**Case Study:**

**For the following questions, please attach the relevant R or Python codes with your submission.**

1. The SENIC dataset attached with this assignment relates to a study carried out by the Centers

of Disease Control and Prevention in the USA to determine whether infection surveillance and

control programs have reduced the rates of nosocomial (hospital acquired) infection in US hospitals.

Here SENIC stands for Study on the Efficacy of Nosocomial Infection Control. This particular data

corresponds to 113 hospitals selected randomly from a master sample of 338 hospitals which were

originally surveyed. A complete list of the variables and corresponding description are included in

the dataset description attached herewith. The following questions relate to various aspects of

the above data:

**[Part A]**

**Suppose the average length of stay in a hospital (Y ) is to be related to infection risk, age and available facilities and services.**

**Questions :-**

1. Regress average length of stay on each of the above predictors separately and write

down the estimated least squares regression model for each.

2. Interpret the slope coefficients of each of the above predictors.

3. Plot each of the estimated models on the corresponding dataset. Does the linear

model appear to provide a good fit in each of the cases ?

4. Determine the predicted average length of stay corresponding to patients with an average

age of 76 years and for hospitals with available facility and services score of 40%.

5. Calculate the coefficient of determination for each of the above three models. Which

of the three predictors (vis-a-vis models) provides the highest share of information on average

length of stay ? Justify.

6. Obtain 95% confidence intervals of the slope coefficients of infection risk

and age and interpret each of those in the context of the problem. Accordingly comment on

the significance of association of each of these predictors on average length of stay.

7. Now regress average length of stay against infection risk for each of the 4 geographic regions. State the estimated regression functions and comment on whether the association pattern of infection risk and average length of stay differs across regions.

8. For each of the models fitted to the geographic regions, perform necessary

regression diagnostics and comment on the validity of the assumptions for each of the four regions/models. If you notice any violation, suggest suitable transformations that would remedy

the situation.

**[Part B]**

In a separate regression, Infection risk (Y ) is to be regressed on length of

stay, age, routine X-ray ratio, number of nurses, available facilities and services and medical school

affiliation (1: affiliated, 0: non-affiliated). Based on this setup, answer the following questions:

1. Fit a multiple regression model to predict Infection risk based on the above

predictors. Accordingly interpret the effect of medical school affiliation and routine X-ray

ratio on Infection risk.

2. Test for the significance of medical school affiliation based on the relevant 95%

confidence interval. Interpret your findings.

3. Using appropriate measures learnt in class (like VIF, R2), comment on the presence (or

absence) of multicollinearity in the above data.

4. Carry out a suitable model selection procedure and identify the optimal model for this

dataset. Justify your findings.

5. Carry out suitable regression diagnostics for the optimal model chosen above and comment

on the validity of the assumptions.