Health Insurance and Factors that Affect the Price of It

Owing to the rising cost of medical treatment, there has been a significant rise in the number of people buying a health insurance policy recently. Health insurance not only covers you during your difficult times but also offers you tax benefits.

While approving your insurance application, the insurance companies generally conduct a thorough assessment of your health profile. Based on their review, they fix the premium charges.

In this project I would like to look at the factors that affect the charges and try to understand what factors are correlated to each other. Also I will try to build a regression model that will help to predict charges based on the factors. You can find the Python code attached to the folder.

Exploratory Data Analysis and Visualisations

Let's take a look on what our data set is about and what features it includes.

	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

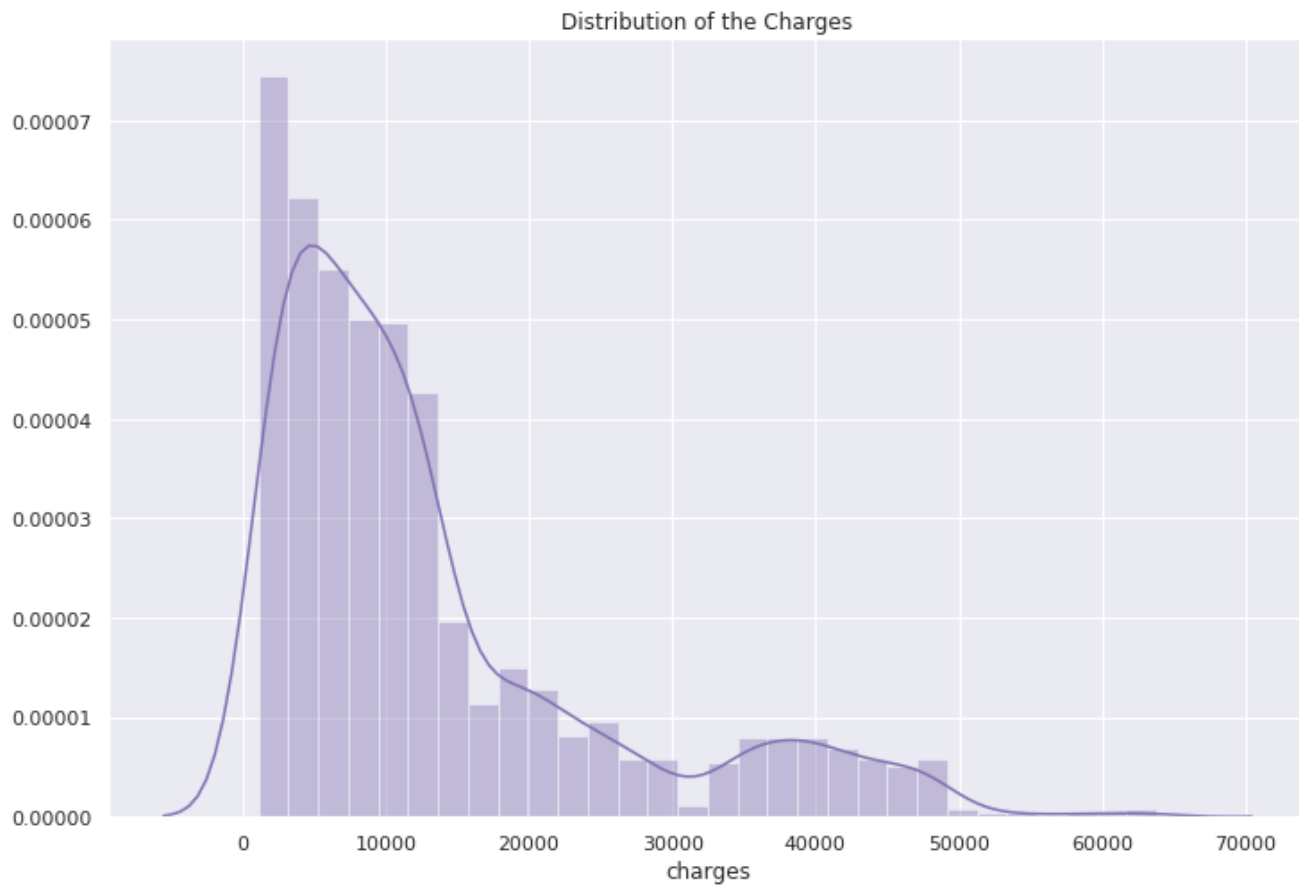
For better understanding what each of the features mean:

- **age:** age of primary beneficiary
- **sex:** insurance contractor gender, female, male
- **BMI – Body Mass Index:** providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- **children:** number of children covered by health insurance / Number of dependents
- **smoker:** smoking
- **region:** the beneficiary's residential area in the US, northeast, southeast, southwest, northwest

There is no NaN values in our data, so we can move on.

age	0
sex	0
bmi	0
children	0
smoker	0
region	0
charges	0

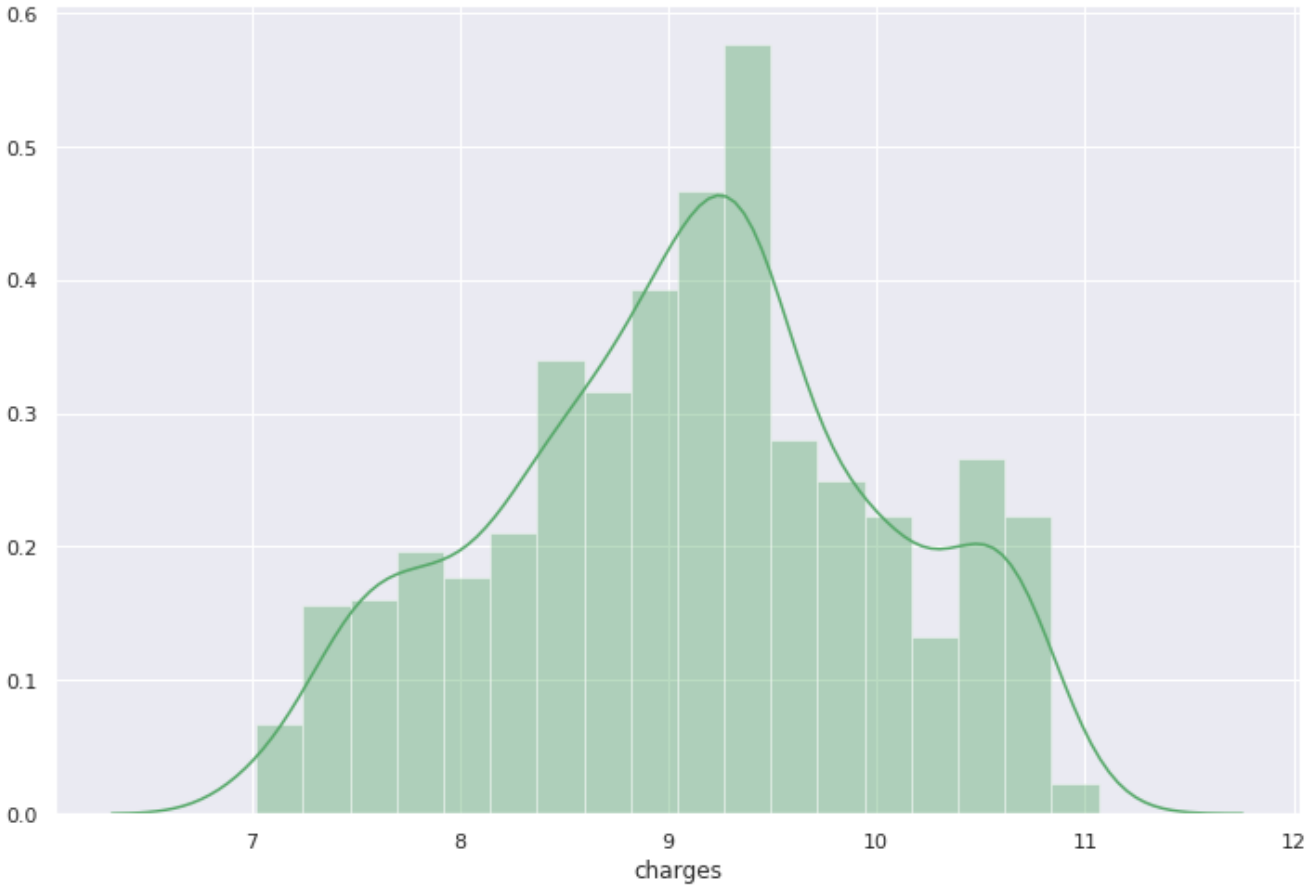
First of all, let's take a look on the distribution of the charges according to the given factors.



The distribution is right-skewed with a skew coefficient: 1.5158796580240388. We can see it from the statistical detailes too.

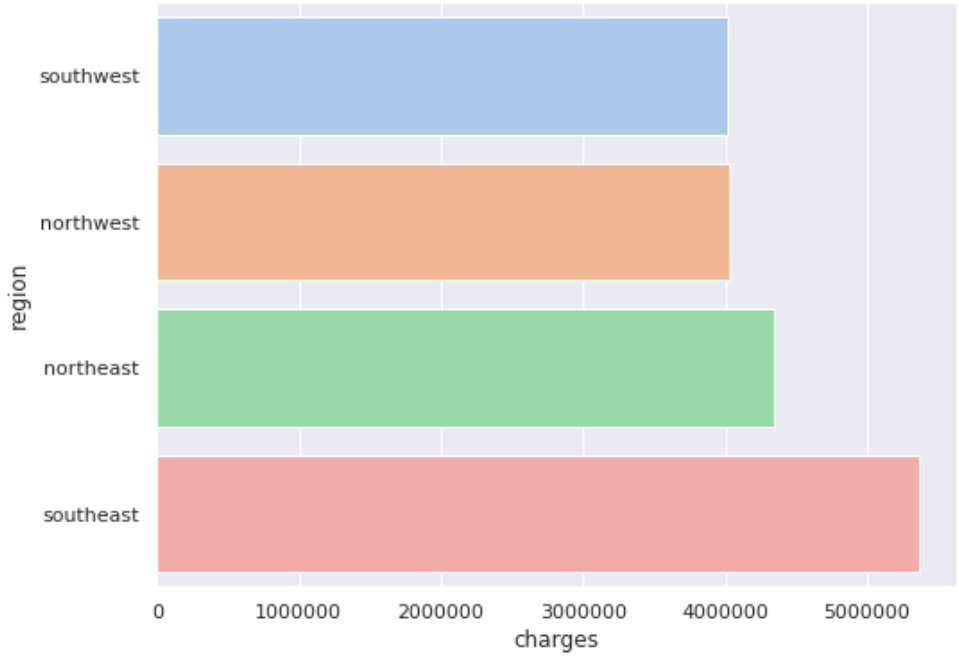
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750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

Real-world data can be messy. Meanwhile, Normal distribution is more reliable to make predictions. To make our data closer to normal we can apply natural log.



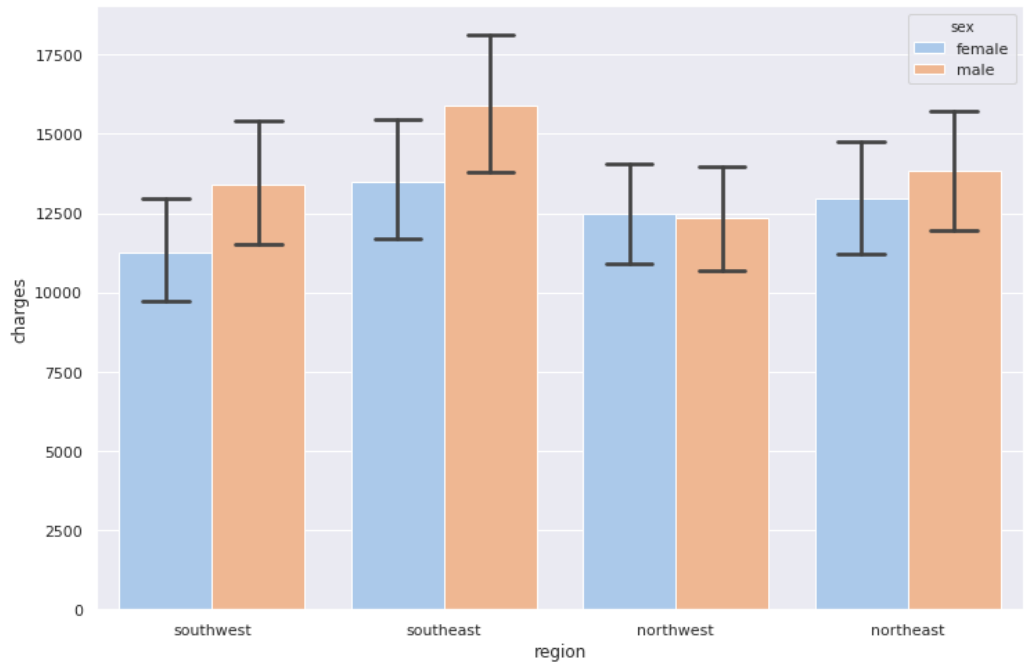
Now it looks much better and closer to the normal. Skew coefficient: -0.09009752473024583.

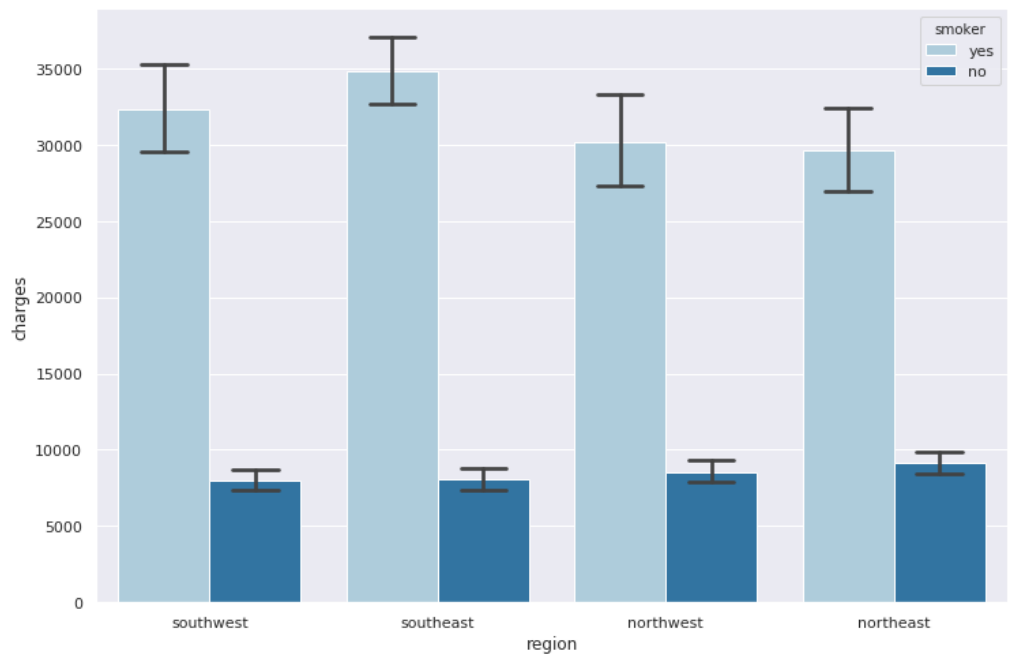
Now let's look at the charges by region.

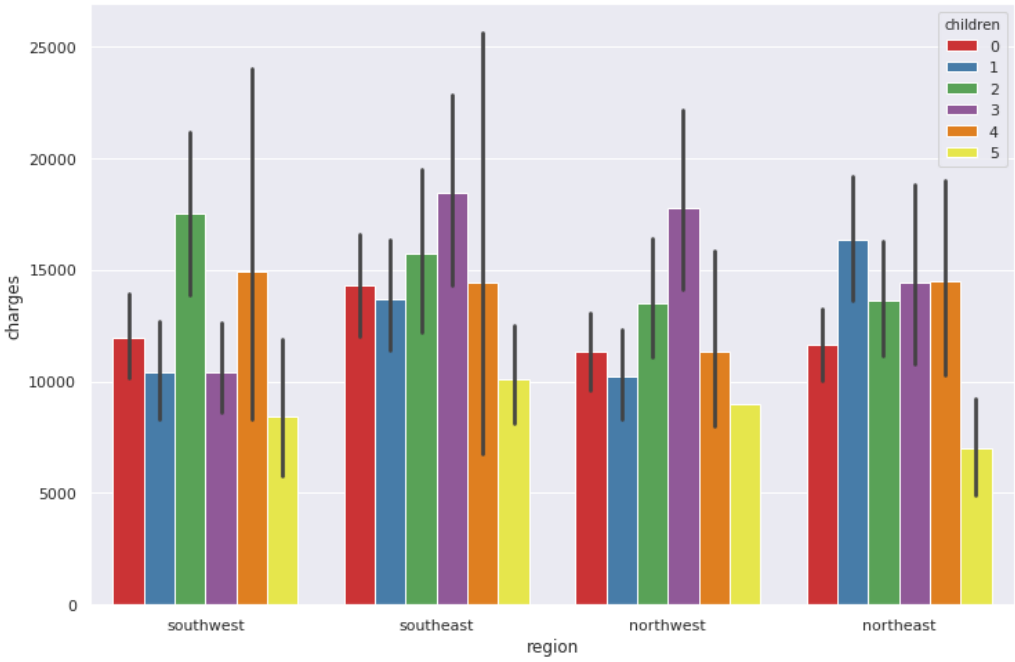


We can see that the charges in southwest and northwest are almost the same, while they are much higher in southeast region.

Now let's see how the charges change not only by region but taking into the account different factors we have in the data set.

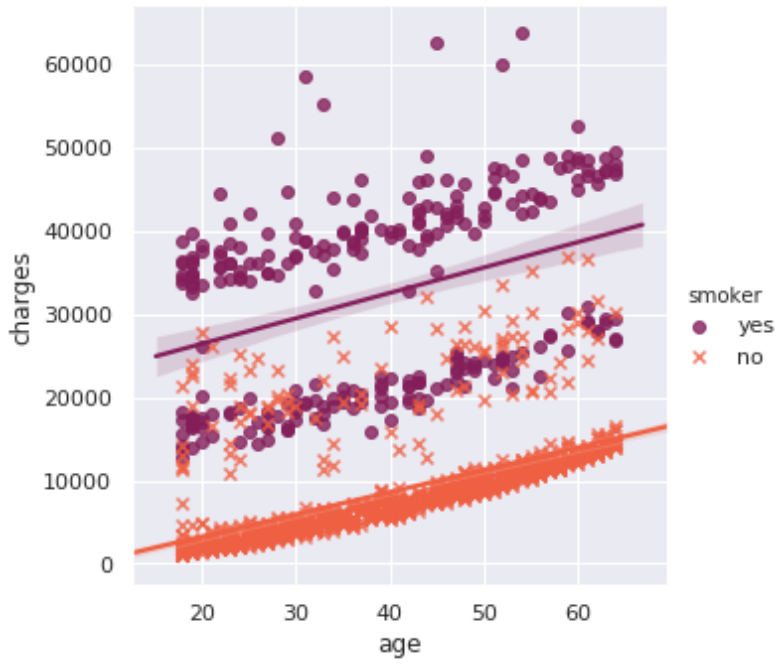


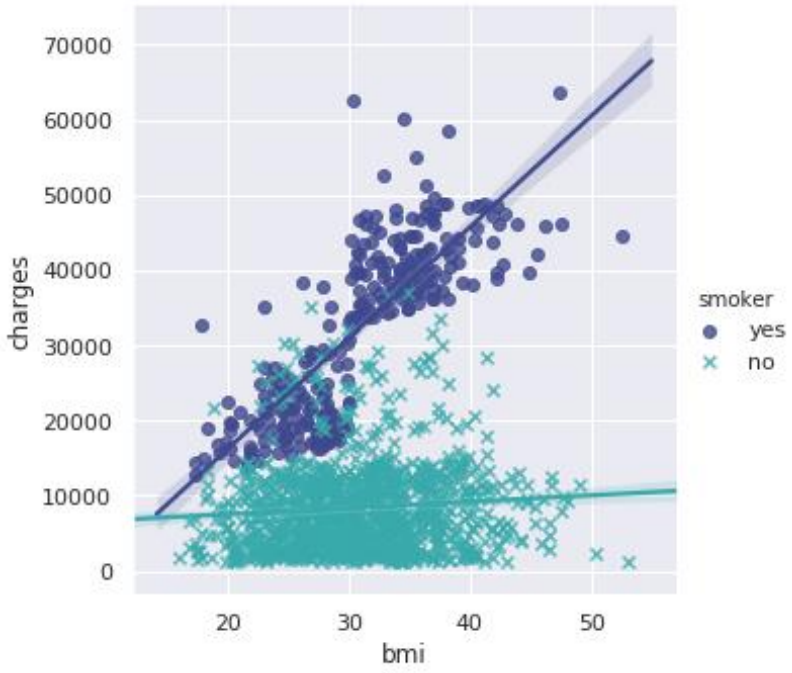


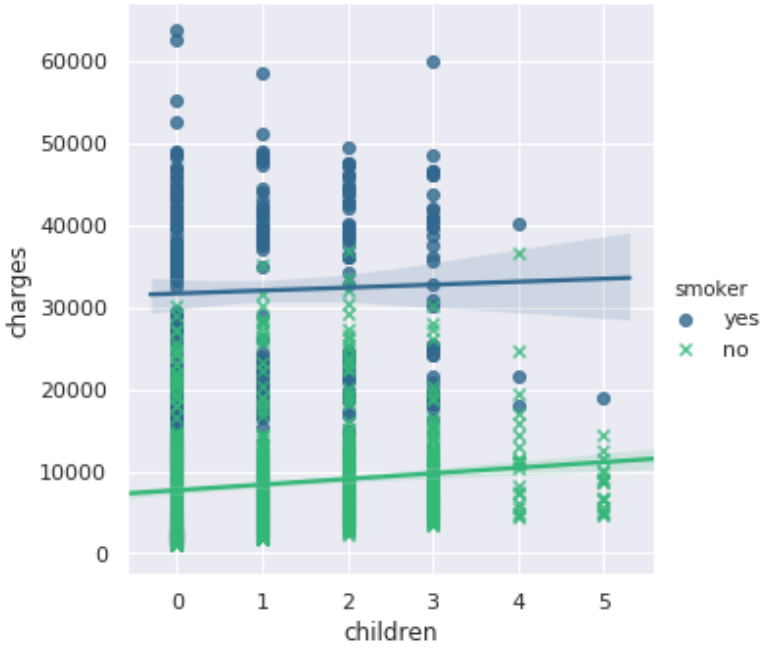


As we can see from the barplots the highest charges based on the smoking criteria are still in the Southeast and the harges for smoking people are much higher in every region, which looks very logical. Charges for male are higher than they are for female in all regions except northwest, where they are almost at the same level. In addition, people with children have higher medical costs overall as well.

Since smoking has the highest impact on medical costs, let's analyze the medical charges by age, BMI and number of children according to the smoking factor.







The costs are growing with age, BMI and children, but smoking has the highest impact on medical costs.

Now let's check if there is any correlations between the factors. For this we need a heatmap.

No correlations, only the smoking again.

Linear Regression

```
from sklearn.model_selection import train_test_split as holdout
from sklearn.linear_model import LinearRegression
from sklearn import metrics
x = df.drop(['charges'], axis = 1)
y = df['charges']
x_train, x_test, y_train, y_test = holdout(x, y, test_size=0.2, random_state=0)
```

Lin_reg = LinearRegression()
Lin_reg.fit(x_train, y_train)
print(Lin_reg.intercept_)
print(Lin_reg.coef_)
print(Lin_reg.score(x_test, y_test))

Results:

```
-11661.983908824435
[  253.99185244   -24.32455098   328.40261701   443.72929547
 23568.87948381  -288.50857254]
0.7998747145449959
```

The results are good enough but it is possible to improve it by removing some unimportant features. Based on the heatmap, we can see that sex and region factors are less important.

Polynomial Regression

from sklearn.preprocessing import PolynomialFeatures
x = df.drop(['charges', 'sex', 'region'], axis = 1)
y = df.charges
pol = PolynomialFeatures (degree = 2)
x_pol = pol.fit_transform(x)
x_train, x_test, y_train, y_test = holdout(x_pol, y, test_size=0.2, random_state=0)
Pol_reg = LinearRegression()
Pol_reg.fit(x_train, y_train)
y_train_pred = Pol_reg.predict(x_train)
y_test_pred = Pol_reg.predict(x_test)
print(Pol_reg.intercept_)
print(Pol_reg.coef_)
print(Pol_reg.score(x_test, y_test))

```
-5325.881705252783
[ 0.00000000e+00 -4.01606591e+01  5.23702019e+02  8.52025026e+02
 -9.52698471e+03  3.04430186e+00  1.84508369e+00  6.01720286e+00
  4.20849790e+00 -9.38983382e+00  3.81612289e+00  1.40840670e+03
 -1.45982790e+02 -4.46151855e+02 -9.52698471e+03]
0.8812595703345235
```

#Evaluating the performance of the algorithm
print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, y_test_pred))
print('Mean Squared Error:', metrics.mean_squared_error(y_test, y_test_pred))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, y_test_pred)))

```
Mean Absolute Error: 2824.4950454776495
Mean Squared Error: 18895160.098780256
Root Mean Squared Error: 4346.85634669243
```

#Predicting the charges
y_test_pred = Pol_reg.predict(x_test)
#Comparing the actual output values with the predicted values
df = pd.DataFrame({'Actual': y_test, 'Predicted': y_test_pred})
print(df)

	Actual	Predicted
578	9724.530000	12101.156323
610	8547.691300	10440.782266
569	45702.022350	48541.022951
1034	12950.071200	14140.067522
198	9644.252500	8636.235727
981	4500.339250	5072.787029
31	2198.189850	3090.494817
1256	11436.738150	13171.361938
1219	7537.163900	9187.612192
1320	5425.023350	7496.320857
613	6753.038000	6653.904925
1107	10493.945800	11893.766490
1263	7337.748000	9291.317273
406	4185.097900	5326.271479
795	19319.742000	25726.734553

Results

- Smoking affects the insurance' price the most.
- Then the age and BMI.
- Polynomial regression provides better results for predictions than a linear one.