OSP projekt - Spotify

36551971 Mark Sarić, 36552071 Kristijan Šagovac

2025-01-30

```
songs <- read_csv("Datasets/spotify_songs.csv")
```

```
## Rows: 32833 Columns: 23
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (10): track_id, track_name, track_artist, track_album_id, track_album_na...
## dbl (13): track_popularity, danceability, energy, key, loudness, mode, speec...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(songs)
glimpse(songs)
```

```
## # A tibble: 6 x 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>
## 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~
```

Koja je razlika u hip hop-u i hip pop-u, osim jednog slova?

Idemo dublje analizirati u čemu se razlikuju, ako se uopće razlikuju.

```
songs_g <- songs %>% filter(playlist_subgenre %in% c("hip hop","hip pop"))
songs_hop <- songs %>% filter(playlist_subgenre %in% c("hip hop"))
songs_pop <- songs %>% filter(playlist_subgenre %in% c("hip pop"))
```

Pogledajmo izvođače koji se najviše pojavljuju iz oba skupa.

```
table(songs_hop$track_artist) %>% sort(decreasing=T) %>% head(7)
table(songs_pop$track_artist) %>% sort(decreasing=T) %>% head(7)
```

```
##
##            Logic Sidhu Moose Wala            Drake           DIVINE
##               62               15               11               10
##      Post Malone       Young Thug           Future
##                9                8                7
##
## Pickin' On Series       Post Malone            Drake           Halsey
##                12                11               10                9
##            Khalid      Selena Gomez            Usher
##                 9                 9                9
```

Vidimo da se neki izvođači pojavljuju u obje kategorije poput Post Malonea i Drakea. Pogledajmo koliko izvođača se pojavljuje u oba skupa.

```
cat(paste("Number of artists in both genres:", length(intersect(songs_hop$track_artist,songs_pop$track_a
cat(paste("Number of artists in hop:", length(unique(songs_hop$track_artist)), "\n"))
cat(paste("Number of artists in pop:", length(unique(songs_pop$track_artist)), "\n"))
```

```
## Number of artists in both genres: 80
## Number of artists in hop: 779
## Number of artists in pop: 797
```

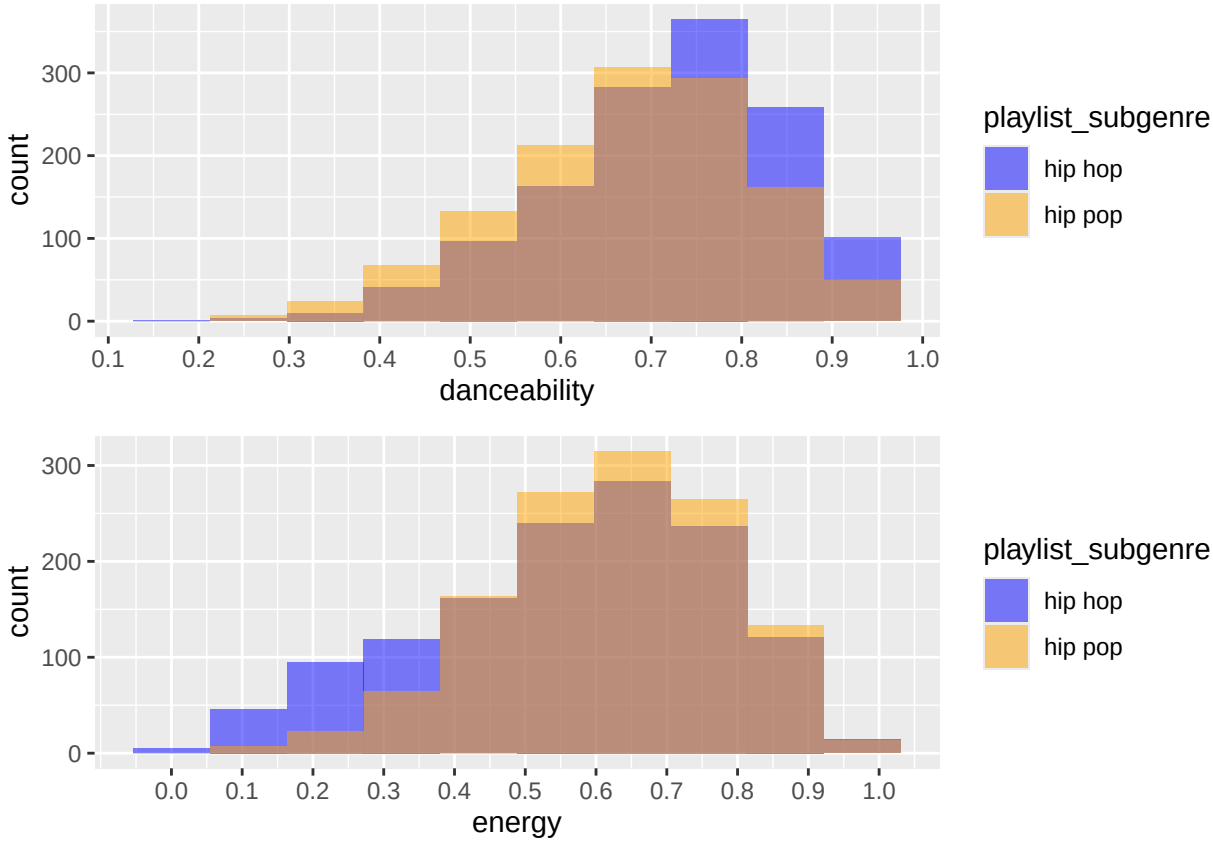
S obzirom da se u obje kategorije nalazi 80 izvođača, a ni u jednoj nema iznad 800, to je indikacija da su te kategorije slične jer se izvođači uglavnom drže svojeg stila pjevanja. Sljedeće ćemo usporediti plesnost

(engl. danceability) i energiju kategorija. Plesnost označava koliko je pjesma prikladna za plesanje, gdje vrijednost 1.0 označava da je najplesnija, a 0.0 obrnuto. Energija se također mjeri u rasponu od 0.0 do 1.0 i predstavlja perceptivnu mjeru intenziteta i aktivnosti. Nacrtat ćemo histograme tih varijabli te vidjeti koliko se preklapaju.

```
g1 <- ggplot(songs_g,aes(x=danceability,fill = playlist_subgenre)) + geom_histogram(bins=10,alpha=0.5,p

g2 <- ggplot(songs_g,aes(x=energy,fill = playlist_subgenre)) + geom_histogram(bins=10,alpha=0.5,positio

grid.arrange(g1,g2)
```



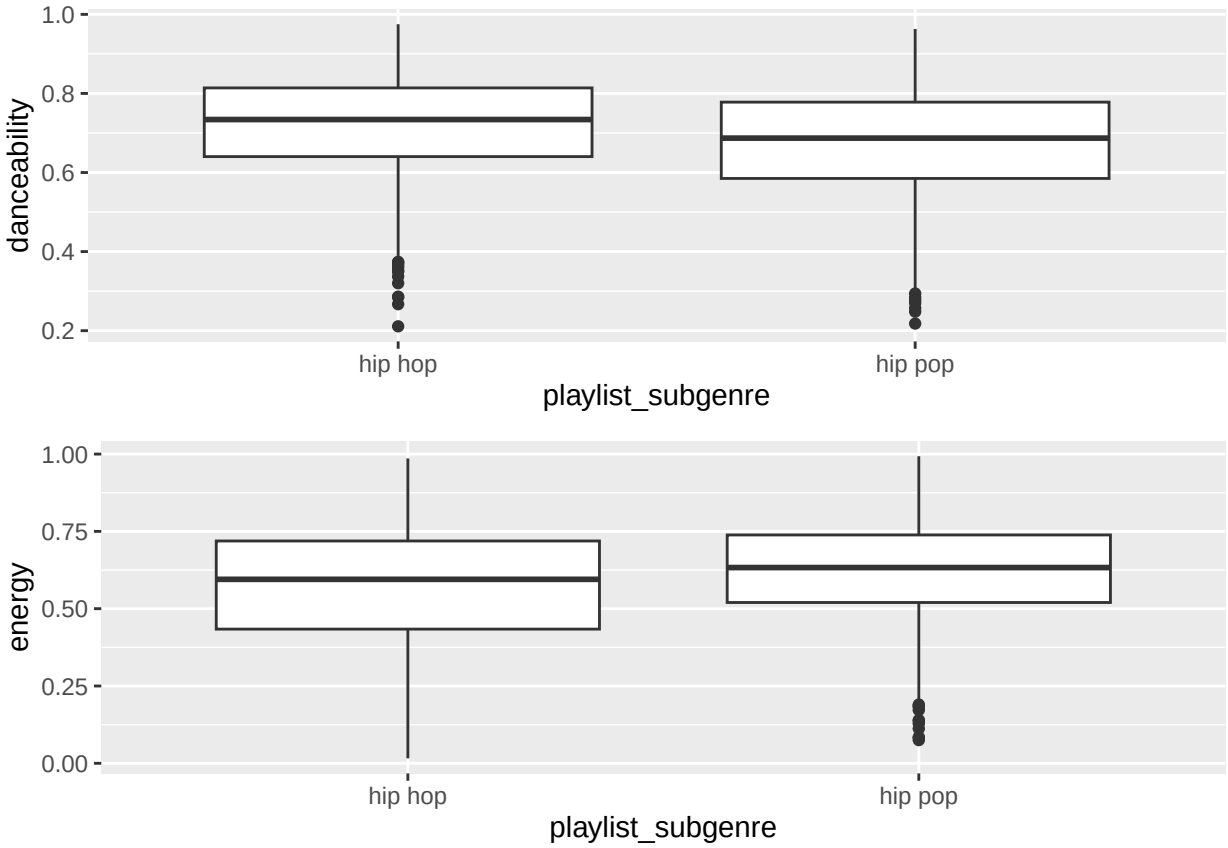
Vidimo da se obje kategorije dosta preklapaju u obje varijable, s tim da hip hop ima više pjesama s većom plesnošću, a hip pop ima više pjesama s većom energijom. Pogledajmo prosjeke.

```
songs_g %>% group_by(playlist_subgenre) %>% summarise(meanDance=mean(danceability),meanEnergy=mean(energ
```

```
## # A tibble: 2 x 3
##   playlist_subgenre meanDance meanEnergy
##   <chr>                <dbl>      <dbl>
## 1 hip hop              0.720      0.566
## 2 hip pop              0.675      0.622
```

U prosjeku se razine plesnosti ne razlikuju toliko, dok se razine energije razlikuju malo više. Još ćemo pogledati boxplot.

```
g1 <- ggplot(songs_g,aes(x=playlist_subgenre,y=danceability)) + geom_boxplot()
g2 <- ggplot(songs_g,aes(x=playlist_subgenre,y=energy)) + geom_boxplot()
grid.arrange(g1,g2)
```

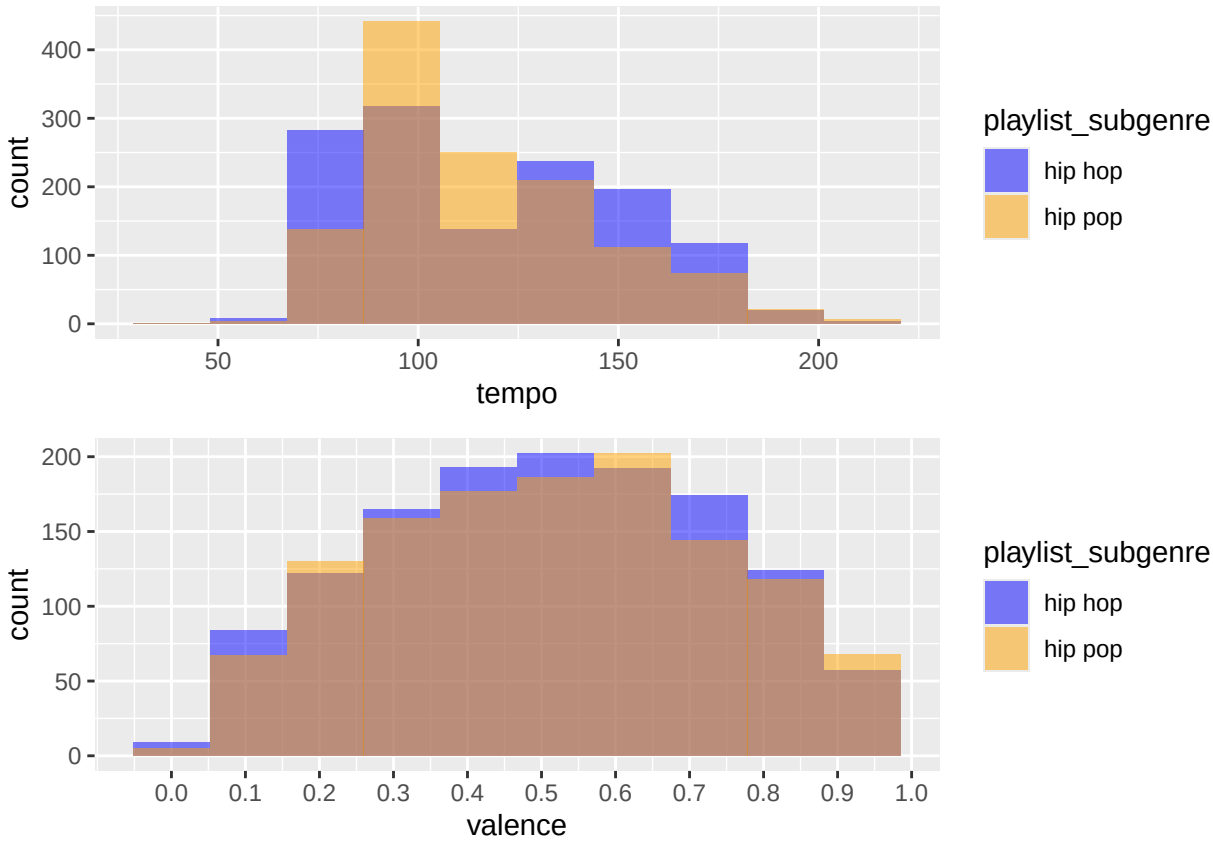


Na temelju analize možemo reći da je hip hop malo prikladniji za ples, dok je hip pop nešto življi. Nastavimo s analizom. Gledat ćemo tempo i ugođaj (engl. valence). Tempo je varijabla koja označava brzinu ritma u otkucajima po sekundi, a ugođaj koliko pozitivnosti pjesma prenosi u rasponu od 0.0 do 1.0.

```
g1 <- ggplot(songs_g,aes(x=tempo,fill = playlist_subgenre)) + geom_histogram(bins=10,alpha=0.5,position

g2 <- ggplot(songs_g,aes(x=valence,fill = playlist_subgenre)) + geom_histogram(bins=10,alpha=0.5,positi

grid.arrange(g1,g2)
```



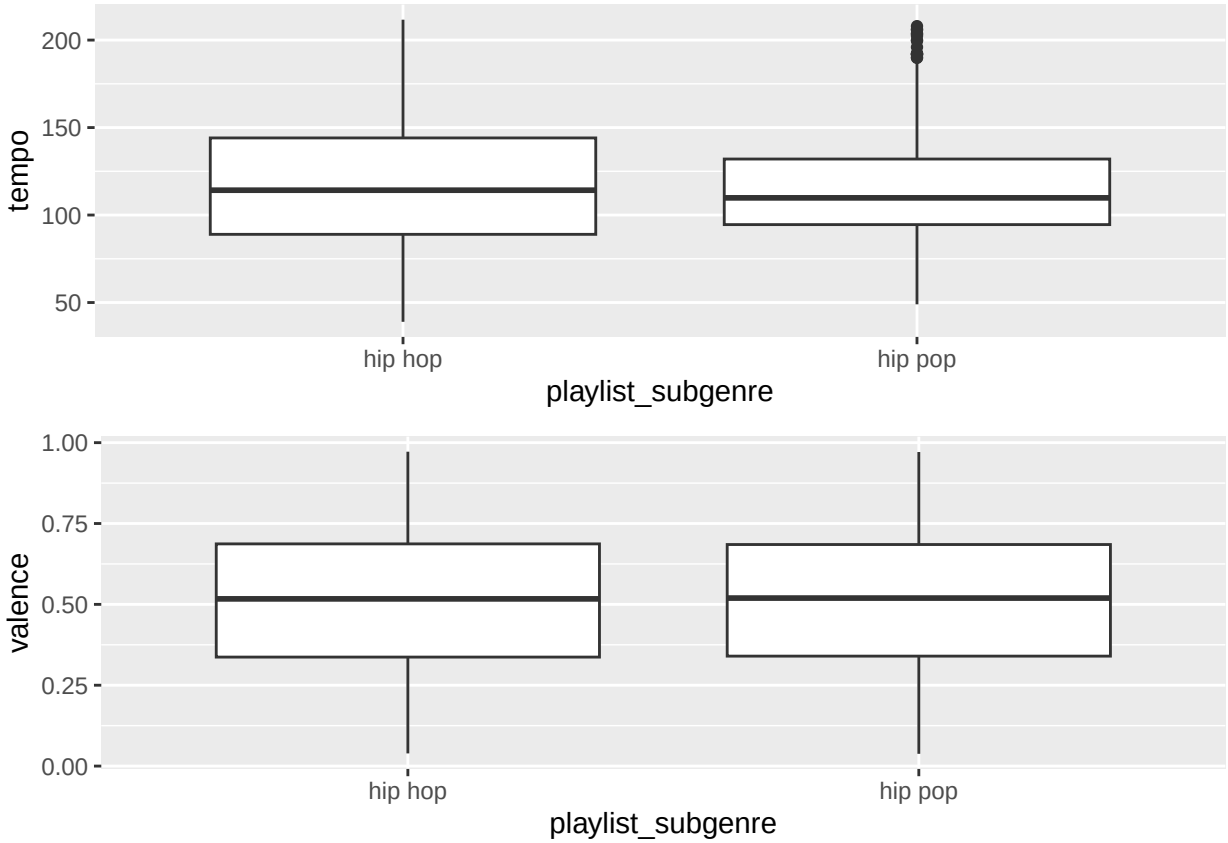
Kod ugođaja gotovo da nema razlike, a kod tempa izgleda kao da hip hop ima sporiji tempo, jer ima više pjesama u lijevom dijelu razdiobe, no ima i više pjesama od desnom dijelu. Za ugođaj već možemo reći da je vrlo vjerojatno identičan, ali za tempo ne. Stoga ćemo pogledati prosjek i boxplotove.

```
songs_g %>% group_by(playlist_subgenre) %>% summarise(meanTempo=mean(tempo),meanValence=mean(valence))
```

```
## # A tibble: 2 x 3
##   playlist_subgenre meanTempo meanValence
##   <chr>                <dbl>       <dbl>
## 1 hip hop               118.       0.509
## 2 hip pop               116.       0.514
```

Prosjeci ugođaja su vrlo slični kao i očekivano. Prosjeci tempa su vrlo slični, što ima smisla. Hip hop ima sličan prosjek jer ako većina pjesama ima spor ili brzi tempo, onda u prosjeku imaju srednji. Zbog toga ćemo pogledati boxplot.

```
g1 <- ggplot(songs_g,aes(x=playlist_subgenre,y=tempo)) + geom_boxplot()
g2 <- ggplot(songs_g,aes(x=playlist_subgenre,y=valence)) + geom_boxplot()
grid.arrange(g1,g2)
```

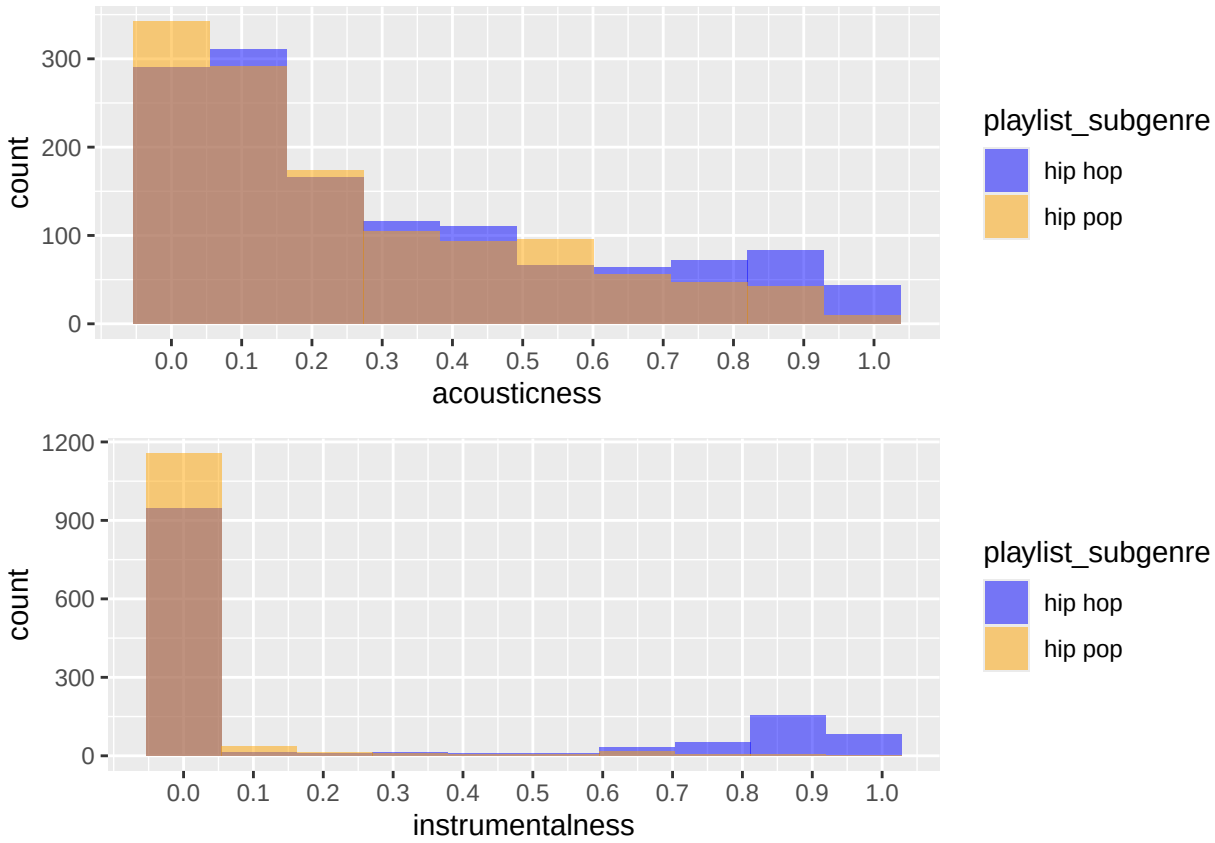


Većina hip pop pjesama ima tempo oko 110 otkucaja po minuti te je varijabilnost manja, dok je kod hip hopa varijabilnost znatno veća. Kod ugođaja su boxplotovi obje kategorije gotovo identični. Trentuno nismo našli značajne razlike između kategorija, stoga nastavljamo dalje, s varijablama: akustičnost i instrumentalnost. Akustičnost govori sadrži li pjesma uglavnom akustične elemente ili je dominirana elektronskim i sintetičkim zvukovima, a instrumentalnost označava vjerojatnost da pjesma nema vokale.

```
g1 <- ggplot(songs_g,aes(x=acousticness,fill = playlist_subgenre)) + geom_histogram(bins=10,alpha=0.5,p

g2 <- ggplot(songs_g,aes(x=instrumentalness,fill = playlist_subgenre)) + geom_histogram(bins=10,alpha=0

grid.arrange(g1,g2)
```



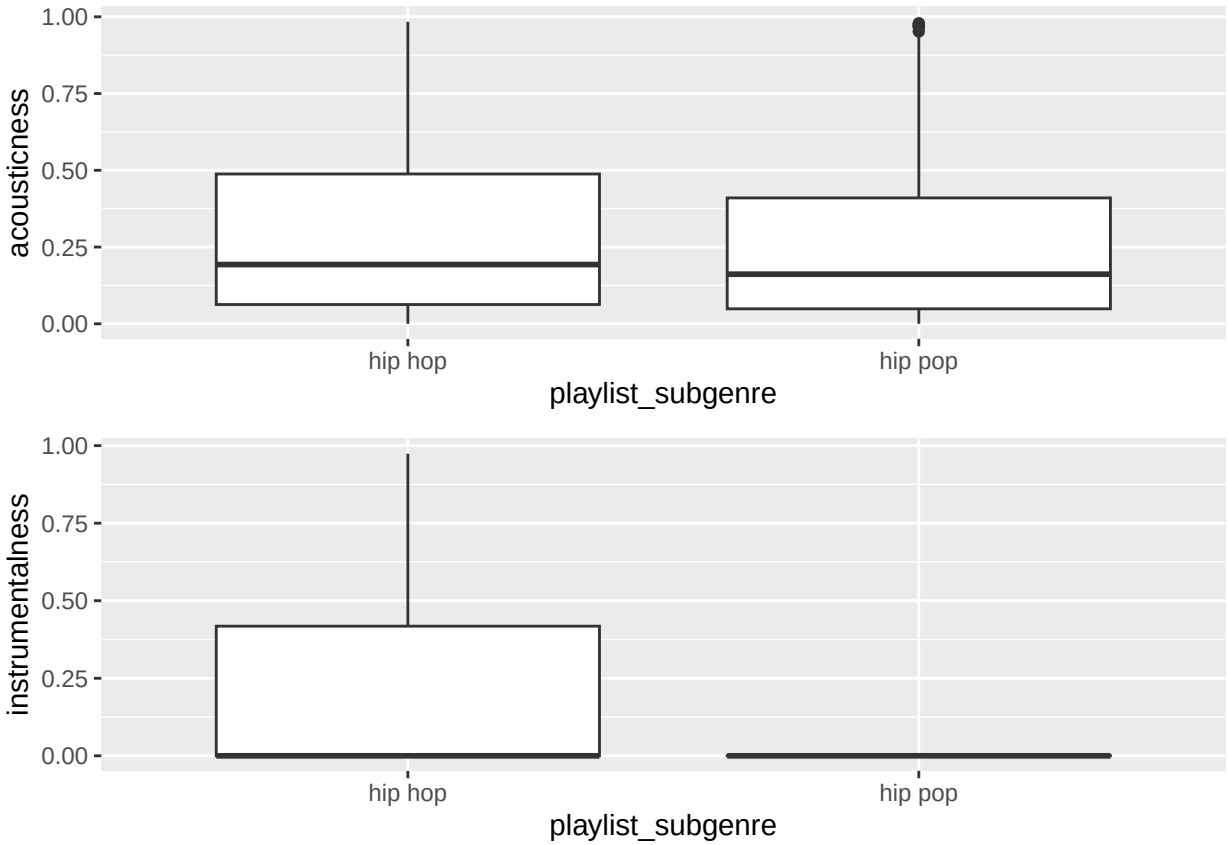
Hip hop ima malo više pjesama koje sadrže akustične elemente nego hip pop. Kod insturmentalnosti vidimo da obje kategorije uglavnom imaju pjesme koje gotovo sigurno sadrže vokale, no hip hop ima i solidan broj pjesama koje su insturmentalne. Zbog čudne razdiobe instrumentalnosti, napravit ćemo prosjek samo za akustičnost.

```
songs_g %>% group_by(playlist_subgenre) %>% summarise(meanAcousticness=mean(acousticness))
```

```
## # A tibble: 2 x 2
##   playlist_subgenre meanAcousticness
##   <chr>                        <dbl>
## 1 hip hop                      0.308
## 2 hip pop                      0.255
```

Kao što smo i pretpostavili, hip hop pjesme u prosjeku imaju manje elektronskih zvukova. Pogledajmo boxplot.

```
g1 <- ggplot(songs_g,aes(x=playlist_subgenre,y=acousticness)) + geom_boxplot()
g2 <- ggplot(songs_g,aes(x=playlist_subgenre,y=instrumentalness)) + geom_boxplot(outlier.shape = NA)
grid.arrange(g1,g2)
```



Vidimo da što se instrumentalnosti tiče, hip hop ima veću varijabilnost, dok hip pop pjesme uglavnom nisu instrumentalne. Maknuli smo stršeće vrijednosti kod instrumentalnosti jer s njima izgleda kao da je puno hip pop pjesama insturmentalno, no to nije istina. Pjesme s instrumentalnošću iznad 0.5 se smatraju instrumentalnim pjesmama.

```
songs_pop %>% filter(instrumentalness>0.5) %>% nrow %>% cat("Broj hip pop pjesama s instrumentalnošću v
cat("Broj ukupno hip pop pjesmama: ",nrow(songs_pop),"\n")

songs_hop %>% filter(instrumentalness>0.5) %>% nrow %>% cat("Broj hip hop pjesama s instrumentalnošću v
cat("Broj ukupno hip hop pjesmama: ",nrow(songs_hop))
```

```
## Broj hip pop pjesama s instrumentalnošću većom od 0.5: 37
## Broj ukupno hip pop pjesmama:  1256
## Broj hip hop pjesama s instrumentalnošću većom od 0.5: 325
## Broj ukupno hip hop pjesmama:  1322
```

Gotovi smo s analizom. Pregledajmo što smo se zaključili. Ima 80 pjevača koji se nalaze u obje kategorije, s tim da u jednoj ima 797, a u drugoj 779 pjevača. Što se tiče plesnosti i energije, postoje neke razlike, hip pop je nešto življi, a hip hop je nešto prikladniji za ples, no razlike nisu baš značajne. Hip hop je varijabilniji u tempu, dok hip pop pokazuje relativnu stabilnost. Što se tiče valence, gotovo su identične. Najveća razlika se vidi u instrumentalnosti. Skoro pa sve hip pop pjesme su neinstrumentalne, dok postoji solidan dio hip hop pjesama koje su instrumentalne. Hip hop pjesme su nešto više akustične nego hip pop, no također nije značajno. Jedan od autora će zaključiti da nema nekakve značajne razlike između ove dvije kategorije te da bi po njemu, kao pravi minimalist, spojio ih u istu stvar.

Tko ima pjesme sretnijeg ugođaja, Bruno Mars ili Adele?

Najdraži pjevači jednom od autora su Bruno Mars i Adele. Želi općenito usporediti pjevače te saznati kojeg pjevača da pusti kad je sretan, a kojeg kad je tužan.

```
songs_both <- songs %>% filter(track_artist %in% c("Bruno Mars","Adele"))
songs_Bruno <- songs %>% filter(track_artist=="Bruno Mars")
songs_Adele <- songs %>% filter(track_artist=="Adele")
```

Pogledajmo prvo žanrove u kojima se pojavljuje Bruno, a u kojima Adele.

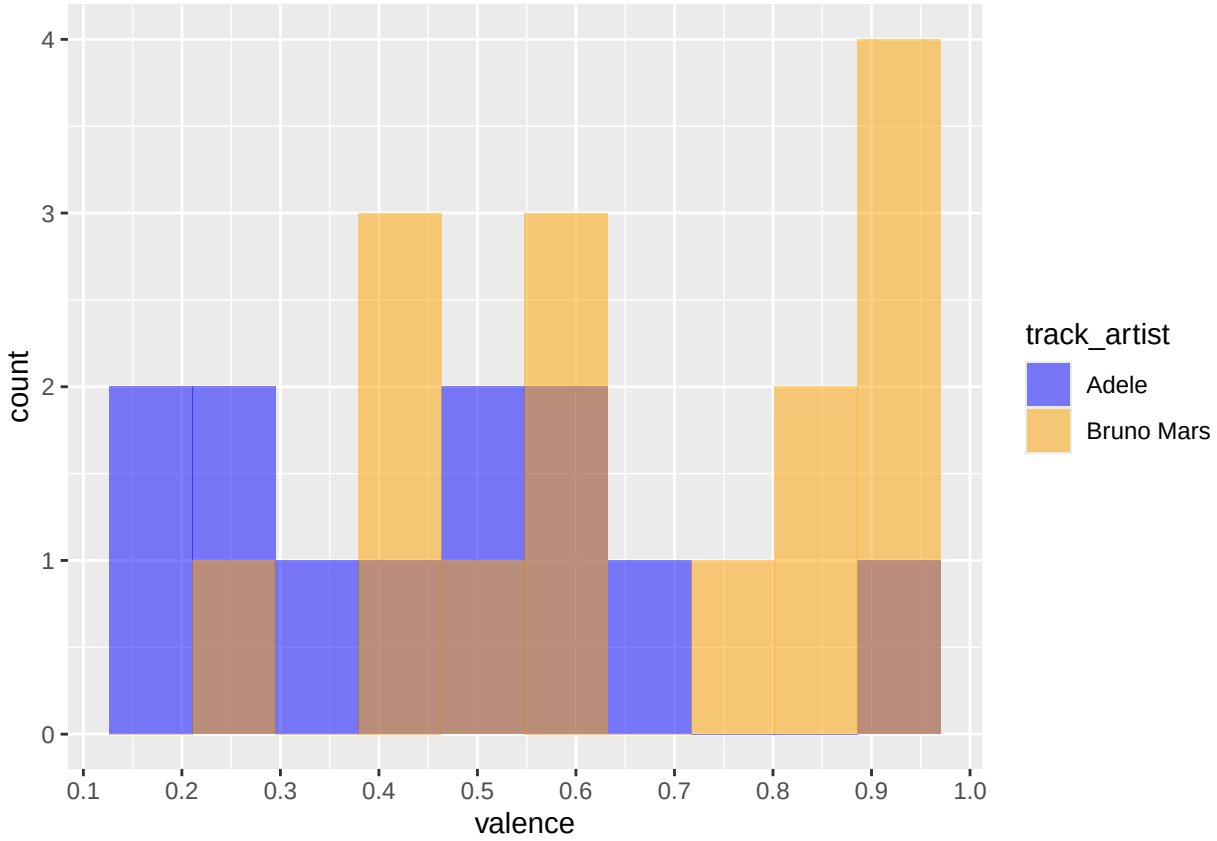
```
songs_Bruno %>% select(playlist_genre) %>% unique()
songs_Adele %>% select(playlist_genre) %>% unique()
```

```
## # A tibble: 4 x 1
##   playlist_genre
##   <chr>
## 1 pop
## 2 latin
## 3 r&b
## 4 edm
## # A tibble: 4 x 1
##   playlist_genre
##   <chr>
## 1 pop
## 2 latin
## 3 r&b
## 4 edm
```

Pojavljuju se u istim žanrovima, i to u nekim neočekivanim. Adele nema veze s latino i EDM-om, kao što Bruno nema veze s EDM-om. Pogledajmo kakve su njihove pjesme što se tiče ugođaja. Pretpostavka autora je da će Adele imati pjesme s tužnijim ugođajem, a Bruno sa sretnijim. S obzirom da se neke pjesme pojavljuju više puta pod različitim žanrovima, a manji je broj pjesama, izbacit ćemo duplikate.

```
songs_both_clean <- songs_both %>% distinct(track_name, .keep_all = TRUE)

ggplot(songs_both_clean,aes(x=valence,fill = track_artist)) + geom_histogram(bins=10,alpha=0.5,position
```



Kao što je bilo pretpostavljeno, Adele se povezuje s tužnim ugođajem, a Bruno sa sretnim. Pogledajmo koje pjesme su najtužnije, a koje najsretnije.

```
songs_both_clean %>% filter(track_artist=="Adele") %>% slice_min(order_by = valence,n = 3) %>% select(t
songs_both_clean %>% filter(track_artist=="Adele") %>% slice_max(order_by = valence,n = 3) %>% select(t

songs_both_clean %>% filter(track_artist=="Bruno Mars") %>% slice_min(order_by = valence,n = 3) %>% sel
songs_both_clean %>% filter(track_artist=="Bruno Mars") %>% slice_max(order_by = valence,n = 3) %>% sel
```

```
## # A tibble: 3 x 2
##   track_name              valence
##   <chr>                     <dbl>
## 1 Melt My Heart to Stone   0.19
## 2 Turning Tables           0.21
## 3 Someone Like You         0.288
## # A tibble: 3 x 2
##   track_name                          valence
##   <chr>                                 <dbl>
## 1 Right As Rain                         0.918
## 2 He Won't Go                           0.71
## 3 Send My Love (To Your New Lover)      0.562
## # A tibble: 4 x 2
##   track_name           valence
##   <chr>                  <dbl>
## 1 Grenade                0.227
## 2 When I Was Your Man    0.387
```

```
## 3 Just The Way You Are   0.434
## 4 Just the Way You Are   0.434
## # A tibble: 3 x 2
##   track_name    valence
##   <chr>           <dbl>
## 1 The Lazy Song   0.949
## 2 Finesse         0.939
## 3 Treasure        0.937
```

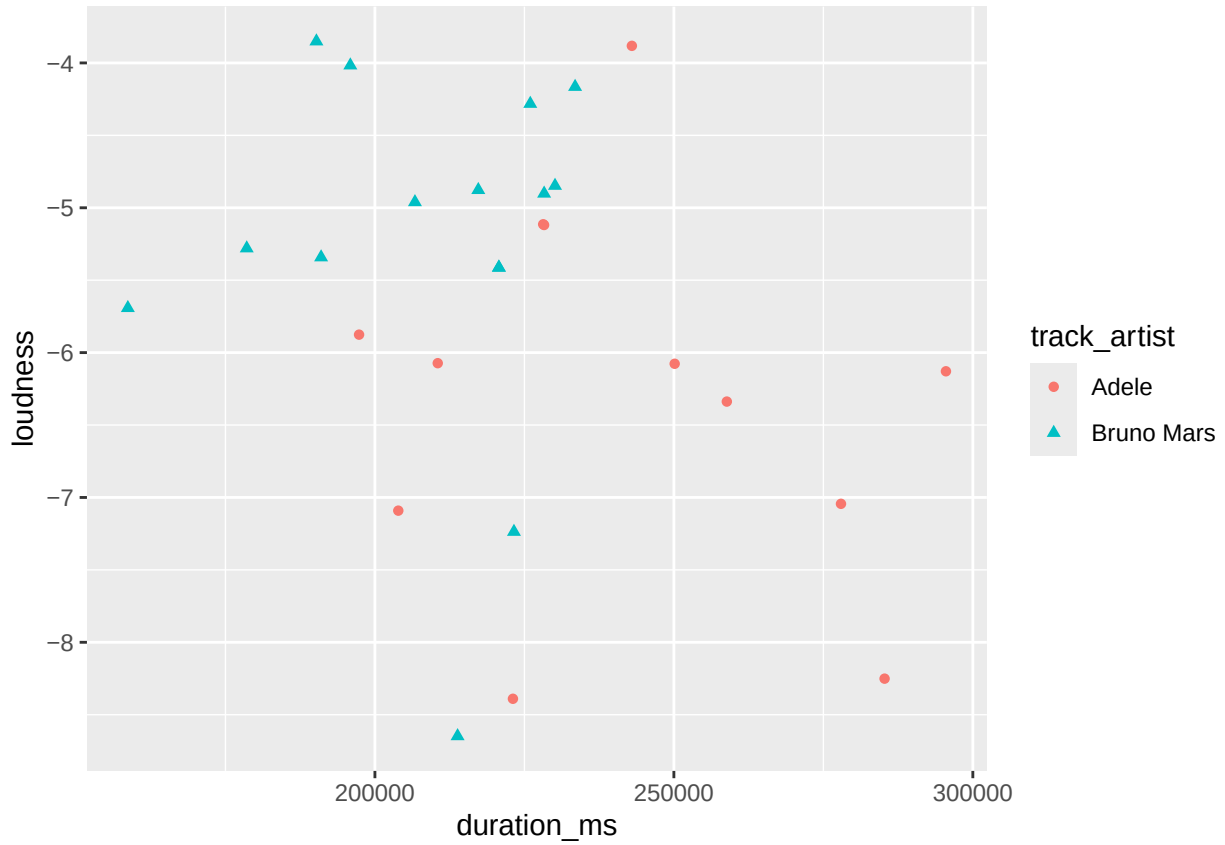
Zanimljivo, krivo je uneseno ime pjesme. Autoru su poznate neke pjesme te se slaže s ugođajima iz skupa podataka. Pogledajmo još samo prosjeke.

```
songs_both_clean %>% group_by(track_artist) %>% summarise(meanValence=mean(valence))
```

```
## # A tibble: 2 x 2
##   track_artist meanValence
##   <chr>              <dbl>
## 1 Adele              0.463
## 2 Bruno Mars         0.668
```

Ako nam se slušaju tužne pjesme, vjerojatno ćemo pustiti Adele, a ako nam se slušaju sretne onda ćemo pustiti Brunu Marsa. Na temelju ovog zaključka ćemo pretpostaviti da Bruno ima glasnije pjesme te da glasnije pjesme traju kraće. Glasnoća se mjeri u decibelima te što je ta veličina manja, to znači da je pjesma glasnija (relativno s obzirom na točku koja predstavlja najveću moguću glasnoću koja se može percipirati bez smetnji, čija vrijednost je 0).

```
ggplot(songs_both_clean,aes(x=duration_ms,y=loudness,shape=track_artist,color=track_artist)) + geom_poi
```



Iako uzorak nije velik, na grafu postoji naznaka da Bruno ima više glasnih pjesama te da one u prosjeku traju kraće, dok Adele ima tiše pjesme i traju dulje. Da zaključimo, Bruno Mars je bolji kad ste u veselom raspolaženju, iako i on ima tužnih pjesama, dok će inače Adele biti bolji izbor.

Na kojeg od najpopularnijih izvođača današnjice ćemo najlakše zaplesati?

Izdvojimo pjesme 5 najpopularnijih izvođača na Spotifyju (prema podacima od 27. siječnja 2025.).

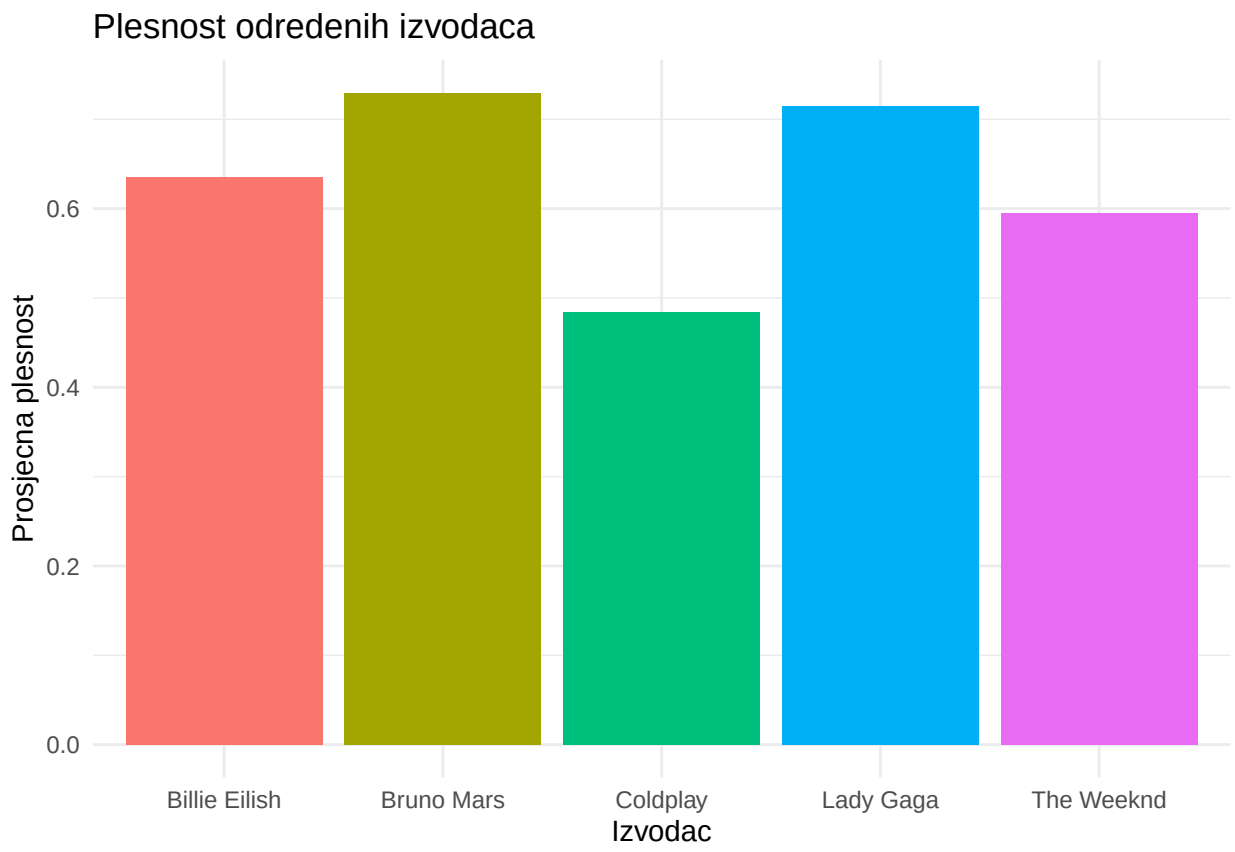
```
top5Izvodjaci <- c("Bruno Mars", "The Weeknd", "Lady Gaga", "Billie Eilish", "Coldplay")
songs %>% filter(track_artist %in% top5Izvodjaci) %>% select(track_name, track_artist, danceability) %>%
songs1
```

```
## # A tibble: 138 x 3
##    track_name                   track_artist  danceability
##    <chr>                        <chr>                <dbl>
##  1 bad guy (with Justin Bieber) Billie Eilish        0.67
##  2 Poker Face                   Lady Gaga            0.851
##  3 bad guy                      Billie Eilish        0.701
##  4 A Sky Full of Stars          Coldplay             0.551
##  5 everything i wanted          Billie Eilish        0.704
##  6 Blinding Lights              The Weeknd           0.513
##  7 Heartless                    The Weeknd           0.531
##  8 The Morning                  The Weeknd           0.652
```

```
##  9 Orphans                    Coldplay            0.503
## 10 Applause                   Lady Gaga           0.669
## # i 128 more rows
```

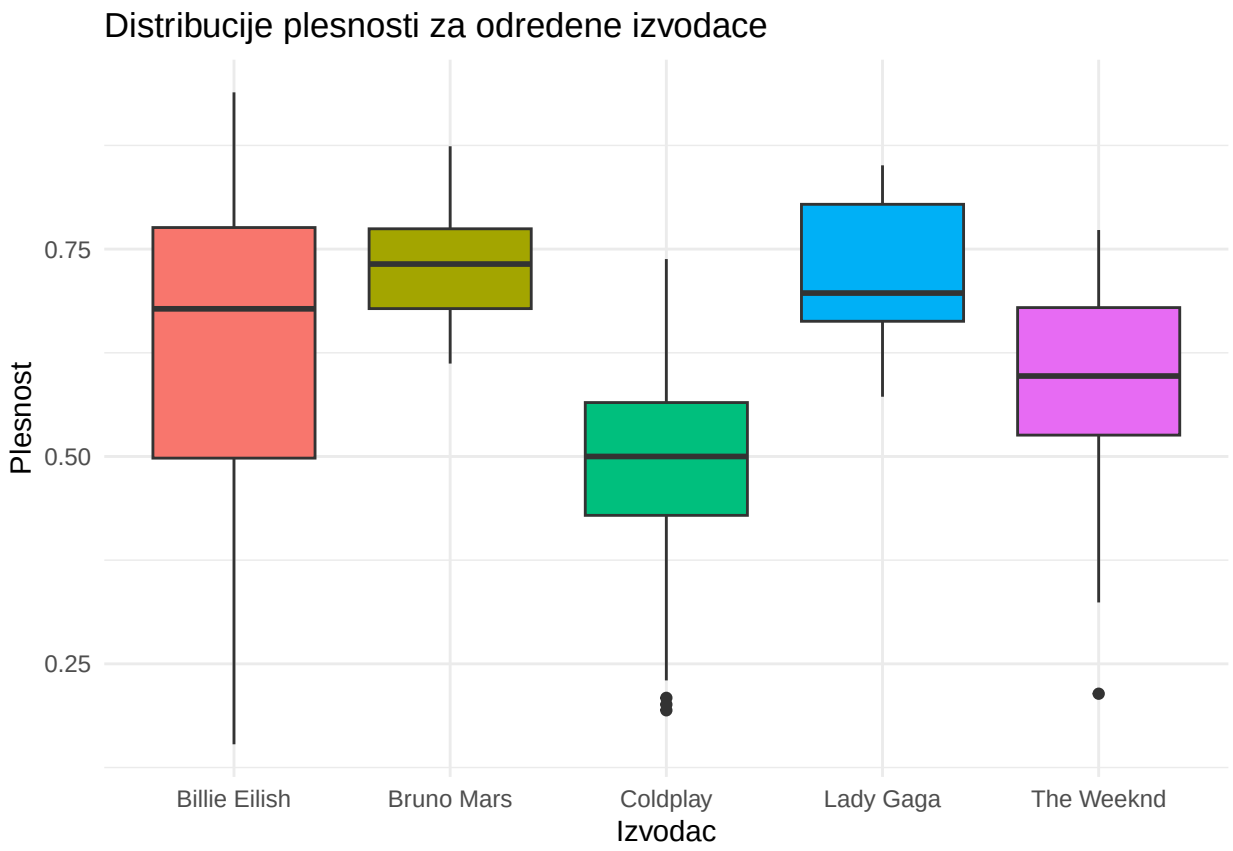
Pogledajmo koji izvođač ima pjesme s najvećom plesnošću.

```
ggplot(songs1, aes(x = track_artist, y = danceability, fill = track_artist)) +
  geom_bar(stat = "summary", fun = "mean") +
  labs(title = "Plesnost određenih izvođača",
       x = "Izvođač",
       y = "Prosječna plesnost") +
  theme_minimal() +
  theme(legend.position = "none")
```



Bruno Mars ima najplesnije pjesme, dok Coldplay, najdraži bend jednog od autora ovog projekta, ima najmanje plesne pjesme. Prikažimo još distribucije njihovih pjesama po plesnosti na box-plot dijagramu.

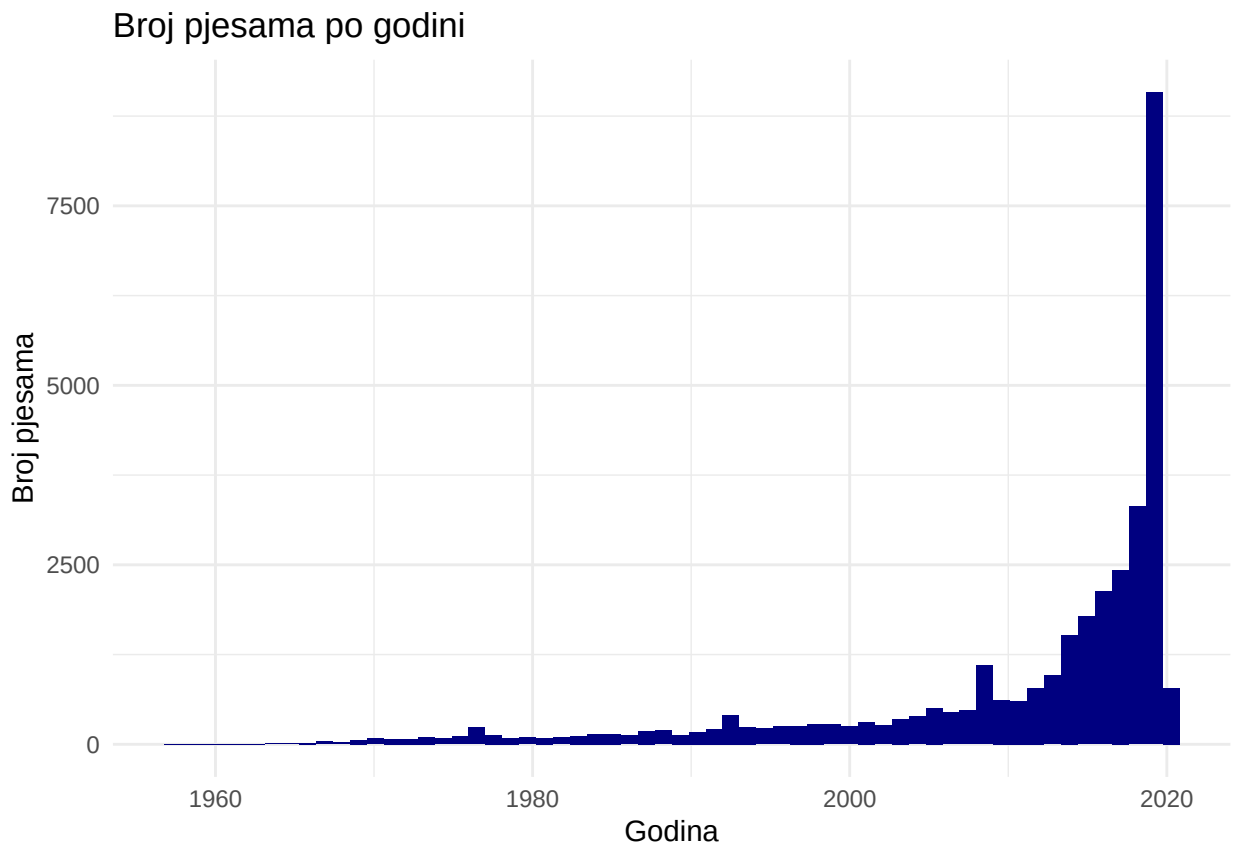
```
ggplot(songs1, aes(x = track_artist, y = danceability, fill = track_artist)) +
  geom_boxplot() +
  labs(title = "Distribucije plesnosti za određene izvođače",
       x = "Izvođač",
       y = "Plesnost") +
  theme_minimal() +
  theme(legend.position = "none")
```



Razlikuju li se nove pjesme od starih?

Možemo li primjetiti promjenu u određenim karakteristikama pjesme (npr. akustičnost, glasnoća ili raspoloženje) kroz godine? Najprije ćemo provjeriti broj pjesama po godini jer i to može imati utjecaja na konačne rezultate.

```
songs %>%
  select(track_name, track_album_release_date, acousticness) %>%
  filter(!is.na(track_album_release_date)) %>%
  mutate(year = as.numeric(substr(track_album_release_date, 1, 4))) %>%
  ggplot(aes(x = year)) + geom_histogram(bins = 60, fill = "navyblue") +
  labs(
    title = "Broj pjesama po godini",
    x = "Godina",
    y = "Broj pjesama"
  ) +
  theme_minimal()
```

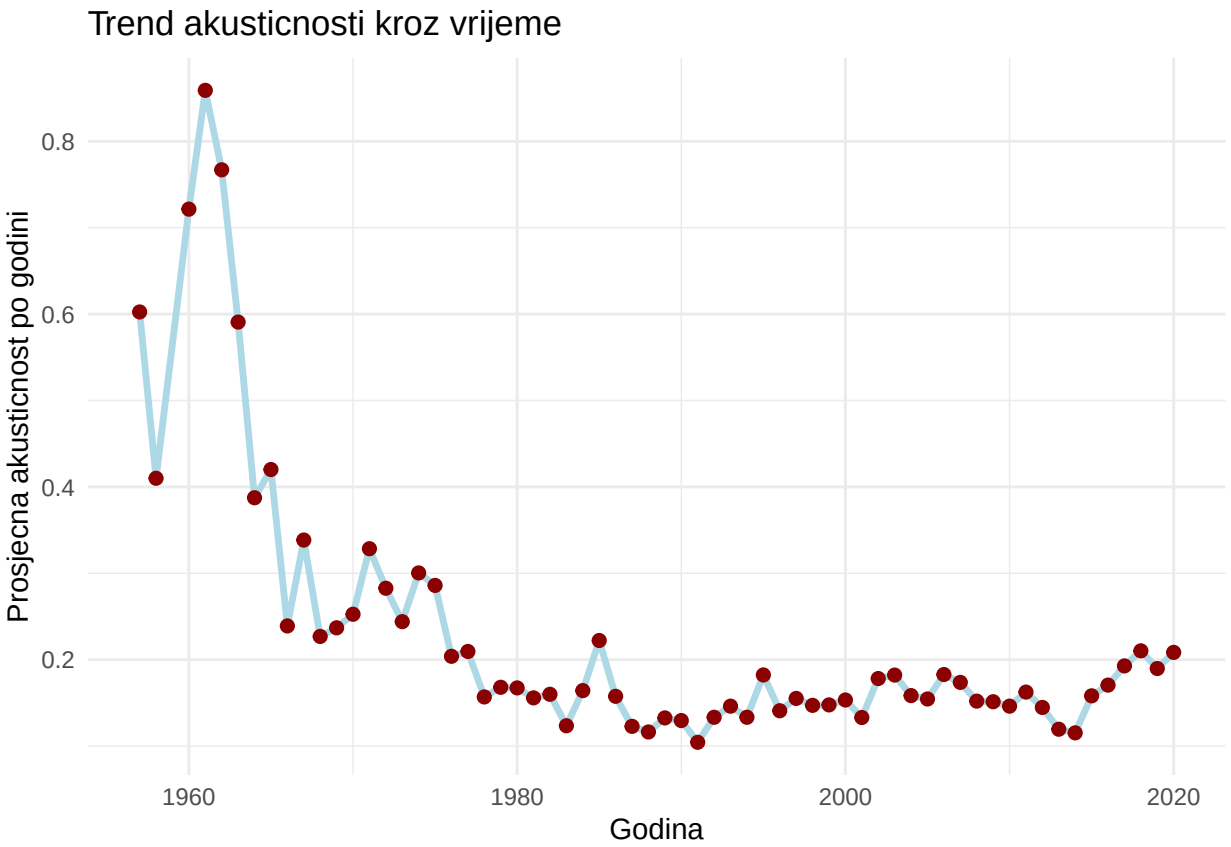


Logično, iz starijih godina postoji puno manje pjesama za analizu. Uzet ćemo to u obzir pri analizi rezultata. Grupirajmo sada vrijednosti parametara po godinama:

```
songs2 <- songs %>%
  select(track_name, track_album_release_date, acousticness, loudness, valence, tempo, duration_ms) %>%
  filter(!is.na(track_album_release_date)) %>%
  mutate(year = as.numeric(substr(track_album_release_date, 1, 4))) %>%
  group_by(year) %>%
  summarise(avg_acousticness = mean(acousticness, na.rm = TRUE),
            avg_loudness = mean(loudness, na.rm = TRUE),
            avg_valence = mean(valence, na.rm = TRUE),
            avg_tempo = mean(tempo, na.rm = TRUE),
            avg_duration_ms = mean(duration_ms, na.rm = TRUE))
```

Sada možemo grafički prikazati vrijednosti određenih parametara po godini. Učinimo to najprije za akustičnost.

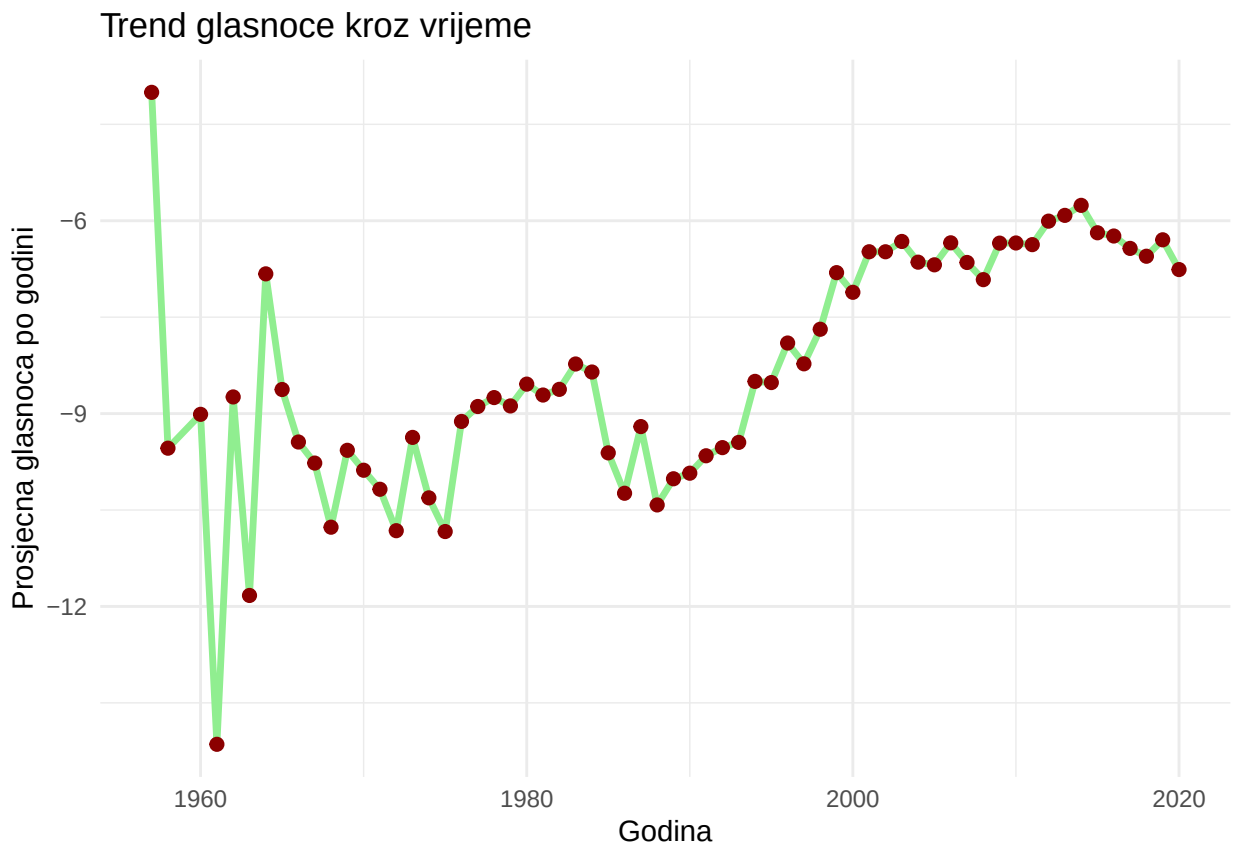
```
ggplot(songs2, aes(x = year, y = avg_acousticness)) +
  geom_line(color = "lightblue", linewidth = 1.2) +
  geom_point(color = "darkred", size = 2) +
  labs(
    title = "Trend akustičnosti kroz vrijeme",
    x = "Godina",
    y = "Prosječna akustičnost po godini"
  ) +
  theme_minimal()
```



Primjetimo kako su malobrojne pjesme iz 60-ih, te donekle 70-ih godina bile akustičnije, dok od 80-ih do danas akustičnost stagnira. To je i očekivano s obzirom na razvoj elektronske i dance glazbe u 80-im godinama.

Analizirajmo sada promjenu glasnoće.

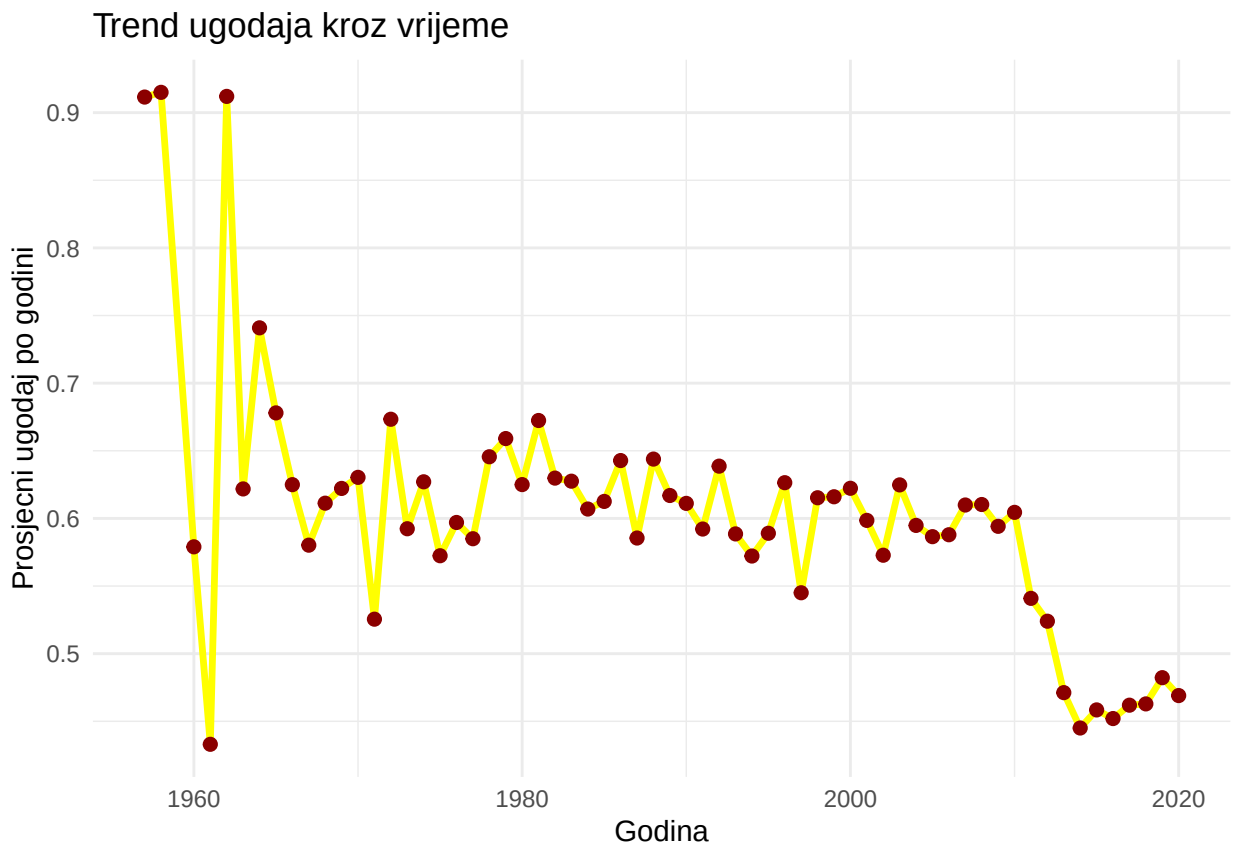
```
ggplot(songs2, aes(x = year, y = avg_loudness)) +
  geom_line(color = "lightgreen", linewidth = 1.2) +
  geom_point(color = "darkred", size = 2) +
  labs(
    title = "Trend glasnoće kroz vrijeme",
    x = "Godina",
    y = "Prosječna glasnoća po godini"
  ) +
  theme_minimal()
```



Od kasnih 80-ih do ranih 2000-ih glasnoća je bila u konstantnom porastu, nakon čega stagnira. Velika varijabilnost u 60-im godinama vjerojatno je posljedica vrlo malog broja pjesama iz tog vremena.

Pogledajmo sada na isti način podatke o ugođaju.

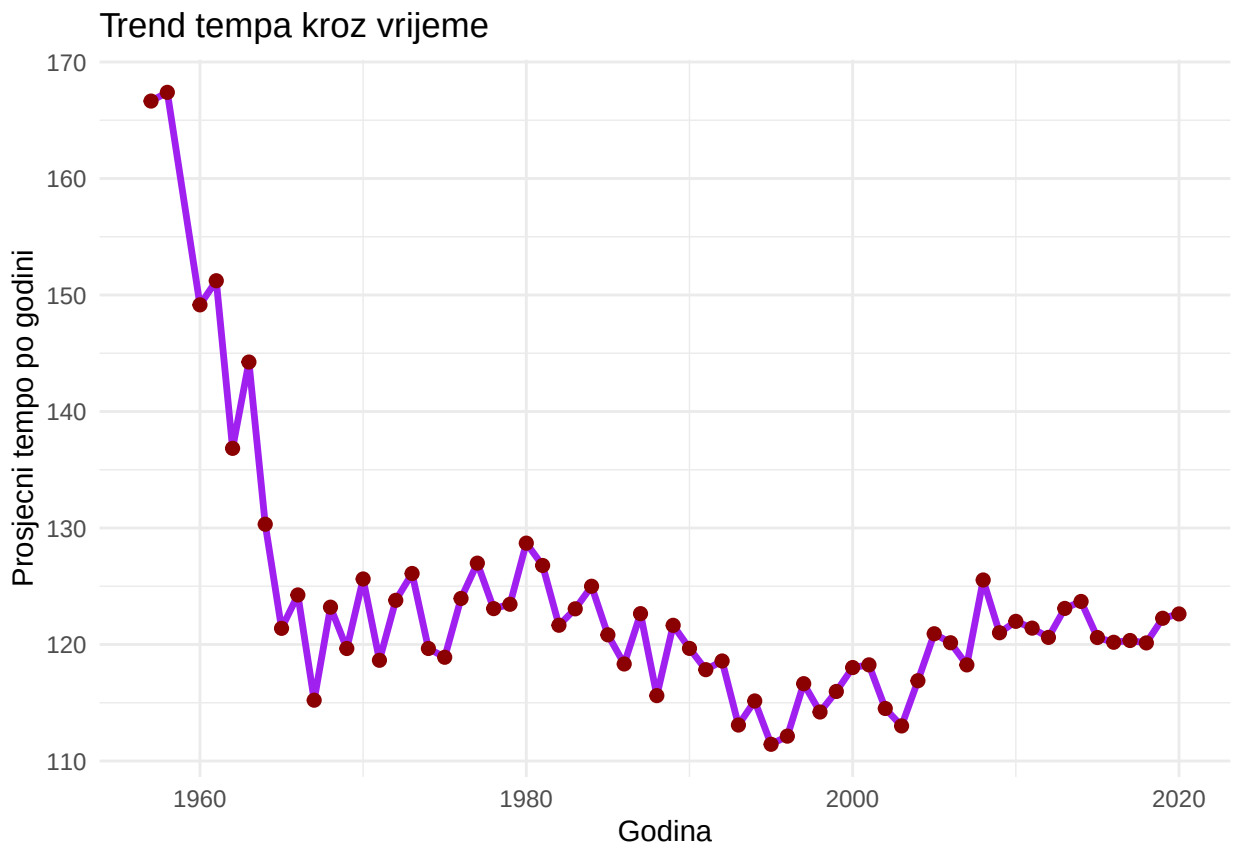
```
ggplot(songs2, aes(x = year, y = avg_valence)) +
  geom_line(color = "yellow", linewidth = 1.2) +
  geom_point(color = "darkred", size = 2) +
  labs(
    title = "Trend ugođaja kroz vrijeme",
    x = "Godina",
    y = "Prosječni ugođaj po godini"
  ) +
  theme_minimal()
```



U 60-ima je ponovno velika varijabilnost pa ne možemo ništa zaključiti za to razdoblje. Raspoloženje pjesama je u prosjeku imali slične vrijednosti sve do ranih 2010-ih. Od 2010. do 2014. primjećujemo vrlo nagli pad u raspoloženju pjesama s 0.6 na 0.4. To sugerira da je u prošlom desetljeću emotivna, tužna glazba postala sve učestalija, te je takav trend ostao i do danas. Ovakvi rezultati su prilično iznenađujući.

Promotrimo sada ovisnost tempa o godini izdanja pjesme.

```
ggplot(songs2, aes(x = year, y = avg_tempo)) +
  geom_line(color = "purple", linewidth = 1.2) +
  geom_point(color = "darkred", size = 2) +
  labs(
    title = "Trend tempa kroz vrijeme",
    x = "Godina",
    y = "Prosječni tempo po godini"
  ) +
  theme_minimal()
```



Malobrojne pjesme oko 1960. godine su bile vrlo brze, nakon čega prosječni tempo približno stagnira. Pjesme su bile najsporije oko 1995. godine nakon čega se primjećuje blagi rast u prosječnom tempu pjesme. Djeluje da prosječni tempo ne ovisi pretjerano (ako i uopće) o godini izdanja pjesme.

Za kraj ovog dijela pogledajmo još mijenja li se trajanje pjesme s obzirom na godinu izdanja.

```
ggplot(songs2, aes(x = year, y = avg_duration_ms)) +
  geom_line(color = "#e1be9b", linewidth = 1.2) +
  geom_point(color = "darkred", size = 2) +
  labs(
    title = "Trend trajanja pjesme kroz vrijeme",
    x = "Godina",
    y = "Prosječno trajanje (u ms) po godini"
  ) +
  theme_minimal()
```

Malobrojne pjesme iz 60-ih i ranijih godina su bile vrlo kratke. Potom se primjećuje trend rasta sve do kasnih 80-ih, nakon čega započinje pad koji je posebno nagao u 2010-ima. Moguće je da je ovo uzrokovano rastom popularnosti streaming servisa, gdje je svako ponovno preslušavanje pjesama važno za njihovu zaradu pa time njihovo trajanje postaje sve kraće.

Možemo li predvidjeti popularnost pjesme na temelju njenih obilježja?

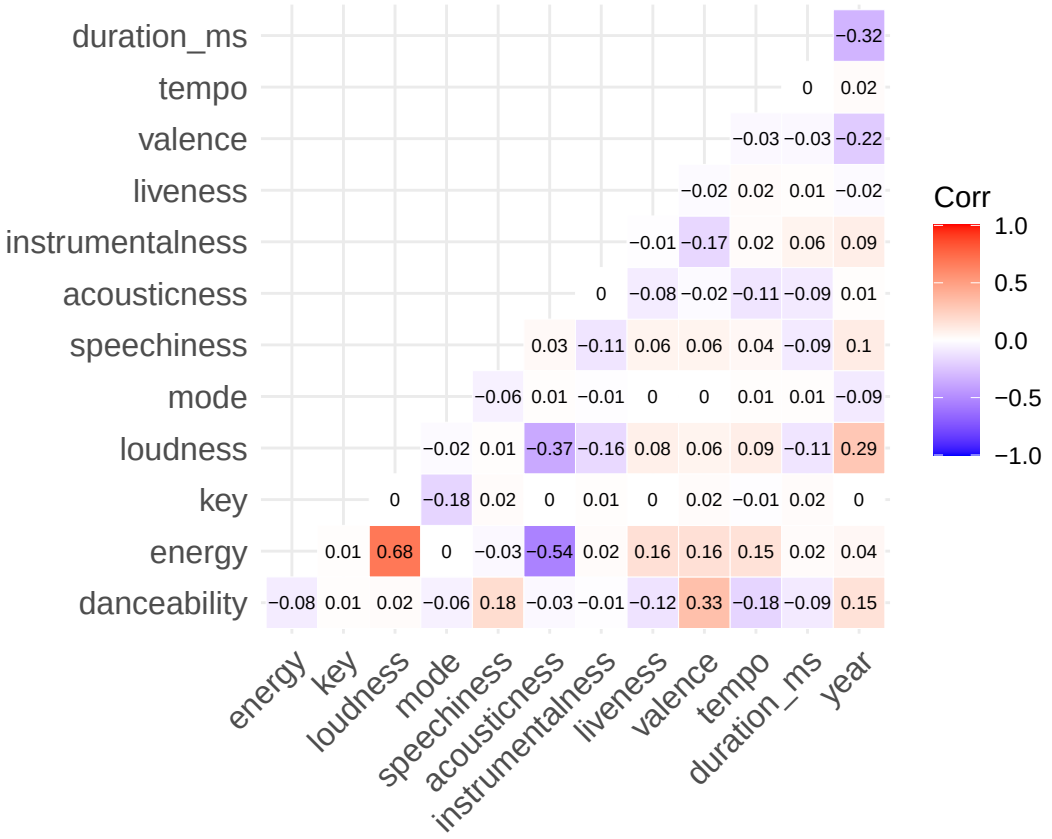
Izdvojimo u zasebni podatkovni okvir one podatke koji će nam trebati za ovo pitanje.

```
songs3 <- songs %>%
  select(track_popularity, playlist_genre, track_album_release_date, danceability:duration_ms) %>%
  mutate(year = as.numeric(substr(track_album_release_date, 1, 4))) %>%
  select(-track_album_release_date) %>%
  unique()
```

Popularnost pjesme u ovom je podatkovnom skupu određena brojem od 0 do 100. Vrijednost 100 predstavlja najpopularniju pjesmu, a 0 najmanje popularnu.

Provjerimo najprije jesu li neki od podataka međusobno korelirani.

```
ggcorrplot(songs3 %>% select(-track_popularity, -playlist_genre) %>% cor, type = "lower", outline.color
```



Primjećujemo da su među većinom obilježja vrlo mali koeficijenti korelacije. Jedino što od toga odskače je parametar "energy" koji je u izrazitoj korelaciji (apsolutne vrijednosti veće od 0.5) s parametrima acousticness i loudness. Stoga tu varijablu ne moramo koristiti u modelu.

Učinimo još tonski rod faktoriziranom varijablom.

```
songs3$mode <- factor(songs3$mode, levels = c(0, 1), labels = c("minor", "major"))
```

Napravimo najprije linearni model koristeći sve varijable i ispišimo njegov rezultat:

```
lm1 <- lm(track_popularity ~ ., songs3)
summary(lm1)
```

```
##
## Call:
## lm(formula = track_popularity ~ ., data = songs3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.439 -17.025   3.078  18.096  64.109
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -3.864e+01  3.250e+01  -1.189  0.23450
## playlist_genrelatin  6.716e+00  5.027e-01  13.360  < 2e-16 ***
## playlist_genrepop    8.975e+00  4.809e-01  18.662  < 2e-16 ***
```

```
## playlist_genrer&b    2.732e+00  5.208e-01   5.246 1.57e-07 ***
## playlist_genrerap    4.514e+00  5.045e-01   8.948  < 2e-16 ***
## playlist_genrerock   9.411e+00  5.799e-01  16.227  < 2e-16 ***
## danceability         1.020e+01  1.156e+00   8.818  < 2e-16 ***
## energy              -2.666e+01  1.248e+00 -21.359  < 2e-16 ***
## key                  7.685e-03  3.775e-02   0.204  0.83870
## loudness             1.454e+00  6.823e-02  21.316  < 2e-16 ***
## modemajor            4.890e-01  2.775e-01   1.762  0.07805 .
## speechiness         -3.075e+00  1.496e+00  -2.056  0.03983 *
## acousticness         3.767e+00  7.392e-01   5.095 3.50e-07 ***
## instrumentalness    -8.025e+00  6.523e-01 -12.302  < 2e-16 ***
## liveness            -3.660e+00  8.904e-01  -4.111 3.96e-05 ***
## valence             -7.314e-01  6.934e-01  -1.055  0.29150
## tempo                2.574e-02  5.176e-03   4.974 6.61e-07 ***
## duration_ms         -4.313e-05  2.399e-06 -17.981  < 2e-16 ***
## year                 5.168e-02  1.592e-02   3.245  0.00117 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.39 on 30353 degrees of freedom
## Multiple R-squared:  0.08306,    Adjusted R-squared:  0.08252
## F-statistic: 152.8 on 18 and 30353 DF,  p-value: < 2.2e-16
```

Primjećujemo da neke varijable ne utječu statistički značajno na popularnost. To su tonalitet (key) i raspoloženje (valence). Pogledajmo kako bi model izgledao bez njih:

```
lm2 <- lm(track_popularity ~ ., songs3 %>% select(-valence, -key))
summary(lm2)
```

```
##
## Call:
## lm(formula = track_popularity ~ ., data = songs3 %>% select(-valence,
##     -key))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.373 -17.023   3.072  18.083  64.187
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -4.586e+01  3.175e+01  -1.444 0.148672
## playlist_genrelatin  6.607e+00  4.917e-01  13.439  < 2e-16 ***
## playlist_genrepop    8.894e+00  4.747e-01  18.737  < 2e-16 ***
## playlist_genrer&b    2.639e+00  5.133e-01   5.142 2.73e-07 ***
## playlist_genrerap    4.468e+00  5.026e-01   8.891  < 2e-16 ***
## playlist_genrerock   9.315e+00  5.727e-01  16.265  < 2e-16 ***
## danceability         9.710e+00  1.061e+00   9.154  < 2e-16 ***
## energy              -2.698e+01  1.209e+00 -22.325  < 2e-16 ***
## loudness             1.458e+00  6.813e-02  21.404  < 2e-16 ***
## modemajor            4.829e-01  2.733e-01   1.767 0.077223 .
## speechiness         -3.114e+00  1.495e+00  -2.083 0.037299 *
## acousticness         3.681e+00  7.346e-01   5.011 5.43e-07 ***
## instrumentalness    -7.949e+00  6.485e-01 -12.259  < 2e-16 ***
```

```
## liveness            -3.655e+00  8.904e-01  -4.105 4.05e-05 ***
## tempo                2.546e-02  5.169e-03   4.925 8.47e-07 ***
## duration_ms         -4.290e-05  2.388e-06 -17.960  < 2e-16 ***
## year                 5.542e-02  1.552e-02   3.571 0.000356 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.39 on 30355 degrees of freedom
## Multiple R-squared:  0.08303,    Adjusted R-squared:  0.08254
## F-statistic: 171.8 on 16 and 30355 DF,  p-value: < 2.2e-16
```

Primjećujemo da se koeficijent determinacije nije promjenio sve do 5.decimale, a prilagođeni koeficijent (koji kažnjava overfitting) je čak i porastao u 5.decimali. Iako su sve preostale varijable statistički značajne za popularnost, koeficijent determinacije je jako mali.

Napravimo sada model bez varijable energy koja je u korelaciji s drugim varijablama:

```
lm3 <- lm(track_popularity ~ ., songs3 %>% select(-valence, -key, -energy))
summary(lm3)
```

```
##
## Call:
## lm(formula = track_popularity ~ ., data = songs3 %>% select(-valence,
##     -key, -energy))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.540 -17.224   2.937  18.234  63.655
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.877e+02  3.136e+01  -5.986 2.18e-09 ***
## playlist_genrelatin  6.968e+00  4.954e-01  14.066  < 2e-16 ***
## playlist_genrepop    9.802e+00  4.768e-01  20.558  < 2e-16 ***
## playlist_genrer&b    4.690e+00  5.091e-01   9.213  < 2e-16 ***
## playlist_genrerap    6.233e+00  5.004e-01  12.457  < 2e-16 ***
## playlist_genrerock   9.550e+00  5.773e-01  16.543  < 2e-16 ***
## danceability         1.130e+01  1.067e+00  10.593  < 2e-16 ***
## loudness             4.939e-01  5.311e-02   9.299  < 2e-16 ***
## modemajor            5.973e-01  2.754e-01   2.168 0.030132 *
## speechiness         -5.357e+00  1.504e+00  -3.562 0.000369 ***
## acousticness         1.000e+01  6.833e-01  14.637  < 2e-16 ***
## instrumentalness    -1.024e+01  6.455e-01 -15.861  < 2e-16 ***
## liveness            -5.999e+00  8.914e-01  -6.731 1.72e-11 ***
## tempo                1.942e-02  5.204e-03   3.731 0.000191 ***
## duration_ms         -4.340e-05  2.408e-06 -18.024  < 2e-16 ***
## year                 1.127e-01  1.543e-02   7.301 2.93e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.58 on 30356 degrees of freedom
## Multiple R-squared:  0.06797,    Adjusted R-squared:  0.06751
## F-statistic: 147.6 on 15 and 30356 DF,  p-value: < 2.2e-16
```

Ovdje primjećujemo blagi pad koeficijenta determinacije. Međutim, bez obzira koji od gornjih modela koristili, koeficijent determinacije je premali što znači da nijedan od tih modela ne objašnjava većinu varijance. Time zaključujemo da temeljem danih obilježja pjesama ne možemo predvidjeti njenu popularnost.

Dakle, među popularnim (i nepopularnim) pjesmama ima velikih raznolikosti: i novih i starih pjesama, i brzih i sporih, i plesnih i mirnih... Ukusi se razlikuju, pa tako i navike slušanja, a u tome i je ljepota glazbe pa tako i umjetnosti općenito.