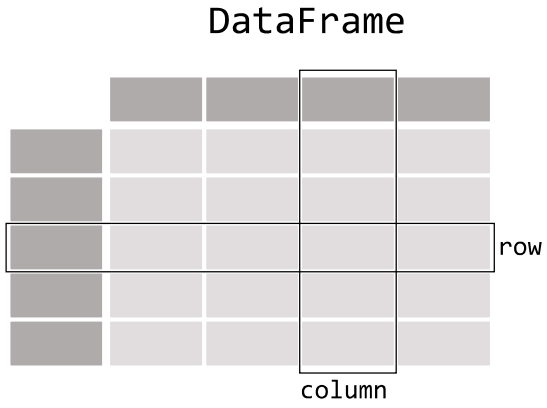
Section 1 - Getting Started with Pandas

Pandas

- It is a Python library used for working with data sets
- It is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool
- It is built on top of another package named Numpy, which provides support for multi-dimensional arrays

Pandas DataFrame

- Pandas is a good tool to use when handling data stored in spreadsheets or databases. In pandas, a data table is called a data frame
- DataFrames are similar to SQL tables or the spreadsheets that you work with in Excel or Calc



What can you do with Pandas DataFrame?

- Data cleansing
- Data fill
- Data normalization
- Merges and joins
- Data visualization
- Statistical analysis
- Data inspection
- Loading and saving data

Installing Pandas

- You can install pandas using PyPI

In [111…
```
!python --version
```

```
Python 3.11.7
```

In [112…
```
# PyPI
!pip install pandas
```

```
Requirement already satisfied: pandas in c:\users\lenovo\anaconda3\lib\site-packages
(2.2.2)
Requirement already satisfied: numpy>=1.23.2 in c:\users\lenovo\anaconda3\lib\site-p
ackages (from pandas) (1.26.4)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\lenovo\anaconda3\l
ib\site-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in c:\users\lenovo\anaconda3\lib\site-pa
ckages (from pandas) (2023.3.post1)
Requirement already satisfied: tzdata>=2022.7 in c:\users\lenovo\anaconda3\lib\site-
packages (from pandas) (2023.3)
Requirement already satisfied: six>=1.5 in c:\users\lenovo\anaconda3\lib\site-packag
es (from python-dateutil>=2.8.2->pandas) (1.16.0)
```

Creating a Pandas DataFrame

I. Import Pandas as pd and Numpy as np

In [113…
```
import pandas as pd
import numpy as np

# NumPy can be used to perform a wide variety of mathematical operations on arrays
```

II-A. Creating a Pandas DataFrame using Dictionaries

In [114…
```
d = {'x': [1, 2, 3], 'y': np.array(['a', 'b', 'c']), 'z': 123}

df1 = pd.DataFrame(d)
df1
```

Out[114…

	x	y	z
0	1	a	123
1	2	b	123
2	3	c	123

II-B. Creating a Pandas DataFrame using Lists

In [115…
```
l = [{'x': 1, 'y': 'a', 'z': 123},
...      {'x': 2, 'y': 'b', 'z': 123},
...      {'x': 3, 'y': 'c', 'z': 123}]
```

```
df2 = pd.DataFrame(l)
df2
```

Out[115…

	x	y	z
0	1	a	123
1	2	b	123
2	3	c	123

II-C. Creating a Pandas DataFrame using NumPy Arrays

In [116…

```
arr = np.array([[1, 'a', 123],
...                 [2, 'b', 123],
...                 [3, 'c', 123]])

df3 = pd.DataFrame(arr, columns=['x', 'y', 'z'])
df3
```

Out[116…

	x	y	z
0	1	a	123
1	2	b	123
2	3	c	123

II-D. Creating a Pandas DataFrame using a CSV File

In [117…

```
# You can get the file from: https://www.kaggle.com/datasets/grosvenpaul/family-inc
# For you to access the file, it is easier if the file is in the same directory as

# After saving the file, you can load it using read_csv()

df4 = pd.read_csv('Family Income and Expenditure.csv')
df4
```

Out[117…

	Total Household Income	Region	Total Food Expenditure	Main Source of Income	Agricultural Household indicator	Bread and Cereals Expenditure
0	480332	CAR	117848	Wage/Salaries	0	42140
1	198235	CAR	67766	Wage/Salaries	0	17329
2	82785	CAR	61609	Wage/Salaries	1	34182
3	107589	CAR	78189	Wage/Salaries	0	34030
4	189322	CAR	94625	Wage/Salaries	0	34820
…	…	…	…	…	…	…
41539	119773	XII - SOCCSKSARGEN	44875	Enterpreneurial Activities	1	23675
41540	137320	XII - SOCCSKSARGEN	31157	Enterpreneurial Activities	1	2691
41541	133171	XII - SOCCSKSARGEN	45882	Enterpreneurial Activities	2	28646
41542	129500	XII - SOCCSKSARGEN	81416	Enterpreneurial Activities	1	29996
41543	128598	XII - SOCCSKSARGEN	78195	Enterpreneurial Activities	1	43485

41544 rows × 60 columns

Selecting a Subset of a DataFrame

In [118…

```
# Since there are a lot of variables in the dataset. We can choose which columns ar
income_region_expenditure = df4[["Total Household Income", "Region", "Total Food Ex
income_region_expenditure
```

Out[118…

	Total Household Income	Region	Total Food Expenditure
0	480332	CAR	117848
1	198235	CAR	67766
2	82785	CAR	61609
3	107589	CAR	78189
4	189322	CAR	94625
…	…	…	…
41539	119773	XII - SOCCSKSARGEN	44875
41540	137320	XII - SOCCSKSARGEN	31157
41541	133171	XII - SOCCSKSARGEN	45882
41542	129500	XII - SOCCSKSARGEN	81416
41543	128598	XII - SOCCSKSARGEN	78195

41544 rows × 3 columns

Filtering and Viewing Specific Rows from a Dataframe

In [119…

```
# We can narrow down the df by specifying a filter
CAR_expenditure = income_region_expenditure[income_region_expenditure["Region"] ==
CAR_expenditure
```

Out[119…

	Total Household Income	Region	Total Food Expenditure
0	480332	CAR	117848
1	198235	CAR	67766
2	82785	CAR	61609
3	107589	CAR	78189
4	189322	CAR	94625
…	…	…	…
40778	90076	CAR	32728
40779	144595	CAR	71342
40780	283262	CAR	41978
40781	146064	CAR	73293
40782	221015	CAR	73259

1725 rows × 3 columns

In [120…
```
# We can view the first 5 columns using head()
CAR_expenditure.head()
```

Out[120…

	Total Household Income	Region	Total Food Expenditure
0	480332	CAR	117848
1	198235	CAR	67766
2	82785	CAR	61609
3	107589	CAR	78189
4	189322	CAR	94625

In [121…
```
# We can view the last 5 columns using tail()
CAR_expenditure.tail()
```

Out[121…

	Total Household Income	Region	Total Food Expenditure
40778	90076	CAR	32728
40779	144595	CAR	71342
40780	283262	CAR	41978
40781	146064	CAR	73293
40782	221015	CAR	73259

Grouping and Getting Summary Statistics

In [122…
```
by_region = income_region_expenditure.groupby('Region').mean().reset_index()
by_region
```

Out[122…

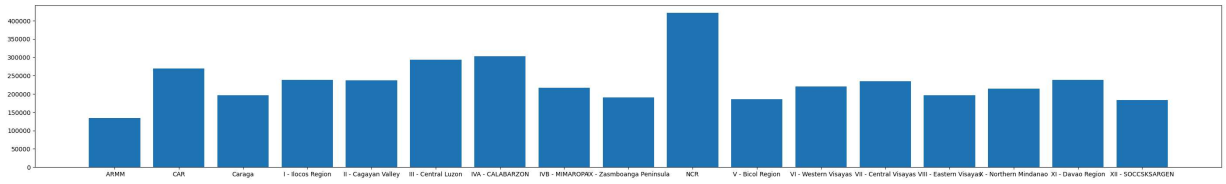
	Region	Total Household Income	Total Food Expenditure
0	ARMM	134746.817616	64931.270463
1	CAR	269540.484638	80352.780290
2	Caraga	196907.376543	71912.659933
3	I - Ilocos Region	238110.084327	80649.937819
4	II - Cagayan Valley	236778.221721	75604.358269
5	III - Central Luzon	292965.181650	99726.701576
6	IVA - CALABARZON	303360.536040	105333.949543
7	IVB - MIMAROPA	216685.124900	70760.293835
8	IX - Zasmboanga Peninsula	191000.908277	69645.318233
9	NCR	420861.861501	127080.456659
10	V - Bicol Region	186105.492718	76811.412217
11	VI - Western Visayas	220481.260260	79829.025956
12	VII - Central Visayas	234909.314050	84307.184179
13	VIII - Eastern Visayas	196736.581087	69833.928969
14	X - Northern Mindanao	214057.779544	64112.585586
15	XI - Davao Region	238115.891251	81126.927228
16	XII - SOCCSKSARGEN	182984.802545	71738.088596

Plotting Your Data

In [123…

```
plt.figure().set_figwidth(35)
plt.bar(by_region['Region'], by_region['Total Household Income'])
```

Out[123… <BarContainer object of 17 artists>



In [ ]: