In this analysis we are supposed to pick a dataset and build a machine learning linear regression model from the dataset to predict continuous data in the dataset. I have picked a dataset from Kaggle which tells us the medical charges of a person depending on the factors on gender, number of children, age, BMI, smoker or not, and the region they live in. I found this dataset to be interesting because I also work in the insurance field in my hospital. From the looks of the "charges" column presented I think the seem to be pretty accurate to my hospital as well. So, I felt like it would be interesting to see what factors determine the charges billed and to predict these charges based on the info provided in the dataset.

To start this lab, I first imported the necessary libraries and read the dataset into a pandas data frame. I looked at the dataset to see what datatypes and how big of a dataset I was dealing with, and also getting its basic statistics. Then, I made a correlation heatmap of the numerical data to see if there are any obvious correlations. Then I checked for any null values and visualizing the categorical data as bar graphs. The gender distribution of this data set is pretty even, the dataset is biased towards non-smokers, and slightly biased towards the southeast region. After that, I had to one hot code the categorical data and make them numerical. After one hot coding the region data, the sex data, and the smoker data, I replaced the original dataset with this new one hot coded data and proceeded to build the linear regression. I made the "X" features values by dropping the "charges" column and made the "y" target by making it the "charges" column. Then, I imported the necessary libraries from scikit learn to separate the data into a 70/30 train to test ratio while making the random state equal to 2 (to replicate my results). After that, I loaded the linear regression model and fitted the training data into the model. Later, I used that model to make predicted charges based on X_test data. Finally, I calculated the coefficients (there are 9 because I had 9 columns in my "X" features), calculated the intercept, R-squared, mean standard error, root mean standard error, and the mean absolute error.

To make the sense of the model, you can set the "y" equal to all the coefficients multiplied by the "X" values plus the intercept and add on the model's error term. My R-squared value is 0.76 which indicates that about 76% of our regression model fits the data. The R-squared value is between 0 and 1 and the closer it is to one the better. It tells how well the regression model fits the data. The mean squared error is 38,108,732.49, this measure shows the average of the sum of the squared differences between the predicted values and the actual values. This value alone doesn't tell much but it will tell us more when we have other regression models to compare it to. The root mean squared error is 6173.23, this measure is the square root of the mean standard error. It is easier to use because the values are not as large as the mean standard error. The mean absolute error in this model is 4292.58, this measure is the average sum of the absolute value of the residuals from the predicted and actual values. It is a more direct representation of sum of errors. The mean squared error, root mean squared error and the mean absolute error are used to compare one regression model to another. The lower these values are the better the regression model.