ta-visualization-eda-using-seaborn

September 25, 2023

1 Seaborn

To implement exploratory data analysis and data visualization using Seaborn. Explore following type of graphs for continuous variables and for categorical variables using seaborn. 1. Correlation matrix using Heatmap 2. sns.jointplot 3. sns.pairplot 4. sns.distplot 5. sns.countplot 6. sns.barplot 7. sns.boxplot 8. sns.violinplot

1.1 plots for continuous variables/ features

[2]:
```
import seaborn as sns
data=sns.load_dataset("tips")
```

[3]:
```
data.head()
```

[3]:

	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]:
```
data.info()
```

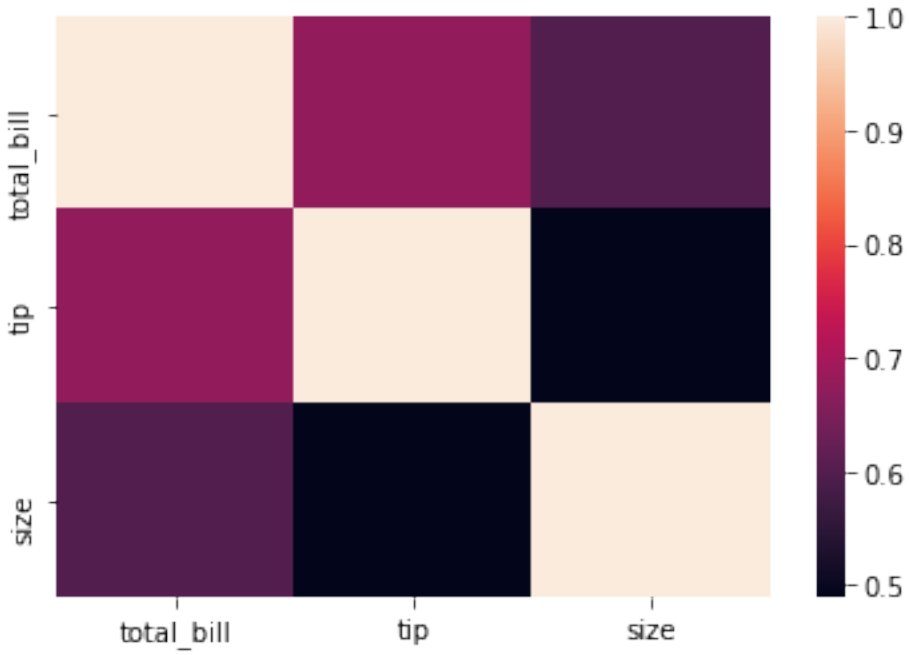
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 244 entries, 0 to 243
Data columns (total 7 columns):
total_bill    244 non-null float64
tip           244 non-null float64
sex           244 non-null category
smoker        244 non-null category
day           244 non-null category
time          244 non-null category
size          244 non-null int64
dtypes: category(4), float64(2), int64(1)
memory usage: 7.2 KB
```

[6]:
```
data.corr()
```

[6]:
```
              total_bill       tip      size
total_bill      1.000000  0.675734  0.598315
tip             0.675734  1.000000  0.489299
size            0.598315  0.489299  1.000000
```

[9]:
```
sns.heatmap(data.corr())
```

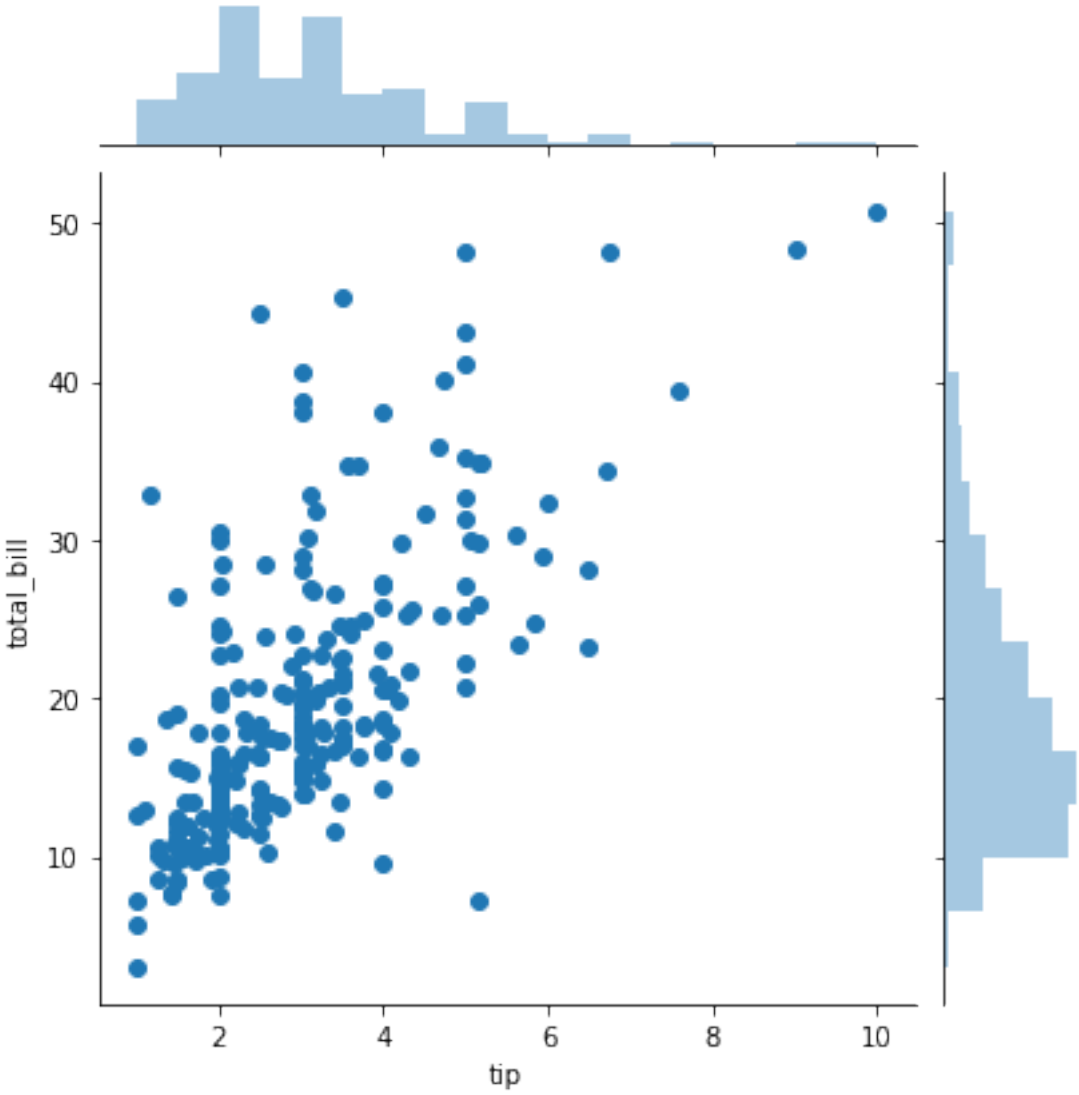
[9]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x22a98451a90>
```



[10]:
```
sns.jointplot(x="tip",y="total_bill",data=data,kind="scatter")
```

```
C:\python3.6.3\lib\site-packages\scipy\stats\stats.py:1706: FutureWarning: Using
a non-tuple sequence for multidimensional indexing is deprecated; use
`arr[tuple(seq)]` instead of `arr[seq]`. In the future this will be interpreted
as an array index, `arr[np.array(seq)]`, which will result either in an error or
a different result.
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
```

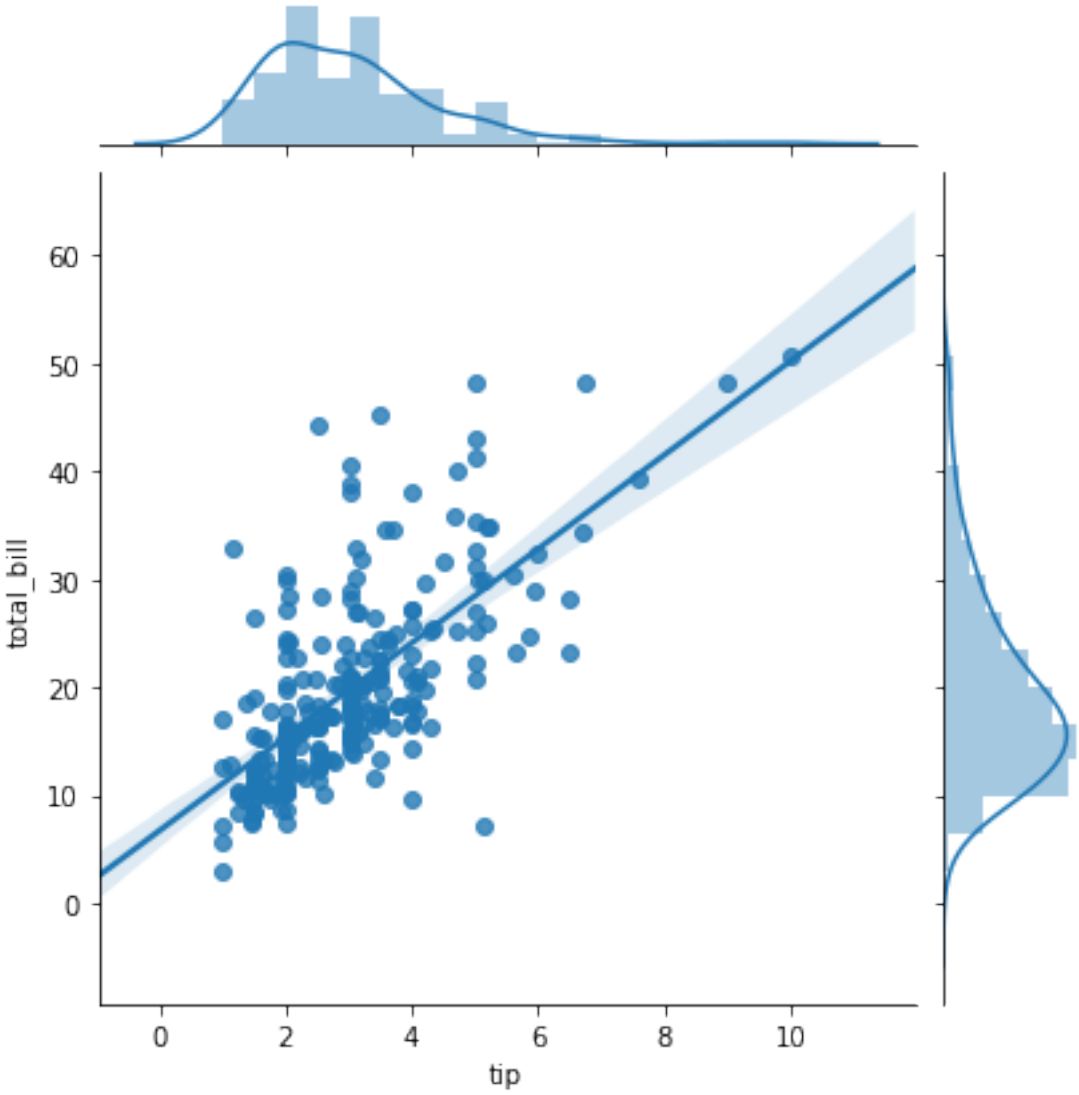
[10]:
```
<seaborn.axisgrid.JointGrid at 0x22a9846b2b0>
```



[11]:
```
sns.jointplot(x="tip",y="total_bill",data=data,kind="reg")
```

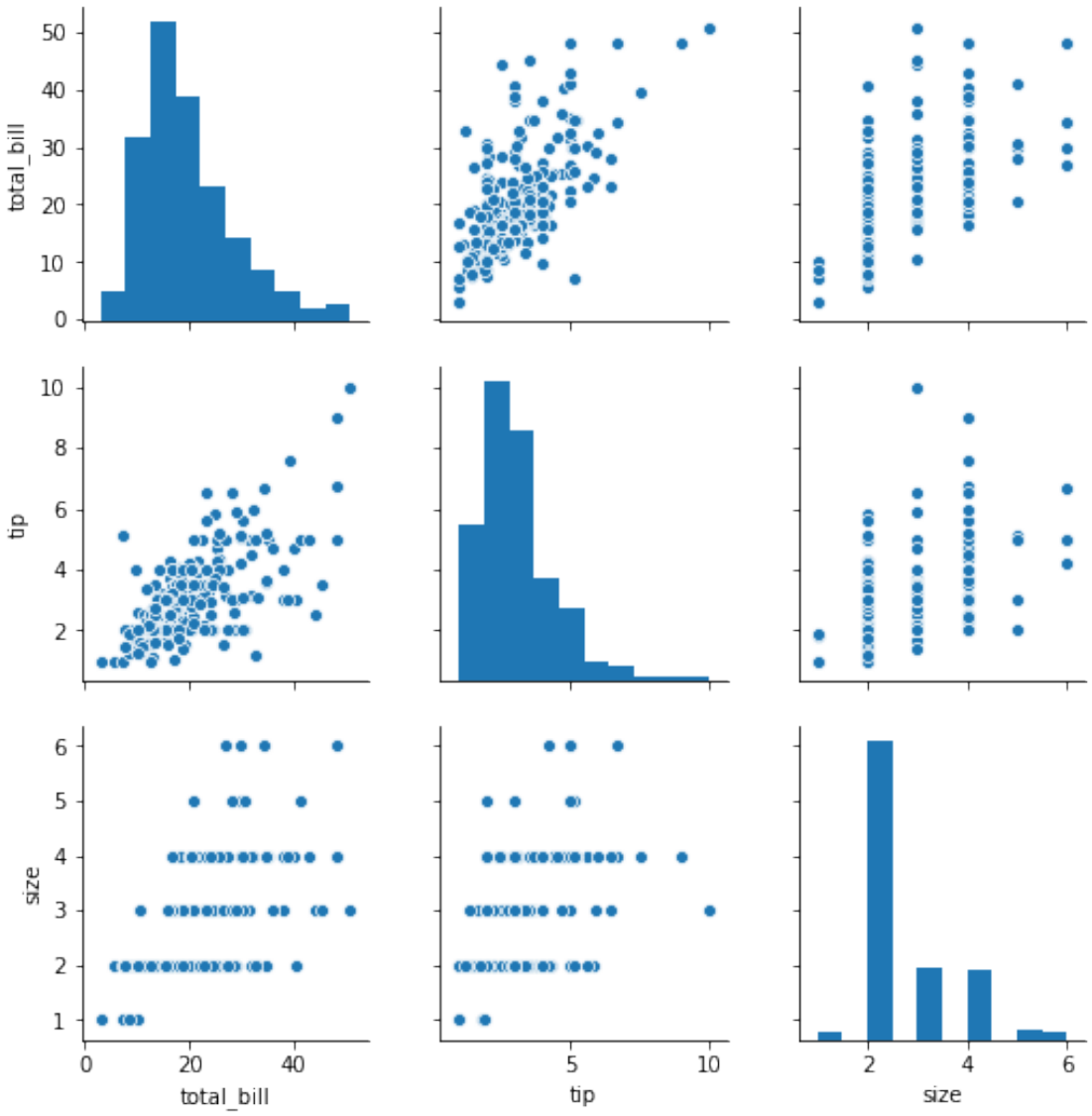
```
C:\python3.6.3\lib\site-packages\scipy\stats\stats.py:1706: FutureWarning: Using
a non-tuple sequence for multidimensional indexing is deprecated; use
`arr[tuple(seq)]` instead of `arr[seq]`. In the future this will be interpreted
as an array index, `arr[np.array(seq)]`, which will result either in an error or
a different result.
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
```

[11]: <seaborn.axisgrid.JointGrid at 0x22a99609f98>



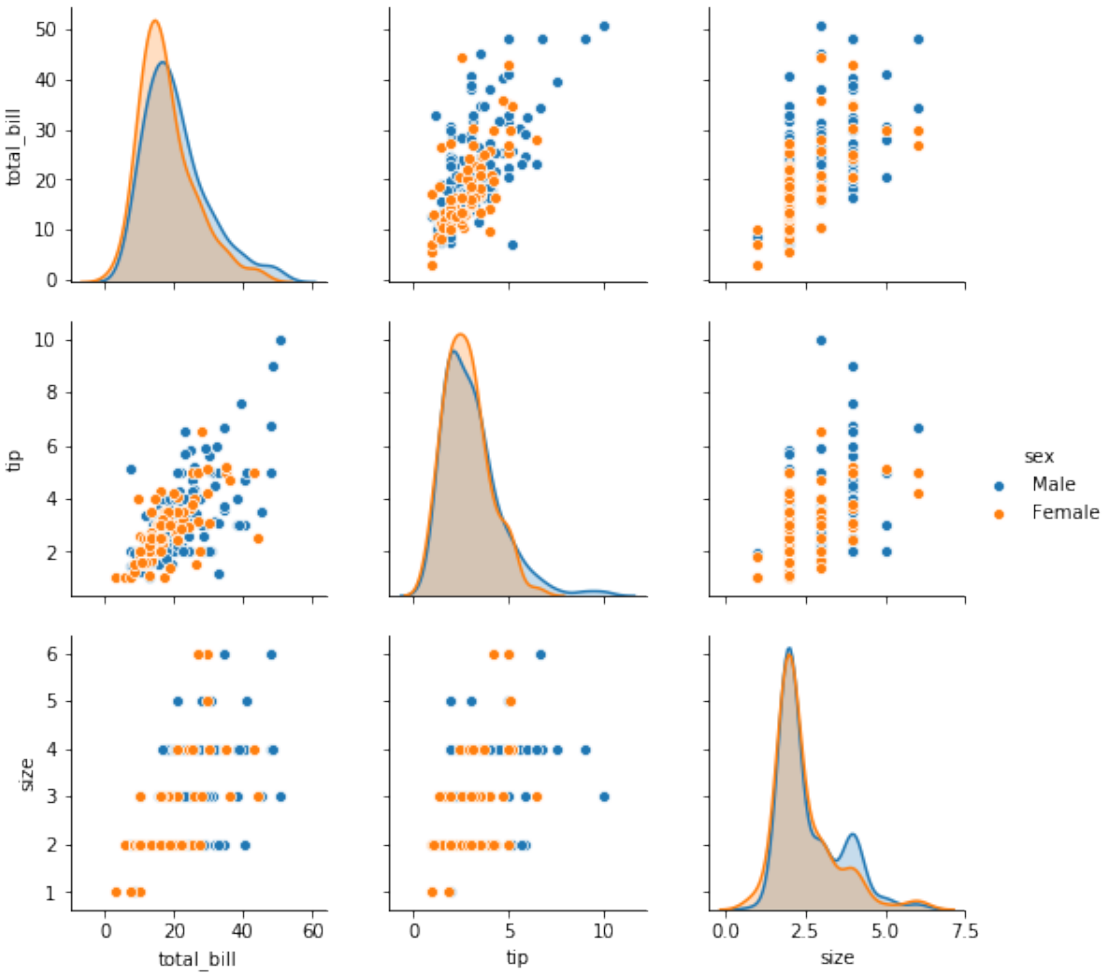
[12]:
```
sns.pairplot(data)
```

[12]: <seaborn.axisgrid.PairGrid at 0x22a9963a710>



[13]:
```
sns.pairplot(data,hue="sex")
```

[13]: <seaborn.axisgrid.PairGrid at 0x22a9b9ac390>

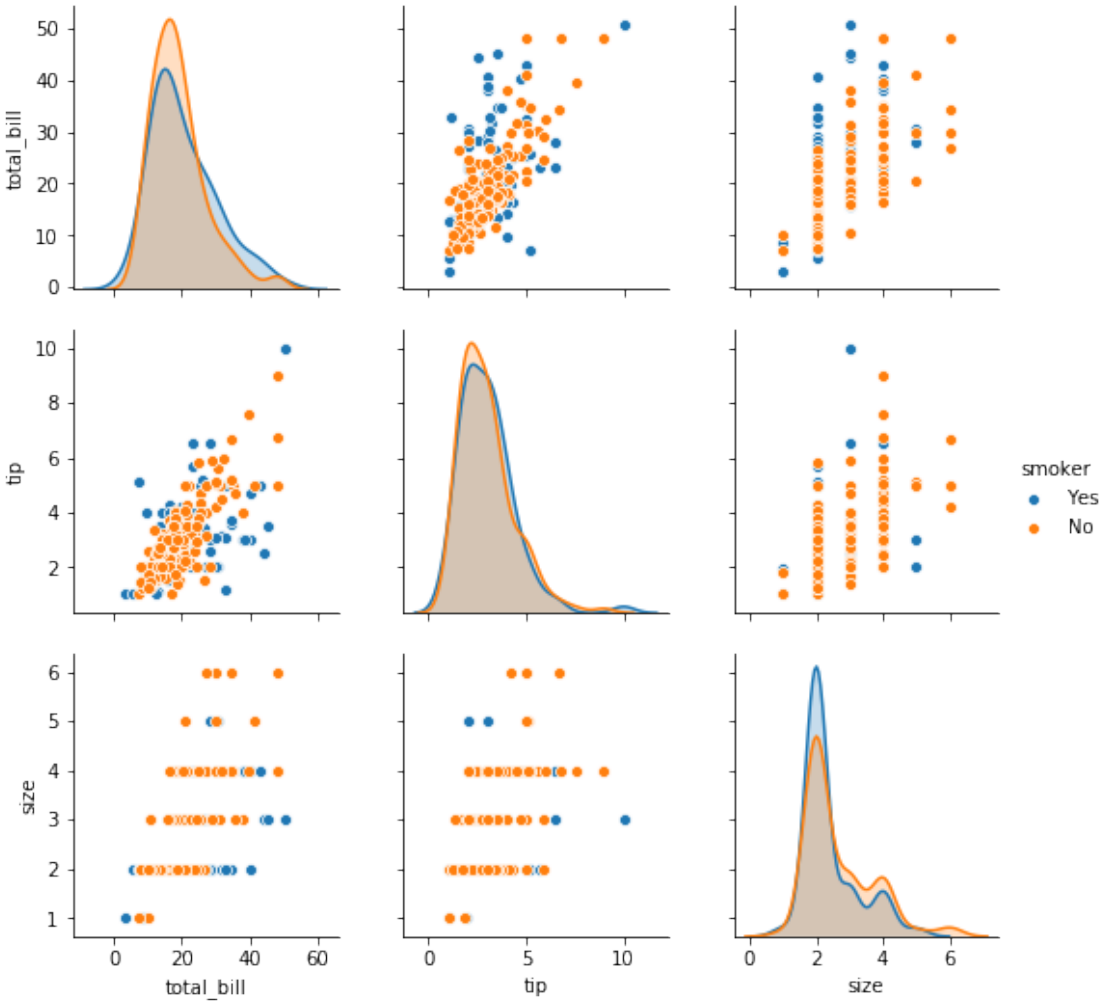


[15]:
```
data['smoker'].value_counts()
```

[15]:
```
No     151
Yes     93
Name: smoker, dtype: int64
```

[16]:
```
sns.pairplot(data,hue="smoker")
```

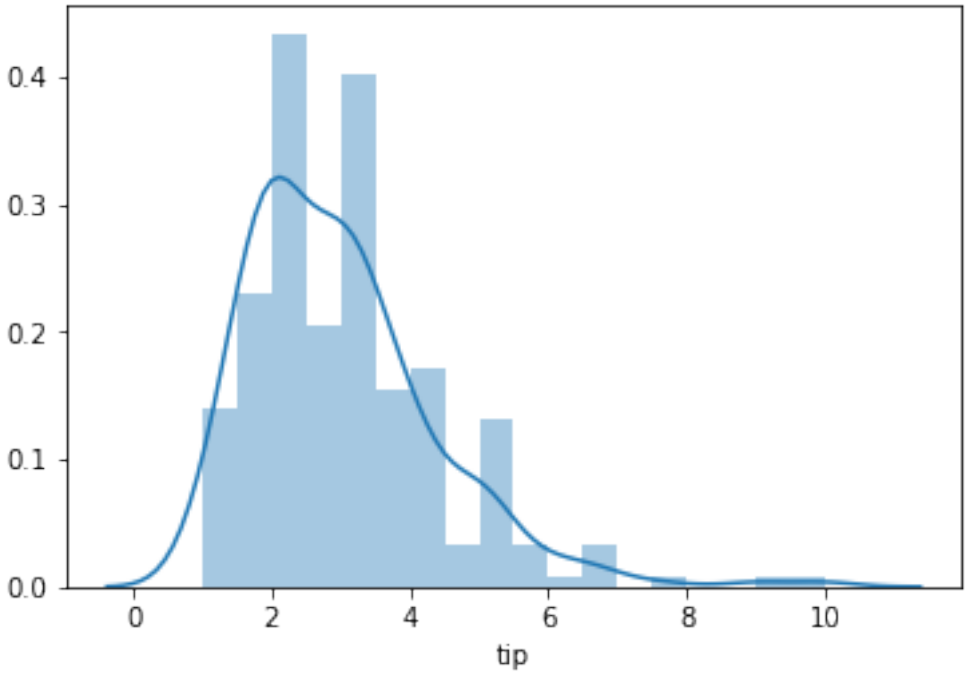
[16]:
```
<seaborn.axisgrid.PairGrid at 0x22a9c46b048>
```



[20]:
```
sns.distplot(data["tip"],kde=True)
```

```
C:\python3.6.3\lib\site-packages\scipy\stats\stats.py:1706: FutureWarning: Using
a non-tuple sequence for multidimensional indexing is deprecated; use
`arr[tuple(seq)]` instead of `arr[seq]`. In the future this will be interpreted
as an array index, `arr[np.array(seq)]`, which will result either in an error or
a different result.
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
```

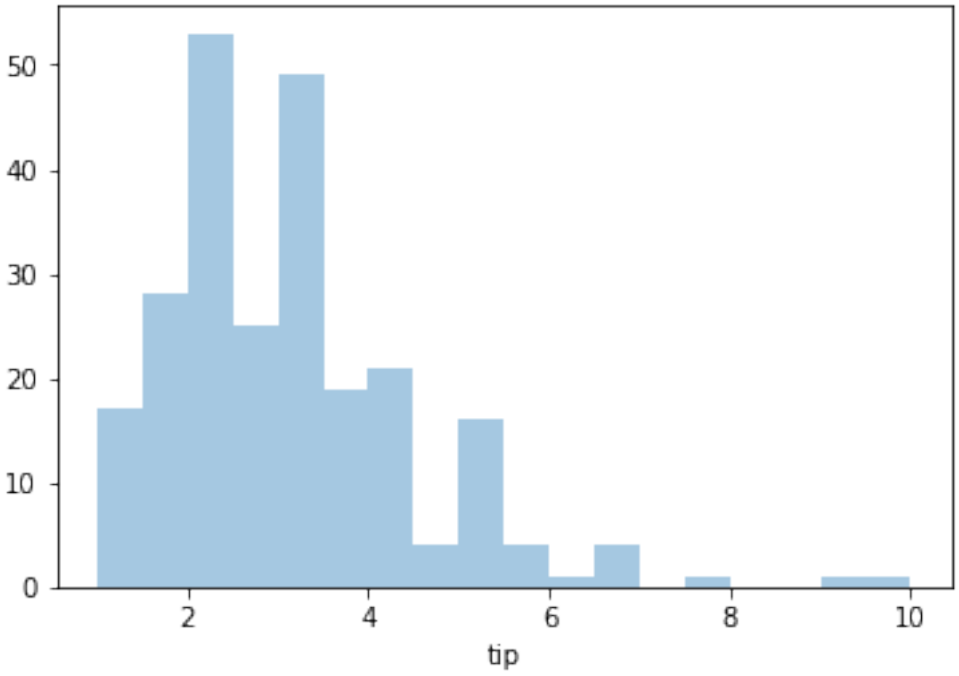
[20]: `<matplotlib.axes._subplots.AxesSubplot at 0x22a9ca266d8>`



[22]:
```
sns.distplot(data["tip"],kde=False)
```

```
C:\python3.6.3\lib\site-packages\scipy\stats\stats.py:1706: FutureWarning: Using
a non-tuple sequence for multidimensional indexing is deprecated; use
`arr[tuple(seq)]` instead of `arr[seq]`. In the future this will be interpreted
as an array index, `arr[np.array(seq)]`, which will result either in an error or
a different result.
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
```

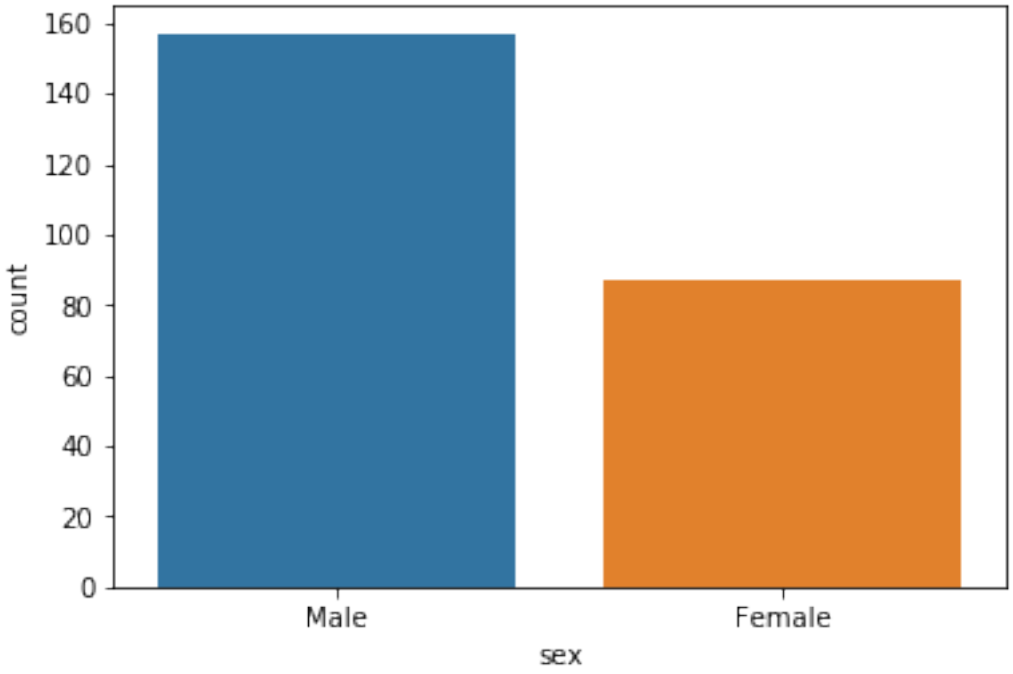
[22]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x22a9dfb2860>
```



1.2 plots for categorical features

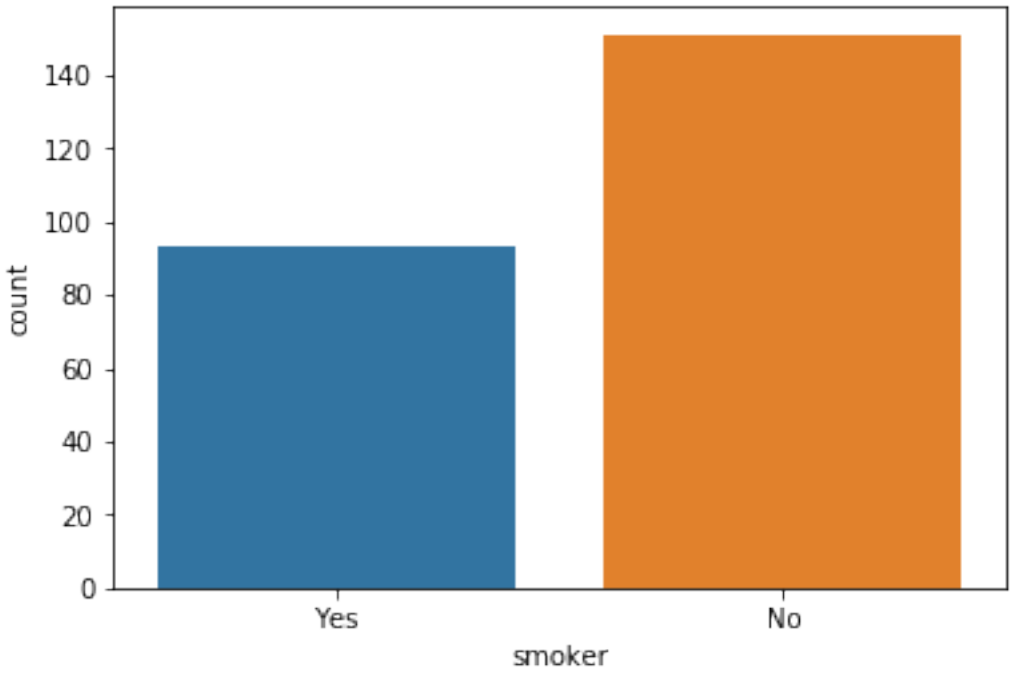
[24]:
```
sns.countplot(data['sex'],data=data)
```

[24]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x22a9dfee668>
```



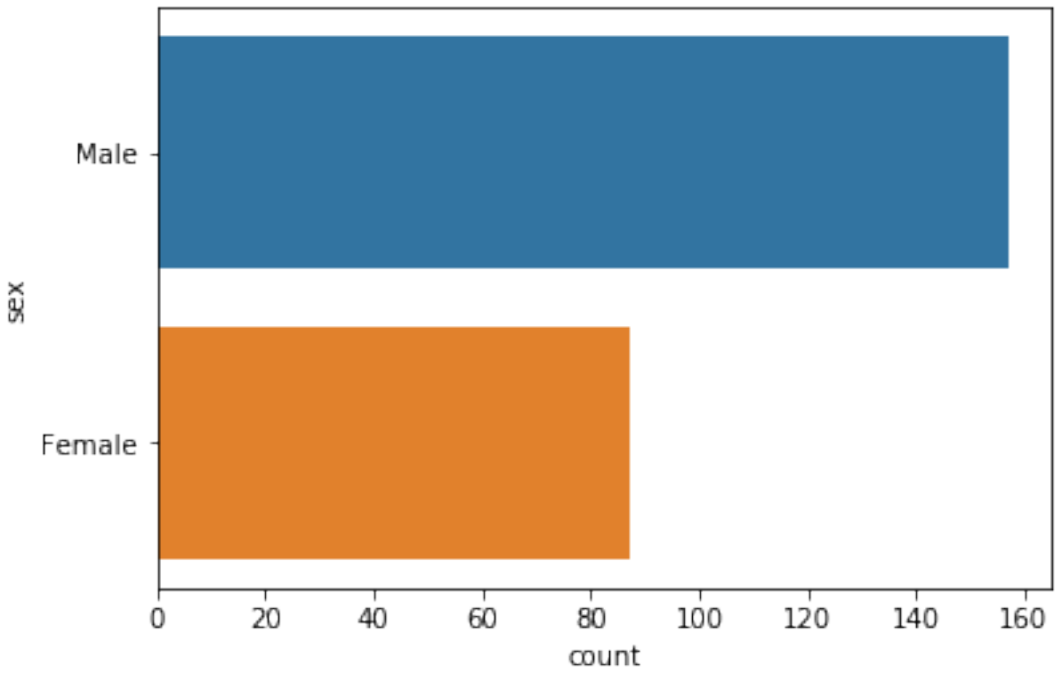
[27]:
```
sns.countplot(data['smoker'],data=data)
```

[27]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x22a9e1587f0>
```



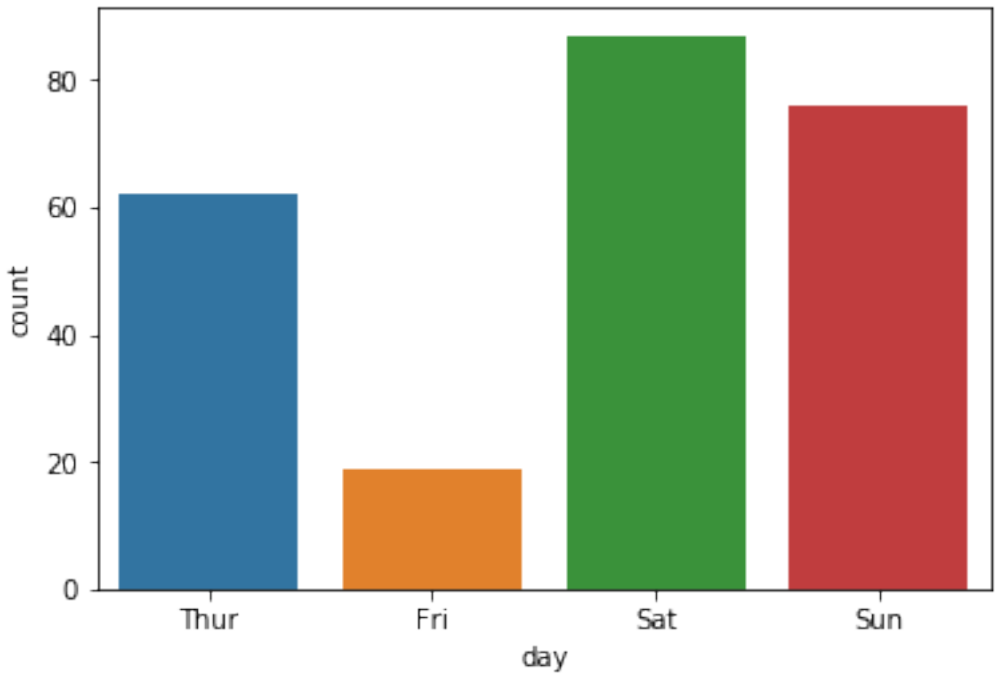
[29]:
```
sns.countplot(y='sex',data=data)
```

[29]: <matplotlib.axes._subplots.AxesSubplot at 0x22a9e1ec198>



[30]:
```
sns.countplot(data['day'],data=data)
```

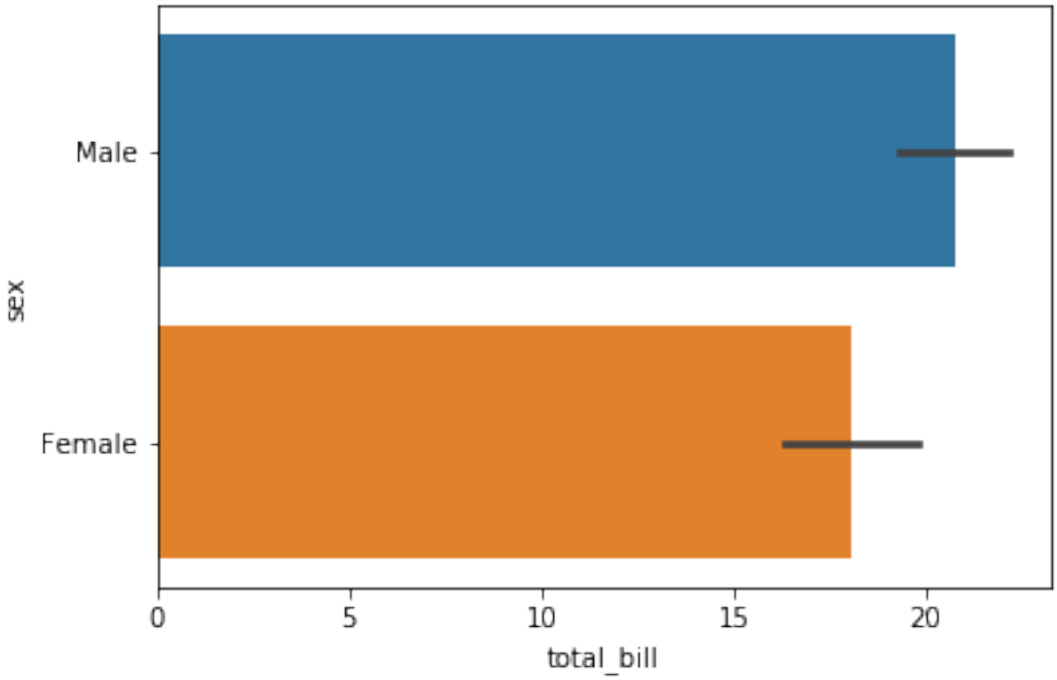
[30]: <matplotlib.axes._subplots.AxesSubplot at 0x22a9e2328d0>



[31]:
```
sns.barplot(x='total_bill',y='sex',data=data)
```

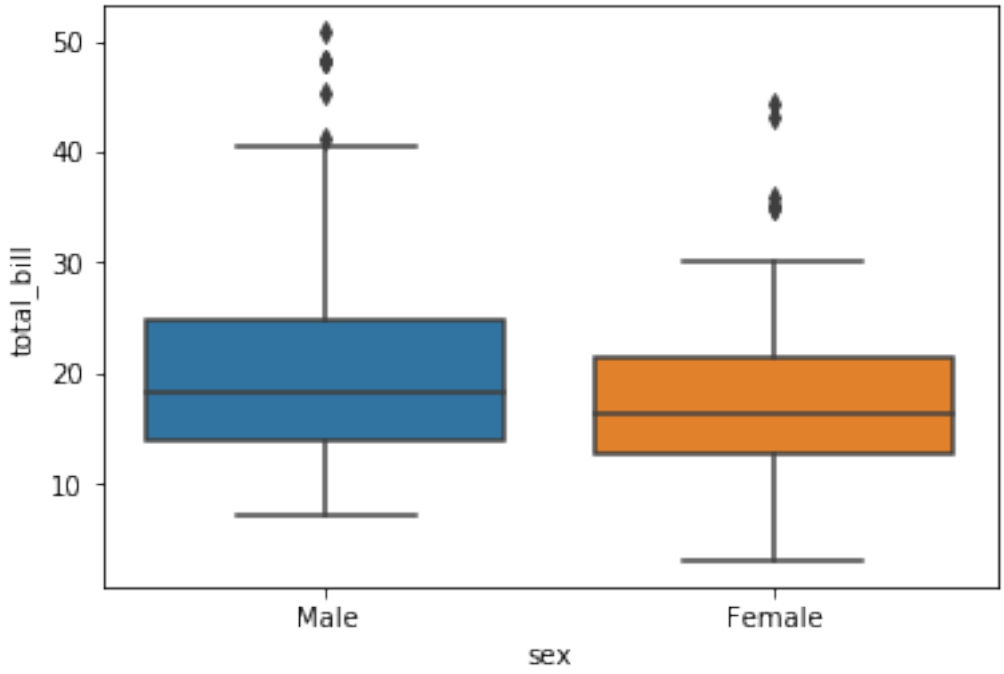
```
C:\python3.6.3\lib\site-packages\scipy\stats\stats.py:1706: FutureWarning: Using
a non-tuple sequence for multidimensional indexing is deprecated; use
`arr[tuple(seq)]` instead of `arr[seq]`. In the future this will be interpreted
as an array index, `arr[np.array(seq)]`, which will result either in an error or
a different result.
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
```

[31]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x22a9e2d26a0>
```



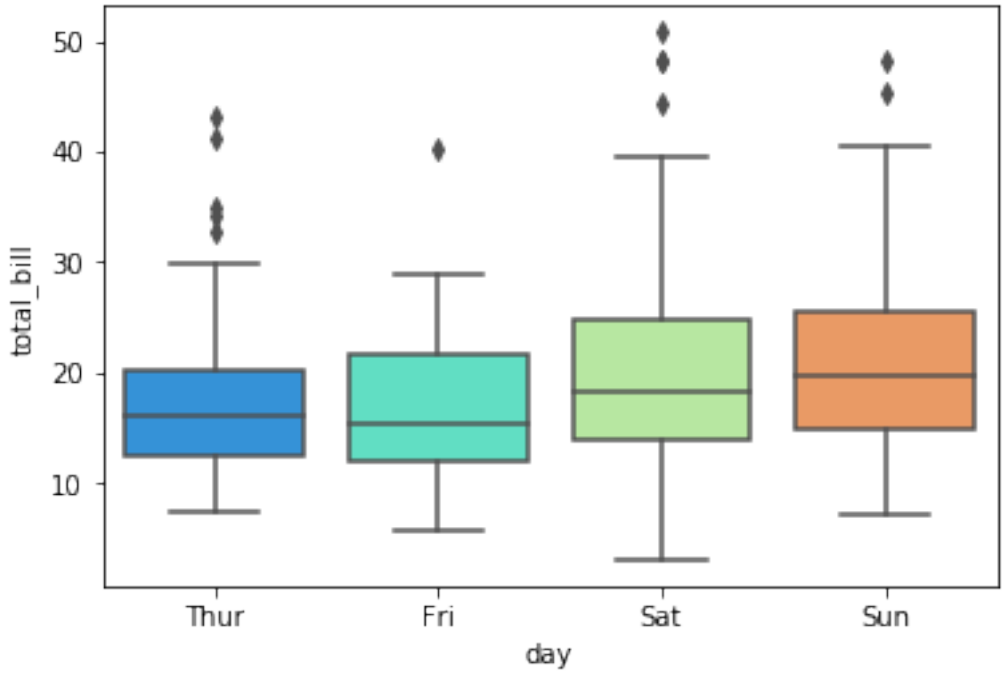
[32]:
```
sns.boxplot(x='sex',y='total_bill',data=data)
```

[32]: <matplotlib.axes._subplots.AxesSubplot at 0x22a9e2a8240>



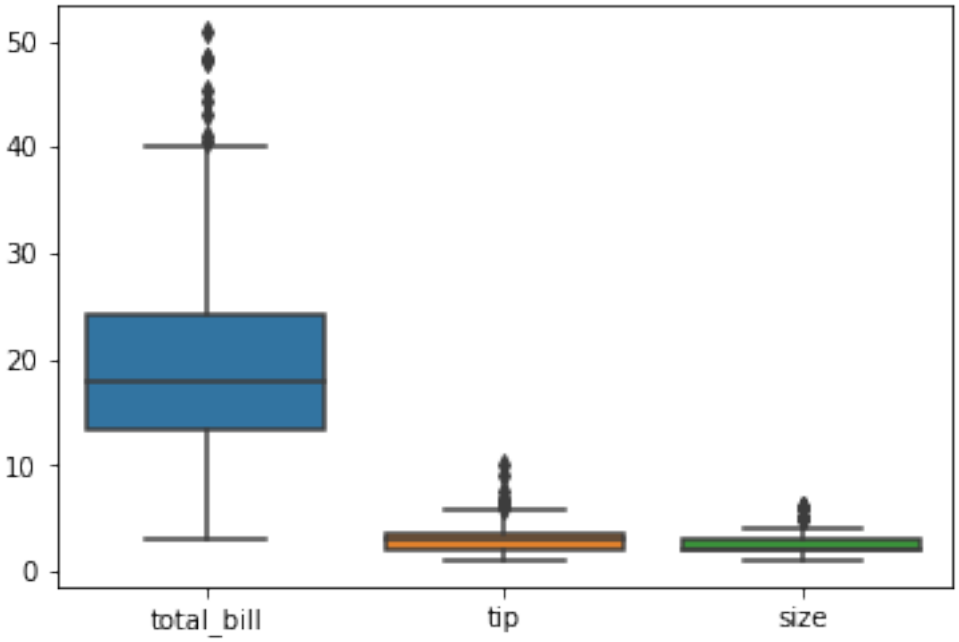
[34]:
```
sns.boxplot(x='day',y='total_bill',data=data,palette='rainbow')
```

[34]: <matplotlib.axes._subplots.AxesSubplot at 0x22a9e3b9b70>



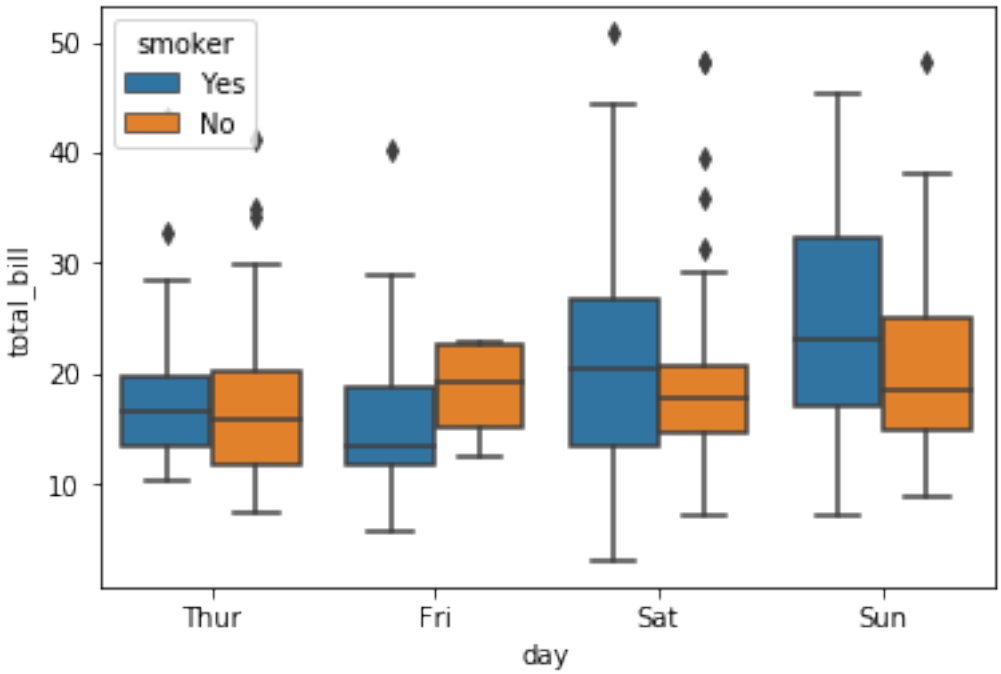
[35]:
```
sns.boxplot(data=data,orient='v')
```

[35]: <matplotlib.axes._subplots.AxesSubplot at 0x22a9e456438>



[36]:
```
sns.boxplot(x='day',y='total_bill',hue='smoker',data=data)
```

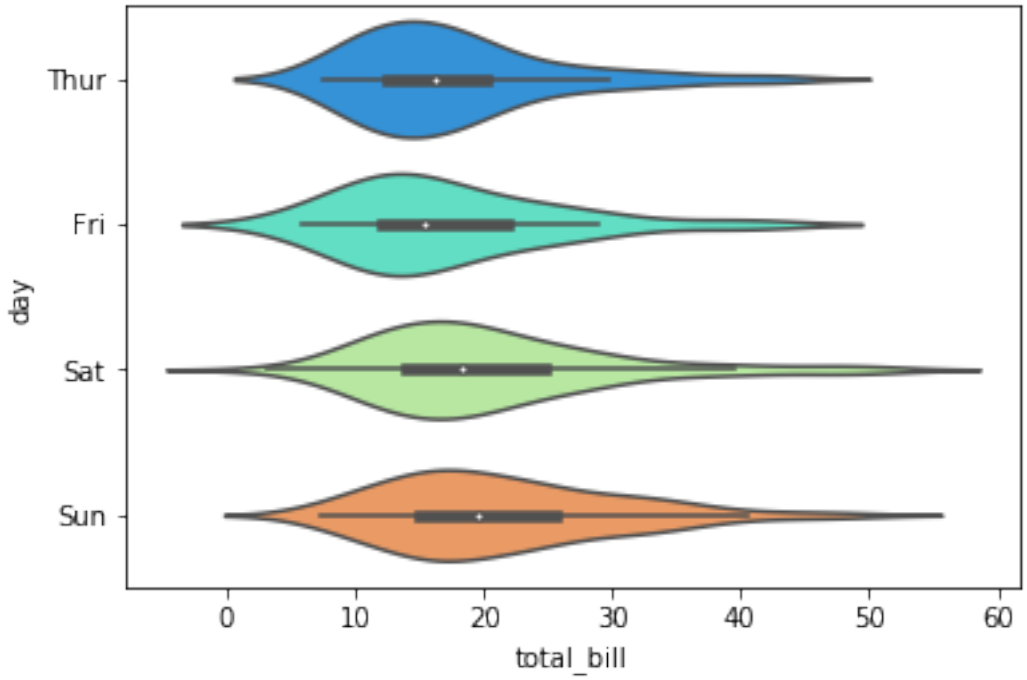
[36]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x22a9e4e0780>
```



[37]:
```
sns.violinplot(x='total_bill',y='day',data=data,palette='rainbow')
```

```
C:\python3.6.3\lib\site-packages\scipy\stats\stats.py:1706: FutureWarning: Using
a non-tuple sequence for multidimensional indexing is deprecated; use
`arr[tuple(seq)]` instead of `arr[seq]`. In the future this will be interpreted
as an array index, `arr[np.array(seq)]`, which will result either in an error or
a different result.
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
```

[37]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x22a9e574cf8>
```



[ ]: