Najlepsze albumy muzyczne

wg użytkowników RateYourMusic.com

Bartosz Łuksza, Rafał Głodek

Wprowadzenie

Muzyka towarzyszy człowiekowi od tysięcy lat. Zawsze stanowiła nieodłączną część naszej kultury. Niestety ograniczenia technologiczne przez długi czas nie pozwalały artystom utrwalić swoich dzieł. Fonografia narodziła się w XIX wieku, a jej największy rozkwit przypada na drugą połowę wieku XX. Z tego względu najstarsze oryginalne dzieła muzyczne, do których mamy obecnie dostęp, pochodzą poprzedniego stulecia. W ostanich latach rynek muzyczny przeżywa niebywały rozwkit. Każdego roku miliardy słuchaczy na całym świecie, przesłuchuje miliony nowych albumów, generując przychody rzędu dziesiątek miliardów dolarów ze sprzedaży nagrań. Rynek muzyczny jest jednak ściśle powiązany z wieloma innymi gałęziami biznesu, takimi jak: film, moda, czy technologie cyfrowe. Szacuje się, że każdego dnia na serwisy streamingowe trafia nawet 120 000 utworów! W tej sytuacji można pokusić się o stwierdzenie, że obecny przemysł muzyczny jest wręcz "przeładowany" muzyką. Warto zadać sobie pytanie, czy za ilością idzie również jakość?

W naszej pracy zajrzymy wgłąb współczesnej historii muzyki i przeanalizujemy bazę pięciu tysięcy najlepiej ocenianych albumów muzycznych przez użytkowników RateYourMusic.com - największego portalu do oceniania muzyki w internecie. Dane pochodzą z 12 grudnia 2021 r. i zostały pobrane z serwisu kaggle.com. Wyodrębniliśmy z nich następujące zmienne:

1. ***Album*** - nazwa albumu
   - Zawiera 4928 unikalne wartości
2. ***Artist Name*** - artysta (imię i nazwisko lub pseudonim artystyczny)
   - Zawiera 2787 unikalne wartości
   - 25 rekordów to *Various Artists*, czyli różni artyści, których jednak nie możemy wyodrębnić, więc pomijamy te wartości
3. ***Release Date*** - dokładna data wydania albumu (dzień/miesiąc/rok), wyodrębniliśmy z niej dwie zmienne:

a) ***Year*** - rok wydania albumu

- najmniejsza wartość: 1947
- największa wartość: 2021
- średnia arytmetyczna: 1987,46
- mediana: 1988
- wariancja: 253,23

b) ***Month*** - miesiąc wydania albumu

4. ***Genres*** - gatunki (lub gatunek), do których należy album

- Niekiedy trudno jest ustalić, do jakiego gatunku należy album. Wówczas określany jest mianem międzygatunkowego i klasyfikuje się go jako przynależnego do każdego z wymienionych gatunków.
- Możliwe wartości to: "Rock", "Hip Hop", "Pop" , "Jazz", "Soul", "Dance", "Techno", "Punk", "Metal", "Folk"

5. ***Descriptors*** - krótki opis albumu

- Opis zawiera kilka przymiotników najlepiej oddających charakter albumu, np. "melancholic, anxious, futuristic, alienation, existential, male vocals, atmospheric, lonely, cold, introspective"
- Opisy będą potrzebne, by sprawdzić jak oceniane są albumy w zależności od nastroju, jaki wywołują w słuchaczu

6. ***Average Rating*** - średnia ocen użytkowników

- Na stronie RateYourMusic.com użytkownicy mogą wystawiać albumom oceny w skali od 0 do 5, z uwzględnieniem "połówek"
- Średnie oceny, które są brane pod uwagę w tym spisie uwzględniają także wagi ocen - oceny użytkowników wykazujących się dużą aktywnością i doświadczeniem mają wyższą wagę niż tych, którzy oceniają muzykę sporadycznie
- najmniejsza wartość: 3,52
- największa wartość: 4,34
- średnia arytmetyczna: 3,771
- mediana: 3,75
- wariancja: 0,0098

7. ***Number of Ratings*** - liczba ocen użytkowników

- najmniejsza wartość: 260

- największa wartość: 70 400
- średnia arytmetyczna: 4084.511
- mediana: 1820
- wariancja: 36016085

8. ***Number of Reviews*** - liczba recenzji użytkowników
   - najmniejsza wartość: 0
   - największa wartość: 1 549
   - średnia arytmetyczna: 71.4492
   - mediana: 34
   - wariancja: 11766.56

Dogłębna analiza pozwoli nam znaleźć korelacje między różnymi zmiennymi ujętymi w zestawieniu i wyciągnąć nieoczywiste wnioski. W ten sposób nie tylko dowiemy się, jak muzyka rozwijała się w ubiegłych dekadach, ale także nakreślimy scieżkę jej dalszego rozwoju.

W jakich latach powstawało najwięcej "dobrych" albumów? Czy istnieje korelacja między średnią oceną użytkowników a datą wydania dzieła? Jakie są średnie ocen dla różnych gatunków muzycznych? Którzy artyści mogą się poszczycić najlepiej ocenianią dyskografią? Na te i wiele innych pytań odpowiemy w naszej pracy.

Analiza danych

Analiza rozkładów lat oraz średnich ocen

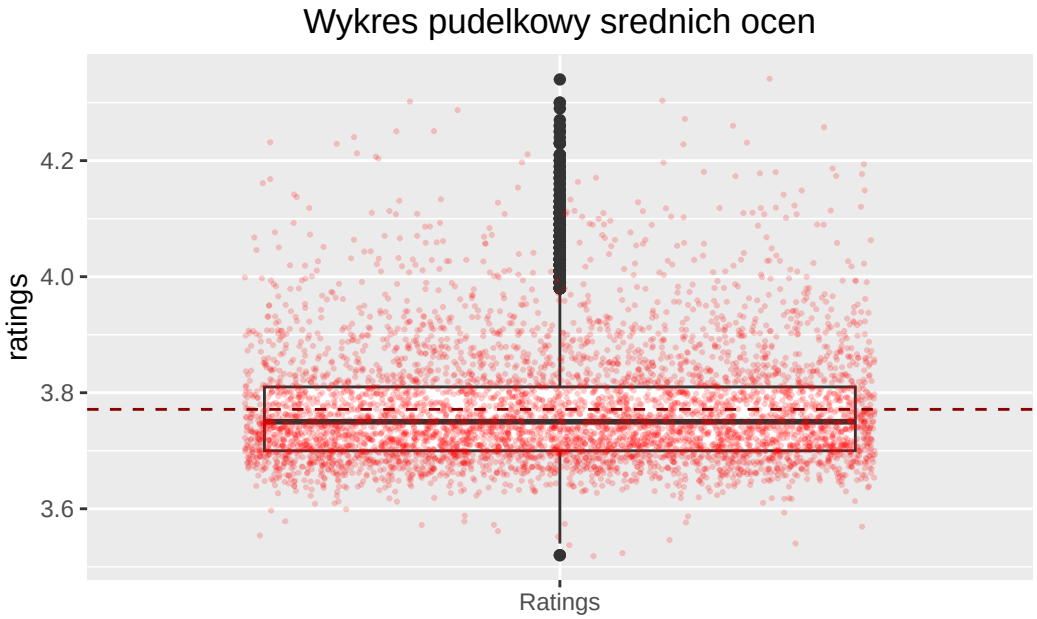
W pierwszej części naszej analizy zbadamy, jak rozkładają się średnie ocen wystawionych przez użytkowników oraz lata wydania albumów. Zaczniemy od wygenerowania wykresów pudełkowych dla każdej z nich oraz wyliczenia jego parametrów - mediany, pierwszego i trzeciego kwartyla, rozstępu międzykwartylowego oraz górnego i dolnego wąsa. Ponadto na wykres pudełkowy nałożymy także realizację naszej zmiennej w postaci punktów oraz jej średnią arytmetyczną.

Wykres pudełkowy dla średnich ocen prezentuje się następująco

```
Warning: package 'hrbrthemes' was built under R version 4.4.2
```

```
Warning: package 'viridis' was built under R version 4.4.2
```

```
Loading required package: viridisLite
```

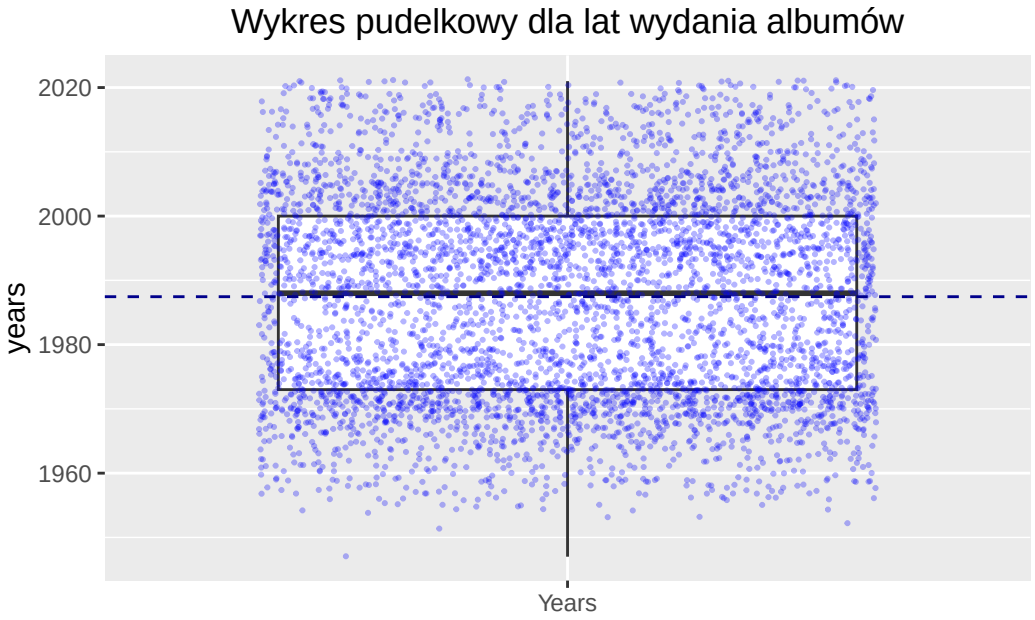


Korzystając z funkcji *boxplot.stats* wydobędziemy z wykresu najważniejsze dane. Przedstawimy je w formie tabeli

Wąs dolny	3.54
Pierwszy kwartyl	3.70
Mediana	3.75
Trzeci kwartyl	3.81
Wąs górny	3.97

Na bazie wykresu oraz tabeli możemy wyciągnąć kilka istotnych wniosków. Zgodnie z ideą wykresu pudełkoweg, zdecydowana większość punktów znajduje się w przedziale między wąsem dolnym a wąsem górnym, czyli $(3.54, 3.97)$. Obserwacje wypadające z niego możemy uznać za odstające. Jeden punkt znajduje się pod wąsem dolnym, natomiast możemy odnaleźć dużo więcej punktów osadzonych ponad wąsem górnym. W kontekście tematyki naszej pracy, możemy interpretować je jako ścisłą czołówkę albumów. Średnia arytmetyczna, będąca nieobciążonym estymatorem wartości oczekiwanej, jest większa niż mediana, a zatem mamy w tym przypadku do czynienia z rozkładem prawoskośnym. Oznacza to dla nas, że wyniki poniżej średniej są w naszej próbce przeważające. Oceny znacznie odbiegające od średniej są zatem dużą rzadkością i tym samym czołówka rankingu zarysowuje się nam coraz mocniej.

Teraz przeprowadzimy analogiczną analizę dla lat wydania albumów. Wygenerujmy dla danych wykres pudełkowy

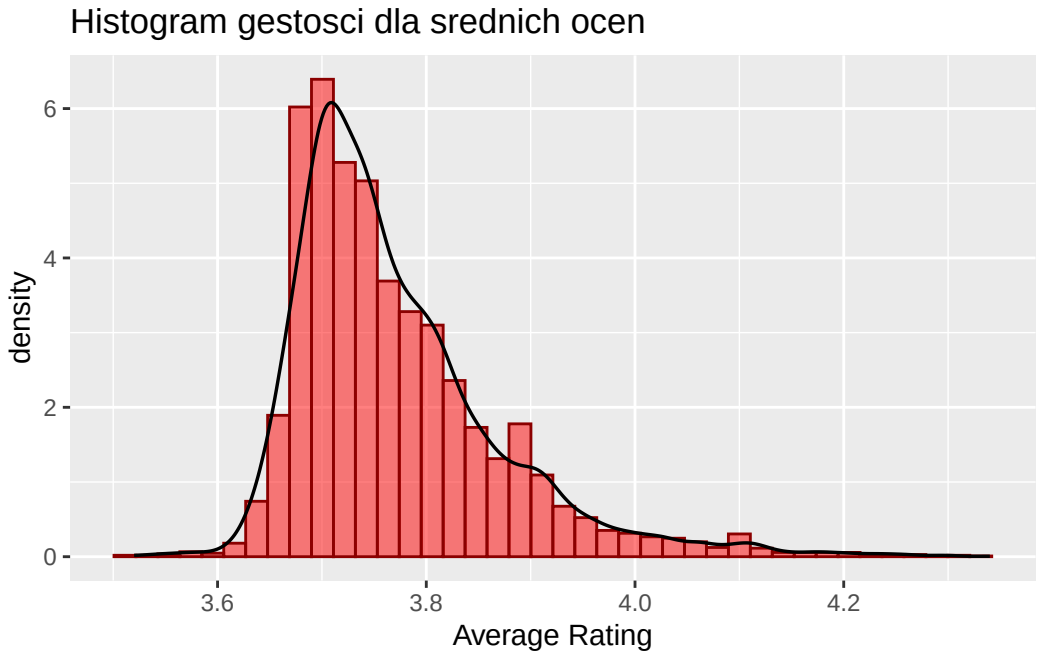


Znów wykorzystamy funkcję *boxplot.stats* i wyliczymy parametry tego wykresu pudełkowego. Zawiera je poniższa tabela.

Wąs dolny	1947
Pierwszy kwartyl	1973
Mediana	1988
Trzeci kwartyl	2000
Wąs górny	2021

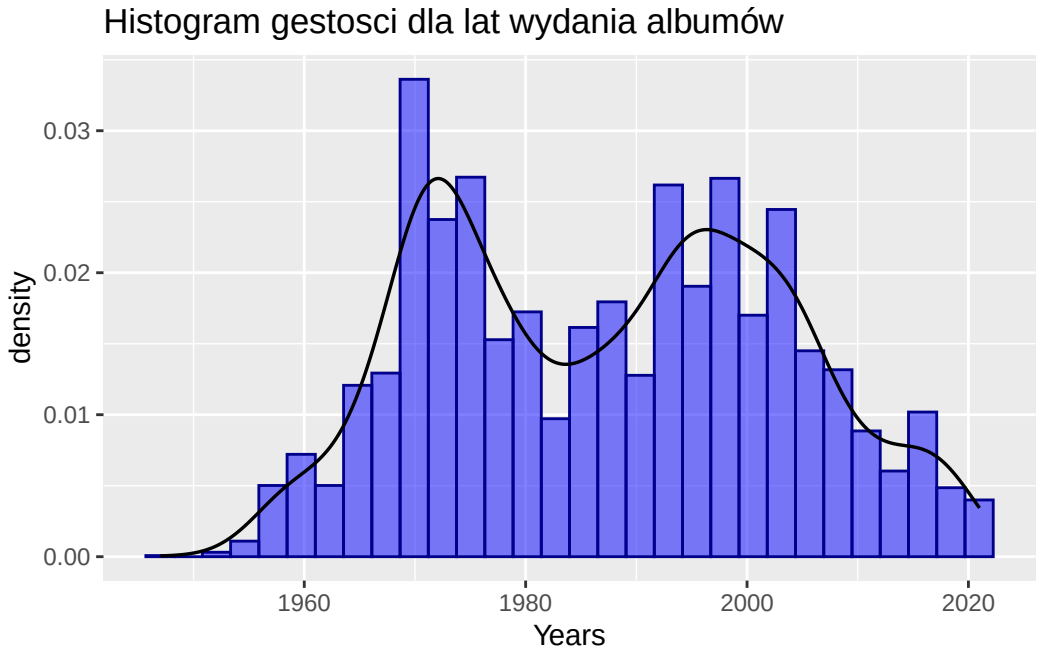
Wyciągnijmy teraz wnioski z tabeli i wykresu. Zauważmy, że wąsy górne i dolne pokrywają się z minimum i maksimum lat wydania albumów. Z tego powodu w rozważanym zbiorze nie występują żadne wartości odstające. Średnia arytmetyczna niemalże pokrywa się z medianą, więc rozkład lat będzie przypominał rozkład symetryczny. Największe zagęszczenie danych występuje między 1973 a 2000 rokiem, czyli między pierwszym a czwartym kwartylem. Oznacza to, że aż 50% wszystkich albumów zakwalifikowanych do rankingu zostało wydanych w okresie tych 27 lat, a w całym zestawieniu rozważamy przedział 74 lat. Można zatem stwierdzić, że najwięcej wysoko ocenianych albumów zostało wydanych w latach 70., 80. i 90. XX wieku.

W następnym kroku dla każdej z rozważanych zmiennych wygenerujemy histogram prawdopodobieństwa *geom_histogram* i dopasujemy do niego jądrowy estymator gęstości używając *geom_density*. Wynik dla średnich ocen prezentuje się następująco



Na bazie wykresu możemy stwierdzić, że oceny użytkowników mają rozkład prawostronnie skośny, z pojedynczą górką (modą) znajdującą się w okolicach punktu 3.7, co sugeruje, że większość albumów uzyskuje taką ocenę. Gęstość empiryczna zaczyna gwałtownie maleć w okolicach punktu 4.1, co potwierdza wniosek wyciągnięty na bazie wykresu pudełkowego — takie oceny są bardzo rzadkie, bowiem charakteryzują jedynie ścisłą czołówkę albumów. Rozkład ten swoim kształtem nieco przypomina rozkład normalny, lecz występuje tu wyraźna asymetria, świadcząca o jego skośności z prawej strony. Można zatem potwierdzić wcześniejszą obserwację, że oceny najczęściej pojawiające się w naszym zestawieniu, znajdują się w przedziale od umiarkowanych do nieco poniżej średniej.

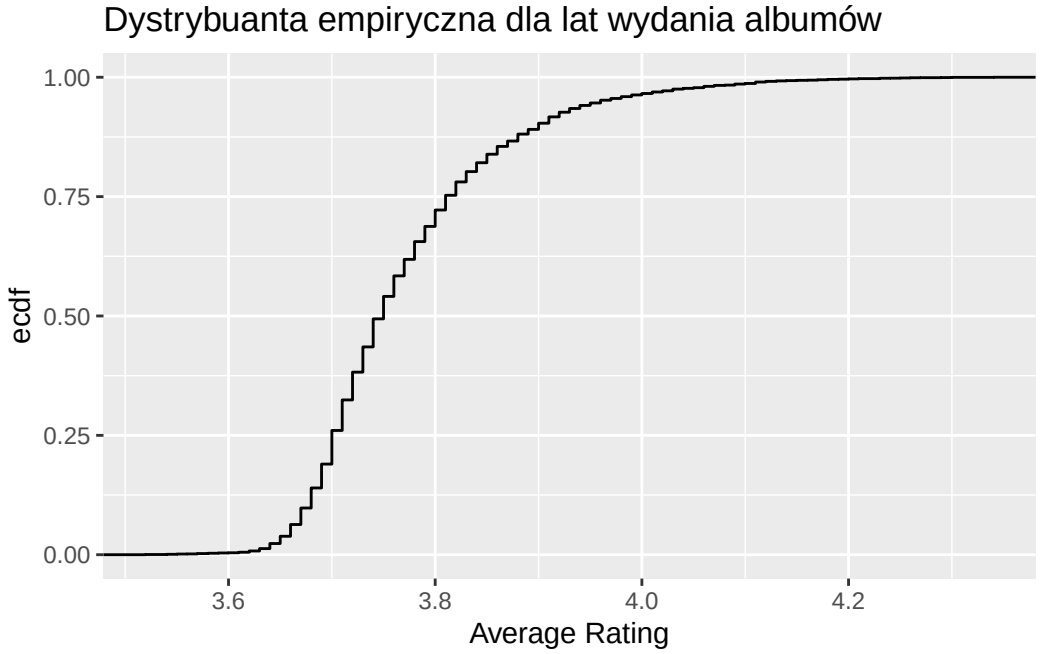
Następnie wygenerujemy analogiczny wykres dla lat wydania albumów.



Rozkład lat wydania albumów ma nieco bardziej skomplikowaną naturę. Można go zaklasyfikować jako dwumodalny, czyli posiadający dwa wyraźne punkty skupienia. Pierwszy wierzchołek tego rozkładu znajduje się w okolicy lat 70., gdy światem muzyki zawładnęły gatunki, takie jak rock, metal i disco. Kolejny szczyt widoczny jest na przełomie lat 90. i 2000., kiedy na światowych scenach dominowały: pop, grunge oraz przede wszystkim — hip hop. Poza tymi okresami występowały znaczące spadki w liczbie dobrze ocenianych albumów. Szokującym może być fakt, że w obecnych czasach obserwowany jest największy spadek jakościowej muzyki od lat 50 z tym, że wtedy wydawano znacznie mniej albumów w porównaniu do dzisiejszych czasów. Warto więc zastanowić się nad pytaniem, czy w czasach współczesnych, pomimo szerokiej dostępności muzyki oraz ogromnych pieniędzy wydawanych na jej produkcję i dystrybucję, stoimy w obliczu największego kryzysu muzycznego?

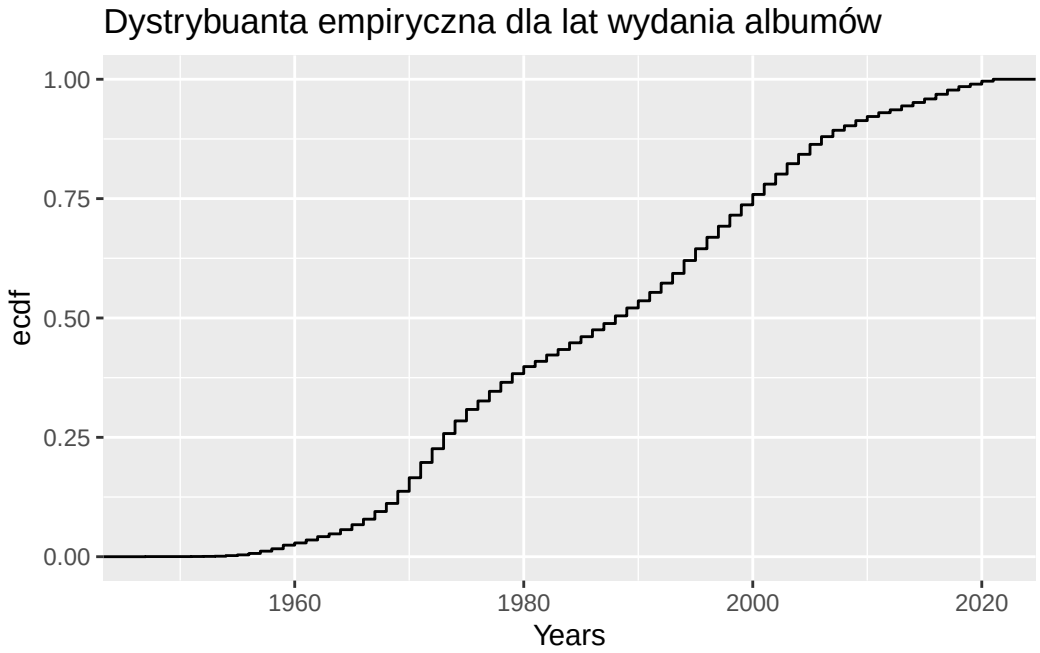
Kolejnym krokiem w analizie rozkładów rozważanych zmiennych będzie wygenerowanie wykresów ich dystrybuant empirycznych za pomocą funkcji $stat\_ecdf$.

Dla średnich ocen wykres ten prezentuje się nastepująco.



Dystrybuanta empiryczna ocen użytkowników potwierdza wnioski wyciągnięte na bazie gęstości empirycznej. Największy wzrost wartości dystrybunaty możemy zaobserwować na przedziale od 3.6 do 3.9, a zatem, gdy oceny oscylują wokół wartości średniej. Dla punktu 3.9 wartość dystrybunaty wynosi około 0.9, a więc prawdopodobieństwo, że losowo wybrana ocena z próbki przekroczy próg 3.9 sięga zaledwie jednej dziesiątej. Ta obserwacja dobrze pokazuje, że albumy należące do ścisłej czołówki najlepiej ocenianych, stanowią bardzo niewielką część całego zestawienia.

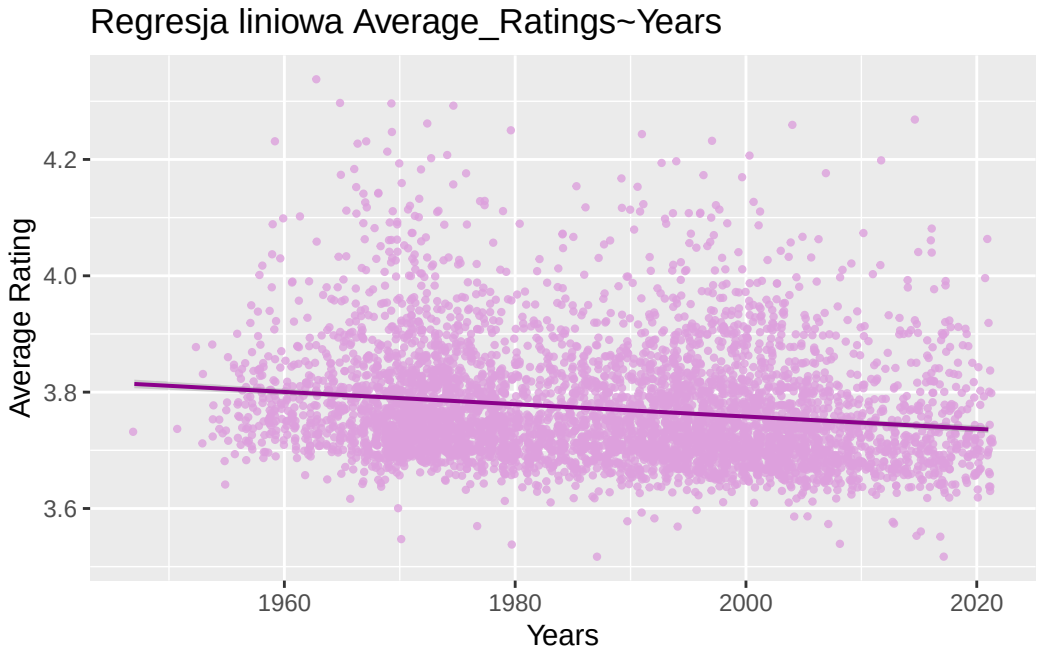
To samo wykonamy dla lat wydania albumów.



Dystrybuanta empiryczna lat wydania albumów jest znacznie bardziej wypłaszczona niż ocen użytkowników, choć nagłe tendencje wzrostowe w okolicach lat 70. i przełomu stuleci, nadal są tu dobrze widoczne, tak jak w przypadku wykresu gęstości empirycznej. Wzrost dystrybuanty zwalnia w dolinach między dwoma szczytami. Zauważmy, że dystrybuanta empiryczna przyjmuje wartość 0.5 mniej więcej w środku przedziału lat, co stwarza wrażenie równomierności rozkładu, dobrze widocznej wcześniej na wykresie pudełkowym.

Wiemy już jak rozkładały się średnie oceny użytkowników oraz lata wydania albumów. Warto w tym miejscu zastanowić się, czy istnieje między tymi zmiennymi jakaś zależność. Poszukamy zatem zależności liniowej stosując z wbudowanej w pakiet R funkcji *lm*, wykorzystującej model regresji liniowej. Gdyby udało się nam znaleźć między nimi silną zależnosć, moglibyśmy stwierdzić, jak zmieniała się muzyka w minionych dekadach oraz postarać się wskazać kierunek jej rozwoju w przyszłości.

Na wykres punktowy “Average Rating” vs “Years” nałożyliśmy prostą dostarczoną przez metodę *lm*. Prezentuje się on w następujący sposób



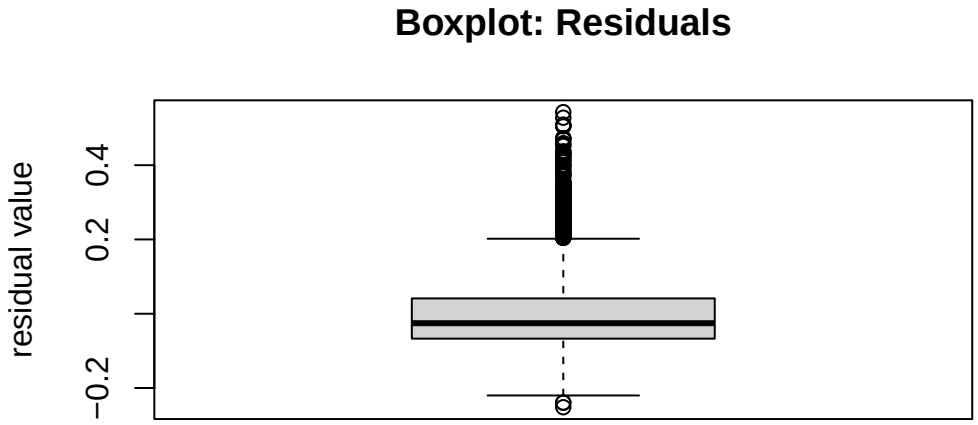
Widzimy, że punkty są dość mocno rozrzucone i nie widać między nimi szukanej liniowej zależności. Model dopasował do danych prostą regresji, lecz nie możemy spodziewać się w tym przypadku dużej skuteczności. Celem poprawnej interpretacji wykresu wydobyliśmy niezbędne informacje o regresji liniowej za pomocą funkcji *summary*. Prezentują się one następująco

```
lm_ratings_and_years = lm(ratings~years)
summary(lm_ratings_and_years)
```

```
Call:
lm(formula = ratings ~ years)

Residuals:
     Min       1Q   Median       3Q      Max
-0.25179 -0.06707 -0.02552  0.04131  0.54285

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.871e+00  1.724e-01   34.05   <2e-16 ***
years       -1.057e-03  8.676e-05  -12.18   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09762 on 4998 degrees of freedom
```

```
Multiple R-squared:  0.02882,   Adjusted R-squared:  0.02863
F-statistic: 148.3 on 1 and 4998 DF,  p-value: < 2.2e-16
```

Na sam początek funkcja przekazuje nam informacje o residuach modelu, czyli różnicach między wartościami obserwowanymi a przewidywanymi przez model. Idealnie chcielibyśmy, żeby rozkład residuów był możliwie najbardziej symetryczny, a często testuje się także jego normalność. Większość punktów byłaby wtedy zbliżona do prostej, a obserwacje bardziej od niej odbiegające są względem niej symetryczne. Tym samym, chcielibyśmy, żeby mediana była zbliżona do zera. Kwartyle oraz wartości maksymalne i minimalne również powinny być symetryczne względem zera oraz możliwie jak najbardziej do siebie zbliżone. Zwizualizujemy rozkład residuów na wykresie pudełkowym.



Widzimy, że mediana jest bliska zeru, co sugeruje że reszty są stosunkowo symetryczne. Wąs górny i dolny są dość symetryczne względem zera, a połowa centralnych reszt — znajdująca się wewnątrz pudełka, między pierwszym i trzecim kwartylem — charakteryzuje się wzlędnie małą rozpiętością. Możemy jednak zaobserwować bardzo dużo wartości odstających, które wskazują na duże trudności w dopasowaniu modelu.

Następną informacją otrzymaną z metody *summary* są wartości parametrów prostej regresji. Zależność liniową między zmienną objaśnianą y, a objaśniającą x możemy zapisać jako

$$y = ax + b$$

Gdzie a jest współczynnikiem kierunkowym prostej, a b — wyrazem wolnym. W naszym wypadku parametry te wynoszą $a = -1.057 \cdot 10^{-3}$ i $b = 5.871$.

Następnie w podsumowaniu umieszczone są błędy standardowe estymatorów a i b. W naszym wypadku wynoszą one odpowiednio 0.1724 oraz $8.676 \cdot 10^{-5}$.

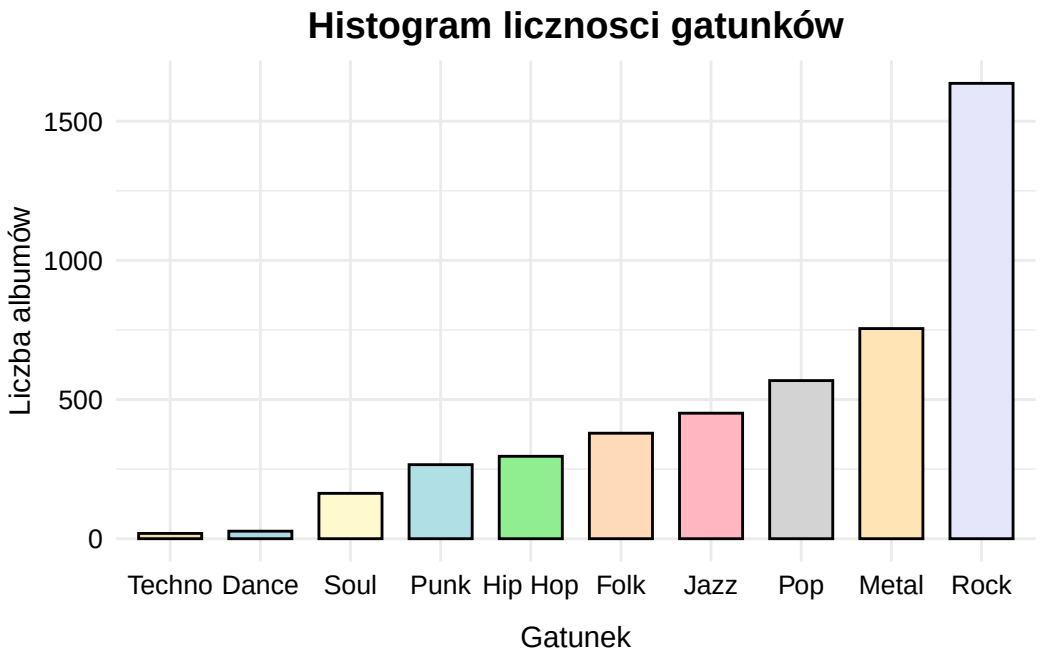
Następnie dostajemy informacje o wykonanych dla parametrów testów t-studenta, przy hipotezie zerowej o średniej równej zero. Podane zostały wartości statystyk t oraz p-wartości przeprowadzonych testów. Te drugie są rzędu 10^{-16} zarówno dla a jak i b, a zatem możemy oba parametry uznać za istotne.

Na końcu podsumowania możemy odczytać wartości błędu standardowego residuów, współczynnika determinacji w wersji standardowej R^2 i skorygowanej R_a^2 oraz wartości statystyki testowej i p-wartośc testu F, badającego istotność co najmniej jednego z parametrów (tę informację już znamy). Warto zwrócić szczególną uwagę na te ostatnie informacje. Współczynnik R^2 mówi nam, jaki procent zmienności w danych jest wyjaśniany przez model. W naszym przypadku wynosi on zaledwie 2.88. Jest to wartość bardzo niska, co oznacza że "Years" ma słabo objaśnia "Average Rating". Skorygowana wartość współczynnika determinacji R_a^2 jest bardzo zbliżona do R^2, co sugeruje, że dodanie większej liczby zmiennych do modelu, nie zmieniłoby jego mocy objaśniającej.

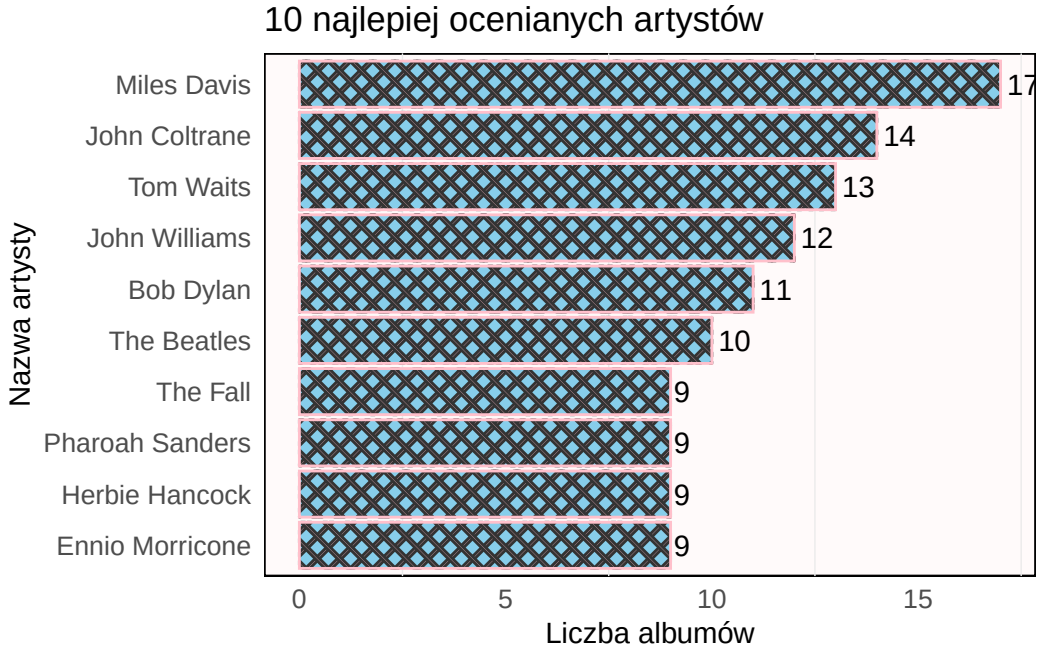
Na bazie przeprowadzonej analizy modelu regresji liniowej możemy wykluczyć obecnośc liniowej zależności między "Average Rating" a "Years". Nie możemy wykluczyć innego rodzaju zależności między nimi, choć wykres punktowy tego nie sugeruje. Należy skorzystać również z faktu, że znamy charakter naszych danych, a nic nie wskazuje na to, żeby musiała istnieć nawet przybliżona matematyczna zależność między latami wydania albumów a ocenami przydzielonymi im przez użytkowników.

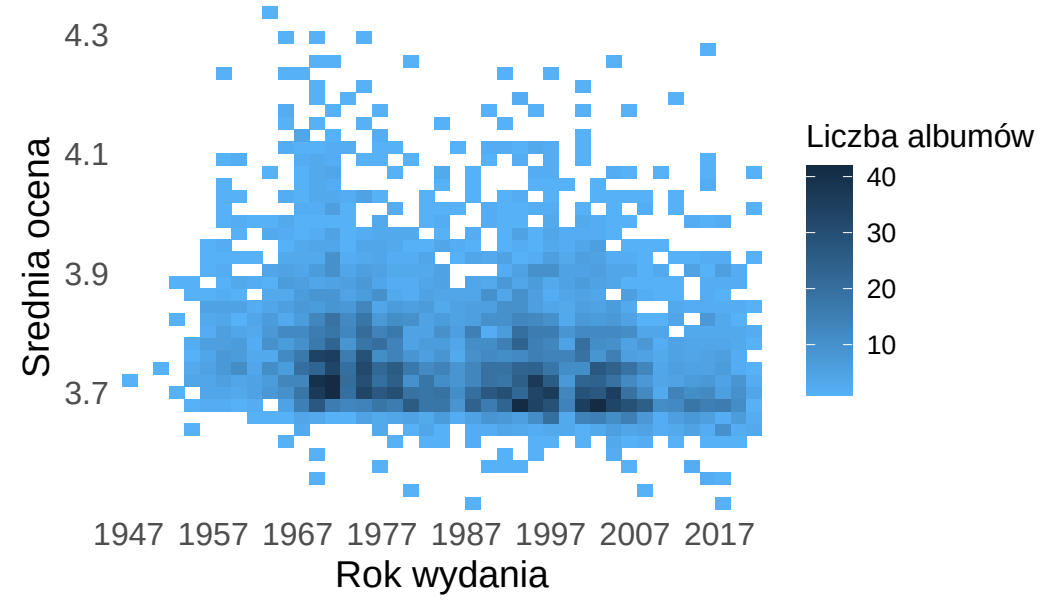
Bardziej ogólnym sposobem może okazać się sprawdzenie korelacji między "Years" a "Average Rating". Wykonajmy test korelacji Pearsona.

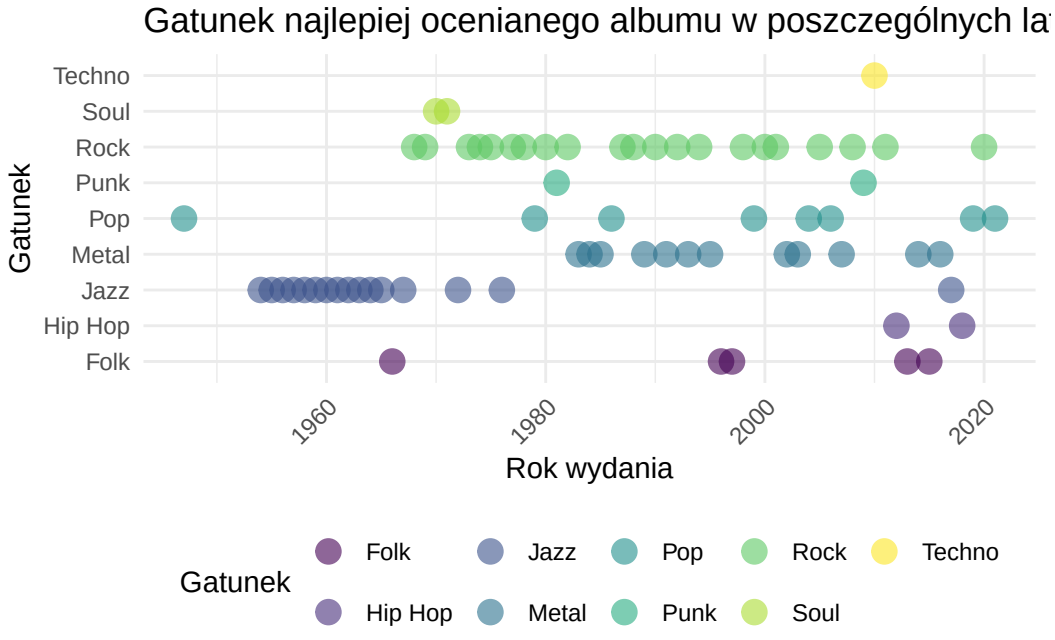
```
    Pearson's product-moment correlation

data:  rym$Years and rym$"Average Rating"
t = -12.179, df = 4998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1965704 -0.1427286
sample estimates:
       cor
-0.1697762
```

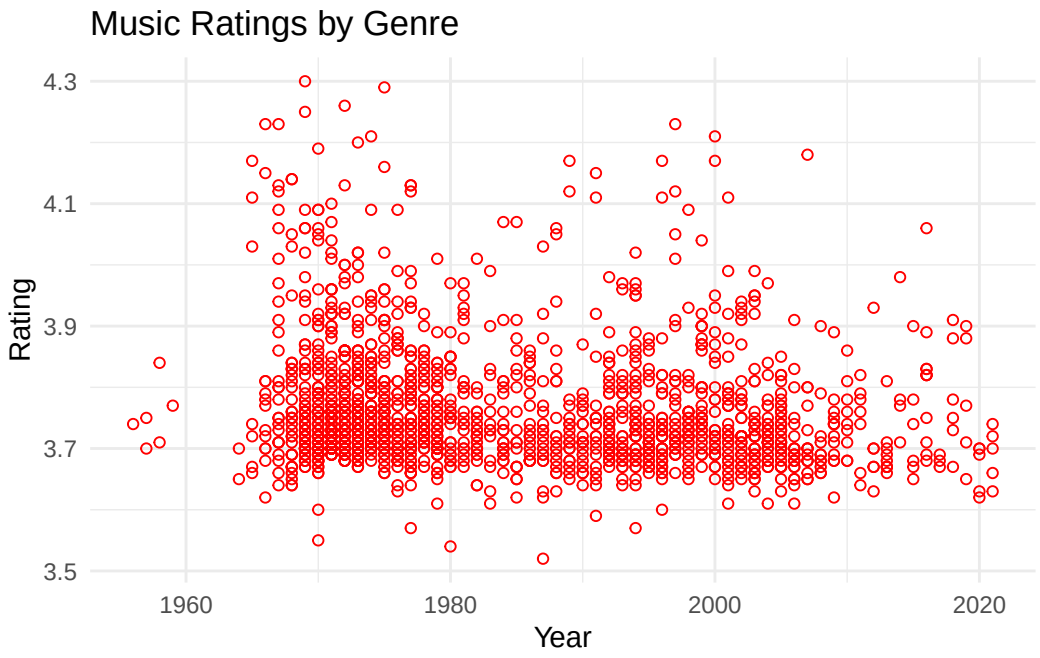


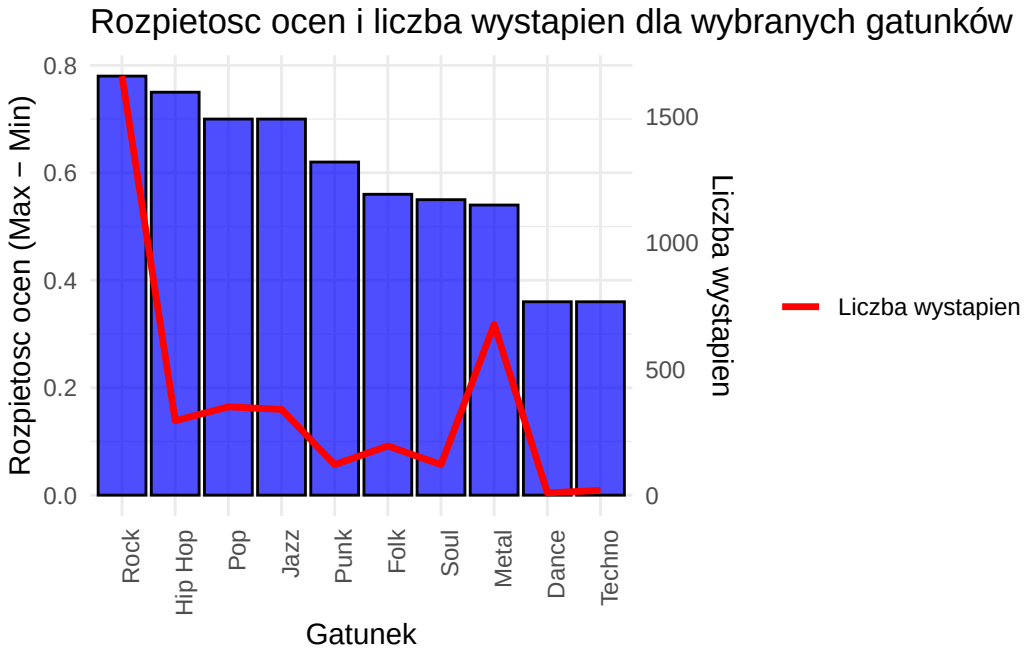

```
Warning: package 'ggpattern' was built under R version 4.4.2
```



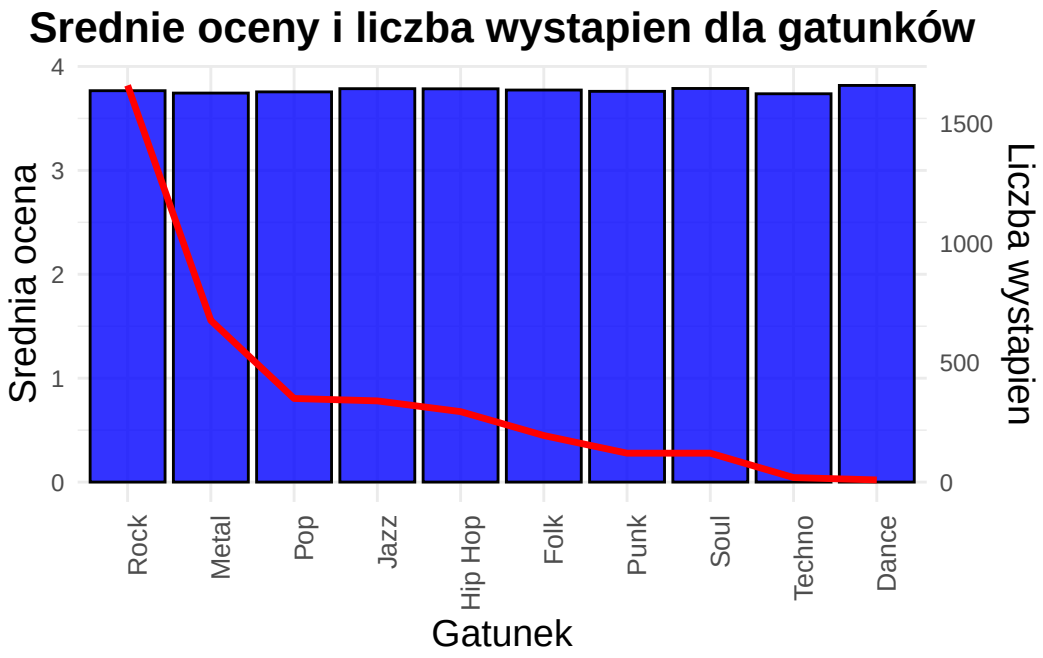

ybucja albumów: Rok wydania vs Srednia ocena
Srednia ocena
4.3
4.1
3.9
3.7
1947 1957 1967 1977 1987 1997 2007 2017
Rok wydania
Liczba albumów
40
30
20
10


Gatunek najlepiej ocenianego albumu w poszczególnych la
Gatunek
Techno
Soul
Rock
Punk
Pop
Metal
Jazz
Hip Hop
Folk
1960
1980
2000
2020
Rok wydania
Gatunek
Folk
Jazz
Pop
Rock
Techno
Hip Hop
Metal
Punk
Soul



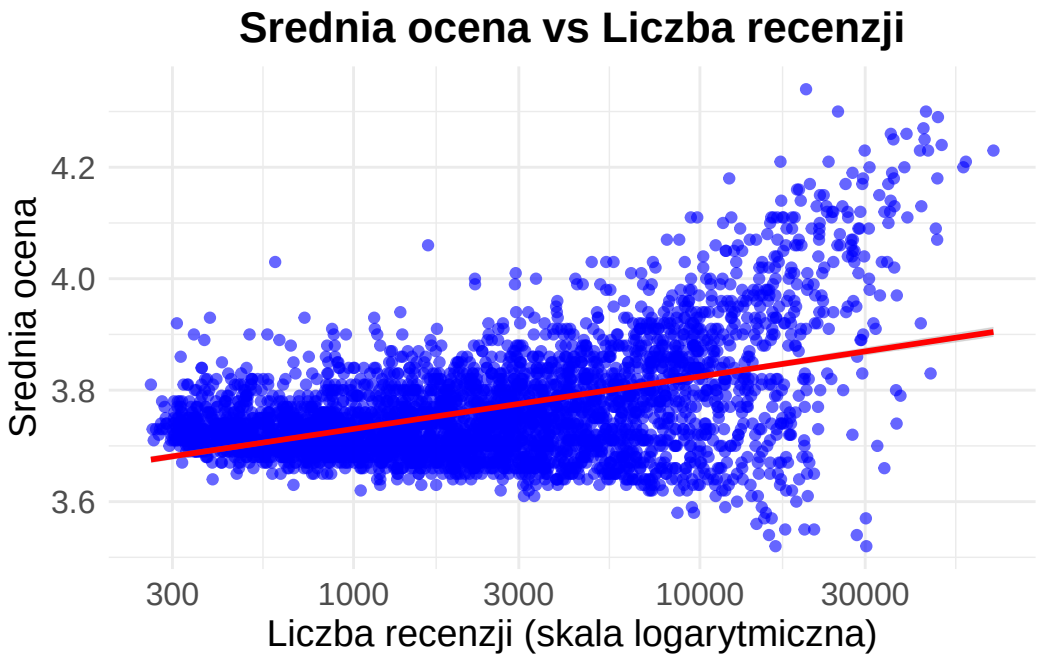



```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



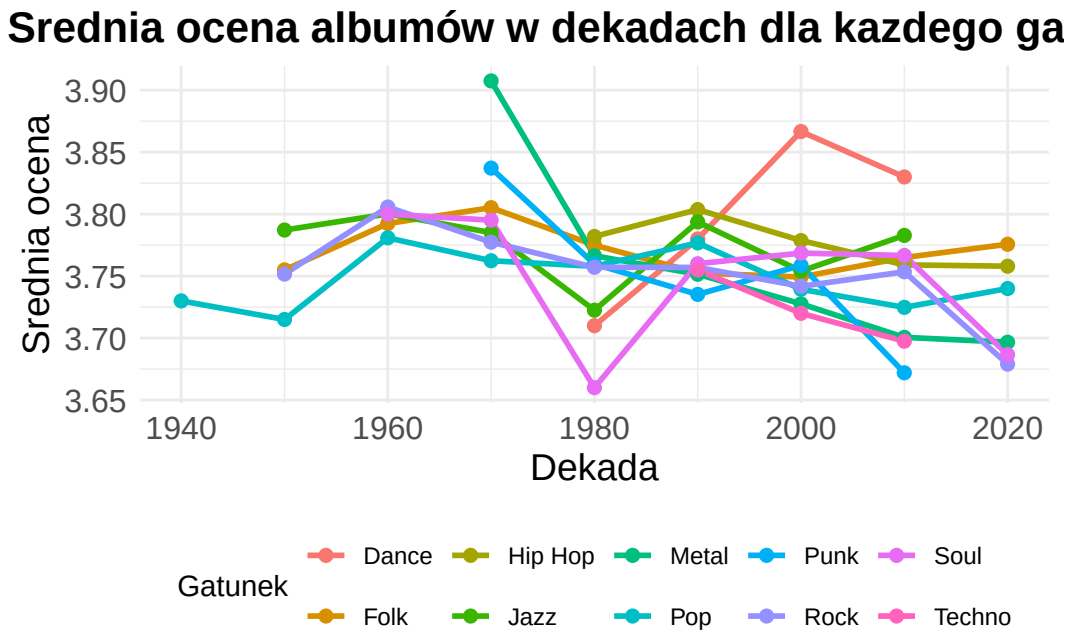

```r
xpp <- rym
xpp$"Number of Ratings" <- as.numeric(gsub(",", "", rym$"Number of Ratings"))

# Filter out missing values correctly
daxpp <- xpp %>%
  filter(!is.na(`Number of Ratings`) & !is.na(`Average Rating`))

# Updated ggplot call with correct references
ggplot(daxpp, aes(x = `Number of Ratings`, y = as.numeric(`Average Rating`))) +
  geom_point(alpha = 0.6, color = "blue") +  # Points
  geom_smooth(method = "lm", color = "red", se = TRUE) +  # Trend line (linear regression)
  scale_x_log10() +  # Logarithmic scale for number of reviews
  labs(
    title = "Średnia ocena vs Liczba recenzji",
    x = "Liczba recenzji (skala logarytmiczna)",
    y = "Średnia ocena"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
    axis.title = element_text(size = 14),
    axis.text = element_text(size = 12)
  )
```

```
`geom_smooth()` using formula = 'y ~ x'
```




```r
# Definicja podstawowych gatunków
library(dplyr)
library(tidyr)
library(stringr)
library(ggplot2)

# Definicja podstawowych gatunków
basic_genres <- c("Rock", "Hip Hop", "Pop", "Jazz", "Soul",
                  "Dance", "Techno", "Punk", "Metal", "Folk")

# Podzielenie albumów na podstawowe gatunki
filtered_data <- xpp %>%
  rowwise() %>%
  mutate(
    Basic_genres_split = list(
      basic_genres[basic_genres %>% sapply(function(genre) str_detect(Basic_Genres, fixed(ge
    )
  ) %>%
  unnest(Basic_genres_split) %>% # Rozdzielenie przypisanych gatunków na osobne wiersze
  rename(Genre = Basic_genres_split) %>% # Zmiana nazwy kolumny na 'Genre'
  ungroup()
```

```r
# Dodanie kolumny z dekadą
filtered_data <- filtered_data %>%
  mutate(Decade = floor(Years / 10) * 10)

# Obliczenie średnich ocen dla każdego gatunku w każdej dekadzie
average_ratings_by_decade <- filtered_data %>%
  group_by(Decade, Genre) %>%
  summarize(Average_Rating = mean(`Average Rating`, na.rm = TRUE), .groups = "drop")

# Wykres
ggplot(average_ratings_by_decade, aes(x = Decade, y = Average_Rating, color = Genre, group =
  geom_line(size = 1) +
  geom_point(size = 2) +
  labs(
    title = "Średnia ocena albumów w dekadach dla każdego gatunku",
    x = "Dekada",
    y = "Średnia ocena",
    color = "Gatunek"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
    axis.title = element_text(size = 14),
    axis.text = element_text(size = 12),
    legend.position = "bottom"
  )
```


Srednia ocena albumów w dekadach dla kazdego ga
Srednia ocena
3.90
3.85
3.80
3.75
3.70
3.65
1940
1960
1980
2000
2020
Dekada
Gatunek
Dance
Hip Hop
Metal
Punk
Soul
Folk
Jazz
Pop
Rock
Techno