SMDA 2018/19 – Lecture L3 - 10/10/2018

Exercise 1: Telco Customer Churn first data analysis using Python (Part 1)
Please, execute the following tasks and provide answers to the proposed questions.

1. Open the Telco Customer Churn dataset page in Kaggle.
- Hint: https://www.kaggle.com/blastchar/telco-customer-churn
- Have a look to the "Overview" tab to understand something more about the dataset

2. Check the main properties of this dataset in the "Data" tab.
- How many samples (rows) does it have?
- How many variables (columns)?
- What does each row/column represent?
- Which is the "target" column? What does it represent?

3. Download the dataset into your computer.
- Which is the extension of the downloaded file?

4. Uncompress the file
- Which is the extension of the uncompressed file?

5. Open the uncompressed file by both a text editor and a spreadsheet software
- Which symbol is used to separate columns?
- Which symbol is used to separate rows?
- Which values can you find for variable SeniorCitizen? And for variable Partner?

6. Generate a new notebook for analyzing this dataset
- Hint: click on "New Kernel", then choose the Notebook kernel type, on the right
- Assign the following title to the notebook: SMDA_L3_ExPython_TelcoCustomerChurn_YourSurname
- Then click on the "Commit" button on top-right to make the notebook ready to be started

7. Open the notebook documentation page to get help if needed
- Hint: click the "Docs" link on the right-bottom of your notebook page

8. Select the first cell (we will call it "Library import cell" in the following), run it
- What is the output of this action?
- What does the code "*import numpy as np*" do? Can you provide a reference website for this library?
- What does the code "*import pandas as pd*" do? Can you provide a reference website for this library?
- What does the code "*import os*" do? Can you provide a reference website for this library?
- How many data files are available? Please provide their names.

9. Add to the first cell new lines to load the following libraries: seaborn, matplotlib.pyplot, sklearn.linear_model (only LogisticRegression)
- Hint: find similar code in the Titanic notebook if needed

10. Select the first cell and add a new cell on top of it
- Hint: use the button on top-right of the cell

11. Select the new cell and transform it in a "Markdown" cell, then copy all the text in this pdf file and paste it in the new Markdown cell

12. Please write your answers to the questions above in the new Markdown cell. From now on you can use the same cell to write your answers as well

13. Select the "Library input cell" and add a new cell below it

14. Use the new cell to load the Telco Customer Churn dataset into a Pandas DataFrame variable called *data*

- Hint: find similar code in the Titanic notebook if needed
- Remind to run the cell after writing the code-box

15. Add the following comment before data loading line: "Data acquisition"

16. Add also a Markdown cell before the data loading cell and write in bold the text "Data acquisition"

- Markdown cells should be used to give a structure to the report, hence they should be added before each new section

17. In a new cell show the number of rows, the number of columns, and the total number of cells in the dataset

- Hint: display the related *parameters* of the Pandas DataFrame
- Hint: use the *print* function to print the results
- You should print, in particular, the following strings:
  - "The number of customers is XXXX"
  - "The number of variables is YYYY"
  - "The total number of cells is ZZZZ"
- Other hints:
  - How can you select a single element from the shape tuple?
  - How can you convert a number to string?
  - How can you concatenate two strings?
  - How can you print the final string?

18. Add the following comment at the beginning of the cell: "Dataset dimension"

19. Add a new markdown cell before this cell and write in it the title "Data Analysis"

20. In a new cell show the names of the variables in the dataset

- Hint: print the *column*'s names of variable *data*

21. In a new cell show the *first* and *last* 10 rows in the dataset

- Hint: find the correct DataFrame *methods* in the Pandas' documentation

22. In a new cell show i) the type of variable *data,* ii) the number of missing values for each variable, iii) the type of each variable, iv) the total memory used to store variable data

- Hint: all this information can be provided by a single method of DataFrame
- How many missing values are there in total?
- Which variables are categorical?
- Which variables are numerical?

23. In a new cell show the following basic statistics for all numerical variables: number of non-missing values, mean, standard deviation, minimum, maximum, median, 1st and 3rd quartiles

- Hint: all this information can be provided by a single method of DataFrame

24. In a new cell show the following basic information for all categorical variables: number of non-missing values, number of unique values, most frequent value and frequency of the most frequent value.

- Hint: all this information can be provided by the DataFrame method used in question 22, using specific arguments
- Can you see any strange value in this result?

25. In a new cell show the histograms of each numeric variable (i.e., column) in the dataset

- Hint: try to find a specific method in the DataFrame API documentation