An overview of machine learning problems

Vladan Devedzic

(Installing and) Loading the required R packages

```
# install.packages('corrplot')
library(corrplot)
library(ggplot2)
```

Linear regression

Reading the dataset

A modified or newly created dataset might have been saved earlier using:

```
write.csv(x = <dataframe>, file = "<filename>", row.names = F)  # do not include the row names (row numbers) column
saveRDS(object = <dataframe or another R object>, file = "<filename>")  # save R object for the next session
```

Restoring the dataset from the corresponding RData file:

```
<dataframe or another R object> <- readRDS(file = "<filename>")         # restore R object in the next session
The Beatles songs dataset has been saved earlier using:
```

saveRDS(the.beatles.songs, “The Beatles songs dataset, v2.1.RData”)

```
the.beatles.songs <- readRDS("The Beatles songs dataset, v2.1.RData")
summary(the.beatles.songs)
```

```
##     Title                Year       Duration     Other.releases
##  Length:310         1963   :67   Min.   : 23.0   Min.   : 0.00
##  Class :character   1968   :45   1st Qu.:133.0   1st Qu.: 0.00
##  Mode  :character   1969   :42   Median :148.0   Median : 9.00
##                     1964   :41   Mean   :159.9   Mean   :10.42
##                     1965   :37   3rd Qu.:174.0   3rd Qu.:16.00
##                     1967   :27   Max.   :502.0   Max.   :56.00
##                     (Other):51
##  Top.50.Billboard
##  Min.   : 0.000
##  1st Qu.: 0.000
##  Median : 0.000
##  Mean   : 4.061
##  3rd Qu.: 0.000
##  Max.   :50.000
##
```

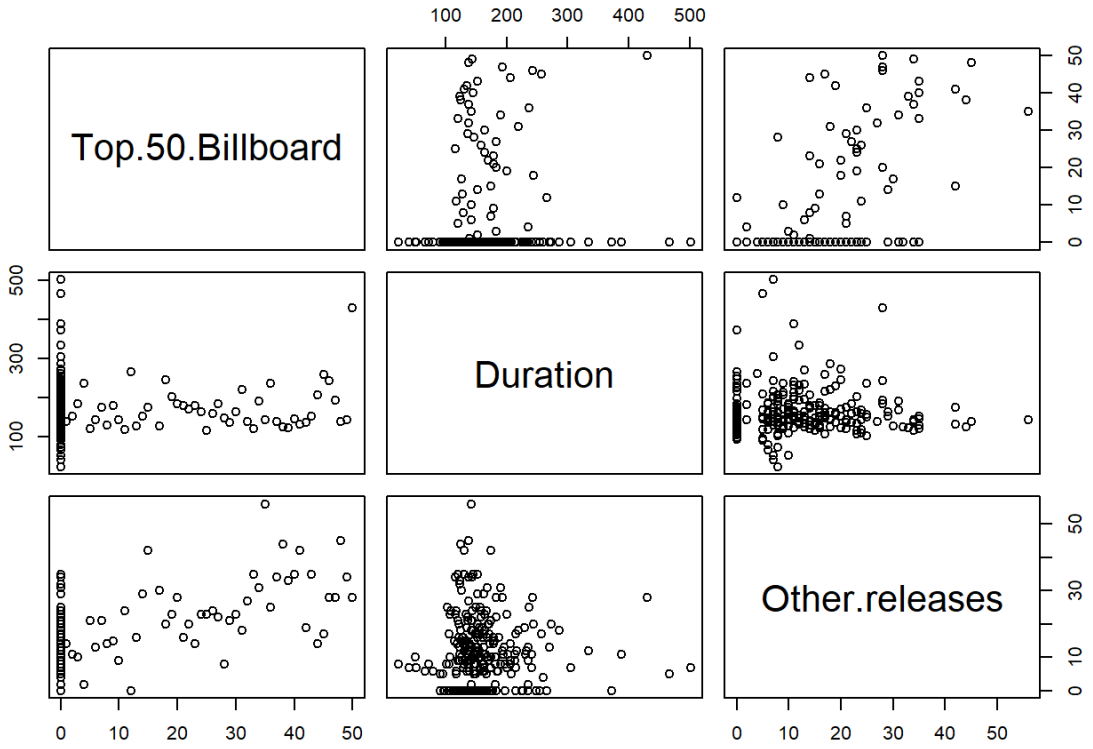
Examining the data

Scatterplot matrices

Scatterplot matrices are uuseful for examining the presence of linear relationship between several pairs of variables:

```
pairs(~<x1> + <x2> + ..., data = <dataframe>)
```

```
pairs(~Top.50.Billboard + Duration + Other.releases, data = the.beatles.songs)
```



Correlation plots

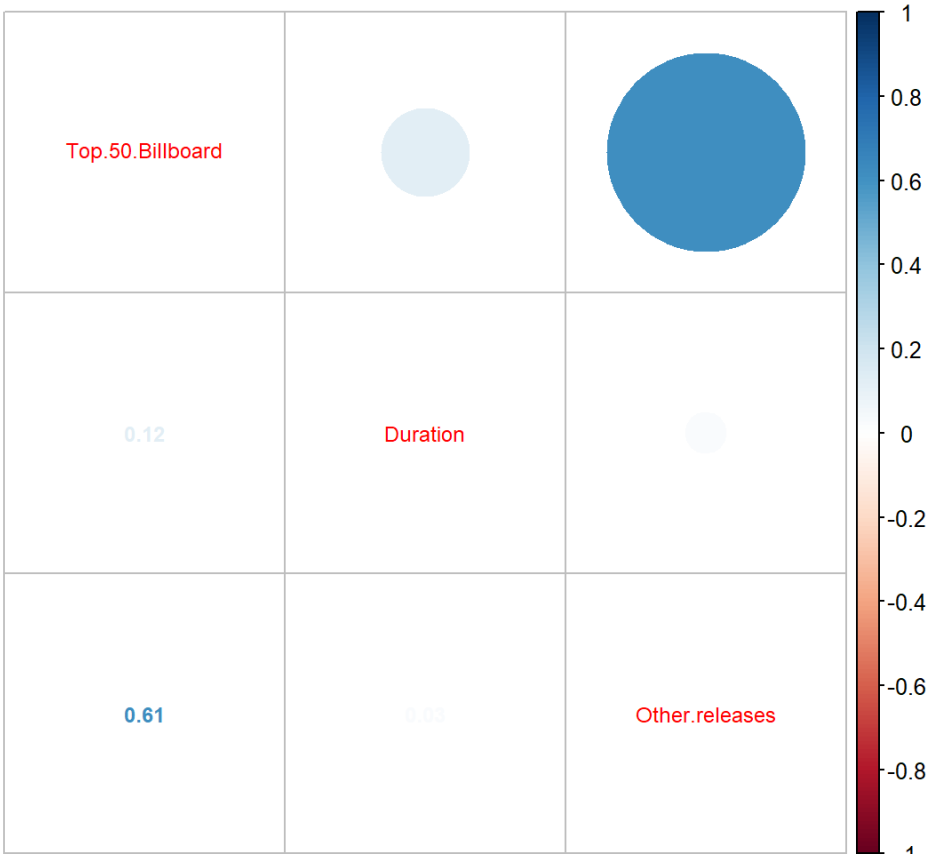
Visual representation of correlations between numeric variables in the dataset:

```
<numeric dataframe> <-                                          # create all-numeric dataframe,
+ data.frame(<num col 1 name> = <dataframe>$<num col 1>,       # leave out all non-numeric columns
+            <num col 2 name> = <dataframe>$<num col 2>,       # from the original dataframe
+            ...)
<correlation matrix> <- cor(<numeric dataframe>)               # all-numeric dataframe
library(corrplot)
corrplot.mixed(<correlation matrix>, tl.cex = <text font size>, number.cex = <number font size>)
```

```
the.beatles.songs.num <- data.frame(Top.50.Billboard = the.beatles.songs$Top.50.Billboard,
                                    Duration = the.beatles.songs$Duration,
                                    Other.releases = the.beatles.songs$Other.releases)
correlation.matrix <- cor(the.beatles.songs.num)
correlation.matrix
```

```
##                 Top.50.Billboard   Duration Other.releases
## Top.50.Billboard       1.0000000 0.12050340     0.61325609
## Duration               0.1205034 1.00000000     0.02617334
## Other.releases         0.6132561 0.02617334     1.00000000
```

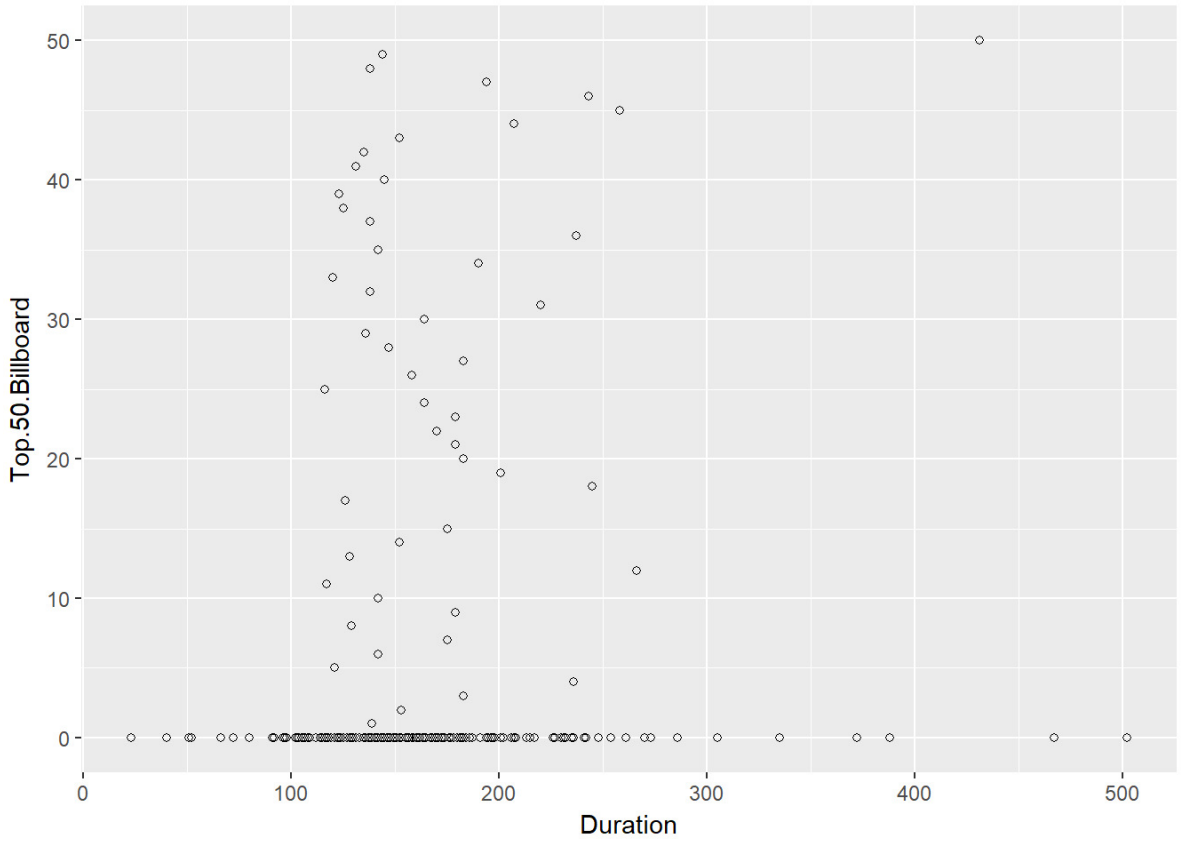
```
library(corrplot)
corrplot.mixed(correlation.matrix, tl.cex = 0.75, number.cex = 0.75)
```



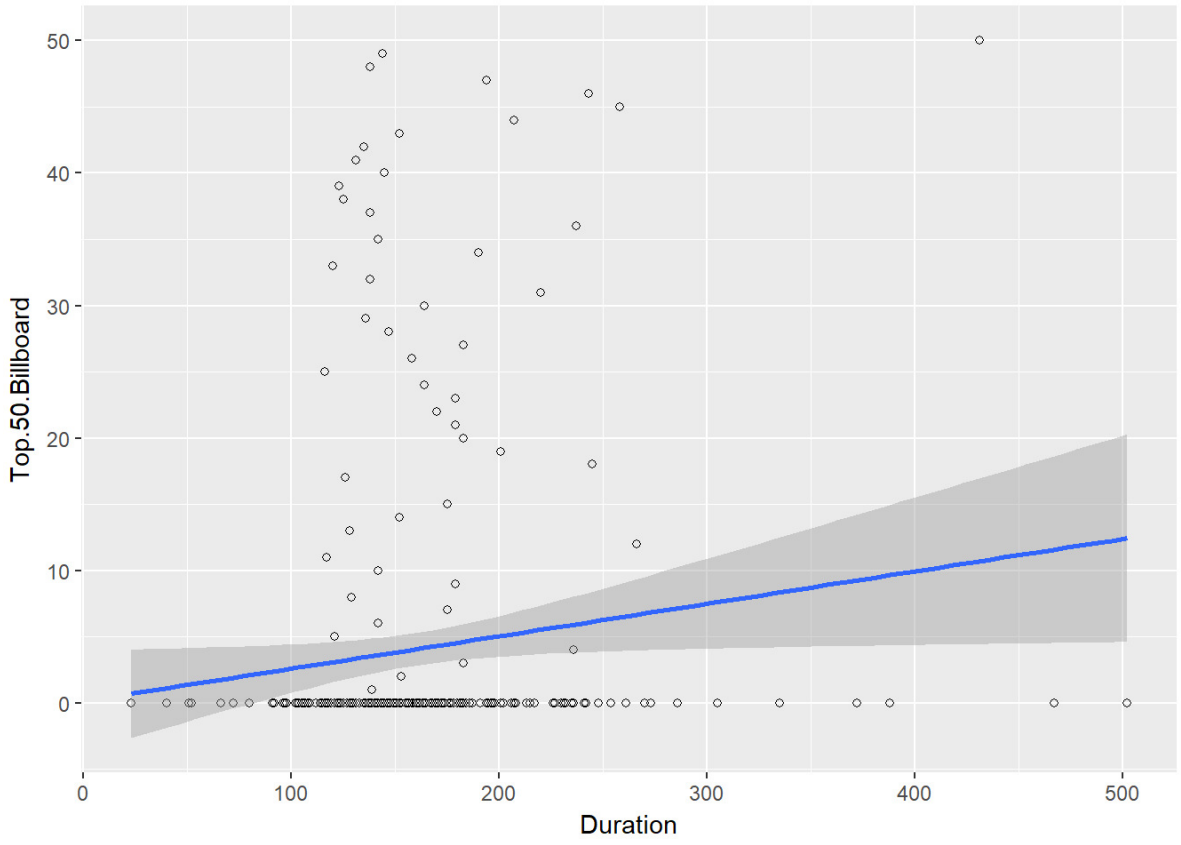
Scatterplots in ggplot2

```
ggplot(<dataset>, aes(x = <num.var.1>, y = <num.var.2>)) +
+ geom_point(shape = <n>,       # <n> = 1: hollow circle
+            fill = <color 1>,  # color of point fill (optional)
+            color = <color 2>, # color of point line (optional)
+            size = <s>) +      # size  of point line (optional)
+ geom_smooth(method = lm,      # add regression line (optional); if left out, nonlinear best-fit line  is shown
+              se=FALSE)         # do NOT show 95% confidence region as a shaded area (optional)
```

```
g1 <- ggplot(the.beatles.songs, aes(x = Duration, y = Top.50.Billboard)) +
  geom_point(shape = 1)
g1
```

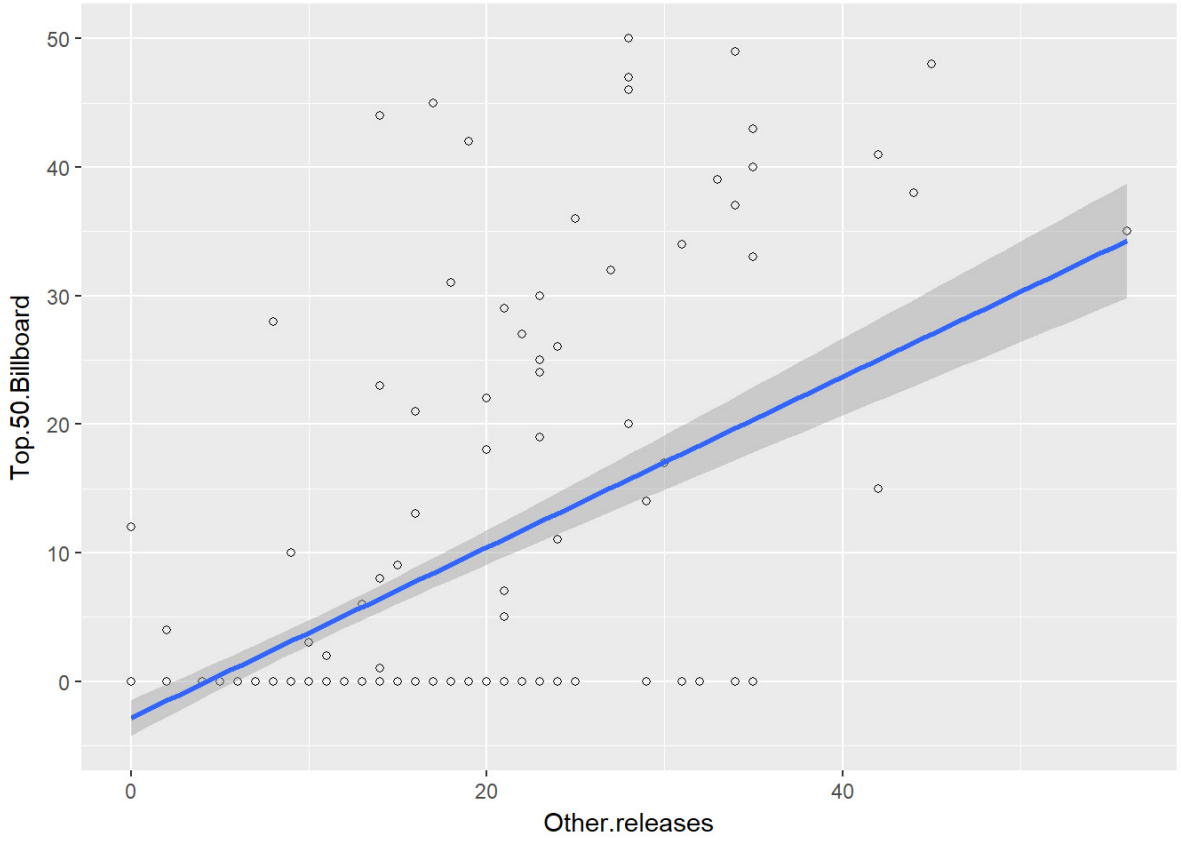



```
g1 <- ggplot(the.beatles.songs, aes(x = Duration, y = Top.50.Billboard)) +
  geom_point(shape = 1) +
  geom_smooth(method = lm)  # linear regression
g1
```



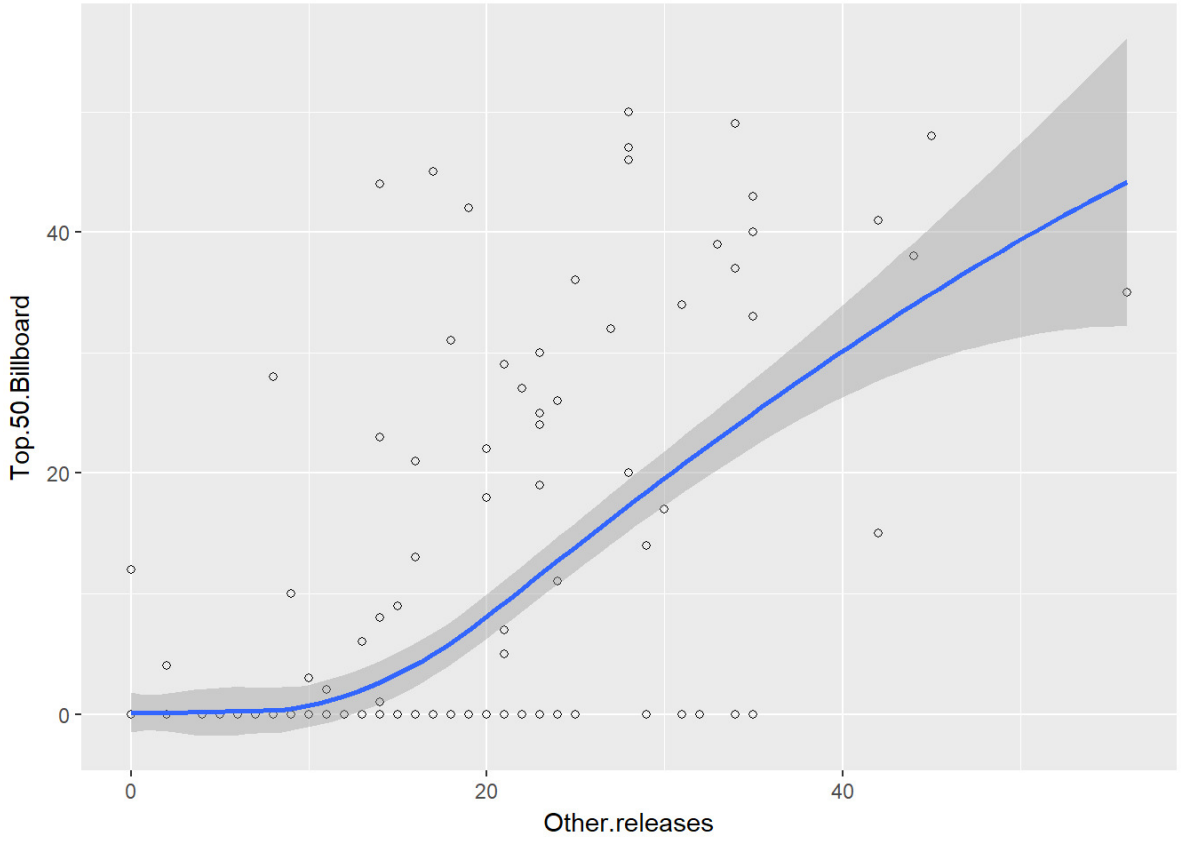
```
# g1 <- ggplot(the.beatles.songs, aes(x = Duration, y = Top.50.Billboard)) +
#   geom_point(shape = 21, size = 2, fill = "green", color = "red") +
#   geom_smooth(method = lm)
# g1

g2 <- ggplot(the.beatles.songs, aes(x = Other.releases, y = Top.50.Billboard)) +
  geom_point(shape = 1) +
  geom_smooth(method = lm)
g2
```




```
g3 <- ggplot(the.beatles.songs, aes(x = Other.releases, y = Top.50.Billboard)) +
  geom_point(shape = 1) +
  geom_smooth()             # non-linear regression
g3
```

```
## `geom_smooth()` using method = 'loess'
```



Build/Fit simple linear regression model and examine it

```
<model> <- lm(<y> ~ <x>,          # build/fit the model over the <dataset>;
+             data = <dataset>)   # <x> and <y> are numeric variables from <dataset>
<model>                   # show the model
coef(<model>)             # show the coefficients of the linear model (intercept and slope)
confint(<model>)          # show the confidence intervals for the estimated intercept and slope
summary(<model>)          # show the model statistics
```

```
lm.fit <- lm(Top.50.Billboard ~ Other.releases, data = the.beatles.songs)
lm.fit
```

```
##
## Call:
## lm(formula = Top.50.Billboard ~ Other.releases, data = the.beatles.songs)
##
## Coefficients:
##    (Intercept)  Other.releases
##         -2.849           0.663
```

```
coef(lm.fit)
```

```
##    (Intercept) Other.releases
##     -2.8486751      0.6629803
```

```
confint(lm.fit)
```

```
##                    2.5 %     97.5 %
## (Intercept)    -4.2436506 -1.4536996
## Other.releases  0.5672378  0.7587228
```

```
summary(lm.fit)
```

```
##
## Call:
## lm(formula = Top.50.Billboard ~ Other.releases, data = the.beatles.songs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.356  -4.444  -1.129   2.849  37.567
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2.84868    0.70894  -4.018 7.38e-05 ***
## Other.releases  0.66298    0.04866  13.626  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.722 on 308 degrees of freedom
## Multiple R-squared:  0.3761, Adjusted R-squared:  0.3741
## F-statistic: 185.7 on 1 and 308 DF,  p-value: < 2.2e-16
```

Make predictions

```
predict(<model>,                          # the model built above
+       <test dataframe>,                 # data frame built over a (small) vector of <x> to predict  <y> for;
+                                         # in the <test dataframe> the name of <x> as in the original dataframe
+       interval = "confidence" |         # include the confidence interval for the predictions (optional)
+                  "predict")             # include prediction intervals (optional)
```

```
the.beatles.songs.num <- data.frame(Other.releases = c(5, 15, 25))
predict(lm.fit, newdata = the.beatles.songs.num, interval = "confidence")
```

```
##          fit        lwr       upr
## 1  0.4662263 -0.6381845  1.570637
## 2  7.0960291  6.0272679  8.164790
## 3 13.7258319 12.0234503 15.428213
```

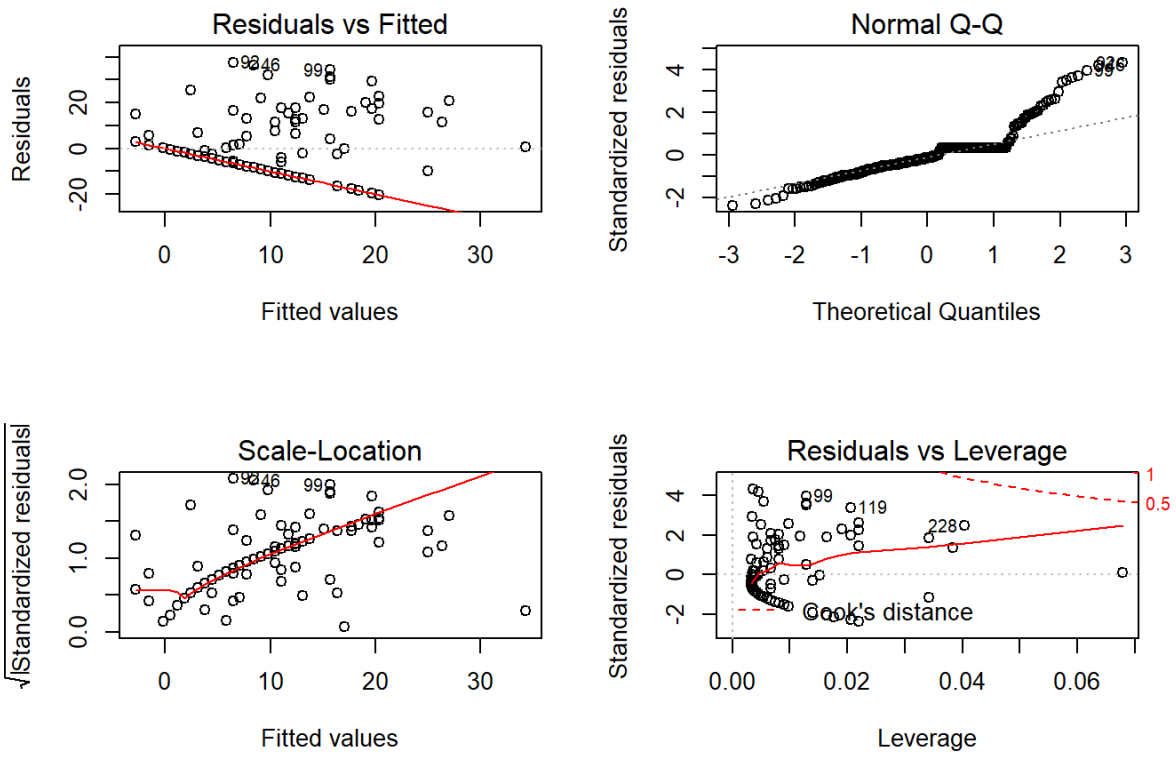
```
predict(lm.fit, newdata = the.beatles.songs.num, interval = "predict")
```

```
##          fit        lwr      upr
## 1  0.4662263 -16.731937 17.66439
## 2  7.0960291 -10.099882 24.29194
## 3 13.7258319  -3.521058 30.97272
```

Check how well the model fits the data

```
par(mfrow = c(2,2))     # set up the plotting panel for 4 graphs
plot(<model>)           # plot the 4 graphs
par(mfrow = c(1,1))     # reset the plotting panel
```

```
par(mfrow=c(2,2))
plot(lm.fit)
```




```
par(mfrow = c(1,1))
```

The 4 graphs:

- Residuals vs Fitted: Is linear assumption justified?
  - Can a non-linear pattern be observed in the residuals? If so, there is a non-linearity not explained by the model…
  - Are the residuals more-or-less evenly (randomly) distributed around the horizontal dotted line? It's better if they are (the linearity assumption is then likely to hold). Ideally (but very unlikely), the red line is overlapped with the horizontal dotted line.
- Q-Q plot: Are the residuals normally distributed (reside on the diagonal line)? It's good if they are.
- Scale-Location: Is the variance of residuals similar (even) along the fitted line?
  - Are the residuals spread evenly along the range(s) of predictor(s)? It's good if they are, and in that case the red line is more-or-less horizontal.
- Residuals vs Leverage: Are there extreme values of a predictor (points of high leverage) that shouldn't be excluded from the analysis?
  - Are there data points outside the dashed red line (having a high Cook's distance score)? If so, they should be given special attention. If they result from erroneous data, they can be excluded from the analysis. Otherwise, they shouldn't be excluded from the analysis, because R-squared and the slope will change a lot. In that case, linear regression is not applicable.

The four plots show potential problematic cases with the row numbers of the data in the dataset. If some cases are identified across all four plots (Residuals vs Leverage being especially critical), or at least on all plots other than Q-Q plot, it is a good idea to take a closer look at them individually. Is there anything special for the subject? Or could they be simply errors in data entry?

Classification - decision trees

Reading the dataset

The dataset has been prepared earlier and saved using:

```
saveRDS(the.beatles.songs, "The Beatles songs dataset, v2.2.RData")
```

Read the dataset using:

```
<dataframe or another R object> <- readRDS(file = "<filename>")        # restore R object / dataset
```

```
the.beatles.songs <- readRDS("The Beatles songs dataset, v2.2.RData")
summary(the.beatles.songs)
```

```
##     Title                Year       Duration     Other.releases
##  Length:310         1963   :66   Min.   : 23.0   Min.   : 0.00
##  Class :character   1968   :45   1st Qu.:133.0   1st Qu.: 0.00
##  Mode  :character   1969   :43   Median :150.0   Median : 9.00
##                     1964   :41   Mean   :159.6   Mean   :10.42
##                     1965   :37   3rd Qu.:172.8   3rd Qu.:16.00
##                     1967   :27   Max.   :502.0   Max.   :56.00
##                     (Other):51
##             Single.certification   Covered.by     Top.50.Billboard
##  No                    :259       Min.   : 0.000   Min.   : 0.000
##  RIAA 2xPlatinum       :  6       1st Qu.: 0.000   1st Qu.: 0.000
##  RIAA 4xPlatinum       :  2       Median : 2.000   Median : 0.000
##  RIAA Gold             : 33       Mean   : 6.752   Mean   : 4.061
##  RIAA Gold, BPI Silver:  2       3rd Qu.: 8.000   3rd Qu.: 0.000
##  RIAA Platinum         :  8       Max.   :70.000   Max.   :50.000
##
##  Top.50
##  No :261
##  Yes: 49
##
##
##
##
##
```

Examining the distribution of the output values

```
table(<dataset>$<output variable>)
prop.table(table(<dataset>$<output variable>))
round(prop.table(table(<dataset>$<output variable>)), digits = 2)
```

```
table(the.beatles.songs$Top.50)
```

```
##
##  No Yes
## 261  49
```

```
prop.table(table(the.beatles.songs$Top.50))
```

```
##
##        No       Yes
## 0.8419355 0.1580645
```

```
round(prop.table(table(the.beatles.songs$Top.50)), digits = 2)
```

```
##
##   No  Yes
## 0.84 0.16
```

Train and test datasets

```
install.packages("caret")
library(caret)
set.seed(<n>)
<train dataset indices> <-                          # stratified partitioning:
+ createDataPartition(<dataset>$<output variable>,  # the same distribution of the output variable in both sets
+                     p = .80,                      # 80/20% of data in train/test sets
+                     list = FALSE)                 # don't make a list of results, make a matrix
<train dataset> <- <dataset>[<train dataset indices>, ]
<test dataset>  <- <dataset>[-<train dataset indices>, ]
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.4.2
```

```
set.seed(333)
train.data.indices <- createDataPartition(the.beatles.songs$Top.50, p = 0.80, list = FALSE)
train.data <- the.beatles.songs[train.data.indices, ]
test.data <- the.beatles.songs[-train.data.indices, ]
```

Building the model / decision tree

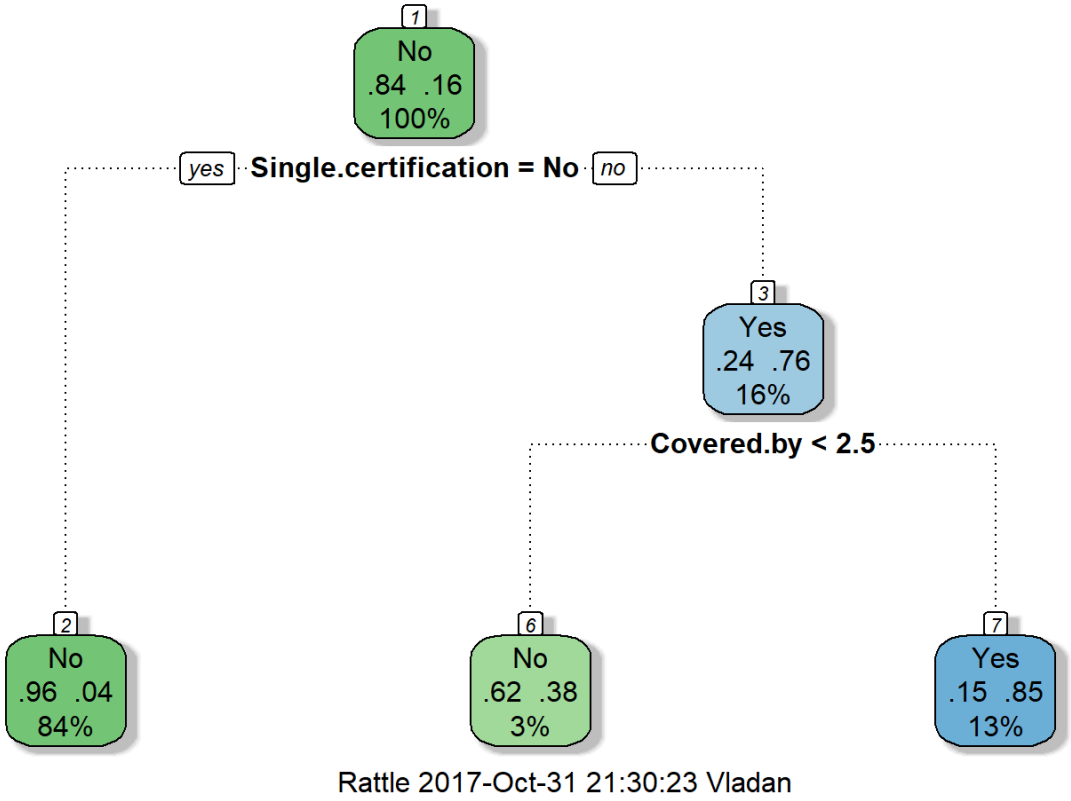
```
# install.packages("rpart")
library(rpart)
<decision tree> <- rpart(<output variable> ~                                    # build the tree
+                        <predictor variable 1> + <predictor variable 2> + ..., # . to include all variables
+                        data = <train dataset>,
+                        method = "class")                                      # build classification tree
```

```
library(rpart)
top.50.tree <- rpart(Top.50 ~ Single.certification + Covered.by + Year,
                     data = train.data,
                     method = "class")
```

Depicting the model

```
# install.packages('rattle')
# install.packages('rpart.plot')
# install.packages('RColorBrewer')
library(rpart)
library(rattle)
library(rpart.plot)
library(RColorBrewer)
fancyRpartPlot(<decision tree>)
```

```
library(rpart)
library(rattle)
library(rpart.plot)
library(RColorBrewer)
fancyRpartPlot(top.50.tree)
```



Making predictions

```
<predictions> <- predict(object = <decision tree>,
+                        newdata = <test dataset>,
+                        type = "class")
<predictions>[<i1>:<ik>]                          # examine some of the predictions
<predictions dataframe> <-
+ data.frame(<observation ID> = <test dataset>$<observation ID column>,
+            <another relevant feature> = <test dataset>$<another relevant feature column>,
+            ...,
+            <predictions feature> = <predictions>)
View(<predictions dataframe>)
```

```
top.50.predictions <- predict(top.50.tree, newdata = test.data, type = "class")
top.50.predictions[1:20]
```

```
##   1   6  22  26  29  32  35  37  44  45  47  50  52  54  65  70  71  72
##  No  No Yes  No  No  No Yes  No  No  No  No Yes  No  No  No  No  No  No
##  76 101
##  No  No
## Levels: No Yes
```

```
top.50.predictions.dataframe <- data.frame(Song = test.data$Title,
                                           Top.50.Billboard = test.data$Top.50.Billboard,
                                           Top.50 = test.data$Top.50,
                                           Prediction = top.50.predictions)
```

```
View(top.50.predictions.dataframe)
```

Clustering - K-Means

Reading the dataset

The dataset has been prepared earlier and saved using:

```
saveRDS(the.beatles.songs, "The Beatles songs dataset, v2.3.RData")
```

Read the dataset using:

```
<dataframe or another R object> <- readRDS(file = "<filename>")         # restore R object / dataset
```

```
the.beatles.songs <- readRDS("The Beatles songs dataset, v2.3.RData")
summary(the.beatles.songs)
```

```
##     Title              Duration     Other.releases    Covered.by
##  Length:310         Min.   : 23.0   Min.   : 0.00   Min.   : 0.000
##  Class :character   1st Qu.:133.0   1st Qu.: 0.00   1st Qu.: 0.000
##  Mode  :character   Median :150.0   Median : 9.00   Median : 2.000
##                     Mean   :159.6   Mean   :10.42   Mean   : 6.752
##                     3rd Qu.:172.8   3rd Qu.:16.00   3rd Qu.: 8.000
##                     Max.   :502.0   Max.   :56.00   Max.   :70.000
##  Top.50.Billboard
##  Min.   : 0.000
##  1st Qu.: 0.000
##  Median : 0.000
##  Mean   : 4.061
##  3rd Qu.: 0.000
##  Max.   :50.000
```

Changing the row names

Optional, run in order to focus on numeric variables in the dataset only (for applying K-Means):

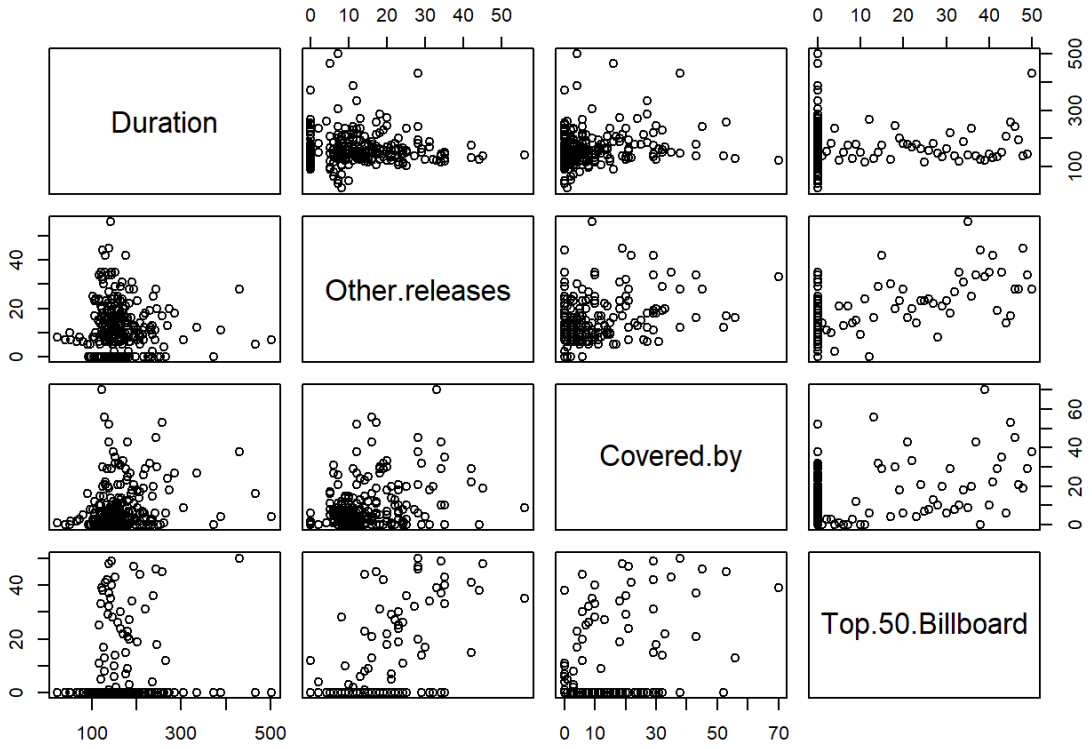
```
rownames(the.beatles.songs) <- the.beatles.songs$Title
the.beatles.songs$Title <- NULL                                    # no longer necessary
```

Examining the data

See if there are some patterns in the data, pairwise, to possibly indicate clusters:

```
pairs(~ <column 1 name> + <column 2 name> + ..., data = <dataframe>)
```

```
pairs(~ Duration + Other.releases + Covered.by + Top.50.Billboard,  # no any striking pattern, i.e.
      the.beatles.songs)                                            # no visual indication of clusters
```

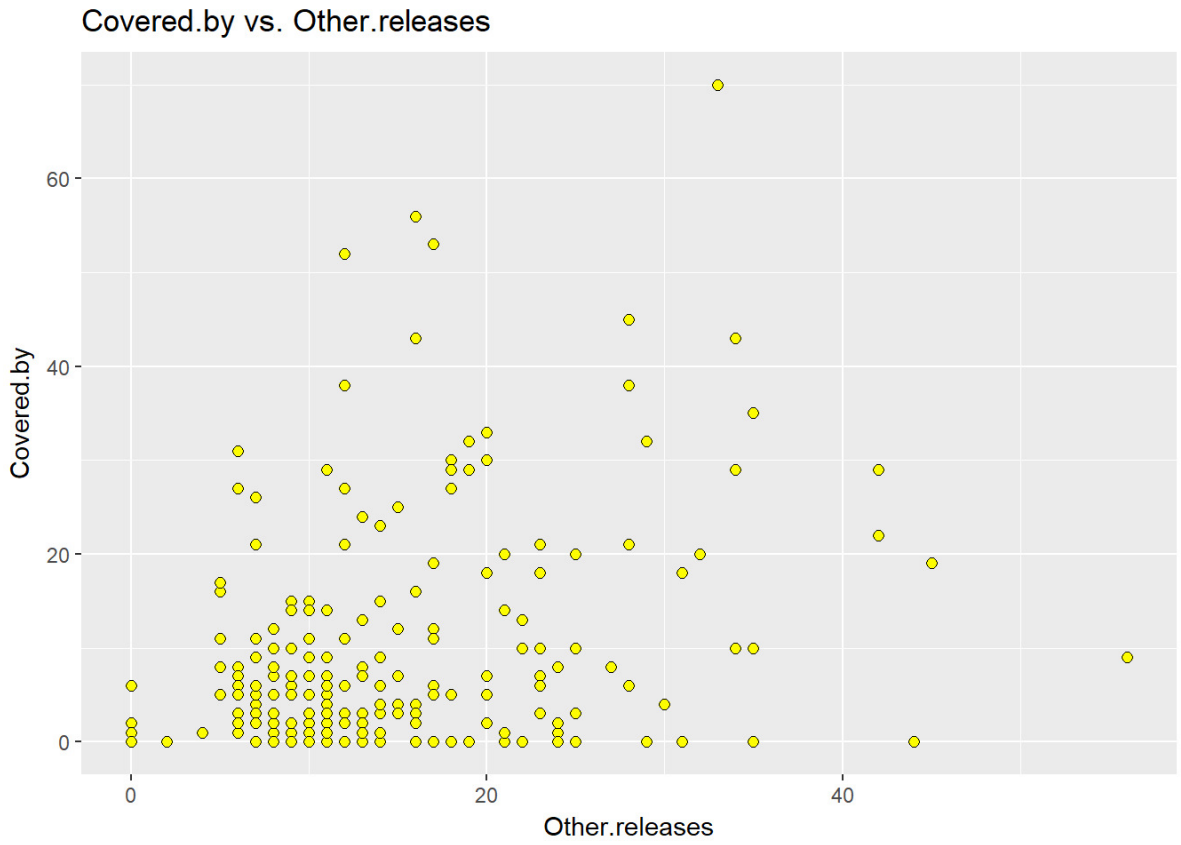


K-Means with 2 variables

Plot the data first:

```
<scatterplot> <-
+ ggplot(<dataset>, aes(x = <num.var.1>, y = <num.var.2>)) +
+   geom_point(shape = <n>,         # <n> = 1: hollow circle, no fill; <n> = 21: circle that can be filled
+              fill = <color 1>,    # color of point fill (optional)
+              color = <color 2>,   # color of point line (optional)
+              size = <s>)          # size  of point line (optional)
<scatterplot> <- <scatterplot> + xlab("<x label>")                    # label/caption on x-axis
<scatterplot> <- <scatterplot> + ylab("<y label>")                    # label/caption on x-axis
<scatterplot> <- <scatterplot> + ggtitle("<scatterplot title>")       # scatterplot title
```

```
scatterplot1 <- ggplot(the.beatles.songs, aes(x = Other.releases, y = Covered.by))
scatterplot1 <- scatterplot1 + geom_point(shape = 21, fill = "yellow", size = 2)
scatterplot1 <- scatterplot1 + xlab("Other.releases")
scatterplot1 <- scatterplot1 + ylab("Covered.by")
scatterplot1 <- scatterplot1 + ggtitle("Covered.by vs. Other.releases")
scatterplot1
```



Subset the original data to include only the variables to be used in K-Means:

```
<new dataframe> <- subset(<dataframe>[, c("<col1 name>", "<col2 name>")])
<new dataframe> <- subset(<dataframe>[, <col1 index>:<col2 index>)])
```

```
the.beatles.songs.2 <- subset(the.beatles.songs[, c("Other.releases", "Covered.by")])
summary(the.beatles.songs.2)
```

```
##  Other.releases    Covered.by
##  Min.   : 0.00   Min.   : 0.000
##  1st Qu.: 0.00   1st Qu.: 0.000
##  Median : 9.00   Median : 2.000
##  Mean   :10.42   Mean   : 6.752
##  3rd Qu.:16.00   3rd Qu.: 8.000
##  Max.   :56.00   Max.   :70.000
```

```
head(the.beatles.songs.2)
```

```
##                           Other.releases Covered.by
## 12-Bar Original                        0          0
## A Day in the Life                     12         27
## A Hard Day's Night                    35         35
## A Shot of Rhythm and Blues             0          0
## A Taste of Honey                      29          0
## Across the Universe                   19         32
```

Data normalization, required by K-Means when the variables have different ranges: `range(<dataframe>$<variable>)`

```
# install.packages("clusterSim")
library(clusterSim)
<dataframe with numeric columns> <-                    # works with vectors and matrices as well
```

```
+ data.Normalization(<dataframe with numeric columns>,
+                    type = "n4",                        # normalization: (x - min(x)) / (max(x) - min(x))
+                    normalization = "column")           # normalization by columns
```

```
range(the.beatles.songs.2$Other.releases)
```

```
## [1]  0 56
```

```
range(the.beatles.songs.2$Covered.by)
```

```
## [1]  0 70
```

```
library(clusterSim)
```

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

```
## Loading required package: MASS
```

```
##
## This is package 'modeest' written by P. PONCET.
## For a complete list of functions, use 'library(help = "modeest")' or 'help.start()'.
```

```
the.beatles.songs.2 <- data.Normalization(the.beatles.songs.2, type = "n4", normalization = "column")
tail(the.beatles.songs.2)
```

```
##                                 Other.releases Covered.by
## You'll Be Mine                      0.00000000 0.00000000
## You're Going to Lose That Girl      0.10714286 0.02857143
## You've Got to Hide Your Love Away   0.21428571 0.30000000
## You've Really Got a Hold on Me      0.03571429 0.00000000
## Young Blood                         0.00000000 0.00000000
## Your Mother Should Know             0.23214286 0.01428571
```

Run K-Means for K = 3:

```
set.seed(<seed>)
<clusters> <- kmeans(x = <normalized dataframe>,
+                    centers = 3,                        # K = 3
+                    iter.max = 20,                      # max number of iterations allowed
+                    nstart = 1000)                      # no. of initial configurations (report on the best one)
<clusters>
```

```
set.seed(888)
clusters.K3 <- kmeans(x = the.beatles.songs.2, centers = 3, iter.max = 20, nstart = 1000)
clusters.K3
```

```
## K-means clustering with 3 clusters of sizes 77, 192, 41
##
## Cluster means:
##   Other.releases Covered.by
## 1     0.35853432 0.06586271
## 2     0.07393973 0.03601190
## 3     0.38763066 0.43693380
##
## Clustering vector:
##                                          12-Bar Original
##                                                        2
##                                       A Day in the Life
##                                                        3
##                                     A Hard Day's Night
##                                                        3
##                                  A Shot of Rhythm and Blues
##                                                        2
##                                        A Taste of Honey
##                                                        1
##                                     Across the Universe
##                                                        3
##                                           Act Naturally
##                                                        1
##                                        Ain't She Sweet
##                                                        2
##                                     All I've Got to Do
##                                                        2
##                                           All My Loving
##                                                        3
##                                     All Things Must Pass
##                                                        2
##                                        All Together Now
##                                                        2
##                                     All You Need Is Love
##                                                        3
##                                           And I Love Her
##                                                        3
##                                  And Your Bird Can Sing
##                                                        2
##                                        Anna (Go to Him)
##                                                        1
##                                           Another Girl
##                                                        2
##                                        Any Time at All
##                                                        1
##                                             Ask Me Why
##                                                        1
##                                          Baby It's You
##                                                        1
##                                        Baby's in Black
##                                                        1
##                                  Baby, You're a Rich Man
##                                                        2
##                                     Back in the U.S.S.R.
##                                                        1
##                                                 Bad Boy
##                                                        2
##                                               Bad to Me
##                                                        2
##                                        Beautiful Dreamer
```

```
##                                                 2
##                            Because I Know You Love Me So
##                                                 2
##                                           Because
##                                                 2
##                         Being for the Benefit of Mr. Kite!
##                                                 2
##                                          Birthday
##                                                 2
##                                         Blackbird
##                                                 3
##                                       Blue Jay Way
##                                                 1
##                                             Boys
##                                                 1
##                                       BÃ©same Mucho
##                                                 2
##                                    Can't Buy Me Love
##                                                 3
##                                             Carol
##                                                 2
##                                    Carry That Weight
##                                                 2
##                                          Catswalk
##                                                 2
##                                           Cayenne
##                                                 2
##                                            Chains
##                                                 1
##                                      Child of Nature
##                                                 2
##                         Christmas Time (Is Here Again)
##                                                 2
##                                           Circles
##                                                 2
##                                        Clarabella
##                                                 2
##                                      Come and Get It
##                                                 2
##                                     Come Together
##                                                 3
##                                        Cry Baby Cry
##                                                 2
##                                    Cry for a Shadow
##                                                 2
##                               Crying, Waiting, Hoping
##                                                 2
##                                         Day Tripper
##                                                 3
##                                      Dear Prudence
##                                                 2
##                                   Devil in Her Heart
##                                                 2
##                                          Dig a Pony
##                                                 2
##                                             Dig It
##                                                 2
##                                      Dizzy, Miss Lizzy
##                                                 1
##                               Do You Want to Know a Secret?
##                                                 1
```

```
##                                          Doctor Robert
##                                                     2
##                                        Don't Bother Me
##                                                     2
##                                       Don't Ever Change
##                                                     2
##                                       Don't Let Me Down
##                                                     3
##                                        Don't Pass Me By
##                                                     2
##                                            Drive My Car
##                                                     1
##                                       Eight Days a Week
##                                                     1
##                                           Eleanor Rigby
##                                                     3
##                                                Etcetera
##                                                     2
##                                     Every Little Thing
##                                                     2
## Everybody's Got Something to Hide Except Me and My Monkey
##                                                     2
##                             Everybody's Trying to Be My Baby
##                                                     1
##                                    Fancy My Chances with You
##                                                     2
##                                           Fixing a Hole
##                                                     2
##                                                 Flying
##                                                     2
##                                              For No One
##                                                     2
##                                            For You Blue
##                                                     1
##                                         Free as a Bird
##                                                     2
##                                         From Me to You
##                                                     1
##                                         From Us to You
##                                                     2
##                                               Get Back
##                                                     3
##                                         Getting Better
##                                                     2
##                                                   Girl
##                                                     1
##                                          Glad All Over
##                                                     2
##                                            Glass Onion
##                                                     2
##                                       Golden Slumbers
##                                                     2
##                                     Good Day Sunshine
##                                                     2
##                                  Good Morning, Good Morning
##                                                     2
##                                                Good Night
##                                                     2
##                                                 Goodbye
##                                                     2
##                                  Got to Get You into My Life
```

```
##                                          2
##                        Hallelujah, I Love Her So
##                                          2
##                          Happiness Is a Warm Gun
##                                          2
##                                       Heather
##                                          2
##                                Hello Little Girl
##                                          2
##                                   Hello, Goodbye
##                                          1
##                                         Help!
##                                          3
##                                   Helter Skelter
##                                          3
##                                     Her Majesty
##                                          2
##                                Here Comes the Sun
##                                          3
##                           Here, There and Everywhere
##                                          3
##                                     Hey Bulldog
##                                          1
##                                        Hey Jude
##                                          3
##                                Hippy Hippy Shake
##                                          2
##                                    Hold Me Tight
##                                          1
##                                     Honey Don't
##                                          1
##                                       Honey Pie
##                                          2
##                                How Do You Do It?
##                                          2
##                                  I Am the Walrus
##                                          3
##                                 I Call Your Name
##                                          1
##                        I Don't Want to Spoil the Party
##                                          2
##                                   I Feel Fine
##                                          1
##                         I Forgot to Remember to Forget
##                                          2
##                                  I Got a Woman
##                                          2
##                             I Got to Find My Baby
##                                          2
##                            I Just Don't Understand
##                                          2
##                              I Lost My Little Girl
##                                          2
##                                       I Me Mine
##                                          2
##                                     I Need You
##                                          2
##                             I Saw Her Standing There
##                                          3
##                           I Should Have Known Better
##                                          1
```

```
##                                  I Wanna Be Your Man
##                                                    1
##                               I Want to Hold Your Hand
##                                                    3
##                                    I Want to Tell You
##                                                    2
##                           I Want You (She's So Heavy)
##                                                    2
##                                                I Will
##                                                    2
##                                           I'll Be Back
##                                                    1
##                                      I'll Be on My Way
##                                                    2
##                                       I'll Cry Instead
##                                                    1
##                                    I'll Follow the Sun
##                                                    1
##                                           I'll Get You
##                                                    1
##                                I'll Keep You Satisfied
##                                                    2
##                                         I'm a Loser
##                                                    1
##                                            I'm Down
##                                                    1
##              I'm Gonna Sit Right Down and Cry (Over You)
##                                                    2
##                          I'm Happy Just to Dance with You
##                                                    2
##                                         I'm In Love
##                                                    2
##                                I'm Looking Through You
##                                                    1
##                                      I'm Only Sleeping
##                                                    2
##                                         I'm So Tired
##                                                    2
##                        I'm Talking About You (Star Club)
##                                                    2
##                              I'm Talking About You (BBC)
##                                                    2
##                                       I've Got a Feeling
##                                                    1
##                                    I've Just Seen a Face
##                                                    2
##                                                If I Fell
##                                                    1
##                                    If I Needed Someone
##                                                    1
##                                    If You've Got Trouble
##                                                    2
##                                                In My Life
##                                                    3
##                               In Spite of All the Danger
##                                                    2
##                                         It Won't Be Long
##                                                    1
##                                      It's All Too Much
##                                                    2
##                                        It's Only Love
```

```
##                                                1
##                                  Jazz Piano Song
##                                                2
##                                   Jessie's Dream
##                                                2
##                                  Johnny B. Goode
##                                                2
##                                            Julia
##                                                2
##                                             Junk
##                                                2
##                        Kansas City/Hey, Hey, Hey, Hey
##                                                2
##                           Keep Your Hands Off My Baby
##                                                2
##                               Komm Gib Mir Deine Hand
##                                                2
##                                      Lady Madonna
##                                                3
##                                 Leave My Kitten Alone
##                                                2
##                                     Lend Me Your Comb
##                                                2
##                                         Let It Be
##                                                3
##                                  Like Dreamers Do
##                                                2
##                                      Little Child
##                                                2
##                             Lonesome Tears in My Eyes
##                                                2
##                                   Long Tall Sally
##                                                1
##                                  Long, Long, Long
##                                                2
##                                     Looking Glass
##                                                2
##                                       Love Me Do
##                                                1
##                                  Love of the Loved
##                                                2
##                                       Love You To
##                                                2
##                                       Lovely Rita
##                                                2
##                                          Lucille
##                                                2
##                          Lucy in the Sky with Diamonds
##                                                2
##                                           Madman
##                                                2
##                                        Maggie Mae
##                                                2
##                              Magical Mystery Tour
##                                                1
##                         Mailman, Bring Me No More Blues
##                                                2
##                                   Martha My Dear
##                                                2
##                                         Matchbox
##                                                1
```

```
##                                 Maxwell's Silver Hammer
##                                                       2
##                                        Mean Mr. Mustard
##                                                       2
##                                     Memphis, Tennessee
##                                                       2
##                                               Michelle
##                                                       3
##                                                 Misery
##                                                       1
##                               Money (That's What I Want)
##                                                       1
##                                            Moonlight Bay
##                                                       2
##                                     Mother Nature's Son
##                                                       2
##                                            Mr. Moonlight
##                                                       1
##                                                My Bonnie
##                                                       1
##                                               No Reply
##                                                       1
##                       Norwegian Wood (This Bird Has Flown)
##                                                       3
##                                      Not a Second Time
##                                                       2
##                                               Not Guilty
##                                                       2
##            Nothin' Shakin' (But the Leaves on the Trees)
##                                                       2
##                                               Nowhere Man
##                                                       3
##                                    Ob-La-Di, Ob-La-Da
##                                                       1
##                                      Octopus's Garden
##                                                       2
##                                            Oh! Darling
##                                                       2
##                                        Old Brown Shoe
##                                                       1
##                                        One After 909
##                                                       1
##                                    One and One Is Two
##                                                       2
##                                   Only a Northern Song
##                                                       2
##                                             Ooh! My Soul
##                                                       2
##                                       P.S. I Love You
##                                                       1
##                                      Paperback Writer
##                                                       1
##                                            Penny Lane
##                                                       1
##                                                Piggies
##                                                       2
##                                    Please Mr. Postman
##                                                       1
##                                      Please Please Me
##                                                       1
##                                         Polythene Pam
```

```
##                                                 2
##                                              Rain
##                                                 1
##                                         Real Love
##                                                 2
##                                       Revolution 1
##                                                 2
##                                       Revolution 9
##                                                 2
##                                        Revolution
##                                                 3
##  Rip It Up/Shake, Rattle, and Roll/Blue Suede Shoes
##                                                 2
##                                 Rock and Roll Music
##                                                 1
##                                       Rocky Raccoon
##                                                 2
##                                 Roll Over Beethoven
##                                                 1
##                                   Run for Your Life
##                                                 2
##                                       Savoy Truffle
##                                                 2
##                                          Searchin'
##                                                 2
##                                September in the Rain
##                                                 2
##                                         Sexy Sadie
##                                                 2
##     Sgt. Pepper's Lonely Hearts Club Band (Reprise)
##                                                 2
##               Sgt. Pepper's Lonely Hearts Club Band
##                                                 1
##                                Shakin' in the Sixties
##                                                 2
##             She Came in Through the Bathroom Window
##                                                 2
##                                      She Loves You
##                                                 3
##                                   She Said She Said
##                                                 2
##                                      She's a Woman
##                                                 1
##                                  She's Leaving Home
##                                                 3
##                                              Shout
##                                                 2
##                                     Sie Liebt Dich
##                                                 2
##                                          Slow Down
##                                                 1
##                     So How Come (No One Loves Me)
##                                                 2
##                Soldier of Love (Lay Down Your Arms)
##                                                 2
##                                     Some Other Guy
##                                                 2
##                                          Something
##                                                 3
##                                    Sour Milk Sea
##                                                 2
```

```
##                          Step Inside Love/Los Paranoias
##                                                       2
##                               Strawberry Fields Forever
##                                                       3
##                                                Sun King
##                                                       2
##                          Sure to Fall (In Love with You)
##                                                       2
##                                     Sweet Little Sixteen
##                                                       2
##                               Take Good Care of My Baby
##                                                       2
##                               Taking a Trip to Carolina
##                                                       2
##                                                  Taxman
##                                                       2
##                                               Teddy Boy
##                                                       2
##                                    Tell Me What You See
##                                                       2
##                                             Tell Me Why
##                                                       1
##                                          Thank You Girl
##                                                       1
##                                      That Means a Lot
##                                                       2
##                                    That'll Be the Day
##                                                       2
##                                 That's All Right (Mama)
##                                                       2
##                           The Ballad of John and Yoko
##                                                       1
##                 The Continuing Story of Bungalow Bill
##                                                       2
##                                               The End
##                                                       2
##                                    The Fool on the Hill
##                                                       3
##                                  The Honeymoon Song
##                                                       2
##                                     The Inner Light
##                                                       2
##                               The Long and Winding Road
##                                                       3
##                                       The Night Before
##                                                       1
##                                             The Saints
##                                                       2
##                                    The Sheik of Araby
##                                                       2
##                                               The Word
##                                                       2
##                                     There's a Place
##                                                       1
##                                    Things We Said Today
##                                                       1
##                                    Think for Yourself
##                                                       2
##                                               This Boy
##                                                       1
##                                     Three Cool Cats
```

```
##                                                 2
##                                  Ticket to Ride
##                                                 3
##                               Till There Was You
##                                                 1
##                                Tip of My Tongue
##                                                 2
##                      To Know Her is to Love Her
##                                                 2
##                           Tomorrow Never Knows
##                                                 3
##                    Too Much Monkey Business
##                                                 2
##                              Twist and Shout
##                                                 1
##                                    Two of Us
##                                                 2
##                                         Wait
##                                                 2
##                               Watching Rainbows
##                                                 2
##                               We Can Work It Out
##                                                 3
##                                    What Goes On
##                                                 2
##                               What You're Doing
##                                                 2
##                    What's The New Mary Jane
##                                                 2
##                                When I Get Home
##                                                 1
##                               When I'm Sixty-Four
##                                                 2
##                  While My Guitar Gently Weeps
##                                                 3
##                 Why Don't We Do It in the Road?
##                                                 2
##                                 Wild Honey Pie
##                                                 2
##                                 Winston's Walk
##                                                 2
##           With a Little Help from My Friends
##                                                 3
##                        Within You Without You
##                                                 2
##                                       Woman
##                                                 2
##                                  Words of Love
##                                                 1
##                                Yellow Submarine
##                                                 1
##                                      Yer Blues
##                                                 2
##                                      Yes It Is
##                                                 1
##                                      Yesterday
##                                                 3
##                                You Can't Do That
##                                                 1
##           You Know My Name (Look Up the Number)
##                                                 2
```

```
##                                     You Know What to Do
##                                                       2
##                                   You Like Me Too Much
##                                                       2
##                             You Never Give Me Your Money
##                                                       2
##                                        You Won't See Me
##                                                       2
##                                          You'll Be Mine
##                                                       2
##                           You're Going to Lose That Girl
##                                                       2
##                         You've Got to Hide Your Love Away
##                                                       3
##                           You've Really Got a Hold on Me
##                                                       2
##                                               Young Blood
##                                                       2
##                                    Your Mother Should Know
##                                                       1
##
## Within cluster sum of squares by cluster:
## [1] 1.772593 1.830571 2.360885
##  (between_SS / total_SS =  66.6 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

Add the vector of clusters to the dataframe:

```
<normalized dataframe>$<new column> <- factor(<clusters>$cluster)   # <clusters>: from the previous step
head(<normalized dataframe>)
```

```
the.beatles.songs.2$Cluster <- factor(clusters.K3$cluster)
head(the.beatles.songs.2)
```

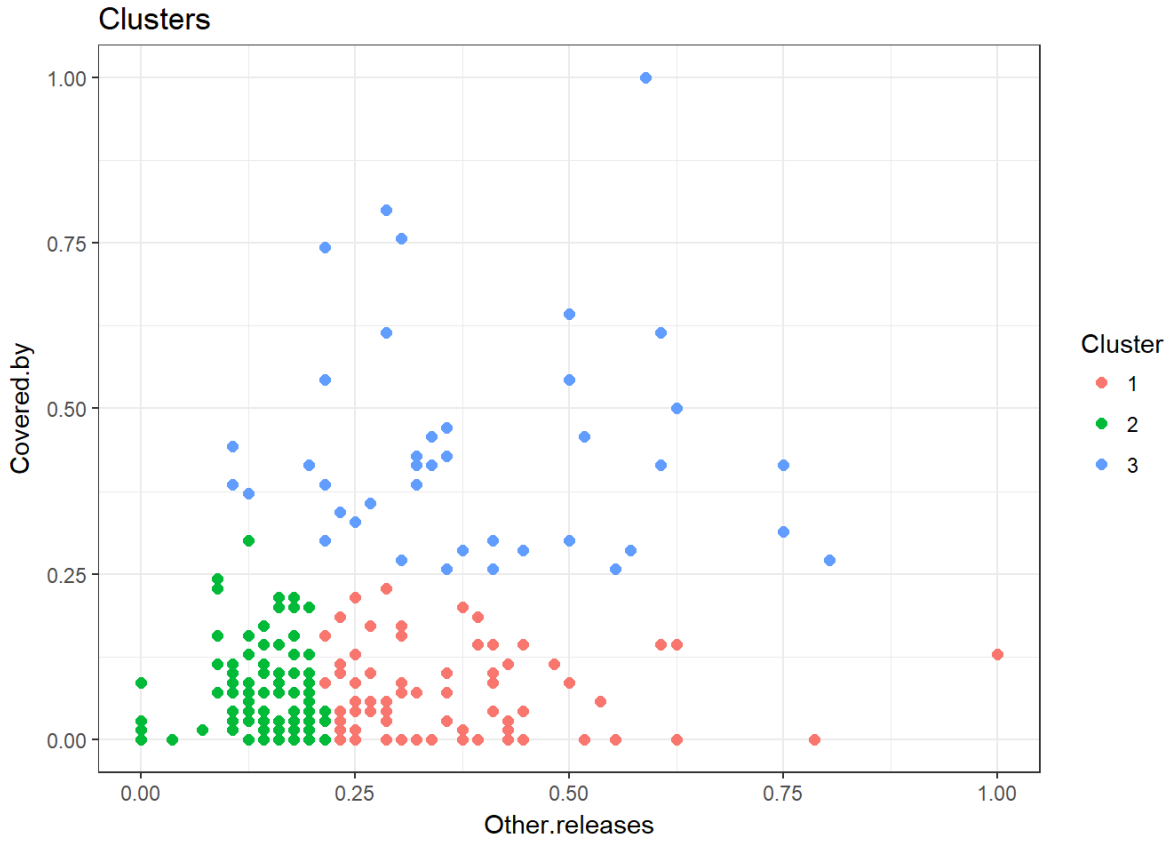
```
##                            Other.releases Covered.by Cluster
## 12-Bar Original                 0.0000000  0.0000000       2
## A Day in the Life               0.2142857  0.3857143       3
## A Hard Day's Night              0.6250000  0.5000000       3
## A Shot of Rhythm and Blues      0.0000000  0.0000000       2
## A Taste of Honey                0.5178571  0.0000000       1
## Across the Universe             0.3392857  0.4571429       3
```

Plot the clusters in a new scatterplot:

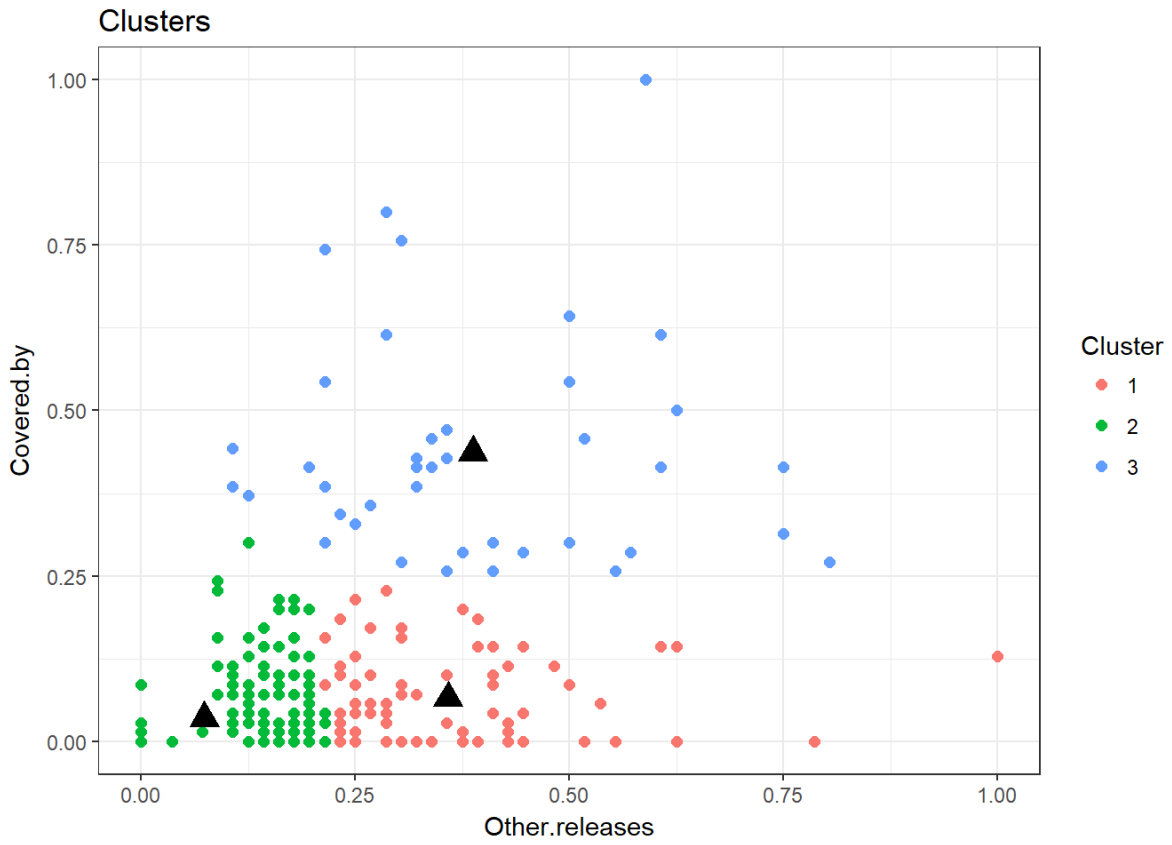
```
<scatterplot> <-
+ ggplot(<dataset with the cluster column>,
+        aes(x = <num.var.1>, y = <num.var.2>,
+            color = <cluster column>)) +       # color clusters differently
<scatterplot> <- <scatterplot> + geom_point()   # fill colors can be added subsequently, see below
<scatterplot> <- <scatterplot> + xlab("<x label>")                 # label/caption on x-axis
<scatterplot> <- <scatterplot> + ylab("<y label>")                 # label/caption on x-axis
<scatterplot> <- <scatterplot> + ggtitle("<scatterplot title>")    # scatterplot title
<scatterplot> <- <scatterplot> +
+ scale_fill_brewer(palette = "Set1",           # palettes:  http://ggplot2.tidyverse.org/reference/scale_brewer.html
+                   name = "<cluster column>")  # legend title
```

```
<scatterplot> <- <scatterplot> + theme_bw()     # white background
<scatterplot> <- <scatterplot> +                # add cluster centers
+ geom_point(data =                            # "data = " MUST be here, otherwise it doesn't work!
+            as.data.frame(<clusters>$centers), # <clusters>: from the previous step
+            color = "<line color>",
+            fill = "<fill color>",
+            size = <size>,                     # frequently used <size>s: 3, 4
+            shape = <shape>)                   # diamond: 23; triangle: 24; circle: 21; ...
```

```
scatterplot2 <- ggplot(the.beatles.songs.2,
                       aes(x = Other.releases, y = Covered.by,
                           colour = Cluster))
scatterplot2 <- scatterplot2 + geom_point(size = 2)
scatterplot2 <- scatterplot2 + xlab("Other.releases")
scatterplot2 <- scatterplot2 + ylab("Covered.by")
scatterplot2 <- scatterplot2 + ggtitle("Clusters")
scatterplot2 <- scatterplot2 +
  scale_fill_brewer(palette = "Set1", name = "Cluster")
scatterplot2 <- scatterplot2 + theme_bw()
scatterplot2
```




```
scatterplot2 + geom_point(data = as.data.frame(clusters.K3$centers),  # add cluster centers
             color = "black",
             fill = "black",
             size = 4,
             shape = 24)
```



Resources, readings, references

The corrplot package: https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html (https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html)
Understanding Diagnostic Plots for Linear Regression Analysis: http://data.library.virginia.edu/diagnostic-plots/ (http://data.library.virginia.edu/diagnostic-plots/)