

Uniwersytet Warszawski
Wydział Nauk Ekonomicznych

Michał Budzyński
Nr albumu: 441444

**When do songs go viral? - Czynniki wpływające na
pozycję utworu w rankingu Spotify „Weekly Top
Songs Global”**

Praca zaliczeniowa
z przedmiotu Ekonometria

Praca wykonana pod kierunkiem
mgr Rafała Walaska
z Katedry Statystyki i Ekonometrii
WNE UW

Warszawa, styczeń 2024

ROZDZIAŁ I

Przegląd literatury

1.1. Spotify jako globalny lider na rynku platform specjalizujących się w streamingu muzyki

Spotify to szwedzki serwis streamingowy oferujący swoim użytkownikom dostęp do nielimitowanej muzyki oraz podcastów. W roku 2017 platforma Spotify zgromadziła ponad 100 milionów użytkowników, w tym 60 milionów użytkowników, którzy wykupili konto premium (Fleischer i Snickars, 2017). Jak stwierdza Hracs oraz Webster (2020), Spotify ukształtowało się jako lider wśród platform specjalizujących się w streamowaniu muzyki, pomimo, iż początkowo budżet aplikacji był niewielki na tle innych firm (m.in. Apple, Amazon, Google), które także posiadały własne platformy służące do streamingu muzyki. Spotify jest jedną z największych spółek na oligopolistycznym rynku muzycznych serwisów streamingowych, generując około 27% przychodów wspomnianego rynku (Tofalvy oraz Koltai, 2023). Taka sytuacja pozwala na określenie jej mianem jednej z wiodących platform streamingowych w dzisiejszej erze.

1.2 Jak atrybuty utworu wpływają na jego popularność

Platforma Spotify wyróżnia szereg charakterystyk, którymi opisane są utwory muzyczne (ang. song features). Zaczynając od relatywnie obiektywnych cech, takich jak tonacja (ang. key), modus (ang. mode) czy tempo, aż po bardzo subiektywne, wyliczane przez algorytm miary, takie jak taneczność (ang. danceability), akustyczność (ang. acousticness) czy pozytywność utworu określana przez zmienną *Valence* (Askin oraz Mauskopf, 2017). Niniejsze charakterystyki stanowią podwaliny oraz fundamenty prężnie rozwijającej się dziedziny nauki określanej jako „Hit Song Science” (HSS) (Oliveira i inni, 2022). Głównym celem przyświecającym HSS jest predykcja popularności utworu bazując na jego cechach. HSS jest także ściśle związane z obszarem „Music Information Retrieval” (MIR), który specjalizuje się w zrozumieniu struktury rynku muzycznego oraz cech utworu, które w sposób istotny przyczyniają się do zwiększenia jego potencjału na komercyjny sukces (Pachet, 2011). Jedną z najbardziej intrygujących kwestii, które stara się rozwiązać HSS, jest wyłonienie tych cech, które są istotne przy oszacowaniu

potencjału danego utworu (ang. hit potential), spośród szerokiego wachlarza pozostałych charakterystyk, które mogą okazać się nieistotne (Oliveira i inni, 2022).

HSS oraz MIR odgrywają kluczową rolę wśród platform specjalizujących się w streamowaniu muzyki (Montecchio, 2019). W kontekście Spotify, są one niezbędne przy tworzeniu spersonalizowanych rekomendacji muzycznych bazujących na głębokim uczeniu maszynowym (ang. deep learning) (Vall i inni, 2019). Ponadto, przewidywanie popularności utworów na podstawie ich cech jest szczególnie istotne dla firm mających na celu rozpoznanie preferencji swoich konsumentów i zwiększenie ich satysfakcji (Çimen oraz Kayis, 2021). MIR pomaga sprostać oczekiwaniom użytkowników w kontekście odkrywania nowych utworów (Herremans i inni, 2014).

Odnosząc się do ekonomii behawioralnej, można stwierdzić, że konsumenci mają ograniczone zasoby czasu i uwagi. Z tego powodu pozycja utworu w rankingu „*Weekly Top Songs Global*” może wynikać z ograniczeń czasowych słuchaczy. Dłuższe utwory mogą wymagać większej inwestycji czasowej, co może wpływać na decyzje słuchaczy. W konsekwencji pojawia się podejrzenie, że istnieje pewna optymalna długość utworu, po której słuchacz traci dalsze zainteresowanie. Z własnych obserwacji wynika również, że utwory, które stają się „virałem” są relatywnie krótkie.

Ponadto, teoria ekonomii behawioralnej wskazuje, że decyzje konsumentów są często kształtowane przez czynniki psychologiczne i emocjonalne. Własne przypuszczenia sugerują, że poziom pozytywności utworu (jego brzmienia) może wpływać na preferencje słuchaczy, ale jednocześnie, oznaczenie treści jako nieprzyzwoitej może modyfikować tę relację. Oprócz brzmienia utworu (pozytywnego lub negatywnego), istotny jest charakter jego treści. Na przykład, utwór może przekazywać wulgarną treść w pozytywnym lub negatywnym brzmieniu. Wtedy, jego pozycja w rankingu „*Weekly Top Songs Global*” pozwoli określić, czy słuchacz preferuje utwory o wulgarnej treści i negatywnym brzmieniu, czy woli otrzymywać wulgarne treści w bardziej wysublimowany sposób.

Co więcej badanie przeprowadzone przez Vox (2022) sugeruje, iż popularność utworu na platformie Spotify może być kształtowana nie tylko przez jego muzyczne atrybuty, ale również przez jego popularność na innych platformach. Wspomniane badanie prezentuje proces określany przez autorów jako „TikTok-to-Spotify Pipeline”. Polega on na przepływie użytkowników TikTok’a na platformę Spotify w celu odnalezienia utworu, który usłyszeli na platformie TikTok. W ten sposób użytkownicy TikTok’a generują kolejne wyświetlenia utworów na platformie Spotify. Co więcej, teoria ekonomii sugeruje, że konsumenci często

kierują się trendami przy podejmowaniu decyzji zakupowych, w tym wyborze muzycznych preferencji. Zatem podążanie za trendami generowanymi na platformie TikTok i odnajdywanie ich na platformie Spotify wydaje się być prawdopodobne.

Na podstawie wykazanych wyżej publikacji oraz luk badawczych wyodrębniono główną hipotezę badawczą oraz dwie hipotezy poboczne:

Hipoteza główna: *Im większa popularność utworu na platformie TikTok, tym jego lepsza pozycja w rankingu „Weekly Top Songs Global” by Spotify.*

Hipoteza 2: *Wpływ poziomu pozytywności utworu na jego pozycję w rankingu „Weekly Top Songs Global”, zależy dodatkowo od tego czy treść utworu oznaczona jest jako nieprzystojna.*

Hipoteza 3: *Relacja pomiędzy długością utworu, a jego pozycją w rankingu „Weekly Top Songs Global” jest nieliniowa. (kwadratowa – do pewnej długości pozycja utworu w rankingu staje się lepsza, po osiągnięciu tej liczby pozycja utworu pogarsza się.)*

ROZDZIAŁ II

Opis danych.

2.1 Bazy danych

Próba służąca do przeprowadzenia badania zawiera utwory, które w latach 2020-2021 przynajmniej raz wystąpiły w zestawieniu „Top 200 Weekly (Global)”, które tworzone jest przez platformę Spotify. Powyższe zestawienie aktualizowane jest co tydzień i zawiera najczęściej odtwarzane utwory na Spotify w danym tygodniu. Baza danych została dodatkowo wzbogacona o indeks popularności utworów na platformie TikTok oraz informację, czy dany utwór jest oznaczony przez Spotify jako wulgarny. Wszystkie wyżej wspomniane zbiory danych pochodzą z platformy Kaggle. Finalna baza danych została oczyszczona z duplikatów i braków danych, w wyniku czego pozostało 101 obserwacji. Za obserwacje uznaje się utwór muzyczny. Spośród wielu zmiennych w bazie danych empirycznie wybrano te najistotniejsze dla niniejszego badania, które wymienione są w poniższej tabeli. Warto nadmienić, iż wszystkie wartości niżej wymienionych zmiennych generowane są przez platformę Spotify, z wyjątkiem zmiennej oznaczającej popularność utworu na platformie TikTok.

Tabela 1. Opis zmiennych.

Nazwa zmiennej	Znaczenie	Rodzaj zmiennej oraz przyjmowane wartości
Highest charting position	Zmienna objaśniana. Najwyższa pozycja osiągnięta przez dany utwór w rankingu „Weekly Top Songs Global”	Zmienna ciągła przyjmująca wartości liczb całkowitych z przedziału [1;200]
TikTok track popularity	Popularność utworu na platformie TikTok. Wewnętrzny indeks wyliczany przez algorytm platformy TikTok.	Zmienna ciągła przyjmująca wartość dowolnej liczby całkowitej z przedziału [0;100]
Explicit	Odpowiada na pytanie czy dany utwór oznaczony jest przez Spotify jako nieprzystojny/brutalny/wulgarny.	Zmienna dychotomiczna. Przyjmuje wartości: 1 - utwór oznaczony jako nieprzystojny 0 -brak oznaczenia jako nieprzystojny

Danceability	Taneczność utworu, określa w jakim stopniu dany utwór jest taneczny.	Zmienna ciągła przyjmująca wartości z przedziału [0.0; 1.0]
Energy	Określa w jakim stopniu dany utwór jest energiczny, intensywny.	Zmienna ciągła przyjmująca wartości z przedziału [0.0; 1.0]
Loudness	Domyślna głośność utworu w decybelach.	Zmienna ciągła, która przyjmuje dowolne wartości. W większości przypadków są to wartości pomiędzy -60 i 0 db.
Speechiness	Opisuje ilość mowy/słów/wokalu w danym utworze.	Zmienna ciągła przyjmująca wartości z przedziału [0.0; 1.0]
Liveness	Ukazuje poziom obecności widowni w nagraniu np. jeżeli zmienna <i>Liveness</i> przyjmuje wartości bliskie 1 to bardzo prawdopodobne jest, że ścieżka dźwiękowa utworu pochodzi z koncertu.	Zmienna ciągła przyjmująca wartości z przedziału [0.0; 1.0]
Duration	Długość utworu w minutach.	Zmienna ciągła przyjmująca wartość dowolnej liczby dodatniej.
Valence	Poziom pozytywność utworu, w jakim stopniu utwór jest pozytywny. Utwory o wysokim indeksie <i>Valence</i> brzmią bardziej radośnie, o niższym - bardziej agresywnie, depresyjnie.	Zmienna ciągła przyjmująca wartości z przedziału [0.0; 1.0]

Źródło: Opracowanie własne na podstawie Spotify API oraz TikTok API

ROZDZIAŁ III

Model ekonometryczny

3.1. Konstrukcja modelu.

W niniejszym badaniu przyjęto poziom istotności 10% dla wszystkich testów.

Za zmienną objaśnianą wskazano zmienną *highest charting position*, podczas gdy pozostałe zmienne wykazane w tabeli 1 uznane zostały za zmienne objaśniające. W związku z powyższym, model opiera się na założeniu, że pozycja utworu w rankingu „Weekly Top Songs Global” nie ma wpływu na jego popularność na TikTok’u. Taka decyzja ma na celu spełnienie założenia o braku wpływu zmiennej objaśnianej *highest charting position* na zmienną objaśniającą *TikTok track popularity*.

W celu przekształcenia zmiennych w kierunku relacji liniowej dokonano logarytmowania zmiennych *Liveness* oraz *Speechiness*. W celu weryfikacji postawionych wcześniej hipotez, zdecydowano również o wprowadzeniu do modelu zmiennej *Duration* w relacji kwadratowej oraz zastosowaniu interakcji między zmiennymi *Valence* i *Explicit*. Ponadto, wszystkie wspomniane transformacje dokonane zostały na podstawie wcześniejszej analizy wykresów rozrzutu zmiennych.

Przed przystąpieniem do selekcji zmiennych wykonano test na łączną nieistotność zmiennych objaśniających. Wartość p-value niniejszego testu wyniosła około 0.07, co pozwoliło odrzucić hipotezę zerową na poziomie istotności 10%. Zatem co najmniej jedna zmienna objaśniana jest statystycznie istotna.

Tabela 2. Test na łączną nieistotność zmiennych objaśniających.

Nazwa testu	Hipoteza zerowa	Statystyka testowa	p-value	Decyzja weryfikacyjna
Test na łączną nieistotność zmiennych.	Zmienne są łącznie nieistotne.	F = 1.776	0.07	Odrzucenie hipotezy zerowej.

Źródło: Opracowanie własne

Następnie na podstawie macierzy korelacji dokonano analizy korelacji między zmiennymi. Zastosowano korelację rang Spearman’a z uwagi na podejrzenie, iż nie wszystkie zmienne posiadają rozkład normalny. W wyniku analizy ujawniono wysoką korelację (około

0.70) między zmiennymi objaśniającymi *Loudness* oraz *Energy*. Należy nadmienić, iż będzie to szczególnie istotne, gdy w finalnym modelu umieszczone zostaną obydwie te zmienne, ponieważ ich wysoka korelacja może zaburzać wyniki modelu i powodować problem współliniowości. Ponadto, z uwagi na wysoką korelację, istnieje podejrzenie, że zmienne mogą opisywać to samo zjawisko. Nie zdecydowano się jednak na usunięcie żadnej ze wspomnianych wcześniej zmiennych na tym etapie. Ewentualne ich usunięcie rozważane będzie w finalnej wersji modelu, gdy okaże się, że po dokonaniu procesu selekcji zmiennych za pomocą testu istotności pozostaną one w modelu. Równie niepokojące okazało się być odkrycie, iż wszystkie zmienne objaśniające wykazują relatywnie niską korelację (około 0.1) ze zmienną objaśnianą *highest charting position*. Jedynie zmienna *tiktok track popularity* wyróżnia się swoją umiarkowaną korelacją wynoszącą około -0.35. Może to wskazywać, iż szacowany na podstawie tych zmiennych objaśniających model nie będzie opisywał w zadowalającym stopniu całkowitej zmienności zmiennej objaśnianej, a zmienne objaśniające okażą się być nieistotne. Jednakże, na tym etapie nie zdecydowano się na usunięcie żadnej ze zmiennych objaśniających.

W dalszej kolejności zmienne objaśniające zostały wyselekcjonowane stosując metodę „od ogólnego do szczegółowego”. Zmienne objaśniające były usuwane pojedynczo. Założono, iż stała musi pozostać w modelu niezależnie od jej statystycznej istotności. Każde usunięcie zmiennej objaśniającej z modelu poprzedzone było wykonaniem testu na istotność danej zmiennej, którego hipoteza zerowa mówi o braku istotności zmiennej. Zatem zmienne z najwyższym p-value dla statystyki testowej „t” były kolejno usuwane z modelu. Za każdym razem przeprowadzany był również test na łączną nieistotność usuniętych z modelu zmiennych objaśniających. Ponadto, przy każdym usunięciu zmiennej objaśniającej analizowano zmianę skorygowanego współczynnika determinacji. Przy usunięciu zmiennej statystycznie nieistotnej oczekuje się wzrostu skorygowanego współczynnika determinacji. Wspomniany proces selekcji przedstawia poniższy tabela.

Tabela 3. Selekcja zmiennych do modelu.

Dependent variable: highest_charting_position								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Danceability	57.501 (54.672)	57.696 (54.264)	56.090 (53.435)	62.150 (51.567)	49.416 (49.382)			
Duration	-25.063 (47.089)	-25.165 (46.794)	-25.100 (46.546)	-28.315 (45.823)	-22.494 (45.272)	-24.583 (45.224)	22.129*** (8.126)	19.286** (7.955)
Duration_sq	7.196 (6.900)	7.222 (6.846)	7.190 (6.808)	7.708 (6.686)	6.970 (6.623)	6.954 (6.623)		
Energy	11.212 (56.343)	9.212 (44.077)						
Explicit	47.086 (32.368)	47.161 (32.163)	47.435 (31.966)	47.290 (31.828)	47.874 (31.779)	52.842* (31.389)	58.902* (30.871)	61.415* (31.028)
Intercept	-23.445 (110.343)	-20.165 (93.921)	-14.278 (89.126)	-6.099 (86.965)	6.912 (85.557)	36.193 (80.398)	-40.736 (33.120)	-58.854* (31.056)
Loudness	-0.249 (4.338)							
Valence	57.205 (45.132)	56.861 (44.485)	60.192 (41.312)	59.734 (41.123)	64.489 (40.705)	81.370** (37.045)	89.004** (36.344)	90.249** (36.573)
Valence:Explicit	-90.859 (55.785)	-90.976 (55.438)	-92.277* (54.798)	-93.613* (54.486)	-100.175* (53.891)	-101.834* (53.866)	-112.217** (52.979)	-112.704** (53.325)
log_Liveness	-4.082 (8.612)	-4.038 (8.530)	-3.896 (8.458)					
log_Speechiness	-7.503 (8.466)	-7.423 (8.304)	-7.404 (8.260)	-7.137 (8.204)				
tiktok_track_popularity	-0.377 (0.253)	-0.377 (0.251)	-0.378 (0.250)	-0.369 (0.248)	-0.339 (0.245)	-0.329 (0.245)	-0.365 (0.243)	
Observations	101	101	101	101	101	101	101	101
R ²	0.180	0.180	0.180	0.178	0.171	0.162	0.152	0.132
Adjusted R ²	0.079	0.089	0.098	0.106	0.109	0.108	0.108	0.096
Residual Std. Error	53.370 (df=89)	53.073 (df=90)	52.794 (df=91)	52.567 (df=92)	52.499 (df=93)	52.499 (df=94)	52.527 (df=95)	52.872 (df=96)
F Statistic	1.776* (df=11; 89)	1.975** (df=10; 90)	2.213** (df=9; 91)	2.485** (df=8; 92)	2.739** (df=7; 93)	3.028*** (df=6; 94)	3.410*** (df=5; 95)	3.648*** (df=4; 96)

Note:

*p<0.1; **p<0.05; ***p<0.01

Źródło: Opracowanie własne

Tabela 4. Test na łączną nieistotność zmiennych objaśniających usuwanych kolejno z modelu.

Nazwa testu	Hipoteza zerowa	Zmienne	Statystyka testowa	p-value	Decyzja weryfikacyjna
Test na łączną nieistotność zmiennych.	Zmienne są łącznie nieistotne.	<i>Loudness</i>	F = 0.003	0.95	Brak podstaw do odrzucenia hipotezy zerowej.
		<i>Loudness</i> <i>Energy</i>	F=0.023	0.98	Brak podstaw do odrzucenia hipotezy zerowej.
		<i>Loudness</i> <i>Energy</i> <i>log_Liveness</i>	F=0.085	0.97	Brak podstaw do odrzucenia hipotezy zerowej.
		<i>Loudness</i> <i>Energy</i> <i>log_Liveness</i> <i>log_Speechiness</i>	F=0.247	0.91	Brak podstaw do odrzucenia hipotezy zerowej.

		<i>Loudness</i> <i>Energy</i> <i>log_Liveness</i> <i>log_Speechiness</i> <i>Danceability</i>	F=0.391	0.85	Brak podstaw do odrzucenia hipotezy zerowej.
		<i>Loudness</i> <i>Energy</i> <i>log_Liveness</i> <i>log_Speechiness</i> <i>Danceability</i> <i>Duration_sq</i>	F=0.504	0.8	Brak podstaw do odrzucenia hipotezy zerowej.
		<i>Loudness</i> <i>Energy</i> <i>log_Liveness</i> <i>log_Speechiness</i> <i>Danceability</i> <i>Duration_sq</i> <i>tiktok_track_popularity</i>	F=0.745	0.63	Brak podstaw do odrzucenia hipotezy zerowej.

Źródło: Opracowanie własne

Selekcja zmiennych objaśniających do modelu następuję zgodnie z procedurą w przypadku modeli od 1 do 3. W przypadku modelu 4 oraz 5 zdecydowano o usunięciu zmiennej o drugim najwyższym p-value, ponieważ w tych modelach zmienna *Duration* posiada najwyższe p-value. Jednakże, zmienna ta występuje również w relacji kwadratowej pod postacią zmiennej *Duration_sq*, która nie posiada już najwyższego p-value, lecz znacznie niższe. Warto nadmienić, iż w modelu nie przewiduje się pozostawienia jedynie zmiennej podniesionej do relacji kwadratowej bez pierwotnej zmiennej, która nie jest podniesiona do kwadratu. Szczególnej uwagi wymaga także usunięcie zmiennej *Danceability* z modelu 5. Wspomniana zmiana spowodowała obniżenie się skorygowanego współczynnika determinacji o 0.001. Z uwagi na relatywnie niską zmianę niniejszego współczynnika, potwierdzenie o statystycznej nieistotności zmiennej *Danceability*, potwierdzonej zarówno za pomocą testu „t” na istotność niniejszej zmiennej, jak i testu „F” na łączną nieistotność zmiennych, zdecydowano o poprawności podjętej selekcji i kontynuowano tworzenie modelu. W przypadku modelu 6 zdecydowano o usunięciu zmiennej *Duration_sq*, ponieważ niniejsza zmienna oraz zmienna *Duration* posiadają dwie najwyższe wartości p-value spośród wszystkich zmiennych objaśnianych znajdujących się w tym modelu. Po usunięciu wspomnianej zmiennej skorygowany współczynnik determinacji w przybliżeniu się nie zmienia. Warto nadmienić, iż wskazuje to na brak podstaw do umieszczenia relacji nieliniowej (kwadratowej) zmiennej *Duration* w finalnym modelu. W następnym kroku

zdecydowano się na usunięcie zmiennej *tiktok_track_popularity* z modelu 7. Na słuszność takiej decyzji wskazywały zarówno test „t” na istotność niniejszej zmiennej, jak i test „F” na łączną nieistotność zmiennych. Jednakże, po usunięciu wspomnianej zmiennej wartość skorygowanego współczynnika determinacji zdecydowanie zmalała. Uznano tę zmianę za szczególnie istotną i z tego powodu zdecydowano o niesłuszności tego kroku w selekcji zmiennych. Z uwagi na powyższe, za najlepszy i finalny model uznano model 7. Niniejszy model posiada relatywnie wysoką wartość skorygowanego współczynnika determinacji (różnica z najwyższą wartością skorygowanego współczynnika determinacji wynosi 0.001 i uznawana jest za marginalną). Mimo, że model ten zawiera zmienną statystycznie nieistotną to ma on wyższy skorygowany współczynnik determinacji niż model 8, zawierający jedynie zmienne statystycznie istotne. Warto jednak podkreślić, iż zawarcie zmiennej statystycznie nieistotnej w modelu może negatywnie rzutować na wyniki testów diagnostycznych oraz oszacowania modelu.

3.2. Interpretacja oszacowań modelu.

Skonstruowany model ekonometryczny przedstawiony jest poniższym równaniem:

$$\text{chart_peak} = -40.736 + 22.129 * \text{Duration} - 0.365 * \text{TikTok} + 58.902 * \text{Explicit} + 89.004 * \text{Valence} - 112.217 * \text{Explicit} * \text{Valence}$$

gdzie:

chart_peak – zmienna objaśniana – najwyższa pozycja osiągnięta przez utwór w zestawieniu Spotify „Top 200 Weekly (Global)”

Duration – długość utworu wyrażona w minutach

TikTok – popularność utworu na platformie TikTok

Explicit – zmienna dychotomiczna określająca czy utwór oznaczony jest przez Spotify jako wulgarny

Valence - Poziom pozytywność utworu, w jakim stopniu utwór jest pozytywny

Przed przystąpieniem do interpretacji oszacowań modelu warto podkreślić specyficzny charakter zmiennej objaśnianej. Niniejsza zmienna odnosi się do najwyższej pozycji osiągniętej w rankingu. Oznacza to, że im mniejsza wartość zmiennej objaśnianej, tym

wyższa pozycja w rankingu. W celu łatwiejszej interpretacji w kontekście zmiennej objaśnianej posłużono się stwierdzeniem „wyższa pozycja w rankingu”, która odnosi się do spadku wartości zmiennej *chart_peak*, oraz „niższa pozycja w rankingu”, która odnosi się do wzrostu wartości wspomnianej zmiennej.

Współczynnik zmienności R^2 dla niniejszego modelu wyniósł 0,152, co oznacza, że wyjaśnione zostało 15,2% całkowitej zmienności zmiennej objaśnianej oznaczającej najwyższą pozycję osiągniętą przez utwór w zestawieniu Spotify „Top 200 Weekly (Global)”. Z modelu wynika, że wzrost długości trwania utworu o minutę przełoży się na jego spadek w rankingu o około 22,129 miejsca *ceteris paribus*. Ponadto, wzrost poziomu pozytywności utworu (*Valence*) o 1, gdy utwór nie jest oznaczony jako wulgarny, przełoży się na jego spadek w rankingu o około 89,004 miejsca *ceteris paribus*. Natomiast, gdy utwór jest oznaczony jako wulgarny to wzrost poziomu pozytywności tego utworu (*Valence*) o 1 przełoży się na jego wyższą pozycję w rankingu o około 23,213 miejsca *ceteris paribus*. Warto jednak nadmienić, iż takie interpretacje zmiennej *Valence* mogą nie mieć swojego praktycznego przełożenia, gdyż niniejsza zmienna przyjmuje jedynie wartości z przedziału od 0 do 1. Zatem wzrost tej zmiennej o 1 jest niemalże niespotykany. W tym przypadku ukazano również słuszność umieszczenia w niniejszym modelu interakcji między zmiennymi *Explicit* oraz *Valence*. Wynika to z faktu, iż jak zostało przedstawione powyżej, wpływ poziomu pozytywności utworu na jego pozycję w rankingu zależy dodatkowo od tego czy jest on oznaczony jako wulgarny. Nie podjęto interpretacji parametru dla popularności utworu na platformie TikTok, ponieważ w poprzednim podrozdziale został on uznany za statystycznie nieistotny.

3.3. Diagnostyka modelu – założenia klasycznego modelu regresji liniowej.

W celu diagnostyki modelu rozważone zostały założenia klasycznego modelu regresji liniowej (KMRL). W przypadku nie spełnienia przez model niektórych założeń wyniki prezentowane przez model mogą być obciążone błędem, a co za tym idzie nieprawdziwie obrazujące rzeczywiste mechanizmy.

Pierwszym z założeń jest założenie o poprawności formy funkcyjnej. Niniejsze założenie analizowano za pomocą testu RESET, którego hipoteza zerowa zakłada liniowość formy funkcyjnej modelu. W wyniku tego testu otrzymano p-value większe od 0.1, co skutkowało w stwierdzeniu braku podstaw do odrzucenia hipotezy zerowej. Zatem forma

funkcyjna modelu jest liniowa. Ponadto, przeprowadzono test Chow'a, którego hipoteza zerowa mówi o stabilności parametrów modelu. W przypadku testu Chow'a otrzymano wartość statystyki F równą około 0.2529, która nie należy do wyliczonego obszaru krytycznego. Na tej podstawie stwierdzono brak podstaw do odrzucenia hipotezy zerowej o stabilności parametrów w modelu. Powyższe testy pozwoliły na potwierdzenie, że niniejszy model spełnia założenie KMRL o poprawności formy funkcyjnej. W przypadku, gdy model nie posiada poprawnej formy funkcyjnej, można spodziewać się, że estymator będzie obciążony. Jednym ze skutków niepoprawnej formy funkcyjnej może być również brak interpretacji parametrów lub łatwość podważenia tej interpretacji oraz brak możliwości udowodnienia własności estymatora.

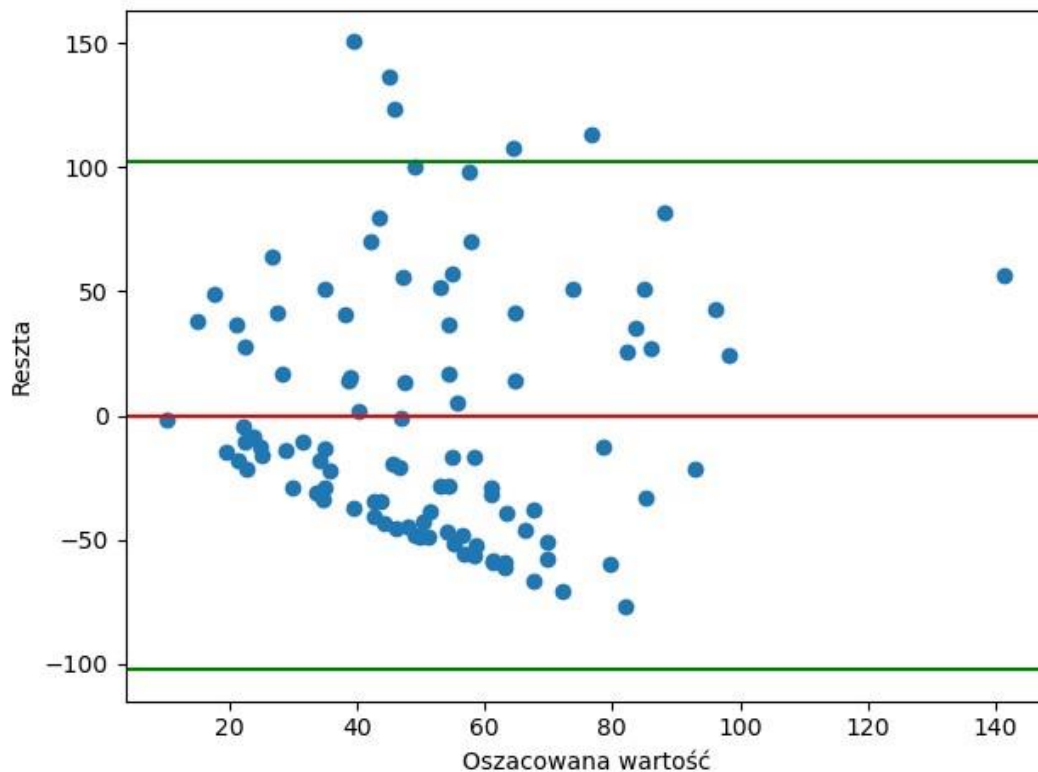
Tabela 5. Testy na poprawność formy funkcyjnej modelu.

Nazwa testu	Testowane założenie KMRL	Hipoteza zerowa	Statystyka testowa	p-value	Decyzja weryfikacyjna
RESET	Poprawność formy funkcyjnej.	Forma funkcyjna modelu jest liniowa.	3.51	0.17	Brak podstaw do odrzucenia hipotezy zerowej.
Chow'a	Poprawność formy funkcyjnej.	Parametru modelu są stabilne.	2.317	-	Brak podstaw do odrzucenia hipotezy zerowej.

Źródło: Opracowanie własne

Kolejnym weryfikowanym założeniem KMRL było założenie o homoskedastyczności składnika losowego, czyli stałości wariancji składnika losowego w modelu. Niespełnienie tego założenia, czyli występowanie heteroskedastyczności, skutkuje obciążeniem macierzy wariancji kowariancji, a co za tym idzie statystyki testowe używane w modelu mogą być obciążone w konsekwencji powodując zwiększenie się wartości błędu standardowego. W celu weryfikacji niniejszego założenia zinterpretowano poniższy wykres rozrzutu składnika losowego.

Rys. 1. Wykres rozrzutu reszt.



Źródło: Opracowanie własne

W sytuacji występowania homoskedastyczności składnika losowego punkty widoczne na wykresie powinny stanowić swoje lustrzane odbicie względem linii czerwonej ($y=0$). Taka sytuacja zdecydowanie nie jest zauważalna na powyższym wykresie zatem możemy podejrzewać występowanie heteroskedastyczności reszt w analizowanym modelu. W celu dalszej weryfikacji wspomnianego założenia przeprowadzono testy statystyczne Breusch-Pagan'a oraz White'a. Hipoteza zerowa obu testów mówi o stałości wariancji składnika losowego w modelu. Dla obu testów uzyskane wartości p-value były większe od poziomu istotności równego 10%, zatem stwierdzono brak podstaw do odrzucenia hipotezy zerowej o homoskedastyczności składnika losowego. Jednakże, zastanawiająca i alarmująca okazuje się być niezgodność wyniku interpretacji wykresu z wynikami testów statystycznych. Wykres rozrzutu reszt wskazuje na występowanie heteroskedastyczności, natomiast testy statystyczne wskazują na homoskedastyczność. W celu dalszej weryfikacji przeprowadzono test Goldfeld-Quandt'a, dla którego p-value wyniosło 0.02, w konsekwencji pozwoliło to na odrzucenie hipotezy zerowej o homoskedastyczności składnika losowego. Finalnie uznano brak

homoskedastyczności składnika losowego i dokonano próby naprawy za pomocą macierzy White'a odpornej na heteroskedastyczność.

Tabela 6. Testy na stałość wariancji składnika losowego.

Nazwa testu	Testowane założenie KMRL	Hipoteza zerowa	Statystyka testowa	p-value	Decyzja weryfikacyjna
Breusch-Pagan	Stałość wariancji reszt.	Homoskedastyczność składnika losowego.	1.594	0.17	Brak podstaw do odrzucenia hipotezy zerowej.
White			0.952	0.51	Brak podstaw do odrzucenia hipotezy zerowej.
Goldfeld-Quandt			1.929	0.02	Odrzucenie hipotezy zerowej.

Źródło: Opracowanie własne

Tabela 7. Model oszacowany za pomocą macierzy odpornej White'a.

<i>Dependent variable: highest_charting_position</i>		
	(1)	(2)
Duration	22.129*** (8.126)	22.129*** (7.124)
Explicit	58.902* (30.871)	58.902** (27.780)
Intercept	-40.736 (33.120)	-40.736 (27.791)
Valence	89.004** (36.344)	89.004*** (26.516)
Valence:Explicit	-112.217** (52.979)	-112.217** (50.010)
tiktok_track_popularity	-0.365 (0.243)	-0.365* (0.195)
Observations	101	101
R ²	0.152	0.152
Adjusted R ²	0.108	0.108
Residual Std. Error	52.527 (df=95)	52.527 (df=95)
F Statistic	3.410*** (df=5; 95)	5.071*** (df=5; 95)
Note:	* p<0.1; ** p<0.05; *** p<0.01	

Źródło: Opracowanie własne

W powyższej tabeli przedstawione są oszacowania modelu bez stosowania macierzy odpornej na heteroskedastyczność (1) oraz modelu oszacowanego za pomocą niniejszej macierzy (2).

Oszacowania parametrów obu modeli w przybliżeniu się od siebie nie różnią. Na tej podstawie możemy wnioskować, iż występowanie heteroskedastyczności w tym modelu nie ma istotnego wpływu na wyniki regresji.

Następnie zweryfikowano założenie KMRL o braku autokorelacji reszt. Brak spełnienia niniejszego założenia może również skutkować obciążeniem macierzy wariancji kowariancji i w konsekwencji zwiększeniu się wartości błędu standardowego. Przeprowadzono test Breusch-Godfrey'a, którego hipoteza zerowa testu mówi o braku autokorelacji pomiędzy wartościami składnika losowego. Hipoteza alternatywna mówi natomiast o występowaniu autokorelacji, i najczęściej występuje w przypadku analizowania danych czasowych lub przestrzennych. W wyniku powyższego testu otrzymano p-value mniejsze od przyjętego poziomu istotności, zatem odrzucono hipotezę zerową o braku autokorelacji reszt. Zatem założenie KMRL o braku autokorelacji reszt nie zostało spełnione i będą wynikać z tego faktu wcześniej wspomniane konsekwencje.

Tabela 8. Test na brak autokorelacji reszt.

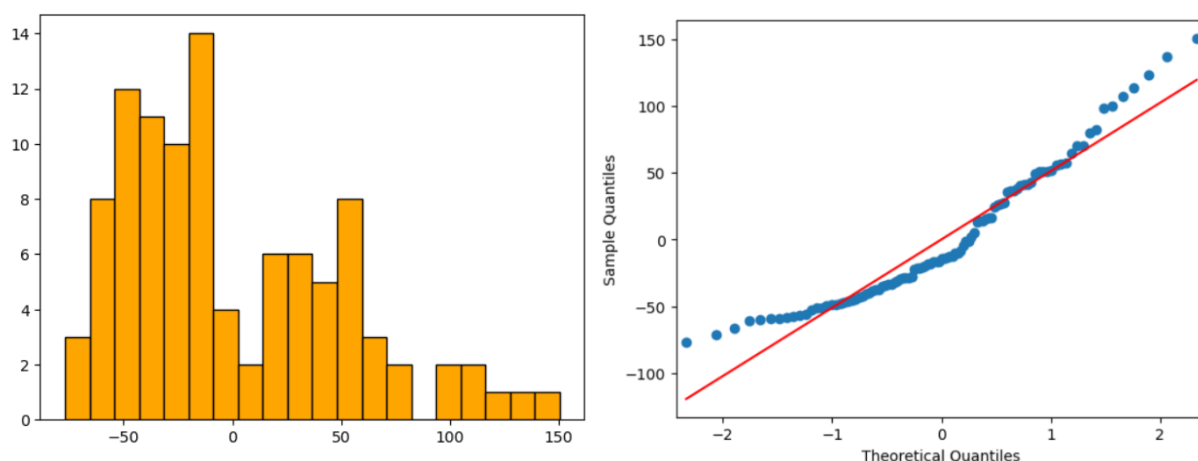
Nazwa testu	Testowane założenie KMRL	Hipoteza zerowa	Statystyka testowa	p-value	Decyzja weryfikacyjna
Breusch-Godfrey	Brak autokorelacji reszt.	Brak autokorelacji reszt.	3.68	0.0004	Odrzucenie hipotezy zerowej.

Źródło: Opracowanie własne

Kolejnym weryfikowanym założeniem KMRL jest założenie o normalności rozkładu składnika losowego. W tym celu przeprowadzono test Jarque-Bera. Hipoteza zerowa testu mówi o normalności rozkładu reszt, natomiast alternatywna o jego braku. W przypadku tego testu, ważne jest zwrócenie uwagi na wielkość próby. Jeżeli próba użyta w badaniu, jest wystarczająco duża, nawet nieprawidłowy wynik testu Jarque – Bera nie daje powodów do obaw, ponieważ nie rzutuje on na wyniki oszacowań modelu. Dodatkowo, poddano analizie poniższe rysunki, które prezentują histogram reszt oraz porównanie rozkładu reszt z rozkładem normalnym. Rysunek 2a przedstawia rozkład, który nie jest zbliżony do rozkładu normalnego, natomiast rysunek 2b potwierdza relatywnie duże różnice między rozkładem reszt modelu oraz rozkładem normalnym. Wyniki testu Jarque-Bera również sugerują brak rozkładu normalnego. Wartość p-value uzyskana w niniejszym teście jest niższa niż przyjęty poziom istotności, co pozwala na odrzucenie hipotezy zerowej o występowaniu normalności

rozkładu. Zatem zarówno z analizy graficznej jak i testu statystycznego wynika, iż założenie KMRL o normalności składnika losowego nie jest spełnione. Jednakże, jak zostało wcześniej wspomniane, ponieważ próba na której szacowany był model jest uznana za dużą to niespełnienie niniejszego założenia nie będzie rzutować na oszacowania i interpretację parametrów modelu.

Rys. 2. Rozkład reszt w modelu: a) histogram rozkładu, b) porównanie z rozkładem teoretycznym



Źródło: Opracowanie własne

Tabela 8. Test na rozkład normalny składnika losowego.

Nazwa testu	Testowane założenie KMRL	Hipoteza zerowa	Statystyka testowa	p-value	Decyzja weryfikacyjna
Jarque-Bera	Normalność rozkładu składnika losowego.	Występuje rozkład normalny.	12.302	0.002	Odrzucenie hipotezy zerowej.

Źródło: Opracowanie własne

3.4. Analiza obserwacji nietypowych oraz wystąpienia problemu współliniowości.

Analizę obserwacji odstających przeprowadzono za pomocą interpretacji rysunku 3, który zawiera wykres obrazujący wartości dźwigni na osi X oraz znormalizowanych reszt na osi Y. Na wykresie za pomocą liczb oznaczone zostały obserwacje z największymi wartościami odległości Cook'a. Pionowa czerwona linia określa wartość krytyczną dźwigni, powyżej której obserwacje uznawane są za nietypowe. Wynosi ona:

$$h_i \geq \frac{2 * (K + 1)}{N}$$

gdzie:

h_i – wartość dźwigni dla i -tej obserwacji

K – liczba zmiennych w modelu, nie licząc stałej

N – liczba obserwacji w modelu

Natomiast poziome czerwone linie wyznaczają granicę dla obserwacji odstających ze względu na reszty standaryzowane. Obserwacje są odstające jeżeli:

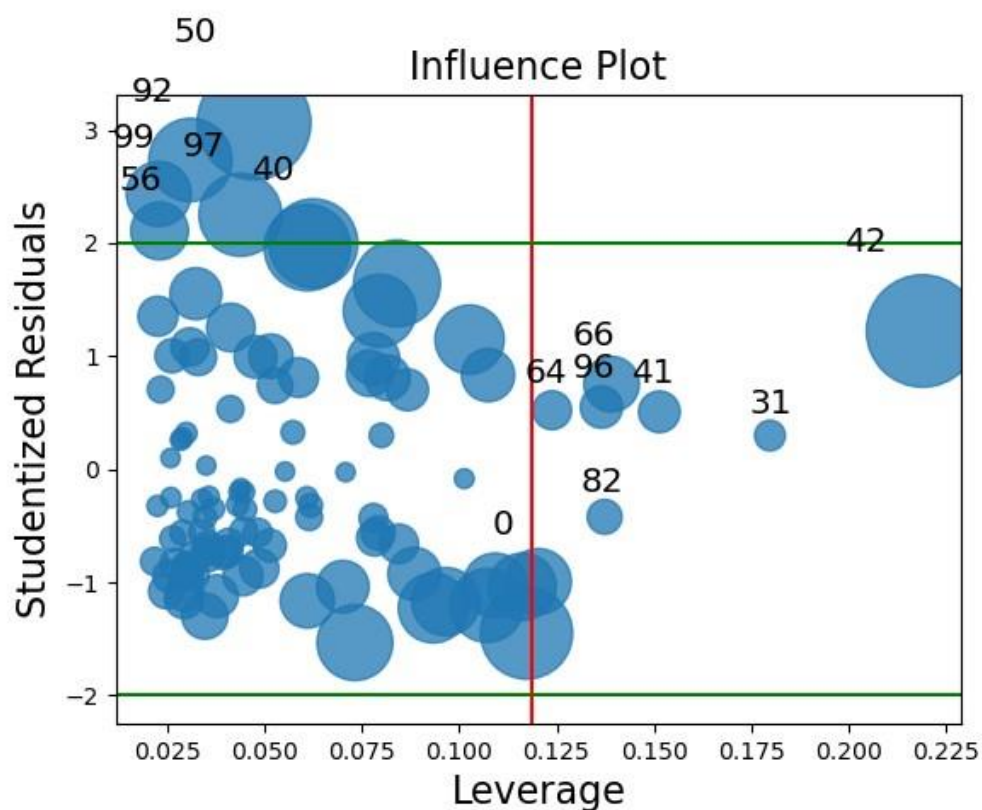
$$|\hat{e}_i| > 2$$

gdzie:

e_i - wartość reszty standaryzowanej do rozkładu normalnego

Łącząc te dwa warunki można stwierdzić, iż nie występują tutaj obserwacje odstające, co ilustruje rysunek 3. Ponadto na podstawie tabeli 9 ukazującej siedem najbardziej odstających obserwacji malejąco według odległości Cook'a, można wnioskować, że wspomniane odległości są relatywnie niskie. Ostatecznie stwierdzono brak obserwacji odstających.

Rys. 3. Wartość dźwigni oraz znormalizowanych reszt.



Źródło: Opracowanie własne

Tabela 9. Odległości Cook'a.

Numer obserwacji	Odległość Cook'a
50	0.071
42	0.069
4	0.046
40	0.043
76	0.041
60	0.040

Źródło: Opracowanie własne

Następnie rozważono wystąpienie problemu współliniowości w niniejszym modelu. Konsekwencją występowania współliniowości zmiennych w modelu jest zwiększenie się wariancji przy pozostałych zmiennych, co skutkuje wzrostem błędu standardowego. Współliniowość można dzielić na dokładną oraz niedokładną. Z współliniowością dokładną mamy do czynienia, gdy jedna z kolumn macierzy obserwacji jest kombinacją liniową pozostałych kolumn – jedna ze zmiennych objaśniających, jest kombinacją liniową innej zmiennej, a co za tym idzie nie wnosi żadnej informacji dodatkowej do modelu. W przypadku

finalnego modelu stwierdzono, iż współliniowość dokładna nie występuje, ponieważ każda zmienna modelu opisuje w pełni inną cechę utworu muzycznego, a co za tym idzie nie może być kombinacją liniową drugiej zmiennej. Natomiast współliniowość niedokładna występuje, gdy zmienne objaśniające są ze sobą bardzo mocno skorelowane. Na podstawie macierzy korelacji stwierdzono, iż silna korelacja między zmiennymi objaśniającymi w tym przypadku nie występuje. Ponadto do analizy problemu współliniowości niedokładnej wykorzystano statystykę VIF, co prezentuje tabela 10. Dla wszystkich zmiennych statystyka VIF kształtowała się na poziomie 1, zatem stwierdzono brak występowania współliniowości niedokładnej.

Tabela 10. Wartości VIF dla poszczególnych zmiennych.

Zmienna	VIF
<i>Duration</i>	1.059
<i>TikTok</i>	1.079
<i>Explicit</i>	1
<i>Valence</i>	1.02

Źródło: Opracowanie własne

PODSUMOWANIE

Cel pracy został osiągnięty poprzez konstrukcję modelu ekonometrycznego, który pozwolił na zidentyfikowanie atrybutów utworu istotnie wpływających na jego pozycję w rankingu „Weekly Top Songs Global” tworzonym przez platformę Spotify. Za pomocą niniejszego modelu udało się negatywnie zweryfikować hipotezę główną mówiącą o wpływie platformy TikTok na pozycję utworu we wspomnianym rankingu. Wykazując jednocześnie brak występowania procesu „TikTok-to-Spotify pipeline” w powyższym badaniu. Odrzucenie hipotezy głównej nastąpiło na podstawie uznania zmiennej określającej poziom popularności utworu na platformie TikTok za statystycznie nieistotną. Ponadto model pozwolił także na zweryfikowanie dwóch hipotez pobocznych postawionych w niniejszym badaniu. Odrzucono hipotezę mówiącą o kwadratowej relacji między długością utworu, a jego pozycją we wspomnianym rankingu. Wykazano, iż występuje brak podstaw do stwierdzenia, że do pewnej długości utworu jego pozycja w rankingu staje się lepsza, natomiast po osiągnięciu tej długości pozycja utworu pogarsza się. Jednakże, pozytywnie zweryfikowano hipotezę wskazującą na to, że wpływ poziomu pozytywności utworu na jego pozycję w rankingu, zależy dodatkowo od tego czy treść utworu oznaczona jest jako nieprzyzwoita. Model ekonometryczny wykazał, że gdy utwór oznaczony jest jako wulgarny to wzrost poziomu pozytywności utworu przełoży się na jego wyższą pozycję w rankingu, natomiast gdy nie ma posiada takiego oznaczenia sytuacja wygląda przeciwnie – wzrost poziomu pozytywności powoduje spadek w rankingu.

Niniejszy model nie spełnia trzech założeń klasycznego modelu regresji liniowej. Jednakże, uznano brak istotnego wpływu niespełnienia założenia o normalności rozkładu składnika losowego oraz homoskedastyczności składnika losowego na oszacowania modelu. Mimo to, do wyników należy jednak podchodzić z rezerwą. Zaprezentowane badanie dotyczyło, relatywnie małej próbki składającej się z 101 obserwacji. Zbadany też został relatywnie krótki czas, bo tylko okres 2 lat. Ponadto, teoria ekonomii behawioralnej sugeruje, iż preferencje konsumentów stale się zmieniają. Zatem model może nie być skuteczny przy analizie bardziej aktualnych danych. W modelu należałoby również rozważyć uwzględnienie kolejnych platform, które mogłyby wpływać istotnie na liczbę odtworzeń utworów na platformie Spotify. Dalsze badania powinny skupić się na weryfikacji czy oprócz istnienia „TikTok-to-Spotify pipeline” zaobserwowane może być również na przykład „Radio-to-Spotify pipeline”.

BIBLIOGRAFIA:

Askin, N. oraz Mauskopf, M. 2017. „What Makes Popular Culture Popular? Product Features and Optimal Differentiation in Music.” *American Sociological Review* 82(5), s. 910-944.

Çimen, A. oraz Kayis, E. 2021. „A longitudinal model for song popularity prediction.” *In Int’l Conf. on Data Science, Technology and Applications (DATA)*, s. 96–104.

Pachet, F. 2011. „Music Data Mining”, CRC Press Taylor & Francis Group, rozdział 10.

Oliveira, G. P., Barbosa, G. R. G., Melo, B. C., Botelho, J. E., Silva, M. O., Seufitelli, D. B. oraz Moro, M. M. 2022. „Musical success in the united states and brazil: novel datasets and temporal analyses.” *Journal of Information and Data Management* 13(1).

Montecchio, N., Roy, P., oraz Pachet, F. 2019. „The skipping behavior of users of music streaming services and its relation to musical structure.”

Vall, A., Dorfer, M., Eghbal-Zadeh, H., Schedl, M., Burjorjee, K. M. oraz Widmer, G. 2019. „Feature-combination hybrid recommender systems for automated music playlist continuation.” *User Modeling and User-Adapted Interaction* 29(2), s. 527-572.

Herremans, D., Martens, D., oraz Sörensen, K. 2014. „Dance hit song prediction.” *Journal of New Music Research* 43(3), s. 291-302.

Gracs, B. oraz Webster, J. 2020. „From selling songs to engineering experiences: exploring the competitive strategies of music streaming platforms.” *Journal of Cultural Economy* 14(2), s. 240-257.

Fleicher, R. oraz Snickars, P. 2017. „Discovering Spotify - A Thematic Introduction.” *Culture Unbound: Journal of Current Cultural Research* 9, s. 130-145.

Vox 2022, „The Unlikely Odds of Making it Big on TikTok”, online: <https://pudding.cool/2022/07/tiktok-story/>, dostęp: 18.01.2024.

Tofalvy, T. oraz Koltai, J. 2023. „“Splendid Isolation”: The reproduction of music industry inequalities in Spotify’s recommendation system.” *New Media & Society* 25(7), s. 1580-1604.