##Project Report ##

**Data set Description**

Dataset is of multiple movies and related columns to that movies

It is divided into two sections :

smaller dataset (10k rows)   
id,original\_language,original\_title,release\_date,vote\_average,vote\_count,genre,overview,revenue,runtime

and

larger dataset (1Million rows)

Id, title, genre, userId, rating, gender, age

Columns are self-explanatory and others are :

Avg\_voting is out of 10.

Revenue is a movie’s collection.

Rating is out of 5\*

Users and movies have different IDs.

# data quality/integrity/ethics issues

* The larger dataset has one to many relationship between the columns like :  
  single movie with its genre and id (1->many) rating, age, gender i.e.one movie was rated by multiple users of any gender and different ratings
* Incomplete Genre Data: The 'genres' column has only 6020 non-null values out of a total of 1000209 rows. This means that genre information is missing for a large portion of the movies in the dataset. This could limit the usefulness of the dataset for certain types of analysis or recommendations.
* Biased User Demographics: The dataset only includes information on gender and age of the users, and it is not clear how representative the sample is of the general population. The dataset is not be inclusive of all types of users, which could lead to biased recommendations or conclusions based on the data.
* Limited Rating Scale: The rating column only ranges from 1 to 5, which could limit the granularity of the analysis and recommendations that can be made from the data.
* Potential for Fraudulent Ratings: There is no information on how the ratings were obtained from the users, which could leave the dataset open to potential fraudulent ratings or other data quality issues.
* Privacy Concerns: The dataset contains personal information such as gender and age, which could raise privacy concerns if the data is not handled appropriately.
* Lack of Contextual Information: The dataset does not provide any contextual information about the movies, such as release date or production budget, which could limit the analysis and insights that can be drawn from the data.
* While in the smaller data there are missing values: There are missing values in almost all columns, with the number of non-null values ranging from 5336 to 10014. This can be problematic for data analysis since missing data can bias results and decrease the representativeness of the dataset.
* Inconsistent data types: Some columns contain data of inconsistent data types. For example, The "release\_date" column is also of type "object" while it will be more appropriate to represent it as a datetime object.
* Data integrity issues: There is a data integrity issues in the dataset, such inconsistencies. For example, Unicode formatting utf-8 is not suitable for many different languages hence data is likely to have movie title as garbage column
* Bias: The dataset may be biased towards certain genres or languages, which can impact the representativeness of the dataset and bias any analysis or modeling based on it.
* Privacy and ethics: The dataset may contain personal information about individuals involved in the movies, such as the actors or crew, that may need to be protected or anonymized to maintain privacy and ethical considerations.
* In summary, the dataset has missing values, inconsistent data, data integrity issues, and potential biases and privacy/ethical considerations. These issues need to be addressed and carefully considered in any data analysis or modeling based on this dataset.

Data preparation/wrangling

#Data-wrangling and Data preparation operations :

* Merging the datasets
* Groupby() & Merge
* Simple reading csv files and identifying the types of data
* Using median and mode in the data columns to fill the null values
* Categorizing using data slice and dice
* Dealing with outliers

*\*All outcomes at the end of the report\**

Data mining fundamentals **(25 points)**

#Data Mining technique used:

* Linear Regression
* Logistic Regression
* Support Vector Machine to find SVM score.
* Correlation and Heatmap to relate different parameters of data.
* KNN for clustering analysis and Silhouette’s coefficient

*\*All outcomes at the end of the report*

Data Visualization:

Please find all the data visualization techniques performed as per the relevance of the data:

(Loading csv)

Text

Description automatically generated

1. Boxplot

Chart

Description automatically generated

Chart

Description automatically generated

**Chart, box and whisker chart

Description automatically generated**

1. Histplot

**Chart

Description automatically generated**

**Chart, histogram

Description automatically generated**

1. **Countplot**

**Chart, bar chart

Description automatically generated**

1. **Scatterplot**

**Chart

Description automatically generated with medium confidence**

**Chart

Description automatically generated**

1. **Piechart**

**Chart, pie chart

Description automatically generated**

1. **Barchart**

**Text

Description automatically generated**

**Chart, bar chart

Description automatically generated**

1. **Density distribution**

**Graphical user interface, text, application

Description automatically generated**

**Chart

Description automatically generated with medium confidence**

1. **Pairplots**

**Chart

Description automatically generated with low confidence**

**Chart

Description automatically generated with medium confidence**

# **Outcomes**

Dealing with outlier :

Text

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, histogram

Description automatically generated

Basic data wrangling operations and their outcomes

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Chart, histogram, box and whisker chart

Description automatically generated

Chart, histogram

Description automatically generated

Categorizing using slicing

Text

Description automatically generated

Visulization using K-means clustering

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

Logistic Regression

Text

Description automatically generated

Text

Description automatically generated

Text, letter

Description automatically generated

As we can see it has only 49% score so Logistic regression model has an accuracy score of around 50%, which means it is not a good model for predicting gender based on movie ratings, vote counts, revenue, and runtime. It also shows that splitting the dataset into training and testing sets did not improve the accuracy of the model.

Linear Regression

Text

Description automatically generated

Text

Description automatically generated

Text, letter

Description automatically generated

As we can see Linear regression has 9% score only and the model may not capture the relationship between the features and the target variable accurately hence its not good method to use for our dataset

**Support Vector Machine**

Text

Description automatically generated

As we can see the SVR model performs poorly with a negative R-squared score, which indicates that the model does not fit the data well. On the other hand, the SVC model performs well with an accuracy score of approximately 97%. However, when looking at the confusion matrix, we can see that the model is predicting only the low revenue category, with all the high revenue movies being classified as low revenue. Hence not good fit for the dataset

K Means Clustering :

Text

Description automatically generated

Graphical user interface, text

Description automatically generated

Result indicates that the Silhouette Coefficient for 9 clusters is 0.7959, which is quite high and suggests that the clustering can be well-defined and the movies are grouped together in a meaningful way based on their features. This could be useful for further analysis or for making recommendations based on the features of the movies in each cluster.

Here is one example of clustering with k=9

Chart, scatter chart

Description automatically generated