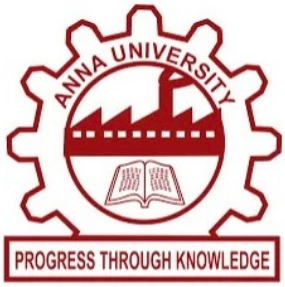
University College of Engineering Villupuram



Department of Computer Science and Engineering

Experience Based Project Learning - IBM (E2324)
conducted by IBM

Personalized content recommendation system
phase - 2 Document

Submitted by

1. S. Akalya
2. K. Harinitha
3. G. Muthazhagi
4. E.S. Mangala yazhini
5. PC .Vaishanavi

Personalized content recommendation system

Phase 2 : Data Wrangling and Analysis for Movie Recommendation system

Introduction :

Data wrangling is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics. It is a crucial process in the data analytics workflow that involves cleaning ,structuring, and enriching raw data to transform it into a more suitable format for analysis. Effectively performing data wrangling ensures that the data used for building recommendation models is clean, relevant ,and suitable for analysis, leading to more accurate and effective recommendations.

Objectives :

Data wrangling follows six major steps: Explore, transform, cleanse, enrich, validate and store.

Explore : Data exploration or discovery is a way to identify patterns, trends, and missing or incomplete information in a dataset.The bulk of exploration happens before creating reports, data visualizations , or training models, but it's common to uncover surprises and insights in a dataset during analysis too.

Cleanse : Data often contains errors as a result of manual entry, incomplete data, data automatically collected from sensors, or even malfunctioning equipment.Data cleansing corrects those entry errors, removes duplicates and outliers (if appropriate), eliminates missing data and improve the data quality .

Transform : Data transformation or data structuring is important; if not done early on, it can compromise the rest of the wrangling process. Data transformation involves putting the raw data in the right shape and format that will be useful for a report, data visualization , or analytic or modeling process.

Enrich : Enrichment or blending makes a dataset more useful by integrating additional sources such as authoritative third-party census, firmographics (Firmographic data is types of information that can be used to categorize organizations) ,or demographic data (Demographic data is information about groups of people according to certain attributes such as age ,education ,and place of residence).

Validate : Validation rules are repetitive programming sequences that verify data consistency, quality, and security.

Store : The last part of the wrangling process is to store or preserve the final product, along with all the steps and transformations that took place so it can be audited, understood, and repeated in the future.

Dataset Description :

The TMDb dataset used to build content based recommendation system .The movie dataset, which is originally from Kaggle, was cleaned and provided by Udacity (educational organization). The TMDb dataset used to build content based recommendation system .The TMDb (The Movie Database) is a comprehensive movie database that provides information about movies.The movie dataset csv file contains movie title , movie id , release date , run time ,vote average , genres ,keywords , tagline (a short description or comment on a movie),overview of the movie etc . The credit dataset csv file contains cast and crew of movie . These datasets are used to build the content based movie recommendation system.

Data Wrangling Techniques :

The dataset usually in the form of excel or csv files are converted into a DataFrame which is a data structure constructed with rows and columns used for machine learning.Pandas is a software library useful in data manipulation and analysis. DataFrame amazing when working with data, including indexing, filtering, grouping, merging, reshaping, and more.A dataset can be loaded from various data sources using relevant Pandas constructs as DataFrame.Once the dataset is loaded , various functions are used to understand the data description .

1. Data Description :
In machine learning or data science projects, we carry out data exploration to understand our data. If we are handling the data with the help of pandas library, we have the advantage of exploring our data easily by using pandas functions such as describe(), head(), unique() and count().
The goal of the Data Description Document is to record all information about the data files and their contents so that someone can use the data in a future research project and understand the data's content and structure.

Load dataset : The dataset can be loaded using python functions .In this project dataset is in the form of csv file , so the dataset is loaded is read_csv() method . The Movies.csv file and Credit.csv file are loaded using read_csv() method .

head method : The head() method returns a specified number of rows, string from the top. The head() method returns the first 5 rows if a number is not specified.

syntax : Movies.head(2)
output :

	budget	genres	homepage	id	keywords	original_language	original_title	overview	popularity	production_companies	production_countries	release_date	revenue	runtime	spoken_languages	status	tagline	title	vote_average	vote_count
0	23700000	[{"id": 28, "name"	http://www.avatarmovie.com/	1999	[{"id": 1463, "name"	en	Avatar	In the 22nd century, a paraplegic	150.437577	[{"name": "Ingenious Film	[{"iso_3166_1": "US", "name"	2009-12-10	27879650	162.0	[{"iso_639_1": "en",	Released	Enter the World	Avatar	7.2	11800
1	30000000	[{"id": 12, "name"	http://disney.go.com/disneypictu	285	[{"id": 270, "name": "ocean"}, {"id": 726	en	Pirates of the Caribbean: At	Captain Barbossa, long believed to	139.082615	[{"name": "Walt Disney Pictures", "id": 2}, {"	[{"iso_3166_1": "US", "name": "United	2007-05-19	96100000	169.0	[{"iso_639_1": "en", "name":	Released	At the end of the world	Pirates of the Cari	6.9	4500

syntax : Credit.head(2)
output :

	movie_id	title	cast	crew
0	19995	Avatar	[{"cast_id": 242, "character": "Jake Sully", "...	[{"credit_id": "52fe48009251416c750aca23", "de...
1	285	Pirates of the Caribbean: At World's End	[{"cast_id": 4, "character": "Captain Jack Spa...	[{"credit_id": "52fe4232c3a36847f800b579", "de...

tail method : The tail() method is used to return a specified number of last rows . The tail() method returns the last 5 rows if a number is not specified.

syntax : Movies.tail(2)
output :

	budget	genres	homepage	id	keywords	original_language	original_title	overview	popularity	production_companies	production_countries	release_date	revenue	runtime	spoken_languag	status	tagline	title	vote_average	vote_count
4801	0	[]	http://shanghaicalling.com/	126186	[]	en	Shanghai Calling	When ambitious New York	0.857008	[]	[{"iso_3166_1": "US", "name":	2012-05-03	0	98.0	[{"iso_639_1": "en",	Released	A New Yorker in Shanghai	Shanghai Calling	5.7	7
4802	0	[{"id": 99, "name": "Documentary"}]	NaN	25975	[{"id": 1523, "name": "obsession"}, {"id": 224	en	My Date with Drew	Ever since the second grade when he first saw	1.929883	[{"name": "rusty bear entertainment", "id": 87	[{"iso_3166_1": "US", "name": "United States	2005-08-05	0	90.0	[{"iso_639_1": "en", "name":	Released	NaN	My Date with Drew	6.3	16

syntax : Credit.tail(2)
output :

	movie_id	title	cast	crew
4801	126186	Shanghai Calling	[{"cast_id": 3, "character": "Sam", "credit_id...	[{"credit_id": "52fe4ad9c3a368484e16a36b", "de...
4802	25975	My Date with Drew	[{"cast_id": 3, "character": "Herself", "credi...	[{"credit_id": "58ce021b9251415a390165d9", "de...

Info method : The info() method prints information about the DataFrame. The information contains the number of columns, column labels, column data types, memory usage, range index, and the number of cells in each column (non-null values).

syntax : Movies.info()
output :

```
<bound method DataFrame.info of       budget                                             genres  \
0       237000000  [{"id": 28, "name": "Action"},
1       300000000  [{"id": 12, "name": "Adventure"},
...           ...                                                ...
4802            0  [{"id": 99, "name": "Documentary"}]

                                               homepage      id  \
0                          http://www.avatarmovie.com/   19995
1        http://disney.go.com/disneypictures/pirates/     285
...                                               ...     ...
4802                                              NaN   25975

                                               keywords original_language  \
0     [{"id": 1463, "name": "culture clash"}, {"id":...                en
1     [{"id": 270, "name": "ocean"}, {"id": 726, "na...                en
...                                                 ...               ...
4802  [{"id": 1523, "name": "obsession"}, {"id": 224...                en

                                 original_title  \
0                                        Avatar
1      Pirates of the Caribbean: At World's End
...                                         ...
```

```
4802                My Date with Drew

                                                overview  popularity  \
0                    In the 22nd century, a paraplegic Marine is di...  150.437577
1                    Captain Barbossa, long believed to be dead, ha...  139.082615
...                                                  ...         ...
4802                 Ever since the second grade when he first saw ...    1.929883
                                          production_companies  \
0                    [{"name": "Ingenious Film Partners", "id": 289...
1                    [{"name": "Walt Disney Pictures", "id": 2}, {"...
...                                                  ...
4802                 [{"name": "rusty bear entertainment", "id": 87...
                                          production_countries release_date  \
0                    [{"iso_3166_1": "US", "name": "United States o...   2009-12-10
1                    [{"iso_3166_1": "US", "name": "United States o...   2007-05-19
...                                                  ...          ...
4802                 [{"iso_3166_1": "US", "name": "United States o...   2005-08-05
                     revenue  runtime                                  spoken_languages  \
0                 2787965087    162.0  [{"iso_639_1": "en", "name": "English"}, {"iso...
1                  961000000    169.0   [{"iso_639_1": "en", "name": "English"}]
...                      ...      ...                                        ...
4802                       0     90.0  [{"iso_639_1": "en", "name": "English"}]
                     status                                  tagline  \
0                  Released                 Enter the World of Pandora.
1                  Released                At the end of the world, the adventure begins.
...                     ...                                      ...
4802               Released                                      NaN
                                                 title  vote_average  vote_count
0                                               Avatar           7.2       11800
1                Pirates of the Caribbean: At World's End           6.9        4500
...                                                 ...           ...         ...
4802                                My Date with Drew           6.3          16
[4803 rows x 20 columns]>
```

syntax : Credit.info()

output :

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4803 entries, 0 to 4802
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   movie_id  4803 non-null   int64
 1   title     4803 non-null   object
 2   cast      4803 non-null   object
 3   crew      4803 non-null   object
dtypes: int64(1), object(3)
memory usage: 150.2+ KB
```

describe method : The describe() method returns description of the data in the DataFrame. If the DataFrame contains numerical data, the description contains these information for each column.

syntax : Movies.describe()
output :

	budget	id	popularity	revenue	runtime	vote_average	vote_count
count	4.803000e+03	4803.000000	4803.000000	4.803000e+03	4801.000000	4803.000000	4803.000000
mean	2.904504e+07	57165.484281	21.492301	8.226064e+07	106.875859	6.092172	690.217989
std	4.072239e+07	88694.614033	31.816650	1.628571e+08	22.611935	1.194612	1234.585891
min	0.000000e+00	5.000000	0.000000	0.000000e+00	0.000000	0.000000	0.000000
25%	7.900000e+05	9014.500000	4.668070	0.000000e+00	94.000000	5.600000	54.000000
50%	1.500000e+07	14629.000000	12.921594	1.917000e+07	103.000000	6.200000	235.000000
75%	4.000000e+07	58610.500000	28.313505	9.291719e+07	118.000000	6.800000	737.000000
max	3.800000e+08	459488.000000	875.581305	2.787965e+09	338.000000	10.000000	13752.000000

syntax : Credit.describe()
output :

	movie_id
count	4803.000000
mean	57165.484281
std	88694.614033
min	5.000000
25%	9014.500000
50%	14629.000000
75%	58610.500000
max	459488.000000

shape : The shape() method is used to fetch the dimensions of Pandas and NumPy type objects in python.

syntax : Movies.shape
output : (4544, 8)

syntax : Credit.shape
output : (4544, 8)

count method : The count() method returns the number of elements with the specified value.

syntax : Movie.count()
output :

```
budget                  4803
genres                  4803
homepage                1712
id                      4803
keywords                4803
original_language       4803
original_title          4803
overview                4800
popularity              4803
production_companies    4803
production_countries    4803
release_date            4802
revenue                 4803
runtime                 4801
spoken_languages        4803
status                  4803
tagline                 3959
title                   4803
vote_average            4803
vote_count              4803
dtype: int64
```

syntax : Credit.count()
output :

```
movie_id    4803
title       4803
cast        4803
crew        4803
dtype: int64
```

2. Null Data Handling :

In a dataset, we often see the presence of empty cells, rows, and columns, also referred to as Missing values. They make the dataset inconsistent and unable to work on. Many machine learning algorithms return an error if parsed with a dataset containing null values. Detecting and treating missing values is essential while analyzing and formulating data for any purpose.

a) Null data Identification : The null values in the dataset are identified using methods provided by pandas library .

isnull( ) : Identifies missing values in a Series or DataFrame.

sum( ) : It provides an inbuilt function sum() which sums up the numbers in the list.

syntax : Movies.isnull().sum()
output :

```
budget                     0
genres                     0
homepage                3091
id                         0
keywords                   0
original_language          0
original_title             0
overview                   3
popularity                 0
production_companies       0
production_countries       0
release_date               1
revenue                    0
runtime                    2
spoken_languages           0
status                     0
tagline                  844
title                      0
vote_average               0
vote_count                 0
dtype: int64
```

syntax : Credit.isnull().sum()
output :

```
movie_id    0
title       0
cast        0
crew        0
dtype: int64
```

notnull() : Detect non-missing values for an array-like object . check for missing values in a pandas Series or DataFrame .

syntax : Movies.notnull()
output :

	genres	id	keywords	title	overview
0	True	True	True	True	True
1	True	True	True	True	True
2	True	True	True	True	True
3	True	True	True	True	True
4	True	True	True	True	True
...	...	...	...	...	...
4798	True	True	True	True	True
4799	True	True	True	True	True
4800	True	True	True	True	True
4801	True	True	True	True	True
4802	True	True	True	True	True

4800 rows × 5 columns

syntax : Credit.notnull()
output :

	movie_id	title	cast	crew
0	True	True	True	True
1	True	True	True	True
2	True	True	True	True
3	True	True	True	True
4	True	True	True	True
...	...	...	...	...
4798	True	True	True	True
4799	True	True	True	True
4800	True	True	True	True
4801	True	True	True	True
4802	True	True	True	True

4803 rows × 4 columns

isna() : similar to notnull() but returns True for missing values and False for non-missing values.

syntax : Movies.isna()
output :

	genres	id	keywords	title	overview
0	False	False	False	False	False
1	False	False	False	False	False
2	False	False	False	False	False
3	False	False	False	False	False
4	False	False	False	False	False
...	...	...	...	...	...
4798	False	False	False	False	False
4799	False	False	False	False	False
4800	False	False	False	False	False
4801	False	False	False	False	False
4802	False	False	False	False	False

4800 rows × 5 columns

syntax : Credit.isna()
output :

	movie_id	title	cast	crew
0	False	False	False	False
1	False	False	False	False
2	False	False	False	False
3	False	False	False	False
4	False	False	False	False
...	...	...	...	...
4798	False	False	False	False
4799	False	False	False	False
4800	False	False	False	False
4801	False	False	False	False
4802	False	False	False	False

4803 rows × 4 columns

b) Null data Imputation : Imputation is the practice of filling missing values, known as null values. In Movies dataset , the columns homepage ,runtime , overview and tagline cannot be filled with appropriate values since most of these are not numerical values .

c) Null data Removal : This step involves removing the rows and column with null values to increase the quality of data thereby enhancing the analysis process .

drop() method : The drop() method removes the specified row or column.By specifying the column axis (axis = "column"), the drop method removes the specified column.By specifying the row axis (axis = "index"), the drop() method removes the specified row.

Removing the columns with null values and that are not necessary for analysis using drop() :

syntax : Movies.drop(columns = ["homepage","tagline","release_date","runtime"],inplace = True)

After droping the columns with null values :

syntax : Movies.isnull().sum()
output :

```
budget                  0
genres                  0
id                      0
keywords                0
original_language       0
original_title          0
overview                3
popularity              0
production_companies    0
production_countries    0
revenue                 0
spoken_languages        0
status                  0
title                   0
vote_average            0
vote_count              0
dtype: int64
```

dropna() method : The dropna() method removes the rows that contains NULL values.The dropna() method returns a new DataFrame object unless the inplace parameter is set to True, in that case the dropna() method does the removing in the original DataFrame instead.

Droping the rows which has null values in overview column :

syntax : Movies.dropna(inplace = True)

After droping the rows with null values in overview column :

syntax : Movies.isnull().sum()
output :

```
budget                  0
genres                  0
id                      0
keywords                0
original_language       0
original_title          0
overview                0
popularity              0
production_companies    0
production_countries    0
revenue                 0
spoken_languages        0
status                  0
title                   0
vote_average            0
vote_count              0
dtype: int64
```

3. Data Validation :

Data validation means checking the accuracy and quality of source data before using, importing or otherwise processing data.Different types of validation can be performed depending on destination constraints or objectives. Data validation is a form of data cleansing.

a) Data Integrity check : Data integrity is a concept and process that ensures the accuracy, completeness, consistency, and validity of an data.Data integrity describes data that's kept complete, accurate, consistent and safe.

Validate the Movies dataset :
syntax :

```
for dtype in Movies.dtypes.items() :
        print(dtype)
```

output :

```
('genres', dtype('O'))
('id', dtype('int64'))
('keywords', dtype('O'))
('title', dtype('O'))
('overview', dtype('O'))
```

Validate the Credit dataset :
syntax :

```
for dtype in Credit.dtypes.items() :
        print(dtype)
```

output :

```
('movie_id', dtype('int64'))
('title', dtype('O'))
('cast', dtype('O'))
('crew', dtype('O'))
```

duplicated() : The duplicated() method returns a Series with True and False values that describe which rows in the DataFrame are duplicated and not.

Identification of Duplicate values in Movies dataset :
syntax : Movies.duplicated().sum()
output : 0

Identification of duplicated values in Credit dataset :

syntax : Credit.duplicated().sum()
output : 0

b) Data Consistency verification :
A consistency check is a type of logical check that confirms the data's been entered in a logically consistent way.Data consistency refers to the state of data in which all copies or instances are the same across all systems and databases. Consistency helps ensure that data is accurate, up-to-date and coherent across different database systems, applications and platforms.

value_counts() : This function returns object containing counts of values .

syntax : Movies["title"].value_counts()
output :

```
title
Batman                                  2
Out of the Blue                         2
The I Inside                            1
Ultramarines: A Warhammer 40,000 Movie  1
Crocodile Dundee                        1
                                       ..
Secondhand Lions                        1
The Age of Adaline                      1
Drag Me to Hell                         1
Southpaw                                1
My Date with Drew                       1
Name: count, Length: 4798, dtype: int64
```

unique() : function returns the unique values that occur at least once.

syntax : Movie["original_title"].unique()
output :

```
array(['Avatar', "Pirates of the Caribbean: At World's End", 'Spectre',
       ..., 'Signed, Sealed, Delivered', 'Shanghai Calling',
       'My Date with Drew'], dtype=object)
```

4. Data Merging :
Pandas DataFrame is two-dimensional size-mutable, potentially heterogeneous tabular data structure with labelled axes (rows and columns). We can join, merge, and concat dataframe using different methods. In Dataframe df.merge(),df.join(), and df.concat() methods help in joining, merging and concatenating different dataframe .

merge() : merging is the process of bringing two datasets together into one, and aligning the rows from each based on common attributes or columns.

syntax :

Movies_data = Movies.merge(Credit , on = "title")
Movies_data.head(2)
output :

	genres	id	keywords	title	overview	movie_id	cast	crew
0	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	19995	[{"id": 1463, "name": "culture clash"}, {"id":...	Avatar	In the 22nd century, a paraplegic Marine is di...	19995	[{"cast_id": 242, "character": "Jake Sully", "...	[{"credit_id": "52fe48009251416c750aca23", "de...
1	[{"id": 12, "name": "Adventure"}, {"id": 14, "...	285	[{"id": 270, "name": "ocean"}, {"id": 726, "na...	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	285	[{"cast_id": 4, "character": "Captain Jack Spa...	[{"credit_id": "52fe4232c3a36847f800b579", "de...

Shape of merged dataset :

syntax : Movies_data.shape

output : (4544, 8)

5. Data Formatting :

Data formatting is the process of converting data into a common format, facilitating data comparison for users.Data formatting is the organization of information according to preset rules, usually in preparation for processes like data migration.

Applying data formatting techniques to the required columns :

The columns genres , cast , crew have string of dictionary data which cannot be used for analysis , so first these columns must be converted into list data format which can then be used for analysis .for building content based recommendation system the required values from each column are chosen.

All the important words are taken from the genre and keyword column , name of the directors are taken from the crew column and three cast members from the cast column .

Import ast module in python for converting columns with string datatype to list :

ast.literal_eval() : By utilizing the ast.literal_eval() function from the ast module, the string is safely evaluated as Python code, converting it into an actual list.

 # Function to convert genre and keywords with string datatype into list

```
def convert(text) :
    l = []
    for i in ast.literal_eval(text) :
        l.append(i["name"])
    return l
```

 # Function to convert the crew column to list of name of the director

```
def convert_crew(text) :
    l = []
    for i in ast.literal_eval(text) :
        if i["job"] == "Director" :
            l.append(i["name"])
    return l
```

 # Function to convert the cast column containing the list of cast members

```
def convert_cast(text) :
    l = []
    count = 0
    for i in ast.literal_eval(text) :
        count += 1
        if count < 4 :
            l.append(i["name"])
    return l
```

apply() : Pandas.apply allow the users to pass a function and apply it on every single value of the Pandas series. It comes as a huge improvement for the pandas library as this function helps to segregate data according to the conditions required due to which it is efficiently used in data science and machine learning.

code :

```
Movies_data["genres"] = Movies["genres"].apply(convert)
Movies_data["keywords"] = Movies["keywords"].apply(convert)
Movies_data["cast"] = Movies_data["cast"].apply(convert_cast)
Movies_data["crew"] = Movies_data["crew"].apply(convert_crew)
```

output :

	genres	id	keywords	title	overview	movie_id	cast	crew
0	[Action, Adventure, Fantasy, Science Fiction]	19995	[culture clash, future, space war, space colon...	Avatar	In the 22nd century, a paraplegic Marine is di...	19995	[Sam Worthington, Zoe Saldana, Sigourney Weaver]	[James Cameron]
1	[Adventure, Fantasy, Action]	285	[ocean, drug abuse, exotic island, east india ...	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	285	[Johnny Depp, Orlando Bloom, Keira Knightley]	[Gore Verbinski]

Check data consistency after merging : To ensure that the merged data is consistent , we use the isnull() function to verify it .

syntax : Movies_data.isnull().sum()

output :

```
genres      3
id          0
keywords    3
title       0
overview    0
movie_id    0
cast        0
crew        0
dtype: int64
```

Replace NaN in genres and keywords column with empty string :

The empty strings are replaced using the fillna() function .

fillna() :

The fillna() method replaces the NULL values with a specified value.The fillna() method returns a new DataFrame object unless the inplace parameter is set to True, in that case the fillna() method does the replacing in the original DataFrame instead.

code :

```
Movies_data["genres"] = Movies_data["genres"].fillna("")

Movies_data["keywords"] = Movies_data["keywords"].fillna("")

Movies_data.isnull().sum()
```

output :

```
genres      0
id          0
keywords    0
title       0
overview    0
movie_id    0
cast        0
crew        0
dtype: int64
```

Convert the overview of string datatype to a list of strings :

A lambda function is used to convert the overview of string datatype to list using the split method .

syntax :

Movies_data["overview"] = Movies_data["overview"].apply(lambda x : x.split())

Movies_data.head(2)

output :

	genres	id	keywords	title	overview	movie_id	cast	crew
0	[Action, Adventure, Fantasy, Science Fiction]	19995	[culture clash, future, space war, space colon...	Avatar	[In, the, 22nd, century,, a, paraplegic, Marin...	19995	[Sam Worthington, Zoe Saldana, Sigourney Weaver]	[James Cameron]
1	[Adventure, Fantasy, Action]	285	[ocean, drug abuse, exotic island, east india ...	Pirates of the Caribbean: At World's End	[Captain, Barbossa,, long, believed, to, be, d...	285	[Johnny Depp, Orlando Bloom, Keira Knightley]	[Gore Verbinski]

Convert the spaces in between the words so as to reduce the error :

 # function to remove the white spaces

code :

```
def remove_space(word) :
    l = []
    for i in word :
        l.append(i.replace(" ",""))
    return l
Movies_data["genres"] = Movies_data["genres"].apply(remove_space)
Movies_data["keywords"] = Movies_data["keywords"].apply(remove_space)
Movies_data["cast"] = Movies_data["cast"].apply(remove_space)
Movies_data["crew"] = Movies_data["crew"].apply(remove_space)
Movies_data.head(2)
```

output :

	genres	id	keywords	title	overview	movie_id	cast	crew
0	[Action, Adventure, Fantasy, ScienceFiction]	19995	[cultureclash, future, spacewar, spacecolony, ...	Avatar	[In, the, 22nd, century,, a, paraplegic, Marin...	19995	[SamWorthington, ZoeSaldana, SigourneyWeaver]	[JamesCameron]
1	[Adventure, Fantasy, Action]	285	[ocean, drugabuse, exoticisland, eastindiatrad...	Pirates of the Caribbean: At World's End	[Captain, Barbossa,, long, believed, to, be, d...	285	[JohnnyDepp, OrlandoBloom, KeiraKnightley]	[GoreVerbinski]

Combine all the columns to prepare the content based recommendation system :

After these columns are modified with appropriate datatypes , now these columns containing the important words are concatenated to make a single column which is used to build the recommendation system using NLP technique .

code :

```
Movies_data["tags"] = Movies_data["genres"] + Movies_data["keywords"] +
Movies_data["overview"] + Movies_data["cast"] + Movies_data["crew"]

data = Movies_data

data.drop(columns = ["genres","keywords","overview","cast","crew","movie_id"], inplace = True)
data.head()
```

output :

id	title	tags
995	Avatar	[Action, Adventure, Fantasy, ScienceFiction, c...
285	Pirates of the Caribbean: At World's End	[Adventure, Fantasy, Action, ocean, drugabuse,...
647	Spectre	[Action, Adventure, Crime, spy, basedonnovel, ...
026	The Dark Knight Rises	[Action, Crime, Drama, Thriller, dccomics, cri...
529	John Carter	[Action, Adventure, ScienceFiction, basedonnov...

Convert all the words in tags to lower case :

converting the words to lowercase to reduce the errors when building the recommendation system based on the content in tags column .

code :

```
def lower_case(text) :
    l = []
    for i in text :
        i = i.lower()
        l.append(i)
    return l
data["tags"] = data["tags"].apply(lower_case)
data.head()
```

output :

	id	title	tags
0	19995	Avatar	[action, adventure, fantasy, sciencefiction, c...
1	285	Pirates of the Caribbean: At World's End	[adventure, fantasy, action, ocean, drugabuse,...
2	206647	Spectre	[action, adventure, crime, spy, basedonnovel, ...
3	49026	The Dark Knight Rises	[action, crime, drama, thriller, dccomics, cri...
4	49529	John Carter	[action, adventure, sciencefiction, basedonnov...

6. Exploratory Data Analysis :

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. Exploratory Data Analysis (EDA) refers to the method of studying and exploring record sets to apprehend their predominant traits, discover patterns, locate outliers, and identify relationships between variables. EDA is normally carried out as a preliminary step before undertaking extra formal statistical analyses or modeling.

Types of EDA :

1. Univariate Analysis :

Univariate analysis is basically the simplest form to analyze data. Uni means one and this means that the data has only one kind of variable. The major reason for univariate analysis is to use the data to describe. The analysis will take data, summarise it, and then find some pattern in the data.

2. Bivariate Analysis :

Bivariate analysis is the simultaneous analysis of two variables. It explores the concept of the relationship between two variable whether there exists an association and the strength of this association or whether there are differences between two variables and the significance of these differences.

3. Multivariate Analysis :

It is an extension of bivariate analysis which means it involves multiple variables at the same time to find correlation between them. Multivariate Analysis is a set of statistical model that examine patterns in multidimensional data by considering at once, several data variable.

Common plots used in Exploratory data analysis :

The types of EDA techniques that can be employed at some stage in information evaluation. The choice of strategies relies upon on the information traits, research questions, and the insights sought from the analysis.

Histogram : A histogram is a bar graph-like representation of data that buckets a range of classes into columns along the horizontal x-axis. The vertical y-axis represents the number count or percentage of occurrences in the data for each column. Columns can be used to visualize patterns of data distributions.

Bar plot : A bar plot or bar chart is a graph that represents the category of data with rectangular bars with lengths and heights that is proportional to the values which they represent. The bar plots can be plotted horizontally or vertically. A bar chart describes the comparisons between the discrete categories. One of the axis of the plot represents the specific categories being compared, while the other axis represents the measured values corresponding to those categories.A bar graph is used to compare discrete or categorical variables in a graphical format whereas a histogram depicts the frequency distribution of variables in a dataset.

Line plot : Line Plots depict the relationship between continuous as well as categorical values in a continuous data point format

Matplotlib and seaborn library are used in python for visualization :

plot() : The plot() function in matplotlib.pyplot is used to draw points (markers) in a diagram. By default, the plot() function draws a line from point to point. The function takes parameters for specifying points in the diagram .

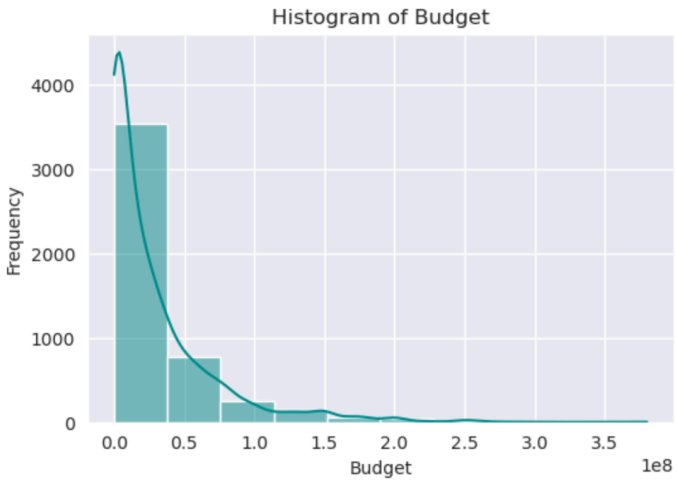
lineplot() : A line plot in seaborn is a relational data visualization showing how one continuous variable changes when another does. It's one of the most common graphs widely used in finance, sales, marketing, healthcare, natural sciences, and more.

histplot() : The **sns.histplot** function in Seaborn is designed for drawing histograms, which are essential for examining the distribution of continuous data. This function is versatile and allows for extensive customization, making it easier to draw meaningful insights from the data.

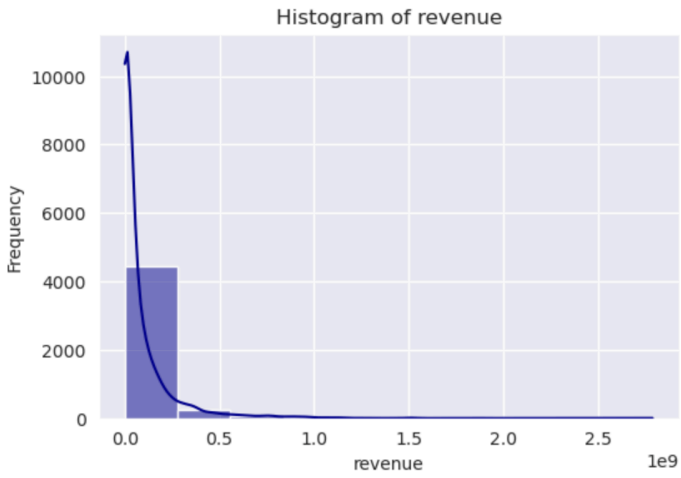
countplot() : Show the counts of observations in each categorical bin using bars.A count plot can be thought of as a histogram across a categorical, instead of quantitative, variable. The basic API and options are identical to those for barplot() so you can compare counts across nested variables.

Histogram :

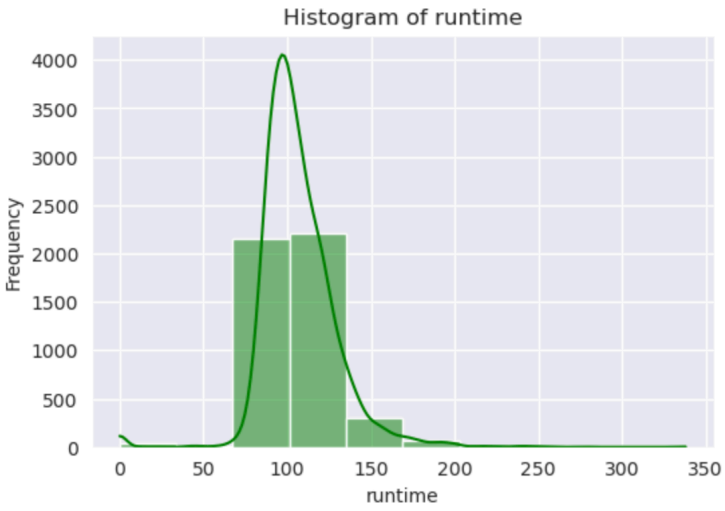
Histogram for Budget :



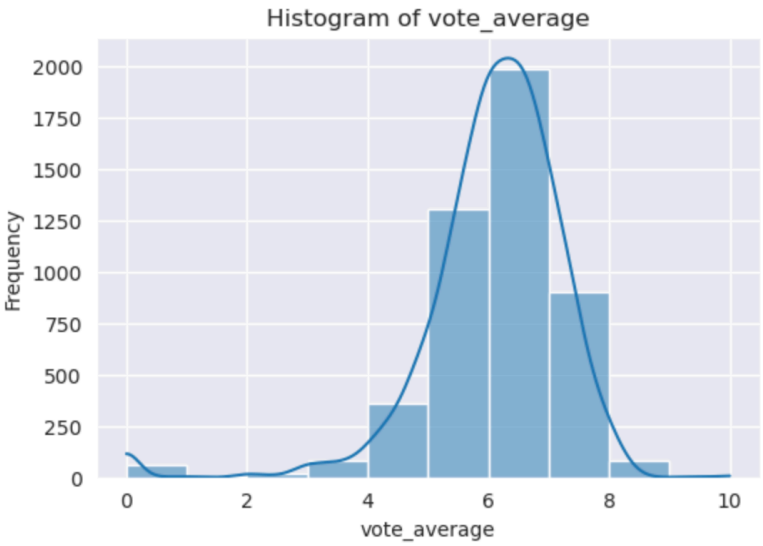
Histogram for revenue :



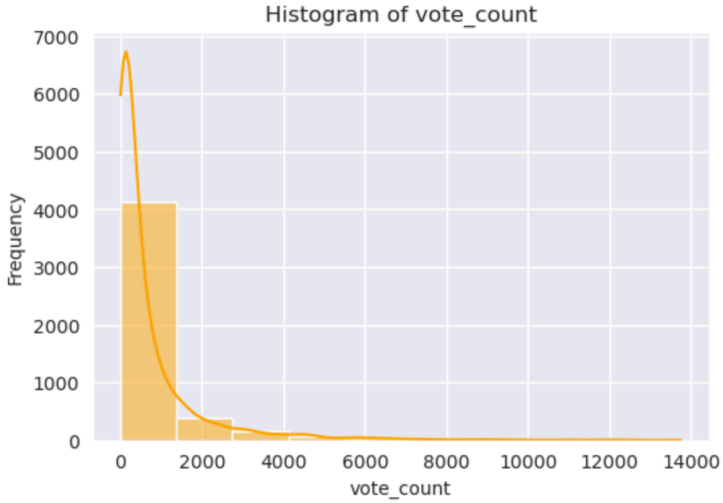
Histogram for runtime :



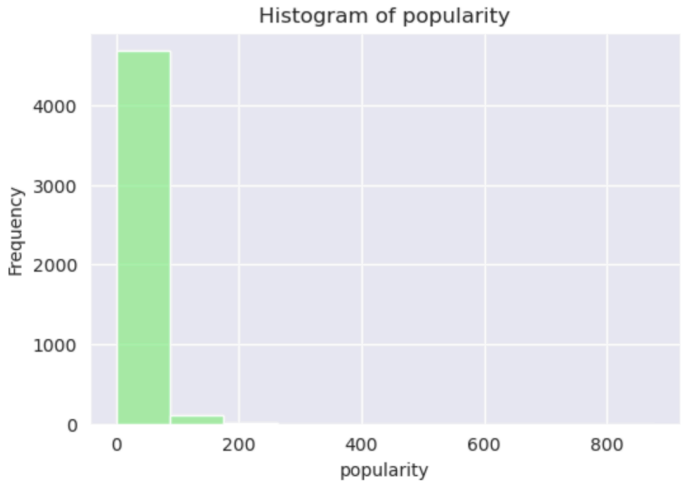
Histogram for vote average :



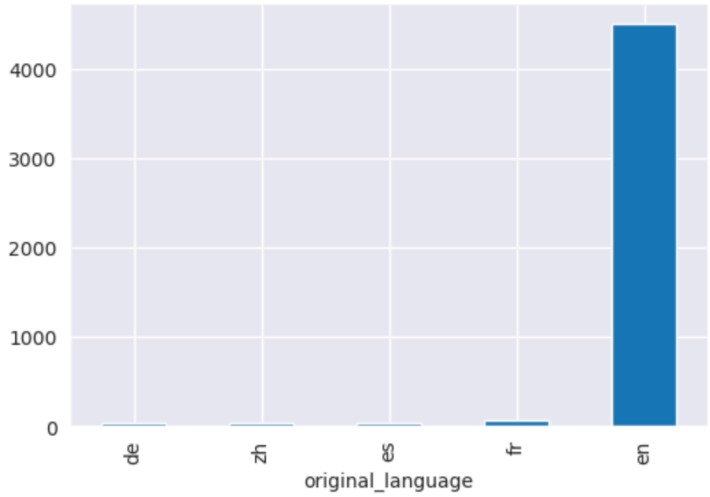
Histogram for vote count :



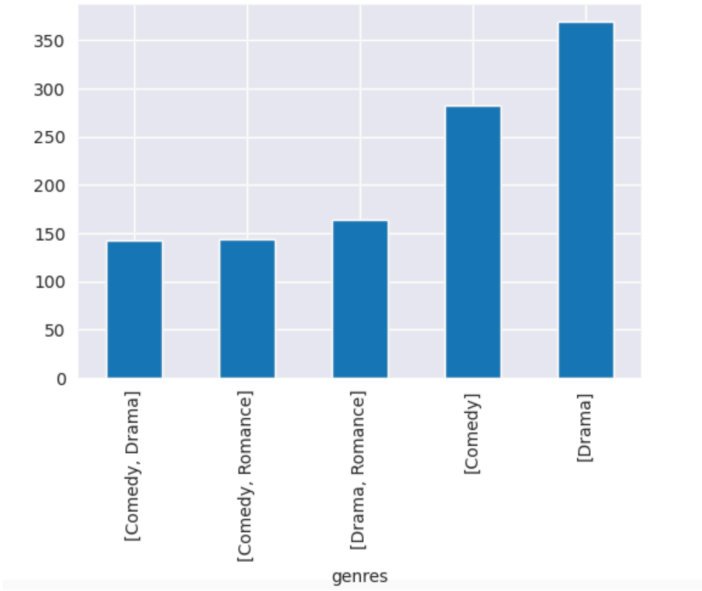
Histogram for popularity :



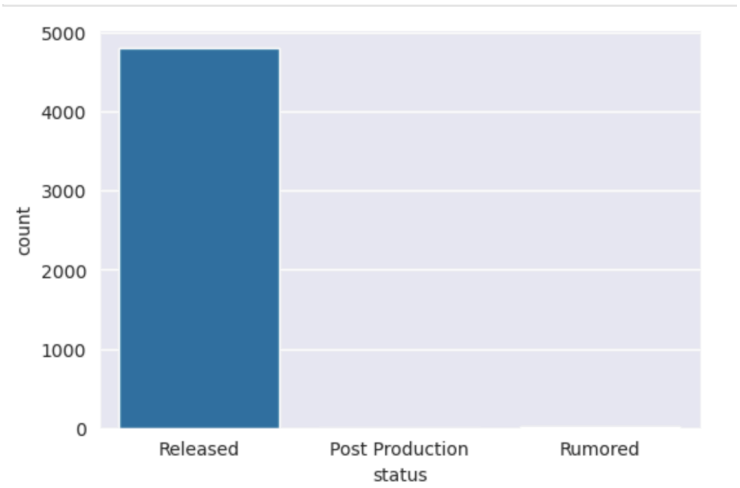
count plot for original language :



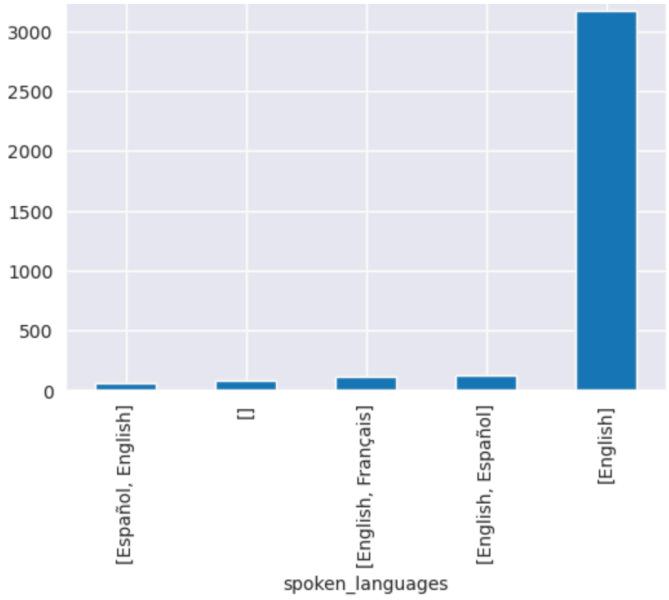
count plot for popularity :



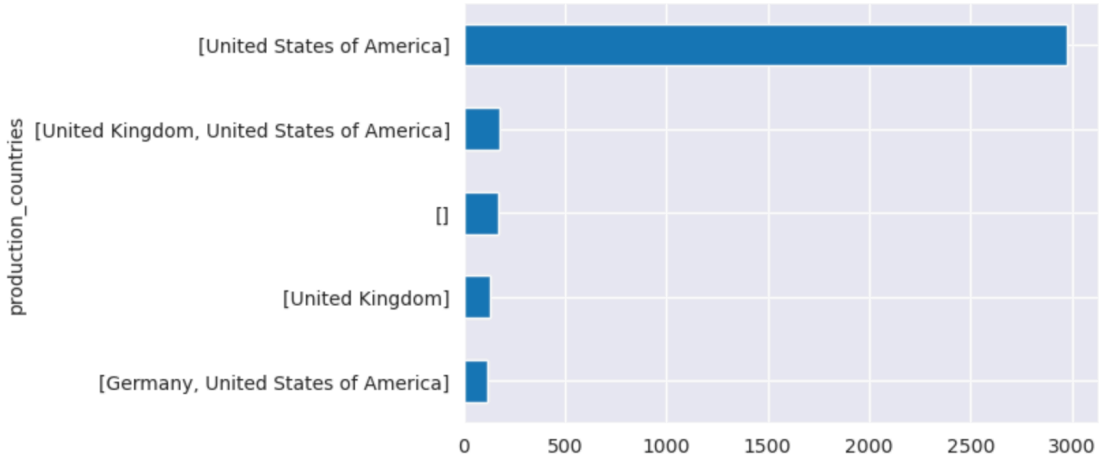
count plot for movie status :



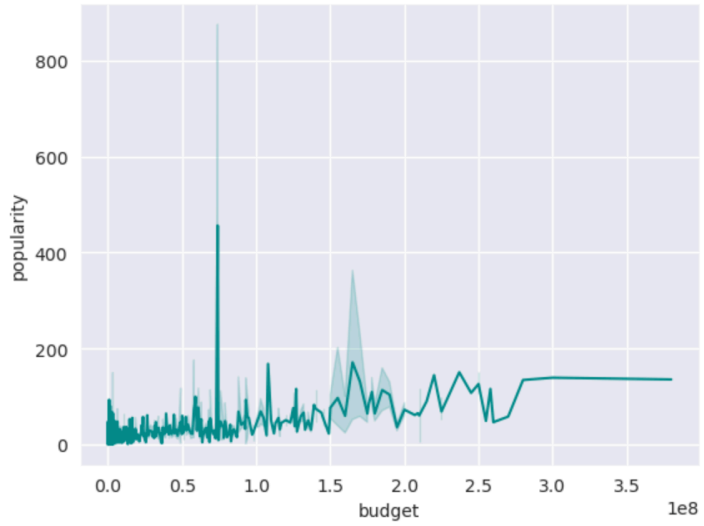
count plot for spoken languages :



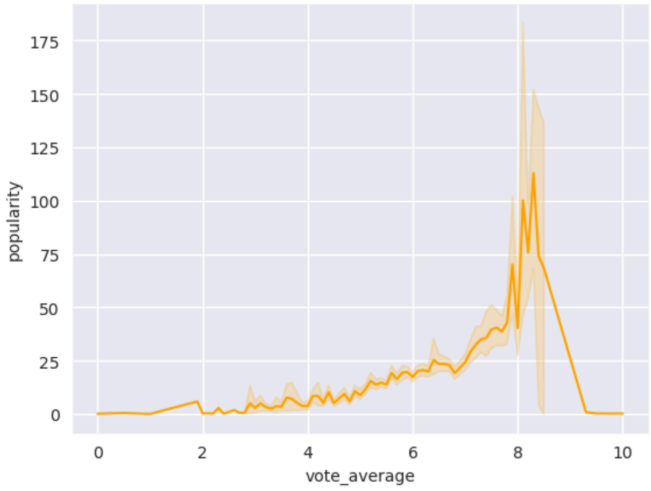
count plot for production countries :



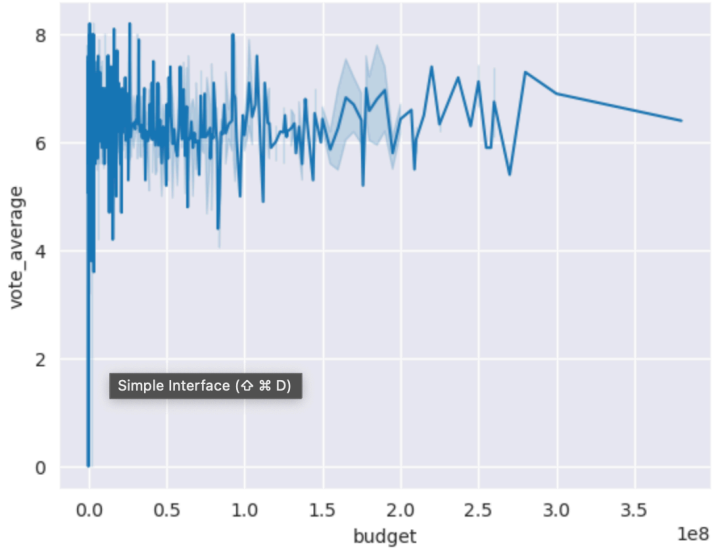
line plot for budget vs popularity :



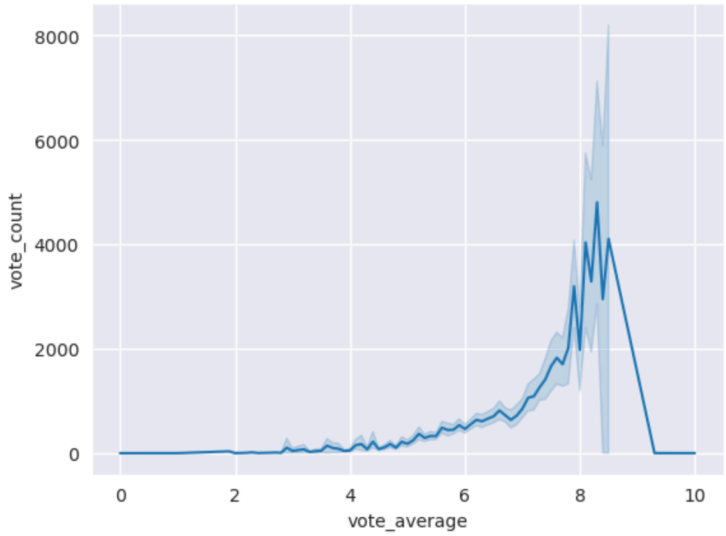
line plot for vote average vs popularity :



line plot for budget vs vote average :



line plot for vote average vs vote count :



These are some of the graphical representation used for Exploratory data analysis of the dataset to understand about the dataset and find the correlation between the columns in the dataset .

Steps After data Preprocessing to build the movie recommendation system :

- Once the data is preprocessed , the next step is embedding . Embedding is the process of creating vectors using deep learning. An "embedding" is the output of this process . A vector that is created by a deep learning model for the purpose of similarity searches by that model. This is used to find the similarity between the movies and recommend the users with similar content .
- A distance measure called cosine similarity to find the resemblance between each bag-of-words. Cosine similarity is a metric that calculates the cosine of the angle between two or more vectors to determine if they are pointing in the same direction.
- Cosine similarity ranges between 0 and 1. A value of 0 indicates that the two vectors are not similar at all, while 1 tells us that they are identical.

Scenario : This project aims to recommend personalized content based movie recommendation system for users based on their past interaction and preferences.

Objective : A movie based recommender system is a software tool that suggests movies to users based on their personal preferences. It uses algorithms and machine learning to analyze data points, such as a user's previous movie choices and ratings, to generate personalized recommendations.

Target Audience : The digital paltform users seeking personalized content recommendations

Conclusion :

The conclusion of data wrangling in a recommendation system is crucial for ensuring the quality and effectiveness of the system. Data wrangling involves various steps such as data collection, cleaning, integration, transformation, and validation.Data wrangling results in reliable data insights, improving the effectiveness of decision-making processes within an organization. Clean data reduces the risk of taking actions based on inaccurate or incomplete information.After data wrangling , this dataset is used to build the recommendation system using NLP techniques provided by python libraries .