chapter 5. 개별 변수 분석하기 범주형 변수

1) 범주형 변수의 기초 통계량

① 성별 데이터가 아래와 같을 때, count를 이용하여 범주별 빈도수와 비율을 확인하시오.

- gender = ['F','M','F','F','F','M','F','M','M']
- 비율 = 범주별 빈도 / 전체 개수

In [ ]:

```
gender = ['F','M','F','F','F','M','F','M','M']

f_cnt = gender.count('F')
m_cnt = gender.count('M')
total_cnt = len(gender)

print('F', f_cnt, f_cnt/total_cnt)
print('M', m_cnt, m_cnt/total_cnt)
```

```
F 5 0.5555555555555556
M 4 0.4444444444444444
```

② 이번에는 pandas의 .value_counts()를 이용하여 범주별 빈도수와 비율을 구해 봅시다.

- gender = ['F','M','F','F','F','M','F','M','M']

In [ ]:

```
import pandas as pd

gender = ['F','M','F','F','F','M','F','M','M']
gender = pd.Series(gender)

print(gender.value_counts())
print(gender.value_counts()/len(gender))
```

```
F    5
M    4
dtype: int64
F    0.555556
M    0.444444
dtype: float64
```

③ 고객 선호 색상 데이터가 아래와 같을 때, 범주별 빈도수와 비율을 구해 봅시다.

- color = ['red','red', 'blue', 'green', 'green', 'red', 'blue', 'blue', 'red', 'green', 'red', 'blue']

In [ ]:

```
import pandas as pd

color = ['red','red', 'blue', 'green', 'green', 'red', 'blue', 'blue', 'red', 'green', 'red', 'b
lue']
color = pd.Series(color)

print(color.value_counts())
print(color.value_counts()/len(color))
```

```
red      5
blue     4
green    3
dtype: int64
red      0.416667
blue     0.333333
green    0.250000
dtype: float64
```

2) 데이터 프레임으로부터 범주형 변수 기초통계량

① Titanic

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

(1) 데이터를 불러와 봅시다.

In [1]:

```
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) 티켓 클래스(Pclass)의 범주별 빈도수와 범주별 비율을 구하시오.

In [2]:

```
import pandas as pd

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

print(titanic['Pclass'].value_counts())
print(titanic['Pclass'].value_counts()/len(titanic['Pclass']))
```

```
3    491
1    216
2    184
Name: Pclass, dtype: int64
3    0.551066
1    0.242424
2    0.206510
Name: Pclass, dtype: float64
```

② Diamond

- 데이터셋 : diamond
- 설명 : 다이아몬드 가격
- url : https://bit.ly/3eHdRrH (https://bit.ly/3eHdRrH)

(1) 데이터를 불러와 봅시다.

In [5]:

```
import pandas as pd

diamond = pd.read_csv('https://bit.ly/3eHdRrH')
print(diamond.head())
```

```
   carat      cut color clarity  depth  table  price     x     y     z
0   0.23    Ideal     E     SI2   61.5   55.0    326  3.95  3.98  2.43
1   0.21  Premium     E     SI1   59.8   61.0    326  3.89  3.84  2.31
2   0.23     Good     E     VS1   56.9   65.0    327  4.05  4.07  2.31
3   0.29  Premium     I     VS2   62.4   58.0    334  4.20  4.23  2.63
4   0.31     Good     J     SI2   63.3   58.0    335  4.34  4.35  2.75
```

(2) color의 범주별 빈도수와 범주별 비율을 구하시오.

In [6]:

```
import pandas as pd

diamond = pd.read_csv('https://bit.ly/3eHdRrH')

print(diamond['color'].value_counts())
print(diamond['color'].value_counts()/len(diamond['color']))
```

```
G    11292
E     9797
F     9542
H     8304
D     6775
I     5422
J     2808
Name: color, dtype: int64
G    0.209344
E    0.181628
F    0.176900
H    0.153949
D    0.125603
I    0.100519
J    0.052058
Name: color, dtype: float64
```

(3) 가공품질(cut)의 범주별 빈도수와 범주별 비율을 구하시오.

In [7]:

```
import pandas as pd

diamond = pd.read_csv('https://bit.ly/3eHdRrH')

print(diamond['cut'].value_counts())
print(diamond['cut'].value_counts()/len(diamond['cut']))
```

```
Ideal        21551
Premium      13791
Very Good    12082
Good          4906
Fair          1610
Name: cut, dtype: int64
Ideal        0.399537
Premium      0.255673
Very Good    0.223990
Good         0.090953
Fair         0.029848
Name: cut, dtype: float64
```

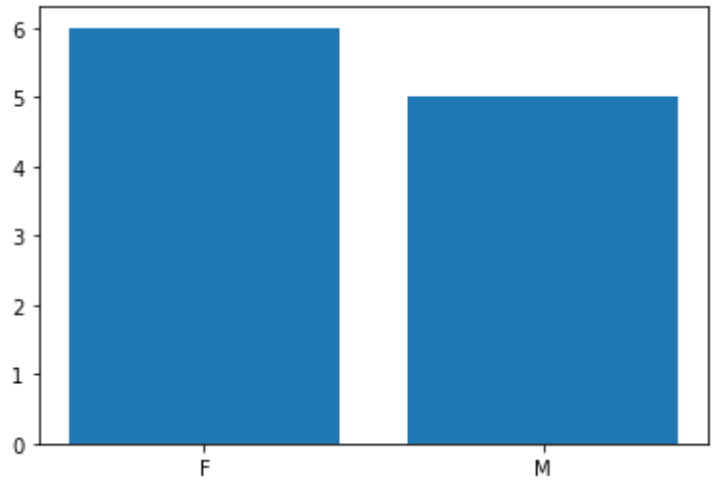
3) 범주형 변수의 시각화 - bar chart

① 다음의 조건에 맞게 bar chart를 그려 봅시다.

- 범주 이름: gender = ['F', 'M']
- 값: cnt = [6, 5]

In [ ]:

```
import matplotlib.pyplot as plt

gender = ['F','M']
cnt = [6, 5]

plt.bar(gender, cnt)
plt.savefig('a.png')
```

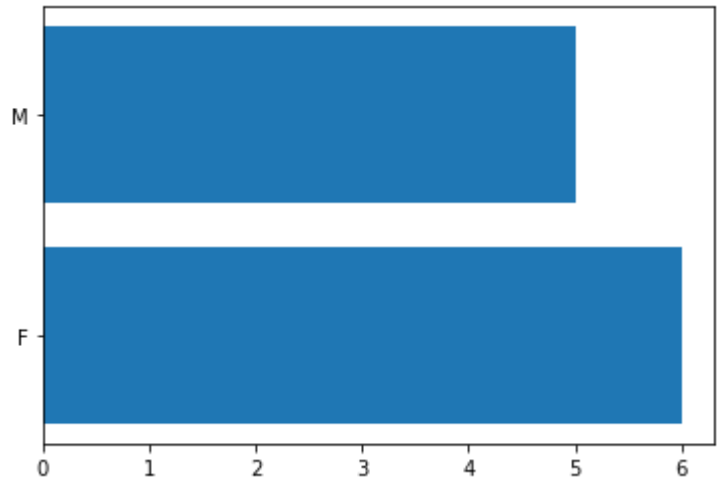


② 다음의 조건에 맞게 bar chart를 그려 봅시다.

- 범주 이름: gender = ['F', 'M']
- 값: cnt = [6, 5]
- 수평 막대 그래프

In [ ]:

```
import matplotlib.pyplot as plt

gender = ['F','M']
cnt = [6, 5]

plt.barh(gender, cnt)
plt.savefig('a.png')
```



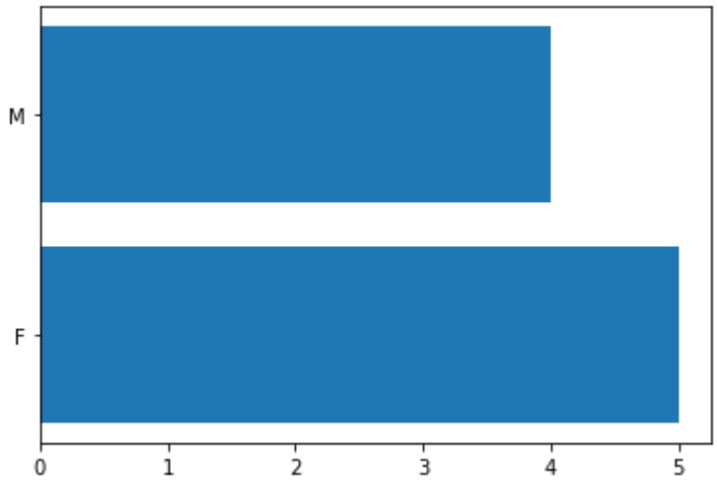
③ 성별 데이터가 아래와 같을 때, value_counts()를 사용하여 빈도수를 집계하고 주어진 결과에 맞게 bar chart를 그리시오.

- gender = ['F','M','F','F','F','M','F','M','M']

In [ ]:

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

gender = ['F','M','F','F','F','M','F','M','M']
gender = pd.Series(gender)

cnt = gender.value_counts()
print(cnt.values)
print(cnt.index)

plt.barh(cnt.index, cnt.values)
plt.savefig('a.png')
```

```
[5 4]
Index(['F', 'M'], dtype='object')
```

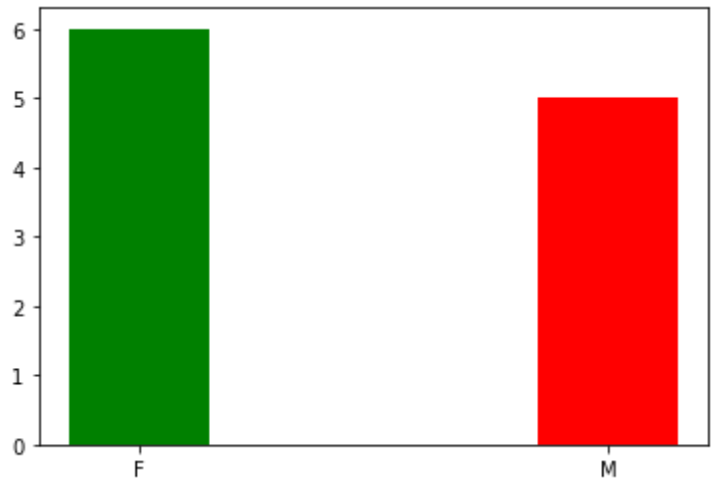


④ ①에서 만든 그래프를 주어진 결과에 맞게 수정해 봅시다.

- width: 0.3

In [ ]:

```
import matplotlib.pyplot as plt

gender = ['F','M']
cnt = [6, 5]

plt.bar(gender, cnt, color = ['g','r'], width = .3)
plt.savefig('a.png')
```



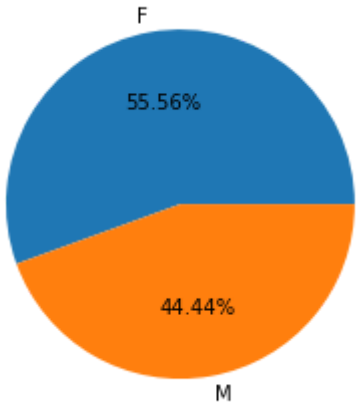
4) 범주형 변수의 시각화 - pie chart

① 성별 데이터가 아래와 같을 때, value_counts()를 사용하여 빈도수를 집계하고 주어진 결과에 맞게 pie chart를 그리시오.

- gender = ['F','M','F','F','F','M','F','M','M']

In [ ]:

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

gender = ['F','M','F','F','F','M','F','M','M']
gender = pd.Series(gender)

cnt = gender.value_counts()

plt.pie(cnt.values, labels = cnt.index, autopct = '%.2f%%')
plt.savefig('a.png')
```

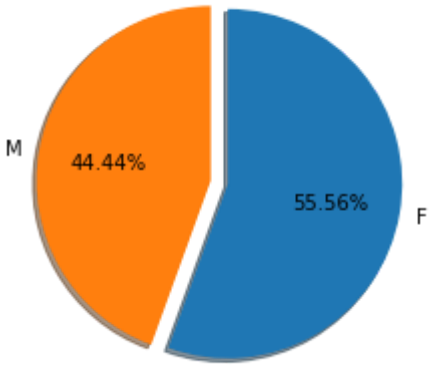


② ①에서 만들었던 그래프를 주어진 결과에 맞게 수정해 봅시다.

- 90도부터 시작
- 시계 방향으로
- 중심으로부터 F, M을 각각 0.05만큼 띄움
- 그림자 추가

In [ ]:

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

gender = ['F','M','F','F','F','M','F','M','M']
gender = pd.Series(gender)

cnt = gender.value_counts()

plt.pie(cnt.values, labels = cnt.index, autopct = '%.2f%%', startangle=90, counterclock=False,
explode = [0.05, 0.05], shadow=True)
plt.savefig('a.png')
```



5) 데이터 프레임으로부터 범주형 변수의 시각화

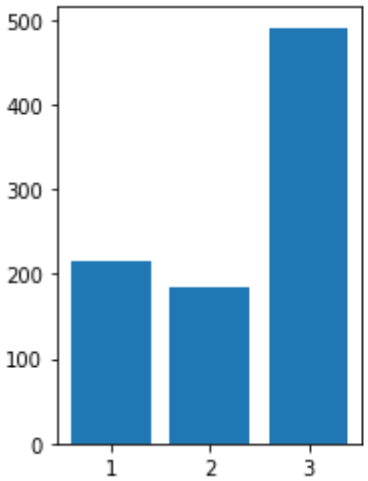
① Titanic

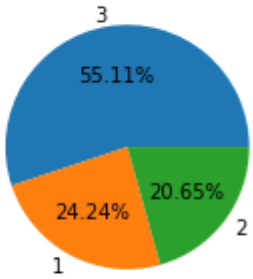
- 데이터셋 : titanic3
- 설명 : NaN 조치된 Titanic
- url : https://bit.ly/3HaMAtZ (https://bit.ly/3HaMAtZ)

(1) 티켓 클래스(Pclass)의 범주별 빈도수를 구하고 bar chart와 pie chart를 주어진 결과에 맞게 그리시오.

In [4]:

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

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

cnt = titanic['Pclass'].value_counts()

plt.subplot(1,2,1)
plt.bar(cnt.index, cnt.values)

plt.subplot(1,2,2)
plt.pie(cnt.values, labels = cnt.index, autopct = '%.2f%%')
plt.savefig('a.png')
```





(2) 분포로 부터 알수 있는 것은?

추가로 더 분석해볼 만한 내용은?

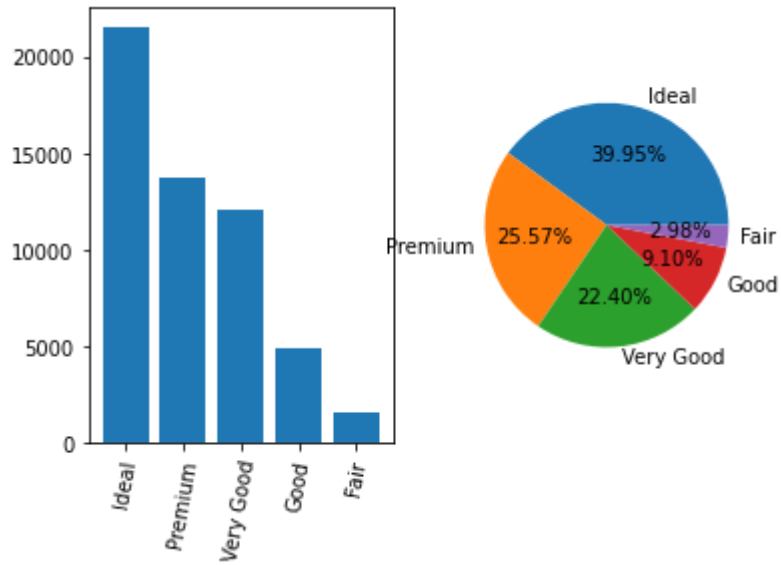
② Diamond

- 데이터셋 : diamond
- 설명 : 다이아몬드 가격
- url : https://bit.ly/3eHdRrH (https://bit.ly/3eHdRrH)

(1) 가공품질(cut)의 범주별 빈도수를 구하고 bar chart와 pie chart를 주어진 결과에 맞게 그리시오.

In [11]:

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

diamond = pd.read_csv('https://bit.ly/3eHdRrH')

cnt = diamond['cut'].value_counts()

plt.subplot(1,2,1)
plt.bar(cnt.index, cnt.values)
plt.xticks(rotation = 80)

plt.subplot(1,2,2)
plt.pie(cnt.values, labels = cnt.index, autopct = '%.2f%%')
plt.savefig('a.png')
```



(2) 분포로 부터 알수 있는 것은?

추가로 더 분석해볼 만한 내용은?