**Association Rule Mining & Clustering:**

**Insights from Retail Transactions & Bank Customer Data**

Sriram Srinivasan

The Jindal School of Management, The University of Texas at Dallas

BUAN 6383/ MIS 6386: Modeling for Business Analytics

Professor Syam Menon, Ph.D.

September 15, 2024

Table of Contents

[Part I: Analyzing Transactions 3](#_Toc177328595)

[2. How many item sets are frequent (using a minimum support threshold of 1%)? 3](#_Toc177328596)

[3. How many association rules are generated (with a minimum confidence of 10%)? 3](#_Toc177328597)

[4. Which rules have the highest lift? Using the results from the previous questions, show exactly how this lift value was calculated for one of the rules with highest lift. (Insert tables/figures as needed) 3](#_Toc177328598)

[5. For the same rule show how leverage and conviction were obtained. 3](#_Toc177328599)

[6. Interpret and discuss the 5 rules with: a) the highest confidence b) the highest lift c) the highest leverage d) the highest conviction. If there are more than five meeting the required criterion, pick any five. Are any of these surprising? Comment on the extent of their redundancy and utility. 4](#_Toc177328600)

[7. Do any of these metrics seem preferable to the others for this dataset? Discuss why or why not. 5](#_Toc177328601)

[8. If you were in charge of these departments (Stationery and Health and Beauty Aids), how would you use the results of this analysis to come up with a strategic plan? 6](#_Toc177328602)

[Part II: Clustering Customers 6](#_Toc177328603)

[2. Apply hierarchical clustering (with Euclidian distance as the measure of distance) using centroid linkage, single linkage, complete linkage, average linkage, and Ward linkage. 6](#_Toc177328604)

[Centroid Linkage 7](#_Toc177328605)

[Single Linkage 10](#_Toc177328606)

[Complete Linkage 12](#_Toc177328607)

[Average Linkage 14](#_Toc177328608)

[Ward Linkage 16](#_Toc177328609)

[Linkage Recommendation 18](#_Toc177328610)

[3. Apply k-means clustering to the dataset. 19](#_Toc177328611)

[Recommendation of the value of ‘k’ 21](#_Toc177328612)

[Comparing hierarchical clustering and K-means clustering 21](#_Toc177328613)

[4. If you were the manager of this bank, how would you use the results of this analysis to come up with a strategic plan? Explain your reasoning. 21](#_Toc177328614)

# Part I: Analyzing Transactions

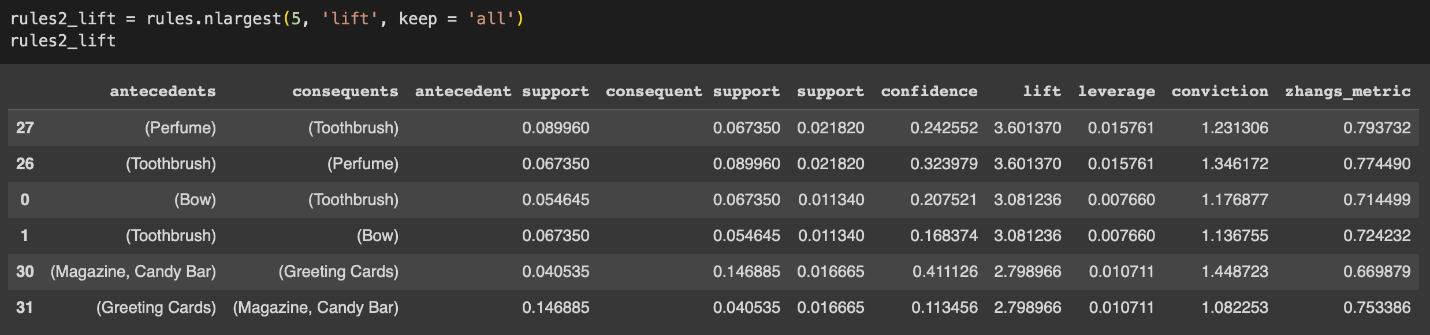
## 2. How many item sets are frequent (using a minimum support threshold of 1%)?

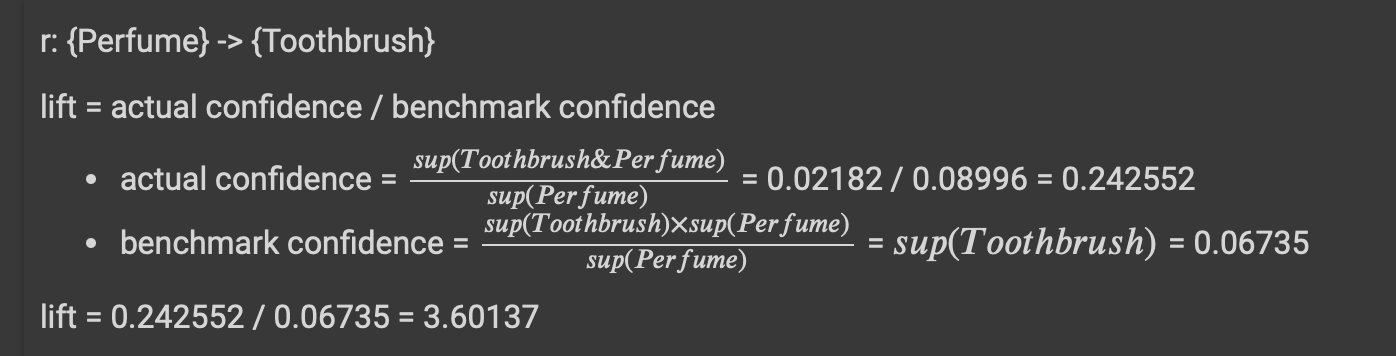
Number of frequent item sets: 40

## 3. How many association rules are generated (with a minimum confidence of 10%)?

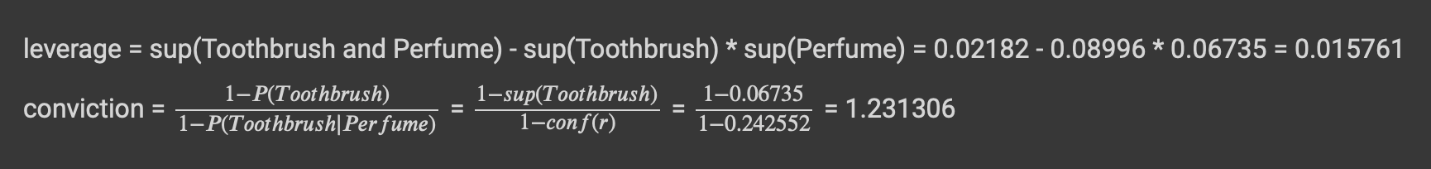
Association rules: 50

## 4. Which rules have the highest lift? Using the results from the previous questions, show exactly how this lift value was calculated for one of the rules with highest lift. (Insert tables/figures as needed)





## 5. For the same rule show how leverage and conviction were obtained.



## 6. Interpret and discuss the 5 rules with: a) the highest confidence b) the highest lift c) the highest leverage d) the highest conviction. If there are more than five meeting the required criterion, pick any five. Are any of these surprising? Comment on the extent of their redundancy and utility.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Criterion** | **Top 5 Rules** | **Interpretation** | **Redundancy** | **Utility** | **Surprising Associations** |
| **Confidence** | 1. (Toothpaste, Pencils) → (Candy Bar)  2. (Greeting Cards, Magazine) → (Candy Bar)  3. (Magazine, Toothpaste) → (Candy Bar)  4. (Magazine, Candy Bar) → (Greeting Cards)  5. (Greeting Cards, Toothpaste) → (Candy Bar) | These rules have high confidence, indicating that when the antecedents are purchased, the consequent is bought with high probability (up to 46.38%). | High redundancy with **Candy Bar** as the consequent in 4/5 rules. | Useful for promotions involving **Candy Bar**, though redundant due to repetitive patterns. | None, as **Candy Bar**is expected to be a frequent co-purchase with smaller items. |
| **Lift** | 1. (Perfume) → (Toothbrush)  2. (Toothbrush) → (Perfume)  3. (Bow) → (Toothbrush)  4. (Toothbrush) → (Bow)  5. (Greeting Cards) → (Magazine, Candy Bar) | Lift indicates the strength of the association beyond random chance. The top rules (lift > 3) show strong co-occurrence of **Perfume** and **Toothbrush**. | Low redundancy for **Toothbrush** and **Perfume**, unlike common items such as **Candy Bar**. | High utility for identifying unexpected relationships between **Toothbrush** and **Perfume**. | The **Toothbrush-Perfume** and **Bow-Toothbrush**combinations are surprising due to their unlikely co-occurrence. |
| **Leverage** | 1. (Greeting Cards) → (Candy Bar)  2. (Candy Bar) → (Greeting Cards)  3. (Toothbrush) → (Perfume)  4. (Perfume) → (Toothbrush)  5. (Candy Bar) → (Toothpaste) | Leverage indicates how much more often the items appear together than by random chance. High leverage is seen for **Candy Bar** and **Toothbrush-Perfume** rules. | **Candy Bar** dominates the first two rules, showing frequent co-occurrence with greeting cards and toothpaste. | High utility for targeted offers, especially with **Toothbrush-Perfume**, which is unexpected. | **Toothbrush-Perfume** stands out as unexpected, while **Candy Bar** associations are predictable and less surprising. |
| **Conviction** | 1. (Toothpaste, Pencils) → (Candy Bar)  2. (Greeting Cards, Magazine) → (Candy Bar)  3. (Magazine, Toothpaste) → (Candy Bar)  4. (Magazine, Candy Bar) → (Greeting Cards)  5. (Greeting Cards, Toothpaste) → (Candy Bar) | Conviction reflects rule reliability when the consequent is true. Most of these rules involve **Candy Bar** as the consequent, showing strong associations. | High redundancy, since **Candy Bar**appears frequently in the consequents. | Moderate utility for predicting frequent co-purchases, but not highly insightful. | No surprising associations, as **Candy Bar** is expected to co-occur with other small items. |

* **Redundancy:** Many rules involve **Candy Bar**, making them **redundant**, especially in terms of confidence and conviction criteria.
* **Utility:** The **Perfume-Toothbrush** association provides valuable insight for targeted promotions due to its unexpected **high lift and leverage**.
* **Surprising Associations:** The **Toothbrush-Perfume** and **Bow-Toothbrush** associations are surprising and merit further investigation for potential marketing opportunities.

## 7. Do any of these metrics seem preferable to the others for this dataset? Discuss why or why not.

Among support, confidence, and lift, lift is the most useful metric for this dataset.

* It normalizes for the frequency of individual items and highlights meaningful associations between items.
* While support shows how frequently items are bought together and confidence measures predictability, lift uncovers relationships that are not just due to popularity. Ex. toothpaste and toothbrush might frequently appear together, but a high lift value confirms a strong relationship beyond just commonality.

## 8. If you were in charge of these departments (Stationery and Health and Beauty Aids), how would you use the results of this analysis to come up with a strategic plan?

* Cross-Promotions: Bundle high-lift itemsets (e.g., toothbrush + toothpaste) and offer discounts on combined purchases.
* Optimize Product Placement: Place frequently bought-together items near each other (e.g., shampoo + deodorant) to boost sales.
* Targeted Promotions: Use high-lift but low-support items (e.g., pain reliever+ magazines) in targeted marketing campaigns.
* Impulse Buys: Place high-lift, low-frequency items near checkout counters (eg., candy bars + magazines) to encourage impulse purchases.

# Part II: Clustering Customers

## 2. Apply hierarchical clustering (with Euclidian distance as the measure of distance) using centroid linkage, single linkage, complete linkage, average linkage, and Ward linkage.

Table : Comparison of Hierarchical Clustering Results Across Different Linkage Methods

|  |  |  |
| --- | --- | --- |
| Linkage Type | # of Clusters Visible | Recommended # of Clusters |
| Centroid | 3 | 2 |
| Single | 2 | 1 |
| Complete | 2 | 2 |
| Average | 2 | 1 |
| Ward | 2 | 2 |

Table : Maximum Distance Between Merged Clusters & Default Threshold for Cluster Formation Across Different Linkage Methods

|  |  |  |
| --- | --- | --- |
| Linkage Type | Maximum distance between merged clusters\* | Default threshold for cluster formation\* |
| Centroid | 0.001268 | 0.000888 |
| Single | 0.000656 | 0.000459 |
| Complete | 0.003250 | 0.002275 |
| Average | 0.002302 | 0.001612 |
| Ward | 0.015312 | 0.010718 |

*\* Rounded to the 6th decimal*

### Centroid Linkage

A screen shot of a graph

Description automatically generated

Figure : Hierarchical Clustering Dendrogram using Centroid Linkage

A graph of a diagram

Description automatically generated with medium confidence

A diagram of a cluster

Description automatically generated

**Data characteristics:**

* From the dendrogram of the hierarchical clusters using centroid linkage: 2 clear clusters (orange and green) and 1 very small red cluster.
  + Record count (largest to smallest): Orange: 519, Green: 79, Red: 2

**Distinguishing characteristics:**

All clusters have a relatively even split on sex

* **Cluster 1:** High income, middle-aged (Avg. 42), married with 1 or more children with highest avg of checking/savings ownership (over 70%)
  + Highest income among the clusters (Avg. $29,388)/ over ½ have cars
  + Little bit less than ½ have a PEP and live in the inner city
  + Less than ½ have a mortgage and live in town with fewer living in rural and suburban
* **Cluster 2:** Middle -aged (Avg. 43), married with a child with middle income ($15,737)
  + More than ½ have Checking/ Savings and live in the inner city
  + Less than ½ have a mortgage, PEP and a car
  + Few live in town, rural and suburban (highest # between the 3 clusters)
* **Cluster 3:** 
  + Youngest of the group (Avg. 35.5) with the lowest income (Avg. $9,149)
  + Not married, or live in rural or suburban regions
  + All have at a car, savings, checking account
  + ½ have children, mortgage, PEP, live in the inner city or town

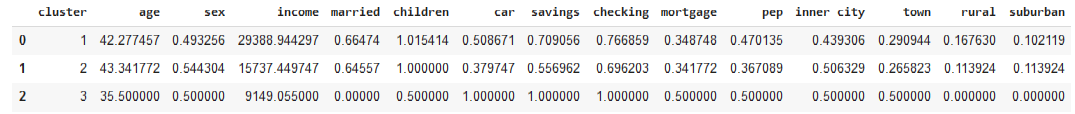


Figure : Average Values of All Columns Grouped by Cluster ID

**Recommendations:**

* Recommend 2 clusters: orange and a combining green + red.
* The distribution of the records among 3 clusters is extremely disproportionate, orange cluster contains over 6 times the number of records compared to the other two combined.
* By grouping red and green together combined makes it 81 (14%), which is a better balance to the larger Orange with 519 (86%) and simplifies the model to make it easier to interpret

### Single Linkage

A screen shot of a barcode

Description automatically generated

Figure : Hierarchical Clustering Dendrogram using Single Linkage

A graph of a number of people

Description automatically generated with medium confidence

A graph of a diagram

Description automatically generated with medium confidence**Data characteristics:**

* There is only 1 clear cluster with almost all records. The blue cluster only includes 2 records.
  + Record count- Orange: 598, Blue: 2
* Given the extreme disproportion, the clusters formed may not be useful. I would not recommend performing clustering because having 2 clusters in this case is almost identical to having just 1 cluster (all records merged into one).
* Potential underfitting, because the model has an imbalance in cluster sizes which doesn't represent any distinct patterns or characteristics.

**Distinguishing characteristics:**

* **Cluster 1:** Middle-aged (Avg. 42.4), with at least 1 child and the highest average income of $27,585
  + More than ½ have checking/ savings, almost ½ have a car and live in the inner city
  + Less than ½ have a PEP and mortgage with a spread living in town, rural and suburban
* **Cluster 2:** Older (avg age 50), single women, don’t have a mortgage and all live in the inner city, have a car, checking/saving and PEP
  + Middle avg. income ($13,283)
* **Cluster 3:** Youngest age avg (21) with the lowest avg. income ($5,014). All live in town, have a car, checking/savings and a mortgage but none have a PEP

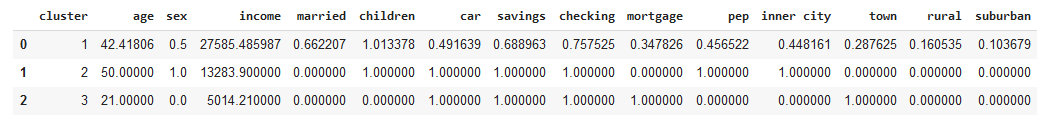


Figure : Average Values of All Columns Grouped by Cluster ID

### Complete Linkage

A white screen with orange and black lines

Description automatically generated

Figure : Hierarchical Clustering Dendrogram using Complete Linkage

A diagram of a cluster of links

Description automatically generated with medium confidence

A diagram of a linkage distance

Description automatically generated

Data characteristics:

* There are 2 clear clusters (orange and green) although the orange cluster still contains significantly more records than the green one.
  + Record count Green: 488, Orange: 112
* Don’t recommend any changes, because the distance between these two main branches is significantly larger than within each cluster, indicating a well-separated grouping.

**Distinguishing characteristics:**

Both clusters have an average age of 42.3 years, balanced sex distribution, over 65% are married have children, have checking/savings account and live in the inner city.

* **Cluster 1:** Higher avg. income between the 2 clusters and likelihood of having financial assets
* **Cluster 2:** Lower income of the two and more likely to live in town

A screenshot of a calculator

Description automatically generated

Figure : Average Values of All Columns Grouped by Cluster ID

### Average Linkage

A screen shot of a computer

Description automatically generated

Figure : Hierarchical Clustering Dendrogram using Average Linkage

A diagram of a cluster

Description automatically generated

A diagram of a graph

Description automatically generated

**Data Characteristics**

* There is only 1 clear cluster with almost all records. The orange cluster contains 2 records.
  + Record count- Green: 598, Orange: 2
* Given the extreme disproportion, the clusters formed may not be useful. I would not recommend performing clustering because having 2 clusters in this case is almost identical to having just 1 cluster (all records merged into one).
* Potential underfitting, because the model has an imbalance in cluster sizes which doesn't represent any distinct patterns or characteristics.

**Distinguishing characteristics:**

Both have balanced distribution of sex at 50%

* **Cluster 1:** Younger than the 2 (avg. 35), all are single, have a car, checking/savings, but the lowest income average ($9,149)
  + 1/2 have children, a mortgage and live in either the inner city or town
* **Cluster 2:** Middle-aged (avg. 42.41) and highest income ($27,585) with mostly living in the inner-city and have a child
  + Majority have a checking/savings, almost ½ have a car and PEP and less than ½ have a mortage

A screenshot of a cell phone

Description automatically generated

Figure : Average Values of All Columns Grouped by Cluster ID

### Ward Linkage

A white screen with colorful lines

Description automatically generated with medium confidence

Figure : Hierarchical Clustering Dendrogram using Ward Linkage

A diagram of a cluster

Description automatically generated

A diagram of a cluster

Description automatically generated

**Data characteristics:**

* There are 2 clear clusters with a relatively balanced split:
  + Record count: Green: 321 (53%), Orange: 279 (47%)
* Don’t recommend any changes, because there is a significant jump in distance from below the threshold to above it. This suggests that two distinct clusters are well-separated, like using Complete linkage.

**Distinguishing characteristics:**

* **Cluster 1:** Represents middle-aged individuals with higher income, moderate car ownership, and extensive use of financial products. They show a diverse geographic spread, including a mix of rural and suburban areas.
* **Cluster 2:** Features slightly younger individuals with lower income, balanced car ownership, and moderate use of financial products. They predominantly reside in towns and have lower levels of rural and suburban residency.

A screenshot of a calculator

Description automatically generated

Figure : Average Values of All Columns Grouped by Cluster ID

### Linkage Recommendation

|  |  |
| --- | --- |
| Linkage | Characteristics |
| Centroid Linkage | Forms clusters based on the centroids of groups. Can be unstable if cluster sizes vary, leading to inconsistent shapes. |
| Single Linkage | Produces long, chain-like clusters with uneven distribution. Not suitable for distinct, well-separated clusters. |
| Complete Linkage | Creates tight, compact clusters but is sensitive to outliers, which merge at higher distances. |
| Average Linkage | Balances between compactness and separation, with clusters moderately spaced. Works well with varying cluster sizes. |
| Ward Linkage | Generates compact, well-separated clusters. Best for identifying distinct customer segments with minimal within-cluster variance. |

Ward linkage is recommended over complete linkage because it produces the most clear and balanced clusters, minimization of error, and reduced sensitivity to outliers, which is ideal for customer segmentation. Based on the analysis of clustering results, the complete linkage and Ward linkage methods produce two well-defined clusters at the 70% threshold. While other linkage methods (single and average) yield two or three clusters, often resulting in one disproportionately large cluster, reducing the interpretability and effectiveness of the clustering process.

* Ward linkage merges records based on the smallest increase in the total error sum of squares, which tends to produce more balanced clusters. This is evident when comparing the distances below the threshold in the dendrogram, where Ward linkage groups clusters at a greater distance than complete linkage. This leads to a more intuitive structure, especially when applying the results to further models like regression.
* Both regression models and Ward linkage focus on minimizing errors. This means that the clusters formed by Ward linkage are likely to align better with regression-based analyses, where the objective is often to reduce the sum of squared errors. In contrast, complete linkage, which groups clusters based on the maximum Euclidean distance, can be more sensitive to outliers. An outlier will directly affect the maximum distance between clusters, potentially skewing the cluster formation and leading to less reliable results.
* Ward linkage is less affected by outliers because it looks at overall variance and minimizes the increase in the error sum of squares. This approach ensures that a single outlier won’t disproportionately influence the clustering process, unlike complete linkage, which is directly impacted by the maximum distance between cluster points.

## 3. Apply k-means clustering to the dataset.

*Try different values of k (4, 5, 6, 7, and 8 at least); make sure you include the number of clusters you decided to use with hierarchical clustering.*

**Are clear clusters visible for any value of k? As before, how many clusters would you recommend, and why? What are some distinguishing characteristics of each cluster?**

Table : Number of visible vs. recommended clusters for different values of 'k'

|  |  |  |
| --- | --- | --- |
| Value of ‘k’ | # of Clusters Visible | # of Recommended Clusters |
| 2 | **2** | **2** |
| 4 | **4** | **4** |
| 5 | **5** | **5** |
| 6 | **6** | **5** |
| 7 | **7** | **6** |
| 8 | **8** | **7** |
| 10 | **10** | **9** |

Table : Distinguishing characteristics of different values of ‘k ‘

|  |  |
| --- | --- |
| Value of ‘k’ | Distinguishing Charactersticis |
| 2 | * Both clusters are clear because they both have a sufficiently large number of records. * The 2 clusters formed by k-means clustering with k = 2 have similar characteristics of the 2 clusters formed by hierarchical clustering with complete or ward linkage. |
| 4 | * All clusters are clear, although Cluster 1 contains roughly half of the records. * Customers in Clusters 0 and 3 tend to be older, married females with more children. They are more likely to own a car and mortgage and familiar with the banking products. * Customers in Clusters 1 and 2 tend to be younger, single males with fewer children. They are less likely to use any banking products and less likely to own a car or have a mortgage. * However, customers in Clusters 1 and 2 earn a slightly higher income compared to customers in Clusters 0 and 3. * Most of the customers in all clusters live in the inner city. |
| 5 | * There are 5 clear clusters. * The mean values of all the variables except income follow the same descending order: Cluster 1 -> 2 -> 4 -> 0 -> 3. Average income is the highest for Clusters 0 and 3. * Higher mean values = older, married, more children, more likely to own a car and/or mortgage, and have a savings, checking account, and personal equity plan. * The only exception is that more customers in Cluster 2 tend to use the personal equity plan than customers in Cluster 1. |
| 6 | * Out of 6 clusters, 5 clusters are clear with a reasonable number of records. Cluster 5 only contains 2 records. * So far, we have observed 2 groups of patterns: * Younger, single, male, higher income, limited use of banking products, do not own a car or mortgage * Older, married, female, lower income more familiar with banking products, own a car or mortgage * Most clusters follow either of these 2 groups, except Cluster 5. Cluster 5 seems to have random patterns (older, female, familiar with banking products, own a car or mortgage but single with no children). Considering that this cluster only contains 2 records, these erratic behaviors are expected. * 6 clusters may be too granular in this case. |
| 7 | * Similar to k-means with 6 clusters, there is 1 cluster with only 2 records. While all other clusters follow either group of patterns, Cluster 6 shows an unexpected behavior. |
| 8 | * Similar to k-means with 6 or 7 clusters, there is 1 cluster with only 2 records. While all other clusters follow either group of patterns, Cluster 6 shows an unexpected behavior. |
| 10 | * Similar to k-means with 6-8 clusters, there is 1 cluster with only 2 records. While all other clusters follow either group of patterns, Cluster 6 shows an unexpected behavior. |

A close up of a text

Description automatically generated

Figure : Cluster sizes for different values of 'k' from 2 to 10

### Recommendation of the value of ‘k’

Recommendation is to use the value of **5** for ‘k’, which produces more meaningful and distinct clusters.

* Using k = 5 results in well-defined clusters, each capturing meaningful segments of the customer base. These clusters provide clear insights into customer behavior and demographics, enabling a more nuanced understanding of customer interactions with banking products.
* Cluster Patterns: The clusters broadly divide into two distinct profiles:
  + Customers with limited usage of traditional banking products.
  + Customers with higher engagement in banking and investment activities.
* At this cluster size, the model successfully identifies two key customer segments that actively use banking products.

### Comparing hierarchical clustering and K-means clustering

**How different are these results from those with hierarchical clustering? Which seems preferable in this case? Explain**

* While both hierarchical and k-means clustering produce clusters with similar characteristics, k-means is preferable in this case because the flexibility in adjusting the k value allows us to capture more inconspicuous patterns.
* Hierarchical clustering is well-suited for capturing hierarchical relationships, outliers, and small, distinct groups within the data. However, it can result in very small clusters that may not always provide useful insights, especially for practical applications where balanced clusters are desired.
* K-means offers more balanced clusters with fewer extreme outliers. It is effective for large datasets and when a consistent number of evenly sized clusters are required, though it may smooth over nuanced or rare data points.

## 4. If you were the manager of this bank, how would you use the results of this analysis to come up with a strategic plan? Explain your reasoning.

If I were the manager of this bank, I would use the **Ward linkage clustering** results at **k = 5** to focus on two key customer groups for targeted promotions: Older, family-oriented, lower-income & Younger, single, higher-income. This approach targets the specific needs of two key customer segments, enhancing their engagement with the bank's products. Traditional products would appeal to risk-averse older customers, while younger, higher-income customers can be targeted with investment opportunities to grow their wealth.

**1. Older, Family-Oriented, Lower Income**

* **Characteristics**: Married, have children, and lower income. They rely heavily on traditional products like savings and checking accounts and are more risk averse.
* **Strategy**:
  + **Promote traditional products**: Focus on savings accounts, checking accounts, and mortgage-related services.
  + **Offer financial planning**: Introduce budgeting tools and long-term savings plans to help them secure their financial future.
  + **Insurance and protection products**: Market life and home insurance to provide financial security for their families.

**2. Younger, Single, Higher Income**

* **Characteristics**: Younger, single, and with higher earning potential. They are more open to investments, including **Personal Equity Plans (PEPs)**.
* **Strategy**:
  + **Subgroup 1: Traditional Product Education**: Educate those unfamiliar with savings/checking accounts. This group could transition into higher investment opportunities over time.
  + **Subgroup 2: Investment-Focused**: Target this subgroup with higher-risk, high-return investment products like stocks, mutual funds, and retirement plans. Leverage digital banking tools and robo-advisors to engage them.

**Additional Considerations:**

* **Data diversification**: Expand the dataset to include more married individuals, families, and other financial behaviors (like property investments or public transportation users).
* **Develop digital banking tools**: Particularly for the younger, higher-income group, focus on mobile-first platforms for investment and personalized financial management.

A screenshot of a computer screen

Description automatically generated

Figure : Distribution of key customer attributes of the dataset