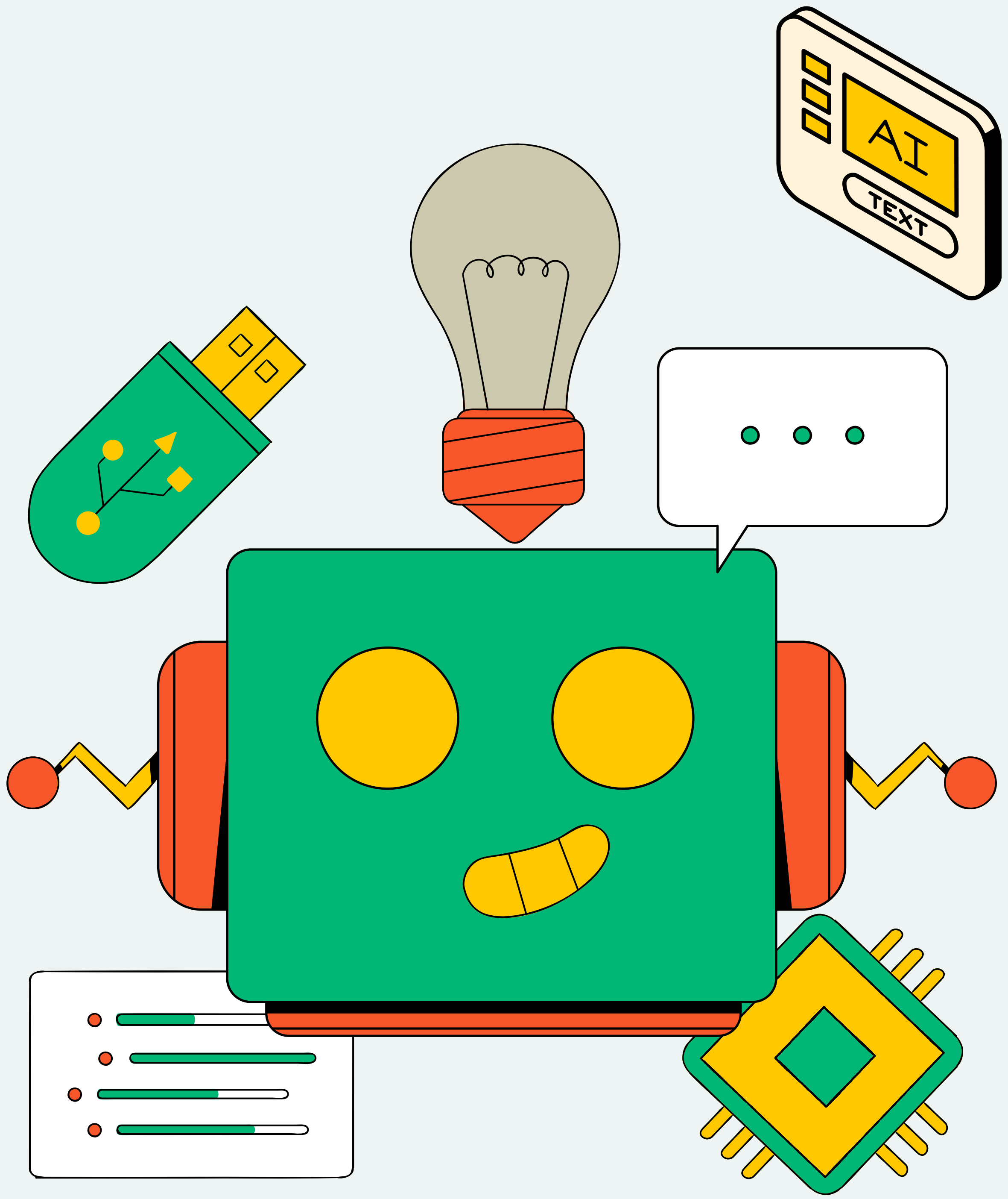
CSE655-NETWORK SCIENCE

MOVIE RATING AND GENRE PREDICTION USING NETWORK SCIENCE ASSISTED BY MACHINE LEARNING



PRESENTED BY: HIMANSHU CHOUDHARY
ANIKET KANOJIA
DIVYAM KHORWAL
KARTIK BANSAL

MOTIVATION

LIMITED DIVERSITY

- We want to create a movie recommendation system that stems from our love for movie and the desire to continually explore new actors and genre.

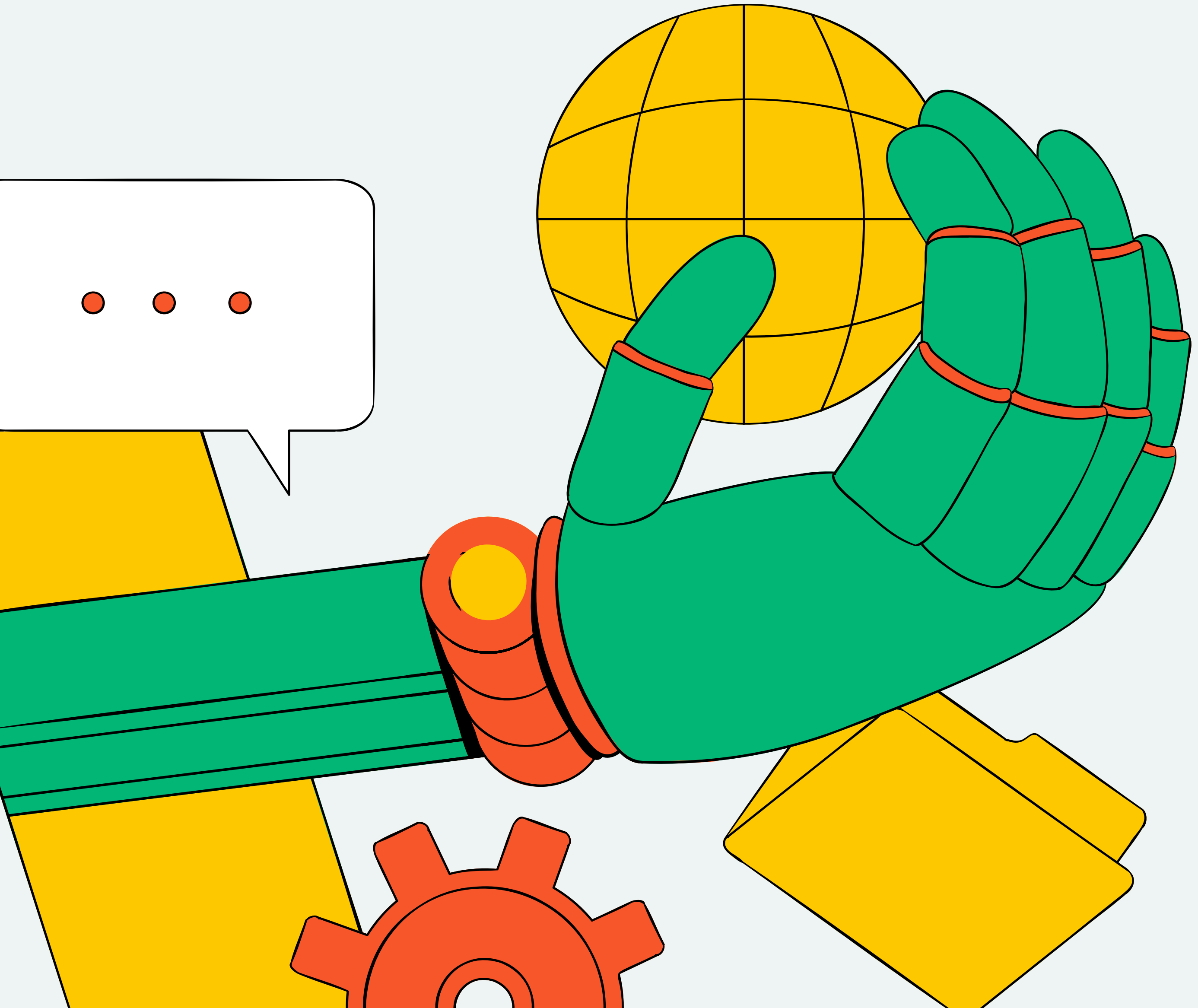
LEARNING

- Building a movie recommendation system helped us build a thorough understanding of various machine learning algorithms and feature engineering techniques.

BIAS

- Existing systems are biased towards popular content and mostly keep revolving with same set of movies.

DATASET DESCRIPTION

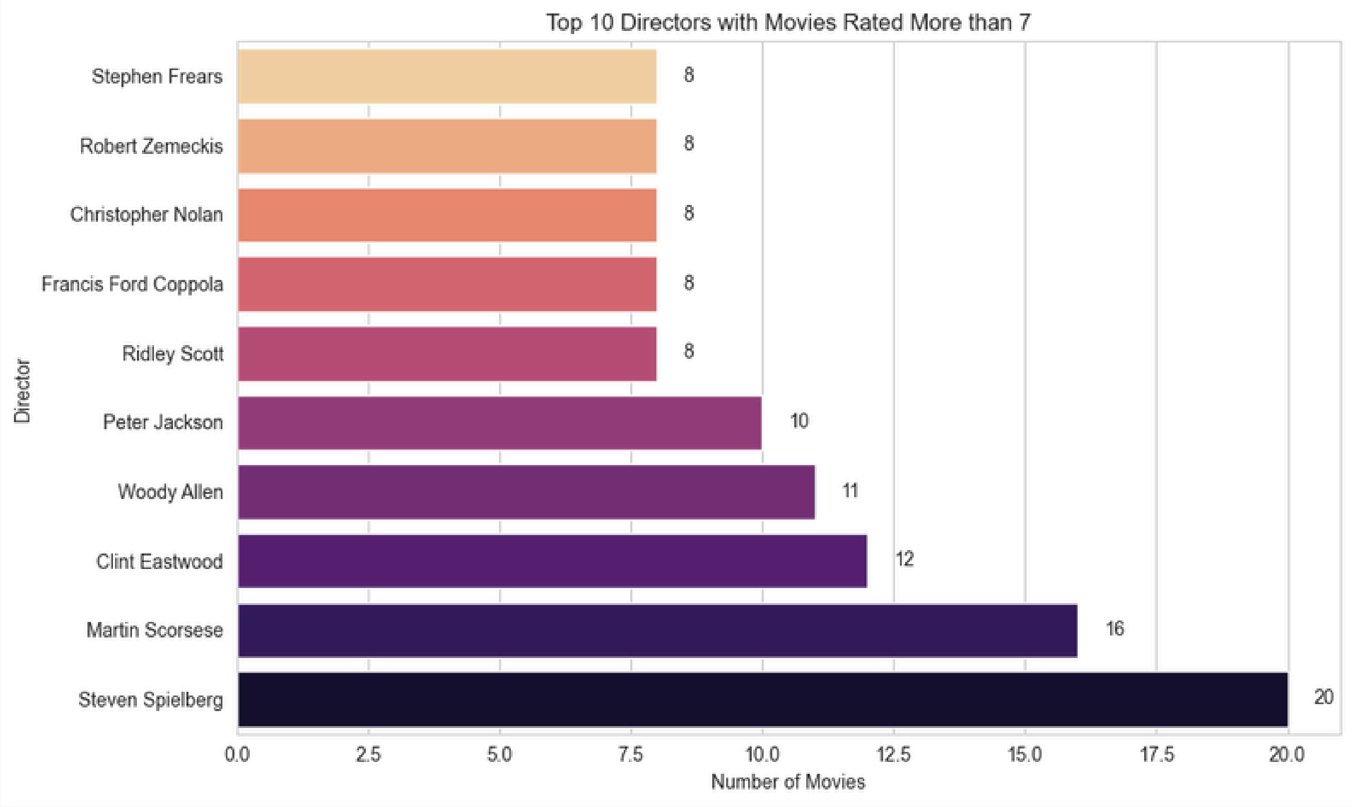


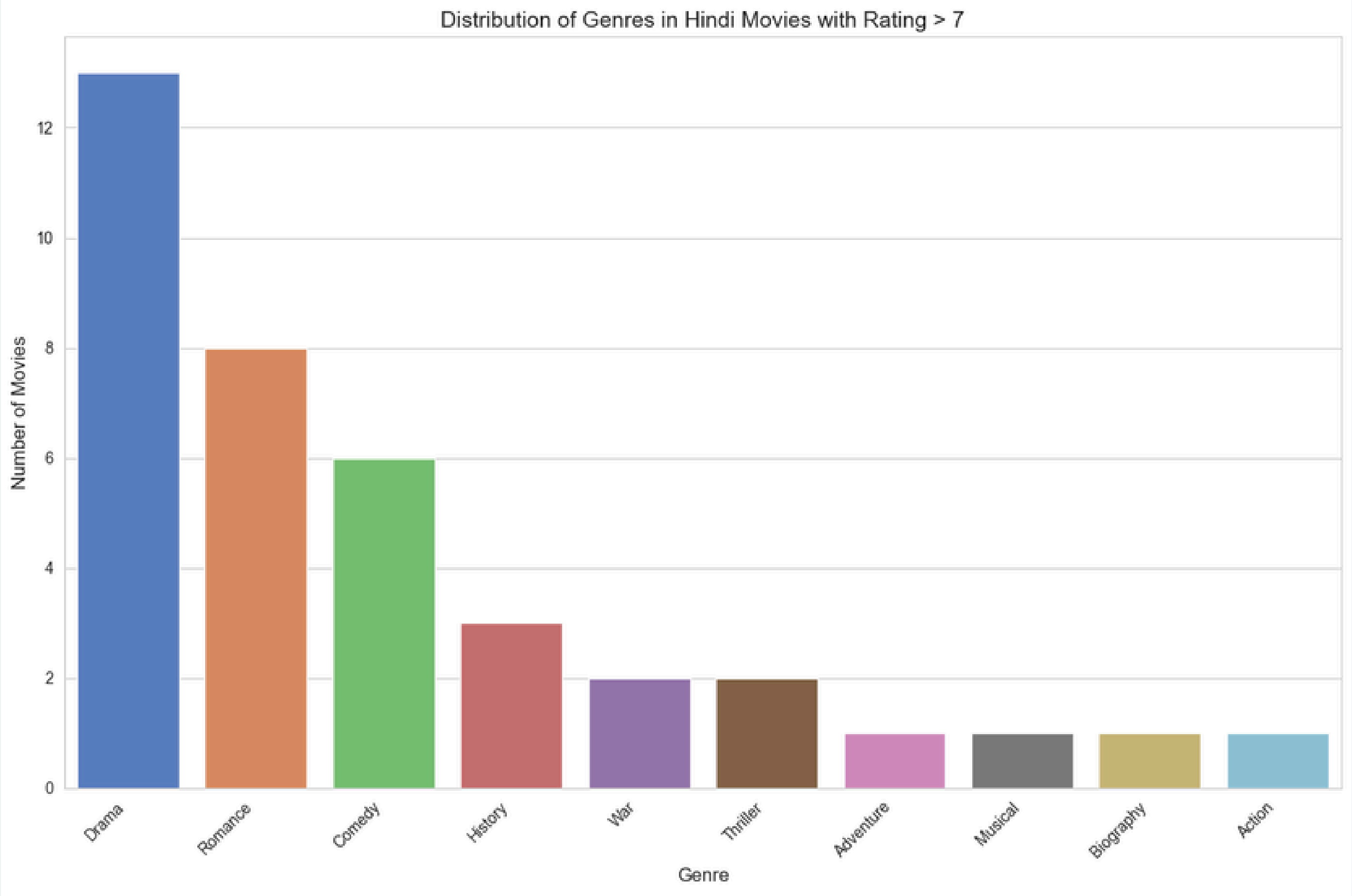
We have used 2 datasets for Movie rating prediction and genre prediction.

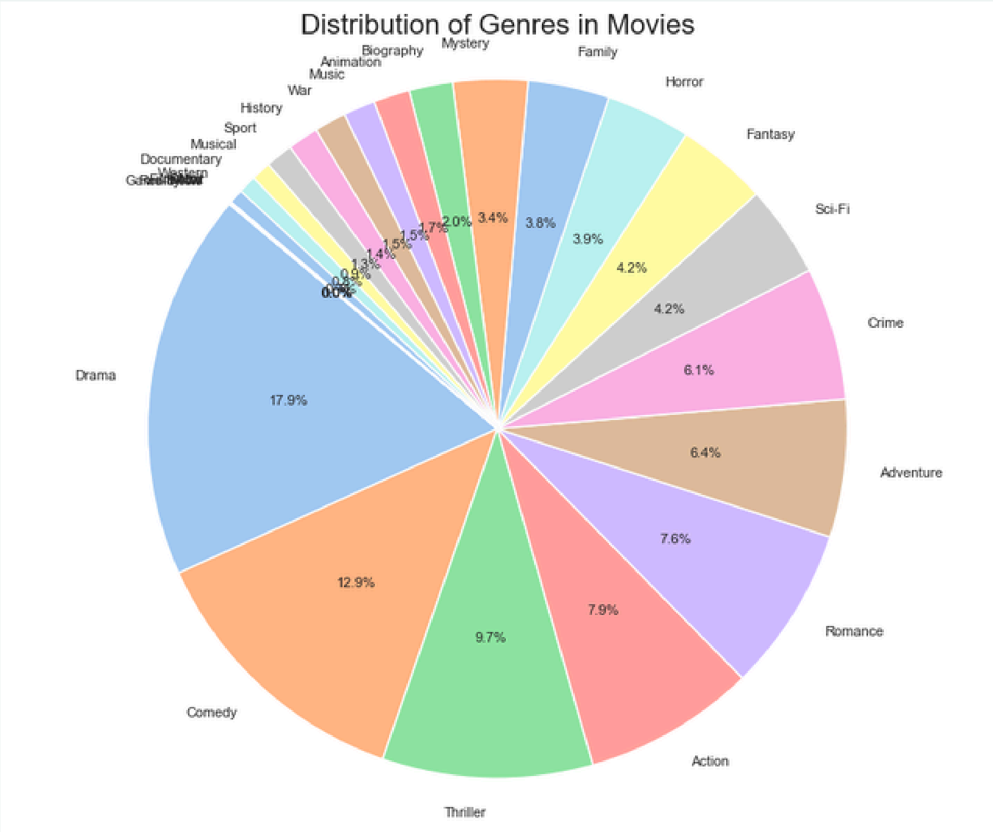
Dataset 1 -:
Link-: https://github.com/yash91sharma/IMDB-Movie-Dataset-Analysis/blob/master/movie_metadata.csv
Size-: 5043
Parameters-: 28
Parameter names-:movie_id,movie_name,rating,plot,etc

Dataset 2-:
link-: https://www.cs.cmu.edu/~ark/personas/
Size-: 42000
Parameters-: 9
Parameter names-:**movie_id,movie_name,genre,plot,etc**

EXPLORATORY DATA ANALYSIS

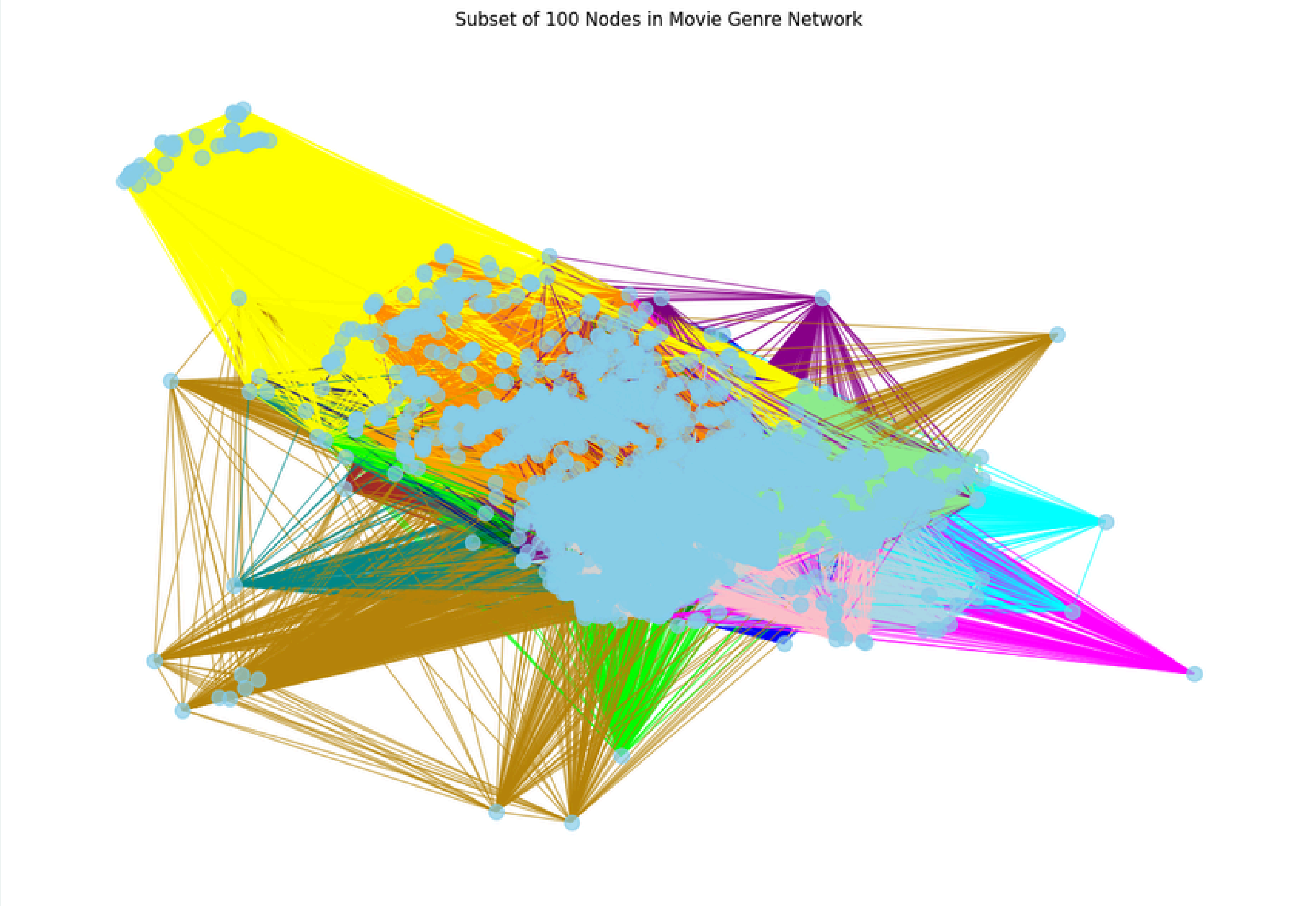


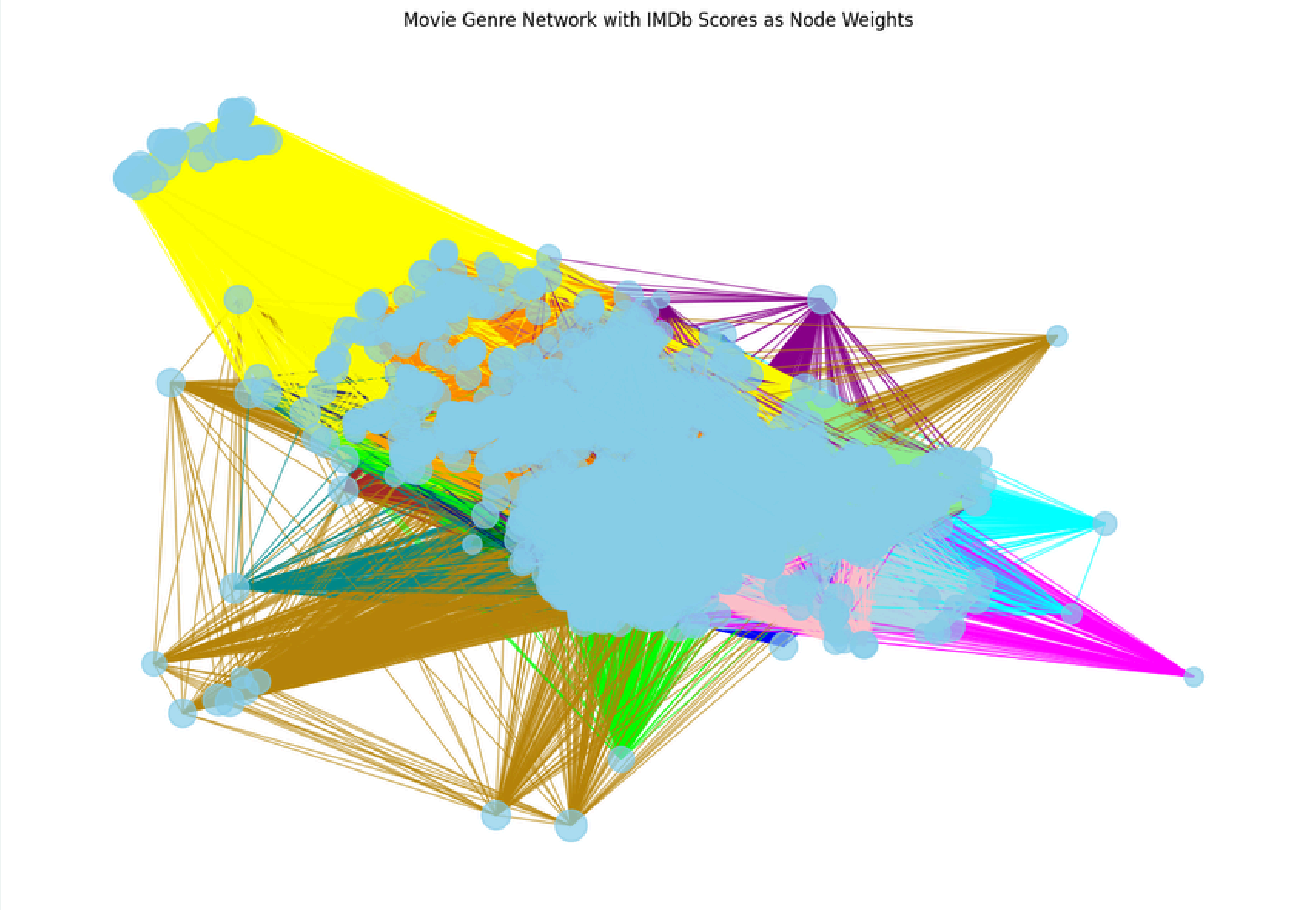




Data in graph shows(Rating) -:

1. Directors whose movies are hit along with number of their hit movies.
2. In pie plot we can see drama ,thriller and comedy movies that are more made in world.
3. Our histogram shows most common genre in movies with hindi language and we can see drama is mostly watched in India.





NUMBER OF NODES: 5043

NUMBER OF EDGES: 6665750

Average degree: 2643.57

AVERAGE SHORTEST PATH LENGTH: 1.08

Density of the Graph: 0.55

THE GRAPH IS NOT WEIGHTED.(GRAPH-1)

THE NODES OF THE GRAPH ARE WEIGHTED.(GRAPH-2)

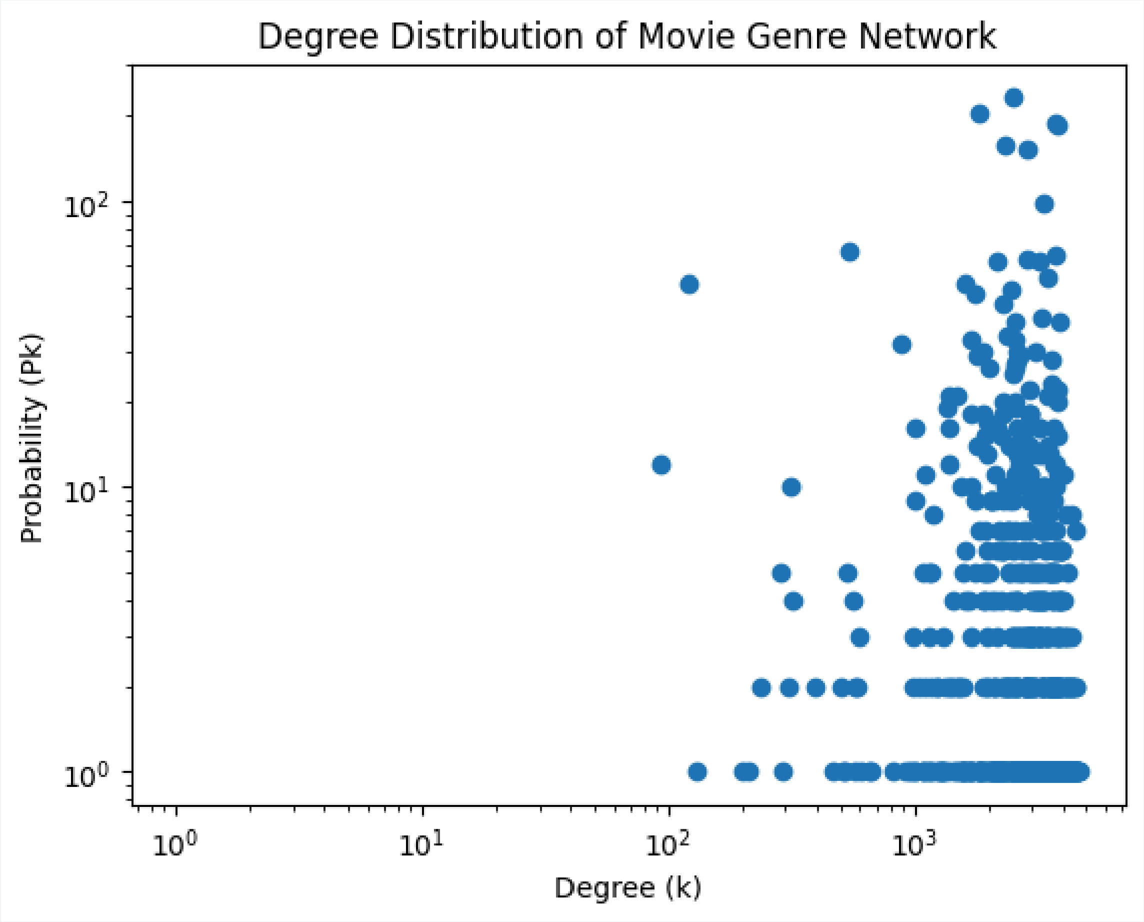
THE GRAPH IS NOT BIPARTITE.

IS CONNECTED: TRUE

Data in graph shows(Rating) -:

1.Edge color shows 2 movies have same genre thus forming clusturs of multiple genre can be seen forming a giant component/network

2. In this nodes can be seen of different sizes dur to their weight which is movie rating.



Node with Highest degree: 4685

DEGREE CENTRALITY OF SOME NODES

AVATAR : DEGREE CENTRALITY = 0.4050

PIRATES OF THE CARIBBEAN: AT WORLD'S END : DEGREE CENTRALITY = 0.3668
SPECTRE : DEGREE CENTRALITY = 0.4668

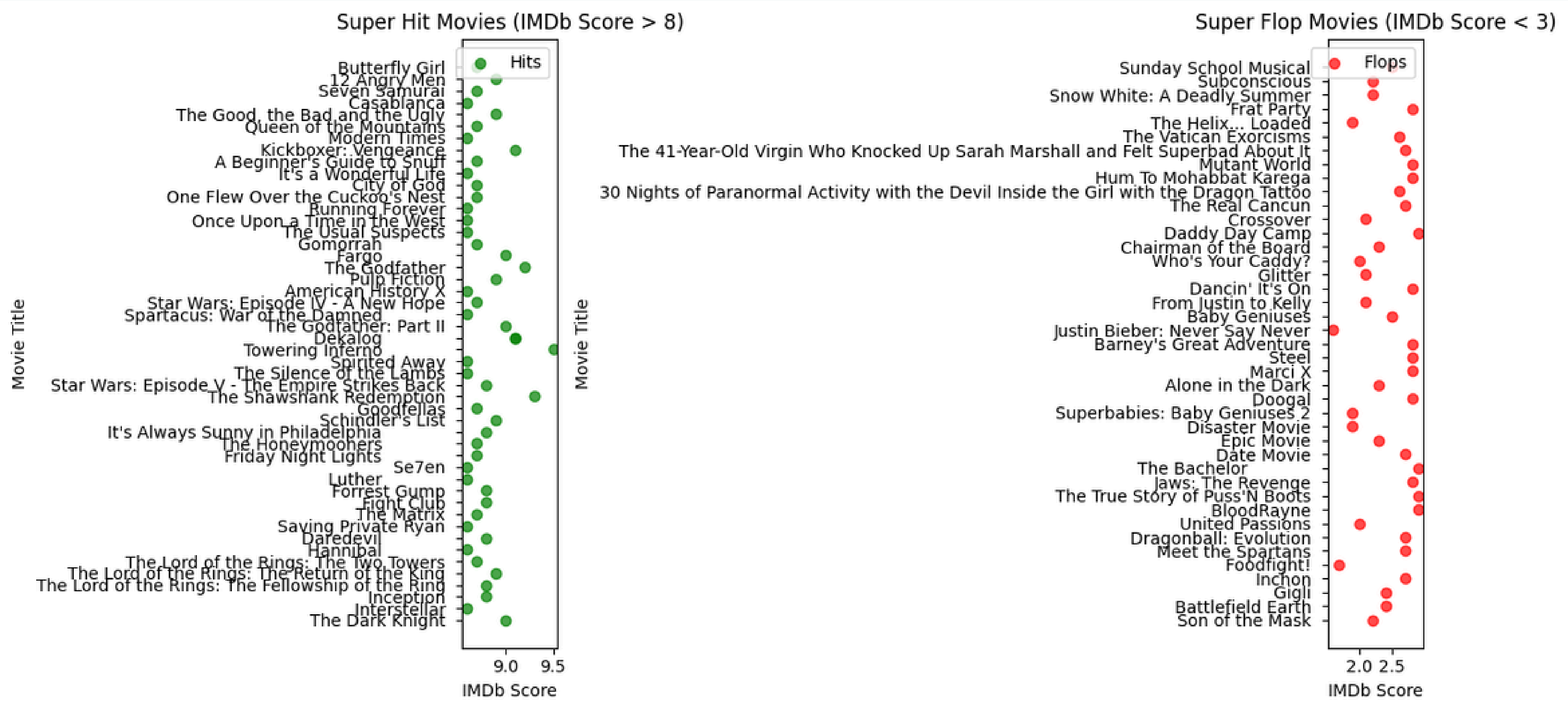
THE DARK KNIGHT RISES : DEGREE CENTRALITY = 0.3875

STAR WARS: EPISODE VII – THE FORCE AWAKENS : DEGREE CENTRALITY = 0.0244

JOHN CARTER : DEGREE CENTRALITY = 0.3572

MODULARITY: 0.19

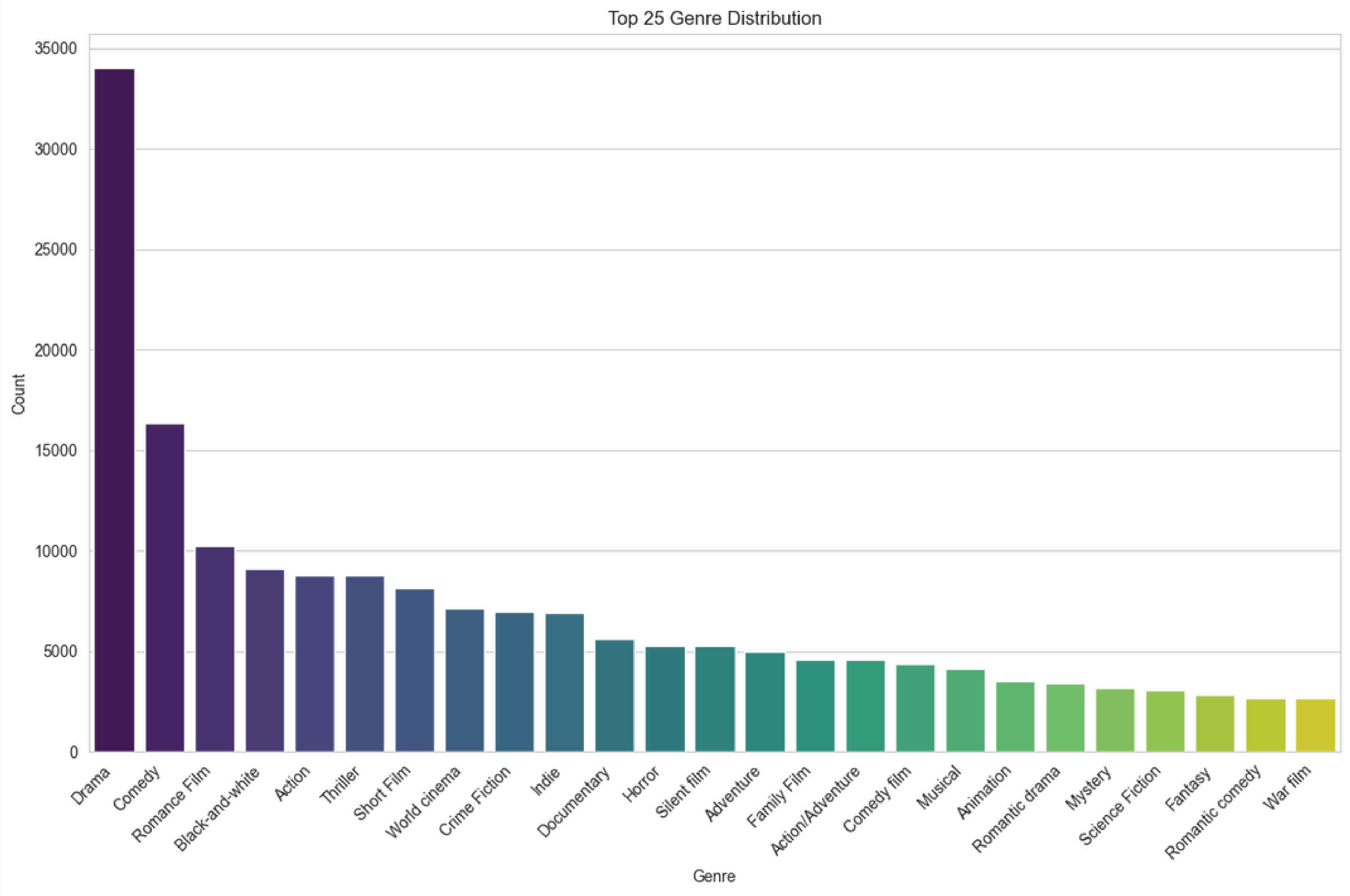
AVERAGE CLUSTERING COEFFICIENT: 0.85



INTERPRETATION:

- **IN THE CONTEXT OF MOVIE GENRES, THIS DISTRIBUTION SUGGESTS THAT MOST GENRES ARE LESS INTERCONNECTED COMPARED TO A FEW POPULAR GENRES THAT ARE HIGHLY CONNECTED.**
- **FOR EXAMPLE, GENRES LIKE “ACTION” OR “DRAMA” MIGHT HAVE MANY CONNECTIONS (HIGH DEGREE), WHILE NICHE GENRES LIKE “FILM NOIR” OR “EXPERIMENTAL” MAY HAVE FEWER CONNECTIONS (LOW DEGREE).**

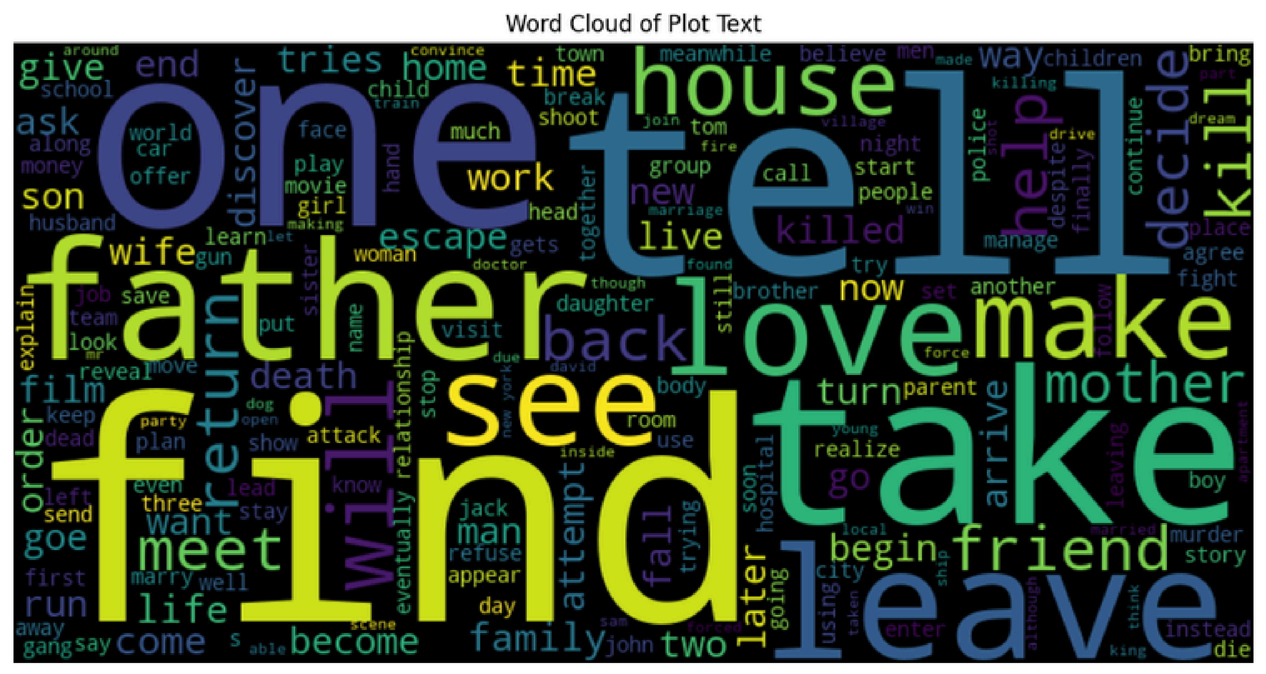
EXPLORATORY DATA ANALYSIS

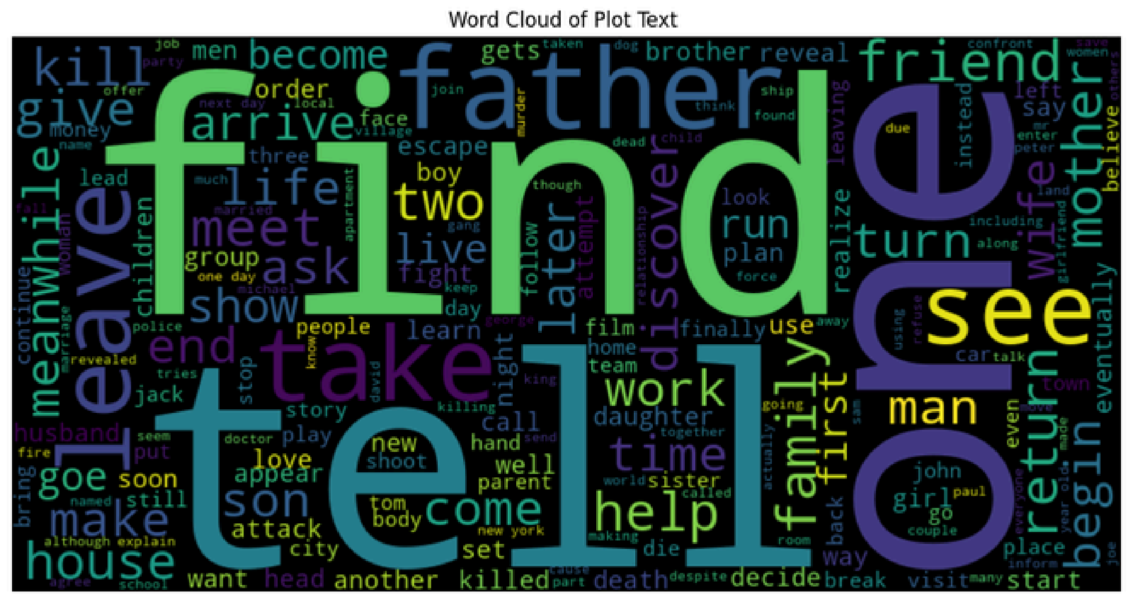


Data in graph shows(Genre) -:

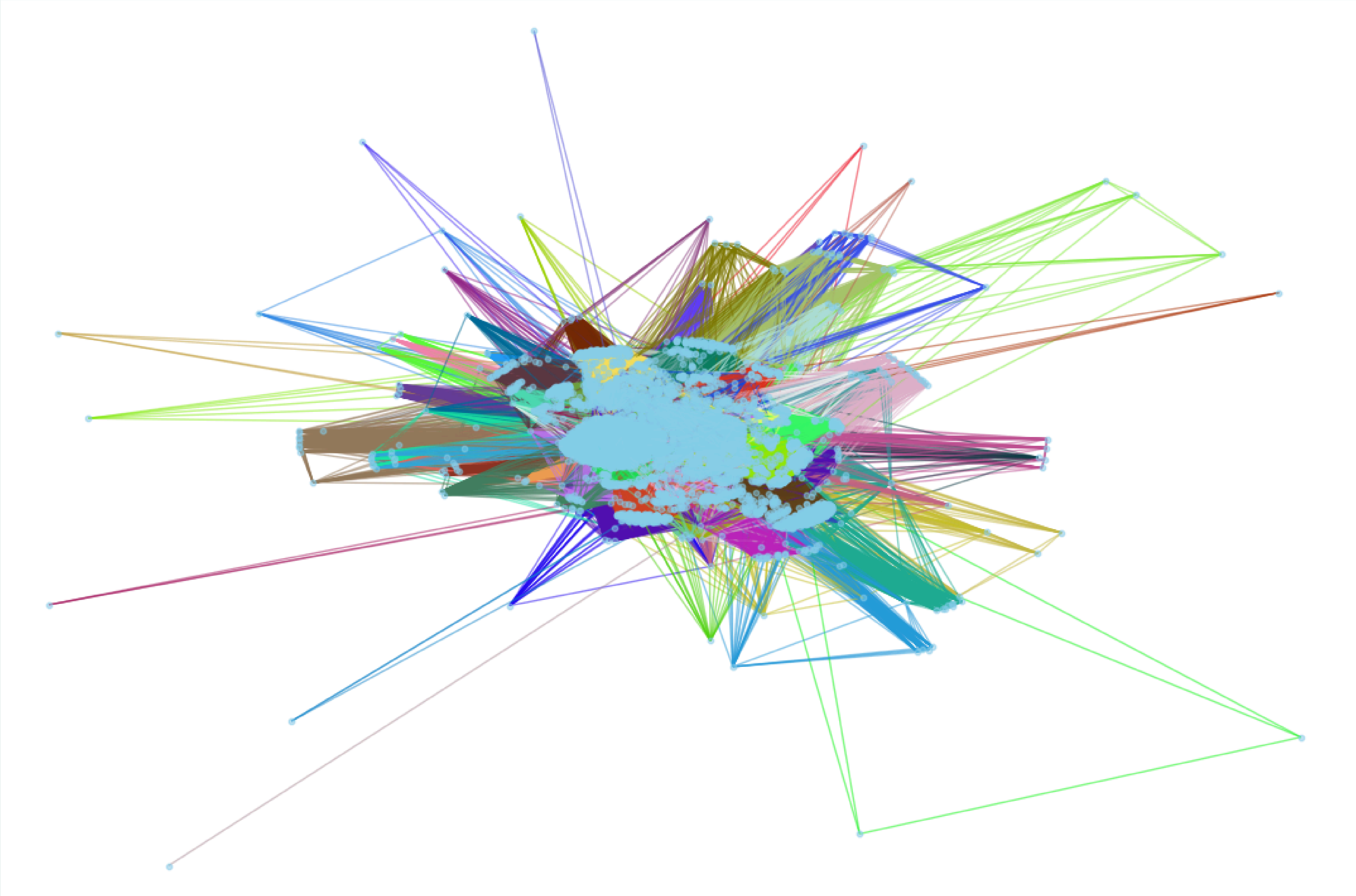
1. Our histogram shows top 25 genre in movies and we can se mostly drama and comedy are seen.

2. Below first word cloud healped in finding that our plots in dataset consisted of multiple stopwords then we removed them as seen is second word cloud with no stop words.





NETWORK VISUALIZATION



NUMBER OF NODES: 42207

NUMBER OF EDGES: 14522763

AVERAGE CLUSTERING COEFFICIENT: 0.78

DENSITY OF THE GRAPH: 0.30

MODULARITY : 0.26

IS CONNECTED: TRUE

THE GRAPH IS NOT BIPARTITE.

HIGHEST DEGREE: 7178

THE GRAPH IS NOT WEIGHTED.

Data in graph shows(Genre) -:

Note*Edge color shows 2 movies have same genre thus forming clusturs of multiple genre can be seen forming a giant component/network

METHODOLOGY

Linear regression

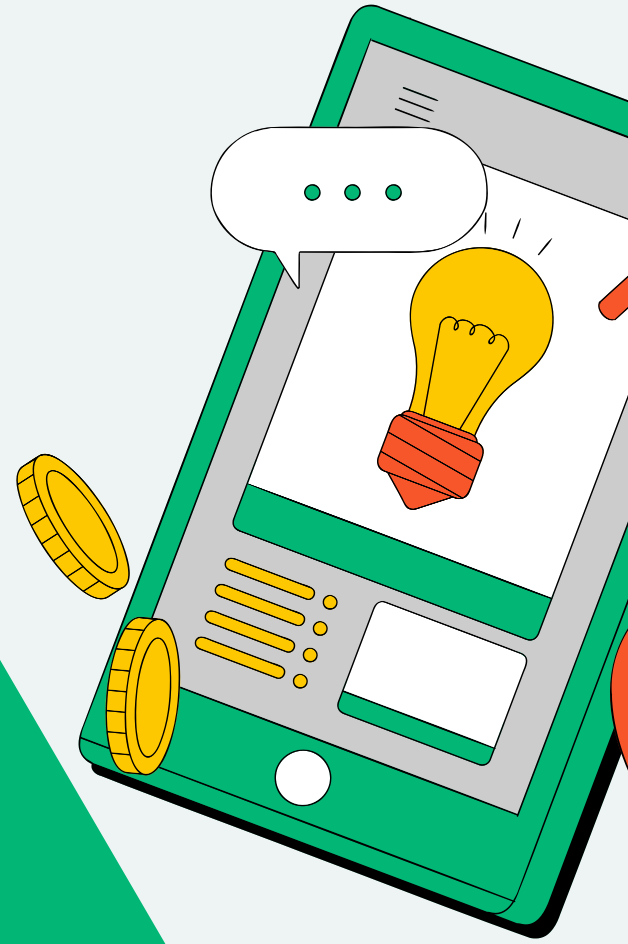
Decision tree

Random forest

Natural language processing TF-IDF

logistic regression

Naive bayes



LINEAR REGRESSION

- we have initialized the linear regression model.
- The model is trained using the features (X_train) and the target variable (y_train).
- The trained model is used to make predictions on the test data (X_test).
- Calculate RMSE (Root Mean Squared Error) for both training and testing data.
- the model accuracy R-squared score.
- Calculated MAPE (Mean Absolute Percentage Error) for both training and testing data.
- Display accuracy.

TRAIN DATA (R-SQUARED SCORE): 0.36
TEST DATA (R-SQUARED SCORE): 0.37

RMSE ON TRAINING DATA: 0.15
RMSE ON TESTING DATA: 0.13

MAPE ON TRAINING DATA: 6.24
MAPE ON TESTING DATA: 5.60

ACCURACY ON TRAINING DATA: 93.75
ACCURACY ON TESTING DATA: 94.39

DECISION TREE

- Initialized a DecisionTreeRegressor model with a random state of 42.
- Fit the Decision Tree model on the training data (X_train, y_train).
- Predicted the target variable for both the training and testing data.
- Evaluate the Decision Tree model:
  - Calculate the R-squared score.
  - Calculate the Root Mean Squared Error (RMSE).
  - Calculate the Mean Absolute Percentage Error (MAPE).
- Print out the evaluation results for the Decision Tree model, including R-squared score, RMSE, and MAPE, for both training and testing datasets.

MAPE (TRAIN): 1.0
MAPE(TEST)-0.13

R-SQUARED (TRAIN): 1.0
R-SQUARED (TEST): -0.13

RMSE (TRAIN): 2.25
RMSE (TEST): 0.17

ACCURACY ON TRAIN DATA : 100.0
ACCURACY ON TEST DATA : 93.20

RANDOM FOREST

- Initialized a Random Forest regressor model.
- Fit the Random Forest model to the training data (X_train, y_train).
- Used the trained Random Forest model to make predictions on both the training and testing data.
- Evaluate the Random Forest model's performance:
  - Calculate the R-squared score.
  - Calculate the Root Mean Squared Error (RMSE).
  - Calculate the Mean Absolute Error (MAE).
  - Calculate the Mean Absolute Percentage Error (MAPE).
- Print out the evaluation results for the Random Forest model, including R-squared score, RMSE, MAE, and MAPE, for both training and testing datasets.

MAPE(TRAIN): 0.92
MAPE (TEST) : 0.51

R-SQUARED (TRAIN): 0.92
R-SQUARED (TEST): 0.51

RMSE (TRAIN): 0.051
RMSE (TEST): 0.116

ACCURACY (TRAIN): 97.95
ACCURACY (TEST): 95.25

MOVIE NAME	PREDICTED GENRE	ACTUAL GENRE
I'LL DO ANYTHING	'DRAMA'	'DRAMA', 'COMEDY', 'DOMESTIC COMEDY'
PRIYA	'DRAMA'	'WORLD CINEMA', 'MUSICAL', 'DRAMA', 'ROMANTIC DRAMA', 'ROMANCE FILM', 'BOLLYWOOD'
CHEERFUL WEATHER FOR THE WEDDING	'DRAMA'	'DRAMA', 'COMEDY'
CREATURE	'HORROR'	'THRILLER', 'SCIENCE FICTION', 'HORROR'
GILIAP	'DRAMA'	'DRAMA'
28 WEEKS LATER	'HORROR', 'THRILLER'	THRILLER', 'SCIENCE FICTION', 'HORROR', 'DOOMSDAY FILM', 'SCI-FI HORROR', 'PLAGUE', 'ZOMBIE FILM'
SAPS AT SEA	'BLACK-AND-WHITE', 'COMEDY', 'SHORT FILM'	'COMEDY', 'BLACK-AND-WHITE
WISE GUYS	'COMEDY'	'CRIME FICTION', 'BUDDY FILM', 'ACTION/ADVENTURE', 'COMEDY', 'BLACK COMEDY', 'ACTION'
EL ACOMPAÑAMIENTO	'DRAMA'	'MUSICAL', 'DRAMA', 'COMEDY'
RELATIVE VALUES	'COMEDY', 'DRAMA'	'ROMANTIC COMEDY', 'ROMANCE FILM', 'COMEDY', 'WORLD CINEMA'

THANKING YOU

ANIKET KANOJIA
DIVYAM KHORWAL
HIMANSHU CHAUDHARY
KARTIK BANSAL