**Lab 3**

**Data Exploration Using R**

We are using this lab to get familiar with the R environment, in which machine learning models will be built in the later parts of the class. Our focus here include import data from a file; manage data with R data structures; and carry out basic statistic calculations and visualization tasks similar to those we have seen in the first two labs.

**Setting up the Stage**

* Managing the working directory (WD)
  + Use getwd() to see what is the default WD, and use setwd() to change it to your lab3 directory, if necessary. (Use “/” for your convenience, “\\” is needed otherwise.)
  + Use ls() to list any saved data sets in memory. (You may remove data sets that you don’t need using rm(), such as rm(list=ls()) removes all of them.)
* Read a data file (in CSV format)
  + Using read.csv(“<file\_name>”). (Path needs to be included if it’s not in your WD.)
  + Assign data set to a data frame identified by a (variable) name. Use naming convention consistently: such as <dataset\_name>.dat for (original) data, .mdl for model, etc.

**Getting to Know Your Data**

* Functions for showing meta data
  + Use the following functions and see what each returns

|  |  |
| --- | --- |
| **Function** | **Purpose** |
| names() |  |
|  | |
| head() |  |
|  | |
| dim() |  |
|  | |
| str() |  |
|  | |

* Basic Stats
  + Summarizing your data – the quick version:

|  |  |
| --- | --- |
| summary() |  |
|  | |

* + For numeric features – just choose one:

|  |  |
| --- | --- |
| sd() | Calculate standard deviation |
| <screenshot> | |
| length(unique()) | Get count of unique values |
| <screenshot> | |
| sum(is.na()) | Count number with missing (NA) values |
| <screenshot> | |

* + For categorical features – just choose one (that’s not binary):

|  |  |  |
| --- | --- | --- |
| table() | Generate frequency table | |
| <screenshot> | | |
| tbl <- table(<frame$col>)  pct <- prop.table(tbl)  round(pct, digits = 1) | | <explain> |
| <screenshot> | | |

**Generate Charts With Appropriate Title and Axis Labels**

* For numeric features – just for the one that you have chosen above
  + Boxplot
  + Histogram
* For categorical features – just for the one that you have chosen above
  + Bar chart

**Saving Results in a File**

* (Frequency) Table & charts

**Due by 2/8, at beginning of class.**