

```
import pandas as pd
import numpy as np
```

```
# Load dataset
df = pd.read_csv("/content/tmdb_5000_movies.csv")

# View first 5 rows
df.head()
```

	budget	genres	homepage	id	keywords	original_language	original_title	overview
0	237000000	[[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}], [{"id": 14, "name": "Fantasy"}]	http://www.avatarmovie.com/	19995	[[{"id": 1463, "name": "culture clash"}, {"id": 1464, "name": "marine"}], [{"id": 1465, "name": "avatar"}]]	en	Avatar	In the 22nd century, a paraplegic Marine is di...
1	300000000	[[{"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}], [{"id": 14, "name": "Fantasy"}]	http://disney.go.com/disneypictures/pirates/	285	[[{"id": 270, "name": "ocean"}, {"id": 726, "name": "pirates"}], [{"id": 727, "name": "pirates of the caribbean"}]]	en	Pirates of the Caribbean: At World's End	Captain Barbosa, long believed to be dead, ha...
2	245000000	[[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}], [{"id": 14, "name": "Fantasy"}]	http://www.sonypictures.com/movies/spectre/	206647	[[{"id": 470, "name": "spy"}, {"id": 818, "name": "bond"}], [{"id": 819, "name": "james bond"}]]	en	Spectre	A cryptic message from Bond's past sends him o...
3	250000000	[[{"id": 28, "name": "Action"}, {"id": 80, "name": "Crime"}], [{"id": 80, "name": "Crime"}]	http://www.thedarkknightises.com/	49026	[[{"id": 849, "name": "dc comics"}, {"id": 853, "name": "batman"}], [{"id": 854, "name": "the dark knight rises"}]]	en	The Dark Knight Rises	Following the death of District Attorney Harve...
4	260000000	[[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}], [{"id": 14, "name": "Fantasy"}]	http://movies.disney.com/john-carter	49529	[[{"id": 818, "name": "based on novel"}, {"id": 819, "name": "john carter"}], [{"id": 820, "name": "john carter"}]]	en	John Carter	John Carter is a war-weary, former military ca...

```
# Shape of dataset
print("Shape:", df.shape)

# Column names
print("\nColumns:")
print(df.columns)

# Info about dataset
print("\nDataset Info:")
print(df.info())

# Statistical summary
print("\nStatistical Summary:")
print(df.describe())
```

```
Columns:
Index(['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'],
      dtype='object')
```

Dataset Info:

```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 4803 entries, 0 to 4802
```

```
Data columns (total 20 columns):
```

#	Column	Non-Null Count	Dtype
0	budget	4803 non-null	int64
1	genres	4803 non-null	object
2	homepage	1712 non-null	object
3	id	4803 non-null	int64
4	keywords	4803 non-null	object
5	original_language	4803 non-null	object
6	original_title	4803 non-null	object
7	overview	4800 non-null	object
8	popularity	4803 non-null	float64
9	production_companies	4803 non-null	object
10	production_countries	4803 non-null	object

```

13 runtime                4801 non-null float64
14 spoken_languages       4803 non-null object
15 status                 4803 non-null object
16 tagline                3959 non-null object
17 title                  4803 non-null object
18 vote_average           4803 non-null float64
19 vote_count             4803 non-null int64

```

dtypes: float64(3), int64(4), object(13)

memory usage: 750.6+ KB

None

Statistical Summary:

	budget	id	popularity	revenue	runtime \
count	4.803000e+03	4803.000000	4803.000000	4.803000e+03	4801.000000
mean	2.904504e+07	57165.484281	21.492301	8.226064e+07	106.875859
std	4.072239e+07	88694.614033	31.816650	1.628571e+08	22.611935
min	0.000000e+00	5.000000	0.000000	0.000000e+00	0.000000
25%	7.900000e+05	9014.500000	4.668070	0.000000e+00	94.000000
50%	1.500000e+07	14629.000000	12.921594	1.917000e+07	103.000000
75%	4.000000e+07	58610.500000	28.313505	9.291719e+07	118.000000
max	3.800000e+08	459488.000000	875.581305	2.787965e+09	338.000000

	vote_average	vote_count
count	4803.000000	4803.000000
mean	6.092172	690.217989
std	1.194612	1234.585891
min	0.000000	0.000000
25%	5.600000	54.000000
50%	6.200000	235.000000
75%	6.800000	737.000000
max	10.000000	13752.000000

Inference:

shape → total rows & columns

info() → data types + null values

describe() → numerical statistics

This helps us understand what cleaning is required.

```
# Count missing values
print("\nMissing Values:")
print(df.isnull().sum())
```

```
Missing Values:
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
```

```
numeric_cols = ['budget', 'popularity', 'revenue',
                'runtime', 'vote_average', 'vote_count']

for col in numeric_cols:
    df[col] = pd.to_numeric(df[col], errors='coerce')
```

Inference:

Converts columns into numeric format

Invalid values become NaN

```
df['homepage'] = df['homepage'].fillna('')  
df['tagline'] = df['tagline'].fillna('')  
df['overview'] = df['overview'].fillna('')
```

Inference:

Text columns can remain empty instead of NaN.

```
df['runtime'] = df['runtime'].fillna(df['runtime'].median())  
df['budget'] = df['budget'].fillna(0)  
df['revenue'] = df['revenue'].fillna(0)
```

Inference:

Runtime → use median (better than mean)

Budget & revenue → 0 (unknown values)

```
df.drop_duplicates(inplace=True)  
  
print("Shape after removing duplicates:", df.shape)
```

Shape after removing duplicates: (4803, 20)

```
df['release_date'] = pd.to_datetime(df['release_date'], errors='coerce')  
  
# Extract release year  
df['release_year'] = df['release_date'].dt.year
```

Inference:

Converts string date to datetime

Extracts year for analysis

```
df['profit'] = df['revenue'] - df['budget']
```

```
df['roi'] = 0  
df.loc[df['budget'] > 0, 'roi'] = df['profit'] / df['budget']
```

```
/tmp/ipython-input-2493819983.py:2: FutureWarning: Setting an item of incompatible dtype is deprecated and will raise an error  
-1.          ]' has dtype incompatible with int64, please explicitly cast to a compatible dtype first.  
df.loc[df['budget'] > 0, 'roi'] = df['profit'] / df['budget']
```

Inference:

Profit → Revenue - Budget

ROI → Profit / Budget

Helps measure movie success

```
# Remove movies with negative values  
df = df[(df['budget'] >= 0) & (df['revenue'] >= 0)]  
  
# Remove very high revenue outliers  
upper_limit = df['revenue'].quantile(0.99)  
df = df[df['revenue'] <= upper_limit]
```

Inference:

Removes unrealistic data

Keeps dataset balanced

```
df.drop(['id', 'imdb_id'], axis=1, inplace=True, errors='ignore')
```

```
print("Final Shape:", df.shape)

print("\nFinal Missing Values:")
print(df.isnull().sum())
```

Final Shape: (4754, 22)

Final Missing Values:

budget	0
genres	0
homepage	0
keywords	0
original_language	0
original_title	0
overview	0
popularity	0
production_companies	0
production_countries	0
release_date	1
revenue	0
runtime	0
spoken_languages	0
status	0
tagline	0
title	0
vote_average	0
vote_count	0
release_year	1
profit	0
roi	0
dtype: int64	

OLAP Operations

```
import pandas as pd
import numpy as np

# Load cleaned dataset (or original if not saved)
df = pd.read_csv("/content/tmdb_5000_movies.csv")
```



```
# Convert numeric columns
numeric_cols = ['budget', 'popularity', 'revenue',
                'runtime', 'vote_average', 'vote_count']

for col in numeric_cols:
    df[col] = pd.to_numeric(df[col], errors='coerce')

# Convert release_date
df['release_date'] = pd.to_datetime(df['release_date'], errors='coerce')
df['release_year'] = df['release_date'].dt.year
df['release_month'] = df['release_date'].dt.month

# Fill missing values
df['budget'].fillna(0, inplace=True)
df['revenue'].fillna(0, inplace=True)
df['runtime'].fillna(df['runtime'].median(), inplace=True)
```

/tmp/ipython-input-1373664512.py:20: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chainable assignment. This behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are se

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] =

```
df['budget'].fillna(0, inplace=True)
```

/tmp/ipython-input-1373664512.py:21: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chainable assignment. This behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are se

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] =

```
df['revenue'].fillna(0, inplace=True)
```

/tmp/ipython-input-1373664512.py:22: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chainable assignment. This behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are se

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] =

```
df['runtime'].fillna(df['runtime'].median(), inplace=True)
```

Movies released in 2015

```
slice_2015 = df[df['release_year'] == 2015]

print("Movies in 2015:", slice_2015.shape)
slice_2015[['title', 'revenue', 'vote_average']].head()
```

Movies in 2015: (216, 22)

	title	revenue	vote_average
2	Spectre	880674609	6.3
7	Avengers: Age of Ultron	1405403694	7.3
28	Jurassic World	1513528810	6.5
44	Furious 7	1506249360	7.3
54	The Good Dinosaur	331926147	6.6

Released between 2010–2015

Revenue > 100M

Rating > 7

```
dice_movies = df[
    (df['release_year'] >= 2010) &
    (df['release_year'] <= 2015) &
    (df['revenue'] > 100000000) &
    (df['vote_average'] > 7)
]

print("Filtered Movies:", dice_movies.shape)
dice_movies[['title', 'release_year', 'revenue', 'vote_average']].head()
```

Filtered Movies: (75, 22)

	title	release_year	revenue	vote_average
3	The Dark Knight Rises	2012.0	1084939099	7.6
6	Tangled	2010.0	591794936	7.4
7	Avengers: Age of Ultron	2015.0	1405403694	7.3
16	The Avengers	2012.0	1519557910	7.4
19	The Hobbit: The Battle of the Five Armies	2014.0	956019788	7.1

Total Revenue Per Year

```
rollup_year = df.groupby('release_year')['revenue'].sum().reset_index()
```

```
rollup_year.head()
```

	release_year	revenue
0	1916.0	8394751
1	1925.0	22000000
2	1927.0	650422
3	1929.0	4358000
4	1930.0	8000000

```
rollup_year_month = df.groupby(
    ['release_year', 'release_month']
)['revenue'].sum().reset_index()
```

```
rollup_year_month.head()
```

	release_year	release_month	revenue
0	1916.0	9.0	8394751
1	1925.0	11.0	22000000
2	1927.0	1.0	650422
3	1929.0	1.0	0
4	1929.0	2.0	4358000

Drill Down into 2015 → Month level

```
df[df['release_year'] == 2015] \
    .groupby('release_month')['revenue'].sum()
```

	revenue
release_month	
1.0	233825738
2.0	1561333527
3.0	1523987541
4.0	3286982086
5.0	1424127021
6.0	4528977085
7.0	2144512007
8.0	781355731
9.0	2413958848
10.0	1414719138
11.0	1765289874
12.0	1695955625

dtype: int64

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import MinMaxScaler, StandardScaler
from scipy.stats import chi2_contingency

# Load dataset
df = pd.read_csv("/content/tmdb_5000_movies.csv")

# Convert numeric columns
numeric_cols = ['budget', 'popularity', 'revenue',
                'runtime', 'vote_average', 'vote_count']

for col in numeric_cols:
```

```
for col in numeric_cols:
    df[col] = pd.to_numeric(df[col], errors='coerce')

# Handle missing values
df['budget'].fillna(0, inplace=True)
df['revenue'].fillna(0, inplace=True)
df['runtime'].fillna(df['runtime'].median(), inplace=True)
df['vote_average'].fillna(df['vote_average'].median(), inplace=True)

df.head()
```



Equal Width Binning

```
# Create 5 equal width bins
df['revenue_bin_equal_width'] = pd.cut(df['revenue'], bins=5)

df[['revenue', 'revenue_bin_equal_width']].head()
```

revenue

revenue_bin_equal_width

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

0	2787965087	(2230372069.6, 2787965087.0]
1	961000000	(557593017.4, 1115186034.8]
2	880674609	(557593017.4, 1115186034.8]
3	1084939099	(557593017.4, 1115186034.8]
4	284139100	(-2787965.087, 557593017.4]

eWarning: A value is trying to be set on a copy of a DataFrame or Series through cha
is inplace method will never work because the intermediate object on which we are se

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = Equal Frequency Binning

```
# Create 5 equal frequency bins
df['revenue_bin_equal_freq'] = pd.qcut(df['revenue'], q=5, duplicates='drop')

df[['revenue', 'revenue_bin_equal_freq']].head()
```

revenue

revenue_bin_equal_freq

df['vote_average'].fillna(df['vote_average'].median(), inplace=True)

Custom Binning

[[{"id": 12,

1	300000000	"name": "Adventure"},	http://disney.go.com/disneypictures/pirates/	285	"name": "ocean"},	en	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed
---	-----------	-----------------------	--	-----	-------------------	----	--	---------------------------------


```
bins = [0, 50000000, 200000000, 500000000, df['revenue'].max()]
labels = ['Low', 'Medium', 'High', 'Blockbuster']

df['revenue_custom_bin'] = pd.cut(df['revenue'], bins=bins, labels=labels)

df[['revenue', 'revenue_custom_bin']].head()
```

	revenue	revenue_custom_bin	name...	name...	sends him
0	2787965087	Blockbuster		["id": 849,	o...
1	961000000	Blockbuster	http://www.thedarkknighttrises.com/	"name": "dc	Following
2	880674609	Blockbuster	49026	comics"}, {"id": 853,	the death of District Attorney Harve...
3	1084939099	Blockbuster		818,	John
4	284139100	High		"name": "based on	Carter is a war-weary, former

Min-Max Normalization

```
min_max_scaler = MinMaxScaler()

df['budget_minmax'] = min_max_scaler.fit_transform(df[['budget']])

df[['budget', 'budget_minmax']].head()
```

	budget	budget_minmax
0	237000000	0.623684
1	300000000	0.789474
2	245000000	0.644737
3	250000000	0.657895
4	260000000	0.684211

Z-Score Normalization

```
standard_scaler = StandardScaler()

df['budget_zscore'] = standard_scaler.fit_transform(df[['budget']])

df[['budget', 'budget_zscore']].head()
```

	budget	budget_zscore
0	237000000	5.107181
1	300000000	6.654402
2	245000000	5.303653
3	250000000	5.426449
4	260000000	5.672039

CHI-SQUARE TEST

```
# Revenue Category
df['revenue_category'] = pd.cut(
    df['revenue'],
    bins=3,
    labels=['Low', 'Medium', 'High']
)

# Vote Category
df['vote_category'] = pd.cut(
    df['vote_average'],
    bins=3,
    labels=['Low Rating', 'Medium Rating', 'High Rating']
)
```

```
contingency_table = pd.crosstab(
```

```

    df['revenue_category'],
    df['vote_category']
)

print(contingency_table)

```

vote_category	Low Rating	Medium Rating	High Rating
revenue_category			
Low	121	3143	1509
Medium	0	6	23
High	0	0	1

```

chi2, p, dof, expected = chi2_contingency(contingency_table)

print("Chi-Square Value:", chi2)
print("P-Value:", p)
print("Degrees of Freedom:", dof)

```

```

Chi-Square Value: 32.379082433072554
P-Value: 1.600427004563815e-06
Degrees of Freedom: 4

```

$p < 0.05 \rightarrow$ Revenue and Rating are related

$p > 0.05 \rightarrow$ No significant relationship

```
!pip install mlxtend
```

```

Requirement already satisfied: mlxtend in /usr/local/lib/python3.12/dist-packages (0.23.4)
Requirement already satisfied: scipy>=1.2.1 in /usr/local/lib/python3.12/dist-packages (from mlxtend) (1.16.3)
Requirement already satisfied: numpy>=1.16.2 in /usr/local/lib/python3.12/dist-packages (from mlxtend) (2.0.2)
Requirement already satisfied: pandas>=0.24.2 in /usr/local/lib/python3.12/dist-packages (from mlxtend) (2.2.2)
Requirement already satisfied: scikit-learn>=1.3.1 in /usr/local/lib/python3.12/dist-packages (from mlxtend) (1.6.1)
Requirement already satisfied: matplotlib>=3.0.0 in /usr/local/lib/python3.12/dist-packages (from mlxtend) (3.10.0)
Requirement already satisfied: joblib>=0.13.2 in /usr/local/lib/python3.12/dist-packages (from mlxtend) (1.5.3)
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.12/dist-packages (from matplotlib>=3.0.0->mlxtend) (1.3.0)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.12/dist-packages (from matplotlib>=3.0.0->mlxtend) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.12/dist-packages (from matplotlib>=3.0.0->mlxtend) (4.53.0)
Requirement already satisfied: kiwisolver>=1.3.1 in /usr/local/lib/python3.12/dist-packages (from matplotlib>=3.0.0->mlxtend) (1.4.7)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.12/dist-packages (from matplotlib>=3.0.0->mlxtend) (24.1)

```

```
Requirement already satisfied: pillow>=8 in /usr/local/lib/python3.12/dist-packages (from matplotlib>=3.0.0->mlxtend) (11.3.0)
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.12/dist-packages (from matplotlib>=3.0.0->mlxtend) (
Requirement already satisfied: python-dateutil>=2.7 in /usr/local/lib/python3.12/dist-packages (from matplotlib>=3.0.0->mlxten
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.12/dist-packages (from pandas>=0.24.2->mlxtend) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.12/dist-packages (from pandas>=0.24.2->mlxtend) (2025.
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.12/dist-packages (from scikit-learn>=1.3.1->mlxt
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.12/dist-packages (from python-dateutil>=2.7->matplotlib>=3.0
```

```
import pandas as pd
import numpy as np
from mlxtend.frequent_patterns import apriori, association_rules
from mlxtend.preprocessing import TransactionEncoder

# Load dataset
df = pd.read_csv("/content/tmdb_5000_movies.csv")

df.head()
```



```

/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)

```

	budget	genres	homepage	id	keywords	original_language	original_title	overview
		{{"id": 28,			{{"id": 1463,			In the 22nd

```
import ast
```

```
# Function to extract genre names
```

```
def extract_names(text):
```

```
    try:
```

```
        data = ast.literal_eval(text)
```

```
        return [item['name'] for item in data]
```

```
    except:
```

```
        return []
```

```
# Apply extraction
```

```
df['genres_list'] = df['genres'].apply(extract_names)
```

```
df[['title', 'genres_list']].head()
```

		{{"id": 12,			"spy"}, {"id": 818,			past sends him o...
		"nam...			"name...			
3	250000000	{{"id": 28,	http://www.thedarkknightises.com/	49026	{{"id": 849,	en	The Dark Knight Rises	Following the death of District Attorney Harve...
		"name": "Action"}, {"id": 80,			"name": "dc comics"}, {"id": 853,...			
		"nam...			{{"id": 818,			John Carter is a
		{{"id": 28,			"name":			

4	260000000	<pre>"name": "Action"}, { "id": 12, "nam...</pre>	http://movies.disney.com/john-carter	49529	<pre>"name": "based on novel"}, { "id":....</pre>	en	John Carter	<pre>war weary, former military ca...</pre>
---	-----------	---	---	-------	---	----	-------------	---

```
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated  
return datetime.datetime.utcnow().replace(tzinfo=utc)  
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated  
return datetime.datetime.utcnow().replace(tzinfo=utc)  
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated  
return datetime.datetime.utcnow().replace(tzinfo=utc)
```

```

/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
    return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
    return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
    return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
    return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
    return datetime.datetime.utcnow().replace(tzinfo=utc)

```

```
transactions = df['genres list'].tolist()
```

```
# Remove empty transactions
```

```
transactions = [t for t in transactions if len(t) > 0]
```

```
len(transactions)
```

[illegible]


```
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated, datetime.datetime.now() is preferred
4 return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated, datetime.datetime.now() is preferred
return datetime.datetime.utcnow().replace(tzinfo=utc)
```

```
te = TransactionEncoder()
te_array = te.fit(transactions).transform(transactions)

df_encoded = pd.DataFrame(te_array, columns=te.columns_)

df_encoded.head()
```



```

/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated

```

```

frequent_itemsets = apriori(
    df_encoded,
    min_support=0.05,
    use_colnames=True
)

```

```

frequent_itemsets.sort_values(by='support', ascending=False).head()

```

```

/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)

```

	Action	Adventure	Animation	Comedy	Crime	Documentary	Drama	Family	Fantasy	Foreign	History	Horror	Music	Mystery	R
0	True	True	False	False	False	False	False	False	True	False	False	False	False	False	
1	True	True	False	False	False	False	False	False	True	False	False	False	False	False	
2	True	True	False	False	True	False	False	False	False	False	False	False	False	False	
3	True	False	False	False	True	False	True	False	False	False	False	False	False	False	
4	True	True	False	False	False	False	False	False	False	False	False	False	False	False	

```

/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated

```

```

return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)

```

	support	itemsets
4	0.481047	(Drama)
2	0.360628	(Comedy)
11	0.266806	(Thriller)
0	0.241675	(Action)
9	0.187225	(Romance)

```

/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)

```

Inference:

Finds frequently occurring genre combinations.

Support = percentage of movies containing that combination.

```
rules = association_rules(  
    frequent_itemsets,  
    metric="confidence",  
    min_threshold=0.6  
)  
  
rules[['antecedents', 'consequents', 'support', 'confidence', 'lift']].head()
```



```

/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)

```

```
df['keywords_list'] = df['keywords'].apply(extract_names)
```

```
transactions_kw = [t for t in df['keywords_list'] if len(t) > 0]
```

```
te_kw = TransactionEncoder()
```

```
te_array_kw = te_kw.fit(transactions_kw).transform(transactions_kw)
```

```
df_encoded_kw = pd.DataFrame(te_array_kw, columns=te_kw.columns_)
```

```
frequent_kw = apriori(df_encoded_kw, min_support=0.03, use_colnames=True)
```

```
frequent_kw.head()
```

```

/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)

```

	antecedents	consequents	support	confidence	lift
0	(Romance)	(Drama)	0.126283	0.674497	1.402143

1 (Mystery) (Thriller) 0.050681 0.695402 2.606394

```
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated  
return datetime.utcnow().replace(tzinfo=utc)
```

```
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated  
return datetime.utcnow().replace(tzinfo=utc)
```

```
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated  
return datetime.utcnow().replace(tzinfo=utc)
```


[illegible]

h.google.com/drive/1XPYwU0HLGjY-x01Zbasp7HsKWsgfhCUK?usp=sharing#printMode=true

[illegible]

[illegible]

```

/usr/local/lib/python3.12/dist-packages/djupyter_client/manager.py:202: DeprecationWarning: datetime.datetime.utcnow() is deprecated, datetime.datetime.now() is preferred

```


[illegible]

<https://colab.research.google.com/drive/1XPYwU0HLGjY-x01Zbasp7HsKWsGfhCUk?usp=sharing#printMode=true>

```

return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.utcnow().replace(tzinfo=utc)

```


[illegible]

[illegible]

[illegible]

[illegible]

h.google.com/drive/1XPYwU0HLGjY-x01Zbasp7HsKWsgfhCUk?usp=sh

[illegible]

[illegible]

```

/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
We: /usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
Extracted genres
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
Converted to transaction format
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
One-hot encoded data
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
Applied Apriori
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
Generated association rules
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
Interpreted support, confidence, lift

```

support itemsets

```

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA

```

```

# Load dataset
df = pd.read_csv("/content/tmdb_5000_movies.csv")

```

```
df.head()
```

```

return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)

```


[illegible]

[illegible]

[illegible]

```

/usr/local/lib/python3.12/dist-packages/duniter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated, datetime.datetime.now() is recommended

```


[illegible]

[illegible]


```
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
```

[illegible]

<https://colab.research.google.com/drive/1XPYwU0HLGjY-x01Zbasp7HsKWsgfhCUk?usp=sharing#printMode=true>

[illegible]

[illegible]

h.google.com/drive/1XPYwU0HLGjY-x01Zbasp7HsKWsgfhCUk?usp=sh

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```

/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)

```

budget	genres	homepage	id	keywords	original_language	original_title	overview
0 237000000	{ "id": 28, "name": "Action"},	http://www.avatarmovie.com/	19995	{ "id": 1463, "name": "Culture	en	Avatar	In the 22nd century, a paralel

```

numeric_cols = ['budget', 'popularity', 'revenue',
                'runtime', 'vote_average', 'vote_count']

# Convert to numeric
for col in numeric_cols:
    df[col] = pd.to_numeric(df[col], errors='coerce')

# Fill missing values
df[numeric_cols] = df[numeric_cols].fillna(df[numeric_cols].median())

# Select features
X = df[numeric_cols]

X.head()

```

	nam...	"name...	senas nim o...
3	250000000	<pre> {"id": 28, "name": "Action"}, {"id": 80, "nam... </pre>	<pre> http://www.thedarkknightises.com/ 49026 </pre>
4	260000000	<pre> {"id": 28, "name": "Action"}, {"id": 12, "nam... </pre>	<pre> http://movies.disney.com/john-carter 49529 </pre>

```

/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is depre
return datetime.datetime.utcnow().replace(tzinfo=utc)

```

```

/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)

```

	budget	popularity	revenue	runtime	vote_average	vote_count
0	237000000	150.437577	2787965087	162.0	7.2	11800
1	300000000	139.082615	961000000	169.0	6.9	4500
2	245000000	107.376788	880674609	148.0	6.3	4466
3	250000000	112.312950	1084939099	165.0	7.6	9106
4	260000000	43.926995	284139100	132.0	6.1	2124

```

/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated
return datetime.datetime.utcnow().replace(tzinfo=utc)

```

PCA works only on numeric data. Missing values are replaced with median.

```

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

```



```

/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)

```

```

pca = PCA()
X_pca = pca.fit_transform(X_scaled)

```

```

/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated,
return datetime.datetime.utcnow().replace(tzinfo=utc)

```

```
return datetime.utcnow().replace(tzinfo=utc)
```

```
explained_variance = pca.explained_variance_ratio_
```

```
print("Explained Variance Ratio:")
```

```
print(explained_variance)
```

```
Explained Variance Ratio:
```

```
[0.54420777 0.19138384 0.11902831 0.07573418 0.04299268 0.02665322]
```

```
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated, use datetime.datetime.now(datetime.UTC).  
return datetime.utcnow().replace(tzinfo=utc)
```

```
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated, use datetime.datetime.now(datetime.UTC).  
return datetime.utcnow().replace(tzinfo=utc)
```

```
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated, use datetime.datetime.now(datetime.UTC).  
return datetime.utcnow().replace(tzinfo=utc)
```

```
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated, use datetime.datetime.now(datetime.UTC).  
return datetime.utcnow().replace(tzinfo=utc)
```

```
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated, use datetime.datetime.now(datetime.UTC).  
return datetime.utcnow().replace(tzinfo=utc)
```

```
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated, use datetime.datetime.now(datetime.UTC).  
return datetime.utcnow().replace(tzinfo=utc)
```

```
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated, use datetime.datetime.now(datetime.UTC).  
return datetime.utcnow().replace(tzinfo=utc)
```

```
/usr/local/lib/python3.12/dist-packages/jupyter_client/session.py:203: DeprecationWarning: datetime.datetime.utcnow() is deprecated, use datetime.datetime.now(datetime.UTC).  
return datetime.utcnow().replace(tzinfo=utc)
```

```
plt.figure()  
plt.plot(np.cumsum(explained_variance))  
plt.xlabel("Number of Components")  
plt.ylabel("Cumulative Explained Variance")  
plt.title("PCA Explained Variance")  
plt.show()
```


