Name: Usman Shah
Student ID: 1006293

Data Analysis, Clustering and Visualisation

TIME SERIES DATA:THE SP500 INDEX:

1. The candles in the upper part of the chart shows the market's open, high, close and low price for each single day. The real body of the candles is the wider area in the middle. Which represents the price range between the open and close of the day's trading.
   The green candlesticks mean that on a particular day the opening price was lower than the closing price. Which means that price moved up on that day. The orange candlesticks mean that the opening price was higher than the closing price. Which means that the price went down on that day.
   For example, in the following chart on 03-Sep-2009, it has a green candle, so the price went up on that day. And on 01-Sep-2009 the orange colour indicates that price went down by quite a bit.

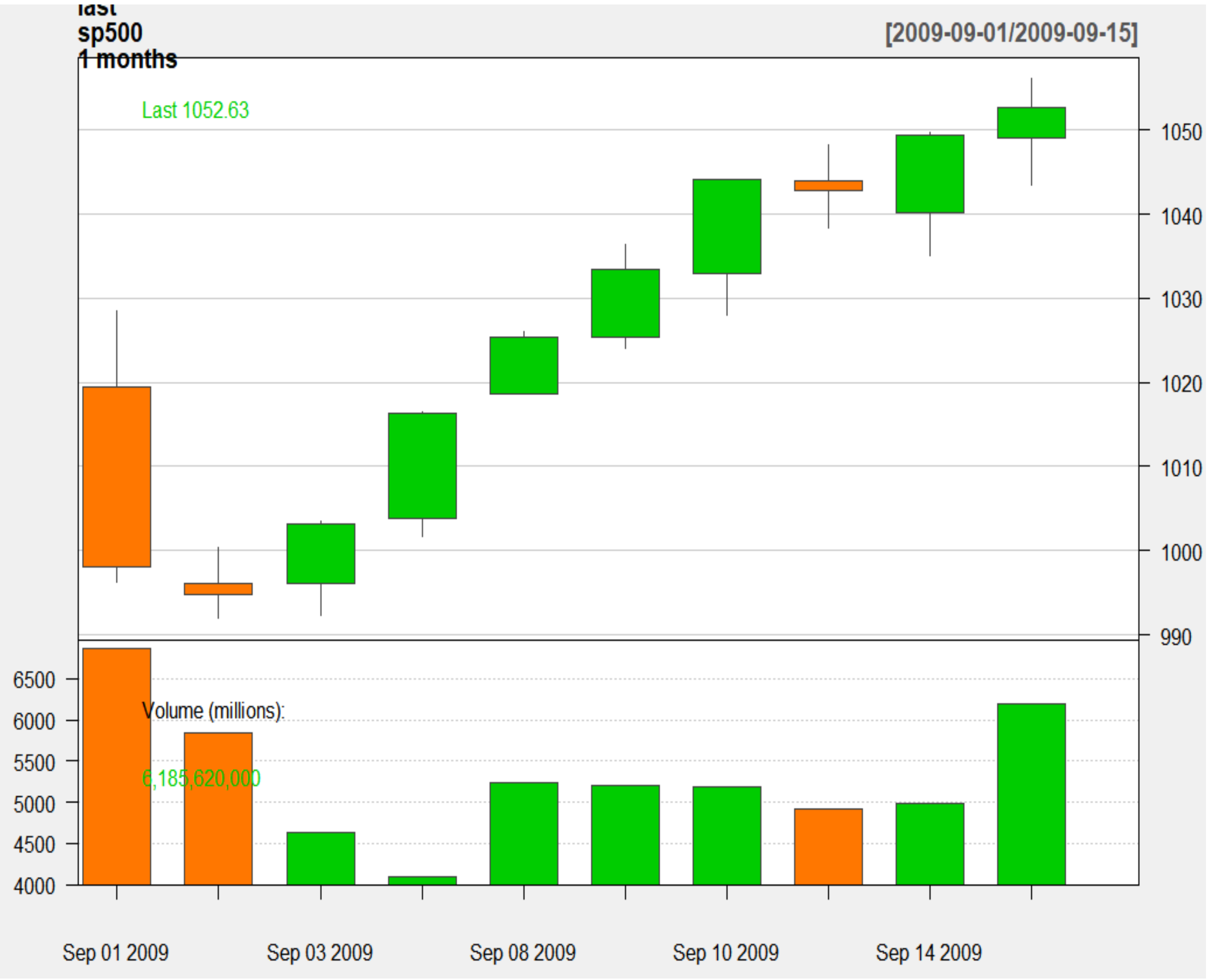
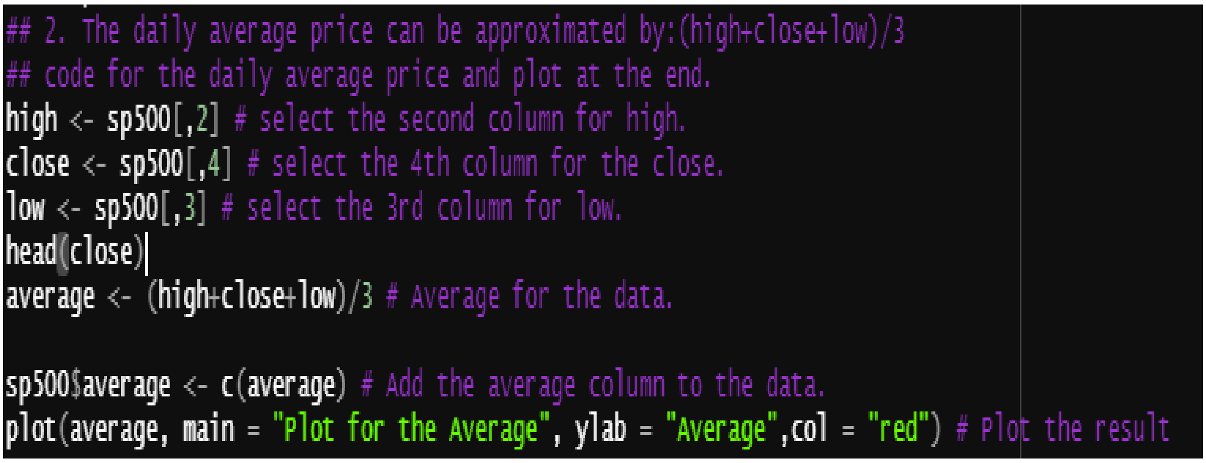


Figure 1 Candle plot

Name: Usman Shah
Student ID: 1006293

2. R code use to calculate average price Av(t) and plot showing average price Av(t):



```
## 2. The daily average price can be approximated by:(high+close+low)/3
## code for the daily average price and plot at the end.
high <- sp500[,2] # select the second column for high.
close <- sp500[,4] # select the 4th column for the close.
low <- sp500[,3] # select the 3rd column for low.
head(close)
average <- (high+close+low)/3 # Average for the data.

sp500$average <- c(average) # Add the average column to the data.
plot(average, main = "Plot for the Average", ylab = "Average",col = "red") # Plot the result
```

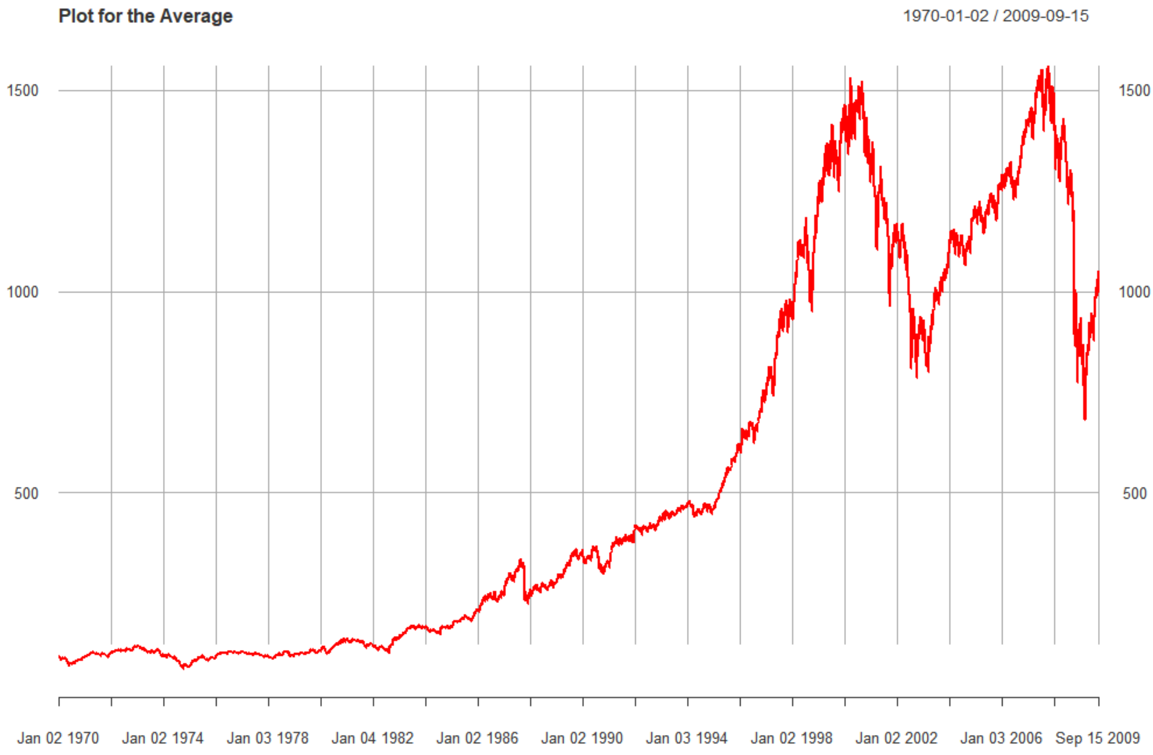
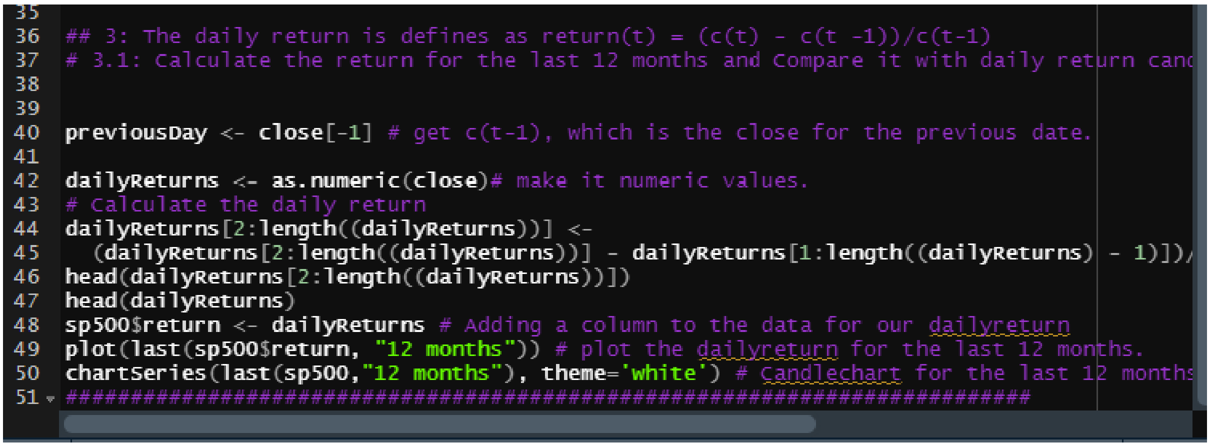


Figure 2 Plot showing average price

Name: Usman Shah
Student ID: 1006293

3. Below figure 3 is the R code for calculating and plotting the return. Figure 4 is the plot for last 12 months return and figure 5 is the candle chart for last 12 months. We can see that that 01-oct-2020 the market started with a very high price and that is why the return for that is very low. I guess its natural than when the price is very high people will not buy products and the return will be very low. At the end of October, the prices are very low and that is why there is a spike in the daily returns, I think that is when the stock market crashed in 2008. From October till March there is quite a bit of change in the price signal by going up and price dropping down and that is why there are big spike and drops in the average return, which shows the market is not stable.
   After March the market is on the way back to normal and there is not big difference in the price signal but a very gradual change and that is why the average return for that time do not have big spikes and drops.



```
## 3: The daily return is defines as return(t) = (c(t) - c(t -1))/c(t-1)
# 3.1: Calculate the return for the last 12 months and Compare it with daily return canc


previousDay <- close[-1] # get c(t-1), which is the close for the previous date.

dailyReturns <- as.numeric(close)# make it numeric values.
# Calculate the daily return
dailyReturns[2:length((dailyReturns))] <-
  (dailyReturns[2:length((dailyReturns))] - dailyReturns[1:length((dailyReturns) - 1)])/
head(dailyReturns[2:length((dailyReturns))])
head(dailyReturns)
sp500$return <- dailyReturns # Adding a column to the data for our dailyreturn
plot(last(sp500$return, "12 months")) # plot the dailyreturn for the last 12 months.
chartSeries(last(sp500,"12 months"), theme='white') # Candlechart for the last 12 months
##########################################################################
```

Figure 3 R Code for Daily Return

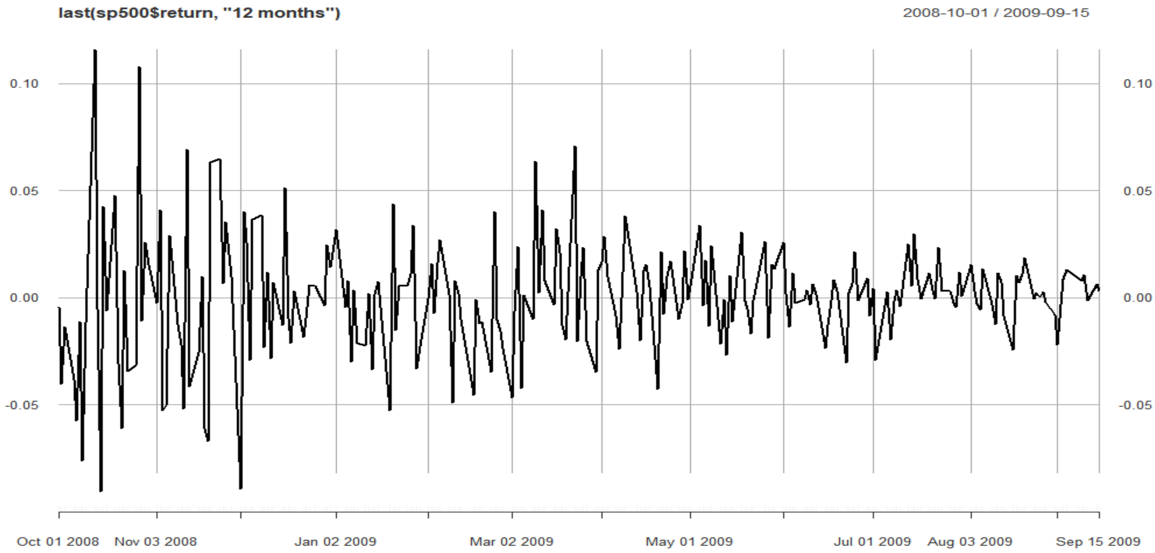


Figure 4 Return for last 12 months.

Name: Usman Shah
Student ID: 1006293

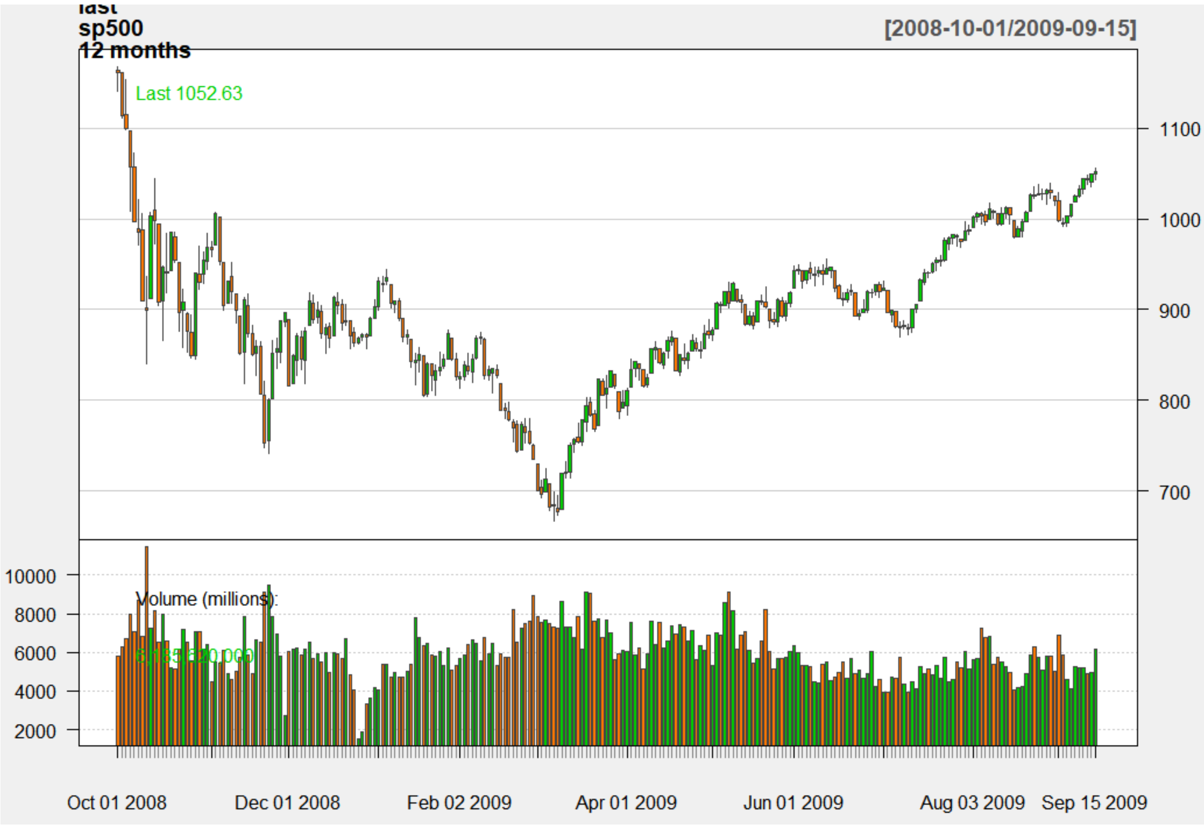


Figure 5 candleChart for last 12 months.

Figure 6 R code for daily return

Degree Distribution : Networks ->

1. The maximum degree for the network is 183 and the minimum is 5.
2. The first plot on the left side of the figure 6 shows the degree distribution of a scale free network of degree. Most of them have a very small frequency. The histogram plot shows that only small number of nodes have a very high frequency means that they have quite a few edges connected to them and at the end there is couple nodes with a very low frequency means they only have a couple links. The line plot sorted by degree shows on how the lower degree (nodes) have so many edges (index) connected to them. Shows that around degree 5 has very high edges connected.

Name: Usman Shah
Student ID: 1006293

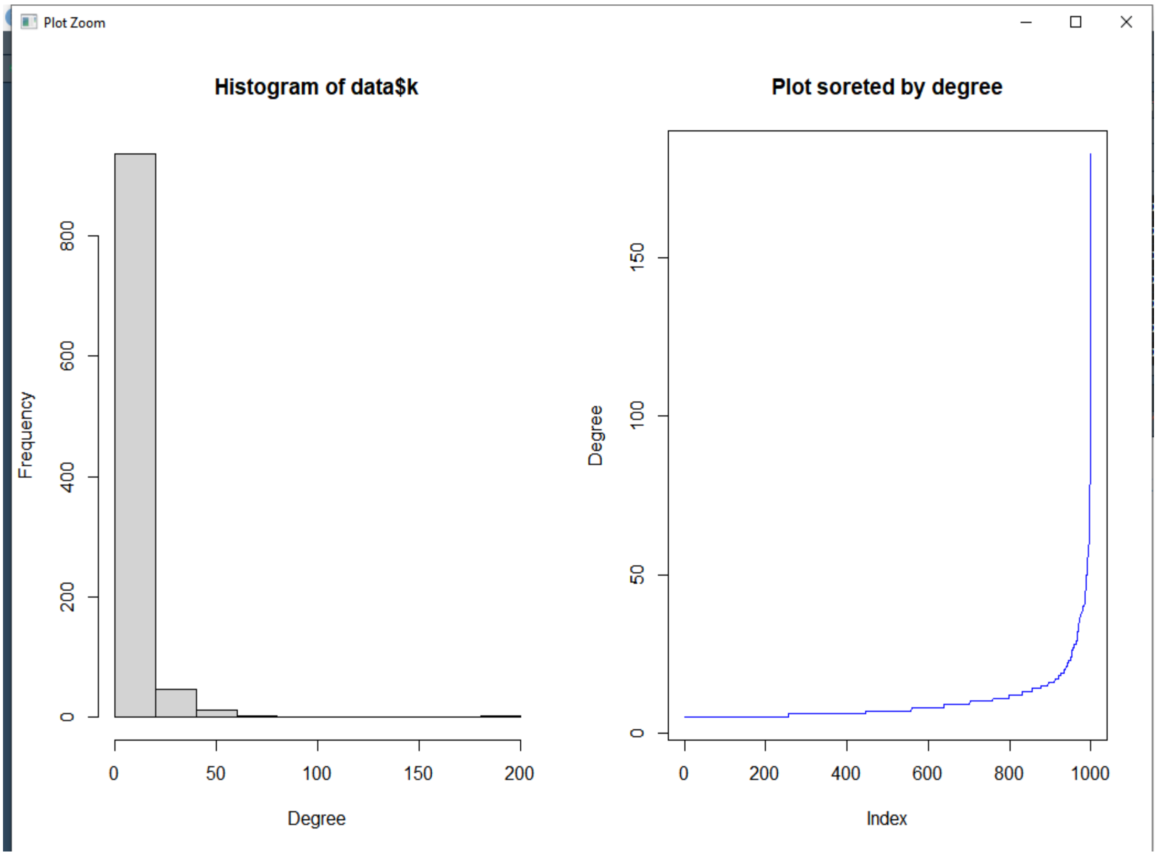
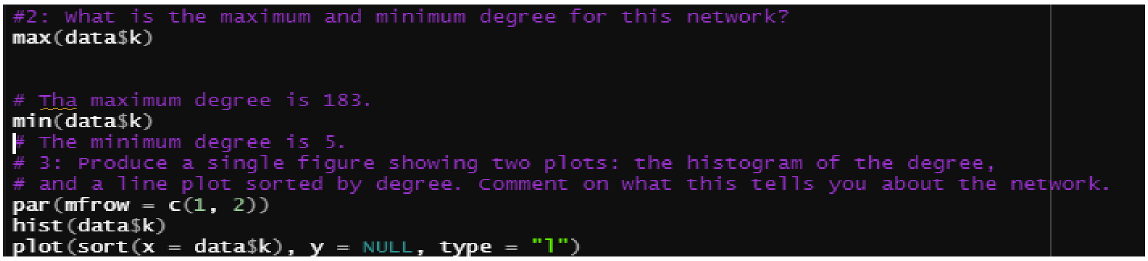


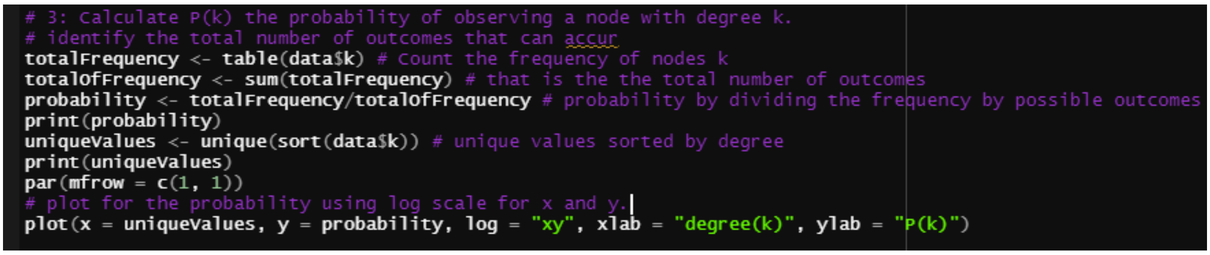
Figure 6 Single figure with two plots.



```
#2: what is the maximum and minimum degree for this network?
max(data$k)


# Tha maximum degree is 183.
min(data$k)
# The minimum degree is 5.
# 3: Produce a single figure showing two plots: the histogram of the degree,
# and a line plot sorted by degree. Comment on what this tells you about the network.
par(mfrow = c(1, 2))
hist(data$k)
plot(sort(x = data$k), y = NULL, type = "l")
```

Figure 7 R Code to calculate the max and min degree

3. R code for calculating p(k) which is the probability of observing a node with degree k and the plot of p(k) versus.



```
# 3: Calculate P(k) the probability of observing a node with degree k.
# identify the total number of outcomes that can accur
totalFrequency <- table(data$k) # Count the frequency of nodes k
totalOfFrequency <- sum(totalFrequency) # that is the the total number of outcomes
probability <- totalFrequency/totalOfFrequency # probability by dividing the frequency by possible outcomes
print(probability)
uniqueValues <- unique(sort(data$k)) # unique values sorted by degree
print(uniqueValues)
par(mfrow = c(1, 1))
# plot for the probability using log scale for x and y.
plot(x = uniqueValues, y = probability, log = "xy", xlab = "degree(k)", ylab = "P(k)")
```

Figure 8 R Code for P(K)

Name: Usman Shah
Student ID: 1006293

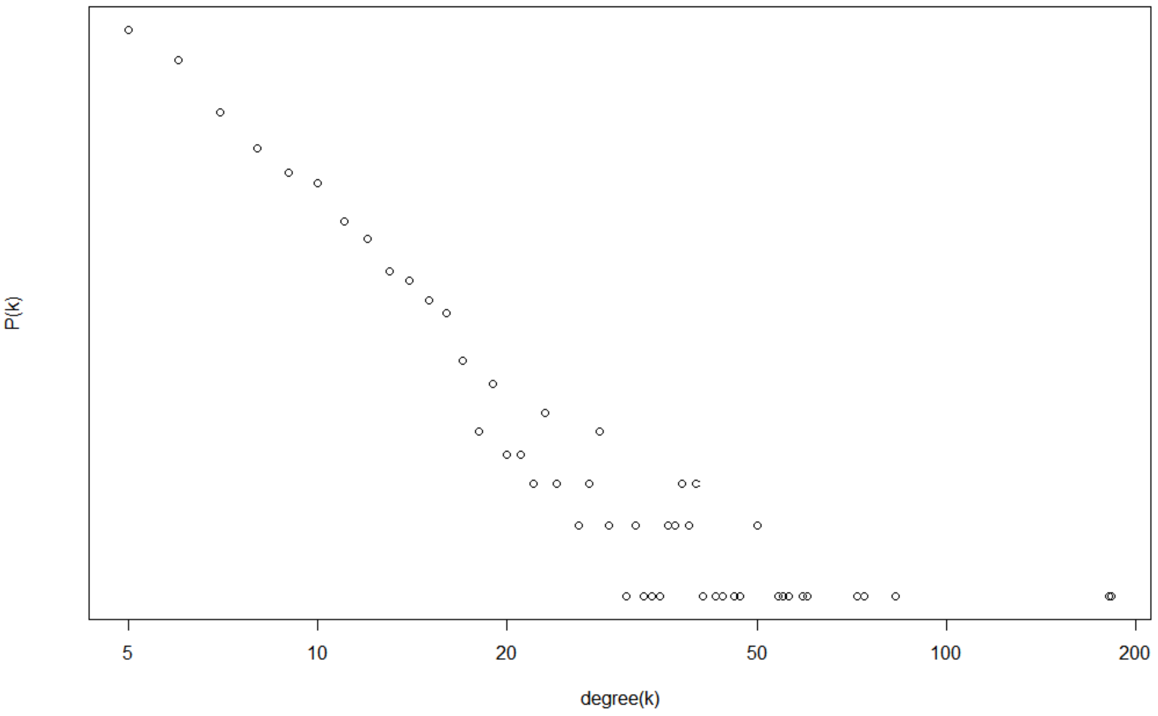


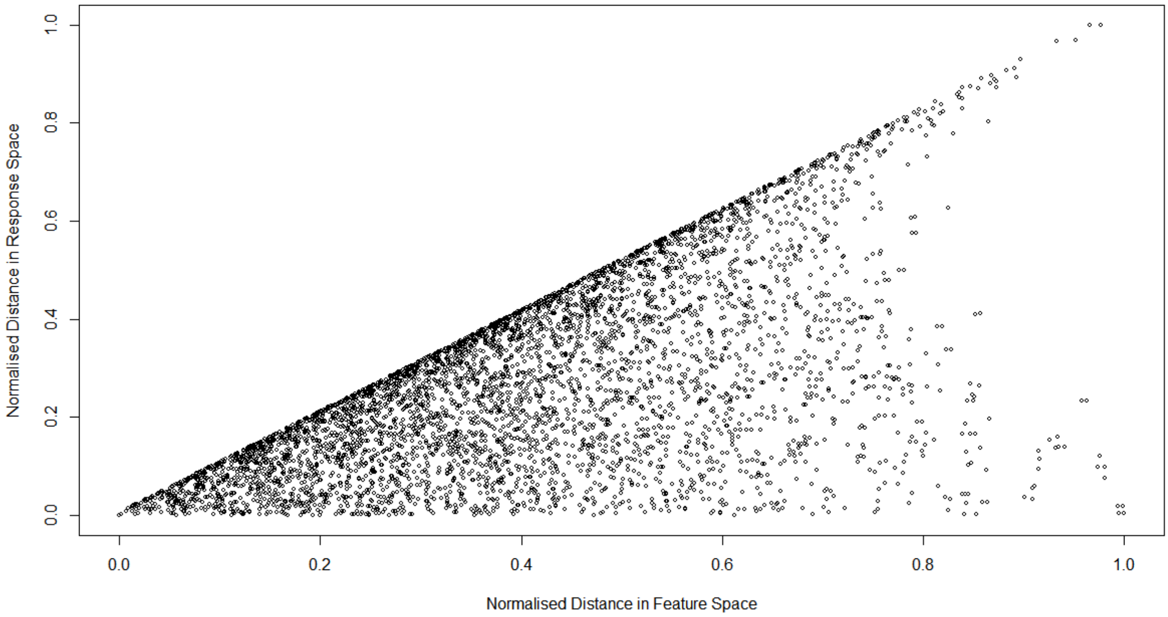
Figure 9 P(k) versus K

The plot above shows that the network is likely to be scale free (with a power law degree distribution) because nodes that already have high number of edges are more likely to see new edges connected to them compared with the nodes with a lower number of edges. It means that when the network grows the underlining structure or properties stays the same. That is the characteristics of this network (Scale free network), they are independent of the size of the network. In the scatter plot above points tend to fall in line and gets a bit messy for large degree at the end of the line.
For example, for a celebrity to establish a connection is very easy as compared to a normal person.

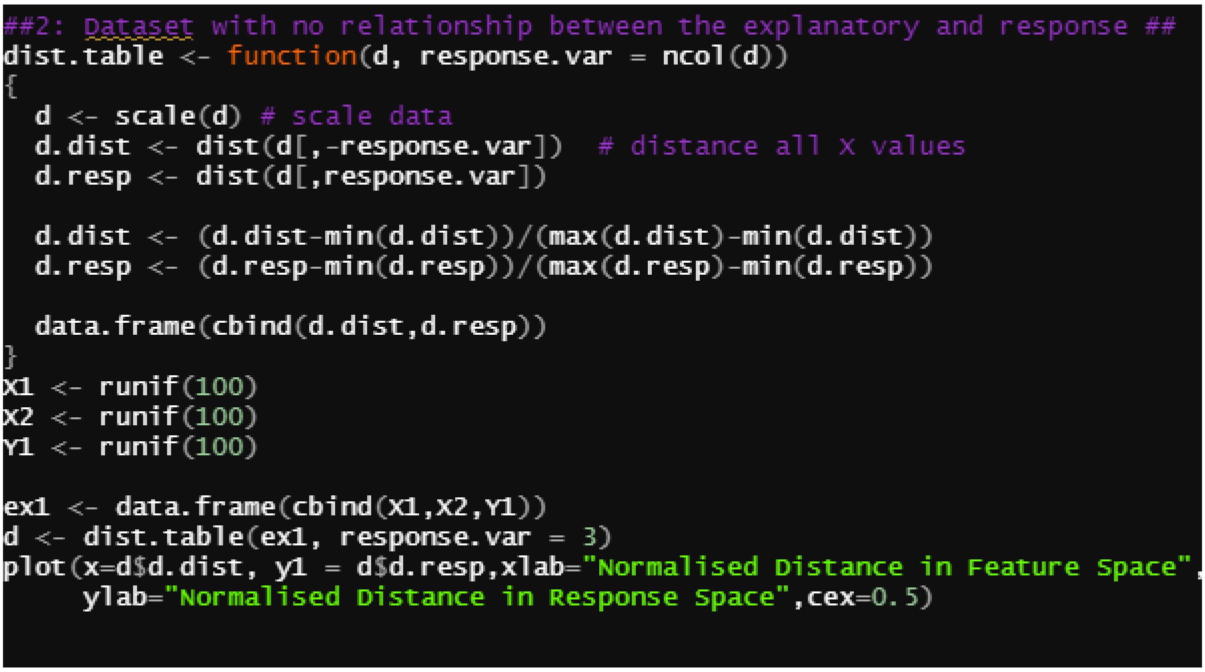
DATA QUALITY AND STRUCTURE:

1. The plot represents the explanatory variables x1, x2 on the x-axis and the response variable y on the y-axis. The plot shows that most of the data is pretty close to the centre middle line and with change in the explanatory variables, there is a change in the response variable. This means that there is a relationship between the explanatory and response variables and that x1 and x2 are good explanatory variables. And is useful method for examining the quality of the dataset.

Name: Usman Shah
Student ID: 1006293



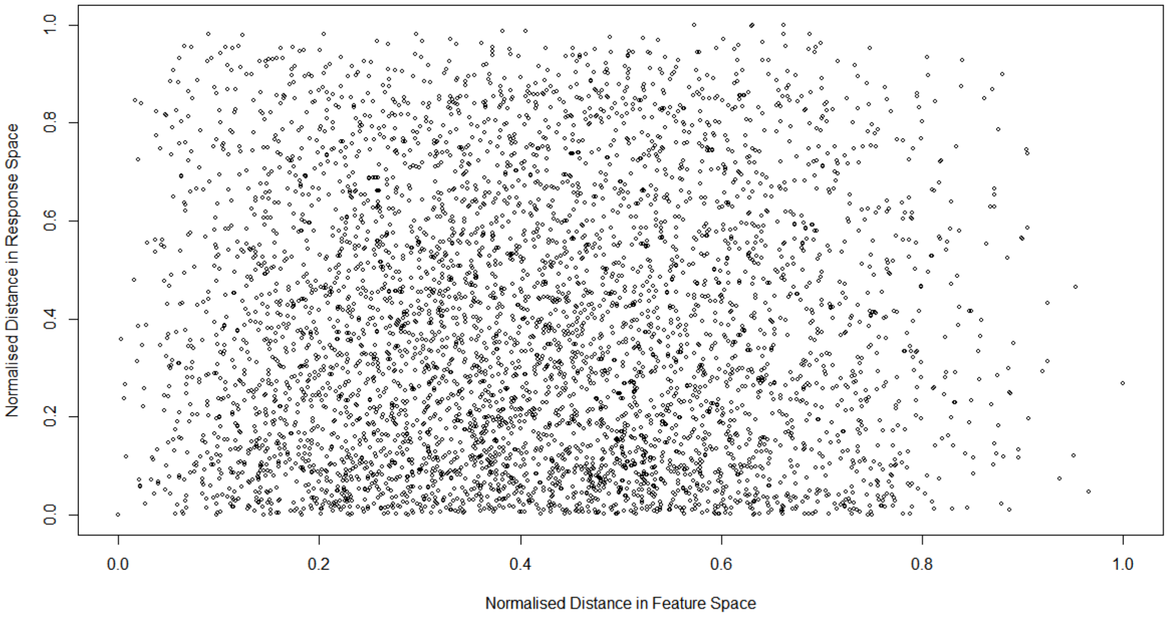
2. The resulting plot shows that x1 and x2 are not good explanatory variables to predict y1 (response variable). The data is everywhere and there is no linear change in the response variable by changing the explanatory variable, which shows that x1 , x2 has no effect on the response variable and that is why are not good explanatory variables and is not a useful method for examining the quality of the dataset.



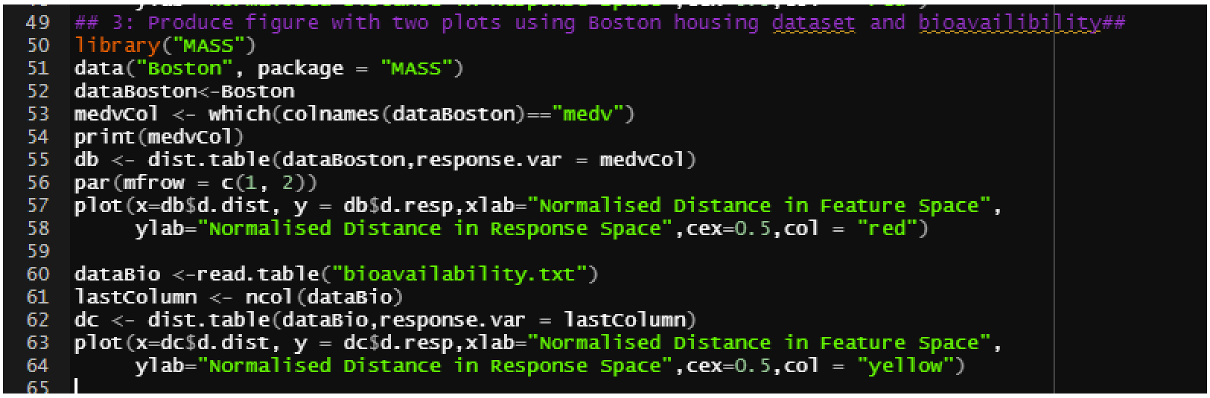
```
##2: Dataset with no relationship between the explanatory and response ##
dist.table <- function(d, response.var = ncol(d))
{
  d <- scale(d) # scale data
  d.dist <- dist(d[,-response.var])  # distance all X values
  d.resp <- dist(d[,response.var])

  d.dist <- (d.dist-min(d.dist))/(max(d.dist)-min(d.dist))
  d.resp <- (d.resp-min(d.resp))/(max(d.resp)-min(d.resp))

  data.frame(cbind(d.dist,d.resp))
}
X1 <- runif(100)
X2 <- runif(100)
Y1 <- runif(100)

ex1 <- data.frame(cbind(X1,X2,Y1))
d <- dist.table(ex1, response.var = 3)
plot(x=d$d.dist, y1 = d$d.resp,xlab="Normalised Distance in Feature Space",
     ylab="Normalised Distance in Response Space",cex=0.5)
```

Figure 7 R Code for producing a plot without a relationship between explanatory and response variables.

Name: Usman Shah
Student ID: 1006293



3. From figure 13 below I believe that Boston dataset will be easier to model because the distance between the data points is not as big compared to bioavailability. You can look at Boston data set as one cluster while bioavailability it is hard to say how many clusters are there.
   The distance function is used to create a distance matrix of our data. That describes the distance between each example row. The idea behind k nearest neighbour is that it computes the distance between the points using Euclidean distances and the point with less distance are grouped with that cluster. Both are related as both methods are using distance as a measure to decide.



```
     ylab="Normalised Distance in Response Space",cex=0.5,col = "red")
## 3: Produce figure with two plots using Boston housing dataset and bioavailibility##
library("MASS")
data("Boston", package = "MASS")
dataBoston<-Boston
medvCol <- which(colnames(dataBoston)=="medv")
print(medvCol)
db <- dist.table(dataBoston,response.var = medvCol)
par(mfrow = c(1, 2))
plot(x=db$d.dist, y = db$d.resp,xlab="Normalised Distance in Feature Space",
     ylab="Normalised Distance in Response Space",cex=0.5,col = "red")

dataBio <-read.table("bioavailability.txt")
lastColumn <- ncol(dataBio)
dc <- dist.table(dataBio,response.var = lastColumn)
plot(x=dc$d.dist, y = dc$d.resp,xlab="Normalised Distance in Feature Space",
     ylab="Normalised Distance in Response Space",cex=0.5,col = "yellow")
```

Figure 8 R Code using Boston and bioavailability data

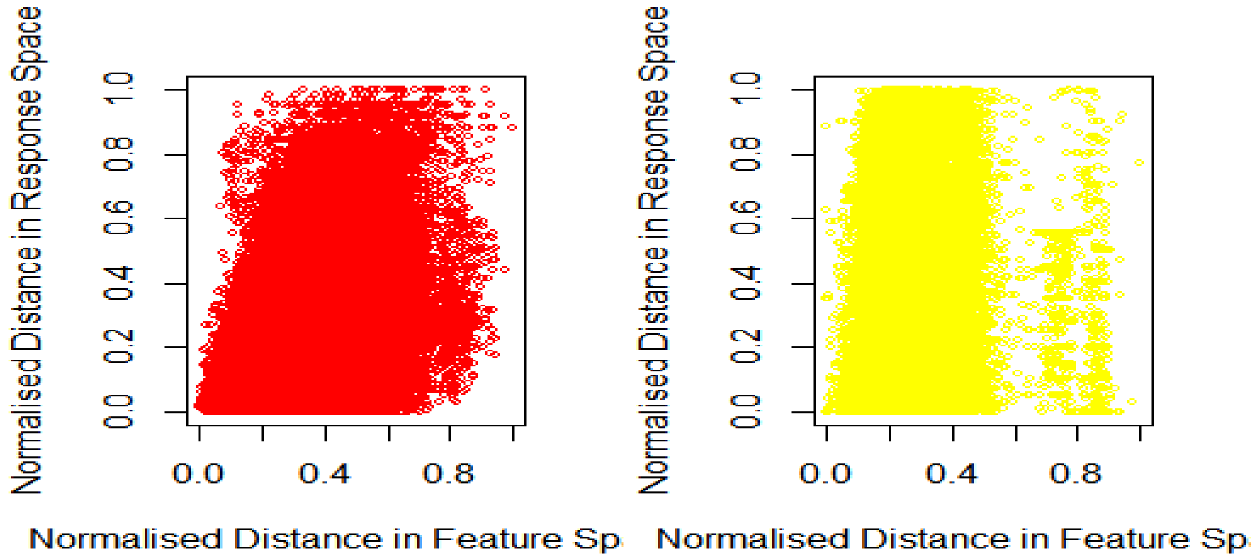
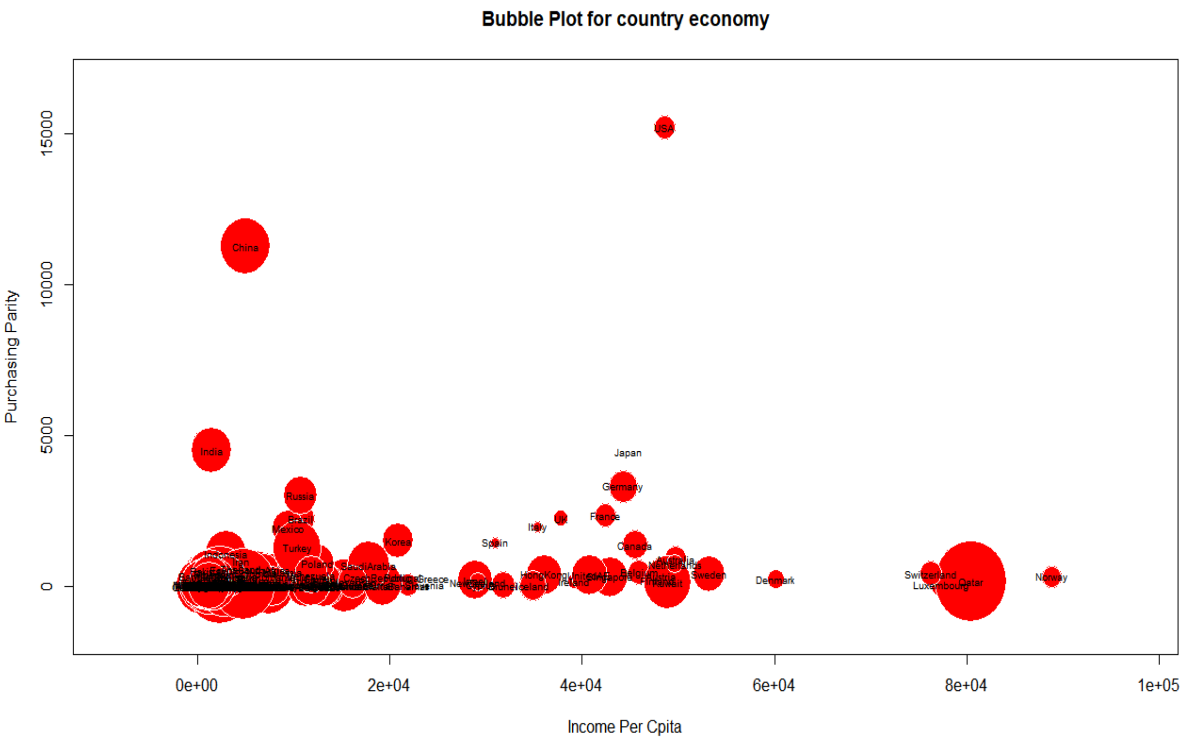


Figure 9 Boston vs bioavailability

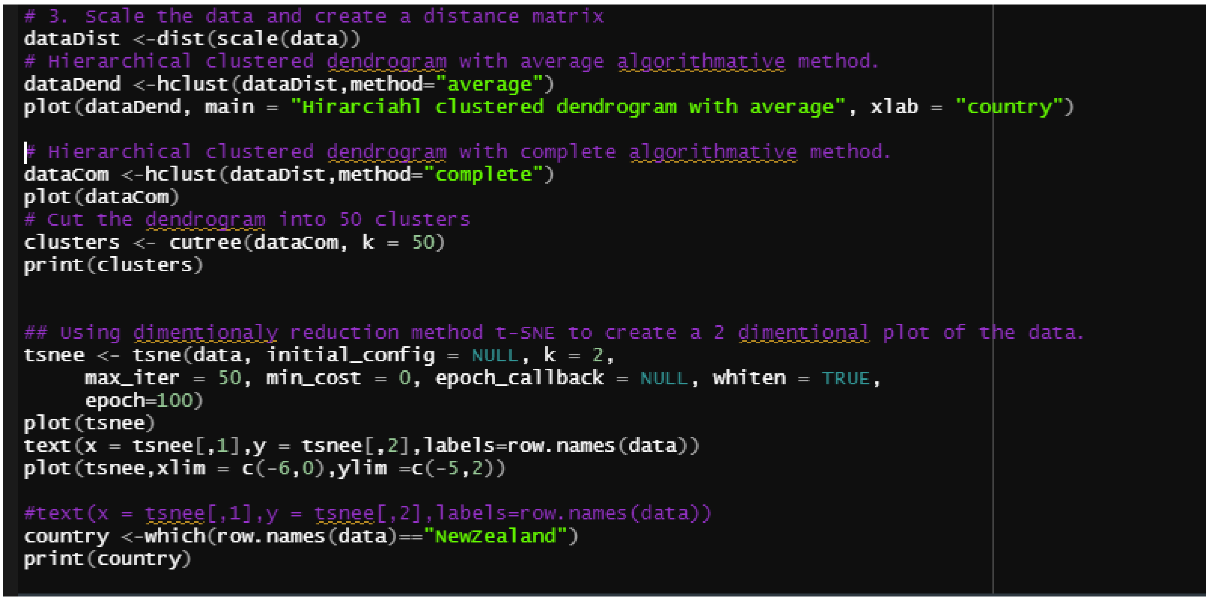
VISUALISATION AND CLUSTERING :

1. Breifly state the meaning of each variable for the data.
   PopDensity = is the measurement of population. In other words how dense an area is term of it population.
   IncomeperCapital = That is the average income earned per person in a given area.
   Purchasing parity = is the measurement of prices in differnt countries.
   changeGDP = is to measure or make an assessment on how the economy of a country is doing.
2. To examine which country is similar and which country is strongest or weakest. I have taken the changeGDP into account which is the measure of how a country is doing economically. I have considered the incomePerCapit and purchasingParity to see which country has high income per person and purchasing parity and what is its effect on the economy of the country. The plot in the figure below shows incomePerCapita on x-axis and purchasing parity on y-axis and the bubble is sized in term of economy. We can see from the plot below that the purchasing parity for China is very high but income per person is not that high and there economy is very good means that the purchasing parity has a big effect on the economy of the country. But on the other hand if you look at Qatar it has a vey low purchasing parity but a very high income per person and has a good economy.
   To answer the question Iceland, Brunei, Switzerland, Norway and Denmark are similar to New Zealand in terms of the economy.
   Italy, UK etc are very weak and countries like Qatar and China have strong economy.

Name: Usman Shah
Student ID: 1006293



3. There are four most used types of linkage in hierarchical clustering.
The following two figures shows two of these types Complete and Average. They are different from each other because in Complete linkage the largest dissimilarities are taken into account for clustering dendrograms while in Average linkage the mean or average of the dissimilarities is recorded to obtain hierarchical clustering dendrogram.
Output from R code at in Figure 17 shows that Brunei, Cyprus, Iceland, Ireland and New Zealand are all in the same cluster based on cutting the dendrogram into 50 clusters.



```
# 3. Scale the data and create a distance matrix
dataDist <-dist(scale(data))
# Hierarchical clustered dendrogram with average algorithmative method.
dataDend <-hclust(dataDist,method="average")
plot(dataDend, main = "Hirarciahl clustered dendrogram with average", xlab = "country")

# Hierarchical clustered dendrogram with complete algorithmative method.
dataCom <-hclust(dataDist,method="complete")
plot(dataCom)
# Cut the dendrogram into 50 clusters
clusters <- cutree(dataCom, k = 50)
print(clusters)


## Using dimentionaly reduction method t-SNE to create a 2 dimentional plot of the data.
tsnee <- tsne(data, initial_config = NULL, k = 2,
     max_iter = 50, min_cost = 0, epoch_callback = NULL, whiten = TRUE,
     epoch=100)
plot(tsnee)
text(x = tsnee[,1],y = tsnee[,2],labels=row.names(data))
plot(tsnee,xlim = c(-6,0),ylim =c(-5,2))

#text(x = tsnee[,1],y = tsnee[,2],labels=row.names(data))
country <-which(row.names(data)=="NewZealand")
print(country)
```

Figure 10 R code for Complete, average dendrograms and cutting the tree into 50 clusters and t-SNE to create two dimensional plot for the country data.

Name: Usman Shah
Student ID: 1006293

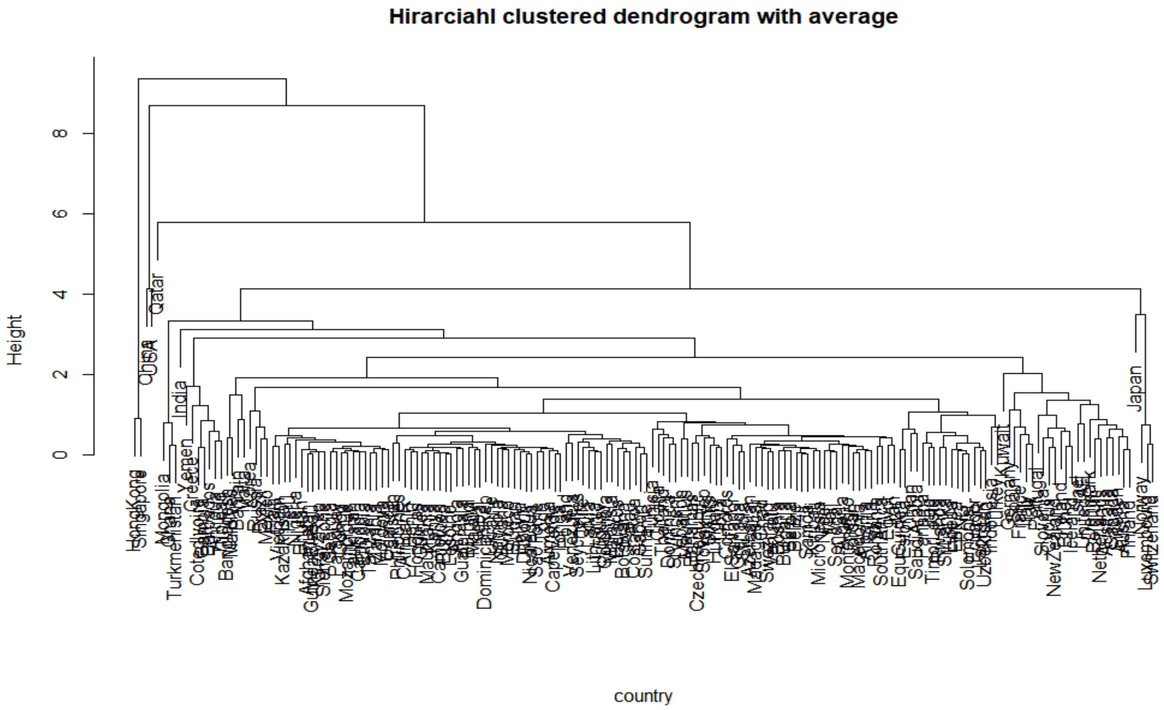


Figure 15 Dendrogram with average

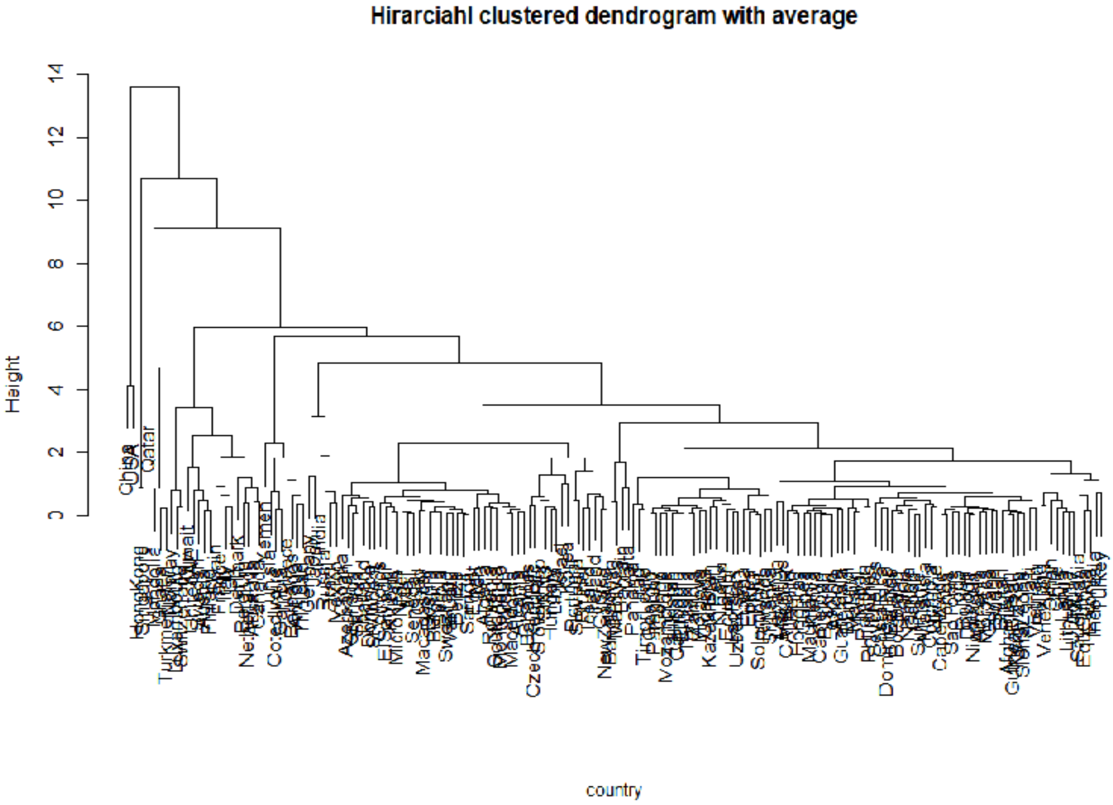
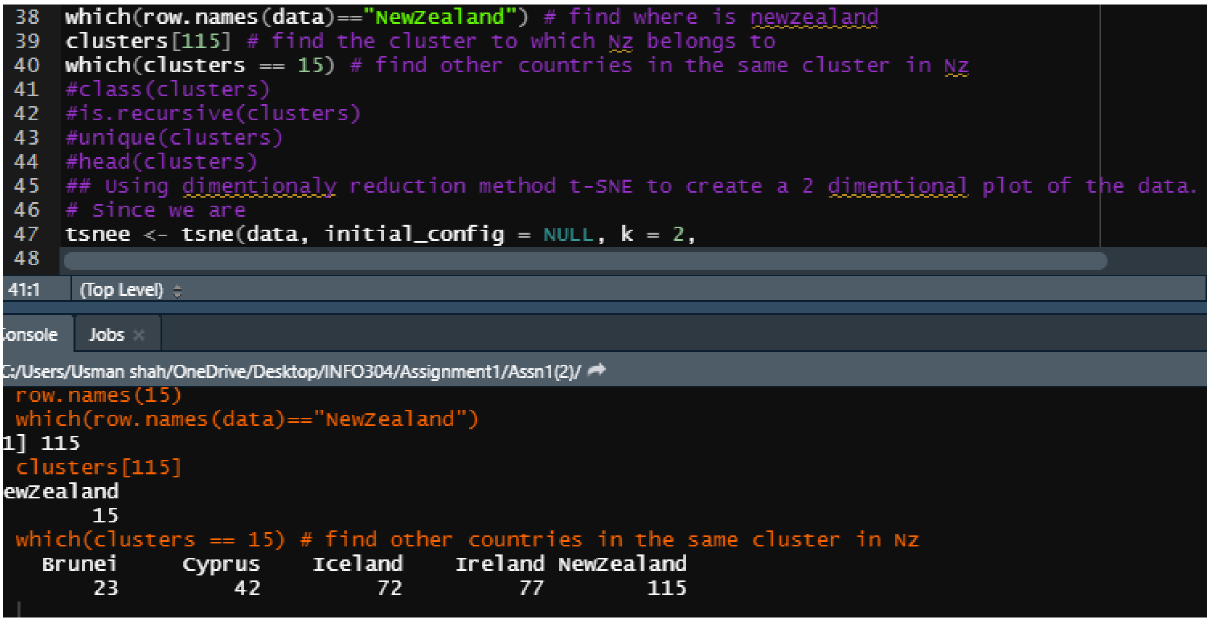


Figure 16 Dendrogram with average

Name: Usman Shah
Student ID: 1006293



```
which(row.names(data)=="NewZealand") # find where is newzealand
clusters[115] # find the cluster to which Nz belongs to
which(clusters == 15) # find other countries in the same cluster in Nz
#class(clusters)
#is.recursive(clusters)
#unique(clusters)
#head(clusters)
## Using dimentionaly reduction method t-SNE to create a 2 dimentional plot of the data.
# Since we are
tsnee <- tsne(data, initial_config = NULL, k = 2,
```

41:1
(Top Level)
onsole
Jobs
:/Users/Usman shah/OneDrive/Desktop/INFO304/Assignment1/Assn1(2)/

```
 row.names(15)
 which(row.names(data)=="NewZealand")
1] 115
 clusters[115]
ewZealand
        15
 which(clusters == 15) # find other countries in the same cluster in Nz
    Brunei      Cyprus     Iceland     Ireland NewZealand
        23          42          72          77        115
```

Figure 17 R code to Find which countries are in the same cluster with NewZealand

4. Figure 18 below is the 2-dimensional plot of the country data created by dimensionality reduction method t-SNE.
Figure 19 below is the plot zoomed in on New Zealand. As we can see that the five countries which are similar to New Zealand are all lying close to it in the plot like Australia and Spain. The reason for that is that they will have properties or explanatory variables which are the same are very close to the values of New Zealand.

Name: Usman Shah
Student ID: 1006293

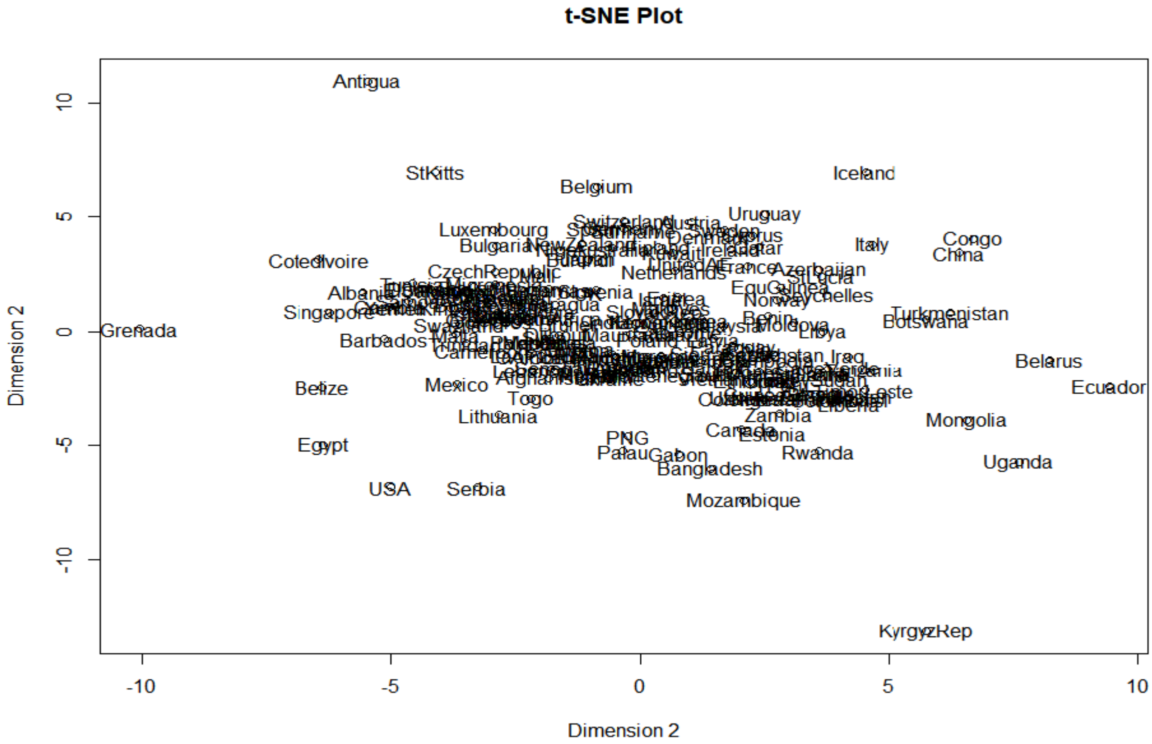


Figure 18 Plot with t-SNE

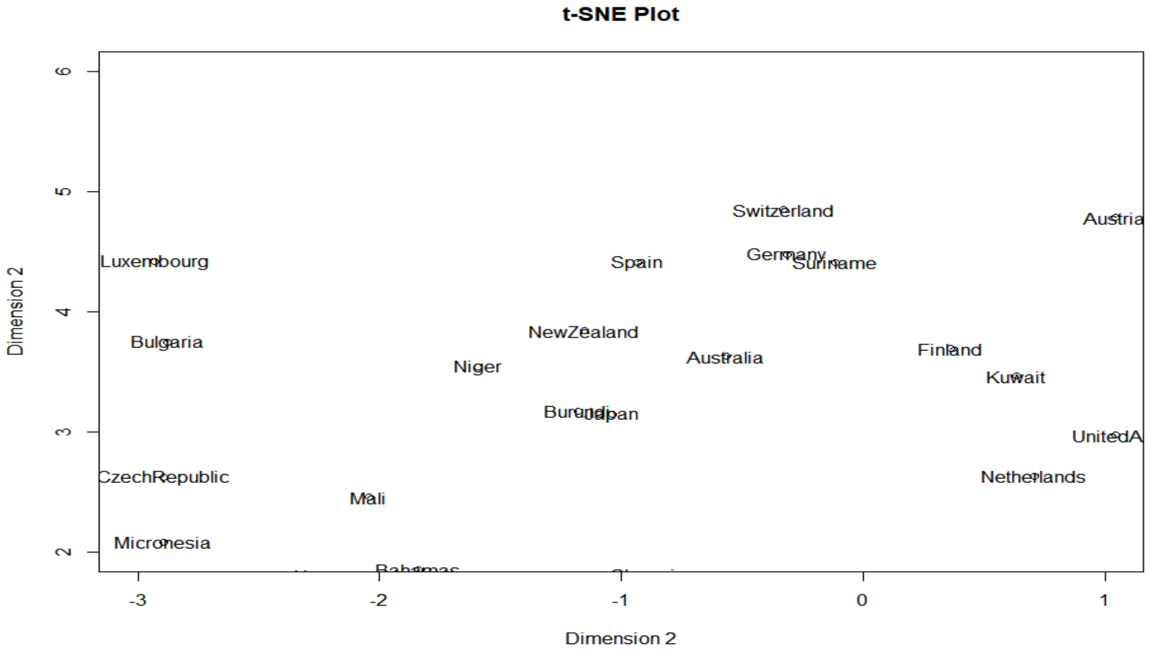


Figure 19 zoom plot for Nz