**P8106 Final Report**

Zhezheng Jin (zj2358)

***Exploratory analysis and data visualization***

A total of 1000 patients were included in our analysis. The data was split into training data (80%) and testing data (20%). From our training dataset, figure 1 showed relationships between the outcome of recovery time from COVID-19 in days and the continuous variables, namely bmi, SBP level, LDL level, age, height, and weight. These associations were shown to be non-linear since their optimized degree of freedom (df) when fitting a smoothing spline are not 2. Then, figure 2 with violin plots portrayed the distribution of outcome stratified by categorical variables. Across all groups, distribution shapes with median values were similar. However, significantly lower maximum values were present in Black population, current smokers, patients with diabetes, and from study A. Next, figure 3 demonstrated a correlation matrix across all variables with some indications of presence of multicollinearity (correlations between predictors). Thus, we should proceed with regression techniques and approaches which could address this issue.

***Model training***

***Results***

***Conclusions***