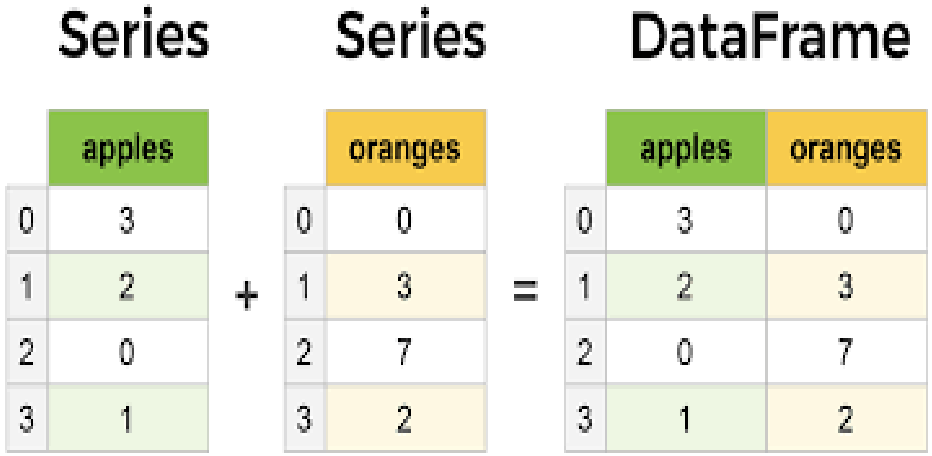
Unit - 1
Data Manipulation with Pandas

Objective

- Data Manipulation with Pandas
- Introduction to Pandas Objects
- Pandas Series Object
- Pandas DataFrame Object
- Data Storage Format
- Reading data from file
- Load CSV file to Pandas
- Groupby Function
- Pandas Data visualization

Data Manipulation with pandas

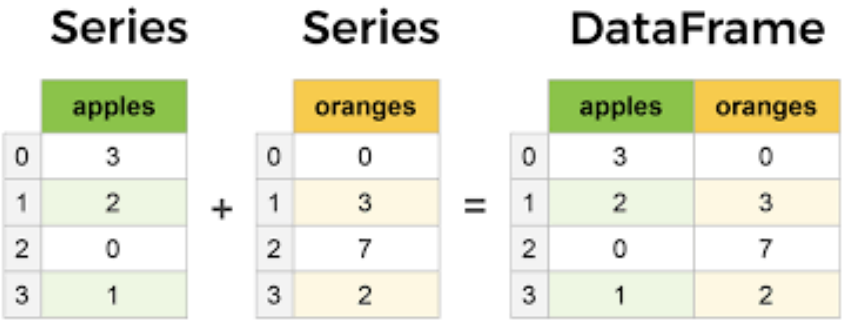
- ***Pandas*** is a fast, powerful, flexible and easy to use open-source data analysis and manipulation tool.
- Pandas is a newer package built on top of NumPy and provides an efficient implementation of a **DataFrame**.



https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcRIlM018KLATJ5zOpUVq0dZmuBCUeBQRts0lg&usqp=CAU

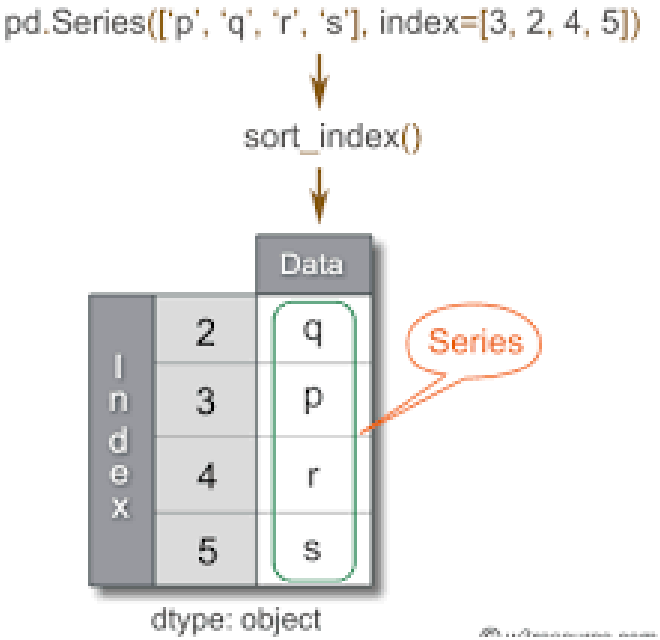
Introducing Pandas Objects

- Pandas objects can be thought of as enhanced versions of NumPy structured arrays in which the rows and columns are identified with labels rather than simple integer indices
- There are three fundamental Pandas data structures:
  - Series
  - DataFrame
  - Index.



The Pandas Series Object

- A Pandas Series is a one-dimensional array of indexed data.
- The Series wraps both a sequence of values and a sequence of indices, which we can access with the values and index attributes.



The Pandas DataFrame Object

- The next fundamental structure in Pandas is the DataFrame.
- The DataFrame can be thought of either as a generalization of a NumPy array, or as a specialization of a Python dictionary.

- ***DataFrame as a generalized NumPy array***
- If a Series is an analog of a one-dimensional array with flexible indices, a DataFrame is an analog of a two-dimensional array with both flexible row indices and flexible column names.

Creating Series from simple datatypes

Creating a Pandas Series

- The Pandas Series can be defined as a one-dimensional array that can store various data types. We can easily convert the list, tuple, and dictionary into series using "series' method.
- The row labels of series are called the index. A Series cannot contain multiple columns. It has the following parameter:

  - data
  - index
  - dtype
  - copy

Data Storage Formats in Pandas

- The different data storage formats available to be manipulated by Pandas library are text, binary and SQL.
- Below is a table containing available ‘readers’ and ‘writers’ functions of the pandas I/O API set with data format and description

Format Type	Data Description	Reader	Writer
text	CSV	read_csv	to_csv
text	Fixed-Width Text File	read_fwf	
text	JSON	read_json	to_json
text	HTML	read_html	to_html
text	Local clipboard	read_clipboard	to_clipboard
binary	MS Excel	read_excel	to_excel
binary	OpenDocument	read_excel	
binary	HDF5 Format	read_hdf	to_hdf
binary	Feather Format	read_feather	to_feather
binary	Parquet Format	read_parquet	to_parquet
binary	ORC Format	read_orc	
binary	Msgpack	read_msgpack	to_msgpack
binary	Stata	read_stata	to_stata
binary	SAS	read_sas	
binary	SPSS	read_spss	
binary	Python Pickle Format	read_pickle	to_pickle
SQL	SQL	read_sql	to_sql
SQL	Google BigQuery	read_gbq	to_gbq

CSV file and JSON file

What is CSV file?

- A CSV is a comma-separated values file, which allows data to be saved in a tabular format.
- CSV files can be used with many spreadsheets program, such as Microsoft Excel or Google Spreadsheets.
- They differ from other spreadsheet file types because you can only have a single sheet in a file, they cannot save cell, column, or row. Also, you cannot save formulas in this format.

Why are .CSV files used?

- CSV files are plain-text files, making them easier for the website developer to create
- Since they're plain text, they're easier to import into a spreadsheet or another storage database, regardless of the specific software you're using.
- To better organize large amounts of data.

How do I save CSV files?

Under the "File name" section in the "Save As" tab, you can select "Save as type" and change it to "CSV (Comma delimited) (*.csv).

What is a JSON file?

- A JSON file is a file that stores simple data structures and objects in JavaScript Object Notation (JSON) format, which is a standard data interchange format.

Structures of JSON

- JSON supports two widely used (amongst programming languages) data structures.
  - A collection of name/value pairs.
  - An ordered list of values.

Reading data from files

Reference: https://realpython.com/pandas-read-write-files

Load CSV files to Python Pandas

- The basic process of loading data from a CSV file into a Pandas DataFrame is achieved using the “read_csv” function in Pandas.

- *pd.read_csv(path/xyz.csv)*

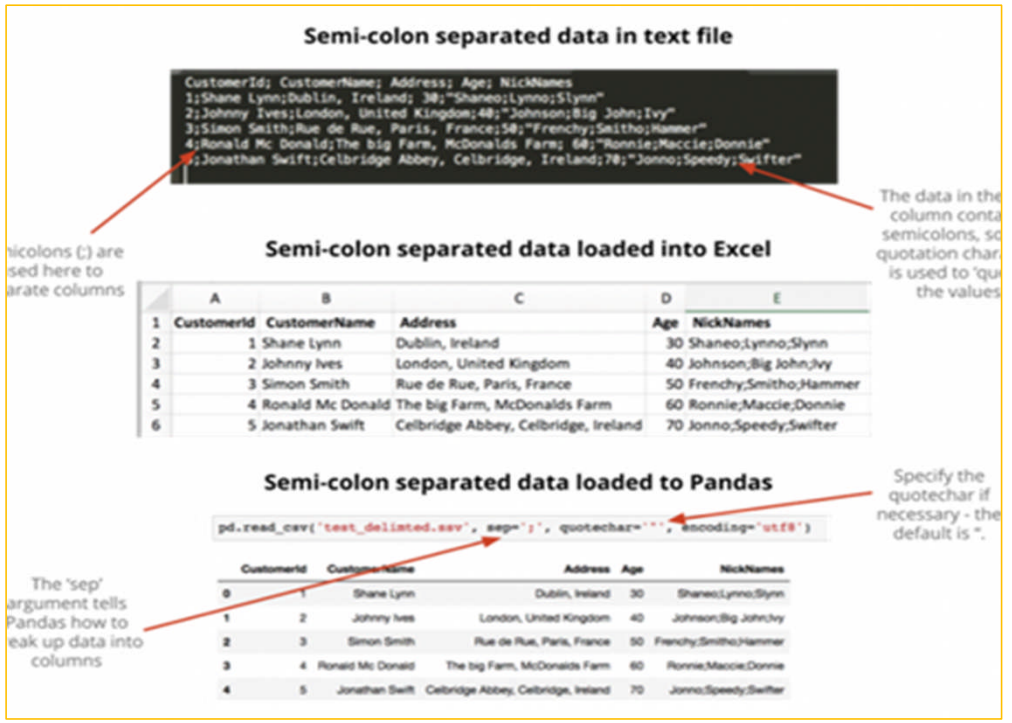
```
In [18]: pd.read_csv("../pokemon.csv", header=[6,3,5,7], squeeze = True,
```

Out[18]:

	Charizard	Fire
	Venusaur	Grass
	Charmeleon	Fire
	Squirtle	Water
0	Wartortle	Water
1	Blastoise	Water
2	Caterpie	Bug
3	Metapod	Bug

Delimiters in Text Fields – Quote char

- The quote character can be specified in Pandas read_csv using the quote char argument.



Reference: https://www.shanelynn.ie/python-pandas-read-csv-load-data-from-csv-files

Python – Paths, Folders, Files

- When you specify a filename to Pandas.read_csv, Python will look in your “current working directory“.

Reference: https://www.shanelynn.ie/python-pandas-read-csv-load-data-from-csv-files

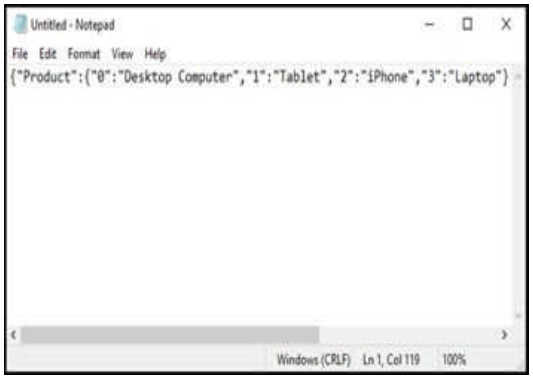
Pandas CSV File Loading Errors

- FileNotFoundError
- UnicodeDecodeError
- pandas.parser.CParserError

Load JSON files to Python Pandas

Product	Price
Desktop Computer	700
Tablet	250
iPhone	800
Laptop	1200

Step 1: Prepare the JSON String



Step 2: Create the JSON File.

```
                Product  Price
0  Desktop Computer      700
1            Tablet      250
2            iPhone      800
3            Laptop     1200
```

Finally, load your JSON file into Pandas DataFrame.

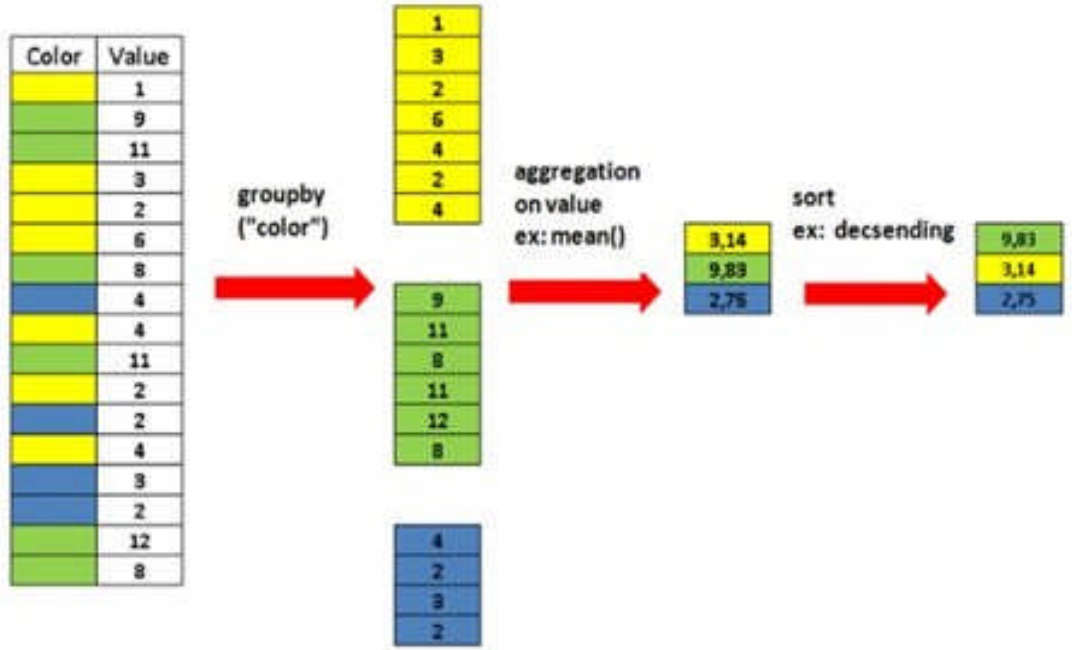
3 different JSON strings

Below are 3 different ways that you could capture the data as JSON strings.

- Index orientation
- Values orientation
- Column's orientation

Groupby Methods

- Pandas dataframe.groupby() function is used to split the data into groups based on some criteria. pandas objects can be split on any of their axes.



Reference: https://towardsdatascience.com/pandas-groupby-explained-453692519d0

Groupby output format – Series or DataFrame?

- As a rule of thumb, if you calculate more than one column of results, your result will be a Data frame.
- For a single column of results, the agg function, by default, will produce a Series.

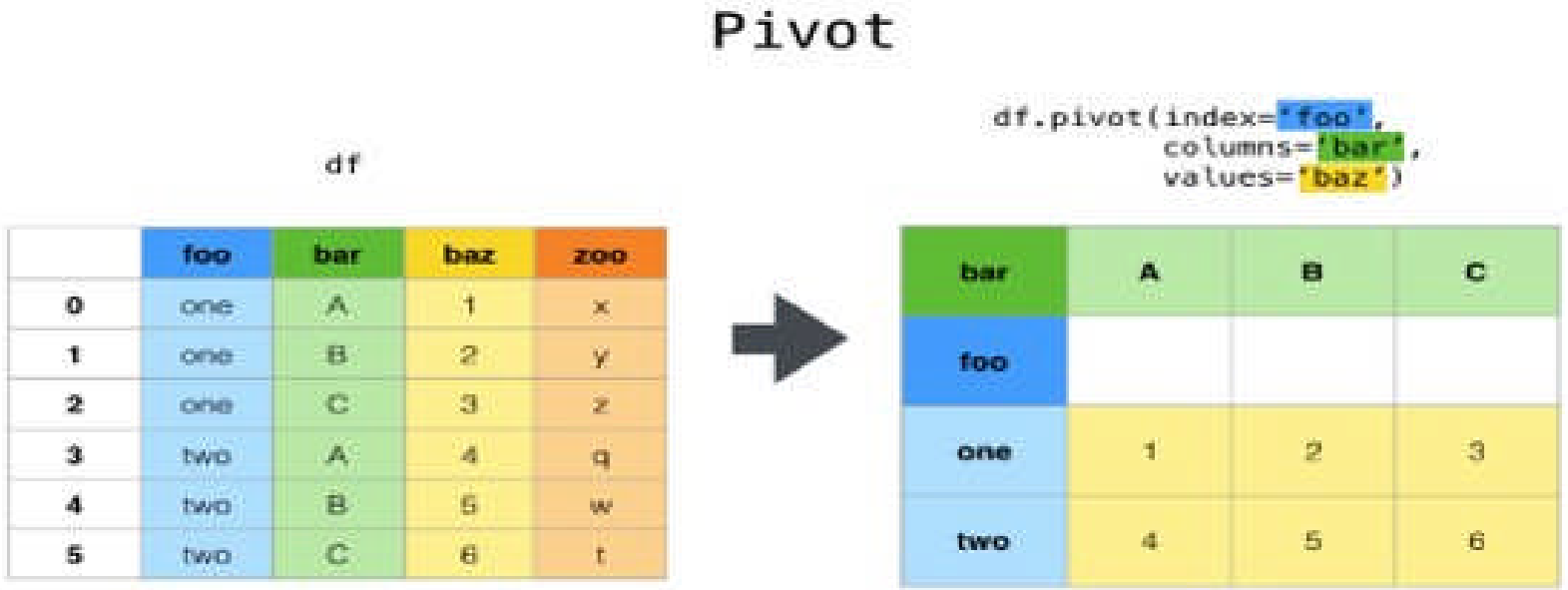
```
In [35]: data.groupby('month', as_index=False).agg({"duration": "sum"})
```

Out[35]:

	month	duration
0	2014-11	26639.441
1	2014-12	14641.870
2	2015-01	18223.299
3	2015-02	15522.299
4	2015-03	22750.441

Pivot Tables

- It’s a table of statistics that helps summarize the data of a larger table by “pivoting” that data.



df

	foo	bar	baz	zoo
0	one	A	1	x
1	one	B	2	y
2	one	C	3	z
3	two	A	4	q
4	two	B	5	w
5	two	C	6	t

bar	A	B	C
foo			
one	1	2	3
two	4	5	6

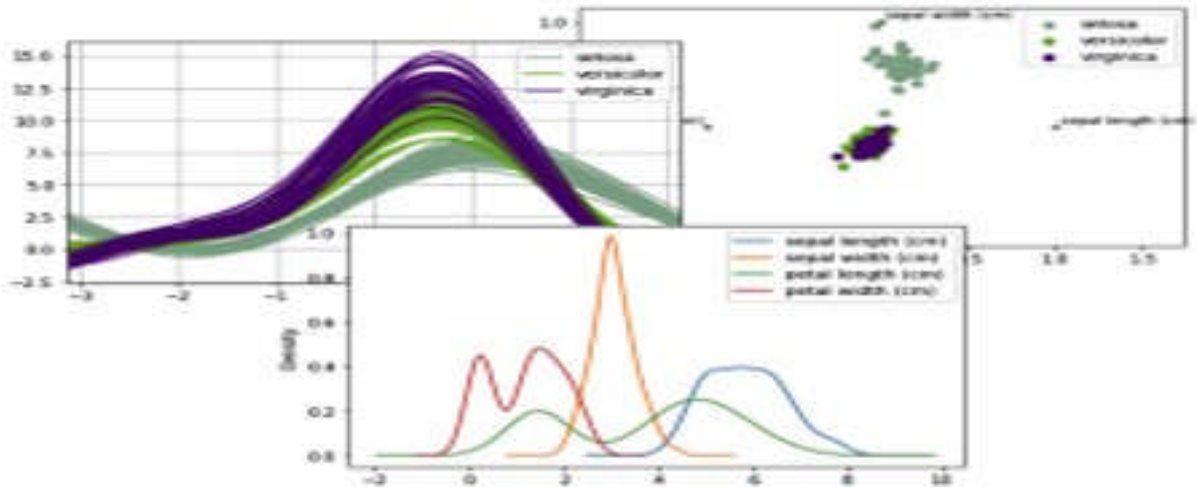
How to Build a Pivot Table in Python

In Pandas, we can construct a pivot table using the following syntax:

```
pandas.pivot_table(data, values=None, index=None, columns=None,
aggfunc='mean', fill_value=None, margins=False, dropna=True,
margins_name='All', observed=False)
```

Pandas Plotting

Plotting in pandas utilises the matplotlib API so in order to create visualisations, you will need to also import this library alongside pandas.



https://towardsdatascience.com/the-best-pandas-plotting-features-c9789e04a5a0

Plot a Scatter Diagram using Pandas

- Scatter plots are used to depict a relationship between two variables.

Step 1: Prepare the data

Unemployment_Rate	Stock_Index_Price
6.1	1500
5.8	1520
5.7	1525
5.7	1523
5.8	1515
5.6	1540
5.5	1545
5.3	1560
5.2	1555
5.2	1565

Plot a Scatter Diagram using Pandas

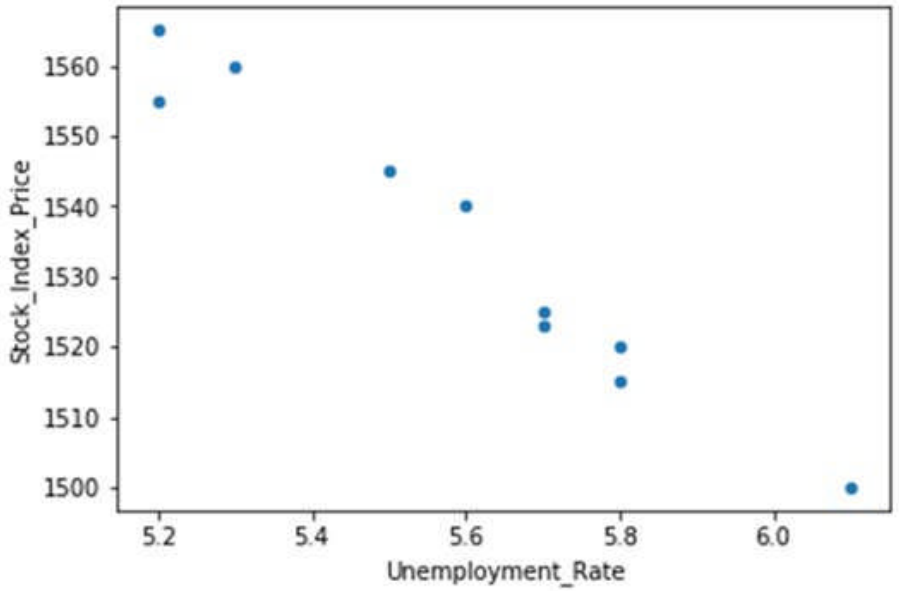
Step 2: Create the DataFrame

	Unemployment_Rate	Stock_Index_Price
0	6.1	1500
1	5.8	1520
2	5.7	1525
3	5.7	1523
4	5.8	1515
5	5.6	1540
6	5.5	1545
7	5.3	1560
8	5.2	1555
9	5.2	1565

Plot a Scatter Diagram using Pandas

Step 3:

Plot the DataFrame using Pandas



Plot a Line Chart using Pandas

- Line charts are often used to display trends overtime.

Step 1: Prepare the data

Year	Unemployment_Rate
1920	9.8
1930	12
1940	8
1950	7.2
1960	6.9
1970	7
1980	6.5
1990	6.2
2000	5.5
2010	6.3

Plot a Line Chart using Pandas

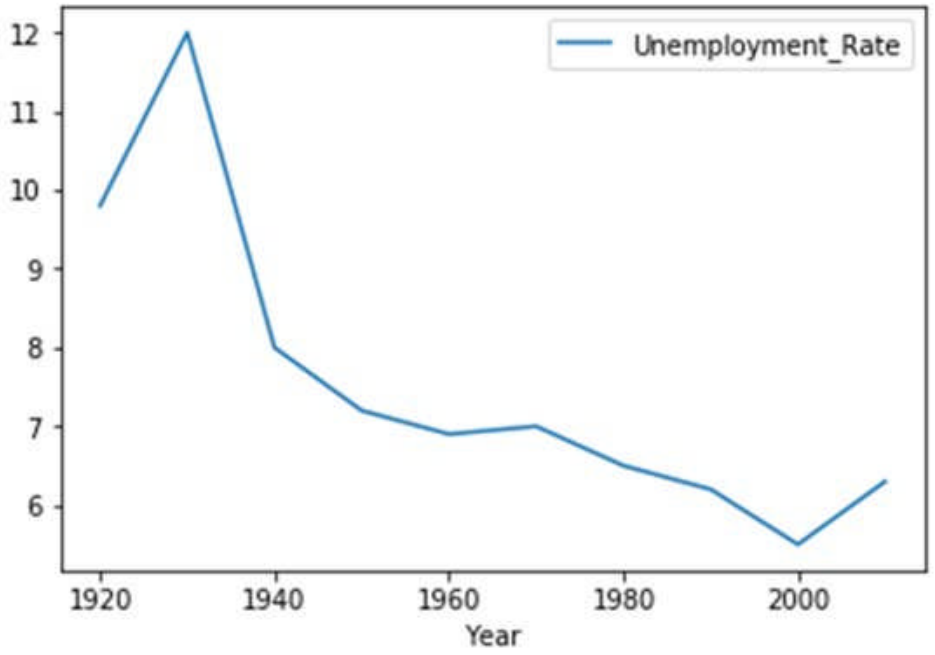
Step 2: Create the DataFrame

	Year	Unemployment_Rate
0	1920	9.8
1	1930	12.0
2	1940	8.0
3	1950	7.2
4	1960	6.9
5	1970	7.0
6	1980	6.5
7	1990	6.2
8	2000	5.5
9	2010	6.3

Plot a Line Chart using Pandas

Step 3:

Plot the DataFrame using Pandas



Plot a Bar Chart using Pandas

Bar charts are used to display categorical data.

Step 1: Prepare the data

Country	GDP_Per_Capita
USA	45000
Canada	42000
Germany	52000
UK	49000
France	47000

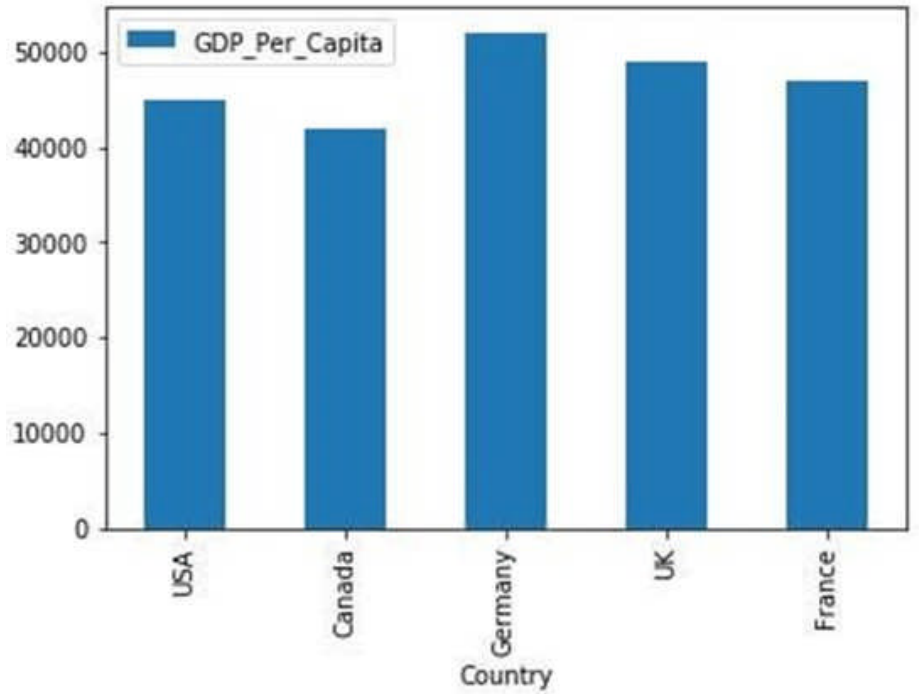
Plot a Bar Chart using Pandas

Step 2: Create the DataFrame

	Country	GDP_Per_Capita
0	USA	45000
1	Canada	42000
2	Germany	52000
3	UK	49000
4	France	47000

Plot a Bar Chart using Pandas

Step 3: Plot the DataFrame using Pandas



Plot a Pie Chart using Pandas

A pie chart (or a circle chart) is a circular statistical graphic, which is divided into slices to illustrate numerical proportion.

Step 1: Prepare the data

Tasks Pending	300
Tasks Ongoing	500
Tasks Completed	700

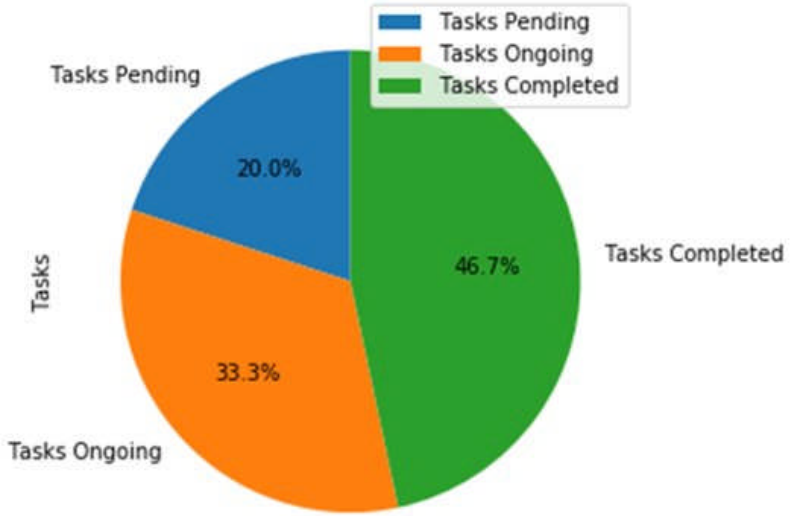
Plot a Pie Chart using Pandas

Step 2: Create the DataFrame

	Tasks
Tasks Pending	300
Tasks Ongoing	500
Tasks Completed	700

Plot a Pie Chart using Pandas

Step 3: Plot the DataFrame using Pandas



References

1. https://pandas.pydata.org/pandas-docs/stable/user_guide/io.html
2. https://www.bigcommerce.com/ecommerce-answers/what-csv-file-and-what-does-it-mean-my-ecommerce-business/#:~:text=A%20CSV%20is%20a%20comma,Microsoft%20Excel%20or%20Google%20Spreadsheets.
3. https://fileinfo.com/extension/json
4. https://www.w3resource.com/JSON/structures.php
5. https://www.shanelynn.ie/python-pandas-read-csv-load-data-from-csv-files/