

PHASE 3: DEVELOPMENT PART 1

PREDICTING IMDB SCORES

INTRODUCTION

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.

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)

Implementation

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.

Code

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 dataset

```
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/Ukrainian/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

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

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

```
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
```

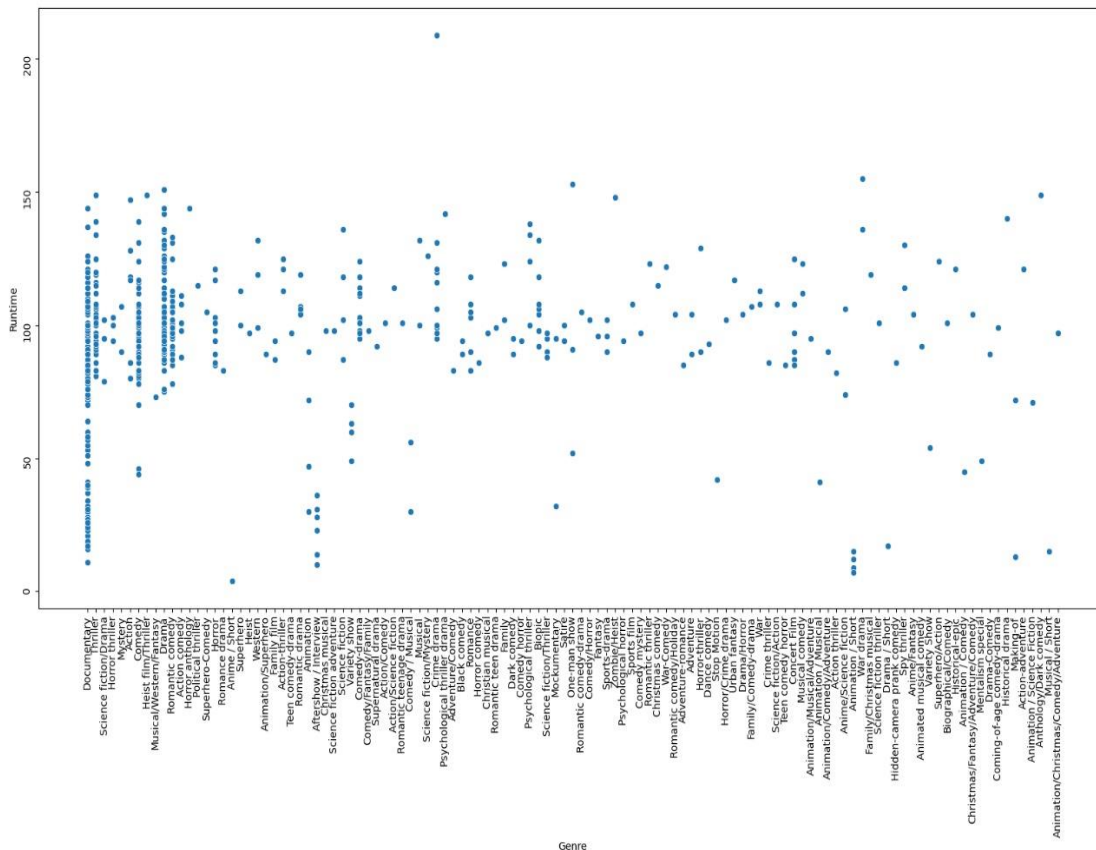
```
df.value_counts(df['Language'])
```

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

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

9.0

```
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.