# Movies Dataset

*Sociology 312, University of Oregon*
*Prof. Gullickson*

## Overview

The movie data contain information about 2,612 movies produced between 2001 and 2013. The data come from the Open Movie Database, which itself contains data from the Internet Movie Database and Rotten Tomatoes. To simplify our analyses, I have limited the analysis to movies that played in the US and received 10 or more reviews. I have also excluded all shorts, documentaries, and foreign language films.

Table 1 provides summary statistics for all the quantitative variables in the movie data.

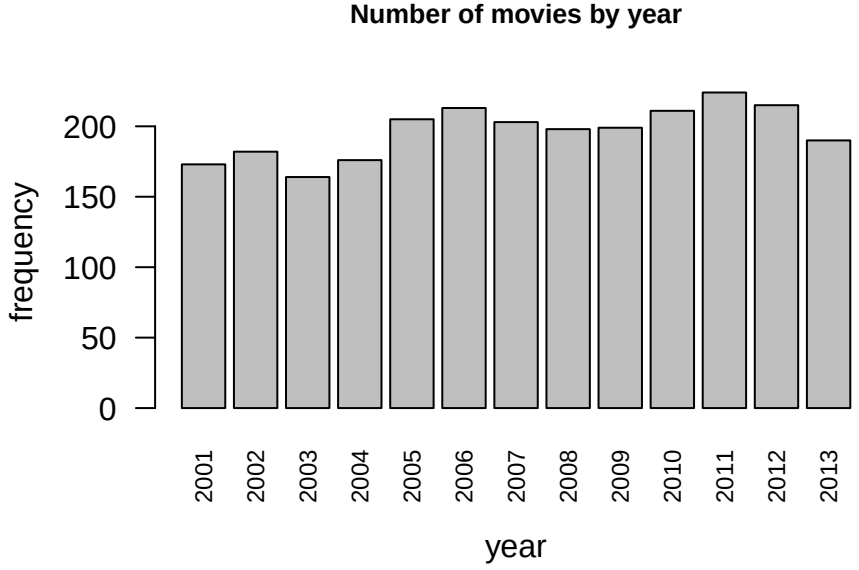
Table 1: Summary statistics for all quantitative variables in the movies dataset

| Statistic | N | Mean | St. Dev. | Min | Pctl(25) | Median | Pctl(75) | Max |
|---|---|---|---|---|---|---|---|---|
| Year | 2553 | 2007.229 | 3.681 | 2001 | 2004 | 2007 | 2010 | 2013 |
| Runtime | 2553 | 105.226 | 16.724 | 45 | 93 | 102 | 114 | 219 |
| Oscars | 2553 | 0.085 | 0.516 | 0 | 0 | 0 | 0 | 11 |
| TomatoRating | 2553 | 5.413 | 1.399 | 1.600 | 4.400 | 5.400 | 6.500 | 9.000 |
| TomatoMeter | 2553 | 47.776 | 26.273 | 0 | 26 | 47 | 71 | 99 |
| BoxOffice | 2553 | 45.156 | 66.517 | 0.0004 | 3.100 | 21.600 | 57.700 | 760.500 |

## Variable Descriptions

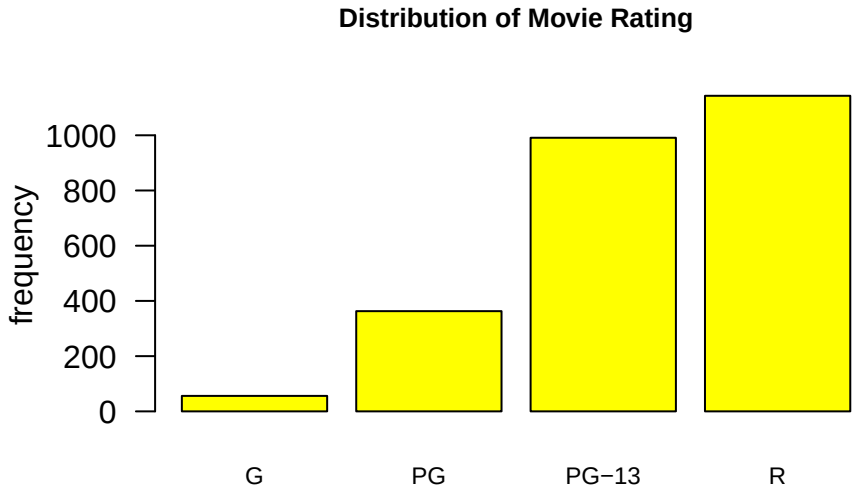
### Year

The calendar year of the film's release. Although this is technically a quantitative variable, I have graphed its distribution below using a barplot, because we often think of such a variable in categorical terms.



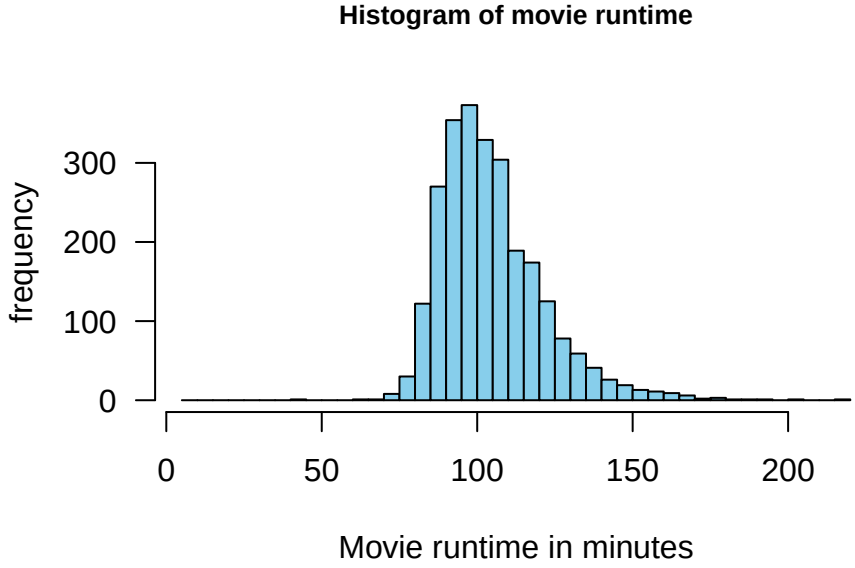
## Rating

The movie's maturity rating. You can see that the R-rated films are the most common and G-rated films are the least common.




## Runtime

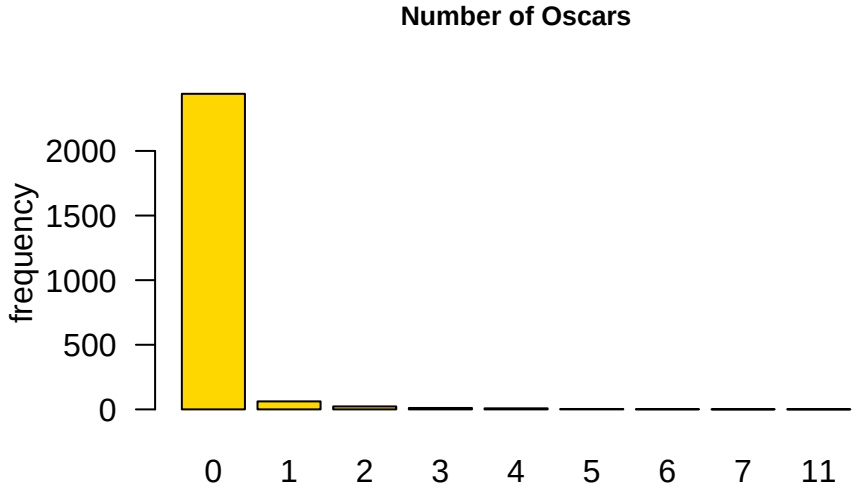
The length of the movie in minutes. The histogram below shows the distribution of this variable with bins of 5 minutes. The distributions is unimodal with slight right skew. The difference between the mean movie length of 105.2 and the median of 102 indicates that this skewness is not severe enough to produce substantial different conclusions about the center.




## Oscars

The number of Oscar awards that this movie received. This includes Oscars that go to individiual actors (leading and supporting), as well as more general awards (best screenplay, editing, cinematography, etc.), and best picture overall.

This is a very heavily right-skewed variable, because the vast majority of movies (95.7%) receive no awards at all.

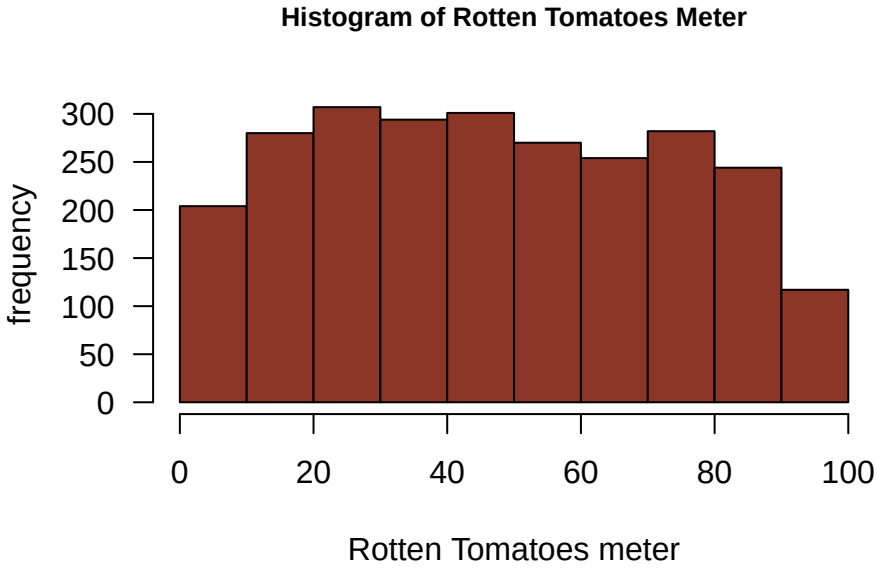



## Rotten Tomatoes Variables

I utilize two different rating metrics employed by Rotten Tomatoes.

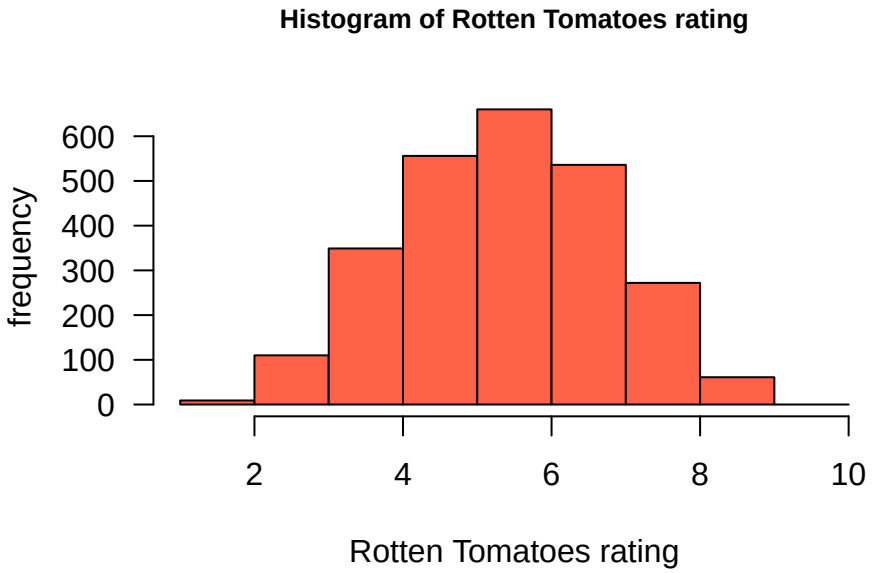
### TomatoMeter

The tomato meter is the percent of reviews that are judged to be positive by Rotten Tomatoes staff. This metric goes from 0 to 100 percent. Note that movies are spread out pretty evenly across the range of this variable, something we call a "uniform" distribution. Note, that this method makes no distinction between how positive a positive review was or how negative a negative review was, so its perfectly possible for two movies with the same Tomato Meter to be viewed very differently by reviewers.



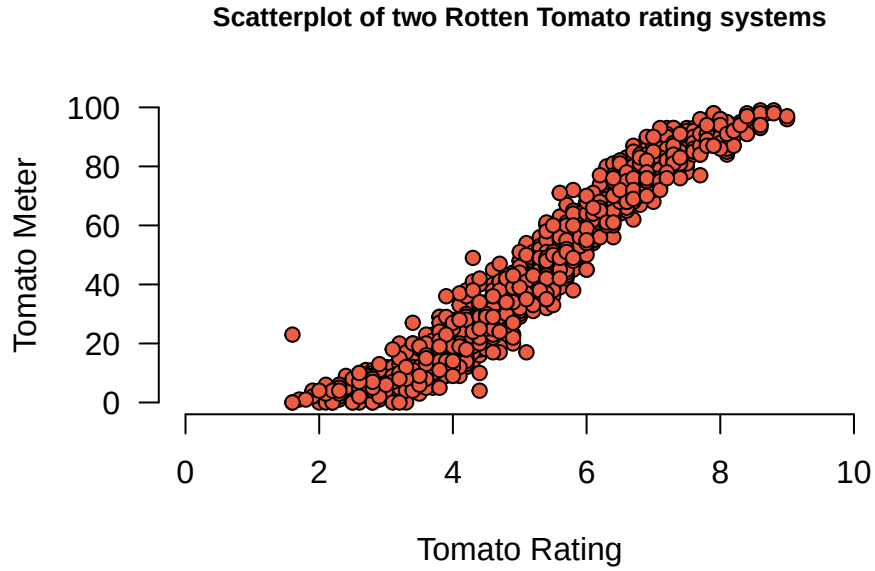
**TomatoRating**

The tomato rating is a combination of all reviews where the review used some kind of numeric rating (e.g. 3 out of 4 stars, 7 out of 10). Rotten Tomatoes "normalizes" these scores so that they are all recorded on the same basis. The scale of this normalized score goes from 1 to 10. Unlike the Tomato Meter, this scale should be capable of distinguishing how strongly positive or negative the review was. The distribution here is more of a bell curve shape.



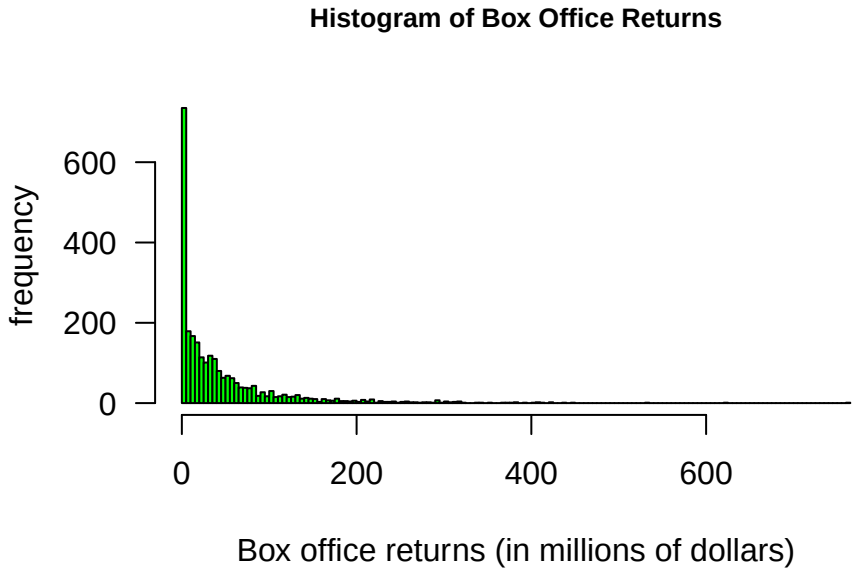

**Relationship between two ratings**

Despite the aforementioned differences in how the rating systems are measured, they do produce very similar results, as the scatterplot below shows. The correlation coefficient between the two scores is 0.974.



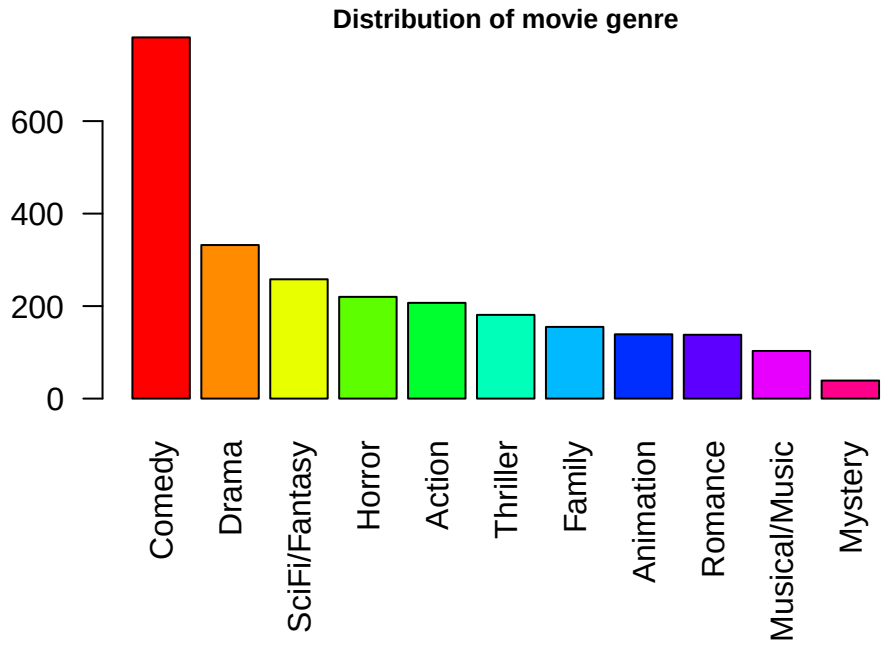
## BoxOffice

This is the total box office returns for the movie, reported in millions of US dollars. The histogram below reveals that this variable is heavily right-skewed with a median of $21.6 million and a mean of $45.2 million. So about half of movies make $21.6 million or less, but a few movies do terrificly well. The highest grossing movie in this time period made $760.5 million (Avatar).




## Genre

This is a nominal variable indicating the genre of the movie. This is a tricky variable to create. In actuality, movies could be listed as multiple genres in the original dataset, with twenty different genres to choose from. For example, "No Country for Old Men" is listed in the genres of crime, drama, and thriller while "Lord of the Rings: Return of the King" is listed as action, adventure, and fantasy. This is probably the best way to treat genres, but for our purposes it adds a lot of complexity. Therefore, I have recoded movies into a single "best" genre based on a decision rule where certain genres trump all others on an ordered basis. For example, comedy trumps romance, so romantic comedies will always show up in this dataset as comedies. The ordering of this system is Animation > Family > Musical > Horror > SciFi/Fantasy > Comedy > Romance > Action > Thriller > Mystery > Drama > All Others. For the most part, this system works well, but you may notice some odd disrepancies for a few movies. For example, The Wolf of Wall Street was originally listed as a crime, comedy, and biography movie, which led to it being classified here as a comedy.


Distribution of movie genre
600
400
200
0
Comedy
Drama
SciFi/Fantasy
Horror
Action
Thriller
Family
Animation
Romance
Musical/Music
Mystery