Project 1 - Group 1: Movie Data Analysis from 1995 to 2015

For this project, our group wanted to analyze movie data from the years 1995 through 2015. We were tasked with grabbing data from the internet, cleaning the data, doing statistical analyses, and creating visualizations on our findings. Before beginning the project, our group was hoping to find the answers to a few questions like what genres are most popular over the timeframe, what films had the biggest ROI, and how the IMDb score correlated with the films with the biggest ROI.

The final cleaned data frame that we used for our analysis and visualizations came from three CSV files that we found through Kaggle, a Google owned website that allows users to find and upload published datasets. The CSV files contained information concerning the movie titles, IMDb score, years the movies released, the budget of the films, the gross income/profit the film made, genre, language, and country. All three CSV files had some of the same information but each had information that the other CSV files did not have, so we had to merge the data sets. After merging all three data sets, we had to drop the N/A values, clean up the column names, and delete duplicate data.

Strengths:

Some strengths of our data is that is has many different variables to analyze and compare against each other such as genre, year, and revenue. The data covers twenty years which gives us enough data to work with and to see trends. Other data scientists use Kaggle as a data source for their work and projects.

Limitations:

There are a few limitations with our data. Our data contains mostly movie information from the United States. The data is mostly comprised of blockbusters and does not contain every indie film released in this time frame. We had to do our analysis between the years of 1995 through 2015 because that is where we had 20 years of most sufficient data.

For this project, our group coded in Jupyter Notebooks. We used Python as the main language, but also used Pandas for its libraries and Matplotlib to create our visualizations. Our group used a few different notebooks so that we could work together on different parts of the project simultaneously. We started out by gathering our three CSV files that contained the data we were wanting to use and merged them into one usable data frame. This proved to be challenging in its own right due to the columns we were wanting to merge on not matching perfectly. For example, we had to split the IMDb ID out of the link for one of the CSVs so that we could merge it with one of the other CSVs on the IMDb ID. On another CSV we had to split the title from its year since it was all one continuous string so that we could merge it with the two other CSVs that had already been merged. We then dropped

N/A values and duplicates and then deleted non necessary columns. All of this got our group to a point of where we could start analyzing the data.

As for our analysis, our group had a few main questions and hypotheses that we wanted to test and answer. Firstly, we had to show that the data that we pulled for this project would be of sufficient quality for our analysis. We determined that we should demonstrate the data had similar trends some other distinct movie dataset in order to validate that it would be of sufficient quality to support our conclusions. We found an additional dataset on Kaggle that contained information regarding movies from the corresponding years and their gross revenues. We looked at the percent change of revenues year over year for each dataset. We were able to observe that our original data (Fg 1, purple line) demonstrates similar revenue trends to the new dataset (Fg 1, blue line). We also had to show that the two datasets were distinct from each other by running a T-test. A p-value of less than 0.05 would mean that in fact we had two distinct datasets. When we compared both datasets, our T-test of 1.46e-31 showed that our original dataset was distinct from the new dataset.

Since we had a movie dataset distinct from another movie dataset and both datasets show similar trends, we concluded that it was sufficient to conduct our analysis. We continued to only use the original dataset because it contained additional categorical information for each movie listed that we needed for the purposes of this analysis.

We started out creating a data frame so that we could visualize the adjusted revenue for the United States for each of the twenty years that we analyzed. We found that the dataset had revenue spikes in 1997, 2003, 2009 and 2015 with the most profitable years being 1997 and 2015 (Fg 2).

We then moved on to analyzing which content rating, on average, made the most revenue in the time period. We started out by creating a data frame that contained the average revenue per content rating. Our analysis and visualization show that G rated movies made the most revenue (Fg 3).

Next, we wanted to see how the five most popular genres trended over the twenty year period that we analyzed and which genre had the greatest change in revenue. We created a sub data frame so that we only had Genres, Release Year, and Gross Revenue to chart out. Based on the line graph that we created (Fg 4), Action films had the greatest increase in gross revenue over the time period over the other five analyzed genres, while Drama films had the greatest decrease in gross revenue. We ran a one tail ANOVA on this data and concluded that the data was statistically significant (p = 2.14-19). Using gross revenue as a marker for popularity at the time of film release, we can assume that action films increased in popularity and drama films decreased in popularity over the 2 decade period.

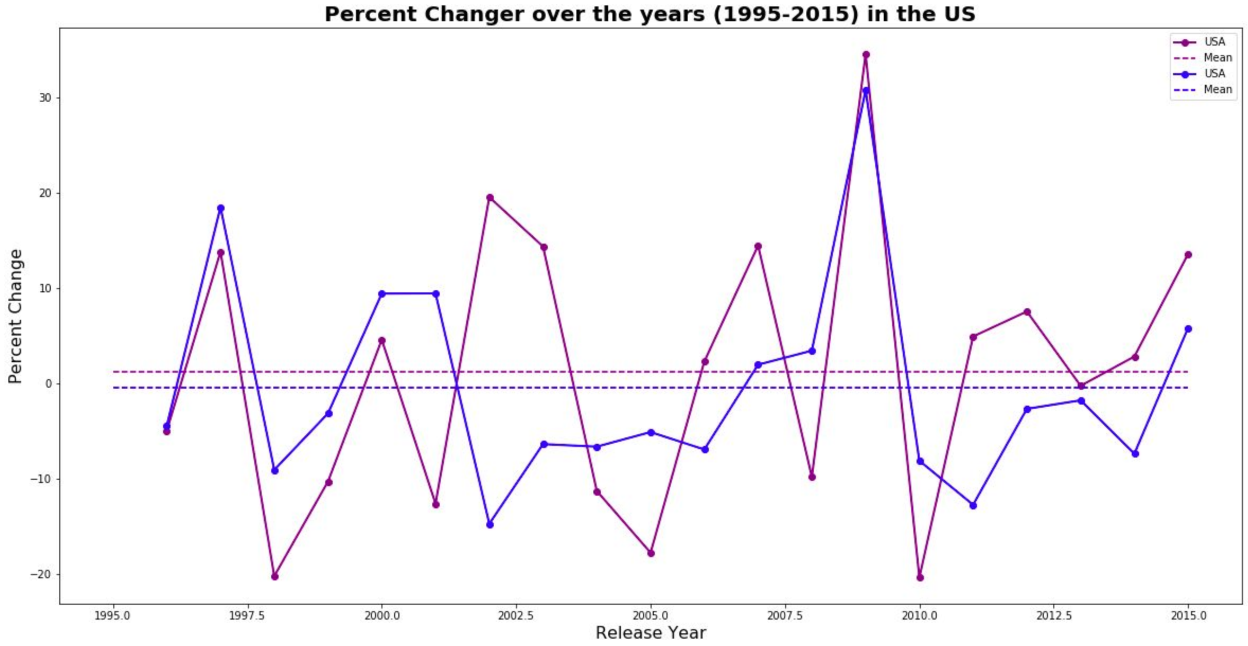
Our next task was to see how IMDb scores trended over each of the five main genres (Fg 5). The highest mean IMDb score belonged to the Biography genre. We also looked at the duration of each film in minutes for each of the genres and found that on average (Fg 6), the Biography genre had the highest mean runtime. Our group determined whether there were any statistically significant differences between the means of two or more independent (unrelated) genres. One-way ANOVA is ideal to understand whether IMDb Score differed based on genres amongst movies, dividing movies into five independent groups. One downside of the one-way ANOVA is an omnibus test statistic and cannot tell us which specific genres were statistically significantly different from each other; it only tells us that at least two genres were different. Since the p-value is 3.12x10-77 which is less than the significance level (α=0.05), you reject the null hypothesis and conclude that not all population means of IMDb Scores are equal. Our group also tested based on the duration. We reject the null hypothesis of equal means of duration between the 5 genres because the p-value is 1.67x10-71 which is less than the significance level (α=0.05).

We also observed that the movies with the highest return on investment were the following in order: Paranormal Activity, The Blair Witch Project, The Gallows, Super Size Me, and In the Company of Men. The films with the highest IMDb score within the time period were The Dark Knight, 12 Angry Men, The Lord of the Rings: The Return of the King, The Lord of the Rings: The Fellowship of the Ring, and Fight Club.

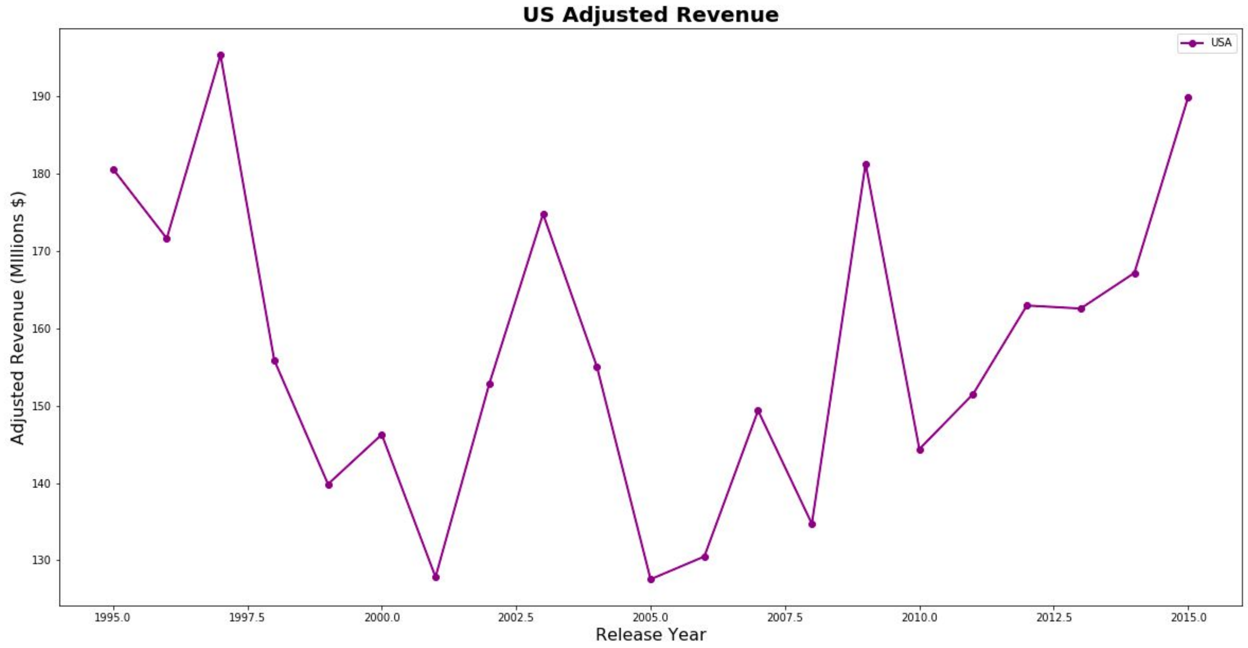
In conclusion, our data is sufficient to support that the strength of our analysis lies on the trends observed. Our analysis shows that the revenue trends fluctuated over the years with increased revenue spikes about every six years. Additionally, the film genre that had the largest positive change in the time frame was Action. The highest average revenue per rating was G rated films. Based on IMDb scores, the Biography genre had the highest mean score over the other genres along with the highest mean duration. Based on the top five highest return on investment films, the horror genre seems to be an untapped genre to make money. They can be extremely cheap to make and can make a rather large sum of revenue.

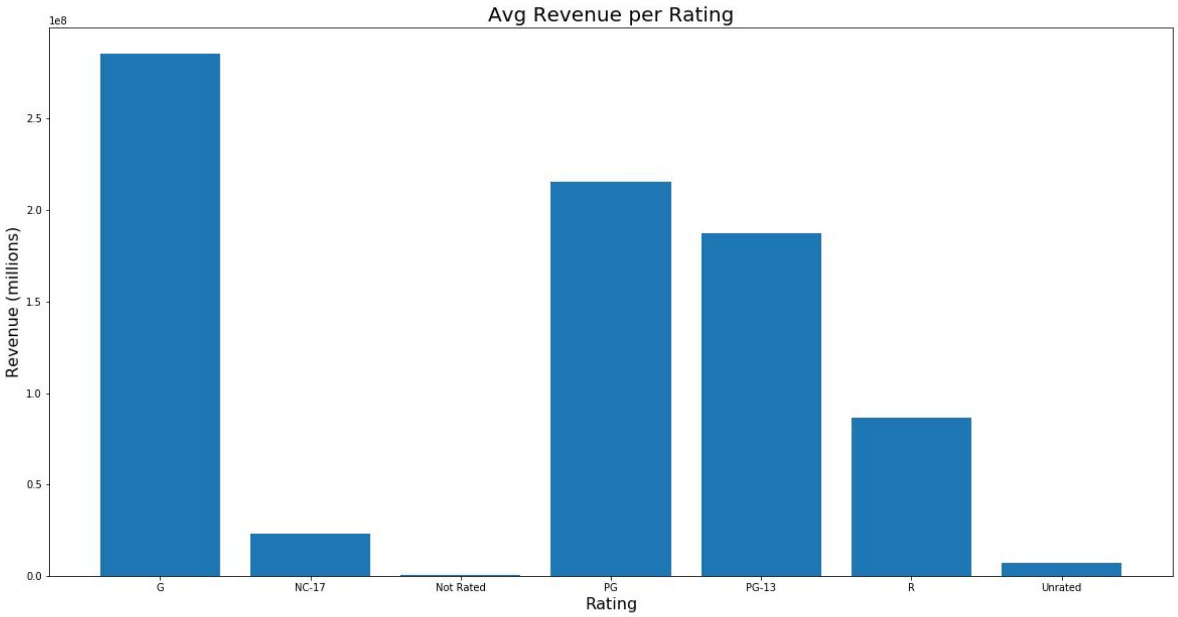
Appendix

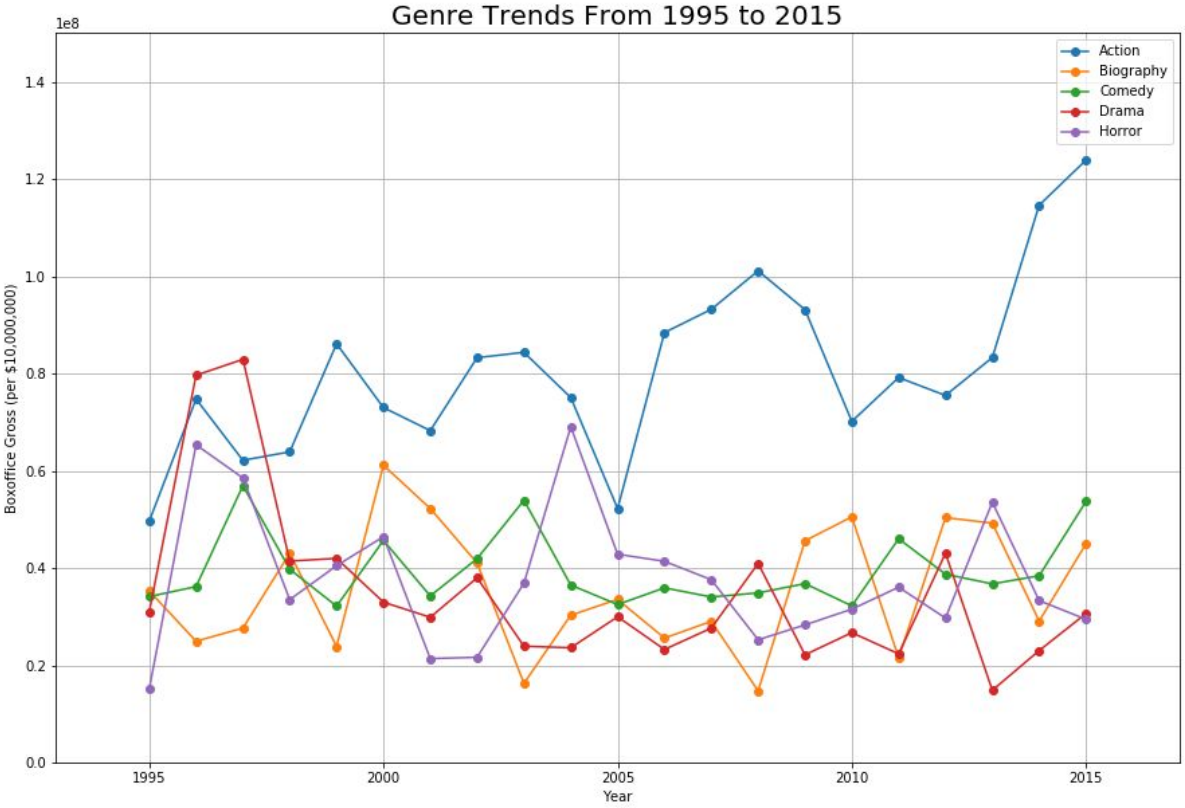
Fg 1

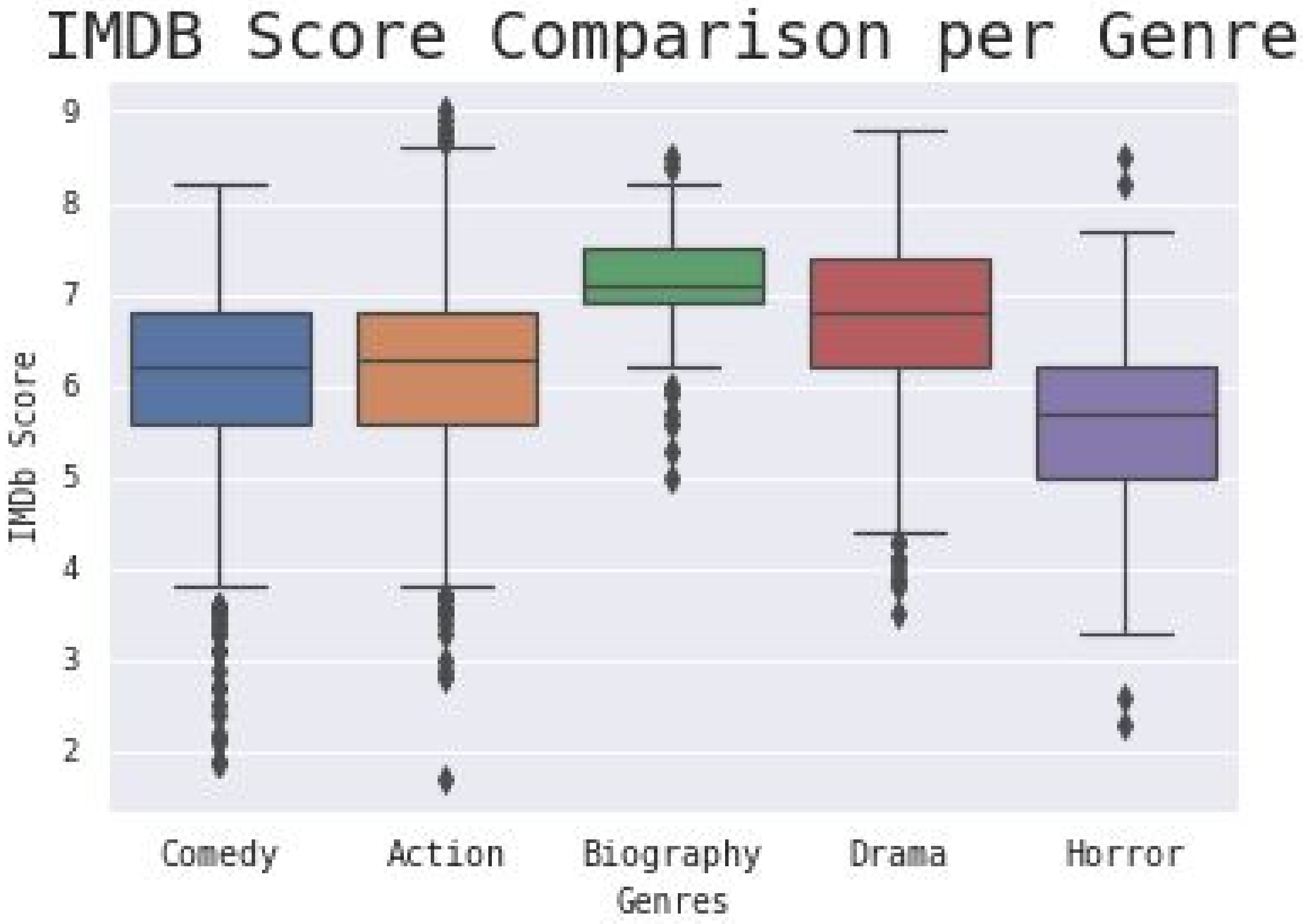


Fg 2

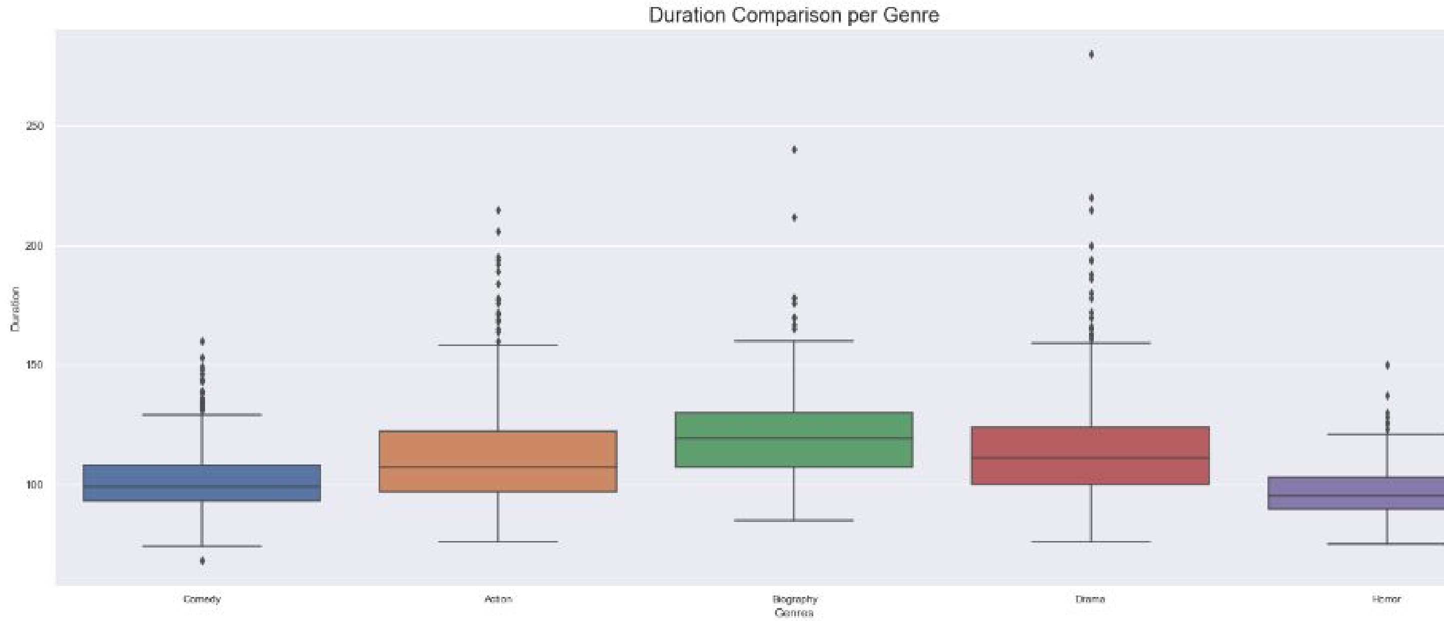



Avg Revenue per Rating
1e8
Revenue (millions)
0.0
0.5
1.0
1.5
2.0
2.5
G
NC-17
Not Rated
PG
PG-13
R
Unrated
Rating

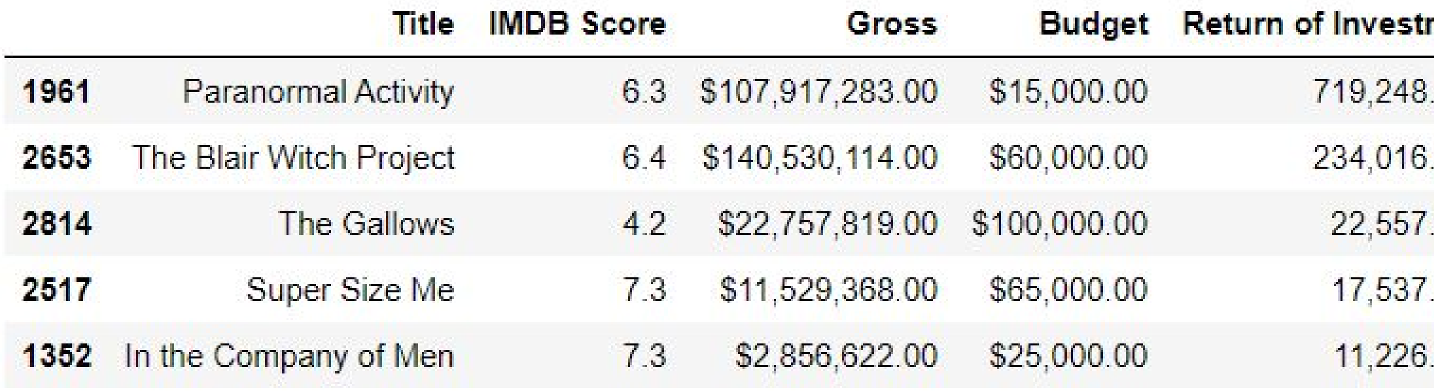

Genre Trends From 1995 to 2015
1e8
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Boxoffice Gross (per $10,000,000)
1995
2000
2005
2010
2015
Year
Action
Biography
Comedy
Drama
Horror


IMDB Score Comparison per Genre
IMDb Score
9
8
7
6
5
4
3
2
Comedy
Action
Biography
Drama
Horror
Genres

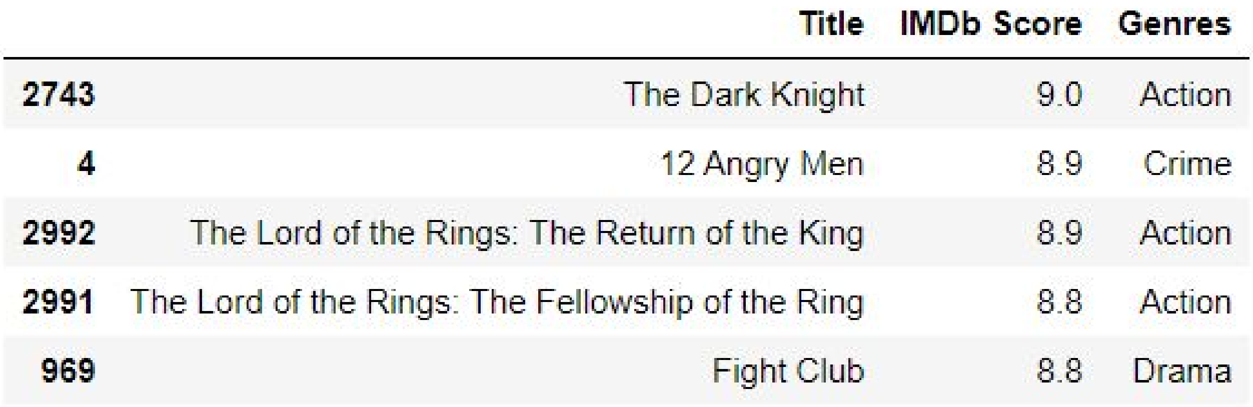
Fg 6



Fg 7



	Title	IMDB Score	Gross	Budget	Return of Investr
1961	Paranormal Activity	6.3	$107,917,283.00	$15,000.00	719,248.
2653	The Blair Witch Project	6.4	$140,530,114.00	$60,000.00	234,016.
2814	The Gallows	4.2	$22,757,819.00	$100,000.00	22,557.
2517	Super Size Me	7.3	$11,529,368.00	$65,000.00	17,537.
1352	In the Company of Men	7.3	$2,856,622.00	$25,000.00	11,226.



	Title	IMDb Score	Genres
2743	The Dark Knight	9.0	Action
4	12 Angry Men	8.9	Crime
2992	The Lord of the Rings: The Return of the King	8.9	Action
2991	The Lord of the Rings: The Fellowship of the Ring	8.8	Action
969	Fight Club	8.8	Drama