**Study IMDB 5000 Movies Dataset**

CA1 submitted in part fulfilment of the requirements for the degree of

MSc in Data Science (Computer Science)

at Technological University Dublin

Shubham Jangir

2020-2021 STUDENT NUMBER: D20124818

15th Nov’20 COURSE: TU059(FULL-TIME)

# Synopsis:

# Problem Statement: To explore the IMDB 5000 movie dataset and gain the insight from the dataset to answer some a series of questions as an analyst.

# Data Preparation:

# Dataset: <https://www.kaggle.com/carolzhangdc/imdb-5000/download>

# Data Import: This dataset contains 5043 rows and 28 columns spanning over 100 years in 66 countries. There are multiple directors and actors/actress names. For our analysis we are considering IMDB Score as our primary variable and other variables which are influencing the IMDB Score.

# Data Cleaning:

# Removing Duplicates: The first part of the data cleaning involves removal of duplicates in the dataset that can cause error in our analysis, so duplicated were deleted and now we have 4998 rows and 28 columns.

# Removing Missing Data: Second part of the cleaning involves removing the missing or NA values as they may cause discrepancies during the analysis, we can’t go ahead and remove all the missing as this might remove some import insight from the data, so we need to go ahead and do the analysis on each of the columns which contains NA.

# First let’s check the gross and budget column as they contain most of the missing values, and we cannot impute anything in place of missing values as most of values are in millions, so we go ahead and remove the NA from gross and budget field, and we are left with 3857 rows and 28 columns.

# Aspect Ration contains the third highest NA values, so we go ahead and check what is the most common aspect ratio i.e., 2.35 and 1.85 and we group the other aspect ratio together and assign 0 to all the NA’s in aspect ratio column and compute the mean of aspect ratio wrt. Gross and we found out that mean is the range of 44Million to 58Million, so we can go ahead and remove aspect ratio as it won’t affect our analysis and now, we are left with 3857 rows and 27 columns.

# Analyzing content rating field for NA we found out that there is very less missing value, so we go ahead and remove them, moreover according to motion pictures M, GP=PG and we have given Unrated, Not Rated, Passed and Approved under R and PG-13 and changed our content rating field to factor with level 5 and now we are left with 3806 rows and 27 columns.

# Analyzing colour column of the dataset, as 95% of the movies are in colour we can go ahead and remove this field as this constant and won’t impact much in our analysis now we are left with 3806 rows and 26 columns.

# Analyzing language column of the dataset, as most of the movies are in English language we can go ahead and remove this column because this will act as constant now, we are left with 3805 rows and 25 columns.

# Analyzing the facenumber\_in\_poster, director\_facebook\_likes and movie\_facebook\_likes columns of the movies, these 3 columns contains zeros so let’s assign NA to all the zeros for analysis purpose and fill NA's with round of mean of their subsequent column values.

# Analyzing the num\_critic\_for\_reviews, actor\_3\_facebook\_likes, actor\_2\_facebook\_likes and actor\_1\_facebook\_likes as these field contains very less NA values we can go ahead and fill the NA with the round of mean of their subsequent column values.

# Analyzing the actor\_1\_name, actor\_2\_name and actor\_3\_name as these field contains very less NA values we can go ahead and remove them because these are categorical fields now, we are left with 3800 rows and 25 columns.

# Analyzing plot keyword contains 21 values and there is no way to fill those missing values so we can go ahead and remove them now, we are left with 3779 rows and 25 columns.

# Removing the | symbols from genres and plot keyword with whitespace as pipe symbol will cause error in our analysis.

# Adding 2 new columns: For our analysis we will add revenue\_generated (gross – budget) and returns\_percent ((revenue\_generated/budget)/100) to our analysis now, we are having 3779 rows and 27 columns as our final dataset without any duplicates and missing values.

# Data Exploration:

# Using names command we found out that there are 27 columns in the dataset (see Rscript 2.1.1).

# Checking the structure of the dataset there are 3779 rows and 27 columns with one factor field i.e., content rating with 5 levels (see R script 2.1.2).

# Checking the summary statistics on the dataset using summary command (see R script 2.1.3).

# Exploring Content Rating visually- Most of the movies have R content rating and after that its PG-13 (see R script 2.1.4).

# Exploring IMDB Score- Most movies fall in the range of 5-7.5 IMDB Rating (see R script 2.1.5).

# Exploring Title year i.e., year in which the movie got released- Most movies released after 2000's (see R script 2.1.6).

# Exploring Country with highest number of movie release- US has produced the greatest number of the movies and after that it's UK (see R script 2.1.7).

# Exploring Percent Return by IMDB Score - Movies with higher IMDB rating has higher percentage return if we leave few outliners (see R script 2.1.8).

# Exploring Budget by IMDB Score- Budget is quite high for movies with higher IMDB Rating (see R script 2.1.9).

# Exploring Gross by IMDB Score - Movies with higher IMDB Rating have higher gross (see R script 2.1.10).

# Exploring Movie Facebook Likes by IMDB Score - Higher IMDB Rating movies have higher Facebook likes (see R script 2.1.11).

# Exploring Critics Review by IMDB Score - If the critics reviews are higher there are more chances that the movie will have higher IMDB Rating (see R script 2.1.12).

# Exploring User Review by IMDB Score -If the user reviews are higher there are more chances that the movie will have higher IMDB Rating (see R script 2.1.13).

# Exploring Content Rating by IMDB Score - Using the boxplot we can see that movies with R and PG-13 Content Rating have the higher IMDB Rating (see R script 2.1.14).

# Exploring Director Facebook likes by IMDB Score - IMDB Rating is not much dependent on Director's Facebook likes (see R script 2.1.15).

# Exploring Revenue Generated by IMDB Score - Apart from few outliners most the movies with higher IMDB Rating have higher Revenue (see R script 2.1.16).

# Exploring Average Budget by Countries-South Korea has the highest budget, but this can be the outliners as most of the movies are produced in USA (see R script 2.1.17).

# Insight from the Dataset:

# Insight 1: What is the trend for number of movies produced annually and their corresponding IMDB Rating?

# There are more movies produced after 1980 and 2000 reaching unto 200 movies per year but the IMDB scores of the movies came down low despite being the fact that a greater number of movies are produced, there is a significant increase in the number of low-quality movies (see R script 3.1.1).

# Insight 2: What is the trend of IMDB Score as per Budget and Gross?

# Movies with higher IMDB Rating have the higher Budget they are positively corelated movies with higher budget do better on box office.

# Movies with IMDB Rating above than 8 have higher gross and there is a significant increase (see R script 3.1.2).

# Insight 3: Top Prolific Directors with highest IMDB Rating and minimum 10 movies directed?

# It is evident from the graph that David Fincher has produced the highest IMDB Rating movies with IMDB Rating more than 7.5 on average (see R script 3.1.3).

# Insight 4: What is the impact of Critics Review and Audience Review together on IMDB Rating?

# It is evident from the graph that as the Critics Review and Audience Review are good for movie it will get the higher IMDB Rating and they are positively corelated (see R script 3.1.4).

# Insight 5: Top 10 movies with Highest Budget and Highest Revenue Generated?

# It is evident from the graph that Avatar, Jurassic Park, Avengers, The Dark night are movies with highest budget and highest revenue generated (see R script 3.1.5).

# Insight 6: Movies with Highest IMDB Rating and Highest Revenue generated after 2000's?

# It is evident from the graph that the dark night rises, The Lord of the rings, Inception are the Highest IMDB Rating and Highest Revenue generated after 2000's (see R script 3.1.6).

# Insight 7: Top 10 prolific directors with highest IMDB Rating and Countries?

# It is evident from the graph that US has the highest number of IMDB movies when it comes to top prolific director that is Cristopher Nolan (see R script 3.1.7).

# Insight 8: Popular content rating with highest IMDB Rating and Revenue Generated?

# It is evident from the graph that PG-13 has the highest IMDB Rating and Revenue generated so director should focus more on PG-13 content rated movies which will be hit on the box office (see R script 3.1.8).

# 