Final Project Report
Predicting Personal Healthcare Cost

Harsh R. Jivani

04/27/2020

Business Understanding

My data analytics project topic is *Predicting Personal Healthcare Costs influenced by various health factors*. We are a startup healthcare solutions provider where we audit healthcare insurance companies across the United States to make sure they are not over charging their customers so that customers will not leave the insurance company.

My motivation in selecting this project is to tell a story on how and why a person should be in healthy shape to avoid increase of medical costs. Within past decade, I have seen many companies investing in smart health technologies to help patients stay healthy and avoid high medical premiums.

The target variable for this dataset is *charges*. The target variable, charges, is a continuous number and the data type is numerical (float). The variable *charges* represent money in this dataset and all data for this variable is calculated in USD or United States Dollar amount.

Business Problem

The business problem is that my company is in charge to audit a healthcare insurance companies who thinks they are losing customers at rapid rate and the business would like to find out if they are leaving because they overcharged them for their medical needs. My company will accurately predict medical costs for current and future members with similar characteristics and find out if the healthcare insurance company is under/over-charging members based on healthy/unhealthy lifestyle. I believe factors such as BMI, age, and someone with smoking history will affect whether they will pay higher medical costs than someone who meets the standard healthy lifestyle will pay lower medical costs.

Dataset

The dataset I will be using has 7 attributes including 1 target attribute. The dataset comes from Kaggle.com and the data is derived from demographic statistics from US Census Bureau. Below, I have provided with the first few rows of the sample dataset for this project. The original dataset contains 1,338 rows; therefore, I have not linked it here.

First 4 instances:

age	sex	bmi	children	smoker	region	charges
19	female	27.9	0	yes	southwest	16884.924
18	male	33.77	1	no	southeast	1725.5523
28	male	33	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.4706
32	male	28.88	0	no	northwest	3866.8552

Source to dataset: https://www.kaggle.com/mirichoi0218/insurance

Proposed Analytics Solution

The following is a potential proposed analytics solution:

- Medical Costs prediction
    - Predict and model medical costs based on age, bmi, smoking.
    - Use the model created to test the results with my family members to see the past and future charges make sense.
    - By using the models, we can use the predicted medical costs and find out how much money the patient would have saved.

Data Exploration and Preprocessing

Here, we are exploring the attributes of our dataset. The data types and data scale for this dataset is explained below.

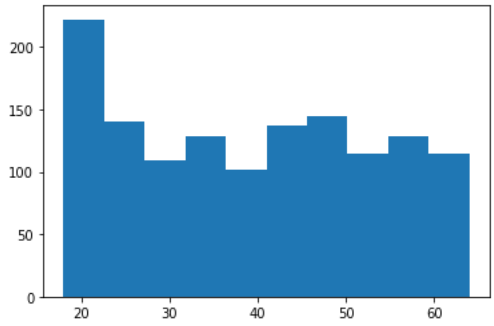
Dataset shows number of rows (1338) for all 7 columns, there are no null values or missing values for this dataset, and data Type for each of the data columns (int, object (string), float)

Attribute	Data Type	Data Scale
age	numerical int	ratio
sex	categorical	nominal
bmi	numerical float	ratio
children	numerical int	ratio
smoker	categorical	nominal
region	categorical	nominal
charges	numerical float	ratio

The dataset attributes are described as:

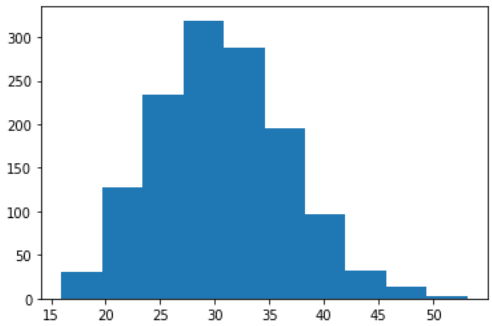
- Age (Continuous): The age of the person.
- Sex (Categorical): The gender (male or female) of the person.
- BMI (Continuous): Body Mass Index, is a measure of body size. It uses person's weight and height to calculate the BMI. The result we get tells us whether the person is underweight if result is < 18.5, healthy weight if result is 18.5 – 24.9, overweight if result is 25.0 – 29.9, or obesity if result is 30 or higher.
- Children (Continuous): Number of dependents that this person has.
- Smoker (Categorical): If person smokes then yes or does not smoke then no.
- Region (Categorical): This is residential area where the person resides. For example: Northwest/Southeast/Southwest, etc.
- Charges (Continuous): This is the medical costs that is billed by the insurance providers.

Histogram for Age:



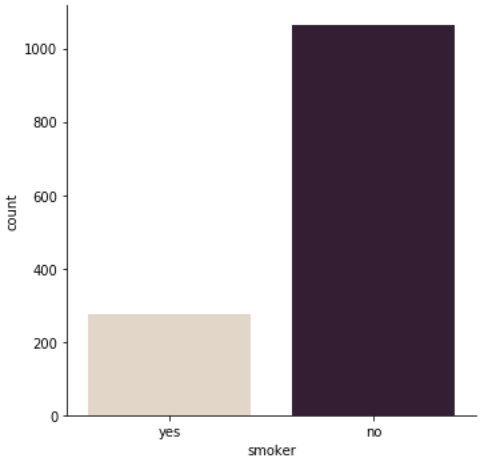
In the histogram above, we see a big spike in the age between 18-25 and then it seems to level off and become balanced for later ages.

Histogram for BMI:



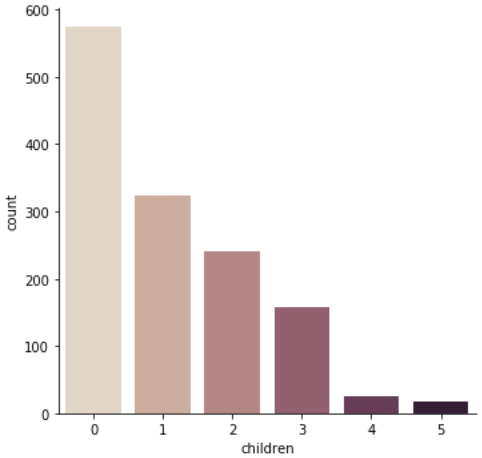
In the above histogram, we see a bell-shaped curve which means that most people have BMI between 25-35.

Bar plot for smoker:



We clearly see that we have majority of people who do not smoke. We have about 79.5% or 1064 of people who do not smoke and 20.5% or 274 that smokes.

Bar plot children:



Majority of people do not have any children. Trend is declining more people with more children.

Data Quality Report

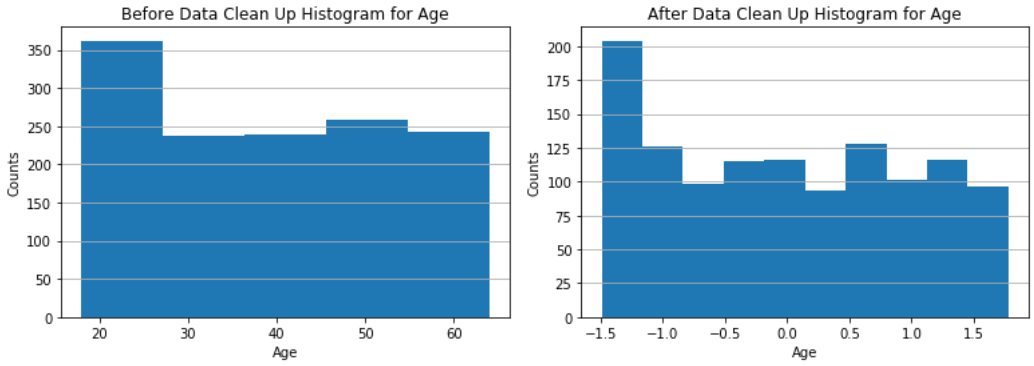
Continuous Features Table –

	Features	Count	%Miss	Card.	Min.	1stQartile	Mean	Median	3rdQartile	Max.	StdDev
0	age	1338	0.0	47	18.0000	27.00000	39.207025	39.000	51.000000	64.00000	14.049960
1	bmi	1338	0.0	548	15.9600	26.29625	30.663397	30.400	34.693750	53.13000	6.098187
2	children	1338	0.0	6	0.0000	0.00000	1.094918	1.000	2.000000	5.00000	1.205493
3	charges	1338	0.0	1337	1121.8739	4740.28715	13270.422265	9382.033	16639.912515	63770.42801	12110.011237

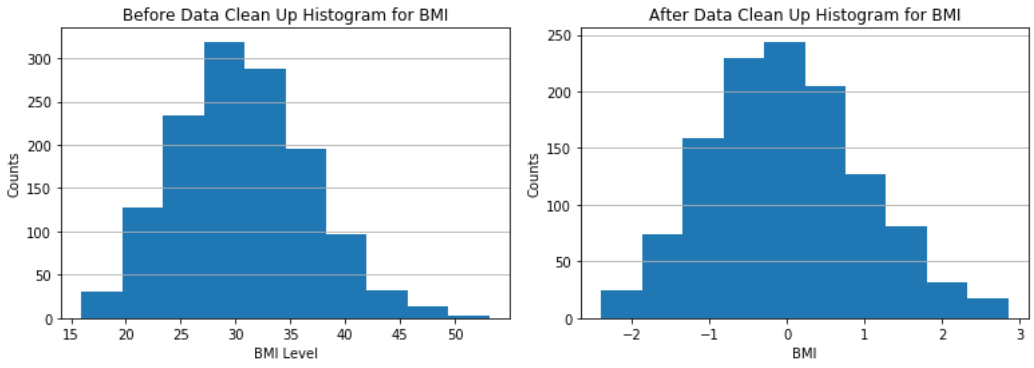
Description: We select age, bmi, children, and charges as our continuous features. All the features have 1338 instances and we do not have any missing values. We have 47 unique ages, 548 unique bmi, 6 unique children categories, and all the charges vary. The minimum values for age are 18, bmi is 15.96, children are 0, and charges is 1121.87. The maximum values for age are 64, bmi is 53.13, children are 5, and charges is 63,770.42. The mean for age is 39, bmi is 30.66, children is 1, and charges is 13,270. Statistics for 1st Quartile, median, 3rd Quartile, and standard deviation can be referenced from table above.

Continuous Features Histogram (before/after data clean up) –

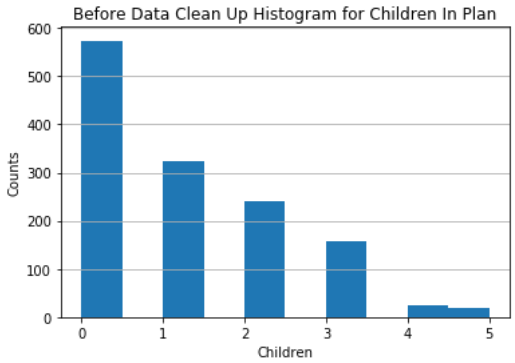
Histogram for age:

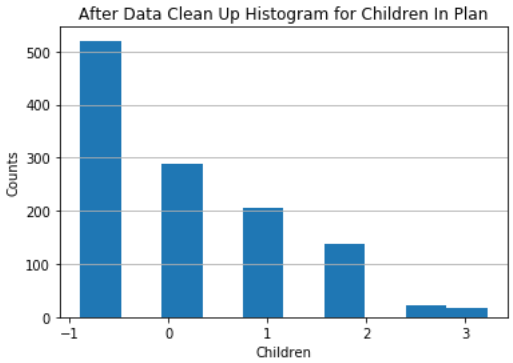


Histogram for BMI:

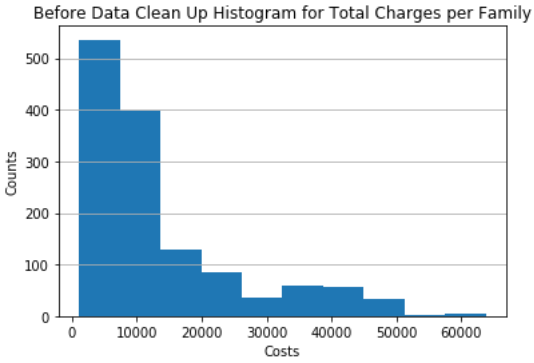


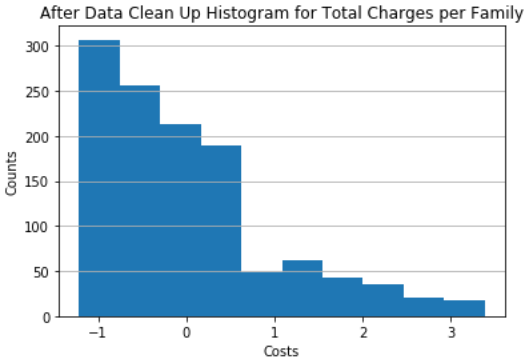
Histogram for children:



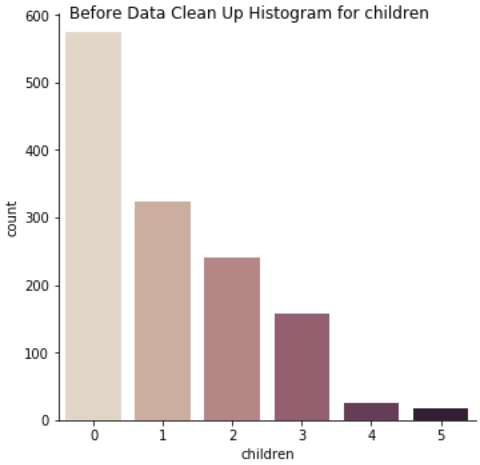


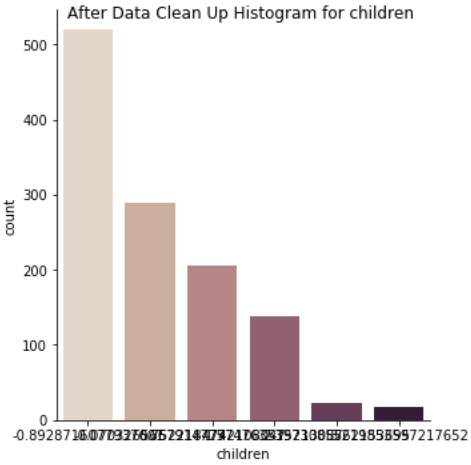
Histogram for charges:





Bar plot for children since cardinality is less than 10:





Categorical Features Table –

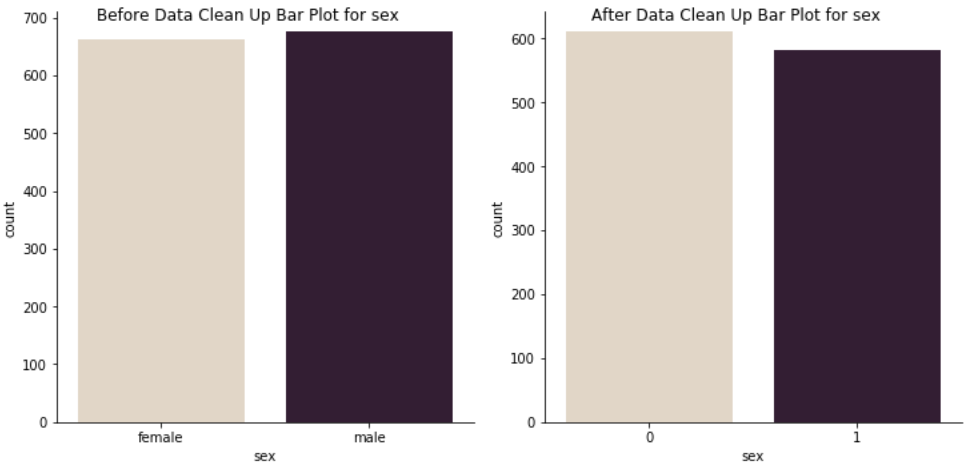
	Features	Count	% Missing	Cardinality	Mode	Mode Freq.	Mode %	2nd Mode	2nd Mode Freq.	2nd Mode %
0	sex	1338	0.0	2	male	676	50.523169	female	662	49.4768
1	smoker	1338	0.0	2	no	1064	79.521674	yes	274	20.4783
2	region	1338	0.0	4	southeast	364	27.204783	---	---	---

Description: We select sex, smoker, and region as our categorical features. All the features have 1338

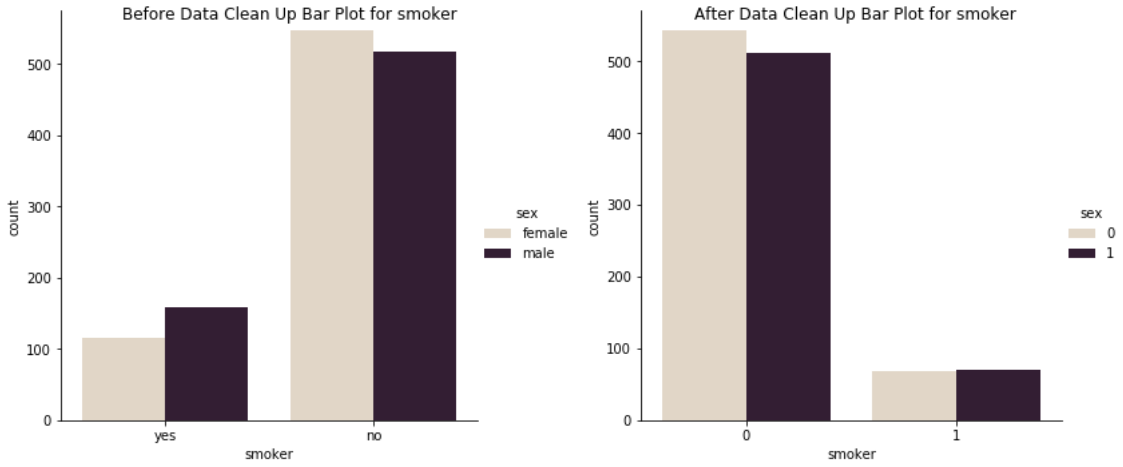
instances and we do not have any missing values. We have 2 unique sex (male/female), 2 unique smoker categories (yes/no), and 4 unique regions categories. The most frequent values for sex is male, smoker is no, and region is southeast. Mode frequency is how often the mode occurs, so for male is 676, mode frequency for non-smoker is 1064, and mode frequency for southeast is 364. We can also describe the frequency in the percentage as mode percentage, so for male is 50.52%, mode percentage for non-smoker is 79.52%, and mode percentage for southeast is 27.20%. Second mode, second mode frequency, and second mode percentage can be found on table above. However, we do not have value for region in the second mode, mode frequency, and mode percentage because all three regions have same values and hence, we cannot find the second mode.

Categorical Features Bar Plot (before/after data clean up) –

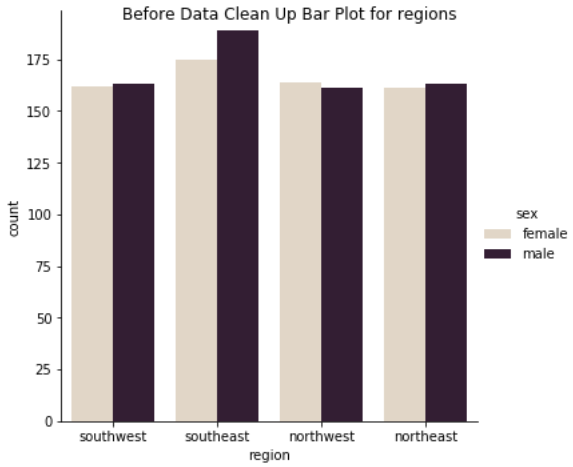
Bar plot for sex:

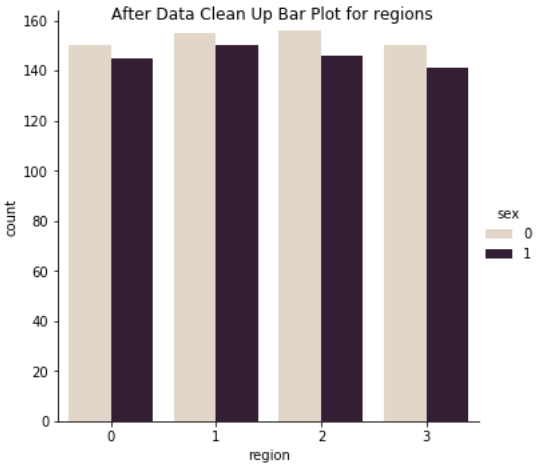


Bar plot for smoker:



Bar plot for region:





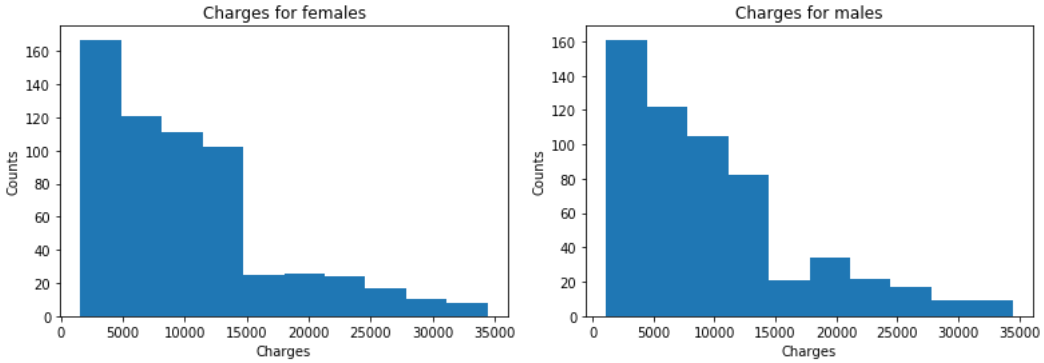
Data Quality Plan –

	Features	Data Quality Issues	Potential Handling Strategies
0	Children	# of childrens vary	Ignore this feature
1	Sex	Does not change the target variable	Ignore this feature
2	Region	Outlier - Southeast	Ignore this feature

Description: Based on my research, investigation, and business problem; I found children, sex, and region features to include in the data quality plan. The issue with children feature was that number of children vary. The issue with sex feature was that including sex feature in the machine learning model was not going to make any difference to predict the target variable because this feature is balanced. Also, the issue with region was that we have southeast has our outlier. So, in order to properly run the machine learning models, I decided to exclude the following three features from my models.

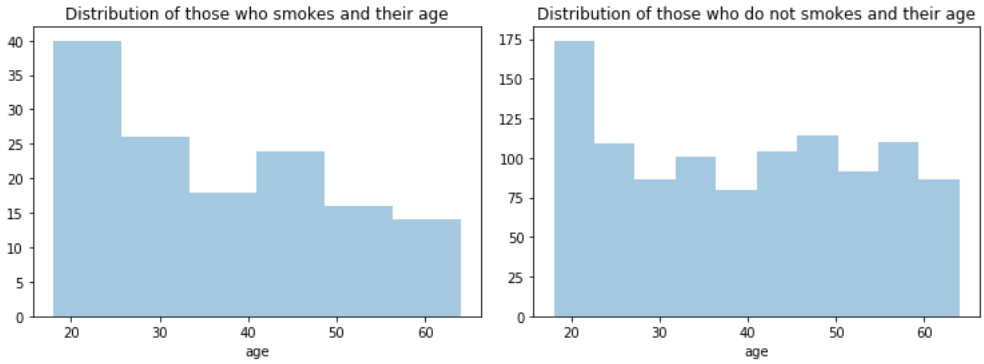
Visualizing relationships between various health factors –

Charges for males and females:



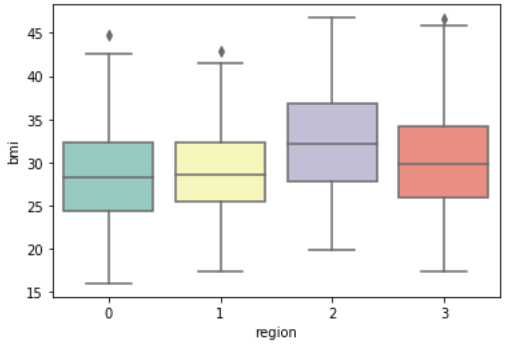
From my analysis, I can see that there is not major difference for male and female charges. Not much information can be extracted from this chart so we will continue to analyze other health factors.

Age vs. Smoking History:



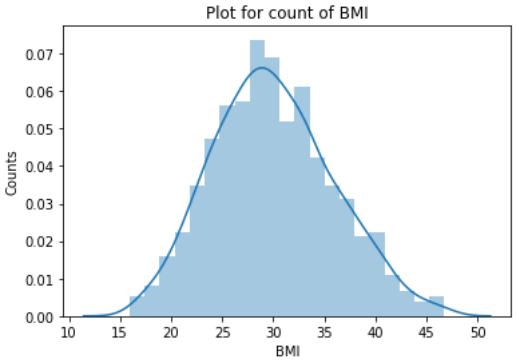
In my analysis, I think this chart is very interesting, because I can see number of people who are between 18-30 and in their 40s do not smoke. However, people in their 20s smoke the most.

BMI per Regions:



In this analysis, I find out that region does not directly impact BMI. However, there is little spike for region 2 (Southeast) and this can be because of many reasons which we do not have data for to analyze further. The average BMI for Northeast, Northwest, and Southwest are similar to each other. However, Southeast has the highest average of BMI.

Counts of BMI:



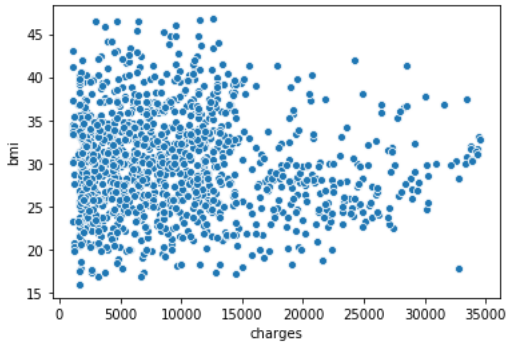
This is important plot to analyze, because it shows perfectly distributed BMI counts and it tells us a lot.

BMI: Body Mass Index, is a measure of body size. It uses person's weight and height to calculate the

BMI. The result we get tells us whether the person is underweight if result is < 18.5, healthy weight if result is 18.5 – 24.9, overweight if result is 25.0 – 29.9, or obesity if result is 30 or higher.

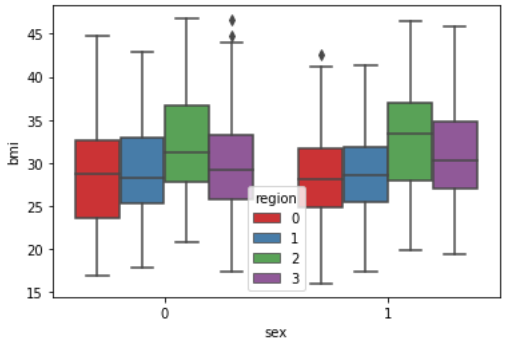
From this plot, we can determine that we have 25 people who are underweight; we have 233 people who are healthy; we have 473 people who are overweight; and we have 462 people who are obesity.

Charges vs BMI:



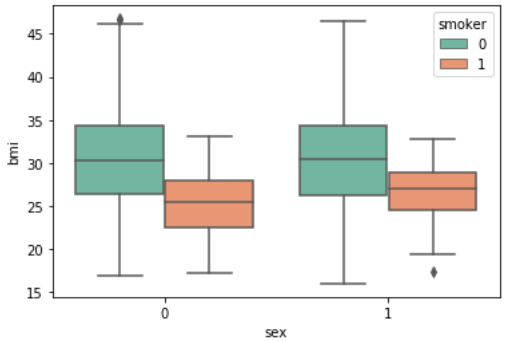
From this scatter plot analysis, I can see there are people who have charges higher than our mean of 9942.26(USD). We will predict and compare if they over overcharged even if they had similar characteristics as people within mean.

Gender and their BMI across regions:



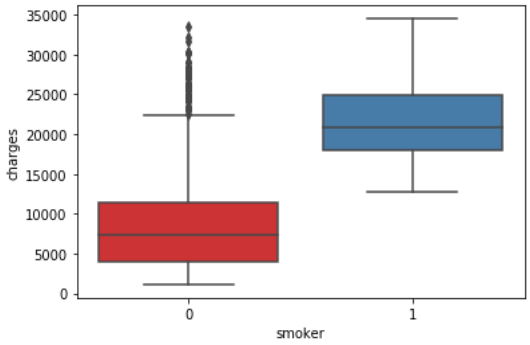
From this analysis, we cannot determine more information other than both gender for Southeast region have higher BMI then other regions. However, because we do not have more information about this region and what causes this spike, we cannot truly determine the factors. Across all the regions, we have average BMI of 29.99 or 30 for simplicity. According to the statistics, average BMI for female and male in region 2 are above the average BMI across all regions.

Gender and their BMI with smoking history:



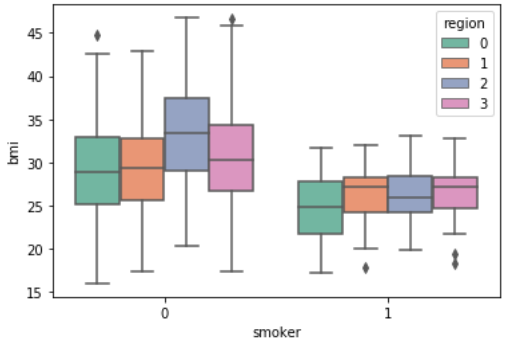
In this analysis, we can see gender really does not impact the BMI. This chart really blew my mind because female who smokes have lower average BMI of 25.15 (overweight) than female who does not some have average BMI of 30.5 (obesity). Male who smokes have lower average BMI of 26.51 (overweight) than male who does not some have average BMI of 30.6 (obesity).

Smoking history and charges:



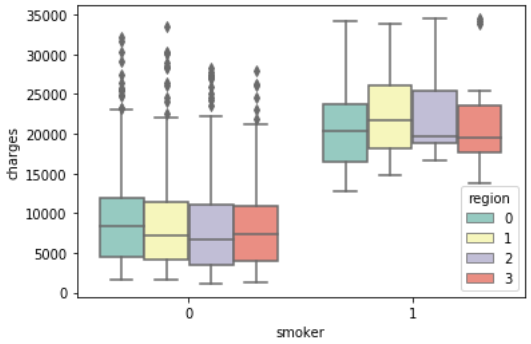
According to the box plot, we clearly see that those who smokes tend to have higher charges. However, those who do not smoke still charged higher than someone who does not smoke. This causes problem for medical payers and insurance providers are profiting because of this issue.

Smoking history and BMI across regions:



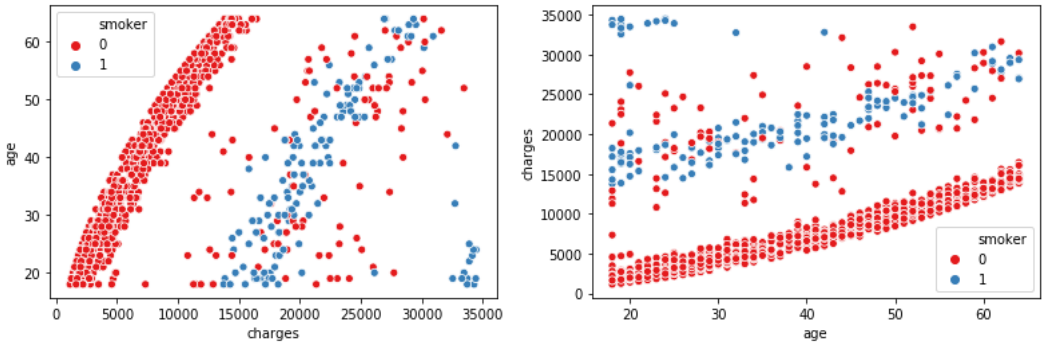
In this analysis, we can see those who smokes for across the regions seems to have variable BMI levels. This means that it's not constant, but it changes widely throughout. On the other hand, for those who do not smoke have similar BMI levels except for Southeast region.

Smoking history and charges across regions:



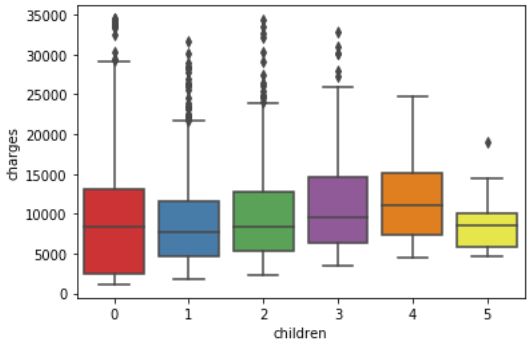
In this analysis, for non-smokers there are people who are beyond the whiskers who are getting charged by the insurance company heavily. On other side, we have smokers and looks like they are charged more and have similar cost for all regions.

Charges and age for those who smokes and those who do not smoke:



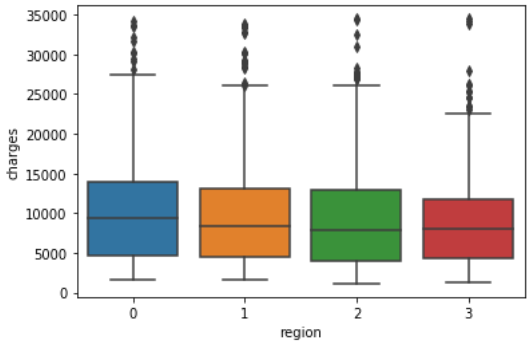
From my analysis, I can see that those who smokes definitely are paying higher than majority of the non-smokers. Also, there is very nice trend for majority of non-smokers who are paying less than 15000(USD) in medical charges. We can also notice non-smokers are paying too much and we will find out what they should be paying based on similar non-smokers from our Machine Learning Models. Yet, there are few smokers who are paying above 30000(USD) and this group is apart from smokers' trend.

Children vs Charges:



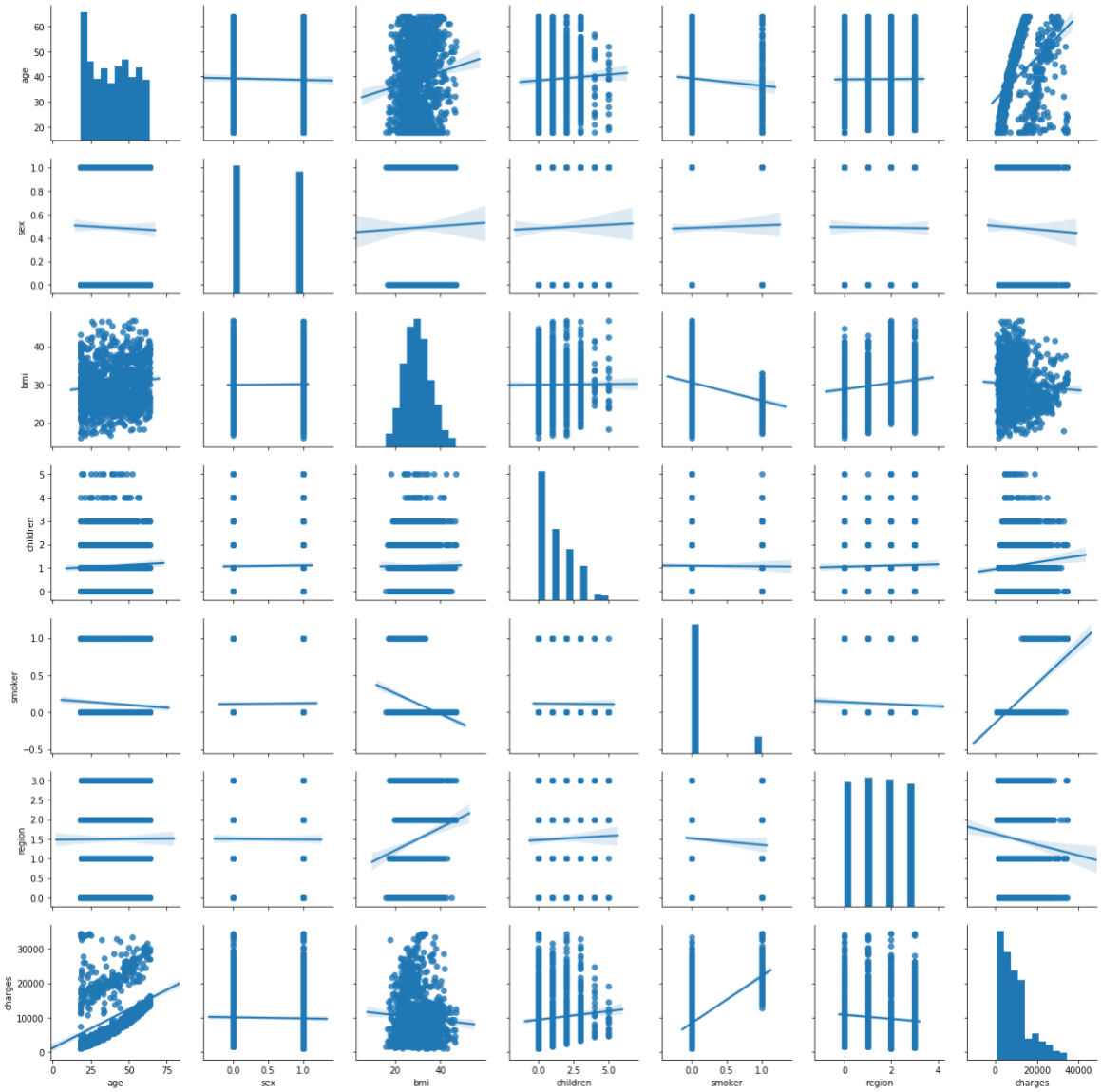
From my analysis, I can see that more the children, the higher the medical charges. However, the majority of the members does not have children and they still pay higher. On the other hand, I see the trend is increasing as children increases.

Region vs Charges:



From my analysis, I can see that regions do not impact the charges. We can see the median charges paid is relatively the same for all regions. However, we can notice just slight difference for region 0 (Northeast) which has quartile 3 higher and region 3 (Southwest) which has quartile 3 lower of all other regions.

Let's visualize using pair plot to get bigger picture of trends



From the above pair plot, we can analyze that we have many features that shows trends.

1) We will focus on charges (target variable):

a) We can make hypothesis that as age increases, the charges also increase.

b) We can make hypothesis that as bmi increases, the charges are decreasing. We see that bmi of around 30 shows charges are higher than those who have lower. We will have to test in our model.

c) We can make hypothesis that those who have more children, the higher the medical cost they are paying.

d) We can make hypothesis that those who smokes, the higher the charges.

e) From my hypothesis, the features we will be using are: Age, BMI, Children, and Smoker. Our target variable to predict is charges based on these features.

Missing Values and Outliers

*****NOTE: Identifying the missing values (Does NOT apply to this dataset):**

However, I verified there were no missing values by using the apply and lambda function and iterating all the instances to find any null values. I found none null or missing values in this dataset.

Identifying the outliers (by IQR Score):

Goal: We will identify and handle the outliers by using the IQR Score method. We identify the outliers for each attribute.

First, we will copy all the attributes to the temporary DataFrame. Next, we will get the 25% quantile from temporary DataFrame and set it equal to Q1. Then, we will do same process for 75% quantile. After this, we will subtract Q1 – Q3 to get the IQR (interquartile range). Next, we compare Q1 – 1.5 * IQR with each attribute and we do same for Q3 + 1.5 * IQR with each attribute. If there exist True value, then this means that there is outlier in that instance. After detecting the outlier, we will remove the outliers and keep only the valid data. To do this, we take the negation of each attribute and compare to see if its less than (Q1 - 1.5 * IQR) or compare each attribute greater than if (Q3 + 1.5 * IQR). After running the code, we get the following final results. 1193 is the number of remaining clean data and the 145 is the data that is outliers and we deleted it.

Normalization

*****NOTE: Normalizing the Dataset (Does NOT apply to this dataset):**

However, I tested analysis with normalized and un-normalized, and the features did not overtake any other features since we are within the similar scale. I still went ahead and normalized my dataset for this dataset just in case I need it later in project or practice normalizing in future.

I normalized the dataset using the Z-Score Normalization. First, I imported the python modules for normalizing. Then, I used preprocessing.StandardScaler() for calculating the standard score of the dataset. Next, I used zs_scaler.fit_transform(df_copy2) to insert my dataset and have the function compute the mean and standard deviation needed in order to get z-score. Finally, I created a separate

DataFrame for my new normalized dataset. First few rows can be seen below:

	age	sex	bmi	children	smoker	region	charges
0	-1.421407	-0.975980	-0.357841	-0.892872	2.764946	1.356873	0.957834
1	-1.492543	1.024611	0.643111	-0.070327	-0.361671	0.456066	-1.133608
2	-0.781184	1.024611	0.511811	1.574764	-0.361671	0.456066	-0.757808
3	-0.425504	1.024611	-1.243693	-0.892872	-0.361671	-0.444740	1.661386
4	-0.496640	1.024611	-0.190732	-0.892872	-0.361671	-0.444740	-0.838186

Feature Selection and Transformations

Impurity-Based Univariate Feature Selection:

For finding the most informative features, I used Information Gain with the Entropy. Since this code was part of my homework 3, I used my own code here and the function output the most informative features. In this case it was bmi, age, children, and region. How it works is that I have IUFS function and it finds k most informative features in the given dataset based on the target variable using information gain with the selected measure. I give the target which is the name of the target variable, dataset which is the DataFrame for the dataset, k which is number of features to return and this must be less than or equal to number of descriptive features in dataset, and measure which can be 'entropy' or 'gini'. In my scenario, my target variable is charges, DataFrame is df_scaled, k = 4, and measure = entropy. How did I get the informative features? Well, it will compute the entropy and information gain for all the features and returns a list in order of k feature names, selected based on univariate selection schema.

Code and more information about the feature selection can be found on the Jupyter notebook.

Applying Transformations:

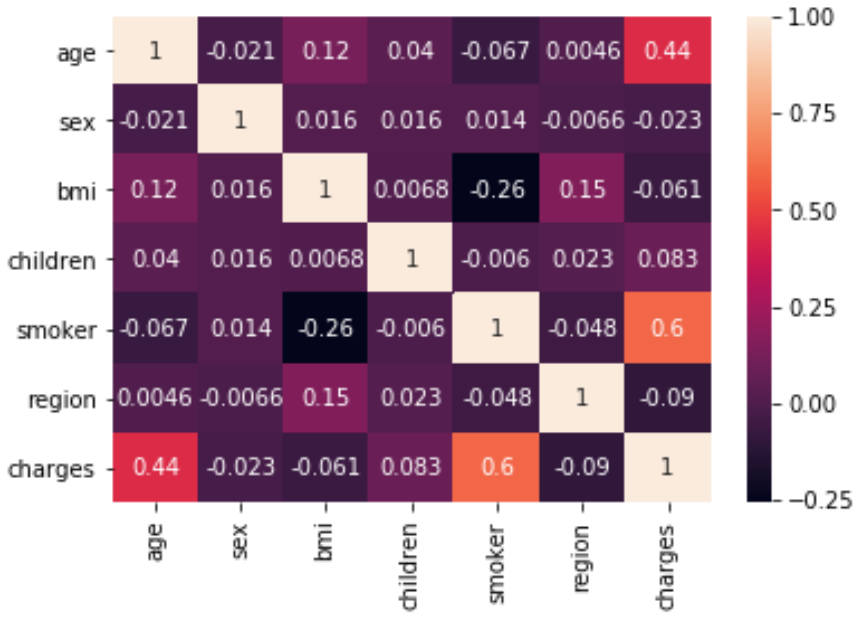
1) I converted categorical feature to numerical feature for normalizing. The steps I took to solve this are as follows: First, I imported the modules to complete this job. Then, I created a variable at set it equal to LabelEncoder(). This object will help me convert the variables by using the built-in function fit_transform. I will perform the conversion for sex, smoker, and region features. After the execution, sex feature will have the following value: female = 0 and male = 1. Smoker feature will have the following value: yes = 1 and no = 0. Region feature will have the following value: Region 0 = Northeast, 1 = Northwest, 2 = Southeast, and 3 = Southwest.

2) I created the equal-width binning for age feature. This will help me analyze the age in the later report. What I did was use numpy and built-in function histogram and set bins = 5. After this, I got the count and divisions. The divisions represent the border for the bins and the count represent the number of people inside the bin. Below, you can see the result it output:

[18. 27.2 36.4 45.6 54.8 64.]

[330 213 209 229 212]

3) Next, I created Correlation Matrix. This Correlation Matrix is helpful in many ways. One way is use it for selecting the most important features. Or we can use the matrix to analyze which features is positive correlated or negative correlated. The steps to solve this was to use the DataFrame from normalized and use the built-in function corr(). Then, we use seaborn and show the heatmap of the correlated matrix. Below is the result we got:



We see high correlation in age and smoker vs charges. We still need to further analyze other features.

Model Selection and Evaluation

In this Fundamentals of Data Science class, we are working with Classification and Regression based models. In my project, I am doing a Regression based machine learning model because we are predicting the continuous target variable, charges. For my machine learning models, I selected these three models (Linear Regression, Decision Tree Regression, and K-Neighbors Regression) and compared each other based on the performance metrics and which can better predict the charges.

Evaluation Metrics

The performance metrics we will be using to compare three machine learning models are:

1) Mean Absolute Error (MAE): This will tell us how bad the predictions were than original data. It will let us know the error only but does not tell us if its over-fitting or under-fitting. If the error is 0 or closer to 0, then it means that there are no errors and the predictions are perfect.

2) Mean Squared Error (MSE): This performance metric is similar to MAE, but it provides the overall idea of error.

3) Root Mean Squared Error (RMSE): This is simply the square root of MSE. This is used instead of using MSE since MSE is long number and for better presentation the RMSE is used.

4) R-Squared: This is used to indicate if the prediction fit the model well. It gives output in the range of 0 to 1, the number closer to 0 indicates model does not fit well and closer to 1 indicates well fit model.

I used the following performance metrics for my business problems because I am predicting the charges for the members if they were over-charged or under-charged. MAE will tell me how bad the overall predicted charges were compared to the original data. The goal is to train the model to get the R-Squared near 1 so that there will be less errors when predicting charges. This is important metrics because we want to find the accurate answer and once we know if the model is predicting the results correctly than we can use it for all the members and we will know the answer. Once we know the details of the members who were over-charged, we can contact the customers and find out more details.

Models

Multiple Linear Regression – Machine Learning Model

Linear regression model is one my favorite since it's easy to implement, and generally it is widely used for predictive analysis. The goal of the linear regression model is takes in one or multiple variables/input and maps a best fit line. This best fir line is in the form of $Y = mx + b$, where Y = predicted score, m = slope of the line, and b is the intercept of Y. In this case, we are trying to predict the charges from given inputs such as age, bmi, and smoker. While the output will be the predicted charges. Once, we get the equation set-up, we can then predict charges based on inputs given and this is how we as an audit company will track people who are over-charged and alert the insurance companies that you have an red flag that the customers might not be happy and leave your membership.

Decision Tree Regression – Machine Learning Model

Decision Tree regression is also easy to implement, but in our case, it was not really helpful. It did not provide me the accurate results that I wanted to see. This model is actually used better in classification problem, but most companies use this model for regression as well. The decision tree regression has same predictors as linear regression model such as: age, bmi, and smoker; the target variable is charges. The tree has decision node and leaf nodes. Because my dataset was big, the tree itself becomes large and we will get many leaf nodes.

K-Neighbors Regression – Machine Learning Model

This machine learning model is easy to implement and learn. The accuracy is somewhat off from other models I tested but it was better than decision tree regression. The algorithm basically uses the neighbors that we input such as n=37 or n=17 and it will use those that are near to each other neighbors. In this project, I decided to test my model based on neighbors that is 37 which grabs 37 closest of all. On the second test, I decided to change to neighbor that is 17 which grabs 17 closest of all. The accuracy really depends on what you set your neighbors to be. From what I learned from this project, the higher you pick the better it will be, but accuracy score will drop. How I got the n=37 was by taking the square root of the dataset, and how I got the n=17 was random number I picked below the 37 and this had the best performance in regard to the accuracy.

Sampling and Evaluation Settings

Then, provide information on how you create your training and testing dataset and evaluate these models; show, using charts and confusion matrices, the performance comparison of your models.

Multiple Linear Regression – Machine Learning Model

First, I read the dataset from the preprocessed csv file that I saved from phase 2. Next, I created an object that would handle the linear regression model. After this, I dropped the unnecessary columns that I do not need to train and test my model. The columns I dropped were sex, children, and region. This action was based on the phase 2 after analyzing the correlation and feature selection. Next, I imported the sklearn modules called model_selection and train_test_split where I can create the variables X_train, X_test, y_train, and y_test. I applied the X variable to be age, bmi, and smoker columns, while y will be charges. I set my test size to be 0.20 so that its 80-20 ratio when testing and training models. Now, I am ready to fit the linear regression to my training set. I used the function called fit and applied X_train and y_train variables. The model was ready, and I found out the slope and the intercept. I then created the simple equation and I used predict function to test my X_test.

```
#Calculate the Residuals/Errors
from sklearn import metrics
from sklearn import model_selection

print('MAE:', metrics.mean_absolute_error(y_test,y_pred))
print('MSE:', metrics.mean_squared_error(y_test,y_pred))
print('RMSE:', np.sqrt(metrics.mean_absolute_error(y_test,y_pred)))
print('Test R-Squared/Score:', metrics.r2_score(y_test,y_pred))
lr_score2 = lr.score(X_train,y_train)
# R-squared of the train
print('Train R-Squared/Score:', lr_score2)
```

```
MAE: 2335.9587408899347
MSE: 15065547.101000844
RMSE: 48.331757063962975
Test R-Squared/Score: 0.6424574261014897
Train R-Squared/Score: 0.5830705751549957
```

Since, I am working on regression models, I cannot measure the confusion matrix. Instead of confusion matrix, I feel more comfortable that my business problem can use performance metrics shown above such as MAE, MSE, and etc.

Our testing accuracy is about 64.2%. It is kind of low, but given the information we had in our dataset, we could do minimal to get better results. I tried various trial and error on adding and removing columns to the machine learning models, but it kept giving me low possible scores. The highest scores were seen on age, bmi, and smoker. If we compare the features, I used with the correlation heatmap, we can see that the red-hot zones are in these features. This suggested that there is major correlation in these features, and we need to train and test on these features.

The MAE (mean absolute error) is derived from 1 divided by the total number of data points then multiplying with the sum of absolute value of actual minus predicted results. In this case, the MAE represents from a single data point to the linear regression line (prediction). MAE here is little bit higher and this means that the model is not trained perfectly. It's not 100% predicting the charges based on the given features.

However, for MSE (mean squared error) the error mostly pertains to the outliers of our model. The further the data points in the model, the higher our MSE will be. In our case, we have about 16 points that can be considered outliers.

RMSE is the square root of the mean squared error (MSE). We can interpret RMSE and MAE based on how they are spread apart. In our case, RMSE indicates that the model had charges missed by 48%.

Here, the insurance companies will use the predict function to predict charges for the patients using age, bmi, children, smoker, and region.

```
print(df.loc[[0]])
#age,bmi,smoker
lr.predict([[19,27.9,1]])
```

```
   age  sex   bmi  children  smoker  region    charges
0   19    0  27.9         0       1       3  16884.924

array([18113.83871377])
```

If you look at the above scenario, we have a person who is 19 years old, has bmi of 27.9, and they smoke. According to the model, the model predicts they should be charged estimated of 18,113(USD) because the person is young, have moderate bmi, and has smoking history. Now, if we say they did not

smoke, then the difference is big, estimated about 3,383(USD). Here, we found out that smoking is a major factor insurance companies consider. Even if we had increased the person's BMI to obesity, it still would not make big difference as it did with smoking.

I converted test, prediction, and the difference of the test and prediction to new DataFrame to create plots and understand the results better.

```
myReal[myReal['Difference']>= 5000].head()
```

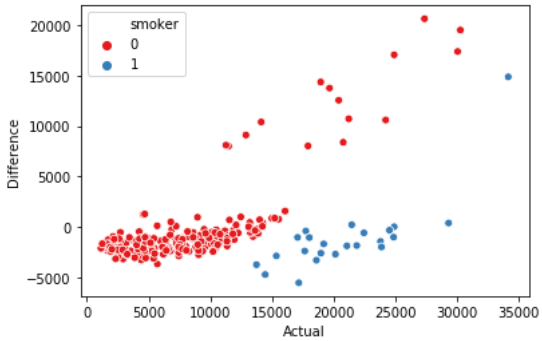
	age	bmi	smoker	Actual	Predicted	Difference
338	50	25.365	0	30284.64294	10750.173433	19534.469507
985	37	29.800	0	20420.60465	7851.254614	12569.350036
1122	55	37.715	0	30063.58055	12657.947952	17405.632598
898	48	29.600	0	21232.18226	10504.573024	10727.609236
998	23	31.400	1	34166.27300	19280.198932	14886.074068

```
myReal[myReal['Difference']>= 5000].describe()
```

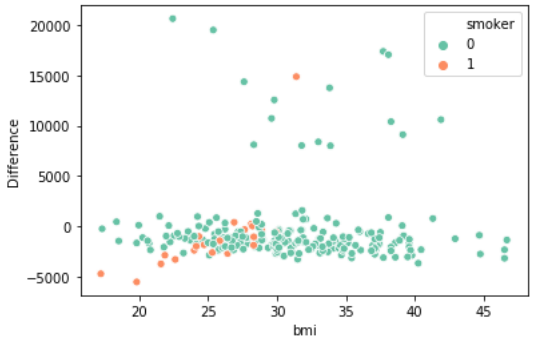
	age	bmi	smoker	Actual	Predicted	Difference
count	16.000000	16.000000	16.0000	16.000000	16.000000	16.000000
mean	35.250000	32.632813	0.0625	21237.686344	8507.749234	12729.937111
std	14.825653	5.413906	0.2500	6933.672312	4552.835966	4244.385310
min	18.000000	22.420000	0.0000	11272.331390	3164.722568	8003.740363
25%	21.750000	29.277500	0.0000	16980.236965	4377.465246	8933.229236
50%	34.500000	32.395000	0.0000	20601.046785	7842.955127	11648.479636
75%	48.500000	37.810000	0.0000	25530.260890	11160.630501	15434.653206
max	58.000000	41.910000	1.0000	34166.273000	19280.198932	20667.622749

For those 16 people who were overcharged (>= 5000(USD)) had mean difference of around 12,730(USD), have mean age of 35, and have high BMI or obesity of 32.6. These people do not smoke, so it means that other health contribution such as they are not doing regular exercise, or they just have unhealth lifestyles.

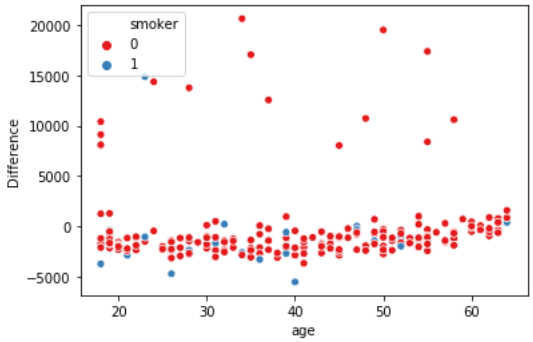
What I did to best visualize and select which models performed the best, I created scatter plot consisting of actual vs difference.



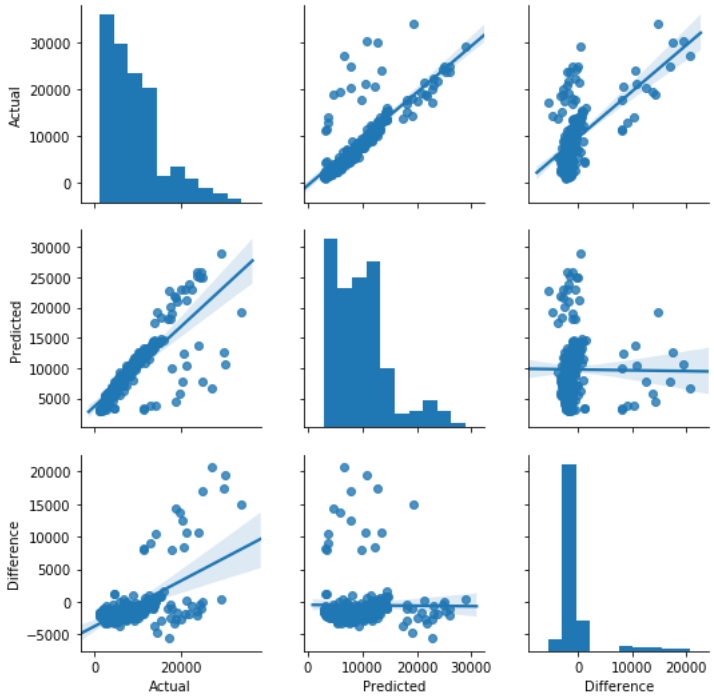
From this plot, we can discuss that those who do not smoke saw few differences in terms of charge than those who did smoke. However, based on the model, the prediction charges for this people are less. From the below chart, we can see that most of them have high BMI level and the age seems to be higher.



This chart shows the relationship between bmi and difference of actual and predicted charges. We see higher differences for the people who have above normal BMI level.



From the above chart, we compare the age and the difference of charges. We can see that most people were charged correctly, and most were undercharged.



From the pair plot, we see the linear relationship between actual and predicted results. Those points that are within or near the line have little to no difference. While the points that are left of the line overcharged while the right of the lines are undercharged.

Decision Tree Regression – Machine Learning Model

First, I read the dataset from the preprocessed csv file that I saved from phase 2. Next, I created an object that would handle the decision tree regression model. After this, I dropped the unnecessary columns that I do not need to train and test my model. The columns I dropped were sex, children, and region. This action was based on the phase 2 after analyzing the correlation and feature selection. Next, I imported the sklearn modules called model_selection and train_test_split where I can create the variables X_train_rg, X_test_rg, y_train_rg, and y_test_rg. I applied the X variable to be age, bmi, and smoker columns, while y will be charges. I set my test size to be 0.20 so that its 80-20 ratio when testing and training models. Now, I am ready to fit the decision tree regression to my training set. I used the function called fit and applied X_train_rg and y_train_rg variables. I also kept all the default inputs such as criterion, splitter, and etc. Once the model was ready, I then used the predict function to test my X_test_rg.

```
#Calculate the Residuals/Errors
from sklearn import metrics

print('MAE:', metrics.mean_absolute_error(y_test_rg,y_pred_rg))
print('MSE:', metrics.mean_squared_error(y_test_rg,y_pred_rg))
print('RMSE:', np.sqrt(metrics.mean_absolute_error(y_test_rg,y_pred_rg)))
print('Test R-Squared/Score:', metrics.r2_score(y_test_rg,y_pred_rg))
rg_score = rg.score(X_train_rg,y_train_rg)
# R-squared of the train
print('Train R-Squared/Score:', rg_score)
```

```
MAE: 3597.046828012553
MSE: 42986093.034901835
RMSE: 59.9753851843617
Test R-Squared/Score: -0.02016596161440476
Train R-Squared/Score: 0.9900214675394683
```

Since, I am working on regression models, I cannot measure the confusion matrix. Instead of confusion matrix, I feel more comfortable that my business problem can use performance metrics shown above such as MAE, MSE, and etc.

Our testing accuracy is about negative 2.01%. It is considered that the testing dataset did not work well with the traning dataset. If its closer to zero percentage, means that testing did not react to the traning set. However, the training dataset score did very well with the results of 99.0%. I tried various trial and error on adding and removing columns to the machine learning models, but it kept giving me lower high possible scores.

The MAE (mean absolute error) is derived from 1 divided by the total number of data points then multiplying with the sum of absolute value of actual minus predicted results. In this case, the MAE is 3597 and MAE on the linear regression was 2335. As an analyst, we can represent that Decission Tree is not good to use since the results were not correctly tested on tranining model. We have to look at other features of the testing dataset that were giving bad results.

However, for MSE (mean squared error) the error mostly pertains to the outliers of our model. The

further the data points in the model, the higher our MSE will be. In our case, we have about 30+ points that can be considered outliers.

RMSE is the square root of the mean squared error (MSE). We can interpret RMSE and MAE based on how they are spread apart. In our case, RMSE indicates that the model had charges missed by 59.9%.

Here, the insurance companies will use the predict function to predict charges for the patients using age, bmi, children, smoker, and region.

```
#print(df.loc[[0]])
#age,bmi,smoker
#rg.predict([[19,27.9,0]])
print(df.loc[[1]])
#age,bmi,smoker
rg.predict([[18,33.77,1]])
```

```
   age  sex    bmi  children  smoker  region    charges
1   18    1  33.77         1       0       2  1725.5523

array([34439.8559])
```

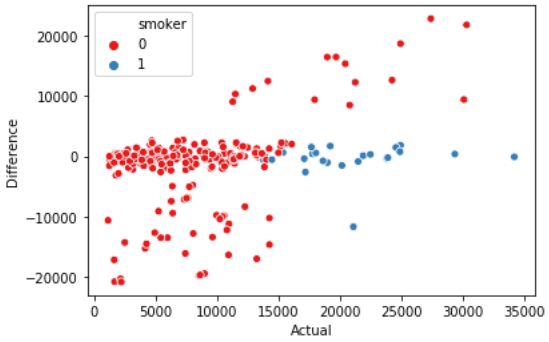
If you look at the above scenario, we have a person who is 18 years old, has bmi of 33.77, and they do not smoke. According to the model, the model predicts they should be charged estimated of 1,136.4(USD) because given the condition that their BMI is too high and chances of incurring future treatments are high. Now, if we say they did smoke, then the difference is big, estimated about 32,714(USD). Because if similar person who is 18, high BMI, and smokes are considered at risk and payment increases. Here, we found out that smoking and BMI is a major factors insurance companies consider.

I converted test, prediction, and the difference of the test and prediction to new DataFrame to create plots and understand the results better.

	Actual	Predicted	Difference
247	4237.12655	3554.2030	682.92355
302	5478.03680	5272.1758	205.86100
406	10381.47870	8125.7845	2255.69420
874	4889.03680	5272.1758	-383.13900
877	19199.94400	17496.3060	1703.63800

From the above table, I have shown the first few rows that my model predicted, and the difference was from the original dataset.

What I did to best visualize and select which models performed the best from the other 2 models, I created scatter plot consisting of actual vs difference.



From this plot, we can discuss that those who do smoke saw few differences in terms of charge than those who did not smoke. However, those who do not smoke, saw major changes to their prediction charges. People who do not smoke had fewer difference in their charges. We can see about 15 people were overcharged and about 30+ people were undercharged. From the below chart, we can see the differences in their lifestyle characteristics.

```
myReal2[myReal2['Difference']>= 5000].head()
```

	age	bmi	smoker	Actual	Predicted	Difference
338	50	25.365	0	30284.64294	8442.66700	21841.97594
985	37	29.800	0	20420.60465	5028.14660	15392.45805
1122	55	37.715	0	30063.58055	20630.28351	9433.29704
898	48	29.600	0	21232.18226	8930.93455	12301.24771
567	35	38.095	0	24915.04626	6196.44800	18718.59826

```
myReal2[myReal2['Difference']>= 5000].describe()
```

	age	bmi	smoker	Actual	Predicted	Difference
count	15.000000	15.000000	15.0	15.000000	15.000000	15.000000
mean	36.066667	32.715000	0.0	20375.780567	6554.864604	13820.915963
std	14.968857	5.593584	0.0	6226.932285	5349.082058	4633.621341
min	18.000000	22.420000	0.0	11272.331390	1137.011000	8512.856670
25%	21.000000	28.955000	0.0	16031.170560	2332.724825	9889.460445
50%	35.000000	33.000000	0.0	20420.604650	5028.146600	12501.369450
75%	49.000000	37.905000	0.0	24571.191750	8725.480275	16495.959550
max	58.000000	41.910000	0.0	30284.642940	20630.283510	22875.565530

For those 15 people who were overcharged (>= 5000(USD)) had mean difference of around 13,821(USD), have mean age of 36, and have high BMI or obesity of 32.7. These people do not smoke, so it means that other health contribution such as they are not doing regular exercise, or they just have unhealth lifestyles.

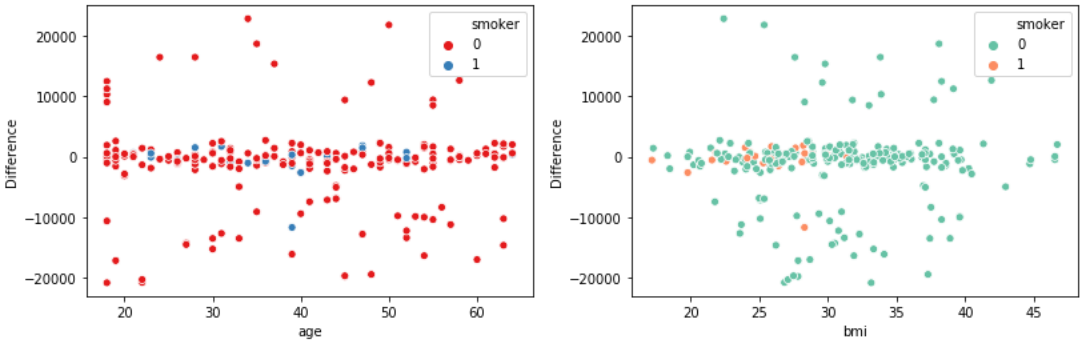
```
myReal2[myReal2['Difference']<= -5000].head()
```

	age	bmi	smoker	Actual	Predicted	Difference
952	48	37.290	0	8978.18510	28468.91901	-19490.73391
303	19	27.835	0	1635.73365	18838.70366	-17202.97001
249	52	31.200	0	9625.92000	23045.56616	-13419.64616
440	31	23.600	0	4931.64700	17626.23951	-12694.59251
581	53	39.600	0	10579.71100	20462.99766	-9883.28666

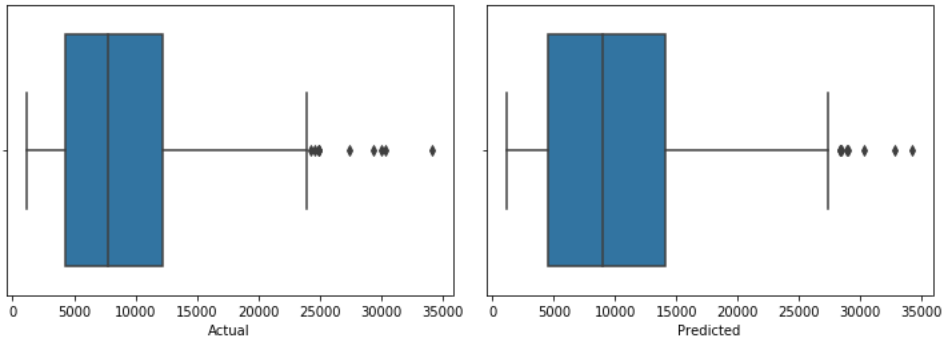
```
myReal2[myReal2['Difference']<= -5000].describe()
```

	age	bmi	smoker	Actual	Predicted	Difference
count	36.000000	36.000000	36.000000	36.000000	36.000000	36.000000
mean	41.694444	30.595556	0.027778	7898.500406	20932.629796	-13034.129390
std	13.166667	5.031864	0.166667	4258.037503	5296.632435	4424.114461
min	18.000000	21.780000	0.000000	1131.506600	11774.159275	-20875.257880
25%	30.750000	27.035000	0.000000	5163.485500	17418.782618	-16206.841850
50%	44.000000	30.220000	0.000000	7681.927500	20546.640585	-12755.306470
75%	52.250000	33.522500	0.000000	10482.841750	23202.970542	-9860.380440
max	63.000000	39.600000	1.000000	21082.160000	32787.458590	-5056.872620

For those 36 people who were undercharged (<= -5000(USD)) had mean difference of around -13,034(USD), have mean age of 41, and have high BMI or obesity of 30.6. Of those 36 people, most of these people do not smoke, only one person who smokes is undercharged is 39-year-old and has BMI of 28.3.



This chart shows the relationship between bmi and difference of actual and predicted charges. We see major differences for the people who have above normal BMI level, and minor difference for normal bmi level that are either overcharged or undercharged.




```
df4['Actual'].describe()
```

```
count      239.000000
mean      9254.477322
std       6504.876336
min       1131.506600
25%       4248.935275
50%       7749.156400
75%      12159.774750
max      34166.273000
Name: Actual, dtype: float64
```

```
df4['Predicted'].describe()
```

```
count      239.000000
mean     10446.242470
std       7309.281511
min       1136.399400
25%       4559.209575
50%       8930.934550
75%      14043.476700
max      34254.053350
Name: Predicted, dtype: float64
```

The difference between predicted mean of Decision Tree Regression and Linear Regression was (10,446 - 9,773 = 673). So far, the difference is not that major. We will compare all three models and pick the best model that has shown best statistics. From the above two boxplots, we see that there are outliers/points beyond the maximum predicted or actual charges. Those are serious points that should be investigated because as an auditor we want to be fair for both sides (consumers/insurance providers). We need more information for these specific people and need description of kind of treatments they did in the past and their health information in order to predict charges better.

K-Neighbors Regression – Machine Learning Model

First, I read the dataset from the preprocessed csv file that I saved from phase 2. Next, I created an object that would handle the K-Neighbors regression model. After this, I dropped the unnecessary columns that I do not need to train and test my model. The columns I dropped were sex, children, and region. This action was based on the phase 2 after analyzing the correlation and feature selection. Next, I imported the sklearn modules called model_selection and train_test_split where I can create the variables X_train_kn35, X_test_kn35, y_train_kn35, and y_test_kn35. I applied the X variable to be age, bmi, and smoker columns, while y will be charges. I set my test size to be 0.20 so that its 80-20 ratio when testing and training models. Now, I am ready to fit the KNeighbors regression to my training set. I used the function called fit and applied X_train_kn35 and y_train_kn35 variables. I used to predict function to test my X_test_kn35.

I repeated the same process, but this I will be checking to see if I change my neighbors to 17 will it

make any difference. Of course, it did make some difference in the scores. So, I went with the n=35 since they were slightly better than this one.

```
#Calculate the Residuals/Errors
from sklearn import metrics

print('MAE:', metrics.mean_absolute_error(y_test_kn35,y_pred_kn5))
print('MSE:', metrics.mean_squared_error(y_test_kn35,y_pred_kn5))
print('RMSE:', np.sqrt(metrics.mean_absolute_error(y_test_kn35,y_pred_kn5)))
print('Test R-Squared/Score:', metrics.r2_score(y_test_kn35,y_pred_kn5))
kn_score5 = kn5.score(X_train_kn35,y_train_kn35)
# R-squared of the train
print('Train R-Squared/Score:', kn_score5)
```

```
MAE: 4200.549262223789
MSE: 31340120.5816488
RMSE: 64.81164449559807
Test R-Squared/Score: 0.25622167559331543
Train R-Squared/Score: 0.26839261943639814
```

Our testing accuracy is about 25.6%. Here we can analyze that our test score and train score is very similar to each other. If it's closer to zero percentage, it means that testing did not react to the training set. However, the training dataset scored about 26.8%. I tried various trial and error on adding and removing columns to the machine learning models, but it kept giving me lower high possible scores.

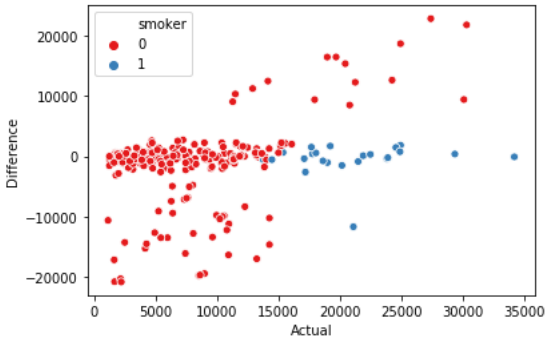
The MAE (mean absolute error) is derived from 1 divided by the total number of data points then multiplying with the sum of absolute value of actual minus predicted results. In this case, the MAE on K-N Regression with 35 neighbors is 4200, MAE on the linear regression was 2335, and MAE on Decision Tree Regression was 3597. As an analyst, we can represent that Decision Tree and K-N Regression models are not good to use since the performance metrics were not impressive on training model. Next approach in getting better and accurate results is to go back to dataset and do more cleaning or change the feature selections.

However, for MSE (mean squared error) the error mostly pertains to the outliers of our model. The further the data points in the model, the higher our MSE will be.

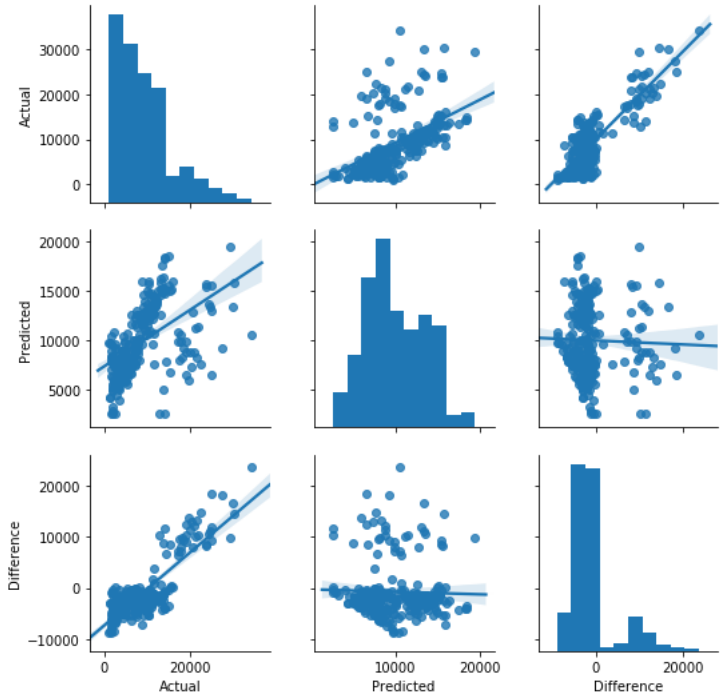
RMSE is the square root of the mean squared error (MSE). We can interpret RMSE and MAE based on how they are spread apart. In our case, RMSE indicates that the model had charges missed by 64.8%.

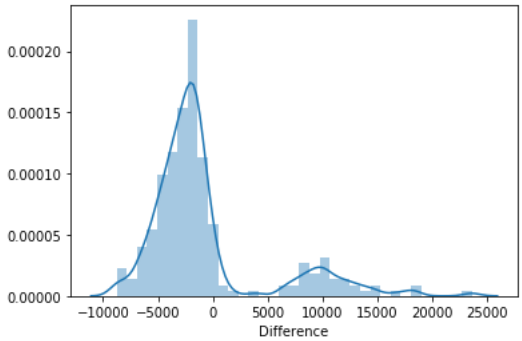
I converted test, prediction, and the difference of the test and prediction to new DataFrame to create plots and understand the results better.

	age	bmi	smoker	Actual	Predicted	Difference
247	30	27.645	0	4237.12655	7702.387421	-3465.260871
302	36	29.920	0	5478.03680	7189.161980	-1711.125180
406	49	36.630	0	10381.47870	12053.109733	-1671.631033
874	36	29.920	0	4889.03680	7189.161980	-2300.125180
877	31	25.900	1	19199.94400	8875.839873	10324.104127



From this plot, we can discuss that those who do smoke saw fewer difference in terms of charge than those who did not smoke. However, those who do not smoke, saw major changes to their prediction charges. We can see about 15 people were overcharged and about 30+ people were undercharged.





Finally, because of how much the difference it shows on the above chart, this model will not be accurate to predict. It will lead to give false results. The line for regression from the above chart references that it is slightly off the trend. The line does not follow the data points like it does with linear regression model.

Evaluation

Based on my basic research and my course knowledge, I learned a lot of different aspects of evaluating and selecting the best models. In real world like the scenario I have chosen about auditing insurance companies, it is critical that there is less errors and more rewards. Insurance companies do not want to deal with complaints and customer churn. They want to make profit, but they should only make profits based on the customer types and not over-charging like we have seen that lot of people were charged extra, even though they had similar characteristics among the group of members, like bmi and age.

On other hands, I would like to pick the linear regression as my best model for deployment. This choice is because given the performance metrics for all the machine learning models, I found that linear regression was prone to have less error and it can predict the charges closely.

Results and Conclusion

In conclusion, from my analytics models and dataset, I found out that age, bmi, and smoking does affect the insurance charges. From the analysis of my linear regression, I found that over 16 people were over charged from the 239 that we tested the model on. Also, in terms of percentage this is only 6.70% of the testing population. From the model, we discovered that majority of the people did not smoke, and they were the once who were over-charged. From the linear regression, I found that about 1 person was under-charged and yet they had history of smoking. I also found out that smoking history is a major factor insurance companies consider when predicting charges. Even if we had increased the person's BMI to obesity, it still would not make big difference as it did with smoking.

I try to use this model and I back tested this model with my family members and friends' medical cost. Since I am using age, bmi, and smoking history, I predicted their results for 3 of my friends and the result I got was somewhat close but not exactly similar to prediction. As I said in the report, linear regression model does show that accuracy was about 64.2% which was higher than other models tested. Due to the nature of this class, I would like to learn more about increasing this accuracy. The features that I did not use such as region and sex was because it was balanced, and it was not correlated to the charges variable.

My recommendation as an auditor for my business is that for those who were over-charged, I would like to know what treatments they did and look at their previous medical bills. I would like the insurance companies to provide them with the discounts to bring the over-priced customers charges down and still make profits. If insurance companies keep on over-charging it will lead to customer churn and they will not be able to make any money from the customers. Also, as an auditor, I would simply re-run the models every few months and analyze if the insurance companies are following these guidelines. On the other hand, I would recommend people should not smoke and exercise daily to stay fit and healthy. We analyzed that average bmi for was about 32.6 and age was 35. This shows that someone in the middle age with 32.6 bmi is not healthy and they are paying about $12,000 more in than what they should have paid.