CLUSTERING ANALYSIS

DENNIS MULUMBI KYALO

01-March-2022

Contents

PART A

1. Clustering is an unsupervised machine learning technique that encompasses a wide range of strategies for identifying subgroups, or clusters, in a data set. Clustering can be applied in the following areas:

   - **Cancer Cells detection**
     Several cancer detection techniques make use of clustering algorithms. It separates the malignant from the non-malignant clusters.
   - **Customer Segmentation**
     When conducting market research, Clustering is performed to categorize consumers according to their interests and preferences.
   - **Search Engines**
     Search engines also use the clustering approach. Based on the proximity of the search query to the item in question, a search result is shown to the user. By grouping comparable data items together in a group that is far from the other different subjects, it accomplishes this objective. If the clustering method is of high quality, the accuracy of a query's result will be improved significantly.

2. In order to implement the K-Means Clustering Algorithm, there is a common approach known as the elbow method that is used to identify the ideal value of K.

   - First, we compute the k-means clustering algorithm for a variety of distinct values of the parameter k, for example, by altering the number of clusters (k) from 1 to 15.
   - We then compute the total within-cluster sum of squares(WCSS) for each k in the list.
   - Plot the WCSS curve based on the amount of clusters k present.
   - The number of clusters needed in a plot is typically determined by the position of the plot's bend (or "elbow").

3. **i. Hierarchical Clustering**

   Advantages

   - The hierarchical structure provides valuable information. As a result, by examining the dendrogram, it is much simpler to determine the number of clusters to be used.
   - Easy to implement.

   Disadvantages

   - Large datasets are not suited for this method because of its time complexity.
   - Highly sensitive to outliers.
   - The sequence in which the data is presented has an effect on the final outcome.

ii. K-Means Clustering

Advantages

- Easy to implement.
- When dealing with a large number of variables, K-Means may be the most computationally effective method.

Disadvantages

- There is a lot of uncertainty on the number of clusters (K-Value) to be used.
- Sensitive to scale: rescaling your datasets will result in a significant change in findings.

4. K-means clustering is the most commonly used clustering algorithm. It's a centroid-based algorithm and the simplest unsupervised learning algorithm.

   Feature selection is used to remove features from your dataset that are unnecessary or redundant. While both feature selection and extraction are important, the primary distinction between the two is that feature selection preserves a subset of the original features. In contrast, feature extraction develops entirely new ones.

   When faced with a huge set of correlated variables, principal components allow us to reduce the set down to a few representative variables that collectively account for the bulk of the variance in the original dataset.

   PCA detects the highest eigenvectors of a covariance matrix and utilizes them to project the data onto a new subspace of equal or fewer dimensions. Feature extraction using PCA, or Principle Component Analysis, is very popular. Applying PCA, the highest-valued eigenvectors of a covariance matrix are identified and used to project the data into a new subspace with an eigenvalue that is equal to or less than the original eigenvalues.

   Advantages of PCA

   - Removes Correlated Features.
   - Reduces the number of dimensions in the algorithm, which improves its performance: The algorithm's training time is greatly reduced with fewer dimensions.
   - Minimizes the chances of overfitting data: Overfitting occurs most often when a dataset has too many variables. As a result, by reducing the number of features, PCA helps in the correction of the overfitting problem.

   Disadvantages of PCA

   - Since PCA only works with numerical data, categorical characteristics need encoding.
   - The loss of information: It is true that Principal Components attempt to account for as much variation in a dataset as possible, but the number of Principal Components we use might overlook important information if we do not choose them appropriately.
   - It is necessary to normalize your data before performing PCA; otherwise, PCA will be unable to identify the ideal Principal Components.

PART B

1. INTRODUCTION (US Arrests Data)

We begin by skimming through the dataset to understand its contexts.

Table 1: Data summary

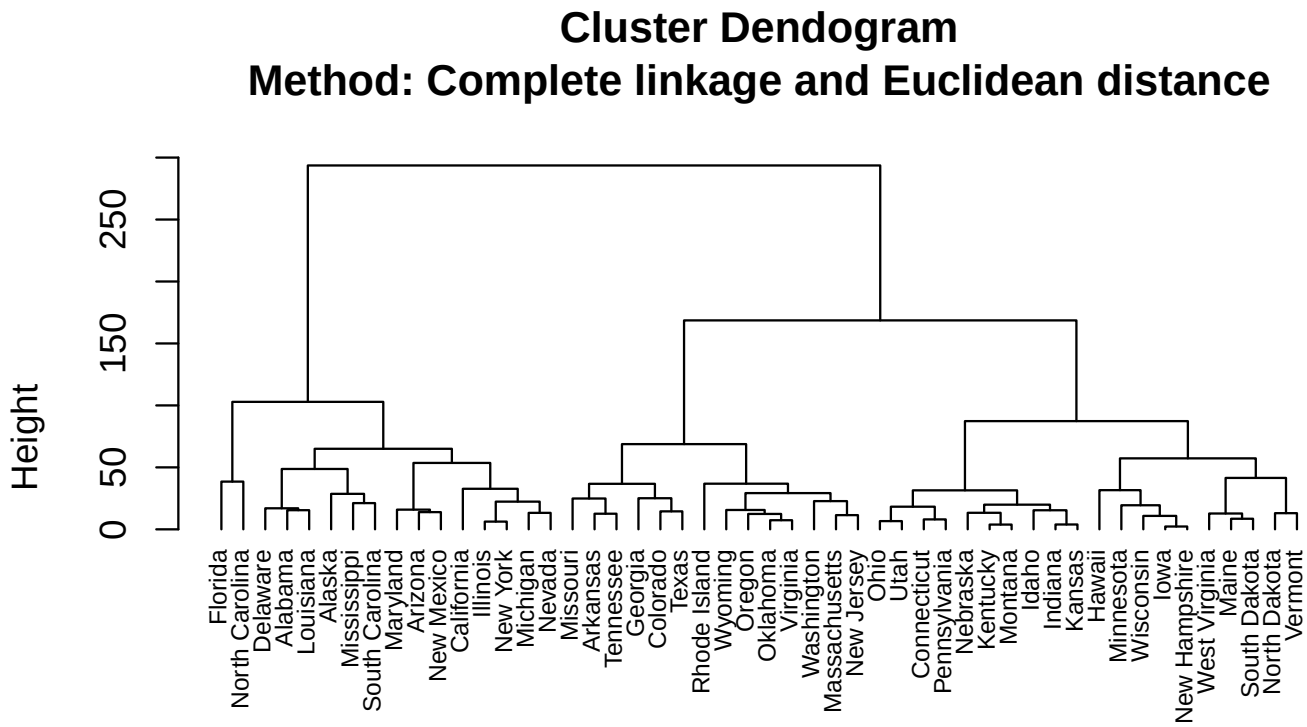
Name	arrests_tbl
Number of rows	50
Number of columns	4
_____	
Column type frequency:	
numeric	4
_____	
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Murder	0	1	7.79	4.36	0.8	4.08	7.25	11.25	17.4
Assault	0	1	170.76	83.34	45.0	109.00	159.00	249.00	337.0
UrbanPop	0	1	65.54	14.47	32.0	54.50	66.00	77.75	91.0
Rape	0	1	21.23	9.37	7.3	15.08	20.10	26.17	46.0

The USArrests dataset contains 50 samples and four variables. The variables include Murder, Assault, UrbanPop and Rape, as seen in the table above.

1a. Hierachical Clustering (No feature scaling)



The cluster dendrogram shown above depicts the clustered states obtained after performing hierarchical clustering with complete linkage and Euclidean distance on the input data. In this figure, we can clearly see the numerous branches formed by the clustering approach. We performed this approach without scaling the dataset.

1b. Cut the dendrogram

We cut the dendrogram at a height that results in three distinct clusters.

Table 3: Cluster Frequency

Cluster	Frequency
1	16
2	14
3	20

Table 4: Clustered States

State	Cluster
Alabama	1
Alaska	1
Arizona	1
Arkansas	2
California	1
Colorado	2
Connecticut	3
Delaware	1
Florida	1
Georgia	2

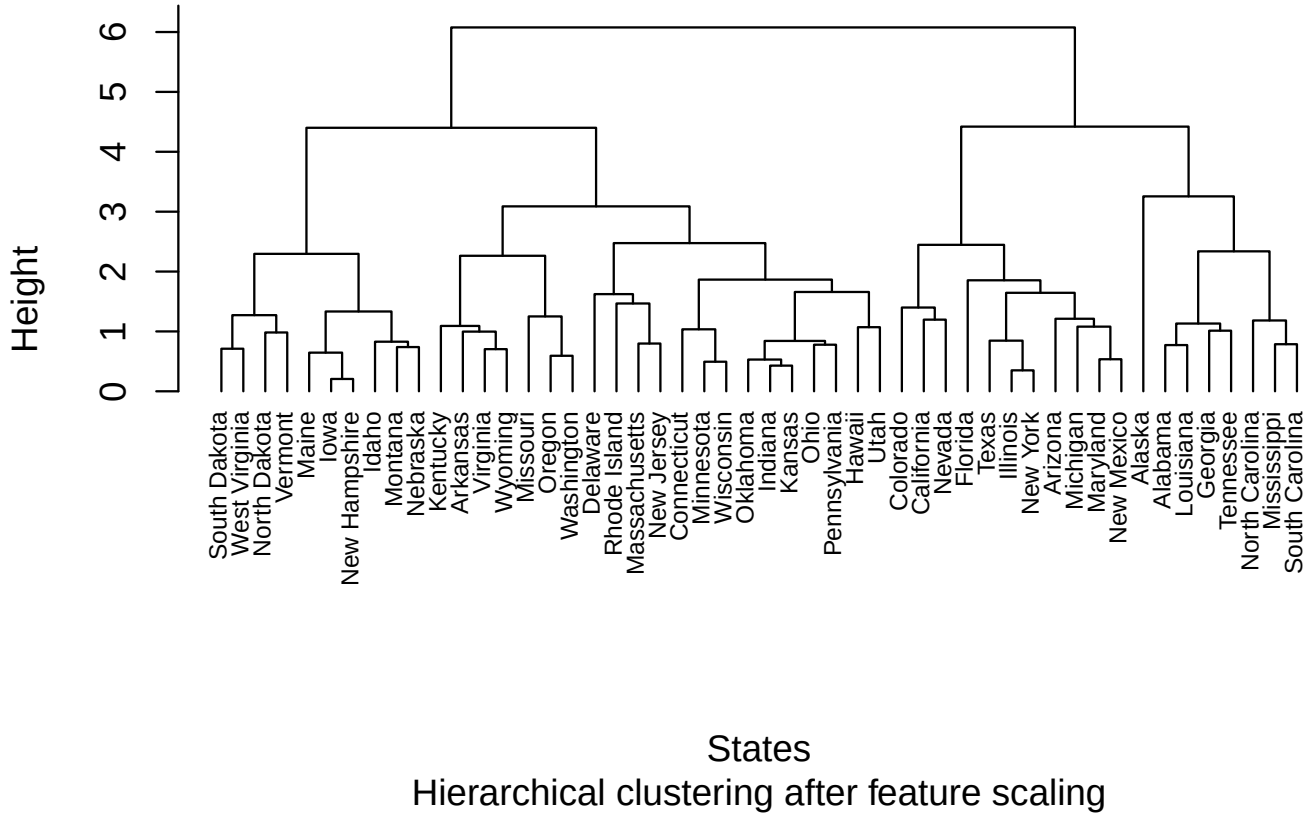
The three clusters obtained after trimming the dendrogram are shown in Table 3 above. Cluster 1 included sixteen states; Cluster 2 contained fourteen states; and Cluster 3 comprised twenty states. Table 4 displays a sample of the dataset's states and the distinct categories to which each state belongs.

1c. Hierachical Clustering (After feature scaling)

We performed feature scaling on the variables with a standard deviation of one and then performed hierarchical clustering using complete linkage and Euclidean distance to cluster the states.

Table 5: Scaled dataset

State	Murder	Assault	UrbanPop	Rape
Alabama	1.2425641	0.7828393	-0.5209066	-0.0034165
Alaska	0.5078625	1.1068225	-1.2117642	2.4842029
Arizona	0.0716334	1.4788032	0.9989801	1.0428784
Arkansas	0.2323494	0.2308680	-1.0735927	-0.1849166
California	0.2782682	1.2628144	1.7589234	2.0678203
Colorado	0.0257146	0.3988593	0.8608085	1.8649672
Connecticut	-1.0304190	-0.7290821	0.7917228	-1.0817408
Delaware	-0.4334739	0.8068381	0.4462940	-0.5799463
Florida	1.7476714	1.9707777	0.9989801	1.1389667
Georgia	2.2068599	0.4828549	-0.3827351	0.4877015



The cluster dendrogram above shows the clustered states with various branches.

1d. Dendrogram Summary

Based on the two dendrograms, the one with scaled variables tends to have more distinct branches and clusters than the one whose dataset was not scaled. In my opinion, the variables should be scaled. If the scale of the variables is not the same, then the model might become biased towards the variables with a higher magnitude.

2. INTRODUCTION (Market Campaign Data)

We now have a glimpse of the market campaign dataset, which will allow us to gain insights from it.

Table 6: Data summary

Name	Piped data
Number of rows	2209
Number of columns	18
Column type frequency:	
character	3
numeric	15
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Education	0	1	3	10	0	5	0
Marital_Status	0	1	5	8	0	5	0
Dt_Customer	0	1	10	10	0	662	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
ID	0	1	5591.62	3245.89	0	2826	5462	8427	11191
Year_Birth	0	1	1968.81	11.98	1893	1959	1970	1977	1996
Income	0	1	52243.98	25198.48	1730	35246	51390	68627	666666
Kidhome	0	1	0.44	0.54	0	0	0	1	2
Teenhome	0	1	0.51	0.54	0	0	0	1	2
Recency	0	1	49.08	28.95	0	24	49	74	99
MntWines	0	1	305.19	337.69	0	24	174	505	1493
MntFruits	0	1	26.35	39.80	0	2	8	33	199
MntMeatProducts	0	1	167.16	224.44	0	16	68	233	1725
MntFishProducts	0	1	37.56	54.58	0	3	12	50	259
MntSweetProducts	0	1	27.07	41.11	0	1	8	33	262
MntGoldProds	0	1	43.85	51.65	0	9	24	56	321
NumWebPurchases	0	1	4.08	2.74	0	2	4	6	27
NumStorePurchases	0	1	5.80	3.25	0	3	5	8	13
Complain	0	1	0.01	0.10	0	0	0	0	1

The dataset contains 2209 samples and 18 variables. The variables are subdivided into three and fifteen character and numeric variables, respectively. The above tables clearly show a summary of the dataset.

2a. Data Preprocessing

Feature extraction & Education Encoding: We extracted two new features from the dataset.

Table 9: Extracted Features: Customer_Age & MembershipDays

ID	Year_Birth	Customer_Age	MembershipDays	Education
11187	1978	43	2964	1
8207	1957	64	2662	2
9723	1960	61	3094	3
2666	1972	49	2702	4
5721	1956	65	2608	5
10652	1957	64	2729	4
1646	1972	49	2664	1
4418	1983	38	3048	4
2656	1971	50	3052	3
7990	1947	74	3255	3
8722	1957	64	3159	2

The above table shows the two new variables, Customer_Age and MembershipDays, and the encoded education variable. We encoded Highschool, Associate, Bachelor, Master and PhD as 1,2,3,4, and 5, respectively.

Encode Marital_Status: The table below shows a sample of the encoded Marital_Status using dummy variables.

Table 10: Encoded Marital_Status

Marital_Status_Married	Marital_Status_Single	Marital_Status_Together	Marital_Status_Widow
0	1	0	0
0	1	0	0
0	0	1	0
0	0	1	0
1	0	0	0
0	0	1	0
0	0	0	0
1	0	0	0
0	0	1	0
0	0	1	0

Exlude Variables and Standardize: We'll now eliminate the variables that aren't needed and standardize the columns.

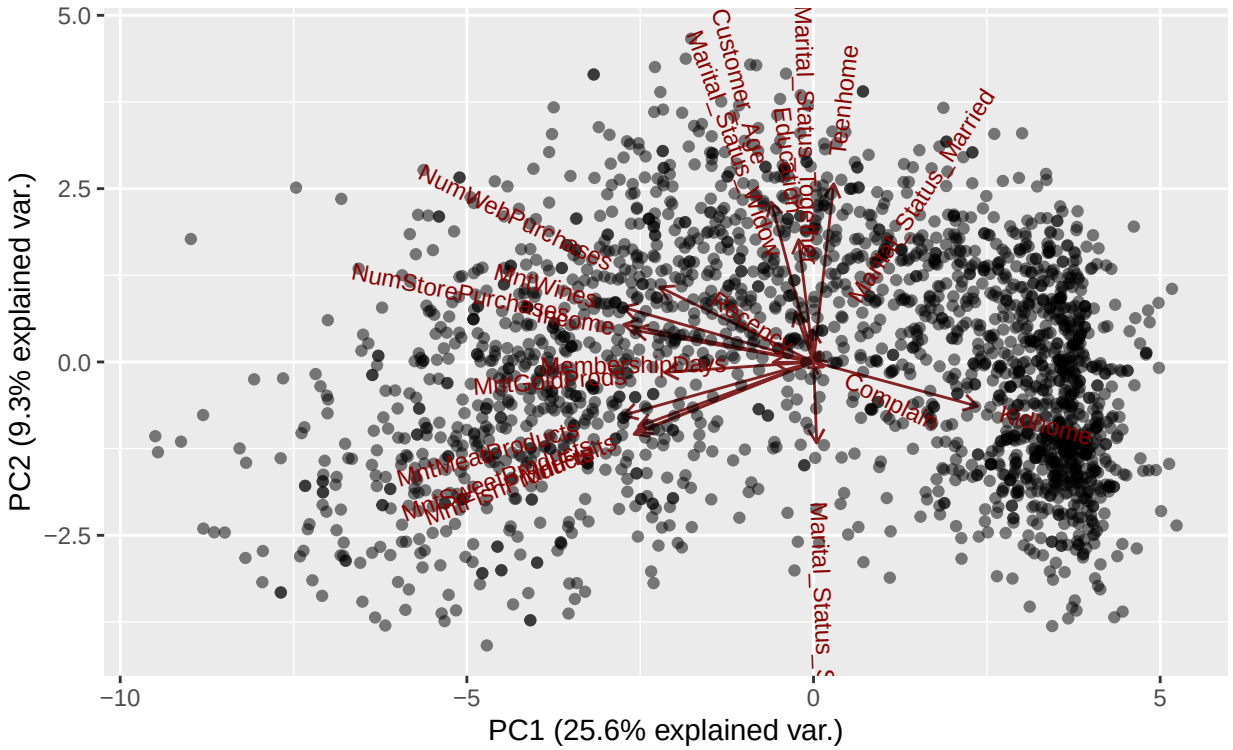
Table 11: Standardized Variables

Income	Kidhome	Teenhome	Recency	MntWines	MntFruits	MntMeatProducts
0.2339039	-0.8227362	-0.9281454	0.3082732	0.9766566	1.5488659	1.6879549
-0.2341403	1.0393789	0.9090170	-0.3826166	-0.8711997	-0.6370558	-0.7180699
0.7686585	-0.8227362	-0.9281454	-0.7971505	0.3577432	0.5689700	-0.1789421
-1.0158542	1.0393789	-0.9281454	-0.7971505	-0.8711997	-0.5616792	-0.6556915
0.2400551	1.0393789	-0.9281454	1.5518749	-0.3914678	0.4182168	-0.2190425
0.4075255	-0.8227362	0.9090170	-1.1425954	0.6361062	0.3930912	-0.3081546
0.1345725	-0.8227362	0.9090170	-0.5207945	-0.2078667	0.9709786	-0.0140849
-0.7456791	1.0393789	-0.9281454	-0.5898835	-0.6787147	-0.4109259	-0.4952898
-0.8688215	1.0393789	-0.9281454	-1.0389619	-0.8623158	-0.6621813	-0.6378691
-1.8491586	1.0393789	0.9090170	0.6537181	-0.8208575	-0.6621813	-0.7180699

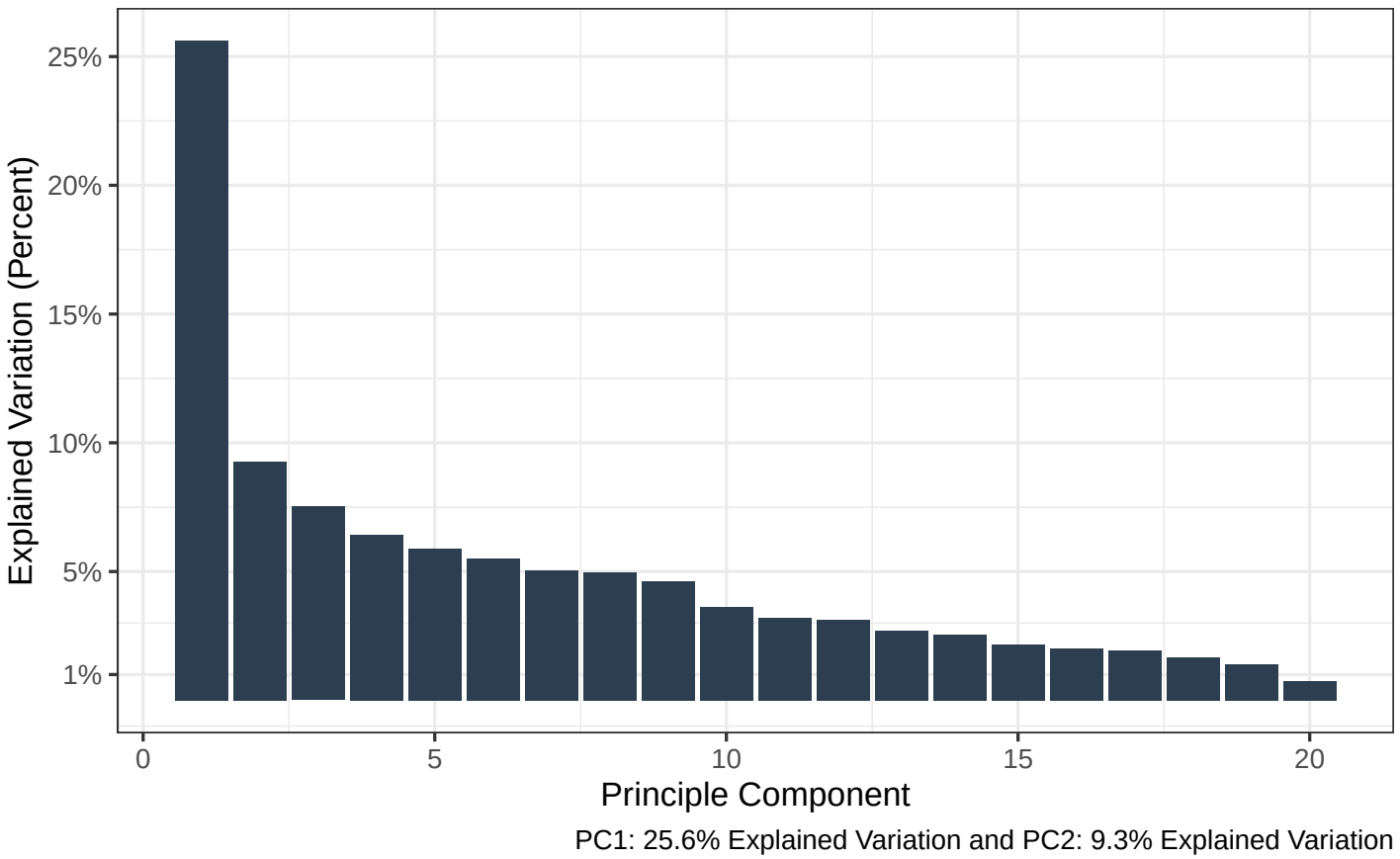
The above table shows a sample of the standardized columns after the ID, Year Birth, Dt Customer, and Marital Status variables have been excluded from the analysis.

2b. Principal Component Analysis

In this study, we did Principal Component Analysis on a standardized dataset and used PC1 and PC2 to show the first 500 observations of the dataset.



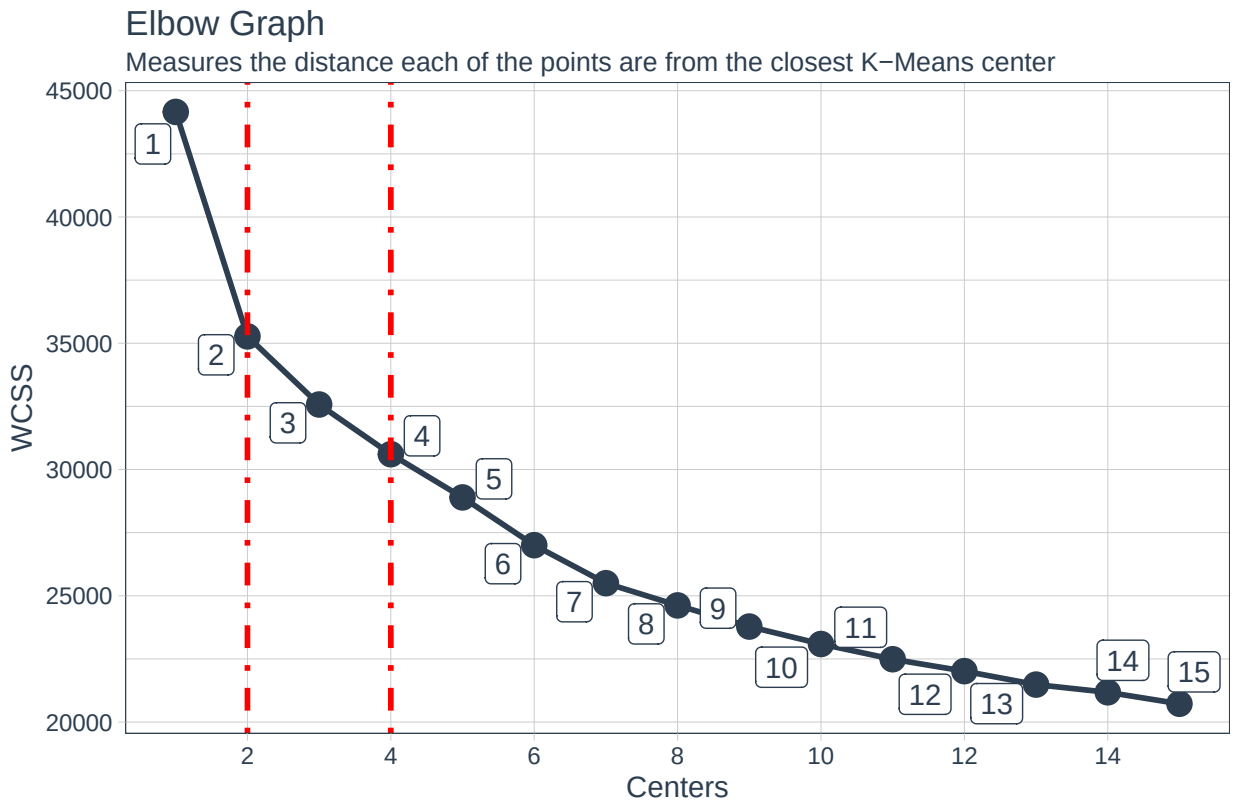
The plot above depicts how much variance each primary component captures from the data set.



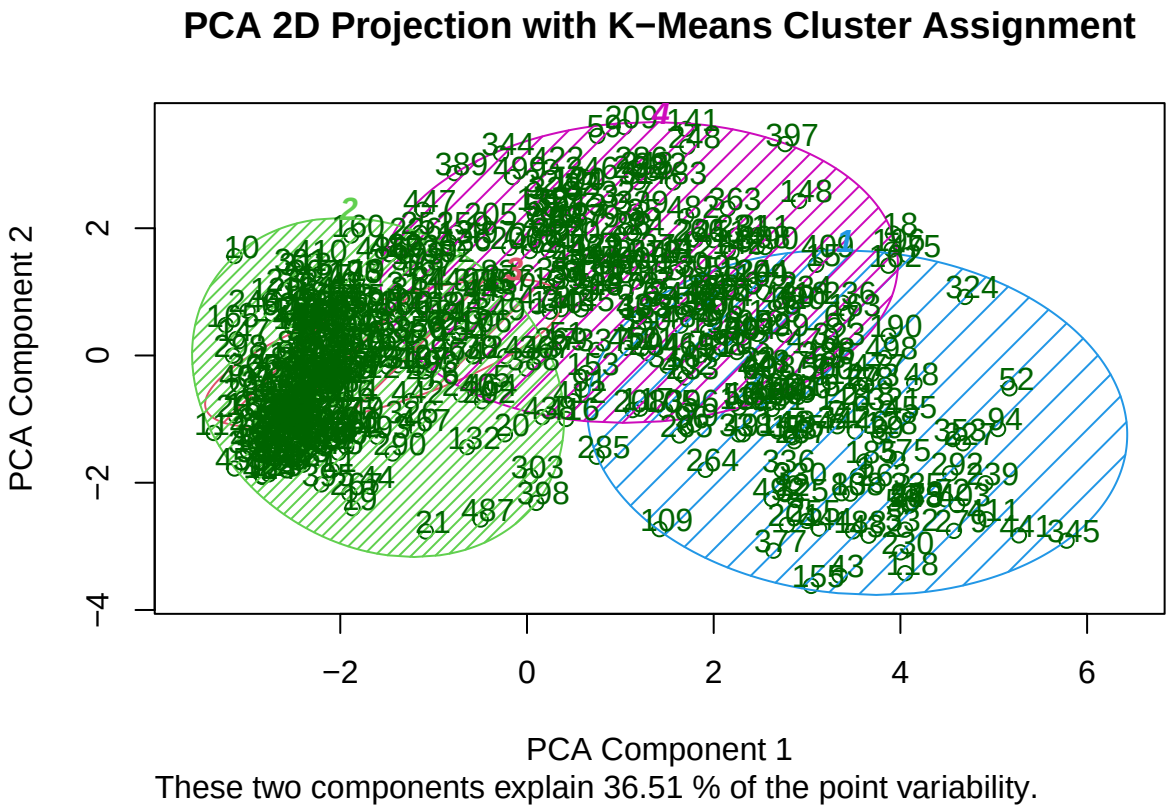
According to the graph above, each chosen component contributes a different proportion of the total variation. PC1 had the highest explainability of 25.6%, followed by PC2, which had an explainability of 9.3%.

2c. Clustering

K-Means Clustering: The assessment of the appropriate number of clusters into which the data may be divided is a critical stage in any unsupervised technique. To calculate this ideal value of k, one of the most common approaches is to use the Elbow Method. In our case we computed the k-means clustering algorithm for a variety of distinct values of the parameter k. We selected fifteen cluster points and computed the within-cluster sum of squares (WCSS).

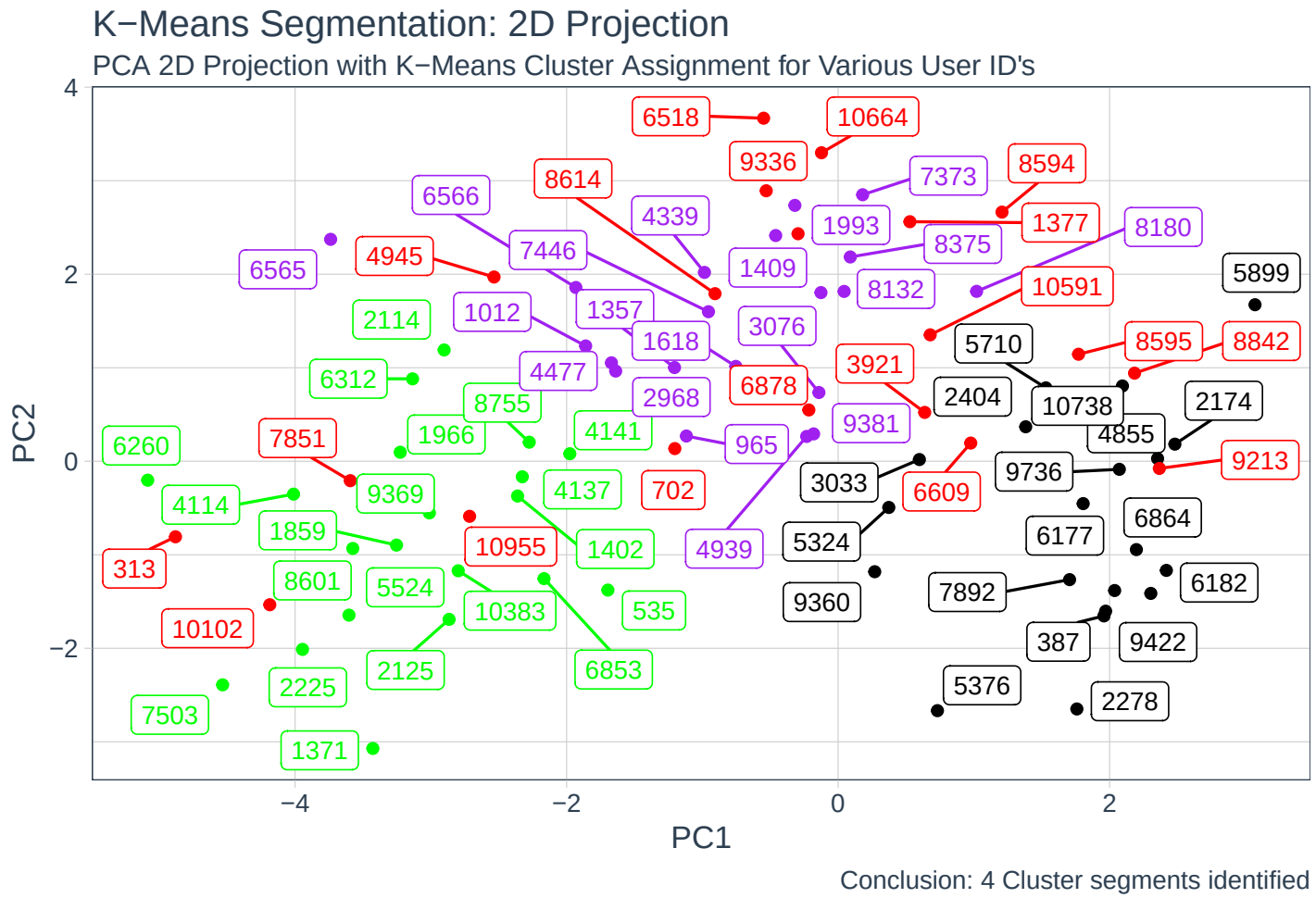


The elbow graph above shows the WCSS as a result of the centres. We then checked for the bends ("elbow") and discovered two bends, 2 and 4. Cluster two is not rational; therefore, we proceeded to choose four clusters.



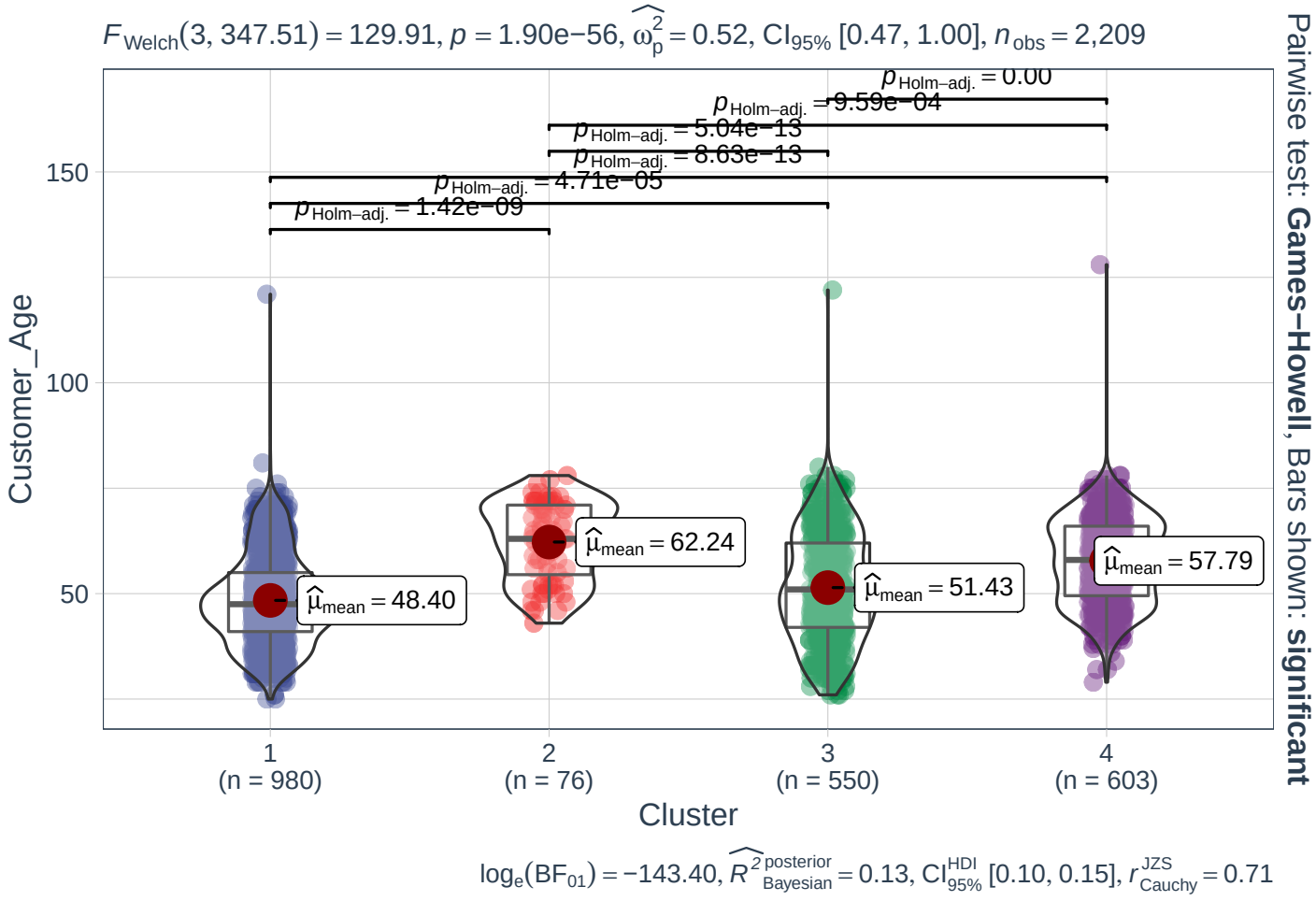
These two components explain 36.51 % of the point variability.

The above visualization shows the first 500 observations using PC1 and PC2.

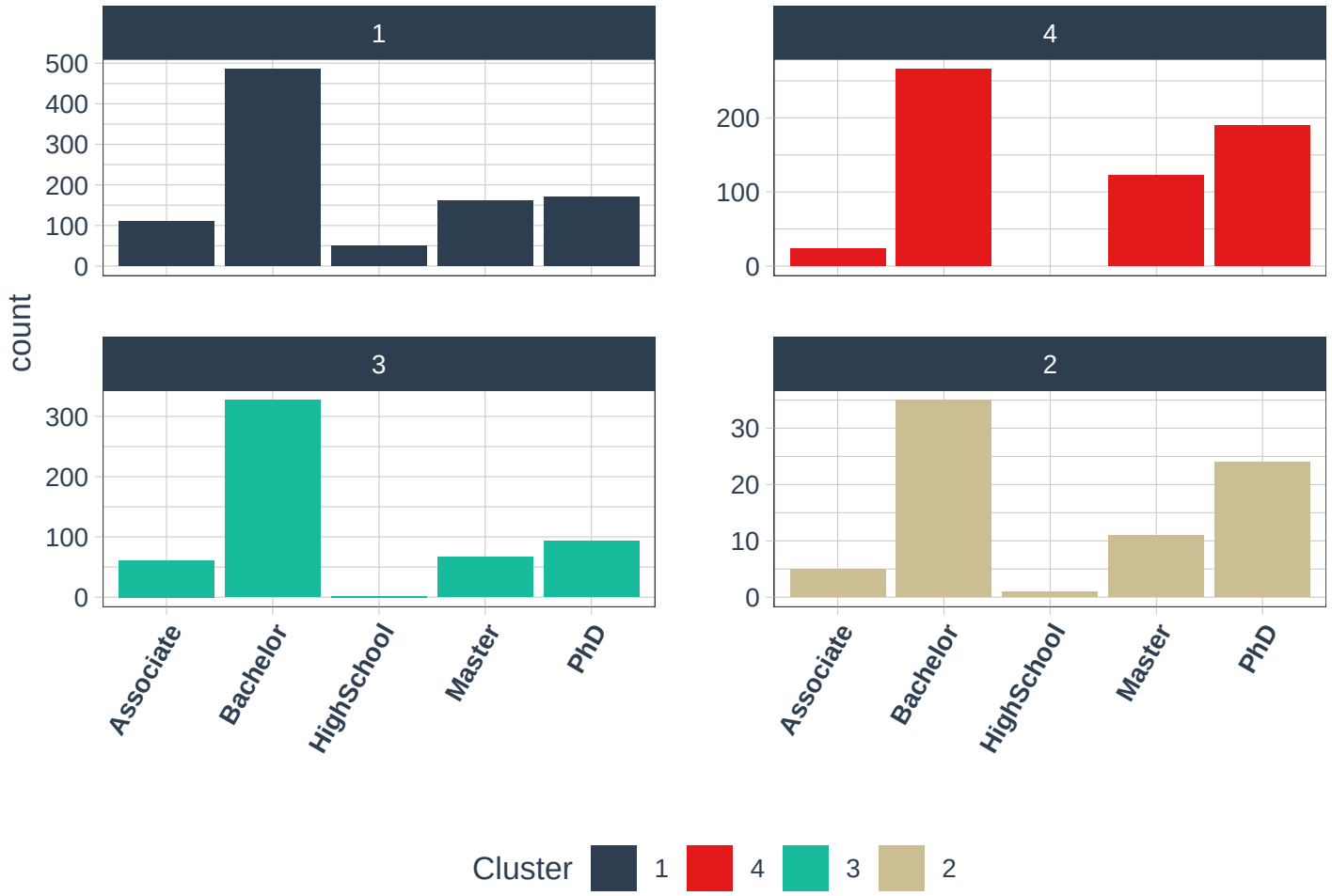


When it comes to viewing, understanding, and analyzing customer clusters, cluster mapping is a cutting-edge method. This information may assist the marketing department in identifying opportunities and making strategic choices based on the insights gained as a result of the process. The 2D projection plot shows the customer ID's mapped to their designated clusters.

Cluster Summary: We now discuss the various clusters with respect to the Customer-Age and Education.



The boxplot above encompassed in a violin plot shows the various mean age distributions of the customers. Cluster one had the highest number of customers followed by cluster four, three and two respectively. The results of the Games–Howell post hoc tests indicate statistically significant differences between all the groups. This is evidenced by the p-values being less than the 0.05 level of significance.



The above histograms clearly show the distribution of the various education levels in various clusters. The Bachelor's degree holders were the majority, so they were evenly distributed in all the categories, with category one having the highest number of customers. Masters and PhD degree holders were concentrated in groups three and four. The Associates and high school holders were less prevalent in category two and four.

REFERENCES

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Kassambara, A. (2017). Practical guide to cluster analysis in R: Unsupervised machine learning (Vol. 1). Sthda.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. Data mining and knowledge discovery, 2(3), 283-304.