빅데이터분석프로그래밍 기말고사 (20241519 조예성)

describe의 max는 boxplot의 max와 같다: O, **X**

IQR = Q3- Q1

Upperfence 구하는 식: Q3 + 1.5 * IQR

빈 데이터프레임 확인 메소드: .empty

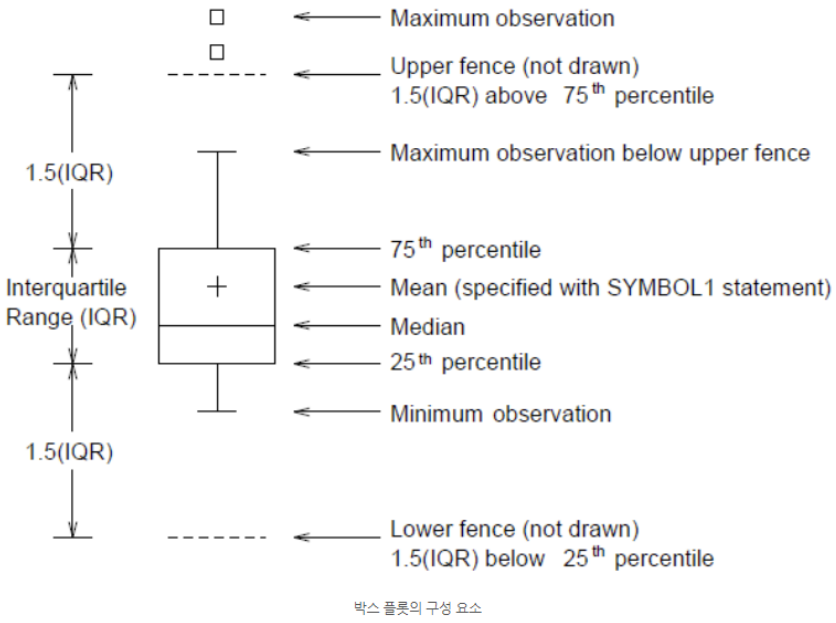
빈값 개수 반환: df.isna().sum().sum()

판다스 기본 소수점 설정: pd.set_option('display.precision', 2)

df['score'] 컬럼에서 값이5 이상인것만 nan으로 변경하고 다른거 유지 코드:

np.where(df['score'] > 5, np.nan, df['score'])

그래프 깔끔하게: %config InlineBackend.figure_format = 'retina'



박스 플롯의 구성 요소

Q1. 다음 데이터셋의 Upper Fence 값을 구하시오.
데이터: [4, 5, 7, 8, 9, 10, 12, 13, 14] 정답: 22.25

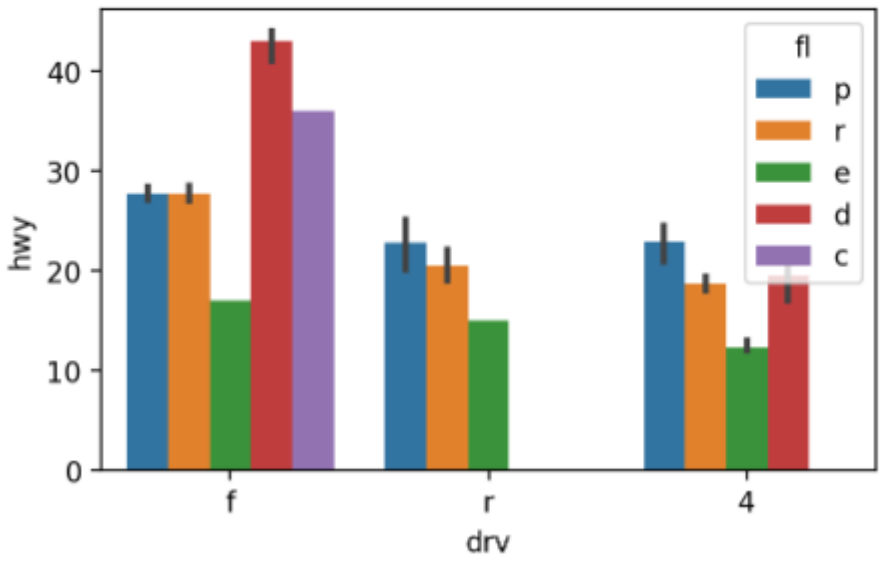
Q2. 다음 데이터셋에 대해 Q1, Q3, IQR, Upper Fence를 구하고,
이상치가 있다면 해당 값을 모두 쓰시오.
데이터: [1, 2, 3, 4, 5, 6, 7, 8, 50, 60]
Q1: 3, **Q3**: 8, **IQR**: 5, **UpperFence**: 15.5, **LowerFence**: -4.5, **Outliers**: 50, 60

시간이 아닌 값: pd.NaT

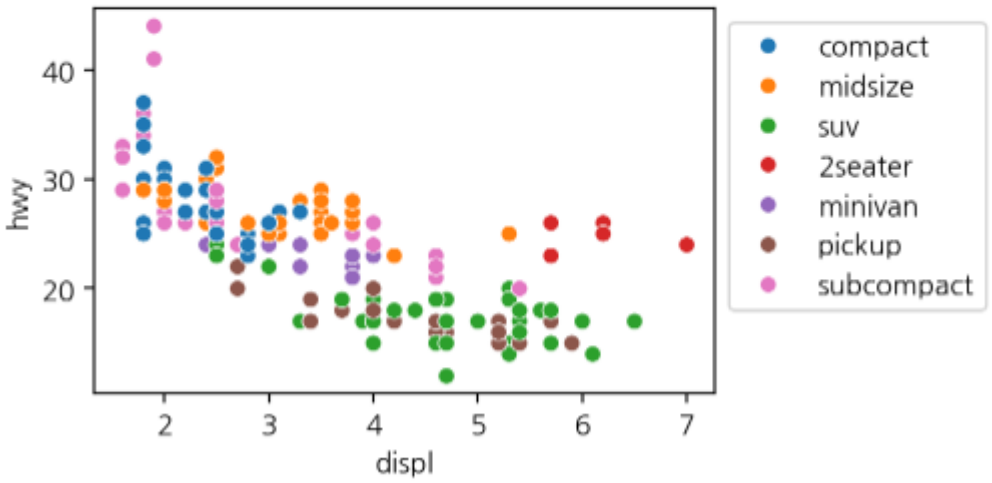
데이터프레임에 관한 정보 주는 메소드: df.info()

수치화된 데이터에 대한 정보 주는 메소드: df.describe()

```
sns.barplot(mpg, x='drv', y='hwy', hue='fl', order=['f', 'r', '4']);
```




```
sns.scatterplot(data = mpg, x = 'displ', y = 'hwy', hue = 'category');
plt.legend(loc='upper left', bbox_to_anchor=(1, 1));
```




```
piv_df = region_ageg[['region', 'ageg', 'proportion']].pivot(index= 'region',
                                                             columns='ageg',
                                                             values = 'proportion')
piv_df
```

ageg region	middle	old	young
강원/충북	31.84	44.15	24.01
광주/전남/전북/제주도	32.48	43.55	23.97
대구/경북	30.44	48.90	20.66
대전/충남	34.51	40.12	25.38
부산/경남/울산	34.15	42.52	23.33
서울	39.01	36.01	24.98
수도권(인천/경기)	39.43	30.87	29.70

파생변수 만들기에 사용하는 메소드: df.assign()

패키지 확인 위한 명령어: pip show

특수문자 제거하는 페키지, 코드: re.sub('[^가-힣]', ' ', text)

명사만 추출하기: hannaum.nouns('문자열')

- import 하기: from konlpy.tag import Hannaum

워드클라우드 import: from wordcloud import WorldCloud

```
from wordcloud import WordCloud
wc = WordCloud(font_path=FONT_PATH, background_color='gray',
               random_state = 1234, width=600, height=300)
img_wc = wc.generate_from_frequencies(dic_word)
```

워드클라우드 이미지 불러오는 라이브러리 불러오기: from PIL import Image

- 이미지 적용시 속성: mask = img

'[^가-힣]'은 str.replace에서도 사용 가능하다: **O**, X

Hannaum.nouns보다 성능 좋지만 시간 오래 걸리는 것: kkma.nouns

	reply
0	[국보, 국보소년단, 소년단]
1	[아줌마]
2	[팩트, 팩트체크, 체크, 보드, 위, 방탄, 방탄소년단, 소년단]
3	[방탄, 방탄소년단, 소년단, 한국, 한국사람, 사람, 자랑, 우리, 하자]
4	[월드, 클래스, 소식, 응원]

이거 풀기: df.explode

	A	B	C
0	[1, 2, 3]	1	[a, b, c]
1	foo	1	NaN
2	[]	1	[]
3	[3, 4]	1	[d, e]

A열 기반으로 explode 하는 코드와 결과: df.explode('A'),

	A	B	C
0	1	1	[a, b, c]
0	2	1	[a, b, c]
0	3	1	[a, b, c]
1	foo	1	NaN
2	NaN	1	[]
3	3	1	[d, e]
3	4	1	[d, e]

A열과 C열을 기반으로 explode 코드, 결과: df.explode(['A', 'C'])

	A	B	C
0	1	1	a
0	2	1	b
0	3	1	c
1	foo	1	NaN
2	NaN	1	NaN
3	3	1	d
3	4	1	e

```
from IPython.display import display_html
def display_side_by_side(*args):
    """여러 데이터프레임 비교가 쉽게 옆쪽으로 표시한다"""
    html_str=''
    for df in args:
        html_str += df.to_html() + ' '*4
    display_html(html_str.replace('table','table style="display:inline"'), raw=True)
```

df1, df2 동시 표기: display_side_by_side(df1, df2)

지도 라이브러리: folium

매개변수	설명
location	지도 중심 좌표 (예: [37.5665, 126.9780] — 서울) **필수**
zoom_start	초기 확대 레벨 (예: 13 정도가 도심 기준)
tiles	배경 지도 스타일 (예: 'OpenStreetMap', 'Stamen Terrain', 'CartoDB positron' 등)
width	지도의 너비 (기본값 '100%', px 또는 %)
height	지도의 높이 (기본값 '100%', px 또는 %)
control_scale	축척 표시 여부 (True면 지도에 축척이 생김)
scrollWheelZoom	마우스 스크롤 확대 허용 여부 (True/False)

지역별로 어떠한 값을 색으로 표현한 통계 지도 명칭: Choropleth

각 열별로 함수 적용하는 메소드: df.apply()