다중회귀분석(Multivariate Regression)

```
import warnings
warnings.filterwarnings('ignore')
```

실습용 데이터 설정

- pandas DataFrame
  - Insurance.csv

```
import pandas as pd

url = 'https://raw.githubusercontent.com/rusita-ai/pyData/master/Insurance.csv'
DF = pd.read_csv(url)

DF.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   expenses  1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

```
DF.head(3)
```

	age	sex	bmi	children	smoker	region	expenses
0	19	female	27.90	0	yes	southwest	16884.9240
1	18	male	33.77	1	no	southeast	1725.5523
2	28	male	33.00	3	no	southeast	4449.4620

I. 탐색적 데이터 분석

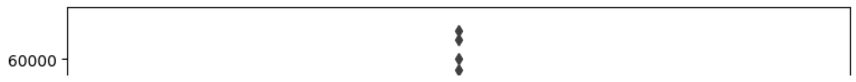
- 시각화 패키지

```
import matplotlib.pyplot as plt
import seaborn as sns
```

1) 전체 의료비 분포

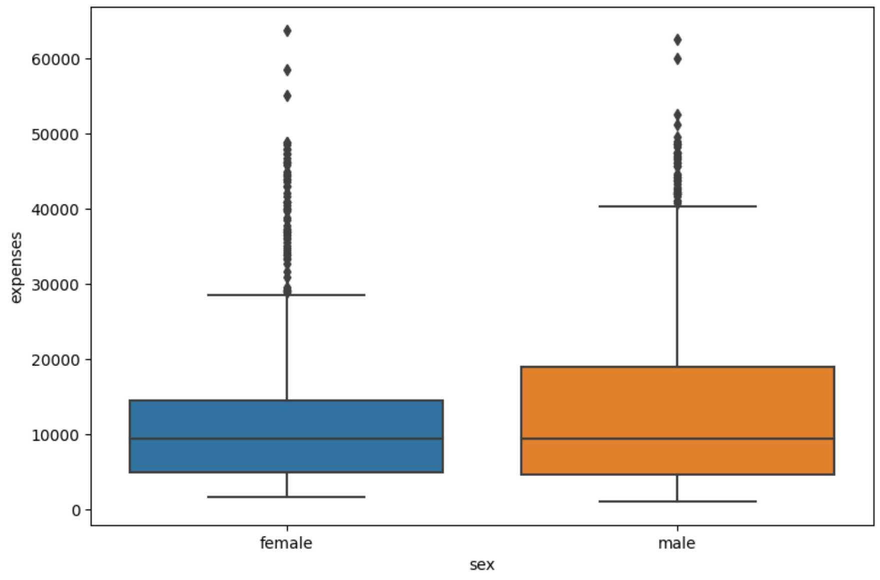
```
plt.figure(figsize = (9, 6))
sns.distplot(DF.expenses,
             hist = True,
             kde = True)
plt.show()
```

```
plt.figure(figsize = (9, 6))
sns.boxplot(y = 'expenses', data = DF)
plt.show()
```



2) 성별 별 의료비 분포

```
plt.figure(figsize = (9, 6))
sns.boxplot(x = 'sex', y = 'expenses', data = DF)
plt.show()
```



```
DF.sex.value_counts()
```

```
male      676
female    662
Name: sex, dtype: int64
```

3) 자녀수 별 의료비 분포

```
plt.figure(figsize = (9, 6))
sns.boxplot(x = 'children', y = 'expenses', data = DF)
plt.show()
```

```
DF.children.value_counts()
```

```
0    574
1    324
2    240
3    157
4     25
5     18
Name: children, dtype: int64
```

4) 흡연여부 별 의료비 분포

```
plt.figure(figsize = (9, 6))
sns.boxplot(x = 'smoker', y = 'expenses', data = DF)
plt.show()
```

```
DF.smoker.value_counts()
```

```
no      1064
yes      274
Name: smoker, dtype: int64
```

5) 거주지역 별 의료비 분포

```
plt.figure(figsize = (9, 6))
sns.boxplot(x = 'region', y = 'expenses', data = DF)
plt.show()
```

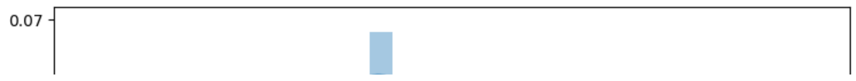
```
DF.region.value_counts()
```

```
southeast    364
southwest    325
northwest    325
northeast    324
Name: region, dtype: int64
```

6) BMI 분포 및 의료비와의 관계

- BMI 분포

```
plt.figure(figsize = (9, 6))
sns.distplot(DF.bmi,
             hist = True,
             kde = True)
plt.show()
```



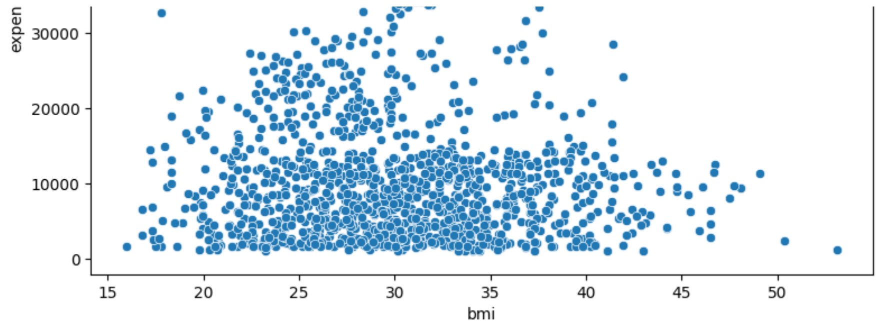
- BMI와 의료비 간의 관계

```
plt.figure(figsize = (9, 6))
sns.scatterplot(x = DF.bmi, y = DF.expenses)
plt.show()
```

###

End Of Document

###


expen
bmi