#### DS5610 EDA: Data Presentation

**Assignment Overview**: In this assignment, you will showcase your proficiency in exploratory data analysis (EDA) by preparing a 5-minute in-class presentation. You will choose a dataset of your preference that involves panel data, spatial data, text data, or network data. Your presentation should encompass both descriptive statistics and exploratory visualizations. The assignment aims to evaluate your ability to effectively communicate your insights using high-quality tables, figures, and the underlying code.

**Due date and submission/format**:

* October 31 – panel and spatial data
* November 2 – panel and text data
* November 7 – panel and network data

The slides (in whatever format you prefer: powerpoint, google slides, pdf, etc) and the raw code used to generate all tables and figures in your slides (as a Jupyter notebook) must be uploaded via the github classroom assignment link (<https://classroom.github.com/a/53z2TfYh>) by the start of class on the day of your presentation. These two documents will be graded jointly, so they must be consistent (as in, don’t change the Python code without also updating the presentation figure!).

Provide well-commented and organized code snippets to accompany each table and figure you present. Your code should be reproducible and clear, showcasing your ability to transform raw data into informative visuals.

**Deliverable**: The presentation should be structured as follows:

* Introduction/motivation (1 slide) - a brief introduction to the dataset and the question(s) that motivate this presentation. The question(s) should be conceptual and open-ended and not prompt a specific analysis. In particular, make sure you understand the difference between a question and an instruction. This is a question: *How has the weight distribution of alpine skiers changed over the years?* This is not a question; it is an instruction: *Make a series of boxplots of the weight of alpine skiers versus the year of the olympics*.
* Data (1 slide) – describe the original data and what parts of the dataset are necessary to answer the motivating question(s). Imagine that your project is a standalone document and the grader has no prior knowledge of the dataset.
* Analysis (2-3 slides) – produce 2-3 tables/figures that best explain and highlight aspects of the data related to your motivation or that motivate future inquiry and thought. Discuss what insights they reveal about the data. These should be exploratory using descriptive statistics and focusing on variables that contribute to your data's context. Your presentation should be explain to explain what the table/figure tells us and it should be clear why that table/figure is the best way to convey that information. I suggest 1 table and 2 figures, but that is not a mandate. The plots should be of different types. At least 1 of your analysis slides should be specific to the type of data to which you have been assigned. Your slides do not need to provide the code that generates your plots, that can be left to the submitted notebook
* Takeaway (1 slide) - interpret the results of your analysis. Identify any trends revealed (or not revealed) by the plots. Speculate about why the data looks the way it does.

All figures and tables in your presentation must have corresponding code. Natural language content in the slides like titles, bullet points of summaries/takeaways, etc is fine and does not need code. Any figures or tables in your presentation shown without corresponding Python code that generated the result will be considered incomplete. 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.)

**Other notes**

All tables and figures must be made in Python and will be graded on the quality of the code that corresponds to each relevant part of your presentation.

Dataset Selection - choose a dataset that has data amenable to one of the two categories for the day of your presentation: panel data (any day), spatial data, text data, or network data. You are encouraged to explore public datasets available on reputable platforms or discussed earlier in the course. Ensure that your dataset is sufficiently complex and meaningful for exploratory analysis.

You are not required to perform any statistical tests in this project, but you may do so if you find it helpful to answer your question. While you're encouraged to explore and apply advanced exploratory techniques, prioritize clarity and effective communication in your presentation.

If you encounter any challenges or require clarification, don't hesitate to reach out to the course instructor or teaching assistants.

This assignment aims to reinforce your EDA skills and your ability to present complex data analysis concisely. Good luck, and we look forward to seeing your insightful presentations!