Name:Atharva Shinde Roll No:54 Div:D15C

EXP 1

Aim:

- Load data in Pandas.
- Description of the dataset.
- Drop columns that aren’t useful.
- Drop rows with maximum missing values.
- Take care of missing data.
- Create dummy variables.
- Find out outliers (manually)
- standardization and normalization of columns

Data preprocessing

Data preprocessing involves transforming raw data into a structured and meaningful format, making it suitable for analysis. It is a crucial step in data mining, as raw data often contains inconsistencies, missing values, or noise. Ensuring data quality is essential before applying machine learning or data mining algorithms to achieve accurate and reliable results.

Why is Data Preprocessing Important?

Data preprocessing is essential for ensuring the quality and reliability of data before analysis. It helps improve the accuracy and efficiency of machine learning and data mining processes. The key aspects of data quality include:

- **Accuracy:** Ensuring the data is correct and free from errors.
- **Completeness:** Checking for missing or unrecorded data.
- **Consistency:** Verifying that data remains uniform across different sources.
- **Timeliness:** Ensuring the data is up-to-date and relevant.
- **Believability:** Assessing whether the data is reliable and trustworthy.
- **Interpretability:** Making sure the data is clear and easy to understand.

Dataset: Top_10000_Movies

Name:Atharva Shinde Roll No:54 Div:D15C

1) Loading Data in Pandas

```
import pandas as pd

df = pd.read_csv("/content/Top_10000_Movies.csv", on_bad_lines="skip")
df.head()
```

	id	original_language	original_title	popularity	release_date	vote_average	vote_count	genre	overview	revenue	runtime	tagline
0	580489.0	en	Venom: Let There Be Carnage	5401.308	2021-09-30	6.8	1736.0	['Science Fiction', 'Action', 'Adventure']	After finding a host body in investigative rep...	424000000.0	97.0	NaN
1	524434.0	en	Eternals	3365.535	2021-11-03	7.1	622.0	['Action', 'Adventure', 'Science Fiction', 'Fa...	The Eternals are a team of ancient aliens who ...	165000000.0	157.0	In the beginning...
2	438631.0	en	Dune	2911.423	2021-09-15	8.0	3632.0	['Action', 'Adventure', 'Science Fiction']	Paul Atreides, a brilliant and gifted young ma...	331116356.0	155.0	Beyond fear, destiny awaits.
3	796499.0	en	Army of Thieves	2552.437	2021-10-27	6.9	555.0	['Action', 'Crime', 'Thriller']	A mysterious woman recruits bank teller Ludwig...	0.0	127.0	Before Vegas, one locksmith became a legend.
4	550988.0	en	Free Guy	1850.470	2021-08-11	7.8	3493.0	['Comedy', 'Action', 'Adventure', 'Science Fic...	A bank teller called Guy realizes he is a back...	331096766.0	115.0	Life's too short to be a background character.

2) Description of the dataset.

```
df.info()

df.describe()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 10014 entries, 0 to 9999
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   id                 10002 non-null  float64
 1   original_language  10002 non-null  object
 2   original_title     10001 non-null  object
 3   popularity         10000 non-null  float64
 4   release_date       9962 non-null   object
 5   vote_average       10000 non-null  float64
 6   vote_count         10000 non-null  float64
 7   genre              10000 non-null  object
 8   overview           9900 non-null   object
 9   revenue            9998 non-null   float64
 10  runtime            9989 non-null   float64
 11  tagline            7079 non-null   object
dtypes: float64(6), object(6)
memory usage: 1.2+ MB
```

	id	popularity	vote_average	vote_count	revenue	runtime
count	10002.000000	10000.000000	10000.00000	10000.000000	9.998000e+03	9989.000000
mean	250003.082683	34.516871	6.29875	1315.084900	5.737536e+07	98.792772
std	261732.329571	100.693958	1.43426	2501.899103	1.480897e+08	28.771525
min	0.000000	6.269000	0.00000	0.000000	0.000000e+00	0.000000
25%	11864.500000	11.908000	5.90000	118.000000	0.000000e+00	89.000000

✓ Connected to Python 3 Google Compute Engine backend

df.info(): Provides an overview of the dataset, including:

- Number of rows and columns.
- Data types of each column (e.g., int, float, object).
- Number of non-null (non-missing) values in each column.

df.describe(): Generates summary statistics for numeric columns, such as:

- count: Number of non-missing values.
- mean: Average value.
- std: Standard deviation.

Name:Atharva Shinde Roll No:54 Div:D15C

- min, 25%, 50% (median), 75%, and max: Percentile values

3) Drop columns that aren't useful: Columns like Invoice ID may not contribute to analysis (it's often just an identifier). Removing irrelevant columns reduces complexity.

Drop Columns

```
df = df.drop(["popularity", "runtime"], axis=1)

df.head()
```

	id	original_language	original_title	release_date	vote_average	vote_count	genre	overview	revenue	tagline
0	580489.0	en	Venom: Let There Be Carnage	2021-09-30	6.8	1736.0	['Science Fiction', 'Action', 'Adventure']	After finding a host body in investigative rep...	424000000.0	NaN
1	524434.0	en	Eternals	2021-11-03	7.1	622.0	['Action', 'Adventure', 'Science Fiction', 'Fa...	The Eternals are a team of ancient aliens who ...	165000000.0	In the beginning...
2	438631.0	en	Dune	2021-09-15	8.0	3632.0	['Action', 'Adventure', 'Science Fiction']	Paul Atreides, a brilliant and gifted young ma...	331116356.0	Beyond fear, destiny awaits.
3	796499.0	en	Army of Thieves	2021-10-27	6.9	555.0	['Action', 'Crime', 'Thriller']	A mysterious woman recruits bank teller Ludwig...	0.0	Before Vegas, one locksmith became a legend.
4	550988.0	en	Free Guy	2021-08-11	7.8	3493.0	['Comedy', 'Action', 'Adventure', 'Science Fic...	A bank teller called Guy realizes he is a back...	331096766.0	Life's too short to be a background character.

Name:Atharva Shinde Roll No:54 Div:D15C

4) Drop rows with maximum missing values.

df.dropna(thresh=int(0.5 * len(df.columns))):

- Drops rows where more than half the columns have missing (NaN) values.
- thresh=int(0.5 * len(df.columns)): Ensures that a row must have at least 50% non-null values to remain.

df = ...: Updates the DataFrame after dropping rows.
print(df.info()): Confirms that rows with excessive missing values have been removed.

Drop Rows with Maximum Missing Values

```
df = df.dropna(thresh=len(df.columns) * 0.5)

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

	0
id	0
original_language	0
original_title	0
release_date	38
vote_average	0
vote_count	0
genre	0
overview	100
revenue	2
tagline	2921

dtype: int64

5) Take care of missing data.

df.fillna(df.mean()): Replaces missing values (NaN) in numeric columns with the mean of that column.

Handle Missing Data

```
[ ] df["release_date"] = pd.to_datetime(df["release_date"], errors="coerce")
```

```
[ ] df.fillna(df.mode().iloc[0], inplace=True)
```

6) Create dummy variables.

pd.get_dummies(): Converts categorical variables into dummy variables (binary indicators: 0 or 1).

columns=['...']: Specifies which columns to convert.
drop_first=True: Avoids the "dummy variable trap" by dropping one dummy variable to prevent multicollinearity.

Name:Atharva Shinde Roll No:54 Div:D15C

```python
df = pd.get_dummies(df, columns=['original_language', 'genre'], drop_first=True)

df.head()
```

	id	original_title	release_date	vote_average	vote_count	overview	revenue	tagline	original_language_be	original_language_bn	...	genre_['Western', 'History', 'War']	genre_['Western', 'History']	genre_['Weste 'Horr 'Thrill
0	580489.0	Venom: Let There Be Carnage	2021-09-30	6.8	1736.0	After finding a host body in investigative rep...	424000000.0	Based on a true story.	False	False	...	False	False	
1	524434.0	Eternals	2021-11-03	7.1	622.0	The Eternals are a team of ancient aliens who ...	165000000.0	In the beginning...	False	False	...	False	False	
2	438631.0	Dune	2021-09-15	8.0	3632.0	Paul Atreides, a brilliant and gifted young ma...	331116356.0	Beyond fear, destiny awaits.	False	False	...	False	False	F
3	796499.0	Army of Thieves	2021-10-27	6.9	555.0	A mysterious woman recruits	0.0	Before Vegas, one locksmith became a	False	False	...	False	False	F

7) Find out outliers (manually)

```python
# Compute Q1 (25th percentile) and Q3 (75th percentile)
Q1 = df["revenue"].quantile(0.25)  # First Quartile (25%)
Q3 = df["revenue"].quantile(0.75)  # Third Quartile (75%)

# Compute Interquartile Range (IQR)
IQR = Q3 - Q1

# Define lower and upper bounds for outliers
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

print(f"Q1 (25th percentile): {Q1}")
print(f"Q3 (75th percentile): {Q3}")
print(f"IQR (Interquartile Range): {IQR}")
print(f"Lower Bound: {lower_bound}")
print(f"Upper Bound: {upper_bound}")
```

```
Q1 (25th percentile): 0.0
Q3 (75th percentile): 47645488.0
IQR (Interquartile Range): 47645488.0
Lower Bound: -71468232.0
Upper Bound: 119113720.0
```

Name:Atharva Shinde Roll No:54 Div:D15C

8) standardization and normalization of columns

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Standardization equation

$$X' = \frac{X - \mu}{\sigma}$$

To standardize your data, we need to import the StandardScalar from the sklearn library and apply it to our dataset.

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Normalization equation

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

- When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0
- On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X' is 1
- If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1

To normalize your data, you need to import the MinMaxScalar from the sklearn library and apply it to our dataset.

Name:Atharva Shinde Roll No:54 Div:D15C

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
df[['vote_average_std', 'vote_count_std', 'revenue_std']] = scaler.fit_transform(df[['vote_average', 'vote_count', 'revenue']])

df.head()
```

	id	original_title	release_date	vote_average	vote_count	overview	revenue	tagline	original_language_be	original_language_bn	...	genre_['Western', 'Horror']	genre_['Western', 'Mystery', 'Thriller', 'Drama']	genre_['Western' 'Romance' 'Drama'
0	580489.0	Venom: Let There Be Carnage	2021-09-30	6.8	1736.0	After finding a host body in investigative rep...	424000000.0	Based on a true story.	False	False	...	False	False	Fals
1	524434.0	Eternals	2021-11-03	7.1	622.0	The Eternals are a team of ancient aliens who ...	165000000.0	In the beginning...	False	False	...	False	False	Fals

```
from sklearn.preprocessing import MinMaxScaler

min_max_scaler = MinMaxScaler()
df[['vote_average_norm', 'vote_count_norm', 'revenue_norm']] = min_max_scaler.fit_transform(df[['vote_average', 'vote_count', 'revenue']])

df.head()
```

	id	original_title	release_date	vote_average	vote_count	overview	revenue	tagline	original_language_be	original_language_bn	...	genre_['Western', 'TV Movie']	genre_['Western', 'Thriller']	genre_['Western'
0	580489.0	Venom: Let There Be Carnage	2021-09-30	6.8	1736.0	After finding a host body in investigative rep...	424000000.0	Based on a true story.	False	False	...	False	False	Fals
1	524434.0	Eternals	2021-11-03	7.1	622.0	The Eternals are a team of ancient aliens who ...	165000000.0	In the beginning...	False	False	...	False	False	Fals
2	438631.0	Dune	2021-09-15	8.0	3632.0	Paul Atreides, a brilliant and gifted young ma...	331116356.0	Beyond fear, destiny awaits.	False	False	...	False	False	Fals
3	796499.0	Army of Thieves	2021-10-27	6.9	555.0	A mysterious woman recruits bank teller	0.0	Before Vegas, one locksmith became a	False	False	...	False	False	Fals

Conclusion:

Thus we have understood how to perform data preprocessing which can further be taken into exploratory data analysis and further in the Model preparation sequence.