Simpsons analiza podataka

Matej Ciglenečki

2020-02-01

Uvod

Odabran dataset sastoji se od 600 zapisa koji opisuju pojedinu Simpsons epizodu sa 13 različitih atributa. U svrhu eksploratorna analiza podataka koristit ćemo sljedećih 8 atributa:

- `title` - ime epizode
- `original_air_date` - datum emitiranje epizode
- `season` - broj sezone u kojoj je sadržana epizoda
- `number_in_season` - redni broj epizode u sezoni
- `number_in_series` - redni broj epizode u seriji
- `us_viewers_in_millions` - broj američkih gledatelja
- `imdb_rating` - imdb ocjena epizode
- `imdb_votes` - broj ocjenjivača epziode

Cilj je pronaći smislene odnose tih atributa koji opisuju trendove u seriji.

Učitavanje podataka

Učitavanje podataka obavljamo funkcijom `read_csv` pošto je dataset zapisan .csv formatu a zatim ćemo pregledati izgled odabranih atributa

```
simpsons <- read_csv("data/dataset_simpsons.csv")
simpsons %>%
  select (-id, -production_code, -views, -video_url, -image_url) %>%
  glimpse(.)
```

```
## Observations: 600
## Variables: 8
## $ title                  <chr> "Homer's Night Out", "Krusty Gets Busted", "...
## $ original_air_date      <date> 1990-03-25, 1990-04-29, 1990-10-11, 1990-11...
## $ season                 <dbl> 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3,...
## $ number_in_season       <dbl> 10, 12, 1, 4, 6, 8, 10, 13, 15, 17, 19, 22, ...
## $ number_in_series       <dbl> 10, 12, 14, 17, 19, 21, 23, 26, 28, 30, 32, ...
## $ us_viewers_in_millions <dbl> 30.3, 30.4, 33.6, 26.1, 25.4, 26.2, 24.8, 26...
## $ imdb_rating            <dbl> 7.4, 8.3, 8.2, 8.1, 8.0, 8.4, 7.8, 8.0, 8.2,...
## $ imdb_votes             <dbl> 1511, 1716, 1638, 1457, 1366, 1522, 1340, 13...
```

Filtriranje podataka

Prvi korak je izbacivanje nepotrebnih atributa koji neće biti korišteni u analizi. Zatim izbacujemo sve redove koji sadrže `na`, `NaN` ili `null` vrijednost za bilo koji atribut.

```
# Remove unnecessary columns
simpsons <- simpsons %>%
  select(-id, -production_code, -views, -video_url, -image_url)

# Handle na/nan/null
simpsons <- simpsons %>% filter_all(all_vars(!is.na(.) & !is.nan(.) & !is.null(.)))
```

Transformacija podataka

Za početak faktorizirao sam stupac `season` koji je do faktorizacija bio numerička vrijednost.

```
# Factorizing
simpsons$season <- as.factor(simpsons$season)
```

Varijabla `episodes` sadrži prosječne vrijednosti gledanosti i ocjene za svaki redni broj epizode u svim sezonama

```
episodes <- simpsons %>%
  group_by(number_in_season) %>%
  summarise(
    us_viewers_in_millions = median(us_viewers_in_millions),
    imdb_rating = median(imdb_rating)
    )
```

Dvije navedene varijable predstavljaju 5 najbolje i najgore ocjenjenih epizoda

```
# Top 5 scored episodes
episodes_bot <- simpsons %>% arrange(imdb_rating) %>% head(5)
episodes_top <- simpsons %>% arrange(desc(imdb_rating)) %>% head(5)
```

Za svaku od 30 sezona varijabla `season` sadrži

- prosječne vrijednosti gledanosti
- standardnu devijaciju ocjena
- ocjene
- dani trajanja sezone

```
# Season table
season <- simpsons %>%
  group_by(season) %>%
  summarise(
    us_viewers_in_millions = median(us_viewers_in_millions),
    imdb_rating_sd=sd(imdb_rating),
    imdb_rating = median(imdb_rating),
    duration=max(original_air_date) - min (original_air_date)
  ) %>%
  filter(duration != 0)
```

Dvije varijable predstavljaju 5 najkontroverznijih i najkonzistetnijih sezona

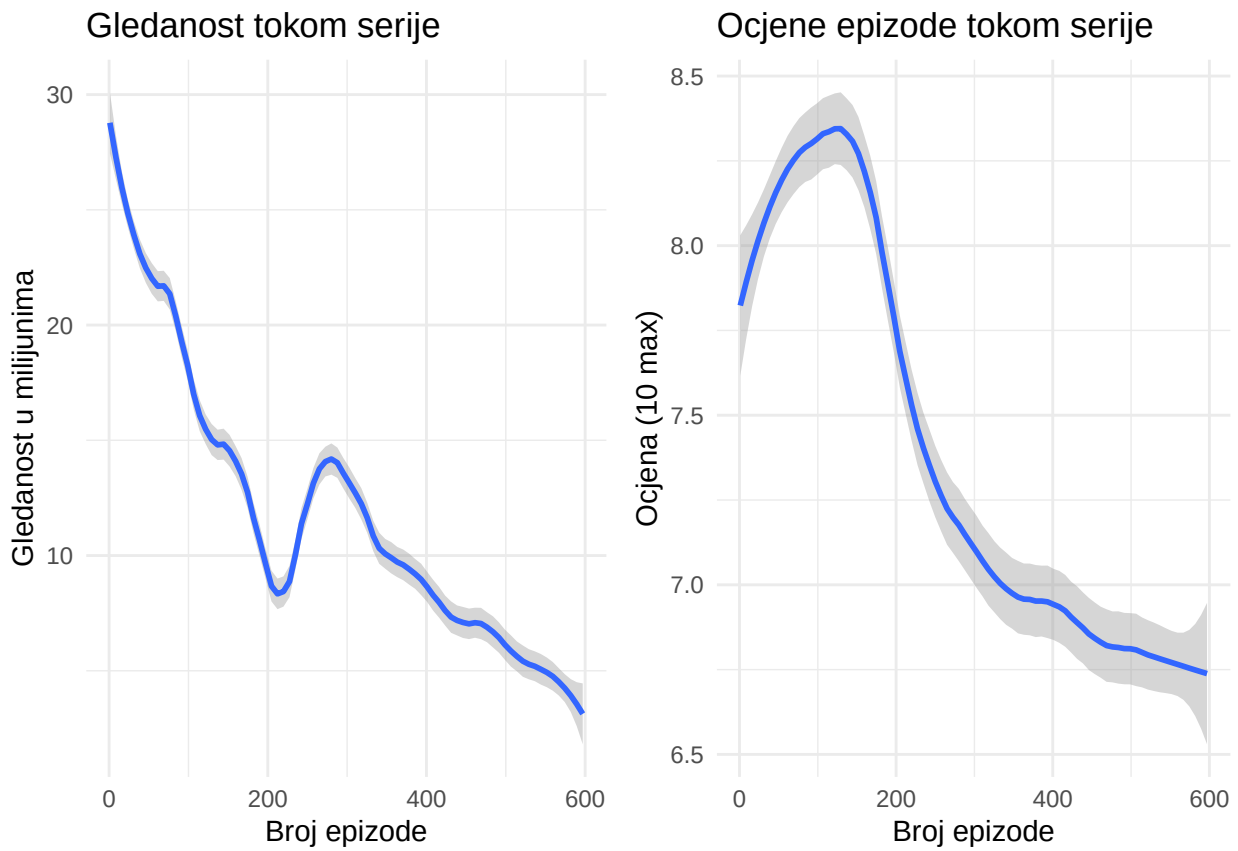
```
# Top 5 contraversiosn
season_sd_top <- season %>% top_n(5,wt=imdb_rating_sd)
season_sd_bot <- season %>% top_n(5,wt=desc(imdb_rating_sd))
```

`previous_episode` varijabla sadrži atribut `diff_viewership` koji predstavlja za koliko se razlikuje gledanost s obzirom na prethodnu epizodu, analogno vrijedi i za varijablu `diff_rating`. 5 najvećih vrijednosti spremili smo u zasebne varijable

```
# Previous episode table
previous_episode <- simpsons %>%
  arrange(number_in_series) %>%
  mutate(diff_rating = imdb_rating - lag(imdb_rating),
         diff_viewership = us_viewers_in_millions - lag(us_viewers_in_millions))

previous_episode_diff_rating_top <- previous_episode %>%
  arrange(diff_rating) %>%
  head(5)
previous_episode_diff_viewership_top <- previous_episode %>%
  arrange(diff_viewership) %>%
  head(5)
```

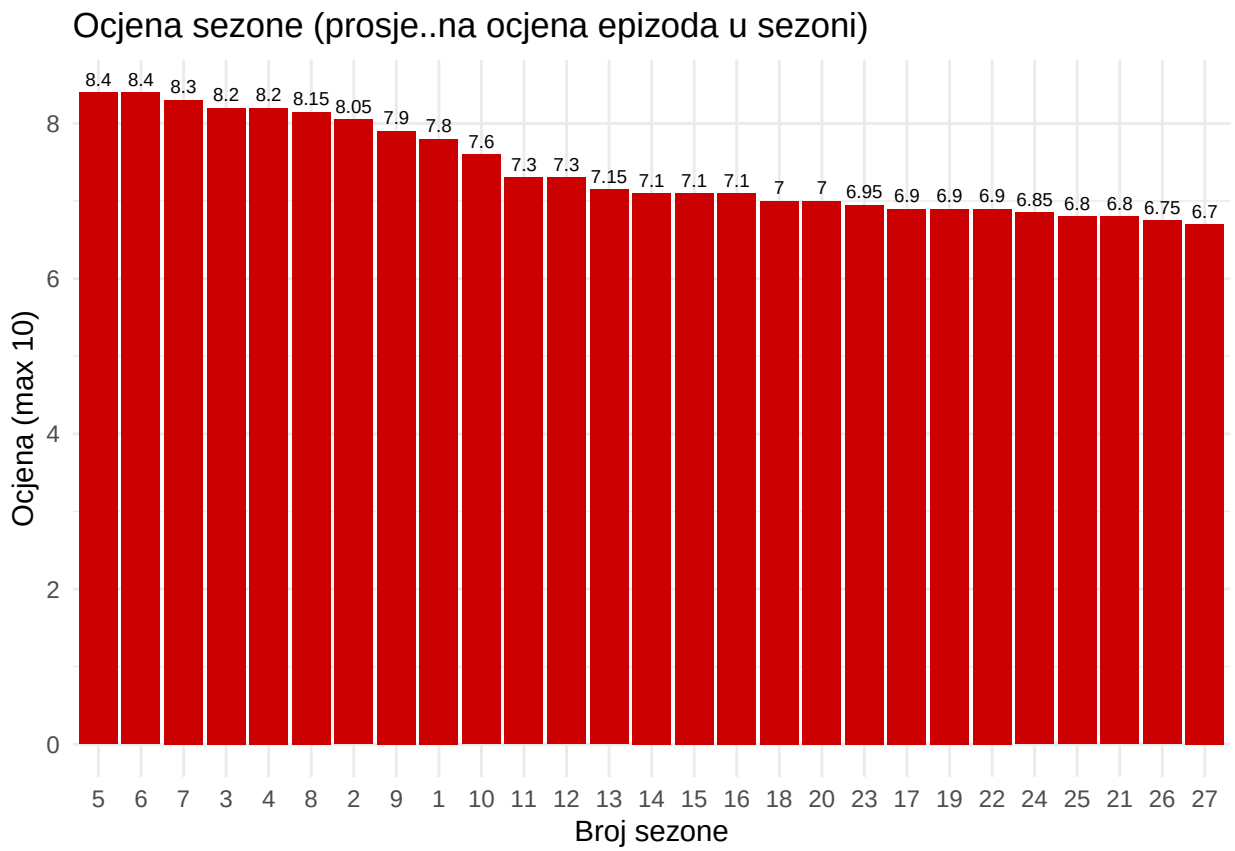
Vizualizacija podataka

```
# Views and rating
gg_view <- ggplot(simpsons, aes(x = number_in_series, y = us_viewers_in_millions)) +
  stat_smooth(span = 0.2, se=TRUE) +
  labs(title="Gledanost tokom serije" ,x = "Broj epizode" , y="Gledanost u milijunima") +
  theme_minimal()

gg_rating <-
  ggplot(simpsons, aes(x = number_in_series, y = imdb_rating)) +
  stat_smooth(span = 0.3, se=TRUE) +
  labs(title = "Ocjene epizode tokom serije", x="Broj epizode",y="Ocjena (10 max)") +
  theme_minimal()

grid.arrange(gg_view, gg_rating, ncol=2)
```



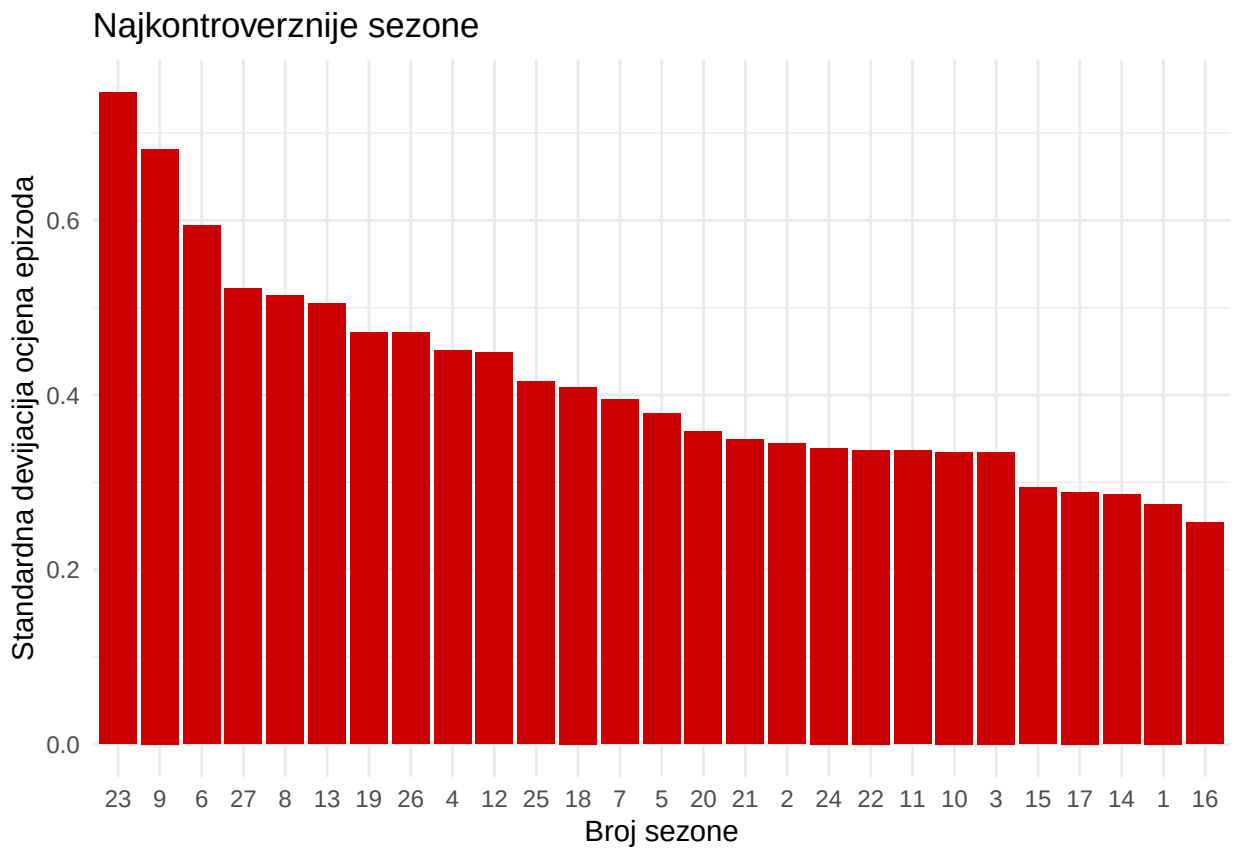
Na prikazanom grafu možemo vidjeti da popularnost Simpsona generalno opada kroz vrijeme. U sredini grafa možemo primjetiti iznenađujuć porast u gledanosti iako nije dugo zaživio što ima smisla s obizrom da kvaliteta serije konzistentno pada, vidljivo na grafu ocjenjivanja.

```
# Highest-lowest rated season
ggplot(season, aes(x = reorder(season,-imdb_rating), y = imdb_rating)) +
  geom_col(fill='#cc0000') +
  labs(title = "Ocjena sezone (prosječna ocjena epizoda u sezoni)",
       x = "Broj sezone",
       y = "Ocjena (max 10)") +
  geom_text(aes(label=imdb_rating), size = 2.5, colour="black", vjust = -0.5) +
  theme_minimal()
```



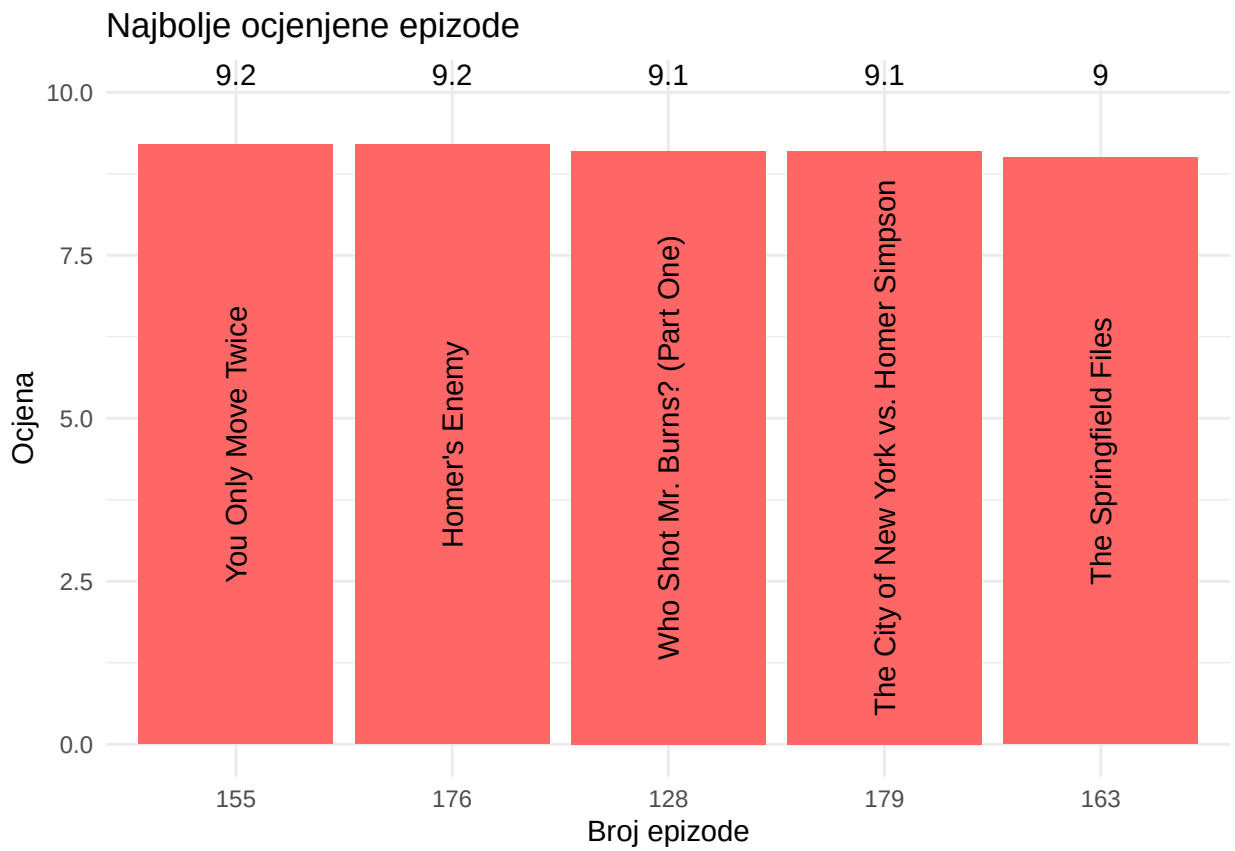
Najbolje Simpsons sezone su tri zaredane sezone 5,6,7 isto tako tri najgore sezone su također zaredane a one su 25,26,27

```
# Highest-lowest consistancy
ggplot(season, aes(x = reorder(season,-imdb_rating_sd), y = imdb_rating_sd)) +
  geom_col(fill='#cc0000') +
  labs(title = "Najkontroverznije sezone",
       x = "Broj sezone",
       y = "Standardna devijacija ocjena epizoda") +
  theme_minimal()
```



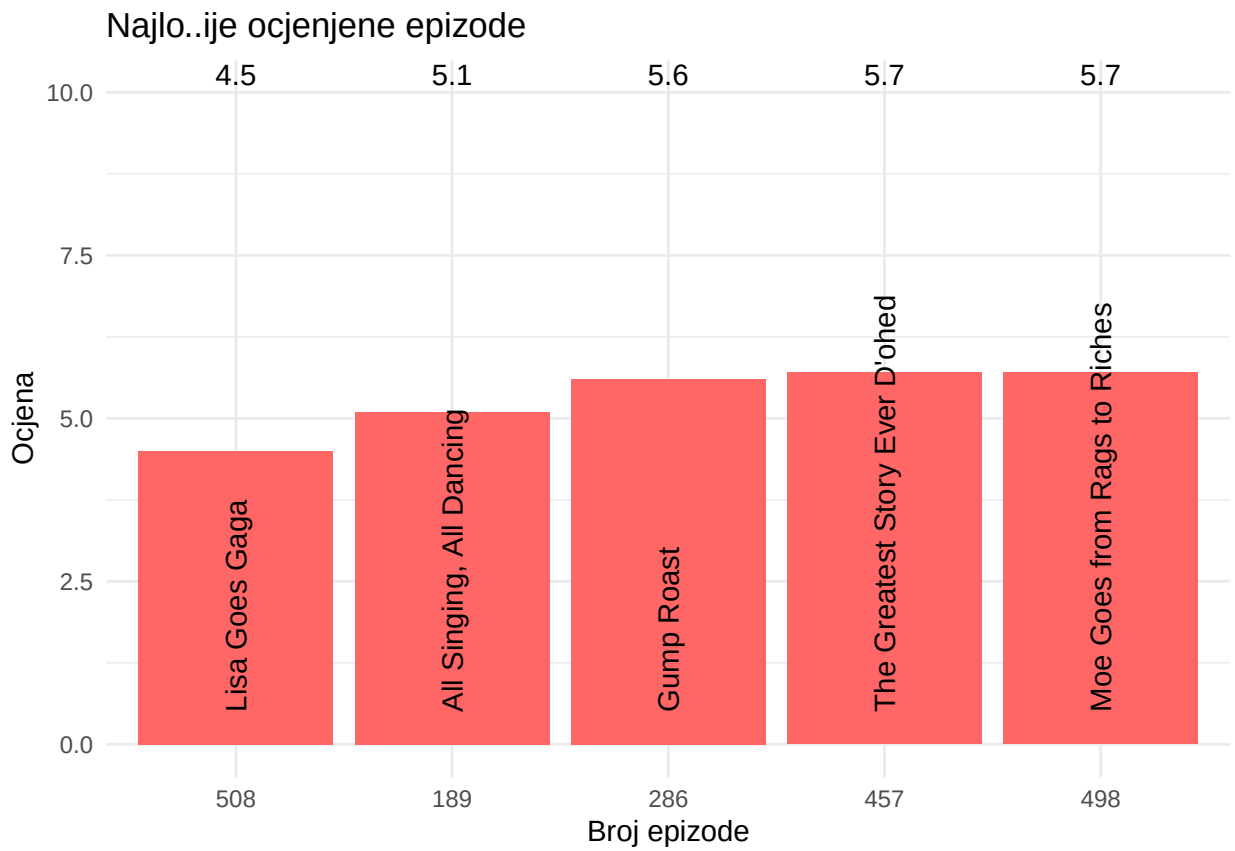
Najkontroverznije sezone su sezone čija je standardna devijacija ocjene epizoda najviša. Suprotno, sezona koja ima najmanju standardnu devijaciju je najkonzistentnija u svojoj kvaliteti.

```
# Highest scored episode
ggplot(episodes_top, aes(x = reorder(number_in_series, -imdb_rating),y=imdb_rating))  +
  geom_col(fill='#ff6666') +
  expand_limits(y=10) +
  geom_text(aes(label=title, fill="black"),
            position = position_stack(vjust = 0.5),
            size = 4,
            angle = 90,
            colour = "black") +
  geom_text(aes(label=imdb_rating),
            size = 4,
            colour="black",
            y = Inf,
            vjust = 1.2) +
  labs(title = "Najbolje ocjenjene epizode",
       x = "Broj epizode",
       y = "Ocjena") +
  theme_minimal()
```



Najbolje epizode prikazuju sjaj Simpsons serijala i s obzirom na prosječne IMDB ocjene, ocjena 9.2 je gotovo pa savršena ocjena. Najbolje ocjenjena epizoda je "Homer's Enemy" (ažurirano) u kojoj se Homer pokuša sprijateljeiti sa zaposlenikom Nuklearne elektrane - Frank Grimsom. Frank iritian i ljut Homerovom nesposobnošću, proglašava Homera svojim neprijateljem te javnim poniženjem Homera dodatno razotkriva Homerove mane.

```
# Lowest scored episode
ggplot(episodes_bot, aes(x = reorder(number_in_series, imdb_rating),y=imdb_rating))  +
  geom_col(fill='#ff6666') +
  expand_limits(y=10) +
  geom_text(aes(label=title, fill="black"),
            size = 4,
            angle = 90,
            colour="black",
            y = 0.5,
            hjust = 'left') +
  geom_text(aes(label=imdb_rating),
            size = 4,
            colour="black",
            y = Inf,
            vjust = 1.2) +
  labs(title = "Najlošije ocjenjene epizode",
       x = "Broj epizode",
       y = "Ocjena") +
  theme_minimal()
```

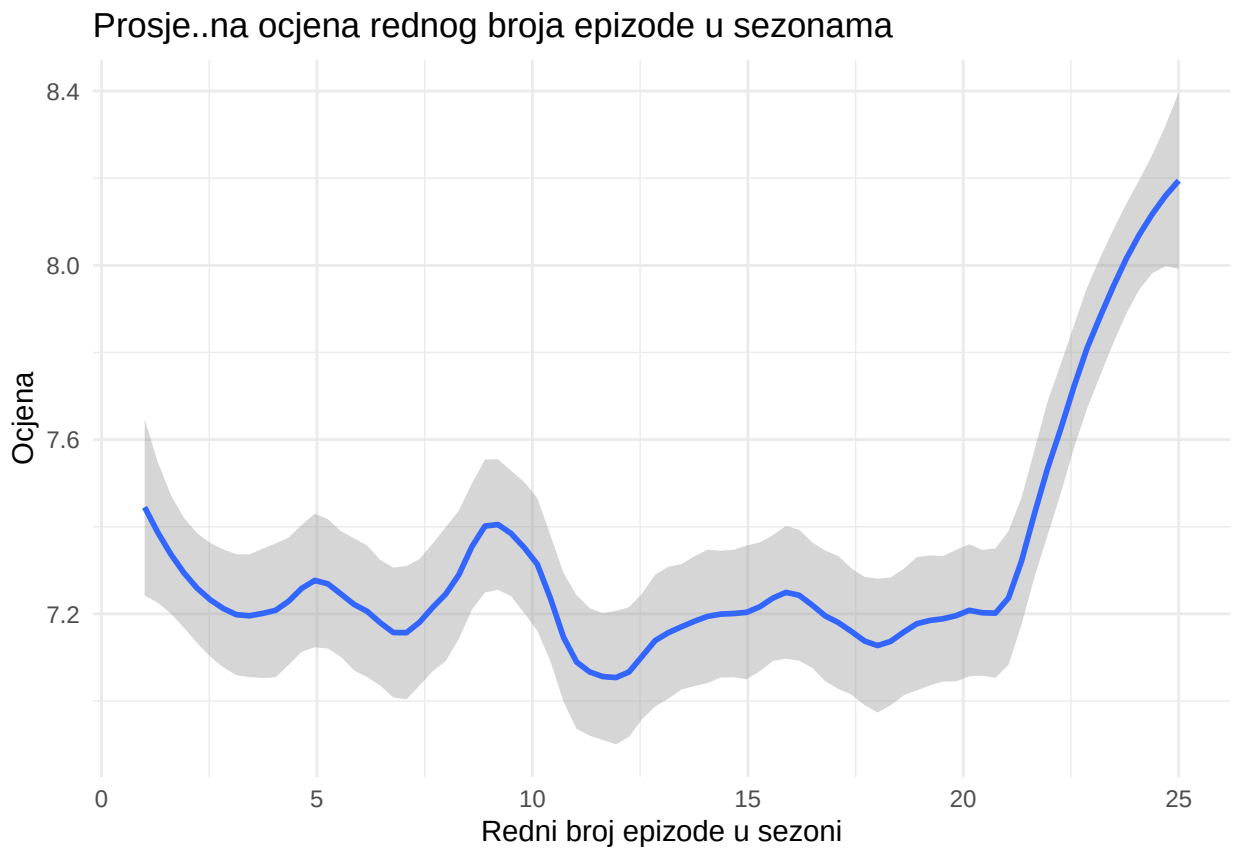


Najlošije Simpsons epizode dolaze u kasnijim brojevima epizode a najgora epizoda je "Lisa Goes Gaga"

```
# Maximum score differnce based on two nehgoubour episodes
ggplot(previous_episode_diff_rating_top ,
       aes(x = reorder(as.factor(number_in_series),
                       diff_rating),
           y=diff_rating)) +
  geom_col(fill='#ff6666') +
  labs(title = "Najveće razlike u gledanosti između dvije epizode",
       x = "Broj epizode u sezoni",
       y = "Razlika u gledanosti u milijunima")+
geom_text(aes(label=title, fill="black"),
          position = position_stack(vjust = 0.5),
          size = 4,
          angle = 90,
          colour="black") +
  theme_minimal()
```

Razlika u gledanosti broj je koji govori koliko je milijuna ljudi manje pogledalo sljedeću epizodu, točnije, koliko je epizoda spustila očekivanja i motivaciju da gledatelj pogleda sljedeću epizodu. 189 najviše je smanjila gledanost sljedeće epizode. Također možemo primjetiti da se 3 od 5 epizoda poklapaju sa najgorim epizodama Simpsona

```
#which period is usually the best in the season (start, middle, ending)
ggplot(episodes, aes(x = number_in_season,y=imdb_rating)) +
  stat_smooth(span = 0.3, se=TRUE) +
  labs(title = "Prosječna ocjena rednog broja epizode u sezonama",
       x = "Redni broj epizode u sezoni",
       y = "Ocjena") +
  theme_minimal()
```



Ovim grafom prikazane su prosječne ocjene za redni broj epizode u svim sezonama. Pri kraju svake sezone kvaliteta epizoda naglo poraste vjerojatno da se gledatlji zainteresiraju za serijal ne bi li nastavili gledati nadolazeću sezonu. Također na početku svake sezone postoje porasti i padovi kojom se stvara dinamika u sezoni nakon čega se ocjena stabilizira.