ECON 104 Group Project #1

Yüksel Polat Akbıyk, Rebecca Zhu, Tori Takeshita, Crystal Huynh

15/04/2021

Question 1

Provide a descriptive analysis of your variables. This should include histograms and fitted distributions, correlation plot, boxplots, scatterplots, and statistical summaries (e.g., the five-number summary). All figures must include comments.

Danceability: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable. e are predicting the dancebility of a song given the following variables:

Energy: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy. (Float)

Key: The key the track is in. Integers map to pitches using standard Pitch Class notation . E.g. 0 = C, 1 = C sharp, 2 = D, and so on. (Integer)

Loudness: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db. (Float)

Mode: Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0. (Integer)

Speechiness: Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks. (Float)

Acousticness: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic. (Float)

Instrumentalness: Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0. (Float)

Liveness: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live. (Float)

Valence: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry). (Float)

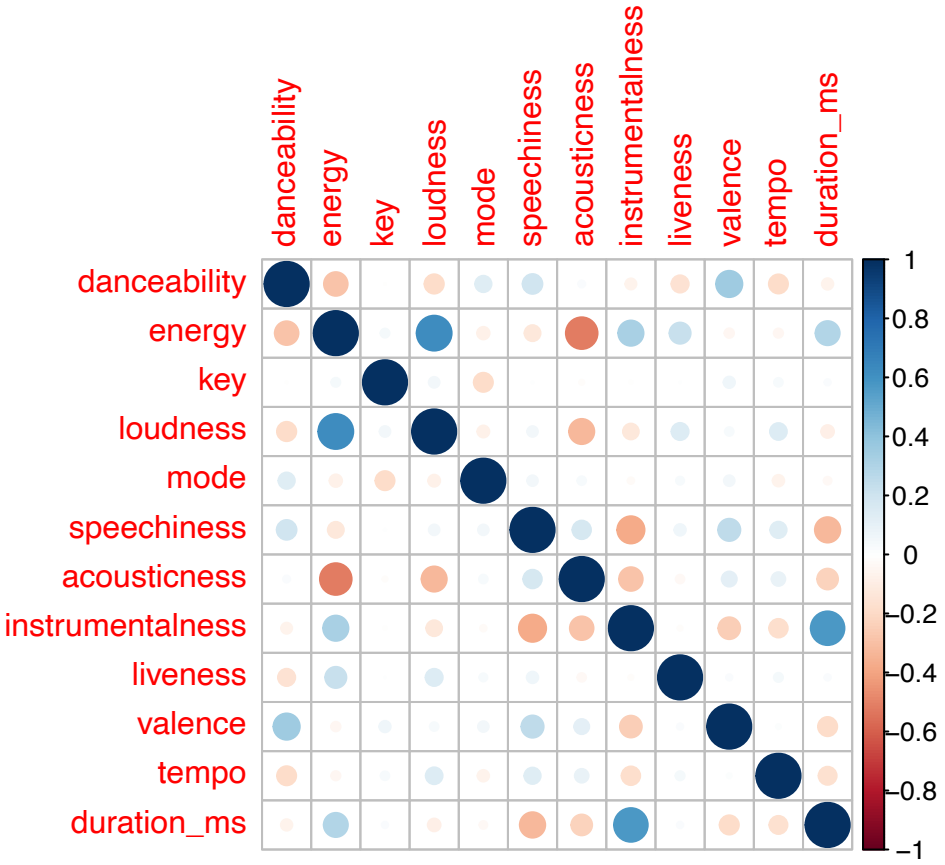
Tempo: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration. (Float)

Duration: The duration of the track in milliseconds. (Integer)

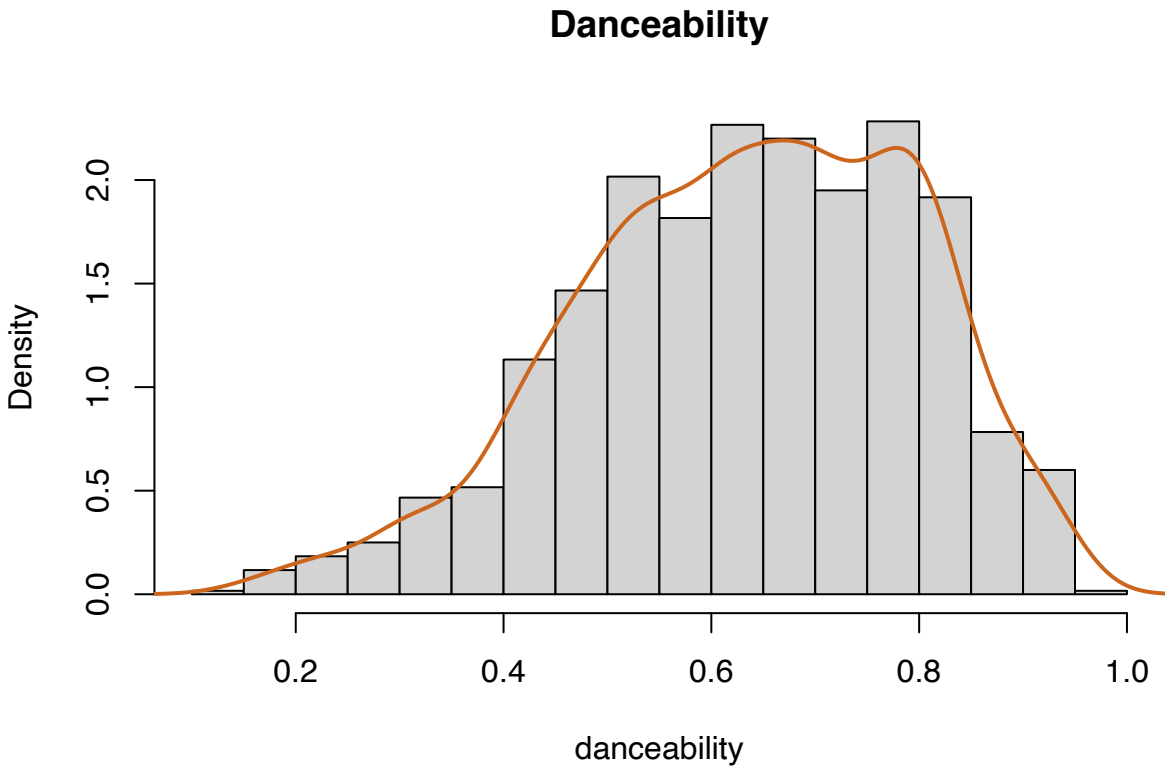
```
library(broom)
genres <- read.csv(file = 'genres_v2.csv')
# take a random sample of size 1200 from a dataset gernes without replacement
set.seed(1)
genres2 <- genres[sample(1:nrow(genres), 1200, replace=FALSE),]
attach(genres2)
library(corrplot)

library(dplyr)


#load the data
genres <- read.csv(file = 'genres_v2.csv')
names <- names(genres)
names
```

```
##  [1] "danceability"     "energy"           "key"              "loudness"
##  [5] "mode"             "speechiness"      "acousticness"     "instrumentalness"
##  [9] "liveness"         "valence"          "tempo"            "type"
## [13] "id"               "uri"              "track_href"       "analysis_url"
## [17] "duration_ms"      "time_signature"   "genre"            "song_name"
## [21] "Unnamed..0"       "title"
```

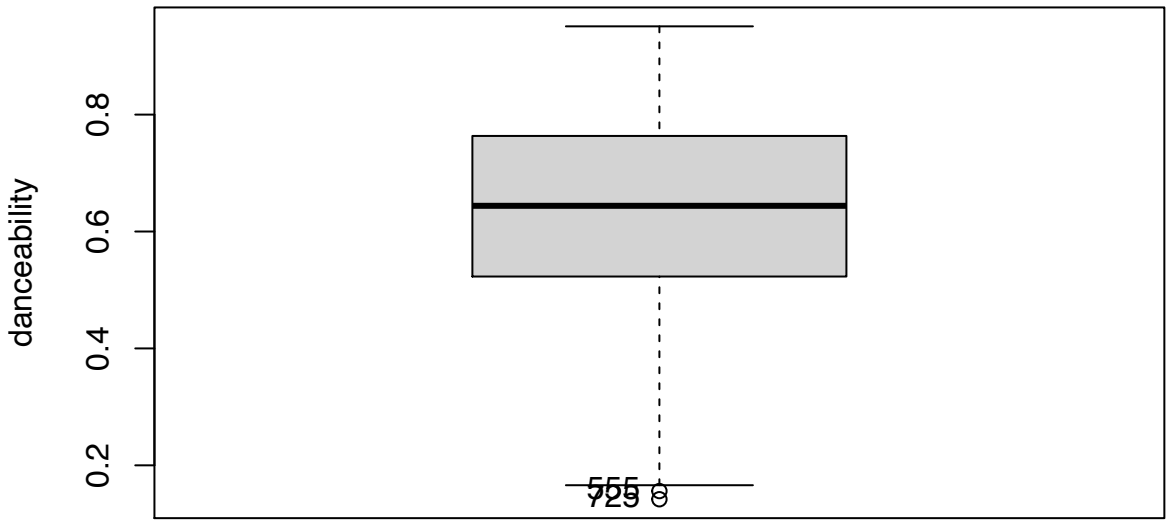
```
# Histogram *
# Fitted Distribution
# Correlation Plot *
# Box Plot *
# Scatter Plot *
# Statistical Summary



# Corrplot
corrplot(cor(genres2[, c("danceability", "energy", "key", "loudness", "mode",
                         "speechiness", "acousticness", "instrumentalness",
                         "liveness", "valence", "tempo", "duration_ms")] ))
```




```
name <-"Danceability"
dance <- "Dancebility"
hist(danceability, main=name, prob=TRUE)
lines(density(danceability), lwd = 2, col = "chocolate3")
```




```
Boxplot(~danceability)
```

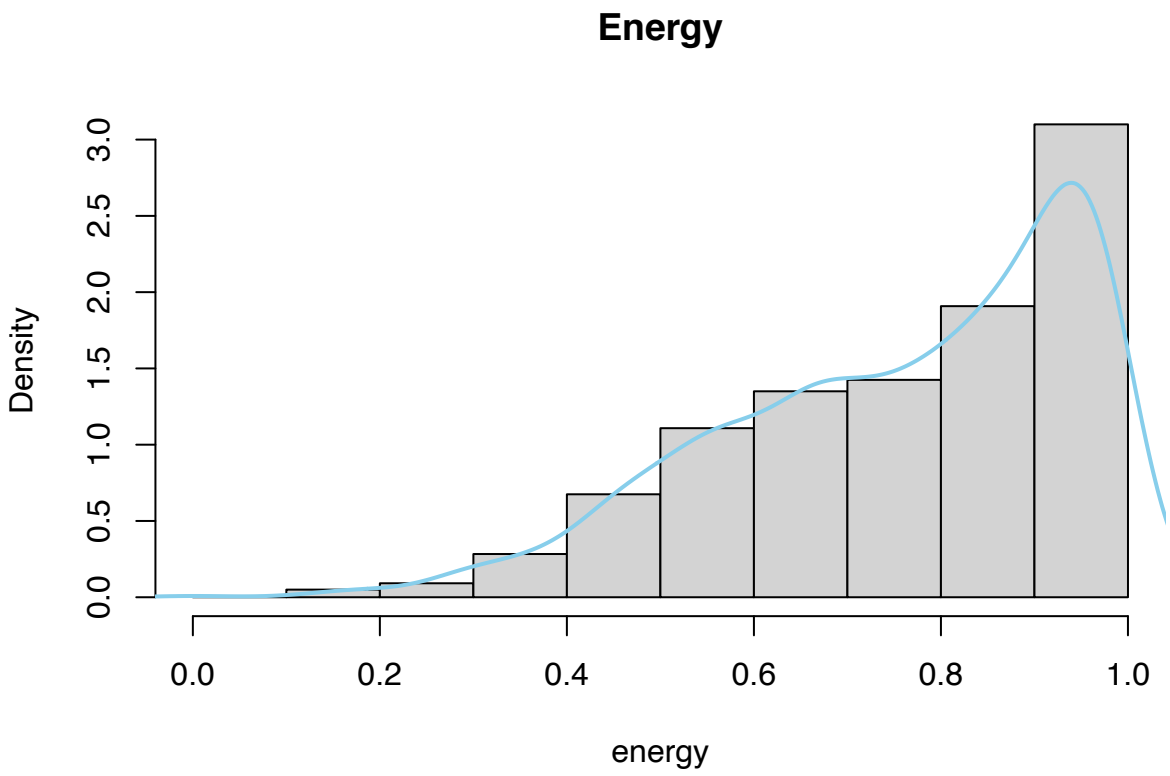



```
## [1] "555" "725"
```

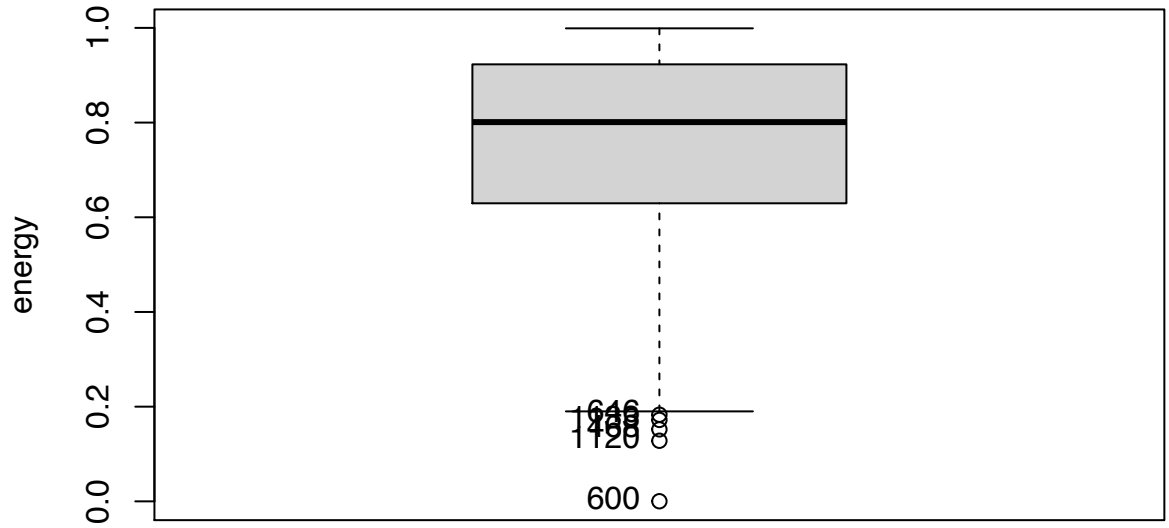
```
summary(danceability)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1420  0.5230  0.6440  0.6346  0.7632  0.9510
```

```
name <- "Energy"
hist(energy, main =name, prob=TRUE)
lines(density(energy), lwd = 2, col = "skyblue")
```

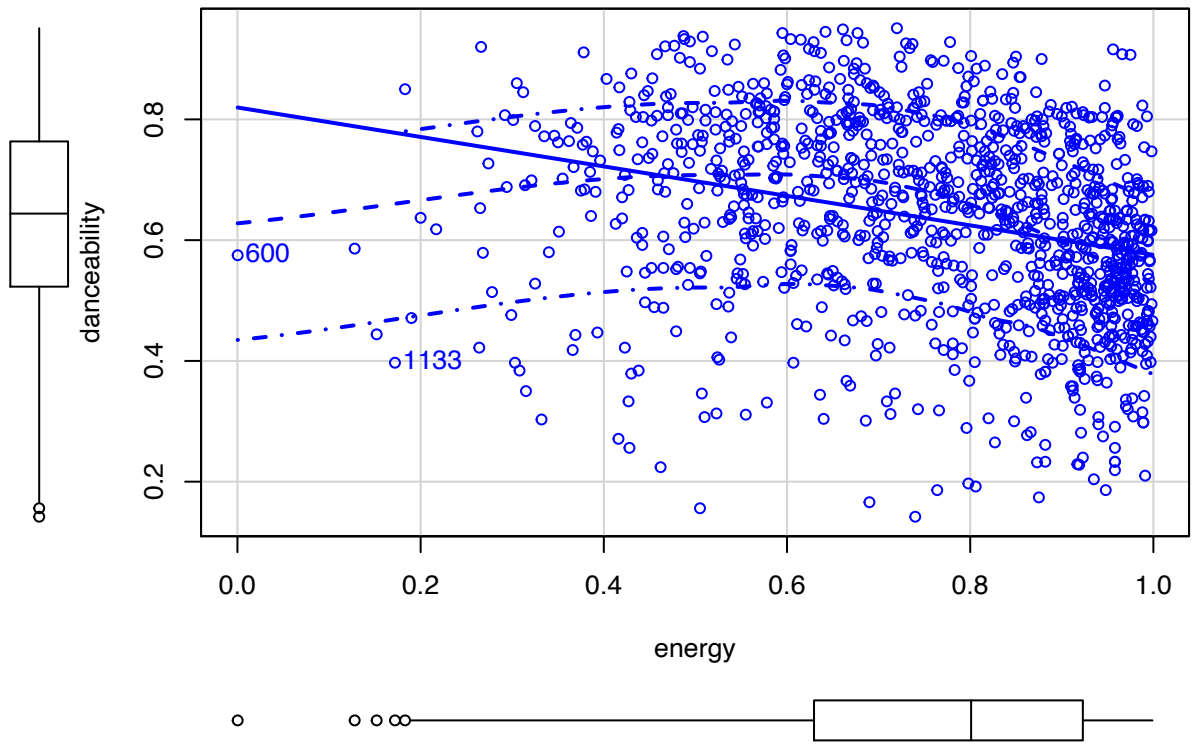



```
Boxplot(~energy)
```




```
## [1] "468"  "600"  "646"  "1120" "1133"
```

```
scatterplot(danceability ~energy, lwd=3, id=TRUE)
```

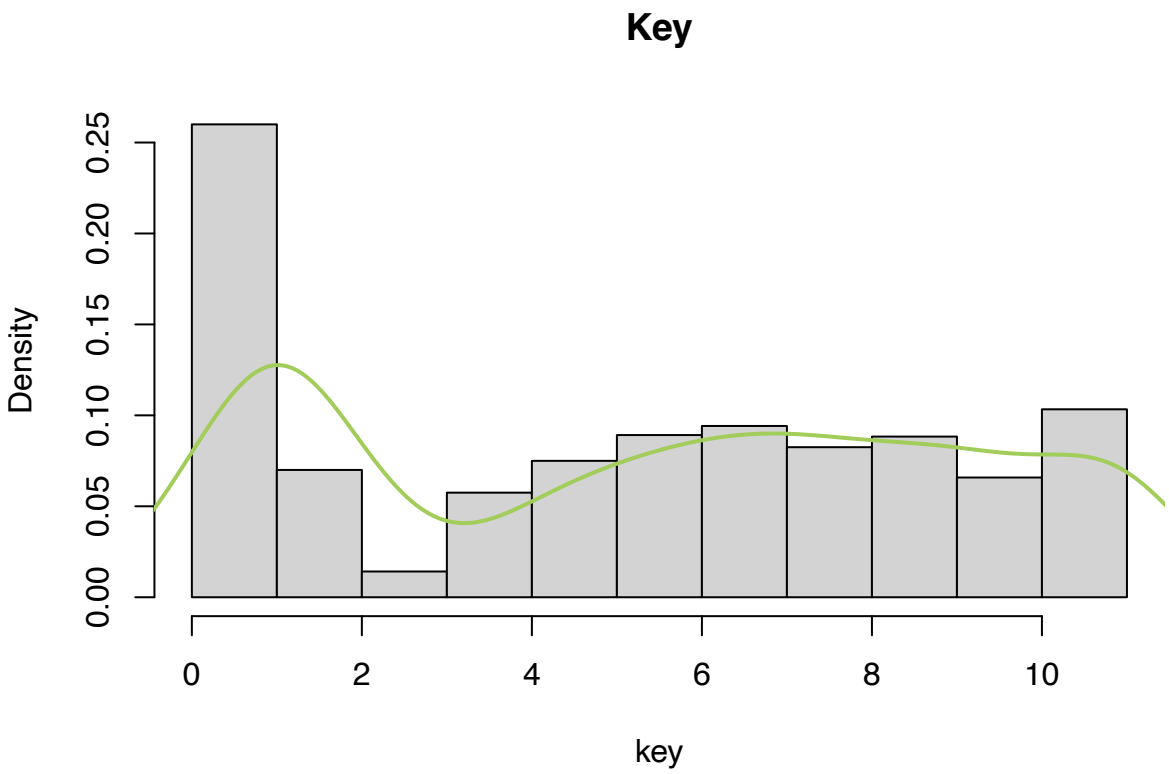



```
## [1]  600 1133
```

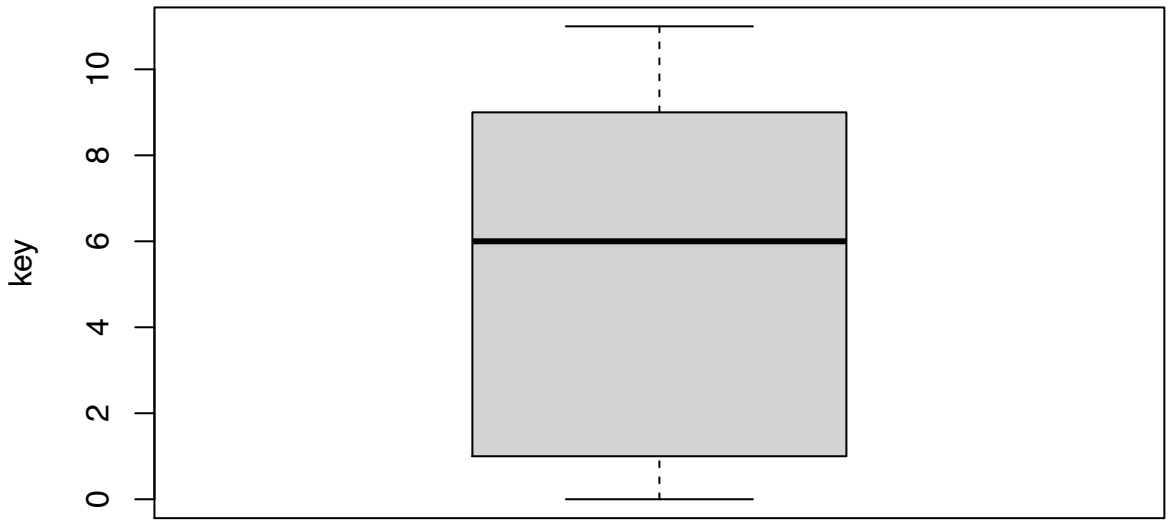
```
summary(energy)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000243 0.629750 0.801000 0.758784 0.923000 0.999000
```

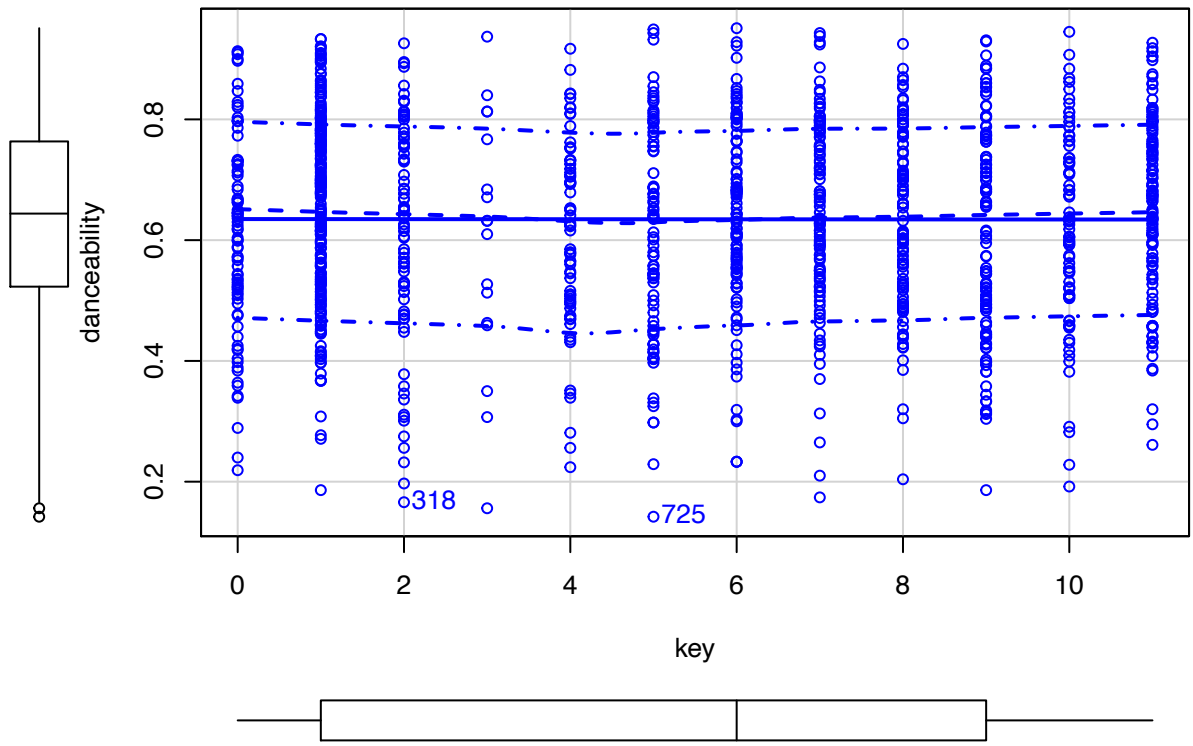
```
name <- "Key"
hist(key, main = name, prob=TRUE)
lines(density(key), lwd = 2, col = "darkolivegreen3")
```




```
Boxplot(~key)
```




```
scatterplot(danceability ~key, lwd=3, id=TRUE)
```



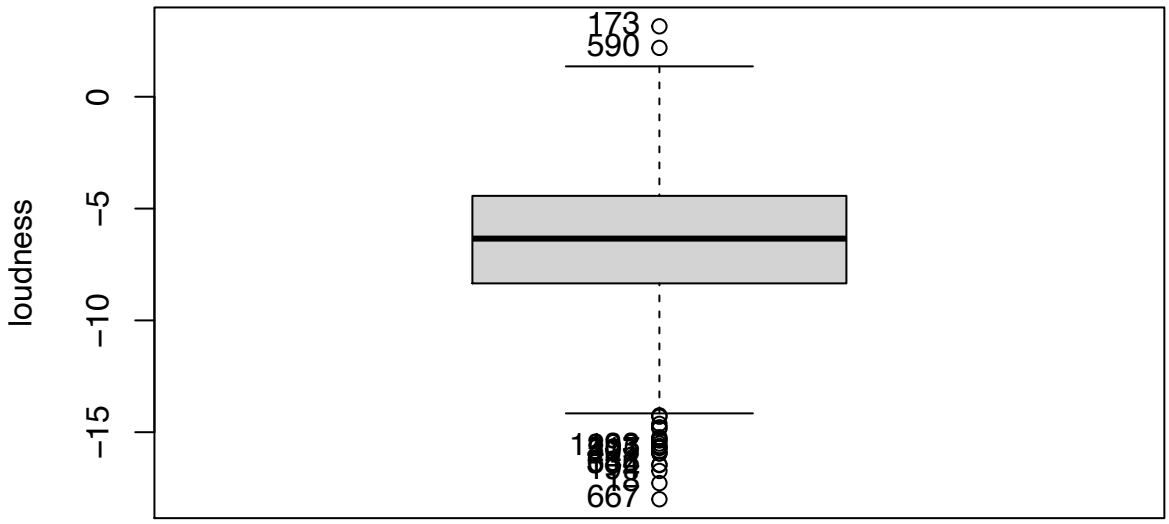

```
## [1] 318 725
```

```
summary(key)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   6.000   5.423   9.000  11.000
```

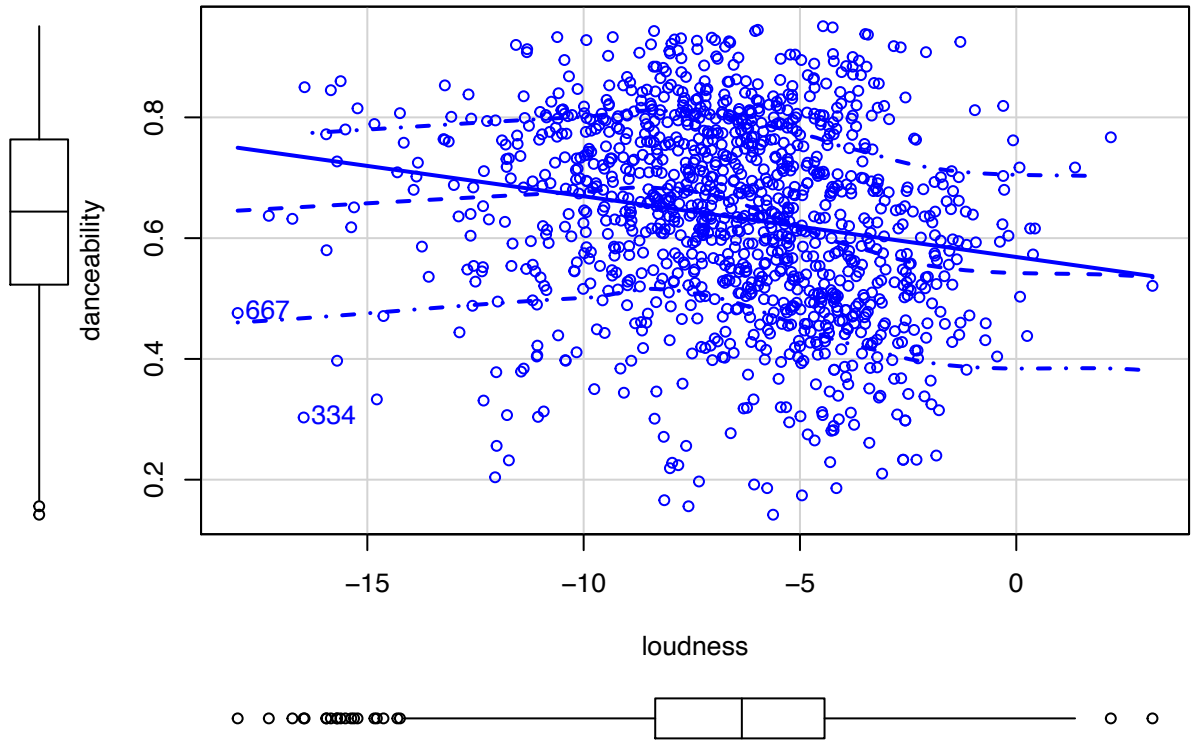
```
name <- "Loudness"
hist(loudness, main = name, prob = TRUE)
lines(density(loudness), lwd = 2, col = "springgreen")
```


```
Boxplot(~loudness)
```




```
##  [1] "667"  "18"   "194"  "334"  "646"  "421"  "205"  "517"  "293"  "1133"
## [11] "173"  "590"
```

```
scatterplot(danceability ~loudness, lwd=3, id=TRUE)
```

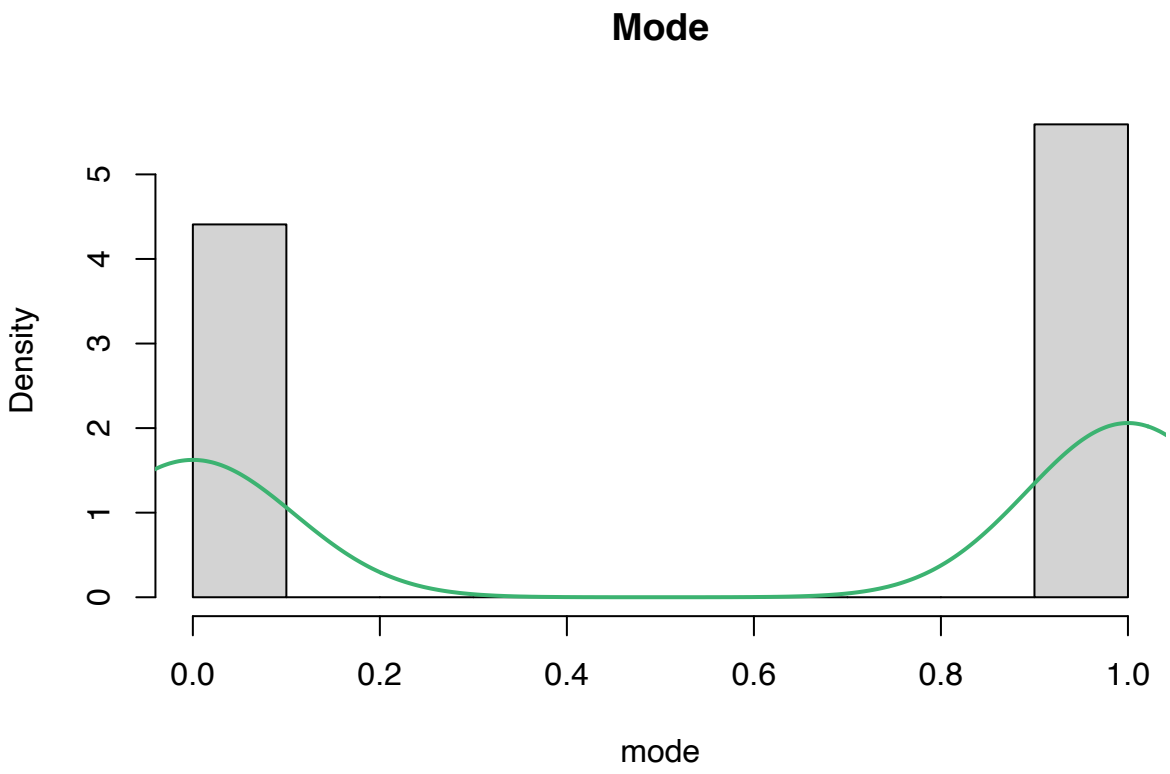



```
## [1] 334 667
```

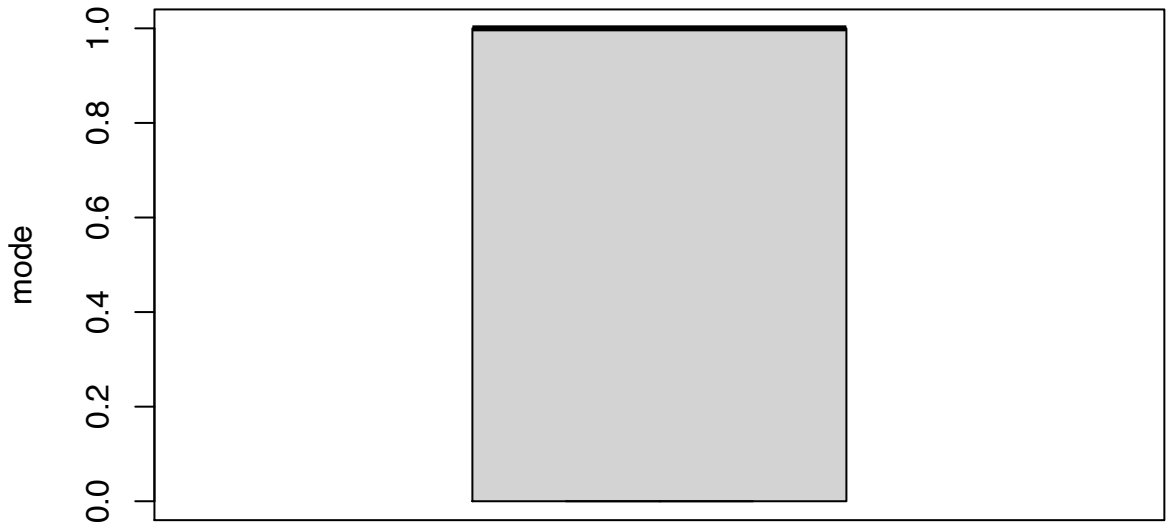
```
summary(loudness)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -17.997  -8.343  -6.343  -6.542  -4.434   3.148
```

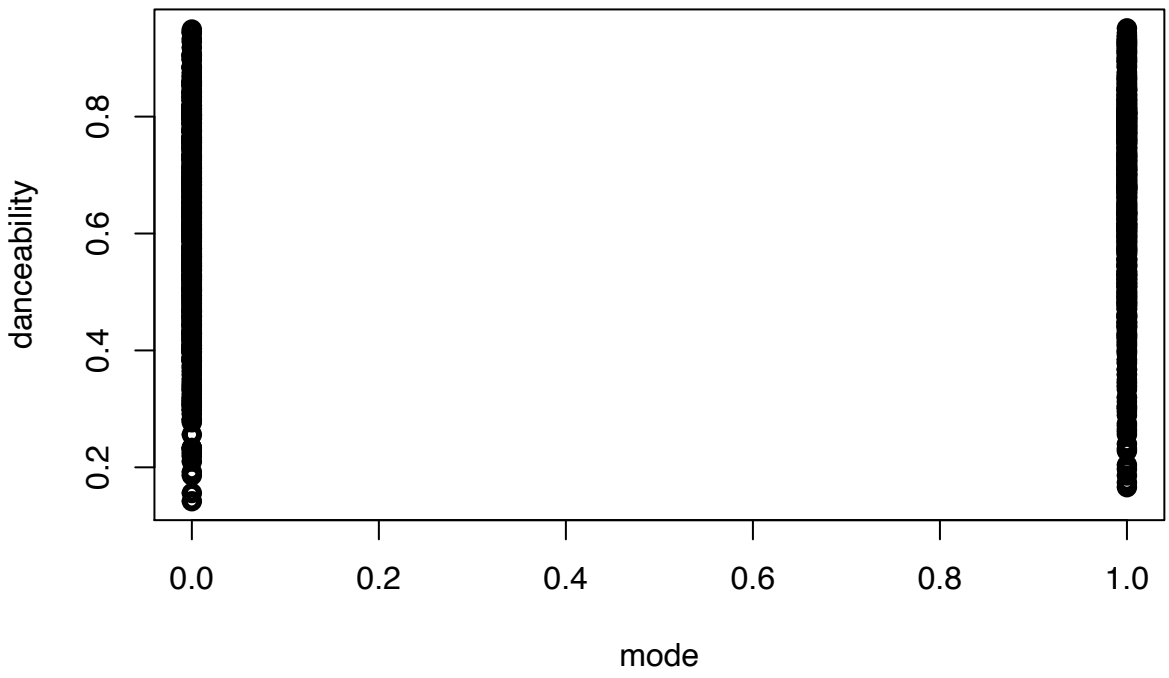
```
name <- "Mode"
hist(mode, main = name, prob = TRUE)
lines(density(mode), lwd = 2, col = "mediumseagreen")
```




```
Boxplot(~mode)
```



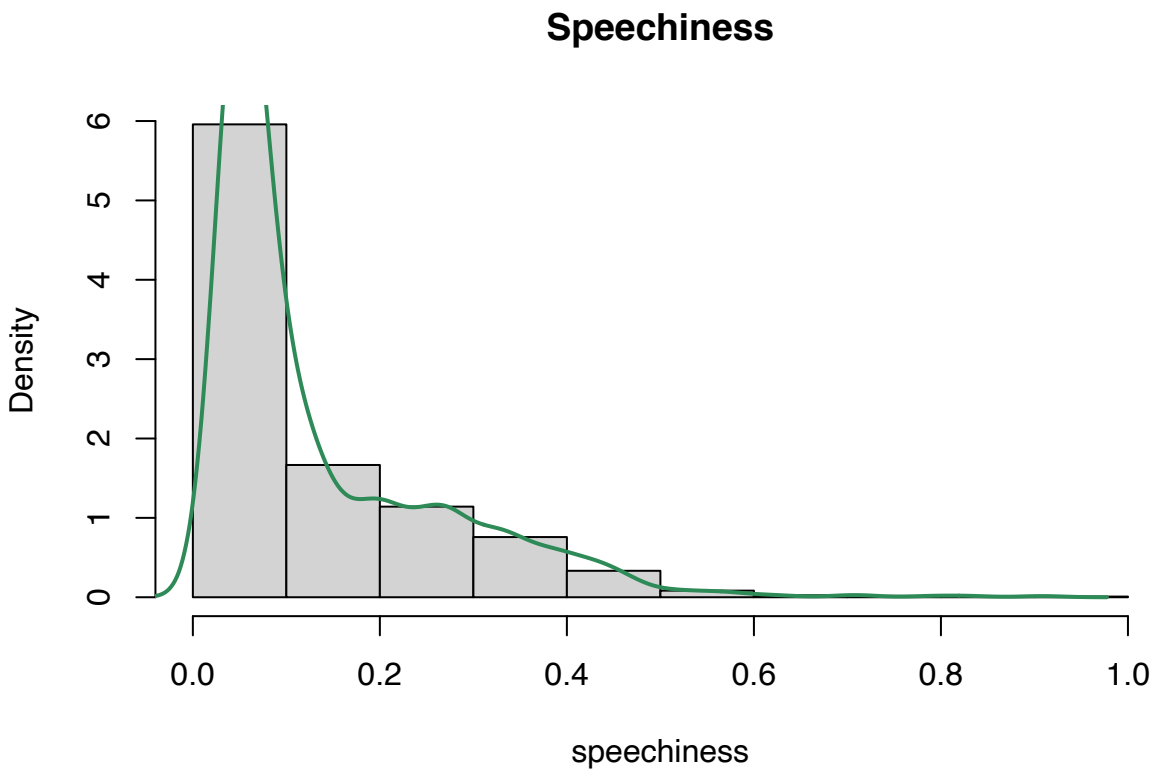

```
plot(danceability ~ mode, lwd=3, id=TRUE)
```



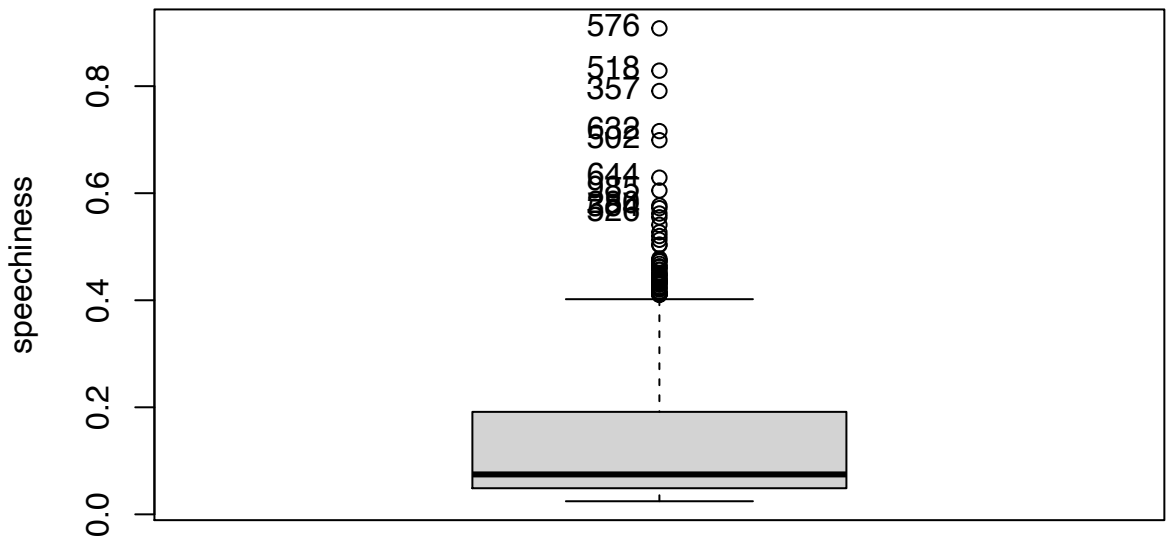

```
summary(mode)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  1.0000  0.5592  1.0000  1.0000
```

```
name <- "Speechiness"
hist(speechiness, main = name, prob = TRUE)
lines(density(speechiness), lwd = 2, col = "seagreen")
```

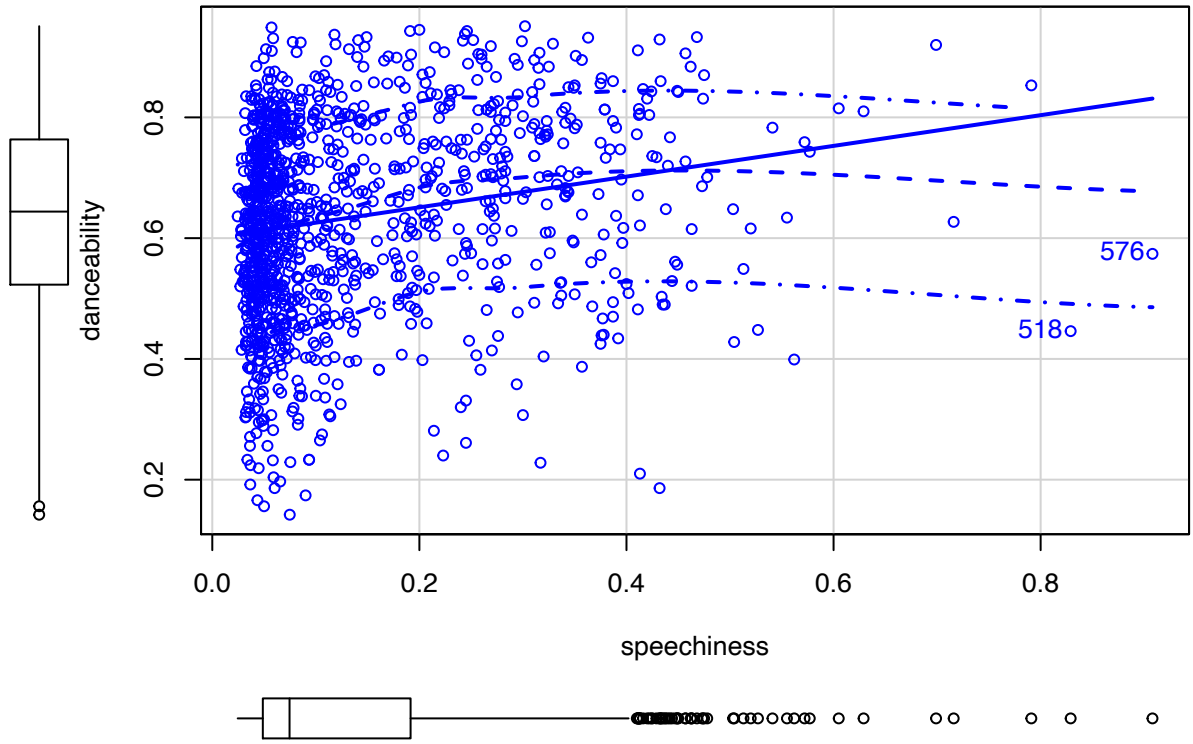



```
Boxplot(~speechiness)
```




```
##  [1] "576" "518" "357" "632" "502" "644" "985" "280" "384" "526"
```

```
scatterplot(danceability ~speechiness, lwd=3, id=TRUE)
```

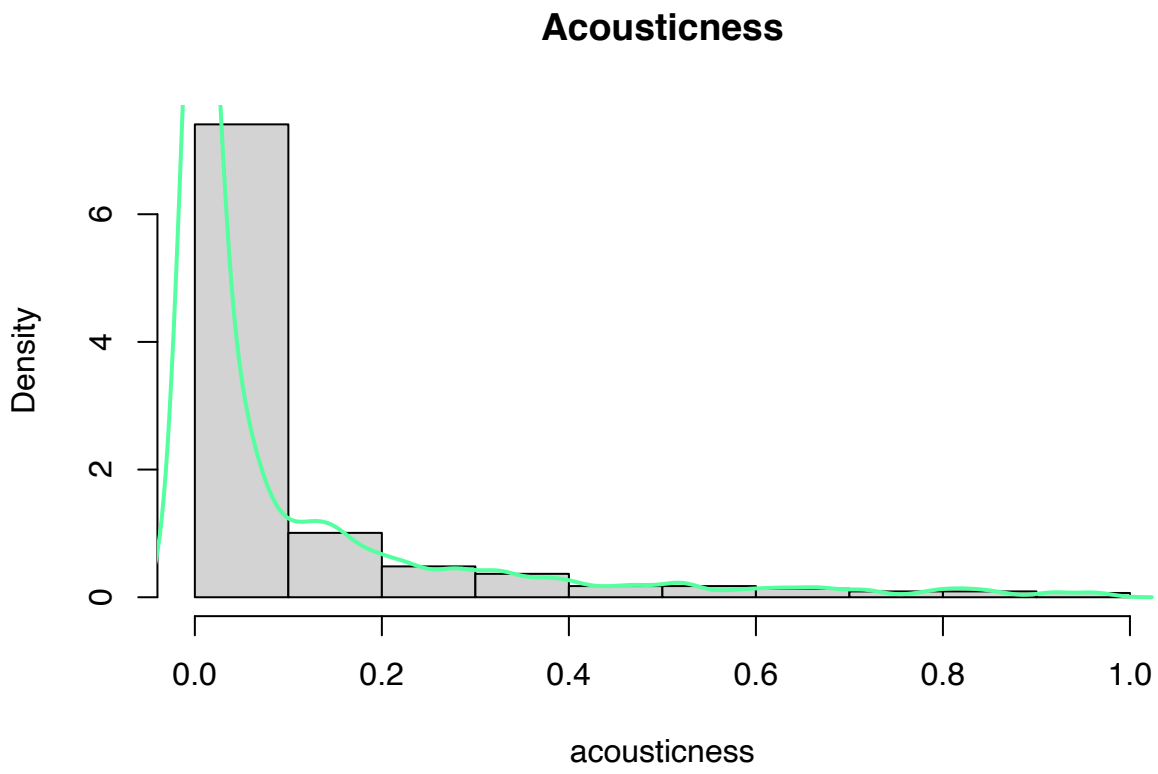



```
## [1] 518 576
```

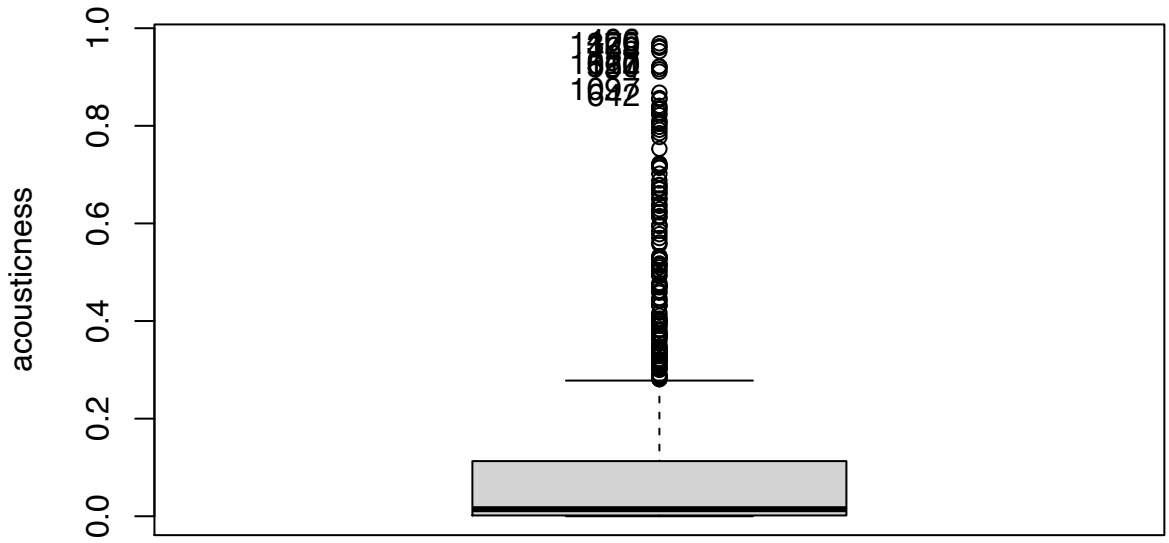
```
summary(speechiness)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0245  0.0488  0.0746  0.1352  0.1913  0.9080
```

```
name <- "Acousticness"
hist(acousticness, main = name, prob = TRUE)
lines(density(acousticness), lwd = 2, col = "seagreen1")
```

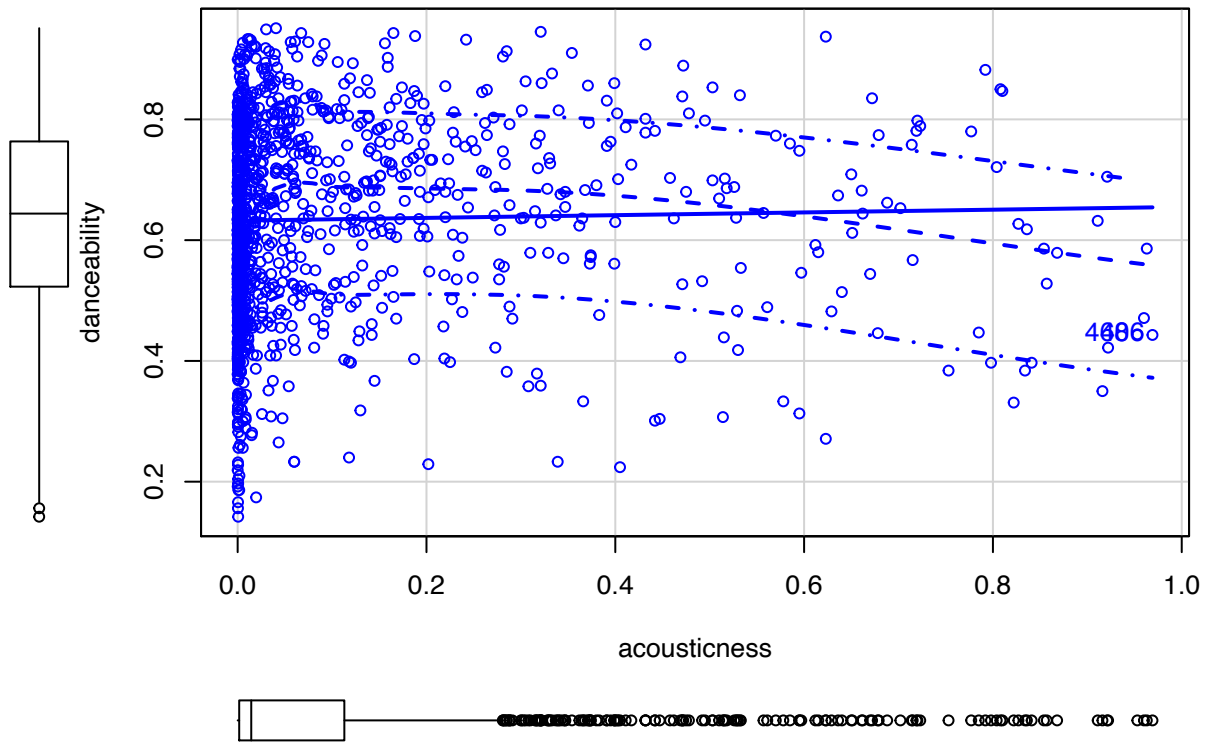



```
Boxplot(~acousticness)
```




```
##  [1] "406"  "1120" "375"  "468"  "327"  "1010" "550"  "194"  "1097" "642"
```

```
scatterplot(danceability ~acousticness, lwd=3, id=TRUE)
```

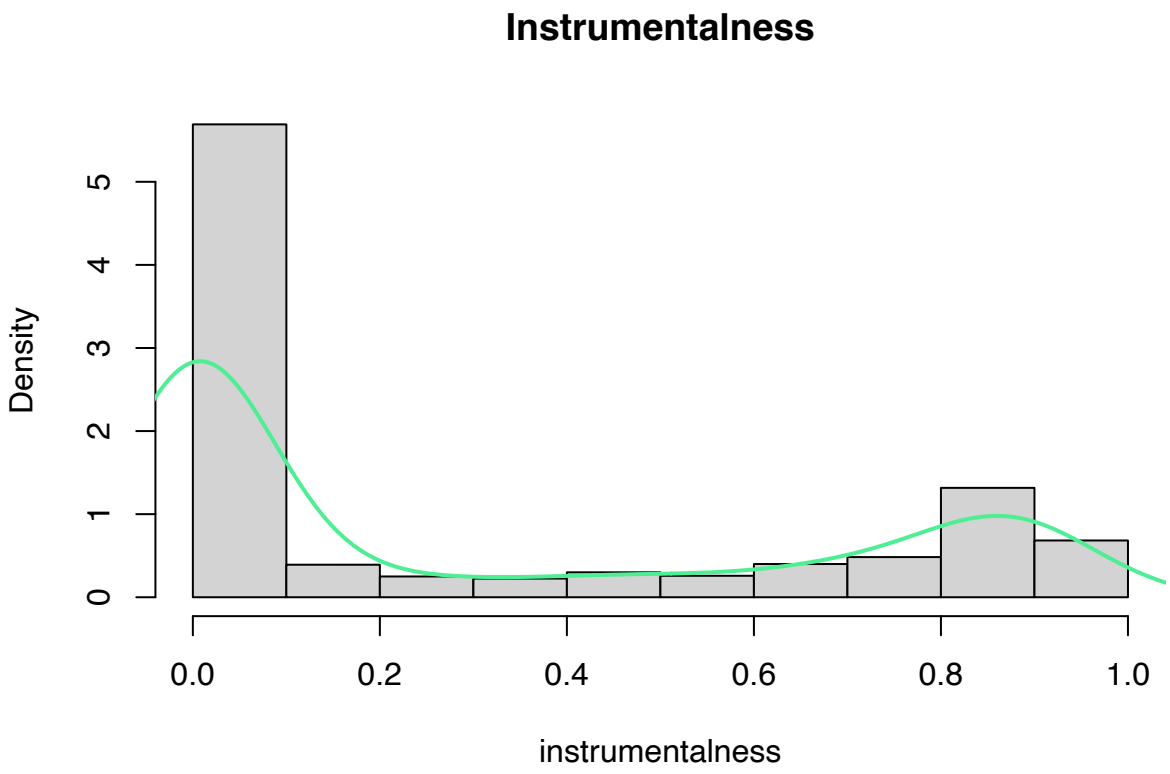



```
## [1] 406 468
```

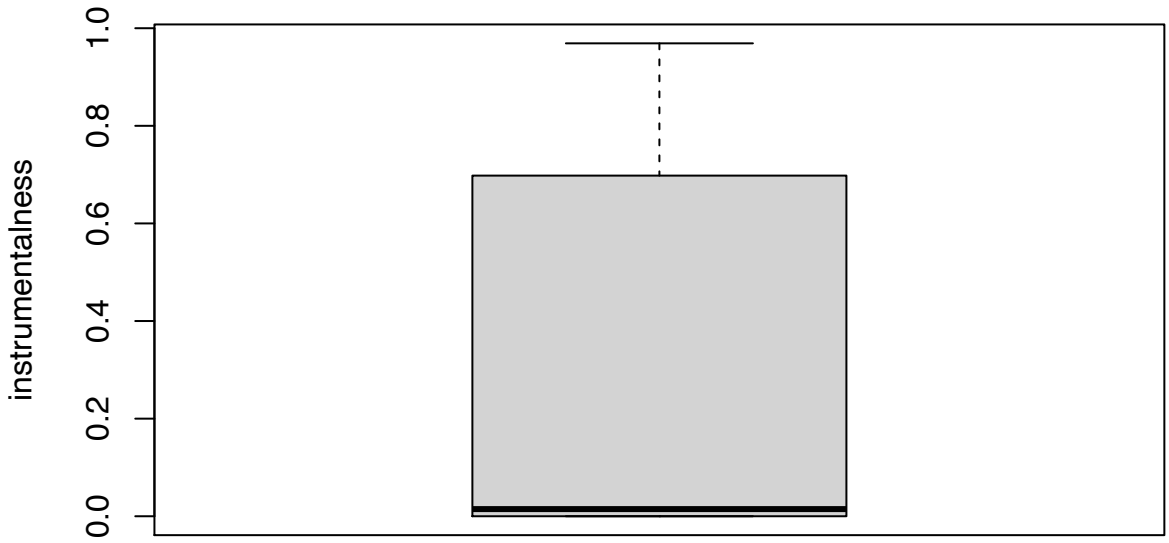
```
summary(acousticness)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0000034 0.0016275 0.0144000 0.0986608 0.1130000 0.9690000
```

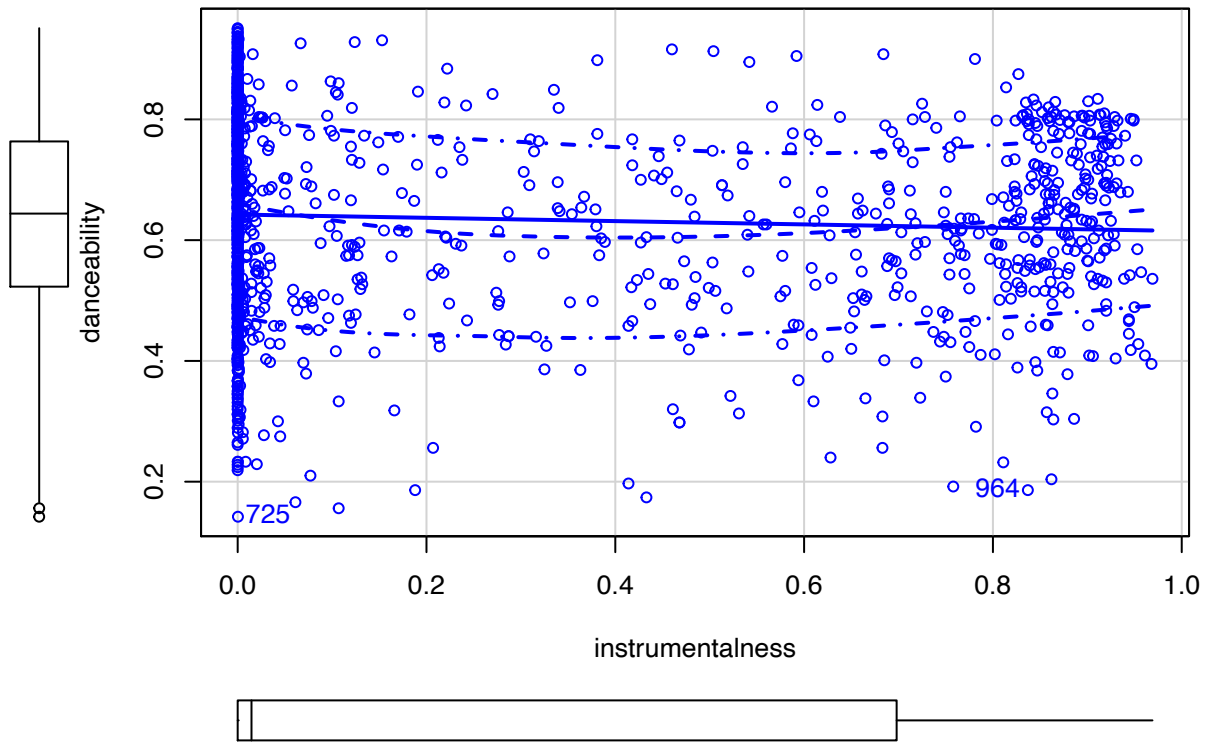
```
name <- "Instrumentalness"
hist(instrumentalness, main = name, prob = TRUE)
lines(density(instrumentalness), lwd = 2, col = "seagreen2")
```




```
Boxplot(~instrumentalness)
```




```
scatterplot(danceability ~instrumentalness, lwd=3, id=TRUE)
```

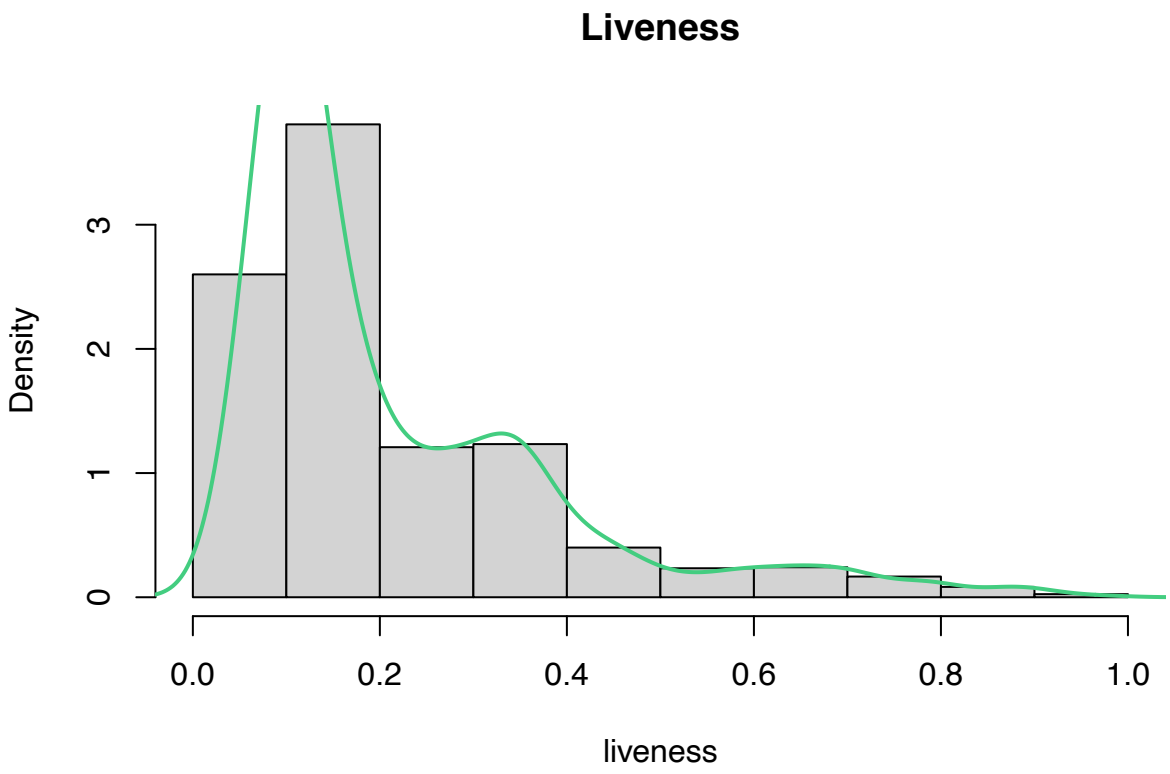



```
## [1] 725 964
```

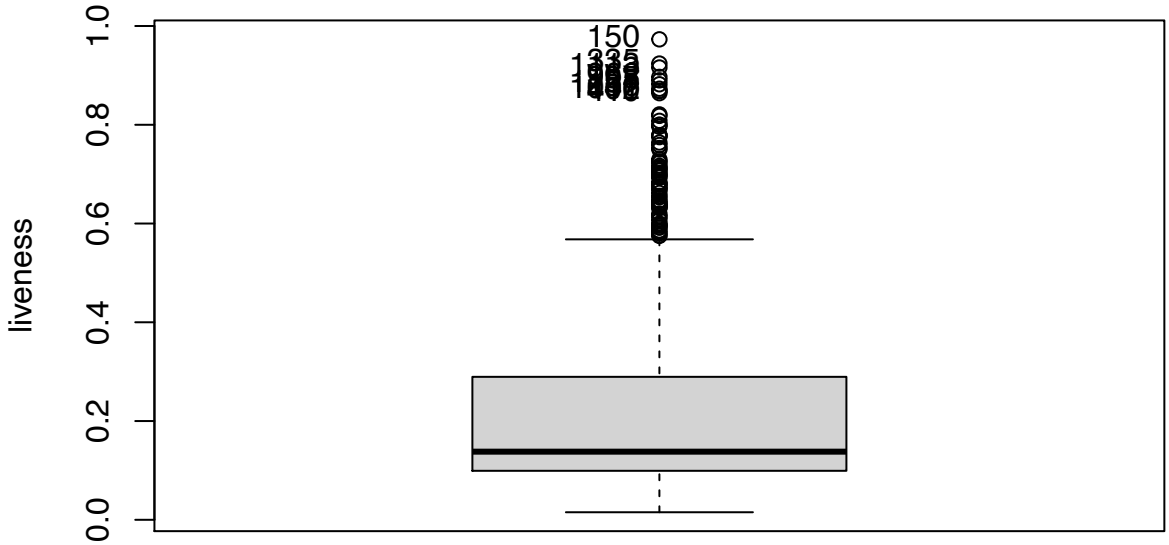
```
summary(instrumentalness)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0000000 0.0000029 0.0146500 0.2887877 0.6975000 0.9690000
```

```
name <- "Liveness"
hist(liveness, main = name, prob = TRUE)
lines(density(liveness), lwd = 2, col = "seagreen3")
```

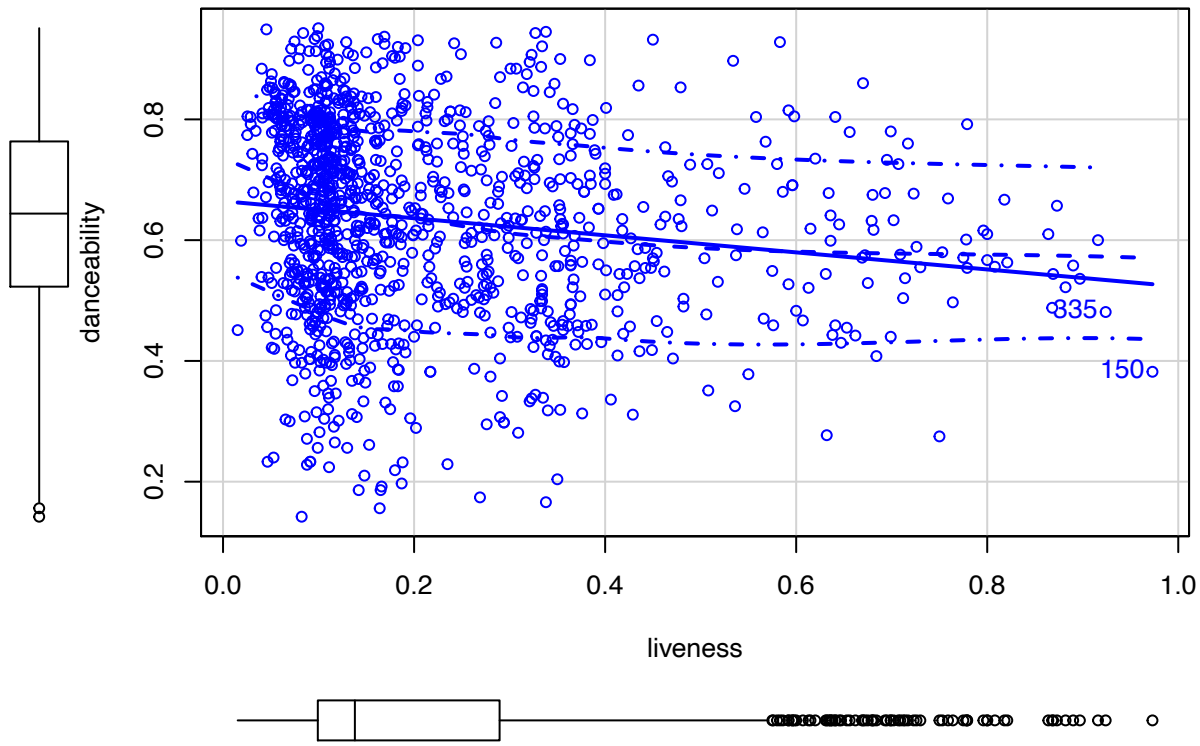



```
Boxplot(~liveness)
```




```
##  [1] "150"  "335"  "1113" "962"  "995"  "1157" "136"  "1041" "402"  "410"
```

```
scatterplot(danceability ~liveness, lwd=3, id=TRUE)
```

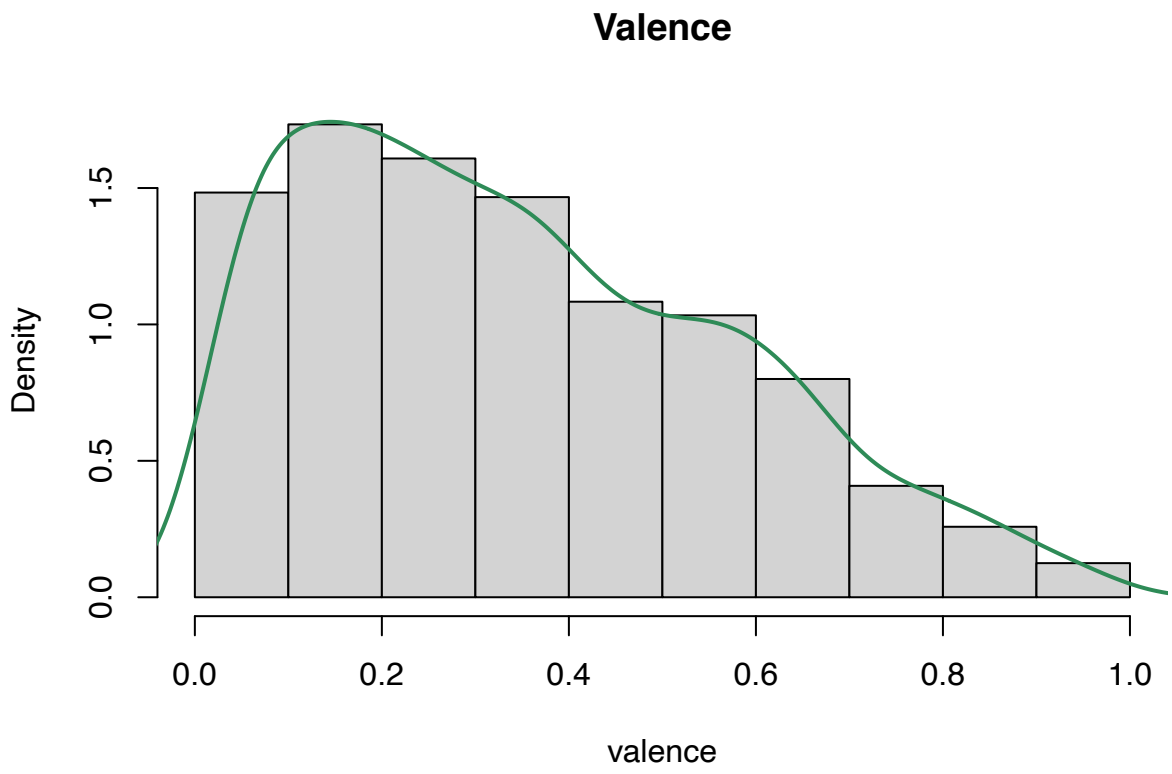



```
## [1] 150 335
```

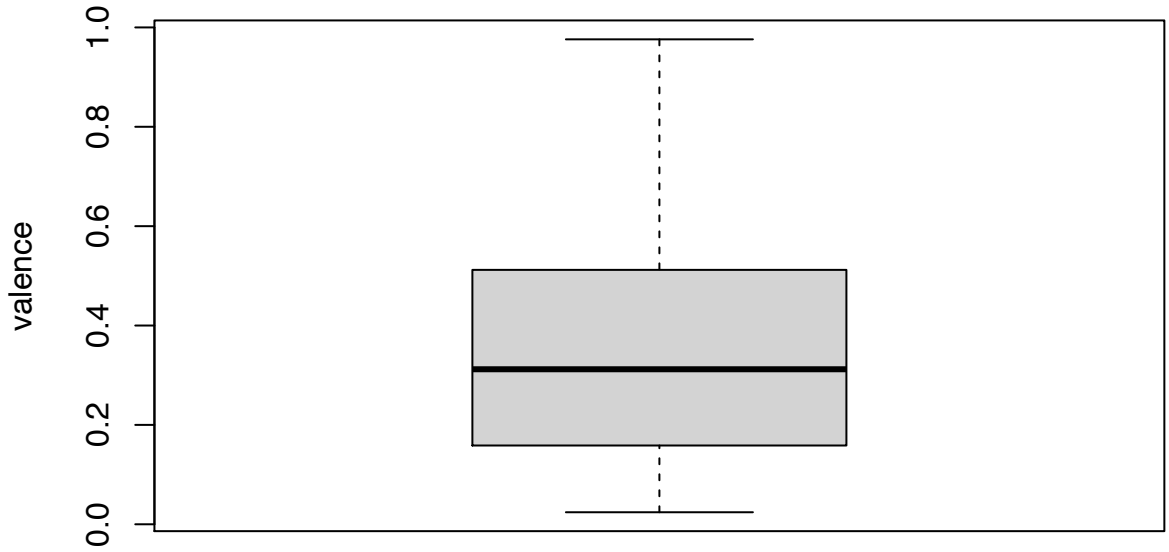
```
summary(liveness)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01530 0.09925 0.13800 0.21204 0.28925 0.97300
```

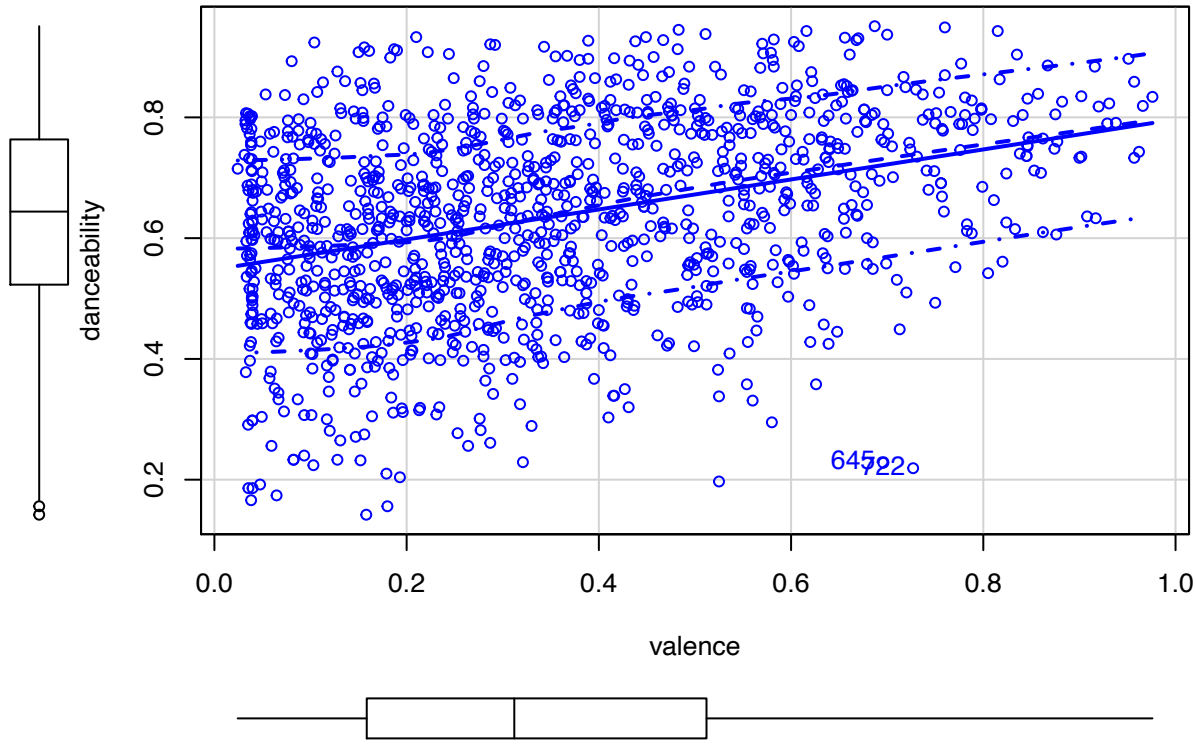
```
name <- "Valence"
hist(valence, main = name, prob = TRUE)
lines(density(valence), lwd = 2, col = "seagreen4")
```




```
Boxplot(~valence)
```




```
scatterplot(danceability ~valence, lwd=3, id=TRUE)
```

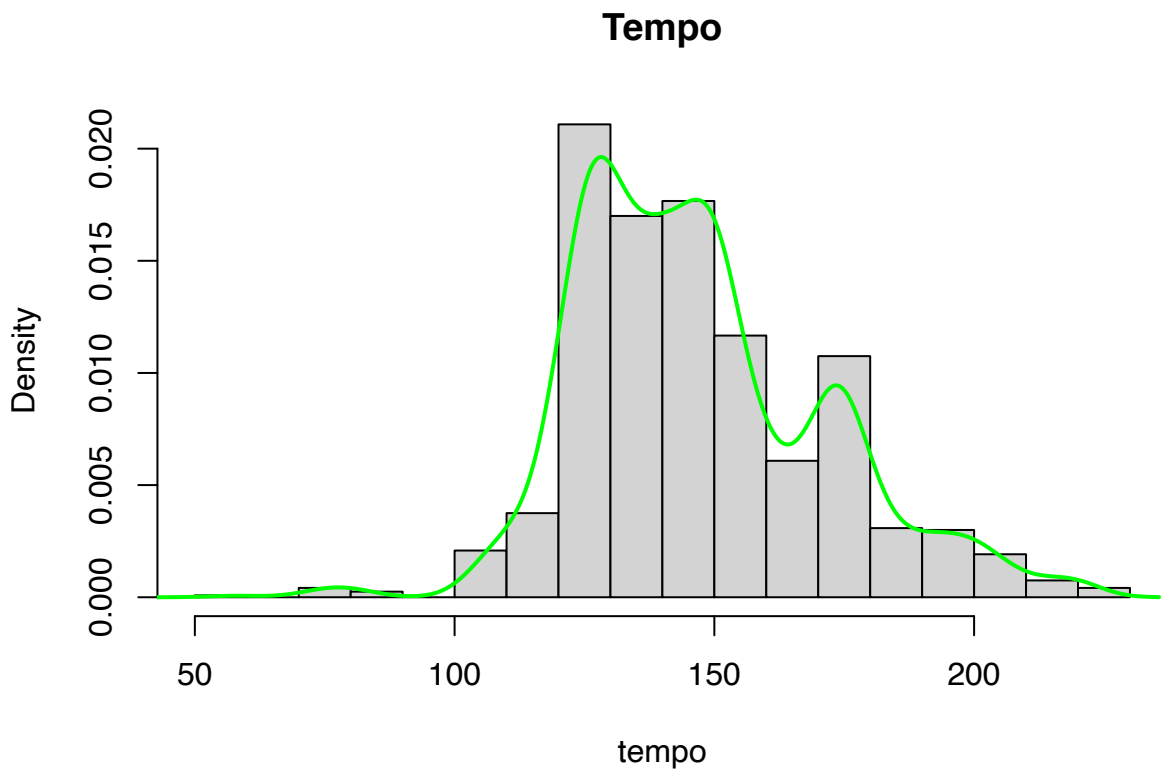



```
## [1] 645 722
```

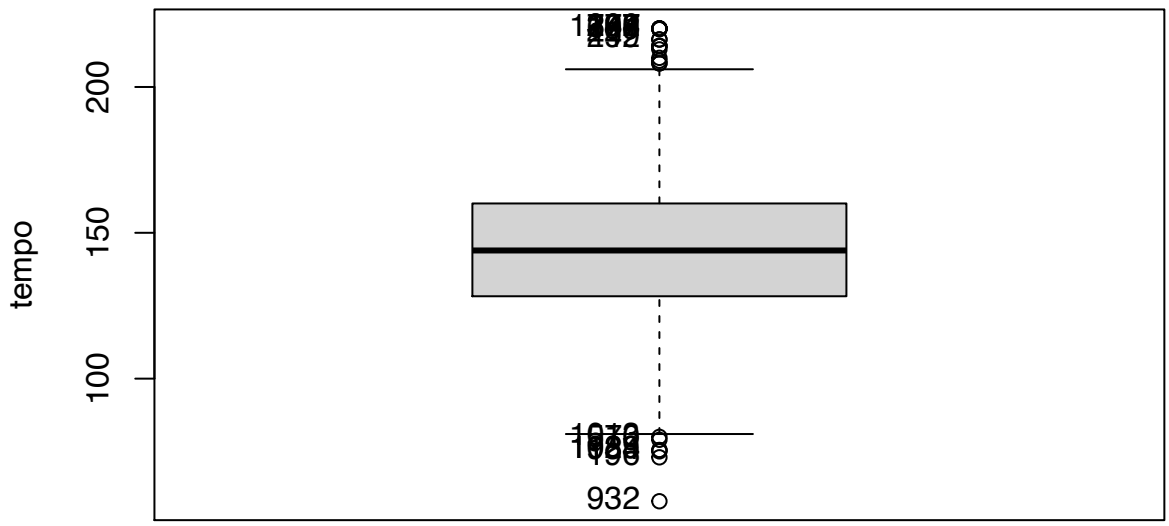
```
summary(valence)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0242  0.1588  0.3120  0.3479  0.5120  0.9760
```

```
name <- "Tempo"
hist(tempo, main = name, prob = TRUE)
lines(density(tempo), lwd = 2, col = "green")
```

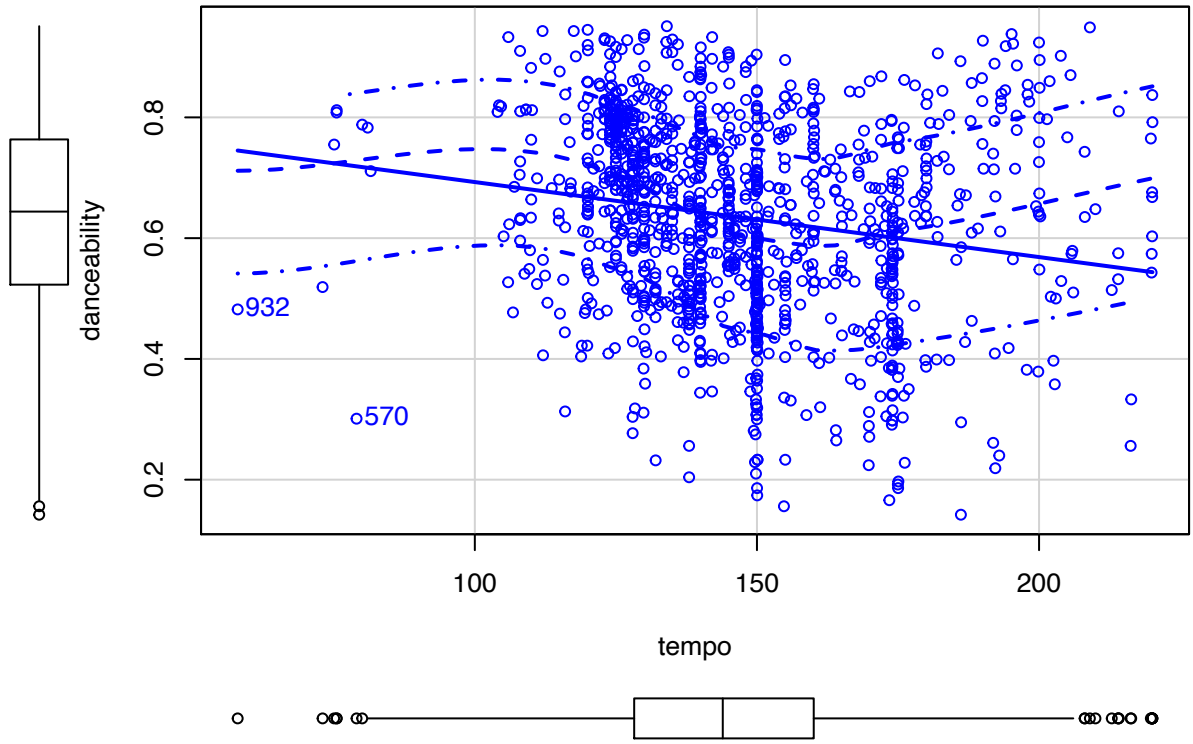



```
Boxplot(~tempo)
```




```
##  [1] "196"  "570"  "932"  "934"  "1013" "1029" "1183" "268"  "177"  "300"
## [11] "242"  "394"  "757"  "1137" "641"  "442"  "299"
```

```
scatterplot(danceability ~tempo, lwd=3, id=TRUE)
```

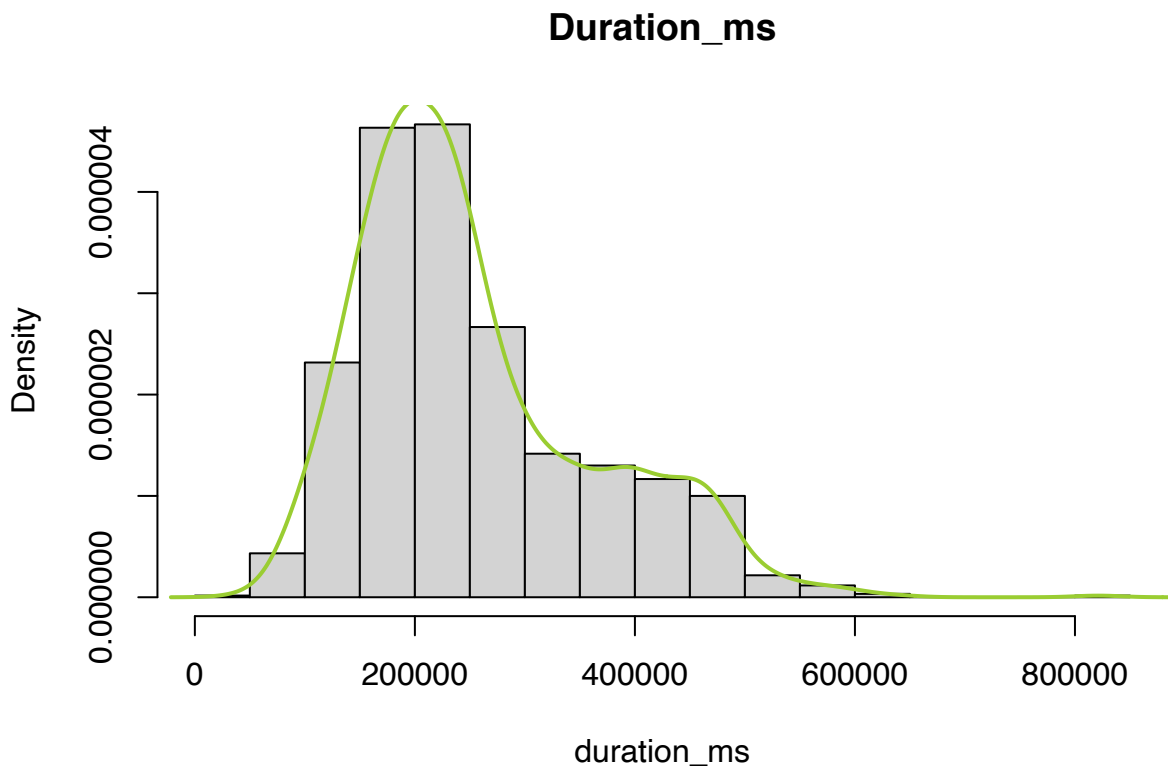



```
## [1] 570 932
```

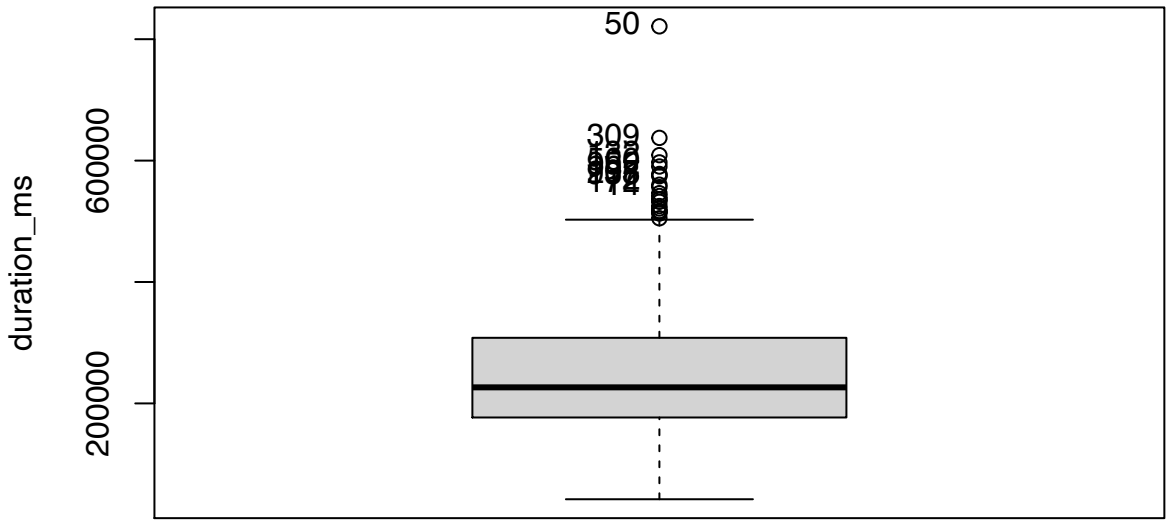
```
summary(tempo)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   57.97  128.31  143.94  146.97  160.06  220.10
```

```
name <- "Duration_ms"
hist(duration_ms, main = name, prob = TRUE)
lines(density(duration_ms), lwd = 2, col = "yellowgreen")
```

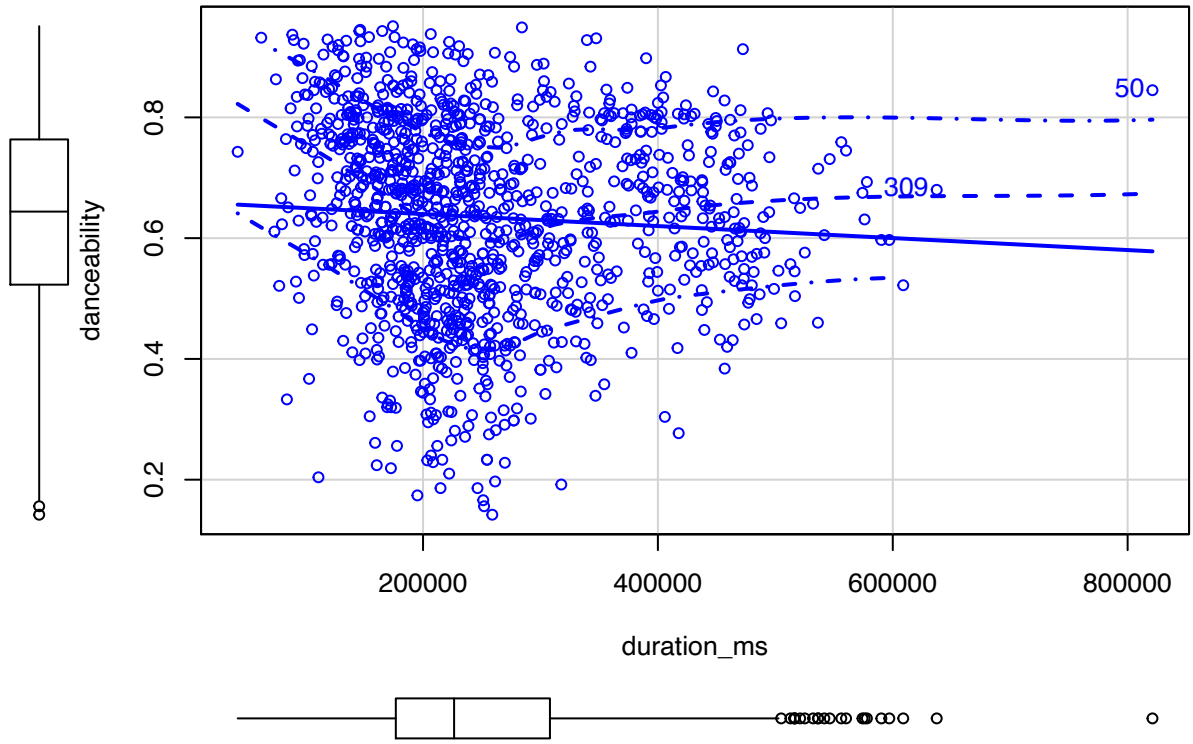



```
Boxplot(~duration_ms)
```




```
##  [1] "50"  "309" "132" "560" "952" "197" "203" "938" "172" "14"
```

```
scatterplot(danceability ~duration_ms, lwd=3, id=TRUE)
```




```
## [1]  50 309
```

```
summary(duration_ms)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   42133  176822  226452  253204  307785  821168
```

Question 2

Estimate a multiple linear regression model that includes all the main effects only (i.e., no interactions nor higher order terms). We will use this model as a baseline. Comment on the statistical and economic significance of your estimates. Also, make sure to provide an interpretation of your estimates.

```
reg.mod.2 <- lm(danceability ~ energy + key + loudness + mode + speechiness +
                  acousticness + instrumentalness + liveness + valence +
                  tempo + duration_ms)
summary(reg.mod.2)
```

```
##
## Call:
## lm(formula = danceability ~ energy + key + loudness + mode +
##     speechiness + acousticness + instrumentalness + liveness +
##     valence + tempo + duration_ms)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48835 -0.08074  0.01301  0.08802  0.40456
##
## Coefficients:
##                       Estimate     Std. Error t value             Pr(>|t|)
## (Intercept)       0.95306627023  0.04545204873  20.969 < 0.0000000000000002 ***
## energy           -0.33592989741  0.03400603879  -9.879 < 0.0000000000000002 ***
## key               0.00038855231  0.00108914396   0.357             0.721342
## loudness          0.00188699529  0.00189143954   0.998             0.318652
## mode              0.02588598445  0.00805070804   3.215             0.001338 **
## speechiness       0.22190586098  0.03461007632   6.412    0.000000000207326 ***
## acousticness     -0.16252899931  0.02582980638  -6.292    0.000000000438990 ***
## instrumentalness  0.04785497552  0.01460535369   3.277             0.001081 **
## liveness         -0.08126345089  0.02357538608  -3.447             0.000587 ***
## valence           0.24269132355  0.01811264319  13.399 < 0.0000000000000002 ***
## tempo            -0.00122340625  0.00016951097  -7.217    0.000000000000944 ***
## duration_ms       0.00000006707  0.00000004598   1.459             0.144867
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1347 on 1188 degrees of freedom
## Multiple R-squared:  0.3035, Adjusted R-squared:  0.2971
## F-statistic: 47.07 on 11 and 1188 DF,  p-value: < 0.00000000000000022
```

```
tidy(reg.mod.2)
```

```
## # A tibble: 12 x 5
##    term                  estimate    std.error statistic  p.value
##    <chr>                    <dbl>        <dbl>     <dbl>    <dbl>
##  1 (Intercept)       0.953        0.0455          21.0   2.61e-83
##  2 energy           -0.336        0.0340          -9.88  3.59e-22
##  3 key               0.000389     0.00109          0.357 7.21e- 1
##  4 loudness          0.00189      0.00189          0.998 3.19e- 1
##  5 mode              0.0259       0.00805          3.22  1.34e- 3
##  6 speechiness       0.222        0.0346           6.41  2.07e-10
##  7 acousticness     -0.163        0.0258          -6.29  4.39e-10
##  8 instrumentalness  0.0479       0.0146           3.28  1.08e- 3
##  9 liveness         -0.0813       0.0236          -3.45  5.87e- 4
## 10 valence           0.243        0.0181          13.4   3.14e-38
## 11 tempo            -0.00122      0.000170        -7.22  9.44e-13
## 12 duration_ms       0.0000000671 0.0000000460     1.46  1.45e- 1
```

Based on the summary above, everything except "key," "mode," "loudness," and "duration_ms," looks significant at 5% level of significance.

Holding all other variables constant, one unit increase in energy decreases expected danceability by 0.3359.

Holding all other variables constant, one unit increase in key increases expected danceability by 0.00038.

Holding all other variables constant, one unit increase in loudness increases expected danceability by 0.0018.

Holding all other variables constant, one unit increase in mode increases expected danceability by 0.0258.

Holding all other variables constant, one unit increase in speechiness increases expected danceability by 0.2219.

Holding all other variables constant, one unit increase in acousticness decreases expected danceability by 0.1625.

Holding all other variables constant, one unit increase in instrumentalness increases expected danceability by 0.04785.

Holding all other variables constant, one unit increase in liveness decreases expected danceability by 0.0812.

Holding all other variables constant, one unit increase in valence increases expected danceability by 0.2426.

Holding all other variables constant, one unit increase in tempo decreases expected danceability by 0.00122.

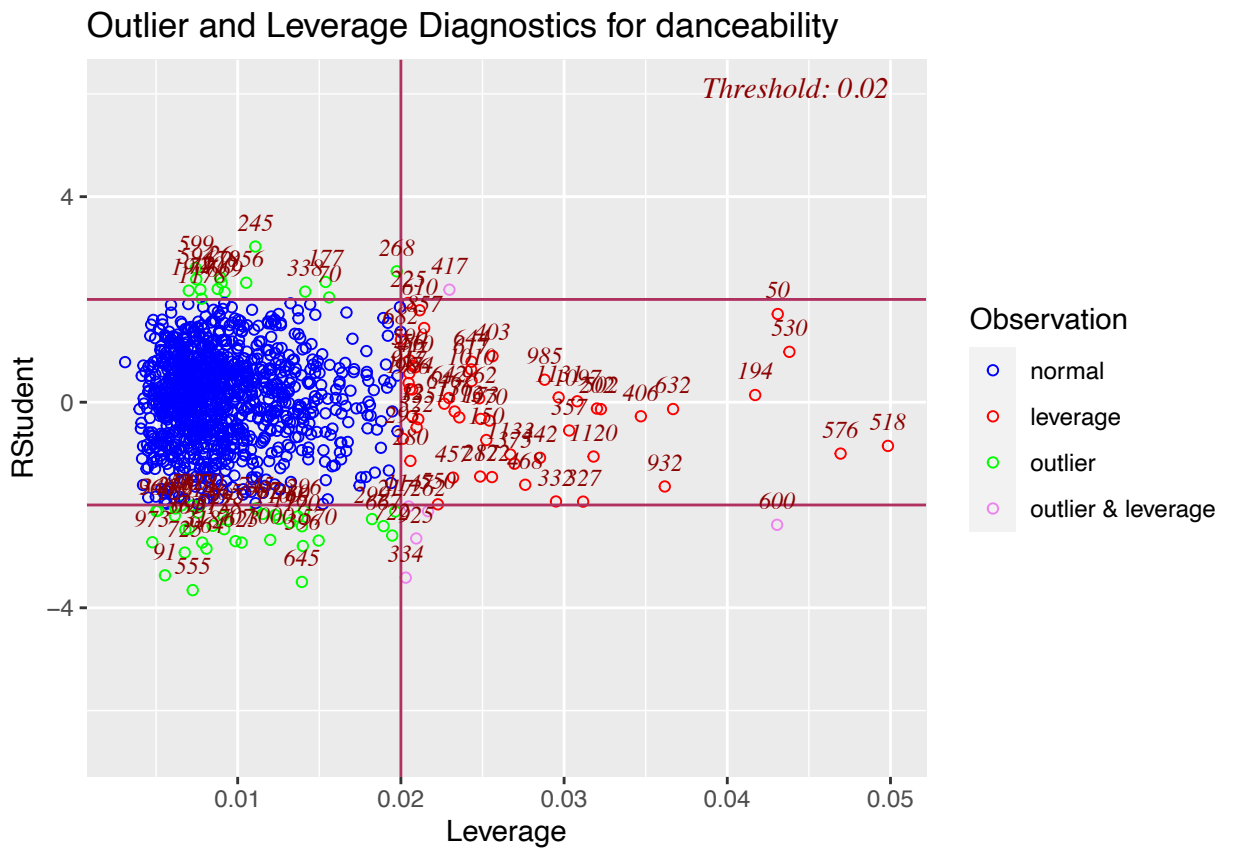
Holding all other variables constant, one millisecond increase in duration increases expected danceability by 0.000000067.

None of them are economically significant because our data does not include any economic values.

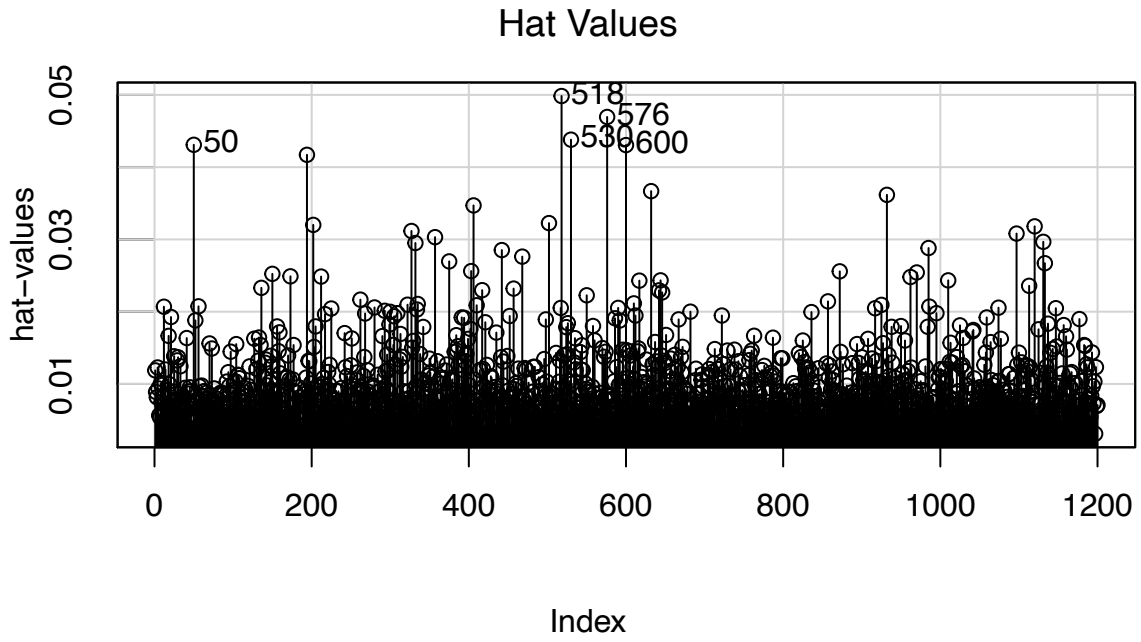
Question 3

Identify if there are any outliers, high leverage, and or influential observations worth removing. If so, remove them but justify your reason for doing so and re-estimate your model.

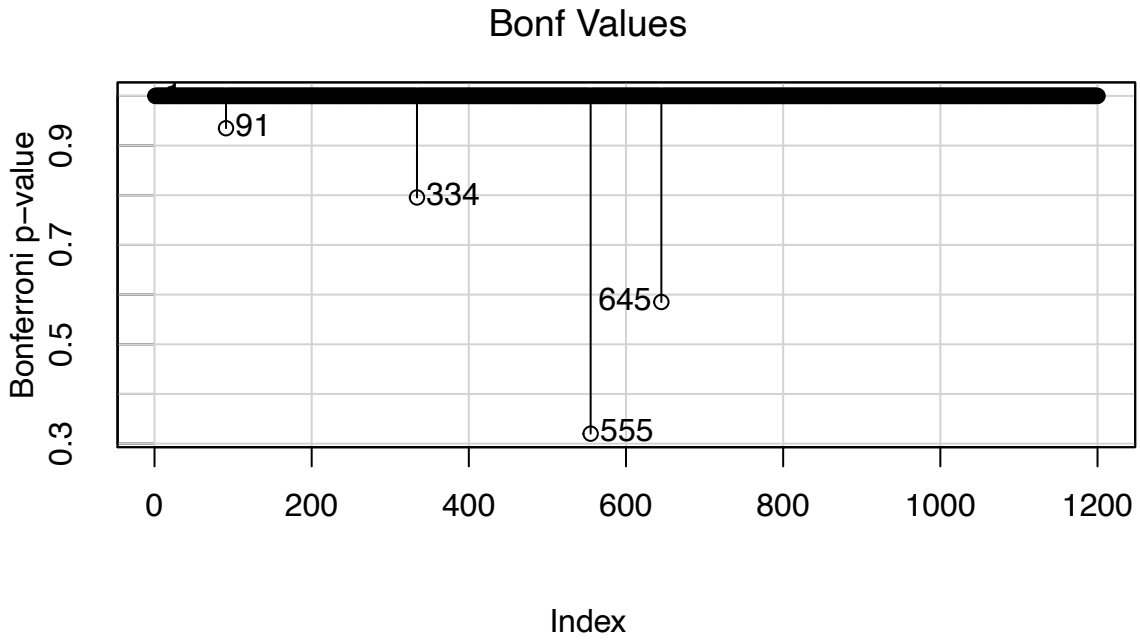
```
library(olsrr)
library(car)

ols_plot_resid_lev(reg.mod.2)
```



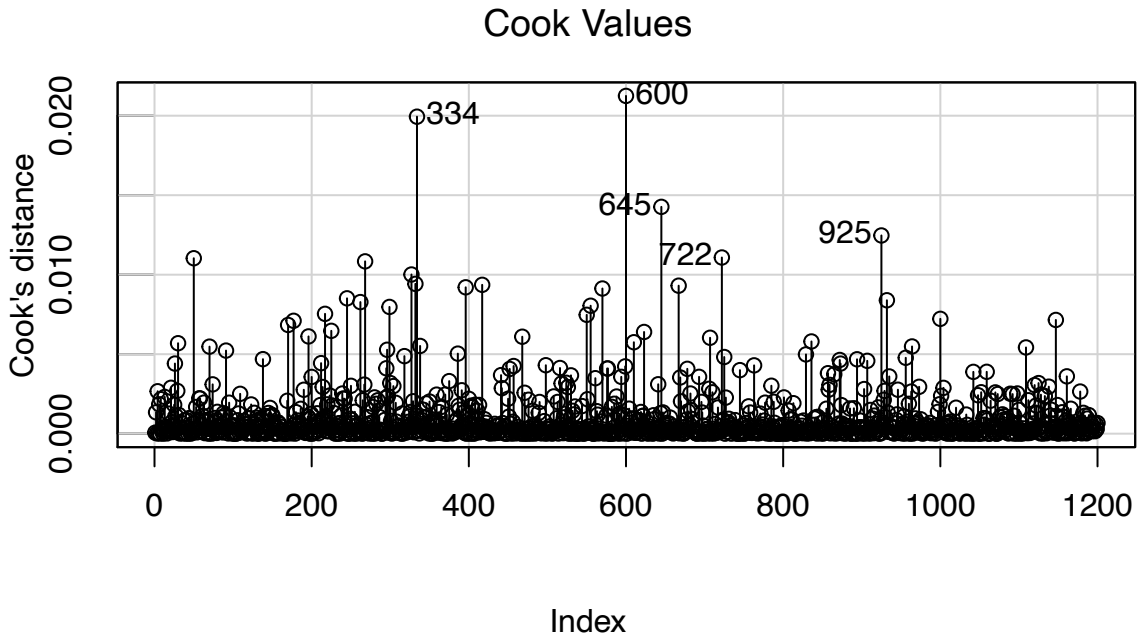
```r
# leverages
influenceIndexPlot(reg.mod.2, id=list(n=5), vars="hat", main="Hat Values")
```



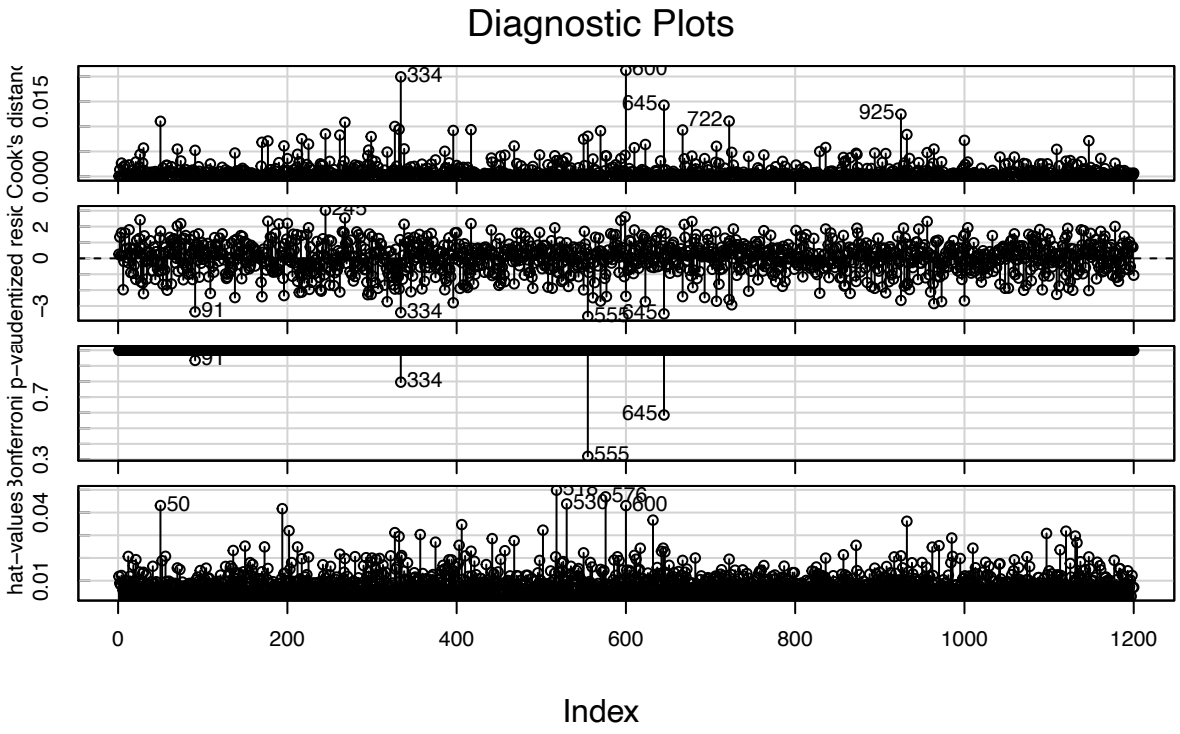

```r
# outliers
influenceIndexPlot(reg.mod.2, id=list(n=5), vars="Bonf", main="Bonf Values")
```




```
# influential
influenceIndexPlot(reg.mod.2, id=list(n=5), vars="cook", main="Cook Values")
```




```
#all
influenceIndexPlot(reg.mod.2, id=list(n=5))
```



Based on the plots above, observations 50, 518, 530, 576 have high leverages.

Based on the plots above, observations 91 and 555 are outliers.

Based on the first plots above, observation 334, 600, 645, 722, and 925 are influential points.

Given that there are many leverage, outlier, and influential points, we are taking out the 5 most significant values (some of the values overlap).

```
# Do again without the leverages, outliers, influential points

reg.mod.3 <- lm(danceability ~ energy + key + loudness + mode + speechiness +
                  acousticness + instrumentalness + liveness + valence + tempo
                + duration_ms, subset= c(-50, -518, -530, -576, -91, -555,
                                         -334, -645, -600, -722, -925))
summary(reg.mod.3)
```

```
##
## Call:
## lm(formula = danceability ~ energy + key + loudness + mode +
##     speechiness + acousticness + instrumentalness + liveness +
##     valence + tempo + duration_ms, subset = c(-50, -518, -530,
##     -576, -91, -555, -334, -645, -600, -722, -925))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39921 -0.08271  0.01243  0.08627  0.40138
```

```
##
## Coefficients:
##                       Estimate     Std. Error t value             Pr(>|t|)
## (Intercept)       0.92868495958  0.04493896112  20.665 < 0.0000000000000002 ***
## energy           -0.33923720287  0.03413861363  -9.937 < 0.0000000000000002 ***
## key               0.00011926850  0.00106597406   0.112              0.910932
## loudness          0.00073667765  0.00187472247   0.393              0.694425
## mode              0.02473399569  0.00787711257   3.140              0.001732 **
## speechiness       0.22901124742  0.03489359287   6.563     0.00000000007880 ***
## acousticness     -0.17969926770  0.02573062073  -6.984     0.00000000000479 ***
## instrumentalness  0.05122649254  0.01438546434   3.561              0.000384 ***
## liveness         -0.08287464988  0.02302142381  -3.600              0.000332 ***
## valence           0.25066498302  0.01778148896  14.097 < 0.0000000000000002 ***
## tempo            -0.00105652206  0.00016670058  -6.338     0.00000000033122 ***
## duration_ms       0.00000005384  0.00000004631   1.163              0.245210
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1311 on 1177 degrees of freedom
## Multiple R-squared:  0.3219, Adjusted R-squared:  0.3155
## F-statistic: 50.78 on 11 and 1177 DF,  p-value: < 0.00000000000000022
```

```
summary(reg.mod.2)
```

```
##
## Call:
## lm(formula = danceability ~ energy + key + loudness + mode +
##     speechiness + acousticness + instrumentalness + liveness +
##     valence + tempo + duration_ms)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48835 -0.08074  0.01301  0.08802  0.40456
##
## Coefficients:
##                       Estimate     Std. Error t value             Pr(>|t|)
## (Intercept)       0.95306627023  0.04545204873  20.969 < 0.0000000000000002 ***
## energy           -0.33592989741  0.03400603879  -9.879 < 0.0000000000000002 ***
## key               0.00038855231  0.00108914396   0.357              0.721342
## loudness          0.00188699529  0.00189143954   0.998              0.318652
## mode              0.02588598445  0.00805070804   3.215              0.001338 **
## speechiness       0.22190586098  0.03461007632   6.412    0.000000000207326 ***
## acousticness     -0.16252899931  0.02582980638  -6.292    0.000000000438990 ***
## instrumentalness  0.04785497552  0.01460535369   3.277              0.001081 **
## liveness         -0.08126345089  0.02357538608  -3.447              0.000587 ***
## valence           0.24269132355  0.01811264319  13.399 < 0.0000000000000002 ***
## tempo            -0.00122340625  0.00016951097  -7.217    0.000000000000944 ***
## duration_ms       0.00000006707  0.00000004598   1.459              0.144867
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1347 on 1188 degrees of freedom
## Multiple R-squared:  0.3035, Adjusted R-squared:  0.2971
## F-statistic: 47.07 on 11 and 1188 DF,  p-value: < 0.00000000000000022
```

After removing the 11 points we listed above, our adjusted R^2 value improved from 0.2971 to .3155 (our R^2 values also improved, but adjusted takes into account multiple predictors.) The standard errors, as a whole, are smaller for reg.mod3, and the predictors are more statistically signifcant.

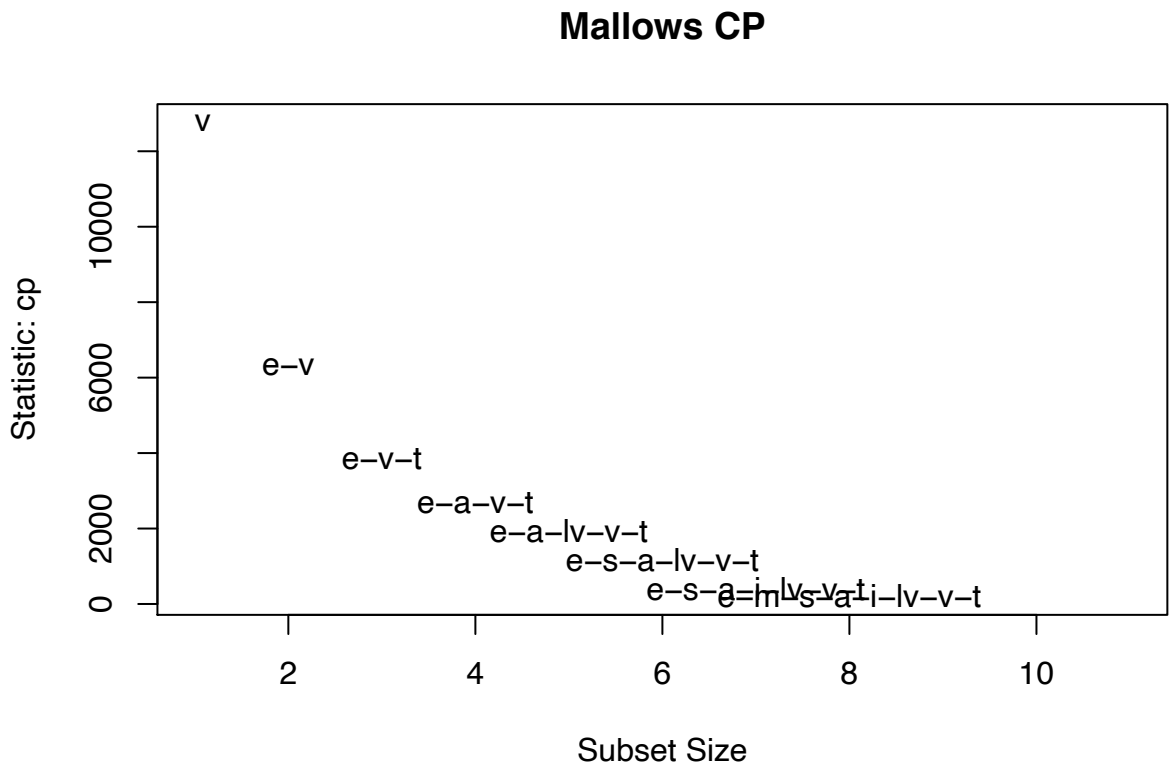
Question 4

Use Mallows Cp for identifying which terms you will keep in the model (based on part 3) and also use the Boruta algorithm for variable selection. Based on the two results, determine which subset of predictors you will keep.

```
library(AER)
library(leaps)

ols_mallows_cp(reg.mod.3, reg.mod.2)
```

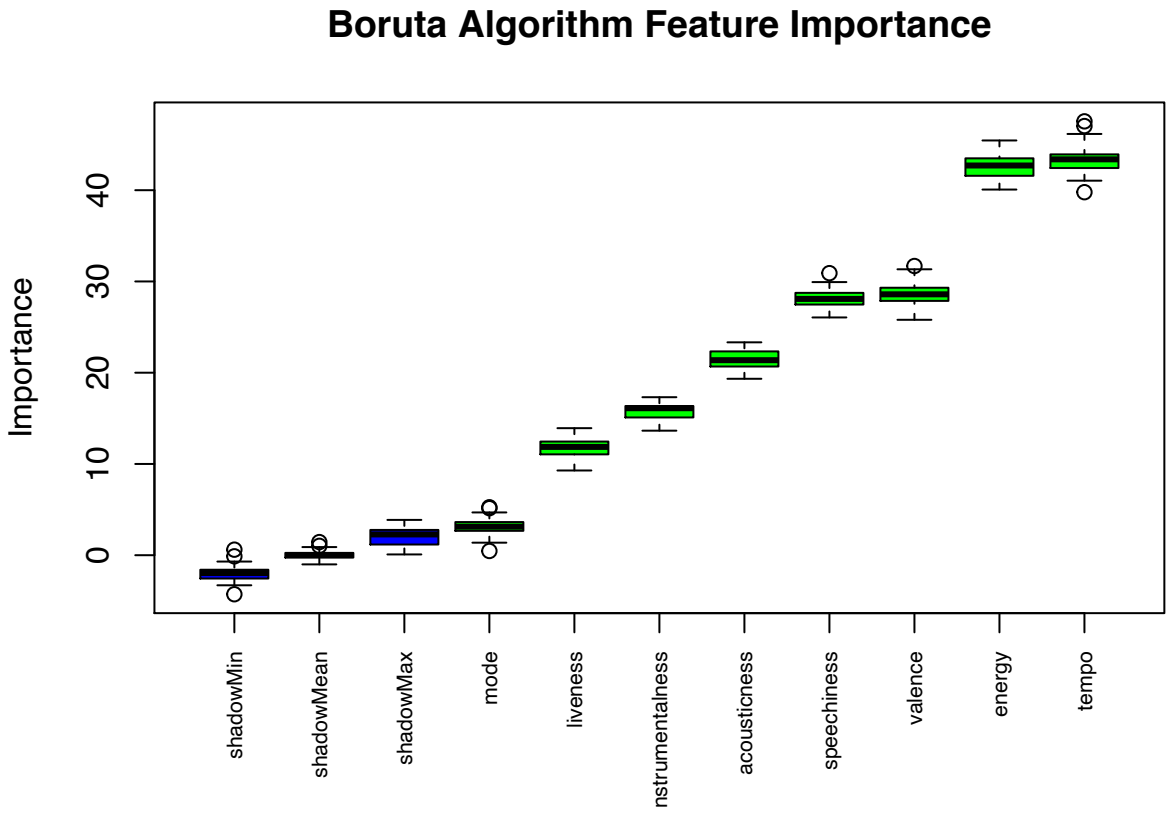
```
## [1] -51.1761
```

```
ss = regsubsets(danceability ~ energy + key + loudness + mode + speechiness +
                  acousticness + instrumentalness + liveness + valence + tempo
                + duration_ms, method = c("exhaustive"), nbest=1, data=genres)

subsets(ss, statistic = "cp", legend = F, main="Mallows CP", col="steelblue4")
```



```
##                  Abbreviation
## energy                      e
## key                         k
## loudness                   ld
## mode                        m
## speechiness                 s
## acousticness                a
## instrumentalness            i
## liveness                   lv
## valence                     v
## tempo                       t
## duration_ms                 d
```

```
library(Boruta)

Bor.res <- Boruta(danceability ~ energy + mode + speechiness + acousticness +
                    instrumentalness + liveness + valence + tempo,
                  data = genres2, doTrace = 2)

# Best model is: energy + mode + speechiness + acousticness +
# instrumentalness + liveness + valence + tempo

plot(Bor.res, xlab = "", xaxt = "n", main="Boruta Algorithm Feature Importance")

lz<-lapply(1:ncol(Bor.res$ImpHistory),function(i)
Bor.res$ImpHistory[is.finite(Bor.res$ImpHistory[,i]),i])
names(lz) <- colnames(Bor.res$ImpHistory)
Labels <- sort(sapply(lz,median))
axis(side = 1,las=2,labels = names(Labels),
at = 1:ncol(Bor.res$ImpHistory), cex.axis = 0.7)
```




```
boruta_signif <- names(Bor.res$finalDecision[Bor.res$finalDecision %in%
                                               c("Confirmed", "Tentative")])
boruta_signif_Conf <- names(Bor.res$finalDecision[Bor.res$finalDecision %in%
                                                    c("Confirmed")])
boruta_signif_Tent <- names(Bor.res$finalDecision[Bor.res$finalDecision %in%
                                                    c("Tentative")])
boruta_signif_Reject <- names(Bor.res$finalDecision[Bor.res$finalDecision %in%
                                                      c("Rejected")])
# energy, mode, speechiness, acousticness..
#..instrumentalness, liveness, valence, tempo
print(boruta_signif_Conf)
```

```
## [1] "energy"           "mode"             "speechiness"      "acousticness"
## [5] "instrumentalness" "liveness"         "valence"          "tempo"
```

```
print(boruta_signif_Tent) #  character(0)
```

```
## character(0)
```

```
print(boruta_signif_Reject) # character(0)
```

```
## character(0)
```

```r
# Look at the statistical attributes in terms of 'variable importance'
# Sort variables by importance (most imp first and least imp last)
sorted_vars = attStats(Bor.res)[order(-attStats(Bor.res)$meanImp),]
print(sorted_vars)
```

```
##                   meanImp medianImp    minImp    maxImp  normHits  decision
## tempo            43.435623 43.382761 39.780843 47.545245 1.0000000 Confirmed
## energy           42.663326 42.706379 40.072094 45.451747 1.0000000 Confirmed
## valence          28.594483 28.587479 25.798190 31.691219 1.0000000 Confirmed
## speechiness      28.094201 28.087053 26.055369 30.909128 1.0000000 Confirmed
## acousticness     21.408843 21.371541 19.334745 23.326949 1.0000000 Confirmed
## instrumentalness 15.769071 16.107302 13.646810 17.314831 1.0000000 Confirmed
## liveness         11.835323 11.865917  9.282023 13.919909 1.0000000 Confirmed
## mode              3.129152  3.127731  0.464100  5.251813 0.7714286 Confirmed
```

```r
# Store variables in a variable (e.g., top 5 variables)
conf_vars = row.names(sorted_vars[1:5,])
print(conf_vars)
```

```
## [1] "tempo"        "energy"       "valence"      "speechiness"  "acousticness"
```

The subset of predictors we will keep, based on the results of the Mallows CP and Boruta functions, include: energy, mode, speechiness, acousticness, instrumentalness, liveness, valence, and tempo.

We have removed key and loudness from the set of predictors.

Question 5

Test for multicollinearity using VIF on the model from (4). Based on the test, remove any appropriate variables, and estimate a new regression model based on these findings.

```r
#from Part 4 --> remove key and loudness
reg.mod.4 <- lm(danceability ~ energy + mode + speechiness + acousticness +
                  instrumentalness + liveness + valence + tempo, subset= c(-50,
                    -518, -530, -576, -91, -555, -334, -645, -600, -722, -925))

summary(reg.mod.4)
```

```
##
## Call:
## lm(formula = danceability ~ energy + mode + speechiness + acousticness +
##     instrumentalness + liveness + valence + tempo, subset = c(-50,
##     -518, -530, -576, -91, -555, -334, -645, -600, -722, -925))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39693 -0.08195  0.01095  0.08691  0.40207
##
## Coefficients:
##                   Estimate Std. Error t value             Pr(>|t|)
```

```
## (Intercept)       0.9287438  0.0321028  28.930 < 0.0000000000000002 ***
## energy           -0.3264775  0.0255286 -12.789 < 0.0000000000000002 ***
## mode              0.0243957  0.0077280   3.157             0.001636 **
## speechiness       0.2237375  0.0345496   6.476     0.00000000013806 ***
## acousticness     -0.1798509  0.0256297  -7.017     0.00000000000381 ***
## instrumentalness  0.0567165  0.0120954   4.689     0.00000306417321 ***
## liveness         -0.0827273  0.0229830  -3.600             0.000332 ***
## valence           0.2497781  0.0176831  14.125 < 0.0000000000000002 ***
## tempo            -0.0010611  0.0001639  -6.476     0.00000000013831 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.131 on 1180 degrees of freedom
## Multiple R-squared:  0.321,  Adjusted R-squared:  0.3164
## F-statistic: 69.74 on 8 and 1180 DF,  p-value: < 0.00000000000000022
```

```
tidy(vif(reg.mod.4))
```

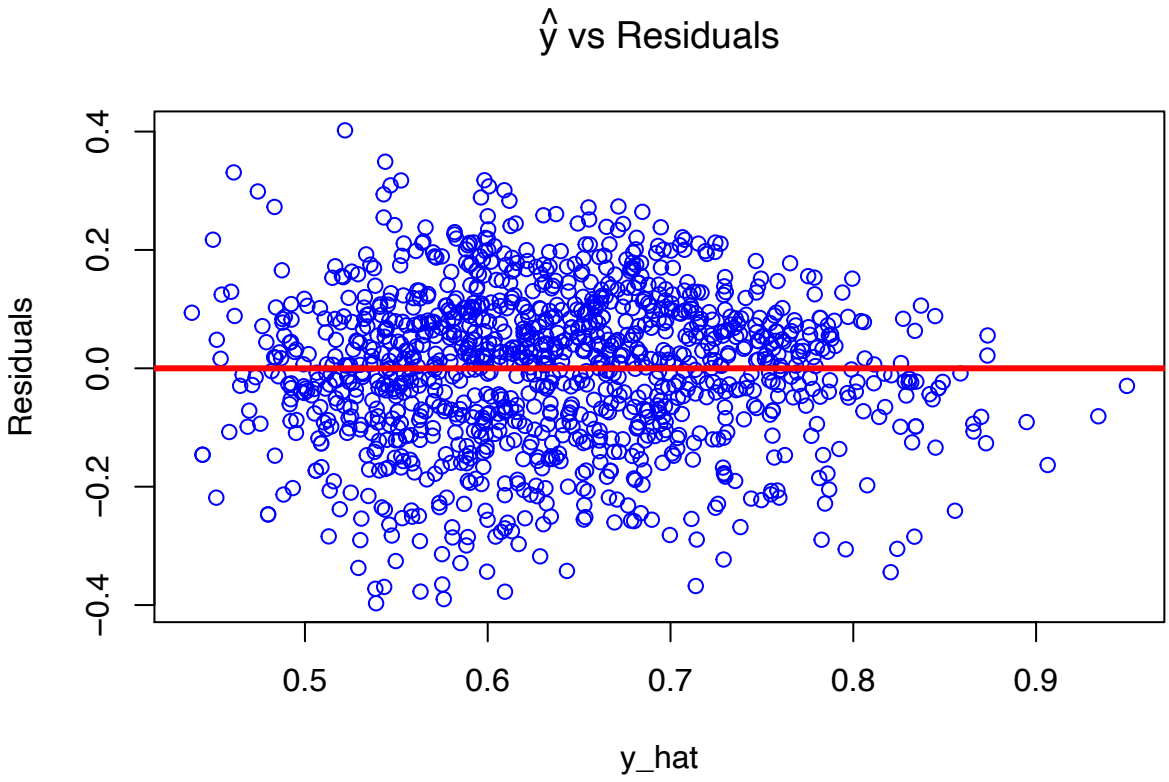
```
## # A tibble: 8 x 2
##   names                x
##   <chr>            <dbl>
## 1 energy            1.57
## 2 mode              1.02
## 3 speechiness       1.23
## 4 acousticness      1.44
## 5 instrumentalness  1.37
## 6 liveness          1.08
## 7 valence           1.12
## 8 tempo             1.05
```

Based on the VIF scores above, since all the CIF values are less than 4, we do not remove any variables. So, there is no need for reestimation.

Question 6

For your model in part (5) plot the respective residuals vs. $\hat{y}$ and comment on your results.

```
#NOTE: No need to change the model from part 4.
library(latex2exp)

y_hat <- predict(reg.mod.4)
plot(y_hat, reg.mod.4$residuals, col="blue", ylab = "Residuals",
     main=TeX('$\\hat{y}$ vs Residuals'))
abline(h=0, col="red", lw=3)
```



Based on the graph above, there is a slight skewness toward the negative side and to the right, but it almost looks like it is centered around 0.

Question 7

For your model in part (5) perform a RESET test and comment on your results.

```
library(lmtest)
# Test if a quadratic model is appropriate  --> does not need
resettest(reg.mod.4 , power=2, type="regressor")
```

```
##
##  RESET test
##
## data:  reg.mod.4
## RESET = 8.0013, df1 = 8, df2 = 1172, p-value = 0.0000000001515
```

According to the RESET Test, since we have a very low p-value, we reject the null and therefore we ignore the quadratic model. This means that we do not need a quadratic model, our model is sufficient.

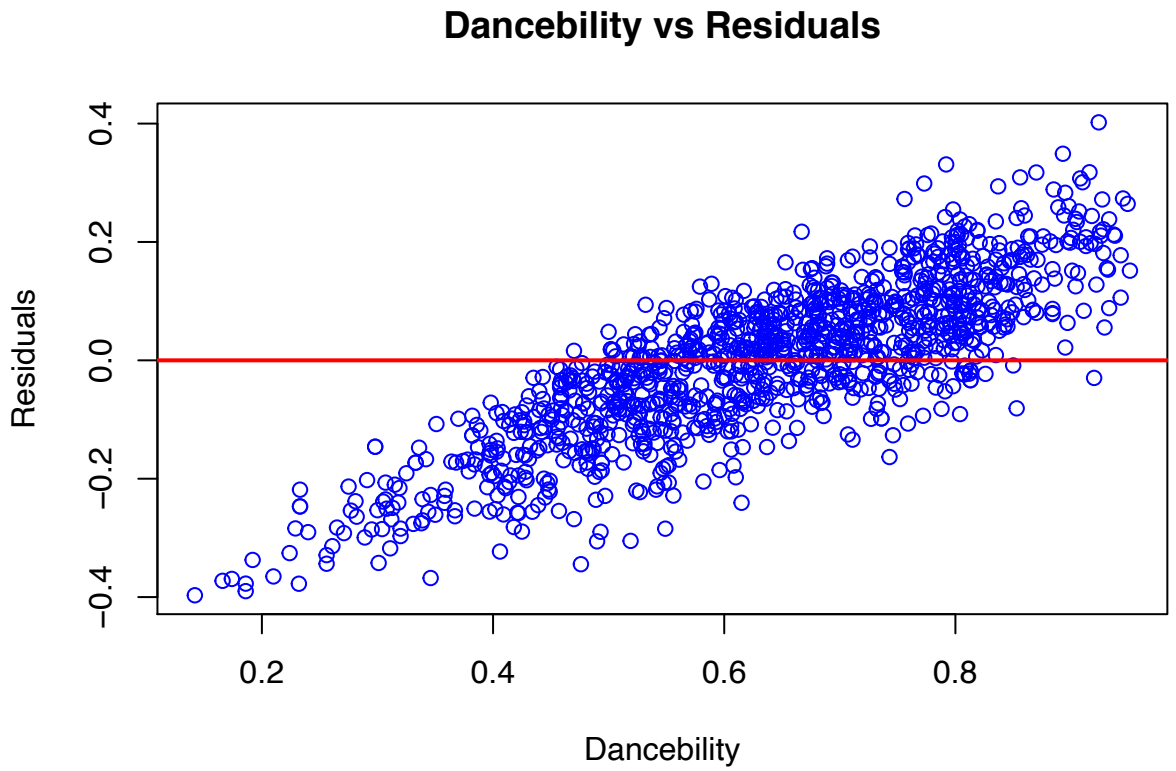
Question 8

For your model in part (5) test for heteroskedasticity and comment on your results. If you identify heteroskedasticy, make sure to account for it before moving on to (9).

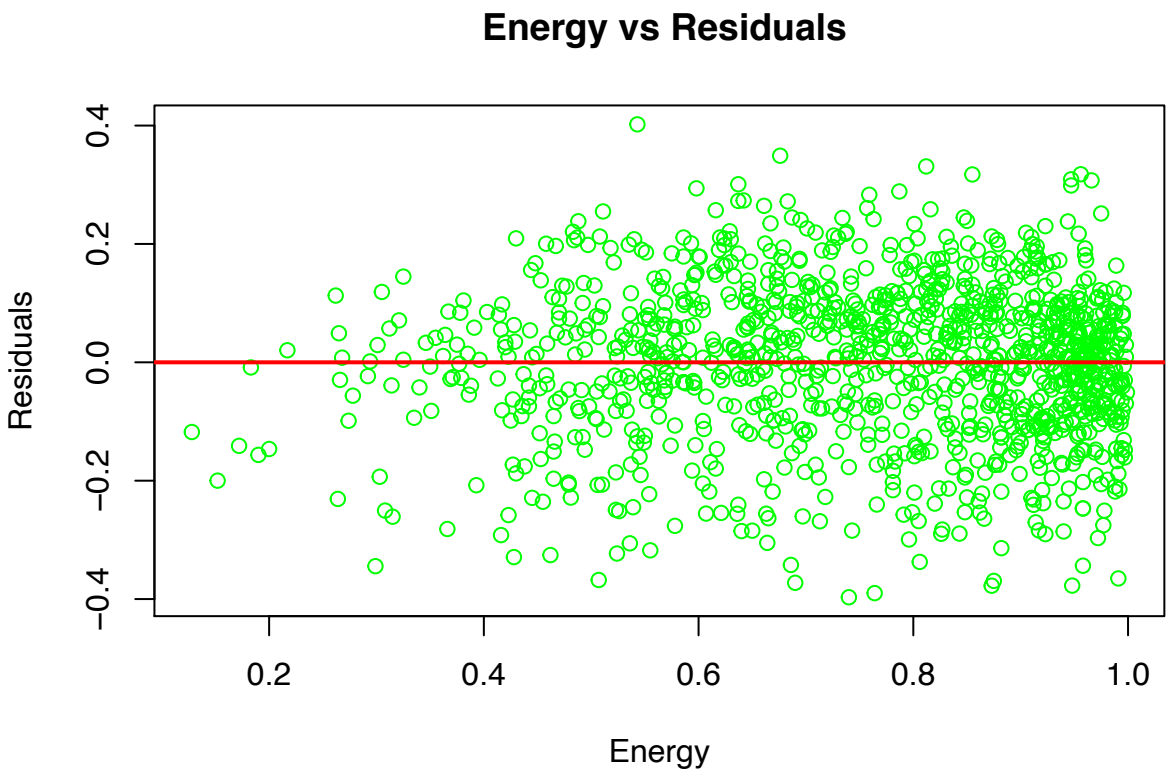
explanation:

If the errors are homoskedastic, there should be no patterns of any sort in the residuals. On the other hand, if the errors are heteroskedastic, they may tend to exhibit greater variation in some systematic way. We can see it by plotting the residuals either against one of the explanatory variables or against $\hat{y}_i$ to see if they vary in a systematic way.

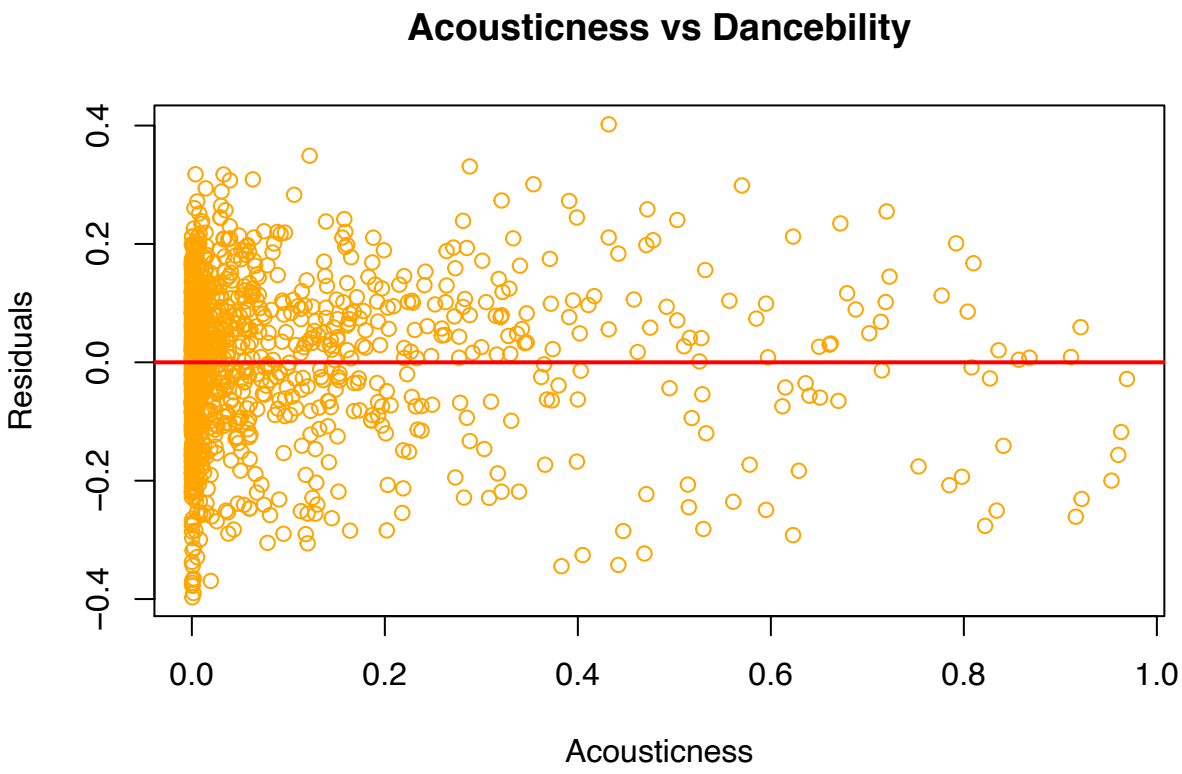
```
#Plot the Residuals to see if there is heteroskedasticity
genres3 <- genres2[-c(50, 518, 530, 576, 91, 555, 334, 645, 600, 722, 925), ]

plot(genres3$danceability, reg.mod.4$residuals, xlab="Dancebility",
     ylab= "Residuals" , main= "Dancebility vs Residuals", col = "blue")
abline(h=0, col="red", lw=2)
```



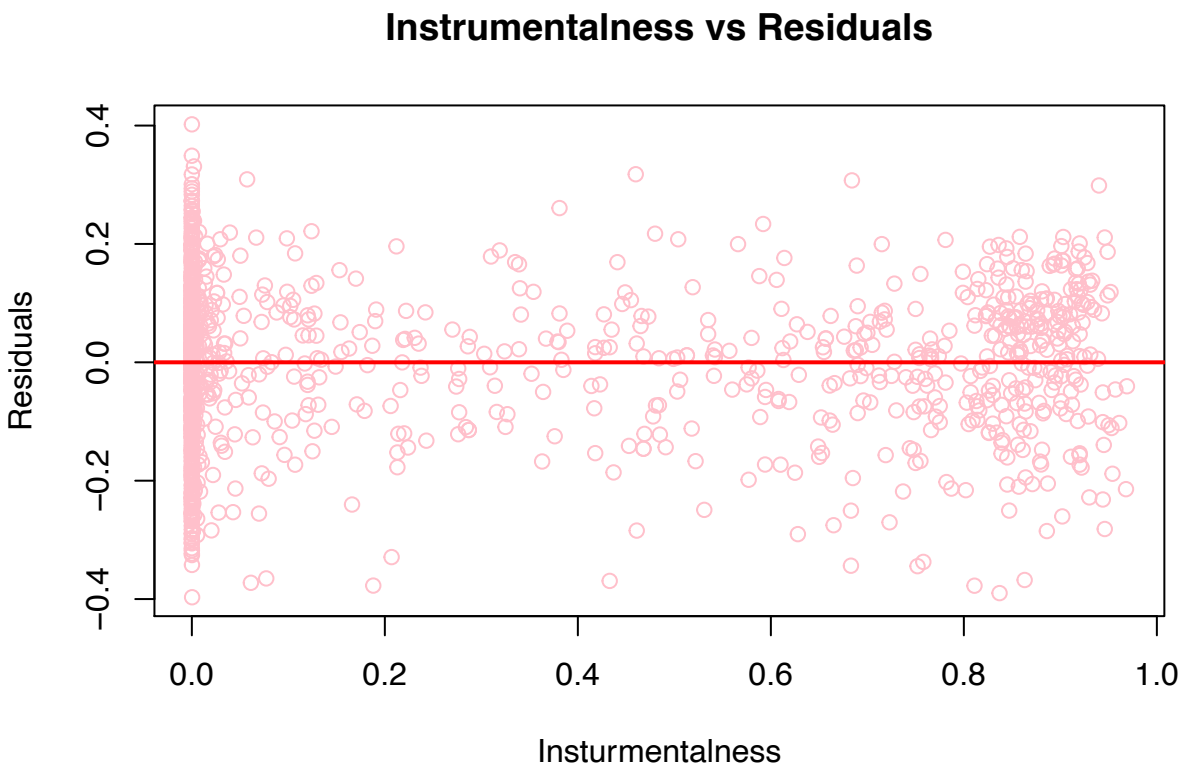

```
plot(genres3$energy, reg.mod.4$residuals, xlab="Energy" , ylab= "Residuals" ,
     main= " Energy vs Residuals", col="green")
abline(h=0, col="red", lw=2)
```



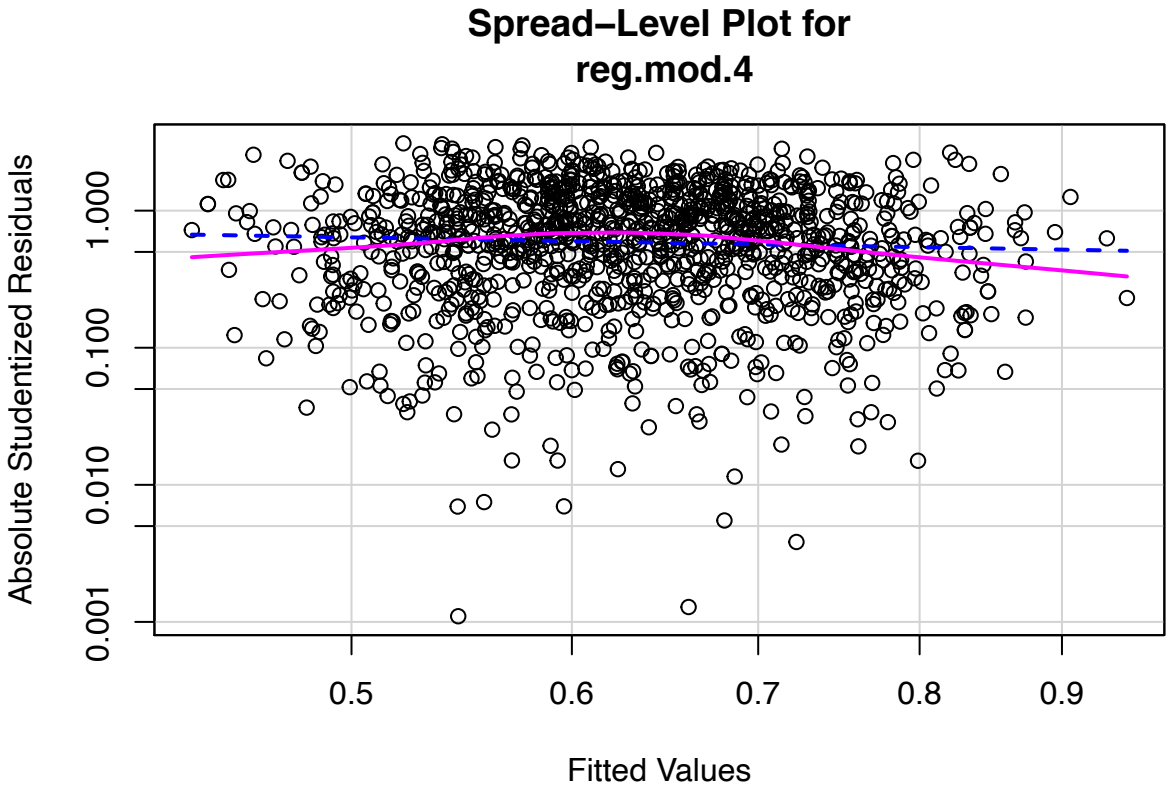

```
plot(genres3$acousticness, reg.mod.4$residuals, xlab="Acousticness" ,
     ylab= "Residuals", main= "Acousticness vs Dancebility", col = "orange")
abline(h=0, col="red", lw=2)
```




```
plot(genres3$instrumentalness,reg.mod.4$residuals, xlab="Insturmentalness",
     ylab="Residuals", main="Instrumentalness vs Residuals", col="pink")
abline(h=0, col="red", lw=2)
```




```
spreadLevelPlot(reg.mod.4)
```




```
##
## Suggested power transformation:  1.350409
```

```
ncvTest(reg.mod.4) # Reject H0
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 8.966593, Df = 1, p = 0.0027496
```

```
bptest(reg.mod.4) # Reject H0
```

```
##
##  studentized Breusch-Pagan test
##
## data:  reg.mod.4
## BP = 50.027, df = 8, p-value = 0.00000004037
```

Looking at the graph of dancebility, energy, acousticness, and instumentalness vs residuals, we can easily see that heteroskedasticity is present. Both the ncvTest abd BP test suggest that we reject the null (because of the low p value), meaning that there is evidence of heteroskedasticity.

Question 9

Estimate a model based on all your findings that also includes interaction terms (if appropriate) and if needed, any higher power terms. Comment on the performance of this model compared to your other models. Make sure to use AIC and BIC for model comparison.

```
# correcting for heteroskedasticity (from part 8):
cov1 <- hccm(reg.mod.4, type="hc1")
mod.HC1 <- coeftest(reg.mod.4, vcov.=cov1)

mod.HC1 # white standard error model
```

```
##
## t test of coefficients:
##
##                    Estimate Std. Error  t value              Pr(>|t|)
## (Intercept)       0.9287438  0.0338087  27.4705 < 0.00000000000000022 ***
## energy           -0.3264775  0.0246364 -13.2518 < 0.00000000000000022 ***
## mode              0.0243957  0.0077619   3.1430              0.001714 **
## speechiness       0.2237375  0.0361562   6.1881       0.0000000008386 ***
## acousticness     -0.1798509  0.0295474  -6.0869       0.0000000015546 ***
## instrumentalness  0.0567165  0.0119371   4.7513       0.0000022697296 ***
## liveness         -0.0827273  0.0211663  -3.9084       0.0000981777807 ***
## valence           0.2497781  0.0164224  15.2096 < 0.00000000000000022 ***
## tempo            -0.0010611  0.0001940  -5.4697       0.0000000549854 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(reg.mod.4) # old model
```

```
##
## Call:
## lm(formula = danceability ~ energy + mode + speechiness + acousticness +
##     instrumentalness + liveness + valence + tempo, subset = c(-50,
##     -518, -530, -576, -91, -555, -334, -645, -600, -722, -925))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39693 -0.08195  0.01095  0.08691  0.40207
##
## Coefficients:
##                    Estimate Std. Error t value             Pr(>|t|)
## (Intercept)       0.9287438  0.0321028  28.930 < 0.0000000000000002 ***
## energy           -0.3264775  0.0255286 -12.789 < 0.0000000000000002 ***
## mode              0.0243957  0.0077280   3.157             0.001636 **
## speechiness       0.2237375  0.0345496   6.476     0.00000000013806 ***
## acousticness     -0.1798509  0.0256297  -7.017     0.00000000000381 ***
## instrumentalness  0.0567165  0.0120954   4.689     0.00000306417321 ***
## liveness         -0.0827273  0.0229830  -3.600             0.000332 ***
## valence           0.2497781  0.0176831  14.125 < 0.0000000000000002 ***
## tempo            -0.0010611  0.0001639  -6.476     0.00000000013831 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.131 on 1180 degrees of freedom
## Multiple R-squared:  0.321,  Adjusted R-squared:  0.3164
## F-statistic: 69.74 on 8 and 1180 DF,  p-value: < 0.00000000000000022
```

```
# fgls model
ehatsq <- resid(reg.mod.4)^2
sighatsq.ols <- lm(log(ehatsq)~log(danceability), data=genres3)
vari <- exp(fitted(sighatsq.ols))
mod.fgls <- lm(danceability ~ energy + speechiness + acousticness +
                instrumentalness + liveness + valence + tempo,
               weights = 1/vari, data = genres3)
summary(mod.fgls)
```

```
##
## Call:
## lm(formula = danceability ~ energy + speechiness + acousticness +
##     instrumentalness + liveness + valence + tempo, data = genres3,
##     weights = 1/vari)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3543 -1.5644 -0.3075  0.9078  6.4349
##
## Coefficients:
##                   Estimate Std. Error t value             Pr(>|t|)
## (Intercept)       0.9396479  0.0281754  33.350 < 0.0000000000000002 ***
## energy           -0.2931780  0.0233258 -12.569 < 0.0000000000000002 ***
## speechiness       0.2094718  0.0312063   6.712      0.0000000000296 ***
## acousticness     -0.1435191  0.0236595  -6.066      0.0000000017628 ***
## instrumentalness  0.0403814  0.0114058   3.540             0.000415 ***
## liveness         -0.0993956  0.0218124  -4.557      0.0000057319008 ***
## valence           0.2063509  0.0159081  12.971 < 0.0000000000000002 ***
## tempo            -0.0008347  0.0001486  -5.618      0.0000000240798 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.804 on 1181 degrees of freedom
## Multiple R-squared:  0.2966, Adjusted R-squared:  0.2925
## F-statistic: 71.15 on 7 and 1181 DF,  p-value: < 0.00000000000000022
```

```
AIC(reg.mod.2, reg.mod.3, reg.mod.4, mod.HC1, mod.fgls)
```

```
##           df       AIC
## reg.mod.2 13 -1391.333
## reg.mod.3 13 -1444.049
## reg.mod.4 10 -1448.616
## mod.HC1   10 -1448.616
## mod.fgls   9 -1580.949
```

```
BIC(reg.mod.2, reg.mod.3, reg.mod.4, mod.HC1, mod.fgls)
```

```
##           df       BIC
## reg.mod.2 13 -1325.162
## reg.mod.3 13 -1377.998
## reg.mod.4 10 -1397.807
## mod.HC1   10 -1397.807
## mod.fgls   9 -1535.221
```

Both the AIC and BIC test tells us that the fgls model (which is corrected for heteroskedasticity) is the best because it is the lowest value by a large margin. Comparing the standard errors of the fgls model to the other models (even the model with the white standard errors), we see that it has the smallest standard errors. While reg.mod 2 has predictors that are not all significant, the predictors for mod.fgls are all significant.

Question 10

Evaluate your model performance (from 9) using cross-validation, and also by dividing your data into the traditional 2/3 training and 1/3 testing samples, to evaluate your out-of-sample performance. Comment on your results.

```
library(lmvar)

fit= lm(danceability ~ energy + speechiness + acousticness +
          instrumentalness + liveness + valence + tempo,
        weights = 1/vari, data = genres3, x = TRUE, y = TRUE)

cv.lm(fit, k = 3)
```

```
## Mean absolute error        :  0.1048076
## Sample standard deviation  :  0.003447216
##
## Mean squared error         :  0.01742306
## Sample standard deviation  :  0.001307108
##
## Root mean squared error    :  0.1319337
## Sample standard deviation  :  0.00498295
```

Based on the output above, $RMSE = 0.13$ which is a little higher than expected since we were aiming for the range between 0.01 and 0.08.

Question 11

Provide a short (1 paragraph) summary of your overall conclusions findings.

In our initial multiple regression model (reg.mod.2), we found that 3 of the 10 predictor variables were not significant at a 5% level. Following that, we found that 11 data points had leverage, were outliers, or were influential, so we removed those points for our reg.mod.3. Using Mallows CP, Boruta, and VIF tests, we removed the 2 predictor variables of key and loudness for our reg.mod.4. The RESET test told us that we did not need higher order terms. Looking at residual graphs, we can see that heteroskedasticity is present in our model, so we used White standard errors for our final model (mod.HC1) and the FGLS test (mod.fgls). Using AIC and BIC tests to compare our 4 models, we found that "mod.fgls" was the most accurate as it had

the lowest AIC and BIC scores at, -1580.949 and -1535.221, respectively. It also had the smallest standard errors. When testing using cross-validation, we found that the RMSE = 0.13, which was slightly higher than expected, but we believe this was still the best model possible.