**Business Problem:**

Analyze the information given in the following ‘Insurance Policy dataset’ to create clusters of persons falling in the same type

**Data:**

Data in the form of numeric data. It contains the following column data.

→ Premium Paid – Amount paid by the person

→ Age – Indicates the age of the person

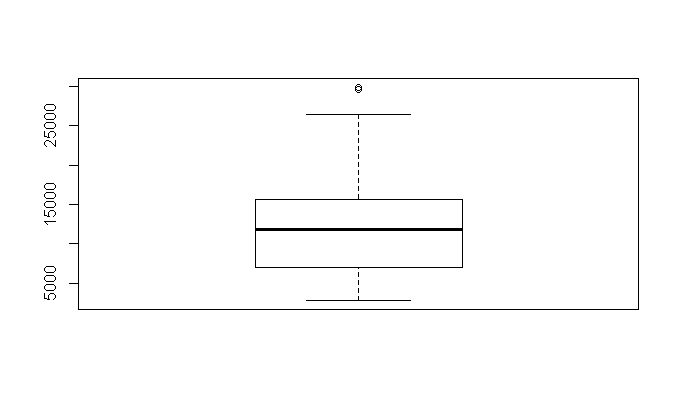
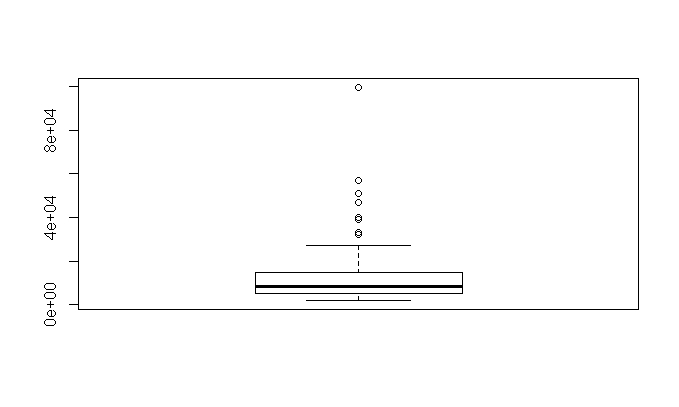
→ Days to renew – Days remaining to renew the policy

→ Claims made – Indicates the claims already made by the person

→ Income – Net income of the person

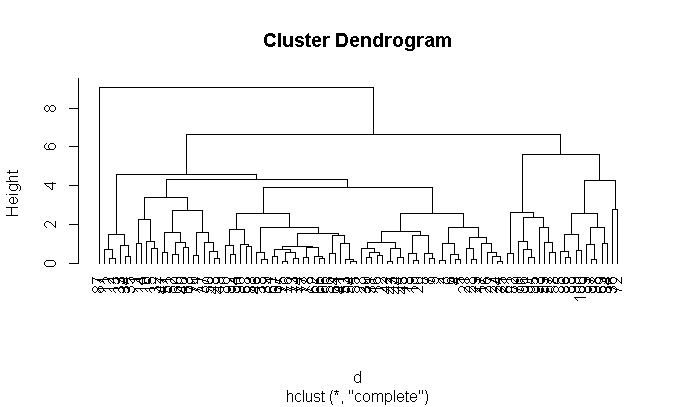
**Pre-processing Data:**

All the features are not in same scale. So first of all convert them all into single scale. There will outlier in the data. Deleted the 87th observation. Please look the box plot.



**Building the Model:**

Build the model using Euclidean distance and complete linkage functions. Please find the Dendrogram.

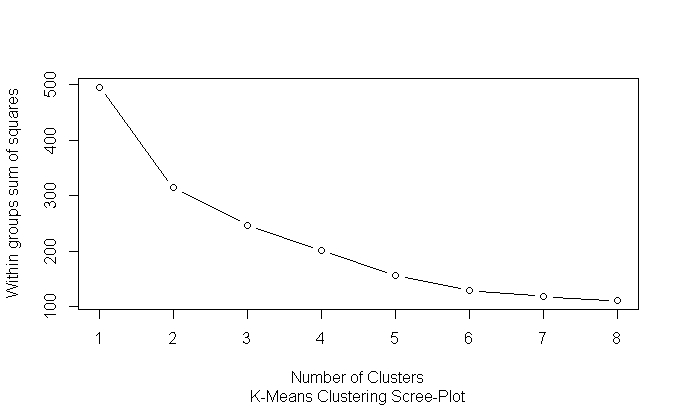


**K-Selection:**

k <- kselection(mydata[,-1], parallel = TRUE, k\_threshold = 0.9, max\_centers=5)

It is **giving k value as 10.**

**Elbow Curve:**



Elbow Curve value is subjective. As per **my analysis 2** will be good value for the K and subjective.

**KMeans:**

Calculate the tot.withinss and betweenss. For a good model tot.withinss should be more and betweenss should be less. If there should be less variation in the difference then we can treat it as final K.

Trails :

km <- kmeans(normalized\_data,2)

str(km)

###$ tot.withinss: num 307

###$ betweenss : num 183

###$ size : int [1:2] 75 25

km <- kmeans(normalized\_data,3)

str(km)

###$ withinss : num [1:3] 98.3 94.5 55.5

###$ tot.withinss: num 248

###$ betweenss : num 247

As per standards tot.withinss should be more and betweenss should be less. when we go to cluster 3 there will be almost equality. **So 2 will be good value for k**.

**KMeans Center and Animation diagram:**

km <- kmeans.ani(normalized\_data, 2)

km$centers

##Premiums.Paid Age Days.to.Renew Claims.made Income

##45 0.9644689 0.6604451 0.3354210 0.7881081 0.8962270

##16 -0.6008167 -0.4114248 -0.2089508 -0.4909526 -0.5583054

km <- kmeans.ani(normalized\_data, 3)

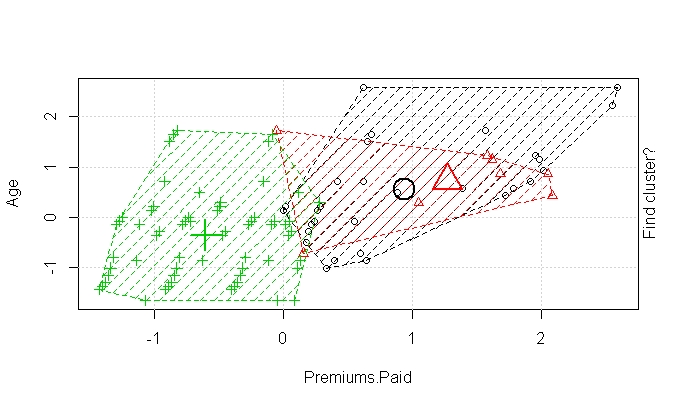
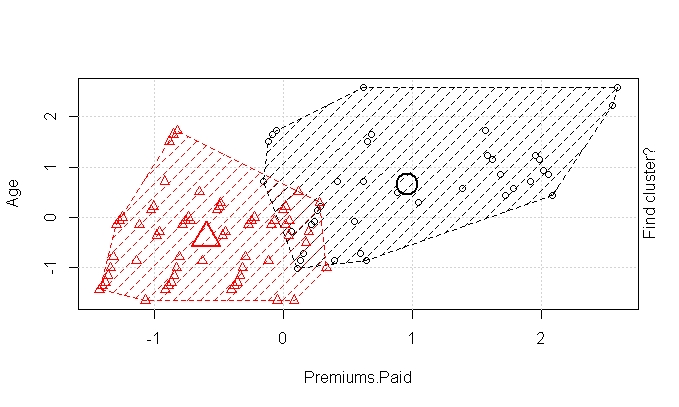
km$centers

## Premiums.Paid Age Days.to.Renew Claims.made Income

##98 0.9396305 0.5547923 -0.4291240 0.09470076 1.0112466

##36 1.2717162 0.7300369 1.6791853 2.76108668 1.0110976

##49 -0.6035970 -0.3536979 -0.0159498 -0.40056477 -0.6034667



By looking both diagrams and centre values, there will be no over lapping for cluster-2 compare than cluster-3.

**Hence K = 2 is good value.**