IMDB MOVIE ANALYSIS

NAME : ANIKET ASHOK UBALE

CLASS : B. TECH

YEAR : 3RD

BRANCH : ARTIFICIAL INTELLIGENCE

AND DATA SCIENCE

PROJECT : IMDB MOVIE ANALYSIS

DATE : 07-05-2023

PROJECT DESCRIPTION

- Problem statement: the dataset provided is related to IMDB movies. A potential problem to investigate could be: "what factors influence the success of a movie on IMDB?" Here, success can be defined by high IMDB ratings. The impact of this problem is significant for movie producers, directors, and investors who want to understand what makes a movie successful to make informed decisions in their future projects.
- Data cleaning: this step involves preprocessing the data to make it suitable for analysis. It includes handling missing values, removing duplicates, converting data types if necessary, and possibly feature engineering.
- Data analysis: here, you'll explore the data to understand the relationships between different variables. You might look at the correlation between movie ratings and other factors like genre, director, budget, etc. You might also want to consider the year of release, the actors involved, and other relevant factors.
- Five 'whys' approach: this technique will help you dig deeper into the problem. For instance, if you find that movies with higher budgets tend to have higher ratings, you can ask "why?" Repeatedly to uncover the root cause. Here's an example:
- Q: "why do movies with higher budgets tend to have higher ratings?"
- A: they can afford better production quality.
- Q: "why does better production quality lead to higher ratings?"
- A: it enhances the viewer's experience.
- Q: "why does an enhanced viewer experience lead to higher ratings?"
- A: viewers are more likely to rate a movie highly if they enjoyed watching it.
- Q: "why are viewers more likely to rate a movie highly if they enjoyed watching it?"
- A: positive experiences lead to positive reviews.
- Q: "why do positive reviews matter?"
- A: they influence other viewers' decisions to watch the movie, increasing its popularity and success.

- Data analytics tasks:
- You are required to provide a detailed report for the below data record mentioning the answers of the questions that follows:
- A. Movie genre analysis: analyze the distribution of movie genres and their impact on the IMDB score.
- Task: determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores.
- Hint: use excel's countif function to count the number of movies for each genre. You might need to manipulate the 'genres' column to separate multiple genres for a single movie. Use excel's functions like AVERAGE, MEDIAN, MODE, MAX, MIN, VAR, and STDEV to calculate descriptive statistics. Compare the statistics to understand the impact of genre on movie ratings.
- B. Movie duration analysis: analyze the distribution of movie durations and its impact on the IMDB score.
- Task: analyze the distribution of movie durations and identify the relationship between movie duration and imdb score.
- Hint: calculate descriptive statistics such as mean, median, and standard deviation for movie durations. Use excel's functions like AVERAGE, MEDIAN, and STDEV. Create a scatter plot to visualize the relationship between movie duration and IMDB score. Add a trendline to assess the direction and strength of the relationship.

- C. Language analysis: situation: examine the distribution of movies based on their language.
- Task: determine the most common languages used in movies and analyze their impact on the imdb score using descriptive statistics.
- Hint: use excel's countif function to count the number of movies for each language. Calculate the mean, median, and standard deviation of the IMDB scores for each language. Compare the statistics to understand the impact of language on movie ratings.
- D. Director analysis: influence of directors on movie ratings.
- Task: identify the top directors based on their average imdb score and analyze their contribution to the success of movies using percentile calculations.
- Hint: calculate the average imdb score for each director. Use excel's PERCENTILE function to identify the directors with the highest scores. Compare the scores of these directors to the overall distribution of scores.
- E. Budget analysis: explore the relationship between movie budgets and their financial success.
- Task: analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.
- Hint: calculate the correlation coefficient between movie budgets and gross earnings using excel's correl function. Calculate the profit margin (gross earnings - budget) for each movie and identify the movies with the highest profit margin using excel's max function.
- Remember, these tasks are designed to progressively explore different aspects of the dataset and uncover meaningful insights. Each task builds upon the previous one to provide a comprehensive analysis of the imdb movie data.

IMDB MOVIE ANALYSIS

- The project is all about analysing the dataset having various columns of different IMDB movies like the movie names, released year, actors, budget, gross, genre etc. Using this we need to analyse the profit, top directors, popular movies number of voted users and other such conclusions.
- From the imdb movie data we can derive meaningful information from the dataset.

IMDB MOVIE ANALYSIS

Description-

- This IMDB movie analysis project is all about analyzing a data, do data cleaning draw insights from a data.
- We will perform exploratory data analysis (eda) where we will

  try to checking a dataset , checking outlier ,filling a null values
- For EDA we can use microsoft excel or python labraries (jupyter notebook.
- The imdb movie data having all the information about movie like actor , director , language etc.

APPROACH-

- First we will be performing our analysis on jupyter notebook using various in-built python libraries such as pandas ,numpy, matplotlib,seaborn etc
- We use EDA understanding columns and rows, identifying missing values,handling missing values, checking outliers ,removing outlier.
- I have used jupyter notebook instead of excel or google sheets because i had already knowledge about datasets, python and jupyter notebook tool which i had learn from college.
- Steps for doing project-
- Download the dataset > understanding the data > find duplicate and null > visualization> data insights

SOFTWARE

- Python
- Jupyter notebook
- Microsoft powerpoint

INSIGHTS

- Performing EDA on the dataset is the first step to build a model or derive conclusion from data. We need to remove outlier so that our model gives higher accuracy or answers.
- Statistical analysis plays very important role in procces analytics.
- Performing eda on the data set of imdb movie has helped me to understand the basics steps involved in exploratory data analysis like cleaning of data and deriving inference from the data by performing various statistical analysis.
- Visualization gives more idea about the data. Visualization is easy to undertand rather than value.
- All the questions asked can be answered through jupyter notebook. I used various functions available on python modules to get the solutions.

DATASET-

➢Table features

Color

Director_name

Num_critic_for_reviews

Duration

Director_facebook_likes

Actor_3_facebook_likes

Actor_2_name

Actor_1_facebook_likes

Gross

Genres

actor_1_name

movie_title

num_voted_users

cast_total_facebook_likes

actor_3_name

acenumber_in_poster

plot_keywords

movie_imdb_link

num_user_for_reviews

language

country

content_rating

title_year

actor_2_facebook_likes

imdb_score

aspect_ratio

movie_facebook_likes

YOUR TASK: CLEAN THE DATA

- Import the dataset
- Understood the dataset
- Remove irrelevant data
- Deal with null and missing data
- Fill the missing data
- Filter out data outlier
- Validate the data

YOUR TASK: FIND THE MOVIES WITH THE HIGHEST PROFIT?

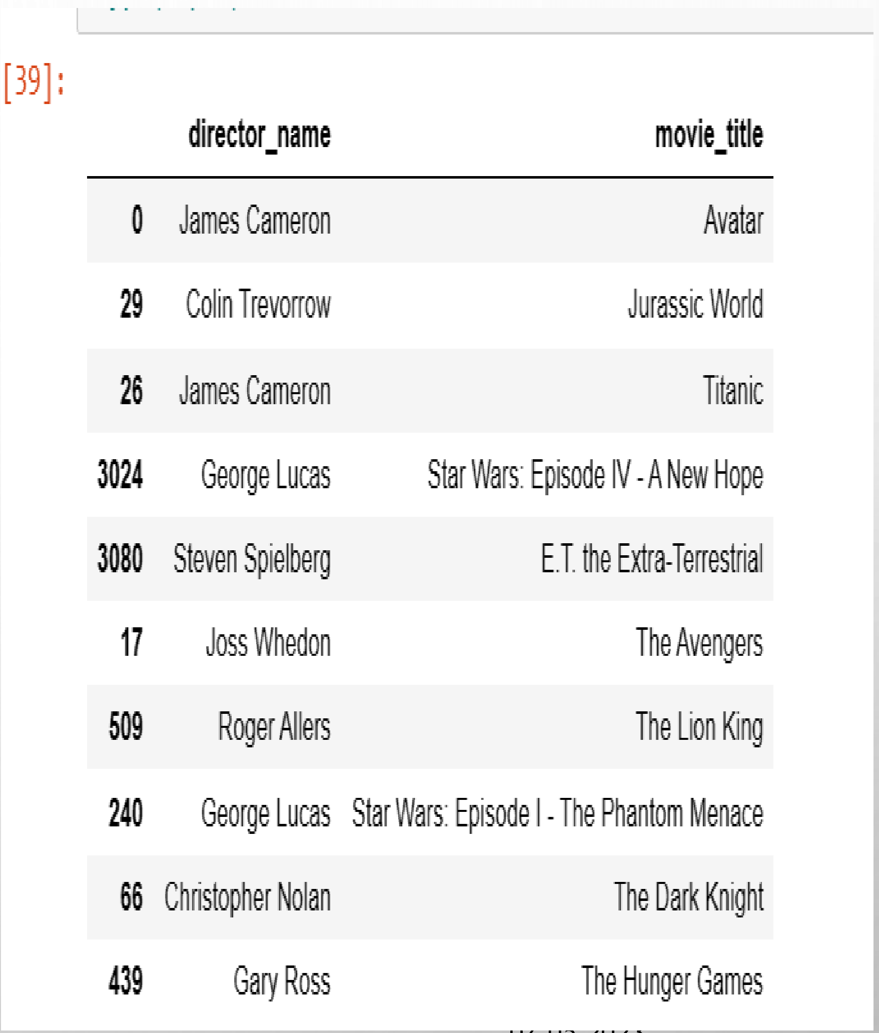
- CODE USED-

Movies['profit']=movies['gross']-movies['budget’]

Movie=movies.Sort_values(by=['profit'],ascending=false)

Top10=movie[['director_name','movie_title']]

Top10.Head(10)



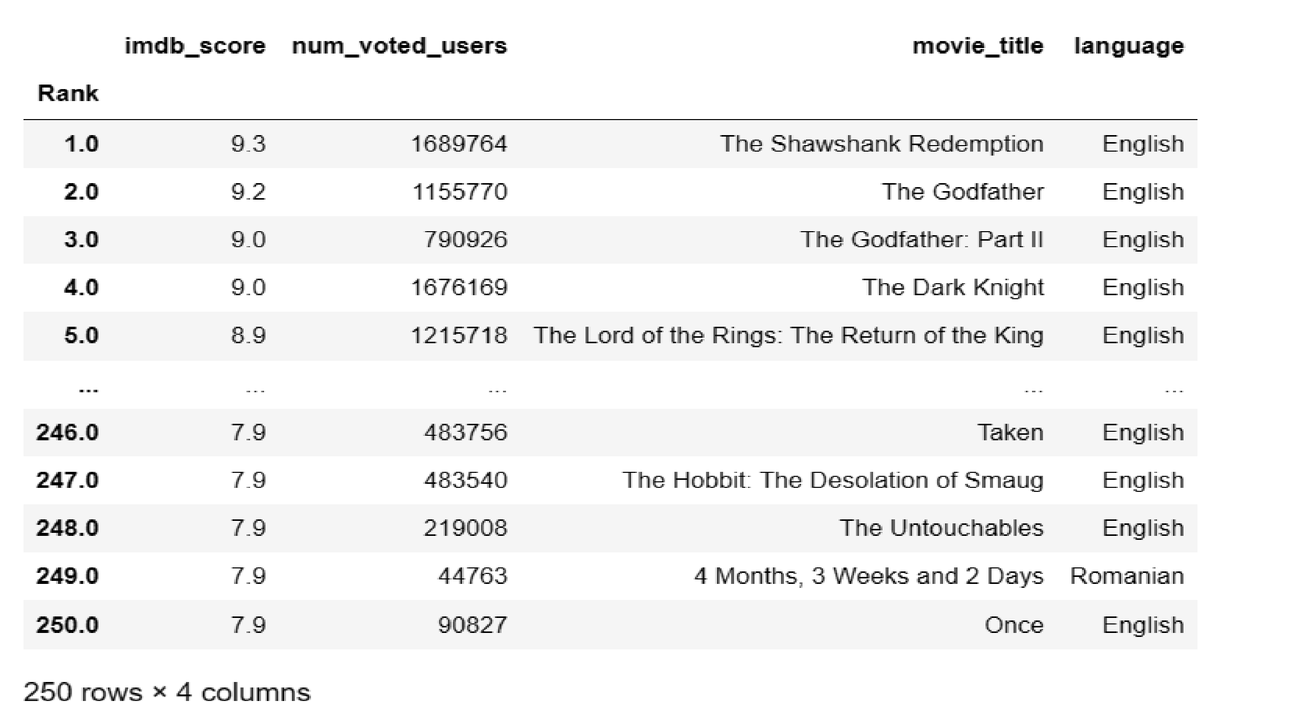
[39]:

	director_name	movie_title
0	James Cameron	Avatar
29	Colin Trevorrow	Jurassic World
26	James Cameron	Titanic
3024	George Lucas	Star Wars: Episode IV - A New Hope
3080	Steven Spielberg	E.T. the Extra-Terrestrial
17	Joss Whedon	The Avengers
509	Roger Allers	The Lion King
240	George Lucas	Star Wars: Episode I - The Phantom Menace
66	Christopher Nolan	The Dark Knight
439	Gary Ross	The Hunger Games

YOUR TASK: FIND IMDB TOP 250

- Code used-

```
Imdb_top_250 = imdb_top_250.Set_index("rank")

Imdb_top_250.Head(250)
```



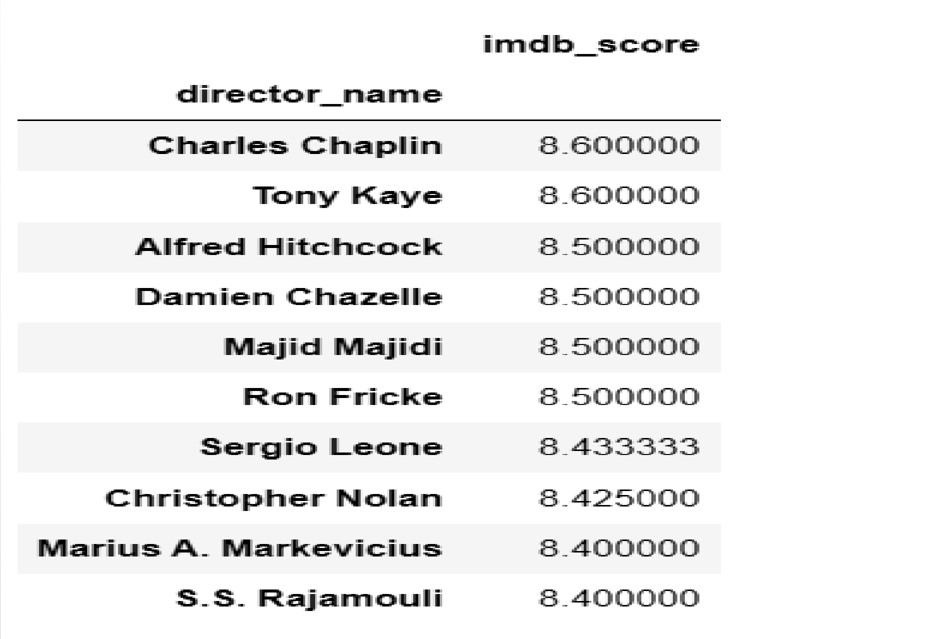
	imdb_score	num_voted_users	movie_title	language
Rank				
1.0	9.3	1689764	The Shawshank Redemption	English
2.0	9.2	1155770	The Godfather	English
3.0	9.0	790926	The Godfather: Part II	English
4.0	9.0	1676169	The Dark Knight	English
5.0	8.9	1215718	The Lord of the Rings: The Return of the King	English
...	...	...	...	...
246.0	7.9	483756	Taken	English
247.0	7.9	483540	The Hobbit: The Desolation of Smaug	English
248.0	7.9	219008	The Untouchables	English
249.0	7.9	44763	4 Months, 3 Weeks and 2 Days	Romanian
250.0	7.9	90827	Once	English

250 rows × 4 columns

YOUR TASK: FIND THE BEST DIRECTORS

- CODE USED-

```
Mov=movies.Groupby('director_name')

Top10director=pd.Dataframe(mov['imdb_score'].Mean().Sort_values(ascending=false))

Top10director=top10director.Head(10)

Top10director=top10director.Sort_values(['imdb_score','director_name'],ascending=(false,true))

Top10director
```

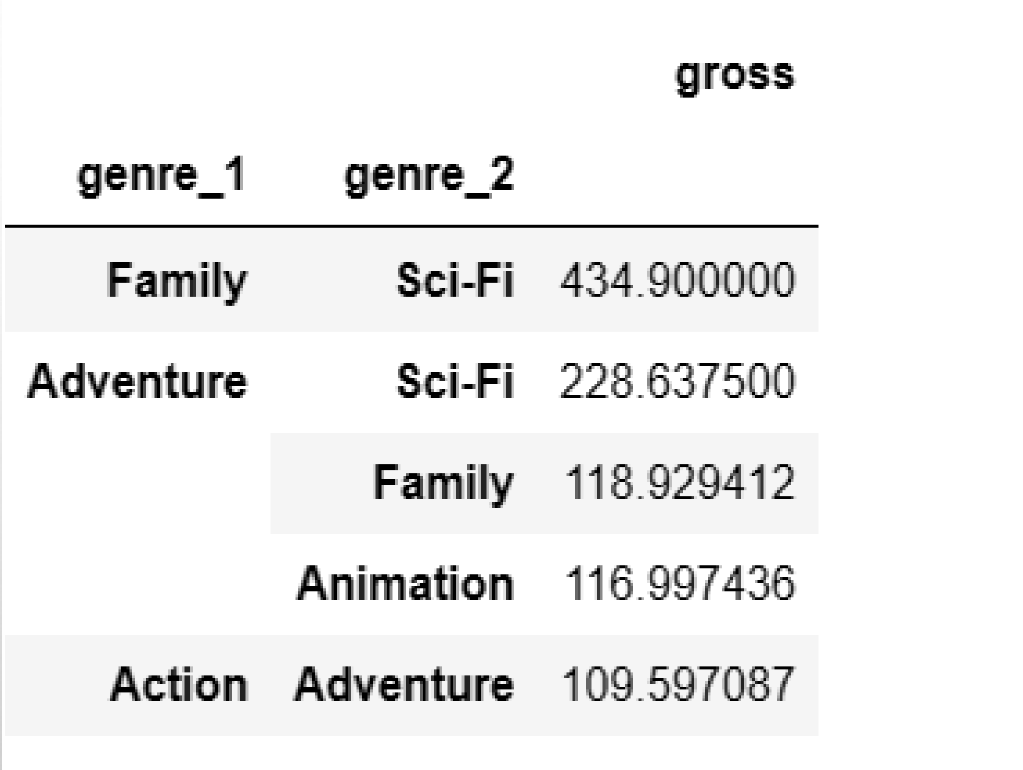


	imdb_score
director_name	
Charles Chaplin	8.600000
Tony Kaye	8.600000
Alfred Hitchcock	8.500000
Damien Chazelle	8.500000
Majid Majidi	8.500000
Ron Fricke	8.500000
Sergio Leone	8.433333
Christopher Nolan	8.425000
Marius A. Markevicius	8.400000
S.S. Rajamouli	8.400000

YOUR TASK: FIND POPULAR GENRES

- CODE USED:-

```
Movies['genres']=movies.Genres.Str.Split('|')

Movies['genre_1']=movies['genres'].Apply(lambda x: x[0])

Movies['genre_2']=movies['genres'].Apply(lambda x: x[1] if len(x)>1 else x[0])

Movies.Head()

..

..

Popgenre=pd.Dataframe(movies_by_segment.Gross.

Mean().Sort_values(ascending=false) )

Popgenre[0:5]
```

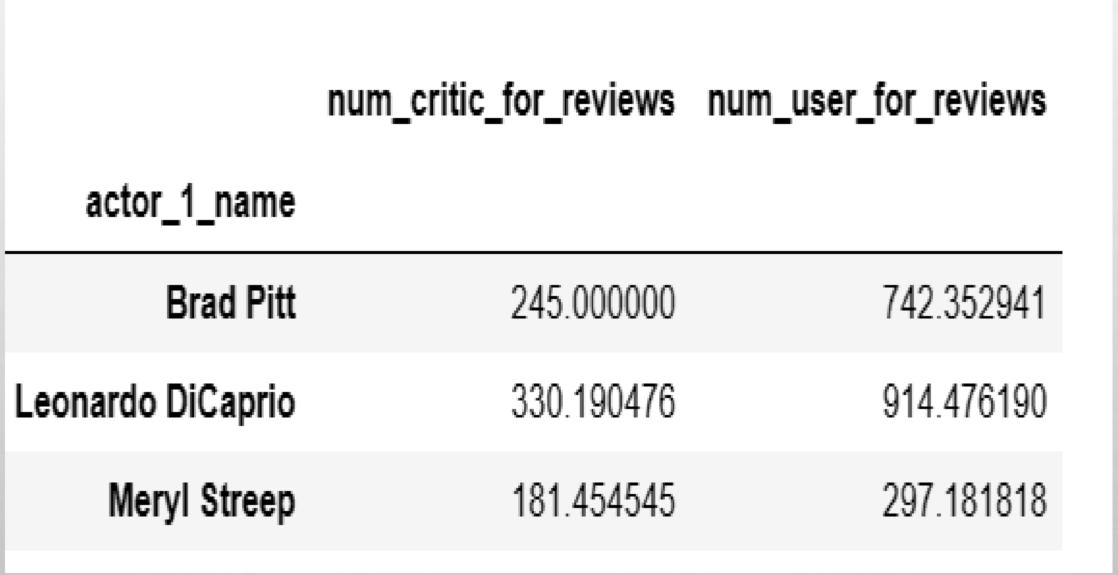


		gross
genre_1	genre_2	
Family	Sci-Fi	434.900000
Adventure	Sci-Fi	228.637500
	Family	118.929412
	Animation	116.997436
Action	Adventure	109.597087

YOUR TASK: FIND THE CRITIC-FAVORITE AN AUDIENCE-FAVORITE ACTORS

- CODE USED:-

```
Combined.Groupby('actor_1_name').Num_user_for_reviews.Mean()

..

..

Combined.Groupby('actor_1_name')[['num_critic_for_reviews','num_user_for_reviews']].Mean()
```



actor_1_name	num_critic_for_reviews	num_user_for_reviews
Brad Pitt	245.000000	742.352941
Leonardo DiCaprio	330.190476	914.476190
Meryl Streep	181.454545	297.181818

RESULT-

- I had the opportunity to deal with real-time datasets. I also had the opportunity to use Python libraries and tools once more. I gained additional knowledge using the Jupyter Notebook.
- I have attempted to draw the necessary graphs, chat in accordance with the requirements and my comprehension, and I have responded to every question in the data set. Gaining practical expertise with real-world data sets and how to clean, modify, visualise, and extrapolate insights from them has been made possible through this project.
- My understanding of this has improved because to the exploratory data analysis section, which comes before more in-depth analytical processing of the data to prepare it for model creation. To ensure that the conclusion generated from the data is a good fit for the subsequent statistical and analytical treatment, we must use EDA to make the data error- and bias-free.
- I realised how to get knowledge from a dataset that is provided to us.How we may extract useful information from the movie data that I have discovered.

Dataset : link

Analysis file : link