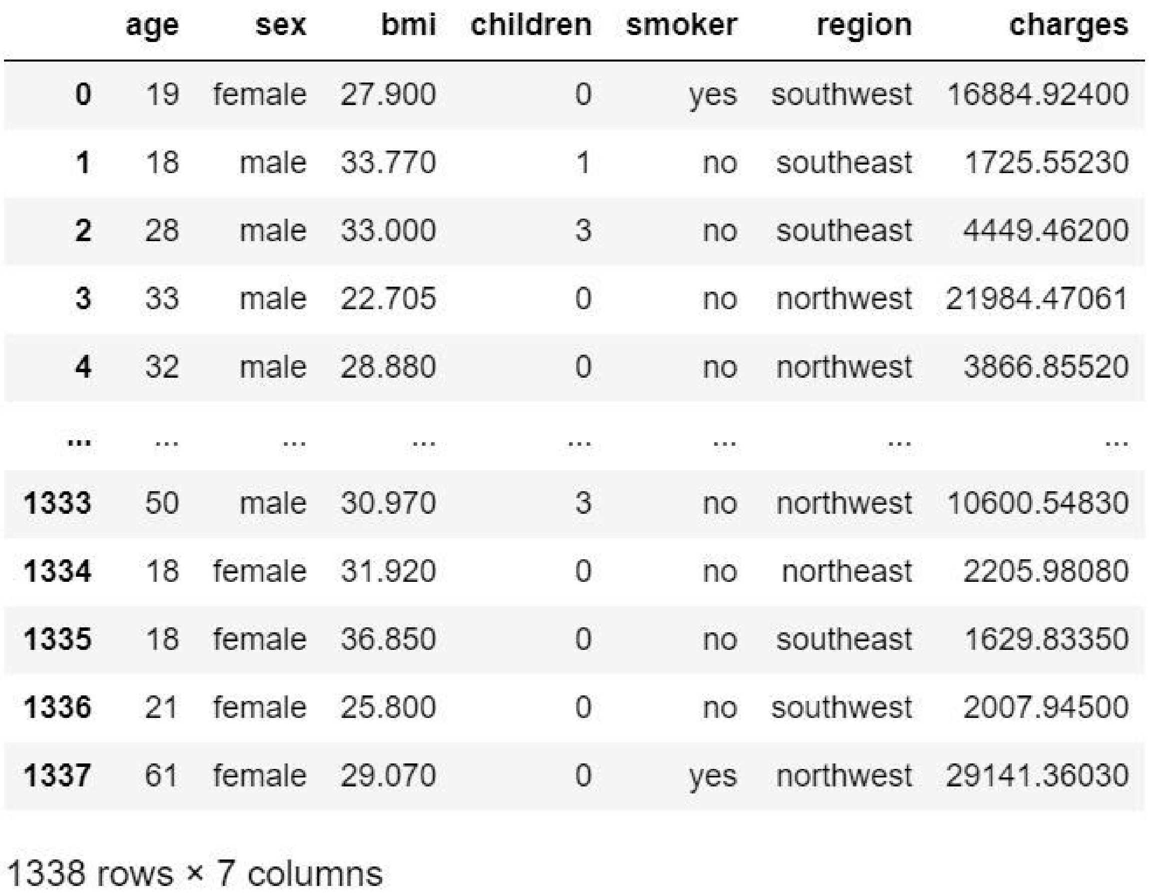
PREDICTING THE LIKELIHOOD OF INSURANCE CHARGES

By Aman Negassi

From a very age, we are constantly taught to be health conscious and that is evident in being told not to do drugs, abuse alcohol consumption, follow your eating habits after the Food Pyramid, and so forth. It is only fitting that there would be a major focus on Healthcare especially in the United States and the topic revolving around coverage for those with pre-existing conditions. The arguments go back and forth in politics where one side argues for increases in regulation believing the U.S. should model their healthcare model after Europe and the other side arguing for deregulation believing that modeling after Europe is not feasible. In the United States, the U.S. Healthcare system is a mix of public and private, for-profit and non-profit insurers and healthcare providers. The United States Federal Government provides funding for programs. There is Medicare which is for adults aged 65 and older and some people with disabilities as well as for various programs for veterans and low-income individuals, including Medicaid and the Children's Health Insurance Program. In 2018, nearly 92% of the population was estimated to have coverage leaving 27.5 million people uninsured. With the dataset provided, I analyzed it as well as to perform models to compare the costs among regions as well as see how much the factors impact the costs.

In the image below is the insurance dataset.



	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
...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	northwest	10600.54830
1334	18	female	31.920	0	no	northeast	2205.98080
1335	18	female	36.850	0	no	southeast	1629.83350
1336	21	female	25.800	0	no	southwest	2007.94500
1337	61	female	29.070	0	yes	northwest	29141.36030

1338 rows × 7 columns

The factors in determining the charges are the age, sex, bmi, children, whether you are a smoker or not, and region in the United States. The region is split into 4 locations: northeast, northwest, southeast, southwest .A useful reference for families.

The Problem Statement

Context

USAA Insurance, a San Antonio-based Fortune 500 Insurance company focused on Insurance, is seeking to improve their revenue margins from $30 billion to $40 billion in the next 3 years. They are looking to improve their predictive modeling with the charges towards insurance policy in order to bring in new customers and compete against other rivaling companies. They also want to make sure they are not overcharging their customers as part of their efforts to increase retention among customers. Their retention rate is roughly 76.4% while customer satisfaction is around 64.2%. As the Chief Data Scientist, I am tasked with this responsibility.

Criteria for Success

USAA maintains a SQL based inquiry system. They distinguish different groups of customers based on their wants and needs as well as satisfaction with the company. The customers that are not satisfied and feel a lack of accommodation will be focused on extensively.

Scope of Solution Space

USAA's inquiry system will be implemented for business use no later than November 5, 2020.

Constraints within Solution Space

There is a likelihood that a segment of customers USAA will be focused on will end up being satisfied with their experience and realize their concerns were misconstrued where the customers they did not focus end up having legitimate concerns about their policy premiums. It's imperative that the communication is proactive between the customers and the company's help support.

Stakeholders to provide key insight

-Data Scientist (Myself)

-Data Engineer-

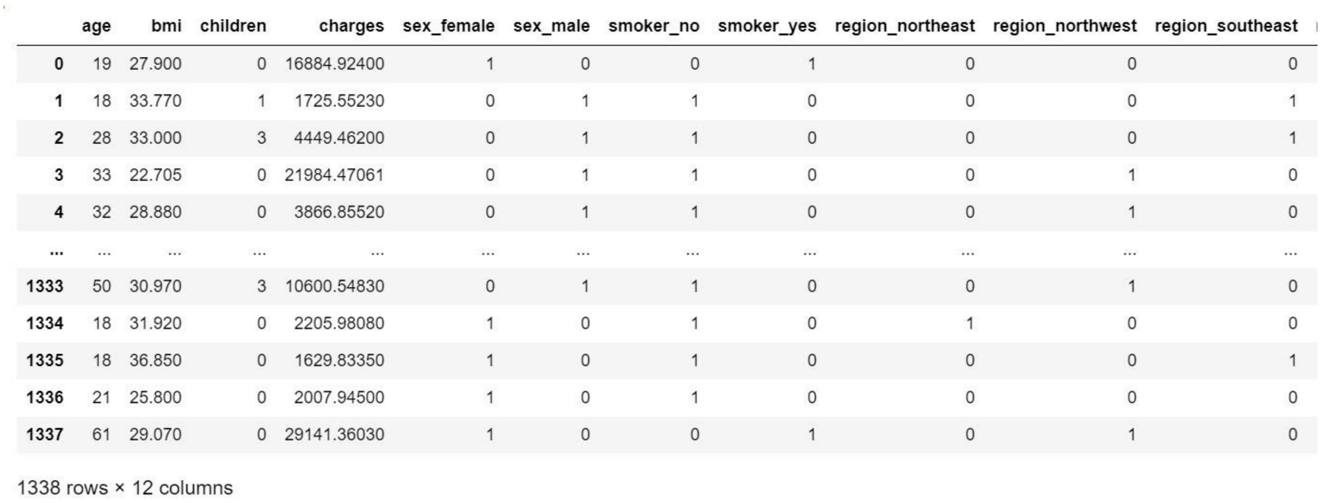
-Data Analyst

What key data sources are required?

-The SQL based Inquiry system

-Tableau for Data Visualization

Data Collection and Wrangling Summary

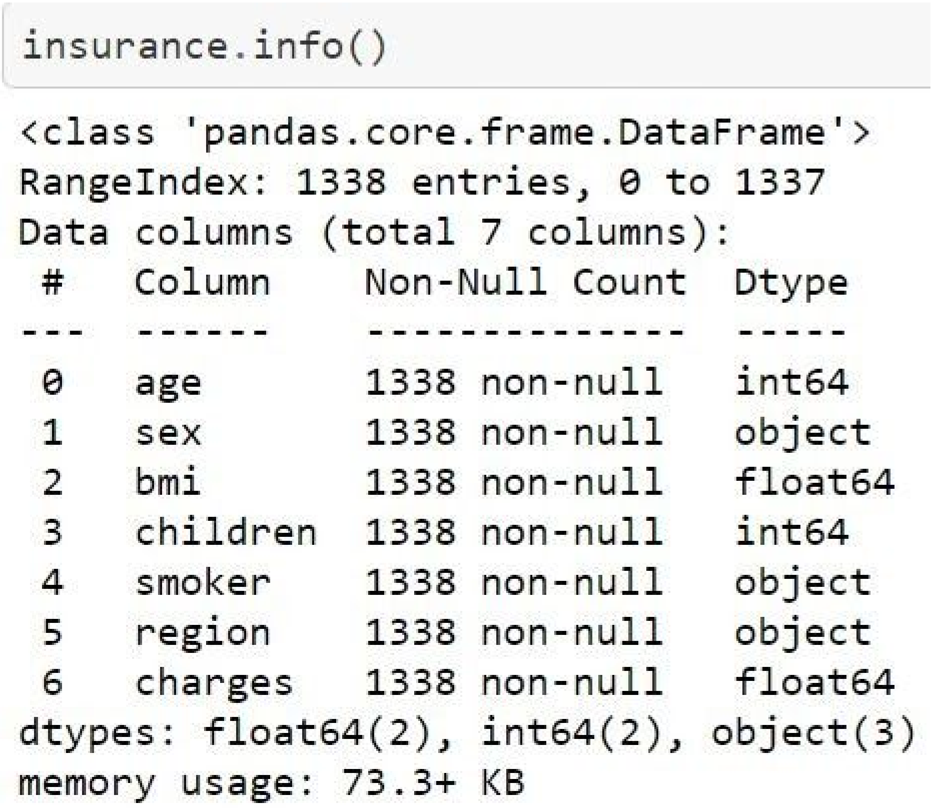


	age	bmi	children	charges	sex_female	sex_male	smoker_no	smoker_yes	region_northeast	region_northwest	region_southeast
0	19	27.900	0	16884.92400	1	0	0	1	0	0	0
1	18	33.770	1	1725.55230	0	1	1	0	0	0	1
2	28	33.000	3	4449.46200	0	1	1	0	0	0	1
3	33	22.705	0	21984.47061	0	1	1	0	0	1	0
4	32	28.880	0	3866.85520	0	1	1	0	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...
1333	50	30.970	3	10600.54830	0	1	1	0	0	1	0
1334	18	31.920	0	2205.98080	1	0	1	0	1	0	0
1335	18	36.850	0	1629.83350	1	0	1	0	0	0	1
1336	21	25.800	0	2007.94500	1	0	1	0	0	0	0
1337	61	29.070	0	29141.36030	1	0	0	1	0	1	0

1338 rows × 12 columns

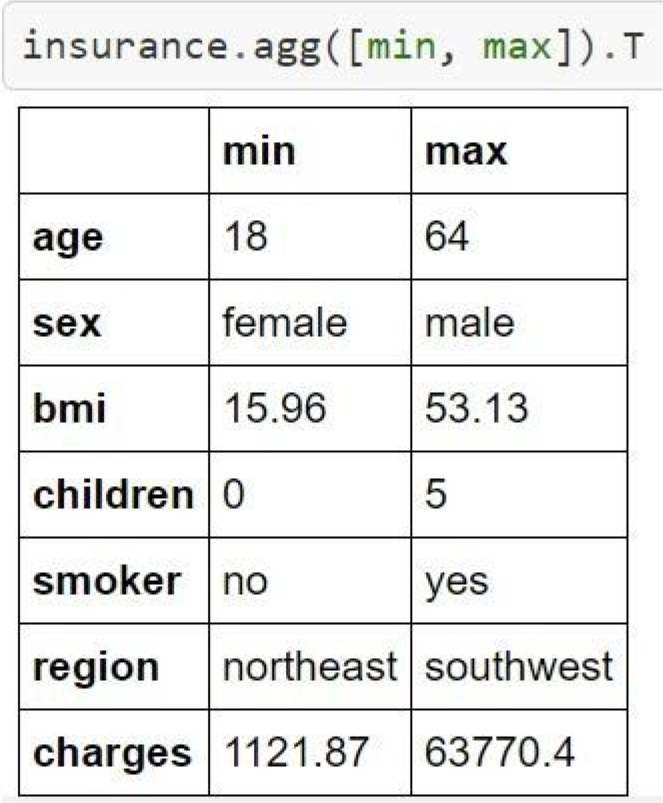
This picture shows the dataset when we create dummy variables, (pd.get_dummies(insurance). Region_southwest is included. The reason this was done was to utilize the categorical and the numerical variables together. Those variables were sex, and region and we made it binary when making it numerical. It becomes more convenient for counting the data to do correlation, scatterplots, and other visualization tools in order to get a better understanding.

Exploratory Data Analysis



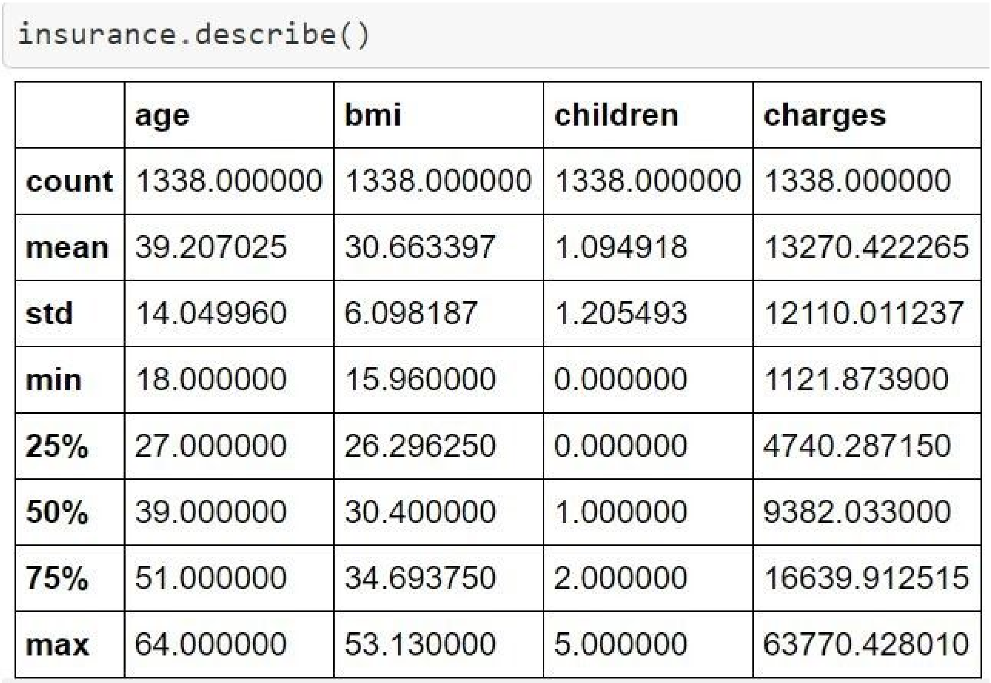
```
insurance.info()
```

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



```
insurance.agg([min, max]).T
```

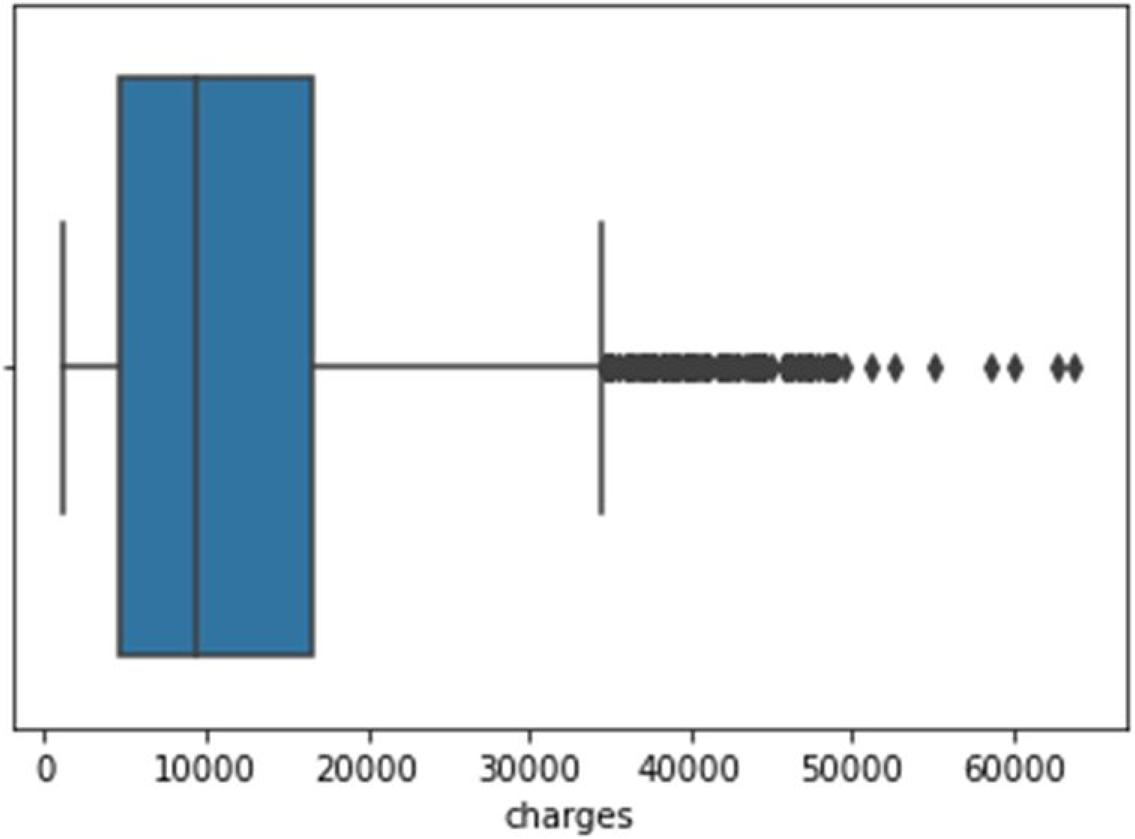
	min	max
age	18	64
sex	female	male
bmi	15.96	53.13
children	0	5
smoker	no	yes
region	northeast	southwest
charges	1121.87	63770.4



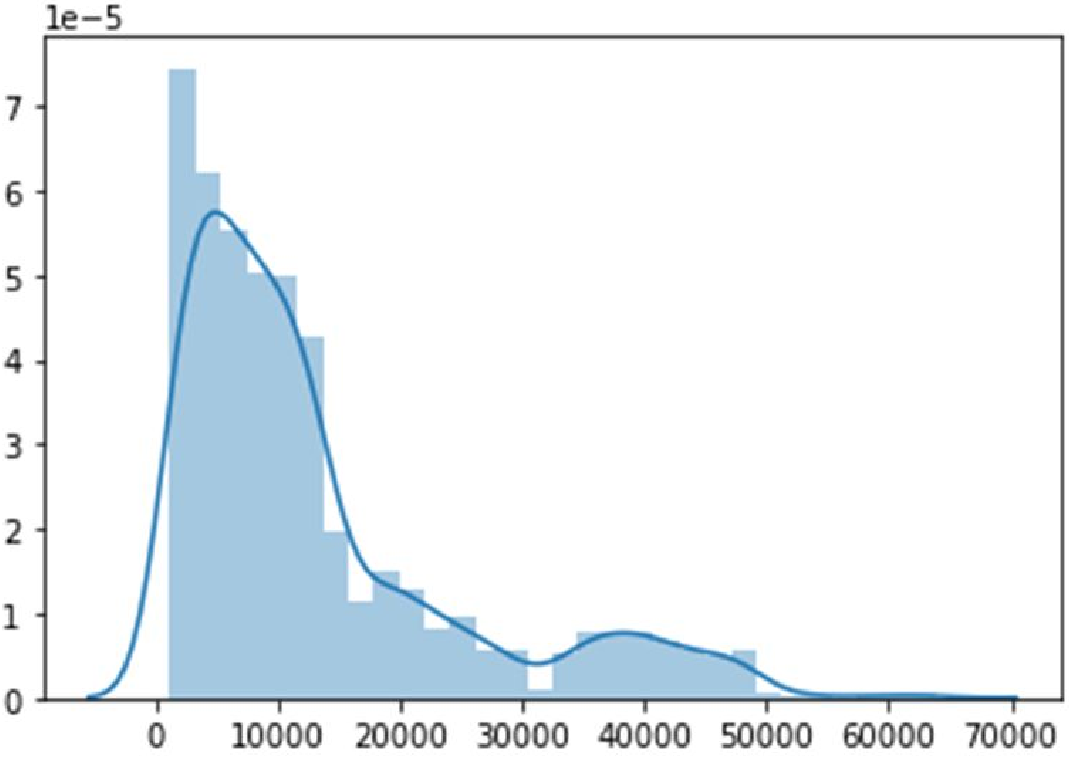
```
insurance.describe()
```

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

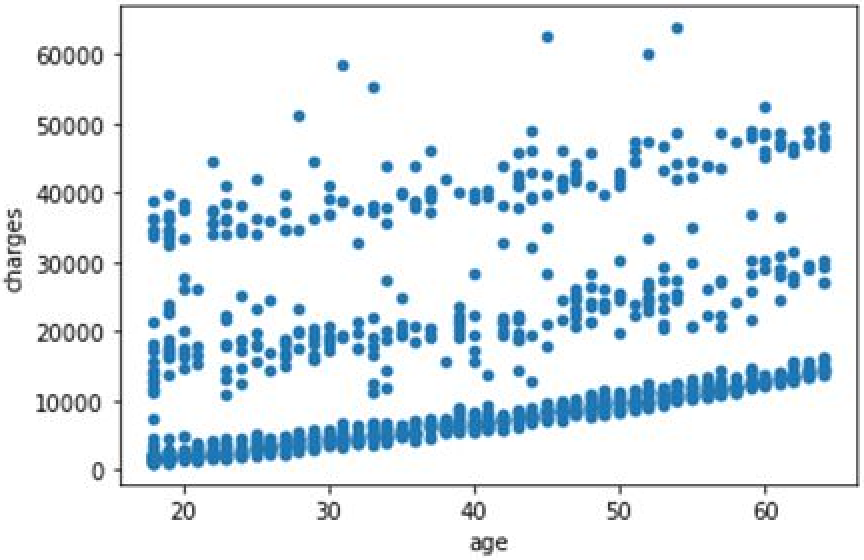
On the left are the columns classified by data types. Followed by that is the minimum and maximum values for each column. On the right is the summary statistics for the columns that are numerical.

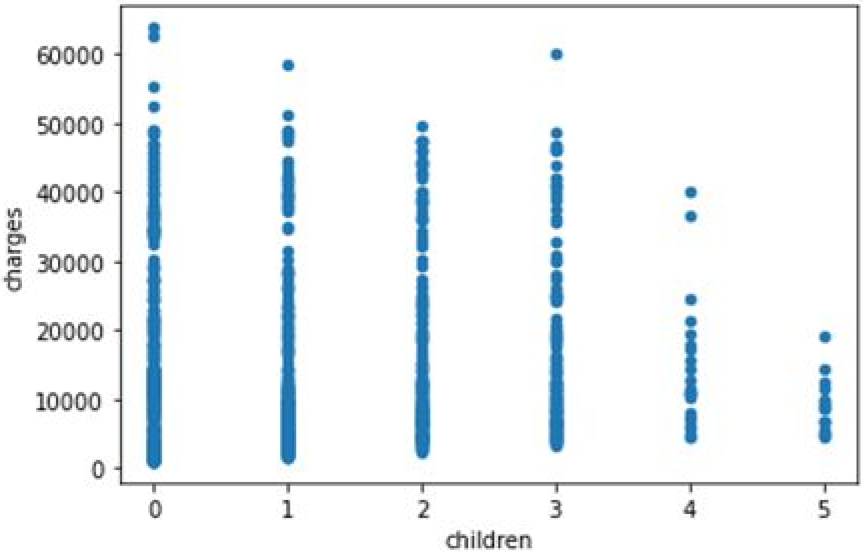


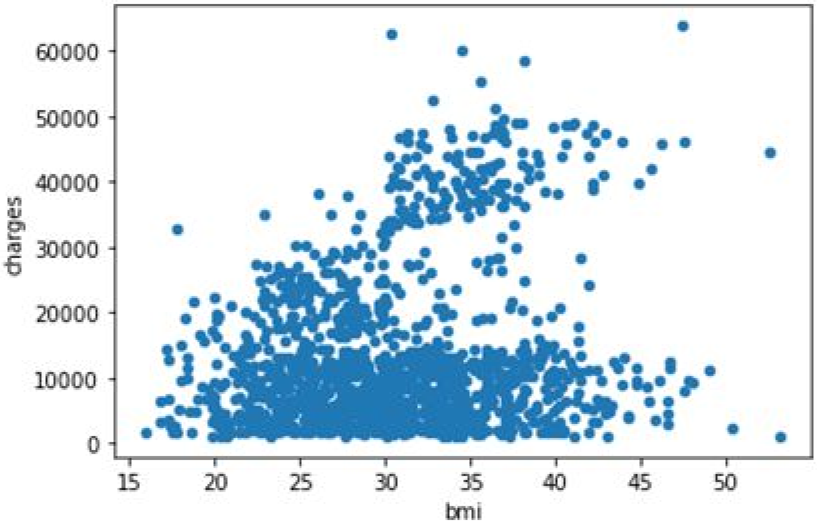
After performing the boxplot, I found the average charges were around $10,000.The outlier seems to start between $30,000-$40,000.



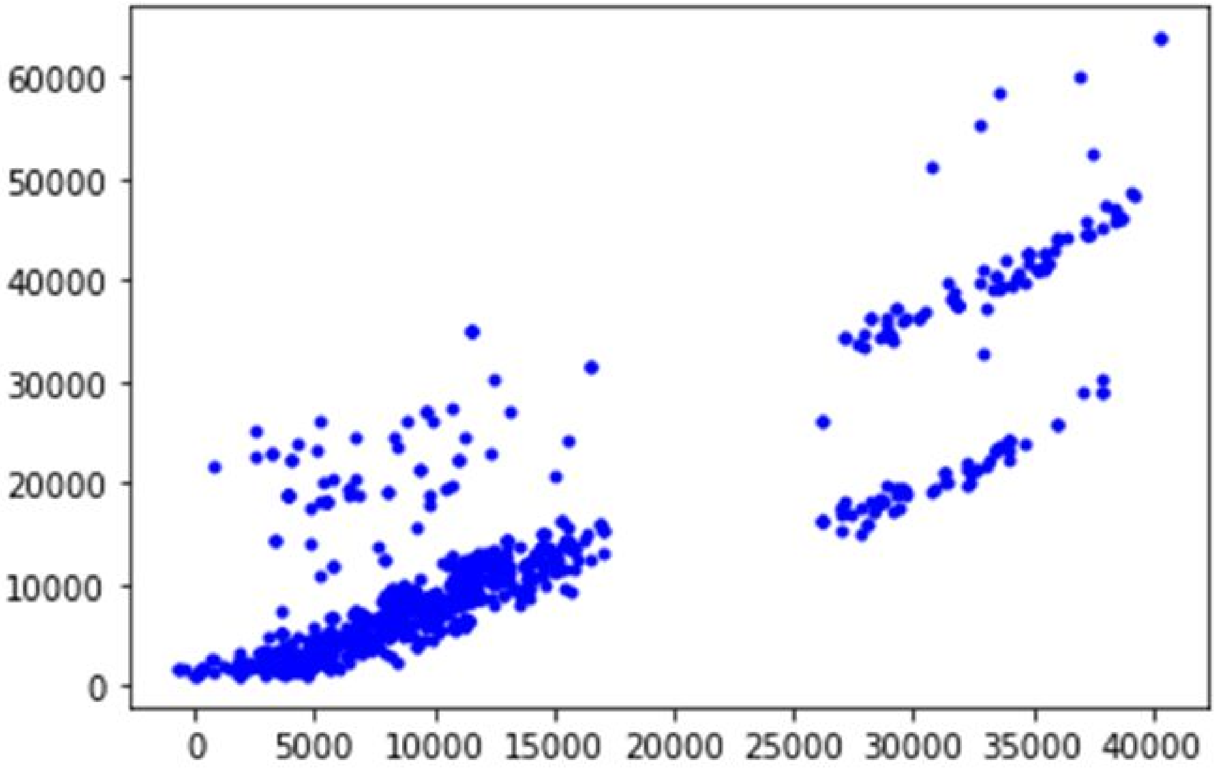
This is the distribution plot of the charges. It skews to the left.







I utilize scatterplots for correlation among the column variables. When taking the scatterplot of the age and charges, as a person gets older, they are likely to pay more in health insurance. Looking at children and charges, we see you are likely to pay less in health insurance as you have more children. With the bmi, the insurance is likely to go up the higher the bmi. Although after a certain point, it does not seem to make a difference.



I built a random number generator followed by picking 60 % of the data for my training set before I will fit the model. It's through the train_test_split and preprocessing that I perform linear regression and utilize the predict method. With the plot, I took the linear regression and subtracted from the training set for the target variable charges.

Linear Regression: As I was conducting Exploratory Data Analysis, Regression had to be utilizing when fitting the column variables. It is also used when building another model that uses both continuous and categorical variables. I also applied the predict to the prediction.

Train_Test_Split: I split the numerical and categorical variables while picking 60% of the data for the training set. Before, I built a random number generator. I separate the variables from the target variable which is charged to determine charges. I create dummy variables, so I am able to utilize the categorical variables turning them into numerical.

Preprocessing.OneHotEncoder: All the variable encoder is doing here is converting our categorical variables to binary. We have three categories, which have two, three, and four possible entries. The encoder uses one binary for each possible response in each category; for example, 'sex' has two columns: one that says yes or no to the individual having 'male' in 'sex', and one that says yes or no to the individual having 'female' in 'sex'. This makes the third category, where we have four possible categories, directly comparable to the others.

Displays the predictions for only the categorical variables

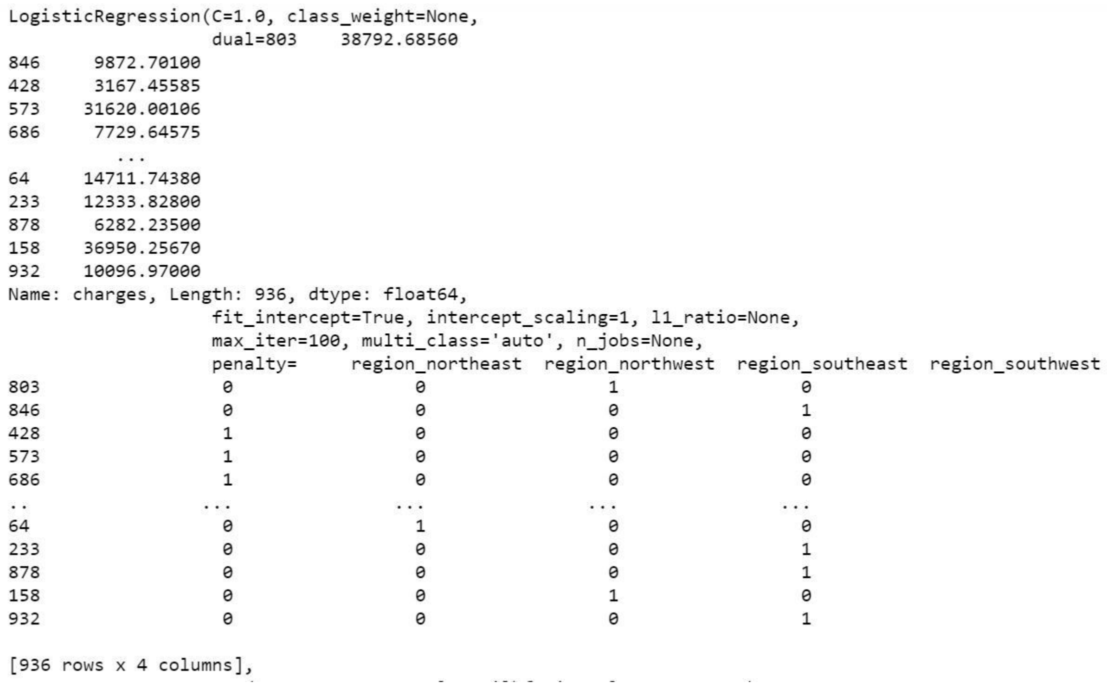
[7393.63001066 8692.45722351 7393.63001066 8214.68123551
7669.71962939 7669.71962939 8210.01839883 32817.16305257....]

Display the predictions for both categorical and numerical variables

array([5.98594069e+03, 1.66025809e+04, 1.92605524e+02, 1.92597268e+03,
        1.34784499e+03, 1.20515083e+04, 1.40021297e+04, 3.28751994e+04...)

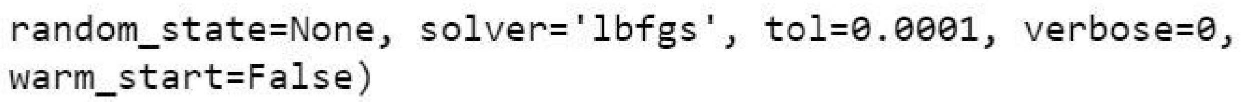
As it can be said, when factoring just sex, and region, the charges are much less than if you factored them all together.

In the image below is the logistic regression of the training set.



```
LogisticRegression(C=1.0, class_weight=None,
                   dual=803    38792.68560
846     9872.70100
428     3167.45585
573    31620.00106
686     7729.64575
          ...
64     14711.74380
233    12333.82800
878     6282.23500
158    36950.25670
932    10096.97000
Name: charges, Length: 936, dtype: float64,
                   fit_intercept=True, intercept_scaling=1, l1_ratio=None,
                   max_iter=100, multi_class='auto', n_jobs=None,
                   penalty=     region_northeast  region_northwest  region_southeast  region_southwest
803                 0                 0                 1                 0
846                 0                 0                 0                 1
428                 1                 0                 0                 0
573                 1                 0                 0                 0
686                 1                 0                 0                 0
..                ...               ...               ...               ...
64                  0                 1                 0                 0
233                 0                 0                 0                 1
878                 0                 0                 0                 1
158                 0                 0                 1                 0
932                 0                 0                 0                 1

[936 rows x 4 columns],
```



```
random_state=None, solver='lbfgs', tol=0.0001, verbose=0,
warm_start=False)
```

REFERENCES

https://www.kaggle.com/annetxu/health-insurance-cost-prediction

https://www.commonwealthfund.org/international-health-policy-center/countries/united-states

https://github.com/Aman101160/Data-Science-Portfolio/blob/master/Predicting_Health_Insurance_Costs.ipynb

https://github.com/Aman101160/Data-Science-Portfolio/blob/master/InsurCostPredPreProcessing%26TrainingDataDevelopment.ipynb