

Second Project Guidelines Supplements

ECO225, Professor Khazra

Please note: It is important to understand that the lecture notes, project guidelines, and Jupyter notebooks *do not cover* all the topics discussed during class. *If students miss a class, they must catch up on the material.* If something is mentioned in class but not included in the project guidelines, such as but not limited to the requirement to report summary statistics tables in a certain way or the language in plot interpretations, it will still be considered and will be graded. Due to the nature of this in-person course, it is impossible to include every lecture detail in the project guidelines and notebooks. Overall, skipping lectures can negatively impact your project grades as you may not receive critical information about the assignments and other subjects.

Please start working on your first project ASAP. You will face problems downloading your data, working with Spyder or Jupyter Notebook, uploading your PDF, etc. You cannot solve them the day before or even a few days before the deadline. During collaboration hours and on Piazza, TAs and other students can help you fix your errors if you start early.

IMPORTANT: Please do not include ML models, or any elements beyond the scope of this project's requirements. Introducing topics for future weeks and mandatory for upcoming projects will disrupt the grading process. Additionally, any extra sections/topics will not be graded, potentially resulting in a lower grade for the essential components of the project. In short, to receive a higher mark and useful feedback, stick to the topics covered in each project. You can go beyond in each category but do not introduce new topics just yet.

The word limit for this project is 1500 new words and is strictly enforced. Less is better, more is not.. You have 1500 word count for P1 and 1500 extra for P2, summing to 3000. Exceeding the limit may result in a penalty of up to 20%. To avoid this, ensure your writing is concise. The word count includes all prose text (introduction, interpretations, conclusions) but excludes code, tables, regression outputs, and references.

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. Any other format, including a .ipynb file, is not 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 projects should have clear “section titles” (e.g. Introduction, Summary Statistics, Summary, Future Steps).
4. Choose a title for your project. What is the question that you would like to answer using this dataset? The question can be your title.
5. You should write your project in Jupyter Notebook (Python) and submit it in pdf format. If you have problems converting your Notebook to PDF, first download it as HTML, then print/save the HTML version in PDF.
6. This is an individual project. However, you are encouraged to check the projects on the Kaggle.com website that use similar data. We have provided some useful links on the data list. You can use these sources, but the coding and explanations **must be yours**. Do not copy and paste the same chunk of code into your project.

Please read the following carefully.

- Please note that if you include a graph or a table, you should **explain what you learn from it**. We do not mark an output not followed by a valid interpretation.
- The final product should be a clean and comprehensive project. Re-read your text to ensure consistency.
- Do not include unnecessary chunks of codes or outputs and errors. You may be penalized for this by up to 50 points. No unnecessary code and errors (including package installation messages) should be printed in your submission PDF.
- Any graph, summary, or output should have an explanation. Why do you include it in your project, and what do you understand? How is it related to your research question?

IMPORTANT: Show your code and results, and provide economic explanations for your results for all parts. We will evaluate your work’s quality and accuracy; simply answering all questions does not guarantee a full mark. These are the minimums; try to go beyond the instructions.

- “Going beyond” means deeper analysis, clearer economic reasoning, better visualization design, or novel ideas (novel variables/data/mechanism).
- All interpretations must explicitly reference variable names, units, sample restrictions, or transformations used in your code/tables.

Mandatory Section Titles in Your Second 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.

Your second project should include your first project. You will have **TWO** main sections for your project (and some subsections that you title yourself). The first main title should be “Project One,” the second main title should be “Project Two.” You can add titles by adding the “#” sign before your title in your markdown cell; check this link for more details on titling: <https://www.datacamp.com/community/tutorials/markdown-in-jupyter-notebook>

1. Project One (Use # in a Markdown Cell to set this as a main section)
 - 1.1 **Introduction** (use ## in Markdown cells to set the following as the subsections)
 - 1.2 Data Cleaning/Loading
 - 1.2.1 you can have a subsection for new dataset(s) that you merge to explain them separately
 - 1.3 Summary Statistics Tables
 - 1.4 Plots, Histograms, Figures
2. Project Two (Use # in a Markdown Cell to set this as a main section)
 - 2.1 The Message
 - 2.2 Maps and Interpretations
 - 2.3 Regressions
 - 2.4 **Conclusion** (note that the entire document has one conclusion, not two or three. Same for the introduction)

We will build upon the first project towards a complete academic paper on a hot topic using real-world data and cool Python techniques; just excellent! Here’s what we are going to do for this project.

Part One

(10 points) Incorporate the comments that you have received in your second project. This is the first part of your third project, which will be graded again. We will check the first section of your projects (your updated previous projects) and grade it again. If you have not received any comments, try to improve your previous project by doing a better job on the introduction or literature review, fine-tuning your visualizations, or adding more meaningful analysis. You will always revisit your past projects until the final project (your masterpiece). In your final project, project 1 and project 2 feedback will be revisited. We need to see significant improvements in your previous projects for this part.

Part Two

From the visualization lecture we learned that there are 4 steps in creating an informative visualization:

1. Identify the message.
2. Describe your visualization.
3. Create a draft of the visualization (and verify your data!).
4. Fine-tune the visualization details.

Your main message plot cannot be a map.

1. **(10 points)** For this part, I want one to two sentences (main message) and a plot (the main message plot). Complicated and uninformative AI/ChatGPT generated main messages receive 0 for this part.

Your message must be a falsifiable empirical claim, not a description (e.g., "X is higher in Y" is insufficient).

Think about the main message/question of your paper. You still don't know the precise answer to that question, but you can use visualization to give us some intuition about the possible answer(s). Please talk to two different TAs and discuss your main message. They can help you think about your main message but cannot tell you what it should be or whether your message achieves full marks. Please write down your message (title it THE MESSAGE!) and go to step two. **Important:** your message cannot be a paragraph. It should be one or two sentences. You can add explanations after but you should be able to articulate your main message in a short format.

2. Use pen and paper to draw what you have in mind. You do *not* need to include your drawing in your project submission but this is a necessary step.
 - Can you plot your Y (dependent variable) by subgroups? Examples: by gender, type (whatever type makes sense for your project), population (create a variable high/low pop), developing or developed countries, characteristics of the firm, income (create a column high/low), the share of minorities, and so on.
 - Notice that you can create histograms, scatter plots, line plots, etc.

What is on the Y axes? What is on the X-axis? What colors are you choosing? What line style will you choose if it is a line plot? Think about all the details and draw them on paper.

3. Code and create the plot that you drew on paper. Please explain what you see in the visualization and how it relates to your main message.
4. **Important:** your main message plot cannot include multiple plots with different variables. But if you want to have two figures with the same X and Y variables, it can

be possible depending on your message. In this exercise, you are practicing to summarize information and visualize as efficiently as possible. Creating four figures for the main message is not what I have in mind.

5. Make sure to fine-tune all graphs in your **first** projects. Please check the fine tuning section in the maps lecture for more details.
6. You can use one of the graphs that you have created in your project one, **however**, it should satisfy all of the above. I recommend creating a new plot that is more thoughtful and relevant to the above instructions.

Part Three

For this part, you will complete and fine-tune your maps from the maps lecture exercise. This is now part of the pdf that you will turn in and should be the last section of your paper (before the conclusion). The guidelines below are very similar to what you saw for the exercise. **(35 points)**

- Create at least **three** *static* maps. One for your outcome variable (Y) and one for your most important independent variable (X). Your X or Y can be per-capita. And the third map can be for a unique variable/ratio that you create.
- If your X or Y variable represents a count (e.g., crime count), divide it by the population to ensure you're mapping the rate rather than just reflecting population density.
- **Important:** Creating more than three maps that are repetitive or uninformative will negatively impact your grade. For example, if you map all of your X variables without a clear and significant contribution to your research question, you may lose up to 20 points for including irrelevant output. Think carefully before duplicating!

Typical strong submissions have 3-5 non-redundant maps; more only if each adds new insight (repetition is penalized).

- Clearly explain your maps. (e.g., What can you learn from the maps? Do you see the same relation between X and Y as before in your maps? How are they related to your previous findings? Do you see any outlier locations?).
- You should know the geographic level that you are working with. Is it a Zip code? Is it county? Is it state? Is it a country? Check the column that contains this information. Is it nice and clean? What is the format? Are there any missing data?
- If your assigned dataset is not at the state level, your maps should reflect a smaller geographic unit. For example:
 - If your data includes lat/long, use county or ZIP code level mapping.
 - If you have county or ZIP code data, these should be the smallest geographic units in your map. Do NOT aggregate to the state level. You lose valuable info.
 - Talk to us if you don't know how to do it!

Note that three is the minimum requirement for a passing grade for this part. I set the minimum requirement since going beyond that is project specific. You should go beyond this requirement. Create innovative indexes, shares, and mixed variables to

plot more meaningful maps, and explain your findings. Full marks for this section will be awarded to papers with three or more well-executed maps of relevant variables. They include clear explanations about the choice and construction of the variables, the relevance of the maps and variables depicted to the research question and the message, and describe the findings and the next steps in clear economic language. However, as I mentioned above: “creating more than three repetitive or uninformative maps will lower your grade.” Talk to us if this is not clear.

- If your dataset does not have enough X variables to build on for mapping, you can map your outcome variable for two periods if it makes sense for your research question.
 - Optional: alternatively, you can merge your data with another relevant and interesting dataset and the same geographic level to map the new variables. **Have a clear subsection and explain the new dataset if you use this.** You can use the subsection “Merging with a New Dataset” and explain the source of the data and why you think it’s valuable and relevant to your research question. Include this subsection in your P2 so we know this is a new dataset.
- Familiarize yourself with the U.S. Census Shapefiles. Look at the Wisconsin voting example in the maps Notebook. I have some links in the map module and some tutorial videos. You can use these links, as in the Wisconsin example, and create your maps.
 - You should create a new data frame with the unique location names. All your datasets are in a panel format, meaning you have more than one row for each location. Therefore you need to first create a dataset with the unique values for each geographic level (whatever value you want to plot) and the unique names or IDs for the geographic groups (zip codes, cities, etc.), put it in a geo-spatial data frame, and then map it. For example, if you have created an average measure of SAT grades for each zip code. You want a data frame with these columns: (zip code, average SAT for each zip). You can then merge this with your zip code shapefile (like we merged county with poll results in the Wisconsin example) and plot. Another strategy is to plot all the points (if you have long/lats for all your observations) and have a heat-map-style visualization.
 - There will be more complications. For instance, the name of the city, county, or state may have extra strings, which may be in lower case or upper case; therefore, your merge with the shapefile may fail. Always check the values in your datasets (containing the original location) and your shapefile. ***Leave 3-4 days for errors and understanding the sample code.***
 - Polishing your map should be easy at this point because you have a draft from your week’s exercise!
 - Important: you can have one or two interactive maps (for your reference and future use) but most of your maps, especially the important maps that are related to your main message should be static similar to examples you saw in class. We grade your static maps only.

Part Four: OLS Regression

(35 points)

Eight regressions are a minimum; quality, novelty, justification, and coherence matter more than the total number.

1. In your first project, you determined your dependent variable and your research question. Do you think the economic relationship between your Y and X is linear or non-linear? Provide economic intuition using economic theories and facts from your data to back up your answer.
2. Choose your Xs based on the existing theories. Why do you think that your Xs should be in your regressions, and why do you think they can explain your Y?
3. Run (a minimum of) eight separate regressions and compare your estimates.
4. These eight regressions can include different Xs, be of various forms (linear, non-linear), include different years or subgroups of observations, dummy variables, interaction variables, etc.
5. Create two or more regression tables and group the relevant regressions in these tables. For example, you can have two tables with 4 regressions in each.

IMPORTANT:

- Each regression must have more than 2 variables (excluding fixed effects) but you can have bi-variate regressions as benchmarks (Y on X). For instance, your table can have 5 columns with the first being Y on the most important X (as a benchmark) but don't count that as one of the regressions, it's just a benchmark. You can then add more columns with more than one X variables.
 - **Your variable choices cannot be random** and you should have a good reason for why a variable should be or should not be included in each regression. Adding variables and dropping them without any story, economic reasoning, or contribution to the research question is not what we are looking for.
 - You can (should) create dummy variables and interact them with your X variables and interpret the joint distribution. Ask questions in class if you need any clarifications.
 - You should put your regressions in a nicely formatted table. **Important:** We don't grade raw python regression outputs.
Use `display(HTML(stargazer.render_html()))` to show your table.
6. Justify why you chose to run these regressions. What is your economic reasoning for running them and what do we learn?
 7. Choose your *preferred specification* and explain why you chose it. In the final project, you will be comparing your results from this preferred regression with your ML results.
 8. How do you evaluate your regressions? What measures should you use to assess the performance of your regressions?

9. What do you understand from your regression results? Briefly explain how these results help you answer your research question.
10. **Super Important:** Simply listing coefficients is not interpreting your results. For example, saying that “the coefficient of my first X is 0.5 and the coefficient of my second X is -0.1” is not enough. You need to tell a story based on these numbers:
 - What do these coefficients mean in the context of your research question?
 - How do they relate to economic theory or real-world implications?
 - What patterns or insights emerge?
 - Are the effects large or small? Significant or not?

I can see the numbers in your table! You don't need to list them again in your text. What I want to know is: What do you learn from them?

Extra 20 bonus points for **valid** causal analysis such as but not limited to valid difference-in-difference and valid IV regressions. These points cannot be appealed. They are awarded to causal analysis that are correctly designed and executed. Historically, no one has been able to get these bonus marks!

Introduction and Conclusion

(10 points) *Your entire document should have one introduction and one conclusion, not two.* The introduction is the first section of your PDF, and the conclusion is the last section. You should update these two sections with your new findings and future next steps in every project. You do **NOT** need two introductions; you only need one introduction for the entire document (PDF). If this needs to be clarified, please ask us during office hours.

Your introduction must include at **least ten to fifteen papers** that are directly relevant to your research. Please cite your papers with APA style. These cannot be news articles and must be academic papers that you have found on sources such as Google Scholar.

Important: Listing papers in your introduction is not what I mean by “citing” papers. I want you to write a story about the literature. Part of the introduction is a review of the literature where you explain (not list) what all the other papers have done (briefly) and then justify why your research is important. I put a podcast in the project module to help you with this part.

Your conclusion must include your future work and be written in a formal language. Include your updated introduction and conclusion. Write about your new exciting findings. You should re-write your intro and conclusion and include the comments you received. If there are no meaningful changes to your previous intro/concl, the 10 points will not be awarded. One more time, you need one introduction and one conclusion for the entire paper (not two intros and two conclusions). You will keep revising these parts in your final project.

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

Enjoy being a creator!