tidytuesday_2023_12_12

Noah Lee

2023-12-18

Tidy Tuesday 2023 Decemeber 12

Read in the data

```
holiday_movies <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/
```

```
## Rows: 2265 Columns: 14
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (6): tconst, title_type, primary_title, original_title, genres, simple_t...
## dbl (4): year, runtime_minutes, average_rating, num_votes
## lgl (4): christmas, hanukkah, kwanzaa, holiday
##
## 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.
```

```
holiday_movie_genres <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/m
```

```
## Rows: 4531 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): tconst, genres
##
## 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.
```

Holdiay film releases per year

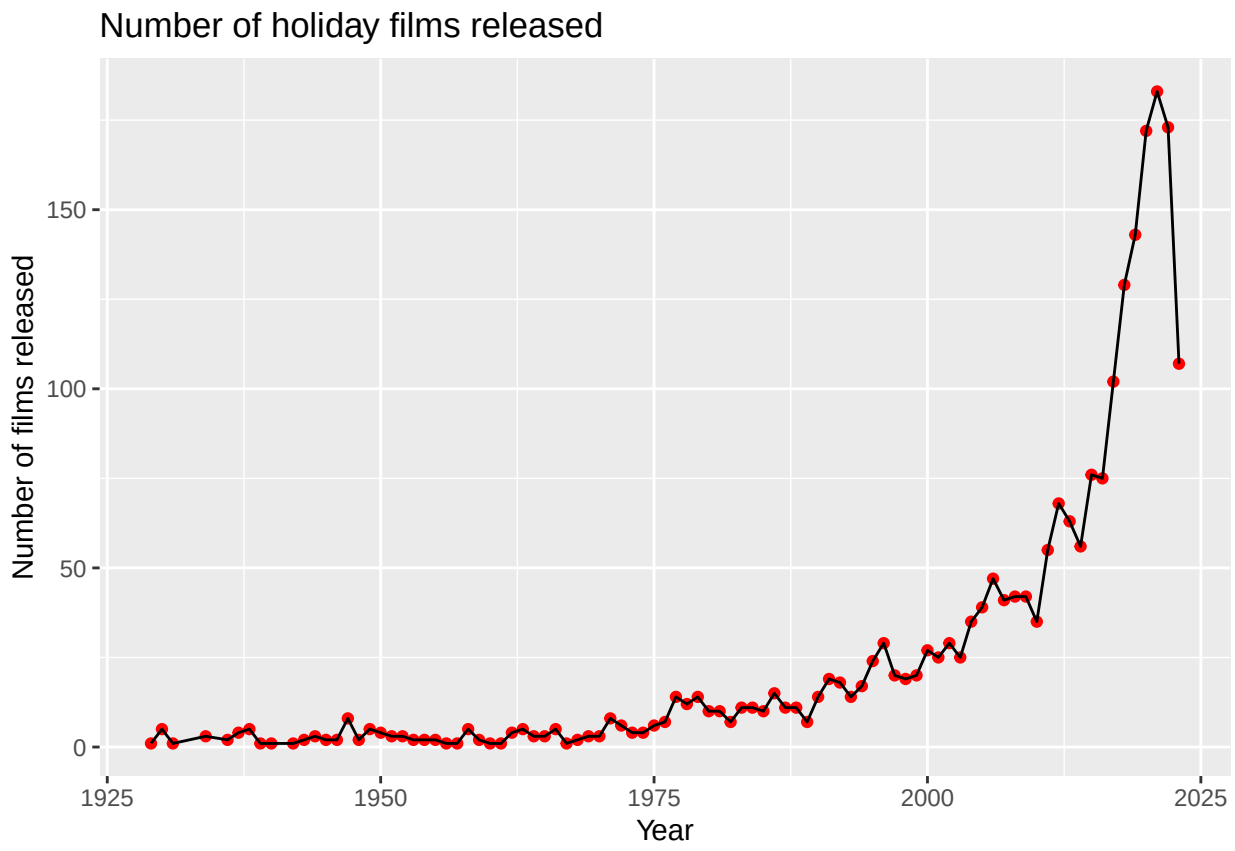
Which years had the most holiday movies?

```
hol_movs_year <- sqldf("SELECT year, COUNT(year) as Num FROM holiday_movies
      GROUP BY year
      ORDER BY Num desc")
head(hol_movs_year,5)
```

```
##   year Num
## 1 2021 183
## 2 2022 173
```

```
## 3 2020 172
## 4 2019 143
## 5 2018 129
```

```
ggplot(data=hol_movs_year, aes(x=year, y=Num)) + geom_point(colour='red') +
  geom_line() + labs(y='Number of films released', x='Year', title='Number of holiday films released')
```



Holiday film releases per decade

Which decade had the most/least number of holiday movies?

```
# "decade_count" takes in an integer that represents a given decade. It returns the number of films rel
decade_count <- function(decade) {
  counting <- 0
  for (i in holiday_movies$year){ # where variable i is the value in the $year column
    if ((i >= decade) & (i< decade + 10)){ # checks if $year i is set in the decade
      counting <- counting + 1
    }
  }
  return (counting)
}
```

```
decades <- c(1930, 1940, 1950, 1960, 1970,
             1980, 1990, 2000, 2010, 2020)
```

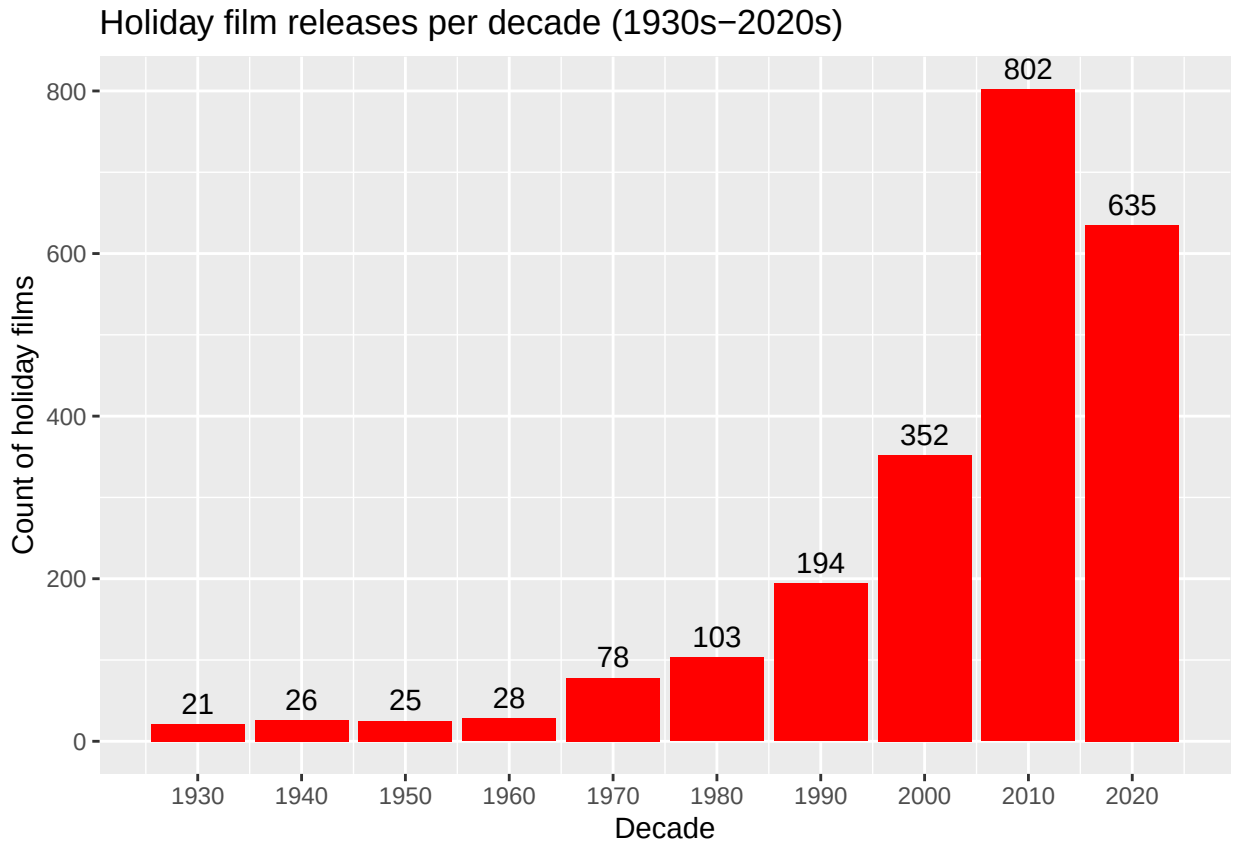
```
decades_rel <- c()

for (dec in decades){
  print(paste("The decade starting with", dec, "saw", decade_count(dec), "holiday films released"))
  decades_rel <- append(decades_rel, decade_count(dec))
}
```

```
## [1] "The decade starting with 1930 saw 21 holiday films released"
## [1] "The decade starting with 1940 saw 26 holiday films released"
## [1] "The decade starting with 1950 saw 25 holiday films released"
## [1] "The decade starting with 1960 saw 28 holiday films released"
## [1] "The decade starting with 1970 saw 78 holiday films released"
## [1] "The decade starting with 1980 saw 103 holiday films released"
## [1] "The decade starting with 1990 saw 194 holiday films released"
## [1] "The decade starting with 2000 saw 352 holiday films released"
## [1] "The decade starting with 2010 saw 802 holiday films released"
## [1] "The decade starting with 2020 saw 635 holiday films released"
```

```
decades_df <- data.frame(decades, decades_rel) #Create a table to do plotting over time

ggplot(data=decades_df, aes(x=decades, y=decades_rel)) + geom_bar(stat='identity', fill="red") +
  geom_text(aes(label=decades_rel), color="black", vjust=-0.5) +
  scale_x_continuous(breaks=seq(1930,2020, by=10)) +
  ylab('Count of holiday films') + xlab('Decade') + ggtitle('Holiday film releases per decade (1930s-20
```

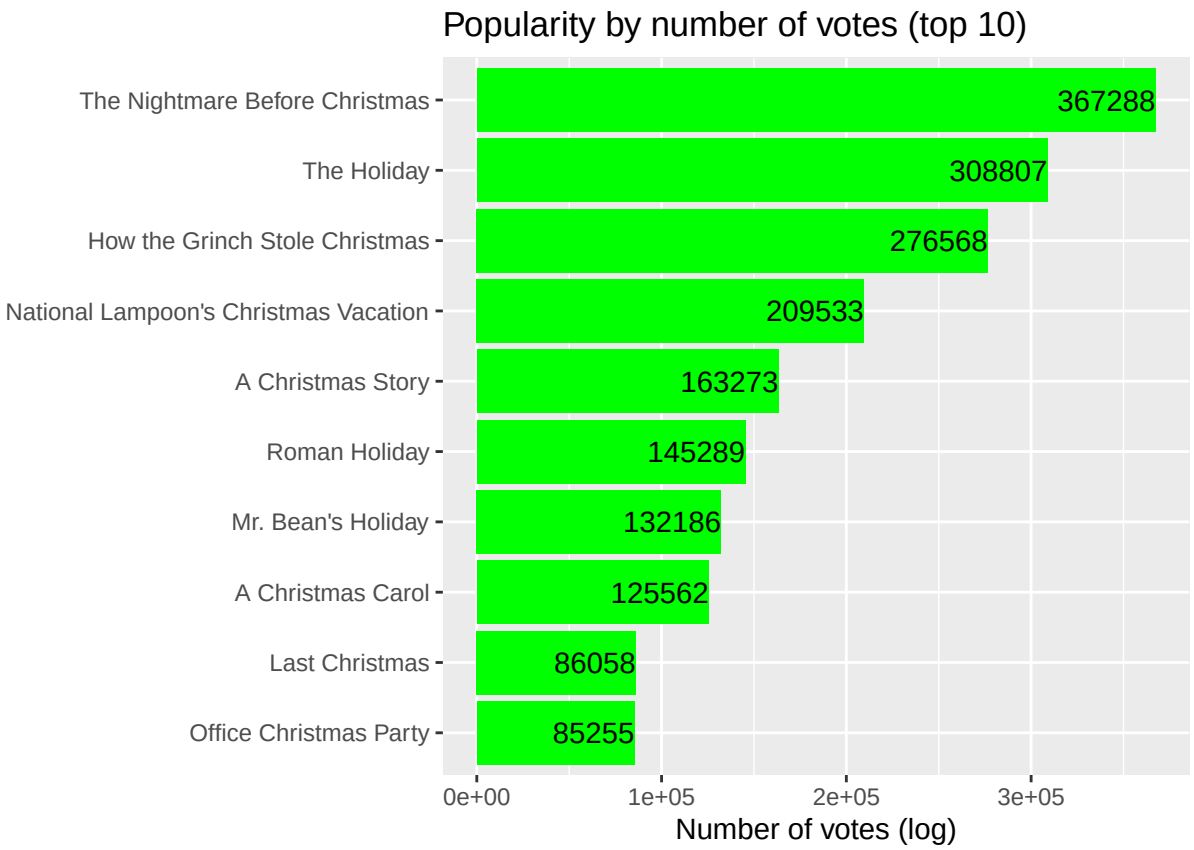


Which holiday movies are most popular?

Answered by number of ratings given

```
ranked_votes <- holiday_movies[order(-holiday_movies$num_votes), ] #sorted dataframe by number of votes

ggplot(data=head(ranked_votes, 10), aes(x=reorder(primary_title,num_votes), y=num_votes)) +
  geom_bar(stat='identity', fill='green' ) + coord_flip() +
  geom_text(aes(label=num_votes), color="black", hjust=1) +
  xlab(' ') + ylab('Number of votes (log)') + ggtitle('Popularity by number of votes (top 10)')
```



Is there a certain genre that gets better ratings?

Use sql joins, then find the average rating per genre.

```
holmov_genres <- sqldf("SELECT holiday_movie_genres.genres, holiday_movie_genres.tconst, average_rating
                       FROM holiday_movies
                       RIGHT JOIN holiday_movie_genres ON holiday_movie_genres.tconst = holiday_movies.
head(holmov_genres, 20)
```

```
##       genres    tconst average_rating
## 1     Comedy tt0020356            5.4
## 2      Drama tt0020823            6.0
## 3    Romance tt0020823            6.0
```

```
## 4     Comedy tt0020985           6.3
## 5      Drama tt0020985           6.3
## 6     Comedy tt0021268           7.4
## 7     Comedy tt0021377           6.1
## 8    Romance tt0021377           6.1
## 9  Adventure tt0021381           6.3
## 10     Crime tt0021381           6.3
## 11   Romance tt0021381           6.3
## 12     Drama tt0023039           6.4
## 13     Crime tt0024869           5.6
## 14     Drama tt0024869           5.6
## 15   Romance tt0024869           5.6
## 16   Western tt0025006           4.8
## 17     Drama tt0025037           6.9
## 18   Fantasy tt0025037           6.9
## 19   Romance tt0025037           6.9
## 20    Comedy tt0027456           5.7
```

```
genre_by_rating <- sqldf("SELECT AVG(average_rating) as average_rating, genres FROM holmov_genres
      GROUP BY genres
      ORDER BY average_rating DESC")
genre_by_rating
```

```
##    average_rating      genres
## 1        9.300000  Reality-TV
## 2        7.192308     History
## 3        7.066667         War
## 4        7.057426 Documentary
## 5        6.900000   Film-Noir
## 6        6.820879       Music
## 7        6.750000   Biography
## 8        6.600000        News
## 9        6.500000       Sport
## 10       6.488542       Short
## 11       6.400373   Animation
## 12       6.256410     Musical
## 13       6.067874       Drama
## 14       6.064356      Family
## 15       6.035414     Romance
## 16       5.933333     Western
## 17       5.921368   Adventure
## 18       5.921081     Fantasy
## 19       5.913756      Comedy
## 20       5.778571      Sci-Fi
## 21       5.754054     Mystery
## 22       5.727273       Crime
## 23       5.690625        <NA>
## 24       5.229032      Action
## 25       5.014286      Horror
## 26       4.953125    Thriller
## 27       4.450000   Talk-Show
```

This looks a bit surprising; why is Reality-TV so high? Why are all these top-10 genres so high? I would expect more comedy-drama-romance based on the many holiday classics there are. I'm thinking of going

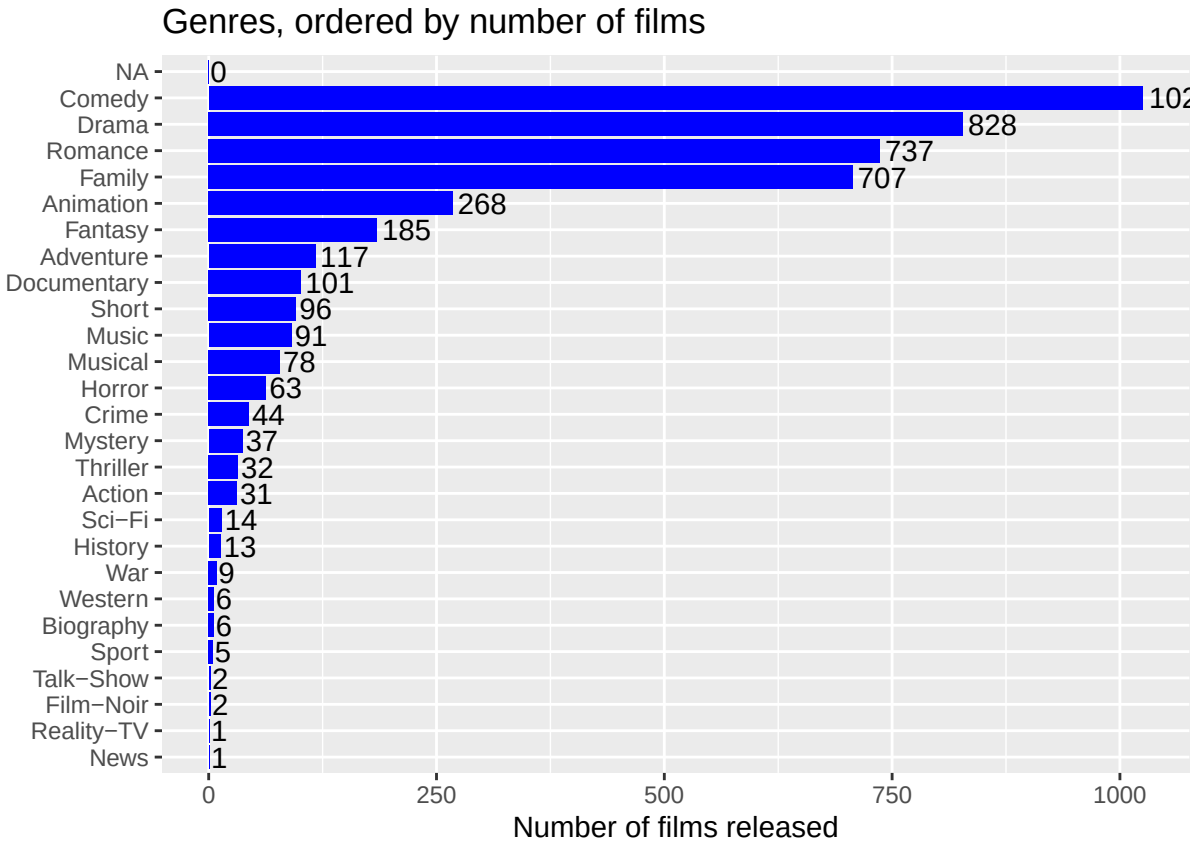
back and adding how many counts per genre - maybe the reasoning of the high genre scores is because of over-saturation of releases?

```
genre_by_rating2 <- sqldf("SELECT AVG(average_rating) as average_rating, COUNT(genres) as genre_count, g
      FROM holmov_genres
      GROUP BY genres
      ORDER BY genre_count DESC")
genre_by_rating2
```

```
##    average_rating genre_count      genres
## 1        5.913756        1025      Comedy
## 2        6.067874         828       Drama
## 3        6.035414         737     Romance
## 4        6.064356         707      Family
## 5        6.400373         268   Animation
## 6        5.921081         185     Fantasy
## 7        5.921368         117   Adventure
## 8        7.057426         101 Documentary
## 9        6.488542          96       Short
## 10       6.820879          91       Music
## 11       6.256410          78     Musical
## 12       5.014286          63      Horror
## 13       5.727273          44       Crime
## 14       5.754054          37     Mystery
## 15       4.953125          32    Thriller
## 16       5.229032          31      Action
## 17       5.778571          14      Sci-Fi
## 18       7.192308          13     History
## 19       7.066667           9         War
## 20       5.933333           6     Western
## 21       6.750000           6   Biography
## 22       6.500000           5       Sport
## 23       4.450000           2   Talk-Show
## 24       6.900000           2   Film-Noir
## 25       9.300000           1  Reality-TV
## 26       6.600000           1        News
## 27       5.690625           0        <NA>
```

This makes more sense, in that more popular genres tend to have lower ratings, compared to less popular genres with higher ratings. Likely to do with over-saturated markets. A visualization.

```
ggplot(data=genre_by_rating2, aes(x=reorder(genres,genre_count), y=genre_count)) +
  geom_bar(stat='identity', fill='blue') +
  coord_flip() +
  geom_text(aes(label=genre_count), color="black", hjust=-0.1) +
  xlab(' ') + ylab('Number of films released') + ggtitle('Genres, ordered by number of films')
```


Genres, ordered by number of films
NA
Comedy
Drama
Romance
Family
Animation
Fantasy
Adventure
Documentary
Short
Music
Musical
Horror
Crime
Mystery
Thriller
Action
Sci−Fi
History
War
Western
Biography
Sport
Talk−Show
Film−Noir
Reality−TV
News
0
102
828
737
707
268
185
117
101
96
91
78
63
44
37
32
31
14
13
9
6
6
5
2
2
1
1
0
250
500
750
1000
Number of films released