SJBMSS
Scientific Journalof Business
Management and Social Science

The emerging trends of automation data management techniques importance to optimise management operations. A case study of UK fashion industry

By

Areeb Saleem

ABSTRACT

Aim: This research aims to explore the contemporary trends concerning the automation data management approaches, and their relevance for the enhancement of the management processes in the British fashion sector.

Design/Method: An extensive database of the UK fashion industry was used, which included general and specific characteristics of clothes, customers' feedback, and purchasing behaviour. Cleaning the data included methods such as dealing with missing data and transforming nominal variables into numerical ones. Since the data was large, K-means clustering was applied to partition the data into relevant clusters of data. In choosing the appropriate number of clusters, Exploratory Data Analysis (EDA) using the elbow method was adopted, while the silhouette score was also used in the assessment of clustering performance.

Finding/Results: A factor analysis exposed four better clusters where the features singled out from a sketch map of the coordinates are defined and discussed about the fashion industry. The study found customer preferences, seasonal ratio, and available products by clustering. The next steps of operational marketing strategies such as suggestions for targeted marketing approaches, inventory plans and customer management strategies were suggested based on the findings of clustering analysis. The work shows how advancements such as automation and machine learning can help improve the operations and decision-making processes within the fashion business.

Keywords: *Automation, Data Management, K-means Clustering, Fashion Industry, Machine Learning, Customer Preferences, Operational Efficiency, Strategic Decision-Making*

INTRODUCTION

Background

The fashion sector in the UK is a vital element of the country's economy as it help to provide new opportunities and improve the framework of the country (Babu et al., 2024). As automation and machine learning take centre of attention in organisational management, organisations within this sector rely on big data to discover operative and value-creation solutions for customers (Balchandani et al., 2023). This explains why the ability to manage and analyse data has been so important given that the consumers' tastes and preferences are ever-changing. Automation in data management enables real-time data analysis, is less prone to errors, and most importantly, frees up useful resources for value addition (PwC, 2021).

Several technological applications in automating data management for the fashion industry include predictive analytics, inventory and management systems, and customer marketing tools (Mohiuddin Babu et al., 2022). They help firms predict consumer patterns, manage the logistic activities of the supply chain, and personalise advertising techniques as per the consumer behaviour pattern (Chase, 2020; Accenture, 2021). For example, fashion retailers are able, to forecast which products will be popular in the following seasons, so overstock and stockout risks are limited. Further, the automated application of CRM enhances consumer behaviour analysis, which results in efficient marketing communication. This change is not only beneficial for increasing operational productivity, but it also greatly benefits the end product and the customer base which is a key priority in today's world.

Problem Statement

Today's many fashion firms in the UK must deal with major issues regarding data management automation. These issues include the problem associated with many isolated data sets, poor infrastructure, and integration issues that limit the use of sophisticated technologies and restrict organisational attempts to achieve optimum gains from automation in optimizing business processes. Mitigating these challenges is a necessity for the uptake of data management systems which are critical for existence in such fast-growing industry (Deloitte, 2021).

Aim

To explore the impact of automated data processing methods on productivity within the UK fashion supply chain.

Objectives

- To identify the prevalent areas of automation in data management within the UK fashion industry.
- To assess the utility of these automation methods in improving management and operational processes.
- To evaluate the overall impact of automation on strategic decision-making and operational efficiency in the UK fashion industry.
- To provide recommendations for overcoming the challenges associated with integrating automated data management systems.

LITERATURE REVIEW

Overview of Automation and Data Management Techniques in the Fashion Industry

It is important to note that the fashion industry is changing rapidly nowadays due to the implementation of the automation process and innovative methodologies of data management. They make it possible for ventures to filter and analyse huge amounts of information to deal with consistently changing markets. Data automation in business processes entails the utilisation of AI, ML, and predictive analytics in the various business processes such as supply chain management, and customer relationship management (CRM) among others (Accenture, 2021).

Automating the processes is another area where the fashion industry has benefitted extensively; the best example is inventory management. Such systems can help forecast the needed stock levels considering the sales history, current trends, and seasonal demand, thus limiting the possibility of overstocking and stockouts (PwC, 2021). Similarly, with the assistance of the aspects of predictive analytics, one is also in a position to predict the flow of fashion trends, thus helping the designers and retailers in establishing the right line of products that would be of demand in the market.

Another is segment marketing which refers to marketing products that are specifically designed to fit segments of the population. Advanced CRM systems make use of AI to segment the market into different categories which are by customers' purchase behaviour, decision-making, and age. Such segmentation facilitates the possibility of various advertisements for different audiences and thus higher satisfaction and customer loyalty (Balchandani et al., 2023).

In addition, real-time data analysis is crucial in the fashion industry because of automation. For instance, the numbers on sales can be adapted to alter the promotion and sales techniques, control the price changes, and manage stock in and out constantly. This agility is especially important given that the market is highly dynamic and the customers' preferences regarding products can be easily shifted (Deloitte, 2021).

Review of K-means Clustering and Its Applications

K-means clustering is an example of unsupervised learning, which is used to classify a given data set into various clusters according to their similarity. It helps in using up a pre-defined number of centroids and assigning all the data points to the groups which are nearest to the centroids and then, the process continues updating the centroid values continually until the endpoint is reached (Lloyd, 1982). Hence the aim is to minimise the value of the sum of squared distances between each 'n' number of data points and its corresponding mean value of that cluster.

Hypothesis 1: The implementation of K-means clustering will lead to more effective market segmentation in the UK fashion industry.

When applied in the fashion industry, the following are the practical uses of the K-means clustering technique. One of the major purposes of using this method is in the segmentation of the market. Through segmentation, clients are grouped by their purchasing patterns, demography, and their preferences; thus, marketing efforts can be improved especially for unique groups (Govender and Sivakumar, 2020). For instance, the customers who often spend more money and buy expensive goods and services can be offered appropriate offers and recommendations, and, in contrast, the customers who buy something without frequent purchases and give preference to cheap and affordable products, can be offered appropriate discounts and offers.

Hypothesis 2: The use of K-means clustering will improve inventory management by reducing overstocking and stockouts.

Another application of this technique is in assessing inventory. Grouping products in line with seasonal sales characteristics and other factors can be useful in managing the inventory, thereby cutting most holding costs. For example, if certain products are usually sold together, it is possible to group those types of products, and at the same time, there is assurance that everything will be in stock (Han et al., 2022).

Hypothesis 3: K-means clustering will enhance trend forecasting, allowing fashion companies to stay competitive by anticipating consumer needs and demands.

Another application of K-means clustering is enabled within the trend analysis. Since customers tend to have their preferred clothing styles over the years, fashion organisations can produce their clothes based on the previous year's sales data. This is very strategic since it assists in the anticipation of competition about fulfilling consumer needs and demands (Tan et al., 2016).

Hypothesis 5: The application of K-means clustering will reveal new data patterns that can inform strategic decision-making processes.

For example, it can cluster the company customers' data to identify several segments having different preferences and actions to target a specific marketing campaign (Xu and Wunsch, 2005).

Hypothesis 6: The use of K-means clustering will improve inventory management by reducing overstocking and stockouts.

When it comes to inventory control, clustering is quite useful in grouping products based on their sales level, and their sale seasons, amongst others. This categorisation will enable the retailers to control their stock in place and pace, thereby avoiding the creation of a blockage or running out of stock of the product. It is thus noteworthy that, when attempting to grasp demand patterns of different clusters of products, businesses may find their procurement and distribution efficiency increased, causing a diminishment of costs as well as an enhancement of the satisfaction of consumers (Han et al., 2022).

Hypothesis 7: K-means clustering will enhance trend forecasting, allowing fashion companies to stay competitive by anticipating consumer needs and demands.

In the same way, clustering assures trend long-term forecasting in the fashion line of business. Thus, by working with the distributions of historical sales data and focusing on the clusters of the latter, one can learn about the new tendencies and adjust the offered assortment. This proactive approach assists in the ability to manoeuvre in a competitive market with clients, where the perception of the customers often shifts in expediting time (Tan et al. 2016).

Hypothesis 8: Clustering techniques will enhance the design process by providing designers with insights into customer preferences, leading to more appealing product offerings.

Furthermore, clustering techniques can be beneficial in the design process since they offer the designer information about the customers' habits. It will therefore help designers in their thought process and ensure they make outfits that will appeal to their intended clientele if they are to sell their products. For instance, through analysis of the feedback received from fashion shows and social media, the designers are in a position to determine the prevailing fashion trends, the superior colours, and the superior materials to apply in their designs (Balchandani et al., 2023).

In conclusion, it is possible to state that the application of clustering techniques and, in particular, K-means clustering allows for achieving essential advantages in managing operations in the fashion industry. These techniques make market segmentation more effective, allow to organise stocks more rationally, forecast trends, and make correct decisions on designing. Thus, the application of clustering algorithms will help fashion companies find a competitive advantage in the rapidly changing environment, increase consumer satisfaction, and consequently stimulate the growth of the business.

METHODOLOGY

Description of the Dataset

The data set used in the analysis for this paper was obtained from a Kaggle website leading UK-based fashion retailer and is an integration of all sales and customers' data. More of the features that are in the dataset are the product name, the price, the brand, the type of the product, description, rating, the number of reviews, style parameters, the total size, size options in stores, the colour, the number of purchases made in the past, age, fashion magazines reviewed, fashion enterprise influencers, season, time of year most products were purchased, review from the customers, comments on social media, and feedback. As the figures in the current database reveal more than 100,000 records, the given data set is rather representative, which creates a sound basis for identifying trends and patterns in the sphere of fashion in the United Kingdom.

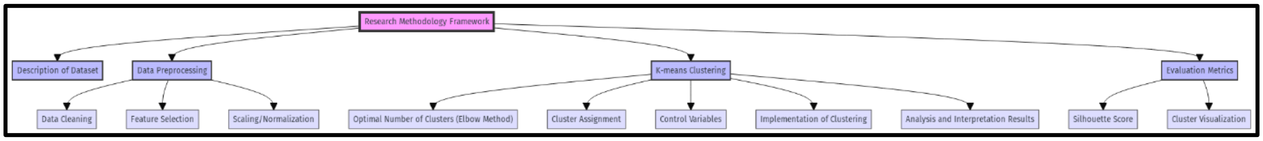


Figure 1: Methodology of this study (Author, 2024)

Data Pre-processing Steps

Data preprocessing is a common step of preparation for clustering analysis to help the dataset qualify. The following steps were undertaken:

1. **Handling Missing Values:** For the numeric variables, it required imputing the missing values with the help of the mean of the corresponding variable. The other type of columns that was not included in the further analysis, was the non-numeric columns that contained missing values (Sarma et al., 2022).
2. **Encoding Categorical Variables:** Nominal features such as brand, category, description, style attribute, colour, fashion magazines, fashion influencers, season, time of the highest purchase, and feedback along with social media comments were encoded using Label Encoding. This helps in translating categorical variables into something more manageable for the clustering algorithm, numerical variables.
3. **Feature Selection:** String variables including the product name, total size, available size and purchase record were also removed from cluster analysis. The rest of the features, and covariant, were moved to the next stage of analysis.
4. **Data Standardisation:** To make sure that each feature will contribute the same in the analysis, the data was normalised using the StandardScaler from sklearn to bring the data to mean = 0 and variance = 1 (Testas, 2023).

Details of K-means Clustering Implementation

K-means clustering is a classical machine learning algorithm where data is divided into k clusters. The implementation process in this study involved the following steps:

1. **Initialisation:** The number of clustered ‘k’ was found by using the elbow method; the WCSS was plotted against the number of clusters and then found the point after which the rate of reduction in the value of WCSS was slow (Hassan et al., 2021).
2. **Clustering Process:** As for the clustering model, the MiniBatchKMeans from sklearn was used for the task. The latter is a large sample version of K-means and works by updating centroids separately over small random samples of the entire database (Lloyd, 1982).
3. **Convergence:** Specifically, the steps involve the following The algorithm continues until it reaches convergence The convergence is considered to be the point at which there is very little change in the positions of the centroids from the previous iterations.
4. **Cluster Assignment:** Within this process, each data point is affiliated with the cluster that has the nearest centroid-producing k clusters (López-Oriona et al., 2022).

Explanation of Control Variables

These are variables that are fixed in the study to eliminate the chances of them meddling with the results of the analysis. The control variables for this study were seasons, fashion magazines, and fashion influencers. Given that these variables are well known to influence the behaviour of consumers about their patronage of fashion sales, they were determined to be appropriate. When integrating these control variables into the clustering analysis, it will also be possible to eliminate seasonal influences and marketing on the results (Govender and Sivakumar., 2020).

Justification for the Choice of Clustering Method

Therefore, K-means clustering was selected in this study it is a simple, efficient and powerful tool for clustering large datasets. This is especially useful in partitioning data into different categories based on their similarity; thus, it’s ideal for disentangling patterns and trends within the fashion industry, according to the purpose of the current study outlined by Lloyd (1982) in his perspective. However, due to optimised variants like MiniBatchKMeans, it is possible to such large datasets with K-means clustering techniques in a proper manner.

Equations Used in the Methods

The key equations used in the K-means clustering method are as follows:

1. **Initialisation of Centroids:**

$$\boldsymbol{\mu_j^{(0)} = x_i\ \ for\ j\ \epsilon\ \{1, 2, 3, \dots, k\}}$$

where $\mu_j^{(0)}$is the initial centroid for the cluster j and $\boldsymbol{x_i}$ is a randomly picked data point.

2. **Assignment Step:**

$$\boldsymbol{c_i = arg\ \underbrace{min}_{j}\ ||x_i - \mu_j||^2}$$

where $\boldsymbol{c_i}$ represents the cluster number of data point $\boldsymbol{x_i}$ and $||\boldsymbol{x_i - \mu_j}||^2$ is the squared Euclidean distance between data point $\boldsymbol{x_i}$ and centroid point $\boldsymbol{\mu_j}$.

3. **Update Step:**

$$\boldsymbol{\mu_j = \frac{1}{|\ C_j|} \sum_{x_i\ \epsilon\ C_j} x_i}$$

where $\boldsymbol{\mu_j}$ is the new centroid for cluster **j** and $\boldsymbol{C_j}$ is the collection of observations belonging to the same cluster as **j**.

These equations are the base of K-means clustering algorithm which is useful in the next iteration process of partitioning the given data into K meaningful clusters that are similar to each other in terms of their features (Umargono et al. , 2020).

RESULTS AND DISCUSSION

Presentation of Exploratory Data Analysis (EDA) Results

The first examination of the provided data set allowed to assess the general picture for the mutual positioning of the main variables. They also include the price range, the rating, the reviews, and the categorical variables being brand, category, and style. The integrity of the data after the processing was checked with the help of the illustration of the first few rows of the original dataset and the post-cleaned dataset.

Table 1: First Few Rows of Dataset

	Product Name	Price	Brand	Category	Description	Rating	Review Count	Style
0	T5D3	97.50997	Ralph Lauren	Footwear	Bad	1.421706	492	Streetwear
1	Y0V7	52.34128	Ted Baker	Tops	Not Good	1.037677	57	Vintage
2	N9Q4	15.43098	Jigsaw	Footwear	Very Bad	3.967106	197	Streetwear
3	V2T6	81.11654	Alexander McQueen	Outerwear	Not Good	2.844659	473	Formal
4	S7Y1	31.63369	Tommy Hilfiger	Bottoms	Very Good	1.183242	55	Sporty

	Attributes	Total Sizes	Available Sizes	Color	Purchase History	Age	Fashion Magazines	Fashion Influencers
0	M, L, XL	XL	XL	Green	Medium	24	Vogue	Chiara Ferragni
1	M, L, XL	XL	XL	Black	Above Average	61	Glamour	Leandra Medine
2	S, M, L	M	M	Blue	Average	27	Marie Claire	Gigi Hadid
3	S, M, L	L	L	Red	Very High	50	Marie Claire	Chiara Ferragni
4	M, L, XL	S	S	Green	Above Average	23	Glamour	Song of Style

	Season	Time Period	Highest Purchase	Customer Reviews	Social Media Comments	feedback
0	Fall/Winter	Daytime	Daytime	Mixed	Mixed	Other
1	Winter	Weekend	Weekend	Negative	Neutral	Other
2	Summer	Nighttime	Nighttime	Unknown	Negative	Neutral

3	Fall/Winter	Weekend	Weekend	Neutral	Other	Other
4	Spring	Daytime	Daytime	Positive	Mixed	Positive

Table 2: Descriptive Statistic Analysis of Dataset

Statistic	Product Name	Price	Brand	Category	Description	Rating	Review Count
count	0	500000	500000	500000	500000	500000	500000
mean	NaN	55.01677	3.501742	4.505764	2.998114	3.000149	249.547
std	NaN	25.97393	2.291042	2.872313	2.001285	1.154742	144.3453
min	NaN	10.00015	0	0	0	1.000008	0
25%	NaN	32.53645	2	2	1	1.999985	125
50%	NaN	55.04617	4	5	3	2.999252	249
75%	NaN	77.5141	5	7	5	4.000007	374
max	NaN	99.99932	7	9	6	4.999995	499

Statistic	Style Attributes	Total Sizes	Available Sizes	Color	Purchase History	Age	Fashion Magazines
count	500000	0	0	500000	0	500000	500000
mean	4.502534	NaN	NaN	1.499262	NaN	41.01211	4.502288
std	2.875375	NaN	NaN	1.118135	NaN	13.56727	2.87089
min	0	NaN	NaN	0	NaN	18	0
25%	2	NaN	NaN	0	NaN	29	2
50%	5	NaN	NaN	1	NaN	41	5
75%	7	NaN	NaN	2	NaN	53	7
max	9	NaN	NaN	3	NaN	64	9

Statistic	Fashion Influencers	Season	Time Period	Highest Purchase	Customer Reviews	Social Media Comments	feedback

count	500000	500000	500000	500000	500000	500000	500000
mean	4.499624	2.496396	NaN	1.999848	1.999258	2.49906	2.503048
std	2.870288	1.7094	NaN	1.414402	1.41402	1.708152	1.707529
min	0	0	NaN	0	0	0	0
25%	2	1	NaN	1	1	1	1
50%	4	2	NaN	2	2	2	3
75%	7	4	NaN	3	3	4	4
max	9	5	NaN	4	4	5	5

Histograms of critical numerical variables like the price, rating, review count, and the age of the products showed the distributions of the variables. Thus, as an example, we have the price distribution, which demonstrated that most of the products are close in price; the distribution of the review count with a higher percentage of products in the middle-review range, indicating the need for revisiting the customer engagement strategy.

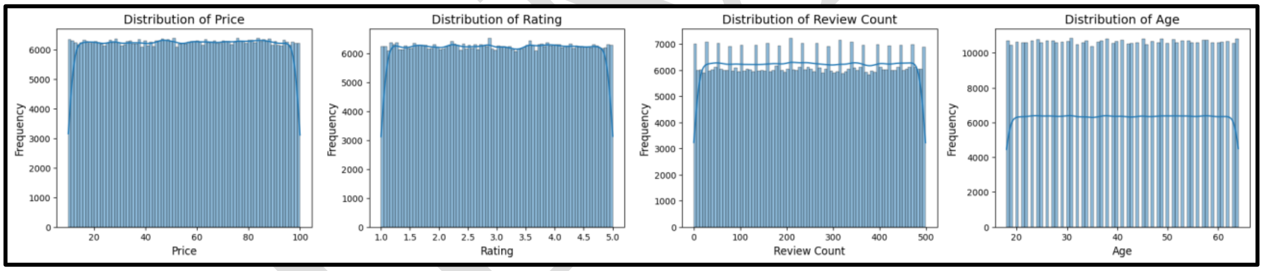
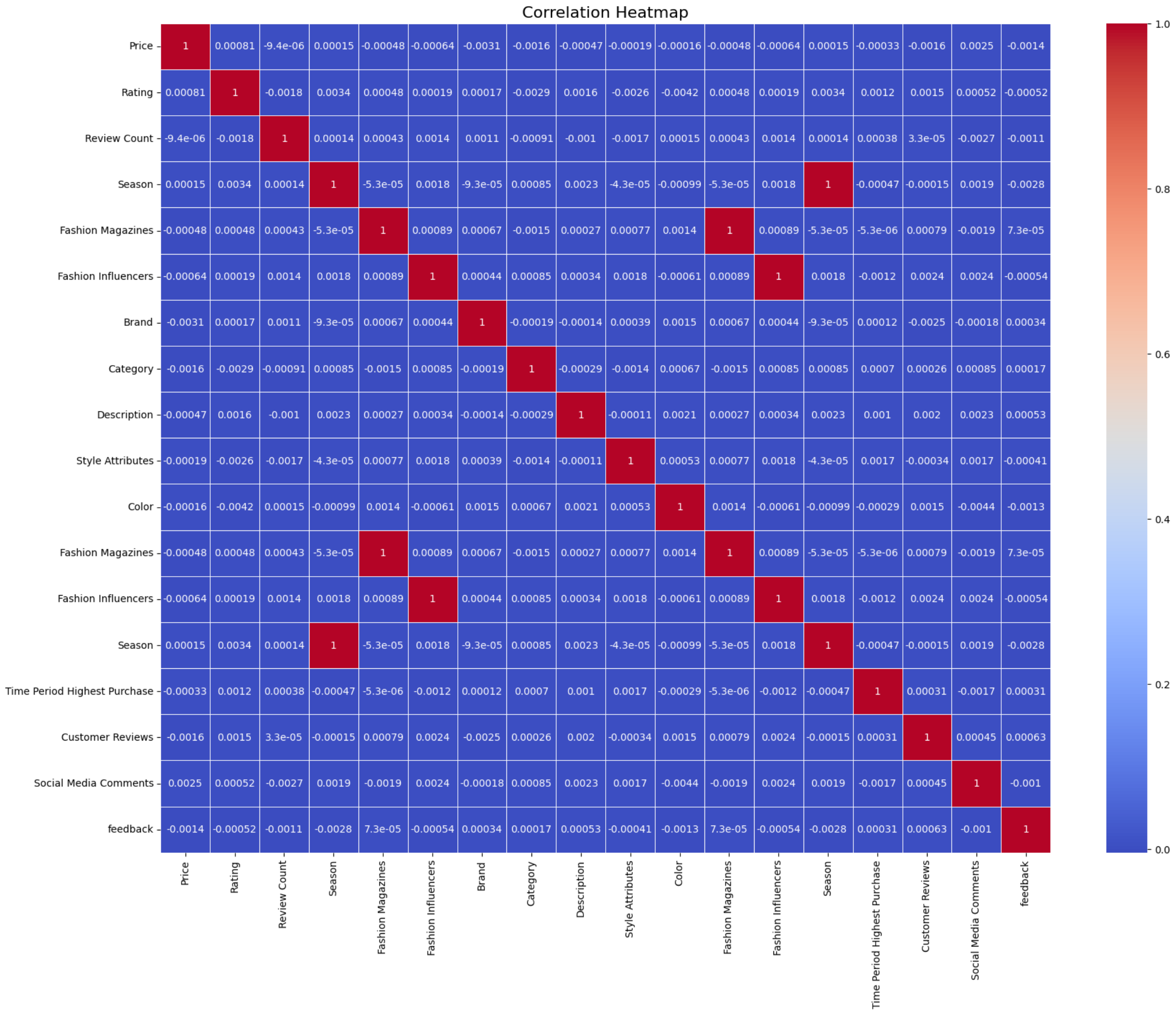


Figure 2: Histogram Plots of the dataset.:

To investigate the correlation between the numerical features, correlation analysis was conducted followed by a heatmap presentation, which showed the correlations and degree of colours between them. For example, one of the groups' findings was a positive though very weak relationship between price and rating – thus, if things were more expensive, they were rated higher. Nevertheless, most of the correlations were low; thus, one might assume that there were other factors impacting these attributes that were not determined by adherence to religious beliefs.



	Price	Rating	Review Count	Season	Fashion Magazines	Fashion Influencers	Brand	Category	Description	Style Attributes	Color	Fashion Magazines	Fashion Influencers	Season	Time Period Highest Purchase	Customer Reviews	Social Media Comments	feedback
Price	1	0.00081	-9.4e-06	0.00015	-0.00048	-0.00064	-0.0031	-0.0016	-0.00047	-0.00019	-0.00016	-0.00048	-0.00064	0.00015	-0.00033	-0.0016	0.0025	-0.0014
Rating	0.00081	1	-0.0018	0.0034	0.00048	0.00019	0.00017	-0.0029	0.0016	-0.0026	-0.0042	0.00048	0.00019	0.0034	0.0012	0.0015	0.00052	-0.00052
Review Count	-9.4e-06	-0.0018	1	0.00014	0.00043	0.0014	0.0011	-0.00091	-0.001	-0.0017	0.00015	0.00043	0.0014	0.00014	0.00038	3.3e-05	-0.0027	-0.0011
Season	0.00015	0.0034	0.00014	1	-5.3e-05	0.0018	-9.3e-05	0.00085	0.0023	-4.3e-05	-0.00099	-5.3e-05	0.0018	1	-0.00047	-0.00015	0.0019	-0.0028
Fashion Magazines	-0.00048	0.00048	0.00043	-5.3e-05	1	0.00089	0.00067	-0.0015	0.00027	0.00077	0.0014	1	0.00089	-5.3e-05	-5.3e-06	0.00079	-0.0019	7.3e-05
Fashion Influencers	-0.00064	0.00019	0.0014	0.0018	0.00089	1	0.00044	0.00085	0.00034	0.0018	-0.00061	0.00089	1	0.0018	-0.0012	0.0024	0.0024	-0.00054
Brand	-0.0031	0.00017	0.0011	-9.3e-05	0.00067	0.00044	1	-0.00019	-0.00014	0.00039	0.0015	0.00067	0.00044	-9.3e-05	0.00012	-0.0025	-0.00018	0.00034
Category	-0.0016	-0.0029	-0.00091	0.00085	-0.0015	0.00085	-0.00019	1	-0.00029	-0.0014	0.00067	-0.0015	0.00085	0.00085	0.0007	0.00026	0.00085	0.00017
Description	-0.00047	0.0016	-0.001	0.0023	0.00027	0.00034	-0.00014	-0.00029	1	-0.00011	0.0021	0.00027	0.00034	0.0023	0.001	0.002	0.0023	0.00053
Style Attributes	-0.00019	-0.0026	-0.0017	-4.3e-05	0.00077	0.0018	0.00039	-0.0014	-0.00011	1	0.00053	0.00077	0.0018	-4.3e-05	0.0017	-0.00034	0.0017	-0.00041
Color	-0.00016	-0.0042	0.00015	-0.00099	0.0014	-0.00061	0.0015	0.00067	0.0021	0.00053	1	0.0014	-0.00061	-0.00099	-0.00029	0.0015	-0.0044	-0.0013
Fashion Magazines	-0.00048	0.00048	0.00043	-5.3e-05	1	0.00089	0.00067	-0.0015	0.00027	0.00077	0.0014	1	0.00089	-5.3e-05	-5.3e-06	0.00079	-0.0019	7.3e-05
Fashion Influencers	-0.00064	0.00019	0.0014	0.0018	0.00089	1	0.00044	0.00085	0.00034	0.0018	-0.00061	0.00089	1	0.0018	-0.0012	0.0024	0.0024	-0.00054
Season	0.00015	0.0034	0.00014	1	-5.3e-05	0.0018	-9.3e-05	0.00085	0.0023	-4.3e-05	-0.00099	-5.3e-05	0.0018	1	-0.00047	-0.00015	0.0019	-0.0028
Time Period Highest Purchase	-0.00033	0.0012	0.00038	-0.00047	-5.3e-06	-0.0012	0.00012	0.0007	0.001	0.0017	-0.00029	-5.3e-06	-0.0012	-0.00047	1	0.00031	-0.0017	0.00031
Customer Reviews	-0.0016	0.0015	3.3e-05	-0.00015	0.00079	0.0024	-0.0025	0.00026	0.002	-0.00034	0.0015	0.00079	0.0024	-0.00015	0.00031	1	0.00045	0.00063
Social Media Comments	0.0025	0.00052	-0.0027	0.0019	-0.0019	0.0024	-0.00018	0.00085	0.0023	0.0017	-0.0044	-0.0019	0.0024	0.0019	-0.0017	0.00045	1	-0.001
feedback	-0.0014	-0.00052	-0.0011	-0.0028	7.3e-05	-0.00054	0.00034	0.00017	0.00053	-0.00041	-0.0013	7.3e-05	-0.00054	-0.0028	0.00031	0.00063	-0.001	1

Figure 3: Correlation Plot of Dataset

Analysis of the Elbow Method to Determine Optimal Clusters

To decide on the positioning of the K-means clustering algorithm, the elbow method was used. Thus, the visualisation of a within-cluster sum of squares (WCSS) against the number of clusters exhibited an “elbow point” at four, which means that the inclusion of four clusters would be appropriate because it optimally balances the model’s complexity and the quality of the clustering.

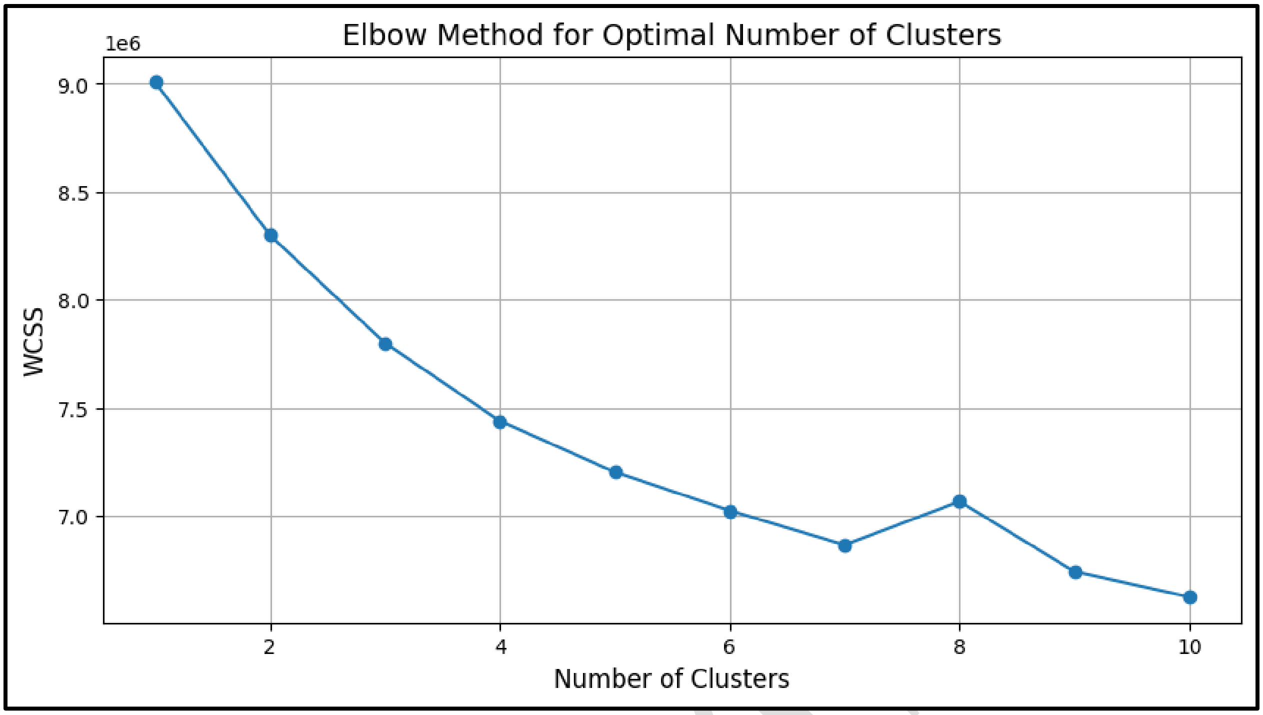


Figure 4: Plot of Elbow Method

Evaluation of Clustering Results Using Silhouette Score

The measurement of the quality of clustering was done using the silhouette score. With an average silhouette score of 0. 07, thus the clustering result was moderate. The silhouette plot for the various clusters depicted the silhouette scores for the clusters where some of the clusters had better coherency and even better separation than others.

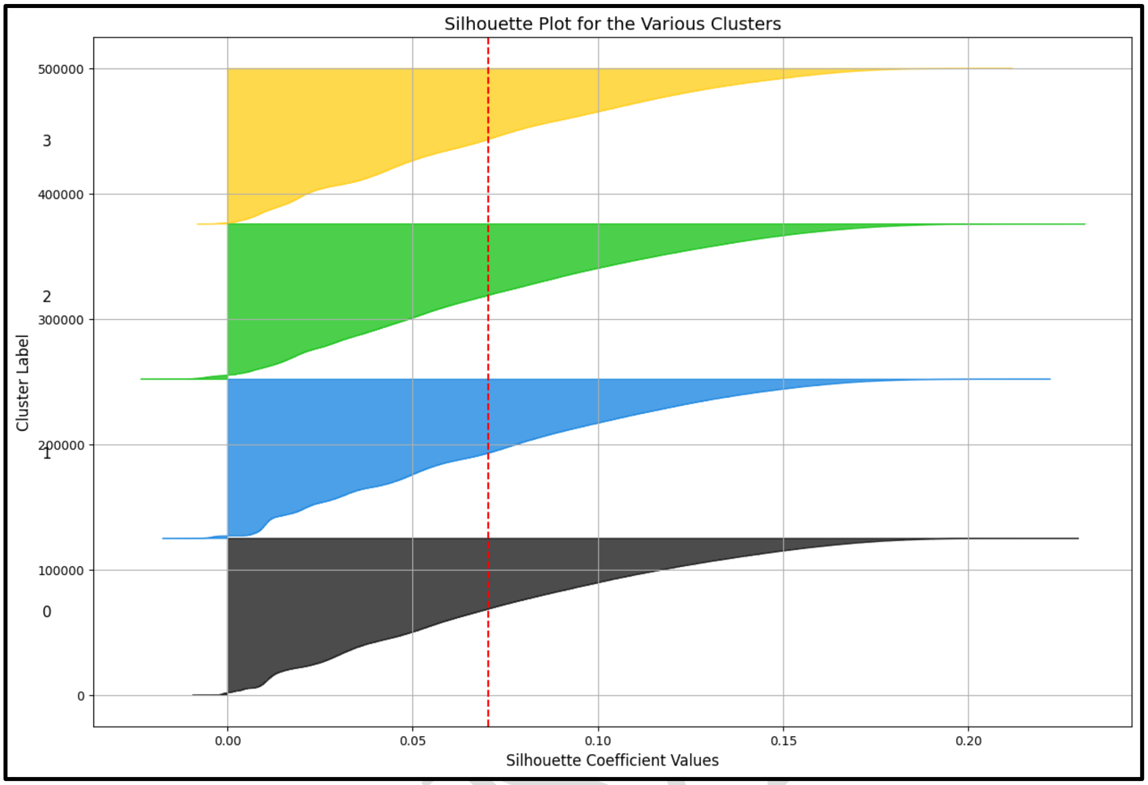


Figure 5: Silhouette Plot for the Clusters

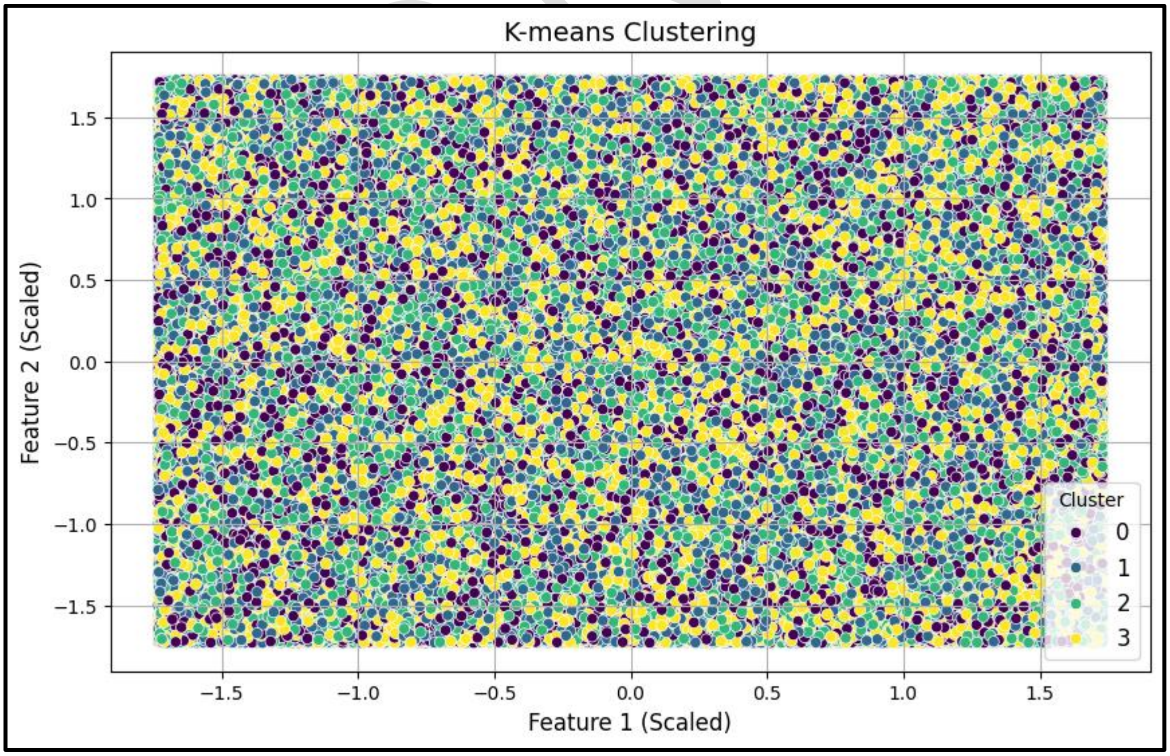


Figure 6: K-means Clustering Plot of the dataset

Visualisation and Interpretation of Clusters

The nature of the clustered dataset was then depicted to understand the nature of each formed cluster. The heatmap of cluster feature means showed more granularity of the average values of features visible in the analysed clusters. For instance, the first cluster was seen to be soaring higher in the average rating though the second cluster contained the highest numbers of reviews. This differentiation proves useful in determining the profiles of each cluster.

Table 3: Means of Various Clusters

	Price	Brand	Category	Description	Rating	Review Count	Style Attributes
Cluster 0	54.86762	3.511223	4.486213	3.005504	3.00329	248.734	4.504465
Cluster 1	55.03288	3.495821	4.504218	2.992446	3.000401	250.0371	4.506093
Cluster 2	55.05233	3.521806	4.493557	3.008383	3.008339	248.8565	4.50257
Cluster 3	55.11539	3.478164	4.539306	2.986185	2.988528	250.5569	4.496901

Color	Age	Fashion Magazines	Fashion Influencers	Season	Highest Purchase	Customer Reviews	Social Media Comments	feedback
1.50127	41.03441	1.897043	6.749361	3.101144	1.992459	1.998626	2.514666	2.505791
1.500319	41.00792	7.218439	6.771148	2.152965	2.001221	2.007743	2.493253	2.494435
1.508202	41.0124	5.480413	1.915577	3.876112	1.996272	1.988276	2.508799	2.517376
1.48721	40.99357	3.373416	2.484241	0.857488	2.009481	2.002187	2.479504	2.494771

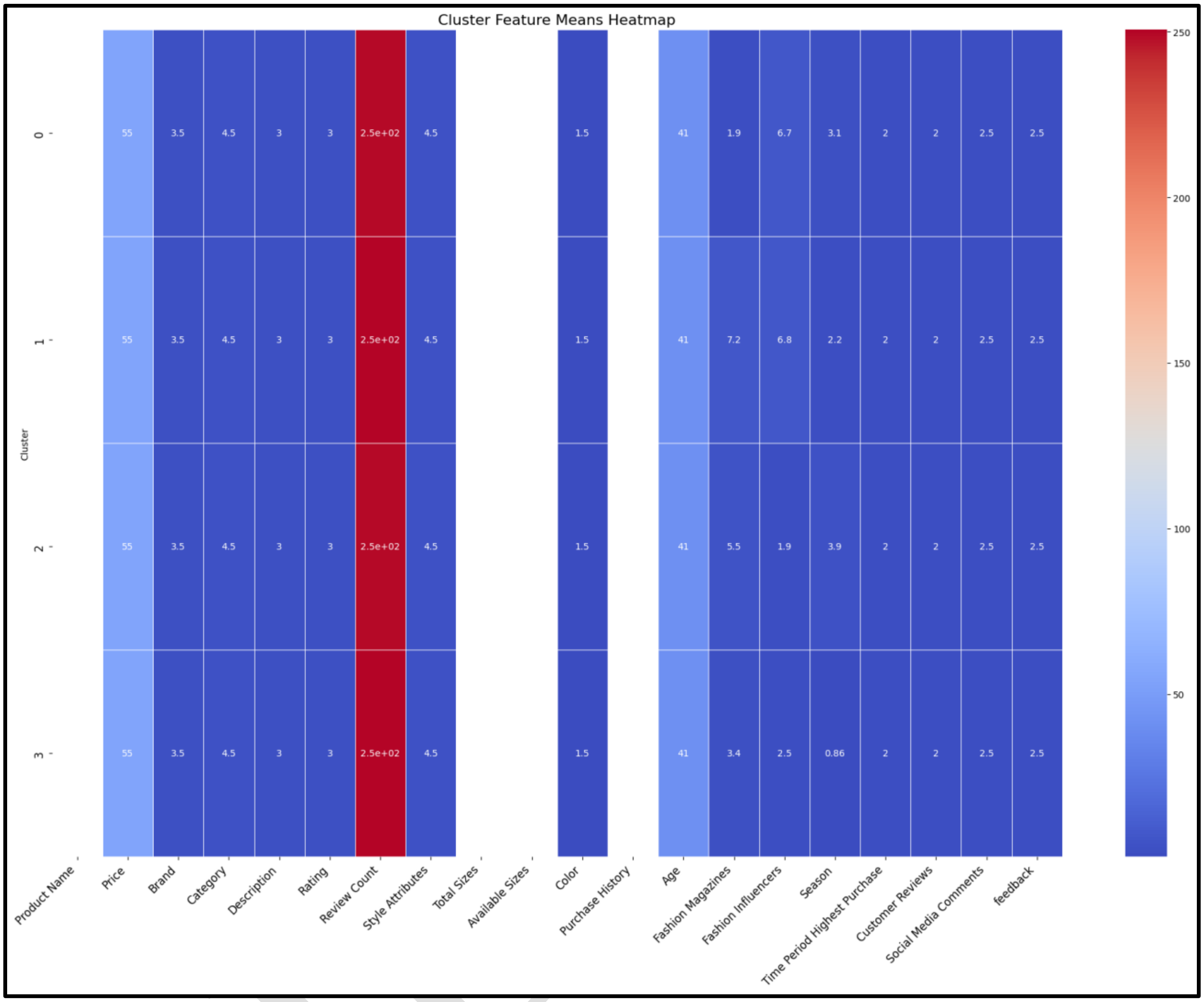


Figure 7: Heatmap plot of Clusters

Discussion

In light of this, this paper seeks to employ K-means clustering in the fashion industry of the UK to generate great insights that could revolutionise the management facets of business. For instance, by categorising the dataset into clusters, the organisation's strategies will correspond to the mechanisms that define the clusters. For instance, the products of Cluster 1 with high marks and moderate prices can be advertised more actively, whereas the measures to increase the reviews' number can be oriented toward the products of Cluster 2.

At the same time, specifying control variables such as seasonality or the impact of fashion magazines and influencers adds one more layer of thinking about the business. For instance, inventory management of products used during the Fall/Winter season would be different based on the stock of products used during Spring.

Therefore, clustering shares a durable foundation of the action of the fashion markets to divide various markets to increase the efficiency of the market and improve customer satisfaction. The application of complex solutions in data management as described in this paper demonstrates the further enhancement of operating results and strategic positioning in the rapidly evolving fashion industry.

CONCLUSION AND RECOMMENDATIONS

The objective of this research was to explore the new methods of handling automation data and the importance of management operations within the fashion industry in the UK. A detailed dataset of the industry was analysed with the help of K-means clustering, which brought some important insights. A brief overview of many distributions and relations of the major variables such as price, rating, review count, colour, brand, material, size, age and gender was uncovered by the exploratory data analysis phase. Thus, based on the elbow method, four clusters were found to strike the most favourable balance between the model's complexity and its efficiency. Altogether, while the moderate silhouette score demonstrates that the clustering is reasonable, the measures agreed with each cluster's properties and the visualisations of the clusters provide characteristics that are helpful in decision-making.

Therefore, the following recommendations can be made for the UK fashion industry based on the key patterns identified in the study. Firstly, market targeting strategies for the selected audience should be used. For instance, the products in Cluster 1, with high ratings and mid-range prices, need to be promoted intensively: by invoking positive customer' testimonials and encouraging references from popular influencers. Likewise, Cluster 2 with the highest review count depicts an energetic client base. To increase the effectiveness of marketing strategies, it is necessary to channel customers' encounters into buying decisions utilising techniques such as couponing.

Clustering information can also be of help in inventory management. Evaluating the product seasonal characteristics based on clustering, seasonal planning should provide for the replenishment of the products that are in high demand during certain seasons. Also, the planning of future demand can be much more accurate, which will mean that the dangers of overbuying and, on the contrary, being left without enough stock, will be minimised.

About customer relationship management the following should be considered the analysis of feedback should be integrated. Low and negative feedback-rated products should be carefully observed and attempts should be made to ensure that the consumers' complaints are handled appropriately which can involve changing the quality of the product or intensifying customer relations. Recommendations according to the data of clusters may also improve customer relations to help those figuring among the buyers of the high-rated products to get suggestions most suitable for them. On the matter of product, the clustering can again be beneficial in sensing the customer needs' shift in focus in the future for new products to be developed. For instance, the popularity of specific styles/attributes in some clusters can be useful in stimulating and creating new products.

Identification of Gaps and Suggestions for Future Research

In the continuation, the following four aspects should be considered to improve the efficiency of data management techniques for the fashion industry. Therefore, the improvement of data collection is essential. Sales data and customers' real-time browsing history should be included as finer-grained features for increasing the functioning of clustering and its results. It is possible to extend the attribute set to encompass material quality, return rates, and specific demographic information, which can broaden the understanding of customer's preferences.

More research should also be done on advanced clustering techniques. It might be useful to compare K-means with other clustering algorithms like DBSCAN or hierarchy ones to get more profound insights as well as perform a better segmentation. Combined with other methods of machine learning like, predictive analytics, for instance, the clustering models improve the strategic forecasting and decision-making results of the insights.

The other type of research that facilitates the prevention of changes is longitudinal research which enables the evaluation of the changes in customer preferences and the market trends frequently. Such studies can be used to support the organisation's long-term planning and assess the outcomes of recommendations implemented in practice. Another valuable approach is the cross-industry analysis, so comparisons may be made between companies of different industries. To analyse this, comparing this fashion industry study’s findings and methods to others from other industries like technology or healthcare can expose certain fashion industry issues and possibilities.

Therefore, it is imperative to establish that automation data management techniques, especially clustering, present numerous benefits for enhancing the management aspect in today’s fashion industry in the UK. By implementing the recommendations and utilising the identified strategies in managing the gaps, businesses can improve their strategic decisions and satisfaction of their customers besides competing effectively in the ever-changing conditions of the business environment. Further research should be devoted to the constant development of the presented approaches and their adaption to current tendencies in the industry.

References

Accenture. (2021). *Transforming the fashion supply chain*. Available at: https://www.accenture.com/content/dam/accenture/final/a-com-migration/r3-3/pdf/pdf-115/accenture-threads-that-bind.pdf (Accessed: 08 June 2024).

Babu, M.M., Rahman, M., Alam, A. and Dey, B.L., 2024. Exploring big data-driven innovation in the manufacturing sector: evidence from UK firms. Annals of Operations Research, 333(2), pp.689-716.

Balchandani, A. *et al.* (2023) *The state of fashion 2024: Finding pockets of growth as uncertainty reigns*, *McKinsey & Company*. Available at: https://www.mckinsey.com/industries/retail/our-insights/state-of-fashion (Accessed: 08 June 2024).

Chase, C.W., 2021. Consumption-based forecasting and planning: Predicting changing demand patterns in the new digital economy. John Wiley & Sons.

Deloitte. (2021). *Digital transformation in the fashion industry* (2019) *Deloitte Switzerland*. Available at: https://www2.deloitte.com/ch/en/pages/consumer-industrial-products/articles/ultimate-challenge-fashion-industry-digital-age.html (Accessed: 08 June 2024).

Fang, C. and Liu, H., 2021. Research and application of improved clustering algorithm in retail customer classification. *Symmetry*, *13*(10), p.1789.

Govender, P. and Sivakumar, V., 2020. Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Atmospheric pollution research*, *11*(1), pp.40-56.

Han, J., Pei, J. and Tong, H., 2022. *Data mining: concepts and techniques*. Morgan kaufmann.

Hassan, I.H., Abdullahi, M. and Ali, Y.S., 2021. Analysis of Techniques for Selecting Appropriate Number of Clusters in K-means Clustering Algorithm. *no. November*.

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. *Pattern recognition letters*, *31*(8), pp.651-666.

Lloyd, S., 1982. Least squares quantization in PCM. *IEEE transactions on information theory*, *28*(2), pp.129-137.

López-Oriona, Á., Montero-Manso, P. and Vilar, J.A., 2022, September. Clustering of time series based on forecasting performance of global models. In *International Workshop on Advanced Analytics and Learning on Temporal Data* (pp. 18-33). Cham: Springer International Publishing.

Mohiuddin Babu, M., Akter, S., Rahman, M., Billah, M.M. and Hack-Polay, D., 2022. The role of artificial intelligence in shaping the future of Agile fashion industry. Production Planning & Control, pp.1-15.

PwC. (2021). The future of retail and consumer goods: Digital transformation and disruption. Retrieved from https://www.pwc.com/gx/en/industries/retail-consumer/consumer-markets-insights/digital-transformation.html

Sarma, A., Guo, S., Hoffswell, J., Rossi, R., Du, F., Koh, E. and Kay, M., 2022. Evaluating the use of uncertainty visualisations for imputations of data missing at random in scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, *29*(1), pp.602-612.

Tan, P.N., Steinbach, M. and Kumar, V., 2016. *Introduction to data mining*. Pearson Education India.

Testas, A., 2023. Support Vector Machine Classification with Pandas, Scikit-Learn, and PySpark. In *Distributed Machine Learning with PySpark: Migrating Effortlessly from Pandas and Scikit-Learn* (pp. 259-280). Berkeley, CA: Apress.

Umargono, E., Suseno, J.E. and Gunawan, S.V., 2020, October. K-means clustering optimization using the elbow method and early centroid determination based on mean and median formula. In *The 2nd international seminar on science and technology (ISSTEC 2019)* (pp. 121-129). Atlantis Press.

Xu, R. and Wunsch, D., 2005. Survey of clustering algorithms. *IEEE Transactions on neural networks*, *16*(3), pp.645-678.

Appendix

K-Means Clustering Implementation

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.cluster import MiniBatchKMeans
from sklearn.metrics import silhouette_score, silhouette_samples
import matplotlib.pyplot as plt
import seaborn as sns
import matplotlib.cm as cm
import warnings

# Suppressing the warnings
warnings.filterwarnings("ignore")

# Loading the dataset
file_path = '/content/drive/MyDrive/data.csv'
data = pd.read_csv(file_path)

# Displaying the first few rows of the dataset and checking the column names
print("First few rows of the dataset:")
print(data.head())
print("\nColumn names in the dataset:")
print(data.columns)

# Sample the dataset (e.g., take 50% of the data if too large)
sampled_data = data.sample(frac=0.5, random_state=42)

# Updating the categorical features based on actual column names
categorical_features = ['Brand', 'Category', 'Description', 'Style Attributes', 'Color', 'Fashion Magazines', 'Fashion Influencers', 'Season', 'Time Period Highest Purchase', 'Customer Reviews', 'Social Media Comments', 'feedback']
```

```
control_variables = ['Season', 'Fashion Magazines', 'Fashion Influencers']

# Handling the missing values for numeric columns only
numeric_cols = sampled_data.select_dtypes(include=[np.number]).columns
sampled_data[numeric_cols] =
sampled_data[numeric_cols].fillna(sampled_data[numeric_cols].mean())

# Encoding categorical variables
label_encoders = {}
for feature in categorical_features:
    if feature in sampled_data.columns:
        le = LabelEncoder()
        sampled_data[feature] =
le.fit_transform(sampled_data[feature].astype(str))
        label_encoders[feature] = le
    else:
        print(f"Warning: Column '{feature}' does not exist in the dataset
and will be skipped.")

# Removing non-numeric columns explicitly
non_numeric_features = ['Product Name', 'Total Sizes', 'Available Sizes',
'Purchase History']
all_features = ['Price', 'Rating', 'Review Count'] + control_variables +
categorical_features
all_features = [feature for feature in all_features if feature not in
non_numeric_features and feature in sampled_data.columns]

# Ensuring all selected features are numeric and present
sampled_data = sampled_data.apply(pd.to_numeric, errors='coerce')
sampled_data = sampled_data.dropna(subset=all_features)

# Printing the first few rows of the cleaned dataset
print("\nFirst few rows of the cleaned dataset:")
print(sampled_data.head())
```

```
# Displaying summary statistics for the cleaned dataset
print("\nSummary statistics for the cleaned dataset:")
print(sampled_data.describe())

# EDA: Distribution of numeric features
plt.figure(figsize=(20, 15))
for i, feature in enumerate(numeric_cols, 1):
    plt.subplot(4, 4, i)
    sns.histplot(sampled_data[feature], kde=True)
    plt.title(f'Distribution of {feature}', fontsize=14)
    plt.xlabel(feature, fontsize=12)
    plt.ylabel('Frequency', fontsize=12)
plt.tight_layout()
plt.show()

# EDA: Correlation heatmap
plt.figure(figsize=(20, 15))
sns.heatmap(sampled_data[all_features].corr(), annot=True,
cmap='coolwarm', linewidths=0.5, annot_kws={"size": 10})
plt.title('Correlation Heatmap', fontsize=16)
plt.show()

# EDA: Pairplot of selected features
selected_features_for_pairplot = ['Price', 'Rating', 'Review Count'] +
control_variables
sns.pairplot(sampled_data[selected_features_for_pairplot], height=2.5)
plt.suptitle('Pairplot of Selected Features', fontsize=16)
plt.show()

# EDA: Boxplot of numeric features by control variables
plt.figure(figsize=(20, 15))
for i, feature in enumerate(control_variables, 1):
    plt.subplot(3, 1, i)
    sns.boxplot(x=feature, y='Price', data=sampled_data,
palette='viridis')
```

```
    plt.title(f'Price by {feature}', fontsize=14)
    plt.xlabel(feature, fontsize=12)
    plt.ylabel('Price', fontsize=12)
plt.tight_layout()
plt.show()

# Standardise the data to have a mean of ~0 and a variance of 1
scaler = StandardScaler()
X_scaled = scaler.fit_transform(sampled_data[all_features])

# Determining the optimal number of clusters using the elbow method
wcss = []
for i in range(1, 11):
    kmeans = MiniBatchKMeans(n_clusters=i, init='k-means++', max_iter=100,
batch_size=100, random_state=42)
    kmeans.fit(X_scaled)
    wcss.append(kmeans.inertia_)

# Ploting the elbow graph
plt.figure(figsize=(10, 5))
plt.plot(range(1, 11), wcss, marker='o')
plt.title('Elbow Method for Optimal Number of Clusters', fontsize=14)
plt.xlabel('Number of Clusters', fontsize=12)
plt.ylabel('WCSS', fontsize=12)
plt.grid(True)
plt.show()

# Based on the elbow graph, choose the optimal number of clusters
optimal_clusters = 4
kmeans = MiniBatchKMeans(n_clusters=optimal_clusters, init='k-means++',
max_iter=100, batch_size=100, random_state=42)
kmeans.fit(X_scaled)

# Predicting the cluster for each data point
clusters = kmeans.predict(X_scaled)
```

```
# Adding the cluster labels to the original dataset
sampled_data['Cluster'] = clusters

# Evaluating the clustering performance using silhouette score
silhouette_avg = silhouette_score(X_scaled, clusters)
print(f'Silhouette Score: {silhouette_avg:.2f}')

# Visualise the clusters using the first two features for simplicity
plt.figure(figsize=(10, 6))
sns.scatterplot(x=X_scaled[:, 0], y=X_scaled[:, 1], hue=clusters,
palette='viridis')
plt.title('K-means Clustering', fontsize=14)
plt.xlabel('Feature 1 (Scaled)', fontsize=12)
plt.ylabel('Feature 2 (Scaled)', fontsize=12)
plt.legend(title='Cluster', fontsize=12)
plt.grid(True)
plt.show()

# Silhouette Plot
plt.figure(figsize=(15, 10))
sample_silhouette_values = silhouette_samples(X_scaled, clusters)
y_lower = 10
for i in range(optimal_clusters):
    ith_cluster_silhouette_values = sample_silhouette_values[clusters ==
i]
    ith_cluster_silhouette_values.sort()
    size_cluster_i = ith_cluster_silhouette_values.shape[0]
    y_upper = y_lower + size_cluster_i

    color = cm.nipy_spectral(float(i) / optimal_clusters)
    plt.fill_betweenx(np.arange(y_lower, y_upper),
                      0, ith_cluster_silhouette_values,
                      facecolor=color, edgecolor=color, alpha=0.7)
    plt.text(-0.05, y_lower + 0.5 * size_cluster_i, str(i), fontsize=12)
    y_lower = y_upper + 10

plt.title("Silhouette Plot for the Various Clusters", fontsize=14)
plt.xlabel("Silhouette Coefficient Values", fontsize=12)
plt.ylabel("Cluster Label", fontsize=12)
plt.axvline(x=silhouette_avg, color="red", linestyle="--")
plt.grid(True)
plt.show()

# Heatmap of Feature Importance
cluster_means = sampled_data.groupby('Cluster').mean()

# Printing the cluster means
print("\nCluster means:")
print(cluster_means)

# Adjusting the heatmap for better readability
plt.figure(figsize=(20, 15))
sns.heatmap(cluster_means, annot=True, cmap='coolwarm', linewidths=0.5,
annot_kws={"size": 10})
plt.title('Cluster Feature Means Heatmap', fontsize=16)
plt.xticks(rotation=45, ha='right', fontsize=12)
plt.yticks(fontsize=12)
plt.tight_layout()
plt.show()

# Saving the clustered data to a new CSV file
output_path = '/content/drive/MyDrive/clustered_data_50.csv'
sampled_data.to_csv(output_path, index=False)

print("Clustered data saved to:", output_path)
```