Movies Data Story

Each year, Hollywood releases vast numbers of movies to diverse audiences all over the world. Of these scores of films, only a scant few are chosen to win an Academy Award, and fewer still win the Award for Best Picture. How might it be possible to predict in advance whether a film will win an Academy Award? Is it possible to assess the quality of a movie based on its relative probability of winning an Academy Award?

Please note that all code used to generate the graphs and output displayed herein is available for review in Appendix B under the appropriate headers.

Initial Preparations

In order to answer these questions, I began by obtaining two datasets containing both general movie data and data indicating whether films were nominated for academy awards. All of these films had at least won an Academy Award for Best Pciture. The general movie dataset contains imdb data obtained from Kaggle at https://www.kaggle.com/antoniobap/imdb-v3/data. The academy award data was obtained from the University of Waterloo at https://cs.uwaterloo.ca/~s255khan/oscars.html.

Before loading the datasets, I loaded the dplyr, tidyr and dummies libraries.

Data Cleaning

Preliminary cleaning procedures

Intending to merge these two datasets into one, I began by changing the names of the columns in the imdb dataset (hereinafter, imdb) to match those in the academy awards dataset (hereinafter, acad_awd). I renamed movie_title, title_year and imdb_score to be name, year and rating, respectively.

I then trimmed all leading and trailing whitespace from the names in both datasets and removed trailing question marks from the names in the imdb set. I also changed the names in both datasets to uppercase for consistency. I changed one title in imdb, Birdman, which used the film's extended title to match the shorter title used in acad_awd. There were two dashes in the nominations column[1] in the acad_awd set that I changed to 1's, signifying that those films had at least won an Oscar for Best Picture.

The na's in certain key columns in imdb, budget, gross and duration, had values imputed to them based on the means of their respective columns as a whole. (The mean for duration was rounded to ensure an integer value.)

A number of duplicated names were observed in the imdb set as well, and these were removed using imdb <- distinct(imdb, name, .keep_all = TRUE). I then subsetted the imdb data for durations greater than 60 in order to eliminate the television shows that were present in the data, and then removed blank names.

The Merge Process

The merge process proved to be fairly complex owing to the different columns in each dataset. In order to prevent loss of data from the acad_awd set, the merge was accomplished in two distinct stages.

First, a merge_acad_awd set was created, dropping all columns except the name and nominations columns. Following this, I created a merged dataset called movies by performing a full join of merge_acad_awd and

[1]The original variable name "nominations" refers to the total number of nominations received by films that won a Best Picture Academy Award. All films listed in the acad_awd dataset were winners of a Best Picture Oscar.

imdb by name (movies <- full_join(merge_acad_awd, imdb, by = "name")). This merge was effective except for those movies that were in the acad_awd set, but not in the imdb set. All data except for their names and nominations were effectively lost in the merge. As there were 17 such films (out of a total of 87 acad_awd films), this proved to be an unacceptable loss.

Accordingly, a second separate merge of the 17 films into movies was necessary in order to at least preserve duration, ratings and genre-related data. In order to make the genre-related data in acad_awd for these films compatible with the movies set, I created binaries from the genre1 and genre2 columns using the dummies library. These binaries had to first be extracted into 2 separate dataframes, binary1 and binary2, in order to prevent overwriting in situations where a film was both a comedy and a drama, for example. The two binary sets were then bound together using cbind and the column of blanks, V1, was removed. The combination of the two sets in this manner created duplicate column names such that, for example, Biography could be added to Biography.1 (as in binaries$Biography <- binaries$Biography + binaries$Biography.1) so that any ones would cancel out any zeroes in order to allow for the presence of two genres simultaneously. The duplicate columns were then removed. I then attached this final set of binaries to acad_awd using cbind and called the resulting dataset merge2_acad_awd. I then dropped all extraneous columns from merge2 (genre1, genre2, release, metacritic and synopsis) and limited this dataset to only those films that did not appear in imdb (using merge2_acad_awd <- merge2_acad_awd[which(merge2_acad_awd$name %!in% imdb$name),] and a custom definition, '%!in%' = Negate('%in%')). I then performed the second merge by doing a full join of merge2 with movies, initially into a test dataset called movies2. After examining movies2, I overwrote movies with movies2. This final merge proved to be successful as the maximum amount of data appeared to be preserved.

Final preparations

Following the final merge, I created two subset dataframes based on the age of the films. The first set consisted of all movies between 1927 and 2014 inclusive (the years common to both of the original datasets). Any films in this set having NA's in the nominations column were changed to 0 to signify that they did not win a Best Picture Oscar. The second set consisted only of newer movies produced from 2015-2016, and naturally had no nominations data. This set could be used for a comparison to the older movies to show how trends in filmmaking, such as budgeting and the types of movie genres commonly produced, may have changed over time. Both sets then required a small amount of additional cleaning to remove additional NA's and irrelevant columns (due to lack of relevancy to older films, such as Facebook likes, for example), after which they were ready for analysis.

Data Analysis

Initial Subsetting and Analysis

I began by subsetting the older movie set, consisting of films produced from 1927 through 2014, into another set called won_movies, indicating those movies that had won a Best Picture Academy Award and received other Oscar nominations, thus allowing for a closer analysis of the characteristics of Oscar-winning 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 older movies, won_movies and the 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-winning films, as illustrated above, all of the movies were in English, typically in either the US or the UK[2].

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-winning 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-winning movies was longer than the length of other movies in general (approximately 140 minutes for Oscar-winning 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 the older movies. However, this 330 minute length in the older movies set 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-winning movies were rated higher than the general population or the newer movies, the average and median ratings being approximately 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-winning movies also had lower variances than the other sets. (0.43 compared to 1.22 and 1.58 for the older movies and the newer movies.) The maxima for all sets was at or near 9.2, but the minima varied. The older movies and the newer 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 older movies set likely because this set also contains the Oscar-winning movies.)

Thus it appears that Oscar-winning films typically have average durations of two to two and a half hours in length with high imdb ratings close to 8 out of 10.

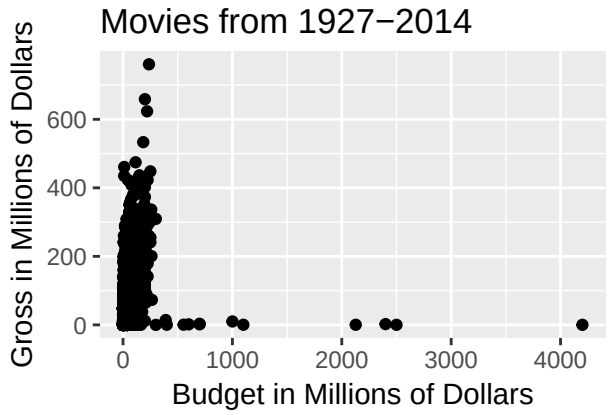
Budget and Gross

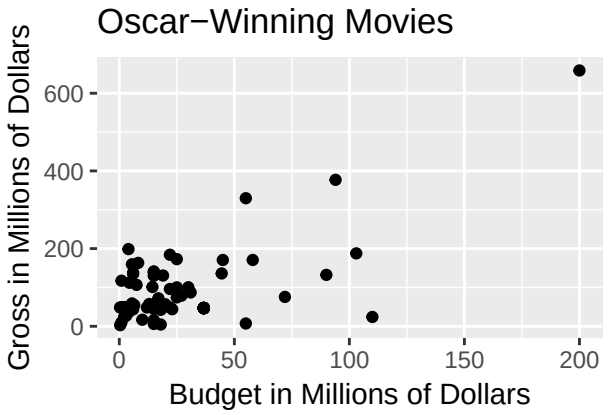
The Oscar-winning movies and the newer movies both displayed a fairly positive correlation between a film's budget and the amount it ultimately grossed, with correlations of 0.63 and 0.64, respectively. Thus, the greater the budget for a film the more likely that such films would gross a large amount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

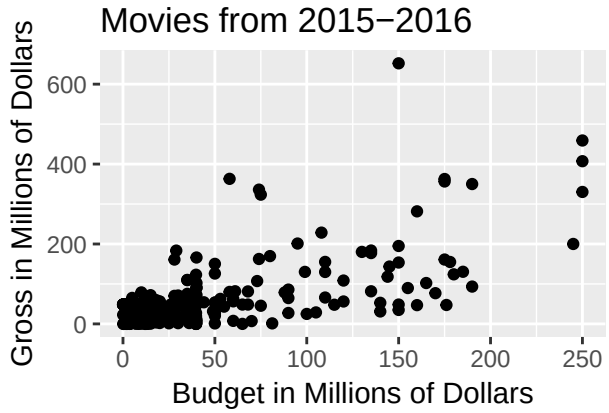
On the other hand, the general population of older movies showed a much lower correlation between these two variables at only 0.20. In this instance, it did not appear to matter much what a film's initial budget was; most films appeared to gross much more, while movies with very large budgets (over a 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. However, it

[2]The more extensive tables for the older and newer movies are available in Appendix A, where the first table corresponds to the older films, inclusive of the films from won_movies, and the second set contains the newer films.

should also be noted that the financial figures in the older movies dataset were not adjusted for inflation, and therefore likely also contributed the lower correlation, and thus, to the form of the scatterplot for the older movies shown here. The relationships between budget and gross scaled in millions of dollars for all three sets are displayed below.





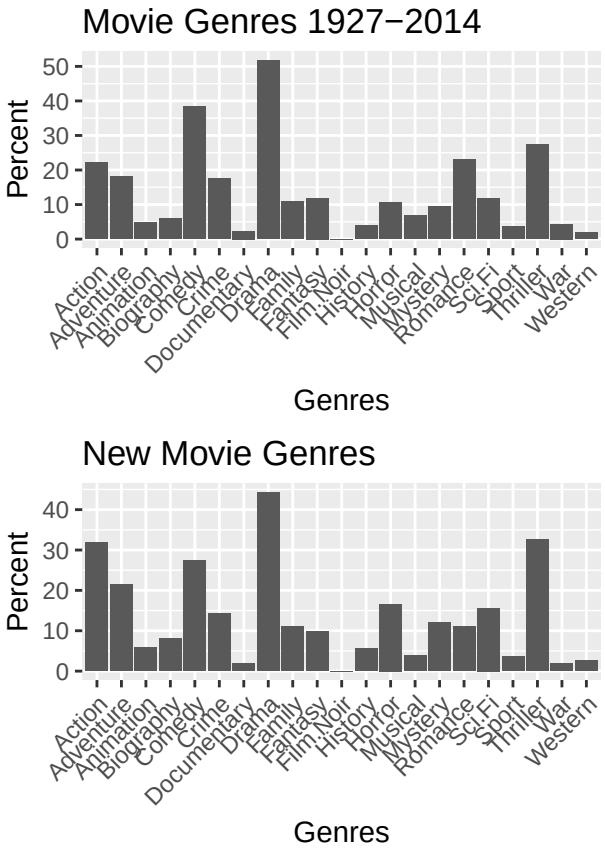


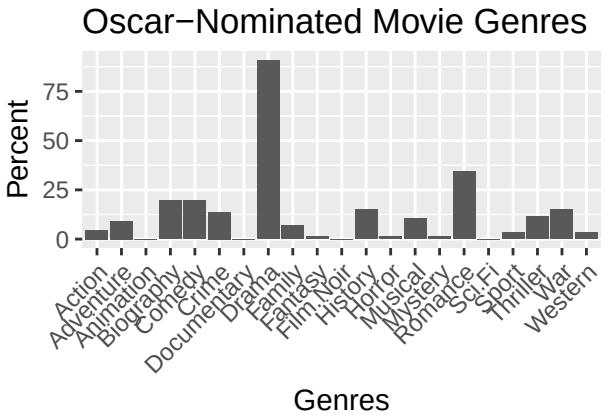
Therefore, it appears that Oscar-winning films typically have budgets not much greater than $100 million and gross not much more than $200 million.

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-winning 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-winning 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-winning movies and newer movies, I found that the top directors for the Oscar-winning movies (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 the Best Picture

Oscar winners at the Motion Picture Academy (although Spielberg did direct one Oscar-winning 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, actors2 <- stack(list(act1 = won_movies$actor_1_name, act2 = won_movies$actor_2_name, act3 = won_movies$actor_3_name)) and removing those entries that were unknown using actors2 <- subset(actors2, actors2$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-winning 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 actors in the Oscar-winning films was Morgan Freeman, although Robert De Niro is in the 90th percentile of top actors. The other top (90th percentile) actors in the Oscar-winning movies 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-winning dataset, a lesser known actor who did indeed appear in the Oscar-winning 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, although this did not occur in every case.

Concluding Remarks

Based on my examination of the characteristics of the films in the movie datasets, it appears that a number of factors characterise successful potentially Oscar-winning movies. Drama films made in English in either the US or UK that are moderately long (approximately two to two and a half hours in length), with high imdb ratings (nearly 8 out of 10) and relatively large budgets (but not much greater than $100 million) appear to be films that are most likely to win Academy Awards for Best Picture.

While the caliber of directors and actors involved in a film's production may be persuasive evidence that a movie will turn out to be successful, it is not determinative. For example, while Steven Spielberg was involved in the production of an Oscar-winning film, he was also involved in many that were not. Likewise, Morgan Freeman (the most frequently appearing actor among the Oscar-winning films) acted not only in Oscar-winning films such as "Unforgiven", but also in less successful films such as "Evan Almighty" and "Ted 2".

Accordingly, I would use variables such as duration, language, country, rating, budget, gross and the genre variables as predictor variables when attempting to make predictions as to whether a film is likely to win an Academy Award for Best Picture, rather than looking to the director and/or actors involved in the film's production. Such variables are likely to also prove useful in assessing the overall quality of a film, assuming that the subjective nature of "quality" may be quantitatively defined as, for example, the number of other awards that a film won or the number of other awards for which they were nominated. These ideas will be further put to the test in my Regression Analysis report.

Appendix A

```
##
##              Afghanistan Argentina Aruba Australia Bahamas Belgium Brazil
##   Aboriginal           0         0     0         1       0       0      0
##   Arabic               0         0     0         0       0       0      0
##   Aramaic              0         0     0         0       0       0      0
##   Bosnian              0         0     0         0       0       0      0
##   Cantonese            0         0     0         0       0       0      0
##   Chinese              0         0     0         0       0       0      0
##   Czech                0         0     0         0       0       0      0
##   Danish               0         0     0         0       0       0      0
##   Dari                 1         0     0         0       0       0      0
##   Dutch                0         0     0         0       0       0      0
##   Dzongkha             0         0     0         1       0       0      0
##   English              0         0     1        47       1       2      1
##   Filipino             0         0     0         0       0       0      0
##   French               0         0     0         0       0       0      0
##   German               0         0     0         0       0       0      0
##   Greek                0         0     0         0       0       0      0
##   Hebrew               0         0     0         0       0       0      0
##   Hindi                0         0     0         0       0       0      0
##   Hungarian            0         0     0         0       0       0      0
##   Icelandic            0         0     0         0       0       0      0
##   Indonesian           0         0     0         0       0       0      0
##   Italian              0         0     0         0       0       0      0
##   Japanese             0         0     0         0       0       0      0
##   Kannada              0         0     0         0       0       0      0
##   Kazakh               0         0     0         0       0       0      0
##   Korean               0         0     0         0       0       0      0
##   Mandarin             0         0     0         0       0       0      0
##   Maya                 0         0     0         0       0       0      0
##   Mongolian            0         0     0         0       0       0      0
##   None                 0         0     0         0       0       0      0
##   Norwegian            0         0     0         0       0       0      0
##   Persian              0         0     0         0       0       0      0
##   Polish               0         0     0         0       0       0      0
##   Portuguese           0         0     0         0       0       0      6
##   Romanian             0         0     0         0       0       0      0
##   Russian              0         0     0         0       0       0      0
##   Spanish              0         4     0         0       0       0      0
##   Swedish              0         0     0         0       0       0      0
##   Thai                 0         0     0         0       0       0      0
##   Vietnamese           0         0     0         0       0       0      0
##   Zulu                 0         0     0         0       0       0      0
##
##              Bulgaria Cambodia Cameroon Canada China Colombia
##   Aboriginal        0        0        0      0     0        0
##   Arabic            0        0        0      0     0        0
##   Aramaic           0        0        0      0     0        0
##   Bosnian           0        0        0      0     0        0
##   Cantonese         0        0        0      0     1        0
##   Chinese           0        0        0      0     2        0
##   Czech             0        0        0      0     0        0
```

```
##   Danish            0        0        0      0     0        0
##   Dari              0        0        0      0     0        0
##   Dutch             0        0        0      0     0        0
##   Dzongkha          0        0        0      0     0        0
##   English           1        1        1    105     3        0
##   Filipino          0        0        0      0     0        0
##   French            0        0        0      6     0        0
##   German            0        0        0      0     0        0
##   Greek             0        0        0      0     0        0
##   Hebrew            0        0        0      0     0        0
##   Hindi             0        0        0      1     0        0
##   Hungarian         0        0        0      0     0        0
##   Icelandic         0        0        0      0     0        0
##   Indonesian        0        0        0      0     0        0
##   Italian           0        0        0      0     0        0
##   Japanese          0        0        0      0     0        0
##   Kannada           0        0        0      0     0        0
##   Kazakh            0        0        0      0     0        0
##   Korean            0        0        0      0     0        0
##   Mandarin          0        0        0      0    18        0
##   Maya              0        0        0      0     0        0
##   Mongolian         0        0        0      0     0        0
##   None              0        0        0      1     0        0
##   Norwegian         0        0        0      0     0        0
##   Persian           0        0        0      0     0        0
##   Polish            0        0        0      0     0        0
##   Portuguese        0        0        0      0     0        0
##   Romanian          0        0        0      0     0        0
##   Russian           0        0        0      0     0        0
##   Spanish           0        0        0      0     0        1
##   Swedish           0        0        0      0     0        0
##   Thai              0        0        0      0     0        0
##   Vietnamese        0        0        0      0     0        0
##   Zulu              0        0        0      0     0        0
##
##              Czech Republic Denmark Dominican Republic Egypt Finland
##   Aboriginal              0       0                   0     0       0
##   Arabic                  0       0                   0     1       0
##   Aramaic                 0       0                   0     0       0
##   Bosnian                 0       0                   0     0       0
##   Cantonese               0       0                   0     0       0
##   Chinese                 0       0                   0     0       0
##   Czech                   1       0                   0     0       0
##   Danish                  0       4                   0     0       0
##   Dari                    0       0                   0     0       0
##   Dutch                   0       0                   0     0       0
##   Dzongkha                0       0                   0     0       0
##   English                 1       6                   0     0       0
##   Filipino                0       0                   0     0       0
##   French                  0       0                   0     0       1
##   German                  0       0                   0     0       0
##   Greek                   0       0                   0     0       0
##   Hebrew                  0       0                   0     0       0
##   Hindi                   0       0                   0     0       0
```

```
##   Hungarian              0       0                 0     0       0
##   Icelandic              0       0                 0     0       0
##   Indonesian             0       0                 0     0       0
##   Italian                0       0                 0     0       0
##   Japanese               0       0                 0     0       0
##   Kannada                0       0                 0     0       0
##   Kazakh                 0       0                 0     0       0
##   Korean                 0       0                 0     0       0
##   Mandarin               0       0                 0     0       0
##   Maya                   0       0                 0     0       0
##   Mongolian              0       0                 0     0       0
##   None                   0       0                 0     0       0
##   Norwegian              0       0                 0     0       0
##   Persian                0       0                 0     0       0
##   Polish                 0       0                 0     0       0
##   Portuguese             0       0                 0     0       0
##   Romanian               0       0                 0     0       0
##   Russian                0       0                 0     0       0
##   Spanish                0       0                 1     0       0
##   Swedish                0       0                 0     0       0
##   Thai                   0       0                 0     0       0
##   Vietnamese             0       0                 0     0       0
##   Zulu                   0       0                 0     0       0
##
##              France Georgia Germany Greece Hong Kong Hungary Iceland India
##   Aboriginal      0       0       0      0         0       0       0     0
##   Arabic          1       0       1      0         0       0       0     0
##   Aramaic         0       0       0      0         0       0       0     0
##   Bosnian         0       0       0      0         0       0       0     0
##   Cantonese       0       0       0      0         8       0       0     0
##   Chinese         0       0       0      0         0       0       0     0
##   Czech           0       0       0      0         0       0       0     0
##   Danish          0       0       0      0         0       0       0     0
##   Dari            0       0       0      0         0       0       0     0
##   Dutch           0       0       0      0         0       0       0     0
##   Dzongkha        0       0       0      0         0       0       0     0
##   English        80       1      77      0         5       1       1     5
##   Filipino        0       0       0      0         0       0       0     0
##   French         60       0       0      0         0       0       0     0
##   German          0       0      14      0         0       0       0     0
##   Greek           0       0       0      1         0       0       0     0
##   Hebrew          0       0       0      0         0       0       0     0
##   Hindi           0       0       0      0         0       0       0    25
##   Hungarian       0       0       0      0         0       1       0     0
##   Icelandic       0       0       0      0         0       0       1     0
##   Indonesian      0       0       0      0         0       0       0     0
##   Italian         0       0       0      0         0       0       0     0
##   Japanese        0       0       0      0         0       0       0     0
##   Kannada         0       0       0      0         0       0       0     1
##   Kazakh          1       0       0      0         0       0       0     0
##   Korean          0       0       0      0         0       0       0     0
##   Mandarin        0       0       0      0         3       0       0     0
##   Maya            0       0       0      0         0       0       0     0
##   Mongolian       0       0       0      0         0       0       0     0
```

```
##   None            0       0       0       0         0       0       0     0
##   Norwegian       0       0       0       0         0       0       0     0
##   Persian         1       0       0       0         0       0       0     0
##   Polish          0       0       0       0         0       0       0     0
##   Portuguese      0       0       0       0         0       0       0     0
##   Romanian        0       0       0       0         0       0       0     0
##   Russian         0       0       0       0         0       0       0     0
##   Spanish         0       0       0       0         0       0       0     0
##   Swedish         0       0       0       0         0       0       0     0
##   Thai            0       0       0       0         0       0       0     0
##   Vietnamese      0       0       0       0         0       0       0     0
##   Zulu            0       0       0       0         0       0       0     0
##
##              Indonesia Iran Ireland Israel Italy Japan Kyrgyzstan Libya
##   Aboriginal         0    0       0      0     0     0          0     0
##   Arabic             0    0       0      0     0     0          0     0
##   Aramaic            0    0       0      0     0     0          0     0
##   Bosnian            0    0       0      0     0     0          0     0
##   Cantonese          0    0       0      0     0     0          0     0
##   Chinese            0    0       0      0     0     0          0     0
##   Czech              0    0       0      0     0     0          0     0
##   Danish             0    0       0      0     0     0          0     0
##   Dari               0    0       0      0     0     0          0     0
##   Dutch              0    0       0      0     0     0          0     0
##   Dzongkha           0    0       0      0     0     0          0     0
##   English            0    1      11      0    12     4          1     1
##   Filipino           0    0       0      0     0     0          0     0
##   French             0    0       0      0     0     0          0     0
##   German             0    0       0      0     0     0          0     0
##   Greek              0    0       0      0     0     0          0     0
##   Hebrew             0    0       0      3     0     0          0     0
##   Hindi              0    0       0      0     0     0          0     0
##   Hungarian          0    0       0      0     0     0          0     0
##   Icelandic          0    0       0      0     0     0          0     0
##   Indonesian         1    0       0      0     0     0          0     0
##   Italian            0    0       0      0     9     0          0     0
##   Japanese           0    0       0      0     0    14          0     0
##   Kannada            0    0       0      0     0     0          0     0
##   Kazakh             0    0       0      0     0     0          0     0
##   Korean             0    0       0      0     0     0          0     0
##   Mandarin           0    0       0      0     0     0          0     0
##   Maya               0    0       0      0     0     0          0     0
##   Mongolian          0    0       0      0     0     0          0     0
##   None               0    0       0      0     0     0          0     0
##   Norwegian          0    0       0      0     0     0          0     0
##   Persian            0    3       0      0     0     0          0     0
##   Polish             0    0       0      0     0     0          0     0
##   Portuguese         0    0       0      0     0     0          0     0
##   Romanian           0    0       0      0     0     0          0     0
##   Russian            0    0       0      0     0     0          0     0
##   Spanish            0    0       0      0     0     0          0     0
##   Swedish            0    0       0      0     0     0          0     0
##   Thai               0    0       0      0     0     0          0     0
##   Vietnamese         0    0       0      0     0     0          0     0
```

```
##   Zulu               0    0       0      0     0     0          0     0
##
##              Mexico Netherlands New Line New Zealand Nigeria Norway Peru
##   Aboriginal      0           0        0           0       0      0    0
##   Arabic          0           0        0           0       0      0    0
##   Aramaic         0           0        0           0       0      0    0
##   Bosnian         0           0        0           0       0      0    0
##   Cantonese       0           0        0           0       0      0    0
##   Chinese         0           0        0           0       0      0    0
##   Czech           0           0        0           0       0      0    0
##   Danish          0           0        0           0       0      0    0
##   Dari            0           0        0           0       0      0    0
##   Dutch           0           4        0           0       0      0    0
##   Dzongkha        0           0        0           0       0      0    0
##   English         1           1        1          11       1      3    1
##   Filipino        0           0        0           0       0      0    0
##   French          0           0        0           0       0      0    0
##   German          1           0        0           0       0      0    0
##   Greek           0           0        0           0       0      0    0
##   Hebrew          0           0        0           0       0      0    0
##   Hindi           0           0        0           0       0      0    0
##   Hungarian       0           0        0           0       0      0    0
##   Icelandic       0           0        0           0       0      0    0
##   Indonesian      0           0        0           0       0      0    0
##   Italian         0           0        0           0       0      0    0
##   Japanese        0           0        0           0       0      0    0
##   Kannada         0           0        0           0       0      0    0
##   Kazakh          0           0        0           0       0      0    0
##   Korean          0           0        0           0       0      0    0
##   Mandarin        0           0        0           0       0      0    0
##   Maya            0           0        0           0       0      0    0
##   Mongolian       0           0        0           0       0      0    0
##   None            0           0        0           0       0      0    0
##   Norwegian       0           0        0           0       0      4    0
##   Persian         0           0        0           0       0      0    0
##   Polish          0           0        0           0       0      0    0
##   Portuguese      0           0        0           0       0      0    0
##   Romanian        0           0        0           0       0      0    0
##   Russian         0           0        0           0       0      0    0
##   Spanish        12           0        0           0       0      0    0
##   Swedish         0           0        0           0       0      0    0
##   Thai            0           0        0           0       0      0    0
##   Vietnamese      0           0        0           0       0      0    0
##   Zulu            0           0        0           0       0      0    0
##
##              Philippines Poland Romania Russia Slovakia South Africa
##   Aboriginal           0      0       0      0        0            0
##   Arabic               0      0       0      0        0            0
##   Aramaic              0      0       0      0        0            0
##   Bosnian              0      0       0      0        0            0
##   Cantonese            0      0       0      0        0            0
##   Chinese              0      0       0      0        0            0
##   Czech                0      0       0      0        0            0
##   Danish               0      0       0      0        0            0
```

```
##   Dari                0      0       0      0        0            0
##   Dutch               0      0       0      0        0            0
##   Dzongkha            0      0       0      0        0            0
##   English             1      1       2      1        1            7
##   Filipino            0      0       0      0        0            0
##   French              0      0       0      0        0            0
##   German              0      0       0      0        0            0
##   Greek               0      0       0      0        0            0
##   Hebrew              0      0       0      0        0            0
##   Hindi               0      0       0      0        0            0
##   Hungarian           0      0       0      0        0            0
##   Icelandic           0      0       0      0        0            0
##   Indonesian          0      0       0      0        0            0
##   Italian             0      0       0      0        0            0
##   Japanese            0      0       0      0        0            0
##   Kannada             0      0       0      0        0            0
##   Kazakh              0      0       0      0        0            0
##   Korean              0      0       0      0        0            0
##   Mandarin            0      0       0      0        0            0
##   Maya                0      0       0      0        0            0
##   Mongolian           0      0       0      1        0            0
##   None                0      0       0      0        0            0
##   Norwegian           0      0       0      0        0            0
##   Persian             0      0       0      0        0            0
##   Polish              0      1       0      0        0            0
##   Portuguese          0      0       0      0        0            0
##   Romanian            0      0       1      0        0            0
##   Russian             0      0       0      8        0            0
##   Spanish             0      0       0      0        0            0
##   Swedish             0      0       0      0        0            0
##   Thai                0      0       0      0        0            0
##   Vietnamese          0      0       0      0        0            0
##   Zulu                0      0       0      0        0            1
##
##              South Korea Soviet Union Spain Sweden Switzerland Taiwan
##   Aboriginal           0            0     0      0           0      0
##   Arabic               0            0     0      0           0      0
##   Aramaic              0            0     0      0           0      0
##   Bosnian              0            0     0      0           0      0
##   Cantonese            0            0     0      0           0      0
##   Chinese              0            0     0      0           0      0
##   Czech                0            0     0      0           0      0
##   Danish               0            0     0      0           0      0
##   Dari                 0            0     0      0           0      0
##   Dutch                0            0     0      0           0      0
##   Dzongkha             0            0     0      0           0      0
##   English              5            0    18      0           2      0
##   Filipino             0            0     0      0           0      0
##   French               0            0     0      0           0      0
##   German               0            0     0      0           1      0
##   Greek                0            0     0      0           0      0
##   Hebrew               0            0     0      0           0      0
##   Hindi                0            0     0      0           0      0
##   Hungarian            0            0     0      0           0      0
```

```
##   Icelandic            0           0     0      0           0      0
##   Indonesian           0           0     0      0           0      0
##   Italian              0           0     0      0           0      0
##   Japanese             0           0     0      0           0      0
##   Kannada              0           0     0      0           0      0
##   Kazakh               0           0     0      0           0      0
##   Korean               6           0     0      0           0      0
##   Mandarin             0           0     0      0           0      1
##   Maya                 0           0     0      0           0      0
##   Mongolian            0           0     0      0           0      0
##   None                 0           0     0      0           0      0
##   Norwegian            0           0     0      0           0      0
##   Persian              0           0     0      0           0      0
##   Polish               0           0     0      0           0      0
##   Portuguese           0           0     0      0           0      0
##   Romanian             0           0     0      0           0      0
##   Russian              0           1     0      1           0      0
##   Spanish              0           0    13      0           0      0
##   Swedish              0           0     0      4           0      0
##   Thai                 0           0     0      0           0      0
##   Vietnamese           0           0     0      0           0      0
##   Zulu                 0           0     0      0           0      0
##
##              Thailand Turkey   UK United Arab Emirates Unknown  USA
##   Aboriginal        0      0    1                     0       0    0
##   Arabic            0      1    0                     1       0    0
##   Aramaic           0      0    0                     0       0    1
##   Bosnian           0      0    0                     0       0    1
##   Cantonese         0      0    0                     0       0    1
##   Chinese           0      0    0                     0       0    0
##   Czech             0      0    0                     0       0    0
##   Danish            0      0    0                     0       0    0
##   Dari              0      0    0                     0       0    1
##   Dutch             0      0    0                     0       0    0
##   Dzongkha          0      0    0                     0       0    0
##   English           2      0  385                     0      20 3375
##   Filipino          0      0    0                     0       0    1
##   French            0      0    0                     0       0    0
##   German            0      0    0                     0       0    0
##   Greek             0      0    0                     0       0    0
##   Hebrew            0      0    0                     0       0    0
##   Hindi             0      0    0                     0       0    1
##   Hungarian         0      0    0                     0       0    0
##   Icelandic         0      0    0                     0       0    0
##   Indonesian        0      0    1                     0       0    0
##   Italian           0      0    1                     0       0    0
##   Japanese          0      0    0                     0       0    1
##   Kannada           0      0    0                     0       0    0
##   Kazakh            0      0    0                     0       0    0
##   Korean            0      0    0                     0       0    0
##   Mandarin          0      0    0                     0       0    0
##   Maya              0      0    0                     0       0    1
##   Mongolian         0      0    0                     0       0    0
##   None              0      0    0                     0       0    1
```

```
##   Norwegian         0      0    0                    0       0    0
##   Persian           0      0    0                    0       0    0
##   Polish            0      0    0                    0       0    0
##   Portuguese        0      0    1                    0       0    0
##   Romanian          0      0    0                    0       0    0
##   Russian           0      0    0                    0       0    0
##   Spanish           0      0    0                    0       0    7
##   Swedish           0      0    0                    0       0    0
##   Thai              3      0    0                    0       0    0
##   Vietnamese        0      0    0                    0       0    1
##   Zulu              0      0    1                    0       0    0
##
##              West Germany
##   Aboriginal             0
##   Arabic                 0
##   Aramaic                0
##   Bosnian                0
##   Cantonese              0
##   Chinese                0
##   Czech                  0
##   Danish                 0
##   Dari                   0
##   Dutch                  0
##   Dzongkha               0
##   English                1
##   Filipino               0
##   French                 0
##   German                 2
##   Greek                  0
##   Hebrew                 0
##   Hindi                  0
##   Hungarian              0
##   Icelandic              0
##   Indonesian             0
##   Italian                0
##   Japanese               0
##   Kannada                0
##   Kazakh                 0
##   Korean                 0
##   Mandarin               0
##   Maya                   0
##   Mongolian              0
##   None                   0
##   Norwegian              0
##   Persian                0
##   Polish                 0
##   Portuguese             0
##   Romanian               0
##   Russian                0
##   Spanish                0
##   Swedish                0
##   Thai                   0
##   Vietnamese             0
##   Zulu                   0
```

```
##
##              Australia Belgium Brazil Canada Chile China Czech Republic
##   Cantonese          0       0      0      0     0     0              0
##   Chinese            0       0      0      0     0     1              0
##   English            2       1      0      6     1     2              1
##   French             0       0      0      0     0     0              0
##   Hebrew             0       0      0      0     0     0              0
##   Hindi              0       0      0      0     0     0              0
##   Japanese           0       0      0      0     0     0              0
##   Korean             0       0      0      0     0     0              0
##   Mandarin           0       0      0      0     0     1              0
##   Panjabi            0       0      0      1     0     0              0
##   Portuguese         0       0      1      0     0     0              0
##   Romanian           0       0      0      0     0     0              0
##   Russian            0       0      0      0     0     0              0
##   Slovenian          0       0      0      0     0     0              0
##   Spanish            0       0      0      0     0     0              0
##   Tamil              0       0      0      0     0     0              0
##   Telugu             0       0      0      0     0     0              0
##   Urdu               0       0      0      0     0     0              0
##
##              France Germany Greece Hong Kong India Ireland Israel Italy
##   Cantonese       0       0      0         1     0       0      0     0
##   Chinese         0       0      0         0     0       0      0     0
##   English         5       1      1         0     0       1      0     0
##   French          3       0      0         0     0       0      0     0
##   Hebrew          0       0      0         0     0       0      1     0
##   Hindi           0       0      0         0     1       0      0     0
##   Japanese        0       0      0         0     0       0      0     0
##   Korean          0       0      0         0     0       0      0     0
##   Mandarin        0       0      0         0     0       0      0     0
##   Panjabi         0       0      0         0     0       0      0     0
##   Portuguese      0       0      0         0     0       0      0     0
##   Romanian        0       0      0         0     0       0      0     0
##   Russian         0       0      0         0     0       0      0     0
##   Slovenian       0       0      0         0     0       0      0     0
##   Spanish         0       0      0         0     0       0      0     1
##   Tamil           0       0      0         0     1       0      0     0
##   Telugu          0       0      0         0     1       0      0     0
##   Urdu            0       0      0         0     0       0      0     0
##
##              Japan Mexico New Zealand Pakistan Panama Romania Russia
##   Cantonese      0      0           0        0      0       0      0
##   Chinese        0      0           0        0      0       0      0
##   English        2      2           1        0      1       0      0
##   French         0      0           0        0      0       0      0
##   Hebrew         0      0           0        0      0       0      0
##   Hindi          0      0           0        0      0       0      0
##   Japanese       1      0           0        0      0       0      0
##   Korean         0      0           0        0      0       0      0
##   Mandarin       0      0           0        0      0       0      0
##   Panjabi        0      0           0        0      0       0      0
##   Portuguese     0      0           0        0      0       0      0
##   Romanian       0      0           0        0      0       1      0
```

```
##   Russian        0      0           0        0      0       0      1
##   Slovenian      0      0           0        0      0       0      0
##   Spanish        0      1           0        0      0       0      0
##   Tamil          0      0           0        0      0       0      0
##   Telugu         0      0           0        0      0       0      0
##   Urdu           0      0           0        1      0       0      0
##
##              Slovenia South Korea Spain Taiwan  UK USA
##   Cantonese         0           0     0      0   0   0
##   Chinese           0           0     0      0   0   0
##   English           0           1     2      0  26 228
##   French            0           0     0      0   1   0
##   Hebrew            0           0     0      0   0   1
##   Hindi             0           0     0      0   0   0
##   Japanese          0           0     0      0   0   0
##   Korean            0           1     0      0   0   0
##   Mandarin          0           0     0      1   0   0
##   Panjabi           0           0     0      0   0   0
##   Portuguese        0           0     0      0   0   0
##   Romanian          0           0     0      0   0   0
##   Russian           0           0     0      0   0   0
##   Slovenian         1           0     0      0   0   0
##   Spanish           0           0     0      0   0   0
##   Tamil             0           0     0      0   0   0
##   Telugu            0           0     0      0   0   0
##   Urdu              0           0     0      0   0   0
```

Appendix B

I. Initial Preparations

The following code is needed as a basis to run the code used to generate the graphs and output displayed in this report.

```
library(dplyr)
library(ggplot2)
library(gridExtra)

options(scipen = 5)

train_movies <- read.csv("train_movies.csv", stringsAsFactors = FALSE)

#Remove invalid country name
train_movies$country[train_movies$country == "Official site"] <- "Unknown"

won_movies <- subset(train_movies, train_movies$nominations != 0)

new_movies <- read.csv("test_movies.csv", stringsAsFactors = FALSE)
```

II. Language and Country

```
table(won_movies$language, won_movies$country)
```

III. Budget and Gross

plot1 <- ggplot(train_movies, aes((budget/10^6), (gross/10^6))) + geom_point() + labs(title="Movies from 1927-2014", x="Budget in Millions of Dollars", y="Gross in Millions of Dollars")
plot2 <- ggplot(won_movies, aes((budget/10^6), (gross/10^6))) + geom_point() + labs(title="Oscar-Winning Movies", x="Budget in Millions of Dollars", y="Gross in Millions of Dollars")
plot3 <- ggplot(new_movies, aes((budget/10^6), (gross/10^6))) + geom_point() + labs(title="Movies from 2015-2016", x="Budget in Millions of Dollars", y="Gross in Millions of Dollars")

grid.arrange(plot1, plot2, plot3, ncol=2)

IV. Genres

genres1 <- train_movies[, c("Action", "Adventure", "Biography", "Comedy", "Crime", "Drama", "Musical", "Western", "Family", "Film.Noir", "History", "Mystery", "Romance", "Sport", "War", "Fantasy", "Sci.Fi", "Thriller", "Documentary", "Horror", "Animation")]
genres2 <- won_movies[, c("Action", "Adventure", "Biography", "Comedy", "Crime", "Drama", "Musical", "Western", "Family", "Film.Noir", "History", "Mystery", "Romance", "Sport", "War", "Fantasy", "Sci.Fi", "Thriller", "Documentary", "Horror", "Animation")]
genres_new <- new_movies[, c("Action", "Adventure", "Biography", "Comedy", "Crime", "Drama", "Musical", "Western", "Family", "Film.Noir", "History", "Mystery", "Romance", "Sport", "War", "Fantasy", "Sci.Fi", "Thriller", "Documentary", "Horror", "Animation")]

genre_plot1 <- as.data.frame(colSums(genres1))
percent1 <- (genre_plot1$`colSums(genres1)` / nrow(train_movies)) * 100
plt1 <- ggplot(genre_plot1, aes(x=rownames(genre_plot1), y=percent1)) + geom_bar(stat = "identity", position = "dodge") + labs(x = "Genres", y = "Percent", title = "Movie Genres 1927-2014") + theme(axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1))

genre_plot2 <- as.data.frame(colSums(genres2))
percent2 <- (genre_plot2$`colSums(genres2)` / nrow(won_movies)) * 100
plt2 <- ggplot(genre_plot2, aes(x=rownames(genre_plot2), y=percent2)) + geom_bar(stat = "identity", position = "dodge") + labs(x = "Genres", y = "Percent", title = "Oscar-Winning Movie Genres") + theme(axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1))

genre_plot_new <- as.data.frame(colSums(genres_new)) percent_new <- (genre_plot_new$`colSums(genres_new)` / nrow(new_movies)) * 100

plt3 <- ggplot(genre_plot_new, aes(x=rownames(genre_plot_new), y=percent_new)) + geom_bar(stat = "identity", position = "dodge") + labs(x = "Genres", y = "Percent", title = "New Movie Genres") + theme(axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1))

grid.arrange(plt1, plt2, plt3, ncol=2)

V. Appendix A

table(train_movies$language, train_movies$country)

table(new_movies$language, new_movies$country)