CUNY Baruch College

**Assignment 5: Exploratory Data Analysis**

Group Submission

Yula Ko

Farzana Manjra

Natalia Gomez

Victoria Wayda

JiaRui (Jessica) Shao

STA 3920 NFA

Yuanfeng Cai

28 November 2018

**1. Using hierarchical clustering with average linkage and Euclidean distance, cluster the states. Cut the dendrogram to obtain four clusters. List the states in each cluster.**

Cluster 1{Alabama, Alaska, Arizona, California, Delaware, Illinois, Louisiana, Maryland, Michigan, Mississippi, Nevada, New Mexico, New York, South Carolina,}

Cluster 2 {Arkansas, Colorado, Georgia, Massachusetts, Missouri, New Jersey, Oklahoma, Oregon, Rhode Island, Tennessee, Texas, Virginia, Washington, Wyoming}

Cluster 3 {Connecticut, Hawaii, Idaho, Indiana, Iowa, Kansas Kentucky, Maine, Minnesota, Montana, Nebraska, New Hampshire, North Dakota, Ohio, Pennsylvania, South Dakota, Utah, Vermont, Virginia, Wisconsin}

Cluster 4 {Florida, North Carolina}

**2.** **Perform K-means clustering with K=4 on the data set. Use the function table() to compare the results obtained in i) and ii). Do they return the same clustering result? For each K-means cluster, what is its within-cluster sum of square? How distinct each cluster is from other clusters?**

No, they didn’t return the same result.

Cluster 1 {Connecticut, Idaho, Indiana, Kansas, Kentucky, Montana, Nebraska, Ohio, Pennsylvania, Utah}

Cluster 2 {Hawaii, Iowa, Maine, Minnesota, New Hampshire, North Dakota, South Dakota, Vermont, West Virginia, Wisconsin}

Cluster 3 {Alabama, Alaska, Arizona, California, Delaware, Florida, Illinois, Louisiana, Maryland, Michigan, Mississippi, Nevada, New Mexico, New York, North Carolina, South Carolina}

Cluster 4 {Arkansas, Colorado, Georgia, Massachusetts, Missouri, New Jersey, Oklahoma, Oregon, Rhode Island, Tennessee, Texas, Virginia, Washington, Wyoming}

Within-cluster sum of square

Cluster 1: 1480.210

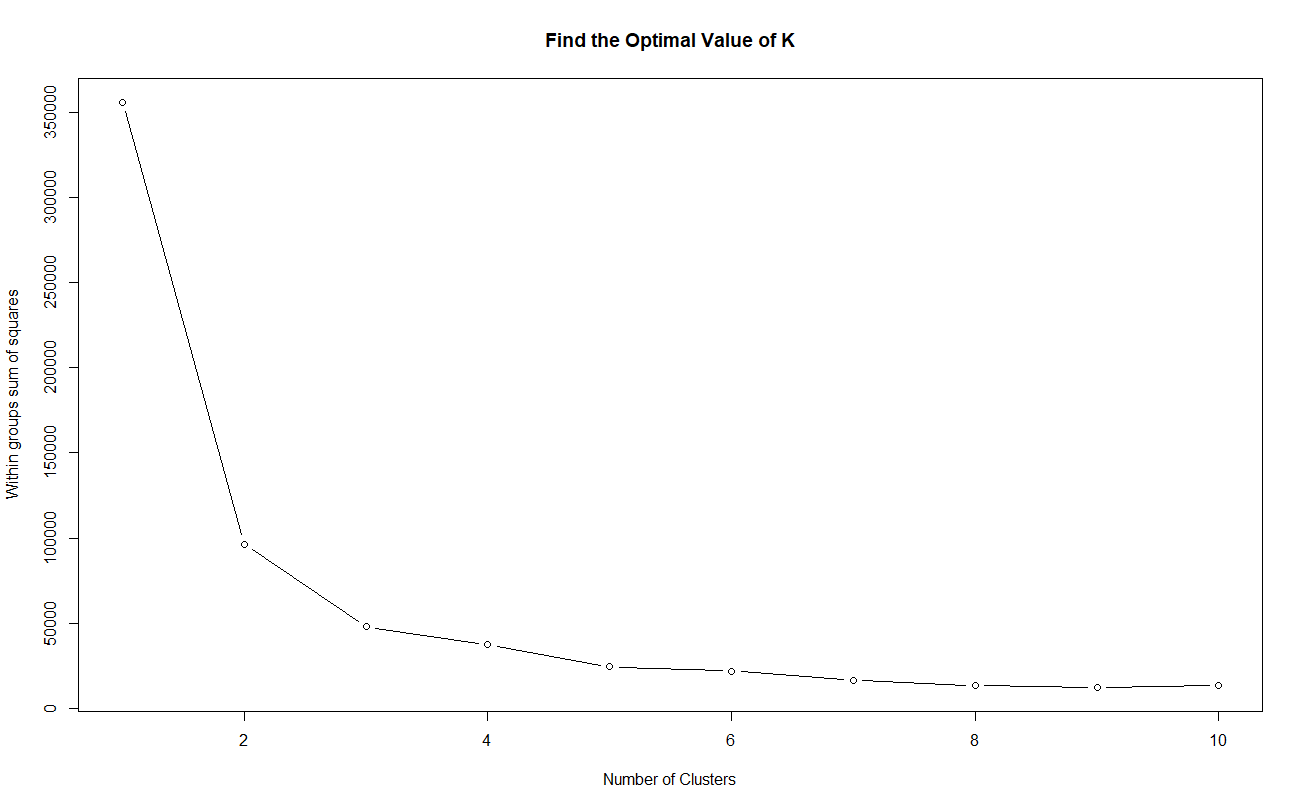
Cluster 2: 4547.914

Cluster 3: 19563.863

Cluster 4: 9136.643

As for the distinctiveness of each cluster, the between-cluster sum of squares we got is 321079.2.

**3. In K-means clustering, could you find the optimal value of K? Justify your answers. Perform K-means clustering using your selected K. List the states in each cluster.**



We can see from the plot above that the optimal value of K is 3.

K-means clustering using k=3:

Cluster 1: {Arkansas, Colorado, Georgia, Massachusetts, Missouri, New Jersey, Oklahoma, Oregon, Rhode Island, Tennessee, Texas, Virginia, Washington, Wyoming}

Cluster 2: {Connecticut, Hawaii, Idaho, Indiana, Iowa, Kansas, Kentucky, Maine, Minnesota, Montana, Nebraska, New Hampshire, North Dakota, Ohio, Pennsylvania  
South Dakota, Utah, Vermont, West Virginia, Wisconsin}

Cluster 3: {Alabama, Alaska, Arizona, California, Delaware, Florida, Illinois, Louisiana, Maryland, Michigan, Mississippi, Nevada, New Mexico, New York, North Carolina, South Carolina}