Movie Predictions

Project Team Members:
Shikha Verma (Team Leader): A20408401
Ayushi Patel: A20407392
Mayur Mehta: A20405901
Abhilash Bhurse: A20404893

1. Project Proposal

- **A formal description of the project with a stated research goal.**
  A movie is not only for entertaining users, but also for a film company to make great profits. There are lot of factors needed for a movie to be a commercial success. The goal of this project is to derive such insights which help making an informed decision for the future generations of movies. For this project, we take IMDB movie dataset from Kaggle website and would like to analyze what kind of movies are more successful or obtained a higher IMDB score than others. The IMDB scores are taken as response variables and would try to obtain predictions based on analyzing them. The results from this project can help the companies to understand the factors of successful movie and answer some interesting questions which are stated below

- **A specific question or set of questions that the project seeks to address.**

  i. Which countries are producing most movies?
  ii. What are the important factors that make a movie more successful than others?
  iii. Which type of movies are profitable in future?
  iv. What kind of movies are most produced?
  v. What genres are the most produced?
  vi. Do IMDB ratings affect the revenue of a movie?

- **A proposed methodology/approach to the analysis that will be performed.**

- First, we will perform preprocessing of data, i.e cleaning the data and preparing it for analysis and predict some outcomes from the data and plot them on graph to derive predictive results for future
- After cleaning of data, data analysis on dataset is performed which can be given as follows

i. Data Preparation:

Data Import:

- We will be importing the data in R import library.
- 5043 observations of 28 variables, spanning across 100 years in 66 countries.

Data Cleaning:

- Removal of spurious characters from the movie title, genre and plot keyword.
- Remove the duplicates in the data, using then "movie_title" column
- To compare financial columns across movies and countries, currency columns are converted in to USD.

ii. Exploratory Data Analysis:

Diverse set of packages, functions and graphical methods will be used to explore the movies dataset, methods included bar chart to statistical heavy distribution fitting

Genre Analysis:

- Cleaning of the Variables
- Frequency Analysis
- Association Analysis

Country:

- Explore the country Variables
- Exploring the budget varies by each country
- Exploring the gross revenue
- Analysis on languages used

IMDB Score Analysis:

- Analysis on mean and the variation in movie score
- Data Distribution
- Analysis of relationship between IMDB score, revenue and budget of the movie
- Analysis of overall profitability

- **A metric or set of metrics which will measure analysis results.**

- For the frequently used, the differences between values which are predicted by a model and the values observed is measured by Root mean square error metric.
- To measure the difference between two continuous variables, Mean absolute error metric will be used.
- Precision, Recall and F1 metrics will be used to determine the performance feature

2. Project Outline

- **Literature review and related work - existing projects, references, papers, and relevant articles, etc.**

Data Repository:

https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset
https://www.kaggle.com/orgesleka/imdbmovies

Supplemental Resources:

http://rstudio-pubs-static.s3.amazonaws.com/337246_91de37b30146468ab2b2684fc570baf5.html
https://data.world/data-society/imdb-5000-movie-dataset
https://webpage.pace.edu/rp84697n/cs641/Portfolio/U01382090_Rakesh_Movie%20Report.pdf

Related Work:

https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a095.pdf
http://cs229.stanford.edu/proj2011/YooKanterCummingsPredictingMovieRevenuesUsingImdbData.pdf

- **All data sources and reference data with descriptions - dataset type (documents, trial data, etc.), dataset size (number of observations, number of dimensions), feature descriptions, other notes (missing data, processing issues, etc.), etc.**

Data Set Characteristics: Multivariate
Associated Tasks: Regression
Number of Observations: 5043
Number of Columns: 28
Missing Values? : Yes
Area: Entertainment Industry

Column Name	Size	Description	Data Type	Other Notes
color	5043	It depicts if the movie is made in color or not	String	Valid Entries: 5024 Mismatched Entries: 19 Unique values: 2
director_name	5043	The name of the director of the movie	String	Valid Entries: 4939 Mismatched Entries: 104 Unique values: 2398
num_critic_for_reviews	5043	The number of critics for reviewing	Numeric	Valid Entries: 4993 Mismatched Entries: 50
duration	5043	The running duration of movie	Numeric	Valid Entries: 5028 Mismatched Entries: 15
director_facebook_likes	5043	The likes on Facebook for director	Numeric	Valid Entries: 4939 Mismatched Entries: 104
actor_3_facebook_likes	5043	The number of likes on Facebook for Actor 3	Numeric	Valid Entries: 5020 Mismatched Entries: 23
actor_2_name	5043	The name of the Actor 2 starring in movie	String	Valid Entries: 5030 Mismatched Entries: 13 Unique Values: 3032
actor_1_facebook_likes	5043	The number of likes on Facebook for Actor 1	Numeric	Valid Entries: 5036 Mismatched Entries: 7
gross	5043	The gross income of movie	Numeric	Valid Entries: 4159 Mismatched Entries: 884
genres	5043	The genre of film	String	Valid Entries: 5043 Mismatched Entries: 0
actor_1_name	5043	The name of Actor 1 starring in movie	String	Valid Entries: 5036 Mismatched Entries: 7 Unique Values:2097
movie_title	5043	The title of movie	String	Valid Entries: 4939 Mismatched Entries: 104 Unique Values: 4917
num_voted_users	5043	The number of users who voted	Numeric	Valid Entries: 5043 Mismatched Entries: 0
cast_total_facebook_likes	5043	The total number of likes on Facebook by the cast	Numeric	Valid Entries: 5043 Mismatched Entries: 0

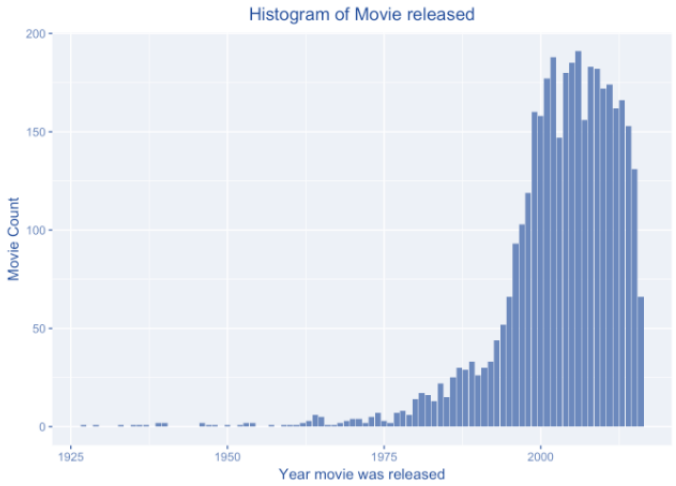
actor_3_name	5043	The name of the Actor 3 starring in movie	String	Valid Entries: 5020 Mismatched Entries: 23 Unique Values: 3521
facenumber_in_poster	5043	The number of faces depicted in poster of movie	Numeric	Valid Entries: 5030 Mismatched Entries: 13
plot_keywords	5043	The keywords used in movie	String	Valid Entries: 4890 Mismatched Entries: 153 Unique Values: 4760
movie_imdb_link	5043	The link on imdb for movie	String	Valid Entries: 5043 Mismatched Entries: 0 Unique Values: 4919
num_user_for_reviews	5043	The number of users who reviewed the movie	Numeric	Valid Entries: 5022 Mismatched Entries: 21
language	5043	The language in which movie is made	String	Valid Entries: 5031 Mismatched Entries: 12 Unique Values: 47
country	5043	The country in which movie is made	String	Valid Entries: 5038 Mismatched Entries: 5 Unique Values: 65
content_rating	5043	The rating given to movie	String	Valid Entries: 4740 Mismatched Entries: 303 Unique Values: 18
budget	5043	The budget for the movie	Numeric	Valid Entries: 4551 Mismatched Entries: 492
title_year	5043	The year in which movie released	Numeric	Valid Entries: 4935 Mismatched Entries: 108
actor_2_facebook_likes	5043	The number of likes on Facebook for Actor 2	Numeric	Valid Entries: 5030 Mismatched Entries: 13
imdb_score	5043	The score on imdb for the movie	Numeric	Valid Entries: 5043 Mismatched Entries: 0
aspect_ratio	5043	The aspect ratio of the movie	Numeric	Valid Entries: 4714 Mismatched Entries: 329
movie_facebook_likes	5043	The number of likes on Facebook for movie	Numeric	Valid Entries: 5043 Mismatched Entries: 0

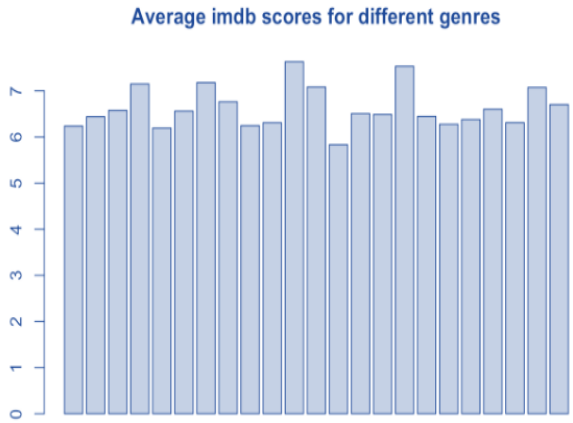
- **Data processing and pipeline - cleaning, imputing, transformation, outlier detection, etc.**

- Data import – Importing the data from kaggle data source.
- Data Cleaning - removing characters from keyword like movie title, genre & plot columns and duplicates.
- Data preview and Analysis based on Data Description after replacing the missing values with suitable value.
- Categorical distribution based on genre, country, IMDB

- **Data stylized facts - distributional analysis, clustering, dimensionality reduction, etc.**
  Some basic example graphs that represent our IMDB movie dataset from Kaggle and will implement more analyzed graphs as the project proceeds

(i) Movie release years in histogram
(ii) Average of IMDB scores of different genres





- **Model selection - feature selection requirements, classification/regression approaches, reference/baseline model, etc.**

➢ **Regression and Predicting Model.**
This model has a goal to predict if the movie is good or bad based on the IMDB score given by the users and profitability analysis of movie to predict the values for future references. Also, as the project proceeds we will compare different attributes (Linear Regression, KNN, etc.) in model to deliver accurate results

- **Software packages, applications, libraries, and associated tools, etc.**

Libraries/Packages:

tibble, DT, knitr, tm, ggplot2, wordcloud, dplyr, fitdistrplus, plotly, plyr will be used for the project.

Softwares:

RStudio, R

Project Management and Source Control:

Kaggle Dataset/Website, Some analysis pdfs and GitHub