Machine Learning

Lab 1

1. Get started

We will use Python for all lab exercises and coursework in this module. Please feel free to use your preferred Python IDEs on your personal laptop, e.g., Jupyter Notebook, Pycharm, etc.

If you do not have a preference or a laptop, you can use Jupyter Notebook which is installed on every PC in the lab. You can access it by opening Anaconda software.

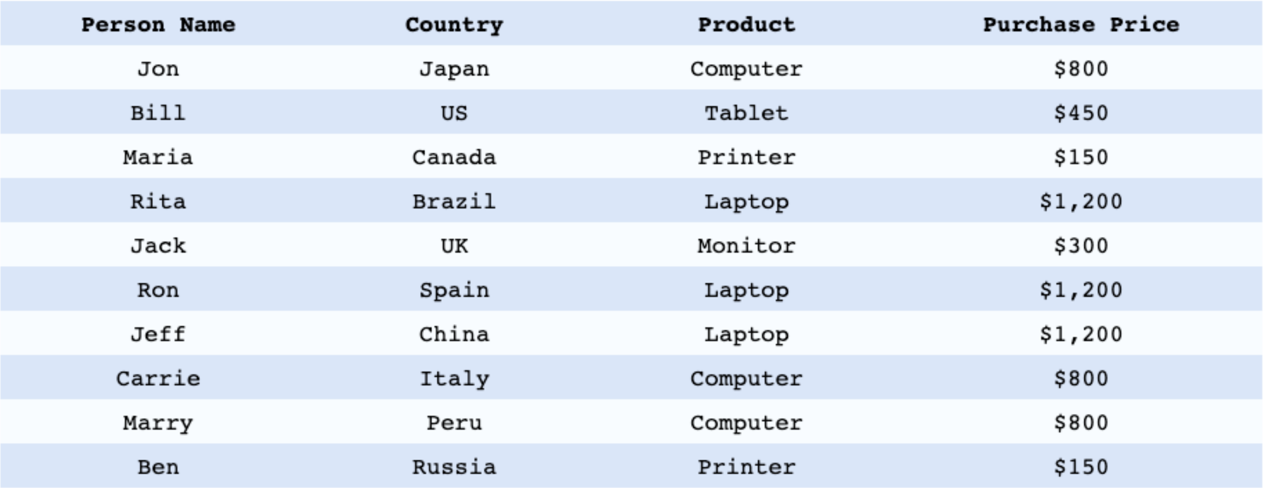
If you are not familiar with Jupyter Notebook or Anaconda, please read the Jupyter Notebook Tutorial and follow the instructions to get started. You can find it next to this lab tutorial on Moodle.

Data is the most important part of all machine learning. Without data, we can’t train any model and all modern research and automation will go in vain. Big enterprises are spending lots of money just to gather as much certain data as possible. In this lab, we will practice how to find a dataset, import it into Python and perform some simple analysis.

2. Import the data into Python

We will practice how to import a dataset into Python. A sample CSV dataset is provided, and you can find it next to this lab tutorial on Moodle. Now, download the dataset and try to follow the instructions below and import it into Python.

So let’s begin with this simple example, where you have the following client list and some additional sales information stored in a CSV file (where the file name is ‘**Clients**‘):



Person Name	Country	Product	Purchase Price
Jon	Japan	Computer	$800
Bill	US	Tablet	$450
Maria	Canada	Printer	$150
Rita	Brazil	Laptop	$1,200
Jack	UK	Monitor	$300
Ron	Spain	Laptop	$1,200
Jeff	China	Laptop	$1,200
Carrie	Italy	Computer	$800
Marry	Peru	Computer	$800
Ben	Russia	Printer	$150

Step (1): Capture the File Path

Firstly, capture the full path where your CSV file is stored.

For example, let’s suppose that a CSV file is stored under the following path:

C:\Users\Ron\Desktop\Clients.csv

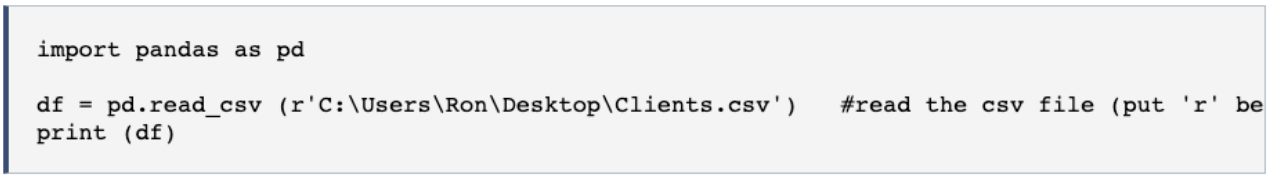
You’ll need to modify the Python code below to reflect the path where the CSV file is stored on *your* computer. Don’t forget to include the:

- File name (as highlighted in green). You may choose a different file name, but make sure that the file name specified in the code matches with the actual file name
- File extension (as highlighted in blue). The file extension should always be ‘.csv’ when importing CSV files

Step (2): Apply the Python code

Type/copy the following code into Python, while making the necessary changes to your path.

Here is the code for our example.

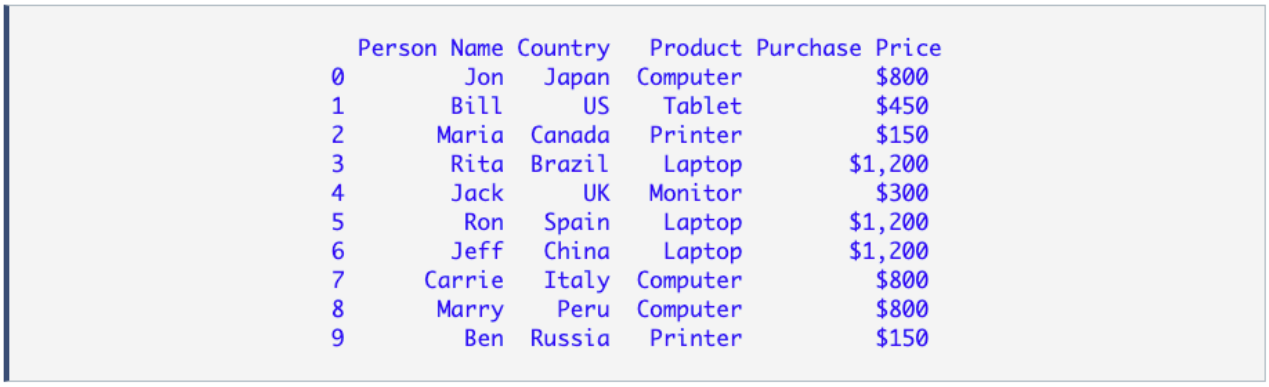


```
import pandas as pd

df = pd.read_csv (r'C:\Users\Ron\Desktop\Clients.csv')   #read the csv file (put 'r' be
print (df)
```

Step (3): Run the Code

Finally, run the Python code and you’ll get:

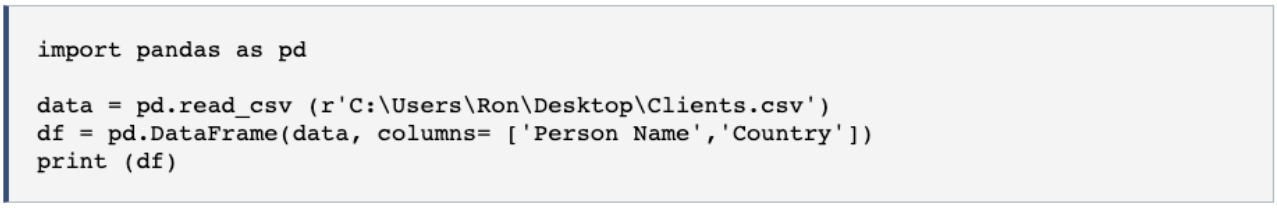


```
  Person Name Country   Product Purchase Price
0         Jon   Japan  Computer           $800
1        Bill      US    Tablet           $450
2       Maria  Canada   Printer           $150
3        Rita  Brazil    Laptop         $1,200
4        Jack      UK   Monitor           $300
5         Ron   Spain    Laptop         $1,200
6        Jeff   China    Laptop         $1,200
7      Carrie   Italy  Computer           $800
8       Marry    Peru  Computer           $800
9         Ben  Russia   Printer           $150
```

Step (4): Select Subset of Columns

Now what if you want to select a subset of columns from the CSV file?

For example, what if you want to select only the *Person Name* and *Country* columns. If that's the case, you can specify those columns names as captured below:

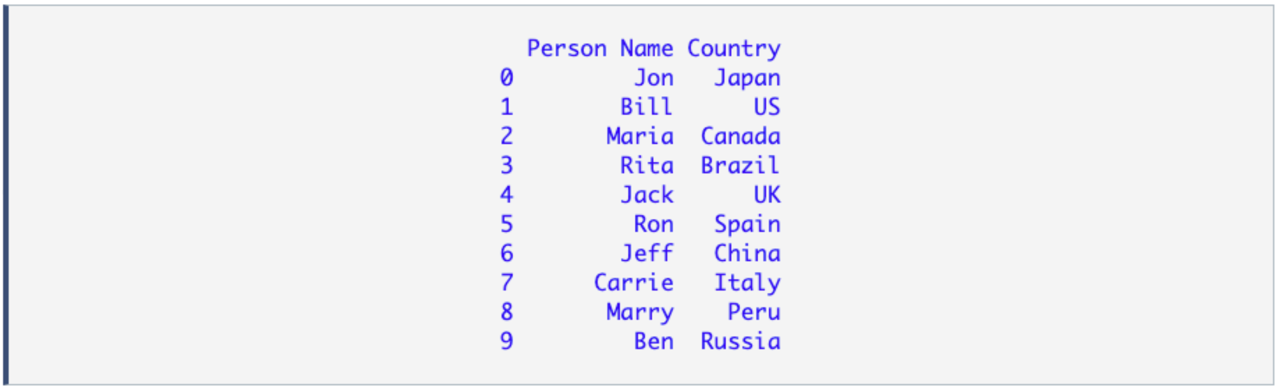


```
import pandas as pd

data = pd.read_csv (r'C:\Users\Ron\Desktop\Clients.csv')
df = pd.DataFrame(data, columns= ['Person Name','Country'])
print (df)
```

You'll need to make sure that the column names specified in the code exactly match with the column names within the CSV file. Otherwise, you'll get NaN value.

Once you're ready, run the code (after adjusting the file path), and you would get only the Person Name and Country columns:



```
  Person Name Country
0         Jon   Japan
1        Bill      US
2       Maria  Canada
3        Rita  Brazil
4        Jack      UK
5         Ron   Spain
6        Jeff   China
7      Carrie   Italy
8       Marry    Peru
9         Ben  Russia
```

3. Find a public dataset and import it into Python

From the online videos, you may now have a brief understanding of machine learning and have found a machine learning application or technique that you are interested. Now, try to find a relevant dataset of that application or technique.

You can explore any public data websites. If you do not know where to start, here are two examples:

- UCI Machine Learning Repository
  https://archive.ics.uci.edu/ml/index.php

- Kaggle dataset
  https://www.kaggle.com/datasets

Download the dataset and try to import it into Python. You could follow the above instructions to import the data if it is stored in CSV file. Or if you would like some challenges, find a dataset stored in other file formats and try to use the right Python packages and functions to import it!