

**PYTHON PROJECT**

**ON**

**TMDB MOVIE DATA ANALYSIS**

DONE

BY

BHARATHWAJ

**Description:** Movies that cost over $100 million can still fail, why so? Movie lovers might have different interests.

A production company wants to analyze a movie dataset to identify what kind of movies perform well in cinemas, which genres they belong to, and so on. It will help the company predict if a movie will be a commercial success, if the movie will be highly rated, etc.

**TMDB Movie Data Dictionary**

1. Budget: The budget of a movie in dollars. (A budget value of 0 means the budget value is unknown.)

2. Genres: The genre of the movie and TMDB id (in JSON format)

3. Homepage: The official homepage URL of a movie

4. Id: IMDB id of a movie (string)

5. Keywords: TMDB id and names of all keywords (in JSON format)

6. Original\_language: Two-digit code of the original language (the language in which the movie was made) such as, en for English, fr for French

7. Original\_title: The original title of a movie. The title and original title of a movie may differ if the original title is not in English.

8. Popularity: Popularity of the movie (in float)

9. Overview: Brief description of the movie

10. Production\_companies: All production companies' names and TMDB id of a movie (in JSON format) 11. Production\_countries: Two-digit code and the full name of the production country (in JSON format) 12. Release\_date: The release date of a movie (in dd/mm/yy format)

13. Revenue: The total revenue earned by a movie (in dollars)

14. Runtime: The total runtime of a movie in minutes (integer)

15. Spoken\_languages: Two-digit code and the full name of the spoken language

16. Status: The status of the movie (postproduction, released, or rumored)

17. Vote\_average: Average vote for a movie

18. Vote\_count: The total vote count for a movie

19. Tagline: Tagline of a movie

20. Title: English title of a movie

* **Task 1:** For loading the dataset in the python notebook I have used pd.read\_csv function and to get the information of number of rows and columns I have used info() function, to get the titles and genres of first 50 movies I have used head() function.
* **Task 2:** To find the null values in the columns I have used df.isnull().sum() function so that we can get the count of the nulls in the existing column wise and after identifying the null values I have done null value treatment and replaced the null values the value count in the column should be more than the null values so that we can replace it with mode by using fillna() function.
* **Task 3:** For displaying the movies with budget greater than 220000 I have used dataframe with column name greater than 220000.
* **Task 4:** To display the movies with categories revenue greater than 961000000 I have filtered the revenue with greater than operator so that we can get filtered data from dataset.
* **Task 5:** For removing the values zero, unknown values in the column revenue and budget I have used drop function with column name so that the zeros will be removed from the dataframe.
* **Task 6:** To get the top 10 movies with highest revenue I have first sorted the movies with revenue and used head function to get first 10 movies from the filtered data.Toget the top 10 movies with least budget first I have sorted the values in ascending order and used head function to get first 10 values from the dataset.
* **Task 7:** To know the popularity of the movie budget first I have take popularity and budget columns into a dataframe and plotted scatter plot using these two columns and the data is positively correlated it means when the budget increase the popularity also increases.
* **Task 8:** To identify get the production company names count, I have used value\_counts() function so that we can get the exact count of how many times each production house is appearing in the dataset.
* **Task 9:** From the previous task we can take dataset of count of production companies, first need to sort the data in descending order and use head function after that to get first 25 values from the sorted data.
* **Task 10:** First sorted the data based on revenue and take top 500 movies using head() function. After that perfomed outlier analysis for the three mentioned columns and identified outliers using box plot and replaced it using iqr method. After removing outliers I have found the mean, median, mode of the three columns.
* **Task 11:** For displaying the movie names along with the runtime of those above average runtime filtered the data using greater than average runtime.