# Movies dataset

Hadley Wickham

June 5, 2006

## 1 Introduction

This document described a new data set specification designed for experimenting with graphical methods for explorin high dimensional continuous and categorical data (and because I was bored of using olive oils data!). Here I document the data set, the collection process, and give some basic univariate statistics for each variable.

The latest version of the this document, the data set, and links to analyses performed by others can be found at had.co.nz.

## 2 Data collection

The internet movie database, imdb.com, is a website devoted to collecting movie data supplied by studios and fan. It claims to be the biggest movie database on the web and is run by amazon. More about information imdb.com can be found online, including information about the data collection process.

IMDB makes their raw data available. Unfortunately, the data is divided into many text files and the format of each file differs slightly. To create one data file containing all the desired information I wrote a script in the ruby to extract the relevent information and store in a database. This data was then exported into csv for easy import into many programs.

The following text files were downloaded and used:

- business.list. Total budget
- genres.list. Genres that a movie belongs to (eg. comedy and action)
- movies.list. Master list of all movie titles with year of production.
- mpaa-ratings-reasons.list. MPAA ratings.
- ratings.list. IMDB fan ratings.
- running-times.list. Movie length in minutes.

Movies were selected for inclusion if they had a known length and had been rated by at least one IMDB user. The tab delimited file contains the following fields:

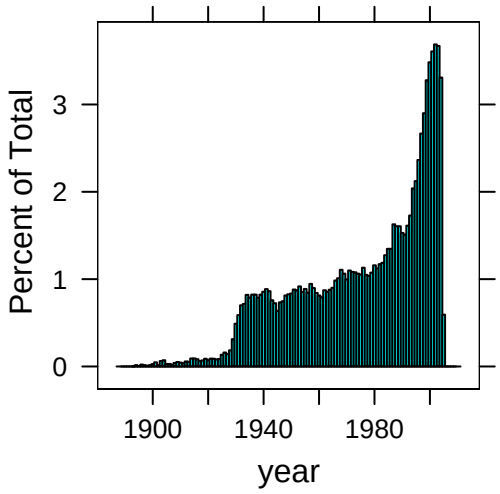
- title. Title of the movie.

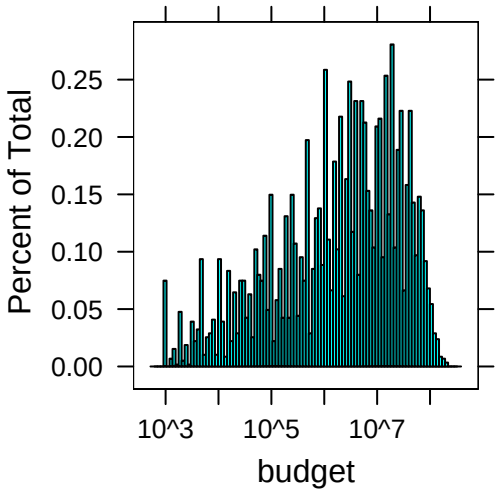
- year. Year of release.
- budget. Total budget (if known) in US dollars
- length. Length in minutes.
- rating. Average IMDB user rating.
- votes. Number of IMDB users who rated this movie.
- r1-10. Distribution of votes for each rating, to mid point of nearest decile: 0 = no votes, 4.5 = 1-9% votes, 14.5 = 11-19% of votes, etc. Due to rounding errors these may not sum to 100.
- mpaa. MPAA rating.
- action, animation, comedy, drama, documentary, romance, short. Binary variables representing if movie was classified as belonging to that genre.

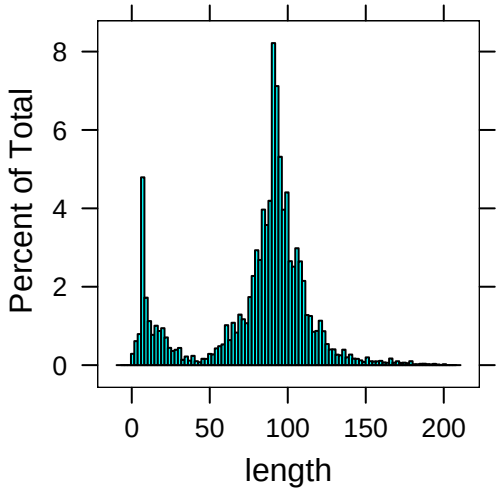
## 3 Data summary

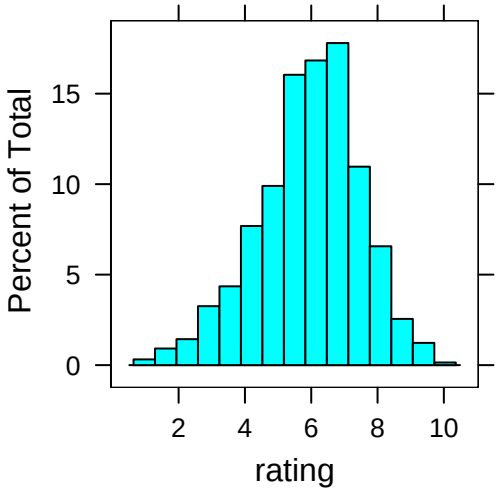
There are a total of 58788 movies from "$" to "xXx: State of the Union".

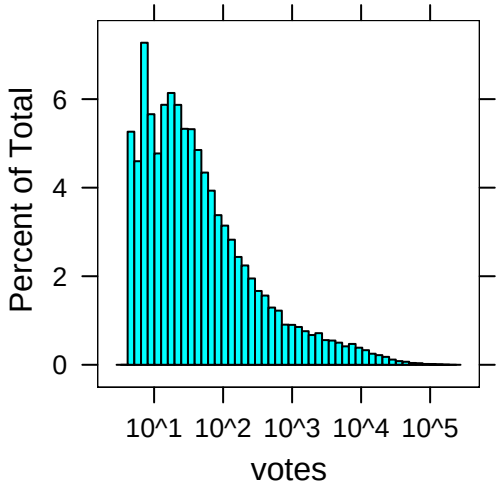
| | Minimum | Maximum | Unique values | Missing values |
|---|---|---|---|---|
| year | 1893 | 2005 | 113 | 0 |
| length | 1 | 5220 | 305 | 0 |
| budget | 0 | 200000000 | 756 | 53573 |
| rating | 1 | 10 | 91 | 0 |
| votes | 5 | 157608 | 4373 | 0 |
| r1 | 0 | 100 | 12 | 0 |
| r2 | 0 | 84 | 10 | 0 |
| r3 | 0 | 84 | 10 | 0 |
| r4 | 0 | 100 | 11 | 0 |
| r5 | 0 | 100 | 11 | 0 |
| r6 | 0 | 84 | 10 | 0 |
| r7 | 0 | 100 | 11 | 0 |
| r8 | 0 | 100 | 11 | 0 |
| r9 | 0 | 100 | 11 | 0 |
| r10 | 0 | 100 | 12 | 0 |

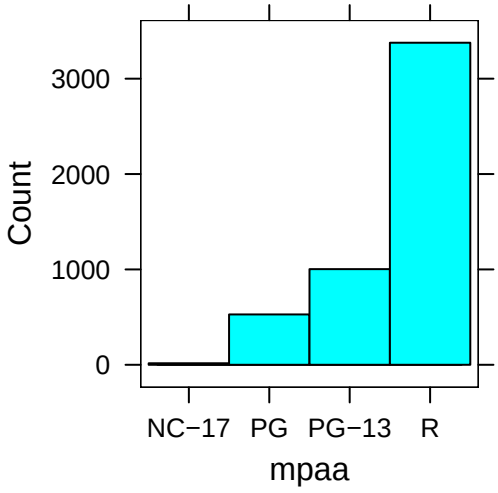

Percent of Total
0
1
2
3
1900
1940
1980
year


Percent of Total
0.00
0.05
0.10
0.15
0.20
0.25
10^3
10^5
10^7
budget


Percent of Total
0
2
4
6
8
0
50
100
150
200
length


Percent of Total
0
5
10
15
2
4
6
8
10
rating


Percent of Total
0
2
4
6
10^1
10^2
10^3
10^4
10^5
votes


Count
0
1000
2000
3000
NC−17
PG
PG−13
R
mpaa