Homework 2 - Solution

Eris Azizaj

October 27, 2018

Problem 1. Your first task will be using exploratory analysis to familiarize with the dataset. Is the dataset complete? Report useful summary statistics (such as mean value) and plots (such as boxplots or histograms) to help the reader understand more details about the dataset (e.g. show 5 states with the highest and lowest crime rate)

Answer: The dataset is complete. I am showing the mean, median, min and max values of all variables in order to have a better understanding of the dataset.

Table 1: Summary statistics

Statistic	Mean	Median	Min	Max
Population	6,279,593	4,265,220	585,501	39,250,017
Violent.crime	24,821.170	17,054.5	989	174,796
Violent.crime.Rate	390.996	370.250	123.800	1,205.900
Murder.and..nonnegligent..manslaughter	344.788	220	14	1,930
Murder.and..nonnegligent..manslaughter.Rate	5.490	5.250	1.300	20.400
Rape.	2,514.846	1,744.5	169	13,702
Rape.Rate	44.173	41.850	5.000	141.900
Robbery	6,449.981	3,504	59	54,789
Robbery.Rate	90.427	80.400	10.100	510.900
Aggravated.assault	15,511.560	10,521.5	691	104,375
Aggravated.assault.Rate	250.900	242.000	71.300	596.500
Property.crime	152,966.100	109,029.5	10,602	1,002,070
Property.crime.Rate	2,475.754	2,582.050	1,031.900	4,802.900
Burglary	29,295.130	19,721.5	1,771	188,304
Burglary.Rate	466.554	429.650	201.700	830.400
Larceny.theft	108,877.300	77,433.5	8,217	637,010
Larceny.theft.Rate	1,782.363	1,754.950	679.000	4,019.800
Motor.vehicle.theft	14,793.670	9,922	282	176,756
Motor.vehicle.theft.Rate	226.837	215.050	45.100	564.300

It is not intuitive to focus on the numbers without taking into consideration the population of each state, so because of this reason I will use only the rate variables through the completion of this assignment. Personally I do value more the value of life than the value of properties. So, I would prefer to live in a state that has lower violent crime rate than a state that has lower property crime rate or the best scenario would be living in a state that has both low rates of violent and property crimes. As we can see from Fig 1 the most **dangerous** states are *District of Columbia, Alaska, New Mexico, Nevada, Tennessee* and the most **peaceful** states are *Maine, Vermont, New Hampshire, Virginia, Puerto Rico.*

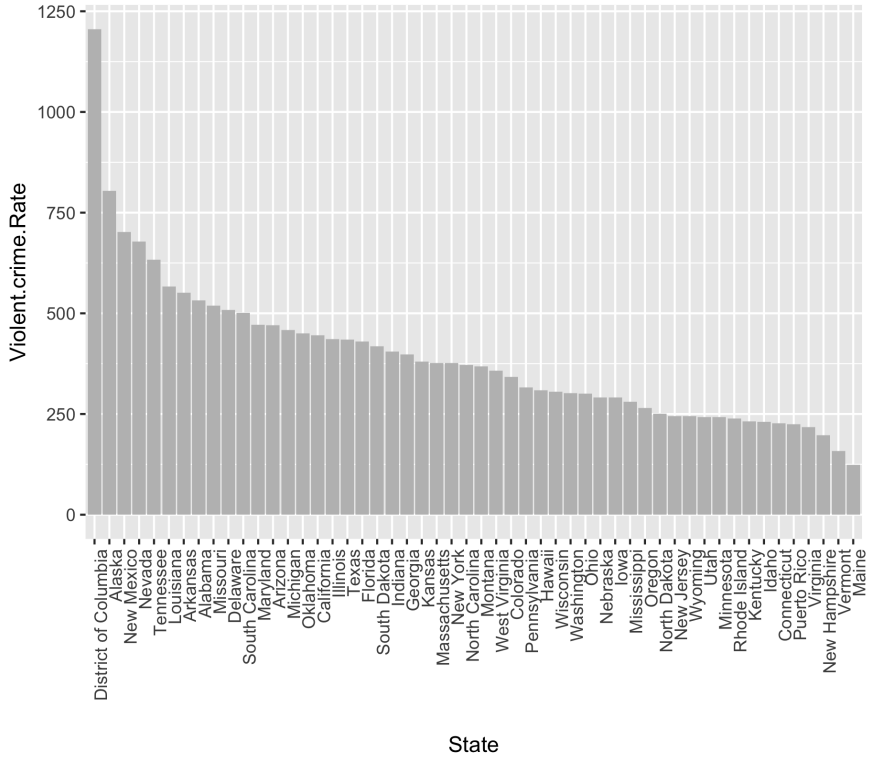

Figure 1: Violent crime rate by State

Regarding the property crime rate, Fig 2 shows that *District of Columbia, New Mexico,Washington, Alaska, Louisiana* are the states with the highest rates of property crime and *Puerto Rico, New Hampshire, New Jersey, New York, Massachusetts* are the safest states. As we can see from these two figures, there are some intersections where states are both with the lowes rates of violent and property crimes. So an ideal safe place to live would be the states of **New Hampshire** and **Puerto Rico**

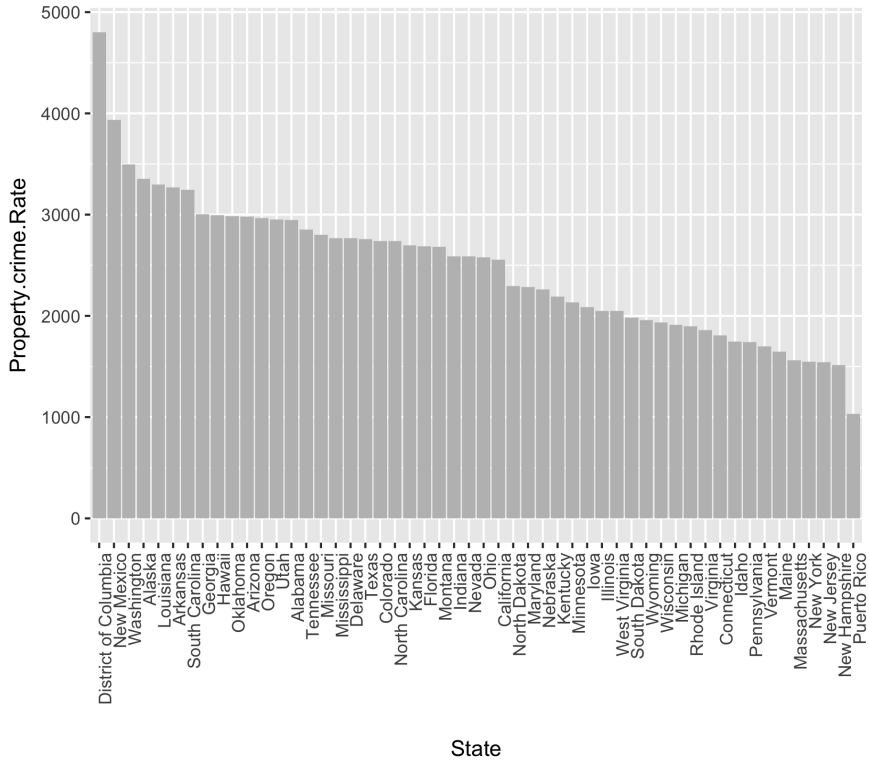

Figure 2: Property crime rate by State

Also another useful plot that might help us understand the dataset is the boxplot for all the rate variables. The following figure shows that violent rate crimes have more similar distribution, while the property crime rates are more spread. The most spread rate seems to be Larceny.theft.Rate. It has a median of 1755 with 25th and 75th percentile of value 1419 and 2052 respectively, 50% of the data lie between these two values. Other rates have smaller ranges.

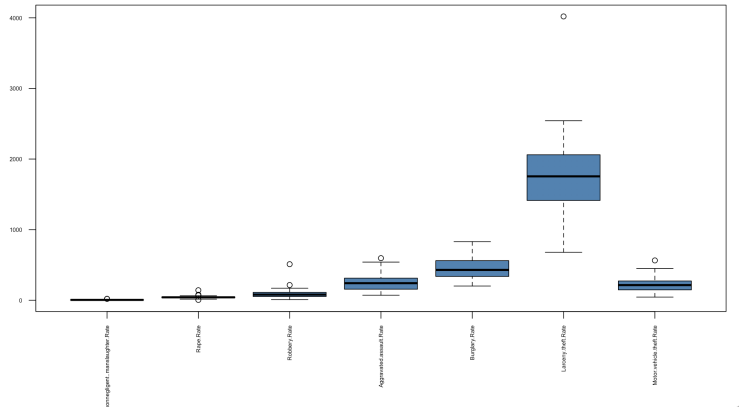

Figure 3: Crime rates

Problem 2. Next, you will need to prepare your data for clustering. Produce a distance matrix to represent the distance between each sample (state), explaining what metric you chose and why. Dont forget to scale your data! You may also decide to exclude some column from this analysis if you think they are not useful or may be detrimental (explain why).

Answer: I am choosing only the rates variable of seven types of crimes because of the reasons explained in Problem 1. First we have to scale the variables(mean=0 and sd=1) and then using the *dist()* function with Euclidean method we find the distance matrix. I am using Eucledian since all variables are numeric. This part is done in the code which is in the Appendix.

Problem 3. Use hierarchical clustering (Wards method) to group the data. Show the resulting dendrogram. By looking at the plot, how many clusters do you suggest are present?

Answer: Fig 4 shows the hierarchical clustering dendrogram. To me it seems that there are 8 clusters. Just by looking at this dendrogram, it seems that District Columbia and Puerto Rico are two outliers since they form separate clusters on their own meaning that they are very diferent from each other and the rest of the states.

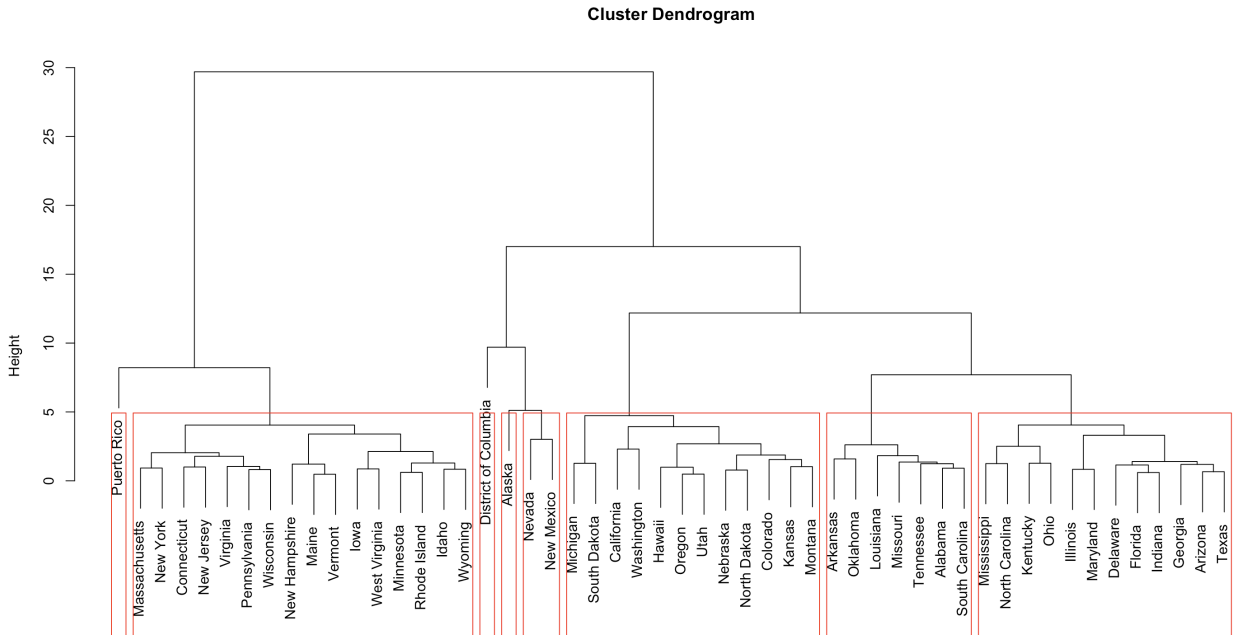

Figure 4: Hierarchical clustering dendrogram

Alaska also form a cluster on its own but still is very similar(close) to Nevada and New Mexico and is far away being an outlier compared to the District of Columbia and Puerto Rico.

Problem 4. Produce a plot of WSS and Calinski-Harabasz index. According to these metrics, what seems to be the most adequate number of clusters?

Answer: Fig 5 gives the values of CH and WSS for all possible values of k from 1 to 10. As we can see from the below figure, CH criterion is maximized at k=8, but also have local maxima at k=3. In order to find the optimal number of clusters from WSS we should look for an "elbow" in the curve. It seems that the "elbow" is at the value k=3 which was also a local maxima for CH index. The intersection of this two is k=3 so i would prefer k=3 as the most adequate number of clusters.

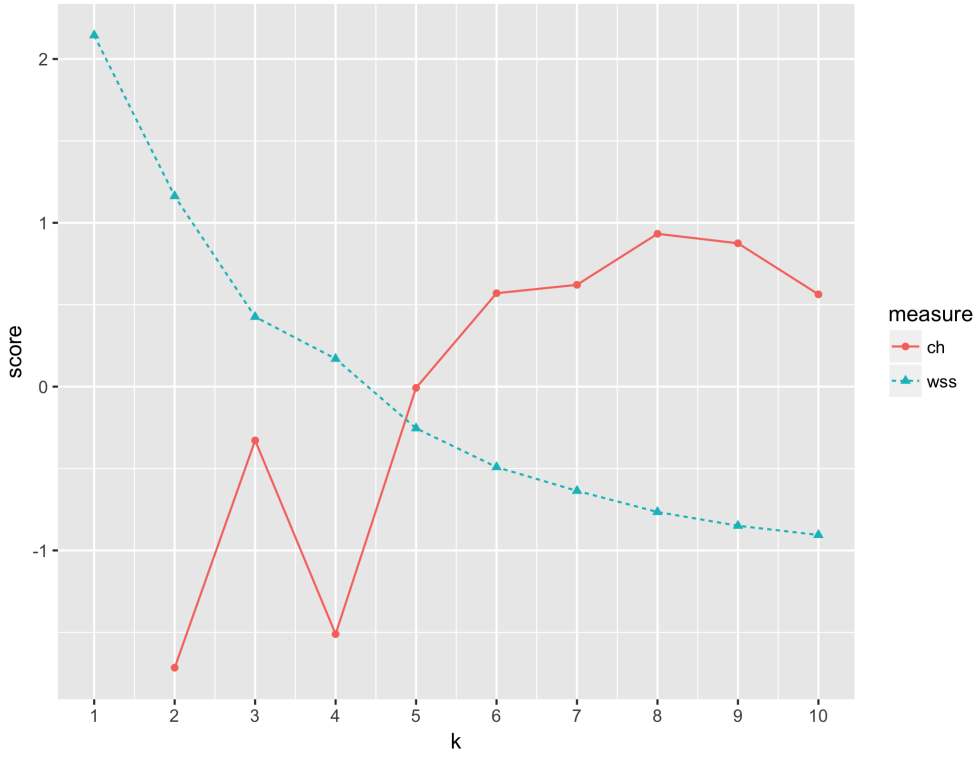

Figure 5: WSS and Calinski-Harabasz index

Problem 5. Use results of points 3 and 4 to pick the best number of clusters. Use that number to do a bootstrap evaluation of cluster stability. Explain your results (what clusters are more stable? How many times have they been dissolved?). Comment on the different clusters: what features do they seem to share?

Answer: By using bootstrap resampling we evaluate how stable a given cluster is, i.e does the cluster represent the actual structure in the data, or is it an artifact of the clustering algorithm? I checked the stability numbers(given by cboot.hclust$bootmean) and the number

of times that the clusters were dissolved(given by cboot.hclust$bootbrd) for two different values of $k \in \{3, 8\}$ and the results are reported in Table 2.

Table 2: Results for $k \in \{3, 8\}$

cluster	1	2	3	4	5	6	7	8	9	10
stability	0.77	0.53	0.82							
dissolvement	7	58	9							
stability	0.80	0.62	0.72	0.63	0.75	0.63	0.67	0.63		
dissolvement	22	44	23	32	19	37	43	37		

Clusters with a stability value less than 0.6 should be considered unstable. Values between 0.6 and 0.75 indicate that the cluster is measuring a pattern in the data, but there is not high certainty about which points should be clustered together. Clusters with stability values above about 0.85 can be considered highly stable (theyre likely to be real clusters). Unfortunately in our results none of the clusters have a stability higher than 0.85(all range between 0.53 - 0.82) but we have stability values very close to it. Referring to Table 2 we see that for k=3 we have higher values of stability and lower number of dissolvement compared to other options of k. This is not surprising since we are specifying the clustermethod to be "hclustCBI" in the function clusterboot().

We expect that states in the same cluster have similar features. Cluster 2 states tend to have high rates of all the crime types but specifically having high *Aggravated.assault.Rate* and high *Motor.vehicle.theft.Rate* compared to other clusters. Cluster 3 has low rates of all variables specially low *Murder.and..nonnegligent..manslaughter.Rate*, low *Rape.Rate* and low *Motor.vehicle.theft.Rate*. Cluster 1 has high rates of *Murder.and..nonnegligent..manslaughter.Rate* and high *Burglary.Rate*.

Cluster 2 includes District of Columbia, Nevada, New Mexico and Alaska which are not very similar with the other as can be seen from the dendrogram. Because of this the stability value is very low(0.53) and dissolvement rate is high(58). So cluster 1 and 3 are more stable, while cluster 3 is taking all the states that does not fit anywhere else.

"cluster 1"

	State	Murder.and..nonnegligent..manslaughter.Rate	Rape.Rate	Robbery.Rate	Aggravated.assault.Rate	Burglary.Rate	Larceny.theft.Rate	Motor.vehicle.theft.Rate
1	Alabama	8.4	39.4	96.4	388.2	700.5	2006.3	241.1
3	Arizona	5.5	47.5	101.8	315.4	544.4	2168.1	265.8
4	Arkansas	7.2	71.7	70.9	401.0	795.5	2233.6	239.4
5	California	4.9	34.9	139.6	265.9	479.8	1623.0	450.3
6	Colorado	3.7	64.2	63.7	211.1	431.4	1955.3	354.0
8	Delaware	5.9	32.4	142.7	327.8	527.6	2078.7	159.7
10	Florida	5.4	36.9	97.9	290.2	486.7	1990.8	209.3
11	Georgia	6.6	34.0	118.4	238.5	614.4	2130.1	259.9
12	Hawaii	2.5	43.3	69.6	193.8	421.2	2175.8	395.8
14	Illinois	8.2	38.3	139.3	250.5	374.9	1518.6	155.5
15	Indiana	6.6	37.7	110.5	249.9	514.0	1853.3	222.1
17	Kansas	3.8	45.1	57.5	274.0	494.1	1962.9	238.6
18	Kentucky	5.9	37.0	75.9	113.5	469.6	1497.4	222.8
19	Louisiana	11.8	38.8	119.1	396.4	740.5	2336.3	220.8
21	Maryland	8.0	29.2	171.0	263.8	410.4	1677.4	196.7
23	Michigan	6.0	71.8	71.7	309.5	398.5	1308.1	203.2
25	Mississippi	8.0	42.7	80.2	149.6	781.4	1842.1	144.7
26	Missouri	8.8	41.9	107.8	360.8	520.4	1978.4	300.3
27	Montana	3.5	55.4	25.5	283.9	377.4	2043.0	263.1
28	Nebraska	2.6	52.1	49.6	186.7	337.9	1677.6	247.8
34	North Carolina	6.7	28.1	92.0	245.5	710.4	1876.2	150.8
35	North Dakota	2.0	45.1	23.9	180.1	427.9	1608.9	259.1
36	Ohio	5.6	48.1	107.8	138.7	575.9	1832.3	169.3
37	Oklahoma	6.2	52.0	80.6	311.0	741.7	1931.4	309.8
38	Oregon	2.8	42.0	55.6	164.1	412.0	2230.0	322.3
41	South Carolina	7.4	48.1	81.3	365.0	664.7	2298.5	280.6
42	South Dakota	3.1	58.8	31.4	325.0	346.6	1460.4	173.6
43	Tennessee	7.3	40.8	117.5	467.3	606.1	2020.7	227.3
44	Texas	5.3	48.0	119.6	261.6	533.8	1978.1	247.8
45	Utah	2.4	49.8	50.5	140.1	420.7	2223.2	307.7
48	Washington	2.7	42.2	77.5	179.7	674.8	2376.3	443.0

"cluster 2"

	State	Murder.and..nonnegligent..manslaughter.Rate	Rape.Rate	Robbery.Rate	Aggravated.assault.Rate	Burglary.Rate	Larceny.theft.Rate	Motor.vehicle.theft.Rate
2	Alaska	7.0	141.9	114.6	540.6	546.3	2394.7	412.1
9	District of Columbia	20.4	78.1	510.9	596.5	346.6	4019.8	436.5
29	Nevada	7.6	58.9	215.6	395.9	641.1	1497.1	448.3
32	New Mexico	6.7	73.3	131.5	491.0	830.4	2542.4	564.3

"cluster 3"

	State	Murder.and..nonnegligent..manslaughter.Rate	Rape.Rate	Robbery.Rate	Aggravated.assault.Rate	Burglary.Rate	Larceny.theft.Rate	Motor.vehicle.theft.Rate
7	Connecticut	2.2	21.3	75.6	128.0	280.9	1328.5	198.7
13	Idaho	2.9	42.7	12.7	172.0	375.4	1245.4	123.4
16	Iowa	2.3	39.8	36.6	212.0	479.5	1447.6	159.0
20	Maine	1.5	30.9	20.0	71.3	300.6	1286.8	58.2
22	Massachusetts	2.0	31.2	78.8	265.0	281.8	1161.0	118.3
24	Minnesota	1.8	42.5	67.5	130.7	337.1	1638.1	158.1
30	New Hampshire	1.3	43.6	32.0	120.7	222.0	1225.7	65.3
31	New Jersey	4.2	16.2	100.4	124.2	282.7	1135.2	126.6
33	New York	3.2	31.7	113.0	228.3	201.7	1271.0	72.9
39	Pennsylvania	5.2	34.7	96.4	180.1	277.8	1362.8	102.1
40	Rhode Island	2.7	41.8	51.1	143.2	358.6	1389.0	151.1
46	Vermont	2.2	28.5	17.0	110.6	336.7	1315.6	45.1
47	Virginia	5.8	32.5	57.1	122.2	238.0	1505.1	116.4
49	West Virginia	4.4	35.9	39.3	278.5	507.9	1402.3	137.0
50	Wisconsin	4.0	34.2	81.4	186.3	336.1	1424.8	172.3
51	Wyoming	3.4	35.0	10.1	195.7	302.5	1518.2	136.6
52	Puerto Rico	19.9	5.0	93.8	105.4	241.9	679.0	111.0

Figure 6: Hierarchical clustering results

Problem 6. Use principal component analysis(PCA) to produce a 2-dimensional plot of the dataset. Are the clusters visible in the plot?

Answer: As we can see from Fig 7 District of Columbia and Puerto rico stand far away from the other states. Alaska and Luisiana also seem to be a little bit of from the other states but not as much as DC and Puerto Rico. Luisiana was not oobvious from the dendrogram but according to PCA it stands abit away from the other states. The clusters are not obvious from this plot.

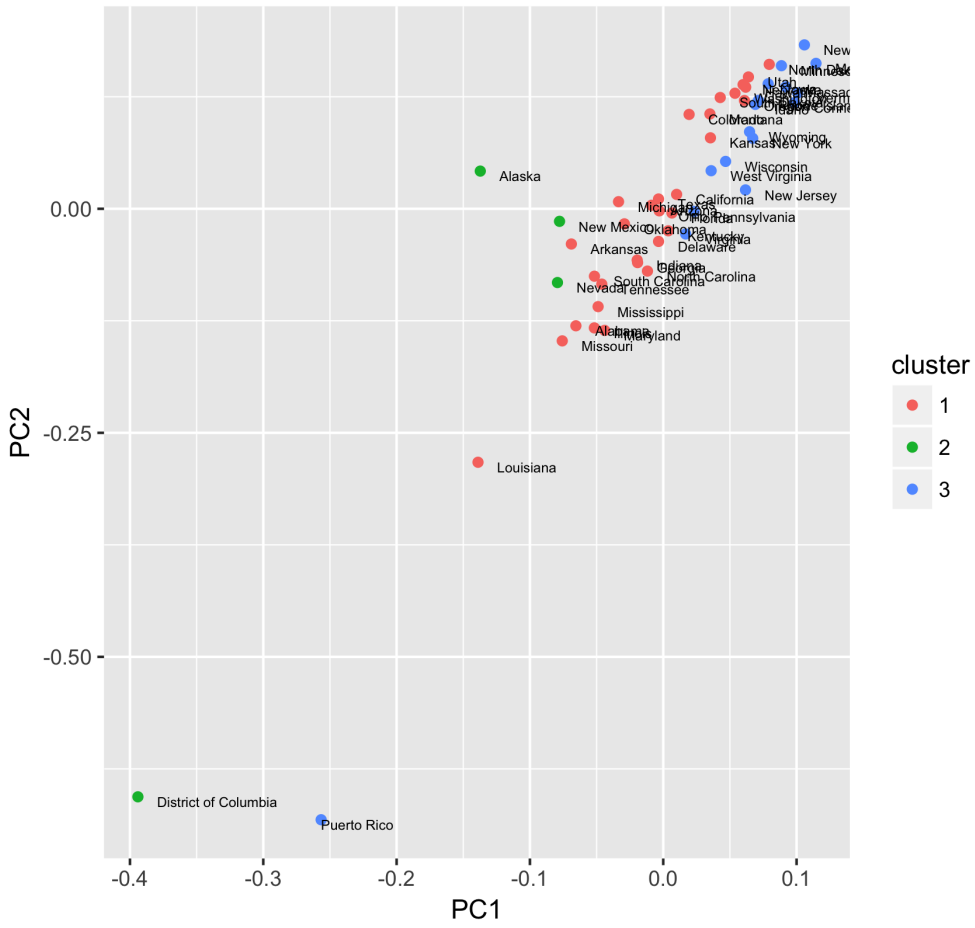

Figure 7: Principal component plot

Problem 7. Use k-means and the number of clusters picked above to group the data again. Comment on the results. Are the cluster the same or there is any visible difference?

Answer: In the previous plots of dendrogram and PCA we realized that the DC and Puerto Rico are outliers. We should exclude them from the k-means since the outliers will pull the centroids of the clusters towards them and make the resulting clusters not accurate. If we compare the hierarchical clusters with the ones of k-means we see that some states have changed cluster. Alaska and New Mexico are in the same cluster while Nevada has moved to another one. Other states of Illinois, Kentucky, Michigan, Nebraska, North Dakota and South Dakota also are part of a different cluster. Referring to Fig 8 we see that cluster 2 has high rates of *Murder.and..nonnegligent..manslaughter.Rate*, high *Burglary.Rate* and high *Motor.vehicle.theft.Rate*. Cluster 3 has generally low rates of almost every variable.

"cluster 1"

	State	Murder.and..nonnegligent..manslaughter.Rate	Rape.Rate	Robbery.Rate	Aggravated.assault.Rate	Burglary.Rate	Larceny.theft.Rate	Motor.vehicle.theft.Rate
2	Alaska	7.0	141.9	114.6	540.6	546.3	2394.7	412.1
32	New Mexico	6.7	73.3	131.5	491.0	830.4	2542.4	564.3

"cluster 2"

	State	Murder.and..nonnegligent..manslaughter.Rate	Rape.Rate	Robbery.Rate	Aggravated.assault.Rate	Burglary.Rate	Larceny.theft.Rate	Motor.vehicle.theft.Rate
1	Alabama	8.4	39.4	96.4	388.2	700.5	2006.3	241.1
3	Arizona	5.5	47.5	101.8	315.4	544.4	2168.1	265.8
4	Arkansas	7.2	71.7	70.9	401.0	795.5	2233.6	239.4
5	California	4.9	34.9	139.6	265.9	479.8	1623.0	450.3
6	Colorado	3.7	64.2	63.7	211.1	431.4	1955.3	354.0
8	Delaware	5.9	32.4	142.7	327.8	527.6	2078.7	159.7
10	Florida	5.4	36.9	97.9	290.2	486.7	1990.8	209.3
11	Georgia	6.6	34.0	118.4	238.5	614.4	2130.1	259.9
12	Hawaii	2.5	43.3	69.6	193.8	421.2	2175.8	395.8
15	Indiana	6.6	37.7	110.5	249.9	514.0	1853.3	222.1
17	Kansas	3.8	45.1	57.5	274.0	494.1	1962.9	238.6
19	Louisiana	11.8	38.8	119.1	396.4	740.5	2336.3	220.8
21	Maryland	8.0	29.2	171.0	263.8	410.4	1677.4	196.7
25	Mississippi	8.0	42.7	80.2	149.6	781.4	1842.1	144.7
26	Missouri	8.8	41.9	107.8	360.8	520.4	1978.4	300.3
27	Montana	3.5	55.4	25.5	283.9	377.4	2043.0	263.1
29	Nevada	7.6	58.9	215.6	395.9	641.1	1497.1	448.3
34	North Carolina	6.7	28.1	92.0	245.5	710.4	1876.2	150.8
36	Ohio	5.6	48.1	107.8	138.7	575.9	1832.3	169.3
37	Oklahoma	6.2	52.0	80.6	311.0	741.7	1931.4	309.8
38	Oregon	2.8	42.0	55.6	164.1	412.0	2230.0	322.3
41	South Carolina	7.4	48.1	81.3	365.0	664.7	2298.5	280.6
43	Tennessee	7.3	40.8	117.5	467.3	606.1	2020.7	227.3
44	Texas	5.3	48.0	119.6	261.6	533.8	1978.1	247.8
45	Utah	2.4	49.8	50.5	140.1	420.7	2223.2	307.7
48	Washington	2.7	42.2	77.5	179.7	674.8	2376.3	443.0

"cluster 3"

	State	Murder.and..nonnegligent..manslaughter.Rate	Rape.Rate	Robbery.Rate	Aggravated.assault.Rate	Burglary.Rate	Larceny.theft.Rate	Motor.vehicle.theft.Rate
7	Connecticut	2.2	21.3	75.6	128.0	280.9	1328.5	198.7
13	Idaho	2.9	42.7	12.7	172.0	375.4	1245.4	123.4
14	Illinois	8.2	38.3	139.3	250.5	374.9	1518.6	155.5
16	Iowa	2.3	39.8	36.6	212.0	479.5	1447.6	159.0
18	Kentucky	5.9	37.0	75.9	113.5	469.6	1497.4	222.8
20	Maine	1.5	30.9	20.0	71.3	300.6	1286.8	58.2
22	Massachusetts	2.0	31.2	78.8	265.0	281.8	1161.0	118.3
23	Michigan	6.0	71.8	71.7	309.5	398.5	1308.1	203.2
24	Minnesota	1.8	42.5	67.5	130.7	337.1	1638.1	158.1
28	Nebraska	2.6	52.1	49.6	186.7	337.9	1677.6	247.8
30	New Hampshire	1.3	43.6	32.0	120.7	222.0	1225.7	65.3
31	New Jersey	4.2	16.2	100.4	124.2	282.7	1135.2	126.6
33	New York	3.2	31.7	113.0	228.3	201.7	1271.0	72.9
35	North Dakota	2.0	45.1	23.9	180.1	427.9	1608.9	259.1
39	Pennsylvania	5.2	34.7	96.4	180.1	277.8	1362.8	102.1
40	Rhode Island	2.7	41.8	51.1	143.2	358.6	1389.0	151.1
42	South Dakota	3.1	58.8	31.4	325.0	346.6	1460.4	173.6
46	Vermont	2.2	28.5	17.0	110.6	336.7	1315.6	45.1
47	Virginia	5.8	32.5	57.1	122.2	238.0	1505.1	116.4
49	West Virginia	4.4	35.9	39.3	278.5	507.9	1402.3	137.0
50	Wisconsin	4.0	34.2	81.4	186.3	336.1	1424.8	172.3
51	Wyoming	3.4	35.0	10.1	195.7	302.5	1518.2	136.6

Figure 8: Kmeans clustering results

Problem 8. Use kmeansruns (criteria ch and asw) to produce a new estimate of the number of clusters. How do they compare to the original one you chose? Comment on the resulting clusters.

Answer: Since kmeansruns uses k-means, for this problem also we are excluding DC and Puerto Rico. Using kmeansruns() function we have the optimal number of clusters k=2 and k=3 for "ch" and "asw" criteria, respectively. Also if we compares the values of clustering.ch$crit and clustcrit$crit we will see that the CH criterion produces different values for kmeans() and hclust() clusterings.

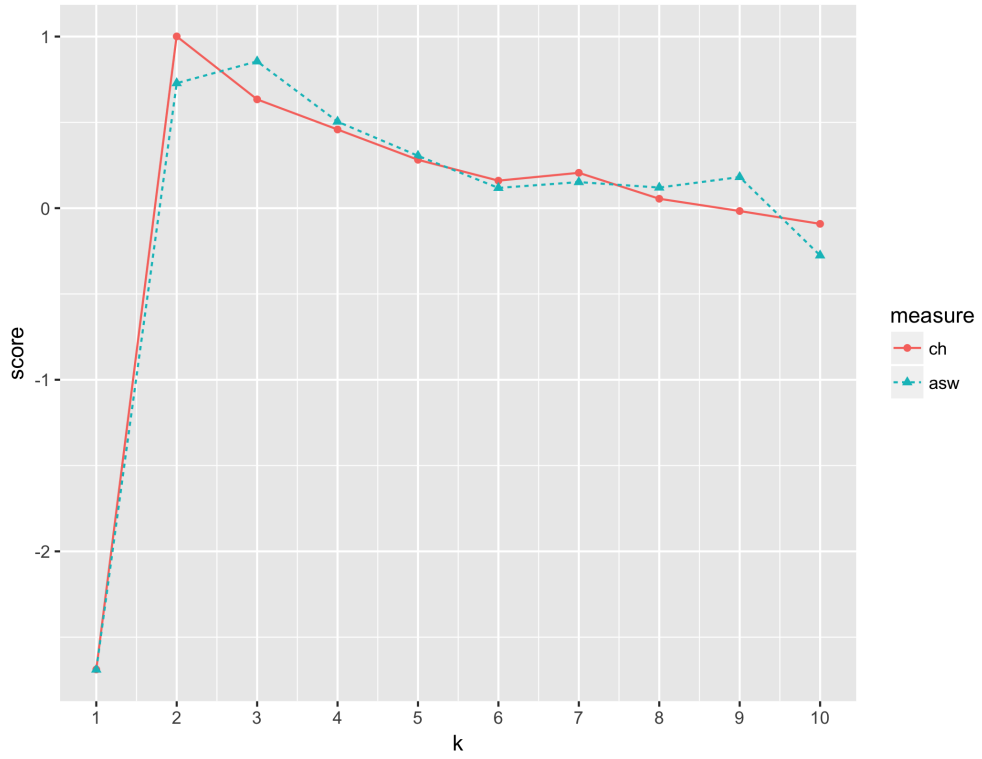

Figure 9: Kmeansruns clustering results

Problem 9. Next, you are going to use Association Rule Mining to find frequent associations between crimes. Since all data is numerical, we will have to bin it and label it. Bin every column into 4 possible levels (Low, Medium, High, Very High). Explain how you chose the binning thresholds. As before, you are allowed to exclude columns from your analysis (explain why).

Answer: I decided to use the rate variables for each crime type since it gives better information on the levels of the crimes in each state without having to worry about the population size. I am binning each variable using the quartiles in order to have a better representation of the data distribution by having equal frequency binning. It would be wrong to have breaks argument that are equally spaced because that would create left or right tail distribution. In Fig 10 which is the density plot of *Murder and nonnegligent manslaughter rate* we see that most of the data is in the lhs, so we should make the bin width so in each category we should have some states for all the variables. The rest of this problem is done in the code.

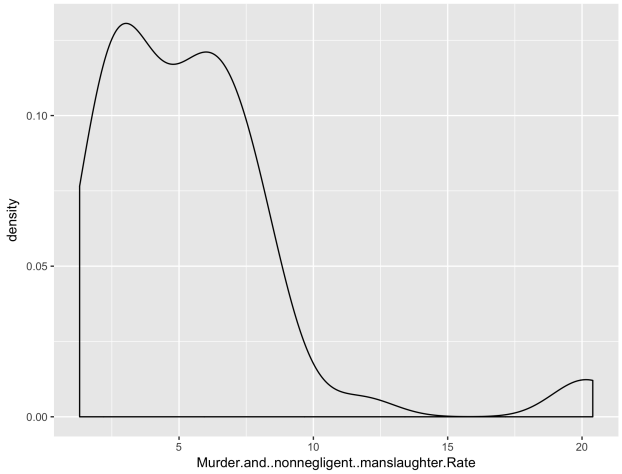

Figure 10: Murder and nonnegligent manslaughter rate distribution

Problem 10. Use the apriori algorithm to find frequent associations in your dataset. Try at least 2 different levels of support (e.g. 5% and 20%) and confidence. Summarize your findings (number of rules, length distribution).

Answer: I tried 3 different support levels {5%, 10%, 20%} with 2 different confidence level {75%, 90%}. The results are all possible combinations are as follows:

(a) Support: 5% & Confidence: 75%

For this scenario we have 446 rules, where most rules contain 3 variables - 2 on the lhs(X) and 1 in the rhs(Y). The quality measures on the rules include support, confidence and lift. As we can see from the summary of quality measures figure, all the rules have a lift greater than 1 which means that there is a good chance that the pattern we are observing is NOT occurring just by chance.

```
set of 446 rules

rule length distribution (lhs + rhs):sizes
  2   3   4   5   6
  4 197 179  60   6

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   3.000   4.000   3.702   4.000   6.000

summary of quality measures:
    support          confidence          lift           count
 Min.   :0.05769   Min.   :0.7500   Min.   :2.786   Min.   : 3.000
 1st Qu.:0.05769   1st Qu.:0.7500   1st Qu.:3.000   1st Qu.: 3.000
 Median :0.05769   Median :0.8571   Median :3.429   Median : 3.000
 Mean   :0.06795   Mean   :0.8822   Mean   :3.527   Mean   : 3.534
 3rd Qu.:0.07692   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.: 4.000
 Max.   :0.19231   Max.   :1.0000   Max.   :4.333   Max.   :10.000

mining info:
        data ntransactions support confidence
 crime_rules           52    0.05       0.75
```

Figure 11: Support: 5% & Confidence: 75%

(b) Support: 5% & Confidence: 90%

By increasing the confidence level to 90% we have a set of 216 rules, where most rules this time contain 4 variables. We see an increase of the min value of lift to 3.74 and confidence at 100%.

```
set of 216 rules

rule length distribution (lhs + rhs):sizes
  3   4   5   6
 63 103  44   6

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   3.000   4.000   3.968   4.000   6.000

summary of quality measures:
    support          confidence      lift           count
 Min.   :0.05769   Min.   :1    Min.   :3.714   Min.   :3.000
 1st Qu.:0.05769   1st Qu.:1    1st Qu.:4.000   1st Qu.:3.000
 Median :0.05769   Median :1    Median :4.000   Median :3.000
 Mean   :0.06357   Mean   :1    Mean   :4.002   Mean   :3.306
 3rd Qu.:0.05769   3rd Qu.:1    3rd Qu.:4.000   3rd Qu.:3.000
 Max.   :0.13462   Max.   :1    Max.   :4.333   Max.   :7.000

mining info:
        data ntransactions support confidence
 crime_rules           52    0.05        0.9
```

Figure 12: Support: 5% & Confidence: 90%

(c) Support: 10% & Confidence: 75%

When we increase the level of support to 10% and keep confidence 75% the set of rules is 26, i.e decreases even more. For this case most rules now have 3 variables, 2 in the lhs(X) and 1 in the rhs(Y). We see a decrease of the min and max value of lift.

```
set of 26 rules

rule length distribution (lhs + rhs):sizes
 2  3  4
 4 18  4

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      2       3       3       3       3       4

summary of quality measures:
    support         confidence          lift           count
 Min.   :0.1154   Min.   :0.7500   Min.   :3.000   Min.   : 6.000
 1st Qu.:0.1154   1st Qu.:0.7692   1st Qu.:3.077   1st Qu.: 6.000
 Median :0.1154   Median :0.8286   Median :3.314   Median : 6.000
 Mean   :0.1354   Mean   :0.8368   Mean   :3.347   Mean   : 7.038
 3rd Qu.:0.1490   3rd Qu.:0.8571   3rd Qu.:3.429   3rd Qu.: 7.750
 Max.   :0.1923   Max.   :1.0000   Max.   :4.000   Max.   :10.000

mining info:
        data ntransactions support confidence
 crime_rules           52     0.1       0.75
```

Figure 13: Support: 10% & Confidence: 75%

(d) Support: 10% & Confidence: 90%

For this case we have only 4 rules and still most rules contain 3 variables. All lift values are 4 and confidence is 100%

```
set of 4 rules

rule length distribution (lhs + rhs):sizes
3 4
3 1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3.00    3.00    3.00    3.25    3.25    4.00

summary of quality measures:
    support         confidence      lift       count
 Min.   :0.1154   Min.   :1    Min.   :4   Min.   :6.0
 1st Qu.:0.1154   1st Qu.:1    1st Qu.:4   1st Qu.:6.0
 Median :0.1250   Median :1    Median :4   Median :6.5
 Mean   :0.1250   Mean   :1    Mean   :4   Mean   :6.5
 3rd Qu.:0.1346   3rd Qu.:1    3rd Qu.:4   3rd Qu.:7.0
 Max.   :0.1346   Max.   :1    Max.   :4   Max.   :7.0

mining info:
       data ntransactions support confidence
 crime_rules            52     0.1        0.9
```

Figure 14: Support: 10% & Confidence: 90%

(e) Support: 20%

If we gonne increase the support to 20% than for different level of support we will have NO rule.

Since the number of items is small in this data, i.e no rare events, we should keep the level of support high enough.

Problem 11. Pick the combination of support and confidence that seems to produce the best results. Report the 5 rules with highest confidence and explain your findings.

Answer: Based on the findings of problem 10, I would prefer a high level of support. So I will pick support=10% and confidence=75% even though there is no much difference with confidence=90%. For this case the 5 rules with highest confidence values are as below:

```
    lhs                                    rhs                                      support confidence     lift count
[1] {Rape.Rate.level=Low,
     Larceny.theft.Rate.level=Low}       => {Burglary.Rate.level=Low}            0.1346154  1.0000000 4.000000     7
[2] {Rape.Rate.level=Low,
     Aggravated.assault.Rate.level=Low}  => {Burglary.Rate.level=Low}            0.1153846  1.0000000 4.000000     6
[3] {Rape.Rate.level=Low,
     Motor.vehicle.theft.Rate.level=Low} => {Burglary.Rate.level=Low}            0.1346154  1.0000000 4.000000     7
[4] {Rape.Rate.level=Low,
     Larceny.theft.Rate.level=Low,
     Motor.vehicle.theft.Rate.level=Low} => {Burglary.Rate.level=Low}            0.1153846  1.0000000 4.000000     6
[5] {Burglary.Rate.level=Low,
     Larceny.theft.Rate.level=Low}       => {Motor.vehicle.theft.Rate.level=Low} 0.1538462  0.8888889 3.555556     8
```

Figure 15: Rules with Support: 10% & Confidence: 75%

The first 4 rules are of length 3. They show that if a state has low level of *Rape.Rate* associated with low level of *Larceny.theft.Rate*, or low level *Aggravated.assault.Rate* or low level of *Motor.vehicle.theft.Rate* or a combination of low level of *Larceny.theft.Rate* AND low level

of *Motor.vehicle.theft.Rate* then that state will have low levels of *Burglary.Rate*. The fifth rule is a combination of length four and claims that if a state has low levels of *Burglary.Rate* and low level of *Larceny.theft.Rate* then it will have low level of *Motor.vehicle.theft.Rate*.

Problem 12. Filter your resulting rules to find rules that have Murder and nonnegligent manslaughter Rate = Very High on the right hand side. How many such rules exist? Report 5 most confident and comment on results.

Answer: If we set support=10% and confidence=75%, there exist only one rule that has Murder and nonnegligent manslaughter Rate = Very High on the right hand side and that is:

```
    items
[1] {Aggravated.assault.Rate.level=Very High,Burglary.Rate.level=Very High}
```

Figure 16: lhs of rules with "murder.rate=Very High" in the rhs

This rule says that if a state has very high levels of *Aggravated.assault.Rate* and very high levels of Murder.and..nonnegligent..manslaughter then we expect this state to have very high levels of *Murder.and..nonnegligent..manslaughter.Rate*.

Problem 13. Write a conclusion (at most 18 sentences!) summarizing the most important findings of the assignment. What did we learn about the dataset? Given your results, in which state would you rather live (safest)?

Answer: In this assignment we learned how to cluster the data using hierarchical clustering and k-means. Also we used different measures to see the performance of different algorithms and their suggestion for the optimal number of clusters. Visualization of clusters in the dendrogram and through principal component analysis suggest that Disrict of Columbia and Puerto Rico are outliers in the data, so we should exclude them for k-means to have unbiased clustering results. Using the association rule mining we found different sets of rules with different levels of support and confidence. Since the number of items in this data is not high, the events happen frequently which tells us to keep the support level high enough. For support=10% the 5 rules with the highest confidence level have low level of *Burglary.Rate* in their rhs. In the beginning of this assignment I claimed that I would prefer to live in a state with low levels of violent crime since I value more the value of life than that the value of properties. Based only on violent crime rate I chose Puerto Rico as the ideal state to live but I was wrong. Through out the completion of this assignment I realized that the value

of *Murder.and..nonnegligent..manslaughter.Rate* for Puerto Rico is almost the same as DC which has the highest rate in the nation and was classified as the most dangerous state to live. Based on all the analysis results of this assignment, I would prefer to live in the New Hampshire, which was also the other most peaceful state to live. Also it has the lowest rate of *Murder.and..nonnegligent..manslaughter.Rate*.

APPENDIX

I was not able to include the code as an Appendix. I have uploaded it as a seperate file.