

# Raport zaliczeniowy

Mateusz Luberda

Czerwiec 5, 2023

## Spis treści

<b>1</b>	<b>Opis problemu</b>	<b>2</b>
<b>2</b>	<b>Pozyskiwanie danych</b>	<b>3</b>
2.1	Scrapowanie . . . . .	3
2.1.1	Billboard . . . . .	3
2.1.2	Wikipedia . . . . .	3
2.2	Pre-processing . . . . .	4
2.3	Wstępna analiza danych . . . . .	5
2.3.1	Wiek . . . . .	6
2.3.2	Państwo . . . . .	7
2.3.3	Gatunki muzyczne . . . . .	7
2.3.4	Lata aktywności . . . . .	8
2.3.5	Płeć . . . . .	8
<b>3</b>	<b>Przewidywania</b>	<b>11</b>
3.1	Znaczenie zmiennych . . . . .	11
3.2	Wykresy przybliżające wynik . . . . .	11
3.3	Wnioski . . . . .	12

# 1 Opis problemu

Celem zadania jest zbadanie zależności słuchalności różnych artystów od czynników takich jak: wiek, kraj urodzenia, główny gatunek muzyczny, liczba różnych gatunków muzycznych grana przez muzyka, lata aktywności oraz płeć. Słuchalność danego artysty może być różnie interpretowana, dlatego na potrzeby tego badania jest ona odwrotnie proporcjonalna z miejscem na liście Billboard. Głównym problemem badawczym jest wyłonienie cech idealnych, czyli takich, które teoretycznie zagwarantują miejsce numer jeden na liście. Oczywiście na taki sukces składa się o wiele więcej czynników, często niemierzalnych, dlatego proszę potraktować wyłonienie idealnych cech jako zbadanie opinii publicznej, a nie jako wyznacznik sukcesu. Dane będą pozyskiwane samodzielnie. Listy artystów pobierane są ze strony <https://www.billboard.com/charts/year-end/top-artists/>, a informacje o nich z wikipedii. W celu uruchomienia programu, proszę zainstalować następujące biblioteki:

- billboard
- pandas
- pandas\_profiling
- wikipedia
- beautifulsoup4
- gender-guesser
- requests
- pandas
- sklearn
- matplotlib

## 2 Pozyskiwanie danych

### 2.1 Scrapowanie

#### 2.1.1 Billboard

W celu pobrania danych ze strony <https://www.billboard.com/charts/year-end/top-artists/> używam biblioteki billboard. Wyróżniłem trzy funkcje:

1. `get_top_artists_from_year(year)` - zwraca listę artystów wraz z ich miejscem w rankingu po podaniu roku utworzenia rankingu.
2. `def get_top_artists_from_year_range(start_year, end_year)` - zwraca listę artystów wraz z ich miejscem w rankingu po podaniu zakresu z jakiego chcemy pobrać dane. Używa `get_top_artists_from_year(year)`
3. `filter_artists(start_year, end_year)` - zwraca listę artystów w rankingu po podaniu zakresu z jakiego chcemy pobrać dane bez powtarzających się nazw. Używa `get_top_artists_from_year_range(start_year, end_year)`

#### 2.1.2 Wikipedia

W celu pobrania informacji na temat artystów z wikipedii użyłem bibliotek wikipedia i bs4. Wyróżniłem funkcje:

1. `prepare_url(artist_name)` - zwraca adres url strony wikipedii na podstawie nazwy artysty podanego w argumencie. Większość adresów na wikipedii ma postać `en.wikipedia.org/wiki/artist_name`, jednak u niektórych dopisany jest komentarz w nawiasie, np. `_(band)` dla zespołów, czy `_(producer)` dla producentów. Jest to zamknięty zbiór takich komentarzy, więc są one zgromadzone w zmiennej globalnej i po kolei przeszukiwane dopóki nie znajdziemy strony.
2. `prepare_infobox(url)` - na podstawie adresu url strony wikipedii zwraca słownik z potrzebnymi do przetworzenia informacjami. Wyłania je, tworząc infobox, czyli szablon z podstawowymi informacjami o danym temacie. Używając `bs4.BeautifulSoup` oraz parsera `html` wyróżniliśmy tylko te dane, których będziemy później używać.

3. `extract_info(infobox)` - na podstawie przefiltrowanego infoboxa zwraca słownik z przetworzonymi danymi. Są to jeszcze dane wymagające pre-processingu. Używa poniższych funkcji:
  - (a) `get_age(infobox)` - oblicza wiek artysty na podstawie daty urodzenia
  - (b) `get_country(infobox)` - znajduje kraj pochodzenia danego artysty
  - (c) `get_genres(infobox)` - tworzy listę gatunków muzycznych jakimi zajmuje się artysta oraz zwraca ich ilość
  - (d) `get_years_active(infobox)` - oblicza jak długo artysta jest aktywny w świecie muzycznym
  - (e) `get_gender(infobox)` - zgaduje płeć artysty na podstawie jego oryginalnego imienia

## 2.2 Pre-processing

Uzupełnianie i poprawa danych w tym przypadku polega na ich ujednoliceniu, ponieważ często te same dane zapisane są innymi słowami. Jeśli danych nie da się poprawić lub nie da się ich znaleźć, usuwamy je. Wyróżniłem funkcje:

1. `preprocessing_age_bands(infobox)` - oblicza średni wiek w zespole muzycznym. Jeśli nie ma informacji na temat niektórych członków zespołu, zakłada, że mają oni zbliżony wiek do tych, których wiek znamy. Jeśli nie ma informacji na temat żadnego członka zespołu, zwraca `None`.
2. `preprocessing_country(country)` - zwraca ujednoliconą nazwę państwa, tzn. usuwa niepotrzebne przecinki, nawiasy, liczby przy nazwie wcześniej pobranej. Sprawdza, czy państwo jest napisane skrótem, czy nie i sprowadza nazwę do jednej postaci.
3. `preprocessing_genres(genres)` - wybiera z listy najbardziej popularny gatunek muzyczny. Lista najbardziej popularnych gatunków jest zbiorem skończonym i zdefiniowanym jako globalna zmienna.
4. `preprocessing_gender(sex)` - dla wartości różnych od `None` ujednolica odpowiedź do "male" albo "female"
5. `preprocessing_gender_bands(infobox)` - dla zespołów wybiera płeć zespołu, czyli taką jaką ma większość członków zespołu

6. `preprocessing_dict(dict)` - na podstawie słownika z danymi o artystach zwraca słownik z przetworzonymi danymi. Używa powyższych funkcji.
7. `erase_leftovers(dict)` - usuwa ze słownika z danymi o artystach wszystkie wiersze, w których wystąpiła wartość `None` i zwraca nowy słownik.

## 2.3 Wstępna analiza danych

Wstępną analizę danych wygenerowano z biblioteki `pandas-profiling`. Po zgromadzeniu danych i pre-processingu otrzymaliśmy tabelę z 420. artystami. Dla każdego z nich obliczyliśmy najwyższe miejsce w rankingu, wiek, określiliśmy kraj pochodzenia, najpopularniejszy gatunek muzyczny grany przez wykonawcę, liczbę różnych gatunków muzycznych jakimi posługuje się artysta, obliczyliśmy lata aktywności na scenie i zapisaliśmy płeć. Nasze ostateczne dane zgromadzone są w postaci pliku `artists.csv`:

	Place	Age	Country	Genres	Number of genres	Years active	Gender
2 Chainz	46	46	U.S.	rap	3	26	male
21 Savage	33	31	U.S.	rap	3	10	female
24kGoldn	45	23	U.S.	pop	4	7	male
3OH!3	45	38	U.S.	rock	5	19	male
5 Seconds Of Summer	22	27	Australia	rock	3	12	male
50 Cent	47	48	U.S.	hip hop	1	27	male
6ix9ine	28	27	U.S.	rap	3	11	male
A Boogie Wit da Hoodie	24	28	U.S.	rap	4	8	male
A Great Big World	85	38	U.S.	rock	2	12	male
A\$AP Rocky	76	35	U.S.	rap	3	16	male
AC/DC	74	70	Australia	rock	3	50	male
AWOLNATION	64	44	U.S.	rock	6	14	male
Adam Lambert	27	41	U.S.	rock	4	22	male
Adele	1	35	UK	pop	2	17	female
Akon	1	50	U.S.	pop	3	27	male
Alessia Cara	27	27	Canada	pop	2	9	female
Alicia Keys	4	42	U.S.	pop	2	30	female
Aloe Blacc	77	44	U.S.	pop	5	28	male
American Authors	63	33	U.S.	rock	5	17	male
Amy Winehouse	41	28	England	r&b	4	9	female

Figure 1: Pierwsze pięć wierszy pliku `artists.csv` z danymi

Macierz korelacji pokazuje nam zależność zmiennych od siebie nawzajem. Lata aktywności oraz wiek są od siebie wzajemnie mocno zależne, co jest dość oczywiste, gdyż większość artystów zaczyna swoją karierę w wieku około 20 lat. Także niezaskakującą zależnością jest państwo od gatunków i na odwrót. Każdy kraj ma dominujący gatunek muzyczny.

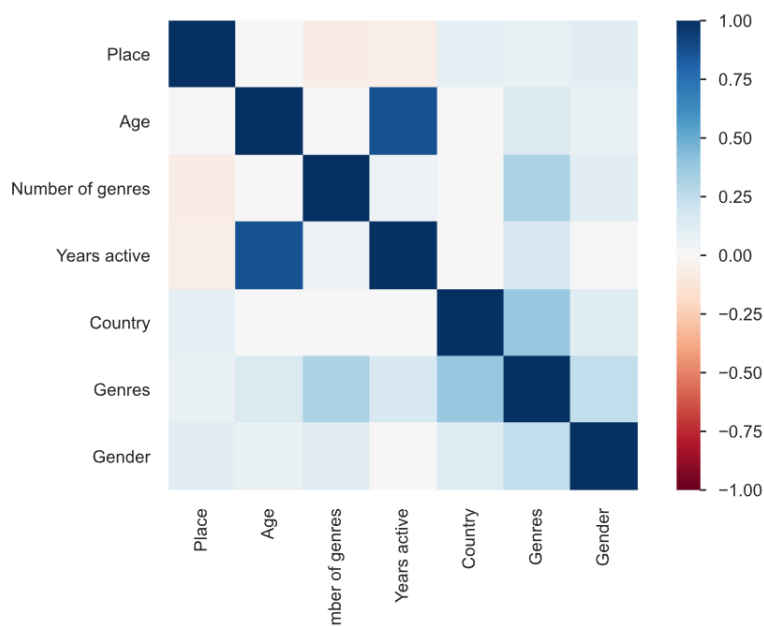


Figure 2: Macierz korelacji

### 2.3.1 Wiek

Najczęściej występującymi artystami w rankingu są wykonawcy w wieku 30-34, jednak tych oscylujących około 40. także nie brakuje. Gdy spojrzymy na zależność miejsca w rankingu od wieku widać, że dane są bardzo porozrzucane. Łatwo jednak zauważyć, że najczęściej pierwsze miejsce w rankingu mają artyści w przedziale 28-50

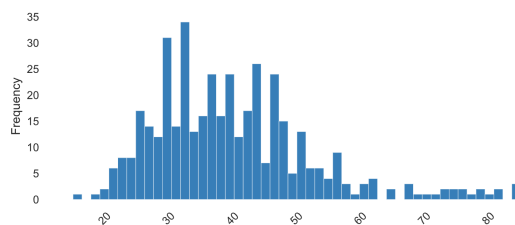


Figure 3: Częstotliwość wieku

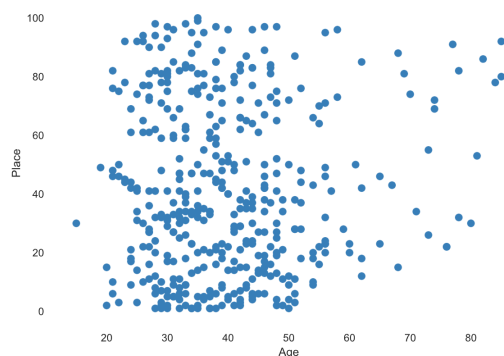


Figure 4: Zależność miejsca w rankingu od wieku

### 2.3.2 Państwo

Najczęściej występującymi artystami w rankingu są wykonawcy anglojęzyczny, którzy 4 pierwsze miejsca, jeżeli chodzi o częstotliwość.

Value	Count	Frequency (%)
U.S.	313	74.7%
England	29	6.9%
Canada	13	3.1%
Australia	11	2.6%
Puerto Rico	5	1.2%
Scotland	5	1.2%
Ireland	4	1.0%
South Korea	3	0.7%
Jamaica	3	0.7%
Sweden	3	0.7%
Other values (22)	30	7.2%

Figure 5: Częstotliwość państw

### 2.3.3 Gatunki muzyczne

Najczęściej występującymi artystami w rankingu są wykonawcy rockowi lub popowi. Jeśli chodzi o różnorodność artysty, to zdecydowanie widać, że najczęściej w rankingu pojawiają się wykonawcy grający muzykę łączącą 3 gatunki. Na wykresie zależności miejsca w rankingu od różnorodności łatwo zauważyć, że tacy artyści też częściej mają pierwsze miejsce na liście.

Value	Count	Frequency (%)
rock	136	32.5%
pop	134	32.0%
rap	57	13.6%
hip hop	55	13.1%
country	17	4.1%
r&b	15	3.6%
electro	3	0.7%
reggae	2	0.5%

Figure 6: Częstość gatunków muzycznych

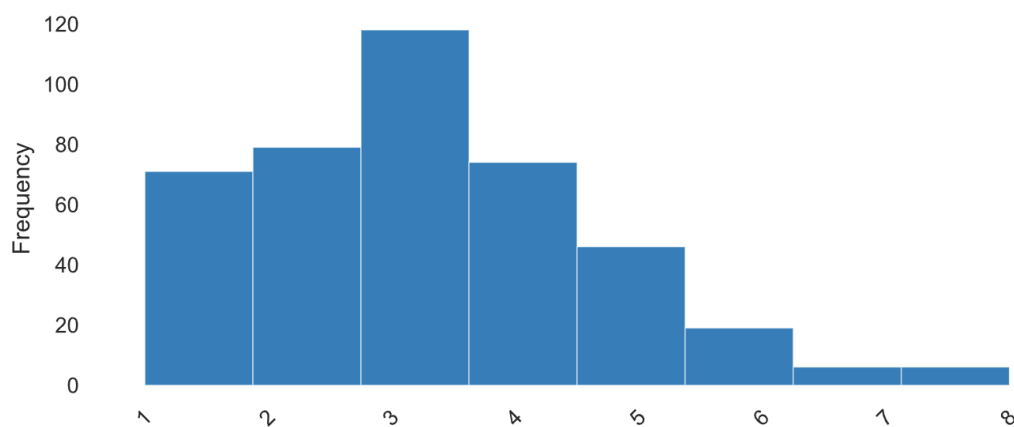


Figure 7: Częstość ilości gatunków muzycznych

#### 2.3.4 Lata aktywności

Lata aktywności są ściśle powiązane z wiekiem, więc wykresy będą przybliżone do tych wcześniej.

#### 2.3.5 Płeć

Rynek muzyczny zdominowali mężczyźni.



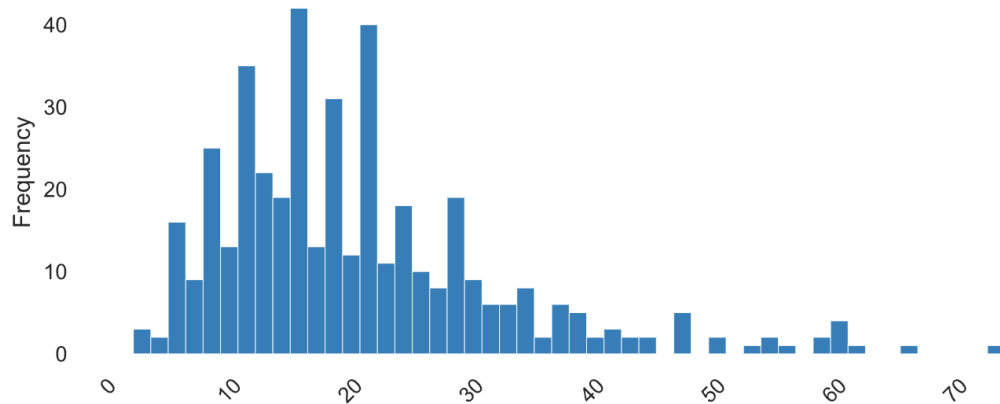


Figure 8: Częstość lat aktywności

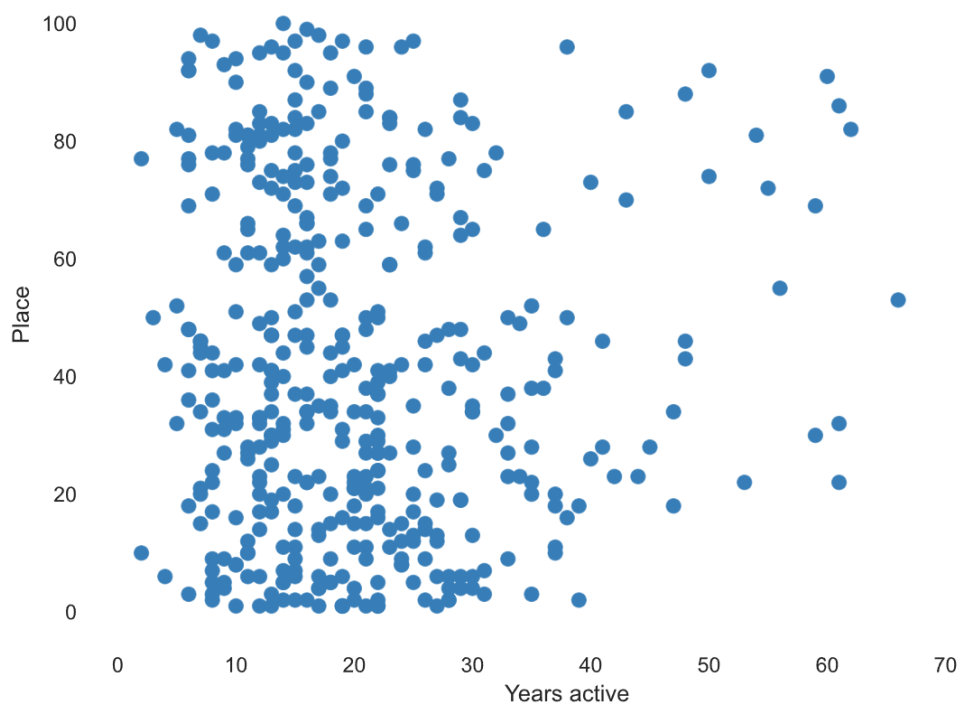


Figure 9: Zależność miejsca w rankingu od lat aktywności

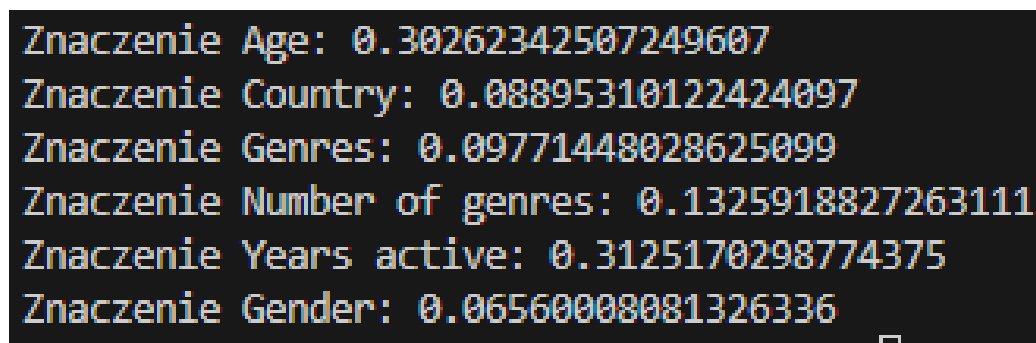
Value	Count	Frequency (%)
male	337	80.4%
female	82	19.6%

Figure 10: Częstość płci

## 3 Przewidywania

### 3.1 Znaczenie zmiennych

Zacznijmy od przebadania, która cecha najbardziej wpływa na wynik. W badaniu wpływu zmiennych na miejsce w rankingu użyłem modelu Decision-TreeClassifier, ponieważ mamy sześć zmiennych i chcemy zbadać, która z nich ma największy wpływ na wynik. Największe znaczenie mają wiek oraz lata aktywności. Potem różnorodność artysty. Najmniejszy wpływ na miejsce w rankingu mają kraj, gatunki muzyczne oraz płeć. Gdy weźmiemy model, w którym rozdzieliliśmy zmienne na cechy kategoryczne otrzymamy bardziej szczegółowe dane. Z państw pochodzenia najbardziej znaczące są USA oraz Anglia. Jeśli chodzi o gatunki muzyczne, to znowu mamy do czynienia z popem oraz rockiem. Płeć wydaje się dość zaskakująca, ponieważ mimo znacznej przewagi w ilości mężczyzn w rankingach, to artystki częściej sięgają po wyższe miejsca.



```
Znaczenie Age: 0.30262342507249607
Znaczenie Country: 0.08895310122424097
Znaczenie Genres: 0.09771448028625099
Znaczenie Number of genres: 0.1325918827263111
Znaczenie Years active: 0.3125170298774375
Znaczenie Gender: 0.06560008081326336
```

Figure 11: Znaczenie zmiennych

### 3.2 Wykresy przybliżające wynik

Tworząc przybliżenia użyłem biblioteki sklearn. Biorąc pod uwagę ilość i różnorodność danych, zdecydowałem się na wielomian 10-ego stopnia, gdyż wszystkie poniżej niewystarczająco przybliżały wyniki. Użyłem modelu LinearRegression oraz w celu podziału cech kategorycznych LabelEncoder.

### 3.3 Wnioski

Dane na jakich operowałem okazały się mało zróżnicowane. Bardzo łatwo wyróżnić cechy artysty, które najczęściej gwarantują czołówkę listy. Ciężko przedstawić przybliżenie wielomianowe w wykresie 7D, dlatego wykresy na pojedynczych cechach przysunęły następujące wnioski:

- Największy wpływ na miejsce w rankingu mają wiek oraz lata aktywności artysty, które są ze sobą ściśle powiązane.
- Wiek, który przybliży artystę do pierwszego miejsca najbardziej to około 50 lat.
- Lata aktywności, które przybliżają artystę do pierwszego miejsca najbardziej to około 32 lata
- Kraj pochodzenia artysty nie wpływa znacząco na miejsce na liście. Wynika to z faktu, że przemysł muzyczny zdominowany jest przez wykonawców ze Stanów Zjednoczonych, którzy mają miejsca w czołówce, jak i w ogonie listy.
- Kraj pochodzenia, który najbardziej zbliży artystę do pierwszego miejsca to Stany Zjednoczone. Na drugim miejscu Anglia.
- Gatunek muzyczny także nie wpływa znacząco na miejsce na liście. Wynika to z faktu, że wykonawcy sięgający po najbardziej popularne gatunki mają miejsca w rankingu w czołówce, jak i w ogonie.
- Gatunek muzyczny najbardziej zbliżający artystę do pierwszego miejsca to rock.
- Różnorodność artysty wpływa na miejsce w rankingu bardzo mocno. Jest to trzecia cecha, zaraz po wieku i latach aktywności, mająca duże znaczenie.
- Liczba różnych gatunków muzycznych, po jakie sięga artysta w swojej dyskografii przybliżająca go do pierwszego miejsca na liście to 7 lub 8. Wynika to z faktu, że niewielu wykonawców miesza ze tak dużo gatunków. Stąd posiadamy mało danych na ich temat, a przeważnie zajmują wysokie miejsce na liście.

- Płeć wpływa na miejsce artysty w rankingu nieznacząco.
- Artystów płci męskiej jest o wiele więcej, jednak to kobiety zajmują wyższe miejsca na liście.

```
Age : 0.26611833713573996
Number of genres : 0.14108693233870545
Years active : 0.2578986402012058
Country_Australia : 0.006244873611214676
Country_Barbados : 0.0
Country_British Virgin Islands : 0.004163249074143117
Country_California : 0.0
Country_Canada : 0.0032971185849420166
Country_Colombia : 0.0
Country_Cuba : 0.0
Country_D.C. : 0.0
Country_Denmark : 0.0
Country_Englan : 0.0
Country_England : 0.01464406341000183
Country_France : 0.003122436805607338
Country_Germany : 0.0032161099097755617
Country_Guyana : 0.0
Country_Iceland : 0.0031651931781410585
Country_Ireland : 0.006146180480682033
Country_Jamaica : 0.007399229036317996
Country_Japan : 0.0032089042863779962
Country_Louisiana : 0.0
Country_Netherlands : 0.0
Country_New Zealand : 0.001040812268535779
Country_Nigeria : 0.0
Country_Puerto Rico : 0.003122436805607338
Country_Scotland : 0.003122436805607338
Country_South Korea : 0.0
Country_St Kevin's CollegeMonash University : 0.0
Country_Sweden : 0.0
Country_Tennessee : 0.0
Country_Texas : 0.003122436805607338
Country_U.S. : 0.03962870793598717
Country_UK : 0.0
Country_United Kingdom : 0.0
Genres_country : 0.0064183423226373
Genres_electro : 0.003122436805607338
Genres_hip hop : 0.016087983922224477
Genres_pop : 0.04812889398420866
Genres_r&b : 0.018106684191195096
Genres_rap : 0.03782811914429617
Genres_reggae : 0.0
Genres_rock : 0.036175660704965015
Gender_female : 0.044588667924921756
Gender_male : 0.01979511232574456
```

Figure 12: Znaczenie zmiennych kategorycznych

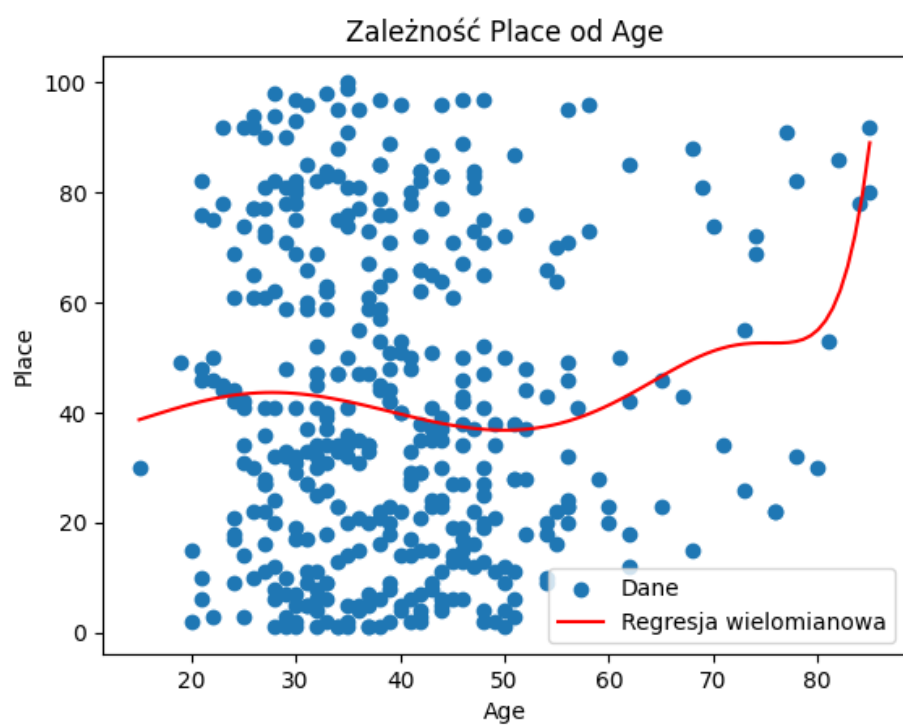


Figure 13: Przybliżenie wielomianem zależności miejsca w rankingu od wieku

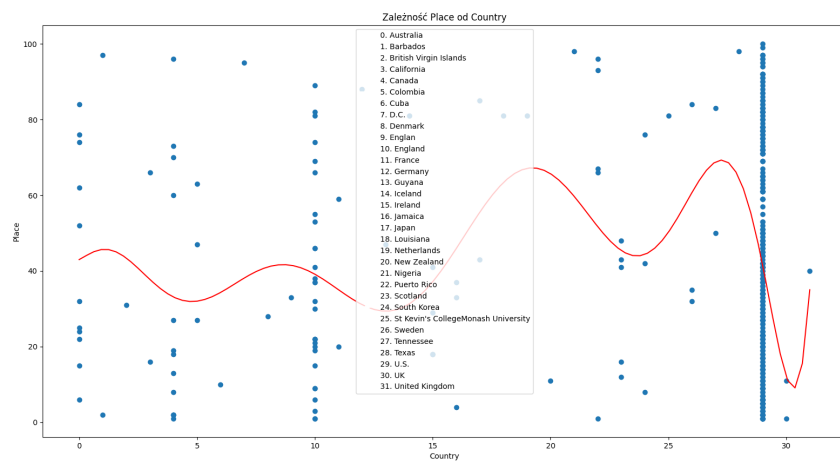


Figure 14: Przybliżenie wielomianem zależności miejsca w rankingu od kraju pochodzenia



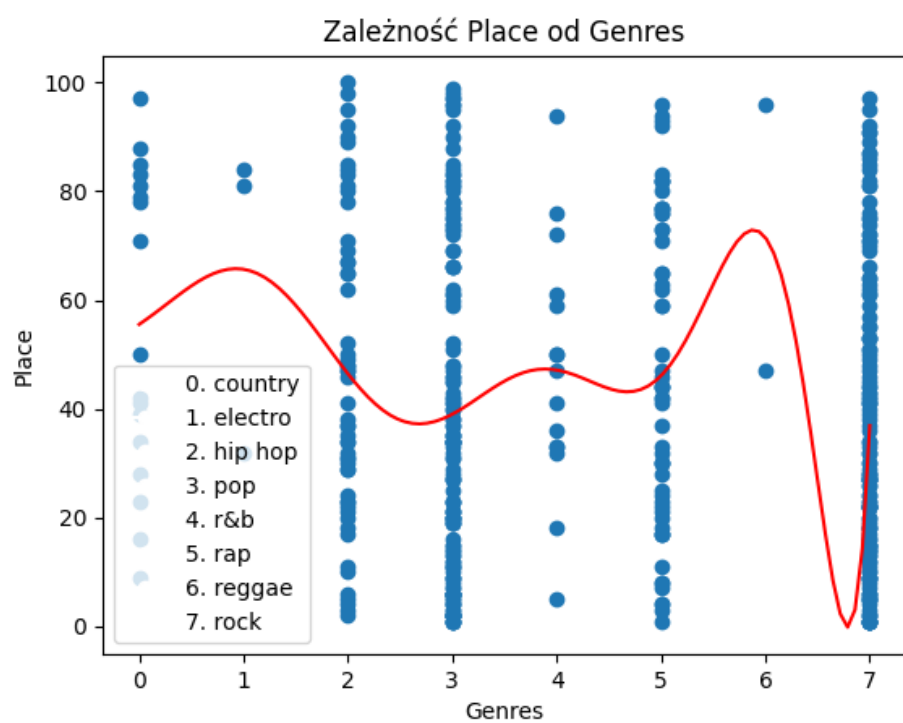


Figure 15: Przybliżenie wielomianem zależności miejsca w rankingu od gatunków muzycznych

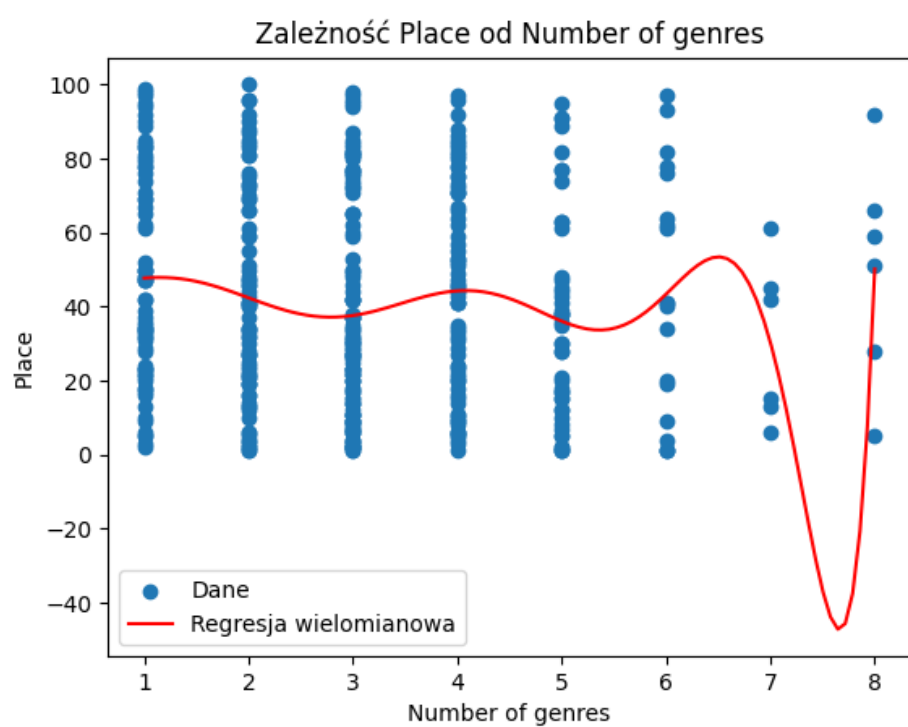


Figure 16: Przybliżenie wielomianem zależności miejsca w rankingu od różnorodności artysty

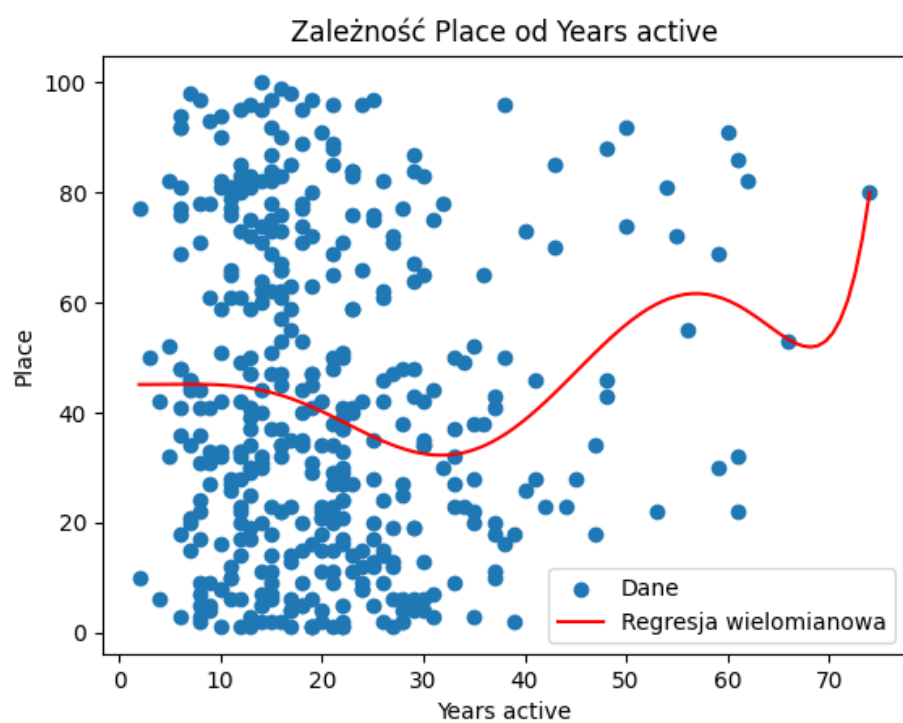


Figure 17: Przybliżenie wielomianem zależności miejsca w rankingu od lat aktywności