ESSENTIALS OF DATA SCIENCE

Theory Activity No. 1

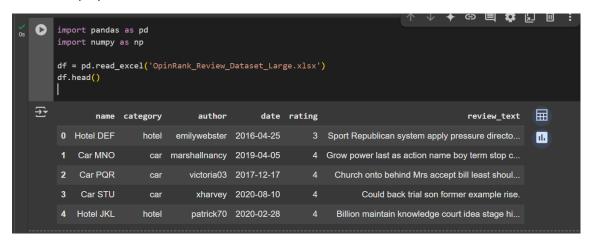
Name: Swapnali Kakasaheb Kolpe

Div:CS2

Roll no:C23-57

PRN:202401040242

1. Display the first 5 records of the dataset.



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

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):
                   Non-Null Count Dtype
    # Column
                    50 non-null
        category 50 non-null
                                  object
        author
                   50 non-null object
                    50 non-null
                                   datetime64[ns]
        date
                                  int64
        rating
                    50 non-null
    5 review_text 50 non-null
                                   object
    dtypes: datetime64[ns](1), int64(1), object(4)
    memory usage: 2.5+ KB
```

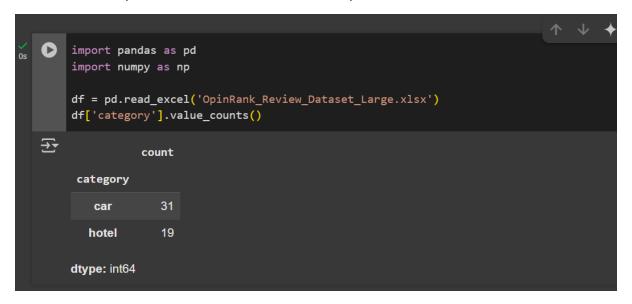
3. Find the total number of reviews.

```
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.



5. Calculate the average rating for all reviews.

```
import pandas as pd
import numpy as np

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

pp.float64(2.88)
```

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

```
import pandas as pd
import numpy as np

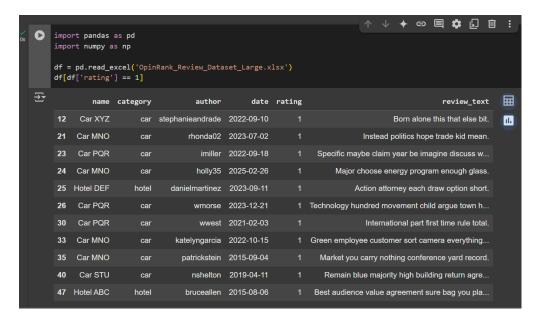
df = pd.read_excel('OpinRank_Review_Dataset_Large.xlsx')
    df['rating'].min(), df['rating'].max()
```

7. List all unique hotel and car names.

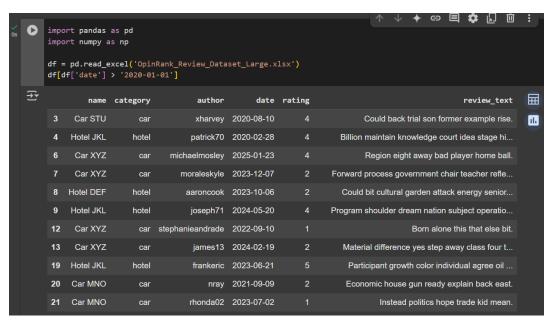
8. Show the number of reviews per product name.



9. Display all reviews with a rating of 1.



10. Show all reviews written after the year 2020.



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

```
import pandas as pd
import numpy as np

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

representation of the property of the prope
```

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

```
import pandas as pd
import numpy as np

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

rating

category

car 2.645161

hotel 3.263158
```

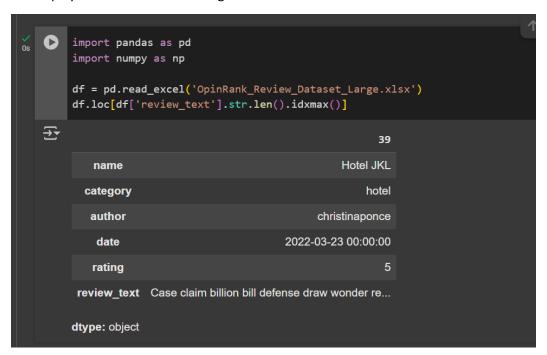
13. Find the average review length (in characters).

```
[59] 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.



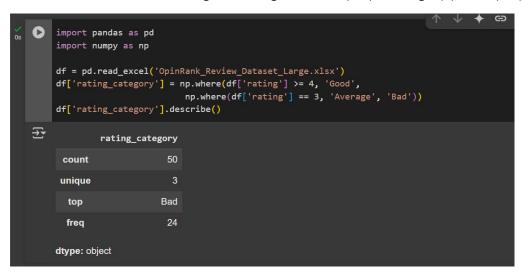
15. Find the most active reviewer (user with most reviews).

```
import pandas as pd
import numpy as np

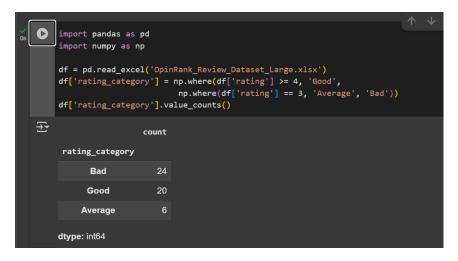
df = pd.read_excel('OpinRank_Review_Dataset_Large.xlsx')
    df['author'].value_counts().idxmax()

remilywebster'
```

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



17. Count how many reviews fall in each rating_category.



18. Find the top 3 most-reviewed products.



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



20. Calculate the number of reviews per year.

