Unsupervised Learning
Clustering Bank Customers

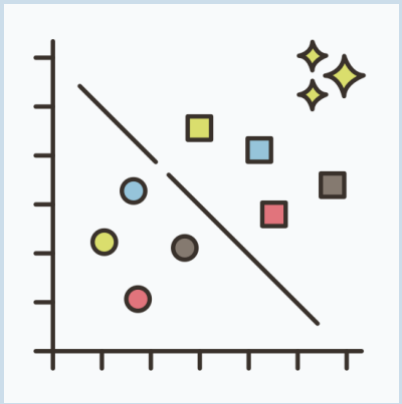


Table of Contents

1. Heirachical Clustering ..... - 3 -

Hierarchical Clustering: Centroid Linkage ..... - 3 -

Dendrogram: Centroid Linkage Distance ..... - 3 -

Hierarchical Clustering Dendrogram using Centroid Linkage Distance (truncated): ..... - 3 -

Cluster number with the numbers of customers in each cluster: ..... - 4 -

The average value for each column by cluster is as follows: ..... - 4 -

Observation ..... - 4 -

2. Hierarchical Clustering: Single Linkage ..... - 5 -

Dendrogram: Single Linkage Distance ..... - 5 -

Hierarchical Clustering Dendrogram using Single Linkage Distance (truncated): ..... - 5 -

Cluster number with the numbers of customers in each cluster: ..... - 6 -

The average value for each column by cluster is as follows: ..... - 6 -

Observations ..... - 6 -

3. Hierarchical Clustering: Complete Linkage ..... - 7 -

Dendrogram: Complete Linkage Distance ..... - 7 -

Hierarchical Clustering Dendrogram using Complete Linkage Distance (truncated): ..... - 7 -

Cluster number with the numbers of customers in each cluster: ..... - 8 -

The average value for each column by cluster is as follows: ..... - 8 -

Observations ..... - 8 -

4. Hierarchical Clustering: Average Linkage ..... - 9 -

Dendrogram: Average Linkage Distance ..... - 9 -

Hierarchical Clustering Dendrogram using Average Linkage Distance (truncated): ..... - 10 -

Cluster number with the numbers of customers in each cluster: ..... - 10 -

The average value for each column by cluster is as follows: ..... - 10 -

Observations: ..... - 10 -

5. Hierarchical Clustering: Ward Linkage ..... - 12 -

Dendrogram: Ward Linkage Distance ..... - 12 -

Hierarchical Clustering Dendrogram using Ward Linkage Distance (truncated): ..... - 12 -

Cluster number with the numbers of customers in each cluster: ..... - 13 -

The average value for each column by cluster is as follows: ..... - 13 -

Observations ..... - 14 -

Conclusion for Hierarchical Clustering ..... - 15 -

6. Applying k-means clustering to the dataset ..... - 16 -

k-means with 3 Clusters ..... - 16 -

Cluster number with the numbers of customers in each cluster: - 16 -

The average value for each column by cluster is as follows: - 16 -

Observations - 16 -

7. k-means with 4 Clusters - 17 -

Cluster number with the numbers of customers in each cluster: - 17 -

The average value for each column by cluster is as follows: - 17 -

Observations - 17 -

8. k-means with 5 Clusters - 18 -

Cluster number with the numbers of customers in each cluster: - 18 -

The average value for each column by cluster is as follows: - 18 -

Observations - 18 -

9. k-means with 6 Clusters - 19 -

Cluster number with the numbers of customers in each cluster: - 19 -

The average value for each column by cluster is as follows: - 19 -

Observations - 20 -

10. k-means with 7 Clusters - 20 -

Cluster number with the numbers of customers in each cluster: - 20 -

The average value for each column by cluster is as follows: - 20 -

Observations - 20 -

11. k-means with 8 Clusters - 21 -

Cluster number with the numbers of customers in each cluster: - 21 -

The average value for each column by cluster is as follows: - 21 -

Observations - 21 -

12. Conclusion of k-means clustering: - 22 -

The Count of Customers Falling into Different Clusters: - 22 -

The Value for the Characteristics of the Clusters: - 23 -

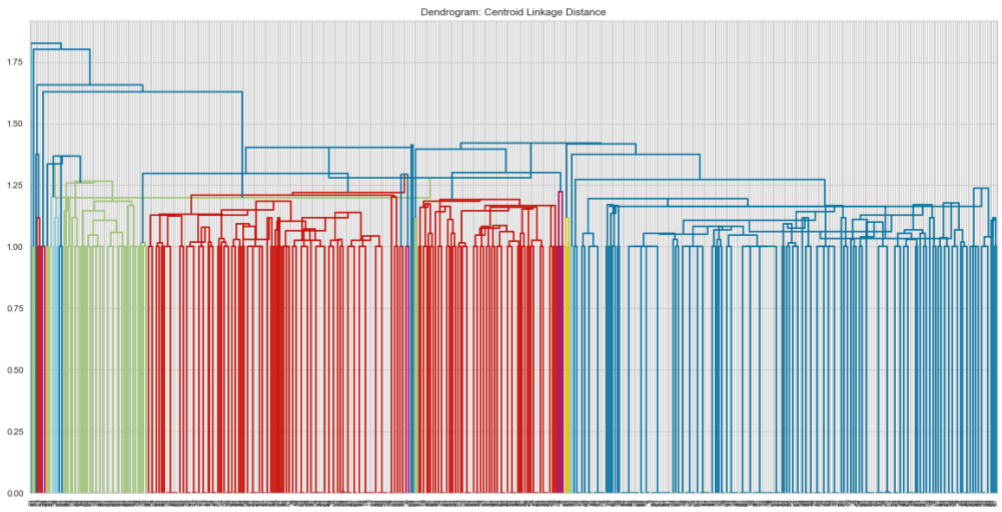
13. Managerial takeaways - 23 -

Clustering bank customers

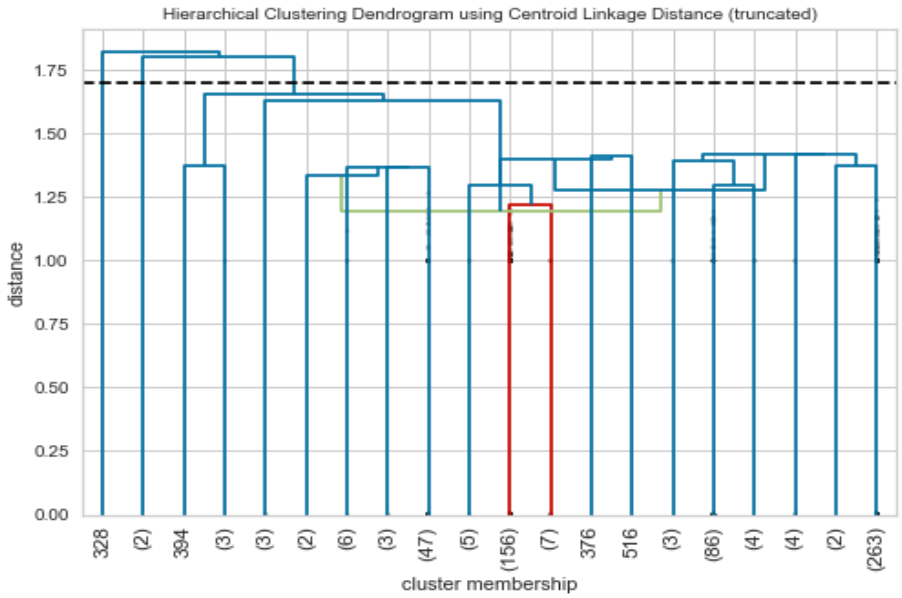
1. Heirachical Clustering

Hierarchical Clustering: Centroid Linkage

Dendrogram: Centroid Linkage Distance



Hierarchical Clustering Dendrogram using Centroid Linkage Distance (truncated):



Cluster number with the numbers of customers in each cluster:

CLUSTER NUMBER	COUNT
1	2
2	597
3	1

The average value for each column by cluster is as follows:

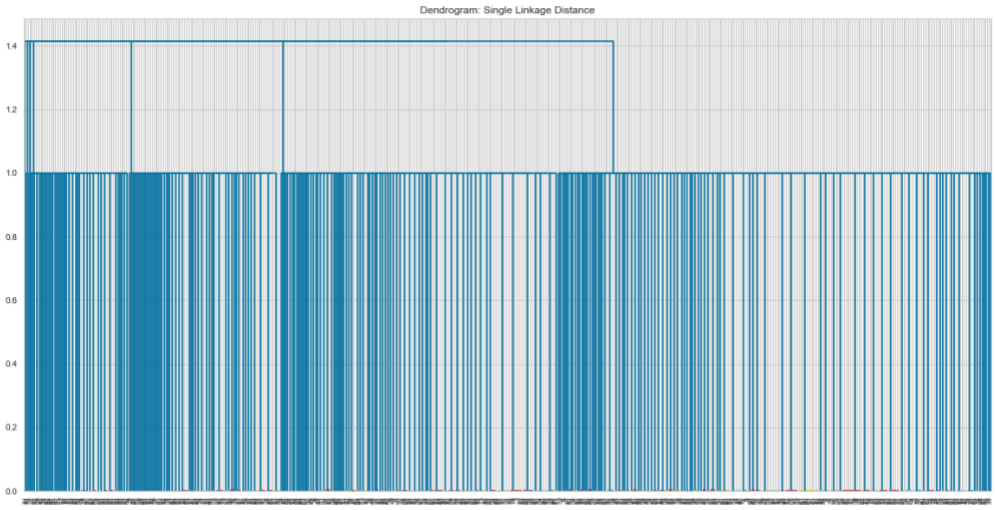
CLUSTER	age	sex	income	married	children	car	savings	checking	mortgage	pep	inner_city	town	rural	suburban
1	33.50	0.00	21459.95	0.00	2.50	1.00	0.00	0.00	0.50	0.00	0.00	0.00	1.00	0.00
2	42.40	0.50	27505.28	0.66	1.01	0.49	0.69	0.76	0.35	0.46	0.45	0.29	0.16	0.10
3	57.00	1.00	50849.20	0.00	1.00	0.00	1.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00

Observation

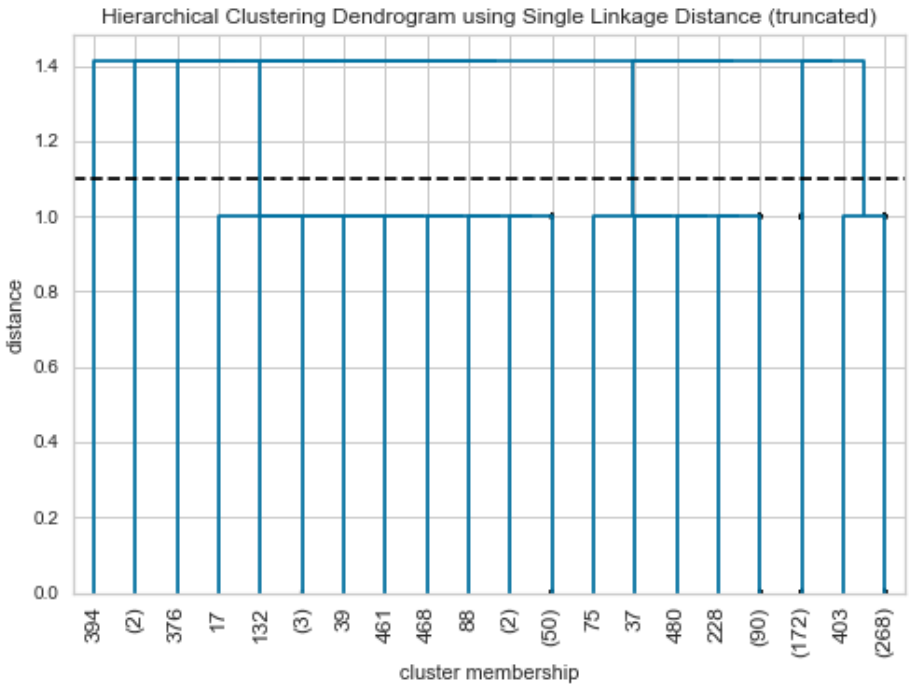
• According to the dendrogram and clusters generated above, we could barely detect clear clusters since most of the customers fall into a large cluster (cluster 2) while only 2 falls into cluster 1 and only 1 customer falls into cluster 3.

2. Hierarchical Clustering: Single Linkage

Dendrogram: Single Linkage Distance



Hierarchical Clustering Dendrogram using Single Linkage Distance (truncated):



Cluster number with the numbers of customers in each cluster:

CLUSTER NUMBER	COUNT
1	2
2	61
3	94
4	172
5	269
6	1
7	1

The average value for each column by cluster is as follows:

CLUSTER	age	sex	income	married	children	car	savings	checking	mortgage	pep	inner_city	town	rural	suburban
1	33.50	0.00	21459.95	0.00	2.50	1.00	0.00	0.00	0.50	0.00	0.00	0.00	1.00	0.00
2	43.49	0.48	28708.62	0.69	0.98	0.39	0.70	0.82	0.34	0.54	0.00	0.00	0.00	1.00
3	43.21	0.50	30209.90	0.65	1.20	0.51	0.74	0.77	0.29	0.49	0.00	0.00	1.00	0.00
4	42.11	0.53	26749.04	0.67	1.00	0.53	0.74	0.74	0.37	0.41	0.00	1.00	0.00	0.00
5	41.99	0.49	26844.00	0.66	0.95	0.48	0.64	0.76	0.35	0.46	1.00	0.00	0.00	0.00
6	58.00	1.00	33204.30	0.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	1.00	0.00	0.00
7	58.00	1.00	25468.50	0.00	0.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	1.00

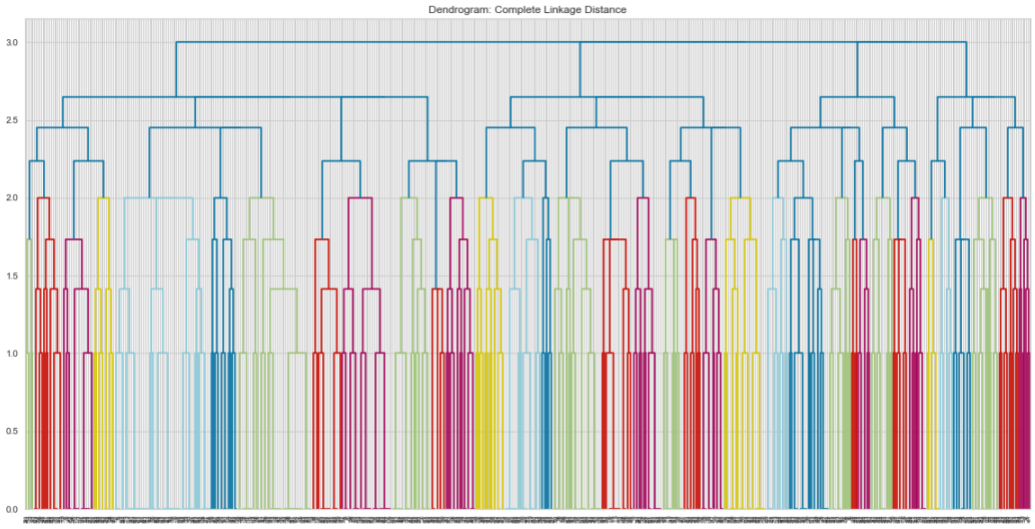
Observations

• According to the dendrogram and number of customers in each cluster, we could detect 7 clusters but the number of customers falling into these clusters differ a lot. Nearly half of the customers fall into CLUSTER 4, nearly 30% of the left customers fall into CLUSTER 3. And there are only 1 or 2 customers falling into CLUSTER 1, 6 and 7.

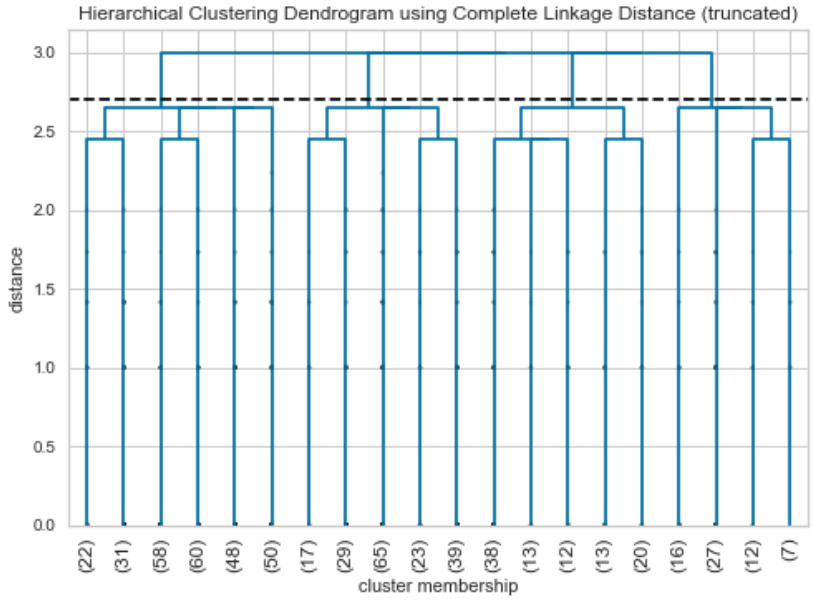
• Based on the characteristics of these clusters, CLUSTER 2, 3, 4 and 5, which cover most of the customers, are not distinguishable. For instance, there is no clear difference in customers' age, sex, income, marriage status, owning car status, savings, checkings, mortgage, pep among these four clusters.

3. Hierarchical Clustering: Complete Linkage

Dendrogram: Complete Linkage Distance



Hierarchical Clustering Dendrogram using Complete Linkage Distance (truncated):



Cluster number with the numbers of customers in each cluster:

CLUSTER NUMBER	COUNT
1	269
2	173
3	96
4	62

The average value for each column by cluster is as follows:

CLUSTER	age	sex	income	married	children	car	savings	checking	mortgage	pep	inner_city	town	rural	suburban
1	41.99	0.49	26844	0.66	0.95	0.48	0.64	0.76	0.35	0.46	1.00	0.00	0.00	0.00
2	42.2	0.53	26786.35	0.66	1	0.53	0.74	0.74	0.38	0.41	0.00	1.00	0.00	0.00
3	43.01	0.49	30027.61	0.64	1.23	0.52	0.73	0.75	0.29	0.48	0.00	0.00	1.00	0.00
4	43.73	0.48	28656.36	0.68	0.97	0.4	0.69	0.81	0.35	0.55	0.00	0.00	0.00	1.00

Observations

- According to the dendrogram above, we could clearly detect 4 clusters. And the distribution of customers among these four clusters are much more evenly compared to the first two methods -- using centroid and single linkage distances.

- The four clusters are mainly divided based on region. Specifically, customers from CLUSTER 1 are from inner city, CLUSTER 2 is from town, CLUSTER 3 from rural, and CLUSTER 4 comes from suburban.

- As for the other characteristics of each cluster, CLUSTER 1 has the lowest average age, average number of children, proportion of having saving accounts; average level in the other characteristics, including proportion of female, average income, proportion of being married, proportion of owning car, having checking accounts, having personal equity plan, etc.

- CLUSTER 2 has the highest proportion of female, owning car, having saving accounts; lowest average income, and proportion of having checking accounts; maintains average level in the other characteristics.

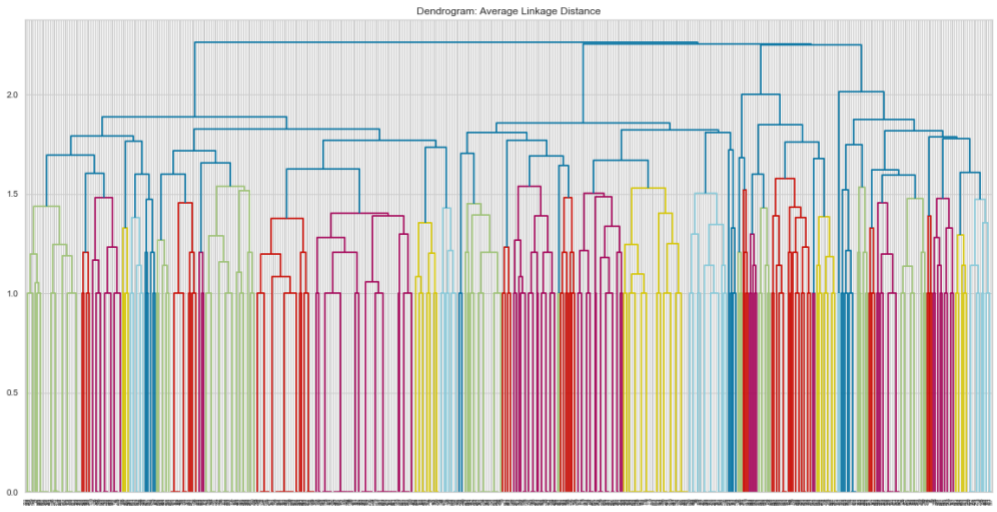
- CLUSTER 3 has highest average income, average number of children, lowest proportion of being married and owning mortgage. The customers have average level in the other measurements.

- CLUSTER 4 has highest average age, being married, having checking accounts, and having personal equity plan; however, has lowest proportion of female and owning car. Besides, the customers from this cluster maintain average level in other aspects.

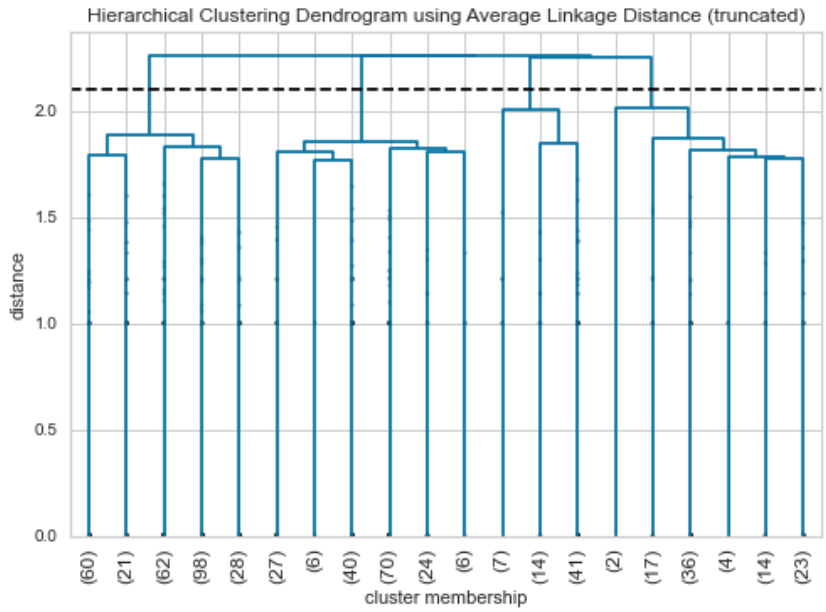
- Overall, the four clusters tend to have higher similarity inside each cluster but have dissimilarity across different clusters.

4. Hierarchical Clustering: Average Linkage

Dendrogram: Average Linkage Distance



Hierarchical Clustering Dendrogram using Average Linkage Distance (truncated):



Cluster number with the numbers of customers in each cluster:

CLUSTER NUMBER	COUNT
1	269
2	173
3	62
4	96

The average value for each column by cluster is as follows:

CLUSTER	age	sex	income	married	children	car	savings	checking	mortgage	pep	inner_city	town	rural	suburban
1	41.99	0.49	26844.00	0.66	0.95	0.48	0.64	0.76	0.35	0.46	1.00	0.00	0.00	0.00
2	42.20	0.53	26786.35	0.66	1.00	0.53	0.74	0.74	0.38	0.41	0.00	1.00	0.00	0.00
3	43.73	0.48	28656.36	0.68	0.97	0.40	0.69	0.81	0.35	0.55	0.00	0.00	0.00	1.00
4	43.01	0.49	30027.61	0.64	1.23	0.52	0.73	0.75	0.29	0.48	0.00	0.00	1.00	0.00

Observations:

- According to the dendrogram above, we could clearly detect 4 clusters. Also, similar to the results using complete linkage distance, the clusters are mainly divided by their region. In details, CLUSTER 1 only contains customers from inner city, CLUSTER 2 is from town, CLUSTER 3 from suburban, and CLUSTER 4 comes from rural.

- As for the other characteristics of each cluster, CLUSTER 1 has the lowest average age, average number of children, proportion of having saving accounts; maintains average level in the other characteristics, including proportion of female, average income, proportion of being married, proportion of owning car, having checking accounts, having personal equity plan, etc.

- CLUSTER 2 has the highest proportion of female, owning car, having saving accounts; lowest average income, and proportion of having checking accounts; maintains average level in the other characteristics.

- CLUSTER 3 has highest average age, being married, having checking accounts, and having personal equity plan; however, has lowest proportion of female and owning car. Besides, the customers from this cluster maintain average level in other aspects.

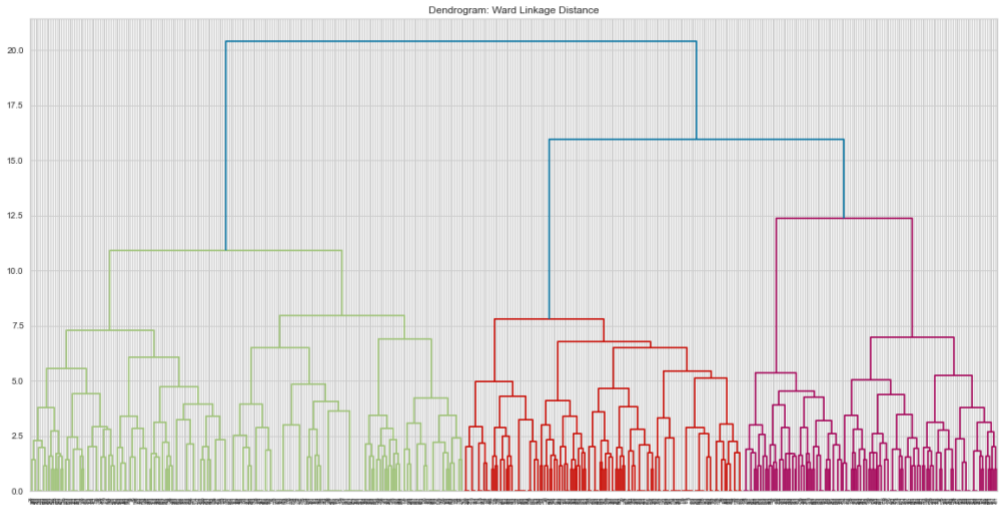
- CLUSTER 4 has highest average income, average number of children, lowest proportion of being married and owning mortgage. The customers have average level in the other measurements.

- Overall, the four clusters tend to have higher similarity inside each cluster but have dissimilarity across different clusters.

- Moreover, the clusters generated based on complete and average linkage distances are totally same, by ignoring the cluster order.

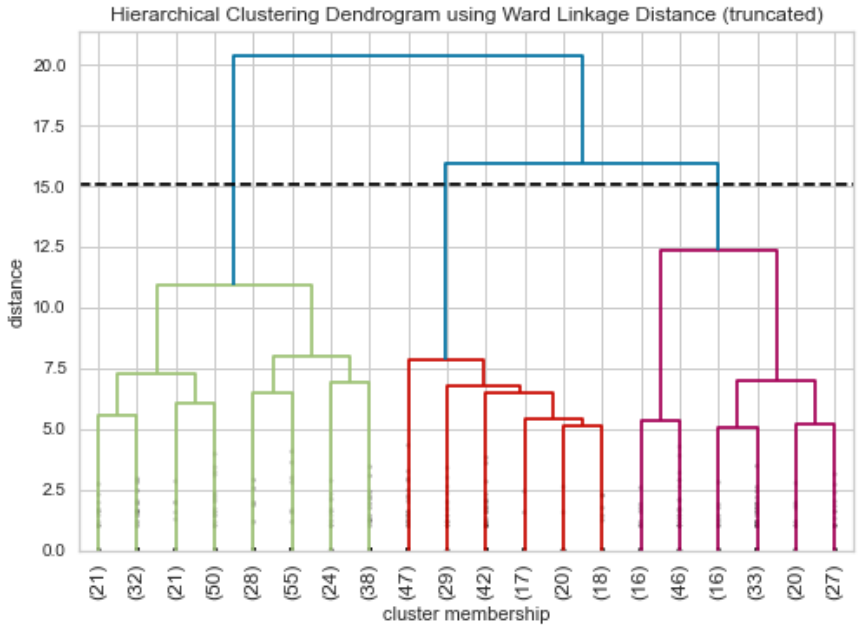
5. Hierarchical Clustering: Ward Linkage

Dendrogram: Ward Linkage Distance

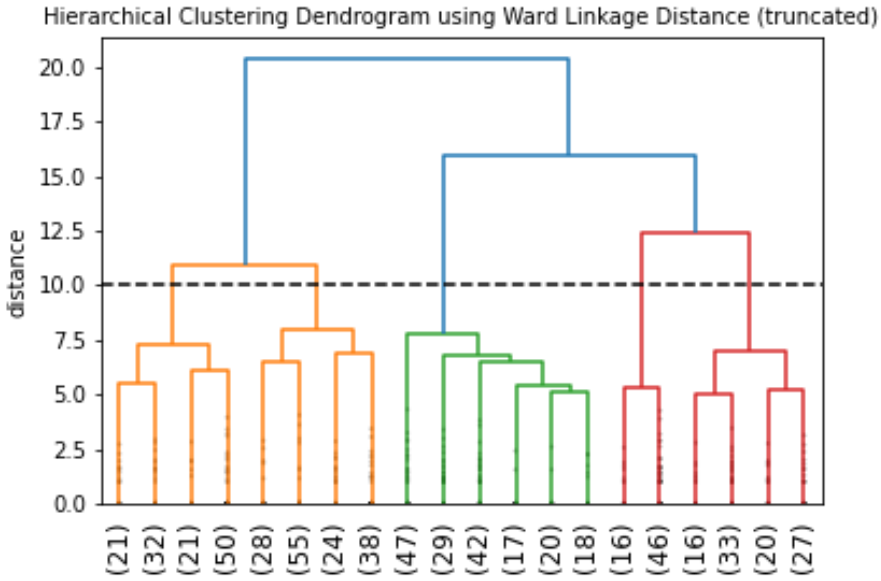


Hierarchical Clustering Dendrogram using Ward Linkage Distance (truncated):

a. 3 Clusters



b. 5 Clusters



Cluster number with the numbers of customers in each cluster:

a. 3 Clusters:

CLUSTER NUMBER	COUNT
1	269
2	173
3	158

b. 5 Clusters:

CLUSTER NUMBER	COUNT
1	124
2	145
3	173
4	62
5	96

The average value for each column by cluster is as follows:

a. 3 Clusters:

CLUSTER	age	sex	income	married	children	car	savings	checking	mortgage	pep	inner_city	town	rural	suburban
1	41.99	0.49	26844.00	0.66	0.95	0.48	0.64	0.76	0.35	0.46	1.00	0.00	0.00	0.00
2	42.20	0.53	26786.35	0.66	1.00	0.53	0.74	0.74	0.38	0.41	0.00	1.00	0.00	0.00
3	43.29	0.49	29489.52	0.65	1.13	0.47	0.72	0.77	0.32	0.51	0.00	0.00	0.61	0.39

b. 5 Clusters:

CLUSTER	age	sex	income	married	children	car	savings	checking	mortgage	pep	Inner_city	town	rural	suburban
1	43.97	0.48	28677.59	0.46	0.99	0.53	0.57	0.77	0.29	0.91	1.00	0.00	0.00	0.00
2	40.30	0.50	25275.96	0.83	0.92	0.44	0.70	0.76	0.40	0.07	1.00	0.00	0.00	0.00
3	42.20	0.53	26786.35	0.66	1.00	0.53	0.74	0.74	0.38	0.41	0.00	1.00	0.00	0.00
4	43.73	0.48	28656.36	0.68	0.97	0.40	0.69	0.81	0.35	0.55	0.00	0.00	0.00	1.00
5	43.01	0.49	30027.61	0.64	1.23	0.52	0.73	0.75	0.29	0.48	0.00	0.00	1.00	0.00

Observations

a. 3 Clusters:

• According to the dendrogram above, we could clearly detect 3 clusters in different colors. And the customers are evenly distributed into these three clusters.

• Based on the characteristics among these clusters, the three clusters tend to have higher similarity inside each cluster but have dissimilarity across different clusters.

• Specially, CLUSTER 1 has lowest average age, largest proportion of male, lowest number of children, lowest proportion of having saving account, from inner city.

• CLUSTER 2 has the highest proportion of female, owning car, having saving account, owning mortagage, but having lowest average income, lowest proportion of having checking account, having personal equity plan. Besides, all of them are living in town.

• CLUSTER 3 has highest average age, income, number of children, highest proportion of having checking accounts and personal equity plan, however lowest proportion of being married, owning car, and owning mortgage. Moreover, all of them are either from rural or suburban.

b. 5 Clusters:

• According to the truncated dendrogram with 5 clusters above, we could also clearly detect 5 clusters. And the distribution of the numbers of customers are evenly distributed into these clusters.

• Based on the characteristics among these clusters, two clusters are from inner city, which is CLUSTER 1 and 2.

• Specially, CLUSTER 1 has lowest marriage rate, lowest proportion of having saving account and mortgage, but 91% of them have personal equity plan, their average age is highest, and their average income ranks second highest among all clusters.

• CLUSTER 2 ranks lowest in average age, average income, the number of children, and having personal equity plan. But they rank highest in marriage rate and the proportion of having mortgage.

• CLUSTER 3 comes from town. They rank highest in proportion of female, owning car, but have second lowest average income, lowest proportion of having checking account.

• CLUSTER 4 is from suburban, which has second highest average age, highest proportion of having checking accounts and personal equity plan, however lowest proportion of owning car.

• CLUSTER 5 contains customers coming from rural, who rank highest in average income, the number of children, but ranked lowest in the proportion of having mortgage.

Conclusion for Hierarchical Clustering

Across all the linkage approaches tried, the dendrogram based on average and complete linkage distances show the same results -- 4 clusters in this example.

According to the ward linkage distances, we could detect 3 clusters roughly but could detect 5 clusters clearly.

We would say 5 clusters using ward linkage distance performs best since this method could clearly divide all customers into as many as clusters as possible while keep the characteristics from each cluster are quite distinguishable.

6. Applying k-means clustering to the dataset

k-means with 3 Clusters

Cluster number with the numbers of customers in each cluster:

CLUSTER NUMBER	COUNT
1	158
2	173
3	269

The average value for each column by cluster is as follows:

CLUSTER	age	sex	income	married	children	car	savings	checking	mortgage	pep	inner_city	town	rural	suburban
1	41.99	0.49	26844.00	0.66	0.95	0.48	0.64	0.76	0.35	0.46	1.00	0.00	0.00	0.00
2	42.20	0.53	26786.35	0.66	1.00	0.53	0.74	0.74	0.38	0.41	0.00	1.00	0.00	0.00
3	43.29	0.49	29489.52	0.65	1.13	0.47	0.72	0.77	0.32	0.51	0.00	0.00	0.61	0.39

Observations

- We could clearly detect 3 clusters and the customers are evenly distributed into these three clusters.

- Based on the characteristics among these clusters, the three clusters tend to have higher similarity inside each cluster but have dissimilarity across different clusters.

- Specially, CLUSTER 1 has highest average age, income, number of children, highest proportion of having checking accounts and personal equity plan, however lowest proportion of being married, owning car, and owning mortgage. Moreover, all of them are either from rural or suburban.

- CLUSTER 2 has the highest proportion of female, owning car, having saving account, owning mortgage, but having lowest average income, lowest proportion of having checking account, having personal equity plan. Besides, all of them are living in town.

- CLUSTER 3 has lowest average age, largest proportion of male, lowest number of children, lowest proportion of having saving account, from inner city.

7. k-means with 4 Clusters

Cluster number with the numbers of customers in each cluster:

CLUSTER NUMBER	COUNT
1	96
2	269
3	173
4	62

The average value for each column by cluster is as follows:

CLUSTER	age	sex	income	married	children	car	savings	checking	mortgage	pep	inner_city	town	rural	suburban
1	43.01	0.49	30027.61	0.64	1.23	0.52	0.73	0.75	0.29	0.48	0.00	0.00	1.00	0.00
2	41.99	0.49	26844.00	0.66	0.95	0.48	0.64	0.76	0.35	0.46	1.00	0.00	0.00	0.00
3	42.20	0.53	26786.35	0.66	1.00	0.53	0.74	0.74	0.38	0.41	0.00	1.00	0.00	0.00
4	43.73	0.48	28656.36	0.68	0.97	0.40	0.69	0.81	0.35	0.55	0.00	0.00	0.00	1.00

Observations

- According to the dendrogram above, we could clearly detect 4 clusters. Also, similar to the results using hierarchical complete and average linkage distances, the clusters are mainly divided by their region. In details, CLUSTER 1 only contains customers from rural, CLUSTER 2 is from inner city, CLUSTER 3 from town, and CLUSTER 4 comes from suburban.

- As for the other characteristics of each cluster, CLUSTER 1 has highest average income, average number of children, lowest proportion of being married and owning mortgage. The customers have average level in the other measurements.

- CLUSTER 2 has the lowest average age, average number of children, proportion of having saving accounts; maintains average level in the other characteristics, including proportion of female, average income, proportion of being married, proportion of owning car, having checking accounts, having personal equity plan, etc.

- CLUSTER 3 has the highest proportion of female, owning car, having saving accounts; lowest average income, and proportion of having checking accounts; maintains average level in the other characteristics.

- CLUSTER 4 has highest average age, being married, having checking accounts, and having personal equity plan; however, has lowest proportion of female and owning car. Besides, the customers from this cluster maintain average level in other aspects.

- Overall, the four clusters tend to have higher similarity inside each cluster but have dissimilarity across different clusters.

- Moreover, the clusters generated based on 4 k-mean clusters are totally same as the clusters generated using hierarchical complete and average linkage distances, by ignoring the cluster order.

8. k-means with 5 Clusters

Cluster number with the numbers of customers in each cluster:

CLUSTER NUMBER	COUNT
1	96
2	146
3	62
4	123
5	173

The average value for each column by cluster is as follows:

CLUSTER	age	sex	income	married	children	car	savings	checking	mortgage	pep	inner_city	town	rural	suburban
1	43.01	0.49	30027.61	0.64	1.23	0.52	0.73	0.75	0.29	0.48	0.00	0.00	1.00	0.00
2	40.12	0.53	25030.38	0.77	1.06	0.49	0.68	0.79	0.34	0.00	1.00	0.00	0.00	0.00
3	43.73	0.48	28656.36	0.68	0.97	0.40	0.69	0.81	0.35	0.55	0.00	0.00	0.00	1.00
4	44.21	0.44	28996.76	0.54	0.82	0.48	0.59	0.73	0.36	1.00	1.00	0.00	0.00	0.00
5	42.20	0.53	26786.35	0.66	1.00	0.53	0.74	0.74	0.38	0.41	0.00	1.00	0.00	0.00

Observations

- When we set the number of k-mean clusters as 5, according to the count in each cluster and the characteristics, we could detect the big cluster with customers coming from inner city are divided into 2 clusters --CLUSTER 2 and 4.

- When we compare these two clusters, we could detect CLUSTER 2 has highest proportion of being married, but lowest average age and income, and 0% of people having personal equity plan compared to the other 4 clusters.

- CLUSTER 4 has highest average age and second highest average income; also, 100% customers have personal equity plan. However, they rank lowest in the proportion of

female, being married, average number of children, proportion of having saving accounts and checking accounts.

- Besides, CLUSTER 1 comes from rural, ranks highest in average income, average number of children, but has lowest proportion of owning mortgage. The customers have average level in the other measurements.

- CLUSTER 3 ranks highest in average age and having checking accounts; however, has lowest proportion of owning car. Besides, the customers from this cluster maintain average level in other aspects.

- CLUSTER 5 comes from town, has the highest proportion of female, owning car, having saving accounts; while ranks second lowest in average income. This cluster maintains average level in the other characteristics.

- Overall, the five clusters tend to have higher similarity inside each cluster but have dissimilarity across different clusters.

- Moreover, the clusters generated based on 5 k-mean clusters could clearly divide all customers into as many as clusters as possible while keep the characteristics from each cluster are quite distinguishable.

9. k-means with 6 Clusters

Cluster number with the numbers of customers in each cluster:

CLUSTER NUMBER	COUNT
1	96
2	146
3	81
4	62
5	92
6	123

The average value for each column by cluster is as follows:

CLUSTER	age	sex	income	married	children	car	savings	checking	mortgage	pep	inner_city	town	rural	suburban
1	43.01	0.49	30027.61	0.64	1.23	0.52	0.73	0.75	0.29	0.48	0.00	0.00	1.00	0.00
2	40.12	0.53	25030.38	0.77	1.06	0.49	0.68	0.79	0.34	0.00	1.00	0.00	0.00	0.00
3	41.56	0.00	26849.57	0.59	1.01	0.57	0.75	0.80	0.41	0.41	0.00	1.00	0.00	0.00
4	43.73	0.48	28656.36	0.68	0.97	0.40	0.69	0.81	0.35	0.55	0.00	0.00	0.00	1.00
5	42.77	1.00	26730.69	0.73	0.99	0.49	0.73	0.68	0.35	0.41	0.00	1.00	0.00	0.00
6	44.21	0.44	28996.76	0.54	0.82	0.48	0.59	0.73	0.36	1.00	1.00	0.00	0.00	0.00

Observations

When we set the number of k-mean clusters as 6, the big cluster with customers from town is divided into 2 clusters -- CLUSTER 3 and 5. According to the characteristics of these two clusters, except for CLUSTER 3 only containing male customers while CLUSTER 5 only containing female customers, the other characteristics between these two clusters are not that distinguishable.

10. k-means with 7 Clusters

Cluster number with the numbers of customers in each cluster:

CLUSTER NUMBER	COUNT
1	96
2	87
3	110
4	92
5	72
6	81
7	62

The average value for each column by cluster is as follows:

CLUSTER	age	sex	income	married	children	car	savings	checking	mortgage	pep	inner_city	town	rural	suburban
1	43.01	0.49	30027.61	0.64	1.23	0.52	0.73	0.75	0.29	0.48	0.00	0.00	1.00	0.00
2	38.16	0.13	23483.53	0.86	0.93	0.15	0.40	0.68	0.44	0.38	1.00	0.00	0.00	0.00
3	42.65	0.69	27885.27	0.86	0.99	0.68	0.95	0.83	0.34	0.23	1.00	0.00	0.00	0.00
4	42.77	1.00	26730.69	0.73	0.99	0.49	0.73	0.68	0.35	0.41	0.00	1.00	0.00	0.00
5	45.61	0.61	29313.74	0.11	0.92	0.58	0.46	0.76	0.26	0.90	1.00	0.00	0.00	0.00
6	41.56	0.00	26849.57	0.59	1.01	0.57	0.75	0.80	0.41	0.41	0.00	1.00	0.00	0.00
7	43.73	0.48	28656.36	0.68	0.97	0.40	0.69	0.81	0.35	0.55	0.00	0.00	0.00	1.00

Observations

- When we set the number of k-mean clusters as 7, the big cluster with customers from inner city is divided into 3 clusters -- CLUSTER 2, 3 and 5. Besides, the big cluster with customers from town is divided into 2 clusters -- CLUSTER 4 and 6.
- Also, according to the characteristics of CLUSTER 4 and 6, except for the differences in gender, the other characteristics between these two clusters are not that distinguishable.

11. k-means with 8 Clusters

Cluster number with the numbers of customers in each cluster:

CLUSTER NUMBER	COUNT
1	75
2	69
3	100
4	67
5	62
6	96
7	58
8	73

The average value for each column by cluster is as follows:

CLUSTER	age	sex	income	married	children	car	savings	checking	mortgage	pep	inner_city	town	rural	suburban
1	40.08	0.81	24152.88	0.93	0.83	0.23	0.71	0.91	0.11	0.11	1.00	0.00	0.00	0.00
2	45.83	0.58	29652.02	0.04	1.01	0.58	0.48	0.75	0.19	0.83	1.00	0.00	0.00	0.00
3	43.50	0.69	27117.70	0.98	0.98	0.50	0.77	0.63	0.27	0.36	0.00	1.00	0.00	0.00
4	44.28	0.39	29785.04	0.87	1.13	1.00	0.84	0.78	0.48	0.31	1.00	0.00	0.00	0.00
5	43.73	0.48	28656.36	0.68	0.97	0.40	0.69	0.81	0.35	0.55	0.00	0.00	0.00	1.00
6	43.01	0.49	30027.61	0.64	1.23	0.52	0.73	0.75	0.29	0.48	0.00	0.00	1.00	0.00
7	37.26	0.07	23585.91	0.81	0.83	0.10	0.53	0.57	0.71	0.64	1.00	0.00	0.00	0.00
8	40.42	0.32	26332.44	0.23	1.03	0.56	0.70	0.89	0.52	0.48	0.00	1.00	0.00	0.00

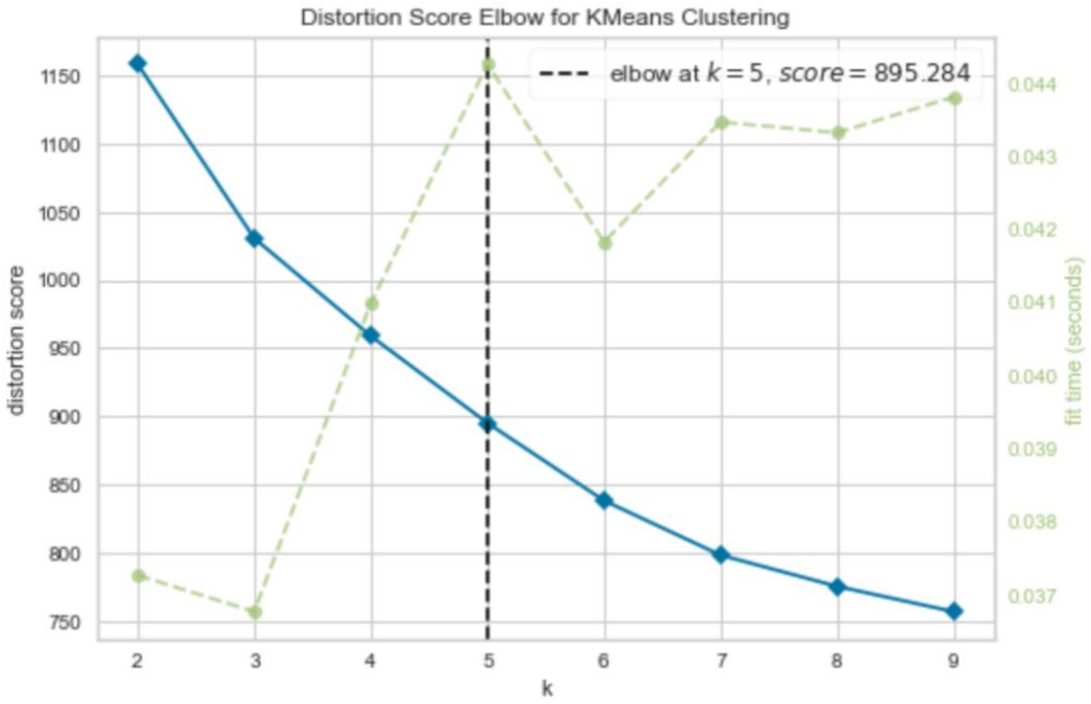
Observations

- When we set the number of k-mean clusters as 8, the big cluster with customers from inner city is divided into 4 clusters – CLUSTER1, 2, 4 and 7. According to the characteristics of these four clusters, these clusters are not so distinguishable.

- Besides, the big cluster with customers from town is divided into 2 clusters -- CLUSTER 4 and 6. According to the characteristics of CLUSTER 4 and 6, except for the differences in gender, the other characteristics between these two clusters are not that distinguishable.

12. Conclusion of k-means clustering:

The K-mean elbow plot is as follows:



According to the plot, we see a smooth curve and the optimal value of K is unclear, because the data is not very clustered. But 5 is the suggested optimal number of clusters.

Besides, based on all the discussion above, 5 is the optimal choice to divide bank customers into different clusters. The better solution according to k-mean clusters is consistent with the optimal choice based on hierarchical clustering. Although the results are slightly different in the number of customers falling into Inner City I and Inner City II, and the values of different columns for these two clusters. The number of customers and their characteristics are totally same for the other three clusters.

The Count of Customers Falling into Different Clusters:

CLUSTER NUMBER	HIERARCHICAL CLUSTERING	K-MEANS CLUSTERING
Inner City I	124	123
Inner City II	145	146
Town	173	173
Suburban	62	62
Rural	96	96

The Value for the Characteristics of the Clusters:

Hierarchical Clustering:

CLUSTER	age	sex	income	married	children	car	savings	checking	mortgage	pep	Inner_city	town	rural	suburban
Inner City I	43.97	0.48	28677.59	0.46	0.99	0.53	0.57	0.77	0.29	0.91	1.00	0.00	0.00	0.00
Inner City II	40.30	0.50	25275.96	0.83	0.92	0.44	0.70	0.76	0.40	0.07	1.00	0.00	0.00	0.00
Town	42.20	0.53	26786.35	0.66	1.00	0.53	0.74	0.74	0.38	0.41	0.00	1.00	0.00	0.00
Suburban	43.73	0.48	28656.36	0.68	0.97	0.40	0.69	0.81	0.35	0.55	0.00	0.00	0.00	1.00
Rural	43.01	0.49	30027.61	0.64	1.23	0.52	0.73	0.75	0.29	0.48	0.00	0.00	1.00	0.00

K-means Clustering:

CLUSTER	age	sex	income	married	children	car	savings	checking	mortgage	pep	Inner_city	town	rural	suburban
Inner City I	44.21	0.44	28996.76	0.54	0.82	0.48	0.59	0.73	0.36	1.00	1.00	0.00	0.00	0.00
Inner City II	40.12	0.53	25030.38	0.77	1.06	0.49	0.68	0.79	0.34	0.00	1.00	0.00	0.00	0.00
Town	42.20	0.53	26786.35	0.66	1.00	0.53	0.74	0.74	0.38	0.41	0.00	1.00	0.00	0.00
Suburban	43.73	0.48	28656.36	0.68	0.97	0.40	0.69	0.81	0.35	0.55	0.00	0.00	0.00	1.00
Rural	43.01	0.49	30027.61	0.64	1.23	0.52	0.73	0.75	0.29	0.48	0.00	0.00	1.00	0.00

13. Managerial takeaways

- As a bank manager, I would like to use 5 clusters to divide my current customers into five categories since all hierarchical, k-mean clusters and the elbow plot show us a division by 5 could make the clusters more distinguishable.

- Then I will label them according to their characteristics and give different categories different customed promotion to increase our bank revenues.

- For Inner City I customers, I will label them as "Investment Oriented Inner-City Seniors" since they have highest average age and second highest average income; also, 90% - 100% customers have personal equity plan. However, they rank lowest in the proportion of female, being married, proportion of having saving accounts. I will promote them with

rewards when they purchase bank financial products or involve into any investment plans provided by my bank.

- For Inner City II customers, I will label them as "Young but Stable Inner-City Married Adults" since they have highest proportion of being married, but lowest average age and income, and only 0-10% of people having personal equity plan compared to the other 4 clusters. I will promote them opening saving accounts, randomly promote the other types of promotions time to time since maintains average level in the other characteristics.

- For Town customers, they have the highest proportion of female, owning car, having saving accounts; while ranks second lowest in average income. This cluster maintains average level in the other characteristics. I with label them as "Town Lady with Limited Purchase Power", and I will mainly promote them opening checking accounts, and send them the other promotions time to time.

- For Suburban customers, they rank highest in proportion of having checking accounts and second highest in average age; however, have lowest proportion of owning car. Besides, the customers from this cluster maintain average level in other aspects. I will label them as "Elderly Suburban People", I will promote them when they buy cars, open more checking accounts, and purchase bank financial products.

- For Rural customers, they rank highest in average income, average number of children, but has lowest proportion of owning mortgage. The customers have average level in the other measurements. I will label them as "Affluent Rural Family", I with mainly promote them with support or discount when they try to upgrade their car, mortgage, or have need of education loan, etc.