DS 203: Programming for Data Science - Exercise 5

Swayam Saroj Patel (22B1816)

March 20, 2025

E5-UCI-Wholesale.csv :

Data Preparation:

Missing Values:

Checking into the data just after loading it I checked on the missing values in it. Luckily al the features are completely filled with no missing values.

Variance:

The variance of each feature was computed on the raw data (i.e., before applying any standardization). The variance values (approximated) for each feature were as follows:

Feature	Variance
Fresh	1.599549×10^8
Milk	5.446997×10^7
Grocery	9.031010×10^7
Frozen	2.356785×10^7
Detergents_Paper	2.273244×10^7
Delicassen	7.952997×10^6

Table 1: Variance of features in the Wholesale Customers dataset.

Based on these values, the top 5 features with the highest variance (before standardization) were selected:

Fresh, Milk, Grocery, Frozen, Detergents_Paper

Correlation after Standardization:

	Fresh	Grocery	Milk	Frozen	Detergents_Paper
Fresh	1.000000	-0.011854	0.100510	0.345881	-0.101953
Grocery	-0.011854	1.000000	0.728335	-0.040193	0.924641
Milk	0.100510	0.728335	1.000000	0.123994	0.661816
Frozen	0.345881	-0.040193	0.123994	1.000000	-0.131525
Detergents_Paper	-0.101953	0.924641	0.661816	-0.131525	1.000000

Table 2: Correlation Matrix of Selected Features

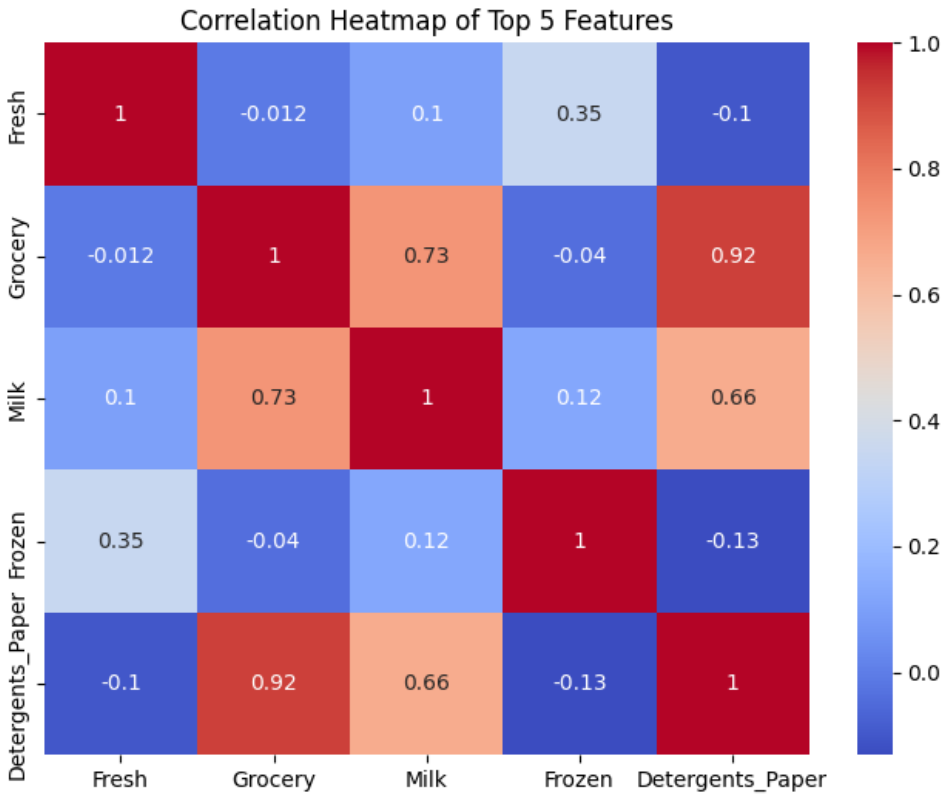

Figure 1: Correlation Heatmap of the Top 5 Features

Clustering Implementation:

K-means:

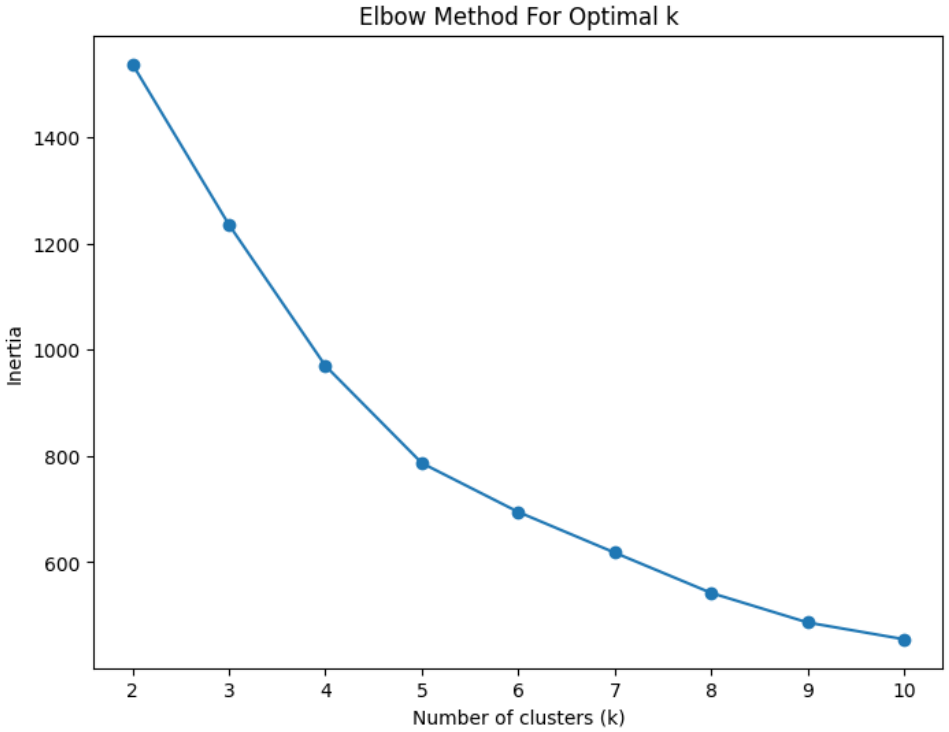

Figure 2: Elbow graph using the top 5 features

From this figure we select the optimal k, and it is chosen where there is a distinct change in the slope. Here i have selected $\mathbf{k = 5}$.
The following the 2-D PCA is :

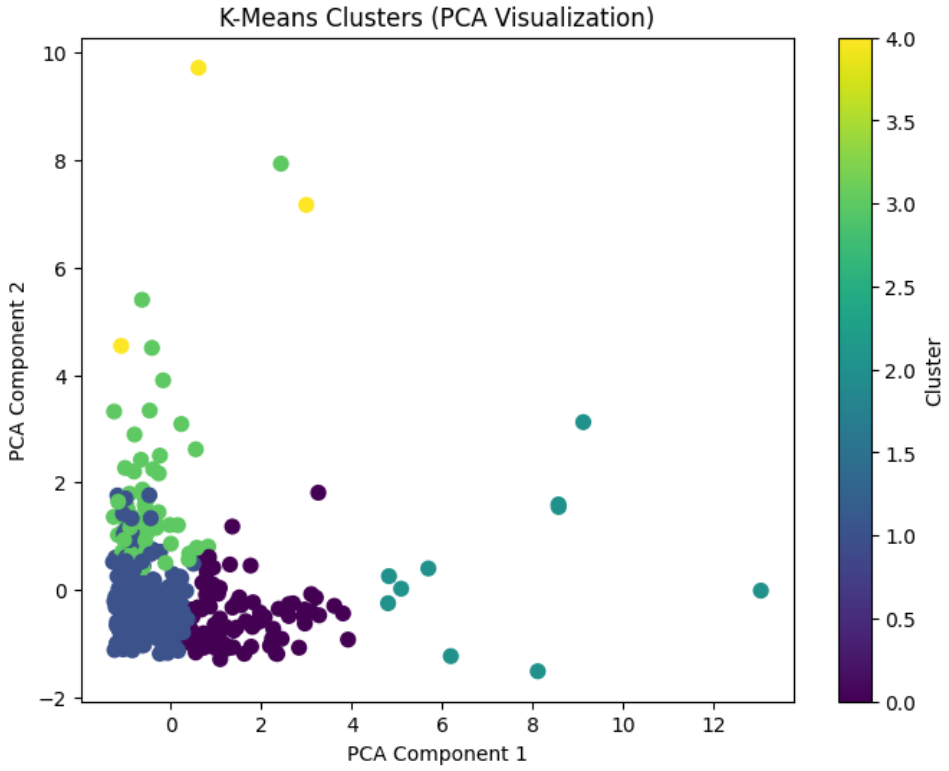

Figure 3: PCA Diagram

In general, silhouette scores range from -1 to 1. Values closer to 1 suggest well-defined, distinct clusters. Scores near 0 indicate overlapping clusters. Negative scores imply that data points may have been assigned to the wrong clusters.
Here the silhouette sore of K-means with optimal k as 5 is **0.397103**. Indicating an overlap in the clusters created.

Hierarchical clustering:

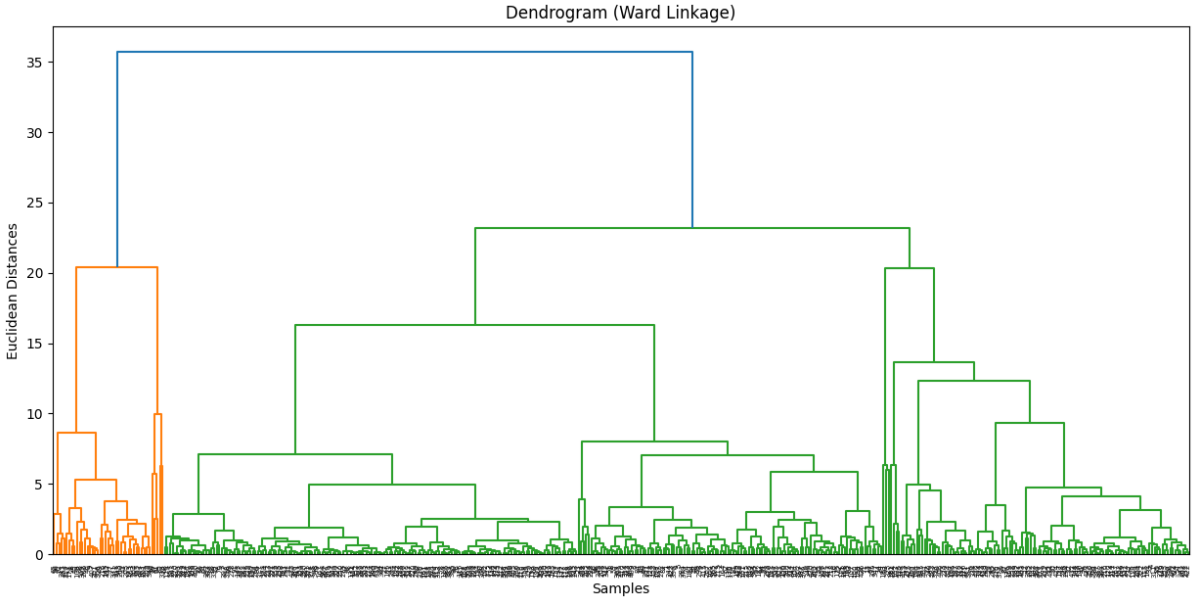

Figure 4: Dendrogram

Here the silhouette score was **0.214461** which indicates that the clusters obtained from hierarchical clustering are not very well separated.

DBSCAN:

Metric	Value
Number of clusters found	2
Percentage of outliers	27.05%
Silhouette Score (excluding outliers)	0.484845

Table 3: DBSCAN Clustering Results

These values suggests that while the clusters themselves are reasonably well-defined, the high outlier rate implies that a considerable amount of data does not meet the density criteria, possibly due to noise or sparsely populated regions. Further parameter tuning or data preprocessing might be necessary to improve the clustering outcome.

Comparative Analysis:

1. *Which algorithm produced the highest silhouette score for your dataset?*

Among the three clustering methods, **DBSCAN** produced the highest silhouette score of approximately 0.484 (excluding outliers). This suggests that, for the points that were clustered, DBSCAN achieved a more compact and well-separated grouping than K-Means (0.397) and Hierarchical Clustering (0.214).

2. *Which method's cluster count seems most appropriate?*

While DBSCAN yielded the highest silhouette score, it also labeled about 27.05% of the data as outliers. In contrast, K-Means was set to produce 5 clusters based on the elbow method and includes all data points. Depending on whether you prioritize having fewer outliers or a higher silhouette score, K-Means may be more appropriate for capturing the complete structure of the dataset.

3. *What percentage of data points were identified as outliers by DBSCAN? What do these outliers represent in your dataset?*

DBSCAN identified approximately 27.05% of the data points as outliers. These outliers likely represent noise or observations that do not belong to any dense region in the dataset—possibly anomalies or transitional cases that do not conform well to the main cluster structures.

4. *For the best-performing algorithm, calculate and report the mean values of each feature for each cluster.*

For the best-performing algorithm (DBSCAN), the following table summarizes the mean values of each selected feature for the two clusters. (Replace the placeholder values with your actual results.)

Feature	**Cluster 0 Mean**	**Cluster 1 Mean**
Fresh	-0.237663	2.302719
Grocery	-0.305082	-0.511538
Milk	-0.285236	-0.496239
Frozen	-0.208437	-0.389627
Detergents_Paper	-0.257709	-0.449334

Table 4: Mean values of each feature for DBSCAN clusters.

Algorithm Comparison:

Metric	**K-Means**	**Hierarchical**	**DBSCAN**
Number of clusters	5	5	2
Silhouette score	0.397103	0.214461	0.484845
Execution time (seconds)	0.135581	0.013421	0.016245
Handles outliers	No	No	Yes

Table 5: Algorithm Comparison Metrics

E5-Credit-Card-Users.csv :

Data Preparation:

Missing values:

Checking on the missing values we get this:

Feature	Value
CUST_ID	0.000000
BALANCE	0.000000
BALANCE_FREQUENCY	0.000000
PURCHASES	0.000000
ONEOFF_PURCHASES	0.000000
INSTALLMENTS_PURCHASES	0.000000
CASH_ADVANCE	0.000000
PURCHASES_FREQUENCY	0.000000
ONEOFF_PURCHASES_FREQUENCY	0.000000
PURCHASES_INSTALLMENTS_FREQUENCY	0.000000
CASH_ADVANCE_FREQUENCY	0.000000
CASH_ADVANCE_TRX	0.000000
PURCHASES_TRX	0.000000
CREDIT_LIMIT	0.011173
PAYMENTS	0.000000
MINIMUM_PAYMENTS	3.497207
PRC_FULL_PAYMENT	0.000000
TENURE	0.000000

Table 6: Feature values for the dataset

so we can see there are some missing values in CREDIT_LIMIT and MINIMUM_PAYMENT. So on how to handle them lets see the how the scatter plot of the features look like.

Figure 5: Scatter plot of the features

Since the plots of the two features we are interested in are kind of random in a certain area we will use **mean** to fill the missing values.

Variance:

Feature	Variance
CREDIT_LIMIT	1.32395×10^7
PAYMENTS	8.381394×10^6
MINIMUM_PAYMENTS	5.431641×10^6
PURCHASES	4.565208×10^6
CASH_ADVANCE	$4.398096e \times 10^6$
BALANCE	4.332775×10^6
ONEOFF_PURCHASES	2.755228×10^6
INSTALLMENTS_PURCHASES	8.178274×10^5
PURCHASES_TRX	6.179027×10^2
CASH_ADVANCE_TRX	$4.657580e \times 10$
TENURE	1.791129
PURCHASES_FREQUENCY	1.610985×10^{-1}
PURCHASES_INSTALLMENTS_FREQUENCY	1.579647×10^{-1}
ONEOFF_PURCHASES_FREQUENCY	8.900441×10^{-2}
PRC_FULL_PAYMENT	8.555578×10^{-2}
BALANCE_FREQUENCY	5.612351×10^{-2}
CASH_ADVANCE_FREQUENCY	4.004857×10^{-2}

Table 7: Variance of features in the dataset

Based on these values, the top 5 features with the highest variance (before standardization) were selected:

CREDIT_LIMIT, PAYMENTS, MINIMUM_PAYMENTS, PURCHASES, and CASH_ADVANCE

Standardization and Correlation:

	CREDIT_LIMIT	PAYMENTS	MINIMUM_PAYMENTS	PURCHASES	CASH_ADVANCE
CREDIT_LIMIT	1.000000	0.421852	0.125134	0.356959	0.303983
PAYMENTS	0.421852	1.000000	0.125046	0.603264	0.453238
MINIMUM_PAYMENTS	0.125134	0.125046	1.000000	0.093515	0.139223
PURCHASES	0.356959	0.603264	0.093515	1.000000	-0.051474
CASH_ADVANCE	0.303983	0.453238	0.139223	-0.051474	1.000000

Table 8: Correlation Matrix of Selected Features

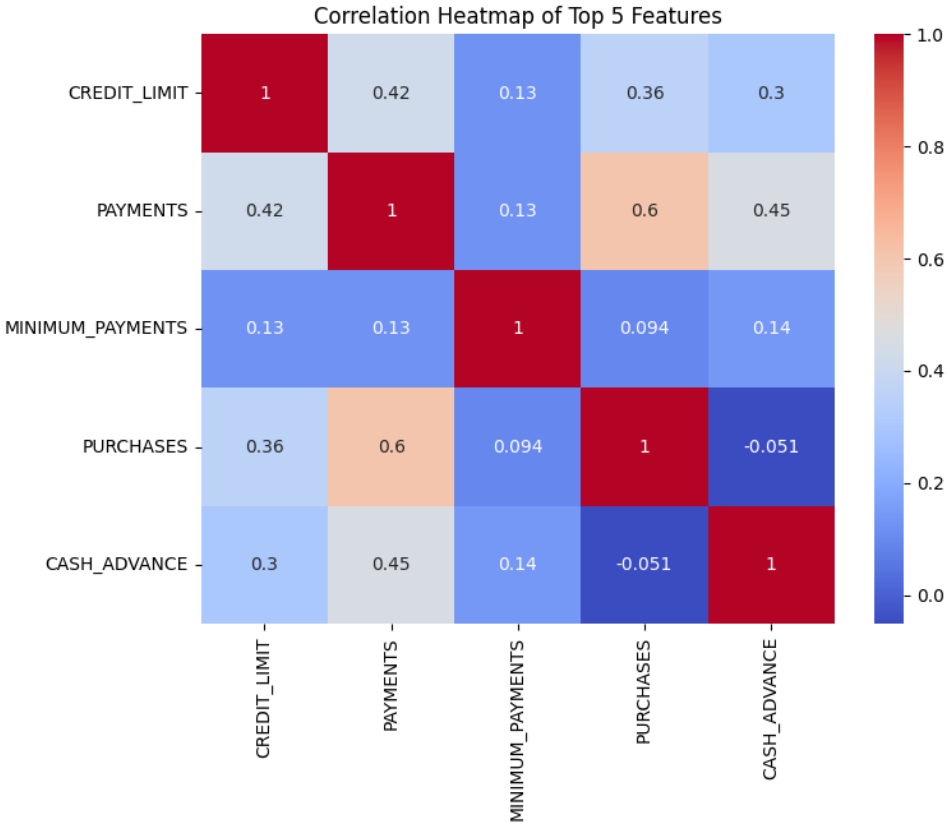

Figure 6: Correlation Map of Selected Features

Clustering Implementation:

K-means:

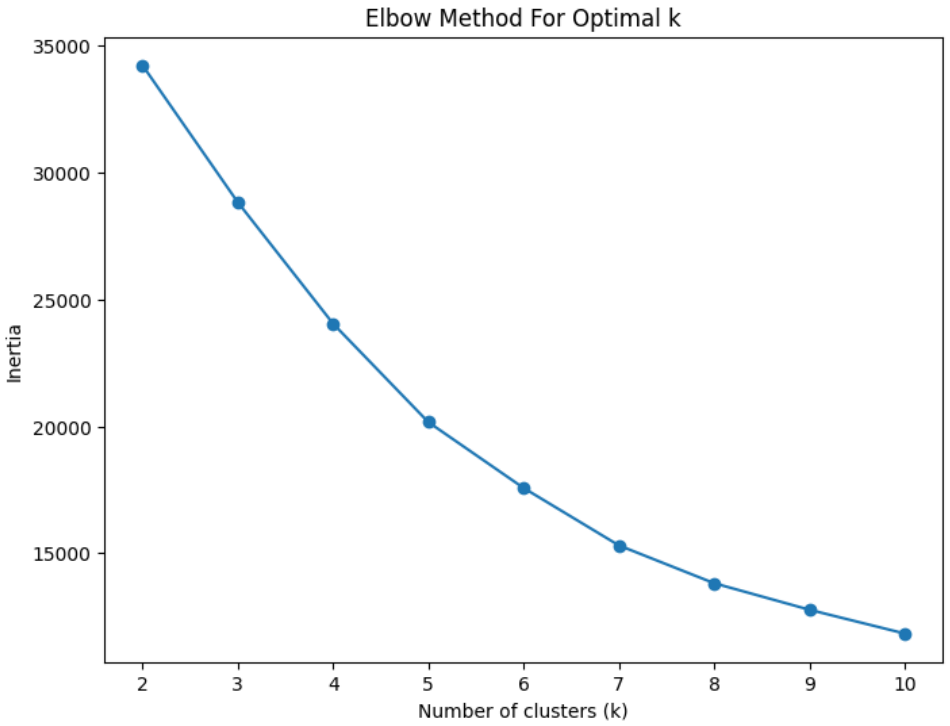

Figure 7: Elbow graph using the top 5 features

From this figure we select the optimal k, and it is chosen where there is a distinct change in the slope. Here i have selected k as 4.
The following the 2-D PCA is :

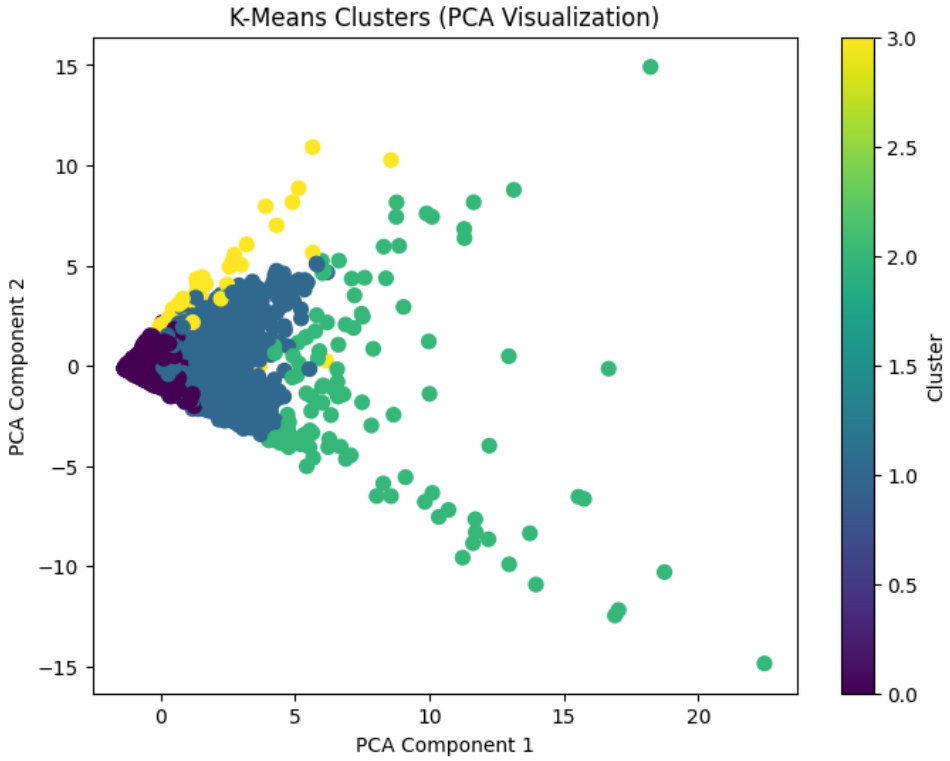

Figure 8: PCA Diagram

Here the silhouette sore of K-means with optimal k as 5 is **0.468732**. Indicating a bit of overlap in the clusters created.

Hierarchical clustering:

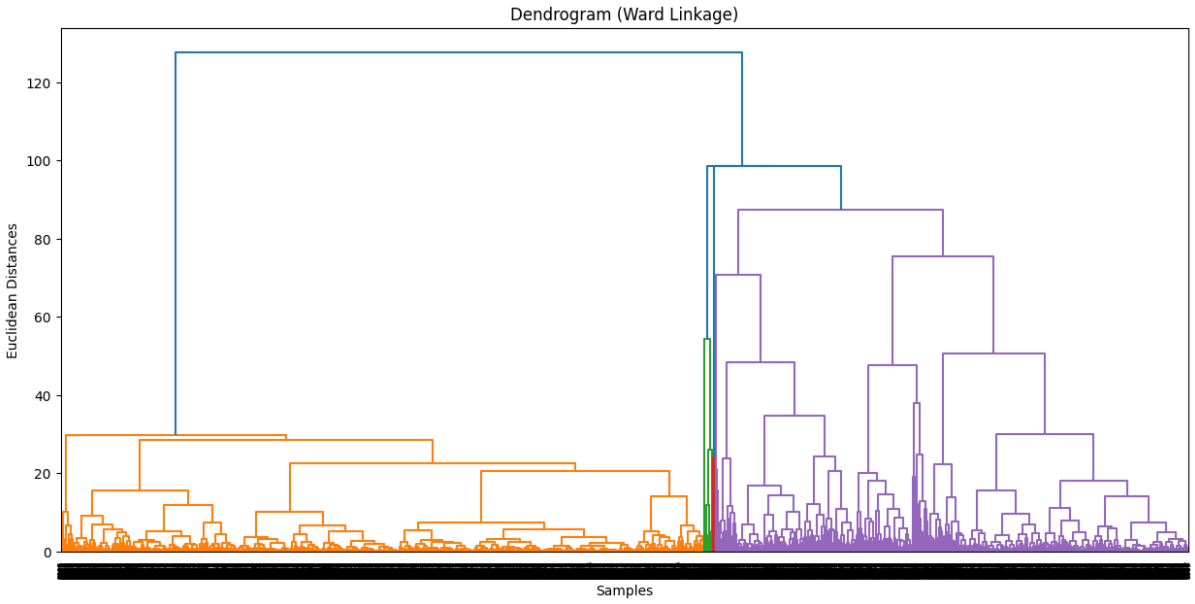

Figure 9: Dendrogram

Here the silhouette score was **0.344612** which indicates that the clusters obtained from hierarchical clustering are not very well separated.

DBSCAN:

Metric	Value
Number of clusters found	11
Percentage of outliers	9.31%
Silhouette Score (excluding outliers)	0.317327

Table 9: DBSCAN Clustering Results

The DBSCAN algorithm identified 11 clusters in the dataset, with a relatively low percentage of data points (approximately 9.31%) labeled as outliers. The silhouette score computed for the inlier clusters (excluding outliers) is around 0.317, indicating that the cluster separation is moderate. Although the low outlier rate suggests that most data points fit within dense regions, the moderate silhouette score implies that the clusters are not highly distinct and may benefit from further tuning of DBSCAN parameters or additional preprocessing.

Comparative Analysis:

1. *Which algorithm produced the highest silhouette score for your dataset?*

K-Means produced the highest silhouette score (0.468732), suggesting that its clusters are more compact and better separated compared to those produced by Hierarchical clustering (0.344612) and DBSCAN (0.317327).

2. *Which method's cluster count seems most appropriate?*

Both K-Means and Hierarchical clustering yielded 4 clusters, which appears to be a more parsimonious segmentation compared to the 11 clusters produced by DBSCAN. Given the higher silhouette score and lower execution time, the 4-cluster solution from K-Means seems more appropriate for capturing the dataset's structure.

3. *What percentage of data points were identified as outliers by DBSCAN? What do these outliers represent in your dataset?*

DBSCAN identified 9.31% of the data points as outliers. These outliers represent observations that do not belong to any dense region and are thus considered noise or anomalies. Their presence indicates that a small portion of the dataset exhibits atypical behavior compared to the main clusters.

4. *For the best-performing algorithm, calculate and report the mean values of each feature for each cluster.*

Since K-Means is the best-performing algorithm based on the silhouette score and execution time, the following table provides a template for reporting the mean values of each feature for the 4 clusters. The actual values should be computed from your dataset.

Cluster	CREDIT_LIMIT	PAYMENTS	MINIMUM_PAYMENTS	PURCHASES	CASH_ADVANCE
0	-0.436066	-0.279021	-0.140731	-0.194120	-0.260351
1	1.227290	0.538084	0.149288	0.336849	0.686210
2	2.214497	5.911589	0.692314	4.775042	2.159087
3	-0.077071	-0.046743	9.312960	0.008617	-0.026540

Table 10: Mean values of selected features for K-Means clusters.

Algorithm Comparison:

Metric	K-Means	Hierarchical	DBSCAN
Number of clusters	4	4	11
Silhouette score	0.468732	0.344612	0.317327
Execution time (seconds)	0.099035	3.850866	1.237302
Handles outliers	No	No	Yes

Table 11: Comparison of Clustering Algorithms