Final Project Instruction:

The final project should be completed **in teams of 2-3 members**. Solo projects are not permitted and will result in an automatic 10% deduction from the final project grade. You may use any Python library for creating your visualizations, Alternatively, you are allowed to use R instead of python, however it is expected that final report submitted as a clean, well-organized PDF document. You are also need to submit your notebook, or markdown file containing inline codes separately. Your final report is due by Tuesday, **May 7, 2023, 11:59 pm.** These two documents will be graded jointly, so they must be consistent (i.e., don't change the Jupiter notebook file without also updating the PDF document!).

All results presented *must* have corresponding code. **Any answers/results given without the corresponding code that generated the result will be considered absent.** All code reported in your final project document should work properly. Please do not include any extraneous code or code which produces error messages. (Code which produces warnings is acceptable, as long as you understand what the warnings mean.)

Each team is also required to prepare a 5-6 minutes presentation that will showcase your datasets, questions, and plots. The presentations are scheduled for **May 1st (for the 591 section) and May 3rd (for the 491 section)**. Providing peer feedback to other teams is a crucial part of this exercise, hence attendance at both sessions is mandatory for all students.

[**491 section**] For your final project, you will be choosing your own dataset, given the following constraints: Pick one of the datasets published after January 1, 2023. You may use UCI repository, or Tidy Tuesday project to find a proper dataset.

[**591 section**] I encourage you to use a dataset related to your research, but you are also allowed to choose your dataset based on 491 criteria.

The final project report should be structured as follows:

- Introduction (1–2 paragraphs, describe the dataset you picked)
- Questions (5 questions you will answer)
- Approach (2–3 paragraphs)
- Analysis (code blocks, figures or computed tables, text/code comments as needed)
- Discussion (1–3 paragraphs)

I encourage you to be concise. A paragraph should typically not be longer than 5 sentences.

Important:

Your project should encompass five distinct categories or topics from the following list:

1. Visualizing Amount
2. Visualizing Distributions
3. Visualizing Proportions
4. Visualizing Quantitative Variables
5. Dimension Reduction (You may present two plots in this category if employing two different dimension reduction techniques)
6. Clustering (You may present two plots in this category if employing two different clustering techniques)
7. Visualizing Time Series
8. Visualizing Trends
9. Visualizing Geospatial Data (Two plots are permissible in this category)

- Be aware of **uncertainty** in your analysis and explain in your analysis how you address that (using confidence interval, error bar, …).
- You are encouraged to produce one figure to answer each question, but it is ok if you need to have more than one plot (or subplots) to answer one question.
- Throughout your analysis, be mindful of **Junk Charts Trifecta Checkup,** which assesses the alignment between the data, the research question, and the visual form of the analysis. Continuously ask yourself if these three elements are well-aligned and if the visualization effectively communicates your findings.
- You are encouraged to visit office hours or send an email to the instructor to help develop your idea.

Project Report

In your **Introduction** section, please provide a concise overview of the dataset, including its context. Discuss what makes this dataset, or its broader topic, particularly intriguing to you, and/or how it ties into your existing research area (591 only). Consider this project as an independent entity, assuming the reader, including the grader, has no previous knowledge of the dataset or its context. Additionally, outline the specific variables within the dataset that you intend to analyze.This foundational information will set the stage for the subsequent sections of your report.

Next, you will formulate your **questions**, which are the most critical part of your project. You might perform some exploratory analysis before refining your questions.

The questions should be conceptual and open-ended, not prompt a specific analysis. Aim to choose questions that captivate both you and your audience. Thus, questions that are too simplistic and fail to spark curiosity will not be effective.

In particular, make sure you understand the difference between a question and an instruction. For example:

A legitimate question is: "How has the weight distribution of alpine skiers changed over the years?" This encourages a broad examination.

Conversely, an instruction would be: "Create a series of boxplots showing the weight of alpine skiers across different Olympic years." This is not an open-ended question.

This is a question that prompts a specific analysis; it is actually an instruction pretending to be a question: "What is the slope value in a regression of skier weight against the year?"

In the **Approach** section, describe what type of data wrangling and pre-processing steps you will perform and what kind of plot(s) you will generate to address your questions. Provide a clear explanation as to why these plots are best for providing the information you are asking about. (You may use “Chapter 5. Directory of Visualizations” for guidance.)

In the **Analysis** section, include the code that is used for data wrangling and generating your plots. It is important to provide enough detail in the code so that the reader can understand the steps that are being taken and how the plots are generated.

When generating plots, ensure that they are clear and easy to understand without the need for extra information (self-explanatory). Pay close attention to the details of your plots, including the labeling of axes, color choices, and any other relevant details that might impact the visual form of the plots.

In **summary**, the Analysis section should include clear and detailed code that is easy to follow, along with explanatory plots that are easy to interpret without additional context. In the Discussion section, interpret the results of your analysis. Identify any trends revealed (or not revealed) by your analysis. Speculate about why the data looks the way it does.

Final Project Grading Rubric

Each aspect of the project will be graded on a competency-based scale ranging from Exceeds expectations to Meets expectations, Needs improvement, Unsatisfactory, and Failing.

	Exceeds expectations	Meets expectations	Needs improvement	Unsatisfactory	Failing
Introduction	The introduction provides a clear explanation of the questions and the dataset used to answer the questions, including a description of all relevant variables in the dataset. **10pts**	Introduction has minor flaws, e.g. is too short or too long. **8pts**	Introduction has one major flaw, such as not describing the relevant variables in the dataset. **6pts**	Introduction has multiple major flaws. **4pts**	Entirely incorrect/not attempted. **0pts**
Justification of approach	The chosen analysis approach and visualizations are clearly explained and justified. **10pts**	Justification of approach has minor flaws, e.g. is too short or too long. **8pts**	Justification of approach has one major flaw, e.g. a visualization is not justified or is incorrect. **6pts**	Justification of approach has multiple major flaws. **4pts**	Entirely incorrect/not attempted. **0pts**
Questions	Questions are interesting, appropriate for the dataset, and conceptual. **10pts**	Questions are conceptual and appropriate for the dataset but lacks depth or insight. **8pts**	Questions are either overly technical or vague, or prompts one particular analysis. **6pts**	Questions are not a question or does not relate to the given dataset. **4pts**	Not attempted, or, using an invalid dataset (outside the specified date boundaries). **0pts**
Code	Code is correct, easy to read, properly formatted, and properly documented. **10pts**	Code is correct but has minor problems with formatting or documentation. **8pts**	Code has minor flaws, is missing data wrangling component, or includes extraneous parts. **6pts**	Code has major flaws. **4pts**	Entirely incorrect/not attempted. **0pts**
Analysis	Analysis approach is appropriate for the given question and dataset. **15pts**	Analysis approach is mostly appropriate but has some minor issues. **12pts**	Analysis approach has one major problem. **9pts**	Analysis approach has multiple problems. **6pts**	Entirely incorrect/not attempted, or, required analysis components as listed in the instructions (PCA, clustering, etc.) are not included. **0pts**
Visualization	The visualizations are appropriate, easy to read, properly labeled, and nicely styled. **15pts**	The visualizations have minor flaws, such as with legibility or labeling, or the chosen geom is suboptimal, or there are minor problems with styling. **12pts**	The visualizations have substantial flaws with legibility or labeling, or are confusing, or have not been styled. **9pts**	The visualizations have major flaws, i.e., are barely comprehensible or entirely inappropriate. **6pts**	Entirely incorrect/not attempted. **0pts**
Discussion of results	Discussion of results is clear and correct, and it has some depth without begin excessively long. **10pts**	Discussion of results is mostly clear and correct, but has minor inaccuracies or lacks some depth. **8pts**	Discussion has one substantial flaw but is otherwise acceptable. **6pts**	Discussion has multiple flaws in logic. **4pts**	Entirely incorrect/not attempted. **0pts**
Reproducibility	All required files are provided. PDF and code. The code parts run without any error **10pts**	All required files except pdf are provided. **8pts**	Code requires minor modification to run, or key datafile is missing. **6pts**	Code requires major modification to run error-free, or is not provided. **4pts**	Not attempted, or, using an invalid dataset (outside the specified date boundaries). **0pts**
Presentation	Entire document is well structured and easy to follow. No extraneous materials. **10pts**	Document is mostly well structured, but some aspects are confusing or difficult to follow. **8pts**	Document has several deficiencies, such as excessive extraneous materials, misplaced figures, code, or text, or is otherwise confusing. **6pts**	Document is near impossible to comprehend. **4pts**	Incomprehensible/not attempted. **0pts**
Points total:	**100**	**80**	**60**	**40**	**0**