MSDS 597 Final Project - Movie Exploratory Data Analysis

Shang-Hao Huang

April 2021

Broad Goals

This project's broad goal of this project is to explore the IMDB movie dataset and address exciting issues such as the genre of movies that are mainly produced, IMDB score analysis, and profitability analysis. Moreover, I want to identify what set of features contribute to a highly rated/profitable film most significantly and try to predict a movie's profitability, if possible.

Load the requried R packages.

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.3
```

```
## Registered S3 methods overwritten by 'tibble':
##   method     from
##   format.tbl pillar
##   print.tbl  pillar
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.0     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
## Warning: package 'tibble' was built under R version 3.6.3
```

```
## Warning: package 'tidyr' was built under R version 3.6.3
```

```
## Warning: package 'readr' was built under R version 3.6.3
```

```
## Warning: package 'purrr' was built under R version 3.6.3
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
```

```
## Warning: package 'stringr' was built under R version 3.6.3
```

```
## Warning: package 'forcats' was built under R version 3.6.3

## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(DT)
```

```
## Warning: package 'DT' was built under R version 3.6.3
```

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 3.6.3
```

```
library(tm)
```

```
## Warning: package 'tm' was built under R version 3.6.3

## Loading required package: NLP

## Warning: package 'NLP' was built under R version 3.6.3

##
## Attaching package: 'NLP'

## The following object is masked from 'package:ggplot2':
##
##     annotate
```

```
library(wordcloud)
```

```
## Warning: package 'wordcloud' was built under R version 3.6.3

## Loading required package: RColorBrewer
```

```
library(fitdistrplus)
```

```
## Warning: package 'fitdistrplus' was built under R version 3.6.3

## Loading required package: MASS

## Warning: package 'MASS' was built under R version 3.6.3

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

## Loading required package: survival
```

```
library(plotly)
```

```
## Warning: package 'plotly' was built under R version 3.6.3
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

Data Preparation

The dataset used in this project is the IMDB 5000 Movie Dataset from Kaggle, which can be accessed via this link: https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset. This dataset recorded information on 5043 movies across 66 countries from 1916 to 2016. The dataset is available in a csv format file and is of size 1MB.

Each record in this dataset has 28 variables, including information such as `Title of the movie'',`Name of the movie director'', `Country the movie was produced in'',`Budget of the movie ($)'', `profitability'', `and`IMDB score for the movie (out of 10)".

Data Import

Import the data and show the dimension and names of all attributes of the data. There are 5043 movies recorded in this dataset, and each of which has 28 attributes. Note that as Kaggle requires a username and password to download the dataset, I am sourcing the same data from my Github repository.

```
url <- "https://raw.githubusercontent.com/TimotheusHuang/msds597finalproj/main/movie_metadata.csv"
movie <- as_tibble(read.csv(url, stringsAsFactors = FALSE))
class(movie)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

```
dim(movie)
```

```
## [1] 5043   28
```

Table 1: Variable Types, Names, and Description

Type	Variable Name	Description
-	ID	ID of each client
credit	LIMIT_BAL	Amount of given credit in NT dollars (includes individual and family/supplementary credit)
demographic	SEX	Gender (1=male, 2=female)
demographic	EDUCATION	(1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
demographic	MARRIAGE	Marital status (1=married, 2=single, 3=others)
demographic	AGE	Age in years
payment	PAY_0	Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, ..., 8=payment delay for eight months, 9=payment delay for nine months and above)
payment	PAY_2	Repayment status in August, 2005 (scale same as above)
payment	PAY_3	Repayment status in July, 2005 (scale same as above)
payment	PAY_4	Repayment status in June, 2005 (scale same as above)
payment	PAY_5	Repayment status in May, 2005 (scale same as above)
payment	PAY_6	Repayment status in April, 2005 (scale same as above)
bill	BILL_AMT1	Amount of bill statement in September, 2005 (NT dollar)
bill	BILL_AMT2	Amount of bill statement in August, 2005 (NT dollar)
bill	BILL_AMT3	Amount of bill statement in July, 2005 (NT dollar)
bill	BILL_AMT4	Amount of bill statement in June, 2005 (NT dollar)
bill	BILL_AMT5	Amount of bill statement in May, 2005 (NT dollar)
bill	BILL_AMT6	Amount of bill statement in April, 2005 (NT dollar)
payment	PAY_AMT1	Amount of previous payment in September, 2005 (NT dollar)
payment	PAY_AMT2	Amount of previous payment in August, 2005 (NT dollar)
payment	PAY_AMT3	Amount of previous payment in July, 2005 (NT dollar)
payment	PAY_AMT4	Amount of previous payment in June, 2005 (NT dollar)
payment	PAY_AMT5	Amount of previous payment in May, 2005 (NT dollar)
payment	PAY_AMT6	Amount of previous payment in April, 2005 (NT dollar)
payment	default.payment.next.month	Default payment (1=yes, 0=no)

```
colnames(movie)
```

```
##  [1] "color"                    "director_name"
##  [3] "num_critic_for_reviews"   "duration"
##  [5] "director_facebook_likes"  "actor_3_facebook_likes"
##  [7] "actor_2_name"             "actor_1_facebook_likes"
##  [9] "gross"                    "genres"
## [11] "actor_1_name"             "movie_title"
## [13] "num_voted_users"          "cast_total_facebook_likes"
## [15] "actor_3_name"             "facenumber_in_poster"
## [17] "plot_keywords"            "movie_imdb_link"
## [19] "num_user_for_reviews"     "language"
## [21] "country"                  "content_rating"
## [23] "budget"                   "title_year"
## [25] "actor_2_facebook_likes"   "imdb_score"
## [27] "aspect_ratio"             "movie_facebook_likes"
```