and self-reported data does not always accurately reflect behavior. Furthermore, surveys frequently fall short in capturing the implicit biases and subconscious motivations that have a big impact on consumer decision-making. The subjectivity of human responses and the possibility of social desirability bias, in which respondents may modify their comments to fit what they believe to be socially acceptable, further compound of the limit of surveys.

An increasing amount of research has been prompted by the insurance industry's rising interest in using AI to analyze consumer behavior. Studies already conducted have shown how AI has the ability to completely transform risk management and customer interaction. Initial studies has concentrated on using machine learning methods to forecast client attrition, with encouraging outcomes [12]. To find important determinants of customer attrition, these researches have used a

range of classification methods, including support vector machines, decision trees, and logistic regression. Although these models have attained a respectable level of accuracy, the intricacy of consumer behavior and the accessibility of thorough data frequently restrict their prediction power.

More recent studies have explored the use of AI for consumer segmentation in greater detail, using clustering algorithms to create discrete client groups according to shared traits and inclinations. These investigations have demonstrated the promise of AI-driven customer segmentation to give insurers more precise control over their marketing campaigns and product offerings. Researchers have also looked into using AI for fraud detection, using anomaly detection methods to find policyholders and claims that seem suspect.

The application of increasingly complex methods, including deep learning, in the insurance industry has gained attention as AI research progresses. A type of machine learning called deep learning has shown impressive results in a range of fields, encompassing natural language processing, predictive analytics, and picture and speech recognition. It has enormous potential for customer behavior analysis in insurance due to its capacity to extract intricate patterns from enormous volumes of data.

Methodology

Dataset Overview:

For the purpose of proper analysis and dividing the customers into several clusters, we have collected datasets and a folder of images from H&M brand. The metrics of data collected from them are shown in the table below:

Name of the Dataset	Instances	Features
Articles	105542	25
Transactions	1371980	7
Customers	31788324	5

Table-1: List of the Datasets

Name of the Folder	Number of Subfolders	Number of Total Images
Image	88	105110

Table 2: Image Folder Details

Before pre-processing and putting the dataset into the algorithm, we analyzed the article and customer dataset to the depth to find out the current trend and how the customer behavior changes according to some valid factors.

There are 25 features in article dataset, 5 features in customer dataset and 7 features in transaction dataset.

	article_id	product_code	prod_name	product_type_no	product_type_name	product_group_name	graphical_appearance_no	graphical_appearance_name	colour_
0	108775015	108775	Strap top	253	Vest top	Garment Upper body	1010016	Solid	
1	108775044	108775	Strap top	253	Vest top	Garment Upper body	1010016	Solid	
2	108775051	108775	Strap top (1)	253	Vest top	Garment Upper body	1010017	Stripe	
3	110065001	110065	OP T-shirt (Idro)	306	Bra	Underwear	1010016	Solid	
4	110065002	110065	OP T-shirt (Idro)	306	Bra	Underwear	1010016	Solid	

Figure-1: Article Dataset

customer_id	FN	Active	club_member_status	fashion_news_frequency	age	postal_code
00000dbacae5abe5e23885899a1fa44253a17956c6d1c3...	NaN	NaN	ACTIVE	NONE	49.0	52043ee2162cf5aa7ee79974281641c6f11a68d276429a...
0000423b00ade91418cceaf3b26c6af3dd342b51fd051e...	NaN	NaN	ACTIVE	NONE	25.0	2973abc54daa8a5f8ccfe9362140c63247c5eee03f1d93...
000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	NaN	NaN	ACTIVE	NONE	24.0	64f17e6a330a85798e4998f62d0930d14db8db1c054af6...
00005ca1c9ed5f5146b52ac8639a40ca9d57aeff4d1bd2...	NaN	NaN	ACTIVE	NONE	54.0	5d36574f52495e81f019b680c843c443bd343d5ca5b1c2...
00006413d8573cd20ed7128e53b7b13819fe5cfc2d801f...	1.0	1.0	ACTIVE	Regularly	52.0	25fa5ddee9aac01b35208d01736e57942317d756b32ddd...
...	...	...	...	...	...	...
ffffbbf78b6eaac697a8a5dfbfd2bfa8113ee5b403e474...	NaN	NaN	ACTIVE	NONE	24.0	7aa399f7e669990daba2d92c577b52237380662f36480b...
ffffcd5046a6143d29a04fb8c424ce494a76e5cdf4fab5...	NaN	NaN	ACTIVE	NONE	21.0	3f47f1279beb72215f4de557d950e0bfa73789d24acb5e...
ffffcf35913a0bee60e8741cb2b4e78b8a98ee5ff2e6a1...	1.0	1.0	ACTIVE	Regularly	21.0	4563fc79215672cd6a863f2b4bf56b8f898f2d96ed590e...
ffffd7744cebcf3aca44ae7049d2a94b87074c3d4ffe38...	1.0	1.0	ACTIVE	Regularly	18.0	8892c18e9bc3dca6aa4000cb8094fc4b51ee8db2ed14d7...
ffffd9ac14e89946416d80e791d064701994755c3ab686...	NaN	NaN	PRE-CREATE	NONE	65.0	0a1a03306fb2f62164c2a439b38c0caa64b40deaae8687...

Figure-2: Customer Dataset

	t_dat	customer_id	article_id	price	sales_channel_id
0	2018-09-20	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	663713001	0.050831	2
1	2018-09-20	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	541518023	0.030492	2
2	2018-09-20	00007d2de826758b65a93dd24ce629ed66842531df6699...	505221004	0.015237	2
3	2018-09-20	00007d2de826758b65a93dd24ce629ed66842531df6699...	685687003	0.016932	2
4	2018-09-20	00007d2de826758b65a93dd24ce629ed66842531df6699...	685687004	0.016932	2
...	...	...	...	...	...

Figure-3: Transaction Dataset

Analysis:

If we look into the 'product_group' feature of the article dataset, we can see that upper body product group is the most popular and Fun product group is the least popular

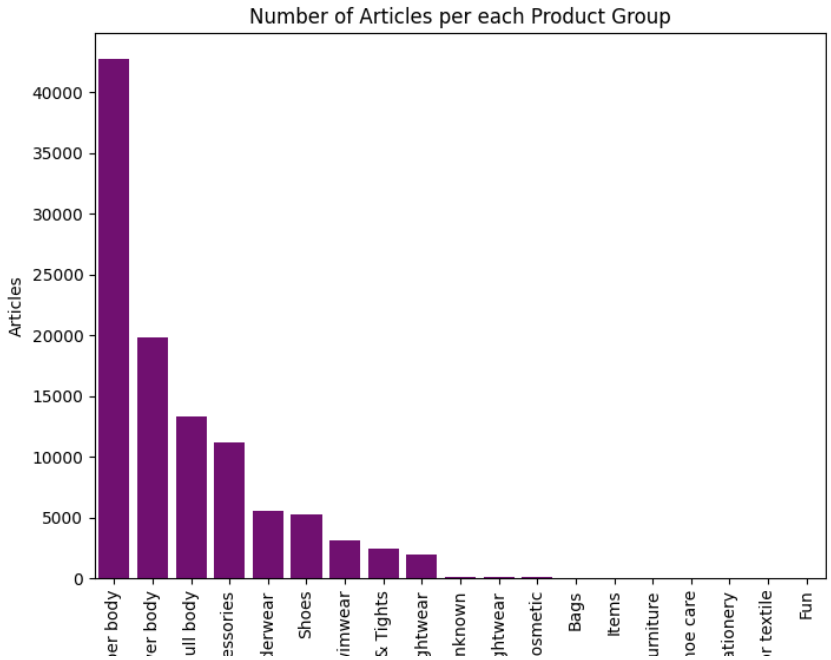


Chart & Graph-1: Product group in article

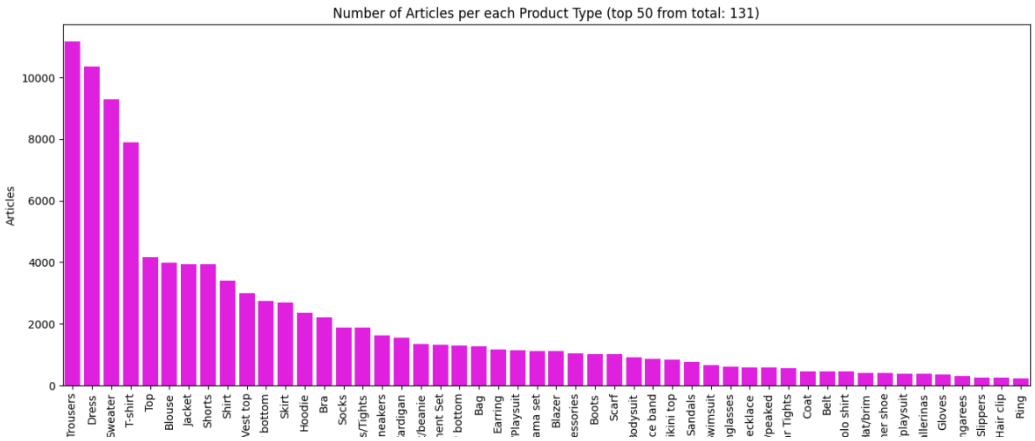


Chart & Graph-2: Product Type in Article

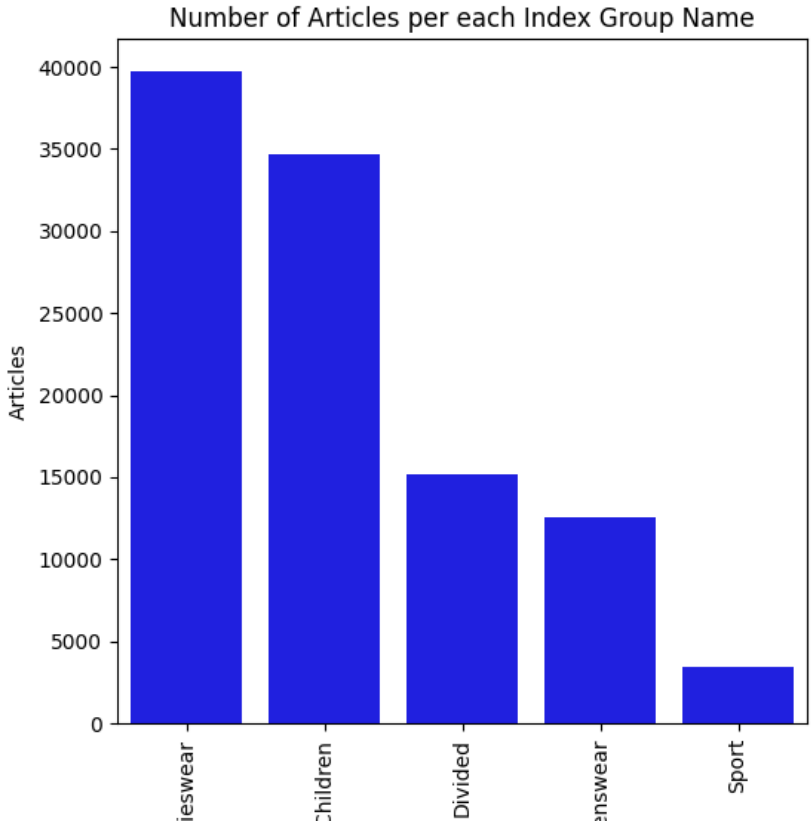


Chart & Graph-3: Index Group name in Article

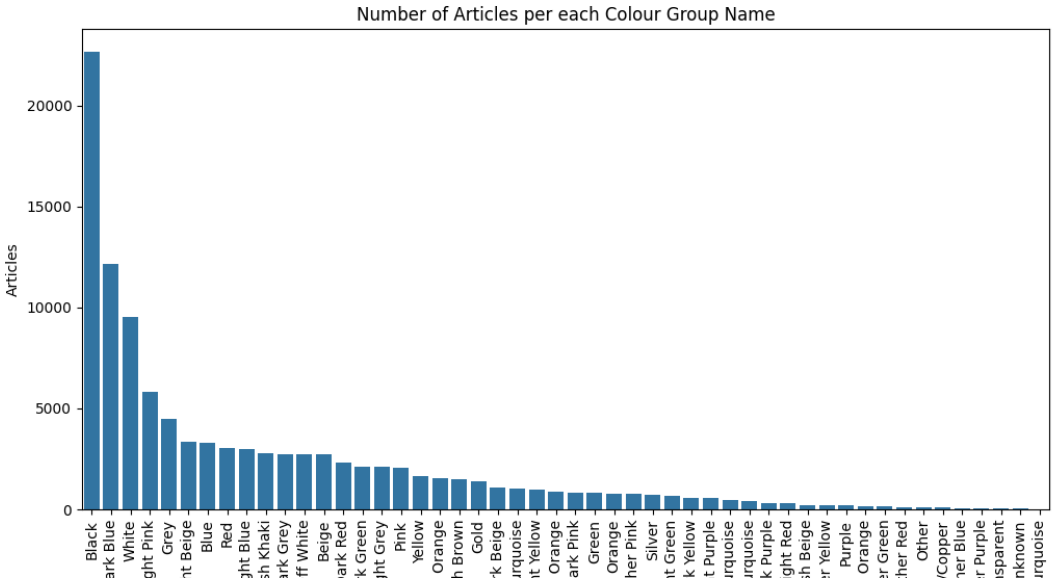


Chart & Graph 4: Colour Group Name in Articles

In customer dataset, we can see that the buyers earlier have been divided into several age groups:

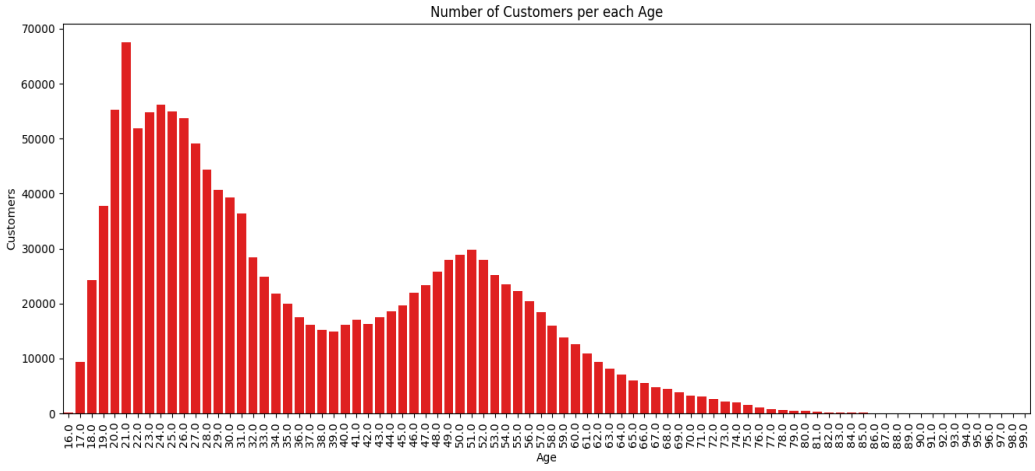


Chart & Graph-5: Age in Customer Dataset

In transaction dataset, how much amount each customer has spent has been featured. In the chart below, we can see that how many transactions have been taken place each day.

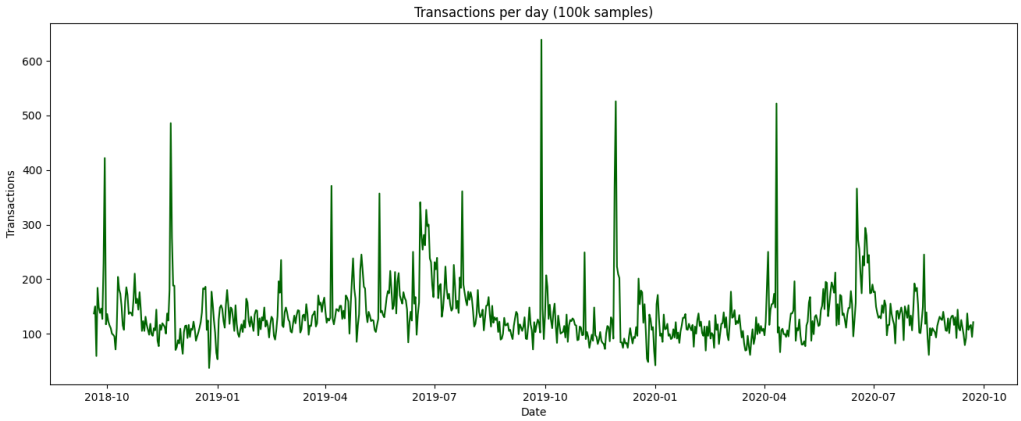


Chart & Graph-6: Transactions Per Day

There are 88 subfolders in the picture folder, each holding different images. Since the article id and the picture name suggest the same thing, we plotted the photos after combining them with the article dataset.

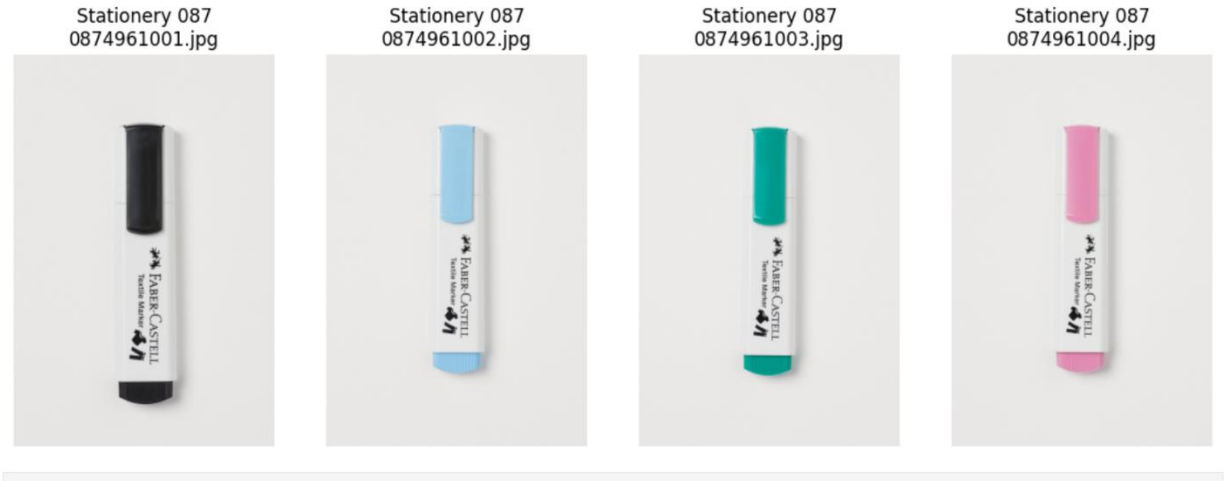


Figure-4: Plotting Images of Stationary Category

Figure-5: Plotting Images of Lower Body Category

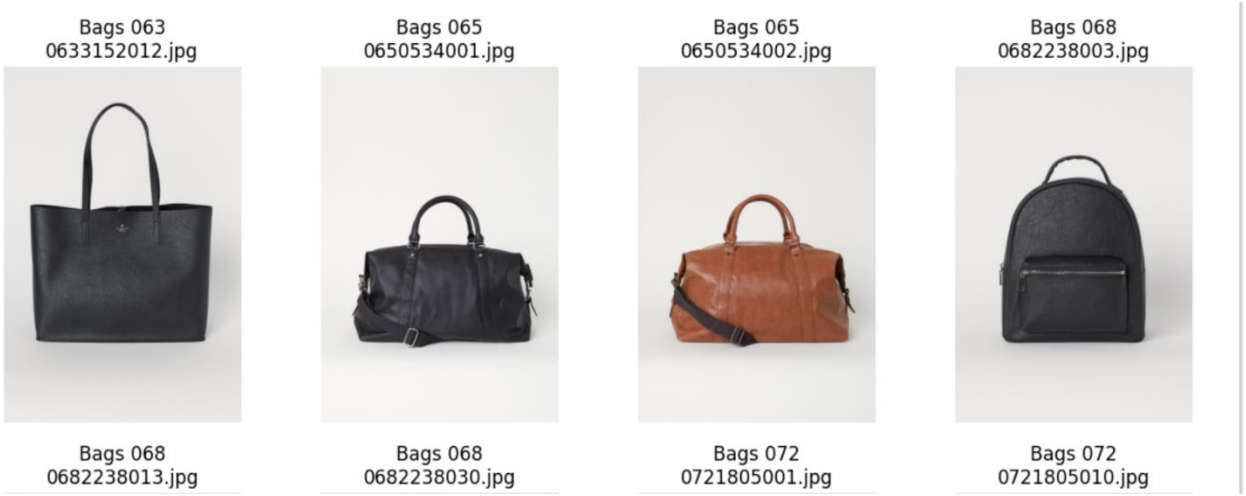


Figure-6: Plotting Images of Bags Category

Data Pre-processing:

We merged the transaction, customer and article dataset having customer_id as the shared feature. We dropped the unnecessary features for fitting them into the clustering algorithm. In the combined dataset, there is a feature called 'index_group_name'. We then created a dataset having different instances of 'customer_id' as features. There are other features named 'price' and 'age' in the new dataset. We can see the final dataset below:

customer_id	Baby/Children	Divided	Ladieswear	Menswear	Sport	price	age
0000423b00ade91418cceaf3b26c6af3dd342b51fd051eec9c12fb36984420fa	0	2	1	0	0	0.042356	25.0
000058a12d5b43e67d225668fa1f8d618c13dc232df0cad8ffe7ad4a1091e318	0	0	2	0	0	0.040661	24.0
00007d2de826758b65a93dd24ce629ed66842531df6699338c5570910a014cc2	0	2	3	0	0	0.017271	32.0
0003abe64294e66a6310c3436fa9e5b754cc5603deef4f26fc8ab8d043af9358	0	3	1	0	0	0.029644	51.0
0004068f54dbe1c7054b23c615edc5f733a508ecc54930bf323209f20410898c	0	0	1	0	1	0.016085	51.0

Figure-7: Dataset to be fitted into the algorithm

Algorithm:

K-means Clustering-

By assigning each data point to the cluster with the closest centroid, the unsupervised machine learning algorithm K-means clustering divides unlabelled data points into a predetermined number of clusters ("k"). This algorithm Iteratively adjusts the centroids until the clusters are optimally separated based on their features [13].

Data can be grouped using K-means clustering according to how similar or close the data points are to one another[8 simpleLearn [14]. Although it's a rather quick and effective procedure, it performs best when the clusters are separated and not overly mixed up. However, determining the appropriate number of clusters (K) in advance is a challenge. Additionally, K Means may not work as well if the data has a lot of noise or overlap [15].

How K-means clustering works-

Finding clusters in the provided input data is the aim of the K-Means algorithm. There are two methods for doing this. By defining the value of K (e.g., 3,4, 5), we can employ the trial and error method. We continue adjusting the value as we go along until we obtain the optimal clusters. To find the value of K, another approach is to apply the Elbow Method. After determining the value of K, the system will randomly assign a certain number of centroids and calculate the separation between each data point and these centroids. As a result, it allocates those points to the centroid that is closest to the distance.

The figure below containing the flow chart of how k-means works is given

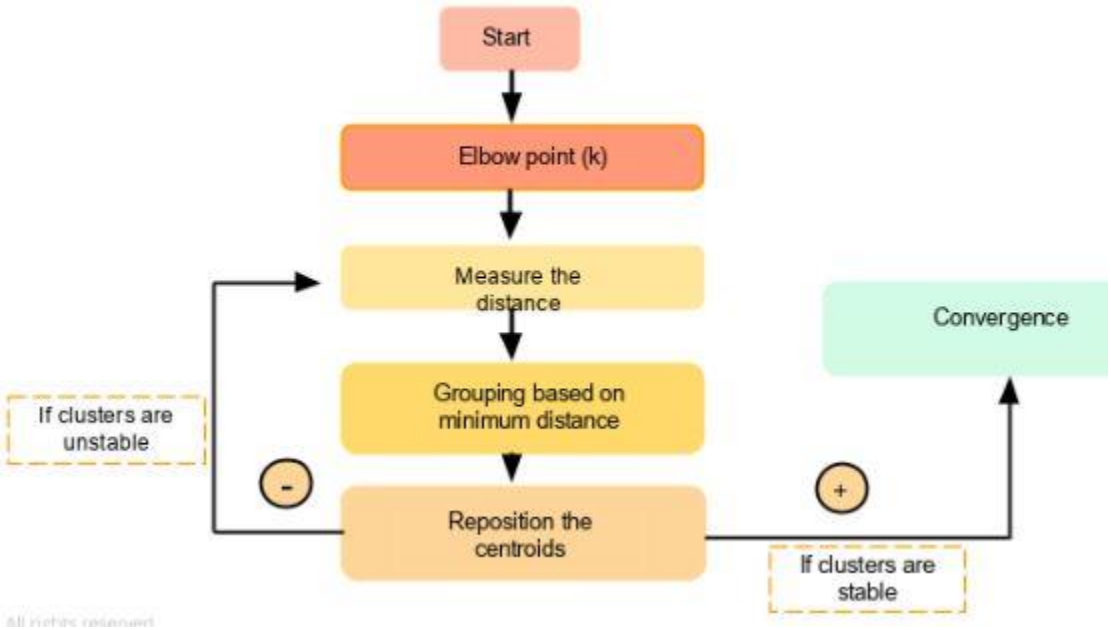
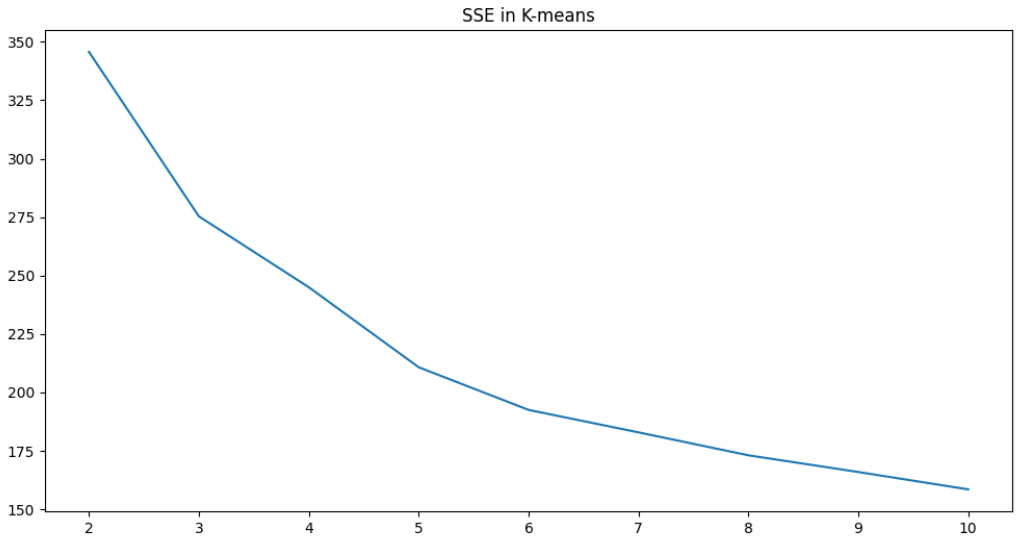


Figure-8: How k-means clustering works

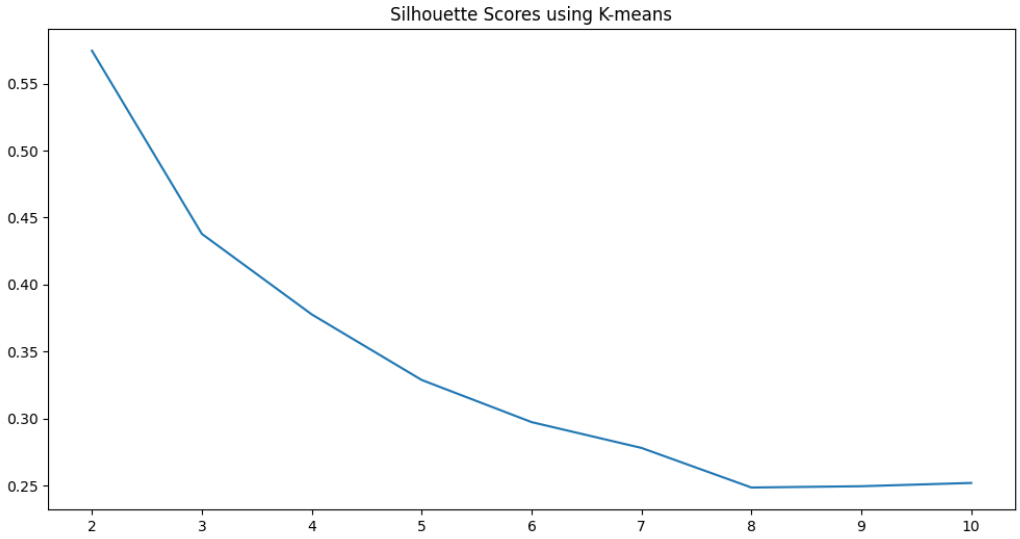
Implementation:

We fitted our final dataset in K-means clustering algorithm. The algorithm shows the SSE :173.7495616072828 and Silhouette Score : 0.2520 after training the dataset.

We checked the value of K starting from 2 to 10 and plotted graphs having the value of K on x-axis and the scores on y-axis.

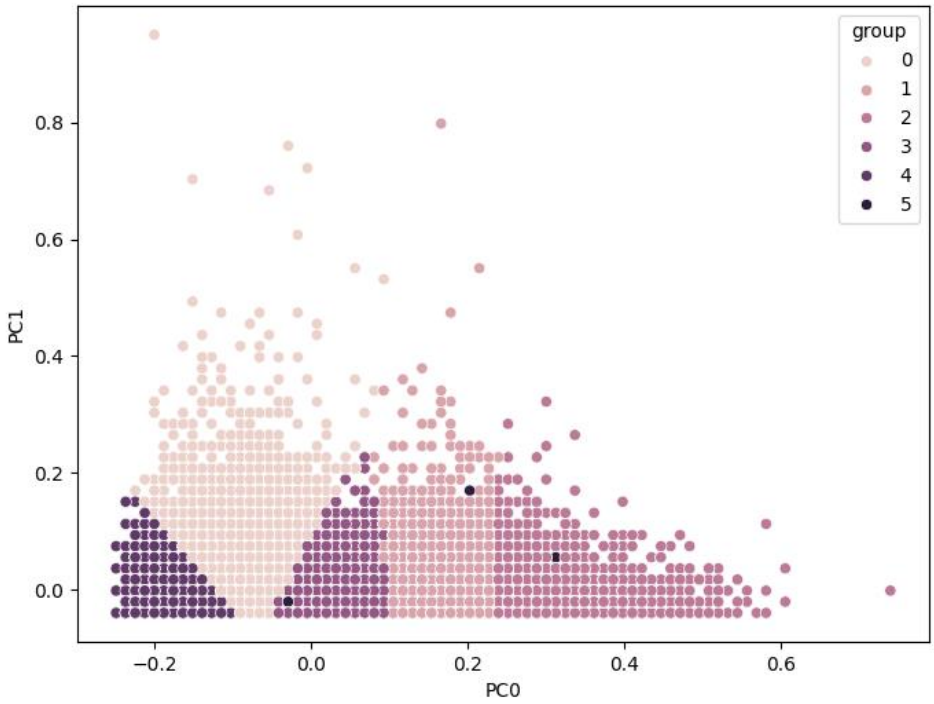


Graph & chart-7: K vs SSE

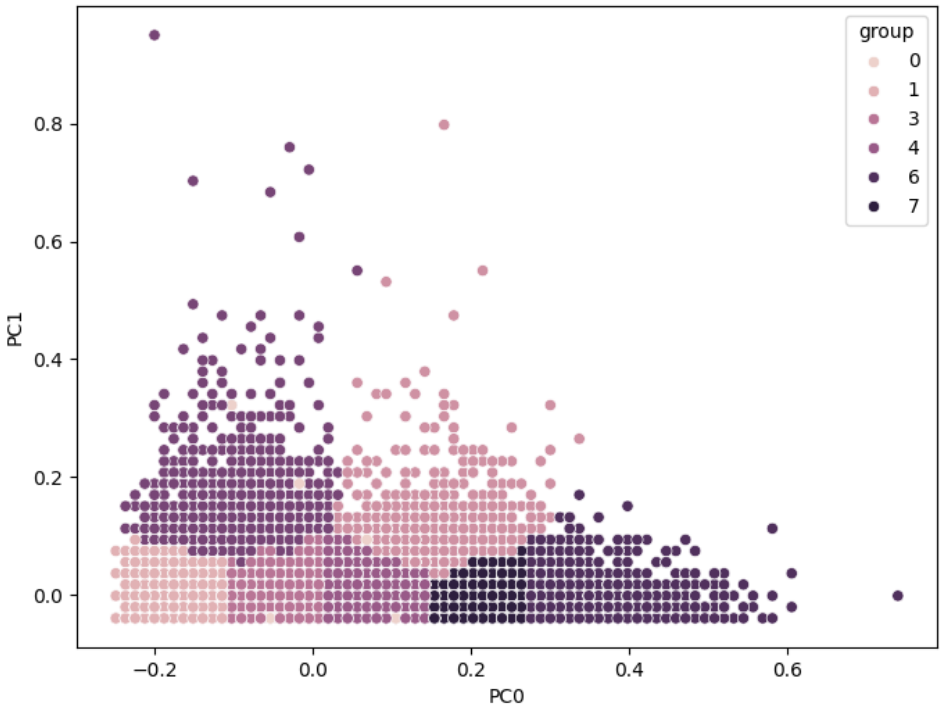


Graph & Chart-8: K vs Silhouette Score

By applying the elbow technique, we can see that the value of K is 6 for SSE and 8 for Silhouette Score. Now we can visualize the clusters for the value of K being 6 and 8 respectively.



Graph & Chart-9: Clusters when k is 6



Graph & Chart-10: Clusters when k is 8

Result:

After the implementation, we can see that the customers have been divided into different eight groups since k=8 looks like the best match.

	PC0	PC1	group
0	-0.151540	-0.020972	2
1	-0.163764	-0.001932	2
2	-0.065976	0.017108	7
3	0.166269	-0.020972	4
4	0.166269	-0.020972	4
...	...	...	...
28133	-0.053753	0.055189	7
28134	0.227387	-0.001932	4
28135	0.117376	-0.001932	4

Figure-9: Assigned groups

We can see the SSE score of 173.7495616072828 value and Silhouette Score of 0.2520 value.

Name of the Score	Value
SSE	173.7495616072828
Silhouette Score	0.2520

Table-3: Scores

Conclusion & Discussion:

The evaluation of our recommendation system yielded an SSE score of 173.75 and a Silhouette Score of 0.2520. The relatively low SSE indicates that the model's predictions are reasonably close to the actual data points, suggesting that the clustering is somewhat effective in capturing user preferences. However, the Silhouette Score of 0.2520 reveals potential issues with cluster separation and cohesion. A Silhouette Score closer to 1 would suggest better-defined clusters, while a score near 0 indicates overlapping clusters. This suggests that while our recommendation system can identify groups of similar users or items, there is room for improvement in distinguishing between these groups more clearly. To enhance the performance of the recommendation system, we could explore alternative clustering algorithms, such as DBSCAN or hierarchical clustering, which may provide better separation of clusters. Since the dataset is constituted with various features, some other features could be used for achieving better output. Additionally, fine-tuning hyperparameters and incorporating more relevant features could lead to improved clustering results. Overall, while the current model shows promise, further refinement is necessary to enhance its effectiveness and user satisfaction in delivering personalized recommendations.

References

[1] A. A. Khade, "Performing Customer Behavior Analysis using Big Data Analytics".

[2] P. D. a. I. Krishnamurthi2, "Customer Behavior Analysis Using Rough Set Approach," 19 September 2012.

[3] b. M. M. C. H. Y. Y.-C. Z. ,. Z.-K. Z. Linyuan Lua, "Recommender System," 06 February 2012.

[4] *. b. F.O. Isinkayea, "Recommendation System: Principles, Methods & Evaluation," 30 June 2015.

[5] D. W. M. M. W. W. G. Z. Jie Lu *, "Recommender system application developments: A survey," p. 21, 2015.

[6] H. R. D. M. N. A. P. &. Mohammad Hossein Moshref Javadi1, "An Analysis of Factors Affecting online Shopping Behavior of Consumers," p. 18, 10 September 2012.

[7] Venkata Siva Prakash Nimmagadda, "Artificial Intelligence for Customer Behavior Analysis in Insurance: Advanced Models, Techniques, and Real-World Applications," p. 18.

[8] G. S. a. A. Gunawardana, "Evaluating Recommendation System," p. 43.

[9] P. H. J. X. X. G. C. G. F. S. C. L. A. P. H. Z. B. Z. Wen Chen, "POG: Personalized Outfit Generation for Fashion Recommendation at Alibaba iFashion," 20 May 2019.

[10] L. K. WEI GONG, "Aesthetics, Personalization and Recommendation: A survey on DeepLearning in Fashion," June 2021.

[11] *. ,. Y. P. 1. a. A. C. 2. Hyeyoung Ko1, "A Survey of Recommendation Systems: Recommendation Models, Techniques, and Application Fields," 3 January 2022.

[12] N. M. a. M. S. S. E. M Sridevi1, "Personalized fashion recommender system with image based neural netwrok".

[13] "K -Means Clustering," [Online]. Available: https://en.wikipedia.org/wiki/K-means_clustering.

[14] "All About K-means Clustering Algorithm," 08 December 2025. [Online]. Available: https://www.simplilearn.com/tutorials/machine-learning-tutorial/k-means-clustering-algorithm.

[15] 15 January 2025. [Online]. Available: https://www.geeksforgeeks.org/k-means-clustering-introduction/.