# Data Mining: Preparation for Practical Assignments

Due on May 8, 2020, 2pm

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. A complete documentation is available here: `http://jupyter.org/documentation`. Using `pip`, you can install *Jupyter Notebook* by running `pip install notebook`. More detailed instructions can be found on `https://jupyter.org/install`. Open a terminal, create a directory (`mkdir a_name`), `cd` into your new directory and start the Jupyter server typing `jupyter notebook`. You can directly start entering Python code or call an existing notebook (`.ipynb`) from your directory.
The following exercises also require the use of several Python libraries, which can be installed by running: `pip install numpy matplotlib`
You must hand in all your solutions as single document (`.pdf`). Feel free to use LaTeX, Word, LibreOffice or anything similar.

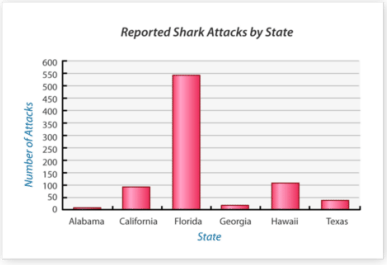
## Task 1

***Data visualization, 30 points.***
In the folder *DAMI2-PlotData* 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. Answer the following questions:

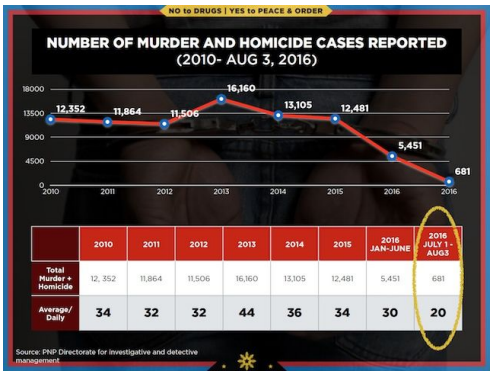
1. What are characteristics of the plot? Especially for a boxplot: which specific measures you can read out from it?
2. Why do you think this plot was best to demonstrate the data?
3. ***Optional:*** Find or generate some other data and try out different plots available in `matplotlib`.

## Task 2

***Plots gone wrong, 10 points.*** You are given the following graphics but, unfortunately, they are quite misleading. Can you spot the flaws and explain them?








| | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 JAN-JUNE | 2016 JULY 1 - AUG3 |
|---|---|---|---|---|---|---|---|---|
| Total Murder + Homicide | 12, 352 | 11,864 | 11,506 | 16,160 | 13,105 | 12,481 | 5,451 | 681 |
| Average/ Daily | 34 | 32 | 32 | 44 | 36 | 34 | 30 | 20 |

# Task 3

***Correlations & Independence, 30 points.***
A very common step in data mining is to explore the 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 $\chi^2$ 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 | 8 () | 16 () | |
| 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, 30 points.***
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 $\beta_0$ and $\beta_1$ 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?*