CASE STUDY – 2 (Python)

TRAINEE NAME: Swathi Baskaran

1. Printing rows of the data

```
[1] # Printing rows of the data
    import pandas as pd
    df = pd.read_csv('/content/drive/MyDrive/annual-enterprise-survey-2023-financial-year-provisional (1).csv')
    df.head()
```

	Year	Industry_aggregation_NZSIOC	Industry_code_NZSIOC	Industry_name_NZSIOC	Units	Variable_code	Variable_name	Variable_category	Value	Industry_code_ANZSIC06
0	2023	Level 1	99999	All industries	Dollars (millions)	H01	Total income	Financial performance	930995	ANZSIC06 divisions A-S (excluding classes K633...
1	2023	Level 1	99999	All industries	Dollars (millions)	H04	Sales, government funding, grants and subsidies	Financial performance	821630	ANZSIC06 divisions A-S (excluding classes K633...
2	2023	Level 1	99999	All industries	Dollars (millions)	H05	Interest, dividends and donations	Financial performance	84354	ANZSIC06 divisions A-S (excluding classes K633...
3	2023	Level 1	99999	All industries	Dollars (millions)	H07	Non-operating income	Financial performance	25010	ANZSIC06 divisions A-S (excluding classes K633...
4	2023	Level 1	99999	All industries	Dollars (millions)	H08	Total expenditure	Financial performance	832964	ANZSIC06 divisions A-S (excluding classes K633...

2. Column names of the dataframe

```
[3] # Printing the column names of the DataFrame
    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')
```

3. Summary of dataframe

```
[4] # Summary of Data Frame
    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
```

4. Descriptive Statistical Measures of a dataframe

```
[5]  # Descriptive Statistical Measures of a DataFrame
     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

5. Missing Data Handling

```
[8]  # Missing Data Handing
     df.isna().sum()
     df.fillna(0, inplace = True)
     df.head()
```

	Year	Industry_aggregation_NZSIOC	Industry_code_NZSIOC	Industry_name_NZSIOC	Units	Variable_code	Variable_name	Variable_category	Value	Industry_code_ANZSIC06
0	2023	Level 1	99999	All industries	Dollars (millions)	H01	Total income	Financial performance	930995	ANZSIC06 divisions A-S (excluding classes K633...
1	2023	Level 1	99999	All industries	Dollars (millions)	H04	Sales, government funding, grants and subsidies	Financial performance	821630	ANZSIC06 divisions A-S (excluding classes K633...
2	2023	Level 1	99999	All industries	Dollars (millions)	H05	Interest, dividends and donations	Financial performance	84354	ANZSIC06 divisions A-S (excluding classes K633...
3	2023	Level 1	99999	All industries	Dollars (millions)	H07	Non-operating income	Financial performance	25010	ANZSIC06 divisions A-S (excluding classes K633...
4	2023	Level 1	99999	All industries	Dollars (millions)	H08	Total expenditure	Financial performance	832964	ANZSIC06 divisions A-S (excluding classes K633...

6. Sorting dataframe values

```
[13] # Sorting DataFrame values
     df.sort_values(by = 'Value')
```

	Year	Industry_aggregation_NZSIOC	Industry_code_NZSIOC	Industry_name_NZSIOC	Units	Variable_code	Variable_name	Variable_category	Value	Industry_code_ANZSIC06
27424	2018	Level 4	QQ111	Hospitals	Percentage	H40	Return on total assets	Financial ratios	-1	ANZSIC06 group Q840
47238	2013	Level 4	CC411	Printing	Percentage	H40	Return on total assets	Financial ratios	-1	ANZSIC06 groups C161 and C162
47202	2013	Level 3	CC41	Printing	Percentage	H40	Return on total assets	Financial ratios	-1	ANZSIC06 groups C161 and C162
3270	2023	Level 4	KK121	Life Insurance	Percentage	H40	Return on total assets	Financial ratios	-1	ANZSIC06 group K631
32685	2016	Level 4	AA131	Dairy Cattle Farming	Percentage	H40	Return on total assets	Financial ratios	-1	ANZSIC06 group A016
...	...	...	...	...	...	...	...	...	...	...
49887	2013	Level 4	LL122	Non-Residential Property Operation	Dollars (millions)	H27	Additions to fixed assets	Financial position	S	ANZSIC06 class L671200
31347	2017	Level 4	LL122	Non-Residential Property Operation	Dollars (millions)	H27	Additions to fixed assets	Financial position	S	ANZSIC06 class L671200
40219	2015	Level 3	KK11	Finance	Dollars (millions)	H26	Fixed tangible assets	Financial position	S	ANZSIC06 groups K621, K622, K623, and K624

7. Apply Function

```
[16] # Apply Function
df['Value'] = pd.to_numeric(df['Value'], errors='coerce')

def categorize_value(x):
  if x < 1000000:
    return 'Low'
  elif x < 10000000:
    return 'Medium'
  else:
    return 'High'

df['ValueCategory'] = df['Value'].apply(categorize_value)
df.head()
```

	Year	Industry_aggregation_NZSIOC	Industry_code_NZSIOC	Industry_name_NZSIOC	Units	Variable_code	Variable_name	Variable_category	Value	Industry_code_ANZSIC06	ValueCatego
0	2023	Level 1	99999	All industries	Dollars (millions)	H01	Total income	Financial performance	930995.0	ANZSIC06 divisions A-S (excluding classes K633...	L
1	2023	Level 1	99999	All industries	Dollars (millions)	H04	Sales, government funding, grants and subsidies	Financial performance	821630.0	ANZSIC06 divisions A-S (excluding classes K633...	L
2	2023	Level 1	99999	All industries	Dollars (millions)	H05	Interest, dividends and donations	Financial performance	84354.0	ANZSIC06 divisions A-S (excluding classes K633...	L
3	2023	Level 1	99999	All industries	Dollars (millions)	H07	Non-operating income	Financial performance	25010.0	ANZSIC06 divisions A-S (excluding classes K633...	L
4	2023	Level 1	99999	All industries	Dollars (millions)	H08	Total expenditure	Financial performance	832964.0	ANZSIC06 divisions A-S (excluding classes K633...	L

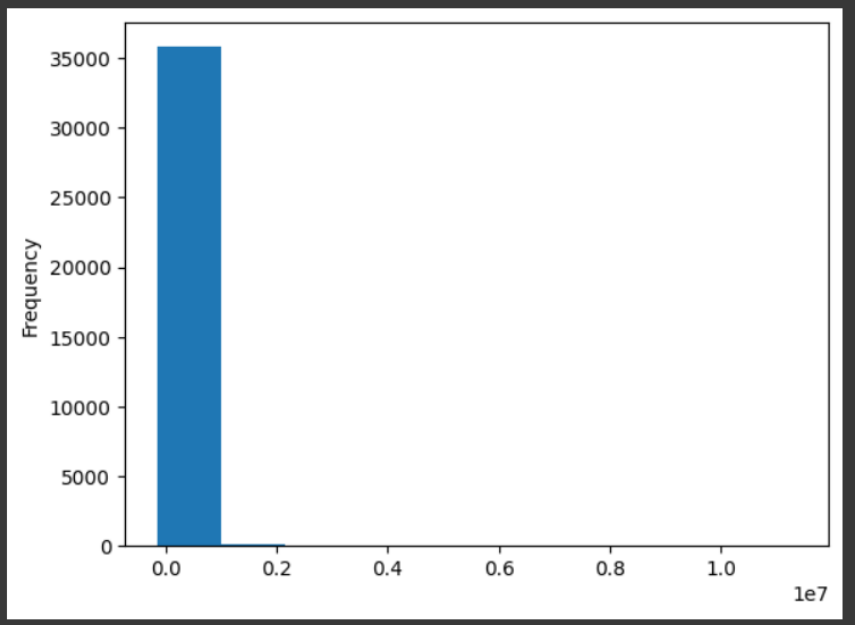
8. Lambda Operator

```
[18] # Lambda Operator

df['ValueCategory'] = df['Value'].apply(
    lambda x: 'Low' if x < 1_000_000 else 'Medium' if x < 10_000_000 else 'High'
)
df.head()
```

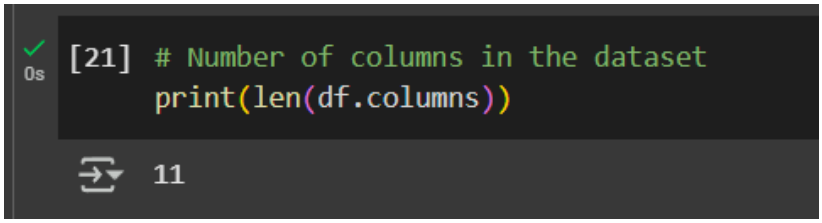
Year	Industry_aggregation_NZSIOC	Industry_code_NZSIOC	Industry_name_NZSIOC	Units	Variable_code	Variable_name	Variable_category	Value	Industry_code_ANZSIC06	ValueCategory
2023	Level 1	99999	All industries	Dollars (millions)	H01	Total income	Financial performance	930995.0	ANZSIC06 divisions A-S (excluding classes K633...	Low
2023	Level 1	99999	All industries	Dollars (millions)	H04	Sales, government funding, grants and subsidies	Financial performance	821630.0	ANZSIC06 divisions A-S (excluding classes K633...	Low
2023	Level 1	99999	All industries	Dollars (millions)	H05	Interest, dividends and donations	Financial performance	84354.0	ANZSIC06 divisions A-S (excluding classes K633...	Low
2023	Level 1	99999	All industries	Dollars (millions)	H07	Non-operating income	Financial performance	25010.0	ANZSIC06 divisions A-S (excluding classes K633...	Low
2023	Level 1	99999	All industries	Dollars (millions)	H08	Total expenditure	Financial performance	832964.0	ANZSIC06 divisions A-S (excluding classes K633...	Low

9. Visualizing Dataframe

```
[20] # Visualizing DataFrame

import matplotlib.pyplot as plt
df['Value'].plot(kind = 'hist')
plt.show()
```

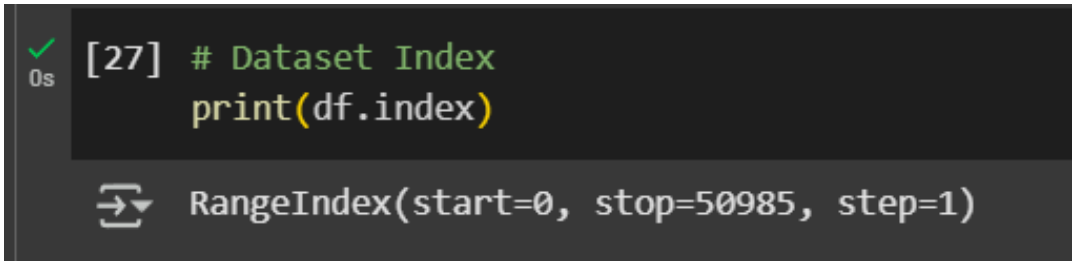


10.Number of columns in the dataset



11.Name of all the columns

12.Dataset Index



13.Number of observations in the dataset