Saivee Phatak

OPSM324

Prof. Deepak Srivastav

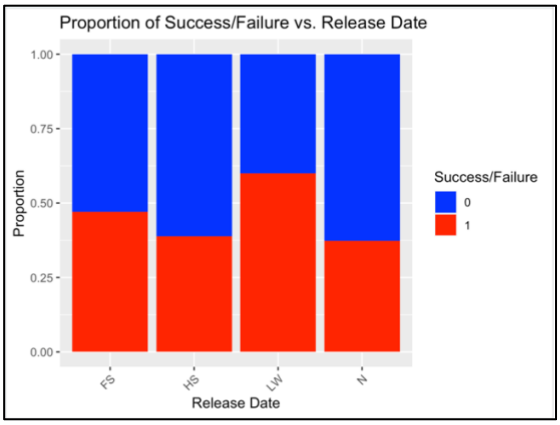
30th April 2024

## Final Submission – Box Office Analysis

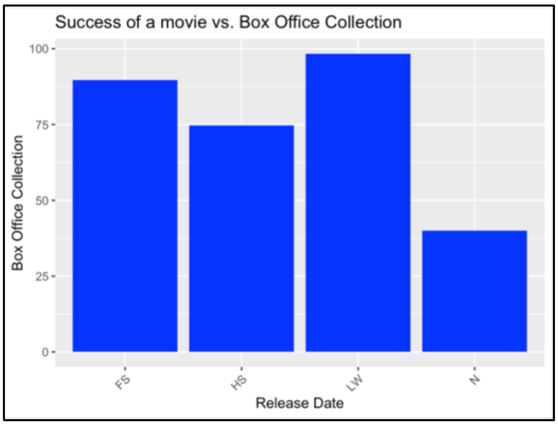
1. Using descriptive analytics find:

- Best season to release a movie. Explain the possible reason behind the same.

To see which the best season is to release a movie, I used descriptive analysis to visualise the following:



*Graph 1.1*



*Graph 1.2*

From both of the above graphs, we can see that *'long weekends'* is the most successful category to release a new movie. The rate of successes, checked in proportion with the rate of failure in Graph 1.1 clearly shows that *'long weekends'* have the highest success rate, followed by *'Festival Season'*.

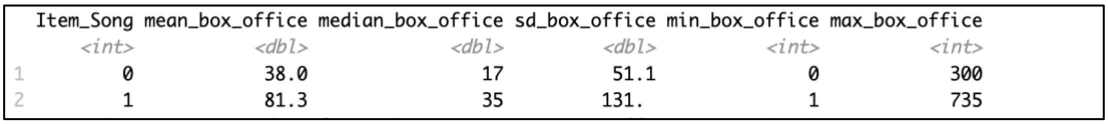
Corroborating that with the Box Office Collections and Release Dates graph, those movies that are released on the *'long weekends'* show the highest box office collection mean, while those that are not released during any season, i.e. *'no season'*, show the lowest earnings and success rates.

*'Long weekends'* proves to be a successful time to release movies especially because people have more free time as compared to just a regular weekend. During this time, people are more willing to spend their time doing leisurely activities such as watching a movie, which also might be considered to be time consuming on any other normal weekend.

Long weekends also sometimes mean that there might be some kind of occasion, where friends and family gather together. During this time, plans are made with a lot more enthusiasm since going to a movie becomes a group outing instead of just an individual one. This can also be the reason why the *'Festival season'* is the second most successful time to release movies.

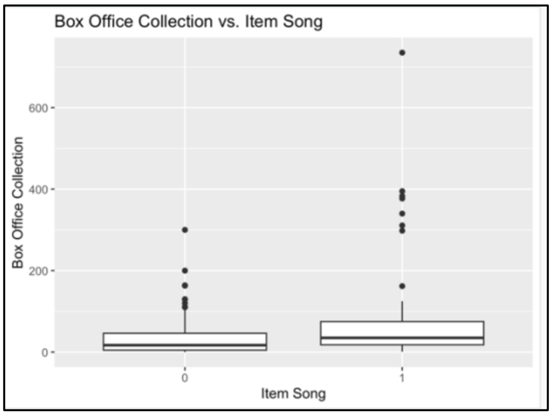
No season is not a good time to release movies since it might clash with important workdays and deadlines, school days, exams, and other tasks, which constitutes for a lack of time to carry out leisurely tasks.

- Does an item song make a difference in the budget/box office collection?



```
  Item_Song mean_box_office median_box_office sd_box_office min_box_office max_box_office
      <int>           <dbl>             <dbl>         <dbl>          <int>          <int>
1         0            38.0                17          51.1              0            300
2         1            81.3                35         131.               1            735
```

*Fig. 1.1*



*Graph 1.3*

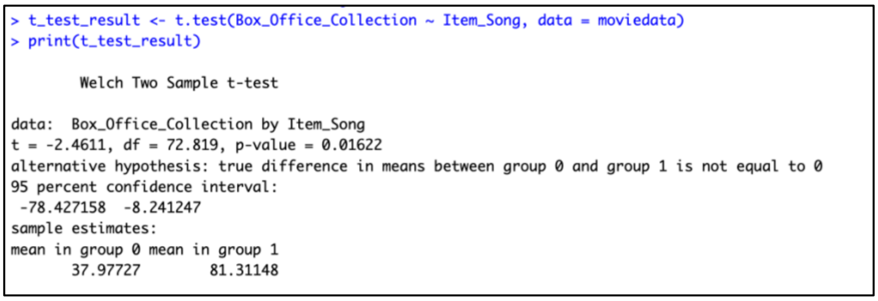
Through Fig. 1.1 and Graph 1.3 we can see that the mean for box office collection with item song is 81.3, and without item song is 38. The maximum box office collection for movies with item songs is also double that of movies with no item songs. To further support the claim that item songs do affect box office collection, a t-test was performed.

The T-Test was performed on the following:

H0: There is no difference in the box office collection of a movie with an item song vs. the box office collection of a movie without an item song

Ha: There is a significant difference in the box office collection of a movie with an item song vs. the box office collection of a movie without an item song

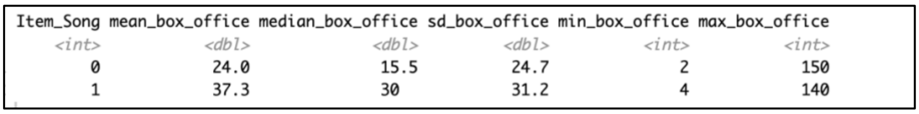
Based on the T-test that was conducted, we get the following results:



```
> t_test_result <- t.test(Box_Office_Collection ~ Item_Song, data = moviedata)
> print(t_test_result)

	Welch Two Sample t-test

data:  Box_Office_Collection by Item_Song
t = -2.4611, df = 72.819, p-value = 0.01622
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -78.427158  -8.241247
sample estimates:
mean in group 0 mean in group 1 
       37.97727        81.31148 
```

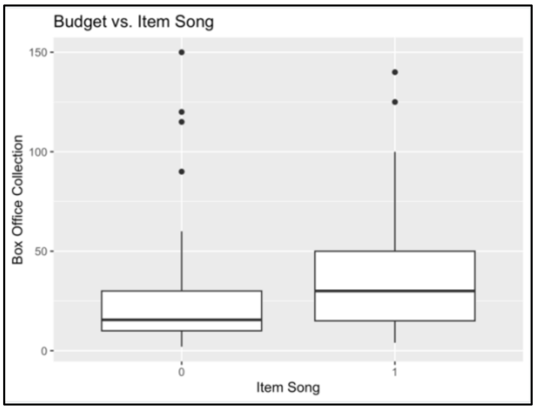
*Fig 1.2*

According to Fig 1.2, the p-value for the t-test is 0.01622, which is lower than the considered threshold of 0.05. This allows us to **reject the null hypothesis**, which states that there is no difference between the box office collection of a movie with an item song vs. the box office collection of a movie without an item song.



```
Item_Song mean_box_office median_box_office sd_box_office min_box_office max_box_office
    <int>           <dbl>             <dbl>         <dbl>          <int>          <int>
        0            24.0              15.5          24.7              2            150
        1            37.3              30            31.2              4            140
```

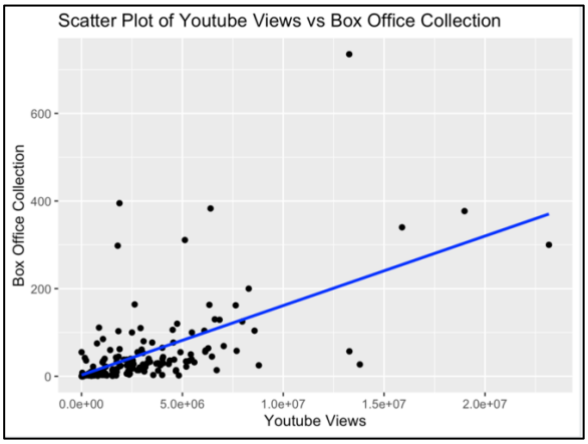
*Fig 1.3*



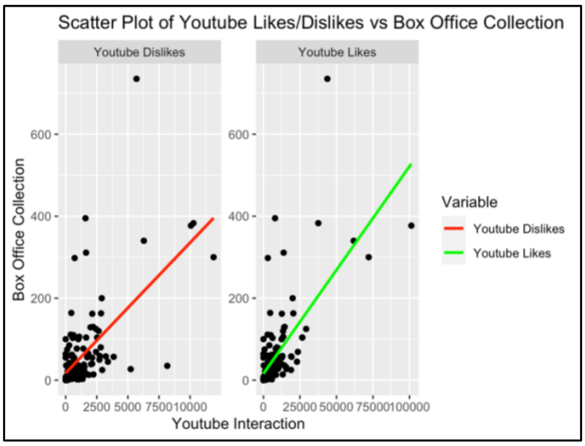
A box plot was also made to see the difference that the inclusion of an item song makes in the budget of producing a film. Through fig 1.3 and Graph 1.4 we can see that the mean does not vastly differ as it did for box office collection. The same can be said for the minimum and maximum budget for a film with or without an item song.

*Graph 1.4*

- How does digital medium impact box office collection?



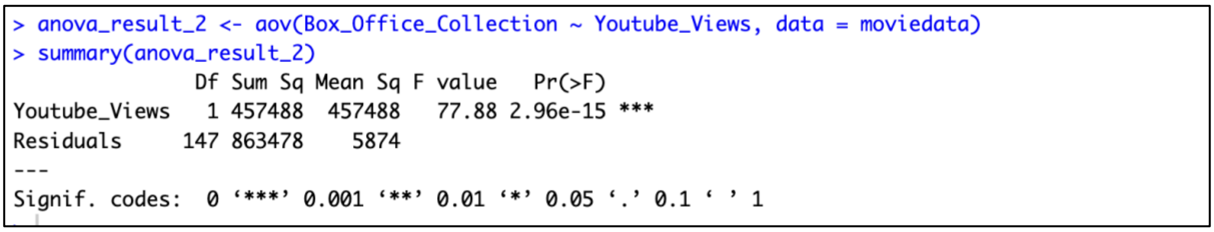
*Graph 1.5*



*Graph 1.6*

Graph 1.6 shows that the trajectory for box office collections based on the likes is steeper than the trajectory for box office collections based on dislikes. This might signify that likes are more important in determining the box office collection of a movie as compared to the dislikes.

Looking at graph 1.5, we can see that the total number of YouTube views has an influence in the trajectory of box office sales. To further test this out, an ANOVA test was carried out.



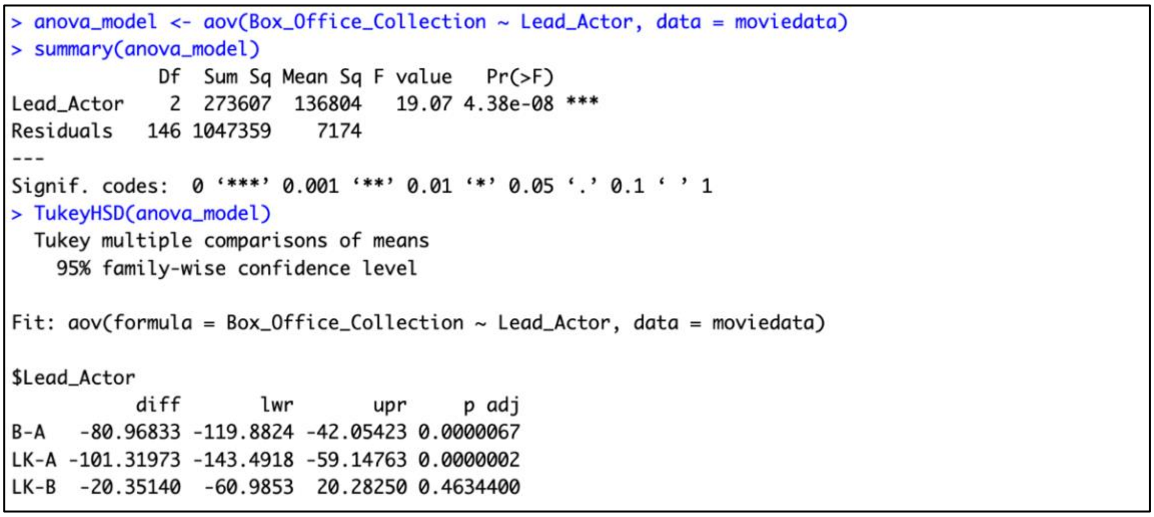
```
> anova_result_2 <- aov(Box_Office_Collection ~ Youtube_Views, data = moviedata)
> summary(anova_result_2)
               Df Sum Sq Mean Sq F value   Pr(>F)    
Youtube_Views   1 457488  457488   77.88 2.96e-15 ***
Residuals     147 863478    5874                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Fig 1.4*

As seen in fig 1.4, the ANOVA test provided a p-value of 2.96e-15, which indicates that YouTube views do in fact impact box office collections. This can be attributed to the popularity of the movie, the interest in the movie as well as the positive or negative sentiment of the general audience towards a movie, which is reflected through likes and dislikes (Graph 1.6).

- What is the estimated difference in box office collection for different lead actor categories?

To find out the estimated differences in the box office collections for different lead actor categories, ANOVA following TukeyHSD was carried out.



```
> anova_model <- aov(Box_Office_Collection ~ Lead_Actor, data = moviedata)
> summary(anova_model)
             Df  Sum Sq Mean Sq F value   Pr(>F)    
Lead_Actor    2  273607  136804   19.07 4.38e-08 ***
Residuals   146 1047359    7174                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(anova_model)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Box_Office_Collection ~ Lead_Actor, data = moviedata)

$Lead_Actor
           diff       lwr       upr     p adj
B-A   -80.96833 -119.8824 -42.05423 0.0000067
LK-A -101.31973 -143.4918 -59.14763 0.0000002
LK-B  -20.35140  -60.9853  20.28250 0.4634400
```

*Fig 1.5*

As can be seen in fig 1.5, the estimated difference between the box office collection of Lead Actor A and Lead Actor B is around $80 million. The p-value is 0.0000067, which indicates that there is a significant difference between the box office collection of Lead Actor A and Lead Actor B.

The estimated difference between the box office collection of Lead Actor A and Lead Actor LK is approximately $101 million. The p-value for the same is 0.0000002, which indicates that there is a significant difference between the box office collection of Lead Actor A and Lead Actor LK.

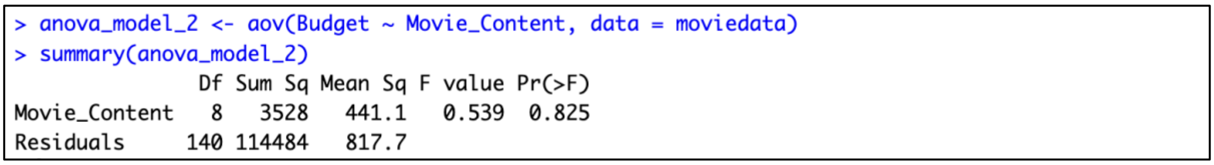
The estimated difference between the box office collection of Lead Actor B and Lead Actor LK is $20 million. The p-value however is 0.46, which indicates that there is not a significant difference between the box office collection of Lead Actor B and Lead Actor LK.

- Is there a significant difference in the budget of different movie types by content?

To answer the above question, ANOVA was carried out on the budget and the movie types by content. Following was the hypothesis used:

H0: There is no difference in the budget of different movie types by content.

Ha: There is a significant difference in the budget of different movie types by content.

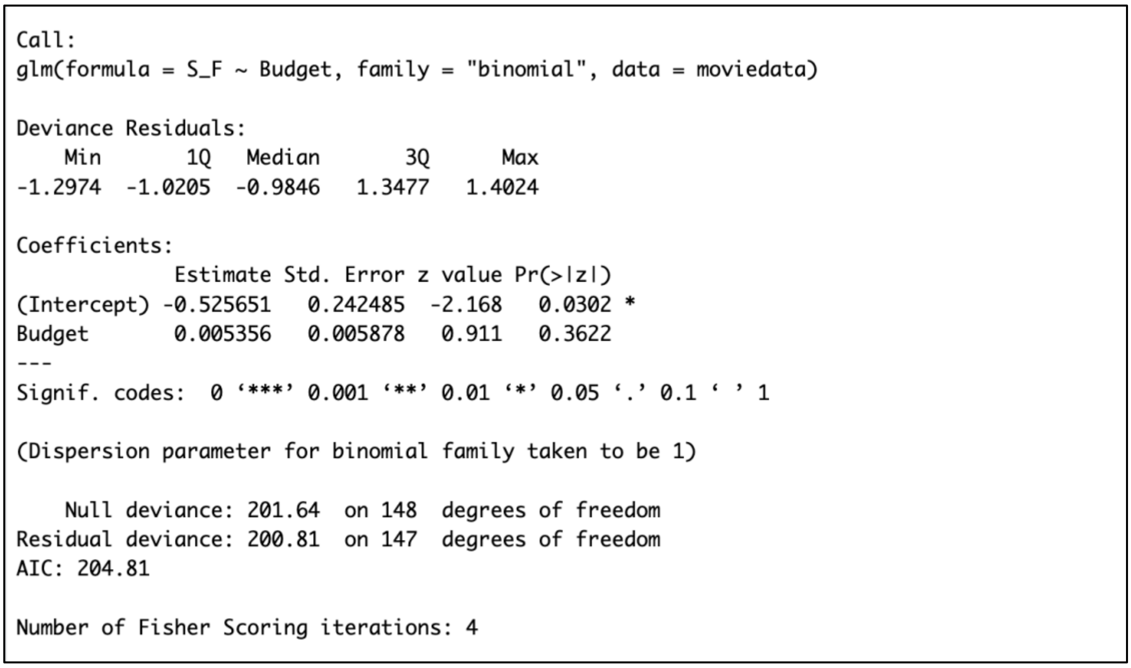


```
> anova_model_2 <- aov(Budget ~ Movie_Content, data = moviedata)
> summary(anova_model_2)
               Df Sum Sq Mean Sq F value Pr(>F)
Movie_Content   8   3528   441.1   0.539  0.825
Residuals     140 114484   817.7
```

*Fig 1.6*

As seen in fig. 1.6, the p-value for movie content is 0.825. The p-value is greater than the significance level of 0.05, and hence we **fail to reject the null hypothesis**, which states that there is no difference in the budget of different movie types by content.

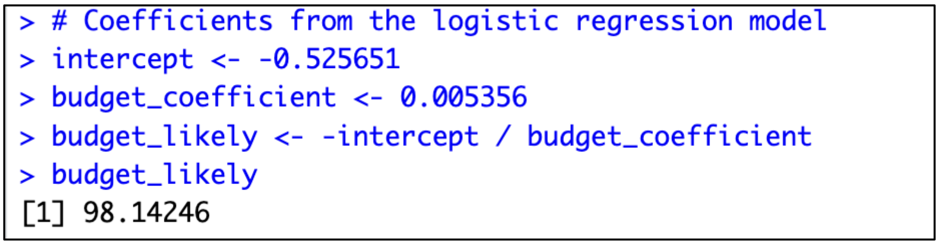
2. Create a logistic regression model using budget as an independent variable and success as a dependent.



```
Call:
glm(formula = S_F ~ Budget, family = "binomial", data = moviedata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2974  -1.0205  -0.9846   1.3477   1.4024

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.525651   0.242485  -2.168   0.0302 *
Budget       0.005356   0.005878   0.911   0.3622
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 201.64  on 148  degrees of freedom
Residual deviance: 200.81  on 147  degrees of freedom
AIC: 204.81

Number of Fisher Scoring iterations: 4
```

*Fig 2.1*

- Calculate the budget for which box office success and failure are equally likely



```
> # Coefficients from the logistic regression model
> intercept <- -0.525651
> budget_coefficient <- 0.005356
> budget_likely <- -intercept / budget_coefficient
> budget_likely
[1] 98.14246
```

*Fig 2.2*

The budget for which box office success and failure is equally likely is: **98.142 crore.**

- Is there sufficient evidence to conclude that higher-budget movies are more likely to fail at the box office?

As seen in fig. 2.1, the coefficient estimate for Budget is 0.005356 and the p-value for the Budget is 0.3622.

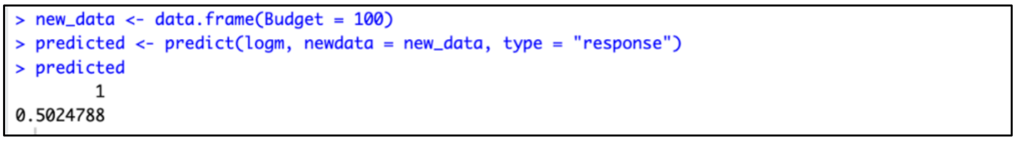
Following is the hypothesis used:

*H0*: There is no effect of the budget on the likelihood of failure of the movie.

*Ha*: There is an effect of the budget on the likelihood of failure of the movie.

Since the p-value is greater than the significance level of 0.05, we fail to reject the null hypothesis, which means that the effect of budget on the likelihood of failure is not statistically significant in this model.

- A production house is making a movie with a 100-crore budget. What is the probability of success for this movie?

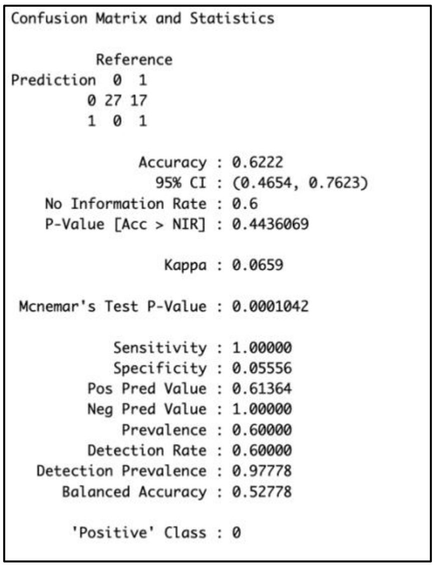


```
> new_data <- data.frame(Budget = 100)
> predicted <- predict(logm, newdata = new_data, type = "response")
> predicted
        1
0.5024788
```

*Fig 2.3*

The probability of success for the movie with a budget of 100 crore is: **50.24%**

- What is the sensitivity and specificity of the classification model used? Interpret your findings.



```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 27 17
         1  0  1

               Accuracy : 0.6222
                 95% CI : (0.4654, 0.7623)
    No Information Rate : 0.6
    P-Value [Acc > NIR] : 0.4436069

                  Kappa : 0.0659

 Mcnemar's Test P-Value : 0.0001042

            Sensitivity : 1.00000
            Specificity : 0.05556
         Pos Pred Value : 0.61364
         Neg Pred Value : 1.00000
             Prevalence : 0.60000
         Detection Rate : 0.60000
   Detection Prevalence : 0.97778
      Balanced Accuracy : 0.52778

       'Positive' Class : 0
```

The sensitivity for the logistic regression model is 1.0.

Sensitivity measures the proportion of actual positives that are correctly identified by the model. In this case, the sensitivity is 1, which means that 100% of the true positives in the model are correctly identified.

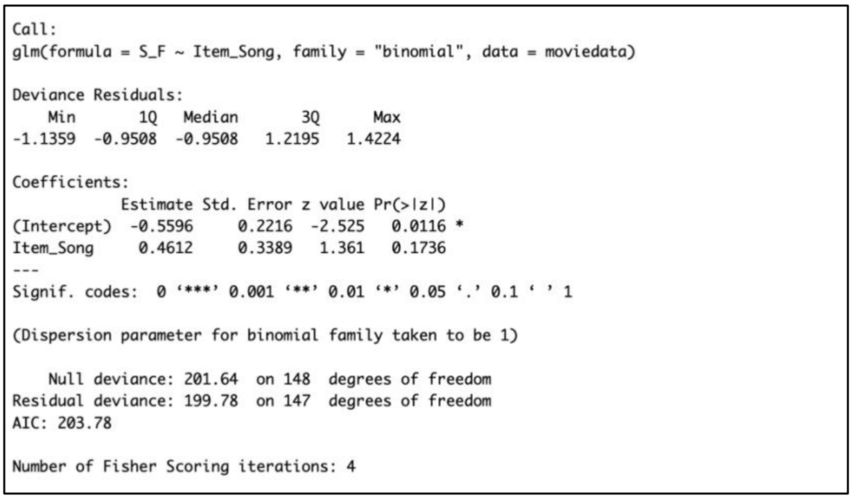
The specificity for the logistic regression model is 0.05

Specificity measures the proportion of actual negatives correctly identified by the model. In this case, the specificity is only 5.56%, which means that only 5.56% of the true negatives are correctly identified by the model.

Here, the model is able to successfully identify movies that will be a success based on the budget but is not able to properly identify movies that will not be a success based on the budget. There are a few reasons for high sensitivity and low specificity, which are as follows:

1. One of the biggest reasons is class imbalance. In this case, the number of failures in the testing dataset are a lot higher than the number of successes in the training dataset. Due to the size of the dataset being quite small, splitting it into training and testing has proved to be a difficult task to validate the models.
2. The model is not able to correctly categories the successes and failures because the budget cannot properly encapsulate the variation that is explained in the dependent variable.

3. Create a logistic regression model using Item song as an independent variable and success as a dependent

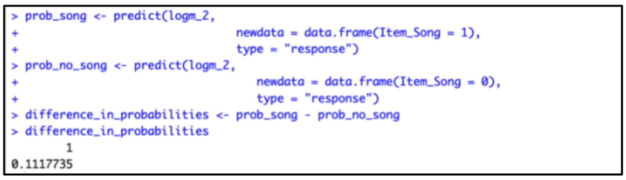


```
Call:
glm(formula = S_F ~ Item_Song, family = "binomial", data = moviedata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1359  -0.9508  -0.9508   1.2195   1.4224

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.5596     0.2216  -2.525   0.0116 *
Item_Song     0.4612     0.3389   1.361   0.1736
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 201.64  on 148  degrees of freedom
Residual deviance: 199.78  on 147  degrees of freedom
AIC: 203.78

Number of Fisher Scoring iterations: 4
```

*Fig 3.1*

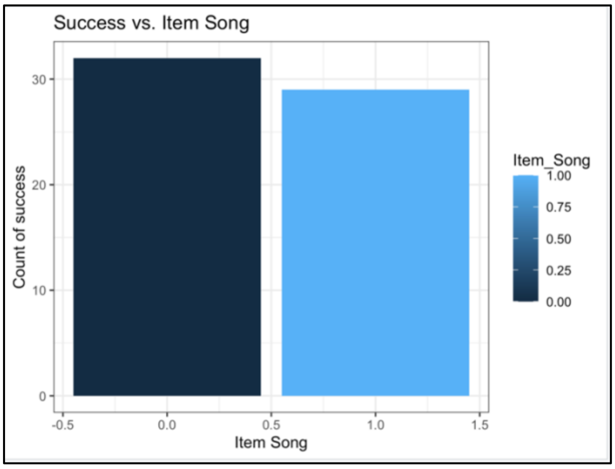
- What is the difference between success probabilities for movies with item songs and movies without item songs? Corroborate your answer with findings of descriptive analytics.

*Fig 3.2*



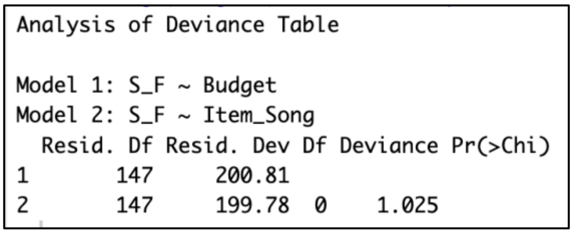
```
> prob_song <- predict(logm_2,
+                                 newdata = data.frame(Item_Song = 1),
+                                 type = "response")
> prob_no_song <- predict(logm_2,
+                                    newdata = data.frame(Item_Song = 0),
+                                    type = "response")
> difference_in_probabilities <- prob_song - prob_no_song
> difference_in_probabilities
        1
0.1117735
```

*Fig 3.3*

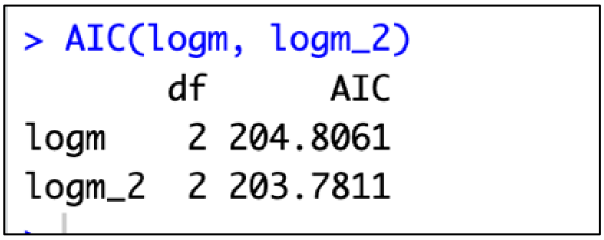


As seen in Fig 3.2, after calculation, the difference in success probabilities for movies with and without an item song is 0.11 or 11.17%. The same can be seen in fig 3.3, where there is not a big difference between the success rate of movies with an item song and movies without an item song. This can be coupled with the logistic regression model in fig 3.1, where the p-value for the item song is 0.17. This shows that there is no significant difference between the success probabilities of movies with item songs and movies without item songs.

- Which is a better model- Budget as an independent variable versus item song as the independent variable? State your reasons and explain the possible causes of the same.



```
Analysis of Deviance Table

Model 1: S_F ~ Budget
Model 2: S_F ~ Item_Song
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       147     200.81
2       147     199.78  0     1.025
```



```
> AIC(logm, logm_2)
       df      AIC
logm    2 204.8061
logm_2  2 203.7811
```

*Fig 3.4*

*Fig 3.5*

For fig 3.4 I tried running an Analysis of Deviance table in order to use hypothesis testing. However, the model could not give me a p-value for the either logistic regression models, therefore not allowing me to see which is the better model.

Next, I compared the AIC of both the models in fig. 3.5. Through the AIC, it can be seen that the model with the item song as the independent variable is marginally better than the model with budget as an independent variable.

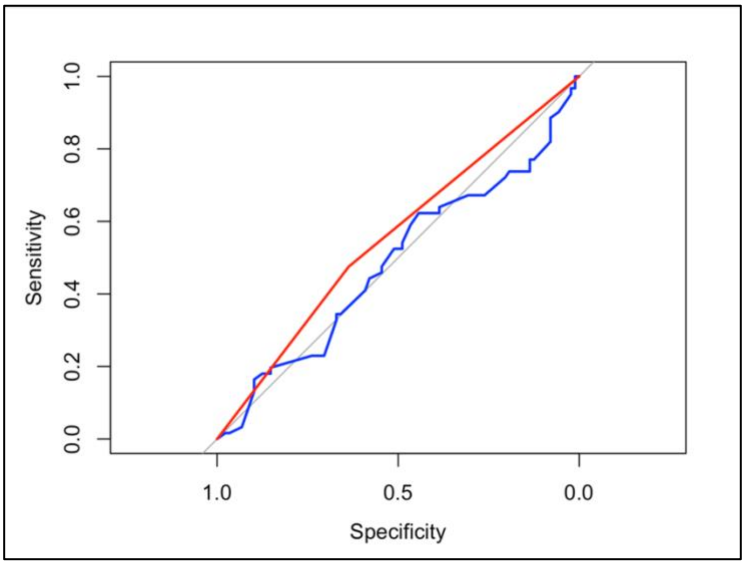
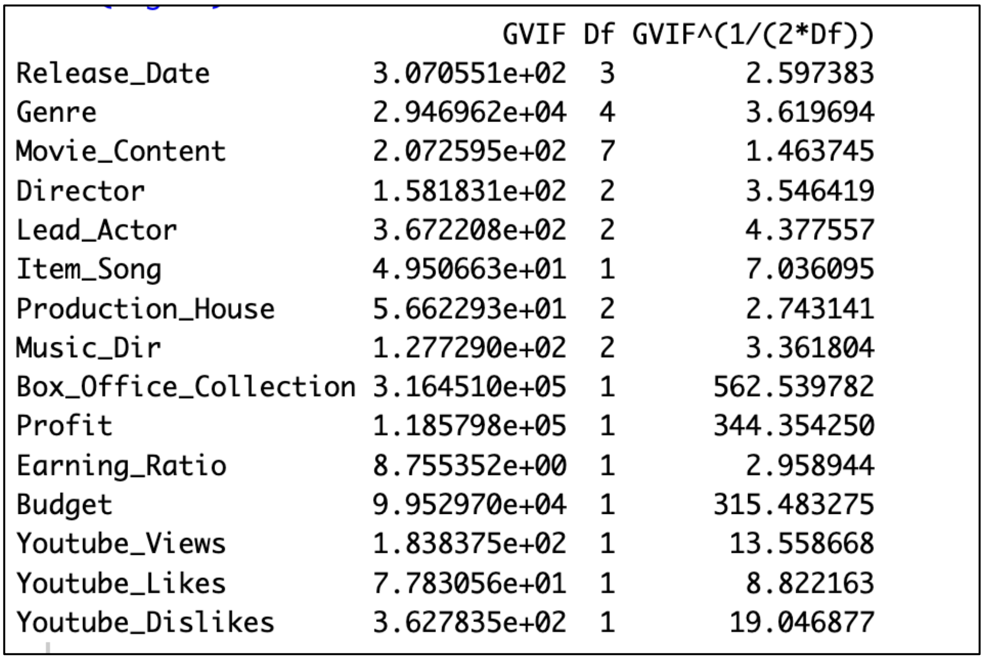


*Fig 3.6*

One more test I ran to compare both the models was the ROC graph. The graph shows the performance of a model over various thresholds, and is particularly used to measure the performance for classification models. Using the ROC graph, we can see that the logistic regression model with item song (red line) covers a larger area under the curve and is hence a better model than the logistic regression model with budget (blue line). However, it can be seen through the ROC graph that the item song model has a tendency to favour predicting the failure of movies, which is not ideal.

Ultimately, the logistic regression model with the item song narrowly edges out the logistic regression model with the budget. However, the difference is not significant enough for me to explicitly choose one model over the other.

4. Develop a model to predict the success of the movie using all the variables provided. Explain the factors affecting the success/failure of a movie. What are the rules that can be used to predict the success or failure



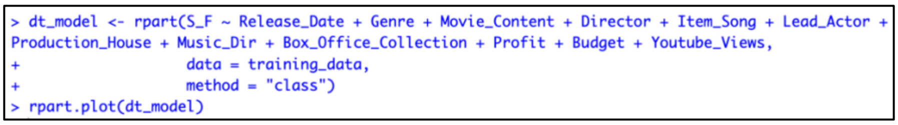
```
                             GVIF Df GVIF^(1/(2*Df))
Release_Date         3.070551e+02  3        2.597383
Genre                2.946962e+04  4        3.619694
Movie_Content        2.072595e+02  7        1.463745
Director             1.581831e+02  2        3.546419
Lead_Actor           3.672208e+02  2        4.377557
Item_Song            4.950663e+01  1        7.036095
Production_House     5.662293e+01  2        2.743141
Music_Dir            1.277290e+02  2        3.361804
Box_Office_Collection 3.164510e+05  1      562.539782
Profit               1.185798e+05  1      344.354250
Earning_Ratio        8.755352e+00  1        2.958944
Budget               9.952970e+04  1      315.483275
Youtube_Views        1.838375e+02  1       13.558668
Youtube_Likes        7.783056e+01  1        8.822163
Youtube_Dislikes     3.627835e+02  1       19.046877
```

*Fig 4.1*

To create the best model to predict the success rate of a movie, it was first important to understand the variables that were collinear in the dataset and remove those that would skew the accuracy of the model.

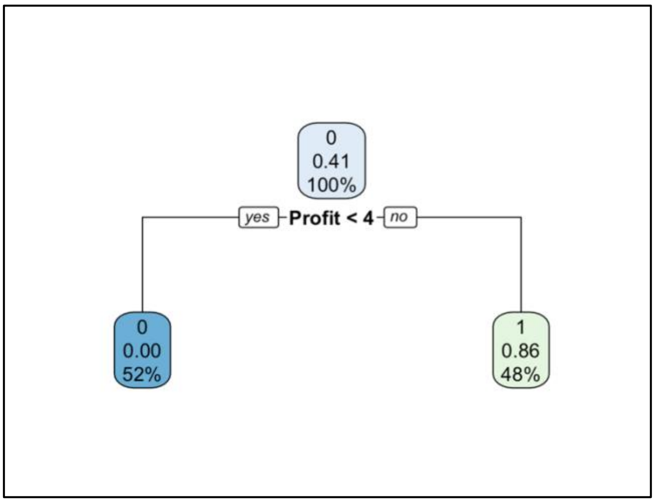
After running the VIF, it was pretty clear that the Box Office Collection, the Profit, and the Budget were closely related to each other. YouTube Views, YouTube likes, and YouTube Dislikes were also related to each other quite heavily.

I decided to use a decision tree to create a model to accurately predict the success of the movie. Taking into consideration the VIF scores, I decided to remove the YouTube likes and YouTube Dislikes from the decision tree model. However, I decided to keep Box Office Collection, the Profit, and the Budget all in the model because I thought they would all play an important role in the accuracy of the model.



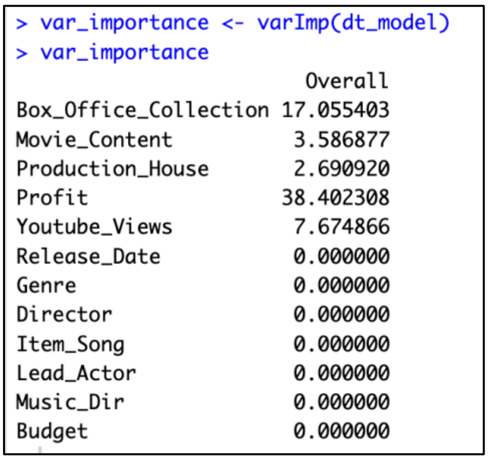
```
> dt_model <- rpart(S_F ~ Release_Date + Genre + Movie_Content + Director + Item_Song + Lead_Actor +
Production_House + Music_Dir + Box_Office_Collection + Profit + Budget + Youtube_Views,
+                     data = training_data,
+                     method = "class")
> rpart.plot(dt_model)
```

*Fig 4.2*



I ran the decision tree with all the variables in fig 4.2, and it gave me the following decision tree, fig 4.3. The model shows that profit is the most important factor in predicting the success of a movie, and a profit of below 4 crores indicated the failure of a movie, while a profit of above 4 crores indicated the success of a movie.

*Fig 4.3*

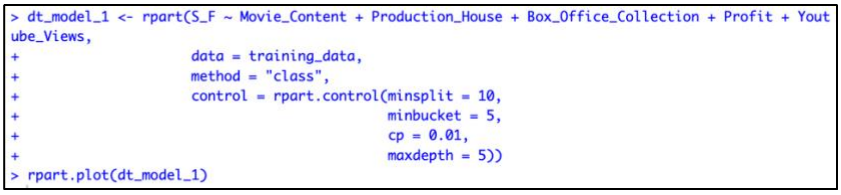


```
> var_importance <- varImp(dt_model)
> var_importance
                       Overall
Box_Office_Collection 17.055403
Movie_Content          3.586877
Production_House       2.690920
Profit                38.402308
Youtube_Views          7.674866
Release_Date           0.000000
Genre                  0.000000
Director               0.000000
Item_Song              0.000000
Lead_Actor             0.000000
Music_Dir              0.000000
Budget                 0.000000
```

After running the first decision tree, I decided to run a variable importance model to see which variables were actually significant in the making of the decision tree. I found out that the Box Office Collection, Movie Content, Production House, Profit and YouTube Views were the most important factors in the decision tree model

*Fig 4.4*

Finally, I ran one last model of the decision tree with only the important variables as well as did some hyper parameter tuning to find out whether there were any other variables that might impact the model.



```
> dt_model_1 <- rpart(S_F ~ Movie_Content + Production_House + Box_Office_Collection + Profit + Yout
ube_Views,
+                       data = training_data,
+                       method = "class",
+                       control = rpart.control(minsplit = 10,
+                                               minbucket = 5,
+                                               cp = 0.01,
+                                               maxdepth = 5))
> rpart.plot(dt_model_1)
```

*Fig 4.5*

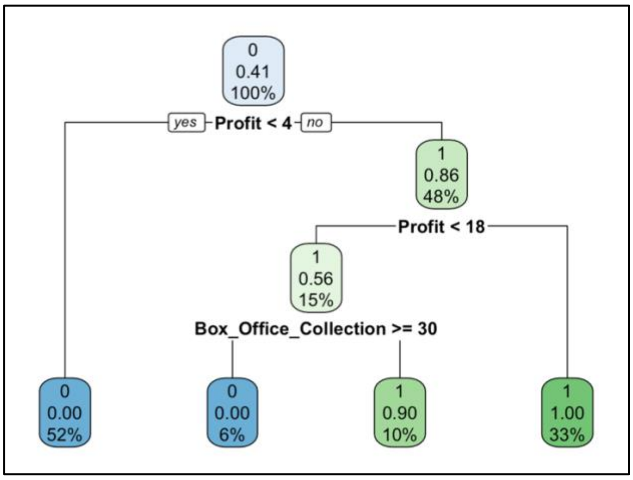
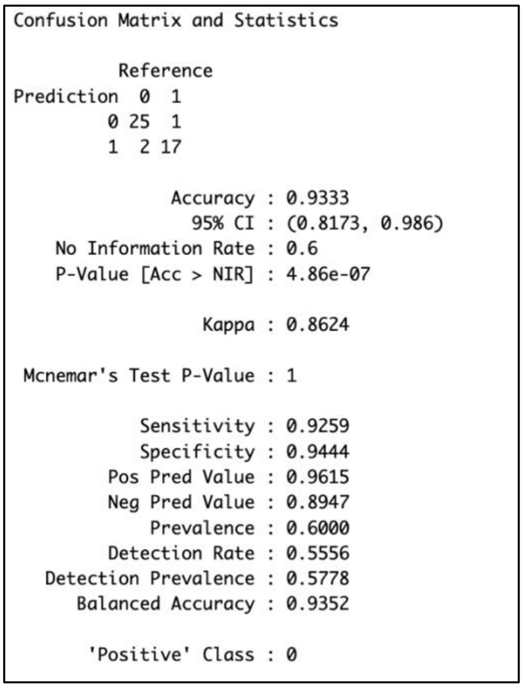


*Fig 4.6*

Fig 4.6 is the decision tree that I got for running the code in fig. 4.5. After some hyper parameter tuning, the profit and the box office collection were both deemed to be deciding factors in the decision tree model.



```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 25  1
         1  2 17

               Accuracy : 0.9333
                 95% CI : (0.8173, 0.986)
    No Information Rate : 0.6
    P-Value [Acc > NIR] : 4.86e-07

                  Kappa : 0.8624

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9259
            Specificity : 0.9444
         Pos Pred Value : 0.9615
         Neg Pred Value : 0.8947
             Prevalence : 0.6000
         Detection Rate : 0.5556
   Detection Prevalence : 0.5778
      Balanced Accuracy : 0.9352

       'Positive' Class : 0
```

*Fig 4.7*

Fig 4.7 was what I received as the confusion matrix for the final decision tree model. The accuracy of the model is 93.33%, indicating that the model is a good fit for explaining the success or failure of a movie.

The Sensitivity for the model is 92%, while the specificity for the model is 94%, indicating that the model is able to accurately predict both the success as well as the failure of a movie based on the give factors. Ultimately, the model has proven to be a success with fewer variables and a higher level of hyper parameter tuning.