# MPP-E1180: Introduction to Collaborative Social Science Data Analysis

## Assignment 1

### Spring 2016

**Version: 01 March 2016**

#### By: Md Mujahedul Islam & Johannes Schulz-Knappe

### Description of the project

We have analysed two seperate example datasets that were pre-installed in R. Md Mujahedul Islam has analysed the **USArrests** datasets while Johannes Schulz-Knappe has analyzed the **nottem** dataset. This repository was created on February 15 for the first pair assignment of the course [MPP-E1180: Introduction to Collaborative Social Science Data Analysis](https://github.com/HertieDataScience/SyllabusAndLectures), taught by [Christopher Gandrud](https://github.com/christophergandrud). This readme file documents our work done for this assignment.

### USArrests dataset by Md Mujahedul Islam

I use the example dataset set up within R called [USArrests](https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/USArrests.html) that provides data on the number of arrests for murder, assault, and rape for each of the 50 states in the United States in year 1973. It also provides data on how many people of a particular state (in percent) live in urban areas. There are thus in total 4 variables named *assault*, *murder*, *rape* and *urbanpop* and 50 rows that contain the names of each states. I visualize the data in four different graphs, showing the distribution of the rate of murder and rape arrests over the different states of the US as well as frequency distributions of the urbanization rates and rates of arrests for violent assaults.

### nottem DATASET by Johannes Schulz-Knappe

**Description of the dataset**

The dataset [nottem](https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/nottem.html) is a time series object containing average air temperatures at Nottingham Castle in degrees Fahrenheit for the years 1920 to 1939. It consists of 240 observations in a single vector.

**Visualisation of the dataset**

In order to analyse it, I convert the data from fahrenheit into degrees celsius. Furthermore I create a new dataframe in order to construct the last graph. I presented it in three diagrams:

* The first graph shows a frequency distribution of the observed temperature levels. The data is reasonably normally distributed.
* The second graph shows a temperature chart over the entire timespan of the dataset.
* In order to show the different levels of variation of the average monthly temperatures across the observed 20 years, I present the data in a third graph consisting of multiple boxplot diagrams. Each boxplot shows the variation of the average monthly temperature of the respective month.

### Accessing the analysis

To repeat all of our analysis, execute the commands contained in the [1\_Analysis\_Main.R](https://github.com/mujahedhertie/Assignment01/blob/master/1_Analysis_Main.R) file. It will automatically run both datasets.

For more detailed insight, directly access:

* [2\_Analysis\_Mujahedul.R](https://github.com/mujahedhertie/Assignment01/blob/master/2_Analysis_Mujahedul.R) for the USArrests dataset analysis
* [3\_Analysis\_Johannes.R](https://github.com/mujahedhertie/Assignment01/blob/master/3_Analysis_Johannes.R) for the nottem dataset analysis

### Pull requests

As part of the Assignment 1, we made a pull request to ...