~~STAT 405 Project 1~~ - An Analysis of Movie Ratings from the Internet Movie Database (IMDb)

October 7, 2010

1 Introduction

The Internet Movie Database, IMDb, was created in October of 1990 and since then has grown to become one of the leading resources for information on movies, television shows, and video games. The users of IMDb are allowed to rate the films between 1-10 and thus, the information on IMDb is an accurate portrayal of the opinions of the general population. The data that was provided from IMDb consisted of 115338 movies produced between 1891 and 2010. We decided to focus our analysis on the ratings of the movies and their relationship with other variables like genre, time, length, and budget.

2 Our Subset

The wealth of data did not make our analysis easy so we decided to focus on a subset of the data in which the movies received more than 1000 votes. We chose to subset the data this way because movies with very few votes may not accurately reflect the quality of the movie, as the small sample size would be susceptible to bias from individual voters. We decided that 1000 votes was a large enough sample to consider the rating to be accurate. We also took all the movies classified as "short" out of the data set because they seemed to cause unnecessary variability due to their inherently low budget and short length. It appeared that because of the budget and length of these movies, they had different standards and viewers. From the data it seemed as if short films were not as well known to the general population anyway, as not many of them had more than 1000 votes to begin with. We ended up taking 155 short films out of the new data set to end up with 10565 movies.

In addition to trimming the data set to better suit our investigation, we created several new variables from the existing ones that would be helpful in our analysis. We created variables for the variance, the length of the title, a variable giving which of 8 quantiles the movie rating fit into, whether or not IMDb had been established when it was produced, the decade it was produced in, the standard deviation, a boolean variable for whether or not it was only one genre, and a boolean variable for whether or not the movie is in the 8th quantile for their it was made.

2.1 Comparing the Original and Subsetted Data

Before we began our investigation, we decided it would be beneficial to compare aspects of our subset to aspects of the original data set to note the differences and be aware of what data had been excluded from our analysis.

Note that there are not many extreme values of rating because with that many people voting, it is hard to reach a consensus that a movie is really good. We can also see that the mean rating is slightly higher for the new data set (6.11 for the original, 6.39 for the subset). We would expect that ratings would follow a

could you overlay them to make comparison easier?

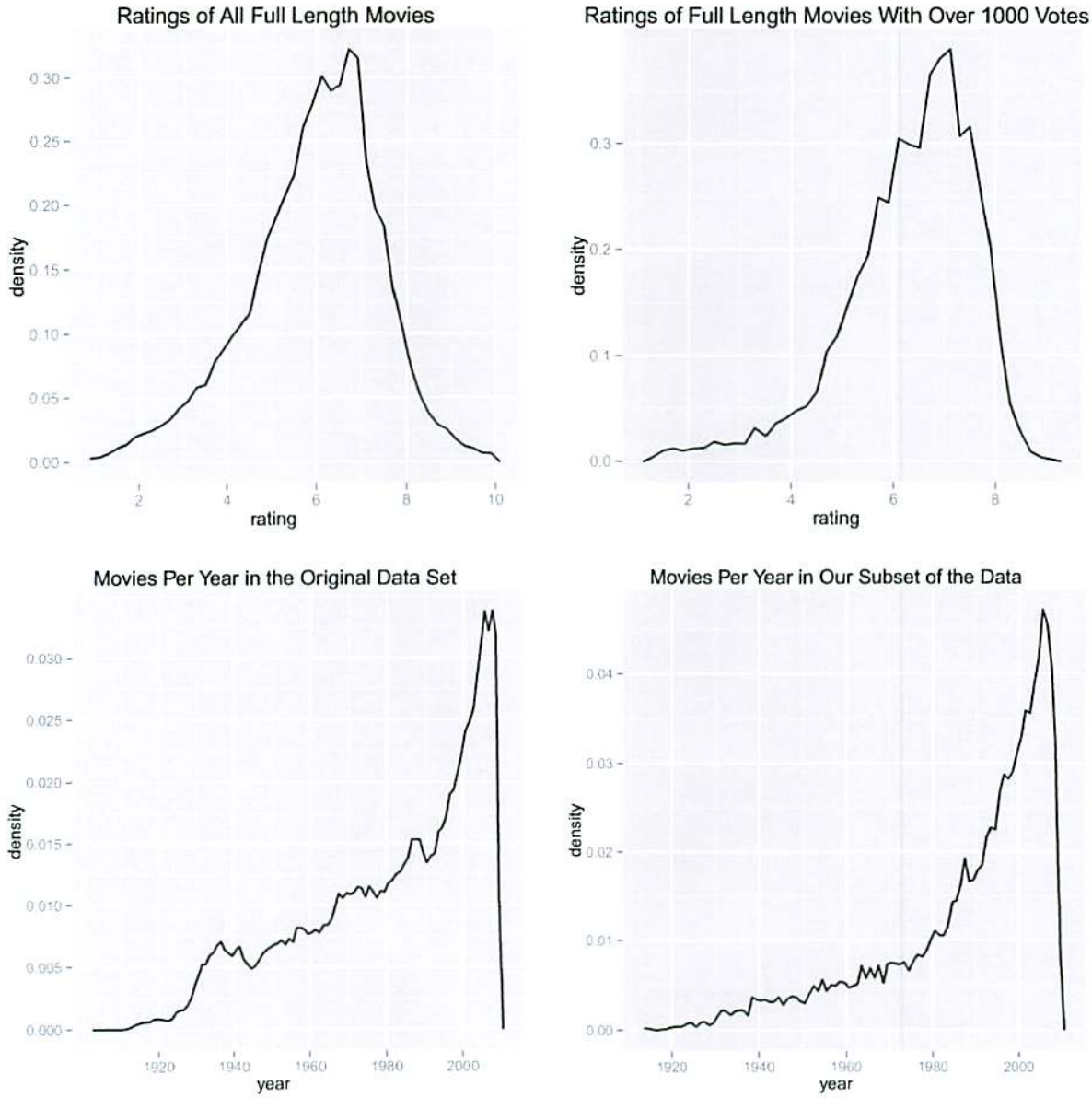

Figure 1: (Top Left) A plot of the distribution of ratings for all of the full length movies. All short movies were removed from this data set for more accurate comparison. (Top Right) A plot of the distribution of ratings for our subset of the data. (Bottom Left) A histogram of the movies produced every year in the original data set (excluding short films). (Bottom Right) The same histogram from our subset of the data.

Table 1: Five Number Summary of Ratings

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Our Subset	1.30	5.70	6.60	6.39	7.30	9.20
All Movies	1.00	5.20	6.30	6.12	7.10	10.00

why?

normal distribution naturally, but we can see that the data is slightly left skewed, and even more so in our subset, indicating the error associated with voting being open to any user. If there was a standardized rating system, for instance, a panel at IMDb that rated every movie, so that every movie would be rated on the same standards, it would make our analysis much easier. This is the sort of data that it would be nice to have in future explorations. As it stands, our data indicates a tendency for users to vote only on movies they like, neglecting the ones that were not as good.

The bottom two plots of Figure 1 are also helpful when comparing our subset to the original data. We can see lots of dips and bumps in the plot of the original data, whereas our subset shows a more consistent upward curve. Why our subset hides the random fluctuations from the norm is somewhat mysterious. It essentially seems to tell us that regardless of random noise in the productiveness of the movie industry, the number of movies that will gain popularity and wide spread recognition is relatively fixed in its growth.

good question!

2.2 Ratings In Our Subset

Ratings in our subset ← please use sentence case.

Figure 1 allows us to see the distributions of the ratings in our subset, but we decided it would also be worthwhile to look at the standard deviations of these ratings, so that we could get an idea of the variability of opinions expressed by the users of IMDb, whose ratings we will be analyzing throughout our investigation.

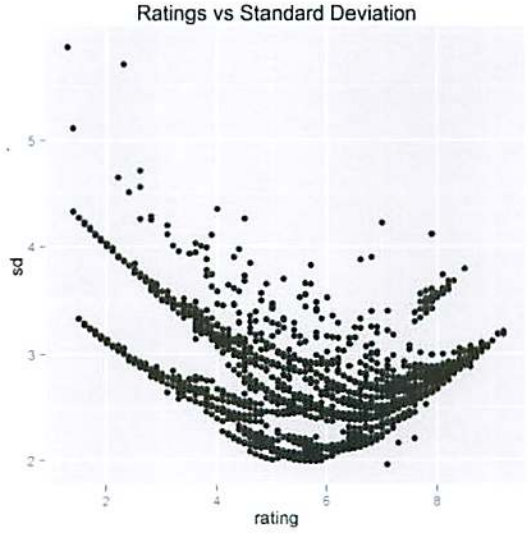

Figure 2: A plot of our standard deviation against the corresponding rating.

how did you compute s.d?

...

your code needs more explanation

...

diagnostic plot of r1-r10 facetted by s.d. would be helpful.

Figure 2 seemed very bizarre to us. In general the trend seems to be that for extreme values of rating, the standard deviation tends to be higher. This would indicate that for very good or very bad movies, people tended to disagree on the rating. This didn't make much sense to us, as it seemed like if people disagreed on

whether a movie was good or bad, then the mean would average out somewhere in the middle, not occur towards either of the extremes. It could be that the standard deviation for values at the extremes would be influenced more by each vote towards the other extreme, explaining the large variability. However, it could also be partially due to the error associated with our calculation of the standard deviation.

The standard deviation was computed based on the variables r1 - r10. However, these values were rounded to the midpoint of the nearest decile, so they are far from exact. As a result, our standard deviations, which were gained by applying the definition of variance using r1 - r10 as the probabilities for each rating, could also be highly inaccurate. Despite the error, this at least gives us a ballpark idea of the variability of IMDb users' opinions, which may prove helpful in our exploration of this data. We will divide our further investigation into four sections. Examining the behavior of movie ratings over time, across genre, by length, and by budget. From this analysis we hope to uncover some of the qualities that cause movies to be rated highly by the users of IMDb.

3 Ratings by Time

One of the major factors contributing to the rating of a movie is probably the time at which it was produced, because as time has passed, movies and the industry that produces them have changed drastically.

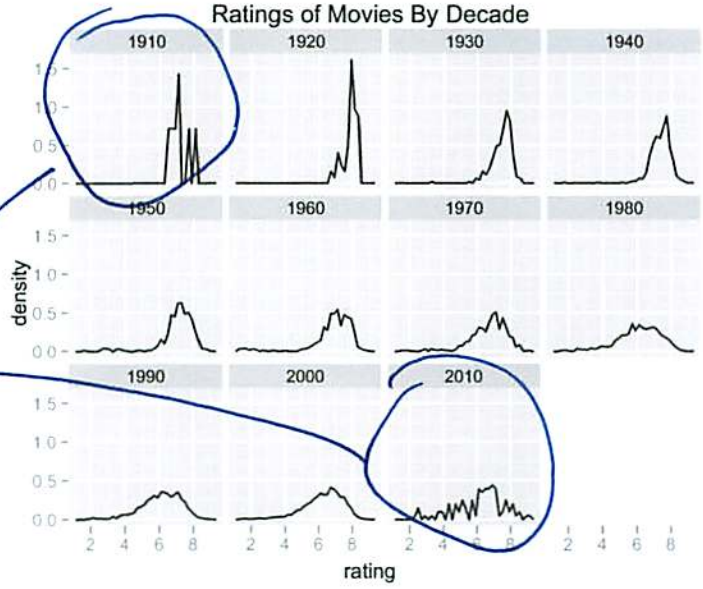

Figure 3: Plots showing the distribution of ratings by decade

We can see that between 1910 and 2010, there have been less and less movies per decade that are as highly ranked. It makes sense to think that if movies from 100 years ago are still being watched and/or voted on today, they must have made a distinct impression, and that usually means to the general population, it was a good movie. But as time progresses, more movies flood the market, and with the increase in the number of movies produced a year, it makes sense that a smaller percentage of those movies would be deemed good in everyone's eyes.

We found it interesting to look at movie ratings during historically important time periods like the Great Depression and World War II. The Great Depression lasted from 1929 to approximately 1940 and ended because the onset of World War II caused a rise in employment rates. One would think that because of

the scarcity of resources and money, movies from that period would be poorly rated due to low budgets in Hollywood, or not even widely known. However, based on the plots from out data subset, we determined that ratings of movies from that period were actually higher than the average rating of the entire subset.

World War II lasted from 1939 to 1945, and with so many countries involved in the war, the majority of each country's resources were diverted to funding the war. The people at home were still suffering the aftermath of the Great Depression and with a shift to a war economy, it was surprising that as many people were able to go see movies. However, based on our plots from the data, it still holds true that, on average, movies from that time period have higher ratings.

However, Figures 3 and 4 must be taken with a grain of salt because the people rating these movies are people living between 1990 and now, not many of the people who were alive when the movies were produced. In this case there may be a selection bias towards higher ratings. People watching movies from this long ago are unlikely to rate them if they were not good, so they may only vote on movies they thought were good, effectively raising the average rating from this time period by excluding potentially negative ratings. The same would hold true for any movie that came out during the World War II time period.

This raises another issue with the data we are provided. While we are given movies that span over a century, we are given ratings that span only over the last 20 years. It is hard to make inferences about the changes in ratings over time when all of the ratings were done in the most recent decades. A data set that included ratings done throughout the same time span as the movies were produced would make our results a lot more reliable. However, there was not really an equivalent system to IMDb in the early 20th century, so this is unfortunately a bit much to ask for.

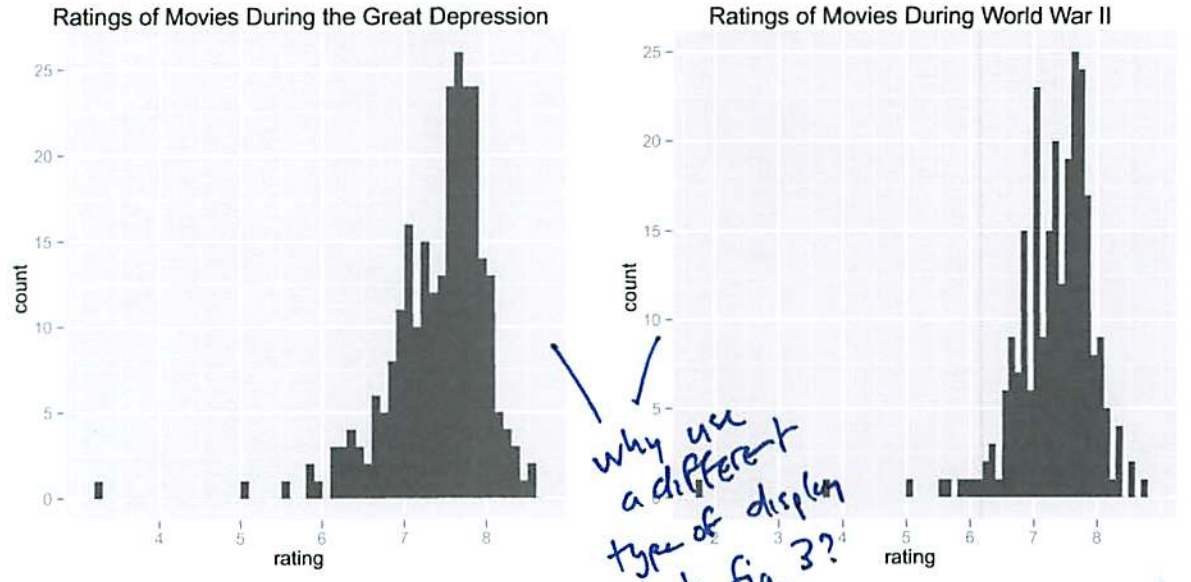

Figure 4: Plots of the rating distribution during the Great Depression and World War II.

4 Ratings by Genre

Movies are most obviously categorized by which genre they belong to because that speaks volumes about what type of people will go see the movie. Production companies will try to catch the interest of certain

groups of people, like teenagers or children or sci-fi fans, because they will go see a certain kind of movie. Our data was split up into six genres: action, animation, comedy, documentary, drama, and romance. The other genre was short movies but we took those movies out of our data set because they seemed too different from our other genres of movie. We subsetted the data again for this section to include movies that were listed under only one genre, to observe how the genre of a movie affected its ratings. We chose to look at movies that were exclusively one genre because if we included the movies that were listed under multiple genres, there would be overlapping data and it would be much more difficult to compare genres without including every possible combination of the 6 genres (i.e. romantic comedies, animated action movies, etc.) This subset left about about 5000 movies from the original subset, which is about half, but certainly still a reasonable amount of data on which to perform our analysis.

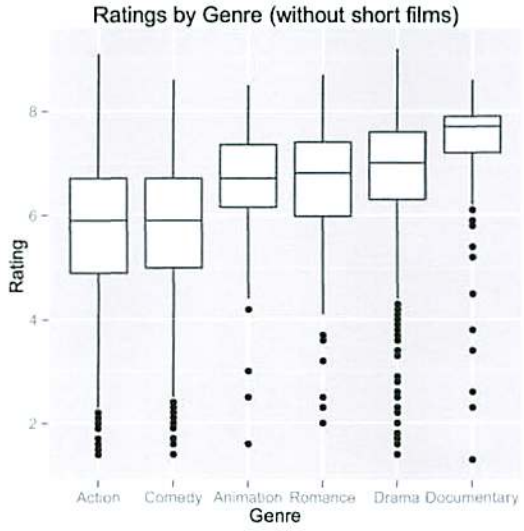

Figure 5: A plot examining the distribution of ratings across genres.

It is easy to see that documentaries get the highest ratings, but with further examination, it becomes apparent that there are less documentaries made than movies of the other genres. Not only are there less documentaries made, the fan base for documentaries is smaller and probably more devoted. For the most part, documentaries seem to target a specific crowd, unlike comedies or action which appeal to a wide range of people. It is probably because comedies and action movies attract so many people that they have such a wide range of ratings. There are many instances of movies in these two genres that are rated very highly, but because there are also so many of them which have low ratings, the average of these two genres is still lower than the other genres. Action films and comedies may be profitable for the production companies and probably not too difficult to produce so there are a large number of them but they are not necessarily all of high quality.

However, there are certainly some limitations to the way we subsetted our data. In terms of count, there are fewer movies that are listed as purely romance, which left us with a smaller sample to deal with. This lack of representation for romance could be due to the fact that romance can be tied in to nearly any story line, and therefore will often be cross listed under another genre as well. It is even possible that the prevalence of romance in other genres of movies makes purely romance movies less interesting because it has saturated the market already.

Another genre that has very few movies is animation. However, the issue with animated movies is not quite the same as romance, because while the total count of movies that are listed under romance (not

necessarily exclusively) is high, the total count for any combination of animated films is also rather low. So the problem is not that it is often listed in conjunction with other genres. It could be because for the most part, animated films are made to target children, who are typically not the ones voting on IMDb. While it's true that often parents have to go along for the ride, it is less likely that the parents wanted to go see the movie themselves, and less likely that they will be voting on it on IMDb. This may have given the animated films fewer votes on average, and therefore not as many made it into our original subset of the data.

5 Ratings by Budget

The final two aspects of the movies we analyzed were the budget and length of each movie. Can a movie get higher ratings if you pump more money into it? And do movies that are longer tend to be more appealing to audiences for their depth, or do they usually drag on and bore viewers into rating it poorly? We begin with a look into the effect budget will have on the rating of a movie.

First we looked at the relationship between the budget of highly rated movies and the average budget per year. For the highly rated movies we picked movies that were in the 8th quantile (88.75% - 100%) because we wanted to guarantee that the 'highly rated movies' included an equal proportion of the total movies for each year. We took a ratio of the average budget for these highly rated movies and the average budget for all movies in that year to see if on average, better movies had higher budgets. Before plotting we took the log of this ratio to restore the symmetry about 1.

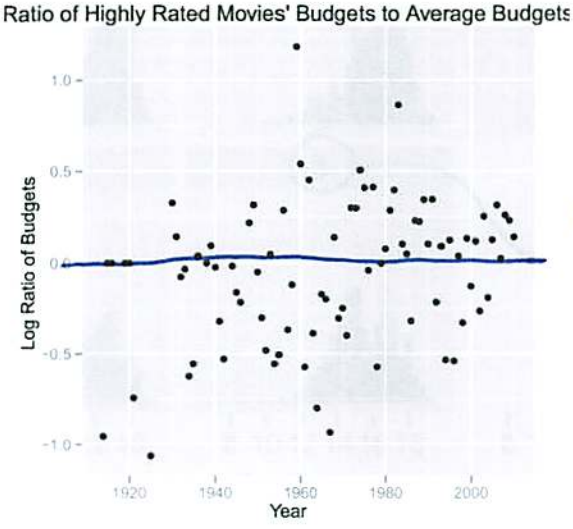

Figure 6: A plot examining the ratio of the budget of highly rated movies (in the 8th quantile for that year) to the average budget in that year.

The data in Figure 6 looks mostly uniform, indicating no permanent correlation between the budget of a movie, and its IMDb rating (The mean ratio was 1.03). This plot leads us to believe that the differences between the means that we observe on a year to year basis can almost certainly be attributed to pure random noise, not to a tendency for better funded movies to do better on IMDb.

There are again some problems with this plot though, primarily in the earlier years of the data. Some of the values there are missing, as no budget data was available in those year, and some of the data is also exactly 0, which it turns out is a result of the lack of movies produced in that year for our data set. There may only have been one movie, so it would simply be divided by itself, yielding one, or a log value of 0.

have *lower* budgets. This relationship is confirmed by the table of means for each quantile.

Table 2: The Average Budget of Movies per Quantile

Quantile	Average Budget
1.00	21026160.44
2.00	23729125.37
3.00	23706380.72
4.00	24574344.43
5.00	22600441.05
6.00	18467816.75
7.00	17910960.88
8.00	17165290.29

Why do these plots tell different stories? Figure 6 seems to indicate no definitive relationship between the budget of a movie and its resulting ratings, while Figure 7 seems to suggest that higher rated movies have lower budgets. It could be due to the fact that in the plot of ratios, movies that are included in the 8th quantile for that year may not be in the 8th quantile for all movies, so the relationships are not going to be quite the same, as they are based on marginally different data. In this case, which one is more accurate? This was a question we were not quite sure how to answer. The conclusion we came to was that Figure 6 gives you a better idea of what's going on as it makes comparisons between more similar movies than our quantile plots. Grouping the movies by year eliminates some of the discrepancies created when comparing movies from across a wide swath of history. This seems especially true when you consider that the budgets for movies has increased drastically since the early 20th century.

6 Ratings by Length

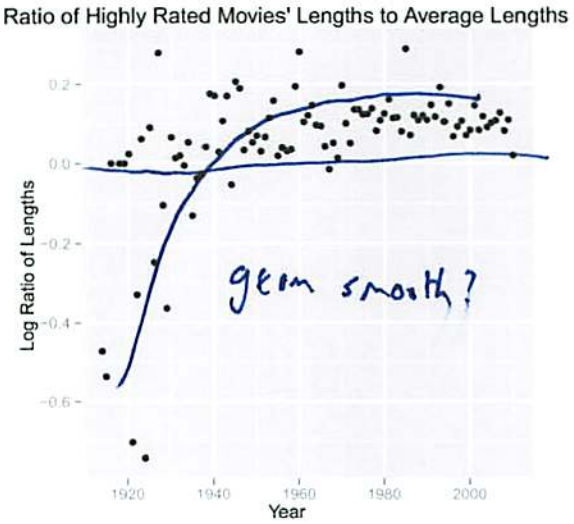

Figure 8: A plot examining the ratio of the length of highly rated movies (in the 8th quantile) to the average length.

When observing length we chose to take the same approach as we did to budget. We took the mean length of movies in the 8th quantile for their respective year, and divided it by the mean length for all of the movies in that year. We again took the log of this ratio for symmetry's sake.

An initial glance at Figure 8 can be very misleading, as it appears as if shorter movies were much more highly ranked in the early years of cinema. However, when we took a closer look at these years, we found that it was often due to there being so few movies that only one movie would be in the 8th quantile. So if that movie happened to be rather short, the ratio would drop quite a bit, so these very low points can probably be considered outliers. After around 1940, the data appears to stabilize a bit, and it appears as if there may be a tendency for longer movies to get better ratings.

Taking the mean ratio we find that it is about 1.06. This makes it seem as if there may be a relationship, but not a strong one and there is certainly no definitive evidence. We decided that removing the earlier years (before 1940) would remove some of the heavily biased data, which could be influencing the mean. When we took the mean ratio for movies after 1940 we got 1.11, a much more convincing figure, that seemed to fit with the data. This was somewhat surprising as we had not expected length to be a factor in the quality of the movie, but it seems as if movies in the 8th quantile tend to be around 10% longer on average. We also took the median of the ratios over the entire time span, as the median would not be as sensitive to the outliers in the earlier years. This yielded 1.09, which again points to about a 10% increase in length between all movies and the highly rated ones. As with budget we will proceed to examine the distributions of the length of movies when facetted by their quantiles.

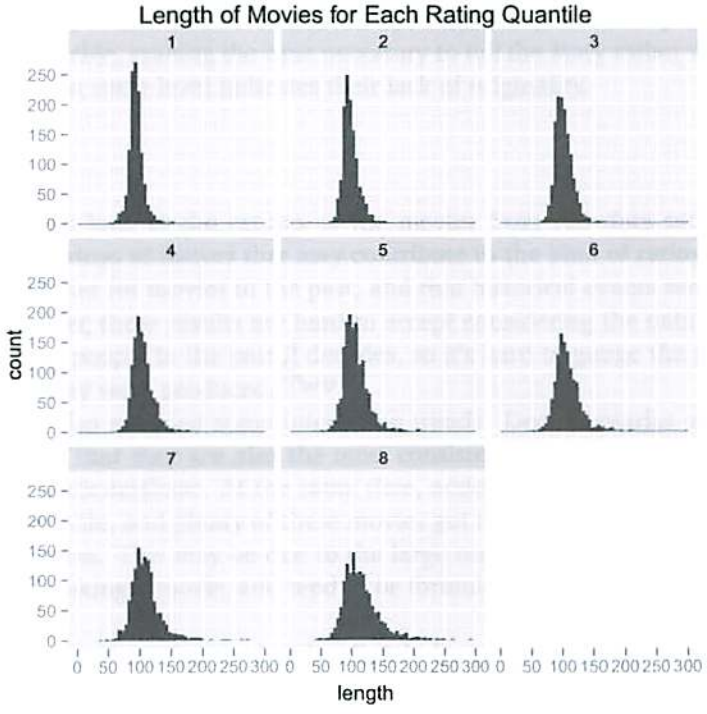

Figure 9: A plot displaying the distribution of the length of movies by the quantile of their ratings.

Figure 9 shows us how the length of movies looks like when broken up by the eight rating quantiles. This plot is very interesting because the distribution looks more or less normal throughout, except with an increasing standard deviation as the ratings get higher. It also appears that the mean length may increase

budgets and more of them produced with what seems like unnecessarily lax financial constraints. When we added time to the equation, though, we observed that there is also evidence to suggest a lack of relationship between budget and movie ratings.

We have done our best to explore the eccentricities of movie ratings with the data we had access to. However, we found the data in the end to be very limiting. Without a standardized rating system we couldn't guarantee that all movies were judged on the same standards, and therefore could not confirm their accuracy in reflecting the quality of the movie. Data was also provided on movies from the late 19th century up until today, but all of the ratings were made in the last two decades, making our analysis of the effect of time on the movie ratings particularly dubious. Having the ratings at the time the movies were produced would make our investigation a lot more reflective of the actual relationships between history and the movie industry. We also ran into trouble observing the variability because the distributions of the ratings for each individual movie were rounded to the midpoint of the decile, which created a lot of error when we calculated the standard deviation. Collection of data that is mentioned here would allow for a much more rigorous and thorough analysis, although these demands are a rather tall order considering they entail traveling back in time.

P 4
C 4 ⇒ 20/25
E 4

A Code

```
movies <- read.csv('movies.csv.bz2')
vote1000 <- read.csv('vote1000.csv')
library(ggplot2)
library(stringr)
library(xtable)

================================================================================

#Collect the non-short films with over 1000 votes

vote1000 <- movies[movies$vote >= 1000, ]
vote1000 <- vote1000[vote1000$Short == 0, ]

================================================================================

#Use r1 - r10 to create the variance and standard deviation of ratings

var1 <- matrix(NA, nrow(vote1000), 10)
for (i in 1:nrow(vote1000)) {
  for (j in 1:10) {
    var1[i, j] <- (vote1000[i, 6 + j] / 100) * (j - vote1000[i, 5]) ^ 2
  }
  vote1000$var[i] <- sum(var1[i, ])
}

vote1000$sd <- sqrt(vote1000$var)

================================================================================

#Collect all of the full length films (Take short movies out of the
#original data set)

full_length <- movies[movies$Short == 0, ]

================================================================================

#Want to know which rating quantile each movie falls into

ptile <- quantile(vote1000$rating, seq(0, 1, .125))
vote1000$Q <- NA

for (i in 1:nrow(vote1000)) {
  for (j in 2:9) {
    if (ptile[j - 1] <= vote1000[i, 5] & vote1000[i, 5] <= ptile[j]) {
      vote1000[i, 27] <- j - 1
    }
```

Or use cut(vote1000$rating, ptile)

```
  }
}
```

```
================================================================================
```

```
#Create a variable for the number of characters in the title of the movie

vote1000 <- transform(vote1000, tlength = str_length(vote1000$title))
```

```
================================================================================
```

```
#Add variable indicating if movie is listed under only one genre

vote1000$one_genre <- NA
for (i in 1:nrow(vote1000)) {
  vote1000$one_genre[i] <- sum(vote1000[i, 18:24]) == 1
}

#Collect the one genre movies into a single data frame

one <- vote1000[vote1000$one_genre == TRUE, ]

#Create a 'genre' variable in the new data set to give the genre of the movie
#This will allow for faceting by genre

one$genre <- NA
for (i in 1:nrow(one)) {
  if (one[i, 18] == 1) {
    one[i, 32] <- 'Action'
  } else if (one[i, 19] == 1) {
    one[i, 32] <- 'Animation'
  } else if (one[i, 20] == 1) {
    one[i, 32] <- 'Comedy'
  } else if (one[i, 21] == 1) {
    one[i, 32] <- 'Drama'
  } else if (one[i, 22] == 1) {
    one[i, 32] <- 'Documentary'
  } else {
    one[i, 32] <- 'Romance'
  }
}
```

```
================================================================================
```

```
#Distinguish between movies produced after IMDB was created and those produced
#before. (IMDB was started in 1990)

vote1000 <- transform(vote1000, IMDB = vote1000$year >= 1990)
```

did you use this variable?

```
#Create a variable to indicate the decade the movie was produced in.
#This will allow us to have some sort of facet to use with year.

vote1000 <- ddply(vote1000, 'year', transform,
  decade = 10 * floor(year / 10), .progress = 'text')

#Collect movies produced during the Great Depression

GD <- vote1000[vote1000$year >= 1929 & vote1000$year <= 1940, ]

#Collect movies produced during World War II

WWII <- vote1000[vote1000$year >= 1939 & vote1000$year <= 1945, ]

#Determine if a movie was rated in the 8th quantile for movies produced in the
#same year

vote1000$year_Q <- NA
vote1000 <- ddply(vote1000, 'year', transform,
  year_Q8 = rating >= quantile(rating, .875))

#Want a collection of ratios of average budget for 8th quantile movies to
#average budget for all movies in each year

bud_rat <- rep(NA, 2010-1913, 1)
for (i in 1:2010 - 1913) {
  bud_rat[i] <- mean(vote1000[vote1000$year == i + 1913 &
  vote1000$year_Q8 == TRUE, ]$budget,
  na.rm = TRUE) / mean(vote1000[vote1000$year == i + 1913, ]$budget,
  na.rm = TRUE)
}

#Want to the same thing with length as with budget

len_rat <- rep(NA, 2010-1913, 1)
```

```
for (i in 1:2010 - 1913) {
  len_rat[i] <- mean(vote1000[vote1000$year == i + 1913 &
  vote1000$year_Q8 == TRUE, ]$length,
  na.rm = TRUE) / mean(vote1000[vote1000$year == i + 1913, ]$length,
  na.rm = TRUE)
}


================================================================================
--------------------------------------------------------------------------------
================================================================================

#Final Plots

#Want to compare our subset of the data to the original data set
#(without short films)

#Figure 1

#Want to look at the rating distribution of the original data and the subset

qplot(rating, ..density.., data = full_length, binwidth = .2,
  geom = 'freqpoly', main = 'Ratings of All Full Length Movies')
ggsave('ratings-full_length.pdf')

qplot(rating, ..density.., data = vote1000, binwidth = .2,
  geom = 'freqpoly',
  main = 'Ratings of Full Length Movies With Over 1000 Votes')
ggsave('ratings-vote1000.pdf')

#Now want to observe distributions of the movies per year for both data sets

qplot(year, ..density.., data = vote1000, geom = 'freqpoly', binwidth = 1,
  main = 'Movies Per Year in Our Subset of the Data')
ggsave('Movies-Year-Subset.pdf')

qplot(year, ..density.., data = full_length, geom = 'freqpoly', binwidth = 1,
  main = 'Movies Per Year in the Original Data Set')
ggsave('Movies-Year-Original.pdf')

#Table 1

#Want a table with the five number summaries of each of these plots

summ_v <- summary(vote1000$rating)
summ_m <- summary(movies$rating)
rating_sum <- matrix(c(summ_v, summ_m), 2, 6, byrow = TRUE)
colnames(rating_sum) <- c("Min.", "1st Qu.", "Median", "Mean", "3rd Qu.",
```

```
  "Max.")
rownames(rating_sum) <- c("Our Subset", "All Movies")
xtable(rating_sum)

#Figure 2

#How are the rating and standard deviation related?

qplot(rating, sd, data = vote1000, main = 'Ratings vs Standard Deviation')
ggsave('Rating-Sd.png')

#Figure 4

#How have ratings for movies changed over the decades?

qplot(rating, ..density.., data = vote1000, binwidth = .2,
  geom = 'freqpoly', main = 'Ratings of Movies By Decade') +
  facet_wrap(~ decade)
ggsave('rating_by_decade.pdf')

#Figure 5

#How did the Great Depression affect the movie industry?

qplot(rating, data = GD, binwidth = 0.1,
  main = 'Ratings of Movies During the Great Depression')
ggsave('rating-gd.pdf')

#Figure 6

#What about World War II?

qplot(rating, data = WWII, binwidth = 0.1,
  main = 'Ratings of Movies During World War II')
ggsave('rating-ww2.pdf')

#Figure 7

#Which genres get higher ratings?

qplot(reorder(genre, rating, median), rating, data = one, geom = 'boxplot')
ggsave('rating-across-genre.pdf')

#Figure 8

#Do highly rated movies tend to have higher budgets?

qplot(1914:2010, log(bud_rat),
```

```
  main = "Ratio of Highly Rated Movies' Budgets to Average Budgets")
ggsave('ratio-high-avg-budget.pdf')

mean(bud_rat, na.rm = T)

#Figure 9

#Another possible way to answer the previous question

qplot(log(budget), data = vote1000) + facet_wrap(~Q)
ggsave('budget-quantiles.pdf')

#Table 10

#Display the mean budget for each quantile of rating

bud_means <- mat.or.vec(8, 2)
for (i in 1:8) {
  bud_means[i, 1] <- i
  bud_means[i, 2] <- mean(vote1000[vote1000$Q == i, ]$budget, na.rm = T)
}
colnames(bud_means) <- c("Quantile", "Average Budget")
bud_means <- data.frame(bud_means)
xtable(bud_means)

#Want to do the same analysis in Figures 9 and 10 and Table 2 with Length

#Figure 11

#First the Ratios:

qplot(1914:2010, log(len_rat),
  main = "Ratio of Highly Rated Movies' Lengths to Average Lengths")
ggsave('ratio-high-avg-length.pdf')

mean(len_rat, na.rm = T)
mean(len_rat[26:length(len_rat)], na.rm = T)
median(len_rat, na.rm = T)

#Figure 12

#Now the distributions at each quantile

qplot(length, data = vote1000, binwidth = 5,
  main = "Length of Movies for Each Rating Quantile") +
  facet_wrap(~ Q) +
  xlim(0, 300)
ggsave('Length-Rating-Quantile.pdf')
```

```
#Table 3

#Show the means at each quantile

len_means <- mat.or.vec(8, 2)
for (i in 1:8) {
  len_means[i, 1] <- i
  len_means[i, 2] <- mean(vote1000[vote1000$Q == i, ]$length, na.rm = T)
}
colnames(len_means) <- c("Quantile", "Average Length")
len_means <- data.frame(len_means)
xtable(len_means)
```