**Data Set Title**

**Exploratory Analysis**

Jacob Redmon, [jredmon@bellarmine.edu](mailto:jredmon@bellarmine.edu)

1. **INTRODUCTION**

This data set contains some basic data from movies in the Internet Movie Database. It contains information about the number of Facebook likes the director and starring actors have, which will provide some interesting analysis. The data set can be found at <https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset>.

1. **DATA SET DESCRIPTION**

This data set contains 5043 samples with 27 columns with various data types. A complete listing is shown in **Table 1**.

**Table 1: Data Types and Missing Data**

|  |  |  |
| --- | --- | --- |
| *Variable Name* | *Data Type* | *Missing Data (%)* |
| movie\_title | nominal/object | 0% |
| title\_year | interval/int64 | 2% |
| country | nominal/object | 0.09% |
| language | nominal/object | 0.2% |
| budget | ratio/float64 | 9.7% |
| gross | ratio/float64 | 17.5% |
| imdb\_score | ratio/float64 | 0% |
| duration | ratio/float64 | 0.3% |
| genres | nominal/object | 0% |
| plot\_keywords | nominal/object | 3% |
| color | nominal/object | 0.3% |
| content\_rating | nominal/object | 6% |
| facenumber\_in\_poster | ratio/float64 | 2.5% |
| num\_critic\_for\_reviews | ratio/float64 | 0.9% |
| num\_user\_for\_reviews | ratio/float64 | 0.4% |
| num\_voted\_users | ratio/int64 | 0% |
| director\_name | nominal/object | 2% |
| actor\_1\_name | nominal/object | 1.4% |
| actor\_2\_name | nominal/object | 0.2% |
| actor\_3\_name | nominal/object | 4.5% |
| director\_facebook\_likes | ratio/int64 | 2% |
| actor\_1\_facebook\_likes | ratio/int64 | 0.1% |
| actor\_2\_facebook\_likes | ratio/int64 | 0.2% |
| actor\_3\_facebook\_likes | ratio/int64 | 0.4% |
| cast\_total\_facebook\_likes | ratio/int64 | 0% |
| movie\_facebook\_likes | ratio/int64 | 0% |
| movie\_imdb\_link | nominal/object | 0% |
| first\_genre | nominal/object | 0% |

1. **Data Set Summary Statistics**

After cleaning up the data by removing any rows with null values and adding an additional column called `first\_genre` that captures just the first genre listed in the `genre` column, looking at summary statistics can help us draw some conclusions about the movies included in the data set. The summary statistics are located in **Table 2** displays some general statistics about the continuous variables. **Tables 3a-3e** showcase the proportions for the categorical variables; the larger tables (the director and actor names, as well as genre and plot keywords contain 1000+ categories) are shown in CSV files in the project folder as they were too large for this document. The correlation matrix (which is too large to fit in this document accurately, so a CSV file is located in the project folder as well) in **Table 4**, and the following heatmap, displays how the continuous data points are correlated with each other. We see a large amount of correlation between a movies gross sales and the amount of reviews it has on IMDB. Additionally, there is significant correlation between actors’ and directors’ Facebook likes, with the highest degree of correlation between the lead actors’ and the movies’ total number of Facebook likes, which is expected.

**Table 2: Summary Statistics for Movie Metadata**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Variable Name* | *Count* | *Mean* | *Standard Deviation* | *Min* |
| title\_year | 3756 | 2002.977 | 9.888108 | 1927 |
| budget | 3756 | 46236850 | 2.26E+08 | 218 |
| gross | 3756 | 52612824 | 70317867 | 162 |
| imdb\_score | 3756 | 6.465282 | 1.056247 | 1.6 |
| duration | 3756 | 110.258 | 22.64672 | 37 |
| facenumber\_in\_poster | 3756 | 1.377263 | 2.041541 | 0 |
| num\_critic\_for\_reviews | 3756 | 167.3783 | 123.452 | 2 |
| num\_user\_for\_reviews | 3756 | 336.8432 | 411.2274 | 4 |
| num\_voted\_users | 3756 | 105826.7 | 152035.4 | 91 |
| director\_facebook\_likes | 3756 | 807.3365 | 3068.172 | 0 |
| actor\_1\_facebook\_likes | 3756 | 7751.339 | 15519.34 | 0 |
| actor\_2\_facebook\_likes | 3756 | 2021.776 | 4544.908 | 0 |
| actor\_3\_facebook\_likes | 3756 | 771.2796 | 1894.25 | 0 |
| cast\_total\_facebook\_likes | 3756 | 11527.1 | 19122.18 | 0 |
| movie\_facebook\_likes | 3756 | 9353.829 | 21462.89 | 0 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Variable Name* | *25th* | *50th* | *75th* | *Max* |
| title\_year | 1999 | 2004 | 2010 | 2016 |
| budget | 10000000 | 25000000 | 50000000 | 1.22E+10 |
| gross | 8270233 | 30093107 | 66881941 | 7.61E+08 |
| imdb\_score | 5.9 | 6.6 | 7.2 | 9.3 |
| duration | 96 | 106 | 120 | 330 |
| facenumber\_in\_poster | 0 | 1 | 2 | 43 |
| num\_critic\_for\_reviews | 77 | 138.5 | 224 | 813 |
| num\_user\_for\_reviews | 110 | 210 | 398.25 | 5060 |
| num\_voted\_users | 19667 | 53973.5 | 128602 | 1689764 |
| director\_facebook\_likes | 11 | 64 | 235 | 23000 |
| actor\_1\_facebook\_likes | 745 | 1000 | 13000 | 640000 |
| actor\_2\_facebook\_likes | 384.75 | 685.5 | 976 | 137000 |
| actor\_3\_facebook\_likes | 194 | 436 | 691 | 23000 |
| cast\_total\_facebook\_likes | 1919.75 | 4059.5 | 16240 | 656730 |
| movie\_facebook\_likes | 0 | 227 | 11000 | 349000 |

Table 3a: Proportions for Country (n=45)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| Afghanistan | 1 | 0.026624 |
| Argentina | 3 | 0.079872 |
| Aruba | 1 | 0.026624 |
| Australia | 39 | 1.038339 |
| Belgium | 1 | 0.026624 |
| Brazil | 5 | 0.133120 |
| Canada | 59 | 1.570820 |
| Chile | 1 | 0.026624 |
| China | 13 | 0.346113 |
| Colombia | 1 | 0.026624 |
| Czech Republic | 3 | 0.079872 |
| Denmark | 8 | 0.212993 |
| Finland | 1 | 0.026624 |
| France | 101 | 2.689031 |
| Georgia | 1 | 0.026624 |
| Germany | 80 | 2.129925 |
| Greece | 1 | 0.026624 |
| Hong Kong | 13 | 0.346113 |
| Hungary | 2 | 0.053248 |
| Iceland | 1 | 0.026624 |
| India | 5 | 0.133120 |
| Indonesia | 1 | 0.026624 |
| Iran | 4 | 0.106496 |
| Ireland | 7 | 0.186368 |
| Israel | 1 | 0.026624 |
| Italy | 11 | 0.292865 |
| Japan | 15 | 0.399361 |
| Mexico | 6 | 0.159744 |
| Netherlands | 3 | 0.079872 |
| New Line | 1 | 0.026624 |
| New Zealand | 11 | 0.292865 |
| Norway | 4 | 0.106496 |
| Official site | 1 | 0.026624 |
| Peru | 1 | 0.026624 |
| Poland | 1 | 0.026624 |
| Romania | 2 | 0.053248 |
| Russia | 3 | 0.079872 |
| South Africa | 3 | 0.079872 |
| South Korea | 8 | 0.212993 |
| Spain | 21 | 0.559105 |
| Taiwan | 2 | 0.053248 |
| Thailand | 4 | 0.106496 |
| UK | 318 | 8.466454 |
| USA | 2987 | 79.526092 |
| West Germany | 1 | 0.026624 |

Table 3b: Proportions for Language (n=34)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| Aboriginal | 2 | 0.053248 |
| Arabic | 1 | 0.026624 |
| Aramaic | 1 | 0.026624 |
| Bosnian | 1 | 0.026624 |
| Cantonese | 7 | 0.186368 |
| Czech | 1 | 0.026624 |
| Danish | 3 | 0.079872 |
| Dari | 2 | 0.053248 |
| Dutch | 3 | 0.079872 |
| English | 3598 | 95.793397 |
| Filipino | 1 | 0.026624 |
| French | 34 | 0.905218 |
| German | 10 | 0.266241 |
| Hebrew | 1 | 0.026624 |
| Hindi | 5 | 0.133120 |
| Hungarian | 1 | 0.026624 |
| Indonesian | 2 | 0.053248 |
| Italian | 7 | 0.186368 |
| Japanese | 10 | 0.266241 |
| Kazakh | 1 | 0.026624 |
| Korean | 5 | 0.133120 |
| Mandarin | 15 | 0.399361 |
| Maya | 1 | 0.026624 |
| Mongolian | 1 | 0.026624 |
| None | 1 | 0.026624 |
| Norwegian | 4 | 0.106496 |
| Persian | 3 | 0.079872 |
| Portuguese | 5 | 0.133120 |
| Romanian | 1 | 0.026624 |
| Russian | 1 | 0.026624 |
| Spanish | 23 | 0.612354 |
| Thai | 3 | 0.079872 |
| Vietnamese | 1 | 0.026624 |
| Zulu | 1 | 0.026624 |

Table 3c: Proportions for Color (n=2)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| Black and White | 124 | 3.301384 |
| Color | 3632 | 96.698616 |

Table 3d: Proportions for Content Rating (n=12)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| Approved | 17 | 0.452609 |
| G | 87 | 2.316294 |
| GP | 1 | 0.026624 |
| M | 2 | 0.053248 |
| NC-17 | 6 | 0.159744 |
| Not Rated | 34 | 0.905218 |
| PG | 566 | 15.069223 |
| PG-13 | 1308 | 34.824281 |
| Passed | 3 | 0.079872 |
| R | 1700 | 45.260916 |
| Unrated | 22 | 0.585729 |
| X | 10 | 0.266241 |

Table 3e: Proportions for First Genre (n=17)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| Action | 959 | 25.53248 |
| Adventure | 369 | 9.824281 |
| Animation | 45 | 1.198083 |
| Biography | 205 | 5.457934 |
| Comedy | 989 | 26.3312 |
| Crime | 255 | 6.789137 |
| Documentary | 26 | 0.692226 |
| Drama | 668 | 17.78488 |
| Family | 3 | 0.079872 |
| Fantasy | 37 | 0.985091 |
| Horror | 164 | 4.366347 |
| Musical | 2 | 0.053248 |
| Mystery | 23 | 0.612354 |
| Romance | 1 | 0.026624 |
| Sci-Fi | 7 | 0.186368 |
| Thriller | 1 | 0.026624 |
| Western | 2 | 0.053248 |

Table 4: Correlation Table/Tables

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | title\_year | budget | gross | imdb\_score | duration | facenumber\_in\_poster | num\_critic\_for\_reviews | num\_user\_for\_reviews | num\_voted\_users | director\_facebook\_likes | actor\_1\_facebook\_likes | actor\_2\_facebook\_likes | actor\_3\_facebook\_likes | cast\_total\_facebook\_likes | movie\_facebook\_likes |
| title\_year | 1 | 0.047138 | 0.054808 | -0.13498 | -0.13021 | 0.070093 | 0.420202 | 0.01947 | 0.023687 | -0.04466 | 0.096194 | 0.12241 | 0.117475 | 0.127045 | 0.306987 |
| budget | 0.047138 | 1 | 0.099496 | 0.02919 | 0.068012 | -0.02167 | 0.104717 | 0.070372 | 0.065927 | 0.018167 | 0.016418 | 0.035715 | 0.039966 | 0.028682 | 0.05249 |
| gross | 0.054808 | 0.099496 | 1 | 0.21474 | 0.245726 | -0.03202 | 0.464187 | 0.544674 | 0.624949 | 0.138351 | 0.14401 | 0.252842 | 0.299864 | 0.235601 | 0.366933 |
| imdb\_score | -0.13498 | 0.02919 | 0.21474 | 1 | 0.366221 | -0.06549 | 0.347886 | 0.325003 | 0.48243 | 0.192314 | 0.093597 | 0.102372 | 0.065544 | 0.106803 | 0.281155 |
| duration | -0.13021 | 0.068012 | 0.245726 | 0.366221 | 1 | 0.026919 | 0.228631 | 0.351595 | 0.339592 | 0.180644 | 0.083632 | 0.128935 | 0.125797 | 0.120179 | 0.215303 |
| facenumber\_in\_poster | 0.070093 | -0.02167 | -0.03202 | -0.06549 | 0.026919 | 1 | -0.03436 | -0.08097 | -0.03221 | -0.04807 | 0.057387 | 0.073098 | 0.106189 | 0.080687 | 0.015185 |
| num\_critic\_for\_reviews | 0.420202 | 0.104717 | 0.464187 | 0.347886 | 0.228631 | -0.03436 | 1 | 0.563684 | 0.592473 | 0.175128 | 0.165986 | 0.253125 | 0.252782 | 0.236533 | 0.705226 |
| num\_user\_for\_reviews | 0.01947 | 0.070372 | 0.544674 | 0.325003 | 0.351595 | -0.08097 | 0.563684 | 1 | 0.778881 | 0.217103 | 0.121935 | 0.187016 | 0.205343 | 0.178675 | 0.370465 |
| num\_voted\_users | 0.023687 | 0.065927 | 0.624949 | 0.48243 | 0.339592 | -0.03221 | 0.592473 | 0.778881 | 1 | 0.299624 | 0.17959 | 0.244822 | 0.267762 | 0.249119 | 0.517751 |
| director\_facebook\_likes | -0.04466 | 0.018167 | 0.138351 | 0.192314 | 0.180644 | -0.04807 | 0.175128 | 0.217103 | 0.299624 | 1 | 0.08963 | 0.116058 | 0.117405 | 0.118547 | 0.161962 |
| actor\_1\_facebook\_likes | 0.096194 | 0.016418 | 0.14401 | 0.093597 | 0.083632 | 0.057387 | 0.165986 | 0.121935 | 0.17959 | 0.08963 | 1 | 0.391166 | 0.252408 | 0.944813 | 0.1303 |
| actor\_2\_facebook\_likes | 0.12241 | 0.035715 | 0.252842 | 0.102372 | 0.128935 | 0.073098 | 0.253125 | 0.187016 | 0.244822 | 0.116058 | 0.391166 | 1 | 0.553755 | 0.642991 | 0.232769 |
| actor\_3\_facebook\_likes | 0.117475 | 0.039966 | 0.299864 | 0.065544 | 0.125797 | 0.106189 | 0.252782 | 0.205343 | 0.267762 | 0.117405 | 0.252408 | 0.553755 | 1 | 0.489795 | 0.271508 |
| cast\_total\_facebook\_likes | 0.127045 | 0.028682 | 0.235601 | 0.106803 | 0.120179 | 0.080687 | 0.236533 | 0.178675 | 0.249119 | 0.118547 | 0.944813 | 0.642991 | 0.489795 | 1 | 0.205564 |
| movie\_facebook\_likes | 0.306987 | 0.05249 | 0.366933 | 0.281155 | 0.215303 | 0.015185 | 0.705226 | 0.370465 | 0.517751 | 0.161962 | 0.1303 | 0.232769 | 0.271508 | 0.205564 | 1 |

A picture containing square

Description automatically generated

1. **DATA SET GRAPHICAL EXPLORATION**

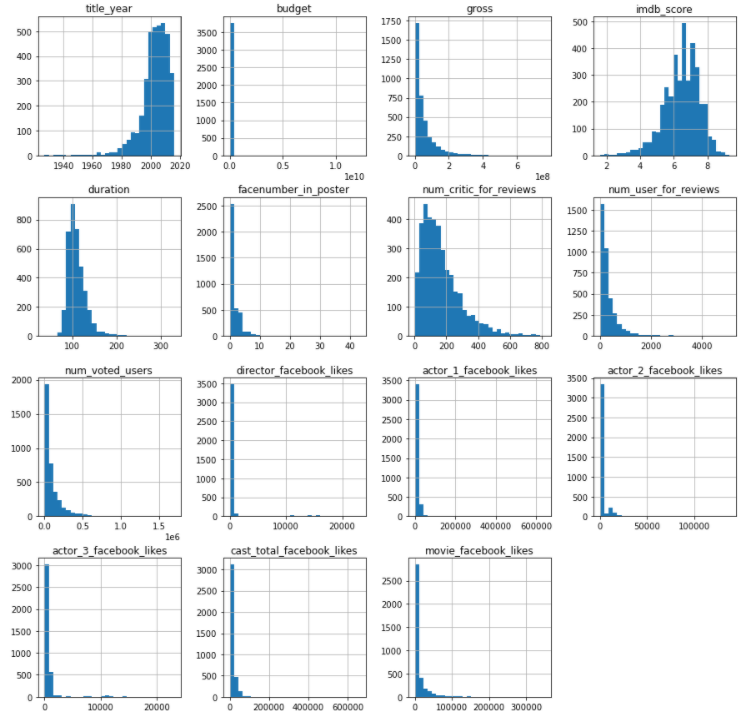
This section contains a graphical exploration of the data set, taking a look at various distributions of the data, scatter plots, and bar charts.

* 1. *Distributions*

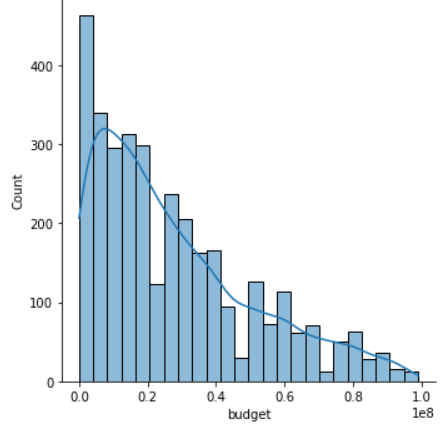
The first look at distributions of the data in **Figure 1** was using Pandas `dataframe.hist()` function to generate histograms of all the continuous variables in the data set. Here, I noticed a significant skew to zero for many variables, most notably `budget` looking empty other than ~98% of the data falling into the first bin, as it caps to nearly 12 billion. Similar disparities are seen in all columns with Facebook like data, and also the numbers of faces in movie posters (which is mainly zero). It should be noted that most movies in this data set seem to be successful since the `gross` column is skewed to zero way less than the `budget` column.

The next distribution I looked, **Figure 2**, at was what that first bin of `budget` actually looks like. I generated a Kernel Density Plot for the budget of movies that had a budget less than $100,000,000, which shows a better look at that column.

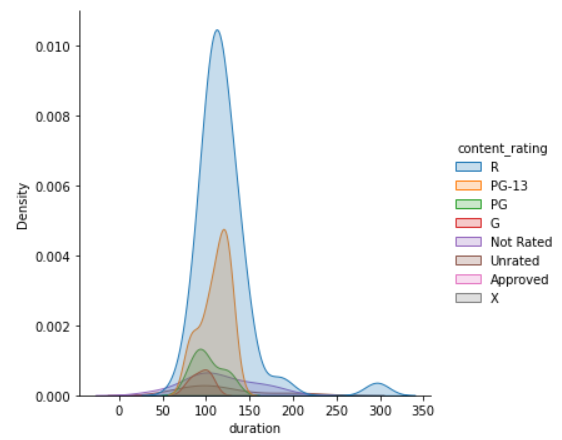
**Figure 3** shows another Kernel Density Plot of the `content\_rating` column for non-English movies. Looking at the data for non-English movies, which is also done later on in the report, is interesting as it cuts out a large portion of the data set (what the Oscars label “foreign” films).



**Figure 1: Histograms of all continuous variables (multiple plots)**



**Figure 2: Kernel Density Plot for movies with a budget under $100,000,000**



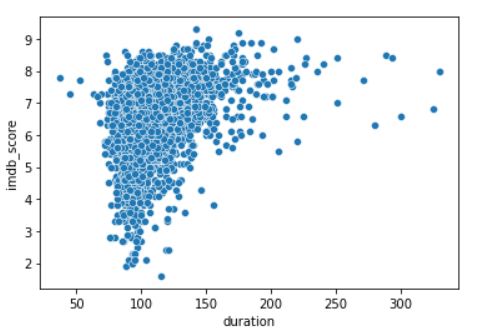
**Figure 3: Kernel Density Plot for content rating compared to duration**

* 1. *Scatter Plots / Pairwise Plots (continuous variables)*

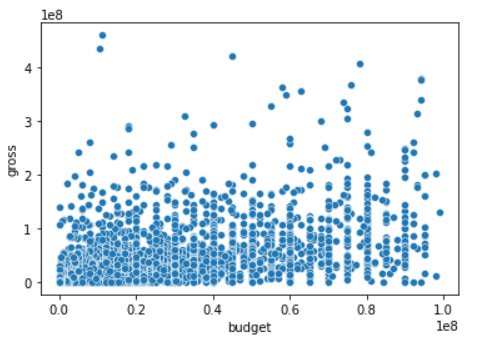
I was most interested in seeing how different variables were distributed with respect to IMDB scores, so many of the following graphs contain data compared to the `imdb\_score` column. **Figure 4** shows a scatter plot of IMDB scores compared to the duration of a movie, where I noticed that duration does not have a significant effect on rating, and that there are, thankfully, few movies longer than three hours.

**Figure 5** shows a scatter plot comparing budget to gross for movies with less than a $100,000,000 budget, which shows just the slightest trend to a larger budget making more money, and some notable outliers that made the most money with a (relatively) smaller budget.

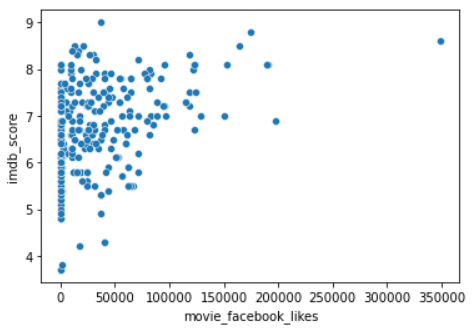
**Figure 6** shows a scatter plot comparing IMDB score to number of Facebook likes for movies with greater than a $100,000,000 budget, which shows the expected trend of a large amount of Facebook likes means the movie is better rated.



**Figure 4: Scatter plot comparing IMDB score to duration**



**Figure 5: Scatter plot comparing budget to gross for movies with < $100,000,000 budget**



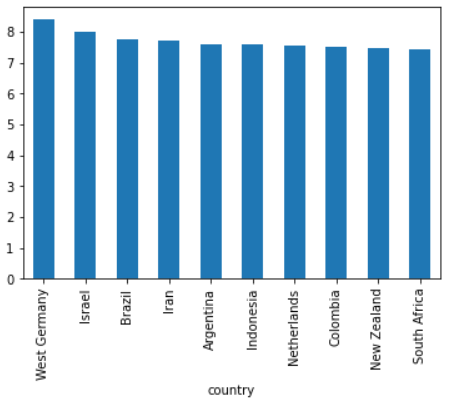
**Figure 4: Scatter plot comparing IMDB score to number of Facebook likes for movies with >$100,000,000 budget**

* 1. *Bar charts (categorical variables)*

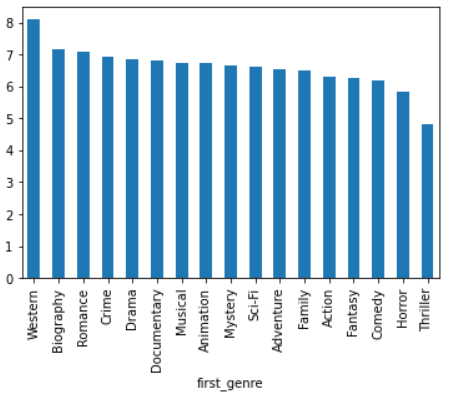
Unfortunately, most of the data set’s categorical variables had many categories, which made it difficult to look at spreads of all the data. I graphed each country’s average IMDB score for and displayed the top ten in **Figure 7**. I was surprised to see these ten countries at the top since I cannot name a movie from any of them (which might show more about my movie knowledge than the actual data).

**Figure 8** shows the genres ranked by average IMDB score, and I was unsurprised to see thrillers rated the lowest because they are generally not good (but so entertaining). Westerns were significantly the highest rated genre, but I am curious if that is accurate and if the data set was not skewed with the best westerns.

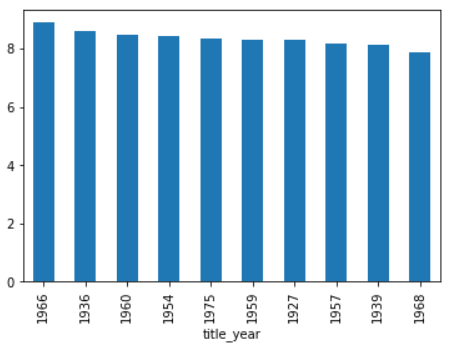
**Figure 9** shows the top ten years ranked by average IMDB score. I was incredibly surprised to see the latest year was 1975 at fifth place, which makes me think if the sheer volume of recent movies brings down the ratings for newer years.



**Figure 7: Bar chart showing the top 10 countries with the highest average IMDB scores**



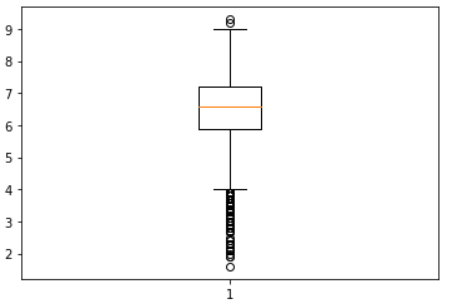
**Figure 8: Bar chart showing the genres ranked by average IMDB score**



**Figure 9: Bar chart showing the top 10 years ranked by average IMDB score**

* 1. *Other Plots*

Out of curiosity, I made a pair plot for the entire data set (not included in this report) which showed some interesting trends though it was far too zoomed out to be useful. I wanted to look at a box-and-whisker plot for the scores to see how they were distributed and found that most movies in this data set were ranked between 6 and 7 on IMDB, with very few approaching a perfect score, seen in **Figure 10**.



**Figure 10: Box-and-whisker plot for IMDB scores**

1. **SUMMARY OF FINDINGS**

This data set showed largely expected results with some movie metadata. Generally, the more Facebook likes a movie, its director, or its actors had, the better the movie both performed financially and was rated on IMDB. This finding is a good argument for the effectiveness of advertising for financial success of movies. That is not to say, however, that that is necessarily the rule. Many movies, usually older movies, performed perfectly fine with their zero likes on a social media platform that did not exist when it was released. Looking at subsets of the data, mainly excluding either movies made in the United States or performed in English, gave a more clear look on how movies performed globally, as both those factors outweigh all other categories by an extremely vast majority.