Class 11. 데이터 시각화

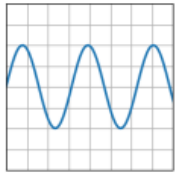
시각화는 데이터 분석의 시작
데이터를 시각화할 수 있다면
먼저 시각화 한 후에 분석에 들어가라

데이터 시각화 도구 matplotlib

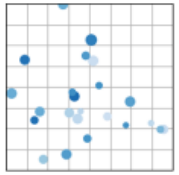
거의 원하는 모든 형태의 그래프를 그릴 수 있음

Basic

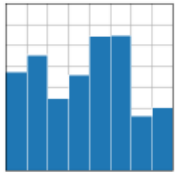
Basic plot types, usually y versus x.



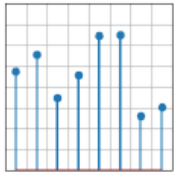
plot(x, y)



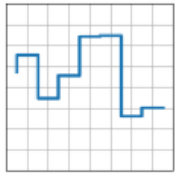
scatter(x, y)



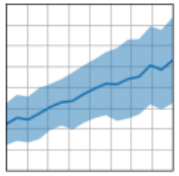
bar(x, height) / barh(y, width)



stem(x, y)



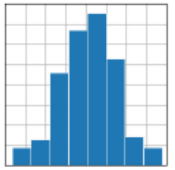
step(x, y)



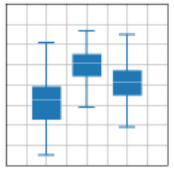
fill_between(x, y1, y2)

Statistics plots

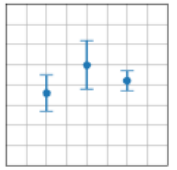
Plots for statistical analysis.



hist(x)

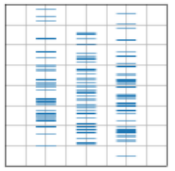


boxplot(X)

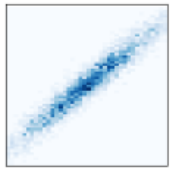


errorbar(x, y, yerr, xerr)

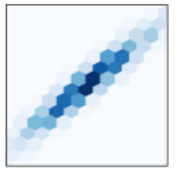
violinplot(D)



eventplot(D)



hist2d(x, y)



hexbin(x, y, C)

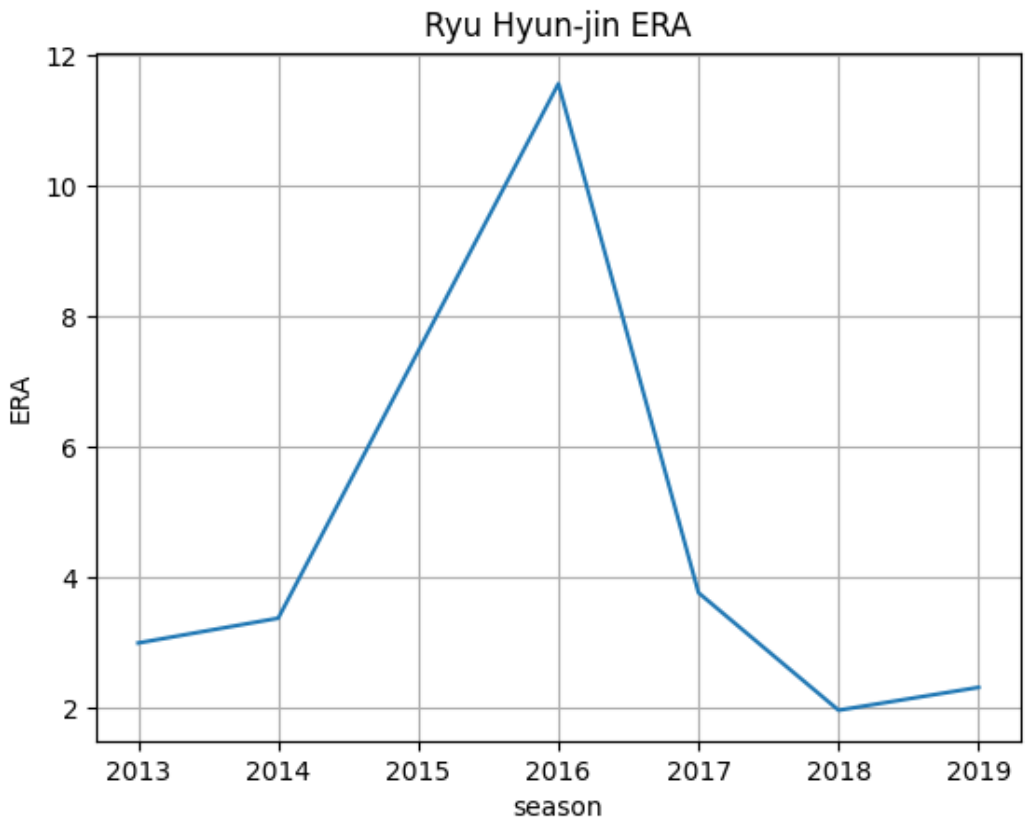
pie(x)

https://matplotlib.org/stable/plot_types/index.html

선 그래프 그리기

1. 류현진 선수의 ERA 추이를 선 그래프로 그려보자

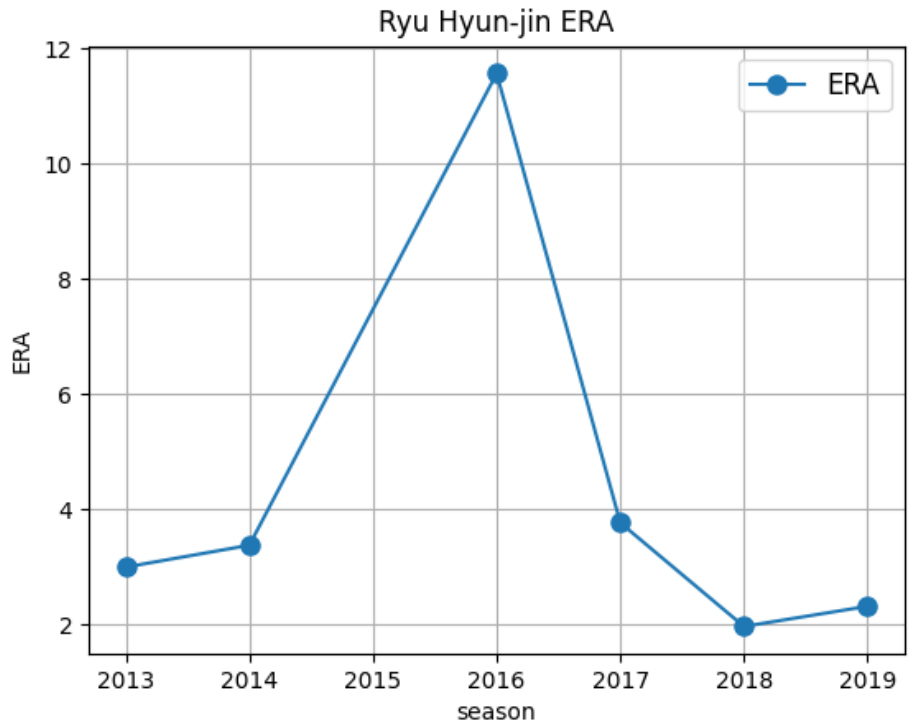
```
import sqlite3
import pandas as pd
import matplotlib.pyplot as plt

with sqlite3.connect("lahmansbaseballdb.sqlite") as con:
    cur = con.cursor()
    cur.execute('''
    SELECT playerID, yearID, ERA
    FROM pitching WHERE playerID = 'ryuhy01';
    ''')
    result = cur.fetchall()

cols = [column[0] for column in cur.description]

df = pd.DataFrame.from_records(data=result, columns=cols)

plt.plot(df['yearID'], df['ERA'])
plt.title('Ryu Hyun-jin ERA')
plt.xlabel('season')
plt.ylabel('ERA')
plt.grid(True)
plt.savefig('ryu_era.png')
```



선 그래프 그리기

2. 류현진 선수의 ERA 추이를 선 그래프로 그려보자(마커, 범례 추가)

```
import sqlite3
import pandas as pd
import matplotlib.pyplot as plt

with sqlite3.connect("lahmansbaseballdb.sqlite") as con:
    cur = con.cursor()
    cur.execute('''
    SELECT playerID, yearID, ERA
    FROM pitching WHERE playerID = 'ryuhy01';
    ''')
    result = cur.fetchall()

cols = [column[0] for column in cur.description] # 컬럼명 가져오기

df = pd.DataFrame.from_records(data=result, columns=cols)

plt.plot(df['yearID'], df['ERA'], marker='o', markersize=8)
plt.legend(labels=['ERA'], loc='best', fontsize=12)
plt.title('Ryu Hyun-jin ERA')
plt.xlabel('season')
plt.ylabel('ERA')
plt.grid(True)
plt.savefig('ryu_era.png')
```



선 그래프 그리기

3. 류현진 선수와 커쇼 선수의 ERA 추이를 선 그래프로 그려보자

```
import sqlite3
import pandas as pd
import matplotlib.pyplot as plt

with sqlite3.connect("lahmansbaseballdb.sqlite") as con:
    cur = con.cursor()
    cur.execute('''
    SELECT playerID, yearID, ERA
    FROM pitching WHERE playerID IN ('ryuhy01', 'kershcl01');
    ''')
    result = cur.fetchall()

cols = [column[0] for column in cur.description] # 컬럼명 가져오기

df = pd.DataFrame.from_records(data=result, columns=cols)
print(df)

df_ker = df[df['playerID']=='kershcl01']
print(df_ker)


df_ryu = df[df['playerID']=='ryuhy01']
print(df_ryu)

plt.plot(df_ker['yearID'], df_ker['ERA'], marker='o', markersize=8)
plt.plot(df_ryu['yearID'], df_ryu['ERA'], marker='o', markersize=8)
plt.legend(labels=['kershaw', 'ryu'], loc='best', fontsize=12)
plt.title('Ryu and Kershaw ERA')
plt.xlabel('season')
plt.ylabel('ERA')
plt.grid(True)
plt.savefig('ryu_kershaw_era.png')
```

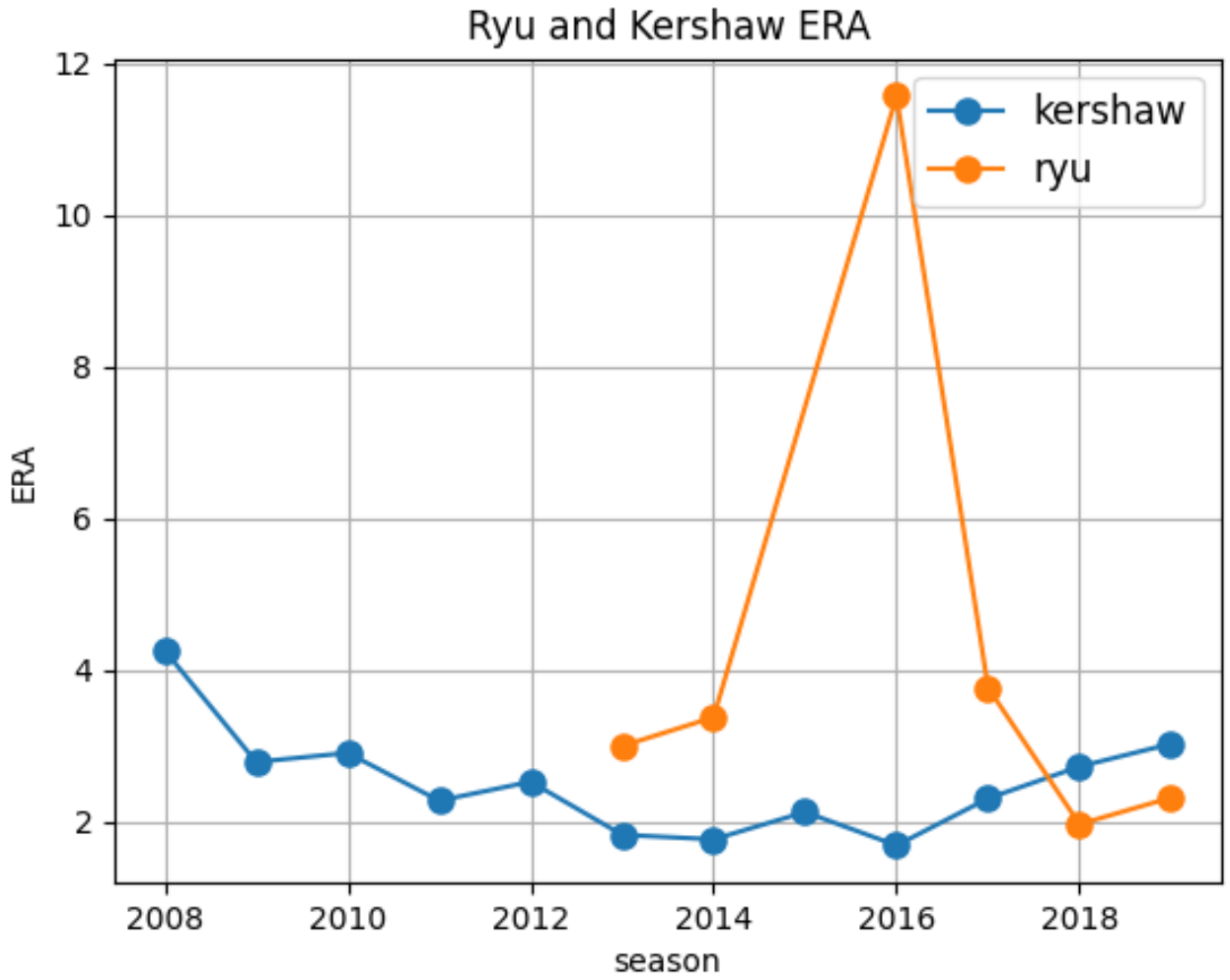
	playerID	yearID	ERA
0	kershcl01	2008	4.26
1	kershcl01	2009	2.79
2	kershcl01	2010	2.91
3	kershcl01	2011	2.28
4	kershcl01	2012	2.53
5	kershcl01	2013	1.83
6	kershcl01	2014	1.77
7	kershcl01	2015	2.13
8	kershcl01	2016	1.69
9	kershcl01	2017	2.31
10	kershcl01	2018	2.73
11	kershcl01	2019	3.03
12	ryuhy01	2013	3.00
13	ryuhy01	2014	3.38
14	ryuhy01	2016	11.57
15	ryuhy01	2017	3.77
16	ryuhy01	2018	1.97
17	ryuhy01	2019	2.32

	playerID	yearID	ERA
0	kershcl01	2008	4.26
1	kershcl01	2009	2.79
2	kershcl01	2010	2.91
3	kershcl01	2011	2.28
4	kershcl01	2012	2.53
5	kershcl01	2013	1.83
6	kershcl01	2014	1.77
7	kershcl01	2015	2.13
8	kershcl01	2016	1.69
9	kershcl01	2017	2.31
10	kershcl01	2018	2.73
11	kershcl01	2019	3.03

	playerID	yearID	ERA
12	ryuhy01	2013	3.00
13	ryuhy01	2014	3.38
14	ryuhy01	2016	11.57
15	ryuhy01	2017	3.77
16	ryuhy01	2018	1.97
17	ryuhy01	2019	2.32

선 그래프 그리기

3. 류현진 선수와 커쇼 선수의 ERA 추이를 선 그래프로 그려보자



막대 그래프 그리기

1. 2019년 LA 다저스 선발 투수 5명(류현진, 워커 뷸러, 클레이튼 커쇼, 켄타 마에다, 리치 힐)의 탈삼진 개수를 막대 그래프로 그려보자

Team Pitching Leag

Rk	Pos	Name
1	SP	Hyun Jin Ryu*
2	SP	Walker Buehler
3	SP	Clayton Kershaw*
4	SP	Kenta Maeda
5	SP	Rich Hill*

```
SELECT people.nameFirst || ' ' || people.nameLast AS
name, pitching.SO FROM pitching JOIN people ON
pitching.playerID = people.playerID WHERE
pitching.yearID = 2019 AND pitching.teamID = 'LAN' AND
people.nameLast IN ('Ryu','Buehler', 'Kershaw','Maeda',
'Hill');
```

JOIN 활용

스포츠 > 야구

"류현진 속한 2019 다저스 선발진, MLB 역대 10위"

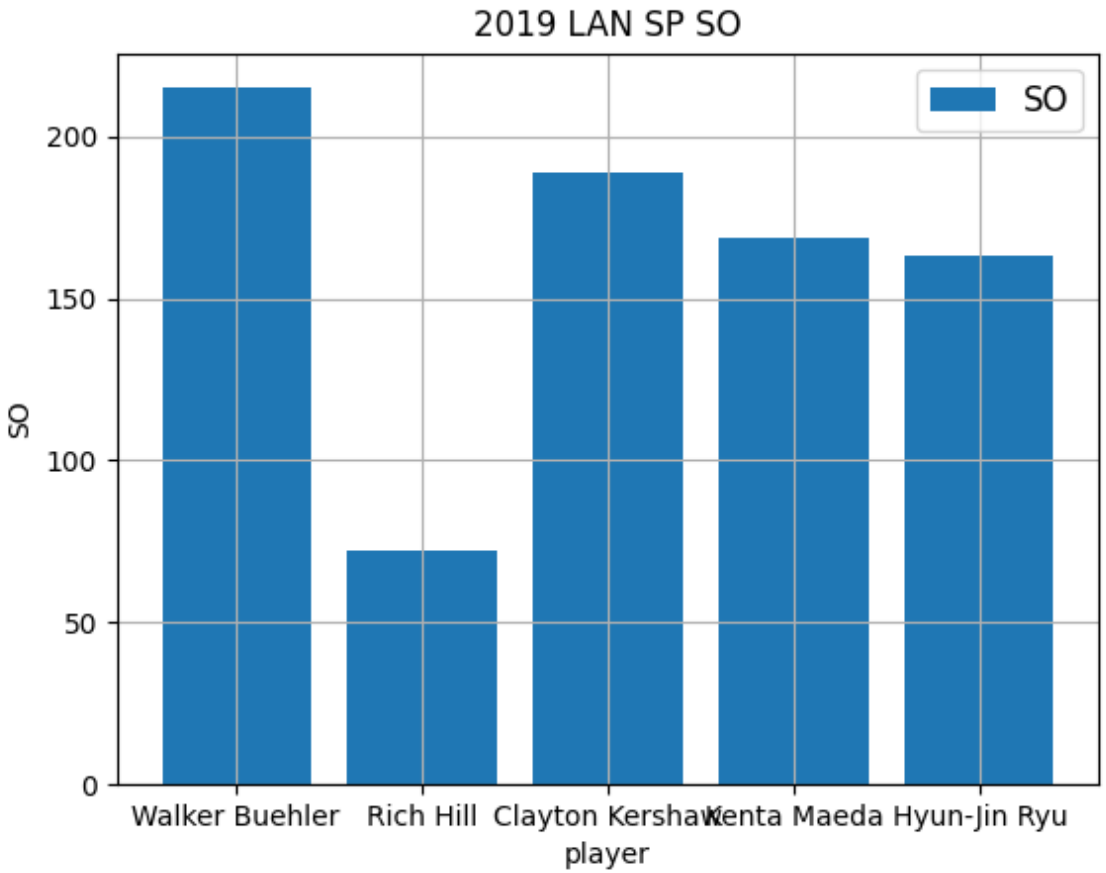
美CBS스포츠 '최고 선발진' 선정

```
name              SO
----------------  ---
Walker Buehler    215
Rich Hill         72
Clayton Kershaw   189
Kenta Maeda       169
Hyun-Jin Ryu      163
```

막대 그래프 그리기

1. 2019년 LA 다저스 선발 투수 5명(류현진, 워커 뷸러, 클레이튼 커쇼, 켄타 마에다, 리치 힐)의 탈삼진 개수를 막대 그래프로 그려보자

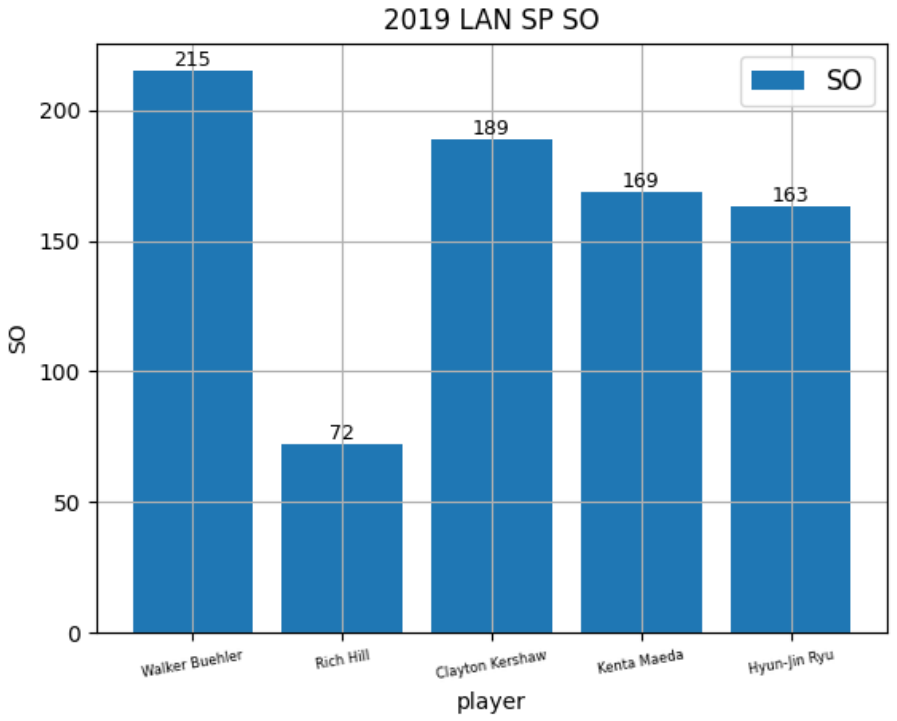
```
import sqlite3
import pandas as pd
import matplotlib.pyplot as plt

with sqlite3.connect("lahmansbaseballdb.sqlite") as con:
    cur = con.cursor()
    cur.execute('''
    SELECT people.nameFirst || ' ' || people.nameLast AS name,
pitching.SO
    FROM pitching JOIN people ON pitching.playerID = people.playerID
    WHERE pitching.yearID = 2019 AND pitching.teamID = 'LAN' AND
people.nameLast IN ('Ryu','Buehler', 'Kershaw','Maeda', 'Hill');
    ''')
    result = cur.fetchall()

cols = [column[0] for column in cur.description] # 컬럼명 가져오기

df = pd.DataFrame.from_records(data=result, columns=cols)

plt.bar(df['name'], df['SO'])
plt.legend(labels=['SO'], loc='best', fontsize=12)
plt.title('2019 LAN SP SO')
plt.xlabel('player')
plt.ylabel('SO')
plt.grid(True)
plt.savefig('LAN2019_SP_SO.png')
```



막대 그래프 그리기

2. 2019년 LA 다저스 선발 투수 5명(류현진, 워커 뷸러, 클레이튼 커쇼, 켄타 마에다, 리치 힐)의 탈삼진 개수를 막대 그래프로 그려보자(막대 위에 값 추가, x축 눈금 라벨 폰트 사이즈 변경 및 회전)

```
import sqlite3
import pandas as pd
import matplotlib.pyplot as plt

with sqlite3.connect("lahmansbaseballdb.sqlite") as con:
    cur = con.cursor()
    cur.execute('''
    SELECT people.nameFirst || ' ' || people.nameLast AS name, pitching.SO
    FROM pitching JOIN people ON pitching.playerID = people.playerID
    WHERE pitching.yearID = 2019 AND pitching.teamID = 'LAN' AND people.nameLast IN
('Ryu','Buehler', 'Kershaw','Maeda', 'Hill');
    ''')
    result = cur.fetchall()

cols = [column[0] for column in cur.description] # 컬럼명 가져오기

df = pd.DataFrame.from_records(data=result, columns=cols)

plt.bar(df['name'], df['SO'])
plt.legend(labels=['SO'], loc='best', fontsize=12)
plt.title('2019 LAN SP SO')
plt.xlabel('player')
plt.ylabel('SO')
plt.grid(True)
plt.xticks(size=6, rotation=10)

for i, v in enumerate(df['name']):
    plt.text(v, df.iloc[i, 1], df.iloc[i, 1],
             fontsize=9, horizontalalignment='center', verticalalignment='bottom')

plt.savefig('LAN2019_SP_SO.png')
```



막대 그래프 그리기

2. 2019년 LA 다저스 선발 투수 5명(류현진, 워커 뷸러, 클레이튼 커쇼, 켄타 마에다, 리치 힐)의 볼넷와 탈삼진 개수를 비교해보자

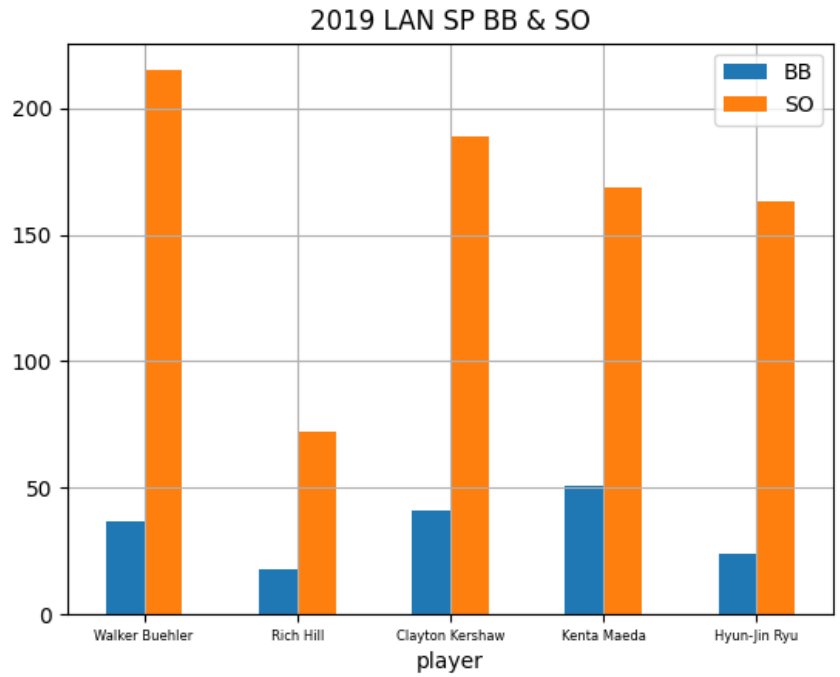
```
SELECT people.nameFirst || ' ' || people.nameLast AS
name, pitching.BB, pitching.SO FROM pitching JOIN
people ON pitching.playerID = people.playerID WHERE
pitching.yearID = 2019 AND pitching.teamID = 'LAN' AND
people.nameLast IN ('Ryu','Buehler', 'Kershaw','Maeda',
'Hill');
```

name	BB	SO
Walker Buehler	37	215
Rich Hill	18	72
Clayton Kershaw	41	189
Kenta Maeda	51	169
Hyun-Jin Ryu	24	163

막대 그래프 그리기

2. 2019년 LA 다저스 선발 투수 5명(류현진, 워커 뷸러, 클레이튼 커쇼, 켄타 마에다, 리치 힐)의 볼넷와 탈삼진 개수를 비교해보자

```
import sqlite3
import pandas as pd
import matplotlib.pyplot as plt

with sqlite3.connect("lahmansballdb.sqlite") as con:
    cur = con.cursor()
    cur.execute('''
    SELECT people.nameFirst || ' ' || people.nameLast AS name, pitching.BB,
pitching.SO
    FROM pitching JOIN people ON pitching.playerID = people.playerID
    WHERE pitching.yearID = 2019 AND pitching.teamID = 'LAN' AND people.nameLast
IN ('Ryu','Buehler', 'Kershaw','Maeda', 'Hill');
    ''')
    result = cur.fetchall()

cols = [column[0] for column in cur.description] # 컬럼명 가져오기

df = pd.DataFrame.from_records(data=result, columns=cols)

df.plot(x="name", y=["BB", "SO"], kind="bar")
plt.xticks(size=6, rotation=0)
plt.title('2019 LAN SP BB & SO')
plt.xlabel('player')
plt.grid(True)
plt.savefig('LAN2019_SP_BB_SO.png')
```

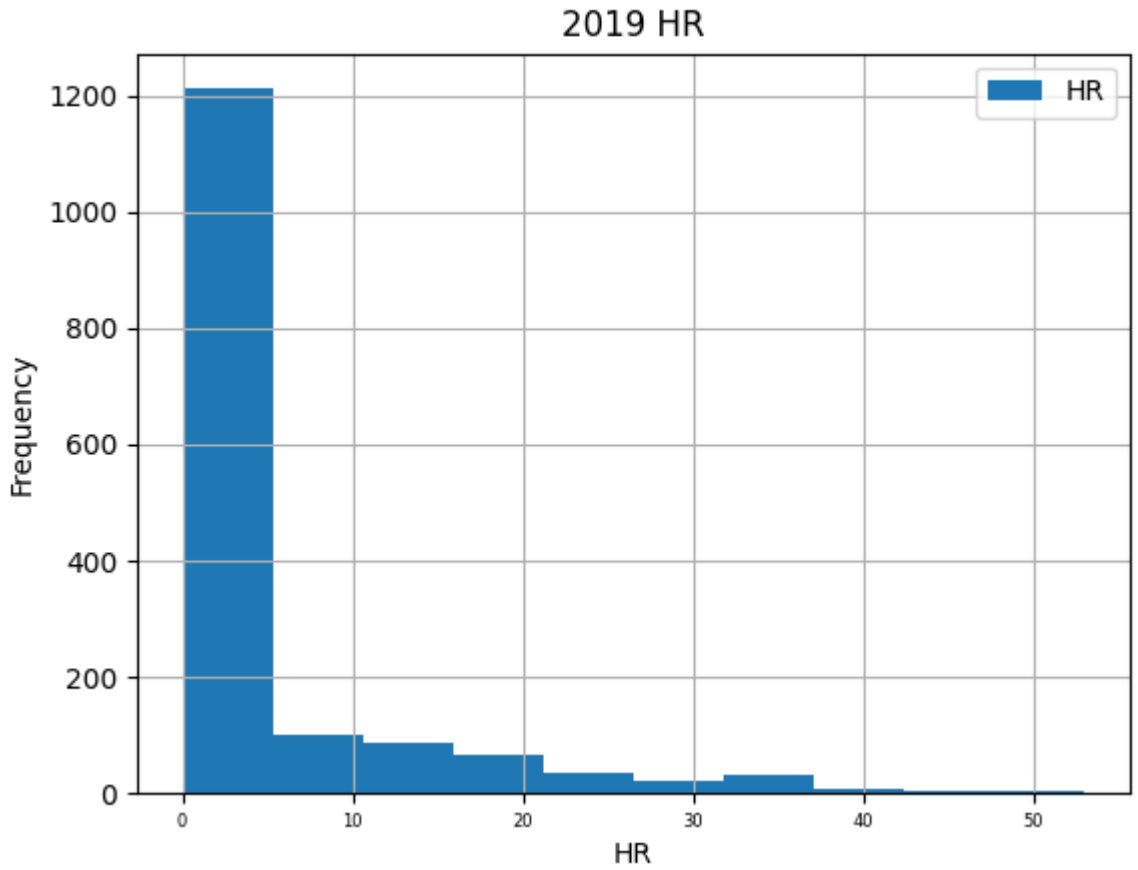


히스토그램 그리기

1. 2019년 MLB 전체 타자들의 홈런수를 히스토그램으로 그려보자

```
SELECT HR FROM batting WHERE yearID = 2019;
```

```
import sqlite3
import pandas as pd
import matplotlib.pyplot as plt

with sqlite3.connect("lahmansbaseballdb.sqlite") as con:
    cur = con.cursor()
    cur.execute('''
    SELECT HR FROM batting WHERE yearID = 2019;
    ''')
    result = cur.fetchall()

cols = [column[0] for column in cur.description]

df = pd.DataFrame.from_records(data=result, columns=cols)

df.plot(kind="hist")
plt.xticks(size=6, rotation=0)
plt.title('2019 HR')
plt.xlabel('HR')
plt.grid(True)
plt.savefig('HR2019.png')
```



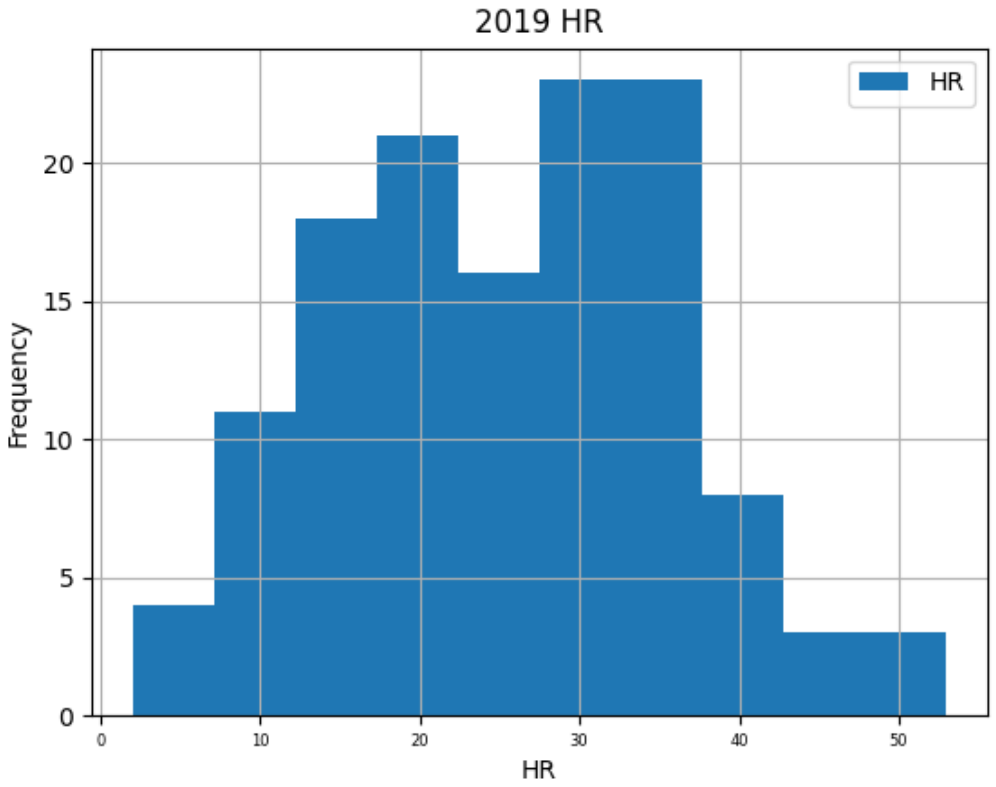
홈런 타자는 희소가치가 높다

히스토그램 그리기

2. 2019년 MLB 규정 타석을 채운 타자들의 홈런수를 히스토그램으로 그려보자

```
SELECT HR FROM batting WHERE yearID = 2019 AND (AB + BB + HBP + SH + SF) >= 502;
```

```
import sqlite3
import pandas as pd
import matplotlib.pyplot as plt

with sqlite3.connect("lahmansbaseballdb.sqlite") as con:
    cur = con.cursor()
    cur.execute('''
    SELECT HR FROM batting WHERE yearID = 2019 AND (AB + BB +
HBP + SH + SF) >= 502;
    ''')
    result = cur.fetchall()

cols = [column[0] for column in cur.description] # 컬럼명 가져오
기

df = pd.DataFrame.from_records(data=result, columns=cols)

df.plot(kind="hist")
plt.xticks(size=6, rotation=0)
plt.title('2019 HR')
plt.xlabel('HR')
plt.grid(True)
plt.savefig('HR2019_regular.png')
```



MLB에서 규정 타석을 채울 기회를 부여 받은 타자는 홈런을 대체로 잘 쳤다

Player Standard Batting Share & Export ▾ When table i

Rk	Name	Age	Tm	Lg	G	PA	AB	R	H	2B	3B	HR
1	Pete Alonso	24	NYM	NL	161	693	597	103	155	30	2	53
2	Eugenio Suarez	27	CIN	NL	159	662	575	87	156	22	2	49
3	Jorge Soler	27	KCR	AL	162	679	589	95	156	33	1	48
4	Cody Bellinger*	23	LAD	NL	156	661	558	121	170	34	3	47
5	Mike Trout	27	LAA	AL	134	600	470	110	137	27	2	45
6	Christian Yelich*	27	MIL	NL	130	580	489	100	161	29	3	44
7	Ronald Acuna Jr.	21	ATL	NL	156	715	626	127	175	22	2	41
8	Nolan Arenado	28	COL	NL	155	662	588	102	185	31	2	41
9	Alex Bregman	25	HOU	AL	156	690	554	122	164	37	2	41
10	Nelson Cruz	38	MIN	AL	120	521	454	81	141	26	0	41
11	George Springer	29	HOU	AL	122	556	479	96	140	20	3	39
12	Freddie Freeman*	29	ATL	NL	158	692	597	113	176	34	2	38
13	Kyle Schwarber*	26	CHC	NL	155	610	529	82	132	29	3	38
14	Gleyber Torres	22	NYY	AL	144	604	546	96	152	26	0	38
15	Josh Bell#	26	PIT	NL	143	613	527	94	146	37	3	37
16	Josh Donaldson	33	ATL	NL	155	659	549	96	142	33	0	37
17	Franmil Reyes	23	TOT	MLB	150	548	494	69	123	19	0	37
18	Matt Chapman	26	OAK	AL	156	670	583	102	145	36	3	36
19	Max Kepler*	26	MIN	AL	134	596	524	98	132	32	0	36
20	J.D. Martinez	31	BOS	AL	146	657	575	98	175	33	2	36
21	Matt Olson*	25	OAK	AL	127	547	483	73	129	26	0	36
22	Joc Pederson*	27	LAD	NL	149	514	450	83	112	16	3	36
23	Eduardo Escobar#	30	ARI	NL	158	699	636	94	171	29	10	35
24	Bryce Harper*	26	PHI	NL	157	682	573	98	149	36	1	35
25	Trey Mancini	27	BAL	AL	154	679	602	106	175	38	2	35
26	Mike Moustakas*	30	MIL	NL	143	584	523	80	133	30	1	35
27	Max Muncy*	28	LAD	NL	141	589	487	101	122	22	1	35
28	Trevor Story	26	COL	NL	145	656	588	111	173	38	5	35
29	Edwin Encarnacion	36	TOT	AL	109	486	418	81	102	18	0	34
30	Paul Goldschmidt	31	STL	NL	161	682	597	97	155	25	1	34
31	Anthony Rendon	29	WSN	NL	146	646	545	117	174	44	3	34
32	Gary Sanchez	26	NYY	AL	106	446	396	62	92	12	1	34
33	Miguel Sano	26	MIN	AL	105	439	380	76	94	19	2	34
34	Carlos Santana#	33	CLE	AL	158	686	573	110	161	30	1	34
35	Juan Soto*	20	WSN	NL	150	659	542	110	153	32	5	34
36	José Abreu	32	CHW	AL	159	693	634	85	180	38	1	33
37	Xander Bogaerts	26	BOS	AL	155	698	614	110	190	52	0	33
38	Kole Calhoun*	31	LAA	AL	152	632	552	92	128	29	1	33
39	Michael Conforto*	26	NYM	NL	151	648	549	90	141	29	1	33
40	Austin Meadows*	24	TBR	AL	138	591	530	83	154	29	7	33
41	Hunter Renfroe	27	SDP	NL	140	494	440	64	95	19	1	33
42	Marcus Semien	28	OAK	AL	162	747	657	123	187	43	7	33
43	Charlie Blackmon*	32	COL	NL	140	634	580	112	182	42	7	32
44	Rafael Devers*	22	BOS	AL	156	702	647	129	201	54	4	32
45	Francisco Lindor#	25	CLE	AL	143	654	598	101	170	40	2	32
46	Manny Machado	26	SDP	NL	156	661	587	81	150	21	2	32
47	Ketel Marte#	25	ARI	NL	144	628	569	97	187	36	9	32
48	Eddie Rosario*	27	MIN	AL	137	590	562	91	155	28	1	32
49	Jose Altuve	29	HOU	AL	124	548	500	89	149	27	3	31
50	Kris Bryant	27	CHC	NL	147	634	543	108	153	35	1	31

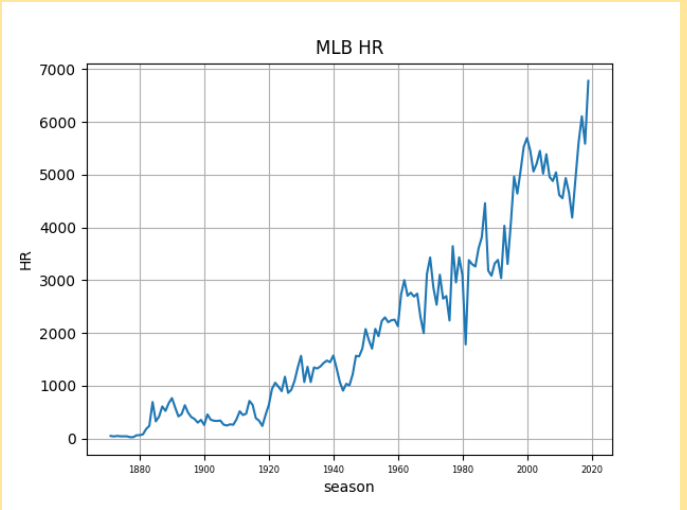
2019 MLB

30 홈런 이상을 친 타자가 무려 58명
20 홈런 이상을 친 타자는 130명
총 홈런 개수 6776개

MLB 시즌 홈런의 추이를 선 그래프로 그려보자

```
SELECT yearID, SUM(HR) AS HR_SUM FROM teams GROUP BY yearID;
```

```
import sqlite3
import pandas as pd
import matplotlib.pyplot as plt

with sqlite3.connect("lahmansbaseballdb.sqlite") as con:
    cur = con.cursor()
    cur.execute('''
    SELECT yearID, SUM(HR) AS HR_SUM FROM teams GROUP BY yearID;
    ''')
    result = cur.fetchall()

cols = [column[0] for column in cur.description]

df = pd.DataFrame.from_records(data=result, columns=cols)

plt.plot(df['yearID'], df['HR_SUM'])
plt.xticks(size=6, rotation=0)
plt.title('MLB HR')
plt.xlabel('season')
plt.ylabel('HR')
plt.grid(True)
plt.savefig('MLB_HR.png')
```



시즌 홈런 추이 그래프에 추세선을 넣어보자

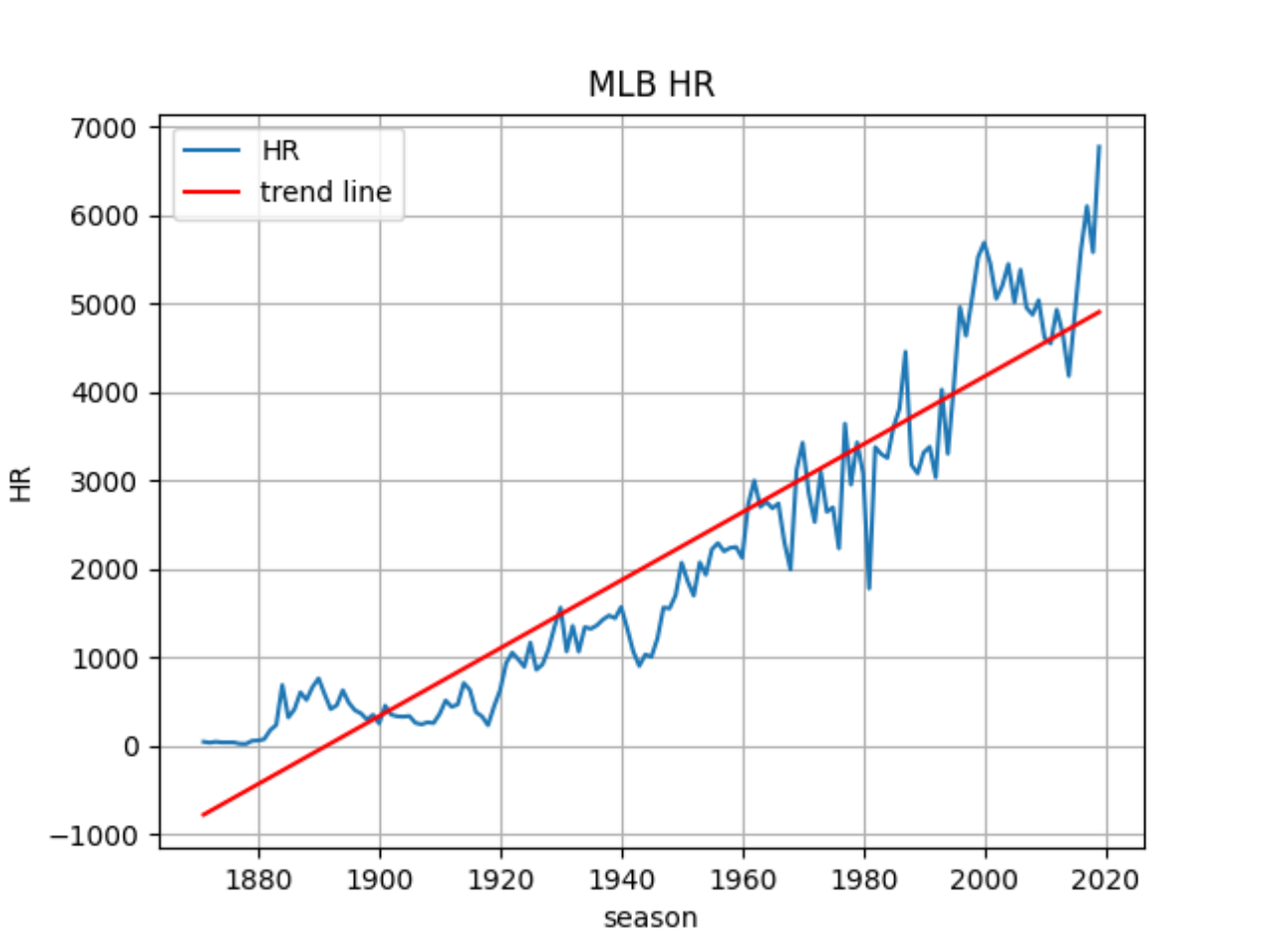
```python
import sqlite3
import pandas as pd
import matplotlib.pyplot as plt
from scipy.stats import linregress

with sqlite3.connect("lahmansbaseballdb.sqlite") as con:
    cur = con.cursor()
    cur.execute('''
    SELECT yearID, SUM(HR) AS HR_SUM FROM teams GROUP BY yearID;
    ''')
    result = cur.fetchall()

cols = [column[0] for column in cur.description]

df = pd.DataFrame.from_records(data=result, columns=cols)

slope, intercept, r_value, p_value, std_err = linregress(df['yearID'], df['HR_SUM'])
print("slope: %f, intercept: %f" % (slope, intercept))
print("R-squared: %f" % r_value**2)

plt.plot(df['yearID'], df['HR_SUM'], label='HR')
plt.plot(df['yearID'], intercept + slope * df['yearID'], 'r', label='trend line')
plt.title('MLB HR')
plt.xlabel('season')
plt.ylabel('HR')
plt.legend()
plt.grid(True)
plt.savefig('MLB_HR_trend.png')
```

사이파이
과학기술계산을 위한 파이썬 패키지

linregress() 메소드: 선형 회귀

선형 회귀 설명!

시즌 홈런 추이 그래프에 추세선을 넣어보자




```
기울기: 38.414638, y절편: -72650.954381
피어슨 상관계수: 0.941227
결정계수(R-squared): 0.885909
```

결정계수

- 회귀식이 얼마나 정확한지를 나타냄
- 회귀 모델에서 독립변수가 종속변수를 얼마만큼 설명해 주는지를 가리키는 지표

1. 전체 기간을 놓고 보면 시즌 홈런은 증가하는 추세

2. 2000년부터 2014년까지는 홈런 개수가 감소하는 추세, WHY? 프로젝트 주제로 삼아보는 건 어떨지?

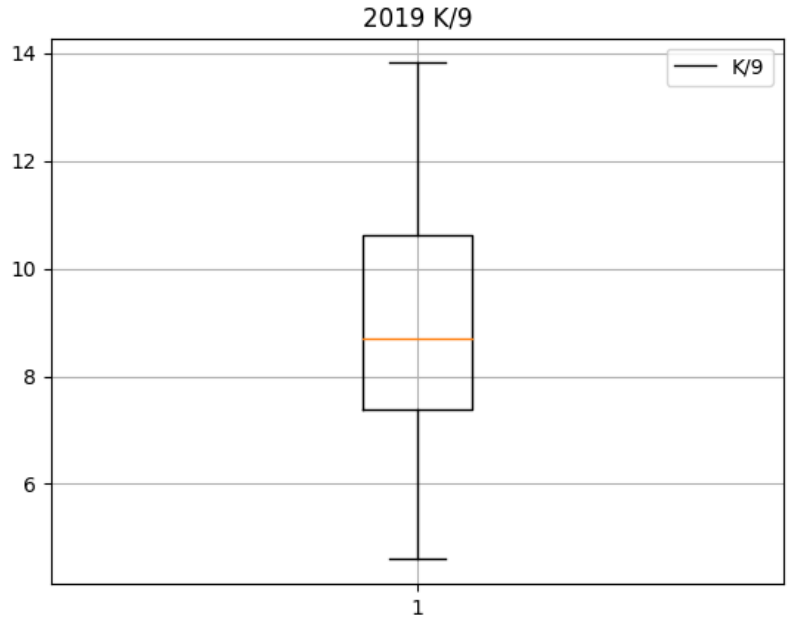
3. 1900-1920년 데드볼 시대라고 불림
- 투고타저 극심
- 데드볼: 반발력이 약한 공을 의미

박스 플롯 그리기

1. 2019년 규정이닝을 채운 MLB 투수들의 K/9를 박스 플롯으로 그려보자

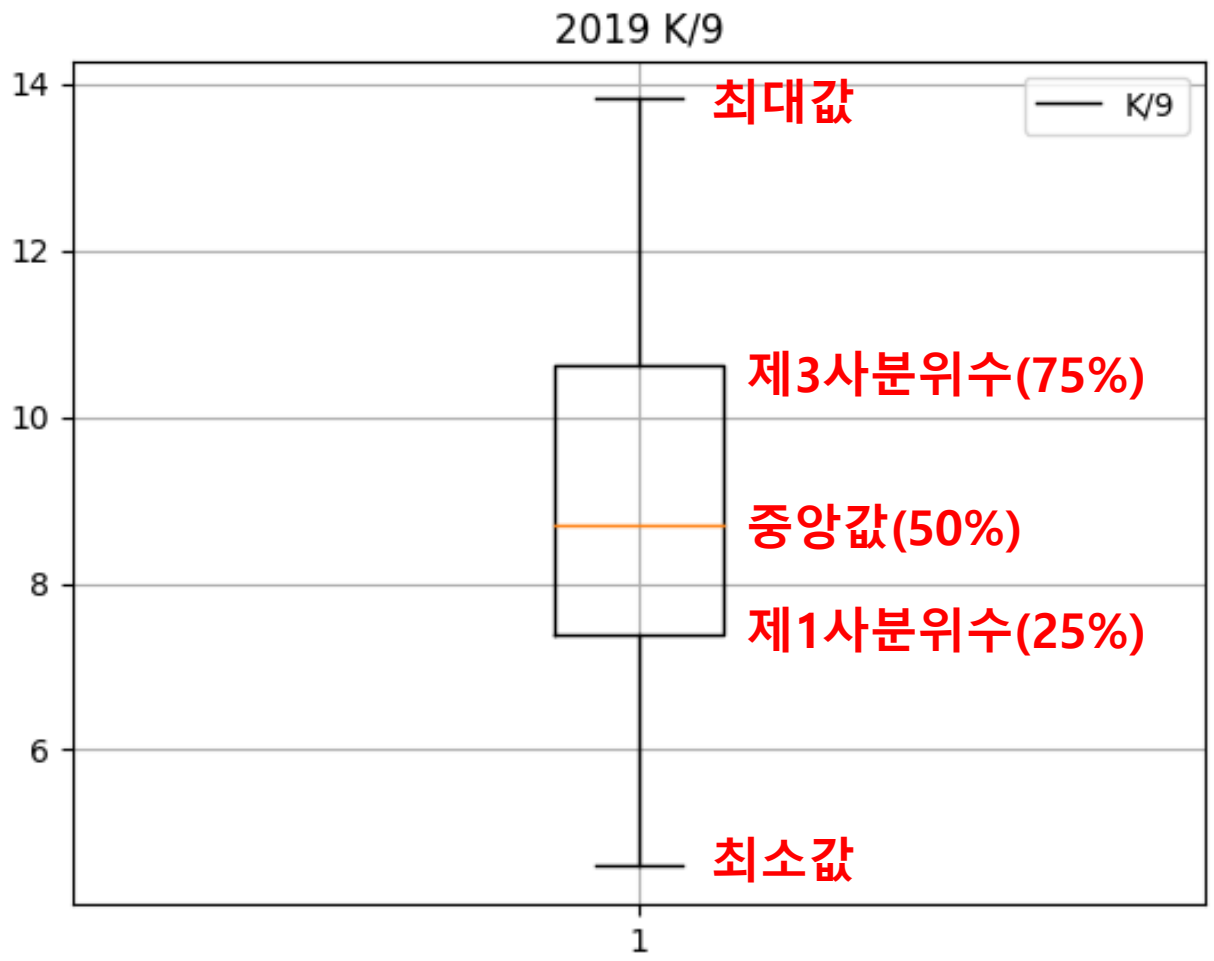
```
SELECT SO*9/(Ipouts/3.0) AS “K/9” FROM pitching WHERE yearID = 2019 AND Ipouts/3.0 >= 162;
```

```python
import sqlite3
import pandas as pd
import matplotlib.pyplot as plt
from scipy.stats import linregress

with sqlite3.connect("lahmansbaseballdb.sqlite") as con:
    cur = con.cursor()
    cur.execute('''
    SELECT SO*9/(Ipouts/3.0) AS "K/9"
    FROM pitching
    WHERE yearID = 2019 AND Ipouts/3.0 >= 162;
    ''')
    result = cur.fetchall()

cols = [column[0] for column in cur.description]

df = pd.DataFrame.from_records(data=result, columns=cols)

plt.boxplot(df['K/9'])
plt.title('2019 K/9')
plt.legend(['K/9'])
plt.grid(True)
plt.savefig('2019_K9.png')
```



규정 이닝을 채운 평균 수준의 MLB 투수들은 9이닝 당 8.5개 정도의 탈삼진을 잡는다

박스 플롯 그리기



중앙값
37
제2사분위수

10 15 19 | 20 33 35 | 39 41 42 | 44 51 52

19.5
제1사분위수

43
제3사분위수

자료를 대표하는 대표값들

평균(mean) vs 중앙값(median) vs 최빈값(mode)

- 평균은 아웃라이어에 취약
- 중앙값, 최빈값은 비교적 아웃라이어의 영향을 덜 받음

TRY

선 그래프, 막대 그래프, 히스토그램, 박스 플롯, 파이 그래프 등을 활용하여
보고 싶은 데이터 3개를 시각화 해보자

이제 슬슬 프로젝트 주제를 선정해보세요