**What kind of cleaning steps did you perform?**

1. Replace “?” and “Unknown/Invalid” with NaN.
2. Re-label the responsive variable.

The current responsive variable 'readmitted' has 3 categories:

readmitted <30 11357

>30 35545

NO 54864

This study is interested in those patients that are readmitted within 30 days after discharge. The <30 category was coded as 1, other two categories were coded as 0.

1. Check missing values in columns and rows.

Drop variables with high percentages (>30%) of missing values.

Drop entries with missing values??

1. Encode categorical variables.

The categorical variables 'diag\_1','diag\_2','diag\_3' were encoded with new categories. The categorical variables 'age', 'max\_glu\_serum', 'A1Cresult', 'change',

'diabetesMed', and 23 medications were encoded to numerical values.

**How did you deal with missing values, if any?**

7 variables have missing values, range from 2% to 97%. The top 3 variables with high missing values were dropped.

weight 96.858479

medical\_specialty 49.082208

payer\_code 39.557416

race 2.233555

diag\_3 1.398306

diag\_2 0.351787

diag\_1 0.020636

The entries with missing values are dropped too.

**Were there outliers, and how did you handle them?**

Eight numerical variables 'time\_in\_hospital', 'num\_procedures', 'number\_diagnoses', 'num\_medications', 'num\_lab\_procedures', 'number\_outpatient', 'number\_inpatient','number\_emergency' were examined for outliers using boxplot.