Statistical Analysis of the Movies Dataset

Initial Subsetting and Analysis

I began by subsetting the older movie set (called train_movies), consisting of films produced from 1927 through 2014, into another set called won_movies, indicating those movies that had received Oscar nominations and thus allowing for a closer analysis of the characteristics of Oscar nominated movies in order to compare them to the set as a whole. As many of the variables were either character variables or binary variables (i.e., the genre variables), the str and summary functions for the sets revealed only scant information concerning the ratings, duration and nominations variables. The maximum number of nominations received by a film was 14. Ratings and duration will be covered in further detail below. Other variables analyzed were the films' language and country of production, their initial budget and amount ultimately grossed at the box office, the varieties and frequencies of film genres and the frequencies of particular directors and actors repersented in the datasets.

Language and Country

I ran tables contrasting language and country for the train_movies, won_movies and new_movies sets, using table(won_movies$language, won_movies$country), for example:

```
##
##           China France Germany UK Unknown USA
##   English     1      1       2 12      18  53
```

Both the older and newer films were primarily English, even where they were made in other countries. In the case of the Oscar nominated films, as illustrated above, all of the movies were in English.

Duration and Ratings

I then ran the mean, median, range, variance (var) and standard deviation (sd) for the durations and ratings of the older, newer and Oscar nominated films. Ratings as used here refer to IMDB ratings given on a scale from 1 (being the worst) to 10 (being the best).

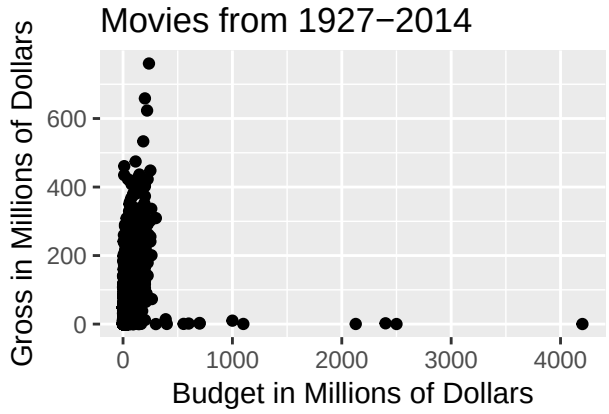
The results for duration indicated that the average length of Oscar-nominated movies was longer than the length of other movies in general (approximately 140 minutes for Oscar-nominated movies compared to 108 and 107 minutes for older and newer movies, respectively.) There was also a greater variance in won_movies (approximately 1303), but a lower maximum at 240 minutes compared to 330 minutes in train_movies. However, this 330 minute length in train_movies appears to be an outlier, since the film occupying the 99th percentile in duration had a length of a more modest 188 minutes. The newer movies had the least variance (at approximately 318) and the lowest average length.

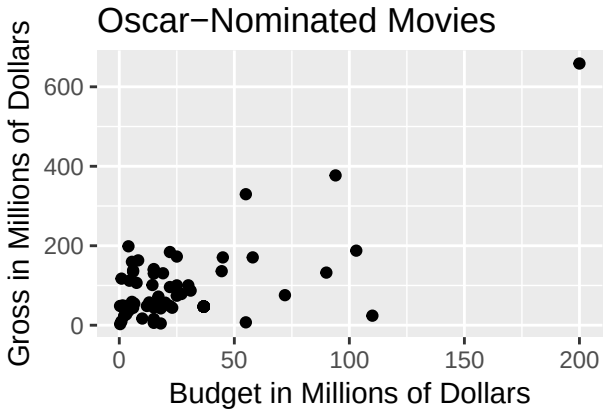
Predictably, the results for rating indicated that the Oscar-nominated movies were rated higher than the general population or the newer movies, the average and median ratings being aproximately 7.8 and 8 respectively compared to 6.4 and 6.6 for the older movies and 6.1 and 6.3 for the newer movies. The Oscar-nominated movies also had lower variances than the other sets. (0.43 compared to 1.22 and 1.58 for train_movies and new_movies.) The maxima for all sets was at or near 9.2, but the minima varied. Train_movies and new_movies had lows of 1.6 and 2.2 respectively, while won_movies had a much greater minimum of 5.8. The smaller variance in won_movies is therefore likely due to those films having been rated consistently high. (The averages for the general population of older movies and newer movies, both here and for duration were substantially the same, with the slightly greater ratings and durations for the train_movies set likely because this set also contains the Oscar-nominated movies.)

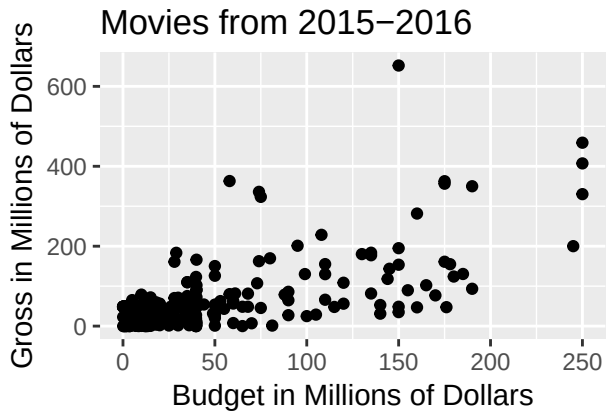
Budget and Gross

The Oscar-nominated movies and the newer movies both displayed a fairly positive correlation between a film's budget and the amount it ultimately grossed, with correlations of 0.63 and 0.64, respectively. Thus, the greater the budget for a film the more likely that such films would gross a large amount at the box office.

On the other hand, the general population of older movies showed a much lower correlation between these two variables at only 0.20. In this instance, it did not appear to matter much what a film's initial budget was; most films appeared to gross much more, while movies with very large budgets (over a billion dollars) grossed very little. Perhaps this was due to older filmmakers knowing how to make a quality film for very little money, stretching their dollar as much as possible, while the extremely large budgets were likely indicative of poor planning and organization which would have contributed to the films' overall commercial failure. The relationships between budget and gross scaled in millions of dollars for all three sets are displayed below.



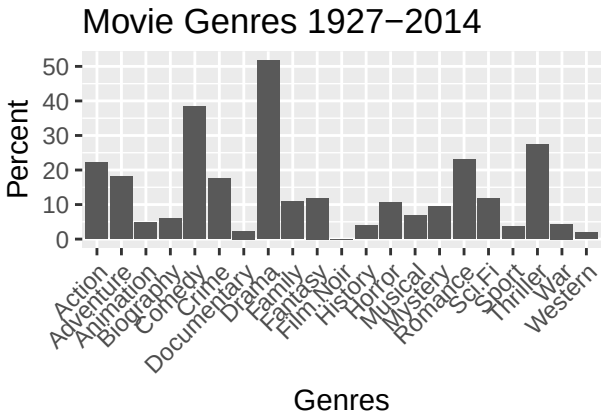


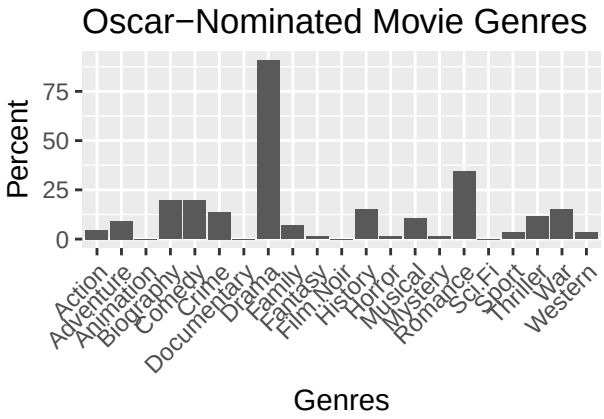


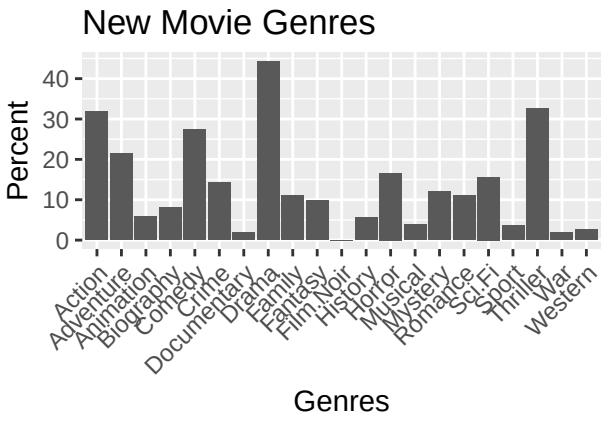
Genres

There were 21 total genres that were counted in each dataset: Action, Adventure, Biography, Comedy, Crime, Drama, Musical, Western, Family, Film Noir, History, Mystery, Romance, Sport, War, Fantasy, Sci Fi, Thriller, Documentary, Horror and Animation. It should be noted, however, that the Film Noir category was not actually represented in the data, despite its being given its own genre category by the creator of the dataset. It may be that those movies falling into this genre were subsumed into other genres in the course of the dataset's creation and the category was never removed.

The most common genre in every case was Drama. The least common genre(s), aside from Film Noir, was Western for the general movie population from 1927-2014, Sci Fi, Documentary, Animation, Fantasy, Horror and Mystery for the Oscar-nominated movies and Documentary for the newer movies. Plots showing the relative percentages of each genre in greater detail are below.







As will be observed, the Oscar-nominated movies are dominated primarily by Drama and Romance movies. This is also the case for the movies from 1927-2014, but with greater representation from Comedies, Thrillers and Action and Adventure films. The influence of changing public tastes may be seen in the New Movies, where Thrillers, Action, Adventure, Sci Fi and Horror (less conventionally "dramatic" genres) are more common than in prior years. Additionally, Animation is a relatively new category at the Academy Awards, thus accounting for its absence from the Oscar-nominated dataset.

Directors and Actors

I then analyzed each of the datasets for the most frequently occuring directors and actors. In the general dataset for Movies from 1927-2014 I created a subset of only movie names and director names called director1 and from there a list called dir_count1 using count(director1, director1$director_name). I then obtained the director name with the maximum count using dir_count1$‘director1$director_name‘[dir_count1$n == max(dir_count1$n)] and the directors occupying above the 99th percentile in frequency using dir_count1$‘director1$director_name‘[dir_count1$n > quantile(dir_count1$n, .99)]. This revealed Steven Spielberg as being the most popular director here, while 16 other directors occupied above the 99th percentile. These other top directors were Barry Levinson, Brian De Palma, Clint Eastwood, Joel Schumacher, Martin Scorsese, Oliver Stone, Renny Harlin, Ridley Scott, Robert Rodriguez, Robert Zemeckis, Ron Howard, Spike Lee, Steven Soderbergh, Tim Burton, Tony Scott and Woody Allen.

Using a similar procedure for the Oscar-nominated movies and newer movies, I found that the top Oscar-nominated directors (all tied for the maximum) were Billy Wilder, Clint Eastwood, David Lean, Elia Kazan, Francis Ford Coppola, Frank Capra, Fred Zinnemann and Milos Forman. Tied for maximum frequency among the directors of the newer movies were David O. Russell, James Wan, Jaume Collet-Serra, Patricia Riggen, Robert Schwentke, Roland Emmerich and, once again, Steven Spielberg. It thus appears that Spielberg is popular among audiences new and old, but less so among those who decide Motion Picture Academy Award nominations (although Spielberg did direct one Oscar-nominated movie in the dataset, "Schindler's List").

In order to analyze the frequency of the actors, I had to stack the three separate actor columns in each dataset (actor_1_name, actor_2_name and actor_3_name) to form a single column outside the datasets using for the first dataset, for example, actors1 <- stack(list(act1 = train_movies$actor_1_name, act2 = train_movies$actor_2_name, act3 = train_movies$actor_3_name)) and removing those entries that were unknown using actors1 <- subset(actors1, actors1$values != "Unknown"). I then used the same procedures as above with the directors in order to obtain the actors with the maximum counts and the actors occupying above the 99th percentile in frequency. (The exception was the Oscar-nominated actors, where I used the 90th percentile since percentiles above the 90th percentile were the same as the maximum).

The most common actor for the movies from 1927-2014 was Robert De Niro. 54 other actors occupied greater than the 99th percentile, appearing in more than 18 films per person. The most common actor among the Oscar-nominated actors was Morgan Freeman, although Robert De Niro is in the 90th percentile of top actors. The other top (90th percentile) Oscar-nominated actors were Al Pacino, Anthony Hopkins, Beth Grant, Claude Rains, Clint Eastwood, Colin Firth, Jack Hawkins, John Gielgud, Judd Hirsch, Karl Malden, Kate Winslet, Leonardo DiCaprio, Marlon Brando, Meryl Streep, Oliver Reed, Ray Walston, Robert Duvall, Robert Shaw, Scoot McNairy and Susannah York. Finally, the most frequent actors appearing in the new movies were Chris Hemsworth and, once again, Robert De Niro, while the other actors appearing above the 99th percentile in frequency were Bradley Cooper, Johnny Depp, Scarlett Johansson and Tom Hardy.

It should be noted here that the compiler of these datasets did not always include the most important actors in a given film, for example leaving out Arnold Schwarzenegger and Sharon Stone from "Total Recall". This may account for the presence of actors here such as "Scoot McNairy" in the Oscar-nominated dataset, a lesser known actor who did indeed appear in the Oscar-nominated movies "12 Years a Slave" and "Argo". Thus, the actors variables may have limited predictive utility for those films that included lesser-known actors at the expense of omitting the main actors, but this did not occur in every case.