Medical insurance cost prediction model:

Executive Summary:

Predicting healthcare costs for individuals using accurate prediction models is important for various stakeholders beyond health insurers, and for various purposes.

The purpose of my study is to investigate different features, observe their relationship, and to come up with the best prediction model based on several features of individual such as age, physical/family condition and location against their existing medical expense to be used for predicting future medical expenses of individuals that help medical insurance to make decision on charging the premium. With the study, I have been able to highlight importance of different input features in predicting insurance charges.

Key stakeholders in these efforts to manage healthcare costs include health insurers, employers, society, and healthcare delivery organizations. Although many researchers have highlighted the importance of predicting people's health costs to improve healthcare budget management, most of them do not address the frequent need to know the reasons behind this prediction, i.e., knowing the factors that influence this prediction. The objective of this research is to accurately predict insurance costs based on people's data, including age, body mass index, smoking or not, etc.

Introduction:

I have used Medical Insurance dataset from Kaggle. The insurance.csv dataset contains 1338 observations (rows) and 7 features (columns). The dataset contains 4 numerical features (age, bmi, children and charges) and 3 nominal features (sex, smoker and region). After pre-processing the data, I developed several models including linear regression, ridge, lasso, neural networks, K nearest neighbours, decision tree, random forest etc on the data. The performance of these models were evaluated using mean squared error and R-squared. The results showed that developed models have a high accuracy in predicting health insurance costs.

Why I chose this research question?

My findings suggest that machine learning algorithms can be effective in predicting health insurance costs and can assist insurance providers in making informed decisions. The study provides insights for further research in developing more accurate models and understanding the factors that influence health insurance costs.

My research:

1. Investigated input features such as age, physical/family condition, and location. Cleaned, and pre-processed data, observed relationship and collinearity between variables.
2. Studied historic data and came up with the best model to predict insurance charges, based on several features of individual.
3. Determined the importance of different features to study their role and influence in deciding insurance charges of an individual.

The insurance.csv dataset was downloaded from the Kaggle site.
https://www.kaggle.com/mirichoi0218/insurance
https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset?resource=download.
https://osf.io/7u5gy

Dataset columns' brief explanation:

- **age**: age of primary beneficiary.
- **sex**: insurance contractor's gender: female, male.
- **bmi**: Body mass index, providing an understanding of the body, weight of the individual relative to his 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 habits of primary beneficiary. yes, no.
- **region**: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- **charges**: Individual medical costs billed by the health insurance.

Exploratory Data analysis:

Data columns (total 7 columns):

#	Column	Count	Non-Null	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

Data Types:

Float64 (2)
Int64 (2)
Object (3)

No null values in any column.

Dataframe.head():

	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

Dataframe.describe(include = 'all'):

	age	sex	bmi	children	smoker	region	charges
count	1338.000000	1338	1338.000000	1338.000000	1338	1338	1338.000000
unique	NaN	2	NaN	NaN	2	4	NaN
top	NaN	male	NaN	NaN	no	southeast	NaN
freq	NaN	676	NaN	NaN	1064	364	NaN
mean	39.207025	NaN	30.663397	1.094918	NaN	NaN	13270.422265
std	14.049960	NaN	6.098187	1.205493	NaN	NaN	12110.011237
min	18.000000	NaN	15.960000	0.000000	NaN	NaN	1121.873900
25%	27.000000	NaN	26.296250	0.000000	NaN	NaN	4740.287150
50%	39.000000	NaN	30.400000	1.000000	NaN	NaN	9382.033000
75%	51.000000	NaN	34.693750	2.000000	NaN	NaN	16639.912515
max	64.000000	NaN	53.130000	5.000000	NaN	NaN	63770.428010

Sex:

The dataset has 676 male (50.5232%) and 662 female (49.4768%)

Children:

574 individuals have 0 children.
324 individuals have 1 child.
240 individuals have 2 children.
157 individuals have 3 children.
25 individuals have 4 children.
18 individuals have 5 children.

Smoker:

SMOKER	COUNT	PERCENTAGE
NO	1064	79.521%
YES	274	20.4783%

Breakdown of smoking habits based on gender:

SEX	SMOKER	COUNT
FEMALE	no	547
	yes	115
MALE	no	517
	yes	159

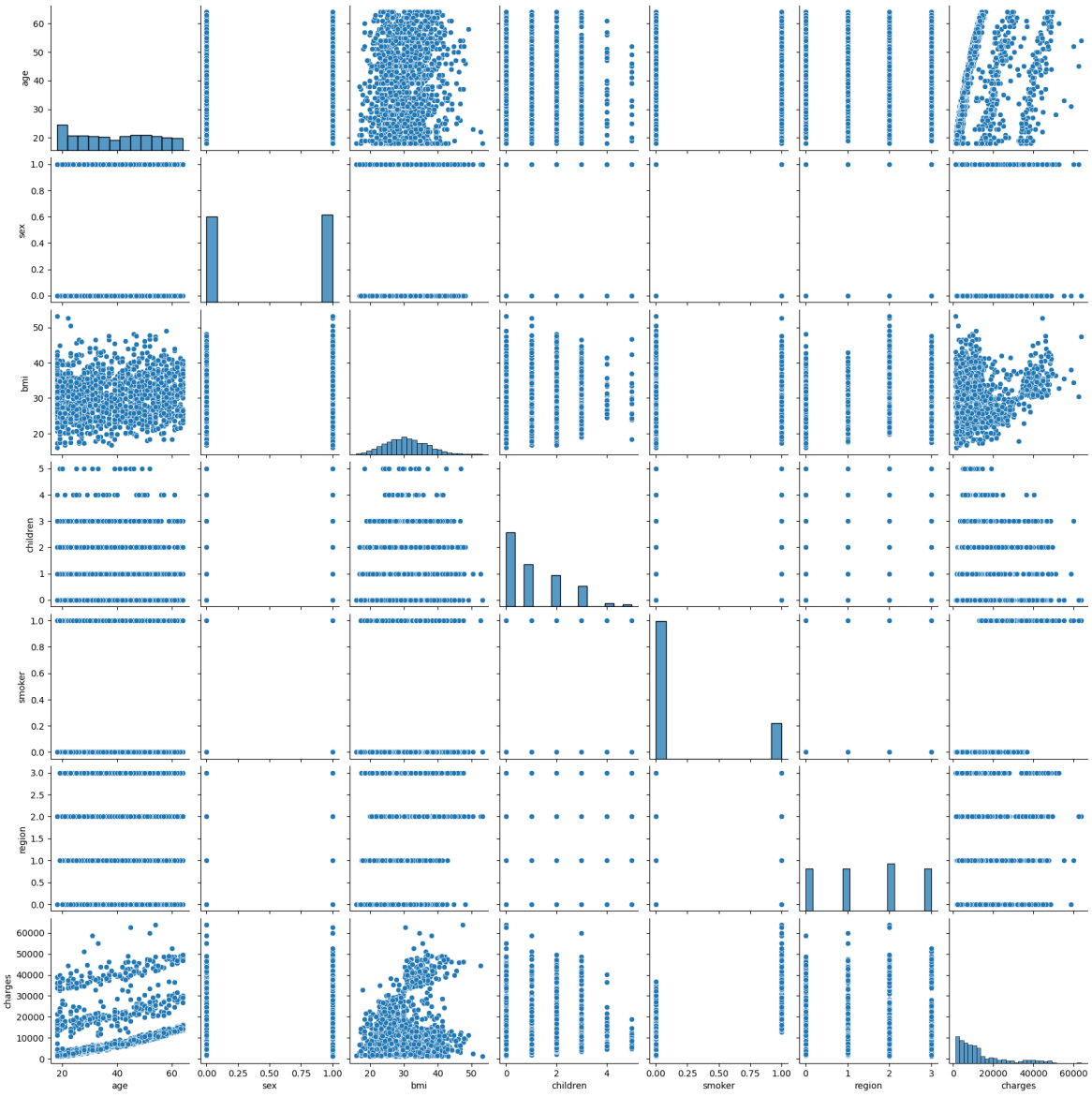
Region Count: Number of individuals from each region in the dataset.

Region	Count	Percentage
Southeast	364	27.20%
Southwest	325	24.29%
Northwest	325	24.29%
Northeast	324	24.21%

Correlation between Variables:

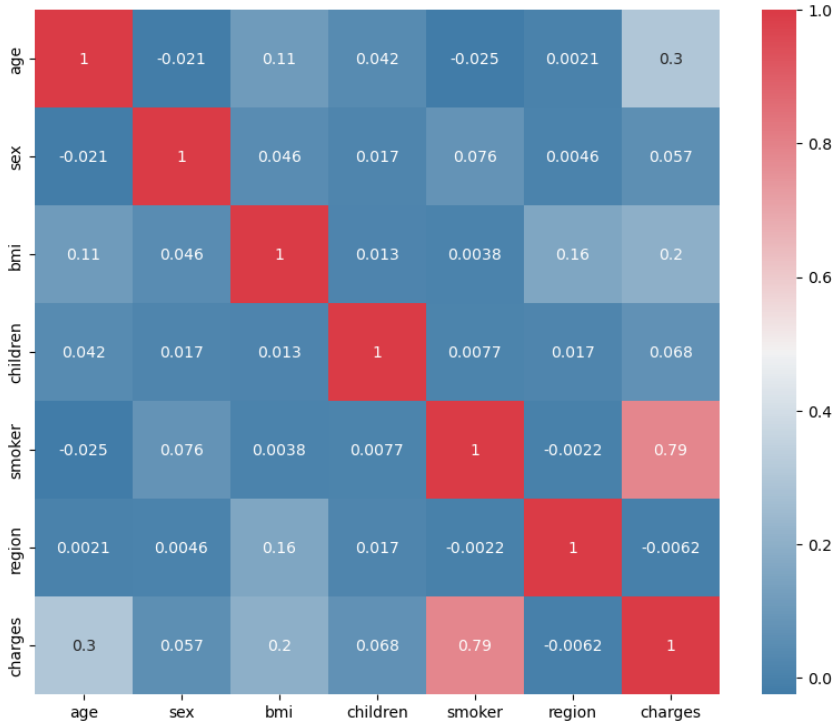
	age	sex	bmi	children	smoker	region	charges
age	1.000000	-0.020856	0.109272	0.042469	-0.025019	0.002127	0.299008
sex	-0.020856	1.000000	0.046371	0.017163	0.076185	0.004588	0.057292
bmi	0.109272	0.046371	1.000000	0.012759	0.003750	0.157566	0.198341
children	0.042469	0.017163	0.012759	1.000000	0.007673	0.016569	0.067998
smoker	-0.025019	0.076185	0.003750	0.007673	1.000000	-0.002181	0.787251
region	0.002127	0.004588	0.157566	0.016569	-0.002181	1.000000	-0.006208
charges	0.299008	0.057292	0.198341	0.067998	0.787251	-0.006208	1.000000

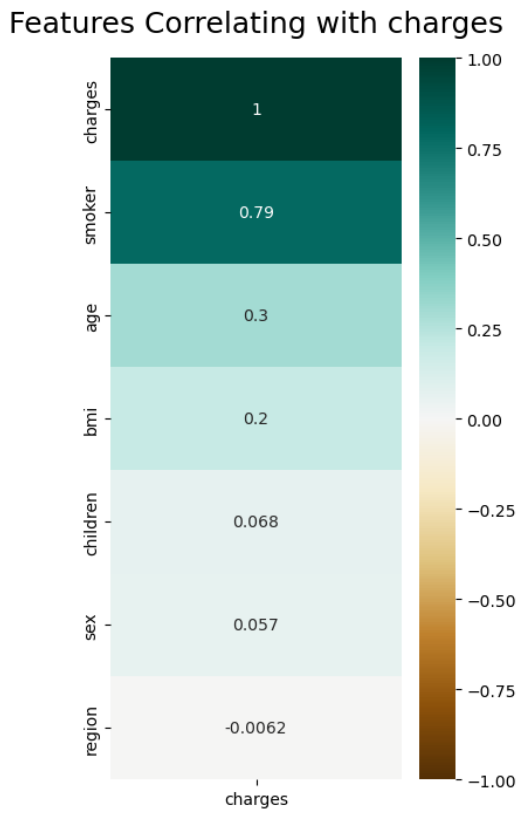
Pairplot:



Correlation values based on ‘Charges’ as target variable:

region	-0.006208
sex	0.057292
children	0.067998
bmi	0.198341
age	0.299008
smoker	0.787251
charges	1.000000

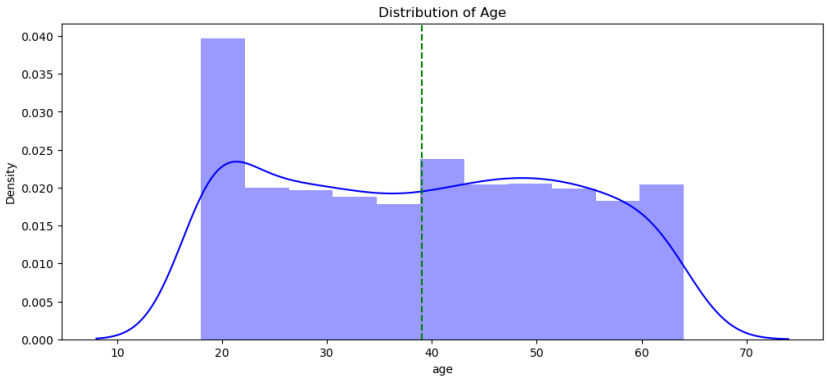




It is apparent from the pairplot and the heatmap, “smoker” feature plays the most important role in deciding insurance charges of an individual, followed by age and BMI.

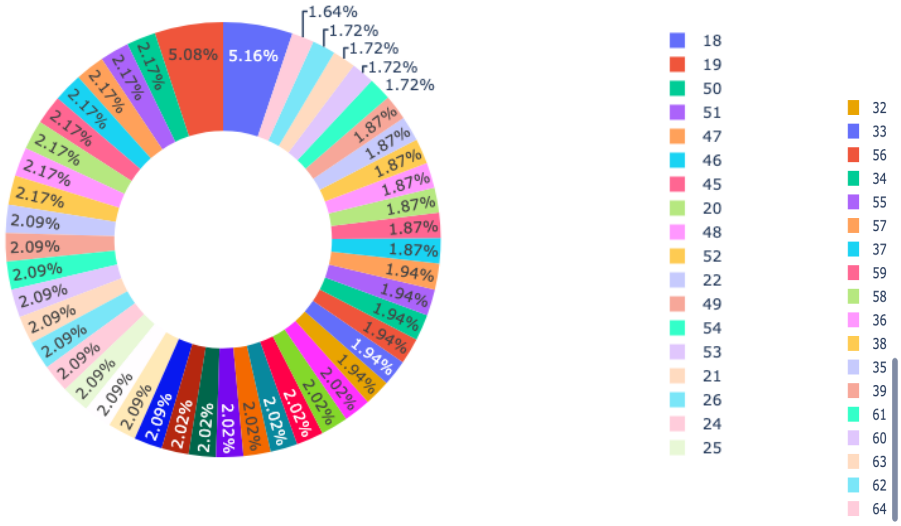
Now I’ll explore each variable one by one in a bit more detail:

1. Age:

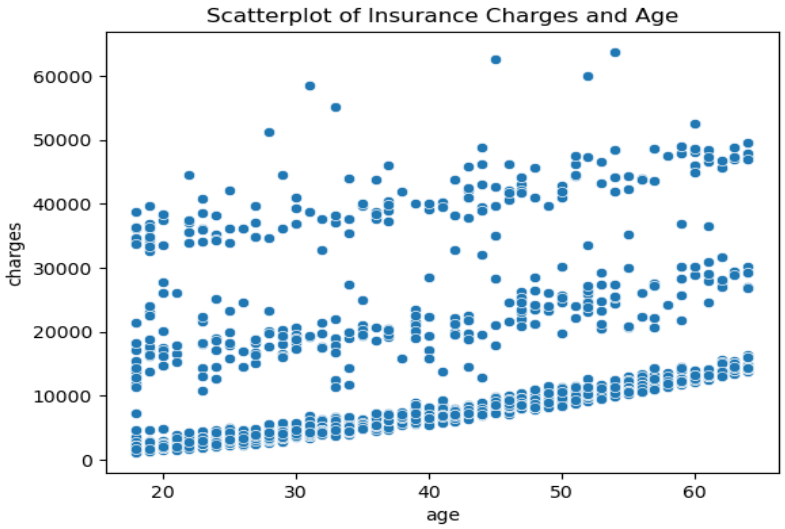


Average age is: 39.20 (value taken from describe method), and is represented by green line in the plot.

Pie chart of ‘Age’:

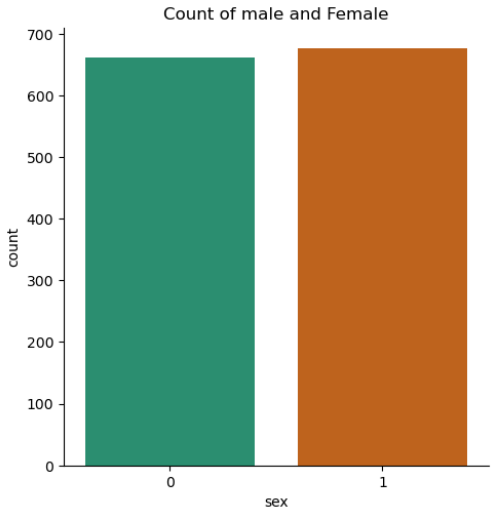


There are individuals under 20 years of age in the data set. This is the minimum age. The maximum age is 64 years.

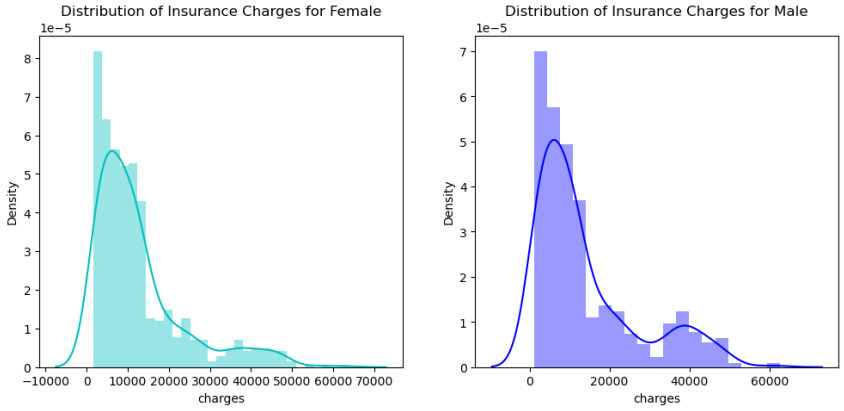


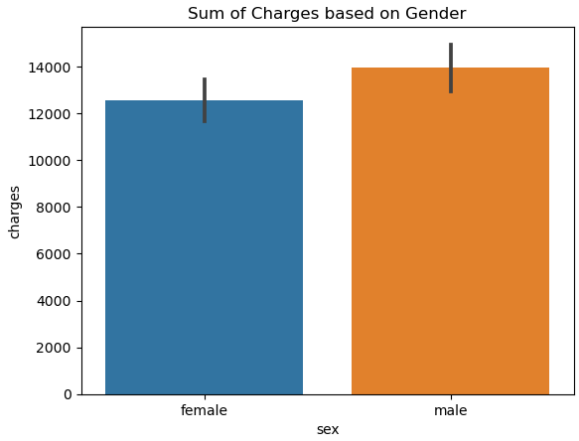
The scatterplot shows, Insurance Charges generally increase with increasing Age other than a few exceptions. That probably has to do with smoking habits of individuals. May be those persons are non-smoker!!

2. Sex:



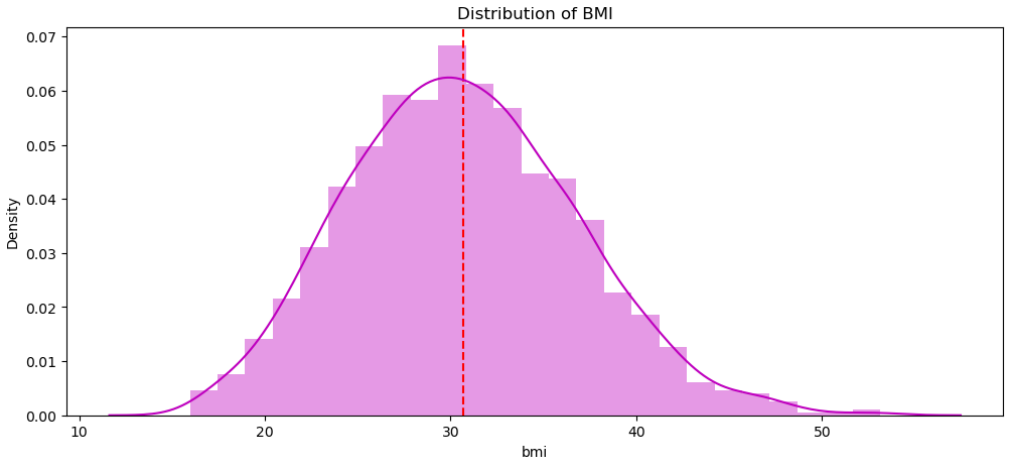
There are more male than female in the dataset which we have already established.



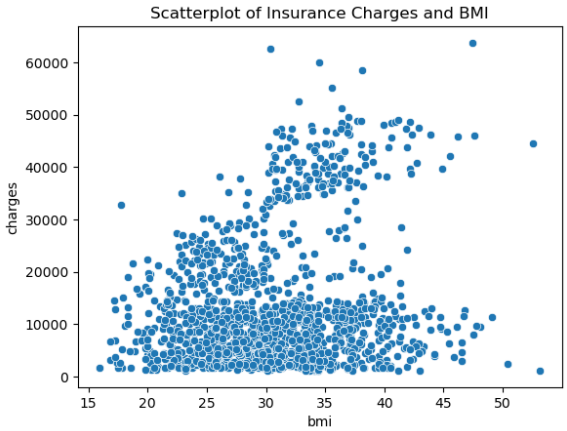


Comparison between Insurance charges of female and male. There isn't a marked difference. That means gender doesn't play much role in deciding Insurance Charges of an individual. Exact value is 0.057, value taken from collinearity table and the heatmap.

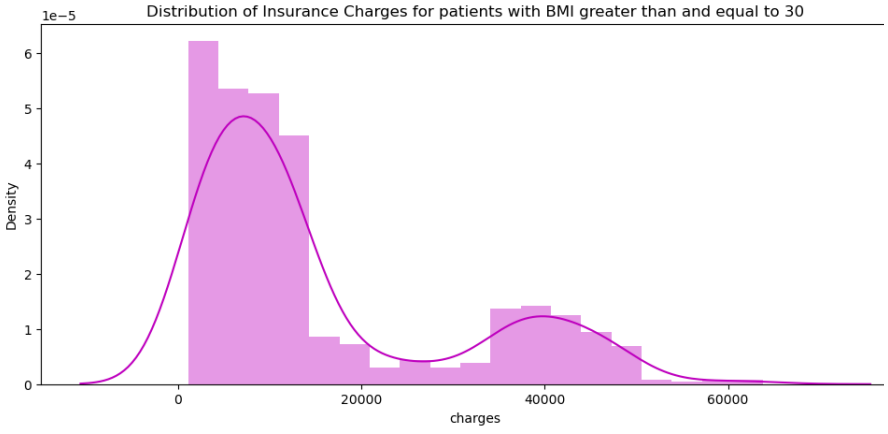
3. BMI:

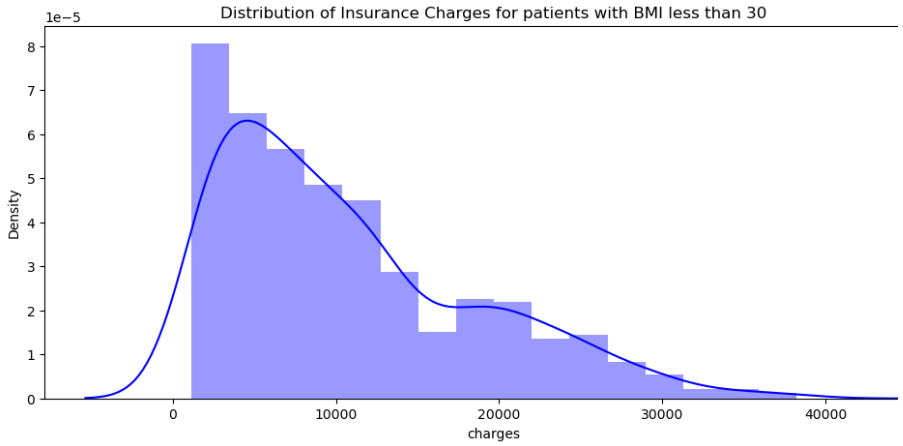


Average BMI is: 30.67 (value taken from describe method), represented by the red line in the graph which is a fairly uniformly distributed graph.



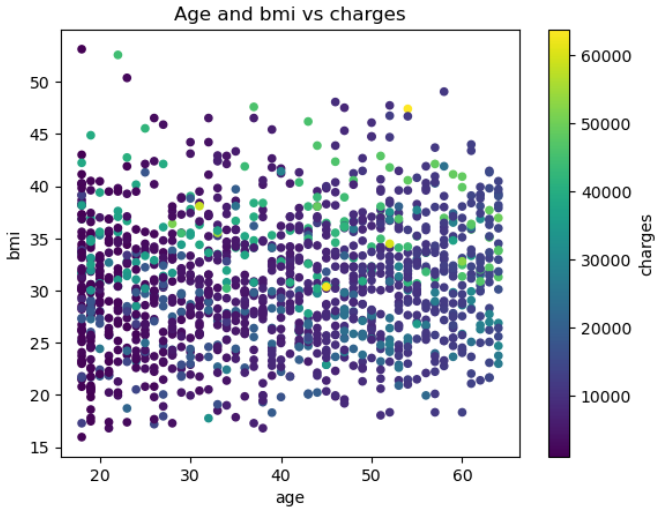
The scatterplot shows a general trend of increase in Insurance Charges with increase in BMI.
Let's explore BMI further and study distribution of Charges in patients with BMI greater and less than 30.





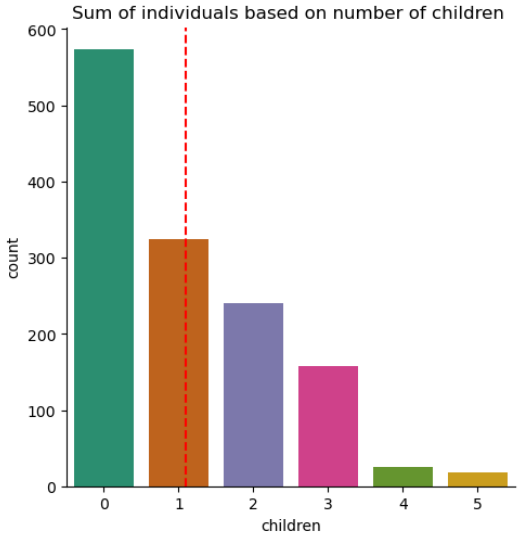
The comparison clearly shows Insurance Charges for individuals with BMI greater than or equal 30 are more than for individuals with BMI less than 30.

Graph to show relationship between Age, BMI, and Insurance Charges.

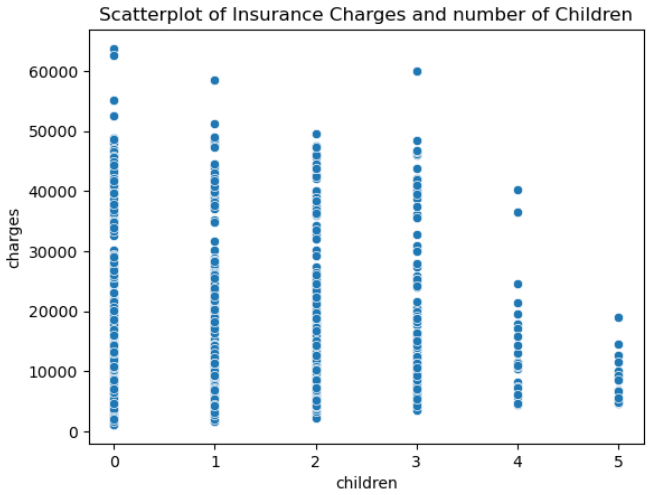


The scatterplot shows very little to no collinearity between Age and BMI when deciding Insurance Charges. The exact value is: 0.1092, value taken from the collinearity table and the heatmap.

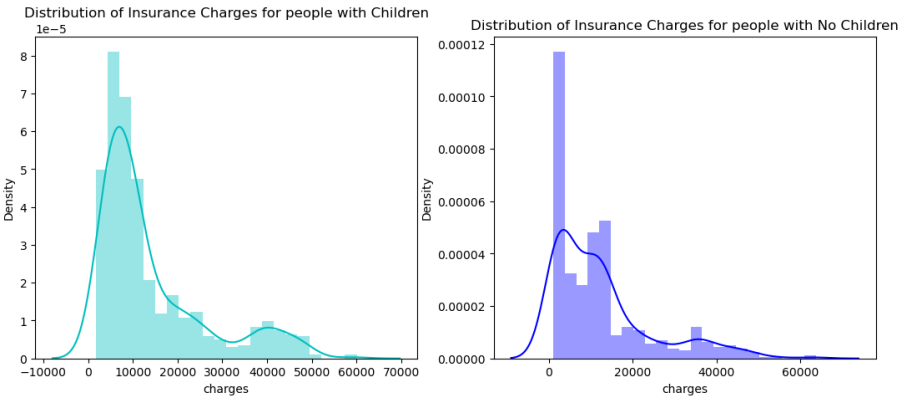
4. Children:



Most individuals have no child, and a very few have 5 children.
Average number of children: 1.09 (value taken from describe method)



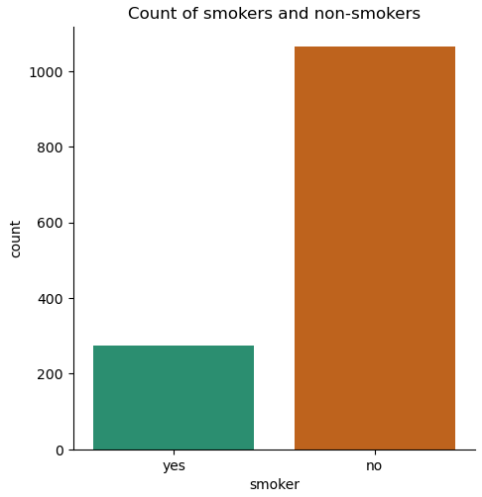
Number of children doesn't play much role in deciding Insurance Charges of an individual. Exact value is: 0.068 (Value taken from the heatmap). This can be further explained by the following subplots:



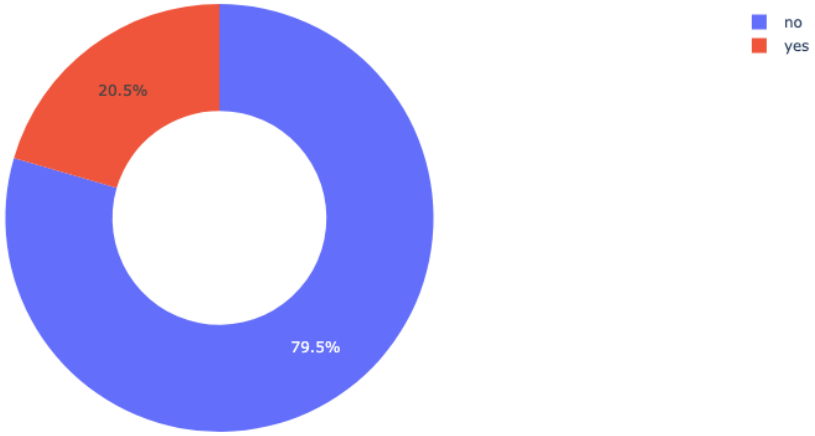
Comparison of Insurance charges of individuals with no child to individuals with children.

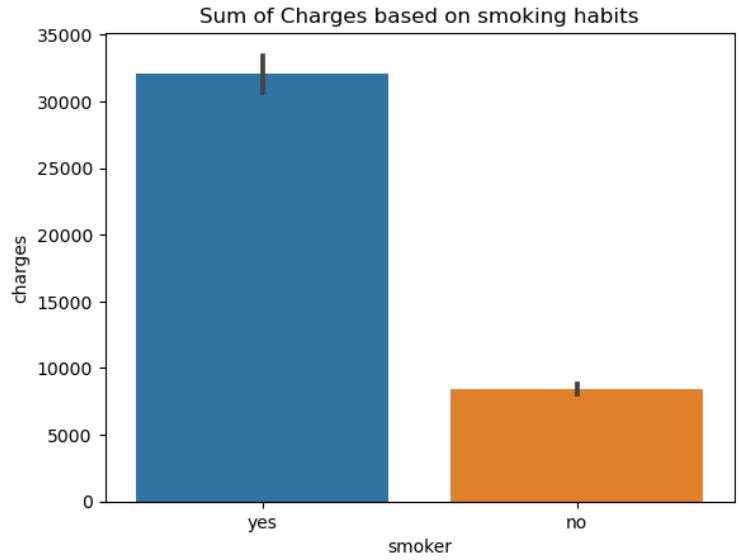
Not much difference.

5. Smoker:

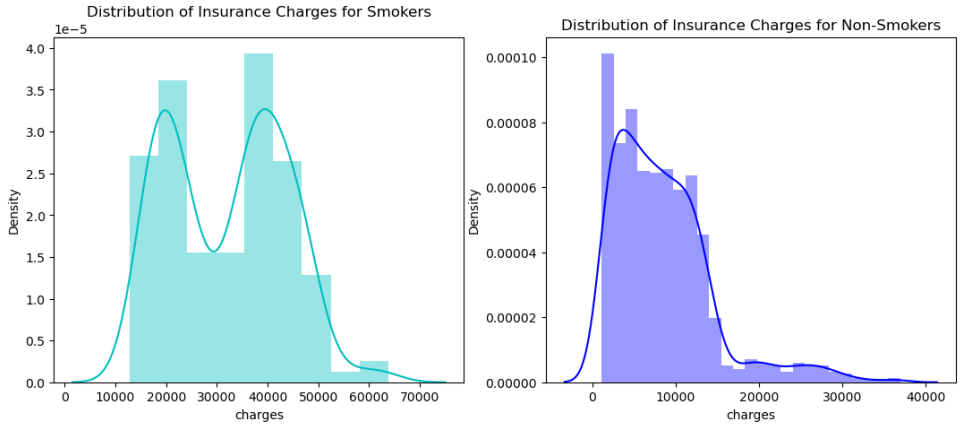


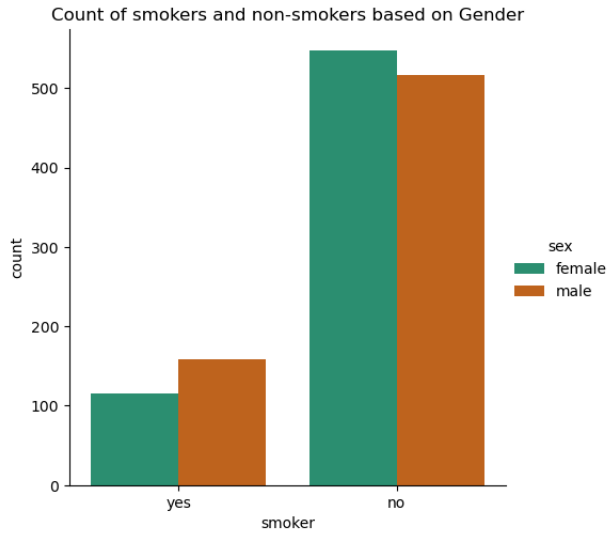
The dataset has more non-smokers than smokers.





Smokers pay much more Insurance Charges than non-smokers. Exact value of relationship between Charges and Smoker is: 0.79 (Value taken from the heatmap) which shows a strong relationship. This can be explained further by the following subplots:

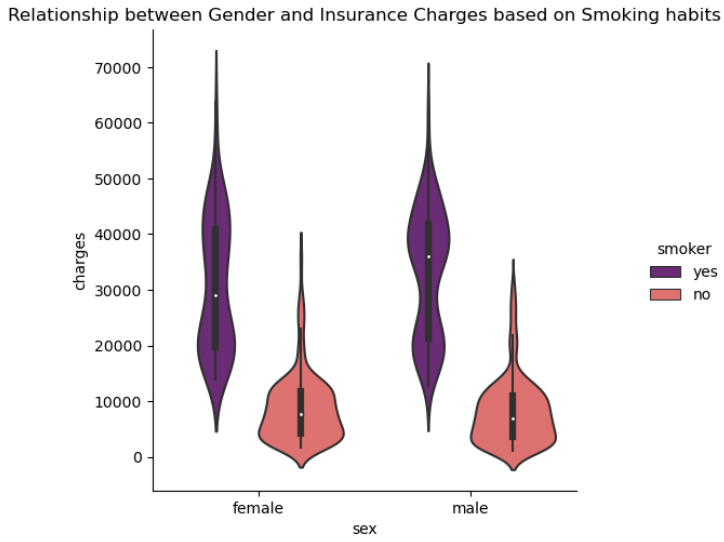




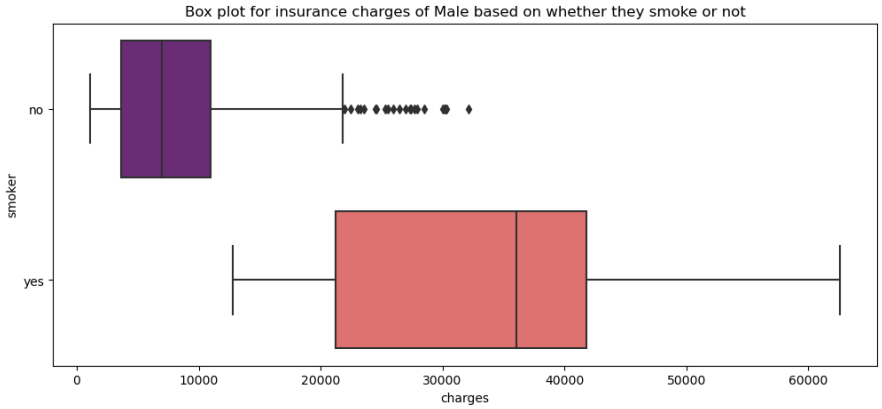
Plot explaining how many males and females in the dataset are smokers and how many are non-smokers.

We can notice that the dataset has more male smokers than women smokers and more female non-smokers than male non-smokers. And that can be explained below:

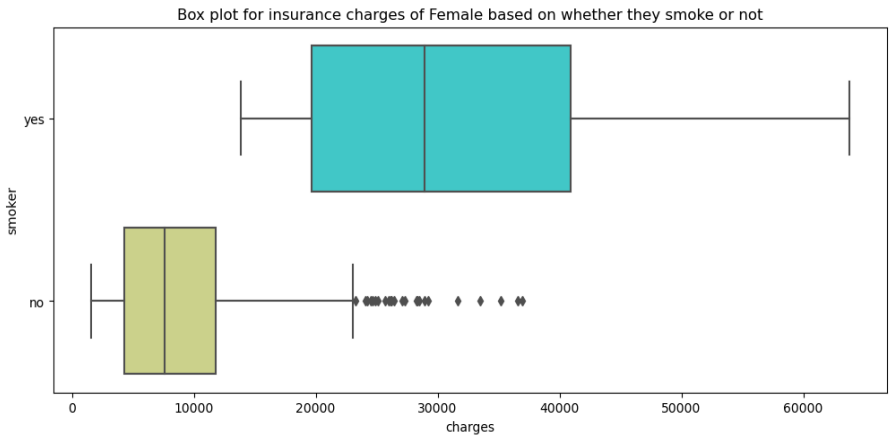
sex	smoker	count
female	no	547
	yes	115
male	no	517
	yes	159



Violin graph explaining relationship between Gender and Insurance Charges for Smokers and Non-Smokers.

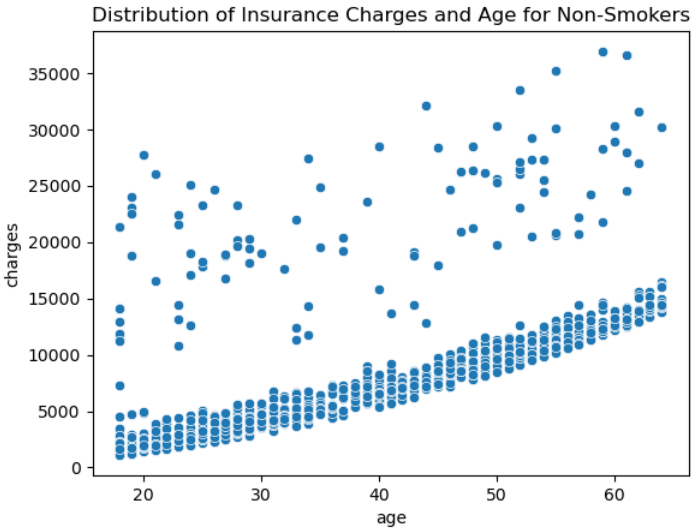


Box plot for Insurance Charges for male based on whether they smoke or not.

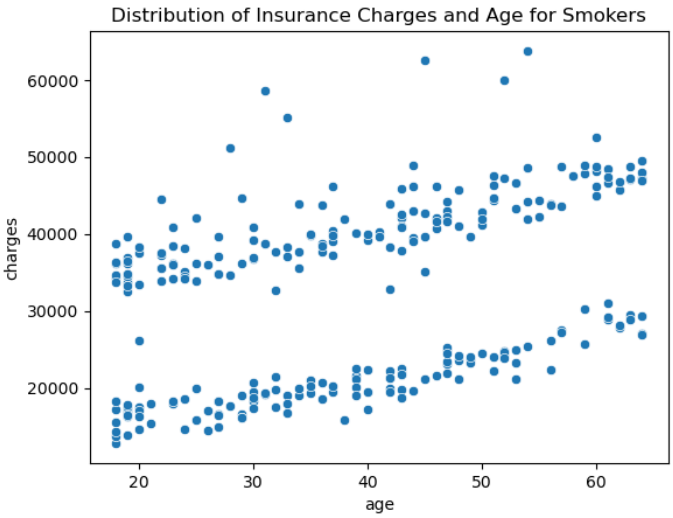


Box plot for Insurance Charges for female based on whether they smoke or not.

Now let's see how the Insurance Charges depends on the age of smokers and non-smokers patients.



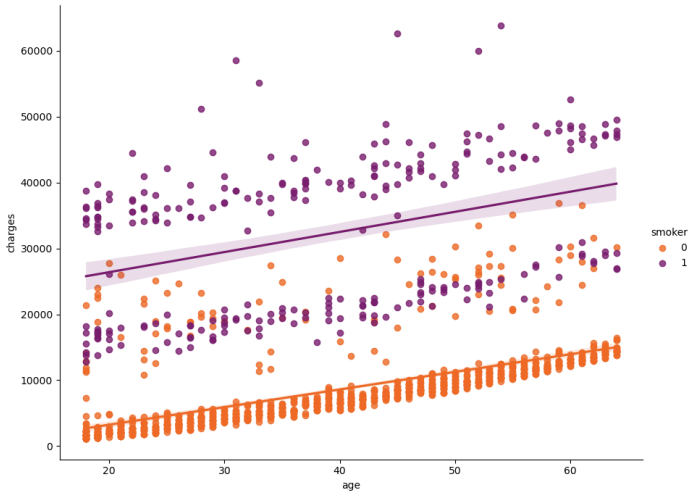
Distribution of Insurance Charges and age for non-smokers.



Distribution of Insurance Charges and age for smokers.

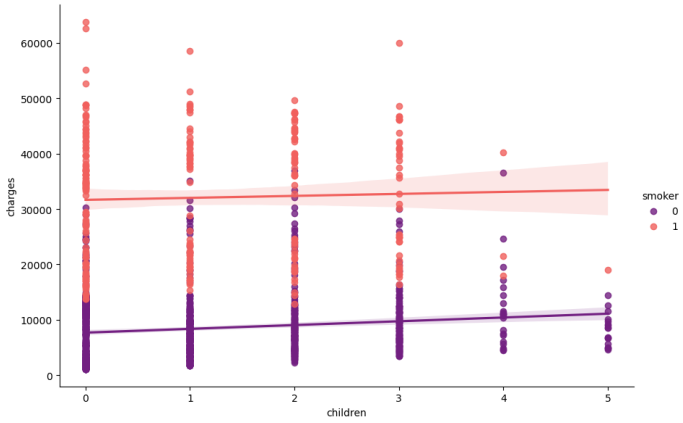
Again, it can be seen non-smokers, no matter how old they are, pay more charges.

Another way to look at it is:



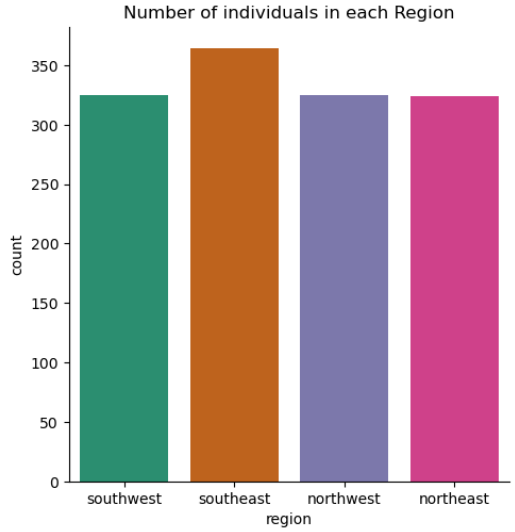
Above is the Linear regression plot between Age and Insurance Charges based on Smoking Habits with linear fit line. Here non-smokers are coded as “0” and smokers are coded as “1”.
The relationship between Charges and Age isn’t a very strong one. It has some effect but that’s not massive.

Above is the Linear regression plot between BMI and Insurance Charges based on Smoking Habits with linear fit line. Again, non-smokers are coded as “0” and smokers are coded as “1”. For smokers, increase in BMI, results in increase in Insurance charges but for non-smokers, there isn’t a strong relationship between BMI and charges.



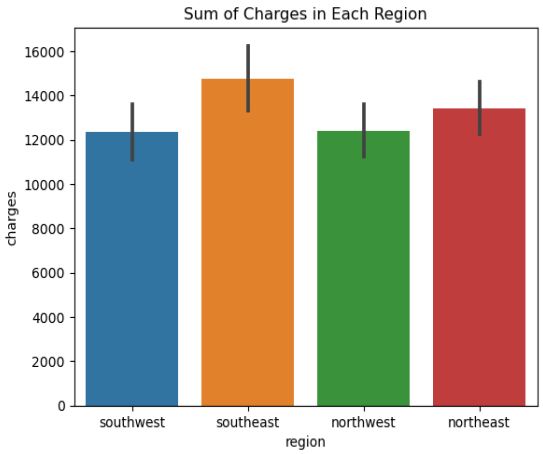
Above is the Linear regression plot between number of Children and Insurance Charges based on Smoking Habits with linear fit line. Non-smokers are coded as “0” and smokers are coded as “1”.
For smokers, we can see number of Children doesn’t play any role in deciding Charges. Its high anyways. But for non-smokers, charges are lower than smokers but then number of children has a slight effect on deciding Charges as can be seen by purple line in the graph.

6. Region:



Count of individuals (in the dataset) in each region.

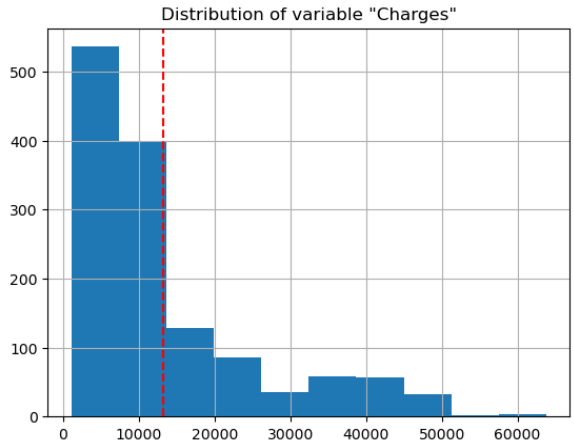
Maximum number of individuals are from southeast region.



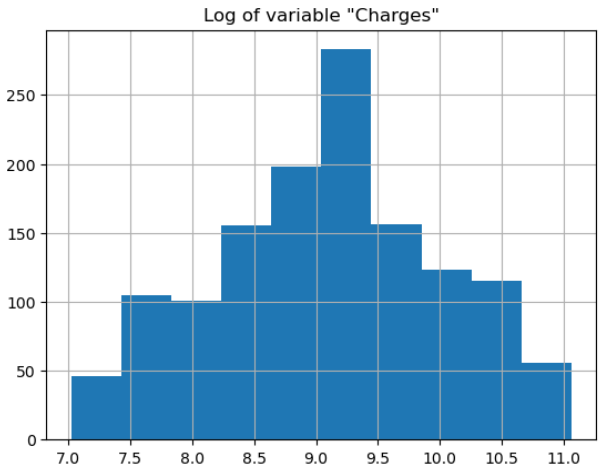
Sum of Insurance charges individuals are paying in each region.

Again, people in Southeast region are paying the most.

7. Charges:



This distribution is left-skewed. To make it closer to normal we can apply natural log.



It's better now. As distribution of target variable matters a lot in regression models.

Methods and Results:

1. First, I uploaded the data in a dataframe.
2. Then scaled the data. Numeric columns: age, bmi, children were scaled with StandardScaler.
3. Category labels were converted into numeric using LabelEncoder.
   Resulting in:

age	int64
sex	int64
bmi	float64
children	int64
smoker	int64
region	int64
charges	float64

4. Then I split the data into X and y. For the first model I just kept "smoker" column as X and "charges" as y, since "smoker" has a strong effect on deciding "charges" of an individual. Applied Linear Regression on it after train_test_split:

	Score
Training data MSE:	7539.347
Testing data MSE:	7170.328
Training data R^2 score:	0.612
Testing data R^2 score:	0.651

   Not very promising.
5. In second attempt, I took three most important features, "BMI", "age", "smoker" as X and "charges" as y. Applied linear regression after trian_test_split and got the following score:

	Score
Training data MSE:	6100.435
Testing data MSE:	6027.434
Training data R^2 score:	0.746
Testing data R^2 score:	0.753

Not promising either.

6. **Baseline Model:** Found out the mean_squared_error (MSE) of the baseline model which came out to be: 12105.485

7. **Mean of Target Variable:** With mean of target variable, y: 13270.422, baseline model cant be c onsidered a good model. With approx. $13,000 as the mean insurance charges, we don't want an aver age error of approximately $12,000 in our model.

8. I used **TransformedTargetRegressor()** on target variable but the results weren't any better in fact MSE score further dropped. So continued with target variable "charges" without TransformedTargetRegressor().

9. **Linear Regression Model:** This time, I took all input features as input variables (X), other than " charges" which was my "y" and fitted Linear Regression model on it.

```
Statistics of Linear Regression model:
```

	Score
Training data MSE:	6066.321
Testing data MSE:	5970.445
Training data R^2 score:	0.749
Testing data R^2 score:	0.758

There is an improvement but I want an even better score.

10. **Linear Regression Model with CV:** So I did cross validation with CV=5 on the data and fitted Linear Regression model again:

Statistics of Linear Regression model with cv=5:

	Score
Training data MSE (mean)	6105.716
Testing data MSE (mean)	6003.466
Training data R^2 score (mean)	0.740
Testing data R^2 score(mean)	0.746

Linear Regression with train_test_split gave better result than with cross validation.

11. **Ridge Model:**

```
Statistics of Ridge model:
```

	Score
Training data MSE:	6066.382
Testing data MSE:	5973.011
Training data R^2 score:	0.749
Testing data R^2 score:	0.758

Almost the same as Linear Regression model.

12. **Ridge Model with CV=5:**

Statistics of Ridge model with CV=5:

	Score
Training data MSE (mean)	6105.558
Testing data MSE (mean)	6005.934
Training data R^2 score (mean)	0.740
Testing data R^2 score(mean)	0.746

Score is the same as that of Linear Regression Model.

13. **Polynomial Features:**

Statistics of Polynomial Features model:

	Score
Training data MSE:	4706.201
Testing data MSE:	4967.868
Training data R^2 score:	0.849
Testing data R^2 score:	0.832

Better result than Linear Regression and ridge model.

14. **Polynomial Features with CV=5:**

Statistics of Polynomial Features model with CV=5:

	Score
Training data MSE (mean)	4767.863
Testing data MSE (mean)	5212.415
Training data R^2 score (mean)	0.842
Testing data R^2 score(mean)	0.806

15. **Polynomial Features after GridSearch:**

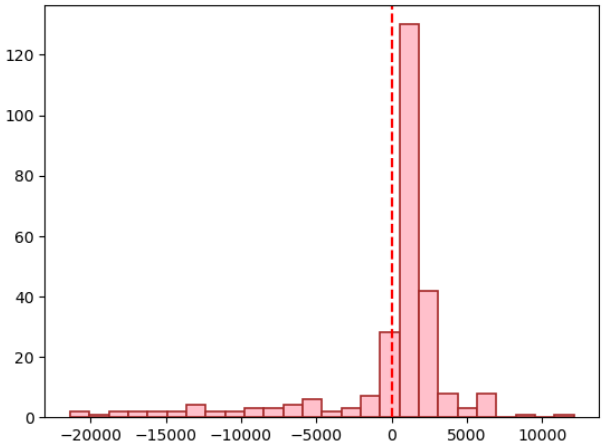
gree	params	split0_test_score	split1_test_score	split2_test_score	split3_test_score	split4_test_score	mean_test_score	std_test_score	rank_test_score
1	{'poly__degree': 1}	-4.537616e+07	-4.003810e+07	-2.659393e+07	-3.401052e+07	-4.193917e+07	-3.759158e+07	6.622070e+06	4
2	{'poly__degree': 2}	-2.925976e+07	-2.571161e+07	-1.387288e+07	-2.043780e+07	-2.615275e+07	-2.308696e+07	5.409410e+06	1
3	{'poly__degree': 3}	-3.183980e+07	-3.043526e+07	-1.616803e+07	-2.220704e+07	-2.738485e+07	-2.560700e+07	5.761501e+06	2
4	{'poly__degree': 4}	-3.335502e+07	-2.949647e+07	-2.464087e+07	-2.850669e+07	-3.296450e+07	-2.979271e+07	3.194789e+06	3
5	{'poly__degree': 5}	-7.430026e+07	-1.048492e+08	-1.181055e+08	-3.872253e+08	-7.280967e+07	-1.514580e+08	1.191720e+08	5
6	{'poly__degree': 6}	-8.748958e+11	-4.951641e+11	-2.597717e+10	-7.461261e+13	-7.790638e+09	-1.520329e+13	2.970641e+13	6
7	{'poly__degree': 7}	-1.245347e+14	-3.696724e+12	-2.327206e+13	-3.632950e+12	-4.935321e+12	-3.201435e+13	4.685532e+13	7
10	{'poly__degree': 10}	-2.691355e+14	-6.515160e+14	-8.668807e+13	-4.279663e+14	-1.498668e+14	-3.170345e+14	2.037069e+14	8

Best Estimator: {'poly__degree': 2}
Best MSE score: 4804.889240253629

	Y_true	Y_pred	diff
1231	20167.33603	13896.259766	-6271.076264
768	14319.03100	14937.904053	618.873053
847	2438.05520	1779.103271	-658.951929
510	11763.00090	13720.580811	1957.579911
363	2597.77900	3775.863770	1178.084770

Actual and predicted value of y along with the difference between them using polynomial Features degree 2.

There is a tendency for the y_true values being overestimated, as we can see in the histogram:



	Count
Underestimation	61
Exact Estimation	0
Overestimation	207

16. **Polynomial Features after GridSearch with CV=5:**

Statistics of Polynomial Features model with CV=5 and degree=2:

	Score
Training data MSE (mean)	4767.863
Testing data MSE (mean)	5212.415
Training data R^2 score (mean)	0.842
Testing data R^2 score(mean)	0.806

17. **Polynomial Features after GridSearch with CV=5("smoker", "age", "BMI" as X)**

Statistics of Polynomial Features model with CV=5, degree=2 and reduced features:

	Score
Training data MSE (mean):	4971.298
Testing data MSE (mean):	4396.937
Training data R^2 score (mean):	0.825
Testing data R^2 score(mean):	0.868

18. **PCA:**

```
Pipeline(steps=[('poly', PolynomialFeatures()), ('pca', PCA(n_components=10)),
                ('model', LinearRegression())])
Best Estimator: {'pca__n_components': 10}
Best score: 5959.234
```

	mean_fit_time	std_fit_time	mean_score_time	std_score_time	param_pca__n_components	params	split0_test_score	split1_test_score	split2_t
0	0.006952	0.002442	0.000892	0.000439	1	{'pca__n_components': 1}	-1.709222e+08	-1.253502e+08	-1.13
1	0.002522	0.000071	0.000492	0.000022	3	{'pca__n_components': 3}	-1.661895e+08	-1.222363e+08	-1.13
2	0.004425	0.002186	0.000731	0.000362	5	{'pca__n_components': 5}	-1.674235e+08	-1.228862e+08	-1.11
3	0.003719	0.001472	0.000477	0.000010	6	{'pca__n_components': 6}	-1.640407e+08	-1.198299e+08	-1.09
4	0.004774	0.000999	0.000933	0.000524	7	{'pca__n_components': 7}	-4.181227e+07	-3.815612e+07	-2.75
5	0.003750	0.000615	0.000669	0.000367	10	{'pca__n_components': 10}	-4.066285e+07	-3.753670e+07	-2.79
6	0.001144	0.000386	0.000000	0.000000	100	{'pca__n_components': 100}	NaN	NaN	

19. **PCA with GridSearch:**

Average MSE of Polynomial Features model with PCA and GridSearch is: 5945.246

20. **DecisionTree Regressor:**

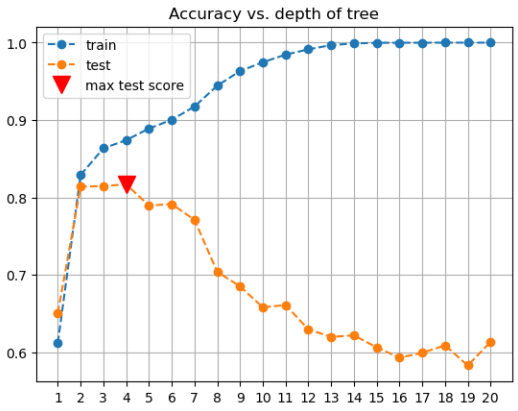
```
Statistics of Tree Regressor model with max_depth=3
```

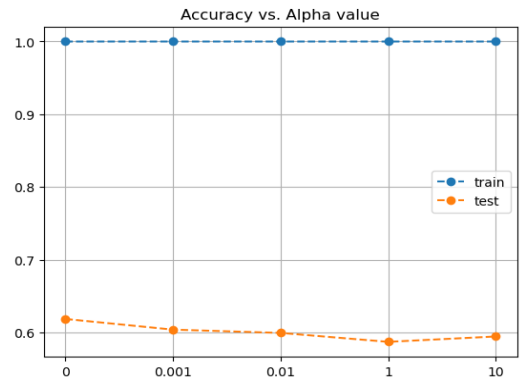
	Score
Training data MSE:	4471.388
Testing data MSE:	5222.981
Training data R^2 score:	0.863
Testing data R^2 score:	0.815

21. **DecisionTree Regressor with CV=5 and max_depth=3:**

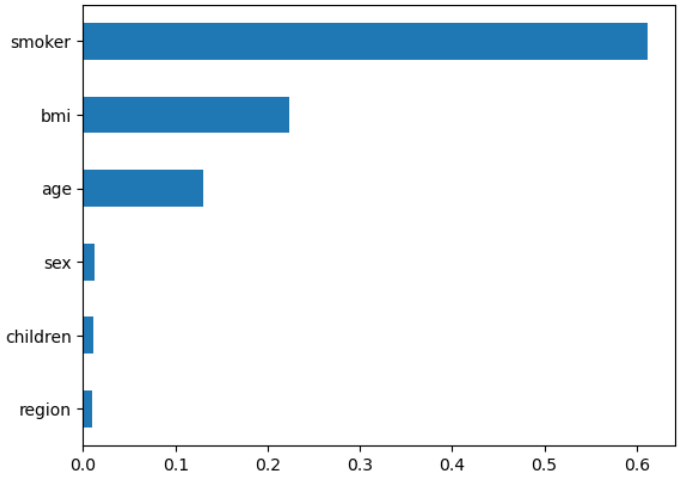
Statistics of Tree Regressor model with CV=5 and max_depth=3

	Score
Training data MSE (mean)	4608.201
Testing data MSE (mean)	5162.296
Training data R^2 score (mean)	0.851
Testing data R^2 score(mean)	0.809





Plotting importance_features:



22. DecisionTree Regressor after GridSearch (max_depth=4):

```
Statistics of DecisionTreeRegressor model with max_depth=4
```

	Score
Training data MSE:	4295.688
Testing data MSE:	5189.529
Training data R^2 score:	0.874
Testing data R^2 score:	0.817

23. DecisionTree Regressor after GridSearch (max_depth=4) and CV=5:

	Score
Training data MSE (mean)	4655.646
Testing data MSE (mean)	5650.149
Training data R^2 score (mean)	0.847
Testing data R^2 score(mean)	0.770

24. GradientBoosting Regressor:

```
Statistics of gboost model:
```

	Score
Training data MSE:	3836.164
Testing data MSE:	4024.666
Training data R^2 score:	0.897
Testing data R^2 score:	0.898

Getting some good results.

25. **GradientBoosting Regressor with reduced Features (“smoker”, “age”, “BMI”):**

```
Statistics of gboost model with 3 Features:
```

	Score
Training data MSE:	3834.743
Testing data MSE:	5012.398
Training data R^2 score:	0.900
Testing data R^2 score:	0.829

Score was better with all input Features.

26. **GradientBoosting Regressor after GridSearch:**

'max_depth': 3, 'min_samples_leaf': 9, 'min_samples_split': 2, 'n_estimators': 50

	Score
Training data MSE:	4233.361
Testing data MSE:	3972.907
Training data R^2 score:	0.875
Testing data R^2 score:	0.901

Best score calculated so far ☺

27. **GradientBoosting Regressor after GridSearch with reduced features (“smoker”, “age”, “BMI”):**

```
Statistics of gboost model with 3 Features and after GridSearch:
```

	Score
Training data MSE:	4188.445
Testing data MSE:	4910.606
Training data R^2 score:	0.880
Testing data R^2 score:	0.836

Previous score was better.

28. **GradientBoosting Regressor with CV=5:**

	Score
Training data MSE (mean)	4756.995
Testing data MSE (mean)	4606.761
Training data R^2 score (mean)	0.840
Testing data R^2 score(mean)	0.857

Result with train_test_split was better.

29. **GradientBoosting Regressor with CV=5 and reduced features (“smoker”, “age”, “BMI”):**

	Score
Training data MSE (mean)	4669.743
Testing data MSE (mean)	5666.585
Training data R^2 score (mean)	0.846
Testing data R^2 score(mean)	0.763

No luck ☹

30. **GradientBoosting Regressor with CV=5 and after GridSearch:**

	Score
Training data MSE (mean)	4628.957
Testing data MSE (mean)	4281.886
Training data R^2 score (mean)	0.848
Testing data R^2 score(mean)	0.875

31. **GradientBoosting Regressor with CV=5 and after GridSearch and reduced Features ("smoker", "age", "BMI"):**

	Score
Training data MSE (mean)	4451.317
Testing data MSE (mean)	5085.594
Training data R^2 score (mean)	0.861
Testing data R^2 score(mean)	0.814

32. **BaggingRegressor:**

```
Statistics of BaggingRegressor model:
```

	Score
Training data MSE:	4388.486
Testing data MSE:	4003.088
Training data R^2 score:	0.866
Testing data R^2 score:	0.899

33. **Postpruned BaggingRegressor:**

n_estimators= 100, ccp_alpha=0.001

```
Statistics of Postpruned Bagging Regressor model:
```

	Score
Training data MSE:	1932.486
Testing data MSE:	4409.670
Training data R^2 score:	0.974
Testing data R^2 score:	0.878

34. **BaggingRegressor with CV=5:**

	Score
Training data MSE (mean)	4767.247
Testing data MSE (mean)	4486.648
Training data R^2 score (mean)	0.839
Testing data R^2 score(mean)	0.865

35. **Postpruned BaggingRegressor with CV=5:**

	Score
Training data MSE (mean)	5124.744
Testing data MSE (mean)	4755.991
Training data R^2 score (mean)	0.814
Testing data R^2 score(mean)	0.847

36. **RandomForest Regressor:**

```
Statistics of Random Forest model:
```

	Score
Training data MSE:	1930.205
Testing data MSE:	4425.672
Training data R^2 score:	0.974
Testing data R^2 score:	0.877

37. **RandomForest Regressor with reduced features (“smoker”, “age”, “BMI”):**

```
Statistics of Random Forest model with 3 Features:
```

	Score
Training data MSE:	1912.864
Testing data MSE:	5757.353
Training data R^2 score:	0.975
Testing data R^2 score:	0.775

38. **RandomForest Regressor after GridSearch:**

```
Statistics of Random Forest model after GridSearch:
```

	Score
Training data MSE:	4395.499
Testing data MSE:	3947.109
Training data R^2 score:	0.865
Testing data R^2 score:	0.902

Good score ☺

39. **RandomForest Regressor after GridSearch and reduced features (“smoker”, “age”, “BMI”):**

```
Statistics of Random Forest model with 3 Features and after GridSearch:
```

	Score
Training data MSE:	4150.845
Testing data MSE:	4956.903
Training data R^2 score:	0.882
Testing data R^2 score:	0.833

Score dropping again ☹

40. **RandomForest Regressor with CV=5:**

	Score
Training data MSE (mean)	5127.437
Testing data MSE (mean)	4780.841
Training data R^2 score (mean)	0.814
Testing data R^2 score(mean)	0.846

We have got better score than this.

41. **RandomForest Regressor with CV=5 and reduced features (“smoker”, “age”, “BMI”):**

	Score
Training data MSE (mean)	4990.118
Testing data MSE (mean)	5451.890
Training data R^2 score (mean)	0.825
Testing data R^2 score(mean)	0.783

42. **RandomForest Regressor with CV=5 after GridSearch:**

	Score
Training data MSE (mean)	4637.763
Testing data MSE (mean)	4310.514
Training data R^2 score (mean)	0.848
Testing data R^2 score(mean)	0.874

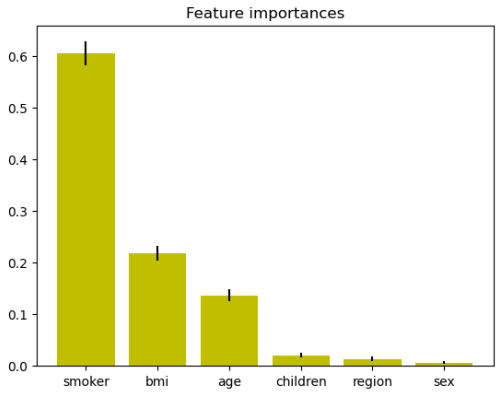
43. **RandomForest Regressor with CV=5 and reduced features (“smoker”, “age”, “BMI”) after GridSearch:**

	Score
Training data MSE (mean)	4463.977
Testing data MSE (mean)	5060.783
Training data R^2 score (mean)	0.860
Testing data R^2 score(mean)	0.814

Feature importance ranking with RandomForest Regressor:

1. smoker. (0.604253)
2. BMI. (0.218423)
3. age. (0.136862)
4. children. (0.020854)
5. region. (0.013593)
6. sex. (0.006015)

This is the same as we had established earlier.



44. **XGBoost Regressor:**

```
Statistics of xgboost model:
```

	Score
Training data MSE:	724.217
Testing data MSE:	5711.582
Training data R^2 score:	0.996
Testing data R^2 score:	0.778

45. **XGBoost Regressor with reduced features (“smoker”, “age”, “BMI”):**

```
Statistics of xgboost model with 3 Features:
```

	Score
Training data MSE:	1305.936
Testing data MSE:	6002.028
Training data R^2 score:	0.988
Testing data R^2 score:	0.755

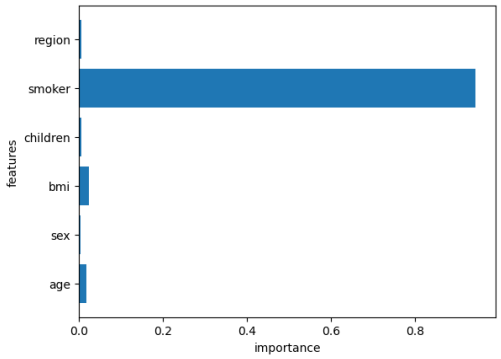
46. **XGBoost Regressor with CV=5:**

	Score
Training data MSE (mean)	5134.179
Testing data MSE (mean)	5833.671
Training data R^2 score (mean)	0.815
Testing data R^2 score(mean)	0.758

47. **XGBoost Regressor with CV=5 and reduced features (“smoker”, “age”, “BMI”):**

	Score
Training data MSE (mean)	5307.836
Testing data MSE (mean)	6046.558
Training data R^2 score (mean)	0.801
Testing data R^2 score(mean)	0.735

Plotting Feature Importance for XGBoost Model:



48. **XGBoost Regressor Hyperparameter tuning with GridSearch:**

Best parameters: {'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 500, 'subsample': 0.7}
Best score: 4513.531

49. **XGBoost Regressor after GridSearch:**

```
Statistics of xgboost model after GridSearch:
```

	Score
Training data MSE:	4029.055
Testing data MSE:	4744.108
Training data R^2 score:	0.889
Testing data R^2 score:	0.847

50. **XGBoost Regressor after GridSearch with reduced features (“smoker”, “age”, “BMI”):**

```
Statistics of xgboost model with 3 Features and after GridSearch:
```

	Score
Training data MSE:	4147.019
Testing data MSE:	4864.881
Training data R^2 score:	0.883
Testing data R^2 score:	0.839

51. **XGBoost Regressor after GridSearch and CV=5:**

	Score
Training data MSE (mean)	4389.679
Testing data MSE (mean)	4996.684
Training data R^2 score (mean)	0.865
Testing data R^2 score(mean)	0.820

52. **XGBoost Regressor after GridSearch and CV=5 with reduced features (“smoker ”, “age”, “BMI”):**

	Score
Training data MSE (mean)	4466.115
Testing data MSE (mean)	5045.722
Training data R^2 score (mean)	0.860
Testing data R^2 score(mean)	0.817

53. **Lasso Regression:**

alpha=0.2, fit_intercept=True, precompute=False, max_iter=1000, tol=0.0001

```
Statistics of Lasso model:
```

	Score
Training data MSE:	6142.441
Testing data MSE:	5643.300
Training data R^2 score:	0.737
Testing data R^2 score:	0.800

54. **Lasso Regression with CV=5:**

alpha=0.2, fit_intercept=True, precompute=False, max_iter=1000, tol=0.0001

	Score
Training data MSE (mean)	6166.919
Testing data MSE (mean)	5605.001
Training data R^2 score (mean)	0.731
Testing data R^2 score(mean)	0.785

55. Lasso Regression coefficients:

	features	coef
2	smoker_no	-2.384504e+04
6	region_southeast	-1.429208e+02
7	region_southwest	-6.836411e+01
1	sex_male	-4.751916e-13
3	smoker_yes	1.891499e-11
0	sex_female	1.290663e+02
5	region_northwest	5.343800e+02
10	children	5.725270e+02
4	region_northeast	8.873688e+02
9	bmi	2.066632e+03

7 features are greater than 0 which are:
['smoker_yes', 'sex_female', 'region_northwest', 'children', 'region_northeast', 'bmi', 'age']

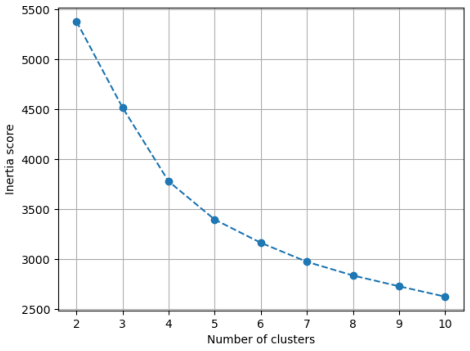
4 features are less than 0 which are:
['smoker_no', 'region_southeast', 'region_southwest', 'sex_male']

56. Neural Netwroks:

	Score
Training data MSE:	11711.695
Testing data MSE:	12155.368
Training data R^2 score:	0.043
Testing data R^2 score:	0.071

No need to explore this regressor any further as score is the worse so far.

57. KMeans Clustering:



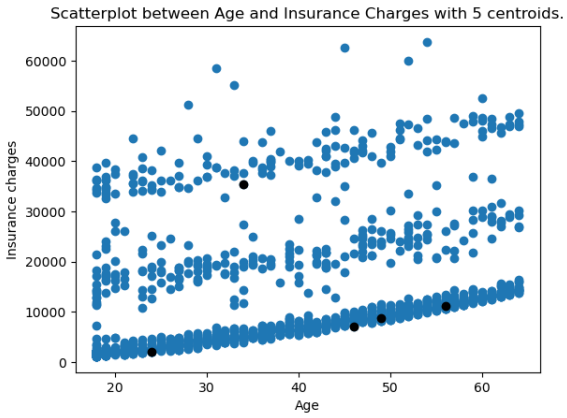
1. As per above graph, number of clusters should be 5.
2. KMeans Clustering inertia is: 3398.2495711048987
3. The number of iterations required to converge: 17
4. First five predicted labels: [0 0 3 4 0]
5. KMeans predicted values: [0 0 3 ... 0 0 4]

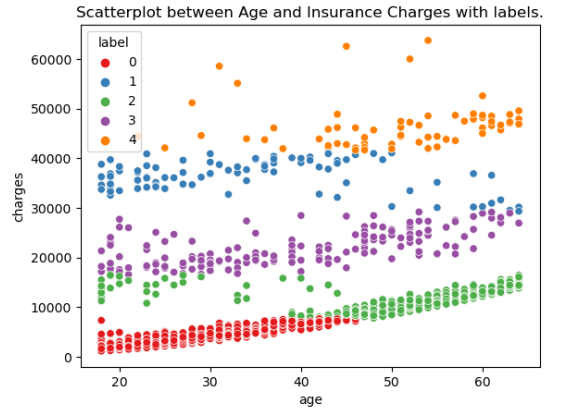
```
cluster
0    377
3    297
4    285
1    223
2    156
```

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

Dataframe after adding labels.

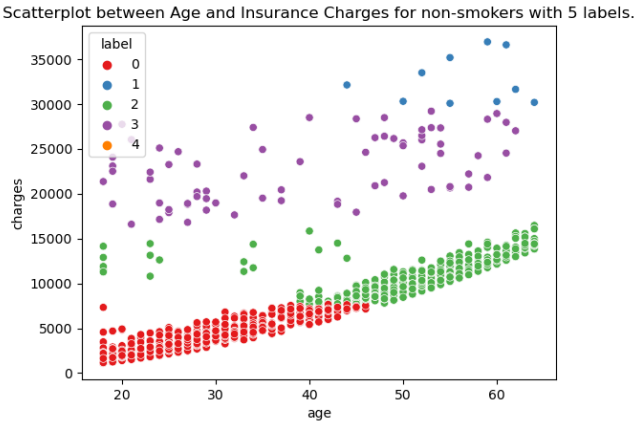
Now to study these 5 clusters in detail.

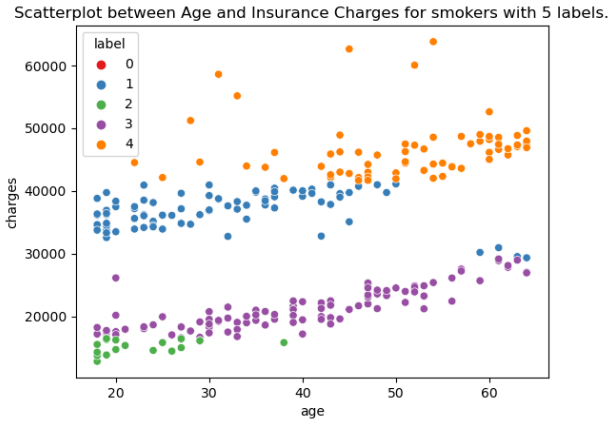




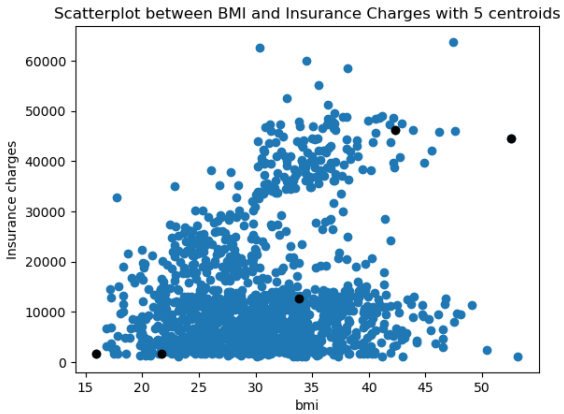
Cluster number 1 has people from age groups of 18 – 45 years roughly with charges under $10,000.
Cluster number 2 has people from all age groups but charges roughly in the range of $30,000 – $42,000.
Cluster number 3 has people from all age groups but charges roughly in the range of $8,000 - $19,000.
Cluster number 4 has people from all age groups but charges roughly in the range of $16,000 - $30,000.
Cluster number 5 has people in all age groups but charges roughly in the range of $40,000-$60,000.

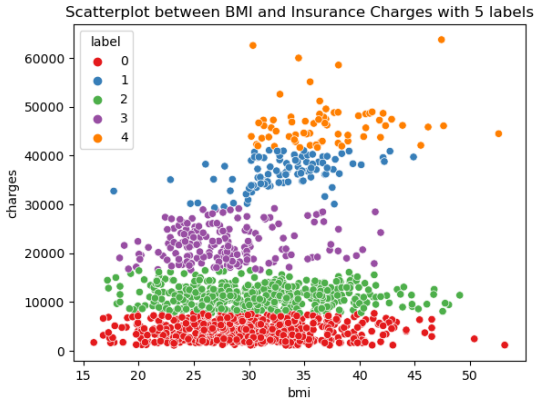
This infers clusters are made mostly based on insurance charges.



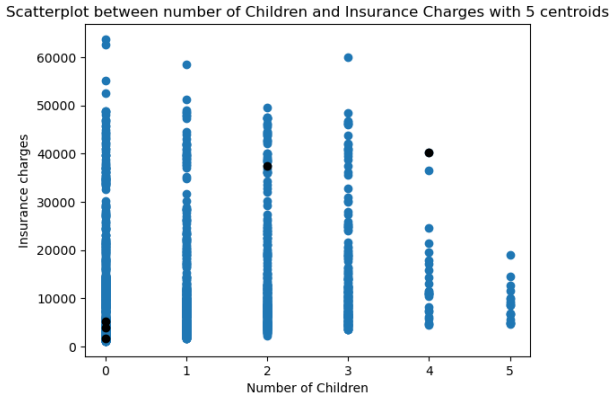


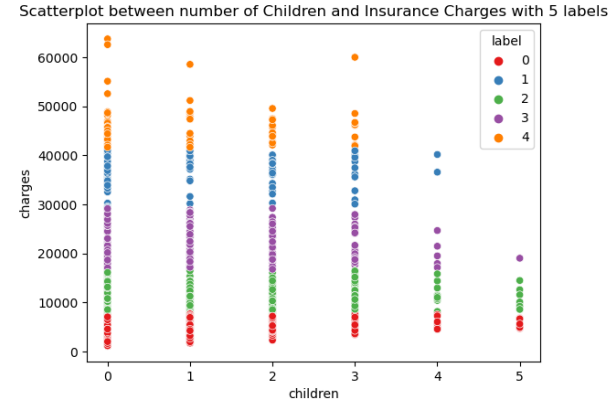
No smokers in cluster number 1.
And no non-smokers in cluster number 5.



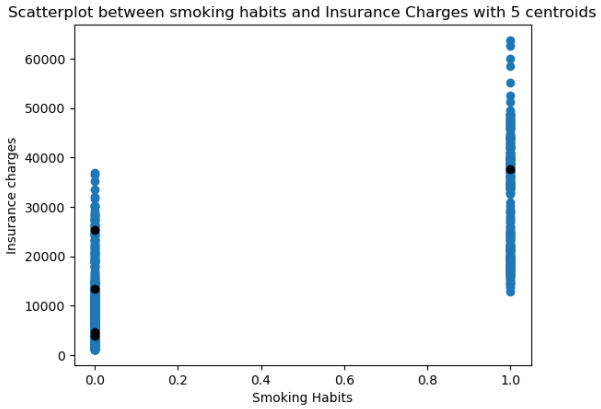


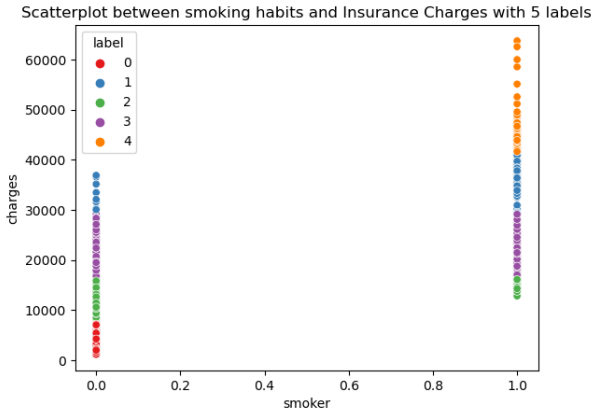
People with BMI >30 are in cluster number 5 and have the highest Charges.
Cluster number 4 has more people in it with BMI<30 so comparatively lesser Charges than people in cluster 5.
Cluster number 2 has more people in it with BMI>30 so higher Charges than people in cluster 4.
Cluster number 1 and 3 have people from both categories of BMI evenly distributed but lower Charges maybe all non-smokers are in these two clusters.





Again cluster 5 have people with highest Charges and have people with 3 or less children.
Cluster 2 have people with high Charges with 4 or a smaller number of children.
Rest of the clusters (1,3,4) have people from both categories that is people with children and no children but overall have lesser charges.





Cluster 1 has all non- smokers in it and so least Charges.
Cluster 5 has all smokers in it so maximum Charges.
Cluster 3 has more non-smokers in it than smokers so comparatively lesser charges.
Cluster 2 and 4 have a mix of both groups almost evenly distributed.

Summary of 5 Clusters:

Cluster 1 (Red):

People from age group 18-45
Charges under $10,000
People from both categories of BMI
People with children or no child
All non-smokers.

Cluster 2 (Blue):

People from all age groups
Charges: $30,000-$42,000
BMI> 30 so higher charges
People with 4 or a smaller number of children
Both smokers and non-smokers

Cluster 3 (Green):

People from all age groups
Charges: $8,000 - $19,000
BMI from both categories
People with children or no child.
Less smokers, more non-smokers so lower Charges.

Cluster 4 (Purple):
People from all age groups
Charges: $16,000-$30,000
More people with BMI < 30 so lesser charges
People with children or no child
Both smokers and non-smokers

Cluster 5 (Orange):
People from all age groups.
Highest charges: $40,000-$60,000
People with BMI>30 so higher Charges
People with 3 or a smaller number of kids
All smokers so high Charges.

Summary of Results:

Sr. No	Model Name	Train test split data	CV (Yes or No)	Hyperparameters	Training MSE	Testing MSE	Training R^2	Testing R^2
1	Linear Regression	Only Smoker column	No		7539.347	7170.328	0.612	0.651
2	Linear Regression	Smoker, Age, BMI	No		6100.435	6027.434	0.746	0.753
3	Baseline model				12105.85			
4	Y mean of target variable				13270.422			
5	Linear Regression	Yes, with all features	No		6066.321	5970.445	0.749	0.758
6	Linear Regression	All Features	CV=5		6105.716	6003.466	0.740	0.746
7	Ridge Regression Model	Yes, will all features	No		6066.382	5973.011	0.749	0.758
8	Ridge Regression Model	All Features	CV=5		6105.558	6005.934	0.740	0.746
9	Polynomial Features	Yes, with all Features	No		4706.201	4967.868	0.849	0.832
10	Polynomial Features	All Features	CV=5		4767.863	5212.415	0.842	0.806
11	Polynomial Features with GridSearch	Yes, All Features	No	Degree=2	4706.201	4967.868	0.849	0.832
12	Polynomial Features with GridSearch	All Features	CV=5	Degree=2	4767.863	5212.415	0.842	0.806
13	Polynomial Features	With Smoker,age, BMI	CV=5	Degree=2	4971.298	4396.937	0.825	0.868
14	PCA	Yes, with all Features				5959.234		
15	PCA with GridSearch	Yes, with all Features	No	n_components (PCA)=10		5945.246		
16	DecisionTreeRegressor	Yes, with all Features	No	max_depth=3	4471.388	5222.981	0.863	0.815
17	DecisionTreeRegressor	With all Features	CV=5	max_depth=3	4608.201	5162.296	0.851	0.809
18	DecisionTreeRegressor	Yes, with all Features		max_depth=4	4295.688	5189.529	0.874	0.817
19	DecisionTreeRegressor	With all Features	CV=5	max_depth=4	4655.646	5650.149	0.847	0.770
20	BaggingRegressor	Yes, with all Features	No		4388.486	4003.088	0.866	0.899
21	BaggingRegressor	With all Features	CV=5		4767.247	4486.648	0.839	0.865
22	PostPruned BaggingRegressor	Yes, with all Features	No	n_estimators=100, ccp_alpha=0.001	1932.486	4409.670	0.974	0.878
23	Postpruned BaggingRegressor	With all Features	CV=5	n_estimators=100, ccp_alpha=0.001	5124.744	4755.991	0.814	0.847
24	Gradient Boosting Model	Yes, with all Features	No		3836.164	4024.666	0.897	0.898
25	Gradient Boosting model	Smoking, Age, BMI	No		3834.743	5012.398	0.900	0.829
26	Gradient Boosting Model	With all Features	CV=5		4756.995	4606.761	0.840	0.857
27	Gradient Boosting Model	Smoking, Age, BMI	CV=5		4669.743	5666.585	0.846	0.763
29	Gradient Boosting Model after GridSearch	Yes, with all Features	No	'max_depth': 3, 'min_samples_leaf': 9, 'min_samples_split': 2, 'n_estimators': 50	4233.361	3972.907	0.875	0.901
30	Gradient Boosting Model after GridSearch	Smoking, age, BMI	No	'max_depth': 3, 'min_samples_leaf': 10, 'min_samples_split': 2, 'n_estimators': 50	4188.445	4910.606	0.880	0.836
31	Gradient Boosting Model after GridSearch	With all Features	CV=5	'max_depth': 3, 'min_samples_leaf': 9, 'min_samples_split': 2, 'n_estimators': 50	4628.957	4281.886	0.848	0.875
32	Gradient Boosting Model after GridSearch	Smoking, Age, BMI	CV=5	'max_depth': 3, 'min_samples_leaf': 10, 'min_samples_split': 2, 'n_estimators': 50	4451.317	5085.594	0.861	0.814

33	Random Forest Regressor	Yes, with all Features	No		1930.205	4425.672	0.974	0.877
34	Random Forest Regressor	Smoking, Age, BMI	No		1912.864	5757.353	0.975	0.775
35	Random Forest Regrerssor	With all Features	CV=5		5127.437	4780.841	0.814	0.846
36	Random Forest Regrerssor	Smoking, age, BMI	CV=5		4990.118	5451.890	0.825	0.783
37	Random Forest Regressor after GridSearch	Yes, with all Features	No	max_depth=4, min_samples_leaf=7, min_samples_split=2, n_estimators=200	4395.499	3947.109	0.865	0.902
38	Random Forest Regressor after GridSearch	Smoking, Age, BMI	No	max_depth=5, min_samples_leaf=7, min_samples_split=2, n_estimators=600	4150.845	4956.903	0.882	0.833
39	Random Forest Regressor after GridSearch	With all Features	CV=5	max_depth=4, min_samples_leaf=7, min_samples_split=2, n_estimators=200	4637.763	4310.514	0.848	0.874
40	Random Forest Regressor after GridSearch	Smoking, age, BMI	CV=5	max_depth=5, min_samples_leaf=7, min_samples_split=2, n_estimators=600	4463.977	5065.783	0.860	0.814
41	XGBoost Model	Yes, with all Features	No		724.217	5711.582	0.996	0.778
42	XGBoost Model	Smoking, Age, BMI	No		1305.936	6002.028	0.988	0.755
43	XGBoost Model	With all Features	CV=5		5134.179	5833.671	0.815	0.758
44	XGBoost Model	Smoking, Age, BMI	CV=5		5307.836	6046.558	0.801	0.735
45	XGBoost model with GridSearch	Yes, with all Features	No	learning_rate=0.01, max_depth=3, n_estimators=500, subsample=0.7	4029.055	4744.108	0.889	0.847
46	XGBoost model with GridSearch	Smoking, Age, BMI	No	learning_rate=0.01, max_depth=2, n_estimators=500, subsample=0.7	4147.019	4864.881	0.883	0.839
47	XGBoost model with GridSearch	With all Features	CV=5	learning_rate=0.01, max_depth=3, n_estimators=500, subsample=0.7	4389.679	4996.684	0.865	0.820
48	XGBoost model with GridSearch	Smoking, Age, BMI	CV=5	learning_rate=0.01, max_depth=2, n_estimators=500, subsample=0.7	4466.115	5045.722	0.860	0.817
49	Lasso model	Yes, with all Features	No	alpha=0.2, fit_intercept=True, precompute=False, max_iter=1000, tol=0.0001	6142.441	5643.300	0.737	0.800
50	Lasso model	With all Features	CV=5	alpha=0.2, fit_intercept=True, precompute=False, max_iter=1000, tol=0.0001	6166.919	5605.001	0.731	0.785
51	Neural Network Regressor	Yes, with all Features	No	alpha=0.0001	11588.598	12009.667	0.063	0.094

Findings:

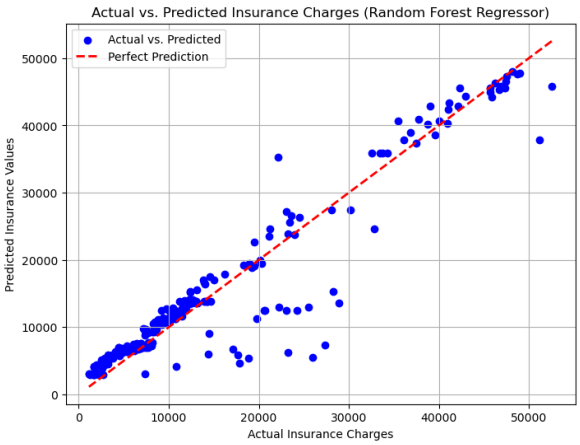
From the scores that I got, following are the three best models:

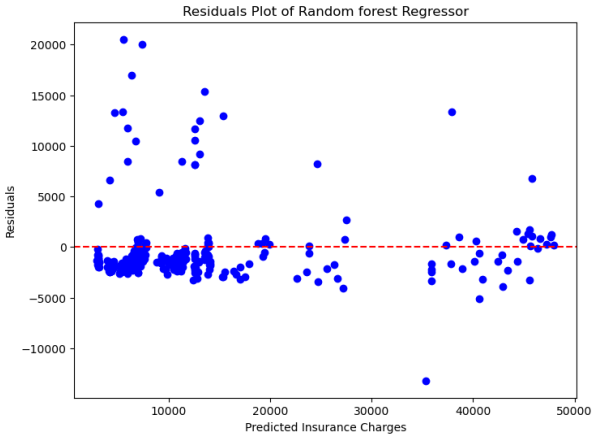
- Random Forest Regressor with mean squared error score: 3947.109
- Gradient Boosting Regressor with mean squared error score: 3972.907
- Bagging Regressor Model with mean squared error score: 4003.088

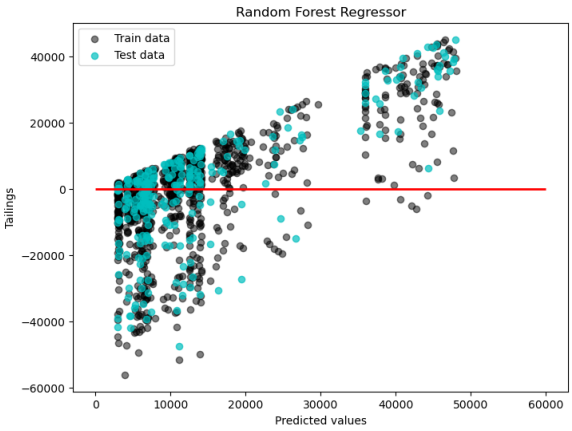
Random Forest Regressor:

	Actual	Predicted	Difference
578	9724.53000	12790.451985	-3065.921985
610	8547.69130	10279.040273	-1731.348973
569	45702.02235	44959.122176	742.900174
1034	12950.07120	13870.043275	-919.972075
198	9644.25250	11079.262089	-1435.009589
...	...	...	...
1084	15019.76005	17023.598473	-2003.838423
726	6664.68595	7136.744908	-472.058958
1132	20709.02034	12543.849642	8165.170698
725	40932.42950	40315.286253	617.143247
963	9500.57305	10695.596504	-1195.023454

268 rows × 3 columns



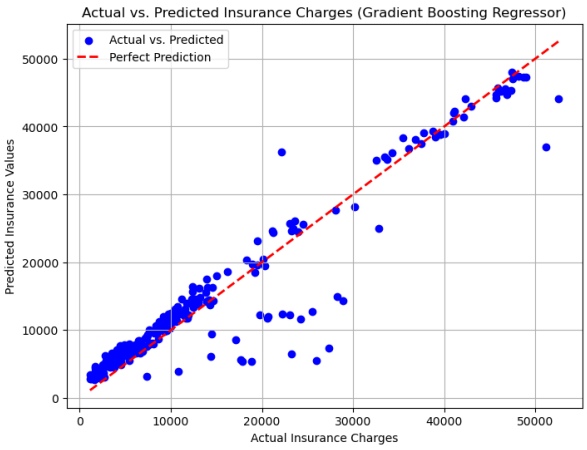


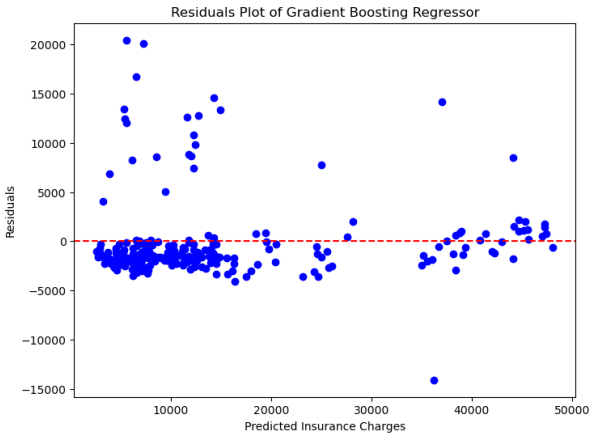


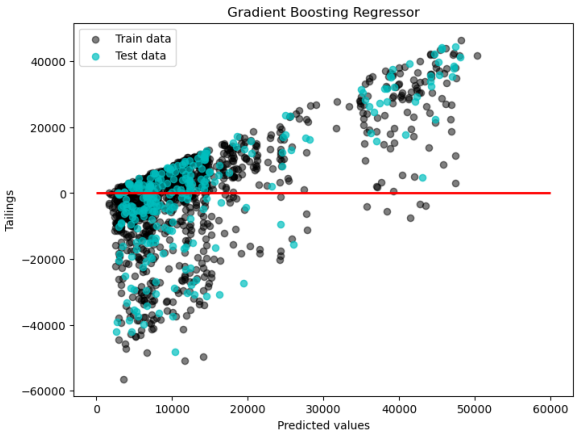
Gradient Boosting Regressor:

	Actual	Predicted	Difference
578	9724.53000	12361.630351	-2637.100351
610	8547.69130	9945.731053	-1398.039753
569	45702.02235	44667.364954	1034.657396
1034	12950.07120	14325.949720	-1375.878520
198	9644.25250	11383.398532	-1739.146032
...	...	...	...
1084	15019.76005	18010.239798	-2990.479748
726	6664.68595	7702.004651	-1037.318701
1132	20709.02034	12061.528818	8647.491522
725	40932.42950	40839.184985	93.244515
963	9500.57305	9964.799572	-464.226522

268 rows × 3 columns



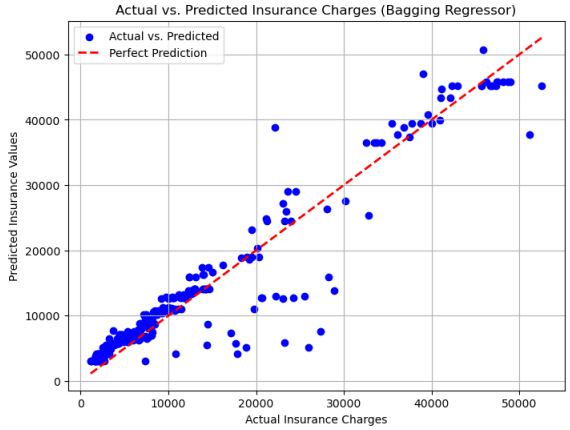


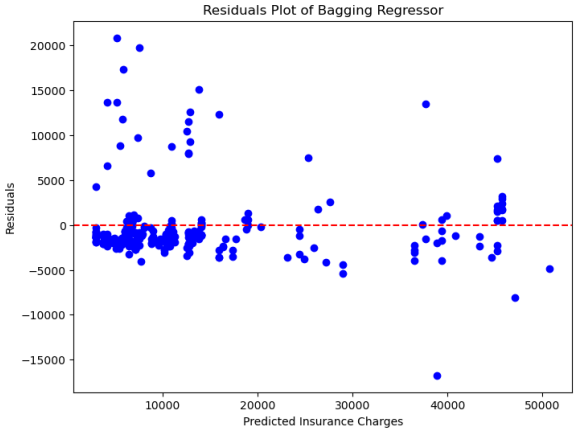


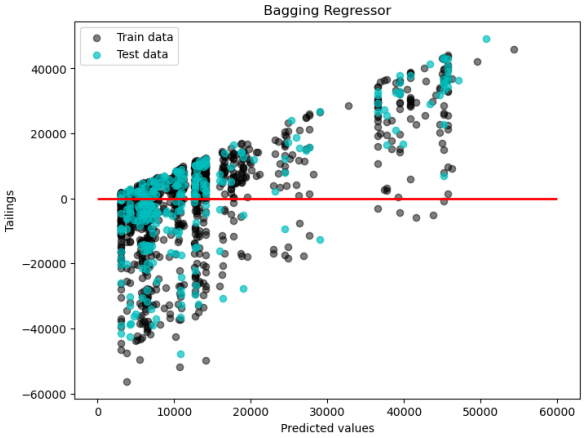
Bagging Regressor Model:

	Actual	Predicted	Difference
578	9724.53000	12799.495736	-3074.965736
610	8547.69130	10492.695297	-1945.003997
569	45702.02235	45211.402486	490.619864
1034	12950.07120	14108.278859	-1158.207659
198	9644.25250	10811.522976	-1167.270476
...	...	...	...
1084	15019.76005	16604.303458	-1584.543408
726	6664.68595	6805.590266	-140.904316
1132	20709.02034	12697.505384	8011.514956
725	40932.42950	39917.805677	1014.623823
963	9500.57305	10823.716503	-1323.143453

268 rows × 3 columns


Actual vs. Predicted Insurance Charges (Bagging Regressor)
Actual vs. Predicted
Perfect Prediction
Predicted Insurance Values
Actual Insurance Charges
0
10000
20000
30000
40000
50000


Residuals Plot of Bagging Regressor
Residuals
Predicted Insurance Charges
20000
15000
10000
5000
0
−5000
−10000
−15000
10000
20000
30000
40000
50000


Bagging Regressor
Train data
Test data
Tailings
Predicted values
40000
20000
0
−20000
−40000
−60000
0
10000
20000
30000
40000
50000
60000

Comparison of the predicted values of three best models:

	Actual	Predicted GBR	Predicted RFR	Predicted BR	Diff GBR	Diff RFR	Diff BR
578	9724.53000	12361.630351	12790.451985	12799.495736	-2637.100351	-3065.921985	-3074.965736
610	8547.69130	9945.731053	10279.040273	10492.695297	-1398.039753	-1731.348973	-1945.003997
569	45702.02235	44667.364954	44959.122176	45211.402486	1034.657396	742.900174	490.619864
1034	12950.07120	14325.949720	13870.043275	14108.278859	-1375.878520	-919.972075	-1158.207659
198	9644.25250	11383.398532	11079.262089	10811.522976	-1739.146032	-1435.009589	-1167.270476
...	...	...	...	...	...	...	...
1084	15019.76005	18010.239798	17023.598473	16604.303458	-2990.479748	-2003.838423	-1584.543408
726	6664.68595	7702.004651	7136.744908	6805.590266	-1037.318701	-472.058958	-140.904316
1132	20709.02034	12061.528818	12543.849642	12697.505384	8647.491522	8165.170698	8011.514956
725	40932.42950	40839.184985	40315.286253	39917.805677	93.244515	617.143247	1014.623823
963	9500.57305	9964.799572	10695.596504	10823.716503	-464.226522	-1195.023454	-1323.143453

268 rows × 7 columns

Here GBR is Gradient Boosting Regressor
RFR is Random Forest Regressor
BR is Bagging Regressor

Conclusion:

The predicted values obtained from Random Forest Regressor, Gradient Boosting Regressor, and Bagging Regressor, are close to actual values with MSE less than $4,000 and it can be seen from the last table that the values are not much different than the actual values.
Also, as previously discussed **smoker** is the most important feature that affects insurance charges of an individual, followed by his **age and BMI**. Number of **children, sex and region** don't play much role in predicting charges.

Recommendations:

With the help of this study, we can seek for ways and factors that control unsustainable increases in healthcare costs. It is imperative that healthcare organizations can predict the likely future costs of individuals, so that care management resources can be efficiently targeted to those individuals at highest risk of incurring significant costs as well as with general business planning in addition to prioritizing the allocation of scarce care management resources. Moreover, for patients, knowing in advance their likely expenditures for the next year could potentially allow them to choose insurance plans with appropriate deductibles and premiums.