MGSC 661 - Midterm Project

Importing data

```
theme_lox <- function() {
  theme(
    panel.grid.major.x = element_line(linewidth = 0.3, colour = "#cbcbcb"),
    panel.grid.major.y = element_line(linewidth = 0.3, colour = "#cbcbcb"),
    plot.title = element_markdown(
      family = "Helvetica",
      size = 22,
      face = "bold",
      color = "#222222"
    ),
    plot.subtitle = element_text(
      family = "Helvetica",
      size = 16,
      margin = margin(2, 0, 2, 0)
    ),
    plot.caption = element_text(family = "Helvetica", face = "bold"),
    axis.text = element_text(
      family = "Helvetica",
      size = 12,
      color = "#222222"
    ),
    axis.title = element_text(
      family = "Helvetica",
      size = 14,
      color = "#222222"
    ),
    legend.text = element_text(family = "Helvetica", size = 12),
    legend.title = element_text(
      family = "Helvetica",
      size = 14,
      face = "bold"
    ),
    legend.position = "right",
    strip.text = element_text(
      family = "Helvetica",
      size = 12,
      hjust = 0.5
    )
  )
}
```

```r
data <- read.csv("./IMDB_data_Fall_2023.csv")
attach(data)
```

Exploratory data analysis

```r
head(data)
```

```
##            movie_title movie_id
## 1 August: Osage County        2
## 2               Radio       12
## 3         Coach Carter       15
## 4       The Possession       20
## 5 Escape from Alcatraz       22
## 6        She's the Man       23
##                                          imdb_link imdb_score movie_budget
## 1 http://www.imdb.com/title/tt1322269/?ref_=fn_tt_tt_1        7.3     25000000
## 2 http://www.imdb.com/title/tt0316465/?ref_=fn_tt_tt_1        6.9     35000000
## 3 http://www.imdb.com/title/tt0393162/?ref_=fn_tt_tt_1        7.2     30000000
## 4 http://www.imdb.com/title/tt0431021/?ref_=fn_tt_tt_1        5.9     14000000
## 5 http://www.imdb.com/title/tt0079116/?ref_=fn_tt_tt_1        7.6      8000000
## 6 http://www.imdb.com/title/tt0454945/?ref_=fn_tt_tt_1        6.4     20000000
##   release_day release_month release_year duration language country
## 1          10           Jan         2014      121  English     USA
## 2          24           Oct         2003      109  English     USA
## 3          14           Jan         2005      136  English     USA
## 4          20           Aug         2012       92  English     USA
## 5          22           Jun         1979      112  English     USA
## 6          17           Mar         2006      105  English     USA
##   maturity_rating aspect_ratio                  distributor nb_news_articles
## 1               R         2.35         The Weinstein Company             2141
## 2              PG         1.85 Columbia Pictures Corporation              331
## 3           PG-13         2.35            Paramount Pictures              223
## 4           PG-13         2.35                     Lionsgate              620
## 5              PG         1.85            Paramount Pictures               97
## 6           PG-13         1.85       Lakeshore International              173
##         director               actor1 actor1_star_meter               actor2
## 1     John Wells Benedict Cumberbatch               259          Meryl Streep
## 2 Michael Tollin        Alfre Woodard              2735           Riley Smith
## 3  Thomas Carter       Channing Tatum               573         Rick Gonzalez
## 4   Ole Bornedal        Kyra Sedgwick              2047      Madison Davenport
## 5     Don Siegel       Clint Eastwood               102       Patrick McGoohan
## 6   Andy Fickman       Channing Tatum               573 Alexandra Breckenridge
##   actor2_star_meter          actor3 actor3_star_meter colour_film
## 1               559   Julia Roberts              513       Color
## 2              3915    Debra Winger             1845       Color
## 3              4793 Robert Ri'chard             6729       Color
## 4              1769   Natasha Calis            11963       Color
## 5              5062       Fred Ward             5451       Color
## 6               370    Laura Ramsey             3711       Color
##                  genres nb_faces
```

```
## 1                 Drama        3
## 2 Biography|Drama|Sport        1
## 3           Drama|Sport        0
## 4       Horror|Thriller        0
## 5 Biography|Crime|Drama        0
## 6        Comedy|Romance        0
##                                                          plot_keywords
## 1 based on play|incestuous relationship|pedophilia|secret|teenage daughter
## 2                          coach|football|football coach|high school|radio
## 3                   basketball|basketball coach|coach|contract|high school
## 4                              basketball coach|box|jewish|rabbi|yard sale
## 5                                    alcatraz|escape|inmate|island|prison
## 6                                    disguise|roommate|school|soccer|twin
##   action adventure scifi thriller musical romance western sport horror drama
## 1      0         0     0        0       0       0       0     0      0     1
## 2      0         0     0        0       0       0       0     1      0     1
## 3      0         0     0        0       0       0       0     1      0     1
## 4      0         0     0        1       0       0       0     0      1     0
## 5      0         0     0        0       0       0       0     0      0     1
## 6      0         0     0        0       0       1       0     0      0     0
##   war animation crime movie_meter_IMDBpro cinematographer    production_company
## 1   0         0     0                4000 Adriano Goldman The Weinstein Company
## 2   0         0     0                8556     Don Burgess    Revolution Studios
## 3   0         0     0                3940    Sharone Meir          Coach Carter
## 4   0         0     0                5452    Dan Laustsen  Ghost House Pictures
## 5   0         0     1                4722   Bruce Surtees    Paramount Pictures
## 6   0         0     0                2446   Greg Gardiner            DreamWorks
```

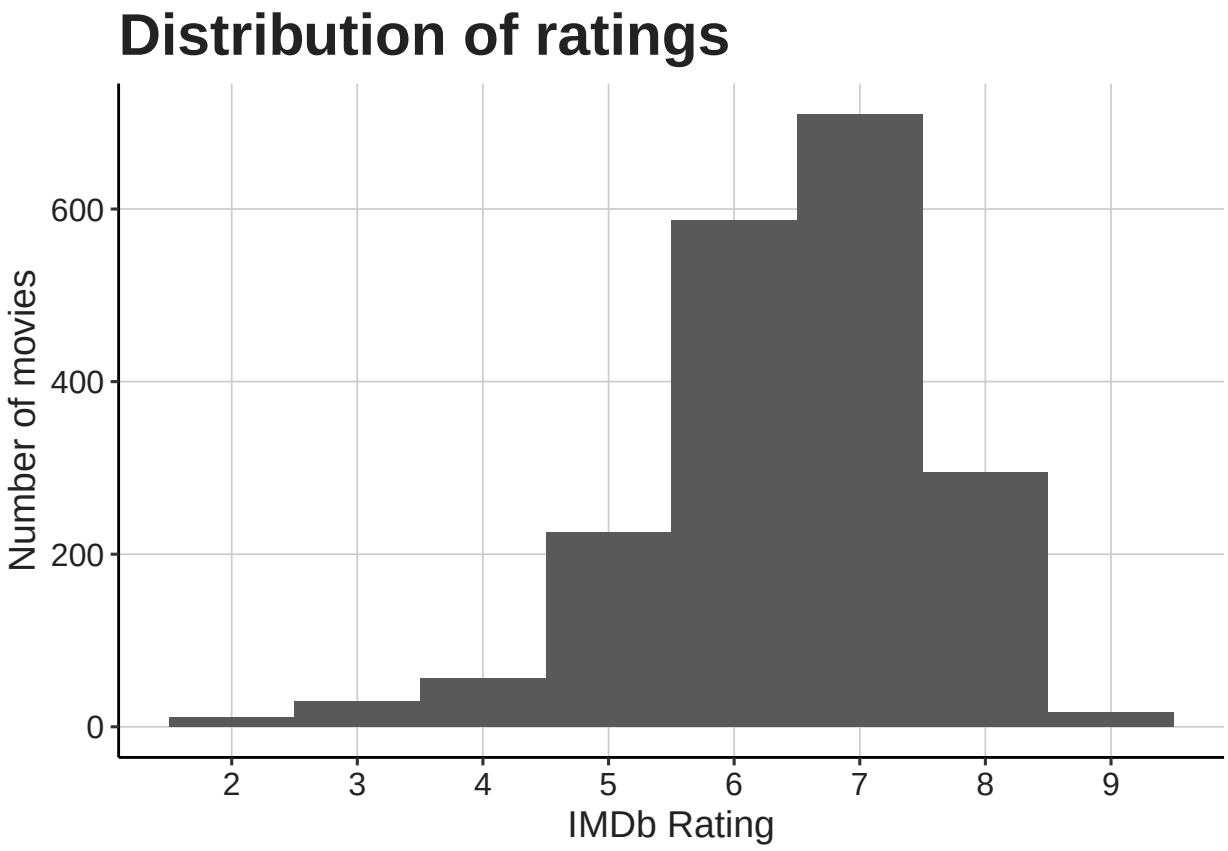
```
summary(data)
```

```
##  movie_title           movie_id      imdb_link           imdb_score
##  Length:1930        Min.   :    2   Length:1930        Min.   :1.900
##  Class :character   1st Qu.: 2528   Class :character   1st Qu.:5.900
##  Mode  :character   Median : 5802   Mode  :character   Median :6.600
##                     Mean   : 7067                      Mean   :6.512
##                     3rd Qu.:10604                      3rd Qu.:7.300
##                     Max.   :21838                      Max.   :9.300
##   movie_budget       release_day    release_month       release_year
##  Min.   :  560000   Min.   : 1.00   Length:1930        Min.   :1936
##  1st Qu.: 8725000   1st Qu.: 9.00   Class :character   1st Qu.:1997
##  Median :18000000   Median :17.00   Mode  :character   Median :2004
##  Mean   :20973774   Mean   :15.95                      Mean   :2001
##  3rd Qu.:30000000   3rd Qu.:23.00                      3rd Qu.:2010
##  Max.   :55000000   Max.   :30.00                      Max.   :2018
##     duration       language           country          maturity_rating
##  Min.   : 37.0   Length:1930        Length:1930        Length:1930
##  1st Qu.: 96.0   Class :character   Class :character   Class :character
##  Median :106.0   Mode  :character   Mode  :character   Mode  :character
##  Mean   :109.7
##  3rd Qu.:118.0
##  Max.   :330.0
##   aspect_ratio   distributor        nb_news_articles    director
##  Min.   :1.180   Length:1930        Min.   :    0.0   Length:1930
##  1st Qu.:1.850   Class :character   1st Qu.:   78.0   Class :character
```

```
##  Median :2.350   Mode  :character   Median :  286.0   Mode  :character
##  Mean   :2.096                     Mean   :  770.6
##  3rd Qu.:2.350                     3rd Qu.:  845.5
##  Max.   :2.760                     Max.   :60620.0
##     actor1          actor1_star_meter    actor2          actor2_star_meter
##  Length:1930        Min.   :      9   Length:1930        Min.   :      3
##  Class :character   1st Qu.:    505   Class :character   1st Qu.:   1895
##  Mode  :character   Median :   1888   Mode  :character   Median :   3986
##                     Mean   :  21190                      Mean   :  17114
##                     3rd Qu.:   4665                      3rd Qu.:   7667
##                     Max.   :8342201                      Max.   :5529461
##     actor3          actor3_star_meter colour_film          genres
##  Length:1930        Min.   :      8   Length:1930        Length:1930
##  Class :character   1st Qu.:   3075   Class :character   Class :character
##  Mode  :character   Median :   5856   Mode  :character   Mode  :character
##                     Mean   :  35469
##                     3rd Qu.:  12250
##                     Max.   :6292982
##     nb_faces     plot_keywords          action         adventure
##  Min.   : 0.00   Length:1930        Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 0.00   Class :character   1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 1.00   Mode  :character   Median :0.0000   Median :0.0000
##  Mean   : 1.44                      Mean   :0.2005   Mean   :0.1264
##  3rd Qu.: 2.00                      3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :31.00                      Max.   :1.0000   Max.   :1.0000
##      scifi           thriller         musical           romance
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.00000   Median :0.0000
##  Mean   :0.1083   Mean   :0.2979   Mean   :0.07047   Mean   :0.2451
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000
##     western            sport             horror          drama
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000   1st Qu.:0.0000
##  Median :0.00000   Median :0.00000   Median :0.000   Median :1.0000
##  Mean   :0.01762   Mean   :0.04819   Mean   :0.113   Mean   :0.5492
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.000   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.000   Max.   :1.0000
##       war            animation           crime        movie_meter_IMDBpro
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :    71
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:  2836
##  Median :0.00000   Median :0.00000   Median :0.0000   Median :  5406
##  Mean   :0.03627   Mean   :0.01036   Mean   :0.2161   Mean   : 11612
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.: 10198
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :849550
##  cinematographer    production_company
##  Length:1930        Length:1930
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

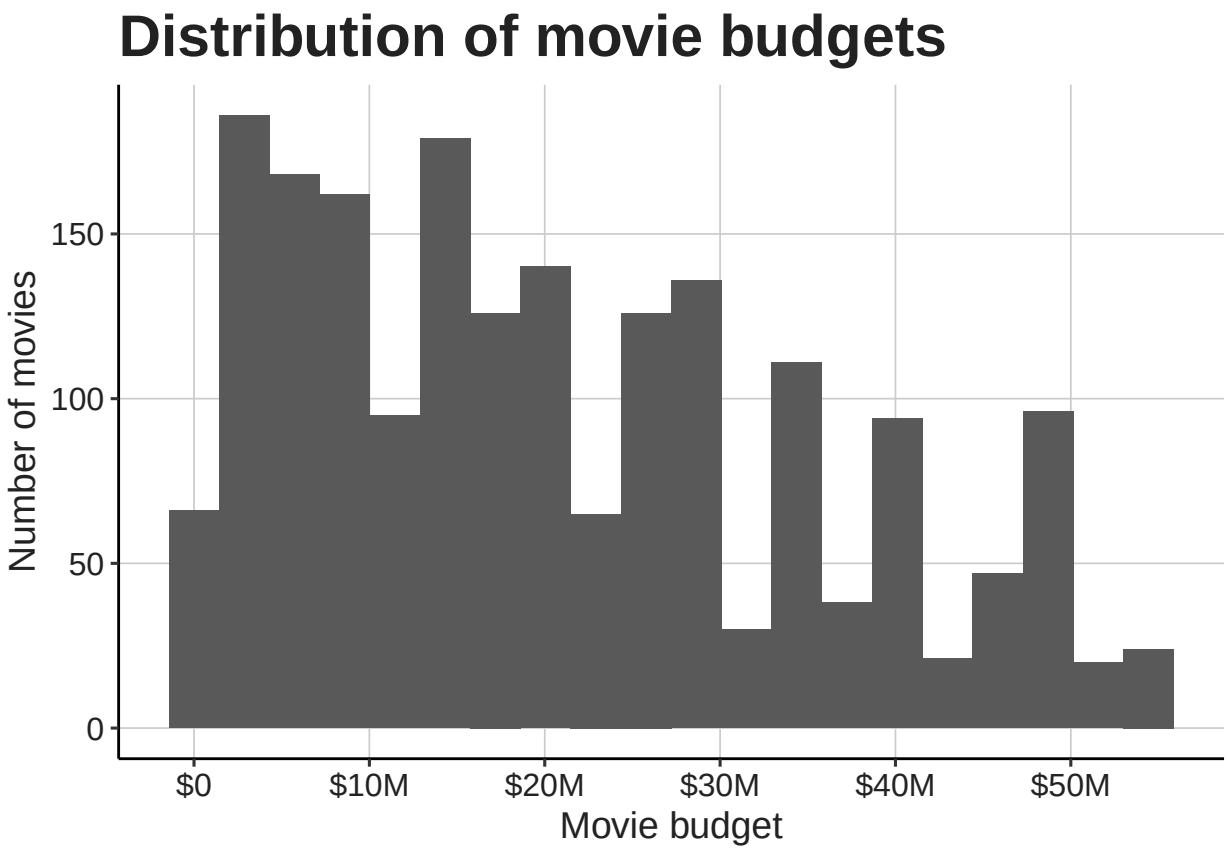
$y =$ **IMDB Score**

```
ggplot(data, aes(x = imdb_score)) +
  geom_histogram(binwidth = 1) +
  scale_x_continuous(breaks = breaks_width(width = 1)) +
  labs(x = "IMDb Rating", y = "Number of movies", title = "Distribution of ratings") +
  theme_pubr() +
  theme_lox()
```



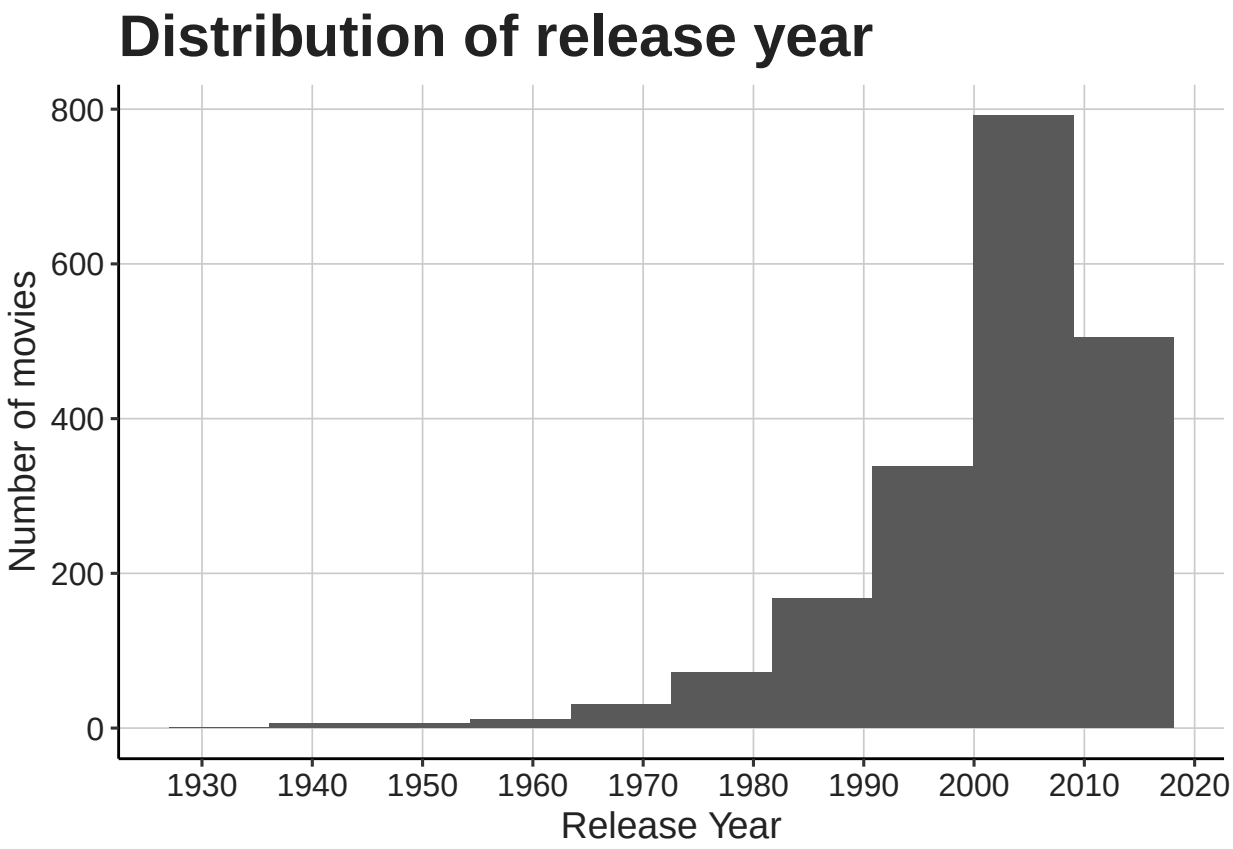
Movie budget

```
ggplot(data, aes(x = movie_budget)) +
  geom_histogram(bins = 20) +
  scale_x_continuous(breaks = breaks_pretty(), labels = label_dollar(scale_cut = cut_short_scale())) +
  labs(x = "Movie budget", y = "Number of movies", title = "Distribution of movie budgets") +
  theme_pubr() +
  theme_lox()
```



Release year

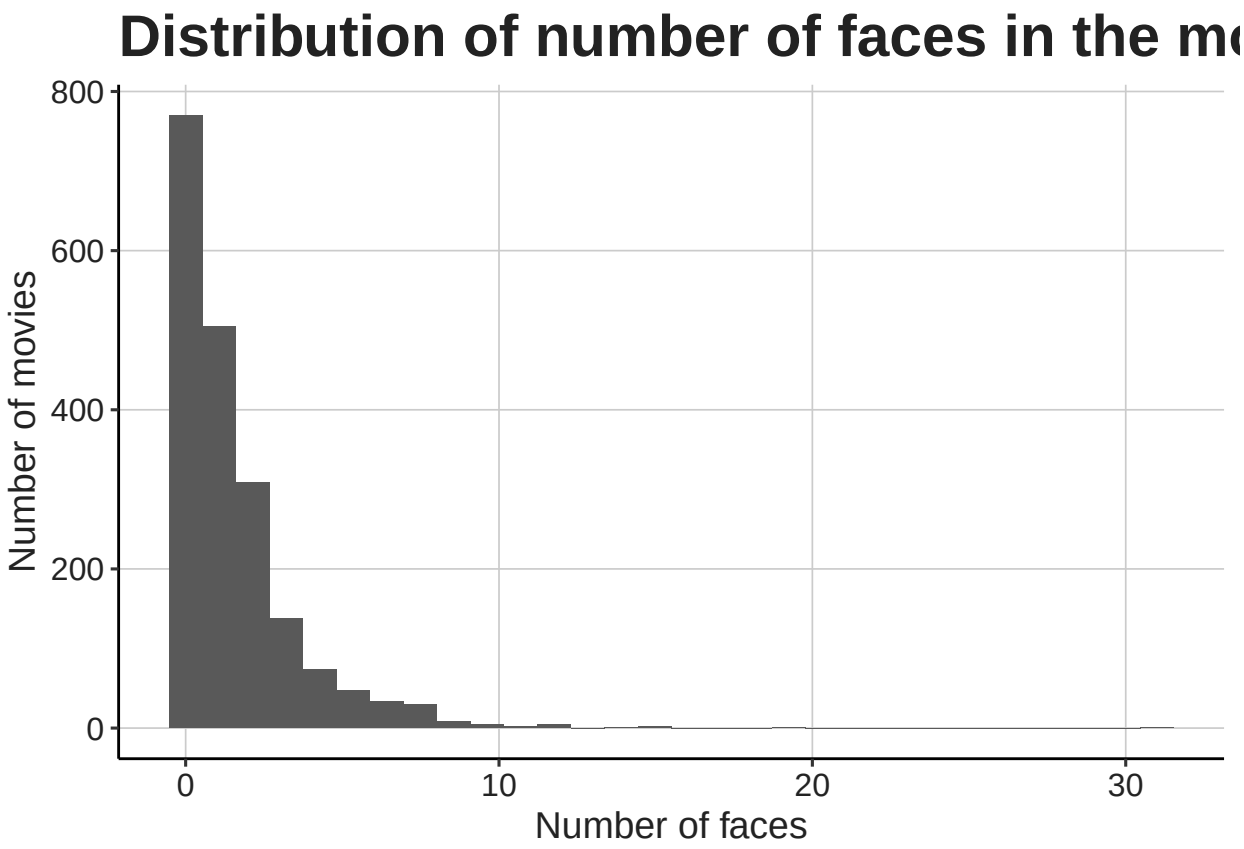
```
ggplot(data, aes(x = release_year)) +
  geom_histogram(bins = 10) +
  scale_x_continuous(breaks = breaks_pretty(n = 10)) +
  labs(x = "Release Year", y = "Number of movies", title = "Distribution of release year") +
  theme_pubr() +
  theme_lox()
```



Number of faces

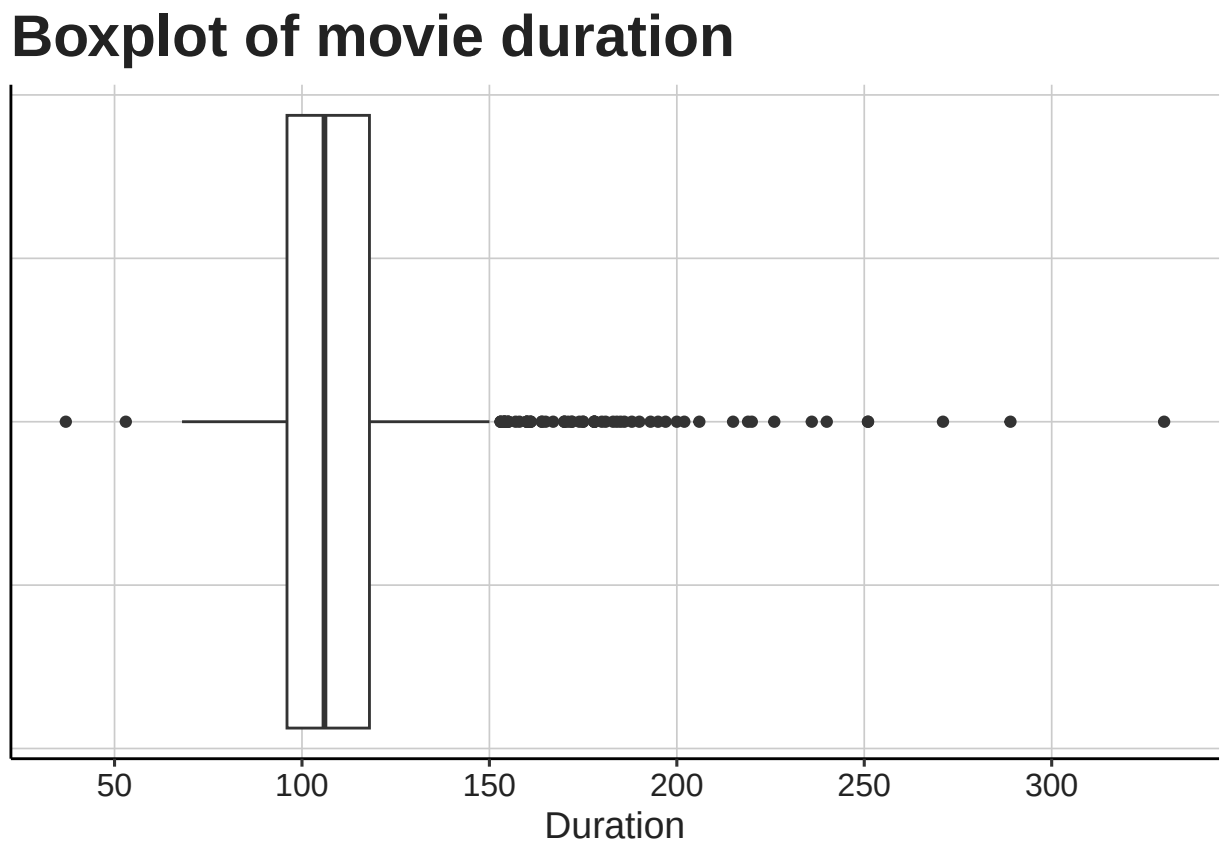
```
ggplot(data, aes(x = nb_faces)) +
  geom_histogram() +
  scale_x_continuous(breaks = breaks_pretty()) +
  labs(x = "Number of faces", y = "Number of movies", title = "Distribution of number of faces in the m
  theme_pubr() +
  theme_lox()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Duration

```
ggplot(data, aes(x = duration)) +
  geom_boxplot() +
  scale_x_continuous(breaks = breaks_pretty()) +
  labs(x = "Duration", title = "Boxplot of movie duration") +
  theme_pubr() +
  theme_lox() +
  theme(axis.ticks.y = element_blank(), axis.text.y = element_blank())
```

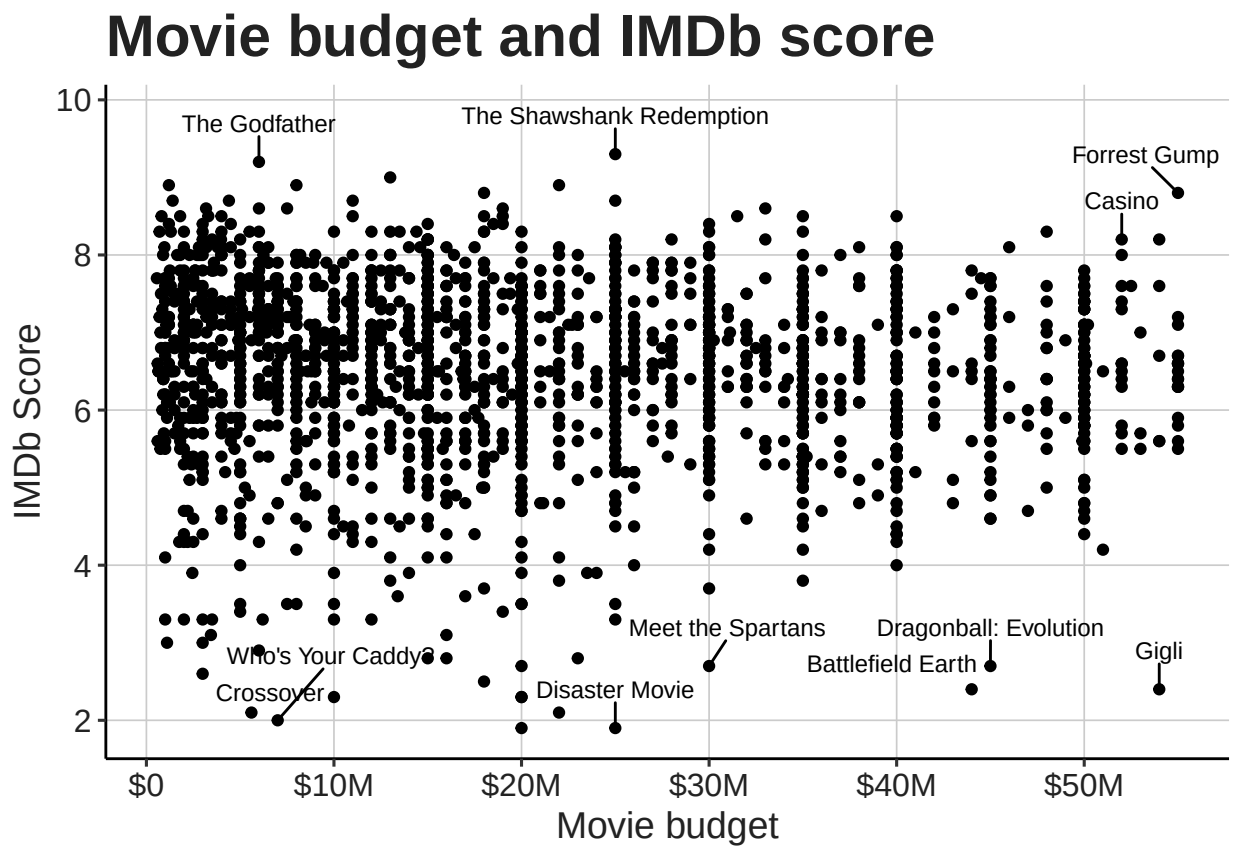


Bivariate distributions

Movie budgets $\sim y$

```
ggplot(data, aes(x = movie_budget, y = imdb_score)) +
  geom_point() +
  geom_text_repel(aes(label = movie_title), size = 3, max.overlaps = 5, nudge_y = 0.5) +
  scale_x_continuous(breaks = breaks_pretty(), labels = label_dollar(scale_cut = cut_short_scale())) +
  labs(x = "Movie budget", y = "IMDb Score", title = "Movie budget and IMDb score") +
  theme_pubr() +
  theme_lox()
```

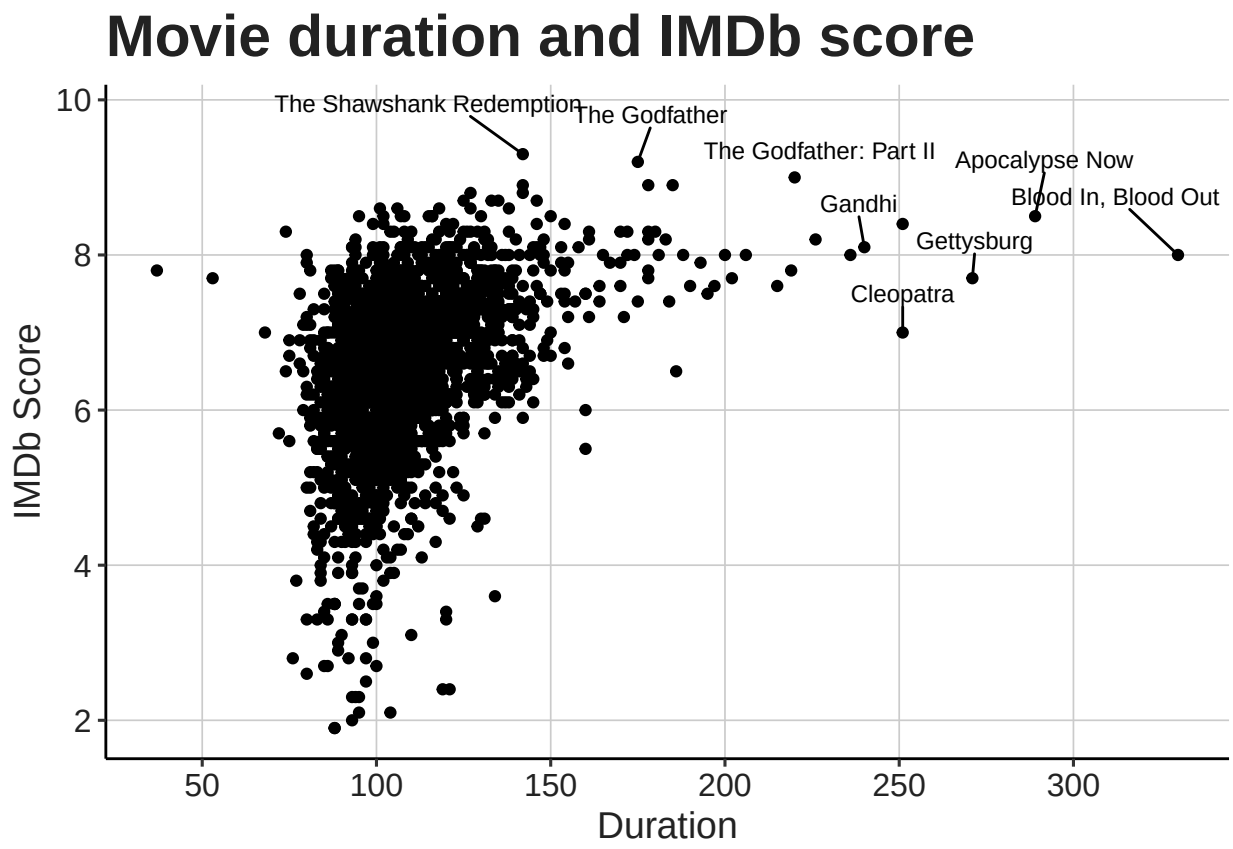
```
## Warning: ggrepel: 1919 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Duration $\sim y$

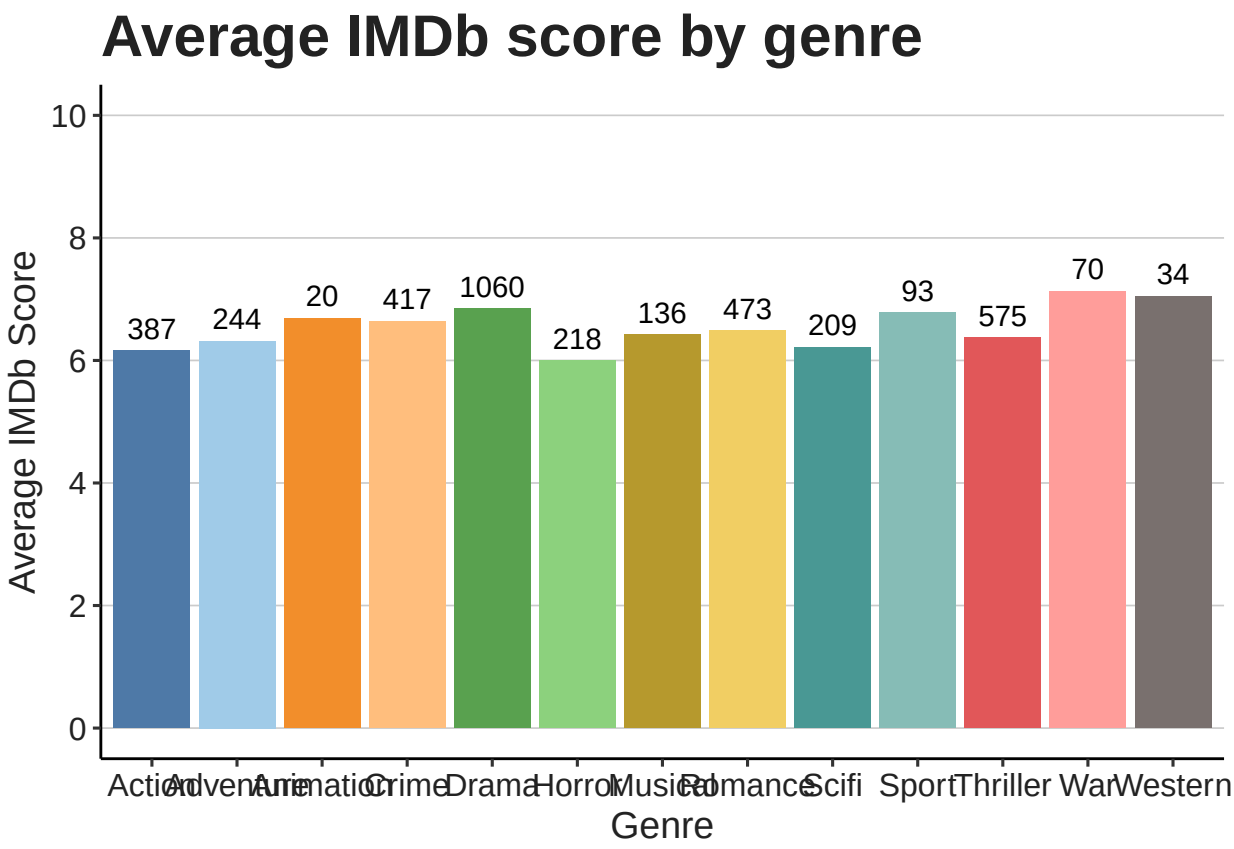
```
ggplot(data, aes(x = duration, y = imdb_score)) +
  geom_point() +
  geom_text_repel(aes(label = movie_title), size = 3, max.overlaps = 10, nudge_y = 0.5) +
  scale_x_continuous(breaks = breaks_pretty(), labels = label_number()) +
  labs(x = "Duration", y = "IMDb Score", title = "Movie duration and IMDb score") +
  theme_pubr() +
  theme_lox()
```

```
## Warning: ggrepel: 1922 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

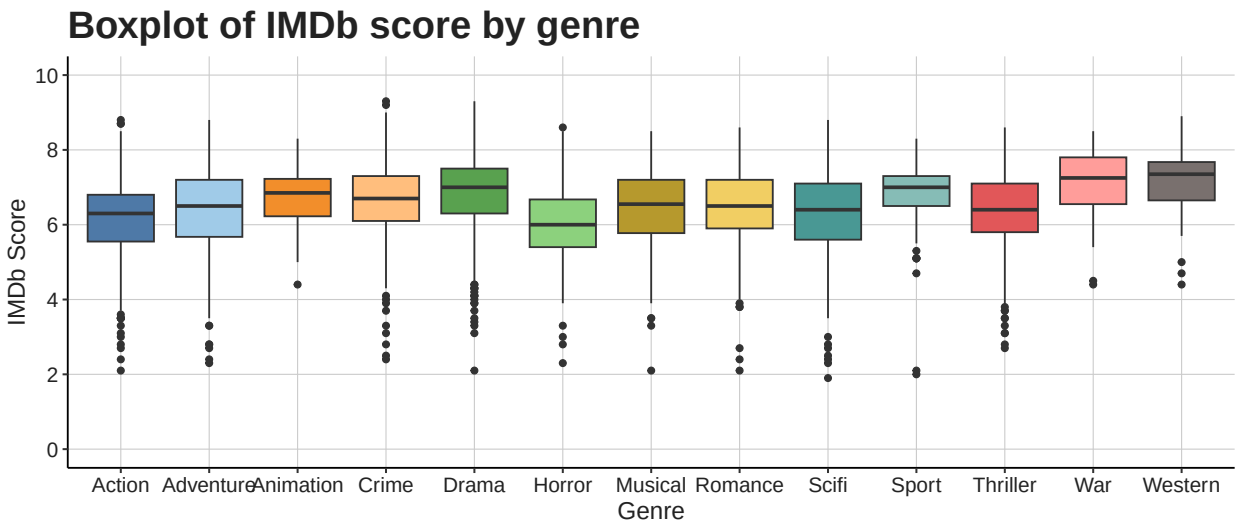


Genre $\sim y$

```
data %>%
  select(movie_id, imdb_score, action:crime) %>%
  pivot_longer(cols = c(-imdb_score, -movie_id), names_to = "genre") %>%
  mutate(genre = str_to_sentence(genre)) %>%
  group_by(genre) %>%
  filter(value == 1) %>%
  summarise(avg_score = mean(imdb_score), num_movies = n()) %>%
  ggplot(aes(x = genre, y = avg_score, fill = genre)) +
  geom_col() +
  geom_text_repel(aes(y = avg_score, label = num_movies), nudge_y = 0.1) +
  scale_fill_tableau(palette = "Tableau 20") +
  scale_y_continuous(breaks = breaks_pretty(), limits = c(0, 10)) +
  guides(fill = "none") +
  labs(x = "Genre", y = "Average IMDb Score", title = "Average IMDb score by genre") +
  theme_pubr() +
  theme_lox() +
  theme(panel.grid.major.x = element_blank())
```



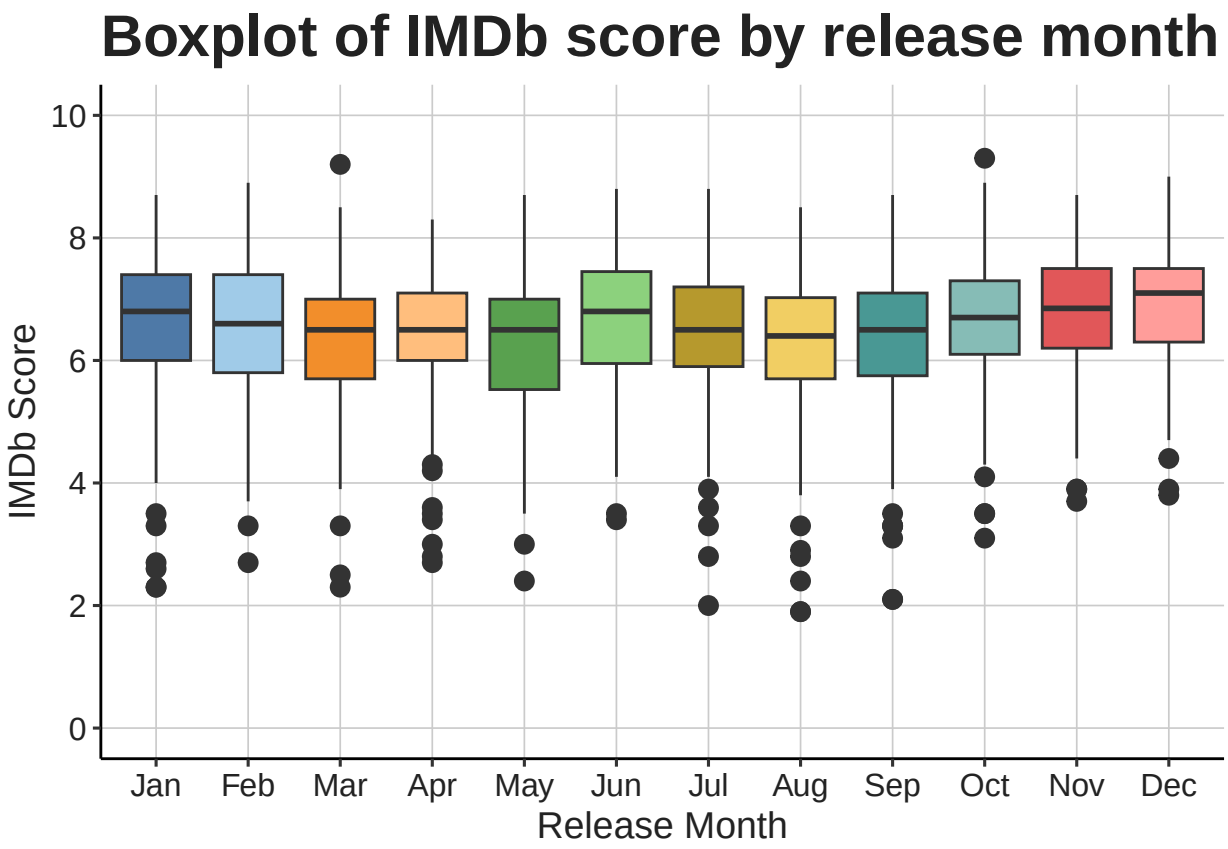

```
data %>%
  select(movie_id, movie_title, imdb_score, action:crime) %>%
  pivot_longer(cols = c(-imdb_score, -movie_id, -movie_title), names_to = "genre") %>%
  mutate(genre = str_to_sentence(genre)) %>%
  group_by(genre) %>%
  filter(value == 1) %>%
  ggplot(aes(x = genre, y = imdb_score, fill = genre)) +
  geom_boxplot() +
  scale_fill_tableau(palette = "Tableau 20") +
  scale_y_continuous(breaks = breaks_pretty(), limits = c(0, 10)) +
  guides(fill = "none") +
  labs(x = "Genre", y = "IMDb Score", title = "Boxplot of IMDb score by genre", caption = "Movies span a
  theme_pubr() +
  theme_lox()
```



Release Month $\sim y$

```
data %>%
  mutate(release_date = dmy(str_c(release_day, release_month, release_year, sep = "-"))) %>%
  mutate(release_month = month(release_date, label = TRUE)) %>%
  ggplot(aes(x = release_month, y = imdb_score, fill = release_month)) +
  geom_boxplot(outlier.size = 3) +
  geom_text_repel(aes(label = movie_title), max.overlaps = 5) +
  scale_fill_tableau(palette = "Tableau 20") +
  scale_y_continuous(breaks = breaks_pretty(), limits = c(0, 10)) +
  guides(fill = "none") +
  labs(x = "Release Month", y = "IMDb Score", title = "Boxplot of IMDb score by release month") +
  theme_pubr() +
  theme_lox()
```

```
## Warning: ggrepel: 1930 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

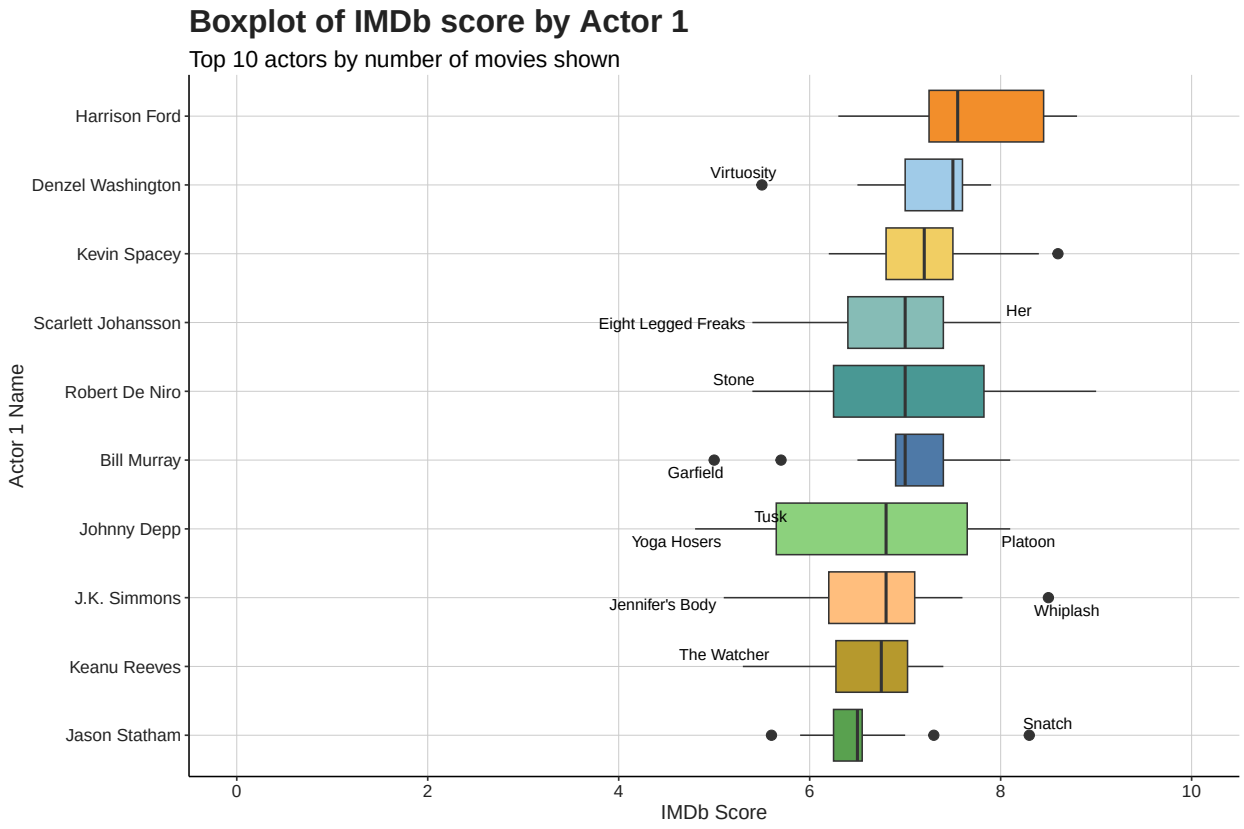


Actor 1 $\sim y$

```
data %>%
  filter(actor1 %in% (
    data %>%
      group_by(actor1) %>%
      count(sort = TRUE) %>%
      head(10)
  )$actor1) %>%
  ggplot(aes(
    x = fct_reorder(actor1, imdb_score, .fun = median),
    y = imdb_score,
    fill = actor1
  )) +
  geom_boxplot(outlier.size = 3) +
  geom_text_repel(aes(label = movie_title), max.overlaps = 7) +
  scale_fill_tableau(palette = "Tableau 20") +
  scale_y_continuous(breaks = breaks_pretty(), limits = c(0, 10)) +
  guides(fill = "none") +
  labs(
    x = "Actor 1 Name",
    y = "IMDb Score",
    title = "Boxplot of IMDb score by Actor 1",
    subtitle = "Top 10 actors by number of movies shown"
```

```
  ) +
  theme_pubr() +
  theme_lox() +
  coord_flip()
```

```
## Warning: ggrepel: 150 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

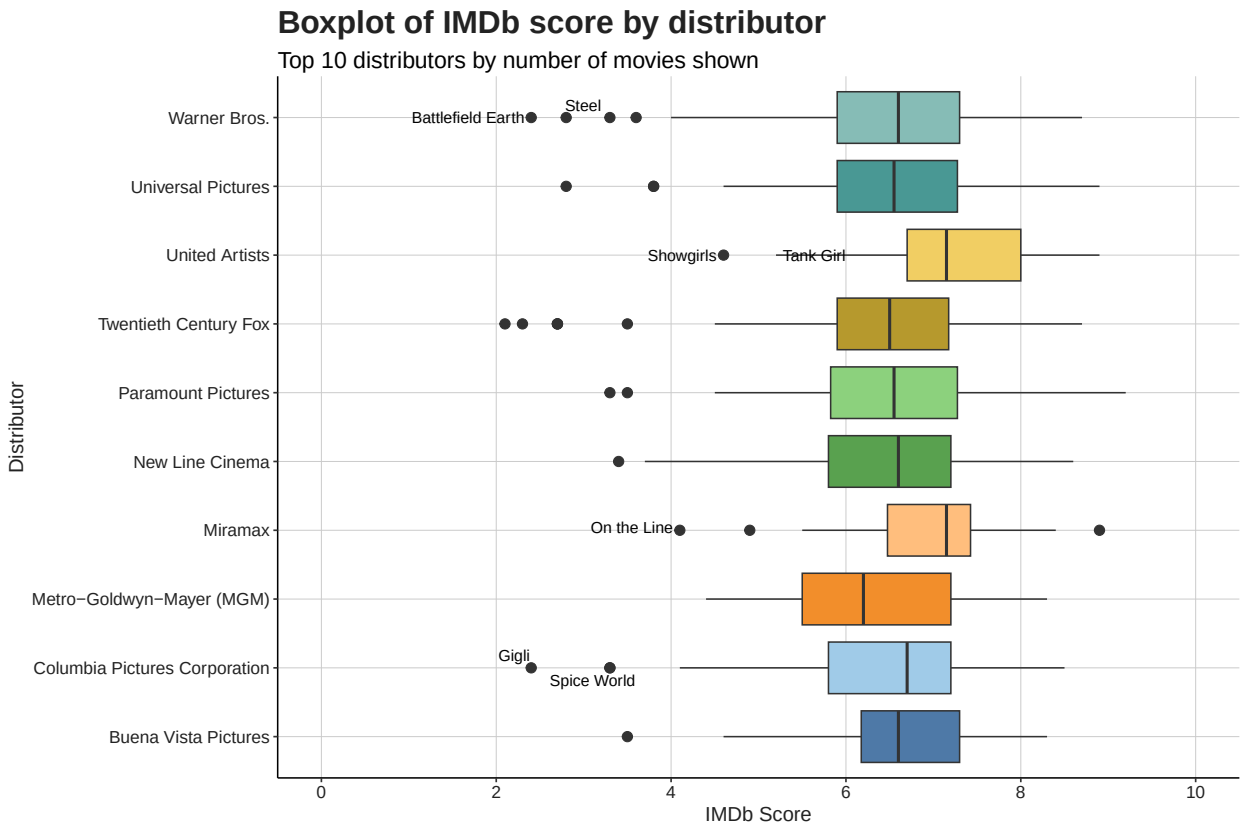


Distributor $\sim y$

```
data %>%
  filter(distributor %in%
           (data %>%
              group_by(distributor) %>%
              count(sort = TRUE) %>%
              head(10))$distributor) %>%
  ggplot(aes(x = distributor, y = imdb_score, fill = distributor)) +
  geom_boxplot(outlier.size = 3) +
  geom_text_repel(aes(label = movie_title), max.overlaps = 5) +
  scale_fill_tableau(palette = "Tableau 20") +
  scale_y_continuous(breaks = breaks_pretty(), limits = c(0, 10)) +
  guides(fill = "none") +
  labs(
    x = "Distributor",
```

```
    y = "IMDb Score",
    title = "Boxplot of IMDb score by distributor",
    subtitle = "Top 10 distributors by number of movies shown"
  ) +
  theme_pubr() +
  theme_lox() +
  coord_flip()
```

```
## Warning: ggrepel: 941 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Data preprocessing

Checking data types

```
data %>% str()
```

```
## 'data.frame':    1930 obs. of  42 variables:
##  $ movie_title        : chr  "August: Osage County" "Radio" "Coach Carter" "The Possession" ...
##  $ movie_id           : int  2 12 15 20 22 23 26 31 38 39 ...
##  $ imdb_link          : chr  "http://www.imdb.com/title/tt1322269/?ref_=fn_tt_tt_1" "http://www.imdb
##  $ imdb_score         : num  7.3 6.9 7.2 5.9 7.6 6.4 7.1 8.1 7.1 6.5 ...
```

```
##  $ movie_budget       : int  25000000 35000000 30000000 14000000 8000000 20000000 22700000 25000000 4
##  $ release_day        : int  10 24 14 20 22 17 24 21 21 14 ...
##  $ release_month      : chr  "Jan" "Oct" "Jan" "Aug" ...
##  $ release_year       : int  2014 2003 2005 2012 1979 2006 1987 2007 1998 2007 ...
##  $ duration           : int  121 109 136 92 112 105 96 122 110 95 ...
##  $ language           : chr  "English" "English" "English" "English" ...
##  $ country            : chr  "USA" "USA" "USA" "USA" ...
##  $ maturity_rating    : chr  "R" "PG" "PG-13" "PG-13" ...
##  $ aspect_ratio       : num  2.35 1.85 2.35 2.35 1.85 1.85 1.85 2.35 2.35 1.85 ...
##  $ distributor        : chr  "The Weinstein Company" "Columbia Pictures Corporation" "Paramount Pictu
##  $ nb_news_articles   : int  2141 331 223 620 97 173 408 4135 1723 378 ...
##  $ director           : chr  "John Wells" "Michael Tollin" "Thomas Carter" "Ole Bornedal" ...
##  $ actor1             : chr  "Benedict Cumberbatch" "Alfre Woodard" "Channing Tatum" "Kyra Sedgwick"
##  $ actor1_star_meter  : int  259 2735 573 2047 102 573 12294 628 547 358742 ...
##  $ actor2             : chr  "Meryl Streep" "Riley Smith" "Rick Gonzalez" "Madison Davenport" ...
##  $ actor2_star_meter  : int  559 3915 4793 1769 5062 370 13732 2450 1054 3086 ...
##  $ actor3             : chr  "Julia Roberts" "Debra Winger" "Robert Ri'chard" "Natasha Calis" ...
##  $ actor3_star_meter  : int  513 1845 6729 11963 5451 3711 8419 3592 3001 642 ...
##  $ colour_film        : chr  "Color" "Color" "Color" "Color" ...
##  $ genres             : chr  "Drama" "Biography|Drama|Sport" "Drama|Sport" "Horror|Thriller" ...
##  $ nb_faces           : int  3 1 0 0 0 0 2 0 1 4 ...
##  $ plot_keywords      : chr  "based on play|incestuous relationship|pedophilia|secret|teenage daught
##  $ action             : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ adventure          : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ scifi              : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ thriller           : int  0 0 0 1 0 0 0 1 0 0 ...
##  $ musical            : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ romance            : int  0 0 0 0 0 1 0 0 0 1 ...
##  $ western            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ sport              : int  0 1 1 0 0 0 0 0 0 0 ...
##  $ horror             : int  0 0 0 1 0 0 0 0 1 0 ...
##  $ drama              : int  1 1 1 0 1 0 0 1 0 0 ...
##  $ war                : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ animation          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ crime              : int  0 0 0 0 1 0 0 1 0 0 ...
##  $ movie_meter_IMDBpro: int  4000 8556 3940 5452 4722 2446 2294 513 697 6854 ...
##  $ cinematographer    : chr  "Adriano Goldman" "Don Burgess" "Sharone Meir" "Dan Laustsen" ...
##  $ production_company : chr  "The Weinstein Company" "Revolution Studios" "Coach Carter" "Ghost House
```

```
data <- data %>%
  mutate(across(.cols = c('language', 'country', 'maturity_rating', 'aspect_ratio', 'distributor', 'dire
```

Checking for multicollinearity

```
vif_model <- lm(imdb_score ~ ., data = (data %>% select(where(is.numeric))))

vif(vif_model)
```

```
##            movie_id        movie_budget         release_day        release_year
##            1.031760            1.252467            1.009593            1.251345
##            duration    nb_news_articles   actor1_star_meter   actor2_star_meter
```

```
##            1.435309            1.044031            1.041778            1.165638
##   actor3_star_meter           nb_faces              action           adventure
##            1.114109            1.062393            1.426011            1.270538
##               scifi            thriller             musical             romance
##            1.226770            1.477230            1.061118            1.175954
##             western               sport             horror               drama
##            1.055618            1.086962            1.330801            1.450746
##                 war           animation               crime movie_meter_IMDBpro
##            1.112120            1.079489            1.373523            1.029943
```

We see that there is no multicollinearity among the numeric variables in the dataset.

Feature engineering

Checking levels for categorical columns

```
categorical_columns <- data %>%
  summarize(across(where(is.factor), ~nlevels(.x))) %>%
  pivot_longer(everything(), names_to = "variable", values_to = "num_levels")

categorical_columns
```

```
## # A tibble: 9 x 2
##   variable           num_levels
##   <chr>                   <int>
## 1 language                   19
## 2 country                    34
## 3 maturity_rating            12
## 4 aspect_ratio               14
## 5 distributor               334
## 6 director                 1115
## 7 colour_film                 2
## 8 cinematographer           737
## 9 production_company        768
```

We see that director, cinematographer, and production company have a lot of unique values. A priori, we expect to drop these columns when building the model.

Let us check the counts for the other variables

```
categorical_columns %>%
  filter(!variable %in% c("director", "cinematographer", "production_company")) %>%
  pull(variable) %>%
  walk(
    ~ data %>%
      group_by_at(.x) %>%
      count(sort = TRUE) %>%
      print()
  )
```

```
## # A tibble: 19 x 2
```

```
## # Groups:   language [19]
##    language       n
##    <fct>      <int>
##  1 English     1892
##  2 French         7
##  3 Spanish        6
##  4 German         3
##  5 Italian        3
##  6 Cantonese      2
##  7 Japanese       2
##  8 Mandarin       2
##  9 None           2
## 10 Zulu           2
## 11 Aboriginal     1
## 12 Aramaic        1
## 13 Dari           1
## 14 Dutch          1
## 15 Hindi          1
## 16 Indonesian     1
## 17 Korean         1
## 18 Mongolian      1
## 19 Portuguese     1
## # A tibble: 34 x 2
## # Groups:   country [34]
##    country       n
##    <fct>     <int>
##  1 USA        1555
##  2 UK          177
##  3 France       40
##  4 Canada       38
##  5 Germany      34
##  6 Australia    23
##  7 Italy         8
##  8 Spain         7
##  9 Ireland       5
## 10 Japan         5
## # i 24 more rows
## # A tibble: 12 x 2
## # Groups:   maturity_rating [12]
##    maturity_rating     n
##    <fct>           <int>
##  1 R                1013
##  2 PG-13             582
##  3 PG                255
##  4 G                  34
##  5 Approved           21
##  6 X                   8
##  7 Passed              4
##  8 NC-17               3
##  9 TV-14               3
## 10 TV-G                3
## 11 GP                  2
## 12 M                   2
## # A tibble: 14 x 2
```

```
## # Groups:   aspect_ratio [14]
##    aspect_ratio     n
##    <fct>        <int>
##  1 2.35           981
##  2 1.85           853
##  3 1.37            28
##  4 1.78            18
##  5 1.66            17
##  6 1.33            11
##  7 2.39             7
##  8 2.2              6
##  9 2.4              3
## 10 1.75             2
## 11 1.18             1
## 12 1.5              1
## 13 2.55             1
## 14 2.76             1
## # A tibble: 334 x 2
## # Groups:   distributor [334]
##    distributor                        n
##    <fct>                          <int>
##  1 Warner Bros.                     169
##  2 Universal Pictures               146
##  3 Paramount Pictures               138
##  4 Twentieth Century Fox            126
##  5 Columbia Pictures Corporation    113
##  6 New Line Cinema                   73
##  7 Buena Vista Pictures              60
##  8 Miramax                           44
##  9 United Artists                    40
## 10 Metro-Goldwyn-Mayer (MGM)         39
## # i 324 more rows
## # A tibble: 2 x 2
## # Groups:   colour_film [2]
##   colour_film         n
##   <fct>           <int>
## 1 Color            1867
## 2 Black and White    63
```

We see that:

- The language is primarily "English"
- The country is primarily "USA"

These features can also be dropped when building the model.

For distributor and plot keywords, we can create binary features for the top 10 values by count.

Top 10 keywords

```
top_10_keywords <- data %>%
  select(plot_keywords) %>%
```

```
  separate_longer_delim(cols = "plot_keywords", delim = "|") %>%
  group_by(plot_keywords) %>%
  count(sort = TRUE) %>%
  head(10)

for (keyword in top_10_keywords$plot_keywords) {
  col_name <- glue("plot_{keyword}")
  data[[col_name]] <-
    as.integer(lapply(data$plot_keywords, function (x) {
      str_detect(x, keyword)
    }))
}
```

Top 10 distributors

```
top_10_distributors <- data %>%
  select(distributor) %>%
  separate_longer_delim(cols = "distributor", delim = "|") %>%
  group_by(distributor) %>%
  count(sort = TRUE) %>%
  head(10)

for (distributor in top_10_distributors$distributor) {
  col_name <- glue("distributor_{distributor}")
  data[[col_name]] <-
    as.integer(lapply(data$distributor, function (x) {
      str_detect(x, distributor)
    }))
}
```

Removing variables

```
columns_to_remove <-
  c(
    "imdb_link",
    "actor1",
    "actor2",
    "actor3",
    "genres",
    "release_year",
    "director",
    "cinematographer",
    "production_company",
    "language",
    "country",
    "distributor",
    "plot_keywords"
  )

data <- data %>% select(-all_of(columns_to_remove))
```

```
data %>% head()
```

```
##            movie_title movie_id imdb_score movie_budget release_day
## 1 August: Osage County        2        7.3     25000000          10
## 2                Radio       12        6.9     35000000          24
## 3         Coach Carter       15        7.2     30000000          14
## 4       The Possession       20        5.9     14000000          20
## 5 Escape from Alcatraz       22        7.6      8000000          22
## 6        She's the Man       23        6.4     20000000          17
##   release_month duration maturity_rating aspect_ratio nb_news_articles
## 1           Jan      121               R         2.35             2141
## 2           Oct      109              PG         1.85              331
## 3           Jan      136           PG-13         2.35              223
## 4           Aug       92           PG-13         2.35              620
## 5           Jun      112              PG         1.85               97
## 6           Mar      105           PG-13         1.85              173
##   actor1_star_meter actor2_star_meter actor3_star_meter colour_film nb_faces
## 1               259               559               513       Color        3
## 2              2735              3915              1845       Color        1
## 3               573              4793              6729       Color        0
## 4              2047              1769             11963       Color        0
## 5               102              5062              5451       Color        0
## 6               573               370              3711       Color        0
##   action adventure scifi thriller musical romance western sport horror drama
## 1      0         0     0        0       0       0       0     0      0     1
## 2      0         0     0        0       0       0       0     1      0     1
## 3      0         0     0        0       0       0       0     1      0     1
## 4      0         0     0        1       0       0       0     0      1     0
## 5      0         0     0        0       0       0       0     0      0     1
## 6      0         0     0        0       0       1       0     0      0     0
##   war animation crime movie_meter_IMDBpro plot_murder plot_love plot_friend
## 1   0         0     0               4000           0         0           0
## 2   0         0     0               8556           0         0           0
## 3   0         0     0               3940           0         0           0
## 4   0         0     0               5452           0         0           0
## 5   0         0     1               4722           0         0           0
## 6   0         0     0               2446           0         0           0
##   plot_death plot_high school plot_police plot_new york city plot_boy
## 1          0                0           0                  0        0
## 2          0                1           0                  0        0
## 3          0                1           0                  0        0
## 4          0                0           0                  0        0
## 5          0                0           0                  0        0
## 6          0                0           0                  0        0
##   plot_drugs plot_school distributor_Warner Bros.
## 1          0           0                       0
## 2          0           1                       0
## 3          0           1                       0
## 4          0           0                       0
## 5          0           0                       0
## 6          0           1                       0
##   distributor_Universal Pictures distributor_Paramount Pictures
## 1                              0                              0
```

```
## 2                              0                              0
## 3                              0                              1
## 4                              0                              0
## 5                              0                              1
## 6                              0                              0
##   distributor_Twentieth Century Fox distributor_Columbia Pictures Corporation
## 1                                 0                                         0
## 2                                 0                                         1
## 3                                 0                                         0
## 4                                 0                                         0
## 5                                 0                                         0
## 6                                 0                                         0
##   distributor_New Line Cinema distributor_Buena Vista Pictures
## 1                           0                                0
## 2                           0                                0
## 3                           0                                0
## 4                           0                                0
## 5                           0                                0
## 6                           0                                0
##   distributor_Miramax distributor_United Artists
## 1                   0                         0
## 2                   0                         0
## 3                   0                         0
## 4                   0                         0
## 5                   0                         0
## 6                   0                         0
##   distributor_Metro-Goldwyn-Mayer (MGM)
## 1                                     0
## 2                                     0
## 3                                     0
## 4                                     0
## 5                                     0
## 6                                     0
```