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

```
##      nconst     primaryName birthYear deathYear
## 1 nm0000001    Fred Astaire      1899      1987
## 2 nm0000002   Lauren Bacall      1924      2014
## 3 nm0000003 Brigitte Bardot      1934       \\N
## 4 nm0000004    John Belushi      1949      1982
## 5 nm0000005  Ingmar Bergman      1918      2007
## 6 nm0000006  Ingrid Bergman      1915      1982
##                    primaryProfession                          knownForTitles
## 1       actor,miscellaneous,producer tt0072308,tt0050419,tt0053137,tt0027125
## 2 actress,soundtrack,archive_footage tt0037382,tt0075213,tt0117057,tt0038355
## 3  actress,music_department,producer tt0057345,tt0049189,tt0056404,tt0054452
## 4      actor,writer,music_department tt0072562,tt0077975,tt0080455,tt0078723
## 5              writer,director,actor tt0050986,tt0083922,tt0050976,tt0069467
## 6        actress,producer,soundtrack tt0034583,tt0038109,tt0036855,tt0038787
```

```
##      tconst averageRating numVotes
## 1 tt0000001           5.7     2086
## 2 tt0000002           5.6      283
## 3 tt0000003           6.5     2090
## 4 tt0000004           5.4      184
## 5 tt0000005           6.2     2824
## 6 tt0000006           5.0      195
```

```
## 'data.frame':    3943633 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,
```

```
## 'data.frame':    1476218 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  2086 283 2090 184 2824 195 888 2231 213 7686 ...
```

```
##    Length     Class      Mode
##   3943633 character character
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       5      11      26    1031     101 2939682
```

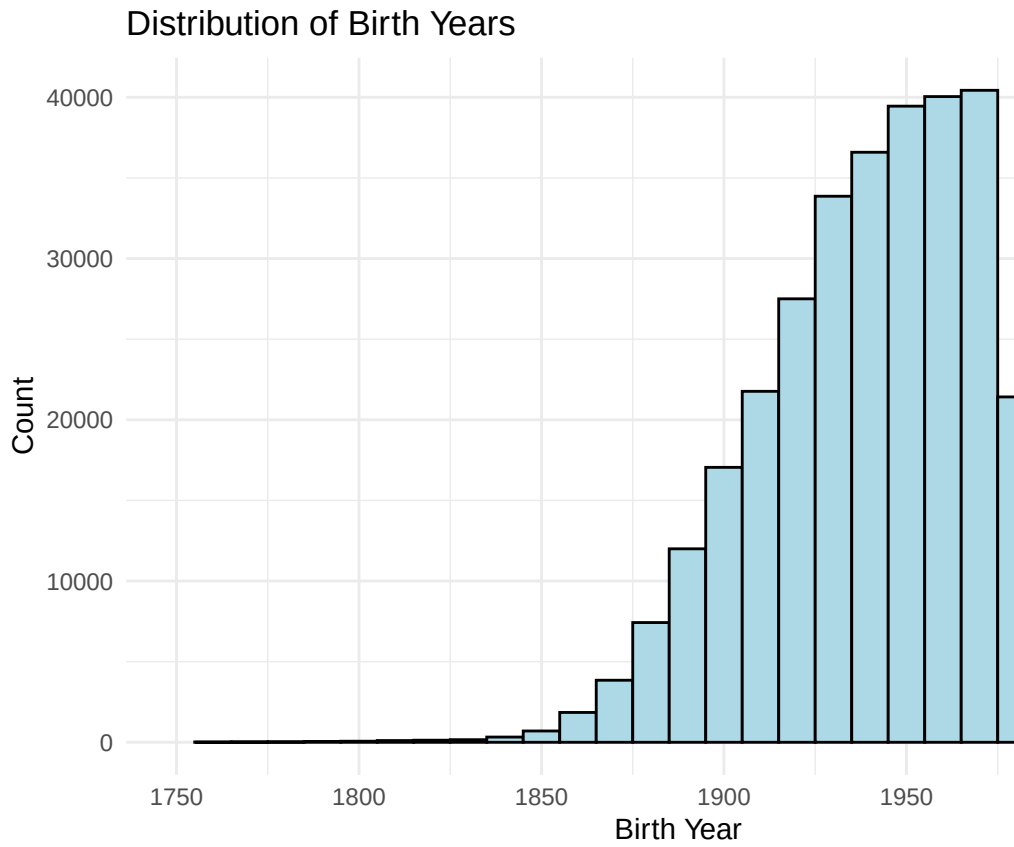
Data exploration

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

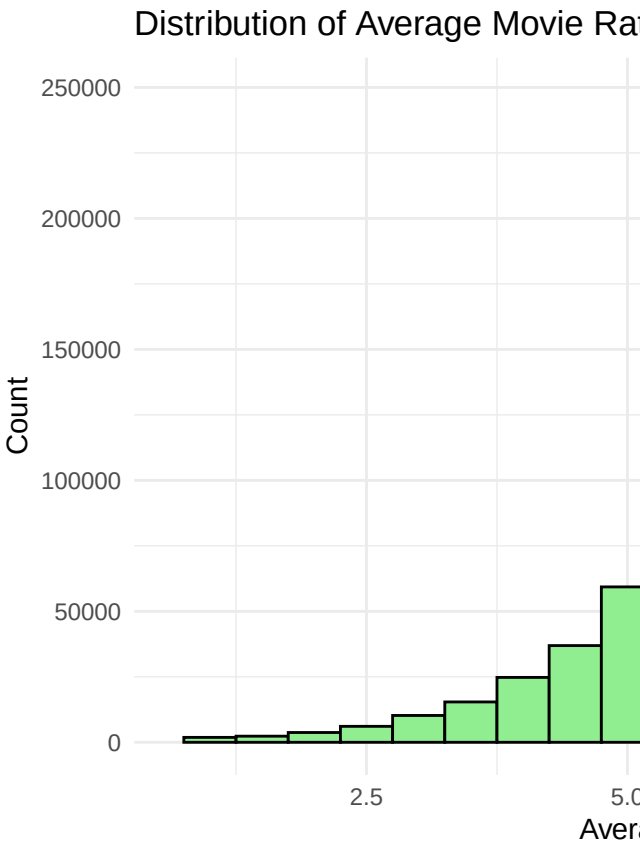
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

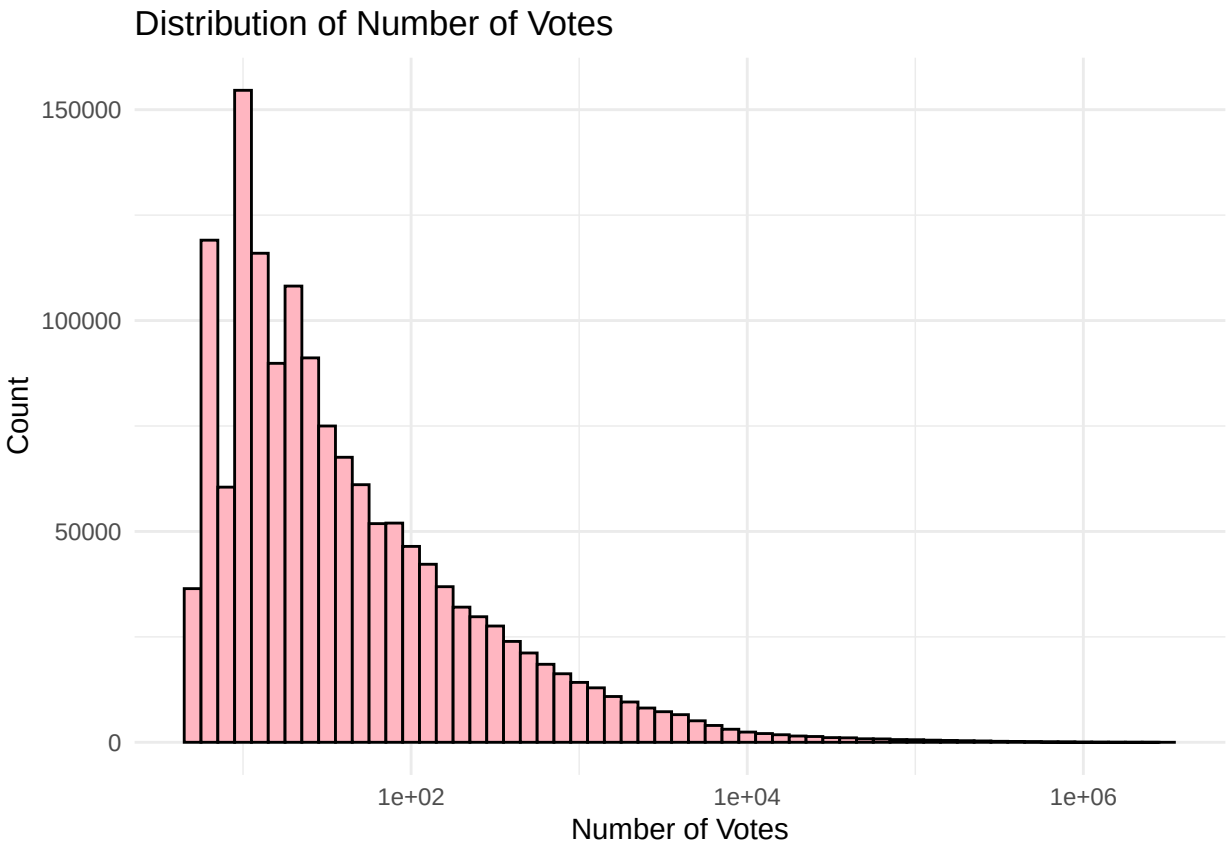
```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.5
## v lubridate 1.9.3     v stringr   1.5.1
## v purrr     1.0.2     v tibble    3.2.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --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



Plot of the distribution of birth years



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


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