Project Milestone 2

Applied Data Science using Python
New York University, Abu Dhabi

|| Out: 26th Oct 2023 || Due: 14th Nov 2023 ||

The second project milestone will be in the form of an **in-class presentation**. Here are some high-level guidelines:

- Each **presentation should be limited to 15 minutes**. This is a strict time limit to accommodate all the groups.

- Presentations should include a **"walkthrough" of your Jupyter Notebook**, and if needed a small deck of powerpoint slides. We are most interested to see the code and techniques used by your group to explore and clean the data.

- All the **code should be written in Python**.

- **Jupyter notebook should be submitted to Brightspace**. Detailed explanation of your thought process in each step should be included in the text cells and as comments in the notebook. Your notebook should be comprehensive. We have provided with this document another example notebook of the **"Journey from raw data to insights (Part 2)"**, which is a walkthrough of the statistical analysis of California Housing dataset. You can use this notebook as an example of how detailed we would like your Jupyter notebooks to be.

In the first milestone of the project, we expected you to (a) choose your team and your dataset (b) develop your hypothesis, (c) study and understand the data, and (d) clean your data.

In this milestone, we expect you to **explore your data**. The objectives of data exploration are three-fold:

(i) **Identifying features:** You need to identify the informative and, (based on your hypothesis) relevant variables in your dataset. You would need to factor in the variables that affect your **outcome** measure. What is your outcome variable? What are the relevant variables that affect your outcome? Are there any variables that may not be explicitly available, but can possibly be inferred? Thinking and discussing these questions should help you accomplish this task. You should be able to justify why the features you chose are important and relevant to your hypothesis in question.

(ii) **Describing and visualizing feature variables:** Once you have identified the relevant features, we would expect you to describe each feature/variable. This means:

- Computing basic statistics for each informative feature variable (eg. min, max, range, median, mean, standard deviation, and count -- if categorical, etc.)
- Plotting histograms or box plots of the most informative continuous variables to describe and understand their distribution.
- Plotting variables across time (if available).
- Plotting variables after stratifying/grouping your data wherever necessary, and visualizing each stratum/group separately or in comparison with each other (such as male vs. female, old vs. young, etc).

You do not need to do all of the above, but choose how to describe your features based on your hypothesis. In describing your features, think of how visualization can help describe your data. The notebook tutorial contains some examples of plots for your help.

(iii) **Statistical Analysis:** Finally, once you have described your data and relevant and informative features, we would like to understand the **relationship** between different variables, and the chosen outcome measure. This would require you to conduct some statistical analysis. Possible examples include:

- Computing and plotting correlations between different variables of interest. This is especially useful if your hypothesis necessitates creation of a predictive model. In such a case, correlations can help indicate possible associations between variables for further exploration.
- In case of a categorical variable, computing the average outcome of each element.
- Using ANOVA if analyzing statistical differences among the means of two or more groups.
- Using pair plots to understand relationships between different variables.

Whichever statistical tests and analysis you conduct, and figures you plot, make sure your notebook and presentation includes some discussion, interpretation, and justification for your choices.

Important Note: *The kind of analysis and visualizations that your group would create would be dependent on what you want to study -- specifically the hypotheses you chose in Milestone 1. Therefore, all your visualizations and analysis should be* ***relevant****. In other words, we are* ***not*** *interested in a jumble of statistical analysis and visualizations, and thus each element (plot, tables, analysis) in your presentation should be justifiable and* ***convey a story.*** *That said, after exploring data, sometimes data scientists find their initial hypotheses infeasible. In that case, you can of course add/adjust your hypotheses as needed.*

Grading

This milestone constitutes **5%** of your final grade, and will be graded from 100 points. The breakdown/rubric is as follows:

Category	Points
Identifying and describing relevant and informative features, pruning uninformative features (if needed).	20
Relevant visualizations (following good visualization principles, and relevant to the hypothesis in study)	30
Relevant and comprehensive statistical analysis (with correct choice of the tests/methods to analyze relationships)	30
Presentation	10
Well explained notebook with all the relevant code, analysis and visualizations	10