Projekt OSP - Spotify Songs

0036539445 Filip Buhinićek, 0036542400 Dominik Zoričić, 0036539188 Nikola Botić

2024-01-25

UVOD

Ovaj skup podataka pruža opsežnu zbirku pjesama s nizom atributa, uključujući, ali ne ograničavajući se na, popularnost pjesme, žanr, pojedinosti o izvođaču i niz akustičnih značajki poput tempa, energije itd. Ovi atributi omogućuju višestruku analizu, od razumijevanja preferencija slušatelja do istraživanja glazbenih trendova tijekom vremena.

Naš projekt ima za cilj istražiti ovaj skup podataka kako bismo otkrili uvide i obrasce u svijetu glazbe na Spotifyju. Za analizu različitih aspekata podataka koristit ćemo niz statističkih tehnika i metoda vizualizacije podataka. Projekt je strukturiran za istraživanje skupa podataka na sveobuhvatan način, počevši od početnog prikupljanja podataka i predobrade, prelazeći kroz istraživačku analizu podataka.

Krajnji cilj ovog projekta je napraviti analizu podatkovnog skupa, koji će biti zanimljiv svim čitateljima a ne samo ljubiteljima glazbe.

OPĆENITO O PODATKOVNOM SKUPU

Paket spotifyr pruža podatke u sirovom, nefiltriranom obliku, dajući slobodu i fleksibilnost u oblikovanju analize. Ova neobrađenost znači da će početni koraci uključivati zadatke pretprocesiranja kao što su čišćenje nepotrebnih vrijednosti, rukovanje duplikatima i strukturiranje podataka na način koji najbolje odgovara analitičkim ciljevima.

Neki od zanimljivih atributa podatkovnog skupa kojeg ćemo eksplicitno prikazivati u našoj obradi su:

- Izvođač (Artist)
- Godina izdanja (Release Year)
- Žanr (Genre)
- Popularnost pjesme (Track Popularity)
- Dužina trajanja (Duration)

UČITAVANJE PODATAKA I PROCESIRANJE

Učitvanje podataka i odbacivanje nepotrebnih stupaca koji se neće koristiti u daljnjoj analizi

Budući da smo odlučili provoditi analizu podataka za koju nam je potrebno samo određeni skup stupaca, odbacujemo nekorištene stupce tj. višak informacija.

Na početku, podaci se učitavaju iz CSV datoteke. Ovaj korak je izveden korištenjem funkcije read_csv unutar R-a, a zanimljivo je da je prilikom učitavanja onemogućena automatska detekcija tipova stupaca. Ovo omogućuje veću kontrolu nad obradom podataka u kasnijim fazama, ali i veću odgovornost samog programera.

Nakon učitavanja, uslijedila je selekcija određenih stupaca iz skupa podataka. Ova operacija je ključna jer omogućuje fokusiranje samo na relevantne i korištene podatke. Stupci koji su zadržani uključuju identifikacijski broj pjesme, ime pjesme, izvođača, popularnost, datum izdanja albuma, žanr playliste i trajanje u milisekundama. Takav odabir stupaca osigurava da su sve bitne informacije zadržane za daljnju analizu.

Dalje, stupci su preimenovani u nazive koji su intuitivniji i lakši za razumijevanje. Na primjer, track_id postaje TrackID, a track_name postaje TrackName.

Jedna od bitnih pretvorbi bila je pretvorba trajanja pjesama iz milisekundi u minute. Ovaj korak je primjer kako podatke možemo prilagoditi tako da budu u formatu koji je više usklađen s uobičajenim načinom percipiranja trajanja pjesama.

Stupci Artist i Genre su transformirani u faktore, što je uobičajena praksa u R-u za rad s kategorijskim podacima. Ovaj korak olakšava analizu i vizualizaciju podataka koji spadaju u određene kategorije.

Važan dio procesa bila je ekstrakcija godine iz datuma izdanja. Razvijena je interna funkcija extractYear koja je mogla obraditi različite formate datuma i izvući relevantnu godinu. Ovo je bilo važno jer formati datuma nisu bili konzistentni unutar podataka. Na kraju, iz skupa podataka su uklonjeni svi redovi gdje godina izdanja nije bila dostupna. Ovo je osiguralo da analiza uključuje samo one zapise gdje su svi relevantni podaci prisutni.

Kroz ovaj proces, sirovi podaci su transformirani u oblik koji je pogodniji za detaljnu analizu. Ovaj postupak čišćenja i transformacije je ključan korak u pripremi podataka, jer osigurava da analiza koja slijedi bude temeljena na preciznim, relevantnim i dobro strukturiranim informacijama.

Slijedi primjer kako podaci izgledaju nakon uvodne transformacije:

```
glimpse(data)
```

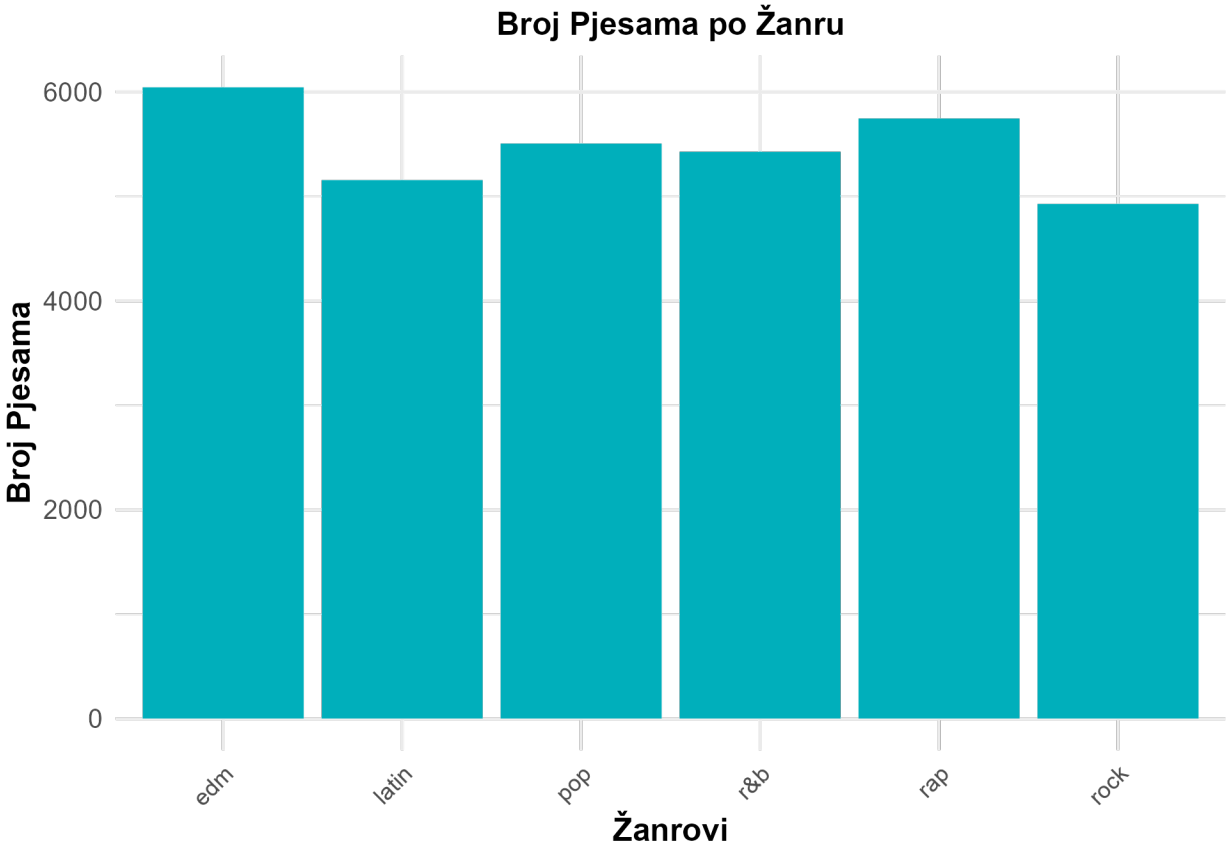
```
## Rows: 32,802
## Columns: 7
## $ TrackID     <chr> "6f807x0ima9a1j3VPbc7VN", "0r7CVbZTWZgbTCYdfa2P31", "1z1Hg~
## $ TrackName   <chr> "I Don't Care (with Justin Bieber) - Loud Luxury Remix", "~
## $ Artist      <fct> "Ed Sheeran", "Maroon 5", "Zara Larsson", "The Chainsmoker~
## $ Popularity  <dbl> 66, 67, 70, 60, 69, 67, 62, 69, 68, 67, 58, 67, 67, 68, 63~
## $ ReleaseYear <dbl> 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2019~
## $ Genre       <fct> pop, pop, pop, pop, pop, pop, pop, pop, pop, pop, pop, pop~
## $ DurationMin <dbl> 3.25, 2.71, 2.94, 2.82, 3.15, 2.72, 3.13, 3.46, 3.22, 4.22~
```

```
head(data)
```

```
## # A tibble: 6 x 7
##   TrackID              TrackName Artist Popularity ReleaseYear Genre DurationMin
##   <chr>                <chr>     <fct>       <dbl>       <dbl> <fct>       <dbl>
## 1 6f807x0ima9a1j3VPbc~ I Don't ~ Ed Sh~         66        2019 pop          3.25
## 2 0r7CVbZTWZgbTCYdfa2~ Memories~ Maroo~         67        2019 pop          2.71
## 3 1z1Hg7Vb0AhHDiEmnDE~ All the ~ Zara ~         70        2019 pop          2.94
## 4 75FpbthrwQmzHlBJLuG~ Call You~ The C~         60        2019 pop          2.82
## 5 1e8PAfcKUYoKkxPhrHq~ Someone ~ Lewis~         69        2019 pop          3.15
## 6 7fvUMiyapMsRRxr07cU~ Beautifu~ Ed Sh~         67        2019 pop          2.72
```

ANALIZA PROCESIRANOG PODATKOVNOG SKUPA

Popularnost žanrova na Spotifyu

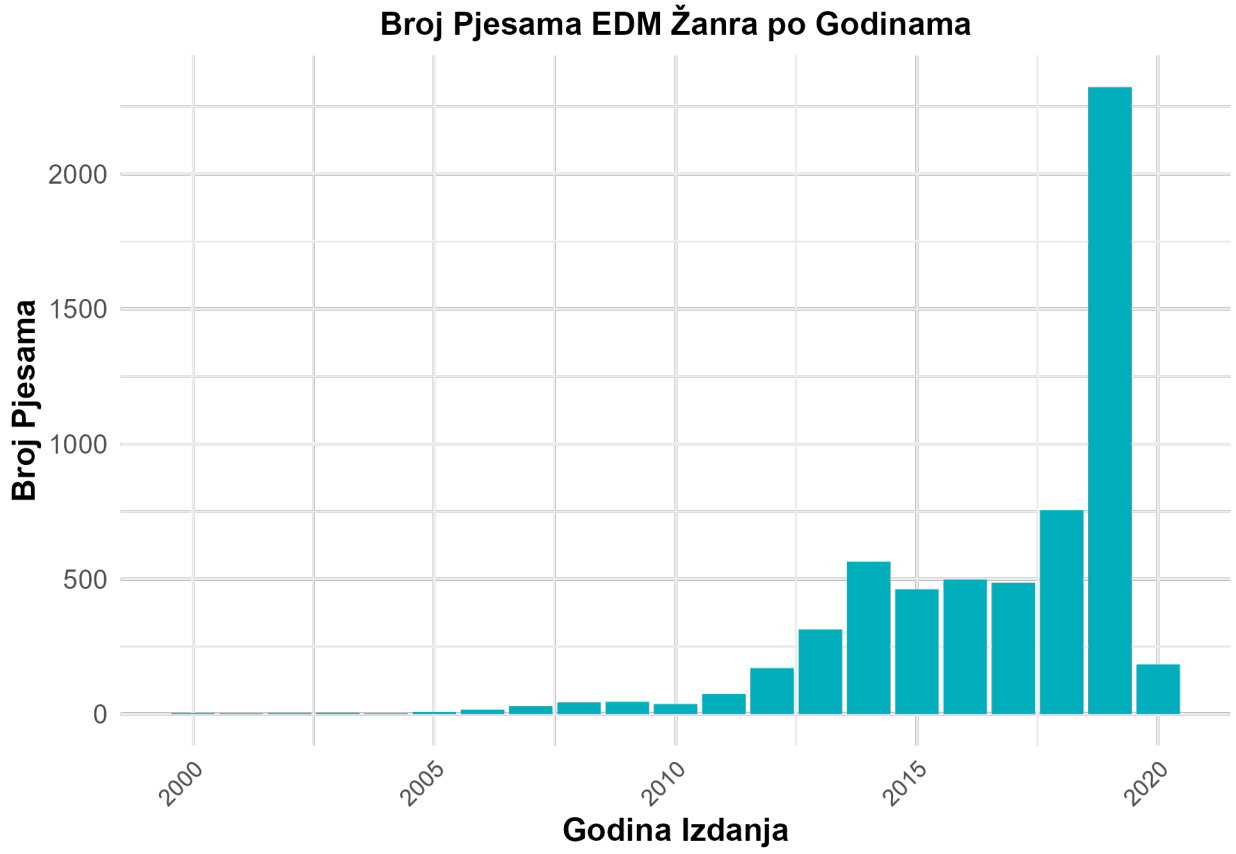


Analiza grafa

Graf "Broj Pjesama po Žanru" vizualno prikazuje raspodjelu glazbenih žanrova Spotify datasetu. Svaki stupac na grafu predstavlja određeni glazbeni žanr, dok visina stupca odražava broj pjesama u tom žanru.

Ovaj graf pruža početni uvid u glazbene preferencije i raznolikost žanrova među pjesmama na Spotifyju. Dominantni žanrovi mogu ukazivati na popularnost određenih glazbenih stilova među korisnicima platforme.

Budući da je najpopulanriji edm (Electro dance music), prikazat ćemo njegovu popularnost kroz godine

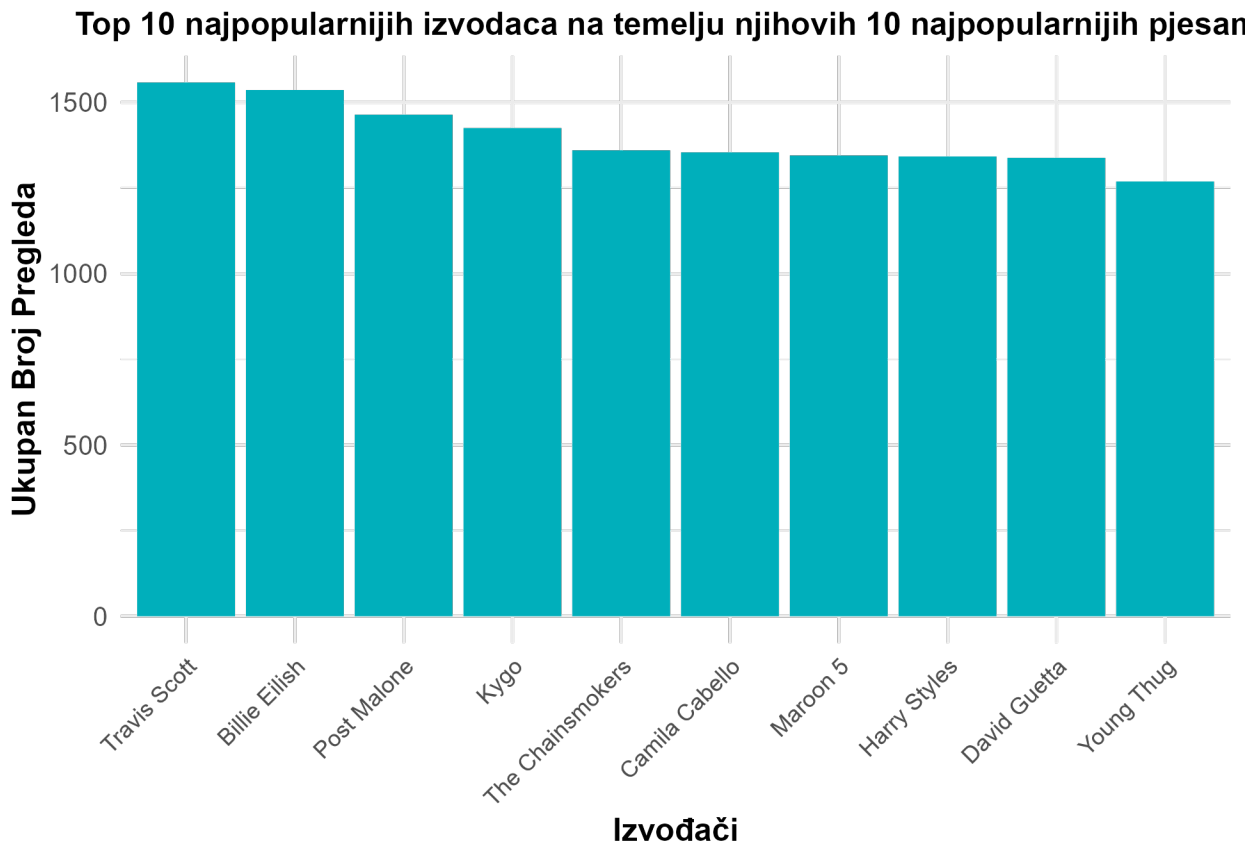


Analiza grafa

Graf "Broj Pjesama EDM Žanra po Godinama" vizualno prikazuje raspodjelu glazbenog žanra EDM (Electro dance music) nad Spotify datasetom. Svaki stupac na grafu predstavlja određenu godinu, dok visina stupca odražava broj pjesama u toj godini.

Iz priloženog grafa može se zaključiti kako je edm (Electro dance music) postao sve više popularan tek od 2012 godine, te od tada sve više preuzima mainstream glazbe na Spotifyu.

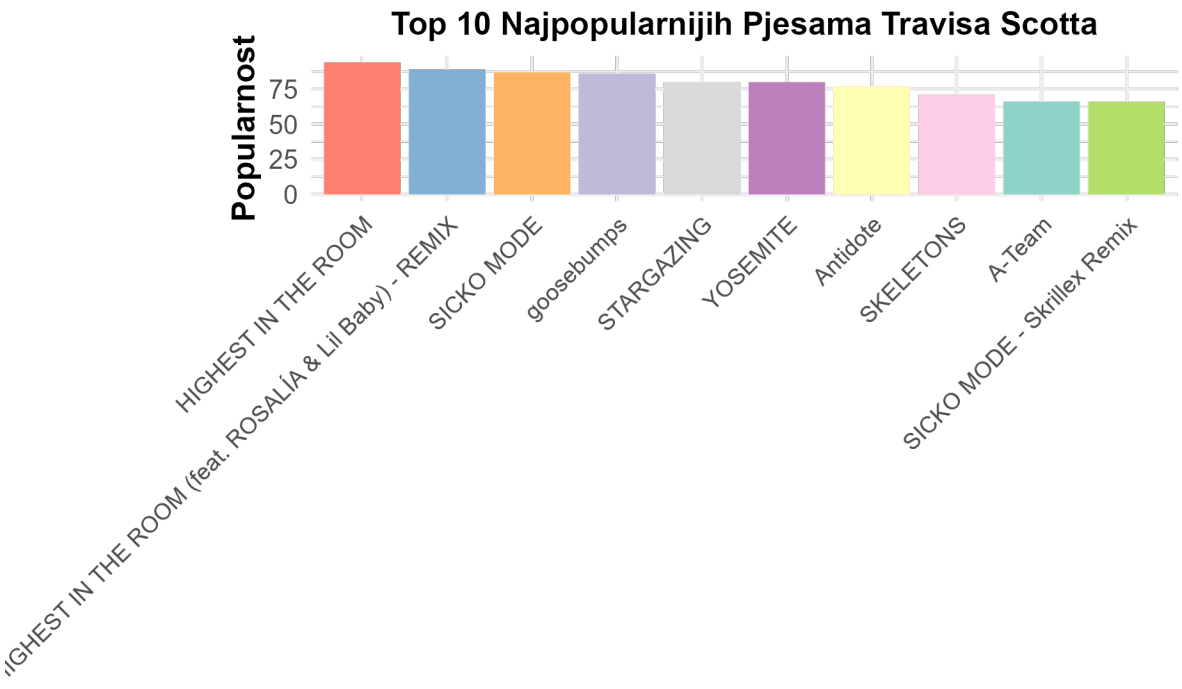
Najpopularniji izvodaci



Analiza grafa

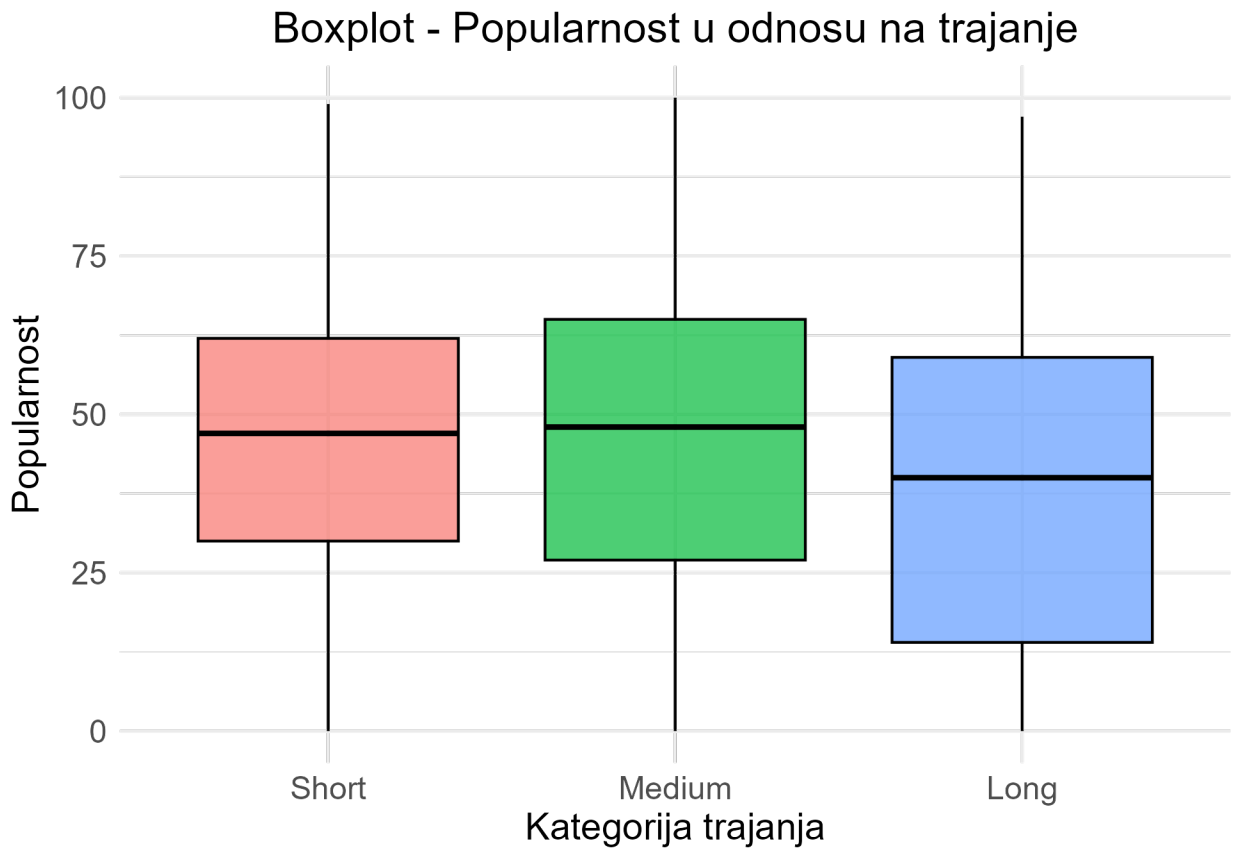
Grafički prikaz ukupnog broja pregleda temeljem 10 najpopularnijih pjesama svakog izvođača pruža zanimljiv uvid u popularnost umjetnika na Spotify platformi. Analizirajući graf, uočavamo da je Travis Scott apsolutni lider među izvođačima, slijedi ga Billie Eilish, Post Malone te ostali.

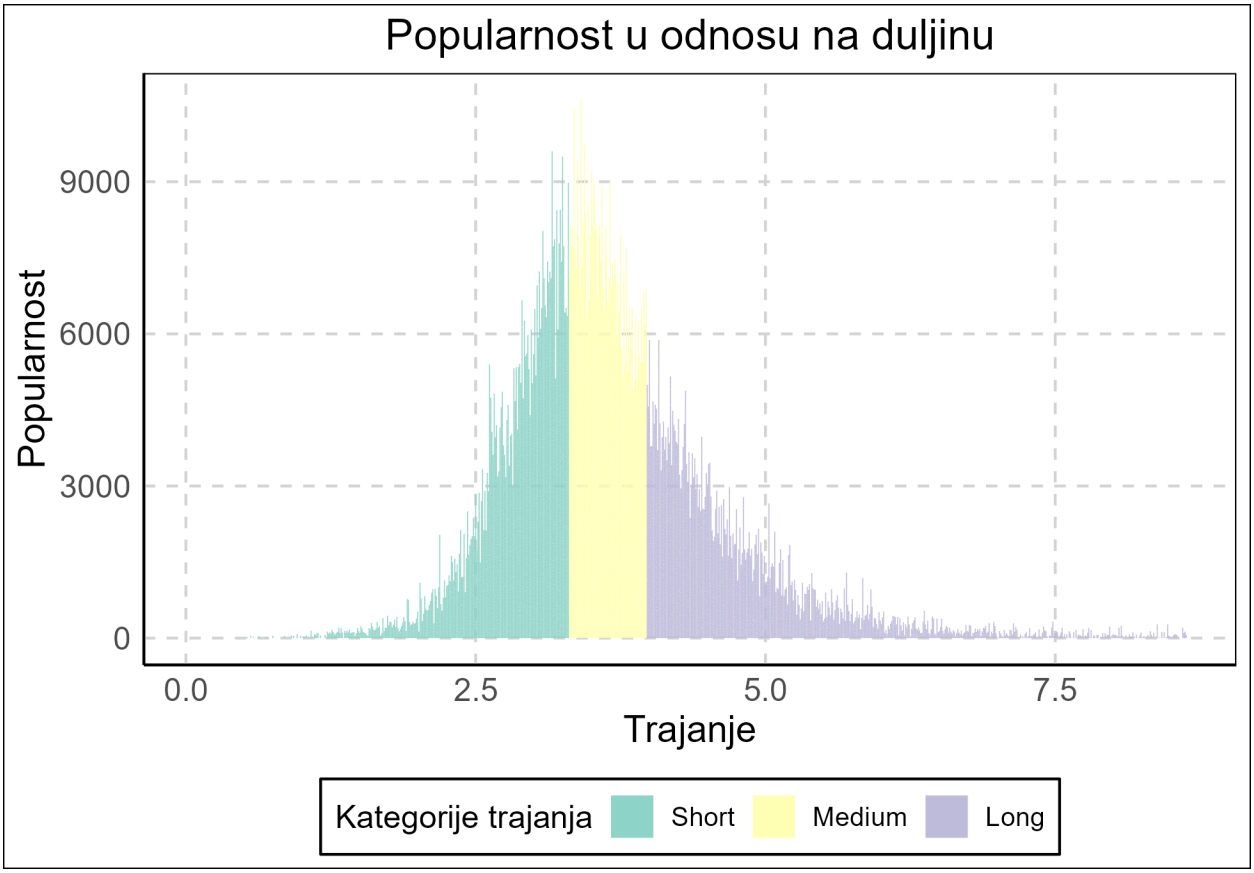
Najpopularnije pjesme Travis Scotta



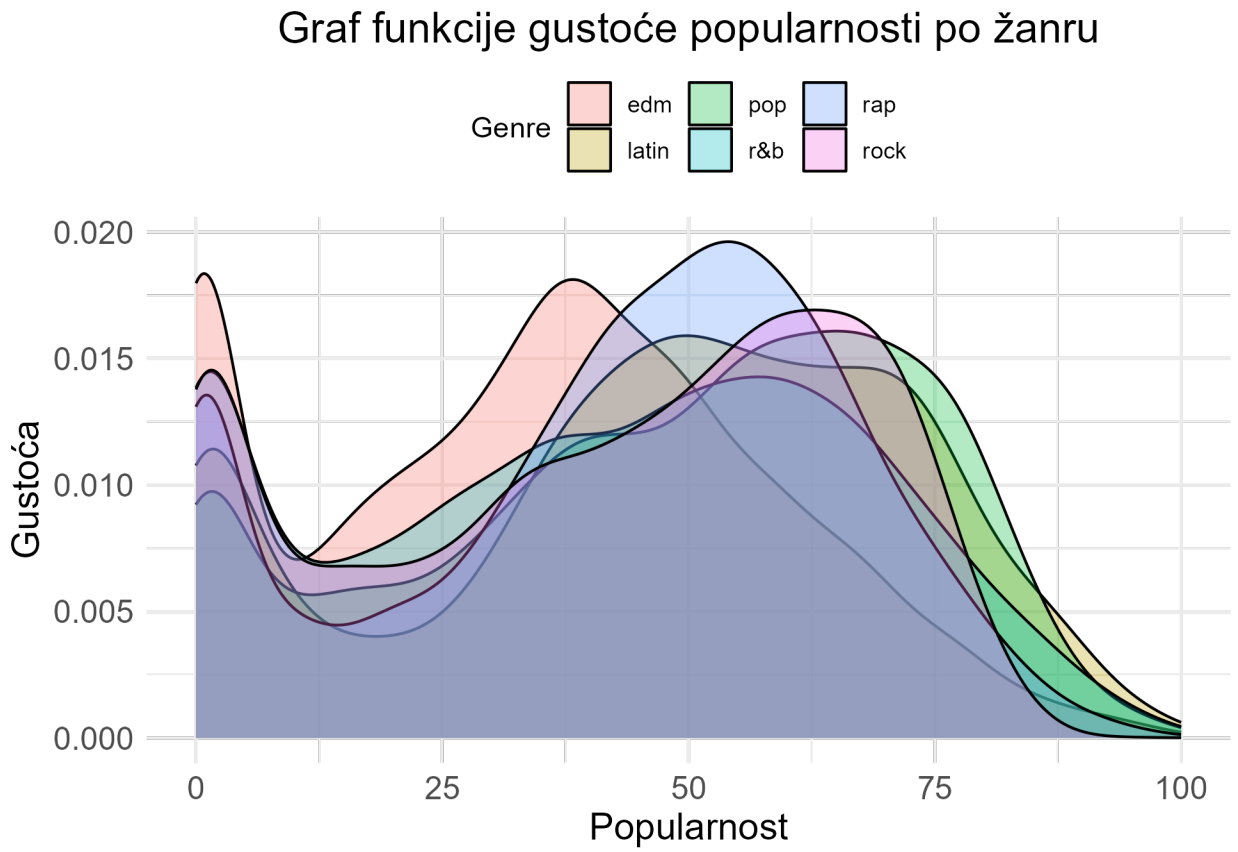
Analiza grafa

Ova analiza naglašava impresivnu popularnost Travisa Scotta na Spotify platformi, pri čemu se ističe njegov utjecaj kroz deset najpopularnijih pjesama. "Highest in the Room" dominira među njima, zatim slijedi istoimeni Remix, "Sicko mode" i ostali. Presudila je njegova konstantnost jer nije samo bljesnuo s jednom pjesmom, već u svima drži jako visoku kvalitetu izvedbe.


Boxplot - Popularnost u odnosu na trajanje
Popularnost
100
75
50
25
0
Short
Medium
Long
Kategorija trajanja



Analiza grafa Podijelili smo trajanje pjesama u 3 jednake kategorije, Short, Medium, Long. Graf "Boxplot - Popularnost u odnosu na trajanje" prikazuje odnose Q1, Q2 i Q3 metrika koje se odnose na popularnost pjesama pripadajuće kategorije. Iz priloženog se vidi da je medijan popularnosti za duge pjesme znatno niži nego što možemo reći sa kratke i srednje. Tu tvrdnju samo podupire graf "Popularnost u odnosu na duljinu" koji prikazuje da su srednje pjesme najpopularnije, na drugom mjestu su kratke, a daleko ispod njih se nalaze duge pjesme



Analiza grafa Iz grafa "Graf funkcije gustoće popularnosti po žanru" možemo zaključiti da većinom funkcije gustoća svih žanrova donekle prate normalnu razdiobu, stime da sve imaju veliku devijaciju u lijevom repu. Naime Žanr "edm" ima najviše malo popularnih pjesama, dok "pop" najmanje. Najviše srednje ocjenjenih pjesama pripadaju žanru "rap", dok u najbolje ocjenjenim pjesmama prednjači žanr "latin"

ZAKLJUČAK

Na samom početku smo napravili potrebne pretvorbe podataka, i pretvorbe tipova podataka u one koje omogućavaju lakšu analizu. Nakon toga nam je ostao podatknovni skup na kojem je puno laške izvoditi analize koje nas zanimaju. Iz grafa "Broj Pjesama po Žanru" vidimo da svakom žanru pripadaju od okvirno 5000 do 6000 pjesama. Za bližu analizu broja novih pjesama kroz godine smo uzeli žanr "edm" i to smo predstavili u grafu "Broj Pjesama EDM Žanra po Godinama". Iz grafa možemo isčitati da je žanr bio relativno nepopularan do 2015. godine, a posebice do 2010. godine. Nakon 2015. godine žanr uživa puno veću popularnost i dostiže brojke od čak više od 2000 novih pjesama godišnje netom prije 2020. godine. Zatim smo obradili popularnost samih izvođača i njihovih pjesama. Grafički prikaz ukupnog broja pregleda temeljem 10 najpopularnijih pjesama svakog izvođača pruža intrigantan uvid u popularnost umjetnika na Spotify platformi. Analizom grafa ističe se da je Travis Scott apsolutni lider među izvođačima. Ova analiza naglašava impresivnu popularnost Travisa Scotta na Spotify platformi, ističući njegov utjecaj kroz deset najpopularnijih pjesama. Posebno se ističe "Highest in the Room" kao dominantna pjesma. Ključna je Travisova konstantnost, ne samo s jednom pjesmom, već s održavanjem visoke kvalitete izvedbe u svim pjesmama.Nakon toga smo se odlučili analizirati odnos popularnosti i trajanja pjesme tako što smo trajanje podjelili u 3 podjednake kategorije.Iz grafova "Boxplot - Popularnost u odnosu na trajanje" i "Popularnost u odnosu na duljinu" vidimo kako su dulje pjesme u glavnom lošije ocjenjene u odnosu na pjesme koje su kratke ili srendnje, koje prednjače u popularnosti. Za kraj odlučili smo se na analizu funckije gustoće popularnosti pojedinih žanrova što smo vizualizirali u grafu "Graf funkcije gustoće popularnosti po žanru". Vidimo da po broju nisko ocjenjenih pjesama pobjedu odnosi žanr "edm" dok žanr "rap" prednjači po srednje ocjenjenim pjesmama, a u najboljim pjesmama se ističe žanr "latin".