NEWFINAL

Lizbeth, Nicole, Jenny, Anabel

2025-11-04

```r
#set working directory
setwd("/cloud/project")
###Importing my original data set to RStudio and calling it MCE
MCE<-read.csv("Mall_Customers_extended.csv", header=TRUE)

#to view variable names
names(MCE)
```

```
##  [1] "Unnamed..0"             "CustomerID"             "Genre"
##  [4] "Age"                    "Annual.Income..k.."     "Spending.Score..1.100."
##  [7] "IncomePerAge"           "SpendingEfficiency"     "AgeCategory"
## [10] "HighSpender"            "IncomeTier"             "OnlineShopFreq"
## [13] "LoyaltyScore"           "Satisfaction"           "CreditUtilization"
```

```r
#######  Step 1: Perform a K-Means Cluster analysis   #########
# I created MCEcluster1 data frame with the only 3 variables needed
# for my cluster: annual income, spending score, age

MCEcluster<-data.frame(MCE)

MCEcluster1<-MCEcluster[ ,c("Annual.Income..k..", "Spending.Score..1.100.", "Age")]

MCEcluster1
```

```
##     Annual.Income..k.. Spending.Score..1.100. Age
## 1                   15                     39  19
## 2                   15                     81  21
## 3                   16                      6  20
## 4                   16                     77  23
## 5                   17                     40  31
## 6                   17                     76  22
## 7                   18                      6  35
## 8                   18                     94  23
## 9                   19                      3  64
## 10                  19                     72  30
## 11                  19                     14  67
## 12                  19                     99  35
## 13                  20                     15  58
## 14                  20                     77  24
## 15                  20                     13  37
## 16                  20                     79  22
## 17                  21                     35  35
## 18                  21                     66  20
## 19                  23                     29  52
```

```
## 20                      23                       98   35
## 21                      24                       35   35
## 22                      24                       73   25
## 23                      25                        5   46
## 24                      25                       73   31
## 25                      28                       14   54
## 26                      28                       82   29
## 27                      28                       32   45
## 28                      28                       61   35
## 29                      29                       31   40
## 30                      29                       87   23
## 31                      30                        4   60
## 32                      30                       73   21
## 33                      33                        4   53
## 34                      33                       92   18
## 35                      33                       14   49
## 36                      33                       81   21
## 37                      34                       17   42
## 38                      34                       73   30
## 39                      37                       26   36
## 40                      37                       75   20
## 41                      38                       35   65
## 42                      38                       92   24
## 43                      39                       36   48
## 44                      39                       61   31
## 45                      39                       28   49
## 46                      39                       65   24
## 47                      40                       55   50
## 48                      40                       47   27
## 49                      40                       42   29
## 50                      40                       42   31
## 51                      42                       52   49
## 52                      42                       60   33
## 53                      43                       54   31
## 54                      43                       60   59
## 55                      43                       45   50
## 56                      43                       41   47
## 57                      44                       50   51
## 58                      44                       46   69
## 59                      46                       51   27
## 60                      46                       46   53
## 61                      46                       56   70
## 62                      46                       55   19
## 63                      47                       52   67
## 64                      47                       59   54
## 65                      48                       51   63
## 66                      48                       59   18
## 67                      48                       50   43
## 68                      48                       48   68
## 69                      48                       59   19
## 70                      48                       47   32
## 71                      49                       55   70
## 72                      49                       42   47
## 73                      50                       49   60
```

```
## 74                   50                     56  60
## 75                   54                     47  59
## 76                   54                     54  26
## 77                   54                     53  45
## 78                   54                     48  40
## 79                   54                     52  23
## 80                   54                     42  49
## 81                   54                     51  57
## 82                   54                     55  38
## 83                   54                     41  67
## 84                   54                     44  46
## 85                   54                     57  21
## 86                   54                     46  48
## 87                   57                     58  55
## 88                   57                     55  22
## 89                   58                     60  34
## 90                   58                     46  50
## 91                   59                     55  68
## 92                   59                     41  18
## 93                   60                     49  48
## 94                   60                     40  40
## 95                   60                     42  32
## 96                   60                     52  24
## 97                   60                     47  47
## 98                   60                     50  27
## 99                   61                     42  48
## 100                  61                     49  20
## 101                  62                     41  23
## 102                  62                     48  49
## 103                  62                     59  67
## 104                  62                     55  26
## 105                  62                     56  49
## 106                  62                     42  21
## 107                  63                     50  66
## 108                  63                     46  54
## 109                  63                     43  68
## 110                  63                     48  66
## 111                  63                     52  65
## 112                  63                     54  19
## 113                  64                     42  38
## 114                  64                     46  19
## 115                  65                     48  18
## 116                  65                     50  19
## 117                  65                     43  63
## 118                  65                     59  49
## 119                  67                     43  51
## 120                  67                     57  50
## 121                  67                     56  27
## 122                  67                     40  38
## 123                  69                     58  40
## 124                  69                     91  39
## 125                  70                     29  23
## 126                  70                     77  31
## 127                  71                     35  43
```

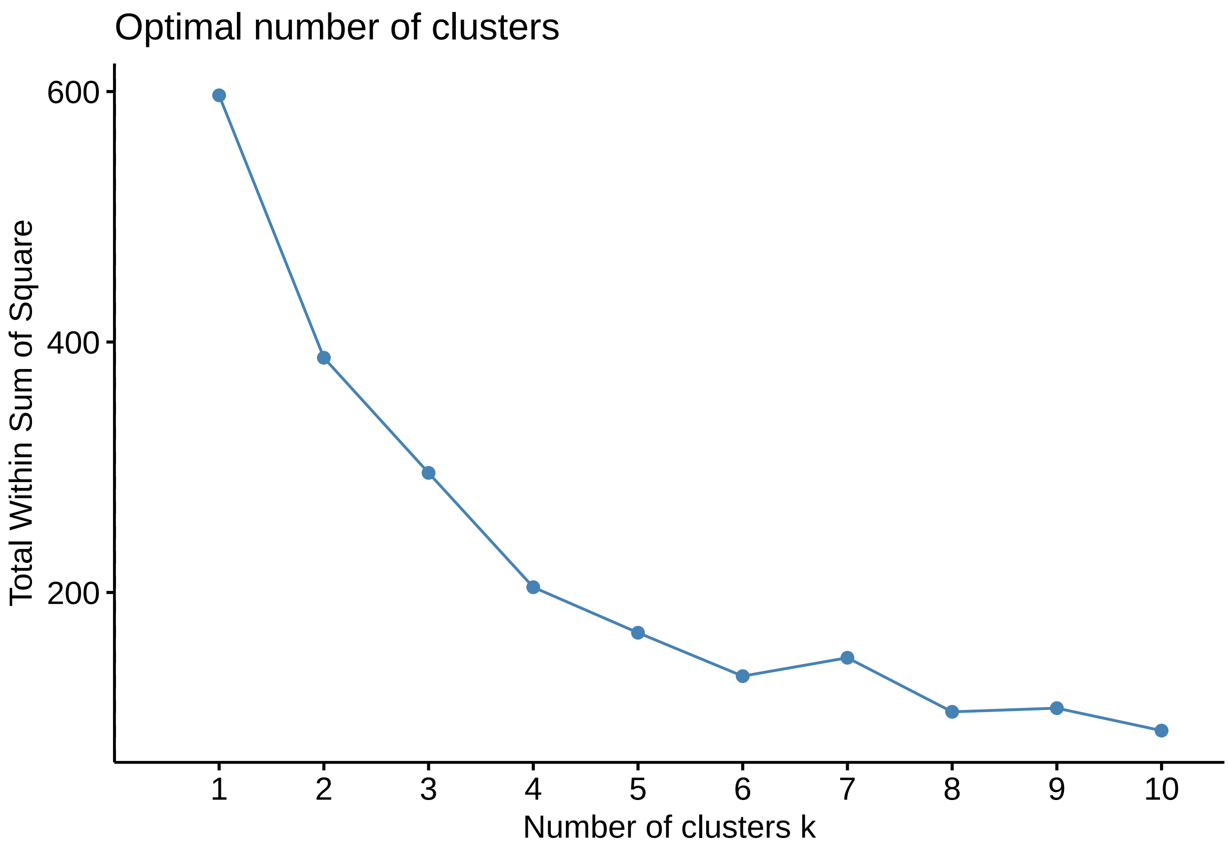
```
## 128                       71                       95  40
## 129                       71                       11  59
## 130                       71                       75  38
## 131                       71                        9  47
## 132                       71                       75  39
## 133                       72                       34  25
## 134                       72                       71  31
## 135                       73                        5  20
## 136                       73                       88  29
## 137                       73                        7  44
## 138                       73                       73  32
## 139                       74                       10  19
## 140                       74                       72  35
## 141                       75                        5  57
## 142                       75                       93  32
## 143                       76                       40  28
## 144                       76                       87  32
## 145                       77                       12  25
## 146                       77                       97  28
## 147                       77                       36  48
## 148                       77                       74  32
## 149                       78                       22  34
## 150                       78                       90  34
## 151                       78                       17  43
## 152                       78                       88  39
## 153                       78                       20  44
## 154                       78                       76  38
## 155                       78                       16  47
## 156                       78                       89  27
## 157                       78                        1  37
## 158                       78                       78  30
## 159                       78                        1  34
## 160                       78                       73  30
## 161                       79                       35  56
## 162                       79                       83  29
## 163                       81                        5  19
## 164                       81                       93  31
## 165                       85                       26  50
## 166                       85                       75  36
## 167                       86                       20  42
## 168                       86                       95  33
## 169                       87                       27  36
## 170                       87                       63  32
## 171                       87                       13  40
## 172                       87                       75  28
## 173                       87                       10  36
## 174                       87                       92  36
## 175                       88                       13  52
## 176                       88                       86  30
## 177                       88                       15  58
## 178                       88                       69  27
## 179                       93                       14  59
## 180                       93                       90  35
## 181                       97                       32  37
```

```
## 182                 97                      86  32
## 183                 98                      15  46
## 184                 98                      88  29
## 185                 99                      39  41
## 186                 99                      97  30
## 187                101                      24  54
## 188                101                      68  28
## 189                103                      17  41
## 190                103                      85  36
## 191                103                      23  34
## 192                103                      69  32
## 193                113                       8  33
## 194                113                      91  38
## 195                120                      16  47
## 196                120                      79  35
## 197                126                      28  45
## 198                126                      74  32
## 199                137                      18  32
## 200                137                      83  30
```

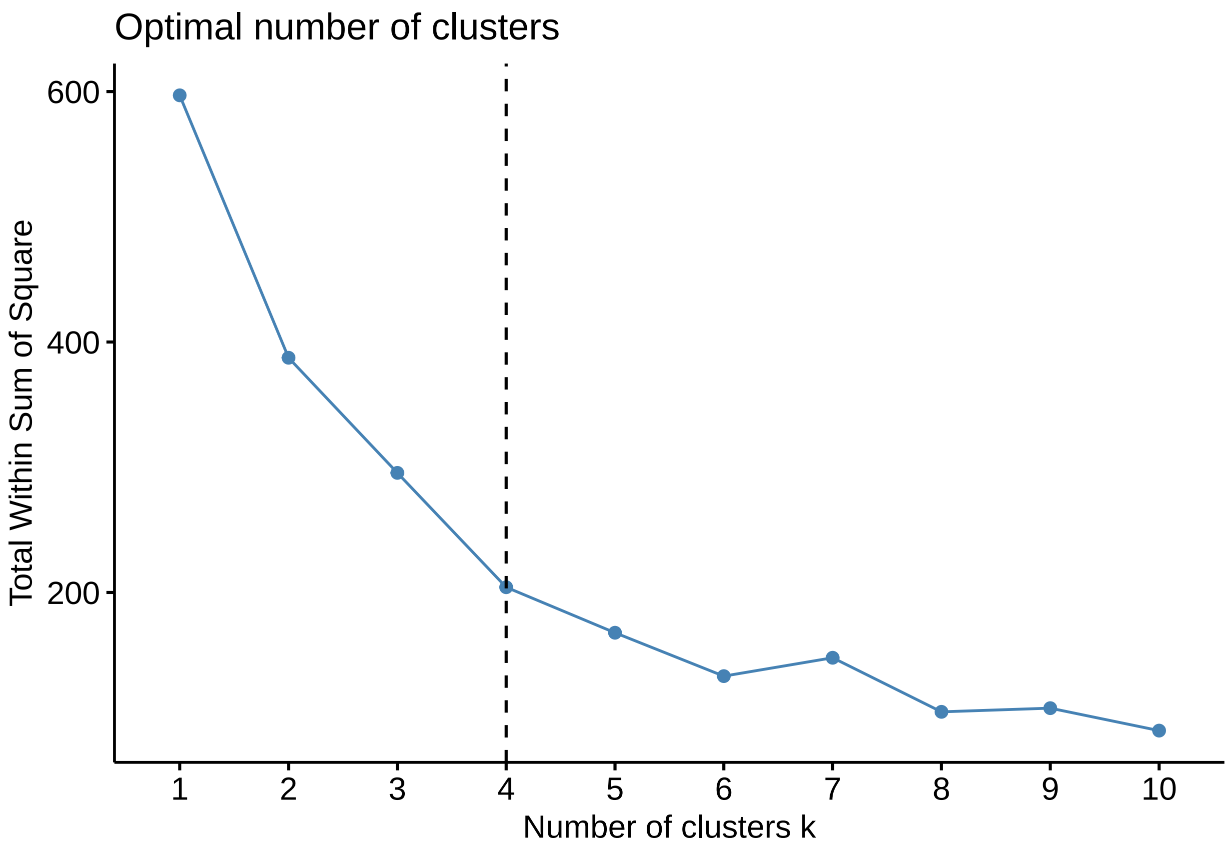
```
#Now, I'm going to create a new file called MCEsv with my 3 values standardized
#so that I can begin running my clusters
MCEcluster2<-data.frame(MCEcluster1)

MCEsv <- scale(MCEcluster2)

#View first 5 observations
head(MCEsv,n=5)
```

```
##      Annual.Income..k.. Spending.Score..1.100.        Age
## [1,]          -1.734646             -0.4337131 -1.4210029
## [2,]          -1.734646              1.1927111 -1.2778288
## [3,]          -1.696572             -1.7116178 -1.3494159
## [4,]          -1.696572              1.0378135 -1.1346547
## [5,]          -1.658498             -0.3949887 -0.5619583
```

```
###### K-Means Algorithm ######################
#The factoextra package creates clusters in R studio

install.packages("factoextra")
library(factoextra)

install.packages("rstatix")
library(rstatix)

#To find the number of clusters needed we use the fviz_nbclust function

fviz_nbclust(MCEsv, kmeans, method="wss") + geom_vline(xintercept = 0, linetype = 2)
```




```
## I'm using 4 clusters based on the graph in my plots box
fviz_nbclust(MCEsv, kmeans, method="wss") + geom_vline(xintercept = 4, linetype = 2)
```

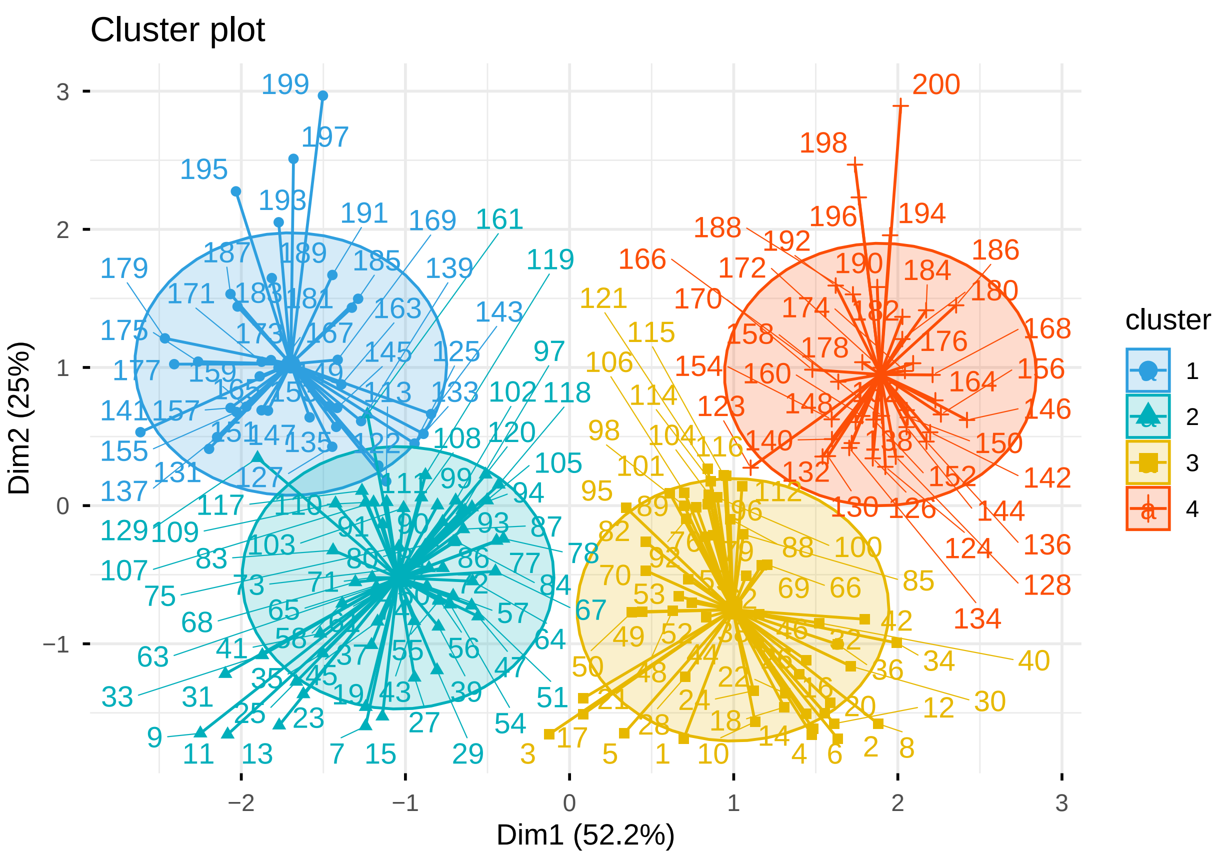



```r
#to obtain descriptive stats on 4 clusters
set.seed(123)

km.res <- kmeans(MCEsv, 4, nstart=25)
##Per homework instructions,  do not run line 52
print(km.res)
```

```
## K-means clustering with 4 clusters of sizes 38, 65, 57, 40
##
## Cluster means:
##   Annual.Income..k.. Spending.Score..1.100.         Age
## 1          0.9876366             -1.1857814  0.03711223
## 2         -0.4893373             -0.3961802  1.08344244
## 3         -0.7827991              0.3910484 -0.96008279
## 4          0.9724070              1.2130414 -0.42773261
##
## Clustering vector:
##   [1] 3 3 3 3 3 3 2 3 2 3 2 3 2 3 2 3 3 3 2 3 3 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2
##  [38] 3 2 3 2 3 2 3 2 3 2 3 3 3 2 3 3 2 2 2 2 2 2 3 2 2 3 2 2 2 3 2 2 3 3 2 2 2 2
##  [75] 2 3 2 2 3 2 2 3 2 2 3 2 2 3 3 2 2 3 2 2 3 3 2 3 2 3 3 2 2 3 2 3 2 2 2 2 2
## [112] 3 1 3 3 3 2 2 2 2 3 1 4 4 1 4 1 4 2 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4
## [149] 1 4 1 4 1 4 1 4 1 4 1 4 2 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1
## [186] 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4
##
## Within cluster sum of squares by cluster:
## [1] 44.01863 74.83280 61.43215 23.91544
##  (between_SS / total_SS =  65.8 %)
```

```
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```r
#####  Step 2: Create the Clustering Visual  ###############

#reating a new dataset using the cbind() function to merge the subset data
#created in step 1, and the km.res$clsuter

dd <- cbind(MCEsv, cluster=km.res$cluster)

fviz_cluster(km.res, data=dd,
             palette = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),
             ellipse.type = "euclid", #Concentration ellipse
             star.plot = TRUE, #Add segments from centroids to items
             repel = TRUE, #Avoid label overplotting
             ggtheme = theme_minimal())
```

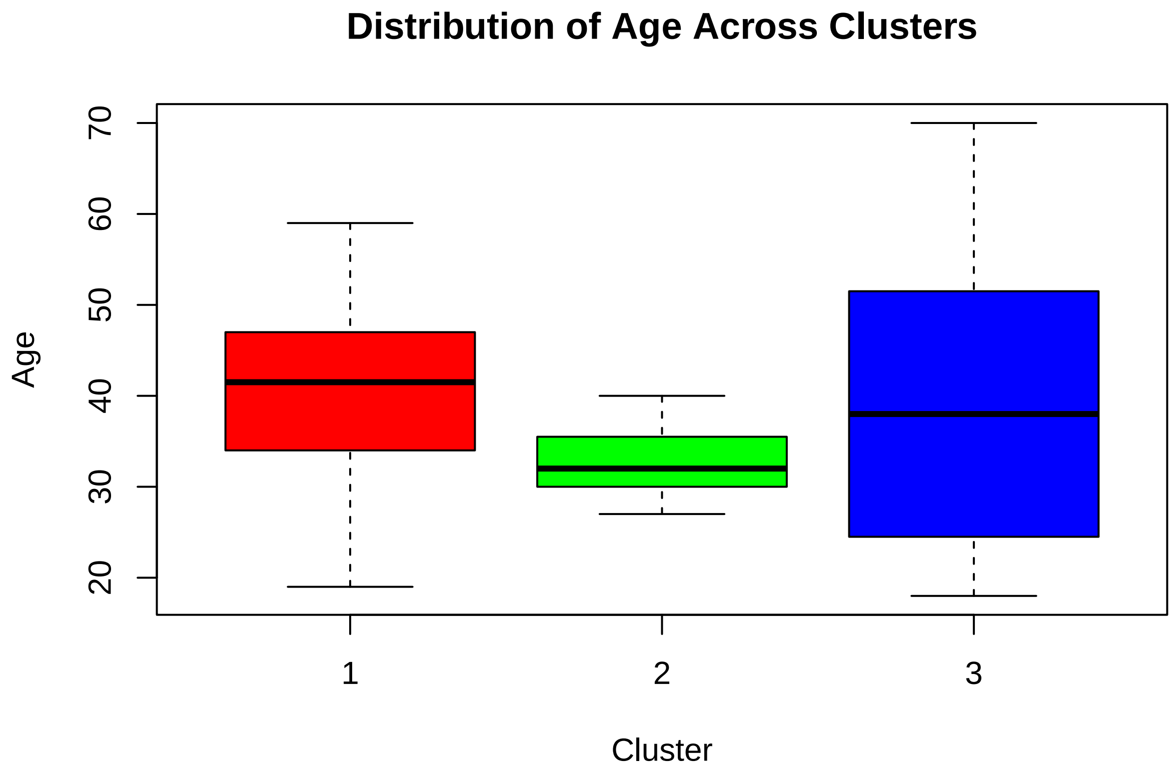



```r
#boxplot
# Step 1: Select columns for clustering
data_for_clustering <- MCE[, c("Age", "Annual.Income..k..", "Spending.Score..1.100.")]

# Step 2: Run k-means clustering

set.seed(123)   # for reproducibility
km.res <- kmeans(data_for_clustering, centers = 3, nstart = 25)
```

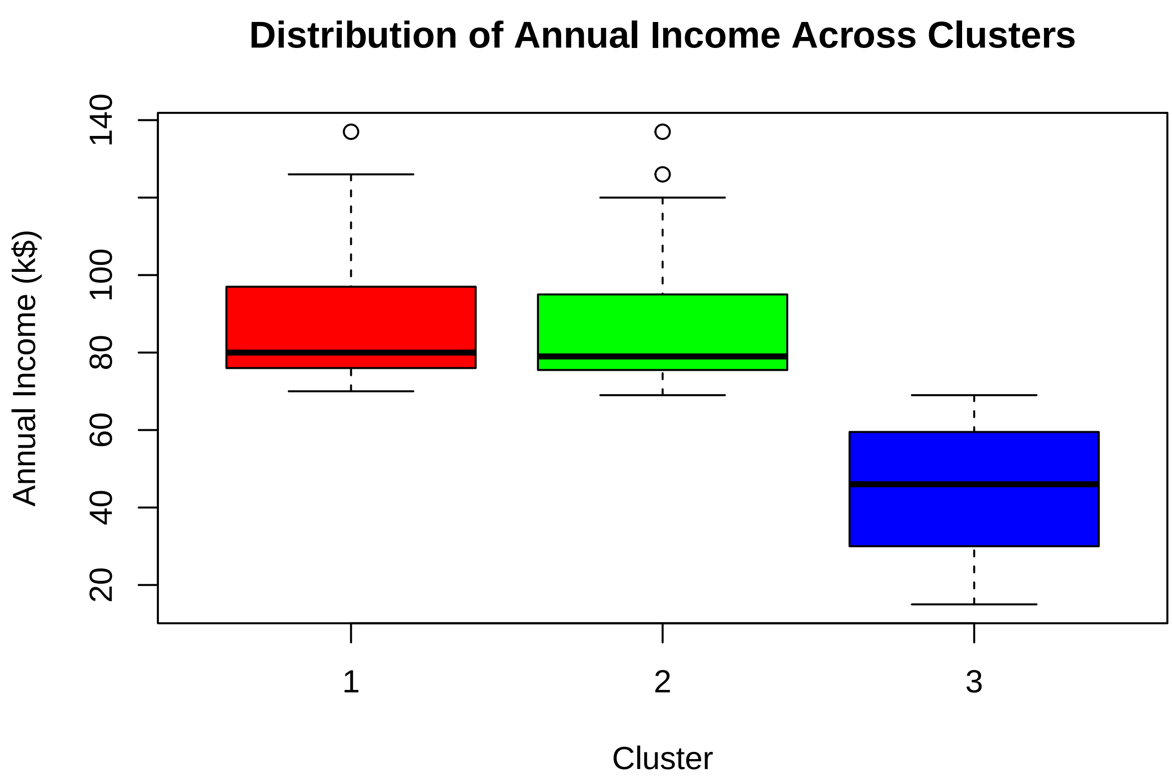
```r
# Step 3:  cluster assigned
km.res$cluster
```

```
##   [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
##  [38] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
##  [75] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
## [112] 3 3 3 3 3 3 3 3 3 3 3 3 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2
## [149] 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1
## [186] 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2
```

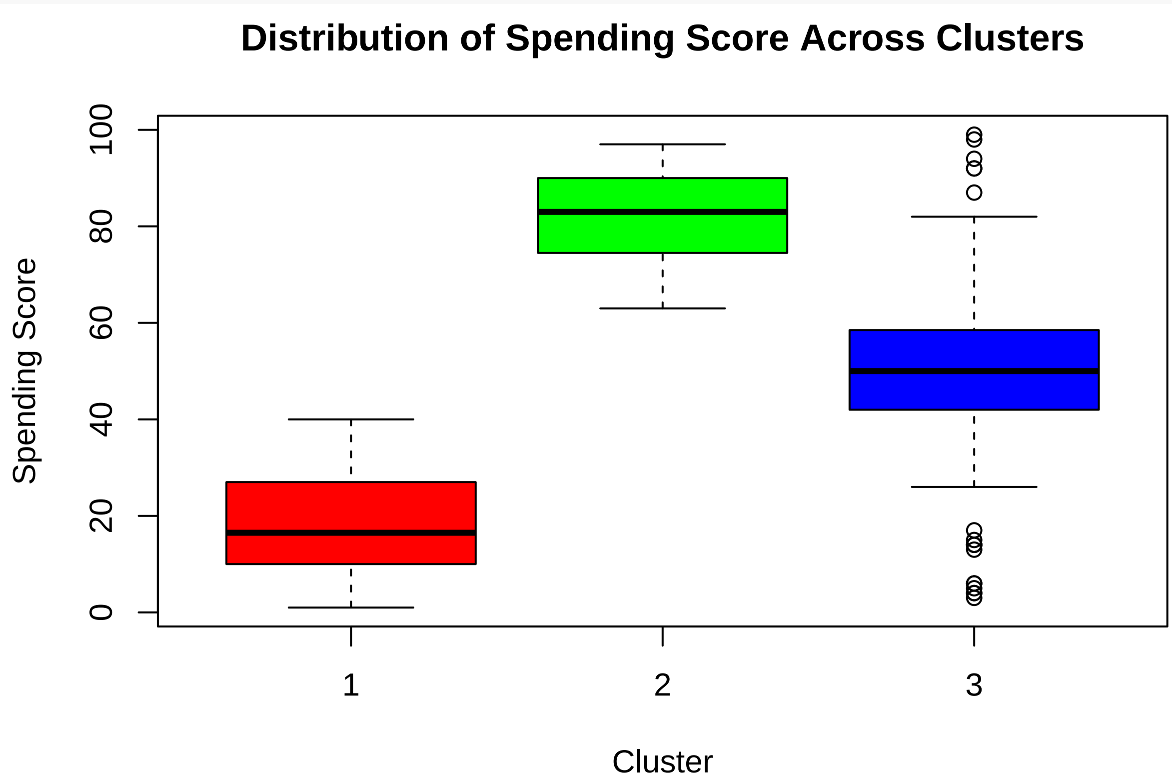
```r
# Merge cluster  with original data
MCE_clustered <- cbind(MCE, Cluster = km.res$cluster)

# Boxplot for Age across clusters
boxplot(Age ~ Cluster, data = MCE_clustered,
        main = "Distribution of Age Across Clusters",
        xlab = "Cluster", ylab = "Age",
        col = c("red", "green", "blue"))
```




```r
# Boxplot for Annual Income across clusters
boxplot(Annual.Income..k.. ~ Cluster, data = MCE_clustered,
        main = "Distribution of Annual Income Across Clusters",
        xlab = "Cluster", ylab = "Annual Income (k$)",
        col = c("red", "green", "blue"))
```




```r
# Boxplot for Spending Score across clusters
boxplot(Spending.Score..1.100. ~ Cluster, data = MCE_clustered,
        main = "Distribution of Spending Score Across Clusters",
        xlab = "Cluster", ylab = "Spending Score",
        col = c("red", "green", "blue"))
```



```r
######################
# descriptive tables #
######################


install.packages("stargazer")
library(stargazer)

cluster1 <- subset(MCE_clustered, Cluster == 1)
cluster2 <- subset(MCE_clustered, Cluster == 2)
cluster3 <- subset(MCE_clustered, Cluster == 3)

variables <- c("Age", "Annual.Income..k..", "Spending.Score..1.100.")

stargazer(cluster1[, variables], type = "text", title = "Cluster 1 Descriptive Stats")
```

```
##
## Cluster 1 Descriptive Stats
## =================================================
## Statistic              N   Mean  St. Dev. Min Max
## -------------------------------------------------
## Age                    38 40.395  11.377  19  59
## Annual.Income..k..     38 87.000  16.271  70  137
## Spending.Score..1.100. 38 18.632  10.916   1  40
## -------------------------------------------------
```

```r
stargazer(cluster2[, variables], type = "text", title = "Cluster 2 Descriptive Stats")
```

```
##
## Cluster 2 Descriptive Stats
## =================================================
## Statistic              N   Mean  St. Dev. Min Max
## -------------------------------------------------
## Age                    39 32.692  3.729   27  40
## Annual.Income..k..     39 86.538  16.312  69  137
## Spending.Score..1.100. 39 82.128  9.364   63  97
## -------------------------------------------------
```

```r
stargazer(cluster3[, variables], type = "text", title = "Cluster 3 Descriptive Stats")
```

```
##
## Cluster 3 Descriptive Stats
## ==================================================
## Statistic               N   Mean  St. Dev. Min Max
## --------------------------------------------------
## Age                    123 40.325  16.114  18  70
## Annual.Income..k..     123 44.154  16.038  15  69
## Spending.Score..1.100. 123 49.829  19.694   3  99
## --------------------------------------------------
```

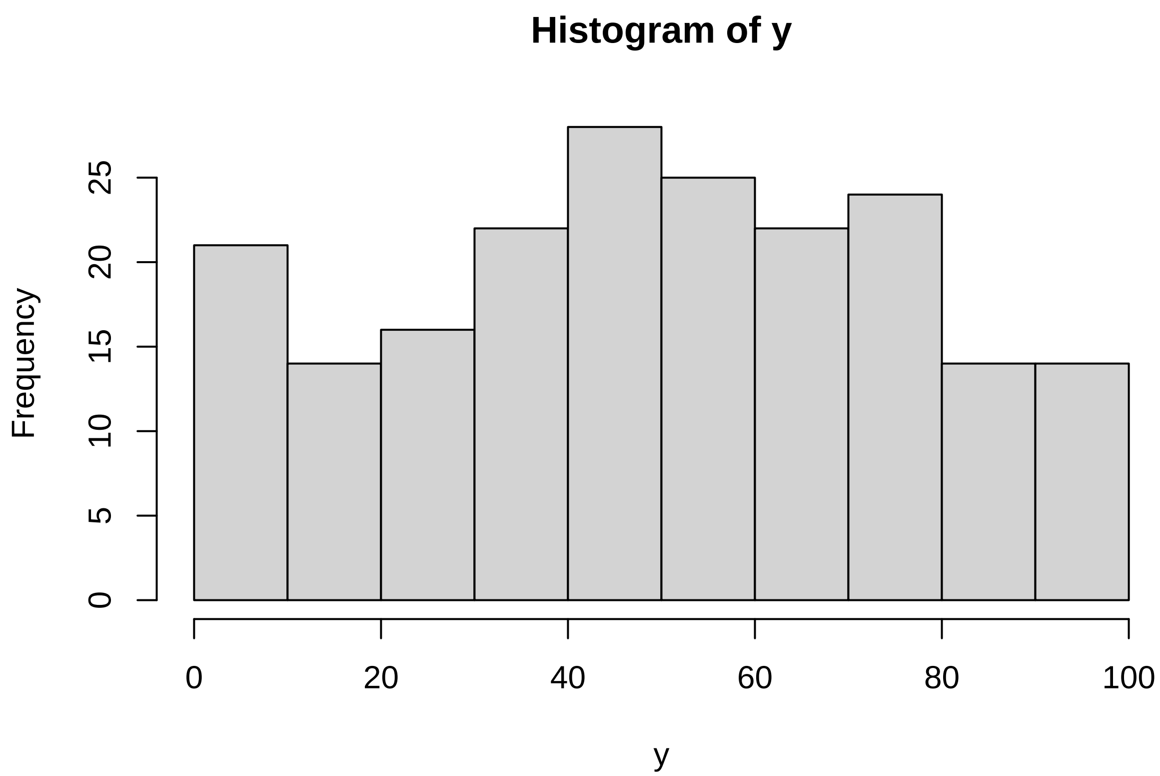
```r
#How to do multiple linear regression in R

MallData <- read.csv("Mall_Customers_extended.csv", header=TRUE)


names(MallData)
```

```
##  [1] "Unnamed..0"            "CustomerID"            "Genre"
##  [4] "Age"                   "Annual.Income..k.."    "Spending.Score..1.100."
##  [7] "IncomePerAge"          "SpendingEfficiency"    "AgeCategory"
## [10] "HighSpender"           "IncomeTier"            "OnlineShopFreq"
## [13] "LoyaltyScore"          "Satisfaction"          "CreditUtilization"
```

```
y <- MallData$y.LoyaltyScore

x1 <- MallData$Age

x2 <- MallData$Annual.Income..k..

x3 <-MallData$Spending.Score..1.100.

str(MallData$LoyaltyScore)
```

```
##  num [1:200] 46 75 28 83.9 39.9 ...
```

```
y <- as.numeric(as.character(MallData$LoyaltyScore))

hist(y)
```




```
model2 <- lm(y ~ x1 + x2 + x3, data=MallData)
model2
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = MallData)
##
## Coefficients:
## (Intercept)           x1           x2           x3
```

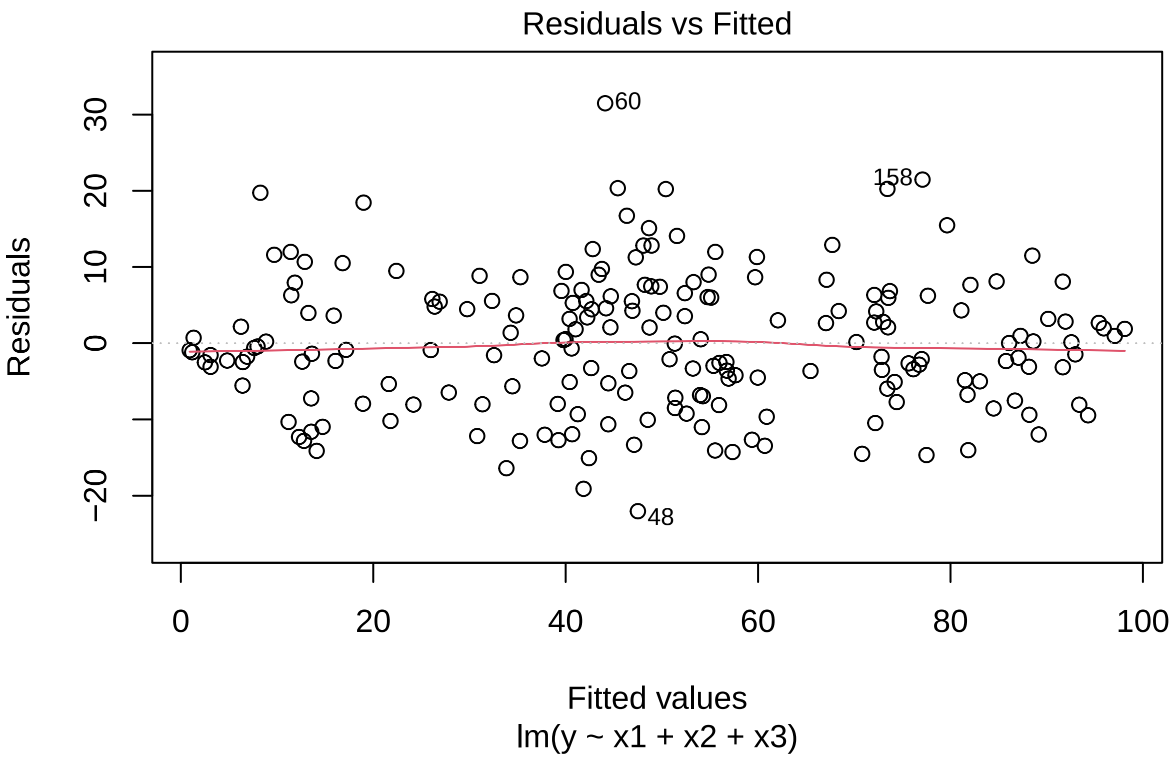
```
##     4.40942     -0.08975     -0.01474      0.98109
```

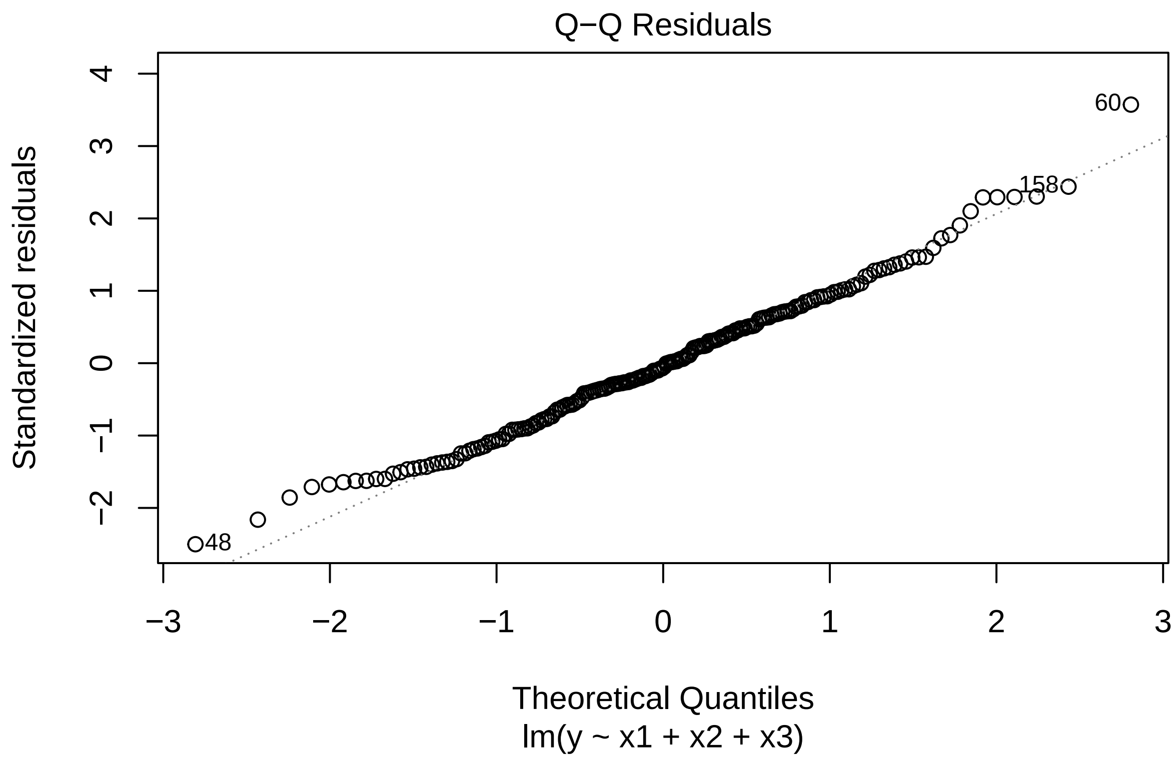
```r
attach(MallData)


summary(model2)
```

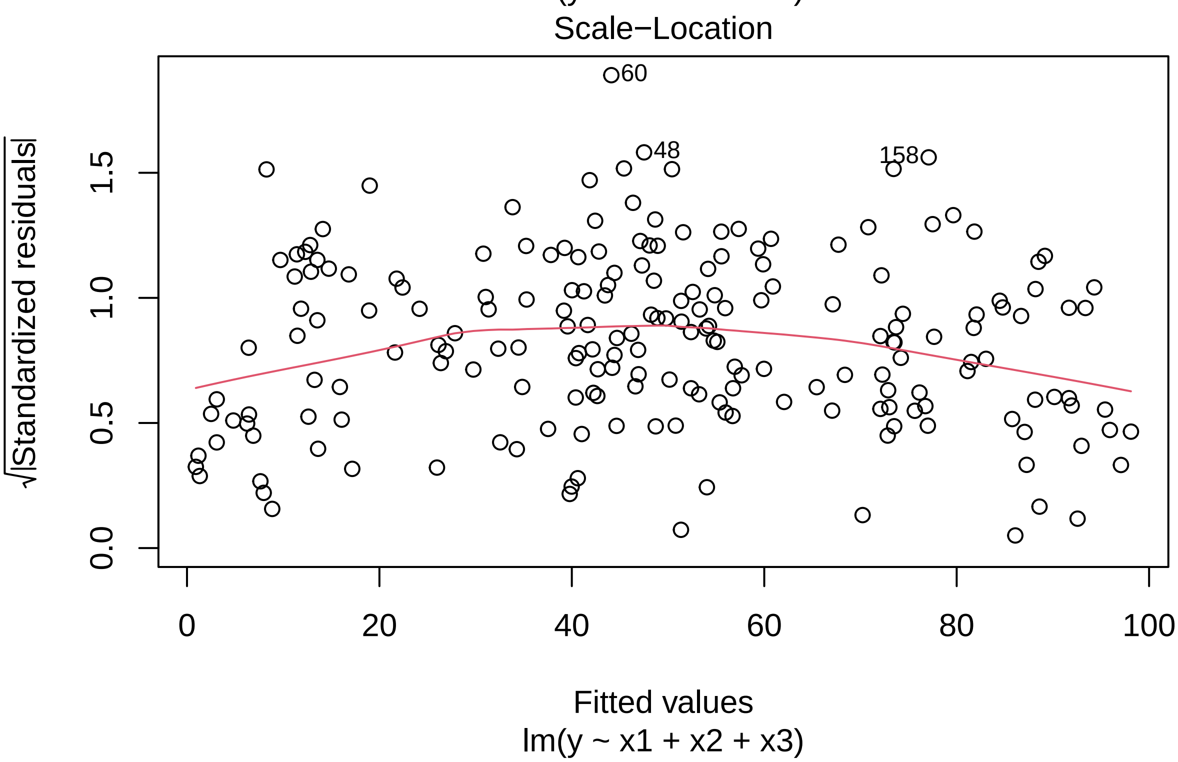
```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = MallData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.0324  -6.4708  -0.5225   5.9657  31.4818
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.40942    3.02939   1.456   0.1471
## x1          -0.08975    0.04761  -1.885   0.0609 .
## x2          -0.01474    0.02393  -0.616   0.5385
## x3           0.98109    0.02575  38.099   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.864 on 196 degrees of freedom
## Multiple R-squared:  0.8956, Adjusted R-squared:  0.894
## F-statistic: 560.7 on 3 and 196 DF,  p-value: < 2.2e-16
```

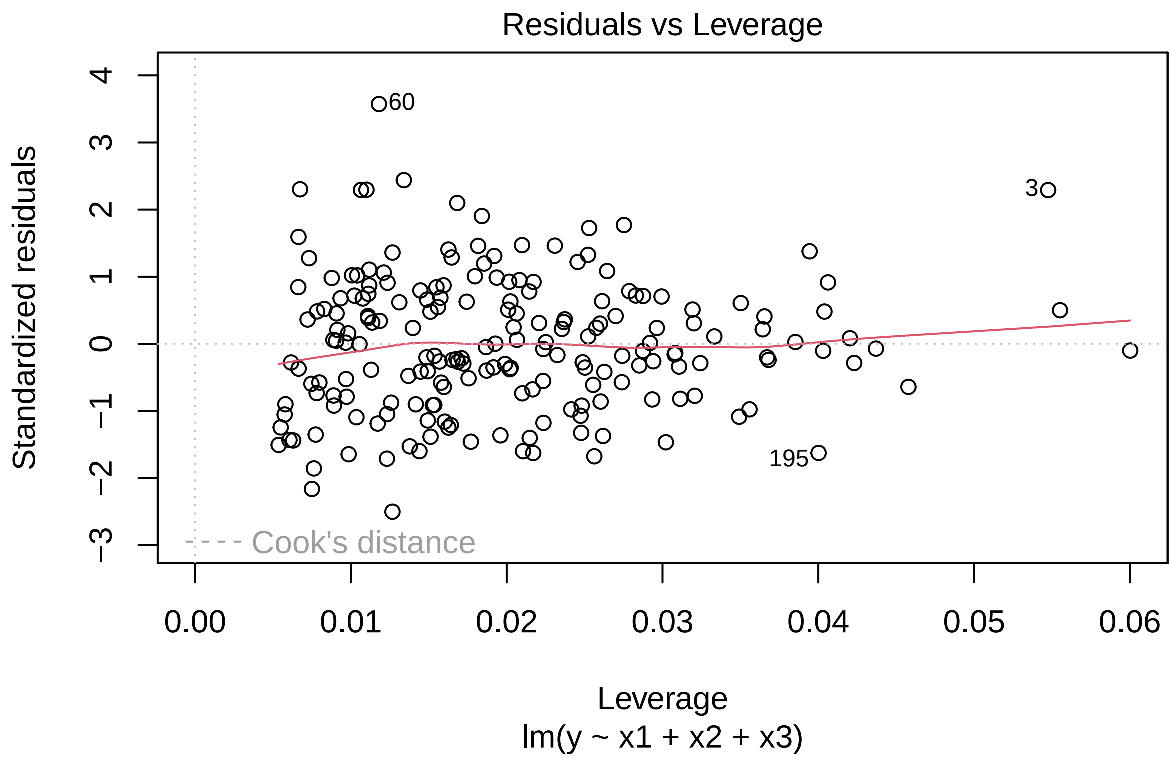
```r
#Intercepts: There are 3 x-values that are independent: these are x1 is the
#intercept for Age. x2 is the intercept for Annual Income K and x3 is the
#intercept for Spending Score.The estimated Beta for x1 is -0.08975, x2 is
#-0.01474 and x3 is 0.98109. There are 3 different variables with different
#outcomes.Age, Annual Income K and Spending Score are all continuous variables.
#For every unit in x1 (Age), there is a decrease of -0.0875 for the outcome
#(LoyaltyScore). Applies for x2, there is a decrease of -0.01474 for the outcome
#and for x3, there is an increase of 0.98108 for the outcome.

#Based on the summary, x1 and x2 are not significant. They do not contain the
#asterick. In this multiple Linear Regression ,x3 significantly predicts the
#outcome. For every unit in x3, the expected outcome increases by 0.98109
#which is adjusting for x1 and x2.

#Mulitple R-Squared: This is used for Multiple Linear Regression to avoid bias
#when trying to find associations of x varibles with outcome. When adding more
#variables R, usually increases, therefore we avoid using Multiple R-squared in
#this situation. Thus, they both provide the proportion of variability that
#the x's will explain on the variance of the outcome. It is ideal for it to be
#high, it should be close to 1. This would mean that all of systemcatic
#variances are being accounted for, in this model, it is high of 0.894.

#F-Statistics: The F-statistics is more than 1, it is 560.7, this means that
#p-value is significant as well. When F-statistics is more than 1, it means that
#the systematic variances are being accounted by these x variables and they
#outweigh the unsystematic variances.
```

```r
#p-Value: This would be considered a good model because it is less than 0.05,
#shows that it is statistically significant.

#Income does not predict more Loyalty stronger than other clusters. The cluster
#that show to have a stronger Loyalty is x3 (Spending Score) with a positive
#beta of 0.98109. The other betas for x1 and x2 variable show decrease with
#association of outcome.




#Assessing fit of model
plot(model2)
```




Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(y ~ x1 + x2 + x3)
60
158
48


Scale−Location
√|Standardized residuals|
Fitted values
lm(y ~ x1 + x2 + x3)
60
48
158


Residuals vs Leverage
Standardized residuals
Leverage
lm(y ~ x1 + x2 + x3)
Cook's distance
60
3
195
4
3
2
1
0
−1
−2
−3
0.00
0.01
0.02
0.03
0.04
0.05
0.06