Oscarovci

Tim #1

2025-11-06

Contents

1 Opis skupa podataka

Skup podataka koji analiziramo u projektu sadrži 571 odgovor/redak.

```
#Prikaz varijabli u skupu podataka i njihovog podatkovnog tipa

data = read_csv2("oscars_dataset.csv")
```

```
## i Using "',' as decimal and "'.'" as grouping mark. Use `read_delim()` for more control.
```

```
## Rows: 571 Columns: 20
## -- Column specification --------------------------------------------------------
## Delimiter: ";"
## chr  (11): Film, Oscar Year, Film Studio/Producer(s), Award, Movie Genre, Ma...
## dbl   (3): Unnamed: 0, Year of Release, Movie Time
## num   (5): IMDB Rating, Tomatometer Rating, Tomatometer Count, Audience Rati...
## date  (1): Original Release Date
##
## 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: 571
## Columns: 20
## $ `Unnamed: 0`              <dbl> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13~
## $ Film                      <chr> "Wings", "7th Heaven", "The Racket", "The Br~
## $ `Oscar Year`              <chr> "1927/28", "1927/28", "1927/28", "1928/29", ~
## $ `Film Studio/Producer(s)` <chr> "Famous Players-Lasky", "Fox", "The Caddo Co~
## $ Award                     <chr> "Winner", "Nominee", "Nominee", "Winner", "N~
## $ `Year of Release`         <dbl> 1927, 1927, 1928, 1929, 1929, 1929, 1928, 19~
## $ `Movie Time`              <dbl> 144, 110, 84, 100, 91, 130, 95, 113, 152, 87~
## $ `Movie Genre`             <chr> "Drama,Romance,War", "Drama,Romance", "Crime~
## $ `Main Genre`              <chr> "Drama", "Drama", "Crime", "Drama", "Action"~
## $ Subgenre                  <chr> "Romance", "Romance", "Drama", "Musical", "C~
## $ `IMDB Rating`             <dbl> 75, 77, 67, 57, 58, 57, 56, 74, 81, 71, 62, ~
## $ `IMDB Votes`              <chr> "12,221", "3,439", "1,257", "6,890", "765", ~
## $ `Content Rating`          <chr> "PG-13", NA, NA, "NR", NA, NA, "NR", NA, NA,~
## $ Directors                 <chr> "William Wellman", NA, NA, "Harry Beaumont",~
## $ `Original Release Date`   <date> 1927-08-12, NA, NA, 1929-02-01, NA, NA, 192~
## $ `Production Company`      <chr> "Unknown", NA, NA, "MGM Home Entertainment",~
## $ `Tomatometer Rating`      <dbl> 930, NA, NA, 330, NA, NA, 560, NA, NA, 750, ~
## $ `Tomatometer Count`       <dbl> 460, NA, NA, 240, NA, NA, 90, NA, NA, 80, NA~
## $ `Audience Rating`         <dbl> 780, NA, NA, 210, NA, NA, 380, NA, NA, 690, ~
## $ `Audience Count`          <dbl> 35300, NA, NA, 18130, NA, NA, 3560, NA, NA, ~
```

```
var_types <- sapply(data, class)
cat(paste(names(var_types), ": ", var_types, sep = "", collapse = "\n"))
```

```
## Unnamed: 0: numeric
## Film: character
```

```
## Oscar Year: character
## Film Studio/Producer(s): character
## Award: character
## Year of Release: numeric
## Movie Time: numeric
## Movie Genre: character
## Main Genre: character
## Subgenre: character
## IMDB Rating: numeric
## IMDB Votes: character
## Content Rating: character
## Directors: character
## Original Release Date: Date
## Production Company: character
## Tomatometer Rating: numeric
## Tomatometer Count: numeric
## Audience Rating: numeric
## Audience Count: numeric
```

2 Analiza žanrova kod pobjednika (Matija)

2.1 Opis problema

Kod analize žanrova kod pobjednika, istražit ćemo postoje li značajne razlike u učestalosti pobjeda među različitim žanrovima filmova nominiranih za Oscara. Želimo utvrditi jesu li neki žanrovi skloniji da osvoje nagrade u usporedbi s drugim žanrovima. Pridjelit ćemo svakome filmu samo jedan žanr i iz podataka ćemo za ovu analizu koristiti stupac `Main Genre`.

2.2 Analiza podataka

2.2.1 Grupiranje žanrova i priprema podataka za analizu:

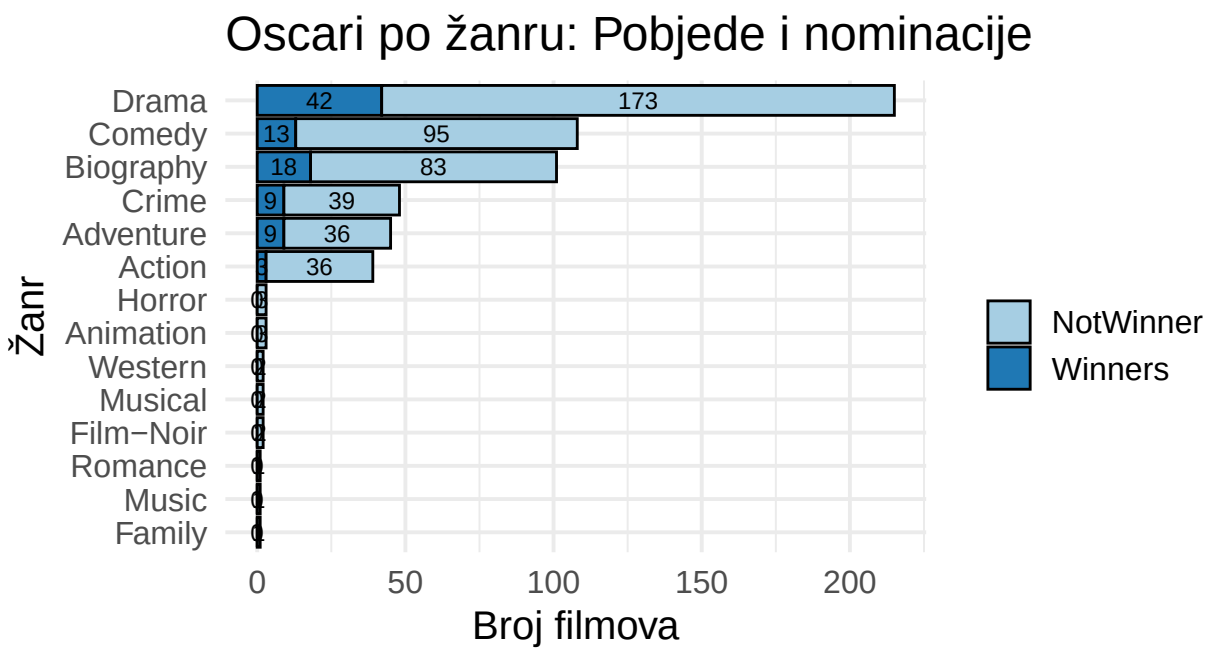
Prvo nego što krenemo raditi bilo kakve statističke testove, moramo se upoznati s podacima.

```
data <- data %>%
  mutate(`Main Genre` = as.character(`Main Genre`))

winners <- data %>% filter(Award == "Winner")
nominees <- data %>% filter(Award %in% c("Nominee", "Winner"))

winner_counts <- winners %>%
  count(`Main Genre`, name = "Winners")

nominee_counts <- nominees %>%
  count(`Main Genre`, name = "Nominations")

genre_table <- nominee_counts %>%
  left_join(winner_counts, by = "Main Genre") %>%
  mutate(
    Winners = replace_na(Winners, 0),
    NotWinner = Nominations - Winners
  ) %>%
  rename(Genre = `Main Genre`)
```

Vizualizacija pobjeda i nominacija po žanru:

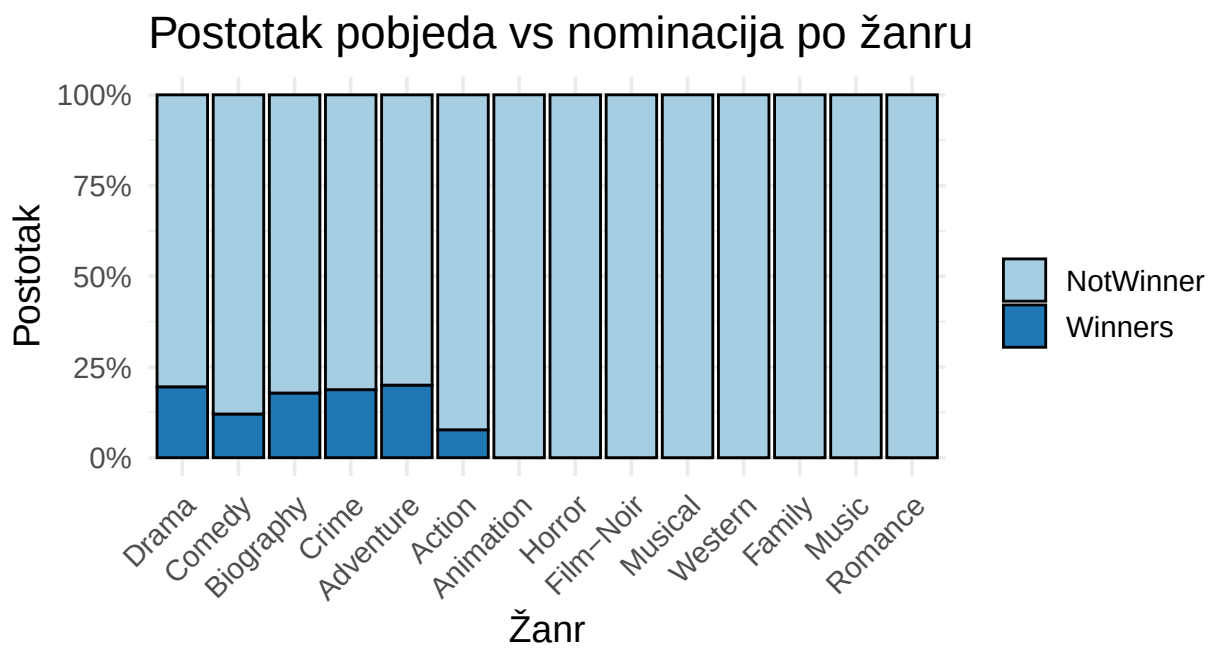
```
genre_plot_data <- genre_table %>%
  arrange(desc(Winners)) %>%
  pivot_longer(
    cols = c(Winners, NotWinner),
    names_to = "Outcome",
    values_to = "Count"
  )


ggplot(genre_plot_data,
       aes(x = Count,
           y = reorder(Genre, Count),
           fill = Outcome)) +
  geom_bar(stat = "identity", position = "stack", color = "black") +
  geom_text(aes(label = Count),
```

```
            position = position_stack(vjust = 0.5),
            size = 3) +
  scale_fill_manual(values = c(
    "Winners" = "#1f78b4",
    "NotWinner" = "#a6cee3"
  )) +
  theme_minimal(base_size = 15) +
  labs(
    title = "Oscari po žanru: Pobjede i nominacije",
    x = "Broj filmova",
    y = "Žanr",
    fill = ""
  )
```



Zbog velike disproporcije između broja pobjeda i nominacija, koristimo i postotne prikaze te logaritamsku skalu za bolju vizualizaciju podataka. ### Postotni prikaz pobjeda vs nominacija po žanru:

```
ggplot(genre_plot_data, aes(x = reorder(Genre, -Count), y = Count, fill = Outcome)) +
  geom_bar(stat = "identity", position = "fill", color = "black") +
  scale_y_continuous(labels = scales::percent_format()) +
  scale_fill_manual(values = c("Winners" = "#1f78b4", "NotWinner" = "#a6cee3")) +
  theme_minimal(base_size = 14) +
  labs(title = "Postotak pobjeda vs nominacija po žanru",
       x = "Žanr", y = "Postotak", fill = "") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

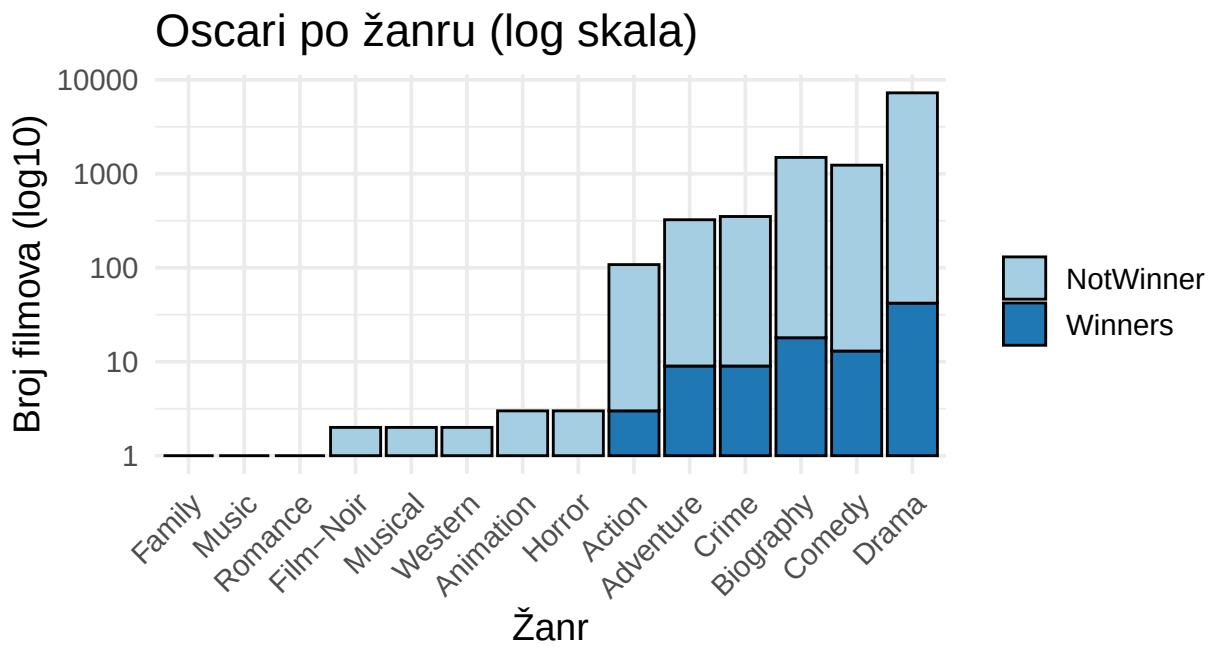


2.2.2 Logaritamska skala za broj pobjeda vs nominacija po žanru:

```
ggplot(genre_plot_data, aes(x = reorder(Genre, Count), y = Count, fill = Outcome)) +
  geom_bar(stat = "identity", position = "stack", color = "black") +
  scale_y_log10() +
  scale_fill_manual(values = c("Winners" = "#1f78b4", "NotWinner" = "#a6cee3")) +
  theme_minimal(base_size = 14) +
  labs(title = "Oscari po žanru (log skala)",
       x = "Žanr", y = "Broj filmova (log10)", fill = "") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning in scale_y_log10(): log-10 transformation introduced infinite values.
```

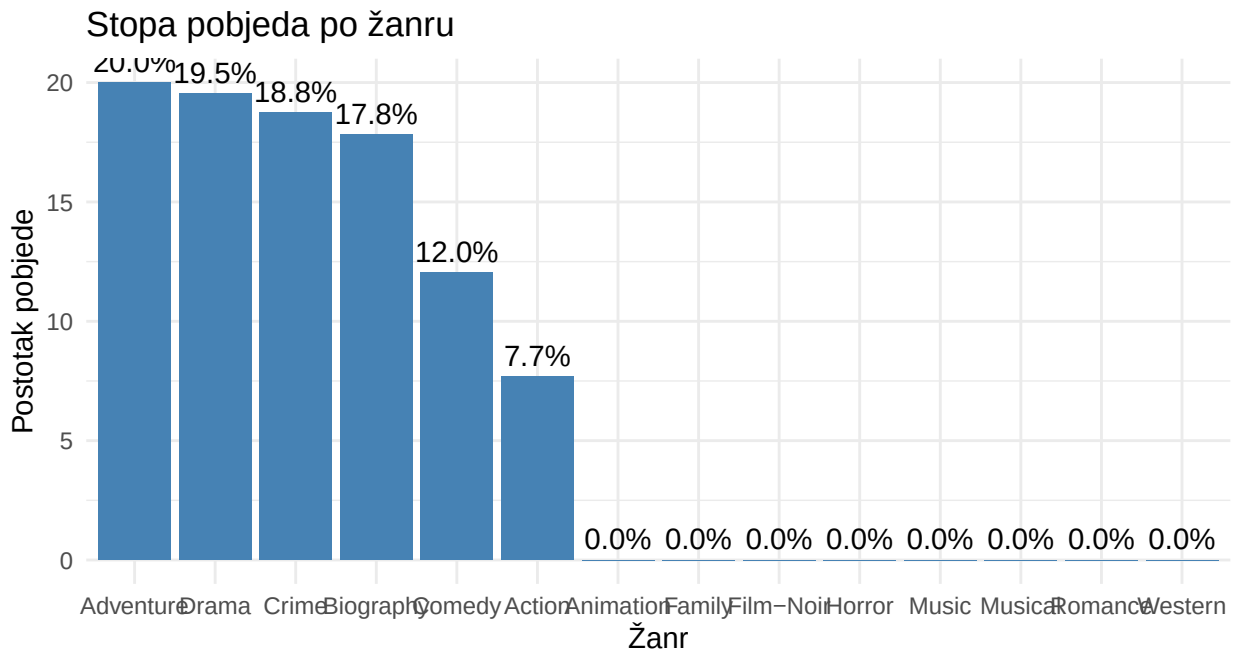
```
## Warning: Removed 8 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```



2.2.3 Stopa uspješnosti po žanru:

Možemo također izračunati stopu uspješnosti (postotak pobjeda u odnosu na broj nominacija) za svaki žanr kako bismo dodatno ilustrirali razlike među žanrovima.

```
genre_table <- genre_table %>%
  mutate(WinRate = (Winners / Nominations) * 100)

ggplot(genre_table, aes(x = reorder(Genre, -WinRate), y = WinRate)) +
  geom_col(fill = "steelblue") +
  geom_text(aes(label = sprintf("%.1f%%", WinRate)), vjust = -0.5) +
  labs(title = "Stopa pobjeda po žanru", x = "Žanr", y = "Postotak pobjede") +
  theme_minimal()
```



2.2.4 Kontigencijska tablica nominiranih i pobjednika po žanrovima:

```
kontigencijska <- genre_table[, c("Winners", "NotWinner")]
row.names(kontigencijska) <- genre_table$Genre
```

```
## Warning: Setting row names on a tibble is deprecated.
```

```
kontigencijska
```

```
## # A tibble: 14 x 2
##    Winners NotWinner
##  *   <int>     <int>
##  1       3        36
##  2       9        36
##  3       0         3
##  4      18        83
##  5      13        95
##  6       9        39
##  7      42       173
##  8       0         1
##  9       0         2
## 10       0         3
## 11       0         1
## 12       0         2
## 13       0         1
## 14       0         2
```

2.3 Hi-kvadrat test i Fisherov test

2.3.1 Hipoteze:

- Nulta hipoteza (H0): Ne postoji značajna razlika u učestalosti pobjeda među različitim žanrovima filmova.
- Alternativna hipoteza (H1): Postoji značajna razlika u učestalosti pobjeda među različitim žanrovima filmova.

2.3.2 Grupiranje žanrova i priprema podataka za analizu:

Hi-kvadrat test ima važan uvjet: očekivane frekvencije u svakoj ćeliji kontingencijske tablice trebaju biti veće od 5. Za žanrove e jako malim brojem filmova (npr. 1–2 pojave), očeivane frekvencije za te vrlo male. Kako bi statistički test bio valjan, grupirat ćemo manje zastupljene žanrove u kategoriju "Ostali" kako bismo osigurali da sve ćelije imaju dovoljne očekivane frekvencije.

```
common_genres <- c("Action", "Adventure", "Biography", "Comedy", "Crime", "Drama")
data$GenreGroup <- ifelse(data$`Main Genre` %in% common_genres,
                          data$`Main Genre`,
                          "Other")
winners <- subset(data, Award == "Winner")
nominees <- subset(data, Award %in% c("Nominee", "Winner"))
winner_counts <- as.data.frame(table(winners$GenreGroup))
colnames(winner_counts) <- c("Genre", "Winners")
nominee_counts <- as.data.frame(table(nominees$GenreGroup))
colnames(nominee_counts) <- c("Genre", "Nominations")
df <- merge(nominee_counts, winner_counts, by="Genre", all.x=TRUE
)
df$Winners[is.na(df$Winners)] <- 0
df$NotWinner <- df$Nominations - df$Winners
kontigencijska <- df[, c("Winners", "NotWinner")]
row.names(kontigencijska) <- df$Genre

kontigencijska
```

```
##           Winners NotWinner
## Action          3        36
## Adventure       9        36
## Biography      18        83
## Comedy         13        95
## Crime           9        39
## Drama          42       173
## Other           0        15
```

2.3.3 Hi-kvadrat test:

Sada radimo Hi-kvadrat test kako bismo vidjeli postoje li značajne razlike u učestalosti pobjeda među različitim žanrovima

```
# Chi-squared test
chisq <- chisq.test(kontigencijska)
```

```
## Warning in chisq.test(kontigencijska): Chi-squared approximation may be
## incorrect
```

```
chisq
```

```
##
##  Pearson's Chi-squared test
##
## data:  kontigencijska
## X-squared = 8.8789, df = 6, p-value = 0.1805
```

Standardizirani reziduali koji pokazuju kako žanrovi odstupaju:

```
chisq$stdres
```

```
##              Winners  NotWinner
## Action    -1.5300621  1.5300621
## Adventure  0.6667449 -0.6667449
## Biography  0.4060704 -0.4060704
## Comedy    -1.3771971  1.3771971
## Crime      0.4465715 -0.4465715
## Drama      1.5385882 -1.5385882
## Other     -1.7423324  1.7423324
```

2.3.4 Fisherov test:

Za dodatnu potvrdu rezultata Hi-kvadrat testa, koristimo Fisherov test koji je prikladniji za manje uzorke ili kada su očekivane frekvencije niske.

```
fisher.test(kontigencijska, simulate.p.value = TRUE, B = 1e6)
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  1e+06 replicates)
##
## data:  kontigencijska
## p-value = 0.1696
## alternative hypothesis: two.sided
```

2.4 Zaključak

Na temelju rezultata Hi-kvadrat testa i Fisherovog testa, možemo zaključiti da ne postoje značajne razlike u učestalosti pobjeda među različitim žanrovima filmova nominiranih za Oscara. Standardizirani reziduali ukazuju na to koji žanrovi značajno odstupaju od očekivanih vrijednosti, što može biti korisno za daljnje istraživanje i razumijevanje trendova u dodjeli nagrada.

3 Analiza multi-label žanrova kod pobjednika (Matija)

3.1 Opis problema

U ovoj analizi istražit ćemo kako se različiti žanrovi filmova pojavljuju među pobjednicima Oscara kada filmovi mogu pripadati više žanrova istovremeno (multi-label), jer je nekada teško jednoznačno odrediti kojemu žanru film pripada. Tako ćemo sljedeće pokušati koristiti kolumnu `Movie Genre`, koja može sadržavati više žanrova odvojenih zarezom. Svaku pojavu žanra za neki film ćemo pribrojati tome žanru.

3.2 Analiza podataka

Priprema podataka za multi-label analizu žanrova:

```
data <- data %>%
  mutate(`Movie Genre` = as.character(`Movie Genre`))
```

Mutiramo podatke kako bismo razdvojili višestruke žanrove u zasebne retke.

```
data_long <- data %>%
  separate_rows(`Movie Genre`, sep = ",") %>%
  mutate(`Movie Genre` = trimws(`Movie Genre`))
```

Izračun broja pobjeda i ne-pobjeda po žanru:

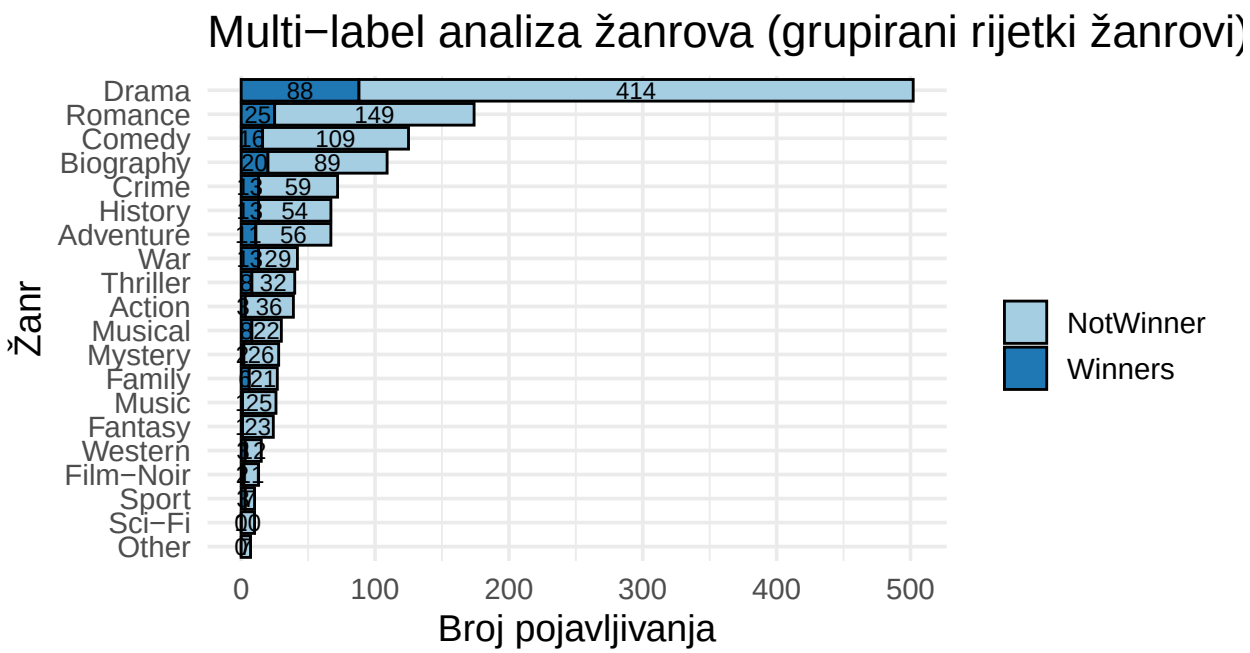
```
genre_counts <- data_long %>%
  mutate(
    Outcome = ifelse(Award == "Winner", "Winner", "NotWinner")
  ) %>%
  count(`Movie Genre`, Outcome) %>%
  pivot_wider(
    names_from = Outcome,
    values_from = n,
    values_fill = 0
  ) %>%
  mutate(
    Total = Winner + NotWinner,
    GenreGroup = ifelse(Total < 5, "Other", `Movie Genre`),
    SuccessRate = Winner / Total
  )
```

Grupiranje rijetkih žanrova u kategoriju “Ostali” i izračun ukupnih vrijednosti po grupama žanrova:

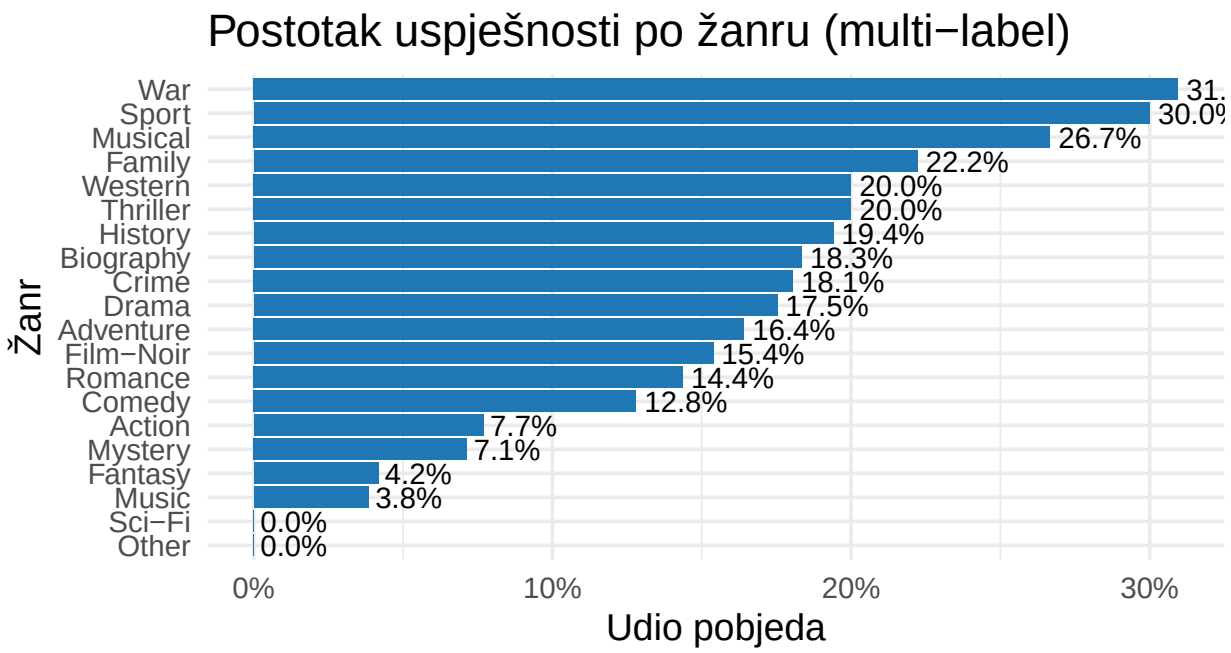
```
genre_table <- genre_counts %>%
  group_by(GenreGroup) %>%
  summarise(
    Winners = sum(Winner),
    NotWinner = sum(NotWinner),
    Total = sum(Total),
    SuccessRate = sum(Winner)/sum(Total),
    .groups = "drop"
  )
```

3.2.1 Vizualizacija rezultata multi-label analize žanrova:

```
genre_melt <- genre_table %>%
  select(GenreGroup, Winners, NotWinner) %>%
  pivot_longer(
    cols = c(Winners, NotWinner),
    names_to = "Outcome",
    values_to = "Count"
  )

# Bar plot (broj pojavljivanja)
ggplot(genre_melt, aes(x = Count, y = reorder(GenreGroup, Count), fill = Outcome)) +
  geom_bar(stat = "identity", color = "black") +
  geom_text(aes(label = Count), position = position_stack(vjust = 0.5), size = 3) +
  scale_fill_manual(values = c("Winners" = "#1f78b4", "NotWinner" = "#a6cee3")) +
  theme_minimal(base_size = 14) +
  labs(title = "Multi-label analiza žanrova (grupirani rijetki žanrovi)",
       x = "Broj pojavljivanja",
       y = "Žanr",
       fill = "")
```




```
# Bar plot (postotak pobjeda)
ggplot(genre_table, aes(x = SuccessRate, y = reorder(GenreGroup, SuccessRate))) +
  geom_col(fill = "#1f78b4") +
  geom_text(aes(label = scales::percent(SuccessRate, accuracy = 0.1)), hjust = -0.1, size = 4) +
  scale_x_continuous(labels = scales::percent_format(), limits = c(0, NA)) +
  theme_minimal(base_size = 14) +
  labs(title = "Postotak uspješnosti po žanru (multi-label)",
       x = "Udio pobjeda",
       y = "Žanr")
```



3.3 Hi-kvadrat test i Fisherov test

Isto kao i kod single-label analize, provodimo Hi-kvadrat test i Fisherov test kako bismo utvrdili postoje li značajne razlike u učestalosti pobjeda među različitim žanrovima u multi-label kontekstu.

3.3.1 Hipoteze:

Hipoteze se ne mijenjaju.

- Nulta hipoteza (H0): Ne postoji značajna razlika u učestalosti pobjeda među različitim žanrovima filmova.
- Alternativna hipoteza (H1): Postoji značajna razlika u učestalosti pobjeda među različitim žanrovima filmova.

3.3.2 Hi-kvadrat test:

```
chisq <- chisq.test(kontigencijska)
```

```
## Warning in chisq.test(kontigencijska): Chi-squared approximation may be
## incorrect
```

```
chisq
```

```
##
##  Pearson's Chi-squared test
##
## data:  kontigencijska
## X-squared = 8.8789, df = 6, p-value = 0.1805
```

Očekivnane frekvencije i standardizirani reziduali:

```
chisq$expected    # provjera očekivanih frekvencija
```

```
##             Winners NotWinner
## Action     6.420315  32.57968
## Adventure  7.408056  37.59194
## Biography 16.626970  84.37303
## Comedy    17.779335  90.22067
## Crime      7.901926  40.09807
## Drama     35.394046 179.60595
## Other      2.469352  12.53065
```

```
chisq$stdres      # standardizirani reziduali
```

```
##               Winners  NotWinner
## Action    -1.5300621  1.5300621
## Adventure  0.6667449 -0.6667449
## Biography  0.4060704 -0.4060704
## Comedy    -1.3771971  1.3771971
## Crime      0.4465715 -0.4465715
## Drama      1.5385882 -1.5385882
## Other     -1.7423324  1.7423324
```

3.3.3 Fisherov test:

```
fisher.test(
  kontigencijska,
  simulate.p.value = TRUE,
  B = 1e6
)
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  1e+06 replicates)
##
## data:  kontigencijska
## p-value = 0.1698
## alternative hypothesis: two.sided
```

```
genre_plot <- genre_table %>%
  pivot_longer(
    cols = c(Winners, NotWinner),
    names_to = "Outcome",
    values_to = "Count"
  )
```

3.4 Zaključak

Na temelju dobivenih podataka zaključujemo da nema povezanosti između