Coding Challenge– 2
Python
Data Processing with Pandas

Annual enterprise survey: 2023 financial year (provisional)

Name: Aathirainathan P

Date: 15-11-2024

1.Printing rows of the Data:

The data is loaded into a dataframe and all rows of the data set are printed.

[5]:
```
#Printing rows of data & display values
print(data) #Display all data
display(data.head())  #first 5 tows
display(data.tail())  #last 5 rows
```

```
       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
...     ...                         ...                  ...
50980  2013                     Level 3                ZZ11
50981  2013                     Level 3                ZZ11
50982  2013                     Level 3                ZZ11
50983  2013                     Level 3                ZZ11
50984  2013                     Level 3                ZZ11

             Industry_name_NZSIOC              Units Variable_code  \
0                All industries  Dollars (millions)           H01
1                All industries  Dollars (millions)           H04
2                All industries  Dollars (millions)           H05
3                All industries  Dollars (millions)           H07
4                All industries  Dollars (millions)           H08
...                         ...                 ...           ...
50980  Food product manufacturing          Percentage           H37
50981  Food product manufacturing          Percentage           H38
50982  Food product manufacturing          Percentage           H39
50983  Food product manufacturing          Percentage           H40
50984  Food product manufacturing          Percentage           H41
```

The Head Function will print the first five rows of the dataset and is shown below:

	Year	Industry_aggregation_NZSIOC	Industry_code_NZSIOC	Industry_name_NZSIOC	Units	Variable_code	Variable_name	Variable_category	Value	Industry_code
0	2023	Level 1	99999	All industries	Dollars (millions)	H01	Total income	Financial performance	930995	ANZSIC06 d (excluding cla
1	2023	Level 1	99999	All industries	Dollars (millions)	H04	Sales, government funding, grants and subsidies	Financial performance	821630	ANZSIC06 d (excluding cla
2	2023	Level 1	99999	All industries	Dollars (millions)	H05	Interest, dividends and donations	Financial performance	84354	ANZSIC06 d (excluding cla
3	2023	Level 1	99999	All industries	Dollars (millions)	H07	Non-operating income	Financial performance	25010	ANZSIC06 d (excluding cla
4	2023	Level 1	99999	All industries	Dollars (millions)	H08	Total expenditure	Financial performance	832964	ANZSIC06 d (excluding cla

The Tail Function will print the last five rows of the dataset and is shown below:

	Year	Industry_aggregation_NZSIOC	Industry_code_NZSIOC	Industry_name_NZSIOC	Units	Variable_code	Variable_name	Variable_category	Value	Industry
50980	2013	Level 3	ZZ11	Food product manufacturing	Percentage	H37	Quick ratio	Financial ratios	52	ANZSI C112, C1
50981	2013	Level 3	ZZ11	Food product manufacturing	Percentage	H38	Margin on sales of goods for resale	Financial ratios	40	ANZSI C112, C1
50982	2013	Level 3	ZZ11	Food product manufacturing	Percentage	H39	Return on equity	Financial ratios	12	ANZSI C112, C1
50983	2013	Level 3	ZZ11	Food product manufacturing	Percentage	H40	Return on total assets	Financial ratios	5	ANZSI C112, C1
50984	2013	Level 3	ZZ11	Food product manufacturing	Percentage	H41	Liabilities structure	Financial ratios	46	ANZSI C112, C1

2.Printing the column names of the DataFrame:

All the columns are fetched using column function, made into a list and printed, as shown below:

```
[6]: #Printing the column names of the DataFrame
print(list(data.columns))
```

```
['Year', 'Industry_aggregation_NZSIOC', 'Industry_code_NZSIOC', 'Industry_name_NZSIOC', 'Units', 'Variable_code', 'Variable_name', 'Variable_category',
'Value', 'Industry_code_ANZSIC06']
```

3.Summary of Data Frame:

The info function is used to fetch the summary of the data frame which gives us the non-null count and the datatype of each column in the dataframe as shown below:

```
[7]: #Summary of Data Frame
     data.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
```

4.Descriptive Statistical Measures of a DataFrame:

For this we use describe function, which will give us the count of non-null values, mean standard deviation, minimum, maximum, 25%,50% and 75% values:

```
[8]: #Descriptive Statistical Measures of a DataFrame
     data.describe()
```

[8]:

	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

5. **Missing Data Handing:**

Here we use dropna function to drop any row that contains a Not-a-Number(NaN), as shown below:

[11]:
```
#Missing Data Handing
data.dropna()
```

[11]:

	Year	Industry_aggregation_NZSIOC	Industry_code_NZSIOC	Industry_name_NZSIOC	Units	Variable_code	Variable_name	Variable_category	Value	Industr
0	2023	Level 1	99999	All industries	Dollars (millions)	H01	Total income	Financial performance	930995	ANZ (exclu
1	2023	Level 1	99999	All industries	Dollars (millions)	H04	Sales, government funding, grants and subsidies	Financial performance	821630	ANZ (exclu
2	2023	Level 1	99999	All industries	Dollars (millions)	H05	Interest, dividends and donations	Financial performance	84354	ANZ (exclu
3	2023	Level 1	99999	All industries	Dollars (millions)	H07	Non-operating income	Financial performance	25010	ANZ (exclu
4	2023	Level 1	99999	All industries	Dollars (millions)	H08	Total expenditure	Financial performance	832964	ANZ (exclu
...	...	...	...	...	...	...	...	...	...	
50980	2013	Level 3	ZZ11	Food product manufacturing	Percentage	H37	Quick ratio	Financial ratios	52	ANZS C112, C
50981	2013	Level 3	ZZ11	Food product manufacturing	Percentage	H38	Margin on sales of goods for resale	Financial ratios	40	ANZS C112, C
50982	2013	Level 3	ZZ11	Food product manufacturing	Percentage	H39	Return on equity	Financial ratios	12	ANZS C112, C
50983	2013	Level 3	ZZ11	Food product manufacturing	Percentage	H40	Return on total assets	Financial ratios	5	ANZS C112, C

6. Sorting DataFrame values:

Here we use the sort_value function to sort the Year column in ascending order and it starts from 2013 and ends by 2023:

[15]:
```
#Sorting DataFrame values
sorted_data = data.sort_values(by='Year',ascending=True)
display(sorted_data)
```

	Year	Industry_aggregation_NZSIOC	Industry_code_NZSIOC	Industry_name_NZSIOC	Units	Variable_code	Variable_name	Variable_category	Value	Indus
50968	2013	Level 3	ZZ11	Food product manufacturing	Dollars (millions)	H25	Current assets	Financial position	9,312	AN C112,
50967	2013	Level 3	ZZ11	Food product manufacturing	Dollars (millions)	H24	Total assets	Financial position	23,102	AN C112,
50966	2013	Level 3	ZZ11	Food product manufacturing	Dollars (millions)	H23	Surplus before income tax	Financial performance	1,227	AN C112,
50965	2013	Level 3	ZZ11	Food product manufacturing	Dollars (millions)	H22	Closing stocks	Financial performance	4,041	AN C112,
50964	2013	Level 3	ZZ11	Food product manufacturing	Dollars (millions)	H21	Opening stocks	Financial performance	3,722	AN C112,
...	...	...	...	...	...	...	...	...	...	
7	2023	Level 1	99999	All industries	Dollars (millions)	H11	Depreciation	Financial performance	30814	AN (exc
3	2023	Level 1	99999	All industries	Dollars (millions)	H07	Non-operating income	Financial performance	25010	AN (exc
0	2023	Level 1	99999	All industries	Dollars (millions)	H01	Total income	Financial performance	930995	AN (exc
19	2023	Level 1	99999	All industries	Dollars (millions)	H29	Other assets	Financial position	1288749	AN (exc
					Dollars		Interest,	Financial		AN

7.Merge Data Frames:

Here for the demonstration of merge function, two dataframes containing the same csv file was merged to get a new dataframe:

[24]:
```
#Merge Data Frames
df1=pd.read_csv('annual-enterprise-survey-2023-financial-year-provisional.csv')
df2=pd.read_csv('annual-enterprise-survey-2023-financial-year-provisional.csv')

df=pd.merge(df1,df2)
print(df)
```

```
       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
...     ...                        ...                 ...
50980  2013                    Level 3                ZZ11
50981  2013                    Level 3                ZZ11
50982  2013                    Level 3                ZZ11
50983  2013                    Level 3                ZZ11
50984  2013                    Level 3                ZZ11

             Industry_name_NZSIOC              Units Variable_code  \
0                 All industries  Dollars (millions)           H01
1                 All industries  Dollars (millions)           H04
2                 All industries  Dollars (millions)           H05
3                 All industries  Dollars (millions)           H07
4                 All industries  Dollars (millions)           H08
...                          ...                ...           ...
50980  Food product manufacturing         Percentage           H37
50981  Food product manufacturing         Percentage           H38
50982  Food product manufacturing         Percentage           H39
50983  Food product manufacturing         Percentage           H40
50984  Food product manufacturing         Percentage           H41

                                      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
...                                              ...                   ...
50980                                   Quick ratio       Financial ratios
50981            Margin on sales of goods for resale       Financial ratios
50982                                 Return on equity       Financial ratios
50983                          Return on total assets       Financial ratios
50984                           Liabilities structure       Financial ratios

        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...
...       ...                                              ...
50980      52  ANZSIC06 groups C111, C112, C113, C114, C115, ...
50981      40  ANZSIC06 groups C111, C112, C113, C114, C115, ...
50982      12  ANZSIC06 groups C111, C112, C113, C114, C115, ...
50983       5  ANZSIC06 groups C111, C112, C113, C114, C115, ...
50984      46  ANZSIC06 groups C111, C112, C113, C114, C115, ...

[50985 rows x 10 columns]
```

8. Apply Function:

Here first we create a column ‘Value_in_Millions’ and add a scale_value function that will convert the Value column data into millions and store it in the new column:

```
[35]: data['Value_in_millions']=0
```

```
[37]: def scale_value(value):
          return value / 1000000


      data['Value_in_Millions'] = data['Value'].apply(scale_value)


      display(data)
```

Industry_name_NZSIOC	Units	Variable_code	Variable_name	Variable_category	Value	Industry_code_ANZSIC06	newColumn	Value_in_millions	Value_in_Millions
All industries	Dollars (millions)	H01	Total income	Financial performance	930995.0	ANZSIC06 divisions A-S (excluding classes K633...	Yes	0	0.930995
All industries	Dollars (millions)	H04	Sales, government funding, grants and subsidies	Financial performance	821630.0	ANZSIC06 divisions A-S (excluding classes K633...	Yes	0	0.821630
All industries	Dollars (millions)	H05	Interest, dividends and donations	Financial performance	84354.0	ANZSIC06 divisions A-S (excluding classes K633...	Yes	0	0.084354
All industries	Dollars (millions)	H07	Non-operating income	Financial performance	25010.0	ANZSIC06 divisions A-S (excluding classes K633...	Yes	0	0.025010
All industries	Dollars (millions)	H08	Total expenditure	Financial performance	832964.0	ANZSIC06 divisions A-S (excluding classes K633	Yes	0	0.832964

9.By using the lambda operator:

Here we use lambda function to create ‘Variable_name_upper’ column which will contain upper case values of ‘Variable_name’ column:

```
•[43]: #lambda funtion to convert Variable name to upper case

       df['Variable_name_upper'] = df['Variable_name'].apply(lambda x: x.upper())

       print("\nDataFrame with Uppercase 'Variable_name':")
       display(df)
```

```
DataFrame with Uppercase 'Variable_name':
```

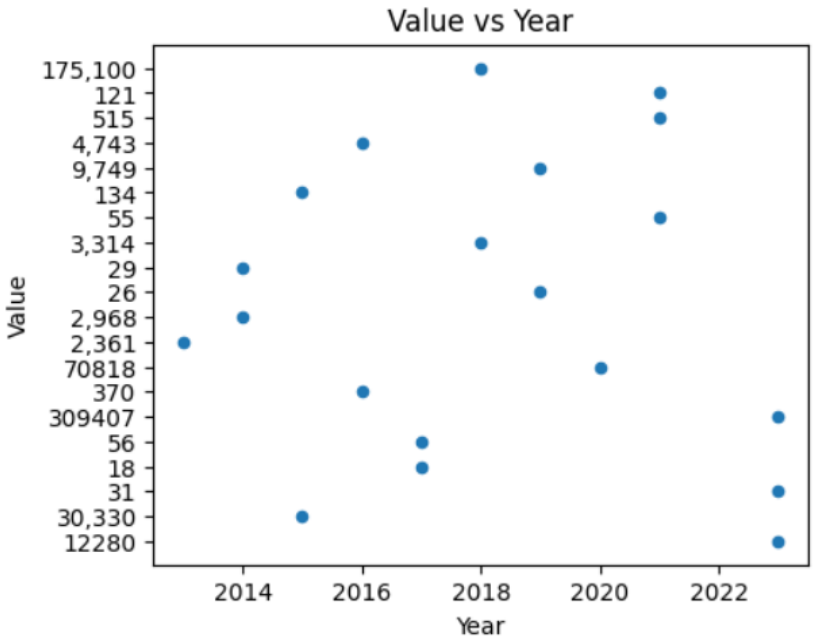
;IOC	Industry_code_NZSIOC	Industry_name_NZSIOC	Units	Variable_code	Variable_name	Variable_category	Value	Industry_code_ANZSIC06	Variable_name_upper
vel 1	99999	All industries	Dollars (millions)	H01	Total income	Financial performance	930995	ANZSIC06 divisions A-S (excluding classes K633...	TOTAL INCOME
vel 1	99999	All industries	Dollars (millions)	H04	Sales, government funding, grants and subsidies	Financial performance	821630	ANZSIC06 divisions A-S (excluding classes K633...	SALES, GOVERNMENT FUNDING, GRANTS AND SUBSIDIES
vel 1	99999	All industries	Dollars (millions)	H05	Interest, dividends and donations	Financial performance	84354	ANZSIC06 divisions A-S (excluding classes K633...	INTEREST, DIVIDENDS AND DONATIONS
vel 1	99999	All industries	Dollars (millions)	H07	Non-operating income	Financial performance	25010	ANZSIC06 divisions A-S (excluding classes K633...	NON-OPERATING INCOME
vel 1	99999	All industries	Dollars (millions)	H08	Total expenditure	Financial performance	832964	ANZSIC06 divisions A-S (excluding classes K633...	TOTAL EXPENDITURE

10.Visualizing DataFrame:

Here 20 rows were taken for sample visualization:

```
•[79]: # Sample 20 rows from the dataset for visualization
df_sample = df.sample(n=20, random_state=42)

# Plot the sample data
df_sample.plot(x='Year', y='Value', kind='scatter', figsize=(5,4))
plt.title("Value vs Year")
plt.xlabel("Year")
plt.ylabel("Value")
plt.show()
```



11.What is the number of columns in the dataset?

```
[62]: num_columns = df.shape[1]
print(f"The number of columns in the dataset is: {num_columns}")
```

```
The number of columns in the dataset is: 11
```

12.Print the name of all the columns.

All the columns are fetched using column function, made into a list and printed, as shown below:

```
[6]: #Printing the column names of the DataFrame
print(list(data.columns))
```

```
['Year', 'Industry_aggregation_NZSIOC', 'Industry_code_NZSIOC', 'Industry_name_NZSIOC', 'Units', 'Variable_code', 'Variable_name', 'Variable_category',
'Value', 'Industry_code_ANZSIC06']
```

13.How is the dataset indexed?

This index function will show us the index information of the dataset.

[63]:
```
# Print the index of the dataset
print(df.index)
```

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

14.What is the number of observations in the dataset?

[64]:
```
# Print the number of observations in the dataset
num_observations = df.shape[0]
print(f"The number of observations in the dataset is: {num_observations}")
```

```
The number of observations in the dataset is: 50985
```

Submitted by:
Aathirainthan P