

First Project Guidelines Supplements

ECO225, Professor Khazra

Please note: The lecture notes, project guidelines, and Jupyter notebooks *do not cover* every topic discussed in class. Attendance and active participation are therefore essential. If a concept, requirement, or convention is explained during lecture but does not explicitly appear in the written guidelines (for example, how to report summary statistics or how to interpret plots), it will still be considered part of the course material and will be graded accordingly. Due to the nature of an in-person course, it is not possible to document every detail discussed in class. As a result, missing lectures may negatively affect your project grades if you miss important clarifications or expectations.

Please start working on your first project as early as possible. You will likely encounter practical issues, such as downloading data, working with Jupyter Notebook, or converting your notebook to a PDF. These problems are difficult to resolve at the last minute. If you begin early, you can use collaboration hours and Piazza to ask questions and receive help from TAs and classmates before the deadline.

IMPORTANT: Do not include regressions, maps, machine learning models, or any methods or topics beyond the scope of this project. Topics scheduled for future weeks will be required in later projects but should not appear here. Including such material will disrupt the grading process. Any additional sections or topics outside the scope of this project will not be graded and may result in a lower score on the required components.

To receive useful feedback and a higher mark, focus on the material covered in this project. You are encouraged to go beyond the minimum requirements *within each category* (for example, through clearer economic interpretation, better variable construction, or more informative visualizations), but you should not introduce new methods such as regressions at this stage.

The word limit for this project is 1500 words and is strictly enforced. Concise writing is expected. Exceeding the word limit may result in a penalty of up to 20%.

Fully AI-generated content that lacks verification, contains citation errors, makes claims unsupported by the data, or shows internal inconsistencies may receive penalties of up to 100%.

Formatting

You will lose at least **5 points** if your submitted project does not meet any of the following requirements:

1. Submitted projects must be in PDF format. Any other format, including a .ipynb file, will not be accepted. Submissions in any other format receive a zero.
2. Use clear and concise language. Avoid unnecessary jargon and overly complex sentences. AI-generated text that uses sophisticated language without clear economic meaning, contains unverified claims, or includes inaccurate citations may receive penalties of up to 100%.
3. The project must have clear and informative section titles (e.g. Introduction, Summary Statistics, Plots, Conclusion).
4. Choose a meaningful title for your project. The title should reflect your research question; the question itself may serve as the title. Talk to the professor about this.
5. You must write your project in a Jupyter Notebook (Python) and submit it in PDF format. If you have difficulty converting your notebook to PDF, first export it as HTML and then save or print the HTML file as a PDF.

Please read the following carefully.

- If you include a graph or a table, you must **explain what you learn from it**. Outputs without a clear and valid interpretation will not be graded.
- The final product should be a clean, coherent, and well-organized report. Re-read your text to ensure internal consistency.
- Do not include unnecessary code, intermediate outputs, or error messages. This includes package installation messages and debugging outputs. Including such material may result in penalties of up to 50 points.
- Tables and summaries or data outputs that extend beyond one page of the PDF will receive a 10-point penalty each.
- Every graph, table, or summary statistic must be motivated: explain why it is included, what it shows, and how it relates to your research question.

IMPORTANT: You must show your code and results and provide clear economic explanations for all parts. We will evaluate the quality, clarity, and accuracy of your analysis. Simply completing all required steps does not guarantee a full mark. These are minimum requirements; higher marks require deeper interpretation and more thoughtful analysis *within the scope of this project*.

Mandatory Section Titles in Your First Project

You can (and should) add more subsections and sections when needed, but all projects must have the following “main” section titles in the same order. We will keep updating this list for future projects.

1. Introduction

2. Data Cleaning/Loading
3. Summary Statistics Tables
4. Plots, Histograms, Figures
5. Conclusion

Coding and interpretation

In this report, you will learn how to load your dataset, clean it, and start building an economic story with data. You will also merge your data with at least one outside dataset (more details below).

Throughout this course, our goal is to study how an explanatory variable (X) relates to an outcome (Y). Strong economic projects also explain **why** X might affect Y through a mechanism or channel (Z). For example, rather than only asking whether education (X) is related to income (Y), you might ask whether education affects income *through* job networks or parental resources (Z). **A key requirement is that your mechanism (Z) must be measurable (or at least proxied) in the data you use.**

Important: This first project is descriptive and exploratory. You are not allowed to run regressions yet. Here, “relation” means understanding patterns in the data using cleaning, summary statistics, and visualizations.

1. Please write a title for your project in title format. Your title can be your research question. “Housing Prices in California” is **NOT** a title. A better title would clearly state X , Y , and (if possible) the channel Z (for example, “Do zoning restrictions (X) affect rents (Y) through housing supply (Z)?”). Please talk to your TAs about choosing an economic research question and title.
2. **(5 points) Write a brief introduction of your project (two to four paragraphs)**, including the source of the data, the literature review, and the key background needed to understand your question (keep it short; you will improve it over time). An outsider should be able to understand what you are trying to do in this project. Your introduction must:
 - state your research question in terms of X , Y , and the mechanism Z ,
 - name your dataset(s) and their source(s),
 - briefly summarize what you found in this project (based on summary tables and plots).

Include a short literature review in your introduction that reads like a story. To help you write a story, your literature review should do **at least two** of the following: (1) explain what the main papers agree on, (2) explain what they disagree on, (3) explain what mechanism(s) (Z) they emphasize, (4) explain what is missing that your descriptive analysis can begin to explore. You should have *at least* five citations, with one of them as your main paper (the closest paper to your work). All citations must appear in your references section in APA or Chicago style.

IMPORTANT: All writing must be in **markdown cells**, and all code must be in code cells, similar to the lecture notebooks. *Writing text in code cells will not be graded as writing.*

3. (30 points) Setting up the project.

- (a) Create a new Jupyter Notebook.
- (b) (5 points) Choose a title for your project and place it at the top of your report (bold and larger font). We will grade your research question and choice of title. Please consult Matthew and Sobia about your title.
- (c) (10 points) Choose an outside dataset to merge with your assigned dataset. The outside dataset should help you measure or proxy an important part of your story (often the mechanism Z) and must add at least **3 new variables** to your final dataset. For example, if you are working with a crime dataset and your story relates crime to mental health services, you might need external data on mental health providers by area. Ask us where/whether you can find it. Download and merge it with your original dataset.

Important: other CSVs in your Kaggle page *do not* count as outside datasets, they should have already been merged into your main database. However, you can use list I provided or other sources of your choice.

* Include a subsection such as “Other datasets” and make it clear what the outside dataset is and where you downloaded it from. Be clear about (1) temporal coverage (daily, monthly, annual) (2) geographical coverage (block, zip code, city, county, state, etc) (3) source of the data (4) the variables you extract and merge with your original dataset.

- (d) Choose one Y variable (outcome) and **at least five** X variables (explanatory variables) can be from either the main or the new datasets you merged. If your dataset allows, choose **up to ten** X variables for stronger analysis.

This choice cannot be random. Your chosen X variables should be motivated by your economic story (and ideally connect to your mechanism Z ; X affects Y through Z). Feel free to get help from your TAs and me.

- (e) (15 points) Clearly explain these variables in paragraph form (no bullet points): why you picked them, what each variable measures, and how they relate to your research question. You should explain why you expect X to be related to Y , and where the mechanism Z fits in your story. **There is a word limit, so write efficiently.**

IMPORTANT: Longitude, latitude, zip code, county code, state name, country name, and other *location variables* do **not** count toward the minimum requirement of five X variables. These variables are geographic identifiers rather than economic covariates and do not represent mechanisms or explanatory factors in your research question.

You should **keep** location variables in your dataset, as they will be used later in the course for mapping and, in future projects, as part of more advanced analyses. However, for this project, they are not considered X variables and *should not be included in your summary statistics tables*. We will use them for mapping and grouping.

4. (35 points) Data cleaning and summaries.

Load your data in your Jupyter Notebook and convert it to a pandas DataFrame. This section is meant to help you understand the structure of your data and assess whether it is informative for your research question.

- (a) Show summary statistics for your chosen X variables and your Y variable. These must be the same variables you explained earlier. Five X variables is the minimum requirement. For a full mark, you are encouraged to work with up to ten X variables, provided they are economically meaningful and relevant to your research question.
- (b) Summary statistics tables must be created using the `stargazer` library and should not be raw outputs from `summary` or `describe` commands. Use `display(HTML(stargazer.render_html()))` to present your table in the notebook and PDF.
- (c) Explain and interpret the summary statistics table in paragraph form. Do not simply restate the numbers. Your interpretation should address questions such as: What do these values tell you about the scale, variation, and distribution of your variables? Are these patterns consistent with your economic story and the proposed mechanism (Z)?
- (d) Create at least one additional summary table based on subgroups of the data. This may involve subsetting the sample or creating indicator (dummy) variables. The choice of subgroups should be motivated by your research question or by the mechanism (Z) you aim to explore.

Please note: (1) We are looking for thoughtful tables that help answer the research question or clarify the underlying economic story. Including many unrelated or repetitive tables will not increase your grade and may result in a formatting penalty. (2) Including location variables, string variables, or identifiers (such as IDs, zip codes, longitude/latitude, or names) in summary tables will result in a penalty. If you are unsure whether a variable is appropriate, ask before submitting.

5. **(25 points) Create at least four plots, figures, or histograms for a passing mark.** Creating four plots is the minimum requirement. Higher marks require a broader set of well-chosen visualizations that reflect different economic intuitions. You must use different types of plots for different purposes; repeating the same type of visualization across variables is not acceptable. For example, creating four bar plots for four different variables does not count as four distinct figures and may be penalized for redundancy. Each visualization should serve a different analytical purpose.

The choice of which variable to visualize and how to visualize it should be deliberate and motivated by your research question and economic story.

The quality, presentation, and interpretation of your plots will be graded. For each figure, you must explain: (1) why you chose this variable, (2) what economic question the plot is addressing, and (3) what you learn from the observed pattern. Plots or tables presented without interpretation will not be graded.

IMPORTANT: For example, if you are working with housing price data, creating simple histograms of price, number of bedrooms, and income will **not** receive a passing mark. While these plots are easy to produce, they are typically uninformative for an economic research question. Your figures should help explore how X relates to Y , and when relevant, how this relationship may operate through a mechanism (Z). If you are unsure whether your plotting strategy is appropriate, discuss your ideas with the instructor or TAs at least one week before the deadline.

- When deciding which variables to visualize, consider constructing economically meaningful measures such as ratios (e.g. crime per capita), differences (e.g. wage gaps), or other transformations that help clarify the underlying mechanism or economic story.
- Plot the distribution of your main X variable and your Y variable. If a variable is continuous, a histogram is appropriate; if it is discrete, consider a boxplot instead. All plots must have clearly labeled axes and informative titles (this is graded). You are encouraged to create new X variables when appropriate, provided they are clearly defined and motivated by your research question.
- Plot the relationship between Y and different X variables separately, with Y on the vertical axis and X on the horizontal axis. Each plot should highlight a distinct economic relationship and must include axis labels and a title (this is graded).
- Use `groupby` or similar grouping methods to create additional plots that compare subgroups of the data. For example, you might compare outcomes across age groups, income groups, education levels, or other relevant categories. These plots should help clarify heterogeneity or provide insight into the potential mechanism (Z).
- Use different visualization types for different questions. The choice of visualization should be appropriate for the variable type and the economic question being addressed.
- **IMPORTANT:** Do not include maps, even if they appear relevant to your project. Mapping and spatial analysis are reserved for Project 2.

6. **(5 points) Conclusion.** Write a brief conclusion in clear and scientific language. The minimum length for the conclusion in this project is *one paragraph*. Your conclusion should synthesize the main findings from your summary statistics and visualizations rather than restating them. In particular, you should summarize what you learned about the relationship between X and Y and whether the descriptive evidence is consistent with the proposed mechanism (Z). You may also briefly note any limitations of the data or questions that remain unanswered at this stage. The conclusion will be expanded and refined in future projects. To guide your writing, you should read conclusions from recent papers in your field and pay attention to how authors summarize evidence without overclaiming.
7. **(Not graded but required) Version control.** Consistently push your code to the assignment repository provided and commit your work regularly. The repository should reflect the development of your project over time, including data loading, cleaning, and analysis steps. While this component is not graded, failure to maintain a repository may result in follow-up questions during grading.

Upload your Jupyter Notebook (your code and explanation) in pdf format on Crowdmark for marking.

We look forward to reading your great work! Good luck!