Homework 15

The USArrests data set available in R was used to perform k-means clustering. The data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

Given the output we had to answer the following questions:

1. What two variables were used in the clustering?
2. Comment on the ability of Urbancat to classify observations defined by the k-means clustering algorithm.

Question 1

As we can see from this snippet of code:

```
USArrests$Urbancat <- as.factor(ifelse(USArrests$UrbanPop > 66,1,0))
```

The Urbancat variable holds the classification of each state. If the selected state is over 66 (unknown units) Urban cat holds a 1.

```
Arrests_z <- data.frame(scale(USArrests[,1:2]))
nc <- NbClust(Arrests_z, min.nc=2, max.nc=15, method="kmeans")
```

In the above snippet we see that the Murder and Assault values of each state are stored as a dataframe into the Arrests_z variable. Then the clustering is done using the **Assault** and **Murder** values in each state.

Question 2

We basically wanted to see if having a greater population (more than the average) contributed to a higher murder or assault rate (per 100k). I don't think Urbancat does a good job in creating clusters and this can be seen on the table result.

Population	< 66 (0)	> 66 (1)
Murder	9	13
Assault	17	11

As we can see neither 0 or 1 values for Urbancat has a really high murder or assault value, which would be needed for a cluster to form. Therefore the Murder and Assault values don't really seem to depend on the population as much as other variables.