Data Cleansing & Visualization

[1] **Pandas**를 활용한 Data Cleansing & Visualization

[2] **Seaborn**을 활용한 Visualization

[3] **Matplotlib**을 활용한 **위젯&애니메이션** 구현

이승한
7/11(목)

1. Pandas를 활용한 Data Cleansing & Visualization

1. **Data Cleansing**

   (column.row indexing , dropping , adding)

   (groupby , sorting , merge/concat , NA imputation)

   (method chaining, lambda, map/apply/applymap, stack/unstack)

2. **Visualization**

   (bar / hist / kde / box / scatter / hexbin / pie)

코드 통해서

Jupyter notebook

(data cleansing & visualization) Pandas

필요한 패키지 : numpy, pandas, matplotlib, seaborn

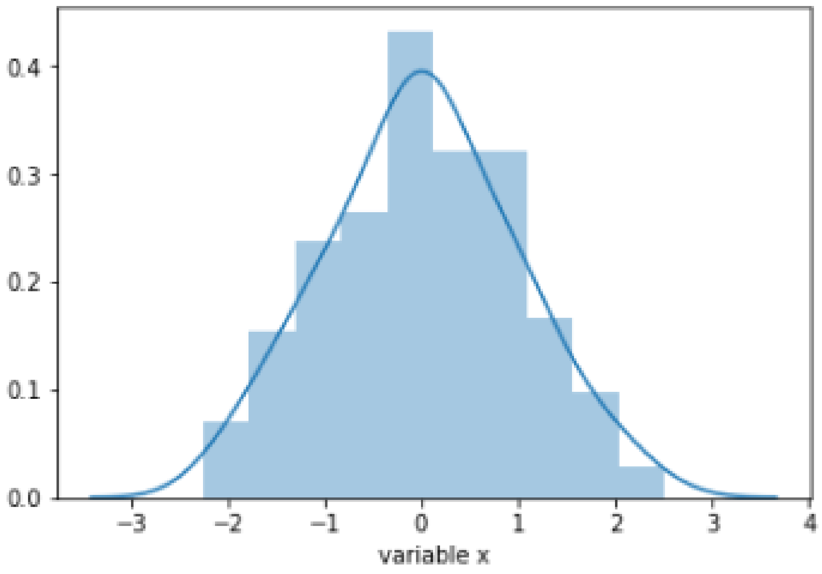
2. Seaborn을 활용한 Visualization

1. **Dist plot**
2. **Box plot**
3. **Strip plot**
4. **Reg plot**
5. **Violin plot**
6. **Heatmap**
7. **Facet grid**
8. **Kde plot**
9. **Pair plot / Pair grid**
10. **Joint plot**

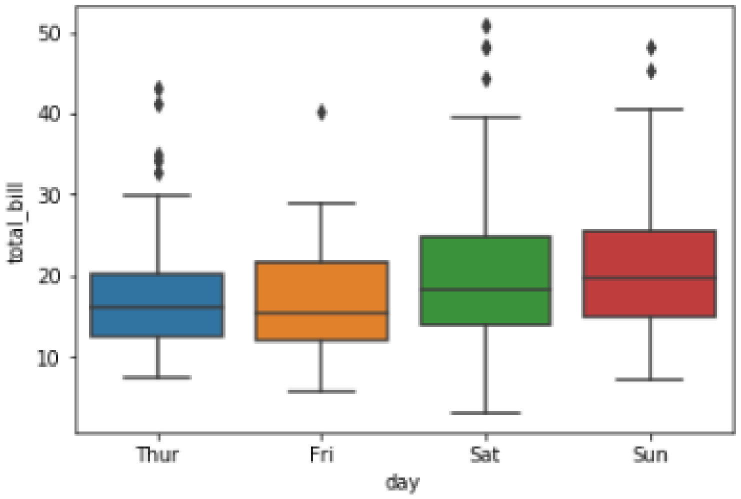
1. **sns.distplot**

변수의 ‘단계’or’정도’에 따라 다른 그래프!

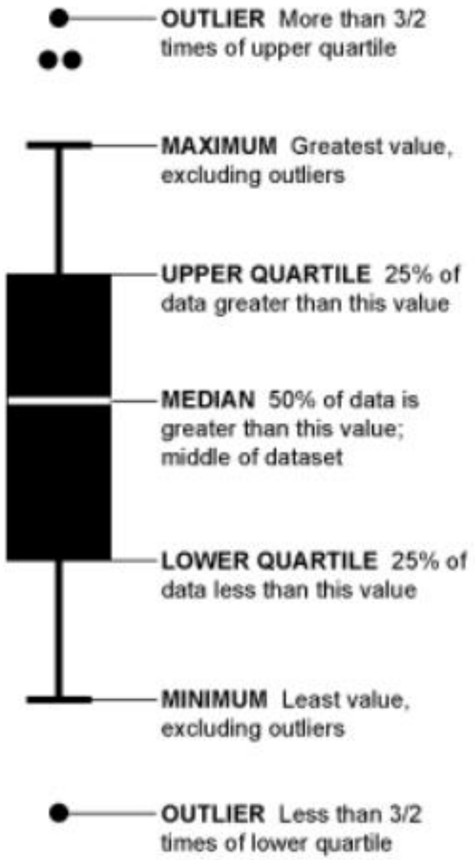
2. **sns.boxplot** (상자수염 그림)



(sns.distplot)

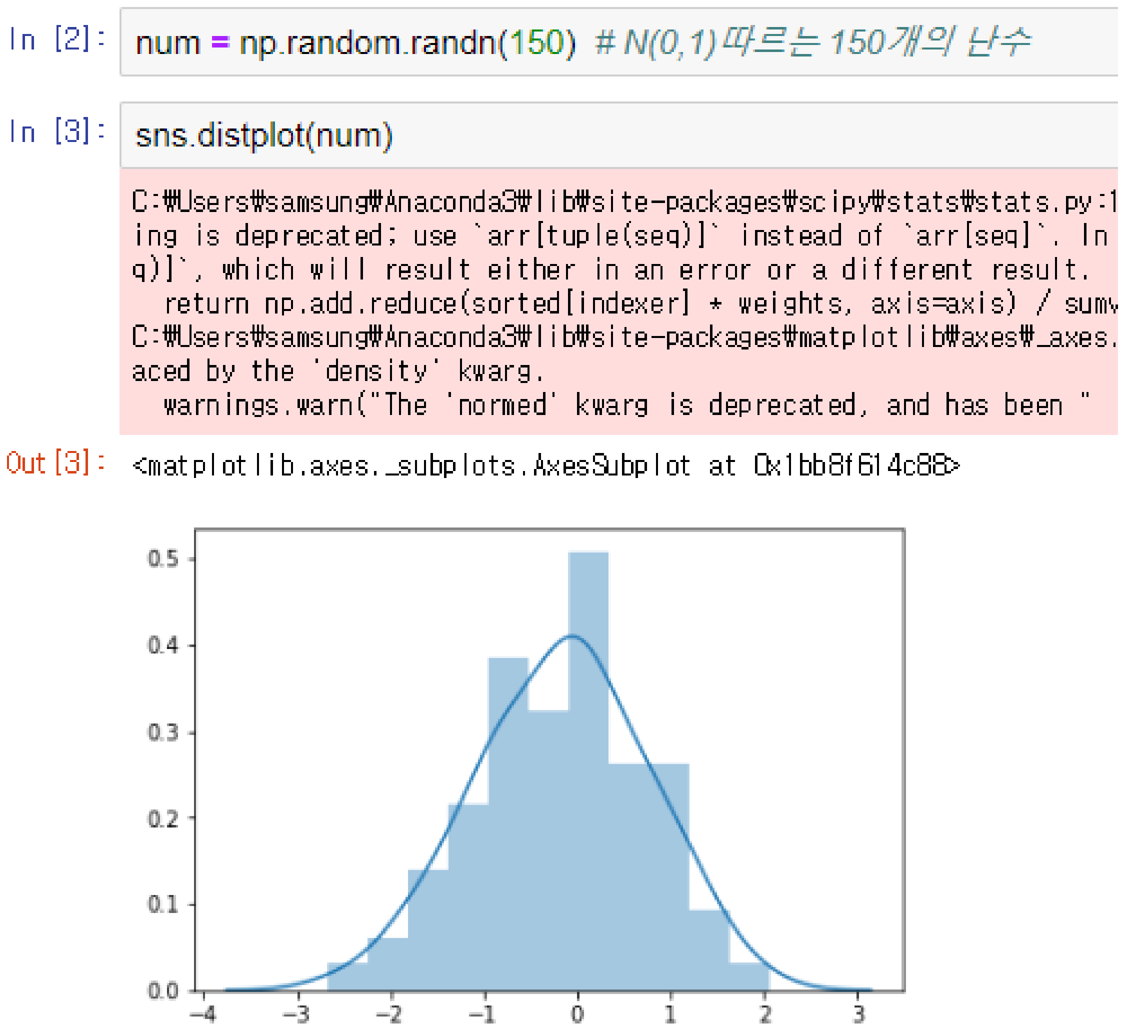


(sns.boxplot)



OUTLIER More than 3/2 times of upper quartile
MAXIMUM Greatest value, excluding outliers
UPPER QUARTILE 25% of data greater than this value
MEDIAN 50% of data is greater than this value; middle of dataset
LOWER QUARTILE 25% of data less than this value
MINIMUM Least value, excluding outliers
OUTLIER Less than 3/2 times of lower quartile

1. sns.distplot

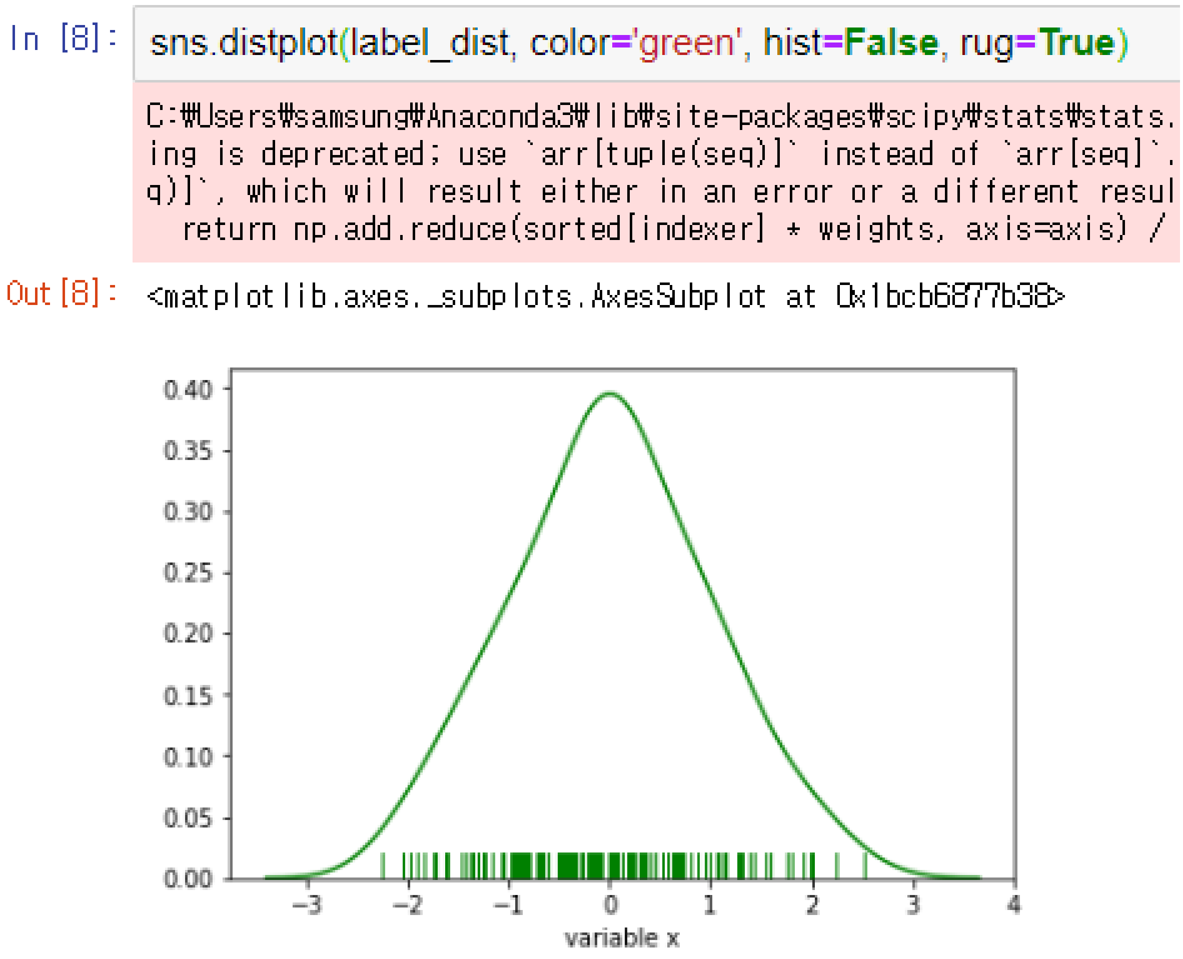


```
In [2]: num = np.random.randn(150) # N(0,1)따르는 150개의 난수
```

```
In [3]: sns.distplot(num)
```

```
C:₩Users₩samsung₩Anaconda3₩lib₩site-packages₩scipy₩stats₩stats.py:1
ing is deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`. In
q)]`, which will result either in an error or a different result.
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumv
C:₩Users₩samsung₩Anaconda3₩lib₩site-packages₩matplotlib₩axes₩_axes.
aced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```

```
Out[3]: <matplotlib.axes._subplots.AxesSubplot at 0x1bb8f614c88>
```

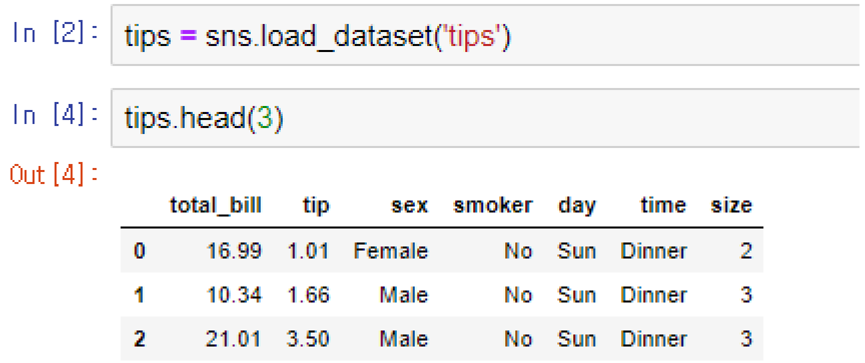


```
In [8]: sns.distplot(label_dist, color='green', hist=False, rug=True)
```

```
C:₩Users₩samsung₩Anaconda3₩lib₩site-packages₩scipy₩stats₩stats.
ing is deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`.
q)]`, which will result either in an error or a different resul
  return np.add.reduce(sorted[indexer] * weights, axis=axis) /
```

```
Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x1bcb6877b38>
```

2. sns.boxplot

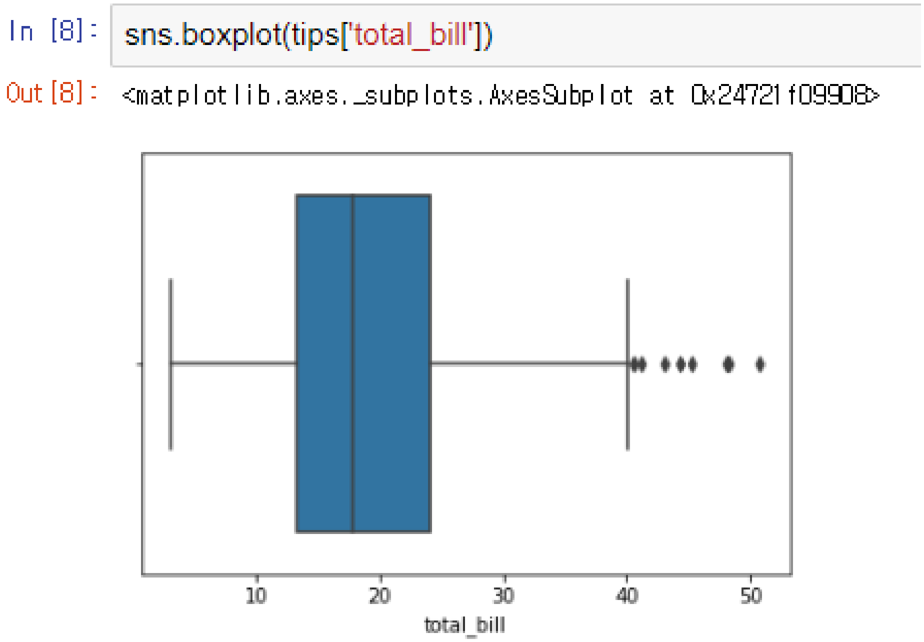


```
In [2]: tips = sns.load_dataset('tips')
```

```
In [4]: tips.head(3)
```

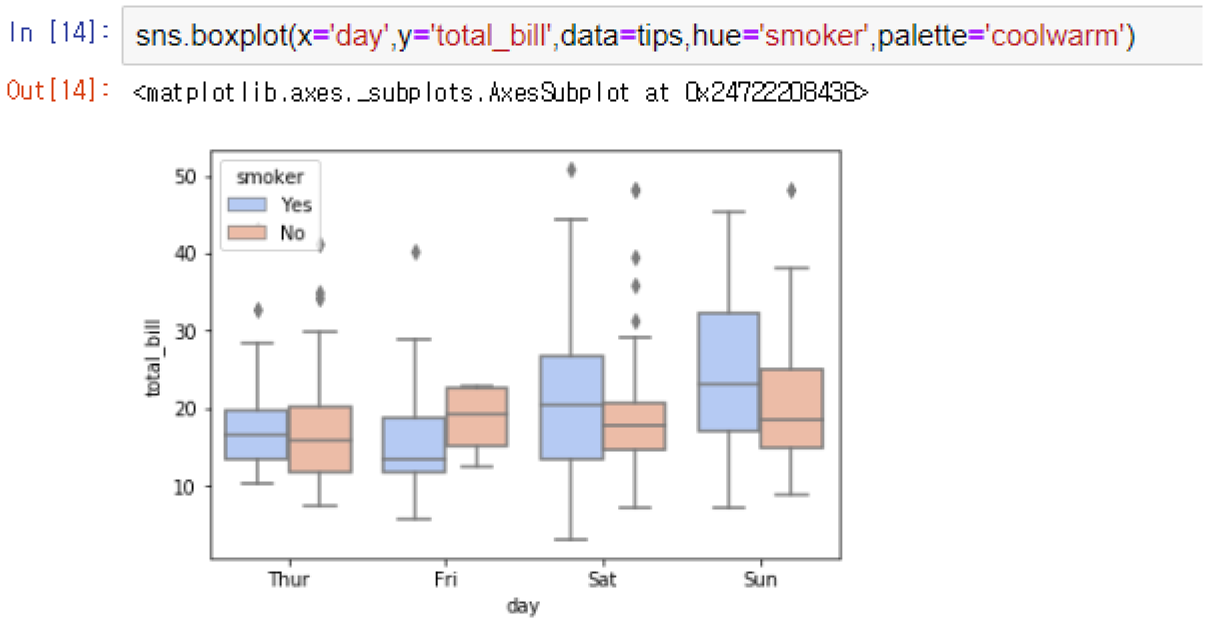
Out [4]:

	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



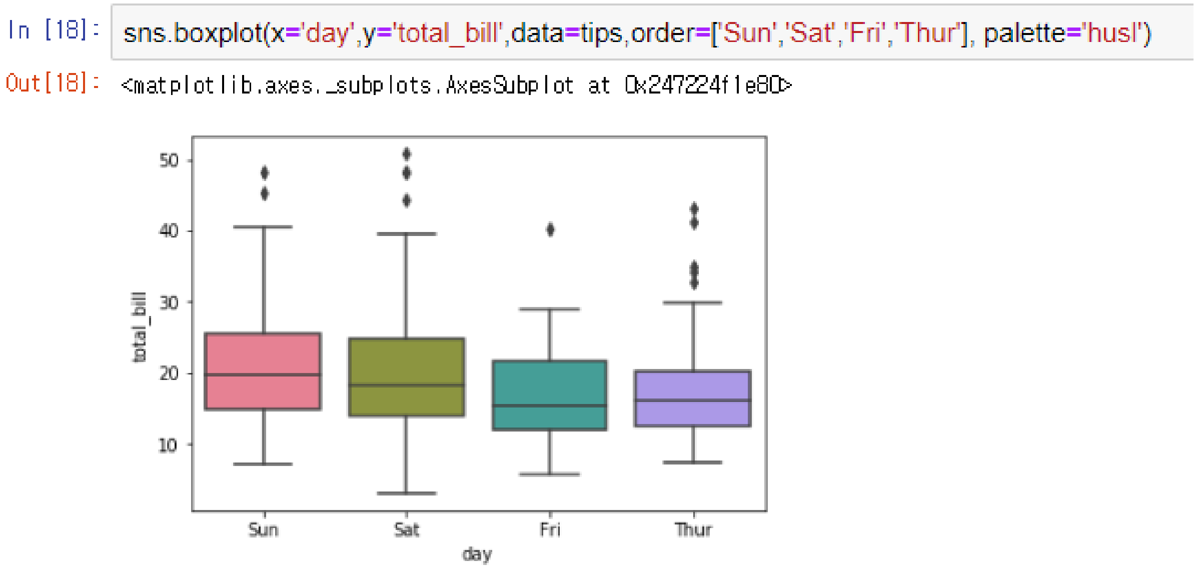
```
In [8]: sns.boxplot(tips['total_bill'])
```

```
Out [8]: <matplotlib.axes._subplots.AxesSubplot at 0x24721f09908>
```



```
In [14]: sns.boxplot(x='day',y='total_bill',data=tips,hue='smoker',palette='coolwarm')
```

```
Out[14]: <matplotlib.axes._subplots.AxesSubplot at 0x24722208438>
```



```
In [18]: sns.boxplot(x='day',y='total_bill',data=tips,order=['Sun','Sat','Fri','Thur'], palette='husl')
```

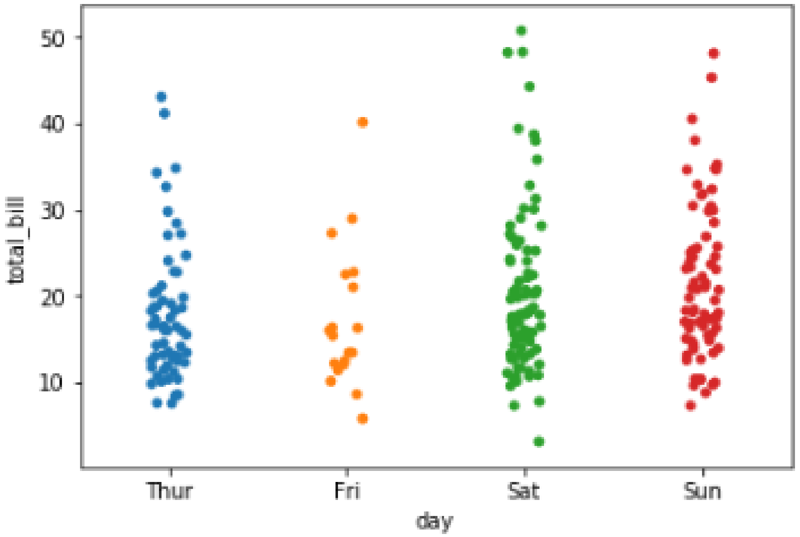
```
Out[18]: <matplotlib.axes._subplots.AxesSubplot at 0x247224f1e80>
```

3. **sns.stripplot** ("띠")

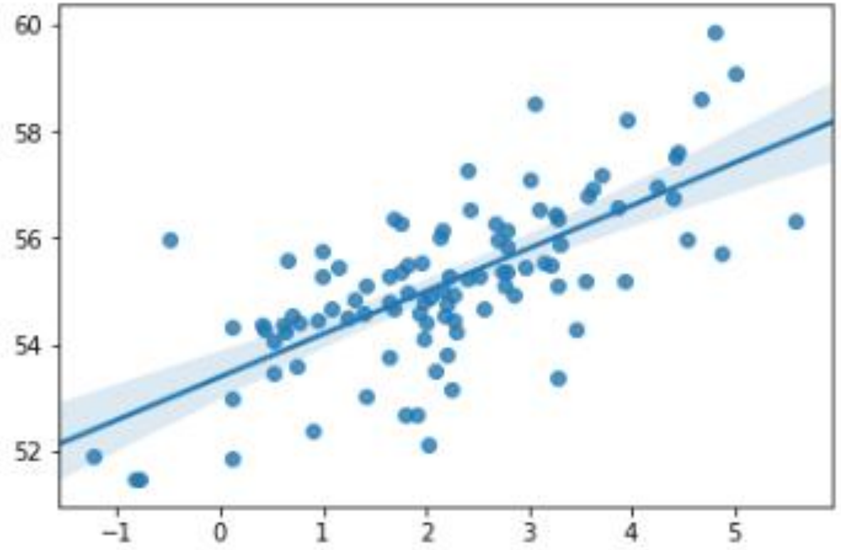
값들의 분포를 "띠" 형태로 확인

4. **sns.regplot**

reg=regression (회귀)

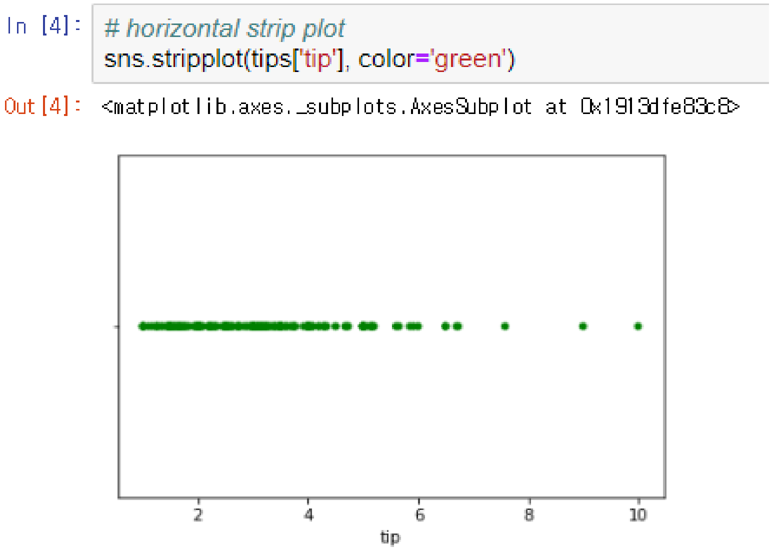


(sns.stripplot)

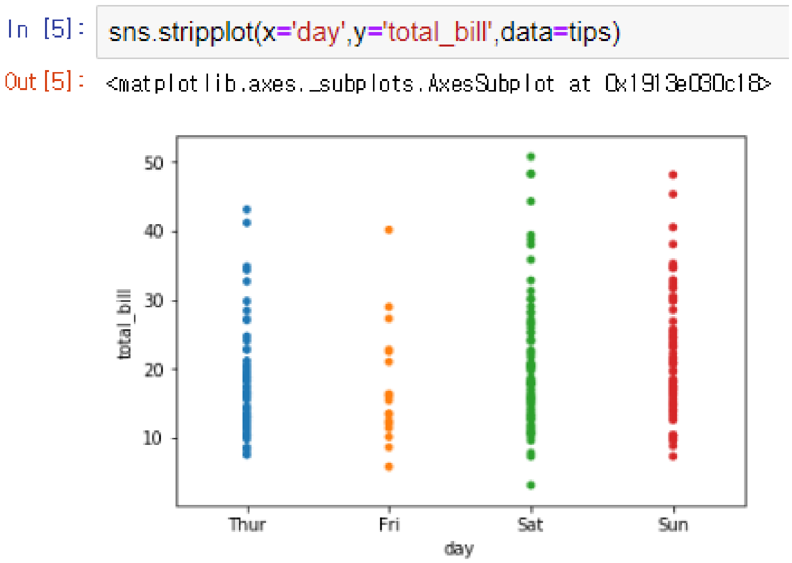


(sns.regplot)

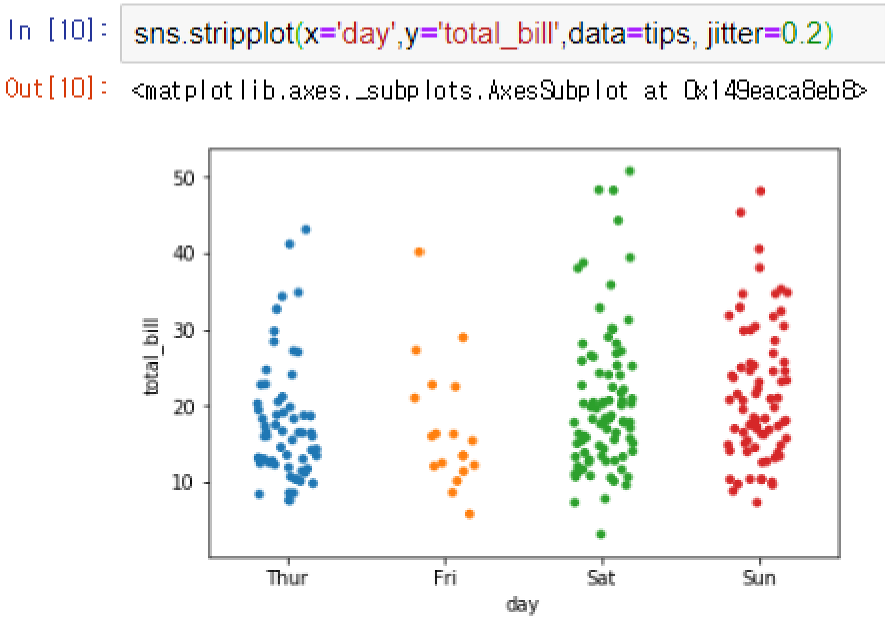
3. sns.stripplot



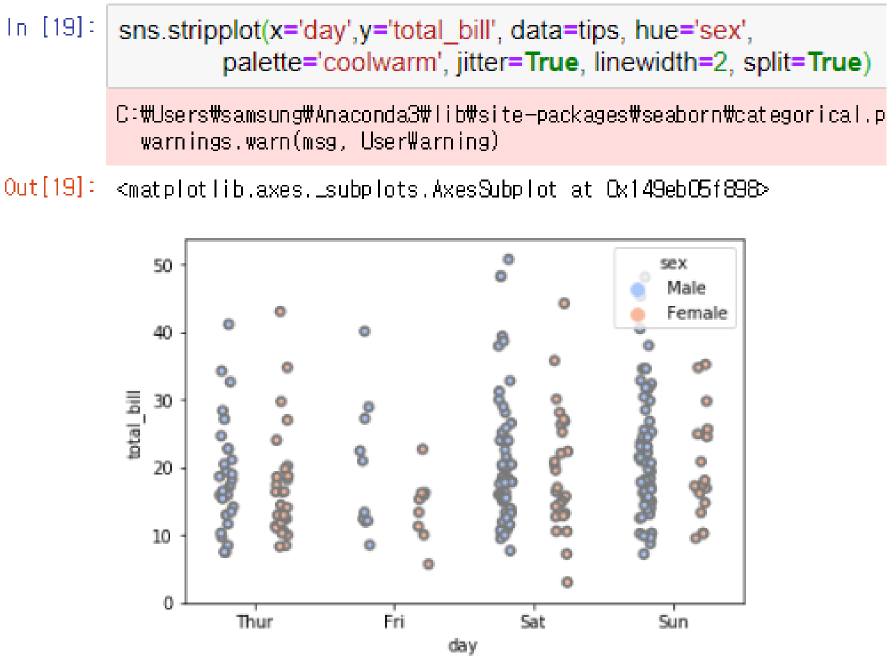
```
In [4]: # horizontal strip plot
        sns.stripplot(tips['tip'], color='green')
Out[4]: <matplotlib.axes._subplots.AxesSubplot at 0x1913dfe83c8>
```



```
In [5]: sns.stripplot(x='day',y='total_bill',data=tips)
Out[5]: <matplotlib.axes._subplots.AxesSubplot at 0x1913e030c18>
```

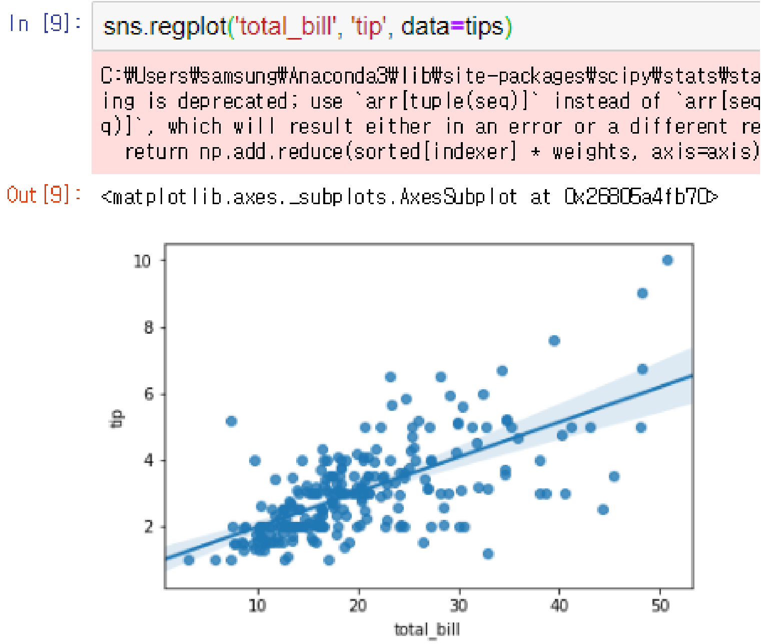


```
In [10]: sns.stripplot(x='day',y='total_bill',data=tips, jitter=0.2)
Out[10]: <matplotlib.axes._subplots.AxesSubplot at 0x149eaca8eb8>
```

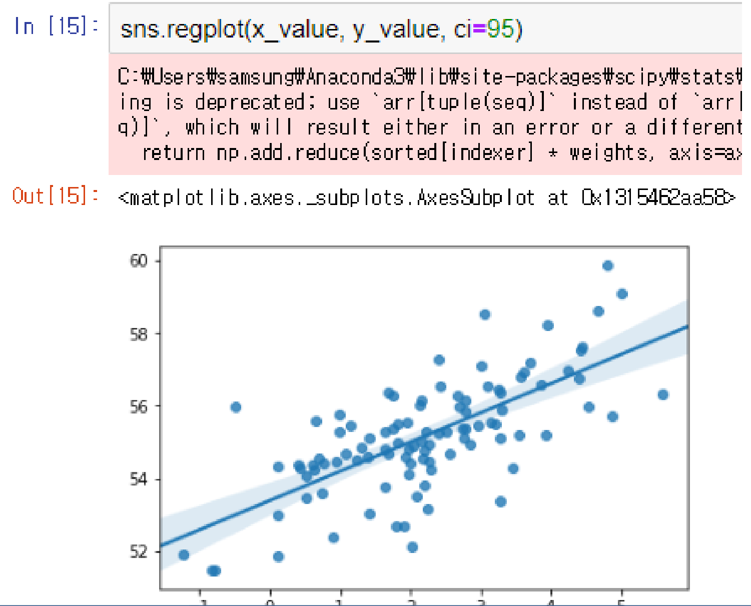


```
In [19]: sns.stripplot(x='day',y='total_bill', data=tips, hue='sex',
                       palette='coolwarm', jitter=True, linewidth=2, split=True)

C:₩Users₩samsung₩Anaconda3₩lib₩site-packages₩seaborn₩categorical.p
  warnings.warn(msg, UserWarning)

Out[19]: <matplotlib.axes._subplots.AxesSubplot at 0x149eb05f898>
```

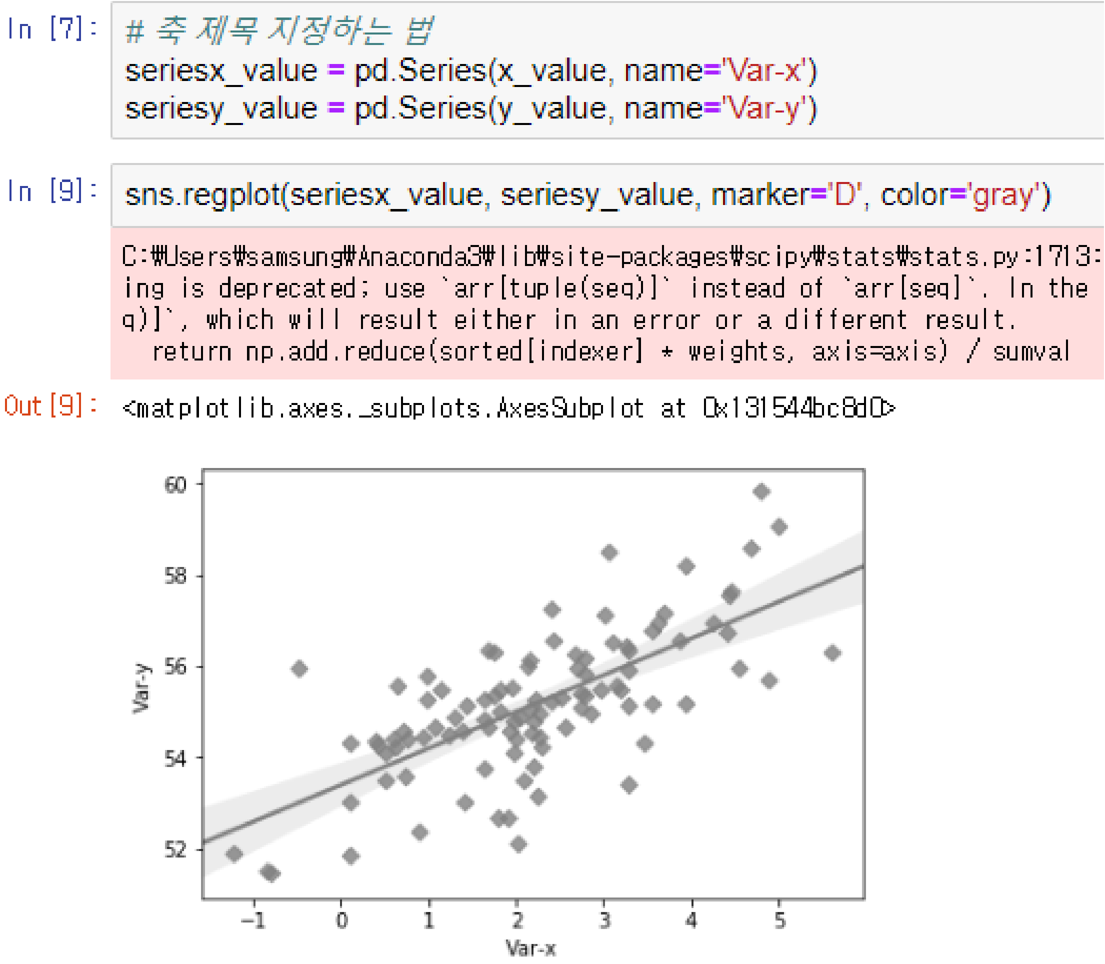
4. sns.regplot



```
In [9]: sns.regplot('total_bill', 'tip', data=tips)

C:₩Users₩samsung₩Anaconda3₩lib₩site-packages₩scipy₩stats₩sta
ing is deprecated; use `arr[tuple(seq)]` instead of `arr[seq
q)]`, which will result either in an error or a different re
  return np.add.reduce(sorted[indexer] * weights, axis=axis)

Out[9]: <matplotlib.axes._subplots.AxesSubplot at 0x26805a4fb70>
```



```
In [15]: sns.regplot(x_value, y_value, ci=95)

C:₩Users₩samsung₩Anaconda3₩lib₩site-packages₩scipy₩stats₩
ing is deprecated; use `arr[tuple(seq)]` instead of `arr[
q)]`, which will result either in an error or a different
  return np.add.reduce(sorted[indexer] * weights, axis=ax

Out[15]: <matplotlib.axes._subplots.AxesSubplot at 0x1315462aa58>
```



```
In [7]: # 축 제목 지정하는 법
        seriesx_value = pd.Series(x_value, name='Var-x')
        seriesy_value = pd.Series(y_value, name='Var-y')

In [9]: sns.regplot(seriesx_value, seriesy_value, marker='D', color='gray')

C:₩Users₩samsung₩Anaconda3₩lib₩site-packages₩scipy₩stats₩stats.py:1713:
ing is deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`. In the
q)]`, which will result either in an error or a different result.
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval

Out[9]: <matplotlib.axes._subplots.AxesSubplot at 0x131544bc8d0>
```

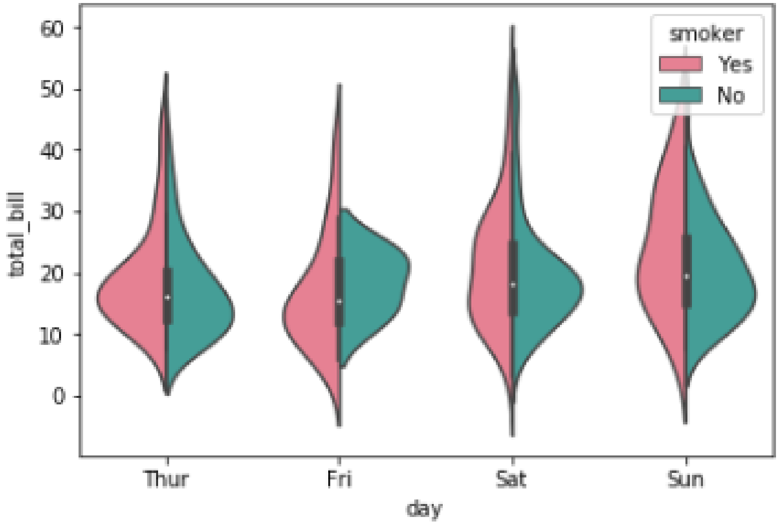
5. **sns.violinplot** (“바이올린 모양”)

boxplot + 확률밀도

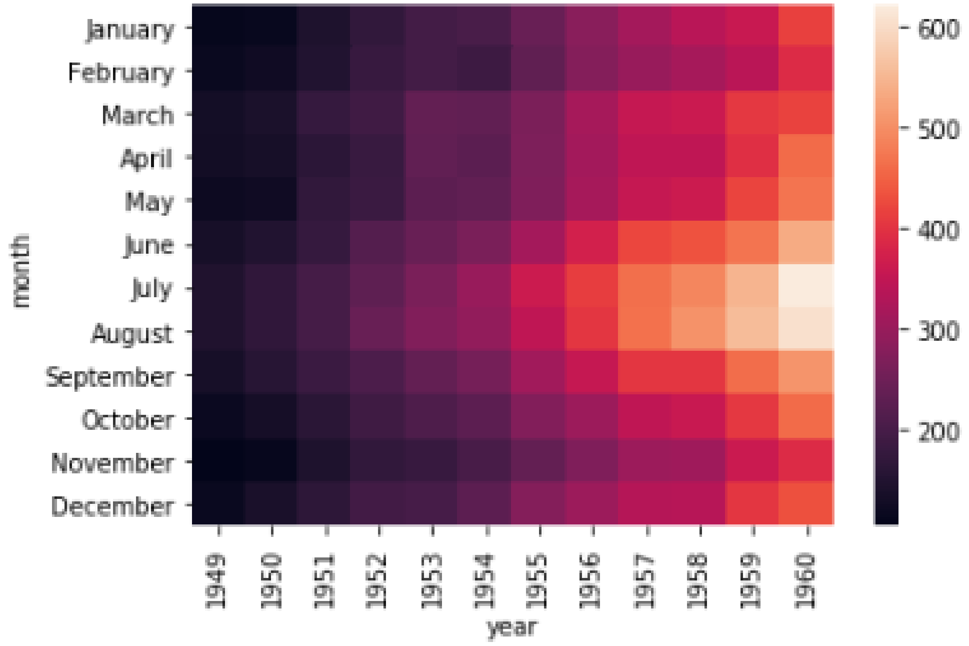
6. **sns.heatmap**

값들을 ‘색깔’로 표현

값들의 변화 양상을 보기에 good

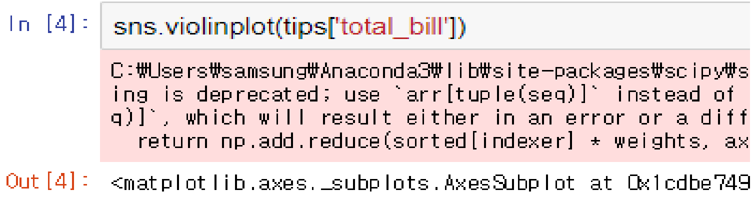


(sns.violinplot)

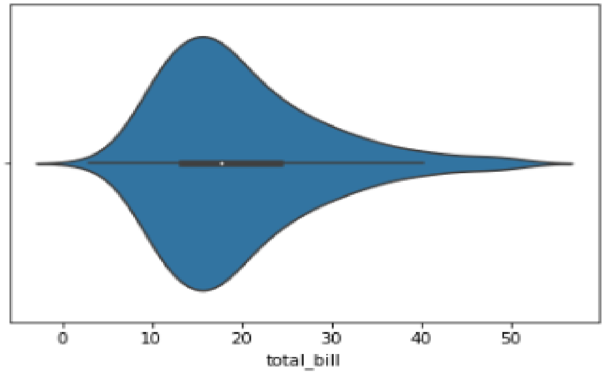


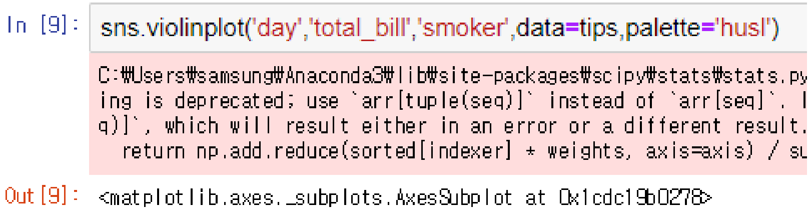
(sns.heatmap)

5. sns.violinplot

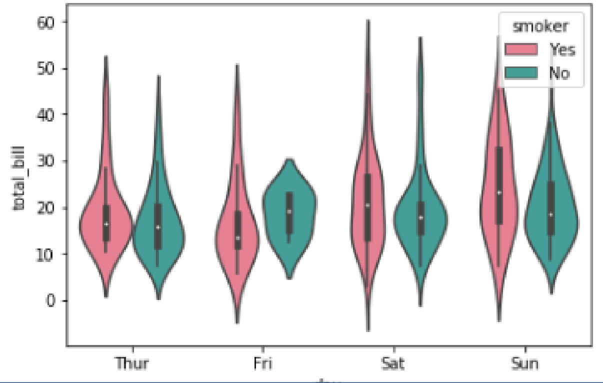


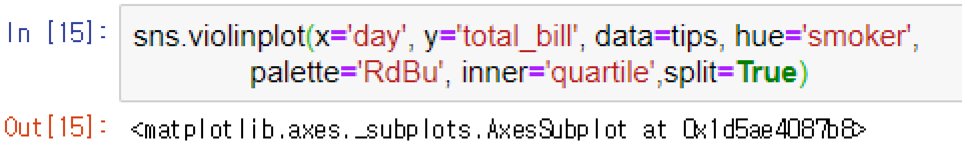
```
In [4]: sns.violinplot(tips['total_bill'])

C:₩Users₩samsung₩Anaconda3₩lib₩site-packages₩scipy₩s
ing is deprecated; use `arr[tuple(seq)]` instead of
q)]`, which will result either in an error or a diff
  return np.add.reduce(sorted[indexer] * weights, ax

Out[4]: <matplotlib.axes._subplots.AxesSubplot at 0x1cdbe749
```



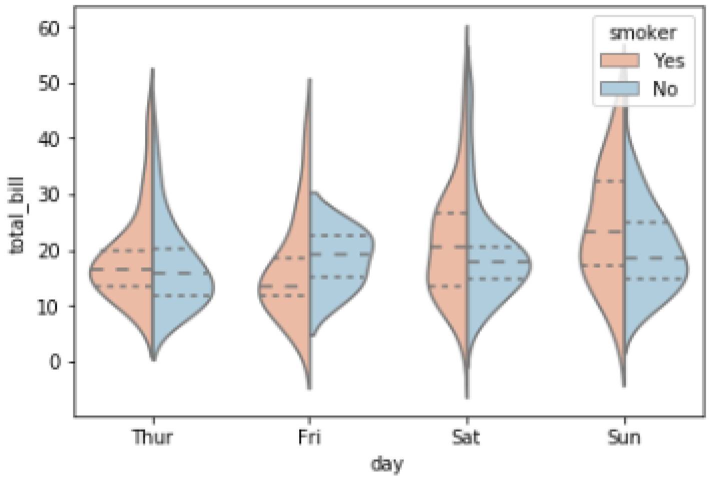


```
In [9]: sns.violinplot('day','total_bill','smoker',data=tips,palette='husl')

C:₩Users₩samsung₩Anaconda3₩lib₩site-packages₩scipy₩stats₩stats.py
ing is deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`. I
q)]`, which will result either in an error or a different result.
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / su

Out[9]: <matplotlib.axes._subplots.AxesSubplot at 0x1cdc19b0278>
```





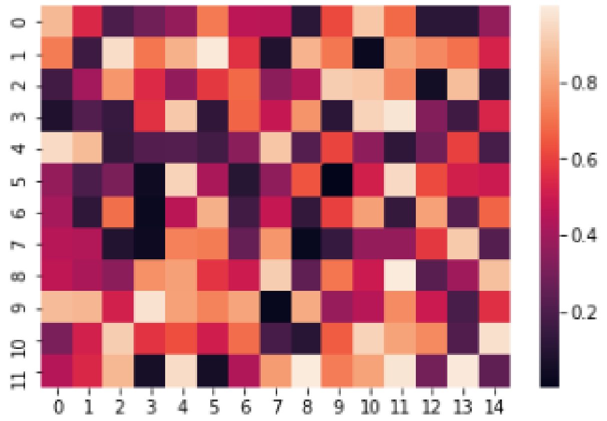
```
In [15]: sns.violinplot(x='day', y='total_bill', data=tips, hue='smoker',
                    palette='RdBu', inner='quartile',split=True)

Out[15]: <matplotlib.axes._subplots.AxesSubplot at 0x1d5ae4087b8>
```



6. sns.heatmap

```
In [3]: # 0~1사이 값
        normal = np.random.rand(12,15)
```

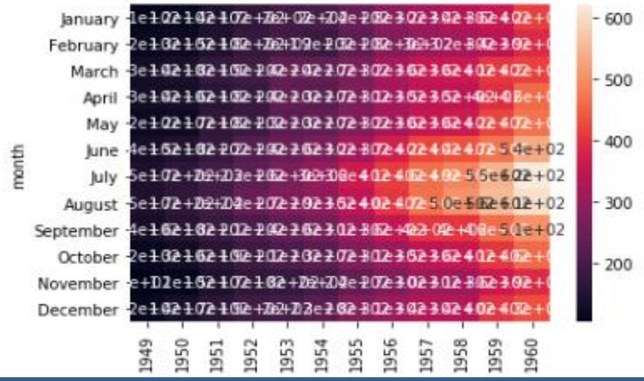
```
In [4]: sns.heatmap(normal)
Out[4]: <matplotlib.axes._subplots.AxesSubplot at 0x1ea0d755828>
```



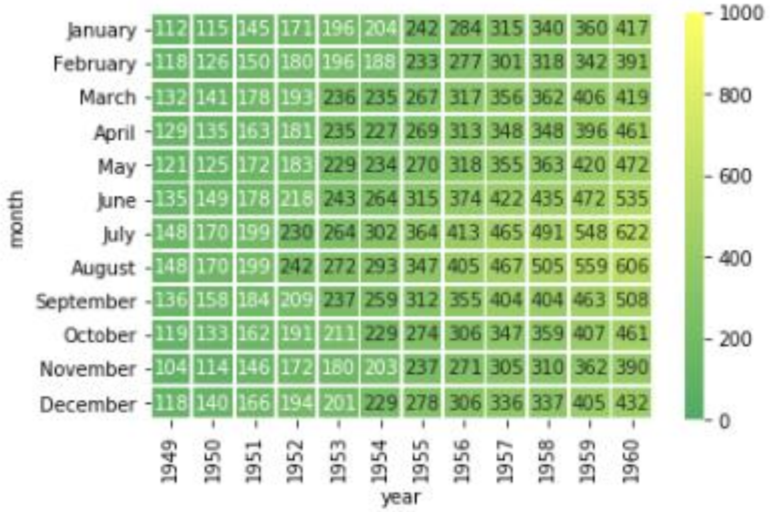
```
In [8]: flights = sns.load_dataset('flights')
```

```
In [9]: flights = flights.pivot('month','year','passengers')
```

```
In [13]: sns.heatmap(flights, annot=True)
Out[13]: <matplotlib.axes._subplots.AxesSubplot at 0x1ea10ee4cf8>
```



```
In [23]: sns.heatmap(flights, linewidths=0.9, cmap='summer',
                     annot=True, fmt='d', vmin=0,vmax=1000, center=flights.loc['June',1954])
Out[23]: <matplotlib.axes._subplots.AxesSubplot at 0x1c164214da0>
```



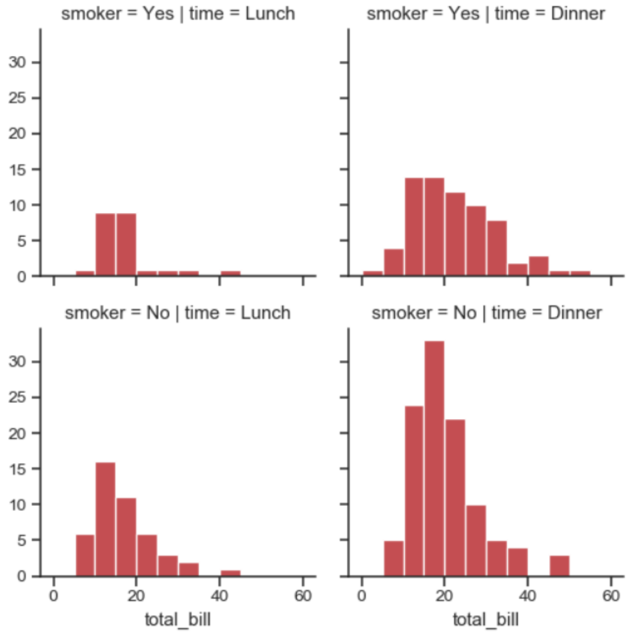
	1949	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960
January	112	115	145	171	196	204	242	284	315	340	360	417
February	118	126	150	180	196	188	233	277	301	318	342	391
March	132	141	178	193	236	235	267	317	356	362	406	419
April	129	135	163	181	235	227	269	313	348	348	396	461
May	121	125	172	183	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

7. sns.facetgrid

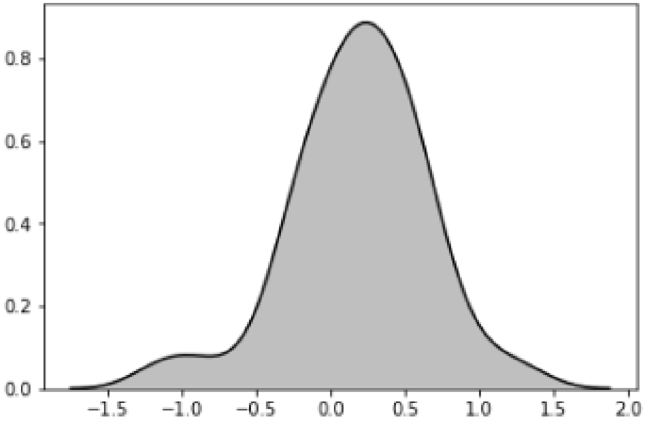
변수의 '단계'or'정도'에 따라 다른 그래프!

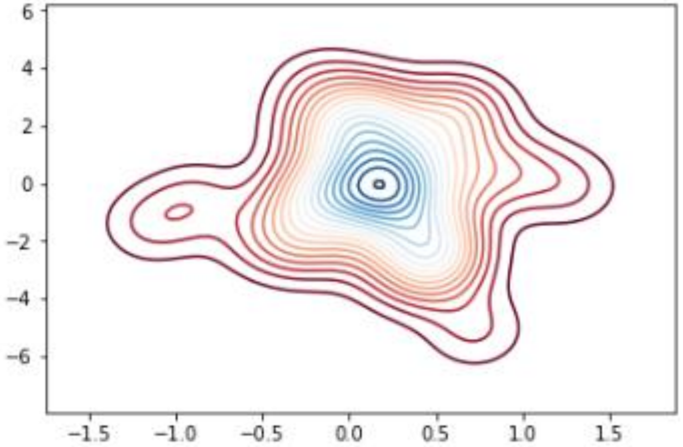
8. sns.kdeplot (kernel density plot)

pdf(확률분포함수)



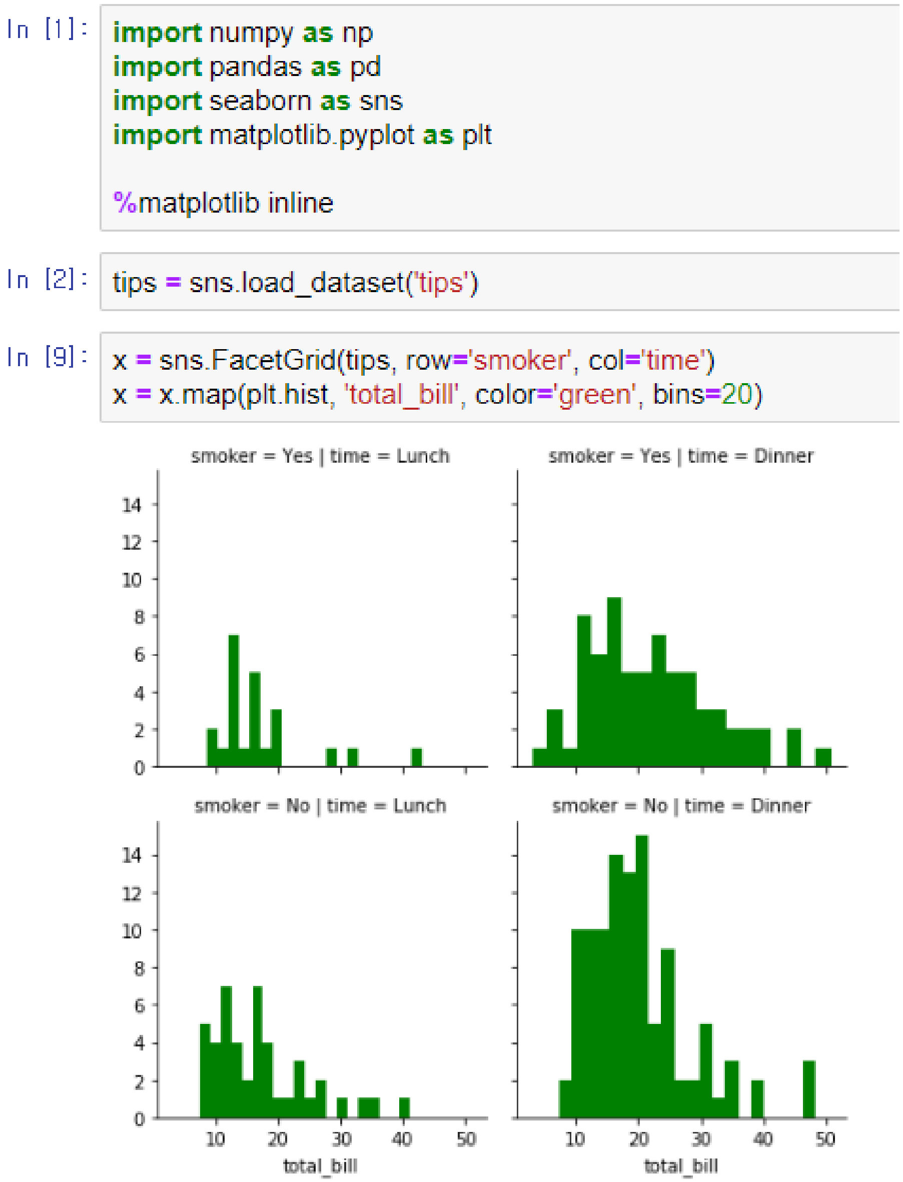
(sns.facetgrid)





(sns.kdeplot)

7. sns.facetgrid

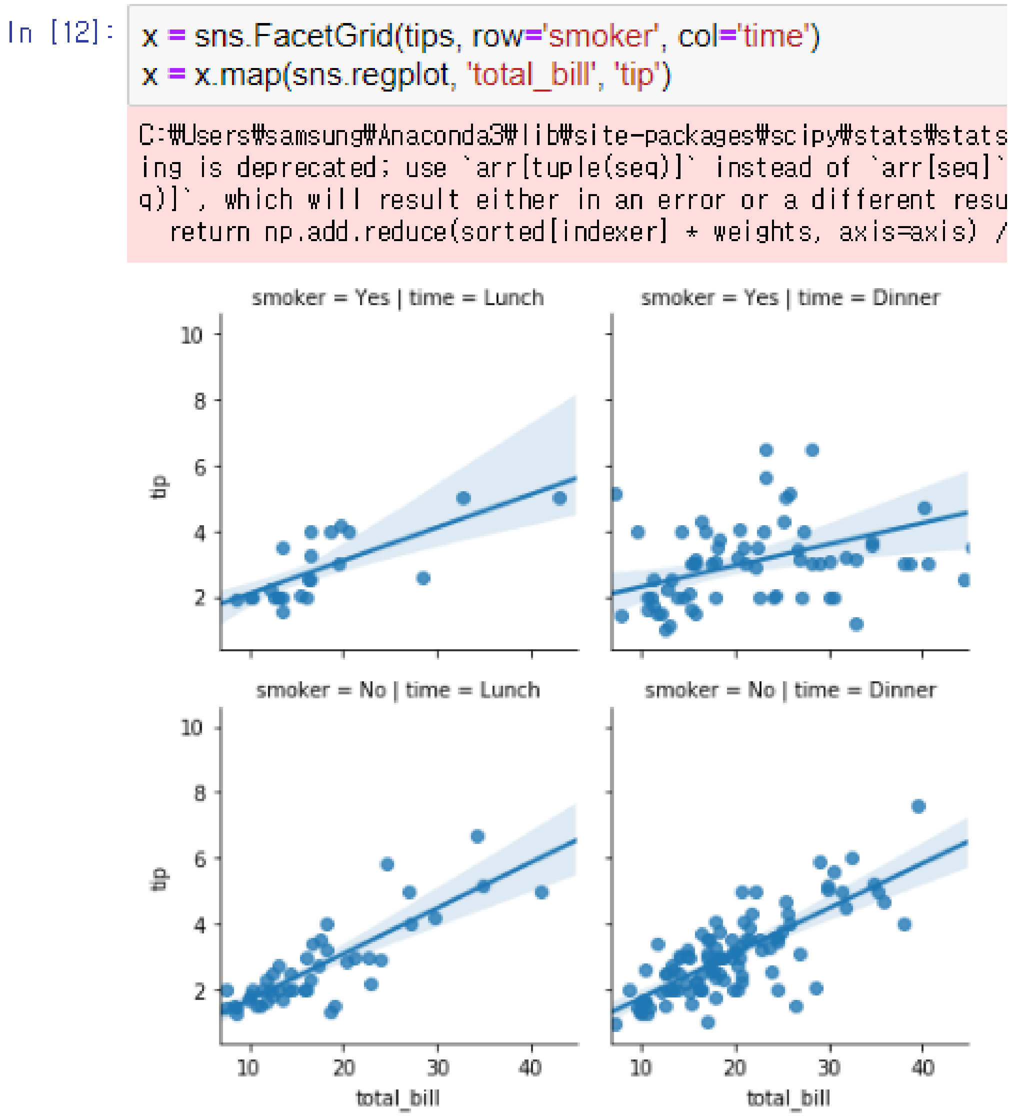


```
In [1]: import numpy as np
        import pandas as pd
        import seaborn as sns
        import matplotlib.pyplot as plt

        %matplotlib inline
```

```
In [2]: tips = sns.load_dataset('tips')
```

```
In [9]: x = sns.FacetGrid(tips, row='smoker', col='time')
        x = x.map(plt.hist, 'total_bill', color='green', bins=20)
```

smoker = Yes | time = Lunch
smoker = Yes | time = Dinner
smoker = No | time = Lunch
smoker = No | time = Dinner
0
2
4
6
8
10
12
14
10
20
30
40
50
total_bill
total_bill

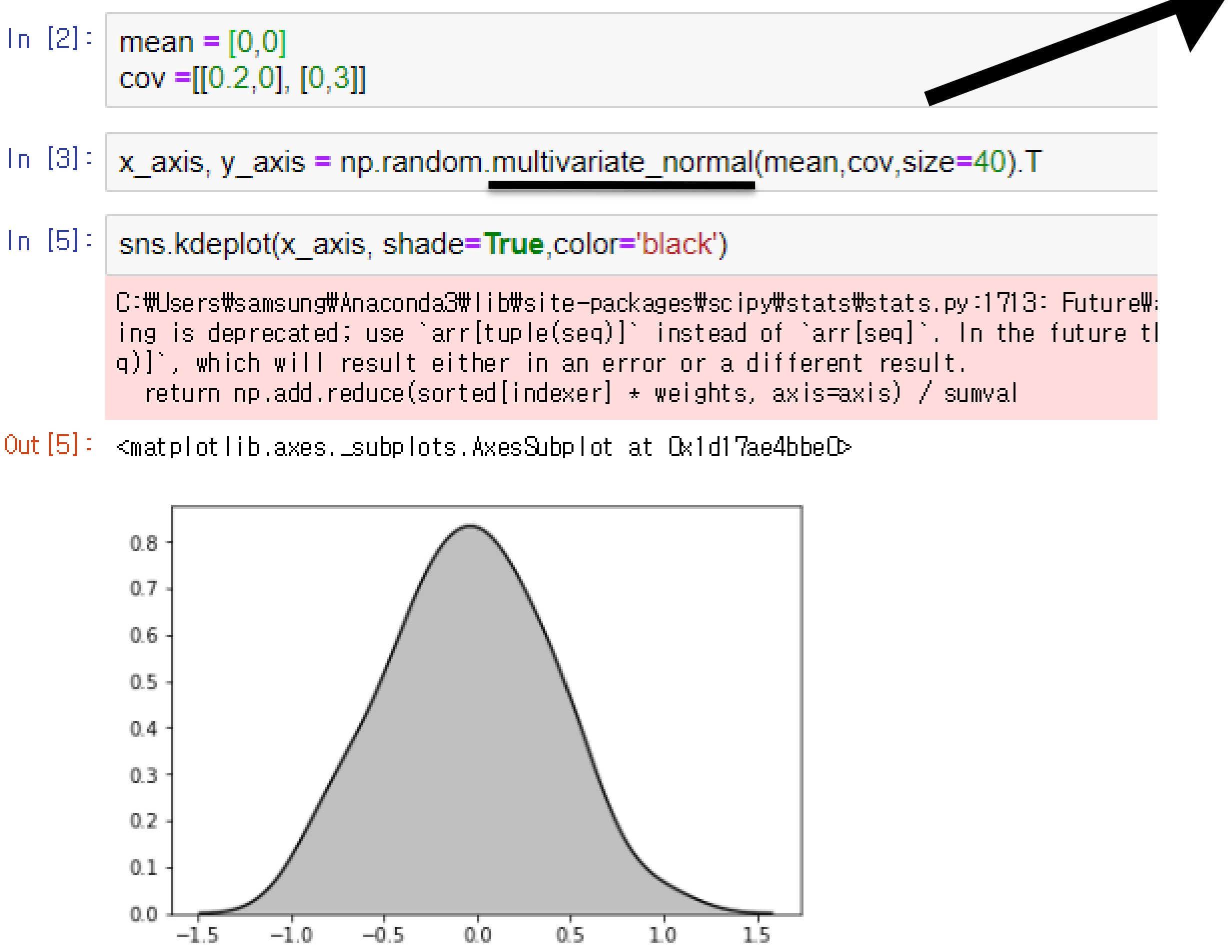


```
In [12]: x = sns.FacetGrid(tips, row='smoker', col='time')
         x = x.map(sns.regplot, 'total_bill', 'tip')
```

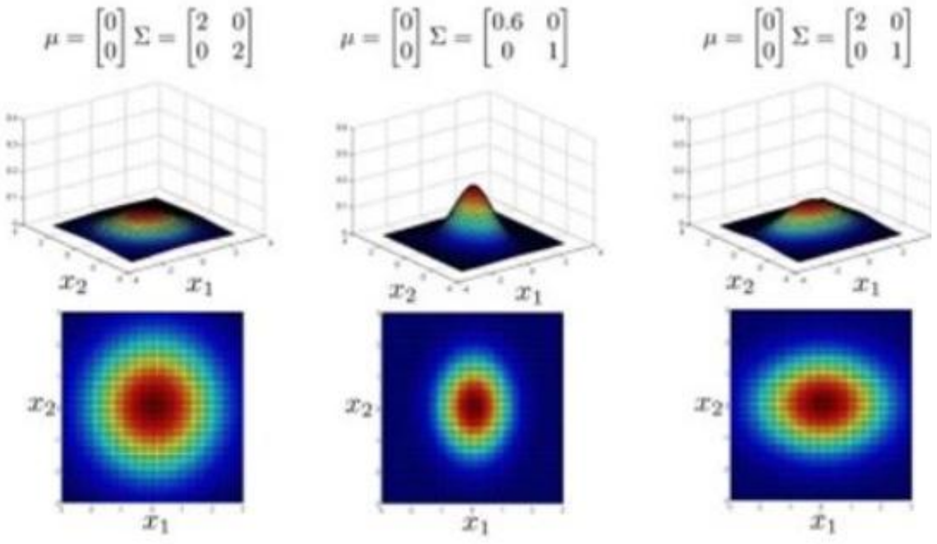
```
C:₩Users₩samsung₩Anaconda3₩lib₩site-packages₩scipy₩stats₩stats
ing is deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`
q)]`, which will result either in an error or a different resu
  return np.add.reduce(sorted[indexer] * weights, axis=axis) /
```

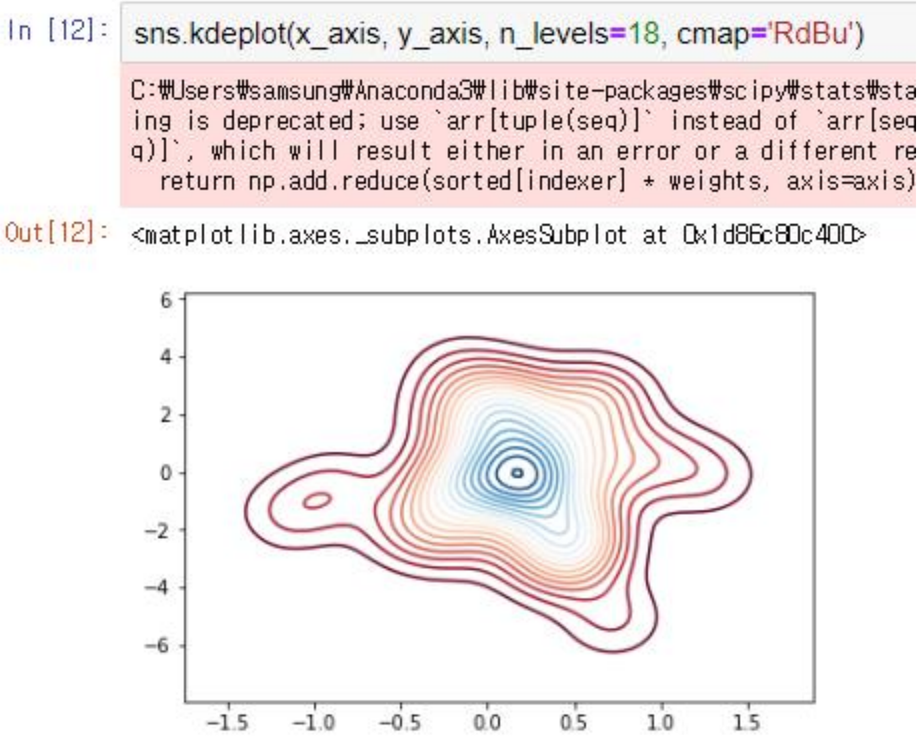
smoker = Yes | time = Lunch
smoker = Yes | time = Dinner
smoker = No | time = Lunch
smoker = No | time = Dinner
tip
tip
2
4
6
8
10
10
20
30
40
total_bill
total_bill

8. sns.kdeplot



```
In [2]:  mean = [0,0]
         cov =[[0.2,0], [0,3]]

In [3]:  x_axis, y_axis = np.random.multivariate_normal(mean,cov,size=40).T

In [5]:  sns.kdeplot(x_axis, shade=True,color='black')

C:₩Users₩samsung₩Anaconda3₩lib₩site-packages₩scipy₩stats₩stats.py:1713: Future₩
ing is deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`. In the future t
q)]`, which will result either in an error or a different result.
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval

Out[5]:  <matplotlib.axes._subplots.AxesSubplot at 0x1d17ae4bbe0>
```





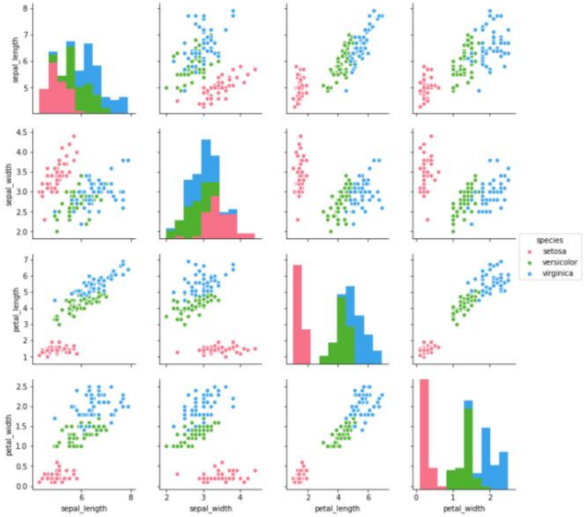
```
In [12]:  sns.kdeplot(x_axis, y_axis, n_levels=18, cmap='RdBu')

C:₩Users₩samsung₩Anaconda3₩lib₩site-packages₩scipy₩stats₩sta
ing is deprecated; use `arr[tuple(seq)]` instead of `arr[seq
q)]`, which will result either in an error or a different re
  return np.add.reduce(sorted[indexer] * weights, axis=axis)

Out[12]:  <matplotlib.axes._subplots.AxesSubplot at 0x1d86c80c400>
```

9. **sns.pairplot & sns.pairgrid**

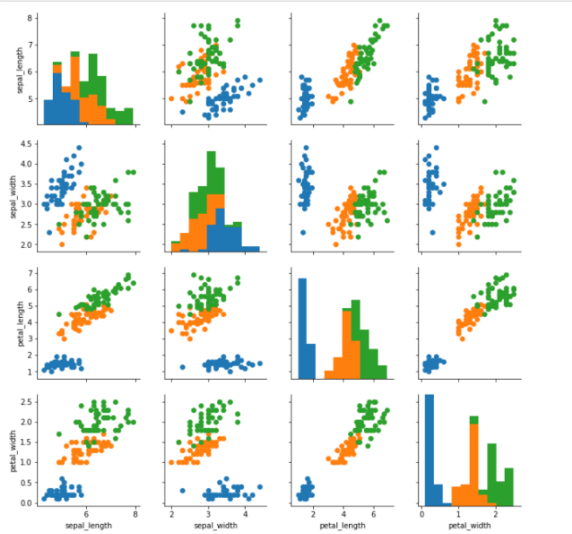
(pair=쌍 -> 두 개의 변수간의 관계)

10. **sns.jointplot**

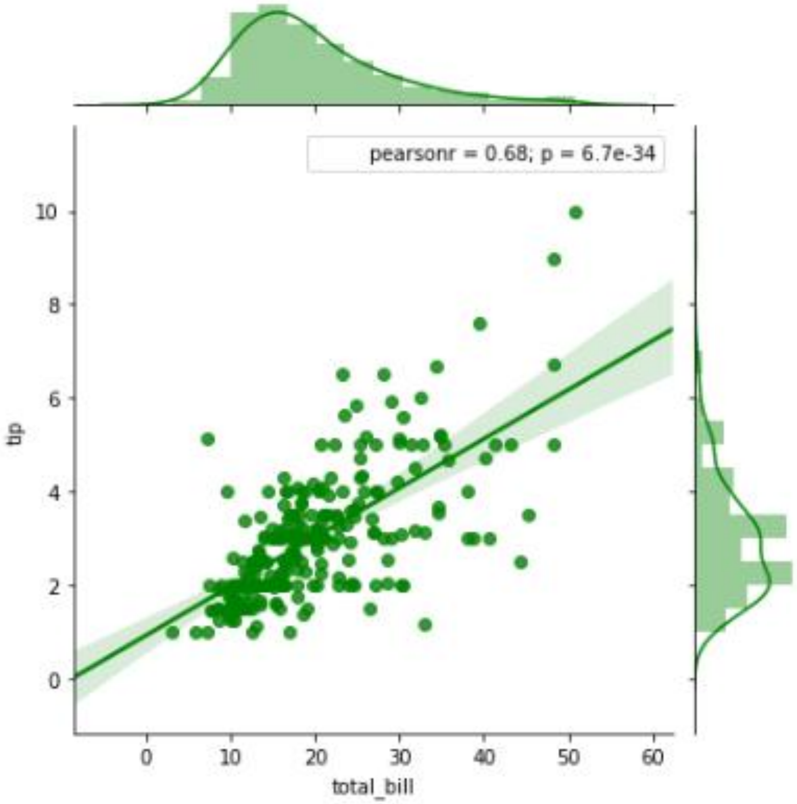
말 그대로 join! 두 종류의 그래프를 같이 표시함



(sns.pairplot)

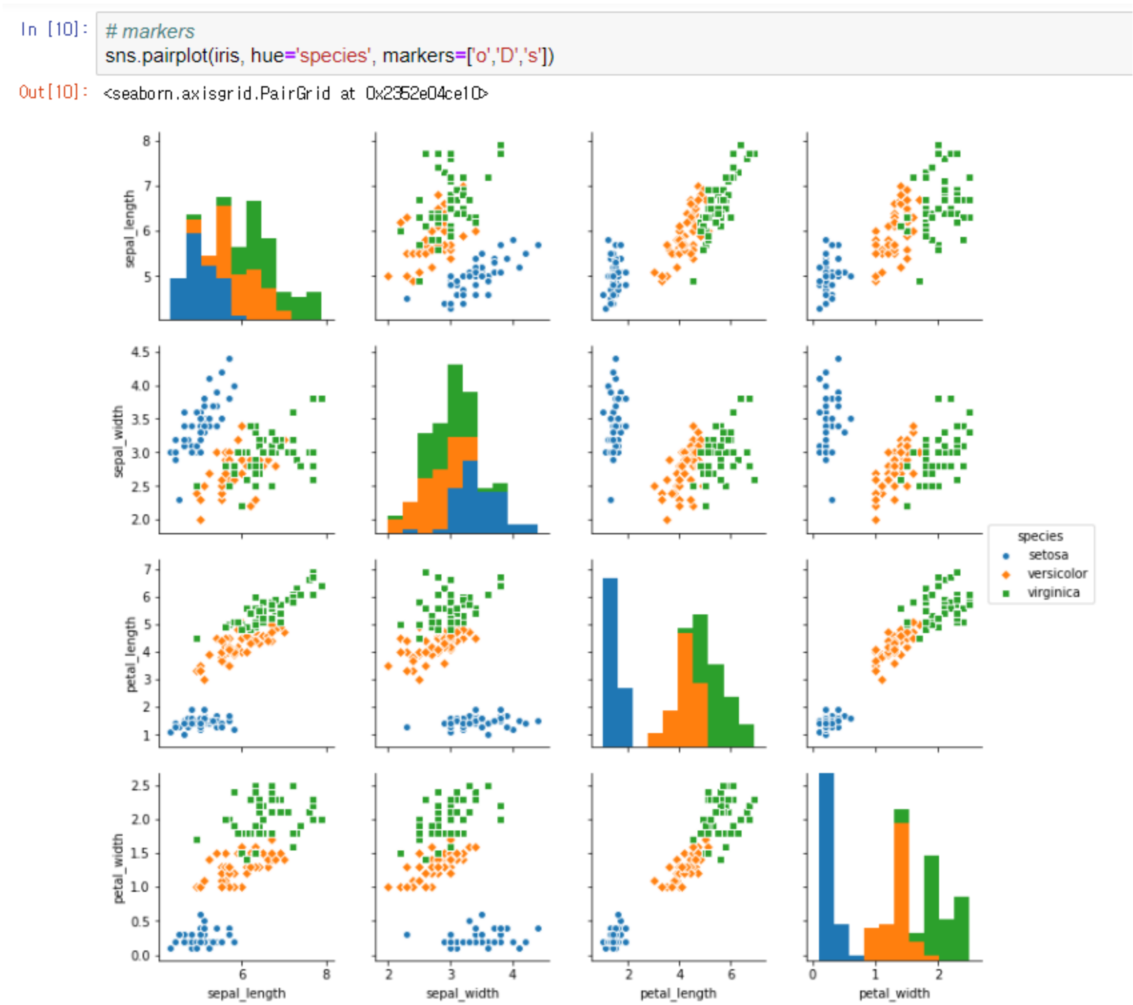


(sns.pairgrid)

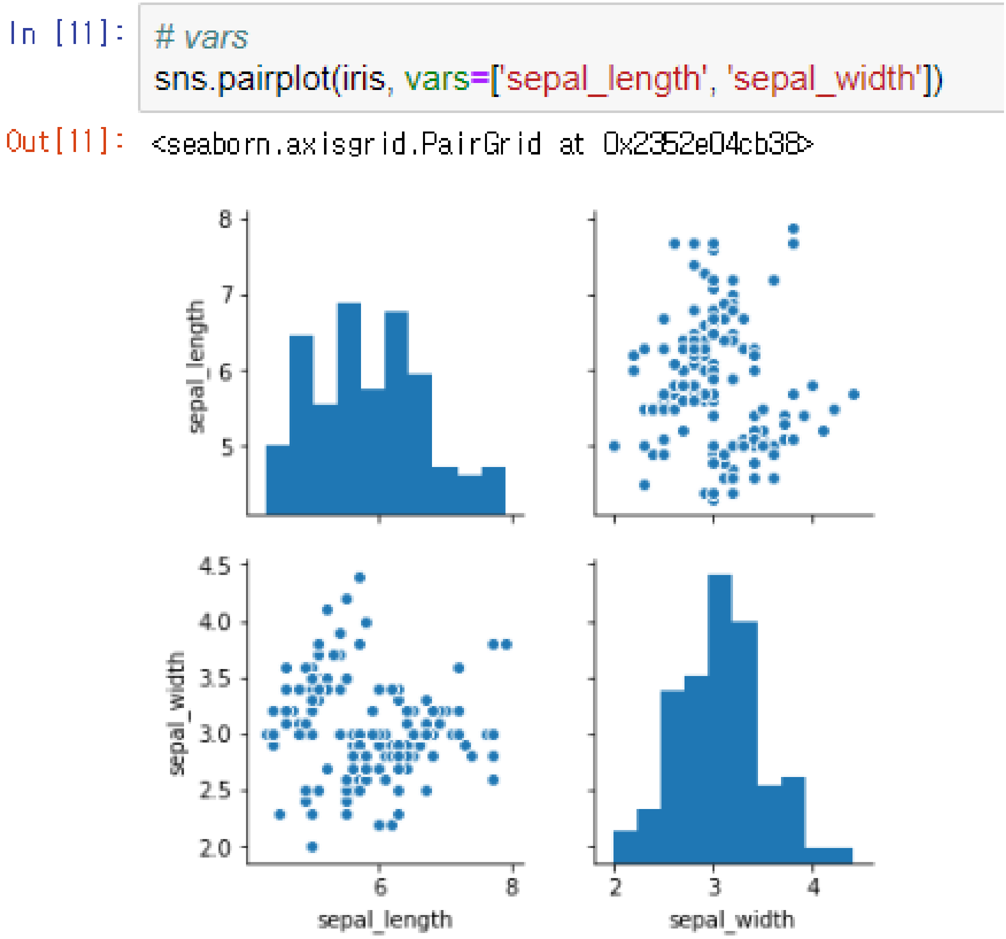


(sns.jointplot)

9 (1) sns.pairplot

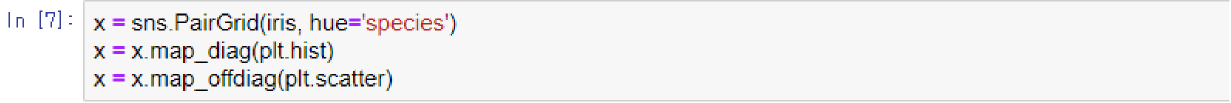


```
In [10]:  # markers
          sns.pairplot(iris, hue='species', markers=['o','D','s'])
Out[10]:  <seaborn.axisgrid.PairGrid at 0x2352e04ce10>
```



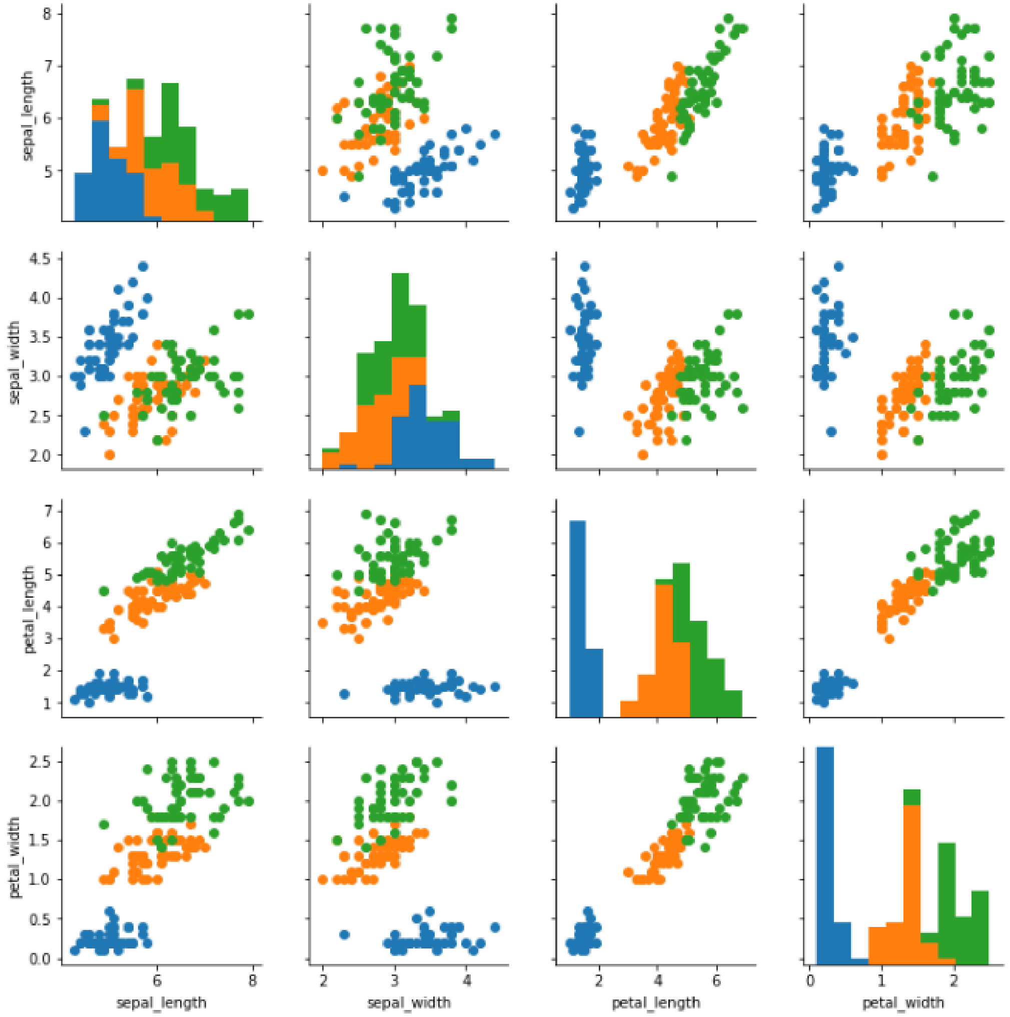
```
In [11]:  # vars
          sns.pairplot(iris, vars=['sepal_length', 'sepal_width'])
Out[11]:  <seaborn.axisgrid.PairGrid at 0x2352e04cb38>
```

9 (2) sns.pairgrid

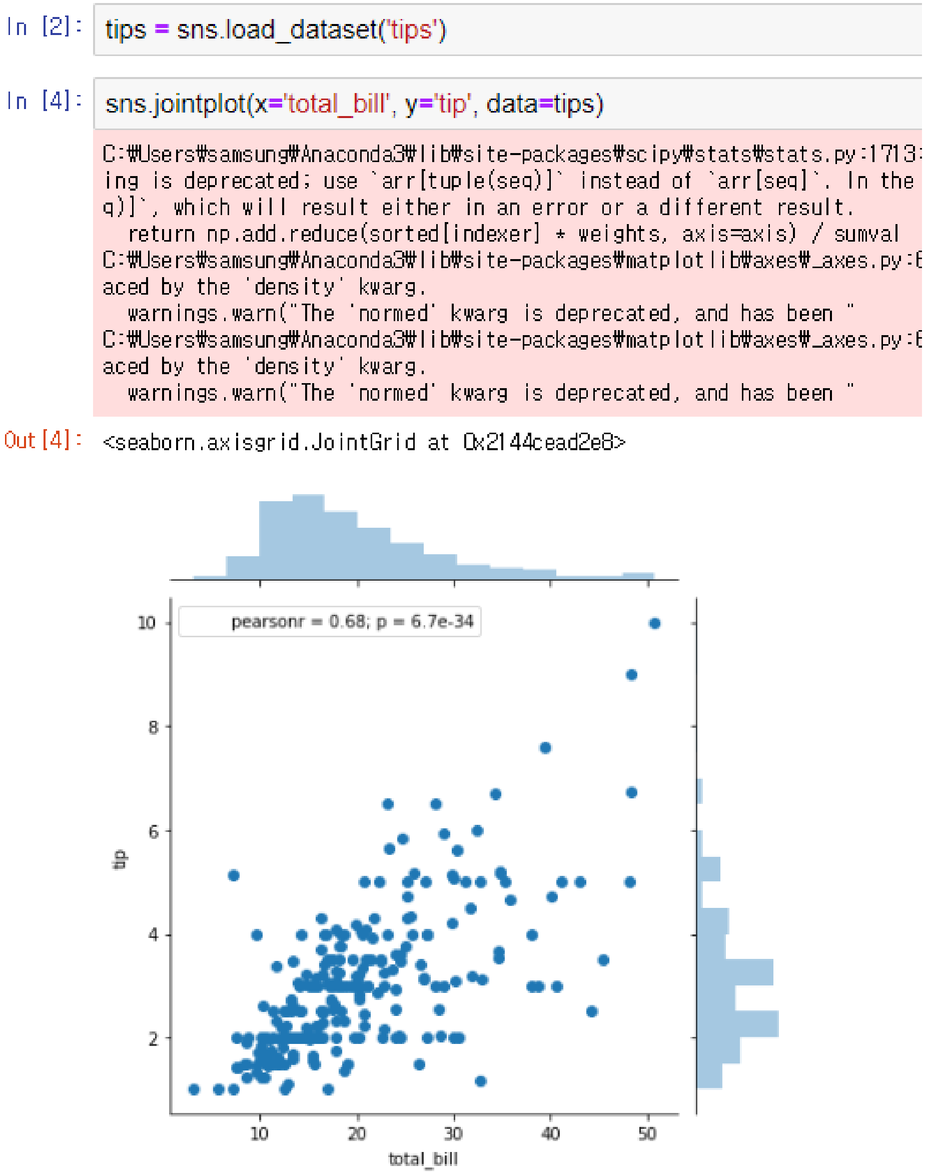


In [7]:

```
x = sns.PairGrid(iris, hue='species')
x = x.map_diag(plt.hist)
x = x.map_offdiag(plt.scatter)
```



10. sns.jointplot

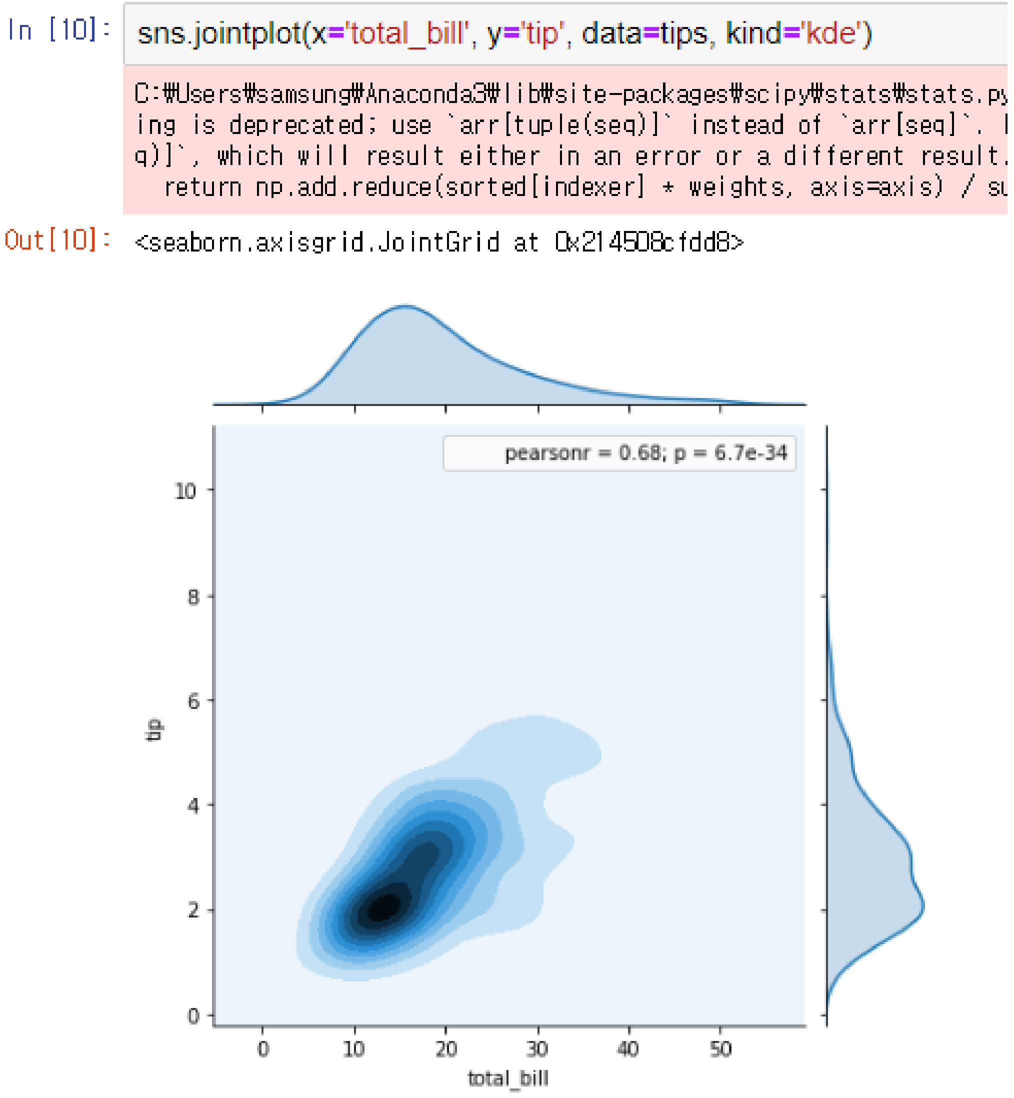


```
In [2]: tips = sns.load_dataset('tips')
```

```
In [4]: sns.jointplot(x='total_bill', y='tip', data=tips)
```

```
C:₩Users₩samsung₩Anaconda3₩lib₩site-packages₩scipy₩stats₩stats.py:1713:
ing is deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`. In the
q)]`, which will result either in an error or a different result.
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
C:₩Users₩samsung₩Anaconda3₩lib₩site-packages₩matplotlib₩axes₩_axes.py:6
aced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
C:₩Users₩samsung₩Anaconda3₩lib₩site-packages₩matplotlib₩axes₩_axes.py:6
aced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```

```
Out[4]: <seaborn.axisgrid.JointGrid at 0x2144cead2e8>
```



```
In [10]: sns.jointplot(x='total_bill', y='tip', data=tips, kind='kde')
```

```
C:₩Users₩samsung₩Anaconda3₩lib₩site-packages₩scipy₩stats₩stats.py
ing is deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`. I
q)]`, which will result either in an error or a different result.
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / su
```

```
Out[10]: <seaborn.axisgrid.JointGrid at 0x214508cfdd8>
```

코드 통해서

Jupyter notebook

(data cleansing & visualization) Seaborn

실습

Jupyter notebook

실습 문제 (wine.csv)