

Eksploratorna analiza Spotify skupa podataka

36558993 Ivan Josip Kardum, 36557629 Damjan Crnković

2026-01-31

U ovom izvještaju provedena je eksploratorna analiza skupa podataka "Spotify songs dataset (Kaggle)", koji sadrži oko 30 000 pjesama s opisanim svojstvima poput popularnosti, plesnosti, tempa i tonaliteta. Cilj analize je statističkom obradom i vizualizacijom podataka ispitati odnose između žanra, audio svojstava i popularnosti pjesama te istražiti kako su se te značajke mijenjale tijekom vremena. Također, izrađeni su modeli za predviđanje popularnosti pjesama na temelju njihovih svojstava i žanra.

Pri učitavanju podataka iz CSV datoteke, nakon čega slijedi osnovna analiza, proučavamo strukturu skupa podataka, prvih nekoliko zapisa te sažetak varijabli kako bismo stekli uvid u njihove karakteristike.

```
data <- read_csv("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.

glimpse(data)

## 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> "2oCs0DGTsR098Gh5ZS12Cx", "63rPS0264uRjW1X5E6~
## $ 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, 0, 0, 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~

# head(data, 10)
# summary(data)

```

Skup podataka sadrži 23 varijable koje opisuju pjesme, uključujući audio značajke, žanr, popularnost te informacije o albumima. Ove varijable omogućuju analizu odnosa između svojstava pjesama i njihove popularnosti.

Priprema podataka

Uočavamo da stupac s datumima treba pretvoriti u ispravan tip te da imamo nekoliko kategoričkih varijabli koje treba faktorizirati kako bismo ih po potrebi mogli koristiti za grupiranje u vizualizacijama i analizi.

```

data$track_album_release_date <- as.Date(data$track_album_release_date)

data$mode <- factor(
  data$mode,
  levels = c(0, 1),
  labels = c("Minor", "Major")
)

factcols <- c("playlist_genre", "playlist_subgenre", "playlist_name", "track_artist")
data[factcols] <- lapply(data[factcols], as.factor)

```

1) Kako se audio svojstva razlikuju po žanrovima?

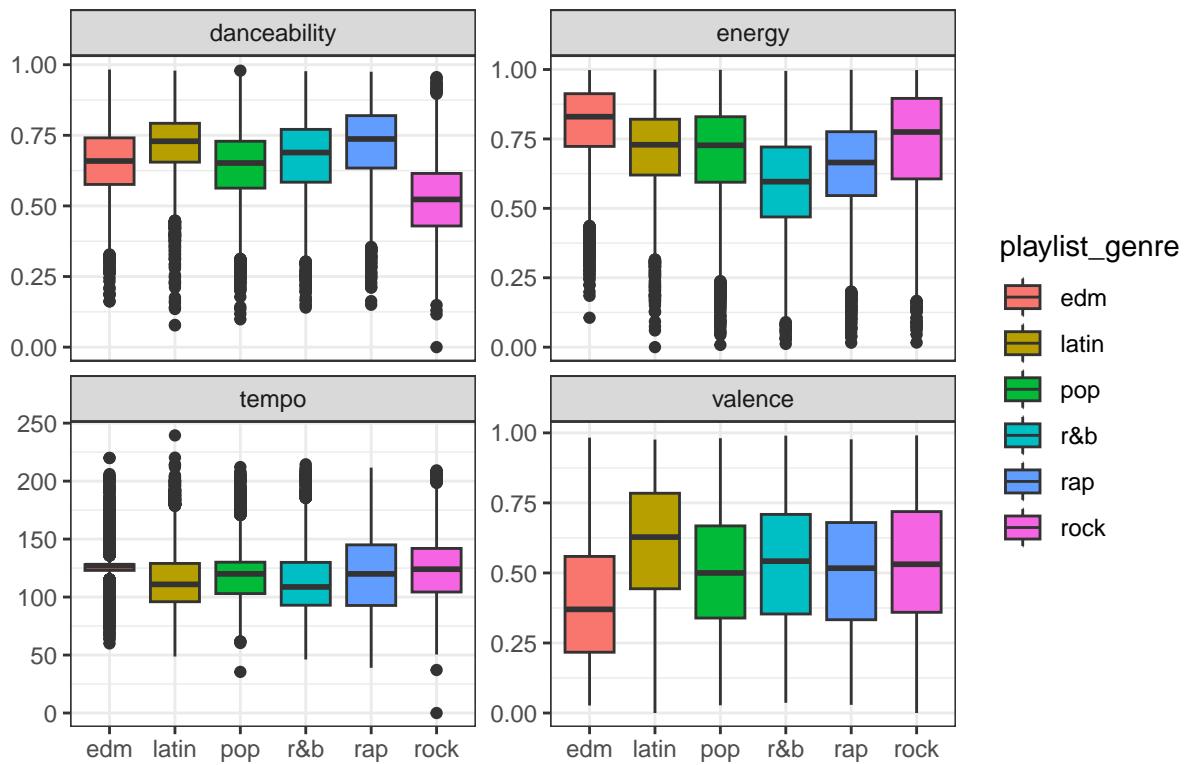
Zanima nas koje audio značajke odlikuju žanrove i koliko se žanrovi razlikuju po tim svojstvima. Provjeravamo odnos energije, plesnosti, pozitivnosti i tempa sa žanrom.

```

data %>%
  dplyr::select(playlist_genre, energy, danceability, valence, tempo) %>%
  pivot_longer(-playlist_genre, names_to = "svojstva", values_to = "vrijednosti") %>%
  ggplot(aes(x = playlist_genre, y = vrijednosti, fill = playlist_genre)) +
  geom_boxplot() +
  facet_wrap(~ svojstva, scales = "free_y") +
  labs(title = "Žanrovi po svojstvima pjesama",
       x = "",
       y = "") +
  theme(legend.position = "none",
        plot.title = element_text(hjust = 0.5)) +
  theme_bw()

```

Žanrovi po svojstvima pjesama



Pregledom grafova vidimo da se žanrovi najviše razlikuju po plesnosti i energiji, gdje su razlike u medijanima vizualno očite, a varijabilnost približno ista (vidljivo po interkvantilnom rangu). Tempo pokazuje male razlike među žanrovima te se čini da nema značajan utjecaj pri klasificiranju pjesme u određeni žanr. Pozitivnost (valence) je također slična među većinom žanrova, uz iznimku EDM žanra koji odstupa s niskom razinom pozitivnosti i latin žanra koji ima blago veću razinu od ostalih. Analiza je provedena na razini playlisti, pri čemu se ista pjesma može pojaviti u više žanrova, što je prihvatljivo jer je cilj analize usporediti karakteristike po žanrovima a ne pojedinim pjesmama.

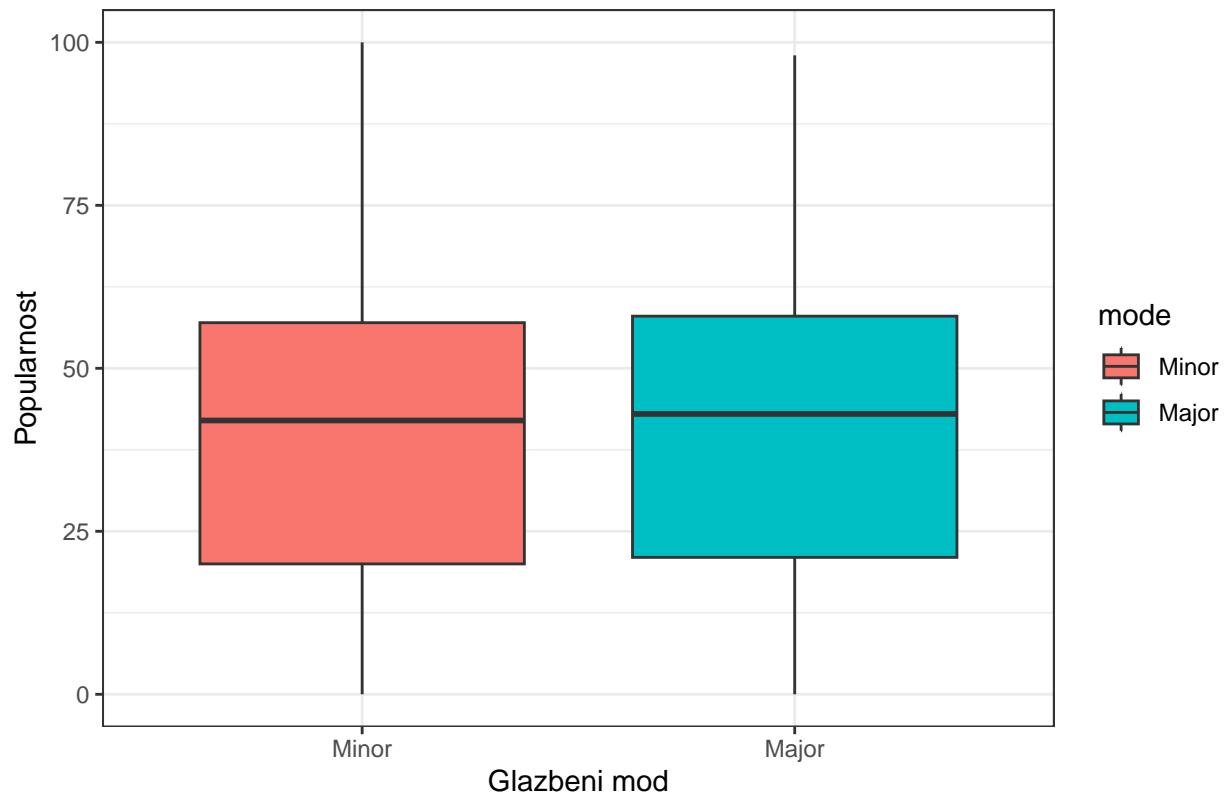
2) Utječe li mode (dur/mol) na popularnost pjesama?

Dur (Major) ljestvice u glazbi zvuče veselije i svijetlijie od pripadajućih mol (Minor) ljestvica, pa analiziramo postoji li razlika u popularnosti pjesama s obzirom na glazbeni mod. Uklonit ćemo višestruka pojavljivanja određenih pjesama kako ne bi negativno utjecala na rezultate.

```
data2 <- distinct(data, track_id, .keep_all = T)

data2 %>% ggplot(aes(x = mode, y = track_popularity, fill = mode)) +
  geom_boxplot() +
  labs(title = "Popularnost pjesama u ovisnosti o glazbenom modu",
       x = "Glazbeni mod",
       y = "Popularnost") +
  theme(legend.position = "none",
        plot.title = element_text(hjust = 0.5)) +
  theme_bw()
```

Popularnost pjesama u ovisnosti o glazbenom modu



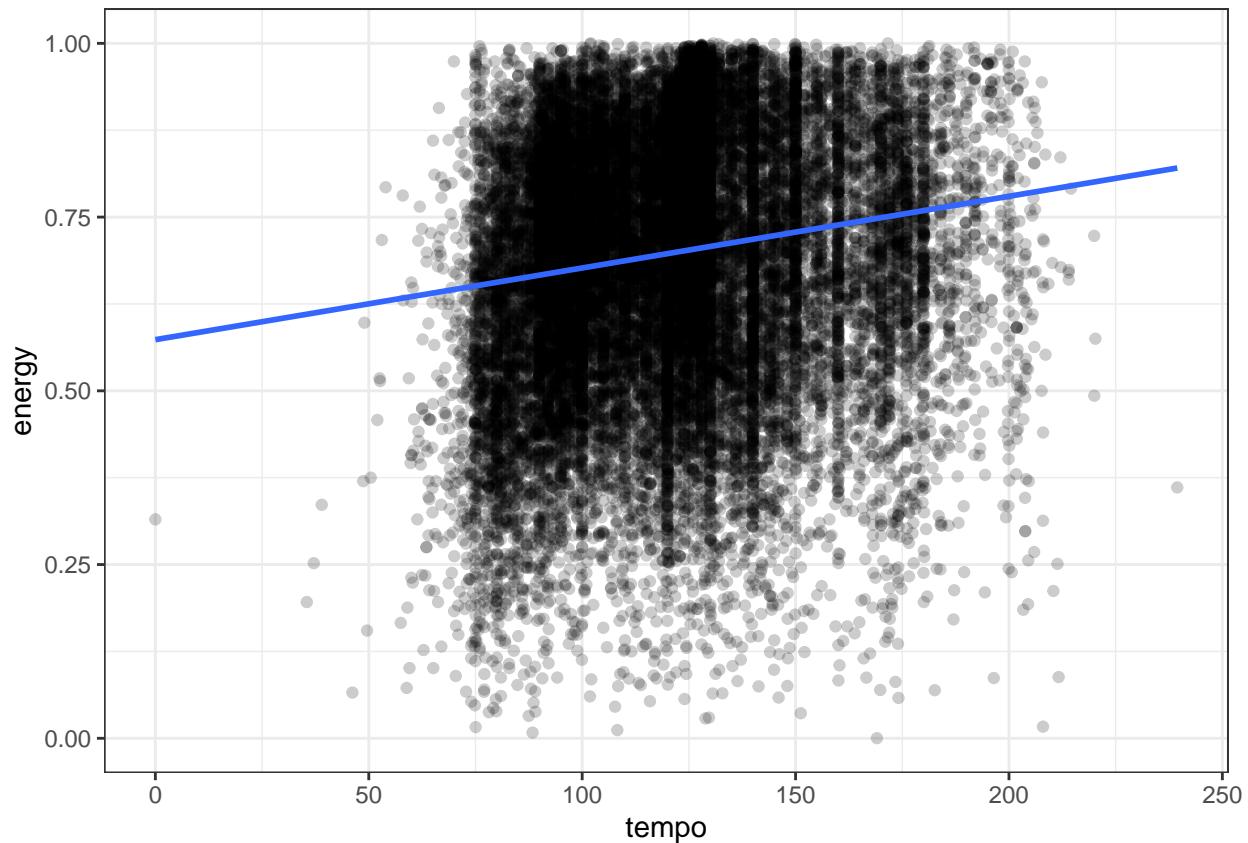
Pregledom grafičkog prikaza ne uočava se jasna razlika u popularnosti između pjesama pisanih u duru i molu. Medijani popularnosti vrlo su slični, a distribucije se u velikoj mjeri preklapaju, što upućuje na to da glazbeni mod sam po sebi nema izražen utjecaj na popularnost pjesama.

3) Je li energija pjesme povezana s tempom i glasnoćom?

Energija pjesme je svojstvo koje je teško objektivno definirati te se često definira pomoću drugih svojstava. Zanima nas postoji li veza između tempa i glasnoće i percipirane energičnosti pjesama. Opet ćemo pri analizi gledati jedinstvene pjesme (bez ponavljanja). U drugom grafu ograničavamo se na 99.9% podataka kako stršeće vrijednosti ne bi remetile izgled grafa.

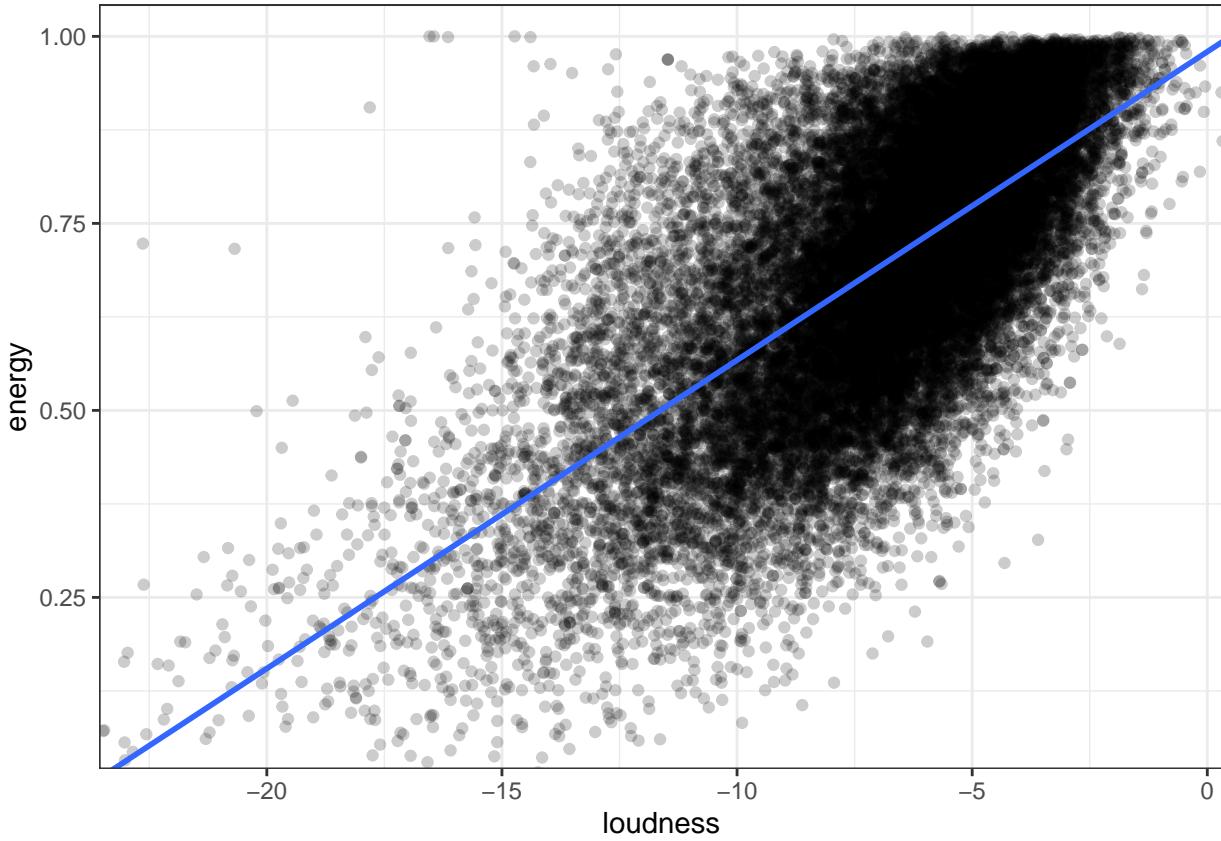
```
ggplot(data2, aes(x = tempo, y = energy)) +
  geom_point(alpha = 0.2) +
  geom_smooth(method = "lm", se = F) +
  theme_bw()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```
ggplot(data2, aes(x = loudness, y = energy)) +
  geom_point(alpha = 0.2) +
  coord_cartesian(
    xlim = quantile(data2$loudness, c(0.001, 0.999)),
    ylim = quantile(data2$energy, c(0.001, 0.999))
  ) +
  geom_smooth(method = "lm", se = F) +
  theme_bw()

## `geom_smooth()` using formula = 'y ~ x'
```



```
model <- lm(energy ~ tempo + loudness, data = data2)
summary(model)
```

```
##
## Call:
## lm(formula = energy ~ tempo + loudness, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.54061 -0.08780  0.00655  0.09096  1.34283 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 0.9048821  0.0042160 214.63   <2e-16 ***
## tempo       0.0005881  0.0000295  19.94   <2e-16 ***
## loudness    0.0407218  0.0002619 155.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1333 on 28353 degrees of freedom
## Multiple R-squared:  0.4727, Adjusted R-squared:  0.4727 
## F-statistic: 1.271e+04 on 2 and 28353 DF,  p-value: < 2.2e-16
```

Linearni regresijski model potvrđuje uočene odnose s grafova. Glasnoća ima snažan i pozitivan utjecaj na energiju pjesme, dok je povezanost između tempa i energije statistički značajna, ali znatno slabijeg intenziteta. Vrijednost R^2 pokazuje da tempo i glasnoća zajedno objašnjavaju oko 47% varijabilnosti energije,

što upućuje na to da energija pjesme ovisi i o drugim svojstvima koja ovdje nisu analizirana. Rezultati su u skladu s očekivanjima, budući da je glasnoća jedna od ključnih komponenti percepcije energičnosti glazbe, iako je utjecaj tempa možda bio manji od očekivanog.

4) Kako su se audio karakteristike pjesama mijenjale kroz vrijeme?

U ovom dijelu analize cilj je ispitati kako su se ključne audio karakteristike pjesama (danceability, energy i valence) mijenjale kroz godine. Budući da nas zanima vremenski trend, prvo je potrebno pripremiti podatke i izdvojiti godinu iz datuma izdanja albuma.

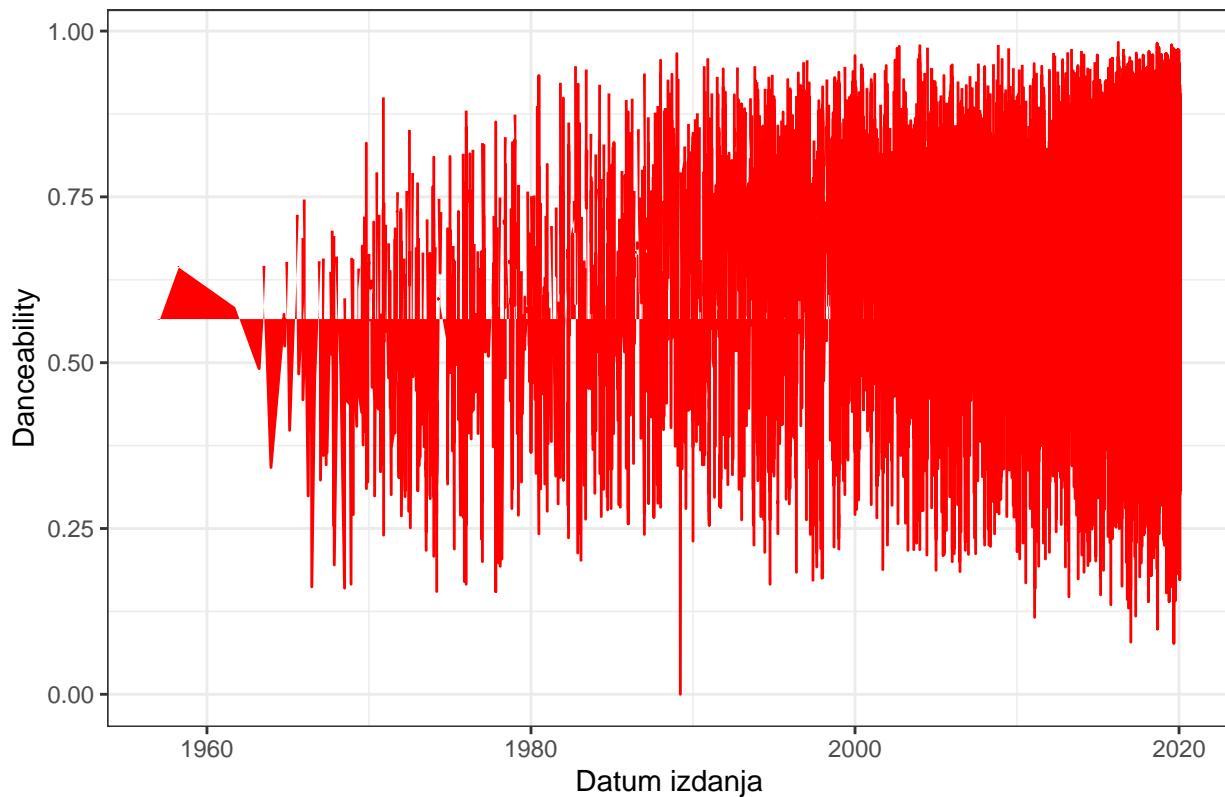
Iz podatkovnog skupa uklanjamo zapise bez poznatog datuma izdanja te iz varijable track_album_release_date izdvajamo godinu izdanja.

```
data_modified <- data %>%
  filter(!is.na(data$track_album_release_date)) %>%
  mutate(release_year = year(track_album_release_date))
```

Kao početni korak, prikazujemo promjenu danceability. Ovakav prikaz pokazuje veliku varijabilnost jer sadrži sve pojedinačne pjesme.

```
ggplot(data_modified, aes(x = track_album_release_date, y = danceability)) +
  geom_line(color = "red") +
  labs(
    title = "Promjena Danceability kroz vrijeme (po pjesmama)",
    x = "Datum izdanja",
    y = "Danceability"
  ) +
  theme_bw()
```

Promjena Danceability kroz vrijeme (po pjesmama)



Iako graf sadrži veliku količinu šuma, služi kao motivacija za agregaciju podataka na godišnjoj razini.

Kako bismo dobili jasniju sliku dugoročnih trendova, agregiramo podatke po godinama i računamo srednje vrijednosti odabranih audio karakteristika:

```
yearly <- data_modified %>%
  group_by(relase_year) %>%
  summarise(
    mean_d = mean(danceability),
    mean_e = mean(energy),
    mean_v = mean(valence)
  )
```

Prije prikaza, pogledajmo broj podataka po desetljeću:

```
data_modified %>%
  mutate(decade = floor(relase_year / 10) * 10) %>%
  count(decade)
```

```
## # A tibble: 8 x 2
##   decade     n
##   <dbl> <int>
## 1 1950      2
## 2 1960     131
## 3 1970     646
## 4 1980     954
## 5 1990    1879
```

```

## 6    2000  3565
## 7    2010 22985
## 8    2020   785

```

Primjećujemo vrlo mali broj podataka u pedesetim i šezdesetim godinama, što može uzrokovati nekonzistentnost i šum na grafičkom prikazu.

U nastavku istovremeno prikazujemo promjene prosječne plesnosti (danceability), energije (energy) i emocionalne pozitivnosti (valence) kroz godine, ograničavajući interval na 1970. godinu nadalje. Podaci su dodatno zaglađeni kako bi se smanjio utjecaj godišnjih oscilacija i jasno istaknuo dugoročni trend, pri čemu se i dalje zadržava pregled stvarnih promjena kroz vrijeme.

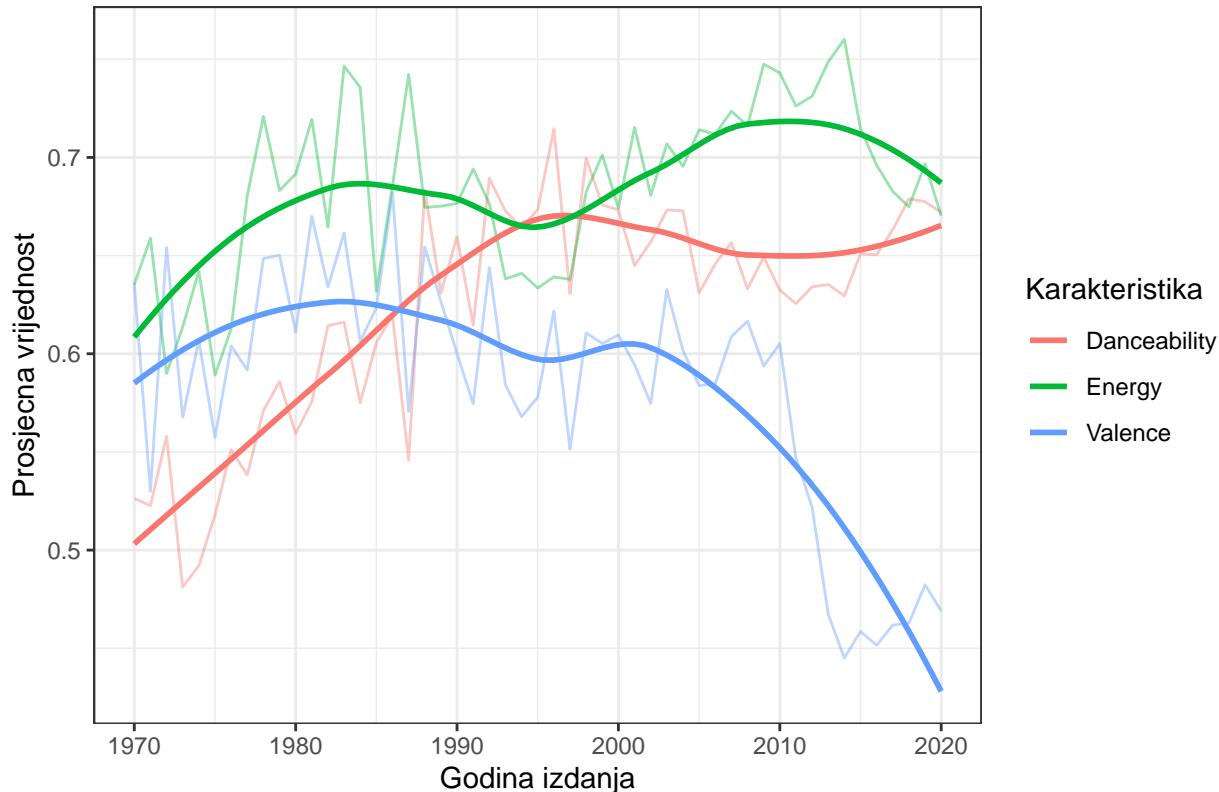
```

ggplot(yearly %>% filter(release_year >= 1970), aes(x = release_year)) +
  geom_line(aes(y = mean_d, color = "Danceability"), alpha = 0.4) +
  geom_line(aes(y = mean_e, color = "Energy"), alpha = 0.4) +
  geom_line(aes(y = mean_v, color = "Valence"), alpha = 0.4) +
  geom_smooth(aes(y = mean_d, color = "Danceability"), se = FALSE) +
  geom_smooth(aes(y = mean_e, color = "Energy"), se = FALSE) +
  geom_smooth(aes(y = mean_v, color = "Valence"), se = FALSE) +
  labs(
    title = "Prosječne audio karakteristike pjesama kroz godine (1970 nadalje)",
    x = "Godina izdanja",
    y = "Prosječna vrijednost",
    color = "Karakteristika"
  ) +
  theme_bw()

## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'

```

Prosječne audio karakteristike pjesama kroz godine (1970 nadalje)



Plesnost (danceability) je doživjela znatan porast u razdoblju od 1970. do 1995. godine, što je vjerojatno povezano s popularnošću plesnih i disco ritmova tog doba, dok nakon 1995. dolazi do stagnacije.

Energija (energy) postepeno raste kroz cijelo promatrano razdoblje, pokazujući kontinuirani trend intenzivnijih i dinamičnijih produkcija u pjesmama.

Emocionalna pozitivnost (valence) od 2000. nadalje strmoglavo opada, što sugerira da novije pjesme, iako ritmički i energetski snažnije, postaju emocionalno ozbiljnije ili manje vedre.

5) Koje varijable utječu na popularnost pjesme?

Cilj ovog dijela analize je istražiti koje karakteristike pjesme utječu na njezinu popularnost. Koristimo **track_popularity** kao ciljnu (zavisnu) varijablu, a sve relevantne numeričke i kategoriske varijable kao prediktore. Prvo primjenjujemo tradicionalnu linearnu regresiju, zatim kompleksnije neuronske metode kako bismo identificirali najbolji model za predikciju.

Prvi korak u izradi prediktivnog modela je podjela podataka na trening set i test set. Trening set koristimo za učenje modela, a test set za provjeru koliko dobro model radi na novim, nepoznatim podacima.

```
set.seed(1234)
train_size <- 0.7 * nrow(data) %>% round
train_ind <- sample(1:nrow(data), train_size)

data_train <- data[train_ind,]
data_test <- data[-train_ind,]
```

Prije modeliranja uklanjamo varijable koje ne pomažu u predviđanju popularnosti, poput ID-eva, imena i datuma.

```
data_train <- data_train %>%
  dplyr::select(-track_id, -track_name, -playlist_name, -track_album_id, -track_album_name, -track_artist_id)
data_test <- data_test %>%
  dplyr::select(-track_id, -track_name, -playlist_name, -track_album_id, -track_album_name, -track_artist_id)
```

Sada koristimo višestruku linearnu regresiju kako bismo procijenili utjecaj svake varijable na popularnost pjesme.

```
lmMod <- lm(track_popularity ~ ., data = data_train)
summary(lmMod)
```

```
##
## Call:
## lm(formula = track_popularity ~ ., data = data_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -59.385 -17.250   3.152  18.531  67.784 
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             6.923e+01  2.055e+00 33.684 < 2e-16 ***
## playlist_genrelatin    8.466e+00  5.871e-01 14.421 < 2e-16 ***
## playlist_genrepop      9.965e+00  5.622e-01 17.726 < 2e-16 ***
## playlist_genrer&b     2.930e+00  6.078e-01  4.820 1.45e-06 ***
## playlist_genrerap      4.579e+00  5.918e-01  7.737 1.06e-14 ***
## playlist_genreroock    9.769e+00  6.285e-01 15.542 < 2e-16 ***
## danceability            1.063e+01  1.354e+00  7.851 4.30e-15 ***
## energy                  -2.809e+01  1.457e+00 -19.281 < 2e-16 ***
## key                      6.892e-02  4.405e-02  1.564  0.11772  
## loudness                1.627e+00  7.683e-02 21.179 < 2e-16 ***
## modeMajor               1.101e-01  3.242e-01  0.340  0.73423  
## speechiness             -1.698e+00  1.771e+00 -0.959  0.33773  
## acousticness            2.787e+00  8.689e-01  3.207  0.00134 ** 
## instrumentalness        -9.126e+00  7.741e-01 -11.790 < 2e-16 ***
## liveness                -3.203e+00  1.041e+00 -3.076  0.00210 ** 
## valence                 -1.066e+00  7.932e-01 -1.344  0.17910  
## tempo                   2.068e-02  6.049e-03  3.419  0.00063 *** 
## duration_ms              -4.566e-05  2.727e-06 -16.743 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.76 on 22965 degrees of freedom
## Multiple R-squared:  0.09029,    Adjusted R-squared:  0.08961 
## F-statistic: 134.1 on 17 and 22965 DF,  p-value: < 2.2e-16
```

Linearni model pokazuje da nekoliko varijabli, poput žanra playlisti, danceability, loudness i instrumentalness, značajno utječe na popularnost pjesama. Neke varijable, poput key, mode i speechiness, nisu statistički značajne i ne doprinose predikciji popularnosti.

Definiramo funkcije za izračun RMSE i R^2 kako bismo ocijenili točnost modela:

```

rmse <- function(pred, stv) {
  sqrt(mean((pred - stv)^2))
}

r_squared <- function(pred, true) {
  1 - sum((pred - true)^2) / sum((true - mean(true))^2)
}

```

Procjenjujemo točnost linearne regresije pomoću RMSE i R² na testnom skupu:

```

data_test$predPopularityLM <- predict(lmMod, data_test)

rmse_lm <- rmse(data_test$predPopularityLM, data_test$track_popularity)
r2_lm <- r_squared(data_test$predPopularityLM, data_test$track_popularity)

cat("Linearna regresija RMSE:", rmse_lm, "\n")

```

Linearna regresija RMSE: 24.00005

```
cat("Linearna regresija R**2:", r2_lm, "\n")
```

Linearna regresija R**2: 0.09059274

Rezultati pokazuju da linearna regresija slabo predviđa popularnost pjesama, s velikom prosječnom pogreškom od oko 24 (skala 1-100). Niska vrijednost R² (0.09) znači da model objašnjava samo mali dio varijacije popularnosti.

Sada koristimo slučajnu šumu za predviđanje popularnosti pjesama, jer nakon linearne regresije želimo poboljšati točnost predikcija. Uz primjenu cross-validation procjenjujemo pouzdanost modela, a zatim predviđamo popularnost na testnom skupu i izračunavamo RMSE i R² za ocjenu točnosti.

```

ctrl <- trainControl(
  method = "repeatedcv",
  number = 5,
  repeats = 2,
  verboseIter = TRUE
)

rfMod <- train(
  track_popularity ~ .,
  data = data_train,
  method = 'ranger',
  tuneLength = 5,
  trControl = ctrl,
  num.trees = 20
)

data_test$predPopularityRF <- predict(rfMod, data_test)

rmse_rf <- rmse(data_test$predPopularityRF, data_test$track_popularity)
r2_rf <- r_squared(data_test$predPopularityRF, data_test$track_popularity)

cat("Slučajna šuma RMSE:", rmse_rf, "\n")

```

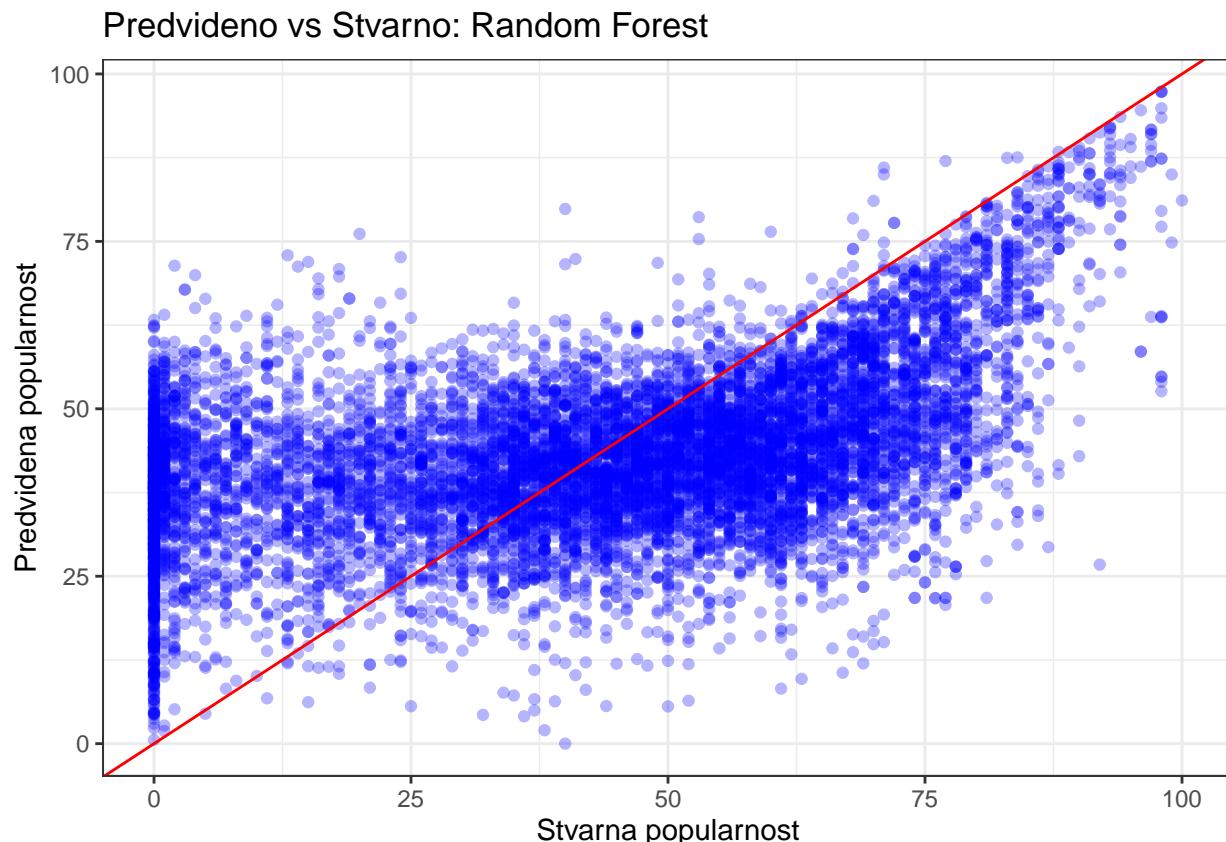
```
## Slučajna šuma RMSE: 22.0656
cat("Slučajna šuma R2:", r2_rf, "\n")
```

```
## Slučajna šuma R2: 0.2312849
```

Model slučajne šume smanjio je RMSE u odnosu na linearnu regresiju, što znači da model bolje predviđa popularnost pjesama u prosjeku. Vrijednost R^2 se više nego udvostručila, što pokazuje da slučajna šuma objašnjava znatno veću varijabilnost popularnosti. To potvrđuje da složeniji model bolje hvata nelinearne i međusobne odnose između varijabli nego model linearne regresije.

Vizualizirajmo za kraj usporedbu stvarnih i predviđenih vrijednosti popularnosti pomoću modela slučajne šume.

```
ggplot(data_test, aes(x = track_popularity, y = predPopularityRF)) +
  geom_point(alpha = 0.3, color = "blue") +
  geom_abline(intercept = 0, slope = 1, color = "red") +
  labs(
    title = "Predviđeno vs Stvarno: Random Forest",
    x = "Stvarna popularnost",
    y = "Predviđena popularnost"
  ) +
  theme_bw()
```



Predviđene vrijednosti popularnosti su prilično raspršene; za mnoge pjesme s popularnošću blizu nule model predviđa znatno više vrijednosti. Iako se neke točke grupiraju oko dijagonale, odstupanja su velika, što se slaže s visokim vrijednostima RMSE i umjerenim R^2 .

Zaključak

1) Kako se audio svojstva razlikuju po žanrovima?

Najveće razlike među žanrovima vidljive su u plesnosti i energiji, dok tempo i valence ostaju relativno slični, uz manje odstupanje određenih žanrova poput EDM-a i latina.

2) Utječe li mode (dur/mol) na popularnost pjesama?

Popularnost pjesama ne ovisi značajno o modu, jer se medijani i distribucije dur i mol pjesama uvelike preklapaju.

3) Je li energija pjesme povezana s tempom i glasnoćom?

Glasnoća snažno utječe na energiju pjesme, tempo ima slabiji, ali statistički značajan utjecaj, a zajedno objašnjavaju čak 47% varijabilnosti energije, što pokazuje da ove dvije karakteristike imaju značajan utjecaj na percepciju energičnosti pjesme.

4) Kako su se audio karakteristike pjesama mijenjale kroz vrijeme?

Plesnost je značajno porasla do 1995. godine, energija postupno raste kroz cijelo razdoblje, dok emocionalna pozitivnost od 2000. nadalje znatno opada, pokazujući trend ritmički i energetski snažnijih, ali emocionalno ozbiljnijih pjesama.

5) Koje varijable utječu na popularnost pjesme?

Linearni model pokazuje da žanr playlisti, danceability, loudness i instrumentalness značajno utječu na popularnost pjesama, dok key, mode i speechiness nemaju značajan utjecaj. Model slučajne šume bolje hvata nelinearne i međusobne odnose među varijablama te objašnjava znatno veću varijabilnost popularnosti.