Lab 01 - Data Cleaning

In this lab, we will discuss how to clean and preprocess data. This is a very important part of the development cycle of a machine learning solution.

Installing packages

- First, you need to install a python environment with all the libraries like numpy, scipy, matplotlib and pandas. If you are an expert, you can install these libraries one by one.
- Or, if you need a simplified installation, you can install the 'Anaconda' development environment. 'Anaconda' is a package manager, an environment manager, a Python/R data science distribution and it contains almost all the libraries you need in this course. Anaconda is free and it offers free community support. It supports Linux, Windows and macOS.
- Click on the relevant link, follow the installation instructions and install anaconda in your machine.
    - Windows: https://docs.anaconda.com/anaconda/install/windows/
    - Linux: https://docs.anaconda.com/anaconda/install/linux/
    - macOS: https://docs.anaconda.com/anaconda/install/mac-os/

Running the Jupyter notebook

- First, download the notebook and the dataset from moodle. Keep both of them in the same directory.
- Linux/macOS: You can start the notebook server from the command line using the following command.

  ```
  $ jupyter notebook
  ```

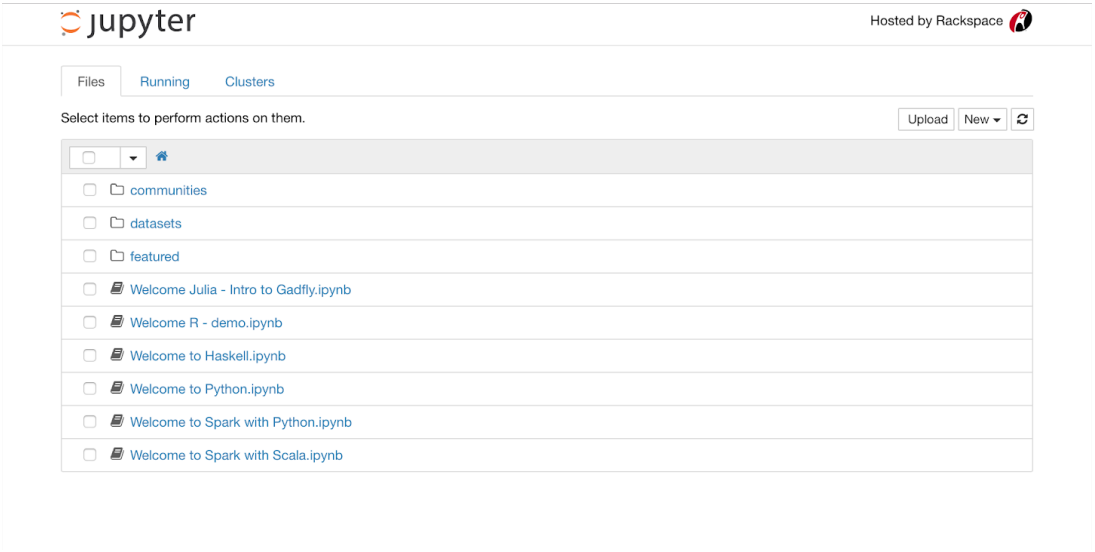
  If everything is correct you should see something similar to the following.

  ```
  [I 08:58:24.417 NotebookApp] Serving notebooks from local directory:
  /Users/rajitha
  [I 08:58:24.417 NotebookApp] 0 active kernels
  [I 08:58:24.417 NotebookApp] The Jupyter Notebook is running at:
  http://localhost:8888/
  [I 08:58:24.417 NotebookApp] Use Control-C to stop this server and
  shut down all kernels (twice to skip confirmation).
  ```

  It will then open your default web browser to this URL.

- Windows: In the start menu, search for 'jupyter'. You should see the jupyter application listed in the search result. Open it and a command prompt will pop up following which your browser will open and show you the interface of Jupiter server

- Here is an image of jupyter server interface.



- The 'Files' tab will allow you to browse the files in your system
- Browse and go to the directory you saved the notebook, using this interface, double click on the notebook file and it will be open in a new tab.

Lab tasks

- Please go through the notebook carefully. If you are unclear about anything please ask a demonstrator attending in your lab. If you need further clarification please contact us by email or rocketchat.
- There are 2 tasks in the notebook for you. You should complete both and upload the modified notebook into the moodle before the deadline. There is a submission link in the lab submissions section.
- Please take note that Data cleaning will be a crucial part of your project, and the experience you gain from this lab will help you a lot.