BTS | Barcelona Technology School

eurecat Centre Tecnològic de Catalunya

Data Science Foundations

Session #1

The following are the tasks that should complete and synchronize with your repository "DataScienceFoundations" **until October 5**. Please notice that none of these tasks is graded, however it's important that you correctly understand and complete them in order to be sure that you won't have problems with further assignments.

Task 1 – Algorithmic thinking

During the class of today (October 2) you solved the following problem:

- Imagine that you acquired the historical data of average traffic intensity per weekdays for the years 2013-2017 for some city.

```
traffic_data = {
  "2013" : [2800,2900,2950,2890,3050,3030,3000],
  "2014" : [2700,2915,2950,2880,3020,3030,3020],
        "2015" : [2750,2910,2955,2820,3030,3080,3010],
        "2016" : [2920,2950,2960,2840,3100,3100,3050],
        "2017" : [2930,2940,2970,2900,3150,3300,3080]
}
```

- Your task is to develop the code that selects the data sub-set of 2016 and finds the minimum value of a traffic intensity. If this value is greater than 2000, then you should "store" the result on the hard drive. Otherwise, you should output a message saying that the data is not saved.

You should push your solution code (that one that you did during the class) into your repository called "DataScienceFoundations".

Task 2 – Algorithmic thinking

During the class you might have insufficient time to correctly finish Task 1. Please finish it at home. Do not forget to add comments to your functions (docstrings).

- Clone the Git repository to get an initial code:

https://github.com/LianaNapalkova/BTS_MasterInBigData.git

- Once you downloaded the repository to your local file system, go to the folder “BTS_MasterInBigData/Session_1

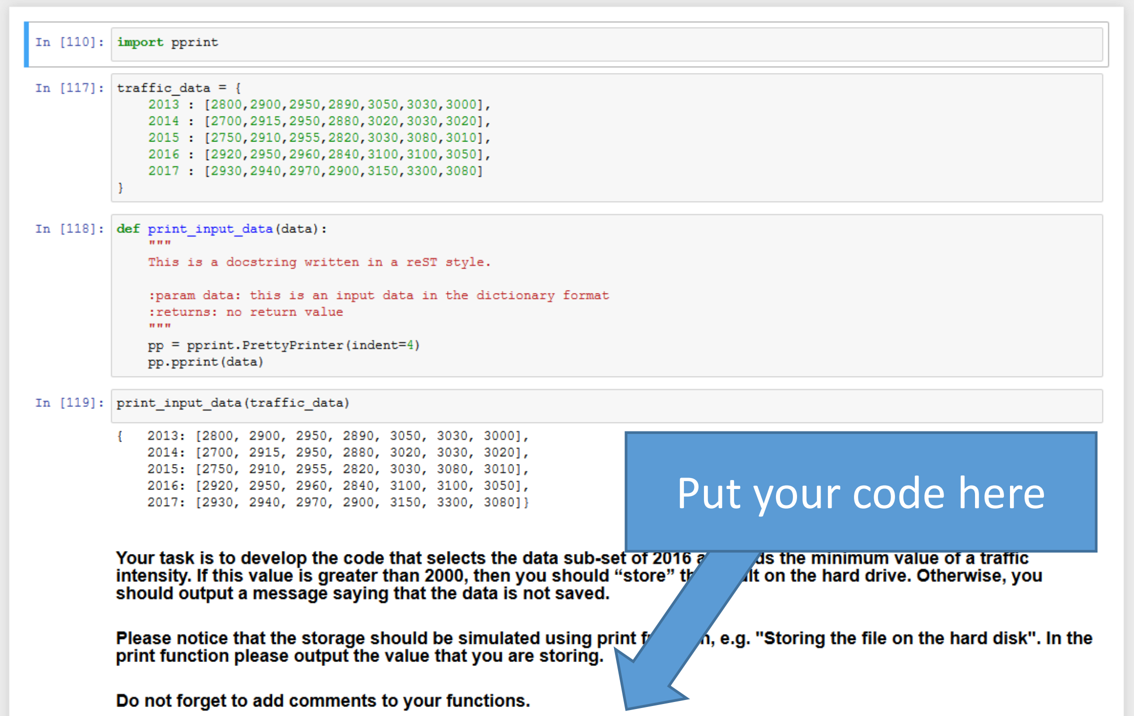
- Copy the folder “Session_1” into your local folder “DataScienceFoundations”.

- In the folder “Session_1” you will see a file called:

  bts_data_science_foundations_session_1_1.ipynb

- Import this file into Jupyter Notebook using the “Upload” button.

- Open the imported script and put your final solution of the Task 1.



```
In [110]: import pprint
```

```
In [117]: traffic_data = {
              2013 : [2800,2900,2950,2890,3050,3030,3000],
              2014 : [2700,2915,2950,2880,3020,3030,3020],
              2015 : [2750,2910,2955,2820,3030,3080,3010],
              2016 : [2920,2950,2960,2840,3100,3100,3050],
              2017 : [2930,2940,2970,2900,3150,3300,3080]
          }
```

```
In [118]: def print_input_data(data):
              """
              This is a docstring written in a reST style.

              :param data: this is an input data in the dictionary format
              :returns: no return value
              """
              pp = pprint.PrettyPrinter(indent=4)
              pp.pprint(data)
```

```
In [119]: print_input_data(traffic_data)

          {   2013: [2800, 2900, 2950, 2890, 3050, 3030, 3000],
              2014: [2700, 2915, 2950, 2880, 3020, 3030, 3020],
              2015: [2750, 2910, 2955, 2820, 3030, 3080, 3010],
              2016: [2920, 2950, 2960, 2840, 3100, 3100, 3050],
              2017: [2930, 2940, 2970, 2900, 3150, 3300, 3080]}
```

Put your code here

Your task is to develop the code that selects the data sub-set of 2016 a ds the minimum value of a traffic intensity. If this value is greater than 2000, then you should “store” th lt on the hard drive. Otherwise, you should output a message saying that the data is not saved.

Please notice that the storage should be simulated using print f n, e.g. "Storing the file on the hard disk". In the print function please output the value that you are storing.

Do not forget to add comments to your functions.

You should push your final solution code into your repository called “DataScienceFoundations”.

Task 3 – Practicing data handling skills with Pandas

- Create a new script in Jupyter Notebook named “data_handling”.

- Load the data from the file “rural_population_data.csv” into pandas DataFrame.

- Do the following tasks with the data:

  1. Find all unique values of the column “Continent”

  2. Calculate a grow rate based on this formula:

     Growrate = (Population2100 - Population2000) / Population2000

3. Create a histogram of “Growrate”.
4. Check the existence of NaN values in the DataFrame.
5. Find the average population size in the years 2000 and 2010.
6. Check how many rows have UrbanRuralDesignation equal to “Urban”.

You should push your solution code into your repository called “DataScienceFoundations”.

STARTING FROM NOW PLEASE ALWAYS PUT ALL YOUR ASSIGNMENTS OF THE MODULE “Data Science Foundations” INTO YOUR PERSONAL REPOSITORY “DataScienceFoundations”. IT MEANS THAT YOU SHOULD ALSO HAVE A FOLDER NAMED AS “DataScienceFoundations” ON YOUR LAPTOP.