Dataset : People habits

Factual and medical information about people.

* **dob**: date of birth
* **gender**: gender of the patient (1=female and 2=male)
* **height**: height of the patient (cm)
* **weight**: weight of patient (kg)
* **gluc:** glucose level (1=normal, 2=above normal, 3=well above normal)
* **smoke:** whether or not the patient smokes
* **alco:** alcohol intake
* **active:** physical activity

**General instructions**:

You will need to upload on Moodle a notebook and an html file. Take care with the presentation and answer the questions carefully, whether in code or with a sentence. For the plots, when you can (and when it adds value), modify the colors, add a title, change the axis titles and you can also change the theme of the plot. Your plots must be as clear and legible as possible, so think carefully about your choice of colors.

For parts II to V, one answer is expected per question. If there is more than one answer, the answer will be considered false, as no choice will be made by the examiner.

Use, as much as possible, *dplyr* package especially for parts II, IV and V.

Any commented line of code will not be taken into account for the marking. And be sure to show all important results in your notebook and in your html file. If you store a result in a variable, remember to print it, otherwise it will not be taken into account.

It is your turn to play. Good luck to all!

Part I – Explore the data (1 point)

Read in the biological dataset, people\_habits\_dataset.csv.

1. What are the dimensions of the data?
2. What are the types of columns? What is the difference between a string and a factor? Show descriptive statistics of the data.
3. Is there any *NA?*

Part II – Data transformation (7 points)

1. Create a new variable called age from the variable dob to obtain the age of individuals based on their date of birth. (Hint: you can use the function *year()* from *lubridate* package to get the year of the date of birth).
2. Calculate the body mass index (BMI) for each of our patients using the formula BMI=Weight/(Height)2, where the weight is measured in kilograms and the height is in meters
3. Create a new variable to indicate which individuals are overweight, i.e. those with a BMI over 25.
4. Some variables are treated as integers when they should be categorical variables. Change them into factors.
5. Change levels of the gender variable, replacing 1 and 2 with *female* and *male*.
6. Group your data by gender and gluc. For each group, compute the mean and standard deviation of weight, then plot out a bar plot of computed means with error bars and use a facet to separate the data for the two genders.

Part III – Plotting (5 points)

1. Plot a jitter plot of the weight of individuals for both categories of active.
2. Plot out a grouped barplot of the number of individuals in the 3 categories of gluc, colored by active.
3. Plot out a boxplot of age for both categories of alco.
4. Plot out a violin plot of the BMI for both males and females.

Part IV – Selecting columns (3 point)

1. Select the columns 1 and 4.
2. Select all the columns ending with the letter “e” .
3. Select all the columns except smoke.

Part V – Filtering rows (4 point)

1. Filter the rows for males aged 64 and over.
2. Filter the rows for alcoholics or smokers.
3. How many people smoke, drink and are active?
4. How many patients are overweight? Show them by gender in a table.