Assignment 4 - Group 4

true

Contents

```
library(tidyverse)
library(dplyr)
library(kableExtra)
library(bookdown)
```

```
dat = read.csv("Data/IMDb movies.csv")
glimpse(dat)
```

```
## Rows: 81,273
## Columns: 22
## $ imdb_title_id         <chr> "tt0000574", "tt0001892", "tt0002101", "tt000...
## $ title                 <chr> "The Story of the Kelly Gang", "Den sorte drÃ...
## $ original_title        <chr> "The Story of the Kelly Gang", "Den sorte drÃ...
## $ year                  <int> 1906, 1911, 1912, 1911, 1912, 1919, 1913, 191...
## $ date_published        <chr> "1906-12-26", "1911-08-19", "1912-11-13", "19...
## $ genre                 <chr> "Biography, Crime, Drama", "Drama", "Drama, H...
## $ duration              <int> 70, 53, 100, 68, 60, 85, 120, 120, 55, 121, 5...
## $ country               <chr> "Australia", "Germany, Denmark", "USA", "Ital...
## $ language              <chr> "", "", "English", "Italian", "English", "Ger...
## $ director              <chr> "Charles Tait", "Urban Gad", "Charles L. Gask...
## $ writer                <chr> "Charles Tait", "Urban Gad, Gebhard SchÃ¤tzle...
## $ production_company    <chr> "J. and N. Tait", "Fotorama", "Helen Gardner ...
## $ actors                <chr> "Elizabeth Tait, John Tait, Norman Campbell, ...
## $ description           <chr> "True story of notorious Australian outlaw Ne...
## $ avg_vote              <dbl> 6.1, 5.9, 5.2, 7.0, 5.7, 6.8, 6.2, 6.7, 5.5, ...
## $ votes                 <int> 537, 171, 420, 2019, 438, 709, 241, 187, 211,...
## $ budget                <chr> "$ 2250", "", "$ 45000", "", "", "", "ITL 450...
## $ usa_gross_income      <chr> "", "", "", "", "", "", "", "", "", "", "", "...
```

Table 1: Metascore summary

	metascore
	Min. : 1.00
	1st Qu.: 43.00
	Median : 56.00
	Mean : 55.76
	3rd Qu.: 69.00
	Max. :100.00
	NA's :68551

```
## $ worlwide_gross_income <chr> "", "", "", "", "", "", "", "", "", "", "", "...
## $ metascore             <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ reviews_from_users    <dbl> 7, 4, 24, 28, 12, 11, 6, 3, 7, 9, 9, 16, 8, 2...
## $ reviews_from_critics  <dbl> 7, 2, 3, 14, 5, 9, 4, 1, 1, 9, 29, 7, 22, 2, ...
```

```
dat %>%
        dplyr::select(metascore) %>%
        summary() %>%
        kable(caption = "Metascore summary") %>%
        kable_styling(bootstrap_options = c("striped", "hover"))
```

According to IMDb, metascore is a single number that represents the overall critical opinion about a movie. Its legitimacy is derived from applying a weighted average to the reviews of the world's most respected critics. That said, metascore is still an extremely vague concept - particularly when it comes to the threshold of what metascore can be considered good.

Therefore, it is important to first consider the distribution of metascore within the data set. This is summarised by Table 1. The most noteworthy features are that the median and mean are very close at 56 and 55.76 respectively and nearly 70,000 movies have not been assigned a metascore. Consequently, any analysis related to the metascore variable must be treated with caution as less than 20% of the movies are represented.

```
f1wn = dat %>%
             ggplot(aes(x = year, y = metascore)) +
             geom_point()
f1wn
```

As expected, the scatter plot depicted in Figure 1 is not particularly informative and potentially misleading. Although the spread is increasing, one can still identify a negative correlation which suggests that the quality of movies has degraded over time. However, it is important to recognise that this dataset was taken from IMDb which means that sampling was far from random. IMDb has much more incentive to review recently

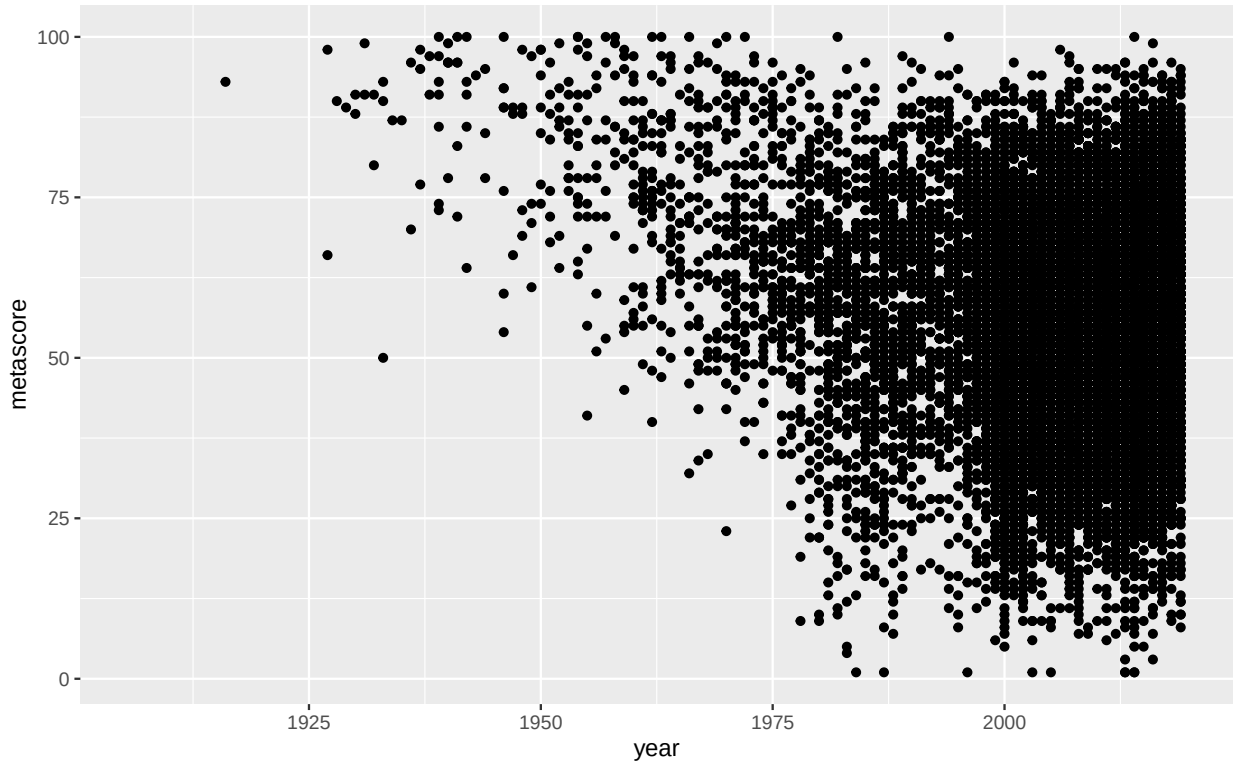

Figure 1: metascore vs time

Table 2: The top 20 movies arranged by metascore

title	year	metascore
The Wizard of Oz	1939	100
Citizen Kane	1941	100
Casablanca	1942	100
Notorious	1946	100
Viaggio in Italia	1954	100
Rear Window	1954	100
Sweet Smell of Success	1957	100
Vertigo	1958	100
Lawrence of Arabia	1962	100
Il gattopardo	1963	100
Au hasard Balthazar	1966	100
Il conformista	1970	100
The Godfather	1972	100
Fanny och Alexander	1982	100
Trois couleurs: Rouge	1994	100
Boyhood	2014	100
City Lights	1931	99
Pinocchio	1940	99
Singin' in the Rain	1952	99
The Night of the Hunter	1955	99

released movies. Hence, when older movies are reviewed, it is likely because it is still immensely popular (i.e. a 'classic').

```
dat2wn <- dat %>%
               dplyr::select(title,
                             year,
                             metascore)


dat3wn <- dat2wn %>%
                arrange(-metascore)


dat3wn %>%
        head(20) %>%
        kable(caption = "The top 20 movies arranged by metascore") %>%
        kable_styling(bootstrap_options = c("striped", "hover"))
```

```
dat3wn %>%
        head(20) %>%
        ggplot(aes(year)) +
```

```
        geom_histogram(binwidth = 5) +
        xlab("Year") +
        ylab("")
```

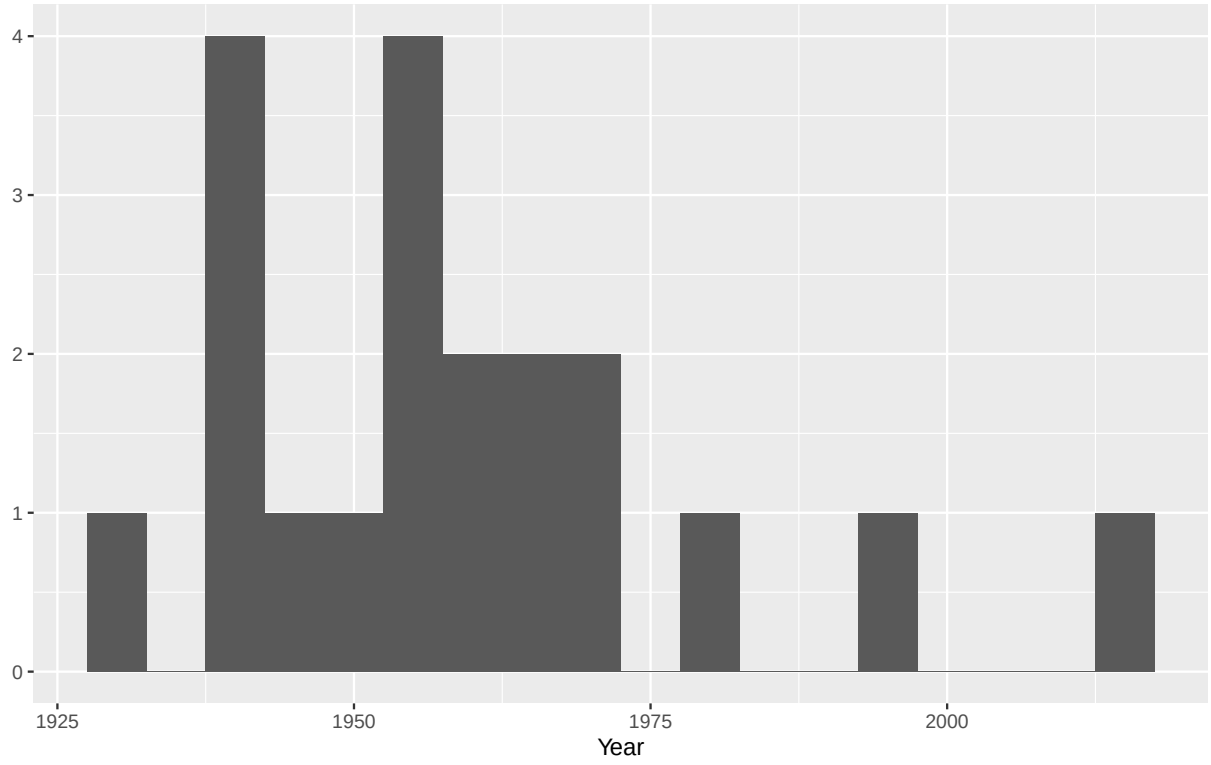

Figure 2: Time distribution of top twenty movies

The aforementioned 'classics bias' is further confirmed by both Table 2 and Figure 2. This list is dominated by the most famous movies from the 1900's such as The Wizard of Oz, Citizen Kane and The Godfather which were all assigned perfect metascores. Moreover, seventeen of the twenty listed movies were released before 1975 and only one was released after 2000.

Assuming that metascore is a reliable representation of a movie's quality, it is easy to see why arguments have been made that movies produced in recent times are not as good. Contrarily, a counterargument can be made that there were just as many bad movies made during previous generations that have been forgotten or overshadowed by the classics. Due to the clear bias within this dataset, it is still unclear which argument is stronger.