

# Fun with Numbers: R and Perl

Thomas G. Moertel

Moertel Consulting

14 June 2007

R is a system for statistical computing and graphics

- Licensed under the GPL
- Design inspired by S and Scheme
- Runs on just about every platform
- Popular research vehicle
- Tons of add-ons via CRAN
- Publication-quality graphics

# The R language

- At the core of R is the *R language*
- It's a serious programming language
- Infix syntax
- Scheme-like underpinnings
- Tailored for statistics

An interesting question ...

*How can I tell  
if a movie  
is worth seeing?*

# Use the Internet!



Earth's Biggest Movie Database™

[NOW PLAYING](#) [MOVIE / TV NEWS](#) [MY MOVIES](#) [NEW ON DVD](#) [IMDb TV](#)

[Home](#) | [Top Movies](#) | [Photos](#) | [Independent F](#)  
[Browse](#) | [Help](#)

search

[IMDb](#) > [Ocean's Thirteen \(2007\)](#)



**Ocean's Thirteen (2007)**

[photos](#) [board](#) [trailer](#) [IMDb PRO details](#)

[Register](#) or [login](#) to rate this title

★★★★★☆☆☆☆

User Rating: 7.5/10 ([7,161 votes](#))

[more](#)

**Photo Gallery** ([see all 148 photos](#))



[more](#)

[+](#) add to My Movies

Quicklinks

Top Links

# But there's a problem

IMDb  
Earth's Biggest Movie Database™

NOW PLAYING MOVIE / TV NEWS MY MOVIES NEW ON DVD IMDb TV

Home | Top Movies | Photos | Independent Film | Browse | Help

search All

IMDb > Ocean's Thirteen (2007)

**Ocean's Thirteen (2007)**

photos board trailer **IMDb PRO** details

Register or login to rate this title

★★★★★☆☆☆☆☆  
User Rating: 7.5/10 (7,161 votes)

more

Photo Gallery (see all 148 photos)

add to My Movies

Quicklinks  
main details

Top Links

← 7.5 *stars* ?

- What does 7.5 stars *mean*?
- Is a star a unit of “goodness”?
- If so, how much is a star worth?

# Goal: Unlock the meaning of star ratings

If we know what stars mean,  
we can interpret online movie ratings.

**We will rule the world!**

- 1 Get movie data from IMDB
- 2 Apply Perl to whip it into shape
- 3 Load data into R
- 4 Extract meaning
- 5 Profit!

# Step 1: Get movie data from IMDB

```
host=ftp://ftp.funet.fi  
file=/pub/mirrors/ftp.imdb.com/pub/ratings.list.gz  
wget -qN $host$file
```

# The data

CRC: 0xDEDD7812 File: ratings.list Date: Fri Jun 8 01:00:00 2007

From: vote@imdb.com (Movie Ratings Report)  
Subject: Movie Ratings Report

For information on the list format, and on submitting your ratings,  
see the end of this posting.

Details on obtaining this posting via FTP are also given below.

Comments, suggestions etc. via e-mail to <help@imdb.com>

TOP 250 MOVIES (1300+ VOTES)

The formula used to calculate the top 250 movies is:

$$\text{weighted rank} = (v/(v+k))*X + (k/(v+k))*C$$

where:

X = average for the movie (mean)

v = number of votes for the movie

k = minimum votes required to be listed in the top 250 (currently 1300)

C = the mean vote across the whole report (currently 6.90)

note: for this top 250, only votes from regular voters are considered.

New	Distribution	Votes	Rank	Title
	0000000015	217316	9.1	Godfather, The (1972)
	0000000115	259045	9.2	Shawshank Redemption, The (1994)
	0000000125	123428	9.0	Godfather: Part II, The (1974)
	0000000124	65394	8.9	Buono, il brutto, il cattivo, Il (1966)
	...	...	...	...
	...	...	...	...

# The data (2)

...  
...

## BOTTOM 10 MOVIES (650+ VOTES)

New	Distribution	Votes	Rank	Title
	8000000001	14276	1.9	Manos: The Hands of Fate (1966)
	5100000001	882	1.8	Snowboard Academy (1996)
	6000000001	3676	1.8	Going Overboard (1989)
	7000000000	1033	1.8	Keloglan kara prens'e karsi (2006)
	7000000001	5425	1.8	SuperBabies: Baby Geniuses 2 (2004)
	7000000001	12659	1.8	From Justin to Kelly (2003)
	6100000001	1601	1.6	Hillz, The (2004)
	6000000002	1307	1.6	Dunyayi kurtaran adam'in oglu (2006)
	7000000001	1904	1.5	Anne B. Real (2003)
	7000000001	4832	1.5	Crossover (2006)

## MOVIE RATINGS REPORT

New	Distribution	Votes	Rank	Title
	0000000016	293	8.3	!Huff (2004) (TV)
	100.000103	22	5.0	"#1 Single" (2006)
	2....4.11	7	4.0	"\$1.98 Beauty Show, The" (1978)
	0..0101102	21	6.9	"\$10,000 Pyramid, The" (1973)
	6..22.....	5	3.0	"\$25 Million Dollar Hoax" (2004)
	2..0.02003	28	6.3	"\$40 a Day" (2002)
	...	...	...	...
	...	...	...	...

## Step 2: Apply Perl to whip it into shape

```
#!/usr/bin/perl -lan

BEGIN {
    print 'Title|Histogram|VoteCount|VoteMean|Year'
}

sub in_the_good_stuff {
    (/^MOVIE RATINGS REPORT/ .. /-----/) &&
    $F[2] =~ /\d/
}

if (in_the_good_stuff()) {
    my $year = /\((\d{4})\D/ ? $1 : "NA";
    print join "|", (split " ", $_, 4)[3,0,1,2], $year;
}
```

# Shape-whipped data

```
Title|Histogram|VoteCount|VoteMean|Year
!Huff (2004) (TV)|0000000016|293|8.3|2004
"#1 Single" (2006)|100.000103|22|5.0|2006
"$1.98 Beauty Show, The" (1978)|2....4.11|7|4.0|1978
"$10,000 Pyramid, The" (1973)|0..0101102|21|6.9|1973
"$25 Million Dollar Hoax" (2004)|6..22....|5|3.0|2004
"$40 a Day" (2002)|2..0.02003|28|6.3|2002
"$treet, The" (2000)|0000000111|41|5.1|2000
"'Allo 'Allo!" (1982)|0000000125|761|8.6|1982
...
...
```

## Step 3: Load data into R

```
## load IMDB ratings data into R

ratings <- read.delim("ratings.psv",
                     header=T,
                     sep="|",
                     comment.char="",
                     as.is=c(T,T,F,F,F))

## we're interested only in movies with lots of ratings

ratings <- subset(ratings, VoteCount > 100)
```

# What's loaded?

```
> summary(ratings)
```

Title	Histogram	VoteCount
Length:34958	Length:34958	Min. : 101
Class :character	Class :character	1st Qu.: 149
Mode :character	Mode :character	Median : 268
		Mean : 2075
		3rd Qu.: 764
		Max. :259045

VoteMean	Year
Min. : 1.000	Min. :1888
1st Qu.: 5.400	1st Qu.:1976
Median : 6.400	Median :1994
Mean : 6.205	Mean :1986
3rd Qu.: 7.300	3rd Qu.:2001
Max. :10.000	Max. :2007

# Step 4: Extract meaning

```
## prepare my prefs for making pretty graphs

require("lattice")

trellis.device(dev=x11, theme="col.whitebg")
toms.bar.fill <- trellis.par.get("bar.fill")
toms.bar.fill$col <- toms.bar.fill$border <- "darkgreen"

tom.plot <- function(f, ...) {
  f(...,
    par.settings = list(bar.fill=toms.bar.fill),
    par.strip.text = list(cex=0.8))
}

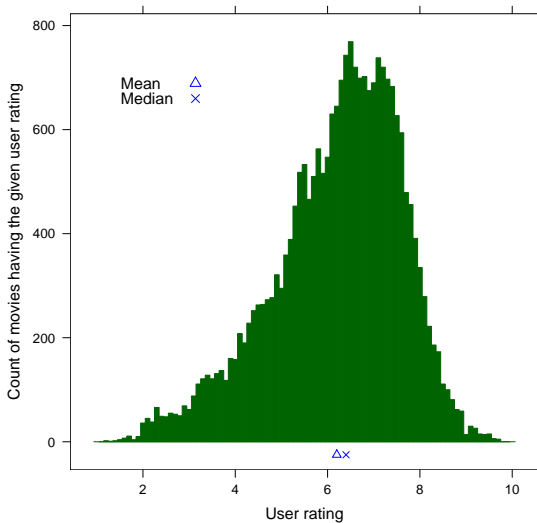
tom.hist <- function(...) {
  tom.plot(histogram, breaks=seq(1,10,0.1), plot.points=F,
  ...)
}
```

# Plot histogram

```
rating.hist <-  
  tom.hist(  
    ~ VoteMean, data=ratings,  
    type="count",  
    main="Distribution of user ratings for IMDB movies",  
    xlab="User rating",  
    ylab="Count of movies having the given user rating",  
    # a few more options ...  
  )  
  
print(rating.hist)
```

# Some meaning

Distribution of user ratings for IMDB movies



# Extracting more meaning

```
## create decoder ring for movie ratings

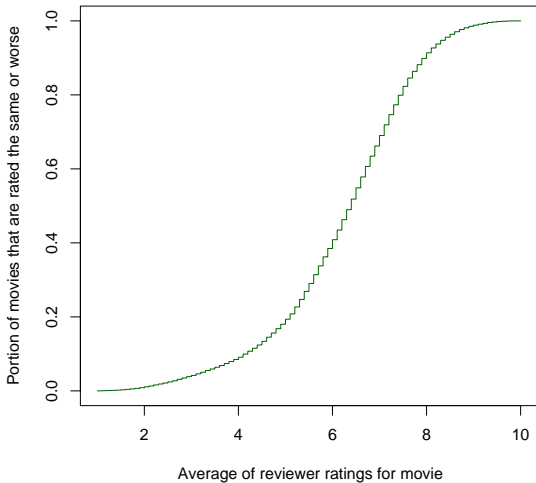
attach(ratings)

F <- ecdf(VoteMean)
x <- seq(1,10,0.1)

plot(x, F(x), type="s",
     main="Cumulative distribution of mean IMDB ...",
     xlab="Average of reviewer ratings for movie",
     ylab="Portion of movies that are rated the ...",
     col="darkgreen")
```

# More meaning

Cumulative distribution of mean IMDB user ratings



# Let's go back to that 7.5 stars

```
> F(7.5)
[1] 0.823102
```

- $F(x)$  places a star rating  $x$  into a **quantile**
- So 7.5 stars means a movie is “better” than about 82 percent of movies ...

... for some definition of *better*

- *Better* means **more highly rated on IMDB.**
- Which isn't the same as **you will enjoy this movie more.**
- But, in practice, “betterness” is a fairly good predictor of enjoyment potential ...

... for some definition of *better*

- *Better* means **more highly rated on IMDB.**
- Which isn't the same as **you will enjoy this movie more.**
- But, in practice, “betterness” is a fairly good predictor of enjoyment potential ...
- (Unless your tastes are weird.)

# Why quantiles work in practice

- Your intuition about **stars** is **weak**:  
*This movie has 7.5 stars.*
- Your intuition about **rankings** is **strong**:  
*This movie is among the best fifth of movies.*
- Quantiles let you convert stars into rankings:

```
> F(7.5)  
[1] 0.823102
```

# Repeat the process for *all* star ratings

- 1 For each star rating  $x$ , compute  $F(x)$   
→ Use R
- 2 Make a table of the results  
→ Use Perl

# Step 5: Profit!

rating quantile

4.00	9
5.00	19
5.25	23
5.50	29
5.75	34
6.00	41
6.25	47
6.50	55
6.75	61
7.00	69
7.25	75
7.50	82
7.75	86
8.00	91
8.25	94
8.50	96
8.75	98
9.00	99

# Apply more R and Perl for ...

?

# More profit!

< GRAND UNIFIED DECODER RING FOR INTERNET MOVIE DATABASE USER RATINGS >

Movie's Genre	Movie's User Rating on Internet Movie Database																	
	4.			5.			6.			7.			8.			9.		
	00	00	25	50	75	00	25	50	75	00	25	50	75	00	25	50	75	00
All genres	11	21	23	29	33	39	44	51	57	65	71	79	83	88	91	95	96	98
Action	17	30	34	39	43	48	54	60	65	72	76	81	84	88	89	93	95	97
Adventure	10	21	24	30	33	38	42	48	54	62	69	77	81	86	88	91	94	96
Animation	5	14	15	22	27	34	38	46	52	61	65	75	81	91	92	95	97	98
Comedy	9	20	24	31	36	43	50	59	66	75	81	88	91	95	96	97	98	99
Documentary	3	7	8	11	12	16	18	23	25	38	47	60	70	85	88	93	95	98
Drama	5	10	13	19	22	29	35	44	52	63	70	80	86	92	95	97	98	99
Family	13	24	29	42	46	54	59	68	75	82	84	91	93	97	97	97	97	97
Fantasy	13	23	26	31	33	42	46	54	59	69	74	81	82	87	89	92	94	95
Horror	35	59	63	70	72	78	80	84	88	91	93	96	97	98	98	98	99	99
Musical	8	12	17	23	29	31	35	42	54	67	77	81	85	92	94	96	98	98
Mystery	8	19	23	31	36	44	50	59	68	76	78	82	87	92	93	96	97	97
Romance	4	12	16	21	26	32	40	52	60	73	80	87	90	95	97	98	98	99
Sci-Fi	25	41	45	48	49	56	58	62	67	73	77	81	83	88	90	95	97	98
Short	4	7	10	12	12	13	15	24	27	34	43	64	73	87	91	94	96	99
Thriller	14	32	37	43	47	55	61	69	74	80	84	90	93	96	97	98	99	99
War	7	11	13	17	20	22	26	34	37	47	54	66	74	83	86	91	96	96
Western	12	12	12	19	31	38	38	38	44	50	56	69	75	88	94	94	94	++

P E R C E N T I L E R A N K

Thank you!

*You rock!*

# For more information

R Project for Statistical Computing  
<http://www.r-project.org/>

Thomas G. Moertel  
Moertel Consulting  
[tgm@moertel.com](mailto:tgm@moertel.com)  
<http://blog.moertel.com/>