Final Project Exploration: TMDB Movies Dataset

Rebecca Hadi

3/7/2018

```
#load packages
library(ggplot2)
library(dplyr)
library(scales)
library(broom)
library(ggthemes)
library(vcd)
```

Here I will conduct exploratory data analysis on the movies data set. At this stage, I have already done some data cleaning and feature engineering with my research question in mind.

Research question: To what extent can the probability that a movie will be profitable be modeled given the predictor variables?

This data set was found on Kaggle at (https://www.kaggle.com/tmdb/tmdb-movie-metadata/data). It contains various data points from the website "The Movie Database" (https://www.themoviedb.org) for 4803 movies. TMDb is a community built movie and TV database. It is not clear how the sample was derived as there are likely more than ~5000 movies that exist.

With my modeling question in mind and uncertainty around how the data were pulled from the site, I wanted to investigate the data set and remove possible sources of bias and skew. The analysis I will present here is after the cleaning, but a summary of that process is contained below. The reason I am analyzing the filtered/engineered data set is because I want to understand the data in the context of how I will be modeling it.

Data removed (movies can meet multiple criteria): * Movies that were not released in English. Upon inspection, the data were highly skewed toward English as a release language. In my project proposal I had included original language as a potential predictor, but after examining my data set I decided that there was not enough data for non-english language movies for this to be a meaningful variable. * Movies that were released prior to the year 2000. It's possible that inflation could skew the input variables of revenue and budget, so I wanted to only look at movies that were released somewhat recently, which I am defining as the year 2000 or greater. * Movies that had zero revenue. These appear to be missing values from the TMDB data set. For example, the movie "Blades of Glory" was listed as having zero revenue (which was consistent with the TMBD site), but a quick google search revealed this movie actually had $146M in revenue.
* Movies that had not yet been released. * Movies with the genre of "TV Movie". This is a different type of movie than what we are trying to model (e.g. revenue from box office)

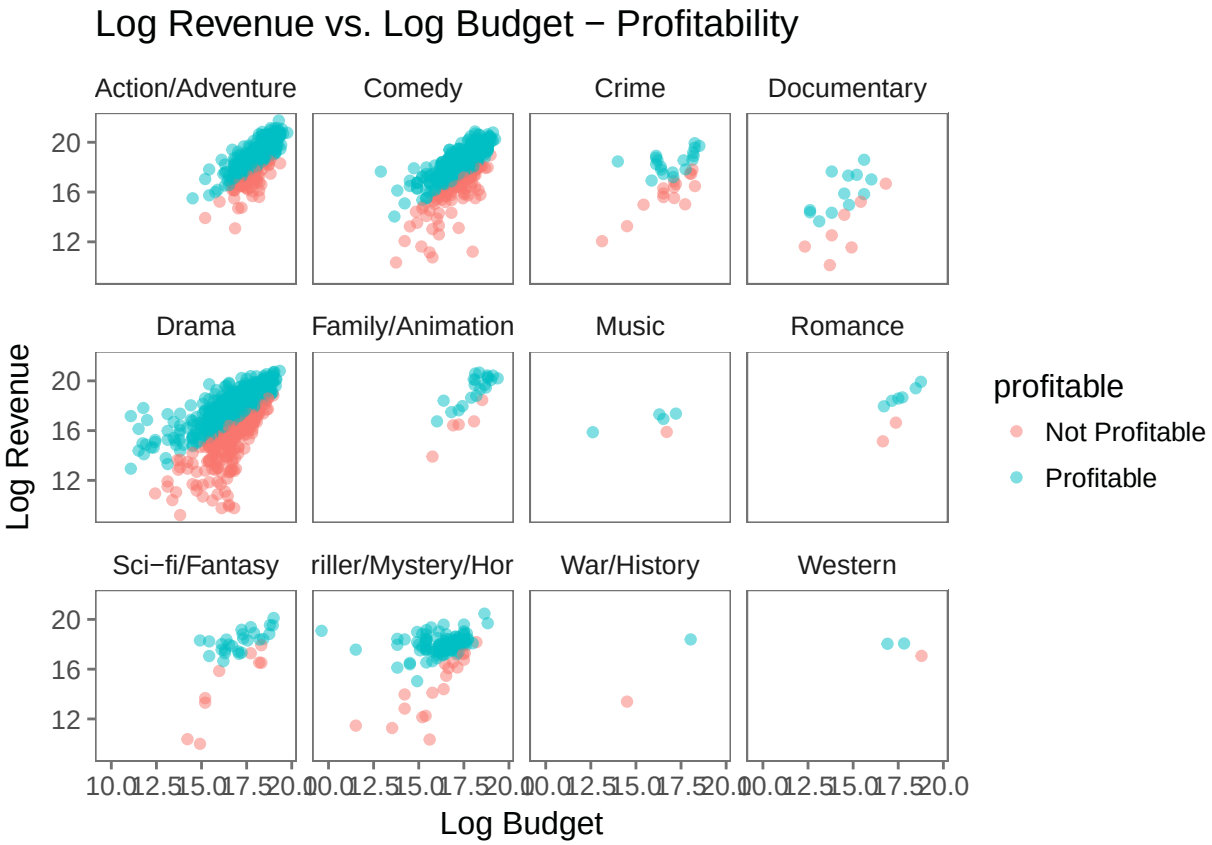
Features engineered: * Profit: Difference between profit and revenue * Profitable: binary indicator whether or not profit was greater than 0. This is the response variable. * Release year: extracted from release date * Genre: The original genre data was in JSON format. Upon extracting, it created a data frame that had one row per genre (making a single movie have as many rows as distinct genres). I had a few ideas on how to handle, ranging from picking a primary/arbitrary genre for each movie to force there to be one row, or allowing the data to have multiple rows. I ended up manipulating the data to create a column specific to each genre (e.g. "f.action" is a 0 or 1 if the movie has action is the genre). Then, I created meaningful grouping of genres based on what the most common genres in the data set were and common groupings that exist in popular culture. * Event: Identifies if the movie was released during a seasonal event (Holiday - Nov/Dec or Summer - June, July, August) or not.

Further, I removed some columns from the data set that were not going to be used in analysis (such as keywords, production company)

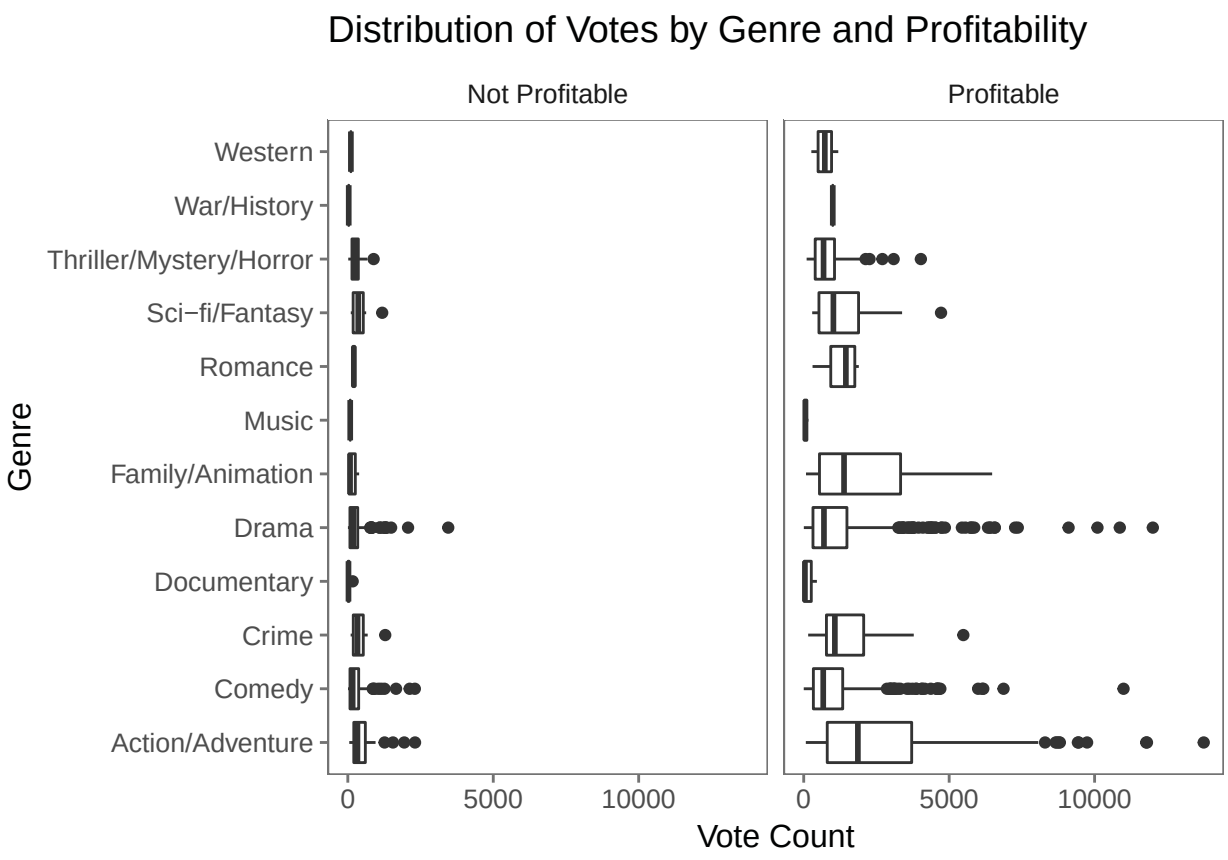
Without further adieu, let's take a look at the final data we are working with. In my final data set I have 2146 movies I am working with.

X	title	budget	revenue	profit	profitable	release_date	release
1	Avatar	2.37e+08	2787965087	2550965087	Profitable	2009-12-10	
2	Pirates of the Caribbean: At World's End	3.00e+08	961000000	661000000	Profitable	2007-05-19	
3	Spectre	2.45e+08	880674609	635674609	Profitable	2015-10-26	
4	The Dark Knight Rises	2.50e+08	1084939099	834939099	Profitable	2012-07-16	
5	John Carter	2.60e+08	284139100	24139100	Profitable	2012-03-07	
6	Spider-Man 3	2.58e+08	890871626	632871626	Profitable	2007-05-01	

X	title	budget	revenue	profit	profi
Min. : 1.0	(500) Days of Summer: 1	Min. : 15000	Min. :1.002e+04	Min. :-165710090	Not Profi
1st Qu.: 537.2	10 Cloverfield Lane : 1	1st Qu.: 15000000	1st Qu.:2.253e+07	1st Qu.: -27521	Profitab
Median :1073.5	102 Dalmatians : 1	Median : 31000000	Median :6.715e+07	Median : 32658323	N
Mean :1073.5	10th & Wolf : 1	Mean : 48643386	Mean :1.388e+08	Mean : 90187012	N
3rd Qu.:1609.8	12 Rounds : 1	3rd Qu.: 65000000	3rd Qu.:1.609e+08	3rd Qu.: 105421398	N
Max. :2146.0	12 Years a Slave : 1	Max. :380000000	Max. :2.788e+09	Max. :2550965087	N
NA	(Other) :2140	NA	NA	NA	N



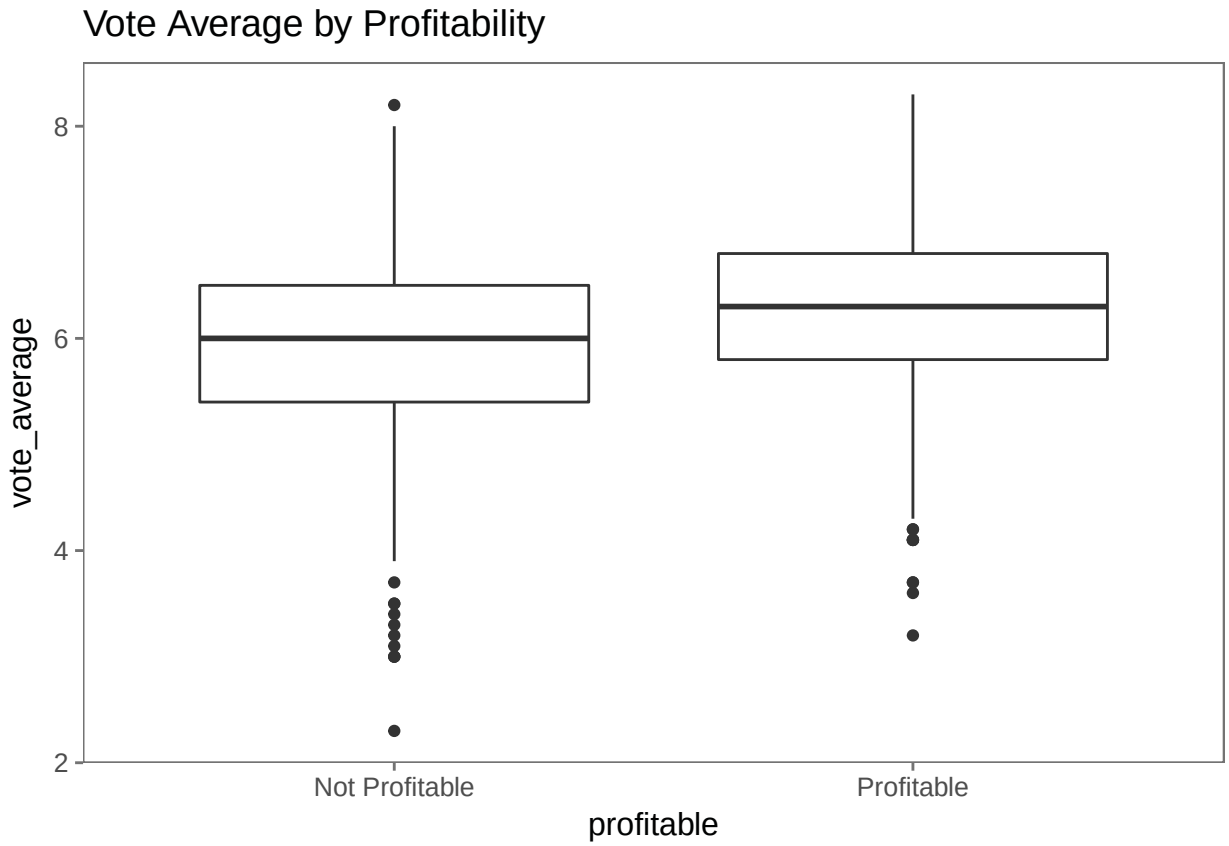
The real measure we are interested in is whether or not the movie was profitable. Let's look at how that varies by some of our potential predictor variables.



Across most genres, the movies that are profitable tend to have a higher number of votes on TMDB. This is not true across all genres, but something we observe in general.

Let's see the relationship between vote average and profitability.

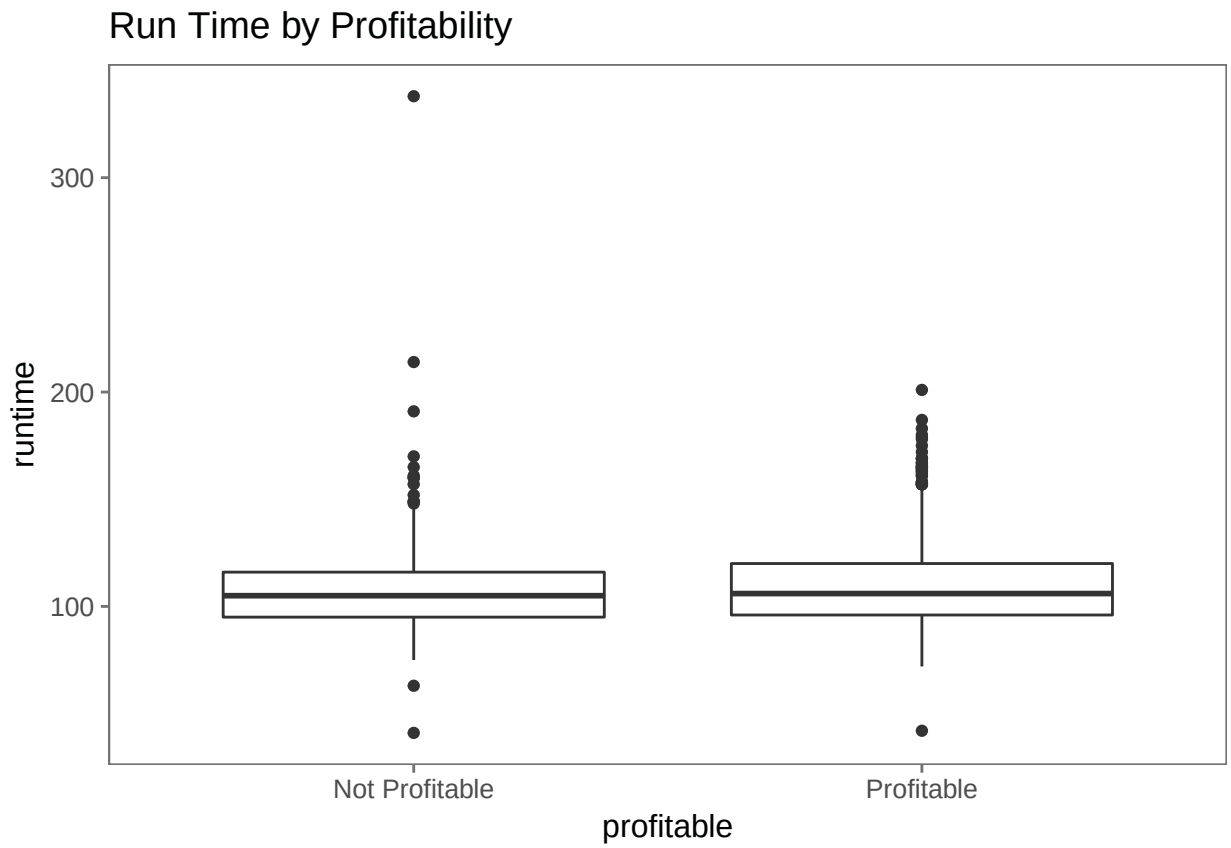
```
ggplot(data = movies.pr, aes(x = profitable, y = vote_average)) +
        geom_boxplot() +
        theme_few() +
        ggtitle("Vote Average by Profitability")
```



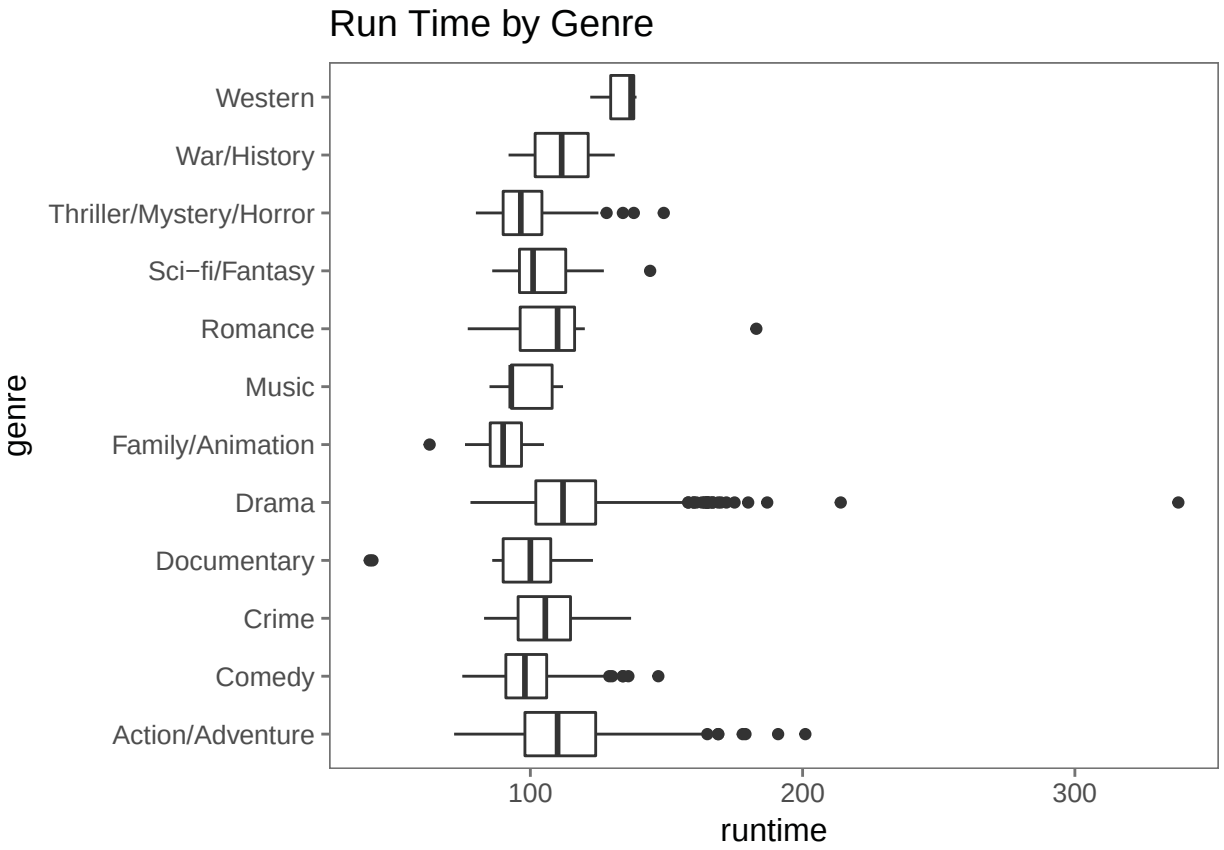
We see that profitable movies have a slightly higher median those those that are not profitable, but they seem to be similarly distributed.

Another predictor variable we can consider is run time.

```
ggplot(data = movies.pr, aes(x = profitable, y = runtime)) +
        geom_boxplot() +
        theme_few() +
        ggtitle("Run Time by Profitability")
```



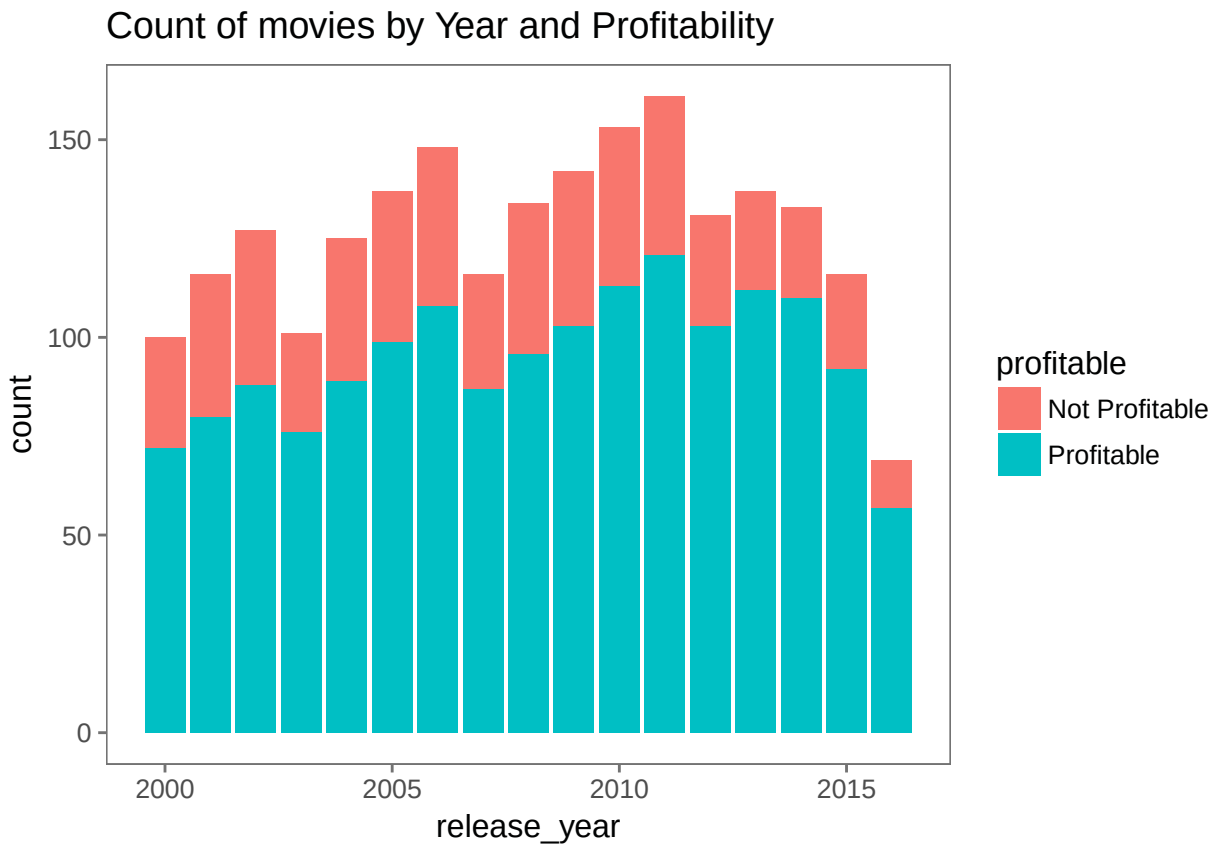

```
#what's the distribution of runtime by genre?
ggplot(data = movies.pr, aes(x = genre, y = runtime)) +
        geom_boxplot() +
        theme_few() +
        coord_flip() +
        ggtitle("Run Time by Genre")
```



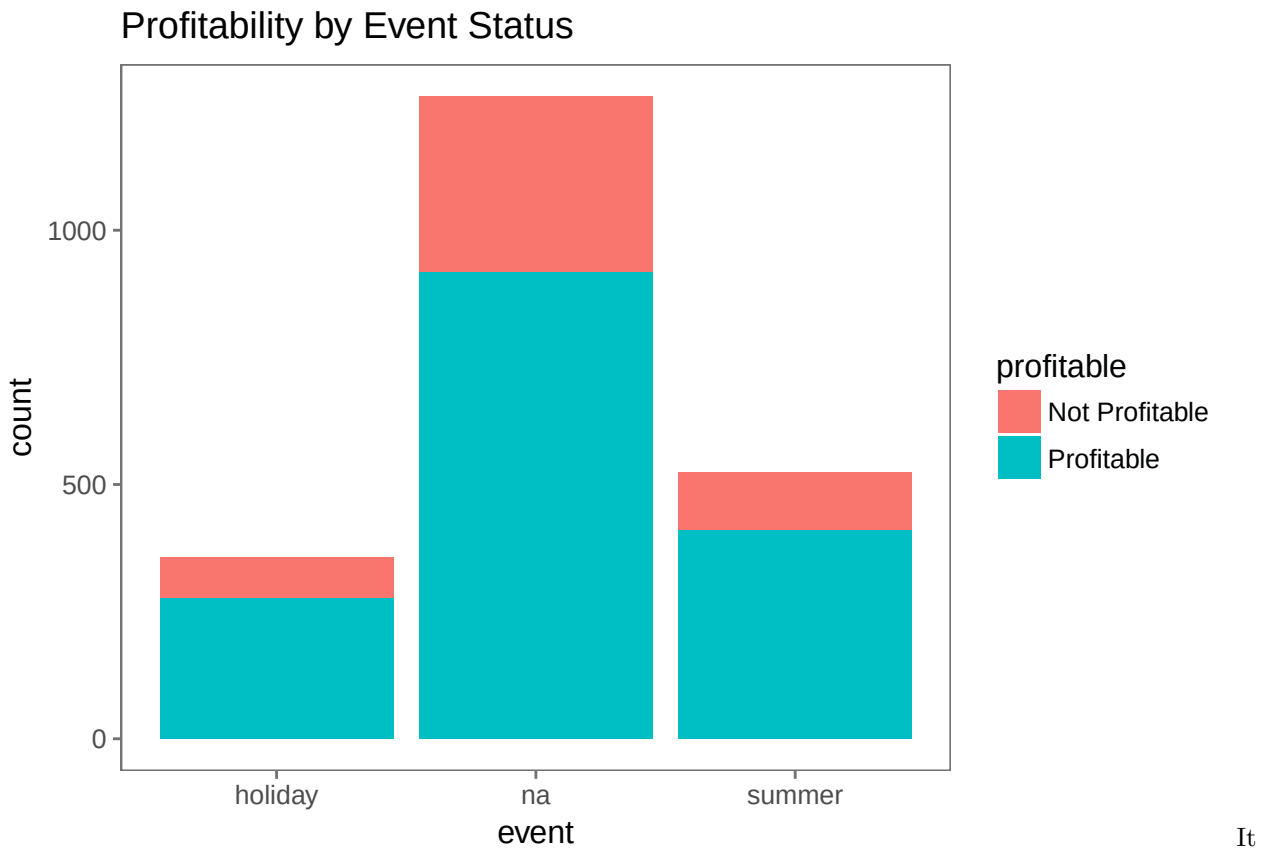
We see that there is some variation in run time by genre. Dramas tend to be longer. Family/Animation tends to be shorter. We also see that there isn't really a difference in run time in terms of profitability.

Now let's see if time has an impact on profitability.

```
ggplot(data = movies.pr, aes(x = release_year, fill = profitable)) +
          geom_bar(stat = "count") +
          theme_few() +
          ggtitle("Count of movies by Year and Profitability")
```



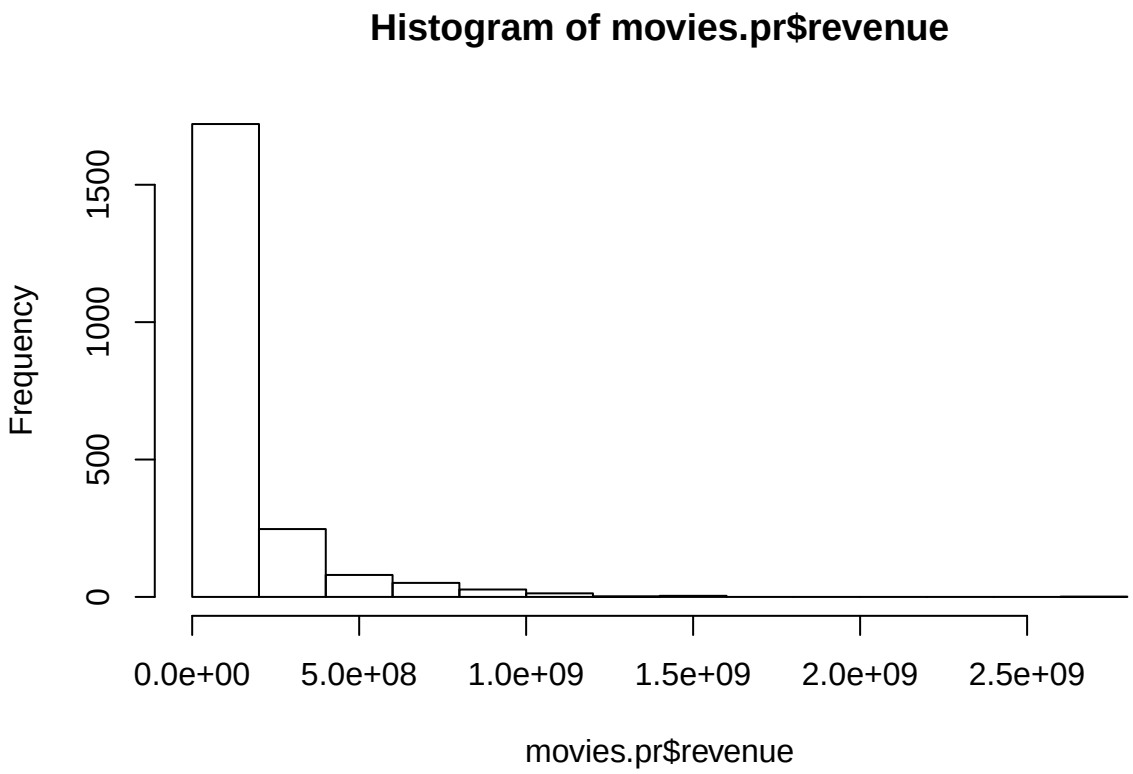

```
#What's the relationship between event and profit?
ggplot(data = movies.pr, aes(x = event, fill = profitable)) +
          geom_bar(stat = "count") +
          theme_few() +
          ggtitle("Profitability by Event Status")
```



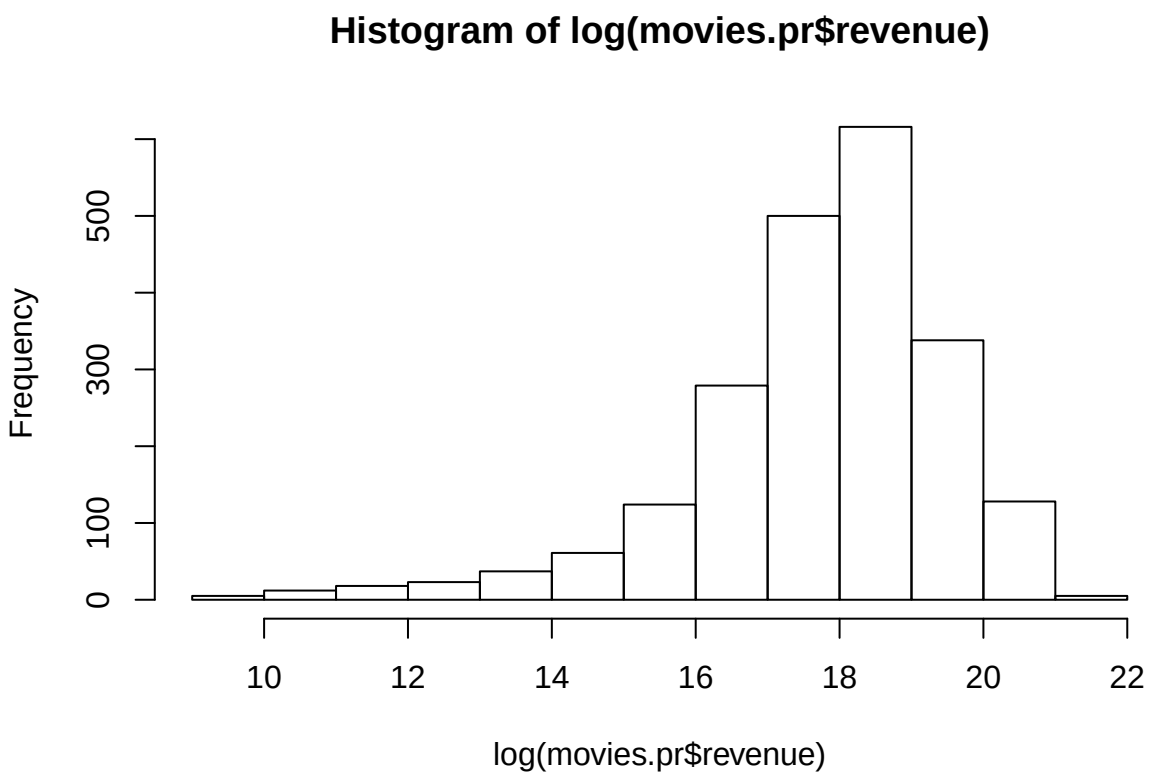
It looks like the proportion of profitable movies is similar between summer and holiday movies, and slightly lower during other points of the year.

So far our focus has been on profitability, but let's get a more broad look at our data set.

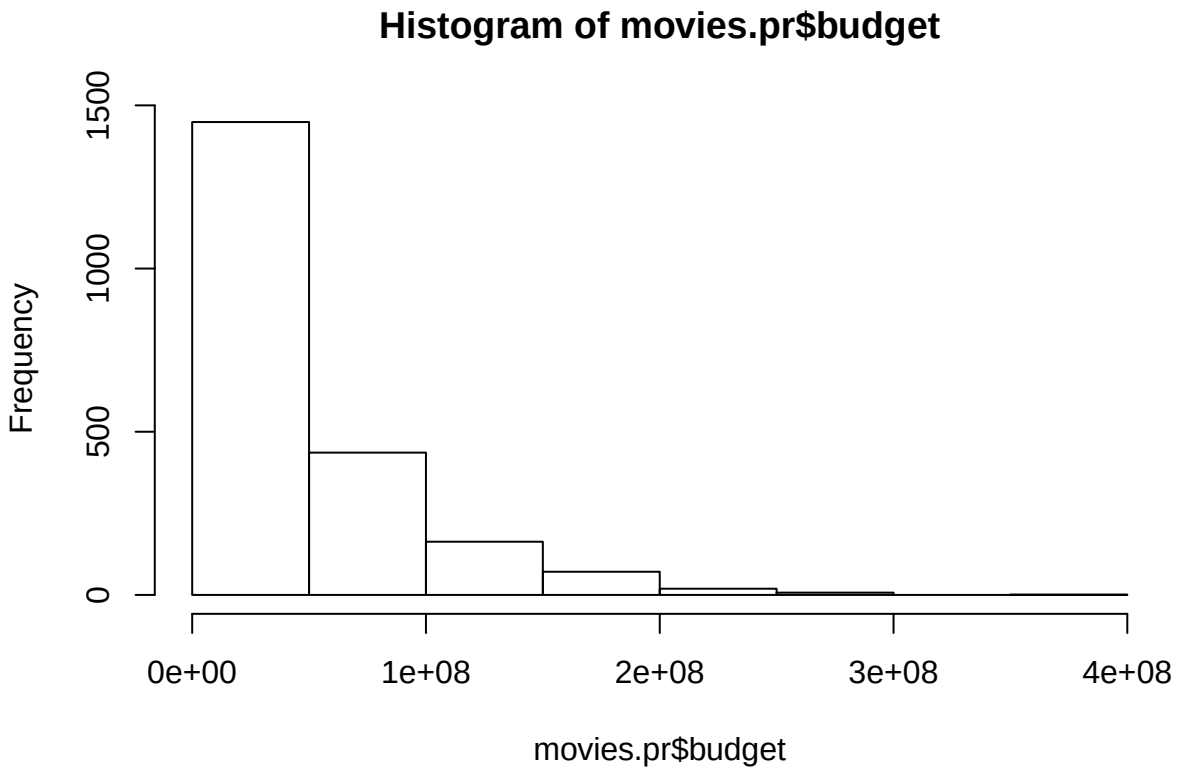
```
#does the log transform improve normality?
hist(movies.pr$revenue)
```



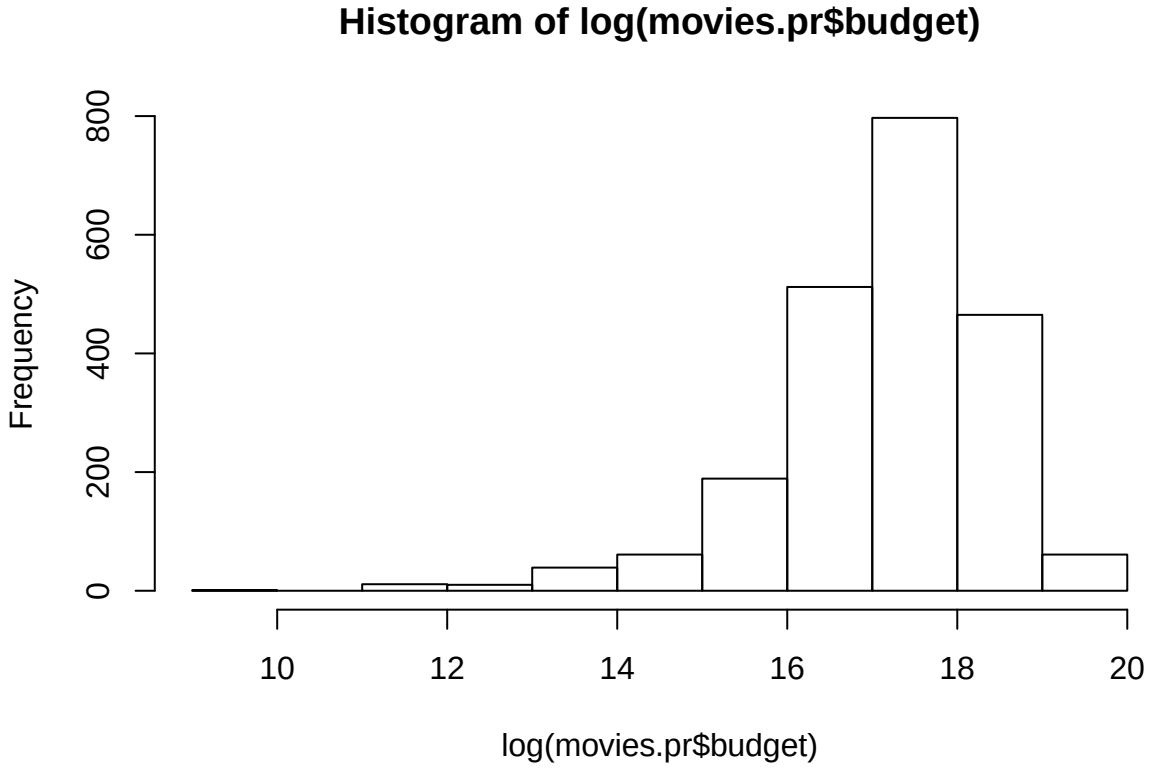

```
hist(log(movies.pr$revenue))
```




```
#what about budget
hist(movies.pr$budget)
```



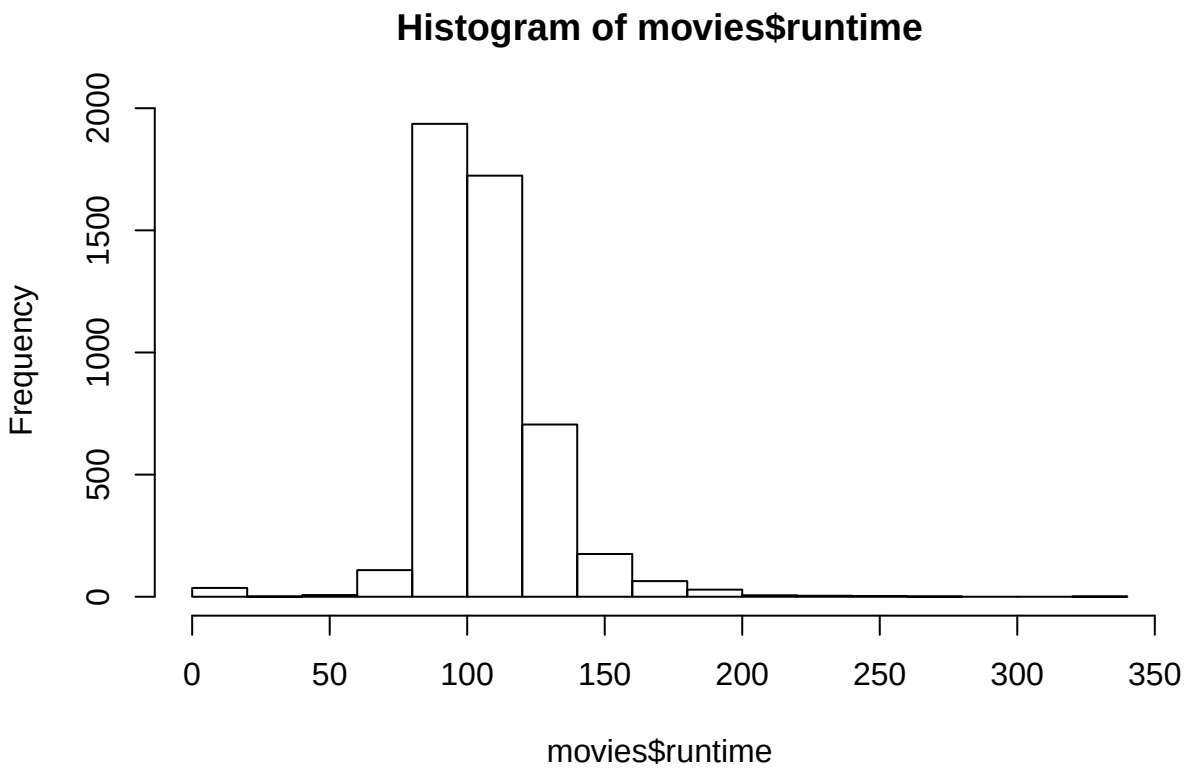

```
hist(log(movies.pr$budget))
```



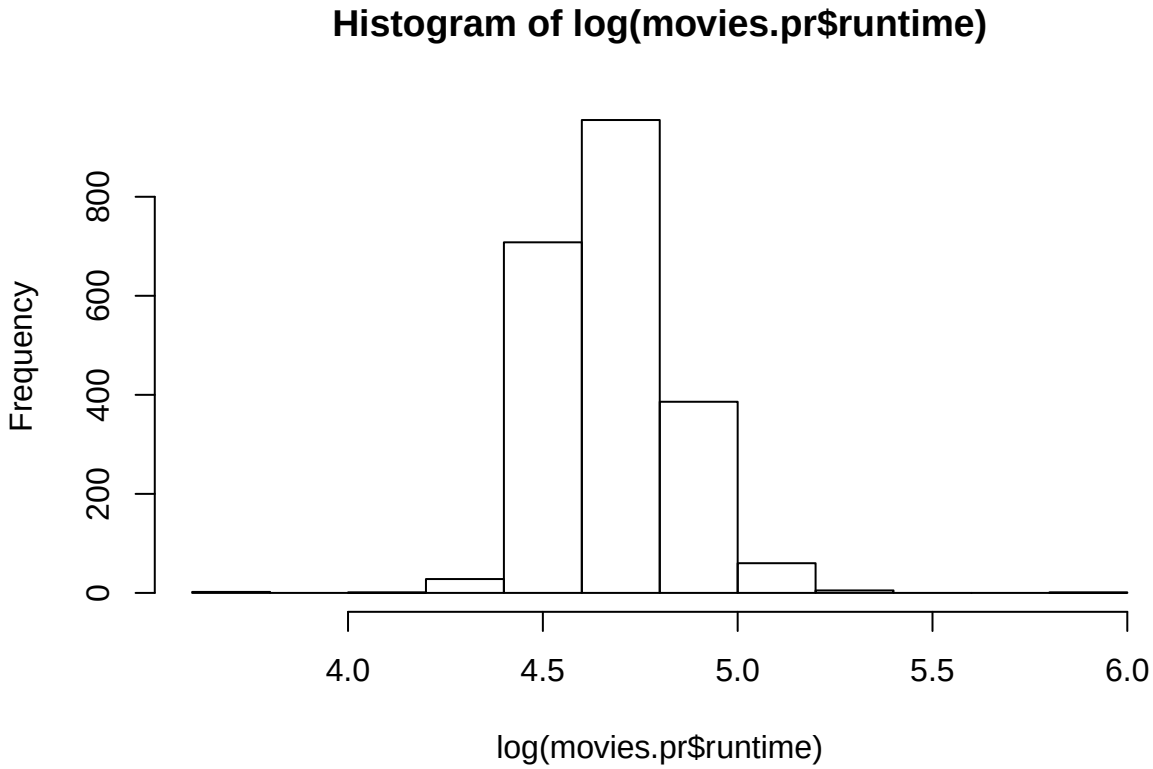
Applying a log transform improves our variables for budget and revenue, but even after the log transform there is some skew in the distribution.

What's the distribution of run time?

```
hist(movies$runtime)
```




```
hist(log(movies.pr$runtime))
```



Suprisingly, a log transform applied to run time improves the normality of this variable.