Data Exploratory Analysis

Learning Outcomes

- Upon successful completion of this lab, you will have demonstrated the abilities to:
  - Generate different data exploratory visualizations
  - Analyze and summarize your findings
  - Know how to handle missing values in a data set

Instructions:

1. Read the tutorial here (http://www.cse.msu.edu/~ptan/dmbook/tutorials/tutorial3/tutorial3.html)
2. Download the following data set from the UCI Machine Learning Repository:
   - https://archive.ics.uci.edu/ml/datasets/Adult (https://archive.ics.uci.edu/ml/datasets/Adult) (adult.data)
   - As the dataset contains missing values (" ?"), you need to use an option in *read_csv* to convert " ?", which is a non-standard missing value representation in python, to standard (**nan**) when you are loading data:

```
missing_values = [" ?"]
data = pd.read_csv('http://archive.ics.uci.edu/ml/machine-learning-database
s/adult/adult.data',header=None, na_values = missing_values)
```

Part I:

1. Apply as many of the different visualization techniques described in the tutorial as possible:
2. Your report should contain (minimum requirement):
   1. For each continuous attribute, calculate its average, standard deviation, minimum, and maximum values.
   2. For the discrete attribute, count the frequency for each of its distinct values.
   3. Draw histogram of the class variable
   4. Draw the distribution of values for a continuous attribute using a histogram.
   5. Draw some scatter plots for a couple of attribute pairs.
   6. Draw a parallel diagram for some attributes in the data set
   7. For each diagram describe your interpretation and insight.

Part II:

1. Identify which attributes have missing values and address the issue by:
   1. Replacing missing values by the *average* or *mod* of the attribute (based on attribute types)

2. Replace missing values by the *average* or *mode* of the attribute in the particular class to which the instance belongs
3. Draw a histogram of the attribute before and after replacing missing values in the previous step 1 and 2

Report:

1. Your report should have a cover letter including the group member names
2. Organize all your *diagrams* and *interpretations* in your lab report *(PDF format)*
3. Include your code and report in a folder (you can zip the folder) and submit it

Resources:

https://medium.com/@roshankg96/handling-missing-data-in-pandas-a3c8dfbd1db **(https://medium.com/@roshankg96/handling-missing-data-in-pandas-a3c8dfbd1db)**

https://towardsdatascience.com/data-cleaning-with-python-and-pandas-detecting-missing-values-3e9c6ebcf78b **(https://towardsdatascience.com/data-cleaning-with-python-and-pandas-detecting-missing-values-3e9c6ebcf78b)**