ANALYZING THE KEY METRICS RELATED TO DISNEY MOVIES

(FINAL REPORT)

By- Debolina Sasmal, Amy Yi, Abdirahman Hussein

(Team 9)

BUSINESS PROBLEM:

The Walt Disney Studios is an American film and entertainment studio, and one of the four business segments of The Walt Disney Company. Disney studios comes under the entertainment industry and is responsible for producing movies (known to us as the Disney movies). It is the third largest segment with a revenue of $2.7 Billion, but over the years, it has experienced a decline in revenue, largely due to heavy competition within the industry. Disney competes with many different media conglomerates across its various business lines. The company's largest competitors are Sony Pictures, Time Warner, Netflix, Prime video, which are pretty well known in the entertainment industry. To keep up in such a competitive environment, Disney should leverage the power of data analytics to differentiate their service and gain a competitive advantage.

For this project, we are using a dataset that includes a comprehensive list of Disney movies and various data pertaining to those movies, such as IMDB ratings, country of release, language used, budget, box office earnings etc. We know that some of the factors mentioned like language used, genre and budget affect the box office earnings and the IMDB ratings. So, we will be analyzing the data to address the basic business problem in the entertainment industry: “How to amplify the box office earnings and score a high IMDB rating?”. This is something almost every person in the entertainment industry will care about, and we will try to analyze the data from Disney’s perspective.

To get started on the project, the CRISP-DM process can be stated as follows:

Business Understanding:

- Identify which factors affect the box office earnings and IMDB ratings.

Data understanding:

- Gathering data pertaining to Disney movies to be able to see a clear trend of how the movies have been perceived by the consumers.

Data preparation:

- Identify the relevant data elements to be used in the prediction models.
- Clean the data and integrate multiple datasets if more relevant data is required for forecasting and decision making.
- The data has been collected from Kaggle and data.world.

Modeling:

- The “regression” method of supervised machine learning can help us predict the correct areas to concentrate on to amplify the earnings and ratings.

- Visualizations can help in assessing the data at a glance.
- How relevant are the variables used in the regression?
- Quality assessment of produced model.

Evaluation:

- Check whether the number of consumers has increased.
- The IMDB ratings or the earnings should increase.
- Check user written reviews.
- Check whether the movies are trending on the social media platforms such as Twitter or Instagram.

Deployment:

- New movies can be made and released based on the target audience. For example- movies in regional language to enter the market of a new country.

DATA WRANGLING:

We decided to use Kaggle as our main data source as it is a large repository of data that was free to use and the Disney movie set was quite robust and in an easy-to-use format for R.

However, the Kaggle dataset for Disney had a few key things missing that are important for our business proposal.

- The Inflated value for both budget and box office earnings were missing.
- Dates were not formatted properly.
- Genre of movies was missing.

Genre is a key component in movie making as it plays into how Disney would mark the movie and types of films to make in the future. There was another Disney movie set on data.world.com that had a list of Disney movies with genre included.

For our project, we merged the two datasets together so we would have one comprehensive dataset to run regression and make visualizations easier by exporting it as a single csv file.

Datasets used:

https://www.kaggle.com/therealsampat/disney-movies-dataset

https://data.world/kgarrett/disney-character-success-00-16

Code used to merge:

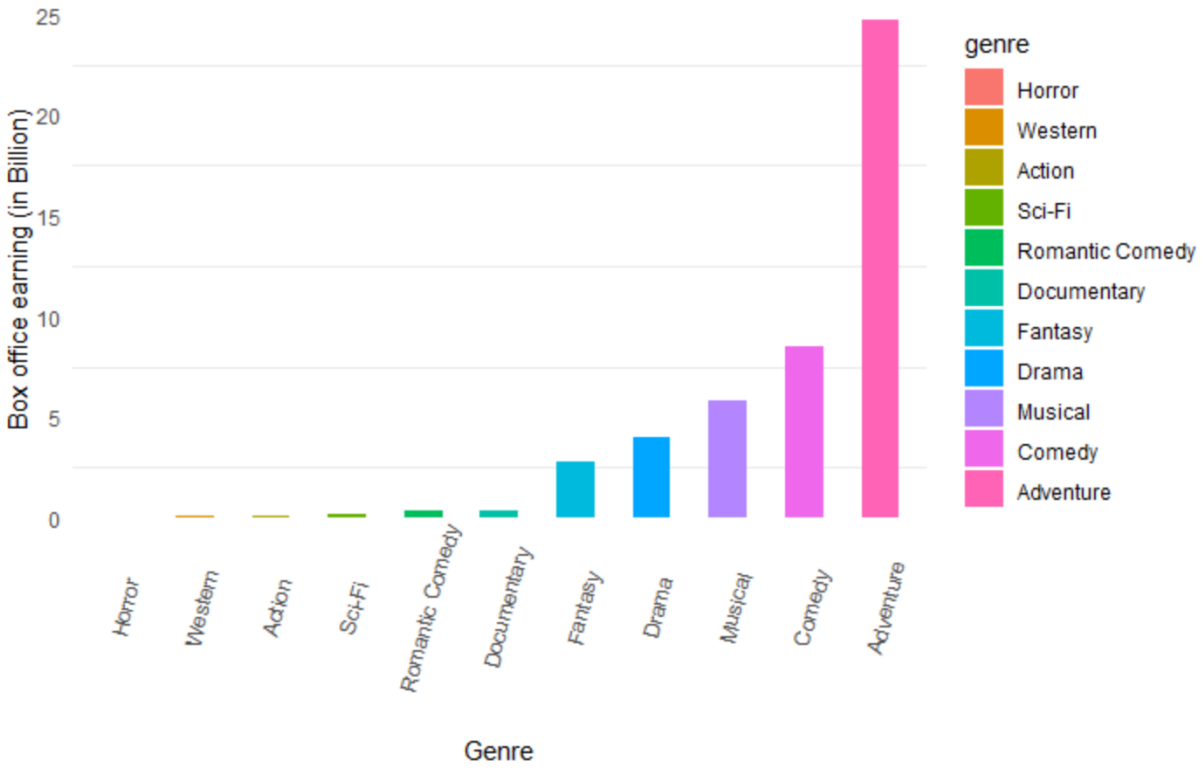
```
df <- read.csv("DisneyMoviesDataset.csv", na.strings ="?")

df1 <- read.csv("disney_movies_total_gross.csv", na.strings ="?")

colnames(df)

colnames(df1)

names(df1)[names(df1) == "ï..movie_title"] <- "title"

merged_df <- merge(df,df1, by.x= "title")
```

This code allowed minimal duplication between the two data sets and allowed us to run regression on a bigger set of data. While there were some movies with NA genre column as the Kaggle Disney data set and the data.world Disney dataset did not have quite the same list of movies, we did have to do some manual data input in order to make our data set as comprehensive as possible.

EXPLORATORY DATA ANALYSIS:

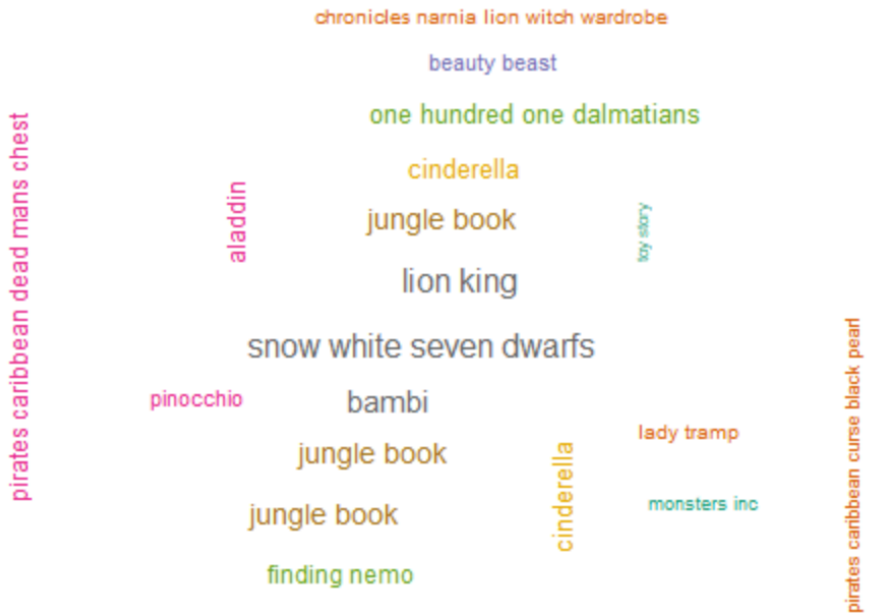
We have generated several visualizations to be able to analyze the data at a glance for better and easier understanding. This helped us decide which variables we should consider for further analysis.

1. **Which movie genre is earning the highest box office revenue?**



This visualization shows the Box Office earnings of the Disney movies for each genre. From this, we can learn that the genre “Adventure” tends to earn the highest profits, surprisingly more that double what the next in line, that is “Comedy” is making. Adventure, Comedy, Musical, Drama, and Fantasy and the most liked genres.

2. **Which movie earned the highest box office earnings?**



The above word cloud shows few of the Disney movies that earned the highest Box Office earnings. As we can see, “Snow white seven dwarfs” is written with the biggest font size, so that movie earns the highest profit among the ones listed. It is interesting to see that several movies such as Pinocchio, Lion King, Cinderella, Jungle Book - these all few of the oldest movies released by Disney, but still have such high profits when compared to the recent ones. This tells us that the competition in the industry has got tough with new entrants like Netflix and Prime Movies. Additionally, those movie have an emotional effect for the millennials, which is why those movies might have reached more consumers.

3. **Top 10 movies in terms of IMDB vs. Metascore vs. Rotten Tomatoes**

```
# A tibble: 10 x 2
   title          imdb
   <chr>          <chr>
 1 Hamilton       8.69999
 2 The Lion King  8.5
 3 WALL-E         8.4
 4 Dangal         8.4
 5 Coco           8.4
 6 Toy Story      8.30000
 7 The Kid        8.30000
 8 Toy Story 3    8.30000
 9 Up             8.19999
10 Finding Nemo   8.1
```

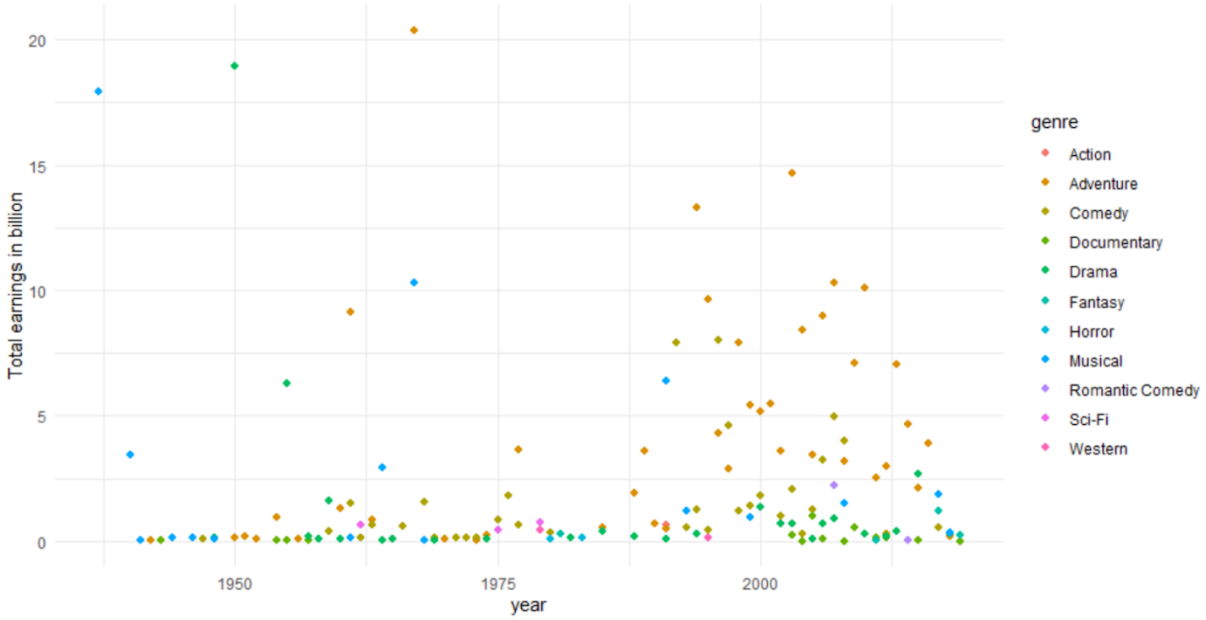
```
# A tibble: 10 x 2
   title                           metascore
   <chr>                           <chr>
 1 Pinocchio                       99
 2 Fantasia                        96
 3 Dumbo                           96
 4 Ratatouille                     96
 5 Snow White and the Seven Dwarfs 95
 6 Beauty and the Beast            95
 7 Toy Story                       95
 8 WALL-E                          95
 9 Inside Out                      94
10 Toy Story 3                     92
```

```
# A tibble: 10 x 2
   title                                    rotten_tomatoes
   <chr>                                              <dbl>
 1 Pinocchio                                              1
 2 Davy Crockett, King of the Wild Frontier               1
 3 Old Yeller                                             1
 4 Darby O'Gill and the Little People                     1
 5 Greyfriars Bobby                                       1
 6 Mary Poppins                                           1
 7 The Many Adventures of Winnie the Pooh                 1
 8 Never Cry Wolf                                         1
 9 Toy Story                                              1
10 Endurance                                              1
```

We can get two inferences from the above tables:

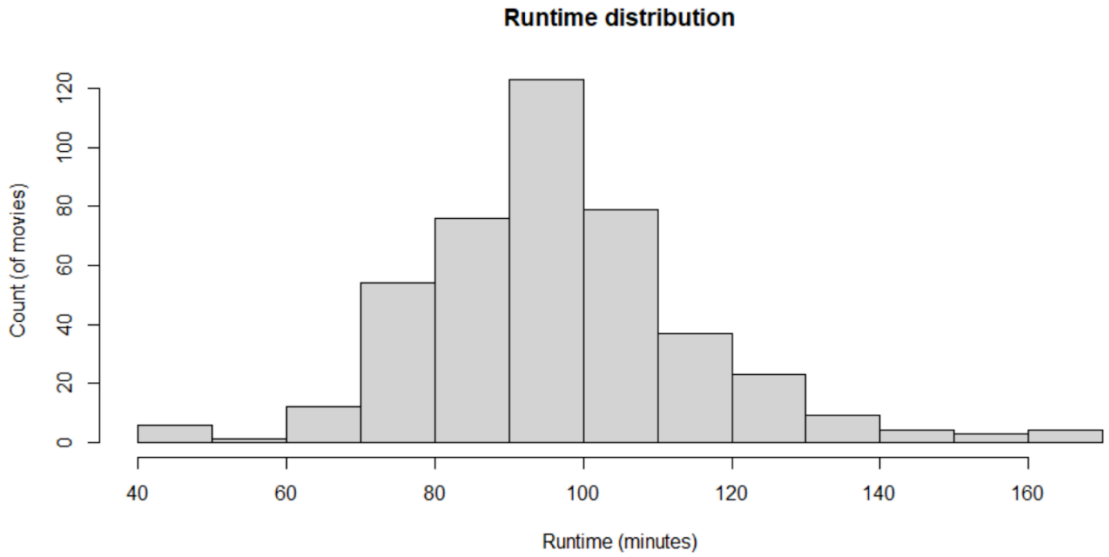
- The only movie that is present in all the three tables is “Toy Story”. That means a high rating in IMDB doesn’t guarantee in a high rating in Rotten Tomatoes or Metascore and vice versa. So, each rating site has their own systematic bias.
- A high IMDB rating doesn’t always mean the movie is earn high Box Office earnings. If we compare these tables to the second visualization titled” Which movie earned the highest box office earnings?”, we can see that out of 10 movies listed on the IMDB table, only three of them are present in the highest earning visualization.

4. What is the trend of the genres of Disney movies over the years?



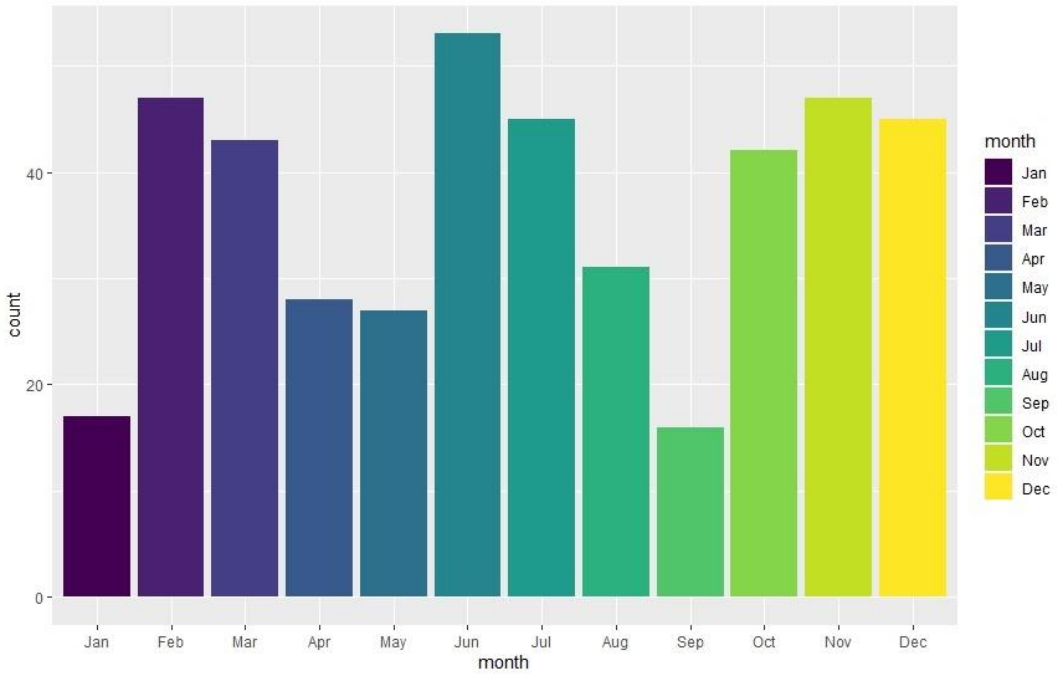
This visualization tells us that before 1975, only a few movies earned high Box office earnings, that too once in ten years or so. Around the 2000s, it is clear that the genre “Adventure” started earning more on a regular basis, when compared to all other genres. Also, more adventure, comedy, drama and musicals were produced in general. The number of movies produced, irrespective of the genre, increased as well.

5. **What is the runtime distribution of the movies?**



According to this visualization, the runtime distribution is of normal type and somewhat skewed to the left. The ideal runtime of the movies is about 90-100 minutes i.e., 1.5-1.7 hours. Back in the earliest days of the film exhibition, the runtime used to be about 30mins (so the few movies with lower runtime are the older ones), but the film industry was threatened by the advent of television as the lower runtime made movies seem almost like a TV show. On the other hand, for the filmmakers and studios who have earned the confidence of the audience, it is fine to make movies of longer runtime. For example, Avengers Endgame was nearly 3hours long, even so it was a blockbuster. But if a first-time producer releases such a long movie, people are more likely to skip it.

6. **Movies are usually released in which months?**



From the figure above, we can say that the movies are usually released in the months of February, March, June, October, November, and December. This is probably because these months are around the holiday seasons, and so people usually have more leisure time to spend their time watching a movie.

REGRESSION MODELLING FOR BOX OFFICE EARNINGS AND IMDB RATING PREDICTIONS

REGRESSION MODEL – 1

In this model, we have used Box Office earnings as the dependent variable. After several trial and errors, the best model we could get was by using the independent variables:

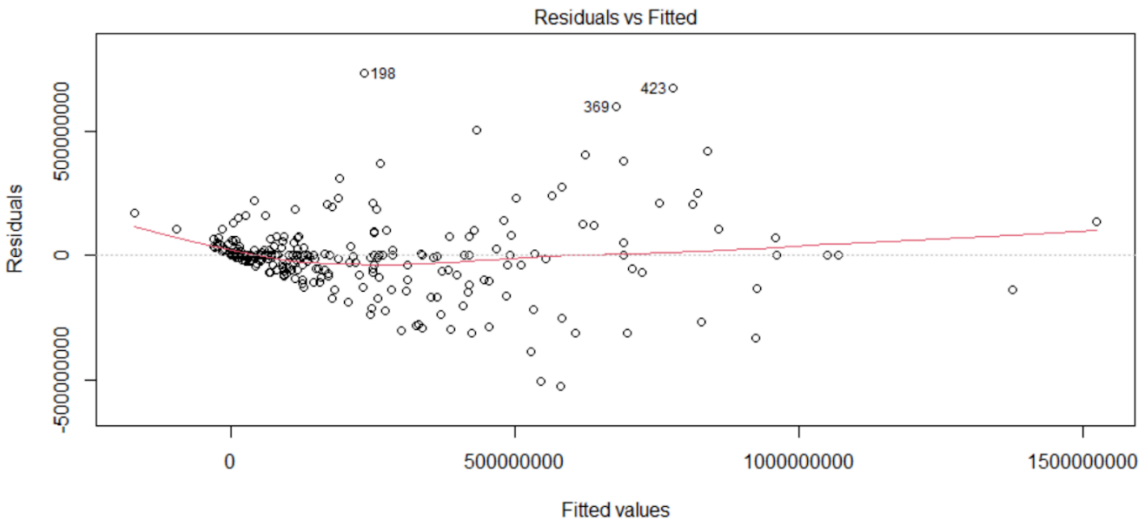
Budget, Runtime, Country of origin, Genre and the Language used.

Below are the quantitative and visual output of the model:

```
Residual standard error: 202300000 on 148 degrees of freedom
  (179 observations deleted due to missingness)
Multiple R-squared:  0.7534,    Adjusted R-squared:  0.565
F-statistic:     4 on 113 and 148 DF,  p-value: 0.000000000000003656
```

For this model, the overall p-value is 3.656×10^{-15}, so it is significant for almost any value of alpha. Also, the multiple R-squared value is 0.7534, so this model explains 75.34% of the variance in the data. Even though we have pretty good p-value and multiple R-squared value for this model, we haven't really found any significant coefficients except the one shown below. We tried to introduce dummy variables for the character type columns, but the output wasn't any good either.

```
Running.time118 minutes
                        **
```



As for the plot above of Residuals vs. Fitted values, it tells us the values are heteroscedastic in nature, that is, the vertical range of the residuals increases as the fitted value increases.

This tells us that even though we have good p-value and R-squared value, the model isn't ideal for prediction and could be made better.

REGRESSION MODEL – 2

In this model, we have used IMDB Rating as the dependent variable. After several trial and errors, the best model we could get was by using the independent variables:

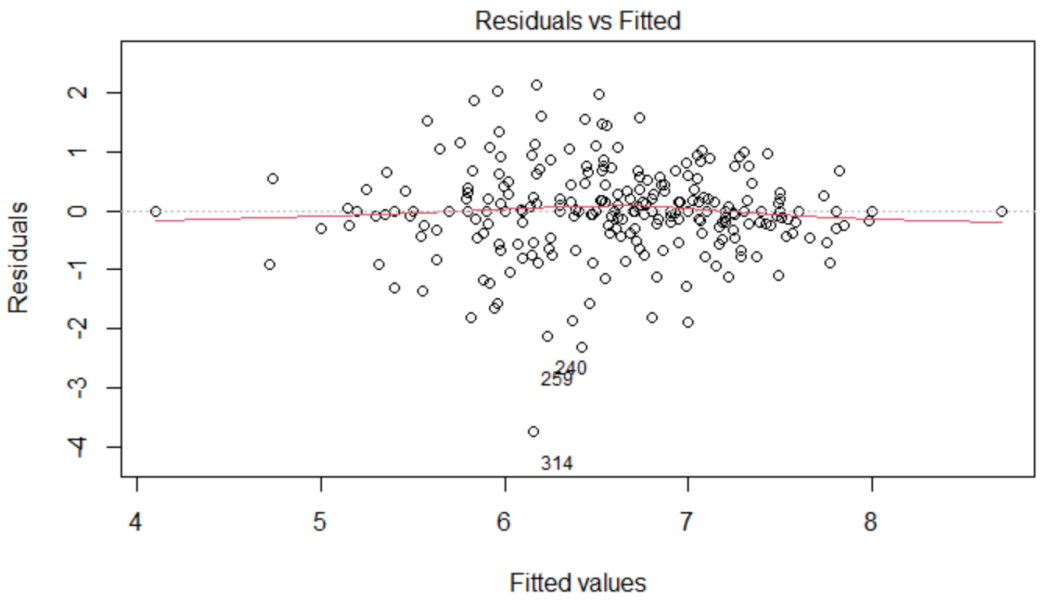
Budget, Runtime, Country of origin, Genre and the Language used.

Below are the quantitative and visual output of the model:

```
Residual standard error: 0.9491 on 159 degrees of freedom
  (169 observations deleted due to missingness)
Multiple R-squared:  0.4765,    Adjusted R-squared:  0.1077
F-statistic: 1.292 on 112 and 159 DF,  p-value: 0.06875
```

For this model, the overall p-value is 0.06875, so it is significant for alpha value 0.1, but not for 0.05 and lower. Also, the multiple R-squared value is 0.4765, so this model explains 47.65% of the variance in the data. One this to note here is that even the Multiple R-squared value is somewhat good, the Adjusted R-squared value is very low, which is a potential discrepancy in the model. As shown below, even for this model we couldn't find many significant variables, making the model not-so-good for predictions.

```
as.numeric(Budget..float.)                                    *
Running.time..int.109                                         *
```



For the plot above of Residuals vs. Fitted values, even though the points are pretty much random, which is a good sign of the model being a good fit to some extent.

PREDICTION USING MODEL-2

For prediction purpose, we have used the second regression model and predicted the IMDB rating by passing random values of the independent variables we have picked from the dataset for the movie “Pinocchio” as follows:

title	Running.time	Country	Language	genre	Budget..float.	imdb
Pinocchio	88	United States	English	Adventure	2600000	7.4

```
newdata <- data.frame(Budget..float. = 2600000, Running.time..int. = "88",
Country = "United States", genre = "Adventure", Language = "English")

predict(model2, newdata, interval = "prediction", level = 0.95)
```

We got the following as output:

```
      fit      lwr      upr
6.873467 4.767679 8.979255
```

The predicted IMDB rating is 6.9, whereas the actual rating according to the dataset is 7.4

So, %Error = (| 6.9 – 7.4 | / 7.4) x 100 = 6.75676%, which is quite low. But the lower and upper limit of the prediction is quite big considering it is on a scale of 10. So, the confidence interval is pretty large, so the predictions might not be accurate, which sets a low confidence.

MANAGERIAL CONCLUSIONS:

There are several recommendations that our team could give to Disney based on our analysis.

Recommendation - 1:

Our first recommendation for Disney is to make an adventure film that is roughly 90-100 minutes long, released in the times of June/July or February/March when people have the most leisure time to see movies. The biggest strength of this strategy is a well-known formula that has worked to generate billions of dollars over the years.

However, this would bring Disney into most competition with outer competitors like Fox, CBS Corp, Warner Bro, and etc. Since the popular market is also the most saturated the movie would need a twist or a hook in the film that would make people pick the Disney film over a competing film.

Recommendation - 2:

Our second recommendation is Disney to go into a niche market. Disney could make a movie in the musical/romance genres that is roughly 90-100 minutes long released in November/December or June/July to get the most exposure to people. The biggest strength of this strategy would be that it would instantly stand out among the usual Disney offerings and draw new audiences to Disney, which can generate more revenue.

However, the biggest risk is that historical data suggests these types of movies have not done well in the past. To make this type of movie a success, Disney may have to spend more in marketing to generate more buzz and get the name out. This is a less trusted money-making strategy.

Recommendation - 3:

Our third recommendation is to break into foreign markets like India as they have a big population with a strong history of movie watching. Disney should release an adventure film with a run time typically for the area released during June/July as there is a lot of festivals and people have more leisure time to go to films. The biggest advantage of this strategy is that it's an untapped market for Western media with potential for Disney to gain a critical foothold in the market. The world is becoming more and more interconnected in terms of business and cultural influence. Foreign markets are an untapped potential because most movies are USA centric.

Disney is already a big household name internationally because of their incredibly strong branding. There would be a big draw for Disney, a big American movie company, making a movie in India. The disadvantage is that there is already a robust Bollywood market which makes competition fierce and costs more resources to make a film in foreign locations as Disney would have to find contacts to find familiar locations and relevant permits. Disney can navigate this issue by partnering with a big Bollywood studio.

The recommendation best suited to what movie to make would largely depend external factors like competitors' strategy and past releases that have done well on the market. Movie making has so many factors and it is hard to quantify a film's potential success solely on data as movies are ultimately a visual expression of the storytelling. The recommendations listed above is to position Disney with the best chance of making a successful film and to increase revenue over time.

APPENDIX A: CODES USED FOR ANALYSIS

Code 1: Libraries used and importing data

```
library(ggplot2)

library(dplyr)

install.packages('tm',dependencies = TRUE)

install.packages("wordcloud")

library('tm')

library("SnowballC")

library("wordcloud")

library("RColorBrewer")

library(lubridate)

library("readxl")

setwd("C:/Users/Debolina/Desktop")

Disney <- read_excel("DisneyProject_CleanedUP.xlsx")
```

Code 2: Which genre movie is earning the highest Box office revenue?

```
Disney$income <- Disney$`Box Office_Inflation Adjusted`

subsetData <- subset(Disney, Disney$genre != "")


group_by_genre <- aggregate(income ~ genre, data = subsetData, FUN = sum)


currency_in_billion <- (group_by_genre$income/1000000000)

group_by_genre$currency <- currency_in_billion

head(group_by_genre)


group_by_genre$genre <- factor(group_by_genre$genre, levels = group_by_gen
re$genre[order(group_by_genre$currency)])


ggplot(group_by_genre, aes(genre,currencyinbillion)) + geom_bar(stat = "id
entity", aes(fill = genre), width = 0.5) + theme_minimal() + xlab("Genre")
+ ylab("Box office earning (in Billion)") + theme(panel.grid.major = eleme
nt_blank(), axis.text.x  = element_text(angle=75,size=9, vjust=0.7))
```

Code 3: Which movie earned the highest income?

```
data <- head(arrange(Disney,desc(`Box Office_Inflation Adjusted`)), n = 20
)

data$Rank <- rank(data$`Box Office_Inflation Adjusted`)

word <- Corpus(VectorSource(data$title))



toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ",
x))



word <- tm_map(word, toSpace, "/")

word <- tm_map(word, toSpace, "@")

word <- tm_map(word, toSpace, "\\|")

word <- tm_map(word, content_transformer(tolower))

word <- tm_map(word, removeNumbers)

word <- tm_map(word, removeWords, stopwords("english"))

word <- tm_map(word, removePunctuation)

word <- tm_map(word, stripWhitespace)



dataframe <- data.frame(text=sapply(word, identity), stringsAsFactors=F)



data$movie_title <- dataframe$text



wordcloud(words = data$movie_title ,freq = data$Rank, min.freq=1,scale = c
(1,0.5), max.words=200,random.order=FALSE,rot.per=0.35,colors=brewer.pal(8
,"Dark2") )
```

Code 4: Top performing movies in terms of IMDB

```
a <- subset(Disney, Disney$imdb != "N/A")

d <- unique(head(arrange(a[,c("title", "imdb")],desc(imdb)),11))

d
```

Code 5: Top performing movies in terms of Rotten tomatoes

```
a2 <- subset(Disney, Disney$rotten_tomatoes != "N/A")

e <- unique(head(arrange(a2[,c("title", "rotten_tomatoes")], desc(rotten_t
omatoes)),10))

e
```

Code 6: Top performing movies in terms of Rotten tomatoes

```
a3 <- subset(Disney, Disney$metascore != "N/A")

e1 <- unique(head(arrange(a3[,c("title", "metascore")], desc(metascore)),1
2))

e1
```

Code 7: What is the trend of movies over the years?

```
trend <- Disney[,c(8,12,14)]



trend$grosstotal <- Disney$income



trend$release_date <- parse_date_time(Disney$Release.date..datetime., orde
rs = c("ymd", "dmy", "mdy"))

trend$year <- year(trend$release_date)



trend$release_date <- NULL



trend1 <- subset(trend,trend$genre != "")



group_by_date <- aggregate(grosstotal ~ year + genre, data = trend1,FUN =
sum)



subset_trend <- subset(group_by_date, group_by_date$grosstotal != 0)



options(scipen=1000000)
```

```
curr <- (subset_trend$grosstotal / 100000000)

subset_trend$sample <- curr



ggplot(subset_trend, aes(x=year, y=curr, color=genre)) + geom_point(size =
1.9) + theme_minimal() + ylab("Total earnings in billion")
```

Code 8: Runtime distribution

```
runtime <- Disney[,c(2,9)]

runtime$num <- as.numeric(runtime$Running.time..int.)

hist(runtime$num, main="Runtime distribution", xlab="Runtime (minutes)", y
lab="Count (of movies)")
```

Code 9: Movies released per month

```
df <- read.csv("DisneyProject_CleanedUP (version 1).csv", na.strings ="?")

names(df)[names(df) == "Release.date..datetime."] <- "Date"

colnames(df)

df$Date <- parse_date_time(df$Date, orders= c( "ymd","dmy", "mdy"))

is.na(df$Date)

df$month <- month(df$Date, label = TRUE)

ggplot(df, aes(x=month, fill = month))+geom_bar()
```

Code 10: Regression model 1

```
model1 <- lm(as.numeric(Box.office..float.) ~ as.numeric(Budget..float.) +

              Running.time + Country + genre + Language, data=Disney)

summary(model1)

plot(model1)
```

Code 11: Regression model 2

```
model2 <- lm((as.numeric(imdb)) ~ as.numeric(Budget..float.) +
```

```
                Running.time..int. + Country + genre + Language, data=Disne
y)

summary(model2)

plot(model2)
```

Code 12: Prediction using Regression model 2

```
newdata <- data.frame(Budget..float. = 2600000, Running.time..int. = "88",
Country = "United States", genre = "Adventure", Language = "English")


predict(model2, newdata, interval = "prediction", level = 0.95)
```

APPENDIX B: DETAILED DESCRIPTION OF EACH GROUP MEMBER'S SPECIFIC CONTRIBUTIONS TO THE PROJECT

Debolina Sasmal:

- Assessing the business problem that can be inferred from the data.
- Exploratory Data Analysis: Creating the visualizations and assessing them.
- Regression Modelling.
- Forecasting using the Regression models.
- For PowerPoint Presentation: All slides except Data wrangling and Managerial conclusions.
- For Final Report: All pages except Data Wrangling and Managerial Conclusions
- For Project Proposal: Business Problem, Planned Analyses

Amy Yi:

- Data gathering from various sites and evaluating them.
- Merging multiple datasets and cleaning them up.
- For Exploratory Data Analysis: Visualization for Movies released per month.
- For PowerPoint Presentation: Wrote the slide on Data wrangling and Managerial conclusions.
- For Final Report: Data Wrangling, Managerial Conclusions
- For Project Proposal: Data Source, Preliminary Work

Abdirahman Hussein:

- For Project Proposal: Risk Assessment

APPENDIX B: REFERENCES

Stack overflow, https://stackoverflow.com/

STHDA, http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization

Aggregating and analyzing data with dplyr, https://datacarpentry.org/R-genomics/04-dplyr.html

STHDA, http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know

Residual Analysis in Regression, https://stattrek.com/regression/residual-analysis.aspx

Disney SWOT Analysis In A Nutshell, https://fourweekmba.com/disney-swot-analysis/

Disney SWOT analysis 2019 | SWOT Analysis of Disney, https://bstrategyhub.com/swot-analysis-of-disney-2019-disney-swot-analysis/

Disney Competitors, https://www.marketing91.com/disney-competitors/

Disney Revenue Breakdown Worldwide (2016 – 2021), https://businessquant.com/disney-revenue-breakdown-worldwide

Walt Disney Studios (division), https://en.wikipedia.org/wiki/Walt_Disney_Studios_(division)

Dataset 1:https://www.kaggle.com/therealsampat/disney-movies-dataset

Dataset 2:https://data.world/kgarrett/disney-character-success-00-16