Medical Cost Personal Datasets

INSURANCE FORECAST BY USING LINEAR REGRESSION

Md Bin Amin Hossan | Data science- CMP020N205S | Date: 09 May 2023

Index

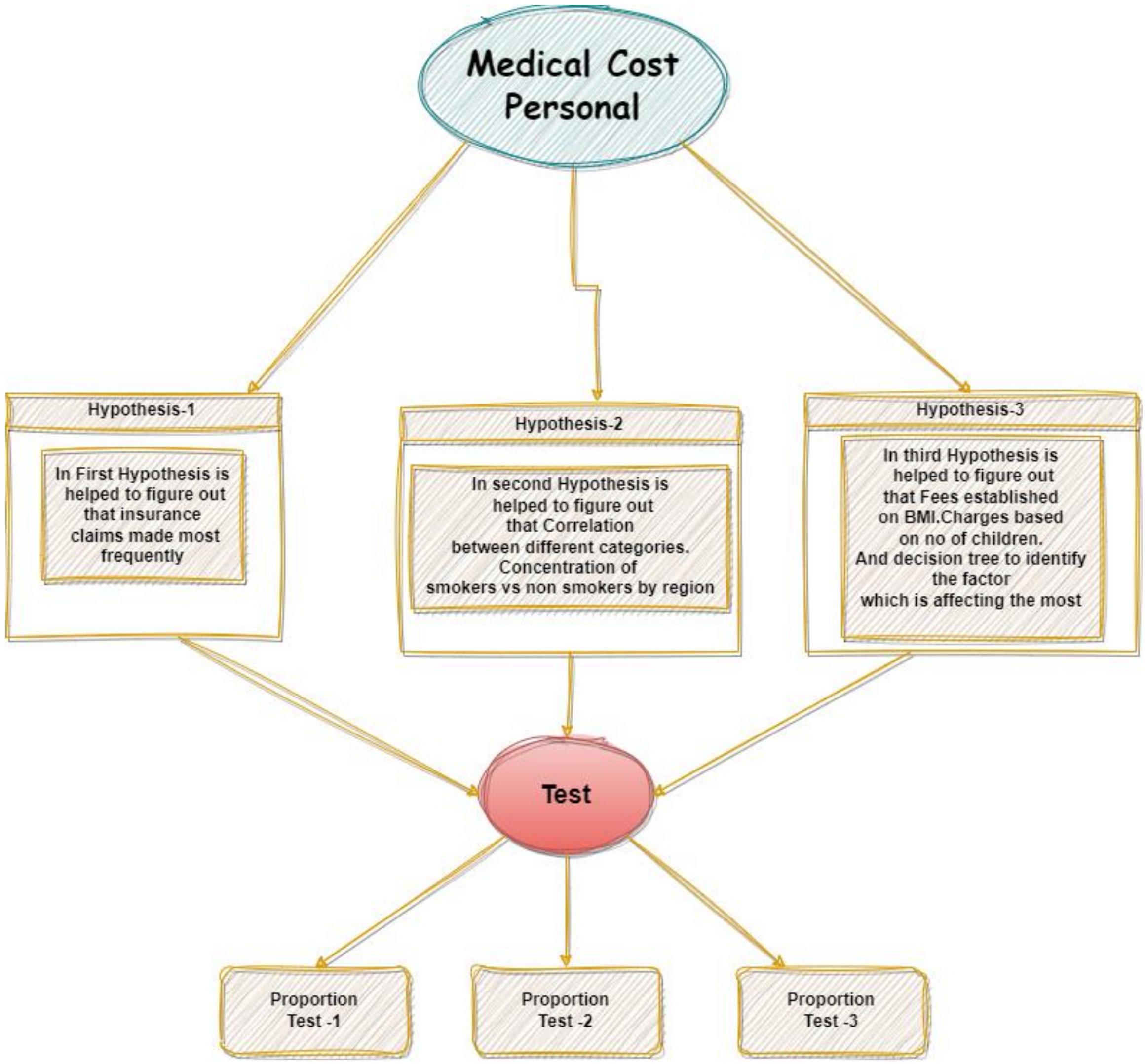
Introduction

A crucial component of healthcare is the medical field, which focuses on the identification, management, and prevention of ailments and diseases. It is a field that is always evolving, leading to better patient outcomes and general health through new discoveries, inventions, and technology. In general, the medical industry is essential to ensuring the health of people, communities, and society at large.

In this report, the data set that we acquired from ('Medical Cost Personal Datasets | Kaggle'), this website will be part of our discussion. This data set's publisher is Brett Lantz. Most of the datasets are in the hands of the general public; they simply had to be recorded and cleaned up to fit with the book's style.

DATA	COLUMNS	DESCRIPTION
Categorical	SEX	Insurance contractor gender, female, male.
Categorical	SOMKER	Smoking
Categorical	REGION	the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
Numerical	AGE	Age of primary beneficiary
Numerical	BMI (objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9)	Body mass index, providing an understanding of body, weights that are relatively high or low relative to height.
Numerical	CHILDREN	Number of children covered by health insurance / Number of dependents.
Numerical	CHARGES	Individual medical costs billed by health insurance.

Hypothesis:


Medical Cost Personal
Hypothesis-1
In First Hypothesis is helped to figure out that insurance claims made most frequently
Hypothesis-2
In second Hypothesis is helped to figure out that Correlation between different categories. Concentration of smokers vs non smokers by region
Hypothesis-3
In third Hypothesis is helped to figure out that Fees established on BMI.Charges based on no of children. And decision tree to identify the factor which is affecting the most
Test
Proportion Test -1
Proportion Test -2
Proportion Test -3

Pre-Processing

Step 1: Firstly, I have eliminated the children column because I no longer use it. Children column has deleted because I do not work on this column.

age	sex	bmi	children	smoker	region	charges
19	female	27.9	0	yes	southwest	16884.92
18	male	33.77	1	no	southeast	1725.552
28	male	33	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.47
32	male	28.88	0	no	northwest	3866.855
31	female	25.74	0	no	southeast	3756.622
46	female	33.44	1	no	southeast	8240.59
37	female	27.74	3	no	northwest	7281.506
37	male	29.83	2	no	northeast	6406.411
60	female	25.84	0	no	northwest	28923.14
25	male	26.22	0	no	northeast	2721.321
62	female	26.29	0	yes	southeast	27808.73
23	male	34.4	0	no	southwest	1826.843
56	female	39.82	0	no	southeast	11090.72
27	male	42.13	0	yes	southeast	39611.76
19	male	24.6	1	no	southwest	1837.237
52	female	30.78	1	no	northeast	10797.34
23	male	23.845	0	no	northeast	2395.172
56	male	40.3	0	no	southwest	10602.39
30	male	35.3	0	yes	southwest	36837.47
60	female	36.005	0	no	northeast	13228.85
30	female	32.4	1	no	southwest	4149.736
18	male	34.1	0	no	southeast	1137.011
34	female	31.92	1	yes	northeast	37701.88

insurance

Step 2: I made the dataset normalize by removing all the duplicate rows. In addition, the dataset I just deleted from has some values missing, even though I double-checked all the names and values to prevent any mistakes or inconsistencies in the analysis.

age	sex	bmi	smoker	region	charges
19	female	27.9	yes	southwest	16884.92
18	male	33.77	no	southeast	1725.552
28	male	33	no	southeast	4449.462
33	male	22.705	no	northwest	21984.47
32	male	28.88	no	northwest	3866.855
31	female	25.74	no	southeast	3756.622
46	female	33.44	no	southeast	8240.59
37	female	27.74	no	northwest	7281.506
37	male	29.83	no	northeast	6406.411
60	female	25.84	no	northwest	28923.14
25	male	26.22	no	northeast	2721.321
62	female	26.29	yes	southeast	27808.73
23	male	34.4	no	southwest	1826.843
56	female	39.82	no	southeast	11090.72
27	male	42.13	yes	southeast	39611.76
19	male	24.6	no	southwest	1837.237
52	female	30.78	no	northeast	10797.34
23	male	23.845	no	northeast	2395.172
56	male	40.3	no	southwest	10602.39
30	male	35.3	yes	southwest	36837.47
60	female	36.005	no	northeast	13228.85
30	female	32.4	no	southwest	4149.736
18	male	34.1	no	southeast	1137.011
34	female	31.92	yes	northeast	37701.88

insurance

Ready Accessibility: Unavailable

To determine how many values are missing,

```
# Calculate the number of missing values in each column
missing_counts = ins.isna().sum(axis=0)

# Print the results
print(missing_counts)
```

```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

```
# Display the structure of the data frame
print(ins.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
None
```

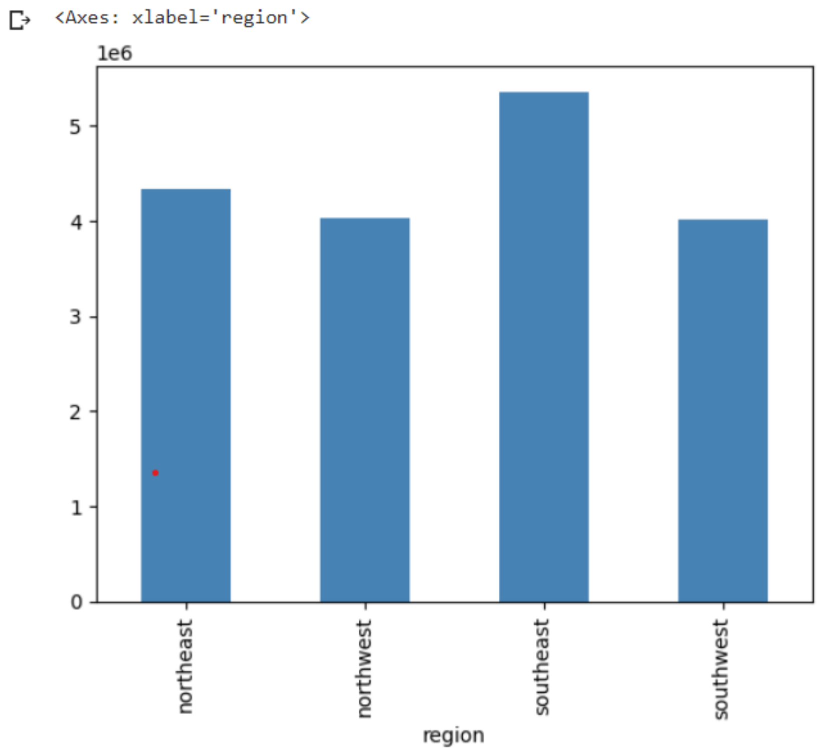
Proposed solution

For this dataset I am using linear regression to solve this problem. Predictive analysis and modelling frequently use linear regression. The relative effects of age, gender, and diet (the predictor factors) on height (the outcome variable), for instance, can be measured using this method.

Hypothesis-1: In first hypothesis is helped to figure out about Insurance prediction. Which region are insurance claims made most frequently?

Solution: For find out hypothesis 1, I am using a programing language (python) to fix this dataset. Where are insurance claims made most frequently? For insurance claims of these areas Northeast, Northwest, Southeast and Southwest are applied for calculation which area are maximum insurance claims in region.

```
# insurance claims are maximum in which region
ins.groupby('region')['charges'].sum().plot(kind='bar',
color='steelblue')
```



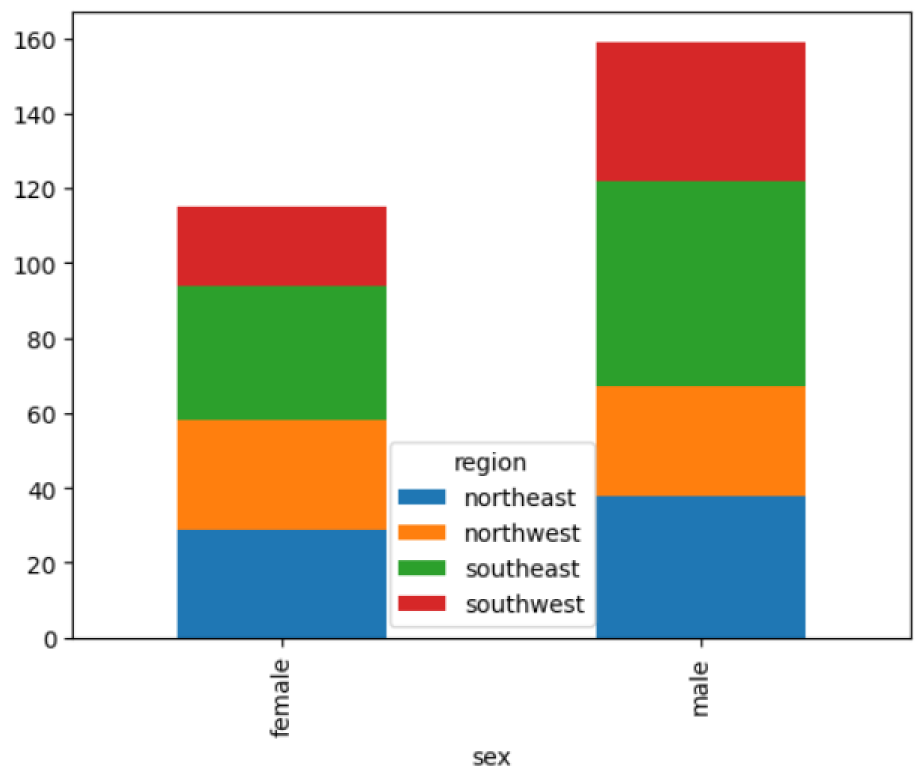
Hypothesis-2: In this second hypothesis-2, Finding somkers and non smokers in region by appling dataset of 'sex', 'region' & 'smoker'. Where is shown in below, female and male gender where male gender were highly smokers.

Relation between different cattagories such as 'AGE','BMI','Children' and ' smoker' as well as 'Bills'. Concentration of smokers vs non smokers by region. Concentration of smokers vs non smokers by region.

Smokers:

```python
# smokers
ins.groupby(['sex', 'region'])['smoker'].apply(lambda x: (x ==
'yes').sum()).unstack().plot(kind='bar', stacked=True)
```

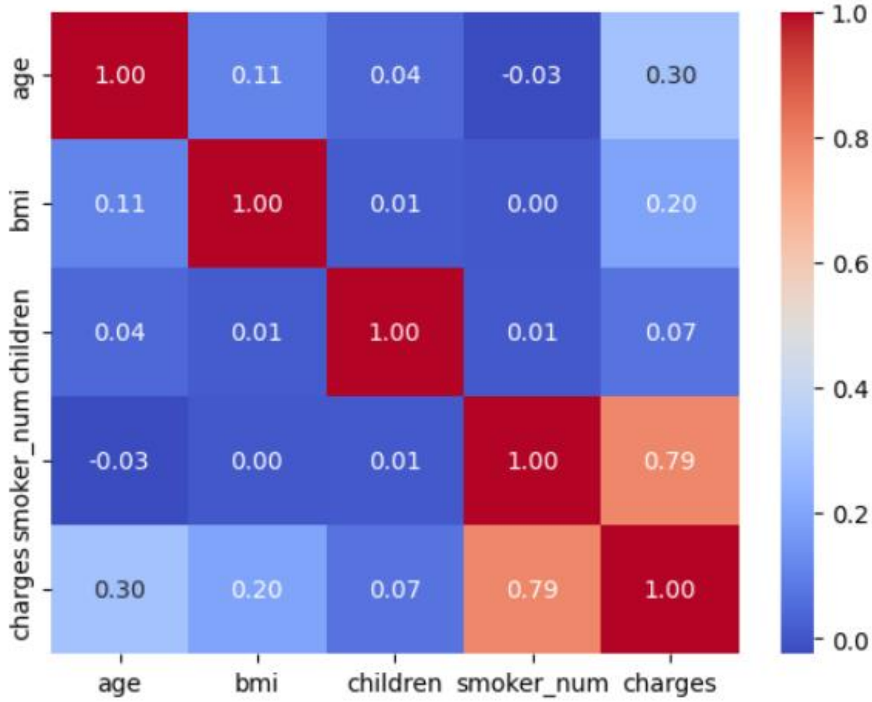
```
<Axes: xlabel='sex'>
```



Correlation between different categories

```python
import seaborn as sns
# correlation between different categories
ins1 = ins[['age', 'bmi', 'children', 'smoker_num', 'charges']]
ins1['smoker_num'] = pd.to_numeric(ins1['smoker_num'])
```

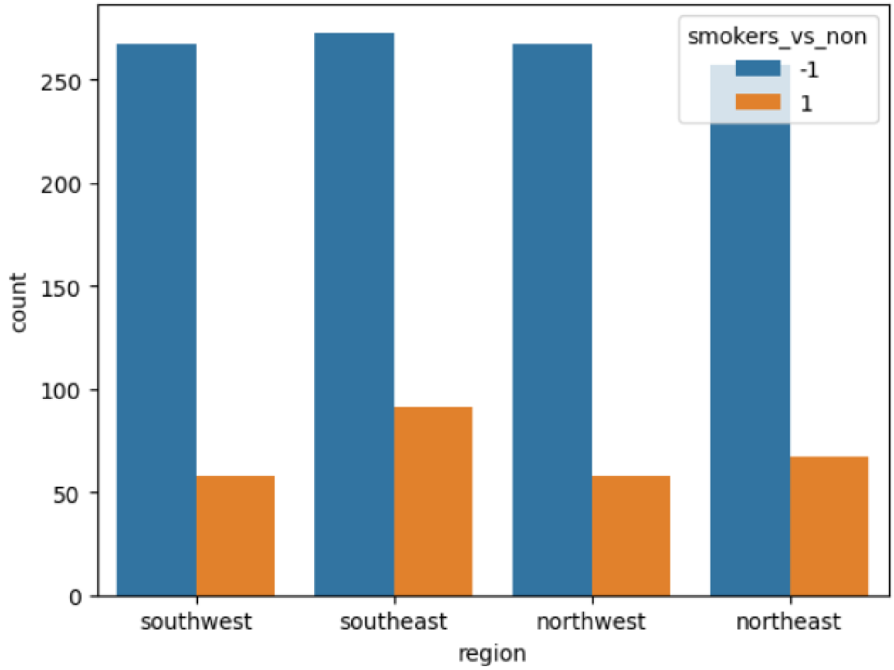
```
corr = ins1.corr()
sns.heatmap(corr, cmap='coolwarm', annot=True, fmt='.2f',
annot_kws={"size": 10})
```



Concentration of smokers vs non smokers by region

```
# concentration of smokers vs non-smokers by region
ins['smokers_vs_non'] = np.where(ins['smoker_num'] == 0, -1, 1)
sns.countplot(x='region', hue='smokers_vs_non', data=ins)
```

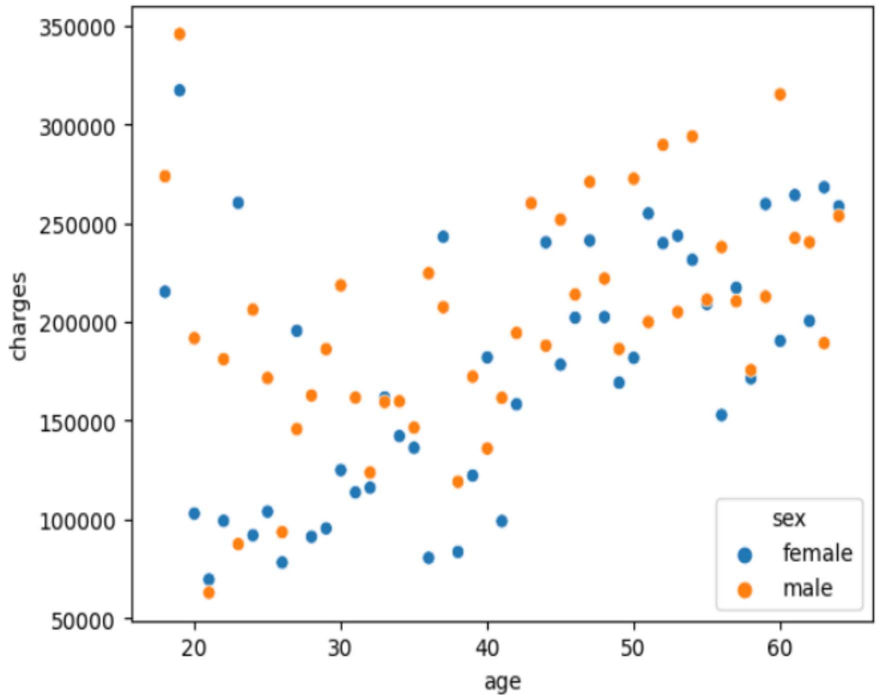
```
<Axes: xlabel='region', ylabel='count'>
```



Charges by different age group

```
# charges by different age group
sns.scatterplot(x='age', y='charges', hue='sex',
data=ins.groupby(['sex', 'age'])['charges'].sum().reset_index())
```

<Axes: xlabel='age', ylabel='charges'>

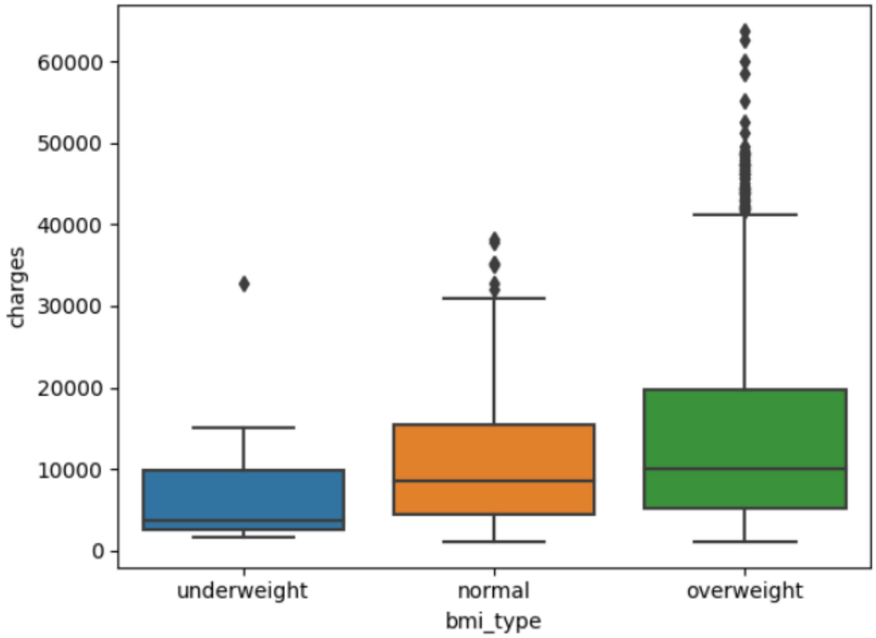


Hypothesis-3: In this hypothesis-3, we are going to finding how much the pay fees by BMI. In Linear regression with scaled data, we calculate RMSE. Decision tree to identify the factor which is affecting the most and find the accuracy as well as Random Forest which predicting charge of medical fees and Actual fees.

Fees established on BMI

```
# charges based on bmi
ins['bmi_type'] = pd.cut(ins['bmi'], bins=[0, 18, 30,
float('inf')], labels=['underweight', 'normal', 'overweight'])
sns.boxplot(x='bmi_type', y='charges', data=ins)
```

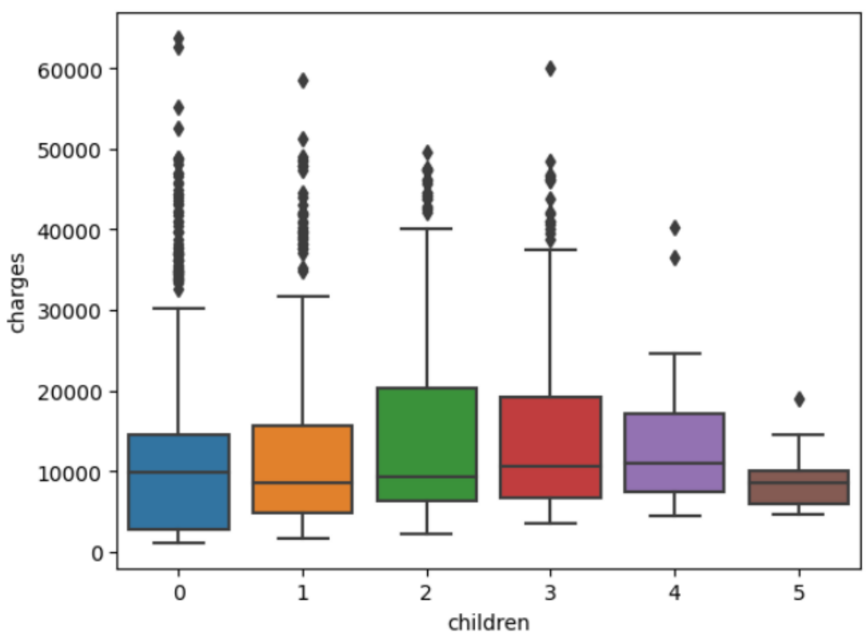
```
<Axes: xlabel='bmi_type', ylabel='charges'>
```



Charges based on no of children

```python
# charges based on no of children
sns.boxplot(x='children', y='charges', data=ins)
```

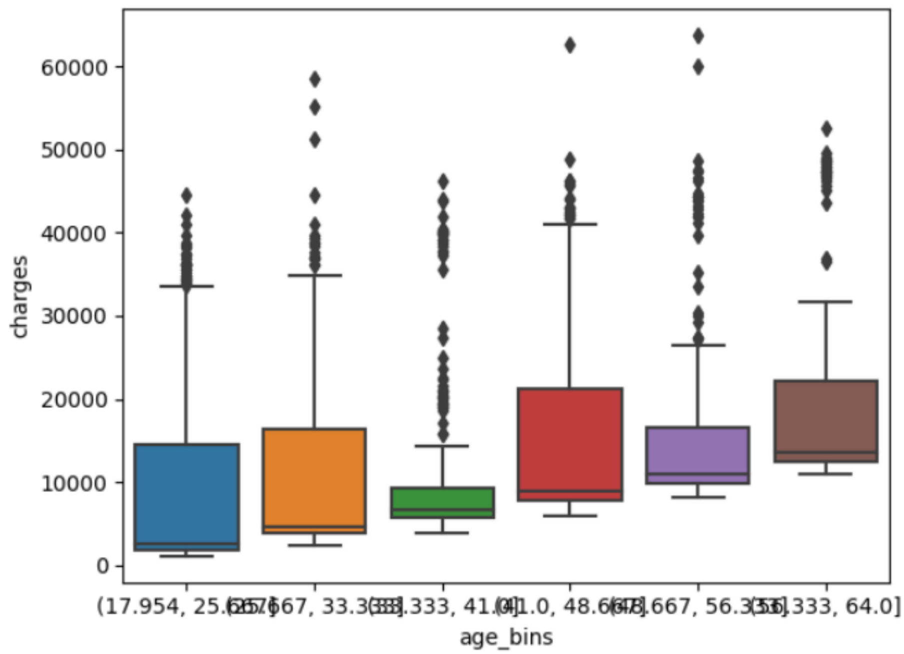
```
<Axes: xlabel='children', ylabel='charges'>
```



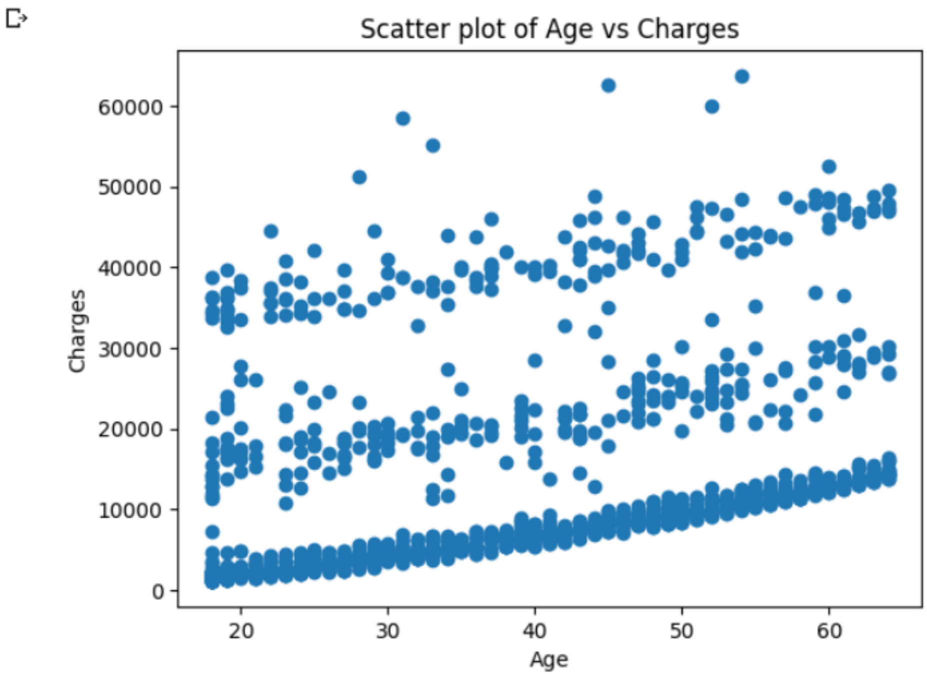
Charges based on age.

```
# charges based on age
ins['age_bins'] = pd.cut(ins['age'], bins=6)
sns.boxplot(x='age_bins', y='charges', data=ins)
```

<Axes: xlabel='age_bins', ylabel='charges'>



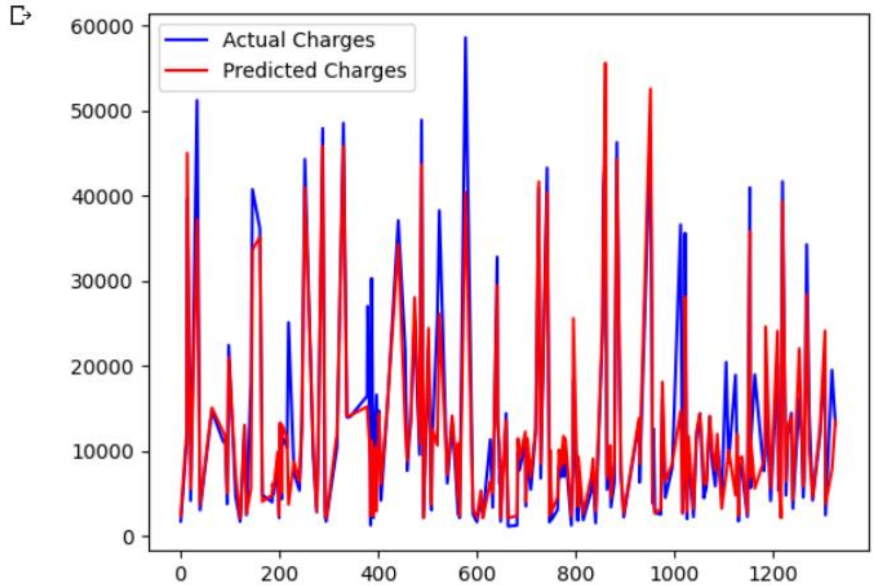

```
# Selecting columns age and charges from the 'ins' DataFrame

selected_data = ins[['age', 'charges']]

# Creating the scatter plot
plt.scatter(selected_data['age'], selected_data['charges'])
plt.xlabel('Age')
plt.ylabel('Charges')
plt.title('Scatter plot of Age vs Charges')
plt.show()
```



Linear regression

```
# Creating smoker_num column
ins['smoker_num'] = np.where(ins['smoker'] == 'yes', 1, 0)

# Creating ins1 DataFrame
ins1 = ins[['age', 'bmi', 'children', 'smoker_num',
'charges']].copy()

# Converting children column to categorical
ins1['child_cat'] = pd.Categorical(ins1['children'])

# Creating ins2 DataFrame
ins2 = ins1.copy()

# Splitting data into training and testing sets
ins_training = ins2.sample(frac=0.85, random_state=42)
```

```
ins_testing = ins2.drop(ins_training.index)

# Creating and fitting linear regression models
m1 = LinearRegression()
m1.fit(ins_training.loc[ins_training['smoker_num'] ==
0].drop(['charges', 'smoker_num'], axis=1),
ins_training.loc[ins_training['smoker_num'] == 0, 'charges'])

m2 = LinearRegression()
m2.fit(ins_training.loc[ins_training['smoker_num'] ==
1].drop(['charges', 'smoker_num'], axis=1),
ins_training.loc[ins_training['smoker_num'] == 1, 'charges'])

# Predicting charges for testing data
ins_testing['pred_hybrid'] = np.where(ins_testing['smoker_num'] ==
0,
m1.intercept_ + m1.coef_[0] * ins_testing['age'] +
m1.coef_[1] * ins_testing['bmi'] +
m1.coef_[2] * ins_testing['children'],
m2.intercept_ + m2.coef_[0] * ins_testing['age'] +
m2.coef_[1] * ins_testing['bmi'] +
m2.coef_[2] * ins_testing['children'])

# Plotting charges vs. predicted charges
plt.plot(ins_testing['charges'], color='blue', label='Actual
Charges')
plt.plot(ins_testing['pred_hybrid'], color='red', label='Predicted
Charges')
plt.legend()
plt.show()

# Calculating RMSE
rmse = np.sqrt(mean_squared_error(ins_testing['charges'],
ins_testing['pred_hybrid']))
print('RMSE:', rmse)
```

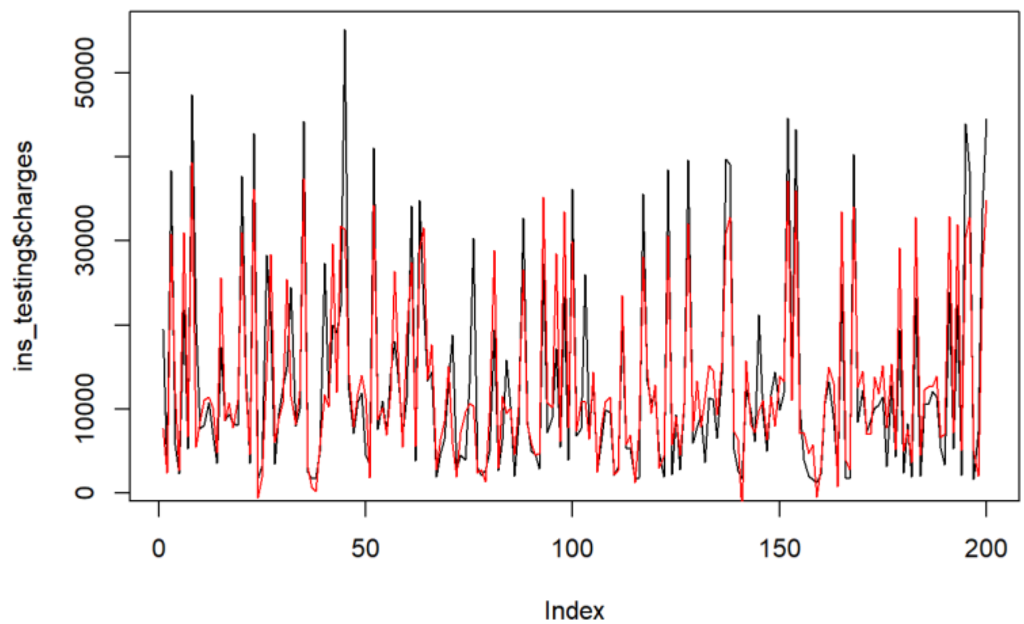


RMSE: 4757.521118649235

Linear regression with scaled data

```
# Creating ins2 DataFrame
ins2 = ins1.copy()

# Standardizing the bmi and charges columns
s_dev = np.std(ins2['charges'])
mean_val = np.mean(ins2['charges'])
ins2['bmi'] = (ins2['bmi'] - np.mean(ins2['bmi'])) /
np.std(ins2['bmi'])
ins2['charges'] = (ins2['charges'] - mean_val) / s_dev

# Splitting data into training and testing sets
ins_training = ins2.sample(frac=0.85, random_state=42)
ins_testing = ins2.drop(ins_training.index)

# Creating and fitting the linear regression model
linear_model = LinearRegression()
```

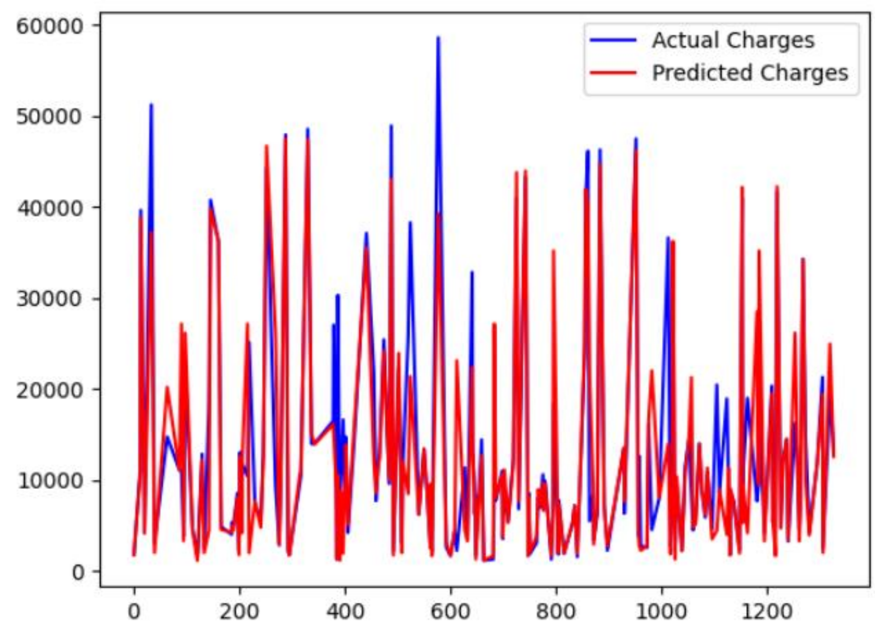
```
linear_model.fit(ins_training.drop('charges', axis=1),
ins_training['charges'])

# Predicting charges for testing data
ins_testing['pred_ins'] =
linear_model.predict(ins_testing.drop('charges', axis=1))

# Scaling charges and predicted charges back to original scale
ins_testing['charges'] = ins_testing['charges'] * s_dev + mean_val
ins_testing['pred_ins'] = ins_testing['pred_ins'] * s_dev +
mean_val

# Plotting charges vs. predicted charges
plt.plot(ins_testing['charges'], color='blue', label='Actual
Charges')
plt.plot(ins_testing['pred_ins'], color='red', label='Predicted
Charges')
plt.legend()
plt.show()

# Calculating RMSE
rmse = np.sqrt(mean_squared_error(ins_testing['charges'],
ins_testing['pred_ins']))
print('RMSE:', rmse)
```



Decision tree to identify the factor which is affecting the most.

```
# Creating ins2 DataFrame
ins2 = ins1.copy()

# Splitting data into training and testing sets
ins_training = ins2.sample(frac=0.85, random_state=42)
ins_testing = ins2.drop(ins_training.index)

# Creating and fitting the decision tree model
mod = DecisionTreeRegressor()
mod.fit(ins_training.drop('charges', axis=1),
ins_training['charges'])

# Predicting charges for testing data
ins_testing['pred'] = mod.predict(ins_testing.drop('charges',
axis=1))

# Plotting charges vs. predicted charges
```

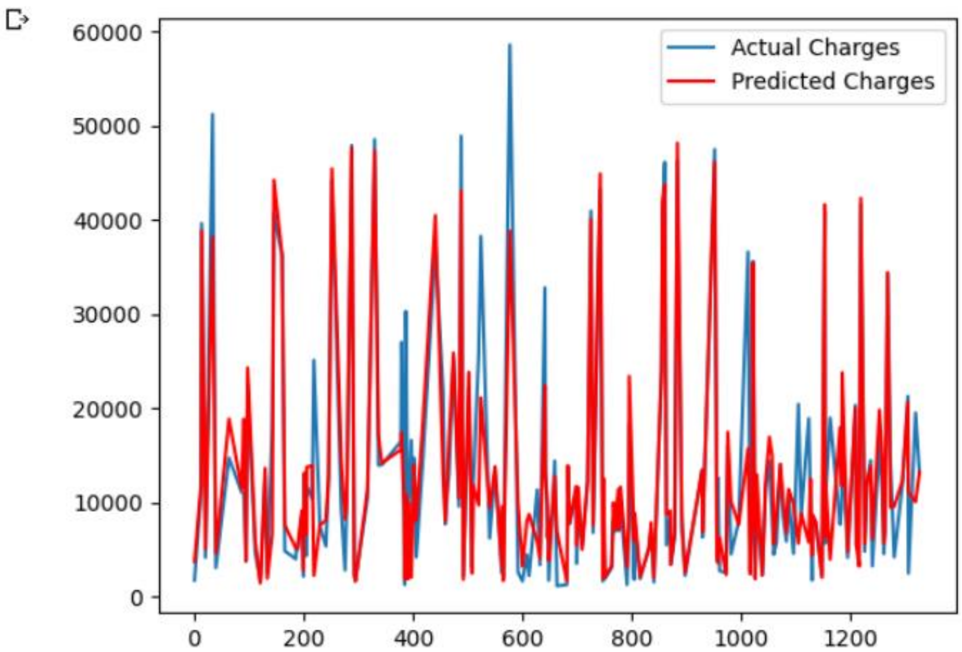
```
plt.plot(ins_testing['charges'], color='blue', label='Actual
Charges')
plt.plot(ins_testing['pred'], color='red', label='Predicted
Charges')
plt.legend()
plt.show()

# Calculating RMSE
rmse = np.sqrt(mean_squared_error(ins_testing['charges'],
ins_testing['pred']))
print('RMSE:', rmse)
```



RMSE: 6292.971637913359

Random forest

```
# Creating ins2 DataFrame
ins2 = ins1.copy()

# Splitting data into training and testing sets
ins_training = ins2.sample(frac=0.85, random_state=42)
ins_testing = ins2.drop(ins_training.index)

# Creating and fitting the random forest model
mod = RandomForestRegressor(n_estimators=500, max_features=3)
mod.fit(ins_training.drop('charges', axis=1),
ins_training['charges'])

# Predicting charges for testing data
ins_testing['pred'] = mod.predict(ins_testing.drop('charges',
axis=1))

# Plotting charges vs. predicted charges
plt.plot(ins_testing['charges'], label='Actual Charges')
plt.plot(ins_testing['pred'], color='red', label='Predicted
Charges')
plt.legend()
plt.show()

# Calculating RMSE
rmse = np.sqrt(mean_squared_error(ins_testing['charges'],
ins_testing['pred']))
print('RMSE:', rmse)
```




```
RMSE: 5041.686035250255
```

Reflection:

In the second part of my coursework, I made some adjustments to my hypotheses 1,2 & 3. I realized that I need to change hypothesis 2 where I provided smoker and non-smokers different in region and predicting that the area of maximum smokers and find which gender of smoker and non-smokers most in region which I aren't mentioned coursework - 1. Now I feel it is more understanding than me before coursework hypothesis 2.

Overall, I successfully addressed all the requirements stated in coursework two and implemented the necessary changes. I gained a deeper comprehension of the variables affecting medical costs as a result of investigating these hypotheses in the Medical Cost Personal Datasets. It emphasized the significance of elements including age, smoking history, and physiological indications in determining healthcare costs. These findings have consequences for those responsible for developing healthcare policies, insurance companies, and those who want to control and lessen the costs of receiving medical care.

Reference:

1. Choi, M. (2018). Medical Cost Personal Datasets. Kaggle. Retrieved from: https://www.kaggle.com/mirichoi0218/insurance.
2. Chen, S., Tung, Y., & Liao, J. (2020). Medical Cost Prediction Using Machine Learning Algorithms with Feature Selection Techniques. Applied Sciences, 10(7), 2387.
3. Pore, P., Mehta, A., & Shah, S. (2020). Predictive Analysis of Medical Insurance Costs using Machine Learning. International Journal of Advanced Science and Technology, 29(9), 4636-4644.
4. Priya, A. S., & Vasuki, A. (2021). Prediction of Medical Insurance Cost using Machine Learning Algorithms. International Journal of Advanced Research in Computer Science, 12(2), 104-109.
5. Chen, L., Zhang, Z., Xu, C., & Qian, H. (2019). Analysis of the Influencing Factors of Medical Insurance Cost Based on Machine Learning. Proceedings of the 9th International Conference on Social Science and Humanity, 249-252.

-Thank you -