Data Mining: Preparation for Practical Assignments

Due on Thu & Fri, April 25-26 2019, 10:15am-13:15 & 14:15am-17:15

For this tutorial we will be using *Jupyter* notebooks. *Jupyter* is a powerful tool to show and execute quickly your Python code in a browser. It is already installed on the PCs in the Data Mining pool rooms (D-114/D-118). A complete documentation is available here: `http://jupyter.org/documentation` Open a terminal, create a directory (`mkdir a_name`), `cd` into your new directory and start the Jupyter server typing `jupyter notebook`. Select the Python 2.7 option on the right. You can directly start entering Python code or call an existing notebook from your directory.

Task 1

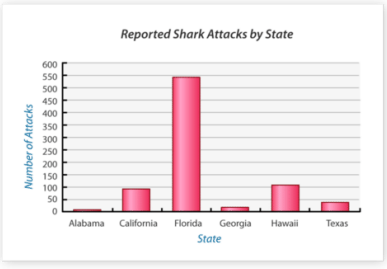
Data visualization, ≈ 45 min.

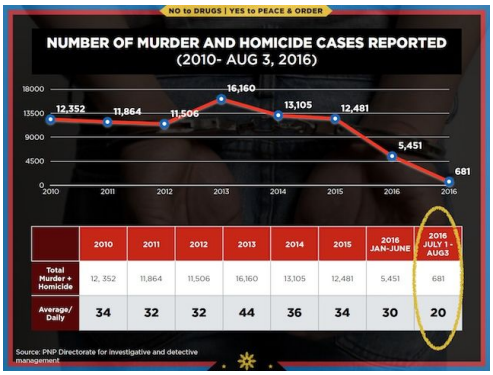
In the folder *DAMI2-Data* you find several files containing data (please note their different file format) which needs an appropriate visualization for data analysis and interpretation. A brief summary of the datasets is provided in the accompanying `datasetDescriptions.txt`. You can choose from a set of different built-in plot functions provided by the `matplotlib` library (at the least the standard plots: bar and pie chart, histogram, and boxplot). You can use the `DAMI2_dataViz.ipynb` template or create your own one. Your approach to solve this task should be guided by the following questions:

1. What are characteristics of the plot? Especially for a boxplot: which specific measures you can read out from it?
2. Why do you think this plot was best to demonstrate the data?
3. ***Optional:*** Find or generate some other data and try out different plots available in `matplotlib`.

Task 2

Plots gone wrong, ≈ 15 min. You are given the following graphics but, unfortunately, they are quite misleading. Can you spot the flaws and explain their possible sources?





Task 3

Correlations & Independence, ≈ 45 min.

A very common step in data mining is to explore possible correlations among features in a dataset. Take the `DAMI2_correlations.ipynb` file and explain, which correlations do you detect for the different data (`x_data, y_data`)?
Another technique in data mining is to test for independence between variables using statistical tests. An example is the χ^2 test (cf. Lecture 3), which you are now asked to perform on the example given below:

	Likes Zombie movies	Does not like Zombie movies	Total
Plays harf	24 ()	6 ()	
Does not play harf	32 ()	22 ()	
Total			

- Calculate the missing values in the table cells and in the parentheses.
- What is the degree of freedom in this example?
- What is the decision of this test assuming $\alpha = 5\%$?
- ***Optional:*** Discuss the following statement: "Uncorrelated random variables are always independent".

Task 4

Regression, ≈ 40 min.
The `DAMI2_simpleRegression.py` serves you as a template to implement a simple regression task. Check the script carefully and add the missing lines of code to perform a regression. Students experienced with Python can set up their own script from scratch.

1. Why is a regression also referred to as *ordinary least square*? (No mathematical derivation)
2. What are the values for β_0 and β_1 and what do they mean given your regression model?
3. For which data distribution a regression would not be a good fit?
4. ***Optional:*** *How would you extend your Python script to perform a multivariate regression?*

@home Task

For the next tutorial, prepare the following topics:

- Decision Trees: Entropy, Classification with decision trees
  "Data Mining, Practical Machine Learning Tools and Techniques", Witten/Frank/Hall/Pal, ***Chapter 6***
  "Machine Learning, Tom Mitchell", ***Chapter 3***
- Neural Networks: Perceptrons, Multi-Layer perceptrons (MLP), neural network training, backpropagation algorithm
  "Neural Networks - A Systematic Introduction", Raul Rojas, ***Chapters 3 (until 3.3.4), 4 (until 4.2.5), 7 (until 7.3.4)****, freely available* `https://page.mi.fu-berlin.de/rojas/neural/`