Data Literacy 2022 Project Guideline

High-level project expectations:
The goal of the project is to go through the process of a "data science" project, starting with a concrete question that can plausibly be tackled (within the limited time available for the project), choosing and implementing appropriate analysis techniques, and communicating your results in a clear way through visualizations and in writing, while acknowledging the limitations of your results. This experience is meant to prepare you for independent research.

There is no expectation that projects should involve any *particular* technique that was covered in the course (correlations, hypothesis tests, etc.), and you are free to use methods not covered in lectures, but the key criterion is that your analyses are **appropriate for your question**, not that they are fancy. This is not an exam to test your understanding of all concepts covered in class, but aims to teach you to apply the right tools in the right situation. We will evaluate the quality of your work and thought process, **not on the outcome of your analysis** (i.e., we don't care whether your hypothesis was wrong or if you had null results).

Specifically, we want you to show that you can …

1. **motivate and design a well-defined research question:** Is the motivation behind the question clearly stated, with a well-defined goal / question that can be addressed through analysis of real data? Is this within-scope for a 4 week project?
2. **collect or identify (and practically handle/curate) an appropriate dataset:** Was your data-collection process appropriate, and is the dataset suitable for addressing your proposed question? Do you have contingency plans if the data collection fails?
3. **identify suitable methods to answer your questions, and clearly describe your analysis:** Are your analysis methods appropriate for addressing your question (e.g., regression, classification), with the kind of data you have (e.g., continuous, categorical)?
4. **be aware of limitations of your data, analysis choices, and results**: Given the time-constraints of the project, what are the limitations / weaknesses, and what could you do to improve it given more time? Note that it's perfectly acceptable to discuss limitations even if you cannot address them given the time constraint.
5. **clearly describe in a written report the result of your investigations:** Does your report have the expected structure (abstract, introduction, methods, results, discussions /limitations)? Does it contain the necessary references? Is the writing clear and free of low-level errors (spelling mistakes, etc.)? One resource for scientific writing is: https://de.coursera.org/lecture/sciwrite/5-8-abstract-ISxVk
6. **clearly visualize the result of your analysis:** do your plots have the necessary components (legends, figure axes, etc.)? Do they convey information correctly and efficiently?

Code: submit your final code as a zip file that contains the entirety of your codebase, and you can optionally include a link to your project GitHub repo in the report itself. Your code will typically not be graded, but we will check it in case we suspect that an analysis is incorrect, or

even plagiarism, e.g., entire code chunks / notebook copied from Kaggle / GitHub repos with no modification.
Your project grade will be informed by how well you accomplish the above points. To help orient you through the process, we provided the table below so you can evaluate yourselves on whether you have accomplished these goals at a level below/at/above expectation.

Project Components	Below Expectation	Meets Expectation	Exceeds Expectation
Research Question & Motivation	Missing, vague, intractable/unfeasible within 4 weeks	Explicit, well-defined goal, tractable/feasible within 4 weeks	An original, interesting and ideally relevant (yet tractable) question/hypothesis
Dataset	Too small, dubious source, not appropriate for research question	Appropriate size, legitimate, contingency plan if data collection involved	Exceptional effort/time spent in collecting or preprocessing dataset
Analysis/methods	Wrong choice, wrong usage or implementation	Choice of suitable methods, analyses	Exceptionally careful design of analysis, possibly even using complementary approaches
Acknowledgement of limitations	No discussion of results and limitations; did not acknowledge clear biases / weaknesses	Interpretation of analyses with respect to research question, respective limitations	Propose ways of addressing raised limitations in data collection and/or analysis
Clarity of written report	Not using an appropriate structure (abstract, introduction, methods, results, discussion), unclear language, lack of precision, lack of references.	Clear structure (likely into 5 sections), clear and concise description of question, analyses, and results, no typos or errors, includes appropriate references	
Visualizations	Small fonts, no legends, too few or many colors, generally unclear plots	All plots correctly labeled, including legends, appropriate choice of colors	
Code	Not shared	Shared	