|  |
| --- |
| Instructions:   * You will have 2 hours to complete this assessment. Submissions are evaluated holistically. * There are 6 tasks. You can use R or Python. * Preferred submission file types include: Rmd, ipynb, html, md, pdf, R, py. * Record your answers either in a markdown format with your code, in comments within your code, or in this document with code attached separately. * 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. Read in the three csv files as data frames. Combine all three data frames using *admission\_id* into one named **all\_df**, retaining all rowsfrom **Admissions.csv**.
2. Impute the missing *age* and *charlson\_comorbidity\_index* values in any way you see fit. Briefly explain how the imputation method(s) might affect a statistical model that uses these variables as predictors.

## Determine if there is a significant difference in *age* between patients who had an *rbc\_transfusion* and patients that did not. Determine if there is a significant difference in *sex* between these groups.

## 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. Interpret the results.

## Create a plot that demonstrates the relationship between *length\_of\_stay* (discharge date and time minus admission date and time) and *charlson\_comorbidity\_index*. Briefly describe how you would change the visualization to add *age* as a third variable.

## Fit a statistical model of your choosing to predict the *result\_value* of the “Hemoglobin” lab tests and evaluate its performance. Explain your choices and how you might improve the model if you had to iterate on it. What other clinical variables would you be interested in adding to the model (they don’t have to be in the provided datasets)?