MS9004 Assignment (50 marks)

BACKGROUND BRIEF

The primary objective of the Study on the Efficacy of Nosocomial Infection Control (SENIC) was to determine whether infection surveillance and control programs have reduced the rates of nosocomial (hospital-acquired) infection in U.S. hospitals.

This data set consists of a random sample of 113 hospitals.

VARIABLE DESCRIPTIONS

Variables	**Description**
risk	Average estimated probability of acquiring infection in hospital (in percent); **response variable.**
length	Average length of stay of all patients in hospital (in days).
age	Average age of patients (in years).
cultures	Ratio of number of cultures performed to number of patients without symptoms of hospital-acquired infection, times 100.
xray	Ratio of number of X-rays performed to number of patients without symptoms of pneumonia, times 100.
bed	Average number of beds in hospital during study period.
patient	Average number of patients in hospital per day during study period.
nurse	Average number of full-time equivalent registered and licensed practical nurses during study period.
facilities	Percent of 35 potential facilities and services that are provided by the hospital.
affiliation	Medical school affiliation; yes or no.
region	Geographic region; NC, NE, S, or W.

INSTRUCTIONS & REQUIREMENTS

1. Explore Data (5 marks)

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

   Describe the data and comment on your observations/findings.

2. Fit Model (5 marks)

   Split the data set into training set and test set in a ratio of 75:25 (approximately).

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

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

3. Evaluate Model (10 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**:

   - Mean-centering of variables
   - Principal component analysis (PCA)
   - Polynomial regression
   - Transformation of variables
   - Interaction of variables
   - Variable selection

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

   [*Remark: There is no universally best model or definitive solution.*]

5. Present Results (6 marks)

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

   The report is to provide comprehensive information without being repetitive, presented with a simple and clear layout. While an elaborate layout design is not required, it should be presented in a clear and reader-friendly manner. Please refrain from submitting your Jupyter file or including direct copies of it within your report.