outlier_analiza korelacji

Krzysztof Kurek

27 02 2020

```
dane <- read.csv("../train_set.csv", stringsAsFactors = FALSE, na.strings="")
dane$X <- NULL

library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(reshape2)
library(zoo)
```

```
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##
##     date
```

```r
library(ggplot2)


dane_numeryczne <- dane %>%
  group_by(playlist_genre) %>%
  select_if(is.numeric)

dane_numeryczne1 <- dane_numeryczne[,2:14]
```

#KORELACJE ##Tabela korelacji

```r
korelacje_liczbowo <- round(cor(dane_numeryczne1),2); korelacje_liczbowo
```

```
##                   track_popularity danceability energy   key loudness  mode
## track_popularity              1.00         0.07  -0.11  0.00     0.06  0.01
## danceability                  0.07         1.00  -0.09  0.01     0.02 -0.06
## energy                       -0.11        -0.09   1.00  0.01     0.68  0.00
## key                           0.00         0.01   0.01  1.00     0.00 -0.18
## loudness                      0.06         0.02   0.68  0.00     1.00 -0.02
## mode                          0.01        -0.06   0.00 -0.18    -0.02  1.00
## speechiness                   0.01         0.18  -0.03  0.03     0.01 -0.06
## acousticness                  0.08        -0.03  -0.54  0.01    -0.36  0.01
## instrumentalness             -0.15         0.00   0.03  0.01    -0.15 -0.01
## liveness                     -0.05        -0.13   0.17  0.00     0.08 -0.01
## valence                       0.04         0.33   0.15  0.02     0.05  0.00
## tempo                         0.00        -0.18   0.15 -0.01     0.09  0.01
## duration_ms                  -0.15        -0.10   0.01  0.02    -0.11  0.01
##                   speechiness acousticness instrumentalness liveness valence
## track_popularity         0.01         0.08            -0.15    -0.05    0.04
## danceability             0.18        -0.03             0.00    -0.13    0.33
## energy                  -0.03        -0.54             0.03     0.17    0.15
## key                      0.03         0.01             0.01     0.00    0.02
## loudness                 0.01        -0.36            -0.15     0.08    0.05
## mode                    -0.06         0.01            -0.01    -0.01    0.00
## speechiness              1.00         0.03            -0.10     0.05    0.07
## acousticness             0.03         1.00            -0.01    -0.08   -0.02
## instrumentalness        -0.10        -0.01             1.00    -0.01   -0.17
## liveness                 0.05        -0.08            -0.01     1.00   -0.02
## valence                  0.07        -0.02            -0.17    -0.02    1.00
## tempo                    0.05        -0.12             0.02     0.02   -0.03
## duration_ms             -0.09        -0.08             0.06     0.01   -0.03
##                   tempo duration_ms
## track_popularity   0.00       -0.15
## danceability      -0.18       -0.10
## energy             0.15        0.01
## key               -0.01        0.02
## loudness           0.09       -0.11
## mode               0.01        0.01
## speechiness        0.05       -0.09
## acousticness      -0.12       -0.08
## instrumentalness   0.02        0.06
## liveness           0.02        0.01
## valence           -0.03       -0.03
```

```
## tempo             1.00        0.00
## duration_ms       0.00        1.00
```

Modyfikacja tabeli korelacji - usunięcie wartości powyżej 1

```
upper <- korelacje_liczbowo
upper[upper.tri(korelacje_liczbowo)] <- ""
upper <- as.data.frame(upper); upper
```

```
##                  track_popularity danceability energy   key loudness  mode
## track_popularity                1
## danceability                 0.07            1
## energy                      -0.11        -0.09      1
## key                             0         0.01   0.01     1
## loudness                     0.06         0.02   0.68     0        1
## mode                         0.01        -0.06      0 -0.18    -0.02     1
## speechiness                  0.01         0.18  -0.03  0.03     0.01 -0.06
## acousticness                 0.08        -0.03  -0.54  0.01    -0.36  0.01
## instrumentalness            -0.15            0   0.03  0.01    -0.15 -0.01
## liveness                    -0.05        -0.13   0.17     0     0.08 -0.01
## valence                      0.04         0.33   0.15  0.02     0.05     0
## tempo                           0        -0.18   0.15 -0.01     0.09  0.01
## duration_ms                 -0.15         -0.1   0.01  0.02    -0.11  0.01
##                  speechiness acousticness instrumentalness liveness valence
## track_popularity
## danceability
## energy
## key
## loudness
## mode
## speechiness                1
## acousticness            0.03            1
## instrumentalness        -0.1        -0.01                1
## liveness                0.05        -0.08            -0.01        1
## valence                 0.07        -0.02            -0.17    -0.02       1
## tempo                   0.05        -0.12             0.02     0.02   -0.03
## duration_ms            -0.09        -0.08             0.06     0.01   -0.03
##                  tempo duration_ms
## track_popularity
## danceability
## energy
## key
## loudness
## mode
## speechiness
## acousticness
## instrumentalness
## liveness
## valence
## tempo                1
## duration_ms          0           1
```

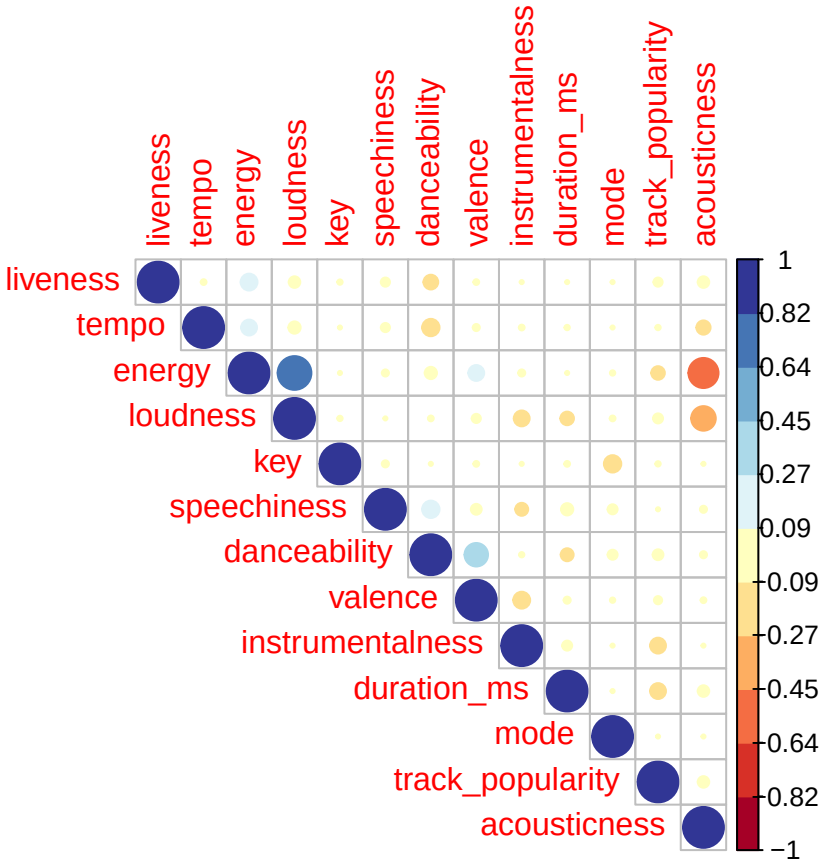
Korelacje - wykres graficzny

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(RColorBrewer)
corplot1 <- corrplot(korelacje_liczbowo, type = "upper", order ="hclust",
         col=brewer.pal(n=13, name ="RdYlBu")); corplot1
```

```
## Warning in brewer.pal(n = 13, name = "RdYlBu"): n too large, allowed maximum for palette RdYlBu is 1
## Returning the palette you asked for with that many colors
```



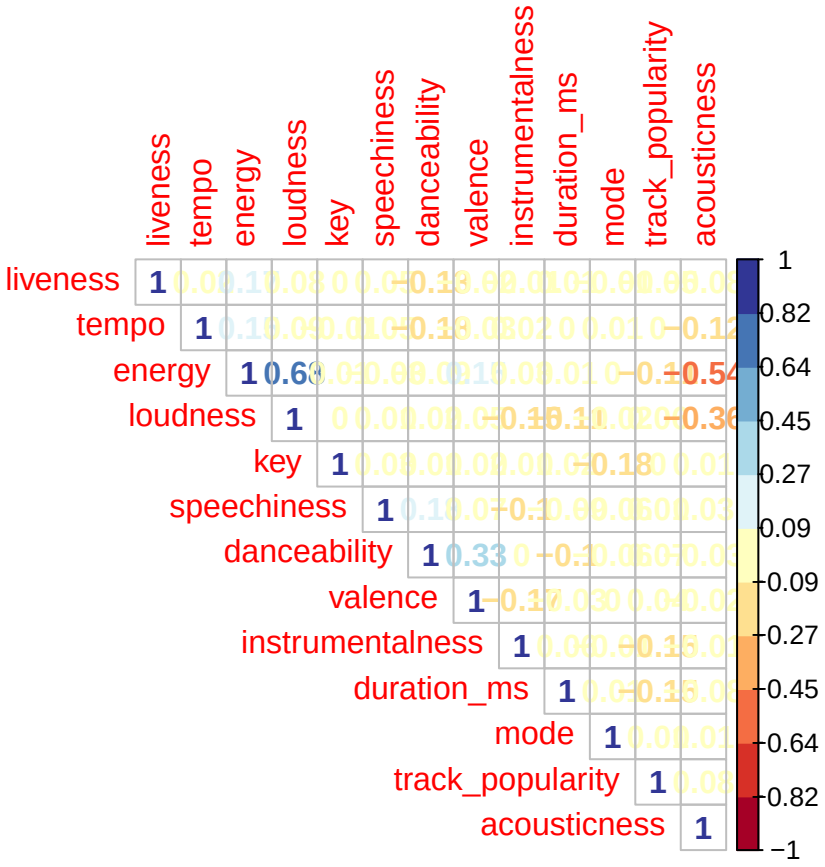

```
##                  liveness tempo energy loudness   key speechiness danceability
## liveness             1.00  0.02   0.17     0.08  0.00        0.05        -0.13
## tempo                0.02  1.00   0.15     0.09 -0.01        0.05        -0.18
## energy               0.17  0.15   1.00     0.68  0.01       -0.03        -0.09
## loudness             0.08  0.09   0.68     1.00  0.00        0.01         0.02
## key                  0.00 -0.01   0.01     0.00  1.00        0.03         0.01
## speechiness          0.05  0.05  -0.03     0.01  0.03        1.00         0.18
## danceability        -0.13 -0.18  -0.09     0.02  0.01        0.18         1.00
## valence             -0.02 -0.03   0.15     0.05  0.02        0.07         0.33
## instrumentalness    -0.01  0.02   0.03    -0.15  0.01       -0.10         0.00
```

```
## duration_ms          0.01  0.00   0.01    -0.11  0.02       -0.09        -0.10
## mode                -0.01  0.01   0.00    -0.02 -0.18       -0.06        -0.06
## track_popularity    -0.05  0.00  -0.11     0.06  0.00        0.01         0.07
## acousticness        -0.08 -0.12  -0.54    -0.36  0.01        0.03        -0.03
##                   valence instrumentalness duration_ms  mode track_popularity
## liveness            -0.02            -0.01        0.01 -0.01            -0.05
## tempo               -0.03             0.02        0.00  0.01             0.00
## energy               0.15             0.03        0.01  0.00            -0.11
## loudness             0.05            -0.15       -0.11 -0.02             0.06
## key                  0.02             0.01        0.02 -0.18             0.00
## speechiness          0.07            -0.10       -0.09 -0.06             0.01
## danceability         0.33             0.00       -0.10 -0.06             0.07
## valence              1.00            -0.17       -0.03  0.00             0.04
## instrumentalness    -0.17             1.00        0.06 -0.01            -0.15
## duration_ms         -0.03             0.06        1.00  0.01            -0.15
## mode                 0.00            -0.01        0.01  1.00             0.01
## track_popularity     0.04            -0.15       -0.15  0.01             1.00
## acousticness        -0.02            -0.01       -0.08  0.01             0.08
##                   acousticness
## liveness                -0.08
## tempo                   -0.12
## energy                  -0.54
## loudness                -0.36
## key                      0.01
## speechiness              0.03
## danceability            -0.03
## valence                 -0.02
## instrumentalness        -0.01
## duration_ms             -0.08
## mode                     0.01
## track_popularity         0.08
## acousticness             1.00
```

Korelacje - wykres numeryczny

```
corrplot(korelacje_liczbowo,method = "number" ,type = "upper", order ="hclust",
         col=brewer.pal(n=13, name ="RdYlBu"))
```

```
## Warning in brewer.pal(n = 13, name = "RdYlBu"): n too large, allowed maximum for palette RdYlBu is 1
## Returning the palette you asked for with that many colors
```



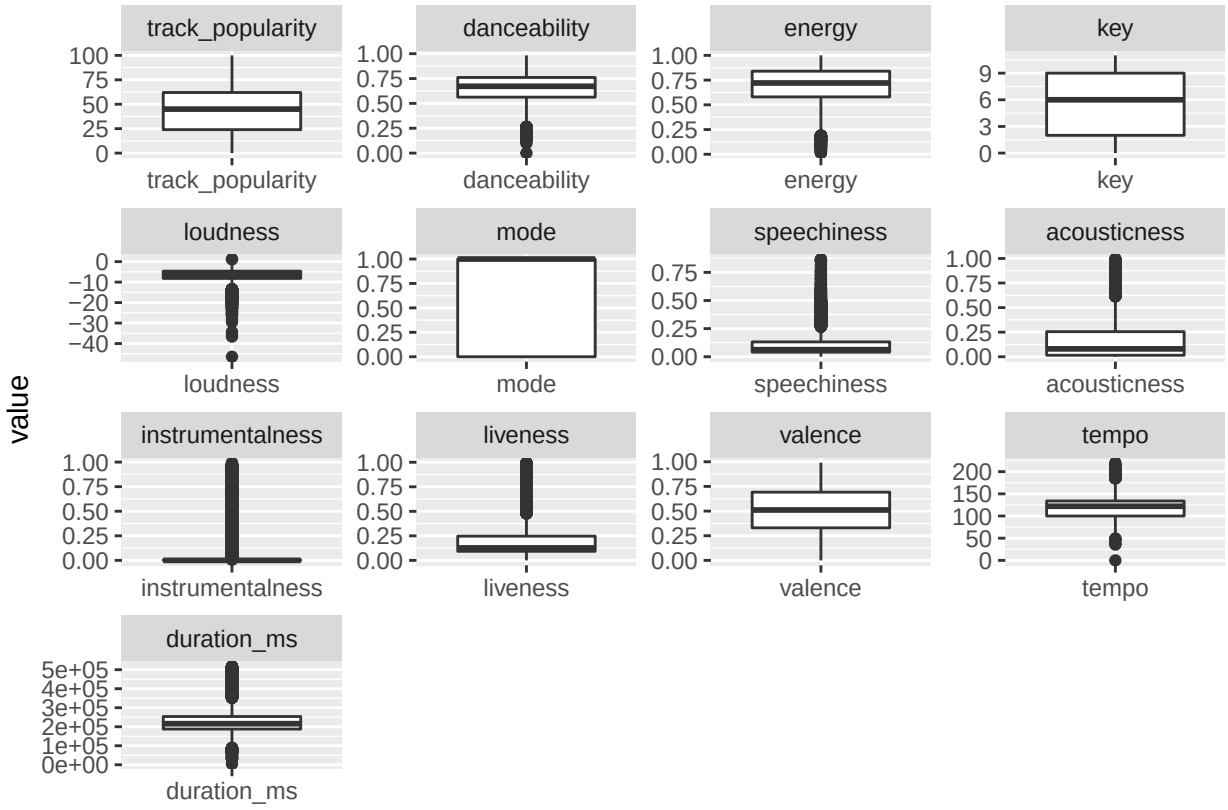
OUTLIERS

wykres boxplot dla każdej zmiennej

```
melt_dane_numeryczne <-melt(dane_numeryczne1)
```

```
## No id variables; using all as measure variables
```

```
ggplot(melt_dane_numeryczne, aes(variable, value))+ geom_boxplot()+
  facet_wrap(~variable, scales = "free")
```



Wyliczenie IQR

```
iqr <- melt_dane_numeryczne %>%
  group_by(variable) %>%
  summarise( IQR= IQR(value));iqr
```

```
## # A tibble: 13 x 2
##    variable                IQR
##    <fct>                 <dbl>
##  1 track_popularity    38
##  2 danceability         0.198
##  3 energy               0.26
##  4 key                  7
##  5 loudness             3.55
##  6 mode                 1
##  7 speechiness          0.091
##  8 acousticness         0.24
##  9 instrumentalness     0.00502
## 10 liveness             0.154
## 11 valence              0.362
## 12 tempo               33.9
## 13 duration_ms      65928.
```

Wyliczenie dolnego przedziału Q - 1.5IQR

```
dolny_przedzial <- melt_dane_numeryczne %>%
  group_by(variable) %>%
  summarise( lower_IQR = quantile(value, 0.25) - 1.5*IQR(value)); dolny_przedzial
```

```
## # A tibble: 13 x 2
##    variable          lower_IQR
##    <fct>                 <dbl>
##  1 track_popularity   -33
##  2 danceability         0.266
##  3 energy               0.190
##  4 key                 -8.5
##  5 loudness           -13.5
##  6 mode                -1.5
##  7 speechiness         -0.0955
##  8 acousticness        -0.345
##  9 instrumentalness    -0.00753
## 10 liveness            -0.139
## 11 valence             -0.213
## 12 tempo               49.0
## 13 duration_ms      88912.
```

```
# Złączenie ponowne w forme wide
library(tidyr)
```

```
##
## Attaching package: 'tidyr'
```

```
## The following object is masked from 'package:reshape2':
##
##     smiths
```

```
dolny_przedzial <-dolny_przedzial %>%
  spread(variable, lower_IQR)
```

Wyliczenie górnego przedziału

```
gorny_przedzial <- as.data.frame(melt_dane_numeryczne %>%
  group_by(variable) %>%
  summarise( lower_IQR = quantile(value, 0.75) + 1.5*IQR(value)));gorny_przedzial
```

```
##            variable    lower_IQR
## 1  track_popularity    119.00000
## 2      danceability      1.05800
## 3            energy      1.23000
## 4               key     19.50000
## 5          loudness      0.67650
```

```
## 6              mode      2.50000
## 7       speechiness      0.26850
## 8      acousticness      0.61500
## 9  instrumentalness      0.01255
## 10         liveness      0.47860
## 11          valence      1.23500
## 12            tempo    184.81200
## 13      duration_ms 352625.37500
```

```
# Złączenie ponowne w forme wide
gorny_przedzial <- gorny_przedzial %>%
  spread(variable, lower_IQR)
```

Usunięcie wartości outlier

```
dane_bez_outlier <- dane_numeryczne1 %>%
  filter(track_popularity >= dolny_przedzial$track_popularity & track_popularity <= gorny_przedzial$tra
         danceability>= dolny_przedzial$danceability & danceability <= gorny_przedzial$danceability,
        energy>= dolny_przedzial$energy & energy <= gorny_przedzial$energy,
        key >= dolny_przedzial$key & key <= gorny_przedzial$key,
        mode >= dolny_przedzial$mode & mode <= gorny_przedzial$mode,
        speechiness >= dolny_przedzial$speechiness & speechiness <= gorny_przedzial$speechiness,
        acousticness >= dolny_przedzial$acousticness & acousticness <= gorny_przedzial$acousticness,
        instrumentalness >= dolny_przedzial$instrumentalness & instrumentalness <= gorny_przedzial$inst
        liveness >= dolny_przedzial$liveness & liveness <= gorny_przedzial$liveness,
        acousticness >= dolny_przedzial$acousticness & acousticness <= gorny_przedzial$acousticness,
        valence >= dolny_przedzial$valence & valence <= gorny_przedzial$valence,
        tempo >= dolny_przedzial$tempo & tempo <= gorny_przedzial$tempo,
        valence >= dolny_przedzial$valence & valence <= gorny_przedzial$valence,
        duration_ms >= dolny_przedzial$duration_ms & duration_ms <= gorny_przedzial$duration_ms)
```

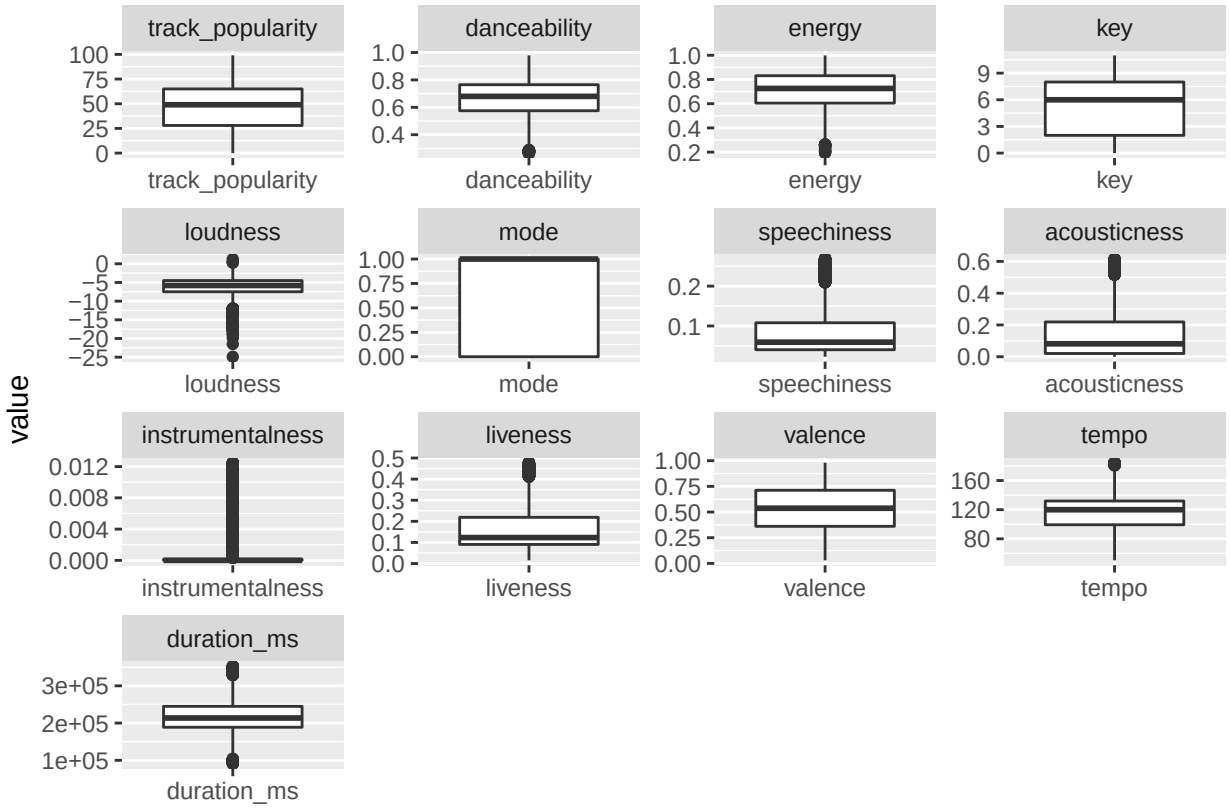
Usunięcie wartości odstających (outlier) spowodowało usunięcie 10 608 rekordów, co stanowiło 40% całości zbioru.

Wykres boxplot dla kązdej zmiennej

```
melt_bez_outlier <-melt(dane_bez_outlier)
```

```
## No id variables; using all as measure variables
```

```
ggplot(melt_bez_outlier, aes(variable, value))+ geom_boxplot()+
  facet_wrap(~variable, scales = "free")
```



Wykres graficzny numeryczny korelacji bez outlier

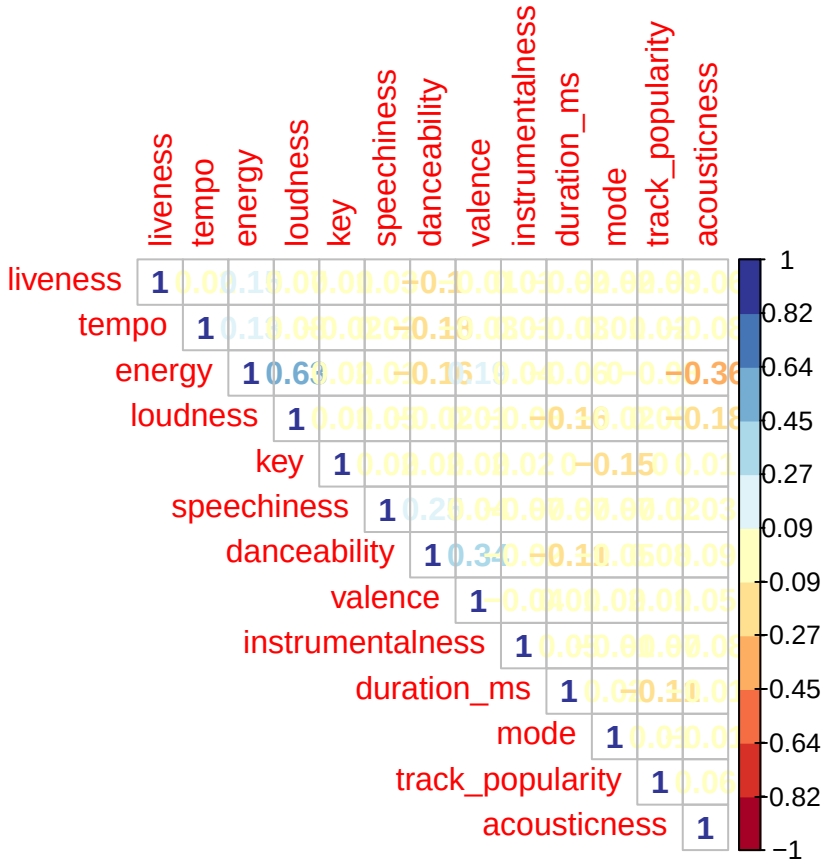
```
korelacje_bez_outlier <- round(cor(dane_bez_outlier),2); korelacje_bez_outlier
```

```
##                   track_popularity danceability energy   key loudness  mode
## track_popularity              1.00         0.08  -0.08  0.00     0.08  0.01
## danceability                  0.08         1.00  -0.16  0.02    -0.02 -0.05
## energy                       -0.08        -0.16   1.00  0.02     0.63  0.00
## key                           0.00         0.02   0.02  1.00     0.01 -0.15
## loudness                      0.08        -0.02   0.63  0.01     1.00 -0.02
## mode                          0.01        -0.05   0.00 -0.15    -0.02  1.00
## speechiness                  -0.02         0.25   0.01  0.02     0.05 -0.07
## acousticness                  0.06         0.09  -0.36  0.01    -0.18 -0.01
## instrumentalness             -0.07        -0.02   0.04  0.02    -0.08 -0.01
## liveness                     -0.03        -0.10   0.15  0.01     0.07 -0.02
## valence                       0.01         0.34   0.19  0.02     0.01  0.02
## tempo                         0.02        -0.18   0.13 -0.02     0.08  0.01
## duration_ms                  -0.11        -0.11  -0.06  0.00    -0.16  0.02
##                   speechiness acousticness instrumentalness liveness valence
## track_popularity        -0.02         0.06            -0.07    -0.03    0.01
## danceability             0.25         0.09            -0.02    -0.10    0.34
## energy                   0.01        -0.36             0.04     0.15    0.19
## key                      0.02         0.01             0.02     0.01    0.02
## loudness                 0.05        -0.18            -0.08     0.07    0.01
## mode                    -0.07        -0.01            -0.01    -0.02    0.02
## speechiness              1.00         0.03            -0.07     0.03    0.04
```

```
## acousticness            0.03        1.00           -0.08    -0.06    0.05
## instrumentalness       -0.07       -0.08            1.00     0.01   -0.04
## liveness                0.03       -0.06            0.01     1.00   -0.01
## valence                 0.04        0.05           -0.04    -0.01    1.00
## tempo                   0.02       -0.08            0.01     0.02   -0.03
## duration_ms            -0.07       -0.01            0.05    -0.02    0.01
##                    tempo duration_ms
## track_popularity    0.02       -0.11
## danceability       -0.18       -0.11
## energy              0.13       -0.06
## key                -0.02        0.00
## loudness            0.08       -0.16
## mode                0.01        0.02
## speechiness         0.02       -0.07
## acousticness       -0.08       -0.01
## instrumentalness    0.01        0.05
## liveness            0.02       -0.02
## valence            -0.03        0.01
## tempo               1.00       -0.03
## duration_ms        -0.03        1.00
```

```
corrplot(korelacje_bez_outlier,method = "number" ,type = "upper", order ="hclust",
         col=brewer.pal(n=13, name ="RdYlBu"))
```

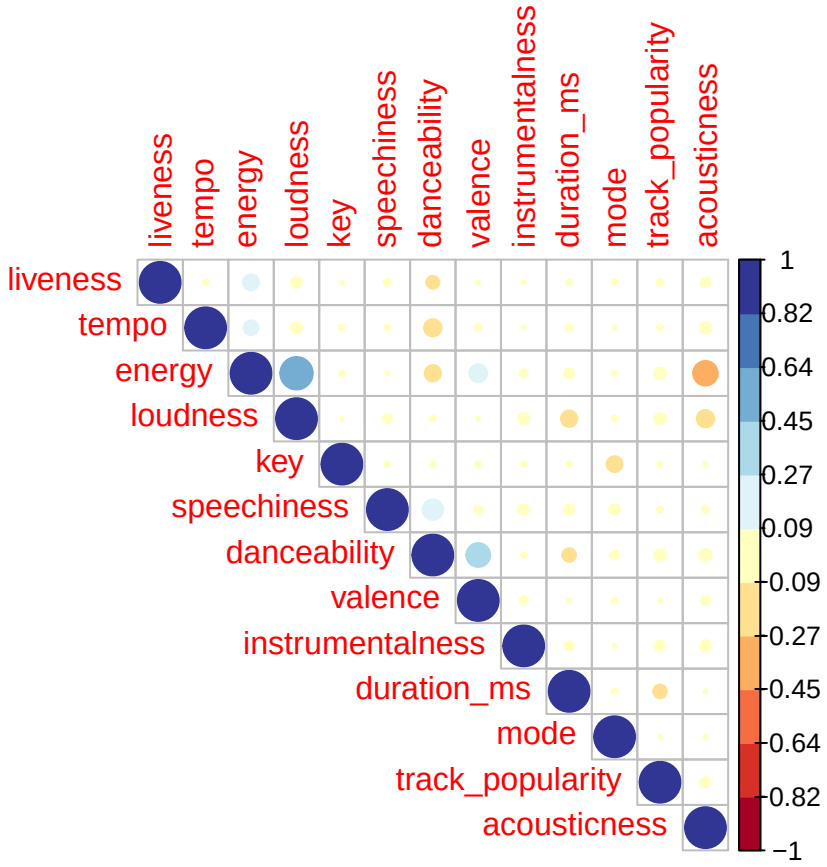
```
## Warning in brewer.pal(n = 13, name = "RdYlBu"): n too large, allowed maximum for palette RdYlBu is 1
## Returning the palette you asked for with that many colors
```



Wykres graficzny kolreacji bez outlier

```
corplot2 <- corrplot(korelacje_bez_outlier, type = "upper", order ="hclust",
         col=brewer.pal(n=13, name ="RdYlBu"));corplot2
```

```
## Warning in brewer.pal(n = 13, name = "RdYlBu"): n too large, allowed maximum for palette RdYlBu is 1
## Returning the palette you asked for with that many colors
```




```
##                  liveness tempo energy loudness   key speechiness danceability
## liveness             1.00  0.02   0.15     0.07  0.01        0.03        -0.10
## tempo                0.02  1.00   0.13     0.08 -0.02        0.02        -0.18
## energy               0.15  0.13   1.00     0.63  0.02        0.01        -0.16
## loudness             0.07  0.08   0.63     1.00  0.01        0.05        -0.02
## key                  0.01 -0.02   0.02     0.01  1.00        0.02         0.02
## speechiness          0.03  0.02   0.01     0.05  0.02        1.00         0.25
## danceability        -0.10 -0.18  -0.16    -0.02  0.02        0.25         1.00
## valence             -0.01 -0.03   0.19     0.01  0.02        0.04         0.34
## instrumentalness     0.01  0.01   0.04    -0.08  0.02       -0.07        -0.02
## duration_ms         -0.02 -0.03  -0.06    -0.16  0.00       -0.07        -0.11
## mode                -0.02  0.01   0.00    -0.02 -0.15       -0.07        -0.05
## track_popularity    -0.03  0.02  -0.08     0.08  0.00       -0.02         0.08
## acousticness        -0.06 -0.08  -0.36    -0.18  0.01        0.03         0.09
##                  valence instrumentalness duration_ms  mode track_popularity
## liveness           -0.01             0.01       -0.02 -0.02            -0.03
## tempo              -0.03             0.01       -0.03  0.01             0.02
```

```
## energy              0.19            0.04       -0.06  0.00            -0.08
## loudness            0.01           -0.08       -0.16 -0.02             0.08
## key                 0.02            0.02        0.00 -0.15             0.00
## speechiness         0.04           -0.07       -0.07 -0.07            -0.02
## danceability        0.34           -0.02       -0.11 -0.05             0.08
## valence             1.00           -0.04        0.01  0.02             0.01
## instrumentalness   -0.04            1.00        0.05 -0.01            -0.07
## duration_ms         0.01            0.05        1.00  0.02            -0.11
## mode                0.02           -0.01        0.02  1.00             0.01
## track_popularity    0.01           -0.07       -0.11  0.01             1.00
## acousticness        0.05           -0.08       -0.01 -0.01             0.06
##                  acousticness
## liveness               -0.06
## tempo                  -0.08
## energy                 -0.36
## loudness               -0.18
## key                     0.01
## speechiness             0.03
## danceability            0.09
## valence                 0.05
## instrumentalness       -0.08
## duration_ms            -0.01
## mode                   -0.01
## track_popularity        0.06
## acousticness            1.00
```