Ex no:3

Date:

Handling missing data in a dataset using Pandas

Aim :

To implement python programs for handling missing data in a dataset using pandas

Car Details Dataset:

data = {

'Car_ID': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],

'Make': ['Toyota', 'Honda', np.nan, 'Ford', 'Chevrolet', 'Toyota', 'Honda', 'Chevrolet', np.nan, 'Ford'],

'Model': ['Camry', 'Civic', 'Accord', np.nan, 'Impala', 'Corolla', 'Civic', 'Cruze', 'Focus', 'Escape'],

'Year': [2015, 2016, 2017, 2018, np.nan, 2020, 2019, 2021, 2022, np.nan],

'Mileage': [50000, 30000, np.nan, 15000, 20000, np.nan, 25000, 12000, 18000, 22000],

'Price': [15000, 12000, 18000, np.nan, 25000, 14000, np.nan, 16000, 19000, 17000],

'Color': ['Red', 'Blue', 'Black', 'White', np.nan, 'Gray', 'Blue', np.nan, 'Black', 'White']

}

Questions :

1. **Write a pandas program to identify missing values in the dataset.**

```
import pandas as pd
import numpy as np

data = {
    'Car_ID': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
    'Make': ['Toyota', 'Honda', np.nan, 'Ford', 'Chevrolet', 'Toyota', 'Honda', 'Chevrolet', np.nan, 'Ford'],
    'Model': ['Camry', 'Civic', 'Accord', np.nan, 'Impala', 'Corolla', 'Civic', 'Cruze', 'Focus', 'Escape'],
    'Year': [2015, 2016, 2017, 2018, np.nan, 2020, 2019, 2021, 2022, np.nan],
    'Mileage': [50000, 30000, np.nan, 15000, 20000, np.nan, 25000, 12000, 18000, 22000],
    'Price': [15000, 12000, 18000, np.nan, 25000, 14000, np.nan, 16000, 19000, 17000],
    'Color': ['Red', 'Blue', 'Black', 'White', np.nan, 'Gray', 'Blue', np.nan, 'Black', 'White']
}

df = pd.DataFrame(data)

# Identifying missing values
missing_values = df.isnull()
print(missing_values)
```

```
   Car_ID   Make  Model   Year  Mileage  Price  Color
0   False  False  False  False    False  False  False
1   False  False  False  False    False  False  False
2   False   True  False  False     True  False  False
3   False  False   True  False    False   True  False
4   False  False  False   True    False  False   True
5   False  False  False  False     True  False  False
6   False  False  False  False    False   True  False
7   False  False  False  False    False  False   True
8   False   True  False  False    False  False  False
9   False  False  False   True    False  False  False
```

2. **How many missing values are there in the Year column?**

```
missing_years = df['Year'].isnull().sum()
print(f"Missing values in 'Year' column: {missing_years}")
```

```
Missing values in 'Year' column: 2
```

3. **How do you use pandas to fill missing values in the Mileage column using the mean?**

```
df['Mileage'].fillna(df['Mileage'].mean(), inplace=True)
print(df)
```

	Car_ID	Make	Model	Year	Mileage	Price	Color
0	1	Toyota	Camry	2015.0	50000.0	15000.0	Red
1	2	Honda	Civic	2016.0	30000.0	12000.0	Blue
2	3	NaN	Accord	2017.0	24000.0	18000.0	Black
3	4	Ford	NaN	2018.0	15000.0	NaN	White
4	5	Chevrolet	Impala	NaN	20000.0	25000.0	NaN
5	6	Toyota	Corolla	2020.0	24000.0	14000.0	Gray
6	7	Honda	Civic	2019.0	25000.0	NaN	Blue
7	8	Chevrolet	Cruze	2021.0	12000.0	16000.0	NaN
8	9	NaN	Focus	2022.0	18000.0	19000.0	Black
9	10	Ford	Escape	NaN	22000.0	17000.0	White

4. **Which panda's technique would you use to replace missing values in the Price column with its median value?**

```
df['Price'].fillna(df['Price'].median(), inplace=True)
print(df)
```

	Car_ID	Make	Model	Year	Mileage	Price	Color
0	1	Toyota	Camry	2015.0	50000.0	15000.0	Red
1	2	Honda	Civic	2016.0	30000.0	12000.0	Blue
2	3	NaN	Accord	2017.0	24000.0	18000.0	Black
3	4	Ford	NaN	2018.0	15000.0	16500.0	White
4	5	Chevrolet	Impala	NaN	20000.0	25000.0	NaN
5	6	Toyota	Corolla	2020.0	24000.0	14000.0	Gray
6	7	Honda	Civic	2019.0	25000.0	16500.0	Blue
7	8	Chevrolet	Cruze	2021.0	12000.0	16000.0	NaN
8	9	NaN	Focus	2022.0	18000.0	19000.0	Black
9	10	Ford	Escape	NaN	22000.0	17000.0	White

5. **How can you fill missing values in the Make column using the mode in pandas?**

```
df['Make'].fillna(df['Make'].mode()[0], inplace=True)
print(df)
```

	Car_ID	Make	Model	Year	Mileage	Price	Color
0	1	Toyota	Camry	2015.0	50000.0	15000.0	Red
1	2	Honda	Civic	2016.0	30000.0	12000.0	Blue
2	3	Chevrolet	Accord	2017.0	24000.0	18000.0	Black
3	4	Ford	NaN	2018.0	15000.0	16500.0	White
4	5	Chevrolet	Impala	NaN	20000.0	25000.0	NaN
5	6	Toyota	Corolla	2020.0	24000.0	14000.0	Gray
6	7	Honda	Civic	2019.0	25000.0	16500.0	Blue
7	8	Chevrolet	Cruze	2021.0	12000.0	16000.0	NaN
8	9	Chevrolet	Focus	2022.0	18000.0	19000.0	Black
9	10	Ford	Escape	NaN	22000.0	17000.0	White

6. **What panda's method can be used to fill missing values in the Model column if it is categorical?**

```
df['Model'].fillna(df['Model'].mode()[0], inplace=True)
print(df)
```

	Car_ID	Make	Model	Year	Mileage	Price	Color
0	1	Toyota	Camry	2015.0	50000.0	15000.0	Red
1	2	Honda	Civic	2016.0	30000.0	12000.0	Blue
2	3	Chevrolet	Accord	2017.0	24000.0	18000.0	Black
3	4	Ford	Civic	2018.0	15000.0	16500.0	White
4	5	Chevrolet	Impala	NaN	20000.0	25000.0	NaN
5	6	Toyota	Corolla	2020.0	24000.0	14000.0	Gray
6	7	Honda	Civic	2019.0	25000.0	16500.0	Blue
7	8	Chevrolet	Cruze	2021.0	12000.0	16000.0	NaN
8	9	Chevrolet	Focus	2022.0	18000.0	19000.0	Black
9	10	Ford	Escape	NaN	22000.0	17000.0	White

7. **How can you use pandas to check if there are any missing values remaining after filling them?**

```
any_missing_values = df.isnull().sum().sum() > 0
print(f"Any missing values remaining: {any_missing_values}")
```

```
Any missing values remaining: True
```

8. **What Python function can you use to check the presence of missing values in the entire dataset?**

```
any_missing_values = df.isnull().sum().sum() > 0
print(f"Any missing values remaining: {any_missing_values}")
```

```
Any missing values remaining: True
```

9. **How would you drop rows where any column has missing values?**

```
df_dropped = df.dropna()
print(df_dropped)
```

	Car_ID	Make	Model	Year	Mileage	Price	Color
0	1	Toyota	Camry	2015.0	50000.0	15000.0	Red
1	2	Honda	Civic	2016.0	30000.0	12000.0	Blue
2	3	Chevrolet	Accord	2017.0	24000.0	18000.0	Black
3	4	Ford	Civic	2018.0	15000.0	16500.0	White
5	6	Toyota	Corolla	2020.0	24000.0	14000.0	Gray
6	7	Honda	Civic	2019.0	25000.0	16500.0	Blue
8	9	Chevrolet	Focus	2022.0	18000.0	19000.0	Black

10. **How can you fill missing values in the Price column using the maximum count (mode)?**

```
df['Price'].fillna(df['Price'].mode()[0])
print(df)
```

	Car_ID	Make	Model	Year	Mileage	Price	Color
0	1	Toyota	Camry	2015.0	50000.0	15000.0	Red
1	2	Honda	Civic	2016.0	30000.0	12000.0	Blue
2	3	Chevrolet	Accord	2017.0	24000.0	18000.0	Black
3	4	Ford	Civic	2018.0	15000.0	16500.0	White
4	5	Chevrolet	Impala	NaN	20000.0	25000.0	NaN
5	6	Toyota	Corolla	2020.0	24000.0	14000.0	Gray
6	7	Honda	Civic	2019.0	25000.0	16500.0	Blue
7	8	Chevrolet	Cruze	2021.0	12000.0	16000.0	NaN
8	9	Chevrolet	Focus	2022.0	18000.0	19000.0	Black
9	10	Ford	Escape	NaN	22000.0	17000.0	White

RUBRICS

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

RESULT

Thus the python programs for handling missing in a dataset using pandas was successfully executed and the output was verified