**Python Project Documentation**

**Project Agenda :-**

* The aim of this project is to conduct an in-depth analysis of the TMDB dataset, focusing on identifying the types of movies that exhibit strong performance at the box office and determining the specific genres to which they belong.
* The project endeavors to assist the company in making predictions about a movie's potential commercial success and audience ratings, thereby providing valuable insights for decision-making.
* By understanding the varying interests of movie lovers, filmmakers and storytellers can tailor their upcoming movies or develop captivating storylines that align with the preferences and interests of the target audience. This allows them to create content that is more likely to resonate with viewers and attract a loyal following.
* Taking into account the diverse interests of movie enthusiasts, production companies can shape their future film projects or craft engaging narratives that cater to the specific preferences and tastes of the movie-going audience.
* This analysis is valuable in understanding the reasons behind the potential failure of movies with a budget of $100 million despite significant investment.

**Dataset Description :-**

* The TMDB (The Movie Database) is that provides information and data related to movies, TV shows, and other forms of visual entertainment.
* It serves as a valuable resource for movie enthusiasts, professionals in the film industry, and researchers interested in studying various aspects of the cinematic world.
* The dataset contains 4803 records and 20 columns related to movies.
* Important columns in this data set for my analysis are 1. Budget, 2. Title, 3. Revenue 4. Runtime, 5. Genres, 6. popularity.
* I analyzed that
* Understand movie popularity trends,
* Explore genre preferences,
* Investigate budget, runtime and revenue outlier analysis.
* By conducting a thorough analysis of the TMDB dataset, we can gain a deeper understanding of the film industry, identify factors that contribute to movie success, and potentially make informed predictions about future movie trends.

**Interesting Questions in my analysis** **:-**

1. Identify the columns that have null values and perform the null value treatment
2. In the dataset, there are some movies for which the budget and revenue columns have the value 0, which mean unknown values, Remove the rows with value = 0 from both the budget and revenue columns.
3. List the top 10 movies with the highest revenues and the top 10 movies with the least budget.
4. How are popularities of movies related with the movie budgets? Are they correlated or totally uncorrelated with each other? Write the interpretation of your analysis.
5. Display the names of the top 25 production companies based on the number of movies they have produced in descending order of the number of movies produced.
6. Identify and display the names of the movies along with their runtimes for those movies that have above average runtime
7. Perform outlier analysis for the (Budget, Revenue and Runtime) three columns

**Task wise Techniques: -**

|  |  |
| --- | --- |
| **TASKS** | **TECHNIQUES** |
| **Task- 1 :-** Load the data set and Display the number of rows and columns and Display first 50 title and genres in the dataset | * Importing pandas library, Json module * Len() * Apply() * Lambda() * For loop * Head() |
| **Task- 2 :-** Identify the columns that have null values and perform the null value treatment(Choose the imputation method based on the type of data in the columns of interest) | * Info( ) * Fillna( ) * Dropna( ) * Import tabulate, statistics library * Mean( ) * Median ( ) * Mode( ) * Isnull( ) * Sum( ) |
| **Task- 3 :-** Display the movie categories that have budget greater than $220,000  **Task 4:-** Display the movie categories where the revenue is greater that $961,000,000  **Task- 11 :-**  Identify and display the names of the movies along with their runtimes for those movies that have above average runtime, using the data from the previous task. | * Filtering(conditions) |
| **Task- 5 :-** In the dataset, there are some movies for which the budget and revenue columns have the value 0, which mean unknown values, Remove the rows with value = 0 from both the budget and revenue columns | * Filtering (conditions) * Drop() |
| **Task- 6 :-**  List the top 10 movies with the highest revenues and the top 10 movies with the least budget. | * Sort\_values( ) * Head( ) * Barh( ) * Title( ) * Show( ) * Import Matplotlib.pyplot |
| **Task -7 :-** How are popularities of movies related with the movie budgets? Are they correlated or totally uncorrelated with each other? Write the interpretation of your analysis. | * Corr( ) * Scatter( ) * Title( ) * Legend( ) * Show( ) * Import Matplotlib.pyplot |
| **Task- 8 :-** Identify and display the names of all production companies along with the number of times they appear in the dataset | * Apply() * Lambda() * For loop * Explode() * Lambda() * Value\_counts() |
| **Task- 9 :-** Display the names of the top 25 production companies based on the number of movies they have produced in descending order of the number of movies produced. | * Head() |
| **Task- 10 :-**  Sort the data in descending order based on revenue and filter the top 500 movies. Find the measures of central tendency for the following columns using the filtered data:  1. Budget,2. Revenue, 3. Runtime.  Perform outlier analysis for the (Budget, Revenue and Runtime) three columns using box plots. | * Import tabulate, statistics library * Insert( ) * Mean( ) * Median ( ) * Mode( ) * Sort\_values( ) * Head( ) * Boxplot( ) * Set\_title( ) * Sub\_plots( ) * Loc[ ] |

**Explanation of above techniques: -**

|  |  |
| --- | --- |
| **Techniques** | **Explanation** |
| 1. Import panda’s library | It is used for data manipulation, analysis, and cleaning tasks, in data frame. |
| 1. Import Json module | It is used for encoding and decoding JSON data, facilitating the interchange of data between different systems |
| 1. Import tabulate(m) | It is used for creating formatted tables from tabular data. |
| 1. Import statistics(m) | It is used for mathematical statistics calculations, such as mean, median, mode, variance, and standard deviation. |
| 1. Import Matplotlib.pyplot | It is used for creating visualizations and plots. |
| 1. Import seaborn | It is used for creating statistical data visualizations. It provides a high-level interface for drawing attractive and informative statistical graphics. |
| 1. Len()(f) | It is used to get the length or number of elements in an object. |
| 1. Apply() | It is used to apply a function to each element or row of a pandas Data Frame or Series. |
| 1. Lambda() | The `lambda` keyword in Python creates anonymous functions, allowing for the creation of small, inline functions without a formal function definition. |
| 1. For loop | It is used to iterate over a sequence |
| 1. Head() | It is used to retrieve the first few rows of data, typically the top five rows by default. |
| 1. Info() | It provides a concise summary of the Data Frame, including information about the column names, data types, number of non-null values, and memory usage. |
| 1. Fillna() | It is used to fill missing or NaN (null) values in a pandas DataFrame or Series |
| 1. Dropna() | It is used to remove rows or columns containing missing or NaN (null) values from a pandas DataFrame |
| 1. Drop() | It removes specified rows or columns from a pandas DataFrame |
| 1. Mean() | It is used to calculate the arithmetic mean or average of a set of numerical values. |
| 1. Median() | It is used to calculate the median value of a set of numerical values. The median is the middle value when the data is sorted in ascending or descending order, or the average of the two middle values if the data has an even number of elements. |
| 1. Mode() | It is used to calculate the mode value of a set of values. The mode represents the most frequently occurring value in the dataset. |
| 1. Isnull() | It checks for null values in a dataset and returns a Boolean mask |
| 1. Sum() | It calculates the sum of numeric values in a sequence or along a specified |
| 1. Filtering | Filtering the dataset according to conditions. |
| 1. Sort\_values() | It sorts the elements or rows of a DataFrame or Series based on specified columns |
| 1. Barh() | It is used to create horizontal bar charts or bar plots in the matplotlib.pyplot module |
| 1. Title() | is used to set the title of a plot or figure in the matplotlib.pyplot module |
| 1. Show() | is used to display the currently active figure or plot in the matplotlib.pyplot module |
| 1. Corr() | It is used to compute the correlation between columns or variables in a DataFrame |
| 1. Scatter() | It creates scatter plots to visualize the relationship between two numerical variables in two lines of code using matplotlib.pyplot. |
| 1. Legend() | It adds a legend to a plot in a single line. |
| 1. Explode() | refers to a method that expands a column containing iterable elements into multiple rows, duplicating the other column values accordingly. It is useful when dealing with nested or hierarchical data structures within a DataFrame |
| 1. Values\_count() | It counts the occurrences of unique values in a column or Series |
| 1. Insert() | It inserts a new element at a specified index position within a list |
| 1. Boxplot() | is used to create box plots, which display the distribution of a dataset using quartiles and potential outliers in the matplotlib.pyplot |
| 1. Set\_title() | In the `matplotlib.pyplot` module, is used to set the title of a plot or figure. |
| 1. Sub\_plots() | It allows for the arrangement of multiple plots or axes in a single figure, providing a convenient way to display multiple visualizations together. |
| 1. Loc[] | It is used to access or modify values in a DataFrame or Series based on label-based indexing. |

**Conclusion :-**

* It is important to acknowledge the limitations of the TMDB dataset, such as unknown values missing or inaccurate information. These limitations might have influenced the outcomes of our analysis. It is recommended to ensure the TMDB dataset is consistently refreshed with the most up-to-date information.(Task 2,5)
* Action, Adventure, and Science fiction genre movies tend to generate high revenue, it is advisable to prioritize the production of films belonging to these genres as they have a higher likelihood of achieving success.(Task 4)
* Some movies with relatively lower budgets have managed to generate significant revenue and achieve financial success(task 6 )
* Considering that Warner Bros and Universal Pictures are the production companies with the highest frequency of movie releases, it is evident that these companies exhibit a strong track record and consistently engage in film production. This observation suggests that there is a positive trend in their performance, indicating that filmmakers may find it beneficial to collaborate with or seek opportunities from these reputable production companies.(task 9)
* The availability of 234 records with a greater than average runtime, In future if they make high runtime movies, there are a few suggestions, considering these things will helpful(task 11)
  + - Review and edit longer movies to streamline the narrative.
    - Analyze the script for areas to tighten without compromising the story.
    - Consider the target audience's attention span and preferences.
    - Gather feedback through test screenings and create alternative versions if necessary.
* Within the action and adventure genres, explore unique themes or settings that differentiate your movies from others. This can help create a distinct identity and attract audience attention.
* The popularity of action and adventure genres, consider developing and producing movies that align with these genres. This can potentially attract a larger audience and increase the chances of commercial success