Statistical Analysis of 5000+ IMDB Movie Database using SAS

Bala Rohit Yeruva 667341113

University of Illinois at Chicago (UIC)

Fall 2016

BUSINESS PROBLEM:

Film Production Companies often invest a lot of money in making films, doing promotions and various other aspects before the release of the movie. Their return on investment (ROI) depends on how well the movie is acclaimed as good by the audience. However, at the box office, every movie doesn't fare as well as the producers want it to be. Hence, the Production Companies can't be sure if their investment in a movie will always be fruitful.

RESEARCH QUESTION:

Apart from the story line, what are the attributes of a movie that makes it successful?

DATASET:

The dataset has been obtained from Kaggle competitions website [1]. The dataset contains information about 5000+ movies scrapped from IMDB website. It has about 28 attributes (13 numerical, 15 categorical) including the names of the director, the actors and the IMDB scores and other attributes of a movie. We have considered IMDB Score to be our dependent variable and all the other, as our independent variables.

DATA PRE-PROCESSSING & MANIPULATION:

Selecting Variables: Out of the 28 variables, a few attributes such as IMDB link and movie name were deemed unimportant and hence we haven't considered them for our analysis. Further, variables such as Gross, Critic Reviews, User Reviews and Voted Users are known only after the movie is released and hence, even these variables we not considered. We were then left with 9 quantitative variables and 13 categorical variables. Further, out of the 5000+ movies, we have remove around 100 odd movies which had very less information with too many missing values. We have then created an index variable just for referring each movie.

Imputing Missing Data: Quantitative variables such as Director Facebook likes, Actor Facebook likes and the budget etc., were replaced by their corresponding medians. Categorical variables such as the name of the director, actors and the movie's content rating etc., were replaced by their corresponding modes.

Creating Dummy Variables: Variables such as Genre have multiple values for each movie and hence we have created 21 indicator variables for all the genre types. We have then created a dummy categorical variable, Score Rating, which discretizes the IMDB Rating into the following categories:

IMDB Rating	**Score Rating**
Low - 5	Poor
5 – 6	Below Average
6 – 7	Above Average
7 - High	Good

Variable Transformations: Variables such as Color of the movie were transformed from categorical variable to binary variable which just indicates if the movie is in color or in black & white.

Removing Special Characters: Many categorical variables such as the Director and the Actors' names, were appended by special characters during the data extraction from the IMDB Database. A few of the names also contain alphabet from other languages. All these special characters were removed or replaced with appropriate English Alphabet.

After cleaning and processing the data, the following are our final variables:

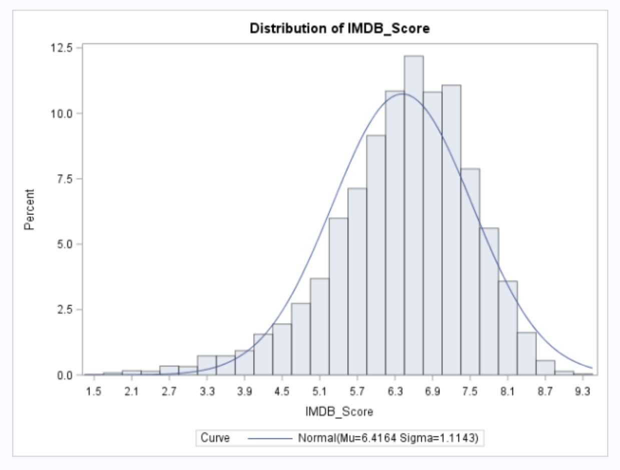
Quantitative Variables: Budget, Movie Facebook Likes, Director Facebook Likes, Actor 1 Facebook Likes, Actor 2 Facebook Likes, Actor 3 Facebook Likes, Total Cast Facebook Likes, Duration, Number of Faces on the Poster.

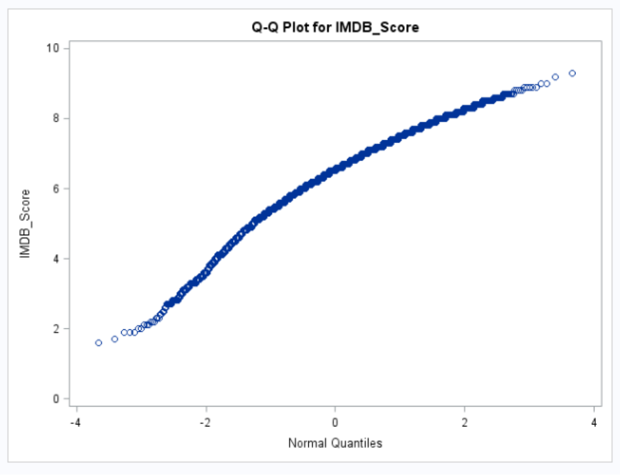
Categorical Variables: Score Rating, Director Name, Actor 1 Name, Actor 2 Name, Actor 3 Name, Language, Country, Year, Content Rating, Aspect Ratio and Plot Keywords.

Binary Variables: 21 Genre Indicator Variables and color.

STATISTICAL ANALYSIS:

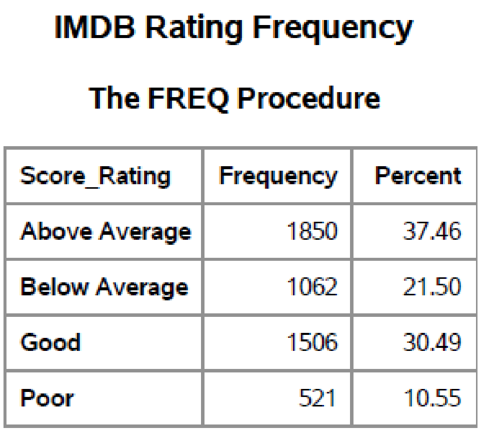
Distribution of the dependent variable: IMDB Score





Many of the statistical methods or significance tests either require or perform well when the corresponding dependent variable have a normal distribution. Fortunately, from the above plot, we could observe that the distribution of the dependent variable is approximately normal with slight left skewness.

Frequency of discretized dependent variable: Score Rating



IMDB Rating Frequency

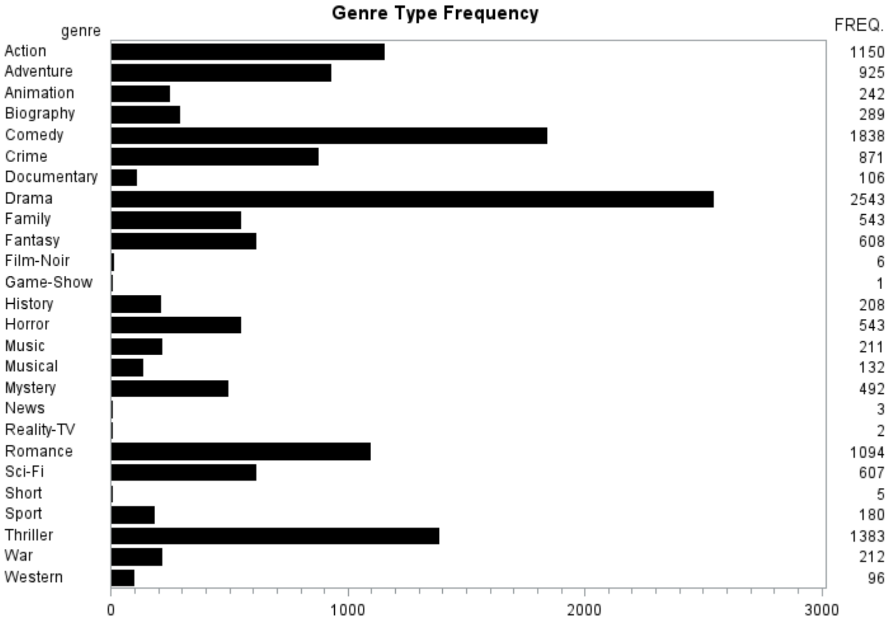
The FREQ Procedure

Score_Rating	Frequency	Percent
Above Average	1850	37.46
Below Average	1062	21.50
Good	1506	30.49
Poor	521	10.55

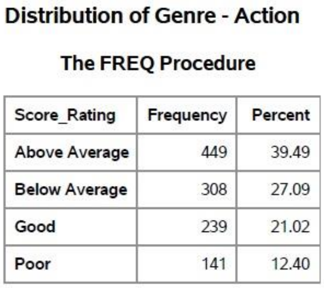
From the above frequency table, we could observe that the data is almost evenly distributed into the four different levels except for the level, ‘Poor’, which anyhow is not as important as the other levels since the main objective of our analysis is to determine the attributes of the movie which makes it fall in the top 2 or 3 levels.

Analysis of Categorical Variables:

1. Analysis of Genre Type:



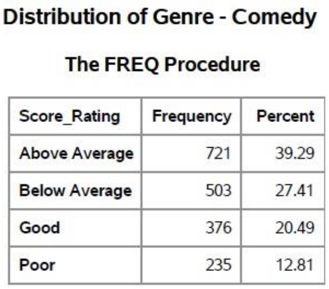
From the frequency of the genre type, we could observe that Drama, Comedy, Thriller, Action and Romance are among the most frequent type of the genres among the available movies. To investigate further, we check the frequency of each genre type within the categorical variable: Score Rating.



Distribution of Genre - Action

The FREQ Procedure

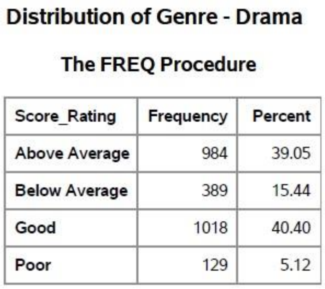
Score_Rating	Frequency	Percent
Above Average	449	39.49
Below Average	308	27.09
Good	239	21.02
Poor	141	12.40



Distribution of Genre - Comedy

The FREQ Procedure

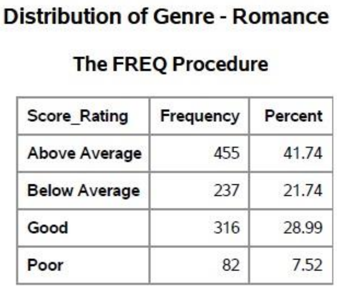
Score_Rating	Frequency	Percent
Above Average	721	39.29
Below Average	503	27.41
Good	376	20.49
Poor	235	12.81



Distribution of Genre - Drama

The FREQ Procedure

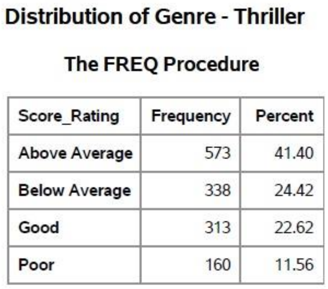
Score_Rating	Frequency	Percent
Above Average	984	39.05
Below Average	389	15.44
Good	1018	40.40
Poor	129	5.12



Distribution of Genre - Romance

The FREQ Procedure

Score_Rating	Frequency	Percent
Above Average	455	41.74
Below Average	237	21.74
Good	316	28.99
Poor	82	7.52



Distribution of Genre - Thriller

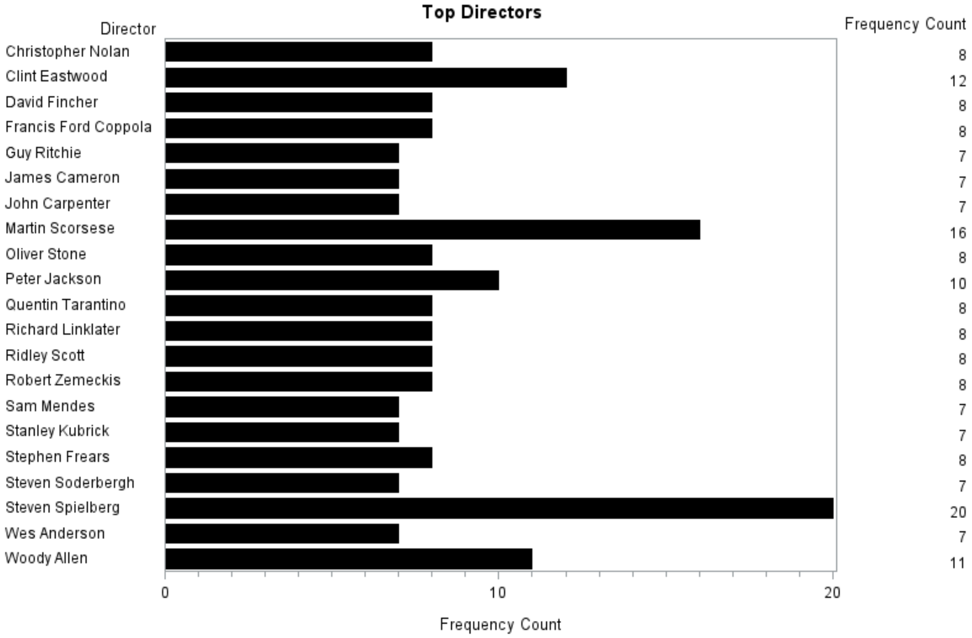
The FREQ Procedure

Score_Rating	Frequency	Percent
Above Average	573	41.40
Below Average	338	24.42
Good	313	22.62
Poor	160	11.56

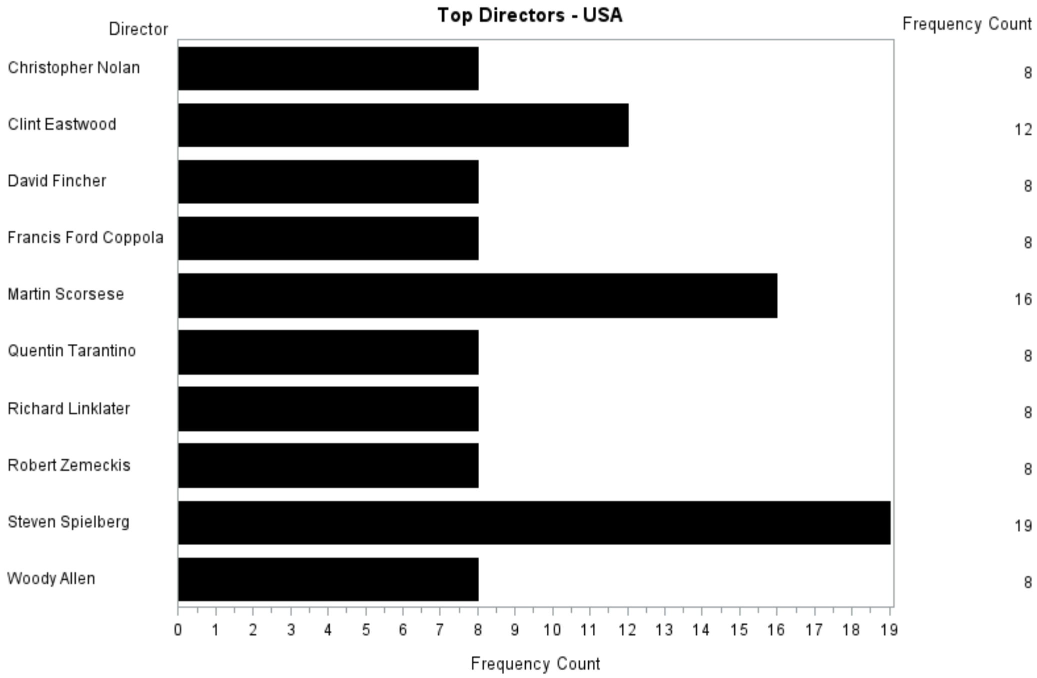
We could observe that genre type, Action, have more Above Average & Below Average movies. A similar pattern is observed for Comedy and Thriller as well. However, for Drama, there are more number of Good movies and relatively very less number of poor movies when compared with the other genres. Even for Romance, there are very less number of poor movies and an evenly distributed average and good movies. Therefore, we could conclude that movies from the genre Drama and Romance have relatively, a higher chance to be a successful movie.

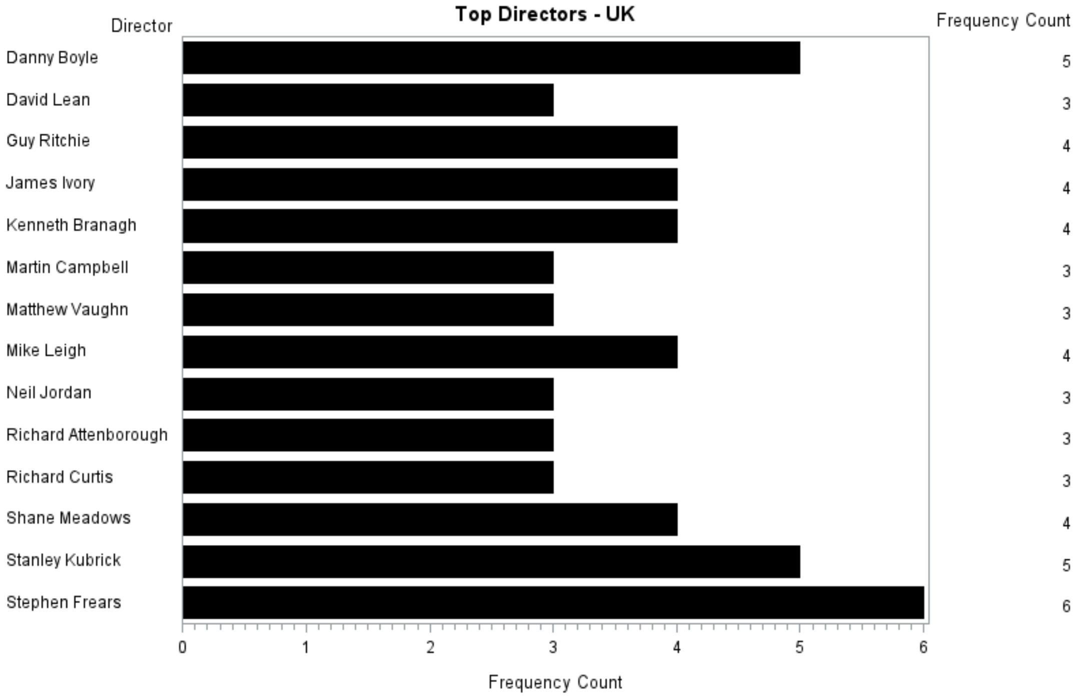
2. Analysis of Directors:

There were about 800 movie directors in total from the dataset. Therefore, we have segregated only those directors who have directed the films which were rated as good. The below figure shows the top 20 directors and the number of movies that they directed that were deemed as good.

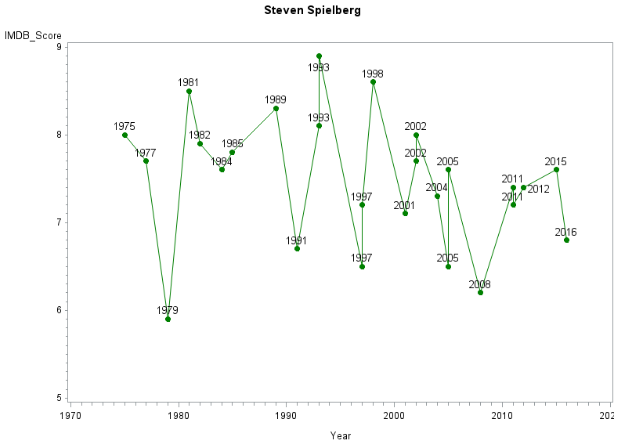


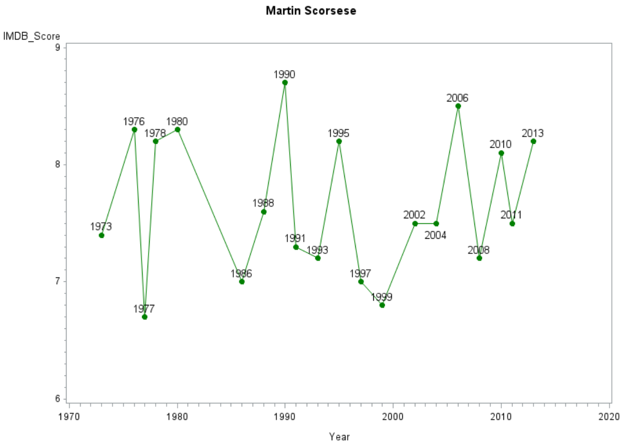
We observe that Steven Spielberg, Martin Scorsese, Clint Eastwood and Woody Allen directed more than ten good movies. This could be considered to be quite an achievement. Further, we divide the movies based on the country where they were released. For our study, we have segregated only those movies which were released in USA and UK as these are the two countries in which most of the movies are released.

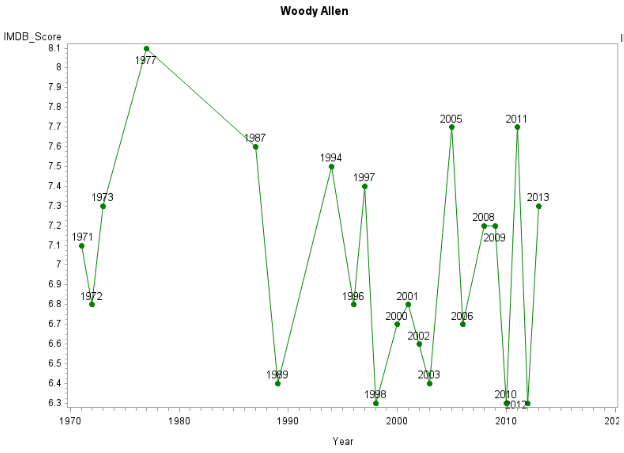


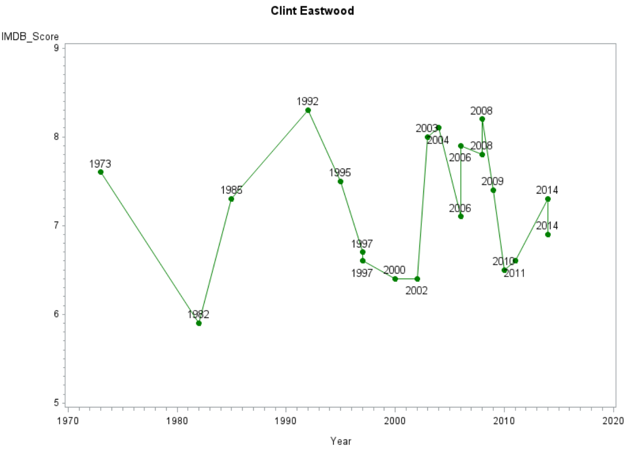


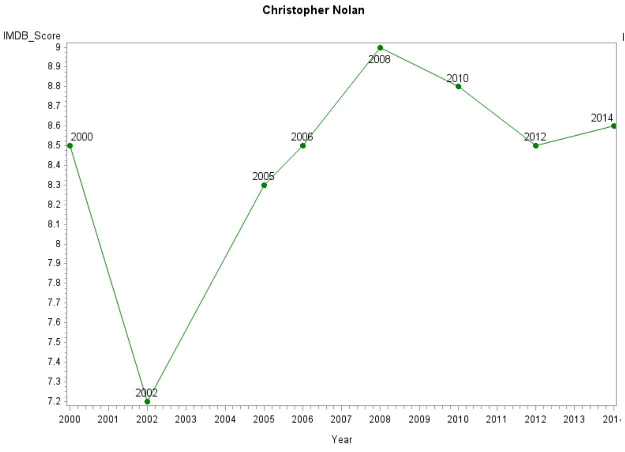
We could observe that the number of good movies directed in USA is relatively high than that in the UK. This could be because of the fact that IMDB is used frequently by the Americans. Hence it could be the result of this confounding variable. Moving on, we have considered the top seven directors to analyze the trend of the good movies directed by them.

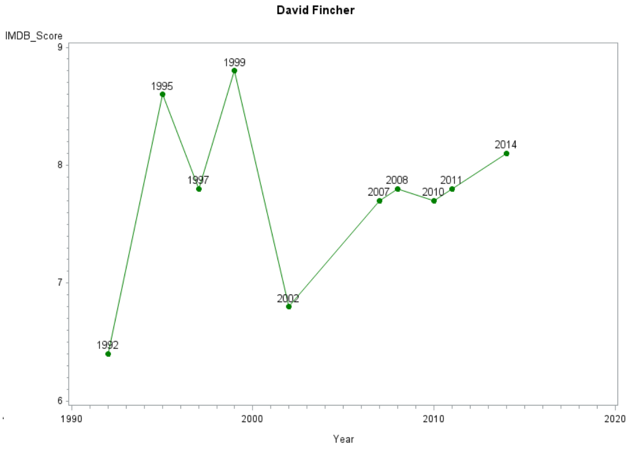








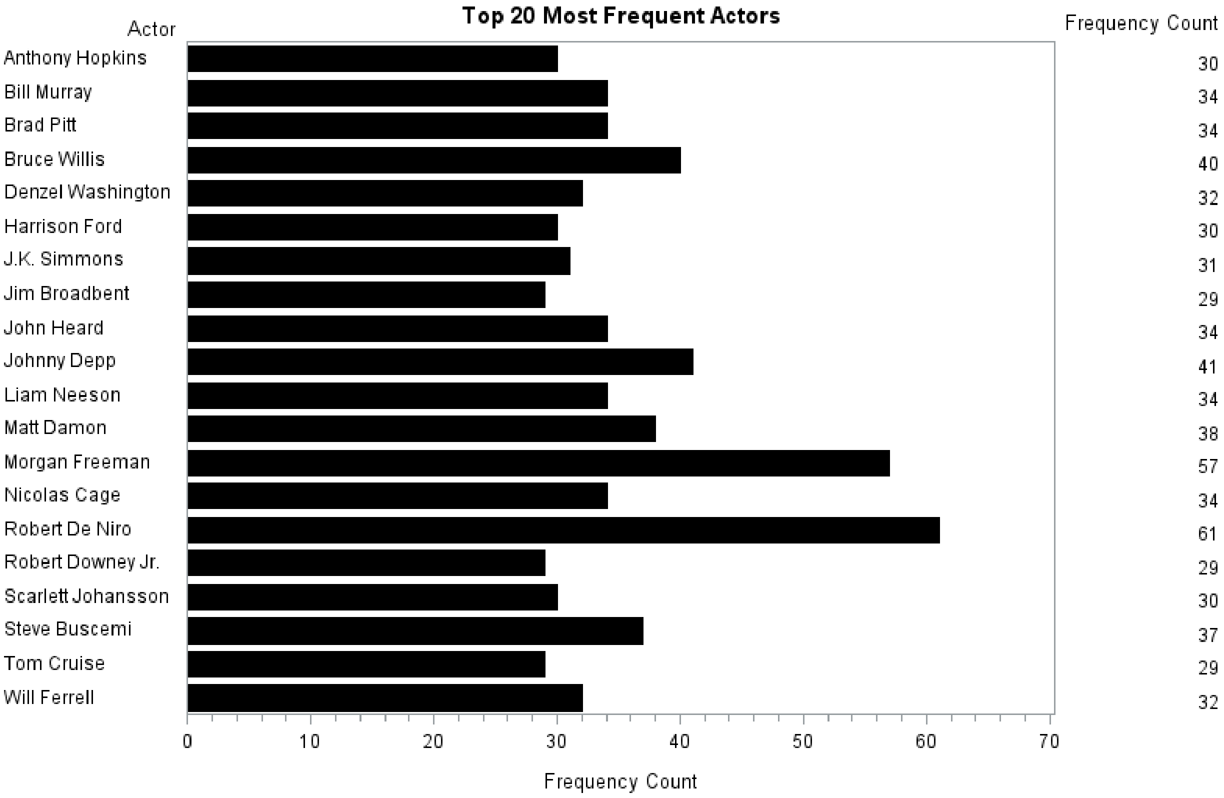




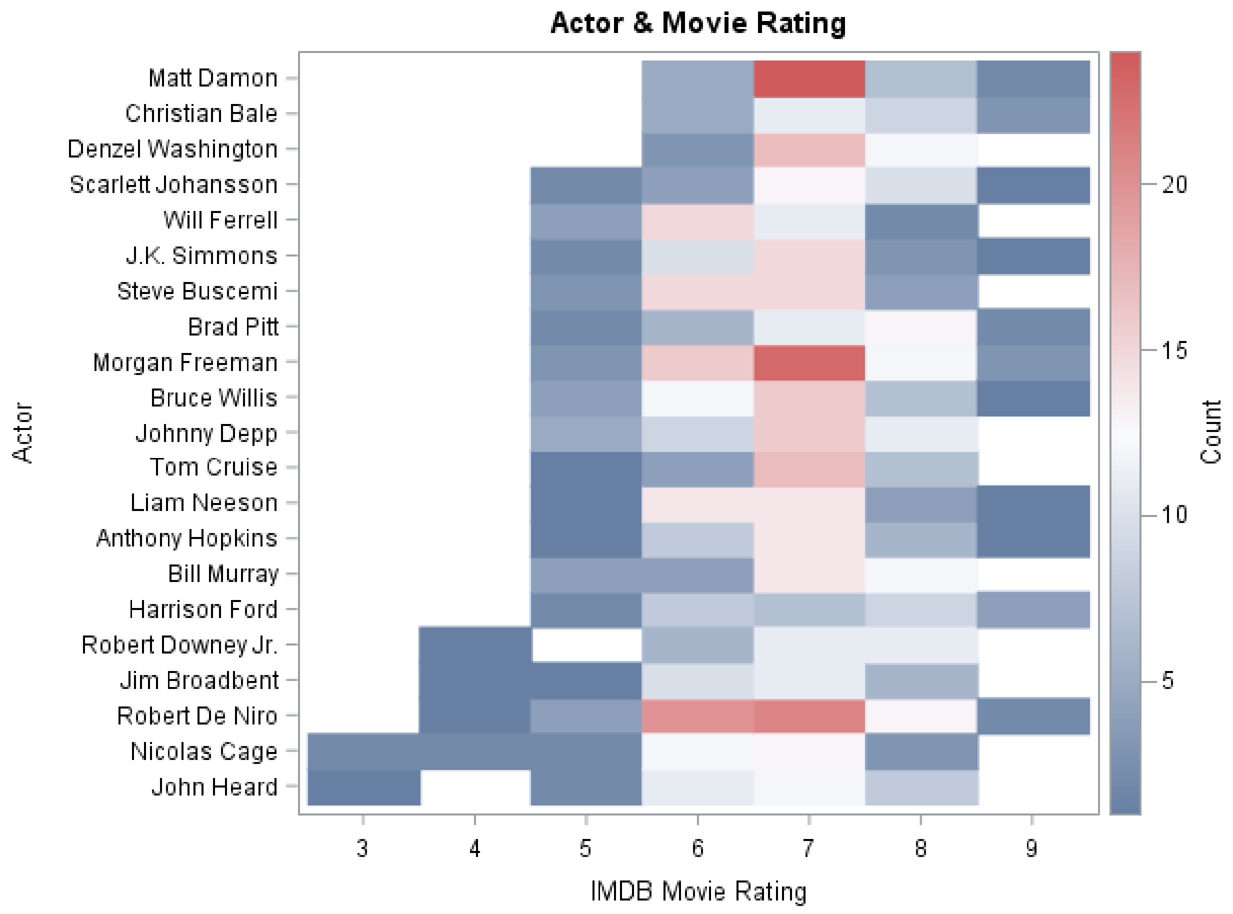
From the above plots, we could see that even though Steven Spielberg directed most number of good movies, their rating is too fluctuating and most of his latest movies only have any average rating around 7. The same could be told about Martin Scorsese, Clint Eastwood and Woody Allen. However, Christopher Nolan directed movies with often high rating with only one exception in 2002. Further, David Fincher appears to be on a raise of late, as the rating of his movies increase almost every time he directs a new movie. Therefore, we could conclude that out of the top directors, Christopher Nolan and David Fincher appear to be more successful than the others.

3. Analysis of Actors:

We have three different variables indicating up to three actors for each movie. For our study we have clubbed all the actors into one categorical variable which eventually had more than 6000 actors from different movies. The following are the top 20 most frequent actors and the number of movies that they appeared.



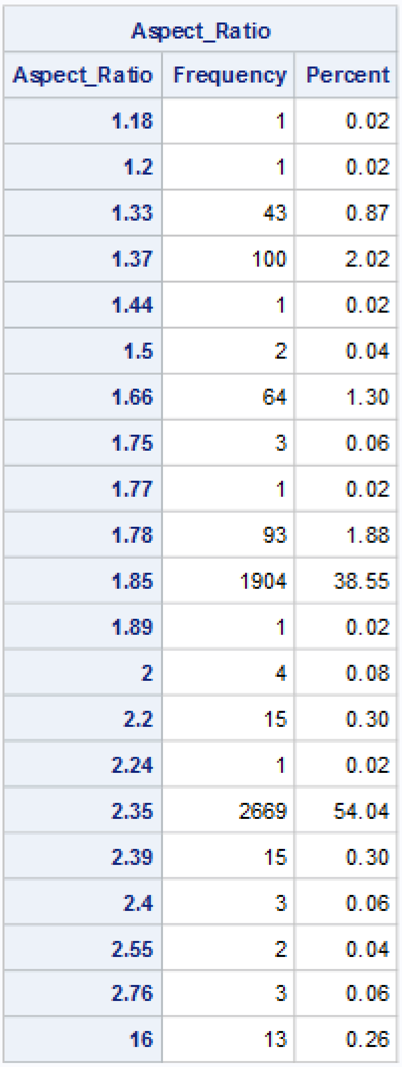
Robert De Niro, Morgan Freeman seems to be the most famous and frequent actors as they appeared in more than 55 movies (all levels). Bruce Wills, Johnny Depp, Matt Damon and Steve Buscsemi are the next frequent actors with featuring in more than 35 movies each. To further analyze the rating of the movies in which they acted, we have built a heat map as shown in the following figure.



From the above heat map, we could observe that majority of the movies including Robert De Niro and Morgan Freeman as well, have an IMDB Rating around 7. However, actors such as Matt Damon, Christian Bale, Scarlett Johansson, J.K. Simmons, Brad Pitt, Bruce Wills, Liam Neeson, Anthony Hopkins and Harrison Ford featured in the movies which were rated between 9 to 10. Therefore, movies with the above mentioned actors could be considered to have a higher probability than the others, to be deemed as a successful movie.

4. Analysis of Aspect Ratio:

Aspect Ratio is the ratio of the width to the height of the screen. Analysis on the effects of this variable on the corresponding movie rating could determine if the users would like to prefer any specific screen size over the others. Below is the frequency distribution of this variable across all the movies in the dataset.

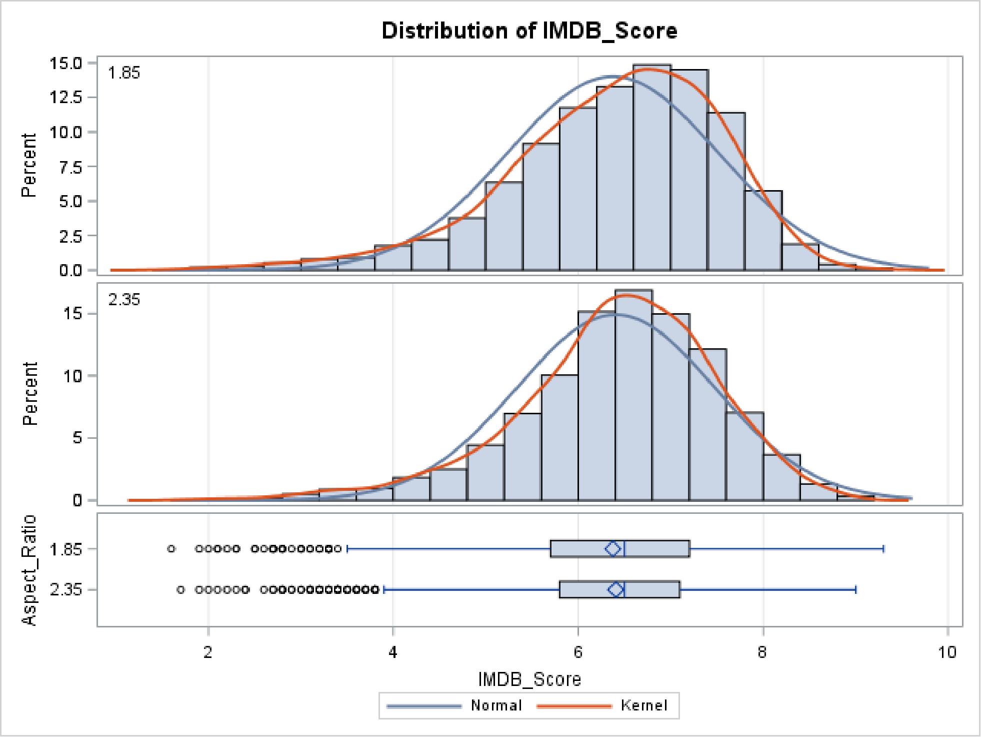


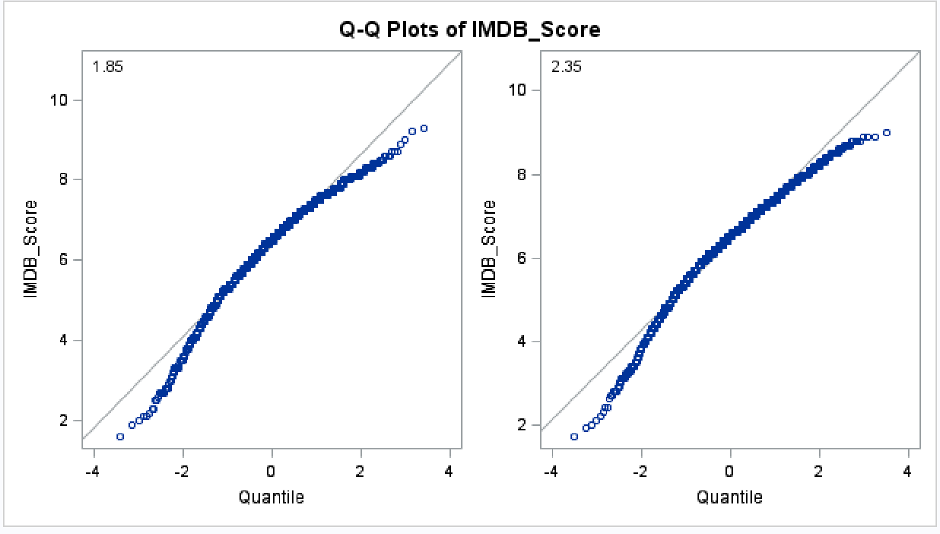
Aspect_Ratio

Aspect_Ratio	Frequency	Percent
1.18	1	0.02
1.2	1	0.02
1.33	43	0.87
1.37	100	2.02
1.44	1	0.02
1.5	2	0.04
1.66	64	1.30
1.75	3	0.06
1.77	1	0.02
1.78	93	1.88
1.85	1904	38.55
1.89	1	0.02
2	4	0.08
2.2	15	0.30
2.24	1	0.02
2.35	2669	54.04
2.39	15	0.30
2.4	3	0.06
2.55	2	0.04
2.76	3	0.06
16	13	0.26

We could observe that around more 90% of the movies either have an aspect ratio of 1.85 or 2.35. Therefore, we have considered only these two levels and proceeded further to investigate if they have an effect on the movie rating.

We have performed a two-sample t-test to investigate if the mean IMDB score for the two levels are equal. However, for performing a two sample t-test, we need the two populations to have same variance and they must be normally distributed. Therefore, assuming that these two populations have the same variance, we could observe that both the samples are approximately normally distributed.



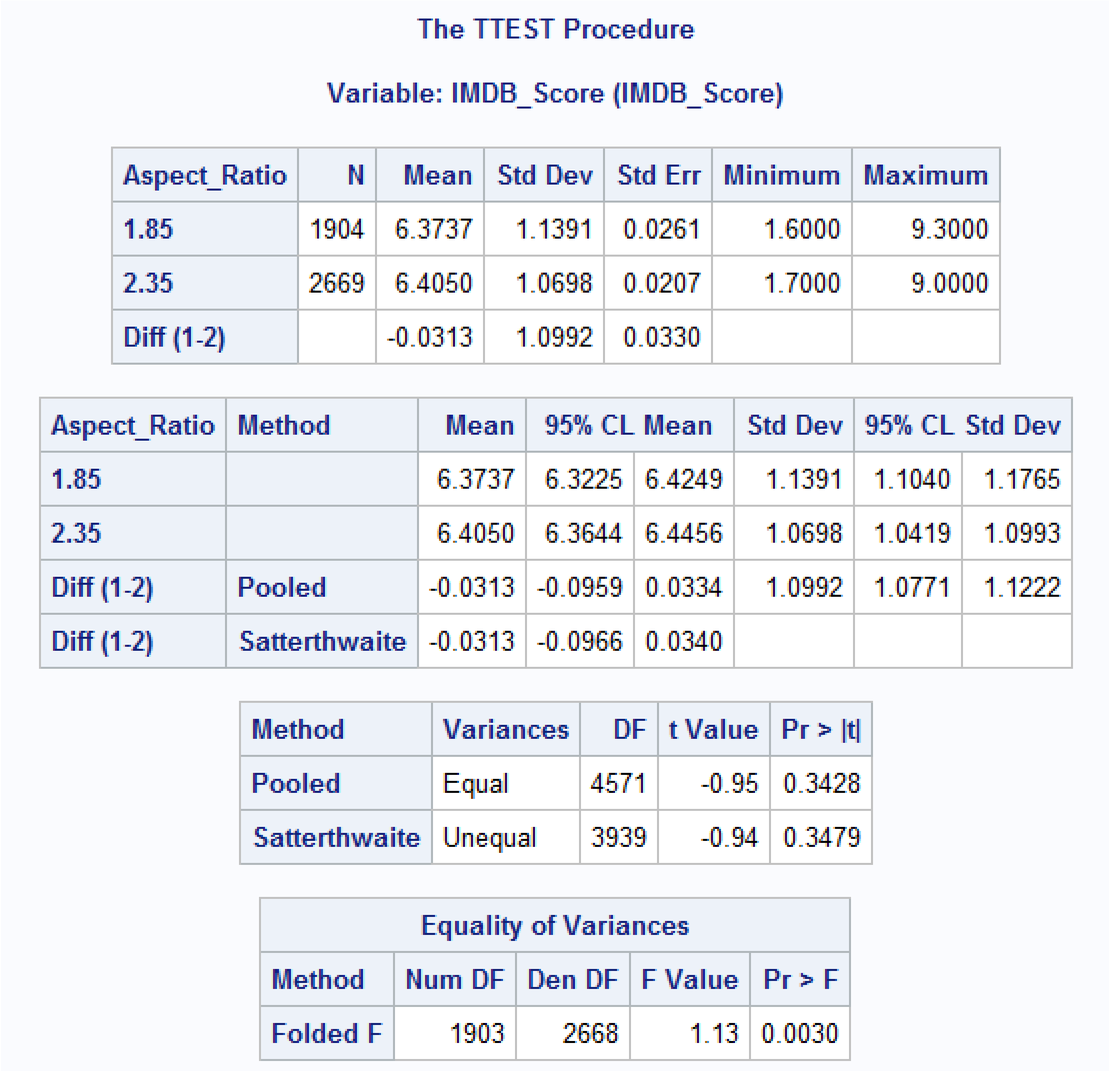


Two Sample t-test:

$$H_0 : \mu_{\text{Aspect Ratio} = 1.85} = \mu_{\text{Aspect Ratio} = 2.35}$$

$$H_a : \mu_{\text{Aspect Ratio} = 1.85} \neq \mu_{\text{Aspect Ratio} = 2.35}$$

Where $\mu_{\text{Aspect Ratio} = 1.85}$ & $\mu_{\text{Aspect Ratio} = 2.35}$ are the mean IMDB scores for movies with Aspect Ration equal to 1.85 and 2.35 respectively. The results of the two sample t-test are as follows:



The TTEST Procedure

Variable: IMDB_Score (IMDB_Score)

Aspect_Ratio	N	Mean	Std Dev	Std Err	Minimum	Maximum
1.85	1904	6.3737	1.1391	0.0261	1.6000	9.3000
2.35	2669	6.4050	1.0698	0.0207	1.7000	9.0000
Diff (1-2)		-0.0313	1.0992	0.0330		

Aspect_Ratio	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1.85		6.3737	6.3225	6.4249	1.1391	1.1040	1.1765
2.35		6.4050	6.3644	6.4456	1.0698	1.0419	1.0993
Diff (1-2)	Pooled	-0.0313	-0.0959	0.0334	1.0992	1.0771	1.1222
Diff (1-2)	Satterthwaite	-0.0313	-0.0966	0.0340			

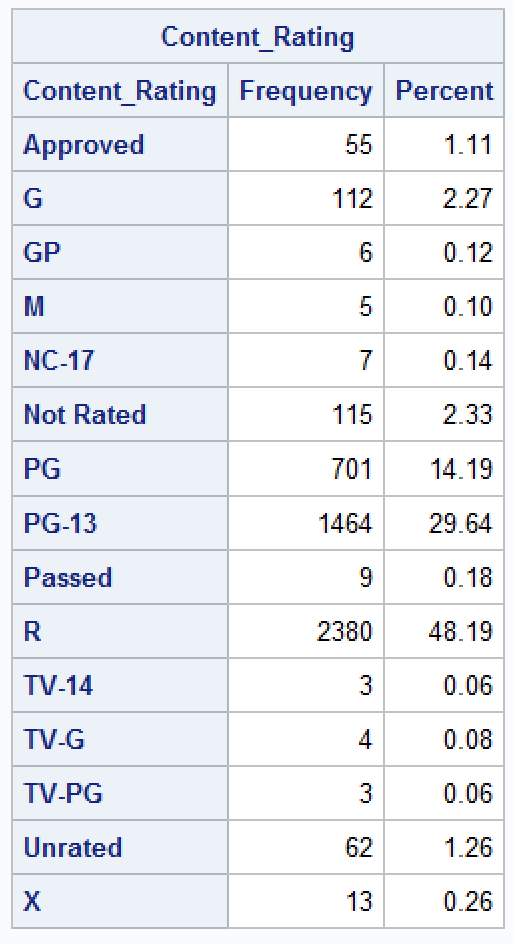
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	4571	-0.95	0.3428
Satterthwaite	Unequal	3939	-0.94	0.3479

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	1903	2668	1.13	0.0030

We could observe that the probability Pr > |t| is equal to 0.3428 which is nowhere significant enough to reject the Null Hypotheses. This means that the probability for two samples having mean IMDB Scores as observed when their populations have equal mean IMDB scores is 34.28%. Thus we could conclude that Aspect Ratio would not have any significant effect on the movie rating.

5. Analysis of Content Rating:

A content rating is an attribute of a movie which tells what age group is suitable to watch the movie. Below is the frequency distribution of the content rating for all the movies in our dataset:

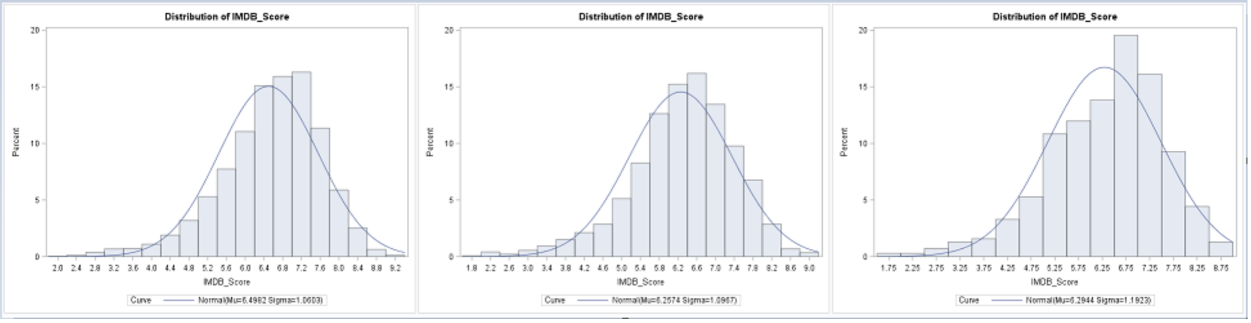


Content_Rating		
Content_Rating	**Frequency**	**Percent**
Approved	55	1.11
G	112	2.27
GP	6	0.12
M	5	0.10
NC-17	7	0.14
Not Rated	115	2.33
PG	701	14.19
PG-13	1464	29.64
Passed	9	0.18
R	2380	48.19
TV-14	3	0.06
TV-G	4	0.08
TV-PG	3	0.06
Unrated	62	1.26
X	13	0.26

From the frequency distribution we could see that around 90% of the movies have either R, PG-13 or PG as their content rating. Therefore, we have considered only these three categories to investigate if the movies with these content rating have an equal average IMDB score. Since there are more than two levels for this variable, we cannot perform a two sample t-test and hence we need to consider a statistical method that allows to equate the means of more than two populations. Hence, the best method to do this analysis would be a one way Anova.

One-way Anova:

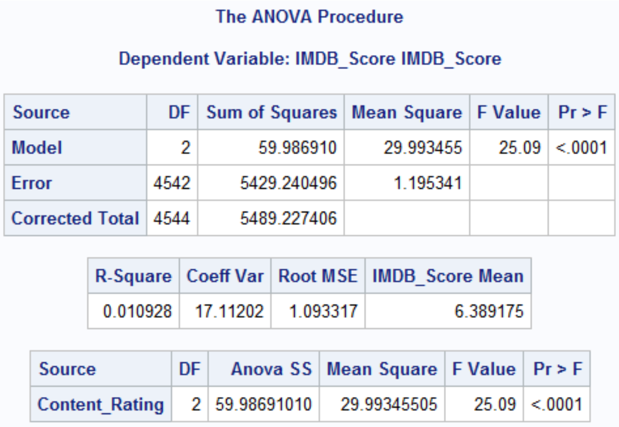
Even One-way Anova has similar assumptions as that of the two-sample T-test. Therefore, assuming that the population of these three categories have equal variance, we could observe from the below figure, that all the three sample populations are approximately normal and hence we can proceed with our analysis.



$$H_0 : \mu_{\text{Content Rating = R}} = \mu_{\text{Content Rating = PG-13}} = \mu_{\text{Content Rating = PG}}$$

$$H_a : \mu_{\text{Content Rating = R}} \neq \mu_{\text{Content Rating = PG-13}} \neq \mu_{\text{Content Rating = PG}}$$

Where $\mu_{\text{Content Rating = R}}$, $\mu_{\text{Content Rating = PG-13}}$ & $\mu_{\text{Content Rating = PG}}$ are the mean IMDB scores for movies with Content Rating equal to R, PG-13 and PG respectively. The results of the one-way Anova are as follows:

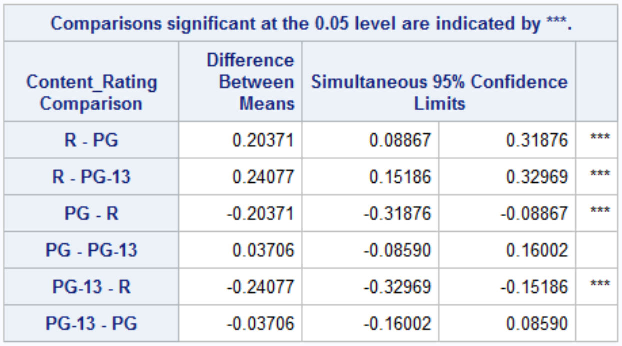

The ANOVA Procedure

Dependent Variable: IMDB_Score IMDB_Score

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	59.986910	29.993455	25.09	<.0001
Error	4542	5429.240496	1.195341		
Corrected Total	4544	5489.227406			

R-Square	Coeff Var	Root MSE	IMDB_Score Mean
0.010928	17.11202	1.093317	6.389175

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Content_Rating	2	59.98691010	29.99345505	25.09	<.0001


Comparisons significant at the 0.05 level are indicated by ***.

Content_Rating Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
R - PG	0.20371	0.08867	0.31876	***
R - PG-13	0.24077	0.15186	0.32969	***
PG - R	-0.20371	-0.31876	-0.08867	***
PG - PG-13	0.03706	-0.08590	0.16002	
PG-13 - R	-0.24077	-0.32969	-0.15186	***
PG-13 - PG	-0.03706	-0.16002	0.08590	

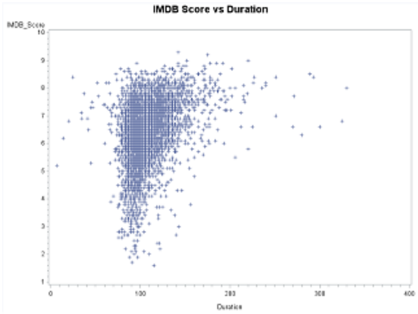
From the above table, we could observe that the probability Pr > |t| for an F statistic to be 25.09 is less than 0.0001 which is significant enough to reject the Null Hypotheses. From the Scheffe's test, we could also see that the score for 'R' rated movies is different from the other two ratings and is higher than that of 'PG' or 'PG-13' rated movies. Thus, we could state that Content Rating of a movie does have an effect on its IMDB score and a movie with 'R' rating is more likely to have a higher IMDB score.

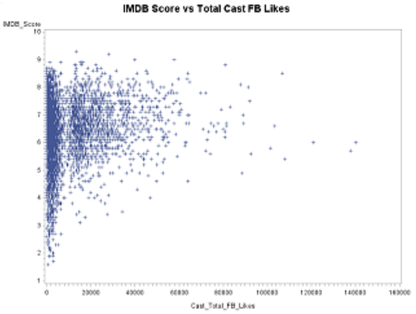
<u>Analysis of Quantitative Variables</u>:

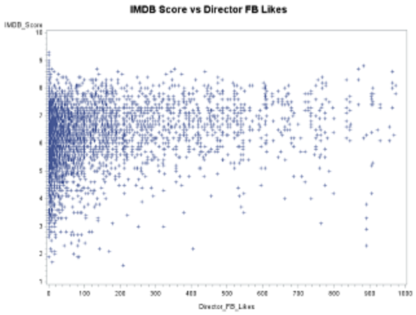
To estimate the IMDB score based on the quantitative attributes of a movie, we have built a simple linear model. Before using all the available independent variables, we have calculated the correlation coefficients for all them against the dependent variable, IMDB score.

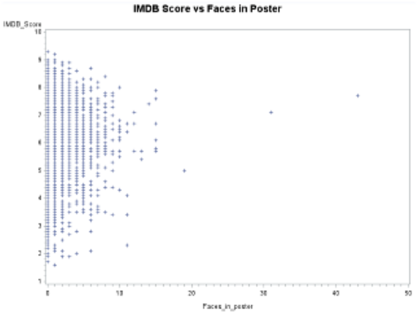
	Budget	Duration	Director FB Likes	Movie FB Likes	Actor 1 FB Likes	Actor 2 FB Likes	Actor 3 FB Likes	Cast FB Likes	Faces on Poster
IMDB Score	0.03029 0.0333	0.34209 <.0001	0.08342 <.0001	0.02587 0.0691	0.03293 0.0207	0.02707 0.0572	0.02691 0.0586	0.08548 <.0001	-0.0696 <.0001

From the above table, we could observe that Duration, Director FB Likes, Cast FB Likes has a relatively high correlation coefficient with the dependent variable. We could also observe that Faces on Poster have a negative correlation which indicates that if there are more the number of faces on the poster, less likely it is to be encouraged by the people. Below are a few of the scatter plots between these correlated variables and the dependent variable.


IMDB Score vs Duration


IMDB Score vs Total Cast FB Likes

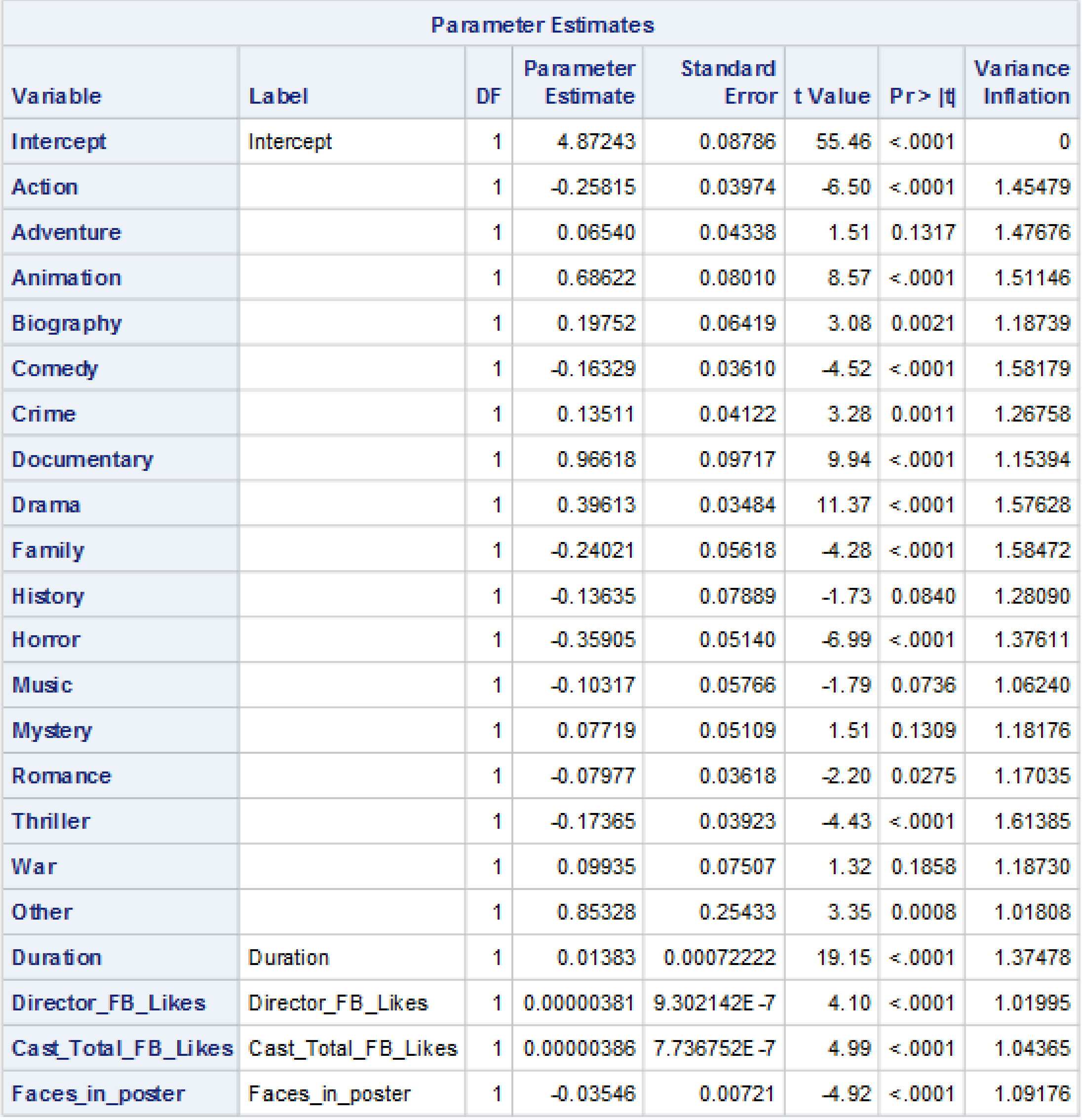

IMDB Score vs Director FB Likes


IMDB Score vs Faces in Poster

From the above plots, even though most of the points are cluttered together, we could observe a slight linear pattern.

Simple Linear Regression Model:

We have then built a simple linear regression model using the above correlated independent variables and the indicator variables created from the genre of the movie. We have used forward selection while choosing the variables and following are the parameter estimates.

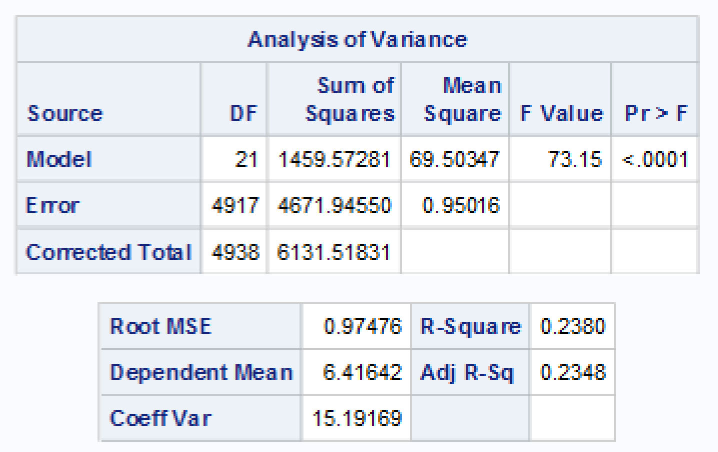


Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	4.87243	0.08786	55.46	<.0001	0
Action		1	-0.25815	0.03974	-6.50	<.0001	1.45479
Adventure		1	0.06540	0.04338	1.51	0.1317	1.47676
Animation		1	0.68622	0.08010	8.57	<.0001	1.51146
Biography		1	0.19752	0.06419	3.08	0.0021	1.18739
Comedy		1	-0.16329	0.03610	-4.52	<.0001	1.58179
Crime		1	0.13511	0.04122	3.28	0.0011	1.26758
Documentary		1	0.96618	0.09717	9.94	<.0001	1.15394
Drama		1	0.39613	0.03484	11.37	<.0001	1.57628
Family		1	-0.24021	0.05618	-4.28	<.0001	1.58472
History		1	-0.13635	0.07889	-1.73	0.0840	1.28090
Horror		1	-0.35905	0.05140	-6.99	<.0001	1.37611
Music		1	-0.10317	0.05766	-1.79	0.0736	1.06240
Mystery		1	0.07719	0.05109	1.51	0.1309	1.18176
Romance		1	-0.07977	0.03618	-2.20	0.0275	1.17035
Thriller		1	-0.17365	0.03923	-4.43	<.0001	1.61385
War		1	0.09935	0.07507	1.32	0.1858	1.18730
Other		1	0.85328	0.25433	3.35	0.0008	1.01808
Duration	Duration	1	0.01383	0.00072222	19.15	<.0001	1.37478
Director_FB_Likes	Director_FB_Likes	1	0.00000381	9.302142E-7	4.10	<.0001	1.01995
Cast_Total_FB_Likes	Cast_Total_FB_Likes	1	0.00000386	7.736752E-7	4.99	<.0001	1.04365
Faces_in_poster	Faces_in_poster	1	-0.03546	0.00721	-4.92	<.0001	1.09176

We see that there are no parameter estimates which have Variance Indicator Factor (VIF) greater than ten and hence we can conclude that there are no collinear variables among the independent variables and thus there is no need for us to further reduce the total number of independent variables.

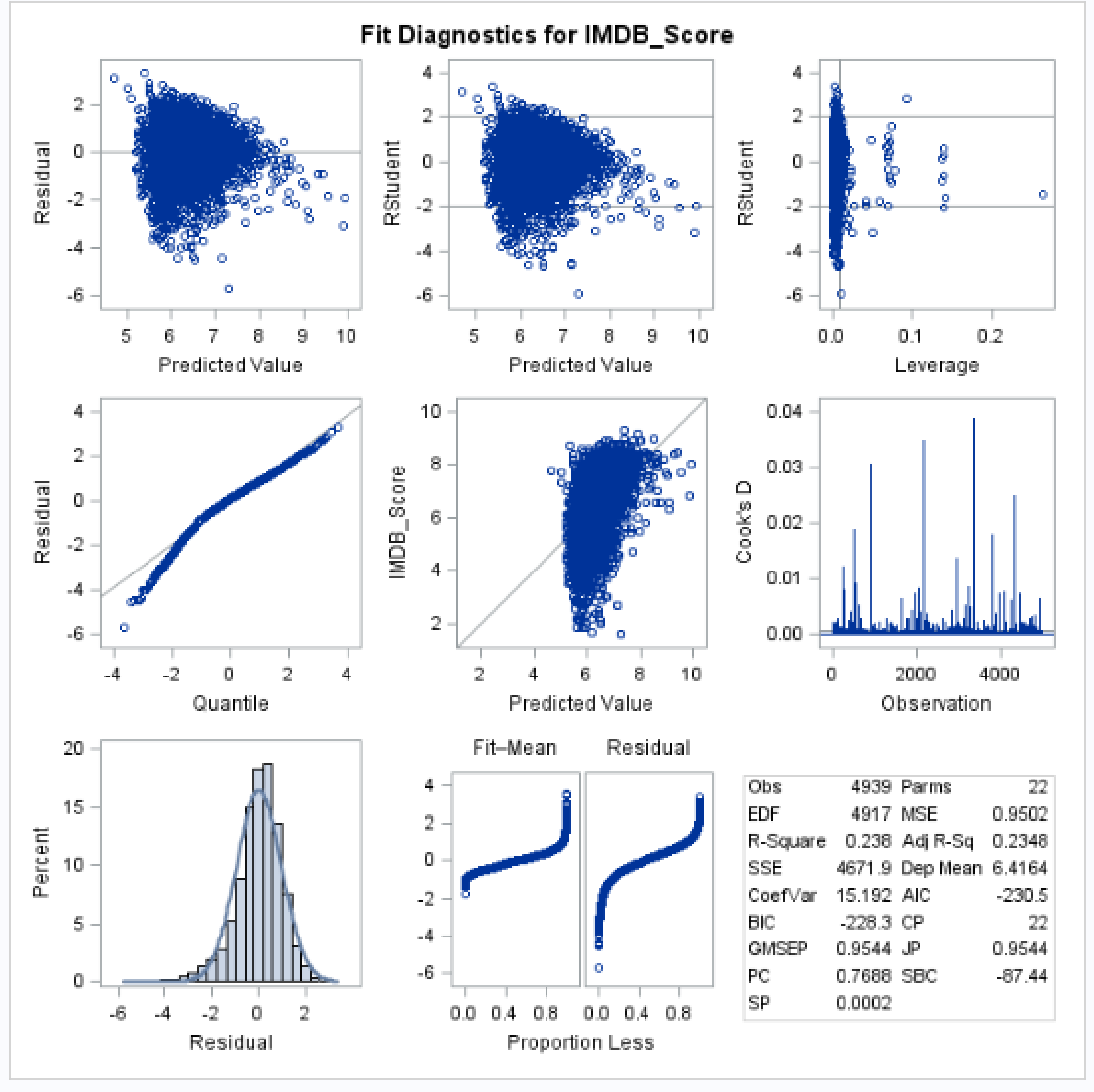
The following plots depict the summary statistics and the residual distribution of our model.



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	21	1459.57281	69.50347	73.15	<.0001
Error	4917	4671.94550	0.95016		
Corrected Total	4938	6131.51831			

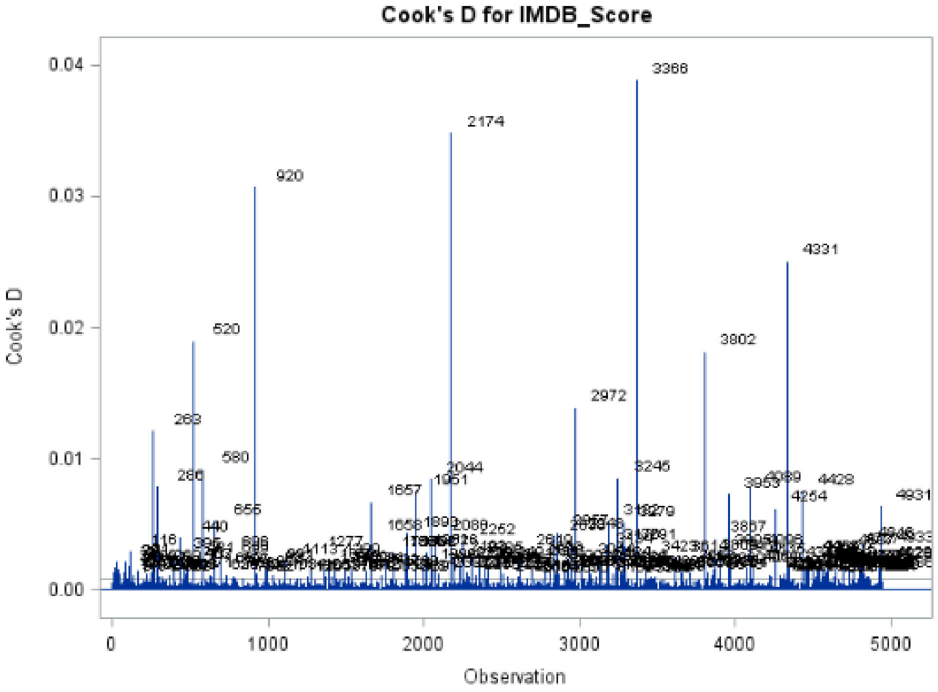
Root MSE	0.97476	R-Square	0.2380
Dependent Mean	6.41642	Adj R-Sq	0.2348
Coeff Var	15.19169		



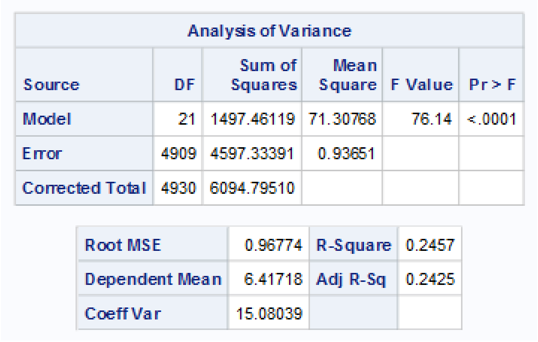
Obs	4939	Parms	22
EDF	4917	MSE	0.9502
R-Square	0.238	Adj R-Sq	0.2348
SSE	4671.9	Dep Mean	6.4164
CoefVar	15.192	AIC	-230.5
BIC	-228.3	CP	22
GMSEP	0.9544	JP	0.9544
PC	0.7688	SBC	-87.44
SP	0.0002		

From the above plots, we could observe that the Adj. R square value is equal to 0.2348 which indicates that only 23.48% of variance in the IMDB Score for a movie is explained by these independent variables. We would expect this, after all the correlations of the independent variables with the dependent variable were all less than 0.5. Further we could observe that the residuals appear to be normal which is essential for a general linear model.

We then observe the Cook's D plot to identify the influential points of the outliers.



From the plot, we have considered the data points 263, 520, 920, 2174, 2972, 3366, 3802 and 4331 as the influential observations and hence we have built the model again after removing these values and observe that the adj. R-square increased from 0.2348 to 0.2425.



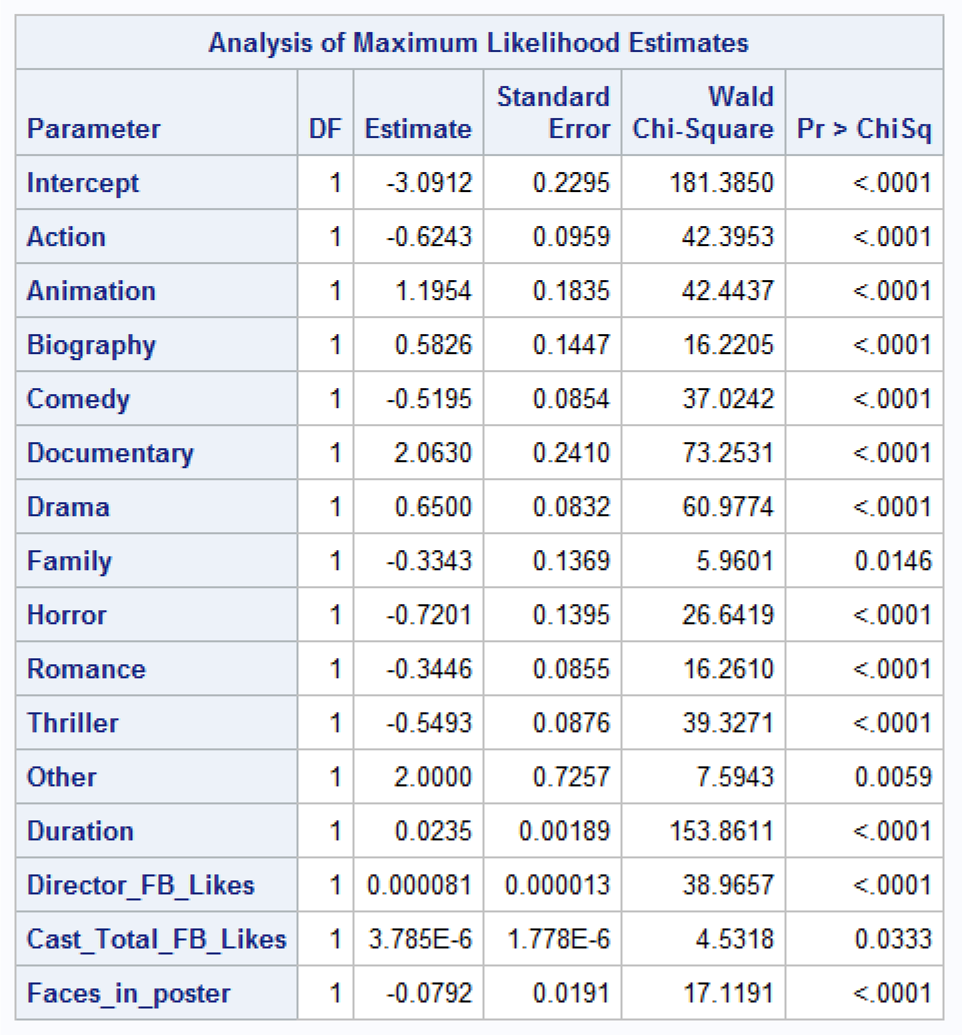
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	21	1497.46119	71.30768	76.14	<.0001
Error	4909	4597.33391	0.93651		
Corrected Total	4930	6094.79510			

Root MSE	0.96774	R-Square	0.2457
Dependent Mean	6.41718	Adj R-Sq	0.2425
Coeff Var	15.08039		

Despite the increase in the R-square value, our simple linear model explains only 24.25% of the variance of the dependent variable, which is still quite less for a good fit.

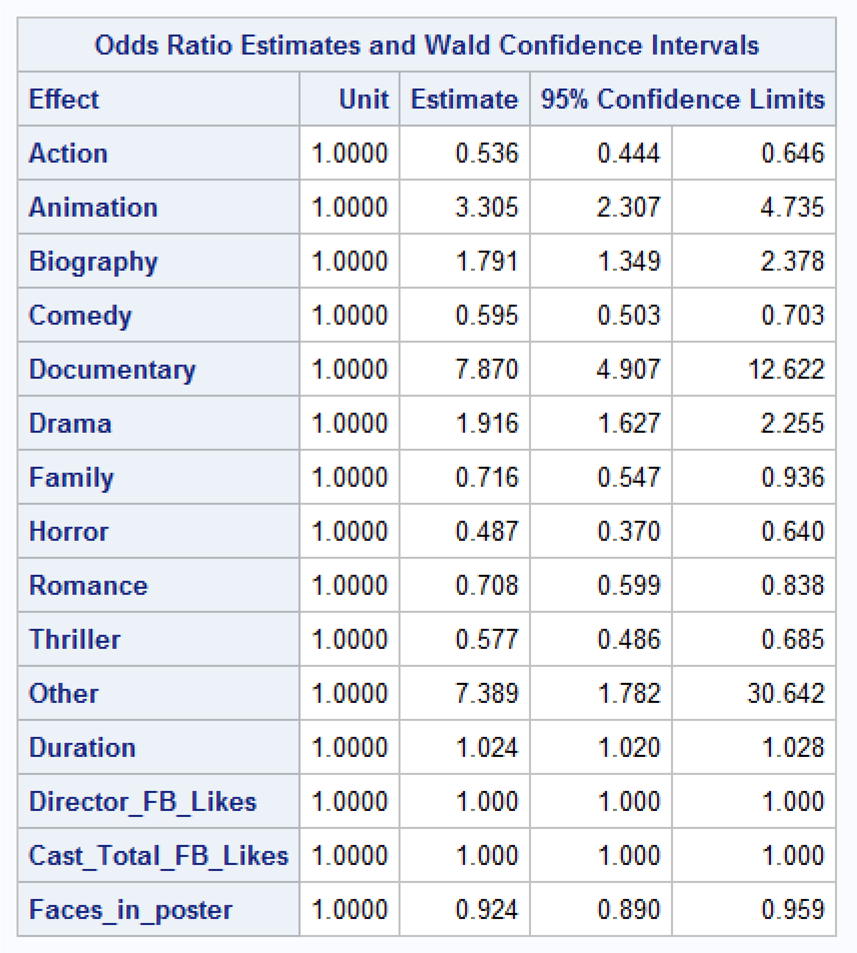
Logistic Regression:

As the above linear model was able to explain only 24.25% of the variance, we have decided to convert the dependent variable into a binary variable and then build a logistic regression. For this, we have created a new variable called 'Success', which takes the value 1 if the rating of the movie is above 7 and 0 otherwise. Using 'Success' as the dependent variable and all the independent variables as used in the linear model along with forward selection procedure, we have built a logistic regression and below are the model results.



Analysis of Maximum Likelihood Estimates

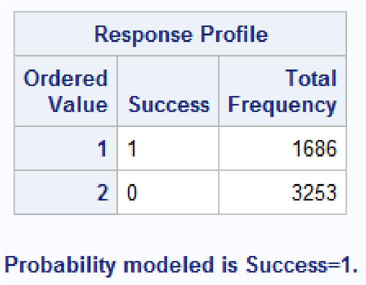
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.0912	0.2295	181.3850	<.0001
Action	1	-0.6243	0.0959	42.3953	<.0001
Animation	1	1.1954	0.1835	42.4437	<.0001
Biography	1	0.5826	0.1447	16.2205	<.0001
Comedy	1	-0.5195	0.0854	37.0242	<.0001
Documentary	1	2.0630	0.2410	73.2531	<.0001
Drama	1	0.6500	0.0832	60.9774	<.0001
Family	1	-0.3343	0.1369	5.9601	0.0146
Horror	1	-0.7201	0.1395	26.6419	<.0001
Romance	1	-0.3446	0.0855	16.2610	<.0001
Thriller	1	-0.5493	0.0876	39.3271	<.0001
Other	1	2.0000	0.7257	7.5943	0.0059
Duration	1	0.0235	0.00189	153.8611	<.0001
Director_FB_Likes	1	0.000081	0.000013	38.9657	<.0001
Cast_Total_FB_Likes	1	3.785E-6	1.778E-6	4.5318	0.0333
Faces_in_poster	1	-0.0792	0.0191	17.1191	<.0001



Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
Action	1.0000	0.536	0.444	0.646
Animation	1.0000	3.305	2.307	4.735
Biography	1.0000	1.791	1.349	2.378
Comedy	1.0000	0.595	0.503	0.703
Documentary	1.0000	7.870	4.907	12.622
Drama	1.0000	1.916	1.627	2.255
Family	1.0000	0.716	0.547	0.936
Horror	1.0000	0.487	0.370	0.640
Romance	1.0000	0.708	0.599	0.838
Thriller	1.0000	0.577	0.486	0.685
Other	1.0000	7.389	1.782	30.642
Duration	1.0000	1.024	1.020	1.028
Director_FB_Likes	1.0000	1.000	1.000	1.000
Cast_Total_FB_Likes	1.0000	1.000	1.000	1.000
Faces_in_poster	1.0000	0.924	0.890	0.959

From the above Odd Ratio Estimates, we could see that Animation, Documentary and Other genres have the highest parameter estimates. However Documentary and Other have a huge confidence interval when compared with Animation. We could also see that the parameter estimates for Director FB Likes and Cast FB Likes are equal to 1. This indicates that they have equal number of successful and unsuccessful movies.

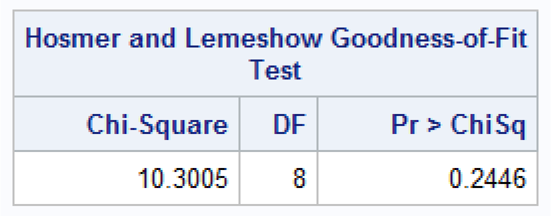
We could also confirm from the following output that we have built the model to estimate the probability of the movie to be successful.



Response Profile		
Ordered Value	Success	Total Frequency
1	1	1686
2	0	3253

Probability modeled is Success=1.

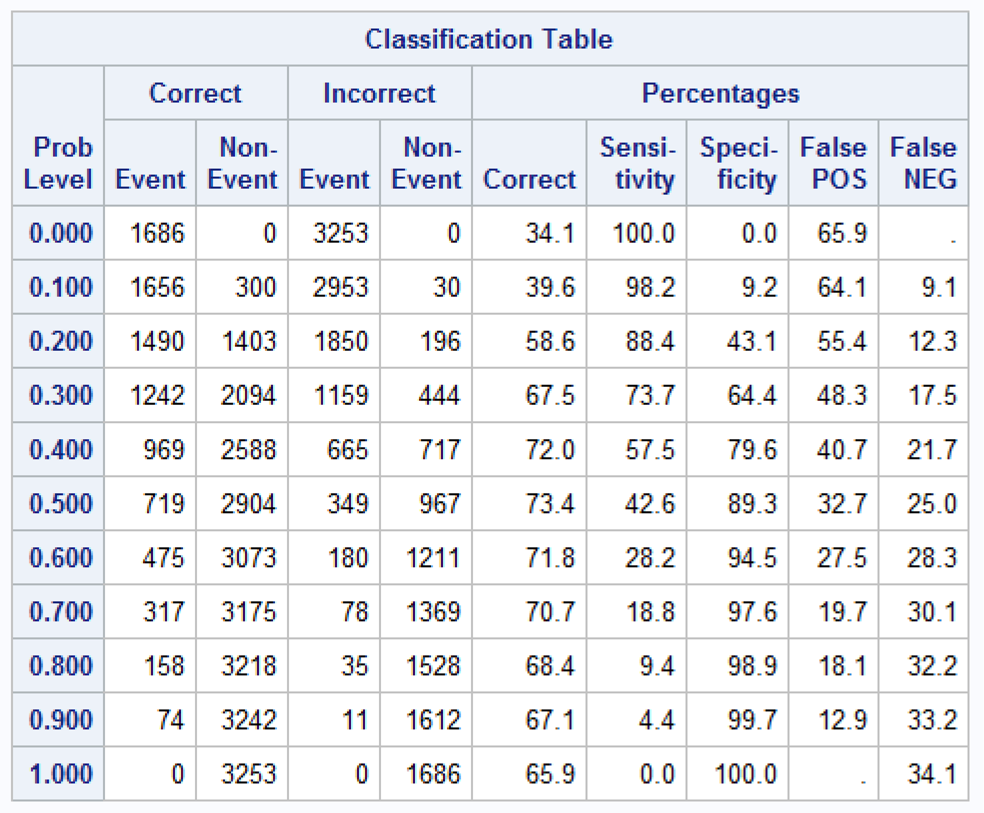
From the below Goodness of Fit test, we could see that we have a high p-value and therefore we won't be rejecting our null hypotheses which states that our data fits the model.



Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
10.3005	8	0.2446

The following is the classification table which provides the actual and predicted values of the outcome.

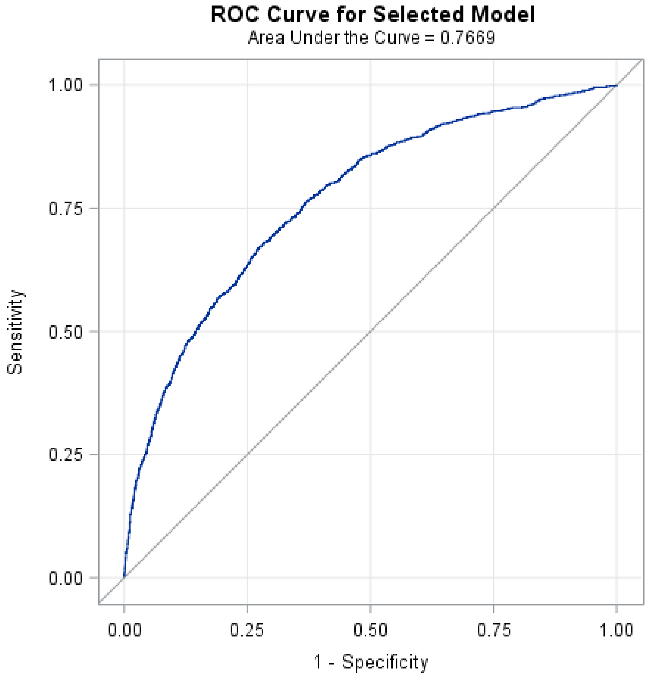


Classification Table

	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.000	1686	0	3253	0	34.1	100.0	0.0	65.9	.
0.100	1656	300	2953	30	39.6	98.2	9.2	64.1	9.1
0.200	1490	1403	1850	196	58.6	88.4	43.1	55.4	12.3
0.300	1242	2094	1159	444	67.5	73.7	64.4	48.3	17.5
0.400	969	2588	665	717	72.0	57.5	79.6	40.7	21.7
0.500	719	2904	349	967	73.4	42.6	89.3	32.7	25.0
0.600	475	3073	180	1211	71.8	28.2	94.5	27.5	28.3
0.700	317	3175	78	1369	70.7	18.8	97.6	19.7	30.1
0.800	158	3218	35	1528	68.4	9.4	98.9	18.1	32.2
0.900	74	3242	11	1612	67.1	4.4	99.7	12.9	33.2
1.000	0	3253	0	1686	65.9	0.0	100.0	.	34.1

Based on the values from the table and our model context, we need to decide the cutoff probability value for considering the movie to be a success. Since we want to predict good movies, we need to concentrate on having a less False positive rate and high specificity along with the accuracy of the model. From the above table, we could see that when the cutoff probability is equal to 0.5, we have the highest accuracy (73.4%) along with reasonable Specificity (89.3) and False positive rate (32.7). Therefore, we could conclude that if the predicted probability is greater than 0.5, then we can predict that the movie would to be successful.

The below figure represents the ROC curve from the above model. This curve depicts the relationship between sensitivity and the false positive rate (1 - specificity).



RESOLVING THE BUSINESS PROBLEM:

Our statistical analysis finds some evidence on how the attributes of a movie could help predict the success of movie. The following are a few suggestions based on our analysis to the film production companies.

- Focus on the genres, Drama and Romance which are more likely to be successful.
- From our logistic regression we could also observe that animation and documentaries could also prove to be successful.
- Hire the directors and actors mentioned in our analysis who have a high success rate.
- Movie's which have a content rating of R seemed to be liked by majority of the population.
- Along with the above attributes, we can use the values of budget, duration, director Facebook likes, total cast Facebook likes along with the type of genre and substitute these values in the logistic regression model and could calculate the probability of the movie becoming a success.