Najlepsze albumy muzyczne

wg użytkowników RateYourMusic.com

Bartosz Łuksza, Rafał Głodek

Spis treści

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.

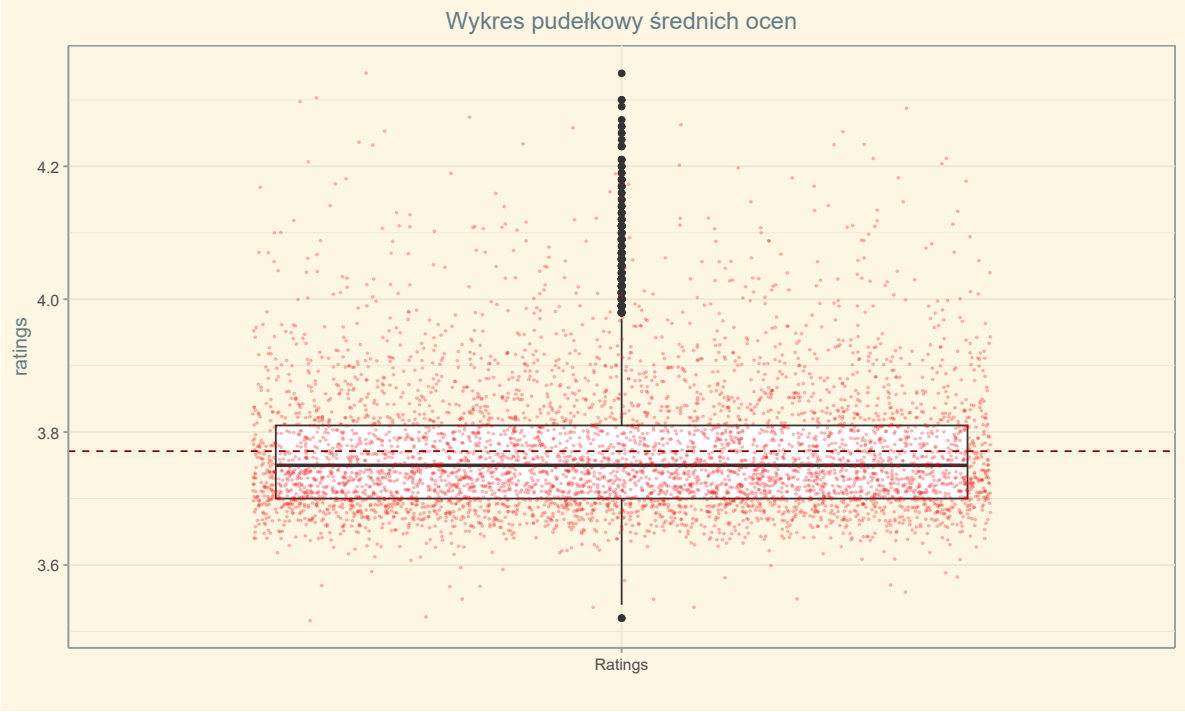
W jakich latach powstawało najwięcej "dobrych" albumów? Czy istnieje korelacja między średnią oceną użytkowników a datą wydania dzieła? Jakie są średnie ocen dla różnych gatunków muzycznych? Którzy artyści mogą się poszczycić najlepiej ocenianią dyskografią? Na te i wiele innych pytań odpowiemy w naszej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

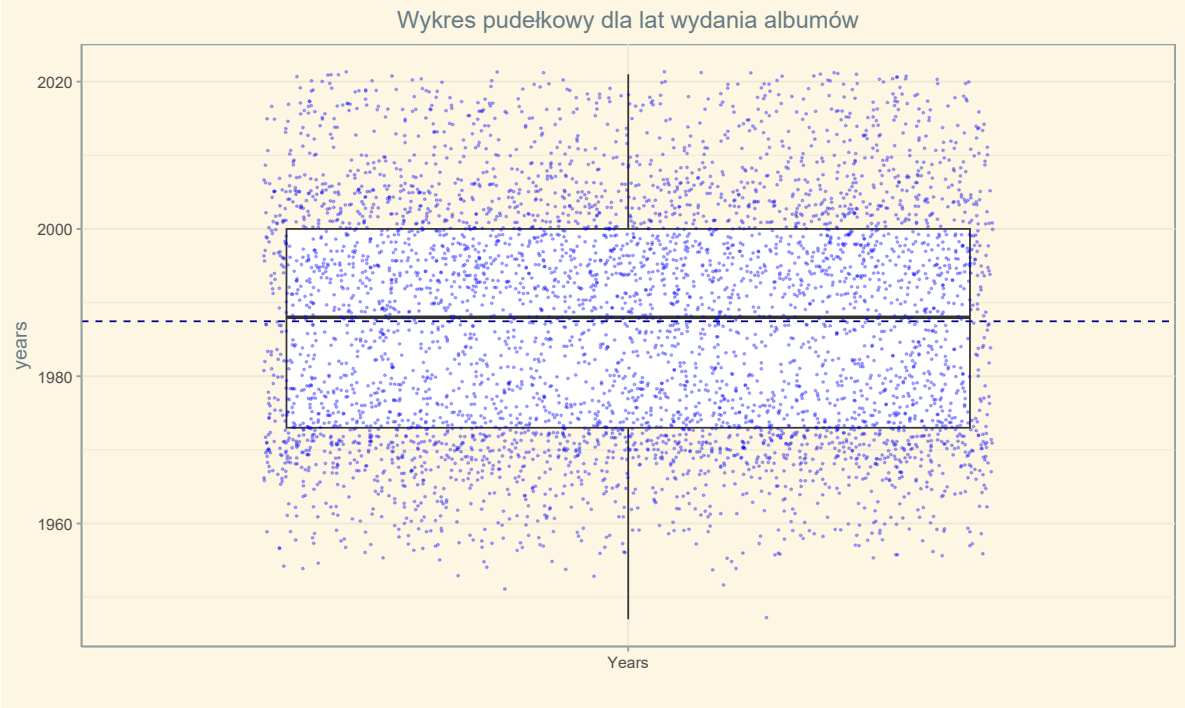


Korzystając z funkcji *boxplot.stats* wydobędziemy z wykresu najważniejsze dane. Przedstawimy je w formie tabeli

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

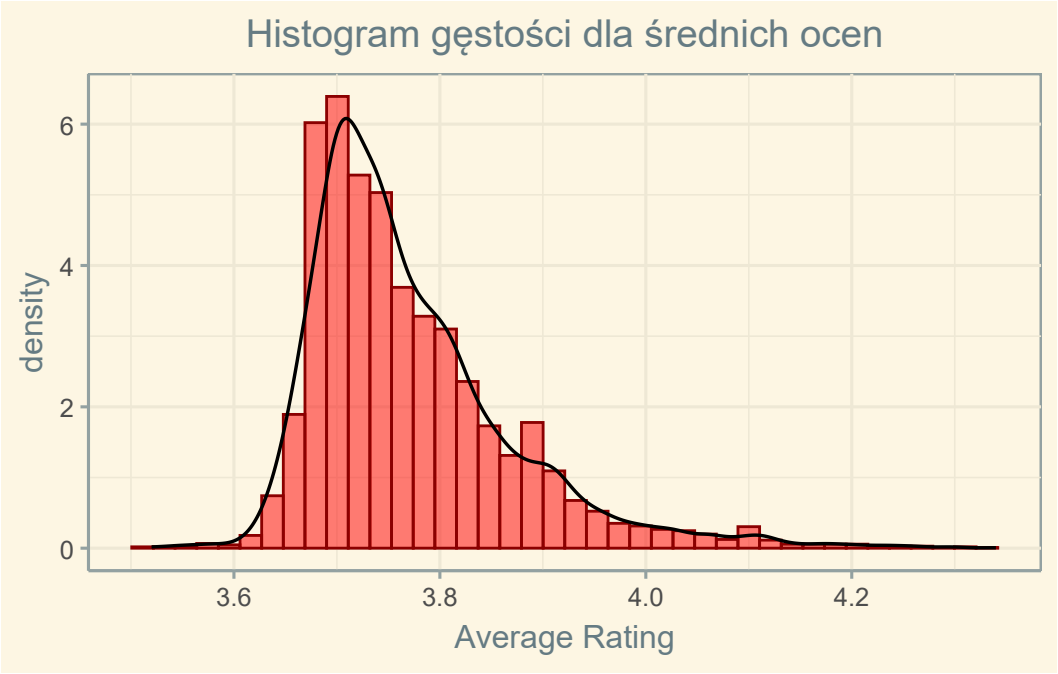


Znów wykorzystamy funkcję *boxplot.stats* i wyliczymy parametry tego wykresu pudełkowego. Zawiera je poniższa tabela.

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.

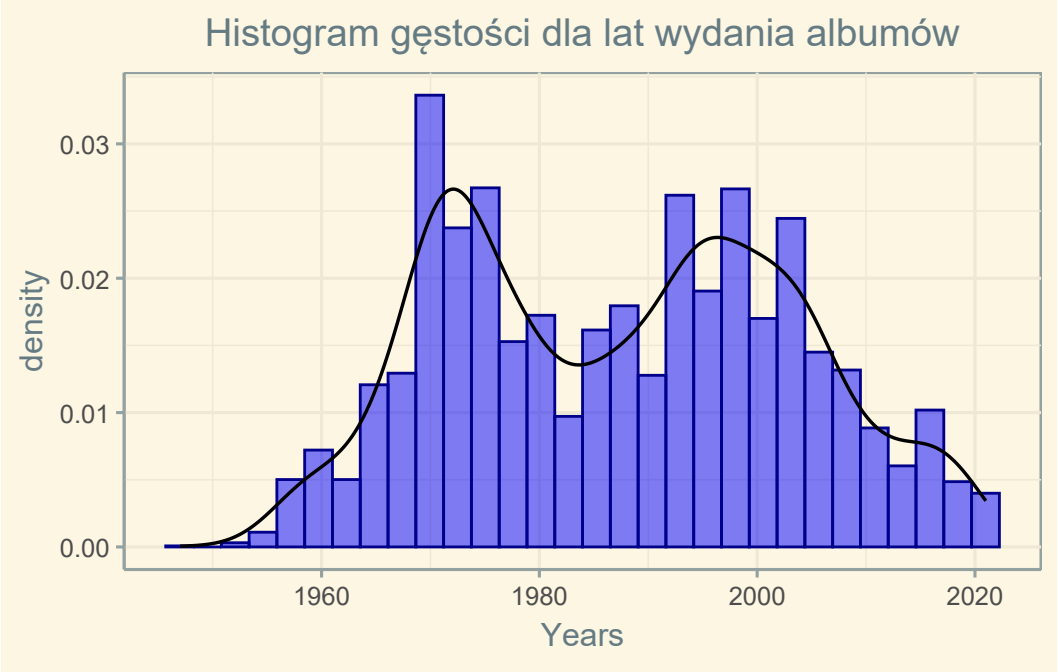
W następnym kroku dla każdej z rozważanych zmiennych wygenerujemy histogram prawdopodobieństwa *geom_histogram* i dopasujemy do niego jądrowy estymator gęstości używając *geom_density*. Wynik dla średnich ocen prezentuje się następująco



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.

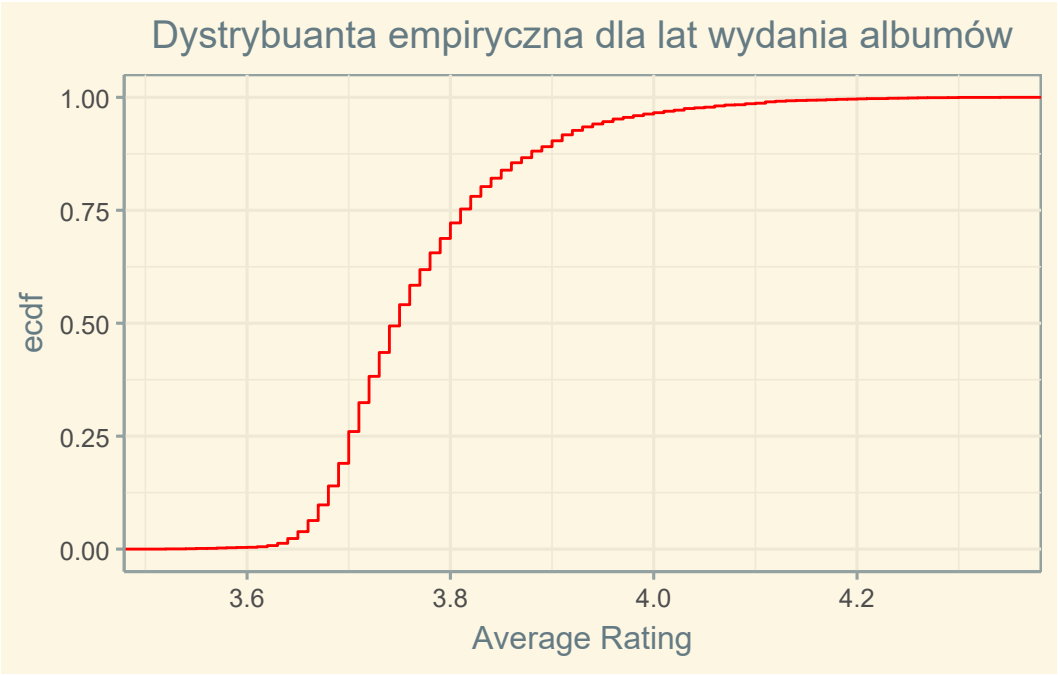
Następnie wygenerujemy analogiczny wykres dla lat wydania albumów.



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?

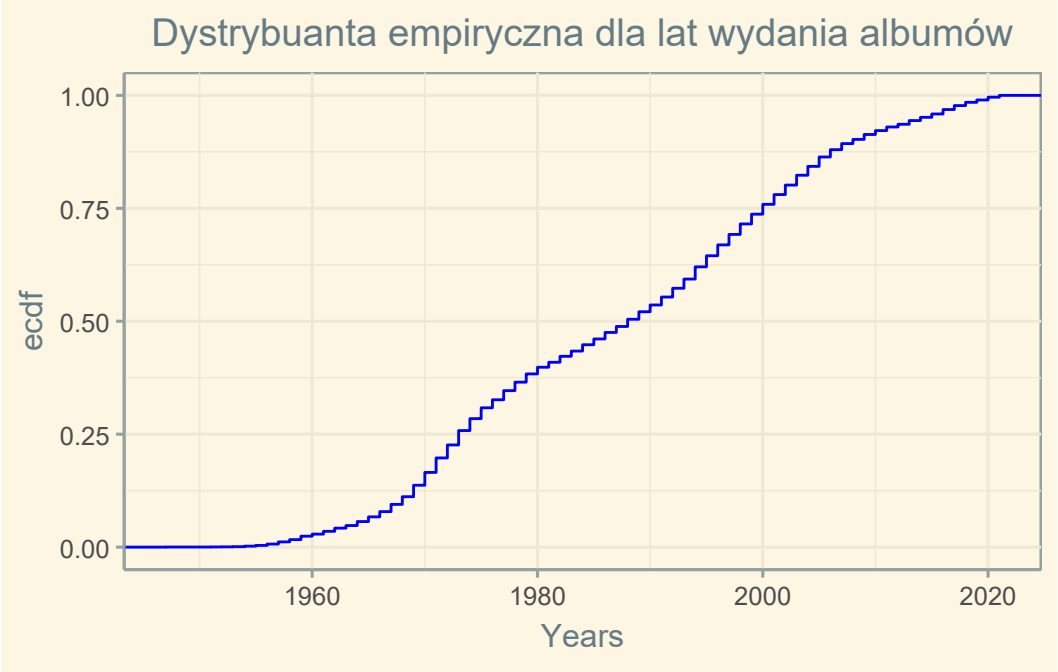
Kolejnym krokiem w analizie rozkładów rozważanych zmiennych będzie wygenerowanie wykresów ich dystrybuant empirycznych za pomocą funkcji $stat\_ecdf$.

Dla średnich ocen wykres ten prezentuje się nastepująco.



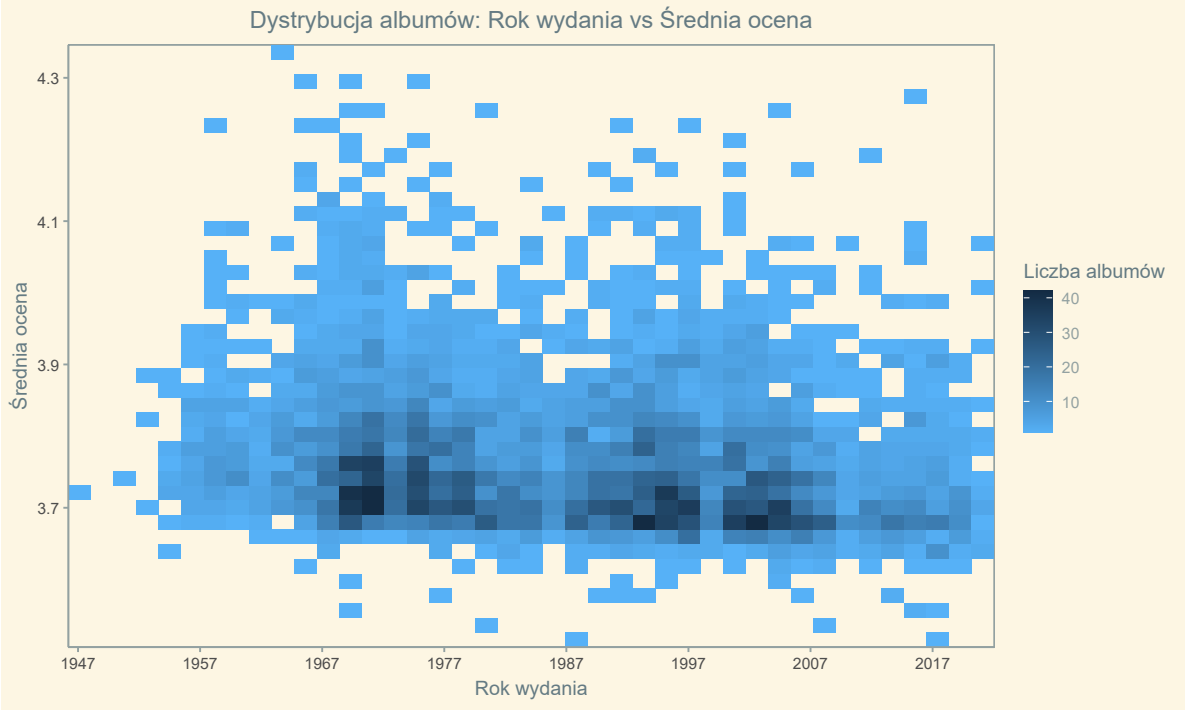
Dystrybuanta empiryczna ocen użytkowników potwierdza wnioski wyciągnięte na bazie gęstości empirycznej. Największy wzrost wartości dystrybunaty możemy zaobserwować na przedziale od 3.6 do 3.9, a zatem, gdy oceny oscylują wokół wartości średniej. Dla punktu 3.9 wartość dystrybunaty wynosi około 0.9, a 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.

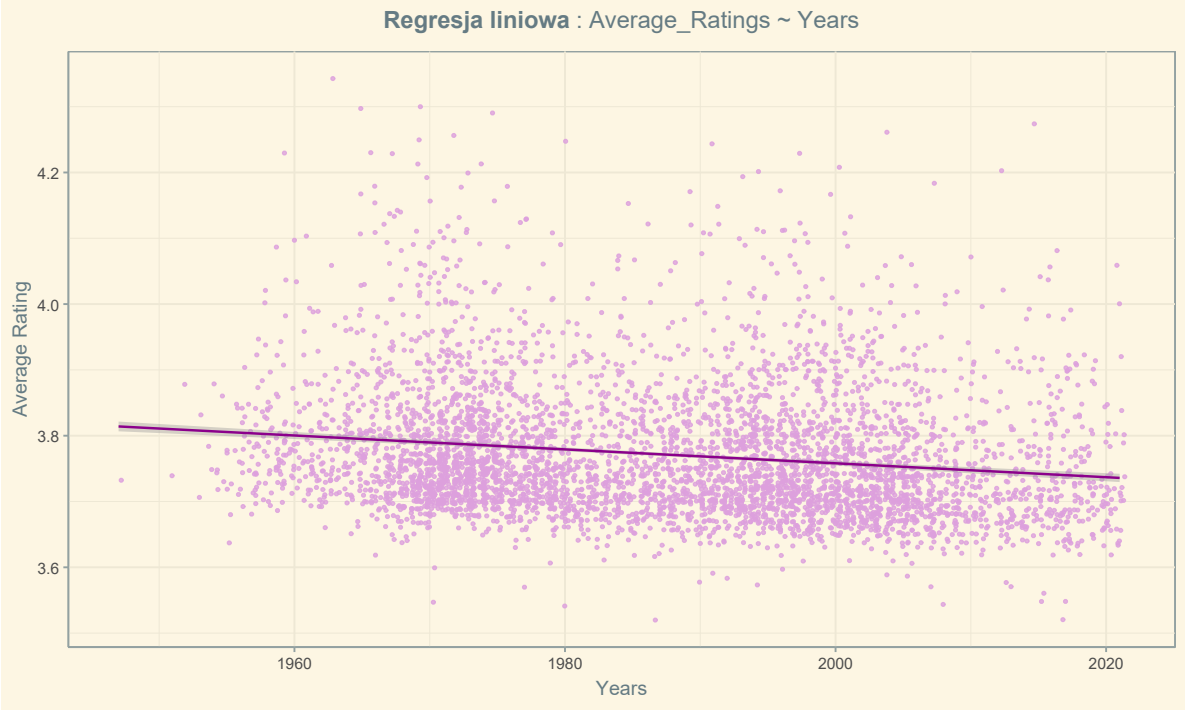
Wiemy już jak rozkładały się średnie oceny użytkowników oraz lata wydania albumów. Warto w tym miejscu zastanowić się, czy istnieje między tymi zmiennymi jakaś zależność. Poszukamy zatem zależności liniowej stosując z wbudowanej w pakiet R funkcji *lm*, wykorzystującej model regresji liniowej. Gdyby 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.



Heatmapa przedstawia zależność między rokiem wydania a średnią ocen. Z łatwością jesteśmy w stanie zauważyć zagęszczenie ilości albumów w okolicach lat siedemdziesiątych i przełomu XX wieku. Również w tych latach mamy największe zagęszczenie albumów o średniej ocenie oscylującej w okolicach 3.8. Jesteśmy w stanie zauważyć znacznie odbiegającą ilość albumów o wysokiej ocenie w latach 70-90, to tutaj powstały jedne z najlepiej ocenianych tytułów w historii — lata te możemy uznać za złotą erę muzki. We współczesnych latach spotykamy się ze znacznym spadkiem jakości albumów. Może to wskazywać na kryzys, do którego doszło w przemyśle muzycznym. Głównym czynnikiem takiego stanu rzeczy są gigantyczne pieniądze, które napłynęły do tej branży. W dzisiejszych czasach muzykę robi się pod "tik-toka" aby poszła w viral i przyniosła dla autora jak największe zyski. Jakość wypuszczanych utworów przeszła na drugi plan. Wykres potwierdza nam, że najwięcej cenionych dzieł muzycznych pochodzi ze złotego okresu, który miał kluczowy wpływ na rozwój kultury muzycznej.

Regresja liniowa średnich ocen i lat wydania albumów

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

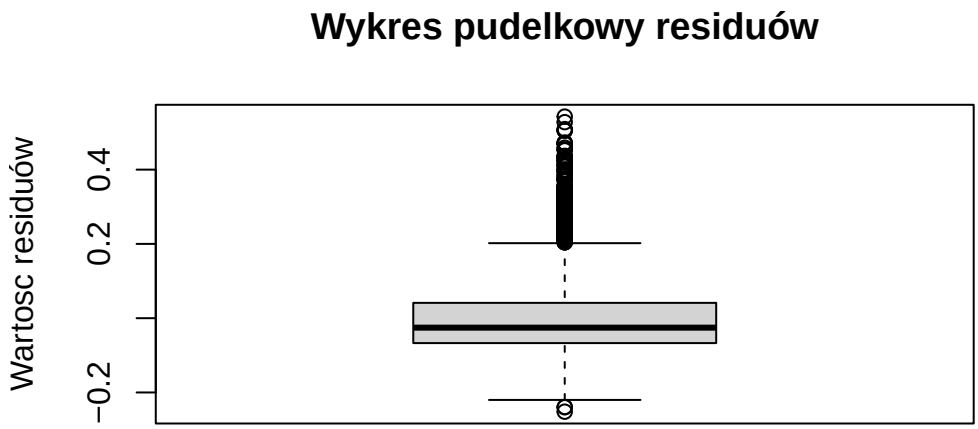


Widzimy, że punkty są dość mocno rozrzucone i nie widać między nimi szukanej liniowej zależności. Model dopasował do danych prostą regresji, lecz nie możemy spodziewać się w tym przypadku dużej skuteczności. Celem poprawnej interpretacji wykresu wydobyliśmy niezbędne informacje o regresji liniowej za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

```
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.

Test korelacji między średnimi ocenami a latami wydania albumów

Bardziej ogólnym sposobem może okazać się sprawdzenie korelacji między "Years" a "Average Rating". Wykonajmy test korelacji Pearsona, przy hipotezie zerowej mówiącej o braku korelacji między zmiennymi ($H_0 : \rho = 0$).

```
    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
```

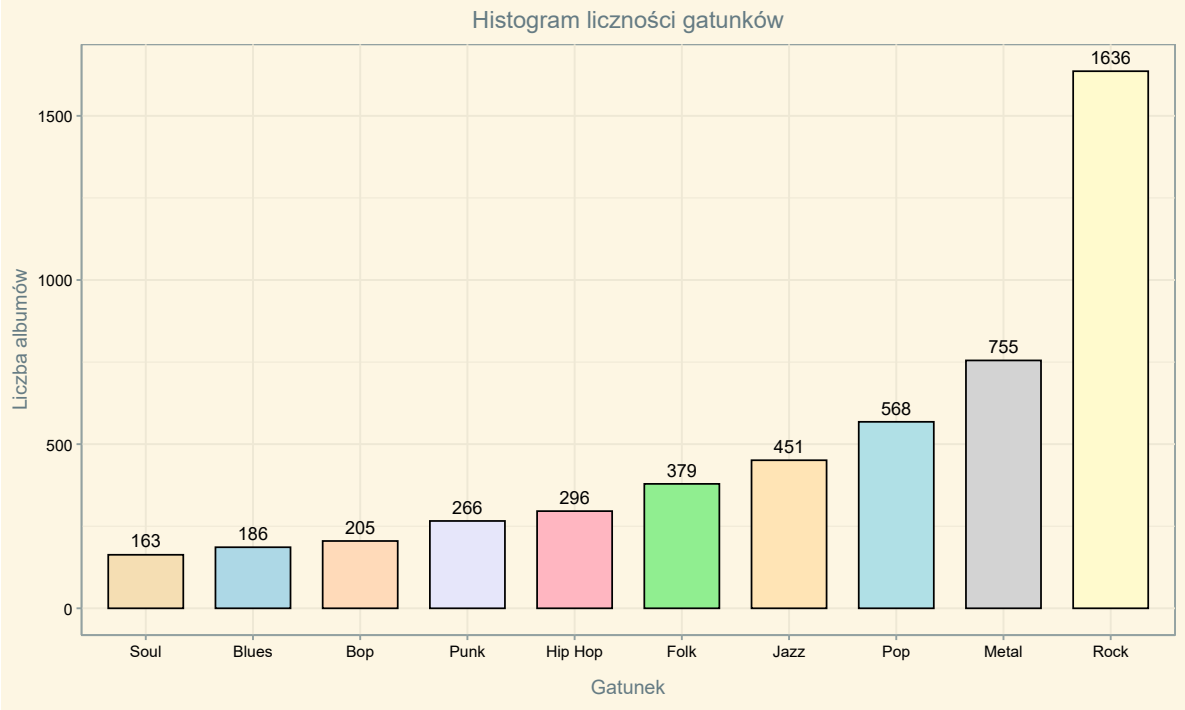
Wyniki testu na początku dostarczają nam informacje o wartości statystyki testowej ($t = -12.179$), liczbie stopni swobody ($n - 2 = 4800$) i p-wartości ($p - value < 2.2 \cdot 10^{-16}$). P-wartość jest na tyle niska, że przy wyborze dowolnej popularnej wartości poziomu istotności α (np. 0.05 lub 0.01) odrzucimy hipotezę zerową o braku korelacji. Następnie dowiadujemy się o

postaci przedziału ufności dla współczynnika korelacji na poziomie istotności $1-\alpha = 0.95$. Jest on równy $(-0.1965704, -0.1427286)$. Na tej podstawie, możemy stwierdzić, że istnieje słaba ujemna korelacja między "Years" a "Average Rating". Próbkowy estymator korelacji Pearsona dla dostarczzonych danych wynosi -0.1697762.

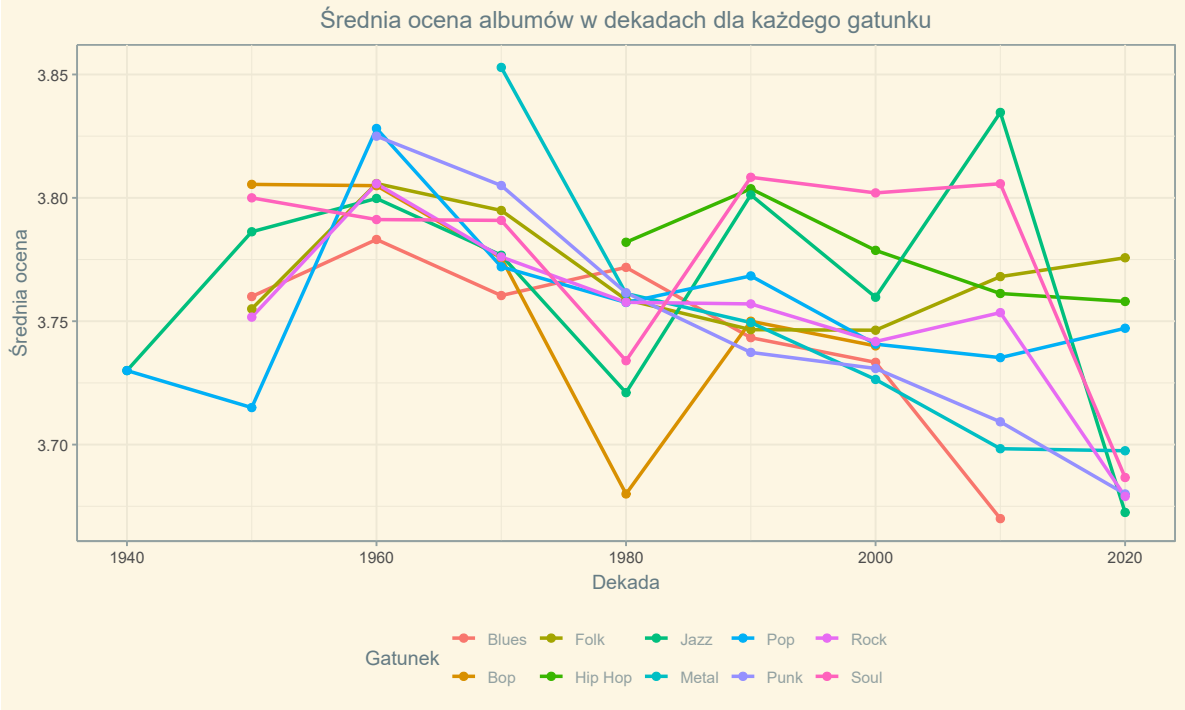
Z dotychczasowej analizy możemy wywnioskować, że zależność między latami wydania albumów a średnimi ocenami użytkowników z pewnością nie jest liniowa, ale zmienne są (słabo) ujemnie skorelowane. Nie można zatem jednoznacznie stwierdzić, że im album jest nowszy, tym niższą średnią ocen otrzymał, ale w ogólności, wraz z upływem czasu oceny użytkowników malały. Oczywiście nasze zestawienie zawiera 5000 najlepiej ocenianych dzieł współczesnej muzyki, więc nie należy także wysuwać wniosku, mówiącego że w obecnych czasach tworzona jest mniej jakościowa muzyka, niż w poprzednim stuleciu. Dobrą sugestią w kontekście interpretacji wyników może być zastąpienie pojęcia "jakości" muzyki pojęciem "ikoniczności". Wówczas znika problem nazywania gorzej ocenianych albumów w rozważanym rankingu jako "Słabych" — w końcu album z najniższą oceną 3.52 dalej jest pieciotysięcznym najlepiej ocenianym dziełem muzyki. Zamiast tego można je uznać za mniej ikoniczne, czyli takie, który odcisnęły mniejsze piętno na historii muzyki. Używajac tej intepretacji, możemy stwierdzić, że w ogólności, nowsze dzieła muzyczne są uznawane za mniej istotne kulturowo. Według nas, ten trend może mieć dwie niewykluczające się przyczyny. Najprostsze, i często przyjmowane przez słuchaczy lubujących się w dwudziestowiecznej muzyce, wyjaśnienie mówi, że muzyka (a nawet cała kultura) swoje lata świetności ma już za sobą i wytwory współczeności nigdy nie dorównają klasycznym poprzednikom. Wizja ta jest dość pesymistyczna, ale z pewnych względów nie można jej wykluczyć. Niepodważalnym faktem jest to, że muzyka faktycznie przeżyła największy przełom w poprzednim stuleciu. Liczba gatunków namnożyła się do poziomu, w którym zliczenie ich wszystkich jest praktycznie niemożliwe, a cyfrowe przetwarzanie dźwięku otworzyło niemalże nieskończone możliwości dla twórców. Był to również pierwszy moment, kiedy światowa publika uświadomiła sobie jak wielki jest wpływ muzyki na różne aspekty ich życia codziennego — w końcu badania wykazują, że muzyka świąteczna w sklepach w okresie bożonarodzeniowym zwiększa wydatki klientów o 12. Drugie wyjaśnienie, które proponujemy, pozostawia słuchaczom więcej nadziei na przyszłość, ponieważ naszym zdaniem nowe wydania nie przeszły jeszcze próby czasu, a jest ona niezbędna żeby zaklasyfikować je jako ikoniczne. Krytycy muzyczni bardzo często odnoszą proces tzw. "starzenia się" do albumów muzycznych. Dzieło, które "dobrze się zestarzało" pozostaje aktualne mimo upływu lat, zarówno pod względem sonicznym, jak i zawartych w nim kontekstów kulturowych. Takie albumy zaliczane są do grona klasyków gatunkowych. W opozycji do nich stoją krążki, które "źle się zestarzały" z przyczyn niewystarczającej głębi lirycznej lub mało wyróżniającej się warstwy muzycznej. Czytelnik z pewnością sam jest w stanie przywołać przykład utworu lub albumu, który w krótkim czasie przeszedł drogę od światowego hitu do dzieła zapomnianego. W dzisiejszych czasach to zjawisko jest bardzo popularne, ponieważ rozwój internetu sprzyja dynamicznemu rozprzestrzenianiu się trendów w kulturze, które, wielokrotnie powielane, szybko stają sie nieaktualne.

Najpopularniejsze gatunki muzyczne

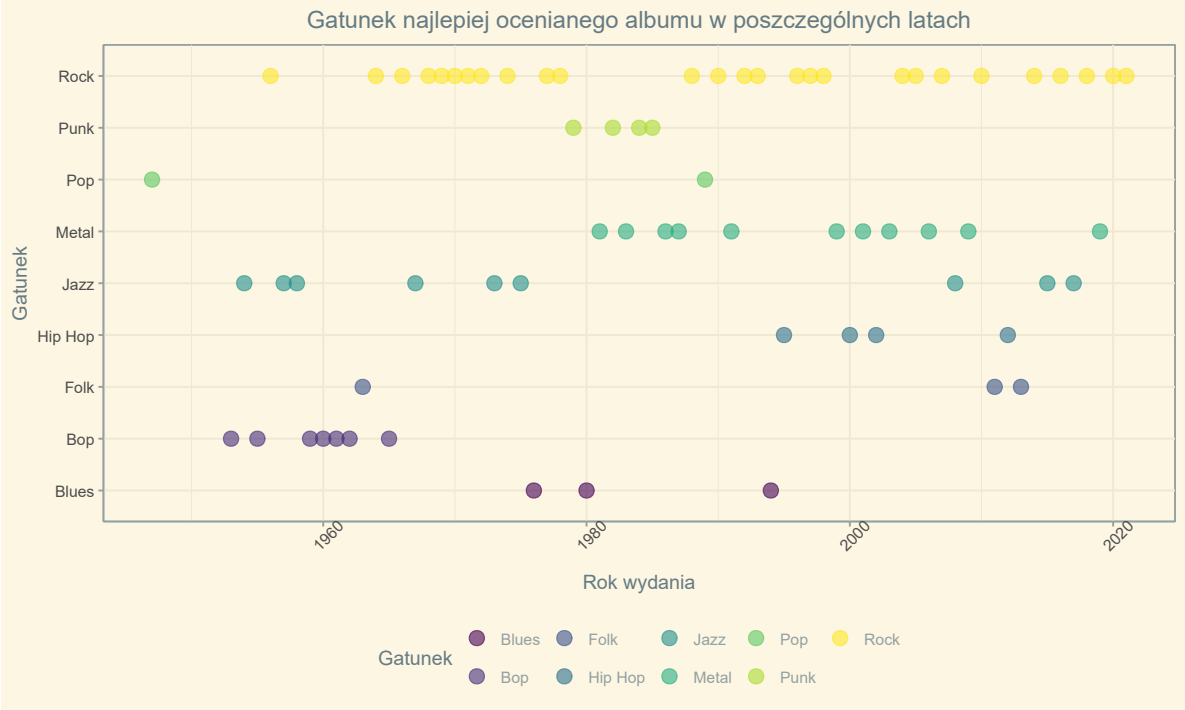
Celem dokładniejszego zagłębienia się w temat i dokonania obszerniejszej analizy, wyodrębniliśmy z pobranego pliku informacje o gatunkach wymienionych w zestawieniu albumów oraz ich autorach. Ze względu na dużą różnorodność gatunków, wzięliśmy pod uwagę 10 najpopularniejszych z nich. Poniższy histogram przedstawia jak dużo spośród wszystkich 5000 albumów należących do rankingu, jest zaklasyfikowanych do każdego z gatunków.



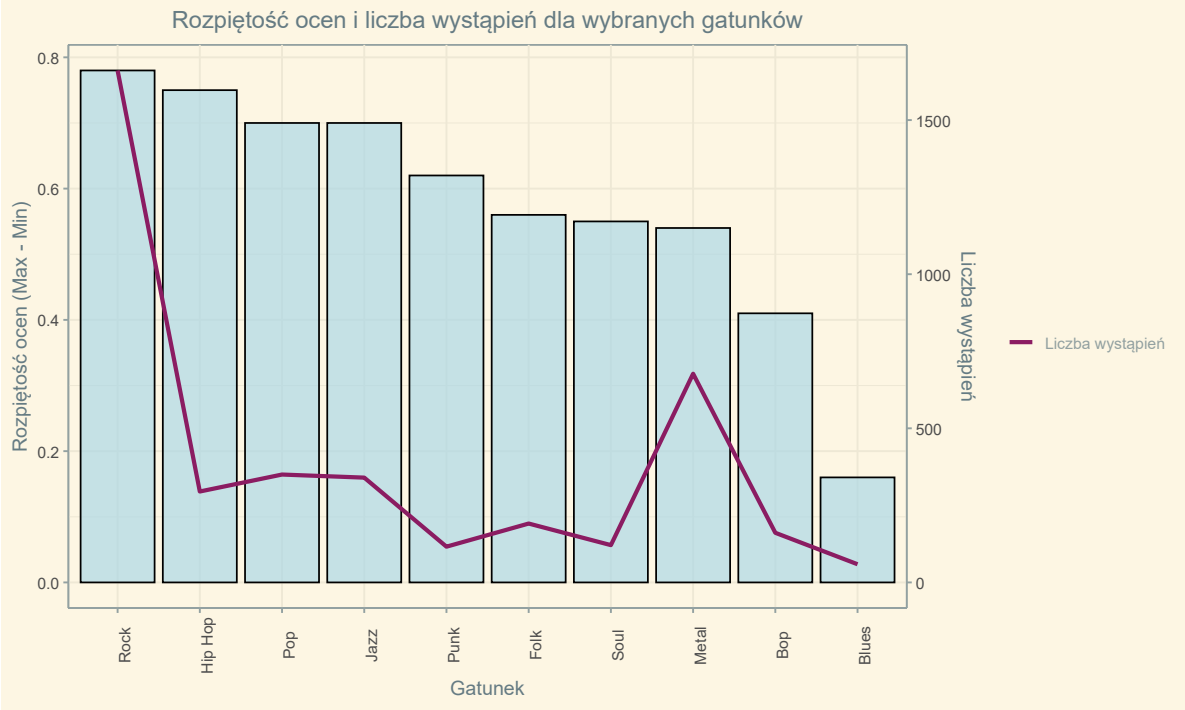
Jak widzimy na wykresie, najliczniejszą grupę spośród wszystkich gatunków ujętych w zestawieniu, stanowią albumy rockowe. Ich liczebność przewyższa ponad dwukrotnie drugi najliczniejszy gatunek - metal. Najrzadziej w rankingu występują "krążki" zaliczane do techno i dance, czyli elektronicznej muzyki tanecznej.



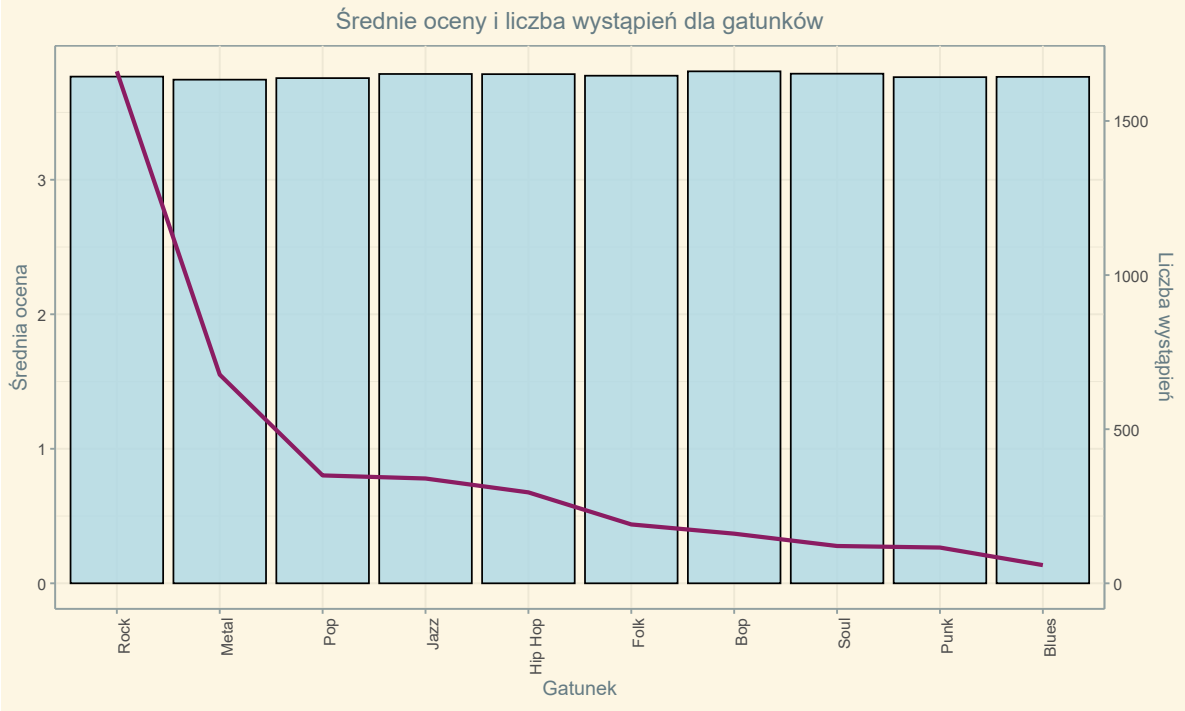
Na wykresie punktowym widzimy jak radziły sobie gatunki w poszczególnych dekadach. Na tej wizualizacji również jesteśmy w stanie zauważyć złote lata muzyki o najlepszym odbiorze przypadające na okres lat 60-70 oraz na lata 90 ubiegło stulecia. Widzimy duże załamanie jakości muzyki w latach 80, które przypadły na recesję i załamanie gospodarcze w USA. Lata 90 były ostatnim dobrym okresem w muzyce. W XXI wieku możemy zobaczyć spadek jakości muzyki zakwalifikowanej do gatunków: blues, metal, rock i jazz.



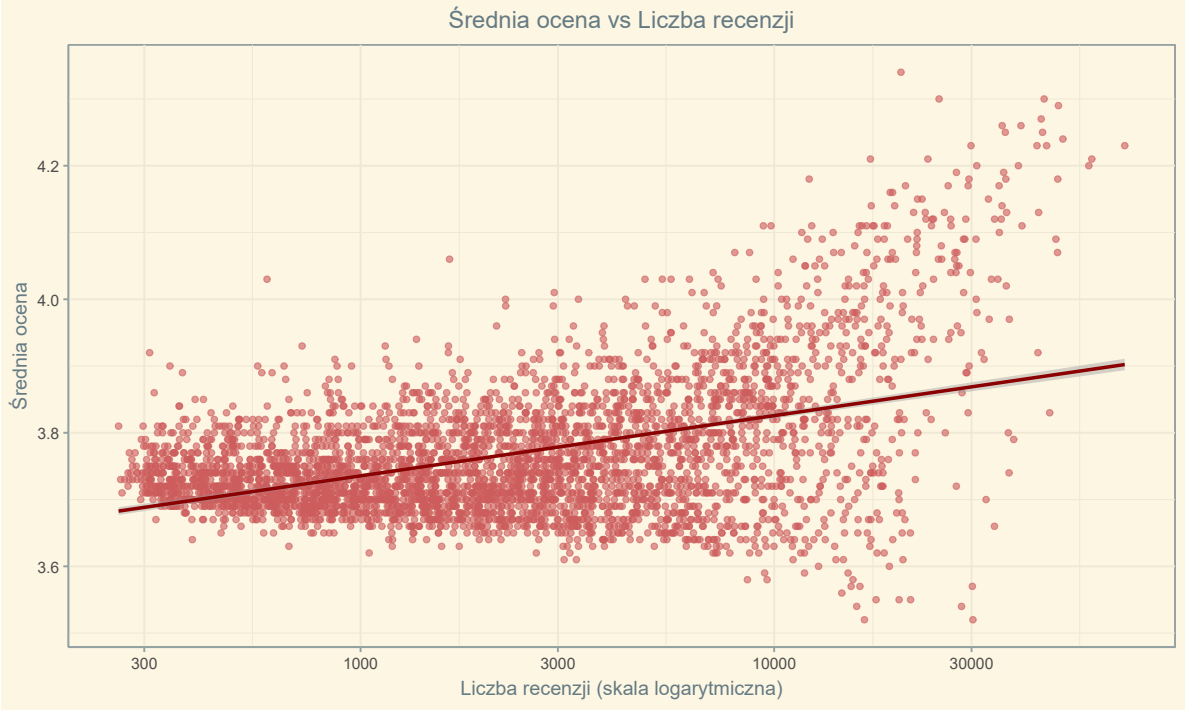
Wykres punktowy pokazuje najlepiej oceniane albumy z danego roku w podziale na gatunki. Na wykresie widzimy, że jazz w pełni zdominował lata 50. i 60. ubiegłego wieku. Z poprzedniego wykresu wiemy już, że w tym okresie walczył o dominację tylko z rockiem i muzyką folk. W kolejnych dekadach to rock wyszedł na prowadzenie i utrzymał się na nim do początku lat 90. W tym okresie widzimy walkę o dominację dwóch gatunków: rocka i metalu, które to konkurowały ze sobą jeszcze w drugim millenium. Ostatni badany okres był rywalizacją między popem a wschodzącym hip-hopem. Przejście od dominacji rocka do bardziej nowoczesnych gatunków odzwierciedla ewolucję gustów muzycznych i zmieniające się trendy w przemyśle muzycznym.



Wykres przedstawia zależność między średnimi ocenami albumów a ich liczbą, w podziale na różne gatunki muzyczne. Wynika z niego, że rozpiętość ocen nie jest skorelowana z liczbą wystąpień danego gatunku. Jeśli przyjrzymy się uważnie wcześniejszemu wykresowi, będziemy w stanie wyciągnąć inną zależność — gatunki o największej zmienności ocen charakteryzują się również najdłuższym "stażem" w naszym rankingu. Duża rozpiętość spowodowana jest zmianą dominacji gatunków na przestrzeni lat, która wynika ze zmieniających się gustów muzycznych. Długi staż rocka zawiera zarówno momenty, w których był on hegemonem, jak i te, w których zainteresowanie tym gatunkiem było praktycznie znikome. Spowodowało to dużą zmienność w ocenie, za to jego długa dominacja sprawiła, że pod względem ilości nie ma sobie równych. Metal również potwierdza nam tę tezę. Jego panowanie dało mu dużą ilość wystąpień, jednak jego popularność szybko przeminęła. Coraz mniejsza ilość albumów tego gatunku dostawała się do naszego zestawienia, nie pozwalając na obniżenie rankingu, który zdobył podczas złotych lat panowania. Drugą największą zmiennością może pochwalić się hip-hop. Wiemy, że jego dominacja przypada na okres ubiegłej dekady, jednak już od lat 80. był on z nami i osiągał wyjątkowo wysokie oceny. Co ciekawe, były to też lata, w których — pod względem średniej ocen — nie miał sobie równych. Jak większość gatunków w latach dwutysięcznych, jego oceny mocno się pogorszyły, jednak pokrywa się to z okresem, w którym dopiero budował swoją pozycję, a ilość albumów w zestawieniu ciągle rosła. Dzięki temu uplasował się na drugiej pozycji pod względem zmienności. Reszta gatunków również uzyskała całkiem wysoką zmienność. Jest ona wypadkową krótkiego okresem popularności i późniejszego szybkiego spadku zainteresowania, lub przeciwnie, długiej historii w rankingu.



Wykres przedstawia nam zależność między średnią ocen a liczbą wystąpień dla gatunku. Analizując wykres jesteśmy w stanie jasno stwierdzić brak korelacji między tymi zmiennymi. Nakładające się na siebie lepsze i gorsze wyniki dały nam średnią w okolicach 3,8. Wynik ten zgadza się z wcześniejszym histogramem.

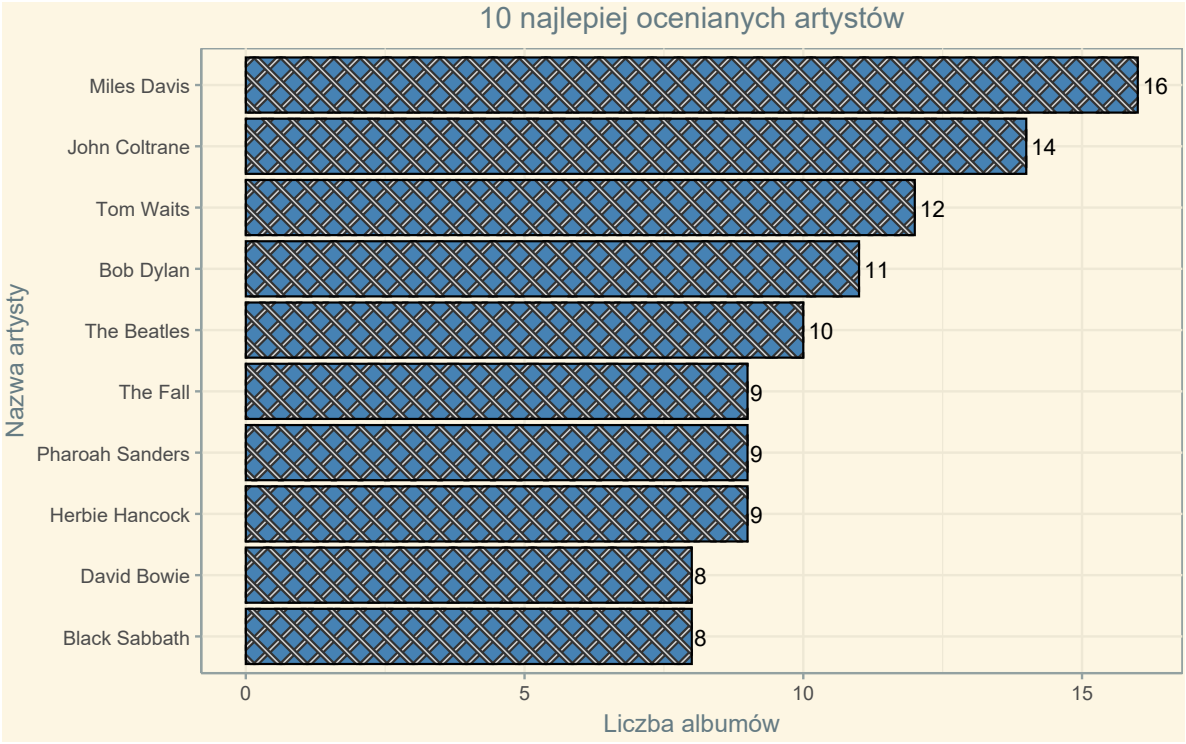


Wykres wizualizuję nam zależność między średnią ocen a liczbą recenzji z nałożoną prostą regresji. Analizując wykres, widzimy duże zagęszczenie punktów w przedziale od 300 do 3000 wystawionych recenzji. Wyniki w tym obszarze są do siebie zbliżone i oscylują w okolicach 3,7. Wraz ze wzrostem ilości recenzji widzimy dużą polaryzację w średniej ocenie. To zjawisko spowodowane jest odczuciami słuchaczy do danego albumu. Recenzenci potrafili chętniej oceniać tylko te albumy, które wywołały w nich skraje uczucia, zarówno te pozytywne, jak i negatywne. Albumy w przedziale od 10000 do 30000 recenzji miały większą liczbę ocen pozytywnych niż negatywnych. Jednak to najczęściej oceniane albumy zdobyły bezkompromisowo najwyższe oceny. Płyty z najlepszym odbiorem przyciągały ludzi, którzy specjalnie głosowali tylko dla nich. Widzimy więc, że średnia ocen jest mocno skorelowana z liczbą recenzji, dając częściej ocenianym albumom zdecydowanie wyższe noty.

Najlepiej ocenani wykonawcy

Kolejny ciekawy wniosek wysuwa się, jeśli policzymy unikalne wartości dla zbioru zawierającego artystów albumów zawartych w rankingu. Okazuje się, że jest ich jedynie 2090. A zatem istnieją wykonawcy, których dyskografie szczególnie przypadły do gustu użytkownikom serwisu. Musieli bowiem stworzyć co najmniej dwa albumy, które znalazły się w tym zestawieniu. Żeby sprawdzić, którzy artyści należą do grona “ulubionych”, policzyliśmy liczbę wystąpień każdego z nich, przyporządkowaliśmy te wartości do konkretnych twórców i posortowaliśmy nasz zbiór

po liczbie wystąpień w kolejności malejącej. Poniżej znajduje się zestawienie dziesięciu najlepiej ocenianych artystów.



Jak możemy zauważyć na wykresie, najczęściej powtarzającym się arystą w naszym rankingu jest amerykański jazzman, Miles Davis. Aż siedemnaście albumów z jego dyskografii zostało zakwalifikowanych do czołówki światowej muzyki. Na drugim miejscu uplasował się John Coltrane, z czternastoma albumami. Coltrane i Miles często ze sobą współpracowali. Stworzyli razem sześć albumów kooperacyjnych osadzonych w gatunku jazz. Gatunek ten jest najczęściej występującym w tym zestawieniu. Jego przedstawicielami, oprócz czołowej dwójki są jeszcze: Pharoah Sanders, Herbie Hancock oraz twórcy muzyki filmowej — John Williams i Ennio Morricone. W najlepszej dziesiątce znalazło się jeszcze trzech reprezentantów muzyki rockowej — Bob Dylan, The Beatles i The Fall, a także pojedyczny twórca bluesowy, Tom Waits.

Podsumowanie

Analiza danych pozyskanych z serwisu RateYourMusic.com umożliwiła nam poznanie statystyk dotyczących muzycznych albumów, i dostarczyła odpowiedzi na zadane przez nas pytania. Wygenerowanie wykresów pudełkowych i dokładne opisanie ich parametrów zapewniły

konkretną bazę pod analizę rozkładów zmiennych “Years” i “Average Rating”, którą następnie rozszerzyliśmy o histogramy pradopodobieństwa z nałożonymi jądrowymi estymatorami gęstości i wizualizację empirycznych dystrybuant. Następnie pochyliliśmy się nad problemem zależności między latami wydania albumów, a ich średnimi ocenami. Metoda regresji liniowej okazała się w tym przypadku mało efektywna, dzięki czemu byliśmy w stanie stwierdzić, że relacji tej nie powinniśmy rozpatrywać w kategoriach liniowości. Przeprowadzony dalej test korelacji Pearsona wykazał, że między rozważanymi zmiennymi występuje słaba ujemna korelacja. Jako możliwe przyczyny zjawiska, zaproponowaliśmy teorię o kryzysie współczesnego rynku muzycznego oraz bardziej optymistyczną wersję — niwystarczającej próby czasu. W dalszej części przeszliśmy do analizy albumów, z uwzględnieniem zmiennej kategorycznej, w postaci gatunków muzycznych. Z bogatego i zróżnicowanego zbioru wyodrębniliśmy dziesięć najbardziej popularnych gatunków. Histogram ich liczności wykazał, że najwięcej albumów spośród najlepszych pięciu tysięcy, należy do gatunku rock, a ich liczba przewyższa ponad dwukrotnie metal, zajmujący drugą pozycję. (analiza gatunkow c.d)

Na poczatku rozważań zastanawialiśmy się także, którzy artyści mogą poszczycić sie najlepiej ocenianą dyskografią. W tym celu stworzyliśmy zestawienie dziesięciu najczęściej występujących artystów w naszym zestawieniu. Okazało się, że ten zaszczytny tytuł przypadł amerykańskiemu jazzmanowi — Miles’owi Davis’owi, który w swojej karierze wydał 60 studyjnych albumów, spośród których aż 16 znalazło się w naszym rankingu. Wśród czołowej dziesiatki znalazło się sześciu przedstawicieli jazzu (w tym dwóch twórców muzyki filmowej), trzech reprezentatów rocka oraz jeden wykonawca bluesa.

\end{document}