Project 1 Requirements

Body fat percentage refers to the relative proportions of body weight in terms of lean body mass (muscle, bone, internal organs, and connective tissue) and body fat. The ideal body fat percentage varies considerably by gender and by age. Persons with higher than the recommended range of body fat are considered to be at increased risk for disease[1].

The most accurate means of estimating body fat percentage are cumbersome and require specialized equipment. The “gold standard” is completely submerging a person in water and calculating the volume of the displaced water. This is what this looks like[2]:

Image from: https://www.fau.edu/education/academicdepartments/eshp/images/underwater.jpg

Physicians *estimate* body fat percentage from anatomical measurements (e.g., abdomen circumference) that are much easier to obtain. In this project, you should use a dataset of 13 measurements from subjects (all men) along with their bodyfat percentage. Your goal is to come up with a formula that can be used to estimate bodyfat percentage using only (some or all of) the 13 measurements. (The “density” column should not be used in this project.)

The challenge is to identify which combination of the 13 predictors will give the most accurate estimate and if transforming some of the variables will increase accuracy. You are encouraged to try a few different combinations of predictors. You can use some domain knowledge to pick the predictors.

---

[1] pennshape.upenn.edu/files/pennshape/Body-Composition-Fact-Sheet.pdf

[2] A full video is here: https://www.youtube.com/watch?v=kgIIcATPQWI

Data

http://staff.pubhealth.ku.dk/~tag/Teaching/share/data/Bodyfat.csv

Note that you can read from the URL directly, like so:
`read.csv("http://staff.pubhealth.ku.dk/~tag/Teaching/share/data/Bodyfat.csv")`
(Information about the dataset is here:
http://staff.pubhealth.ku.dk/~tag/Teaching/share/data/Bodyfat.html)

Hints

There are two major steps in this project:

1) Exploratory analysis to identify variables, remove outliers, identify candidate variables for modeling body fat percentage, possible transformations, etc.
2) Modeling body fat percentage. You will need to decide if regression (linear modeling), classification (k-NN), or clustering (k-means) is the appropriate approach. You should evaluate models using an appropriate metric (e.g., R2, precision, or recall)

Submission:

1. Write a short report listing the different combinations of predictor variables you tried, and if you tried transforming any of the variables. The report should include a plot that shows the evaluation metric of the different models (i.e., a comparison of the R2 values). [A PDF file]
2. A listing of your R code [.R file]
3. An R function of the following form that returns your best body fat percentage prediction [filename must be `bodyfatpercentage.R` and must include a function with the same name that can take all these named parameters]:

```
bodyfatpercentage <- function(Age, Weight, Height, Neck, Chest,
Abdomen, Hip, Thigh, Knee, Ankle, Biceps, Forearm, Wrist) {
# your code goes here
}
```

The function must be a stand-alone function, i.e., do not read from any file inside it.

Due date:

Tuesday 12/3, 5:30pm on Titanium. Submit three files: the PDF report, full code (.R), and bodyfatpercentage function (.R).

Group work:

You may work in groups of 1-3. Include all group member names in the PDF file. Only one person in the group needs to submit to Titanium.