Pre-processing and Clustering Assignment

Sakib Uddin
Dr. Elodi Lugez
CPS803: Machine Learning
Toronto Metropolitan University

November 30, 2025

Table of Contents

Introduction

This assignment applies machine learning pre-processing and clustering techniques to the “E-shop Clothing 2008 clickstream dataset”, an online browsing dataset containing user interactions with a fashion retail website [1]. This dataset is particularly valuable because it captures real-world consumer behavior, including products viewed, what users tend to click on, and how different factors affect a customer staying or leaving an e-store site.

Background

Dataset

The dataset contains 165,474 rows and 14 features describing user browsing sessions.

Table 1: Definitions and Translations of Features in Dataset

Feature	**Definition**	**Values**	**Datatype**
Year	Year of visit.	2008	Int64
Month	Month of visit.	Ranges from (4-8) 4 = April 5 = May 6 = June 7 = July 8 = August	Int64
Day	Day of visit.	1-30 representing calendar day.	Int64
Order	Sequence of clicks during a session.	1-X: X being the number of total clicks.	Int64
Country	Variable from 1-47 representing country codes.	1-Australia 2-Austria 3-Belgium 4-British Virgin Islands 5-Cayman Islands 6-Christmas Island 7-Croatia 8-Cyprus 9-Czech Republic 10-Denmark 11-Estonia 12-unidentified 13-Faroe Islands	Int64

		14-Finland 15-France 16-Germany 17-Greece 18-Hungary 19-Iceland 20-India 21-Ireland 22-Italy 23-Latvia 24-Lithuania 25-Luxembourg 26-Mexico 27-Netherlands 28-Norway 29-Poland 30-Portugal 31-Romania 32-Russia 33-San Marino 34-Slovakia 35-Slovenia 36-Spain 37-Sweden 38-Switzerland 39-Ukraine 40-United Arab Emirates 41-United Kingdom 42-USA 43-biz (*.biz) 44-com (*.com) 45-int (*.int) 46-net (*.net) 47-org (*.org)	
Session ID	Variable indicating the session the clicks are associated with.	1-X: X being the final session ID.	Int64
Page 1 (Main category)	Category of product clicked.	1-trousers 2-skirts 3-blouses 4-sale	Int64
Page 2 (Clothing Model)	Product code of respective category.	1-217: each representing a different product.	String

Colour	Color of the product.	1-beige 2-black 3-blue 4-brown 5-burgundy 6-gray 7-green 8-navy blue 9-of many colors 10-olive 11-pink 12-red 13-violet 14-white	Int64
Location	Variable representing the location of the product on the page.	1-top left 2-top in the middle 3-top right 4-bottom left 5-bottom in the middle 6-bottom right	Int64
Model Photography	Boolean indicating whether there is a model in the product page.	1-en face 2-profile	Int64
Price	Price in USD.	X : X representing dollar amount.	Int64
Price 2	Boolean indicating whether the price of a product is higher than average price of its product category.	1-yes 2-no	Int64
Page	Page number within the e-store website.	1-5 representing page number within e-store.	Int64

Technologies

- Python 3
- Pandas: data loading and cleaning
- NumPy: numerical computations
- scikit-learn
    - StandardScaler: standardizing numeric features
    - OneHotEncoder: encoding categorical variables
    - ColumnTransformer: combining preprocessing steps
    - K-Means & Agglomerative Clustering: clustering algorithms
    - Silhouette Score: evaluating cluster quality
    - PCA: cluster validity visualization
- Matplotlib: plotting the elbow curve and PCA cluster visualization

Using Pandas, the number of missing values for each feature is checked:

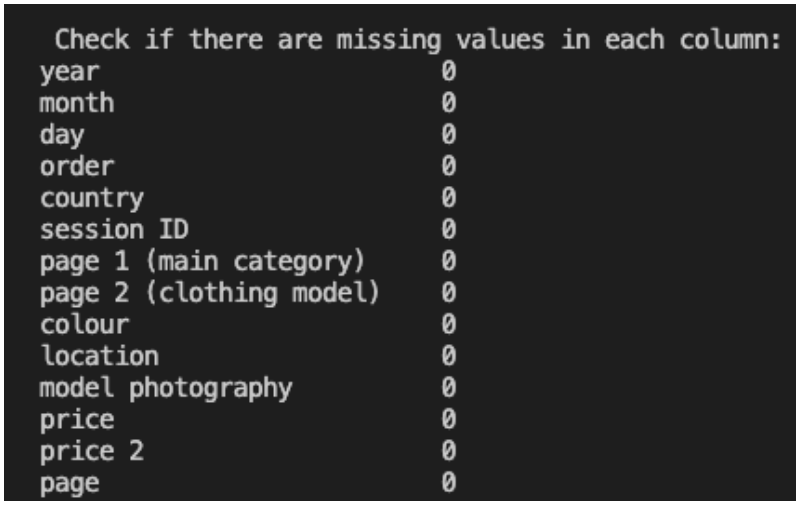

Figure 1: Panda Dataframe representing number of missing values per feature.

Methods

Pre-processing

To prepare the dataset for clustering, several preprocessing steps were applied. Since the dataset contains no missing values, no missing-value handling was required.

1. Remove Irrelevant Data

Several features in the dataset do not contribute to understanding user shopping behavior and therefore were removed prior to clustering:

- Year
- Month
- Day
- Order
- Session ID

These variables are either constant or identifiers, that do not describe how a user interacts with products. Removing them ensures the clustering algorithm focuses only on behavioral features.

2. Remove Duplicates

Duplicate rows were removed using Pandas to ensure that repeated interactions do not bias the clustering process. This step prevents the algorithm from over-representing certain browsing sessions and ensures that each instance in the dataset contributes equally to the overall cluster shape.

3. Frequency Encoding Features

The feature “page 2 (clothing model)” contains 217 unique product models. One-hot encoding this feature would create 217 additional columns, which is computationally expensive. To work around this, frequency encoding was used. Frequency encoding maps each category to the number of times it appears in the dataset; capturing its popularity [2]. A new feature, “page_2_freq”, was created to hold these popularity scores, and the original categorical column was removed.

4. One-Hot Encode Categorical Features

Several features are categorical and must be encoded numerically for clustering algorithms to process them:

- Country
- Page 1 (main category)
- Colour
- Location
- Model Photography
- Page

5. Standardize Numeric Features

The following numeric features vary widely in scale:

- Price
- Price 2

- Page_2_freq

Because clustering algorithms such as K-Means rely on distance calculations, features with large numeric ranges can dominate the clustering process. To prevent this imbalance, StandardScaler was applied.

Formula : z = (x - u) / s
x = the original feature value
u = the mean of that feature
s = the standard deviation of that feature
z = the standardized value

After applying all preprocessing steps: including dropping irrelevant features, removing duplicates, encoding categorical variables, and scaling numeric features, the dataset was reduced from:

- Original shape: (165,474 rows, 14 columns)
- Preprocessed shape: (3,141 rows, 81 columns)

This reduction removed unnecessary clutter while preserving the statistical structure of the data.

Clustering Methods

Two clustering methods were applied to the preprocessed dataset: K-Means Clustering and Hierarchical Clustering. Using two different algorithms allows us to compare performance and ensure the best clustering method is selected.

K-Means Clustering:

To select the most suitable number of clusters, we applied the Elbow Method. The “elbow point” represents a value of k where the rate of improvement sharply decreases. The optimal number of clusters was found to be 4. This indicates that the dataset naturally forms four meaningful behavioral groups.

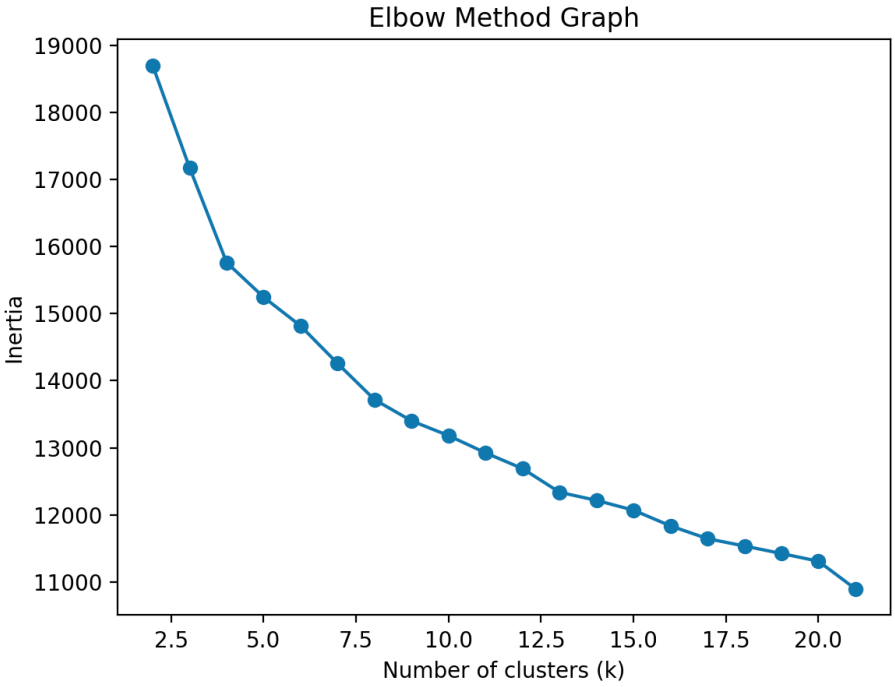

Figure 2: Relationship between number of clusters (k) and sum of squared distances (Inertia).

Hierarchical Clustering

This method is particularly useful for understanding the hierarchical structure of the data, even though it is more computationally expensive for large datasets. Agglomerative Hierarchical Clustering was utilized with k = 4 to maintain consistency with K-Means.

Cluster Validity

Cluster validity was assessed using two methods: Silhouette Score and PCA visualization. These tools help determine how well-separated and meaningful the clusters are.

Silhouette Score

The silhouette score measures both cluster cohesion and separation [3]. The formula for the silhouette score is:

s = b - a / max(a, b)

- a = average distance between a point and all other points in the same cluster
- b = average distance between a point and all points in the nearest different cluster
- s = silhouette value

Silhouette results:

- K-means : 0.1335
- Hierarchical : 0.1213

K-Means achieved a higher silhouette score and therefore produced more coherent clusters.

Principal Component Analysis (PCA)

PCA was applied after preprocessing to reduce the dataset into just two dimensions. This reduction allows us to visualize the structure of the dataset and assess how well the clustering algorithm separates the data into meaningful groups. PCA helped verify that the clustering algorithm is not being misled by noise or redundant features.

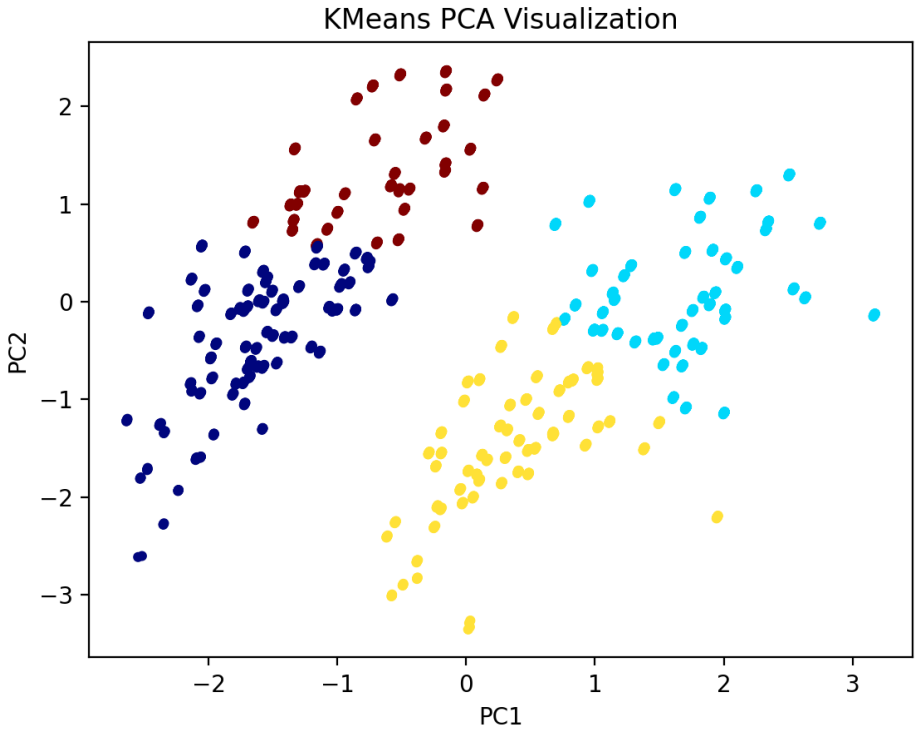

Figure 3: PCA visualization of K-Means.

According to the PCA graph, there are clear boundaries that indicate that the clusters follow trends that are meaningfully different from each other.

Results

Table 2: Final Results Yielded by K-Means Clustering

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Cluster Size	906	879	688	668
Average Price	31.81	55.76	46.08	35.51
Average Price 2	2.0	1.0	1.0	2.0
Average page_2_freq (Clothing model)	12.87	19.53	12.55	19.99
Countries	29 (8.5%), 24 (8.3%), 46 (7.6%), 16 (6.8%), 44 (6.7%)	9 (5.2%), 16 (5.2%), 24 (5.2%), 29 (5.2%), 41 (5.2%)	29 (8.7%), 24 (8.3%), 16 (7.6%), 46 (7.6%), 44 (7.4%)	29 (5.2%), 9 (5.1%), 24 (5.1%), 41 (5.1%), 44 (5.1%)
Page 1 (Main category)	4 (41.5%), 3 (29.5%), 1 (19.1%), 2 (9.9%)	2 (38.0%), 3 (25.4%), 1 (25.0%), 4 (11.6%)	4 (52.3%), 3 (37.8%), 1 (5.8%), 2 (4.1%)	1 (46.4%), 2 (23.8%), 3 (19.3%), 4 (10.5%)
Colours	6 (21.3%), 2 (14.7%), 14 (13.5%), 4 (11.5%), 1 (7.6%)	2 (22.2%), 3 (15.7%), 9 (12.1%), 12 (12.1%), 4 (9.1%)	2 (27.5%), 14 (16.4%), 4 (15.8%), 9 (8.9%), 6 (8.0%)	3 (24.7%), 2 (15.0%), 4 (10.5%), 1 (8.7%), 7 (8.7%)
Locations	1 (20.1%), 3 (20.1%), 5 (19.9%), 4 (16.6%), 6 (12.9%)	1 (28.7%), 5 (18.2%), 2 (17.2%), 3 (13.1%), 6 (12.8%)	2 (26.6%), 3 (19.3%), 4 (15.9%), 1 (14.7%), 6 (12.9%)	2 (28.9%), 4 (20.7%), 6 (16.9%), 1 (15.1%), 5 (12.7%)
Model Photography	Type 1: 56.8%, Type 2: 43.2%	Type 1: 87.5%, Type 2: 12.5%	Type 1: 52.3%, Type 2: 47.7%	Type 1: 80.8%, Type 2: 19.2%
Page Depth	pg2 (42.5%), pg3 (30.5%), pg4 (11.5%), pg1 (11.0%), pg5 (4.5%)	pg1 (61.3%), pg2 (30.6%), pg3 (5.9%), pg4 (2.2%)	pg2 (32.8%), pg1 (20.8%), pg3 (19.6%), pg4 (17.6%), pg5 (9.2%)	pg1 (77.8%), pg2 (11.2%), pg3 (10.9%)

Cluster 0:

- Strong interest in sale items (41.5% of their visits)
- Browses many blouses and skirts
- Frequently views products in gray, black, and white
- Often look at items positioned on the top or center of the webpage
- Lower average price: $31.81
- More likely to view items that are below the average price of their category (price2 = no)
- They explore many pages (page 2 and page 3 are dominant)

Cluster 1:

- Strong bias toward blouses (38%) and skirts (25%)
- Tend to choose black, blue, or multicolor products
- Prefer products shot with a model (87.5%)
- Highest average price: $55.76
- More likely to view items priced above average in their category (price2 = yes)
- 61% stay on page 1

Cluster 2:

- Very strong interest in the sale category (52.3%)
- Prefer black, white, and brown items
- Usually view items priced above average (price2 = yes)
- Wide exploration across pages 1-5

Cluster 3:

- Mostly browse trousers (46%) and skirts
- Often pick blue, red, or green products
- Prefer products shot with a model (80.8%)
- Strong tendency to browse items priced below average (price2 = no)
- Extremely focused on page 1 (78%)

Conclusion

Cluster 0 represents high-engagement bargain browsers. These users concentrate heavily on sale items, prefer neutral colours, and explore multiple pages. Their behaviour shows a clear preference for lower-priced items. Cluster 1 captures a focus on blouses and skirts, users tend to lean toward black or blue colours, and overwhelmingly prefer items displayed with a model. They also browse the most expensive items and stay almost exclusively on page 1. Cluster 2 users show a strong preference for the sale category but tend to click on items priced above the average. They explore across all five pages, indicating an incentive to explore sale items. Cluster 3 shows strong interest in trousers and skirts, favouring colours like blue, red, and green. They prefer model-shot photography and tend to look at lower-priced options. K-means was the strongest clustering method as it produced clusters with stronger internal cohesion and clearer separation between groups.

References

[1] Clickstream Data for Online Shopping [Dataset]. (2019). UCI Machine Learning Repository. https://doi.org/10.24432/C5QK7X.

[2] Ninja, N. (2024, April 1). *Frequency encoding: Counting categories for representation*. Let's Data Science. https://letsdatascience.com/frequency-encoding/

[3] *Silhouette_score*. scikit. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html