chapter 4. 개별 변수 분석하기 숫자형 변수 - 시각화

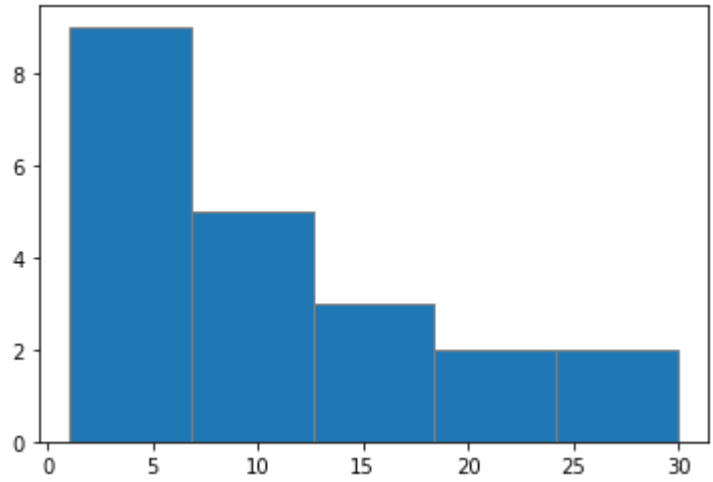
1) 히스토그램

① 아래와 같은 총 21개의 값이 있다고 할 때, 히스토그램을 그리시오.

- 1, 2, 3, 3, 4, 4, 4, 5, 6, 7, 8, 9, 10, 11, 15, 17, 17, 19, 20, 25, 30
- 구간의 개수: 5, edgecolor: grey

In [ ]:

```
import matplotlib.pyplot as plt

a = [1, 2, 3, 3, 4, 4, 4, 5, 6, 7, 8, 9, 10, 11, 15, 17, 17, 19, 20, 25, 30]
plt.hist(a, bins = 5, edgecolor = 'grey')
plt.savefig('a.png')
```

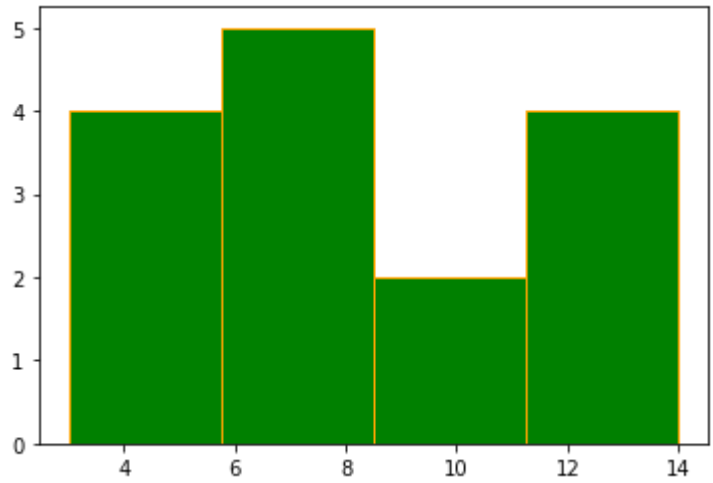


② 아래와 같은 총 15개의 값이 있다고 할 때, 히스토그램을 그리시오.

- 3, 4, 5, 5, 6, 6, 6, 8, 8, 9, 11, 12, 12, 13, 14
- 구간의 개수: 4 , color: green, edgecolor: orange

In [ ]:

```
import matplotlib.pyplot as plt

a = [3, 4, 5, 5, 6, 6, 6, 8, 8, 9, 11, 12, 12, 13, 14]
plt.hist(a, bins = 4, color = 'green', edgecolor = 'orange')
plt.savefig('a.png')
```



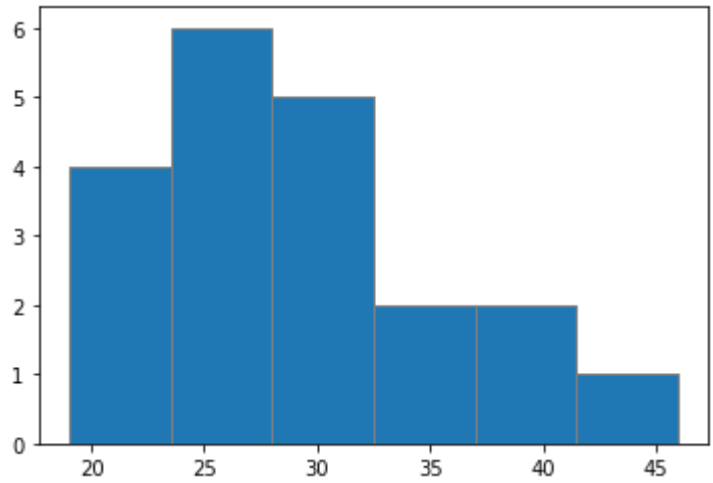
③ 고객의 나이 정보 age가 아래와 같을 때, 히스토그램을 그리고 결과를 저장하여 구간과 빈도수를 확인하시오.

- age = [19, 20, 23, 46, 21, 25, 26, 25, 28, 31, 37, 24, 28, 34, 38, 33, 32, 29, 27, 24]
- 구간의 개수: 6, edgecolor: grey

In [1]:

```
import matplotlib.pyplot as plt

age = [19,20,23,46,21,25,26,25,28,31,37,24,28,34,38,33,32,29,27,24]
hist1 = plt.hist(age, bins = 6, edgecolor = 'grey')
plt.savefig('a.png')

print(hist1)
print('-' * 50)
print('빈도수 : ', hist1[0])
print('구간값 : ', hist1[1])
```

```
(array([4., 6., 5., 2., 2., 1.]), array([19. , 23.5, 28. , 32.5, 37. , 41.5, 46.
]), <a list of 6 Patch objects>)
--------------------------------------------------
빈도수 :  [4. 6. 5. 2. 2. 1.]
구간값 :  [19.  23.5 28.  32.5 37.  41.5 46. ]
```



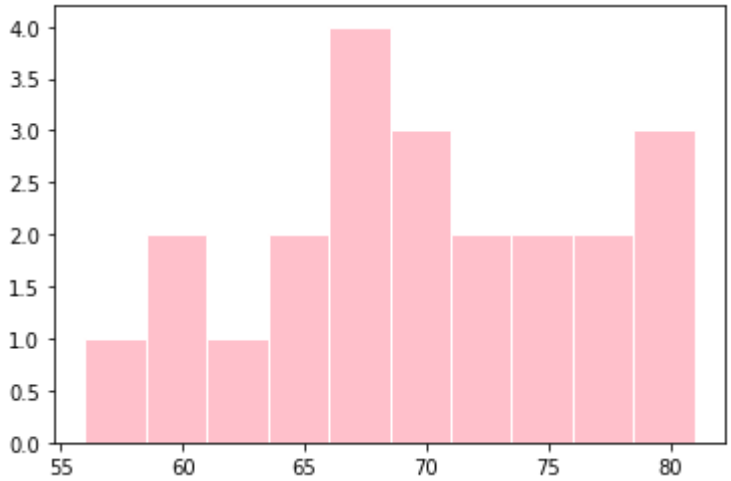
④ 몸무게 데이터 weight가 아래와 같을 때, 히스토그램을 그리고 결과를 저장하여 구간과 빈도수를 확인하시오.

- weight = [68, 81, 64, 56, 78, 74, 61, 77, 66, 68, 59, 71, 80, 59, 67, 81, 69, 73, 69, 74, 70, 65]
- 구간의 개수: 10, color: pink, edgecolor: white

In [2]:

```
import matplotlib.pyplot as plt

weight = [68, 81, 64, 56, 78, 74, 61, 77, 66, 68, 59, 71, 80, 59, 67, 81, 69, 73, 69, 74, 70, 65
]
hist1 = plt.hist(weight, bins = 10, color = 'pink', edgecolor = 'white')
plt.savefig('a.png')

print(hist1)
print('-' * 50)
print('빈도수 : ', hist1[0])
print('구간값 : ', hist1[1])
```

```
(array([1., 2., 1., 2., 4., 3., 2., 2., 2., 3.]), array([56. , 58.5, 61. , 63.5, 6
6. , 68.5, 71. , 73.5, 76. , 78.5, 81. ]), <a list of 10 Patch objects>)
--------------------------------------------------
빈도수 :  [1. 2. 1. 2. 4. 3. 2. 2. 2. 3.]
구간값 :  [56.  58.5 61.  63.5 66.  68.5 71.  73.5 76.  78.5 81. ]
```



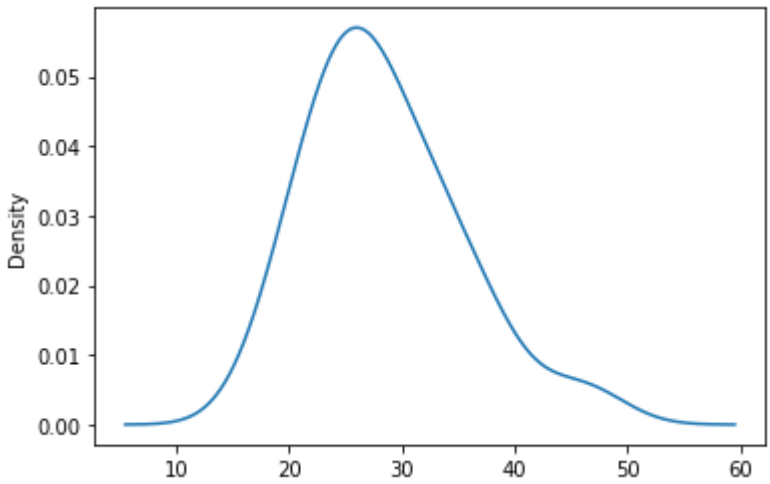
2) 밀도함수 그래프

① 고객의 나이 정보 age가 아래와 같을 때, 밀도함수 그래프를 그리시오.

- age = [19, 20, 23, 46, 21, 25, 26, 25, 28, 31, 37, 24, 28, 34, 38, 33, 32, 29, 27, 24]

In [ ]:

```
import matplotlib.pyplot as plt
import pandas as pd

age = [19,20,23,46,21,25,26,25,28,31,37,24,28,34,38,33,32,29,27,24]
age = pd.Series(age)
age.plot(kind = 'kde')
plt.savefig('a.png')
```

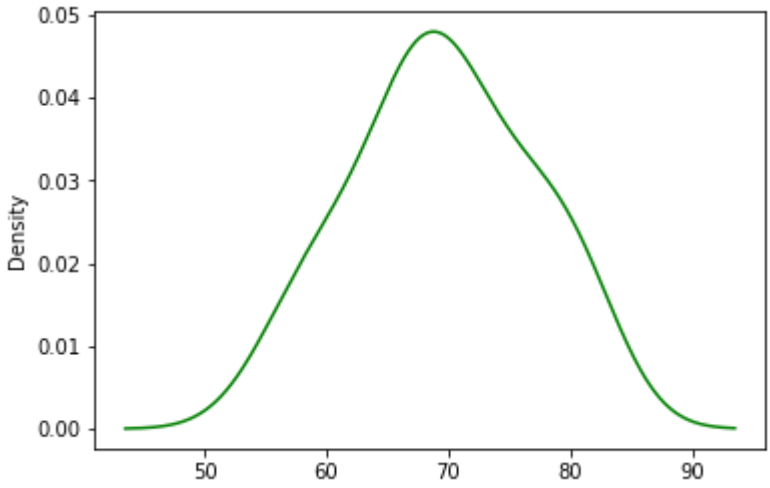


② 몸무게 데이터 weight가 아래와 같을 때, 밀도함수 그래프를 그리시오.

- weight = [68, 81, 64, 56, 78, 74, 61, 77, 66, 68, 59, 71, 80, 59, 67, 81, 69, 73, 69, 74, 70, 65]
- color: green

In [ ]:

```
import matplotlib.pyplot as plt
import pandas as pd

weight = [68, 81, 64, 56, 78, 74, 61, 77, 66, 68, 59, 71, 80, 59, 67, 81, 69, 73, 69, 74, 70, 65
]
weight = pd.Series(weight)
weight.plot(kind = 'kde', color = 'green')
plt.savefig('a.png')
```



3) Boxplot

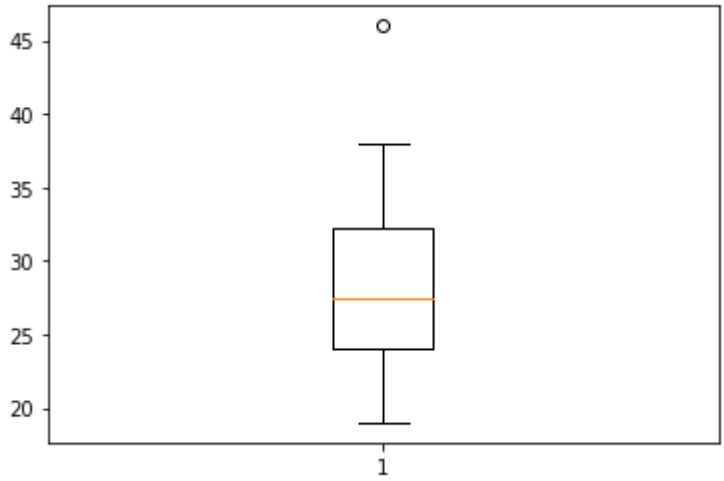
① 고객의 나이 정보 age가 아래와 같을 때, 박스플롯을 세로로 그리고 결과를 저장, 아래/위 수염의 min, max를 구해 봅시다.

- age = [19, 20, 23, 46, 21, 25, 26, 25, 28, 31, 37, 24, 28, 34, 38, 33, 32, 29, 27, 24]

In [4]:

```
import matplotlib.pyplot as plt

age = [19,20,23,46,21,25,26,25,28,31,37,24,28,34,38,33,32,29,27,24]
box1 = plt.boxplot(age)
plt.savefig('a.png')

print(box1['whiskers'])
print(box1['whiskers'][0].get_ydata()) # 아래쪽 수염의 max, min
print(box1['whiskers'][1].get_ydata()) # 위쪽 수염의 min, max
```

```
[<matplotlib.lines.Line2D object at 0x7fd8f18a0ad0>, <matplotlib.lines.Line2D obje
ct at 0x7fd8f18a7050>]
[24. 19.]
[32.25 38.  ]
```



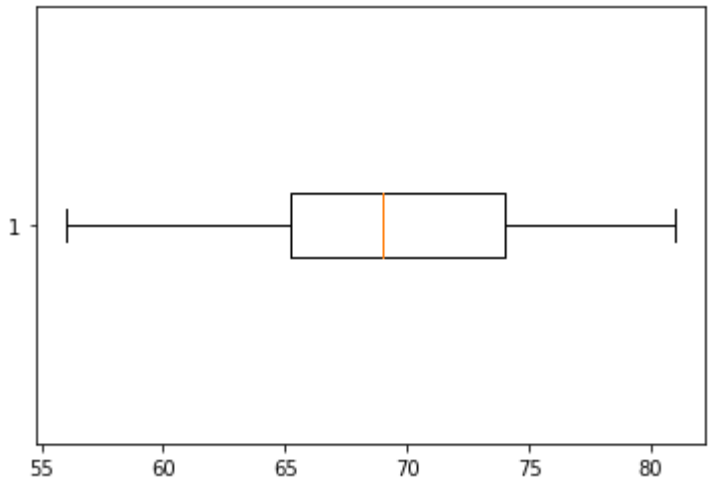
② 몸무게 데이터 weight가 아래와 같을 때, 박스플롯을 가로로 그리고 결과를 저장, 아래/위 수염의 min, max를 구해 봅시다.

- weight = [68, 81, 64, 56, 78, 74, 61, 77, 66, 68, 59, 71, 80, 59, 67, 81, 69, 73, 69, 74, 70, 65]

In [5]:

```
import matplotlib.pyplot as plt

weight = [68, 81, 64, 56, 78, 74, 61, 77, 66, 68, 59, 71, 80, 59, 67, 81, 69, 73, 69, 74, 70, 65
]
box1 = plt.boxplot(weight, vert = False)
plt.savefig('a.png')

print(box1['whiskers'])
print(box1['whiskers'][0].get_xdata()) # 왼쪽 수염의 max, min
print(box1['whiskers'][1].get_xdata()) # 오른쪽 수염의 min, max
```

```
[<matplotlib.lines.Line2D object at 0x7fd8f1893e50>, <matplotlib.lines.Line2D obje
ct at 0x7fd8f18973d0>]
[65.25 56.  ]
[74. 81.]
```



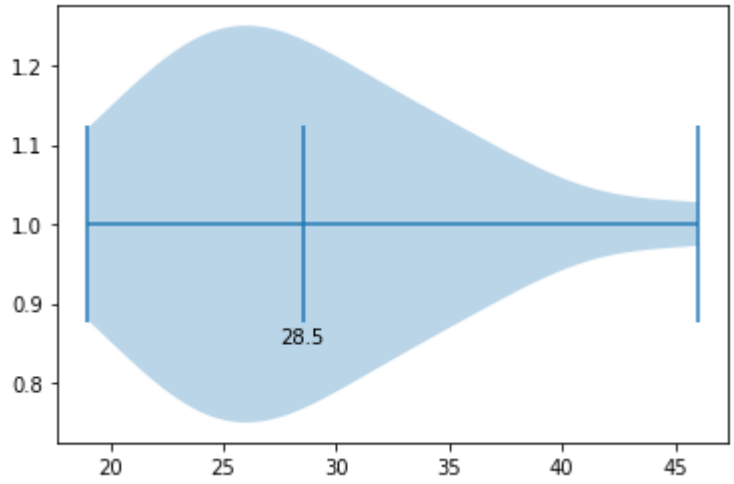
4) 바이올린 플롯

① 고객의 나이 정보 age가 아래와 같을 때, 바이올린 플롯을 가로로 그리고 평균값을 텍스트로 나타내시오.

- age = [19, 20, 23, 46, 21, 25, 26, 25, 28, 31, 37, 24, 28, 34, 38, 33, 32, 29, 27, 24]

In [ ]:

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

age = [19,20,23,46,21,25,26,25,28,31,37,24,28,34,38,33,32,29,27,24]
age_m = np.mean(age)

plt.violinplot(age, vert = False, showmeans=True)
plt.text( age_m - 1, 0.85 , age_m)
plt.savefig('a.png')
```

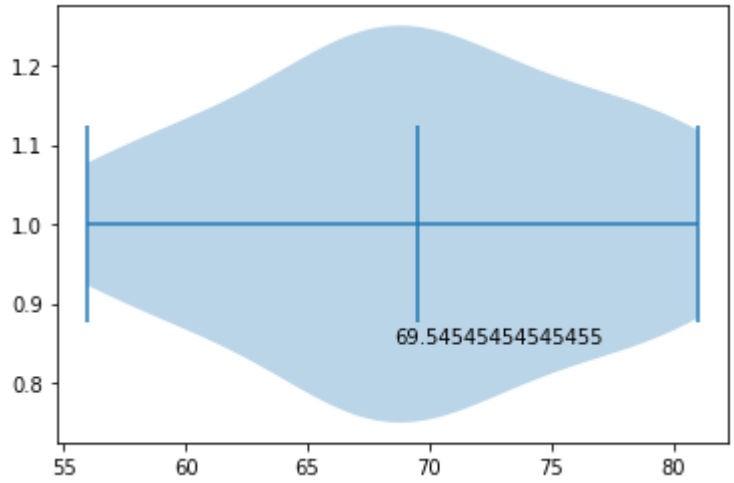


② 몸무게 데이터 weight가 아래와 같을 때, 바이올린 플롯을 세로로 그리고 평균값을 텍스트로 나타내시오.

- weight = [68, 81, 64, 56, 78, 74, 61, 77, 66, 68, 59, 71, 80, 59, 67, 81, 69, 73, 69, 74, 70, 65]

In [ ]:

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

weight = [68, 81, 64, 56, 78, 74, 61, 77, 66, 68, 59, 71, 80, 59, 67, 81, 69, 73, 69, 74, 70, 65
]
weight_m = np.mean(weight)

plt.violinplot(weight, vert=False, showmeans=True)
plt.text(weight_m - 1, 0.85 , weight_m)
plt.savefig('a.png')
```



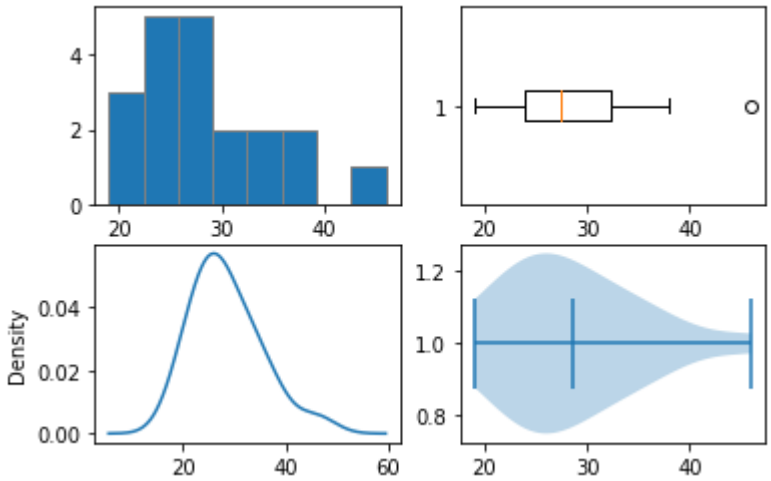
5) 그래프 비교하기

① 고객의 나이 정보 age의 값들을 활용하여 앞에서 다루었던 네 가지 그래프(히스토그램, 밀도 함수 그래프, 박스플롯, 바이올린 플롯)을 주어진 결과에 맞게 한꺼번에 그리고 비교해 봅시다.

- age = [19,20,23,46,21,25,26,25,28,31,37,24,28,34,38,33,32,29,27,24]
- 힌트: subplot() 사용

In [ ]:

```
import matplotlib.pyplot as plt
import pandas as pd

age = [19,20,23,46,21,25,26,25,28,31,37,24,28,34,38,33,32,29,27,24]

plt.subplot(2,2,1)
plt.hist(age, bins = 8, edgecolor = 'grey')

plt.subplot(2,2,2)
plt.boxplot(age, vert = False)

plt.subplot(2,2,3)
pd.Series(age).plot(kind = 'kde')

plt.subplot(2,2,4)
plt.violinplot(age, vert = False, showmeans=True)

plt.savefig('a.png')
```

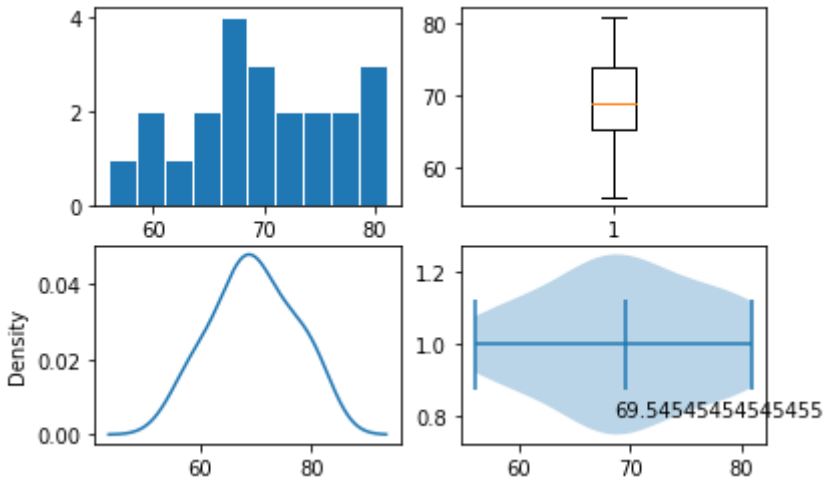


② 몸무게 데이터 weight를 활용하여 앞에서 다루었던 네 가지 그래프(히스토그램, 밀도 함수 그래프, 박스플롯, 바이올린 플롯)을 주어진 결과에 맞게 한꺼번에 그리고 비교해 봅시다.

- weight = [68, 81, 64, 56, 78, 74, 61, 77, 66, 68, 59, 71, 80, 59, 67, 81, 69, 73, 69, 74, 70, 65]
- 힌트: subplot() 사용

In [ ]:

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

weight = [68, 81, 64, 56, 78, 74, 61, 77, 66, 68, 59, 71, 80, 59, 67, 81, 69, 73, 69, 74, 70, 65
]

plt.subplot(2,2,1)
plt.hist(weight, bins = 10, edgecolor = 'white')

plt.subplot(2,2,2)
plt.boxplot(weight, vert = True)

plt.subplot(2,2,3)
pd.Series(weight).plot(kind = 'kde')

plt.subplot(2,2,4)
plt.violinplot(weight, vert = False, showmeans=True)
plt.text(np.mean(weight) - 1, 0.8, np.mean(weight))

plt.savefig('a.png')
```



6) 데이터 프레임의 숫자형 변수 시각화

① Titanic

- 데이터셋 : titanic3
- 설명 : NaN 조치된 Titanic
- url : https://bit.ly/3HaMAtZ (https://bit.ly/3HaMAtZ)

(1) 데이터를 불러와서 상위 5개 행을 조회하시오.

In [ ]:

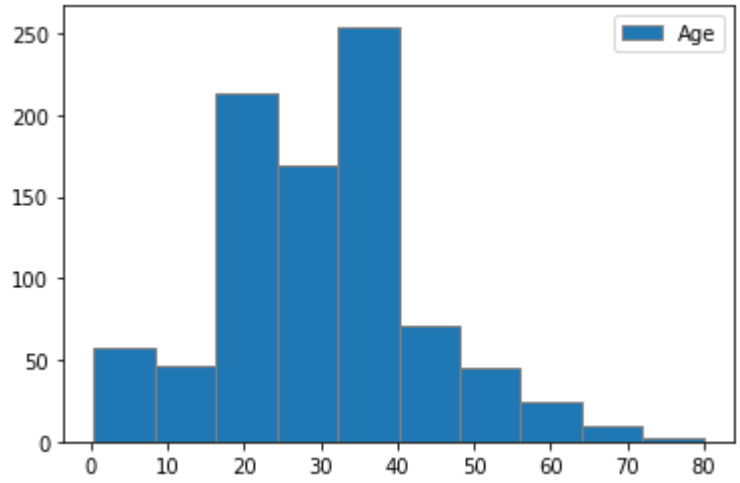
```
import pandas as pd
titanic = pd.read_csv('https://bit.ly/3HaMAtZ')
print(titanic.head())
```

```
   Survived  Pclass     Sex   Age  ...  Embarked  AgeGroup  Family Age_scale1
0         0       3    male  22.0  ...         S  Age21_30       2   0.271174
1         1       1  female  38.0  ...         C  Age31_40       2   0.472229
2         1       3  female  26.0  ...         S  Age21_30       1   0.321438
3         1       1  female  35.0  ...         S  Age31_40       2   0.434531
4         0       3    male  35.0  ...         S  Age31_40       1   0.434531

[5 rows x 11 columns]
```

(2) 나이(Age)의 분포를 확인하기 위한 히스토그램을 주어진 결과에 맞게 그리시오.

In [ ]:

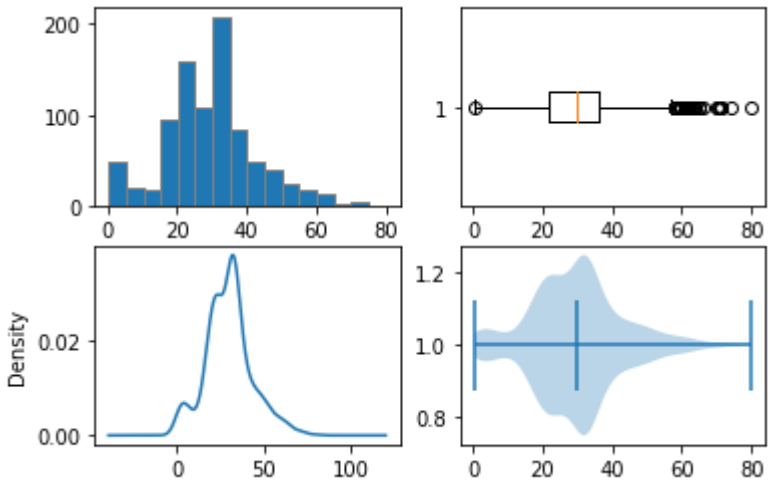
```
import matplotlib.pyplot as plt
import pandas as pd

titanic = pd.read_csv('https://bit.ly/3HaMAtZ')

plt.hist(titanic['Age'], bins = 10, edgecolor = 'grey', label = 'Age')
plt.legend()
plt.savefig('a.png')
```



(3) 나이(Age)의 분포를 확인하기 위한 네 가지 그래프(히스토그램, 밀도 함수 그래프, 박스플롯, 바이올린 플롯)을 주어진 결과에 맞게 한꺼번에 그리고 비교해 봅시다.

In [ ]:

```
import pandas as pd
import matplotlib.pyplot as plt

titanic = pd.read_csv('https://bit.ly/3HaMAtZ')

plt.subplot(2,2,1)
plt.hist(titanic['Age'], bins = 16, edgecolor = 'grey')

plt.subplot(2,2,2)
plt.boxplot(titanic['Age'], vert = False)

plt.subplot(2,2,3)
titanic['Age'].plot(kind = 'kde')

plt.subplot(2,2,4)
plt.violinplot(titanic['Age'], vert = False, showmeans=True)

plt.savefig('a.png')
```



(4) 위 분포로 부터 알수 있는 것은 무엇인가요?

더 살펴보고 싶은 부분은 무엇인가요?

② New York Air Quality

- 데이터셋 : airquality
- 설명 : 뉴욕 공기오염도
- url : https://bit.ly/3qmthqZ (https://bit.ly/3qmthqZ)

(1) 데이터를 불러와서 상위 5개 행을 조회하시오.

In [ ]:

```
import pandas as pd
air = pd.read_csv('https://bit.ly/3qmthqZ')
print(air.head())
```

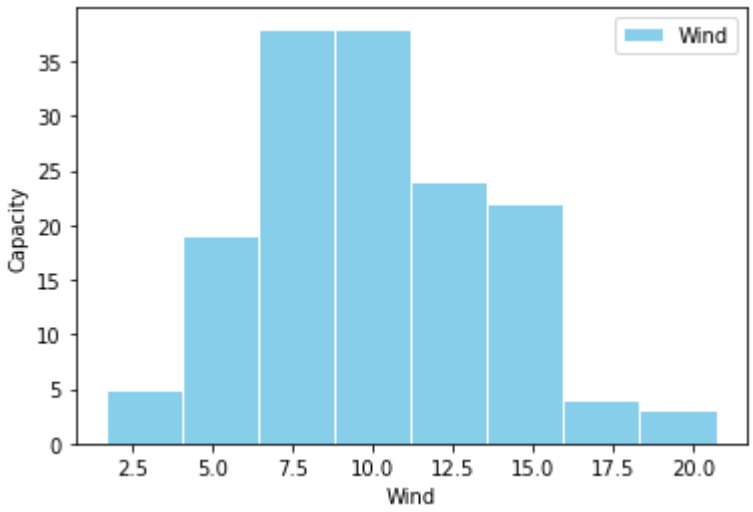
```
   Ozone  Solar.R  Wind  Temp        Date
0     41    190.0   7.4    67  1973-05-01
1     36    118.0   8.0    72  1973-05-02
2     12    149.0  12.6    74  1973-05-03
3     18    313.0  11.5    62  1973-05-04
4     19      NaN  14.3    56  1973-05-05
```

(2) 풍량(Wind)의 분포를 확인하기 위한 히스토그램을 주어진 결과에 맞게 그리시오.

- color: skyblue
- edge color: white

In [ ]:

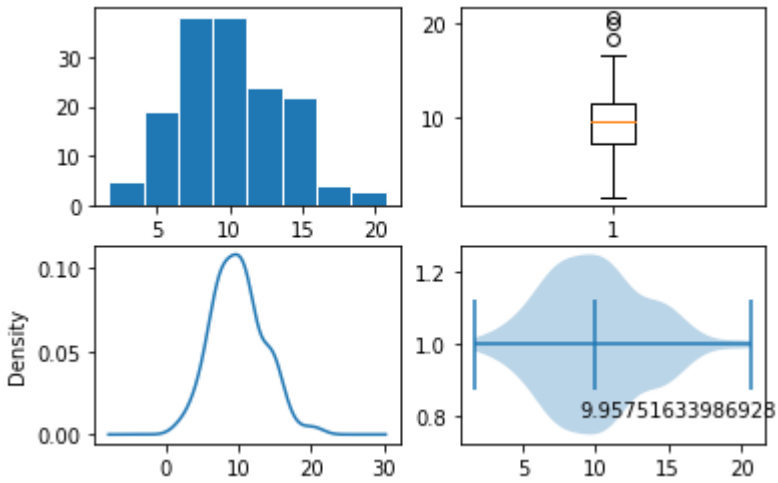
```
import pandas as pd
import matplotlib.pyplot as plt

air = pd.read_csv('https://bit.ly/3qmthqZ')
plt.hist(air['Wind'], bins = 8, color = 'skyblue', edgecolor = 'white', label = 'Wind')

plt.xlabel('Wind')
plt.ylabel('Capacity')
plt.legend()
plt.savefig('a.png')
```



(3) 풍량(wind)의 분포를 확인하기 위한 네 가지 그래프(히스토그램, 밀도 함수 그래프, 박스플롯, 바이올린 플롯)을 주어진 결과에 맞게 한꺼번에 그리고 비교해 봅시다.

In [ ]:

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

air = pd.read_csv('https://bit.ly/3qmthqZ')

plt.subplot(2,2,1)
plt.hist(air['Wind'], bins = 8, edgecolor = 'white')

plt.subplot(2,2,2)
plt.boxplot(air['Wind'])

plt.subplot(2,2,3)
air['Wind'].plot(kind = 'kde')

plt.subplot(2,2,4)
plt.violinplot(air['Wind'], vert = False, showmeans=True)
plt.text(np.mean(air['Wind']) - 1, 0.8, np.mean(air['Wind']))

plt.savefig('a.png')
```



(4) 위 분포로 부터 알수 있는 것은 무엇인가요?

더 살펴보고 싶은 부분은 무엇인가요?