Numerical Analysis of Disney Company Film Ratings and Relation to Revenue

Peter Way
CPSC 5240 Principles of Data Analytics
University of Tennessee at Chattanooga

I. Introduction

Disney films are some of the most iconic, with many individuals having at least one 'favorite' Disney film. But, do the success of these beloved films hinge on the rating given to them by the Motion Picture Association of America (MPAA)? This report seeks to answer this question, as well as develop a stronger understanding of the relationship between a films monetary success and rating. This single question is the goal for three different methods, those being Word Cloud, HeatMaps, and Correlograms. These individual methods are explored in Sections III, IV, and V. However, before one begins analysis, it is importance to review the data itself.

II. Data

The data used for this analysis was retrieved from Data World's database, from user Kelly Garrett [1].

No analysis of a data set is complete without an understanding of just what that data set includes. For the Disney films data, there were initially: the film's title, the film's genre, the film's MPAA rating at time of release, and the total gross income, as well as adjusted gross income for inflation. This last column will allow us to understand how well older films performed compared to those of today, without the uncertainty of currency values.

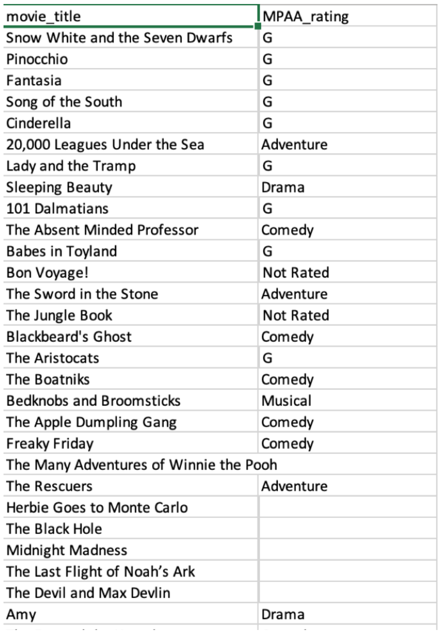
While the data was mostly complete, some adjustments and additions were required for analysis. For example, not all films had their MPAA rating included, so additional research was required to fill in those gaps. Any missing fields from the MPAA rating column were filled based off of their IMDB listing [2]. For an example of the mentioned missing values, please see Figure 1.

Additionally, both MPAA rating and genre were initially listed as string type variables and as such unable to be understood as factors. This created the necessity to convert those fields to numerical values to properly test their relationships to the other variables. The numeric values of ratings and genres are displayed in Table II.

III. Word Cloud

A. Goal Definition

As stated in Section I, understanding the relationship between how much success a film achieves and the rating passed down by the MPAA is the goal of this report.



movie_title	MPAA_rating
Snow White and the Seven Dwarfs	G
Pinocchio	G
Fantasia	G
Song of the South	G
Cinderella	G
20,000 Leagues Under the Sea	Adventure
Lady and the Tramp	G
Sleeping Beauty	Drama
101 Dalmatians	G
The Absent Minded Professor	Comedy
Babes in Toyland	G
Bon Voyage!	Not Rated
The Sword in the Stone	Adventure
The Jungle Book	Not Rated
Blackbeard's Ghost	Comedy
The Aristocats	G
The Boatniks	Comedy
Bedknobs and Broomsticks	Musical
The Apple Dumpling Gang	Comedy
Freaky Friday	Comedy
The Many Adventures of Winnie the Pooh	
The Rescuers	Adventure
Herbie Goes to Monte Carlo	
The Black Hole	
Midnight Madness	
The Last Flight of Noah's Ark	
The Devil and Max Devlin	
Amy	Drama

Fig. 1. Missing Values in Data

TABLE I
Ratings and Genres Numerically Adjusted

Genre as Listed	Numeric Value
Action	0
Adventure	1
Comedy	2
Concert/Performance	3
Crime Drama	4
Dark Comedy	5
Documentary	6
Drama	7
Fantasy	8
Horror	9
Musical	10
Romantic Comedy	11
Thriller	12
Western	13

MPAA Rating	Numeric Value
G	0
PG	1
PG-13	2
R	3
Not Rated	4

B. Acquiring and Cleaning of Data

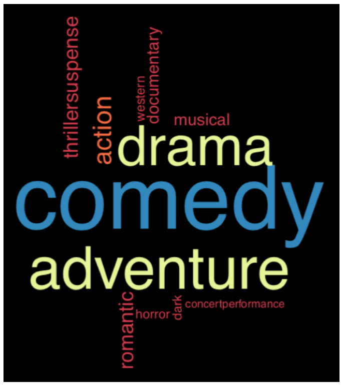
As stated in Section II, missing fields of the MPAA rating column were addressed via the aid of IMDB. MPAA Rating and Genre were both altered to list as numeric factors instead of their original string representation.

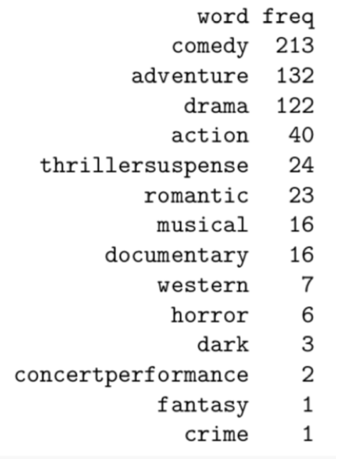
C. Method

Word clouds are useful in understanding the meanings of large sections of text. For this analysis, we will be looking into the original text-styled genre and title of the films Disney has released. Both traditional word clouds and tables of word counts figured into said clouds will be presented.

D. Method Creation, Analysis, and Decisions Made

The first word cloud up for analysis is the one created from the genres of Disney films. As Figure 2 demonstrates, the highest number of films were comedies, with adventure and drama second and third respectively. While this information by itself does not directly lend itself to understanding the relationship between revenue and ratings, we now understand that the majority of Disney films are indeed termed ’Family Friendly’ to at least some degree. As one can observe in the attached table, genres such as Horror and Dark Comedies are much further down the listings than Musicals.

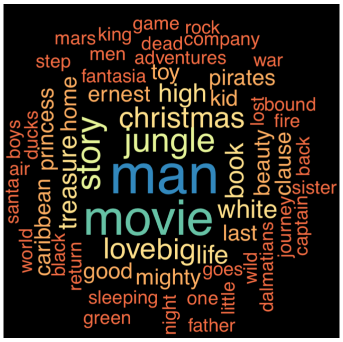


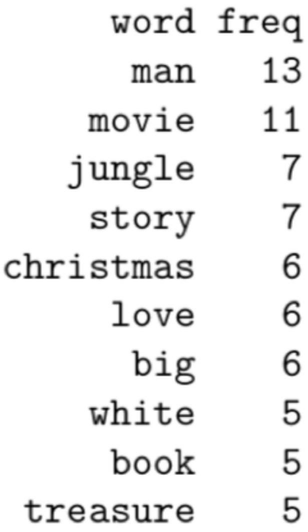


word	freq
comedy	213
adventure	132
drama	122
action	40
thrillersuspense	24
romantic	23
musical	16
documentary	16
western	7
horror	6
dark	3
concertperformance	2
fantasy	1
crime	1

Fig. 2. Word Cloud and Table for Genres

Continuing on to the second word cloud, shown in Figure 3 we now observe the count regarding the titles of Disney’s films. As shown, the top ten listings for titles are all generally positively associated words. All of the top ten listed words in titles are those considered to be attractive to younger audiences, or those looking for an exciting but ’safe’ option. The top ranking word in Disney film titles is ’man’, followed closely by ’movie’. However, if one ignores the word ’movie’, then the next ranking word would be ’jungle’. This invokes positive excitement in younger viewers, and a sense for adventure, as explored in Saif Mohammad’s work with emotions [3].





word	freq
man	13
movie	11
jungle	7
story	7
christmas	6
love	6
big	6
white	5
book	5
treasure	5

Fig. 3. Word Cloud and Table for Titles

Only one of the two versions of code will be shown here, in interest of conserving space. See Figure 4 for line by line code.

IV. Heat Map

A. Goal Definition

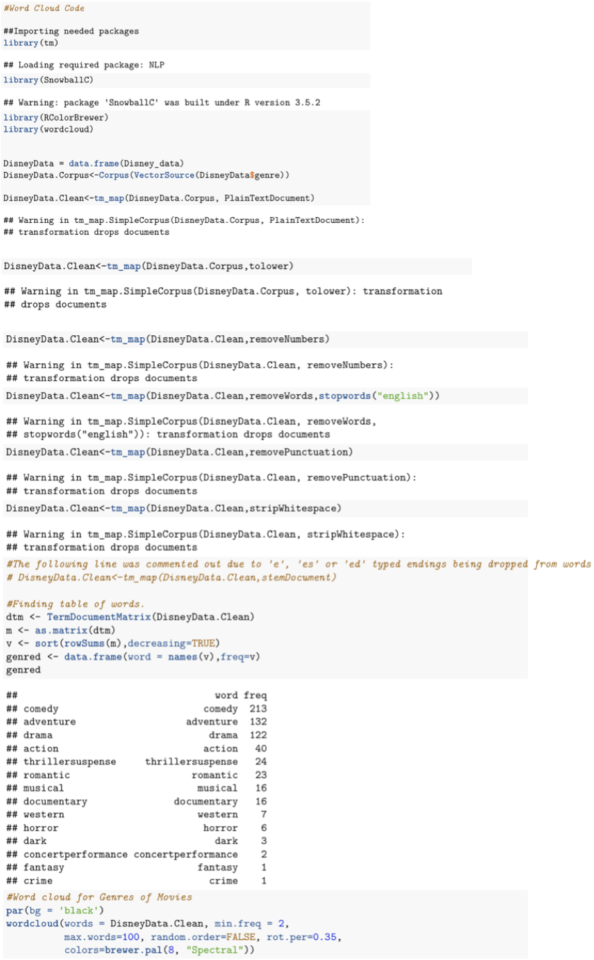
As stated in Section I, understanding the relationship between how much success a film achieves and the rating passed down by the MPAA is the goal of this report.

B. Acquiring and Cleaning of Data

As stated in Section II, missing fields of the MPAA rating column were addressed via the aid of IMDB. MPAA Rating and Genre were both altered to list as numeric factors instead of their original string representation.

C. Method

Another avenue for investigation in our exploration of Disney’s film archive is the use of Heat Maps. Heat maps



```
#Word Cloud Code

##Importing needed packages
library(tm)

## Loading required package: NLP
library(SnowballC)

## Warning: package 'SnowballC' was built under R version 3.5.2
library(RColorBrewer)
library(wordcloud)


DisneyData = data.frame(Disney_data)
DisneyData.Corpus<-Corpus(VectorSource(DisneyData$genre))

DisneyData.Clean<-tm_map(DisneyData.Corpus, PlainTextDocument)

## Warning in tm_map.SimpleCorpus(DisneyData.Corpus, PlainTextDocument):
## transformation drops documents


DisneyData.Clean<-tm_map(DisneyData.Corpus,tolower)

## Warning in tm_map.SimpleCorpus(DisneyData.Corpus, tolower): transformation
## drops documents


DisneyData.Clean<-tm_map(DisneyData.Clean,removeNumbers)

## Warning in tm_map.SimpleCorpus(DisneyData.Clean, removeNumbers):
## transformation drops documents
DisneyData.Clean<-tm_map(DisneyData.Clean,removeWords,stopwords("english"))

## Warning in tm_map.SimpleCorpus(DisneyData.Clean, removeWords,
## stopwords("english")): transformation drops documents
DisneyData.Clean<-tm_map(DisneyData.Clean,removePunctuation)

## Warning in tm_map.SimpleCorpus(DisneyData.Clean, removePunctuation):
## transformation drops documents
DisneyData.Clean<-tm_map(DisneyData.Clean,stripWhitespace)

## Warning in tm_map.SimpleCorpus(DisneyData.Clean, stripWhitespace):
## transformation drops documents
#The following line was commented out due to 'e', 'es' or 'ed' typed endings being dropped from words
# DisneyData.Clean<-tm_map(DisneyData.Clean,stemDocument)

#Finding table of words.
dtm <- TermDocumentMatrix(DisneyData.Clean)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
genred <- data.frame(word = names(v),freq=v)
genred

##                                    word freq
## comedy                           comedy  213
## adventure                     adventure  132
## drama                             drama  122
## action                           action   40
## thrillersuspense       thrillersuspense   24
## romantic                       romantic   23
## musical                         musical   16
## documentary                 documentary   16
## western                         western    7
## horror                           horror    6
## dark                               dark    3
## concertperformance concertperformance    2
## fantasy                         fantasy    1
## crime                             crime    1
#Word cloud for Genres of Movies
par(bg = 'black')
wordcloud(words = DisneyData.Clean, min.freq = 2,
          max.words=100, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Spectral"))
```

Fig. 4. Code for Creation of Genre Cloud

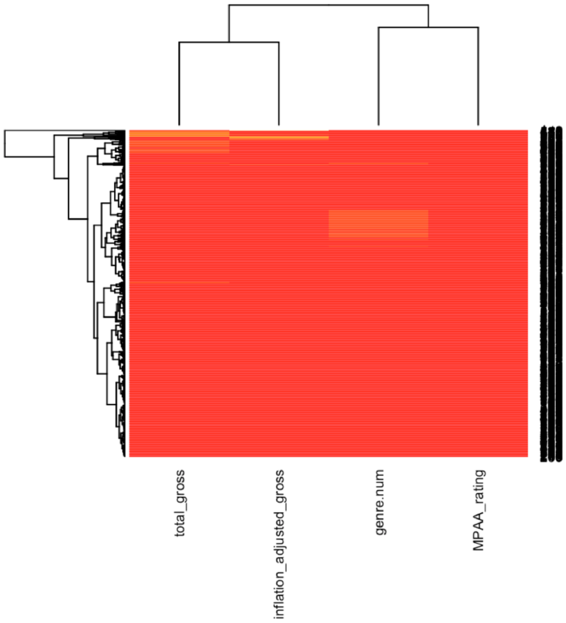
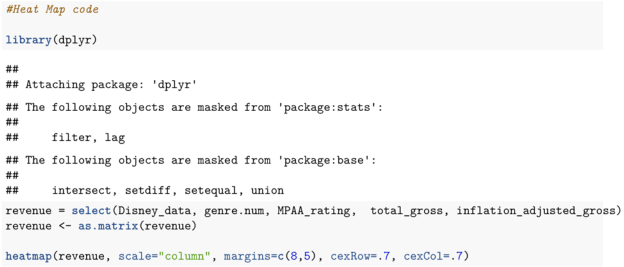


Fig. 5. Heat Map of Complete Numeric Data Set (Scaled)



```
#Heat Map code

library(dplyr)

##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
revenue = select(Disney_data, genre.num, MPAA_rating,  total_gross, inflation_adjusted_gross)
revenue <- as.matrix(revenue)

heatmap(revenue, scale="column", margins=c(8,5), cexRow=.7, cexCol=.7)
```

Fig. 6. Heat Map Code

are used to view where individual points lie on the respective column based on a 'heat' color.

D. Method Creation, Analysis, and Decisions Made

However, this particular data set is not quite conducive to heat map usage, as shown in Figure 5. The related code to the aforementioned heat map is shown in Figure 6. So, further testing was performed, narrowing the dataset down to the first one hundred entries. This new heat map is shown in Figure 7. Once again, not much useful data can be discerned from this style of output. Therefore, we must move along to correlograms.

V. Correlogram

A. Goal Definition

As stated in Section I, understanding the relationship between how much success a film achieves and the rating passed down by the MPAA is the goal of this report.

B. Acquiring and Cleaning of Data

As stated in Section II, missing fields of the MPAA rating column were addressed via the aid of IMDB. MPAA Rating and Genre were both altered to list as numeric factors instead of their original string representation.

C. Method

Correlograms are visual representations of how variables relate to one another. For example, if one were to look at a data set involving altitude and oxygen levels, one would observe a negative correlation. That is, as altitude increases, oxygen levels decrease. This visualized relationship dynamic allows us to see if there is indeed a correlation between income generated by Disney films and their ratings.

D. Method Creation, Analysis, and Decisions Made

The Correlogram created from the Disney film data set is shown in Figure 8, with its code shown in Figure 9.

As shown in the figure, there is a slight negative correlation between the inflation adjusted gross income of Disney films. This correlogram also demonstrates that the higher rating films are those that are less made. This is because we found the mass number of films created by Disney are those deemed Comedy, Adventure, and Drama. These are found earlier in the listings,

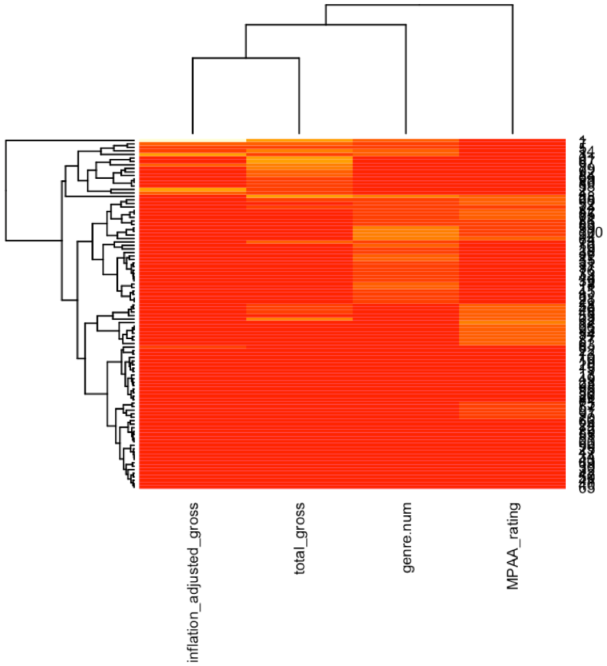


Fig. 7. Heat Map of Complete Numeric Data Set (Scaled)

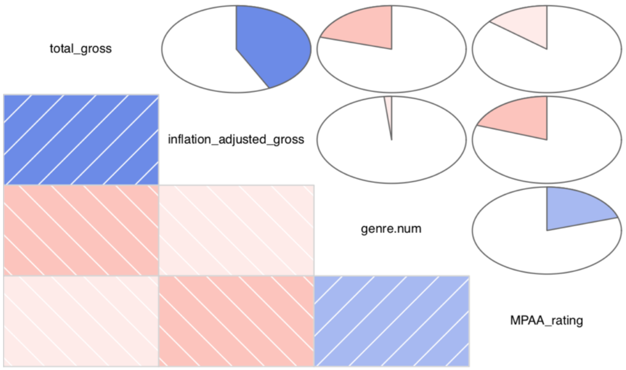
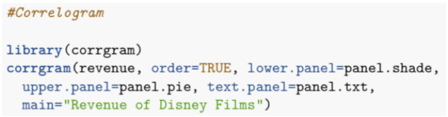


Fig. 8. Correlogram of Data

and as such have lower designation numbers than those found further into the alpha-listing.

An alternate approach to the correlogram is shown in Figure 10, with its code shown in Figure 11. This correlogram presents an easier to read format, with larger/darker circles meaning more of a relationship. Again we see that negative relationship between rating and revenue, but it is not particularly significant.



```
#Correlogram

library(corrgram)
corrgram(revenue, order=TRUE, lower.panel=panel.shade,
  upper.panel=panel.pie, text.panel=panel.txt,
  main="Revenue of Disney Films")
```

Fig. 9. Correlogram Code

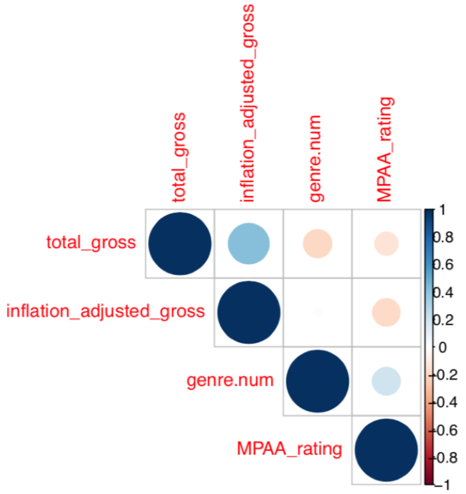
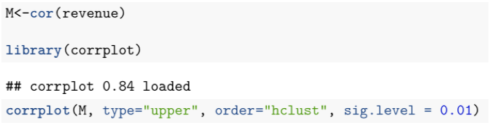


Fig. 10. Alternate Correlogram, Scaled.



```
M<-cor(revenue)

library(corrplot)

## corrplot 0.84 loaded
corrplot(M, type="upper", order="hclust", sig.level = 0.01)
```

Fig. 11. Alternate Correlogram Code

VI. Conclusions

This analysis of Disney films sought to prove there was a relationship between the rating that a film garners and its revenue generated. While there is some evidence to support this claim, more research should be conducted before this claim can become commonly accepted.

References

[1] Disney Character Success, K. Garrett. Disney characters, Box Office Success and Annual Gross Income. https://data.world/kgarrett/disney-character-success-00-16

[2] The Complete List of Disney Movies. IMDB. https://www.imdb.com/list/ls068561553/

[3] Mohammad, Saif M. and Turney, Peter D., Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon