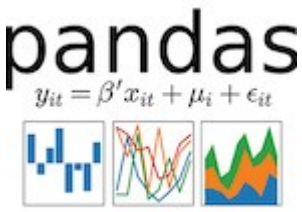
Importing Data in CSV Files with Pandas

This lesson will focus on how to import data from CSV files in Python using Pandas - a library for handling datasets in Python.

WE'LL COVER THE FOLLOWING

Pandas

Pandas is a very popular Python library that is used for data analysis, which is the second step of the Data Science lifecycle. It offers functionalities to handle and manipulate data very efficiently. It has been widely adopted by people who are not computer scientists or programmers as it makes them move beyond Excel for analyzing data. We will be using Pandas in this course for all of our analyses.



We will start by learning how to *import* CSV data using Pandas.

Importing CSV files

These days almost all of the data that is acquired by companies is entered in spreadsheets using different software such as Excel. The data is in the form of tables. This data is stored in CSV (Comma Separated Values) files that have **".csv"** at the end of their name. A spreadsheet is shown in the below image.

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms
1	-122.23	37.88	41.0	880.0	129.0
2	-122.22	37.86	21.0	7099.0	1106.0
3	-122.24	37.85	52.0	1467.0	190.0
4	-122.25	37.85	52.0	1274.0	235.0
5	-122.25	37.85	52.0	1627.0	280.0
6	-122.25	37.85	52.0	919.0	213.0

The above spreadsheet can be represented in a CSV file as:

```
longitude,latitude,housing_median_age,total_rooms,total_bedrooms
-122.23,37.88,41.0,880.0,129.0
-122.22,37.86,21.0,7099.0,1106.0
-122.24,37.85,52.0,1467.0,190.0
-122.25,37.85,52.0,1274.0,235.0
-122.25,37.85,52.0,1627.0,280.0
-122.25,37.85,52.0,919.0,213.0
```

To work with CSV files in Python, we have to import them first. Below is an example of how to import CSV files. The file *housing.csv* is the Census data of housing blocks in California. You can download it to view the contents by clicking on it. Run the code below.

housing.csv

```
# import pandas library to use it
import pandas as pd

# read the file
df = pd.read_csv('housing.csv')

# Shape of the dataframe
print('Shape of dataframe',df.shape)

# information about the data set
print('\n',df.info())

# print the first 5 rows
print('\n\n',df.head())
```

To use Pandas, we import it in the 1st line. Then we read the file using the function, `read_csv`. You only have to provide the name of the file to this function as done above in **line 5**. Pandas stores the file in an object called **dataframe** that stores the data in the form of rows and columns like a 2D array. We name our dataframe `df`. To find the number of rows and columns, we print `df.shape` which gives us the number of rows and columns in the format `(rows, columns)`.

To retrieve general information about the columns of a dataset, we use the function `info()` in **line 11**. The `info` function gives the name, the number of values, and the data type of each column.

To view the first 5 rows, we have written `print(df.head())` in **line 14**. The function `head()` returns the first 5 rows of the dataframe and `print` displays them on the screen.

> Sometimes because of the long length of the table, it cannot be completely displayed on the screen and it breaks. Some columns are shown together and then the rest of the columns are shown below, as you can see from the output of **line 14** in the above code.

In the next lesson, we will look at how to pick and choose data from a dataframe.