SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTING

BIG DATA AND DATA ANALYTICS
LAB PROJECT 1
(DATASET DESCRIPTION & PREPARATION)

This lab project is based on a dataset from the National Institute of Diabetes and Digestive and Kidney Disease, which is available from the UCI Machine Learning Repository (Lichman, 2013).

PREPARATION

In preparation for lab project 1, load the pimadata.csv dataset into R and use it to practice the following topics:

- Load and explore *.csv datasets
- Use "?" to explore the R documentation (e.g., "?mean", "?median", "?summary")
- Search for online information on the topics "data frame" and "vector" (in R). How do data frames and vectors relate to each other in R?
- Find out more about the concept of a data frame and the function "data.frame()" on the website *stackoverflow* (http://stackoverflow.com/) and on *YouTube*
- Selecting subsets of data frames (e.g., autodata$weight[autodata$origin=="Japanese"])
- Simple statistics: mean, median, quartiles, percentiles, variance, standard deviation
- Writing custom functions (e.g., to calculate the standard error)

REFERENCES

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

DATASET

Pimadata	*Pima Indians Diabetes Database*

Description

A diabetes dataset. All patients here are females at least 21 years old of Pima Indian heritage. Note: Even though the dataset donors made no such statement, it seems very likely that several zero values encode missing data for several variables.

Usage

Pimadata

Format

A data frame with 768 observations on the following 9 variables.

`timesPregnant`	Number of times pregnant
`PCG`	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
`DBP`	Diastolic blood pressure (mm Hg)
`TSFT`	Triceps skin fold thickness (mm)
`insulin`	2-Hour serum insulin (mu U/ml)
`BMI`	Body mass index (weight in kg/(height in m)^2)
`DPF`	Diabetes pedigree function. It provides some data on diabetes mellitus history in relatives and the genetic relationship of those relatives to the patient. This measure of genetic influence gives an idea of the hereditary risk one might have with the onset of diabetes mellitus.
`age`	Age (Years)
`diabetes`	1 tested positive for diabetes 0 tested negative for diabetes

Source

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.