Final assignment

Due date: Probably in January, at the latest. Send it to this email address: dimitri.scronias@univ-amu.fr.

Send me your code, your data and a written report.

What should my report look like? What do I put in it?

For the report, you can either:

- Write it on Word/LibreOffice, export it to pdf (LaTeX if you are very fancy)
- Write it in a RMarkdown file

Both are fine. Do whatever you prefer.

If you do not have an idea of a structure, an IMRaD structure always works. This is a typical structure for research papers:

- **Introduction**. What is the question you want to ask? why it is relevant? Can you give some context? Some other work on the same subject?
- **Methodology**. What data will you use? Where does it come from? How was it gathered? How many observations, variables? What variables? What analyses will you do? Please talk about your data cleaning/any modifications you have made to your data here : what did you have to modify, were there missing variables, what did you do with your missing variables, were there weird values, or outliers, etc... Also, please talk about any initial hypotheses you have (e.g. we expect variable X to be positively correlated with Y, or negatively correlated, a non-linear relationship between X and Y, no significant impact of X on Y, higher sales during the holiday period, higher vaccine hesitancy among women, lower life expectancy in poorer neighbourhoods...).
- **Results**. Descriptive statistics, regressions, tables, graphs, etc., go here. Describe these results. At this stage, students usually write about *every single result they find*: it's not necessary. We are able fully capable of reading your tables, you know. Just put your tables, and write about **what is important** in your results.
- **Discussion**. Here, interpret your results. How do your analyses help answer your initial question? If there is other research on the same subject, do you have similar results? Different ones? Was their methodology different? Talk about the limits of your analyses, potential biases.

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** (the one with the emergency data). 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 a question that you find interesting to answer and start laying out the structure of your report.

In the class I have mostly covered situations where your dependent variable is in your data (so, supervised Machine Learning/regression analysis). If you want to try exploratory analyses, you can do clustering instead. It is not that complicated, and there are a lot of resources online. You could try hierarchical clustering

4. Clean the data: check for missing data, aberrant values, think about potential biases from your cleaning.

See **Class 2**.

5. Perform analyses to answer your initial question (univariate, bivariate, multivariate if needed). For IMRaD: this would be your results section.

See Lab 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. Do some data visualisation. Interpret your results. Put data visualisations in the results section. Any interpretation go in the Discussion section.

See **Lab 2 & 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
- Census income dataset This is weighted data. If you use it, please do your descriptive statistics using the MARSUPWRT as a weighting variable. You can use the survey package plus srvyr for this. Use the function `as_survey_design(weights = MARSUPWRT)`
- Airbnb listings
- Mass mobilization data, you can combine it with the democracy ratings dataset and the HDI data

Grading

Up to 10 points for the final report. Up to 10 points for the oral presentation.

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. Up to 5 persons per group.

Q. What will you be grading? **A.** I'm mainly grading how you tackled your research question, your reasoning in your data cleaning, your interpretations of the models, the variety of analyses. Don't hesitate to explain what you did. I won't be grading how accurate your model is, its validity, the quality of your results.

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.