Exercise 1 : Data Acquisition

Workflow

1. Create a folder on your Desktop and name it Cx1015_[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

Try to solve the problems on your own. Take help and hints from the "Preparation" codes and the 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.

Note : Don't forget to import the Essential Python Libraries required for solving the Exercise. Write code in the usual "Code" cells, and notes/comments in "Markdown" cells of the Notebook. Check the preparation notebooks for guidance.

Preparation

M1 DataAcquisition.ipynb	Practice acquiring data in Jupyter notebook from various sources 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) You will need the CSV data file pokemonData.csv to use this code

Problems

Problem 1

Download the dataset **train.csv** posted with this Exercise. This dataset is collected from Kaggle. You may also want to download it directly from the following Kaggle Competition (Login > Go to "Data" > "Download All" > train.csv). Either way, read the competition description (no login required) to get an idea about what the target Data Science task is.

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

a) Import the "train.csv" data you downloaded (either from NTU Learn or Kaggle) 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 data/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) Extract the main table, "2016 Summer Olympics medal table", and store it as a new Pandas DataFrame.
e) Extract the TOP 20 countries from the medal table, as above, and store these rows as a new DataFrame.

Bonus Problems

A. Download the “Census Income” dataset (source : https://archive.ics.uci.edu/ml/datasets/Census+Income) from the UCI Machine Learning Repository (in the “Data Folder”), and import it in Jupyter Notebook as a DataFrame.

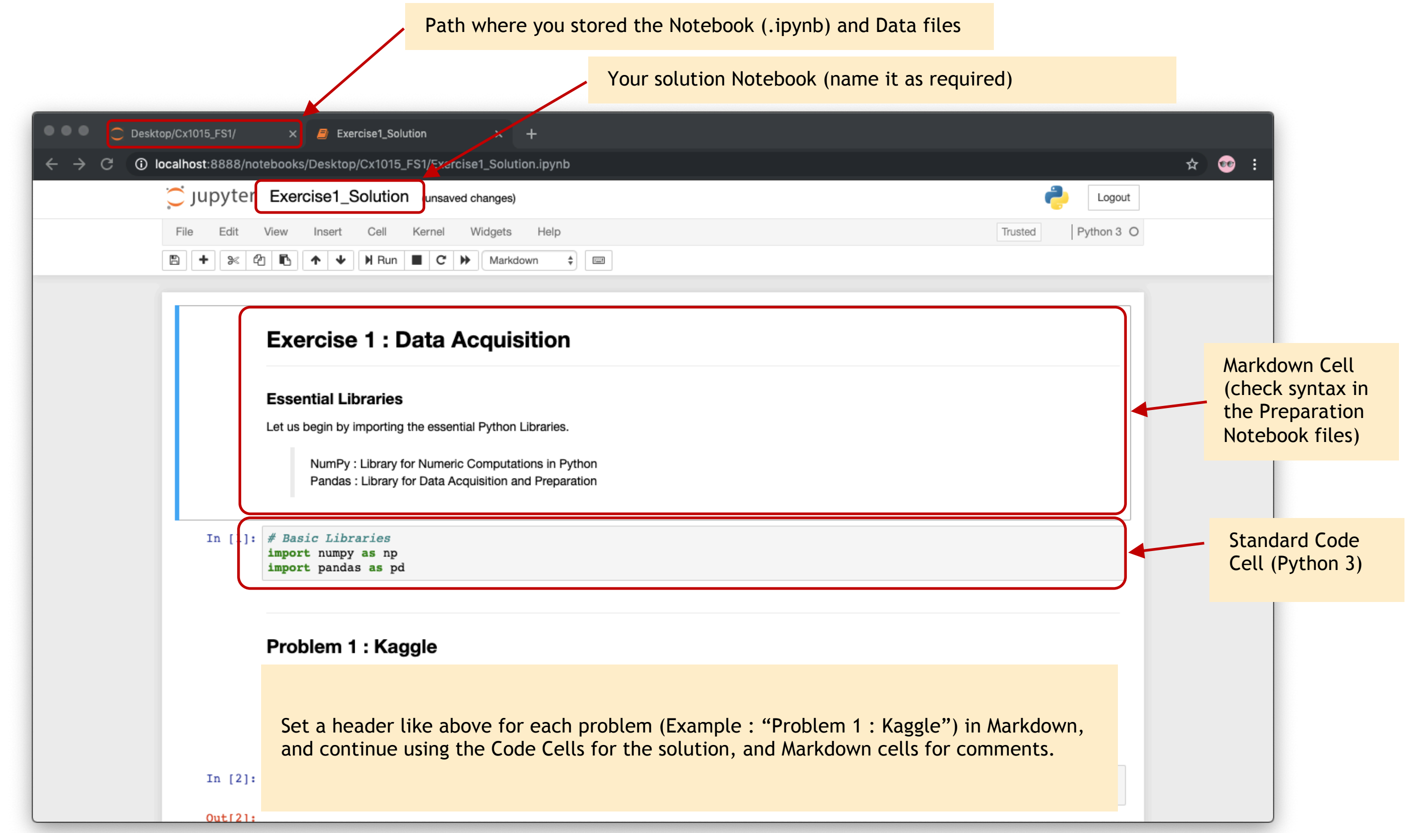
Explore the dataset using .shape, .info() and .describe(), exactly as you did in Problem 1 above. Do you spot anything interesting while exploring this dataset? Discuss amongst friends or talk to the Instructor, if you did.

B. Note that the Summer Olympic medal tally on Wikipedia follows a really nice structure for the URL, where you can simply change the year in https://en.wikipedia.org/wiki/2016_Summer_Olympics_medal_table to fetch any Summer Olympic page. Try changing 2016 in the URL to 2012 or 2008 or 2004 to see for yourself. This allows us to fetch the Olympics medal table from all these years (in fact, any year) quite easily. Let’s try the following.

Write a loop to extract the main tables, “20XX Summer Olympics medal table”, from 2000 to 2016, that is, for the five consecutive Olympics in 2000, 2004, 2008, 2012 and 2016. Store all five tables in respective DataFrames. Now, extract the TOP 20 countries from each of these medal tables, and store these rows as new DataFrames.

Notebook

Your Notebook setup may look something like the following example. Seek help from the Instructor if you face problems.



Path where you stored the Notebook (.ipynb) and Data files

Your solution Notebook (name it as required)

Exercise 1 : Data Acquisition

Essential Libraries

Let us begin by importing the essential Python Libraries.

> NumPy : Library for Numeric Computations in Python
> Pandas : Library for Data Acquisition and Preparation

```
In [ ]: # Basic Libraries
        import numpy as np
        import pandas as pd
```

Problem 1 : Kaggle

```
In [2]:

Out[2]:
```

Markdown Cell (check syntax in the Preparation Notebook files)

Standard Code Cell (Python 3)

Set a header like above for each problem (Example : “Problem 1 : Kaggle”) in Markdown, and continue using the Code Cells for the solution, and Markdown cells for comments.