PYTHON CODING CHALLENGE

NAME : TEJASWINI GOKANAKONDA

ROLL No: DE142

DATE : 15-11-2024

Printing Rows of the Data

```
import pandas as pd

df = pd.read_csv("dataset.csv")

print(df.head())
```

```
   Year Industry_aggregation_NZSIOC Industry_code_NZSIOC Industry_name_NZSIOC  \
0  2023                     Level 1                99999       All industries
1  2023                     Level 1                99999       All industries
2  2023                     Level 1                99999       All industries
3  2023                     Level 1                99999       All industries
4  2023                     Level 1                99999       All industries

                Units Variable_code  \
0  Dollars (millions)           H01
1  Dollars (millions)           H04
2  Dollars (millions)           H05
3  Dollars (millions)           H07
4  Dollars (millions)           H08

                                    Variable_name      Variable_category  \
0                                    Total income  Financial performance
1  Sales, government funding, grants and subsidies  Financial performance
2                Interest, dividends and donations  Financial performance
3                           Non-operating income  Financial performance
4                              Total expenditure  Financial performance

    Value                            Industry_code_ANZSIC06
0  930995  ANZSIC06 divisions A-S (excluding classes K633...
1  821630  ANZSIC06 divisions A-S (excluding classes K633...
2   84354  ANZSIC06 divisions A-S (excluding classes K633...
3   25010  ANZSIC06 divisions A-S (excluding classes K633...
4  832964  ANZSIC06 divisions A-S (excluding classes K633...
```

Printing the Column Names of the DataFrame

```
print(df.columns)
```

```
Index(['Year', 'Industry_aggregation_NZSIOC', 'Industry_code_NZSIOC',
       'Industry_name_NZSIOC', 'Units', 'Variable_code', 'Variable_name',
       'Variable_category', 'Value', 'Industry_code_ANZSIC06'],
      dtype='object')
```

Summary of Data Frame

```
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50985 entries, 0 to 50984
Data columns (total 10 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Year                        50985 non-null  int64
 1   Industry_aggregation_NZSIOC  50985 non-null  object
 2   Industry_code_NZSIOC        50985 non-null  object
 3   Industry_name_NZSIOC        50985 non-null  object
 4   Units                       50985 non-null  object
 5   Variable_code               50985 non-null  object
 6   Variable_name               50985 non-null  object
 7   Variable_category           50985 non-null  object
 8   Value                       50985 non-null  object
 9   Industry_code_ANZSIC06      50985 non-null  object
dtypes: int64(1), object(9)
memory usage: 3.9+ MB
None
```

Descriptive Statistical Measures of a DataFrame

```
print(df.describe())
```

```
                Year
count  50985.000000
mean    2018.000000
std        3.162309
min     2013.000000
25%     2015.000000
50%     2018.000000
75%     2021.000000
max     2023.000000
```

Missing Data Handling

```
# Filling it with 0
print(df.isnull().sum())
df = df.fillna(0)
# Deleting null values
df = df.dropna()
```

```
Year                          0
Industry_aggregation_NZSIOC   0
Industry_code_NZSIOC          0
Industry_name_NZSIOC          0
Units                         0
Variable_code                 0
Variable_name                 0
Variable_category             0
Value                         0
Industry_code_ANZSIC06        0
dtype: int64
```

Sorting DataFrame Values

```
df_sorted = df.sort_values(by="Year", ascending=True)
print(df_sorted.head())
```

```
       Year Industry_aggregation_NZSIOC Industry_code_NZSIOC  \
50984  2013                    Level 3                ZZ11
47889  2013                    Level 4               CC822
47890  2013                    Level 4               CC822
47891  2013                    Level 4               CC822
47892  2013                    Level 4               CC822

             Industry_name_NZSIOC              Units Variable_code  \
50984  Food product manufacturing          Percentage          H41
47889     Machinery Manufacturing  Dollars (millions)          H09
47890     Machinery Manufacturing  Dollars (millions)          H10
47891     Machinery Manufacturing  Dollars (millions)          H11
47892     Machinery Manufacturing  Dollars (millions)          H12

                 Variable_name      Variable_category Value  \
50984    Liabilities structure       Financial ratios    46
47889   Interest and donations  Financial performance    36
47890           Indirect taxes  Financial performance     9
47891             Depreciation  Financial performance    72
47892  Salaries and wages paid  Financial performance   908

                                Industry_code_ANZSIC06
50984  ANZSIC06 groups C111, C112, C113, C114, C115, ...
47889               ANZSIC06 groups C245, C246, and C249
47890               ANZSIC06 groups C245, C246, and C249
47891               ANZSIC06 groups C245, C246, and C249
47892               ANZSIC06 groups C245, C246, and C249
```

Merging Data Frames

```
df1 = df[['Industry_code_ANZSIC06', 'Industry_name_NZSIOC', 'Variable_name', 'Value']]

merged_df = pd.merge(df, df1, on="Industry_code_ANZSIC06", suffixes=('_left', '_right'))

print(merged_df.head())
```

```
   Year Industry_aggregation_NZSIOC Industry_code_NZSIOC  \
0  2023                     Level 1                99999
1  2023                     Level 1                99999
2  2023                     Level 1                99999
3  2023                     Level 1                99999
4  2023                     Level 1                99999

  Industry_name_NZSIOC_left              Units Variable_code  \
0            All industries  Dollars (millions)           H01
1            All industries  Dollars (millions)           H01
2            All industries  Dollars (millions)           H01
3            All industries  Dollars (millions)           H01
4            All industries  Dollars (millions)           H01

  Variable_name_left      Variable_category Value_left  \
0       Total income  Financial performance     930995
1       Total income  Financial performance     930995
2       Total income  Financial performance     930995
3       Total income  Financial performance     930995
4       Total income  Financial performance     930995

                              Industry_code_ANZSIC06  \
0  ANZSIC06 divisions A-S (excluding classes K633...
1  ANZSIC06 divisions A-S (excluding classes K633...
2  ANZSIC06 divisions A-S (excluding classes K633...
3  ANZSIC06 divisions A-S (excluding classes K633...
4  ANZSIC06 divisions A-S (excluding classes K633...

  Industry_name_NZSIOC_right                              Variable_name_right  \
0             All industries                                     Total income
1             All industries  Sales, government funding, grants and subsidies
2             All industries                Interest, dividends and donations
3             All industries                             Non-operating income
4             All industries                                Total expenditure

  Value_right
0      930995
1      821630
2       84354
3       25010
4      832964
```

Applying a Function

```
import numpy as np

df['Value'] = pd.to_numeric(df['Value'], errors='coerce')

#function to increase each value by 10%
def increase_by_percentage(value):
    return value * 1.10 if not np.isnan(value) else value

df['Value'] = df['Value'].apply(increase_by_percentage)

print(df[['Value']].head())
```

```
        Value
0  1024094.5
1   903793.0
2    92789.4
3    27511.0
4   916260.4
```

Using Lambda Operator

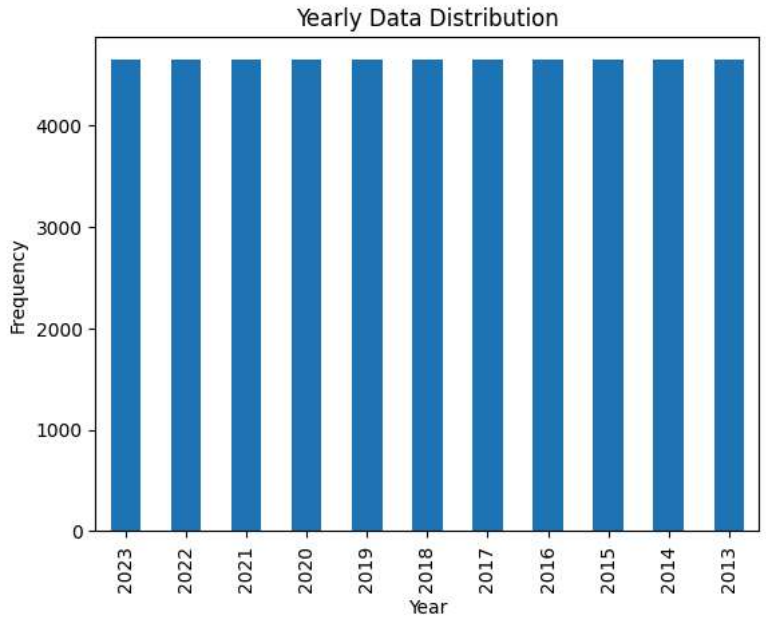
```
df['Adjusted_Value'] = df['Value'].apply(lambda x: x * 1.1 if x > 1000 else x)
print(df.head())
```

```
   Year Industry_aggregation_NZSIOC Industry_code_NZSIOC Industry_name_NZSIOC  \
0  2023                     Level 1                99999       All industries
1  2023                     Level 1                99999       All industries
2  2023                     Level 1                99999       All industries
3  2023                     Level 1                99999       All industries
4  2023                     Level 1                99999       All industries

                Units Variable_code  \
0  Dollars (millions)           H01
1  Dollars (millions)           H04
```

```
2  Dollars (millions)           H05
3  Dollars (millions)           H07
4  Dollars (millions)           H08

                                    Variable_name      Variable_category  \
0                                    Total income  Financial performance
1  Sales, government funding, grants and subsidies  Financial performance
2                Interest, dividends and donations  Financial performance
3                           Non-operating income  Financial performance
4                              Total expenditure  Financial performance

       Value                              Industry_code_ANZSIC06  \
0  1024094.5  ANZSIC06 divisions A-S (excluding classes K633...
1   903793.0  ANZSIC06 divisions A-S (excluding classes K633...
2    92789.4  ANZSIC06 divisions A-S (excluding classes K633...
3    27511.0  ANZSIC06 divisions A-S (excluding classes K633...
4   916260.4  ANZSIC06 divisions A-S (excluding classes K633...

   Adjusted_Value
0      1126503.95
1       994172.30
2       102068.34
3        30262.10
4      1007886.44
```

Visualizing DataFrame

```
import matplotlib.pyplot as plt

df['Year'].value_counts().plot(kind='bar')
plt.xlabel("Year")
plt.ylabel("Frequency")
plt.title("Yearly Data Distribution")
plt.show()
```



Number of Columns in the Dataset

```
print("Number of columns:", df.shape[1])
```

```
Number of columns: 11
```

Printing the Name of All Columns

```
print("Column names:", df.columns.tolist())
```

```
Column names: ['Year', 'Industry_aggregation_NZSIOC', 'Industry_code_NZSIOC', 'Industry_name_NZSIOC', 'Units', 'Variable_code', 'Variabl
```

Dataset Indexing

```
print("Dataset index:", df.index)
```

```
Dataset index: RangeIndex(start=0, stop=50985, step=1)
```

Number of Observations in the Dataset

```
print("Number of observations:", df.shape[0])
```

```
Number of observations: 50985
```