Data Storage Formats in Pandas

The different data storage formats available to be manipulated by Pandas library are text, binary and SQL.

Disclaimer: The content is curated for educational purposes only.
© Edunet Foundation. All rights reserved.

Format Type	Data Description	Reader	Writer
text	CSV	read_csv	to_csv
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	HDF5 Format	read_hdf	to_hdf
binary	Feather Format	read_feather	to_feather
binary	Parquet Format	read_parquet	to_parquet
binary	Msgpack	read_msgpack	to_msgpack
binary	Stata	read_stata	to_stata
binary	SAS	read_sas	
binary	Python Pickle Format	read_pickle	to_pickle
SQL	SQL	read_sql	to_sql
SQL	Google Big Query	read_gbq	to_gbq

Reference: https://www.programmersought.com/article/4943826582

© Edunet Foundation. All rights reserved.

CSV file and JSON file

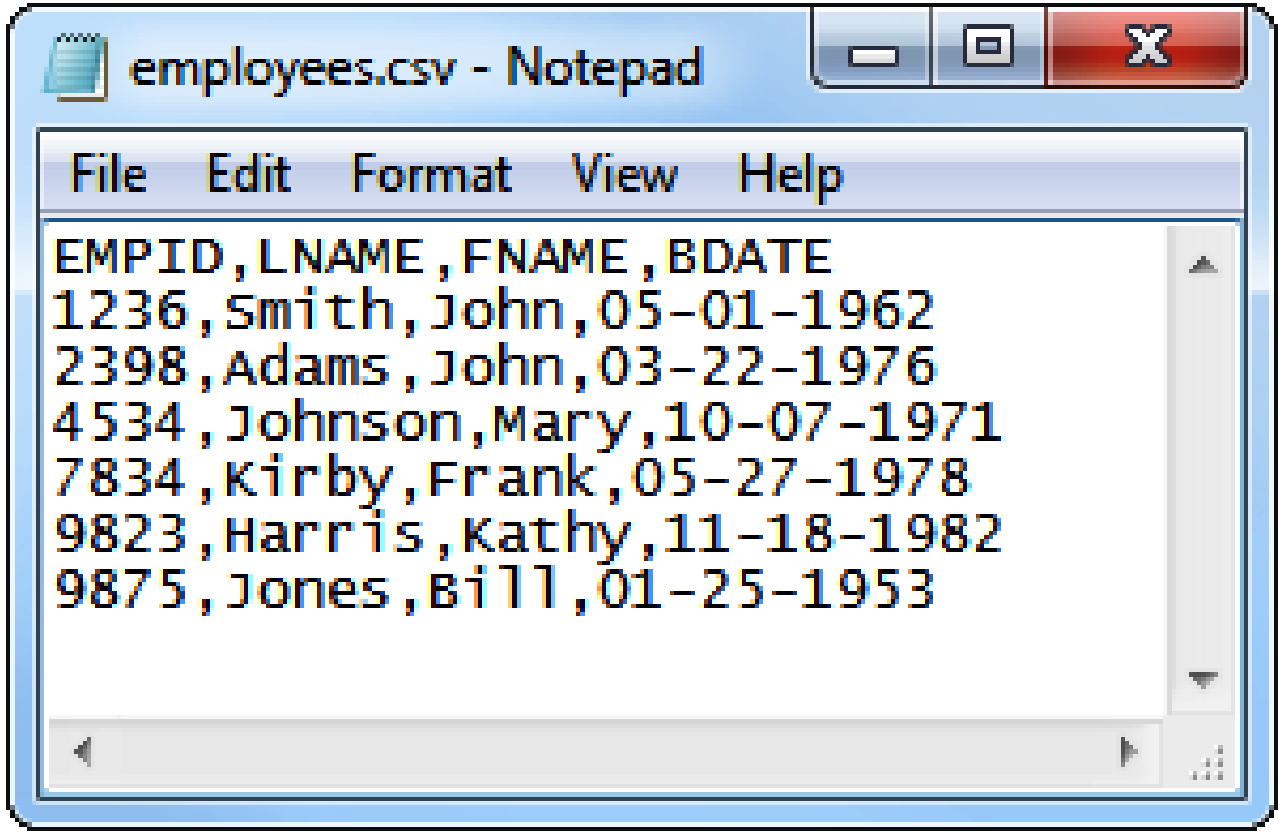
https://pngtree.com/freepng/csv-file-document-icon_4175842.html

https://pngtree.com/freepng/json-file-document-icon_4172477.html

© Edunet Foundation. All rights reserved.

What is CSV file?

A CSV is a comma-separated values file, which allows data to be saved in a tabular format.

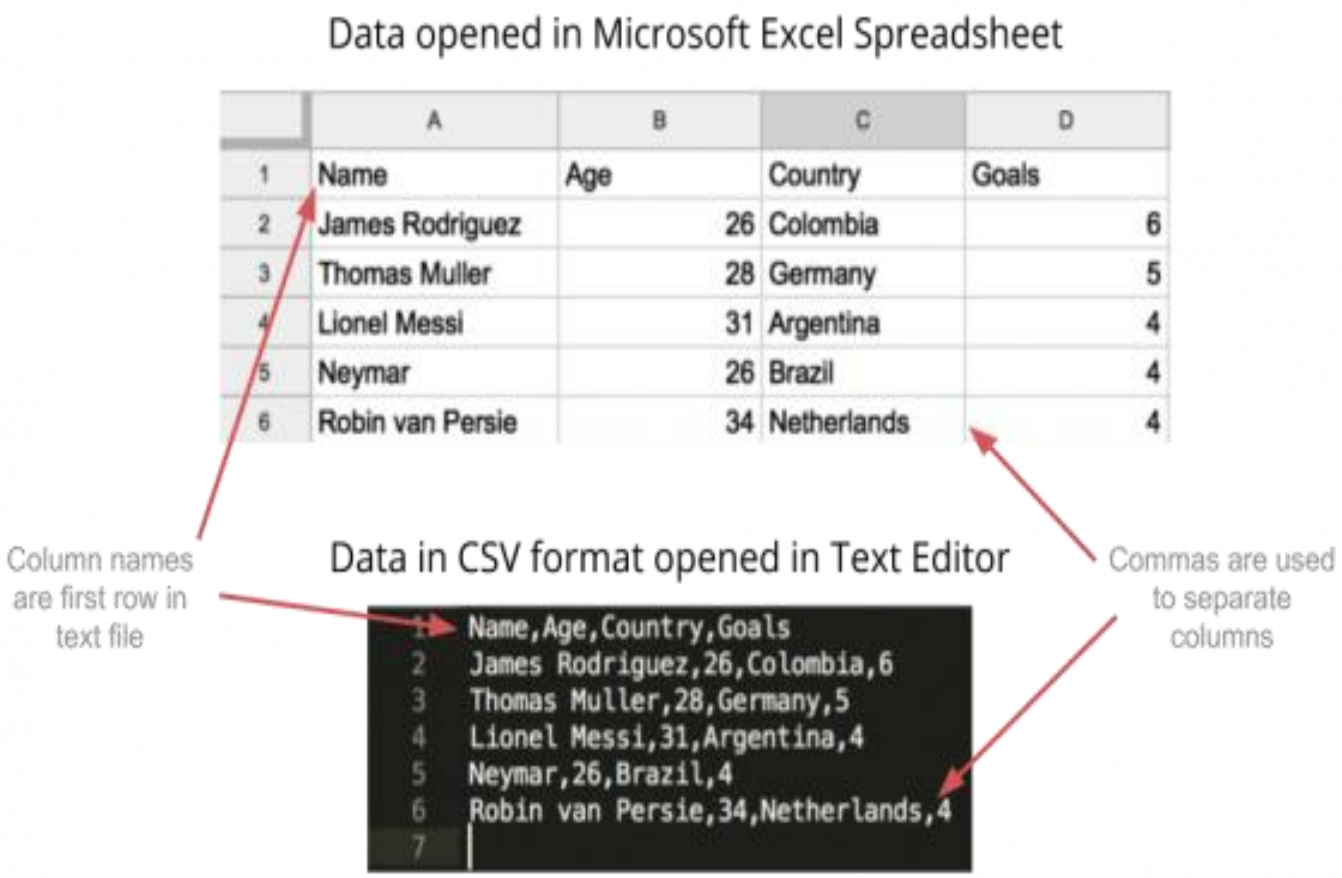


Reference: https://www.goanywhere.com/managed-file-transfer/more/tutorials/how-to-import-csv-file-into-database

© Edunet Foundation. All rights reserved.

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.



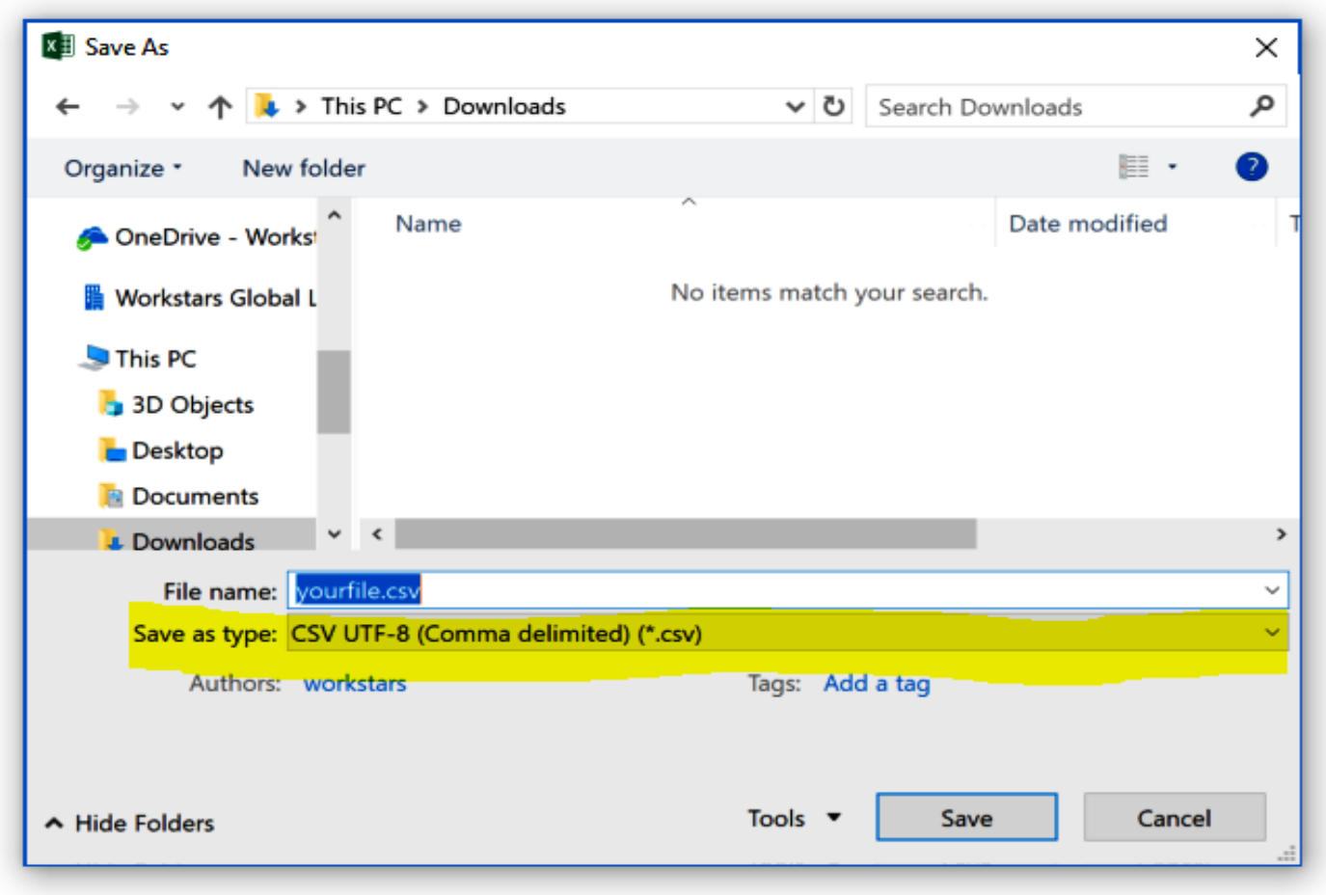
	A	B	C	D
1	Name	Age	Country	Goals
2	James Rodriguez	26	Colombia	6
3	Thomas Muller	28	Germany	5
4	Lionel Messi	31	Argentina	4
5	Neymar	26	Brazil	4
6	Robin van Persie	34	Netherlands	4

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

© Edunet Foundation. All rights reserved.

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).



Reference: https://docs.workstars.com/en/latest/howto/save-csv-utf8.html

© Edunet Foundation. All rights reserved.

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.

```
{
    "Product": {
        "0": "Desktop Computer",
        "1": "Tablet",
        "2": "iPhone",
        "3": "Laptop"
    },
    "Price": {
        "0": 700,
        "1": 250,
        "2": 800,
        "3": 1200
    }
}
```

Reference: https://datatofish.com/export-pandas-dataframe-json

© Edunet Foundation. All rights reserved.

Structures of JSON

JSON supports two widely used (amongst programming languages) data structures.

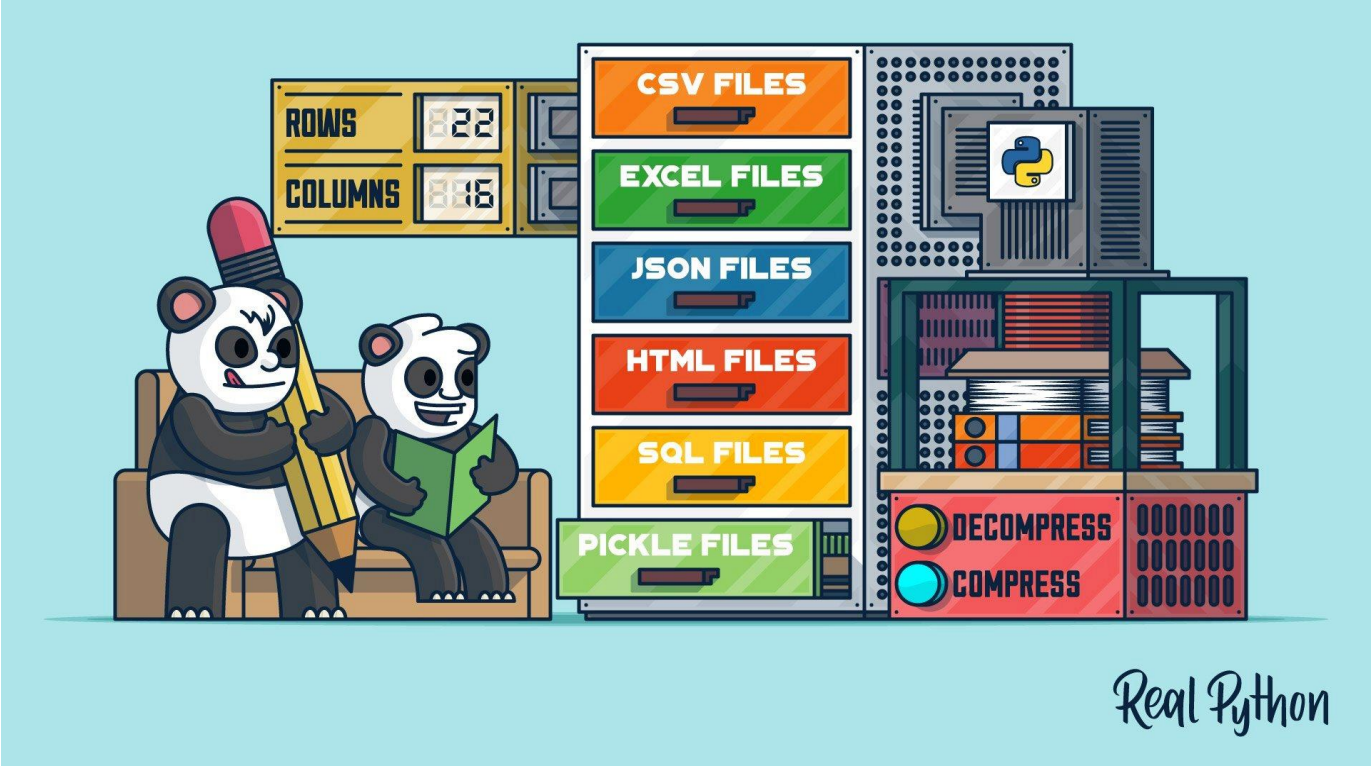
- A collection of name/value pairs.
- An ordered list of values.

```
{                                                        Object Starts
    "Title": "The Cuckoo's Calling"
    "Author": "Robert Galbraith",
    "Genre": "classic crime novel",
    "Detail": {                                          Object Starts
        "Publisher": "Little Brown"                      Value string
        "Publication_Year": 2013,                        Value number
        "ISBN-13": 9781408704004,
        "Language": "English",
        "Pages": 494
    }                                                    Object ends
    "Price": [                                           Array starts
        {                                                Object Starts
            "type": "Hardcover",
            "price": 16.65,
        }                                                Object ends
        {                                                Object Starts
            "type": "Kindle Edition",
            "price": 7.03,
        }                                                Object ends
    ]                                                    Array ends
}                                                        Object ends
```

Reference: https://www.w3resource.com/JSON/introduction.php

© Edunet Foundation. All rights reserved.

Reading data from files



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

© Edunet Foundation. All rights reserved.

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.

```
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

Reference: https://www.geeksforgeeks.org/python-read-csv-using-pandas-read_csv

© Edunet Foundation. All rights reserved.

Delimiters in Text Fields – Quotechar

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

Semi-colon separated data in text file

```
CustomerId; CustomerName; Address; Age; NickNames
1;Shane Lynn;Dublin, Ireland; 30;"Shaneo;Lynno;Slynn"
2;Johnny Ives;London, United Kingdom;40;"Johnson;Big John;Ivy"
3;Simon Smith;Rue de Rue, Paris, France;50;"Frenchy;Smitho;Hammer"
4;Ronald Mc Donald;The big Farm, McDonalds Farm; 60;"Ronnie;Maccie;Donnie"
5;Jonathan Swift;Celbridge Abbey, Celbridge, Ireland;70;"Jonno;Speedy;Swifter"
```

nicolons (;) are sed here to arate columns

The data in the column conta semicolons, so quotation char is used to 'qu the values

Semi-colon separated data loaded into Excel

	A	B	C	D	E
1	CustomerId	CustomerName	Address	Age	NickNames
2	1	Shane Lynn	Dublin, Ireland	30	Shaneo;Lynno;Slynn
3	2	Johnny Ives	London, United Kingdom	40	Johnson;Big John;Ivy
4	3	Simon Smith	Rue de Rue, Paris, France	50	Frenchy;Smitho;Hammer
5	4	Ronald Mc Donald	The big Farm, McDonalds Farm	60	Ronnie;Maccie;Donnie
6	5	Jonathan Swift	Celbridge Abbey, Celbridge, Ireland	70	Jonno;Speedy;Swifter

Semi-colon separated data loaded to Pandas

Specify the quotechar if necessary - the default is ".

```
pd.read_csv('test_delimted.ssv', sep=';', quotechar='"', encoding='utf8')
```

The 'sep' argument tells Pandas how to eak up data into columns

	CustomerId	CustomerName	Address	Age	NickNames
0	1	Shane Lynn	Dublin, Ireland	30	Shaneo;Lynno;Slynn
1	2	Johnny Ives	London, United Kingdom	40	Johnson;Big John;Ivy
2	3	Simon Smith	Rue de Rue, Paris, France	50	Frenchy;Smitho;Hammer
3	4	Ronald Mc Donald	The big Farm, McDonalds Farm	60	Ronnie;Maccie;Donnie
4	5	Jonathan Swift	Celbridge Abbey, Celbridge, Ireland	70	Jonno;Speedy;Swifter

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

© Edunet Foundation. All rights reserved.

Python – Paths, Folders, Files

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

```
In [26]: pd.read_csv('file_not_in_right_place.csv')

---------------------------------------------------------------------------
FileNotFoundError                         Traceback (most recent call last)
<ipython-input-26-f3609a36b9ff> in <module>()
----> 1 pd.read_csv('file_not_in_right_place.csv')

~/Envs/analysis/lib/python3.6/site-packages/pandas/io/parsers.py in parser_f(filepath_or_buffer, sep, delimiter, head
er, names, index_col, usecols, squeeze, prefix, mangle_dupe_cols, dtype, engine, converters, true_values, false_value
s, skipinitialspace, skiprows, nrows, na_values, keep_default_na, na_filter, verbose, skip_blank_lines, parse_dates,
infer_datetime_format, keep_date_col, date_parser, dayfirst, iterator, chunksize, compression, thousands, decimal, li
neterminator, quotechar, quoting, escapechar, comment, encoding, dialect, tupleize_cols, error_bad_lines, warn_bad_li
nes, skipfooter, doublequote, delim_whitespace, low_memory, memory_map, float_precision)
    676                     skip_blank_lines=skip_blank_lines)
    677
--> 678         return _read(filepath_or_buffer, kwds)
    679
    680     parser_f.__name__ = name

~/Envs/analysis/lib/python3.6/site-packages/pandas/io/parsers.py in _read(filepath_or_buffer, kwds)
    438
    439     # Create the parser.
--> 440     parser = TextFileReader(filepath_or_buffer, **kwds)
    441
    442     if chunksize or iterator:

~/Envs/analysis/lib/python3.6/site-packages/pandas/io/parsers.py in __init__(self, f, engine, **kwds)
    785             self.options['has_index_names'] = kwds['has_index_names']
    786
--> 787         self._make_engine(self.engine)
    788
    789     def close(self):
```

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

© Edunet Foundation. All rights reserved.

Pandas CSV File Loading Errors

- FileNotFoundError
- UnicodeDecodeError
- pandas.parser.CParserError

© Edunet Foundation. All rights reserved.

Load JSON files to Python Pandas

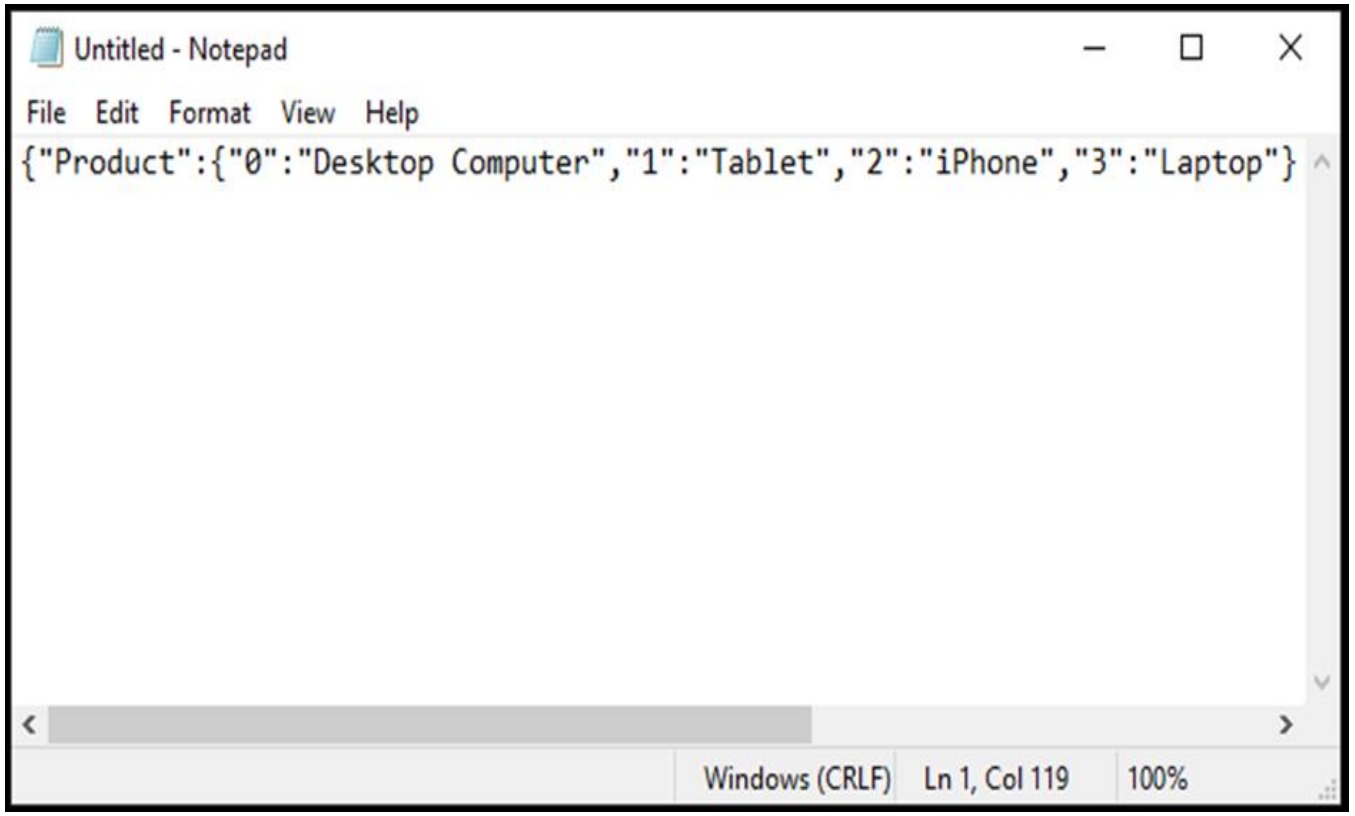
Step 1: Prepare the JSON String

Product	Price
Desktop Computer	700
Tablet	250
iPhone	800
Laptop	1200

© Edunet Foundation. All rights reserved.

Load JSON files to Python Pandas

Step 2: Create the JSON File.

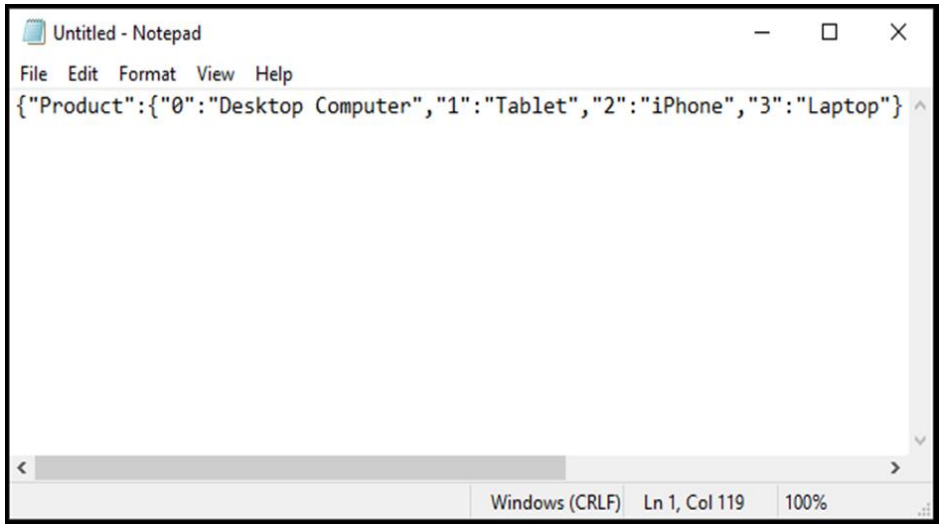


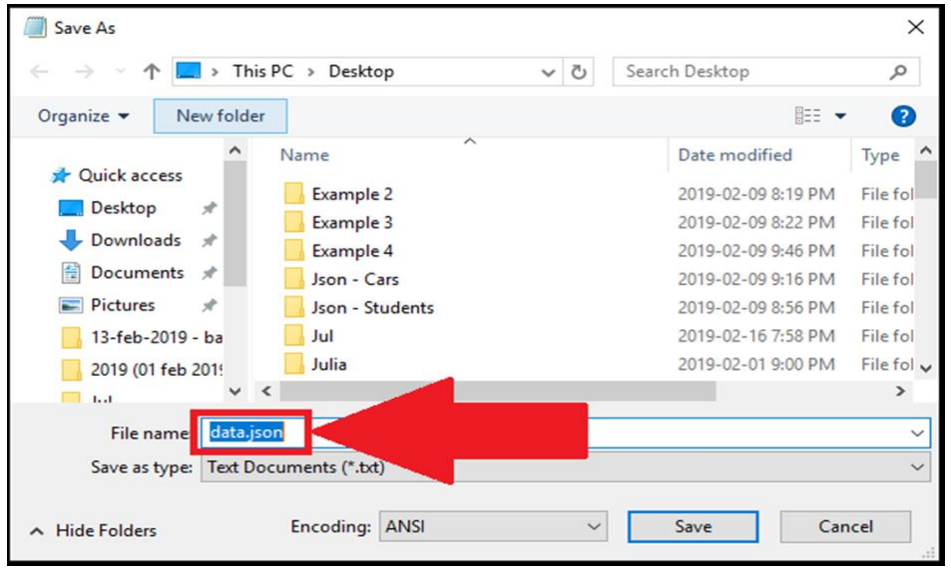
https://datatofish.com/load-json-pandas-dataframe

© Edunet Foundation. All rights reserved.

Load JSON files to Python Pandas

Step 2: Create the JSON File.





https://datatofish.com/load-json-pandas-dataframe

© Edunet Foundation. All rights reserved.

Load JSON files to Python Pandas

Finally, load your JSON file into Pandas DataFrame.

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

© Edunet Foundation. All rights reserved.

3 different JSON strings

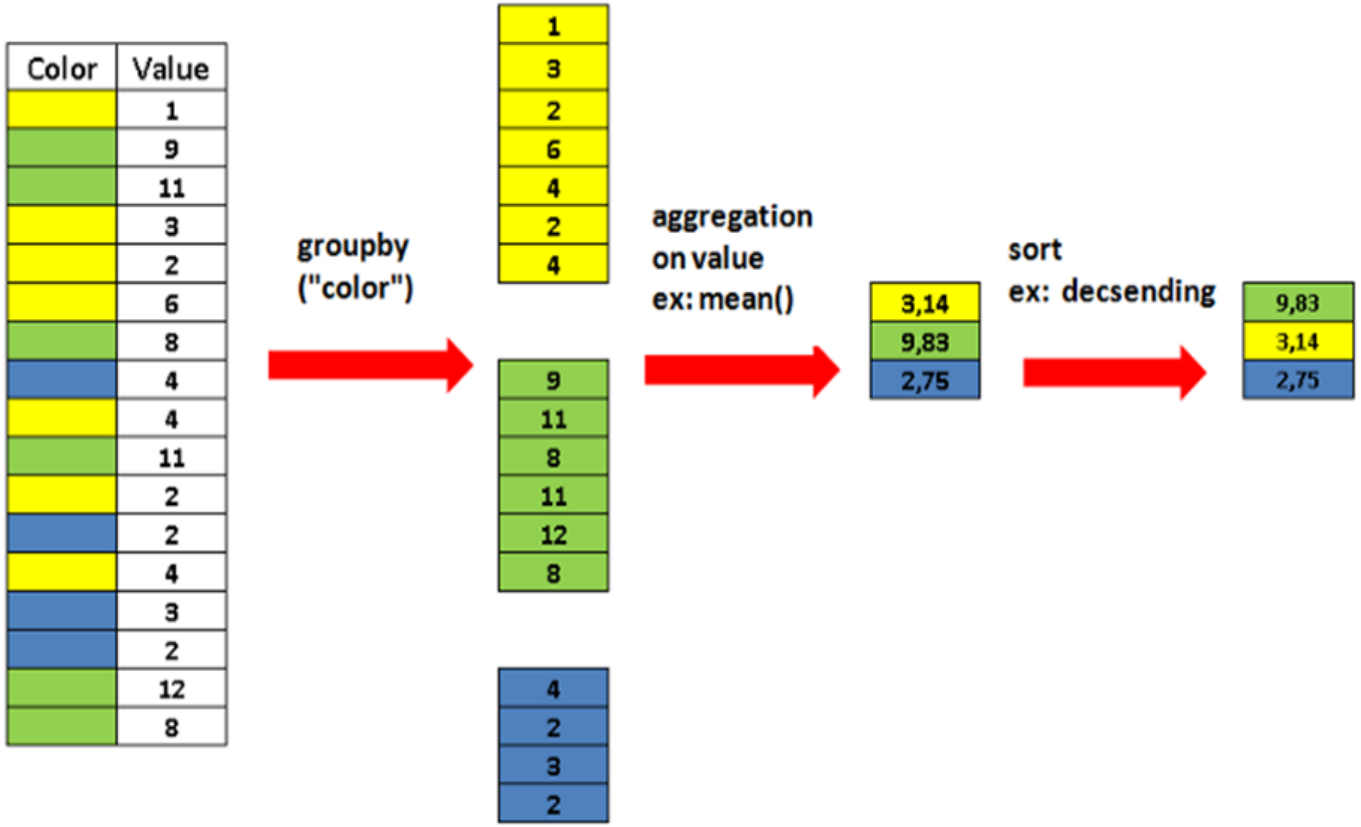
Below are 3 different ways that you could capture the data as JSON strings.

- Index orientation
- Values orientation
- Column's orientation

© Edunet Foundation. All rights reserved.

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

© Edunet Foundation. All rights reserved.

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 Dataframe. 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

Reference: https://www.shanelynn.ie/summarising-aggregation-and-grouping-data-in-python-pandas

© Edunet Foundation. All rights reserved.

Pivot Tables

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

Pivot

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

```
df.pivot(index='foo',
         columns='bar',
         values='baz')
```

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

https://pandas.pydata.org/pandas-docs/version/0.25.3/user_guide/reshaping.html

© Edunet Foundation. All rights reserved.

How to Build a Pivot Table in Python

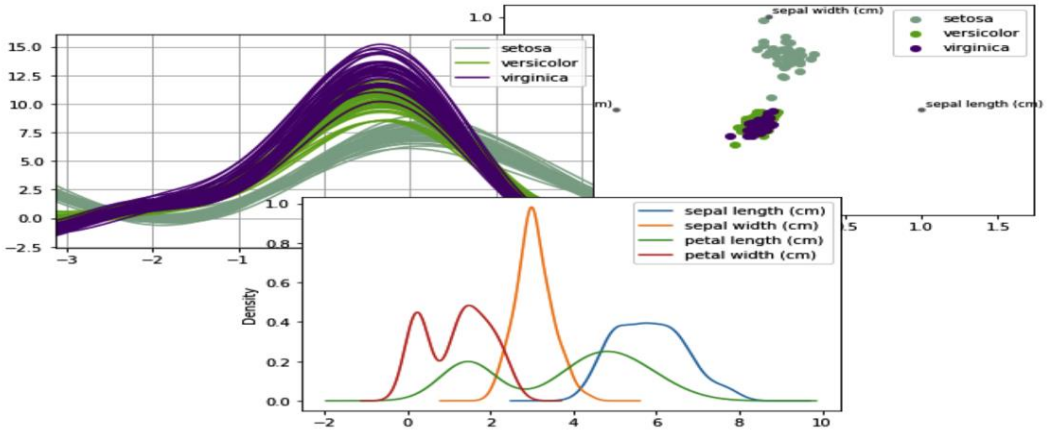
In Pandas, we can construct a pivot table using the following syntax:

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

© Edunet Foundation. All rights reserved.

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

© Edunet Foundation. All rights reserved.

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

© Edunet Foundation. All rights reserved.

Plot a Scatter Diagram using Pandas

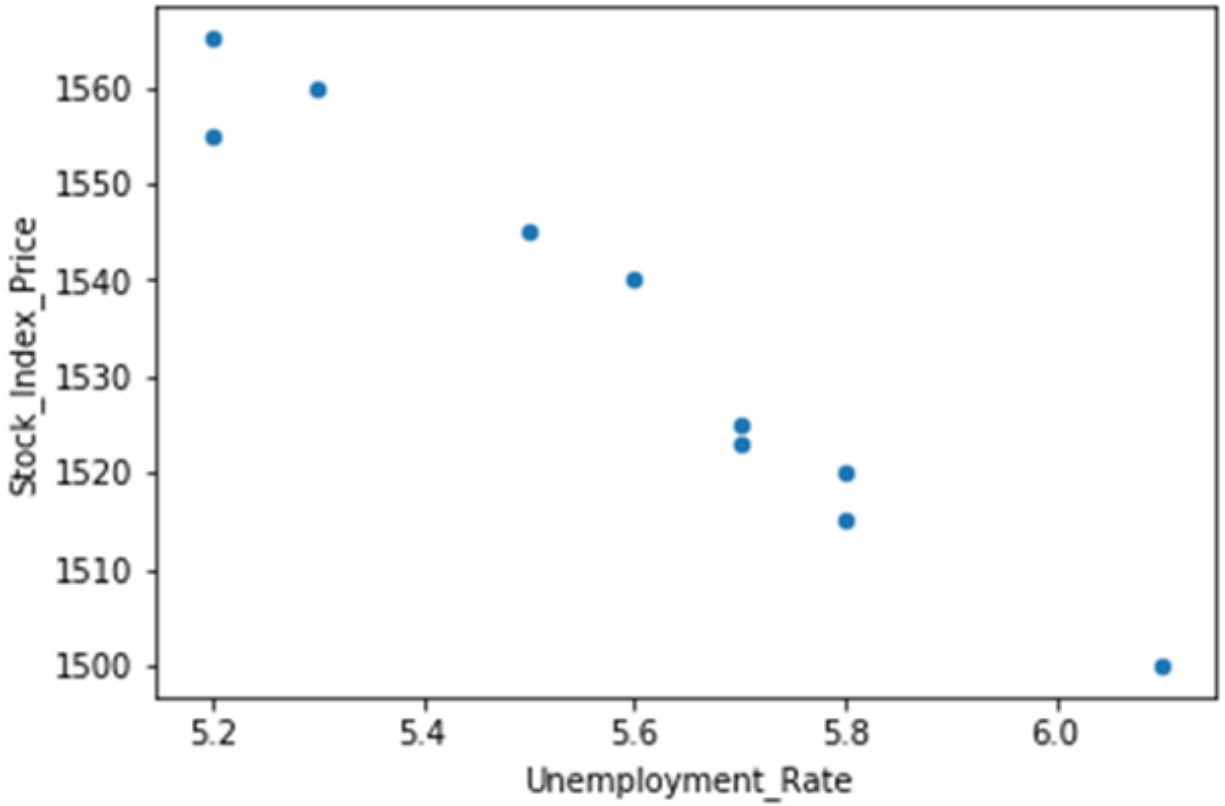
Step 2: Create the DataFrame

	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

© Edunet Foundation. All rights reserved.

Plot a Scatter Diagram using Pandas

Step 3: Plot the DataFrame using Pandas



© Edunet Foundation. All rights reserved.

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

© Edunet Foundation. All rights reserved.

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

© Edunet Foundation. All rights reserved.

Plot a Line Chart using Pandas

Step 3: Plot the DataFrame using Pandas

© Edunet Foundation. All rights reserved.

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

© Edunet Foundation. All rights reserved.

Plot a Bar Chart using Pandas

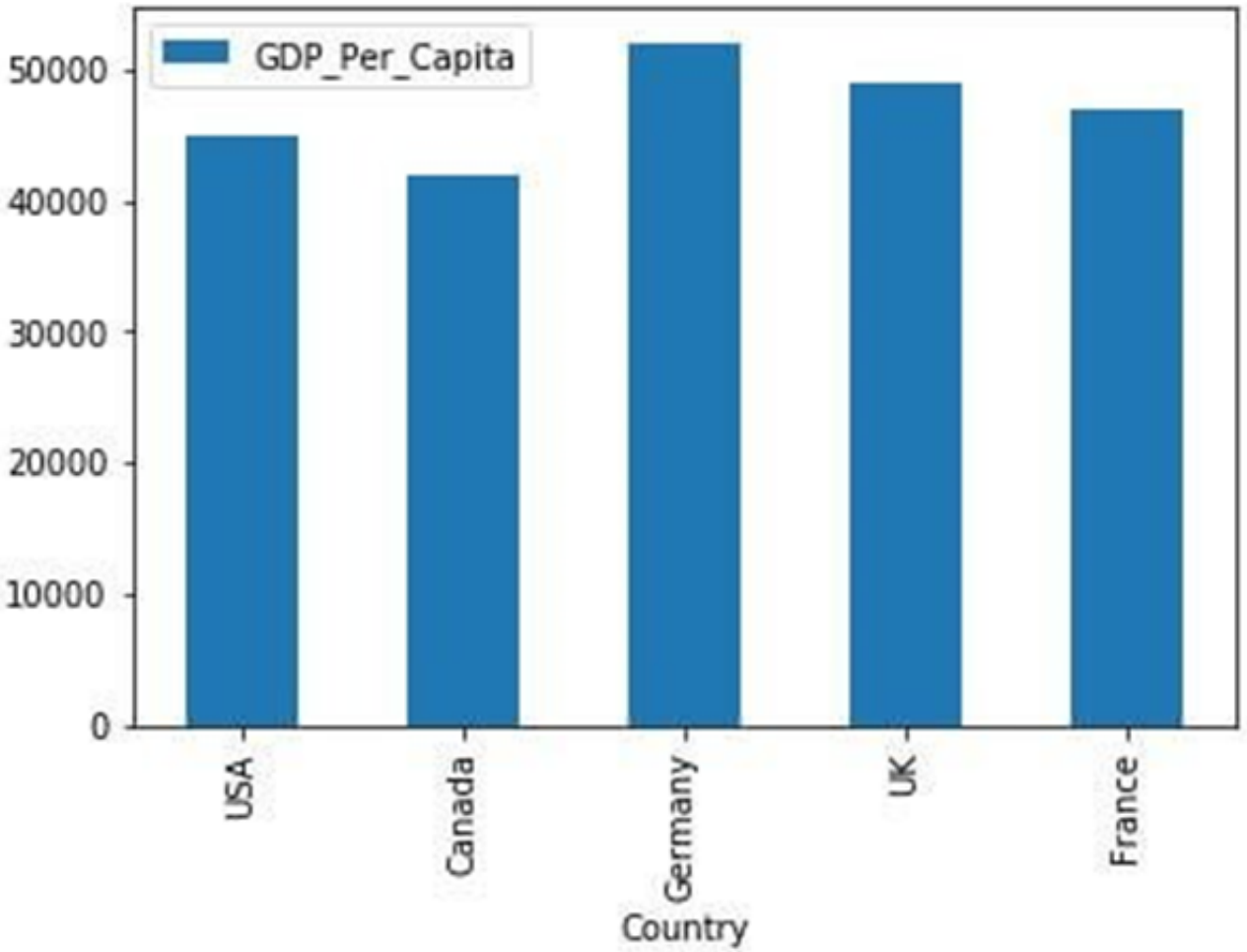
Step 2: Create the DataFrame

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

© Edunet Foundation. All rights reserved.

Plot a Bar Chart using Pandas

Step 3: Plot the DataFrame using Pandas



© Edunet Foundation. All rights reserved.

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

© Edunet Foundation. All rights reserved.

Plot a Pie Chart using Pandas

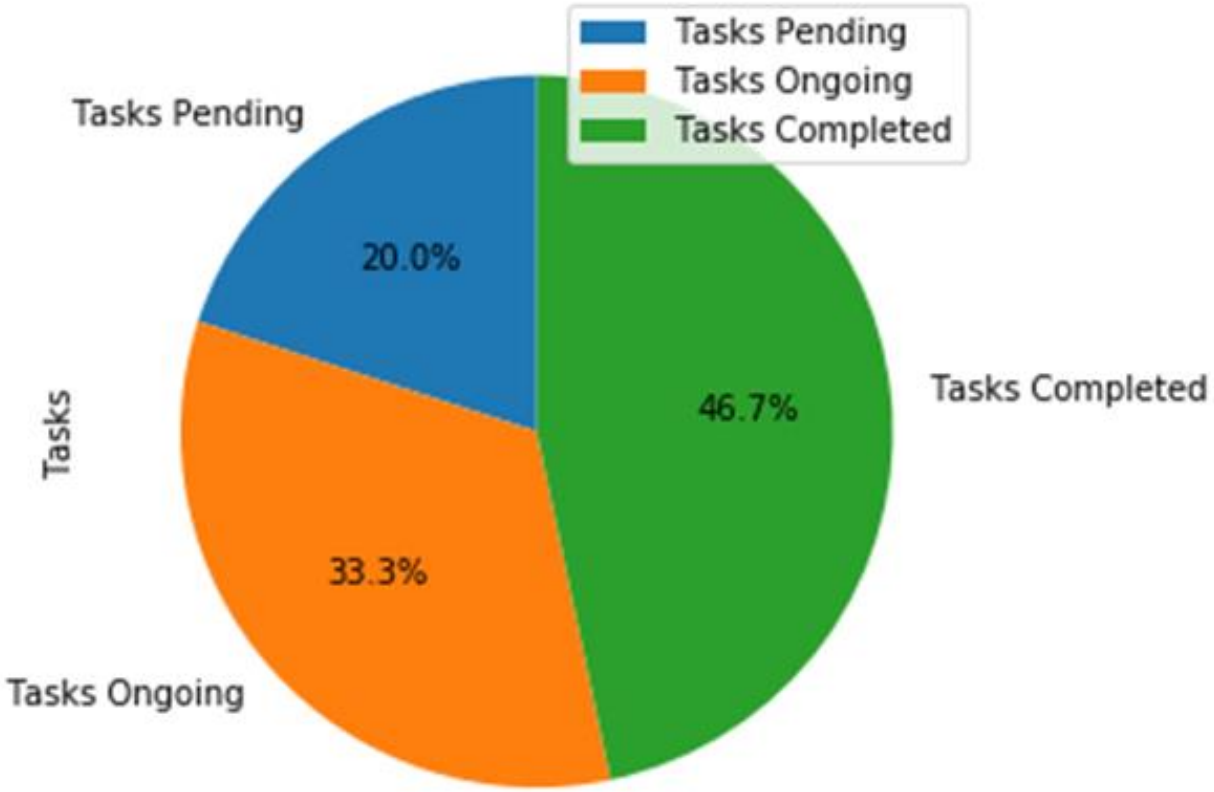
Step 2: Create the DataFrame

	Tasks
Tasks Pending	300
Tasks Ongoing	500
Tasks Completed	700

© Edunet Foundation. All rights reserved.

Plot a Pie Chart using Pandas

Step 3: Plot the DataFrame using Pandas



© Edunet Foundation. All rights reserved.

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/

© Edunet Foundation. All rights reserved.

REFERENCES

6. https://datatofish.com/load-json-pandas-dataframe
7. https://www.shanelynn.ie/summarising-aggregation-and-grouping-data-in-python-pandas
8. https://towardsdatascience.com/pandas-groupby-explained-453692519d0
9. https://towardsdatascience.com/a-step-by-step-guide-to-pandas-pivot-tables-e0641d0c6c70
10. https://pandas.pydata.org/docs/reference/api/pandas.pivot_table.html
11. https://datatofish.com/plot-dataframe-pandas

© Edunet Foundation. All rights reserved.

THANK YOU