Requirements

You need to produce an R or Python notebook that covers the full scope of the data science courses, from exploring data to optimizing machine learning model performance. Throughout each stage of the process, thoroughly explain your thought process. For example, perhaps you chose to ignore a certain variable because it is too related to another feature, or because regularization indicated it was not useful.

* **Exploratory Data Analysis**: Summarize variables, visualize distributions and relationships. Generate a few interesting questions about the data and explore them with some visualizations.
* **Research Methods**: Calculate the sample correlation between at least one pair of variables. Come up with a hypothesis and calculate the *p*-value.
* **Data Cleaning and Preparation**: Apply any appropriate preprocessing steps, such as removing duplicates, missing values, outliers, and scaling data as appropriate (note that which model(s) is/are chosen may determine whether scaling is necessary).
* **Feature Engineering:** Create new features or transform existing ones to improve performance. Even if you decide not to use these features (e.g., they don’t affect performance or make it worse), keep the code and an explanation of what you tried in your notebook.
* **Model Selection:** Try various models (at least 3), showing your evaluation process. Clearly indicate which metrics you used and the performance of each model. Be sure to address any imbalance in the data, as well as using an appropriate train/test data split.
* **Performance Optimization:** Use regularization, hyperparameter tuning, or other techniques to further optimize your chosen model and/or help select the best model.

At the end of your notebook, provide a brief summary (one paragraph) of your model – what it is, what preprocessing, feature engineering, and optimization you did, and the final accuracy (or another appropriate metric).