Case Study #5 - Cluster Analysis

1. Upload, explore, clean, and preprocess data for clustering.
   a. Dimensions and first 10 records.

```
The dimensions of the rates_df is (50, 8). 50 rows and 8 columns.
```

```
rates_df.head(10)
```

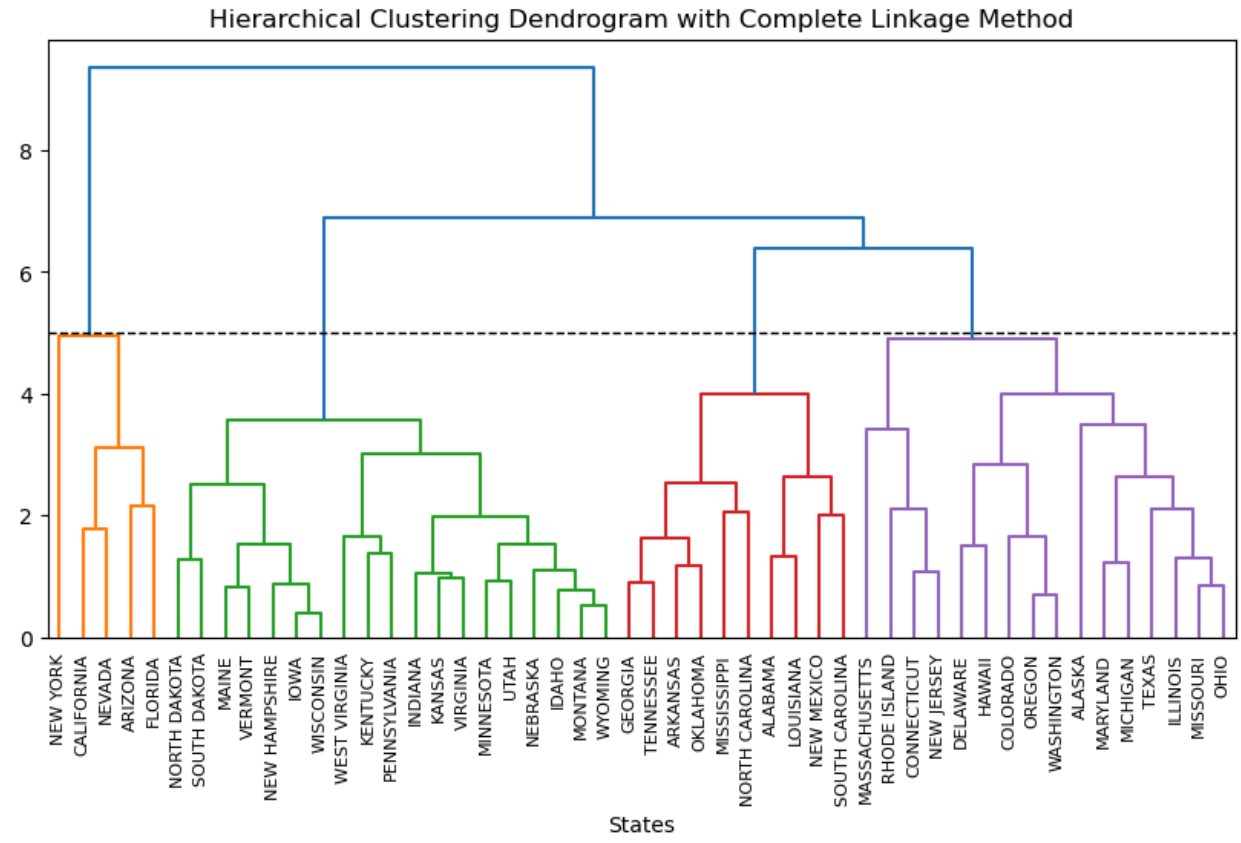
	States	murder	rape	robbery	assault	burglary	larceny	auto
0	ALABAMA	14.2	25.2	96.8	278.3	1135.5	1881.9	280.7
1	ALASKA	10.8	51.6	96.8	284.0	1331.7	3369.8	753.3
2	ARIZONA	9.5	34.2	138.2	312.3	2346.1	4467.4	439.5
3	ARKANSAS	8.8	27.6	83.2	203.4	972.6	1862.1	183.4
4	CALIFORNIA	11.5	49.4	287.0	358.0	2139.4	3499.8	663.5
5	COLORADO	6.3	42.0	170.7	292.9	1935.2	3903.2	477.1
6	CONNECTICUT	4.2	16.8	129.5	131.8	1346.0	2620.7	593.2
7	DELAWARE	6.0	24.9	157.0	194.2	1682.6	3678.4	467.0
8	FLORIDA	10.2	39.6	187.9	449.1	1859.9	3840.5	351.4
9	GEORGIA	11.7	31.1	140.5	256.5	1351.1	2170.2	297.9

   b. The reason why we normalize data in clustering is because raw distance measures highly influence clustering algorithms. If we do not normalize the data prior to training then higher values influence clustering in a manner that would produce inaccurate results. Normalized data also speeds up the clustering process. The normalization of data is done through z-scores of the data.

```
Normalized Input Variables for Five Utilities
```

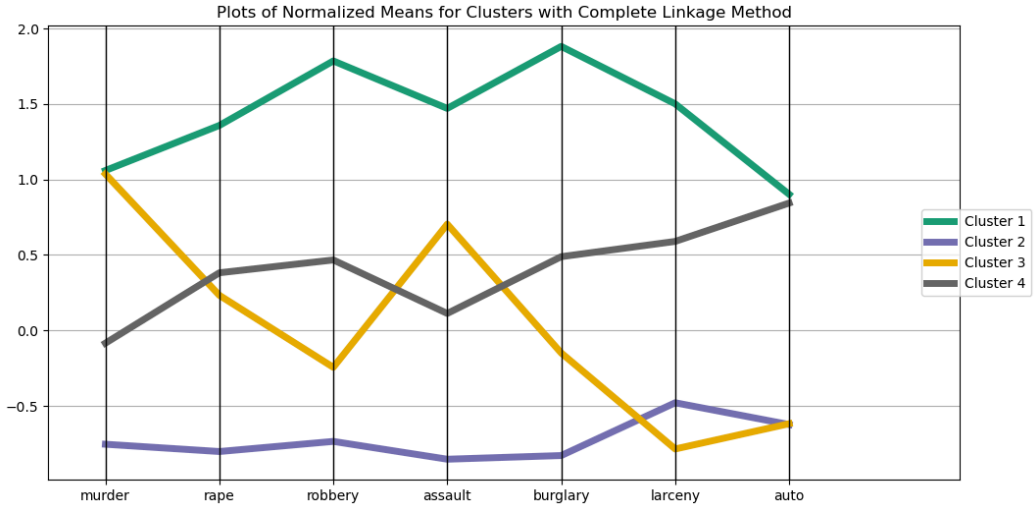
States	murder	rape	robbery	assault	burglary	larceny	auto
ALABAMA	1.747195	-0.049630	-0.308913	0.668309	-0.361665	-1.087448	-0.500666
ALASKA	0.867908	2.403986	-0.308913	0.725165	0.092023	0.962259	1.943045
ARIZONA	0.531710	0.786830	0.159686	1.007451	2.437697	2.474295	0.320454
ARKANSAS	0.350680	0.173426	-0.462848	-0.078801	-0.738351	-1.114724	-1.003783
CALIFORNIA	1.048938	2.199518	1.843924	1.463297	1.959729	1.141345	1.478709
COLORADO	-0.295854	1.511762	0.527547	0.813940	1.487542	1.697062	0.514875
CONNECTICUT	-0.838943	-0.830326	0.061212	-0.792993	0.125090	-0.069689	1.115203
DELAWARE	-0.373438	-0.077512	0.372479	-0.170568	0.903436	1.387381	0.462650
FLORIDA	0.712740	1.288706	0.722230	2.371998	1.313420	1.610687	-0.135092
GEORGIA	1.100661	0.498716	0.185719	0.450859	0.136883	-0.690291	-0.411729

2. Apply hierarchical clustering to classify the states into clusters based on the normalized crime data.
   a. There are four clusters in the dendrogram below. See cluster membership for list of states.




```
Cluster Membership for 4 Clusters Using Average Linkage Method
1  :  ARIZONA , CALIFORNIA , FLORIDA , NEVADA , NEW YORK
2  :  IDAHO , INDIANA , IOWA , KANSAS , KENTUCKY , MAINE , MINNESOTA , MONTANA , NEBRASKA , NEW HAMPSHIRE , NORTH DAKOTA , PENN
SYLVANIA , SOUTH DAKOTA , UTAH , VERMONT , VIRGINIA , WEST VIRGINIA , WISCONSIN , WYOMING
3  :  ALABAMA , ARKANSAS , GEORGIA , LOUISIANA , MISSISSIPPI , NEW MEXICO , NORTH CAROLINA , OKLAHOMA , SOUTH CAROLINA , TENNES
SEE
4  :  ALASKA , COLORADO , CONNECTICUT , DELAWARE , HAWAII , ILLINOIS , MARYLAND , MASSACHUSETTS , MICHIGAN , MISSOURI , NEW JER
SEY , OHIO , OREGON , RHODE ISLAND , TEXAS , WASHINGTON
```

b. Profile Plot of Normalized Means for each cluster. Respective cluster mean values displayed below in the dataframe



We can use mean values of each cluster to characterize how the cluster was formed. Clusters with higher or lower means for a variable shows that there is an emphasis on a certain feature for that particular cluster.

Normalized Means of Input Variables for Clusters with Complete Linkage Method

	murder	rape	robbery	assault	burglary	larceny	auto	Cluster
1	1.059	1.357	1.785	1.470	1.881	1.500	0.902	Cluster 1
2	-0.753	-0.801	-0.734	-0.852	-0.828	-0.479	-0.623	Cluster 2
3	1.036	0.233	-0.244	0.704	-0.149	-0.785	-0.617	Cluster 3
4	-0.084	0.382	0.467	0.113	0.489	0.590	0.844	Cluster 4

c.

Cluster 1 represents the following states: ARIZONA , CALIFORNIA , FLORIDA , NEVADA and NEW YORK which exhibit the highest crime for robbery and burglary but lowest for murder and auto incidents.

Cluster 2 represents the following states: IDAHO , INDIANA , IOWA , KANSAS , KENTUCKY , MAINE , MINNESOTA , MONTANA , NEBRASKA , NEW HAMPSHIRE , NORTH DAKOTA , PENNSYLVANIA , SOUTH DAKOTA , UTAH , VERMONT , VIRGINIA , WEST VIRGINIA , WISCONSIN and WYOMING which exhibit high larceny but low rates of violent crime.

Cluster 3 represents the following states: ALABAMA , ARKANSAS , GEORGIA , LOUISIANA , MISSISSIPPI , NEW MEXICO , NORTH CAROLINA , OKLAHOMA , SOUTH CAROLINA and TENNESSEE which exhibit high rates of violent crime but low rates of non-violent crime.

Cluster 4 represents the following states: ALASKA , COLORADO , CONNECTICUT , DELAWARE , HAWAII , ILLINOIS , MARYLAND , MASSACHUSETTS , MICHIGAN , MISSOURI , NEW JERSEY , OHIO , OREGON , RHODE ISLAND , TEXAS and WASHINGTON which exhibit lower non-violent crime but higher theft and auto crime.

3. Apply k-means clustering to classify the states into clusters based on the crime data.
   a. The two main differences between the algorithms used in hierarchical and k-means clustering:
   1) k-means requires a prior definition of how many clusters to be computed whereas hierarchical does not require a predefined number of clusters.

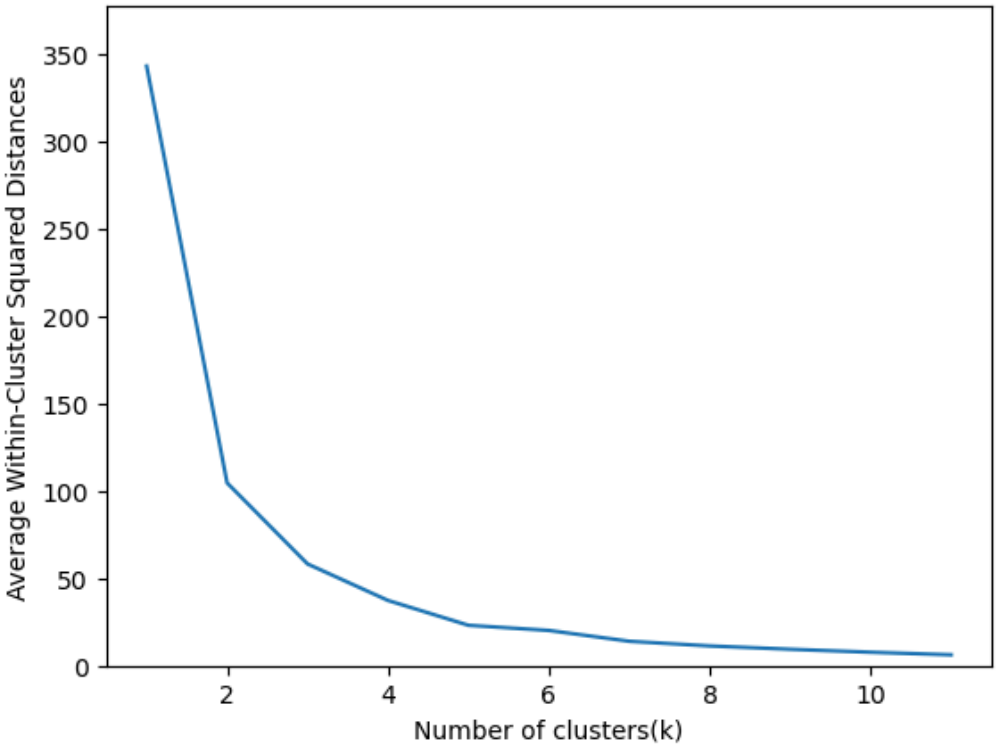
   2) hierarchical clustering is less stable and computationally expensive when using large data sets compared to k-means clustering.

   There are many other differences between these two but I saw these as the most important to consider before building either of the models.

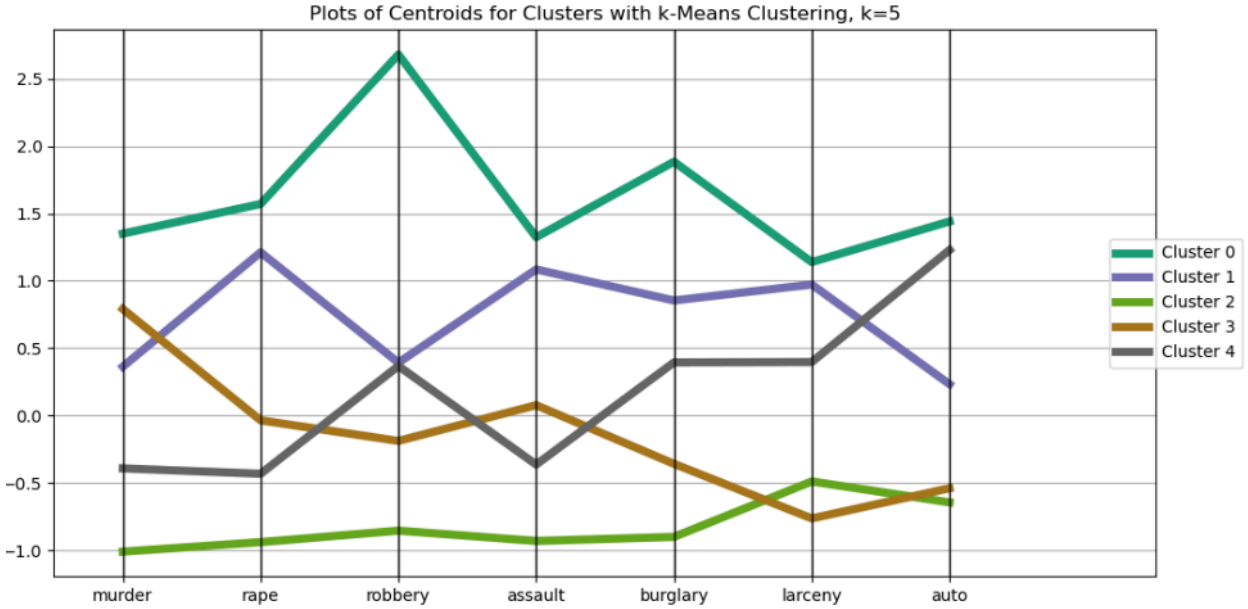
Cluster Membership for 5 Clusters Using k-Means Clustering

```
0 :  CALIFORNIA, NEVADA, NEW YORK
1 :  ALASKA, ARIZONA, COLORADO, FLORIDA, MARYLAND, MICHIGAN, NEW MEXICO, OREGON, SOUTH CAROLINA, TEXAS, WASHINGTON
2 :  IDAHO, IOWA, MAINE, MINNESOTA, MONTANA, NEBRASKA, NEW HAMPSHIRE, NORTH DAKOTA, PENNSYLVANIA, SOUTH DAKOTA, UTAH, VERMONT,
WEST VIRGINIA, WISCONSIN, WYOMING
3 :  ALABAMA, ARKANSAS, GEORGIA, INDIANA, KANSAS, KENTUCKY, LOUISIANA, MISSISSIPPI, MISSOURI, NORTH CAROLINA, OKLAHOMA, TENNESS
EE, VIRGINIA
4 :  CONNECTICUT, DELAWARE, HAWAII, ILLINOIS, MASSACHUSETTS, NEW JERSEY, OHIO, RHODE ISLAND
```

b. k=5 is appropriate but k=7 is worth trying out. Seven clusters will likely make it harder to explain each cluster so we will keep k=5.



c. In k-means clustering, clusters can be characterized by their respective centroids since centroids represent the mean of all data points within each cluster, serving as a central representative point. Profile Plot of centroids for each cluster. Respective cluster centroid values displayed below in the dataframe



Cluster Centroids for k-Means Clustering with k = 5

	murder	rape	robbery	assault	burglary	larceny	auto	Cluster
0	1.351	1.571	2.680	1.324	1.884	1.139	1.441	Cluster 0
1	0.362	1.212	0.392	1.085	0.854	0.973	0.232	Cluster 1
2	-1.011	-0.941	-0.855	-0.932	-0.902	-0.490	-0.646	Cluster 2
3	0.790	-0.036	-0.189	0.076	-0.359	-0.765	-0.540	Cluster 3
4	-0.393	-0.433	0.367	-0.365	0.394	0.397	1.229	Cluster 4

d.

Cluster 0 represents the following states: CALIFORNIA, NEVADA and NEW YORK which experience the highest rates of robbery and burglary but low rates of violent crime.

Cluster 1 represents the following states: ALASKA, ARIZONA, COLORADO, FLORIDA, MARYLAND, MICHIGAN, NEW MEXICO, OREGON, SOUTH CAROLINA, TEXAS and WASHINGTON which experience high rates of some violent crimes but lower rates of some non-violent crimes.

Cluster 2 represents the following states: IDAHO, IOWA, MAINE, MINNESOTA, MONTANA, NEBRASKA, NEW HAMPSHIRE, NORTH DAKOTA, PENNSYLVANIA, SOUTH DAKOTA, UTAH, VERMONT, WEST VIRGINIA, WISCONSIN and WYOMING which experience higher rates of non-violent crimes than violent crime.

Cluster 3 represents the following states: ALABAMA, ARKANSAS, GEORGIA, INDIANA, KANSAS, KENTUCKY, LOUISIANA, MISSISSIPPI, MISSOURI, NORTH CAROLINA, OKLAHOMA, TENNESSEE and VIRGINIA which experience high rates of violent crime and low rates of non-violent.

Cluster 4 represents the following states: CONNECTICUT, DELAWARE, HAWAII, ILLINOIS, MASSACHUSETTS, NEW JERSEY, OHIO and RHODE ISLAND which experience high rates of non-violent crime compared to violent crime.

4. Compare the clusters from parts 2 and 3. From your standpoint, which clustering, hierarchical or k-means, provides more useful insights of the states' crime rates.

K-means clustering is the winner in my eyes. The main reason I choose k-means over hierarchical clustering is because the separation between centroids is well fragmented making it easier to interpret differences between clusters. High and low values are easy to identify right away whereas in hierarchical clustering, the values of means lie in a smaller range of values making it harder to quickly scan through. This makes sense since the value of k in k-means was optimized/visualized in the elbow chart to produce the best separation without making the model too complex.