ANALYZING TRENDS IN IMDB MOVIE DATASET

COLLEGE OF
INFORMATION
STUDIES

INST 627: Data Analytics for Information Professionals

Project Paper

INSTRUCTOR: Dr. Yla Tausczik

Fall 2016

Submitted by Team Bazinga

Arpit Chandra
Shashank Kava
Khushal Navani
Prerak Sheth

Table of Contents

Motivation

There is unlimited information about various movies in form of reviews and news that floats around the internet, printed media, audio channels and innumerable other sources. Rarely can anyone separate the legitimate information from the hoax information. For someone who is trying to figure out what movie to go for, too much information can be frustrating. Same is the case with production houses where they are trying to find out a potential investment opportunity but are unable to reach an agreement due to enormous amount of available information. Movie reviews seem to be a very subjective measure of grading a movie, and critics, although are supposed to be fair and just, are by accords of human nature, biased. People and their reviews are too erratic, and too subjective. In spite of a low critic score, often the movie does well at the box office. The challenge is to find a quantifiable value, by the virtue of which we can gauge how successful the movie could be. Not only does this affect the audience, but also the production houses who are heavily invested in the industry. Using data driven decisions to gauge a success of the movie and finding out what kind of parameters contribute to this decision can aid both, the production houses and the viewers. These decisions can also provide meaningful insights to think tanks who advise the production houses to produce better movies. A varied data set with information ranging from a directors social media presence to the number of actors that form a part of the movie will provide interesting and structured insights in this highly irregular movie industry.

Research Questions

1. Does higher number of likes for all the actors and directors suggest a higher gross revenue?
2. What is the effect of number of user and critic reviews on the IMDB score?
3. Does more number of genres for a movie appeal a larger audience (gross revenue)?

Data Collection

For the purpose of analysis, we have selected an open dataset from https://www.kaggle.com. This is a dataset of 5000+ movies scrapped from the IMDB website, which consists of movies across various geography, languages and time. The data contains multiple attributes ranging from the color of the movie, director names; to Facebook likes on the directors page. It also has data on number of people on various movie posters which were collected by using face recognition algorithms and many more numeric entities. Data was collected using a python library called scrapy.

Population

The population consists of all the movies released across the world till date and listed on IMDB.

Sample

The sample is the set of 5043 movies collected using a scrapy library in python out of the population that consists of all the movies listed on IMDB. This is a random sample, since every movie listed on the IMDB website has an equal and independent chance of being selected. Since random sampling is carried out, this sample set is an appropriate representation of the population. The level of analysis is a movie which has been listed, reviewed and rated on IMDB.

Study Design

It is an observational study, since the data has been collected from the IMDB website, which consists of legitimate facts and figures. It is not a random assignment since we are not dividing movies in groups based on any criteria. Since the study is observational and there are no control groups for specific experiments, we do not have random assignment.

Statistical Analysis

Question 1:

Does higher number of Facebook likes for all the actors and directors suggest a higher gross revenue?

Importance: In the age of social media, this question essentially points at the importance of a having a social media footprint for every actor and director. If there exists a relation, it can encourage the actors and directors of the movie to be actively present on social media.

Data Cleaning:

There are 20 tuples which have null values for the number of likes. The data set was scraped on August 29, 2016 and as the number of likes for all the entities have changed over time, the data set cannot be updated. Hence, we eliminate these rows. We also delete 776 rows for movies having no data about gross earnings.

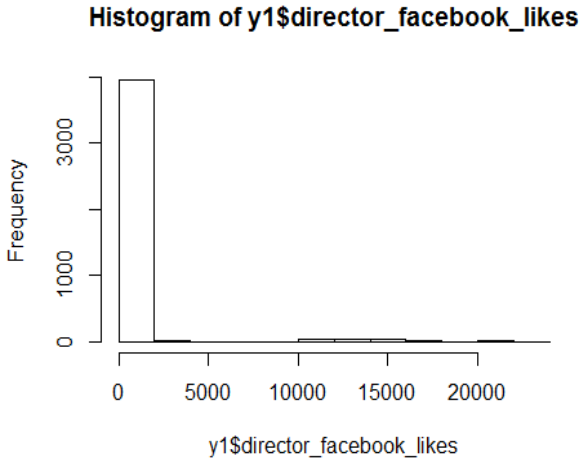
Variables and Scale of Measurement:

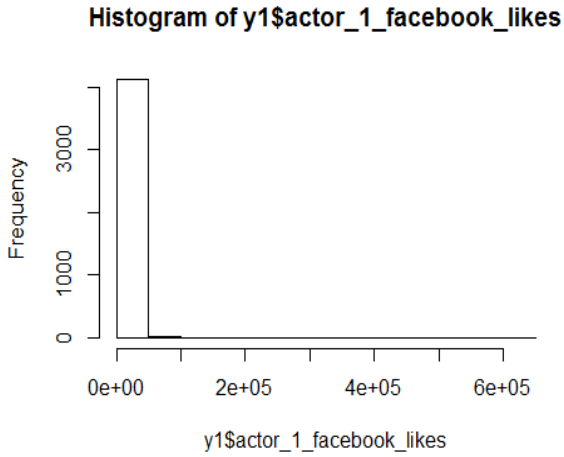
Independent Variables: FB likes for actor 1 (Ratio), FB likes for actor 2 (Ratio), FB likes for actor 3 (Ratio), FB likes for the entire cast (Ratio), FB likes for director (Ratio)

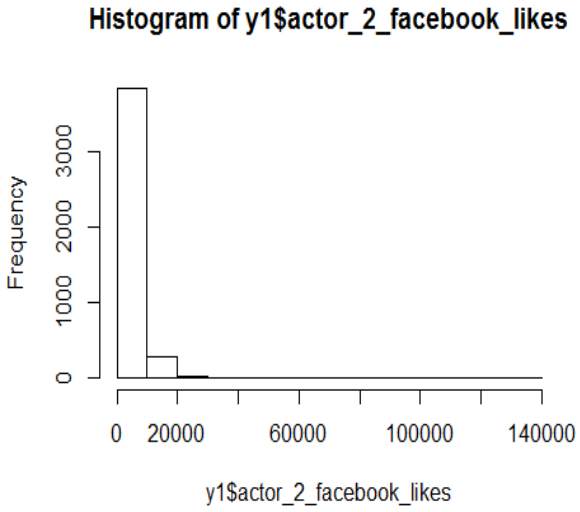
Dependent Variables: Gross Revenue (Ratio)

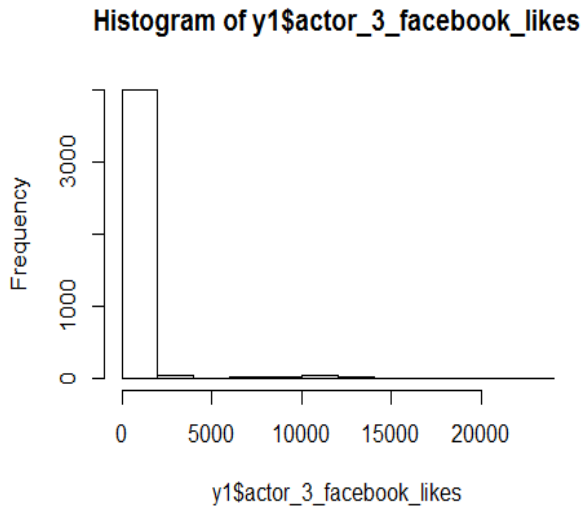
Descriptive Statistics and graphs of variables:

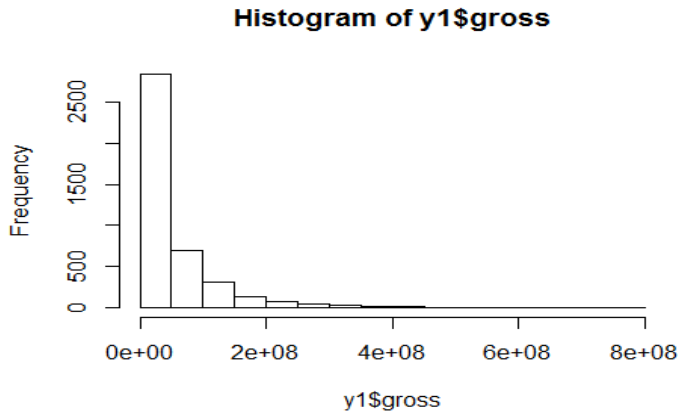
```
hist(y1$director_facebook_likes)
hist(y1$actor_1_facebook_likes)
hist(y1$actor_2_facebook_likes)
hist(y1$actor_3_facebook_likes)
hist(y1$gross)
```


Histogram of y1$director_facebook_likes
Frequency
0
1000
3000
0
5000
10000
15000
20000
y1$director_facebook_likes


Histogram of y1$actor_1_facebook_likes
Frequency
0
1000
3000
0e+00
2e+05
4e+05
6e+05
y1$actor_1_facebook_likes


Histogram of y1$actor_2_facebook_likes
Frequency
0
1000
2000
3000
0
20000
60000
100000
140000
y1$actor_2_facebook_likes


Histogram of y1$actor_3_facebook_likes
Frequency
0
1000
3000
0
5000
10000
15000
20000
y1$actor_3_facebook_likes

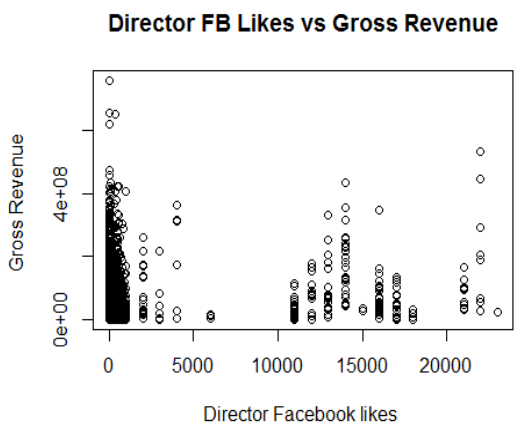

Histogram of y1$gross
Frequency
0
500
1500
2500
0e+00
2e+08
4e+08
6e+08
8e+08
y1$gross

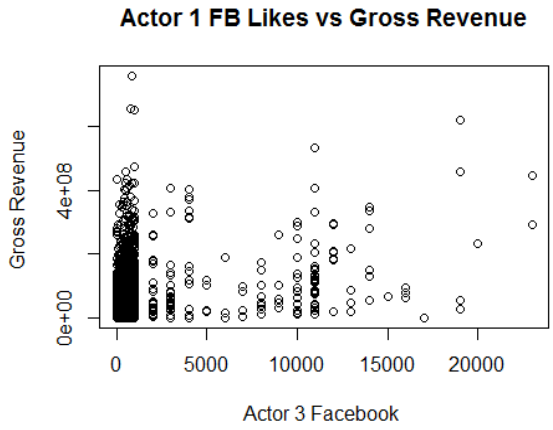
As the distribution for all the variables are skewed we consider median to be the measure of the central tendency.

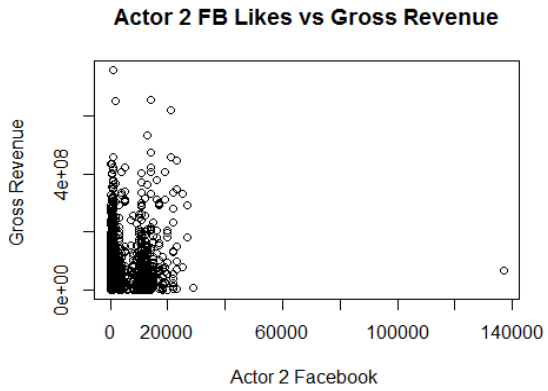
```
median(y1$director_facebook_likes)

 [1] 57

median(y1$actor_1_facebook_likes)

 [1] 1000

median(y1$actor_2_facebook_likes)

 [1] 651

median(y1$actor_3_facebook_likes)

 [1] 416

median(y1$gross)

 [1] 25592632
```

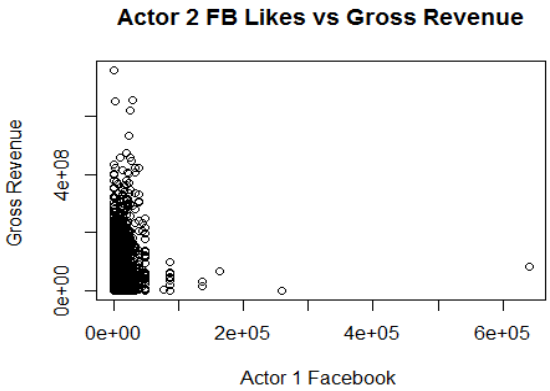
DV vs IV plots:

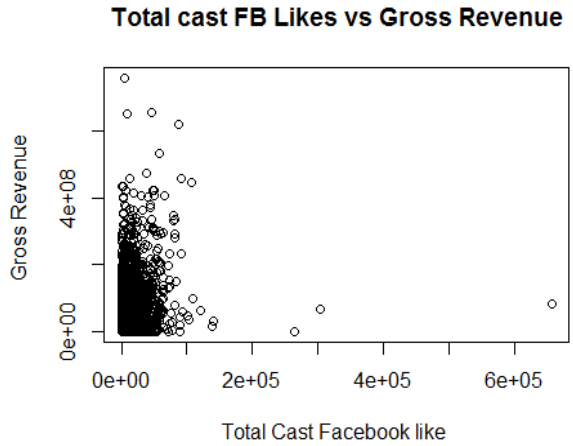
```
plot(y1$director_facebook_likes,y1$gross, main=" Director FB Likes vs Gross Revenue", ylab="Gross R
evenue ", xlab="Director Facebook likes ")

plot(y1$actor_3_facebook_likes,y1$gross, main=" Actor 1 FB Likes vs Gross Revenue ", ylab="Gross R
evenue ", xlab="Actor 3 Facebook ")

plot(y1$actor_2_facebook_likes, y1$gross, main=" Actor 2 FB Likes vs Gross Revenue ", ylab="Gross R
evenue ", xlab="Actor 2 Facebook ")

plot(y1$actor_1_facebook_likes, y1$gross, main=" Actor 2 FB Likes vs Gross Revenue ", ylab="Gross R
evenue ", xlab="Actor 1 Facebook ")

plot(y1$cast_total_facebook_likes, y1$gross, main=" Total cast FB Likes vs Gross Revenue ", ylab="Gro
ss Revenue ", xlab="Total Cast Facebook like ")
```


Director FB Likes vs Gross Revenue
Gross Revenue
0e+00
4e+08
0
5000
10000
15000
20000
Director Facebook likes


Actor 1 FB Likes vs Gross Revenue
Gross Revenue
0e+00
4e+08
0
5000
10000
15000
20000
Actor 3 Facebook


Actor 2 FB Likes vs Gross Revenue
Gross Revenue
0e+00
4e+08
0
20000
60000
100000
140000
Actor 2 Facebook


Actor 2 FB Likes vs Gross Revenue
Gross Revenue
0e+00
4e+08
0e+00
2e+05
4e+05
6e+05
Actor 1 Facebook


Total cast FB Likes vs Gross Revenue
Gross Revenue
0e+00
4e+08
0e+00
2e+05
4e+05
6e+05
Total Cast Facebook like

Stating the hypothesis:

Null Hypothesis (Ho): There is no correlation between the gross revenue and the predictor variables mentioned above.

Alternative Hypothesis (Ha): There exists a correlation between the gross revenue and at least one of the predictor variables.

Assumptions:

1. Collinearity:

We calculate the correlation between the independent variables in R and check for collinearity between the variables.

	y$director_facebook_likes	y$actor_1_facebook_likes	y$actor_2_facebook_likes	y$actor_3_facebook_likes
y$actor_1_facebook_likes	0.09167696	x	x	x
y$actor_2_facebook_likes	0.1194725	0.3923905	x	x
y$actor_3_facebook_likes	0.1208131	0.2551741	0.5552208	x
y$cast_total_facebook_likes	0.1213189	0.94573	0.6423117	0.4901311

As the variable likes for the entire cast has a strong positive correlation with the other independent variables, it violates the assumption. We also eliminate the variable from the regression model.

2. Outliers:

The presence of outliers can be detected by recording observations which are 3 or more standard deviations away from the predicted values.

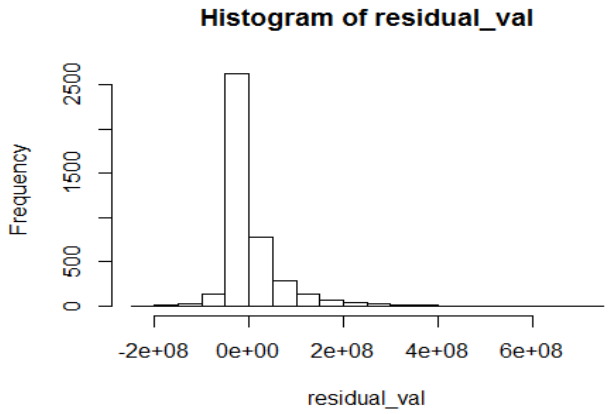
```
m=lm(y1$gross~y1$director_facebook_likes+y1$actor_1_facebook_likes+y1$actor_2_facebook_likes+y
1$actor_3_facebook_likes)
pred_val=m$fitted.values
residual_val=m$residuals
residual_sd=sd(residual_val)
length(residual_val[abs(residual_val)>=3*residual_sd])
```

```
 [1] 90
```

3. Normality:

It assumes that the error between the observed and the predicted values is normally distributed. This assumption can be checked by plotting the residual values on a histogram.

```
hist(residual_val)
```

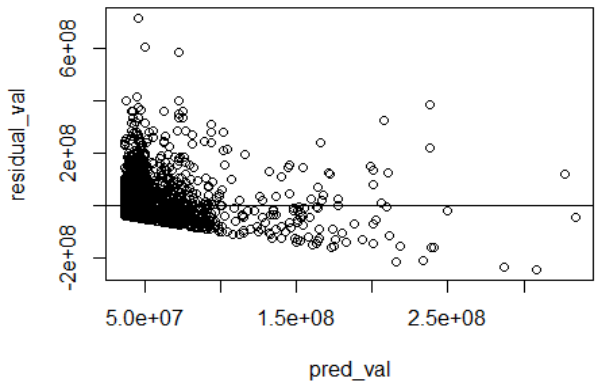


Since, the plot is normally distributed the assumption of normality is not violated.

4. Constant Error:

The constant error is used to test that variances of the errors for all the predicted values is equal. This is done by plotting the predicted values with the residuals in R. As shown below, the residuals do not vary evenly, hence there is a variation of constant error.

```
plot(pred_val,residual_val)
abline(0,0)
```



Statistical Analysis:

To find the correlation between the outcome and the predictor variables, we need to make use of multiple regression model. The aim of the research question is to build the best possible linear model to explain the gross revenue.

Results:

Using the multiple regression model to find a correlation we arrive at the value of the statistic and p-values,

```
m=lm(y1$gross~y1$director_facebook_likes+y1$actor_1_facebook_likes+y1$actor_2_facebook_likes+y
1$actor_3_facebook_likes)
summary(m)


Call:
lm(formula = y1$gross ~ y1$director_facebook_likes + y1$actor_1_facebook_likes +
    y1$actor_2_facebook_likes + y1$actor_3_facebook_likes)

Residuals:
      Min        1Q    Median        3Q       Max
-242188861  -36195451  -20400053   13198825  715448231

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              3.602e+07  1.170e+06  30.784  < 2e-16 ***
y1$director_facebook_likes 2.334e+03  3.424e+02   6.817 1.07e-11 ***
y1$actor_1_facebook_likes  2.052e+02  7.232e+01   2.837  0.00458 **
y1$actor_2_facebook_likes  1.695e+03  2.896e+02   5.855 5.13e-09 ***
y1$actor_3_facebook_likes  8.480e+03  6.659e+02  12.735  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 64360000 on 4138 degrees of freedom
Multiple R-squared:  0.1189, Adjusted R-squared:  0.1181
F-statistic: 139.6 on 4 and 4138 DF,  p-value: < 2.2e-16
```

Conclusion:

1. With an increase in the number of like for the director by 1, the gross revenue for the movie increases by 2334.
2. With an increase in the number of like for the first actor by 1, the gross revenue for the movie increases by 205.2.
3. With an increase in the number of like for the second actor by 1, the gross revenue for the movie increases by 1695.

4. With an increase in the number of like for the third actor by 1, the gross revenue for the movie increases by 8480.

Question 2:

What is the effect of number of user and critic reviews on the IMDB score?

Importance: Viral posts and video have taken over the internet recently. The more people talk about something, the more publicity it earns. Similarly, this question can help us link the importance of the number of user and critic reviews with the success of the movie

Data Cleaning:

There are a few blank values present in the number of critic reviews and number of user reviews columns which have been replaced with 0.

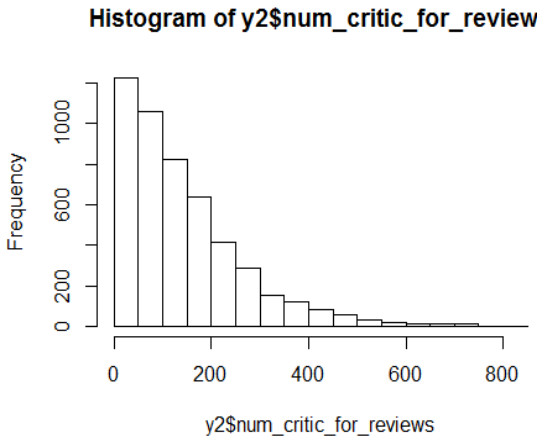
Variables and Scale of Measurement:

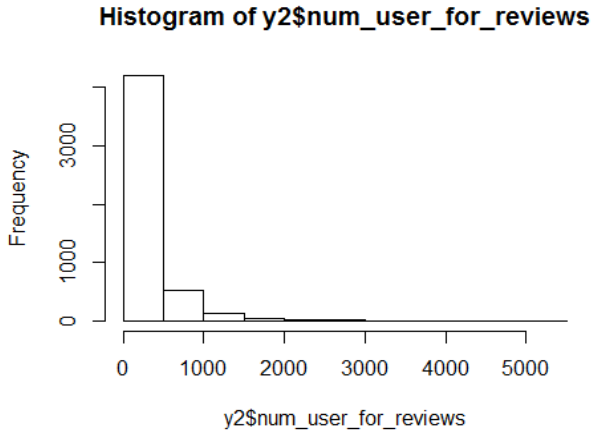
Independent Variables: Number of user reviews (Ratio), Number of critic reviews (Ratio)

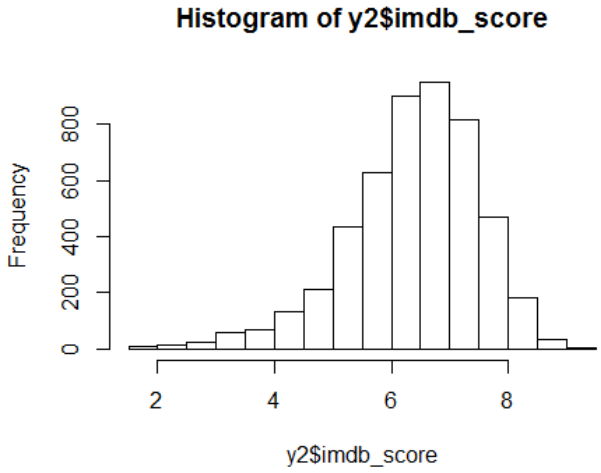
Dependent Variables: IMDB Score (Interval)

Descriptive Statistics and graphs of variables:

```
hist(y2$num_critic_for_reviews)
hist(y2$num_user_for_reviews)
hist(y2$imdb_score)
```





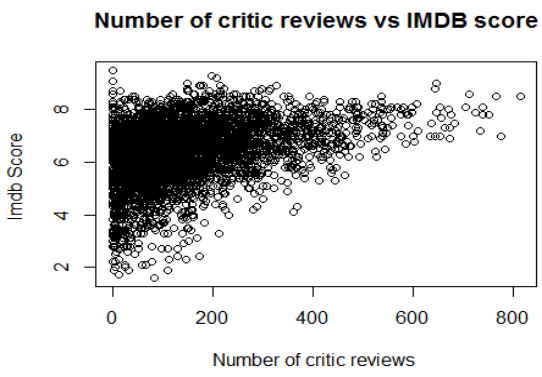


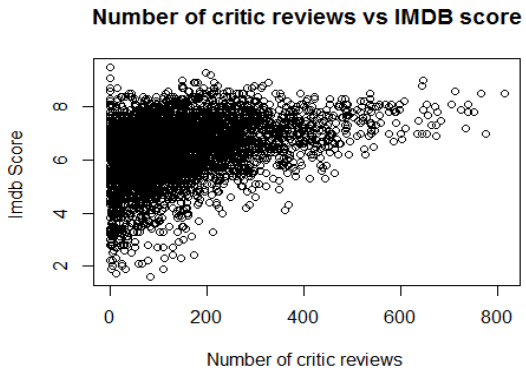
As the distribution for all the variables are skewed we consider median to be the measure of the central tendency.

```
median(y2$num_critic_for_reviews)
 [1] 111
median(y2$num_user_for_reviews)
 [1] 159
median(y2$imdb_score)
 [1] 6.5
```

DV vs IV plots:

```
plot(y2$num_critic_for_reviews, y2$imdb_score, main=" Number of critic reviews vs IMDB score ", ylab="Imdb Score ", xlab="Number of critic reviews ")

plot(y2$num_user_for_reviews, y2$imdb_score, main=" Number of user reviews vs IMDB score ", ylab="Imdb Score ", xlab="Number of user reviews ")
```





Stating the hypothesis:

Null Hypothesis (Ho): There is no correlation between the IMDB score (outcome) and Number of user reviews, Number of critic reviews(Predictor).

Alternative Hypothesis (Ha): There exists a correlation between the outcome variable and predictor variable.

Assumptions:

1. Collinearity:

We calculate the correlation between the independent variables in R and check for collinearity between the variables.

```
cor.test(y2$num_user_for_reviews,y2$num_critic_for_reviews)


 Pearson's product-moment correlation

data:  y2$num_user_for_reviews and y2$num_critic_for_reviews
t = 54.024, df = 4937, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5917044 0.6267704
sample estimates:
      cor
0.6095355
```

As the variable Numbers of User reviews has a strong positive correlation with the other independent variables, it violates the assumption. We also eliminate the variable from the regression model.

2. Outliers:

The presence of outliers can be detected by recording observations which are 3 or more standard deviations away from the predicted values.

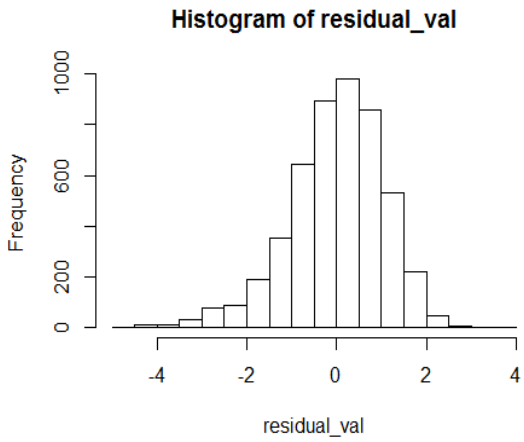
```
m=lm(y2$imdb_score~y2$num_critic_for_reviews)
pred_val=m$fitted.values
residual_val=m$residuals
residual_sd=sd(residual_val)
length(residual_val[abs(residual_val)>=3*residual_sd])
```

```
[1] 44
```

3. Normality:

It assumes that the error between the observed and the predicted values is normally distributed. This assumption can be checked by plotting the residual values on a histogram.

```
hist(residual_val)
```

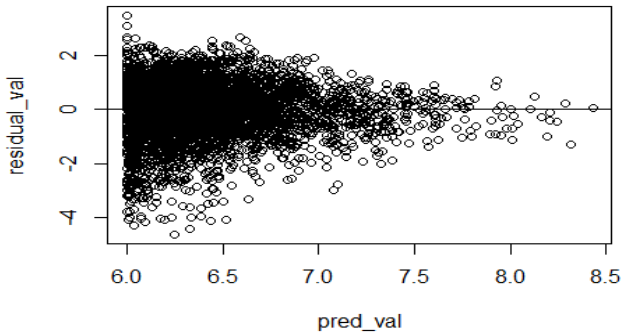


Since, the plot is normally distributed the assumption of normality is not violated.

4. Constant Error:

The constant error is used to test that variances of the errors for all the predicted values is equal. This is done by plotting the predicted values with the residuals in R. As shown below, the residuals do not vary evenly, hence there is a variation of constant error.

```
plot(pred_val,residual_val)
abline(0,0)
```



Statistical Analysis:

To find the correlation between the imdb score and the number of critic reviews, we need to make use of linear regression model. The aim of the research question is to build the best possible linear model to explain the imdb score.

Results:

Using the linear regression model to find a correlation we arrive at the value of the statistic and p-values,

```
m=lm(y2$imdb_score~y2$num_critic_for_reviews)
summary(m)


Call:
lm(formula = y2$imdb_score ~ y2$num_critic_for_reviews)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6473 -0.6082  0.0875  0.7242  3.5044

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                5.9955881  0.0229653  261.07   <2e-16 ***
y2$num_critic_for_reviews 0.0029965  0.0001231   24.34   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.054 on 4937 degrees of freedom
Multiple R-squared:  0.1071, Adjusted R-squared:  0.1069
F-statistic: 592.3 on 1 and 4937 DF,  p-value: < 2.2e-16
```

Conclusion:

With an increase in the number of critic reviews by 1, there is an increase in the IMDB score by 0.0029.

Question 3:

Does more number of genres for a movie appeal a larger audience (gross revenue)?

Importance: Every movie has some tagged genre such as comedy, drama, thriller, etc. This question tries to establish a relation between the number of genres and gross revenue. This can help a production house tag genres for movies such that it can appeal a larger audience which will eventually result in higher revenues.

Data Cleaning:

The genre column contain pipe separated genres list of a movie. We create a column count_genres which contains the count of the number of genres. We count the number of occurrences of the pipe operator (|) and add a count of one to it. To do so, we use the formula stated below.

count_genres = LEN([cell])-LEN(SUBSTITUTE([cell],"|",""))+1, where cell = address of the first cell in the column genre.

The genre column did not have any blank values, had a single genre as the least possible value, thus, the count_genres has a minimum value of 1 and a maximum value of 8.

We attribute the column gross as a measure of larger audience under the assumption that if the movie appeals to greater number of people, it will result in higher collections at the box office.

We delete the rows which have a gross of zero as that would be wrong indication that the movie made zero sales. Absence of a value indicates the data wasn't scraped properly or is unknown. Hence we remove such rows that add up to 783.

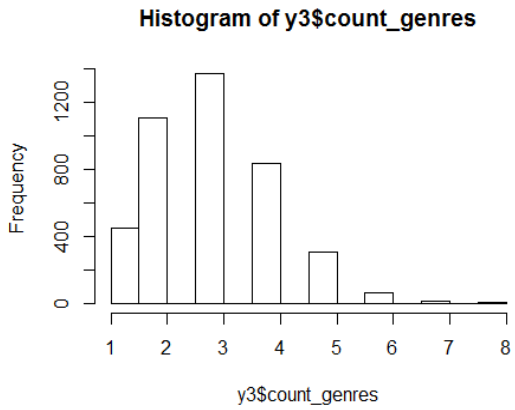
Variables and Scale of Measurement:

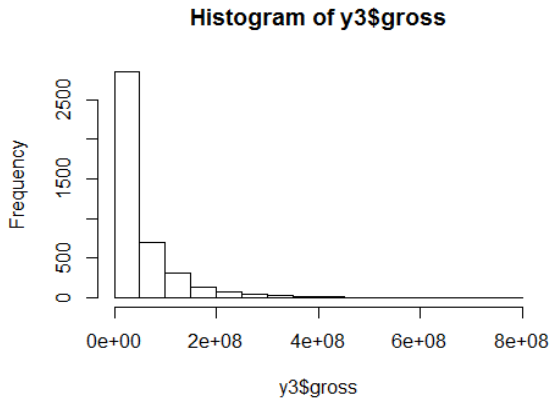
Dependent Variable: Gross Revenue (Ratio)

Independent Variable: Count of genres (interval as a zero value does not exist, a movie cannot not have a genre)

Descriptive Statistics and graphs of variables:

```
hist(y3$count_genres)
hist(y3$gross)
```



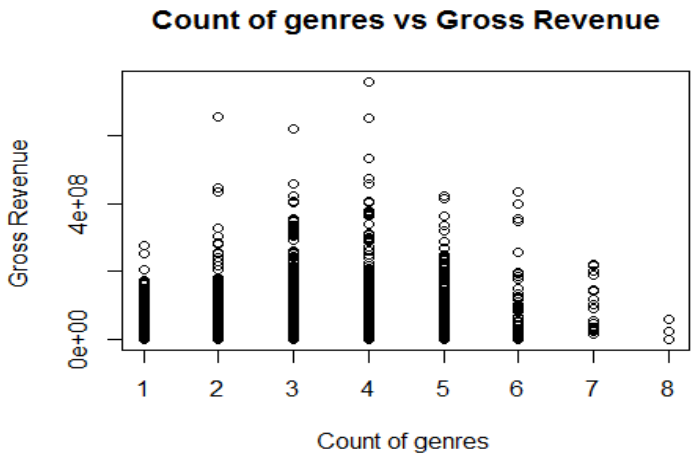


As the distribution for all the variables are skewed we consider median to be the measure of the central tendency.

```
median(y3$count_genres)

 [1] 3

median(y3$gross)

 [1] 25529690
```

DV vs IV plots:

```
plot(y3$count_genres,y3$gross, main=" Count of genres vs Gross Revenue ",
   ylab="Gross Revenue ", xlab="Director Facebook likes ")
```



Stating the hypothesis:

Null Hypothesis (Ho): There is no correlation between number of genres and its gross revenue earning i.e. r = 0

Alternate Hypothesis (Ha): There exists a correlation between the number of genres and its gross revenue earning i.e. r0

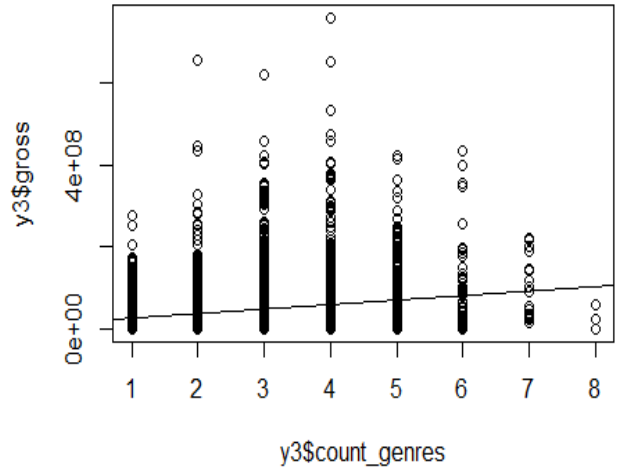
Statistical analysis:

To find the correlation between the variables, we need to make use of Pearson's coefficient r which we can calculate using the cor.test(DV~IV) function as shown below

```
cor.test(y3$count_genres,y3$gross)


 Pearson's product-moment correlation

data:  y3$count_genres and y3$gross
t = 12.643, df = 4154, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1630382 0.2215956
sample estimates:
      cor
0.1924883
```

To visually represent the relationship, we will plot the points using a scatter plot and draw a correlation line through the points.

```
plot(y3$count_genres,y3$gross)
abline(lm(gross~count_genres,data=y3))
```



Conclusion:

Comparing the p-value with the level of significance, we reject the Null Hypothesis. The gross earnings of a movie have a weak correlation i.e. directly related with the number of genres by a moderate positive correlation factor of 0.19.

Limitations

There are certain limitations present in our data set and in our understanding of it. We work on the assumption that IMDB has information about all movies which have released till date. The data set was scraped using a third party tool called Scrappy which has scraped data randomly. As a result we have some outlier data, like movies released in black and white, movies released in pre 2000 years. We have eliminated such rows so that they do not affect the resultant data. We have used the number of likes on Facebook has a predictor variable in our tests, but there is always a temporal inconsistency in this matter. The number of likes are never constant and keep on increasing over time. For a movie which just released before the algorithm scraped the data, the number of likes in the data set will be less than the current likes. Also there are certain actors and directors for which the number of likes are zero and also they might not have a Facebook page. In the gross revenue collected, the amount scrapped is measured just as a numerical entity without mention of the currency. Although most of the monetary measures (budget, gross revenue) is in dollars, there are certain movies which have them measured in their local currency. The absence of the currency factor makes it difficult to compare all the values of the same scale, but we have worked with the assumption that all such monetary measures are in US dollars.

Future Scope

In addition to the data used for our research, there is tons of data available on movies and its success on websites similar to IMDB. Using this data and conducting similar tests between different combinations of variables, we can create a Decision Support System that can aid various categories of people who are affected by this industry. The two primary users of this system can be the viewers/audience and the production houses. Prediction of the movie ratings based on different variables like the actors, number of actors, fame of the director and so on. This can help the viewers plan ahead on the movies to watch prior to its release. The same feature can be used by the production house for the purpose of casting to ensure better revenues. Additionally, we found out from our research how the actors and the director’s social media presence has an impact on the movies success. Production house can use such data driven decision to emphasize on the importance of using social media and similar aspects which can contribute to maximize the

revenues. Also, such decisions can further structure the process of fund allocations which can help optimize the budget and spending.

Point Allocation

We have agreed to divide the points evenly among all members

Group Member	Percentage Distribution
Arpit Chandra	25%
Shashank Kava	25%
Khushal Navani	25%
Prerak Sheth	25%

References

1. Inc, K. (2016). IMDB 5000 movie Dataset. Retrieved December 4, 2016, from Kaggle, https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset

2. Free statistics book. Retrieved December 4, 2016, from Online Statistics Education: An Interactive Multimedia Course of Study, http://onlinestatbook.com/