Osnove statističkog programiranja

Ak. god. 2023./2024.

Spotify Songs

Petra Buršić, Diego Mišetić, Ante Sorić, Lovro Vuletić

Sadržaj

1. Uvod

U današnje, digitalno doba, glazbene platforme poput Spotifya postale su dio svakodnevnog života ljubitelja glazbe. Spotify je platforma koja pruža ogroman katalog pjesama te sakuplja značajne količine podataka o korisničkim preferencijama i glazbenim trendovima. Analiza ovih podataka postaje ključna kako bismo bolje razumjeli obrasce ponašanja slušatelja, usmjeravali marketinške strategije, te optimizirali glazbene ponude.

Ovaj projekt usredotočit će se na eksploratornu analizu podataka vezanih uz glazbu na Spotifyu, s fokusom na skup podataka koji uključuje različite informacije o pjesmama i playlistama. Stupci poput "track_name", "track_artist", "track_popularity" i mnogi drugi pružaju bitne informacije o karakteristikama pjesama.

Kroz analizu ovih podataka, istražit ćemo pitanja poput koje vrste glazbe dominira na određenim playlistama, kako se popularnost pjesama mijenja tijekom vremena, te kako određene glazbene karakteristike (npr., danceability, energy) utječu na ukupnu popularnost pjesme. Pritom ćemo razmotriti kako se zajednički elementi među najuspješnijim pjesmama na platformi mogu povezati s određenim glazbenim žanrovima.

Ovaj seminar pružit će uvid u kompleksnost podataka koji okružuju glazbene platforme poput Spotifya i istaknuti važnost eksploratorne analize u otkrivanju ključnih uzoraka i informacija koje mogu koristiti glazbenoj industriji, marketinškim stručnjacima i ljubiteljima glazbe diljem svijeta.

2. Opis projekta

3. Eksploratorna analiza

3.1 Opis atributa

Atribut	Tip podatka	Opis
track_id	character	Jedinstveni ID pjesme
track_name	character	Naziv pjesme
track_artist	character	Izvođač pjesme
track_popularity	double	Popularnost pjesme (0-100)
track_album_id	character	Jedinstveni ID albuma
track_album_name	character	Naziv albuma na kojem se nalazi pjesma
track_album_release_date	character	Datum izlaska albuma
playlist_name	character	Naziv playliste
playlist_id	character	Jedinstveni ID playliste
playlist_genre	character	Žanr playliste
playlist_subgenre	character	Podžanr playliste
danceability	double	Plesnost (koliko je pjesma prikladna za plesanje u rasponu 0.0-1.0)
energy	double	Energičnost (perceptualna mjera intenziteta i aktivnosti u rasponu 0.0-1.0)
key	double	Ukupni tonalitet pjesme
loudness	double	Glasnoća pjesme u decibelima
mode	double	Modus pjesme (1 - veliki, 0 - mali)
speechiness	double	Prisutnost izgovorenih riječi u pjesmi
acousticness	double	Mjera povjerenja je li pjesma akustična u rasponu od 0.0 do 1.0
instrumentalness	double	Sadrži li pjesma vokale
liveness	double	Detektira prisutnost publike u snimci
valence	double	Mjera od 0.0 do 1.0 koja opisuje glazbenu pozitivnost koju prijenosi pjesma
tempo	double	Ukupno procijenjeni tempo pjesme u udarcima po minuti (BPM)
duration_ms	double	Trajanje pjesme u milisekundama

Tablica 3.1: Opis tablice podataka o glazbi

3.2 Proces učitavanja i prilagodbe podataka

Proces učitavanja podataka

Učitavanje podataka:

```
```{r setup, include=FALSE}
library(dplyr)    # Biblioteka za manipulaciju podacima
library(readr)    # Biblioteka za čitanje podataka
library(stringr)  # Biblioteka za rad s nizovima znakova
library(tidyr)    # Biblioteka za rad s podacima u širem formatu
library(lubridate) # Biblioteka za rad s datumima
library(ggplot2)   # Biblioteka za vizualizaciju podataka

```

```{r}
df <- read.csv("spotify_songs.csv")  # Učitavanje podataka iz CSV datoteke "spotify_songs.csv"
head(df)                             # Prikaz prvih nekoliko redova podataka
glimpse(df)                          # Prikaz sažetih informacija o podacima
summary(df)                          # Prikaz osnovnih statističkih informacija o podacima

```
```

Slika 3.1: **Učitavanje podataka u R-u**

U prikazanom kodu sa slike, koristimo različite R pakete kako bismo pripremili i istražili skup podataka "spotify_songs.csv". Prvo, koristimo pakete poput **readr**, **dplyr** i **stringr** za čitanje i manipulaciju podacima. Nakon toga, prikazujemo prvih nekoliko redova podataka pomoću funkcije **head** kako bismo dobili inicijalni uvid u strukturu podataka.

Zatim, koristimo funkciju **glimpse** za detaljniji pregled strukture podataka, prikazujući informacije o varijablama, njihovim tipovima podataka i prvim redovima podataka. Na kraju, koristimo funkciju **summary** kako bismo dobili osnovne statističke informacije o numeričkim varijablama u skupu podataka.

Ovi koraci omogućuju nam osnovni uvid u strukturu podataka prije nego što nastavimo s daljnjom analizom i vizualizacijom.

Proces prilagodbe podataka

Stupce *playlist_genre* te *playlist_subgenre* pretvorili smo u u faktore s obzirom da postoji određen broj kategorija jedne i druge varijable. U podatkovnom okviru

također postoji 5 redaka s null vrijednostima, koji su izbačeni radi lakšeg rada s grafovima.

3.3 Vizualizacija Podataka

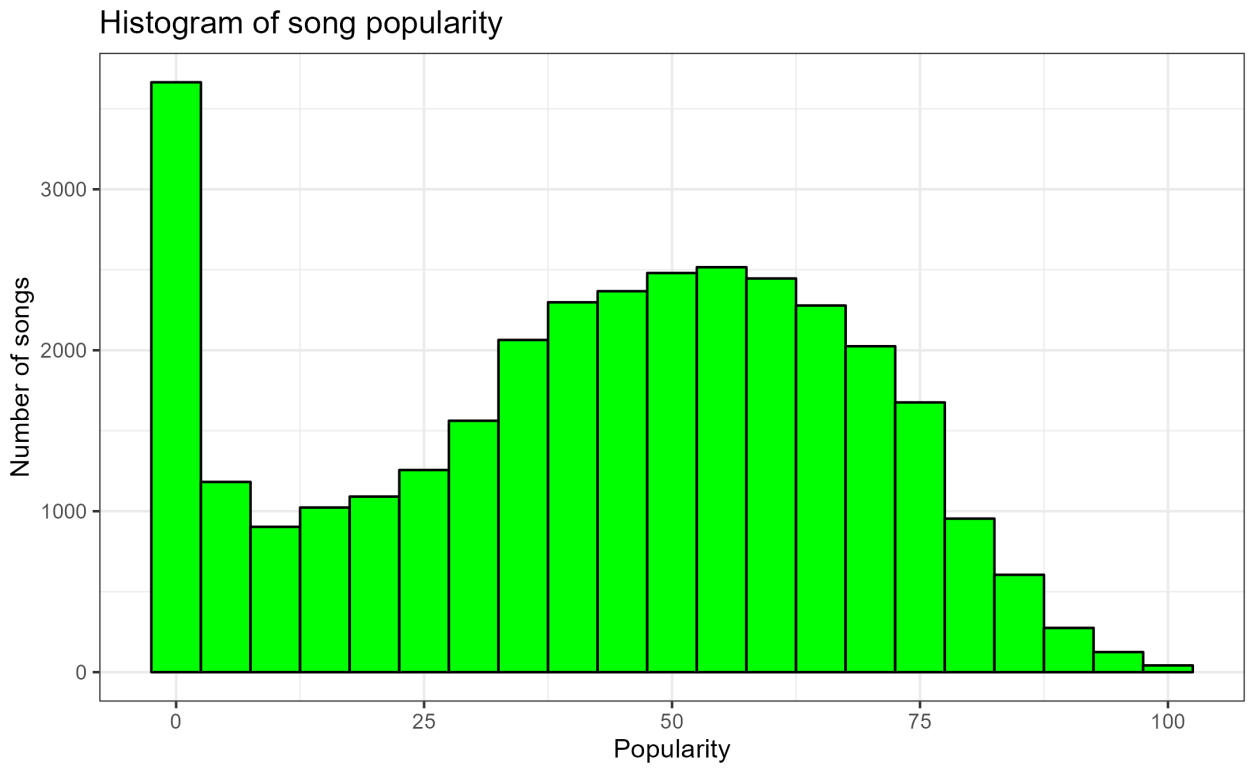
Vizualizacija podataka postaje ključna komponenta analize i interpretacije kompleksnih skupova podataka. U ovom podpoglavlju istražujemo moć vizualizacije u kontekstu glazbene platforme Spotify, prezentirajući neke od grafova kako bismo bolje razumjeli glazbene obrasce, preferencije slušatelja te dinamiku glazbene industrije.

1) Histogram of song popularity

Opis grafa:

Ovaj graf prikazuje histogram popularnosti. Prikazuje distribuciju popularnosti pjesama. Na x-osi nalaze se razine popularnosti pjesama, a y-osi broj pjesama koje se nalaze u pojedinoj razini popularnosti. Ovaj histogram omogućava vizualni pregled koje su razine popularnosti češće, a koje rjeđe.

Slika grafa:



Slika 3.2: Histogram of song popularity

2) Top 10 Artists Based on Popularity

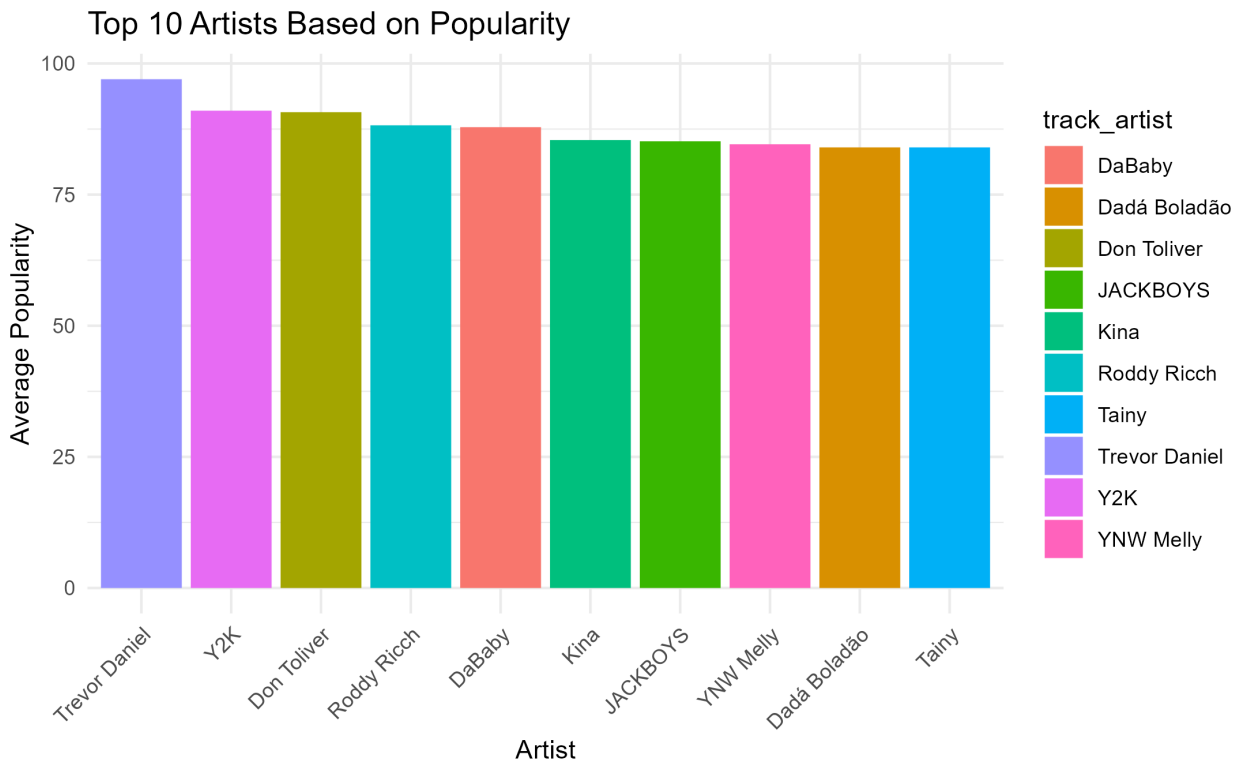
Opis grafa:

Ovaj graf prikazuje deset najpopularnijih glazbenih izvođača temeljem prosječne popularnosti njihovih pjesama. Izračunata je srednja vrijednost popularnosti za svakog izvođača, a zatim su odabrani najbolji deset izvođača prema toj mjeri popularnosti.

Na x-osi su navedeni izvođači, poredani prema visini prosječne popularnosti, dok y-os prikazuje prosječnu popularnost. Svaki šareni stupac predstavlja jednog izvođača, a visina stupa označava njegovu prosječnu popularnost.

Ovaj graf pruža brz i pregledan način usporedbe popularnosti izvođača, omogućujući identifikaciju najboljih deset temeljem prosjeka popularnosti njihovih pjesama.

Slika grafa:



Slika 3.3: Top 10 Artists Based on Popularity

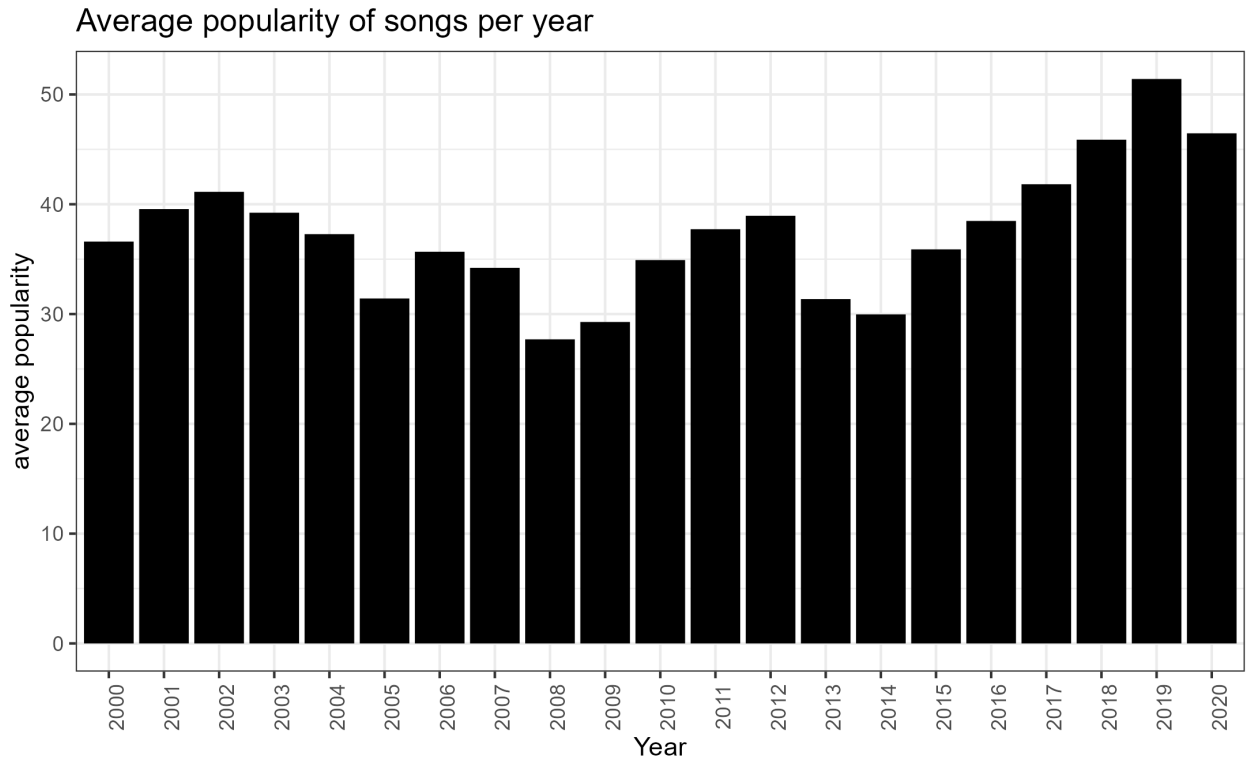
3) Average popularity of songs per year

Opis grafa:

Ovaj stupčasti graf prikazuje prosječnu popularnost pjesama po godinama u

razdoblju od 2000. godine do 2020. godine. X-os ovog grafa su godine u navedenom razdoblju (svaki stupac predstavlja jednu godinu), dok y-os predstavlja prosječnu popularnost. Uvidom u ovaj graf možemo jednostavno vidjeti u kojoj su godini pjesme imale najveću popularnost, te vidjeti kako se popularnost mijenjala tokom tih 20 godina.

Slika grafa:



Slika 3.4: Average popularity of songs per year

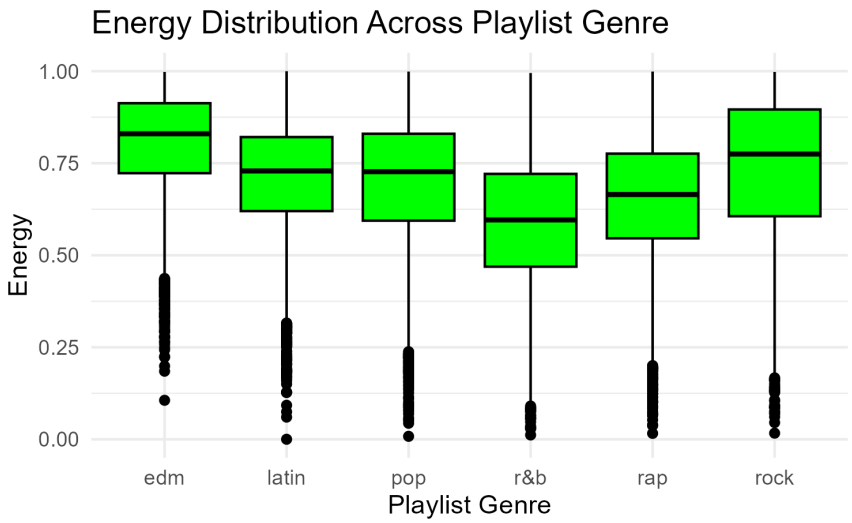
4) Energy Distribution Across Playlist Genre

Opis grafa:

Ovaj graf prikazuje distribuciju energije (y-os) na temelju različitih žanrova playlista (x-os). Svaki boxplot predstavlja jedan žanr, a njegova visina odražava raspon energije unutar tog žanra. Unutar svakog boxplota nalazi se pravokutnik koji predstavlja interkvartilni raspon, a linija unutar pravokutnika označava medijan energije.

Dodatno, postojanje "notcha" u sredini svakog boxplota pruža informaciju o razlikama u medijanima između žanrova.

Slika grafa:



Slika 3.5: Energy Distribution Across Playlist Genre

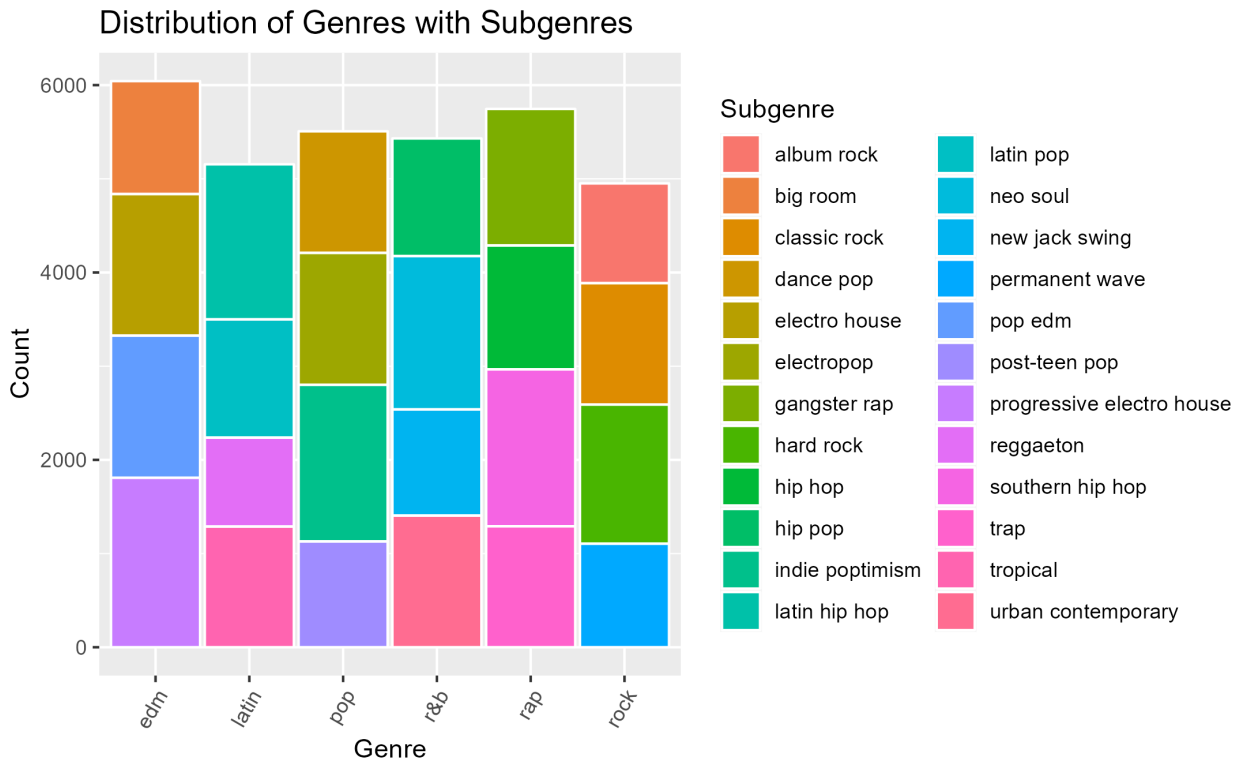
5) Distribution of Genres and Subgenres

Opis grafa:

Ovaj graf prikazuje broj playlista unutar određenih glavnih žanrova, razdijeljenih prema podžanrovima. Na x-osi su navedeni glavni žanrovi playlista, dok y-os pokazuje broj playlista. Svaki šareni segment na stupcu predstavlja određeni podžanr unutar glavnog žanra.

Stupci su složeni jedan na drugi kako bi se vizualno prikazala distribucija podžanrova u okviru svakog glavnog žanra.

Slika grafa:



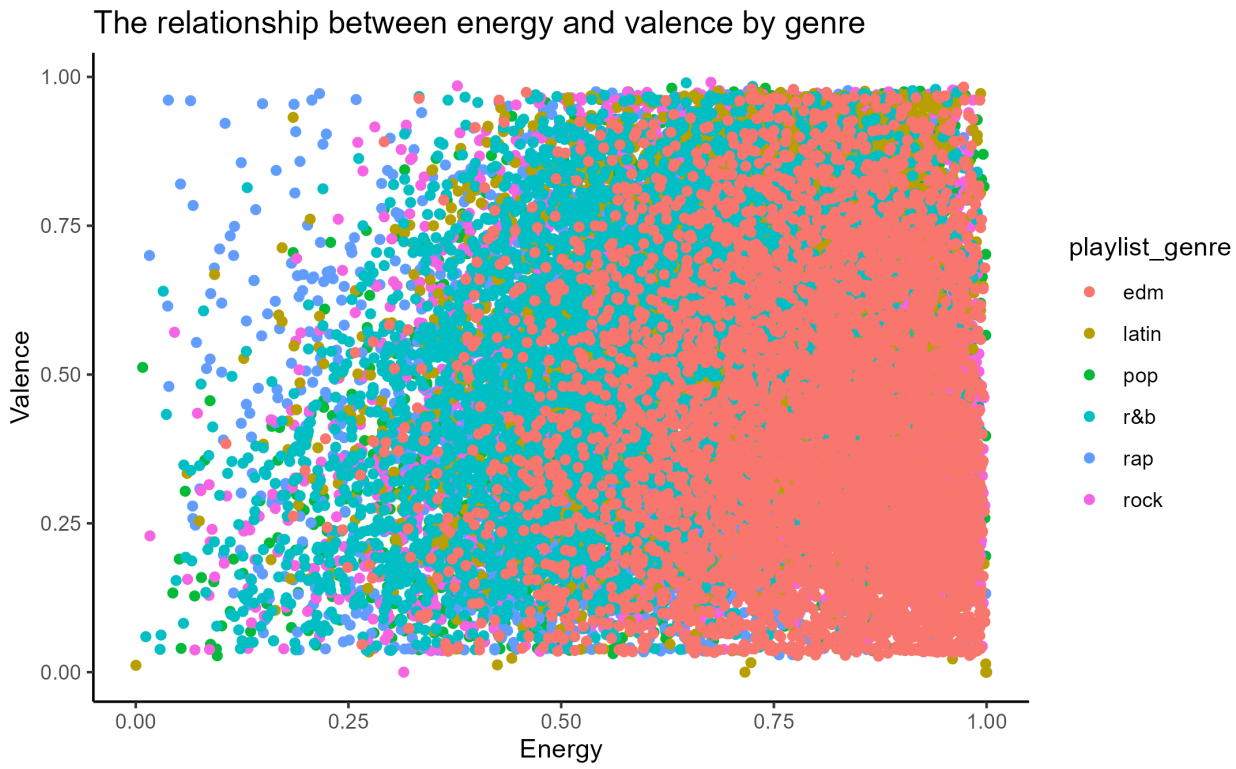
Slika 3.6: Distribution of Genres and Subgenres

6) The relationship between energy and valence by genre

Opis grafa:

Ovaj graf prikazuje odnos između energije i valencije. N a x-osi nalazi se energija koja može biti u rasponu između 0 i 1, a na y-osi nalazi se valencija koja može biti u isto rasponu kao i energija. Svaka točka na grafu prikazuje jednu pjesmu , a njezina pozicija prikazuje odnos energija-valencija. Svaka boja točke prikazuje različiti žanr.

Slika grafa:



Slika 3.7: The relationship between energy ad valence by genre

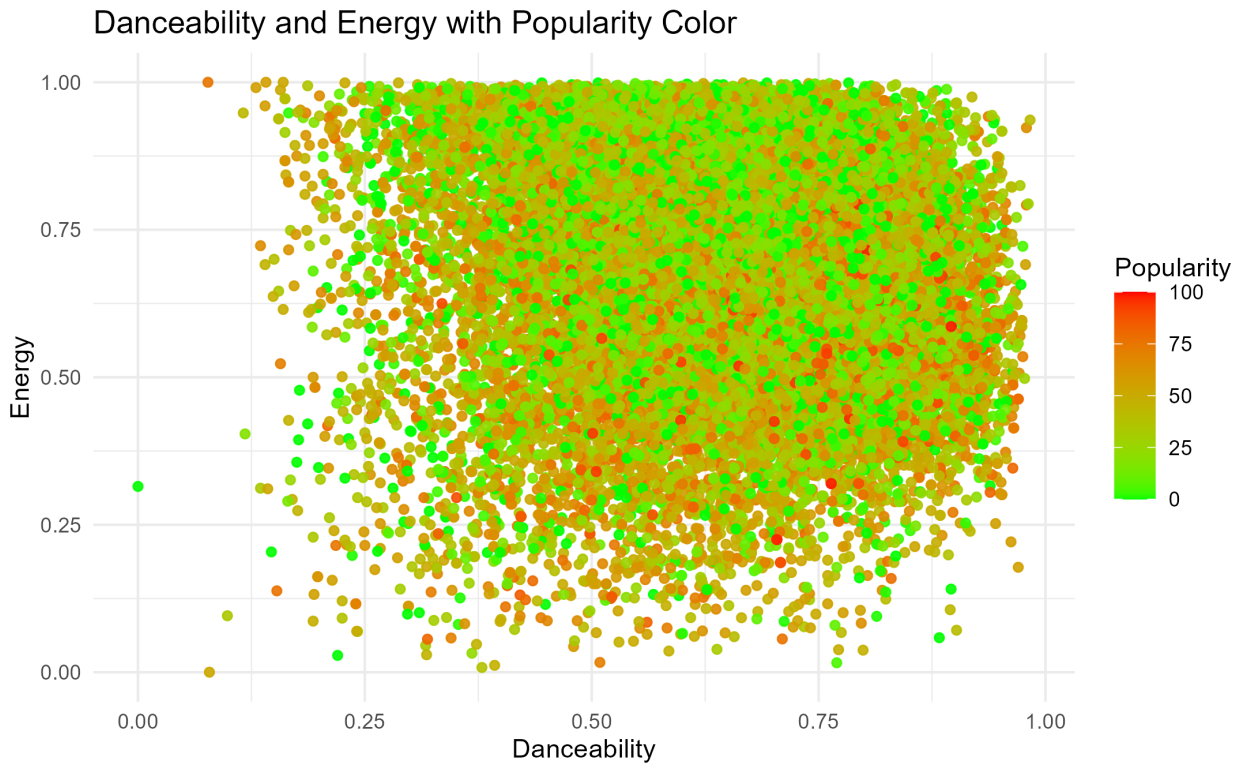
7) Danceability and Energy with Popularity

Opis grafa:

Ovaj šareni graf prikazuje odnos između plesnosti (x-os) i energije (y-os) za različite glazbene pjesme. Svaka točka na grafu predstavlja pojedinu pjesmu, a njezina boja označava razinu popularnosti. Tamnije crvene nijanse označavaju popularnije pjesme, dok svjetlije plave nijanse ukazuju na manju popularnost.

Graf pruža uvid u raznolikost glazbenih preferencija te naglašava da glazbene osobitosti kao što su plesnost i energija nisu nužno ključni faktori koji određuju popularnost pjesama na temelju analize ovog skupa podataka.

Slika grafa:



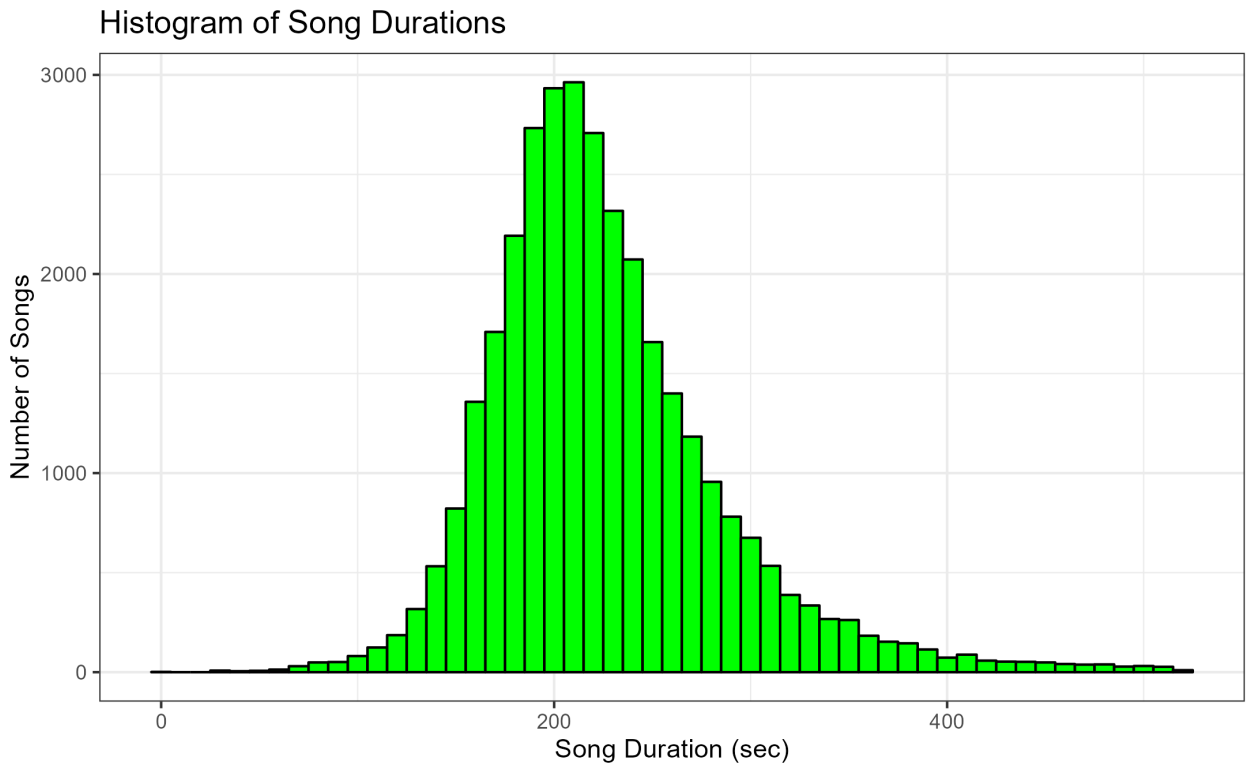
Slika 3.8: Danceability and Energy with Popularity

8) Histogram of song durations

Opis grafa:

Ovaj graf pruža uvid u distribuciju trajanja pjesama. Na x-osi nalaze se različite razine trajanju u sekundama, dok y-os predstavlja broj pjesma u pojedinoj razini. Ovaj zanimljiv histogram omogućava vizualni o najčešćem trajanju pjesama.

Slika grafa:



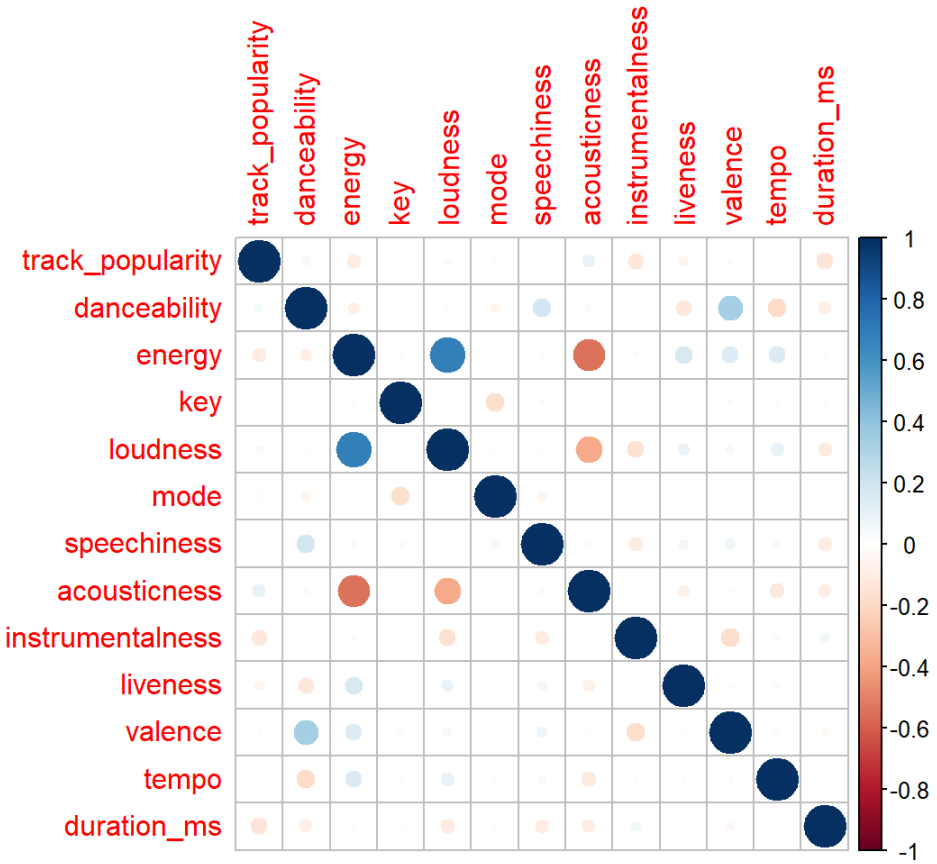
Slika 3.9: Histogram of song duration

3.4 Prediktivni modeli

U nastavku bit će opisan rad s osnovnim prediktivnim modelima: linearna regresija te kNN klasifikacija.

3.4.1 Linearna regresija

Linearna regresija koristi se za predviđanje vrijednosti varijable s obzirom na vrijednosti jedne ili više drugih. Za određivanje koeficijenata smjera koristi se metoda najmanjih kvadrata. U nastavku pokušat ćemo predvidjeti vrijednost varijable *Energy* s pomoću ostalih numeričkih varijabli iz našeg podatkovnog skupa, tj.koristit ćemo višestruku (multiplu) linearnu regresiju. Za početak provjerit ćemo vrijednosti kolinearnosti ulaznih varijabli.



Slika 3.10: Correlation of input variables

Možemo primijetiti da nema prevelikih kolinearnosti, a vidimo da npr. *valence* i *danceability* imaju pozitivnu korelaciju, što nam govori da je moguće da "plesne"

pjesme imaju veću valenciju, tj. pozitivnost. S druge strane, *acousticness* i *loudness* imaju negativnu korelaciju, što znači da akustične trake većinom imaju manju glasnoću.

Za provjeru moguće multikolinearnosti koristit ćemo VIF mjeru koju ćemo izračunati i čiji rezultat se nalazi u nastavku:

```
danceability            key       loudness           mode
    1.246744       1.033497       1.202153       1.038367
 speechiness   acousticness instrumentalness       liveness
    1.067333       1.182559       1.076243       1.032821
     valence          tempo
    1.169310       1.066050
```

Slika 3.11: VIF

Na temelju izračunatih vrijednosti možemo zaključiti da vrlo vjerojatno nema multikolinearnosti ulaznih varijabli

Sada ćemo konačno stvoriti linearni model. Kao što je prije navedeno, varijabla *energy* bit će izlaz, a preostale numeričke varijable ulaz. Podatkovni okvir razdijeljen je u 2 dijela, jedan za treniranje te jedan za testiranje modela. Veličina okvira za treniranje je 70% originalnog podatkovnog okvira. U nastavku prvo slijedi sažetak korištenog linearnog modela.

```
Call:
lm(formula = energy ~ danceability + key + loudness + mode +
    speechiness + acousticness + instrumentalness + liveness +
    valence + tempo, data = spotify.train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.49278 -0.07111  0.00451  0.07526  0.73624

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.9680694  0.0065572 147.635  < 2e-16 ***
danceability     -0.1826222  0.0060672 -30.100  < 2e-16 ***
key               0.0005211  0.0002225   2.342   0.0192 *
loudness          0.0342869  0.0002856 120.031  < 2e-16 ***
mode             -0.0003402  0.0016265  -0.209   0.8343
speechiness      -0.0065974  0.0079198  -0.833   0.4048
acousticness     -0.2717628  0.0038724 -70.180  < 2e-16 ***
instrumentalness  0.1177738  0.0035557  33.122  < 2e-16 ***
liveness          0.0914302  0.0051532  17.743  < 2e-16 ***
valence           0.1498768  0.0036640  40.905  < 2e-16 ***
tempo             0.0002102  0.0000303   6.939 4.06e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1115 on 19838 degrees of freedom
Multiple R-squared:  0.6319, Adjusted R-squared:  0.6317
F-statistic:  3405 on 10 and 19838 DF,  p-value: < 2.2e-16
```

Slika 3.12: Linear regression summary

Možemo vidjeti da su *danceability*, *loudness*, *acousticness*, *instrumentalness*, *livenes* i *valence* jako statistički značajni u ovome modelu, *key* ima manje statistički značajan utjecaj, dok *mode* i *speechiness* vrlo vjerojatno nemaju statističkog utjecaja na ovaj model, zbog visoke p-vrijednosti. Rezidualna standardna pogreška (RSE) nam govori koliko u prosjeku model promašuje kod predviđanja izlazne varijable *energy*, a vidimo da iznosi 0.1115. *Multiple R-squared* i *Adjusted R-squared* iznose 0.6319, odnosno 0.6317. R-kvadrat nam govori o količini varijabilnosti koja je objašnjena modelom, a u našem modelu možemo zaključiti da je otprilike 63% varijabilnosti varijable *energy* objašnjeno modelom. Kako smo koristili višestruku linearnu regresiju više pažnje obratit ćemo na prilagođenu R-kvadrat vrijednost koja prilagođava R-kvadrat mjeru zbog broja prediktora u modelu, međutim vidimo da su otprilike iste. Konačno, F-statistika s p-vrijednosti manjom od 2.2×10^{-16} nam govori da je model statistički značajan u objašnjavanju izlazne varijable.

Sada ćemo pomoću koristeći model koji je istreniran nad trening skupom pokušati

procijeniti vrijednosti varijable *energy* u testnom podatkovnom okviru. Kako bismo provjerili kvalitetu našeg modela, iskoristit ćemo RMSE (*engl. root mean square error*) mjeru koja iznosi 0.1138, što je otprilike isto kao kod rezidualne pogreške

3.4.2 kNN klasifikacija

Ovaj klasifikacijski prediktivni model koristit ćemo za predviđanje kategorijske varijable *genre*. Kako se određena pjesma može nalaziti na više playlista, a svaka playlista može imati drugačiji žanr, za početak ćemo odrediti da ako se pjesma nalazi na više playlista, sve te retke ćemo spojiti u jedan koji će pjesmi pridružiti onaj žanr na kojem pjesma ima najviše playlista.

Kao i kod linearne regresije podatkovni okvir bit će razdijeljen u trening i test ovkir. Također, kako je ovo klasifikacijski model podrazumijeva se da je ciljna varijabla *genre* faktorizirana varijabla. Također sve numeričke varijable bit će normalizirane i svedene na istu skalu vrijednosti.

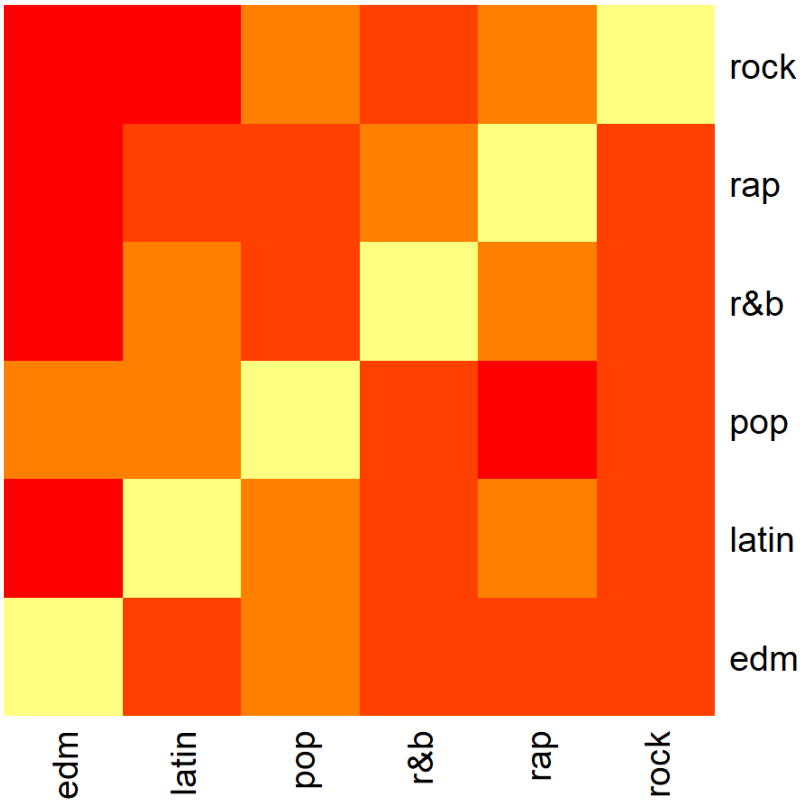
Kako je već navedeno, izlazna varijabla bit će *genre*, a ulazne sve numeričke varijable u okviru, tj. *track_popularity*, *danceability*, *energy*, *key*, *loudness*, *mode*, *speechiness*, *acousticness*, *instrumentalness*, *liveness*, *valence*, *tempo*, *duration_ms*

Kako bismo provjerili uspješnost modela koristit ćemo matricu konfuzije čiji izgled se nalazi u nastavku. Retci označavaju stvarne vrijednosti, a stupci predviđene vrijednosti.

	edm	latin	pop	r&b	rap	rock
edm	910	153	255	108	125	145
latin	132	352	227	196	202	160
pop	204	206	278	196	184	189
r&b	95	224	203	439	270	199
rap	121	229	240	296	478	225
rock	134	140	197	186	231	378

Slika 3.13: Confusion matrix

Sada ćemo prikazati ovu matricu kao *heatmap*



Slika 3.14: Confusion matrix heatmap

Svjetlije boje označavaju veći broj predikcija za tu kombinaciju retka i stupca. Vidimo kako je dijagonala najsvjetlija, što označava točne predikcije. Također postoje mnoge krive predikcije, pogotovo između pop i latin te rap i r&b glazbe, što je donekle očekivano s obzirom na mnoge sličnosti između navedenih tipova glazbe.

4. Zaključak