

Eksploratorna analiza Spotify skupa podataka

36558993 Ivan Josip Kardum, 36557629 Damjan Crnković

2026-01-31

Pri učitavanju podataka iz csv datoteke treba pripaziti na separator i vrstu encodinga.

```
data <- read.csv("spotify_songs.csv", sep = ",", fileEncoding = "ISO-8859-1")

glimpse(data)
head(data, 10)
summary(data)
```

Preparacija 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.

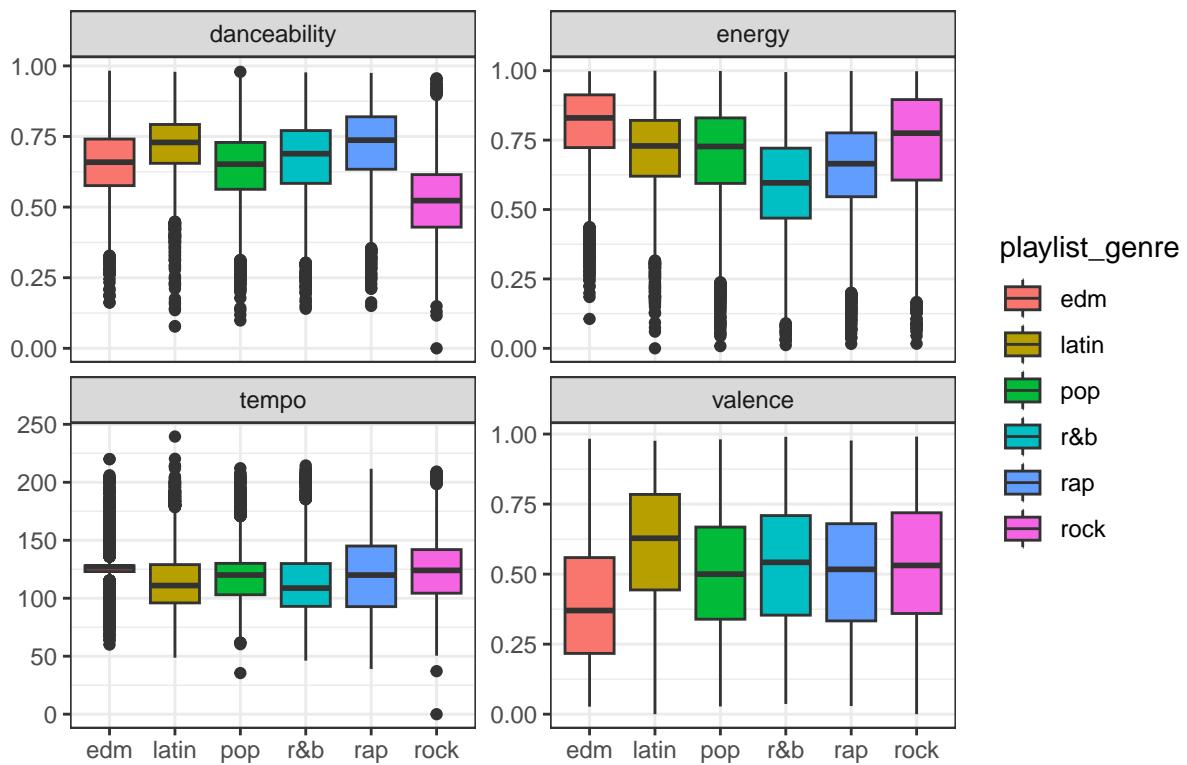
```
data %>%
  select(playlist_genre, energy, danceability, valence, tempo) %>%
  pivot_longer(-playlist_genre, names_to = "svojstva", values_to = "vrijednosti") %>%
  ggplot(aes(x = playlist_genre, y = vrijednosti, fill = playlist_genre)) +
  geom_boxplot() +
  facet_wrap(~ svojstva, scales = "free_y") +
  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 svijetlige 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")

```

```

y = "Popularnost") +
theme(legend.position = "none",
      plot.title = element_text(hjust = 0.5)) +
theme_bw()

```



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

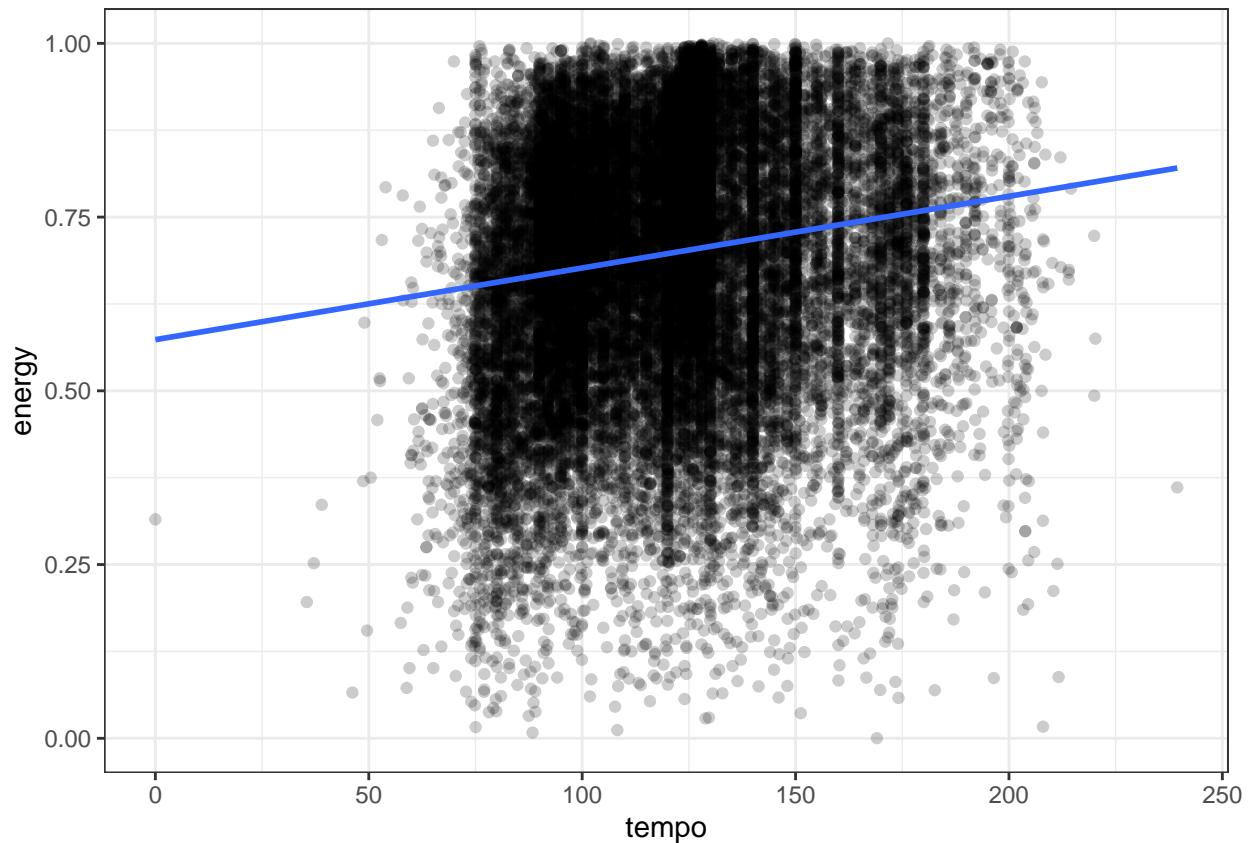
Energija pjesme je svojstvo koje je teško objektivno definirati te se često definira pomoću drugih svojstava. Zanima nas postoji li veza između tempa i glasnoće i percipirane energičnosti pjesama. Opet ćemo pri analizi gledati jedinstvene pjesme (bez ponavljanja). U drugom grafu ograničavamo se na 99.9%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

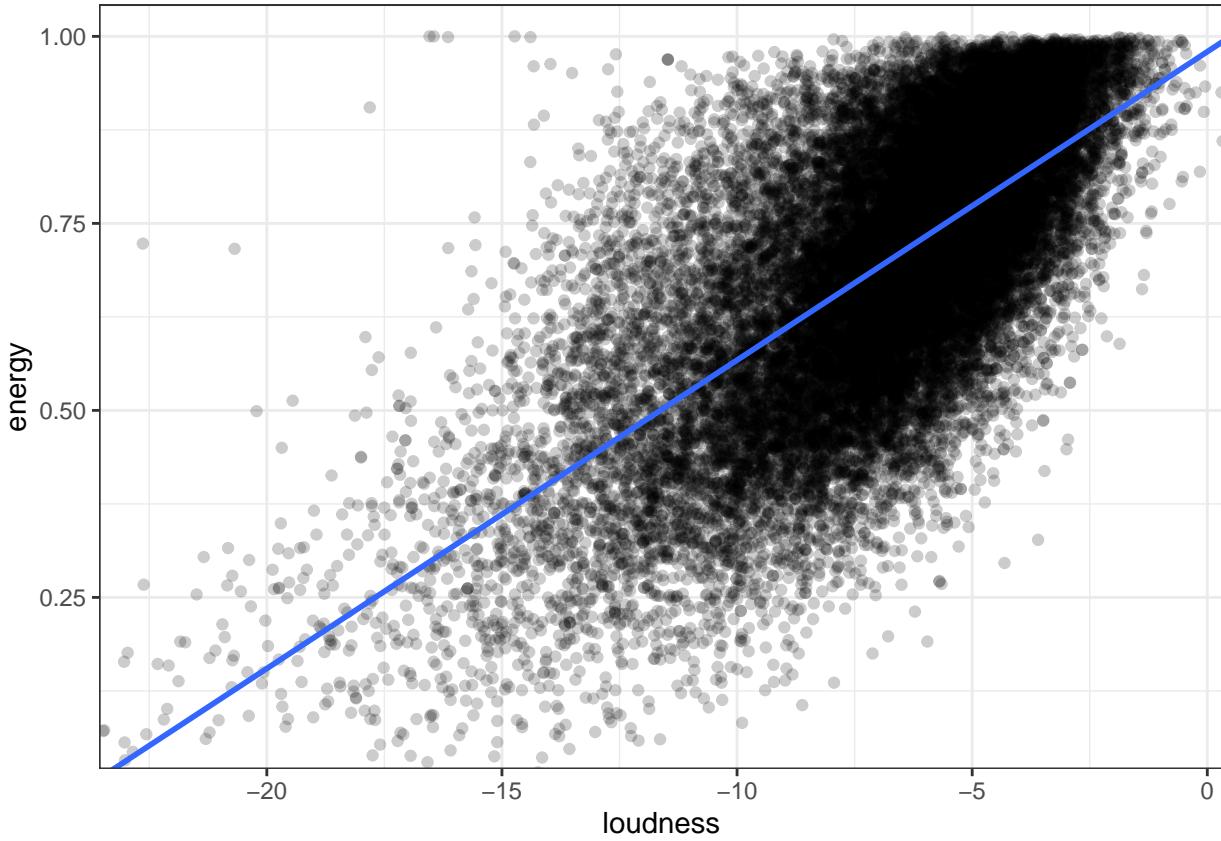
```

```
## `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.

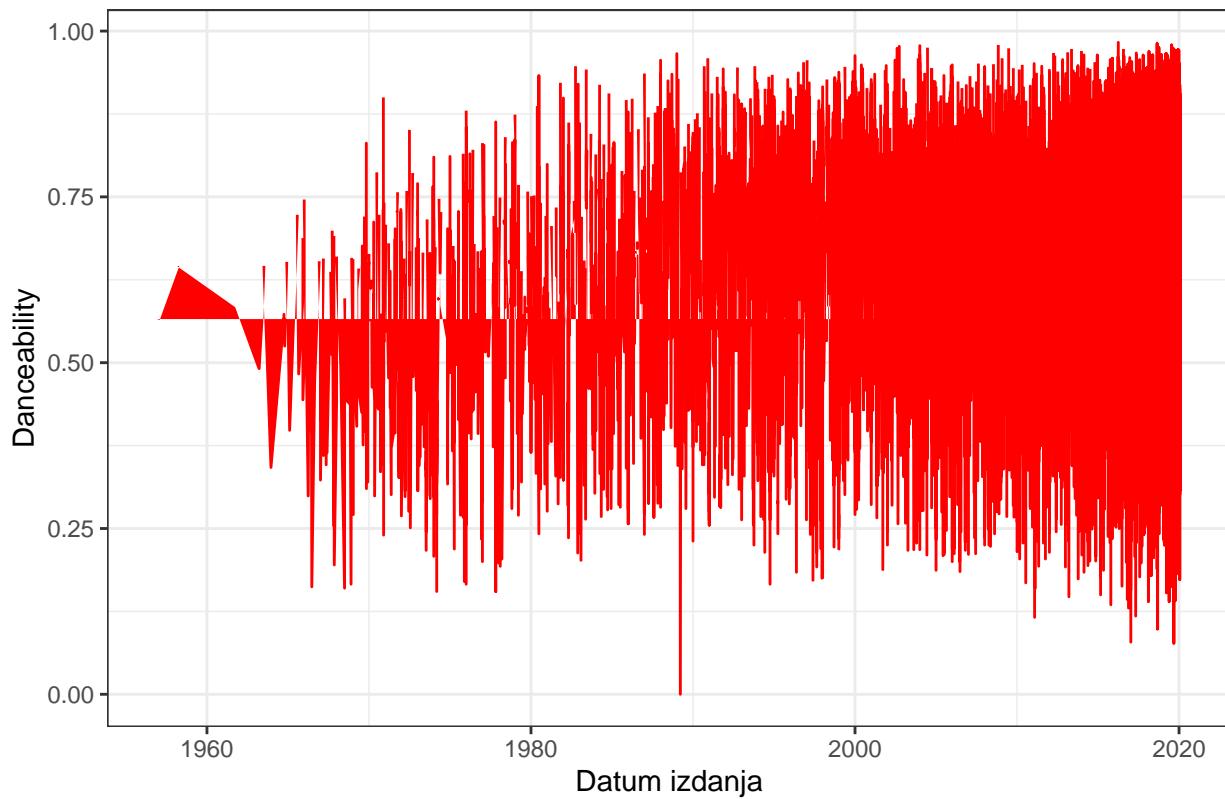
Iz podatkovnog skupa uklanjamo zapise bez poznatog datuma izdanja te iz varijable track_album_release_date izdvajamo godinu izdanja.

```
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)
```

```
##   decade     n
## 1   1950      2
## 2   1960     131
## 3   1970     646
## 4   1980    954
## 5   1990   1879
## 6   2000   3565
## 7   2010  22985
```

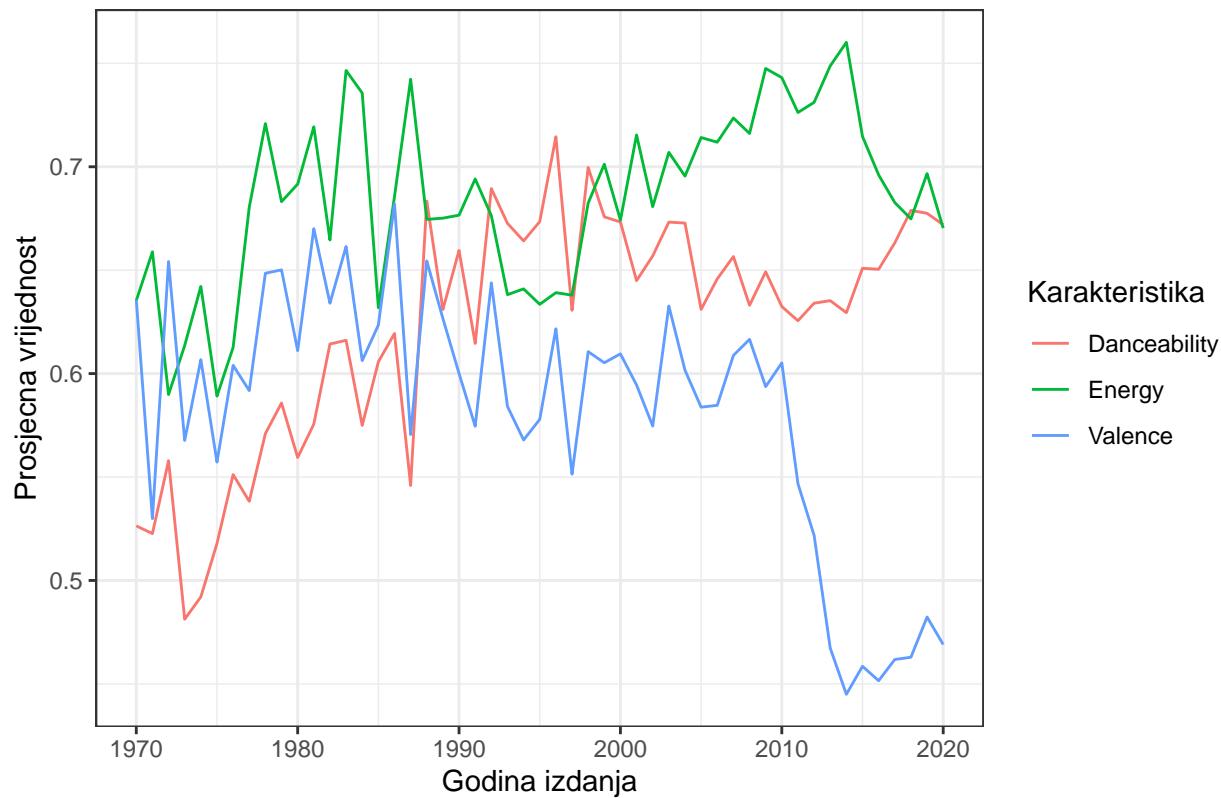
8 2020 785

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

U nastavku istovremeno prikazujemo promjene prosječne danceability, energy i valence kroz godine, ograničavajući interval na 1970. godinu nadalje.

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

Prosječne audio karakteristike pjesama kroz godine (1970 nadalje)



Podaci su stabilni, s vidljivim, iako sporim, ali svakako postojanim rastom mjera danceability i energy, dok je valence doživio nagli pad unazad od 2010. godine. Ovo možemo interpretirati kao da emocionalna pozitivnost novijih pjesama značajno opada.