Final assignment

Due date: 31 January 2025. Send it to this email address.

What do I do?

1. Find a messy dataset with at least 10 variables (features), and >1K observations (instances).

What do I mean by messy dataset? Look at the one we used for **Lab 2**. It doesn't have to be as messy and full of biases and missing data, but I'm expecting you to do a bit of data cleaning, to think about imputation, etc.

2. Have a quick look at the data.

Just check the variables and if there is missing data, to be sure than you can use it for this assignment.

3. Find some question that you find interesting to answer and do a data analysis plan

In the class I have mostly covered situations where your dependent variable is in your data (so, supervised Machine Learning/regression analysis). You could do clustering instead if you would like.

See **Class 1** as to what you can put in your data analysis plan. I'm just expecting you to write a bit about the dataset **before you start analysing it**: what variables there are, how many observations, what it is about, what question you are trying to answer, what analyses you will do (descriptive, some type of regression?), what kind of results/associations/groups you are expecting to find (i.e. your a priori hypotheses...).

4. Clean the data: check for missing data, aberrant values, think about potential biases if you can

See **Class 2 and 3**. Please detail what problems you have found in your dataset.

5. Perform analyses to answer your initial question (univariate, bivariate, multivariate if needed)

See Lab 1 and 2. Start with a quick, overall description of the distributions of the variables. Then you can start looking at associations. So this progression: univariate analysis (1 variable) > bivariate analysis (2 variables) > multivariate analysis (multiple variables).

6. Interpret your results and do some data viz

See **Class 4 and Lab 3**. You can do a R Shiny if you want (use the skeleton we did for Lab 3). If you manage to make it look good, it can be a good deliverable to show to potential recruiters!

Datasets

Websites where you may find datasets:

- Scraped Airbnb data
- UC Irvine ML repo
- AskAManager Salary Survey
- Financial well-being survey data (great source, don't forget to download the file codebook, on top of the data)
- Data science dojo

Grading

Up to 15 points for the final report. 5 points if you send me the 3 labs (5/3 points per lab).

Final note is out of 20, in the typical French grading system.

Frequently asked questions

Q. Can I use a dataset on which I have previously worked on? (e.g. for another project) **A.** Yes. Just don't use the one from Lab 2, I want to use it as an example.

Q. Can we be in groups? **A.** Yes. Please do.

Q. What will you be grading? **A.** I'm not grading how accurate your model is, its validity, etc. etc.

Q. What format? **A.** Just send me: the data, your code and some written document. Other than that, do what works best for you. You can do everything in a single Quarto/RMarkdown document if you want. Or it can be multiple code files + a pdf file if you want to be tidy or if you don't like Quarto/RMarkdown (good for you). Keep the code in a R file or in Quarto/RMarkdown though, please don't put that in a pdf file (it would drive me crazy).

Q. I'm stuck somewhere. Can I ask you for help? **A.** Yes. I am paid for this.

Q. My question is not on this list. **A.** Send me an email.