EX NO: 3

DATE:

Implementation of Data Cleaning, Transformation and Indexing Techniques

AIM:

To implement data cleaning, transformation and Indexing techniques in python using jupyter notebook

DATASET:

```
data = {

'Title': ['The Great Gatsby', '1984', 'To Kill a Mockingbird', 'Pride and Prejudice', 'The Catcher in the Rye'],

'Author': ['F. Scott Fitzgerald', 'George Orwell', 'Harper Lee', 'Jane Austen', 'J.D. Salinger'],

'Publication Year': [1925, 1949, 1960, 1813, 1951],

'Genre': ['Fiction', 'Dystopian', 'Fiction', 'Romance', 'Fiction'],

'Price': [10.99, 8.99, 7.99, 6.99, 9.99],

'Pages': [218, 328, 281, 279, 277],

'Publisher': ['Scribner', 'Secker & Warburg', 'J.B. Lippincott & Co.', 'T. Egerton', 'Little, Brown and Company']
```

QUESTIONS:

1. How would you handle missing values in the Price column?

```
import pandas as pd
df['Price'].fillna(value=0)
print(df)
                   Title
                                      Author Publication Year
                                                                   Genre \
0
        The Great Gatsby F. Scott Fitzgerald
                                                       1925
                                                                Fiction
                                                       1949 Dystopian
                   1984 George Orwell
  To Kill a Mockingbird
                              Harpe.
Jane Austen
2
                                Harper Lee
                                                       1960 Fiction
3 Pride and Prejudice
4 The Catcher in the Rye
                                                       1813 Romance
                              J.D. Salinger
                                                       1951 Fiction
                              Publisher
  Price Pages
         218
0 10.99
                                Scribner
         218 Scribner
328 Secker & Warburg
  8.99
           Secker & Warburg
281 J.B. Lippincott & Co.
279
  7.99
   6.99
3
           279
                             T. Egerton
4 9.99 277 Little, Brown and Company
```

2. Replace missing values in the Price column with the median value of this column.

```
median_price = df['Price'].median()
df['Price'].fillna(value=median_price)

0  10.99
1  8.99
2  7.99
3  6.99
4  9.99
Name: Price, dtype: float64
```

3. Check for duplicate entries based on the Title column in the dataset.

```
duplicates = df[df.duplicated(subset='Title')]
print(duplicates)

Empty DataFrame
Columns: [Title, Author, Publication Year, Genre, Price, Pages, Publisher]
Index: []
```

4. Write a pandas function to convert the Publication Year column to integer format.

```
df['Publication Year'] = df['Publication Year'].astype(int)
print(df)
                   Title
                                      Author Publication Year
                                                                   Genre \
0
        The Great Gatsby F. Scott Fitzgerald
                                                        1925
                                                                 Fiction
                                                         1949 Dystopian
                    1984
                             George Orwell
2 To Kill a Mockingbird
                                 Harper Lee
                                                        1960
                                                                 Fiction
     Pride and Prejudice
                                                         1813
                                                                 Romance
                                 Jane Austen
4 The Catcher in the Rye
                               J.D. Salinger
                                                         1951
                                                                 Fiction
```

5. Drop rows from the dataset where the Publisher column has missing values.

```
df.dropna(subset=['Publisher'], inplace=True)
print(df)
                                    Author Publication Year
                  Title
                                                                Genre
0
        The Great Gatsby F. Scott Fitzgerald
                                                      1925
                                                              Fiction
                        George Orwell
                  1984
                                                      1949 Dystopian
1
   To Kill a Mockingbird
                                                       1960
                                Harper Lee
                                                              Fiction
3
     Pride and Prejudice
                                Jane Austen
                                                       1813
                                                              Romance
4 The Catcher in the Rye
                            J.D. Salinger
                                                       1951
                                                              Fiction
  Price Pages
                              Publisher
0 10.99
          218
                              Scribner
                      Secker & Warburg
   8.99
          328
  7.99 281
2
                 J.B. Lippincott & Co.
3 6.99 279
                            T. Egerton
4 9.99 277 Little, Brown and Company
```

6. Add a new column named PriceInDollars to the dataset, which multiplies the Price by 1.

```
df['PriceInDollars'] = df['Price'] * 1
print(df)
                                    Author Publication Year
        The Great Gatsby F. Scott Fitzgerald
                                               1925
                                                              Fiction
                         George Orwell
                                                      1949 Dystopian
                   1984
1
   To Kill a Mockingbird
                                Harper Lee
                                                      1960
                                                              Fiction
                               Jane Austen
                                                      1813
    Pride and Prejudice
                                                              Romance
4 The Catcher in the Rye
                            J.D. Salinger
                                                      1951
                                                              Fiction
                             Publisher PriceInDollars
  Price Pages
                              Scribner
0 10.99
          218
                                               10.99
   8.99
           328
                       Secker & Warburg
                                                 8.99
   7.99
           281
                   J.B. Lippincott & Co.
                                                 7.99
                                                6.99
   6.99
           279
                            T. Egerton
   9.99
        277 Little, Brown and Company
                                                9.99
```

7. Apply a function to increase the Price column values by 10%.

```
df['Price'] = df['Price'] * 1.10
print(df)
                                    Author Publication Year
                  Title
                                                                Genre
        The Great Gatsby F. Scott Fitzgerald
                                              1925
                                                              Fiction
                  1984 George Orwell
                                                     1949 Dystopian
                                                     1960 Fiction
2
   To Kill a Mockingbird
                              Harper Lee
3
    Pride and Prejudice
                               Jane Austen
                                                      1813
                                                              Romance
  The Catcher in the Rye
                            J.D. Salinger
                                                      1951
                                                              Fiction
   Price Pages
                               Publisher PriceInDollars
0 12.089
           218
                                Scribner
                                                10.99
                        Secker & Warburg
   9.889
           328
                                                 8.99
1
2
   8.789
           281
                   J.B. Lippincott & Co.
                                                  7.99
3
   7.689
           279
                             T. Egerton
                                                 6.99
         277 Little, Brown and Company
4 10.989
                                                  9.99
```

8. Create a new column called PriceCategory that categorizes Price as 'Expensive' if greater than 9.99 and 'Affordable' otherwise.

```
df['PriceCategory'] = df['Price'].apply(lambda x: 'Expensive' if x > 9.99 else 'Affordable')
print(df)
                                   Author Publication Year
                 Title
                                                              Genre \
       The Great Gatsby F. Scott Fitzgerald
                                          1925
                                                            Fiction
                                                    1949 Dystopian
1
                  1984 George Orwell
                                                    1960
2
  To Kill a Mockingbird
                              Harper Lee
                                                            Fiction
3
                                                    1813
    Pride and Prejudice
                              Jane Austen
                                                            Romance
                                                    1951
                                                            Fiction
4 The Catcher in the Rye
                            J.D. Salinger
   Price Pages
                              Publisher PriceInDollars PriceCategory
0 12.089
                                          10.99
         218
                              Scribner
                                                       Expensive
         328
  9.889
                       Secker & Warburg
                                                8.99
                                                       Affordable
                                               7.99
  8 789
         281
                                                       Affordable
                   J.B. Lippincott & Co.
3
  7.689
         279
                                                6.99
                                                      Affordable
                            T. Egerton
4 10.989 277 Little, Brown and Company
                                                9.99
                                                        Expensive
```

9. Convert the Pages column to integer type.

```
df['Pages'] = df['Pages'].astype(int)
print(df)
                     Title
                                          Author Publication Year
                                                                          Genre
         The Great Gatsby F. Scott Fitzgerald
0
                                                               1925
                                                                       Fiction
                            .__zgerald
George Orwell
,,
1
                     1984
                                                               1949 Dystopian
                                 Harper Lee
    To Kill a Mockingbird
                                                              1960
                                                                       Fiction
  Pride and Prejudice
The Catcher in the Rye
                                    Jane Austen
                                                               1813
                                                                        Romance
                                J.D. Salinger
4
                                                               1951
                                                                        Fiction
                                   Publisher PriceInDollars PriceCategory
    Price Pages
             218 Scribner
328 Secker & Warburg
281 J.B. Lippincott & Co.
                                                 10.99
0 12.089
                                                                   Expensive
                                                        8.99
7.99
1
    9.889
                                                                  Affordable
                                                                Affordable
Affordable
    8.789
3
   7.689
             279
                                  T. Egerton
                                                         6.99
  10.989
           277 Little, Brown and Company
                                                        9.99
                                                                  Expensive
```

10. Calculate the total sum of the Price column in the dataset.

```
total_price = df['Price'].sum()
print(total_price)
```

49.445000000000001

11. Create a new column named YearsSincePublication that calculates the number of years since the Publication Year.

```
current_year = 2024
df['YearsSincePublication'] = current_year - df['Publication Year']
print(df)
                     Title
                                           Author Publication Year
                                                                            Genre
         The Great Gatsby F. Scott Fitzgerald
                                                                          Fiction
                                                                 1925
   To Kill a Mockingbird Harper Lee
Pride and Prejudice Jane Austen
The Catcher in the Rye J.D. Salinger
                                                                 1949
                                                                       Dystopian
                                                                 1960
                                                                         Fiction
                                                                 1813
                                                                          Romance
                                  J.D. Salinger
                                                                 1951
  The Catcher in the Rye
                                    Publisher PriceInDollars PriceCategory
           Pages
                                                10.99
                                                                    Expensive
Affordable
0
   12.089
              218
                                      Scribner
              328
                            Secker & Warburg
                                                           8.99
    9.889
                    J.B. Lippincott & Co.
    8.789
              281
                                                           7.99
                                                                    Affordable
                                                           6.99
    7.689
              279
                                    T. Egerton
                                                                    Affordable
   10.989
              277 Little, Brown and Company
                                                                     Expensive
   YearsSincePublication
1
                        75
3
                       211
```

12. Add a new column called GenreGroup that categorizes Genre as 'Classic' if 'Fiction' or 'Romance' and 'Modern' otherwise.

```
df['GenreGroup'] = df['Genre'].apply(lambda x: 'Classic' if x in ['Fiction', 'Romance'] else 'Modern')
print(df['GenreGroup'])

0    Classic
1    Modern
2    Classic
3    Classic
4    Classic
Name: GenreGroup, dtype: object
```

13. Replace all occurrences of missing values in the Publisher column with 'Unknown'.

14. Create a new column PriceDifference which is the difference between the maximum price and each book's price.

```
max_price = df['Price'].max()
df['PriceDifference'] = max_price - df['Price']
print(df['PriceDifference'])

0     0.0
1     2.2
2     3.3
3     4.4
4     1.1
Name: PriceDifference, dtype: float64
```

15. Aggregate the dataset to find the total number of pages by Genre

```
pages_by_genre = df.groupby('Genre')['Pages'].sum()
print(pages_by_genre)

Genre
Dystopian 328
Fiction 776
Romance 279
Name: Pages, dtype: int64
```

16. Retrieve the rows for books published after 1950.

```
books_after_1950 = df[df['Publication Year'] > 1950]
print(books_after_1950)
                                Author Publication Year
                                                                   Price
                                                           Genre
                           Harper Lee
   To Kill a Mockingbird
                                                  1960 Fiction
                                                                   8.789
4 The Catcher in the Rye J.D. Salinger
                                                   1951 Fiction 10.989
  Pages
                        Publisher PriceInDollars PriceCategory
           J.B. Lippincott & Co.
                                                   Affordable
                                            7.99
    277 Little, Brown and Company
                                            9.99
                                                    Expensive
  YearsSincePublication GenreGroup PriceDifference
2
                     64
                          Classic
                                              3.3
4
                     73
                          Classic
```

17. Remove rows where the Price is less than 8.00.

```
df = df[df['Price'] >= 8.00]
print(df)
                 Title
                                   Author Publication Year
                                                              Genre
        The Great Gatsby F. Scott Fitzgerald
                                                     1925
                                                             Fiction
1
                  1984
                        George Orwell
                                                     1949 Dystopian
2
  To Kill a Mockingbird
                              Harper Lee
                                                     1960
                                                            Fiction
4 The Catcher in the Rye
                            J.D. Salinger
                                                     1951
                                                            Fiction
   Price Pages
                              Publisher PriceInDollars PriceCategory
                                           10.99
0 12.089
         218
                              Scribner
                                                        Expensive
  9.889
           328
                                               8.99
1
                       Secker & Warburg
                                                        Affordable
                 J.B. Lippincott & Co.
2
  8.789
           281
                                                7.99
                                                        Affordable
4 10.989
         277 Little, Brown and Company
                                                9.99
                                                        Expensive
```

18. Filter the dataset to show books that have both Price greater than 9.00 and Pages greater than 270.

```
filtered_books = df[(df['Price'] > 9.00) & (df['Pages'] > 270)]
print(filtered_books)
                  Title
                               Author Publication Year
                                                           Genre
                                                                   Price \
                   1984 George Orwell
                                      1949 Dystopian
                                                                   9.889
4 The Catcher in the Rye J.D. Salinger
                                                 1951
                                                         Fiction 10.989
                        Publisher PriceInDollars PriceCategory \
  Pages
1
    328
                 Secker & Warburg
                                    8.99
                                                  Affordable
    277 Little, Brown and Company
                                          9.99
                                                   Expensive
  YearsSincePublication GenreGroup PriceDifference
                    75
                         Modern
1
                    73
4
                          Classic
                                             1.1
```

19. Create a new column PublicationDecade from the Publication Year column in 'YYYYs' format (e.g., '1920s')

```
df['PublicationDecade'] = (df['Publication Year'] // 10 * 10).astype(str) + 's'
print(df)
                   Title
                                      Author Publication Year
                                                                  Genre \
         The Great Gatsby F. Scott Fitzgerald
                                                                 Fiction
                   1984
                                                         1949 Dystopian
1
                          George Orwell
                                                         1960
2
    To Kill a Mockingbird
                                 Harper Lee
  The Catcher in the Rye
                              J.D. Salinger
                                                         1951
                                                                 Fiction
                                Publisher PriceInDollars PriceCategory
    Price Pages
                                            10.99
a
  12.089
            218
                                 Scribner
                                                             Expensive
                         Secker & Warburg
    9.889
                                                    8.99
                                                            Affordable
2
   8.789
            281
                     J.B. Lippincott & Co.
                                                    7.99
                                                           Affordable
            277 Little, Brown and Company
  10.989
                                                    9.99
                                                             Expensive
   YearsSincePublication GenreGroup PriceDifference PublicationDecade
0
                     99
                          Classic
                                              0.0
1
                     75
                           Modern
                                              2.2
                                                              19405
2
                     64
                           Classic
                                              3.3
                                                              1960s
4
                     73
                          Classic
                                              1.1
                                                              1950s
```

20. How do you check for duplicates based on the Author column?

<pre>author_duplicates = df[df.duplicated(subset='Author', keep=False)]</pre>	
<pre>print(author_duplicates)</pre>	

Empty DataFrame

Columns: [Title, Author, Publication Year, Genre, Price, Pages, Publisher, PriceInDollars, PriceCategory, YearsSincePublication, GenreGroup, PriceDifference, PublicationDecade]

Index: []

RUBRICS

Problem Understanding (10)	Implementation (20)	Viva (10)	Time Management (10)	Total (50)

RESULT

Thus the implementation of data cleaning, transformation and Indexing techniques in python using jupyter notebook was successfully executed and the output was verified.