4장 파이썬 업그레이드(시각화, 데이터처리)

시각화

In [0]:

```
import numpy as np
import matplotlib.pyplot as plt
```

In [21]:

```
x = np.linspace(0,14,100)   # 0부터 14까지 동일한 간격의 값 100개 만들기
x
```

Out[21]:

```
array([ 0.        ,  0.14141414,  0.28282828,  0.42424242,  0.56565657,
        0.70707071,  0.84848485,  0.98989899,  1.13131313,  1.27272727,
        1.41414141,  1.55555556,  1.6969697 ,  1.83838384,  1.97979798,
        2.12121212,  2.26262626,  2.4040404 ,  2.54545455,  2.68686869,
        2.82828283,  2.96969697,  3.11111111,  3.25252525,  3.39393939,
        3.53535354,  3.67676768,  3.81818182,  3.95959596,  4.1010101 ,
        4.24242424,  4.38383838,  4.52525253,  4.66666667,  4.80808081,
        4.94949495,  5.09090909,  5.23232323,  5.37373737,  5.51515152,
        5.65656566,  5.7979798 ,  5.93939394,  6.08080808,  6.22222222,
        6.36363636,  6.50505051,  6.64646465,  6.78787879,  6.92929293,
        7.07070707,  7.21212121,  7.35353535,  7.49494949,  7.63636364,
        7.77777778,  7.91919192,  8.06060606,  8.2020202 ,  8.34343434,
        8.48484848,  8.62626263,  8.76767677,  8.90909091,  9.05050505,
        9.19191919,  9.33333333,  9.47474747,  9.61616162,  9.75757576,
        9.8989899 , 10.04040404, 10.18181818, 10.32323232, 10.46464646,
       10.60606061, 10.74747475, 10.88888889, 11.03030303, 11.17171717,
       11.31313131, 11.45454545, 11.5959596 , 11.73737374, 11.87878788,
       12.02020202, 12.16161616, 12.3030303 , 12.44444444, 12.58585859,
       12.72727273, 12.86868687, 13.01010101, 13.15151515, 13.29292929,
       13.43434343, 13.57575758, 13.71717172, 13.85858586, 14.        ])
```

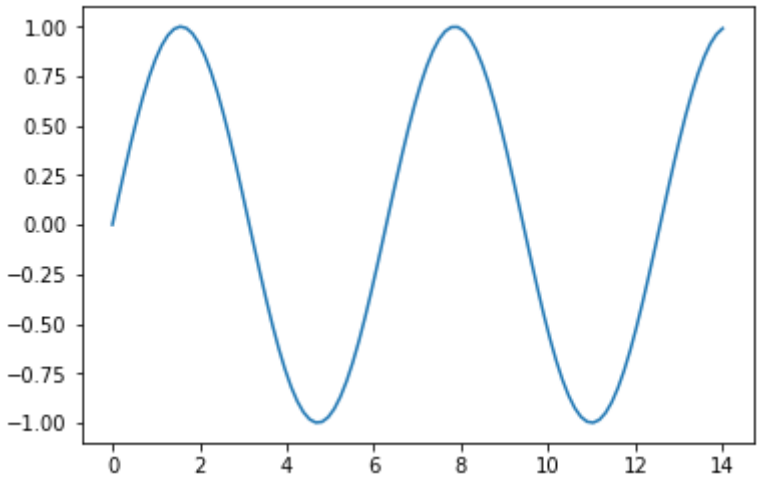
In [22]:

```
# x의 값을 이용하여 y값 (sin) 구하기
y = np.sin(x)

# 그래프 그리기
plt.plot(x,y)
```

Out[22]:

```
[<matplotlib.lines.Line2D at 0x7fc2d3faec18>]
```



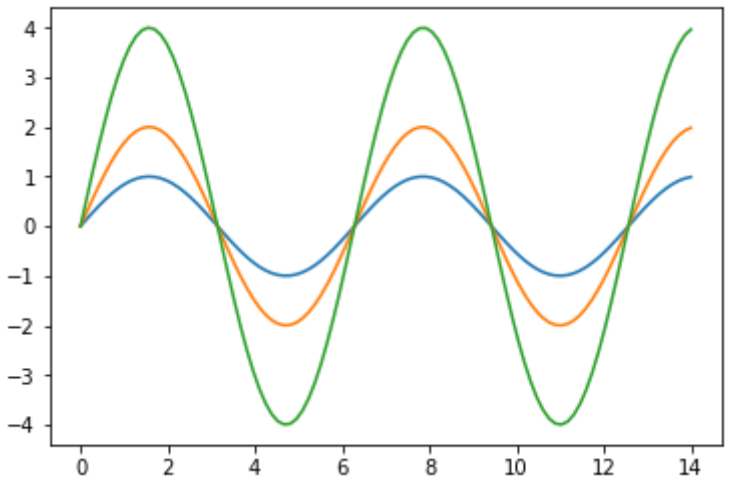
In [23]:

```
# 여러개의 그래프 그려보기
y1 = np.sin(x)
y2 = np.sin(x) * 2
y3 = np.sin(x) * 4

# 그래프 그리기
plt.plot(x,y1, x, y2, x, y3)
```

Out[23]:

```
[<matplotlib.lines.Line2D at 0x7fc2d3f194a8>,
 <matplotlib.lines.Line2D at 0x7fc2d3f19630>,
 <matplotlib.lines.Line2D at 0x7fc2d3f199b0>]
```

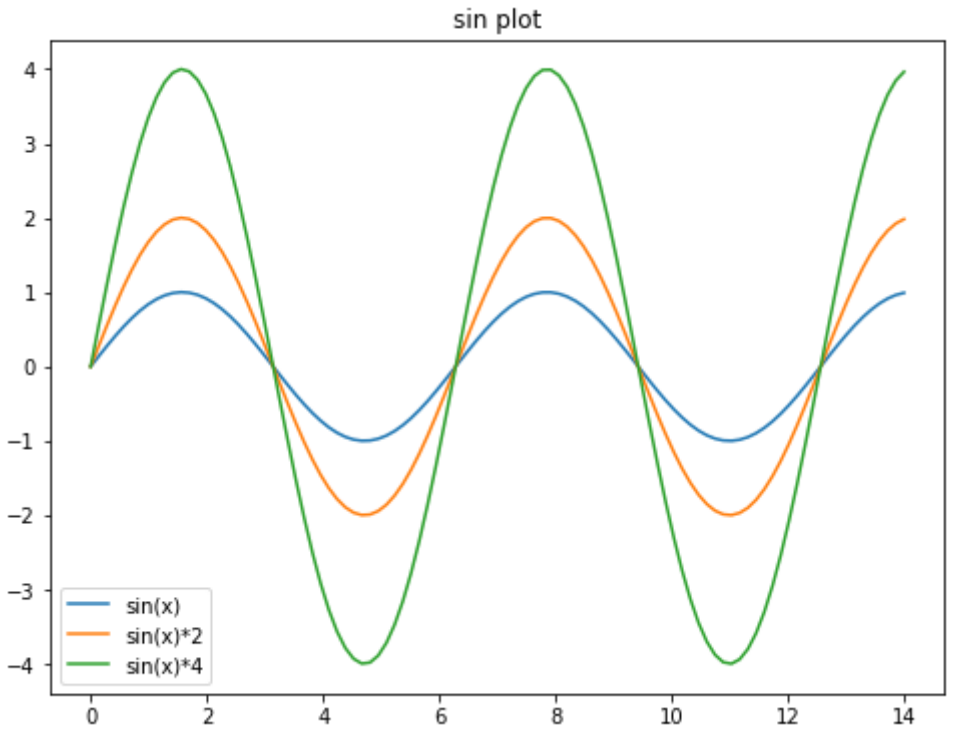


In [29]:

```
## 그래프에 범례 표시해보기
plt.figure(figsize=(8,6))
plt.plot(x, y1, label="sin(x)")
plt.plot(x, y2, label="sin(x)*2")
plt.plot(x, y3, label="sin(x)*4")
plt.legend()
plt.title("sin plot")
```

Out[29]:

```
Text(0.5, 1.0, 'sin plot')
```



In [0]:

```
import seaborn as sns
```

In [31]:

```
tips = sns.load_dataset("tips")
tips
```

Out[31]:

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4
5	25.29	4.71	Male	No	Sun	Dinner	4
6	8.77	2.00	Male	No	Sun	Dinner	2
7	26.88	3.12	Male	No	Sun	Dinner	4
8	15.04	1.96	Male	No	Sun	Dinner	2
9	14.78	3.23	Male	No	Sun	Dinner	2
10	10.27	1.71	Male	No	Sun	Dinner	2
11	35.26	5.00	Female	No	Sun	Dinner	4
12	15.42	1.57	Male	No	Sun	Dinner	2
13	18.43	3.00	Male	No	Sun	Dinner	4
14	14.83	3.02	Female	No	Sun	Dinner	2
15	21.58	3.92	Male	No	Sun	Dinner	2
16	10.33	1.67	Female	No	Sun	Dinner	3
17	16.29	3.71	Male	No	Sun	Dinner	3
18	16.97	3.50	Female	No	Sun	Dinner	3
19	20.65	3.35	Male	No	Sat	Dinner	3
20	17.92	4.08	Male	No	Sat	Dinner	2
21	20.29	2.75	Female	No	Sat	Dinner	2
22	15.77	2.23	Female	No	Sat	Dinner	2
23	39.42	7.58	Male	No	Sat	Dinner	4
24	19.82	3.18	Male	No	Sat	Dinner	2
25	17.81	2.34	Male	No	Sat	Dinner	4
26	13.37	2.00	Male	No	Sat	Dinner	2
27	12.69	2.00	Male	No	Sat	Dinner	2
28	21.70	4.30	Male	No	Sat	Dinner	2
29	19.65	3.00	Female	No	Sat	Dinner	2
...	...	...	...	...	...	...	...
214	28.17	6.50	Female	Yes	Sat	Dinner	3
215	12.90	1.10	Female	Yes	Sat	Dinner	2
216	28.15	3.00	Male	Yes	Sat	Dinner	5

	total_bill	tip	sex	smoker	day	time	size
217	11.59	1.50	Male	Yes	Sat	Dinner	2
218	7.74	1.44	Male	Yes	Sat	Dinner	2
219	30.14	3.09	Female	Yes	Sat	Dinner	4
220	12.16	2.20	Male	Yes	Fri	Lunch	2
221	13.42	3.48	Female	Yes	Fri	Lunch	2
222	8.58	1.92	Male	Yes	Fri	Lunch	1
223	15.98	3.00	Female	No	Fri	Lunch	3
224	13.42	1.58	Male	Yes	Fri	Lunch	2
225	16.27	2.50	Female	Yes	Fri	Lunch	2
226	10.09	2.00	Female	Yes	Fri	Lunch	2
227	20.45	3.00	Male	No	Sat	Dinner	4
228	13.28	2.72	Male	No	Sat	Dinner	2
229	22.12	2.88	Female	Yes	Sat	Dinner	2
230	24.01	2.00	Male	Yes	Sat	Dinner	4
231	15.69	3.00	Male	Yes	Sat	Dinner	3
232	11.61	3.39	Male	No	Sat	Dinner	2
233	10.77	1.47	Male	No	Sat	Dinner	2
234	15.53	3.00	Male	Yes	Sat	Dinner	2
235	10.07	1.25	Male	No	Sat	Dinner	2
236	12.60	1.00	Male	Yes	Sat	Dinner	2
237	32.83	1.17	Male	Yes	Sat	Dinner	2
238	35.83	4.67	Female	No	Sat	Dinner	3
239	29.03	5.92	Male	No	Sat	Dinner	3
240	27.18	2.00	Female	Yes	Sat	Dinner	2
241	22.67	2.00	Male	Yes	Sat	Dinner	2
242	17.82	1.75	Male	No	Sat	Dinner	2
243	18.78	3.00	Female	No	Thur	Dinner	2

244 rows × 7 columns

In [32]:

```
## 앞의 데이터만 살펴보기
tips.head()
```

Out[32]:

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

In [33]:

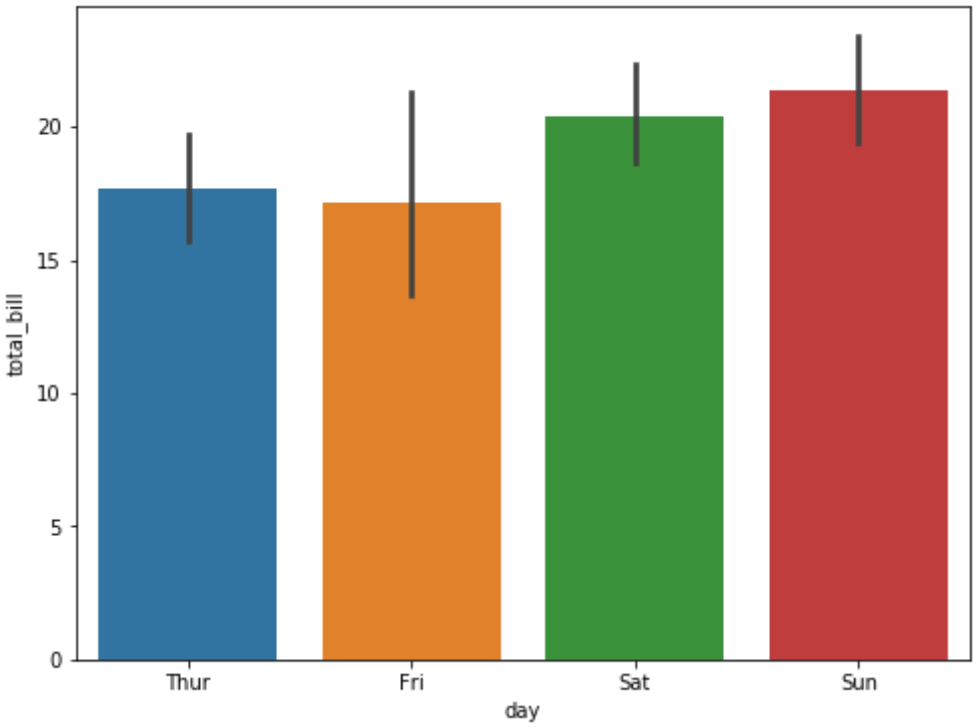
```
## 뒤의 데이터 살펴보기 - tail()
tips.tail()
```

Out[33]:

	total_bill	tip	sex	smoker	day	time	size
239	29.03	5.92	Male	No	Sat	Dinner	3
240	27.18	2.00	Female	Yes	Sat	Dinner	2
241	22.67	2.00	Male	Yes	Sat	Dinner	2
242	17.82	1.75	Male	No	Sat	Dinner	2
243	18.78	3.00	Female	No	Thur	Dinner	2

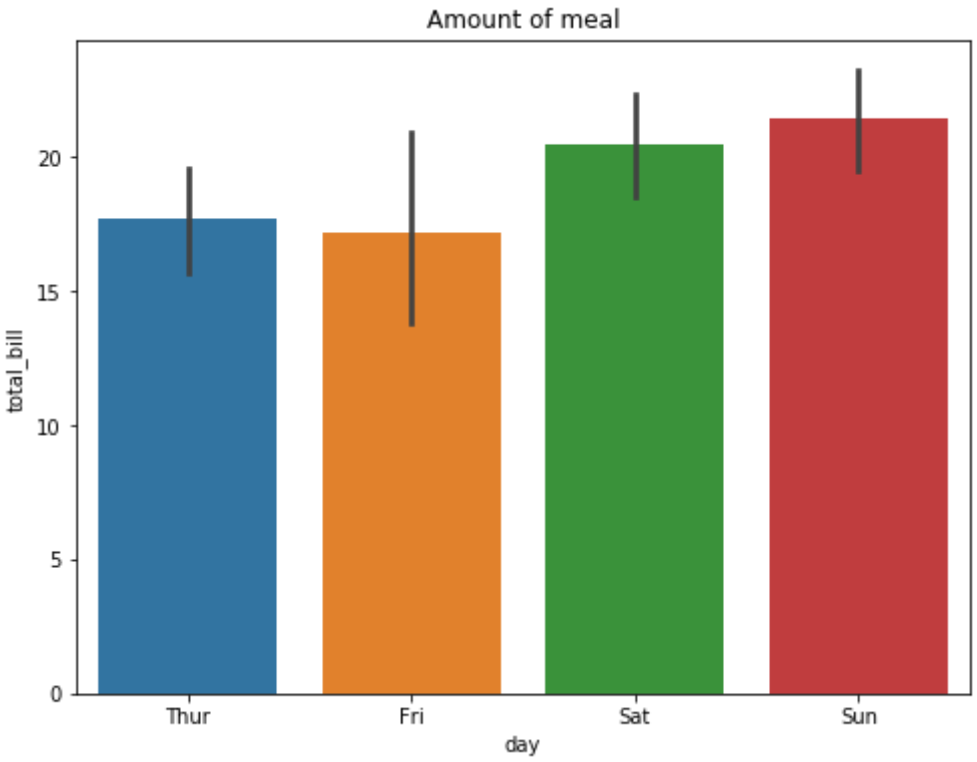
In [35]:

```
### 요일별 식사 금액은 얼마나 될까?
plt.figure(figsize=(8,6))
sns.barplot(x="day", y="total_bill", data=tips)
plt.show()
```



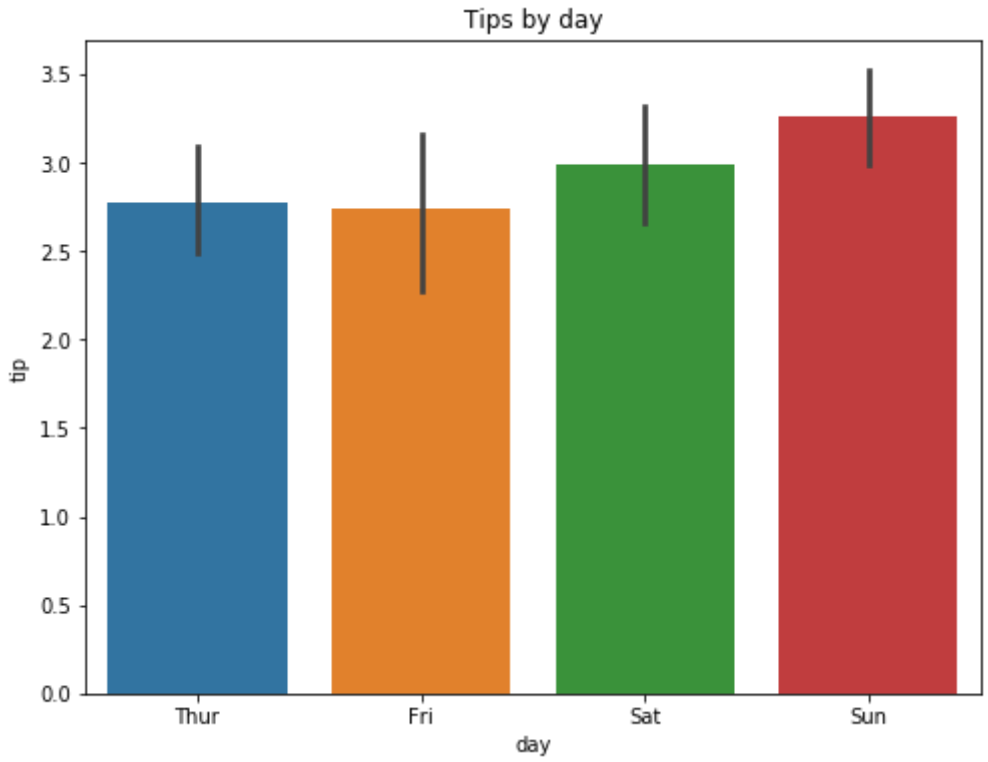
In [36]:

```
### 요일별 식사 금액은 얼마나 될까?
plt.figure(figsize=(8,6))
sns.barplot(x="day", y="total_bill", data=tips)
plt.title("Amount of meal")
plt.show()
```



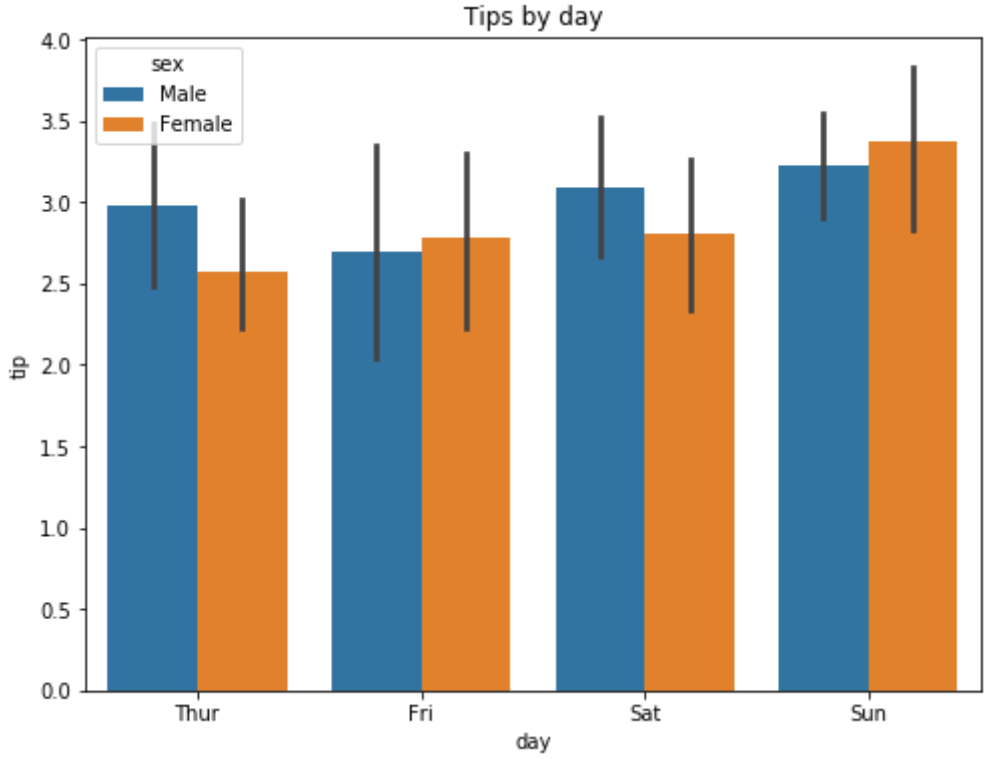
In [38]:

```
### (직접 해보기) 그렇다면 요일별 Tips은 얼마나 될까?
plt.figure(figsize=(8,6))
sns.barplot(x="day", y="tip", data=tips)
plt.title("Tips by day")
plt.show()
```



In [39]:

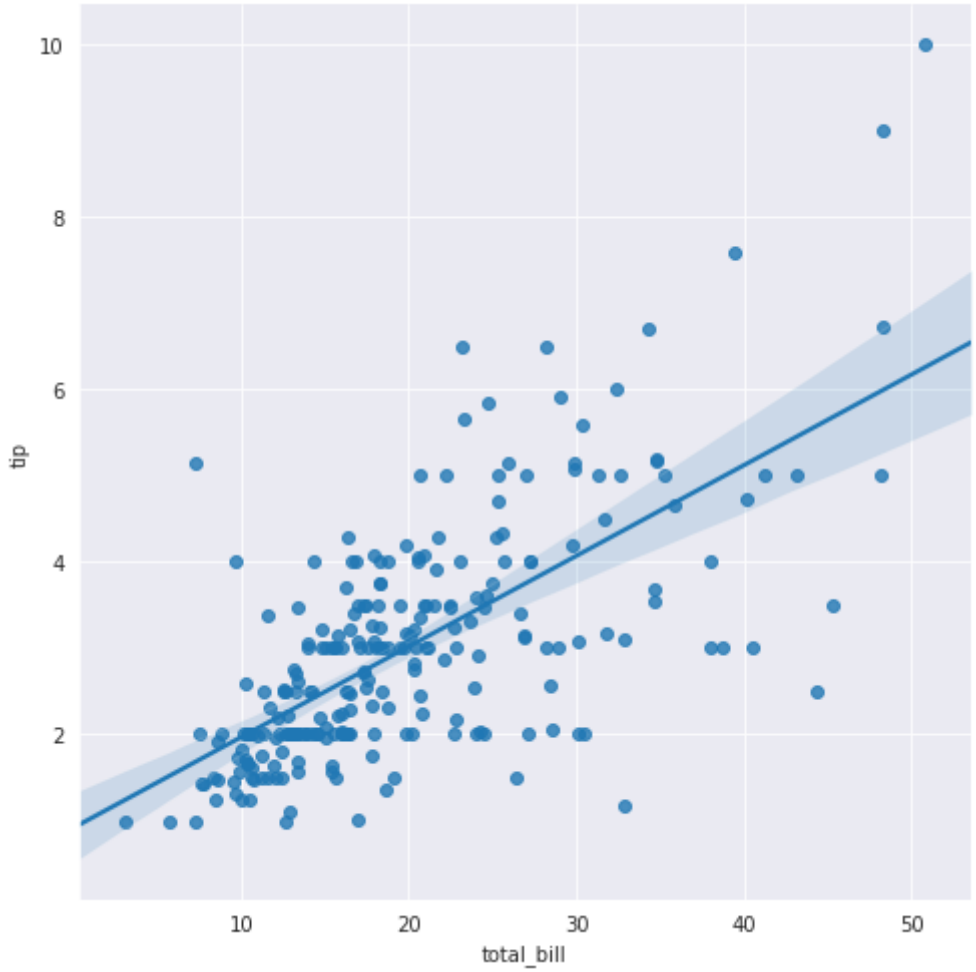
```
# 요일별 tip은 남성과 여성은 어떠할까?
plt.figure(figsize=(8,6))
sns.barplot(x="day", y="tip", hue="sex", data=tips)
plt.title("Tips by day")
plt.show()
```



In [43]:

```
# 산점도와 lmplot이용해보기
sns.set_style("darkgrid")   # 종류 : darkgrid, whitegrid, dark, white, ticks
sns.lmplot(x="total_bill", y="tip", data=tips, size=7)
plt.show()
```

```
/usr/local/lib/python3.6/dist-packages/seaborn/regression.py:546: UserWarning: The `
size` paramter has been renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```



In [44]:

```
tips.columns
```

Out[44]:

```
Index(['total_bill', 'tip', 'sex', 'smoker', 'day', 'time', 'size'], dtype='object')
```

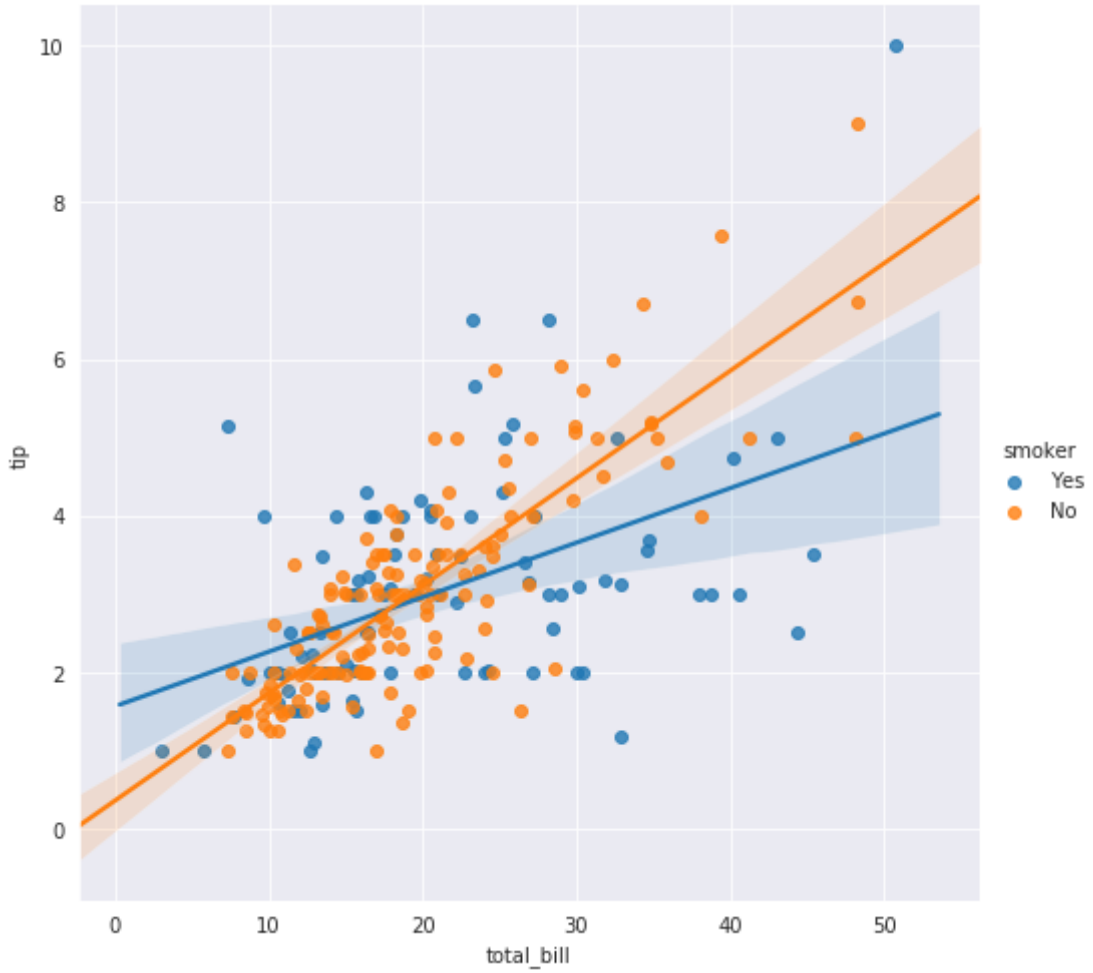
In [53]:

```
sns.lmplot(x="total_bill", y="tip", hue='smoker', data=tips, size=7)
```

```
/usr/local/lib/python3.6/dist-packages/seaborn/regression.py:546: UserWarning: The `
size` paramter has been renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```

Out[53]:

```
<seaborn.axisgrid.FacetGrid at 0x7fc2c890d400>
```



03 항공 데이터 시각화

In [54]:

```
fg = sns.load_dataset("flights")
fg.head()
```

Out[54]:

	year	month	passengers
0	1949	January	112
1	1949	February	118
2	1949	March	132
3	1949	April	129
4	1949	May	121

In [64]:

```
print(fg.shape)   # 데이터의 행과 열
print(fg.columns) # 데이터 제목열 이름
print(fg.describe() ) # 데이터 요약
```

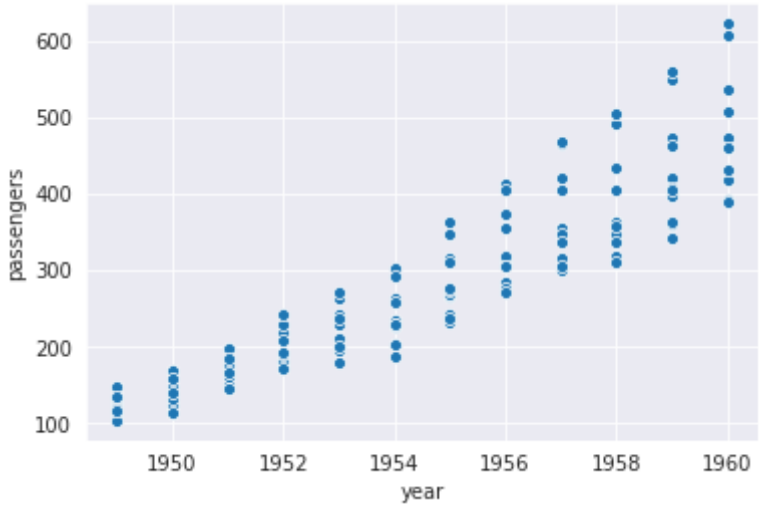
```
(144, 3)
Index(['year', 'month', 'passengers'], dtype='object')
              year  passengers
count   144.000000  144.000000
mean   1954.500000  280.298611
std       3.464102  119.966317
min    1949.000000  104.000000
25%    1951.750000  180.000000
50%    1954.500000  265.500000
75%    1957.250000  360.500000
max    1960.000000  622.000000
```

In [56]:

```
sns.scatterplot(x="year", y="passengers", data=fg)
```

Out[56]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fc2c883af28>
```

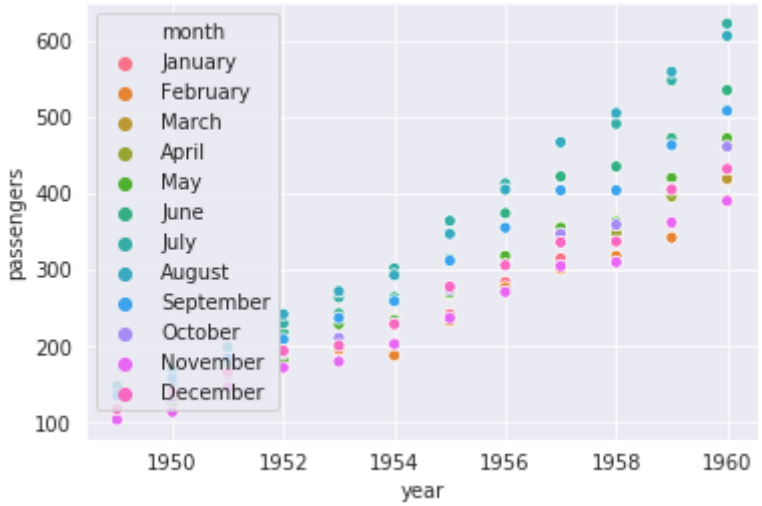


In [59]:

```
sns.scatterplot(x="year", y="passengers", hue="month", data=fg)
```

Out[59]:

<matplotlib.axes._subplots.AxesSubplot at 0x7fc2c879f128>

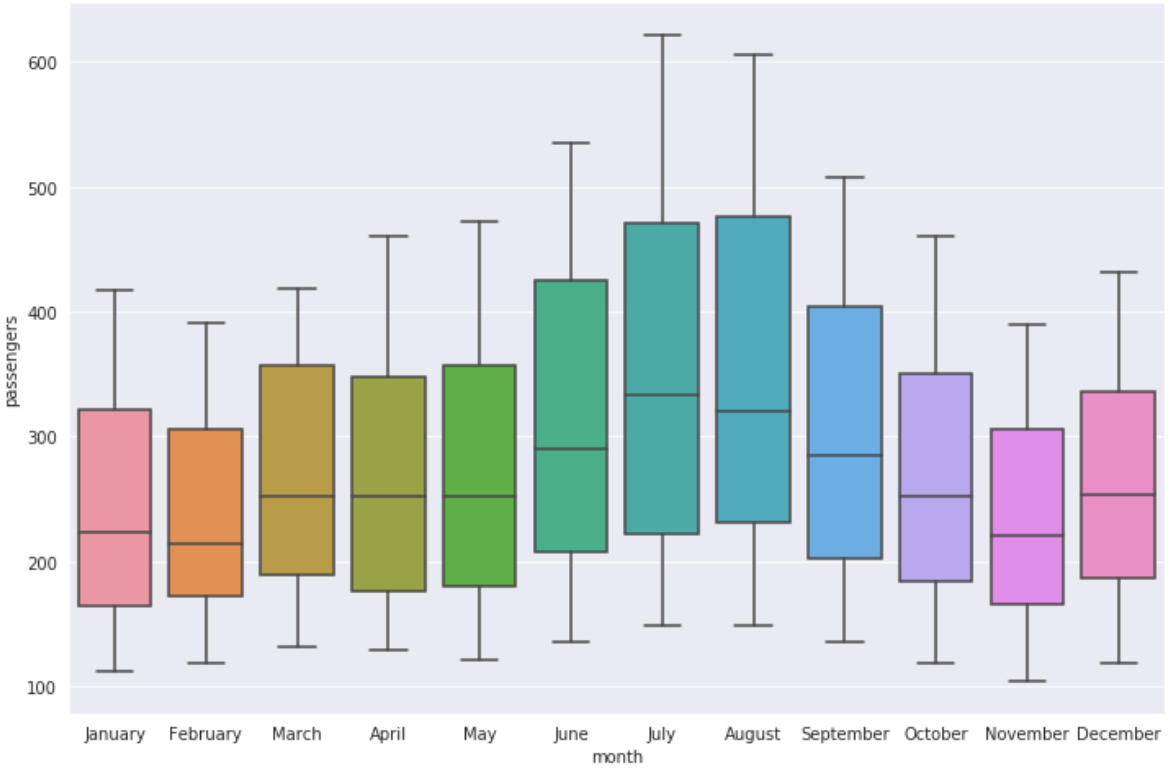


In [63]:

```
plt.figure(figsize=(12,8))
sns.boxplot(x="month", y="passengers", data=fg)
```

Out[63]:

<matplotlib.axes._subplots.AxesSubplot at 0x7fc2c8d05470>



데이터를 요약한 형태로 볼 수 없을까?

In [65]:

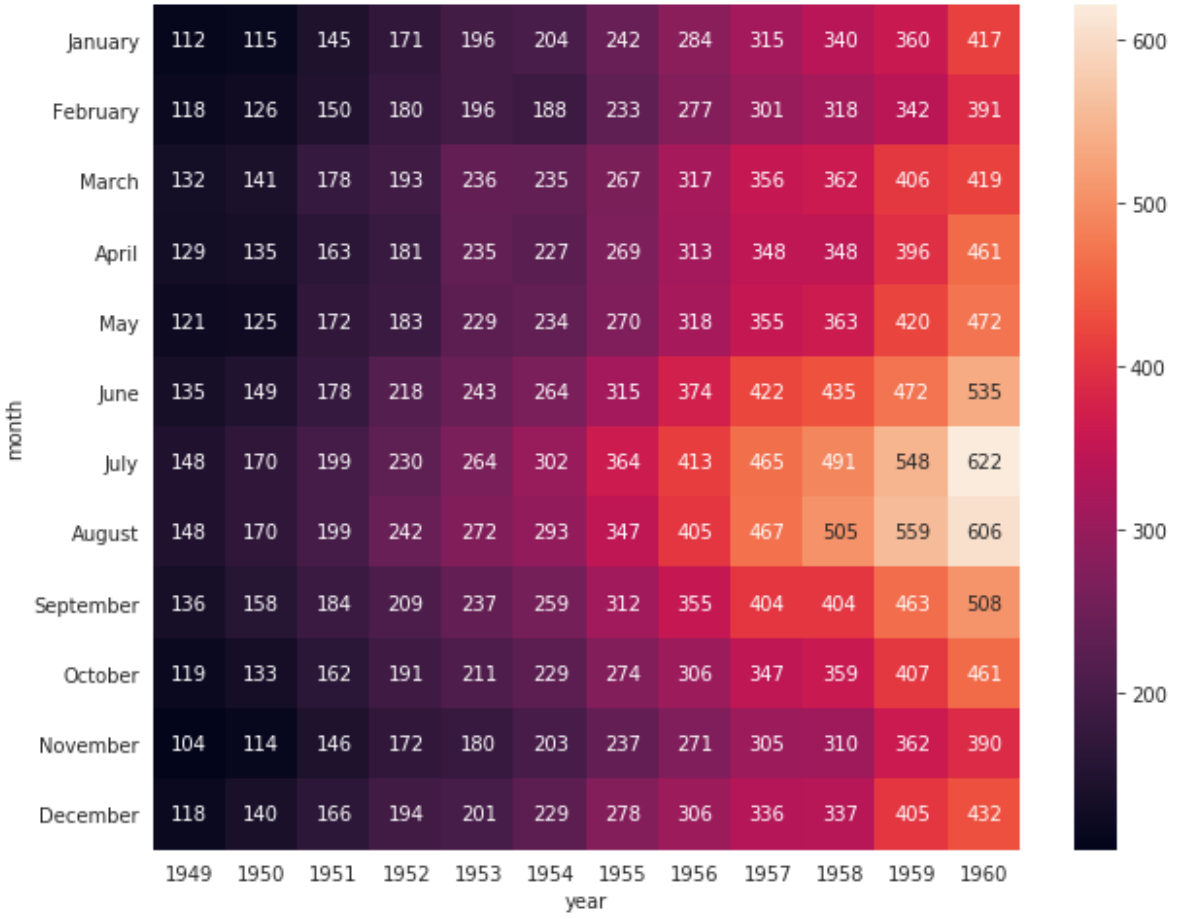
```
fgpivot = fg.pivot("month", "year", "passengers")
fgpivot
```

Out[65]:

year month	1949	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960
January	112	115	145	171	196	204	242	284	315	340	360	417
February	118	126	150	180	196	188	233	277	301	318	342	391
March	132	141	178	193	236	235	267	317	356	362	406	419
April	129	135	163	181	235	227	269	313	348	348	396	461
May	121	125	172	183	229	234	270	318	355	363	420	472
June	135	149	178	218	243	264	315	374	422	435	472	535
July	148	170	199	230	264	302	364	413	465	491	548	622
August	148	170	199	242	272	293	347	405	467	505	559	606
September	136	158	184	209	237	259	312	355	404	404	463	508
October	119	133	162	191	211	229	274	306	347	359	407	461
November	104	114	146	172	180	203	237	271	305	310	362	390
December	118	140	166	194	201	229	278	306	336	337	405	432

In [66]:

```
plt.figure(figsize=(10,8))
sns.heatmap(fgpivot, annot=True, fmt="d")
plt.show()
```



In [69]:

```
iris = sns.load_dataset("iris")
iris.head()
```

Out[69]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

In [0]:

```
## csv 데이터 셋 만들기
iris.to_csv("iris.csv", index=False)

## xlsx 파일 만들기
iris.to_excel("iris.xlsx", index=False)
```

In [0]:

```
import pandas as pd
```

In [71]:

```
iris_csv = pd.read_csv("iris.csv")
iris_csv.head()
```

Out[71]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

In [72]:

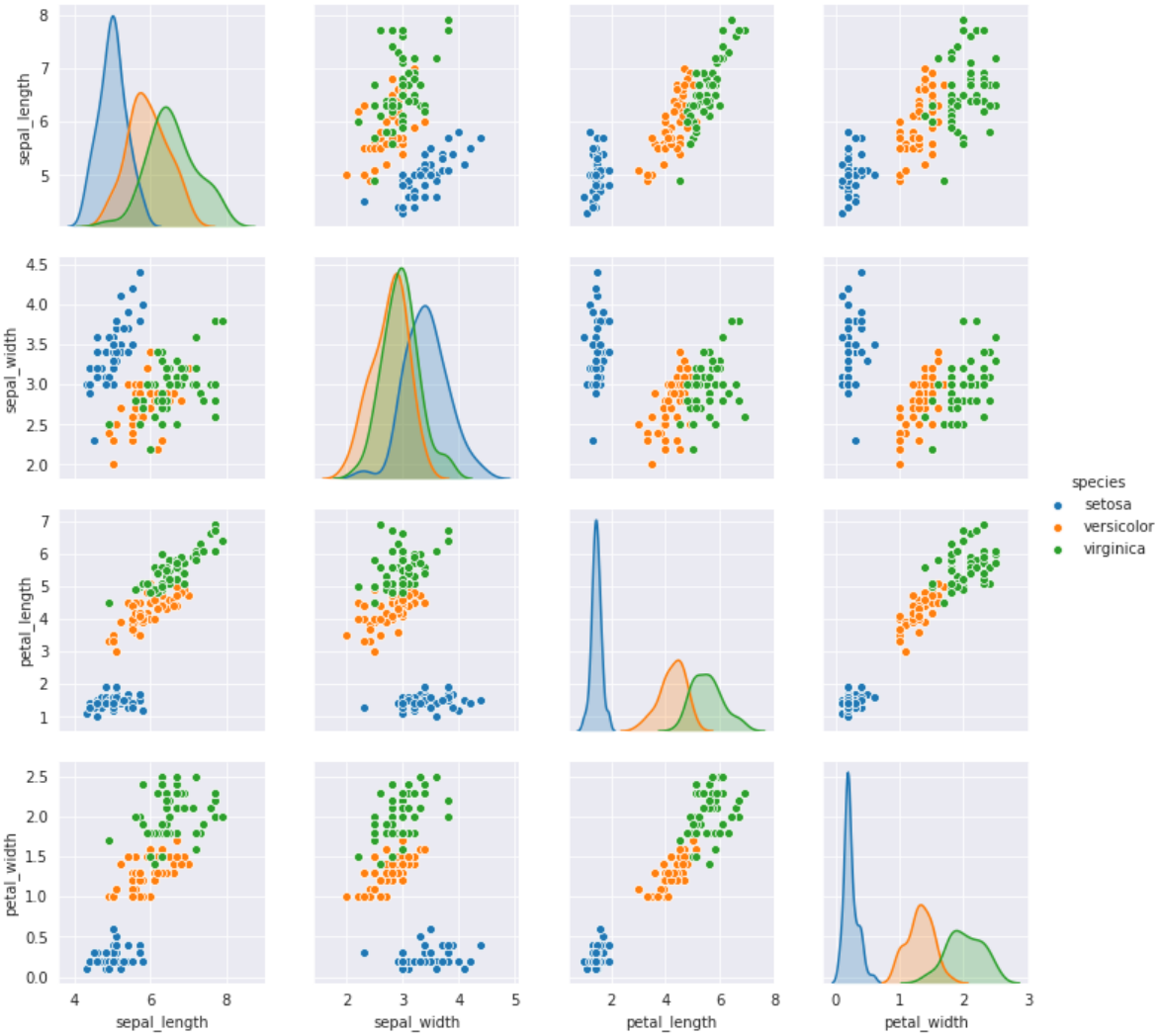
```
iris_excel = pd.read_excel("iris.xlsx")
iris_excel.head()
```

Out[72]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

In [73]:

```
sns.pairplot(iris_excel, hue="species")
plt.show()
```



In [0]:

```
```