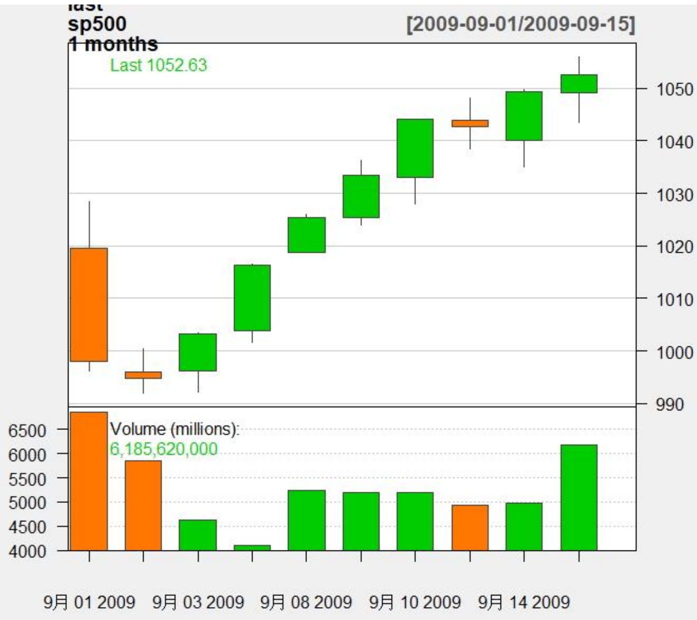
Jiawei Liu   9752650

QUESTION ONE

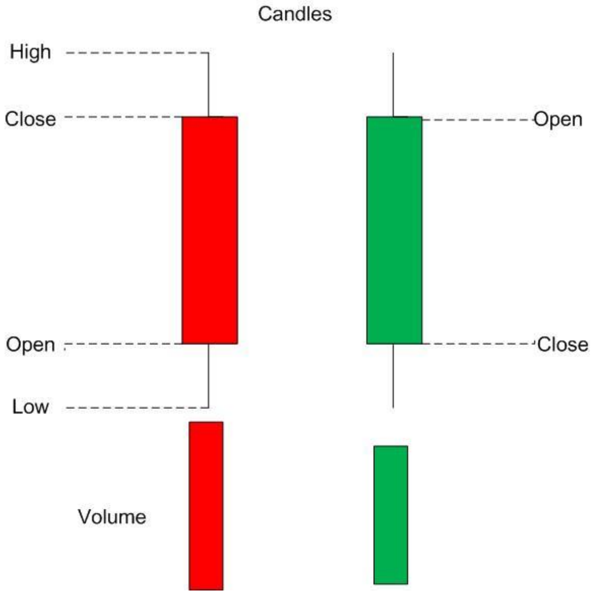
1. (1) Describe what the candles are showing?

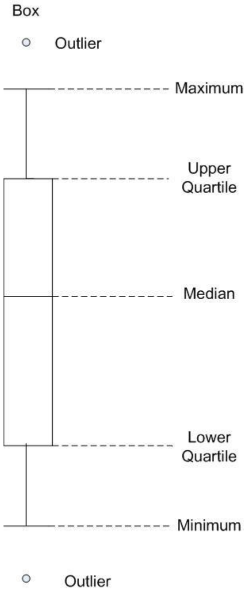


The candle plot is usually used to demonstrate the price change and trading volume day by day. Every candle is used to describe the price change in every single day and the bar under the candle shows turnover in that day. The further details will be introduced below.

(2) Explain how candle plots and boxplots differ?

Two pictures below show the basic elements in candle plots and boxplots:

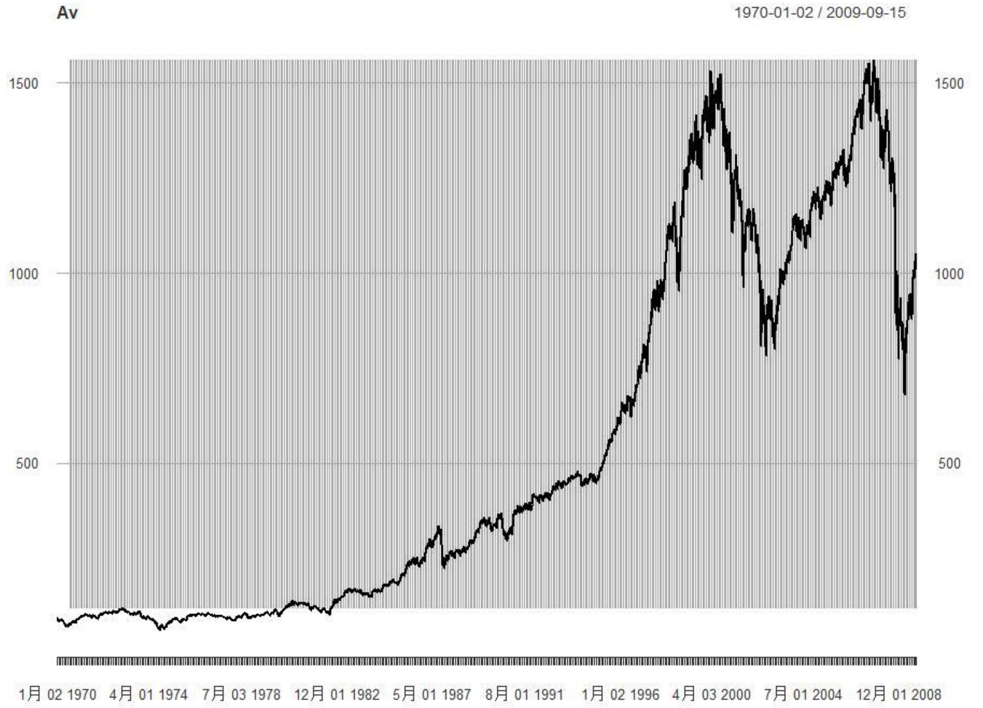




It can be seen that in first graph, there are five terms here. 'High' means the highest price in that day and 'Low' is the lowest. Meanwhile, 'Close' means the final price in that day and ‘Open’ means the beginning price. ‘Volume’ means the trading volume in that day. If the colour is red, it usually means that price increases. Therefore, close price is higher than open price. If it is green, the price decreases and open price is higher than open price.

When it comes to boxplot, if can be found that there are six points. ‘Maximum’ means the greatest value while ‘Minimum’ means the least. ‘Median’ is the middle value of dataset which means 50% of data is greater than this value. ‘Upper quartile’ is the value which is less than 25% of data and ‘Lower quartile’ is the value which is less than 75% of data. Sometimes there are also some upper and lower outliers in boxplots.

2. A plot showing Av(t).



R code used to calculate Av(t):

```
 Av <- (sp500$High + sp500$Close + sp500$Low)/3
 plot(Av)
 sp500 <- cbind(sp500,Av)
 head(sp500)
```

references:
https://en.wikipedia.org/wiki/Candlestick_chart
https://en.wikipedia.org/wiki/Box_plot

QUESTION TWO

1. Load in the dataset network.csv.

2. Maximum and minimum degree.

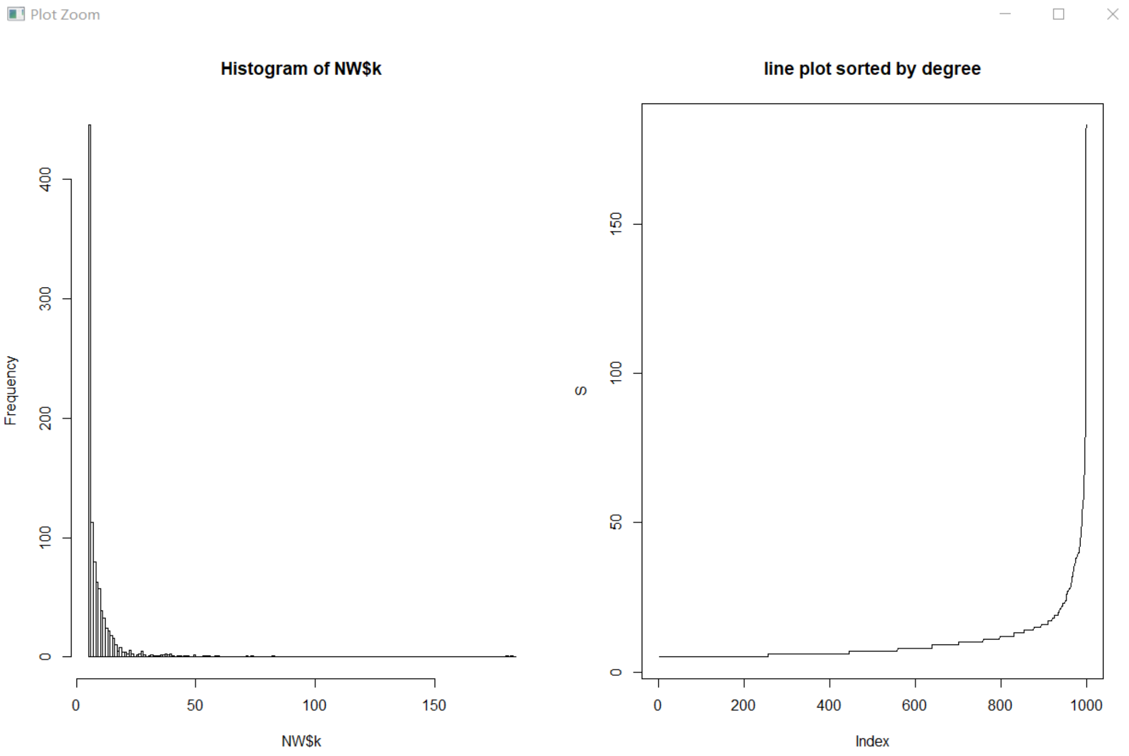
```
NW <- read.csv("D:/Assn1/network.csv",header=TRUE)
Max <- max(NW$k)
Min <- min(NW$k)
```

   Maximum degree is 183

   Minimum degree is 5

3. A single figure with 2 plots.

```
par(mfrow = c(1,2))
NW.hist <- hist(NW$k,breaks=seq(from=min(NW$k),to=max(NW$k)+1,by=1))
S <- sort(NW$k)
plot(S,main = "line plot sorted by degree",type = 'l')
```

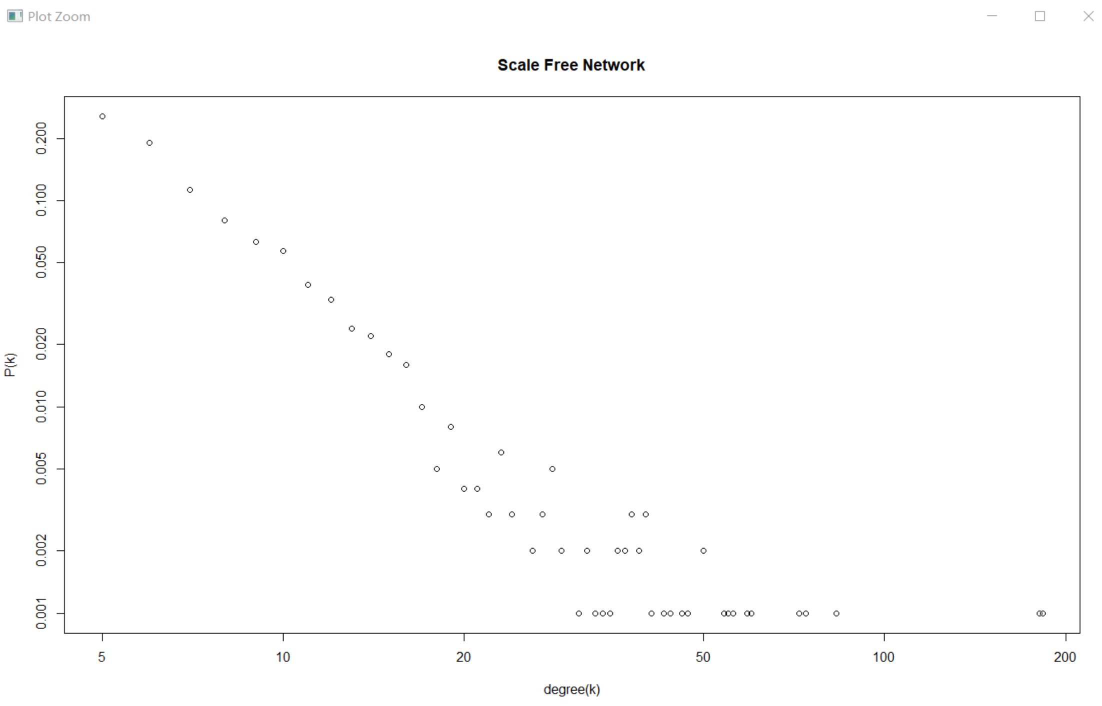


   These two plots show me the distribution of degree in network. It can be seen that the majority of nodes' degree locates around 5 to 10. There are only a few nodes/outliers which are more than 150.

4. Calculating the P(K) and plot it.

```
library(plyr)
T <- count(NW$k)
Pk <- T/1000
```

```
X <- unique(T)
plot(X$x,Pk$freq,log = 'xy',main = "Scale Free Network",xlab = "degree(k)",ylab = "P(k)")
```



5. Why is the network likely to be scale-free?
   There are two common types of network now, one is random network and one is scale-free network. Scale-free network is the network which follows the power law degree distribution and the most famous example for it is the World Wide Web. There are two terminologies which intend to explain the reasons for the scale-free network, preferential attachment and fitness model. To put it simple, there are some famous or popular websites/nodes at the beginning and most of people tend to visit these websites. Then, although more websites/nodes appear, the people who used this network before as well as others who are not familiar with this network also tend to visit these super “hot” websites/nodes rather than new websites/nodes. This phenomenon might be the reason for the scale-free network. Therefore, the plot in step 4 show us that most of nodes are in very low degree and only a few of them, representing those “hot” websites/nodes, are in very high degree,

References:
https://en.wikipedia.org/wiki/Scale-free_network
http://mathworld.wolfram.com/Scale-FreeNetwork.html

QUESTION THREE

1. Briefly state the meaning of each variable for the countrystats data.

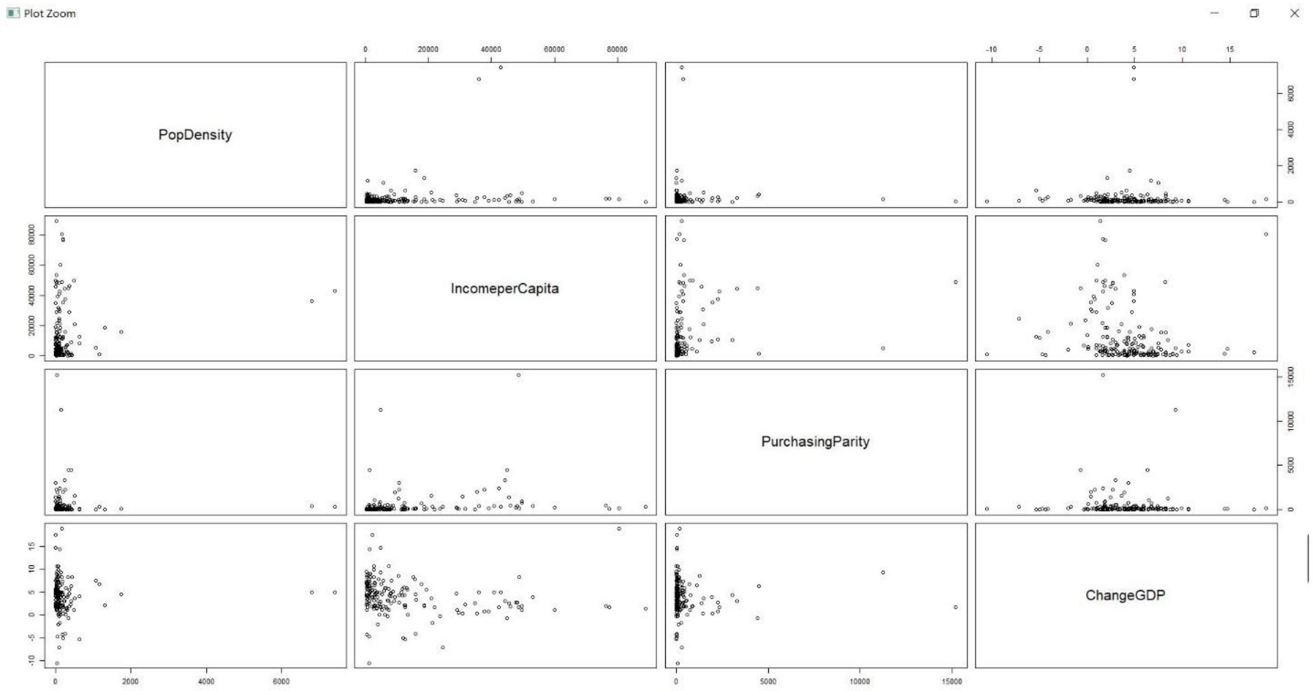
   PopDensity: the variable PopDensity shows how many people per km^2 in a particular country.

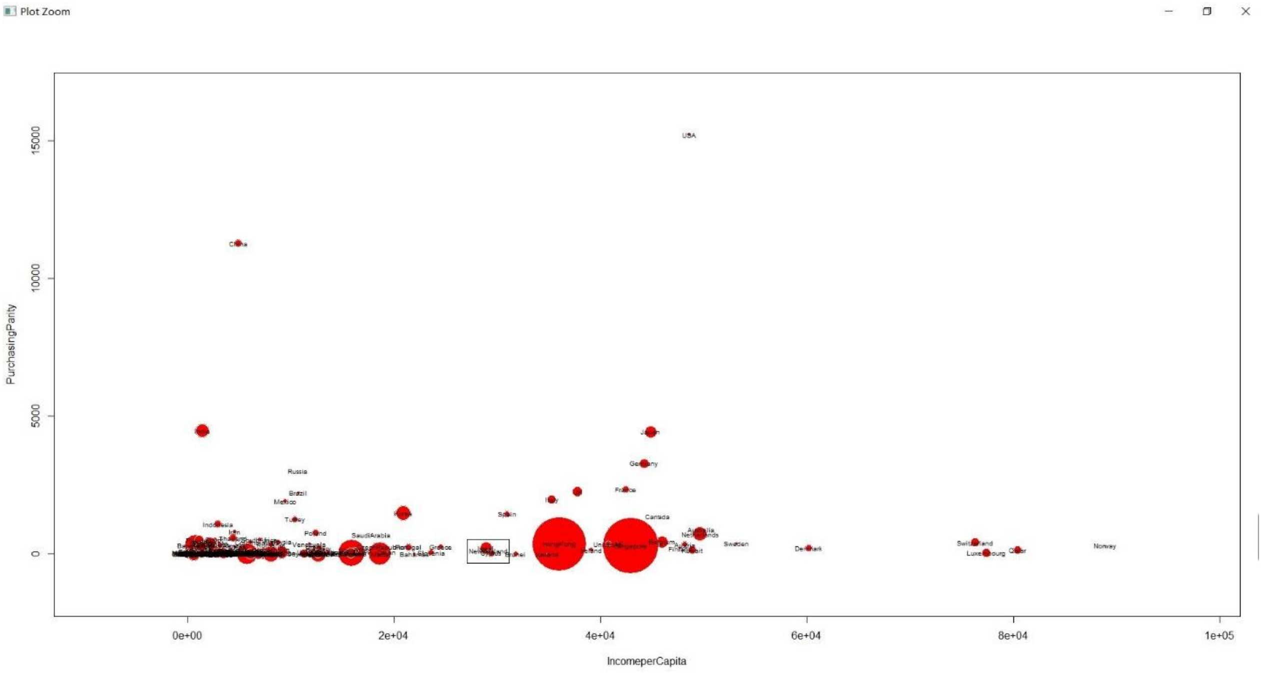
   IncomeperCapita: the variable IncomeperCapita shows the amount of income per capita in a particular country.

   PurchasingParity: the variable PurchasingParity shows the currencies' purchasing power of different countries respectively. This variable is determined by not only currencies themselves but also the prices of countries' goods.

   ChangeGDP:the variable ChangeGDP is simple to understand. It shows the change of GDP in a particular country.

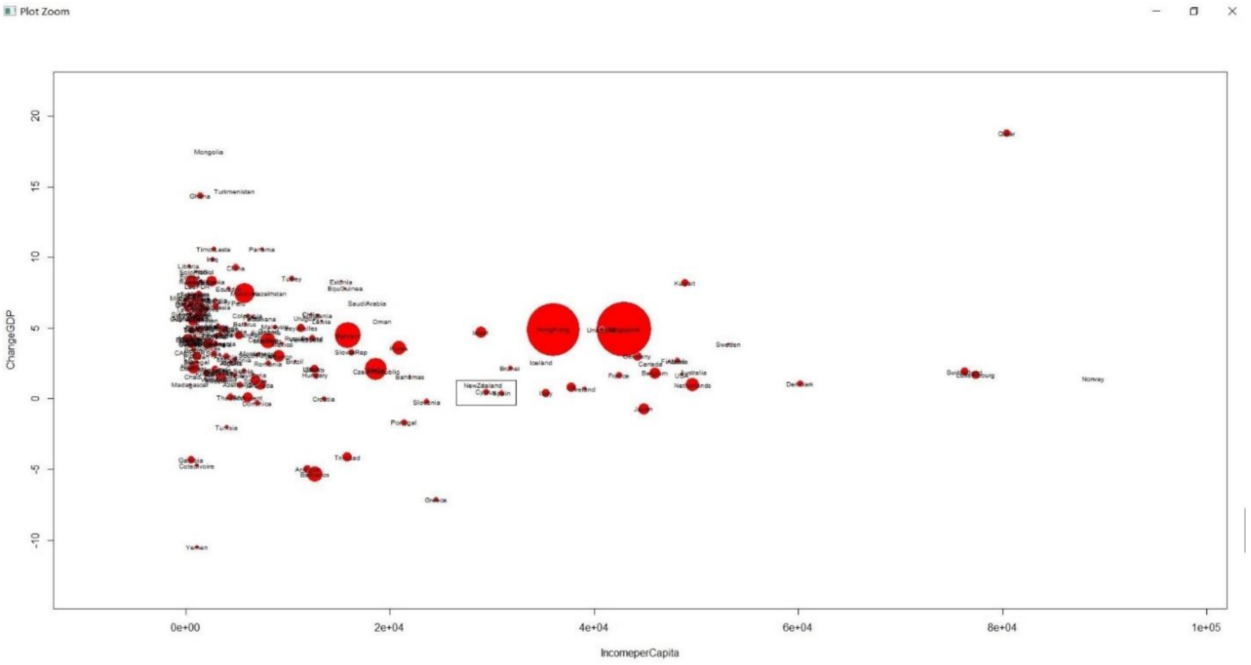
2. Visualize the countrystats data to examine the relationships between countries. (1) Which countries appear to be the similar to New Zealand?



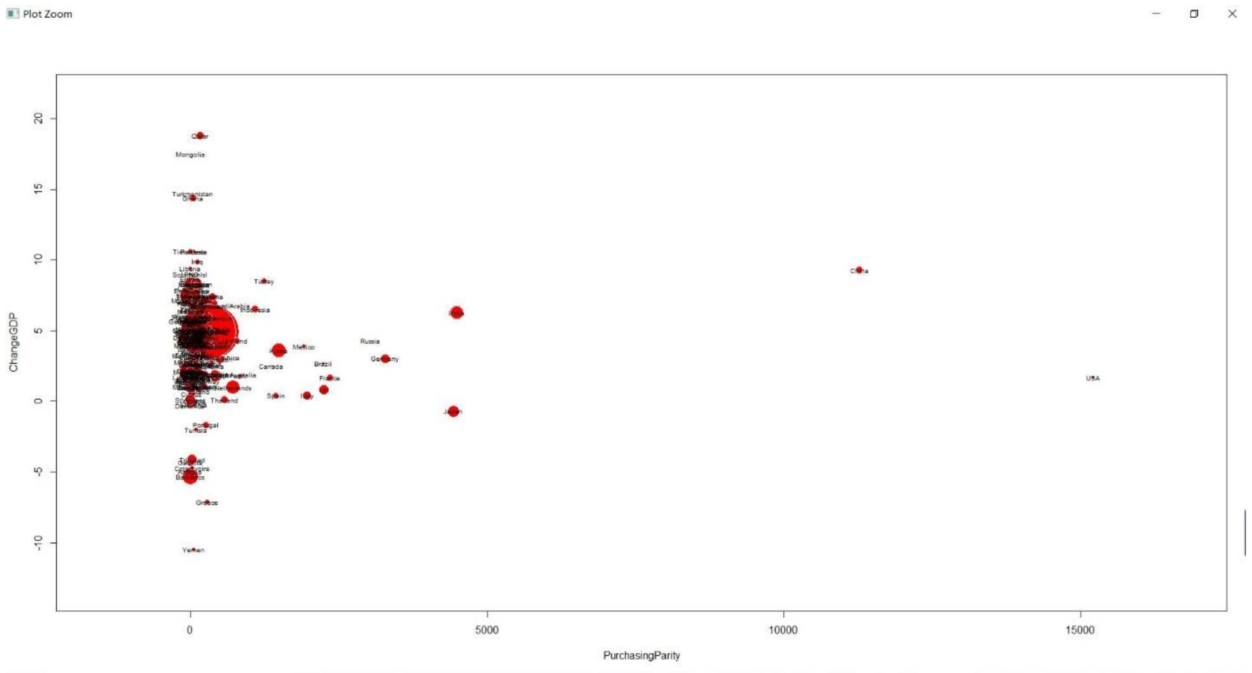


```
popradius= sqrt(C$PopDensity/pi)
symbols(C$IncomeperCapita, C$PurchasingParity, circles=popradius,
            inches=0.45, fg="white", bg="red", xlab="IncomeperCapita",
ylab="PurchasingParity")
text(C$IncomeperCapita, C$PurchasingParity, rownames(C), cex=0.6)
```

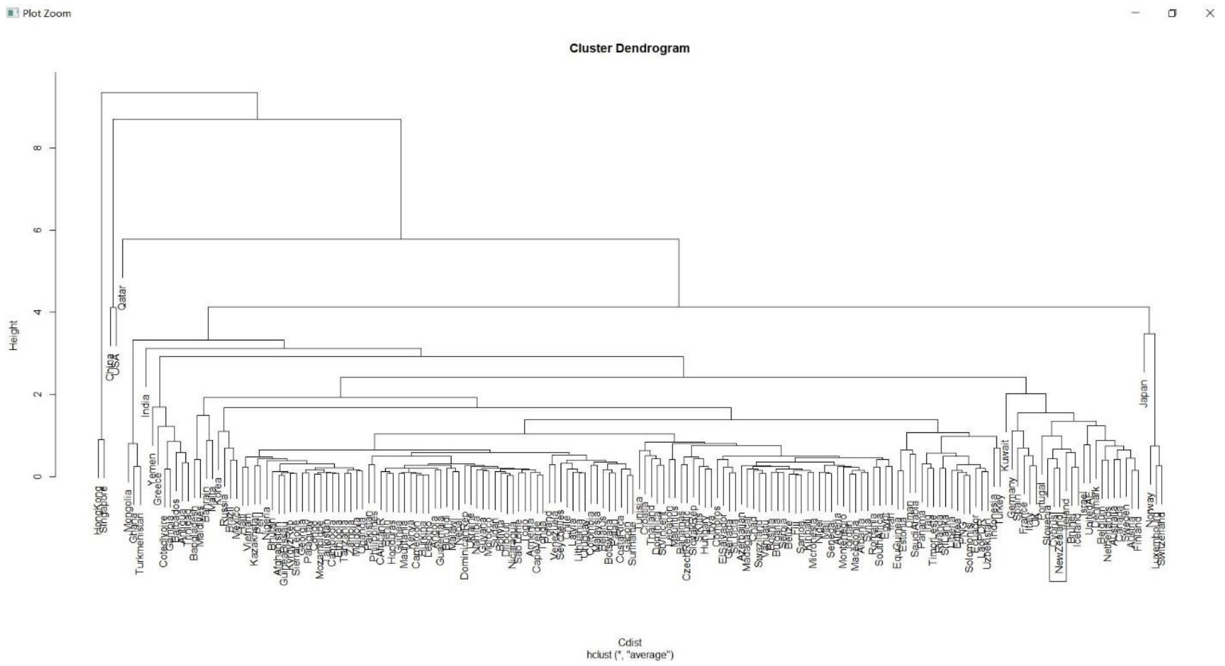
This is a plot of IncomeperCapita (x-axis) versus PurchasingParity (y-axis) and the size of bubble shows the PopDensity. It is not clear here and I have already find New Zealand and other familiar countries and put them in a box. In this plot, the most familiar country is Cyprus and Israel.



This is a plot of IncomeperCapita (x-axis) versus ChangeGDP (y-axis). It is also not clear here and I have already find New Zealand and other familiar countries and put them in a box. In this plot, the most familiar country is Cyprus and Spain.



This is a plot of PurchasingParity (x-axis) versus ChangeGDP (y-axis). It can be seen that in this plot, most of countries get together and I cannot tell that which one is New Zealand.



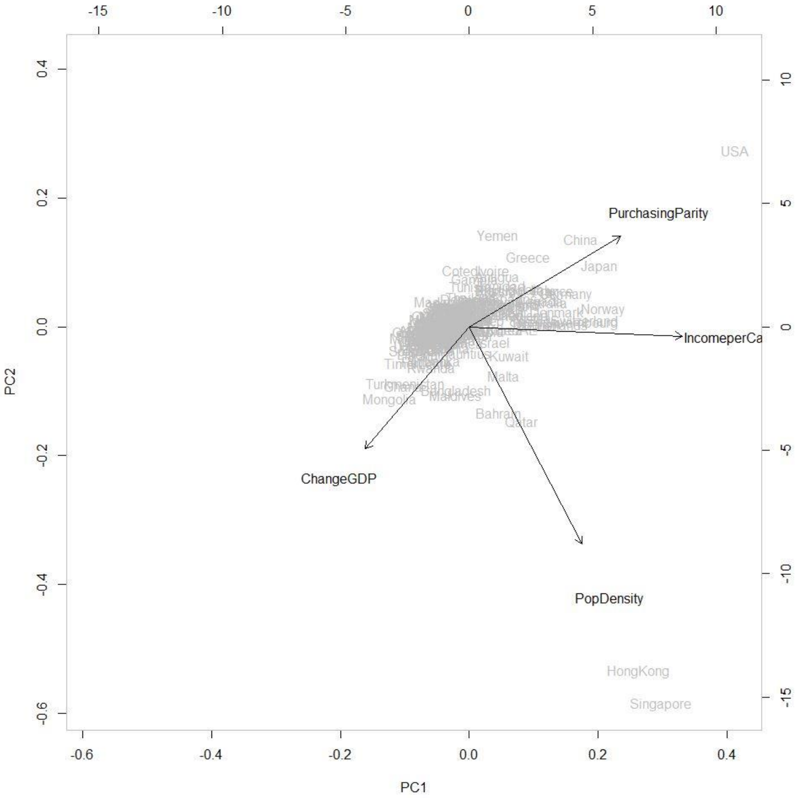
C <- scale(C)
Cdist <- dist(C)
Cdend <- hclust(Cdist,method="average")
plot(Cdend)

This is the dendrogram of the countrystats data using average linkage after scaling the data. It can be seen that Cyprus is the most similar country to New Zealand. Then it is followed by Slovenia, Ireland, Brunel and Iceland. The result here is also quite close to above plots.

It can be concluded that Cyprus is the most similar country to New Zealand.

(2) Which countries are the strongest/weakest?
It is generally believed that for a strong country, PopDensity should not be too large or too small. IncomeperCapita could be important but PurchasingParity may be more significant. Finally, ChangeGDP should be high but if it is too high, such as 15%-20%, it may also mean that country is not in a good situation.
In conclusion, the USA may be the strongest country and Yemen may be the weakest.

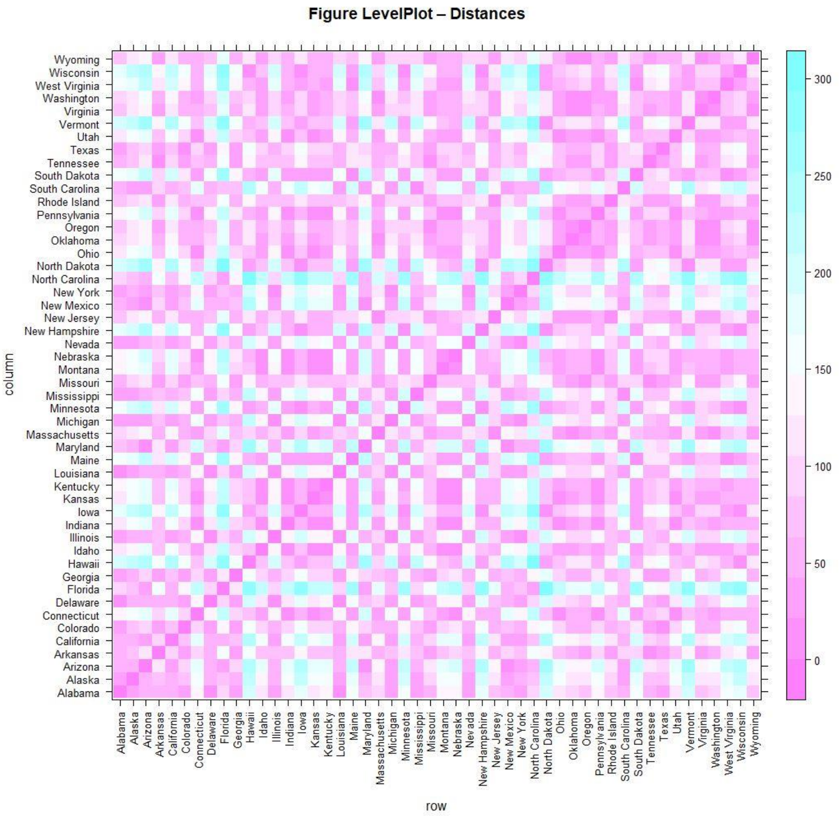
3. Produce a figure showing the biplot and discuss/interpret the principal component biplot.



```
C.pca = prcomp(C,scale=TRUE)
print(C.pca)
summary(C.pca)
plot(C.pca)
biplot(C.pca, col=c("gray","black"))
```

This plot shows the first two principal components of these data. The grey country names represent the score for the first two principal components. The black arrows indicate the first two principal component loading vectors. It can be seen that the loadings for ChangeGDP on the first and second principal components are negative numbers. Therefore, this variable is far from the other three variables and it means ChangeGDP variable is less correlated with others. For the rest of three variables, it can be seen that the loading for IncomeperCapita on first principal component is largest and loading for PurchasingParity on second principal component is largest. Therefore, it can be easily found that the USA have the largest positive scores on both first and second principal components and it may mean the USA is the strongest. Meanwhile, Mongolia might be the weakest.

4. (1) Create a distance matrix (without scaling) for the dataset USArrests and visualise it using levelplot. Name 3 states that appear to be dissimilar (different) to New York.

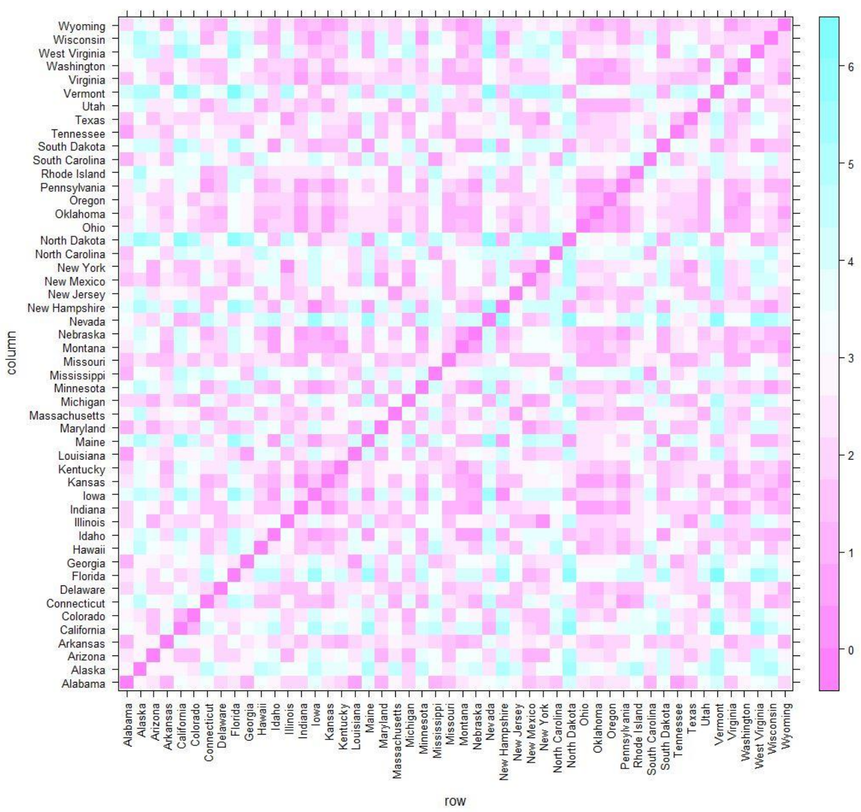


```
U <- USArrests
Udist <- dist(U)
levelplot(as.matrix(Udist), main = "Figure LevelPlot – Distances", scales=list(x=list(rot=90)))
```

From this levelplot, it can be seen that North Dakota, Vermont and Lowa seem to be different from New York.

(2) Explain why scaling may change the distance relationships and demonstrate that this is true.

Figure LevelPlot – Distances(scaled)

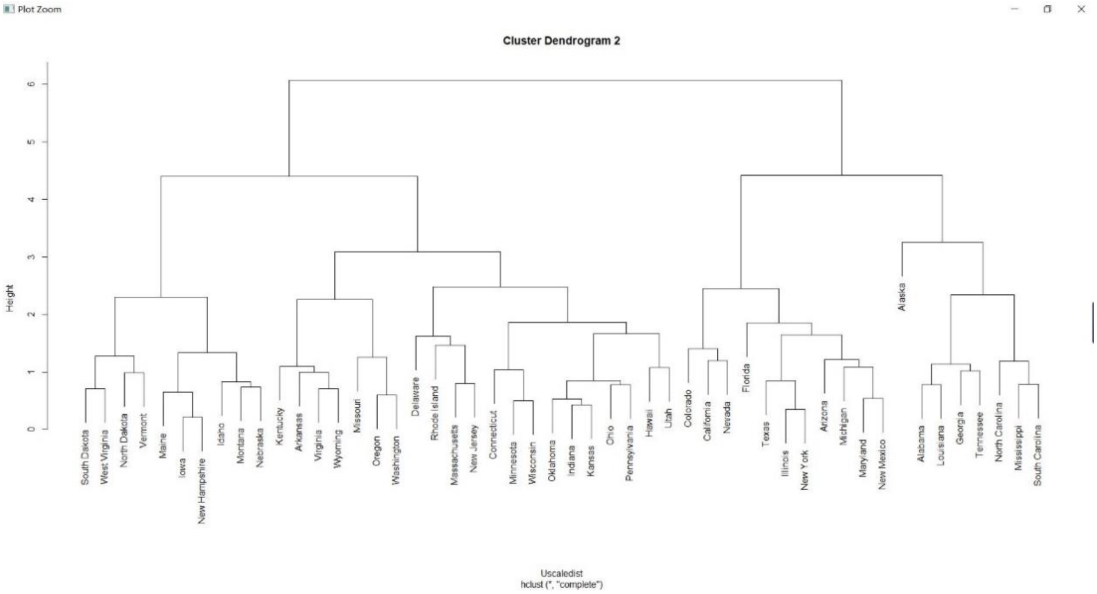


After the data is scaled, it can be seen that here, North Dakota, Vermont and West Virginia seem to be different from New York. The reason for the change is that the scaling changes the distance relationships.

(3) Explain why the scaled versus unscaled levelplots are different.

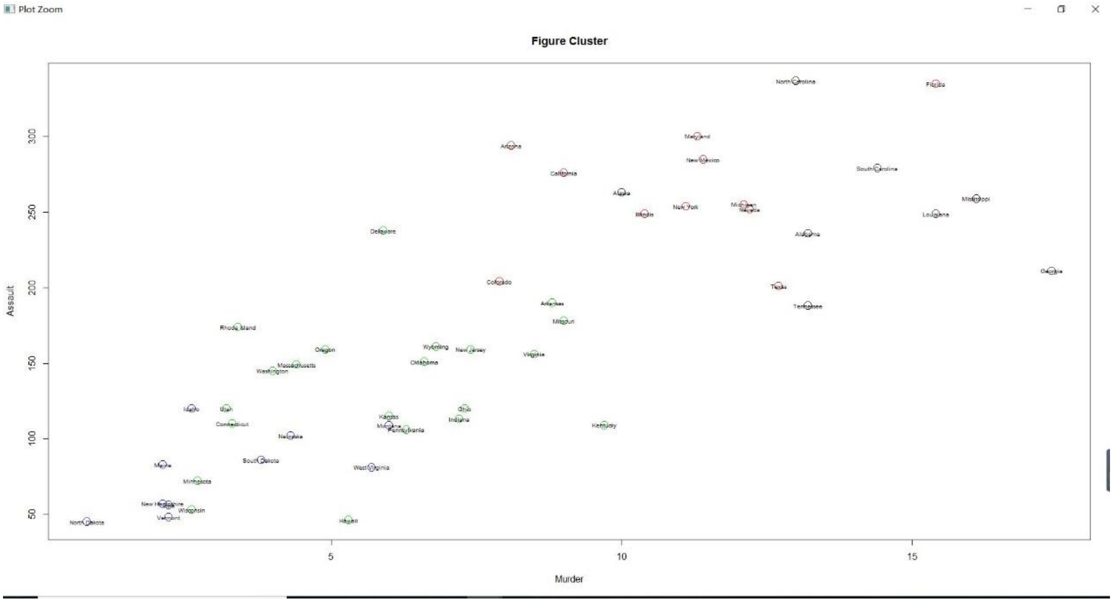
The reason for this is that the value formats or ranges of four variables are different. There may be some extreme big or small values and they may affect the results of distance relationships. Therefore, levelplots with scaled data and unscaled data are different.

5. (1) Hierarchically cluster (dendrogram) the USArrests data using complete linkage after scaling the data. Show this dendrogram as figure Dendrogram 2



U <- USArrests
Uscale <- scale(U)
Uscaledist <- dist(Uscale)
Uscaledend <- hclust(Uscaledist,method="complete")
plot(Uscaledend)

(2) Cut the dendrogram into 4 clusters and produce a point plot with the x-axis as Murder, the y-axis as Assault, and colour each point by cluster number. Label this plot as Figure Cluster.

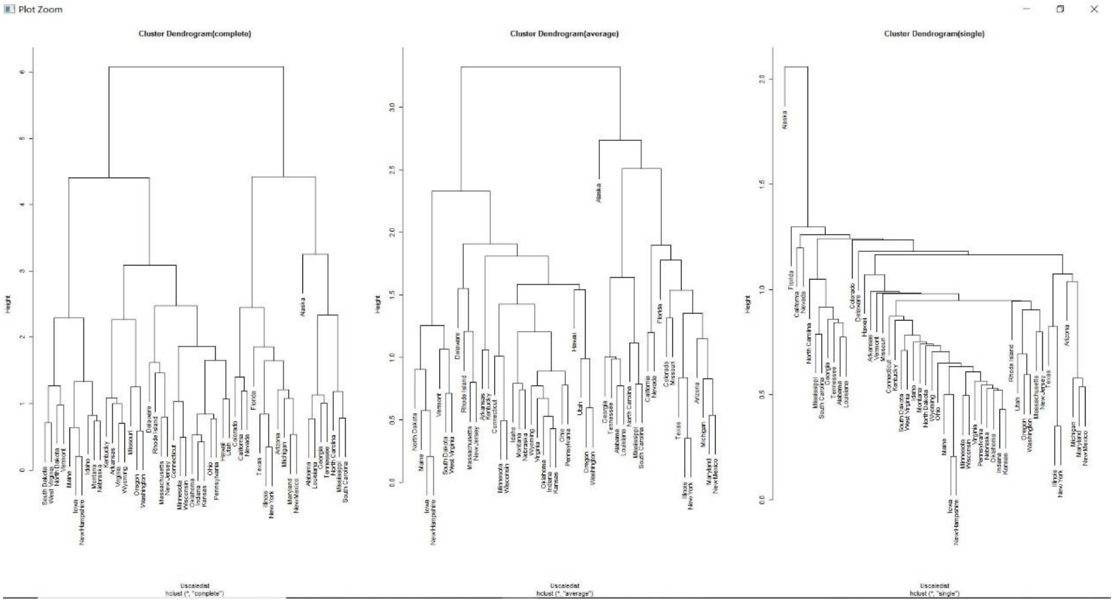


clusters <- cutree(Uscaledend, k = 4)
plot(U$Murder, U$Assault, main = "Figure Cluster", xlab = "Murder", ylab = "Assault", col = clusters, cex = 2)

```
text(U$Murder, U$Assault, rownames(U), cex=0.6)
```

(3) Explain why (or why not) you observe patterns in this plot.
I observe patterns in this plot because after splitting the dendrogram into 4 clusters and plotting them, it can be clearly seen the distribution of these states and values of Murder and Assault for each state.

6. (1) Create 3 dendrograms using the scaled data from the previous question with the agglomeration methods single, complete and average. Produce a figure showing each of the dendrograms.



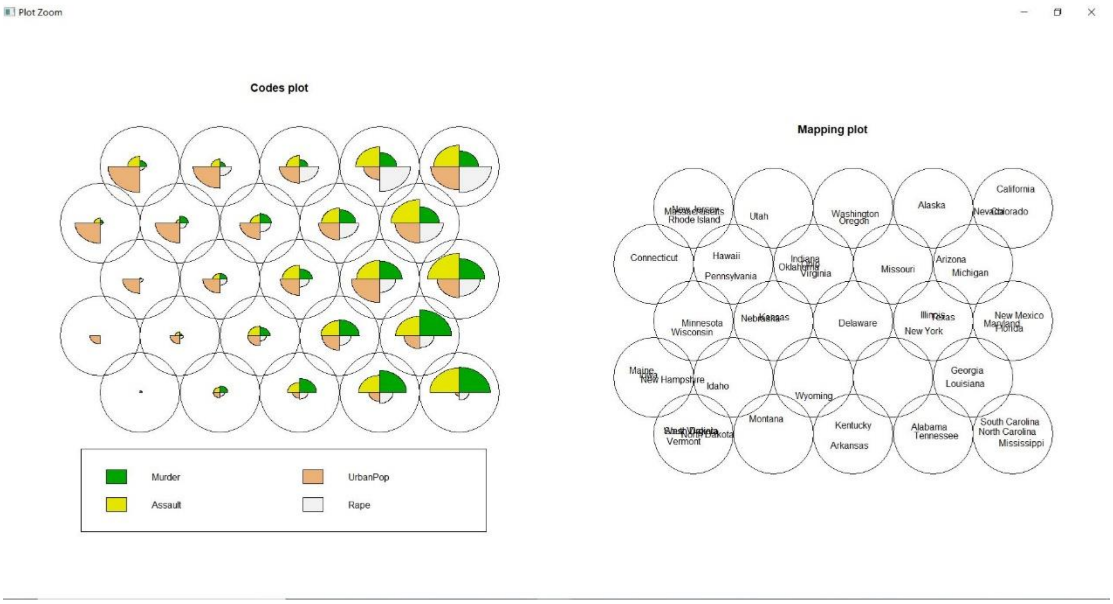
```
par(mfrow = c(1,3))
U <- USArrests
Uscale <- scale(U)
Uscaledist <- dist(Uscale)
Uscaledend <- hclust(Uscaledist,method="complete")
plot(Uscaledend, main = "Cluster Dendrogram(complete)")

Uscaledend <- hclust(Uscaledist,method="average")
plot(Uscaledend, main = "Cluster Dendrogram(average)")

Uscaledend <- hclust(Uscaledist,method="single")
plot(Uscaledend, main = "Cluster Dendrogram(single)")
```

(2) explain why they differ – in other words, what is the agglomeration method doing and why does this alter the dendrogram?
There are three linkage methods which are used in the hierarchical clustering. They are different. Average uses the mean of intercluster dissimilarity, which means that it calculates all pairwise dissimilarities between the observations of two clusters and then records the mean of these dissimilarities. Complete

uses the maximal intercluster dissimilarity, which means that it calculates all pairwise dissimilarities between the observations of two clusters and then records the largest of these dissimilarities. Single uses the minimal intercluster dissimilarity, which means that it calculates all pairwise dissimilarities between the observations of two clusters and then records the smallest of these dissimilarities.

It can be concluded that they use different methods to calculate the dissimilarity. Therefore, it may change the components in one cluster and may also change the distribution of various clusters or the height of two clusters' connection node.

7. (1) Create a self-organising map (SOM) using the scaled USArrests data. (HINT: create a 5x5 hexagonal SOM). Create two visualisations of the SOM – one that shows the relative values of each prototype vector, and one that is labelled with the state names.



```
library(kohonen)
U <- USArrests
par(mfrow = c(1,2))
Uscale <- scale(U)
Uscalesom <- som(as.matrix(Uscale),grid = somgrid(5,5,"hexagonal"))
plot(Uscalesom)
plot(Uscalesom,type="mapping", labels=row.names(Uscale))
```

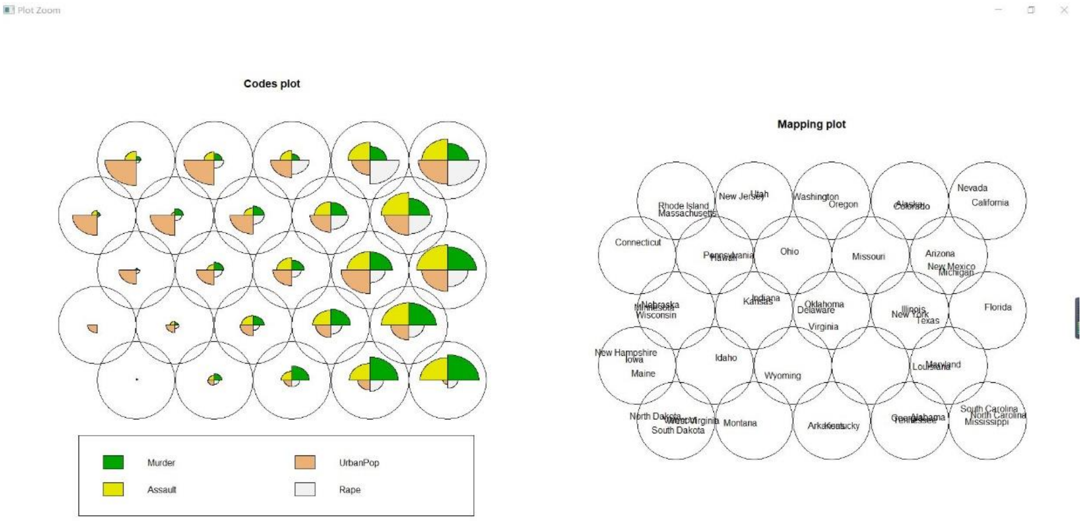
(2) What does the prototype vector tell you about the important variables used to cluster around New York versus Georgia?

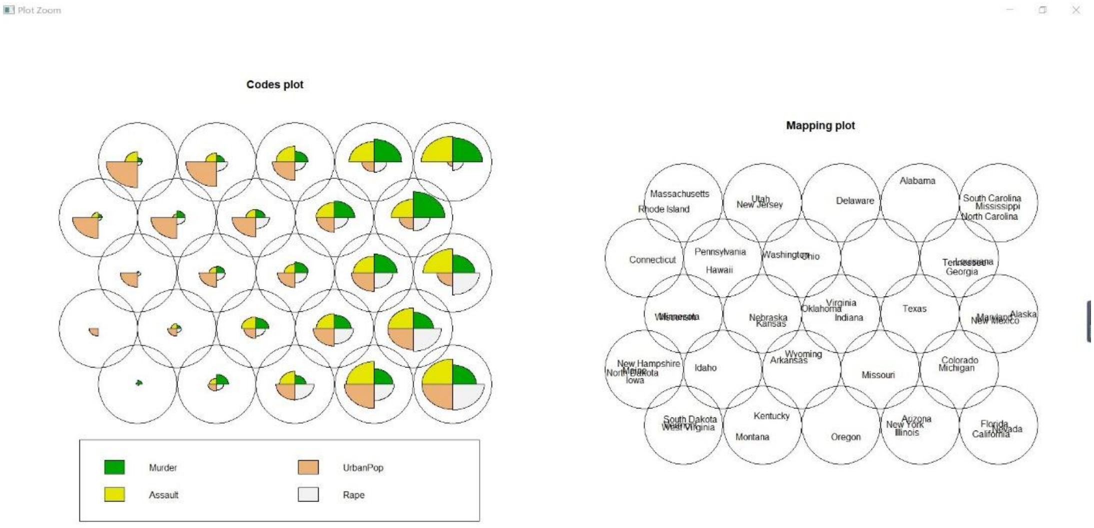
By visualising the prototype vectors across the map, it can be seen the distribution of samples and variables. To be more precise, these fan diagrams in each node visualise the prototype vectors and we can easily find that Murder and UrbanPop are two important variables used to cluster around

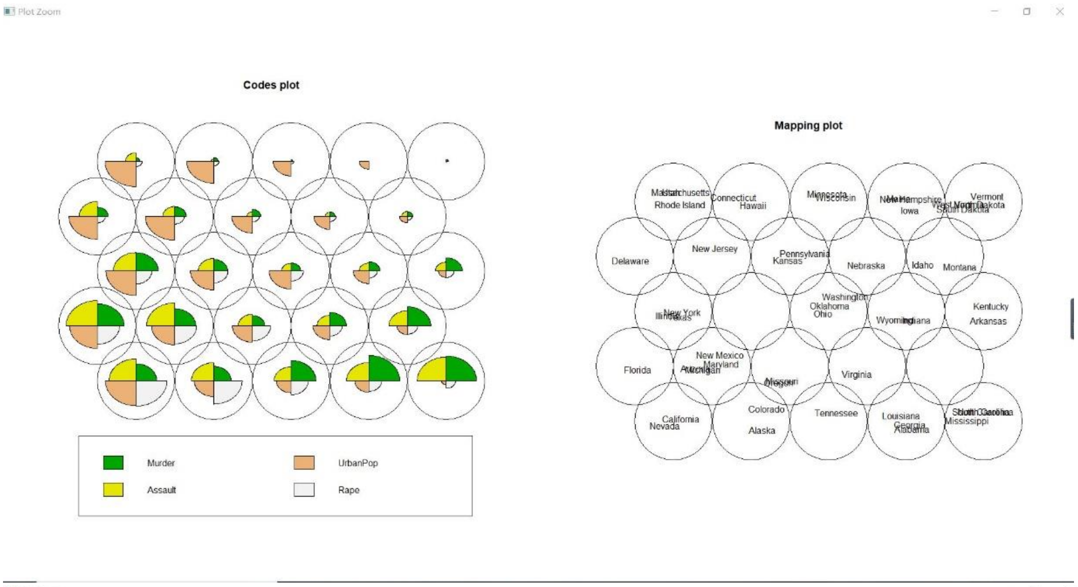
New York versus Georgia. Compared with Georgia, New York has lower Murder and higher UrbanPop. The rest of variables, Assault and Rape are similar for these two places.

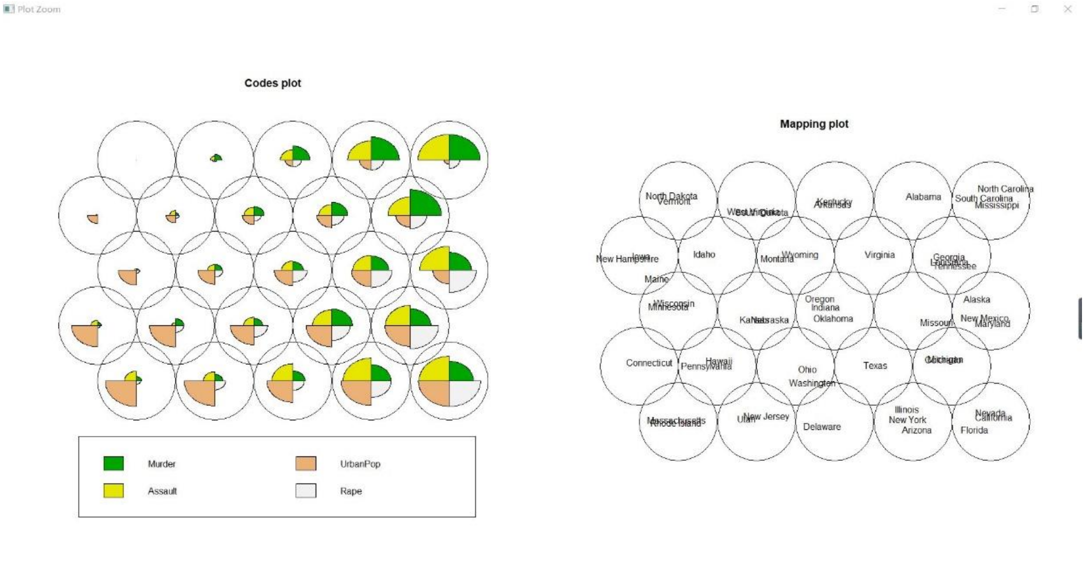
(3) Which countries appear to be similar to Georgia for the SOM, and do they agree with the dendrogram?
From this SOM, Louisiana seems to be the most similar state to Georgia. We can see that, for all three dendrograms above, Louisiana is not the most similar one, but it is still true that the Louisiana is quite similar to Georgia in dendrograms.

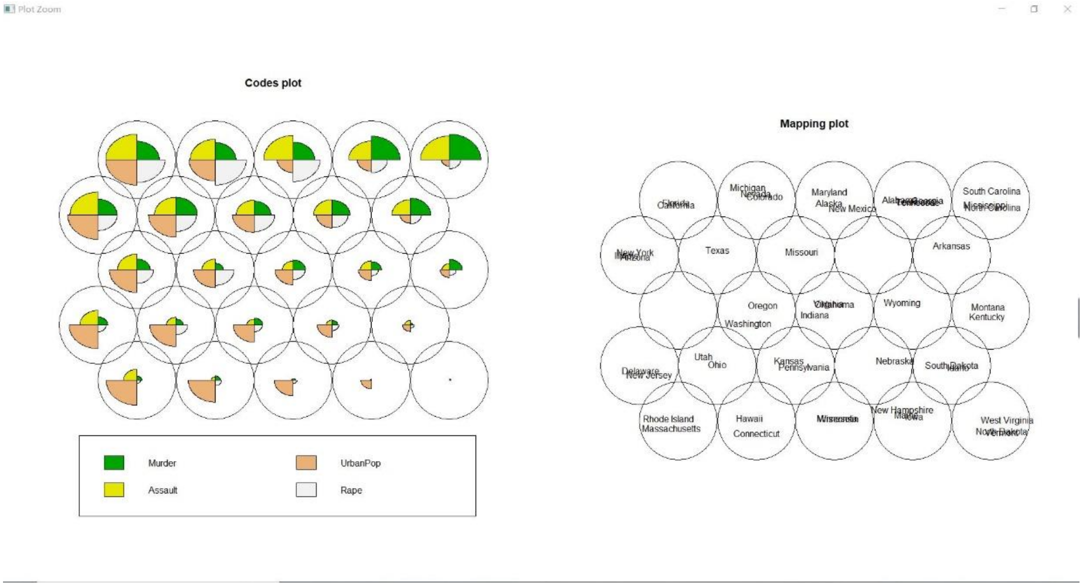
8. Run the SOM a few times – do you always get the same map when you plot? Comment on whether you prefer the dendrogram or SOM representation for clustering and why.






Plot Zoom
Codes plot
Murder
Assault
UrbanPop
Rape
Mapping plot
Massachusetts
Rhode Island
Connecticut
Hawaii
Minnesota
Wisconsin
New Hampshire
Iowa
Vermont
North Dakota
South Dakota
West Virginia
Delaware
New Jersey
Pennsylvania
Kansas
Nebraska
Idaho
Montana
Illinois
New York
Texas
Washington
Oklahoma
Ohio
Wyoming
Indiana
Kentucky
Arkansas
Florida
New Mexico
Maryland
Arizona
Michigan
Missouri
Oregon
Virginia
California
Nevada
Colorado
Alaska
Tennessee
Louisiana
Georgia
Alabama
South Carolina
North Carolina
Mississippi


Plot Zoom
Codes plot
Murder
Assault
UrbanPop
Rape
Mapping plot
North Dakota
Vermont
West Virginia
South Dakota
Kentucky
Arkansas
Alabama
North Carolina
South Carolina
Mississippi
New Hampshire
Iowa
Maine
Idaho
Montana
Wyoming
Virginia
Georgia
Louisiana
Tennessee
Wisconsin
Minnesota
Kansas
Nebraska
Oregon
Indiana
Oklahoma
Missouri
Alaska
New Mexico
Maryland
Connecticut
Pennsylvania
Hawaii
Ohio
Washington
Texas
Colorado
Michigan
Massachusetts
Rhode Island
Utah
New Jersey
Delaware
Illinois
New York
Arizona
Nevada
California
Florida


Plot Zoom
Codes plot
Murder
Assault
UrbanPop
Rape
Mapping plot
Florida
California
Michigan
Nevada
Colorado
Maryland
Alaska
New Mexico
Alabama
Georgia
Tennessee
South Carolina
Mississippi
North Carolina
New York
Arizona
Texas
Missouri
Arkansas
Oregon
Washington
Oklahoma
Indiana
Wyoming
Montana
Kentucky
Delaware
New Jersey
Utah
Ohio
Kansas
Pennsylvania
Nebraska
South Dakota
Idaho
Rhode Island
Massachusetts
Hawaii
Connecticut
Minnesota
Wisconsin
New Hampshire
Maine
Iowa
West Virginia
North Dakota
Vermont

I run the SOM many times and the results for each time is slightly different. I prefer SOM representation for clustering because it can be clearly seen not only clusters with their components but also the vectors which are used to cluster them.

References:
lecture 2 slides
lecture 3 slides
lecture 4 slides
https://www.r-bloggers.com/self-organising-maps-for-customer-segmentation-using-r/
An Introduction to Statistical Learning