GEMINI DS Assessment

Instructions:

- You will have 2 hours to complete this assessment. There are 6 tasks.
- You can use R or Python and any packages. Preferred submission file types include: Rmd, ipynb, html, md, pdf, R, py.
- Submissions are evaluated holistically. Please briefly explain your thought process regarding any major choices you make.
- Include your name on any scripts or documents you submit.
- Upload your answers using this form: https://forms.gle/aT1mHpXFTK7yFi9A6

Admissions.csv simulates administrative data where each row represents a unique admission to a hospital.
Lab.csv simulates results for patients who had laboratory testing (e.g. blood counts) in their admission.
Transfusions.csv simulates information on patients who underwent a blood transfusion in their admission.

1. Impute the missing *charlson_comorbidity_index* values in any way you see fit, with the intention that this variable will be used as a predictor in a statistical model.

2. Determine if there is a significant difference in *sex* between patients who had an *rbc_transfusion* and patients that did not.

3. Fit a linear regression model using the *result_value* of the “Platelet Count” lab tests as the dependent variable and *age*, *sex*, and *hospital* as the independent variables. Briefly interpret the results.

GEMINI DS Assessment

4. Create one or multiple plots that demonstrate the relationships between *length_of_stay* (discharge date and time minus admission date and time), *charlson_comorbidity_index*, and *age.*

5. You are interested in evaluating the effect of platelet transfusions on a disease. The patients with *platelet_transfusion* represent the selected treatment group. Select a control group in any way you see fit.
   How could you improve your selection if you had more data and access to any clinical variable you can think of?

6. Fit a first-iteration statistical model of your choosing to predict the *result_value* of the “Hemoglobin” lab tests and evaluate its performance.
   How could you improve the model if you had more data and access to any clinical variable you can think of?