#2018107115 고지영

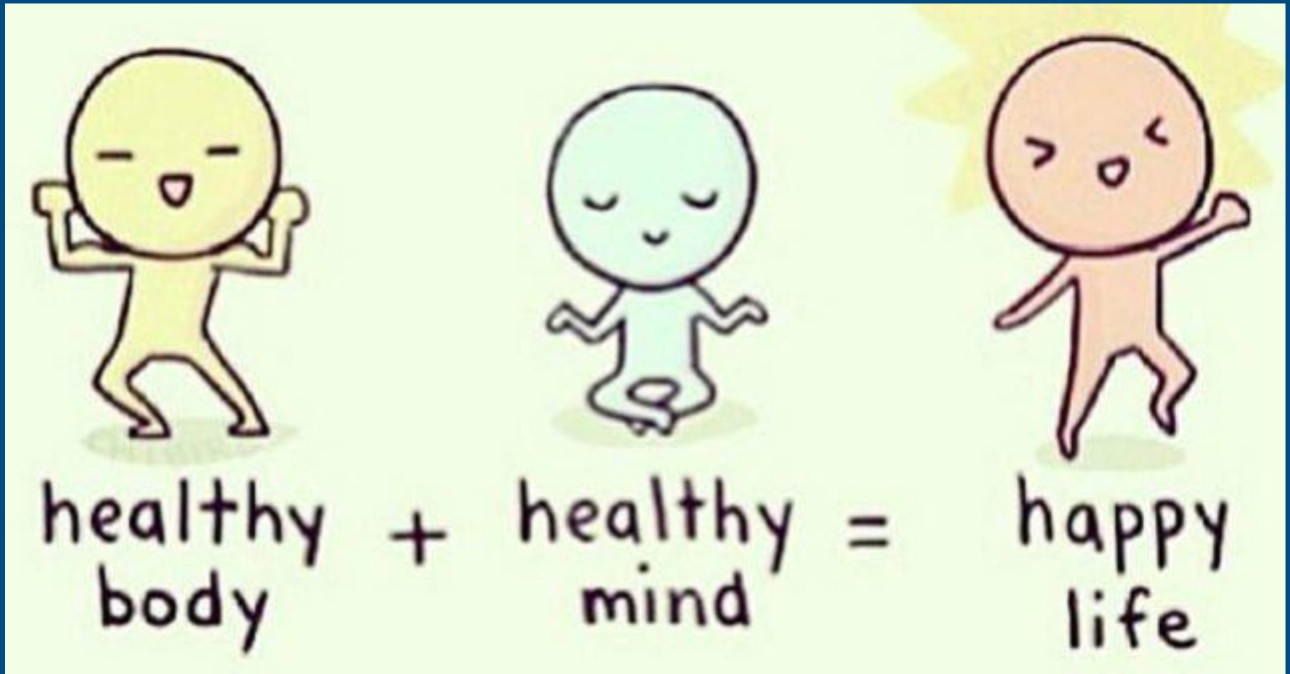
인공지능 발표

목차

A table of Contents

#1 주제

여러 요인에 따라 달라지는 치료비용 예측



#1 사용한 데이터

Data Tasks (2) Code (646) Discussion (12) Activity Metadata | Download (56 kB) | New Notebook

Usability 8.8 | License Database: Open Database, Contents: Database Contents | Tags education, health, finance, insurance, healthcare

Description

Context

Machine Learning with R by Brett Lantz is a book that provides an introduction to machine learning using R. As far as I can tell, Packt Publishing does not make its datasets available online unless you buy the book and create a user account which can be a problem if you are checking the book out from the library or borrowing the book from a friend. All of these datasets are in the public domain but simply needed some cleaning up and recoding to match the format in the book.

Content

Columns

- 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.
- charges: Individual medical costs billed by health insurance

Acknowledgements

The dataset is available on GitHub here.

Inspiration

Can you accurately predict insurance costs?

의료비 개인 데이터 세트
(선형회귀를 통한 보험료 예측)

컬럼

- 나이
- 성별(남,여)
- Bmi(체질량 지수 – 정상:18.5~24.9)
- 자녀(자녀의 수)
- 지역(미국)
- 흡연(흡연자,비흡연자)
- 비용(개인 의료비)

BMI(체질량지수) 계산

$$체중(kg) \div \left\{\frac{키(cm)}{100}\right\}^2$$

BMI	정상 여부
18.5 미만	저체중
18.5 – 24.9	정상
25 – 29.9	과체중
30 이상	비만

#2 데이터 읽어오기

```python
import numpy as np
import pandas as pd
import os
import matplotlib.pyplot as pl
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
data = pd.read_csv('../input/insurance.csv')
```

모듈 import
데이터 읽어 오기

```python
data.head(10)
```

	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
5	31	female	25.740	0	no	southeast	3756.62160
6	46	female	33.440	1	no	southeast	8240.58960
7	37	female	27.740	3	no	northwest	7281.50560
8	37	male	29.830	2	no	northeast	6406.41070
9	60	female	25.840	0	no	northwest	28923.13692

상위 10개 데이터

```python
from sklearn.preprocessing import LabelEncoder
#sex
le = LabelEncoder()
le.fit(data.sex.drop_duplicates())
data.sex = le.transform(data.sex)
# smoker or not
le.fit(data.smoker.drop_duplicates())
data.smoker = le.transform(data.smoker)
#region
le.fit(data.region.drop_duplicates())
data.region = le.transform(data.region)
```

+ Code + Markdown

```python
print(data.region[:10])
print(le.classes_)
```

```
0    3
1    2
2    2
3    1
4    1
5    2
6    2
7    1
8    0
9    1
Name: region, dtype: int64
['northeast' 'northwest' 'southeast' 'southwest']
```

```python
print(data.smoker[:10])
```

```
0    1  ← 흡연자
1    0
2    0
3    0
4    0
5    0  ← 비흡연자
6    0
7    0
8    0
9    0
Name: smoker, dtype: int64
```

```python
print(data.sex[:10])
```

```
0    0  ← 여성
1    1
2    1
3    1
4    1  ← 남성
5    0
6    0
7    0
8    1
9    0
Name: sex, dtype: int64
```

[labelEncoder] 을 사용하여
성별, 흡연여부,지역과 같은
데이터를 수치로 바꿔준다.

#3 읽어온 데이터 표시-히트맵

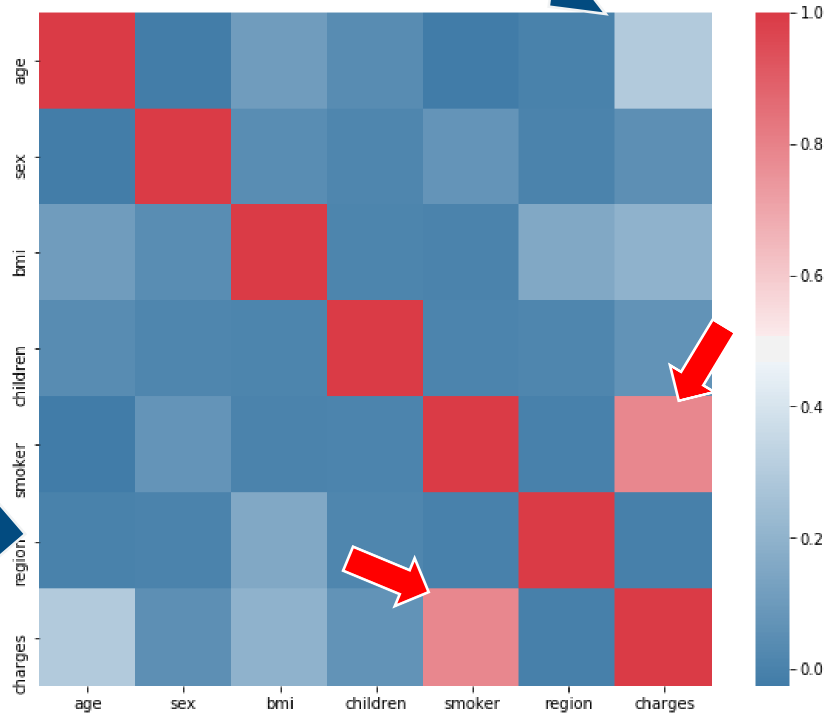
```
data.isnull().sum()
```

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

널값이 있는지 확인

```
f, ax = pl.subplots(figsize=(10, 8))
corr = data.corr()
sns.heatmap(corr, mask=np.zeros_like(corr, dtype=np.bool), cmap=sns.diverging_palette(240,10,as_cmap=True),
            square=True, ax=ax)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f23b
```



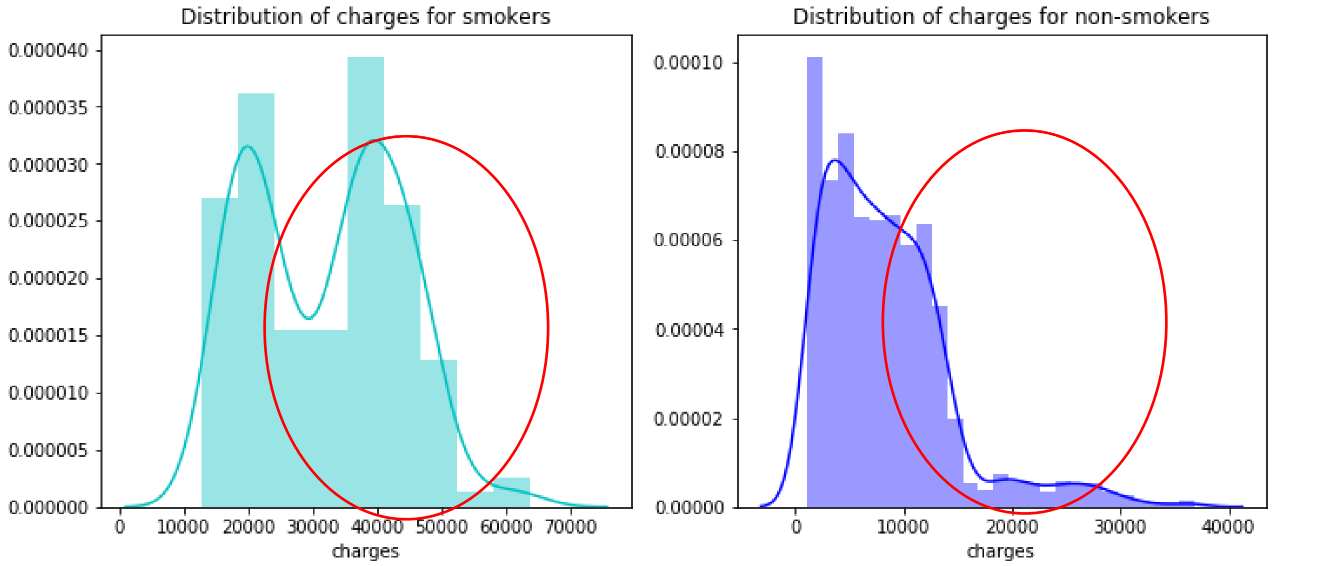
흡연자가 상관관계가
높은 것을 볼 수 있다.

반면 지역은 상관관계가
거의 없다

#3 읽어온 데이터 표시-히스토그램

```
from bokeh.io import output_notebook, show
from bokeh.plotting import figure
output_notebook()
import scipy.special
from bokeh.layouts import gridplot
from bokeh.plotting import figure, show, output_file

f= pl.figure(figsize=(12,5))

ax=f.add_subplot(121)
sns.distplot(data[(data.smoker == 1)]["charges"],color='c',ax=ax)
ax.set_title('Distribution of charges for smokers')

ax=f.add_subplot(122)
sns.distplot(data[(data.smoker == 0)]['charges'],color='b',ax=ax)
ax.set_title('Distribution of charges for non-smokers')
```

```
Text(0.5,1,'Distribution of charges for non-smokers')
```



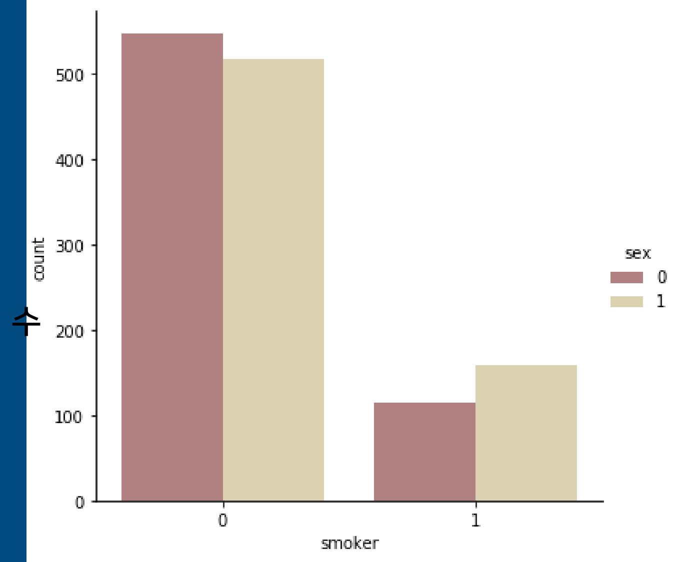
흡연자 vs 비흡연자

흡연자중에 의료비용이 높은 사람이 많다

#3 읽어온 데이터 표시-catplot , violin

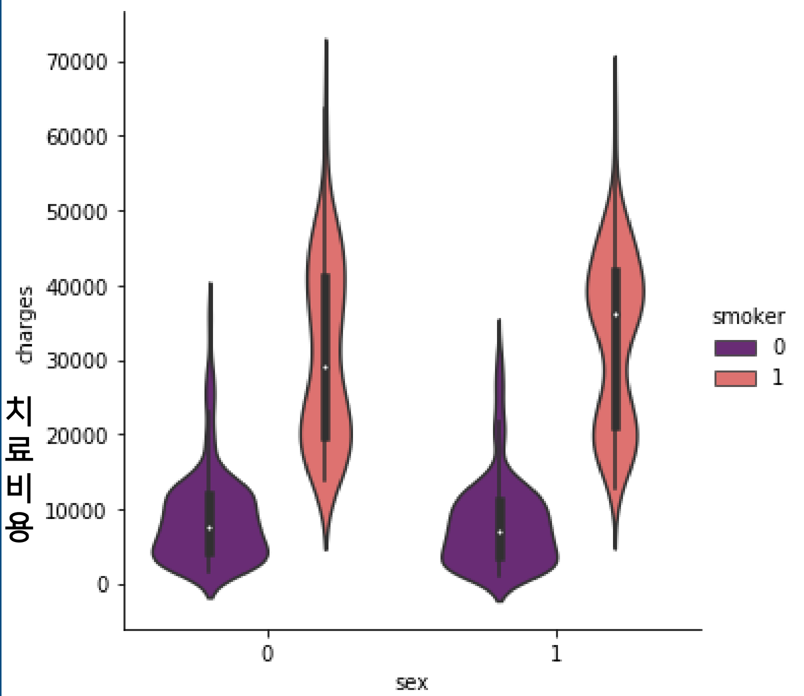
```
sns.catplot(x="smoker", kind="count",hue = 'sex', palette="pink", data=data)
```

<seaborn.axisgrid.FacetGrid at 0x7f23bd9246d8>



여성:0
남성:1

<seaborn.axisgrid.FacetGrid at 0x7f23b60da390>



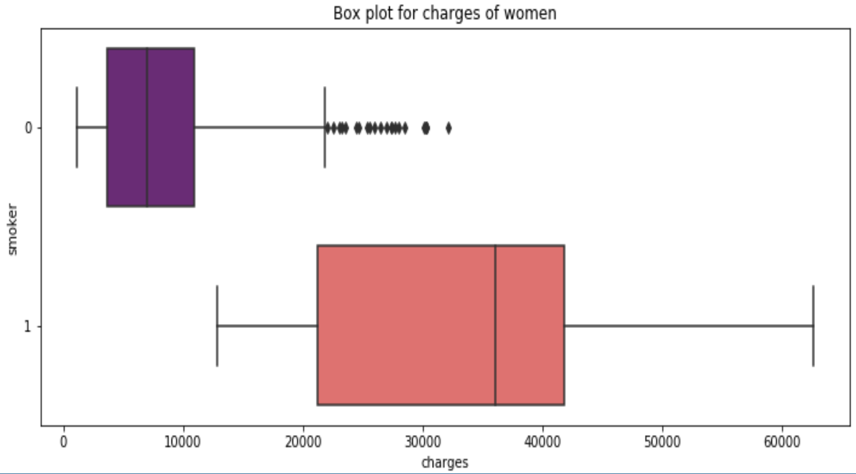
<성별에 따른 흡연자>
여성보다 남성흡연자의수가 많다

흡연자가 치료비용이 더 높다

#3 읽어온 데이터 표시-boxplot

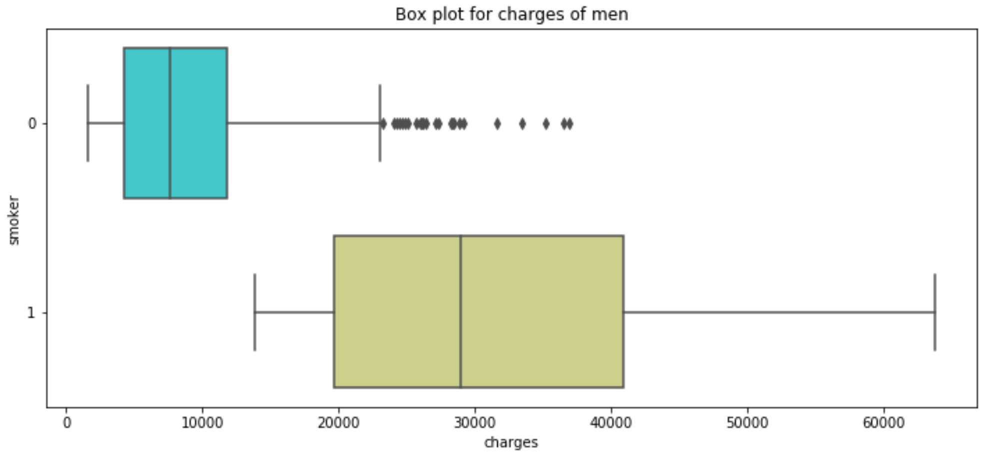
```
pl.figure(figsize=(12,5))
pl.title("Box plot for charges of women")
sns.boxplot(y="smoker", x="charges", data =  data[(data.sex == 1)] , orient="h", palette = 'magma')
```

<matplotlib.axes._subplots.AxesSubplot at 0x7f23b6069c18>




```
pl.figure(figsize=(12,5))
pl.title("Box plot for charges of men")
sns.boxplot(y="smoker", x="charges", data =  data[(data.sex == 0)] , orient="h", palette = 'rainbow')
```

<matplotlib.axes._subplots.AxesSubplot at 0x7f23b5ff10b8>



<여성>

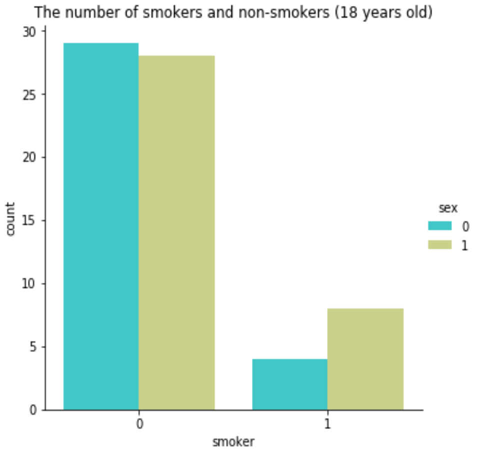
<남성>

둘 다 흡연자의 의료 비용의 평균이
더 높다

#3 18세의 데이터

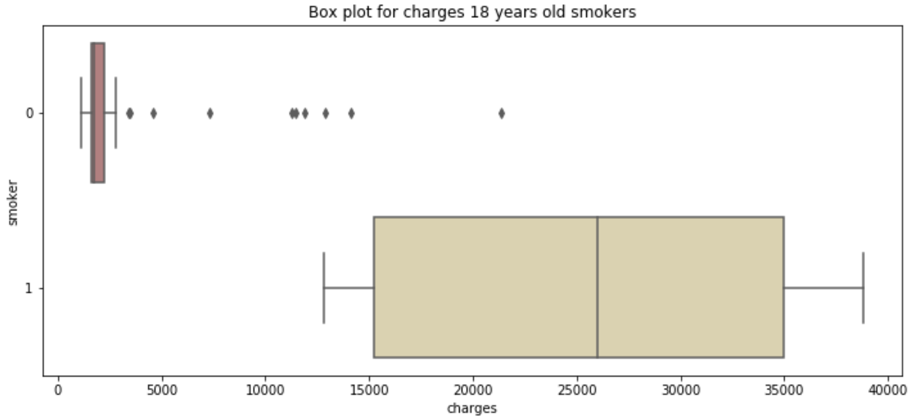
```python
sns.catplot(x="smoker", kind="count",hue = 'sex', palette="rainbow", data=data[(data.age == 18)])
pl.title("The number of smokers and non-smokers (18 years old)")
```

```
Text(0.5,1,'The number of smokers and non-smokers (18 years old)')
```




```python
pl.figure(figsize=(12,5))
pl.title("Box plot for charges 18 years old smokers")
sns.boxplot(y="smoker", x="charges", data = data[(data.age == 18)] , orient="h", palette = 'pink')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f23b5fd8080>
```



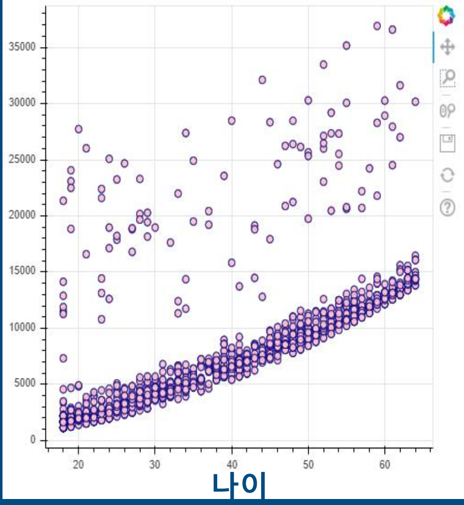
18살 흡연자의 비

⇨ 성인과 비슷

흡연자와 비흡연장의 의료비용 차이가 크다

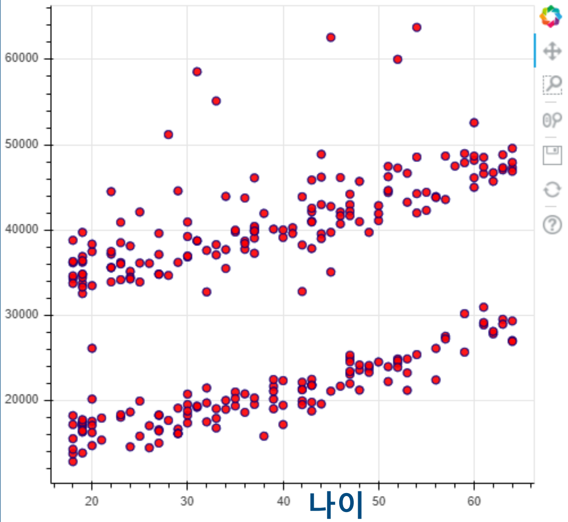
#3 비흡연자 vs 흡연자 – 나이에 따른 의료비용

```python
#non - smokers
p = figure(plot_width=500, plot_height=450)
p.circle(x=data[(data.smoker == 0)].age,y=data[(data.smoker == 0)].charges, size=7, line_color="navy", fill_color="pink", fill_alpha=0.9)

show(p)
```



<비흡연자>

```python
#smokers
p = figure(plot_width=500, plot_height=450)
p.circle(x=data[(data.smoker == 1)].age,y=data[(data.smoker == 1)].charges, size=7, line_color="navy", fill_color="red", fill_alpha=0.9)
show(p)
```



<흡연자>

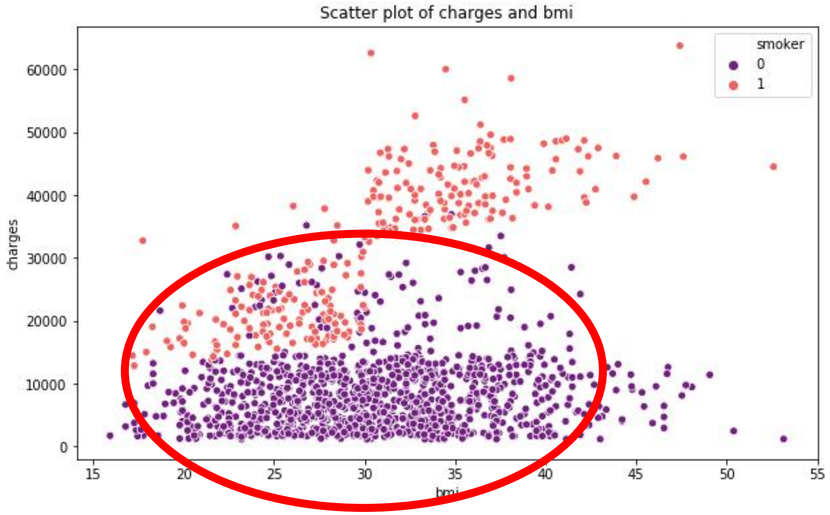
둘 다 나이에 따라 의료비용 증가

흡연자가 의료비용 더 높음

#3 Bmi(체질량 지수) 에 따른 의료 비용

```
pl.figure(figsize=(10,6))
ax = sns.scatterplot(x='bmi',y='charges',data=data,palette='magma',hue='smoker')
ax.set_title('Scatter plot of charges and bmi')

sns.lmplot(x="bmi", y="charges", hue="smoker", data=data, palette = 'magma', size = 8)
```

<seaborn.axisgrid.FacetGrid at 0x7f23b5844f98>



Bmi보다는 흡연여부가
더 큰 영향을 미침

#4 학습 및 테스트(LinearRegression)

```
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import PolynomialFeatures
from sklearn.metrics import r2_score,mean_squared_error
from sklearn.ensemble import RandomForestRegressor
```

```
x = data.drop(['charges'], axis = 1)      #문제부분
y = data.charges                            #답부분

x_train,x_test,y_train,y_test = train_test_split(x,y, random_state = 0)
lr = LinearRegression().fit(x_train,y_train)  #학습

y_train_pred = lr.predict(x_train)
y_test_pred = lr.predict(x_test)

print(lr.score(x_test,y_test))
```

```
0.7962732059725786
```
#점수

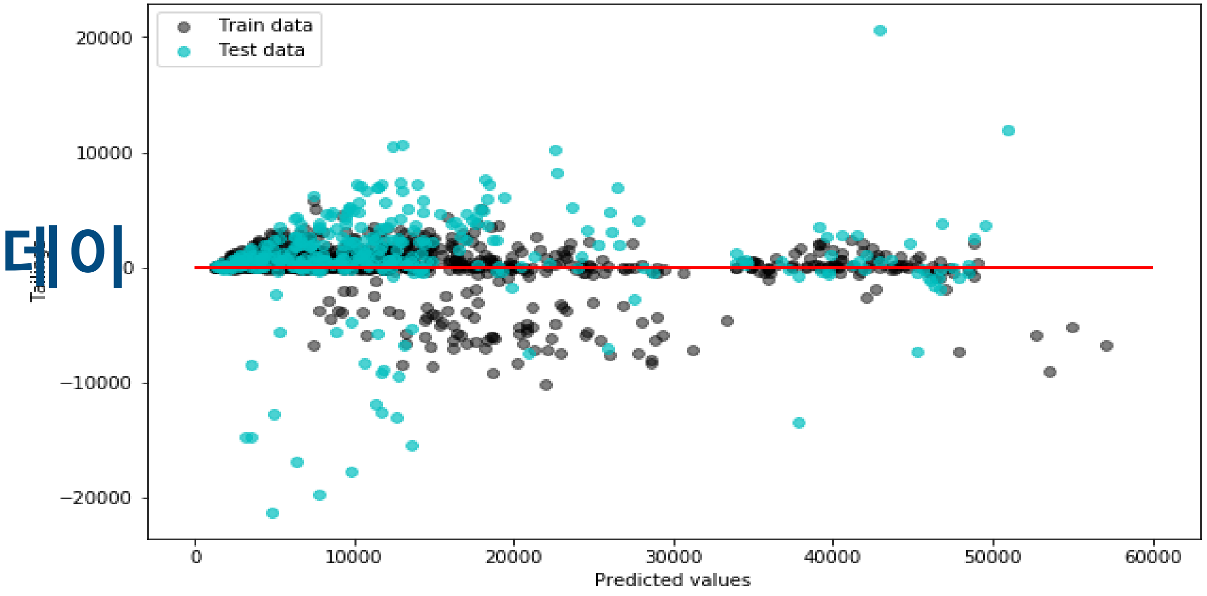
#4 학습 및 테스트(RandomForestRegressor)

```
RF = RandomForestRegressor(random_state = 0)
RF.fit(x_train,y_train)                     #학습

score = RF.score(x_train,y_train)
print('Score:', format(score,'.3f'))


pred = RF.predict(x_test)
print('Predicted:', pred)
print('Correct answer:\n', y_test)
```

```
Score: 0.964
```

#점수

```
pl.figure(figsize=(10,6))

pl.scatter(forest_train_pred,forest_train_pred - y_train,
           c = 'black', marker = 'o', s = 35, alpha = 0.5,
           label = 'Train data')
pl.scatter(forest_test_pred,forest_test_pred - y_test,
           c = 'c', marker = 'o', s = 35, alpha = 0.7,
           label = 'Test data')
pl.xlabel('Predicted values')
pl.ylabel('Tailings')
pl.legend(loc = 'upper left')
pl.hlines(y = 0, xmin = 0, xmax = 60000, lw = 2, color = 'red')
pl.show()
```



<결과>학습용데이터와 테스트데이터가 거의 비슷하다

#4 느낀점

- 이번 발표를 준비하면서 머신러닝 학습시키는데 필요한 데이터와 그렇지 않은 데이터를 구분하는 과정을 알게 되었고 이 주제를 찾기까지 다양한 주제들이 있었는데 머신러닝으로 정말 다양한 것을 할 수 있다는 것을 알게 되었습니다.
- 알고리즘마다 정확도가 다르다는 것을 알게 되었고 데이터를 직접 시각화 하는 과정에서 그래프가 어떤 의미를 갖는지에 대해서 생각해보고 컬럼 간에 어떤 관계가 있는지도 알 수 있게 되었습니다.

감사합니다.