Group Assignment Skills: Data Prep.&Workflow Mgt

Team 5

2024-09-06

Team 5

#Sophie van Hest #Eveline Cai #Mette Swanenberg #Tyamo van der Ceelen

Research Motivation

Our research question is: Is an individual's fame related to his/her birth year? By examining movie ratings and number of votes as a proxy for fame, this study seeks to explore whether people born in certain time periods are more likely to achieve fame in the film industry

Data

Data1 includes:

- nconst (string) - alphanumeric unique identifier of the name/person
- primaryName (string)– name by which the person is most often credited
- birthYear – in YYYY format
- deathYear – in YYYY format if applicable, else '\N'
- primaryProfession (array of strings)– the top-3 professions of the person
- knownForTitles (array of tconsts) – titles the person is known for

Data2 includes:

- tconst (string) - alphanumeric unique identifier of the title
- averageRating – weighted average of all the individual user ratings
- numVotes - number of votes the title has received

```
data1 <- read.csv("name.basics.tsv", sep = "\t")
```

```
## Warning in scan(file = file, what = what, sep = sep, quote = quote, dec = dec,
## : EOF within quoted string
```

```
data2 <- read.csv("title.ratings.tsv", sep = "\t")
```

```
str(data1)
```

```
## 'data.frame':    3944029 obs. of  6 variables:
##  $ nconst           : chr  "nm0000001" "nm0000002" "nm0000003" "nm0000004" ...
##  $ primaryName      : chr  "Fred Astaire" "Lauren Bacall" "Brigitte Bardot" "John Belushi" ...
##  $ birthYear        : chr  "1899" "1924" "1934" "1949" ...
##  $ deathYear        : chr  "1987" "2014" "\\N" "1982" ...
##  $ primaryProfession: chr  "actor,miscellaneous,producer" "actress,soundtrack,archive_footage" "actr
##  $ knownForTitles   : chr  "tt0072308,tt0050419,tt0053137,tt0027125" "tt0037382,tt0075213,tt0117057,
```

```
str(data2)
```

```
## 'data.frame':    1472885 obs. of  3 variables:
##  $ tconst       : chr  "tt0000001" "tt0000002" "tt0000003" "tt0000004" ...
##  $ averageRating: num  5.7 5.6 6.5 5.4 6.2 5 5.4 5.4 5.4 6.8 ...
##  $ numVotes     : int  2081 280 2078 181 2816 194 885 2225 212 7671 ...
```

```
summary(data1$birthYear)
```

```
##    Length     Class      Mode
##   3944029 character character
```

```
summary(data2$averageRating)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   6.200   7.200   6.962   7.900  10.000
```

```
summary(data2$numVotes)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       5      11      26    1032     101 2935976
```

Data exploration

```
#Libraries:
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(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.1
```

```r
library(tidyr)
library(tidyverse)
```

```
## Warning: package 'readr' was built under R version 4.4.1
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.5
## v lubridate 1.9.3     v stringr   1.5.1
## v purrr     1.0.2     v tibble    3.2.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Plot of the distribution of birth years

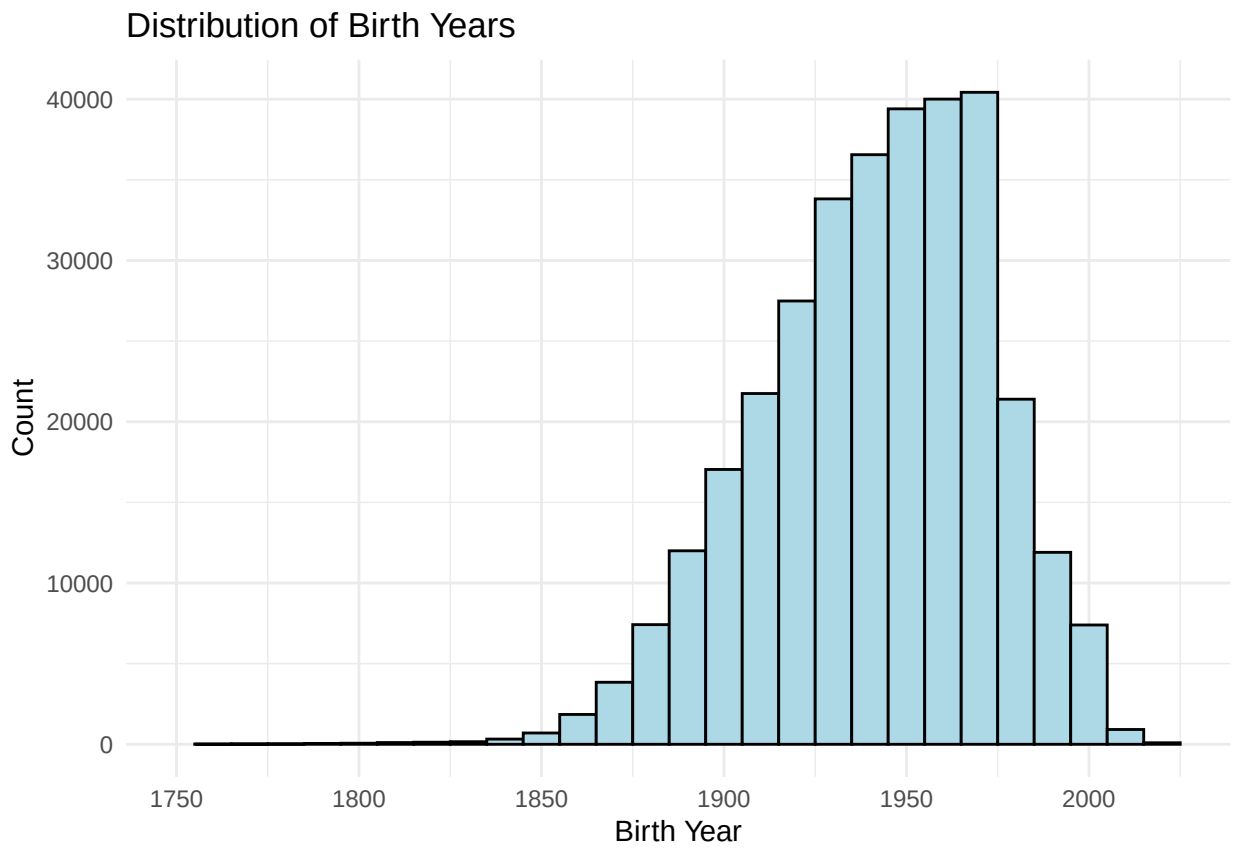
```r
data1$birthYear <- as.numeric(data1$birthYear)
```

```
## Warning: NAs introduced by coercion
```

```r
ggplot(data1, aes(x = birthYear)) +
  geom_histogram(binwidth = 10, fill = "lightblue", color = "black") +
  labs(title = "Distribution of Birth Years", x = "Birth Year", y = "Count") +
  theme_minimal() +
  scale_x_continuous(limits = c(1750, NA))
```

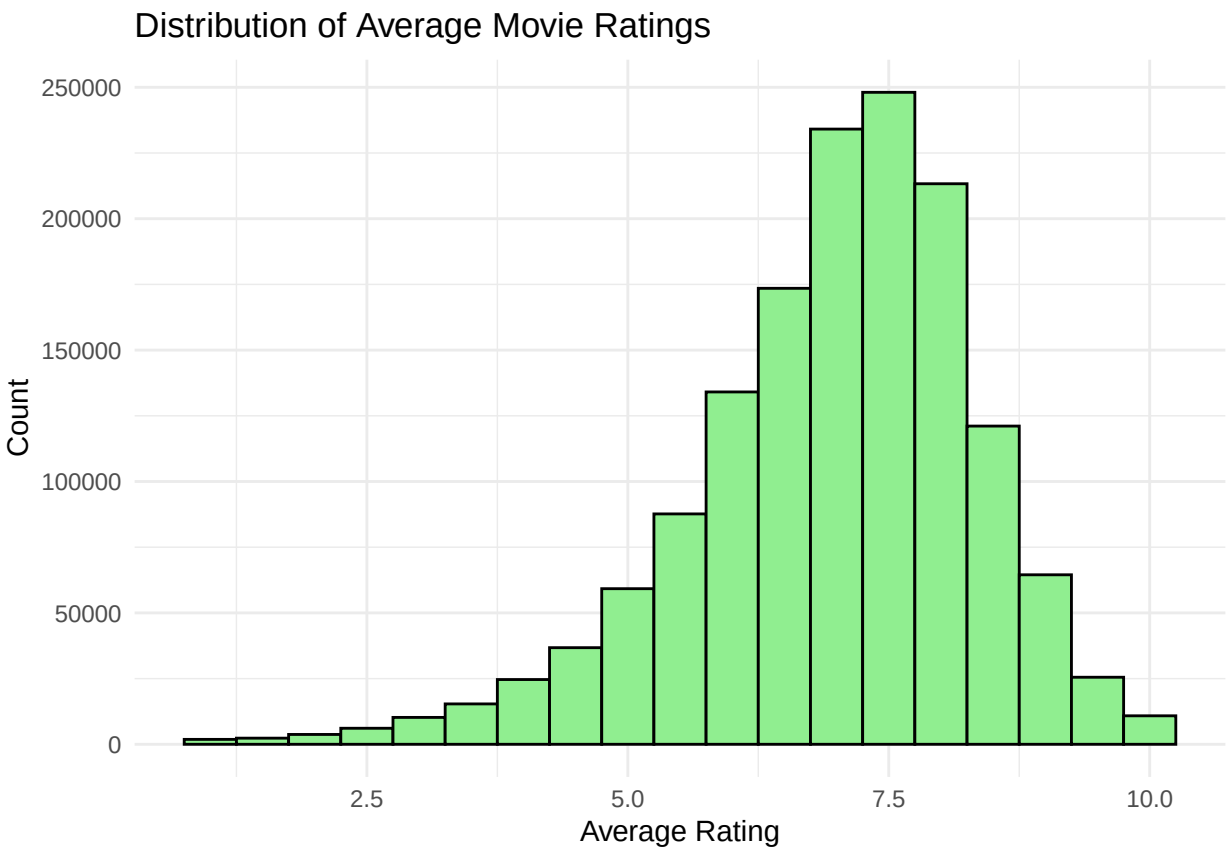
```
## Warning: Removed 3619120 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_bar()`).
```

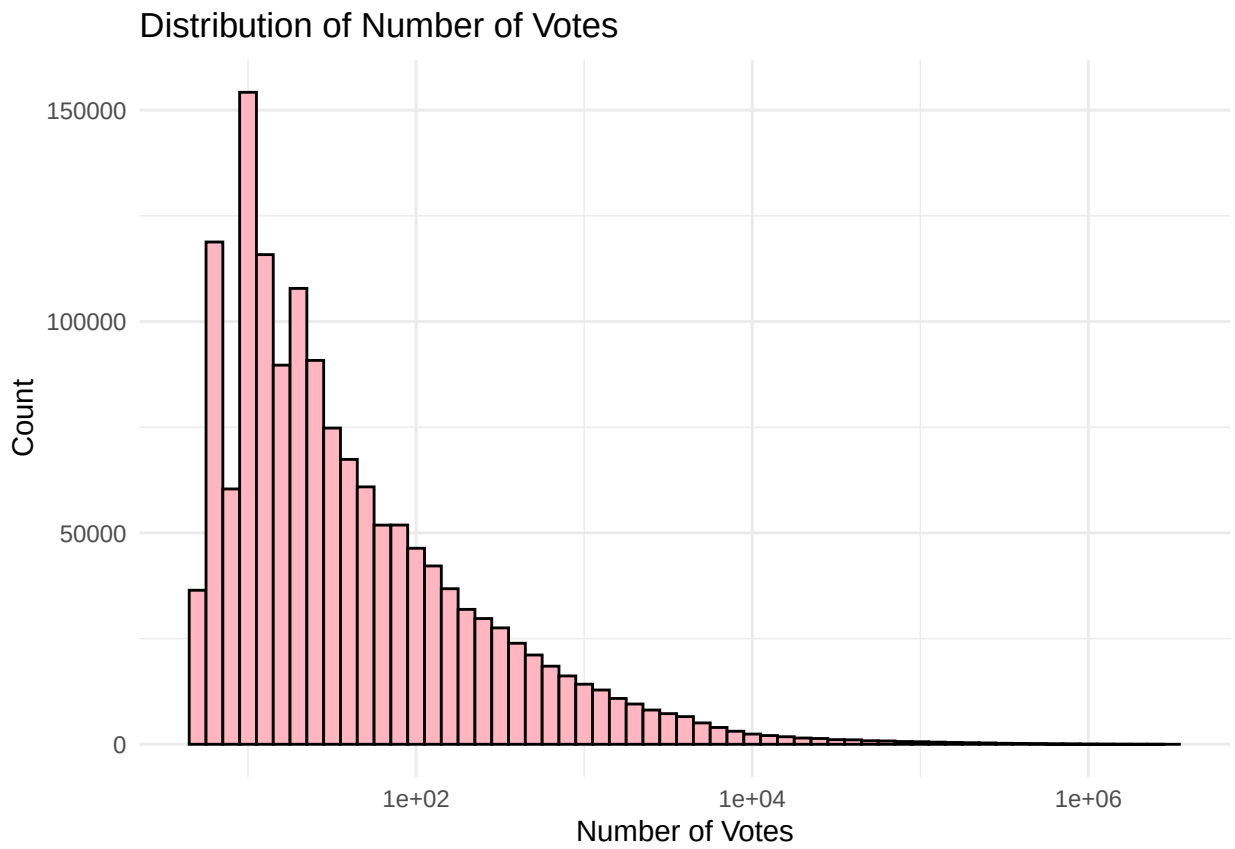


Plots of the distribution of average movie ratings and number of votes

```r
# Plotting distribution of average movie ratings
ggplot(data2, aes(x = averageRating)) +
  geom_histogram(binwidth = 0.5, fill = "lightgreen", color = "black") +
  labs(title = "Distribution of Average Movie Ratings", x = "Average Rating", y = "Count") +
  theme_minimal()
```




```r
# Plotting distribution of number of votes
ggplot(data2, aes(x = numVotes)) +
  geom_histogram(binwidth = 0.1, fill = "lightpink", color = "black") +
  labs(title = "Distribution of Number of Votes", x = "Number of Votes", y = "Count") +
  theme_minimal() +
  scale_x_log10()
```


Distribution of Number of Votes
150000
100000
50000
0
Count
1e+02
1e+04
1e+06
Number of Votes