Eksploratorna analiza podataka

Podatkovni skup “Simpsons episodes”

Dominik Pavelić 0036543749

2024-01-25

```
library(tidyverse)
library(ggplot2)
library(dplyr)
library(readr)
knitr::opts_chunk$set(results = 'hold')
```

Učitavanje i priprema podataka

```
simpsons_episodes <- read_csv("./Data/simpsons_episodes.csv", show_col_types = F)
missing <- colSums(is.na(simpsons_episodes))
missing
```

```
##                     id                  title      original_air_date
##                      0                      0                      0
##        production_code                 season       number_in_season
##                      0                      0                      0
##       number_in_series us_viewers_in_millions                  views
##                      0                      6                      4
##            imdb_rating            imdb_votes              image_url
##                      3                      3                      4
##              video_url
##                      4
```

```
simpsons_episodes <- drop_na(simpsons_episodes)
simpsons_episodes %>% select(-image_url, -video_url) -> simpsons_episodes
```

Podatkovni skup ”Simpsons episodes” sadrži 600 redaka i 13 stupaca. Stupci podatkovnog skupa su:

1. id: jedinstveni identifikator epizode
2. title: naslov epizode
3. original air date: datum prvog emitiranja
4. production code: produkcijski kod epizode
5. season: sezona kojoj epizoda pripada

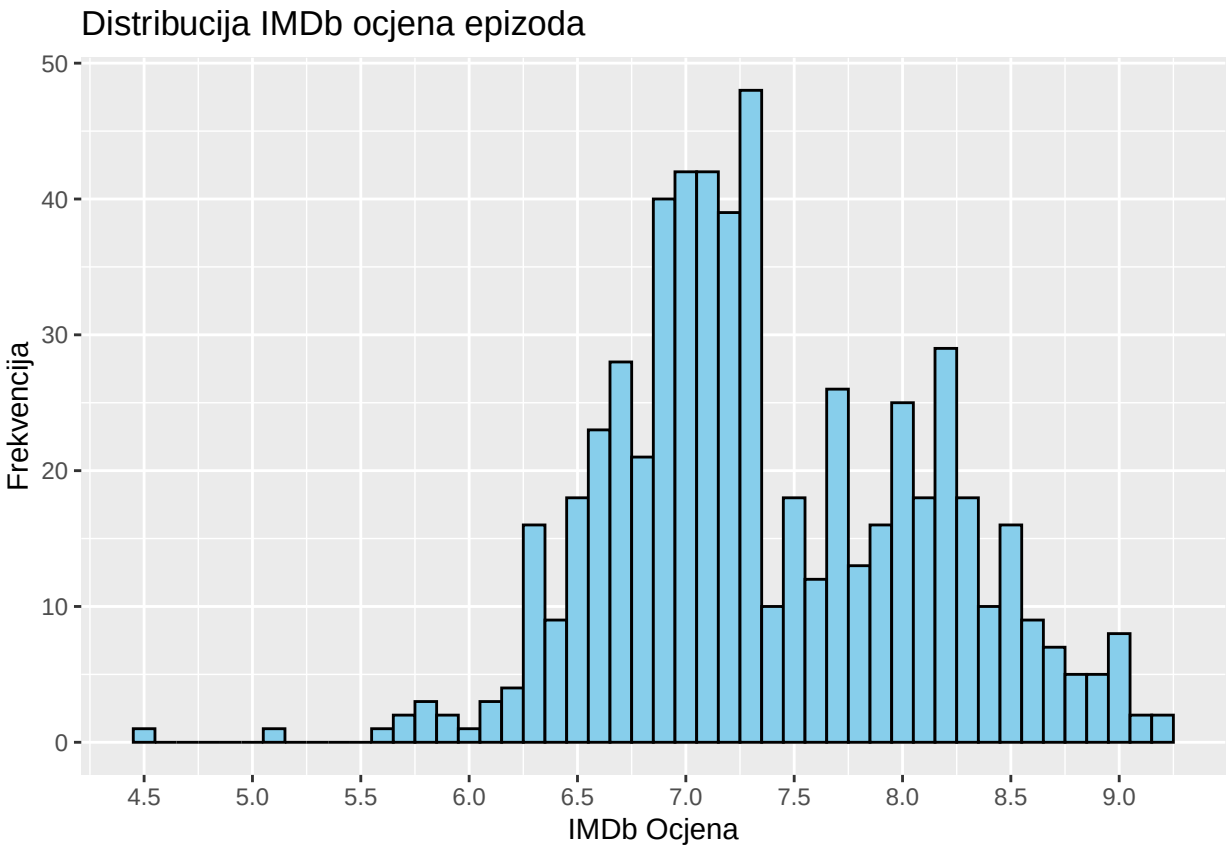
6. number in season: redni broj epizode u sezoni
7. number in series: redni broj epizodfe u cijeloj seriji
8. us viewers in millions: broj gledatelja u SAD-u u milijunima
9. views: broj geldatelja cijelom svijetu
10. imdb rating: IMDb ocjena epizode
11. imdb votes: broj glasova dobivenih na IMdb-u
12. image url: URL slike epizode
13. video url: URL videa epizode

U podatkovnom skupu postoje nedostaju´ci podatci u stupcima "us viewers in millions", "views", "imdb rating", "imdb votes", "image url" i "video url" , od kojih ve´cina nedostaje u stupcima "image url" i "video url", no ti stupci nisu ključni za ovu analizu.

Vizualizacije

Distribucija IMDb ocjena

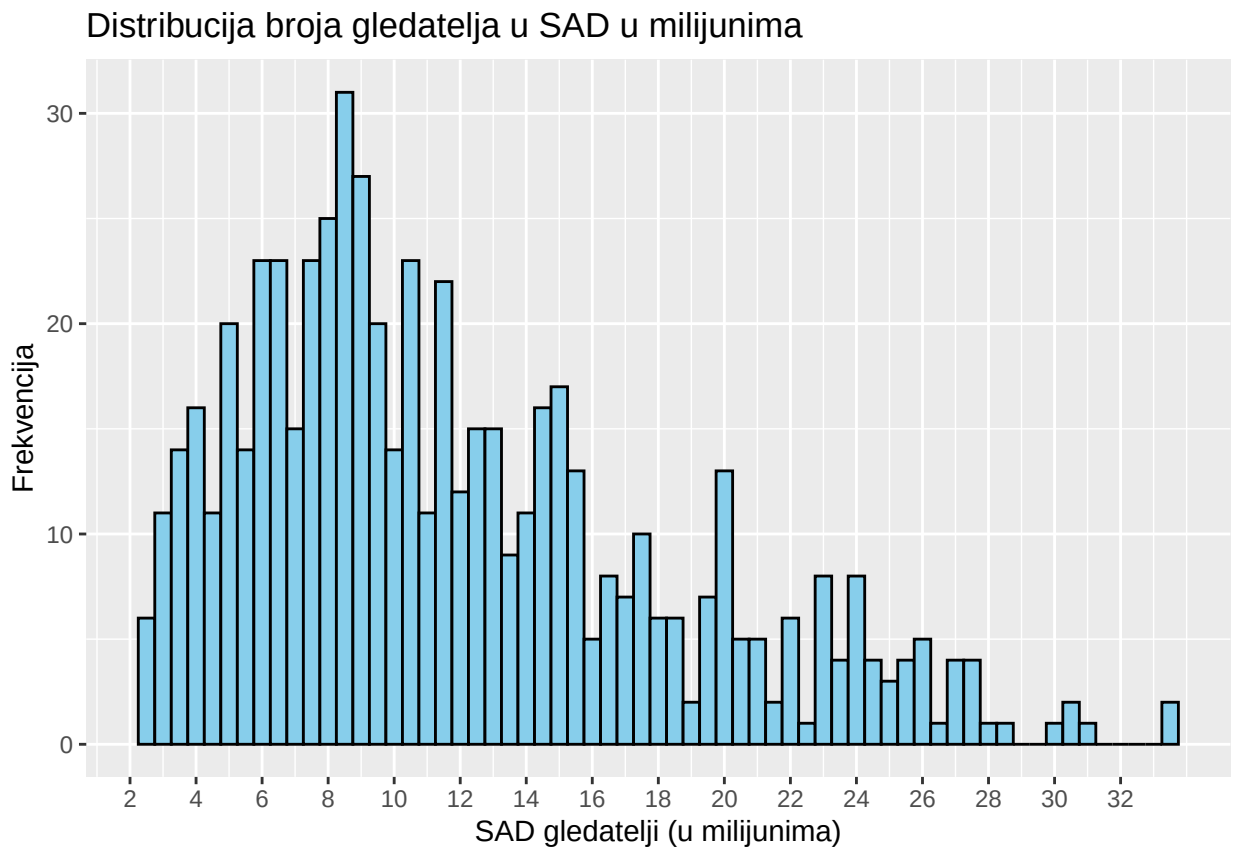
```
ggplot(simpsons_episodes, aes(x = imdb_rating)) +
  geom_histogram(binwidth = 0.1, fill = "skyblue", color = "black") +
  scale_x_continuous(breaks = seq(0, max(simpsons_episodes$imdb_rating), by = 0.5))+
  labs(title="Distribucija IMDb ocjena epizoda", x="IMDb Ocjena", y="Frekvencija")
```



Visina stupaca ukazuje na broj epizoda s određenim ocjenama. Najviši stupci su koncentrirani oko ocjena između 6.5 i 7.5, što sugerira da većina epizoda ima ocjenu u tom rasponu. Epizode s vrlo visokim (blizu 9) ili vrlo niskim (ispod 6) su rijetke, što se vidi iz manje visine stupaca na krajnje lijevoj, odnoso desnoj strani grafa.

Distribucija broja gledatelja

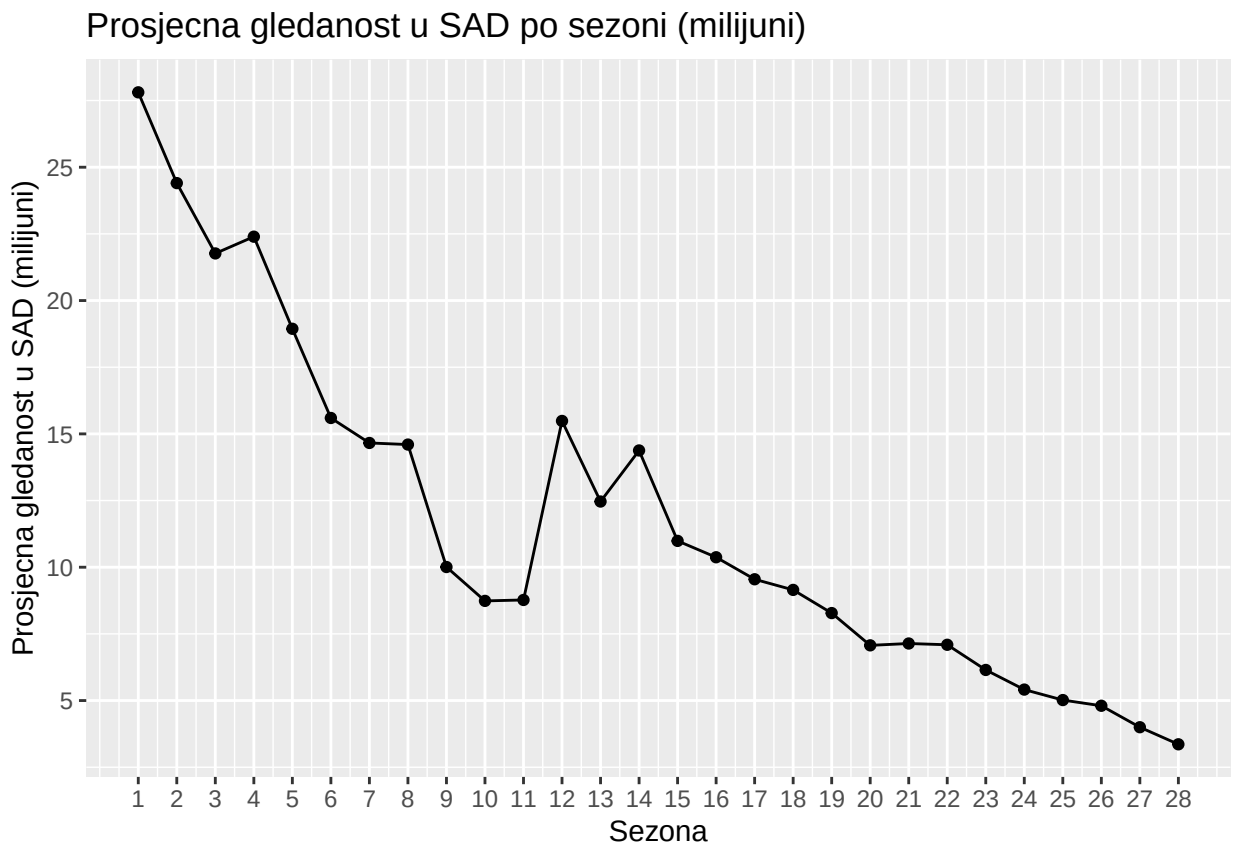
```
ggplot(simpsons_episodes, aes(x = us_viewers_in_millions)) +
  geom_histogram(binwidth = 0.5, fill = "skyblue", color = "black") +
  scale_x_continuous(breaks =
                       seq(0, max(simpsons_episodes$us_viewers_in_millions), by = 2))+
  labs(title = "Distribucija broja gledatelja u SAD u milijunima",
       x = "SAD gledatelji (u milijunima)", y = "Frekvencija")
```



Visina stupaca na grafu ukazuje na frekvenciju određenog raspona broja gledatelja. Stupci su najviši u lijevom dijelu grafa, što ukazuje na to da većina eipzoda ima od 4 do 12 milijuna gledatelja. Postoji nekoliko epizoda s izuzetno visokim brojem gledatelja, što je vidljivo iz manjeg broja visokih stupaca na desnoj strani grafa.

Proječna gledanost po sezoni u SAD

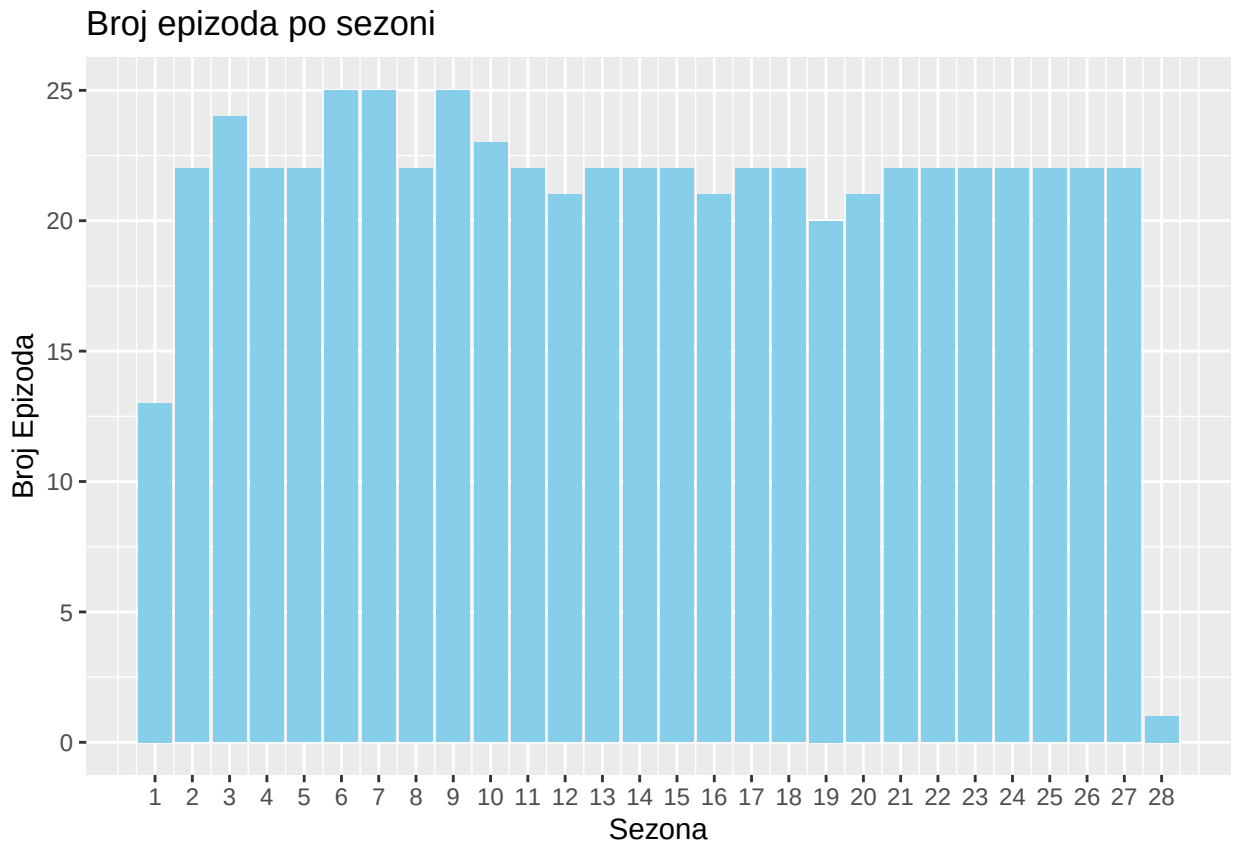
```
average_viewership <- simpsons_episodes %>%
  group_by(season) %>%
  summarise(average_viewership = mean(us_viewers_in_millions))

ggplot(average_viewership, aes(x = season,
                               y = average_viewership)) +
  geom_line(color = "black") +
  geom_point(color = "black") +
  scale_x_continuous(breaks =
                       seq(min(average_viewership$season),
                           max(average_viewership$season))) +
  scale_y_continuous(breaks = seq(0, max(average_viewership$average_viewership), by = 5)) +
  labs(title = "Prosječna gledanost u SAD po sezoni (milijuni)",
       x = "Sezona", y = "Prosječna gledanost u SAD (milijuni)")
```



Graf prikazuje trend pada gledanosti sa svakom novijom sezonom, postoji povećanje gledanosti oko dvanaeste sezone, no nakon toga gledanost i dalje nastavlja opadati.

Broj epizoda po sezoni

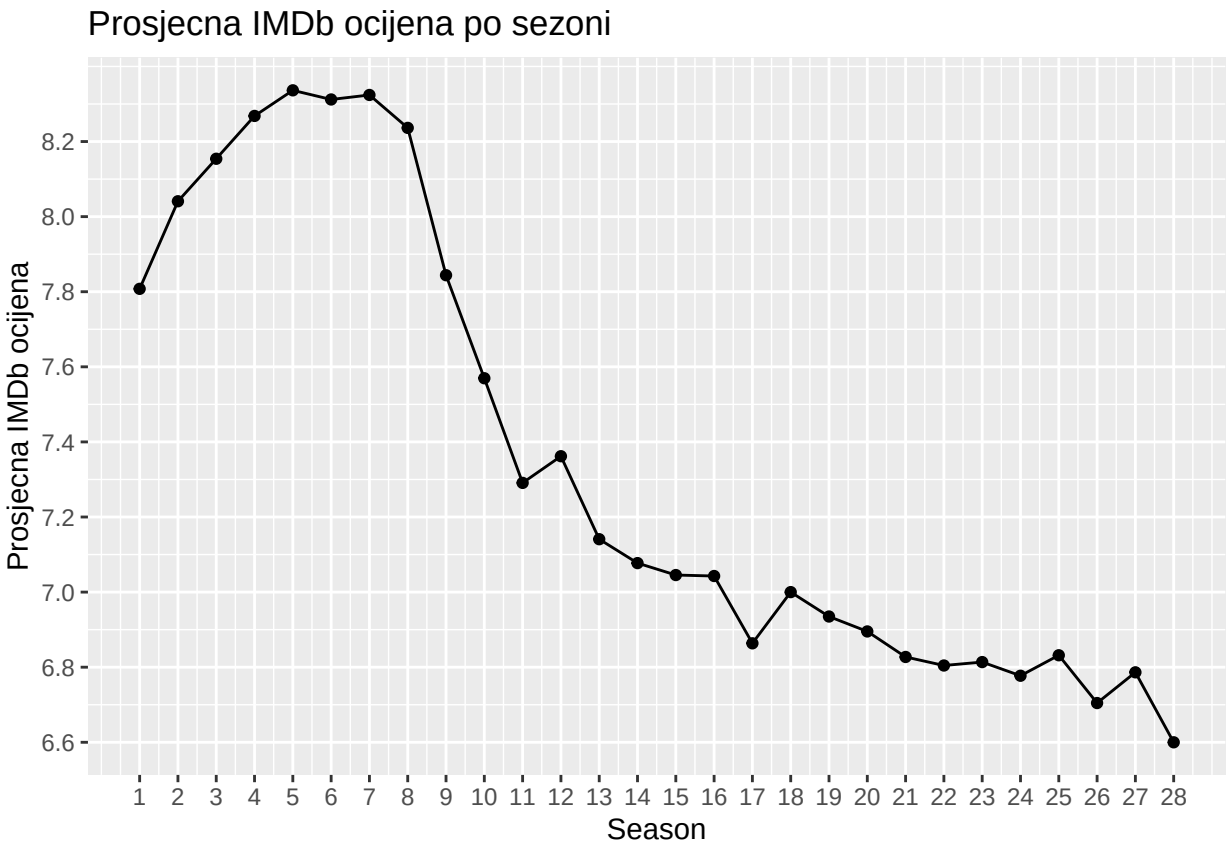
```
episodes_per_season <- simpsons_episodes %>%
  group_by(season) %>%
  summarise(episodes = n())

ggplot(episodes_per_season, aes(x = season, y = episodes)) +
  scale_x_continuous(breaks = seq(min(average_viewership$season), max(average_viewership$season))) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Broj epizoda po sezoni",
       x = "Sezona", y = "Broj Epizoda")
```



Graf prikazuje da je broj epizoda u većini sezona približno jednak. Iako neke sezone imaju veći broj epizoda, a neke manji, razlike u duljini sezona su gotovo zanemarive. Jedidno prva sezona ima gotovo upola manje epizoda od ostalih sezona.

Projsečna IMDb ocijena po sezoni

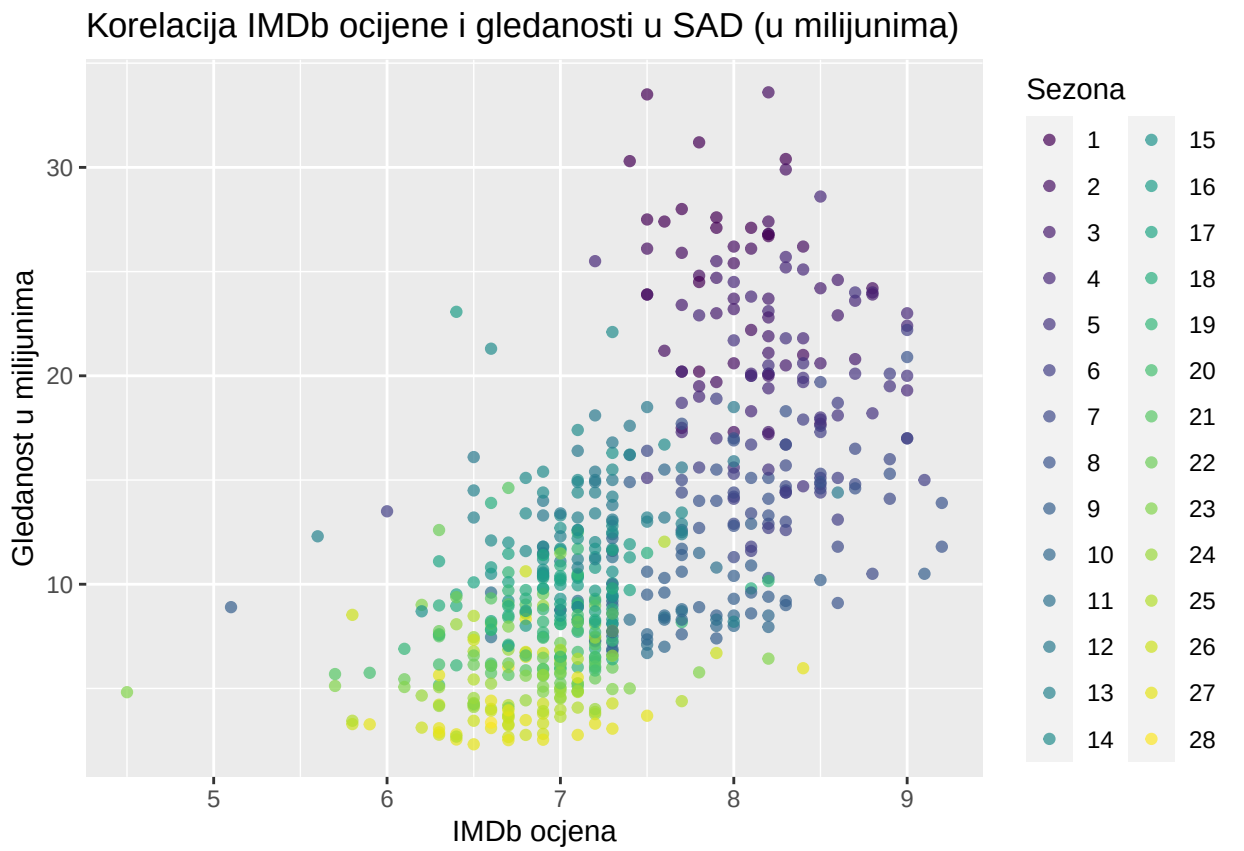
```
average_rating_per_season <- simpsons_episodes %>%
  group_by(season) %>%
  summarise(average_rating = mean(imdb_rating))

ggplot(average_rating_per_season, aes(x = season, y = average_rating)) +
    scale_x_continuous(breaks = seq(min(average_viewership$season), max(average_viewership$season))) +
  scale_y_continuous(breaks = seq(0, max(average_rating_per_season$average_rating), by = 0.2)) +
  geom_line(color = "black") +
  geom_point(color = "black") +
  labs(title = "Prosječna IMDb ocijena po sezoni",
       x = "Season", y = "Prosječna IMDb ocijena")
```



Graf prikazuje proječnu IMDb ocijenu po sezonama. U ranijim sezona ocijena sezone raste te dostiže vrhunac od oko 8.3, a zatim slijedi nagli pad u ocjeni do jedaneaste sezone nakon koje pad polako usporava, no prosječna ocijena sezone se i dalje smanjuje.

Korelacija gledanosti i IMDb ocijene

```
ggplot(simpsons_episodes, aes(x = imdb_rating, y = us_viewers_in_millions, color = as.factor(season))) +
  geom_point(alpha = 0.7) +
  scale_color_viridis_d() +
  labs(title = "Korelacija IMDb ocijene i gledanosti u SAD (u milijunima)",
       x = "IMDb ocjena", y = "Gledanost u milijunima", color = "Sezona")
```



Postoji nejasna korelacija između IMDb ocijena i gledanosti. Iako se čini da epizode s višim IMDb ocijenama imaju veću gledanost, postoje iznimke, što ukazuje na to da kvaliteta epizode i pritom ocijena nisu jedini faktor koji utiječe na gledanost. Različite boje točaka omogućuju vizualizaciju kako su se gledanost i ocijene mijenjale iz sezone u sezonu. Novije sezone imaju manju gledanost i niže ocijene, što se može vidjeti iz koncentracije žutih točaka u sredini grafa blizu x-osi.

Zaključak

Popularnost serije "Simpsoni" mijenja se kroz vrijeme. Neke su sezone bile izrazito popularne s visokom gledanošću i visokim ocijenama, druge su doživjele pad u oba aspekta. Iako postoji određena korelacija između IMDb ocijena i gledanosti epizoda, ona nije uvijek izravna ili konzistentna, što ukazuje na to da gledanost nije isključivo vezana za precepciju kvalitete. Broj epizoda po sezoni pokazuje konzistentnost u produkciji, ali postoje i varijacije koje mogu biti rezultat različitih vanjskih faktora. Varijacije u gledanosti i ocijenama kroz različite sezone mogu biti povezana s mnogo vanjskih faktora kao što su promjene u TV industriji, konkurencija s drugim serijama te prelazak gledatelja na streaming platforme.