Introduction

In this lab, we get some initial experience with using some of the main python tools for this course, including Numpy, Matplotlib and Pandas. We also load some datasets, compute some basic statistics on them and plot them.

The lab assumes that you are familiar with Python. Please complete `Week 01 lab: Introduction to python` before attempting this lab.

The lab can be executed on either your own machine (with anaconda installation) or lab computer.

- Please refer canvas for instructions on installing anaconda python and Jupyter Notebook.

Objective

- Continue to familiarise with Python and Jupyter Notebook
- Load dataset and examine the dataset
- Learn to compute basic statistics to understand the dataset more
- Plot the datasets to visually investigate the dataset

Dataset

We examine two regression based datasets in this lab. The first one is to do with house prices, some factors associated with the prices and trying to predict house prices. The second dataset is predicting the amount of share bikes hired every day in Washington D.C., USA, based on time of the year, day of the week and weather factors. These datasets are available in 'housing.data.csv' and 'bikeShareDay.csv' in the code repository.

First, ensure the two data files are located within the Jupyter workspace.

- If you are on the local machine copy the two data data directories ('BostonHousingPrice','Bike-Sharing-Dataset') to your current folder.

Load dataset to Python Notebook

Next we examine how to load these into Python and Jupyter notebooks. We will first analyse the House prices dataset, then you'll repeat the process to analyse the bike hire dataset.

First we need to import a few packages that will be used for our data loading and analysis. In python notebook you can load packages just before it is called (no need to load them at the start of the program).

Pandas is a great Python package for loading data. We will use Matplotlib to visualise some of the distributions. Numpy is a numeric library that has many useful matrices and mathematical functionality.

```
1 import pandas as pd
```

```
1 import matplotlib.pyplot as plt
```

```
1 import numpy as np
```

```
1 housedata = pd.read_csv(r"Lab\housing.data.csv", delimiter="\s+")
```

Replace the filename with the relative or absolute path to your files. We strongly encourage you to look up the documentation of the functions we use in the lab, in this case examine Pandas read_csv documentation.

The `read_csv()` command loads the input file, which is a csv formatted file delimited by tabs, into a **Pandas dataframe** (which can be thought of as a table). A dataframe can store the column names as well as the data. Examine what has been loaded into the dataframe `housedata`.

```
1 print(housedata)
```

If you are interested in checking only the first few rows of the dataframe to see if you have read the data in correctly, you can use the head method in dataframe.

```
1 housedata.head()
```

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296.0	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242.0	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222.0	18.7	396.90	5.33	36.2

Now we have loaded the data into a data frame and printed it out, next we will compute some very basic statistics. The abbreviated column names:

- CRIM: per capita crime rate by town
- ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS: proportion of non-retail business acres per town
- CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- NOX: nitric oxides concentration (parts per 10 million)
- RM: average number of rooms per dwelling
- AGE: proportion of owner-occupied units built prior to 1940
- DIS: weighted distances to five Boston employment centres
- RAD: index of accessibility to radial highways
- TAX: full-value property-tax rate per USD10,000
- PTRATIO: pupil-teacher ratio by town
- B: 1000 (Bk - 0.63)^2 where Bk is the proportion of blacks by town
- LSTAT: lower status of the population
- MEDV: Median value of owner-occupied homes in USD1000's

The target column is **MEDV** and all the other columns are attributes.

Study the variables carefully and understand what they represent before moving to the next section.

Exploratory Data Analysis (EDA)

Often the first step in developing a machine learning solution for a given dataset is the EDA. EDA refers to the critical process of performing initial investigations on data so as to:

- Maximize insight into a data set;
- Uncover underlying structure;
- Extract important variables;
- Detect outliers and anomalies;
- Test underlying assumptions;
- Develop parsimonious models; and
- Determine optimal factor settings.

with the help of summary statistics and graphical representations. The particular graphical techniques employed in EDA are often quite simple, consisting of various techniques of:

- Plotting the raw data (such as data traces, histograms, bi-histograms, probability plots, lag plots, block plots, and Youden plots.
- Plotting simple statistics such as mean plots, standard deviation plots, box plots, and main effects plots of the raw data.
- Positioning such plots so as to maximize our natural pattern-recognition abilities, such as using multiple plots per page.

[ ] ↳ *33 cells hidden*

Exercise: Analyse the Bike Share Data

[ ] ↳ *17 cells hidden*