KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING
School of Data Science and Analytics

Individuality Is Too Much Work!
How to Make a Good Unoriginal Movie

Nia Taylor
Dr. Ruvini Jayamaha, Dr. Sudhashree Sayenju

INTRODUCTION

Hundreds of movies are released every year with varying degrees of success, especially in terms of ratings. The inspiration of these movies also varies greatly, with some being more original than others. My goal for this project was to explore how originality of movies has changed over time and how ratings compare between original and non-original movies. I also wanted to find out what highly rated non-original movies have in common.

The dataset I used for this project was found on data.world and contained data scraped from hydramovies.com. The original dataset contained 3940 observations of 13 variables. I created a variable to determine whether a movie was original or not. Information to create the new variable was gathered from IMDB.com.

METHODS

I firstly conducted exploratory analysis to see how many original vs. non-original movies were in the dataset as well as the percentage of original vs non-original movies by year. Next, I calculated the mean rating of movies based on if a movie was original or not. Since the data was not normally distributed, I conducted a Kruskal Wallis test to determine whether the mean rating on ranks was statistically different based on a movie's originality.

I then conducted further exploratory analysis to find out which genre was most prevalent in highly rated non-original movies.

I conducted text mining analysis to determine which words were present most often in the short summaries for highly rated (ratings over 7) non-original movies.

Finally, I conducted sentiment analyses to determine which emotional sentiments were present most often for highly rated non-original movies.

RESULTS

There was almost twice as many original movies as non-original movies released from 2004 to 2018.

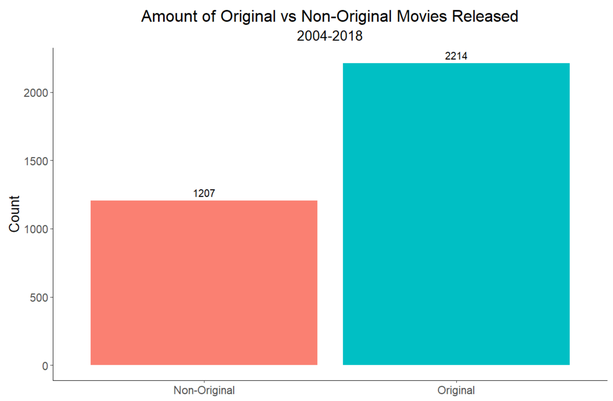
The percentage of original to non-original movies has been nearly constant with only a slight increase.

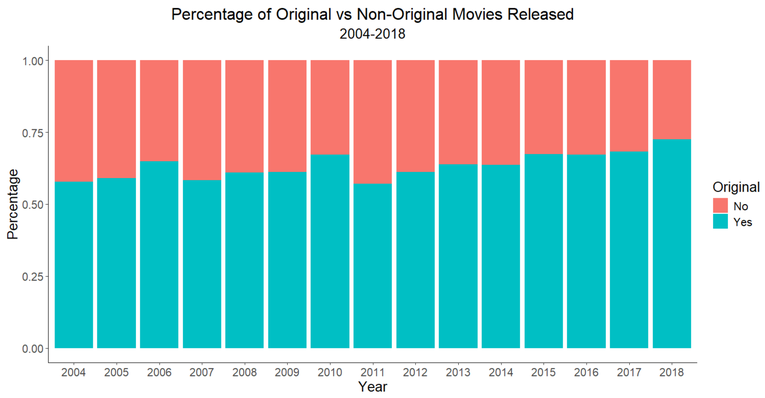
The Kruskal-Wallis test conducted to determine whether the mean rating on ranks was statistically different based on a movie's originality had a p-value of less than 0.05 meaning that original movies have a statistically significantly higher mean.

Action movies were the most frequent genre of highly rated non-originals.

Text mining analysis of the short summaries of highly rated non-original movies shows that the most common words were: young, life, new, man, and world (descending order).

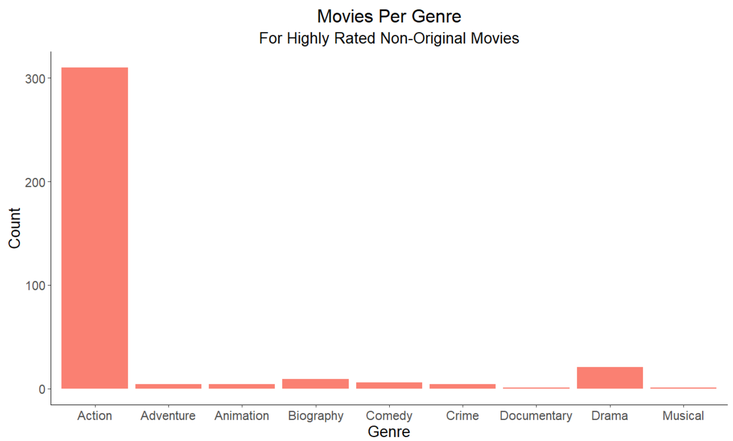
Emotional sentiment analysis of the short summaries of highly rated non-original movies shows that the most frequent sentiment was fear.

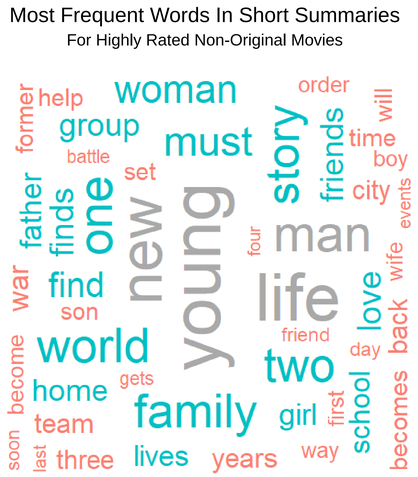


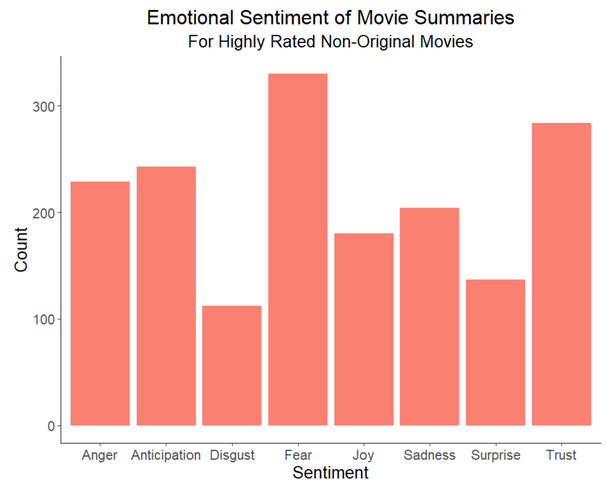


Original	Mean Rating
Yes	6.583679
No	6.494219

Kruskal-Wallis Test Summary	Rating by Original
chi-squared	7.3872
df	1
p-value	0.006569







DISCUSSION

Non-original movies have been less frequently released than original movies; Possibly due to their lower average ratings. Awareness of what attributes are common amongst highly rated non-original movies can help push filmmakers in the right direction if they are looking to make a sequel, adaptation, remake, etc., that is highly rated.

Future research could include more recent releases to see how the trend between original vs non-original movie releases is changing over time. Also, further research could include exploring originality's impact on a movies financial profit.

CODE SNIPPET

```
kruskal.test(Rating ~ Original, data = movies_2004_2018)

TextDoc<-highratedunoriginals[,c(1,4,13)]TextDoc  <-
Corpus(VectorSource(movies_2004_2018$Short.Summary))
TextDoc_dtm <- TermDocumentMatrix(TextDoc)dtm_m <-
as.matrix(TextDoc_dtm)dtm_v <-
sort(rowSums(dtm_m),decreasing=TRUE)dtm_d <-
data.frame(word = names(dtm_v),freq=dtm_v)head(dtm_d, 5)

syuzhet_vector <-
get_sentiment(highratedunoriginals$Short.Summary,
method="syuzhet")
vectorhead(syuzhet_vector)
vectorsummary(syuzhet_vector)
```

References

https://www.imdb.com/
https://data.world/iliketurtles/movie-dataset