PREDICTING IMDB SCORES

Phase 3 Document submission

Name : E. Tharun

Department : CSE-III

Register Number : 410121104058

NM ID : au410121104039

Domain : Applied Data Science

Project Title : Prediction IMDB Score

College : Adhi college of engineering and technology, Kanchipuram.

PREDICTING IMDB SCORES

About the Dataset

This dataset consists of all Netflix original films released as of June 1st, 2021. Additionally, it also includes all Netflix documentaries and specials. The data was web scraped from this Wikipedia page, which was then integrated with a dataset consisting of all of their corresponding IMDB scores. IMDB scores are voted on by community members, and most of the films have 1,000+ reviews.

Content

Included in the dataset is:

- Title of the film
- Genre of the film
- Original premiere date
- Runtime in minutes
- IMDB scores (as of 06/01/21)
- Languages currently available (as of 06/01/21)

Dataset Link: https://www.kaggle.com/datasets/luiscorter/netflix-original-films-imdb-scores

Implementation

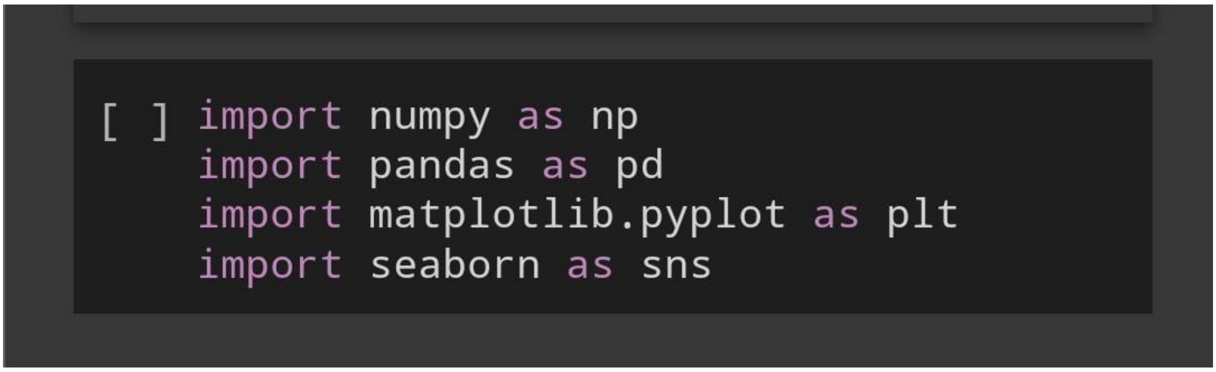
The Dataset is downloaded and loaded by mounting the Google Drive in collab and the necessary preprocessing steps are done such as missing values, null values and duplicated values etc.

Since there were no null values and duplicated values found, the data analysis process is carried forward.

Data preprocessing is a crucial step in any data analysis. This process involves cleaning and transforming the raw data to make it suitable for analysis. In this report, we will outline the key steps and techniques for data preprocessing in Python using various libraries, primarily Pandas and NumPy.

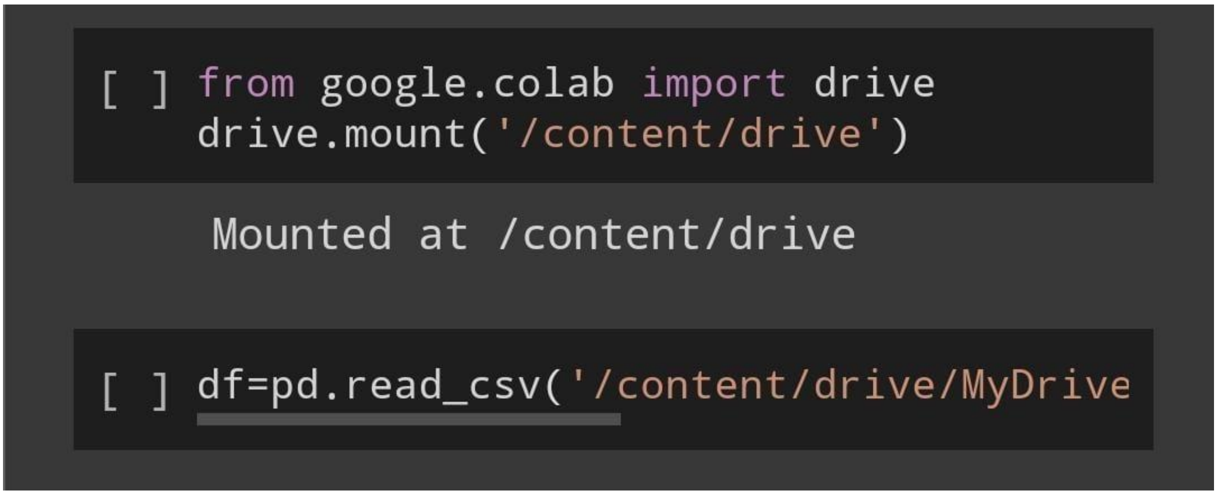
Coding

1. Loading the required modules



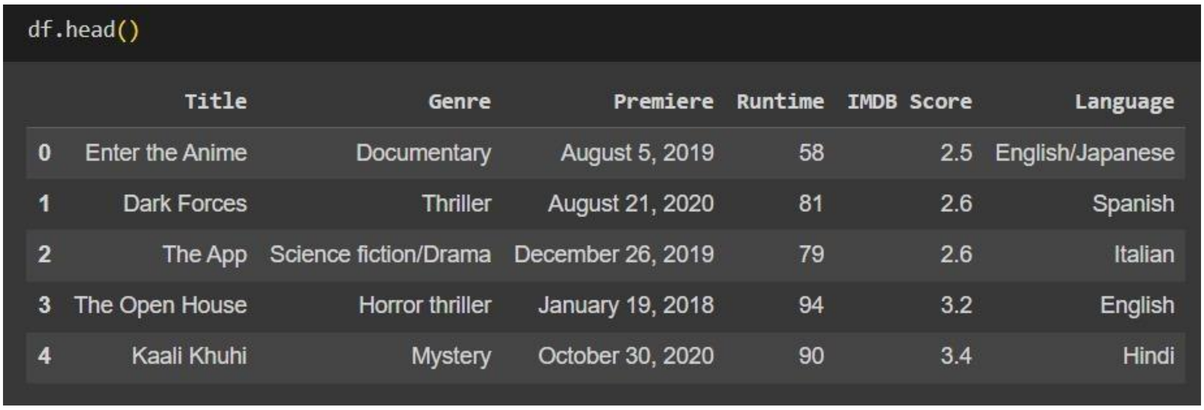
```
[ ] import numpy as np
    import pandas as pd
    import matplotlib.pyplot as plt
    import seaborn as sns
```

2. Loading the Dataset



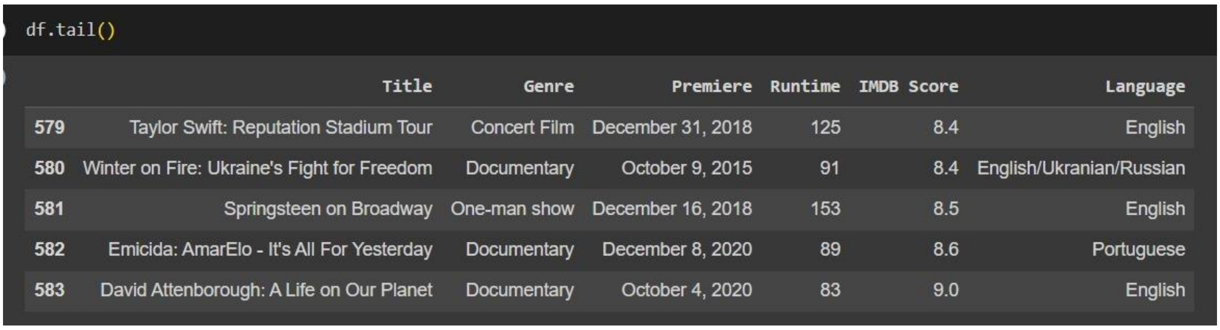
```
[ ] from google.colab import drive
    drive.mount('/content/drive')

     Mounted at /content/drive

[ ] df=pd.read_csv('/content/drive/MyDrive
```

3. Understanding the data



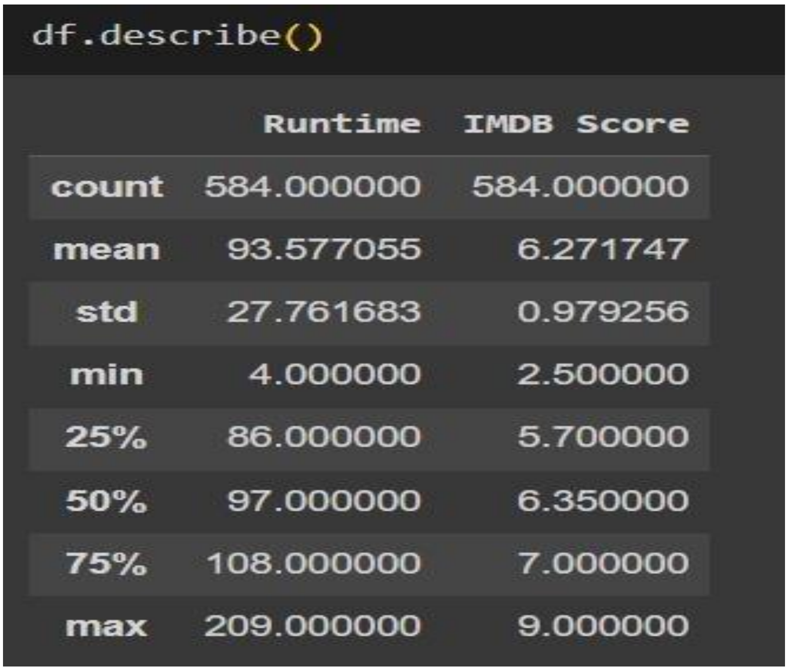
```
df.head()
```

	Title	Genre	Premiere	Runtime	IMDB Score	Language
0	Enter the Anime	Documentary	August 5, 2019	58	2.5	English/Japanese
1	Dark Forces	Thriller	August 21, 2020	81	2.6	Spanish
2	The App	Science fiction/Drama	December 26, 2019	79	2.6	Italian
3	The Open House	Horror thriller	January 19, 2018	94	3.2	English
4	Kaali Khuhi	Mystery	October 30, 2020	90	3.4	Hindi



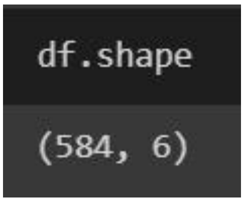
```
df.tail()
```

	Title	Genre	Premiere	Runtime	IMDB Score	Language
579	Taylor Swift: Reputation Stadium Tour	Concert Film	December 31, 2018	125	8.4	English
580	Winter on Fire: Ukraine's Fight for Freedom	Documentary	October 9, 2015	91	8.4	English/Ukranian/Russian
581	Springsteen on Broadway	One-man show	December 16, 2018	153	8.5	English
582	Emicida: AmarElo - It's All For Yesterday	Documentary	December 8, 2020	89	8.6	Portuguese
583	David Attenborough: A Life on Our Planet	Documentary	October 4, 2020	83	9.0	English



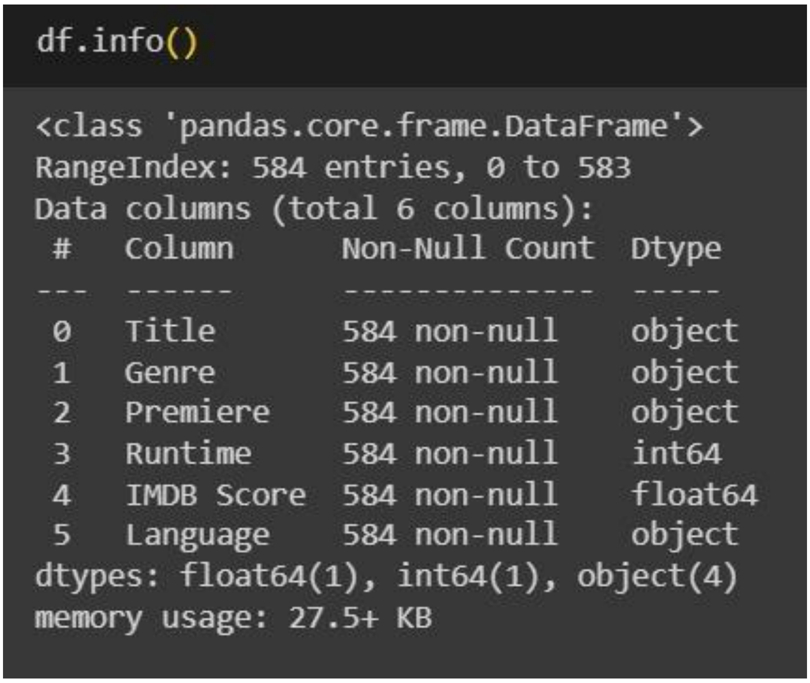
```
df.describe()
```

	Runtime	IMDB Score
count	584.000000	584.000000
mean	93.577055	6.271747
std	27.761683	0.979256
min	4.000000	2.500000
25%	86.000000	5.700000
50%	97.000000	6.350000
75%	108.000000	7.000000
max	209.000000	9.000000



```
df.shape
```

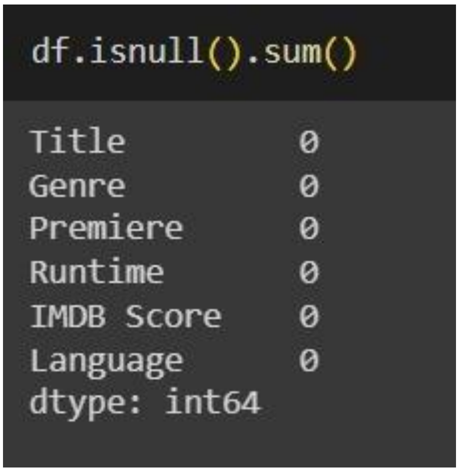
```
(584, 6)
```



```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 584 entries, 0 to 583
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   Title       584 non-null    object 
 1   Genre       584 non-null    object 
 2   Premiere    584 non-null    object 
 3   Runtime     584 non-null    int64  
 4   IMDB Score  584 non-null    float64
 5   Language    584 non-null    object 
dtypes: float64(1), int64(1), object(4)
memory usage: 27.5+ KB
```

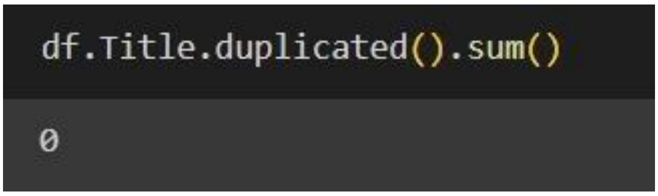
4.Checking for null values



```
df.isnull().sum()
```

```
Title          0
Genre          0
Premiere       0
Runtime        0
IMDB Score     0
Language       0
dtype: int64
```

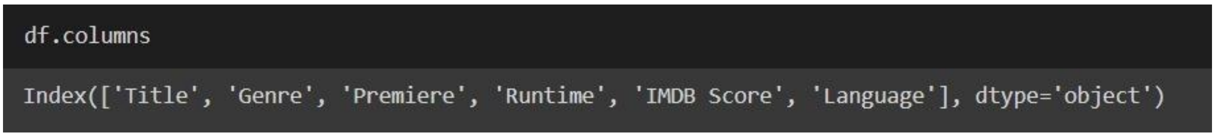
5.Checking for duplicate data



```
df.Title.duplicated().sum()
```

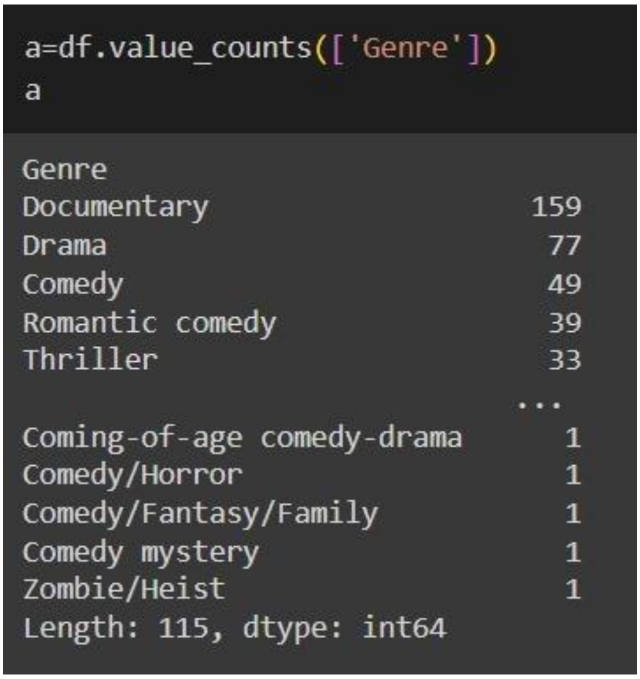
```
0
```

6.Performing analysis



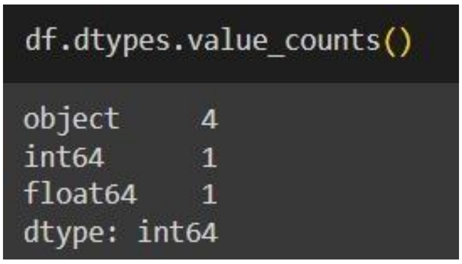
```
df.columns
```

```
Index(['Title', 'Genre', 'Premiere', 'Runtime', 'IMDB Score', 'Language'], dtype='object')
```



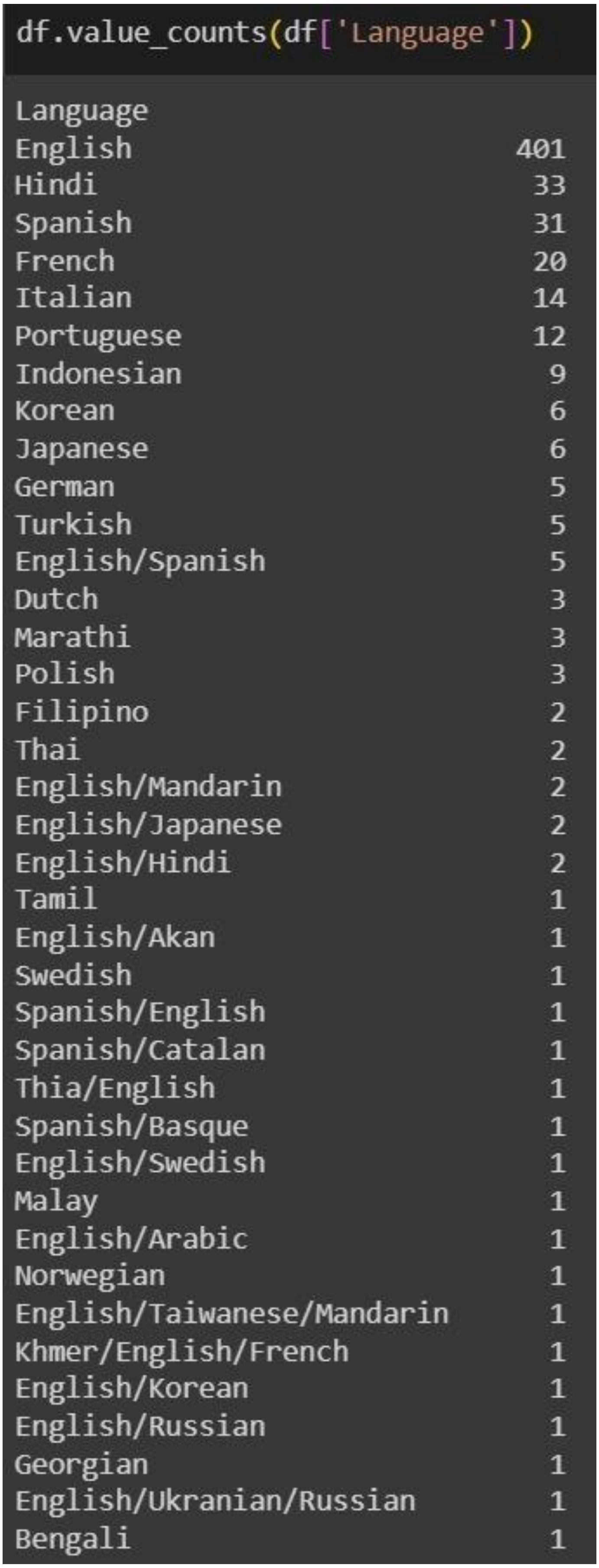
```
a=df.value_counts(['Genre'])
a
```

```
Genre
Documentary                 159
Drama                        77
Comedy                       49
Romantic comedy              39
Thriller                     33
                           ... 
Coming-of-age comedy-drama    1
Comedy/Horror                 1
Comedy/Fantasy/Family         1
Comedy mystery                1
Zombie/Heist                  1
Length: 115, dtype: int64
```



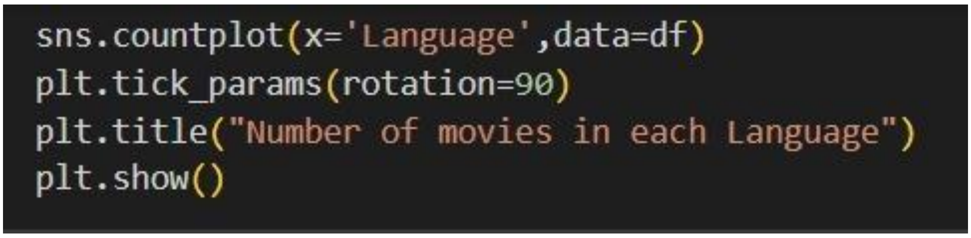
```
df.dtypes.value_counts()
```

```
object     4
int64      1
float64    1
dtype: int64
```

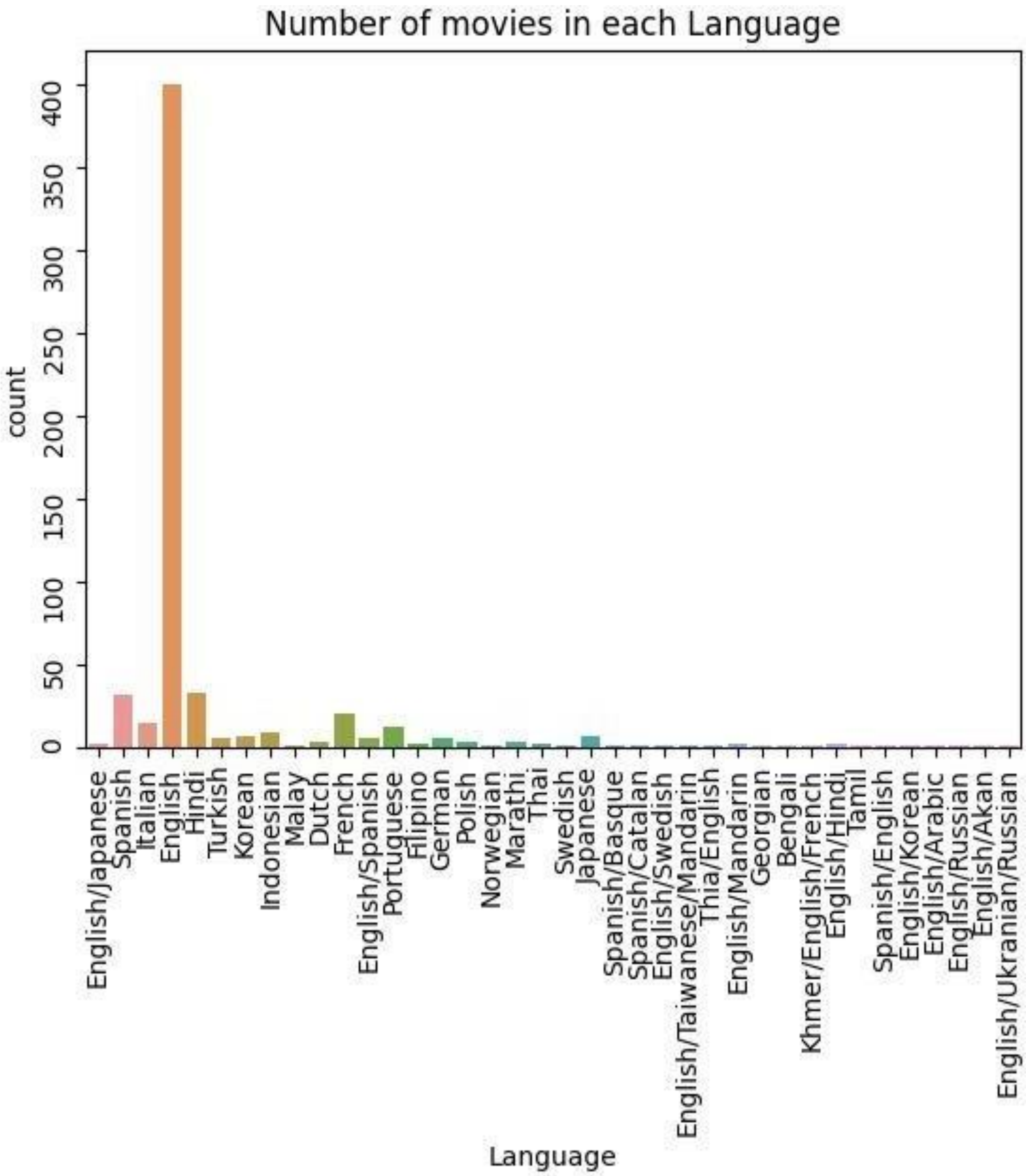


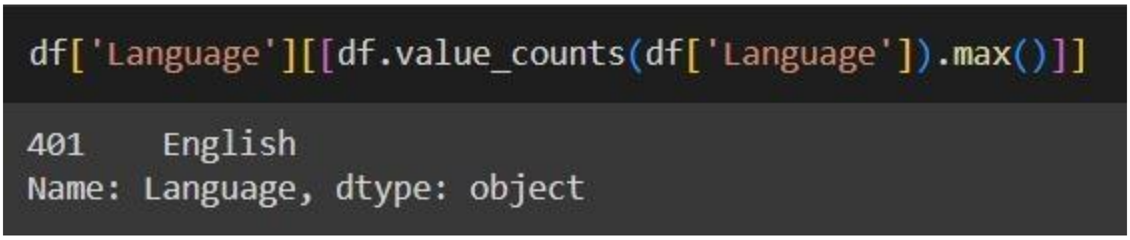
```
df.value_counts(df['Language'])
```

```
Language
English                       401
Hindi                          33
Spanish                        31
French                         20
Italian                        14
Portuguese                     12
Indonesian                      9
Korean                          6
Japanese                        6
German                          5
Turkish                         5
English/Spanish                 5
Dutch                           3
Marathi                         3
Polish                          3
Filipino                        2
Thai                            2
English/Mandarin                2
English/Japanese                2
English/Hindi                   2
Tamil                           1
English/Akan                    1
Swedish                         1
Spanish/English                 1
Spanish/Catalan                 1
Thia/English                    1
Spanish/Basque                  1
English/Swedish                 1
Malay                           1
English/Arabic                  1
Norwegian                       1
English/Taiwanese/Mandarin      1
Khmer/English/French            1
English/Korean                  1
English/Russian                 1
Georgian                        1
English/Ukranian/Russian        1
Bengali                         1
```



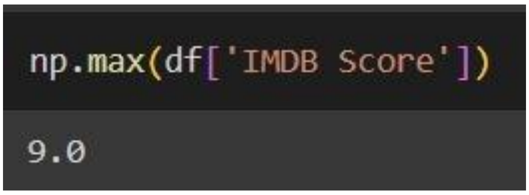
```
sns.countplot(x='Language',data=df)
plt.tick_params(rotation=90)
plt.title("Number of movies in each Language")
plt.show()
```





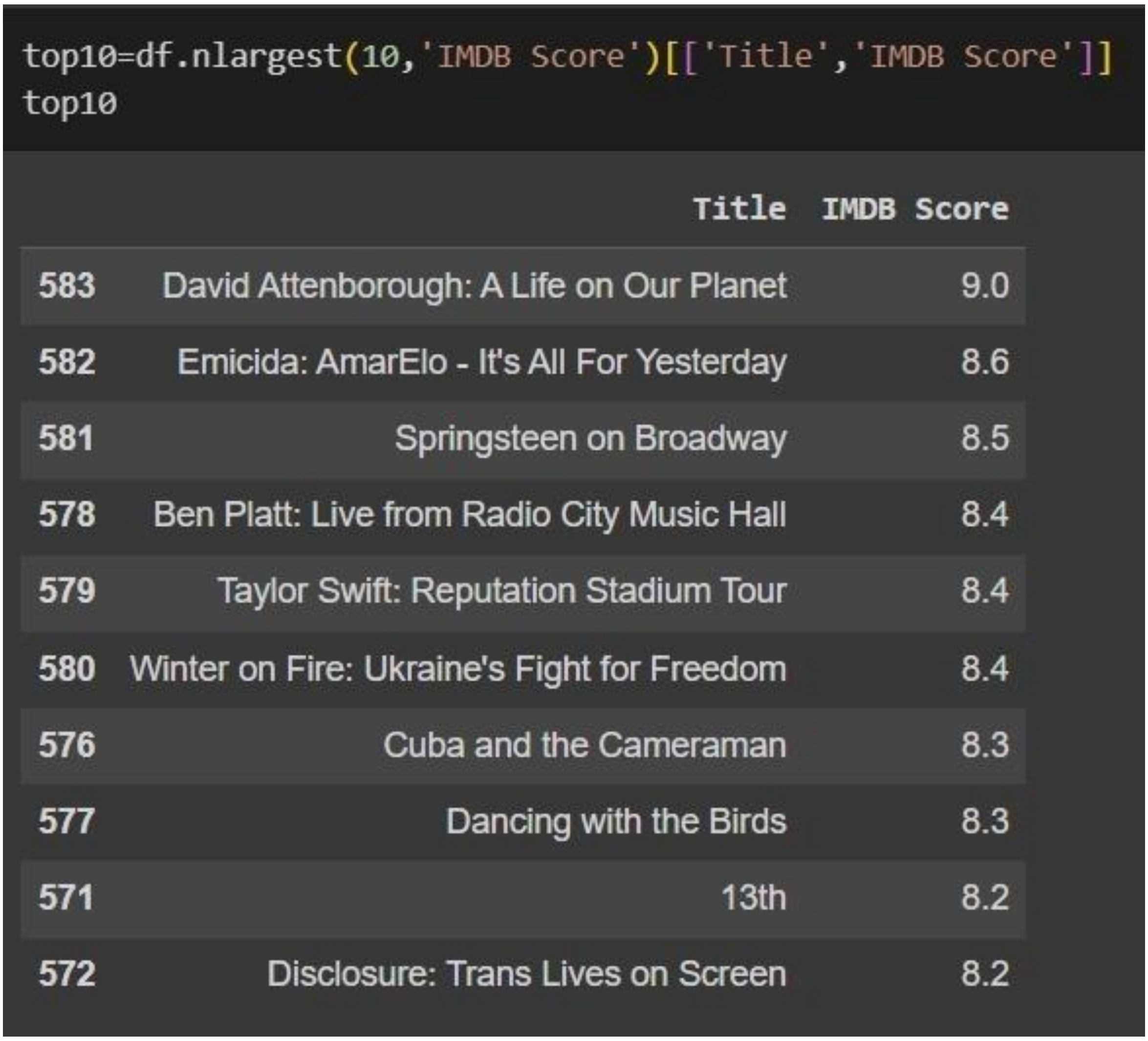
```
df['Language'][[df.value_counts(df['Language']).max()]]
```

```
401    English
Name: Language, dtype: object
```



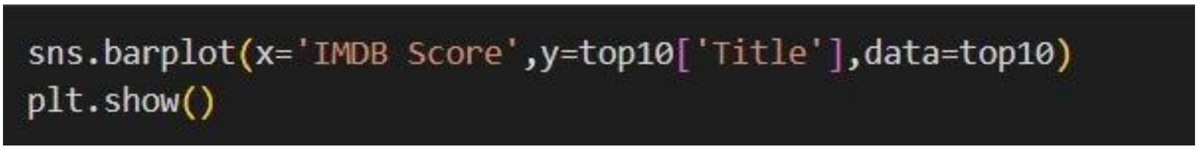
```
np.max(df['IMDB Score'])
```

```
9.0
```

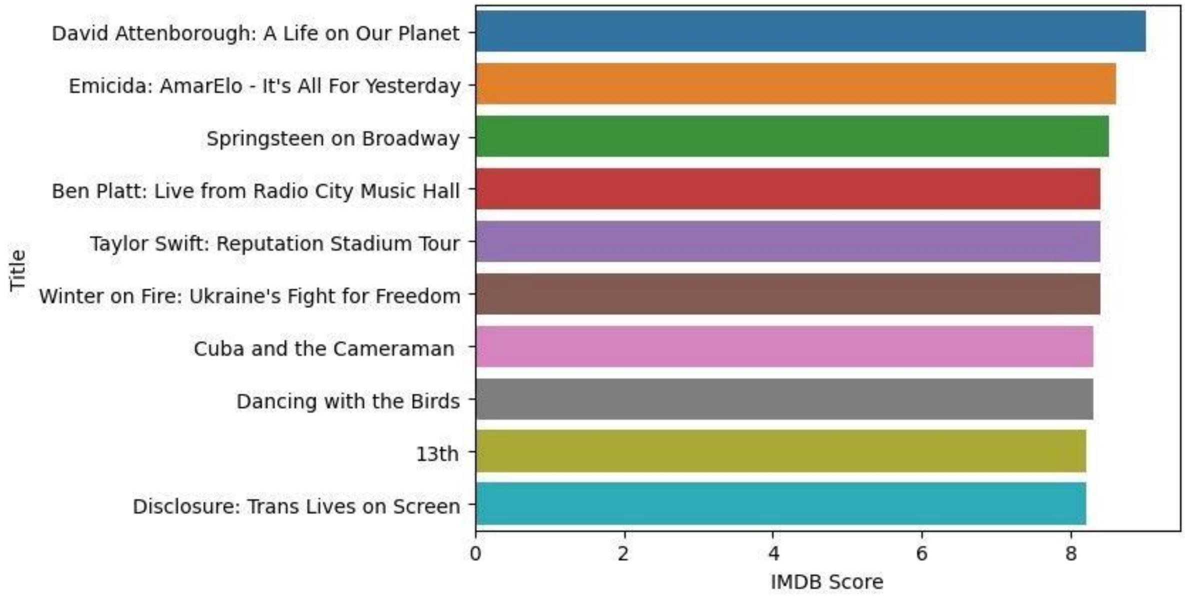


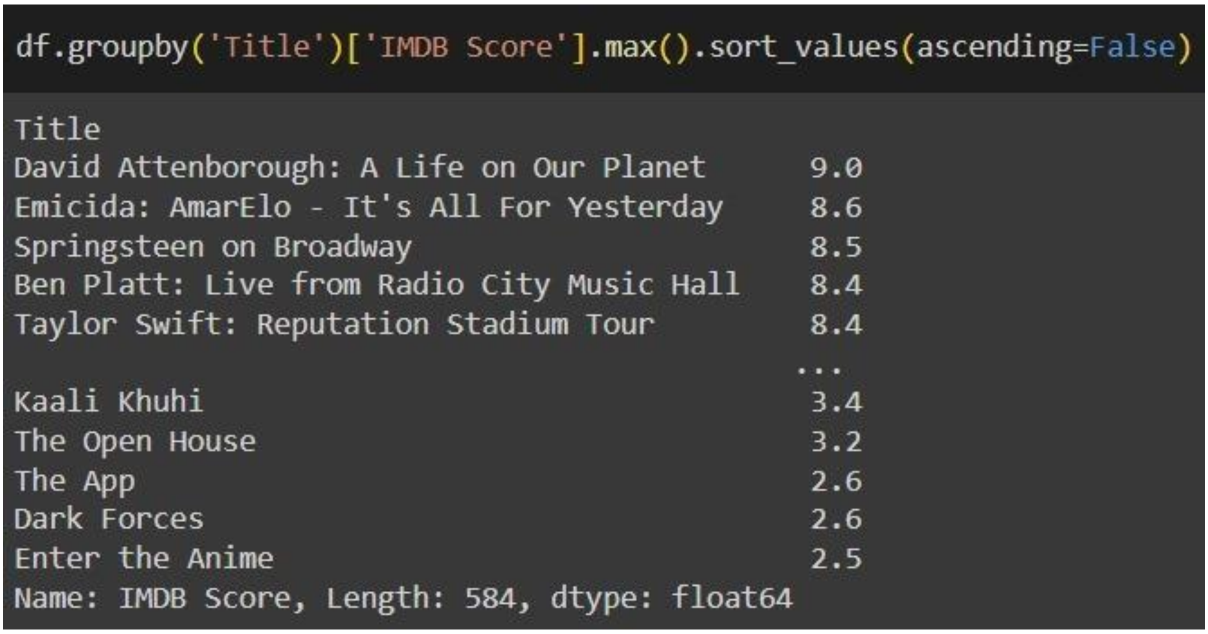
```
top10=df.nlargest(10,'IMDB Score')[['Title','IMDB Score']]
top10
```

	Title	IMDB Score
583	David Attenborough: A Life on Our Planet	9.0
582	Emicida: AmarElo - It's All For Yesterday	8.6
581	Springsteen on Broadway	8.5
578	Ben Platt: Live from Radio City Music Hall	8.4
579	Taylor Swift: Reputation Stadium Tour	8.4
580	Winter on Fire: Ukraine's Fight for Freedom	8.4
576	Cuba and the Cameraman	8.3
577	Dancing with the Birds	8.3
571	13th	8.2
572	Disclosure: Trans Lives on Screen	8.2



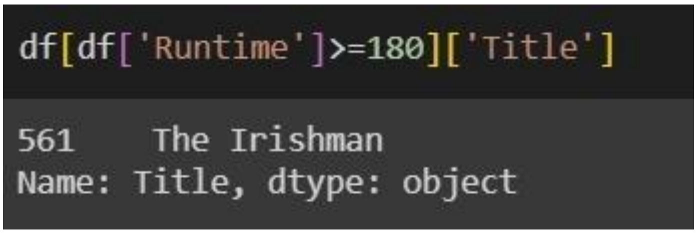
```
sns.barplot(x='IMDB Score',y=top10['Title'],data=top10)
plt.show()
```





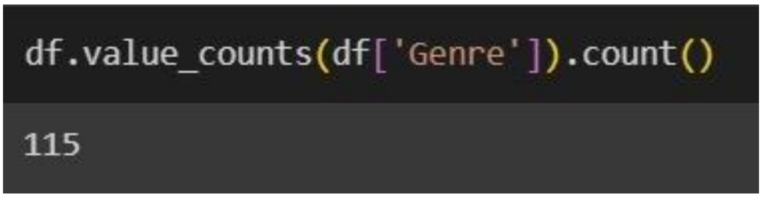
```
df.groupby('Title')['IMDB Score'].max().sort_values(ascending=False)
```

```
Title
David Attenborough: A Life on Our Planet      9.0
Emicida: AmarElo - It's All For Yesterday     8.6
Springsteen on Broadway                       8.5
Ben Platt: Live from Radio City Music Hall    8.4
Taylor Swift: Reputation Stadium Tour         8.4
                                             ...
Kaali Khuhi                                   3.4
The Open House                                3.2
The App                                       2.6
Dark Forces                                   2.6
Enter the Anime                               2.5
Name: IMDB Score, Length: 584, dtype: float64
```



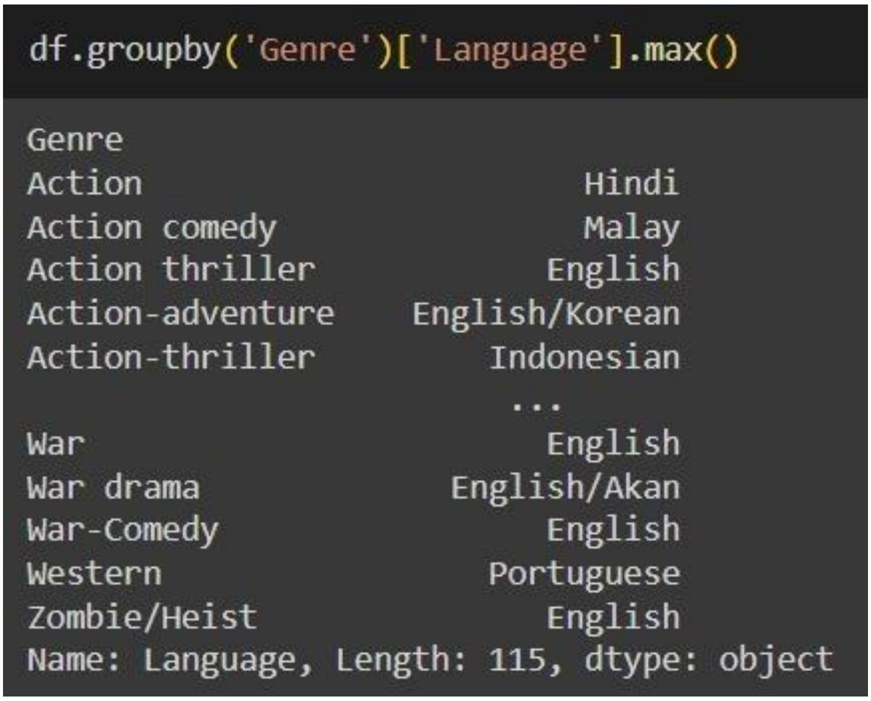
```
df[df['Runtime']>=180]['Title']
```

```
561    The Irishman
Name: Title, dtype: object
```



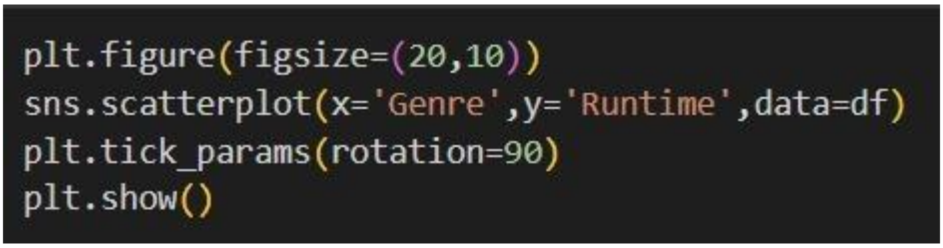
```
df.value_counts(df['Genre']).count()
```

```
115
```

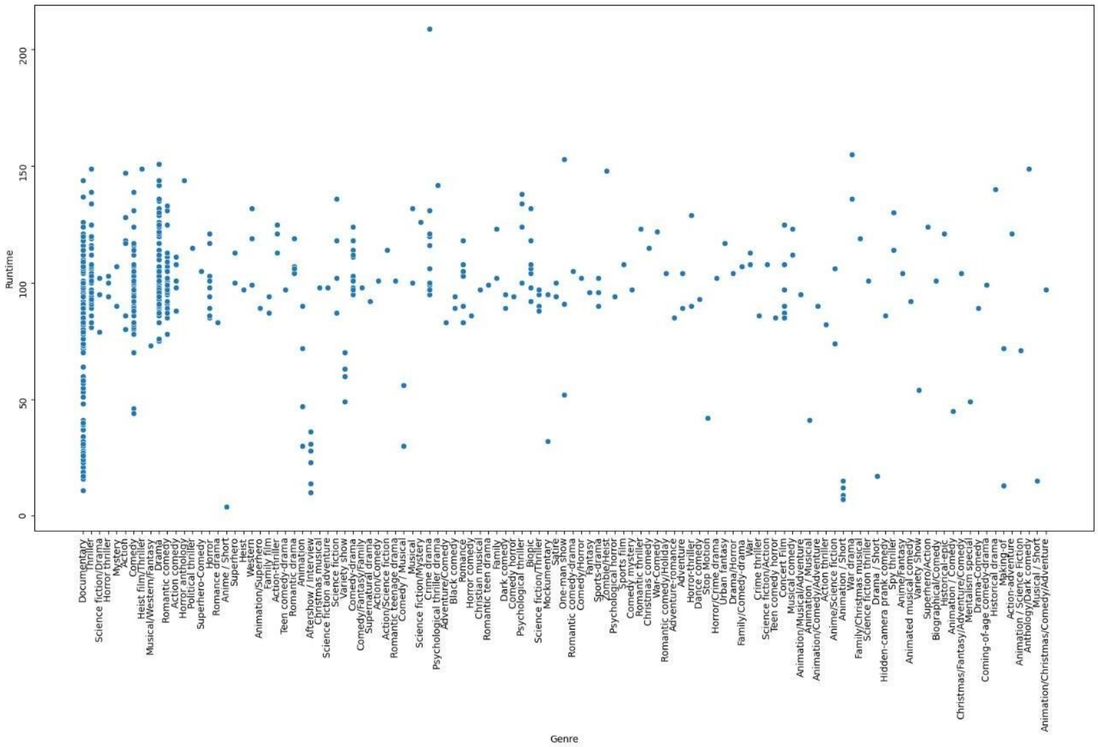


```
df.groupby('Genre')['Language'].max()
```

```
Genre
Action                           Hindi
Action comedy                    Malay
Action thriller                English
Action-adventure        English/Korean
Action-thriller             Indonesian
                             ...
War                            English
War drama                 English/Akan
War-Comedy                     English
Western                     Portuguese
Zombie/Heist                   English
Name: Language, Length: 115, dtype: object
```



```
plt.figure(figsize=(20,10))
sns.scatterplot(x='Genre',y='Runtime',data=df)
plt.tick_params(rotation=90)
plt.show()
```



Conclusion

Data preprocessing ensures that the data is clean, consistent, and suitable for the tasks at hand. By following the steps outlined in this report and using the appropriate techniques and libraries, the given dataset has been pre-processed in python for accurate and meaningful analysis.