The core requirements are:

1. There are well-defined questions or purposes to the analysis: Done
2. It should involve some data preparation and exploration
3. You define a baseline performance with a simple model
4. You will make use of at least two analysis/prediction techniques from the unit
5. Develop some kind of visualization of the data or results

**Requirement 1**: done

**Requirement 2**:

+ Preparing data for analysis. 🡪 Done

+ Exploration Data Analysis: Pending

🡪 Requirements form EDA tasks:

|  |  |  |
| --- | --- | --- |
| 1 | variable identification | define each variable and its role in the dataset |
| 2 | univariate analysis | for continuous variables, build box plots or histograms for each variable independently; for categorical variables, build bar charts to show the frequencies |
| 3 | bi-variate analysis | determine the interaction between variables by building visualization tools |
|  |  | 3.1. Continuous and Continuous: scatter plots |
|  |  | 3.2. Categorical and Categorical: stacked column chart |
|  |  | 3.3. Categorical and Continuous: boxplots combined with swarmplots |
| 4 | Missing value treatment | Detect and treat missing values |
| 5 | Detect and treat outliers |  |

**Requirement 3: Pending**

**Requirement 4:** asked to use more than one analysis technique. For example, you might use clustering to find groups within the data and then perform a linear regression on some variables within the groups. Or, you might use logistic regression to establish a baseline classification performance and then apply a neural network to see if you can improve performance. This would also satisfy requirement 3.

**Requirement 5:** can be involved in any part of the project, such as data itself, data exploration, data analysis. You may use a bar to visualize a categorical variable or a histogram for a numerical variable.

Here are few suggestions for the advanced project:

1. Make use of linear regression as a predictive model and improving it using polynomial regression. Find important features using RFE technique.
2. Make use of various classification/prediction/clustering techniques from the unit
3. Use various criteria (or metrics) for evaluation: e.g. use of Mean Square Error (MSE), Mean Absolute Error (MAE), and R-squared (r2) for regression problem. Use of accuracy, F-score, Area Under the ROC curve (AUC) for classification problem.
4. First implement a simple algorithm (or model) as a baseline and then improving the baseline using more complex models/techniques.
5. Do parameter analysis to find out which configuration of parameters give best model’s performance. For example, the performance under different k for the KNN algorithm.