# Porject One: Part I- EDA and Feature Engineering

STA 511 - Foudations of Data Science

## Contents



## 1 Data Set

Choose a data set that has at least four categorical variables and four numerical variables. The sample size should be at least 200. You can find a data set either from my teaching data repository or other data sources. The data set should be cross-sectional (i.e., each of the data points must be observed/collected/generated at the same time).

The data set must have both continuous and (ideally binary) category variables so you can perform linear and logistic regression modeling.

## 2 Description of Data

The following information of the data should be provided in the report:

- A brief description of the data source.
- How the data set is generated or collected.
- Number of variables and their type (categorical or numerical) and size of the data set.
- List the variable names and their description/definitions.

## 3 Problem Statements and Candidate Models

Formulate at least two practical questions based on the continuous and categorical response variables. Please make clear statements of the practical questions and convert them into unambiguous analytic questions so you can identify candidate models with sufficient justification to address the practical questions.

Write model formulas and assumptions of all candidate models explicitly.

# 4 Exploratory Data Analysis and Feature Engineering

Perform the standard EDA to serve the following major purposes:

- Inspecting data issues such as missing values, mistakenly recorded data values, inconsist data formats, etc. and fix them;
- Identifying new patterns/insights to improve subsequent modeling;
- Checking assumptions of candidate models and perform appropriate feature engineering methods

To present your EDA in clear logical order, you are encouraged to use subsections to organize your work.

For each EDA and associated representation, you should

- open a paragraph with one or few sentence to describe the reasons for the specific EDA before actual analysis;
- After the analysis, interpret what you observed and the implication of potential feature engineering;
- Perform feature engineering (if necessary) based on EDA findings, and thoroughly document all steps to ensure reproducibility.

# 5 Creating Analytical Data

Create an analytical data set that includes

- all original feature variables if no feature engineering is not needed
- all feature engineered featured variables and exclude the corresponding orginal variables

All variables will be called directly in subsequent models. Note that all numerical feature variables needs to be standardized for predictive modeling.

# 6 Wrapping Feature Engineering Code

Wrapping feature engineering code into **reusable functions** for predictive modeling. This ensures that the same transformations applied during training can be seamlessly applied to new raw data during inference.

- **Modularity**: Each feature engineering step should be a separate function.
- **Consistency**: Transformations must behave identically on training and new data.
- **Stateful Transformations**: A term refers to storing learned parameters (e.g., imputation values, scalers) during training for reuse on new data.

# 7 Reporting and format

Use the suggested reporting template (the RMarkdown Source can be found at https://pengdsci.github.io/STA551/w01/w01-ReportingRMarkdoenSource.txt) and the report component at (https://pengdsci.github.io/STA551/w02/w02-AssignSunmission.html)