SMDA 2018/19 – Lecture L4 - 12/10/2018

Exercise 1: Telco Customer Churn first data analysis using Python (Part 2)
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

...after having developed points 1 to 23...

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

26. In a new cell show the box-plots of each numeric variable (i.e., column) in the dataset

- Hint: try to find a specific method in the DataFrame API documentation
- Does this chart provide a good visualization? Why?
- Try to generate one box-plot for each numerical variable
- Try to put all three charts in the same figure using the subplot function

27. In a new cell show the histograms of the categorical variables in the dataset

- Hint: try to use a function from the Seaborn library which counts the number of time each element appears and makes a related bar plot
- Hint: use the subplot function to put all the charts in the same figure
- Hint: resize the figure so that to avoid overlapping and enable a clear visualization of all charts

28. In a new cell generate a new DataFrame called data1 and containing only variables gender, Partner, MonthlyCharges, Churn

- Hint: you could try also other selections

29. In a new cell show the first 5 rows of the new dataset

30. Convert categorical values in data1 to numeric as follows:

- **gender: Male=0, Female=1**
- **Partner: No=0, Yes=1**
- **Churn: No=0, Yes=1**

- Hint: find similar code in the Titanic notebook if needed

31. Generate a separate Series variable called data1Churn for the dependent (churn) variable and drop it from DataFrame data1

- Hint: Series is a data structure defined in Pandas, try to find its documentation page
- Hint: each column of a DataFrame is a Series
- Hint: learn how to drop columns from a dataset in the Titanoc notebook
- What is the difference between data1[[‘Churn’]] and data1[‘Churn’]?
- When single square brackets are used with Pandas DataFrame? When double brackets are used instead?

32. Generate a linear logistic model using data1 as dependent variables and data1Churn as independent variable, then show the model “score”

- Hint: try to find a function for linear logistic model learning in the sklearn library
- Hint: find similar code in the Titanic notebook if needed

33. Show the parameters of the linear logistic model computed above. Which variable seems to be more related to customer churn?

- Hint: find similar code in the Titanic notebook if needed

34. If you want, click on the *Sharing* field on the right hand side menu and share the notebook with me (Kaggle user: albertocastellini)

- No score/evaluation will be given, don’t worry :-)