PYTHON CODING CHALLENGE

HARSHINI V

QUESTIONS:

Annual enterprise survey: 2023 financial year (provisional) – CSV Comma Separated Values, 7.7 MB

1. Printing rows of the Data
2. Printing the column names of the DataFrame
3. Summary of Data Frame
4. Descriptive Statistical Measures of a DataFrame
5. Missing Data Handing
6. Sorting DataFrame values
7. Merge Data Frames
8. Apply Function
9. By using the lambda operator
10. Visualizing DataFrame
11. What is the number of columns in the dataset?
12. print the name of all the columns.
13. How is the dataset indexed?
14. What is the number of observations in the dataset?

DATASET USED:

annual-enterprise-survey-2023-financial-year-provisional.csv

ANSWERS:

1. Printing rows of the Data

```
import pandas as pd
filepath='C:\\Users\\harsh\\Downloads\\annual-enterprise-survey-2023-financial-year-provisional.csv'
data = pd.read_csv(filepath)
display(data.head())
```

EXPLANATION:

The code snippet demonstrates how to load and preview the first few rows of a dataset stored in a CSV file using the Python pandas library. The import pandas as pd statement imports the pandas library, a powerful Python tool for data manipulation and analysis. It is aliased as pd for convenience, allowing concise code. The variable filepath is assigned the full path to the CSV file (annual-enterprise-survey-2023-financial-year-provisional.csv).
The pd.read_csv(filepath) function is called to read the data from the specified file path into a pandas DataFrame. The display(data.head()) function outputs the first five rows of the DataFrame data.

```
# 1. Printing rows of the Data
import pandas as pd
filepath='C:\\Users\\harsh\\Downloads\\annual-enterprise-survey-2023-financial-year-provisional.csv'
data = pd.read_csv(filepath)
display(data.head())
```

	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 ((excluding cl
1	2023	Level 1	99999	All industries	Dollars (millions)	H04	Sales, government funding, grants and subsidies	Financial performance	821630	ANZSIC06 ((excluding cl
2	2023	Level 1	99999	All industries	Dollars (millions)	H05	Interest, dividends and donations	Financial performance	84354	ANZSIC06 ((excluding cl
3	2023	Level 1	99999	All industries	Dollars (millions)	H07	Non-operating income	Financial performance	25010	ANZSIC06 ((excluding cl
4	2023	Level 1	99999	All industries	Dollars (millions)	H08	Total expenditure	Financial performance	832964	ANZSIC06 ((excluding cl

2. Printing the column names of the DataFrame

```
print(data.columns)
```

EXPLANATION:

The code print(data.columns) is used to display the names of all columns in a pandas DataFrame. By passing data.columns to the print() function, the code outputs these column names to the console.

```
# 2. Printing the column names of the DataFrame
print(data.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')
```

3. Summary of Data Frame

print(data.info())

EXPLANATION:

The code print(data.info()) is used to display a concise summary of the DataFrame's structure and content.

```
#3. Summary of DataFrame
print(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
None
```

4. Descriptive Statistical Measures of a DataFrame

data.describe()

EXPLANATION:

The code data.describe() is used to generate and display descriptive statistical measures for the numerical columns in the DataFrame. The describe() method provides a summary of essential statistics, including:

1. **Count**: The number of non-null values in each column.
2. **Mean**: The average value of the column.
3. **Standard Deviation (**std**)**: A measure of the spread or variability of the data.
4. **Minimum (**min**)**: The smallest value in the column.
5. **25th Percentile (**25%**)**: The value below which 25% of the data falls (first quartile).
6. **50th Percentile (**50%**)**: The median value, where half the data falls below this point.
7. **75th Percentile (**75%**)**: The value below which 75% of the data falls (third quartile).
8. **Maximum (**max**)**: The largest value in the column.

This method focuses only on numeric columns and excludes non-numeric data types by default.

```
# 4. Descriptive Statistical Measures of a DataFrame
data.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

5. Missing Data Handing

data.dropna()

EXPLANATION:

The data.dropna() method in pandas is used to handle missing data by removing rows or columns that contain NaN (Not a Number) values. By default, this method drops all rows where at least one element is missing, effectively cleaning the dataset of incomplete records.

```
# 5. Missing Data Handling
data.dropna()
```

	Year	Industry_aggregation_NZSIOC	Industry_code_NZSIOC	Industry_name_NZSIOC	Units	Variable_code	Variable_name	Variable_category	Value	Indust
0	2023	Level 1	99999	All industries	Dollars (millions)	H01	Total income	Financial performance	930995	AN (excl
1	2023	Level 1	99999	All industries	Dollars (millions)	H04	Sales, government funding, grants and subsidies	Financial performance	821630	AN (excl
2	2023	Level 1	99999	All industries	Dollars (millions)	H05	Interest, dividends and donations	Financial performance	84354	AN (excl
3	2023	Level 1	99999	All industries	Dollars (millions)	H07	Non-operating income	Financial performance	25010	AN (excl
4	2023	Level 1	99999	All industries	Dollars (millions)	H08	Total expenditure	Financial performance	832964	AN (excl
...	...	...	...	...	...	...	...	...	...	
50980	2013	Level 3	ZZ11	Food product manufacturing	Percentage	H37	Quick ratio	Financial ratios	52	AN; C112,
50981	2013	Level 3	ZZ11	Food product manufacturing	Percentage	H38	Margin on sales of goods for resale	Financial ratios	40	AN; C112,
50982	2013	Level 3	ZZ11	Food product manufacturing	Percentage	H39	Return on equity	Financial ratios	12	AN; C112,
50983	2013	Level 3	ZZ11	Food product manufacturing	Percentage	H40	Return on total assets	Financial ratios	5	AN; C112,
50984	2013	Level 3	ZZ11	Food product manufacturing	Percentage	H41	Liabilities structure	Financial ratios	46	AN; C112,

50985 rows × 10 columns

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

print(len(data.columns))

EXPLANATION:

The code print(len(data.columns)) is used to calculate and display the total number of columns in a pandas DataFrame.

```
# 11. What is the number of columns in the dataset?
print(len(data.columns))
```

```
10
```

12. Print the name of all the columns

list(data.columns)

EXPLANATION:

The code list(data.columns) is used to convert the column names of a pandas DataFrame into a Python list and then print them to the console.

```
# 12. Print the name of all the columns
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?

data.index

EXPLANATION:

The data.index attribute in pandas is used to access the index (or row labels) of a DataFrame. The index represents the labels or identifiers for the rows of the dataset, similar to how column names identify the features of the data. The data.index provides an Index object that contains the row labels, and it allows you to view, modify, or work with the indexing system of the DataFrame.

```
# 13. How is the dataset indexed?
data.index
```

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

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

len(data)

EXPLANATION:

The code len(data) is used to determine the number of rows in a pandas DataFrame. When applied to a DataFrame, the len() function returns the total count of rows, which corresponds to the number of observations or records in the dataset. This is helpful for understanding the size of the dataset and can be useful when performing data analysis or preprocessing tasks.

```
# 14. What is the number of observations in the dataset?
len(data)
```

```
50985
```