Université Grenoble Alpes
M1 SSD/ M1 MSIAM /Applied probabiliy

Lab session 2

Data Cleaning

The `forestfires.txt` dataset is a collection of observations from the Monteshinho park in Portugal, that were used to build a predictive model of forest fire areas. The columns of the data frame are:

- `X` - x-axis spatial coordinate within the Montesinho park map: 1 to 9
- `Y` - y-axis spatial coordinate within the Montesinho park map: 2 to 9
- `month` - month of the year: "jan" to "dec"
- `day` - day of the week: "mon" to "sun"
- `temp` - temperature in degrees Celsius
- `RH` - relative humidity in %: 15.0 to 100
- `wind` - wind speed in km/h
- `rain` - outside rain in mm/m2
- `area` - the burned area of the forest (in ha)

1. Your dataset is given in the website https://toltex.u-ga.fr/M1Math, under

   `<groupXX>_forestfires.csv`.

   Read the dataset into `R`. The dataset is in a comma separated format, and has a header line.

2. Print the dataset out on the screen using the `print` command. Scroll through the dataset. With a small dataset such as this one, scrolling through the dataset can be useful to familiarize yourself with the data and spot problems. Can you see anything in this dataset that might pose a problem for future analysis? Can you see any entries that appear to be errors? If so, how would you propose fixing those errors? (Don't copy-past the entire dataframe).

3. Now summarize the dataset using the `summary` command. Does anything strike you as unusual in the data summaries? Investigate further and propose fixes for any errors or irregularities.

4. Create a scatterplot matrix of the variables and once again look for anomalies in the plots. Can you see any plots which might suggest outlying or extreme observations? Investigate these with individual scatterplots and/or boxplots. Are the outliers reasonable, or do they signify an error?

5. Provide a cleaned version of the dataset as an appendix at the end of assignment file, attempting to preserve as much data as possible. Justify each change you make.

6. Provide a statistical description of the data: each variable and each pair of variables. Comment.

Please make sure that your answers are as specific as possible. For example, do not write "I deleted some of the records because some entries appear to be errors." Instead write record no... was deleted because field... appears to be incorrect - it contains a value which is ..." Unspecific answers will not be fully graded. Include all the `R` commands, plots, summaries and any other output of `R` you consider relevant. Submit your work and the full script file.