Report

Case Study #5: Crime Rate

Name: Prathyush Kaparthi
Net id: bs9845

Question 1:

a)

```
Out[52]: (50, 7)
```

Out[53]:

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

The data frame is created by uploading the original data set. There are total of 50 rows and 7 columns and also the first 10 records are displayed.

b)

Normalized Input Variables

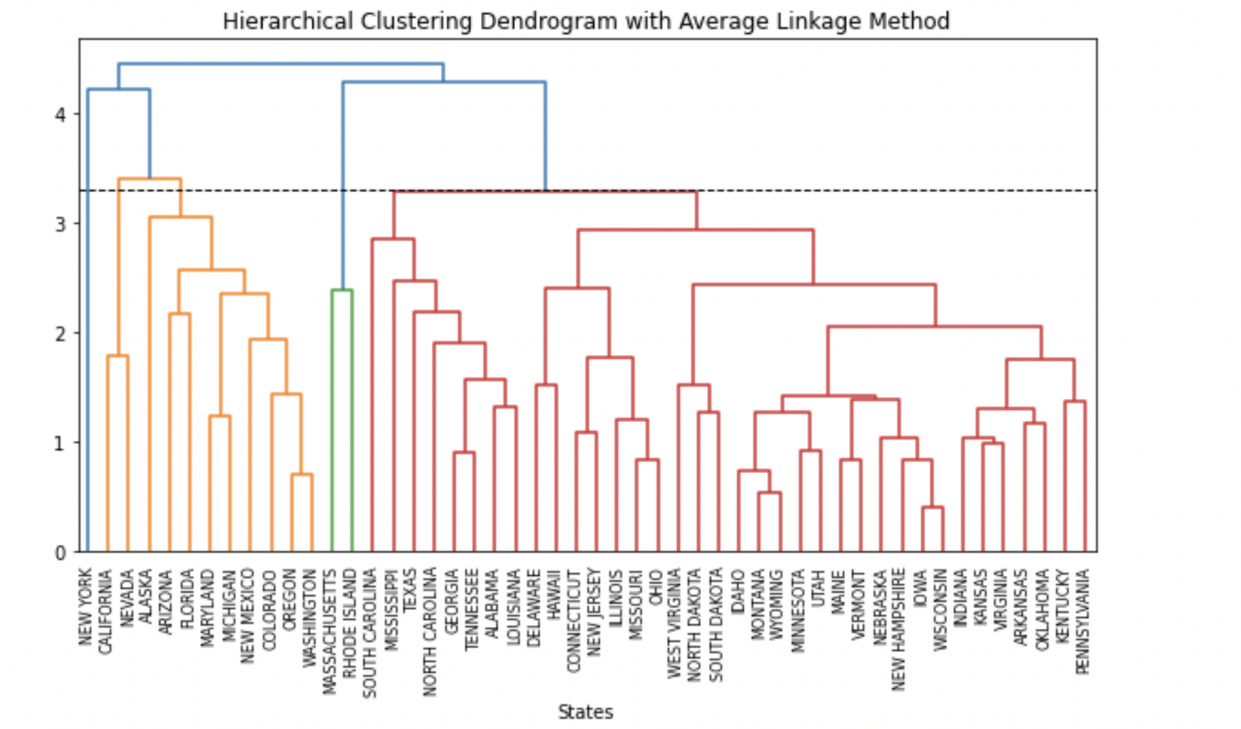
Out[54]:

States	murder	rape	robbery	assault	burglary	larceny	auto
ALABAMA	1.75	-0.05	-0.31	0.67	-0.36	-1.09	-0.50
ALASKA	0.87	2.40	-0.31	0.73	0.09	0.96	1.94
ARIZONA	0.53	0.79	0.16	1.01	2.44	2.47	0.32
ARKANSAS	0.35	0.17	-0.46	-0.08	-0.74	-1.11	-1.00
CALIFORNIA	1.05	2.20	1.84	1.46	1.96	1.14	1.48
COLORADO	-0.30	1.51	0.53	0.81	1.49	1.70	0.51
CONNECTICUT	-0.84	-0.83	0.06	-0.79	0.13	-0.07	1.12
DELAWARE	-0.37	-0.08	0.37	-0.17	0.90	1.39	0.46
FLORIDA	0.71	1.29	0.72	2.37	1.31	1.61	-0.14
GEORGIA	1.10	0.50	0.19	0.45	0.14	-0.69	-0.41

The scale of measurements can significantly affect raw distance measures, but this issue can be resolved through data normalization, which involves calculating Z-scores. Normalized values are preferred in clustering as they provide a consistent scale for all variables. The process of normalization involves subtracting the mean and dividing by the standard deviation.

Question 2:

a)



There are 6 clusters Classified on the dendrogram, they are

Cluster Membership for 5 Clusters Using Average Linkage Method
1 : CALIFORNIA , NEVADA
2 : ALASKA , ARIZONA , COLORADO , FLORIDA , MARYLAND , MICHIGAN , NEW MEXICO , OREGON , WASHINGTON
3 : NEW YORK
4 : MASSACHUSETTS , RHODE ISLAND
5 : ALABAMA , ARKANSAS , CONNECTICUT , DELAWARE , GEORGIA , HAWAII , IDAHO , ILLINOIS , INDIANA , IOWA , KANSAS , KENTUCKY , LOUISIANA , MAINE , MINNESOTA , MISSISSIPPI , MISSOURI , MONTANA , NEBRASKA , NEW HAMPSHIRE , NEW JERSEY , NORTH CAROLINA , NORTH DAKOTA , OHIO , OKLAHOMA , PENNSYLVANIA , SOUTH CAROLINA , SOUTH DAKOTA , TENNESSEE , TEXAS , UTAH , VERMONT , VIRGINIA , WEST VIRGINIA , WISCONSIN , WYOMING

In [7]:
```
# Develop cluster membership for agglomerative clustering using average
# linkage method. The number of clusters is assigned to be 6 as shown
# in the dendrogram with average linkage.
memb_ave = fcluster(hi_average, 5, criterion='maxclust')
memb_ave = pd.Series(memb_ave, index=df_norm.index)

# Display cluster memberships for 5 clusters.
print('Cluster Membership for 5 Clusters Using Average Linkage Method')
for key, item in memb_ave.groupby(memb_ave):
    print(key, ' : ',' , '.join(item.index))
```

```
Cluster Membership for 5 Clusters Using Average Linkage Method
1  :  CALIFORNIA , NEVADA
2  :  ALASKA , ARIZONA , COLORADO , FLORIDA , MARYLAND , MICHIGAN , NEW MEXICO , OREGON , WASHINGTON
3  :  NEW YORK
4  :  MASSACHUSETTS , RHODE ISLAND
5  :  ALABAMA , ARKANSAS , CONNECTICUT , DELAWARE , GEORGIA , HAWAII , IDAHO , ILLINOIS , INDIANA , IOWA , KANSAS ,
KENTUCKY , LOUISIANA , MAINE , MINNESOTA , MISSISSIPPI , MISSOURI , MONTANA , NEBRASKA , NEW HAMPSHIRE , NEW JERSEY
, NORTH CAROLINA , NORTH DAKOTA , OHIO , OKLAHOMA , PENNSYLVANIA , SOUTH CAROLINA , SOUTH DAKOTA , TENNESSEE , TEXA
S , UTAH , VERMONT , VIRGINIA , WEST VIRGINIA , WISCONSIN , WYOMING
```

b)

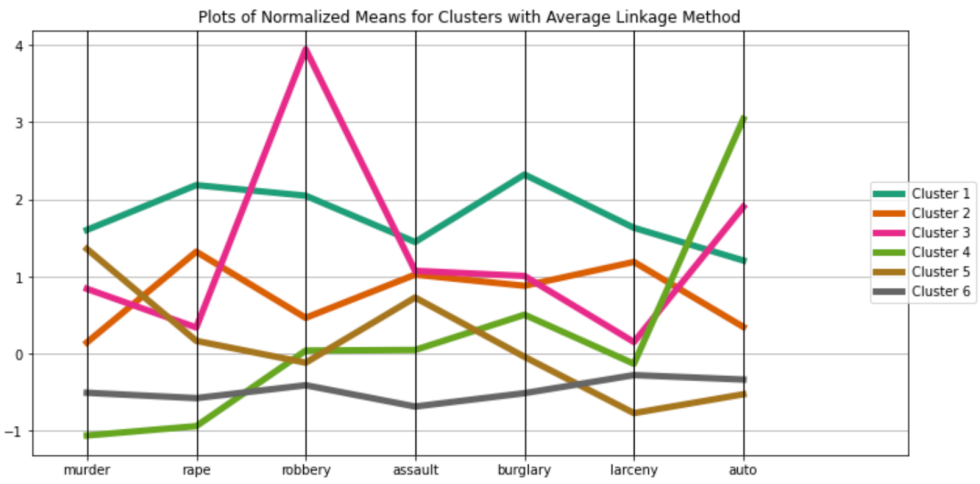
In [8]:
```
# 2) b

# Identify  and display cluster normalized mean values
# for each of 8 input variables (measurements).

# Create data frame with normalized cluster means for each
# cluster and each input variable (measurement).
clust_mean_norm = df_norm.groupby(memb_ave).mean()

# Add cluster titles (Cluster 1, Cluster 2, ...) to the
# cluster_mean_norm data frame with means and get precision
# of 3 decimals.
clust_mean_norm['Cluster'] = ['Cluster {}'.format(i) for i in clust_mean_norm.index]

# Use display.precision to reduce the number of decimals to 3.
pd.set_option('display.precision', 3)

# Display the data frame with normalized mean values and cluster titles.
print('Normalized Means of Input Variables for Clusters with Average Linkage Method')
clust_mean_norm
```

```
Normalized Means of Input Variables for Clusters with Average Linkage Method
```

Out[8]:

	murder	rape	robbery	assault	burglary	larceny	auto	Cluster
1	1.605	2.186	2.048	1.448	2.322	1.632	1.209	Cluster 1
2	0.147	1.323	0.466	1.025	0.881	1.190	0.348	Cluster 2
3	0.842	0.341	3.945	1.075	1.008	0.153	1.904	Cluster 3
4	-1.059	-0.937	0.042	0.050	0.506	-0.129	3.042	Cluster 4
5	-0.090	-0.410	-0.342	-0.369	-0.405	-0.385	-0.376	Cluster 5

The clusters can be described based on the amount of crime they have. For example, Cluster 0 has the lowest rates of murder, rape, robbery, assault, burglary, and auto crimes, with cluster centroid values of -1.011 for murder, -0.941 for rape, -0.855 for assault, -0.902 for burglary, and -0.646 for auto crimes. Cluster 1, in contrast, has the highest rates of all crimes, including murder, rape, robbery, assault, burglary, and larceny. Clusters 2 and 3 both have medium levels of crime. Cluster 4 has the lowest rates of murder but the highest rates of auto crimes, with a cluster centroid value of -0.985 for murder and 2.399 for auto crimes. Finally, Cluster 5 has medium rates of murder, rape, robbery, assault, and burglary, but has the lowest rates of larceny and auto crimes, with cluster centroid values of -0.800 for larceny and -0.586 for auto crimes.



c)

Cluster 1 - largest number of murder , rape, assault, burglary and larceny crimes
Cluster 2- medium crimes (murder, robbery, assault, burglary etc.)
Cluster 3- largest number of robbery crimes
Cluster 4- lowest murder crimes and largest auto crimes
Cluster 5- medium number of murder crimes and lowest number of (robbery, assault, burglary, larceny and auto crimes)

Question 3:

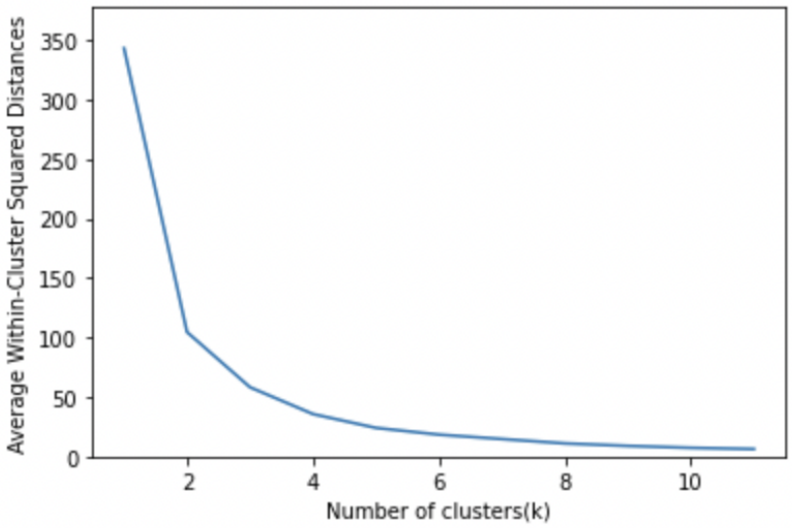
a)

```
Cluster Membership for 6 Clusters Using k-Means Clustering

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

Hierarchical clustering using the agglomerative method begins with n clusters (one for each observation) and combines similar clusters sequentially until a single cluster is formed. In contrast, the divisive method starts with a single cluster that encompasses all observations and divides it into smaller, more distinct clusters.

With k-means clustering, the process starts by selecting k initial clusters, often based on randomly chosen k centroids. At each step, each record is assigned to the cluster with the closest centroid. Then, the centroids of clusters that have lost or gained a record are recomputed, and the process repeats. The algorithm stops when moving records between clusters increases the within-cluster dispersion.

b)



When using k-means clustering on crime data, increasing the number of clusters results in cluster members being closer to each other. However, once the number of clusters exceeds 6, the improvement in cluster homogeneity becomes less significant. Therefore, it is appropriate to set k=6 as the number of clusters in the k-means clustering of the crime data.

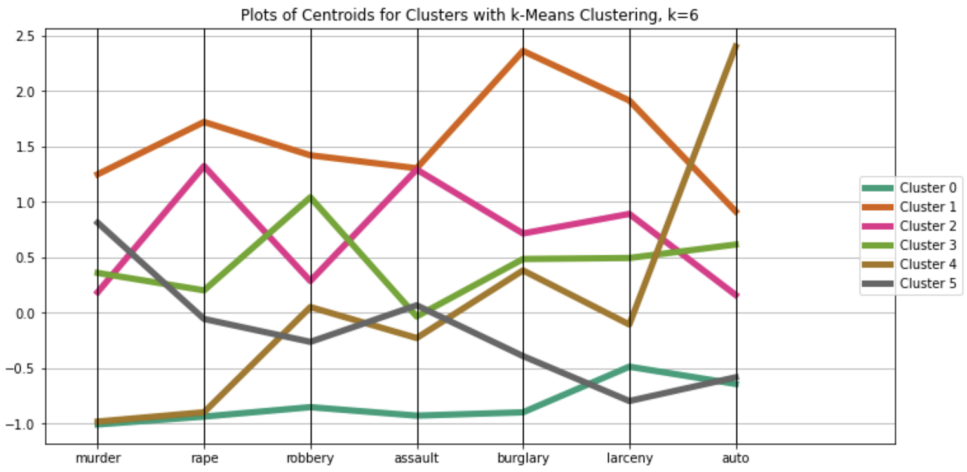
c)

Out[61]:

Cluster Centroids for k-Means Clustering with k = 6

	murder	rape	robbery	assault	burglary	larceny	auto	Cluster
0	-1.011	-0.941	-0.855	-0.932	-0.902	-0.490	-0.646	Cluster 0
1	1.247	1.719	1.419	1.301	2.361	1.913	0.913	Cluster 1
2	0.183	1.321	0.284	1.290	0.713	0.889	0.158	Cluster 2
3	0.356	0.199	1.039	-0.036	0.482	0.491	0.613	Cluster 3
4	-0.985	-0.902	0.048	-0.231	0.379	-0.109	2.399	Cluster 4
5	0.810	-0.059	-0.266	0.064	-0.394	-0.800	-0.586	Cluster 5

In the crime data clustering, Cluster 0 has the fewest instances of murder and auto crimes and has a cluster centroid value of -1.011. On the other hand, Cluster 1 has the highest occurrences of crimes, including murder, rape, robbery, assault, burglary, and larceny. Cluster 2 is characterized by a moderate number of crimes, while Cluster 3 is similar to Cluster 2, also having a medium number of crimes. Cluster 4 has the lowest number of murder crimes but has the highest number of auto crimes, with a cluster centroid value of -0.985 for murder crimes and 2.399 for auto crimes. Finally, Cluster 5 has a moderate number of crimes, including murder, rape, robbery, assault, and burglary. However, it has the lowest occurrences of larceny and auto crimes, with a cluster centroid value of -0.800 for larceny and -0.586 for auto crimes.



d)

Cluster 0 : lowest murder and auto crimes
Cluster 1 : highest murder , rape ,robbery, assault , burglary and larceny crimes
Cluster 2 : medium number of crimes
Cluster 3 : medium number of crimes
Cluster 4 : lowest murder crimes and highest auto crimes
Cluster 5 : medium murder,rape ,robbery,assault and burglary crimes and lowest larceny and auto crimes.

Question 4:

After analyzing the clusters generated in parts 2 and 3, it is evident that both clustering techniques offer valuable information regarding the crime rates in various states. Hierarchical clustering produced dendrograms that grouped states based on their crime rates and provided a visual representation of their similarities. In contrast, k-means clustering grouped the states into five clusters based on crime rates, which offered a quantitative and objective approach to clustering. While both methods are useful, I personally favor hierarchical clustering because it offers a visual representation of the dendrogram, which aids in identifying the optimal number of clusters. Additionally, it is robust in handling outliers and can identify nested clusters.