Interim report (May 12, 2024 – Sunday till 11:59 pm): The report should be **at most 10 A4 pages** written in Times New Roman font 12.

1. A brief statement of the aim of the project
2. Source of the data and variables, including which is the dependent variable
3. Explain the data cleaning and tidying steps
4. EDA with missing values (**Please include descriptive statistics**, such as means, standard deviations, frequencies with their interpretations and correlations, histograms, matrix scatter plots, etc. Also, use nice visualization techniques, mostly multivariate plots. Use ggplot2, aesthetics, facets, and so on.
4. Exploration of the missingness mechanism and impute missing values and validate it.
5. Data manipulation and feature engineering and dimension reduction (not recommended) if possible
6. Confirmatory Data Analysis for the questions that you raised in EDA. Please combine the EDA and CDA results. Do not give them separately.
7. Cross-validation, train-test split using specific seed number (choose only one CV method). You will use the same train set for all methods that you will add in the final stage.
8. Statistical modeling for prediction or classification includes validating the model assumptions on the train set and measuring the model performance on the test set (provide train set performance to see the possibility of overfitting). **Please include the model output, its interpretation (coefficients are logical or not, and so on), and validation of the model assumptions. So do proper statistical modeling.**

This report is merely a description of work in progress. You should assemble a description of the data exploration you have done thus far. The purpose of this report is to help you avoid an end-of-semester rush to complete the project on time, but you need to finish all the steps given above. This report **will also be graded**.