DZIENNIK Z REALIZACJI PROJEKTU GRUPOWEGO
TOP LISTY SPOTIFY

Grupa wykonywująca:
Jakub Goleń
Michał Król
Amelia Jonarska

Inżynieria i Analiza Danych 2021/2022

Temat naszego projektu, który wykonaliśmy posługując się językiem programowania R, opiera się na analizie danych pobranych ze szwedzkiego serwisu strumieniowego Spotify, który oferuje dostęp do muzyki oraz podcastów. Zawiera ponad sześćdziesiąt milionów utworów pochodzących od wytwórni płytowych i firm medialnych. My natomiast swoją uwagę skupiliśmy na analizie danych 51 wybranych państw.

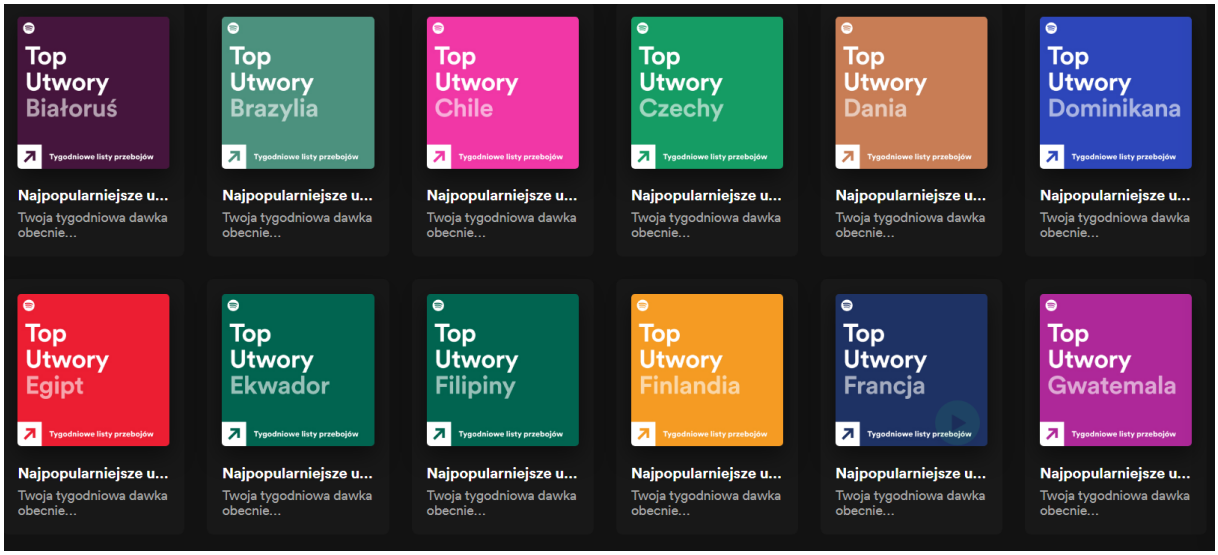
19 maja 2022, 20 maja 2022, 21 maja 2022

Swoją pracę nad projektem zaczęliśmy od poszukiwań gotowych zbiorów danych. Te które znaleźliśmy nie satysfakcjonowały nas, więc zebraliśmy dane za pomocą pakietu Spotifyr – Spotify API. API to metoda komunikacji między aplikacjami sieciowymi oraz wymiany danych między oddzielnymi systemami. API umożliwia rozszerzenie funkcjonalności aplikacji sieciowych poprzez gromadzenie danych ze źródeł zewnętrznych. Spotify Web API w R opiera się na funkcji search_spotify, która jest zapytaniem do serwera. Do funkcji podaje się poszukiwaną treść np. search_spotify('radiohead', 'artist').

```
search_spotify(
  q,
  type = c("album", "artist", "playlist", "track"),
  market = NULL,
  limit = 20,
  offset = 0,
  include_external = NULL,
  authorization = get_spotify_access_token(),
  include_meta_info = FALSE
)
```

q – zapytanie (działa jak wyszukiwanie na aplikacji Spotify), type – wektor typów wyszukiwania, zwraca albumy oraz utwory, limit – maksymalna liczba wyników do zwrócenia, authorization - ważny i wymagany token dostępu z usługi konta Spotify. Uzyskuje się go na stronie spotifyfordevelopers.com.

23 maja 2022, 24 maja 2022, 4 czerwca 2022, 5 czerwca 2022



Screen pochodzący z aplikacji Spotify

Kolejnym etapem w tworzeniu projektu było napisanie głównej funkcji, na której cały czas bazowaliśmy - get_playlist_top_50_spotify. Jest to funkcja na podstawie wektora, który zawiera nazwy krajów. Zbiera piosenki, które znajdują się na playlistach top 50 poszczególnych krajów, które wybraliśmy. Lista zawiera pięćdziesiąt jeden pozycji, głównie znajdują się tam kraje Europy. Funkcja zwraca listę ramek danych. Zebrała czterdzieści tysięcy danych w mniej więcej pół godziny.

Ramka danych zawiera w sumie około trzydzieści osiem tysięcy danych, z informacjami takimi jak nazwa artysty, nazwa gatunku, nazwa piosenki, id artysty, id piosenki, danceability, energy, loudness, speechIness, aucosticness, liveness, tempo, valance. Funkcja ta wstępnie wyszukuje playlisty w pętli. Następnie na podstawie playlist wyszukuje artystów. Następnie na bazie artystów wyszukuje ich dane takie jak id artysty, gatunek. W tym momencie tworzenia projektu napotkaliśmy pierwszy problem. Niektóre gatunki wypisywały się jako lista o długości zero. Na podstawie wielu warunków, podmienialiśmy puste listy na gatunek „other”.

Po uzyskaniu danych natrafił się kolejny problem – różnorodność gatunkowa. Kolumna zawierająca dane na temat gatunku danych artystów przedstawiała pozycje takie jak polish hip-hop, france hip-hop, które znaczyły ten sam gatunek, czyli w tym przypadku hip-hop. Pozycje zawierały niepotrzebną informację na temat kraju. Aby zapobiec temu problemowi, zunifikowaliśmy do jednych gatunków, calem

zwiększenia prawdopodobieństwa trafności algorytmu do klasyfikacji, a później też chartowo, aby uniknąć tak zwanego ciasta pokrojonego na milion kawałków.

Kolejny problem stanowił fakt, że jedna kolumna zawierała tytuł piosenki i artystę wykonującego ją. Z tym poradziliśmy sobie za pomocą regex.

```
list_of_song_artist <- substring(unlist(stri_extract_all_regex(
  splited[[i]][2],"(by {1}.*)")), 4)
artist <- append(artist, list_of_song_artist)
```

Próbowaliśmy również zebrać dane za pomocą webscrapingu - funkcja get_spotify_charts_data. Jej zadaniem było odczytanie ze strony spotifycharts.com danych dotyczących najbardziej popularnych piosenek w danym dniu.

```
get_spotify_charts_data <- function(
    region = c("global","pl","de","ua","us","sk"),
    timestamp = c("daily", "weekly"),
    date = "latest"
)
```

Pojawił się kolejny problem - trzy funkcje z czterech utworzonych nie działają, ponieważ design strony został zmieniony w czasie tworzenia naszego projektu. Funkcje, które nie działają: get_spotify_charts_data, artist_data, regional_artist_data.

6 czerwca 2022, 7 czerwca 2022

Tego dnia skupiliśmy się na wizualizacji danych, które wcześniej odpowiednio przygotowaliśmy. Za pomocą biblioteki ggplot2 stworzyliśmy wykresy pudełkowe, które przedstawiają charakterystyczne dla piosenek cechy.

Danceability - opisuje w jakim stopniu utwór nadaje się do tańca na podstawie kombinacji elementów muzycznych, takich jak tempo, stabilność rytmu, siła uderzenia i ogólna regularność. Wartość 0,0 oznacza najmniej taneczny utwór, a 1,0 najbardziej taneczny.

Energy - jest miarą od 0,0 do 1,0 i stanowi percepcyjną miarę intensywności i aktywności. Zazwyczaj utwory energetyczne są szybkie, głośne i hałaśliwe.

Loudness – wykres dzięki któremu możemy się dowiedzieć, w którym kraju słucha się najgłośniejszych piosenek.

Speechiness – wykrywa obecność słów mówionych w utworze. Im bardziej nagranie przypomina wyłącznie mowę (np. talk show, książka audio, poezja), tym bardziej wartość atrybutu zbliża się do 1,0. Wartości powyżej 0,66 opisują utwory, które prawdopodobnie w całości składają się ze słów mówionych. Wartości pomiędzy 0,33 a 0,66 opisują utwory, które mogą zawierać zarówno muzykę, jak i mowę, w sekcjach lub warstwach, w tym takie przypadki jak muzyka rap. Wartości poniżej 0,33 najprawdopodobniej reprezentują muzykę i inne utwory nie zawierające mowy.

Liveness - wykrywa obecność publiczności w nagraniu. Wyższe wartości współczynnika (powyżej 0,8) oznaczają zwiększone prawdopodobieństwo, że utwór został wykonany na żywo.

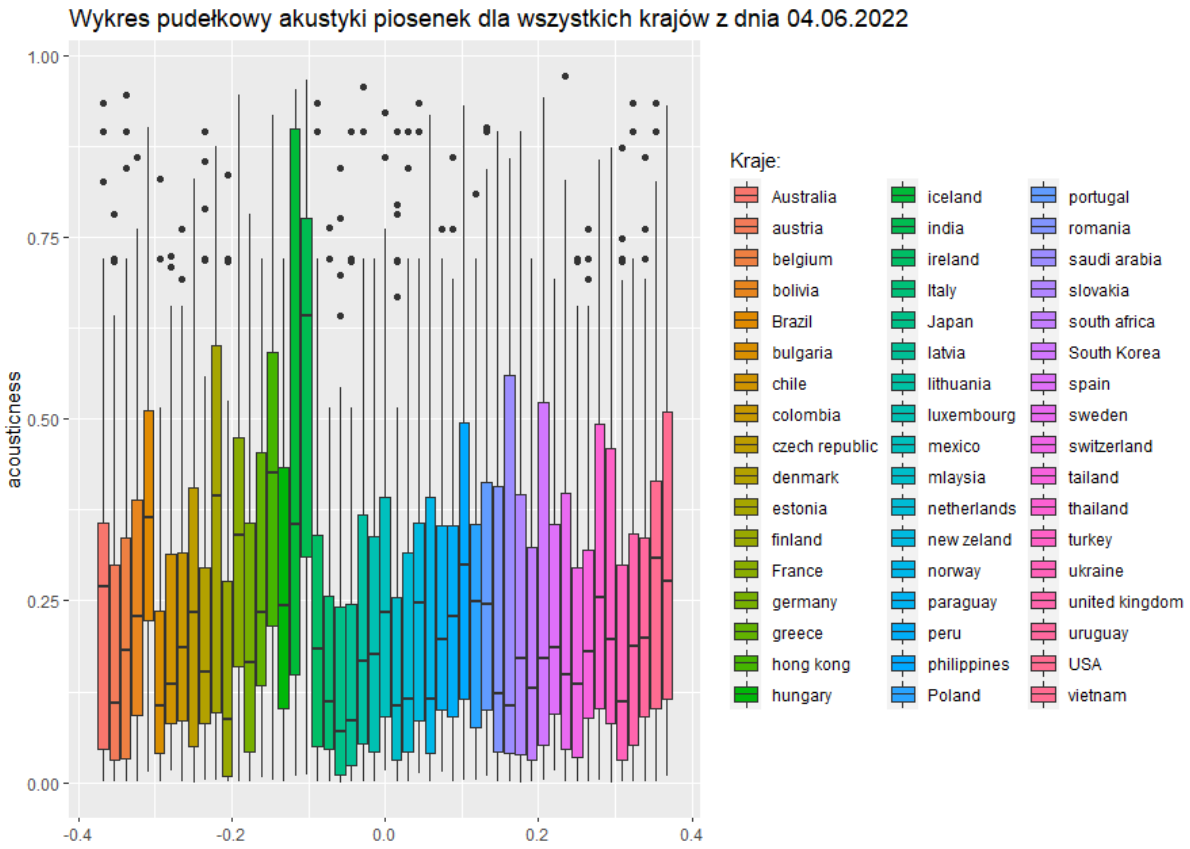
Acousticness - miara pewności w zakresie od 0,0 do 1,0, określająca, czy utwór jest akustyczny. 1,0 oznacza wysoką pewność, że utwór jest akustyczny.

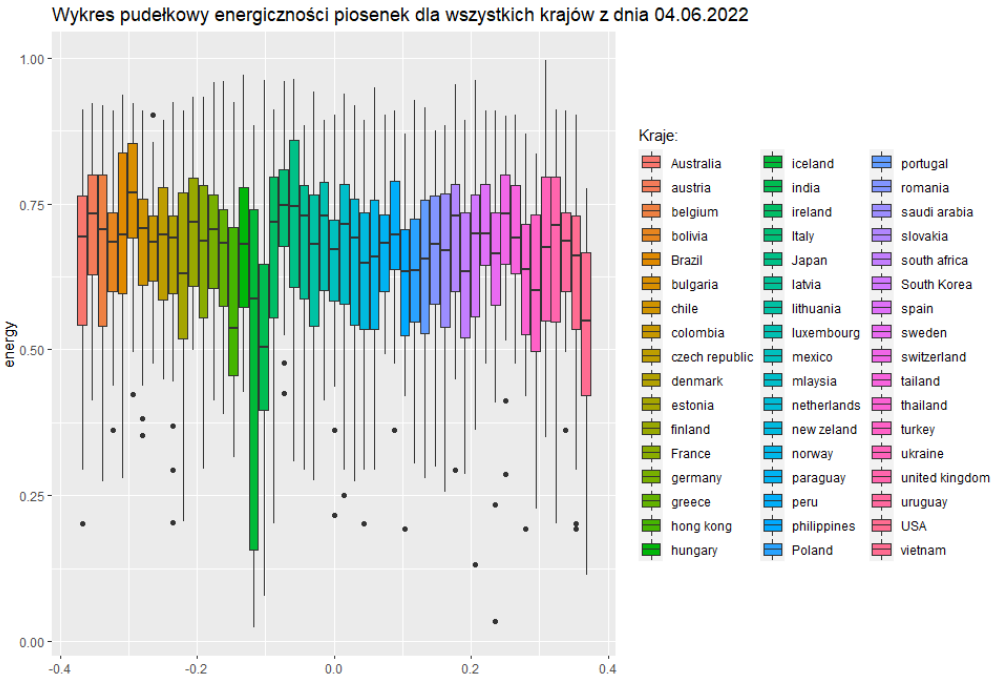
Tempo - ogólne szacunkowe tempo utworu w bitach na minutę. W terminologii muzycznej tempo jest szybkością lub tempem danego utworu i wynika bezpośrednio ze średniego czasu trwania taktu.

Valence - miara od 0,0 do 1,0 określająca pozytywność muzyczną przekazywaną przez utwór. Utwory o tym wysokim współczynniku brzmią bardziej pozytywnie (np. radośnie, wesoło, euforycznie), natomiast utwory o niskim współczynniku valence brzmią bardziej negatywnie (np. smutno, przygnębiająco, gniewnie).

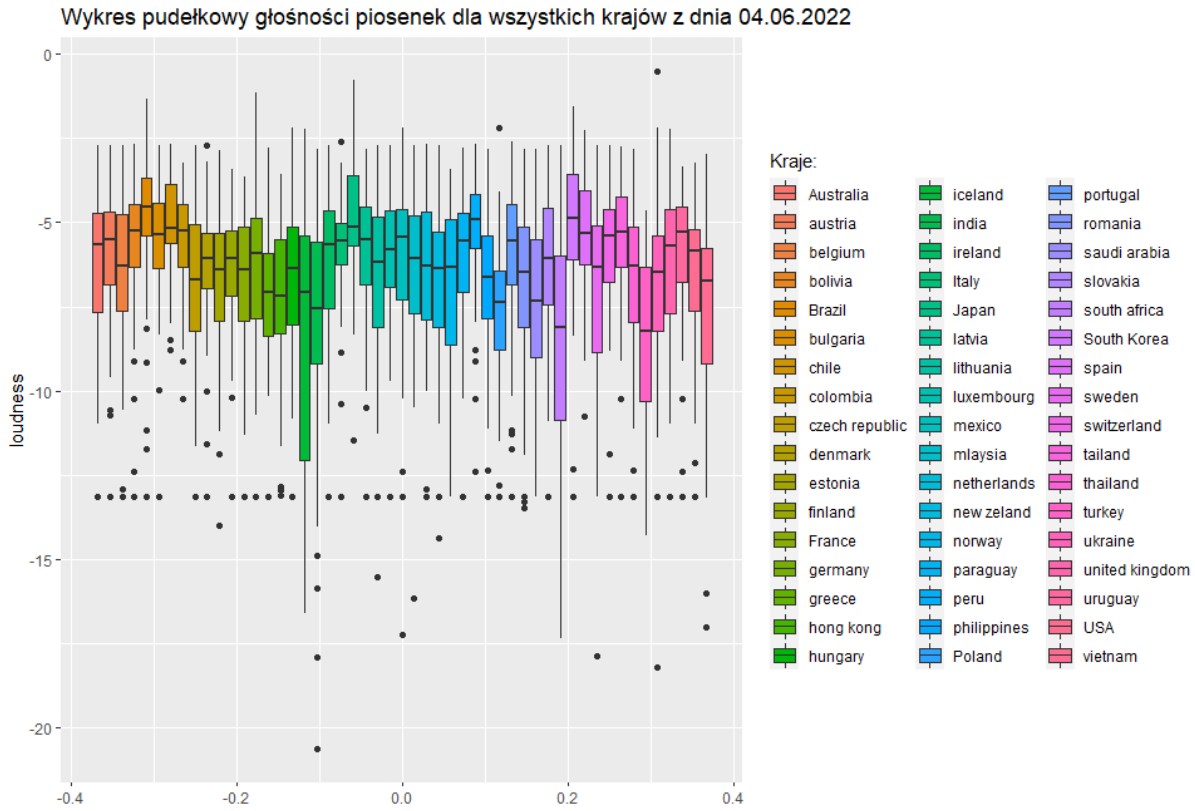
Powyższe dane przedstawiliśmy za pomocą wykresów pudełkowych:

Można zauważyć, że muzyka akustyczna góruje wśród słuchaczy Islandii oraz Indii, tak samo jak na poniższym wykresie, który uwzględnia te państwa jako największych fanów muzyki spokojnej.

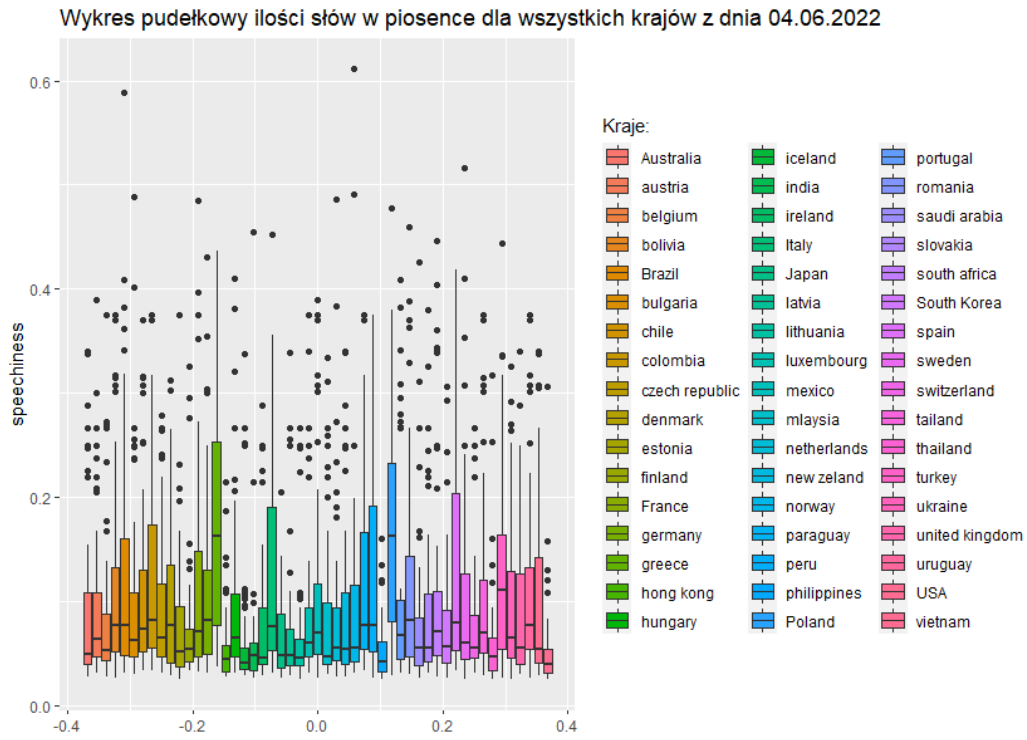




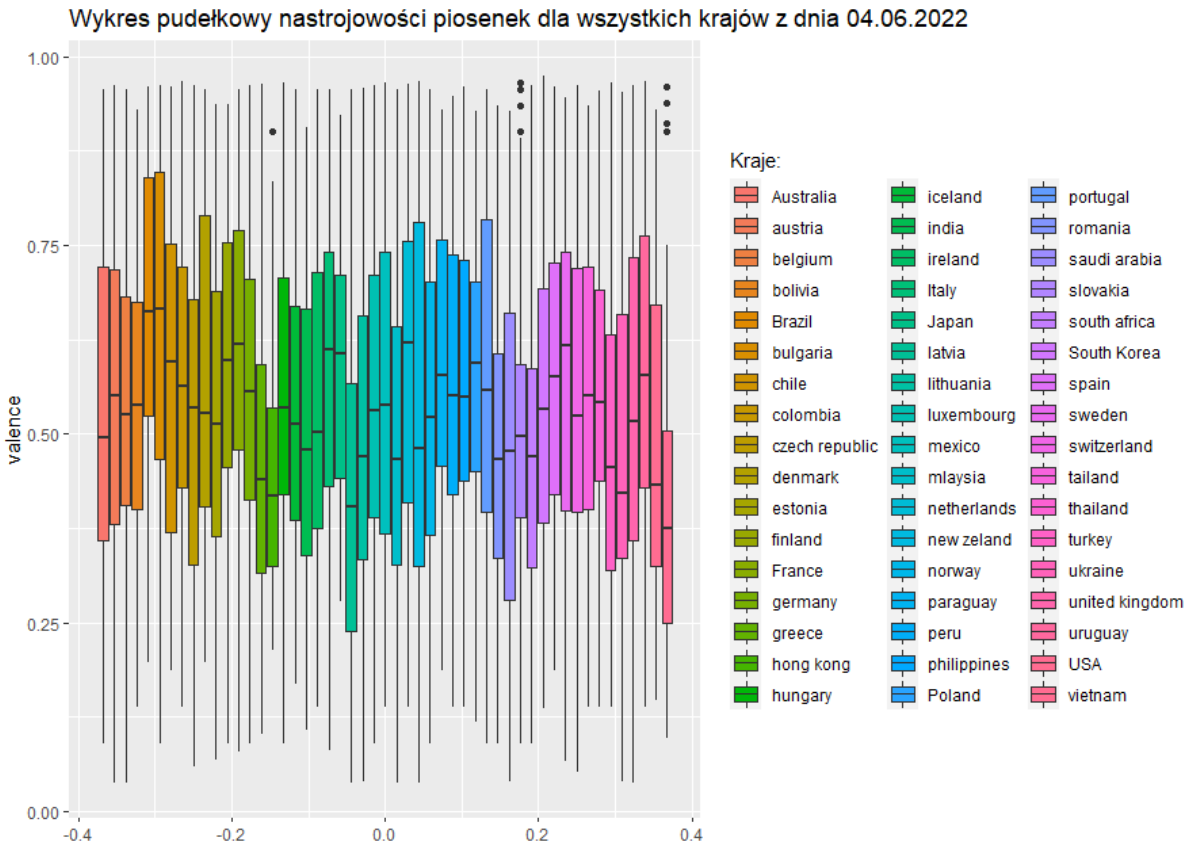
Najcichsze brzmienia występują w słuchawkach obywateli Islandii, najgłośniejsze natomiast w słuchawkach Japończyków.



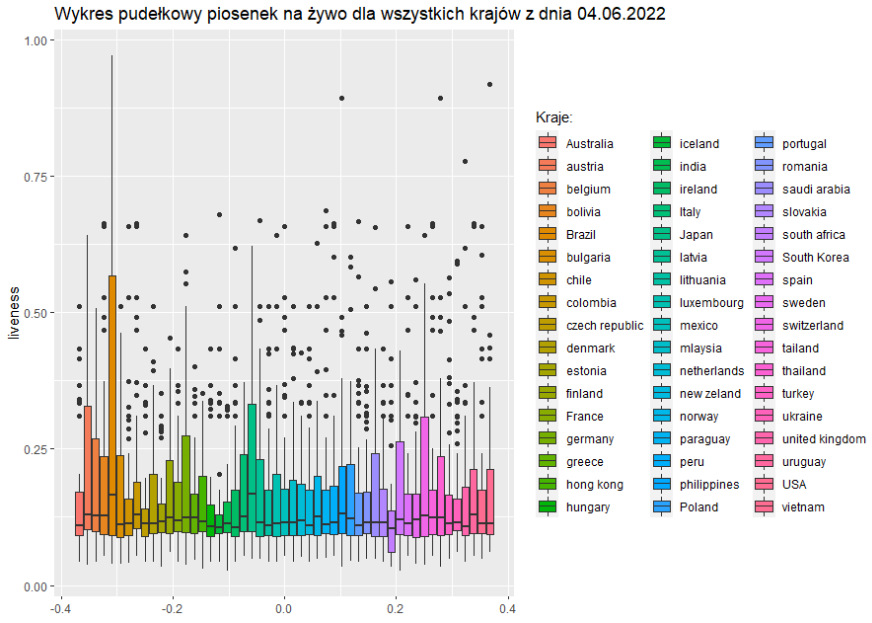
Piosenki z obecnością słów mówionych w utworze w największych ilościach występują w Grecji.



Najbardziej nastrojowe piosenki górują w playliście Top 50 w Bułgarii oraz Brazylii. Najmniej nastrojowe natomiast na Litwie.

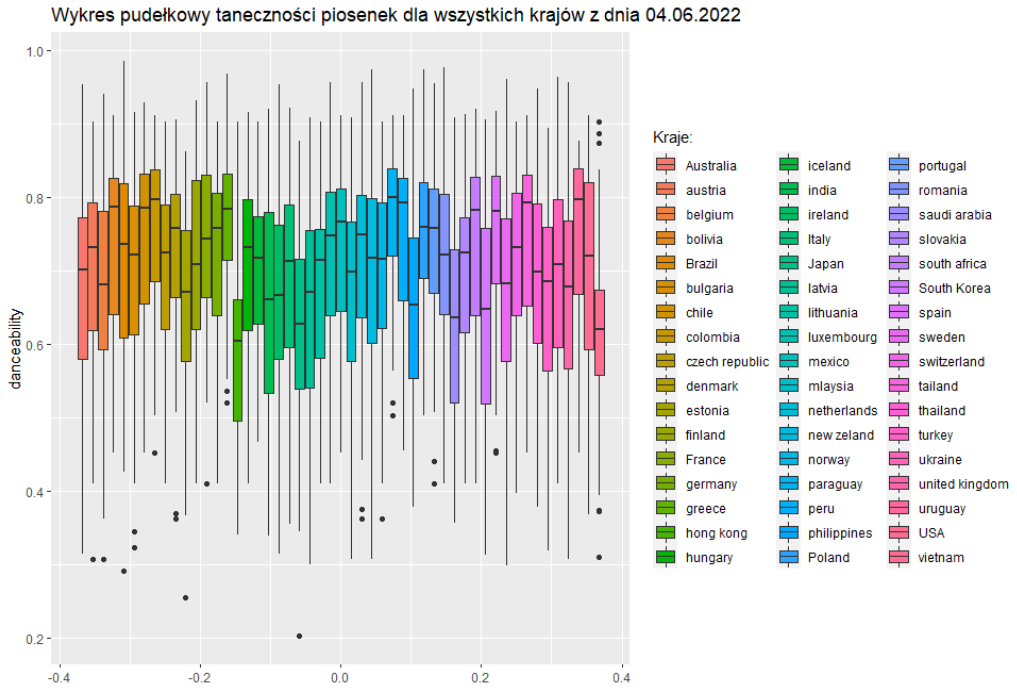


Piosenki ze zwiększonym prawdopodobieństwem nagrania na żywo, górują na play liście w Brazylii.

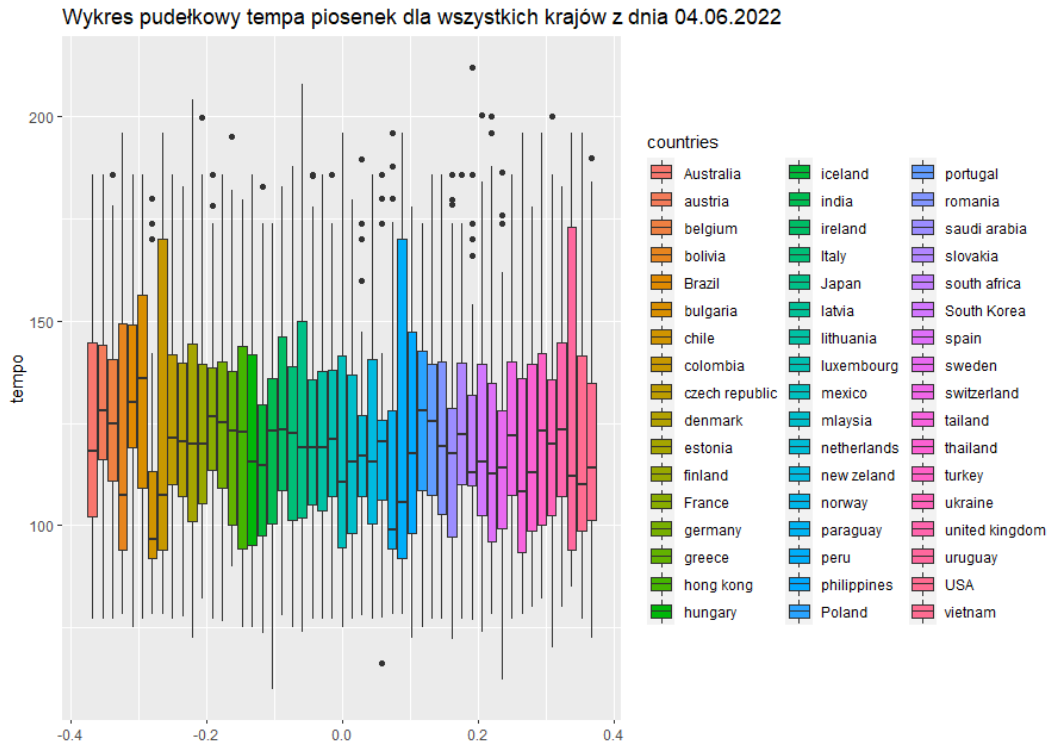


Jak można zauważyć, wykres taneczności piosenek pokazuje, że praktycznie we wszystkich krajach ten współczynnik waha się wokoło podobnych wartości. Należy

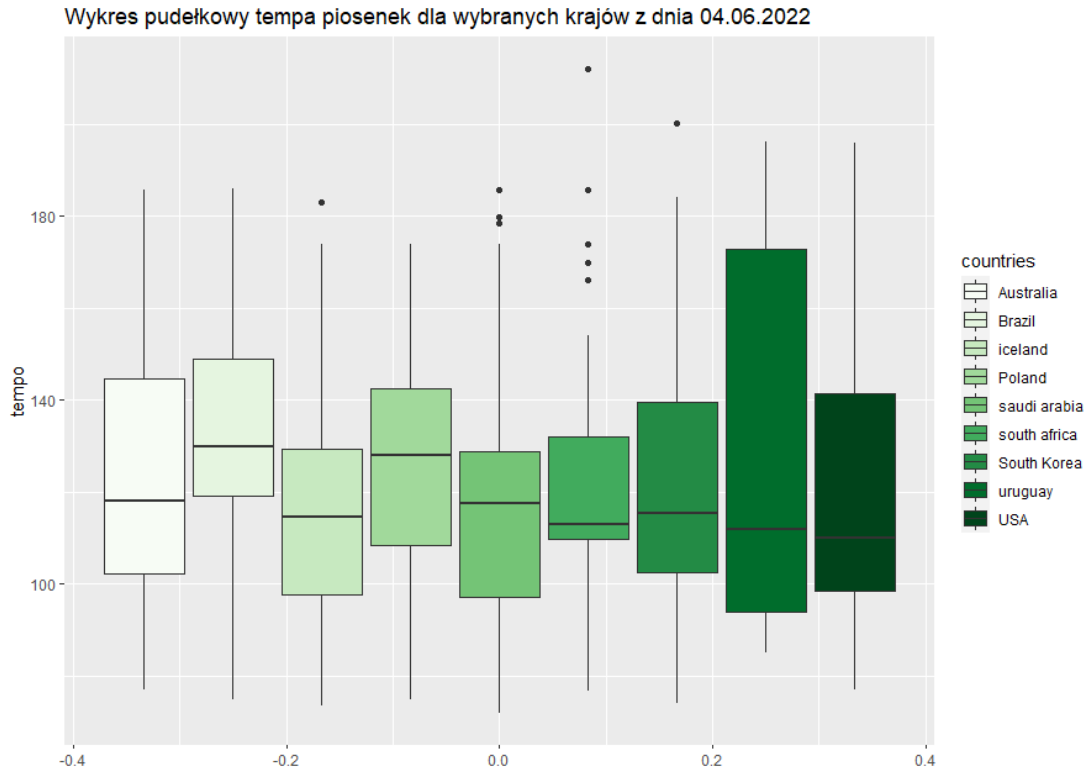
zaznaczyć jedynie, że znów Islandia obraca się wśród mniej ruchliwych i energicznych piosenek, bardziej spokojnych oraz jak poniższy wykres wskazuje mniej tanecznych.

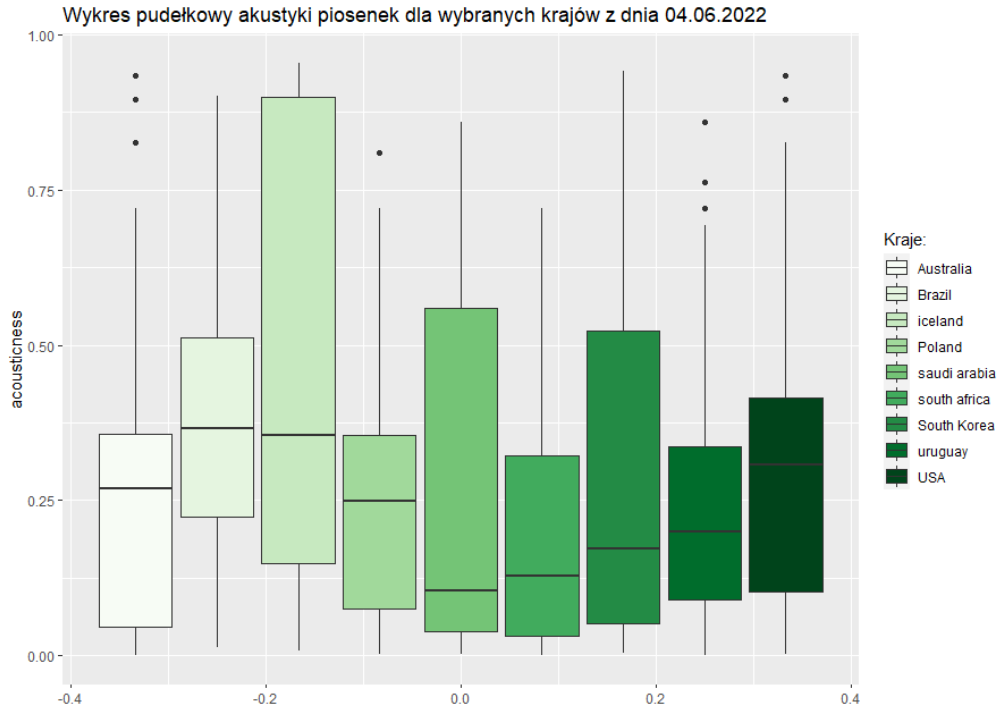


Piosenki ze szybkim tempem górują na playliście w Kolumbii oraz Urugwaju.

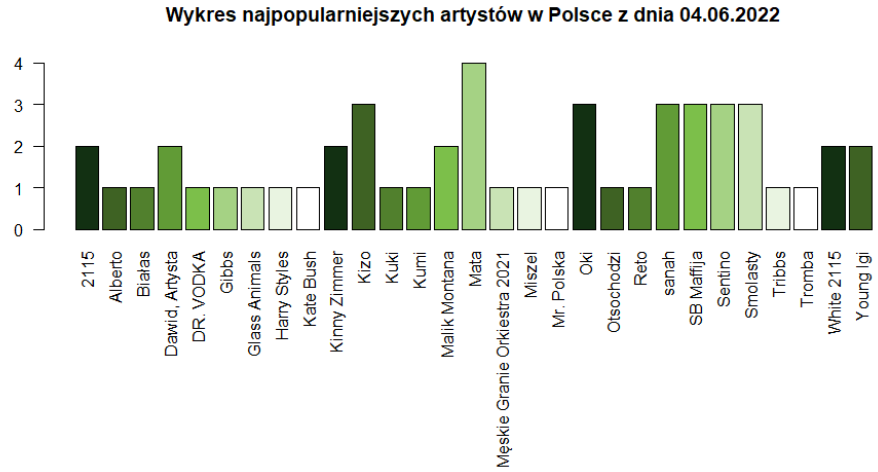


Poniższe przedstawienie danych na wykresie uczyniliśmy dla wybranych przez siebie krajów z różnych kontynentów, w celu dostrzeżenia różnicy kulturowej. Przykładowe wykresy pudełkowe:

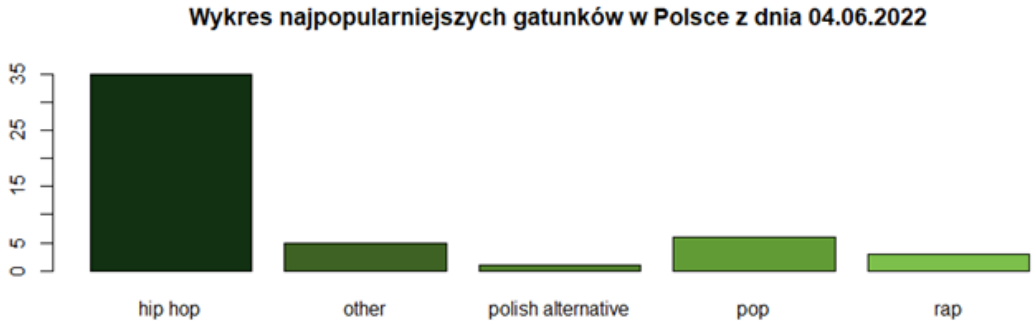


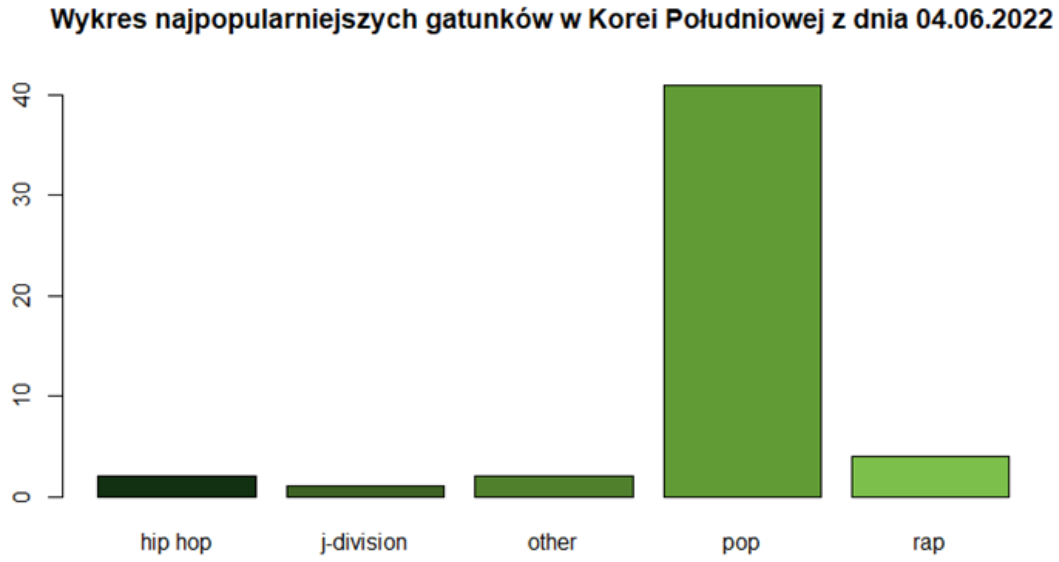


Dokonaliśmy wizualizacji najpopularniejszych artystów w Polsce:



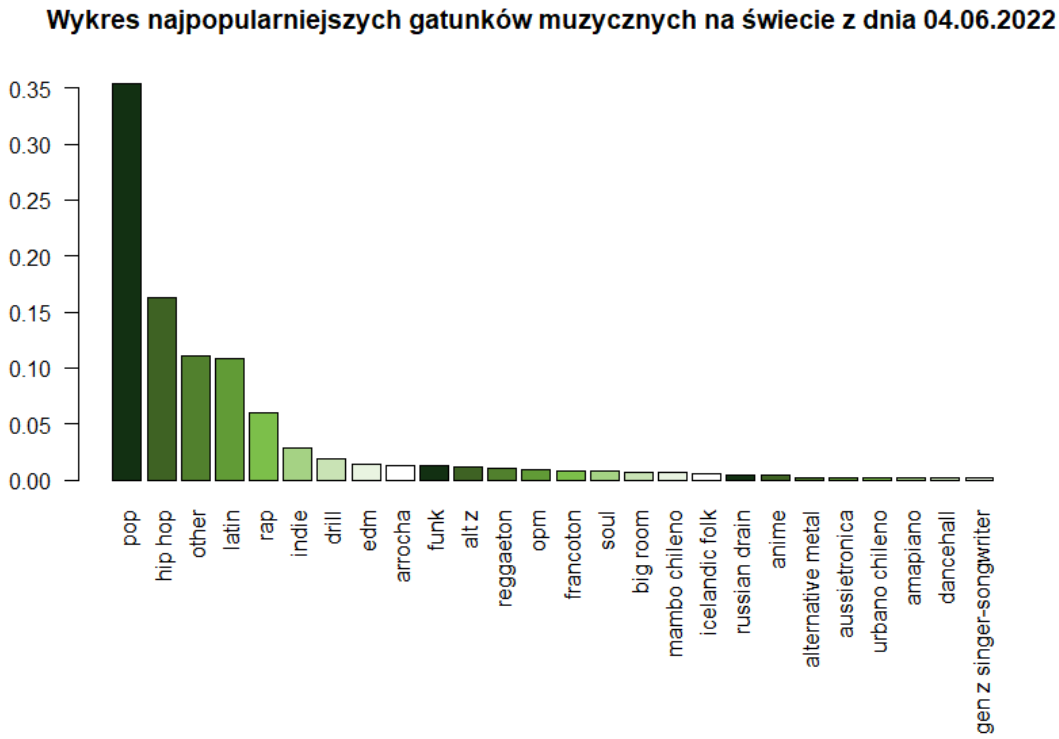
Oraz najpopularniejszych gatunków w Polsce i Korei Południowej:





W Polsce z ogromną przewagą góruje hip hop, natomiast w Korei Południowej pop.

Również sprawdziliśmy, jakie gatunki górują wśród upodobań słuchaczy biorąc pod uwagę wszystkie kraje:



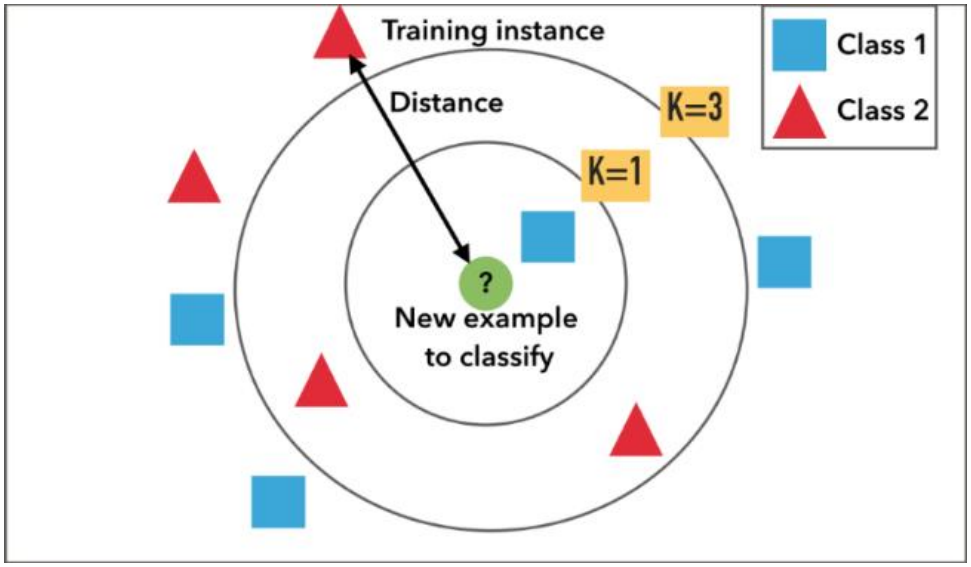
Wniosek jaki nasuwa się po przejrzeniu wszystkich wykresów jest taki, że wśród wszystkich playlist Top 50, przewija się sporo piosenek o gatunku muzycznym takim jak hip hop oraz pop. Większość słuchaczy gustuje również w piosenkach o weselszych brzmieniach oraz bardziej energicznych i tanecznych.

Klasyfikacja

Jednym z najważniejszych zadań sztucznej inteligencji jest klasyfikacja, czyli taki problem decyzyjny, który odpowiada do jakiej kategorii należy badany obiekt. Jednym z przykładów jest rozpoznanie rodzaju schorzenia na podstawie objawów. Dodać trzeba, że bardzo często algorytmy sztucznej inteligencji sprawdzają się tu bardzo dobrze. Do projektu został wykorzystany algorytm k-najbliższych sąsiadów.

K najbliższych sąsiadów jest algorytmem dość prostym w zrozumieniu. W pewnej przestrzeni algorytm szuka najbliższego sąsiada (lub sąsiadów) rekordu na podstawie wartości pewnych zmiennych. W przypadku uczenia maszynowego dzieli się dane na

uczące i testowe. (Najczęściej w stosunku 70:30 lub 80:20). Dane uczące to dane na których algorytm się “uczy się”. Gdzie dane testowe to dane które są porównywane do danych uczących.



Przykład działania algorytmu KNN

Dane na których odbyła się klasyfikacja

Danceability, energy, loudness, speechiness, acousticness, liveness, valence, tempo.

Klasyfikacja popularności czyli predykcja popularności piosenki

Pierwszym planem było sklasyfikowanie piosenek na podstawie popularności. Nie było to dobrym pomysłem, ponieważ precyzja wynosiła około 1-3%. Wynika to z prostego względu, że zmienne typu danceability, valence czy energy nie ma wpływu na popularność. Bardzo popularne są piosenki spokojne jak i energiczne, taneczne jak i spokojne oraz smutne jak i bardzo wesołe.

Klasyfikacja gatunku muzycznego czyli predykcja gatunku muzycznego piosenki

Drugim planem było sklasyfikowanie piosenek na podstawie gatunku muzycznego. Tym razem precyzja wyniosła około 20% co jak dla nas było bardzo dobrym wynikiem w porównaniu do poprzedniego podejścia. Problemem w tym wypadku była nierównomierna ilość gatunków w ramce danych (ogromna ilość popu w porównaniu do innych gatunków) co skutkuje dość sfałszowanym wynikiem.

Końcowa konkluzja

Najlepszym sposobem byłoby wybrać pewną ilość piosenek z każdego gatunku i wtedy dokonać klasyfikacji. Jednak nie było na to czasu więc musieliśmy tą część zadania zostawić na ewentualne poprawki w swoim wolnym czasie. Dodatkowo pojawił się problem z web api spotify jakim jest ban na 4 dni.

Źródła:

https://rdrr.io/cran/spotifyr/man/search_spotify.html