CPS 844 Lab 1: Data Exploration

Data exploration and statistics enable to understand various data characteristics. In turn, this may help to select appropriate preprocessing and analysis techniques. This assignment makes use of the Iris sample data of three Iris species: Setosa, Versicolour, and Virginica. Each iris flower is described by five attributes:

- sepal length (in cm)
- sepal width (in cm)
- petal length (in cm)
- petal width (in cm)
- class (one of the tree species)

In this lab, you will get familiar with Python, you will learn how to load a CSV file into a Pandas DataFrame object, and compute various statistics from the DataFrame.

To do:

1) Obtain an Integrated Development Environment (IDE) for Python 3. It is recommended you use Spyder, which you can download for free. The best is to get the Anaconda python distribution, as Spyder is included by default in it. https://www.anaconda.com/products/individual
2) Look for Spyder, and open it.
3) You will need to use the pandas library for the data manipulation and analysis. To first being able to use it, you need to install it. Type into your Spyder console: pip install pandas.
4) Download and check the format of the data:
   http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data
   as well as the description of the data:
   http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.names
5) Using one of the functions from the pandas library, load the 'iris.data' csv file into a DataFrame. The size of your DataFrame should consist of 150 records of 5 attributes.
6) You should have noticed that the data did not have headers. Assign new headers to the DataFrame. The headers should be relevant to the attribute description in the file 'iris.names'.
7) For each quantitative attribute, calculate its average, standard deviation, minimum, and maximum values. Use the methods associated with the DataFrame; e.g. the average can be computed using the method described here:
   https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.mean.html
8) Count and print the frequency for each of its distinct class values.