

# Analiza cech piosenek na Spotify

**407616, Bartłomiej Chwast**, poniedziałek 16<sup>15</sup>

*AGH, Wydział Informatyki Elektroniki i Telekomunikacji  
Rachunek prawdopodobieństwa i statystyka 2021/2022*

Kraków, January 20, 2022

*Ja, niżej podpisany własnoręcznym podpisem deklaruję, że przygotowałem przedstawiony do oceny projekt samodzielnie i żadna jego część nie jest kopią pracy innej osoby.*

.....

## 1 Streszczenie raportu

Raport powstał, w oparciu o analizę danych dotyczących cech prawie 19 000 utworów spośród dostępnych na platformie Spotify. Cechy te zostały zmierzone przez API Spotify (więcej informacji o sposobie mierzenia na stronie <https://developer.spotify.com/documentation/web-api/reference/#/operations/get-audio-features>)

## 2 Opis danych

Dane do projektu pochodzą... ze strony <https://www.kaggle.com/ektanegi/spotifydata-19212020>. Dotyczą one cech prawie 19 000 utworów na Spotify. Zmienne analizowane w projekcie:

- danceability - jak bardzo utwór nadaje się do tańczenia od 0 do 1
- energy - jak bardzo energetyczny jest utwór od 0 do 1
- loudness - ogólna głośność utworu od -60dB do 0dB
- tempo - tempo utworu w BPM
- valence - jak bardzo pozytywnie brzmi utwór od 0 do 1

## 3 Analiza danych

### 3.1 Przygotowanie danych do analizy

W projekcie wykorzystałem następujące biblioteki:

```
> library(ggplot2)
> library(ggcormrplot)
> library(moments)
> library(nortest)
> library(grid)
> library(gridExtra)
> library(hrbrthemes)
```

Pracę rozpoczęłem od wczytania danych do R, po czym wybrałem kolumny do analizy i zapisałem je do osobnego pliku.

```
> filePath <- "data.csv"
> data <- read.csv(filePath)
> dataSet <- as.data.frame(cbind(data$danceability, data$energy, data$loudness,
+                                     data$tempo, data$audio_valence))
> colnames(dataSet) [1] <- "danceability"
> colnames(dataSet) [2] <- "energy"
> colnames(dataSet) [3] <- "loudness"
> colnames(dataSet) [4] <- "tempo"
> colnames(dataSet) [5] <- "valence"
> attach(dataSet)
> write.csv(dataSet, "./dataSet.csv", row.names = TRUE)
```

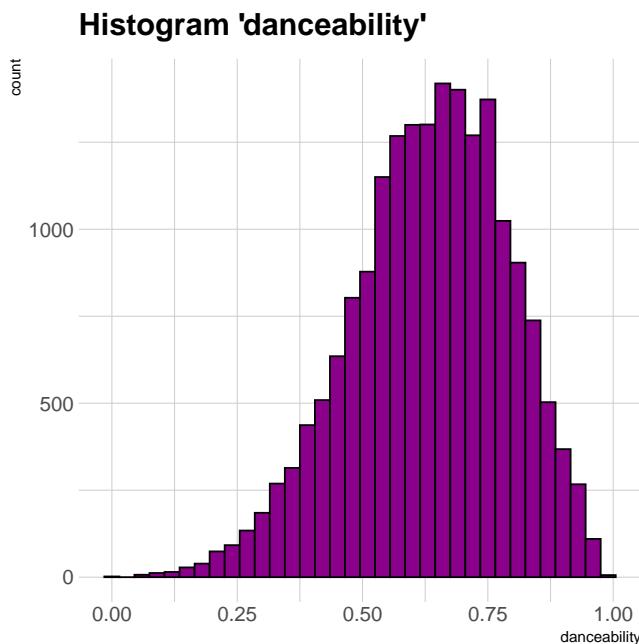
### 3.2 Wydobywanie podstawowych informacji z danych

Dla każdej ze zmiennych wyznaczyłem przedział wartości, średnią, kwartyle, wariancję, kurtozę i współczynnik asymetrii. Kolejno dla każdej wygenerowałem histogram i boxplot. Następnie wygenerowałem wykres QQ oraz wykonalem test Lillieforsa w celu sprawdzenia czy dane pochodzą z rozkładu normalnego.

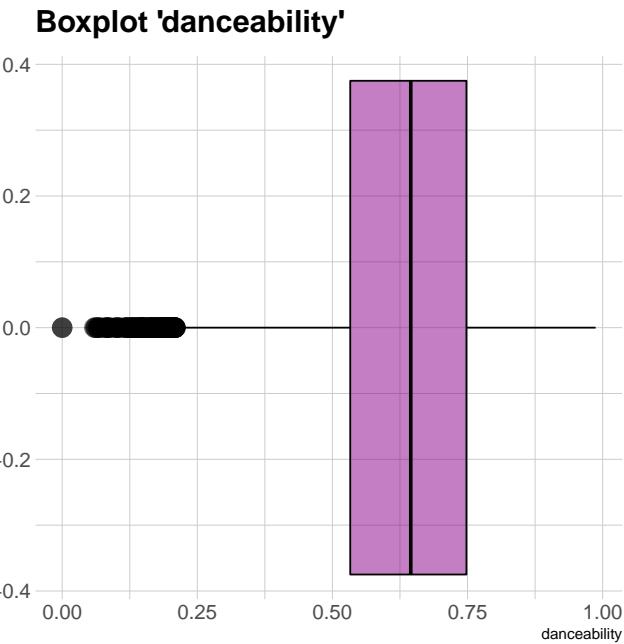
### 3.3 Zmienna 'danceability'

Na początek zbadałem zmienną 'danceability'.

```
> histogram <- ggplot(dataSet, aes(x = danceability), xlab="danceability") +  
+   geom_histogram(binwidth=0.03, color="black", fill="darkmagenta") +  
+   theme_ipsum(base_family = 'sans') +  
+   labs(title="Histogram 'danceability'")  
> histogram
```



```
> boxplot <- ggplot(dataSet, aes(x = danceability), xlab="danceability") +  
+   geom_boxplot(color="black", fill="darkmagenta", alpha=0.5, outlier.size=5) +  
+   theme_ipsum(base_family = 'sans') +  
+   labs(title="Boxplot 'danceability'")  
> boxplot
```



```

> range(danceability)
[1] 0.000 0.987
> mean(danceability)
[1] 0.6333481
> quantile(danceability, c(0.25, 0.5, 0.75))
 25%   50%   75%
0.533 0.645 0.748
> var(danceability)
[1] 0.02456201
> sd(danceability)
[1] 0.1567227
> skewness(danceability)
[1] -0.3916879

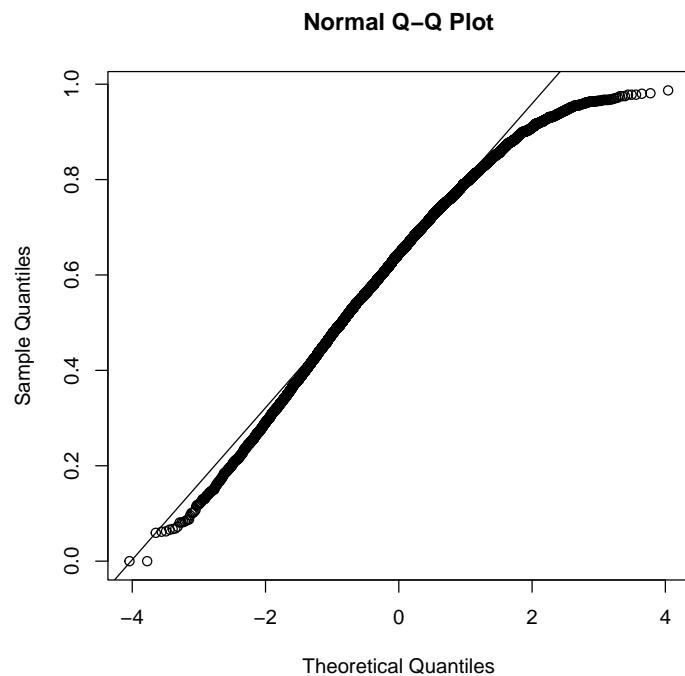
```

```
> kurtosis(danceability)
```

```
[1] 2.924905
```

Wartości zmiennej sięgają od 0.000 do 0.987, ich średnia to 0.633, wariancja to 0.0246, a odchylenie standardowe to 0.157. Współczynnik asymetrii jest ujemny, zatem rozkład jest lewostronnie skośny. Współczynnik wyostrzenia również jest ujemny, więc rozkład jest platykurtyczny.

```
> qqnorm(danceability)
> qqline(danceability)
```



```
> lillie.test(danceability)
```

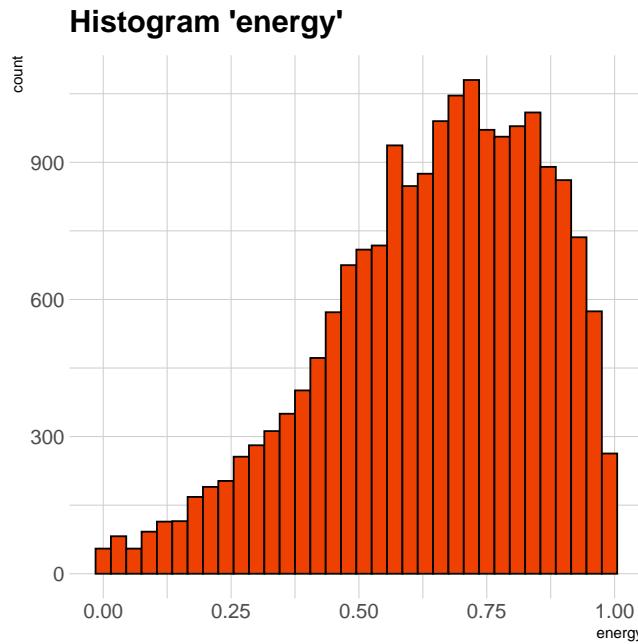
```
Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data: danceability
D = 0.031847, p-value < 2.2e-16
```

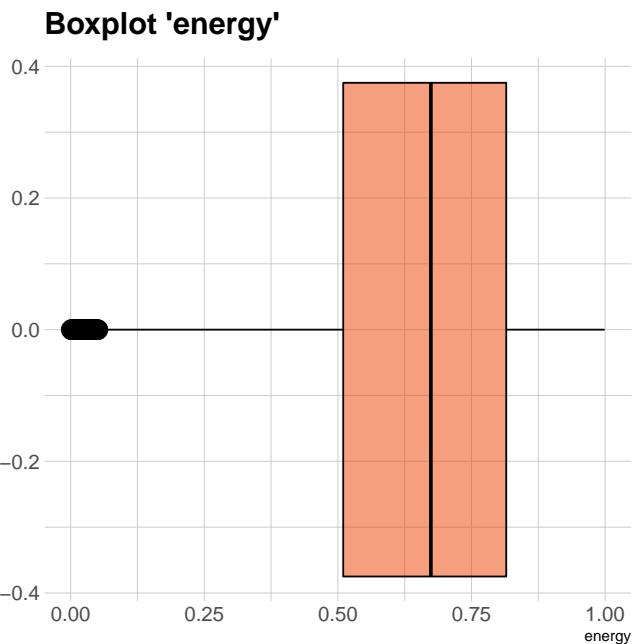
Zarówno wykres QQ, jak i test Lillieforsa wskazują, że rozkład zmiennej 'danceability' jest daleki od rozkładu normalnego, kształt histogramu jednakże przypomina krzywą dzwonową

### 3.4 Zmienna 'energy'

```
> histogram <- ggplot(dataSet, aes(x = energy), xlab="energy") +  
+   geom_histogram(binwidth=0.03, color="black", fill="orangered2") +  
+   theme_ipsum(base_family = 'sans') +  
+   labs(title="Histogram 'energy'")  
> histogram
```



```
> boxplot <- ggplot(dataSet, aes(x = energy), xlab="energy") +  
+   geom_boxplot(color="black", fill="orangered2", alpha=0.5, outlier.size=5) +  
+   theme_ipsum(base_family = 'sans') +  
+   labs(title="Boxplot 'energy'")  
> boxplot
```



```

> range(energy)
[1] 0.00107 0.99900

> mean(energy)
[1] 0.6449948

> quantile(energy, c(0.25, 0.5, 0.75))
 25%   50%   75%
0.510 0.674 0.815

> var(energy)
[1] 0.04583913

> sd(energy)
[1] 0.2141008

> skewness(energy)
[1] -0.6206881

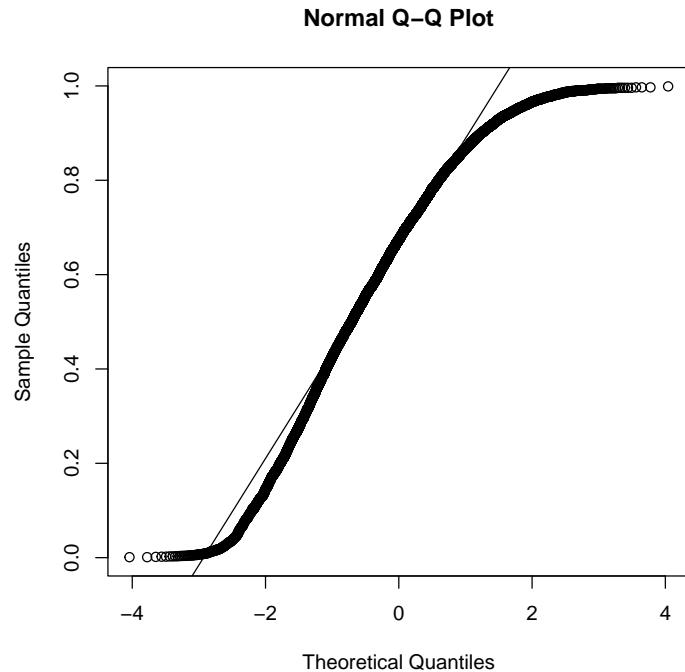
```

```
> kurtosis(energy)
```

```
[1] 2.861844
```

Wartości zmiennej sięgają od 0.00107 do 0.999, ich średnia to 0.645, wariancja to 0.0458, a odchylenie standardowe to 0.214. Współczynnik asymetrii jest ujemny, zatem rozkład jest lewostronnie skośny. Współczynnik wyostrzenia również jest ujemny, więc rozkład jest platykurtyczny.

```
> qqnorm(energy)
> qqline(energy)
```



```
> lilliefors.test(energy)
```

```
Lilliefors (Kolmogorov-Smirnov) normality test
```

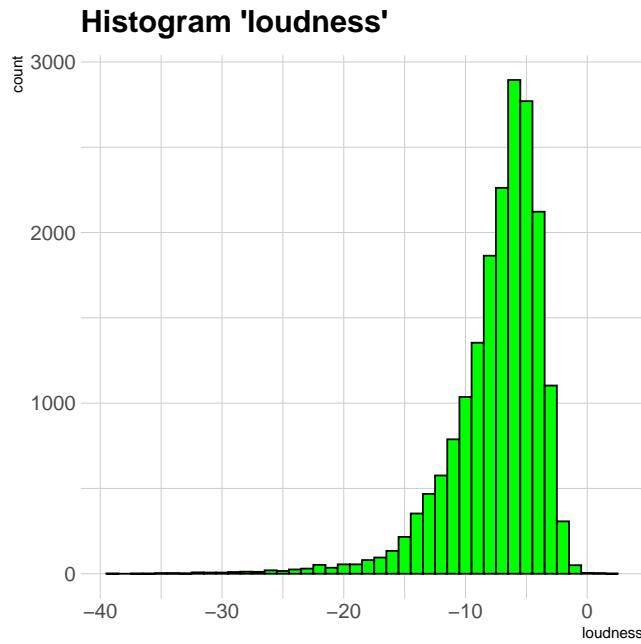
```
data: energy
```

```
D = 0.058258, p-value < 2.2e-16
```

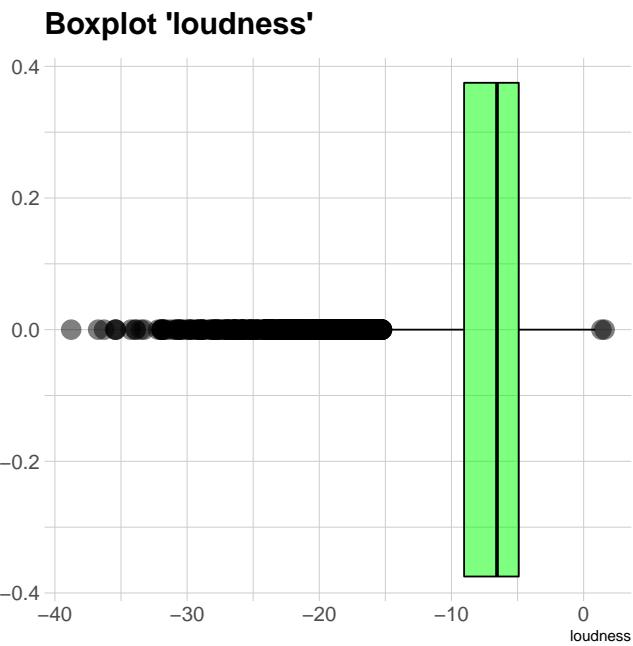
Zarówno wykres QQ, jak i test Lillieforsa wskazują, że rozkład zmiennej 'energy' jest daleki od rozkładu normalnego.

### 3.5 Zmienna 'loudness'

```
> histogram <- ggplot(dataSet, aes(x = loudness), xlab="loudenss") +
+   geom_histogram(binwidth=1, color="black", fill="green1") +
+   theme_ipsum(base_family = 'sans') +
+   labs(title="Histogram 'loudness'")  
> histogram
```



```
> boxplot <- ggplot(dataSet, aes(x = loudness), xlab="loudness") +
+   geom_boxplot(color="black", fill="green1", alpha=0.5, outlier.size=5) +
+   theme_ipsum(base_family = 'sans') +
+   labs(title="Boxplot 'loudness'")  
> boxplot
```



```

> range(loudness)
[1] -38.768   1.585

> mean(loudness)
[1] -7.447435

> quantile(loudness, c(0.25, 0.5, 0.75))
 25%      50%      75%
-9.044 -6.555 -4.908

> var(loudness)
[1] 14.65229

> sd(loudness)
[1] 3.827831

> skewness(loudness)
[1] -1.929357

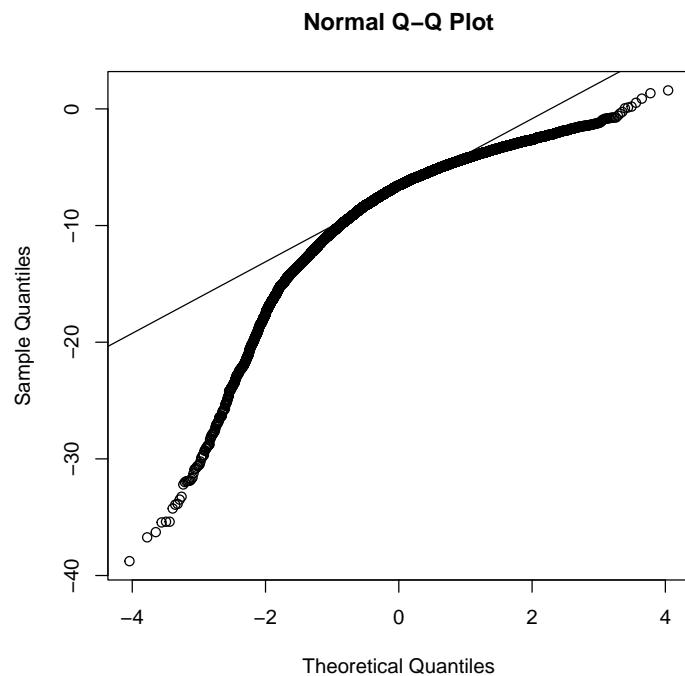
```

```
> kurtosis(loudness)
```

```
[1] 9.52043
```

Wartości zmiennej sięgają od -38.768 do 1.585, ich średnia to -7.447, wariancja to 14.652, a odchylenie standardowe to 3.828. Współczynnik asymetrii jest ujemny, zatem rozkład jest lewostronnie skośny. Współczynnik wyostrzenia jest dodatni, więc rozkład jest leptokurtyczny.

```
> qqnorm(loudness)
> qqline(loudness)
```



```
> lilliefors.test(loudness)
```

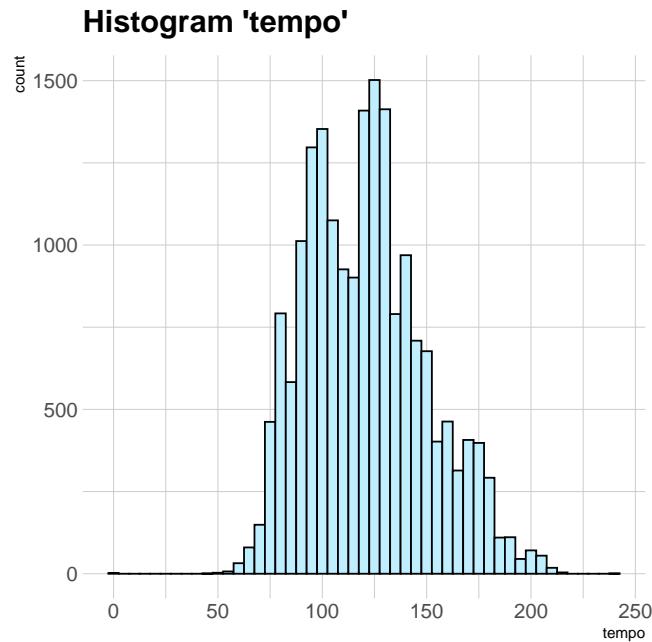
```
Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data: loudness
D = 0.10797, p-value < 2.2e-16
```

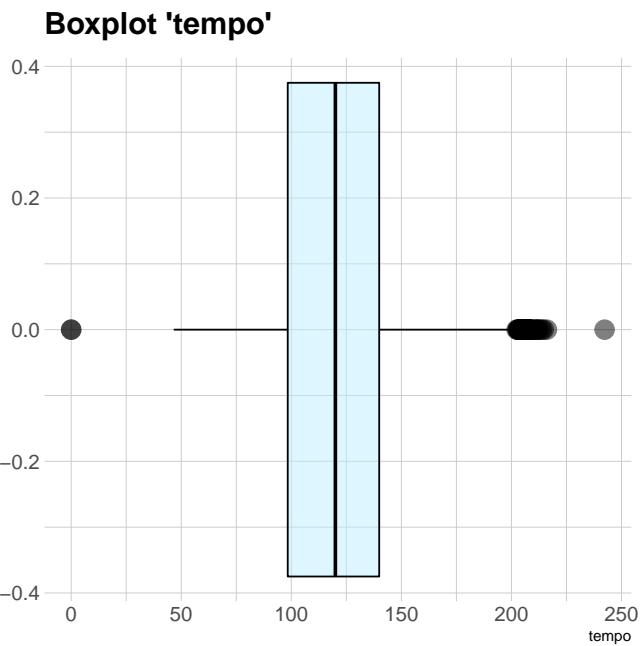
Zarówno wykres QQ, jak i test Lillieforsa wskazują, że rozkład zmiennej 'loudness' jest daleki od rozkładu normalnego.

### 3.6 Zmienna 'tempo'

```
> histogram <- ggplot(dataSet, aes(x = tempo), xlab="tempo") +  
+   geom_histogram(binwidth=5, color="black", fill="lightblue1") +  
+   theme_ipsum(base_family = 'sans') +  
+   labs(title="Histogram 'tempo'")  
> histogram
```



```
> boxplot <- ggplot(dataSet, aes(x = tempo), xlab="tempo") +  
+   geom_boxplot(color="black", fill="lightblue1", alpha=0.5, outlier.size=5) +  
+   theme_ipsum(base_family = 'sans') +  
+   labs(title="Boxplot 'tempo'")  
> boxplot
```



```

> range(tempo)
[1] 0.000 242.318
> mean(tempo)
[1] 121.0732
> quantile(tempo, c(0.25, 0.5, 0.75))
 25%      50%      75%
98.368 120.013 139.931
> var(tempo)
[1] 824.52
> sd(tempo)
[1] 28.71446
> skewness(tempo)
[1] 0.4428193

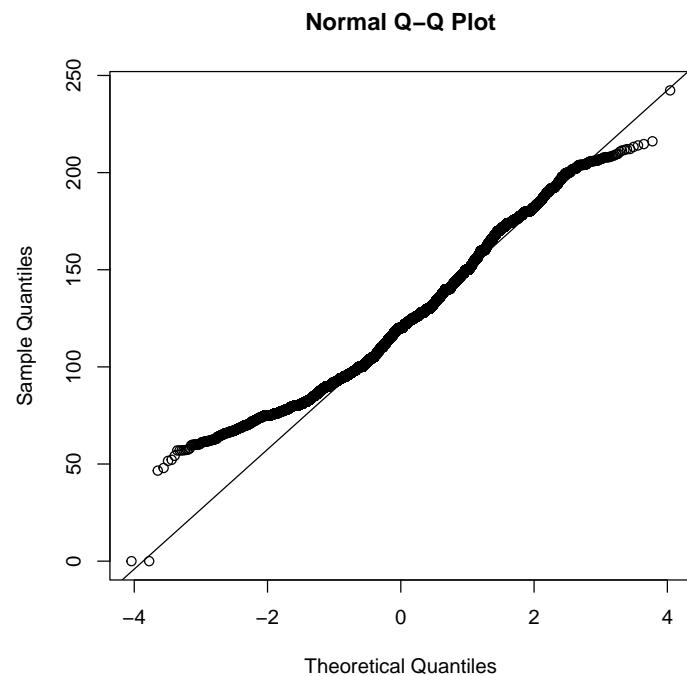
```

```
> kurtosis(tempo)
```

```
[1] 2.782223
```

Wartości zmiennej sięgają od 0 do 242.32, ich średnia to 121.07, wariancja to 824.52, a odchylenie standardowe to 28.714. Współczynnik asymetrii jest dodatni, zatem rozkład jest prawostronnie skośny. Natomiast współczynnik wyostrzenia jest ujemny, czyli rozkład jest platykurytyczny.

```
> qqnorm(tempo)
> qqline(tempo)
```



```
> lillie.test(tempo)
```

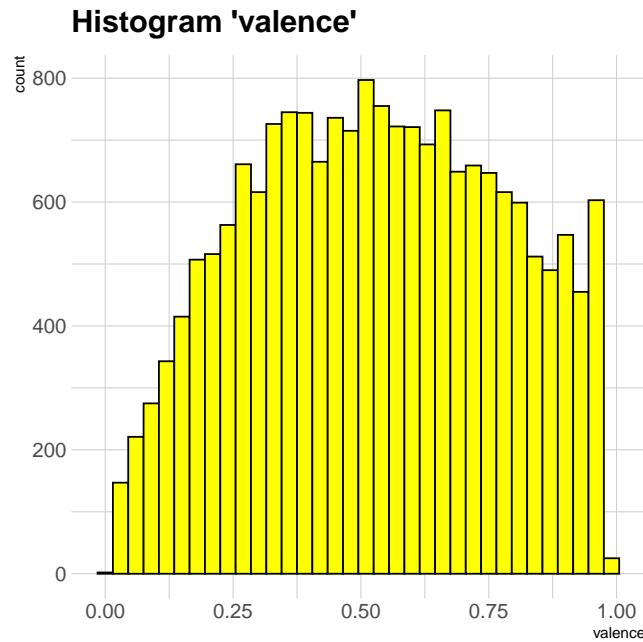
```
Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data: tempo
D = 0.055522, p-value < 2.2e-16
```

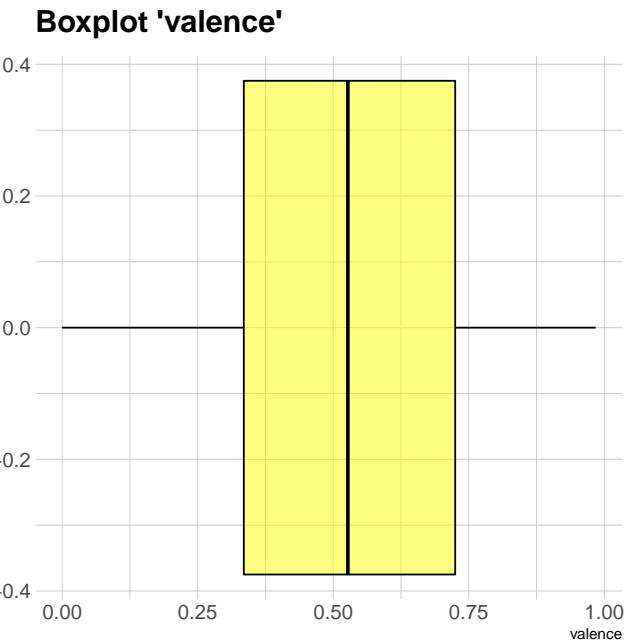
Zarówno wykres QQ, jak i test Lillieforsa wskazują, że rozkład zmiennej 'tempo' jest daleki od rozkładu normalnego.

### 3.7 Zmienna 'valence'

```
> histogram <- ggplot(dataSet, aes(x = valence), xlab="valence") +  
+   geom_histogram(binwidth=0.03, color="black", fill="yellow1") +  
+   theme_ipsum(base_family = 'sans') +  
+   labs(title="Histogram 'valence'")  
> histogram
```



```
> boxplot <- ggplot(dataSet, aes(x = valence), xlab="valence") +  
+   geom_boxplot(color="black", fill="yellow1", alpha=0.5, outlier.size=5) +  
+   theme_ipsum(base_family = 'sans') +  
+   labs(title="Boxplot 'valence'")  
> boxplot
```



```

> range(valence)
[1] 0.000 0.984
> mean(valence)
[1] 0.5279669
> quantile(valence, c(0.25, 0.5, 0.75))
 25%   50%   75%
0.335 0.527 0.725
> var(valence)
[1] 0.05984466
> sd(valence)
[1] 0.2446317
> skewness(valence)
[1] -0.01642191

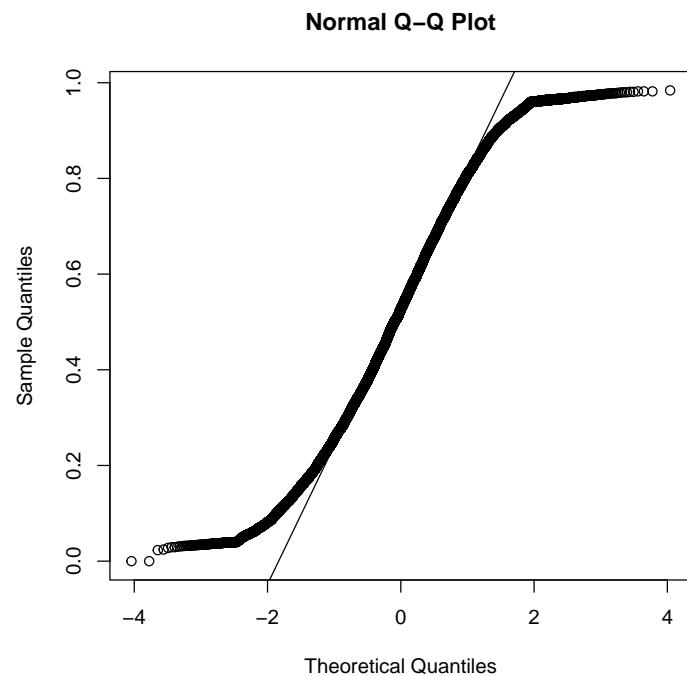
```

```
> kurtosis(valence)
```

```
[1] 2.02227
```

Wartości zmiennej sięgają od 0 do 0.984, ich średnia to 0.528, wariancja to 0.0598, a odchylenie standardowe to 0.245. Współczynnik asymetrii jest ujemny, zatem rozkład jest lewostronnie skośny. Współczynnik wyostrzenia jest ujemny, więc rozkład jest platykurytyczny.

```
> qqnorm(valence)
> qqline(valence)
```



```
> lilliefors.test(valence)
```

```
Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data: valence
D = 0.041885, p-value < 2.2e-16
```

Zarówno wykres QQ, jak i test Lillieforsa wskazują, że rozkład zmiennej 'valence' jest daleki od rozkładu normalnego.

## 4 Analiza zależności pomiędzy zmiennymi

### 4.1 Kowariancja i korelacja między wszystkimi zmiennymi

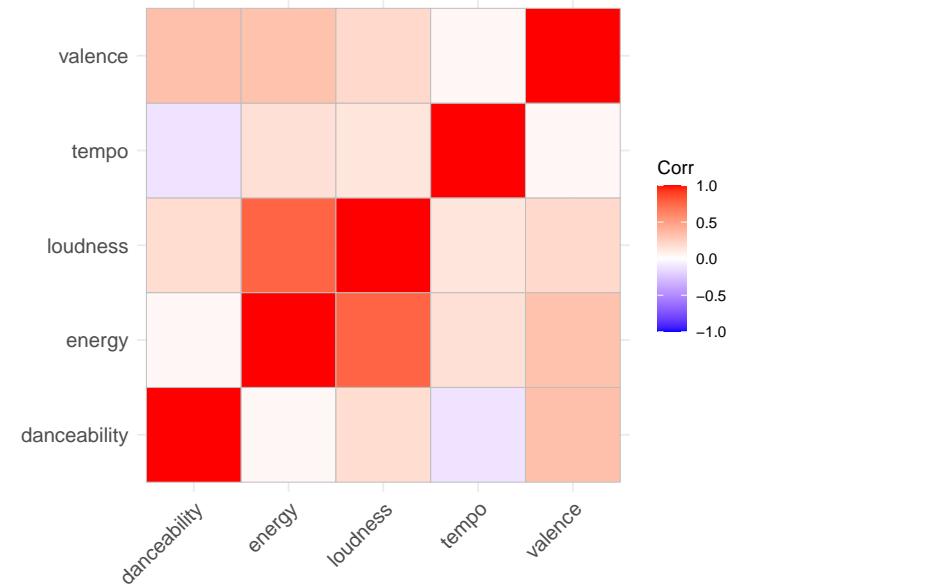
```
> cov(dataSet[,c(1,2,3,4,5)])
```

|              | danceability | energy      | loudness   | tempo       | valence    |
|--------------|--------------|-------------|------------|-------------|------------|
| danceability | 0.024562006  | 0.001488922 | 0.1065669  | -0.5458132  | 0.01272921 |
| energy       | 0.001488922  | 0.045839134 | 0.6191764  | 0.9986978   | 0.01658961 |
| loudness     | 0.106566874  | 0.619176350 | 14.6522916 | 14.2820088  | 0.18699703 |
| tempo        | -0.545813232 | 0.998697790 | 14.2820088 | 824.5199676 | 0.26458976 |
| valence      | 0.012729213  | 0.016589609 | 0.1869970  | 0.2645898   | 0.05984466 |

```
> cor(dataSet[,c(1,2,3,4,5)])
```

|              | danceability | energy     | loudness  | tempo       | valence    |
|--------------|--------------|------------|-----------|-------------|------------|
| danceability | 1.00000000   | 0.04437331 | 0.1776387 | -0.12128625 | 0.33201441 |
| energy       | 0.04437331   | 1.00000000 | 0.7555155 | 0.16244835  | 0.31674170 |
| loudness     | 0.17763868   | 0.75551551 | 1.0000000 | 0.12993793  | 0.19969593 |
| tempo        | -0.12128625  | 0.16244835 | 0.1299379 | 1.00000000  | 0.03766689 |
| valence      | 0.33201441   | 0.31674170 | 0.1996959 | 0.03766689  | 1.00000000 |

```
> ggcorrplot(cor(dataSet[,c(1,2,3,4,5)]))
```

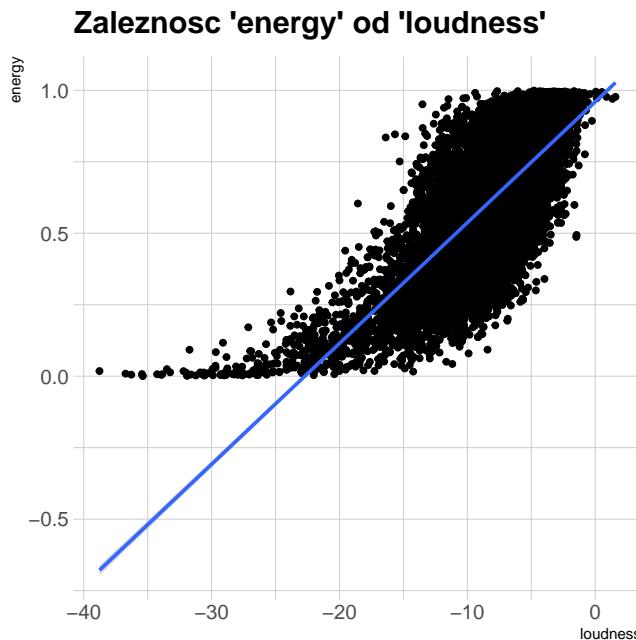


Na podstawie uzyskanych danych można stwierdzić, że istnieje całkiem silna korelacja pomiędzy zmiennymi 'energy' i 'loudness'. Ponadto pomiędzy zmiennymi 'valence' i 'danceability' oraz 'valence' i 'energy' istnieją zauważalne korelacje.

#### 4.2 Badanie zgodności rozkładów 'energy' i 'loudness'

Rozpoczęłem od wygenerowania wykresu zależności 'energy' od 'loudness'

```
> plotle <- ggplot(dataSet, aes(x = loudness, y = energy), xlab="loudness",
+   ylab="energy") +
+   geom_point() +
+   geom_smooth(formula = y ~ x, method = "lm") +
+   theme_ipsum(base_family = 'sans') +
+   labs(title="Zależność 'energy' od 'loudness'")  
> plotle
```

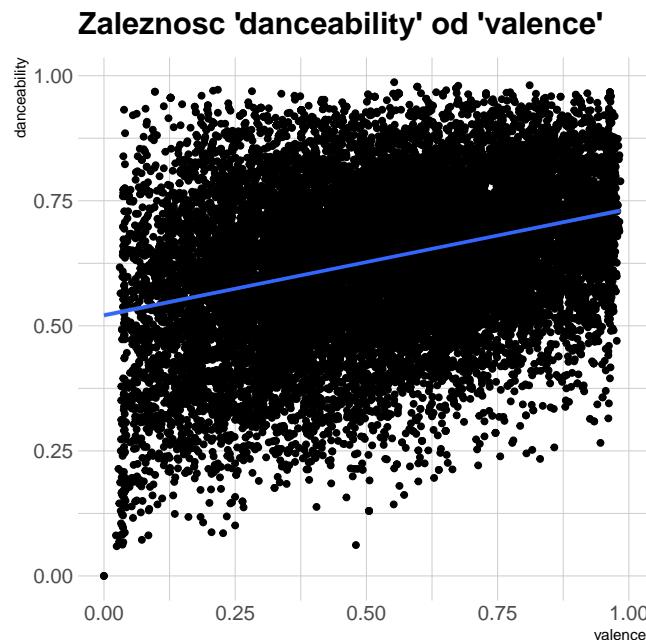


Wykres nie wskazał obecności zależności liniowej pomiędzy 'energy' a 'loudness' zatem dalsza analiza nie jest sensowna.

### 4.3 Badanie zgodności rozkładów 'valence' i 'danceability'

Wygenerowałem wykres zależności 'danceability' od 'valence'

```
> plotvd <- ggplot(dataSet, aes(x = valence, y = danceability), xlab="valence",
+   ylab="danceability") +
+   geom_point() +
+   geom_smooth(formula = y ~ x, method = "lm") +
+   theme_ipsum(base_family = 'sans') +
+   labs(title="Zależność 'danceability' od 'valence'")
> plotvd
```



Widoczna na wykresie zależność liniowa jest wystarczająco zauważalna, zatem przechodzę do analizy regresji liniowej

Testowane zmienne są niezależne.

Można przyjąć, że zmienna 'danceability' ma rozkład zbliżony do rozkładu normalnego.

```

> reg <- lm(danceability ~ valence)
> summary(reg)

Call:
lm(formula = danceability ~ valence)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.56145 -0.09842  0.00466  0.10494  0.42630 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.521047  0.002562 203.3   <2e-16 ***
valence     0.212704  0.004403  48.3   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1478 on 18833 degrees of freedom
Multiple R-squared:  0.1102,    Adjusted R-squared:  0.1102 
F-statistic: 2333 on 1 and 18833 DF,  p-value: < 2.2e-16

```

Z otrzymanych danych można wywnioskować, że istnieje istotna pozytywna relacja pomiędzy 'valence' a 'danceability' (p-value < 0.001). Można zaobserwować także wzrost 'valence' o 0.213 związany z jednostkowym wzrostem 'danceability'

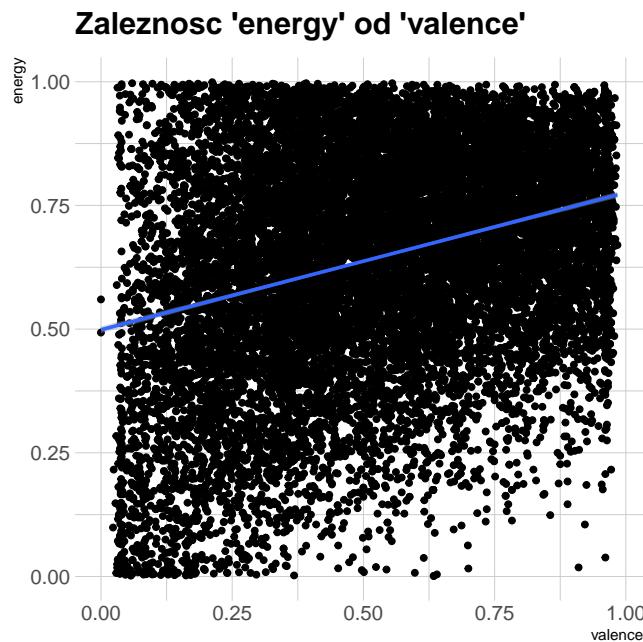
Jednakże niski wskaźnik  $R^2$  mówi o tym, że testowane zmienne mają jednak dość ograniczony wpływ na siebie.

#### 4.4 Badanie zgodności rozkładów 'valence' i 'energy'

Finalnie wygenerowałem wykresu zależności 'energy' od 'valence'

```
> plotve <- ggplot(dataSet, aes(x = valence, y = energy), xlab="valence",
+   ylab="energy") +
+   geom_point() +
+   geom_smooth(formula = y ~ x, method = "lm") +
+   theme_ipsum(base_family = 'sans') +
+   labs(title="Zależność 'energy' od 'valence'")
```

> plotve



Zależność liniowa ukazana na wykresie nie jest na tyle widoczna, aby warto było przeprowadzać dalsze badanie.

## **5 Wnioski**

Wnioski płynące z przeprowadzonej analizy, są następujące:

- rozkład 'danceability' jest zbliżony kształtem do rozkładu normalnego,
- wykonane testy Lillieforsa wskazują jednakże, że badane rozkłady nie są w rzeczywistości zbliżone do rozkładu normalnego,
- istnieje silna dodatnia korelacja pomiędzy zmiennymi 'energy' i 'loudness',
- widoczna jest zależność liniowa pomiędzy 'valence' a 'danceability'