UNIVERSITY OF ECONOMICS AND LAW - VNUHCM

--------☙ 🕮 ❧--------

THE FINAL REPORT:

DATA MINING

Topic:

INSURANCE PREMIUM ANALYSIS REPORT

Lecture: Phan Huy Tam

Student: Luong Thi My Tam

MSSV: K204141929

Class: K20414C

TP Ho Chi Minh tháng 1 năm 2023

UNIVERSITY OF ECONOMICS AND LAW - VNUHCM

--------&📖&--------

THE FINAL REPORT:

DATA MINING

Topic:

INSURANCE PREMIUM ANALYSIS REPORT

Lecture: Phan Huy Tam

Student: Luong Thi My Tam

MSSV: K204141929

Class: K20414C

TP Ho Chi Minh tháng 1 năm 2023

TABLE OF CONTENTS

LIST OF TABLES/FIGURES

I. DESCRIBE THE DATA SET AND THE PROBLEM STATEMENT, THE NATURE OF THIS CASE:

1.1. The description of the data set:

Our data set consists of 1338 entries including:

- **Age:** Age of the customer in years (18-64). The mean age of the customer is 39 years old.
- **Sex:** Gender of the customer (male/female)
- **Bmi:** Body mass index, providing an understanding of body, weight 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. The mean of the customer's Bmi is 30.66
- **Children:** Number of children the customer is having (0-5). The mean is 1 people
- **Smoker:** indicate whether the customer is a smoker or a non-smoker (yes/no)
- **Region:** the customer's residential area in the US (northeast, southeast, southwest, northwest)
- **Charges:** Individual medical costs billed by health insurance. The mean of charges is about 13270.42.

1.2.The problem statement:

An insurance firm wants to give their customers different premiums based on their nature of risk.

1.3.The nature of this case:

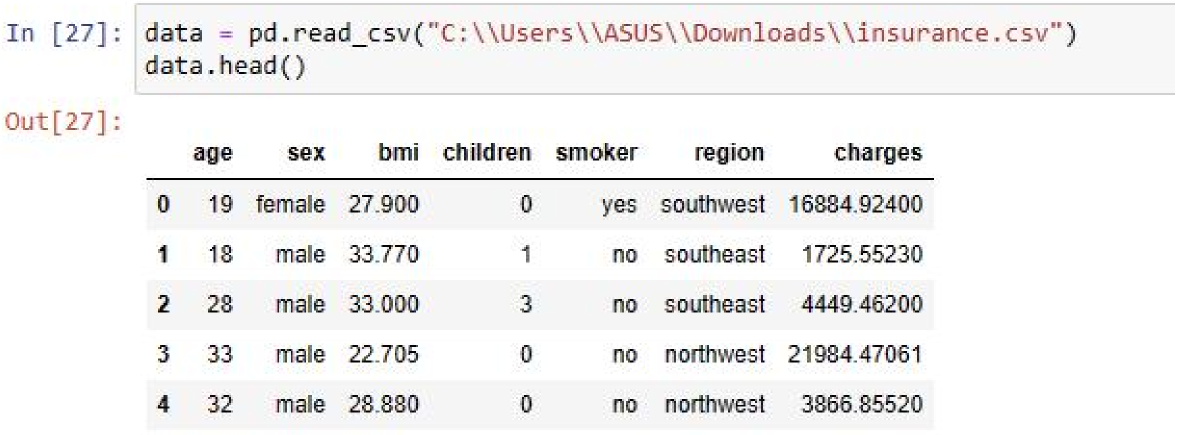
The essence of the problem is to estimate the insurance premium for each customer because each customer will have a different level of risk, resulting in different insurance premiums for each person. The higher the risk, the higher the premium. For example, The premiums for the elderly are usually higher than for the young. We will determine the value in the “charges” column using columns such as age, gender, bmi, number of children, smoking and region. Using algorithms to train the model and use the most optimal model to forecast premium. If we can do the same for historical data, then we can also estimate fees for new customers, just by asking for information like their age, gender, BMI, number of children, smoking habits and their region.

II. THE DEFECTS OR ISSUES OF DATA AND THE SOLUTION:

2.1.Finding defects or issues of data:

- **Input Data**

Table 2.1: Showing the data set

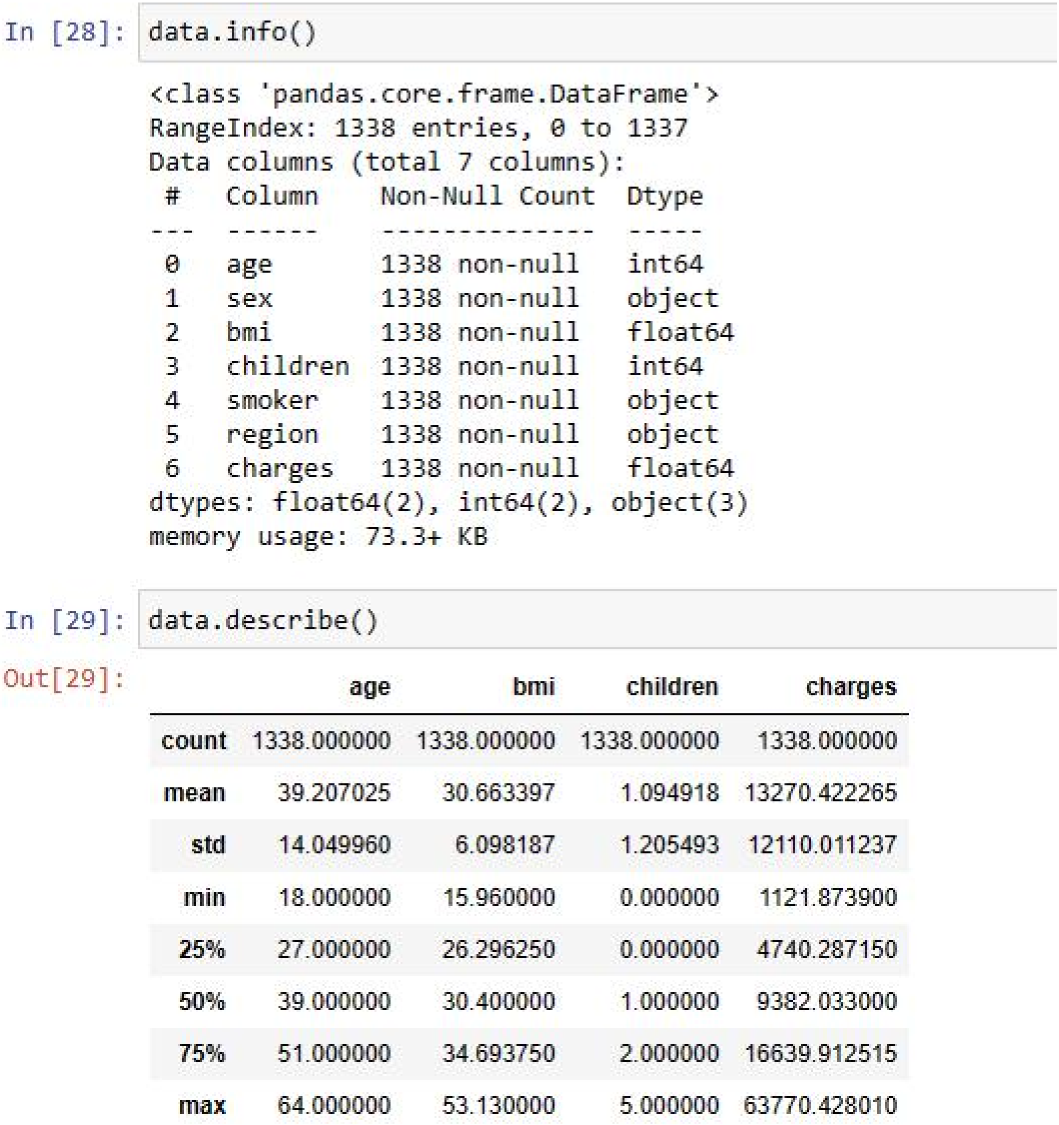


```
In [27]: data = pd.read_csv("C:\\Users\\ASUS\\Downloads\\insurance.csv")
         data.head()
```

Out[27]:

	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

Table 2.2: The Descriptive statistics of the data set



```
In [28]: data.info()

<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

In [29]: data.describe()
```

Out[29]:

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

Table 2.2 shows "age", "bmi" and "charges" are numerics, whereas "sex", "smoker" and "region" are categories. However, “children” contains discrete values, so this column is categorical (object datatype).

- **The statistics for the numerical columns**

The Age column is ranging from 18 to 64 ,with mean of 38.2 and standard deviation of 14.04

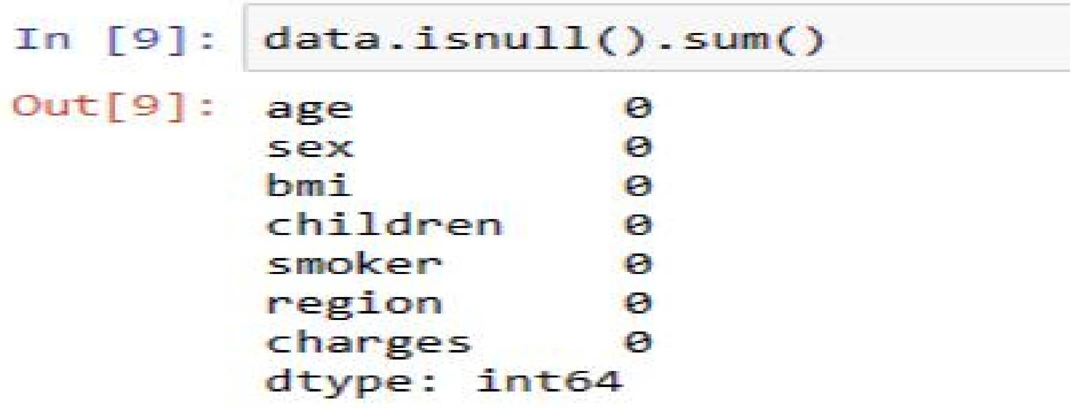
The BMI column is ranging from 15.96 to 53.13 , with mean of 30.6 and standard deviation of 6.09

The Charges column is ranging from 1121 to 63770 , with mean of 13270 and standard deviation of 12110

The Age column, the BMI column and the Charges column have Mean > Median implies the data is right skewed

- **Checking missing value**

Table 2.3: Checking missing values



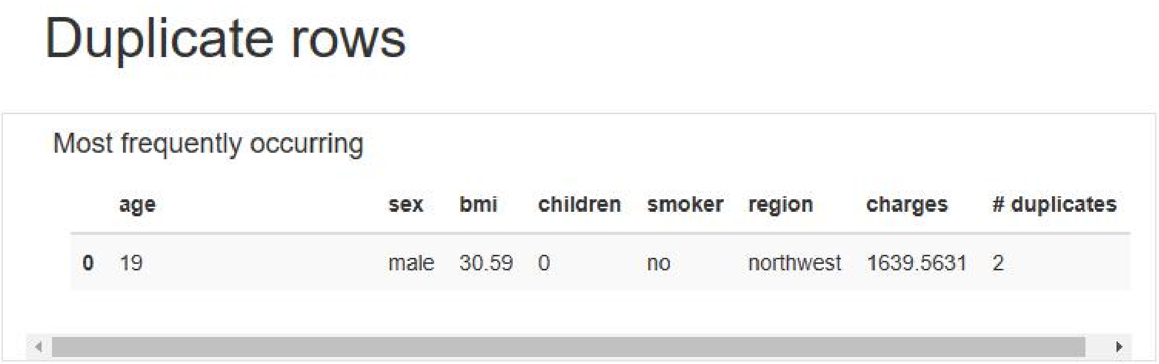
```
In [9]:  data.isnull().sum()

Out[9]:  age         0
         sex         0
         bmi         0
         children    0
         smoker      0
         region      0
         charges     0
         dtype: int64
```

As we can see in the table above, there is no missing value in the data set.

- **Checking Duplicate**

Table 2.4: Checking Duplicate



Duplicate rows

Most frequently occurring

	age	sex	bmi	children	smoker	region	charges	# duplicates
0	19	male	30.59	0	no	northwest	1639.5631	2

From Table 2.4, we know the dataset has a repeating row.

- **Outlier**

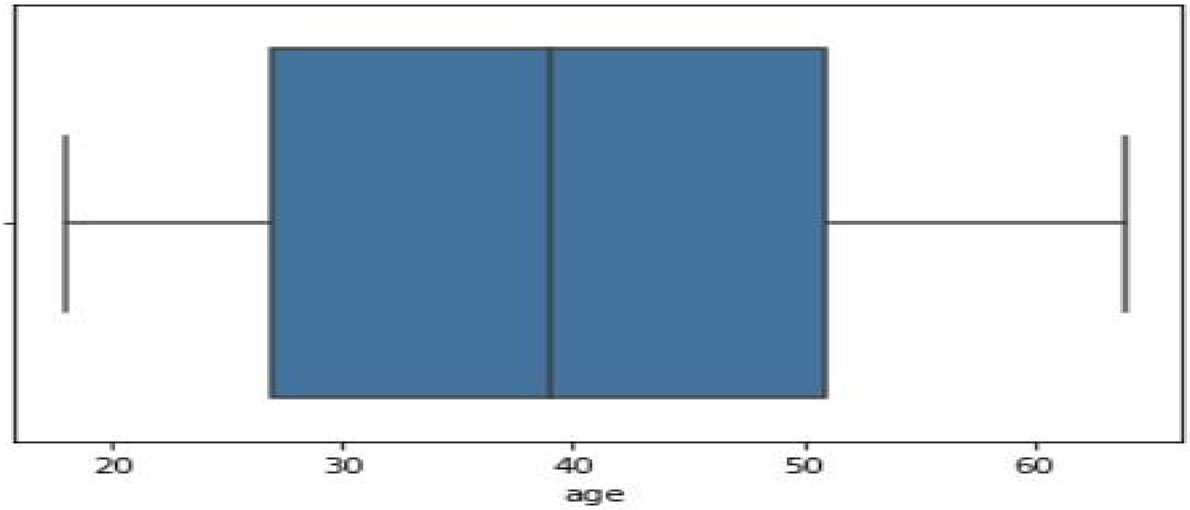


Figure 2.5: Boxplot of the age column

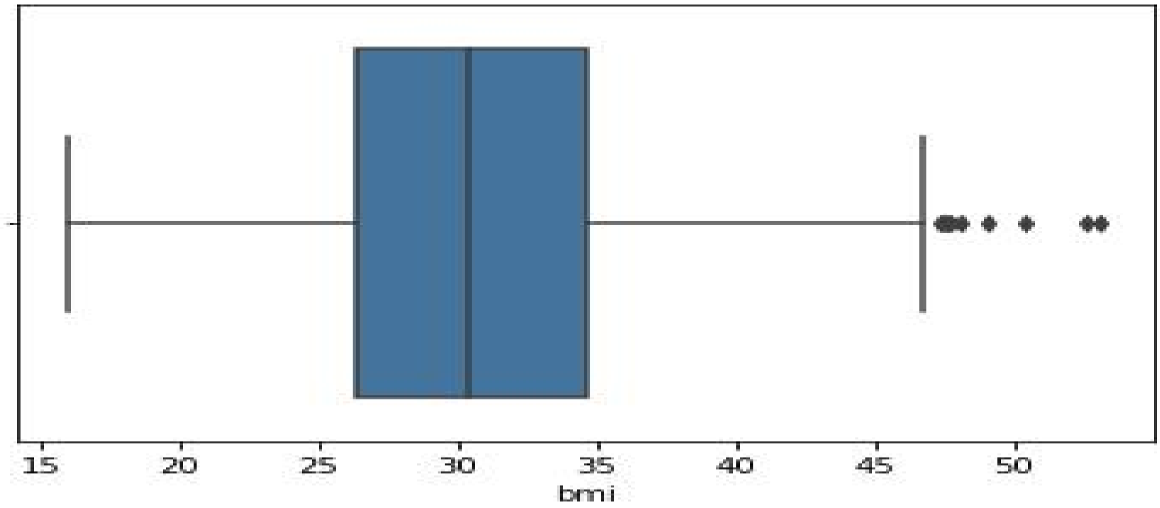


Figure 2.6: Boxplot of the BMI column

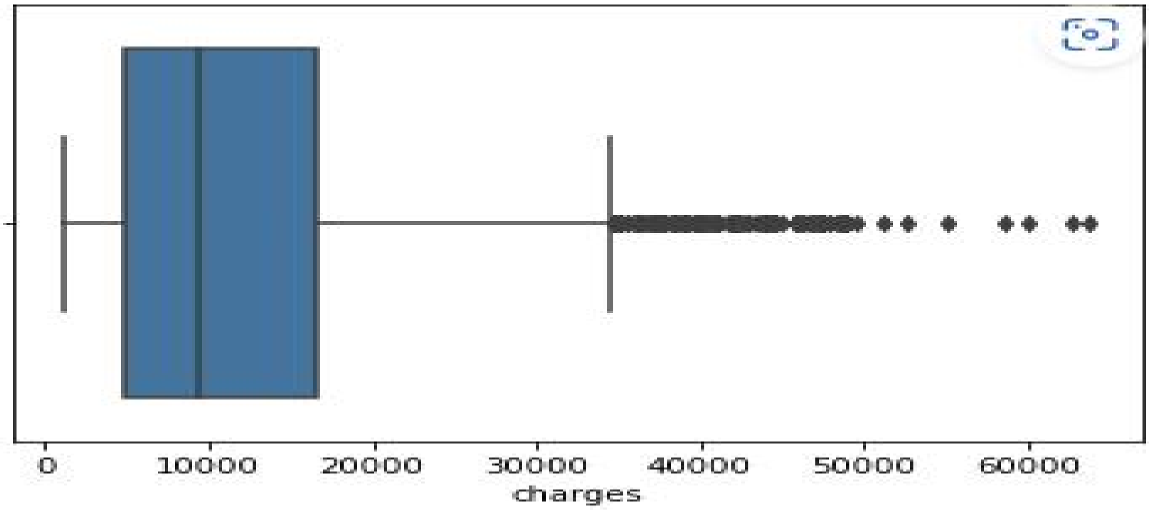


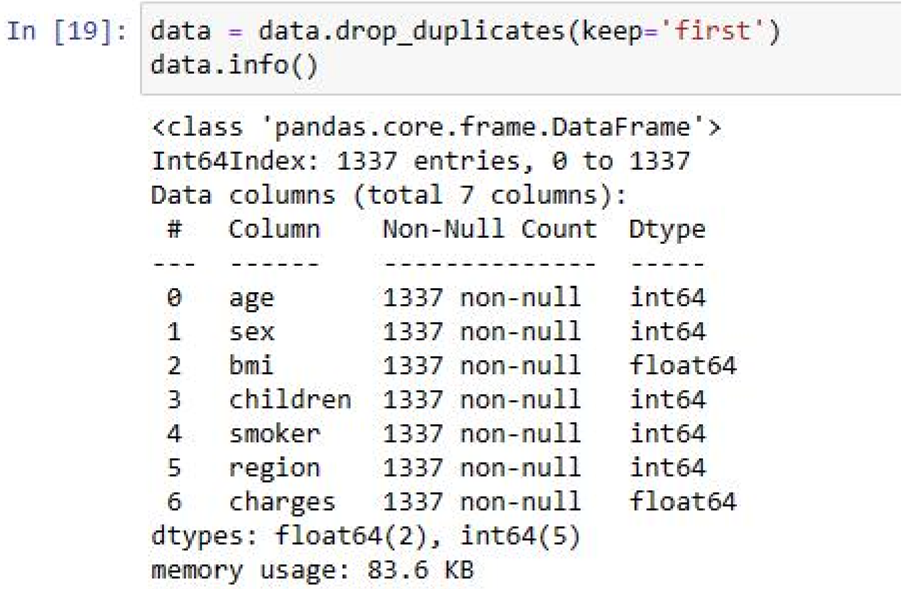
Figure 2.7: Boxplot of the Charges column

In summary, the Data set has been cleaned quite well. There is no missing value but there is a duplicate that needs to be processed and we need to convert the Sex column, the Smoker column and the Region column to numeric and the Children column should be converted to category (object datatype). Besides, the BMI column and the Charges column contain outliers and the Charges column appears to be significantly skewed because the median is much lower than the maximum value.

2.2.The solution:

For Duplicate, It’s unlikely that two people have the same age, sex, BMI, and children from the same region, both non-smokers, and have exactly the same medical charges. We can drop this duplicated row. After that, the data set will have only 1337 entries in the table below.

Table 2.8: The result after dropping duplicated row

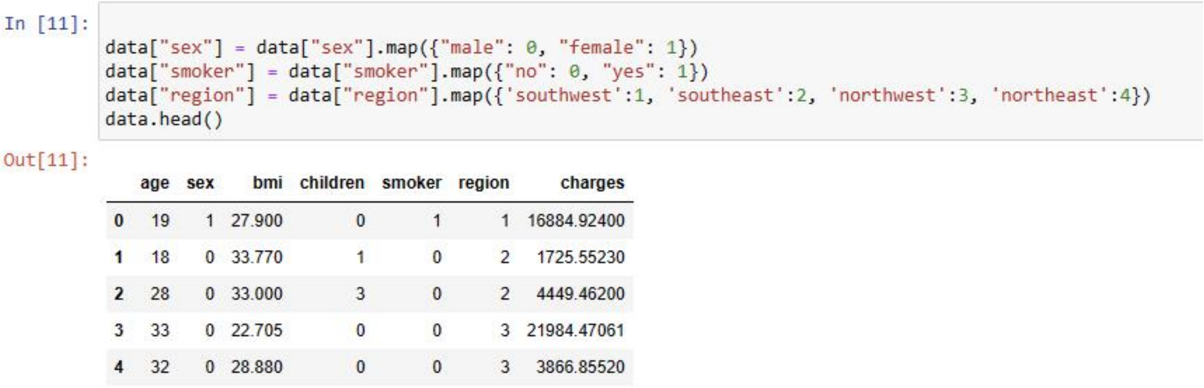


```
In [19]: data = data.drop_duplicates(keep='first')
         data.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 1337 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       1337 non-null   int64  
 1   sex       1337 non-null   int64  
 2   bmi       1337 non-null   float64
 3   children  1337 non-null   int64  
 4   smoker    1337 non-null   int64  
 5   region    1337 non-null   int64  
 6   charges   1337 non-null   float64
dtypes: float64(2), int64(5)
memory usage: 83.6 KB
```

Next, we will convert Sex column, Smoker column and Region column to numeric. For the Sex column, if “male” will assign 0 and “female” will assign 1. For the Smoker column, if “no” will assign 0 and “yes” will assign 1. And the Region column, “southwest” will assign 1, “southeast” will assign 2, “northwest” will assign 3, and “northeast” will assign 4.

Table 2.9: The result after data transformation of columns



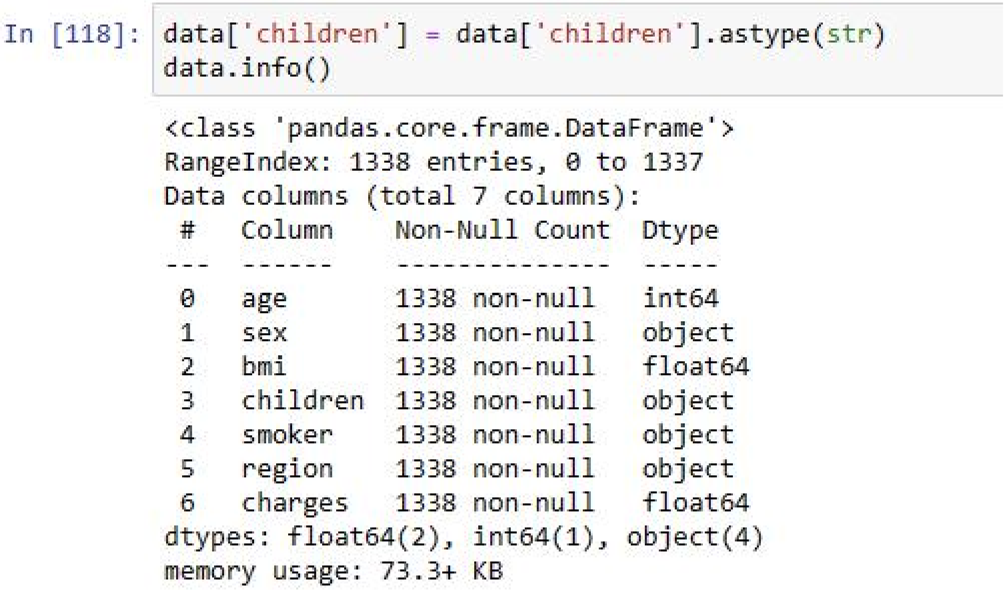
```
In [11]:
data["sex"] = data["sex"].map({"male": 0, "female": 1})
data["smoker"] = data["smoker"].map({"no": 0, "yes": 1})
data["region"] = data["region"].map({'southwest':1, 'southeast':2, 'northwest':3, 'northeast':4})
data.head()
```

Out[11]:

	age	sex	bmi	children	smoker	region	charges
0	19	1	27.900	0	1	1	16884.92400
1	18	0	33.770	1	0	2	1725.55230
2	28	0	33.000	3	0	2	4449.46200
3	33	0	22.705	0	0	3	21984.47061
4	32	0	28.880	0	0	3	3866.85520

Finally, we will change the data type of the children column.

Table 2.10: The result after changing the data type for the Children column



```
In [118]: data['children'] = data['children'].astype(str)
          data.info()
```

```
<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   object 
 4   smoker    1338 non-null   object 
 5   region    1338 non-null   object 
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(1), object(4)
memory usage: 73.3+ KB
```

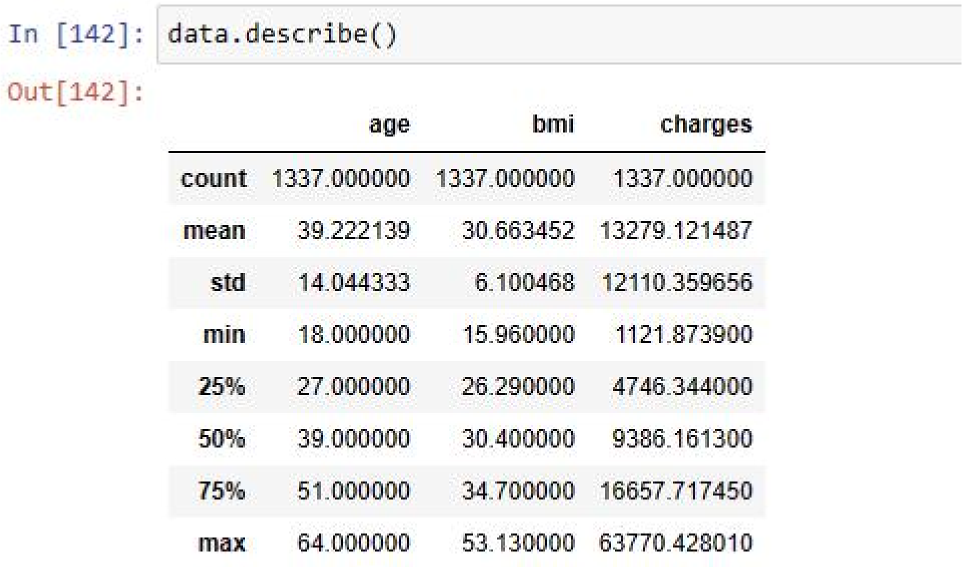
III. WHAT KIND OF MODEL COULD BE USED IN THIS CASE? EXPLAIN:

The goal of the problem is to estimate the value in the charges column for each customer through other data columns. As we know, the charges column is given and the output values as continuous data. Therefore, we see predictive modeling is the most suitable model. Because if it is cluster analysis, it will be unsupervised learning and will not know in advance what the label is. And classification analysis is supervised learning and given labels in advance, but the output data is discrete data. Anomaly detection is used for problems that detect anomalies. Anomalies here can refer to deviations, anomalies, noises and outliers. This technique can be used in many different fields. Such as intrusion detection or health monitoring, so it is not also suitable. Finally, association analysis is concerned with understanding the relationships between variables in the database, so this model is also not suitable. Based on the above reasons, it is emphasized that predictive modeling is the most suitable model and specifically, a regression model will be applied to solve this problem because the forecasting technique when used with regression will be used to forecast continuous values.

IV. PERFORM DATA EXPLORATORY ANALYSIS:

4.1.The Descriptive statistics:

Table 4.1: The Descriptive statistic of numerical columns



```
In [142]: data.describe()
Out[142]:
               age          bmi       charges
count  1337.000000  1337.000000   1337.000000
mean     39.222139    30.663452  13279.121487
std      14.044333     6.100468  12110.359656
min      18.000000    15.960000   1121.873900
25%      27.000000    26.290000   4746.344000
50%      39.000000    30.400000   9386.161300
75%      51.000000    34.700000  16657.717450
max      64.000000    53.130000  63770.428010
```

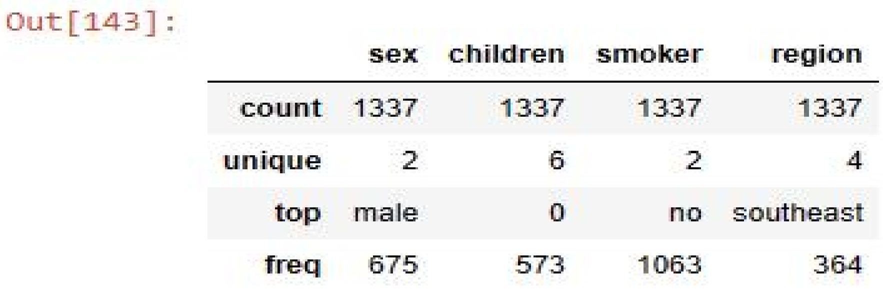
The Age column is ranging from 18 to 64 ,with mean of 39.2 and standard deviation of 14.04.

The BMI column is ranging from 15.96 to 53.13 , with mean of 30.6 and standard deviation of 6.1

The Charges column is ranging from 1121 to 63770 , with mean of 13279 and standard deviation of 12110

The Age column, the BMI column, the Children column and the Charges column have Mean > Median implies the data is right skewed

Table 4.2: The statistic of four categorical columns

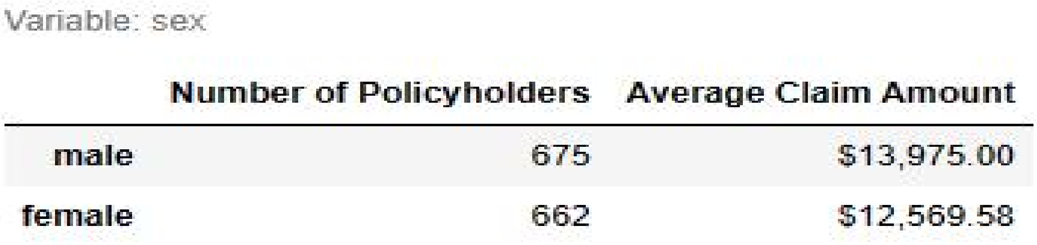


Out[143]:

	sex	children	smoker	region
count	1337	1337	1337	1337
unique	2	6	2	4
top	male	0	no	southeast
freq	675	573	1063	364

From table 12, we can see that the sex column and the smoker column have two attributes (male/female, yes/no) and the children column will have 6 attributes (from 0 to 5 children) and the region column has 4 attributes (southwest, southeast, northwest, northeast). Besides, men will account for more than women with 675 people. Customers without children are the largest group of policyholders with 573 people. Non-smokers accounted for the majority when compared to smokers with 1063, and customers from the southeast were the most engaged with 364.

Table 4.3: The statistic of the sex column

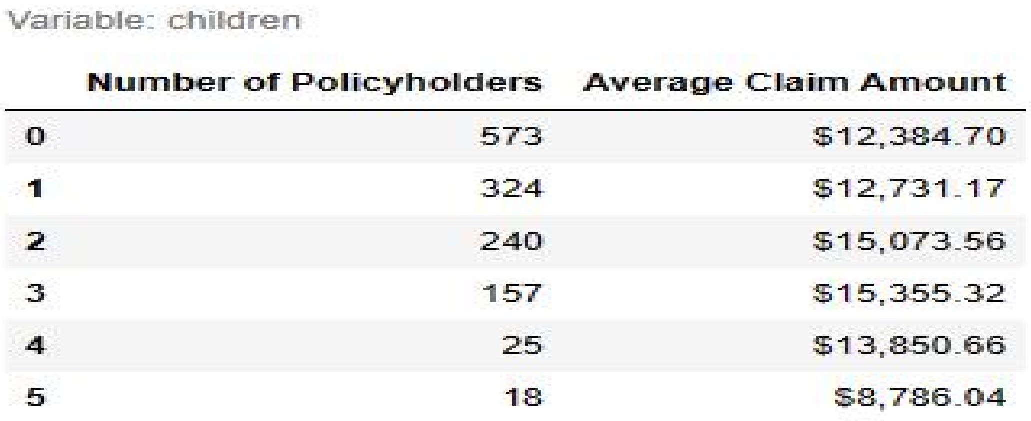


Variable: sex

	Number of Policyholders	Average Claim Amount
male	675	$13,975.00
female	662	$12,569.58

From Table 13, it is easy to see that there is not a big difference between the number of male and female customers. And the Average claim amount for men is higher than for women and this is true in fact, Research says that women generally live longer than men. This means that life insurance companies interpret men as more at-risk than women. Therefore, insurance premiums for women are also slightly lower than for men.

Table 4.4: The statistic of the children column

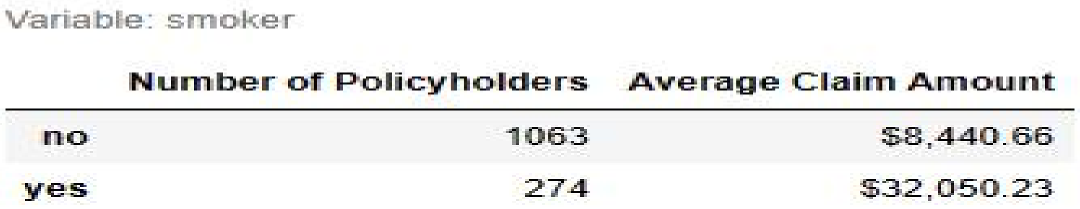


Variable: children

	Number of Policyholders	Average Claim Amount
0	573	$12,384.70
1	324	$12,731.17
2	240	$15,073.56
3	157	$15,355.32
4	25	$13,850.66
5	18	$8,786.04

From the table above, we see that customers with more children tend to buy less insurance than customers with fewer children. Average claim amount will gradually increase as the customer has more children but the average claim amount tends to decrease when the customer's number of children is 4 or more children.

Table 4.5: The statistic of the smoker column

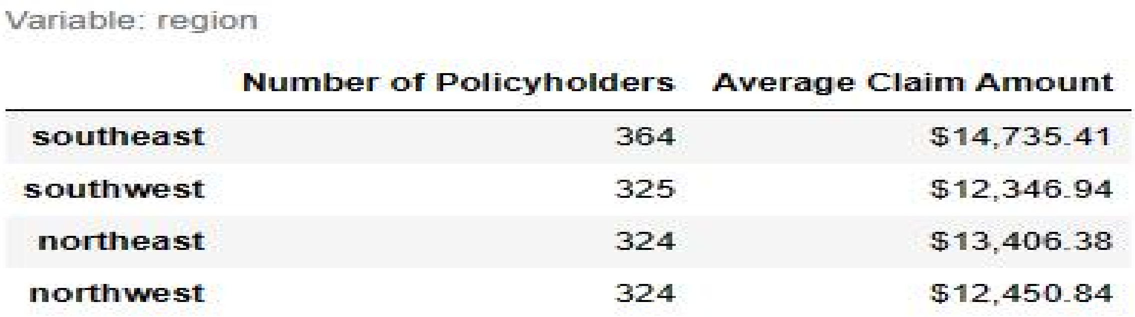


Variable: smoker

	Number of Policyholders	Average Claim Amount
no	1063	$8,440.66
yes	274	$32,050.23

From the table we see that the number of smokers participating in insurance is much less than that of non-smokers. And the average claim amount of smokers is much higher than that of non-smokers. This is very obvious because smoking will harm the health of the user, so the risk is therefore also higher leading to higher insurance premiums.

Table 4.6: The statistic of the region column



Variable: region

	Number of Policyholders	Average Claim Amount
southeast	364	$14,735.41
southwest	325	$12,346.94
northeast	324	$13,406.38
northwest	324	$12,450.84

From the table we can see that there is quite a similarity in the number of customers from 4 regions. Customers from the southeast will pay the highest average claim amount while customers from the southwest pay the lowest average claim amount of the 4 regions.

4.2.Visualization:

- **Univariate Analysis**

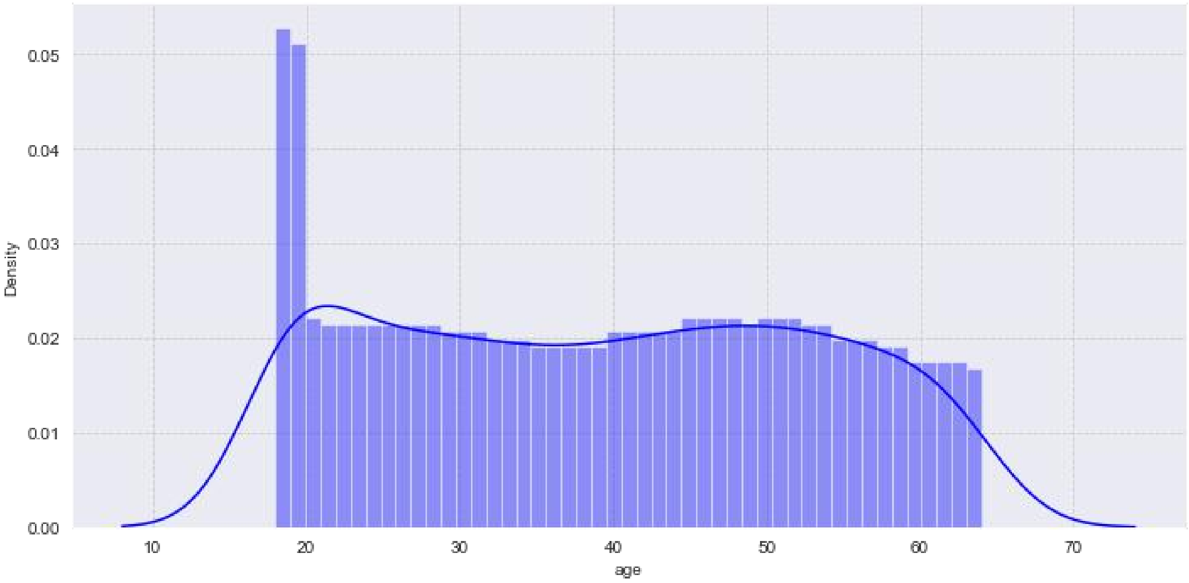


Figure 4.7: Distribution chart of the age column

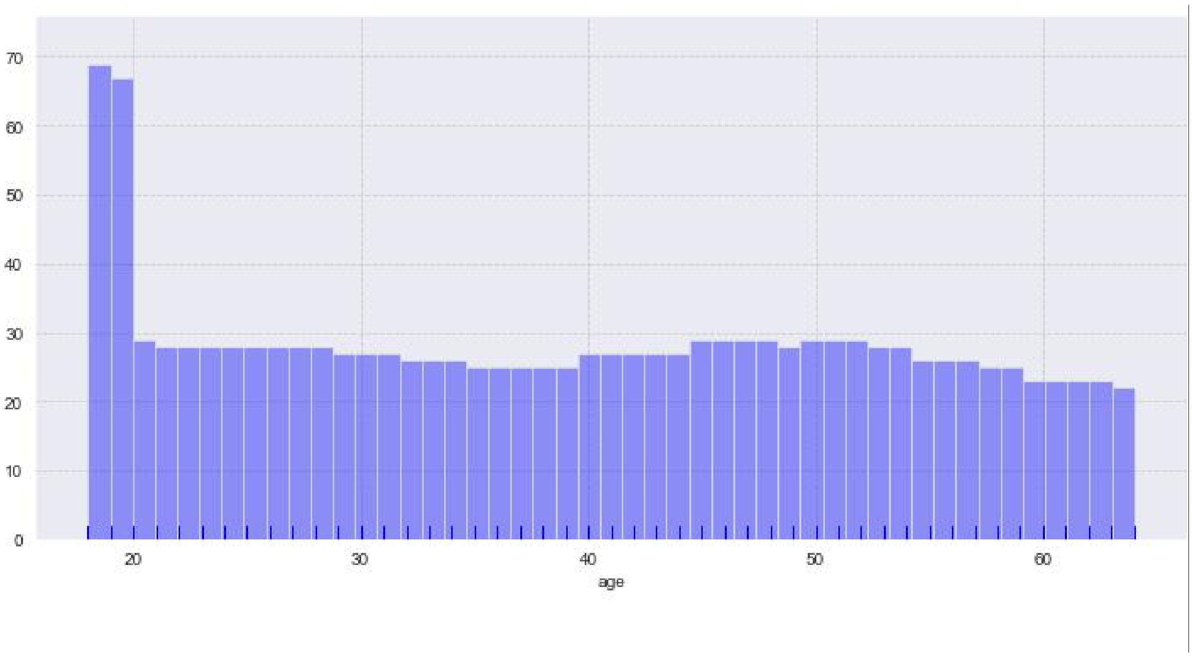


Figure 4.8: Distribution chart of the age column

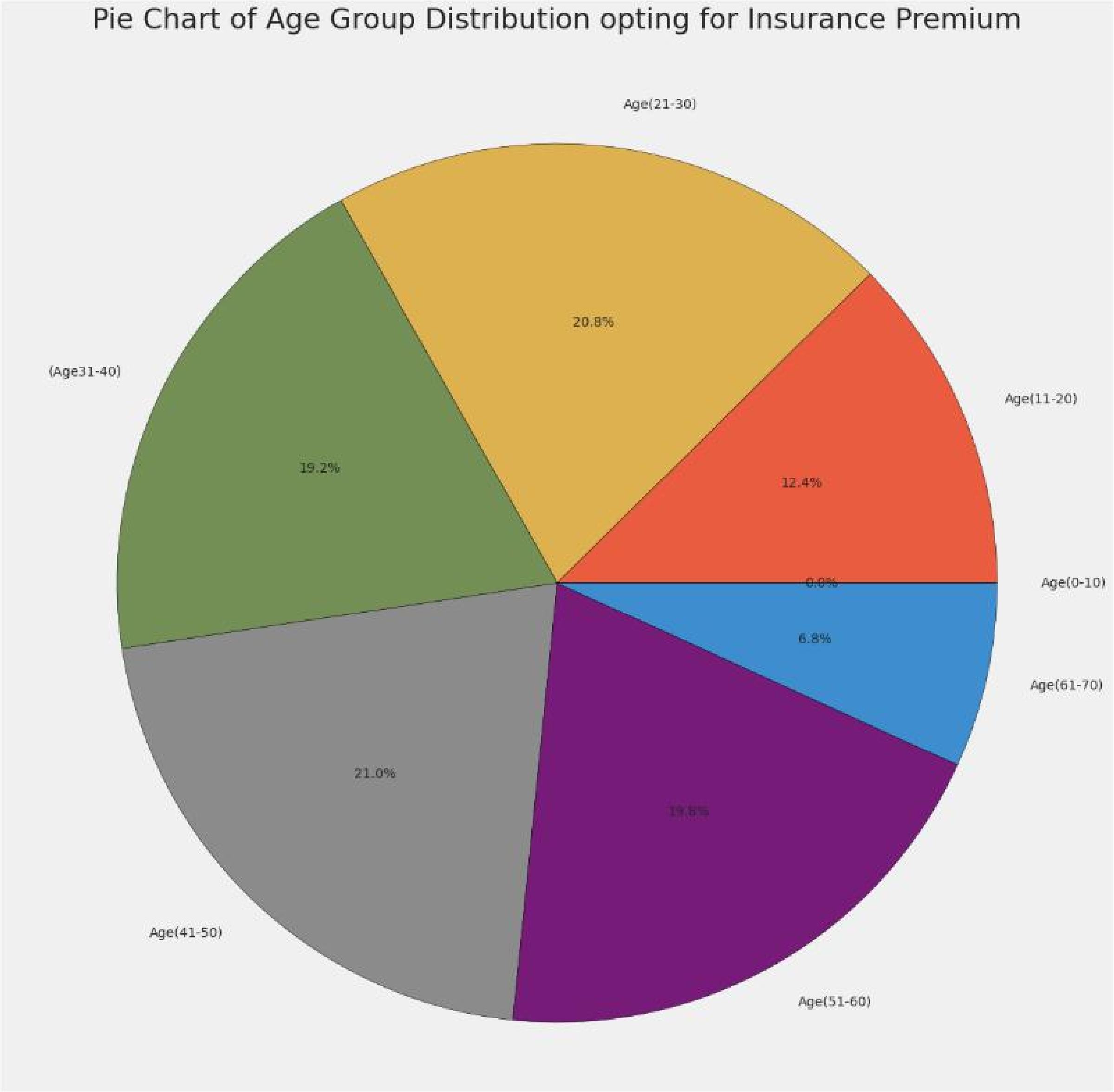


Figure 4.9: Pie chart of age group distribution

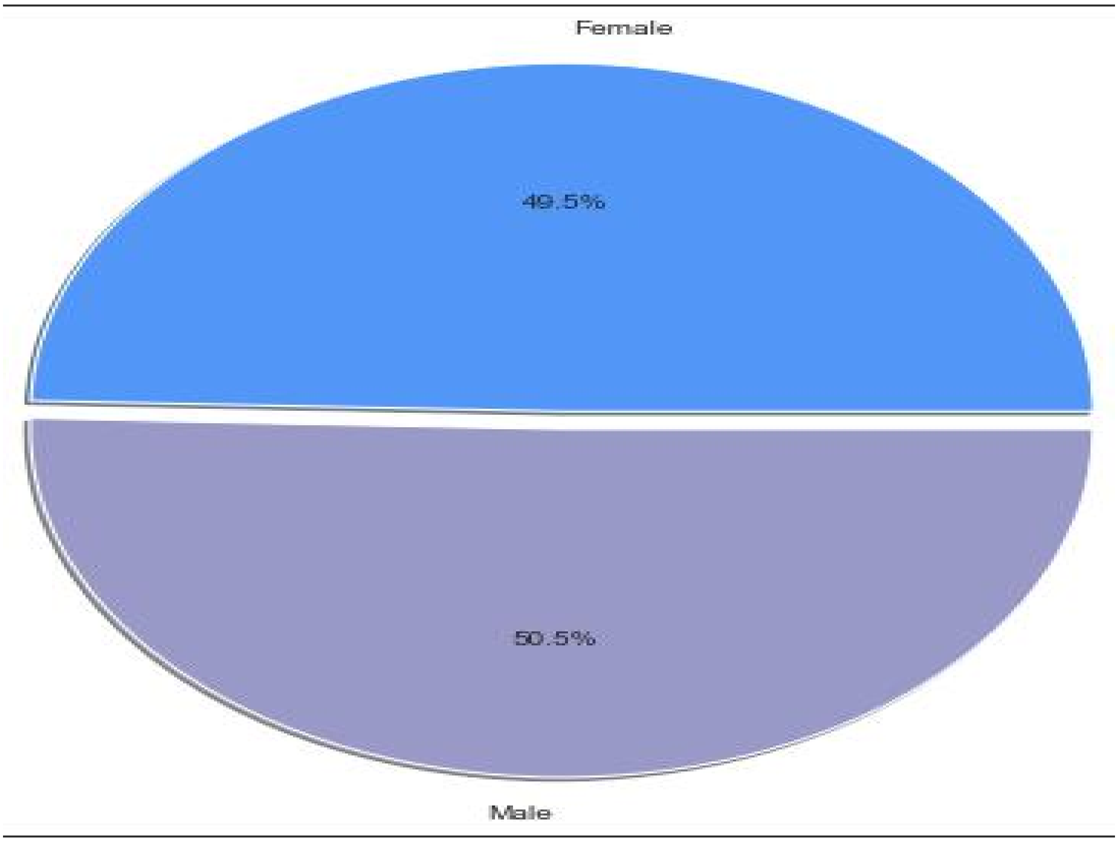


Figure 4.10: Pie chart of the sex column

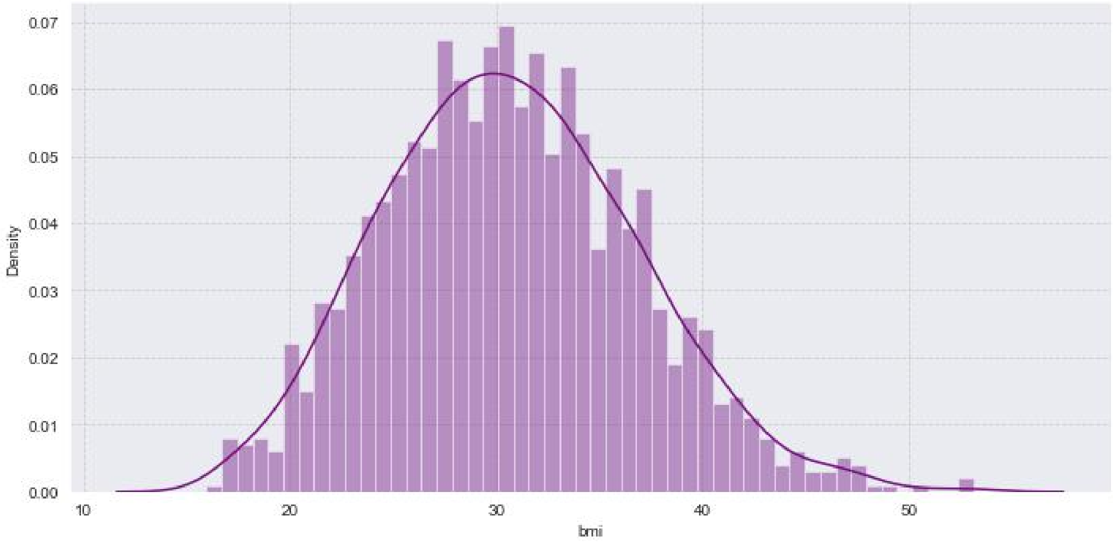


Figure 4.11: Distribution chart of the BMI column

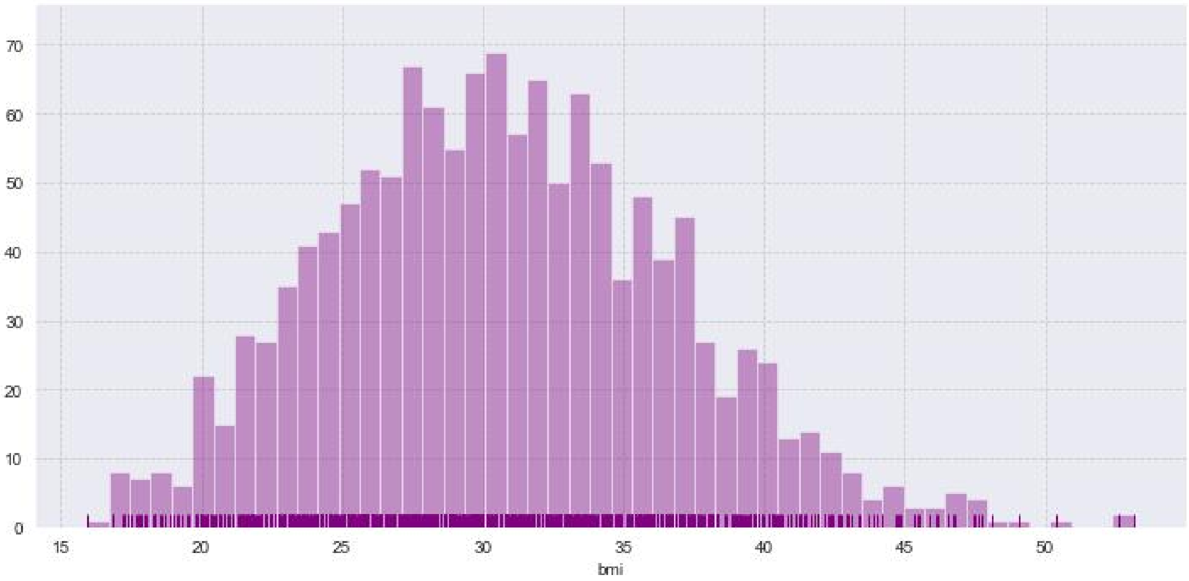


Figure 4.12: Distribution chart of the BMI column

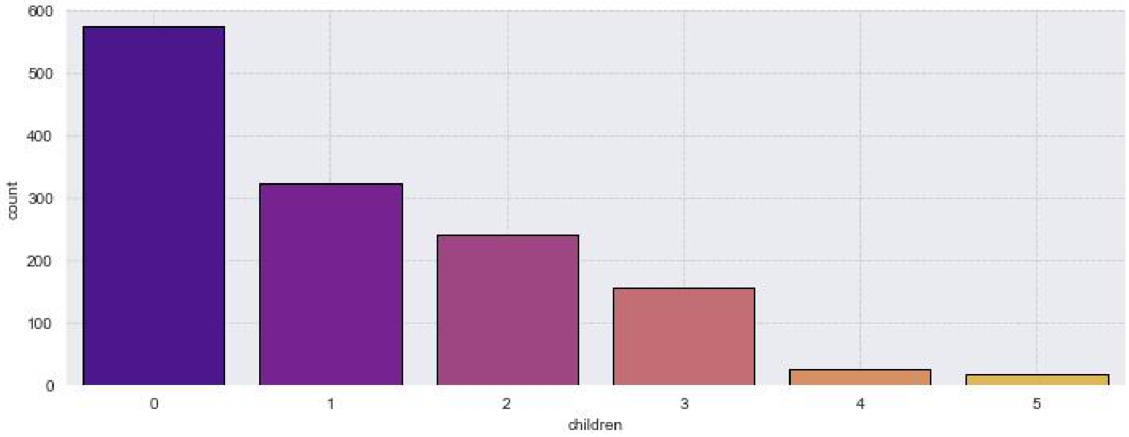


Figure 4.13: Distribution chart of the children column

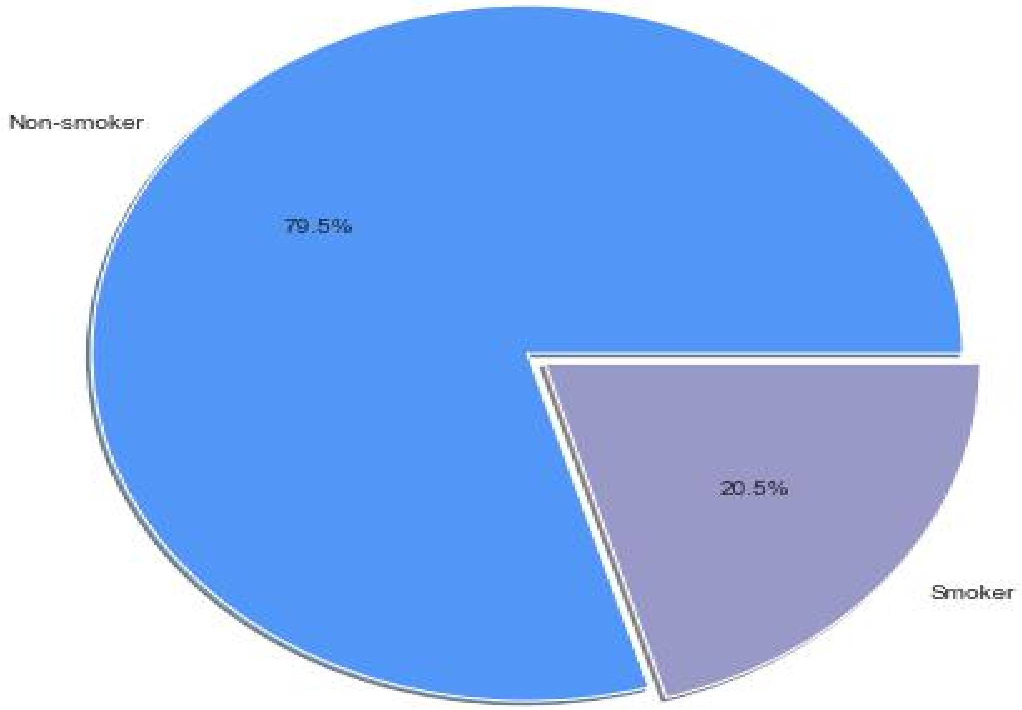


Figure 4.14: Pie chart of the smoker column

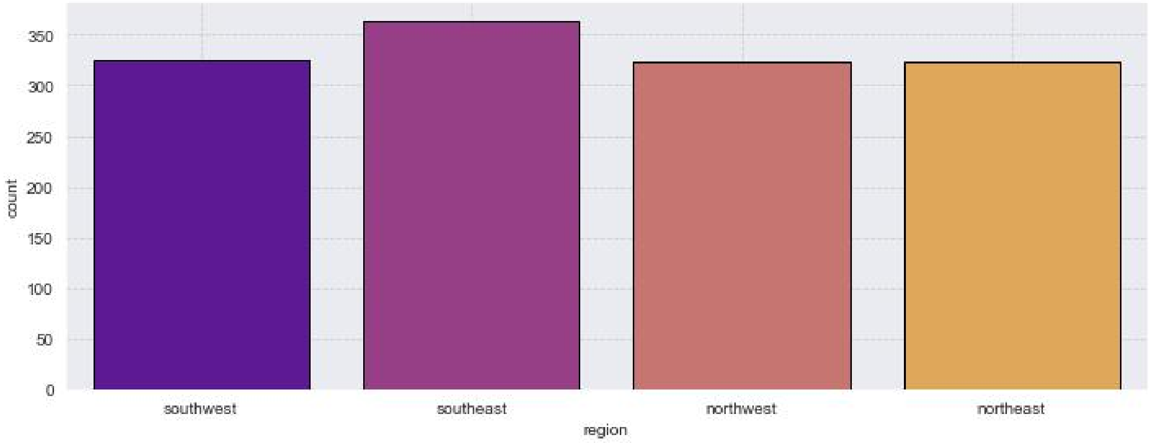


Figure 4.15: Distribution chart of the region column

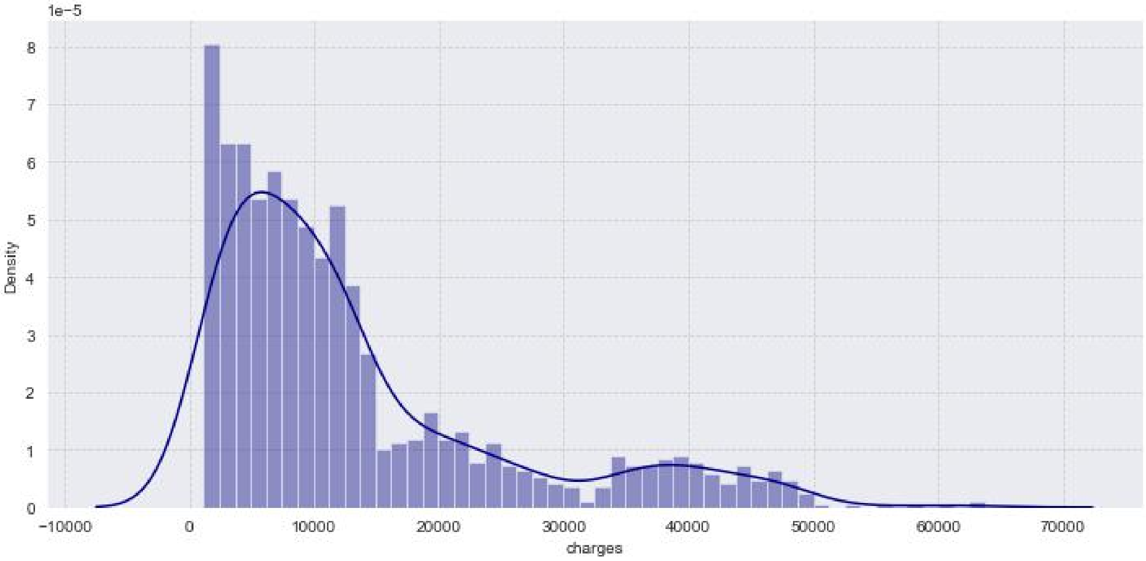


Figure 4.16: Distribution chart of the charges column

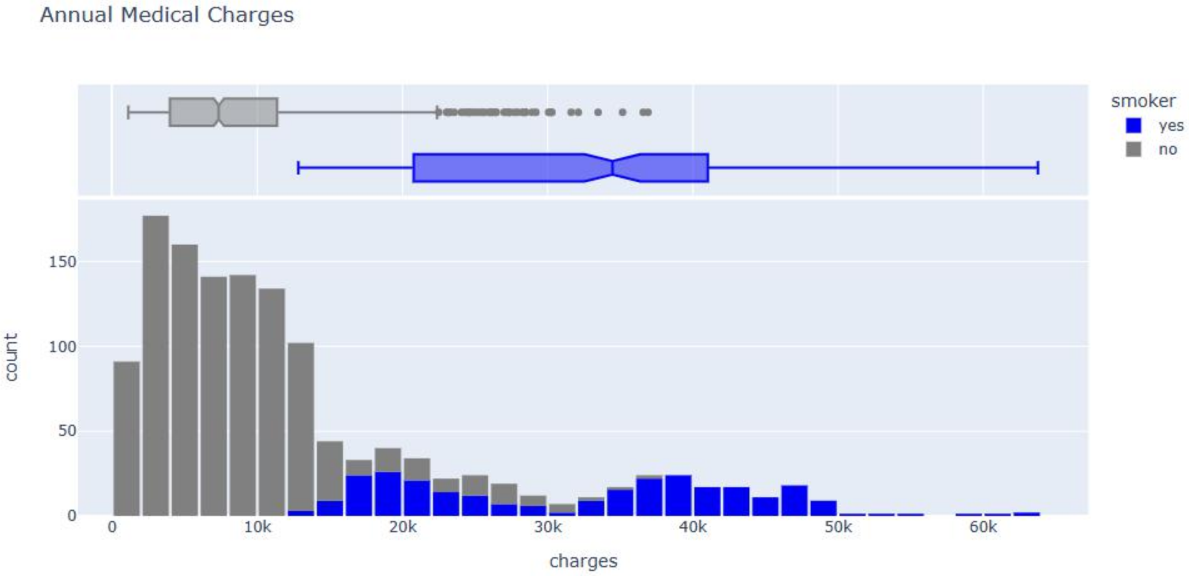


Figure 4.17: Distribution chart of charges and smoker

From the above graphs, the following observations can be drawn:

For the age column, it can be seen that the distribution of the histogram is skewed and customers aged 18 and 19 years have the most frequency, 69 people and 68 people, respectively, more than twice the frequency of other ages. And the number of customers in other ages evenly distributed from 22 to 29 people. However, looking at the pie chart, we can see that in terms of age group, the number of customers in the age group from 41 to 50 years old is the largest segment (21%) and the number of customers from the age group from 61 to 70 years old is the lowest segment (6.8%). The reason why the number of 18-year-old and 19-year-old customers is more than double that of other ages is because young people have fewer illnesses than the elderly, so in the same premium, insurance companies often pay less money, less medical bills for young people than older people. Otherwise, every age group in the US has equivalent population density.

For the sex column, The pie chart gives the data visualization results that are consistent with the descriptive statistics that we did above. The proportions between men and women are approximately equal at 50.5% for men and 49.5% for women, respectively. There is no data imbalance.

For the BMI column, we can see that the distribution of the histogram is skewed to the right. The BMI with the most frequency is 32.3. A person with a normal BMI will range between 18.5 - 24.9, this number indicates that you are at your ideal weight. However, when looking at the chart, we see that the majority of customers have a BMI of 25 or more. This shows that there are many health-risk customers, therefore the insurance company will have to pay more medical bills for this type of customers and hence they provide these types of people the same health insurance at a higher price

For the children column, There is a trend that is easy to see when looking at the graph that customers with more children are less likely to buy insurance than customers with fewer children. Specifically, the number of customers without children is the largest with 574 people while the number of customers with 5 children is 18 people. This trend is similar to the result of the Descriptive statistics above.

For the smoker column the region column, The data visualization results are consistent with the results of the descriptive statistics table for these two columns above.

For the charges column, We can see that the distribution of the histogram is skewed to the right and the majority of customers' annual insurance costs are in the range of less than $10,000, only a small number of customers have higher insurance costs. The cause may be due to an accident, serious illness or genetic disease. Besides, we can see in the next chart that the insurance costs of smokers are higher than that of non-smokers.

- **Multivariate Analysis**

Categorical columns with target column

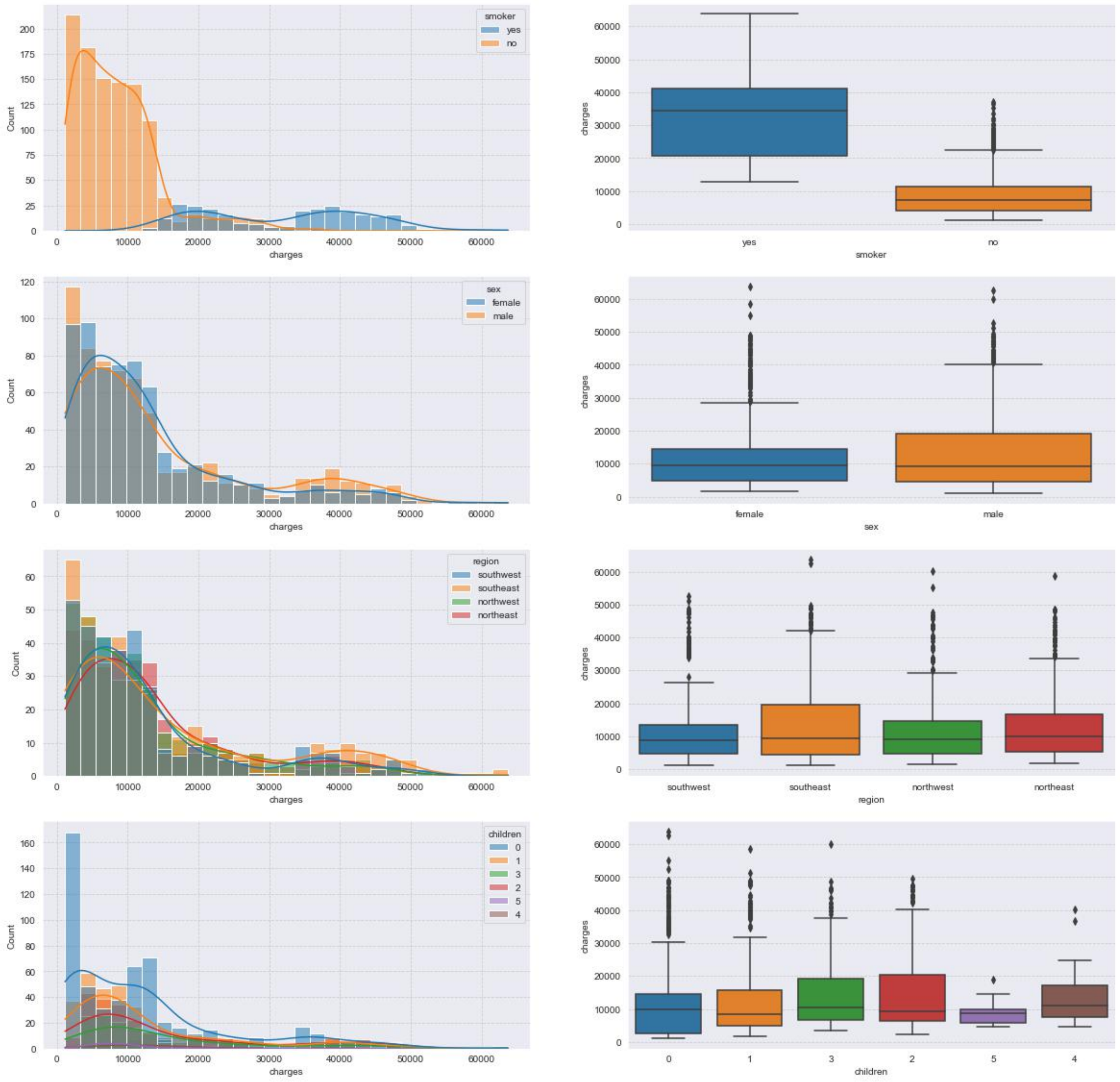


Figure 4.18: The charts of categorical columns with target column

The distribution of charges over smoker, we can see that smokers have higher premiums than non-smokers due to the higher health risks.

The distribution of charges over gender, we will see that male will pay higher premiums than female because male are more adventurous and risk-taking than female and this increases the risk of being exposed to more dangers than female. And in fact, in the US it is clear that men will tend to buy insurance more than women.

The distribution of charges over region, southeast is the region that pays the most insurance premiums compared to other regions, but most customers from all regions of the US are only charged from 0-20 000 USD.

The distribution of charges over the number of children, It seems the majority of our customers who have 0 or 1 child will have the average fees range from 8500 to 11000 USD. We can also conclude that those with more children are given less priority in terms of discounts.

Numerical columns with target column

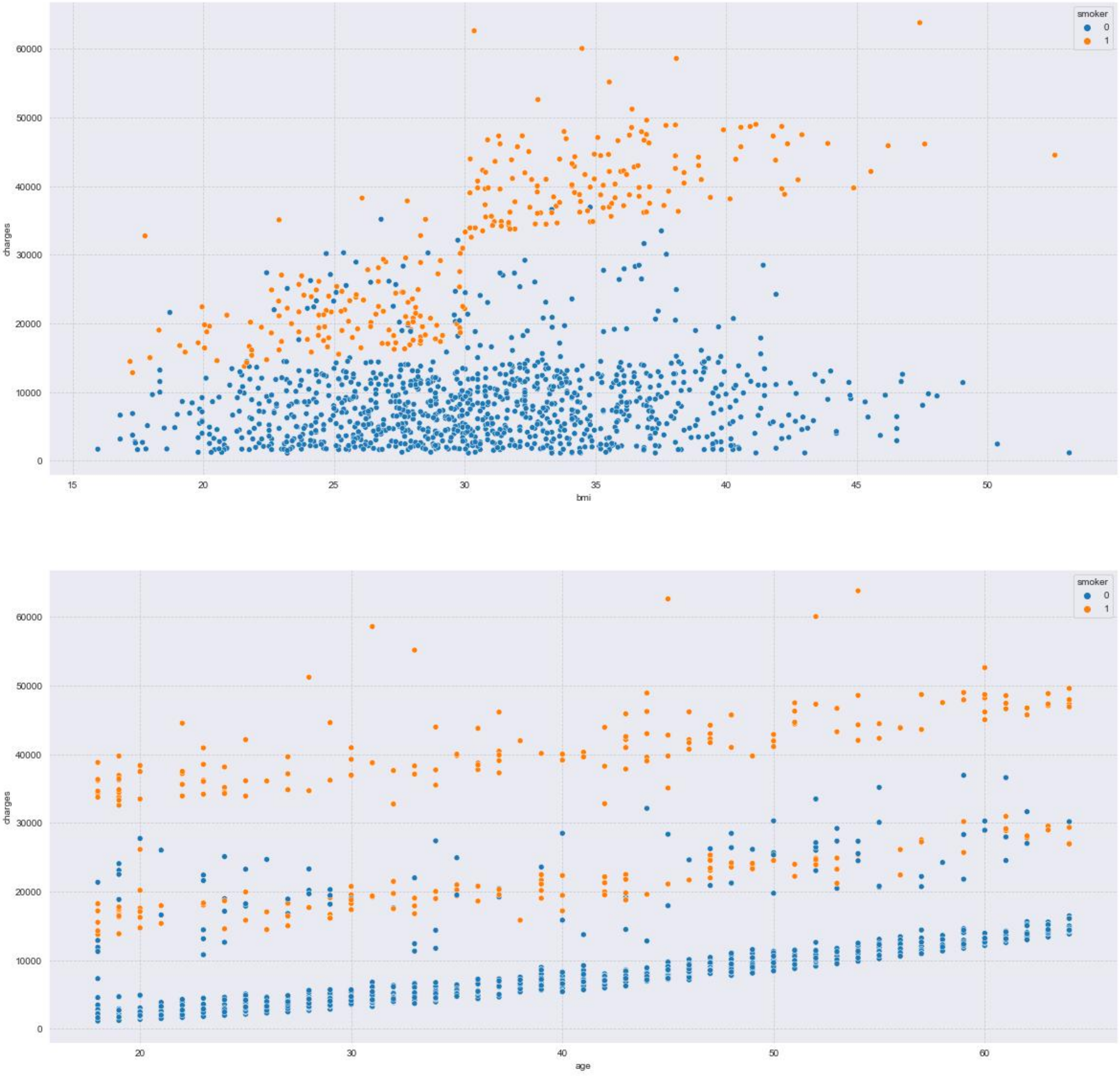


Figure 4.19: The charts of numerical columns with target column

According to the distribution of charges over BMI, smokers with a BMI index greater than 30 are likely to have much higher medical costs. For non-smokers, an increase in BMI did not appear to have an impact on insurance costs.

The distribution of charges over age, we find that whether a smoker or a non-smoker, as their age increases, so does the cost of insurance. This is a very obvious one. However, there may be a mix of smokers but no serious health problems and non-smokers with the disease. So in short, whether you have the disease or not, a smoker will still bear a much higher premium than a non-smoker even if a non-smoker has the disease. The difference is around 5000 to 1000 USD.

Categorical columns with numerical columns

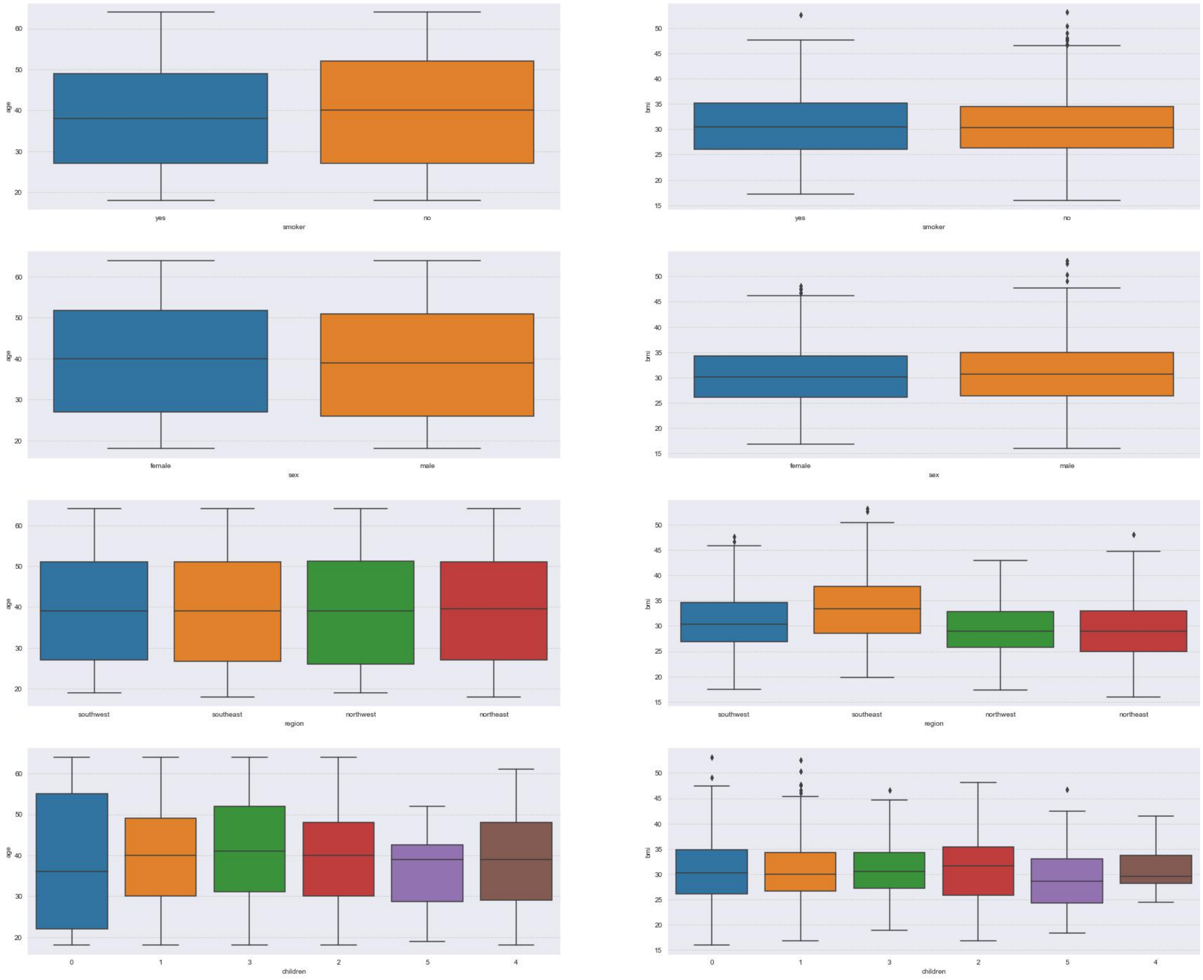
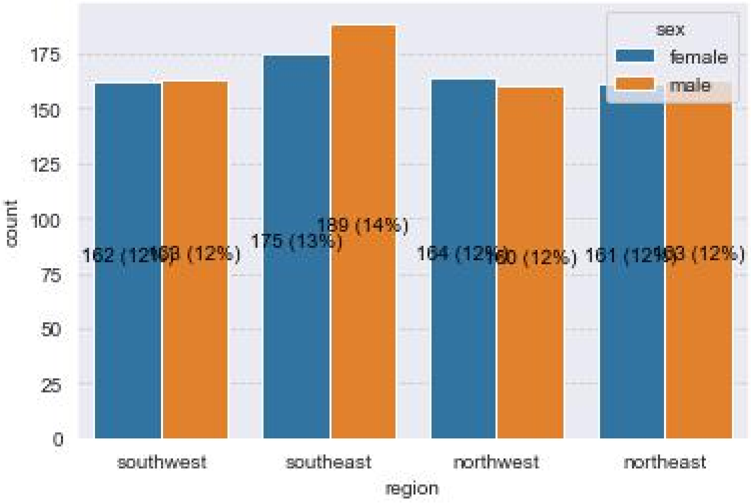


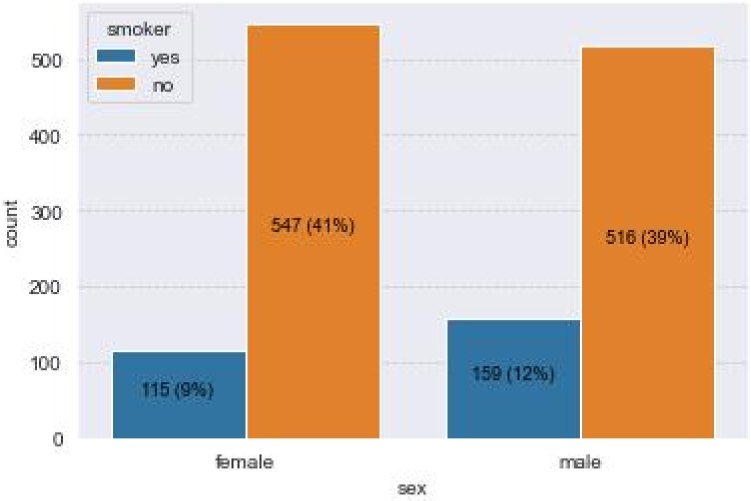
Figure 4.20: The charts of categorical columns with numerical columns

The distribution of age over smoker, we see that non-smokers will age older than smokers. Next, There are no major differences between the ages of men and women as well as those of customers from 4 regions. In addition, customers without children will appear to be of all ages while customers with 5 children will only be concentrated in the age group from 30 years old to 40 years old.

According to the distribution of BMI over smoker, there was no difference in BMI of smokers and non-smokers as well as between male and female. Meanwhile, customers from the southeast had a higher BMI than other regions.

Categorical columns with each other





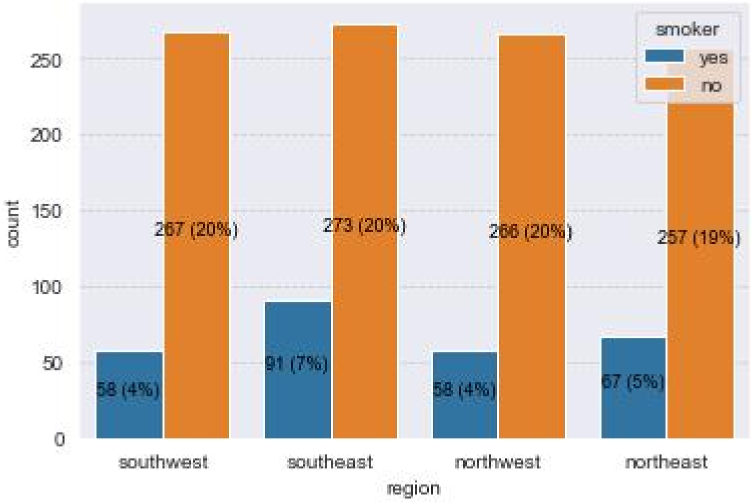
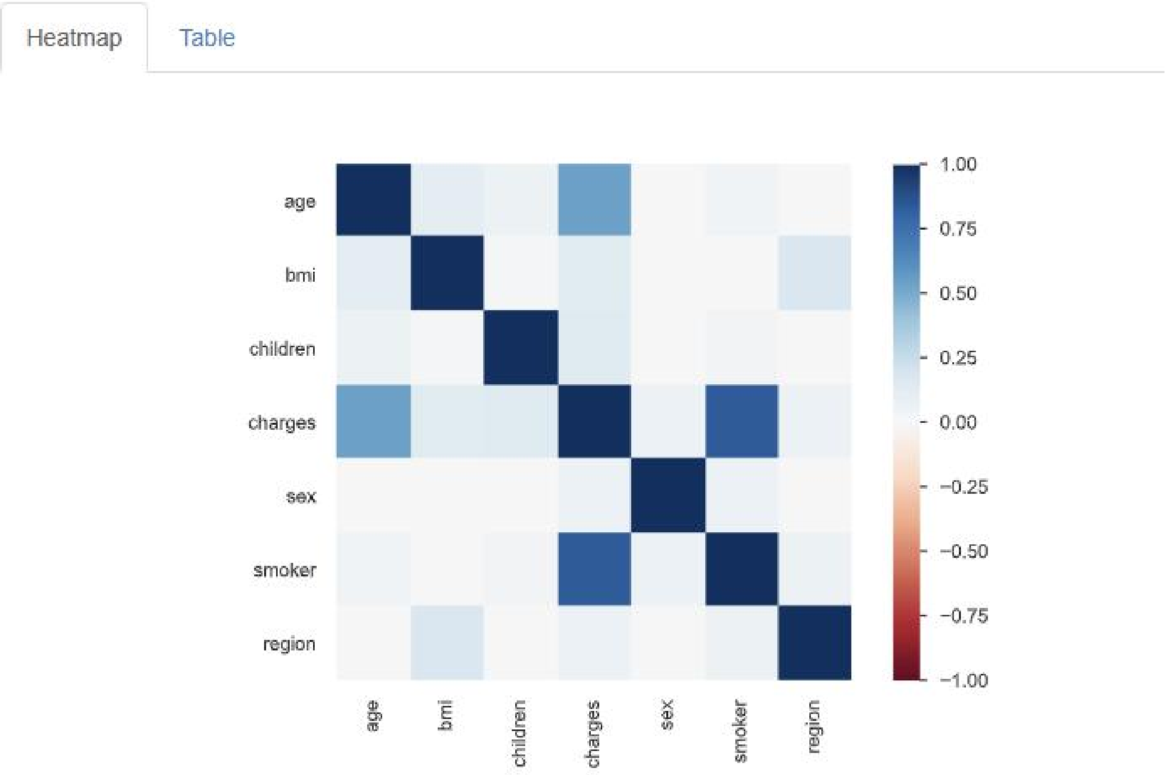
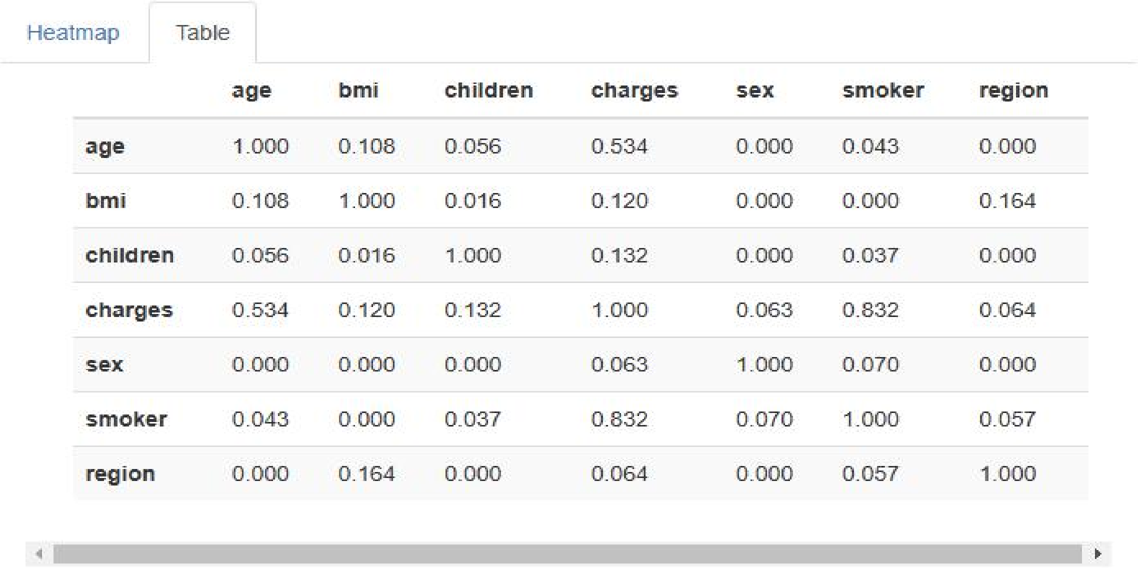


Figure 4.21: The charts of categorical columns with each other

The three graphs above are yet another confirmation that these gender data are perfectly balanced. The first section shows data roughly aggregated by 50% men and 50% women from each region, with the southeast having a higher number of males than females. This allows us to confirm, by considering the second and third cases, that there are the same number of smokers in both sexes and in all regions.

CORRELATION





	age	bmi	children	charges	sex	smoker	region
age	1.000	0.108	0.056	0.534	0.000	0.043	0.000
bmi	0.108	1.000	0.016	0.120	0.000	0.000	0.164
children	0.056	0.016	1.000	0.132	0.000	0.037	0.000
charges	0.534	0.120	0.132	1.000	0.063	0.832	0.064
sex	0.000	0.000	0.000	0.063	1.000	0.070	0.000
smoker	0.043	0.000	0.037	0.832	0.070	1.000	0.057
region	0.000	0.164	0.000	0.064	0.000	0.057	1.000

Figure 4.22: Correlation Matrix for Attributes

The correlation matrix shows that the target variable is mainly related to age and smoking characteristics and there is no strong collinearity between the independent variables

V. THE SPECIAL POINT OR POTENTIAL ISSUE OF DATA:

After exploratory analysis of the data to understand the data set, we found that all columns have balanced data however there is an imbalance of data for the smoker column with a ratio of 80:20 for the Non-smokers and smokers. Besides, based on the data explorations we have done above, we all see that our male customers are paying more bills than female customers but it is interesting that women in the region Northwest are paying more medical bills. Why this happens requires further research.

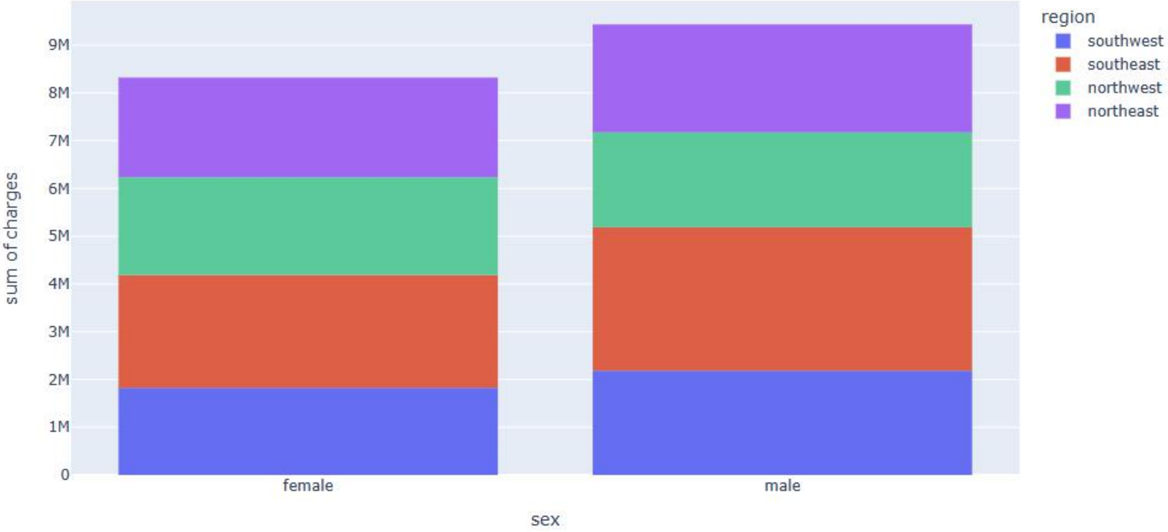


Figure 5.1: The chart between charges, sex and region

VI. PERFORM THE MODEL TO SOLVE THE PROBLEM, DISCUSS THE RESULT, MAKE CONCLUSIONS OR RECOMMENDATIONS:

- **The steps to build the model include:**

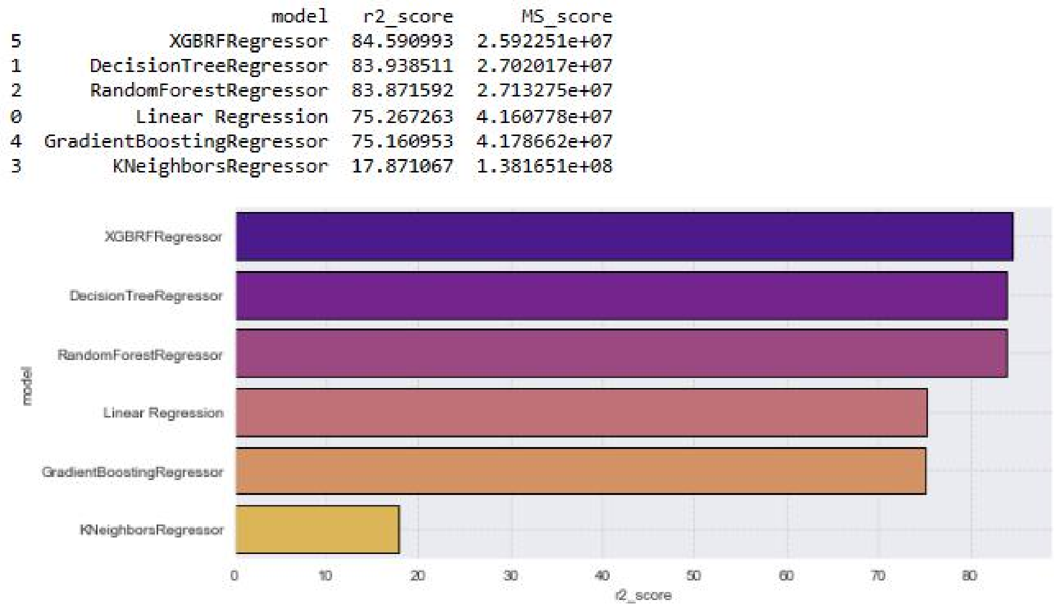
Step 1: Splitting the Dependent and Independent features

Step 2: Splitting the dataset into Training Set and Test Set

Step 3: Model training with multiple algorithms to find out the optimize model

Step 4: Models evaluation

- **The result and Discussion:**



```
                     model  r2_score      MS_score
5           XGBRFRegressor  84.590993  2.592251e+07
1    DecisionTreeRegressor  83.938511  2.702017e+07
2    RandomForestRegressor  83.871592  2.713275e+07
0        Linear Regression  75.267263  4.160778e+07
4 GradientBoostingRegressor  75.160953  4.178662e+07
3      KNeighborsRegressor  17.871067  1.381651e+08
```

Figure 5.1: The result of models

The error between the actual charge and the predicted charge was calculated using the mean squared error (MSE). To check the accuracy of the model The R-squared value is also calculated. The performance of different methods is based on Mean

Absolute Error and R-squared values are depicted in Figure above. As we can see, there are 6 models used in which there are 3 models with R^2 prediction efficiency over 80% which are XGBRFRegressor, DecisionForestRegressor and RandomForestRegressor. And the model XGBRFRegressor is the model that we will choose because it has the highest R^2 of 84.59% which means this model can correctly predict the premium up to 84.59% and its mean square error is also the lowest compared to other models.

- **Conclusion:**

The report uses various regression models to forecast health insurance costs based on specific attributes that influence medical costs. The findings summarized in the figure above showed that XGBRFRegressor performed the best, with an accuracy of 84.59. Therefore, XGBRFRegressor can be used to estimate insurance costs with better performance than other regression models.

The data set used in the report is not large, but it is quite similar in reality as men will have higher insurance costs than women or non-smokers will have much less insurance costs than smokers. In the future, if there is an opportunity and cost to scale up to a larger data set, there will be more and more precise and in-depth studies.

Insurance price forecasting supports certain factors that help insurance providers attract customers and save individual planning time. This greatly reduces these individual efforts in policy making, as metric capacity models can calculate costs in a very short time, while someone else would take a long time. to perform continuous tasks. This can help businesses improve profitability. Metric capacity unit models can even handle huge amounts of data.