Liana Cramb

ECO 520 Homework 3:

Table of Contents

[Hierarchical Method Cluster Analysis: 1](#_Toc188737377)

[Pseudo T-Squared Line Plot: 3](#_Toc188737378)

[Dendrogram: 3](#_Toc188737379)

[5 Cluster Analysis 4](#_Toc188737380)

[4 Cluster Analysis 6](#_Toc188737381)

[Non-Hierarchical Cluster (K-Means) Analysis: 8](#_Toc188737382)

[5 Cluster Analysis 8](#_Toc188737383)

[4 Cluster Analysis 12](#_Toc188737384)

[ANOVA Test on Approval Rate: 16](#_Toc188737385)

[Potential Issues with Outliers and Overall Problems with the Data: 18](#_Toc188737386)

[Conclusion: 22](#_Toc188737387)

# Hierarchical Method Cluster Analysis:

| **Cluster History** | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Number of Clusters** | **Clusters Joined** | | **Freq** | **Semipartial R-Square** | **R-Square** | **Approximate Expected R-Square** | **Cubic Clustering Criterion** | **Pseudo F Statistic** | **Pseudo t-Squared** | **Tie** |
| **15** | CL39 | CL23 | 381 | 0.0012 | .986 | .974 | 33.5 | 15E3 | 495 |  |
| **14** | CL29 | CL26 | 465 | 0.0013 | .984 | .972 | 32.7 | 15E3 | 486 |  |
| **13** | CL31 | CL48 | 808 | 0.0013 | .983 | .970 | 32.5 | 15E3 | 1919 |  |
| **12** | CL22 | CL28 | 195 | 0.0014 | .982 | .968 | 32.8 | 15E3 | 187 |  |
| **11** | CL16 | CL40 | 189 | 0.0025 | .979 | .965 | 30.6 | 14E3 | 175 |  |
| **10** | CL15 | CL27 | 974 | 0.0037 | .976 | .961 | 26.7 | 14E3 | 1403 |  |
| **9** | CL12 | CL10 | 1169 | 0.0049 | .971 | .957 | 22.4 | 13E3 | 629 |  |
| **8** | CL20 | CL18 | 112 | 0.0050 | .966 | .952 | 20.3 | 12E3 | 225 |  |
| **7** | CL14 | CL19 | 578 | 0.0063 | .959 | .945 | 18.5 | 12E3 | 878 |  |
| **6** | CL7 | CL54 | 600 | 0.0091 | .950 | .935 | 15.9 | 12E3 | 521 |  |
| **5** | CL9 | CL13 | 1977 | 0.0122 | .938 | .922 | 13.9 | 12E3 | 1522 |  |
| **4** | CL8 | CL11 | 301 | 0.0245 | .914 | .901 | 6.24 | 11E3 | 581 |  |
| **3** | CL6 | CL17 | 787 | 0.0450 | .869 | .854 | 5.04 | 1E4 | 1659 |  |
| **2** | CL5 | CL4 | 2278 | 0.1904 | .678 | .720 | -8.1 | 6460 | 6663 |  |
| **1** | CL3 | CL2 | 3065 | 0.6783 | .000 | .000 | 0.00 | . | 6460 |  |

Based on the cluster analysis, I observed a significant improvement in the R-squared value when the number of clusters increased to 4 or 5. This led me to test both options to determine which provided a better fit for the dataset. Since we are working with three variables, I chose to use the 'Ward's procedure' because this method provided better results based on the R-Square values versus the centroid method. Ward's procedure focuses on minimizing variance within clusters, leading to compact and well-separated groups.

## Pseudo T-Squared Line Plot:

A graph with blue lines

AI-generated content may be incorrect.

As displayed by the Pseudo t-squared line plot, we observe an 'elbow' at 4 clusters. This suggests that 4 clusters might be the better choice, as it represents a natural division in the data.

## Dendrogram:

A black and white drawing of a rectangular object

AI-generated content may be incorrect.

The dendrogram also supports the choice of 4 clusters, as it reveals four distinct groupings. This visualization highlights the natural divisions within the data based on the three variables.

## 5 Cluster Analysis

**Average Income vs. Population**

**A graph of a number of bars

AI-generated content may be incorrect.**

**Average Income vs. Minority**

A chart of different colored circles

AI-generated content may be incorrect.

**Population vs. Minority**

A graph of a cluster of dots

AI-generated content may be incorrect.

After performing the cluster scatter plots with the number of clusters set to 5, the first two graphs—Average Income vs. Population and Average Income vs. Minority—show five distinct cluster groups. In these graphs, the clusters are clearly structured, with the cluster group one ranging from $0 to just over $70,000 in average income, cluster group four centered around $75,000, cluster group two spanning $80,000 to $85,000, cluster group three around $95,000, and cluster group five near $100,000 in average income.

In contrast, the Population vs. Minority scatter plot shows the clusters predominantly clumped together, indicating minimal variation or distinction between population and minority when grouped alone. While this lack of differentiation might pose challenges if used by itself, the inclusion of average income in the other two scatter plots results in clear and meaningful clusters. This reinforces the idea that average income is a key variable driving the clustering and information needed for the analysis.

## 4 Cluster Analysis

**Average Income vs. Population**

**A graph of a number of people

AI-generated content may be incorrect.**

**Average Income vs. Minority**

A graph of different colored circles

AI-generated content may be incorrect.

**Population vs. Minority**

A graph of a cluster of dots

AI-generated content may be incorrect.

After performing the cluster scatter plots with the number of clusters set to 4, the first two graphs—Average Income vs. Population and Average Income vs. Minority— show four distinct cluster groups. In these graphs, the clusters are also clearly structured, with cluster group one ranging from $0 to just over $70,000 in average income, cluster group four centered around $75,000, the cluster group two spanning $80,000 to $85,000, and cluster group three ranging $95,000 to $100,000 in average income. For the Population vs. Minority graph, even after reducing the number of clusters, there was very little improvement in distinguishing distinct groups between these two variables. This supports the argument that population and minority have limited variation when analyzed together.

After reviewing both cluster variations, I determined that 4 clusters were the better choice with the hierarchical clustering. Looking at the Average Income vs. Population and Average Income vs. Minority graphs, I noticed that the cluster with an average income around $95,000 didn’t really stand out from the one starting at $100,000, so it seemed unnecessary. Combining them made more sense, as they were closely aligned. Additionally, both the dendrogram and the R-squared values from the cluster analysis also supported that 4 clusters provided the best fit for the data.

# Non-Hierarchical Cluster (K-Means) Analysis:

## 5 Cluster Analysis

| **Cluster Means** | | | |
| --- | --- | --- | --- |
| **Cluster** | **avg\_income** | **population** | **minority** |
| **1** | 84424.03990 | 4168.31222 | 48.27205 |
| **2** | 68473.27001 | 3654.62687 | 14.64033 |
| **3** | 99096.18056 | 4915.74653 | 34.44514 |
| **4** | 55600.00000 | 3482.59091 | 18.97273 |
| **5** | 91653.84615 | 16594.53846 | 31.65154 |

The means of the clusters provides insights into the central tendencies of the three variables for each cluster:

* **Cluster 1:** Has a mean average income of $84,424, an average population of around 4,168 and an average minority percentage of about 48%.
* **Cluster 2:** Has a mean average income of $68,473, an average population of around 3,655 and an average minority percentage of about 15%.
* **Cluster 3:** Has a mean average income of $99,096, an average population of around 4,916 and an average minority percentage of about 34%.
* **Cluster 4:** Has a mean average income of $55,600, an average population of around 3,483 and an average minority percentage of about 19%.
* **Cluster 5:** Has a mean average income of $91,653, an average population of around 16,595 and an average minority percentage of about 32%.

| **Statistics for Variables** | | | | |
| --- | --- | --- | --- | --- |
| **Variable** | **Total STD** | **Within STD** | **R-Square** | **RSQ/(1-RSQ)** |
| **avg\_income** | 9367 | 2658 | 0.919582 | 11.435043 |
| **population** | 1938 | 1728 | 0.206149 | 0.259682 |
| **minority** | 32.18266 | 28.86034 | 0.196860 | 0.245112 |
| **OVER-ALL** | 5523 | 1831 | 0.890281 | 8.114218 |

The R-Squared values from this table measure the proportion of the variance in each variable that is explained by the clustering. Average income has an R-squared value of 0.92 (92%), which is very high, meaning that the clustering can explain most of the variance in average income based on the clusters. This indicates that the clusters are largely differentiated by average income. In contrast, population has a much lower R-squared value of 0.21 (21%), meaning that the differences between the clusters are less distinct with respect to population. Similarly, the minority rate has a low R-squared value of 0.20 (20%), suggesting that the clusters do not differ much in terms of the proportion of the minority population. Lastly, the overall R-squared value of 0.89 (89%) indicates the combined variance explained by the clusters across the three variables. This high value suggests that the clusters explain a significant portion of the variance across average income, population, and minority rate. The high overall r-squared value is also an indication that the choice of the number of clusters is a good fit for the data.

**Average Income vs. Population**

A graph of different colored lines

AI-generated content may be incorrect.

**Average Income vs. Minority**

A chart of different colored circles

AI-generated content may be incorrect.

**Population vs. Minority**

A graph of a cluster of dots

AI-generated content may be incorrect.

Compared to the hierarchical clustering scatter plots shown earlier in this report, the graphs of Average Income vs. Population and Average Income vs. Minority do not differ significantly in structure and show only small changes in how the clusters are grouped. The most notable change observed was in cluster group four, which was centered around an average income of $75,000 in the hierarchical scatter plots. In the k-means scatter plots, this group was divided into two distinct clusters, with the left portion reassigned to cluster group two and the right portion to cluster group one.

Similarly, as with the hierarchical scatter plot for Population vs. Minority, the k-means scatter plot shows that the data points are mostly clumped together, with no distinct cluster groups. This supports the observation that population and minority exhibit minimal variability when analyzed together. This finding aligns with the earlier R-squared values, which show that average income has an R-squared value of 0.92, indicating it explains 92% of the variance in the clustering. This highlights that average income is the primary factor driving the differentiation between cluster groups. This conclusion is further supported by the scatter plots, which show that distinct clusters only appear when average income is one of the variables.

## 4 Cluster Analysis

| **Cluster Means** | | | |
| --- | --- | --- | --- |
| **Cluster** | **avg\_income** | **population** | **minority** |
| **1** | 90517.64706 | 15487.52941 | 34.84824 |
| **2** | 66165.66667 | 3579.38500 | 13.77047 |
| **3** | 83759.55576 | 4137.97038 | 46.11170 |
| **4** | 99117.77003 | 4898.72474 | 34.22429 |

* **Cluster 1:** Has a mean average income of $90,518, an average population of around 15,488 and an average minority percentage of about 35%.
* **Cluster 2:** Has a mean average income of $66,166, an average population of around 3,579 and an average minority percentage of about 14%.
* **Cluster 3:** Has a mean average income of $83,760, an average population of around 4,138 and an average minority percentage of about 46%.
* **Cluster 4:** Has a mean average income of $99,118, an average population of around 4,899 and an average minority percentage of about 34%.

The cluster means for the four-cluster group versus the five-cluster group are not drastically different; rather, the reduction to four clusters results in the means of the three variables being more generalized. For example, in the 5-cluster analysis, Cluster 4, which represented the lowest-income group with a low population and minority rate, appears to have been absorbed into other clusters in the 4-cluster analysis. The advantage of using the 4-cluster solution is that the broader, more consolidated clusters make it easier to interpret overall trends without adding unnecessary complexity to the analysis.

| **Statistics for Variables** | | | | |
| --- | --- | --- | --- | --- |
| **Variable** | **Total STD** | **Within STD** | **R-Square** | **RSQ/(1-RSQ)** |
| **avg\_income** | 9367 | 2943 | 0.901365 | 9.138413 |
| **population** | 1938 | 1712 | 0.220767 | 0.283314 |
| **minority** | 32.18266 | 29.56664 | 0.156792 | 0.185947 |
| **OVER-ALL** | 5523 | 1966 | 0.873412 | 6.899666 |

In the 4-cluster analysis, average income has an R-squared value of 0.90 (90%), which supports the observation that the differences between cluster groups are primarily explained by average income. In contrast, the R-squared values for population (0.22 or 22%) and minority rate (0.16 or 16%) are much smaller, indicating that these variables contribute minimally to the differentiation between clusters. The overall R-squared value of 0.87 (87%) suggests that the clusters still explain a significant portion of the variance across all three variables. While the overall R-squared value and the R-squared for average income are slightly lower than in the 5-cluster analysis, the high values indicate that the 4-cluster solution is still a good fit for the data. Opting for the 4-cluster analysis simplifies the interpretation and helps in avoiding any added complexity.

**Average Income vs. Population**

**A graph of a number of people

AI-generated content may be incorrect.**

**Average Income vs. Minority**

**A graph of different colored circles

AI-generated content may be incorrect.**

**Population vs. Minority**

A diagram of a cluster of dots

AI-generated content may be incorrect.

Interestingly, the color and size of Cluster Group One remained consistent between the hierarchical and k-means scatter plots for the graphs of Average Income vs. Population and Average Income vs. Minority. However, Cluster Group Three in the k-means model retained its size from the hierarchical model but also merged with Cluster Group Four, which was distinct in the hierarchical model. In the k-means model for Average Income vs. Population, Cluster Group One is distinctly different, encompassing data points with average incomes ranging from $85,000 to $95,000 and population sizes over 10,000. In contrast, in the Average Income vs. Minority graph, Cluster Group One appears intertwined with Cluster Group Four and includes data points with average incomes of around $95,000 and varying minority rates. Cluster Group Three maintains a structure like that in the hierarchical scatter plots for both graphs, except for some data points that shifted to Cluster Group One in the k-means model.

Lastly, it’s not surprising that the Population vs. Minority scatter plot shows minimal difference between the hierarchical method and the k-means model, as well as between the 5-cluster scatter plots in both methods. As discussed earlier, population and minority exhibit minimal variability when analyzed together, which explains the lack of distinct separation in the clusters. While I prefer the hierarchical scatter plots for displaying the cluster groups in the 4-cluster analysis, I still support using four clusters. This approach yields results like the 5-cluster analysis but offers greater simplicity in the overall analysis of the data.

# ANOVA Test on Approval Rate:

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Model** | 3 | 0.46777546 | 0.15592515 | 6.09 | 0.0004 |
| **Error** | 3061 | 78.39844224 | 0.02561204 |  |  |
| **Corrected Total** | 3064 | 78.86621770 |  |  |  |

| **R-Square** | **Coeff Var** | **Root MSE** | **approval\_rate Mean** |
| --- | --- | --- | --- |
| 0.005931 | 19.85891 | 0.160038 | 0.805873 |

| **Source** | **DF** | **Anova SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **CLUSTER** | 3 | 0.46777546 | 0.15592515 | 6.09 | 0.0004 |

* **The null hypothesis (H0):** The clusters are **not** **related** to the approval rate, meaning the approval rate does not differ significantly across the clusters.
* **The alternative hypothesis (HA):** The clusters **are related** to the approval rate, meaning at least one cluster has a significant different approval rate compared to others.

Since the p-value of 0.0004 is smaller than 0.05, we reject the null hypothesis and conclude that the clusters are related to the approval rate. The cluster's ANOVA sum of squares (0.46777546) represents the portion of variability in approval rates explained by the clusters. While this value is relatively small compared to the total sum of squares (78.86621770), it still indicates that the clusters explain some variability in approval rates. The R-squared value of 0.005931 (or 0.5931%) shows that clusters account for a small portion of the variance in approval rates. Although this is a small percentage, it still suggests a relationship between clusters and approval rates. There may be other factors, not captured by the clustering, that contribute to the variability in approval rates.

**A diagram of a distribution of approval rate

AI-generated content may be incorrect.**

This box plot visualized the distribution of the approval rate variable across the four cluster groups.

* **Cluster 1**: Approval rates exhibit very little variation, with values generally high (around 0.9). There are only two minor outliers.
* **Cluster 2:** Approval rates show significant variability, with the widest range among the clusters. The median approval rate is approximately 0.8, but the cluster includes numerous outliers, some with extremely low values close to 0.
* **Cluster 3:** Approval rates display substantial variability, with a median like Cluster 2. However, this cluster has the most outliers, many of which have very low approval rates.
* **Cluster4:** Approval rates have lower variability compared to Clusters 2 and 3, with a higher median. There are only a few outliers, with values ranging from 0.4 to 0.6.

This box plot supports the argument that approval rates are related to the clusters. If the clusters were unrelated to approval rates, all the box plots would have similar distributions, medians, and variability. However, as we can see, each cluster group shows distinct differences. For example, Cluster 1 is associated with higher and more consistent approval rates, while Clusters 2 and 3 exhibit greater variability and lower approval rates. Like the three variables used in the earlier clustering analysis, approval rates demonstrate meaningful groupings. This indicates that the clusters are effectively dividing the data into groups with distinct differences in approval rates.

# Potential Issues with Outliers and Overall Problems with the Data:

**Hierarchical Cluster Plots**

A graph showing a number of data

AI-generated content may be incorrect.

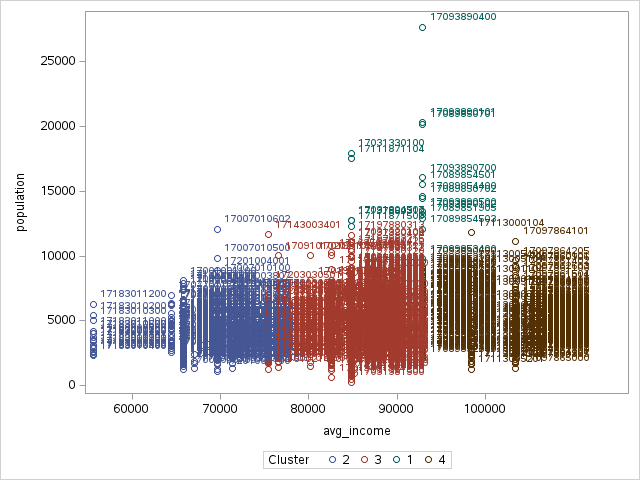


A graph of numbers and a cluster of numbers

AI-generated content may be incorrect.



**K-Means Cluster Plots**





A graph of numbers and a number

AI-generated content may be incorrect.



The only notable outlier observed in both the hierarchical and k-means cluster plots is the data point corresponding to census tract 17093890400. This outlier is due to the population value exceeding the maximum range of 25,000 shown in the graphs. Despite the presence of this outlier, it is not necessary to remove it, as it likely reflects a valid data point. Extreme values are common in many datasets and can provide meaningful variation.

**Outlier Included**

| **Analysis Variable : population** | | | |
| --- | --- | --- | --- |
| **N** | **Mean** | **Median** | **Std Dev** |
| 3065 | 4162.81 | 3899.00 | 1938.25 |

**Outlier Not Included**

| **Analysis Variable : population** | | | |
| --- | --- | --- | --- |
| **N** | **Mean** | **Median** | **Std Dev** |
| 3064 | 4155.15 | 3899.00 | 1891.63 |

To ensure that the outlier does not heavily affect the population variable, I ran summary statistics both with and without the outlier. Comparing the results, there were no significant changes in the mean or standard deviation, confirming that the outlier is a valid data point, it is likely an extreme value within the population variable, which caused it to appear farther from other data points in the graphs analyzing population.

The Population vs. Minority graphs, both hierarchical and k-means, initially posed a problem as the cluster groups were tightly clumped together, making it difficult to distinguish the clusters or their structures. However, this does not invalidate the other cluster analyses but rather suggests that population and minority have limited variation when analyzed together. The k-means analysis revealed that the average income variable had an R-squared value of 0.901 (90%), indicating that the differences between cluster groups are primarily explained by average income. This is evident in the analyses of Average Income vs. Population and Average Income vs. Minority, where the cluster groups are more clearly distinguishable in both hierarchical and k-means clustering.

# Conclusion:

After running both hierarchical and k-means cluster analysis on the Home Mortgage Disclosure Act (HMDA) dataset in SAS to determine the optimal number of cluster groups, **I concluded that four clusters would be the best choice.** During the hierarchical cluster analysis, the R-squared values significantly improved around four or five clusters. Additionally, the Pseudo T-Squared line plot and dendrogram both supported four cluster groups as the ideal number. To further validate this choice, I also ran the analysis with five clusters for comparison. Both the hierarchical and k-means analyses showed that the structures of the four and five cluster groups were not significantly different, nor were the k-means summary statistics. Since there was no major change between the two, I believe that using the four-cluster analysis simplifies interpretation and avoids unnecessary complexity. The tables and graphs provided throughout this report support this conclusion.