Ass3- 368036jk-371474ts

368036jk and 371474ts

4/8/2017

Q1

Q1A: get the centroid for each cluster

```
set.seed(1)
x = cbind(c(1, 1, 0, 5, 6, 4), c(4, 3, 4, 1, 2, 0))

labels = sample(2, nrow(x), replace=T)

centroid1 = c(mean(x[labels==1, 1]), mean(x[labels==1, 2]))
centroid2 = c(mean(x[labels==2, 1]), mean(x[labels==2, 2]))
```

Q1B: Assign for each observation to a cluster closest to the observation

```
euclid = function(a, b) {
  return(sqrt((a[1] - b[1])^2 + (a[2]-b[2])^2))
}
assign_labels = function(x, centroid1, centroid2) {
  labels = rep(NA, nrow(x))
  for (i in 1:nrow(x)) {
    if (euclid(x[i,], centroid1) < euclid(x[i,], centroid2)) {
      labels[i] = 1
    } else {
      labels[i] = 2
    }
  }
  return(labels)
}
labels = assign_labels(x, centroid1, centroid2)
```

Q1C: repeat the steps until cluster stops changing

```
last_labels = rep(-1, 6)
while (!all(last_labels == labels)) {
  last_labels = labels
  centroid1 = c(mean(x[labels==1, 1]), mean(x[labels==1, 2]))
  centroid2 = c(mean(x[labels==2, 1]), mean(x[labels==2, 2]))
  print(centroid1)
  print(centroid2)
  labels = assign_labels(x, centroid1, centroid2)
}
```

```
## [1] 0.6666667 3.6666667
## [1] 5 1
```

Q2

perform a hierachrical clustering on the same date (x):

```
dist_matrix <- dist(x, method = "euclidean")
print(dist_matrix)
```

```
##          1        2        3        4        5
## 2 1.000000
## 3 1.000000 1.414214
## 4 5.000000 4.472136 5.830952
## 5 5.385165 5.099020 6.324555 1.414214
## 6 5.000000 4.242641 5.656854 1.414214 2.828427
```

```
dendogram <- hclust(dist_matrix)

groups <- cutree(dendogram, k=5)

dist <- dist(groups, method = "euclidean")
print(dist)
```

```
##   1 2 3 4 5
## 2 0
## 3 1 1
## 4 2 2 1
## 5 3 3 2 1
## 6 4 4 3 2 1
```

Q3

First we log into Twitter by using the authentication

```
require("twitteR")
```

```
## Loading required package: twitteR
```

```
consumer_key <-"2gE9wSUFIse7OdHeFQnBhaVE3"
consumer_secret <- "KP8BqE7aVqDcDo1A4eJyR9iFGPSQD7efrzWA7n2hDyB3Bp4w8y"
access_token <-"500773091-44pCm84DjQr05kYisvm4FS5Mrw3vl3RSx6MOHJT2"
access_secret<-"UTfRMU30evSfRdzJ5drshOgvx4o1n0XqaXIYgA2sQv00J"
setup_twitter_oauth(consumer_key,consumer_secret,access_token,access_secret)
```

```
## [1] "Using direct authentication"
```

Hereafter we collect the Tweets related to Deep Learning.

```
AItweets <- searchTwitter('#DeepLearning OR #machinelearning OR #ai',lang="en",geocode = "51.92,4.48,10
```

These tweets are stripped of retweets then and put into a dataframe

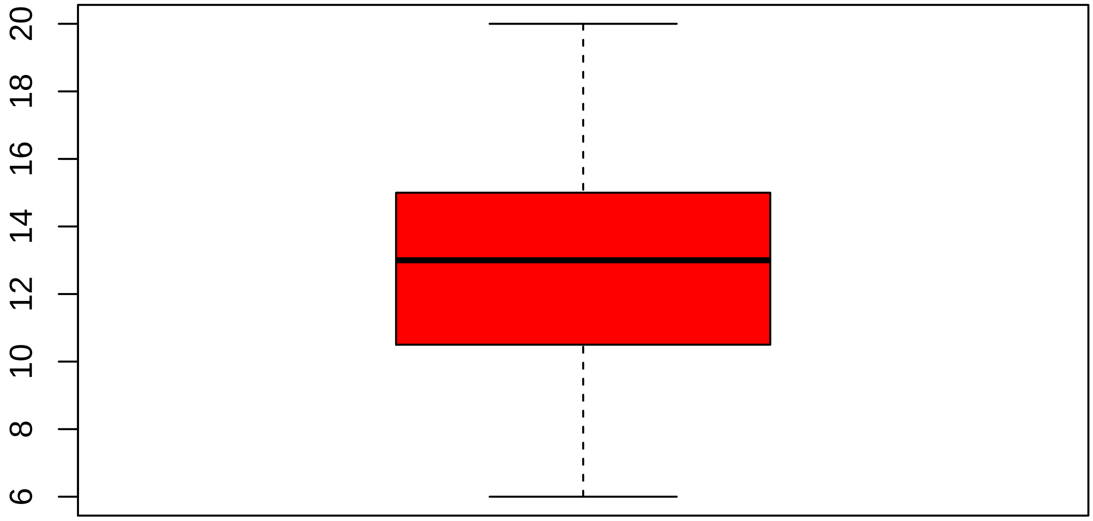
```
AItweets <- strip_retweets(AItweets)
AI <- twListToDF(AItweets)
```

Now the data is prepared for the graphical representation of the word count in each tweet in order to analyse the word count.

```
require("stringr")
```

```
## Loading required package: stringr
```

```
word_count <- str_count(AI$text,"\\S+")
boxplot(word_count,col = "red")
```



Now we explore some users who tweeted about AI, deep learning and machine learning

```
tweetsUser <- getUser("Fueladdicts")
tweetsUser2 <- getUser("AndreSpeek")
tweetsUser3 <- getUser("bigdataned")
```

First the users are analyzed

```
str(tweetsUser)
```

```
## Reference class 'user' [package "twitteR"] with 18 fields
##  $ description      : chr "EnergyPlatform Fueladdicts a https://t.co/Igc1anIq93 product https://t.co/
##  $ statusesCount    : num 5575
##  $ followersCount   : num 7639
##  $ favoritesCount   : num 5043
##  $ friendsCount     : num 277
##  $ url              : chr "https://t.co/ABkkOCGj5h"
##  $ name             : chr "Fueladdicts ( WWEP )"
##  $ created          : POSIXct[1:1], format: "2014-06-28 09:50:49"
##  $ protected        : logi FALSE
##  $ verified         : logi FALSE
##  $ screenName       : chr "Fueladdicts"
##  $ location         : chr "Rotterdam, Niederlande"
##  $ lang             : chr "en"
##  $ id               : chr "2592853261"
##  $ lastStatus       :Reference class 'status' [package "twitteR"] with 17 fields
##   ..$ text         : chr "Hot off the press! The Smartcity Rotterdam Daily is out!... https://t.co/o
##   ..$ favorited    : logi FALSE
##   ..$ favoriteCount: num 0
##   ..$ replyToSN    : chr(0)
##   ..$ created      : POSIXct[1:1], format: "2017-04-10 09:16:56"
##   ..$ truncated    : logi FALSE
##   ..$ replyToSID   : chr(0)
##   ..$ id           : chr "851363358158860288"
##   ..$ replyToUID   : chr(0)
##   ..$ statusSource : chr "<a href=\"http://www.facebook.com/twitter\" rel=\"nofollow\">Facebook</a>"
##   ..$ screenName   : chr "Unknown"
##   ..$ retweetCount : num 0
##   ..$ isRetweet    : logi FALSE
##   ..$ retweeted    : logi FALSE
##   ..$ longitude    : chr(0)
##   ..$ latitude     : chr(0)
##   ..$ urls         :'data.frame':    1 obs. of  5 variables:
##   .. ..$ url         : chr "https://t.co/o7v7H7PBBU"
##   .. ..$ expanded_url: chr "http://fb.me/wm2SSLrL"
##   .. ..$ display_url : chr "fb.me/wm2SSLrL"
##   .. ..$ start_index : num 60
##   .. ..$ stop_index  : num 83
##   ..and 53 methods, of which 39 are  possibly relevant:
##   ..  getCreated, getFavoriteCount, getFavorited, getId, getIsRetweet,
##   ..  getLatitude, getLongitude, getReplyToSID, getReplyToSN,
##   ..  getReplyToUID, getRetweetCount, getRetweeted, getRetweeters,
##   ..  getRetweets, getScreenName, getStatusSource, getText, getTruncated,
##   ..  getUrls, initialize, setCreated, setFavoriteCount, setFavorited,
##   ..  setId, setIsRetweet, setLatitude, setLongitude, setReplyToSID,
##   ..  setReplyToSN, setReplyToUID, setRetweetCount, setRetweeted,
##   ..  setScreenName, setStatusSource, setText, setTruncated, setUrls,
##   ..  toDataFrame, toDataFrame#twitterObj
##  $ listedCount      : num 151
##  $ followRequestSent: logi FALSE
##  $ profileImageUrl  : chr "http://pbs.twimg.com/profile_images/826366212544278532/bQed4Trq_normal.jpg
##  and 59 methods, of which 45 are  possibly relevant:
##    getCreated, getDescription, getFavorites, getFavoritesCount,
##    getFavouritesCount, getFollowRequestSent, getFollowerIDs, getFollowers,
##    getFollowersCount, getFriendIDs, getFriends, getFriendsCount, getId,
##    getLang, getLastStatus, getListedCount, getLocation, getName,
##    getProfileImageUrl, getProtected, getScreenName, getStatusesCount,
##    getUrl, getVerified, initialize, setCreated, setDescription,
##    setFavoritesCount, setFollowRequestSent, setFollowersCount,
##    setFriendsCount, setId, setLang, setLastStatus, setListedCount,
##    setLocation, setName, setProfileImageUrl, setProtected, setScreenName,
##    setStatusesCount, setUrl, setVerified, toDataFrame,
##    toDataFrame#twitterObj
```

```
str(tweetsUser2)
```

```
## Reference class 'user' [package "twitteR"] with 18 fields
##  $ description      : chr "Living in the Netherlands \u221e Try to be creative in music, software de
##  $ statusesCount    : num 24454
##  $ followersCount   : num 1014
##  $ favoritesCount   : num 74
##  $ friendsCount     : num 792
##  $ url              : chr "http://t.co/JUrZe5p2U7"
##  $ name             : chr "Andre Speek"
##  $ created          : POSIXct[1:1], format: "2009-08-17 09:09:17"
##  $ protected        : logi FALSE
##  $ verified         : logi FALSE
##  $ screenName       : chr "AndreSpeek"
##  $ location         : chr "Capelle aan den IJssel"
##  $ lang             : chr "en"
##  $ id               : chr "66315955"
##  $ lastStatus       :Reference class 'status' [package "twitteR"] with 17 fields
##   ..$ text         : chr "Bring Predictability to Innovation - https://t.co/IIqr35UuUS #in"
##   ..$ favorited    : logi FALSE
##   ..$ favoriteCount: num 0
##   ..$ replyToSN    : chr(0)
##   ..$ created      : POSIXct[1:1], format: "2017-04-10 07:09:01"
##   ..$ truncated    : logi FALSE
##   ..$ replyToSID   : chr(0)
##   ..$ id           : chr "851331166116265984"
##   ..$ replyToUID   : chr(0)
##   ..$ statusSource : chr "<a href=\"http://klout.com\" rel=\"nofollow\">Post with Klout</a>"
##   ..$ screenName   : chr "Unknown"
##   ..$ retweetCount : num 0
##   ..$ isRetweet    : logi FALSE
##   ..$ retweeted    : logi FALSE
##   ..$ longitude    : chr(0)
##   ..$ latitude     : chr(0)
##   ..$ urls         :'data.frame':    1 obs. of  5 variables:
##   .. ..$ url         : chr "https://t.co/IIqr35UuUS"
##   .. ..$ expanded_url: chr "http://klou.tt/rgihzijyqhjn"
##   .. ..$ display_url : chr "klou.tt/rgihzijyqhjn"
##   .. ..$ start_index : num 37
##   .. ..$ stop_index  : num 60
##   ..and 53 methods, of which 39 are  possibly relevant:
##   ..  getCreated, getFavoriteCount, getFavorited, getId, getIsRetweet,
##   ..  getLatitude, getLongitude, getReplyToSID, getReplyToSN,
##   ..  getReplyToUID, getRetweetCount, getRetweeted, getRetweeters,
##   ..  getRetweets, getScreenName, getStatusSource, getText, getTruncated,
##   ..  getUrls, initialize, setCreated, setFavoriteCount, setFavorited,
##   ..  setId, setIsRetweet, setLatitude, setLongitude, setReplyToSID,
##   ..  setReplyToSN, setReplyToUID, setRetweetCount, setRetweeted,
##   ..  setScreenName, setStatusSource, setText, setTruncated, setUrls,
##   ..  toDataFrame, toDataFrame#twitterObj
##  $ listedCount      : num 89
##  $ followRequestSent: logi FALSE
##  $ profileImageUrl  : chr "http://pbs.twimg.com/profile_images/812107001228619776/pSA-XofG_normal.jp
##  and 59 methods, of which 45 are  possibly relevant:
##    getCreated, getDescription, getFavorites, getFavoritesCount,
##    getFavouritesCount, getFollowRequestSent, getFollowerIDs, getFollowers,
##    getFollowersCount, getFriendIDs, getFriends, getFriendsCount, getId,
##    getLang, getLastStatus, getListedCount, getLocation, getName,
##    getProfileImageUrl, getProtected, getScreenName, getStatusesCount,
##    getUrl, getVerified, initialize, setCreated, setDescription,
##    setFavoritesCount, setFollowRequestSent, setFollowersCount,
##    setFriendsCount, setId, setLang, setLastStatus, setListedCount,
##    setLocation, setName, setProfileImageUrl, setProtected, setScreenName,
##    setStatusesCount, setUrl, setVerified, toDataFrame,
##    toDataFrame#twitterObj
```

```r
str(tweetsUser3)
```

```
## Reference class 'user' [package "twitteR"] with 18 fields
##  $ description      : chr "Big Data | Machine Learning | Deep Learning | Blockchain & Artificial inte
##  $ statusesCount    : num 154
##  $ followersCount   : num 204
##  $ favoritesCount   : num 4
##  $ friendsCount     : num 176
##  $ url              : chr(0)
##  $ name             : chr "Big Data News"
##  $ created          : POSIXct[1:1], format: "2017-03-08 11:28:36"
##  $ protected        : logi FALSE
##  $ verified         : logi FALSE
##  $ screenName       : chr "bigdataned"
##  $ location         : chr "Rotterdam, Nederland"
##  $ lang             : chr "nl"
##  $ id               : chr "839437691678822404"
##  $ lastStatus       :Reference class 'status' [package "twitteR"] with 17 fields
##   ..$ text         : chr "Four Trends in Artificial Intelligence That Affect Enterprises - Hortonwork
##   ..$ favorited    : logi FALSE
##   ..$ favoriteCount: num 3
##   ..$ replyToSN    : chr(0)
##   ..$ created      : POSIXct[1:1], format: "2017-04-07 13:03:05"
##   ..$ truncated    : logi TRUE
##   ..$ replyToSID   : chr(0)
##   ..$ id           : chr "850333105416265728"
##   ..$ replyToUID   : chr(0)
##   ..$ statusSource : chr "<a href=\"http://bufferapp.com\" rel=\"nofollow\">Buffer</a>"
##   ..$ screenName   : chr "Unknown"
##   ..$ retweetCount : num 1
##   ..$ isRetweet    : logi FALSE
##   ..$ retweeted    : logi FALSE
##   ..$ longitude    : chr(0)
##   ..$ latitude     : chr(0)
##   ..$ urls         :'data.frame':    2 obs. of  5 variables:
##   .. ..$ url         : chr [1:2] "https://t.co/jFMYDiFwzU" "https://t.co/DzEG1F6RZg"
##   .. ..$ expanded_url: chr [1:2] "http://buff.ly/2oHBU1i" "https://twitter.com/i/web/status/850333105
##   .. ..$ display_url : chr [1:2] "buff.ly/2oHBU1i" "twitter.com/i/web/status/8\u2026"
##   .. ..$ start_index : num [1:2] 77 116
##   .. ..$ stop_index  : num [1:2] 100 139
##   ..and 53 methods, of which 39 are  possibly relevant:
##   ..  getCreated, getFavoriteCount, getFavorited, getId, getIsRetweet,
##   ..  getLatitude, getLongitude, getReplyToSID, getReplyToSN,
##   ..  getReplyToUID, getRetweetCount, getRetweeted, getRetweeters,
##   ..  getRetweets, getScreenName, getStatusSource, getText, getTruncated,
##   ..  getUrls, initialize, setCreated, setFavoriteCount, setFavorited,
##   ..  setId, setIsRetweet, setLatitude, setLongitude, setReplyToSID,
##   ..  setReplyToSN, setReplyToUID, setRetweetCount, setRetweeted,
##   ..  setScreenName, setStatusSource, setText, setTruncated, setUrls,
##   ..  toDataFrame, toDataFrame#twitterObj
##  $ listedCount      : num 4
##  $ followRequestSent: logi FALSE
##  $ profileImageUrl  : chr "http://pbs.twimg.com/profile_images/843852315299921920/vd13MqlU_normal.jpg
##  and 59 methods, of which 45 are  possibly relevant:
##    getCreated, getDescription, getFavorites, getFavoritesCount,
##    getFavouritesCount, getFollowRequestSent, getFollowerIDs, getFollowers,
##    getFollowersCount, getFriendIDs, getFriends, getFriendsCount, getId,
##    getLang, getLastStatus, getListedCount, getLocation, getName,
##    getProfileImageUrl, getProtected, getScreenName, getStatusesCount,
##    getUrl, getVerified, initialize, setCreated, setDescription,
##    setFavoritesCount, setFollowRequestSent, setFollowersCount,
##    setFriendsCount, setId, setLang, setLastStatus, setListedCount,
##    setLocation, setName, setProfileImageUrl, setProtected, setScreenName,
##    setStatusesCount, setUrl, setVerified, toDataFrame,
##    toDataFrame#twitterObj
```

```
tweetsUser$getDescription()
```

```
## [1] "EnergyPlatform Fueladdicts a https://t.co/Igc1anIq93 product https://t.co/o831h7UGwr https://t.
```

```
tweetsUser2$getDescription()
```

```
## [1] "Living in the Netherlands <U+221E> Try to be creative in music, software design, photography &
```

```
tweetsUser3$getDescription()
```

```
## [1] "Big Data | Machine Learning | Deep Learning | Blockchain & Artificial intelligence News..  Stay
```

```
tweetsUser$getFollowersCount()
```

```
## [1] 7639
```

```
tweetsUser2$getFollowersCount()
```

```
## [1] 1014
```

```
tweetsUser3$getFollowersCount()
```

```
## [1] 204
```

```
tweetsUser$getLastStatus()
```

```
## [1] "Unknown: Hot off the press! The Smartcity Rotterdam Daily is out!... https://t.co/o7v7H7PBBU"
```

```
tweetsUser2$getLastStatus()
```

```
## [1] "Unknown: Bring Predictability to Innovation - https://t.co/IIqr35UuUS #in"
```

```
tweetsUser3$getLastStatus()
```

```
## [1] "Unknown: Four Trends in Artificial Intelligence That Affect Enterprises - Hortonworks https://t
```

Now their timelines are analyzed

```
tweetsuserTimeline1 <- userTimeline("Fueladdicts")
tweetsuserTimeline2 <- userTimeline("AndreSpeek")
tweetsuserTimeline3 <- userTimeline("bigdataned")
# --- Analyze contents
str(tweetsuserTimeline1[[5]])
```

```
## Reference class 'status' [package "twitteR"] with 17 fields
##  $ text         : chr "The latest Fueladdicts World Wide Energy Platform ( WWEP )! https://t.co/LM2b
##  $ favorited    : logi FALSE
##  $ favoriteCount: num 0
##  $ replyToSN    : chr(0)
##  $ created      : POSIXct[1:1], format: "2017-04-10 07:33:34"
##  $ truncated    : logi FALSE
##  $ replyToSID   : chr(0)
##  $ id           : chr "851337342472384512"
##  $ replyToUID   : chr(0)
##  $ statusSource : chr "<a href=\"http://paper.li\" rel=\"nofollow\">Paper.li</a>"
##  $ screenName   : chr "Fueladdicts"
##  $ retweetCount : num 0
##  $ isRetweet    : logi FALSE
##  $ retweeted    : logi FALSE
##  $ longitude    : chr(0)
##  $ latitude     : chr(0)
##  $ urls         :'data.frame':   1 obs. of  5 variables:
##   ..$ url         : chr "https://t.co/LM2bdZ3JVK"
##   ..$ expanded_url: chr "http://paper.li/Fueladdicts/1468308853?edition_id=00435af0-1dc0-11e7-802f-0
##   ..$ display_url : chr "paper.li/Fueladdicts/14\u2026"
##   ..$ start_index : num 60
##   ..$ stop_index  : num 83
##  and 53 methods, of which 39 are  possibly relevant:
##    getCreated, getFavoriteCount, getFavorited, getId, getIsRetweet,
##    getLatitude, getLongitude, getReplyToSID, getReplyToSN, getReplyToUID,
##    getRetweetCount, getRetweeted, getRetweeters, getRetweets,
##    getScreenName, getStatusSource, getText, getTruncated, getUrls,
##    initialize, setCreated, setFavoriteCount, setFavorited, setId,
##    setIsRetweet, setLatitude, setLongitude, setReplyToSID, setReplyToSN,
##    setReplyToUID, setRetweetCount, setRetweeted, setScreenName,
##    setStatusSource, setText, setTruncated, setUrls, toDataFrame,
##    toDataFrame#twitterObj
```

```r
tweetsuserTimeline1[[5]]$getText()
```

```
## [1] "The latest Fueladdicts World Wide Energy Platform ( WWEP )! https://t.co/LM2bdZ3JVK Thanks to @
```

```r
str(tweetsuserTimeline2[[5]])
```

```
## Reference class 'status' [package "twitteR"] with 17 fields
##  $ text         : chr "Have a nice Sunday evening... ;-) https://t.co/mrIuE9g8bZ"
##  $ favorited    : logi FALSE
##  $ favoriteCount: num 0
##  $ replyToSN    : chr(0)
##  $ created      : POSIXct[1:1], format: "2017-04-09 18:19:13"
##  $ truncated    : logi FALSE
##  $ replyToSID   : chr(0)
##  $ id           : chr "851137437292167168"
##  $ replyToUID   : chr(0)
##  $ statusSource : chr "<a href=\"http://twitter.com\" rel=\"nofollow\">Twitter Web Client</a>"
##  $ screenName   : chr "AndreSpeek"
##  $ retweetCount : num 0
##  $ isRetweet    : logi FALSE
##  $ retweeted    : logi FALSE
##  $ longitude    : chr(0)
##  $ latitude     : chr(0)
##  $ urls         :'data.frame':   0 obs. of  4 variables:
##   ..$ url         : chr(0)
##   ..$ expanded_url: chr(0)
##   ..$ dispaly_url : chr(0)
##   ..$ indices     : num(0)
##  and 53 methods, of which 39 are  possibly relevant:
##    getCreated, getFavoriteCount, getFavorited, getId, getIsRetweet,
##    getLatitude, getLongitude, getReplyToSID, getReplyToSN, getReplyToUID,
##    getRetweetCount, getRetweeted, getRetweeters, getRetweets,
##    getScreenName, getStatusSource, getText, getTruncated, getUrls,
##    initialize, setCreated, setFavoriteCount, setFavorited, setId,
##    setIsRetweet, setLatitude, setLongitude, setReplyToSID, setReplyToSN,
##    setReplyToUID, setRetweetCount, setRetweeted, setScreenName,
##    setStatusSource, setText, setTruncated, setUrls, toDataFrame,
##    toDataFrame#twitterObj
```

```r
str(tweetsuserTimeline2[[5]]$getText())
```

```
##  chr "Have a nice Sunday evening... ;-) https://t.co/mrIuE9g8bZ"
```

```r
str(tweetsuserTimeline3[[5]])
```

```
## Reference class 'status' [package "twitteR"] with 17 fields
##  $ text         : chr "#Artificial Intelligence: Yes, machines are going to steal your jobs https://
##  $ favorited    : logi FALSE
##  $ favoriteCount: num 0
##  $ replyToSN    : chr(0)
##  $ created      : POSIXct[1:1], format: "2017-04-07 08:50:00"
##  $ truncated    : logi FALSE
##  $ replyToSID   : chr(0)
##  $ id           : chr "850269415291494400"
##  $ replyToUID   : chr(0)
##  $ statusSource : chr "<a href=\"http://bufferapp.com\" rel=\"nofollow\">Buffer</a>"
##  $ screenName   : chr "bigdataned"
##  $ retweetCount : num 0
##  $ isRetweet    : logi FALSE
##  $ retweeted    : logi FALSE
##  $ longitude    : chr(0)
##  $ latitude     : chr(0)
##  $ urls         :'data.frame':   1 obs. of  5 variables:
##   ..$ url         : chr "https://t.co/1wE1pmUk6P"
##   ..$ expanded_url: chr "http://buff.ly/2oHwK5C"
##   ..$ display_url : chr "buff.ly/2oHwK5C"
##   ..$ start_index : num 69
##   ..$ stop_index  : num 92
##  and 53 methods, of which 39 are  possibly relevant:
##    getCreated, getFavoriteCount, getFavorited, getId, getIsRetweet,
##    getLatitude, getLongitude, getReplyToSID, getReplyToSN, getReplyToUID,
##    getRetweetCount, getRetweeted, getRetweeters, getRetweets,
##    getScreenName, getStatusSource, getText, getTruncated, getUrls,
##    initialize, setCreated, setFavoriteCount, setFavorited, setId,
##    setIsRetweet, setLatitude, setLongitude, setReplyToSID, setReplyToSN,
##    setReplyToUID, setRetweetCount, setRetweeted, setScreenName,
##    setStatusSource, setText, setTruncated, setUrls, toDataFrame,
##    toDataFrame#twitterObj
```

```r
str(tweetsuserTimeline3[[5]]$getText())
```

```
##  chr "#Artificial Intelligence: Yes, machines are going to steal your jobs https://t.co/1wE1pmUk6P |
```

Finally these are transferred to the Dataframe

```r
tweetsuserTimeline1 <-twListToDF(tweetsuserTimeline1)
tweetsuserTimeline2 <-twListToDF(tweetsuserTimeline2)
tweetsuserTimeline3 <-twListToDF(tweetsuserTimeline3)
```

Q4

First, import the data.

```
data <- read.csv(file = "/Users/jckrick/Documents/BIM/Big_Data_and_Business_Analytics/Data/Assignment3/
```

Then the working directory is returned and set `getwd()` `setwd(...)`

Preprocessing

Now, we preprocess the data.First, we add the column names.

```
colnames(data) <- c("cultivar","Alcohol","Malic Acid","Ash","Alcalinity of ash","Magnesium","Total phen
```

Inspect the data

```
summary(data)
```

```
##     cultivar        Alcohol        Malic Acid        Ash       
##  Min.   :1.000   Min.   :11.03   Min.   :0.74   Min.   :1.360  
##  1st Qu.:1.000   1st Qu.:12.36   1st Qu.:1.60   1st Qu.:2.210  
##  Median :2.000   Median :13.05   Median :1.87   Median :2.360  
##  Mean   :1.944   Mean   :12.99   Mean   :2.34   Mean   :2.366  
##  3rd Qu.:3.000   3rd Qu.:13.67   3rd Qu.:3.10   3rd Qu.:2.560  
##  Max.   :3.000   Max.   :14.83   Max.   :5.80   Max.   :3.230  
##  Alcalinity of ash   Magnesium      Total phenols     Flavanoids   
##  Min.   :10.60     Min.   : 70.00   Min.   :0.980   Min.   :0.340  
##  1st Qu.:17.20     1st Qu.: 88.00   1st Qu.:1.740   1st Qu.:1.200  
##  Median :19.50     Median : 98.00   Median :2.350   Median :2.130  
##  Mean   :19.52     Mean   : 99.59   Mean   :2.292   Mean   :2.023  
##  3rd Qu.:21.50     3rd Qu.:107.00   3rd Qu.:2.800   3rd Qu.:2.860  
##  Max.   :30.00     Max.   :162.00   Max.   :3.880   Max.   :5.080  
##  Nonflavanoid phenols Proanthocyanins Color intensity       Hue       
##  Min.   :0.1300       Min.   :0.410   Min.   : 1.280   Min.   :0.480  
##  1st Qu.:0.2700       1st Qu.:1.250   1st Qu.: 3.210   1st Qu.:0.780  
##  Median :0.3400       Median :1.550   Median : 4.680   Median :0.960  
##  Mean   :0.3623       Mean   :1.587   Mean   : 5.055   Mean   :0.957  
##  3rd Qu.:0.4400       3rd Qu.:1.950   3rd Qu.: 6.200   3rd Qu.:1.120  
##  Max.   :0.6600       Max.   :3.580   Max.   :13.000   Max.   :1.710  
##  OD280/OD315 of diluted wines    Proline      
##  Min.   :1.270                Min.   : 278.0  
##  1st Qu.:1.930                1st Qu.: 500.0  
##  Median :2.780                Median : 672.0  
##  Mean   :2.604                Mean   : 745.1  
##  3rd Qu.:3.170                3rd Qu.: 985.0  
##  Max.   :4.000                Max.   :1680.0
```

```
str(data)
```

```
## 'data.frame':    177 obs. of  14 variables:
##  $ cultivar                    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Alcohol                     : num  13.2 13.2 14.4 13.2 14.2 ...
##  $ Malic Acid                  : num  1.78 2.36 1.95 2.59 1.76 1.87 2.15 1.64 1.35 2.16 ...
##  $ Ash                         : num  2.14 2.67 2.5 2.87 2.45 2.45 2.61 2.17 2.27 2.3 ...
##  $ Alcalinity of ash           : num  11.2 18.6 16.8 21 15.2 14.6 17.6 14 16 18 ...
##  $ Magnesium                   : int  100 101 113 118 112 96 121 97 98 105 ...
##  $ Total phenols               : num  2.65 2.8 3.85 2.8 3.27 2.5 2.6 2.8 2.98 2.95 ...
##  $ Flavanoids                  : num  2.76 3.24 3.49 2.69 3.39 2.52 2.51 2.98 3.15 3.32 ...
##  $ Nonflavanoid phenols        : num  0.26 0.3 0.24 0.39 0.34 0.3 0.31 0.29 0.22 0.22 ...
##  $ Proanthocyanins             : num  1.28 2.81 2.18 1.82 1.97 1.98 1.25 1.98 1.85 2.38 ...
##  $ Color intensity             : num  4.38 5.68 7.8 4.32 6.75 5.25 5.05 5.2 7.22 5.75 ...
##  $ Hue                         : num  1.05 1.03 0.86 1.04 1.05 1.02 1.06 1.08 1.01 1.25 ...
##  $ OD280/OD315 of diluted wines: num  3.4 3.17 3.45 2.93 2.85 3.58 3.58 2.85 3.55 3.17 ...
##  $ Proline                     : int  1050 1185 1480 735 1450 1290 1295 1045 1045 1510 ...
```

```
colSums(is.na(data))
```

```
##                     cultivar                      Alcohol 
##                            0                            0 
##                   Malic Acid                          Ash 
##                            0                            0 
##            Alcalinity of ash                    Magnesium 
##                            0                            0 
##                Total phenols                   Flavanoids 
##                            0                            0 
##         Nonflavanoid phenols              Proanthocyanins 
##                            0                            0 
##              Color intensity                          Hue 
##                            0                            0 
## OD280/OD315 of diluted wines                      Proline 
##                            0                            0
```

There is no missing data and the data seems to be intact.

In order to produce good results we need to scale the data.

```
data <- scale(data[2:13])
```

Cluster Analysis

Start with 5 clusters

```
require("rpart")
```

```
## Loading required package: rpart
```

```
require("cluster")
```

```
## Loading required package: cluster
```

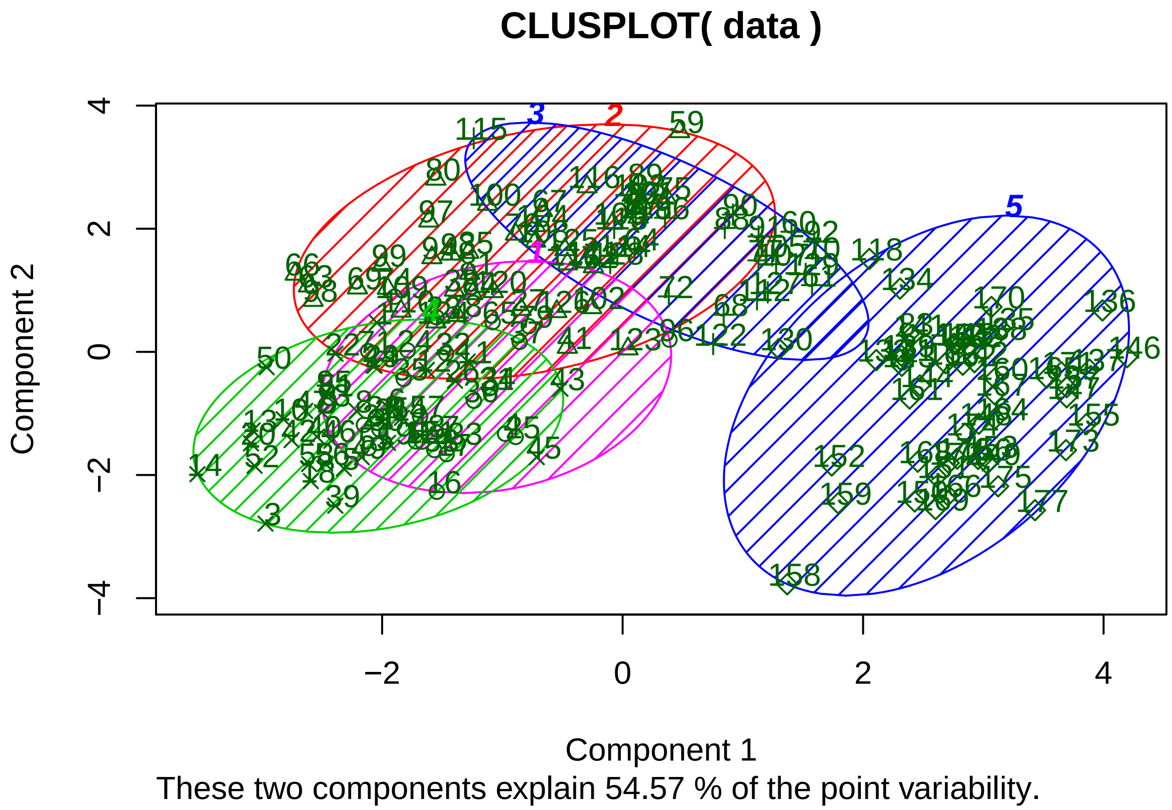
```
require("rpart.plot")
```

```
## Loading required package: rpart.plot
```

```
require("ape")
```

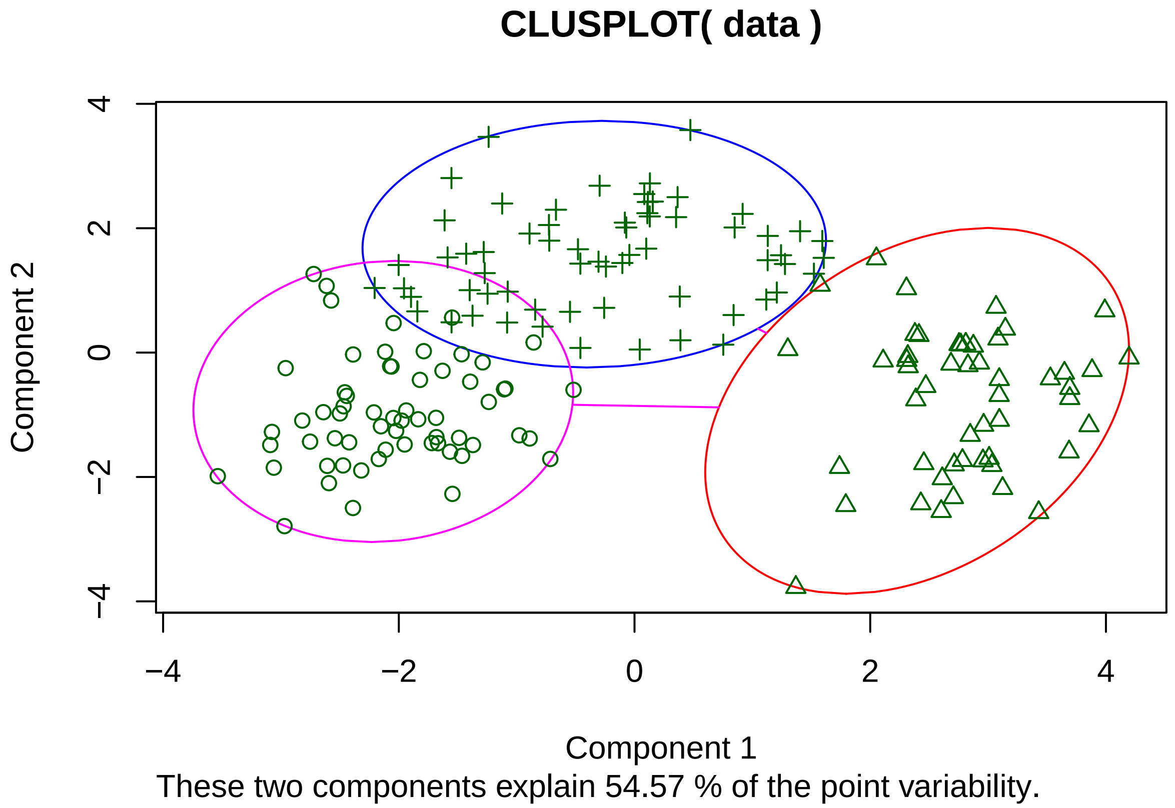
```
## Loading required package: ape
```

```
cluster_model <- kmeans(data,5)
clusplot(data, cluster_model$cluster, color = TRUE,shade = TRUE,labels = 2,lines=0)
```



We see that there is quite some overlap and it hints to 3 or 2 means clustering.

```
cluster_model <- kmeans(data,3)
clusplot(data, cluster_model$cluster, color = TRUE)
```



This time, it makes sense and there is no substantial overlap. This means that there are three clusters in this dataset.

We then add the clusters value to the original dataset and view the new DataFrame.

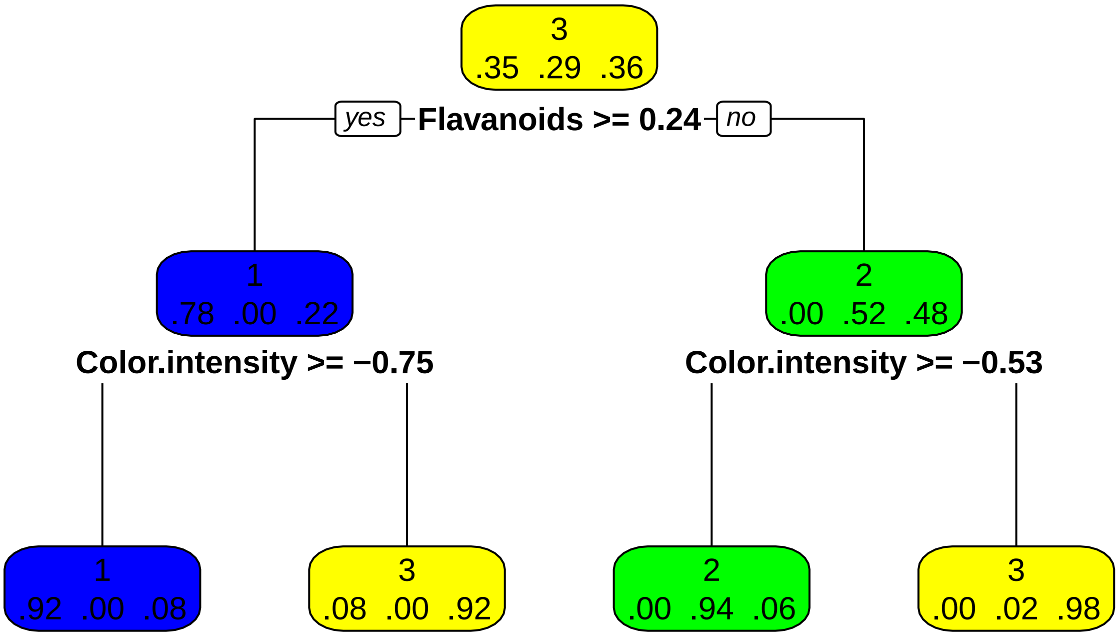
```
data <- data.frame(data, cluster=as.factor(cluster_model$cluster))
View(data)
```

DecisionTree

To find out what's behind each cluster one must employ a decision tree

```
tree<-rpart(cluster ~ Alcohol + Malic.Acid + Ash + Alcalinity.of.ash + Magnesium + Total.phenols + Flava

rpart.plot(tree,box.col=c("blue","green","yellow","grey")[tree$frame$yval],extra=4)
```



Here, we can see that the wine was seperated into three categories according to its Flavanoids and its Color intensity. In Category one there are only wines which have high Flavanoids and have an intense color. In Category two there are wines that have low flavanoids and medium color intensity. In Category three there are wines that are very low in color intensity and do not depend on Flavanoids.