Eksploraturna podatkovnog skupa Spotify

Luka Žmak

2024-02-24

Uvod

Cilj ovog projekta je provesti analizu ulaznog skupa Spotify pjesama kroz prizmu različitih parametara, kao što su žanrovi, tempo, energija, popularnost izvođača i te brojni drugi. Planiram istražiti obrasce u glazbenim preferencama korisnika, prepoznati trendove koji oblikuju glazbenu industriju te razumjeti ključne karakteristike pjesama koje čine neku skladbu popularnom ili privlačnom publici.

Prvo i osnovno, učitat ćemo naše ulazne podatke:

```
## 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~
```

Pretvorba stupaca u odgovarajuće tipove podataka:

```
Spotify$track_album_release_date <- as.Date(Spotify$track_album_release_date,
                                            format = '%Y-%m-%d')
Spotify$mode <- factor(Spotify$mode, levels = c(0, 1))
Spotify$key <- factor(Spotify$key, levels = -1:11)
Spotify$playlist_genre <- factor(Spotify$playlist_genre,
                         levels = c("pop", "rap", "rock", "latin","r&b", "edm"))
Spotify$playlist_subgenre <- factor(Spotify$playlist_subgenre,
                                    levels = unique(Spotify$playlist_subgenre))
```

```
Spotify$mode <- sapply(Spotify$mode, function(num){
  if(num==0) return("mol")
  if(num==1) return("dur")
})

Spotify$mode <- as.factor(Spotify$mode)
Spotify$decade <- as.factor(10 * (year(Spotify$track_album_release_date) %/% 10))
```

Dodavanje stupca koji prikazuje koliko pojedina pjesma traje u formatu "minute:sekunde":

```
Spotify$duration_min <-  sprintf("%02d.%02d", floor(Spotify$duration_ms / 60000),
round(((Spotify$duration_ms / 60000) - floor(Spotify$duration_ms / 60000)) * 60))
```

Pregled žanrova, izvođača i albuma

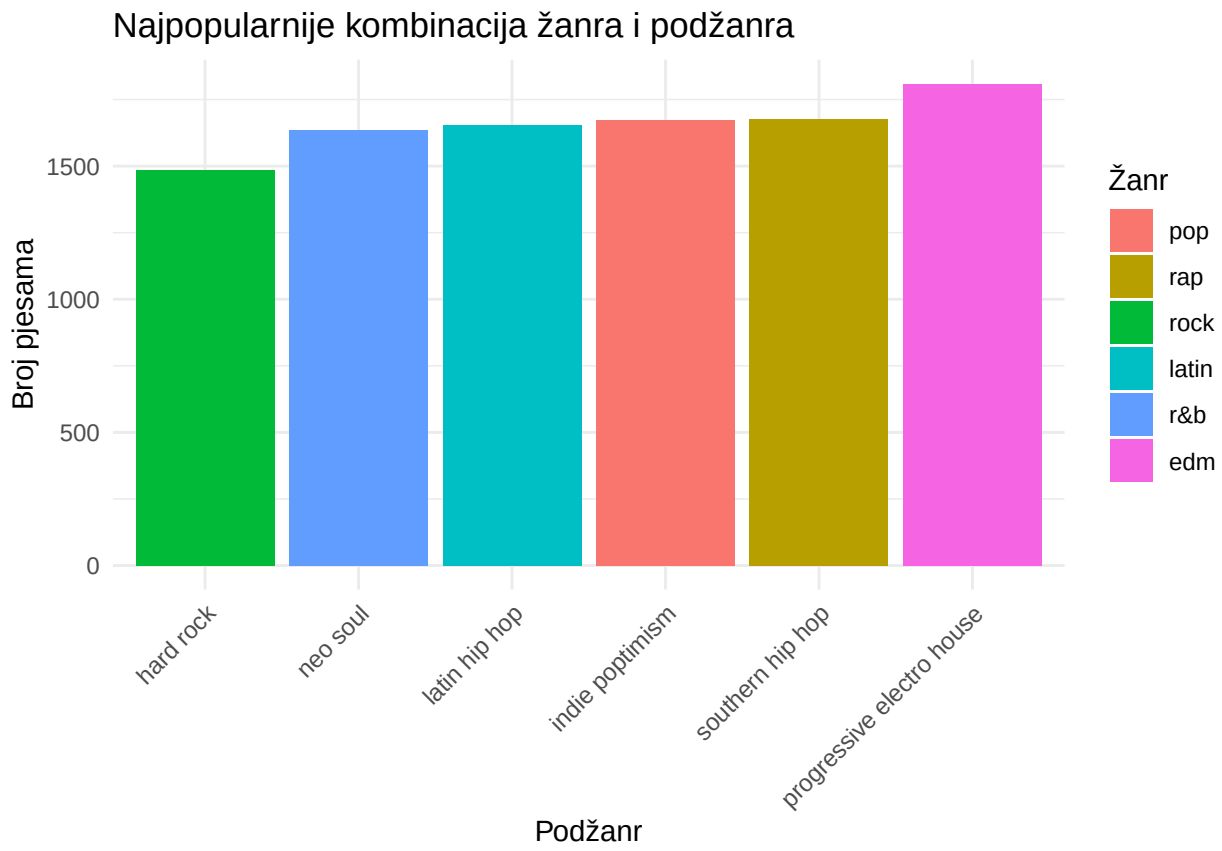
Pogledajmo koji sve žanrovi glazbe su u ovom skupu podataka.

```
## [1] "pop"   "rap"   "rock"  "latin" "r&b"   "edm"
```

Naš podatkovni skup osim žanra za pojedinu pjesmu ima definiran i podžanr. Ovdje je popis svih kombinacija žanrova poredan po broju pjesama:

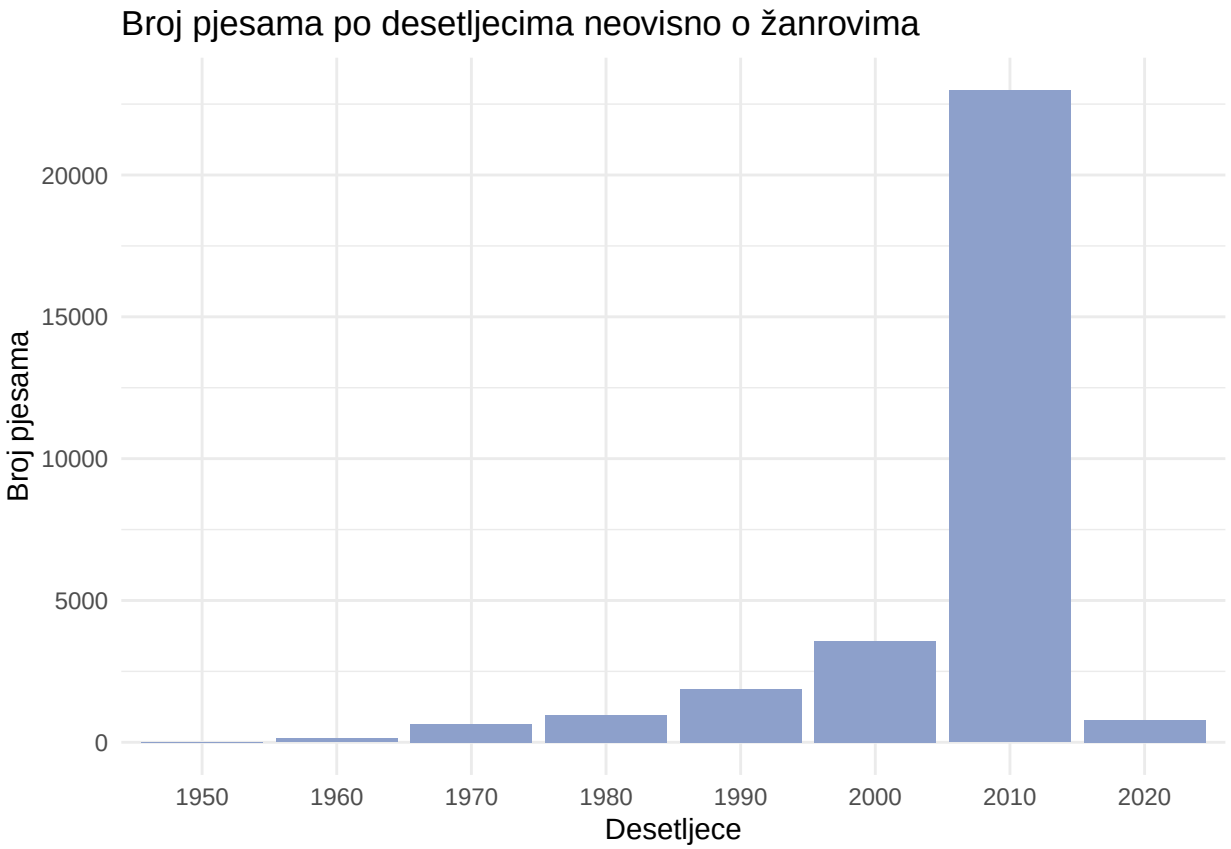
```
## # A tibble: 24 x 3
##    playlist_genre playlist_subgenre         count
##    <fct>          <fct>                     <int>
##  1 edm            progressive electro house  1809
##  2 rap            southern hip hop           1675
##  3 pop            indie poptimism            1672
##  4 latin          latin hip hop              1656
##  5 r&b            neo soul                   1637
##  6 edm            pop edm                    1517
##  7 edm            electro house              1511
##  8 rock           hard rock                  1485
##  9 rap            gangster rap               1458
## 10 pop            electropop                 1408
## # i 14 more rows
```

Grafički ćemo prikazati koje su najzastupljenije kombinacije pojedinog žanra s njegovim podžanrom:

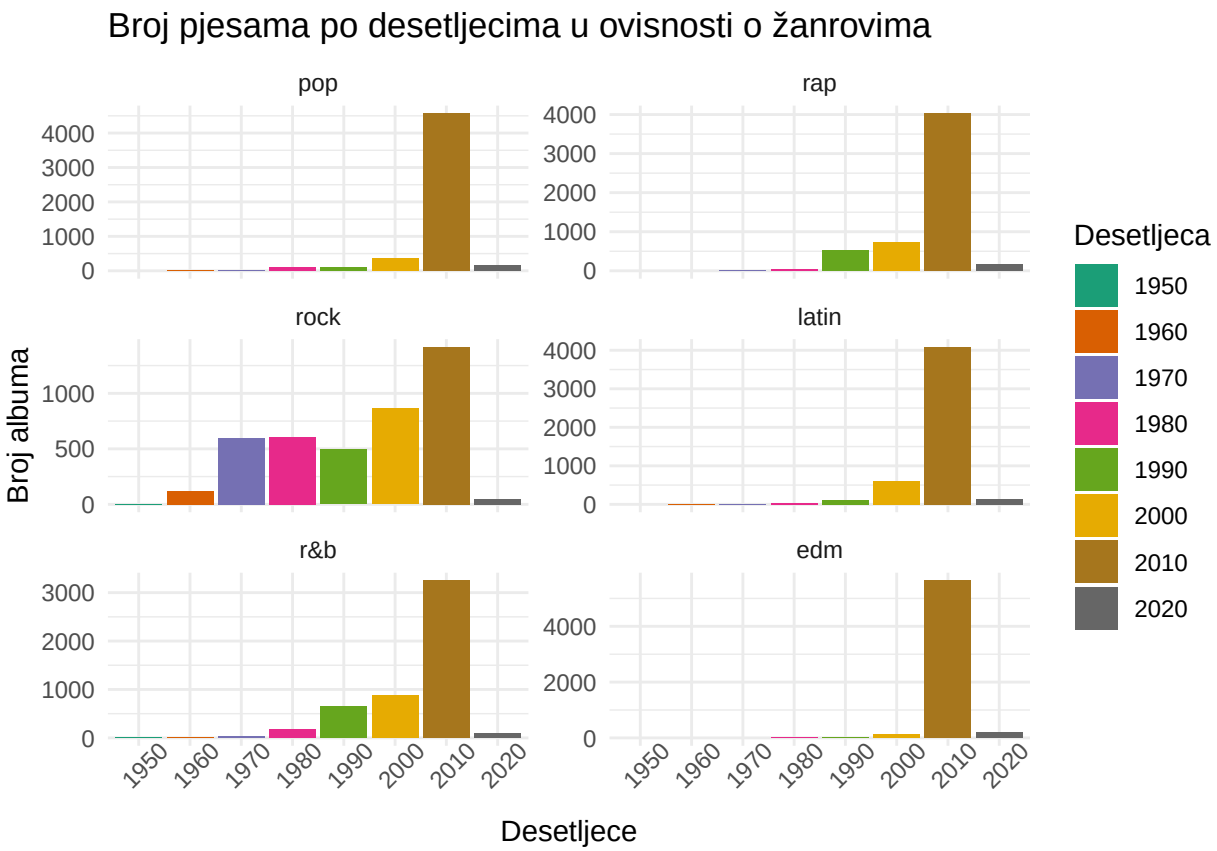


Trendovi kroz različita razdoblja

Korisnici često pretražuju pjesme prema razdoblju nastanka, stoga im ovaj graf može biti vrlo zanimljiv:



Htio smo provjeriti u kojem su se desetljeću smjestilo najviše pjesama, međutim gornji graf nam ukazuje da je prevelik uzorak pjesama nastao u 2010-im godinama. Stoga sam se odlučili fokusirati na popularnost pojedinog žanra po desetljećima. Ukoliko želimo vidjeti i raspodjelu pjesama po žanru kroz povijest, dobivamo ove rezultate:



Nastavno na prošli graf, rezultati žanrova pop-a, edm-a te rock-a me ne iznenađuje. Slavno doba roka bilo je 80-ih godina, rap se počeo razvijati 90-ih godina, dok je sada vrhunac žanrova edm i pop-a. Međutim, rezultati za r&b te latino muziku je učinio da se zapitam je li uzorak reprezentativan jer smo očekivao da je njihov vrhunac u nekim davnijim vremenima (r&b 50-ih godina, a latino muzika je bila sveprisutna kroz prošla razdoblja).

Stoga odlučujem provjeriti koji su najpopularniji izvođači pojedinog desetljeća s minimalno tri pjesme kako bi uzorak bio reprezentativan:

- Najpopularniji pjevači 1950-ih godina:

```
## # A tibble: 2 x 4
##   track_artist  total_popularity total_songs average_popularity
##   <chr>                    <dbl>       <int>              <dbl>
## 1 Elvis Presley               73           1                 73
## 2 Ray Charles                 59           1                 59
```

Budući da u razdoblju 50-ih imamo samo dvije pjesme, nije moguće donijeti neke kvalitetne zaključke.

- Najpopularniji pjevači 1960-ih godina:

```
## # A tibble: 5 x 4
##   track_artist        total_popularity total_songs average_popularity
##   <chr>                          <dbl>       <int>              <dbl>
## 1 The Beatles                      628           9               69.8
## 2 Jimi Hendrix                     481           7               68.7
## 3 Led Zeppelin                     668          10               66.8
## 4 Buffalo Springfield              262           4               65.5
## 5 Bob Dylan                        323           5               64.6
```

- Najpopularniji pjevači 1970-ih godina:

```
## # A tibble: 5 x 4
##   track_artist             total_popularity total_songs average_popularity
##   <chr>                               <dbl>       <int>              <dbl>
## 1 Eagles                                536           7               76.6
## 2 AC/DC                                 749          10               74.9
## 3 Electric Light Orchestra              279           4               69.8
## 4 Deep Purple                           416           6               69.3
## 5 The Beatles                           346           5               69.2
```

U razdoblju između 1960. - 1979. pretežito je najpopularniji žanr bio rock, a to odgovara i najpopularnijim izvođačima tog razdoblja.

- Najpopularniji pjevači 1980-ih godina:

```
## # A tibble: 5 x 4
##   track_artist      total_popularity total_songs average_popularity
##   <chr>                        <dbl>       <int>              <dbl>
## 1 a-ha                           332           4               83
## 2 AC/DC                          979          13               75.3
## 3 Bruce Springsteen              366           5               73.2
## 4 Luis Miguel                    286           4               71.5
## 5 Eurythmics                     495           7               70.7
```

Iako je 80-ih godina i dalje prevladavao rock, vidimo puno veću raznolikost među najpopularnijim pjevačima. Tako se na samom vrhu našao jedan pop band, dva rock izvođača te jedan predstavnik latino žanra.

- Najpopularniji pjevači 1990-tih godina:

```
## # A tibble: 5 x 4
##   track_artist          total_popularity total_songs average_popularity
##   <chr>                            <dbl>       <int>              <dbl>
## 1 Green Day                          462           6               77
## 2 Red Hot Chili Peppers              684           9               76
## 3 AC/DC                              369           5               73.8
## 4 Foo Fighters                       514           7               73.4
## 5 Destiny's Child                    365           5               73
```

Iako je najveći uzorak pjesama r&b, među pet najpopularnijih izvođača ne nalazimo niti jednog izvođača iz tog žanra. Rock i dalje ostaje najomiljeniji žanr.

- Najpopularniji pjevači 2000-tih godina:

```
## # A tibble: 5 x 4
##   track_artist          total_popularity total_songs average_popularity
##   <chr>                            <dbl>       <int>              <dbl>
## 1 Red Hot Chili Peppers              309           4               77.2
## 2 Green Day                          385           5               77
## 3 My Chemical Romance                307           4               76.8
## 4 MGMT                               454           6               75.7
## 5 Coldplay                           890          12               74.2
```

Slična priča je se nastavlja i u 2000-tim, iako je to razdoblje r&b, tog žanra nema na vidiku, te su u prosijeku pjesme rock izvođača najpopularnije.

- Najpopularniji pjevači 2010-ih godina:

```
## # A tibble: 5 x 4
##   track_artist  total_popularity total_songs average_popularity
```

```
##   <chr>                    <dbl>       <int>              <dbl>
## 1 Trevor Daniel              582           6               97
## 2 Y2K                        637           7               91
## 3 Don Toliver                635           7               90.7
## 4 Roddy Ricch               1676          19               88.2
## 5 DaBaby                    1230          14               87.9
```

2010-e godine donose promjene u trendu. Iako je najzastupljeniji žanr edm, njega ne nalazimo među 5 najpopularnijih izvođača. Uglavnom prevladavaju rap pjesme.

- Najpopularniji pjevači 2020-ih godina:

```
## # A tibble: 5 x 4
##   track_artist  total_popularity total_songs average_popularity
##   <chr>                    <dbl>       <int>              <dbl>
## 1 Justin Bieber              665           7               95
## 2 Future                     651           7               93
## 3 Selena Gomez              1255          15               83.7
## 4 Khalid                     415           5               83
## 5 Halsey                     656           8               82
```

Iako uzorak za 2020-e godine nije prevelik, jasno je da današnje generacije najviše uživaju slušajući pop.

Budući da uzorak nije ravnomjerno raspoređen po svim desetljećima, moja pretpostavka je da ukoliko bi izjednačili broj pjesama za svako razdoblje, ne bi vidjeli veliko odstupanje od gore navedenih rezultata.

Analiza parametara pjesama

Zadani podatkovni skup sadrži brojne parametre koji nam daju dobru predodžbu kakva je pojedina pjesma bez da je poslušamo. Stoga odlučujem provjeriti koliko su koji parametri međusobno povezani:

```
## Warning: package 'ggcorrplot' was built under R version 4.3.2
```

Uočavamo da većina parametara međusobno ne korelira, ali energija pjesme podosta ovisi o glasnoći i akustičnosti. Stoga sam odlučio napraviti jednostavni linearni model kako bi vidjeli koliko je dobro ta veza opisana.

```
linMod <- lm(energy~loudness+acousticness, data = Spotify)
summary(linMod)
```

```
##
## Call:
## lm(formula = energy ~ loudness + acousticness, data = Spotify)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49827 -0.07862  0.00613  0.08281  1.28352
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.9746657  0.0016742  582.18   <2e-16 ***
## loudness      0.0342723  0.0002481  138.16   <2e-16 ***
## acousticness -0.2742082  0.0033244  -82.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1195 on 30944 degrees of freedom
## Multiple R-squared:  0.5625, Adjusted R-squared:  0.5625
## F-statistic: 1.989e+04 on 2 and 30944 DF,  p-value: < 2.2e-16
```

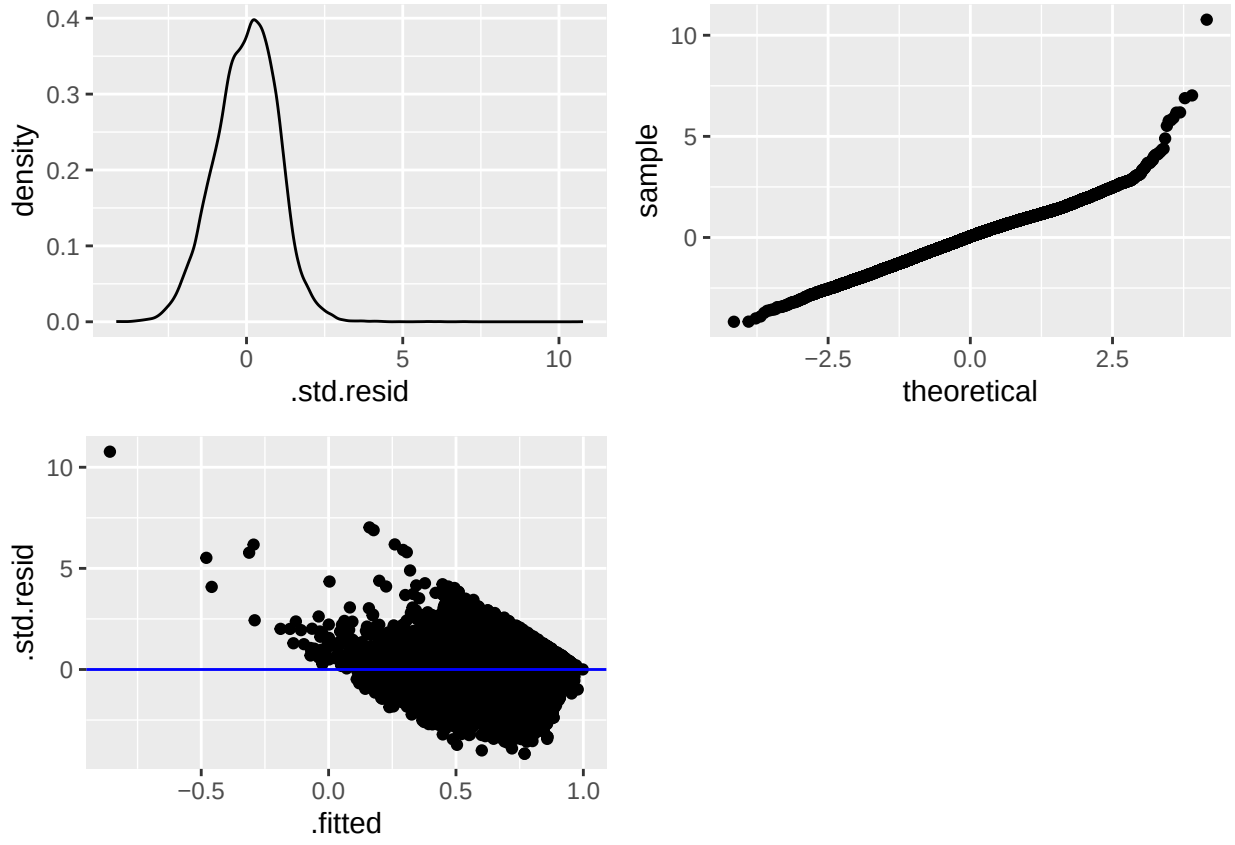
R-kvadrat parametar iznosi 0.5625 te **p-vrijednost** 2e-16. **R-kvadrat** nam predstavlja količinu varijabilnosti

koja je objašnjena modelom. S druge strane, **p-vrijednost**, tj. procjena vjerojatnosti da je kolinearnost uočena slučajno, a ovdje vidimo da je ona iznimno mala. Trenutno brojke obećavaju, ali i dalje ne možemo znati je li model dobar, stoga sam odlučilo preispitati ga:

```
##    PraveVrijednosti ProcijenjeneVrijednosti
## 1             0.916                   0.856
## 2             0.815                   0.785
## 3             0.931                   0.835
## 4             0.930                   0.837
## 5             0.833                   0.793
## 6             0.919                   0.768
## 7             0.856                   0.759
## 8             0.903                   0.883
## 9             0.935                   0.846
## 10            0.818                   0.803
```

U gornjoj tablici vidimo da razlika između procijenjenih i pravih vrijednosti ne varira previše. Želimo analizirati ponašanje reziduala, odnosno grešaka.

```
## Warning: package 'gridExtra' was built under R version 4.3.2
```



Iz gornjih grafova zaključujemo da se greške ponašaju prema normalnoj razdiobi, što je dobar znak za naš model.

Odlučujem provjeriti i koliko iznosi prosječna greška procijenjenih i pravih vrijednosti:

```
allresults <- predict(linMod, Spotify[, c("loudness", "acousticness")])
average_error <- (allresults - Spotify$energy) %>%  abs() %>% sum() %>%
  `/`(nrow(Spotify))
```

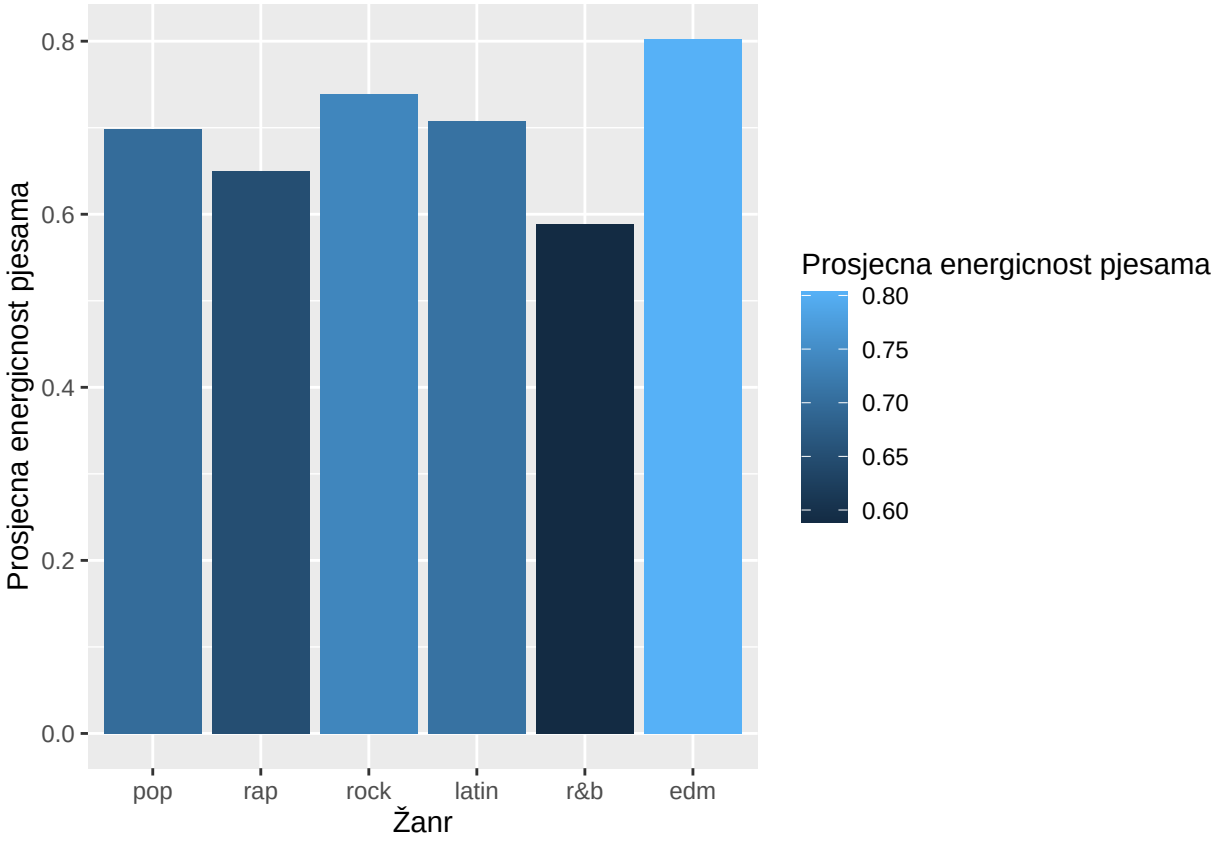
```
average_error
```

```
## [1] 0.09522884
```

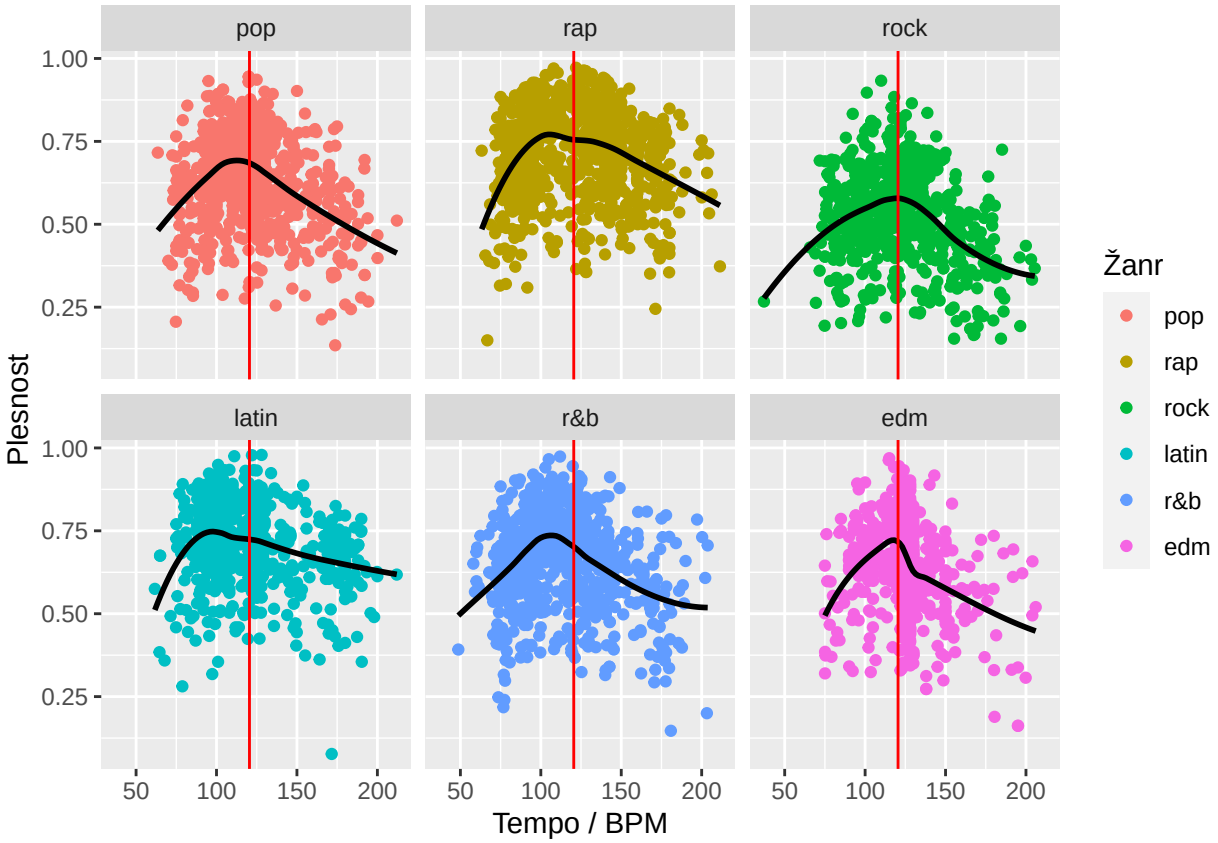
Smatram da ovakva greška dovoljno mala da bi generalno mogli procijeniti pjesmu kao energičnu ili ne energičnu.

Analiza pjesama prema njezinim karakteristikama

Mnogima je prilikom odabira playliste bitno koliko je ona energična. Prikazat ćemo prosječnu ocjenu energičnosti pjesama u svakom žanru.



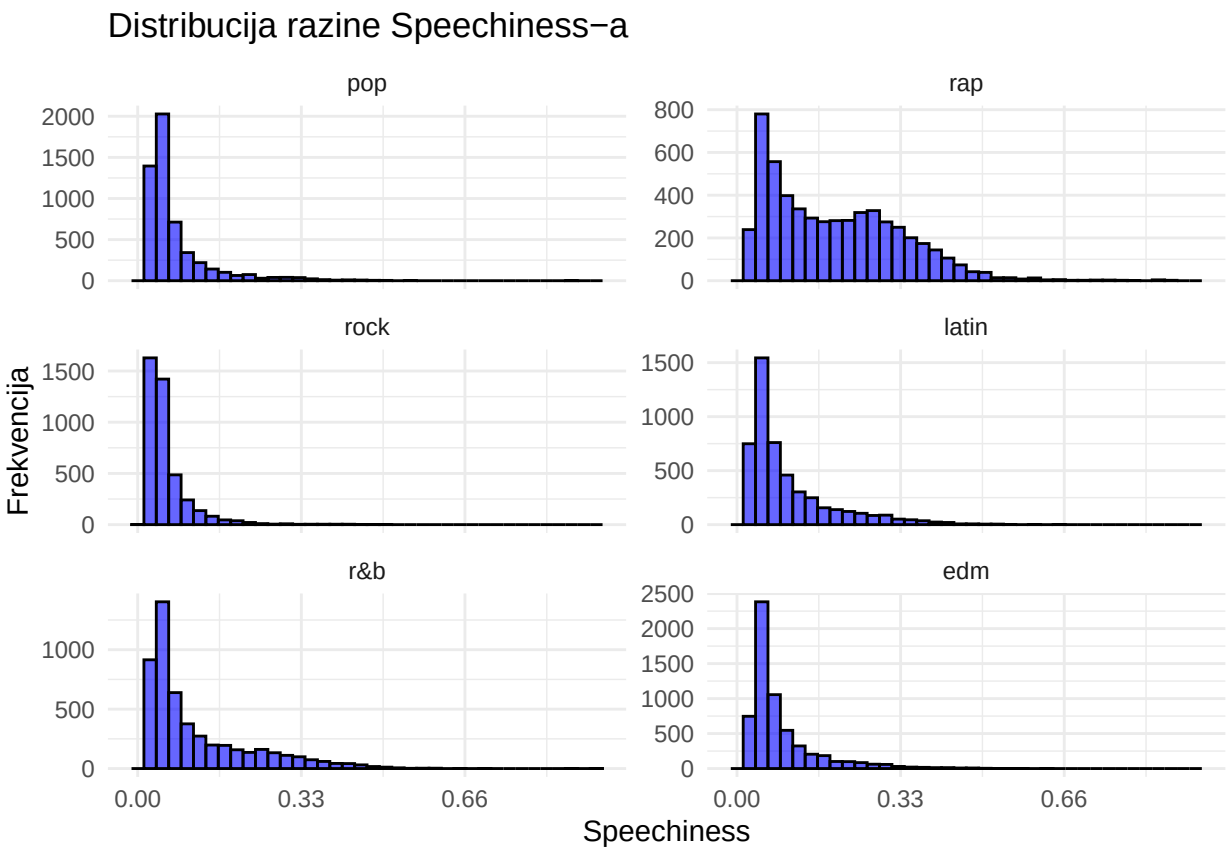
Očekivano, EDM (Electronic dance music) predvodi u ovoj kategoriji, dok se blues nalazi na posljednjem mjestu.



Omjer tempa i prikladnosti pjesme za plesanje slijedi model kvadratne funkcije sa negativnim vodećim članom. Ovakav model se potvrđuje kroz sve žanrove. Srednja vrijednost tempa svih žanrova iznosi 122 BPM (bitova po minuti) te se ta vrijednost otprilike poklapa s vrijednostima u kojima je prikladnost za ples najveća.

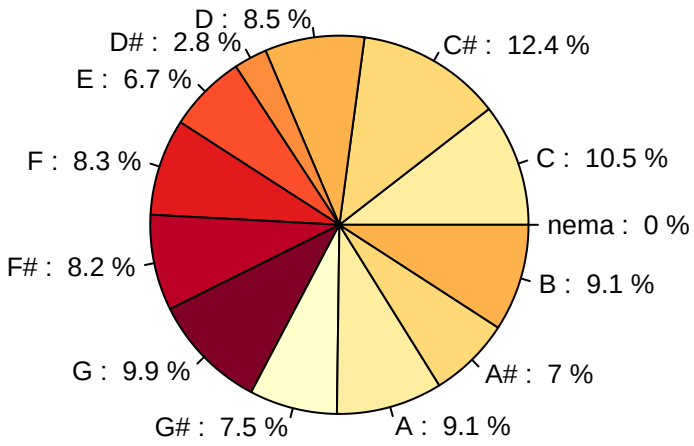
Speechiness detektira prisutnost govornih riječi u glazbenom zapisu. Što je snimka sličnija govoru (npr. razgovor, podcast), to će njegova vrijednost biti bliža 1.0. Vrijednosti iznad 0.66 opisuju zapise koji su vjerojatno potpuno sastavljeni od govornih riječi. Vrijednosti između 0.33 i 0.66 opisuju zapise koji mogu sadržavati i glazbu i govor, ili u određenim dijelovima ili slojevima, uključujući slučajeve poput rap glazbe. Vrijednosti ispod 0.33 najvjerojatnije predstavljaju glazbu i druge zapise koji nisu slični govoru.

Htio sam provjeriti razinu tog atributa za sve žanrove primjenjujući histogram:



Rezultati su očekivani. Većina pjesama svih žanrova ima razinu ispod 0.33, tj. radi se o zapisima koji predstavljaju glazbu. Iznimka je jedino rap, koji sadrži puno riječi te ne čudi činjenica da ima povišenu razinu tog parametra.

Uz pomoć key podataka možemo zaključiti koji je glavni tonalitet pojedine pjesme. Stoga me je zanimalo koji su tonaliteti najzastupljeniji u ovom uzorku:



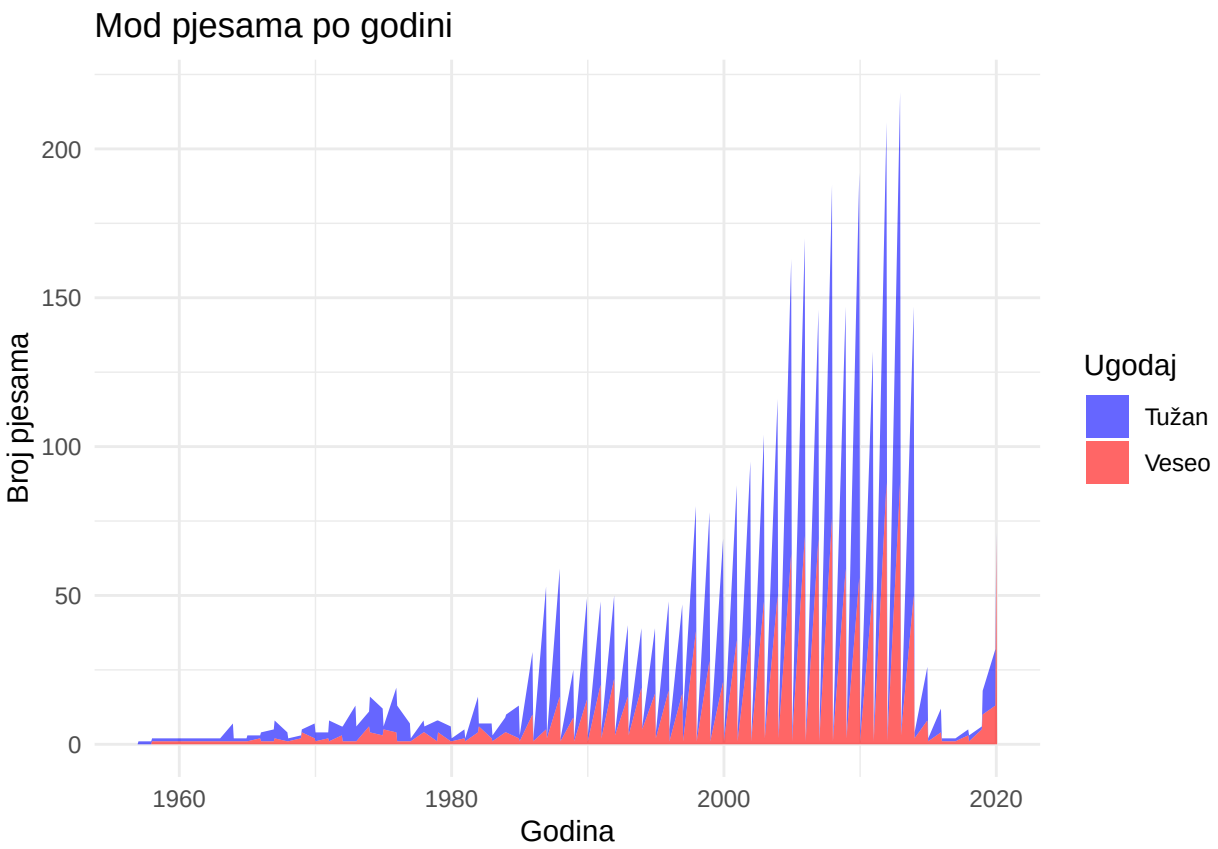

```
## nema    C   C#    D   D#    E    F   F#    G   G#    A   A#    B
##    0 3246 3833 2632  881 2061 2557 2538 3075 2323 2806 2164 2831
```

Analizirajući internetom zastupljenost tonaliteta, pronašao sam podatak da trećina svih pjesama na svijetu sadrži ključeve: **C#**, **D#**, **G#** i **A#**. Zanimljivost je da u našem skupu taj zbroj iznosi približno isto.

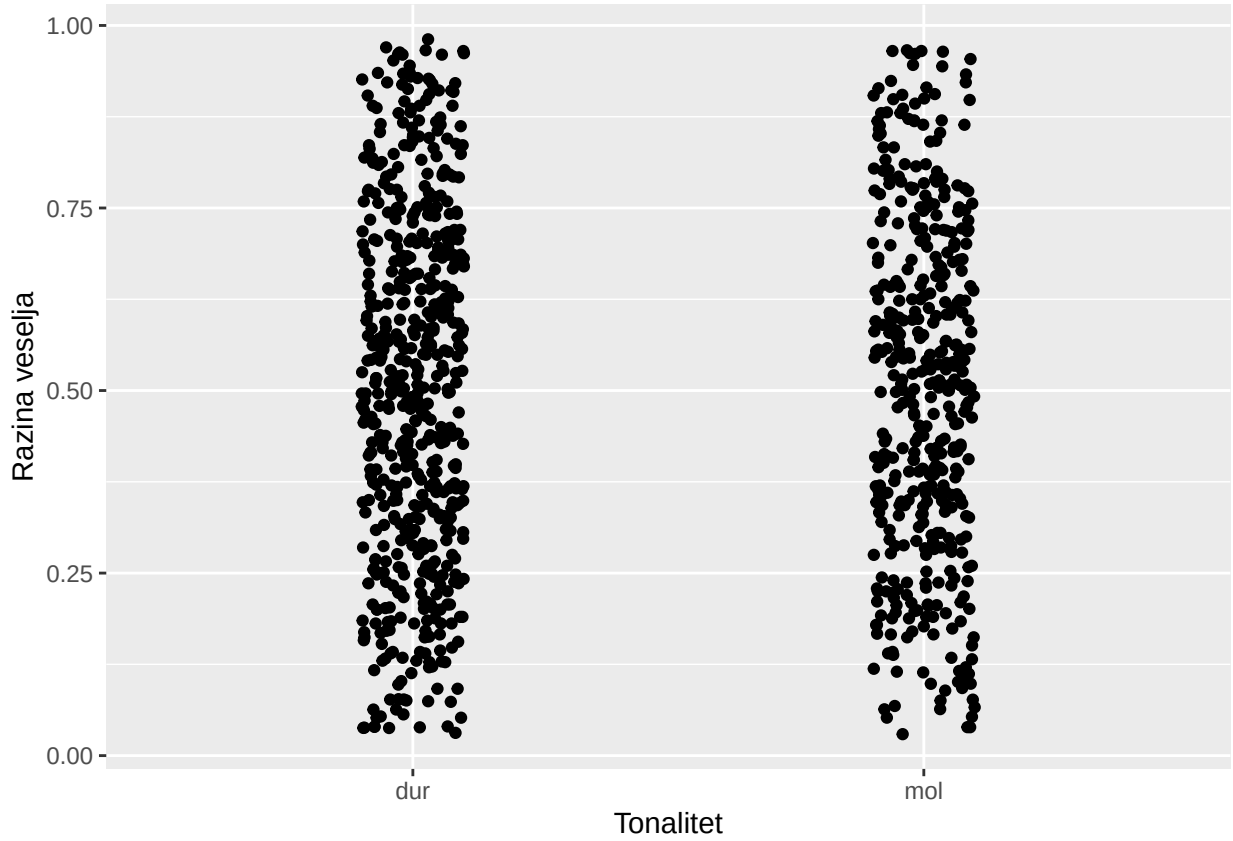
```
paste(round(sum(key_count["C#"], key_count["D#"], key_count["G#"], key_count["A#"]) /
  sum(key_count[1:13]) * 100, 2), "%")
```

```
## [1] "29.73 %"
```

Pogledajmo koliko se po desetljećima stvaralo vesele, a koliko tužne glazbe:



Možemo reći da je generalno trend da se više stvara **tužne** glazbe. No, koliko je to zapravo istinito. Na satovima glazbenog su nas jednostavno učili kako odabir tonaliteta određuje je li glazbeno djelo veselo ili tužno. Dur smo poistovjećivali sa sretnom, a mol sa tužnom glazbom.

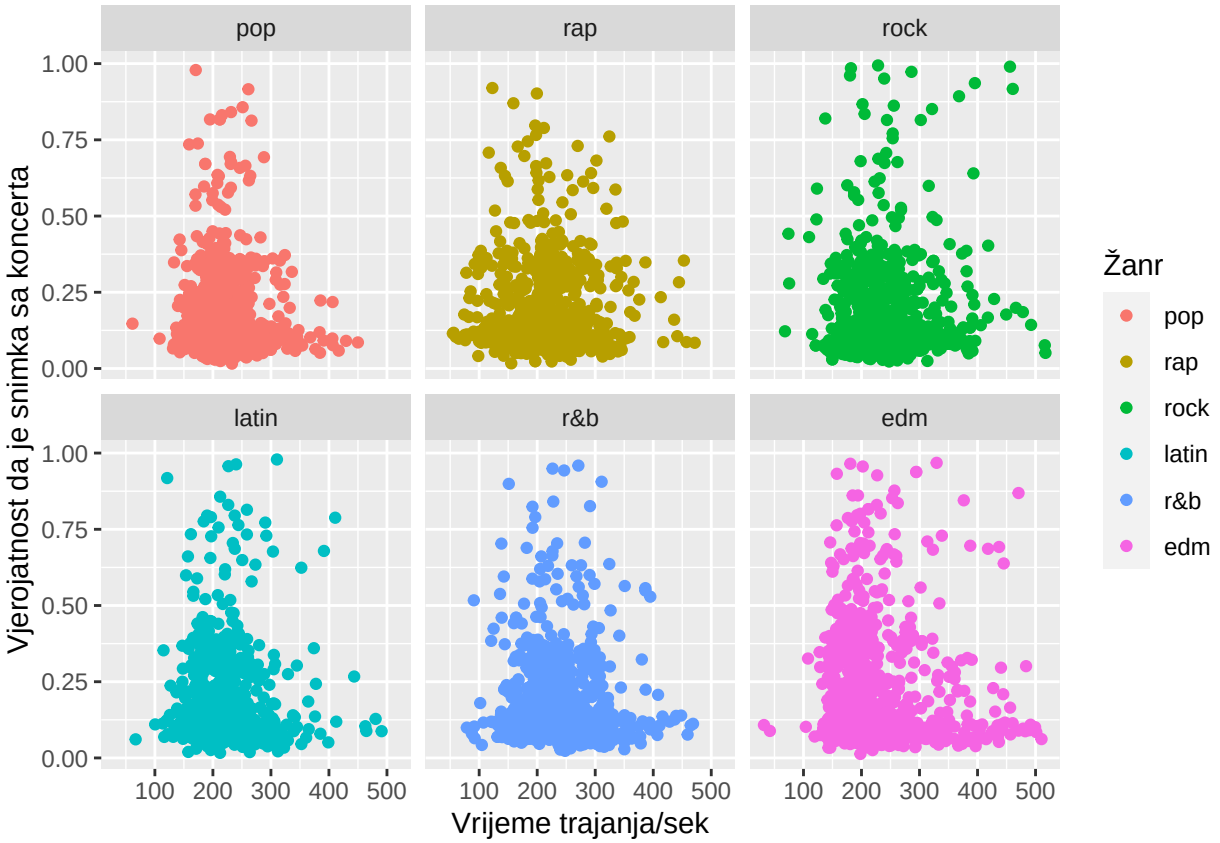


Vidimo da razine veselja i u mol ljestvici i u dur ljestvici postižu i visoke i niske razine veselja. Također, naziremo najveću gustoću rezultata oko srednje razine veselja. Vizualno zaključujemo slične rezultate za obje ljestvice. Jesu li nas varali u školi? Provjerimo tezu izradom linearnog regresijskog modela.

```
##
## Call:
## lm(formula = valence ~ mode, data = Spotify)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50525 -0.17925  0.00124  0.18175  0.48575
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.5052452  0.0017665 286.009   <2e-16 ***
## modemol     -0.0004885  0.0026660  -0.183    0.855
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2327 on 30945 degrees of freedom
## Multiple R-squared:  1.085e-06,  Adjusted R-squared:  -3.123e-05
## F-statistic: 0.03357 on 1 and 30945 DF,  p-value: 0.8546
```

Vidimo da je p-vrijednost za varijablu tonaliteta jako velika (0.855) što nas vodi do zaključka da je ovisnost između tonaliteta i razine veselja jako mala.

Za očekivati je da snimke koncerata dulje traju jer izvođač ima interakciju s publikom te brojni bendovi ubacuju dugačke solo dionice gitarista/bubnjara.



Očekivali smo da pjesme koje su najduže da će imati najveću razinu liveness-a, međutim taj trend se jedino nazire za žanrove rock i edm-a, dok kod ostalih prevladava da pjesme srednje duljine trajanja karakterizira najveća razina tog parametra.

Fun facts:

- Koliko ima ukupno dueta

```
## [1] "Ukupni broj dueta: 3293"
```

- Najdulje prosječno trajanje pjesme

```
## [1] "Newcleus"
```

Iz podataka se vidi kako izvođač Newcleus ima prosječno najdulje pjesme te prosjek trajanja njegovih pjesama je 8 minuta i 40 sekundi.

- Najplesniji album

```
## [1] "Quality Control: Control The Streets Volume 2"
```

Ako ste raspoloženi za ples, preporučujem album "Quality Control: Control The Streets Volume 2"

- "Najužurbaniji" album (najbrži tempo)

```
## [1] "Hola (Remix)"
```

Ako negdje žurite, pustite si album "Hola (Remix)". Sigurno će vas ubrzati. Pazite na ograničenja u prometu :)

- Kralj publike

```
## [1] "Soda Stereo"
```

Ne volite pjevati sami? Pridružite se bendu Soda Stereo i njihovoj publici u jednoj od brojnih snimki koncerata.