MGT 256 Netflix jr

Analysis of Movie Dataset
to Know insights of the movie world

CONTENTS

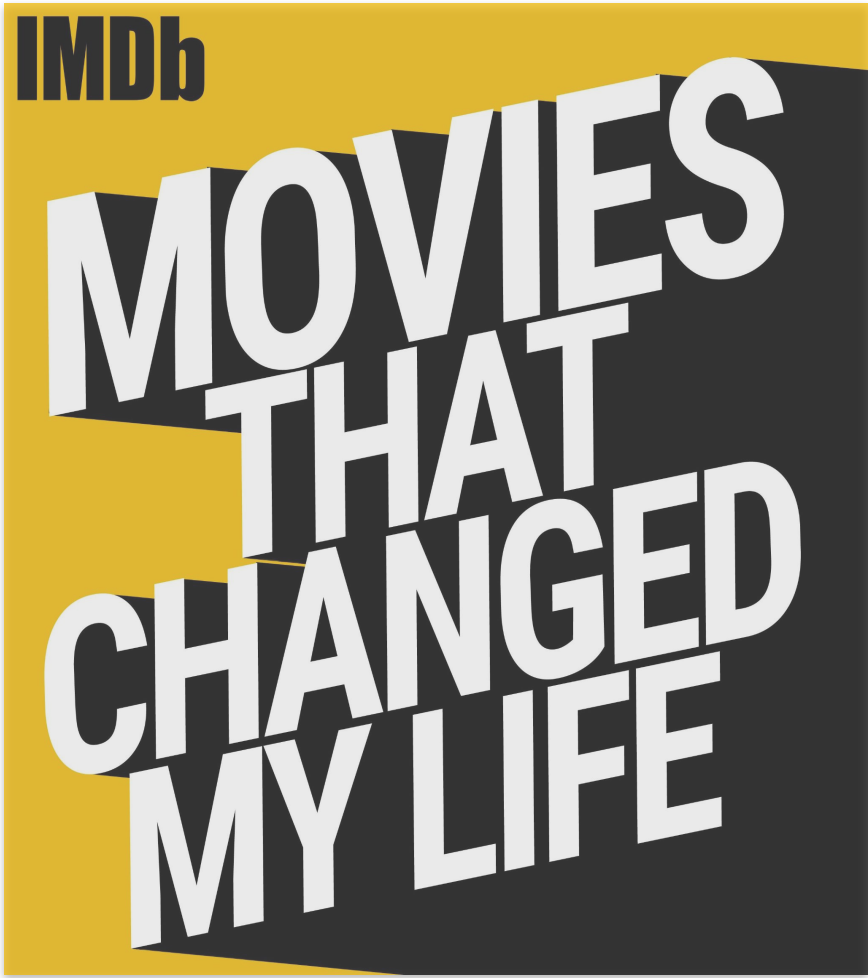
Research Problem

- What variables are fundamental to a film's ability to earn high revenues?

- The relationship between the face value of the actor/director and the rating of IMDB.

- Can we predict the rating of an upcoming movie based on specific variable values?

Our Data



Our Data from https://www.kaggle.com/datasets/carolzhangdc/imdb-5000-movie-dataset

Data Overview: explore the data we have

```
> glimpse(movie)
Rows: 5,043
Columns: 28
$ color                    <chr> "Color", "Color", "Color", "Color"…
$ director_name            <chr> "James Cameron", "Gore Verbinski",…
$ num_critic_for_reviews   <int> 723, 302, 602, 813, NA, 462, 392, …
$ duration                 <int> 178, 169, 148, 164, NA, 132, 156, …
$ director_facebook_likes  <int> 0, 563, 0, 22000, 131, 475, 0, 15,…
$ actor_3_facebook_likes   <int> 855, 1000, 161, 23000, NA, 530, 40…
$ actor_2_name             <chr> "Joel David Moore", "Orlando Bloom…
$ actor_1_facebook_likes   <int> 1000, 40000, 11000, 27000, 131, 64…
$ gross                    <int> 760505847, 309404152, 200074175, 4…
$ genres                   <chr> "Action|Adventure|Fantasy|Sci-Fi",…
$ actor_1_name             <chr> "CCH Pounder", "Johnny Depp", "Chr…
$ movie_title              <chr> "Avatar ", "Pirates of the Caribbe…
$ num_voted_users          <int> 886204, 471220, 275868, 1144337, 8…
$ cast_total_facebook_likes <int> 4834, 48350, 11700, 106759, 143, 1…
$ actor_3_name             <chr> "Wes Studi", "Jack Davenport", "St…
$ facenumber_in_poster     <int> 0, 0, 1, 0, 0, 1, 0, 1, 4, 3, 0, 0…
$ plot_keywords            <chr> "avatar|future|marine|native|parap…
$ movie_imdb_link          <chr> "http://www.imdb.com/title/tt04995…
$ num_user_for_reviews     <int> 3054, 1238, 994, 2701, NA, 738, 19…
$ language                 <chr> "English", "English", "English", "…
$ country                  <chr> "USA", "USA", "UK", "USA", "", "US…
$ content_rating           <chr> "PG-13", "PG-13", "PG-13", "PG-13"…
$ budget                   <dbl> 237000000, 300000000, 245000000, 2…
$ title_year               <int> 2009, 2007, 2015, 2012, NA, 2012, …
$ actor_2_facebook_likes   <int> 936, 5000, 393, 23000, 12, 632, 11…
$ imdb_score               <dbl> 7.9, 7.1, 6.8, 8.5, 7.1, 6.6, 6.2,…
$ aspect_ratio             <dbl> 1.78, 2.35, 2.35, 2.35, NA, 2.35, …
$ movie_facebook_likes     <int> 33000, 0, 85000, 164000, 0, 24000,…
```

outcome

- 5043 movie observations
- 28 variables
- across 100 years
- 66 countries

Main Variables We Choose

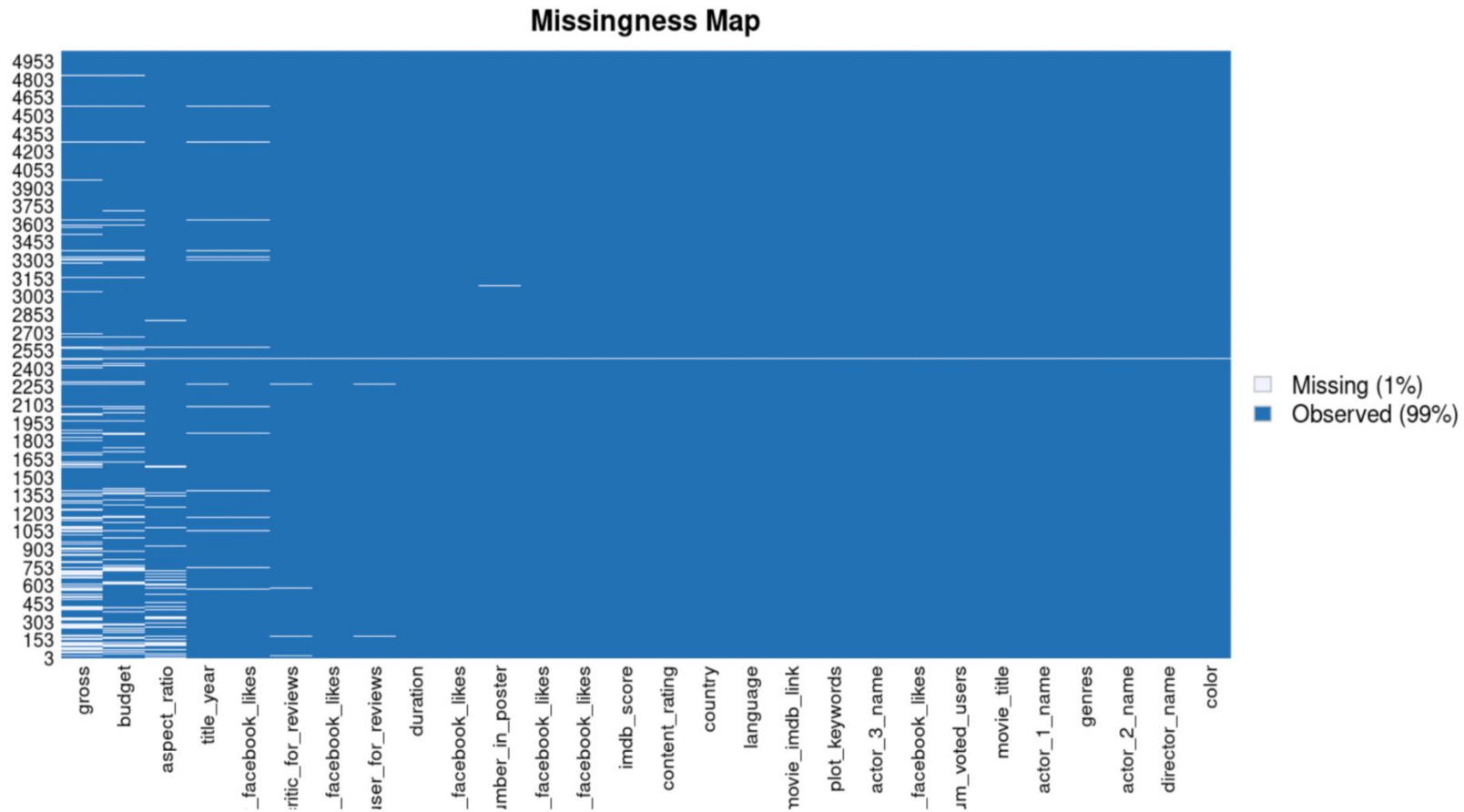
Main Variables

IMDB score	the score of the movie on IMDB.com
Director_name	the names of the directors
Num_critic_for_reviews	numbers of Critic reviews
Duration	the time of the movie
Director_facebook_likes	number of likes of the director on his/her Facebook Page
Actor_facebook_likes	number of likes of the actor on his/her Facebook Page
Movie_facebook_likes	number of likes of the movie on the Facebook Page
Gross	the total revenue of the movie
Budget	the budget of the movie

Cleaning the dataset

Missing Data Summary

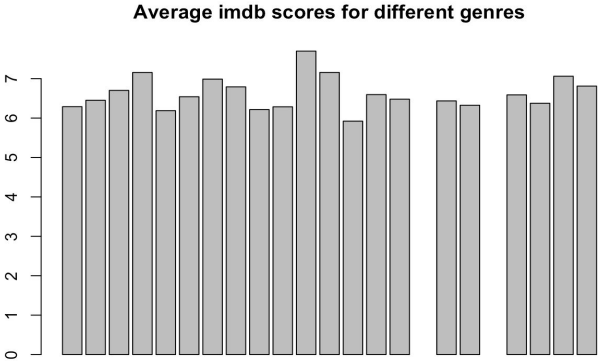
Missing data in selected sub sample.

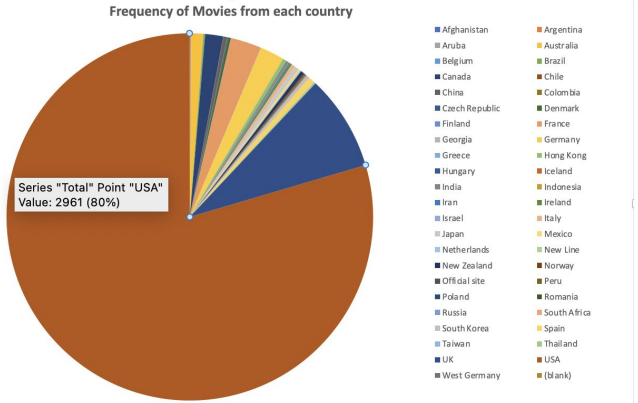


Clean the data (from 5043 rows to 2961)

Things we did:

- Deal with NA & 0s : Replace with average
- Delete some columns:
  (facenumber_in_poster)(plot_keywords)(aspect_ratio)
  (actor_2_name)(actor_3_name)(genres)
- Aggregate the columns: actor_facebook_likes to
  actor_1+2+3_facebook likes
- delete the movies from other countries, focus on USA market
  (80+% of the dataset!)
- Clean some outliers





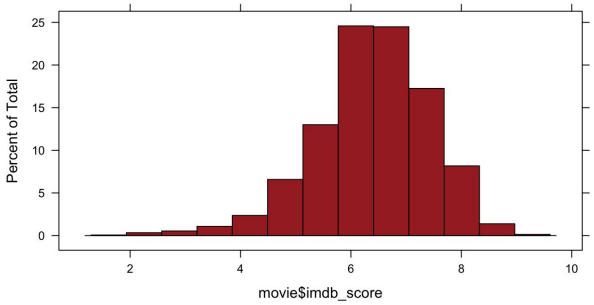
Before VS After

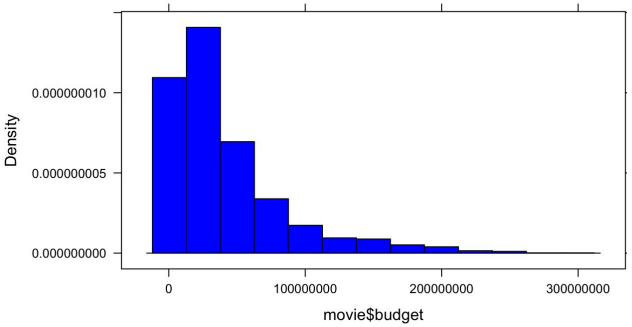
```
> glimpse(movie)
Rows: 5,043
Columns: 28
$ color                    <chr> "Color", "Color", "Color", "Color"…
$ director_name            <chr> "James Cameron", "Gore Verbinski",…
$ num_critic_for_reviews   <int> 723, 302, 602, 813, NA, 462, 392, …
$ duration                 <int> 178, 169, 148, 164, NA, 132, 156, …
$ director_facebook_likes  <int> 0, 563, 0, 22000, 131, 475, 0, 15,…
$ actor_3_facebook_likes   <int> 855, 1000, 161, 23000, NA, 530, 40…
$ actor_2_name             <chr> "Joel David Moore", "Orlando Bloom…
$ actor_1_facebook_likes   <int> 1000, 40000, 11000, 27000, 131, 64…
$ gross                    <int> 760505847, 309404152, 200074175, 4…
$ genres                   <chr> "Action|Adventure|Fantasy|Sci-Fi",…
$ actor_1_name             <chr> "CCH Pounder", "Johnny Depp", "Chr…
$ movie_title              <chr> "Avatar ", "Pirates of the Caribbe…
$ num_voted_users          <int> 886204, 471220, 275868, 1144337, 8…
$ cast_total_facebook_likes <int> 4834, 48350, 11700, 106759, 143, 1…
$ actor_3_name             <chr> "Wes Studi", "Jack Davenport", "St…
$ facenumber_in_poster     <int> 0, 0, 1, 0, 0, 1, 0, 1, 4, 3, 0, 0…
$ plot_keywords            <chr> "avatar|future|marine|native|parap…
$ movie_imdb_link          <chr> "http://www.imdb.com/title/tt04995…
$ num_user_for_reviews     <int> 3054, 1238, 994, 2701, NA, 738, 19…
$ language                 <chr> "English", "English", "English", "…
$ country                  <chr> "USA", "USA", "UK", "USA", "", "US…
$ content_rating           <chr> "PG-13", "PG-13", "PG-13", "PG-13"…
$ budget                   <dbl> 237000000, 300000000, 245000000, 2…
$ title_year               <int> 2009, 2007, 2015, 2012, NA, 2012, …
$ actor_2_facebook_likes   <int> 936, 5000, 393, 23000, 12, 632, 11…
$ imdb_score               <dbl> 7.9, 7.1, 6.8, 8.5, 7.1, 6.6, 6.2,…
$ aspect_ratio             <dbl> 1.78, 2.35, 2.35, 2.35, NA, 2.35, …
$ movie_facebook_likes     <int> 33000, 0, 85000, 164000, 0, 24000,…
```

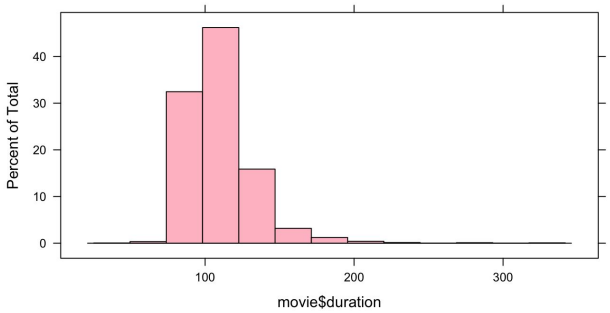
```
> glimpse(movie)
Rows: 2,961
Columns: 13
$ movie_title              <chr> "Avatar ", "Pirates of the Caribbean: At World's End ", "The …
$ director_name            <chr> "James Cameron", "Gore Verbinski", "Christopher Nolan", "Andr…
$ actor_1_name             <chr> "CCH Pounder", "Johnny Depp", "Tom Hardy", "Daryl Sabara", "J…
$ num_critic_for_reviews   <int> 723, 302, 813, 462, 392, 324, 635, 673, 434, 313, 450, 733, 2…
$ duration                 <int> 178, 169, 164, 132, 156, 100, 141, 183, 169, 151, 150, 143, 1…
$ director_facebook_likes  <int> 976, 563, 22000, 475, 976, 15, 976, 976, 976, 563, 563, 976, …
$ actor_facebook_likes     <int> 2791, 46000, 73000, 1802, 39000, 1636, 66000, 21000, 28903, 4…
$ movie_facebook_likes     <int> 33000, 17372, 164000, 24000, 17372, 29000, 118000, 197000, 17…
$ gross                    <int> 760505847, 309404152, 448130642, 73058679, 336530303, 2008072…
$ num_user_for_reviews     <int> 3054, 1238, 2701, 738, 1902, 387, 1117, 3018, 2367, 1832, 711…
$ country                  <chr> "USA", "USA", "USA", "USA", "USA", "USA", "USA", "USA", "USA"…
$ budget                   <int> 237000000, 300000000, 250000000, 263700000, 258000000, 260000…
$ imdb_score               <dbl> 7.9, 7.1, 8.5, 6.6, 6.2, 7.8, 7.5, 6.9, 6.1, 7.3, 6.5, 7.2, 6…
>
```

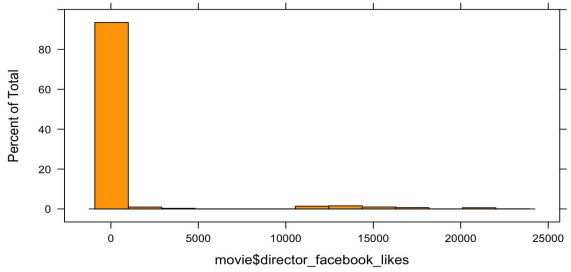
Data description & Visualization

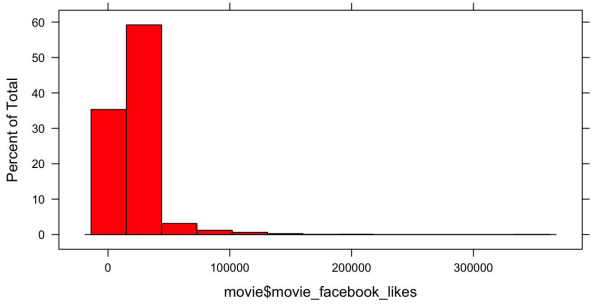
Histogram Data visualization and description

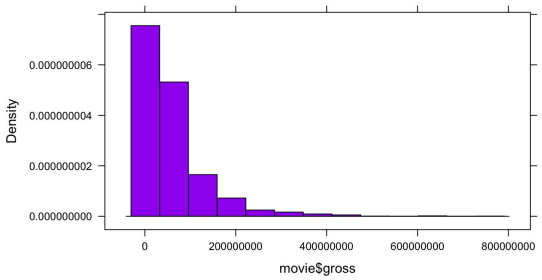












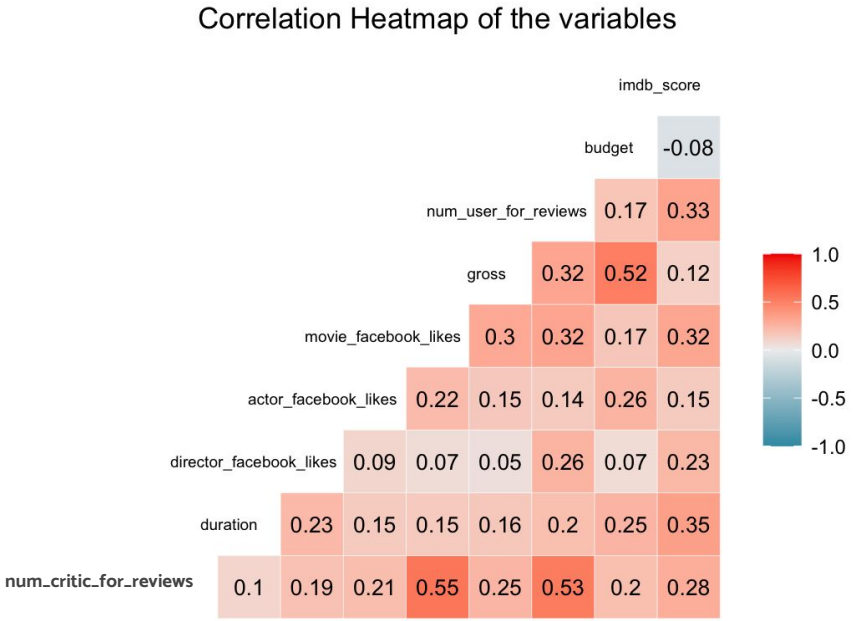
Glimpse of the movie dataset : **1st quartile - 3rd quartile(interquartile range)**

movie duration: [95-120] Min
gross revenue: [13 M-76 M]$

iMDB Score: [5-8] score
budget [11 M -50 M]$

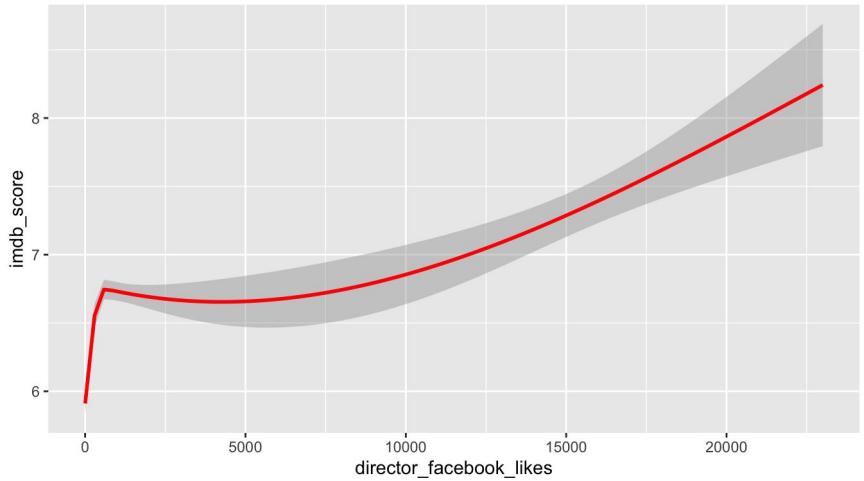
director facebook likes: [41-800]
movie facebook likes[5000-17372]

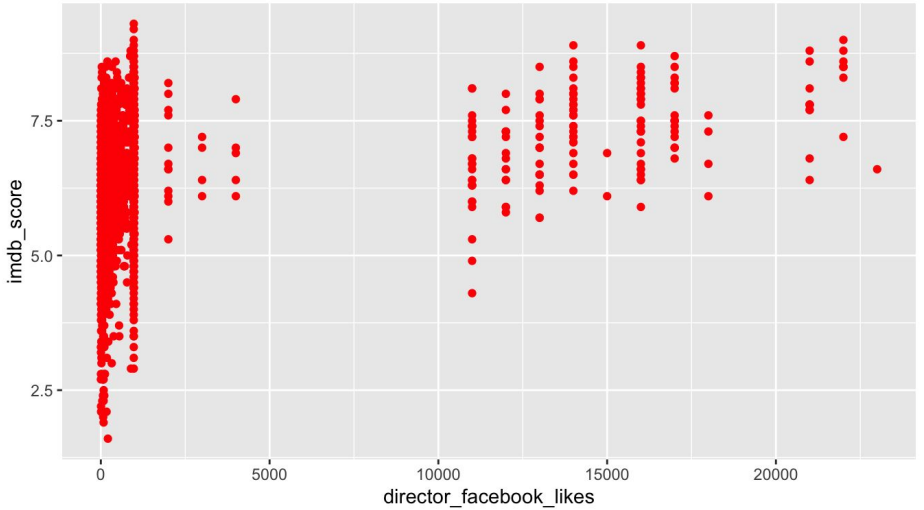
Correlation Heatmap



- Not so many high correlations between each variables, the maximum correlation score is 0.55 num_critic_for_reviews VS Movie facebook_likes
- Good News: No indicates the obvious presence of multicollinearity

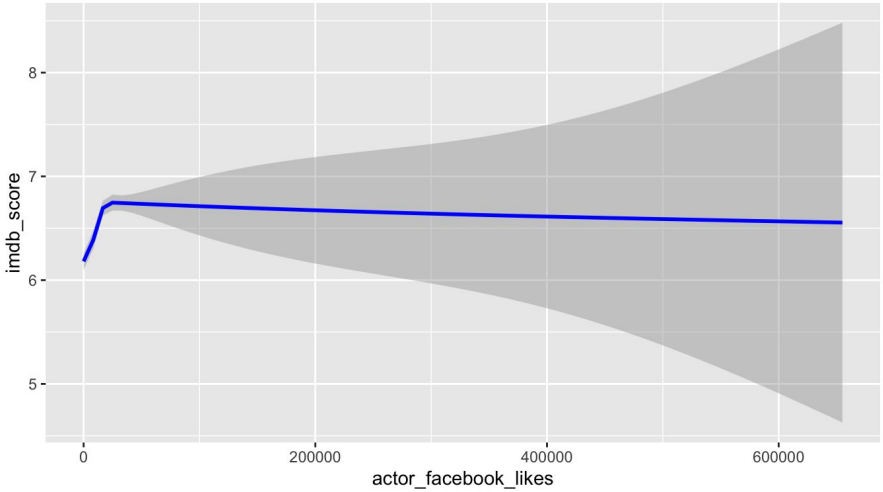
Directors' Effect on IMDB score

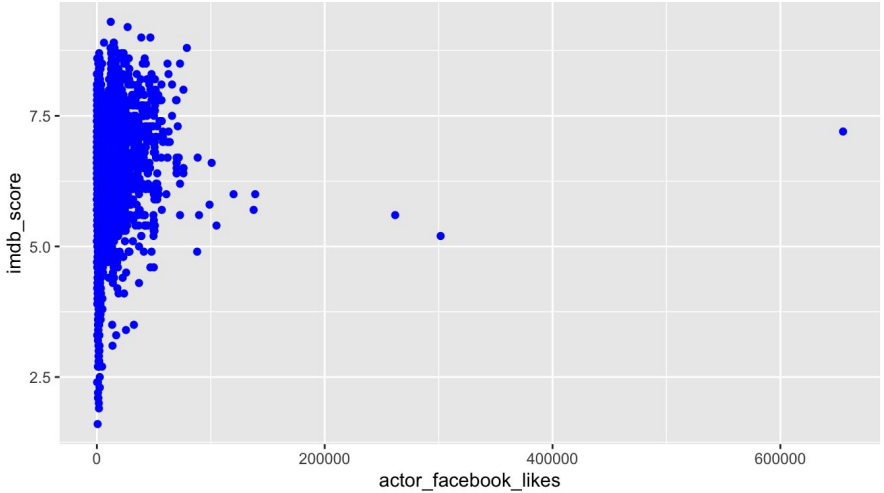




- X=director_facebook_likes, Y=imdb_score
  Use ggplot geom_point and smooth to identify the relationship between Director's effectiveness on iMDB score
- Positive impact and linear relationship, big jump at the beginning, right skewed
- Dead data if directors have no social media account,some between (0-1000 facebook likes)

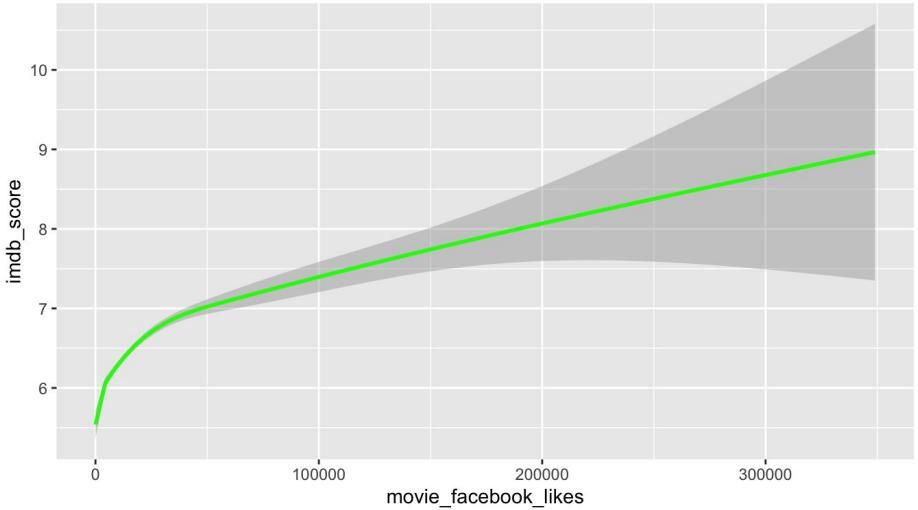
Actors' Effect on IMDB score (including all the leading actors)

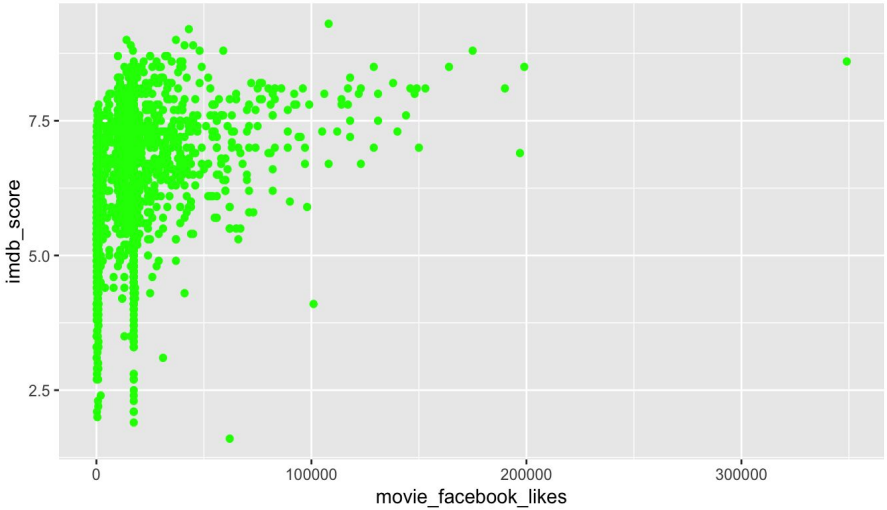




- Interesting! There is not so much relationship between actors' social media effectiveness and iMDB scores.
- Most actors' facebook likes are between (0-10,000)

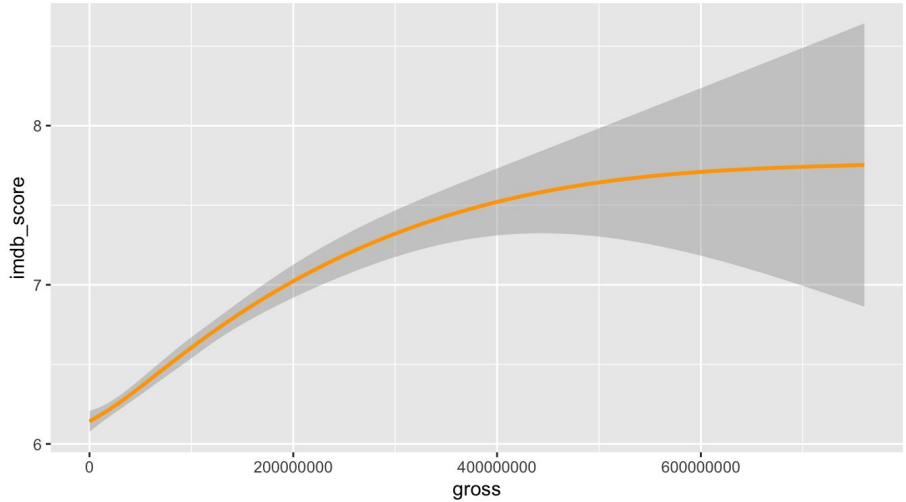
Movies' Effect on IMDB score

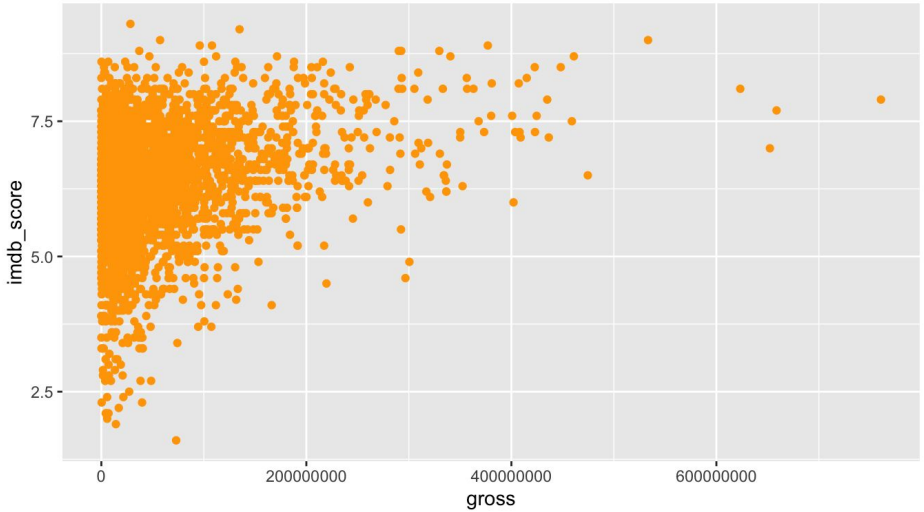




- Different from actors' effectiveness on iMDB scores. The movies facebook likes have a positive impact on iMDB scores. Reasonable !
- Dead data if movies has no facebook account,most likes are between (5000-15000)

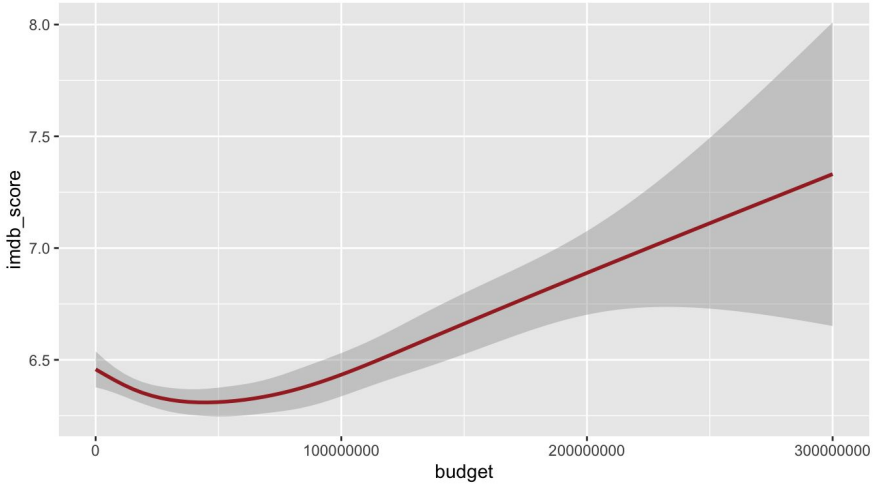
Does the movies' revenue effect iMDB score?

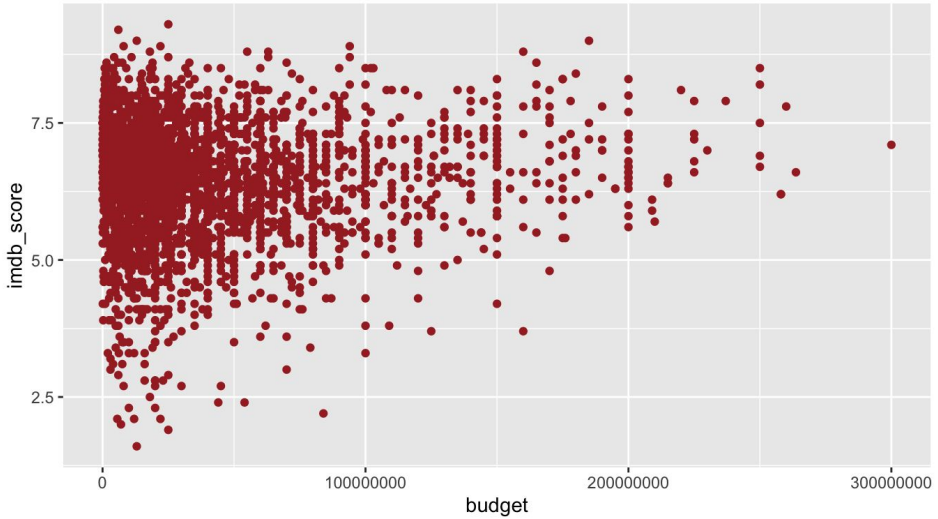




- Obviously, the revenue of the movies has a positive impact on iMDB scores which is very make sense!
- The iMDB scores will be more stable after the movie generate more than $400 Million in revenue

Let's explore the budget part, will it effect iMDB score?





- As our expectation, the budget does have an influence on the iMDB scores and shows the positive linear relationship
- Interesting! if the budget is between 0- 100 Million. There will be no correlation with iMDB scores
- if you invest more than 100 million dollars, the movie will have more chance to be successful.

Descriptive multiple linear regression

Estimate the full model = β0 + β1 ×(num_critic_for_reviews) + β2 ×(duration) + β3 ×(director_facebook_likes)+ β4 ×(actor_facebook_likes)+β5(movie_facebook_likes)+β6(gross revenue)+β7(num_user_for reviews) + β8(budget) + ϵ. .

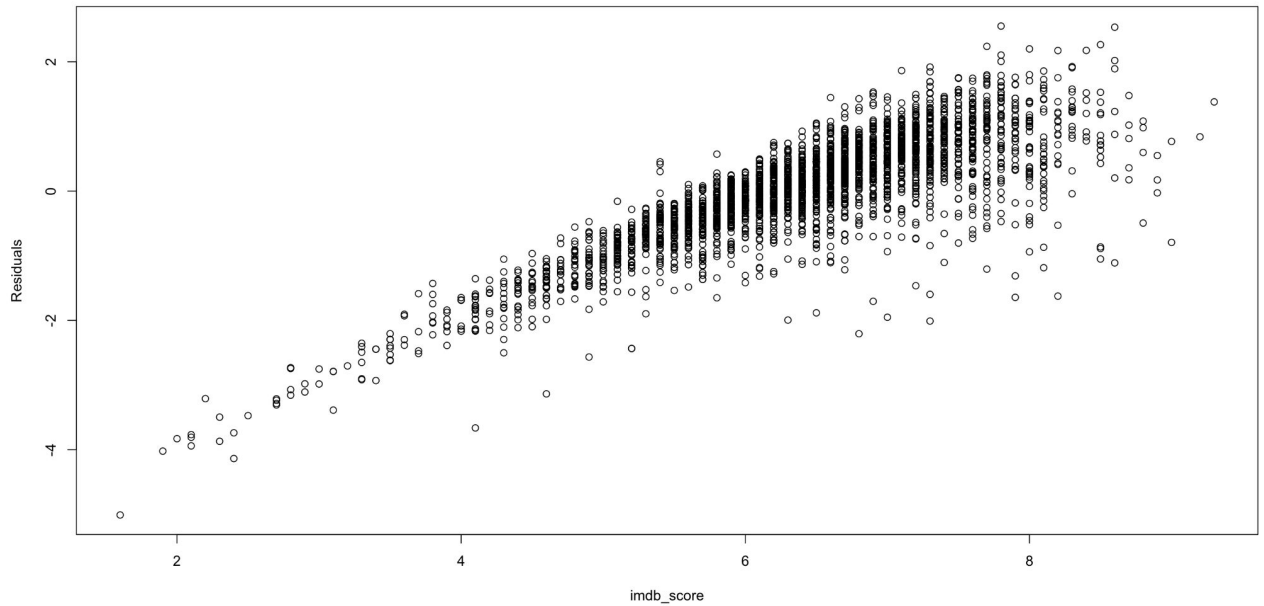
Using R

```
full.mod <- lm(imdb_score~.,data=movie_data)
summary(full.mod)
```

- Outcome: imdb_score 8 Predictors
- The reported p-values for most of the predictor variables are very close to zero.
- p-value: < 2.2e-16
- Therefore they are significant at 1% level. we can reject the null and conclude that the model is significant
- Multiple R-squared: 0.3044, Adjusted R-squared: 0.3025
- There is no evidence of multicollinearity. All of the signs are logical

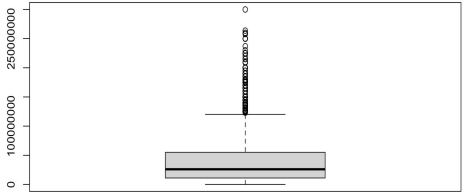
```
##
## Call:
## lm(formula = imdb_score ~ ., data = movie_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0094 -0.4698  0.0966  0.5921  2.5529
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               4.496e+00  8.550e-02  52.582  < 2e-16 ***
## num_critic_for_reviews    2.025e-03  2.065e-04   9.807  < 2e-16 ***
## duration                  1.404e-02  8.048e-04  17.445  < 2e-16 ***
## director_facebook_likes   2.905e-05  5.081e-06   5.717 1.19e-08 ***
## actor_facebook_likes      2.264e-06  8.381e-07   2.701 0.006951 **
## movie_facebook_likes      2.640e-06  1.122e-06   2.352 0.018743 *
## gross                     2.956e-09  3.192e-10   9.262  < 2e-16 ***
## num_user_for_reviews      1.932e-04  5.336e-05   3.621 0.000298 ***
## budget                   -7.914e-09  5.048e-10 -15.677  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.876 on 2952 degrees of freedom
## Multiple R-squared:  0.3044, Adjusted R-squared:  0.3025
## F-statistic: 161.5 on 8 and 2952 DF,  p-value: < 2.2e-16
```

Graph the residuals against imdb_score

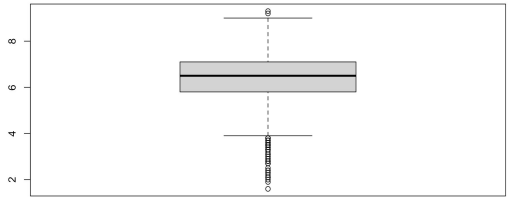


- A plot graph of the residuals VS the outcome values imdb_score
- The residuals seem to spread out across the predicted values,looks good!

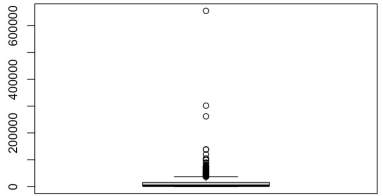
monitor some outliers in the dataset



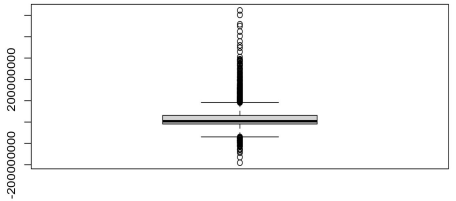
budget



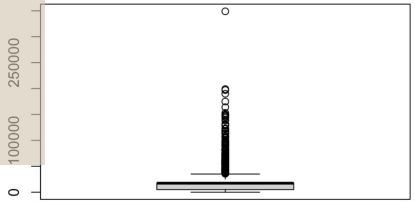
imdb score



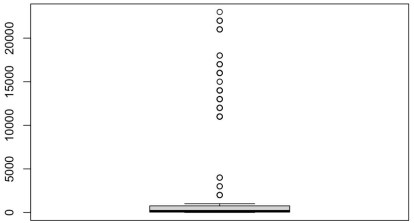
Actors' Facebook Likes



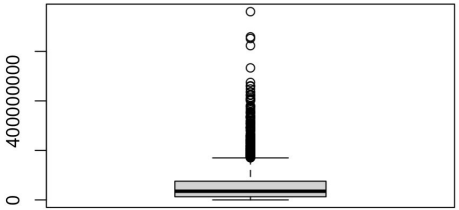
profit



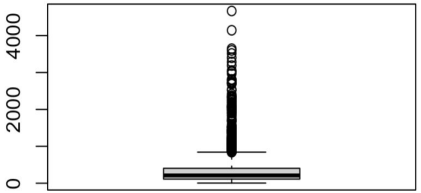
Movices' Facebook Likes



Directors' Facebook Likes



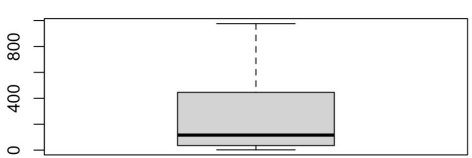
Gross Revenue



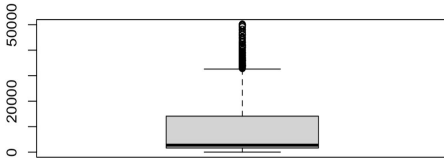
Num users for Reviews

- Most of the outliers are on the top side which shows for example: some movies have very high reviews, facebook likes, gross revenues, etc.
- the profit has both lower and higher outliers
- The imdb score has lower outliers which means some movies have lower scores

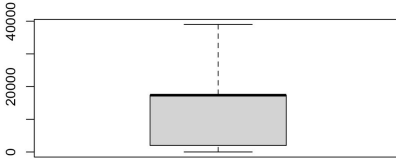
Clean the data: after clean some outliers looks better!



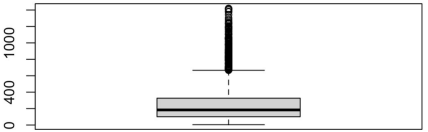
Directors' Facebook Likes



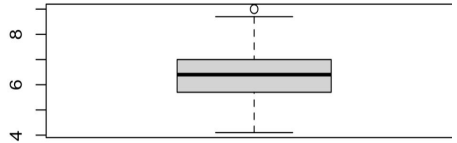
Actors' Facebook Likes



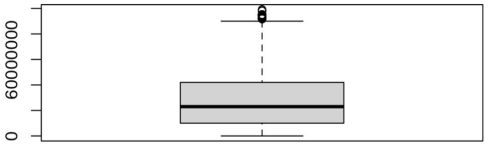
Movices' Facebook Likes



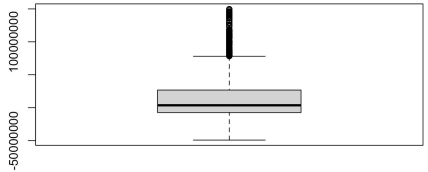
Num users for Reviews



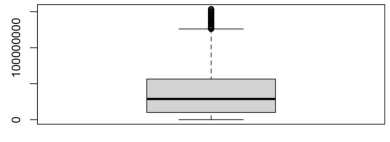
imdb score



budget



profit



Gross Revenue

- Before we clean the outliers, check the data manually in the CSV sheet to understand if it is a bad data or not.
- It is important to understand the reasons for the outliers before cleaning them.

```
> glimpse(movie)
Rows: 2,961
Columns: 13
$ movie_title              <chr> "Avatar ", "Pirates of the Caribbean: At World's End ", "The …
$ director_name            <chr> "James Cameron", "Gore Verbinski", "Christopher Nolan", "Andr…
$ actor_1_name             <chr> "CCH Pounder", "Johnny Depp", "Tom Hardy", "Daryl Sabara", "J…
$ num_critic_for_reviews   <int> 723, 302, 813, 462, 392, 324, 635, 673, 434, 313, 450, 733, 2…
$ duration                 <int> 178, 169, 164, 132, 156, 100, 141, 183, 169, 151, 150, 143, 1…
$ director_facebook_likes  <int> 976, 563, 22000, 475, 976, 15, 976, 976, 976, 563, 563, 976, …
$ actor_facebook_likes     <int> 2791, 46000, 73000, 1802, 39000, 1636, 66000, 21000, 28903, 4…
$ movie_facebook_likes     <int> 33000, 17372, 164000, 24000, 17372, 29000, 118000, 197000, 17…
$ gross                    <int> 760505847, 309404152, 448130642, 73058679, 336530303, 2008072…
$ num_user_for_reviews     <int> 3054, 1238, 2701, 738, 1902, 387, 1117, 3018, 2367, 1832, 711…
$ country                  <chr> "USA", "USA", "USA", "USA", "USA", "USA", "USA", "USA", "USA"…
$ budget                   <int> 237000000, 300000000, 250000000, 263700000, 258000000, 260000…
$ imdb_score               <dbl> 7.9, 7.1, 8.5, 6.6, 6.2, 7.8, 7.5, 6.9, 6.1, 7.3, 6.5, 7.2, 6…
>
```

Before

VS

After

```
> glimpse(movie_nooutlier)
Rows: 2,175
Columns: 9
$ num_critic_for_reviews   <int> 196, 175, 174, 188, 208, 97, 169, 200, 255, 82, 308, 198, 181, 143, 308, …
$ duration                 <int> 113, 127, 121, 101, 88, 110, 93, 152, 85, 89, 135, 107, 94, 88, 121, 128,…
$ director_facebook_likes  <int> 719, 57, 976, 976, 20, 342, 976, 976, 116, 10, 976, 70, 368, 7, 976, 323,…
$ actor_facebook_likes     <int> 1130, 26000, 2595, 3230, 222, 1524, 13465, 3377, 8660, 3361, 1743, 1519, …
$ movie_facebook_likes     <int> 17372, 17372, 2000, 16000, 33000, 578, 17372, 10000, 18000, 17372, 16000,…
$ gross                    <int> 119412921, 101643008, 66862068, 55673333, 130174897, 10200000, 59475623, …
$ num_user_for_reviews     <int> 391, 498, 524, 160, 155, 263, 407, 1103, 177, 139, 593, 125, 643, 222, 57…
$ budget                   <int> 65000000, 90000000, 83000000, 80000000, 99000000, 10000000, 90000000, 920…
$ imdb_score               <dbl> 7.5, 6.5, 5.7, 7.0, 7.2, 7.3, 5.2, 7.7, 6.2, 6.9, 6.7, 6.6, 5.5, 7.5, 7.0…
```

Without outlier

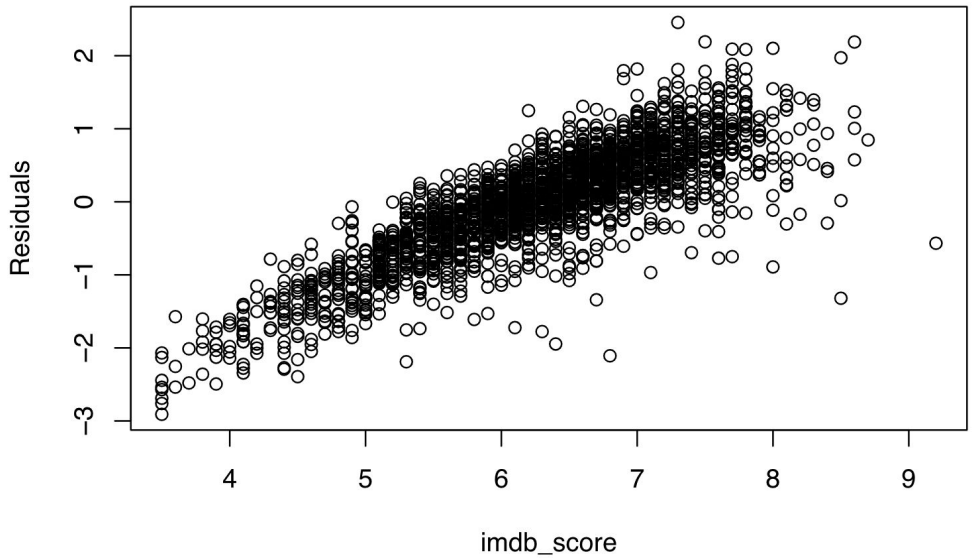
Descriptive multiple linear regression (no the outliers)

Estimate the full model = β0 + β1 ×(num_critic_for_reviews) + β2 ×(duration) + β3 ×(director_facebook_likes)+ β4 ×(actor_facebook_likes)+β5(movie_facebook_likes)+β6(gross revenue)+β7(num_user_for reviews) + β8(budget) + ϵ. .

- Outcome: imdb_score 8 Predictors
- Standard error getting increased
- R-squared is the same =0.3084
- Not so different from the last model(with outliers)
- same p-value: < 2.2e-16
- the predictors: movie_facebook_likes and actor_facebook_likes are more significant

```
## Call:
## lm(formula = imdb_score ~ ., data = movie_nooutlier)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.90950 -0.43701  0.08441  0.52654  2.45602
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               4.448e+00  9.019e-02  49.314  < 2e-16 ***
## num_critic_for_reviews    7.377e-04  2.430e-04   3.036  0.00243 **
## duration                  1.403e-02  8.654e-04  16.214  < 2e-16 ***
## director_facebook_likes   2.532e-04  4.647e-05   5.448 5.68e-08 ***
## actor_facebook_likes      8.915e-06  2.192e-06   4.068 4.91e-05 ***
## movie_facebook_likes      1.774e-05  2.252e-06   7.876 5.29e-15 ***
## gross                     2.369e-09  5.986e-10   3.957 7.84e-05 ***
## num_user_for_reviews      5.434e-04  7.585e-05   7.163 1.07e-12 ***
## budget                   -1.251e-08  9.122e-10 -13.718  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7743 on 2166 degrees of freedom
## Multiple R-squared:  0.3084, Adjusted R-squared:  0.3058
## F-statistic: 120.7 on 8 and 2166 DF,  p-value: < 2.2e-16
```

Graph the residuals against imdb_score (without the outliers)



- more residuals , might be the increase of the standard errors.
- the regression model doesn't get improved for the dataset without the outliers

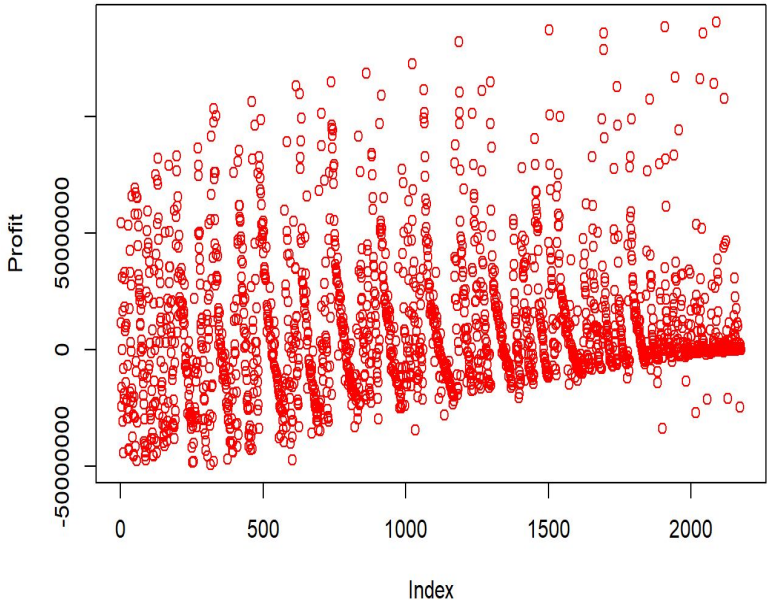
KNN

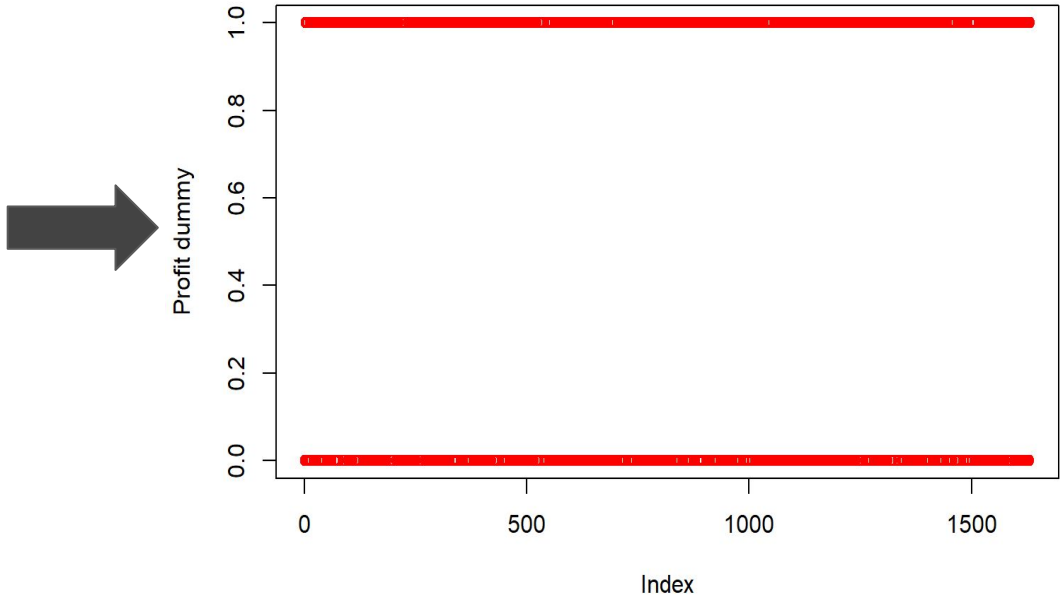

TOP 100 MOVIES WATCHLIST
THE HIGHEST-RATED IMDb CROWD FAVORITES
POP CORN

Predictors

IMDB score	the score of the movie on IMDB.com
Number Users reviews	number of regular imdb user reviews
Number critic reviews	numbers of Critic reviews
Duration	the time of the movie
Director facebook likes	number of likes of the director on his/her Facebook Page
Actor facebook likes	number of likes of the Actor on his/her Facebook Page
movie facebook likes	number of likes of the movie on the Facebook Page
Gross	the total revenue of the movie
Budget	the budget of the movie

Gross - Budget was used to make Profit variable





- Visualization of profit dummy variable for classification
- 2 classes: profit > $0 “1”
  profit < $0 “0”

KNN method

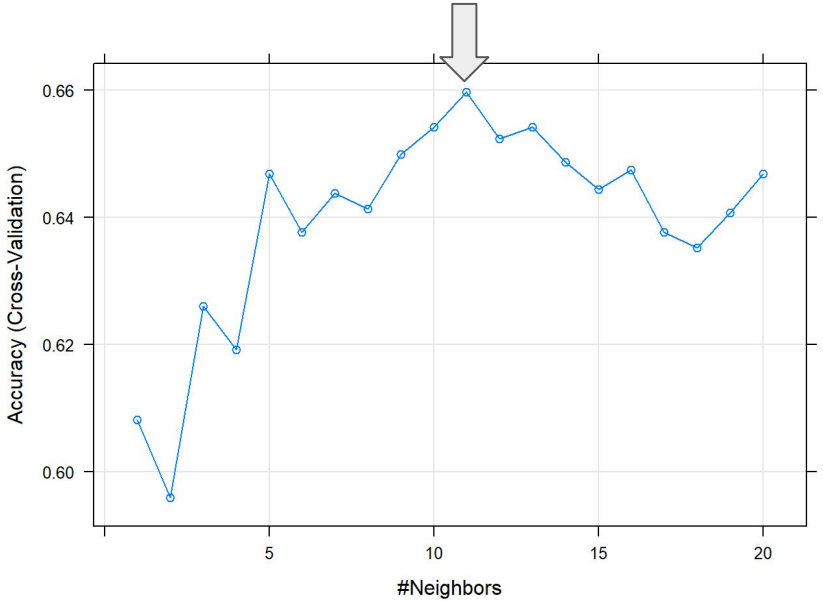
Goal: predict if a film will be profitable using 8 predictors

Parameters:

- 70/30 split
- 2 classes: profit > $0 “1”
  profit < $0 “0”
- Tested for k=20
- 10 folds
- Optimal K = 11

```
k-Nearest Neighbors

1631 samples
   8 predictor
   2 classes: '0', '1'

Pre-processing: centered (8), scaled (8)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1468, 1468, 1467, 1468, 1468, 1468, ...
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   1  0.6081633  0.1984277
   2  0.5959230  0.1742815
   3  0.6259922  0.2257276
   4  0.6192323  0.2104030
   5  0.6468549  0.2649707
   6  0.6376638  0.2450833
   7  0.6437989  0.2560643
   8  0.6413149  0.2515469
   9  0.6498887  0.2665147
  10  0.6541981  0.2741789
  11  0.6597198  0.2833732
  12  0.6523353  0.2678341
  13  0.6541757  0.2711618
  14  0.6486805  0.2586902
  15  0.6443859  0.2482090
  16  0.6474272  0.2573613
  17  0.6376636  0.2358597
  18  0.6352059  0.2306217
  19  0.6407313  0.2416061
  20  0.6468438  0.2535160

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 11.
```

Training set



Confusion Matrix		
	Predicted Class	
Actual Class	0	1
0	398	159
1	301	773

Statistics:	
Accuracy	0.718
95% CI:	(0.69 : 0.74)
Sensitivity	0.8294
Specificity	0.5694
Pos Pred Value	0.7197
Neg Pred Value	0.7145

- 71.8% accuracy on training set
- Sensitivity at 82.94%
- Low Specificity at 56.94%

Test Set

.5 cut off

Confusion Matrix		
	Predicted Class	
Actual Clas	0	1
0	121	89
1	126	208

Statistics:	
Accuracy	0.6048
95% CI:	(0.56 : 0.65)
Sensitivity	0.7003
Specificity	0.4899
Pos Pred Value	0.6228
Neg Pred Value	0.5762

.7 cut off

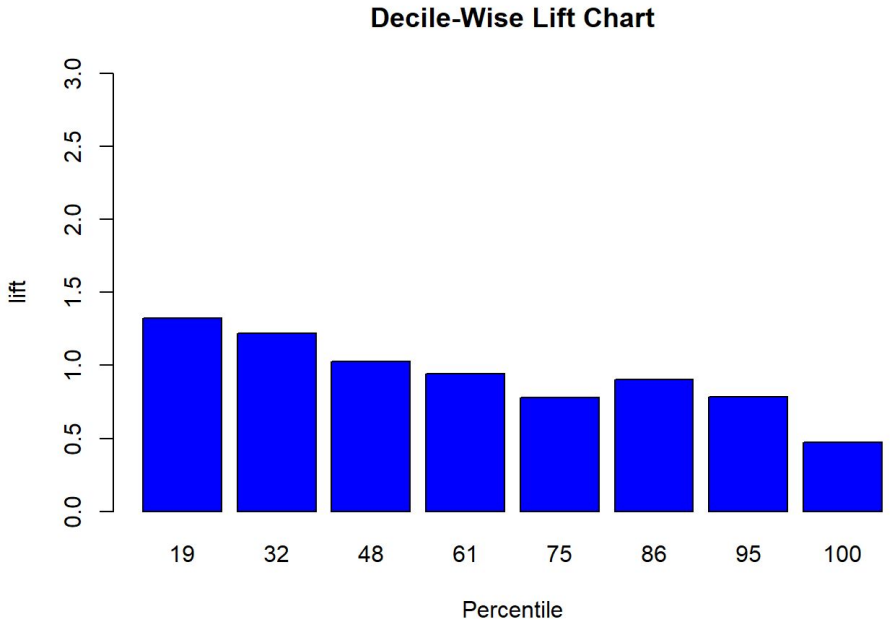
Confusion Matrix		
	Predicted Class	
Actual Class	0	1
0	79	58
1	168	239

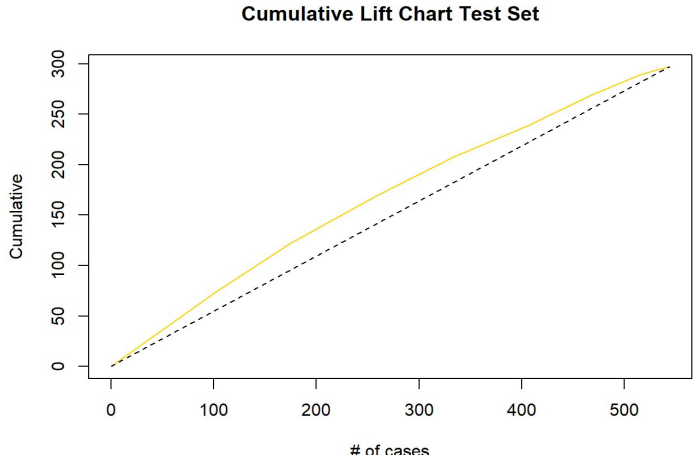
Statistics:	
Accuracy	0.5846
95% CI:	(0.54 : 0.63)
Sensitivity	0.41
Specificity	0.7854
Pos Pred Value	0.6989
Neg Pred Value	0.5272

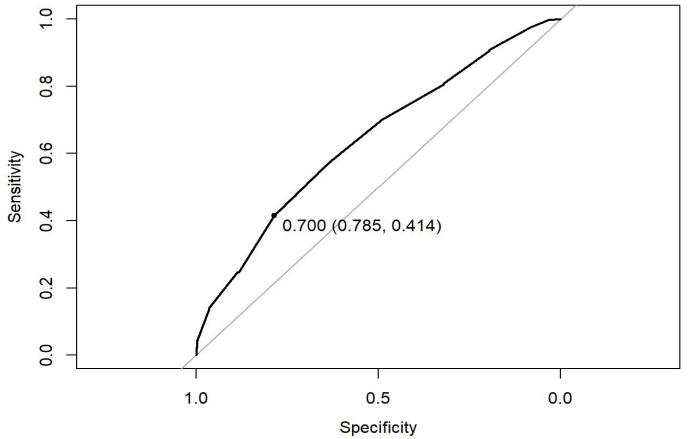
- Accuracy is lower on test data
- Sign of overfitting

- .7 cutoff lowers accuracy but raised specificity
- when predicting profit, ensure more true negatives or positives?

Test data





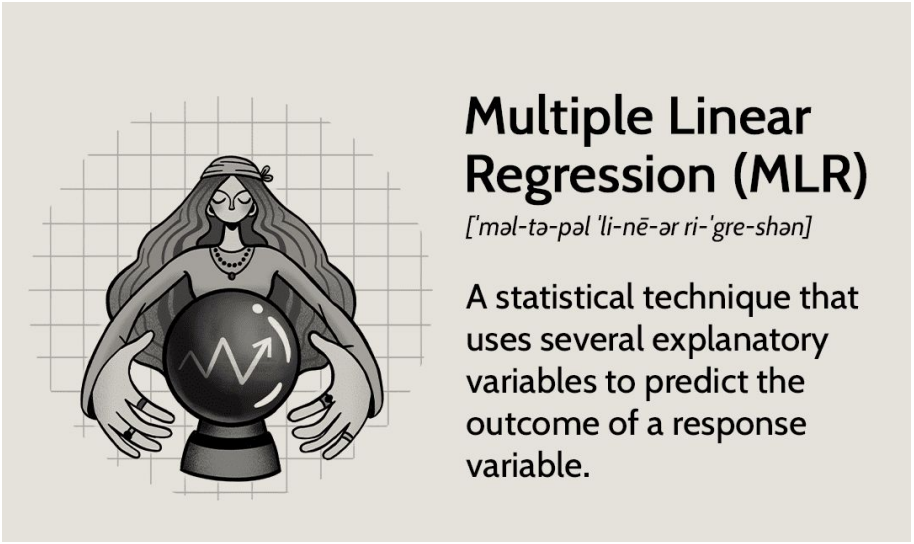


- Model performed slightly better than baseline
- Area under curve = .6403

Predictive Multiple Linear Regression Model

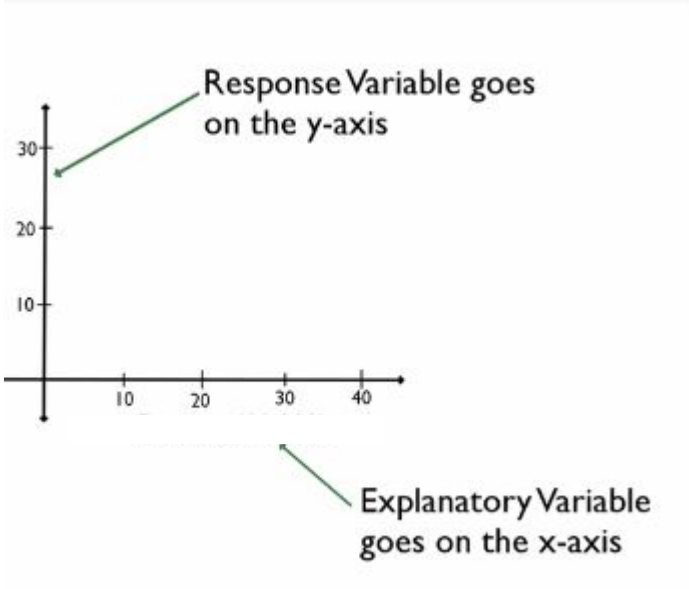
Multiple linear regression refers to a statistical technique that **uses two or more independent variables to predict the outcome of a dependent variable**.

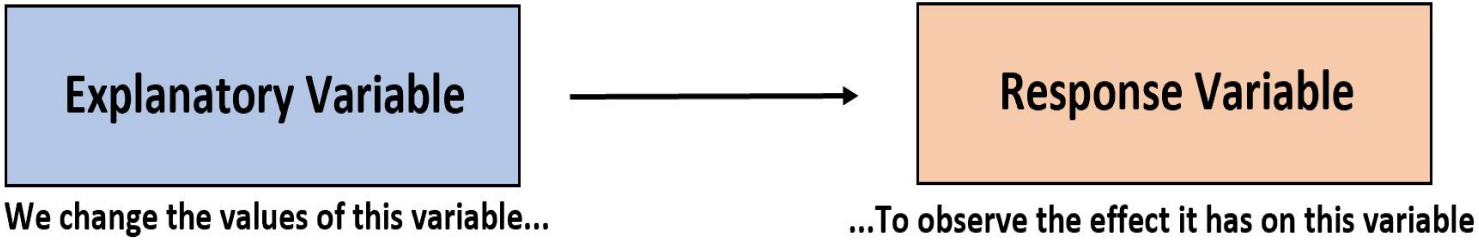
The technique enables analysts to determine the variation of the model and the relative contribution of each independent variable in the total variance

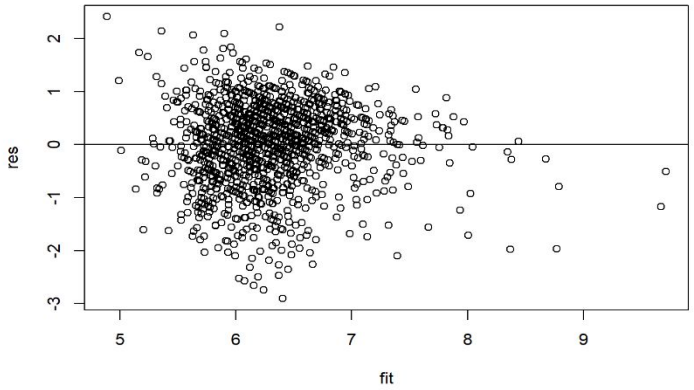


Following prediction variables are in used the linear regression model

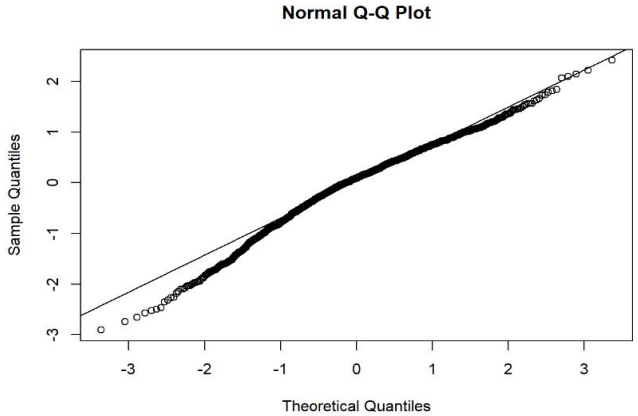
- IMDB Score (Response Variable)
- Number Users Reviews
- Number critic reviews
- Duration
- Director's Facebook likes
- Actor's Facebook likes
- Movie's Facebook likes
- Gross Revenue
- Budget of the Movie

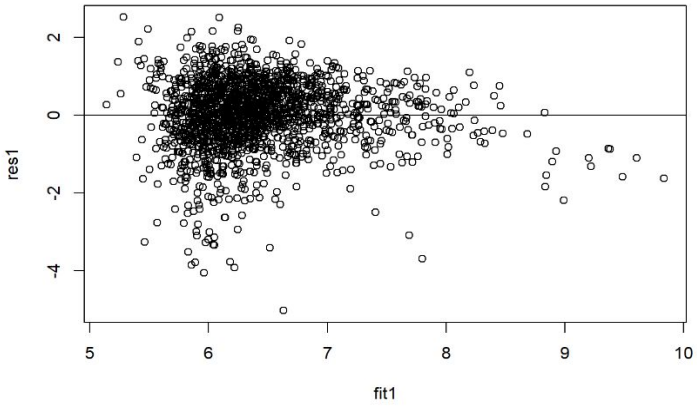




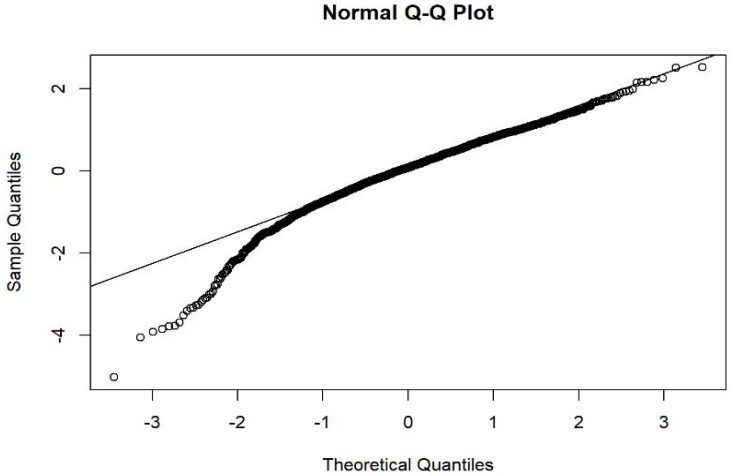


Without Outlier

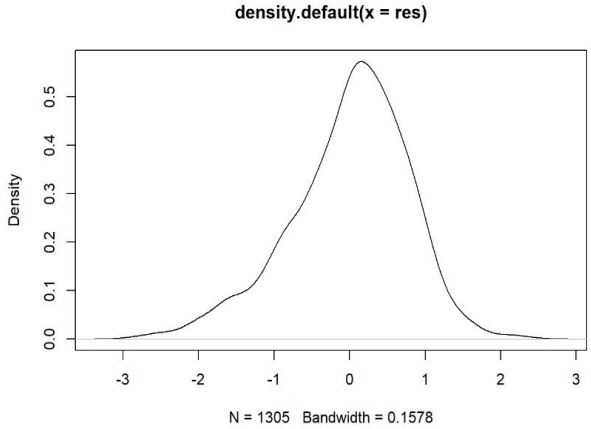




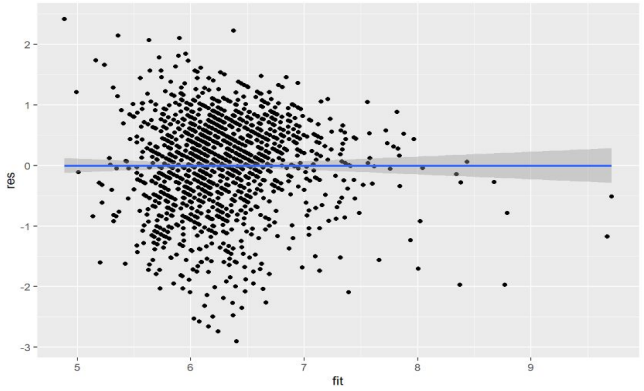
With Outlier

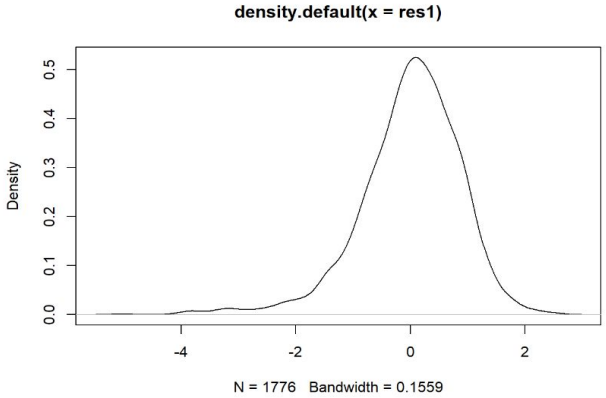


- A residual plot shows the difference between the observed response and the fitted response values.
- Residuals tend to stay close to the plotted line, indicating they are normally distributed.
- Previous slide graphs and plots indicates that the data is some-what non-linear.
- No Constant Variance
- The random pattern in the residuals indicates that the deterministic portion (predictor variables) of the model is capturing some explanatory information that is “leaking” into the residuals which shows the sign of a good model.

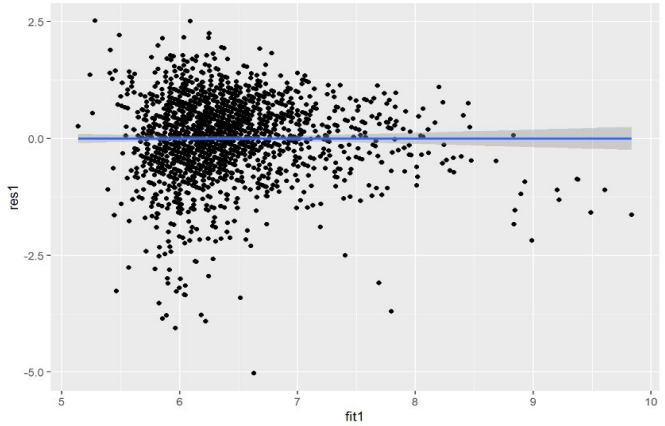


Without Outlier



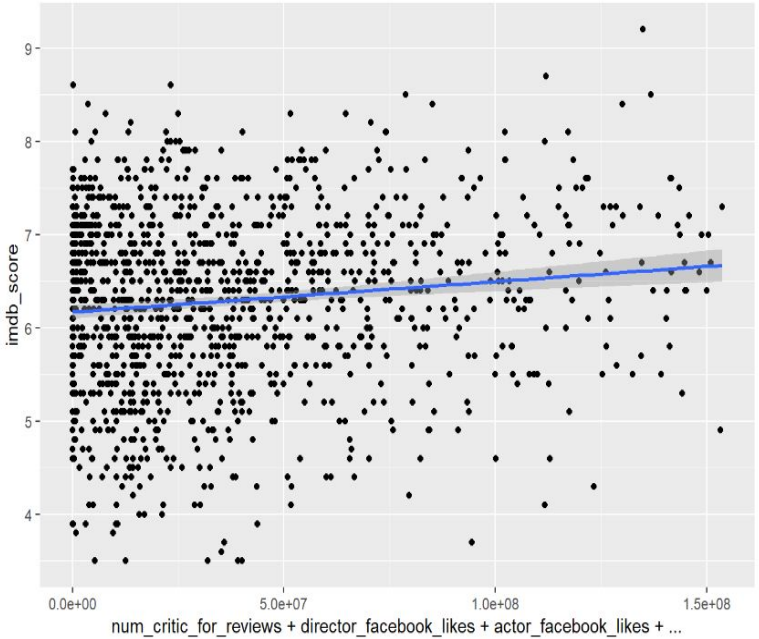


With Outlier

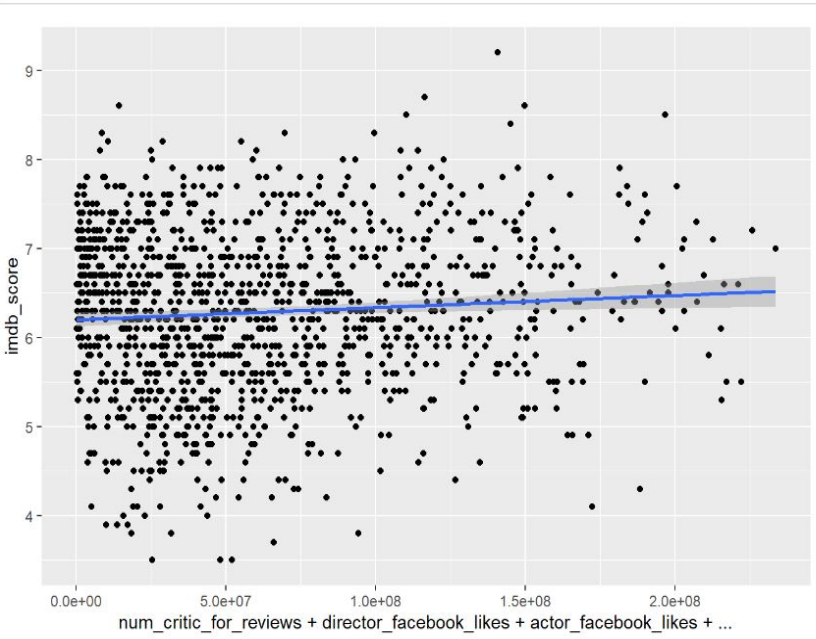


The density plot also shows a normal distribution
The density plot shows a rough bell-shaped symmetry with some values skewed to the left

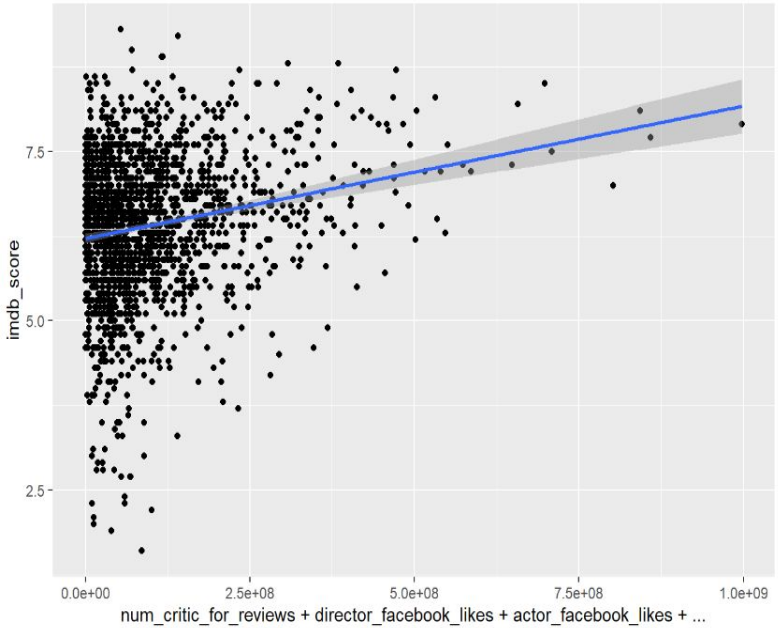
Choosing the best two models with and without outliers



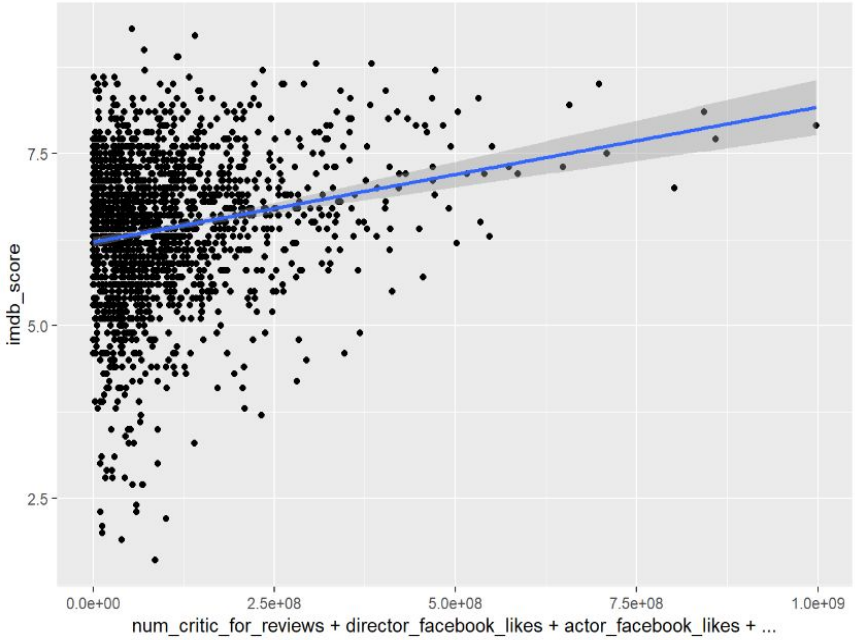
Without outliers




```
ggplot(model3a, aes(x = num_critic_for_reviews
                    +director_facebook_likes+actor_facebook_likes+movie_facebook_likes+num_user_for_reviews+budget+gross+dur
ation, y =imdb_score)) +
  geom_point() +
  stat_smooth(method = "lm")
```



With Outlier



Adjusted R2 is a corrected goodness-of-fit (model accuracy) measure for linear models which means the higher the adjusted R2, the better the fit.

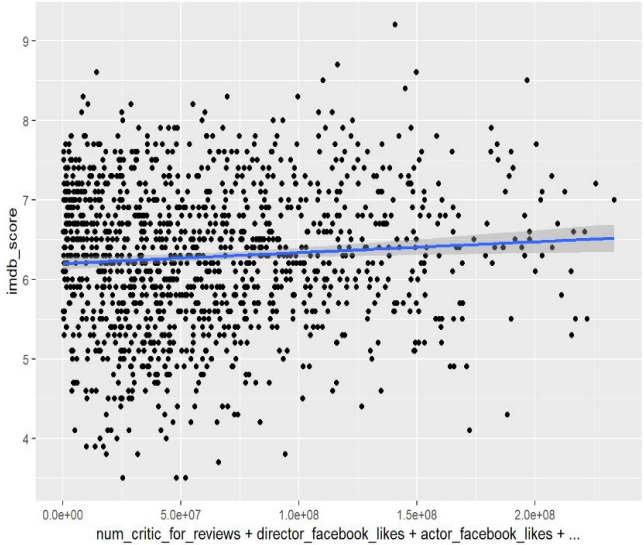
The four models that we selected two from outliers, two from without outliers are Model 2 and Model 3 which have highest Adjusted R square i.e. 0.23, 0.30, 0.23, and 0.29 respectively.
By this we can conclude that outliers does not have an effect on the models.

Choosing the Best Model

```
summary(model3)$adj.r.squared
```

```
## [1] 0.3071031
```

```
ggplot(model3, aes(x = num_critic_for_reviews
                   +director_facebook_likes+actor_facebook_likes+movie_facebook_likes+num_user_for_reviews+budget+gross+dura
tion, y =imdb_score)) +
  geom_point() +
  stat_smooth(method = "lm")
```

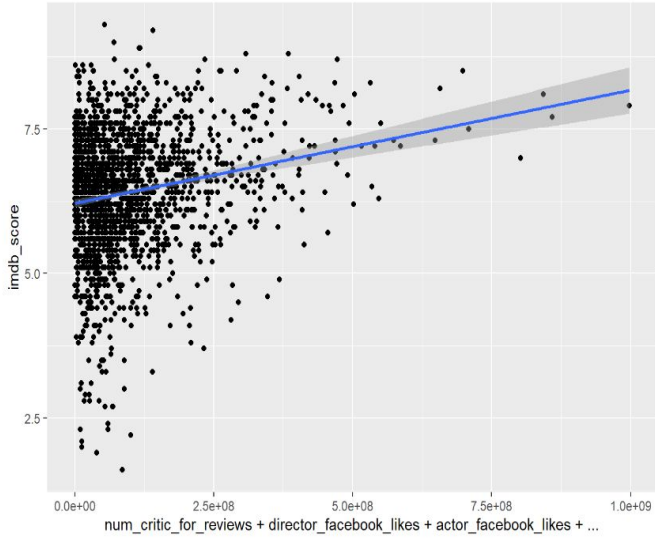



```
##                    ME      RMSE       MAE       MPE      MAPE
## Test set 0.007889586 0.803427 0.6163956 -1.878191 10.69455
```

```
summary(model3a)$adj.r.squared
```

```
## [1] 0.2992361
```

```
ggplot(model3a, aes(x = num_critic_for_reviews
                   +director_facebook_likes+actor_facebook_likes+movie_facebook_likes+num_user_for_reviews+budget+gross+dur
ation, y =imdb_score)) +
  geom_point() +
  stat_smooth(method = "lm")
```




```
##                    ME      RMSE       MAE       MPE      MAPE
## Test set -0.04389083 0.868644 0.6666109 -3.277308 11.94164
```

Conclusion


TOP 100 MOVIES WATCHLIST
THE HIGHEST-RATED IMDb CROWD FAVORITES
POP CORN

- **According to the dataset, the social media presence of directors and movies, revenue, and budget variables have more impact on the IMDB score compared to other variables.**

- **With the help of KNN model we can predict if a movie will be profitable with 60% accuracy given we have all the 8 predictors used in the model. More data types could raise the accuracy.**

- **Linear regression analysis might be a good way to describe and predict the model.**

- **According to the predictive multiple linear regression model, we can conclude that when we include all the variables with or without outliers(model 3), the more accurate the predicting model would be.**

- **In essence it is complex to get an accurate prediction of the movies because there are so many different factors we need to consider and explore.**

THANK YOU!

Welcome to ask questions!

Contact:

Suyash Saxena <ssaxe015@ucr.edu>