607 Week 4 Assignment

Nathan Groom

September 22, 2014

First I used the tbl\_df() function on "movies." Then I had to decide what criteria I'd use to determine what a "best popular" movie is. I decided on rating and number of votes because these are the attributes that determine (in my eyes) if a movie is popular (votes) and if it's the one of the "best" (rating) I'm choosing 112 votes because that's the 3rd quartile cutoff of the "votes" metric. a few of the movies have thousands of votes which really skews the mean, so I'm going "above the 3rd quartile" as the cutoff of what is popular. As for rating, I chose 8.0, which is a bit more subjective. The median of the rating metric is 5.9, the median is 6.1 and the 3rd quartile is 7.0. However, in my personal experience of looking at IMDB, I can't consider a movie to be considered one of the "best" if it has less than an 8.0 rating. In my opinion 8.0 is the cutoff between a very good movie and an excellent movie. It's an unscientific opinion but the ratings on IMDB are unscientific too.

(require(dplyr))

## Loading required package: dplyr  
##   
## 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

## [1] TRUE

(require(ggplot2))

## Loading required package: ggplot2

## [1] TRUE

(movies\_df<-tbl\_df(movies))

## Source: local data frame [58,788 x 24]  
##   
## title year length budget rating votes r1 r2 r3  
## 1 $ 1971 121 NA 6.4 348 4.5 4.5 4.5  
## 2 $1000 a Touchdown 1939 71 NA 6.0 20 0.0 14.5 4.5  
## 3 $21 a Day Once a Month 1941 7 NA 8.2 5 0.0 0.0 0.0  
## 4 $40,000 1996 70 NA 8.2 6 14.5 0.0 0.0  
## 5 $50,000 Climax Show, The 1975 71 NA 3.4 17 24.5 4.5 0.0  
## 6 $pent 2000 91 NA 4.3 45 4.5 4.5 4.5  
## 7 $windle 2002 93 NA 5.3 200 4.5 0.0 4.5  
## 8 '15' 2002 25 NA 6.7 24 4.5 4.5 4.5  
## 9 '38 1987 97 NA 6.6 18 4.5 4.5 4.5  
## 10 '49-'17 1917 61 NA 6.0 51 4.5 0.0 4.5  
## .. ... ... ... ... ... ... ... ... ...  
## Variables not shown: r4 (dbl), r5 (dbl), r6 (dbl), r7 (dbl), r8 (dbl), r9  
## (dbl), r10 (dbl), mpaa (fctr), Action (int), Animation (int), Comedy  
## (int), Drama (int), Documentary (int), Romance (int), Short (int)

(movies10<-filter(movies\_df, rating>=8.0, votes>=112))

## Source: local data frame [690 x 24]  
##   
## title year length budget rating votes r1  
## 1 12 Angry Men 1957 96 340000 8.7 29278 4.5  
## 2 12 stulev 1971 161 NA 8.9 252 4.5  
## 3 2001: A Space Odyssey 1968 156 10500000 8.3 64982 4.5  
## 4 21 Grams 2003 124 20000000 8.0 21857 4.5  
## 5 39 Steps, The 1935 86 NA 8.0 7931 4.5  
## 6 A zori zdes tikhie 1972 188 NA 8.6 137 0.0  
## 7 AC/DC: Let There Be Rock 1980 95 NA 8.5 159 4.5  
## 8 Abenteuer des Prinzen Achmed, Die 1926 66 NA 8.4 169 4.5  
## 9 Adventures of Robin Hood, The 1938 102 1900000 8.2 7359 4.5  
## 10 African Queen, The 1951 105 NA 8.1 13765 4.5  
## .. ... ... ... ... ... ... ...  
## Variables not shown: r2 (dbl), r3 (dbl), r4 (dbl), r5 (dbl), r6 (dbl), r7  
## (dbl), r8 (dbl), r9 (dbl), r10 (dbl), mpaa (fctr), Action (int),  
## Animation (int), Comedy (int), Drama (int), Documentary (int), Romance  
## (int), Short (int)

(movies11<-select(movies10, year, rating, votes))

## Source: local data frame [690 x 3]  
##   
## year rating votes  
## 1 1957 8.7 29278  
## 2 1971 8.9 252  
## 3 1968 8.3 64982  
## 4 2003 8.0 21857  
## 5 1935 8.0 7931  
## 6 1972 8.6 137  
## 7 1980 8.5 159  
## 8 1926 8.4 169  
## 9 1938 8.2 7359  
## 10 1951 8.1 13765  
## .. ... ... ...

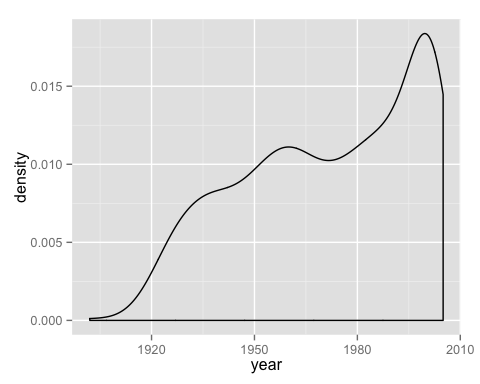
This line of code in dplyr selects only the columns "year", "rating" and "votes" from the already filtered movies dataframe.

(movies11<-select(movies10, year, rating, votes))

## Source: local data frame [690 x 3]  
##   
## year rating votes  
## 1 1957 8.7 29278  
## 2 1971 8.9 252  
## 3 1968 8.3 64982  
## 4 2003 8.0 21857  
## 5 1935 8.0 7931  
## 6 1972 8.6 137  
## 7 1980 8.5 159  
## 8 1926 8.4 169  
## 9 1938 8.2 7359  
## 10 1951 8.1 13765  
## .. ... ... ...

This line of code in dplyr selects only the columns "year", "rating" and "votes" from the already filtered movies dataframe.

(qplot(year, data=movies11, geom="density"))

 this uses ggplot2 to create a density plot with the data points in the above. We can see that the peak is slightly before 2010, meaning more recent years had the most "best popular" movies. Since the question is what year has the most "best popular" movies, to figure out the exact year, I ran this code:

(summary(movies11))

## year rating votes   
## Min. :1902 Min. :8.00 Min. : 112   
## 1st Qu.:1952 1st Qu.:8.10 1st Qu.: 210   
## Median :1975 Median :8.20 Median : 538   
## Mean :1972 Mean :8.28 Mean : 11480   
## 3rd Qu.:1997 3rd Qu.:8.40 3rd Qu.: 9623   
## Max. :2005 Max. :9.30 Max. :157608

From the summary we can tell that the latest year (max) on here is 2005. Also, from the graph I can tell that the most common year is close to the latest year, probably after 2000. To figure out what year had the most "best popular" movies, I used filter starting with the latest year:

(filter(movies11, year==2005))

## Source: local data frame [4 x 3]  
##   
## year rating votes  
## 1 2005 8.4 1013  
## 2 2005 8.0 198  
## 3 2005 8.3 23473  
## 4 2005 8.3 306

But only 4 movies here so that's probably not it.

(filter(movies11, year==2004))

## Source: local data frame [33 x 3]  
##   
## year rating votes  
## 1 2004 8.3 8789  
## 2 2004 8.0 886  
## 3 2004 8.4 154  
## 4 2004 8.0 1453  
## 5 2004 8.3 1660  
## 6 2004 8.2 850  
## 7 2004 8.4 880  
## 8 2004 8.1 152  
## 9 2004 8.0 350  
## 10 2004 8.0 8479  
## 11 2004 8.6 46240  
## 12 2004 8.1 17715  
## 13 2004 8.1 23814  
## 14 2004 8.0 2476  
## 15 2004 8.0 889  
## 16 2004 8.5 8328  
## 17 2004 8.2 350  
## 18 2004 8.3 30749  
## 19 2004 8.3 185  
## 20 2004 8.3 44189  
## 21 2004 8.4 449  
## 22 2004 8.5 117  
## 23 2004 8.3 20195  
## 24 2004 8.7 215  
## 25 2004 8.2 120  
## 26 2004 8.0 17051  
## 27 2004 8.3 399  
## 28 2004 8.0 535  
## 29 2004 8.5 1792  
## 30 2004 8.2 114  
## 31 2004 8.4 8039  
## 32 2004 8.2 201  
## 33 2004 8.0 368

33 movies from 2004 so this could be it

(filter(movies11, year==2003))

## Source: local data frame [33 x 3]  
##   
## year rating votes  
## 1 2003 8.0 21857  
## 2 2003 8.0 267  
## 3 2003 8.1 31525  
## 4 2003 8.2 2829  
## 5 2003 8.4 133  
## 6 2003 8.2 3170  
## 7 2003 8.4 2106  
## 8 2003 8.0 14307  
## 9 2003 8.1 130  
## 10 2003 8.2 147  
## 11 2003 8.3 41846  
## 12 2003 8.4 3416  
## 13 2003 8.2 114  
## 14 2003 8.2 369  
## 15 2003 8.4 496  
## 16 2003 8.3 65026  
## 17 2003 8.0 522  
## 18 2003 8.0 134  
## 19 2003 9.0 103631  
## 20 2003 8.0 42450  
## 21 2003 8.0 144  
## 22 2003 8.8 239  
## 23 2003 8.0 28969  
## 24 2003 8.0 171  
## 25 2003 8.4 8983  
## 26 2003 8.0 60812  
## 27 2003 8.0 1139  
## 28 2003 8.6 140  
## 29 2003 8.1 7112  
## 30 2003 8.0 115  
## 31 2003 8.1 3671  
## 32 2003 8.2 2121  
## 33 2003 8.2 154

Another 33 movies from 2003 so this is tied with 2004

(filter(movies11, year==2002))

## Source: local data frame [25 x 3]  
##   
## year rating votes  
## 1 2002 8.5 36747  
## 2 2002 8.0 440  
## 3 2002 8.7 25964  
## 4 2002 8.6 406  
## 5 2002 8.3 349  
## 6 2002 8.1 716  
## 7 2002 8.5 690  
## 8 2002 8.1 12609  
## 9 2002 8.1 8371  
## 10 2002 8.0 4631  
## 11 2002 8.8 114797  
## 12 2002 8.1 326  
## 13 2002 8.0 4350  
## 14 2002 8.0 115  
## 15 2002 8.3 583  
## 16 2002 8.5 30467  
## 17 2002 8.2 453  
## 18 2002 8.2 1829  
## 19 2002 8.7 242  
## 20 2002 8.0 1073  
## 21 2002 8.5 224  
## 22 2002 8.5 139  
## 23 2002 8.0 10252  
## 24 2002 8.0 4065  
## 25 2002 8.1 24128

In 2002 there were only 25 of these "best popular" movies. We reached an absolute maximum and now we're going down. Therefore, it's a tie between the years 2003 and 2004 as to what year has the most "best popular" movies based on my (subjective) rating system.