**Assignment 4 (5% CA):**

**Due at 11:59pm through Moodle on the 2nd of December**

**Provide all the code and results from R and report the answers in a document along with your interpretation.**

**K-means Clustering**

**Rock:**

Load in the ‘rock’ dataset which contains information about rock samples from a petroleum reservoir

1. Explore the data, describing the variables and making suitable plots.

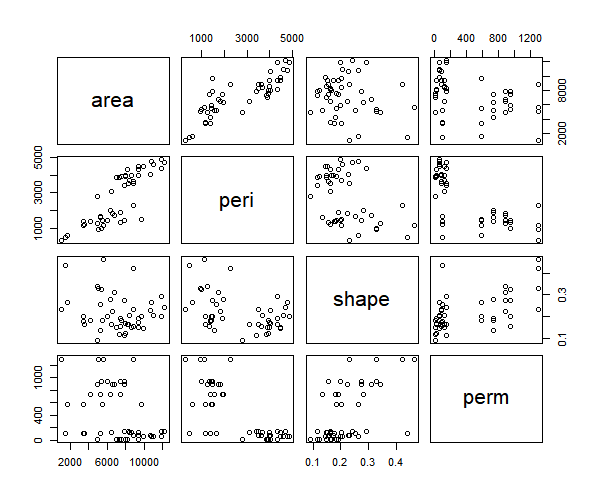
*Measurements on 48 rock samples from a petroleum reservoir.*

*A data frame with 48 rows and 4 numeric columns.*

*[,1] area area of pores space, in pixels out of 256 by 256*

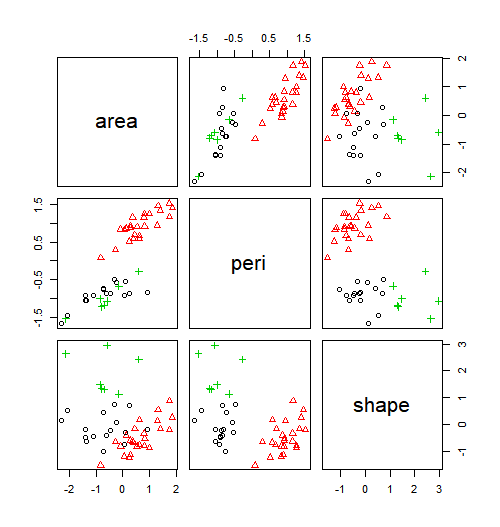
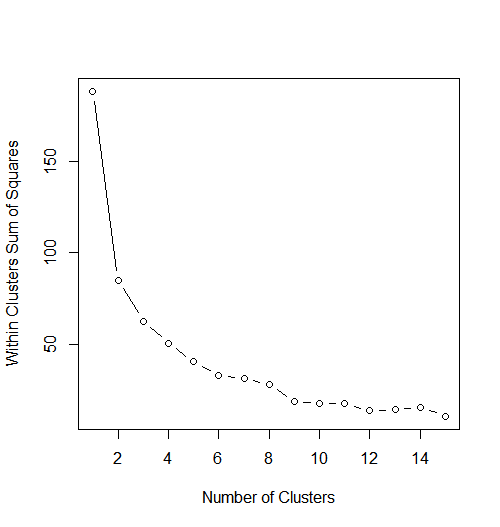
*[,2] peri perimeter in pixels*

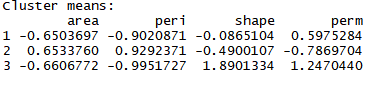
*[,3] shape perimeter/sqrt(area)*

*[,4] perm permeability in milli-Darcies*

1. Investigate if clustering would be suitable for this dataset, and if so how many clusters do you think is appropriate based on data?

*From examining the data it is made clear that a lot of the data on many plots is clustered with a rough estimate of around 3 clear clusters together which would suggest that clustering would be suitable for the dataset.*

1. Carry out a k-means cluster analysis on the data. Justify your choice of k and detail how this choice was made. Produce a suitable graph to show



1. Interpret the output from the cluster analysis.

*Using three clusters:*

*Green and black cluster tend to be correlated more closely than the red cluster.*

*Cluster 1:*

* *Area : Fairly below average area of pores space*
* *Perimeter: Really below average perimeter in pixels*
* *Shape: Really Below average perimeter/sqrt(area)*
* *Permeability: Fairly above average Permeability*

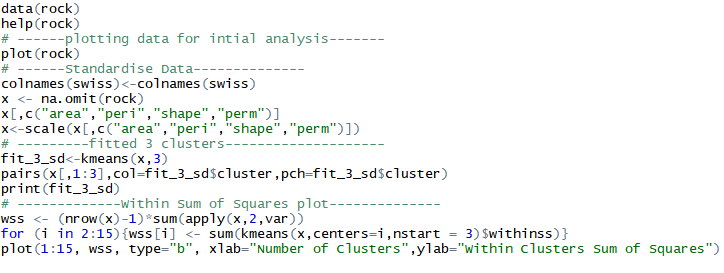
*Cluster 2:*

* *Area : Above average area of pores space*
* *Perimeter: Really above average perimeter in pixels*
* *Shape: Fairly Below average perimeter/sqrt(area)*
* *Permeability: Really Below average Permeability*

*Cluster 3:*

* *Area : Fairly below average area of pores space*
* *Perimeter: Really below average perimeter in pixels*
* *Shape: Extremely above average perimeter/sqrt(area)*
* *Permeability: Extremely above average Permeability*

*Elbow Curve agrees with our original estimate that 3 seems to be the correct number of clusters therefore k =3.*

*Code: Included R file in Zip with Word Document*