Title: data science project

Author:

1-Mohammed Emad

2-Fares Saleh

3-Mohammed Ali

4-Nour El Din

5-Abdelrahman Mahmoud

6-Abdelrahman Kamal

The dataset:

The TMDB Movies dataset contains information about various movies, including their genres, release dates, budgets, revenues, and ratings. The goal is to explore this dataset and extract meaningful information to understand trends, patterns, and factors influencing the success of movies.

Significance and Relevance:

Understanding the TMDB Movies dataset is crucial for several stakeholders in the movie industry. By analyzing this dataset, we can uncover valuable insights that can aid filmmakers, production companies, distributors, and movie enthusiasts in making informed decisions.

Problem Definition:

the problem may vary depending on the specific analysis and goals.

Methods:

1. Data Preprocessing:

a. Handling Missing Values: handled using techniques such as imputation (mode) or deletion depending on the nature of the missing data. so in column “Imbd” & “ overview “ we drop the missing value and in “cast“, “tagline“, “director” , “keywords”,”generes” &”production” we fill the null value with the most-frequent value in column

b. Data Cleaning: The dataset is cleaned by removing duplicate entries, check the column type by print the unique value and handle it if needed , drop unnecessary column (“homepage”)

1. Analysis:

a. Exploratory Data Analysis (EDA): EDA is conducted to understand the distribution, relationships, and summary statistics of the variables in the dataset. we checked for the correlation ,we found strong positive correlation between popularity and vote count.

b. Descriptive Statistics: Key statistical measures such as mean, median, standard deviation, and percentiles are calculated to summarize the dataset and gain insights into its central tendencies and variability. By print dataset.describe()

3. Visualization:

Data Visualization: Visual representations such as bar charts, histograms, scatter plots, and heatmaps are created to effectively communicate patterns, trends, and relationships within the dataset.

using Python libraries like Matplotlib, Seaborn, and Plotly.

1. Machine Learning:

Create machine learning model ( Recommendation System Model ) to recommend movies according to your mood.

1. Dashboard power bi:

We did the necessary visualization

Experiment:

1. Dataset Acquisition:

The TMDB Movies dataset is obtained from a reliable source.

1. Data Preprocessing:

a. Python Libraries/Modules: Pandas and NumPy are used for data manipulation and preprocessing tasks. Pandas provides powerful data structures and functions for handling data, while NumPy offers efficient numerical computations.

b. Data Cleaning: missing values, and data quality issues are addressed using Pandas and NumPy functions, such as dropna(), fillna().

1. Analysis and Visualization:

a. Python Libraries/Modules: Matplotlib, Seaborn, and Plotly are used for data visualization. Matplotlib provides a wide range of plotting options, Seaborn offers enhanced statistical visualizations, and Plotly allows interactive and dynamic visualizations.

1. Machine Learning:

a. Python Libraries/Modules: sklearn.feature & sklearn.metrics are essential components of scikit-learn, allowing users to preprocess their data and evaluate the performance of their machine learning models.

1. Dashboard Creation:

a. Power BI: Power BI is used to create a dashboard for visualizing and presenting the insights gained from the TMDB Movies dataset.