**Customer Segmentation for Targeted Promotions**

**BUAN 6383.001**

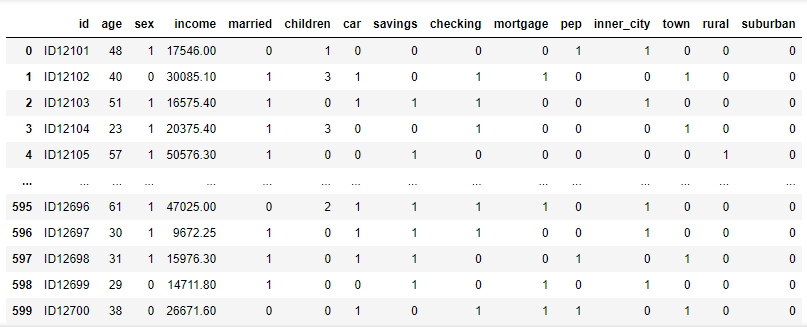
**Syam Menon**

Mihir Hirave (MDH230000)

**CLUSTERING CUSTOMERS**

1. **Notice that region is categorical; we need to do what is referred to as “one-hot encoding” – convert it into separate (binary) variables, one for each possible value of region. So you will need to create 4 new variables, corresponding to inner city, town, rural, and suburban (a 1 in a column would represent being from the associated region). You can do this either explicitly by writing your own code, or by using the OneHotEncoder option available in sklearn (preprocessing). Read in the data, create the four new columns, and drop region.**

**We use pd.get\_dummies to encode “region” variables and we later dropped the original and added new variables to our original dataset**

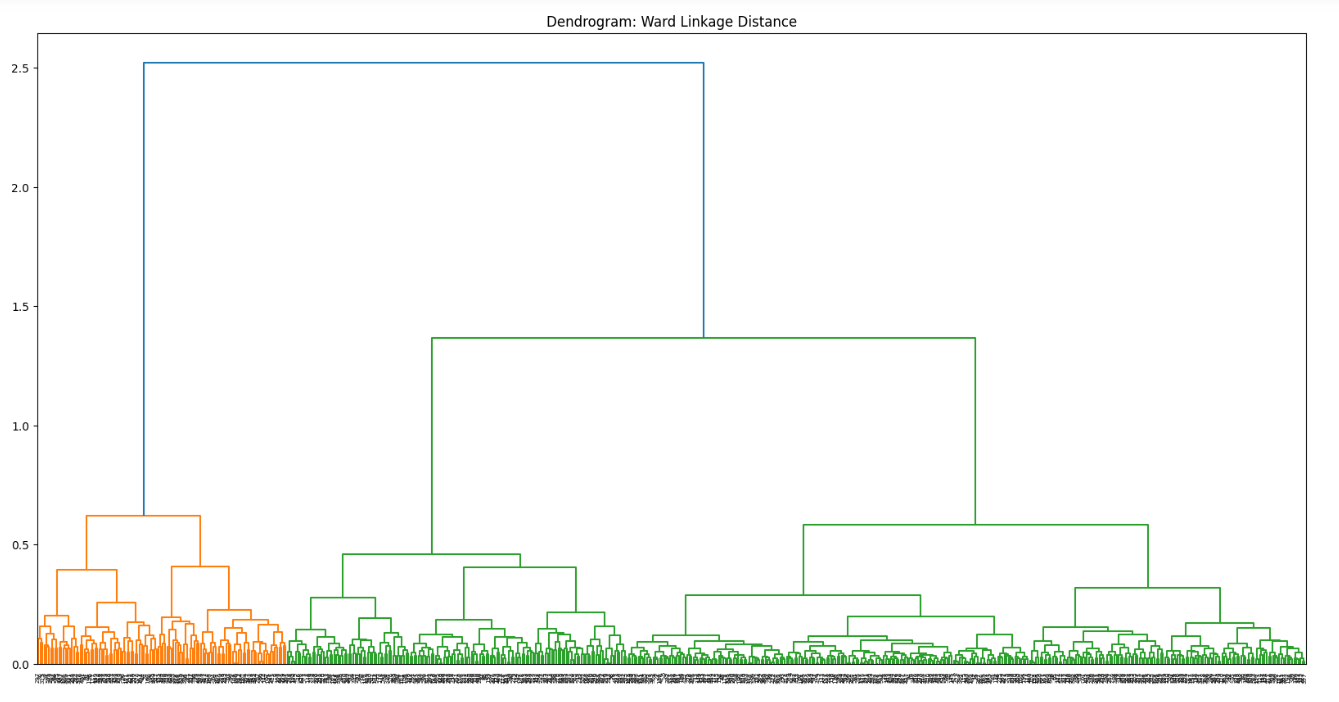


1. **Apply hierarchical clustering (with Euclidian distance as the measure of distance) to the dataset using (i) centroid linkage, (ii) single linkage, (iii) complete linkage, (iv) average linkage, and (v) Ward linkage. For each of these, comment on whether you see any clear clusters, and how many clusters you would recommend (and why). Across all the linkage approaches tried, which one has worked best in this example (provide your reasoning)? What are some distinguishing characteristics of each cluster?**

We decided to apply a logarithmic scale to “income” since we found the distribution of income to be right-skewed.

We used all 5 methods of hierarchical clustering and found out the most optimum cluster to be Ward Linkage.

The reason for choosing this method was due to the intra-cluster distance being minimum (the vertical lines at the bottom of the graph indicate a small distance between each cluster) & inter cluster distance being maximum (the vertical lines above ~0.8 are significantly longer, showing a bigger distance between clusters).



This way we could easily pick our cut-off range (0.8) which determined our number of clusters

1. Centroid Linkage:

When using centroid linkage distance, we find 2 distinct clusters with large distances between the two centroids. This is an okay cluster

1. Single Linkage:

This is the worst cluster by far. It seems to have very disproportionate clusters with most of the records falling into a single cluster. Such clusters are not helpful.

1. Complete Linkage:

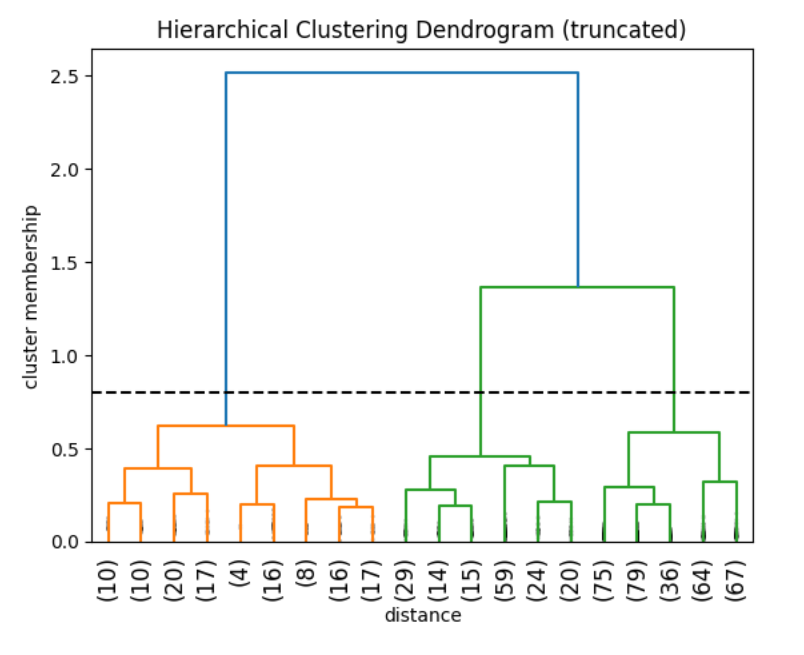
This is a great cluster and the best one till now. All three clusters show distinct differences.

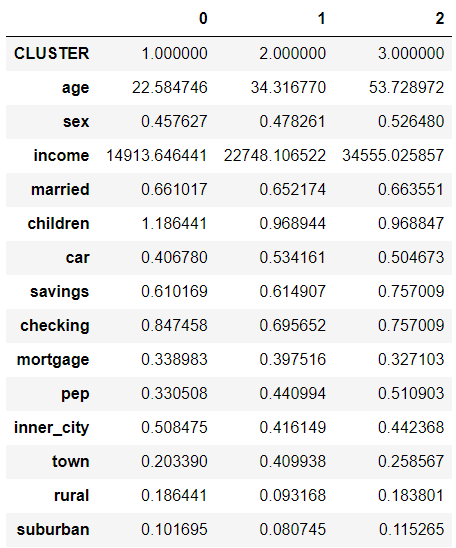
1. Average Linkage:

This cluster seems to be skewed towards only one cluster. It isn’t as good as the Complete linkage cluster but is better than the first two clusters (centroid & single).

1. Ward Linkage:

This is our best cluster so far, since it has the highest inter cluster distance and the lowest intra cluster distance. We shall choose this one from the 5 clusters.





Cluster 0: Represents the young population

Cluster 1: Represents the middle-aged population

Cluster 2: Represents the people near the age of retirement i.e. nearing 60

Distinguishing Characteristics:

We decided to focus on demographic information to distinguish and characterize our 3 clusters.

Cluster 0: represented the young population, with the lowest income (~$15K). A little over half of this cluster is married, and they have 1 child on average.

Cluster 1: represented the adults in their 30’s, with an average income (~$23K). A little over half of this cluster is married, and they have less children than young people.

Cluster 2: represented the people nearing their retirement age of 60, with the highest income (~$35K). A little over half of this cluster is married, and they have about the same number of children as adults in their 30’s

The clustering method we picked is “WARD LINKAGE METHOD” because it had the maximum inter-cluster distance and minimum intra-cluster distance.

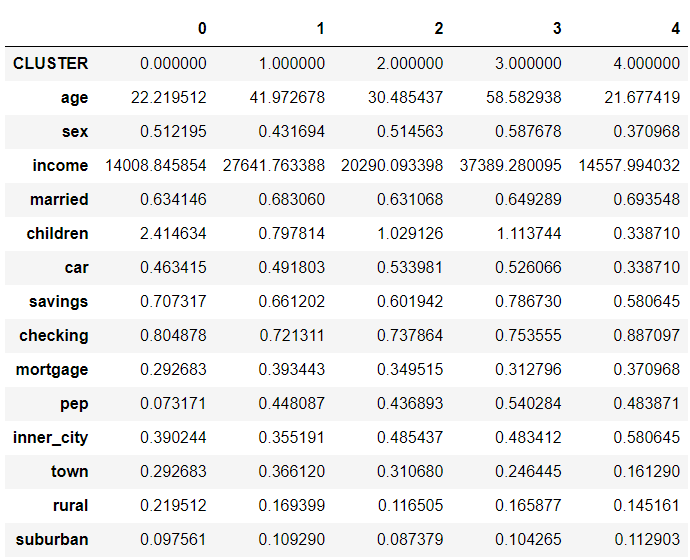
In this cluster we found that:

* More middle – aged people tend to have a car than older people in general
* Older people massively preferred having a savings account over the other clusters while on the other hand the youth preferred having a checking account
* Middle- aged people tend to take on a mortgage compared to others which makes sense since that’s the age when most people staring earning a stable income
* Personal Equity plan does not show any significant dissimilarity, but old people prefer it more than the others
* Young people prefer to stay in the inner city, while the middle-aged folks prefer staying in the town.

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

We applied k-means clustering on the dataset with different values of k [3, 4, 5, 6, 7, and 8] and found that k=5 gave us the most distinct clusters; hence, we would recommend 5 clusters. This is more than the clusters we decided to include when using hierarchical clustering, which was only 3.

The results of k-means clustering for k=5 is shown below:



Distinguishing Characteristics:

Cluster 3: represented majority women nearing retirement.

Cluster 2: represented adults in their late 20’s or early 30’s.

Cluster 0: represented the young parents in their early 20s.

Cluster 4: represented predominantly young men in their early 20’s

Cluster 1: represented majority middle-aged men in their 40’s.

We can see that all clusters seem to have people who opt-in to have a checking account, but the young men seem to top this category.

On the other hand, they aren't as likely to open a savings account. Here the older women top this category. Adults generally avoid opening a savings account almost like young men.

Here we could identify young parents who don’t want a pep. These people are less likely to have a personal equity plan because they may prioritize more immediate financial needs such as education and health expenses of their children. Here, targeting family-oriented financial products for these individuals might be effective. For example, college funds for their children.

Surprisingly Pep was a major consideration in the oldest and the youngest of people. Retirees might have good personal equity for their retirement and young adults, who are fresh out of college are investing in their 401Ks and utilizing their 401K employee match.

Young men had the highest mortgage of them all, while young parents did not have or did not risk have one. As mentioned before, it might be because young families tend to focus on fulfilling the immediate needs, rather than taking out a mortgage.

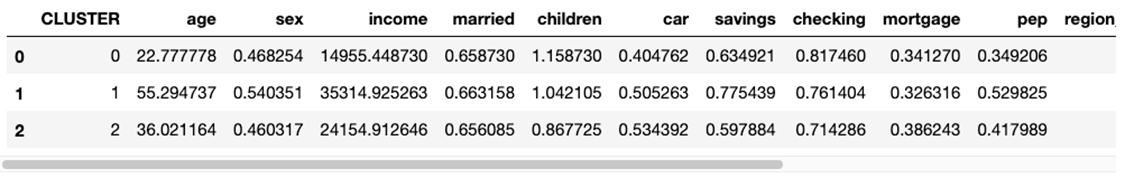
Young parents & Middle-aged men generally avoid living in inner parts of cities, while the young men prefer it.

K-means clustering vs hierarchical clustering

We will first compare k-means clustering when k=3 and hierarchical clustering. Then, we will compare the insights we can gather when comparing k-means clustering for k=5 and hierarchical clustering.

When comparing the results of k-means clustering for k=3 vs the results of hierarchical clustering at a threshold of 0.8, we found the following:

1. **k-means clustering, for k =3**



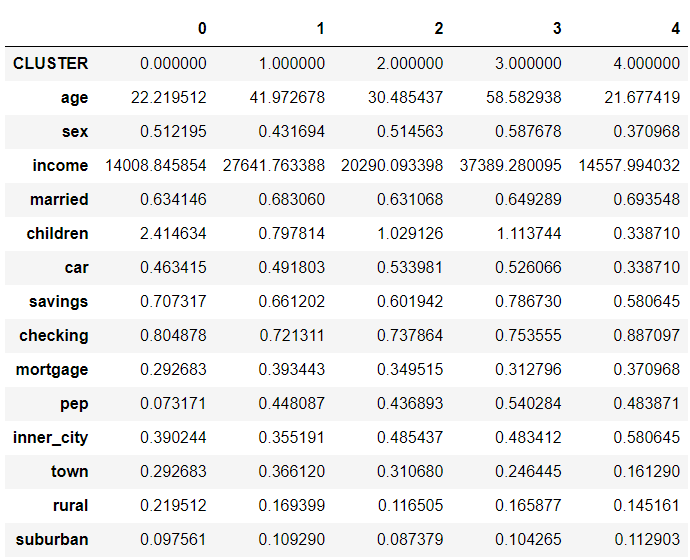
1. **Hierarchical clustering, Ward Linkage, 3 clusters**



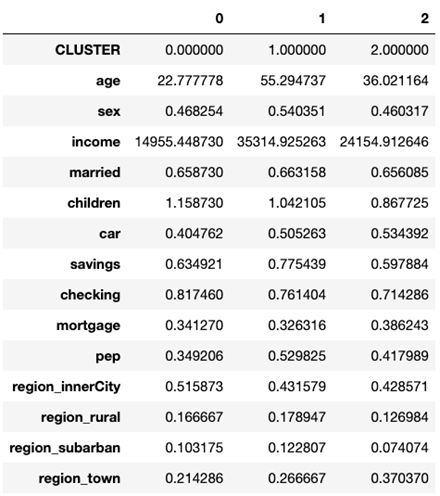
We see that most clusters look similar, with only minimal variation amongst the numbers.

However, when we compare the findings of k-means clustering for k=5, and the findings of hierarchical clustering, we have a lot more to comment on:

1. **K-means clustering for k=5**



1. **Hierarchical Clustering, 3 clusters**



* The two cluster techniques show us a significant difference in the type of people we are clustering. In hierarchical clustering we could identify that we are looking at a diverse age group of people. While in k-means with k=5 we could pinpoint, on top of each cluster’s average age, their sex, their family structure (are they married? Do they have children?) and can better make an estimate on what their financial priorities are.
* For example, we could not distinguish which clusters had kids or who was married in hierarchical clustering, but we have a much more insightful division with k-means clustering. This information was important because it can be a driving factor of why people were not interested in getting a pep. K-means clustering gave us better answers to questions such as: Which group has a high mortgage? Who is likely to earn/save more? How could location impact spending/saving habits?
* Because we are more interested in knowing what the financial needs and habits of our customers are, we would choose k-means clustering over hierarchical clustering.

1. **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. As in Part I, this question is open ended, and I am looking for innovative thinking.**

If we were managers of this bank, we would mainly focus on promoting PEP and mortgage to cluster 4 and cluster 2, respectively.

First, we will recommend PEP to cluster 4, because most individuals in this cluster have not invested in PEP; most probably due to lack of awareness of such products. They have higher income than individuals in cluster 3, however, invest less in PEP. We might further promote PEP to these individuals by informing them that these financial products have tax incentives. This will lower their taxes and hence give them more disposable income to either invest more in PEP or use it for their personal expenses.

Secondly, we will promote mortgages to cluster 2 because their income is similar to cluster 3 but their mortgage is much lower than cluster 3. We could give special mortgage rates to these individuals according to their credit scores and encourage them to take mortgages from us.

In addition, cluster 3 has the largest number of children (2.38), thus it’s more likely that they will need to buy a house. But, because of their immediate expenses, they are concerned that their earnings may not be adequate to cover mortgage payments. We could start financial counseling for young families in our bank to help families assess their financial situation, understand mortgage options, and create a plan that aligns with their budget and goals. Furthermore, we could inform families about the government-sponsored programs such as FHA loans for first-time homebuyers. Furthermore, we could also encourage young families to start college funds for their kids.