# Exercise 1 : Data Acquisition

## Setup

1. Create a folder on your Desktop and name it MA0218_[LabGroup], where [LabGroup] is the name of your Group
2. Download the .ipynb files and data files posted corresponding to this exercise and store in the aforesaid folder
3. Open Jupyter Notebook (already installed on the Lab computer) and navigate to the aforesaid folder on Desktop
4. Open and explore the .ipynb files (notebooks) that you downloaded, and go through "Preparation", as follows
5. The walk-through videos posted on NTU Learn (under Course Content) may help you with this "Preparation" too

6. Create a new Jupyter Notebook, name it Exercise1_solution.ipynb, and save it in the same folder on the Desktop
7. Solve the "Problems" posted below by writing code, and corresponding comments, in Exercise1_solution.ipynb

Note : Don't forget to import the Essential Libraries required for solving the Exercise (check the preparation notebooks)

## Preparation

| | |
|---|---|
| M1 DataAcquisition.ipynb | Practice acquiring data in Jupyter notebook from various sources<br>You will need the data folder (posted as data.zip) to use this code |
| M2 BasicStatistics.ipynb | Check how to import the Pokemon data (Statistics not yet required)<br>You will need the CSV data file pokemonData.csv to use this code |

## Problems

### Problem 1

Download the dataset from the following Kaggle Competition (login required) – Go to "Data", and "Download All".

House Prices Competition : https://www.kaggle.com/c/house-prices-advanced-regression-techniques

a) Import the "train.csv" data from the downloaded data folder (has four files) in Jupyter Notebook.
b) How many observations (rows) and variables (columns) are in the above dataset? Check the "shape".
c) What are the data types ("dtypes") – Numeric/Categorical – of the variables (columns) in the dataset?
d) What does the .info() method do? Use the .info() method on the imported dataset to check this out.
e) What does the .describe() method do? Use the .describe() method on the imported dataset to check.

### Problem 2

Check Summer Olympic 2016 medal tally : https://en.wikipedia.org/wiki/2016_Summer_Olympics_medal_table

a) Import the Wikipedia page in Jupyter Notebook (check M1 DataAcquisition.ipynb for hints about this).
b) How many tables are in this Wikipedia page? Check the "len" of the imported page to find this out.
c) Which one is the actual "2016 Summer Olympics medal table"? Explore all tables in the data to know.
d) Store the main table, that is, "2016 Summer Olympics medal table", as a new Pandas DataFrame.
e) Extract the TOP 20 countries from the medal table, and store these rows as a new DataFrame.

### Important

Try to solve the problems on your own. Take help/hints from the "Preparation" codes and walk-through videos.

If you are still stuck, talk to your friends in the Lab to get help/hints. If that fails too, approach the Lab Instructor.