Osnove statističkog programiranja

Ak. god. 2023./2024.

Spotify Songs

Petra Buršić, 0036539882

Diego Mišetić, 0036543343

Ante Sorić, 0036539765

Lovro Vuletić, 0036542213

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:

```
spotify <- read_csv("spotify_songs.csv")
```

Slika 3.1: **Učitavanje podatkovnog okvira *spotify_songs***

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.

```
head(spotify)
```

```
## # A tibble: 6 × 23
##   track_id               track_name track_artist track_popularity track_album_id
##   <chr>                  <chr>      <chr>                   <dbl> <chr>
## 1 6f807x0ima9a1j3VPbc7VN I Don't C… Ed Sheeran                 66 2oCs0DGTsRO98…
## 2 0r7CVbZTWZgbTCYdfa2P31 Memories … Maroon 5                   67 63rPSO264uRjW…
## 3 1z1Hg7Vb0AhHDiEmnDE79l All the T… Zara Larsson               70 1HoSmj2eLcsrR…
## 4 75FpbthrwQmzHlBJLuGdC7 Call You … The Chainsm…              60 1nqYsOef1yKKu…
## 5 1e8PAfcKUYoKkxPhrHqw4x Someone Y… Lewis Capal…              69 7m7vv9wlQ4i0L…
## 6 7fvUMiyapMsRRxr07cU8Ef Beautiful… Ed Sheeran                67 2yiy9cd2QktrN…
## # ℹ 18 more variables: track_album_name <chr>, track_album_release_date <chr>,
## #   playlist_name <chr>, playlist_id <chr>, playlist_genre <chr>,
## #   playlist_subgenre <chr>, danceability <dbl>, energy <dbl>, key <dbl>,
## #   loudness <dbl>, mode <dbl>, speechiness <dbl>, acousticness <dbl>,
## #   instrumentalness <dbl>, liveness <dbl>, valence <dbl>, tempo <dbl>,
## #   duration_ms <dbl>
```

Slika 3.2: Rezultat poziva *head* funckije

```
glimpse(spotify)
```

```
## Rows: 32,833
## Columns: 23
## $ track_id                <chr> "6f807x0ima9a1j3VPbc7VN", "0r7CVbZTWZgbTCYdfa…
## $ track_name              <chr> "I Don't Care (with Justin Bieber) - Loud Lux…
## $ track_artist            <chr> "Ed Sheeran", "Maroon 5", "Zara Larsson", "Th…
## $ track_popularity        <dbl> 66, 67, 70, 60, 69, 67, 62, 69, 68, 67, 58, 6…
## $ track_album_id          <chr> "2oCs0DGTsRO98Gh5ZSl2Cx", "63rPSO264uRjW1X5E6…
## $ track_album_name        <chr> "I Don't Care (with Justin Bieber) [Loud Luxu…
## $ track_album_release_date <chr> "2019-06-14", "2019-12-13", "2019-07-05", "20…
## $ playlist_name           <chr> "Pop Remix", "Pop Remix", "Pop Remix", "Pop R…
## $ playlist_id             <chr> "37i9dQZF1DXcZDD7cfEKhW", "37i9dQZF1DXcZDD7cf…
## $ playlist_genre          <chr> "pop", "pop", "pop", "pop", "pop", "pop", "po…
## $ playlist_subgenre       <chr> "dance pop", "dance pop", "dance pop", "dance…
## $ danceability            <dbl> 0.748, 0.726, 0.675, 0.718, 0.650, 0.675, 0.4…
## $ energy                  <dbl> 0.916, 0.815, 0.931, 0.930, 0.833, 0.919, 0.8…
## $ key                     <dbl> 6, 11, 1, 7, 1, 8, 5, 4, 8, 2, 6, 8, 1, 5, 5,…
## $ loudness                <dbl> -2.634, -4.969, -3.432, -3.778, -4.672, -5.38…
## $ mode                    <dbl> 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, …
## $ speechiness             <dbl> 0.0583, 0.0373, 0.0742, 0.1020, 0.0359, 0.127…
## $ acousticness            <dbl> 0.10200, 0.07240, 0.07940, 0.02870, 0.08030, …
## $ instrumentalness        <dbl> 0.00e+00, 4.21e-03, 2.33e-05, 9.43e-06, 0.00e…
## $ liveness                <dbl> 0.0653, 0.3570, 0.1100, 0.2040, 0.0833, 0.143…
## $ valence                 <dbl> 0.518, 0.693, 0.613, 0.277, 0.725, 0.585, 0.1…
## $ tempo                   <dbl> 122.036, 99.972, 124.008, 121.956, 123.976, 1…
## $ duration_ms             <dbl> 194754, 162600, 176616, 169093, 189052, 16304…
```

Slika 3.3: Rezultat poziva *glimpse* funckije

```
summary(spotify)
```

```
##    track_id          track_name        track_artist       track_popularity
##  Length:32833       Length:32833       Length:32833       Min.   :  0.00
##  Class :character   Class :character   Class :character   1st Qu.: 24.00
##  Mode  :character   Mode  :character   Mode  :character   Median : 45.00
##                                                           Mean   : 42.48
##                                                           3rd Qu.: 62.00
##                                                           Max.   :100.00
##  track_album_id     track_album_name   track_album_release_date
##  Length:32833       Length:32833       Length:32833
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  playlist_name      playlist_id        playlist_genre     playlist_subgenre
##  Length:32833       Length:32833       Length:32833       Length:32833
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##   danceability        energy              key             loudness
##  Min.   :0.0000   Min.   :0.000175   Min.   : 0.000   Min.   :-46.448
##  1st Qu.:0.5630   1st Qu.:0.581000   1st Qu.: 2.000   1st Qu.: -8.171
##  Median :0.6720   Median :0.721000   Median : 6.000   Median : -6.166
##  Mean   :0.6548   Mean   :0.698619   Mean   : 5.374   Mean   : -6.720
##  3rd Qu.:0.7610   3rd Qu.:0.840000   3rd Qu.: 9.000   3rd Qu.: -4.645
##  Max.   :0.9830   Max.   :1.000000   Max.   :11.000   Max.   :  1.275
##       mode          speechiness      acousticness    instrumentalness
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000000
##  1st Qu.:0.0000   1st Qu.:0.0410   1st Qu.:0.0151   1st Qu.:0.0000000
##  Median :1.0000   Median :0.0625   Median :0.0804   Median :0.0000161
##  Mean   :0.5657   Mean   :0.1071   Mean   :0.1753   Mean   :0.0847472
##  3rd Qu.:1.0000   3rd Qu.:0.1320   3rd Qu.:0.2550   3rd Qu.:0.0048300
##  Max.   :1.0000   Max.   :0.9180   Max.   :0.9940   Max.   :0.9940000
##     liveness         valence           tempo         duration_ms
##  Min.   :0.0000   Min.   :0.0000   Min.   :  0.00   Min.   :  4000
##  1st Qu.:0.0927   1st Qu.:0.3310   1st Qu.: 99.96   1st Qu.:187819
##  Median :0.1270   Median :0.5120   Median :121.98   Median :216000
##  Mean   :0.1902   Mean   :0.5106   Mean   :120.88   Mean   :225800
##  3rd Qu.:0.2480   3rd Qu.:0.6930   3rd Qu.:133.92   3rd Qu.:253585
##  Max.   :0.9960   Max.   :0.9910   Max.   :239.44   Max.   :517810
```

Slika 3.4: Rezultat poziva *summary* funckije

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.

```
spotify %>% filter(!is.na(spotify$track_name)) -> spotify
spotify$playlist_genre <- as.factor(spotify$playlist_genre)
spotify$playlist_subgenre <- as.factor(spotify$playlist_subgenre)
```

Slika 3.5: Prilagodba podataka

3.3 Vizualizacija Podataka

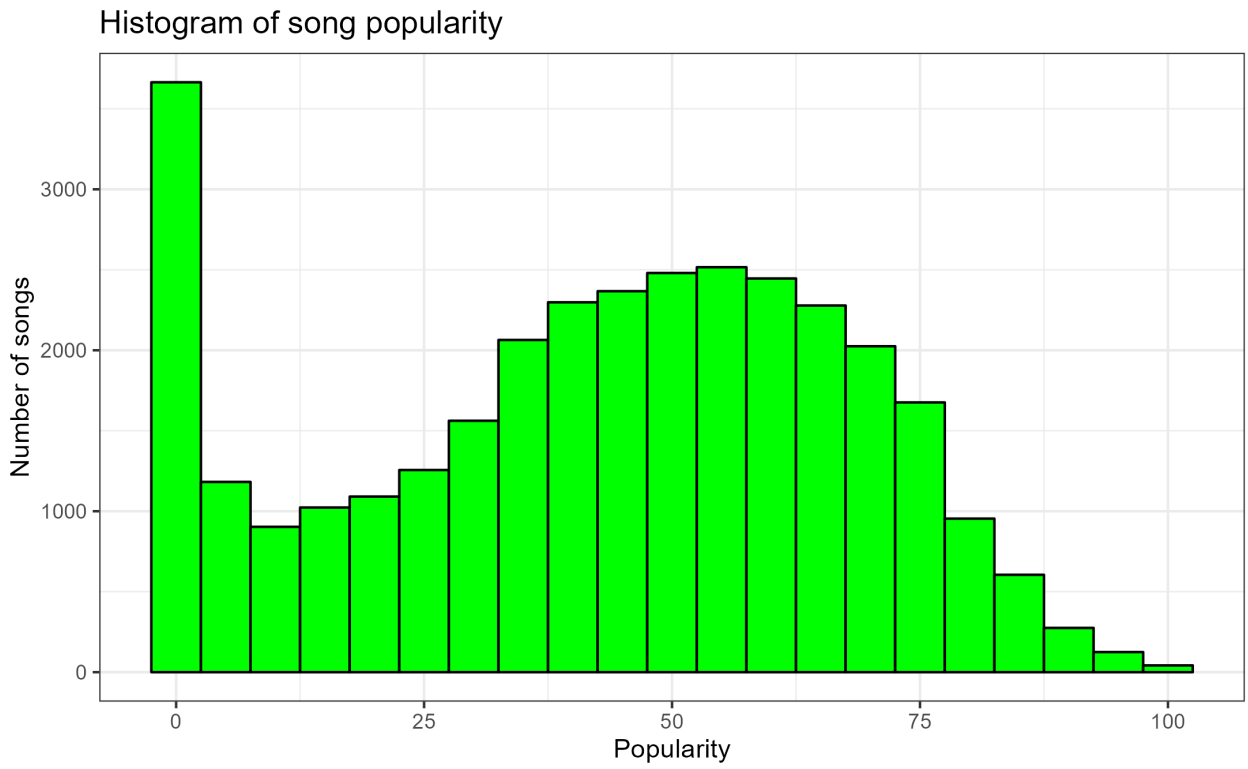
Vizualizacija podataka postaje ključna komponenta analize i interpretacije kompleksnih skupova podataka. U ovom podpoglavlju istražujemo moć vizualizacije u kontekstu glazbene platforme Spotify, prezentirajući neke od grafova kako bismo bolje razumjeli glazbene obrasce, preferencije slušatelja te dinamiku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

1) Histogram popularnosti pjesama

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 3.6: Histogram popularnosti pjesama

2) Top 10 umjetnika po popularnosti

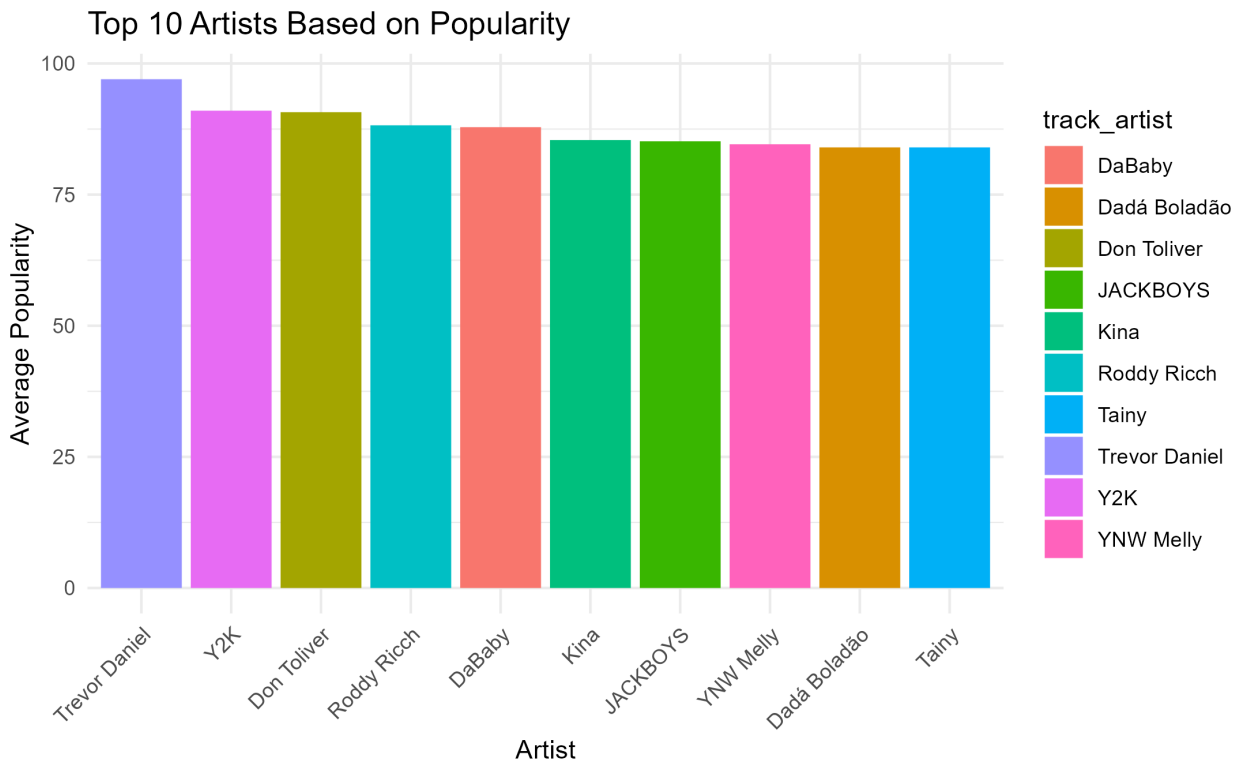
Opis grafa:

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 3.7: Top 10 umjetnika po popularnosti

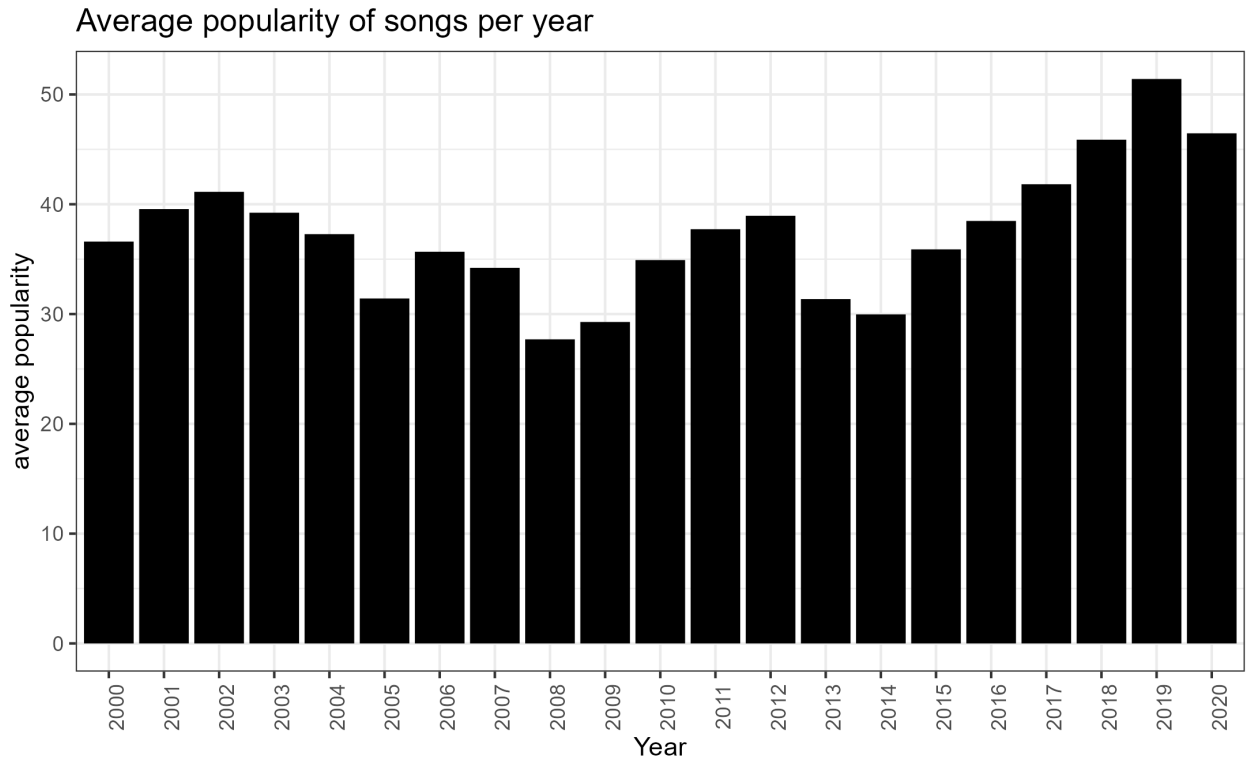
3) Prosječna popularnost pjesama po godinama

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 3.8: Prosječna popularnost pjesama po godinama

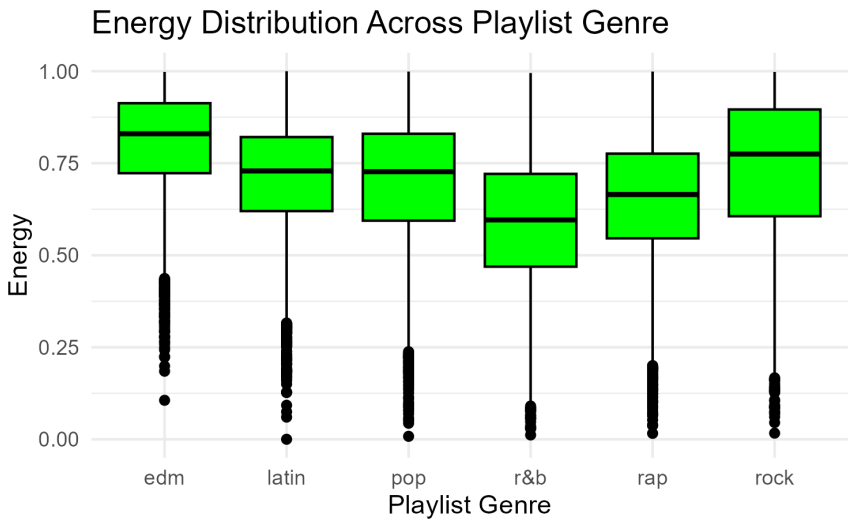
4) Distribucija energije kroz žanrove playlista

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:



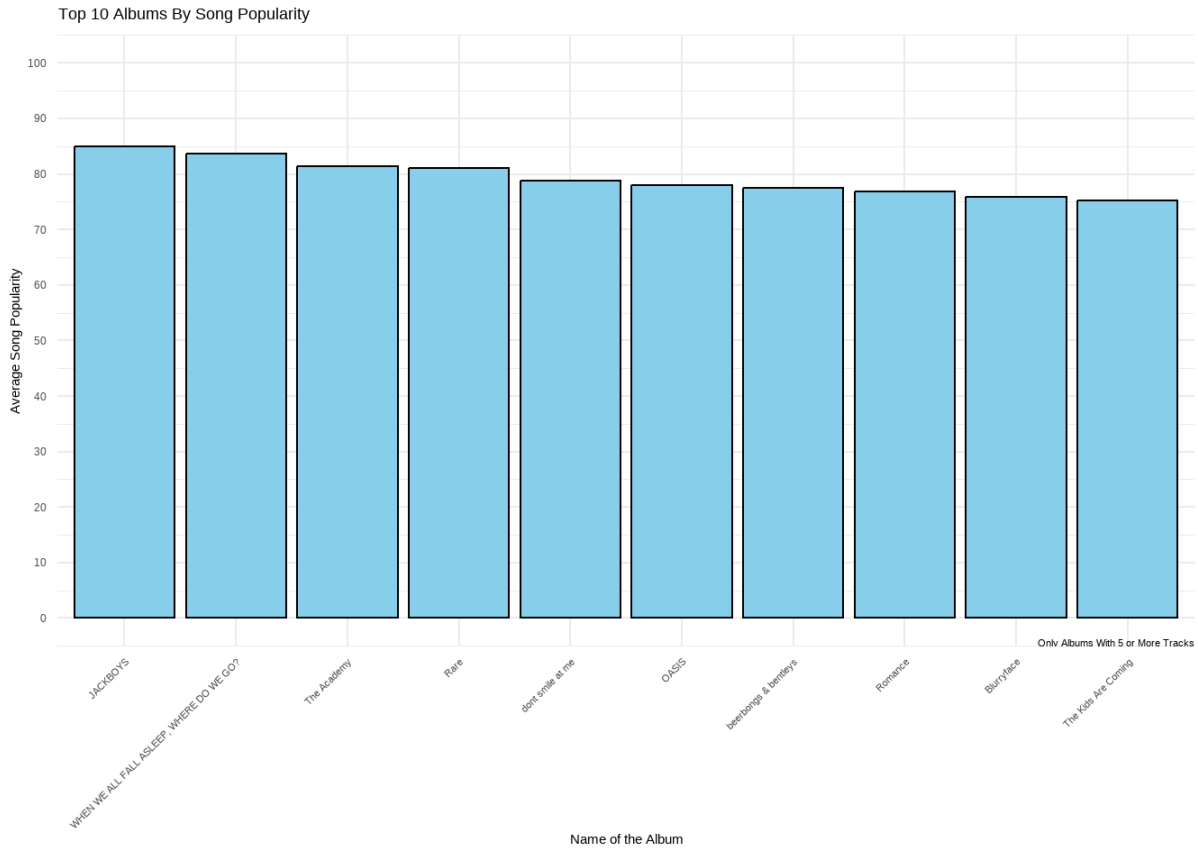
Slika 3.9: Distribucija energije kroz žanrove playlista

5) Top 10 albuma prema popularnosti pjesama

Opis grafa:

Na x osi nalazi se ime albuma, dok se na y osi nalazi prosječna popularnost pjesme u tom albumu. Vidimo kako su u prosjeku pjesme s albuma *JACKBOYS*, *WHEN WE ALL FALL ASLEEP, WHERE DO WE GO?*, *The Academy*, *Rare* i ostali najpopularnije.

Slika grafa:



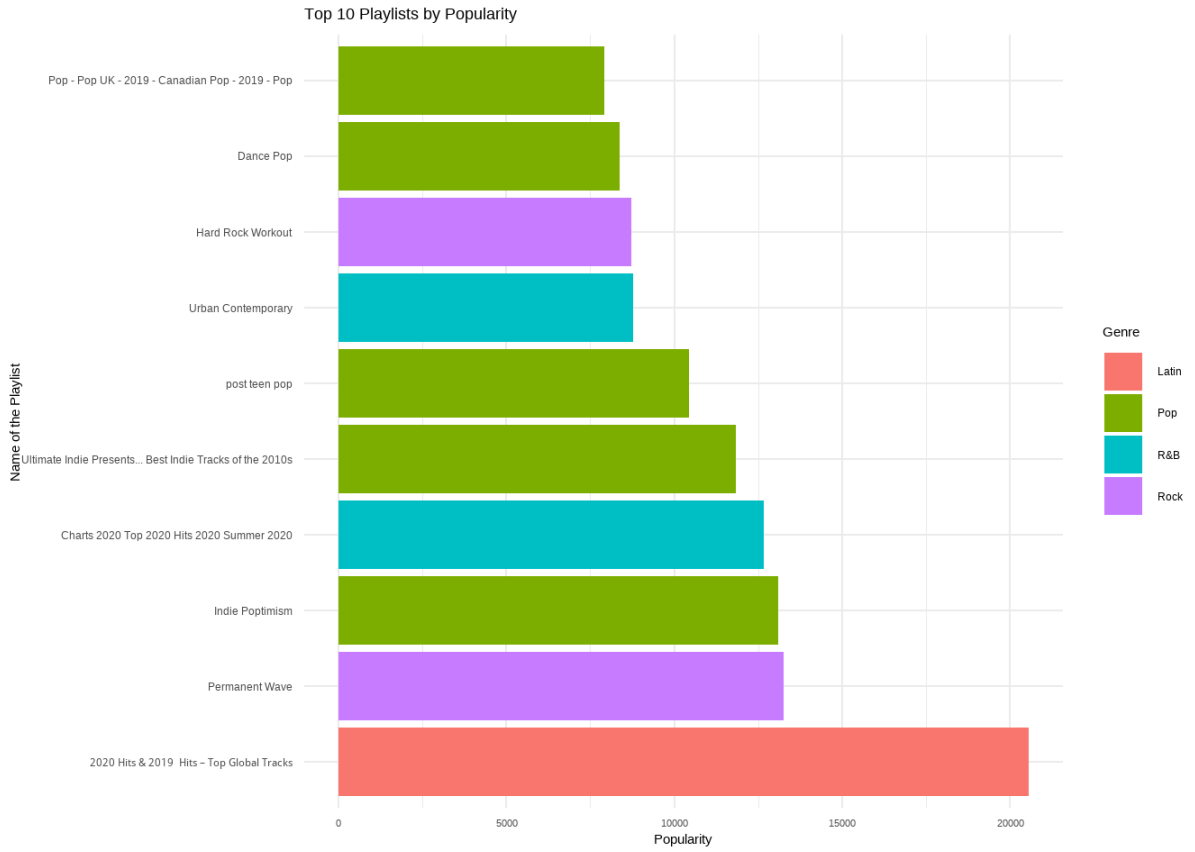
Slika 3.10: Top 10 albuma prema popularnosti pjesama

6) Top 10 playlista po popularnosti

Opis grafa:

Na y osi nalaze se 10 najpopularnijih playlista, a na y os nam govori zbroj popularnosti svih pjesama u toj playlisti. Legenda prikazuje žanr pojedinih playlista.

Slika grafa:



Slika 3.11: Top 10 playlista po popularnosti

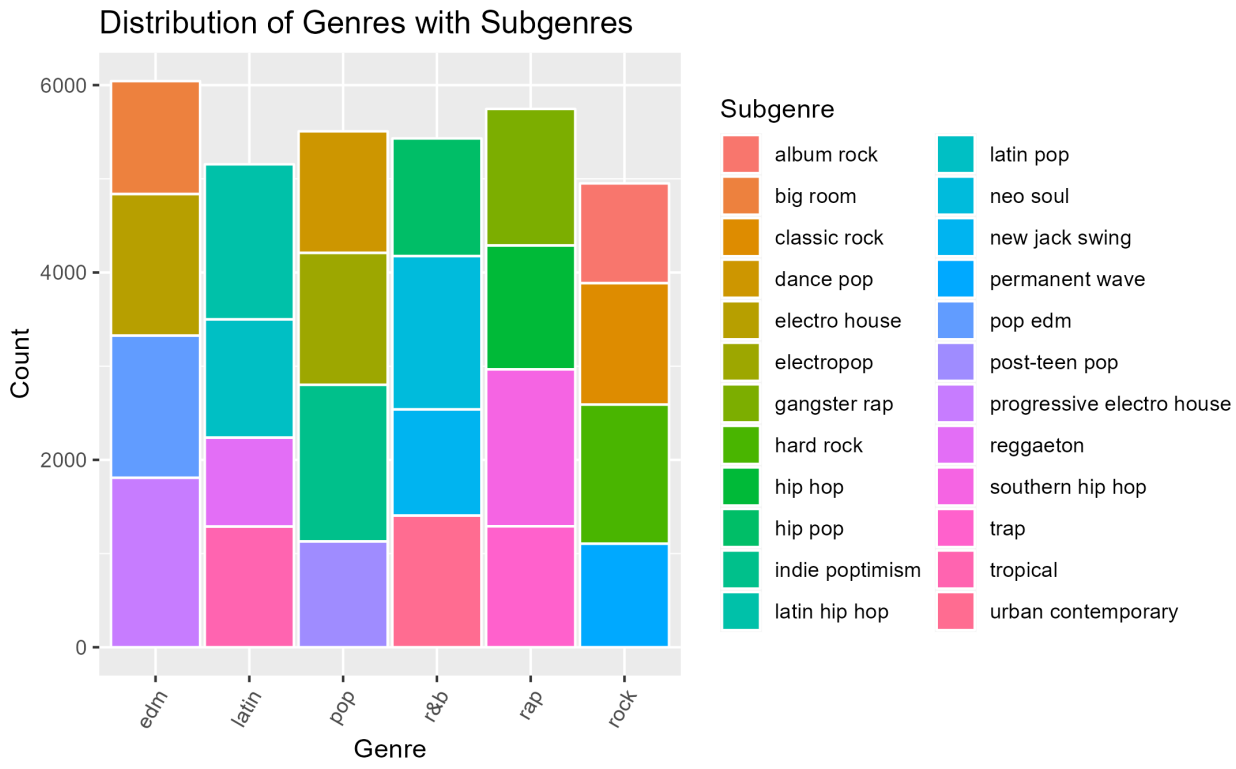
7) Distribucija žanrova i podžanrova

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:



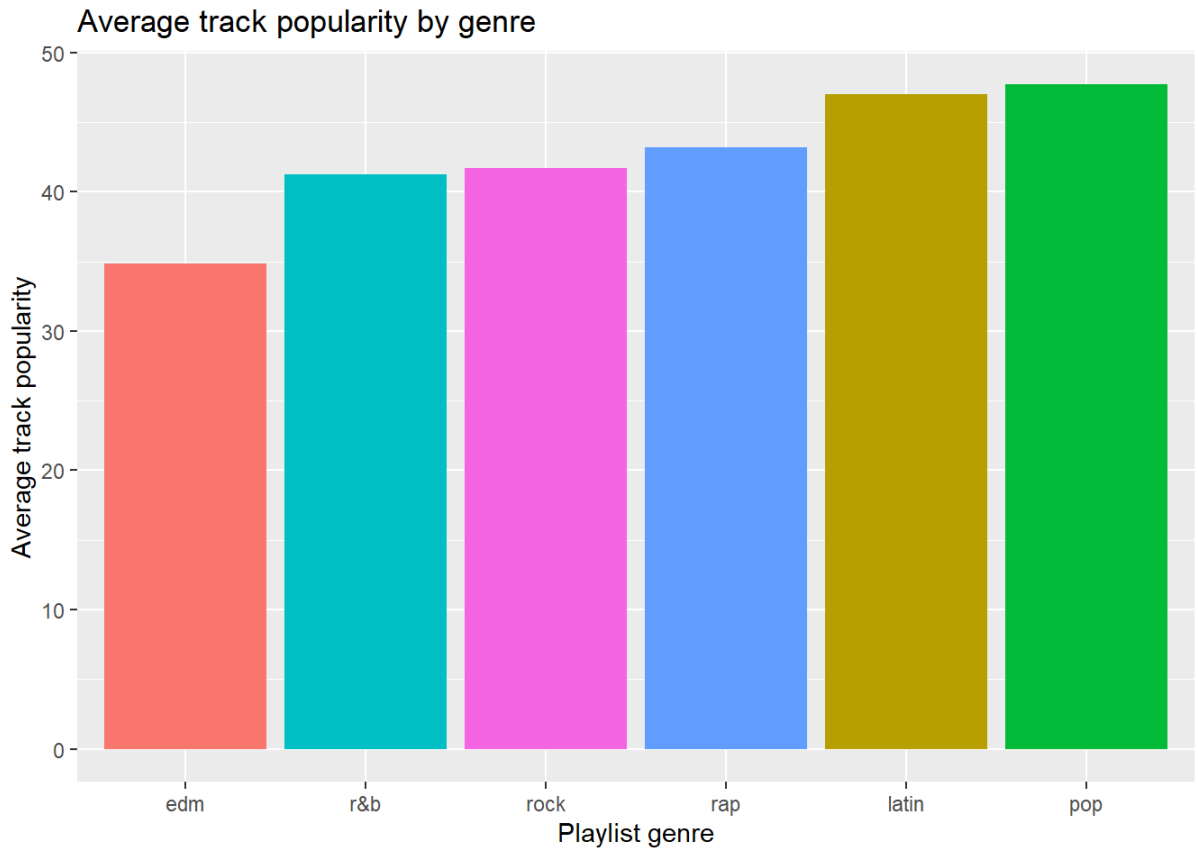
Slika 3.12: Distribucija žanrova i podžanrova

8) Prosječna popularnost pjesme po žanru

Opis grafa:

Na grafu možemo vidjeti prosječnu popularnost pjesama grupiranih na osnovu žanra playliste u kojoj se nalaze. Možemo vidjeti kako su pop i latin najpopulaniji, dok je edm najmanje popularan među slušačima.

Slika grafa:



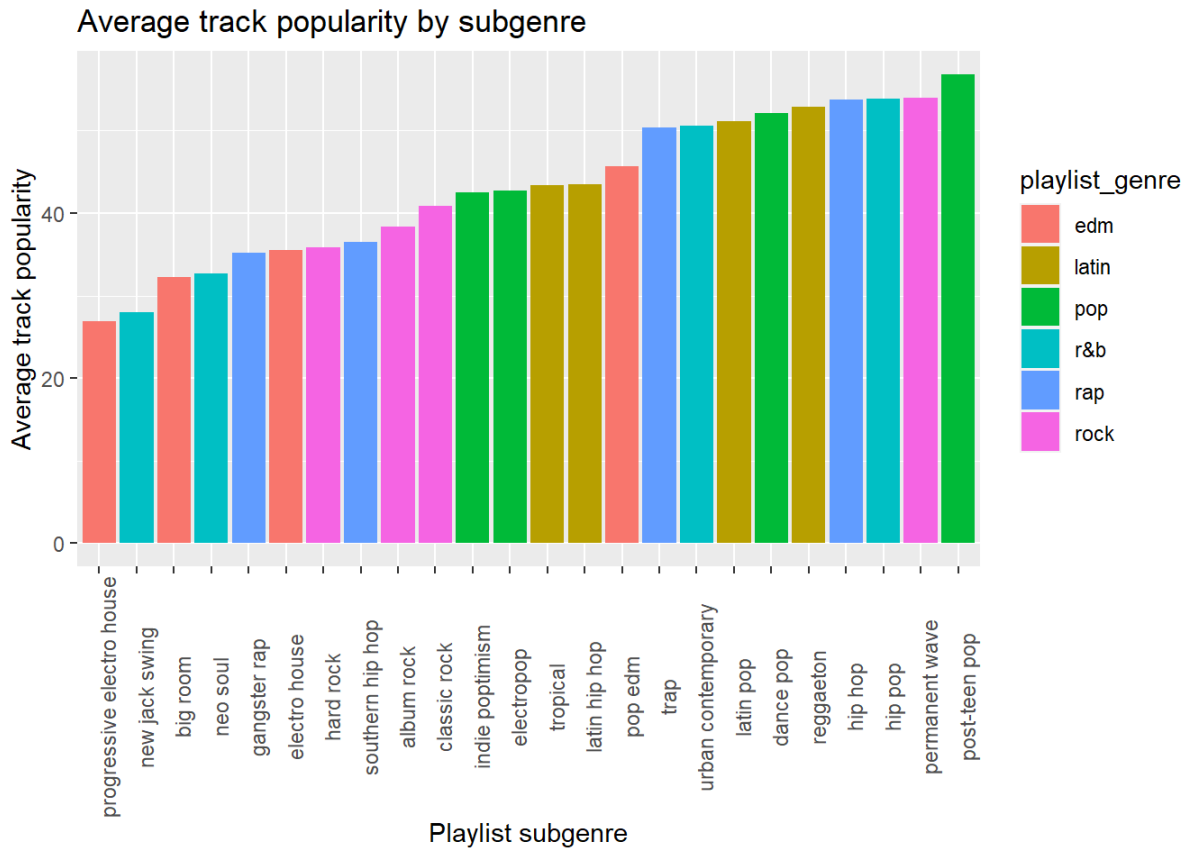
Slika 3.13: Prosječna popularnost pjesme po žanru

9) Prosječna popularnost pjesme po podžanru

Opis grafa:

Na ovome grafu možemo vidjeti prikaz sličan kao i na prethodnom, međutim sada su pjesme grupirane na temelju podžanra. Boje stupaca označavaju vrstu žanra, kako bismo bolje vidjeli popularnost pojedinih podžanrova s obzirom na njihov žanr. Vidimo kako se ističe post-teen pop, iz čega bi mogli zaključiti da su velik broj korisnika Spotifyja mladu ljudi nakon tinejdžerske dobi.

Slika grafa:



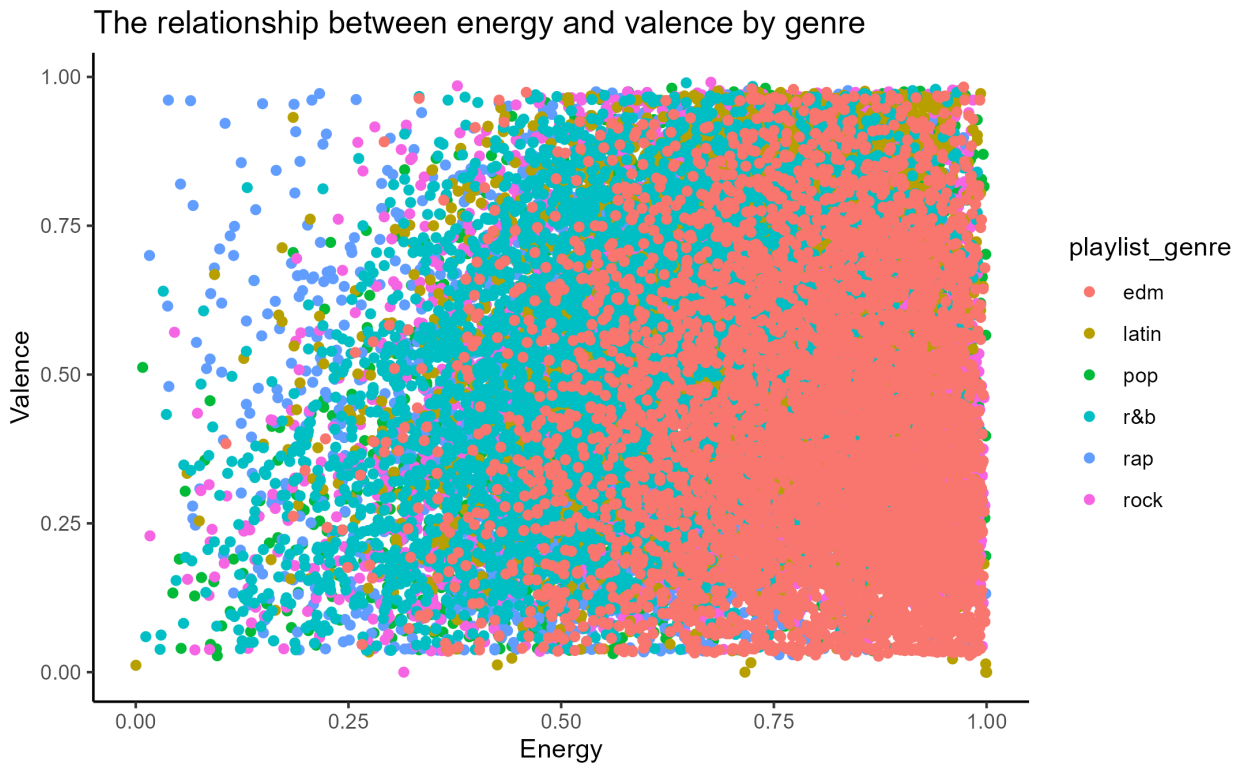
Slika 3.14: Prosječna popularnost pjesme po podžanru

10) Odnos između energije i valencije u odnosu na žanr

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 3.15: Odnos između energije i valencije u odnosu na žanr

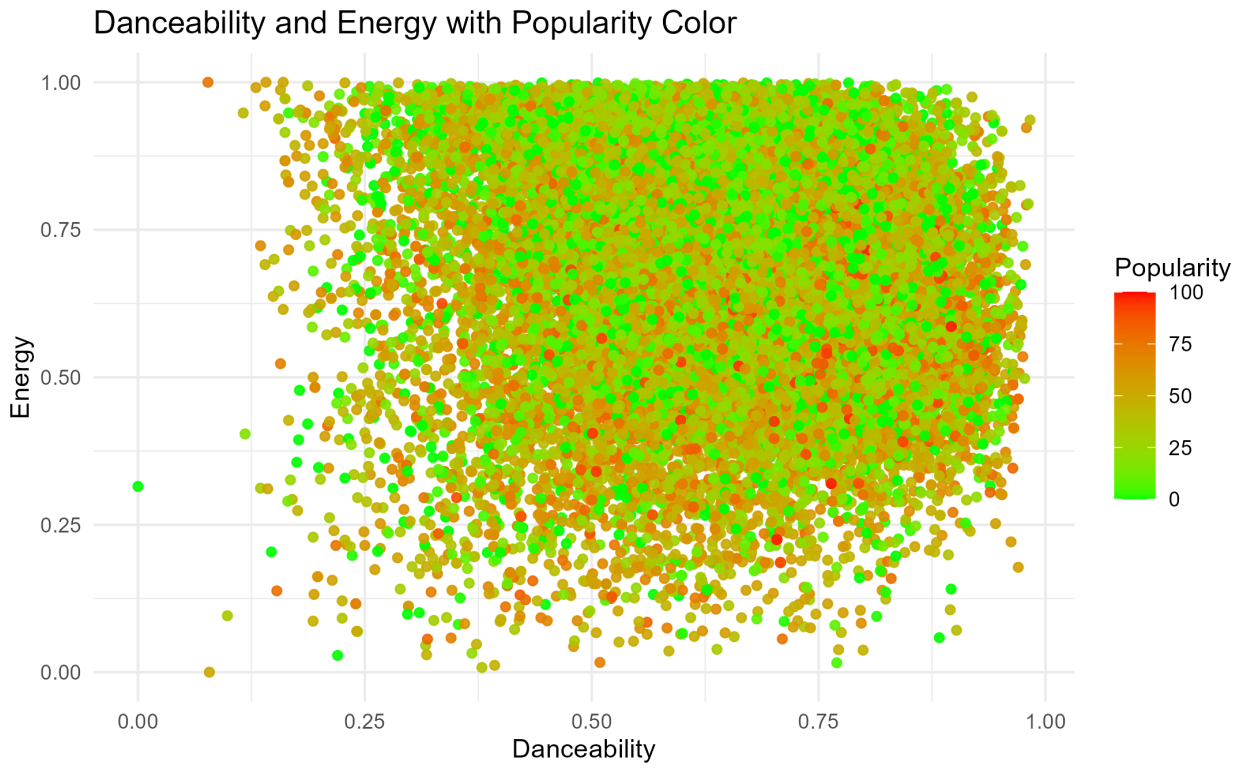
11) Odnos energije i plesnosti za različite popularnosti pjesama

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:



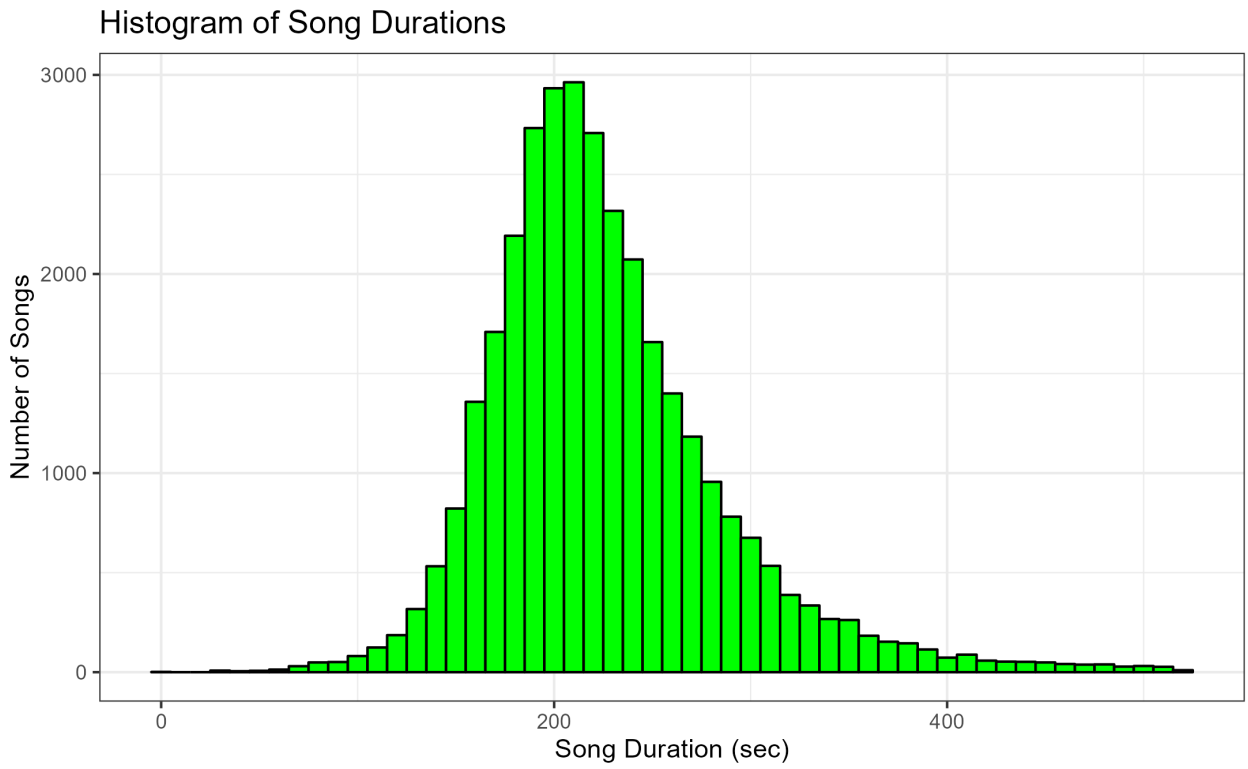
Slika 3.16: Odnos energije i plesnosti za različite popularnosti pjesama

12) Histogram trajanja pjesama

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 vizualnu interpretaciju o najčešćem trajanju pjesama.

Slika grafa:



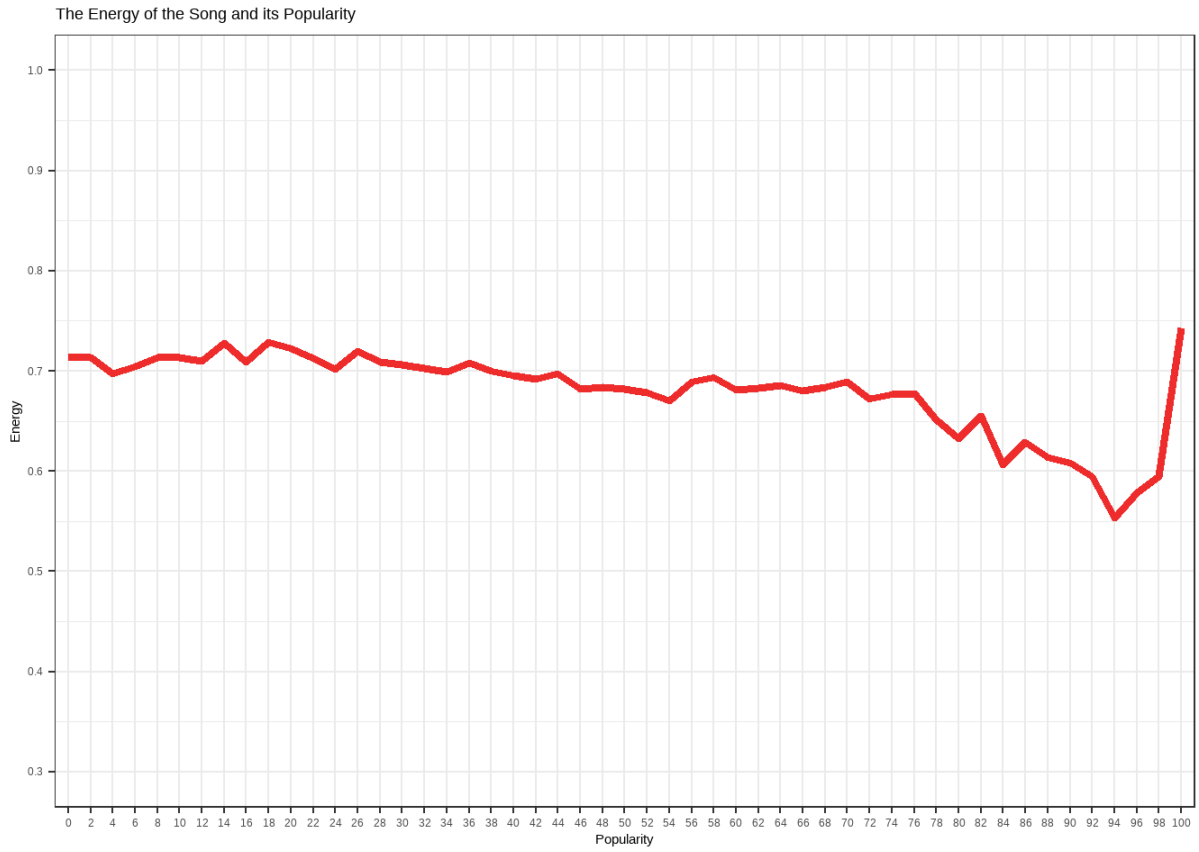
Slika 3.17: Histogram trajanja pjesama

13) Odnos energije pjesme i njene popularnosti

Opis grafa:

Graf prikazuje odnos pupularnosti pjesme s njenom energijom. Vidimo kako se linija energije uglavnom kreće oko vrijednosti 0.7, dok je vidljiv mali pad do vrijednosti 0.6 prema kraju x-osi, te ponovno porast do vrijednosti 0.7 na kraju grafa. Ove veće oscilacije na kraju grafa možemo pripisati manjem broju pjesama s velikom vrijednosti varijable *popularity*.

Slika grafa:



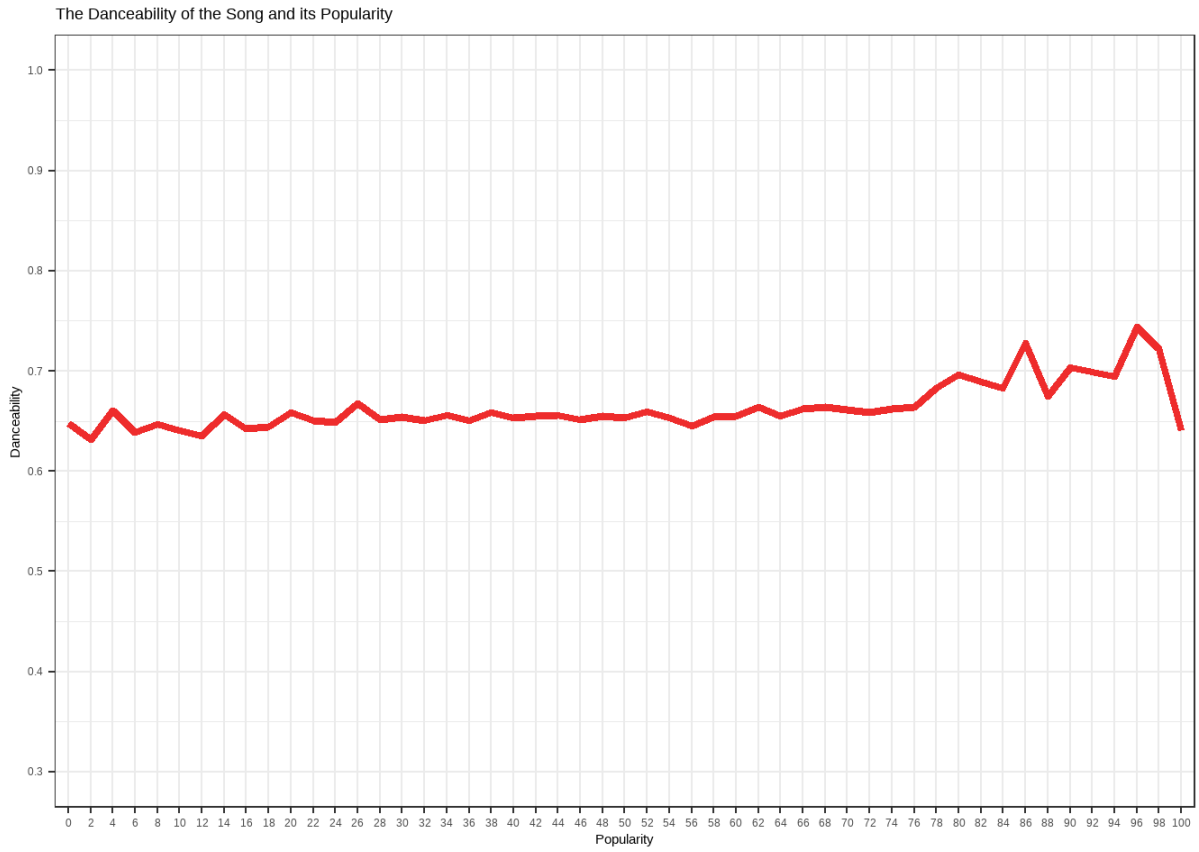
Slika 3.18: Odnos energije pjesme i njene popularnosti

14) Odnos energije pjesme i njezine plesnosti

Opis grafa:

Graf prikazuje odnos popularnosti pjesme s njenom plesnošću. Ovdje je funkcija koja prikazuje plesnost uglavnom ravnomjerno raspoređena oko vrijednosti 0.65, dok je s porastom popularnosti vidljiv blagi porast funkcije.

Slika grafa:



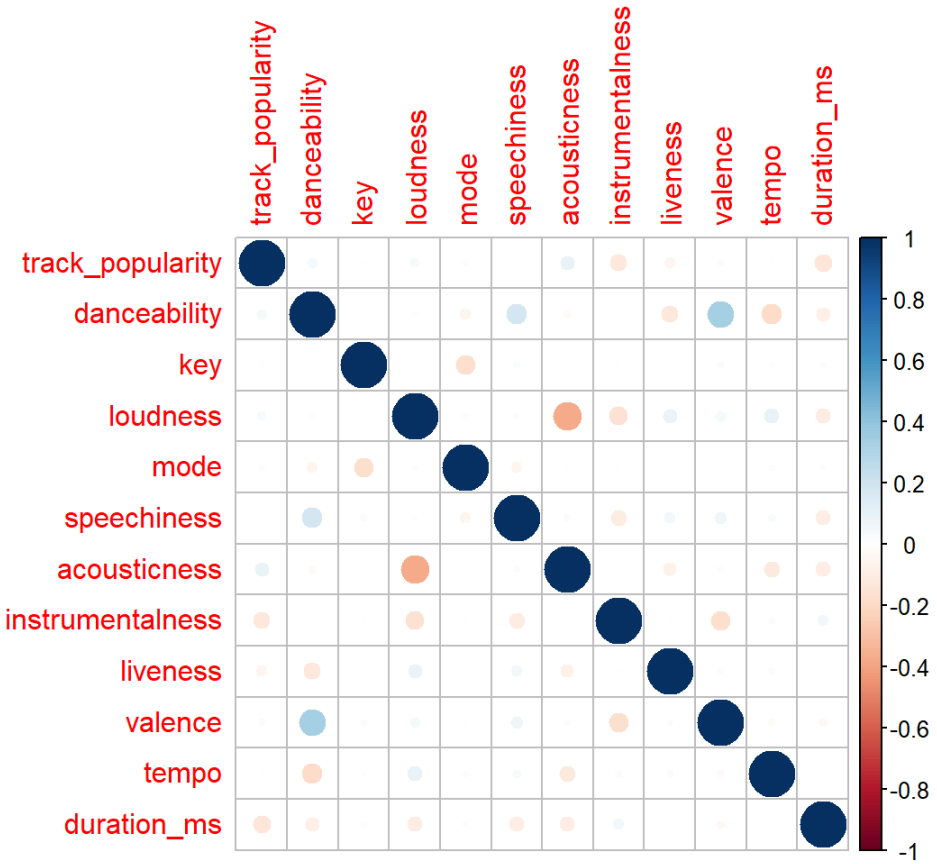
Slika 3.19: Odnos energije pjesme i njezine plesnosti

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 ćemo pokušati 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 3.20: Korelacije ulaznih varijabli

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 3.21: VIF

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

Sada ćemo konačno stvoriti linearni model. Kao što je već navedeno, varijabla *energy* bit će izlaz, a preostale numeričke varijable ulazi u modelu. 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 3.22: Sažetak linearnog modela

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.

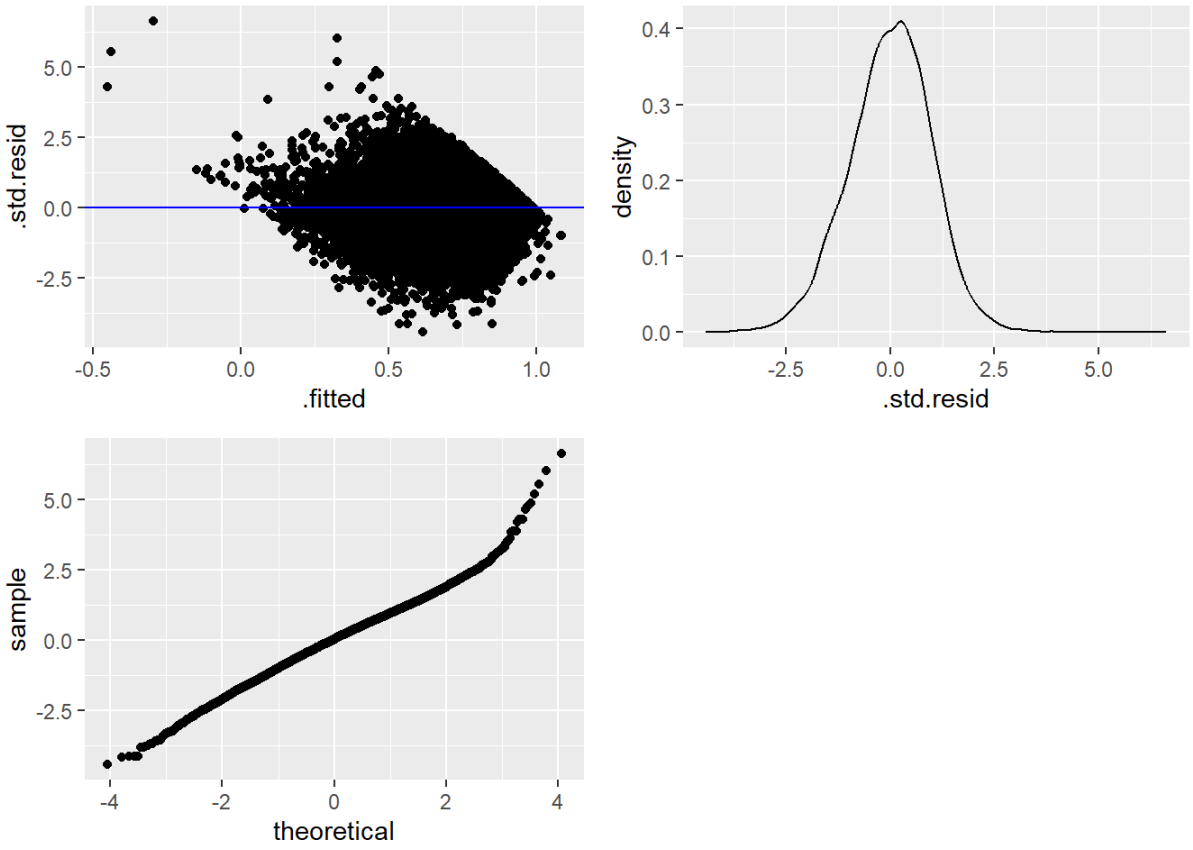
Konačno, F-statistika s p-vrijednosti manjom od 2.2×10^{-16} nam govori da je model statistički značajan u objašnjavanju izlazne varijable.

Sada ćemo 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.

U nastavku ćemo iskoristit metodu *augment* iz paketa *broom* koja na osnovu prosljeđenog prediktivnog modela vraća originalni podatkovni okvir korišten za stvaranje modela, ali proširen sa nizom korisnih stupaca:

- `.fitted` - Predikcije dobivene primjenom modela
- `.se.fit` - Standardna greška pojedine predikcije
- `.resid` - Iznos reziduala
- `.std.resid` - Reziduali standardizirani na interval [0,1]
- `.hat` - Mjera "ekstremnosti" ulazne varijable obzervacije
- `.cooksd` - Mjera "utjecajnosti" obzervacije na model

U nastavku prikazano je nekoliko grafova koji su redom: točkasti graf sa predikcijama na osi x i (standardiziranim) rezidualima na osi y, graf funkcije gustoće razdiobe standardiziranih reziduala, kvantil-kvantil graf standardiziranih reziduala reziduala.



Slika 3.23: Vizualizacije linearnog modela

Na temelju ovih vizualizacija te već prethodno navedenih svojstava linearnog modela možemo zaključiti da je linearni model koji smo dobili vrlo vjerojatno dobar dobar izbor za stvaranje predikcija.

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.

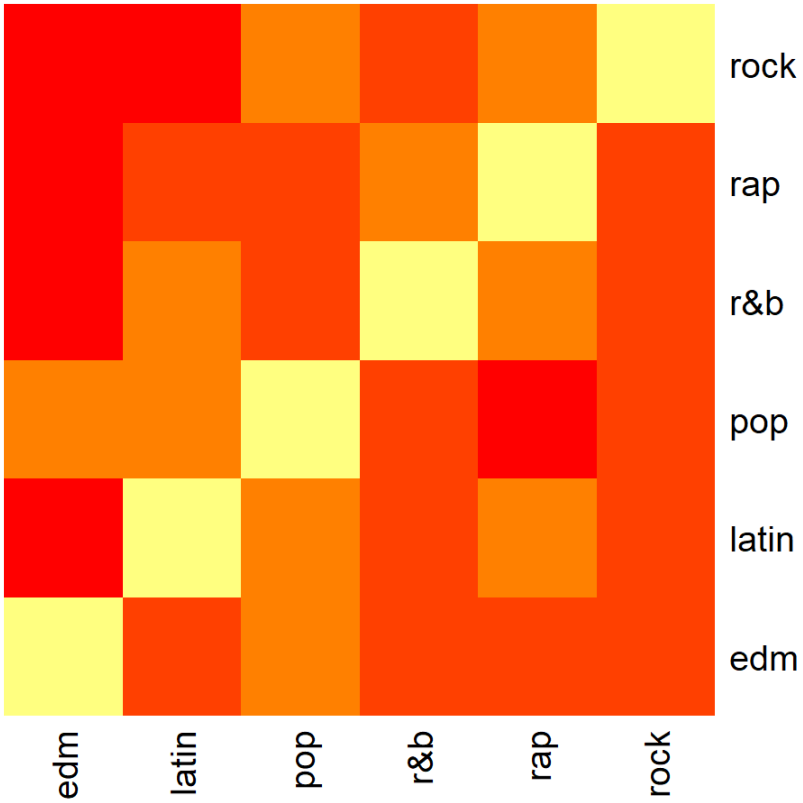
Kako je već navedeno, izlazna varijabla bit će *genre*, a ulazne sve numeričke varijable u okviru, tj. *track_popularity, danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, duration_ms*

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 3.24: Confusion matrix

Sada ćemo prikazati ovu matricu kao *heatmap*



Slika 3.25: 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

Kroz ovu eksploratornu analizu obradili su mnogo stvari, od zanimljivih vizualizacija do jednostavnih prediktivnih modela. Proučili smo razne ovisnosti varijabli pjesama, žanrova i slično.

Vidjeli smo kako u prosjeku Trevor Daniel ima najpopularnije pjesme, 2019. je prosječna popularnost pjesama bila najviša, pop i latin imaju najpopularniju glazbu, vidjeli smo donekle normalnu razdiobu trajanja pjesama i još mnogo toga.

Stvorili smo linearni model za predikciju energije pjesama te klasifikacijski model za predikciju žanrova pjesama. Uočili smo kako se energije može dosta dobro predvidjeti kroz ostale ulazne varijable, dok za žanrove ipak postoji više grešaka u predikcijama, međutim okvirno model radi dosta dobro.