IUM etap pierwszy

Marcin Połosak, Antek Grajek

listopad 2024

1 Definicja problemu biznesowego

Platforma streamingowa dąży do zwiększania zaangażowania użytkowników i wyróżniania się na tle konkurencji, oferując innowacyjne funkcje dostosowane do potrzeb słuchaczy. Jednym z kluczowych wyzwań w tym kontekście jest zdolność do przewidywania, które utwory zyskają na popularności w najbliższym czasie, jeszcze zanim staną się hitami. Dzięki takiej funkcjonalności platforma może nie tylko przyciągnąć nowych użytkowników, ale także zwiększyć satysfakcję obecnych, dostarczając im spersonalizowane rekomendacje trafnie odzwierciedlające ich preferencje oraz trendy muzyczne.

1.1 Jawna definicja problemu biznesowego

Celem biznesowym projektu jest opracowanie systemu, który z tygodniowym wyprzedzeniem będzie przewidywał listę najpopularniejszych utworów.

1.2 Kryterium sukcesu

Kluczowym wskaźnikiem sukcesu projektu będzie zdolność modelu do wygenerowania listy 10 lub 20 utworów, które znajdą się na rzeczywistej liście najpopularniejszych piosenek w kolejnym tygodniu, przy czym wszystkie proponowane utwory muszą znajdować się w rankingu top 50 popularności z przyszłego tygodnia.

Oczekuje się również, że model będzie wykazywał odporność na dynamicznie zmieniające się preferencje użytkowników oraz ewoluujące trendy muzyczne, co umożliwi długoterminowe wykorzystanie rozwiązania przy minimalnej potrzebie utrzymania.

1.2.1 Analityczne Kryterium sukcesu

Analityczne kryterium sukcesu projektu zostanie określone na podstawie dokładności przewidywań modelu. Model uznaje się za skuteczny, jeśli co najmniej 90 procent utworów z przewidywanej listy 10 lub 20 utworów faktycznie znajdzie się w top 50 najpopularniejszych piosenek w kolejnym tygodniu. Poprawnie przewidziane utwory to te, które znajdują się zarówno na wygenerowanej przez model liście, jak i w rankingu top 50 przyszłego tygodnia.

Dodatkowo, aby uznać model zaawansowany za analitycznie skuteczny, jego wyniki muszą przewyższać wyniki modelu bazowego, stanowiącego punkt odniesienia. Porównanie dokładności przewidywań z modelem bazowym pozwoli na obiektywną ocenę jakości zastosowanego podejścia.

1.3 Zadania modelowania

Zadanie opierać będzie na przewidywaniu wartości popularności, którą osiągnie utwór, na podstawie tendencji wzrostowej uzyskanej z odczytu sesji W ramach projektu przewidziane jest opracowanie dwóch głównych wariantów modeli: bazowego oraz zaawansowanego, które różnią się podejściem do predykcji.

Model bazowy Model bazowy nie wykorzystuje technik uczenia maszynowego. Działa na prostych zasadach:

- Playlisty są tworzone na podstawie najpopularniejszych piosenek z poprzedniego tygodnia.
- Nie uwzględnia analizy cech utworów ani danych historycznych.
- Stanowi szybkie, proste rozwiązanie, które nie wymaga trenowania modelu ani dużej mocy obliczeniowej.

Model ten służy jako punkt odniesienia do oceny skuteczności bardziej zaawansowanych metod predykcyjnych.

Model zaawansowany Model zaawansowany koncentruje się wyłącznie na wykorzystaniu technik szeregów czasowych lub rekurencyjnych sieci neuronowych (RNN), takich jak LSTM, w celu dokładniejszego przewidywania popularności utworów w przyszłym tygodniu. Jego kluczowe cechy to:

- Analiza trendów popularności utworów na podstawie danych historycznych (jeśli są dostępne).
- Modelowanie dynamiki popularności utworów za pomocą analizy szeregów czasowych.
- Implementacja sieci neuronowych LSTM do:
  - wykrywania wzorców sekwencyjnych w danych,
  - modelowania długoterminowych zależności,
  - przewidywania nagłych zmian w popularności utworów.

Model ten jest zaprojektowany do badania i prognozowania dynamicznych zmian w preferencjach użytkowników.

Podsumowanie

- Model bazowy: oparty na najpopularniejszych utworach z poprzedniego tygodnia, bez wykorzystania uczenia maszynowego.
- Model zaawansowany: wykorzystuje techniki analizy szeregów czasowych lub LSTM w celu modelowania i prognozowania trendów popularności.

Porównanie wyników obu modeli pozwoli na ocenę wartości dodanej zaawansowanego podejścia względem prostego rozwiązania bazowego.

2 Analiza danych

W ramach tej części zdecydowaliśmy się na dokładną analizę danych w celu określenia struktury otrzymywanych informacji wraz z ich rozkładem oraz korelacjami.

Na początku zauważyliśmy, że otrzymane od klienta dane są wybrakowane. W niektórych przypadkach brak jest informacji o ,np. popularności utworu albo gatunku muzyki tworzonej przez artystę. Po konsultacji z klientem dostarczył on danę, które mają poprawione te ubytki.

W przypadku wartości popularności można by próbować własnoręcznie uzupełniać te nieprawidłowości. Poprawnym podejściem mogłoby na przykład założenie w takich przypadkach, że popularność utworu jest równa śreniej popularności utworów danego artysty. Oczywiście wiąże się to z ryzykiem że zostaną zatracone próbki, które odsawały znacząco od średniej, co mogłoby zaburzyć wynik.

2.1 Poszukiwane zależności - przewidywana zmienna

W celu określenia, które utwory staną się najpopularniejsze należy przewidzieć jak będzie się kształtować popularność utworu na platformie ,tj. suma odtworzeń utworu w aplikacji. Sprowadza się to do potencjalnej analizy dwóch zachowań tej właśnie zmiennej:

- Stała wartość popularności - przewidywanie na podstawie parametrów utworu oczekiwanej wartości tego parametru w przyszłości. Potencjalnie, niektóre cechy utworu będą sprawiać, że będzie on lepszym kandydatem do osiągnięcia sukcesu
- Zmiana warrtości popularności - na podstawie nasilenia odczytów uruchomień utworów w ostatnim przedziale czasu można zakładać, że widoczna jest rosnąca tendencja popularności utworu w najbliższym czasie.

Dodatkowo w celu lepszego przewidywania, jak ta wartość może zmieniać się w przyszłości dobrym porównaniem będzie globalna wartość popularności. Mimo, że nie dostarcza ona informacji o zmianach zaobserwowanych na platformie klienta, to jest ona dobrym prognostykiem, które długoterminowe trendy powinny być widoczne ,np. jeśli globalnie pop w ostatnich latach jest najpopularniejszy, to trend ten powinien również obowiązywać na platformie. Dlatego też, zależności te, również będą przeanalizowane w dalszej części tekstu.

2.2 Potencjalnie używane atrybuty

W pierszej kolejności ustaliliśmy listę parametrów, które mogą nieść informacje o przyszłej popularności utworu. Dla każdego z dostarczonych zbiorów można

wyróżnić takie interesujące parametry:

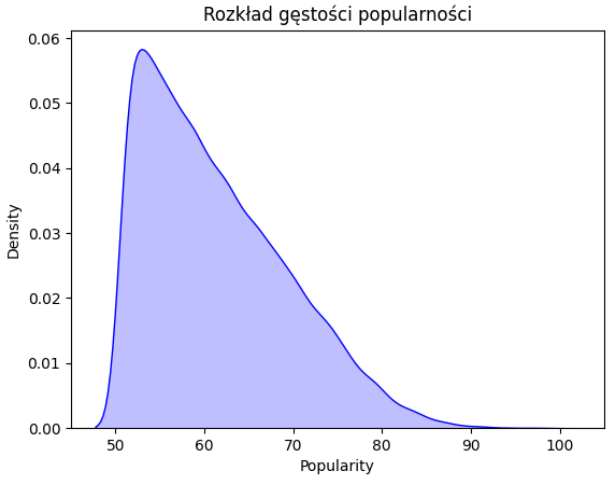
- Sesja - timestamp, track id, event type
- Artyści - id, genres
- Utwory - id, popularity, duration ms, explicit, id artist, danceability, energy, key, loudness, speechiness, acousticness, instrumentalness, liveness, valence, tempo

Wskazane parametry wynikają jedynie ze wstępnej analizy danych. W dalszej części skupiliśmy się na analizie wpływu wskazanych parametrów w celu określenia, czy mogą one być przydatne. W pierwszym etapie przyjrzeliśmy się dwóm ostatnim zbiorom, które powinny odpowiedzieć na pytanie: Co świadczy o popularności utworu?

W przypadku sesji zdecydowaliśmy, że interesujące z punktu rozwiązania zadania powinny być głównie interakcje odtworzenia utworu, ponieważ powinny świadczyć o rosnącej popularności utworu w danym okresie czasu. To właśnie odtworzenia utworu świadczą o jego popularności.

2.3 Analiza utworów

Analizę utworów rozpoczęliśmy od analizy gęstości popularności globalnej dla różnych przedziałów czasu



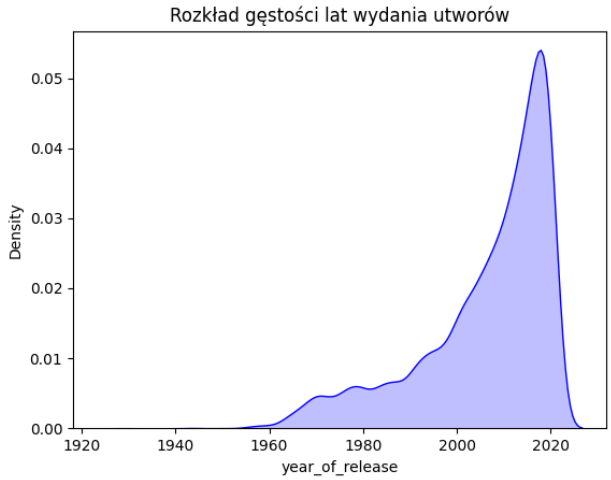
Rysunek 1: Gęstość popularności globalnej dla całości danych

Rysunek 2: Gęstość popularności globalnej dla danych po 2000 roku

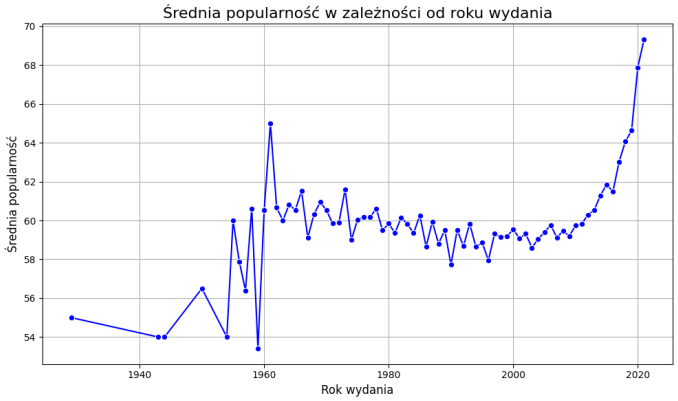
Można zauważyć, że ograniczenie przedziału danych nie wpłynęło na rozkład popularności utworów. Popularność utworów kształtuje się średnio na wartośic 60 i maksymalnie osiąga wartość około 95. Możemy wstępnie zauważyć ,że szukane utwory, które osiągną dużą popularność, będą osiągały tą wartość w przedziale 70-100. W dalszej części to właśnie te utwory będą analizowane pod względem ich cech, ponieważ powinny one przekładać się na najpopularniejsze utwory na platformie użytkownika.

2.3.1 Analiza lat wydania

Jako że interesują nas aktualne trendy utworów warto również przeanalizować wpływ roku wydania utworu na popularność globalną, oraz udział starych danych w dostarczonej próbce. Po dokonaniu analizy można zauważyć, że znaczą-



Rysunek 3: Gęstość lat wydania utworów

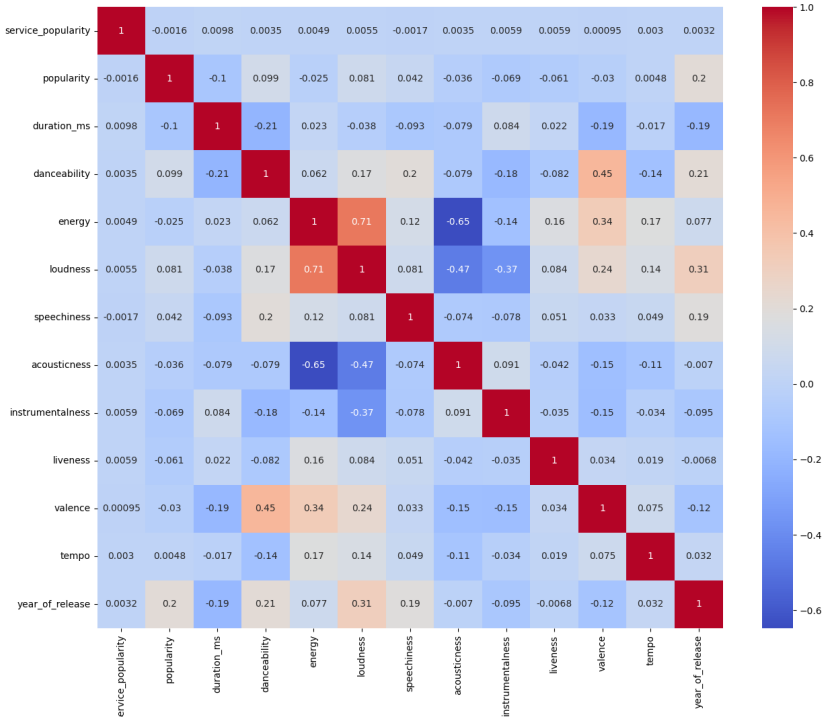


Rysunek 4: średnia popularność globalna utworów w danym roku

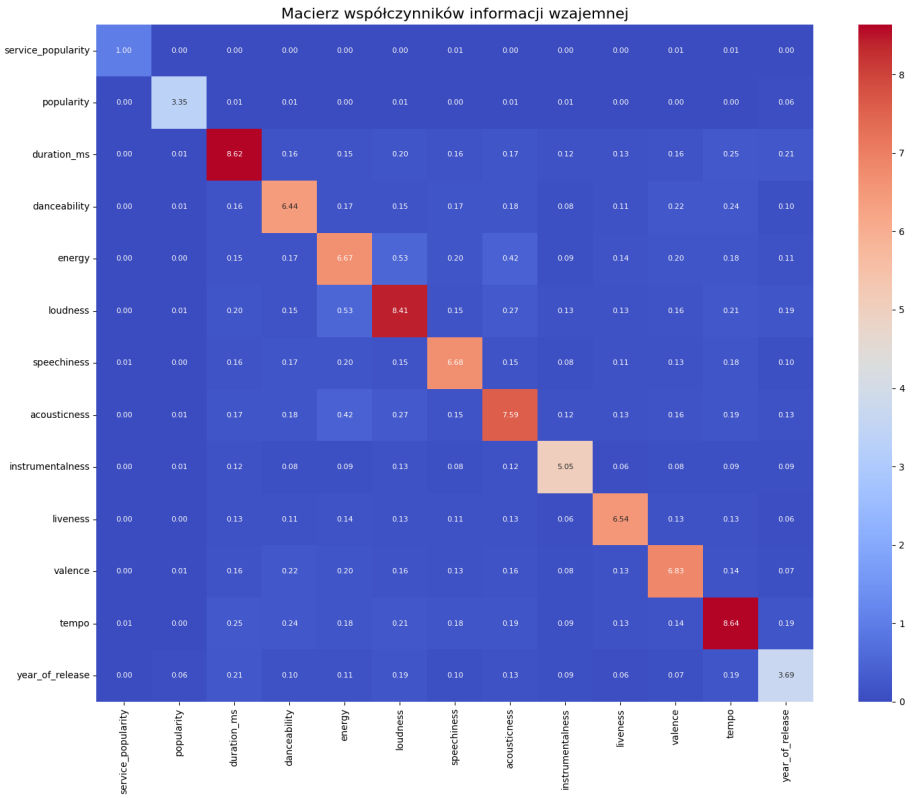
cą część dostarczonych danych stanowią utwory nowsze. Warto również zwrócić uwagę na to, że osiągają one zdecydowanie większą średnią popularność globalną. Należy o tym pamiętać, ponieważ nie możemy przyjąć średnich kryteriów oceny utworów na podstawie całości danych.

2.4 Wpływ atrybutów

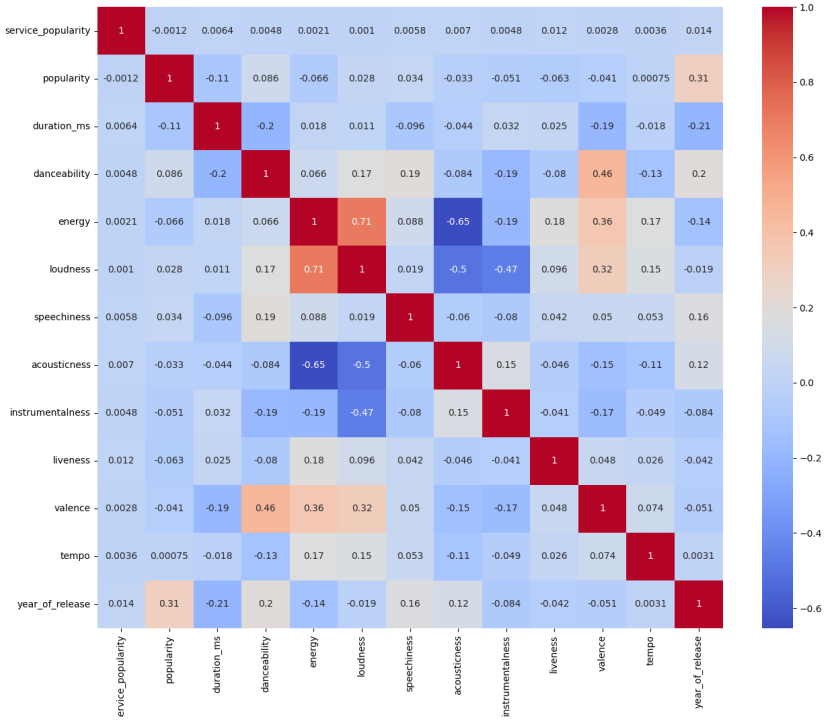
W celu zbadania wpływu technicznych parametrów utworu skorzystaliśmy z macierzy korelacji oraz informacji wzajemnej. Otrzymane zależności wydzieliliśmy dla całości danych oraz określonego przedziału. Analizie poddany zostanie zarówno wpływ na popularność globalną jak i na platformie klienta.



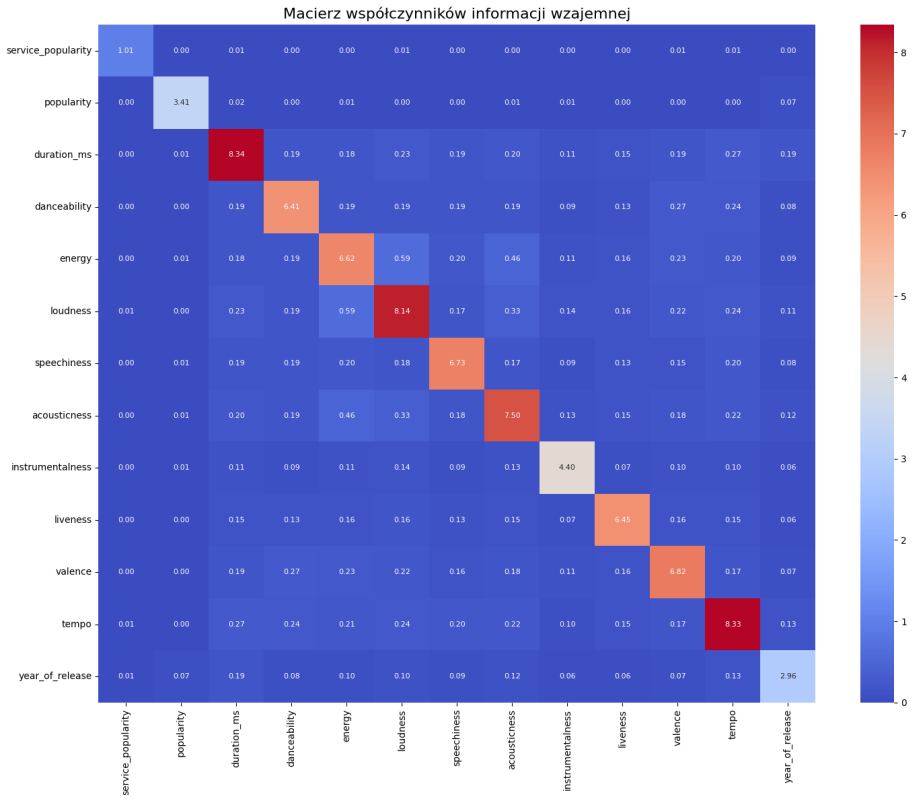
Rysunek 5: Korelacja dla całych danych



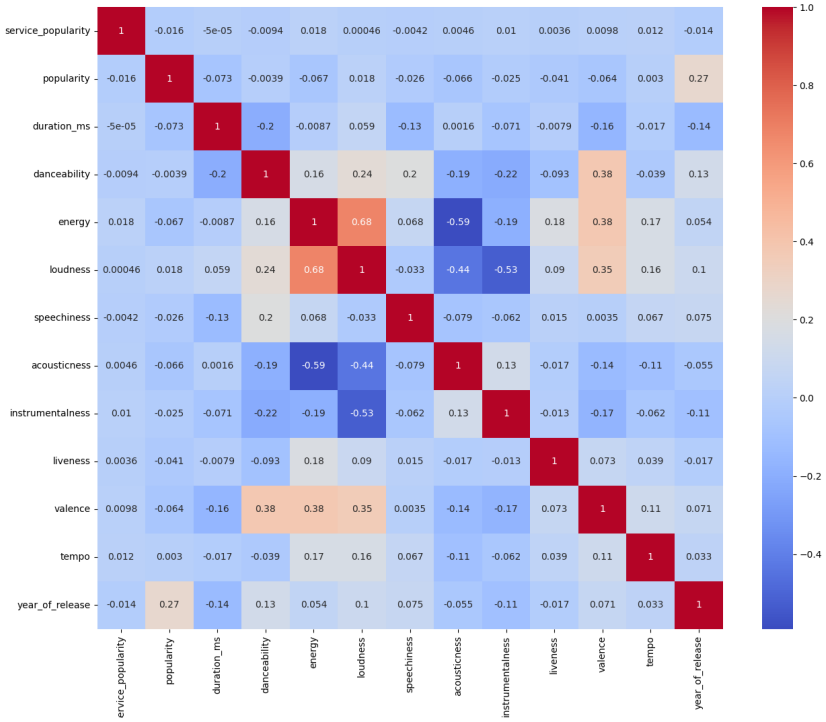
Rysunek 6: informacja wzajemna dla całych danych



Rysunek 7: korelacja po 2000



Rysunek 8: informacja wzajemna po 2000



Rysunek 9: korelacja po 2015

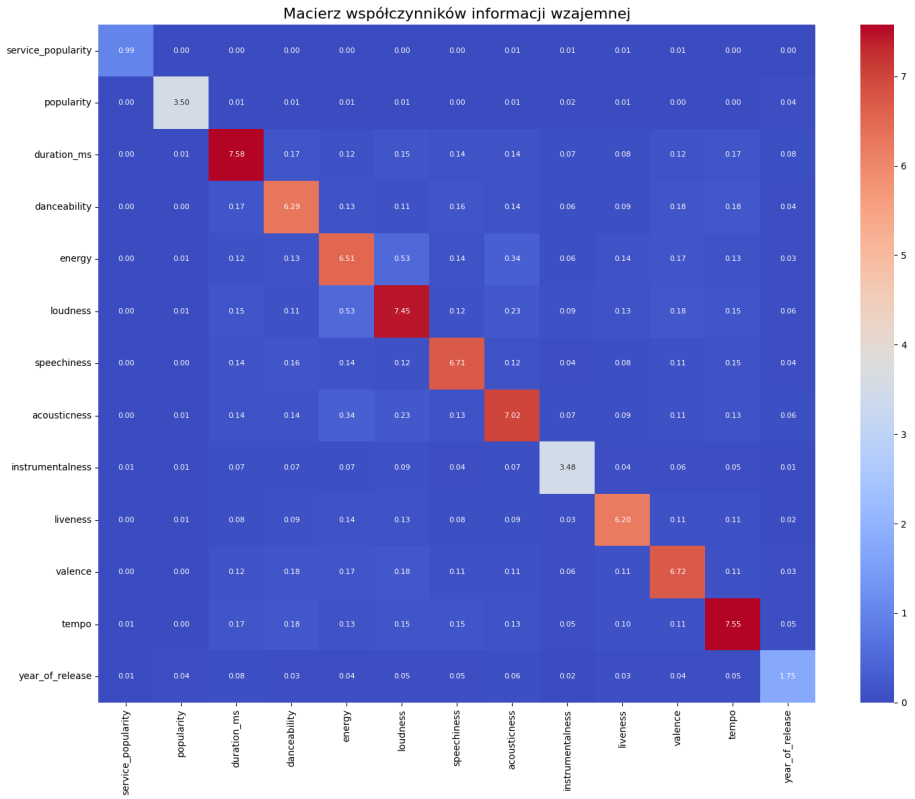
2.4.1 Wpływ na popularność globalną:

Można zauważyć, że otrzymane korelacje są niewielkie. Warto również zwrócić uwagę na to, że dla nowszych utworów które są dla nas najbardziej istotne niektóre zależności maleją. Dzieje się tak na przykład w przypadku korelacji między dance ability a popularnością globalną.

2.4.2 Wpływ na popularność w aplikacji:

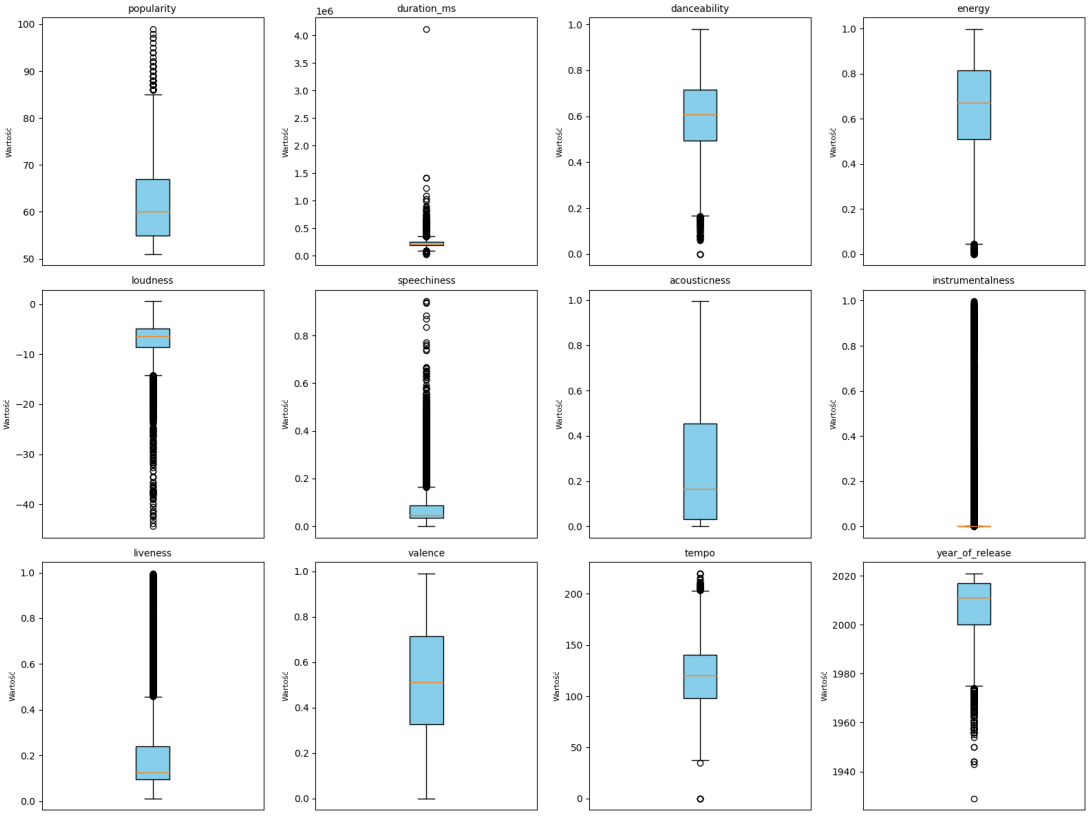
W przypadku analizy związku parametrów utworu z popularnością na platformie zależności są jeszcze mniejsze. Ciężko o wyraźne zaznaczenie zależności między parametrami wejściowymi i wyjściowymi.

Dokonaliśmy również analizy wykresów pudełkowych, które pokazują, że dla niektórych parametrów ciężko otrzymywane wartości są bardzo skoncentrowane, a dla innych ciężko o ustalenie jakiejś wartości przewidywanej. Z racji znikomego wpływu wartości loudness oraz jego silnej korelacji z energy można zdecydować

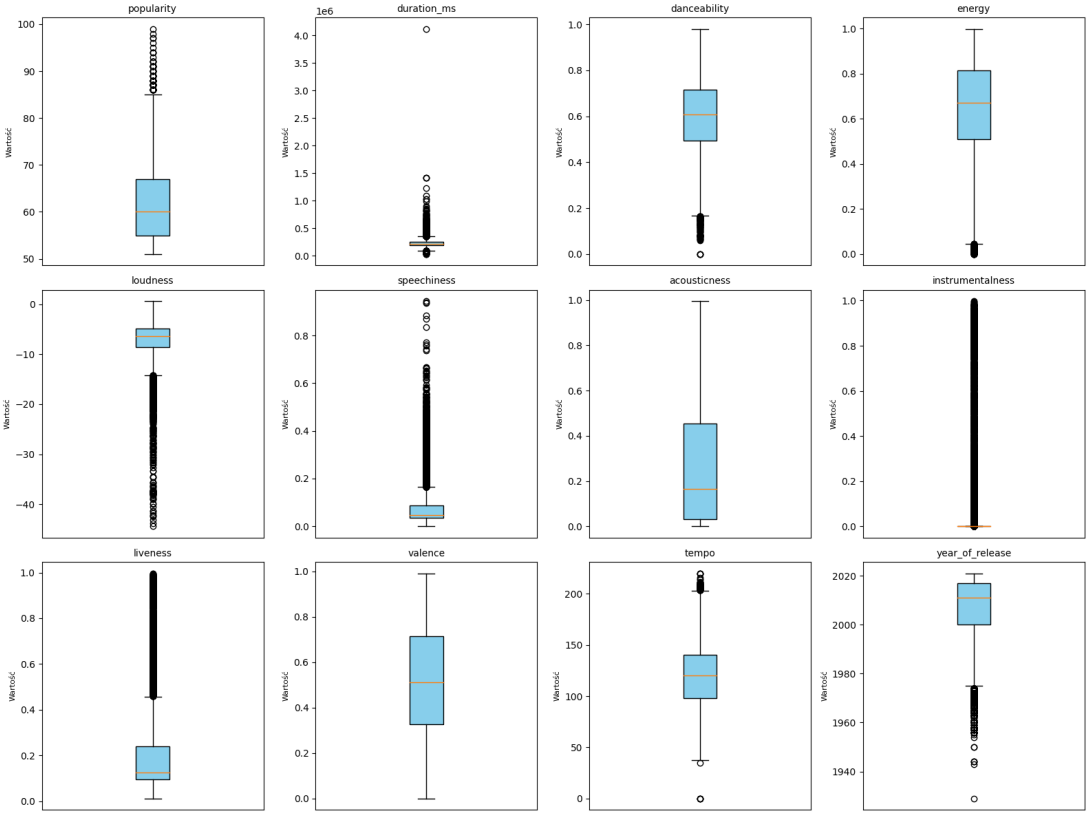


Rysunek 10: informacja wzajemna po 2015

o porzuceniu tego parametru. Niepokojąca jest również mała korelacja z popularnością globalną, co zostanie rozwinięte we wnioskach końcowych. Jeśli chodzi o współczynniki informacji wzajemniej, to są one marginalne.



Rysunek 11: Wykresy pudełkowe dla całych danych

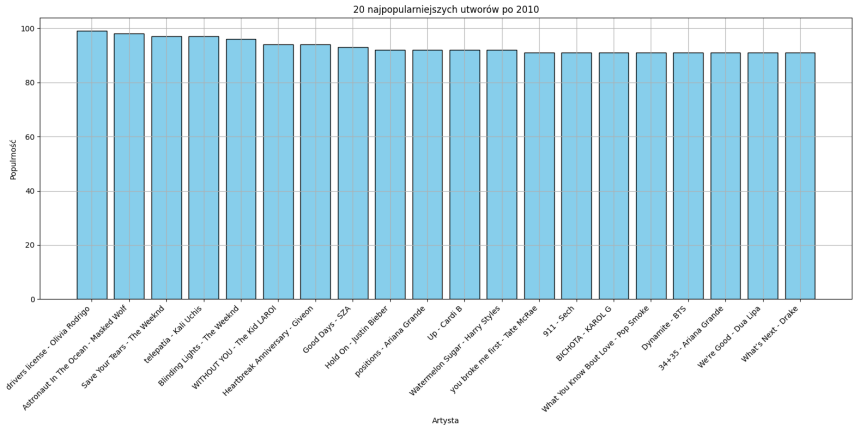


Rysunek 12: Wykresy pudełkowe dla danych po 2015

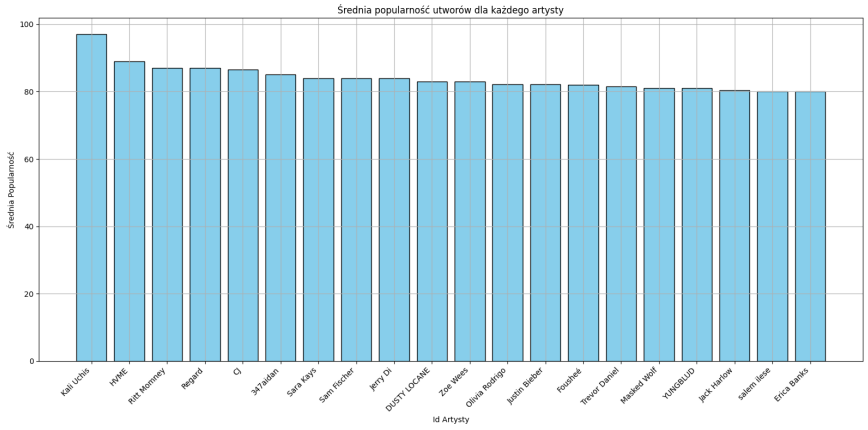
2.5 Analiza artystów i gatunków

W tej części analizy skupiliśmy się na analizie artystow utworów. Założyliśmy, że znajomość artysty danego utworu może nam dawać przewagę w przewidywaniu, czy dany utwór będzie popularny. Podobnej zależności spodziewaliśmy się w przypadku gatunku utworu. Otrzymane gatunki były szeroko zróżnicowane, dlatego zdecydowaliśmy się przekształcić parametr w postaci listy na kilka parametrów binarnych, np. czy utwór jest z gatunku pop. Wydzieliliśmy tylko podstawowe gatunki, które najczęściej występowały w danych.

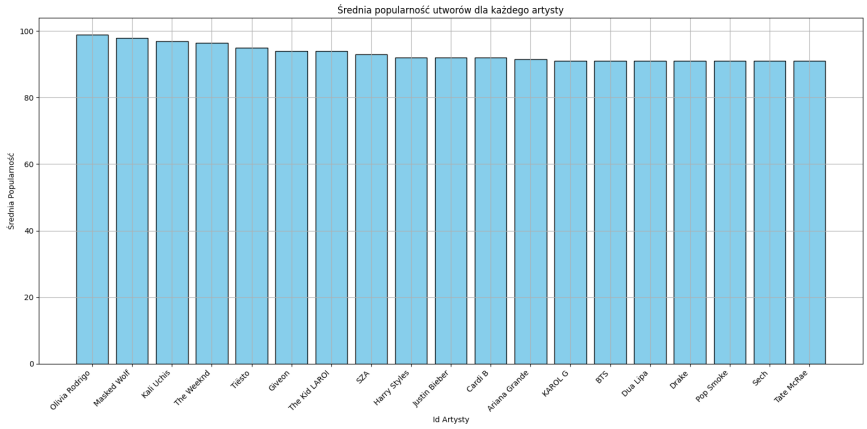
2.5.1 Analiza artystów



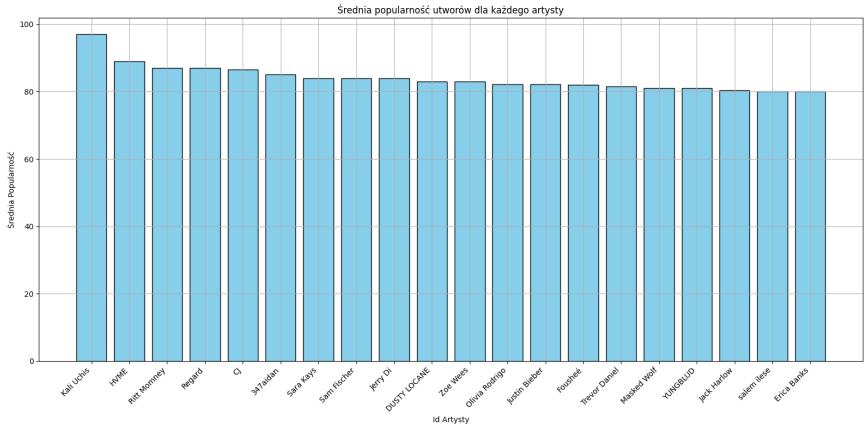
Rysunek 13: 20 najpopularniejszych utworów wydanych po 2010 roku



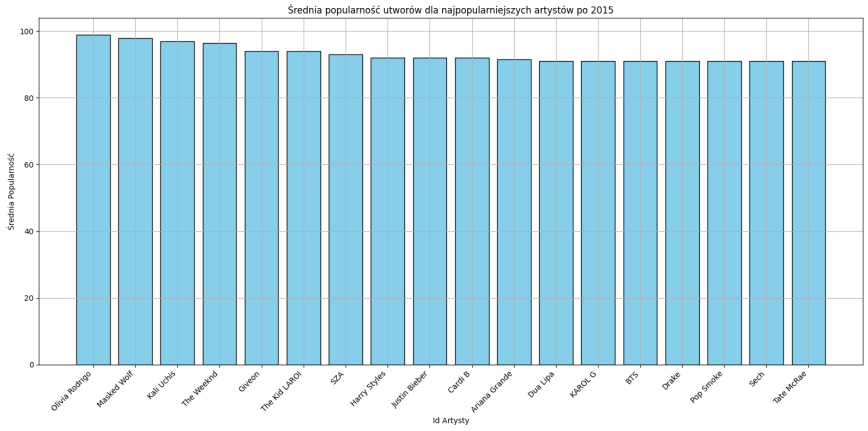
Rysunek 14: Średnia popularność utworów artysty



Rysunek 15: średnia popularność hitowych utworów artysty



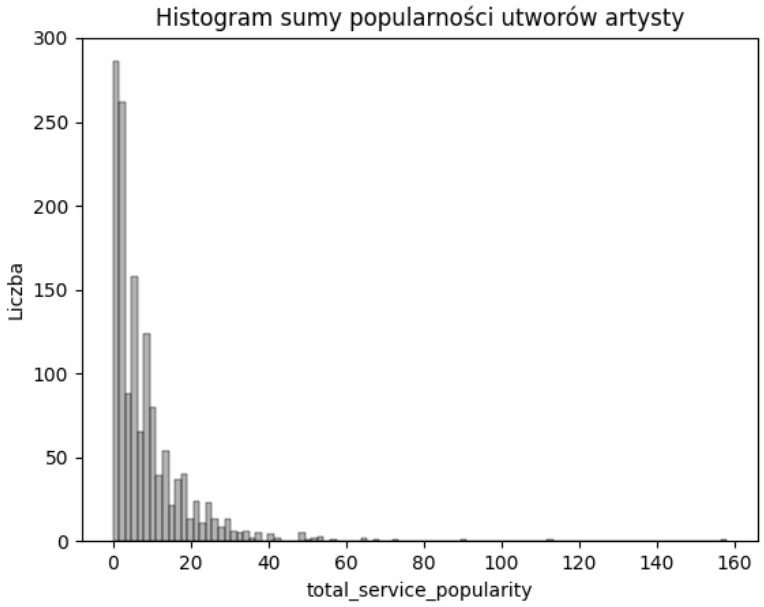
Rysunek 16: średnia popularność utworów artystów po 2015



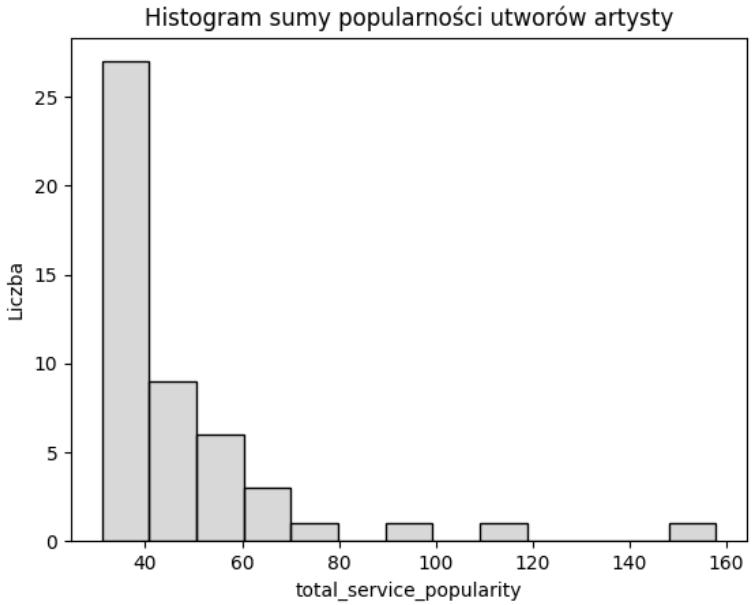
Rysunek 17: średnia popularność hitowych utworów artystów po 2015

W przypadku badania utworów hitowych braliśmy pod uwagę tylko utwory o popularności globalnej większej niż 90. Miało to na celu określenie kto w danym przedziale czasowym "królował na szczycie".

Można zauważyć, że niektórzy artyści osiągają zdecydowanie większe zasięgi od pozostałych. Informacje tą można wykorzystać w celu przewidzenia, że ich utwory najprawdopodobniej utrzymają pewien poziom popularności. Podobny trend można również delikatnie zauważyć w aplikacji, gdzie zbadano na histogramie rozkład sumy odtworzeń utworów artystów:



Rysunek 18: Suma popularności utworów artysty

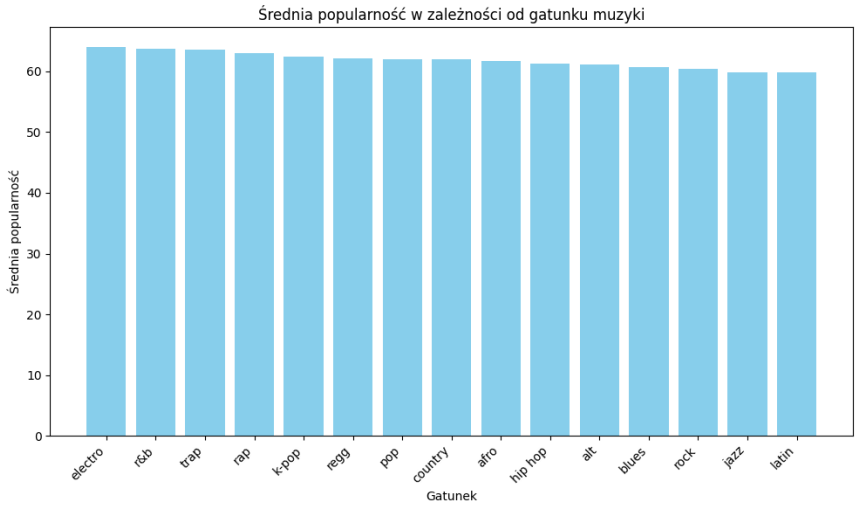


Rysunek 19: Suma popularności utworów artysty dla najpopularniejszych artystów

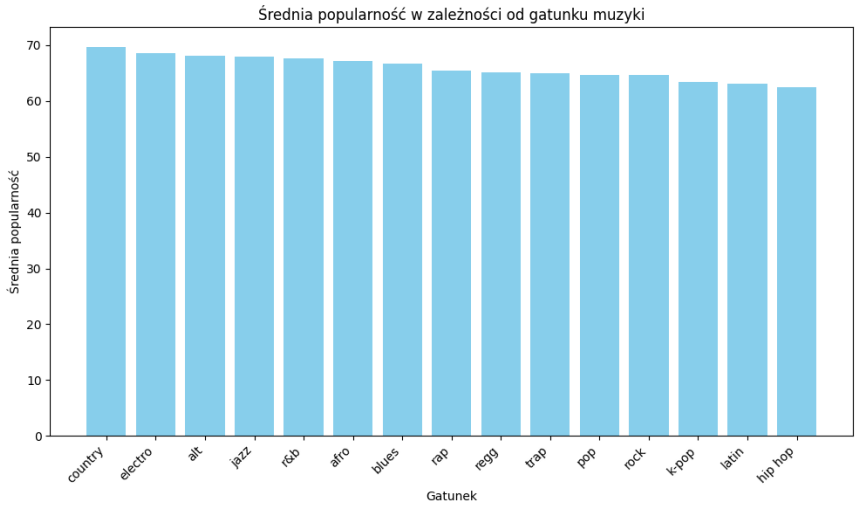
Również w aplikacji można już zauważyć trend, świadczący silnie o tym, że niektórzy artyści są znacznie popularniejsi od pozostałych, co świadczy o tym, że w ich przypadku znacznie bardziej prawdopodobne stanie się, że utwór będzie jednym z popularniejszych w przyszłości.

2.5.2 Analiza gatunków

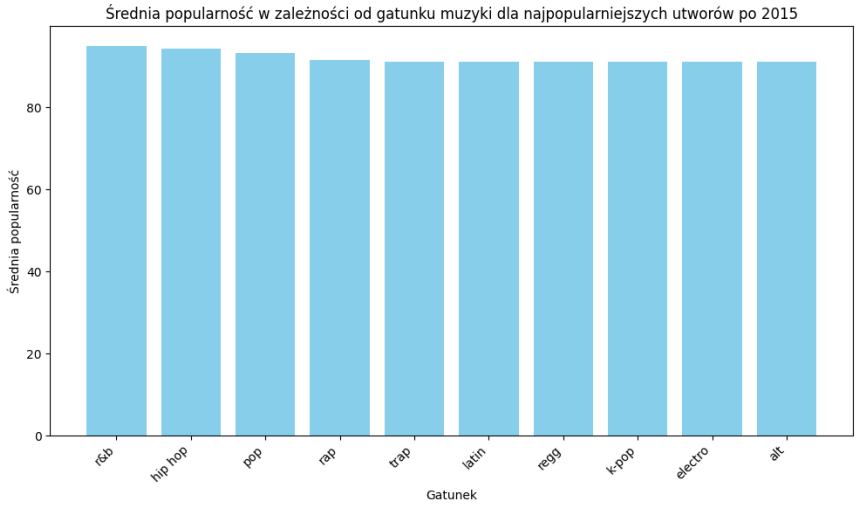
Tak jak zostało to wspomniane wcześniej również część związana z gatunkami utworów będzie poddana analizie pod względem popularności globalnej. W tym przypadku ciężko o jakąś przewagę określonego gatunku muzyki, ale warto przyjrzeć się, jak te wartości kształtują się dla utworów hitowych.



Rysunek 20: średnia popularność gatunku dla wzsystkich danych



Rysunek 21: średnia popularność gatunku po 2015 roku



Rysunek 22: średnia popularność gatunku dla utworów hitowych po 2015 roku

2.5.3 Wnioski

Warto zauważyć, że na przykład w przypadku muzyki country posiada ona największą średnią popularność globalną po 2015 roku, ale nie "króluje" w gronie utworów hitowych. Sprawia to, że ciężko zakładać że utwór country prześcignie inne gatunki i osiągnie jeden z najlepszych wyników.

3 Wnioski odnośnie przydatności danych

Uzyskane od klienta dane, pozwalają na szukanie wpływu ogólnych cech na globalną popularność utworu.

Niestety, znacznie gorzej wygląda to w przypadku próby przewidywania popularności na platformie klienta. Otrzymane dane wydają się nie reprezentatywne, co można potwierdzić tym, że jest zdecydowanie za mała korelacja, między popularnością globalną a tą w aplikacji. Utwory najpopularniejsze osiągnęły dopiero wartość 6.

W przypadku niektórych parametrów ich korelacja z wyjściem modelu zdaje się być już wystarczająca, ale dokładne działanie predykcji będzie możliwe tylko w przypadku znacznego powiększenia zbioru danych sesji.

Udało się za to znaleźć ogólny wzór utworu, który powinnien również na platformie osiągać najważniejsze wartości, tj. można zauważyć jak jego gatunek/artysta wpłynie na popularność.