Week 2: Data Munging

Download from SurreyLearn in the lab 2 material section the “lab2_DataMunging.ipynb” and the required data sets in “cars.csv”, “census.csv”, “log.csv”, and “olympics.csv” in the same folder. Load the python notebook and follow the instructions to understand every section in the code. For any missing module use the pip install module and try again.

Then, for an introduction to big data, download “lab2_PySpark_1.ipynb” for PySpark RDD introduction, and “lab2_PySpark_2.ipynb” for PySpark DataFrames introduction. Again you will need to pip install any missing module. A number of data files are zipped in Data.zip to be used in these notebooks, please download them in the same folder.