

## **Class 4. 상관관계**

**득점은 많이 하고 실점은 적게 하는 팀이 자주 이긴다**

대부분의 구기 종목에 해당되는 명제

**정말로 득점은 많이 하고 실점은 적게 하는 팀이 자주 이겨?**

**너의 주장을 증명해봐!**

# 정말 득점을 많이 하고 실점은 적게 하는 팀의 승률이 좋을까?

우선 SQLite3를 이용해서 2019년 MLB 팀들의 팀승률과 득실점비율을 조회해보자

$$\text{팀승률} = \frac{\text{승}}{\text{승} + \text{패}}$$

※ 참고로 MLB에는 무승부가 없음

$$\text{득실점비율} = \frac{\text{득점}^2}{\text{득점}^2 + \text{실점}^2}$$

```
SELECT CAST(W as REAL)/(W+L) AS WIN_ratio,  
CAST(R*R as REAL)/(R*R + RA*RA) AS RS_RA_ratio  
FROM teams where yearID = 2019;
```

WIN_ratio	RS_RA_ratio
0.3333333333333333	0.355764016917905
0.518518518518518	0.542145807524451
0.447204968944099	0.419999195634979
0.574074074074074	0.578059843012287
0.291925465838509	0.288043103910119
0.660493827160494	0.673885350318471
0.364197530864198	0.387363889920999
0.444444444444444	0.439743899582461
0.623456790123457	0.607984074327162
0.635802469135803	0.619525975880783
0.598765432098765	0.606944769109803
0.419753086419753	0.418774457676421
0.592592592592593	0.578802717439711
0.481481481481481	0.459780908545576
0.41358024691358	0.434644500519519
0.524691358024691	0.544896283484664
0.598765432098765	0.569745003016194
0.518518518518518	0.563103974300684
0.462962962962963	0.492918202226453
0.438271604938272	0.431721206769829
0.654320987654321	0.67627469815165
0.351851851851852	0.366820904887939
0.549382716049383	0.501954389929014
0.530864197530864	0.535296231338495
0.5	0.487246972782698
0.425925925925926	0.409092300736569
0.432098765432099	0.427643210790285
0.475308641975309	0.434807366159894
0.561728395061728	0.571164647630976
0.574074074074074	0.592494781564248

# 정말 득점을 많이 하고 실점은 적게 하는 팀의 승률이 좋을까?

우선 SQLite3를 이용해서 2019년 MLB 팀들의 팀승률과 득실점비율을 조회해보자

## ROUND 함수 처리

```
SELECT ROUND(CAST(W as REAL)/(W+L), 3) AS  
WIN_ratio, ROUND(CAST(R*R as REAL)/(R*R + RA*RA), 3)  
AS RS_RA_ratio FROM teams where yearID = 2019;
```

### ROUND() 함수

숫자 반올림에 사용되는 함수.

Ex1) SELECT ROUND(1.6543); → 2.0

ex2) SELECT ROUND(1.6543, 1); → 1.7

ex3) SELECT ROUND(1.6543, 2); → 1.65

ex4) SELECT ROUND(1.6543, 3); → 1.654

WIN_ratio	RS_RA_ratio
0.333	0.356
0.519	0.542
0.447	0.42
0.574	0.578
0.292	0.288
0.66	0.674
0.364	0.387
0.444	0.44
0.623	0.608
0.636	0.62
0.599	0.607
0.42	0.419
0.593	0.579
0.481	0.46
0.414	0.435
0.525	0.545
0.599	0.57
0.519	0.563
0.463	0.493
0.438	0.432
0.654	0.676
0.352	0.367
0.549	0.502
0.531	0.535
0.5	0.487
0.426	0.409
0.432	0.428
0.475	0.435
0.562	0.571
0.574	0.592

대략적으로 봤을 때 승률이 높은 경우 득실점 비율도 높다  
하지만 좀 더 정확하게 판단을 하기 위해서는 상관관계를 살펴봐야함

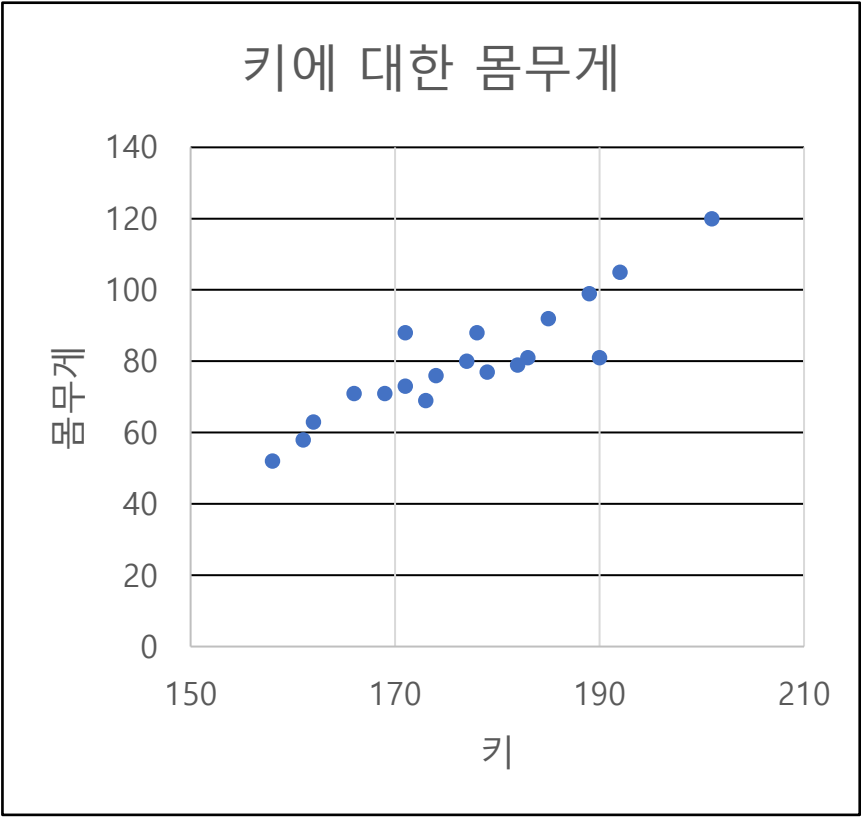
# 상관계수(correlation coefficient)

상관계수: 두 데이터간 연관성을 판단할 때 사용, 상관성의 정도를 하나의 수치로 나타낸 것

키(cm)	몸무게(kg)
169	71
182	79
192	105
173	69
177	80
171	88
162	63
185	92
166	71
178	88
179	77
161	58
189	99
183	81
158	52
174	76
171	73
201	120
190	81

(x, y)

산점도(scatter plot): 데이터 간의 관계를 나타내는 그래프

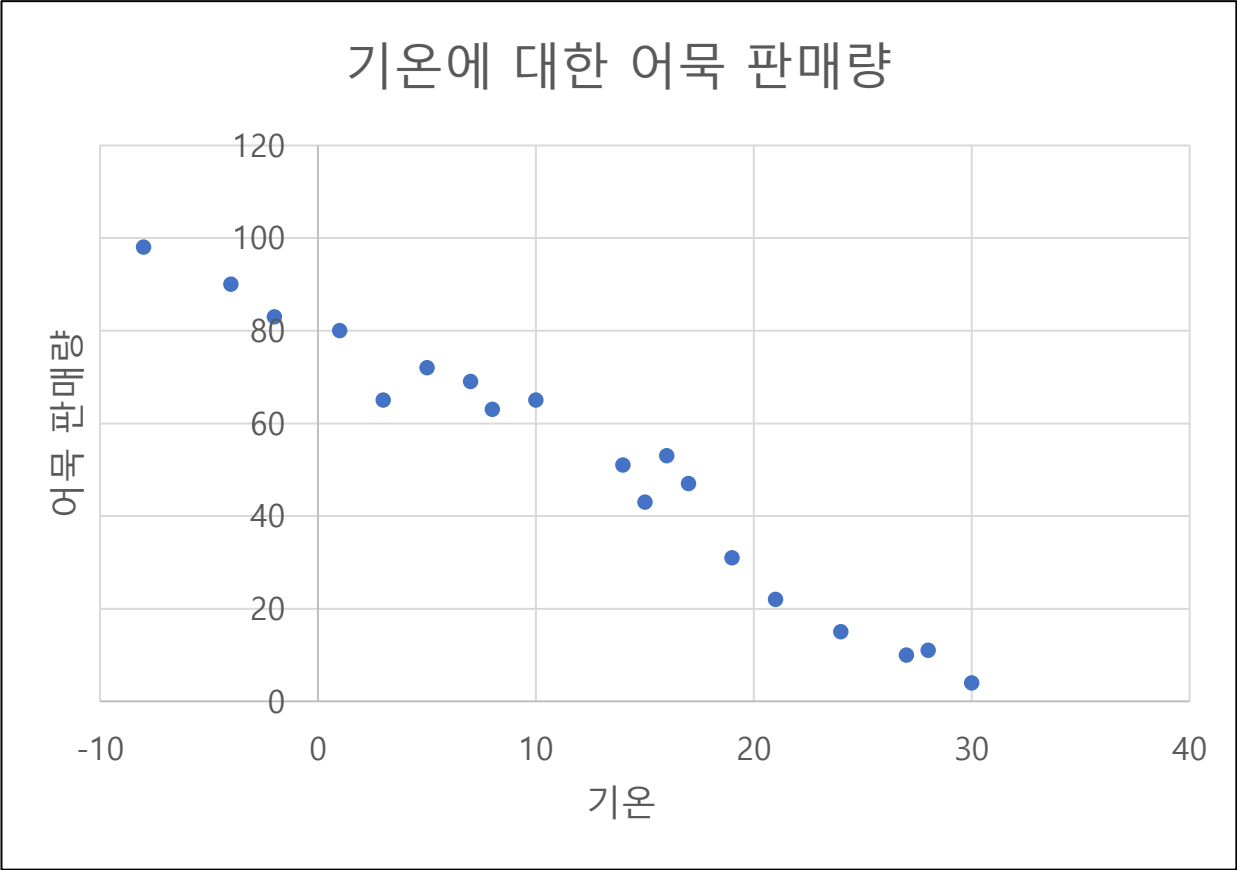


키와 몸무게는 상관관계가 강하다(강한 양의 상관관계)

# 상관계수(correlation coefficient)

기온	포장마차 어묵 판매량
30도	4개
28도	11개
-8도	98개
5도	72개
10도	65개
14도	51개
27도	10개
21도	22개
15도	43개
1도	80개
7도	69개
24도	15개
8도	63개
-4도	90개
-2도	83개
19도	31개
17도	47개
3도	65개
16도	53개

(x, y)



기온과 어묵판매량은 상관관계가 강하다(강한 음의 상관관계)

# 상관계수(correlation coefficient)

피어슨 상관계수, 스피어만 상관계수, 켄달 상관계수 등이 있음

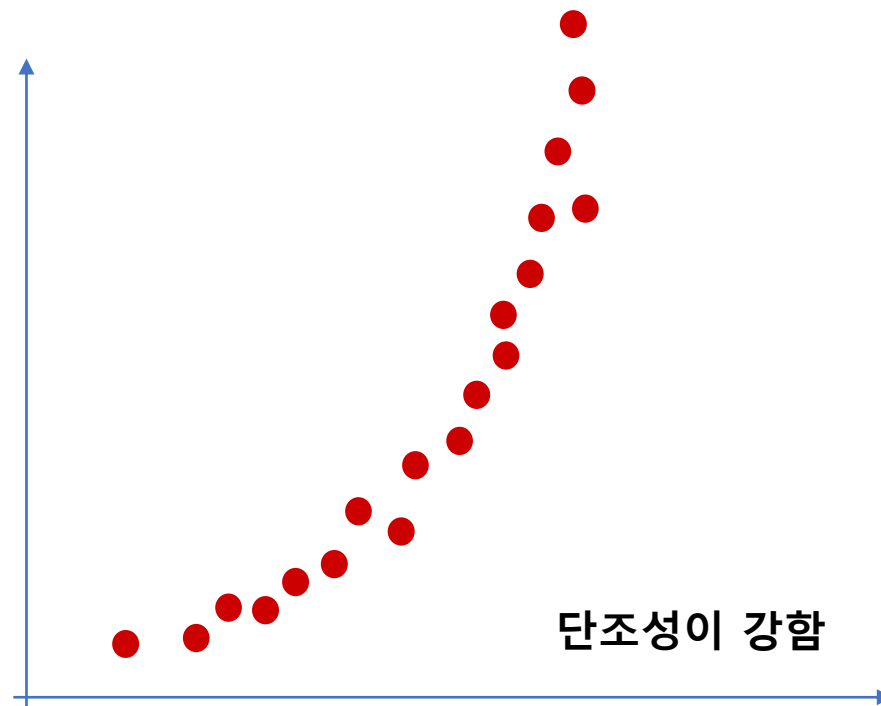
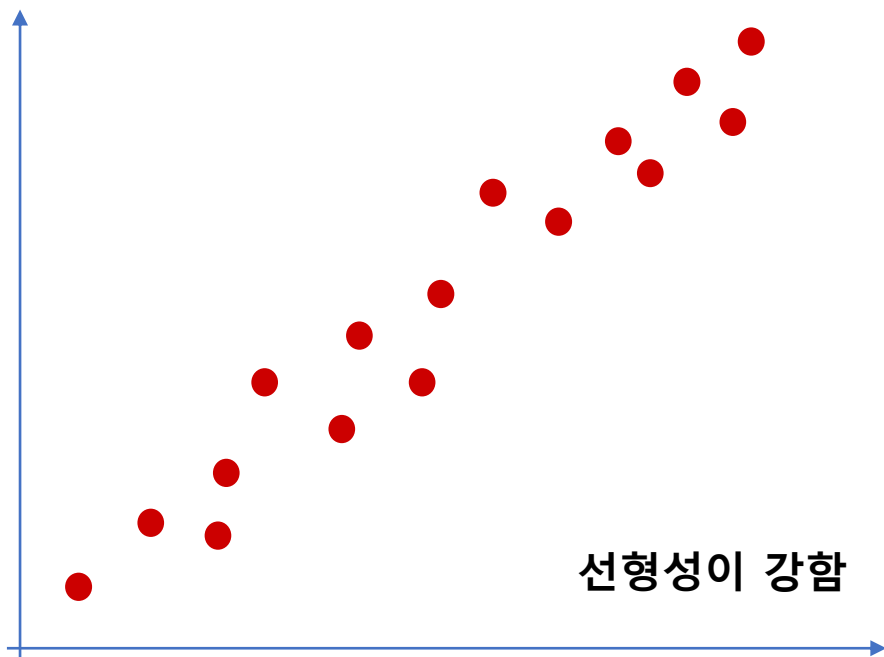
## 1) 피어슨 상관계수

두 데이터 간 선형성(linearity)이 얼마나 강한지 측정하기 위해 사용됨

## 2) 스피어만 상관계수, 켄달 상관계수

두 데이터 간 단조성(monotonicity)이 얼마나 강한지 측정하기 위해 사용됨

상관계수의 절대값	정도
0 ~ 0.2	상관관계 거의 없음
0.2 ~ 0.4	약한 상관관계
0.4 ~ 0.7	다소 강한 상관관계
0.7 ~ 0.9	강한 상관관계
0.9 ~ 1.0	매우 강한 상관관계





상관계수를 구하기 위해  
sqlite3에서 조회한 결과를 스프레드시트로 가져가보자

**\*스프레드시트(sheet):**

표 형식으로 데이터의 저장, 분석을 가능하게 하는 프로그램

ex) 마이크로소프트 엑셀, 리브레오피스 캘크, 로터스 1-2-3, 한글과컴퓨터 한셀, 구글 스프레드시트 등

## 조회 결과 CSV 형식으로 보기

```
1,abercda01,1871,1,TR0,8,NA,1,,4,0,0,0,0,0,0,0,0,0,,,0  
2,addybo01,1871,1,RC1,7,NA,25,,118,30,32,6,0,0,13,8,1,4,0,,,0  
3,allisar01,1871,1,CL1,3,NA,29,,137,28,40,4,5,0,19,3,1,2,5,,,1  
4,allisdo01,1871,1,WS3,9,NA,27,,133,28,44,10,2,2,27,1,1,0,2,,,0  
5,ansonca01,1871,1,RC1,7,NA,25,,120,29,39,11,3,0,16,6,2,2,1,,,0
```

## CSV(Comma-Separated Values)

컬럼을 쉼표로 구분한 텍스트 데이터 파일

[illegible][illegible]

```
1abercda0118711TRO8NA140000000002addybo0118711RC17NA25118303260013814003allisar0118711CL13NA29137284045019312514allisd0118711WS39NA271332844102227110205ansonca0118711RC17NA25120293911301662210
```

```
<TD>17D1</TD>  
<TD>abercda01</TD>  
<TD>18T1</TD>  
<TD>1</TD>  
<TD>TR0</TD>  
<TD>8</TD>  
<TD>-NA</TD>  
<TD>1</TD>  
<TD>C</TD>  
<TD>4</TD>  
<TD>8</TD>  
<TD>8</TD>  
<TD>8</TD>  
<TD>8</TD>  
<TD>8</TD>  
<TD>8</TD>  
<TD>8</TD>  
<TD>8</TD>  
<TD>C</TD>  
<TD>C</TD>  
<TD>C</TD>  
<TD>8</TD>  
<TR>  
<TR><TD>2</TD>  
<TD>addybo01</TD>  
<TD>18T1</TD>  
<TD>1</TD>  
<TD>RC1</TD>
```

## html|모드

조회 결과 출력 형식
ascii
column
csv
html
Insert
list
quote
tabs
tcl

## 기본 출력 모드

# 조회 결과 스프레드시트에서 보기

## 조회 결과 CSV 파일로 저장

```
.mode csv
.output data.csv
```

.output 자신이원하는파일이름.csv

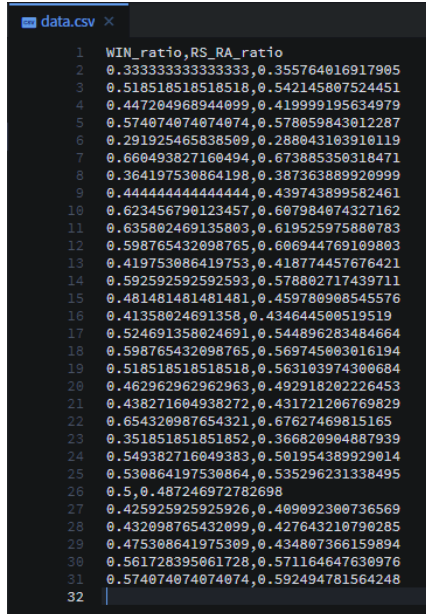
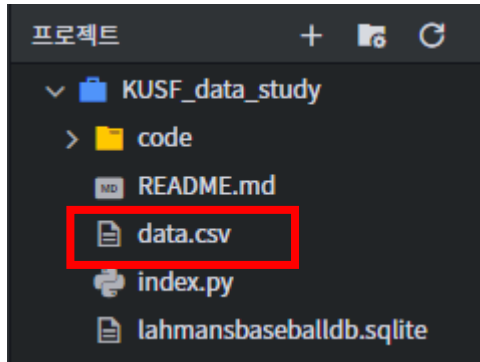
```
.output stdout
```

조회 결과를 화면에 출력

이렇게 세팅한 이후에 쿼리를 실행하면 조회 결과가 data.csv에 저장됨. 화면에 표출되지 않음.

```
SELECT CAST(W as REAL)/(W+L) AS WIN_ratio, CAST(R*R as REAL)/(R*R + RA*RA) AS RS_RA_ratio FROM
teams where yearID = 2019;
```

위 쿼리문을 실행하면, 현재 작업 디렉토리 내에 생성된 data.csv라는 파일 내에 조회 결과가 csv의 형태로 저장됨.

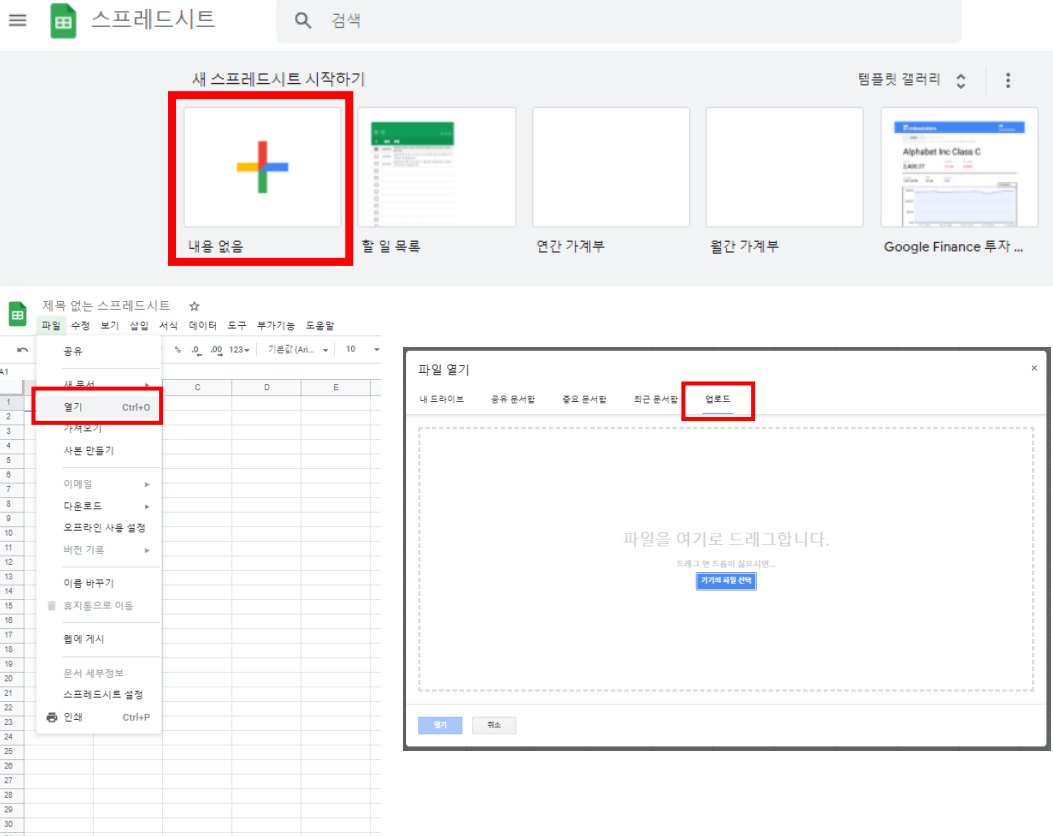
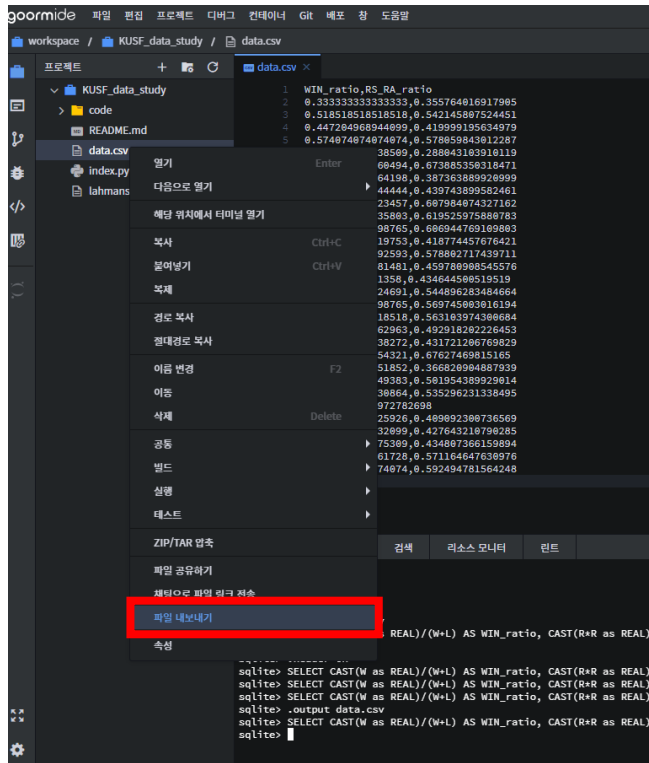


만약 .header on을 안 했으면 컬럼명은 없을 것

이제 data.csv 파일을 스프레드시트 프로그램에서 열어보자

# 조회 결과 스프레드시트에서 보기

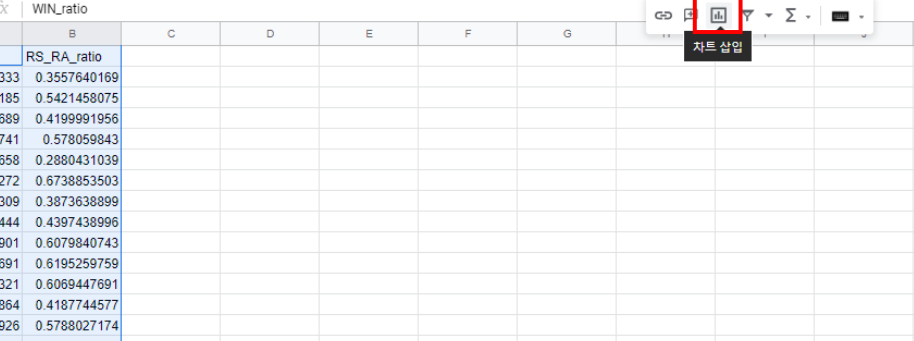
구글스프레드시트에서 해당 csv 파일을 열어보자



data					
파일 수정 보기 삽입 서식 데이터 도구 부가기능 도움말					
100% 123 Arial					
A1	-	fx	WIN_ratio		
	A	B	C	D	E
1	WIN_ratio	RS_RA_ratio			
2	0.3333333333	0.3557640169			
3	0.5185185185	0.5421458075			
4	0.4472049689	0.4199991956			
5	0.5740740741	0.578059843			
6	0.2919254658	0.2880431039			
7	0.6604938272	0.6738853503			
8	0.3641975309	0.3873638899			
9	0.4444444444	0.4397438996			
10	0.6234567901	0.6079840743			
11	0.6358024691	0.6195259759			
12	0.5987654321	0.6069447691			
13	0.4197530864	0.4187744577			
14	0.5925925926	0.5788027174			
15	0.4814814815	0.4597809085			
16	0.4135802469	0.4346445005			
17	0.524691358	0.5448962835			
18	0.5987654321	0.569745003			
19	0.5185185185	0.5631039743			
20	0.462962963	0.4929182022			
21	0.4382716049	0.4317212068			
22	0.6543209877	0.6762746982			
23	0.3518518519	0.3668209049			
24	0.549382716	0.5019543899			
25	0.5308641975	0.5352962313			
26	0.5	0.4872469728			
27	0.4259259259	0.4090923007			
28	0.4320987654	0.4276432108			
29	0.475308642	0.4348073662			
30	0.5617283951	0.5711646476			
31	0.5740740741	0.5924947816			
32					
33					
34					
35					

구글스프레드시트에서 csv 파일 열기 성공

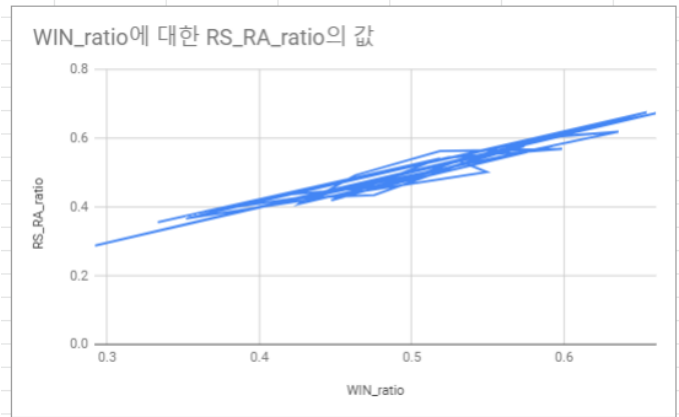
## 승률과 득실점 비율에 대한 산점도를 그려보자



데이터 영역 드래그 한 후에 차트 삽입 클릭

데이터 영역 드래그 한 후에 차트 삽입 클릭

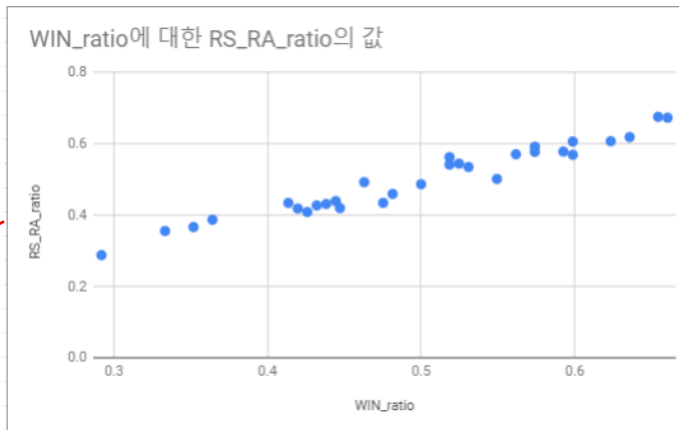
산점도를 통해 승률과 득실점비율은  
강한 상관관계를 갖고 있음을 파악할 수 있음.  
그 정도를 수치화하려면 상관계수를 구해야함.



## 기본 차트인 선 차트가 그려질 것



## 분산형 차트로 바꾸기



## 산점도가 그려질 것

# 승률과 득실점 비율 사이의 상관계수를 구해보자

상관계수 구하는 함수:

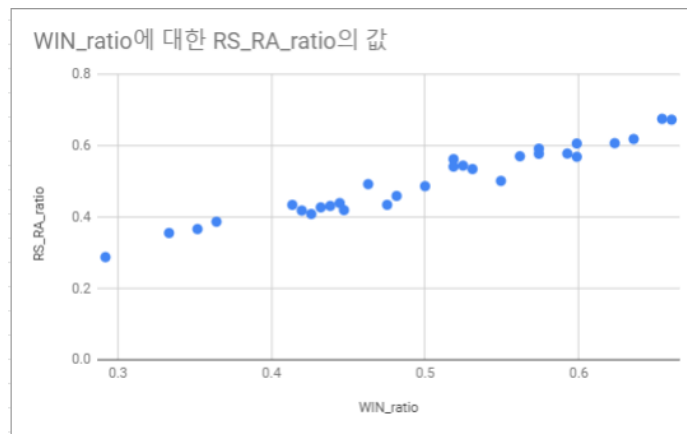
CORREL(첫번째 변수 범위, 두번째 변수 범위)

data ☆ 📄 🔄			
파일 수정 보기 삽입 서식 데이터 도			
100% 🔍 📏 📄 🔄			
B33	=CORREL(A2:A31,B2:B31)		
	A	B	C
1	WIN_ratio	RS_RA_ratio	
2	0.3333333333	0.3557640169	
3	0.5185185185	0.5421458075	
4	0.4472049689	0.4199991956	
5	0.5740740741	0.578059843	
6	0.2919254658	0.2880431039	
7	0.6604938272	0.6738853503	
8	0.3641975309	0.3873638899	
9	0.4444444444	0.4397438996	
10	0.6234567901	0.6079840743	
11	0.6358024691	0.6195259759	
12	0.5987654321	0.6069447691	
13	0.4197530864	0.4187744577	
14	0.5925925926	0.5788027174	
15	0.4814814815	0.4597809085	
16	0.4135802469	0.4346445005	
17	0.524691358	0.5448962835	
18	0.5987654321	0.569745003	
19	0.5185185185	0.5631039743	
20	0.462962963	0.4929182022	
21	0.4382716049	0.4317212068	
22	0.6543209877	0.6762746982	
23	0.3518518519	0.3668209049	
24	0.549382716	0.5019543899	
25	0.5308641975	0.5352962313	
26	0.5	0.4872469728	
27	0.4259259259	0.4090923007	
28	0.4320987654	0.4276432108	
29	0.475308642	0.4348073662	
30	0.5617283951	0.5711646476	
31	0.5740740741	0.5924947816	
32			
33	상관계수	0.9750523791	
34			

fx =CORREL(A2:A31,B2:B31)

A2:A31의 의미 → A컬럼의 2번째 행부터 31번째 행까지

B2:B31의 의미 → B컬럼의 2번째 행부터 31번째 행까지



피어슨 상관계수 0.975, 매우 강한 상관관계  
득점을 많이 하고 실점은 적게 하는 팀의 승률이 좋다고 말할 수 있다

**정말로 득점은 많이 하고 실점은 적게 하는 팀이 자주 이겨?**

**너의 주장을 증명해봐!**

**ANSWER:**

2019년 MLB 팀의 승률과 득실점 비율 사이의 상관계수를 구했더니 0.975로 매우 강한 상관관계를 갖고 있습니다.

따라서, 득점을 많이 하고 실점은 적게 하는 팀이 이길 확률이 매우 높습니다.

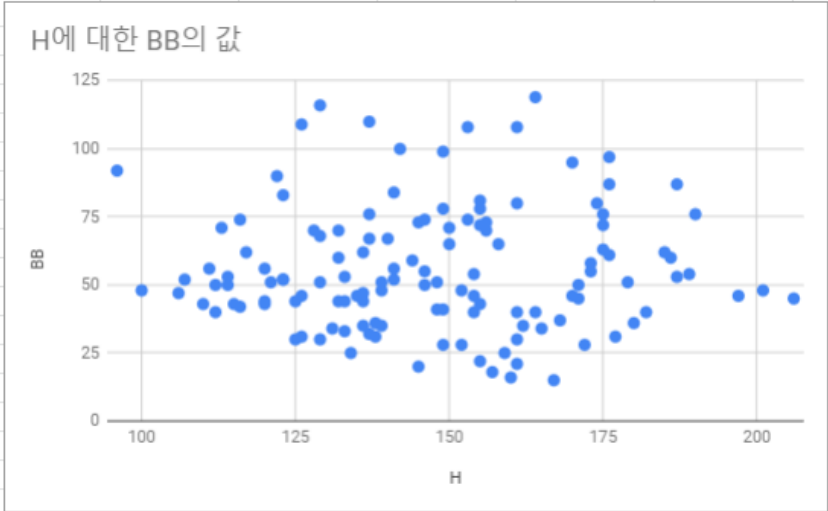
# 실습1: 타자의 볼넷 개수와 안타 개수는 어떠한 상관관계를 갖고 있을까?

2019년 MLB 타자들의 볼넷 개수와 안타 개수를 조회해보자(규정타석 채운 선수만)

```
SELECT H, BB FROM batting WHERE yearID = 2019 and (AB + BB + HBP + SH + SF) >= 502;
```

규정타석 조건

	A	B	C	D	E	F	G	H	I
1	H	BB							
2		180	36						
3		175	76						
4		135	46						
5		141	52						
6		160	16						
7		189	54						
8		155	72						
9		149	41						
10		120	44						
11		167	15						
12		165	34						
13		110	43						
14		185	62						
15		149	28						
16		170	95						
17		146	74						
18		123	83						
19		144	59						
20		176	97						
21		182	40						
22		190	76						
23		111	56						
24		179	51						
25		131	34						
26		164	119						



안타를 치는 능력과 볼넷을 고르는 능력은 거의 상관관계가 없음

상관계수  
0.01476661443

피어슨 상관계수 0.0148, 상관관계 거의 없음

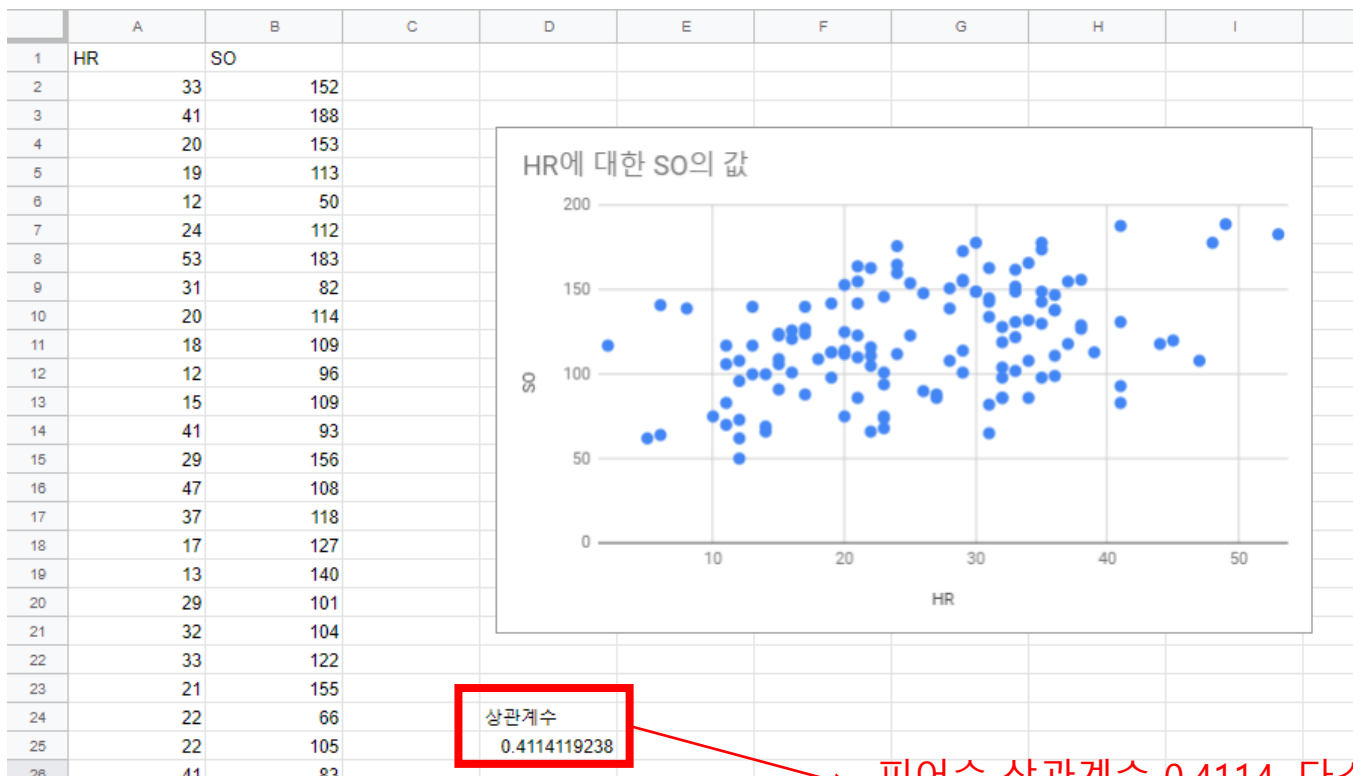
`=CORREL(A2:A131, B2:B131)`



## 실습2: 타자의 홈런 개수와 삼진 개수는 어떠한 상관관계를 갖고 있을까?

2019년 MLB 타자들의 볼넷 개수와 안타 개수를 조회해보자(규정타석 채운 선수만)

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



홈런을 잘 치는 타자는 삼진을 많이 당하는 경향이 있음

피어슨 상관계수 0.4114, 다소 강한 상관관계

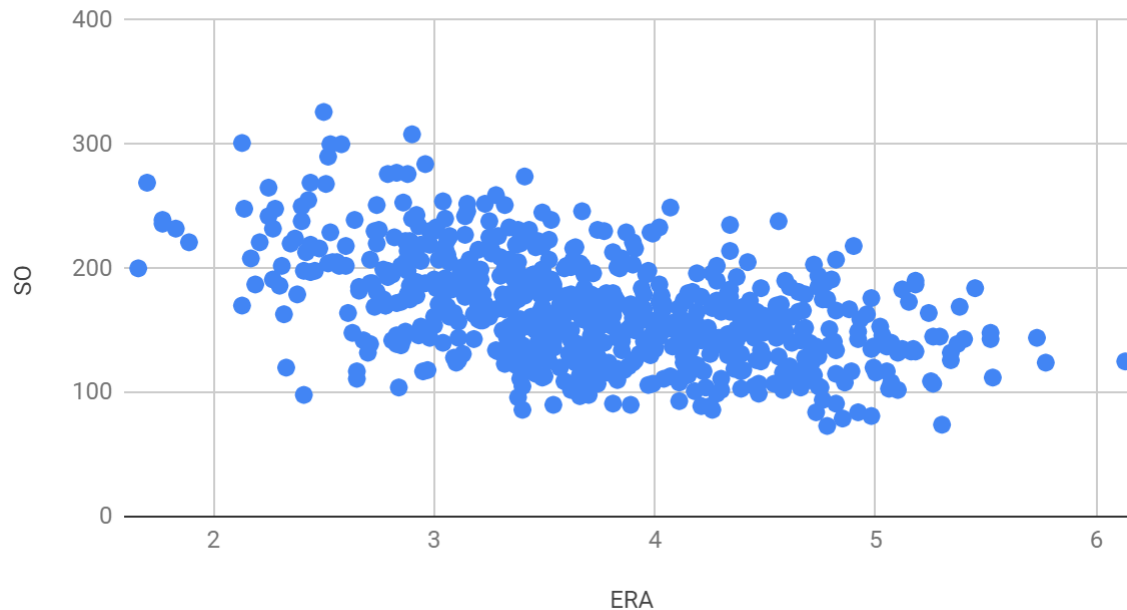
fx | =CORREL(A2:A131, B2:B131)

# Quiz

2010년부터 2019년까지 규정이닝 이상을 던진 투수들의 ERA와 탈삼진 개수 간의 상관계수를 구하라. 산점도도 그려라.

정답:

ERA에 대한 SO의 값



피어슨 상관계수  $-0.5025$ , 다소 강한 음의 상관관계

탈삼진을 잘 잡아내는 투수의 ERA가 좋은 경향이 있다

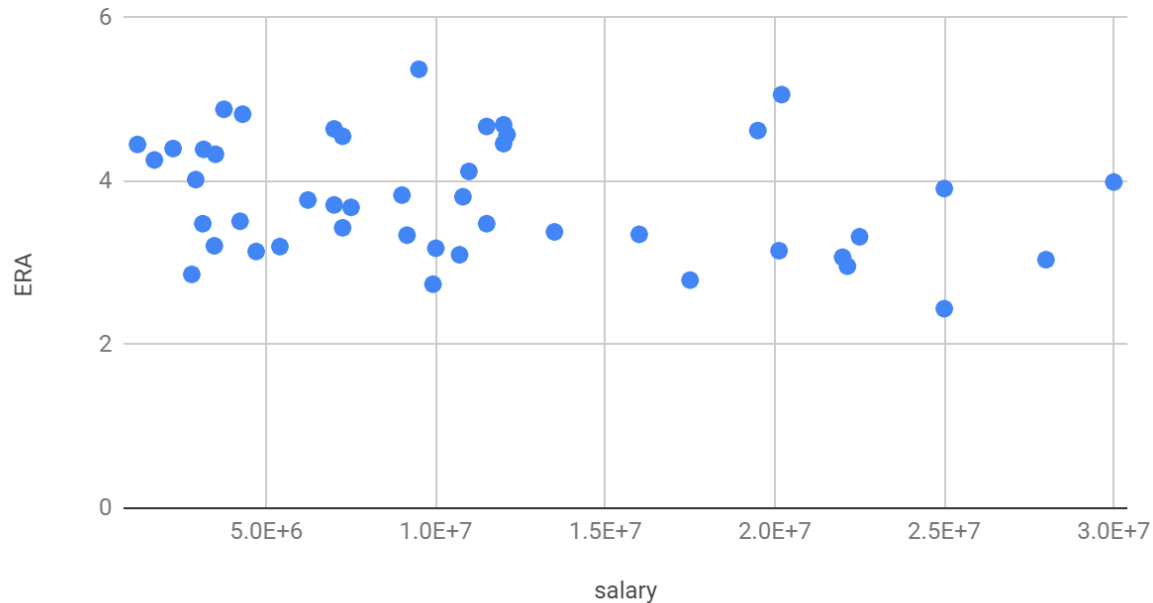
# Quiz 심화

2016년 투수 연봉과 ERA 사이의 상관관계를 파악하라. 산점도를 그리고, 상관계수도 구하라.  
규정이닝 이상 던졌고, 연봉 100만 달러 이상인 선수만.

힌트: JOIN을 사용하는데, JOIN의 조건을 두 개 사용해야함

정답:

salary에 대한 ERA의 값



피어슨 상관계수 -0.2597, 약한 상관관계

연봉과 ERA는 상관관계가 있긴 하지만 약하다.  
즉, 몸값이 비싸다고 꼭 ERA가 좋은 건 아니다.

## 과제#4

2019년 타자 도루와 득점 사이의 상관관계를 파악하라. (규정타석 채운 선수들만)  
산점도를 그리고, 상관계수도 구하라.

**코드와 실행 결과 캡처 화면을 word로 정리해서 [kyohoonsim@gmail.com](mailto:kyohoonsim@gmail.com) 으로 보내주세요~!**

**문서 제목 양식:**

KUSF데이터분석\_과제4\_이름.docx

ex) KUSF데이터분석\_과제4\_심교훈.docx