First Project Guidelines

ECO225, Professor Khazra

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

Formatting

You will lose **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.

2. The projects should have clear "section titles" (e.g. Introduction, Summary Statistics, Summary, Future Steps).

3. 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.

4. 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.

5. 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 in your project.

Please note that if you include a graph or a table, you should **explain what you learn from it.** Do not add an output without any explanation. All reports should have an introduction and a conclusion. Suppose you want to send your project to a company or school that you are applying to. The final product should be a clean and comprehensive report.

- Do not include unnecessary chunk of codes or outputs and errors.

- Any graph, summary, or output should have an explanation following. Why do you include it in your report, and what do you understand from it?

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 instruction.**

Coding and interpretation

At the end of this course, we want to find the relation between an outcome (Y) and multiple predictive covariates (X) by applying different methods. To start, the goal of this report is to learn about your data-set, and upload it, and start working on it. You do not have to merge your data with other data-sets at this point.

- (40 points) Create a new Jupyter Notebook. 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.

  Write a brief introduction of your project (one or two paragraphs), source of the data, and important background necessary to understand your project (Keep it short. You will complete it over time). An outsider should be able to understand what you are trying to do in this project. Clearly explain Y (outcome) and at least two X.

- (25 points) Read the data on your Jupyter Notebook. It helps you if you change it to a data frame format afterwards. (use DataFrame() in pandas)

- (25 points) Choose at least two X that you think can better explain the Y. Show the summary statistics of at least two X and Y (use sum() in pandas library).

- (10 points): Do at least two of the following tasks. The quality, presentation, and interpretation of your plots will be graded. Why did you choose this variable? What do you understand from the observed patterns?
  - Plot the histogram of your Xs and the histogram Y. If your X is discrete, you can show a boxplot instead of histogram.
  - Plot the relation between Y and different Xs separately (Y on the vertical axis and X on the horizontal axis)
  - Use Groupby to add more innovative graphs to better explain the covariates and their correlation with the outcome. For example, you can show the distribution of covariates in different subgroups

- (Not graded:) Create a GitHub repository for your project and keep track of its different versions.

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!