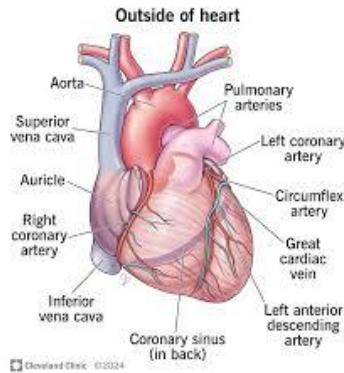# Part 1 : Help doctors saving lives



You are a brilliant young doctor trained in data analysis. You are attending the annual conference for general practitioners and would like to present a new protocol for detecting heart disease. The aim of this protocol is to determine whether a patient has heart disease with a minimum of tests.

To create this protocol, you will use a dataset : Heart_Disease_Prediction.csv
and follow the steps mentioned below:

1) Check your dataset : check the number of rows and variable types, convert the variable if not right type, check for missing data and replace, check for duplicated rows.
2) Descriptive analysis : for all variables, provide some descriptive analysis (Q1, Q3, mean, median, variance), histograms and distributions graphs. Provide an analysis based on this.
3) Multivariate analysis : build the correlation matrix and conclude. Provide an analysis based on this.
4) For each variable, perform a statistical test (select the test based on the variable type) to know if it can be cause of heart disease
5) Build your protocol based on those test
6) BONUS 1 : Build a classification tree to have a more precise protocol and compare with the one you built manually
7) BONUS 2 : Share your analysis and the final protocol in a convincing presentation

# Part 2 : Dashboard for surfers !

As a data analyst, you have been contacted by a Surf School to help them to collect and visualize data relative to the sea conditions. They want to have a dashboard in the school showing the best moment to practice surf in the week.

The school is located near Bordeaux. In this area, the best conditions to practice this activity is when the sea generates high waves and the wind speed is strong coming from the North. So, you will have to build appropriate KPIs relative to this information.

## Step 1 : Data extraction

School's instructors use the following webpage to get information about the sea : https://www.surf-report.com/meteo-surf/lacanau-s1043.html

### SAMEDI 22 OCTOBRE

| | | Vagues | Houle | | | Autres houles | | | Vent | | Temps | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 08:00 | ★☆☆ | 0.8m - 1.3m | 1.3m | 7s | ↑ | 0.8m | 5s | ↙ | 3km/h | ◀ | ☁ | 20°C |
| 11:00 | ☆☆☆ | 0.7m - 1.2m | 1.2m | 7s | ↑ | 0.7m | 6s | ↙ | 15km/h | ▼ | ☁ | 21°C |
| 14:00 | ★☆☆ | 0.7m - 1.1m | 1.1m | 6s | ↑ | 0.7m | 6s | ↙ | 20km/h | ▶ | ☁ | 22°C |
| 17:00 | ★☆☆ | 0.6m - 1.0m | 1.0m | 6s | ↑ | 0.7m | 5s | ↙ | 21km/h | ▶ | ⛅ | 22°C |
| 20:00 | ★☆☆ | 0.5m - 0.9m | 0.9m | 6s | ↑ | | | | 23km/h | ▼ | ☁ | 21°C |

On this webpage for each day (7 days) displayed, you will capture 5 kind of data :
- The date (Here : "Samedi 22 Octobre")
- The time, in the 1st column (example : 08:00)
- The wave size, in the 3rd column : (Vagues : "0.8m - 0.7m")
- The wind speed (colored number - example : 3km/h)
- The wind direction (the arrow next to the wind speed)

What you have to do :
1) Extract the data

2) Put the extracted data in a dataframe. It should look something like that :

| | day | hour | waves_size | wind_speed | wind_direction |
|---|---|---|---|---|---|
| 0 | Samedi 22 Octobre | 08:00 | 0.8 - 1.3 | \n3\n\n | Orientation vent Est |
| 1 | Samedi 22 Octobre | 11:00 | 0.7 - 1.2 | \n15\n\n | Orientation vent Est Sud Est |
| 2 | Samedi 22 Octobre | 14:00 | 0.7 - 1.1 | \n20\n\n | Orientation vent Sud Est |
| 3 | Samedi 22 Octobre | 17:00 | 0.6 - 1.0 | \n21\n\n | Orientation vent Sud Sud Est |
| 4 | Samedi 22 Octobre | 20:00 | 0.5 - 0.9 | \n23\n\n | Orientation vent Sud Est |
| 5 | Dimanche 23 Octobre | 08:00 | 0.7 - 1.1 | \n27\n\n | Orientation vent Est Sud Est |
| 6 | Dimanche 23 Octobre | 11:00 | 0.7 - 1.1 | \n35\n\n | Orientation vent Est Sud Est |
| 7 | Dimanche 23 Octobre | 14:00 | 0.7 - 1.2 | \n43\n\n | Orientation vent Sud Est |
| 8 | Dimanche 23 Octobre | 17:00 | 0.7 - 1.2 | \n46\n\n | Orientation vent Sud Est |
| 9 | Dimanche 23 Octobre | 20:00 | 0.8 - 1.3 | \n45\n\n | Orientation vent Sud Est |
| 10 | Lundi 24 Octobre | 08:00 | 0.7 - 1.1 | \n31\n\n | Orientation vent Sud Est |

3) Save the dataframe a csv file
4) Create a python library called "surf_scrap"(the user can import with the standard python import command. In this library there will be only one function. This function will allow the user to select a specific url (ex : https://www.surf-report.com/meteo-surf/carcans-plage-s1013.html; https://www.surf-report.com/meteo-surf/moliets-plage-centrale-s102799.html). and save the related dataframe in csv file in the location, he/she wants)
5) Create a python (.py) script allowing to execute the library.

Observation : you can ask for a piece of code allowing to extract a variable from me. It will cost 2 points by variable.


**Step 2 : Dashboard creation**

Congratulations, you should have got the data allowing you to create the dashboard. Now let's build it.
You will do it using the flexdashboard package in R. This dashboard will have the following features :

- Run the Python script
- Import the dataset just saved after the script run
- Prepare the data allowing to compute KPIs
- Contain only one tab

Your client want the following KPIs in the dashboard :
- One graph containing the wave size (let's take the mean) over the time
- One graph containing the wind speed over the time
- A table containing the day, the hour, wave size and direction
- A box containing the best moment to practice surf during the coming week

- A box containing the highest wave of the week
- A gauge containing a grade of the sea quality for the best moment to practice.

For this last KPI, you can create your own scale.

We assume that a good wave is up to 1.0m. A wind speed up to 50km/h is ok and should coming from the North (presence of the word "Nord" in the wind direction variable is ok)

Find the best organization for your dashboard. It should be as userfriendly as possible !

BONUS : provide a clear documentation of your solution

Good Luck !