Based on our choice of Death Event as our response variable, we decided to use logistic regression for prediction, first we loaded the data and took a cursory look at it, we had 12 Explanatory Variables and the last column, Death Event as our Response Variable. In the model selection step, we will use the AIC/BIC method, and then choose the better one between these two methods

The next step to linear regression. we first make the correlation coefficient between each explanatory variable into a heat map, we can see that the correlation coefficient between most variables is not high. In the subsequent model selection step, we can optimize our model more easily through heatmap.

We split the data into two training datasets of the same size and a testing dataset, and we first used the AIC method to fit the two datasets we got in the previous step, and we can see from the results of the fitting that several explanatory variables have a significant impact on our response. On the other hand, anaemia, high blood pressure, and serum sodium, the three variables seem less important in our model.

Because we have more than a dozen explanatory variables in our data, in order to avoid overfitting of our model that is too complex and affects the results of the AIC method, we will use the BIC method to further analyze our data

The results of the BIC method show us five optimal models, including four 4 variables models and one 3 variables model. The values of BIC of the 5 model we get are very close, and we will compare the posterior probability between them. For ease of view, we show the value of ‘post prob’ of 5 models in a histogram, and we can see that the post prob with 3 variables is about 3 to 5 times higher than the other four models, we choose this model and compare it with the previous model chosen by AIC