Findings

- Initially two data frames were combined Hospital-length-of-stay and Data-at-admission.
- All the data column from dataset Data-at-admission were taken as the independent variables and the column ‘Hospital-length-of-stay’ from the dataset Hospital-length-of-stay was taken as the dependent variable
- **Initially** we had Total columns: 750, Total rows: 508
- They were matched on the basis on the column ‘id’
- **Handling null values:**
    a. Column ‘height’ and ‘weight’ null values were handled by setting the null values as their mean.
    b. Column ‘smoking_history’s null value was handled by filling it with 0 because it was observed that there were 1’s for rows that had a smoking history.
    c. The columns ['year_they_quit','Methylprednisolone Acetate Injectable Suspension Usp','Prozac Capsules','Rybelsus'] were dropped as 'year_they_quit' wasnt seen as a significant column that would provide insight on the model and the rest of the columns had 508 null values.
- **Total rows:** 508, **Total columns:** 746
- **Types of columns**

Type	Counts
Medications	694
Reasons for admit	27
Comorbitities	18
Personal	7

- Y is the target variable - ‘Hospital-length-of-stay ’, X is the independent variable apart from Y
- The number of columns were reduced in dimension with the help of PCA. We saw an elbow point at 6-8 however it explained around 42 percent of the variance of the data.

Therefore to achieve 80,85 and 90 percent of the data variance we had to reduce to 80, 105 and 140 dimensions respectively.

a. Findings at 80 percent variance:
   i. Using Lasso Regression, alpha=0.1:
      1. Mean Squared Error for training: 113.43747827442814
      2. Mean Squared Error for testing: 176.7316408907595

b. Findings at 85 percent variance:
   i. Using Lasso Regression, alpha=0.1:
      1. Mean Squared Error for training: 110.509238038359
      2. Mean Squared Error for testing: 175.94147441787882

c. Findings at 90 percent variance:
   i. Using Lasso Regression, alpha=0.1:
      1. Mean Squared Error for training: 106.83501460014685
      2. Mean Squared Error for testing: 177.86817100344763

- Findings for Lasso regression for the initial dataset:
  a. 22 important columns were calculated with the help of the coefficients
  b. The model was created around those 22 features where:
     i. Alpha = 0.1
     ii. Mean Squared Error train: 107.48457315351651
     iii. Mean Squared Error test: 177.5778499672496
  c. It was also found that an increase in the alpha value decreased the error in the training set however increased the testing set error
- Findings for Ridge regression for the initial dataset:
  a. Findings for alpha value = 1
     i. Mean Squared Error train: 37.39640727980008
     ii. Mean Squared Error test: 222.43033854644258
  b. Findings for alpha value = 1000
     i. Mean Squared Error train: 127.3814799415317
     ii. Mean Squared Error test: 171.9431982126735