ESSENTIALS OF DATA SCIENCE

Theory Activity No. 1

Name : Swapnali Kakasaheb Kolpe

Div:CS2

Roll no:C23-57

PRN:202401040242

1. Display the first 5 records of the dataset.

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

df = pd.read_excel('OpinRank_Review_Dataset_Large.xlsx')
df.head()
```

	name	category	author	date	rating	review_text
0	Hotel DEF	hotel	emilywebster	2016-04-25	3	Sport Republican system apply pressure directo...
1	Car MNO	car	marshallnancy	2019-04-05	4	Grow power last as action name boy term stop c...
2	Car PQR	car	victoria03	2017-12-17	4	Church onto behind Mrs accept bill least shoul...
3	Car STU	car	xharvey	2020-08-10	4	Could back trial son former example rise.
4	Hotel JKL	hotel	patrick70	2020-02-28	4	Billion maintain knowledge court idea stage hi...

2. Show basic information about the dataset (columns, data types, non-null counts).

Python

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

df = pd.read_excel('OpinRank_Review_Dataset_Large.xlsx')
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 6 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   name         50 non-null     object
 1   category     50 non-null     object
 2   author       50 non-null     object
 3   date         50 non-null     datetime64[ns]
 4   rating       50 non-null     int64
 5   review_text  50 non-null     object
dtypes: datetime64[ns](1), int64(1), object(4)
memory usage: 2.5+ KB
```

3.Find the total number of reviews.

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

df = pd.read_excel('OpinRank_Review_Dataset_Large.xlsx')
len(df)
```

```
50
```

4. Count how many reviews are for hotels and how many for cars.

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

df = pd.read_excel('OpinRank_Review_Dataset_Large.xlsx')
df['category'].value_counts()
```

category	count
car	31
hotel	19

dtype: int64

5. Calculate the average rating for all reviews.

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

df = pd.read_excel('OpinRank_Review_Dataset_Large.xlsx')
df['rating'].mean()
```

```
np.float64(2.88)
```

6. Find the minimum and maximum ratings in the dataset.

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

df = pd.read_excel('OpinRank_Review_Dataset_Large.xlsx')
df['rating'].min(), df['rating'].max()
```

```
(1, 5)
```

7. List all unique hotel and car names.

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

df = pd.read_excel('OpinRank_Review_Dataset_Large.xlsx')
df['name'].unique()
```

```
array(['Hotel DEF', 'Car MNO', 'Car PQR', 'Car STU', 'Hotel JKL',
       'Hotel GHI', 'Car XYZ', 'Hotel ABC'], dtype=object)
```

8. Show the number of reviews per product name.

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

df = pd.read_excel('OpinRank_Review_Dataset_Large.xlsx')
df['name'].value_counts()
```

name	count
Car PQR	9
Car MNO	8
Car STU	8
Hotel JKL	6
Hotel ABC	6
Car XYZ	6
Hotel DEF	4
Hotel GHI	3

dtype: int64

9. Display all reviews with a rating of 1.

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

df = pd.read_excel('OpinRank_Review_Dataset_Large.xlsx')
df[df['rating'] == 1]
```

	name	category	author	date	rating	review_text
12	Car XYZ	car	stephanieandrade	2022-09-10	1	Born alone this that else bit.
21	Car MNO	car	rhonda02	2023-07-02	1	Instead politics hope trade kid mean.
23	Car PQR	car	imiller	2022-09-18	1	Specific maybe claim year be imagine discuss w...
24	Car MNO	car	holly35	2025-02-26	1	Major choose energy program enough glass.
25	Hotel DEF	hotel	danielmartinez	2023-09-11	1	Action attorney each draw option short.
26	Car PQR	car	wmorse	2023-12-21	1	Technology hundred movement child argue town h...
30	Car PQR	car	wwest	2021-02-03	1	International part first time rule total.
33	Car MNO	car	katelyngarcia	2022-10-15	1	Green employee customer sort camera everything...
35	Car MNO	car	patrickstein	2015-09-04	1	Market you carry nothing conference yard record.
40	Car STU	car	nshelton	2019-04-11	1	Remain blue majority high building return agre...
47	Hotel ABC	hotel	bruceallen	2015-08-06	1	Best audience value agreement sure bag you pla...

10. Show all reviews written after the year 2020.

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

df = pd.read_excel('OpinRank_Review_Dataset_Large.xlsx')
df[df['date'] > '2020-01-01']
```

	name	category	author	date	rating	review_text
3	Car STU	car	xharvey	2020-08-10	4	Could back trial son former example rise.
4	Hotel JKL	hotel	patrick70	2020-02-28	4	Billion maintain knowledge court idea stage hi...
6	Car XYZ	car	michaelmosley	2025-01-23	4	Region eight away bad player home ball.
7	Car XYZ	car	moraleskyle	2023-12-07	2	Forward process government chair teacher refle...
8	Hotel DEF	hotel	aaroncook	2023-10-06	2	Could bit cultural garden attack energy senior...
9	Hotel JKL	hotel	joseph71	2024-05-20	4	Program shoulder dream nation subject operatio...
12	Car XYZ	car	stephanieandrade	2022-09-10	1	Born alone this that else bit.
13	Car XYZ	car	james13	2024-02-19	2	Material difference yes step away class four t...
19	Hotel JKL	hotel	frankeric	2023-06-21	5	Participant growth color individual agree oil ...
20	Car MNO	car	nray	2021-09-09	2	Economic house gun ready explain back east.
21	Car MNO	car	rhonda02	2023-07-02	1	Instead politics hope trade kid mean.

11. Count how many users gave a 5-star rating.

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

df = pd.read_excel('OpinRank_Review_Dataset_Large.xlsx')
(df['rating'] == 5).sum()
```

```
np.int64(9)
```

12. Find the average rating by product category (hotel/car).

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

df = pd.read_excel('OpinRank_Review_Dataset_Large.xlsx')
df.groupby('category')['rating'].mean()
```

category	rating
car	2.645161
hotel	3.263158

13.Find the average review length (in characters).

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

df = pd.read_excel('OpinRank_Review_Dataset_Large.xlsx')
df['review_text'].str.len().mean()
```

```
np.float64(57.72)
```

14. Display the review with the longest review text.

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

df = pd.read_excel('OpinRank_Review_Dataset_Large.xlsx')
df.loc[df['review_text'].str.len().idxmax()]
```

	39
name	Hotel JKL
category	hotel
author	christinaponce
date	2022-03-23 00:00:00
rating	5
review_text	Case claim billion bill defense draw wonder re...

dtype: object

15. Find the most active reviewer (user with most reviews).

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

df = pd.read_excel('OpinRank_Review_Dataset_Large.xlsx')
df['author'].value_counts().idxmax()
```

```
'emilywebster'
```

16. Create a new column to categorize ratings as 'Good' (4-5), 'Average' (3), 'Bad' (1-2).

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

df = pd.read_excel('OpinRank_Review_Dataset_Large.xlsx')
df['rating_category'] = np.where(df['rating'] >= 4, 'Good',
                                 np.where(df['rating'] == 3, 'Average', 'Bad'))
df['rating_category'].describe()
```

	rating_category
count	50
unique	3
top	Bad
freq	24

dtype: object

17. Count how many reviews fall in each rating_category.

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

df = pd.read_excel('OpinRank_Review_Dataset_Large.xlsx')
df['rating_category'] = np.where(df['rating'] >= 4, 'Good',
                                 np.where(df['rating'] == 3, 'Average', 'Bad'))
df['rating_category'].value_counts()
```

rating_category	count
Bad	24
Good	20
Average	6

dtype: int64

18. Find the top 3 most-reviewed products.

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

df = pd.read_excel('OpinRank_Review_Dataset_Large.xlsx')
df['name'].value_counts().head(3)
```

name	count
Car PQR	9
Car MNO	8
Car STU	8

dtype: int64

19. Show all reviews that contain the word "good".

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

df = pd.read_excel('OpinRank_Review_Dataset_Large.xlsx')
df[df['review_text'].str.contains("good", case=False)]
```

	name	category	author	date	rating	review_text
33	Car MNO	car	katelyngarcia	2022-10-15	1	Green employee customer sort camera everything...

20. Calculate the number of reviews per year.

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

df = pd.read_excel('OpinRank_Review_Dataset_Large.xlsx')
df['year'] = pd.to_datetime(df['date']).dt.year
df['year'].value_counts().sort_index()
```

year	count
2015	4
2016	3
2017	2
2018	4
2019	9
2020	3
2021	6
2022	5
2023	6