Real world data science

17th July

Parwez Alam
Katharina Ecker

1. Creating a GitHub repository on https://github.ibm.com/ for us to share code.
   - If you do not have an account you need to create one.
   - Give it a reasonable name :)
   - Share the repository with me and give me access rights (if it is not public).
   - Git cheat sheet: https://github.github.com/training-kit/downloads/github-git-cheat-sheet.pdf

2. Create a folder structure and add a first Jupiter notebook.
   - For now it is okay if you create a directory notebooks and save a notebook with a reasonable name in it. We will use it to do some data science.
   - We also need a data/raw/ directory.
   - Read: http://drivendata.github.io/cookiecutter-data-science/
   - What is the project about? Why is it important to use a common structure for your data science projects?

3. Find data you want to look at, some excel or csv (maybe from kaggle) or some built in data set from scikit learn. (Hint: if you are stuck you can use the iris data set from scikit learn).
   - In case you have a csv or excel file: Save the data into the data/raw directory.
   - Load the data into a pandas data frame in your Jupiter notebook.

4. Perform data exploration: Try to formulate two questions you want to have answered at the end of the day and make plots that answer those questions.
   - Show me some nice plots :)
   - If you are daring, use seaborn, they have a nice plotting functionality

5. Do supervised or unsupervised machine learning