**Part 1 – Dataset Presentation (15 marks)**

The first part of your notebook must describe your dataset. Start by describing your variables and contextualizing the dataset. The minimal requirements to have full marks in this part are:

* • You must generate attribute’s descriptors for all variables and discuss your results.
* • You must generate plots that can give hints about the data distribution and a discussion about the results must be done.
* • The decision of any descriptor statistics or plots must be justified. You will lose marks if plots are made without any justification.
  1. • Does your data contain missing data? Are outliers present in your data? **The answer must be yes**, you should point it out the attribute(s) and how you will handle it. o If you are using your own dataset, please make sure you create a procedure to remove data and ‘simulate’ that data is missing or to replace some values with outliers.
  2. o At least two attributes must contain missing data (between 10%-20%) and outliers (3%-5%). Make sure you use a proper technique to ‘decide’ which rows will be considered outliers.
  3. • Make sure you remove the missing data rows and the outliers from your DataFrame.

**Part 2 – Data scaling (25 marks)**

The second part of your project will consist of exploring data scaling techniques. The following tasks must be performed in this part:

* • Choose at least 3 data scaling techniques and discuss why you are testing them. At least one data scaling technique must change the attributes’ distribution.
* • Plot the result, side by side, of the original and the data scaling techniques. Discuss the figure.

**Part 3 – Handling missing data and outliers (25 marks)**

The third part of your project must advance your pre-processing strategy to handle the missing data and outliers selected in Part 1.

* • Choose one technique provided by pandas for filling missing data (e.g., bfill, ffill, etc) and fill the missing data. Plot the data and discuss potential problems with it.
* • Create models using the two classifiers or regressors discussed in class (e.g., linear or KNN). Validate them and decide which one must be chosen. You must decide the best combination Classifier, Regressor + Data scaling technique. Discuss why you chose the combination. Don’t forget to justify which evaluation metric you will use to make such a decision.
* • Fill the missing data and the outliers using the decided Classifier, Regressor + Data scaling technique. Make sure the rows are properly scaled before you forecast the values