Assignment DPREP Group 5

Eveline Verhage

2025-09-04

Contents

1 Question 1.1

How do audience ratings of “escapist” movie genres (fantasy, comedy, romance) compare to “heavy” genres (drama, thriller) across different historical release periods — Pre-War (1940), Post-War Boom (1950–1970), Pre-Digital/Pre-2000 (1970–1990), and the Modern Era (1990–2025)?

2 part 2.1

```r
# load the data
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(tidyr)

# 1) Download a sample directly from IMDb (fast & reproducible)
basics <- read_tsv(
  "https://datasets.imdbws.com/title.basics.tsv.gz",
  na = "\\N",
  col_select = c(tconst, titleType, startYear, genres),
  n_max = 100000
)
```

```
## Rows: 100000 Columns: 4

## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (3): tconst, titleType, genres
## dbl (1): startYear
##
## 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.
```

```r
ratings <- read_tsv(
  "https://datasets.imdbws.com/title.ratings.tsv.gz",
  na = "\\N",
  col_select = c(tconst, averageRating, numVotes),
  n_max = 100000
)
```

```
## Rows: 100000 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (1): tconst
## dbl (2): averageRating, numVotes
##
## 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.
```

```r
# 2) Merge on tconst
imdb_sample <- basics %>%
  inner_join(ratings, by = "tconst")

# 1) Movies filtered by startyear
imdb_movies <- imdb_sample %>%
  filter(titleType == "movie", !is.na(startYear)) %>%
  mutate(period = case_when(
    startYear <= 1939 ~ "Pre-war (<=1939)",
    startYear >= 1940 & startYear <= 1949 ~ "War decade (1940-1949)",
    startYear >= 1950 & startYear <= 1964 ~ "Post-war (1950-1964)",
    startYear >= 1965 & startYear <= 1979 ~ "Modern I (1965-1979)",
    startYear >= 1980 & startYear <= 2025 ~ "Modern II (1980-2025)",
    TRUE ~ NA_character_
  )) %>%
  filter(!is.na(period))

# 2 Filter by numVotes to reduce noise
imdb_movies <- imdb_movies %>%
  filter(numVotes >= 1000)

# 3) Map genre families (Escapist vs Heavy)
imdb_movies_genrefam <- imdb_movies %>%
  filter(!is.na(genres)) %>%
  separate_rows(genres, sep = ",") %>%
  mutate(genre_family = case_when(
    genres %in% c("Fantasy", "Comedy", "Romance")  ~ "Escapist",
    genres %in% c("Drama", "Thriller")             ~ "Heavy",
    TRUE                                           ~ NA_character_
  )) %>%
  filter(!is.na(genre_family))

# 4) Distribution by period × genre family
period_genre_counts <- imdb_movies_genrefam %>%
  count(period, genre_family, name = "n") %>%
  group_by(period) %>%
  mutate(period_total = sum(n),
         share_within_period = n / period_total) %>%
  ungroup()

period_genre_counts
```

```
## # A tibble: 10 x 5
##    period                 genre_family     n period_total share_within_period
##    <chr>                  <chr>        <int>        <int>               <dbl>
##  1 Modern I (1965-1979)   Escapist      1625         3932               0.413
##  2 Modern I (1965-1979)   Heavy         2307         3932               0.587
##  3 Modern II (1980-2025)  Escapist      2244         4402               0.510
##  4 Modern II (1980-2025)  Heavy         2158         4402               0.490
##  5 Post-war (1950-1964)   Escapist      1183         2915               0.406
##  6 Post-war (1950-1964)   Heavy         1732         2915               0.594
##  7 Pre-war (<=1939)       Escapist       822         1584               0.519
##  8 Pre-war (<=1939)       Heavy          762         1584               0.481
##  9 War decade (1940-1949) Escapist       688         1492               0.461
## 10 War decade (1940-1949) Heavy          804         1492               0.539
```

```r
# 5) Quality overview
period_genre_summary <- imdb_movies_genrefam %>%
  group_by(period, genre_family) %>%
  summarise(
    n_titles   = n(),
    avg_rating = mean(averageRating, na.rm = TRUE),
    med_rating = median(averageRating, na.rm = TRUE),
    avg_votes  = mean(numVotes, na.rm = TRUE),
    med_votes  = median(numVotes, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  arrange(period, genre_family)

period_genre_summary
```

```
## # A tibble: 10 x 7
##    period        genre_family n_titles avg_rating med_rating avg_votes med_votes
##    <chr>         <chr>           <int>      <dbl>      <dbl>     <dbl>     <dbl>
##  1 Modern I (19~ Escapist         1625       6.49        6.6    11096.      3089
##  2 Modern I (19~ Heavy            2307       6.66        6.8    15277.      3029
##  3 Modern II (1~ Escapist         2244       6.18        6.3    23094.     4514.
##  4 Modern II (1~ Heavy            2158       6.45        6.6    22878.     3980.
##  5 Post-war (19~ Escapist         1183       6.71        6.8    10440.      2912
##  6 Post-war (19~ Heavy            1732       6.89        6.9    13466.      2961
##  7 Pre-war (<=1~ Escapist          822       6.91        6.9     7823.     2208.
##  8 Pre-war (<=1~ Heavy             762       6.86        6.9     7071.     2138.
##  9 War decade (~ Escapist          688       6.83        6.8     8517.     2584.
## 10 War decade (~ Heavy             804       6.89        6.9    10608.     2868.
```