MIT | Academy of Engineering

(An Autonomous Institute Affiliated to Savitribai Phule Pune University)

MAKE UP SESSION

NAME: PUSHKAR HARSHKUMAR PARAKH

PRN: 202401040313

CLASS:CS4

ROLL NO.:54

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

# Load the dataset
df = pd.read_csv("Book Reviews.csv")

# Display basic info
df.info()
df.head()
```

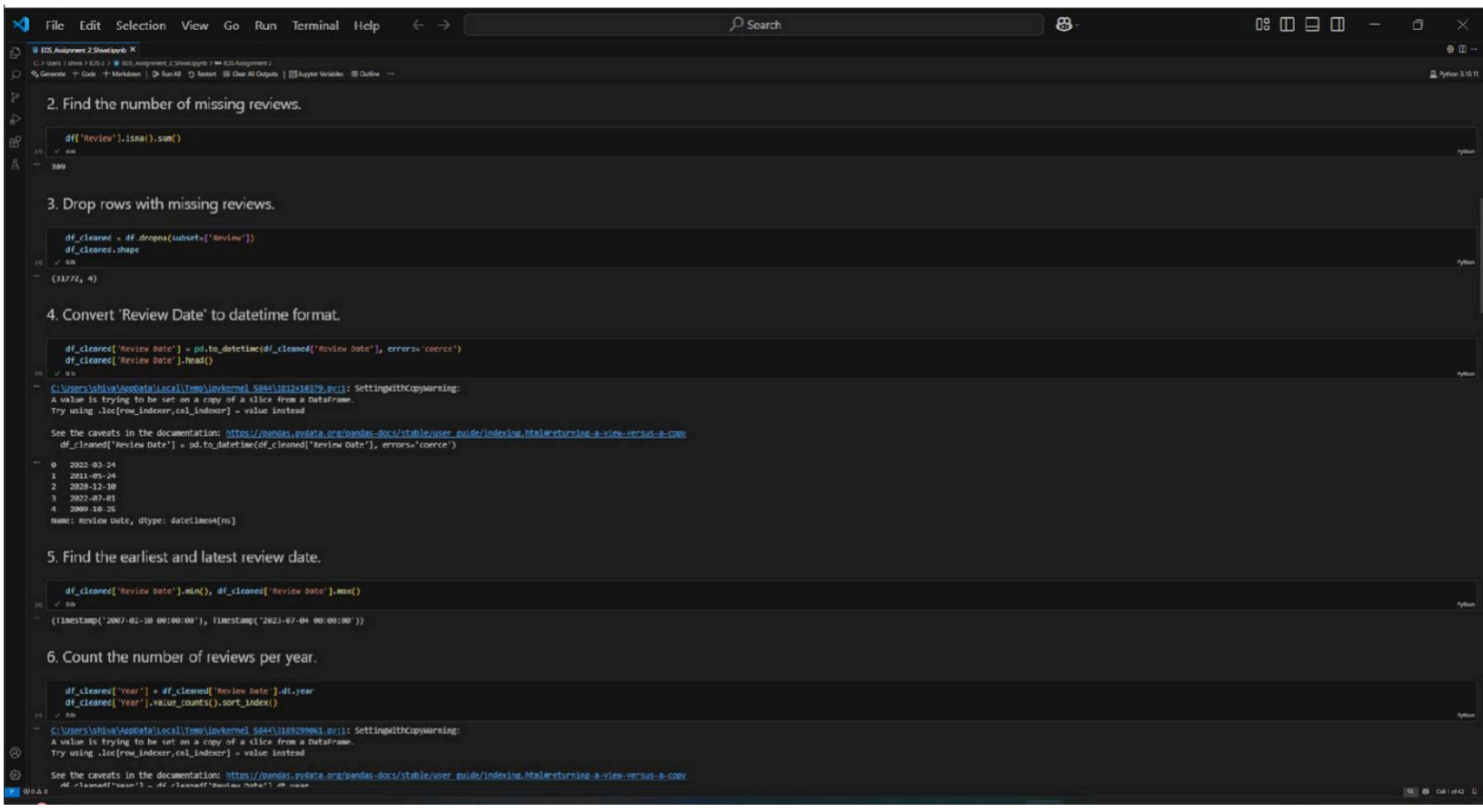
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32081 entries, 0 to 32080
Data columns (total 4 columns):
 #   Column       Non-Null Count  Dtype 
---  ------       --------------  ----- 
 0   Unnamed: 0   32081 non-null  int64 
 1   Book         32081 non-null  object
 2   Review       31772 non-null  object
 3   Review Date  32081 non-null  object
dtypes: int64(1), object(3)
memory usage: 1002.7+ KB
```

	Unnamed: 0	Book	Review	Review Date
0	0	To Kill a Mockingbird	/// gentle reminder that this is not the time ...	March 24, 2022
1	1	To Kill a Mockingbird	\n\n5.0 stars. I know I am risking a serious...	May 24, 2011
2	2	To Kill a Mockingbird	\n\nLooking for a new book but don't want to...	December 10, 2020
3	3	To Kill a Mockingbird	To Kill a Mockingbird. Harper Lee[To Kill a Mo...	July 1, 2022
4	4	To Kill a Mockingbird	Why is it when I pick up [To Kill A Mockingbi...	October 25, 2000

1. Display the total number of reviews per book.

```python
df['Book'].value_counts()
```

```
Book
To Kill a Mockingbird                 30
Burmese Onyx                          30
Amok and Other Stories                30
Rabbit Redux (Rabbit Angstrom, #2)    30
Sentimental Education                 30
                                      ..
The Triple Mirror of the Self          5
The Trusting and the Maimed            4
Farbenmul - volumul I                  3
태백산맥 세트                                2
Halbzeit                               2
Name: count, Length: 1096, dtype: int64
```



2. Find the number of missing reviews.

```python
df['Review'].isna().sum()
```

```
309
```

3. Drop rows with missing reviews.

```python
df_cleaned = df.dropna(subset=['Review'])
df_cleaned.shape
```

```
(31772, 4)
```

4. Convert 'Review Date' to datetime format.

```python
df_cleaned['Review Date'] = pd.to_datetime(df_cleaned['Review Date'], errors='coerce')
df_cleaned['Review Date'].head()
```

```
C:\Users\shiva\AppData\Local\Temp\ipykernel_5044\1812410379.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_cleaned['Review Date'] = pd.to_datetime(df_cleaned['Review Date'], errors='coerce')
0   2022-03-24
1   2011-05-24
2   2020-12-10
3   2022-07-01
4   2000-10-25
Name: Review Date, dtype: datetime64[ns]
```

5. Find the earliest and latest review date.

```python
df_cleaned['Review Date'].min(), df_cleaned['Review Date'].max()
```

```
(Timestamp('2007-02-10 00:00:00'), Timestamp('2023-07-04 00:00:00'))
```

6. Count the number of reviews per year.

```python
df_cleaned['Year'] = df_cleaned['Review Date'].dt.year
df_cleaned['Year'].value_counts().sort_index()
```

```
C:\Users\shiva\AppData\Local\Temp\ipykernel_5044\3189299061.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

File Edit Selection View Go Run Terminal Help

Search

EDS_Assignment_2_Shivat.ipynb

Generate + Code + Markdown | Run All Restart Clear All Outputs | Jupyter Variables Outline

Python 3.10.11

6. Count the number of reviews per year.

```
df_cleaned['Year'] = df_cleaned['Review Date'].dt.year
df_cleaned['Year'].value_counts().sort_index()
```

0.0s Python

```
C:\Users\shiva\AppData\Local\Temp\ipykernel_5044\3189299061.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_cleaned['Year'] = df_cleaned['Review Date'].dt.year
```

```
Year
2007.0     446
2008.0     739
2009.0     568
2010.0     525
2011.0     876
2012.0    1177
2013.0    1068
2014.0    1086
2015.0    1613
2016.0    2174
2017.0    2605
2018.0    2569
2019.0    2663
2020.0    3340
2021.0    3388
2022.0    3602
2023.0    2151
Name: count, dtype: int64
```

7. Get top 5 most reviewed books.

```
df_cleaned['Book'].value_counts().head()
```

0.0s Python

```
Book
The Pilgrim's Progress                        30
The Purloined Letter (C. Auguste Dupin, #3)   30
Kidnapped (David Balfour, #1)                 30
Of Love and Shadows                           30
The Garden Party and Other Stories            30
Name: count, dtype: int64
```

8. Count how many reviews contain the word 'love'.

```
df_cleaned['Review'].str.contains('love', case=False, na=False).sum()
```

0.0s Python

```
9176
```

File Edit Selection View Go Run Terminal Help

Search

EDS_Assignment_2_Shivat.ipynb

Generate + Code + Markdown | Run All Restart Clear All Outputs | Jupyter Variables Outline

Python 3.10.11

9. Add a column with the length of each review.

```
df_cleaned['Review Length'] = df_cleaned['Review'].str.len()
df_cleaned[['Review', 'Review Length']].head()
```

0.0s Python

```
C:\Users\shiva\AppData\Local\Temp\ipykernel_5044\1644116654.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_cleaned['Review Length'] = df_cleaned['Review'].str.len()
```

	Review	Review Length
0	/// gentle reminder that this is not the time ...	8167
1	\n\n\n6.0 stars. I know I am risking a serious...	4269
2	\n\n\nLooking for a new book but don't want to...	2779
3	To Kill a Mockingbird, Harper Lee(To Kill a Mo...	3365
4	Why is it when I pick up	to Kill A Mockingbi...	2542

10. Find average review length per book.

```
df_cleaned.groupby('Book')['Review Length'].mean().sort_values(ascending=False).head()
```

0.0s Python

```
Book
Cloud Atlas           6511.400000
Infinite Jest         6299.633333
Gravity's Rainbow     5301.800000
The Glass Bead Game   5155.833333
The Recognitions      4964.500000
Name: Review Length, dtype: float64
```

11. Show standard deviation of review lengths using NumPy.

```
np.std(df_cleaned['Review Length'].dropna())
```

0.0s Python

```
2284.845754323991
```

12. Create a column for word count in each review.

```
df_cleaned['Word Count'] = df_cleaned['Review'].apply(lambda x: len(str(x).split()))
df_cleaned[['Review', 'Word Count']].head()
```

0.0s Python

```
C:\Users\shiva\AppData\Local\Temp\ipykernel_5044\3503576763.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_cleaned['Word Count'] = df_cleaned['Review'].apply(lambda x: len(str(x).split()))
```