Question 3 Report: Analysis by Linear Regression and Alternative Methods for the Estimation of Continuous Target Variable

- **First and Last Name**: Fahrettin Solak
- **School Number**: 201401053
- **The Subject of the Problem**: Analysis by Linear Regression and Alternative Methods for Estimation of Continuous Target Variable
- **The Data Set and Link Used**:
    - **Data Set Source**: Kaggle
    - **Link**: Medical Cost Personal Datasets

Methods Used:

- Linear Regression
- Ridge Regression
- Lasso Regression

1. Introduction

The main objective of this project is to compare the performance of these models by using different regression models to estimate health insurance premiums. Firstly, we created a basic prediction model using the linear regression model. Then, we aimed to prevent excessive compliance and increase the overall accuracy of the model by performing regularization in the model using Ridge and Lasso regressions.

The data set is available on Kaggle as an "Insurance Dataset" and contains several basic features that are used to determine individuals' health insurance premiums. This data set includes variables such as age, gender, body mass index (BMI), number of children, smoking and region. These features help us to understand the impact on each person's insurance premium and more accurately predict future insurance claims.

In this study about the data set, the following stages were followed:

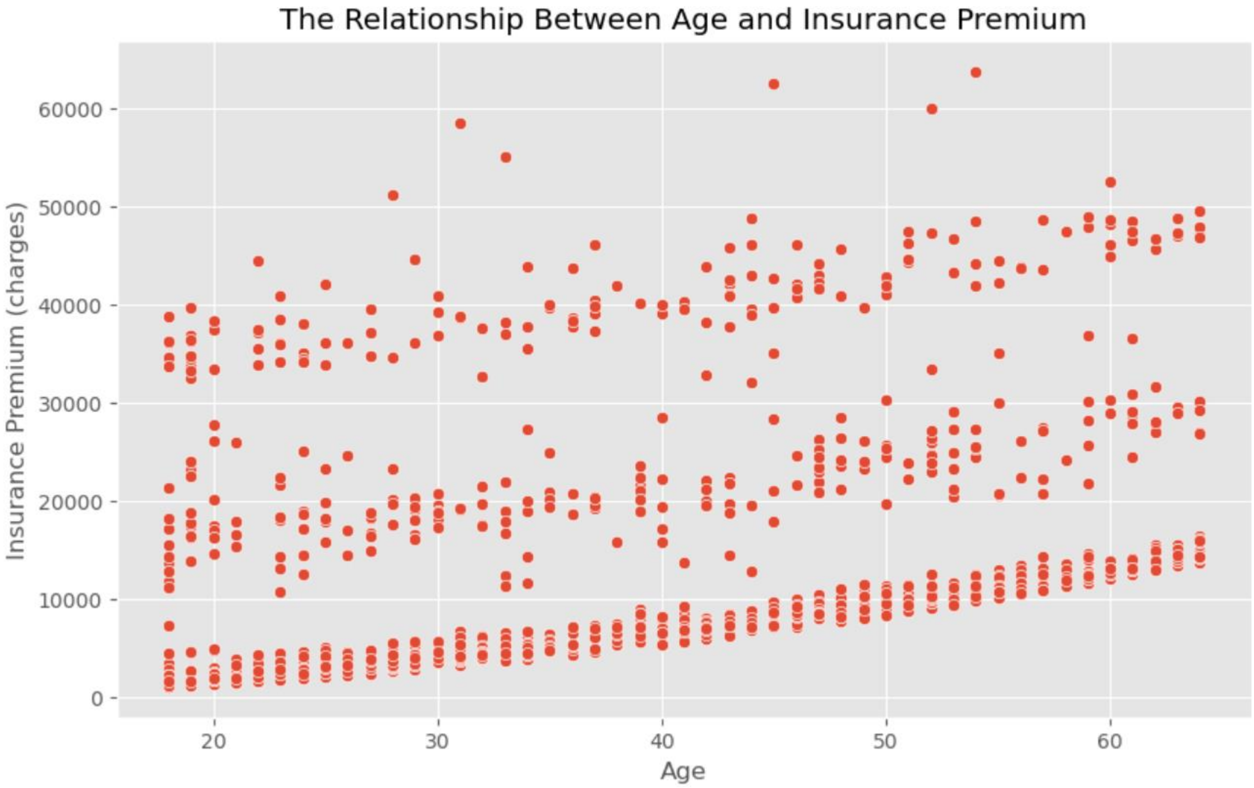
- **Examination of the Data Set**: We tried to understand what each property means and its effects on the target variable.
- **Data Pre-Processing**: We have performed the necessary operations to analyze the data correctly. At this stage, operations such as checking missing values, converting categorical variables and scaling data took place.
- **Model Training and Evaluation**: Different regression models were trained, and the performances of these models were analyzed by comparing.

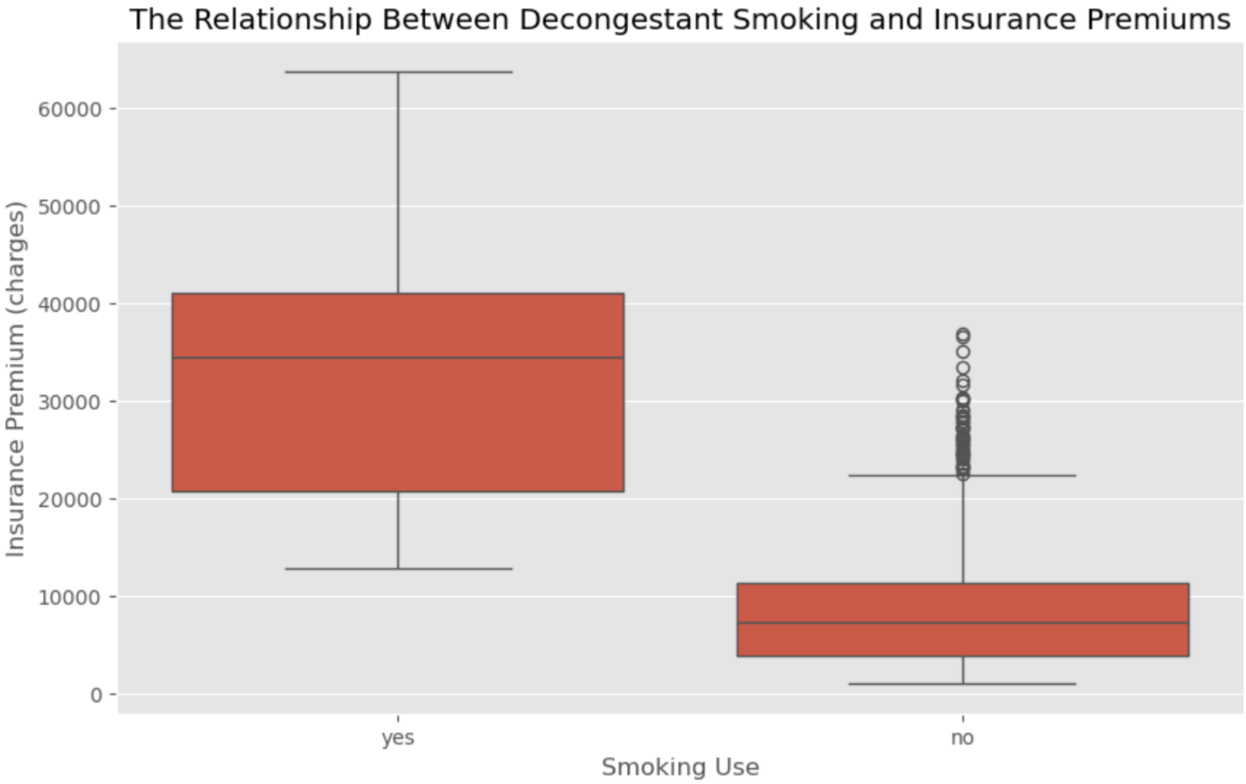
This project aimed to evaluate the impact and utility of regulatory techniques in estimating health insurance premiums. This analysis reveals findings that may be useful in the risk analysis and pricing strategies of insurance companies.

---

2. Data Set and Characteristics

Before starting the data set review, we need to understand the meaning of each variable in the data set and analyze the effects of these variables on the target variable that we will predict. The data set used consists of a total of 1338 rows and 7 columns, and each row in this data set represents an individual's health insurance information. Below is a detailed description of each feature:

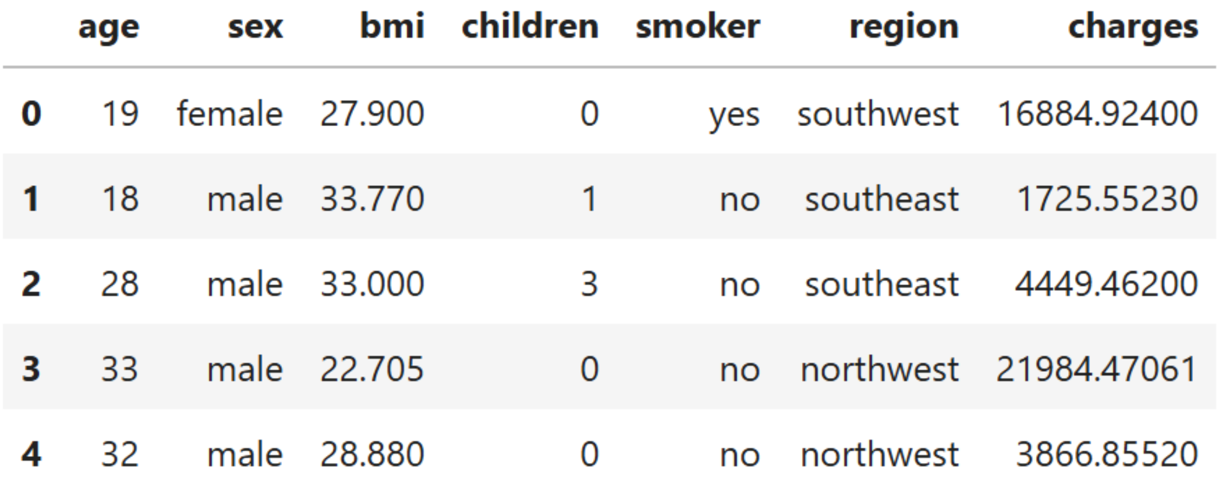

The Relationship Between Age and Insurance Premium
Insurance Premium (charges)
60000
50000
40000
30000
20000
10000
0
20
30
40
50
60
Age


The Relationship Between Decongestant Smoking and Insurance Premiums
Insurance Premium (charges)
60000
50000
40000
30000
20000
10000
0
yes
no
Smoking Use

Data Set Characteristics:

- **age**: This variable refers to the person's age. Age plays an important role on health risks, which can lead to increased insurance premiums. For example, as age progresses, health problems increase, which means higher insurance premiums.
- **sex**: Gender information contains two categories, as male or female. Gender can have different effects on health risks. For example, the incidence of some diseases is higher in women, while some are more common in men.
- **bmi**: Body mass index is the ratio of a person's weight to the square of his height and indicates whether a person is overweight or not. A high BMI indicates that a person is obese, and therefore their health risks are higher. This situation has an impact on the insurance premium.
- **children**: The number of children owned. As the number of children increases, health care costs may also increase. Insurance premiums may rise in this case.
- **smoker**: The smoking situation. It is divided into categories as "yes" and "no". The health risks of individuals who smoke are much higher, and this directly affects insurance premiums. This variable is one of the variables that seriously affects the prediction performance of the model.
- **region**: The area where the individual lives. This variable refers to the geographical region where the person lives. The region where you live is important for access to health services, and this can affect insurance premiums.
- **charges**: The insurance premium, that is, the target variable. This variable refers to the fee that an individual pays for health insurance, and it will be tried to estimate it with the help of all other characteristics.

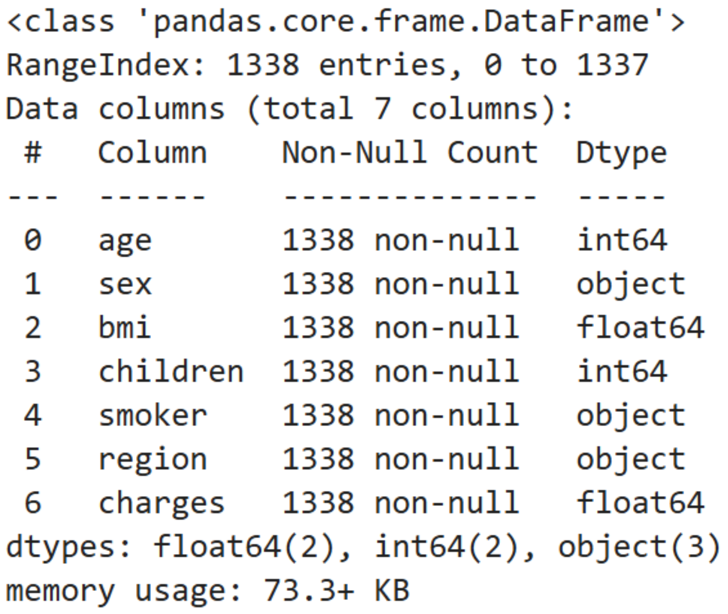
The First Five Lines of the Dataset: This table will help us understand the structure of the data set and how each variable contains data:



	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

3. Data Pre-Processing

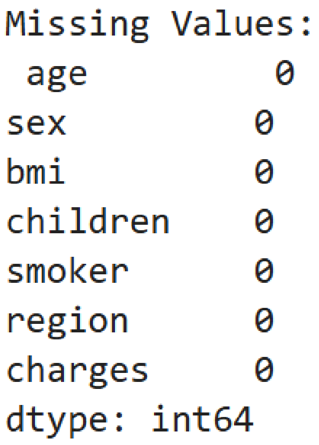
Data preprocessing is a very important step in machine learning projects, because any errors or omissions in the data set can negatively affect the success of the model. In this section, I will cover the data preprocessing stages in detail.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       1338 non-null   int64  
 1   sex       1338 non-null   object 
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64  
 4   smoker    1338 non-null   object 
 5   region    1338 non-null   object 
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

3.1 Incomplete Value Control

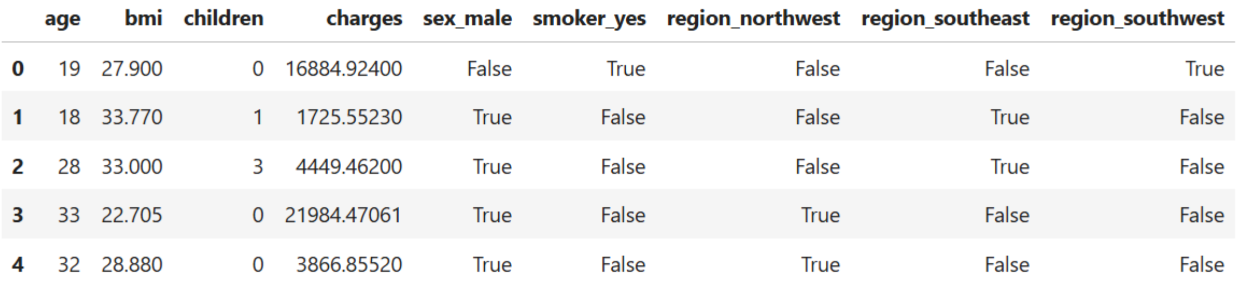
As a first step, we checked whether there is a missing value in the data set. Since the presence of missing values in the data set may cause deviations in our analyses, it is necessary to address the missing values. It has been determined that there is no missing value in this data set.



```
Missing Values:
 age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

3.2 Transformation of Categorical Variables

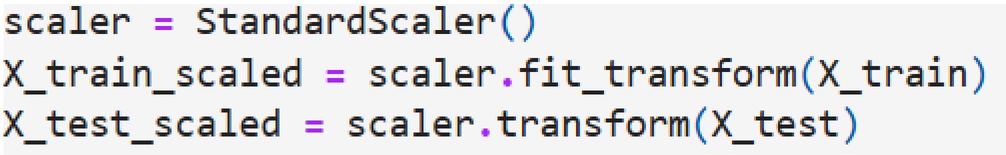
There are some categorical variables in the data set (sex, smoker, region). These variables need to be converted to numerical values in order to be used in models. For this purpose, pd.get_dummies() function. For example, the sex variable contains the categories "male" and "female", and this variable is numerically encoded in the form of two new columns.



	age	bmi	children	charges	sex_male	smoker_yes	region_northwest	region_southeast	region_southwest
0	19	27.900	0	16884.92400	False	True	False	False	True
1	18	33.770	1	1725.55230	True	False	False	True	False
2	28	33.000	3	4449.46200	True	False	False	True	False
3	33	22.705	0	21984.47061	True	False	True	False	False
4	32	28.880	0	3866.85520	True	False	True	False	False

3.3 Data Scaling

Since each variable in the dataset is in a different unit and scale, this can make it difficult to train the model. For example, the age variable ranges from 0-100 Dec, while the charges variable may have much higher values. Therefore, we performed data scaling using **StandardScaler** to ensure that all variables are at similar scales.



```
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

4. Model Training and Evaluation

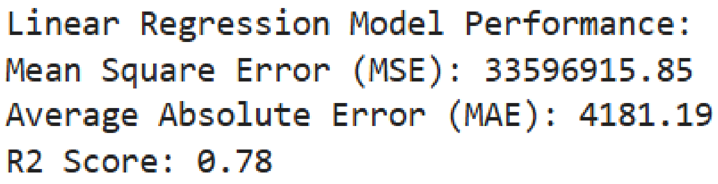
In this section, linear regression, Ridge and Lasso regression models were trained and evaluated. In order to understand the performance of the models, metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE) and R2 score were used. These metrics give information about the forecasting ability of the model.

44.1 Linear Regression Model

Linear regression assumes a linear relationship between the independent variables and the target variable. Dec. First, starting with this model, we tried to estimate insurance premiums. This model expresses the effect of independent variables on the target variable by a simple linear equation.

Performance Evaluation:

- **MSE**: 33,596,920.00 - This value refers to the average of the square of the prediction errors of the model.
- **MAE**: 4181.19 - The average of the absolute values of prediction errors.
- **R2 Score**: 0.78 - Shows how much of the variance of the target variable is explained by the model.



```
Linear Regression Model Performance:
Mean Square Error (MSE): 33596915.85
Average Absolute Error (MAE): 4181.19
R2 Score: 0.78
```

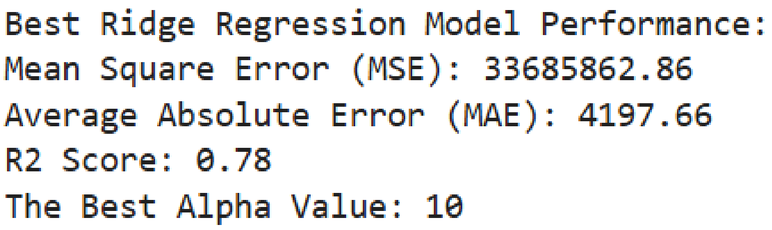
4.2 The Ridge Regression Model

Ridge regression tries to prevent over-matching by adding L2 editing on top of linear regression. L2 editing makes the model more balanced by reducing the influence of features with large coefficients.

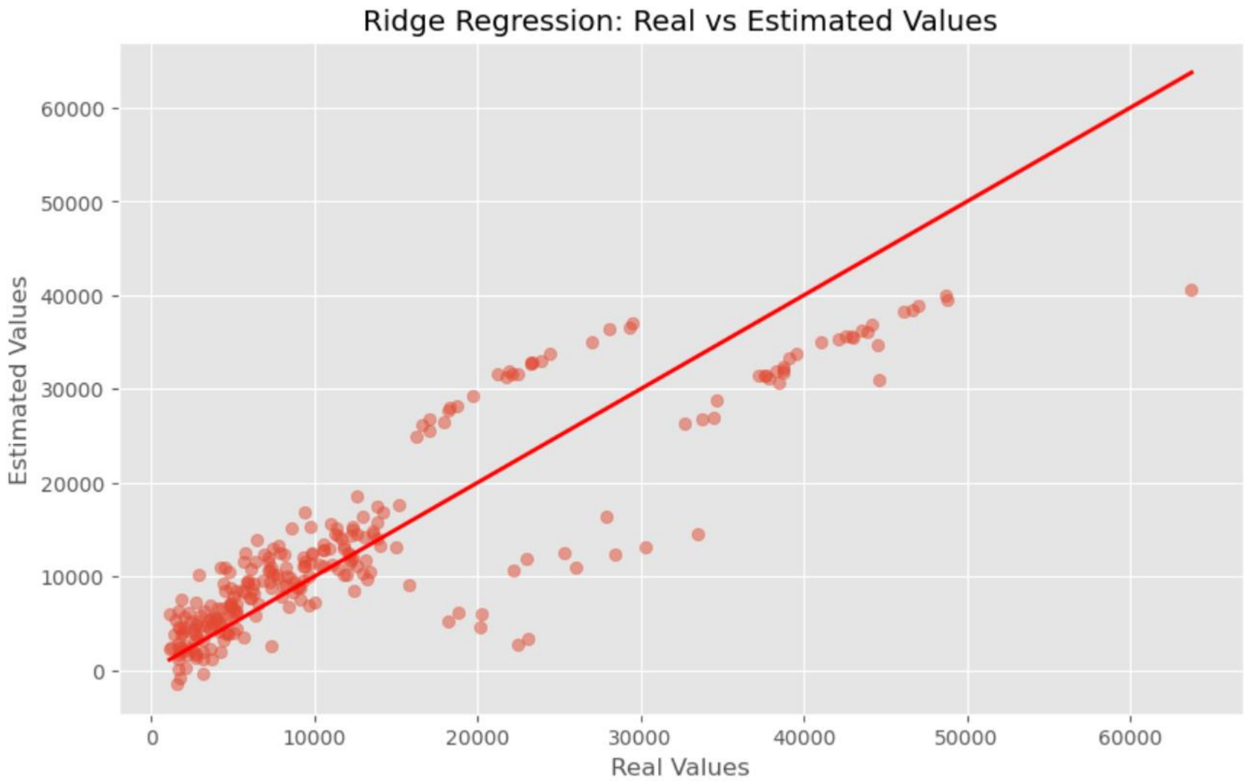
- **Hyperparameter Adjustment**: Alpha, the editing parameter of the Ridge model, has been optimized using GridSearchCV. The model was trained again by selecting the most appropriate alpha value.

Performance Evaluation:

- **MSE**: 33,685,862.86
- **MAE**: 4197.66
- **R2 Score**: 0.78



```
Best Ridge Regression Model Performance:
Mean Square Error (MSE): 33685862.86
Average Absolute Error (MAE): 4197.66
R2 Score: 0.78
The Best Alpha Value: 10
```



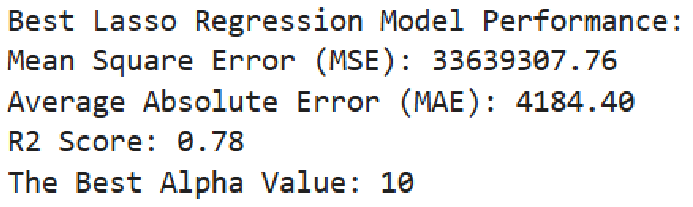
4.3 The Lasso Regression Model

Lasso regression can reduce the coefficients of some properties to zero using L1 editing and make the model simpler. This in turn increases the interpretability of the model and eliminates unnecessary features, providing a better predictive power.

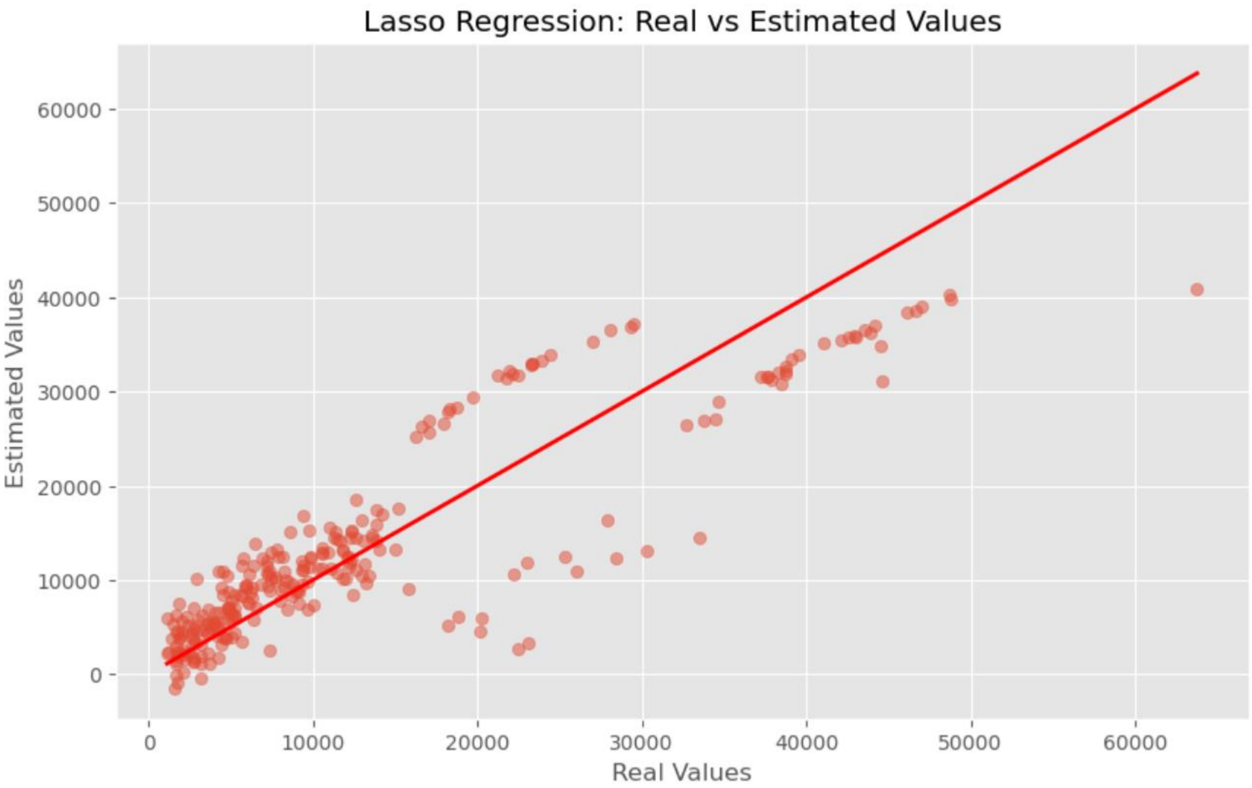
- **Hyperparameter Adjustment**: The alpha value of the Lasso model has also been optimized with GridSearchCV. The most appropriate amount of regulation has been determined by this method.

Performance Evaluation:

- **MSE**: 33,639,307.76
- **MAE**: 4184.40
- **R2 Skoru**: 0.78



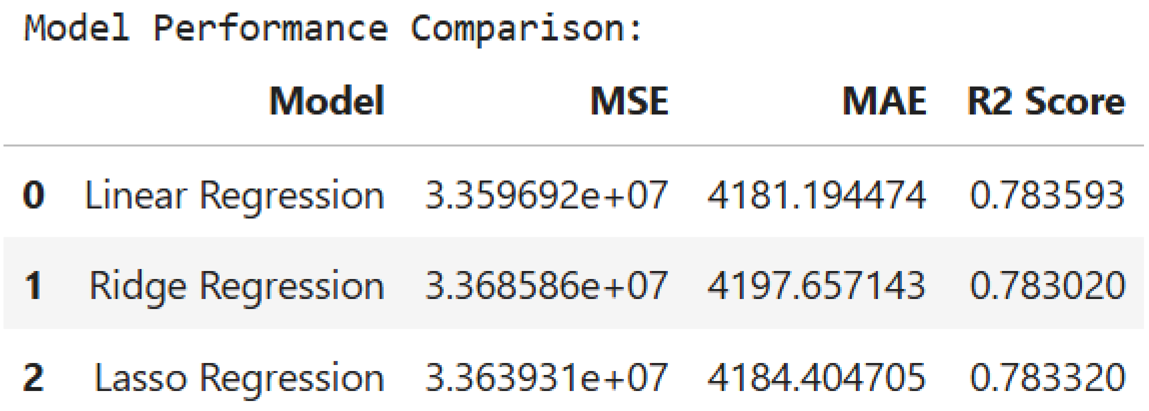
```
Best Lasso Regression Model Performance:
Mean Square Error (MSE): 33639307.76
Average Absolute Error (MAE): 4184.40
R2 Score: 0.78
The Best Alpha Value: 10
```



---

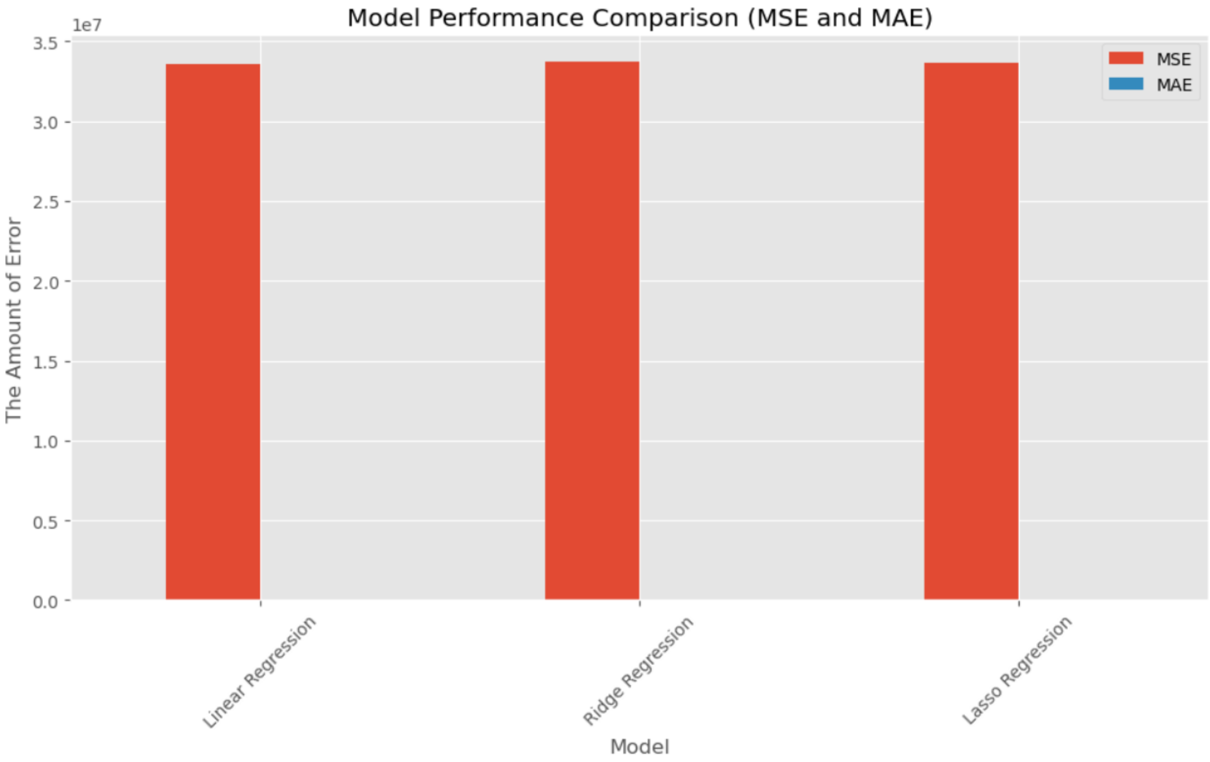
5. Model Performance Comparison

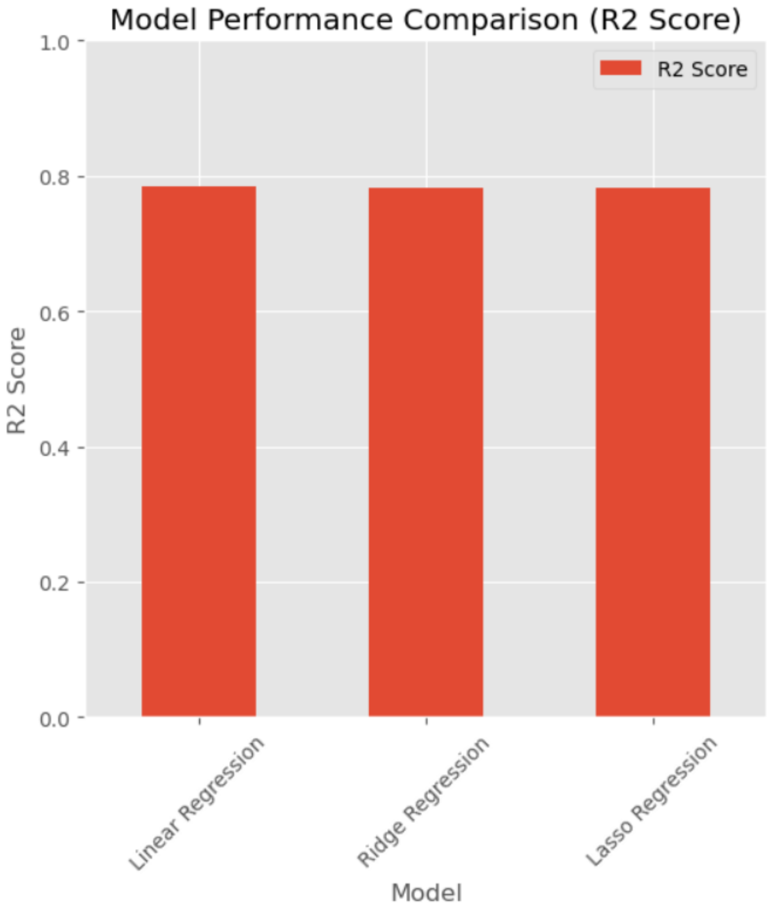
In this section, we evaluated the performance of linear regression, Ridge and Lasso regression models by comparing them. Considering the advantages and disadvantages of each model, it has been seen that the Ridge and Lasso models prevent excessive compliance with regulation and provide more balanced results.



```
Model Performance Comparison:
```

	Model	MSE	MAE	R2 Score
0	Linear Regression	3.359692e+07	4181.194474	0.783593
1	Ridge Regression	3.368586e+07	4197.657143	0.783020
2	Lasso Regression	3.363931e+07	4184.404705	0.783320


Model Performance Comparison (MSE and MAE)


Model Performance Comparison (R2 Score)

6. Results and Learnings

In this project, the performances of these models were compared by using different regression models in estimating insurance premiums. It has been observed that Ridge and Lasso, which are editing techniques, help the model avoid over-fitting and give more general results. We saw the advantages of Ridge and Lasso in this project by experiencing them in particular.

- **Linear Regression**: As a basic model, it allowed us to understand the effects of all properties.
- **Ridge Regression**: L2 offered a more balanced model by reducing excessive compliance with regulation.
- **Lasso Regression**: By removing unnecessary features from the model with L1 editing, it simplified the model and made it more interpretable.

In this project, I tried to estimate health insurance fees using different regression models. I started with Linear Regression and then used Ridge and Lasso regressions to improve the performance of the model. The Ridge and Lasso regressions helped prevent over-compliance by applying penalties, which made the model more balanced and reliable. With Ridge Regression, I was able to reduce the effect of large coefficients, while Lasso helped me identify and remove unnecessary features, making the model simpler and easier to interpret. In general, I have learned that editing techniques are very useful in improving model performance and making predictions more general