SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTING

BIG DATA & DATA ANALYTICS
LAB PROJECT 3

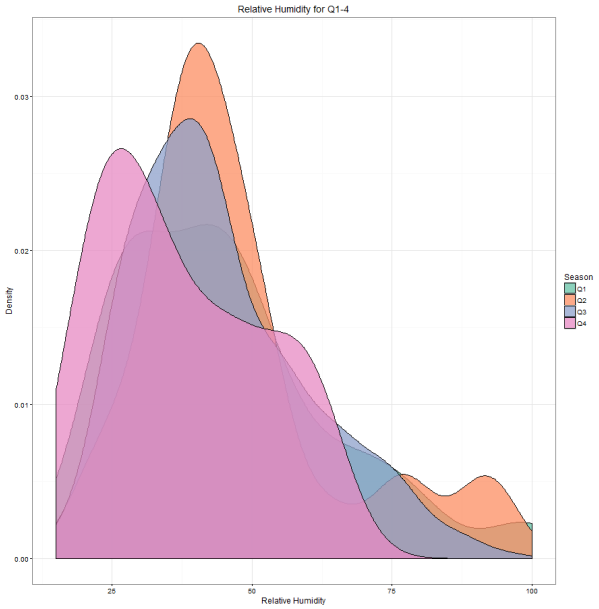
This lab project is based on a meteorological dataset about forest fires in the northeast region of Portugal. The dataset is available from the UCI Machine Learning Repository (Lichman, 2013):

http://archive.ics.uci.edu/ml/datasets/Forest+Fires

EXERCISE 1 (2 MARKS) *[R-CODE]*

Create a new column "quarter" in the dataframe firedata that describes the meteorological quarter of the observation based on the month. Distinguish between the following categories: "Q1", "Q2", "Q3", "Q4", referring to the four quarters of a calendar year. In particular, the year is divided into 4 quarters:

- Q1: January 1 to March 31
- Q2: April 1 to June 30
- Q3: July 1 to September 30
  Q4: October 1 to December 31



Use ggplot() to create a density plot for the relative humidity across different quarters (using the newly created quarter variable) and directly export this plot as png-file titled "rh_quarters.png" with a resolution of 900x900.

EXERCISE 2 (2 MARKS) *[R-CODE]*

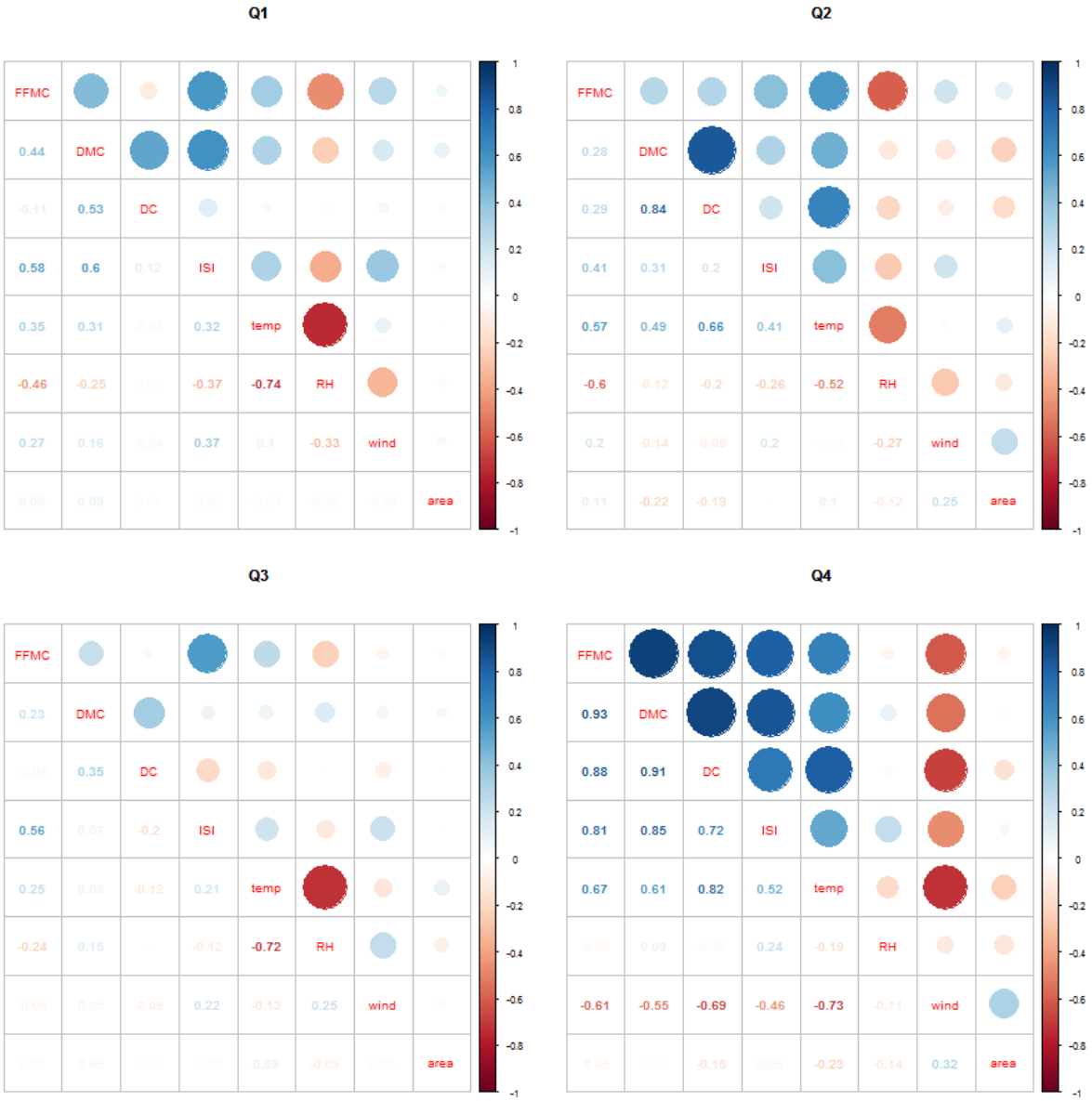
Use the ddply() function of the package "plyr" to create a data frame "summarystat" with the means of FFMC, DMC, DC, ISI, temp, RH, wind, rain, and area. Also include the number of observations N for each of the four quarters. The output should look like this:

	quarter	N	ffmc_avg	dmc_avg	dc_avg	isi_avg	temp_avg	RH_avg	wind_avg	rain_avg	area_avg
1	Q1	76	86.70	27.10	70.72	5.97	11.97	45.42	4.57	0.00	4.75
2	Q2	28	88.11	63.72	203.05	9.21	17.36	47.25	4.33	0.00	7.78
3	Q3	388	91.77	135.61	666.83	9.83	20.78	44.29	3.82	0.03	15.06
4	Q4	25	88.04	34.38	539.72	5.58	12.36	37.56	5.01	0.00	8.78

Use the package "rtf" to create an rtf document called "output.rtf" and include the newly created data frame "summarystat" as a table in this document.

EXERCISE 3 (2 MARKS) *[R-CODE]*

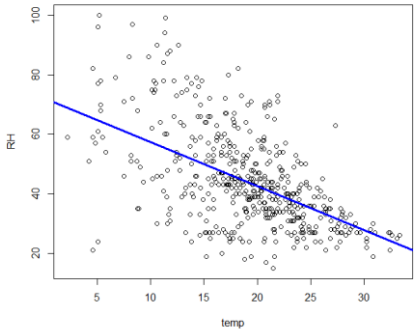
Use the functions “tapply” and “cor” to create a Pearson correlation matrix for FFMC, DMC, DC, ISI, temp, RH, wind, and area, separately for each of the four quarters (Q1-4). Use the newly created correlation matrices to create a set of mixed correlation plots with the corrplot command. Save the result as a PDF file “corrplots.pdf” with a 9x9 resolution. Identify and interpret at least three differences across two or more quarters.



Optional: Try using “par(mfrow =c(2,2))” before creating the plots to display all four plots into one bigger plot.

EXERCISE 4 (1 MARK) *[R-CODE]*

Use “plot” to draw a scatterplot showing RH and temp. Add a blue regression line to the plot.



EXERCISE 5 (2 MARKS) *[R-CODE]*

Use R to create a simple linear regression that regresses RH on temp. Interpret the coefficients. Retrieve and interpret the R^2.

EXERCISE 6 (1 MARK) *[R-CODE]*

Use the function “skewness” of the package “moments” to investigate the distribution of the variable area. Interpret the result.

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

Forestfires	*Forest Fires Dataset*

Description

A meteorological dataset about forest fires in the northeast region of Portugal. Some of the data includes measures from the Fire Weather Index (FWI).

Usage

```
Forestfires
```

Format

A data frame with 517 observations on the following 13 variables.

`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'
`FFMC`	FFMC (Fine Fuel Moisture Code) index from the FWI system: 18.7 to 96.20. This is a numerical rating of the moisture content of surface litter and other cured fine fuels. It shows the relative ease of ignition and flammability of the fine fuels.
`DMC`	DMC (Duff Moisture Code) index from the FWI system: 1.1 to 291.3. The DMC is a numerical rating of the average moisture content of loosely compacted organic layers of moderate depth.
`DC`	DC (Drought Code) index from the FWI system: 7.9 to 860.6. is is a numerical rating of the moisture content of deep, compact, organic layers. It is a useful indicator of seasonal drought and shows the likelihood of fire involving the deep duff layers and large logs.
`ISI`	ISI (Initial Spread Index) index from the FWI system: 0.0 to 56.10. This indicates the rate fire will spread in its early stages.
`temp`	temperature in Celsius degrees: 2.2 to 33.30
`RH`	relative humidity in %: 15.0 to 100
`wind`	wind speed in km/h: 0.40 to 9.40
`rain`	outside rain in mm/m2 : 0.0 to 6.4
`area`	the burned area of the forest (in ha): 0.00 to 1090.84

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.

Cortez, P. & Morais, A. (2007). A Data Mining Approach to Predict Forest Fires using Meteorological Data. In J. Neves, M. F. Santos and J. Machado Eds., *New Trends in Artificial Intelligence*, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimaraes, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9.

FWI (Fire Weather Index) descriptions are taken from: http://www.malagaweather.com/fwi-txt.htm