Fundamentals of Data Science in Business and Engineering Project

Goal. The goal of this project is to give you the opportunity to explore, analyse and derive conclusions about some real dataset of your own choosing. You should demonstrate proficiency in the techniques we have covered in this class by applying them to the chosen dataset in a meaningful way.

Part 1: Questions. Come up with a theme that you care about and that you wish to address. It should contain at least three interesting questions that your analysis seek to resolve with significant response impact. What questions and/or concerns do you have about your dataset and your project?

Part 2: Data. You may gather the data yourself using a survey or by conducting an experiment, or use another data source. You can choose the data based on your interests or based on work in other courses or research projects. You will submit your data. Include in your writeup project informations about the source of your data, as much as you know about the method of collection, and any concerns you have for the data in terms of bias, etc. Be clear about what variables were measured and how they were measured.

In order for you to have the greatest chance of success with this project it is important that you choose a manageable dataset. This means that the data should be readily accessible and large enough that multiple relationships can be explored. As such, your dataset must have at least 100 observations and between 4 to 10 variables (you should include categorical variables, discrete numerical variables, and continuous numerical variables).

Do not reuse datasets used in examples, homework assignments, or labs in the class.

The list below of data repositories might be of interest to browse. You are not limited to these resources, and in fact you are encouraged to go beyond them.

Kaggle datasets	OpenML	Awesome Public Datasets
data.europa.eu	World Bank Open Data	HealthData.gov
TidyTuesday	NHS Scotland Open Data	UK Gov Data

Part 3: Exploratory Data Analysis. You should present numerical and graphical information that summarises the data. For example, you may wish to report appropriate sample means or standard deviations, or present graphs like histograms or scatterplots. Be very clear about what the information that you include is representing, and carefully label your graphs. The goal here is to get a handle on some basic features of the data set before you delve into the proper data analysis.

Part 4: Data Analysis. Perform an appropriate analysis of your data using the methods learned in class. How exactly this looks will vary widely depending on the kind of data you have and what questions you would like to answer. The goal is not to do an exhaustive data analysis i.e., do not calculate every statistic and procedure you have learned for every variable, but rather demonstrate proficiency at asking meaningful questions and answering them, interpreting and presenting the results of data analysis. Focus on methods that help you begin to answer your research questions. Also, critique your own methods and provide suggestions for improving your analysis. Issues pertaining to the appropriateness of the statistical analysis should be discussed here.

Some examples:

- **Single Variable Exploration**: Examine a single variable, understand its underlying distribution, and choose the appropriate summary statistics for it.
- **Pair-Wise Exploration**: Identify possible relationships between variables and compute correlations and linear fits.
- **Estimation and Hypothesis Testing**: Think of a hypothesis you want to prove or disprove, and then think of the data you need to retrieve to answer it and the tools you need to properly answer your question.
- **Resampling and Bootstrap**: Find standard errors on your estimators or confidence intervals for unknown parameters, in case the standard methods are not applicable (e.g. data not normal).
- **Bayesian Inference**: Justify any prior distributions that you use, with the goal of making them acceptable to a skeptical audience.
- **Visualization:** You should create some kind of compelling visualization(s) of this data. You do not need to visualize all of the data. A single high quality visualization will count more than a large number of poor quality visualizations.
- **Learning from data**: Reduce dimensionality while preserving information, cluster data and discover patterns.

The project is very open ended. There is no limit on what tools or packages you may use. Use data visualization as a tool for examining data and communicating results. Also pay attention to your presentation. Neatness, coherency, and clarity will count.

Part 5: Results. Conclude your project with a clear summary of the results of your analysis, and interpret your analysis in the context of the theme that you wished to address.

Output. The project will be submitted to the Moodle webpage of the course and will contain the following files:

name_report.pdf	Your written report (3000+ words) in PDF format
name_presentation.pdf	Your presentation slides in PDF format
name_data.csv	Your dataset in CSV format and your data dictionary
name_code.extension	Your code[1]

[1]The code should be clean, with supporting comments. This code needs to be reproducible.

Presentation. You may team up for a common project (at most two members per project). You will present your project to the class. There is no limit to how many slides you can use, just a time limit (10 minutes total). Each team member should get a chance to say something substantial during the presentation. Your presentation should not just be an account of everything you tried ("then we did this, then we did this, etc."), instead it should convey what choices you made, and why, and what you found.

The presentations will be planned for the last two classes of the semester (14.01.2025 and 21.01.2025). You will indicate your preferred date and time slot in this Google Sheet, on a first-come, first-served basis. Regardless of your choice, you will have to attend all other presentations (both dates) and to provide feedback in the form of peer evaluations.

Peer evaluation. You will be asked to fill out a survey where you rate your colleagues' projects out of 10 points. Filling out the survey is a prerequisite for getting credit on the individual project. If your peer evaluation is below 2 points, please provide some explanation.

Grading.

Total	30 pts
Project report & presentation	20 pts
Classmates' peer evaluation	10 pts