Analiza wpływu danych związanych z utworem na jego popularność

Michał Żądełek, 266519
266519@student.pwr.edu.pl
K01-21c

10 czerwca 2023

Spis treści

1 Opis analizowanego problemu

Celem projektu jest zbadanie wpływu różnych informacji związanych z utworem, takich jak:

- Długość jego trwania
- Zawartość treści dla dorosłych
- Popularność albumu, w którym się znajduje
- Średnia liczba obserwujących przypadająca na wykonawcę
- Średnia popularność przypadająca na wykonawcę

względem aktualnej popularności utworu. W zbiorach danych udostępnianych przez popularne serwisy strumieniowe, takie jak Spotify, analiza popularności utworu zazwyczaj opiera się na liczbie odsłuchań oraz aktualności utworu, rozumianej jako liczba jego odtworzeń w określonym czasie. W niniejszym raporcie zostanie zbadane alternatywne podejście do oceny popularności utworu, uwzględniające wymienione wyżej czynniki.

2 Zgromadzenie i wstępne przetworzenie danych

2.1 Zgromadzenie danych

Dane związane z utworami, albumami oraz artystami zostały zgromadzone za pomocą Web API - Spotify for Developers udostępniającego użytkownikowi spełniającemu określone wymagania użytkowe treści o różnorodnej tematyce w formacie .json. Wspomnianymi wyżej wymaganiami są:

- Aktywne konto na platformie Spotify
- Posiadanie powiązanej z kontem aplikacji deweloperskiej
- Aktywny token, wygenerowany na podstawie związanych z aplikacją informacji określanych jako *Client ID* oraz *Client Secret*

Taki użytkownik nabywa uprawnień do uzyskania danych zarówno powiązanych ze swoim kontem, jak i tych ogólnodostępnych na platformie, których stan datowany jest na ten sam dzień, w którym użytkownik korzysta z interfejsu programowania aplikacji.

Utwory do analizy zostały wyszukane za pomocą rekomendacji, które można było dostosować za pomocą filtrów takich jak:

- *seed genres* - lista gatunków, z których utworów poszukujemy (maksymalnie 5)
- *limit* - maksymalna liczba zwróconych obiektów utworów (100 dla pojedynczej prośby o dane)
- *market* - lista rynków, w których utwory są dostępne
- *min popularity* - minimalna popularność utworów
- *max popularity* - maksymalna popularność utworów

Dostosowanie filtrów:

- Z faktu wysokiej dostępności utworów jedynym rynkiem został wybrany rynek *USA*
- Zważając na ograniczenia ilościowe dostępnych rekomendacji tj. średnio 700 utworów na każde 5 dostępnych gatunków muzycznych do analizy użyto utworów z 25 gatunków muzycznych
- *limit* został ustawiony na jego maksymalną wartość
- z racji popularności utworu będącej w przedziale [0;100] *min popularity* ustawiono na 0
- z racji popularności utworu będącej w przedziale [0;100] *max popularity* ustawiono na 100

Każdy obiekt pobrany z Web API można zidentyfikować na podstawie jego *id* oraz charakterystycznych dla niego cech. W przypadku obiektu utworu jego atrybuty, takie jak między innymi:

- *track name* - nazwa utworu
- *artists* - lista wykonawców
- *artists ids* - lista id wykonawców
- *album name* - nazwa albumu
- *album id* - id albumu
- *duration [ms]* - czas trwania utworu w milisekundach
- *explicit* - zawartość treści dla dorosłych

okazały się **niewystarczające** do uzyskania sensownych wyników w badaniu jego popularności.

Z racji wystąpienia w instancji utworu informacji identyfikujących album oraz wykonawców niezbędnym okazało się dobranie dodatkowych treści, którymi okazały się być:

- *album popularity* - popularność albumu, do którego należy utwór
- *artist popularity* - popularność wykonawcy
- *artist followers* - liczba użytkowników obserwujących wykonawcę

Tak zebrane dane wydawały się być **wystarczające** do podjęcia się ich dalszego przetwarzania.

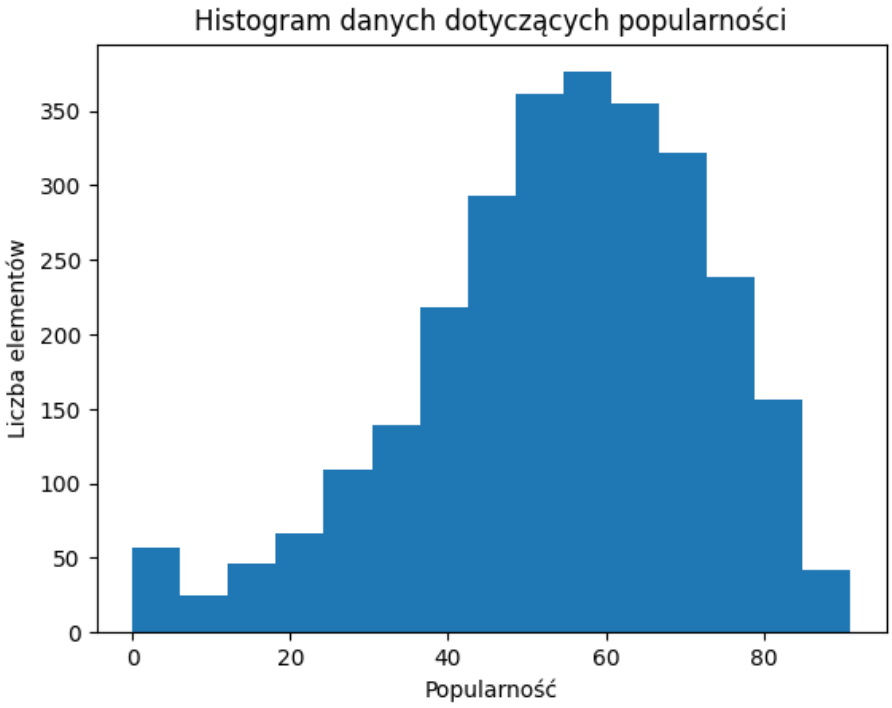
2.2 Wstępne przetworzenie danych

Podczas przetwarzania danych podjęto następujące kroki:

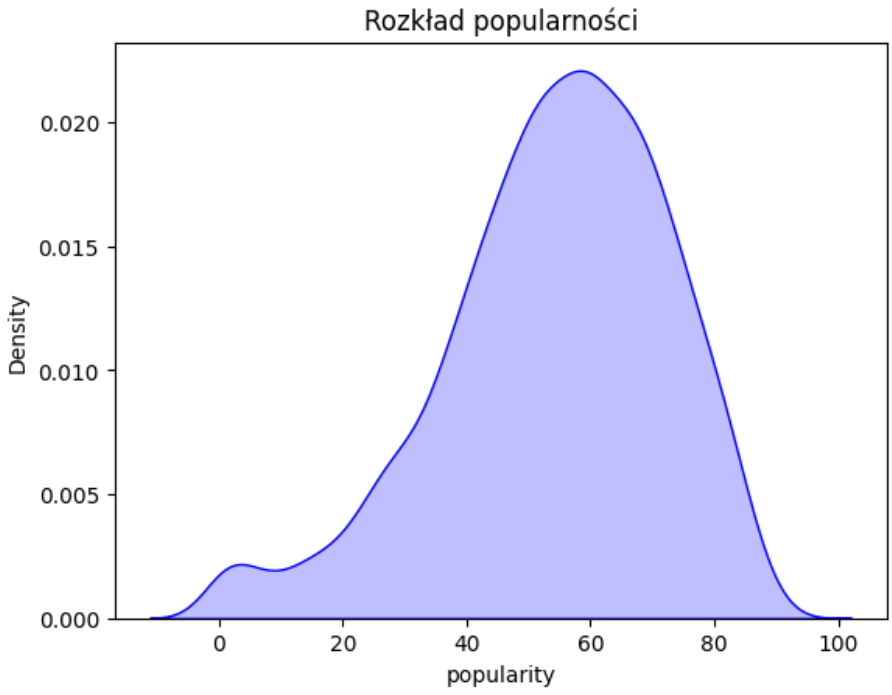
- Na wstępie dokonano połączenia określonych w punkcie wyżej danych utworu z wspomnianymi dodatkowymi informacjami uzyskanymi z obiektów albumu oraz wykonawców, tworząc kompletne rekordy dotyczące poszczególnych utworów.
- Zważając na fakt, iż jeden utwór może być wykonywany przez wielu artystów odpowiednią praktyką wydawało się wyciągnięcie średniej wartości popularności oraz liczby obserwujących przypadających na jednego artystę.
- W Zbiorze danych po scaleniu nie odnotowano żadnych brakujących wartości w kolumnach.
- Dane z kolumny *album popularity* mimo znacznego udziału wystąpień wartości 0 (42%) zostały niezmienione, uznając tak częste wystąpienia tej wartości za charakterystyczną właściwość badanego zbioru danych.
- Dokonano usunięcia kolumn, których rolą była jedynie identyfikacja obiektów utworu, albumu i artystów.
- Zastąpiono wartości "True"oraz "False"w kolumnie *explicit* odpowiednio wartościami 1 oraz 0, aby można było stworzyć macierz korelacji między cechami, a zmienną celu.

3 Wstępna analiza danych

3.1 Analiza danych dotyczących popularności utworu



Rysunek 1: Histogram popularności utworów



Rysunek 2: Rozkład danych o popularności utworów

Wartość min.	Wartość maks.	Średnia	Odchylenie standardowe
0	91	54.07	18.21

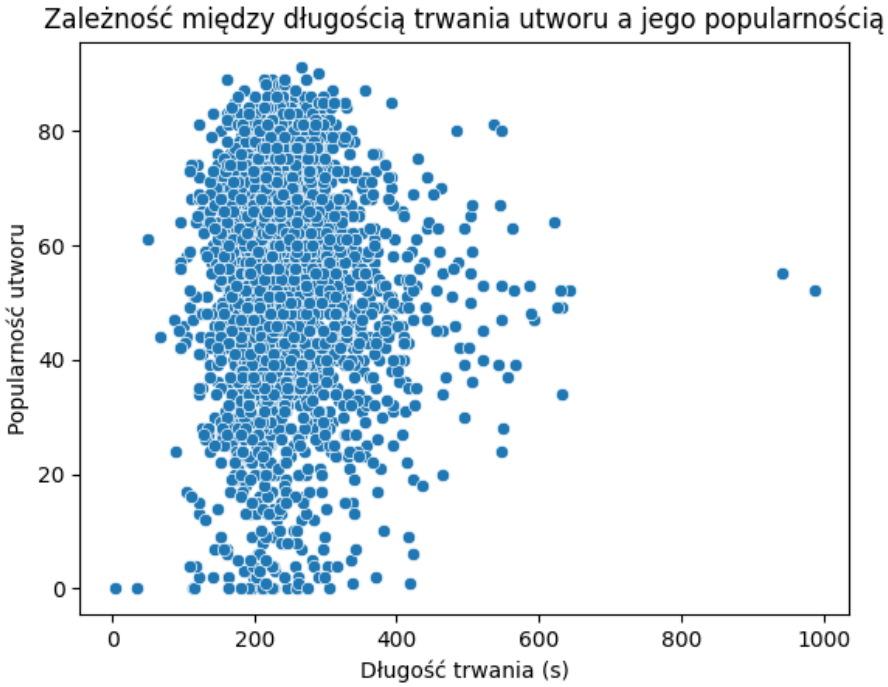
Tabela 1: Ważne statystyki popularności (jej wartości nie mają jednostki)

Można zauważyć, iż rozkład popularności przypomina rozkład normalny. W związku z tym najczęściej można spotkać się z wartościami popularności utworów wynoszącymi około 54. Ten fakt może przyczynić się do późniejszego trafniejszego przewidywania popularności utworów o średniej popularności (30-70), jednak może okazać się problemem przy ocenie popularności utworów o niskiej jak i wysokiej popularności ze względu na niewystarczającą liczbę danych uczących.

3.2 Analiza zależności między zmienną celu a cechami

3.2.1 Wpływ długości trwania utworu na jego popularność

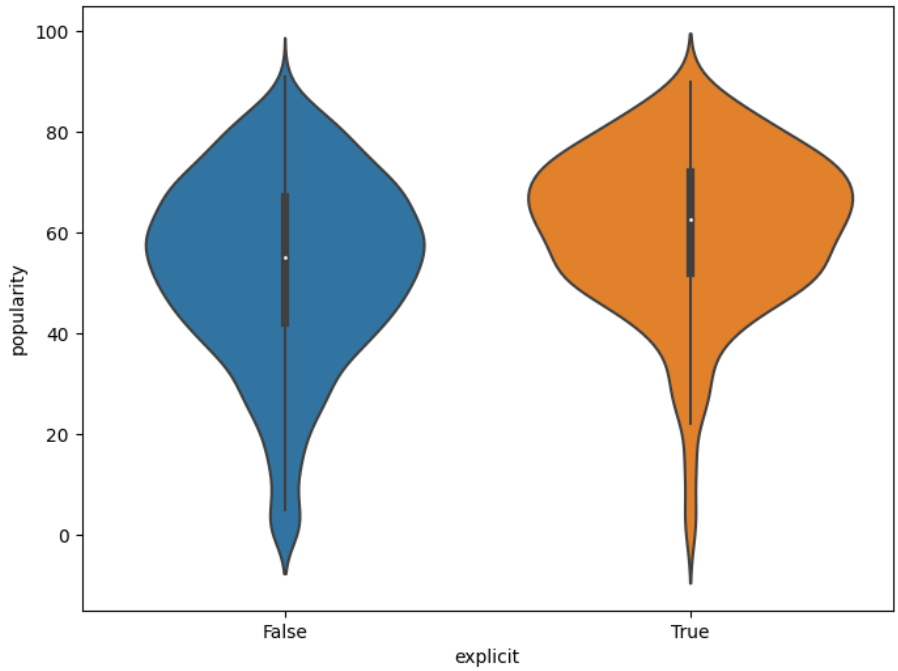
Z wykresu 3 można zauważyć, iż długość utworu ma nikły wpływ na popularność utworu - wiele utworów o podobnej długości ma diametralnie różną popularność. Zasadnym wydaje się odrzucenie tego parametru w dalszej pracy na danych.



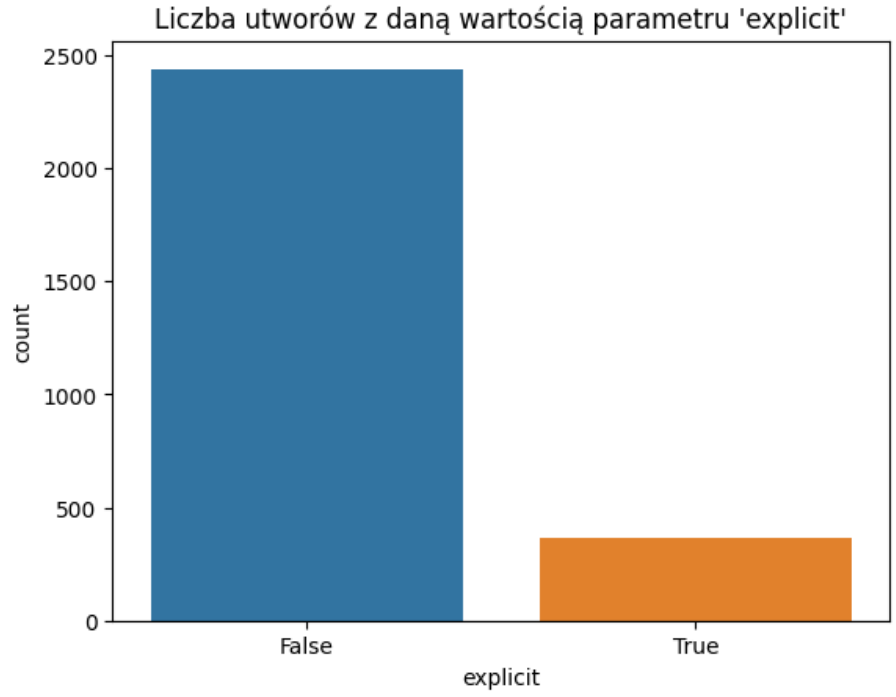
Rysunek 3: Wykres popularności utworu od długości trwania utworu (w sekundach)

3.2.2 Wpływ zawartości treści dla dorosłych w utworze na jego popularność

Na podstawie wykresu 4 można zauważyć, iż jest nieznacznie mniej utworów z niską popularnością (do 40), a także niewiele więcej utworów z wysoką (od 70) oraz średnią popularnością (40-70) posiadających *explicit=True*. Fakt ten może świadczyć o pewnej korelacji między tym parametrem a popularnością utworu. Należy jednak zauważyć, iż utworów z *explicit=False* jest kilkukrotnie więcej - co obrazuje wykres 5. Możliwym jest, iż taka dysproporcja danych bezpośrednio wpływa na kszałt wykresu pierwszego i wnioski z punktu wyżej w rzeczywistości nie są słuszne, czyniąc ten parametr niepewnym punktem dalszej analizy.



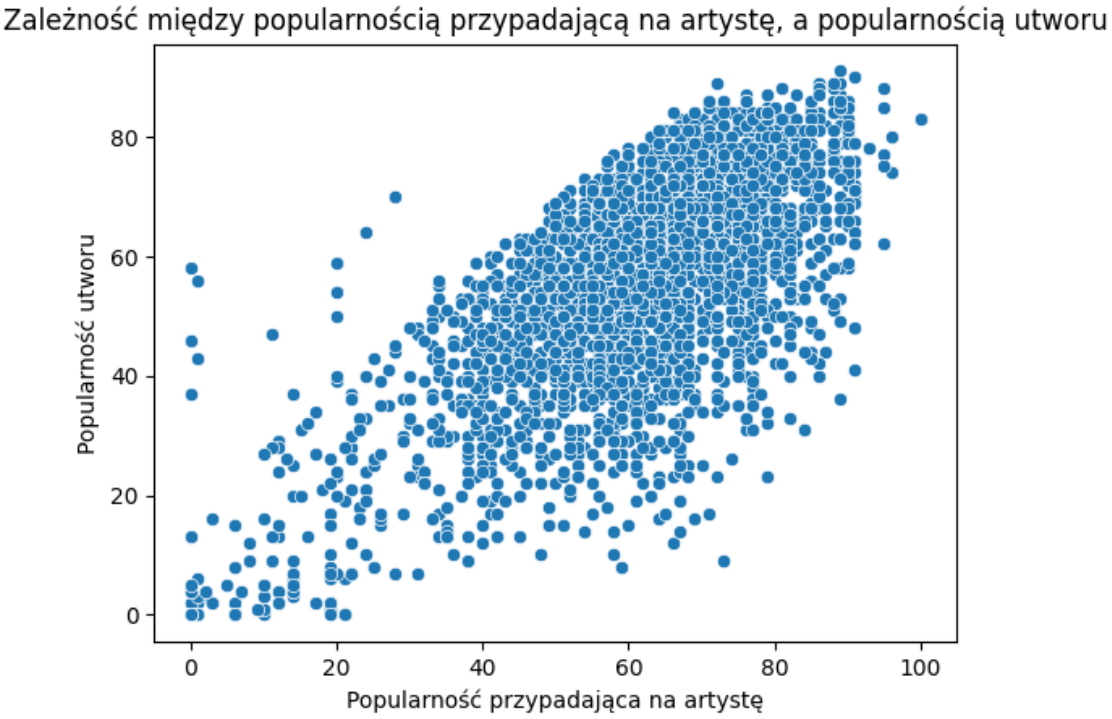
Rysunek 4: Wykres rozkładów popularności utworów w zależności od wartości parametru explicit



Rysunek 5: Wykres słupkowy obrazujący liczność konkretnych wystąpień parametru explicit

3.2.3 Wpływ popularności przypadającej na wykonawcę utworu na jego popularność

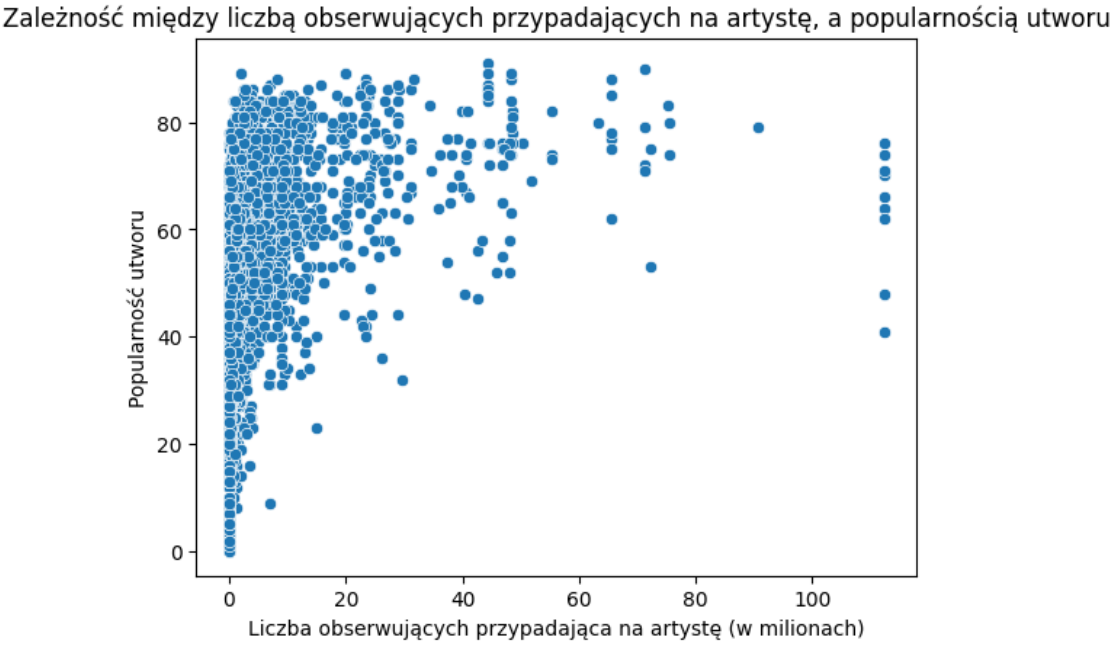
W relacji popularności przypadającej na artystę oraz popularności utworu umieszczonej na wykresie 6 można dojrzeć korelację - im wyższa popularność przypadająca na artystę tym także wedle liniowego trendu wyższa popularność utworu.



Rysunek 6: Wykres popularności utworu od popularności przypadającej na artystę

3.2.4 Wpływ liczby obserwujących przypadającej na wykonawcę utworu na jego popularność

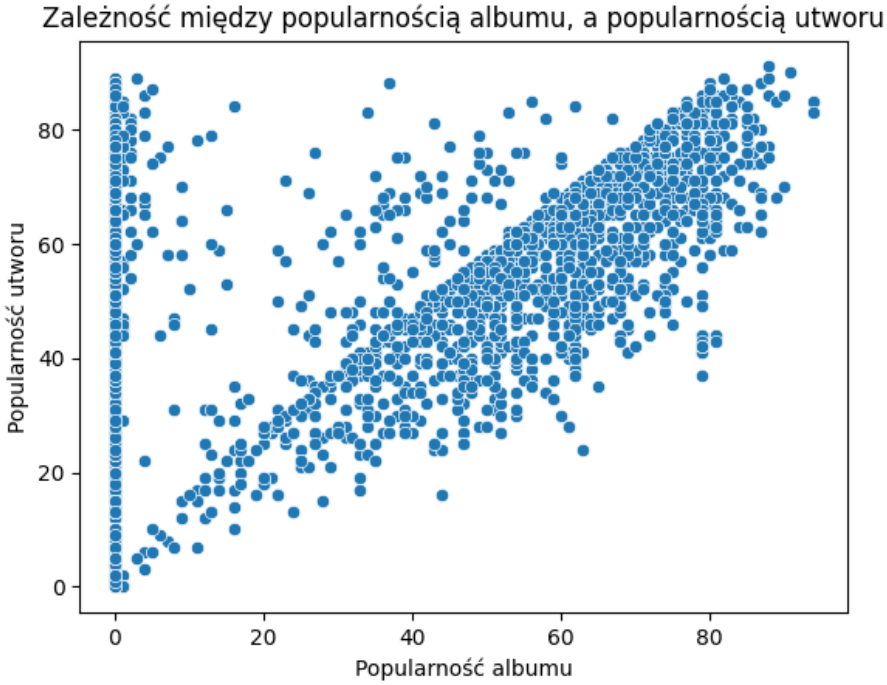
Zależność popularności utworu od średniej liczby obserwujących (w milionach) umieszczona na wykresie 7 wydaje się być logarytmiczna. Może być to jednak spowodowane konkretnym podziałem zbioru danych na testowy i treningowy, dla którego wspomniana zależność akurat występuje.



Rysunek 7: Wykres popularności utworu od liczby obserwujących przypadających na artystę

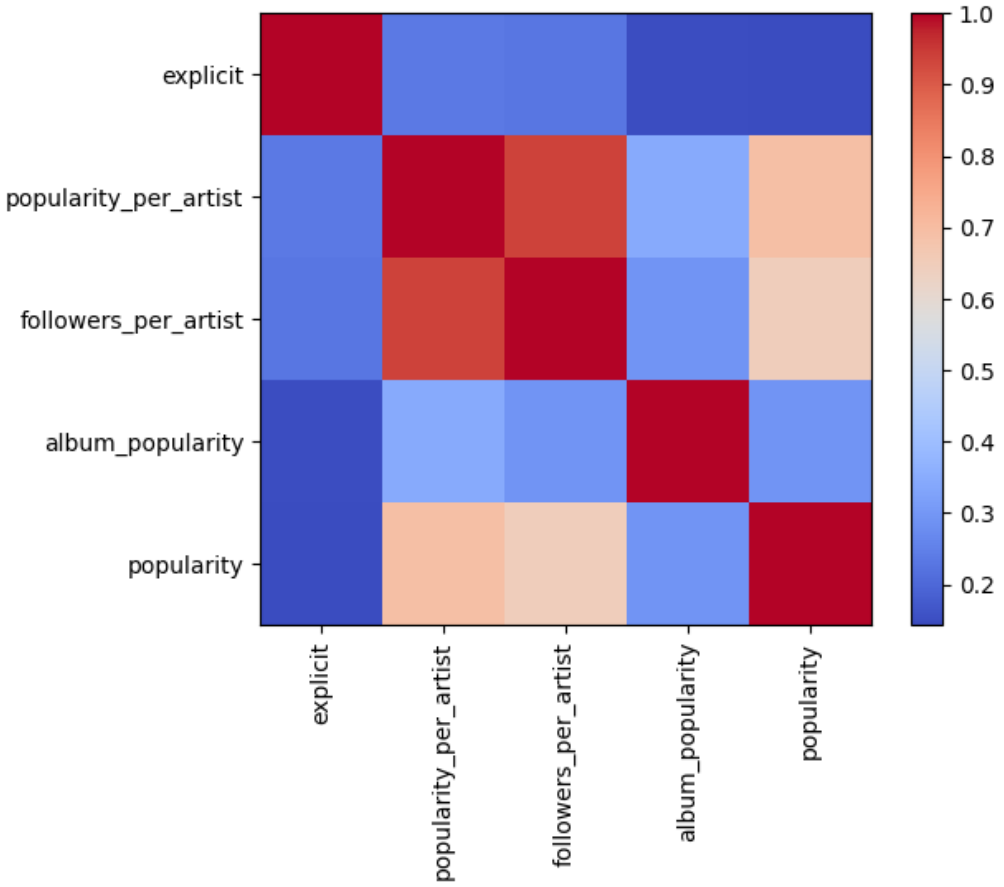
3.2.5 Wpływ popularności albumu na popularność utworu

Badając wpływ popularności albumu na popularność utworu, który w nim jest zawarty (wykres 8) można dostrzec trend - wraz ze wzrostem popularności albumu często rośnie także popularność utworu. Problem jednak może stanowić podzbiór wartości popularności albumu równych 0, dla których popularność utworu jest rozłożona na całą skalę możliwych wartości - z danych wynika, iż niepopularny album może zawierać bardzo lub wcale niepopularne utwory. Prawdopodobnie są to wszystkie te przypadki, gdzie z niszowego albumu wyłania się mniejszy, bądź większy "hit".



Rysunek 8: Wykres popularności utworu od popularności albumu

3.3 Macierz korelacji



Rysunek 9: Macierz korelacji między parametrami

Jak widać z powyższego wykresu, następuje duża rozbieżność między korelacjami parametrów:

- Parametr *explicit* wedle przewidywań ma niewielki wpływ na wartość popularności utworu, jak i nie jest skorelowany z żadną inną cechą ze zbioru cech.
- Parametr *album popularity* mimo znacznie wyższego wskaźnika korelacji dalej pozostaje wyraźnie daleki od znaczącego wpływania na popularność utworu.
- Parametr *followers per artist* na macierzy korelacji został zlogarytmizowany - w takiej postaci wykazuje zauważalną korelację z popularnością utworu.
- Parametr *popularity per artist* według przewidywań jest najbardziej znaczącym atrybutem jeżeli chodzi o przewidywanie popularności utworu osiągając najwyższą, jednak dalej daleką od oczekiwanej wartości korelacji z zmienną *popularity*.

Dodatkowo można dostrzec, iż parametry *followers per artist* oraz *popularity per artist* są ze sobą znacząco skorelowane, co może sugerować, iż warto pominąć jeden z nich przy trenowaniu modelu, gdyż niosą porównywalną informację o popularności utworu.

4 Modele

4.1 Sposoby oceny modeli

Jakość modelu określana jest za pomocą trzech metryk:

- R^2 Score - będącą proporcją wariancji zmiennej zmiennej docelowej/przewidywanej, którą można wyjaśnić parametrami modelu. Wyższa wartość wskazuje na lepsze dopasowanie, dziedzina to: $R^2 \in [0, 1]$.
- Błąd średniokwadratowy (MSE, ang. Mean Square Error) - mierzącą średnią wartość kwadratu różnicy między wartościami przewidywanymi przez model a wartościami rzeczywistymi. Im mniejsza wartość MSE, tym lepiej dopasowany jest model do danych. Jej najmniejsza możliwa wartość to 0.
- Pierwiastek błędu średniokwadratowego (RMSE, ang. Root Mean Square Error) - mierzącą średnią różnicę między wartościami przewidywanymi przez model a wartościami rzeczywistymi.Im mniejsza wartość RMSE, tym lepiej dopasowany jest model do danych. RMSE jest wyrażane w tych samych jednostkach co zmienna zależna, co ułatwia interpretację. Z racji, iż jest to pierwiastek z MSE, najmniejsza możliwa wartość również wynosi 0.

4.2 Przygotowanie danych do modeli

4.2.1 Selekcja cech

Z faktu, iż analiza danych wykazała różną siłę zależności pomiędzy cechami a zmienną celową warto zdecydować się na wybór tych cech, które mogą być najbardziej użyteczne w trenowaniu modeli. Takie działanie ma na celu przede wszystkim redukcję wymiarowości danych oraz poprawę wydajności modelu. Wybór zostaje dokonany poprzez użycie obiektu klasy SelectKBest, która:

- Przeprowadza test chi^2 dla każdej cechy w zbiorze danych.
- Przydziela każdej cesze wartość znaczenia na podstawie obliczonej metryki.
- Wybiera 2 najlepsze cechy, które mają najwyższe wartości znaczenia.

Dwoma parametrami z najwyższymi wartościami znaczenia na podstawie tego testu okazały się *album popularity* oraz *popularity per artist*.

4.2.2 Podział danych

Dane zostały podzielone na testowe i trenigowe w stosunku 1:4. Zgodnie z tą proporcją modele trenowano na 2803 próbkach, natomiast do testu użyto 701 utworów.

4.3 Dobór modeli

Z racji tego, iż problem dotyczy predykowania wartości popularności utworów można go zaklasyfikować jako problem regresji. Z tego faktu do jego rozwiązania zostały wybrane modele regresji liniowej (LinearRegression) oraz regresji wektorów nośnych (SVR). Model regresji liniowej polega na znalezieniu najlepszego dopasowania linii do danych, aby przewidywać wartości zależnej zmiennej (celu) na podstawie niezależnych zmiennych (cech), natomiast SVR na znalezieniu hiperpłaszczyzny w przestrzeni o jak największym marginesie, która najlepiej dopasowuje dane i przewiduje wartości zależnej zmiennej (celu). Do celów zadania użyto implementacji tychże modeli zawartych w bibliotece *sklearn*.

4.4 Trenowanie modeli

Do znalezienia najlepszych parametrów modeli posłużył obiekt GridSearchCV, który automatycznie przeszukuje przestrzeń hiperparametrów i wybiera optymalne parametry dla modelu na podstawie określonej miary oceny - w przypadku tejże analizy korzystając z metryki R^2.

W przypadku modelu Linear Regression pula parametrów była następująca:

- "fit intercept": [True, False] - określający czy obliczać przecięcie (wartość wyrazu wolnego) dla tego modelu.

- "copy X": [True, False] - jeśli True, kopiowany jest zbiór cech X; w przeciwnym razie jest nadpisywany.
- "positive": [True, False] - gdy ustawione na True, wymusza dodatniość współczynników

Dla modelu SVR parametrami do wyboru określono:

- "C": [1.0, 2.0, 4.0] - parametr regularyzacji.
- "gamma": [0.001, 0.1, 1.0, 10.0] - współczynnik jądra (kernel) dla 'rbf'

Ostateczne optymalne dobory uzyskano po 10-stopniowej walidacji krzyżowej. Były to:

- *fit intercept = True, copy X = True, positive = True* dla Linear Regression
- *C = 4.0, gamma = 0.001* - dla SVR

4.5 Określenie jakości modeli

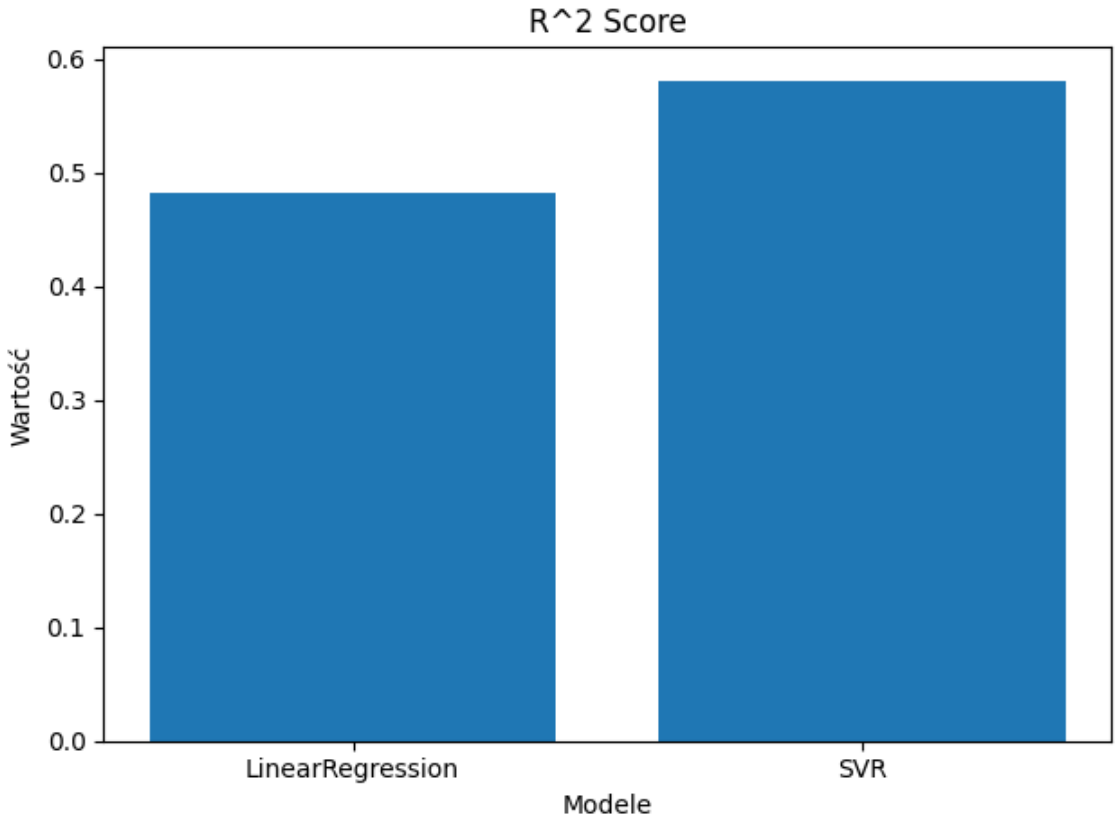
Aby wyeleminować dopasowania modelu do konkretnych podziałów danych wyniki zostały osiągnięte na podstawie średniej z przeprowadzonej 5-stopniowej walidacji krzyżowej.

	R^2 Score	MSE	RMSE
Linear Regression	0.48	171.07	13.08
SVR	0.58	138.48	11.76

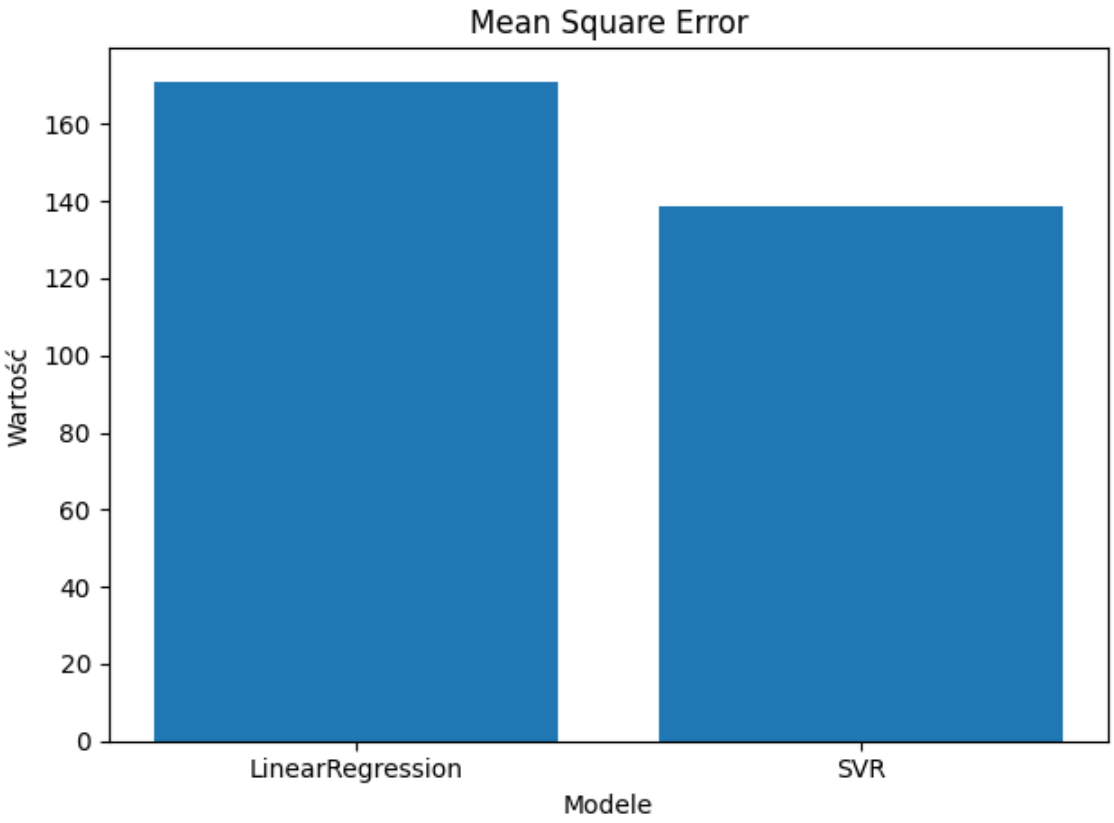
Tabela 2: Wartości oceny modeli w zależności od przyjętej metryki

5 Podsumowanie

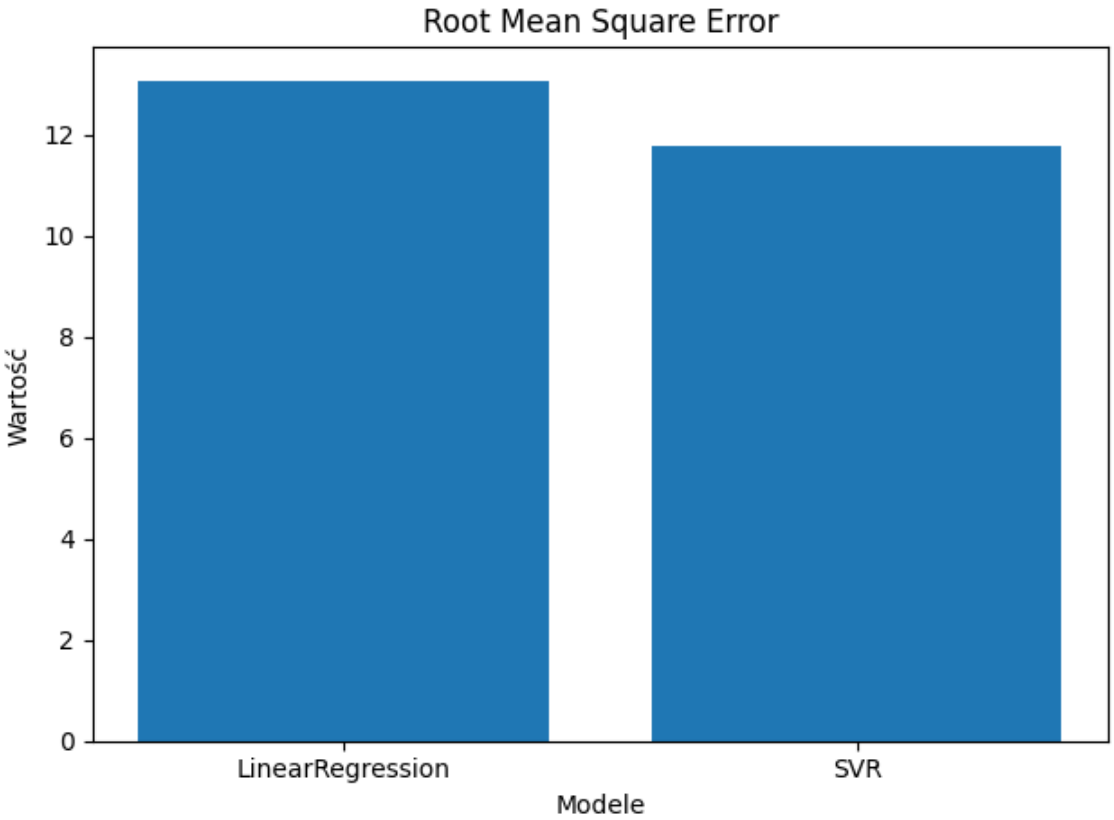
5.1 Wizualizacja wyników modeli



Rysunek 10: Porównanie wyników modeli za pomocą metryki R^2 Score



Rysunek 11: Porównanie wyników modeli za pomocą metryki MSE



Rysunek 12: Porównanie wyników modeli za pomocą metryki RMSE

5.2 Wnioski

- Zgodnie z przewidywaniami znaleziona została zależność między niektórymi danymi związanymi z utworem a jego popularnością.
- Można stwierdzić, że dla podanych wyżej parametrów i cech model SVR (Support Vector Regression) osiąga lepsze wyniki niż model LinearRegression w problemie oceny popularności utworów ze Spotify.
- Posiada on wyższy R^2 Score (0.5809 w porównaniu do 0.4816) oraz niższe wartości MSE i RMSE (138.4832 i 11.7640 w porównaniu do 171.0678 i 13.0785), co wskazuje na mniejsze błędy predykcji i lepsze dopasowanie do danych.
- Oba modele mają wartości R^2 Score wyższe od 0, co sugeruje, że mają pewne zdolności przewidywania popularności piosenek. Jednakże, wartości te wskazują, że nadal istnieje znaczna wariancja, której modele nie są w stanie wyjaśnić.
- W przypadku obu modeli, wartości MSE oraz RMSE są stosunkowo wysokie, co sugeruje, że modele mają tendencję do popełniania znacznych błędów w przewidywaniu popularności piosenek.
- Uzyskany model SVR objaśnia około 58% wariancji w popularności utworów (48% w przypadku LinearRegression) oraz myli się z predykcją jego popularności o średnio 11.764 (13.078 LinearRegression).
- Mimo wszystko uzyskane wyniki można uznać za zadowalające, zważając na fakt, iż jest to alternatywne podejście do oceny popularności utworów do tego proponowanego przez autorów zbioru danych z Web API Spotify.

5.3 Ulepszenia

- Podstawowym czynnikiem, który mógłby znacząco ulepszyć modele byłby równomierny rozkład popularności oraz idąca za tym większa liczba badanych utworów - jest szansa, iż modele poprawniej oceniałyby skrajne wartości popularności, niż ma to miejsce obecnie.
- Wpłynąć na ostateczne wyniki mogłaby także 'obróbka' powtarzających się wartości *0* w parametrze *album popularity* - przykładowo wstawiając w ich miejsce średnią wartość odnotowaną u pozostałych próbek lub całkowicie odrzucając te wartości ze zbioru danych.
- Kwestią do przemyślenia pozostaje także wpływ parametru *followers per artist* na wyniki eksperymentu - prawdopodobnym jest, iż zastosowanie innego podejścia do jego przetworzenia niż zaproponowane w niniejszym raporcie dałoby zauważalną poprawę wydajności modeli.
- Możliwe jest również, że przewidywanie okazałoby się skuteczniejsze w przypadku zastosowania innych, bardziej skomplikowanych modeli.