MS9004 Assignment (50 marks)

BACKGROUND BRIEF

The data set is collected from an experiment using mice in a biological lab environment. The aim of the experiment is to assess the effect of medication in recovering the ability to learn in mice with chromosome abnormality.

In this assignment, we boil the question down to predicting the expression level of a certain protein, which produces detectable signals in the brain cortex of mice. Predictors include expression levels of a number of other proteins, and some qualitative attributes.

VARIABLE DESCRIPTIONS

1. Response Variable

   Y: Expression level of the protein named *pCASP9*.

2. Quantitative Predictors

   Expression levels of 10 other proteins:

Variable Name	Protein Name		Variable Name	Protein Name
X1	*ERBB4*		X6	*PSD95*
X2	*IL1B*		X7	*SYP*
X3	*nNOS*		X8	*BRAF*
X4	*pNR2A*		X9	*DYRK1A*
X5	*P70S6*		X10	*pELK*

3. Qualitative Predictors

Variable Name	Levels and Descriptions
Genotype	“Ts65Dn” if the mouse has the chromosome abnormality trisomy, “Control” if otherwise.

Treatment	“Memantine” if the mouse is injected with the medication memantine, “Saline” if the mouse is injected with only saline.
Behavior	“C/S” (context-shock) if the mouse is stimulated to learn, “S/C” (shock-context) if the mouse is not stimulated to learn.

INSTRUCTIONS & REQUIREMENTS

1. Explore Data (6 marks)

   Perform exploratory analysis on the variables using the whole data set.

   Describe the data and comment on your observations/findings.

2. Fit Model (6 marks)

   Split the data set into training set and testing set in a (approximate) ratio 75:25.

   Set random state/seed using the last 4 digits of your SP admission number.

   Fit the **full** additive MLR model on the training set.

3. Evaluate Model (12 marks)

   Conduct relevant diagnostics on the full MLR model fitted.

   Evaluate the model from the perspectives of model fit, prediction accuracy, model/predictor significance, and checking of assumptions.

4. Improve Model (24 marks)

   Improve the model using at least 4 of the following techniques **where appropriate**:

   - Removing outlier(s) (if any)
   - Centering and/or standardizing of variables
   - Principal component analysis (PCA)
   - Transformation of variables
   - Interaction of variables
   - Variable selection

   Explain how the model is improved after applying each of the techniques.

   [*Remark: There is no “best” model or “standard” solution.*]

5. Present Results (2 marks)

   Present and explain your works for the above, including relevant graphs, figures, and/or tables which may support your analysis, in a report of no more than 12 pages.

   The report is expected to be detailed but not redundant. Fantastic layout design is not necessary, but the report shall be clear and easy to read. Do not submit your jupyter file or copy-and-paste your jupyter file in your report.