Cluster Analysis

Dean D'souza

2016-07-16

## Solutions:

* **Using the mtcars data set**

We setup the data as follows:

data(mtcars)  
head(mtcars)

## mpg cyl disp hp drat wt qsec vs am gear carb  
## Mazda RX4 21.0 6 160 110 3.90 2.620 16.46 0 1 4 4  
## Mazda RX4 Wag 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4  
## Datsun 710 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1  
## Hornet 4 Drive 21.4 6 258 110 3.08 3.215 19.44 1 0 3 1  
## Hornet Sportabout 18.7 8 360 175 3.15 3.440 17.02 0 0 3 2  
## Valiant 18.1 6 225 105 2.76 3.460 20.22 1 0 3 1

dat<-mtcars  
remove(mtcars)  
set.seed(596979)

* **Create a kmeans object from the first, second, and third columns**

We create the kmeans object after subsetting as follows:

cars123<-mtcars[,1:3]  
cars\_km1<-kmeans(cars123, centers = 3)

* **What is the size of each cluster?**

We can obtain the size of the three clusters as follows:

cars\_km1$size

## [1] 14 9 9

* **What are the centers of each cluster?**

We can obtain the centers of the clusters as follows:

cars\_km1$centers

## mpg cyl disp  
## 1 15.10000 8.000000 353.10000  
## 2 20.60000 5.555556 174.52222  
## 3 27.34444 4.000000 96.55556

* **What is the average disp, wt, and qsec of each cluster?**

We obtain the details as follows:

cars\_478\_avg<-aggregate(mtcars, by=list(cars\_km1$cluster), FUN = mean)  
cars\_478\_avg[,c(4,7,8)]

## disp wt qsec  
## 1 353.10000 3.999214 16.77214  
## 2 174.52222 3.128889 18.74889  
## 3 96.55556 2.089222 18.62333

* **Describe each cluster in English.**

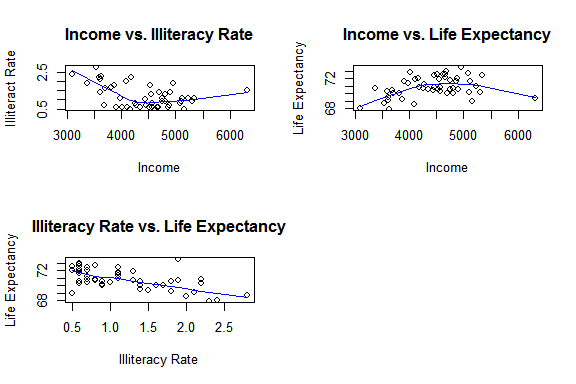
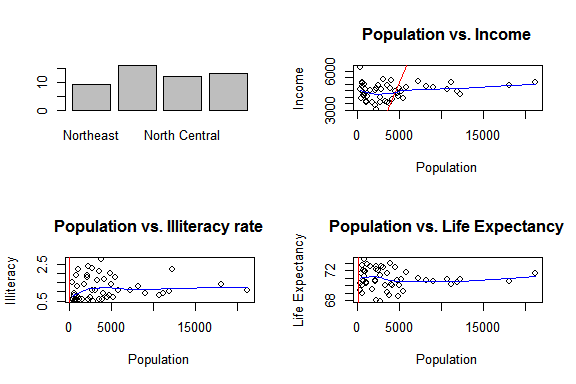
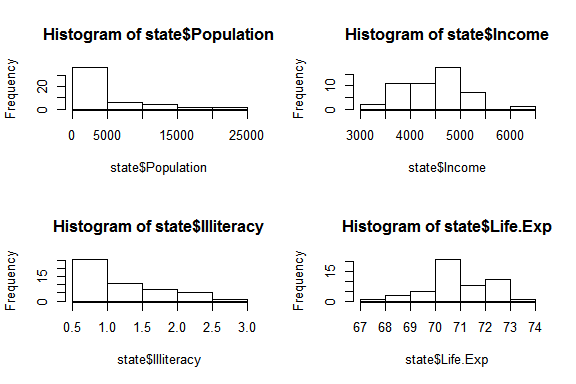
In the exercise above we have generated 3 clusters of sizes 14, 9 and 9 observations respectively. These clusters are centered around observations of seemingly increasing miles per gallon and decreasing displacement.

* **Find a data set with at least 4 columns of numeric data and a categorical column**

For the purpose of this assignment I chose to combine the first four columns of State.x77 and the State.region data into a single data frame as follows:

state<-na.omit(data.frame(state.x77[,1:4],state.region))

* **Run several scatter plots of the data**
* Please note in the following plots, the red line represents the regression line while the blue line represents the lowess line.



* **Create a kmeans object from the numeric data, you can pick K to be whatever you want**

We create the Kmeans object as follows:

state1234<-state[,1:4]  
state\_km1<-kmeans(state1234,centers = 3)

* **Determine the size of each cluster**

The size of the clusters are as follows:

state\_km1$size

## [1] 27 6 17

* **Determine the centers of each cluster**

The centers of the clusters are as follows:

state\_km1$centers

## Population Income Illiteracy Life.Exp  
## 1 1476.963 4389.741 1.074074 71.13074  
## 2 14217.167 4720.333 1.233333 70.75833  
## 3 5125.882 4408.529 1.300000 70.52059

* **Compare the clusters to the categorical data column as we did with the iris$Species column**

We compare the clusters to the categorical data as follows:

state\_play<-state  
state\_play$cluster\_id<-state\_km1$cluster  
head(state\_play)

## Population Income Illiteracy Life.Exp state.region cluster\_id  
## Alabama 3615 3624 2.1 69.05 South 3  
## Alaska 365 6315 1.5 69.31 West 1  
## Arizona 2212 4530 1.8 70.55 West 1  
## Arkansas 2110 3378 1.9 70.66 South 1  
## California 21198 5114 1.1 71.71 West 2  
## Colorado 2541 4884 0.7 72.06 West 1

table(state\_play$state.region,state\_play$cluster\_id)

##   
## 1 2 3  
## Northeast 5 2 2  
## South 6 1 9  
## North Central 5 2 5  
## West 11 1 1

* **For your chosen data set,**
* **Describe what each row of data represents**

Each row of the data set contains observations about the various states in the U.S.A. which includes average life expectancy, population, average income, percent illiteracy and region.

* **Describe each of your columns used - give a one sentence description of the column**

We now continue to describe the data with it's variables:

|  |  |  |
| --- | --- | --- |
|  | Variable | Description |
| 1 | Population | population estimate as of July 1, 1975 |
| 2 | Income | per capita income (1974) |
| 3 | Illiteracy | illiteracy (1970, percent of population) |
| 4 | Life Exp. | life expectancy in years (1969-71) |
| 5 | state.region | factor giving the region (Northeast, South, North Central, West) that each state belongs to. |

* **If you know it, describe how the data was generated**
* The State.x77 data set and state.region vector were combined together for the purpose of this assignment. These are provided with base R package of datasets. As all observations are arranged according to alphabetical order of states we only needed to combine the two normally.
* **For the clusters**
* **Describe the size and means of clusters**
* The clusters generated from the data consists of one relatively large cluster of size 27 observations, a medium cluster of 17 observations and a small cluster of 16 observations. We can also see from the table that a number of observations in the first cluster come from the Western states and a few from the other regions. Most of the southern states fall under cluster 3 whilw north central sates seem to favour clusters 1 and 3. Northeastern states seem to favour cluster 1.  
  Overall we can see that for the most part the clusters are formed over population densities and hence we name them in the following question appropriately.
* **Give a one- or two-word description to each cluster - in other words, give each cluster a label or name**
* Cluster 1 <- Low Population  
  Cluster 2 <- High Population  
  Cluster 3 <- Medium Population