[MS3] EDA, Planning, and Setting Goals [4/5-4/18]

Key dates:
EDA, Adjusted Project Plan and Setting Goals Due: Friday, April 18, 2025 at 9:59pm.

Note that **submissions are due at 9:59pm** the day of the deadline (submission window closes at 10pm sharp). There are no late days for any of the project milestones. It is better to submit what you have than nothing at all!

Cite all use of generative AI agents (if used) as well as any outside resources used in your `.ipynb` or PDF file. See **here** and **here** for our citing expectations as well as our course syllabus regarding our policies. Violation of these policies will result in a reduction in your group's score with no option of a regrade.

Objective

In this milestone, you will focus on inspecting your data through Exploratory Data Analysis (EDA). The goal is to take the insights that you gain from the analysis to inform your project plan and to set group goals and workflows for Milestones 4 and 5.

Deliverables

Please submit a 2-3 page written document PDF **and** the `.ipynb` file that was used to create the EDA and visualizations, which will include the items outlined below. The `.ipynb` should be a different notebook than your MS2 submission. Your 2-3 page limit includes reasonably sized figures and images and the submission should include:

- Canvas Project number
- Group members' names
- Data Description
- Summary of the Data + Data Analysis + Meaningful Insights
- Clean and Labeled Visualizations
- Summary of Findings
- Clear Research Question
- Baseline Model or Baseline Model Implementation Plan

Please refer to the "General Guidelines" below for more insights. Make sure that any plot on your report is meaningful in the sense that it affects your decisions regarding your project direction and your plots are labeled correctly. **DO** include an explanation of how each plot or meaningful insight you have on your report shapes the project direction. It could affect how you decided to handle specific features, the baseline model you chose, the feature engineering you did, etc.

At this point, please work with your group members to develop a plan for MS4 and MS5, making sure everyone's roles are clearly defined. The descriptions for MS4 and MS5 will be released soon, and there's no need to submit your plan to the grader.

FYI, MS5 includes an evaluation of each member's time and effort, along with a ranking score. This will influence your individual final grade, as it's intended to prevent any member who contributed significantly less from receiving the same grade as those who worked more. Make sure you are working as a team.

General Guidelines

Your submission should include the following components. Note that these guidelines should serve as a suggestion for what to include. If you have any concerns, please contact your TA/TF/CA.

- **Data Description:** Provide any missing information from Milestone 2 based on the feedback you received from your grader.
- **Summary of the Data:** Provide the shape of the data, data types, and descriptive statistics such as mean, max, and dtypes. Additionally, provide a summary of the features of the data, including histograms, correlation plots, and clustering plots as appropriate.
- **Data Analysis:** Identify patterns, trends, and outliers in the data. Additionally, explore the relationships between variables and identify any potential confounding variables that may impact the analysis.
- **Meaningful Insights:** Based on your analysis of the data, provide meaningful insights. Meaningful insights are those that connect back to your problem and are relevant to your specific context. Any insights should be well-supported by the data, provide actionable recommendations, and have a brief justification for why or how it's important to the project.
- **Clean and Labeled Visualizations:** Visualizations are important components of EDA and should be clean, labeled, and well-presented. You need to ensure that your visualizations are easy to understand and can be included in their final presentation slides or report. Anyone that reads your EDA should be able to understand what is depicted in the plots just by looking at them. Viz's that are not labeled correct will result in a loss of pts.
- **Summary of findings:** Summarize your findings in a clear and concise manner. This can be achieved through the use of visualizations and captions that highlight the most important insights gained through the analysis.
- **(Revised) Project Question:** Based on the insights gained through EDA, you should develop a clear project/research question that will guide your analysis. This question should be well-defined, specific to the problem at hand, and an improvement from the initial project proposal.

Baseline Model or Implementation Plan: Finally, you should include a baseline model **or** a clear plan for its implementation. This can include details on the model architecture, train/test

data, and the evaluation metrics. Your baseline model can be as simple as a clustering algo or a FFNN.

Note: Please leave detailed and useful #comments on your code as well as proper and clear function '''docstrings'''. (i.e. you don't need a #comment to tell us you are plotting something.)

Your submission should demonstrate a thorough understanding of the dataset, exploration of the dataset, the beginnings of model conversation, and readiness for further analysis and higher level modeling. This is what your TA/TF/CA will be grading you on. Please submit your document on time.