Box Office Blues

STAT-420, Team: Summer Proj, A Arora, S Dani, G Shrivastava

July 2020

Team

The names of the students who will be contributing to the group project:

- **amana4** (Aman Arora)
- **dani4** (Savvy Dani)
- **gaurav4** (Gourav Shrivastava)

Project

A tentative title for the project: **Box Office Blues**

Dataset

The data file is a cvs file with 4803 records and 20 columns. It contains metadata and revenue information for over 5000 movies sourced from Kaggle TMDB 5000 Movie Dataset (https://www.kaggle.com/tmdb/tmdb-movie-metadata?select=tmdb_5000_movies.csv) . A few variables of interest are:

- **Original_title**: Name of the movie
- **Budget**: Budget of movies in USD (numeric)
- **Revenue**: Revenue of movie in USD (numeric)
- **Original Language**: The language in which movie was originally produced (factor variable)
- **Genres**: Genre of the movie (factor variable)
- **Popularity**: A numeric metric to measure popularity of the movie (numeric)
- **Vote Average**: A numeric metric to measure average vote from audience (numeric)

- **Runtime** : A numeric metric for the total runtime(in min) of the movie (numeric)
- **Production Companies** : A categorical for the production companies name (factor)

Goals

In 2018, the global box office was worth $41.7 billion. In 2019, total earnings at the North American box office amounted to $11.32 billion. The magic movies create in our daily lives is undeniable, but more interesting to us is the story the data tells us.

As part of this project, we will like to predict the 'Revenue' of the movie. We will be exploring different features like 'genres', 'runtime', 'budget', 'vote_average','vote_count','production_companies' to find the best possible model. We will validate our model performance by holding a 'validation' dataset.

Sample DataSet with Preliminary Data Analysis

- Evidence that the data can be loaded into `R` . Load the data, and print the first few values of the response variable as evidence.

```
tmdb_movies = read_csv("tmdb_5000_movies.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   budget = col_double(),
##   id = col_double(),
##   popularity = col_double(),
##   release_date = col_date(format = ""),
##   revenue = col_double(),
##   runtime = col_double(),
##   vote_average = col_double(),
##   vote_count = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```
#Select the relevant columns
col_sel = c( "original_title","revenue", "budget","popularity", "vote_a
verage","runtime","genres","production_companies","original_language")
tmdb_movies_small = tmdb_movies[,col_sel]
```

Data Snippet

Here is a snippet of data with only the columns considered.

```
ft <- flextable(head(tmdb_movies_small,n=10))
ft <- autofit(ft)
ft
```

original_title	revenue	budget	popularity	vote_average	runtime	genres	production_companies	original_language
Avatar	2787965087	237000000	150.44	7.2	162	[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}, {"id": 878, "name": "Science Fiction"}]	[{"name": "Ingenious Film Partners", "id": 289}, {"name": "Twentieth Century Fox Film Corporation", "id": 306}, {"name": "Dune Entertainment", "id": 444}, {"name": "Lightstorm Entertainment", "id": 574}]	en
Pirates of the Caribbean: At World's End	961000000	300000000	139.08	6.9	169	[{"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}, {"id": 28, "name": "Action"}]	[{"name": "Walt Disney Pictures", "id": 2}, {"name": "Jerry Bruckheimer Films", "id": 130}, {"name": "Second Mate Productions", "id": 19936}]	en
Spectre	880674609	245000000	107.38	6.3	148	[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 80, "name": "Crime"}]	[{"name": "Columbia Pictures", "id": 5}, {"name": "Danjaq", "id": 10761}, {"name": "B24", "id": 69434}]	en

original_title	revenue	budget	popularity	vote_average	runtime	genres	production_companies	original_language
The Dark Knight Rises	1084939099	250000000	112.31	7.6	165	[{"id": 28, "name": "Action"}, {"id": 80, "name": "Crime"}, {"id": 18, "name": "Drama"}, {"id": 53, "name": "Thriller"}]	[{"name": "Legendary Pictures", "id": 923}, {"name": "Warner Bros.", "id": 6194}, {"name": "DC Entertainment", "id": 9993}, {"name": "Syncopy", "id": 9996}]	en
John Carter	284139100	260000000	43.93	6.1	132	[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 878, "name": "Science Fiction"}]	[{"name": "Walt Disney Pictures", "id": 2}]	en
Spider-Man 3	890871626	258000000	115.70	5.9	139	[{"id": 14, "name": "Fantasy"}, {"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}]	[{"name": "Columbia Pictures", "id": 5}, {"name": "Laura Ziskin Productions", "id": 326}, {"name": "Marvel Enterprises", "id": 19551}]	en
Tangled	591794936	260000000	48.68	7.4	100	[{"id": 16, "name": "Animation"}, {"id": 10751, "name": "Family"}]	[{"name": "Walt Disney Pictures", "id": 2}, {"name": "Walt Disney Animation Studios", "id": 6125}]	en
Avengers: Age of Ultron	1405403694	280000000	134.28	7.3	141	[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 878, "name": "Science Fiction"}]	[{"name": "Marvel Studios", "id": 420}, {"name": "Prime Focus", "id": 15357}, {"name": "Revolution Sun Studios", "id": 76043}]	en

original_title	revenue	budget	popularity	vote_average	runtime	genres	production_companies	original_language
Harry Potter and the Half-Blood Prince	933959197	250000000	98.89	7.4	153	[{"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}, {"id": 10751, "name": "Family"}]	[{"name": "Warner Bros.", "id": 6194}, {"name": "Heyday Films", "id": 7364}]	en
Batman v Superman: Dawn of Justice	873260194	250000000	155.79	5.7	151	[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}]	[{"name": "DC Comics", "id": 429}, {"name": "Atlas Entertainment", "id": 507}, {"name": "Warner Bros.", "id": 6194}, {"name": "DC Entertainment", "id": 9993}, {"name": "Cruel & Unusual Films", "id": 9995}, {"name": "RatPac-Dune Entertainment", "id": 41624}]	en

```
#Find rows with 0 values and set to NA
tmdb_movies_small[tmdb_movies_small$budget == 0, "budget" ] = NA
tmdb_movies_small[tmdb_movies_small$revenue == 0, "revenue" ] = NA
tmdb_movies_small[tmdb_movies_small$popularity == 0, "popularity" ] = N
A
#tmdb_movies_small[tmdb_movies_small$vote_count == 0, "vote_average" ]
= NA
tmdb_movies_small$original_language = as.factor(tmdb_movies_small$origi
nal_language)

#remove invalid rows
tmdb_movies_small =  na.omit(tmdb_movies_small)
```

```
#Fit a simple  model and print summary
boxoffice_model_1 = lm(revenue ~  budget+popularity , tmdb_movies_smal
l)
summary(boxoffice_model_1)
```

```
##
## Call:
## lm(formula = revenue ~ budget + popularity, data = tmdb_movies_smal
l)
##
## Residuals:
##       Min         1Q    Median        3Q       Max
## -1.04e+09 -4.66e+07 -7.09e+06  2.35e+07  1.99e+09
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.68e+07   2.94e+06   -9.13   <2e-16 ***
## budget       2.30e+00   5.14e-02   44.69   <2e-16 ***
## popularity   1.88e+06   6.31e+04   29.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 117000000 on 3226 degrees of freedom
## Multiple R-squared:  0.606,  Adjusted R-squared:  0.606
## F-statistic: 2.49e+03 on 2 and 3226 DF,  p-value: <2e-16
```

A simple additive model above was used to predict revenue based on popularity and budget of the movie. The predictors seems to be significant, our objective is to improve this simplistic model in final report of this project.