Movie Data Analysis for Producers

Group 10

Group Members: Elliana Raines, Mukta Marathe, Evan Sharpe, Nicolas Peterson, Chancellor Harper

Project Description

The goal of this project is to gain a real-world understanding of a data set, in our case, movie data. Our team at Arizona State University is going to take our selected data and gain an insightful understanding of it using in-class tools such as Excel and JMP to develop an in-depth analysis. The goal of this proposal is to create and present recommendations for movie producer companies based on the movie data set.

Our data set contains information about various movies. For each movie, there is a director name, names of two actors, gross total made from the movie, a genre and key plot phrasing, link to an IMDB page as well as a score, and some other relevant data.

We wanted to address the following points:

1. Which factor is the best predictor of a movie's success?
2. What is the extent of external reception versus internal production on the movies' success/popularity? For example, how much do critic and actor choices influence movie Facebook likes?
3. What would we tell our marketing team given the task of advertising a certain movie, how would we do it, and what factors would be most important to touch on/what matters to audiences the most?
4. What are the most common or most popular plots and genres for movies?

Approaches to Analyze Data

In order to analyze the data, we used a combination of regression and ANOVA tests as well as running the data and creating tables in Excel and visualizations in Tableau to better understand trends. The regression and ANOVA tests were used to find significance of certain variables to answer our questions, and Excel and Tableau were used to analyze and find trends. By using the regression tests to find the correlation levels between certain variables, we should be able to find what variables best predict movie success based on the highest correlation (r squared number).

Results:

Question 1: Best predictors of a movie's success based on gross income?

For the first question, assuming that movie success is determined by gross income, we tested various different factors that could determine movie success.

1. Gross income VS Budget

If you refer to Figure 5 in the appendix, you can see that the r squared value from the bivariate model gives us an r squared value of 0.5773, which shows a fairly certain correlation between gross income and budget.

2. Gross income vs Facebook Likes

As you can see in figure 5, movie Facebook likes correlates less at an r squared value of 0.4607, lower than the correlation for budget. This shows that facebook likes is not in as strong correlation with a successful movie with high gross income. So, budget is a better determinant of movie success (based on gross income) than movie FB likes.

3. Gross Income VS IMDB score

As you can see in figure 7 from the appendix, there is still not a super strong correlation (r squared of 0.4494), and this is our lowest r squared value thus far, so it shows us that IMDB does not have the best prediction of movie success, so it does not determine movie success as strongly as our first variable, movie budget.

4. Gross Income VS Director Name

As you can see in figure 17, the director's name has a correlation of 0.503, which while not extremely high, is closer to our highest correlation of budget. However, since the p value is slightly greater than 0.01, this is not extremely significant but it would still be a partial potential predictor of movie success.

Here, we address our original null hypothesis:

H_0: Budget is not significant in predicting a film's success (gross income).

To continue with our testing, we identified the p values to ensure they were significant. As you can see according to figure 8, for the budget vs gross income Anova test, the p value is extremely significant as it is less than 0.01. We can conclude, based on our regression analyses and Anova test, that out of the three factors tested, movie budget is the strongest predictor of movie success (based on gross income). The r squared value of budget and gross income, at 0.5773 tells us that 57.53% of the gross income can be explained by the budget of the movie, which while not a super strong correlation, it does show that this model does explain some of the results in gross profit. We can reject the null hypothesis that budget is not significant in predicting a movie's success based on gross profit.

Question 2: External reception VS internal production

1. Relationship between movie Facebook likes versus IMDB scores:

According to Figure 9, it seems that there is a discrepancy between what critics and audiences like, showing that the lower IMDB scores have higher Facebook likes. This shows that possibly IMDB score, which is an internal factor, does not necessarily predict external outcomes (how much viewers like the movie). Additionally, according to Figure 10, we see that with the r squared value of 0.061, there is not much correlation between movie Facebook likes and IMDB scores, matching our observations.

2. Relationship between gross income and IMDB score.

Interestingly enough, according to Figure 11, the movies that made the highest gross income actually fall between about 6.5-8 scores in IMDB, though closer to 10 is considered a really well made movie. This supports the finding that an IMDB score does not necessarily predict outcomes. Therefore, critics' views vary from audience views and thus gross income can be high even if critic reviews are not extremely high. This fits with our results from our first question, which concluded that critics' results are not necessarily the best predictor of movie gross income. Additionally, if you look at Figure 12, we see that the low r squared value of IMDB score and gross income at 0.0074, shows there is very little to no correlation between gross income and IMDB score.

3. Relationship between actors chosen and IMDB score.

According to Figure 13, we can see that the 1 way Anova test between Actor 1 name and IMDB score has a decent r squared value of 0.46, which does not show a high correlation, but more of a correlation than gross income and IMDB score. Therefore, we can say that the main actor chosen does have slight influence (not heavy) over the critic reviews.

4. Relationship between actors chosen and Facebook likes.

According to Figure 14, we can see that the 1 way Anova test between Actor 1 name and Facebook likes has a lower r squared value than that from 3 (above) at 0.33, so there is not a strong correlation between actor chosen and Facebook likes.
Therefore, we can conclude that internal choices like the lead actor name and IMDB scores do not predict movie outcomes, and are very unlikely to ever predict movie outcomes (based on Facebook likes).

Question 3: Marketing team considerations
According to our findings from questions 1 and 2, we understand that a movie actor name does not necessarily have a strong correlation with movie success. However, according to Figure 17, we did find that the director's name does have a slight but not extremely significant correlation with movie success based on gross profit. As a result, we would recommend our marketing team to include the director's name in big letters (or a noticeable way) on the movie ad. Marketers can include the actor name and IMDB score, however this may not have a significant influence over viewer and movie success, but it would not hurt the results.

Question 4: Most Common Genres
According to Figure 15, you can evidently see that action, adventure and sci fi get the majority of Facebook likes by a great amount. If you take a look at gross income in Figure 16, (same chart just with gross income instead of Facebook likes), action, adventure, and sci-fi get the majority of gross income compared to other genres (a total of over 9,000,000!).

Roadblocks:

We had some issues figuring out how to best incorporate an ANOVA test into our analysis and results. However, with some teamwork, we realized that. We also had a group member who could not access JMP (the app was not working), but after re-installation we were able to get past this issue. It was a bit difficult due to working remotely and not being able to meet in class to easily discuss plans to work on the analysis, however we were able to communicate through our group text and worked everything out.

How can the results contribute to the final recommendations?

Based on our tests from our first question, we understand that movie success as defined by gross income is predicted not completely, but partially by budget. Based on this, we would recommend that movie production groups ensure that they have a strong budget to create a high quality movie consumers will enjoy.

Based on question 3, we recommend to include the director name in big letters on an advertisement, and can include the main/lead actor name and IMDB score, though this may not have a significant influence on movie success based on gross profit, but it will not hurt viewer turnout or profit levels.

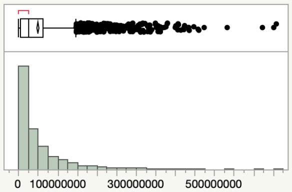
Additionally, based on question 4, we understand that action, adventure, and sci-fi genres tend to get the most Facebook likes and the majority of gross income compared to other genres by a significant amount, so these genres will be crucial in movie production.

Appendices

Figure 1

Distributions

Gross Income



Quantiles

100.0%	maximum	658672302
99.5%		401018018.69
97.5%		242385210.3
90.0%		124984113.2
75.0%	quartile	62118747.5
50.0%	median	25482931
25.0%	quartile	5321182.5
10.0%		381217.2
2.5%		23836.45
0.5%		4014.07
0.0%	minimum	162

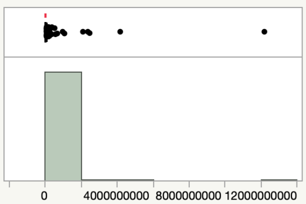
Summary Statistics

Mean	48227826
Std Dev	67424410
Std Err Mean	1045748.2
Upper 95% Mean	50278052
Lower 95% Mean	46177600
N	4157

Figure 2

Distributions

Budget



Quantiles

100.0%	maximum	12215500000
99.5%		250000000
97.5%		165000000
90.0%		85000000
75.0%	quartile	45000000
50.0%	median	20000000
25.0%	quartile	6000000
10.0%		1400000
2.5%		200000
0.5%		20000
0.0%	minimum	218

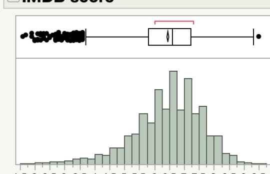
Summary Statistics

Mean	39752620
Std Dev	206114898
Std Err Mean	3055314.8
Upper 95% Mean	45742521
Lower 95% Mean	33762720
N	4551

Figure 3

Distributions

IMDB score



Quantiles

100.0%	maximum	9.5
99.5%		8.7
97.5%		8.3
90.0%		7.8
75.0%	quartile	7.2
50.0%	median	6.6
25.0%	quartile	5.8
10.0%		5
2.5%		3.7
0.5%		2.7
0.0%	minimum	1.6

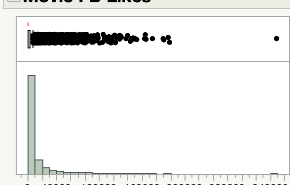
Summary Statistics

Mean	6.4420071
Std Dev	1.1251893
Std Err Mean	0.0158462
Upper 95% Mean	6.4730725
Lower 95% Mean	6.4109418
N	5042

Figure 4

Distributions

Movie FB Likes



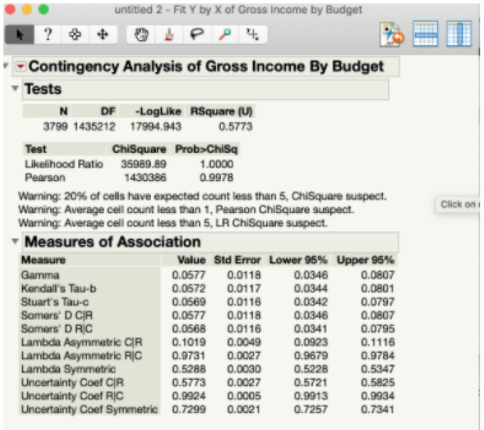
Quantiles

100.0%	maximum	349000
99.5%		123000
97.5%		62000
90.0%		24000
75.0%	quartile	3000
50.0%	median	166
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	7527.4572
Std Dev	19322.071
Std Err Mean	272.11485
Upper 95% Mean	8060.9206
Lower 95% Mean	6993.9938
N	5042

Figure 5



untitled 2 - Fit Y by X of Gross Income by Budget

Contingency Analysis of Gross Income By Budget

Tests

N	DF	-LogLike	RSquare (U)
3799	1435212	17994.943	0.5773

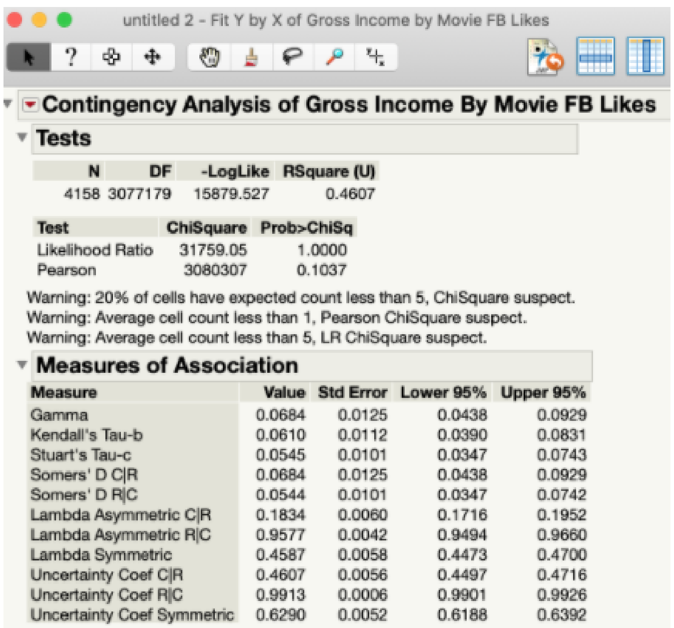
Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	35989.89	1.0000
Pearson	1430386	0.9978

Warning: 20% of cells have expected count less than 5, ChiSquare suspect.
Warning: Average cell count less than 1, Pearson ChiSquare suspect.
Warning: Average cell count less than 5, LR ChiSquare suspect.

Measures of Association

Measure	Value	Std Error	Lower 95%	Upper 95%
Gamma	0.0577	0.0118	0.0346	0.0807
Kendall's Tau-b	0.0572	0.0117	0.0344	0.0801
Stuart's Tau-c	0.0569	0.0116	0.0342	0.0797
Somers' D C|R	0.0577	0.0118	0.0346	0.0807
Somers' D R|C	0.0568	0.0116	0.0341	0.0795
Lambda Asymmetric C|R	0.1019	0.0049	0.0923	0.1116
Lambda Asymmetric R|C	0.9731	0.0027	0.9679	0.9784
Lambda Symmetric	0.5288	0.0030	0.5228	0.5347
Uncertainty Coef C|R	0.5773	0.0027	0.5721	0.5825
Uncertainty Coef R|C	0.9924	0.0005	0.9913	0.9934
Uncertainty Coef Symmetric	0.7299	0.0021	0.7257	0.7341

Figure 6



untitled 2 - Fit Y by X of Gross Income by Movie FB Likes

Contingency Analysis of Gross Income By Movie FB Likes

Tests

N	DF	-LogLike	RSquare (U)
4158	3077179	15879.527	0.4607

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	31759.05	1.0000
Pearson	3080307	0.1037

Warning: 20% of cells have expected count less than 5, ChiSquare suspect.
Warning: Average cell count less than 1, Pearson ChiSquare suspect.
Warning: Average cell count less than 5, LR ChiSquare suspect.

Measures of Association

Measure	Value	Std Error	Lower 95%	Upper 95%
Gamma	0.0684	0.0125	0.0438	0.0929
Kendall's Tau-b	0.0610	0.0112	0.0390	0.0831
Stuart's Tau-c	0.0545	0.0101	0.0347	0.0743
Somers' D C|R	0.0684	0.0125	0.0438	0.0929
Somers' D R|C	0.0544	0.0101	0.0347	0.0742
Lambda Asymmetric C|R	0.1834	0.0060	0.1716	0.1952
Lambda Asymmetric R|C	0.9577	0.0042	0.9494	0.9660
Lambda Symmetric	0.4587	0.0058	0.4473	0.4700
Uncertainty Coef C|R	0.4607	0.0056	0.4497	0.4716
Uncertainty Coef R|C	0.9913	0.0006	0.9901	0.9926
Uncertainty Coef Symmetric	0.6290	0.0052	0.6188	0.6392

Figure 7

untitled 2 - Fit Y by X of Gross Income by IMDB score

Contingency Analysis of Gross Income By IMDB scor

Tests

N	DF	-LogLike	RSquare (U)
4158	298442	15491.987	0.4494

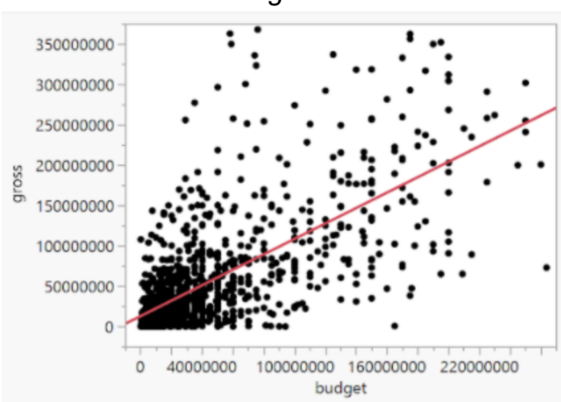
Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	30983.97	1.0000
Pearson	296909.9	0.9765

Warning: 20% of cells have expected count less than 5, ChiSquare suspect.
Warning: Average cell count less than 1, Pearson ChiSquare suspect.
Warning: Average cell count less than 5, LR ChiSquare suspect.

Measures of Association

Measure	Value	Std Error	Lower 95%	Upper 95%
Gamma	0.0344	0.0108	0.0133	0.0555
Kendall's Tau-b	0.0339	0.0106	0.0131	0.0547
Stuart's Tau-c	0.0339	0.0106	0.0131	0.0547
Somers' D C|R	0.0344	0.0108	0.0133	0.0555

Figure 8
Gross Income V Budget



Summary Statistics

	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	0.093308	0.06169	0.124739	<.0001*
Covariance	1.44e+15			
Count	3799			

Variable	Mean	Std Dev
Budget	39752620	2.061e+8
Gross Income	48227826	67424410

Figure 9

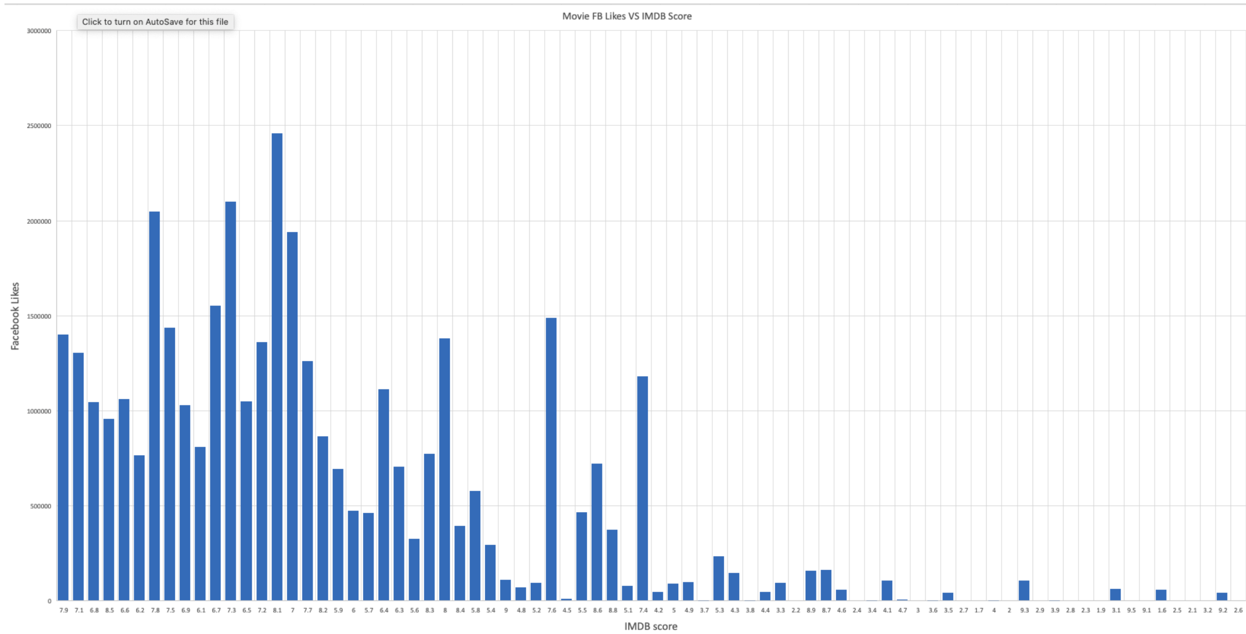


Figure 10

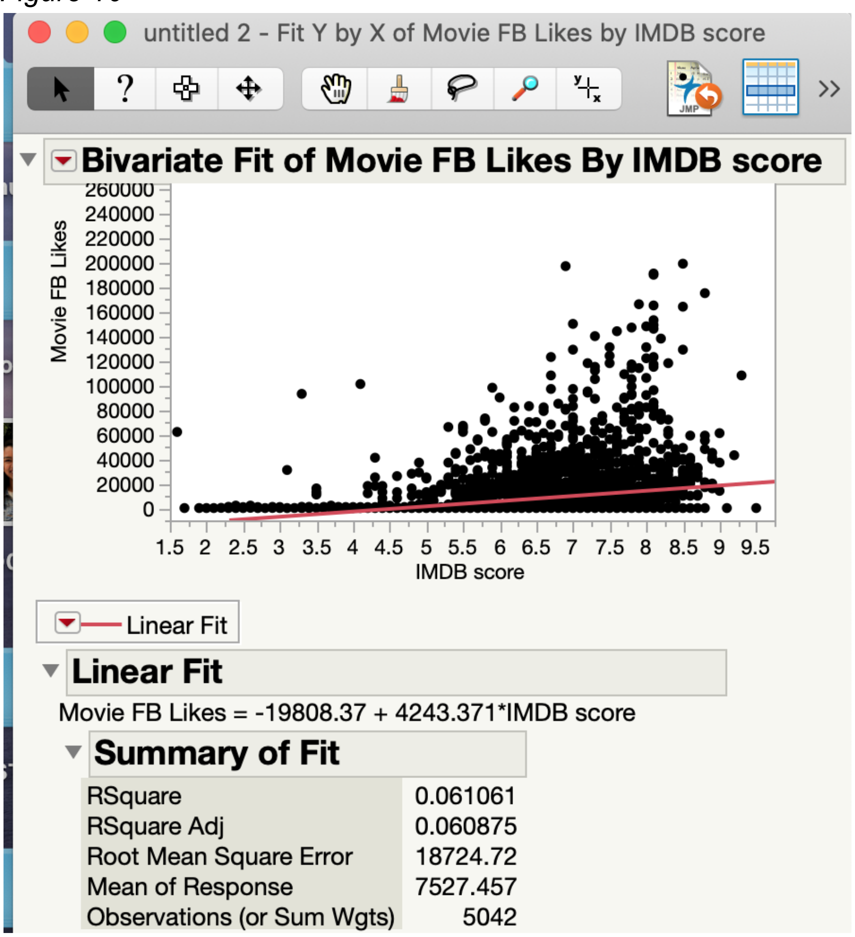


Figure 11

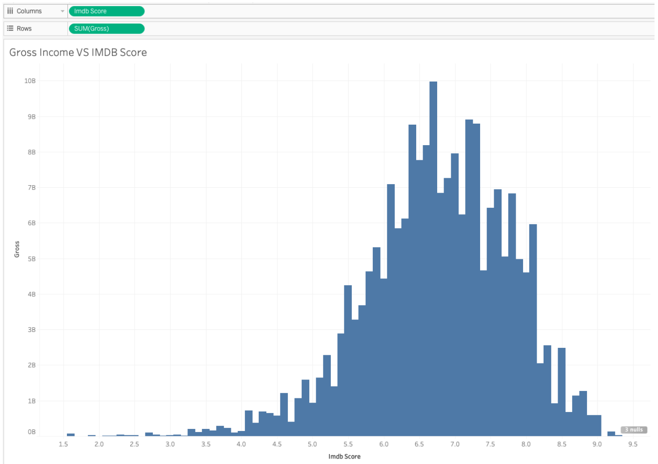
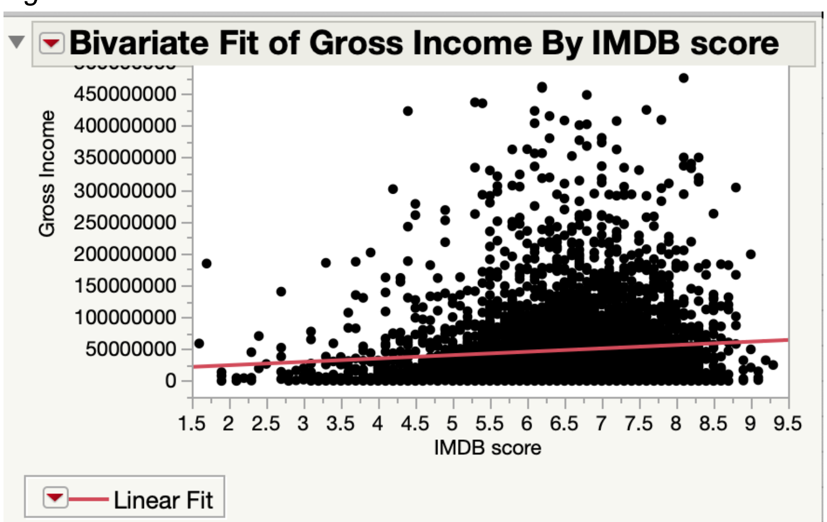


Figure 12



Linear Fit

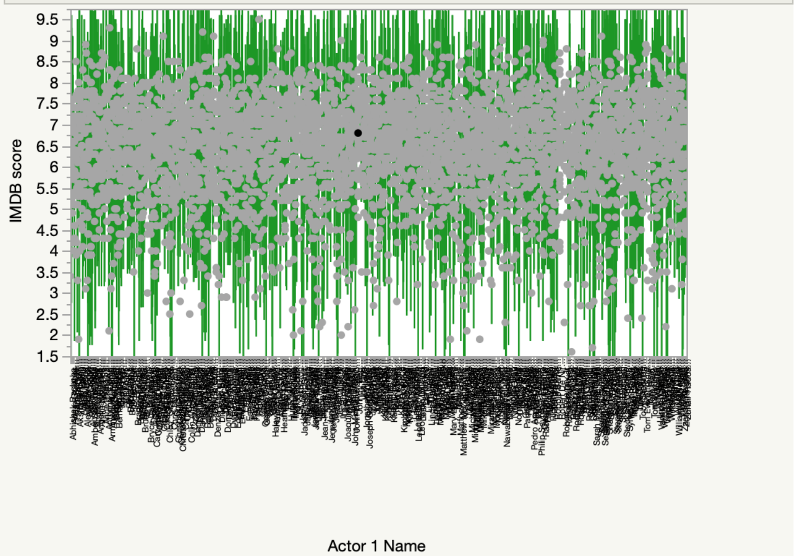
Gross Income = 14008024 + 5292141.9*IMDB score

Summary of Fit

RSquare	0.007414
RSquare Adj	0.007175
Root Mean Square Error	67182077
Mean of Response	48227826
Observations (or Sum Wgts)	4157

Figure 13

Oneway Analysis of IMDB score By Actor 1 Name



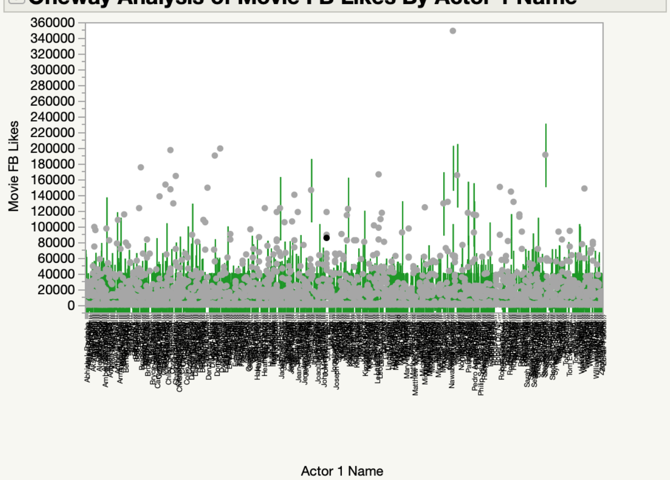
Oneway Anova

Summary of Fit

Rsquare	0.461759
Adj Rsquare	0.078087
Root Mean Square Error	1.080625
Mean of Response	6.442205
Observations (or Sum Wgts)	5035

Figure 14

Oneway Analysis of Movie FB Likes By Actor 1 Name



Oneway Anova

Summary of Fit

Rsquare	0.336744
Adj Rsquare	-0.13604
Root Mean Square Error	20606.89
Mean of Response	7537.161
Observations (or Sum Wgts)	5035

Figure 15

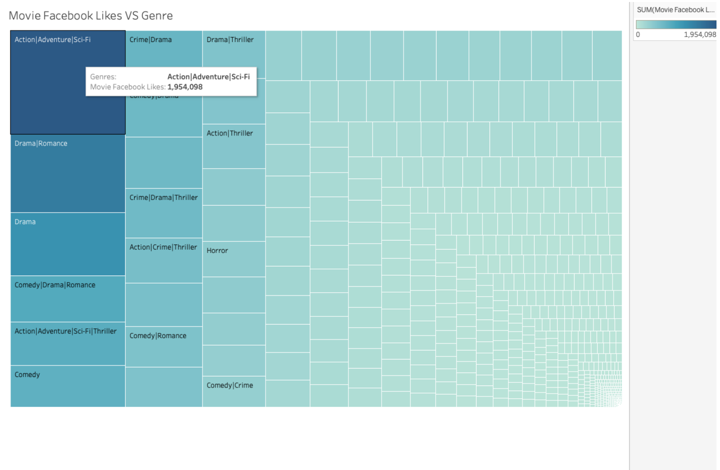


Figure 16

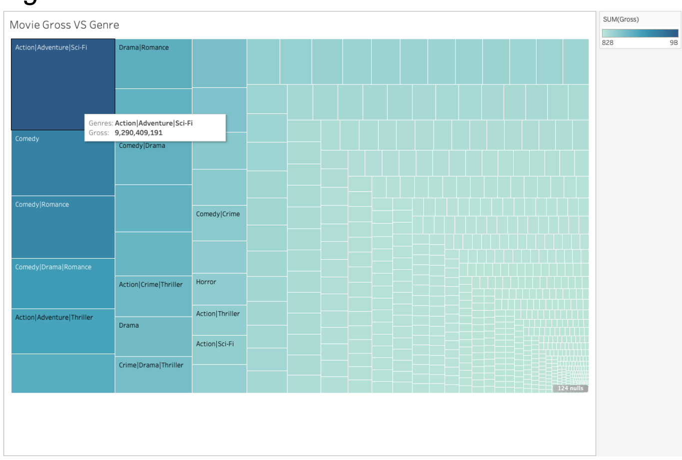
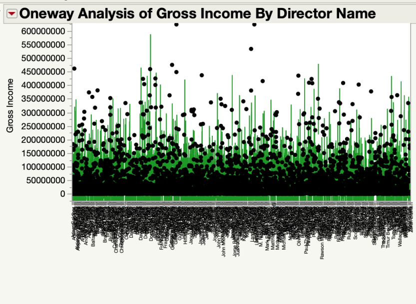


Figure 17

Fit Group

Oneway Analysis of Gross Income By Director Name



Oneway Anova

Summary of Fit

Rsquare	0.503412
Adj Rsquare	0.046644
Root Mean Square Error	66140345
Mean of Response	48476473
Observations (or Sum Wgts)	4071

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Director Name	1950	9.4015e+18	4.821e+15	1.1021	0.0142*