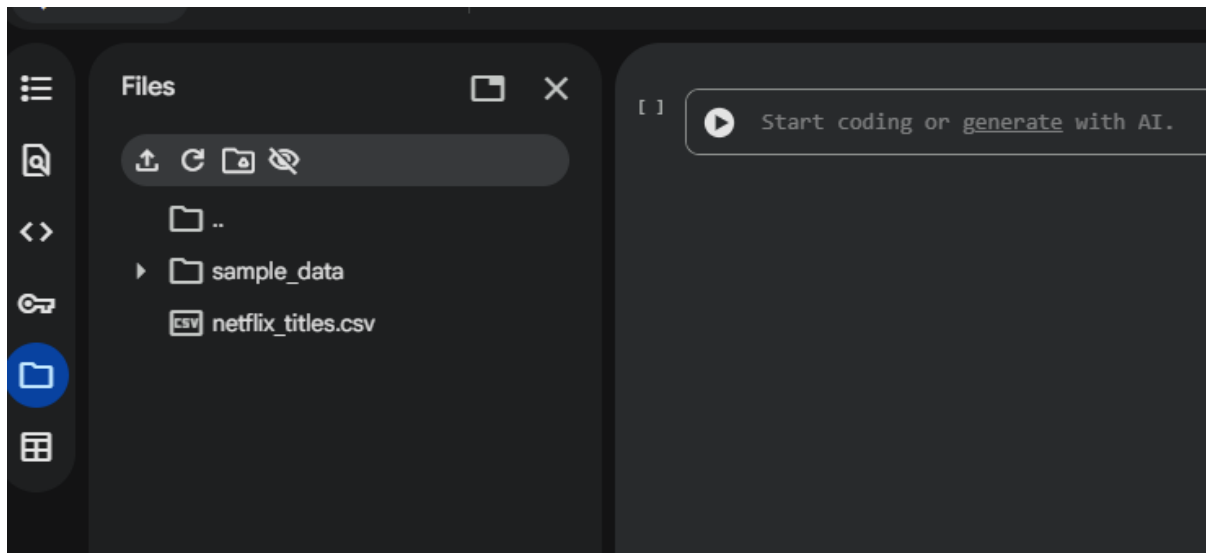


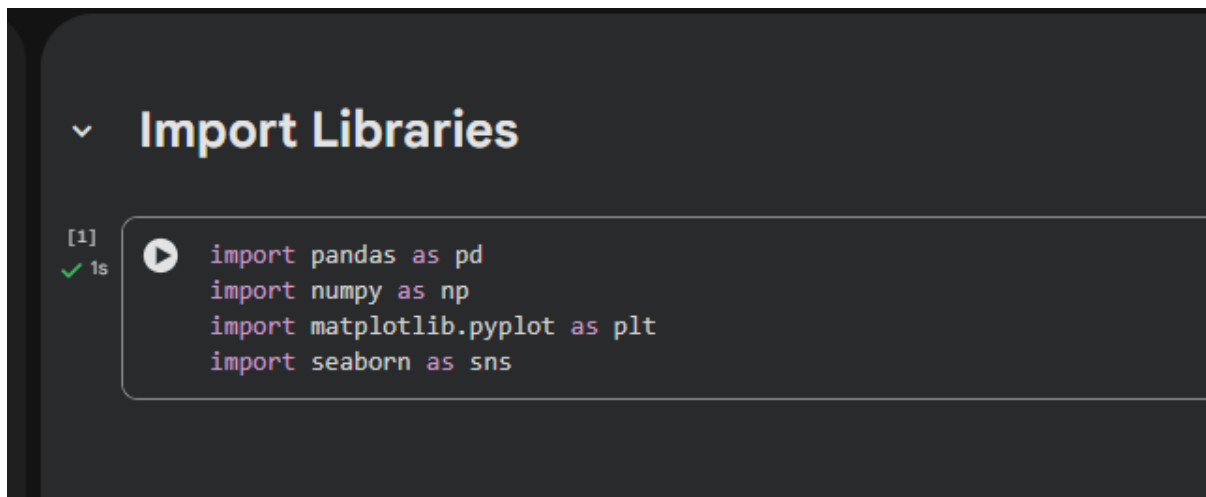
Name: **Sumbul**
Abdul wahid

Data Science
(internship)

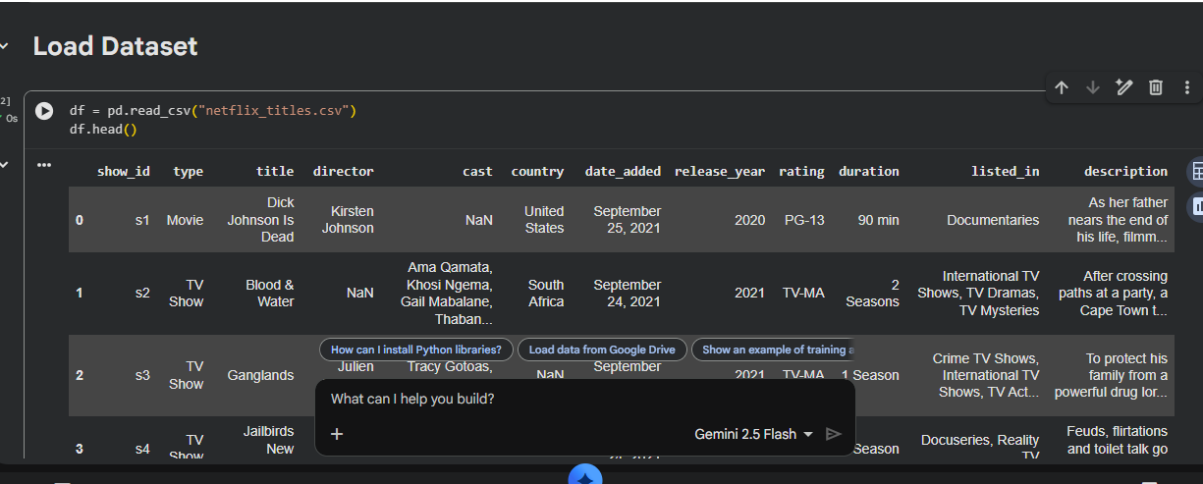
Introduction: In this task, advanced data cleaning and preprocessing techniques were applied to a real-world dataset as part of Week 1 of the Data Science Internship. The goal was to transform raw and messy data into a clean, analysis-ready format using Python



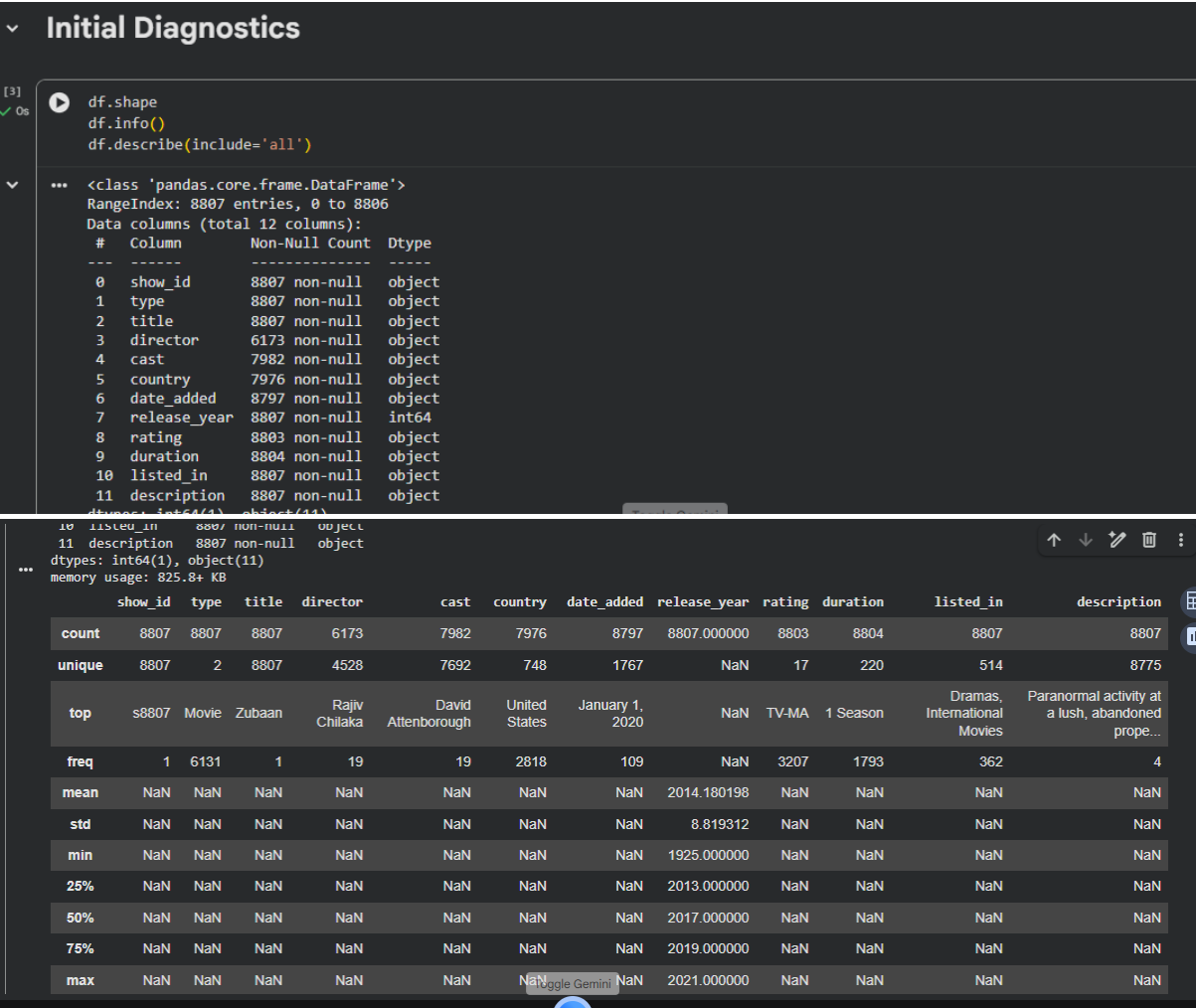
Import Libraries:



Dataset Description: The Netflix Movies and TV Shows dataset was selected for this task. It contains information such as title, type, director, cast, country, release year, rating, and duration. The dataset consists of more than 8,000 records and includes missing values and inconsistent data types.



Initial Diagnostics:Initial data inspection was performed to understand the structure, size, and data types of the dataset. This step helped identify missing values and columns requiring preprocessing.



Name: **Sumbul**
Abdul wahid

Data Science
(internship)

Missing Value Analysis: Missing values were analyzed using both tabular and visual methods. Several columns such as director, cast, country, and rating contained missing data.

Missing Values Analysis

[4]
✓ Os

df.isnull().sum()

	0
show_id	0
type	0
title	0
director	2634
cast	825
country	831
date_added	10
release_year	0
rating	4
duration	3

	0
show_id	0.000000
type	0.000000
title	0.000000
director	0.000000
cast	0.000000
country	0.000000
date_added	0.000000
release_year	0.000000
rating	0.045418
duration	0.000000
listed_in	0.000000
description	0.000000

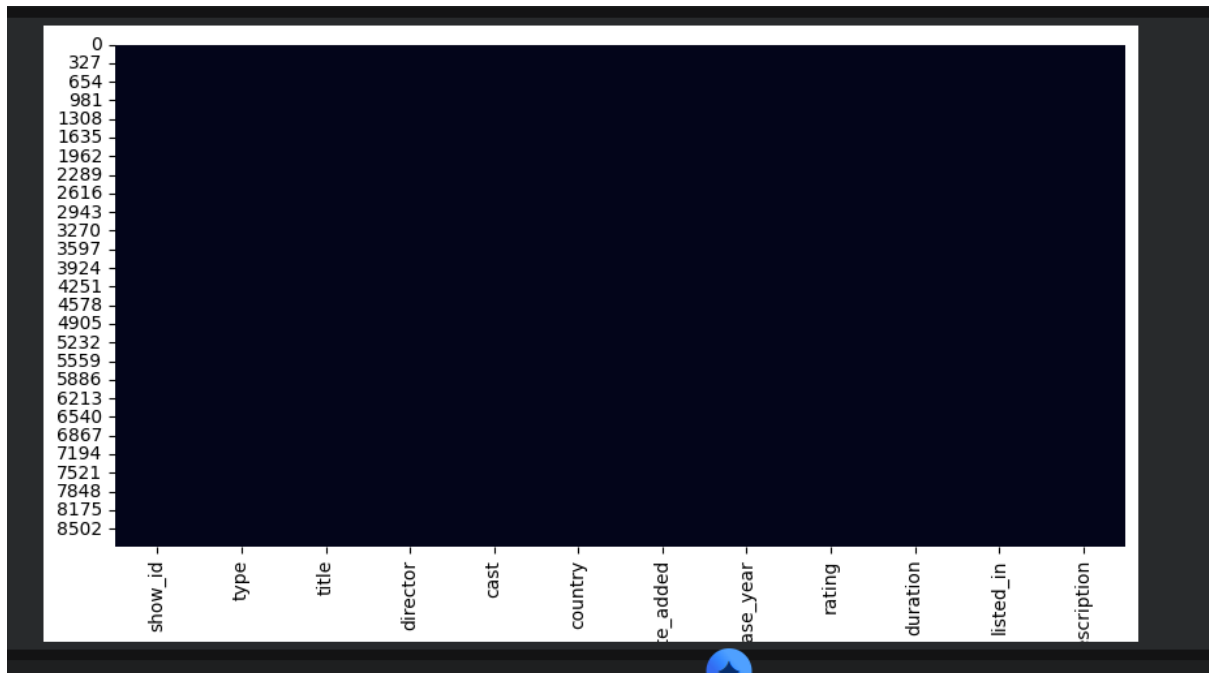
dtype: float64

Name: **Sumbul**

Abdul wahid

Data Science

(internship)



Missing Value Imputation: Two different imputation techniques were applied. Statistical methods such as mean and mode were used for numerical and categorical features. Additionally, KNN Imputation was applied as an advanced machine-learning-based technique to handle missing values more effectively.

Name: **Sumbul**

Abdul wahid

Data Science

(internship)

```
Mean Imputation

[ ] df['release_year'].fillna(df['release_year'].mean(), inplace=True)

... /tmp/ipython-input-1443942658.py:1: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an in
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behave
For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) inst

df['release_year'].fillna(df['release_year'].mean(), inplace=True)

Mode Imputation

[ ] df['rating'].fillna(df['rating'].mode()[0], inplace=True)

... /tmp/ipython-input-2284108631.py:1: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an in
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behave
For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) inst

df['rating'].fillna(df['rating'].mode()[0], inplace=True)

Missing value Imputation(Method 2 - KNN)

[9]
✓ 0s from sklearn.impute import KNNImputer

imputer = KNNImputer(n_neighbors=5)
df[['release_year']] = imputer.fit_transform(df[['release_year']])

[10]
✓ 0s df['release_year'].isnull().sum()

np.int64(0)
```

Data Type Correction & Date Parsing: Data type corrections were performed to ensure consistency across the dataset. The date_added column was converted into a proper datetime format for better usability.

Name: **Sumbul**

Abdul wahid

Data Science

(internship)

```
▼ Data Parsing Data Type Correction

[ ] df['date_added'] = pd.to_datetime(df['date_added'], errors='coerce')
df.info()

▼ <class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 12 columns):
#   Column          Non-Null Count  Dtype
---  -
0   show_id         8807 non-null   object
1   type            8807 non-null   object
2   title           8807 non-null   object
3   director        8807 non-null   object
4   cast            8807 non-null   object
5   country         8807 non-null   object
6   date_added      8807 non-null   datetime64[ns]
7   release_year    8807 non-null   float64
8   rating          8807 non-null   object
9   duration        8807 non-null   object
10  listed_in       8807 non-null   object
11  description     8807 non-null   object
dtypes: datetime64[ns](1), float64(1), object(10)
memory usage: 825.8+ KB
```

Outlier Detection & Treatment: Outliers in the release_year column were detected using the Interquartile Range (IQR) method. Identified outliers were removed to improve data quality and reliability.

```
▼ IQR Method(Outlier Detection)

[13] ✓ Os Q1 = df['release_year'].quantile(0.25)
Q3 = df['release_year'].quantile(0.75)

IQR = Q3 - Q1
lower = Q1 - 1.5 * IQR
upper = Q3 + 1.5 * IQR

▼ Outlier Removal

[14] ✓ Os ▶ df = df[(df['release_year'] >= lower) & (df['release_year'] <= upper)]
df.shape

▼ ... (8088, 12)
```

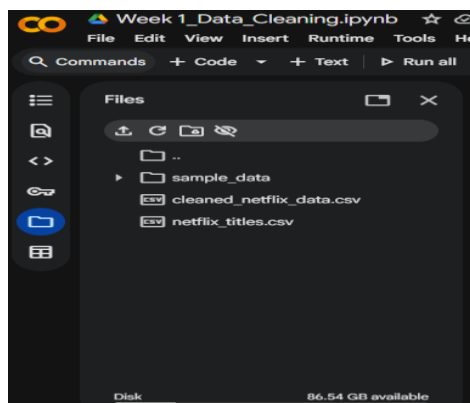
Name: **Sumbul**
Abdul wahid

Data Science
(internship)

Final Cleaned Dataset: After applying all cleaning and preprocessing steps, a final clean dataset was obtained. Sample rows were reviewed to verify the results.

```
df.head()
df.tail()
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	2019-11-20	2007.0	R	158 min	Cult Movies, Dramas, Thrillers	A political cartoonist, a crime reporter and a...
8803	s8804	TV Show	Zombie Dumb	Unknown	Not Available	Unknown	2019-07-01	2018.0	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies	While living alone in a spooky town, a young g...
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	2019-11-01	2009.0	R	88 min	Comedies, Horror Movies	Looking to survive in a world taken over by zo...
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	2020-01-11	2006.0	PG	88 min	Children & Family Movies, Comedies	Dragged from civilian life, a former superhero...
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chahan...	India	2019-03-02	2015.0	TV-14	111 min	Dramas, International Movies, Music & Musicals	A scrappy but poor boy worms his way into a ty...



Conclusion: This task provided hands-on experience in handling real-world data challenges. Advanced preprocessing techniques were successfully applied, resulting in a clean and structured dataset ready for further analysis.