PROJECT REPORT

Medical Insurance Cost with Linear Regression

By Somashree Sahoo

Introduction

Health insurance in India is a growing segment of India's economy. The Indian health system is one of the largest in the world, with the number of people it concerns: nearly 1.3 billion potential beneficiaries. The health industry in India has rapidly become one of the most important sectors in the country in terms of income and job creation. In 2018, one hundred million Indian households (500 million people) do not benefit from health coverage. In 2011, 3.9% [1] of India's gross domestic product was spent in the health sector.

According to the World Health Organization (WHO), this is among the lowest of the BRICS (Brazil, Russia, India, China, South Africa) economies. Policies are available that offer both individual and family cover. Out of this 3.9%, health insurance accounts for 5-10% of expenditure, employers account for around 9% while personal expenditure amounts to an astounding 82%.

In the year 2016, the NSSO released the report “Key Indicators of Social Consumption in India: Health” based on its 71st round of surveys. The survey carried out in the year 2014 found out that, more than 80% of Indians are not covered under any health insurance plan, and only 18% (government funded 12%) of the urban population and 14% (government funded 13%) of the rural population was covered under any form of health insurance.

India's public health expenditures are lower than those of other middle-income countries. In 2012, they accounted for 4% of GDP, which is half as much as in China with 5.1%. In terms of public health spending per capita, India ranks 184th out of 191 countries in 2012. Patients' remaining costs represent about 58% of the total.[4] The remaining costs borne by the patient represent an increasing share of the household budget, from 5% of this budget in 2000 to over 11% in 2004-2005.[5] On average, the remaining costs of poor households as a result of hospitalization accounted for 140% of their annual income in rural areas and 90% in urban areas.

This financial burden has been one of the main reasons for the introduction of health insurance covering the hospital costs of the poorest.

Data Description:

The data at hand contains medical costs of people characterized by certain attributes.

Domain: Healthcare

Context:

Leveraging customer information is paramount for most businesses. In the case of an insurance company, attributes of customers like the ones mentioned below can be crucial in making business decisions. Hence, knowing to explore and generate value out of such data can be an invaluable skill to have.

Attribute Information:

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- children: Number of children covered by health insurance /Number of dependents
- smoker: Smoking
- region : the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges : Individual medical costs billed by health insurance.

Scope:

- Exploratory data analysis
- Data Pre-processing
- Training linear regression model with OLS method for prediction
- Tuning the model to improve the performance

Problem Statement:

Predict how much could be the insurance charges for a beneficiary based on the data provided using Linear Regression.

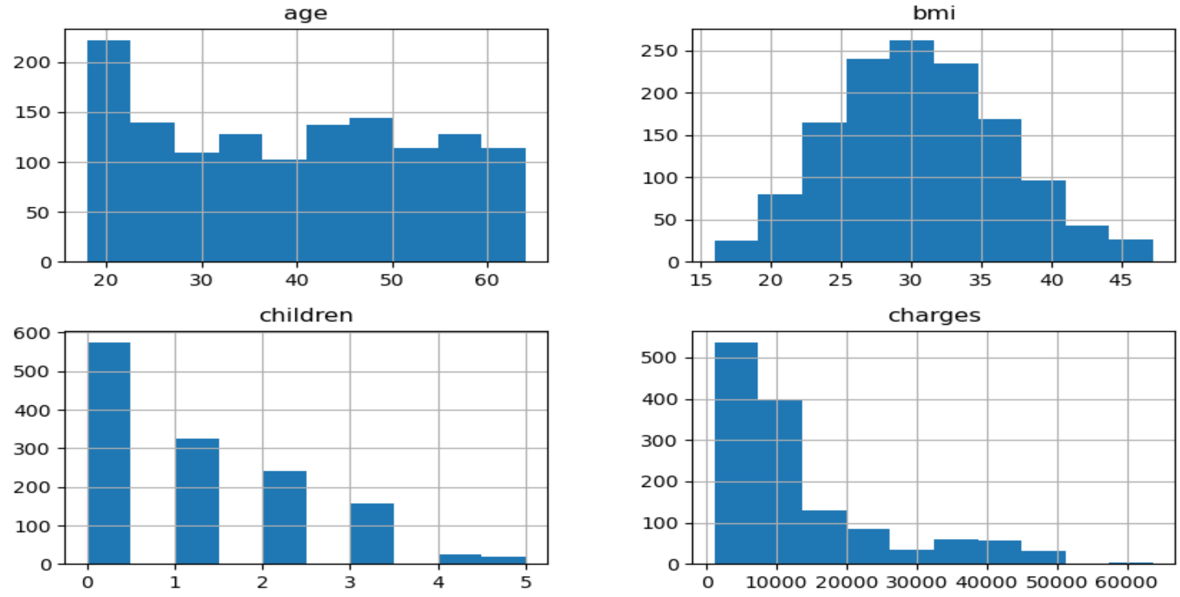
EDA and Business Implication

Data Information and Data Types

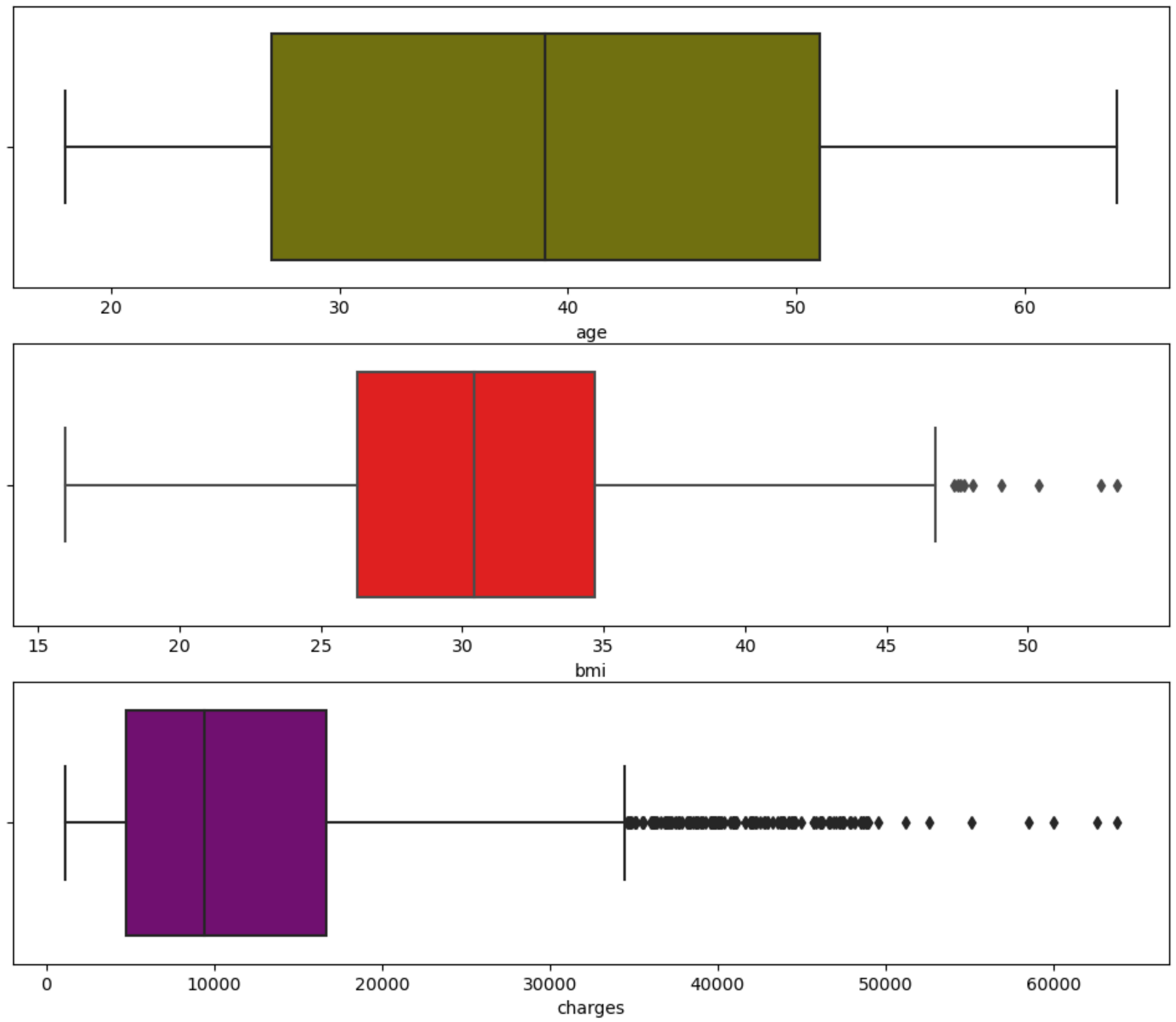
```
#   Column    Non-Null Count dtype

---  ------    --------------  -----

 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)
```

The data types are appropriately listed. Age, bmi, children, charges are numeric data type and the rest are object data type. We could also observe the presence of no Null values. There are no unwanted punctuation marks, spaces, prefixes or suffixes. The existing names holds good, renaming might not be necessary.

Visualizing and analyzing the data:

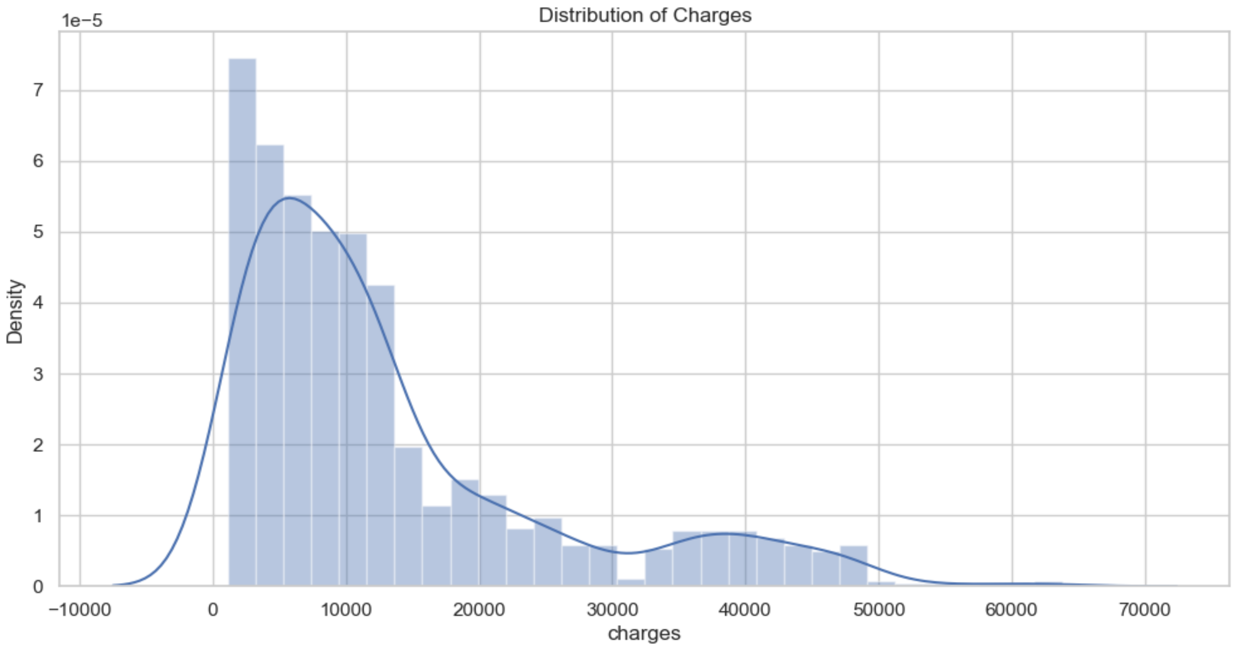


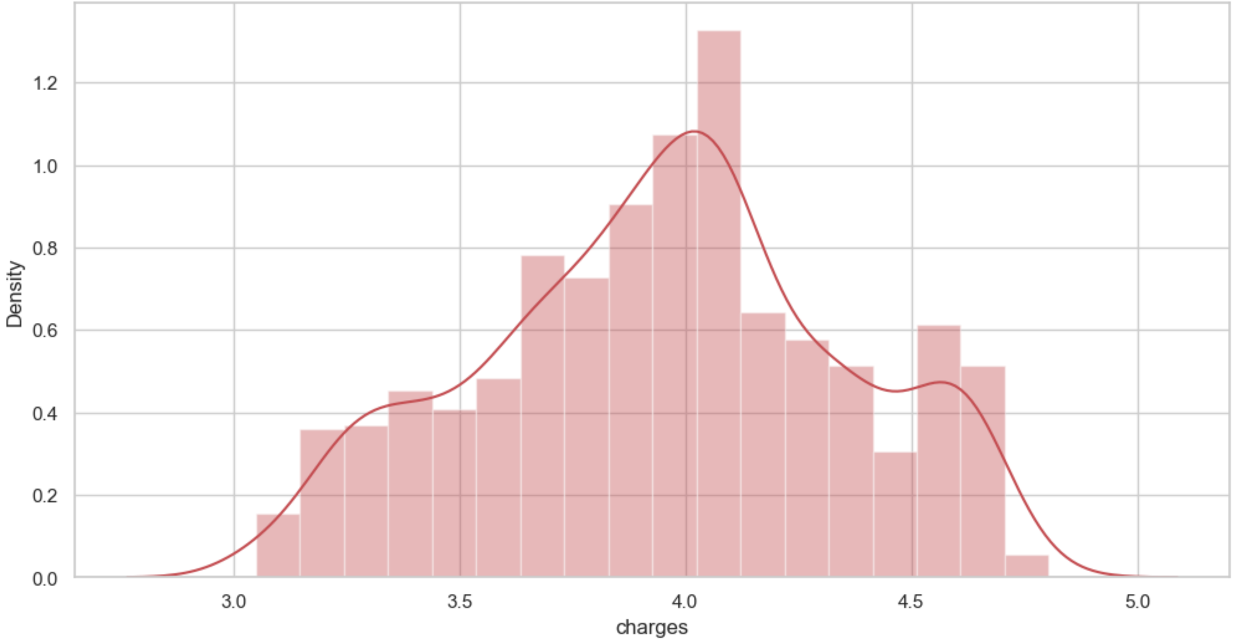
Checking the outliers:



We can observe a fair amount of positive skewness in charges and notice significant outliers in bmi and charges. Here the charges col is our dependent variable so we are not removing the outliers on charges col. Outliers can have huge adverse impact on the linear regression hence it is necessary to remove the outliers before we split the data into train and test

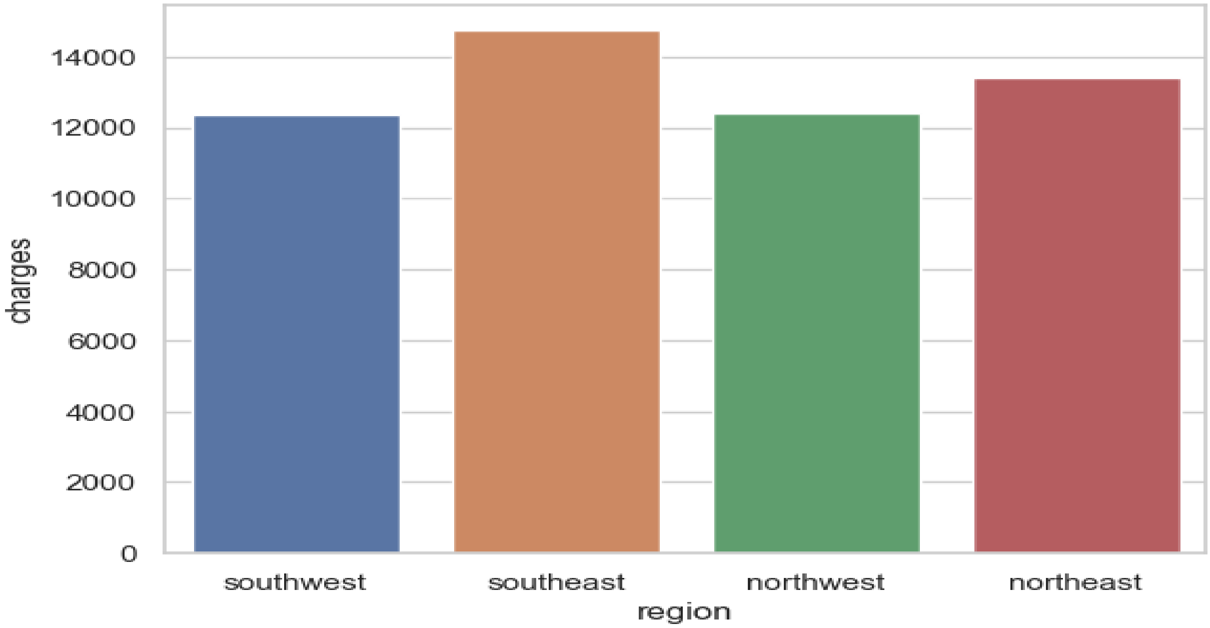
Histplot for Charges:



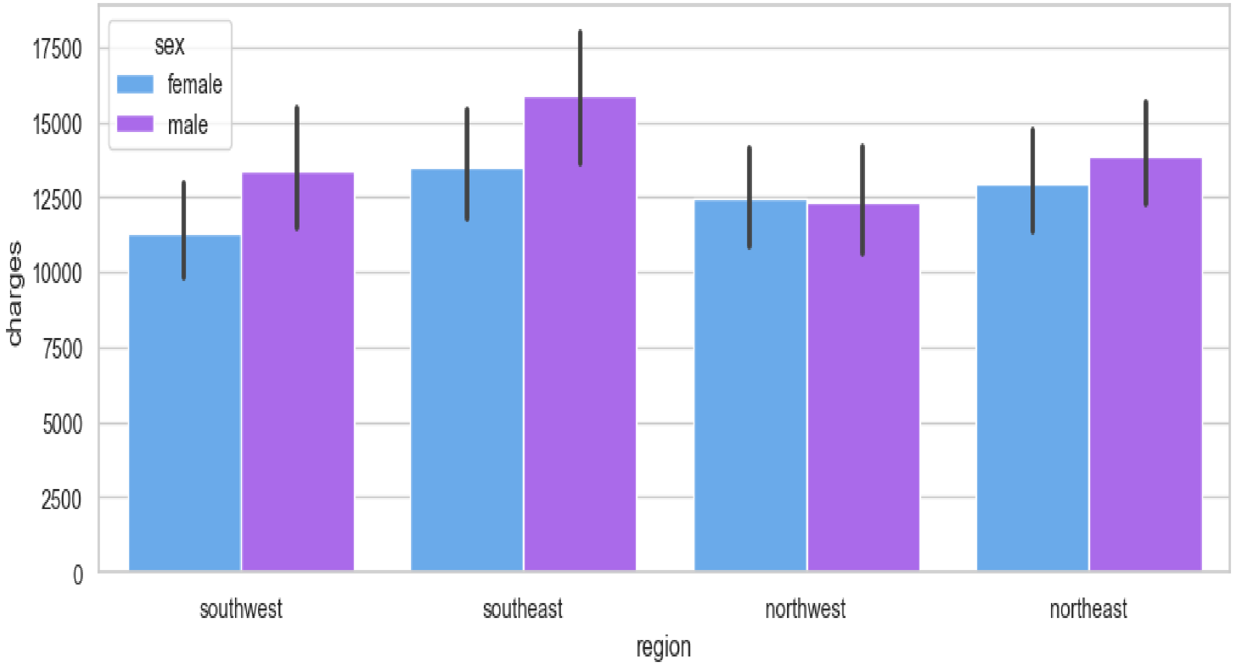


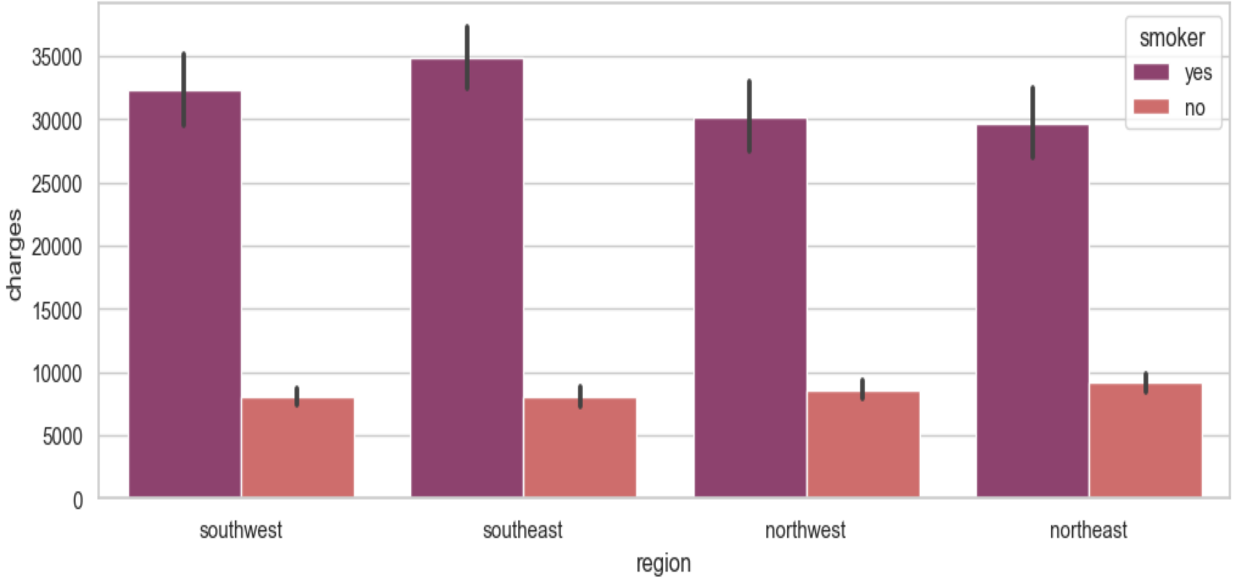
This above distribution is right-skewed and the below diagram is the normal distribution of charges.
To make it closer to normal we applied natural log"

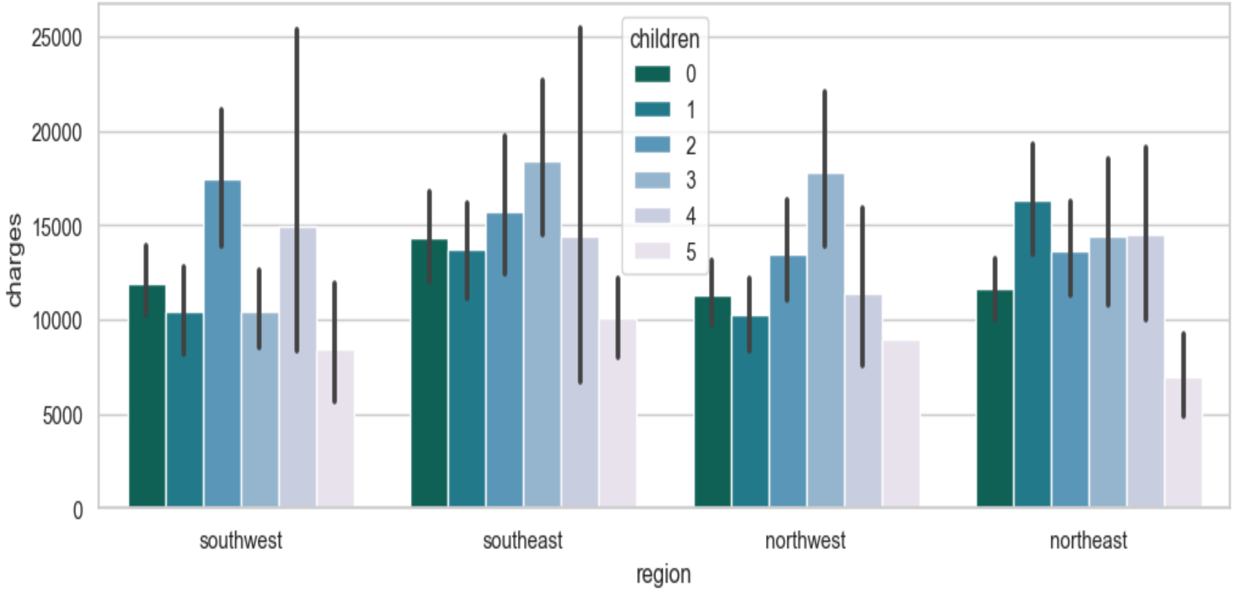
Now let's look at the charges by region:



So overall the highest medical charges are in the Southeast and the lowest are in the Southwest. Taking into account certain factors (sex, smoking, having children) let's see how it changes by region

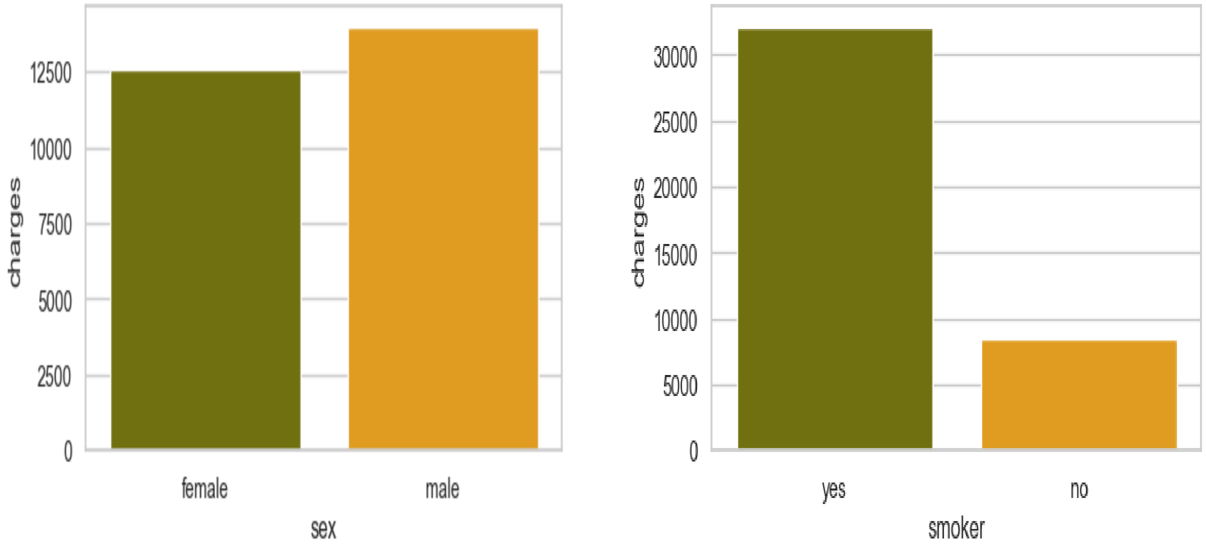


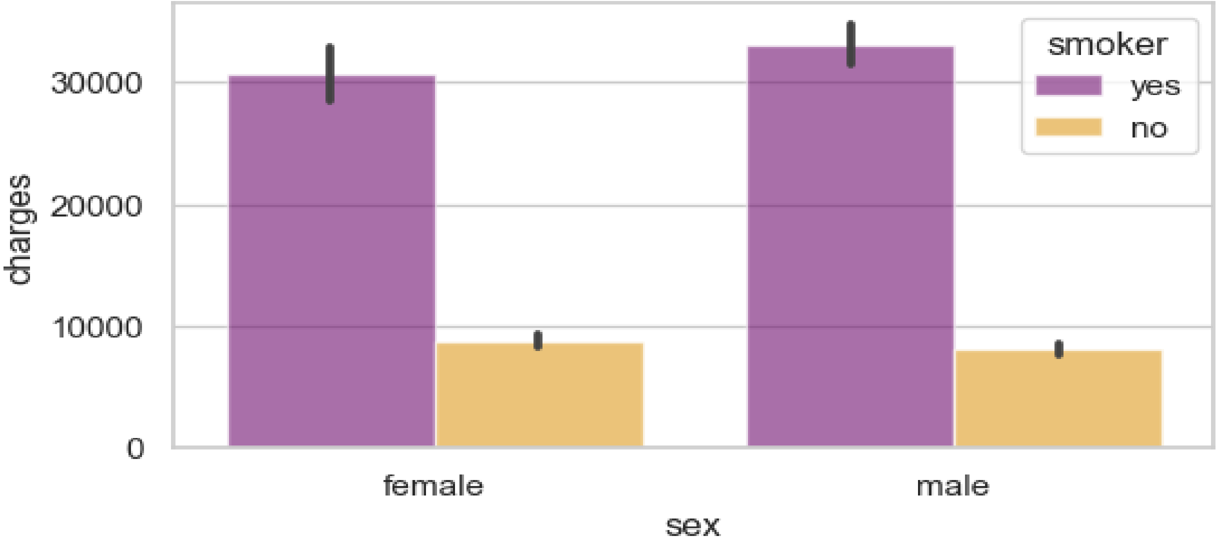




As we can see from these barplots the highest charges due to smoking are still in the Southeast but the lowest are in the Northeast. People in the Southwest generally smoke more than people in the Northeast, but people in the Northeast have higher charges by gender than in the Southwest and Northwest overall. And people with children tend to have higher medical costs overall as well.

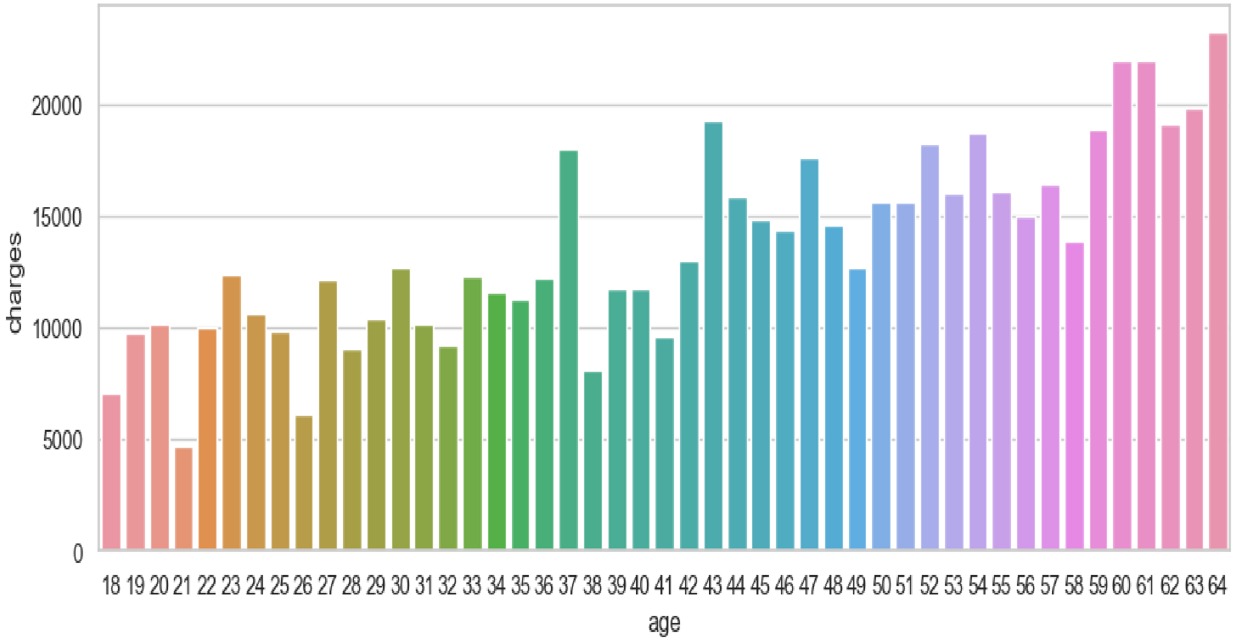
Now let's analyze the medical charges by gender and smoker:

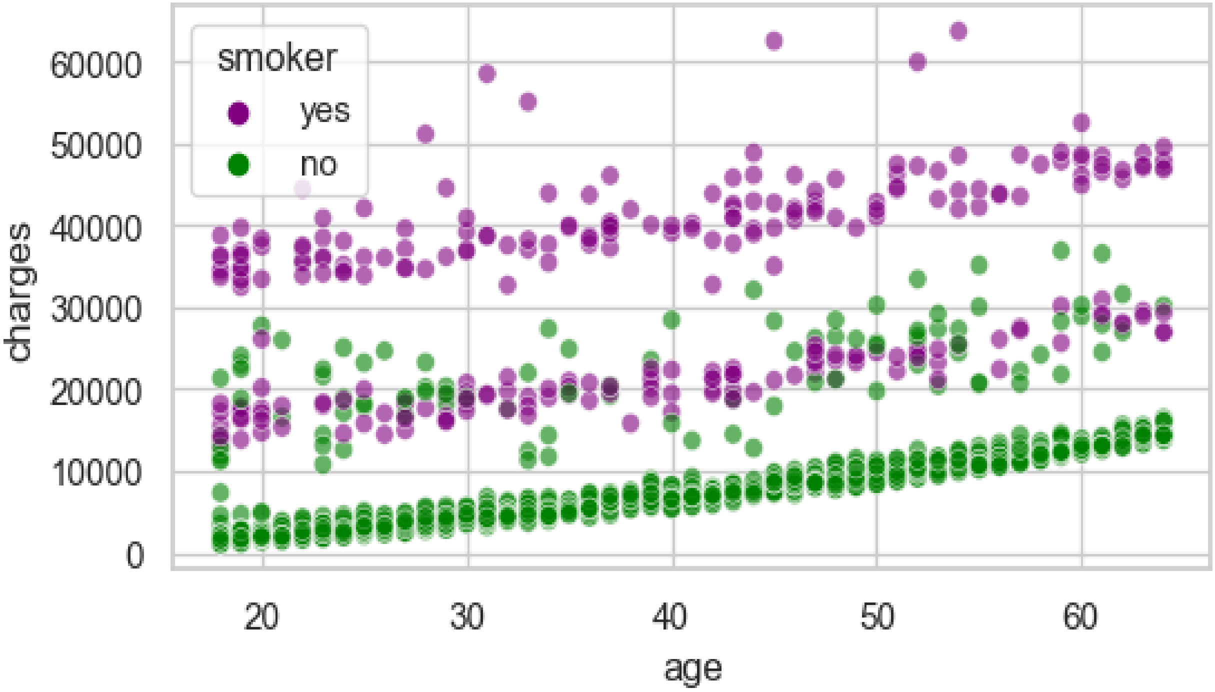


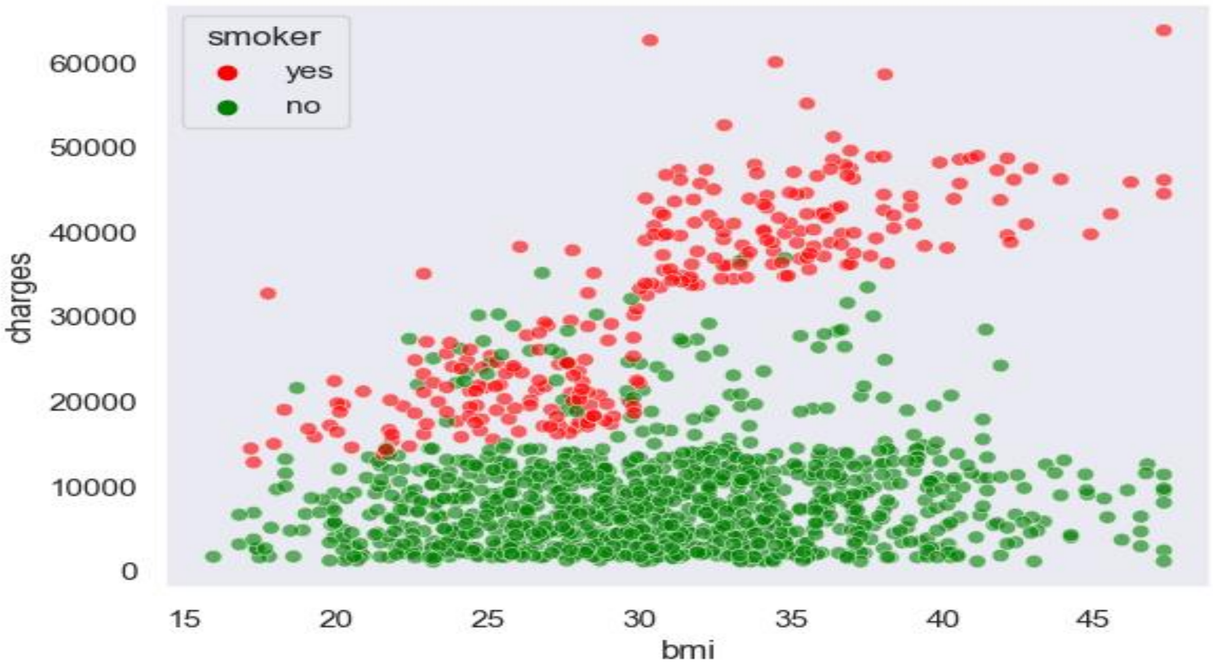


The insurance charges is higher for male compared to female and From the above barplot it is obvious than the insurance charges for smokers are very high. Hence the male smokers are having to face more charges for their insurance

Analyzing the medical charges by age and bmi according to the smoking factor:

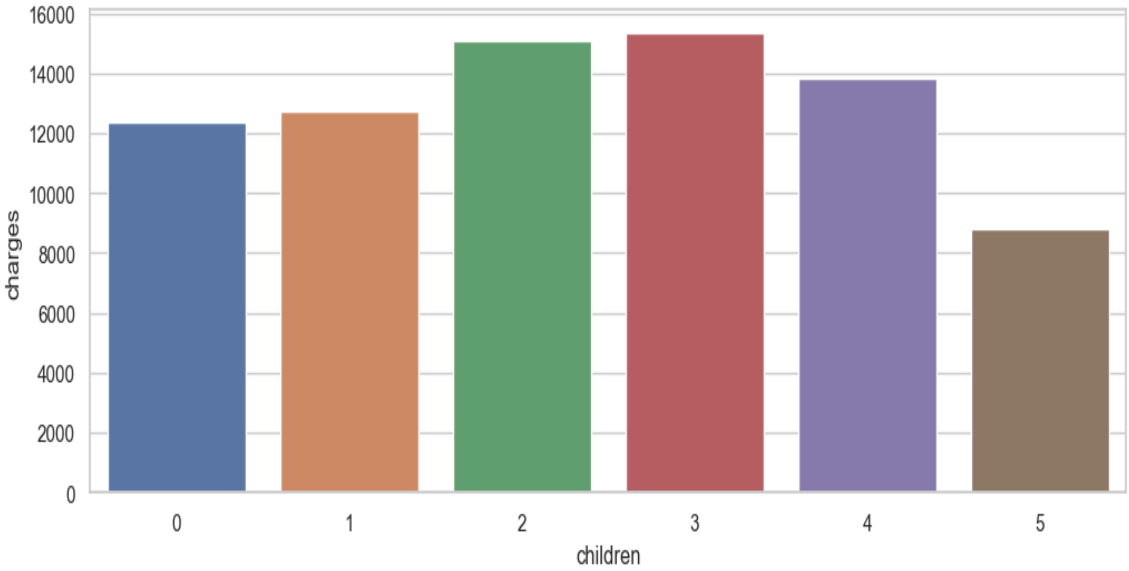


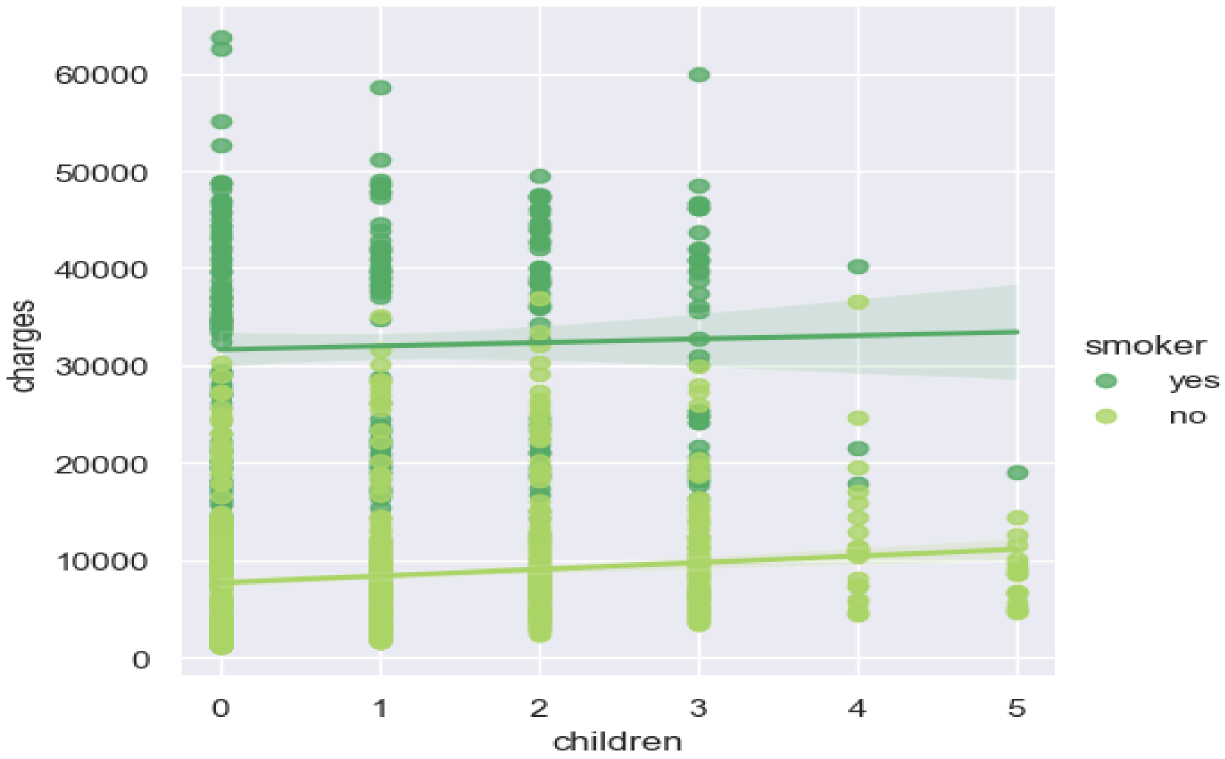


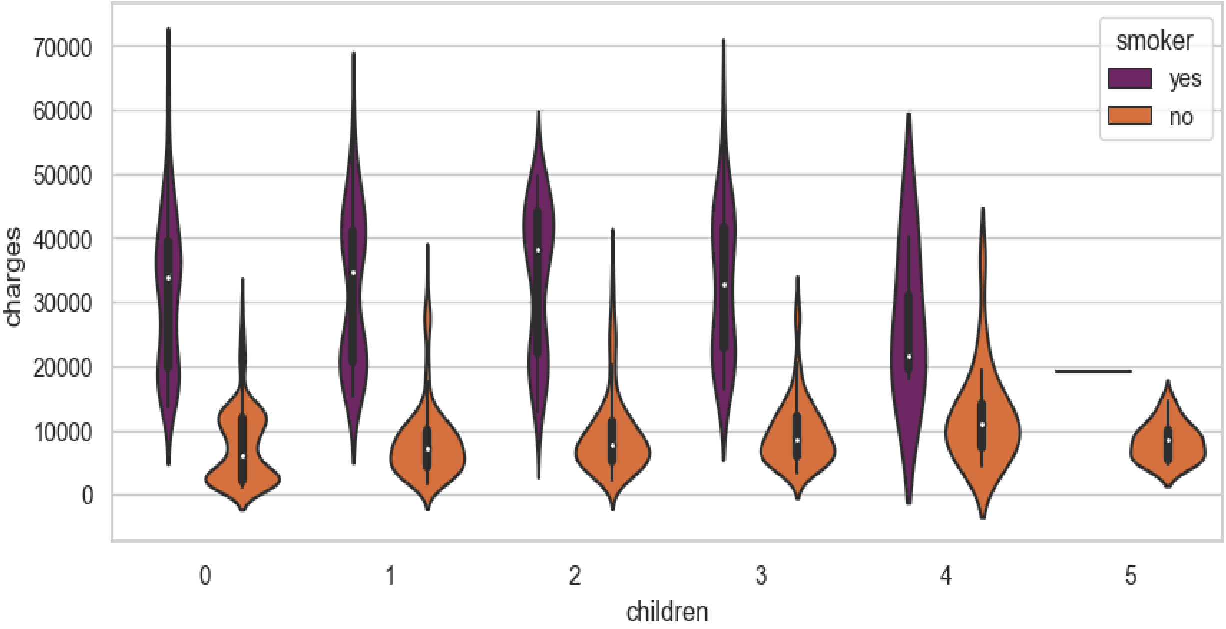


People of age more than 60 years and have smoking habits are having more insurance charges. But the insurance charges increase with the body mass index of smokers than the nonsmoker bmi.

Analyze the medical charges by children according to the smoking factor:

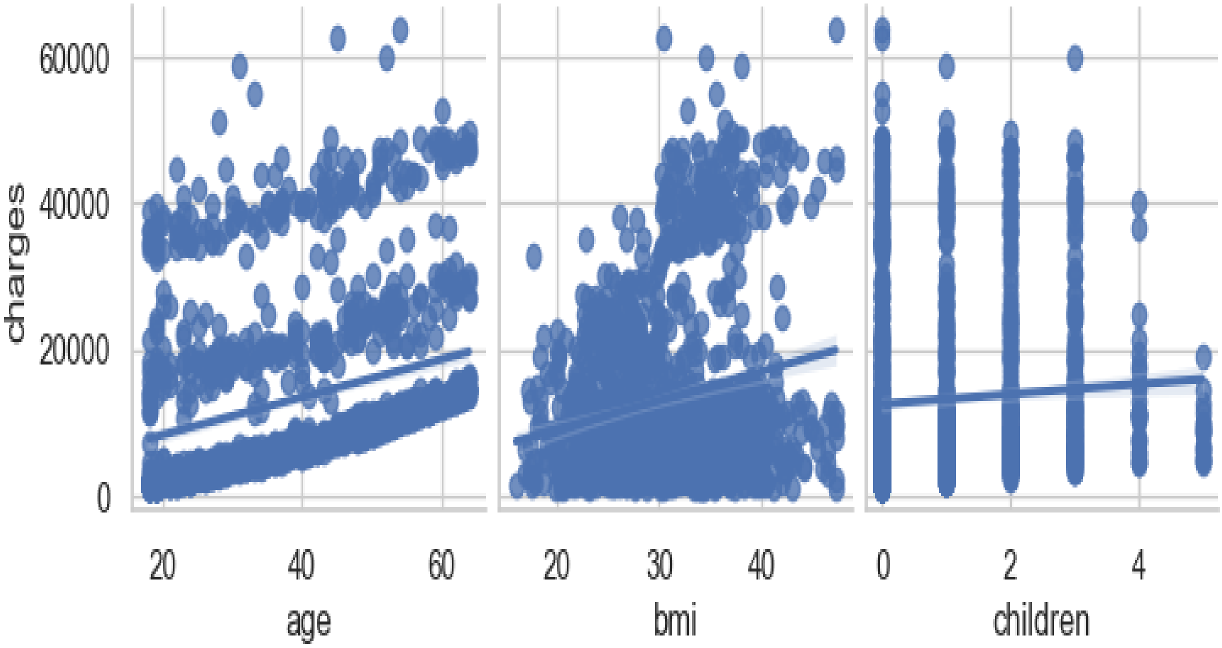



charges
60000
50000
40000
30000
20000
10000
0
0
1
2
3
4
5
children
smoker
yes
no


charges
70000
60000
50000
40000
30000
20000
10000
0
0
1
2
3
4
5
children
smoker
yes
no

People having more than 2 or 3 children have to pay large amount(charges). Also people who have children generally smoke less, which the following violinplots shows too.

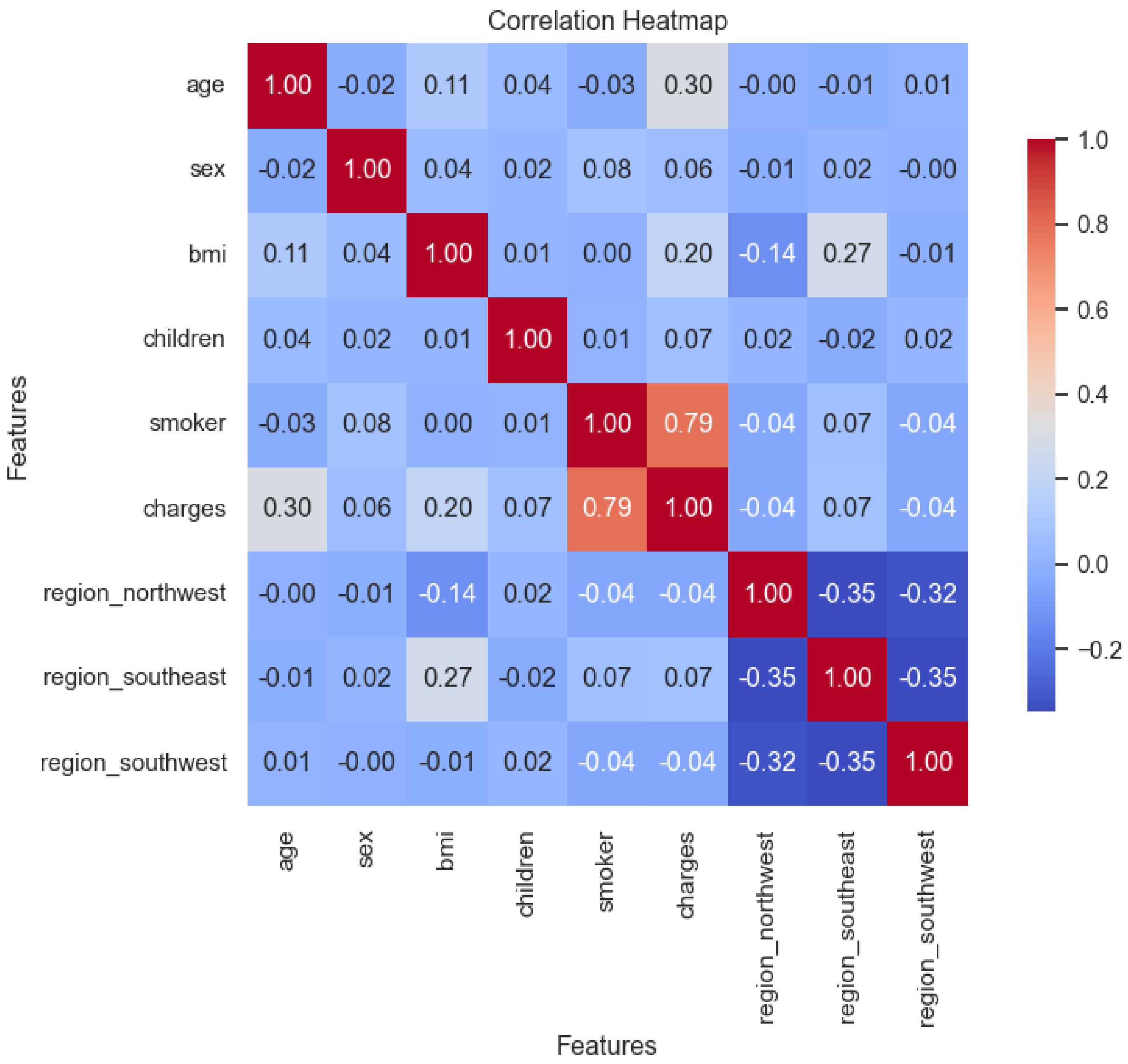
Pair plot of age, bmi, children w.r.t charges:



Smoking has the highest impact on medical costs, even though the costs are growing with age, bmi and children.

Multivariate Analysis:

Correlation using Heat Map:



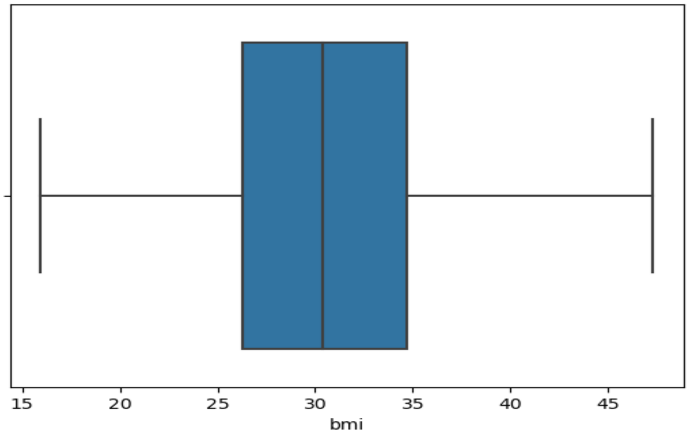
According to the correlation matrix, smoking has a very strong positive correlation with insurance costs, with a coefficient of 0.79. This suggests that smoking is a much stronger predictor of insurance costs than either age or BMI. However, all three features have a significant impact on insurance costs, and should be taken into consideration when developing a predictive model.

3. Data Cleaning and Pre-processing:

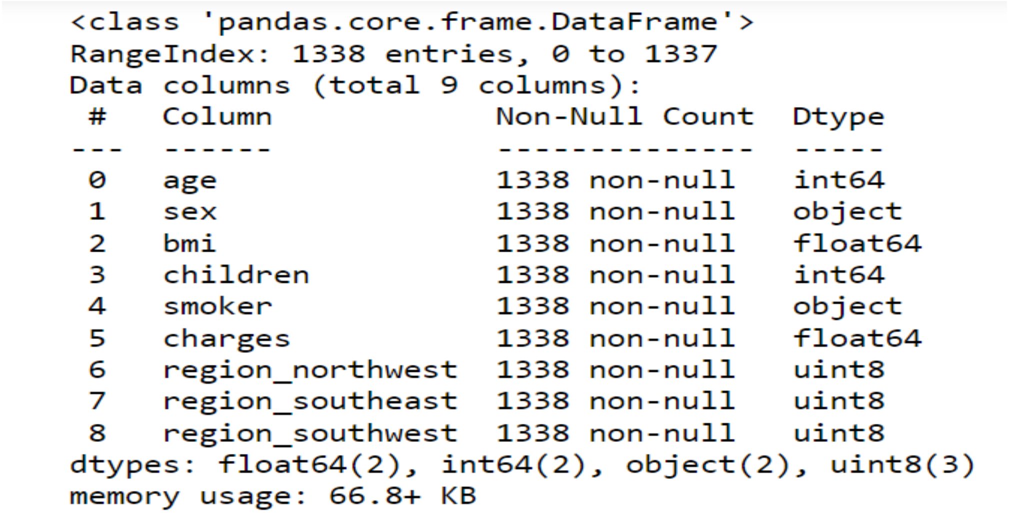
Approach used for identifying and treating missing values and outlier treatment.

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

As there is no null values present in the data so we don’t need to fill the missing values. But we treated the outliers of bmi so there is no outlier present in the bmi col.



Need for variable transformation:

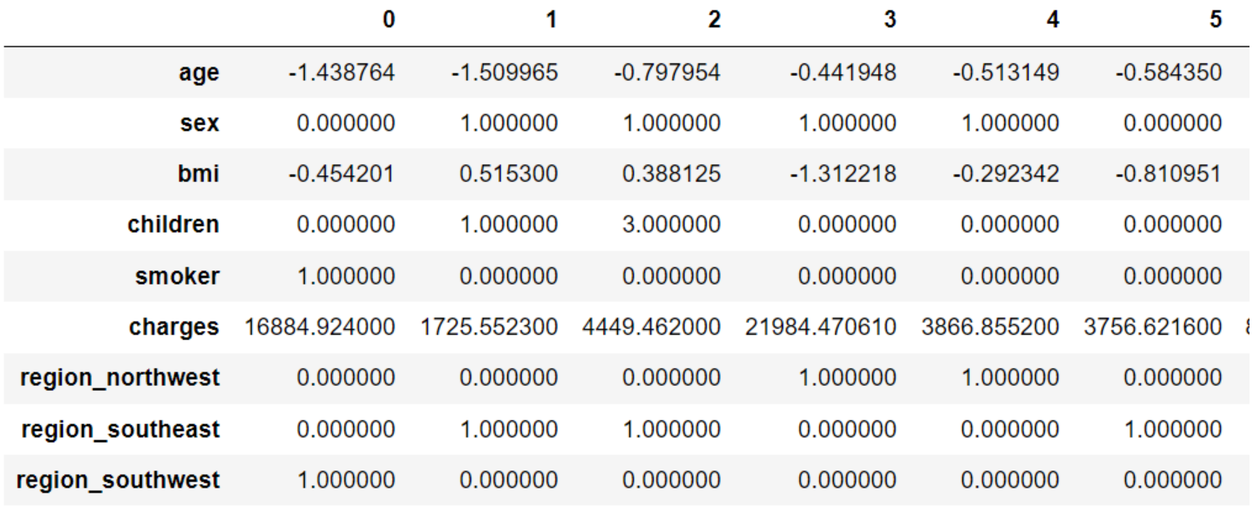


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 9 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   age               1338 non-null   int64  
 1   sex               1338 non-null   object 
 2   bmi               1338 non-null   float64
 3   children          1338 non-null   int64  
 4   smoker            1338 non-null   object 
 5   charges           1338 non-null   float64
 6   region_northwest  1338 non-null   uint8  
 7   region_southeast  1338 non-null   uint8  
 8   region_southwest  1338 non-null   uint8  
dtypes: float64(2), int64(2), object(2), uint8(3)
memory usage: 66.8+ KB
```

In this data preprocessing step, we will be converting object data types to numerical using two techniques - get dummies and label encoding. Get dummies creates binary columns for each unique value in the object column(region) , whereas label encoding assigns a numerical label to each unique value in the object column (sex) , (smoker).

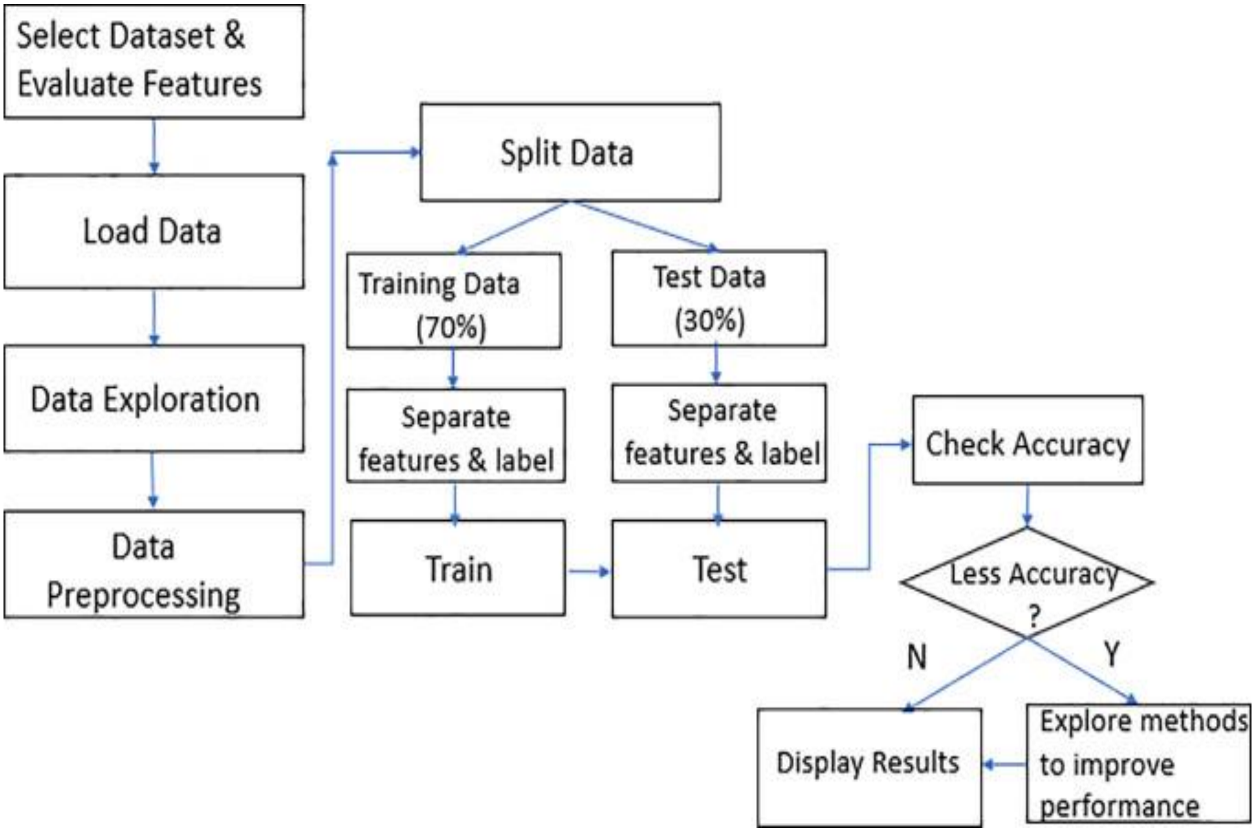
Additionally, we will be standardizing the continuous numerical data of (age) and (bmi) using z-score normalization. This will transform the age and bmi values to a common scale with a mean of 0 and standard deviation of 1. This helps to remove any potential bias or variation in the data due to differences in the original scales of the variables.

Overall, these steps will help prepare the data for machine learning models that require numerical inputs, while also ensuring that the numerical data is standardized and ready for further analysis.



	0	1	2	3	4	5
age	-1.438764	-1.509965	-0.797954	-0.441948	-0.513149	-0.584350
sex	0.000000	1.000000	1.000000	1.000000	1.000000	0.000000
bmi	-0.454201	0.515300	0.388125	-1.312218	-0.292342	-0.810951
children	0.000000	1.000000	3.000000	0.000000	0.000000	0.000000
smoker	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
charges	16884.924000	1725.552300	4449.462000	21984.470610	3866.855200	3756.621600
region_northwest	0.000000	0.000000	0.000000	1.000000	1.000000	0.000000
region_southeast	0.000000	1.000000	1.000000	0.000000	0.000000	1.000000
region_southwest	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000

Model building:



The objective of the study is to predict how much could be the insurance charges for a beneficiary based on the data. We have a dependent variable which is 'charges' column and one or more independent variables.

Hence, we will use Supervised learning method to build linear regression models. The data is split into train and test at 70:30 ratio. From the **Scikit learn** package, we shall begin with a simple linear regression model which will be our base model and build several other popular algorithms R-squared, MSE, RMSE, MAPE, Ridge and Lasso.

Linear Regression:

Linear regression can help to quantify the relationship between two variables, and understand how changes in one variable affect changes in another. If the goal is prediction, forecasting, or error reduction, then linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables as here we have independent variables such as age, bmi, children, etc and the dependent one is the charges. A linear regression model is a good fit for this data because it can help us understand the linear relationship between the independent variables and the dependent variable (charges). The aim of the model is to find the best linear relationship that describes the data.

In summary, our project utilizes linear regression to build a predictive model that estimates medical charges based on various input variables. We aim to develop a model that accurately predicts medical charges, which can be utilized by healthcare providers and policymakers to make informed decisions.

Train accuracy: 75%

Test accuracy: 74%

Train/Test	**Train**	**Test**
R-Squared	75%	74%
MSE	36464396.71	36742890.45
RMSE	6038.57	6061.59

A training accuracy of 75% means that the model has correctly predicted the target variable for 75% of the data in the training dataset. This suggests that the model has learned the patterns and relationships present in the data to a reasonable extent. However, it is important to note that a high training accuracy does not necessarily indicate that the model is good at generalizing to new, unseen data.

On the other hand, the test accuracy of 74% indicates that the model is able to predict the target variable correctly for 74% of the data in the testing dataset. This suggests that the model is performing decently well on new, unseen data. However, the fact that the test accuracy is slightly lower than the training accuracy could indicate that the model is slightly overfitting to the training data and not generalizing well to new data.

Overall, based on the given information, the model seems to be performing decently well in predicting the target variable. However, further analysis and evaluation are required to determine if the model is indeed a good fit for the data and to identify any potential areas for improvement.

MAPE= MEAN ABSOLUTE PERCENTAGE ERROR

MAPE for Train= 42%

MAPE for Test=41%

The MAPE is a commonly used metric to measure the accuracy of a predictive model. It represents the average percentage difference between the actual values and the predicted values of the target variable. In this case, since the MAPE is higher than 10%, it indicates that the model's predictions are not accurate. However, the fact that the MAPE is similar for both the train and test datasets indicates that the model is not overfitting to the train dataset.

Overall, further analysis may be required to determine the causes of the errors and to improve the accuracy of the model.

Here are some additional analyses that could be performed:

Cross-validation: Cross-validation can be used to check the stability of the model's performance**.**

Average MSE: 36975239.885134675
Average MAPE: 0.42376571529498264

The average MSE value of 36975239.885134675 indicates that the model has a relatively high error in its predictions. The MSE measures the average squared difference between the actual and predicted values, so a higher value means that the model is less accurate in its predictions.

The average MAPE value of 0.42376571529498264 indicates that, on average, the model's predictions are off by around 42% of the actual value. This means that the model's predictions are not very precise and could be improved

Regularization techniques:

Regularization techniques such as Ridge and Lasso regression can be used to improve the model's performance by reducing overfitting. This involves adding a penalty term to the loss function to constrain the values of the coefficients**.**

R-Squared	Train	Test
Ridge(L1)	0.7546	0.7409
Lasso(L2)	0.7546	0.7407

Based on the results of the Ridge and Lasso regressions, it can be observed that the R-squared values for both the train and test sets are similar to those obtained from the Linear Regression model. The R-squared values for both Ridge and Lasso are 75% for the train set and 74% for the test set. This indicates that both Ridge and Lasso models are performing similarly to the Linear Regression model in terms of explaining the variance in the dependent variable**.**

Overall, these values suggest that the model may need to be refined or a different model altogether may need to be used to obtain better predictions**.**

VIF= variance inflation factor:

Variance Inflation Factor (VIF) is a statistical measure used to identify multicollinearity in regression analysis. In regression analysis, multicollinearity occurs when there is a high correlation between two or more independent variables in a regression model. VIF is used to quantify the degree of multicollinearity among the independent variables.

Age= 1.0172517155172904

Sex= 1.7276329710022946

Bmi= 1.0894238500263063

Children= 1.6171413000944357

Smoker= 1.2349638036669421

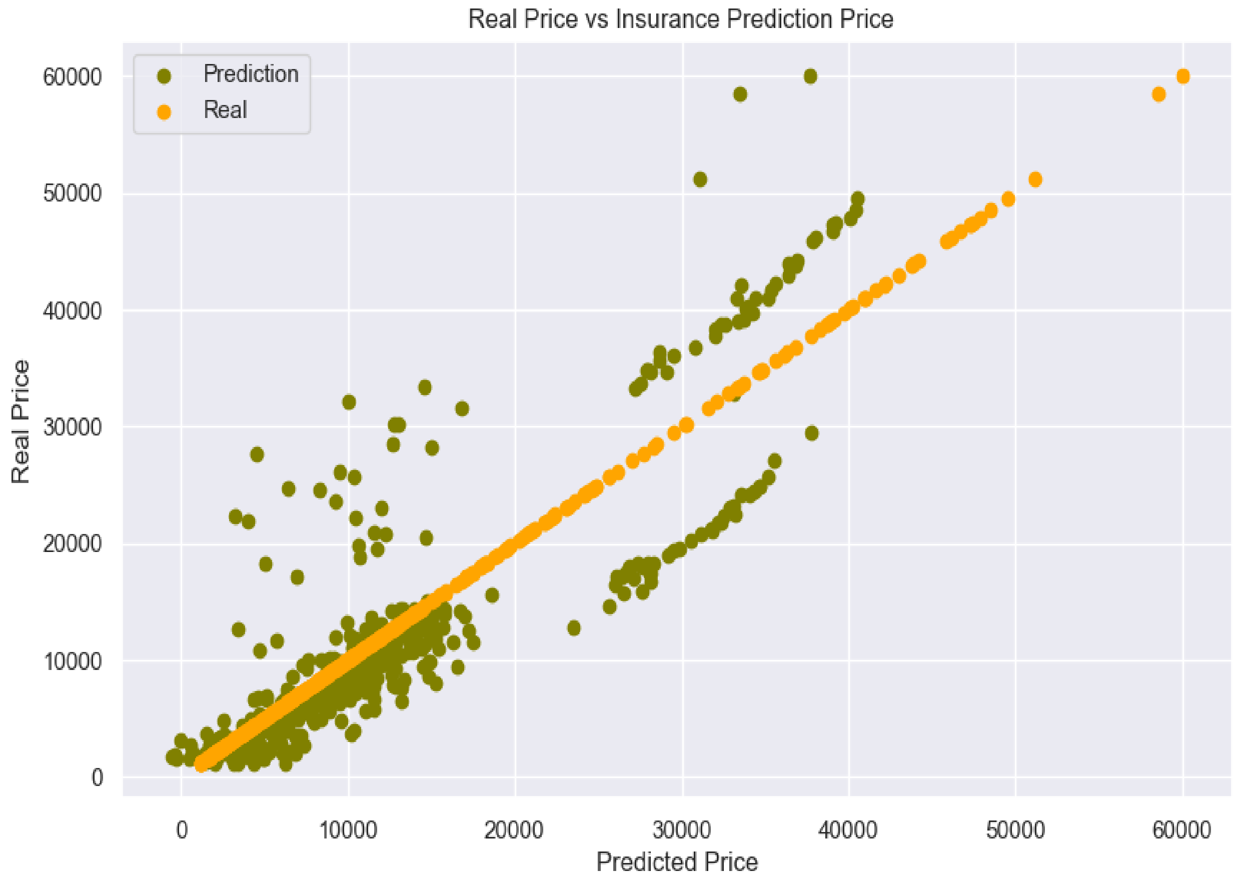
region_northwest= 1.3404356004807794

region_southeast= 1.4612111423444158

region_southwest= 1.3329706938450583

From the values you provided, none of the VIF values exceed 2, indicating that there is relatively low collinearity among the predictor variables in your regression model. Therefore, the model is likely to provide reliable estimates of the effects of the predictor variables on the outcome variable.

Comparison Between Real Insurance Price and Prediction Price:



Result:

- From the regression analysis, we find that region and gender do not bring significant difference on charges.
- Age, BMI, number of children and smoking are the ones that drive the charges
- Smoking seems to have the most influence on the medical charges.

Life expects you to embrace uncertainties and
ensure you're young & responsible than old and helpless

Conclusion

In conclusion, we have successfully built a linear regression model to predict the insurance charges based on various features such as age, bmi, sex, etc. The model achieved a reasonable level of accuracy with an R-squared value of 0.74 on the test set. The insights and trends identified in the exploratory data analysis have proven to be valuable in understanding the impact of different features on the insurance charges. The model can be used by insurance companies to predict the charges for potential customers and to identify the key drivers of insurance charges. Further work can be done to improve the model's performance and explore other machine learning algorithms for comparison.

Suggestion:

- Encouraging individuals to adopt a healthier lifestyle, such as quitting smoking, maintaining a healthy weight, and exercising regularly, to reduce their healthcare costs in the long run.

- Encouraging insurance companies to consider a more personalized approach to determining premiums, taking into account factors such as lifestyle, family history, and other relevant health information, in addition to age, gender, and geographic location.

- Suggesting that policymakers consider implementing programs to promote healthy living and provide resources to those who may not have access to them, such as low-income individuals or those living in underserved communities.

- Highlighting the importance of regular check-ups and preventive care in maintaining good health and reducing healthcare costs in the long run.