Data Science Challenge

1) What is the correlation between columns B and I? Explain the role of correlation in a typical machine learning pipeline design. Which columns have the maximum amount of correlation?

2) Plot the distribution of column U. Comment on the nature of the distribution qualitatively. Compute statistical properties of this distribution. What can you do to make it more like a normal distribution?

3) Quantify the interdependence between columns D and H?

4) How would you perform the feature selection on this dataset?

5) Plot the feature importances and elaborate on the results.

6) Predict column ‘y’ and evaluate your model performance.

Submit the results as a Jupyter notebook (.ipynb, for reproducibility) and its rendered HTML export (.html, for presentation) which contains both the code blocks and their outputs.