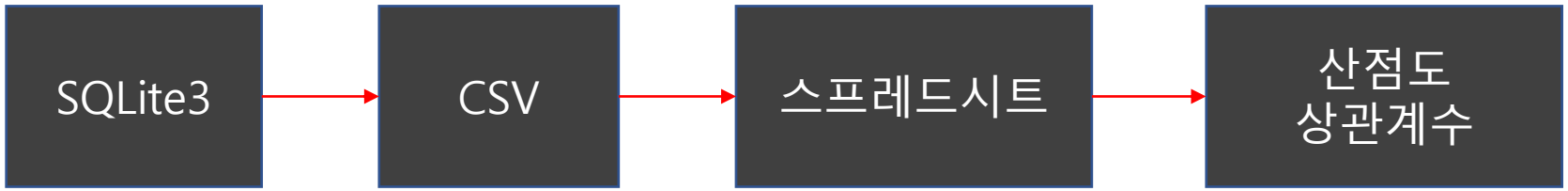
Class 9. 성적 평가에 좋은 스탯

지난 시간에 sqlite3에서 조회한 내용을
csv 파일로 export해서
스프레드시트로 가져온 후에
산점도를 그리고 상관계수를 구하는 분석을 진행했음



이 과정이 꽤 번거롭고 귀찮음

코딩을 하게 되는 주된 동기

다음과 같은 파이썬 코드면 그 모든 과정이 한번에 실행됨

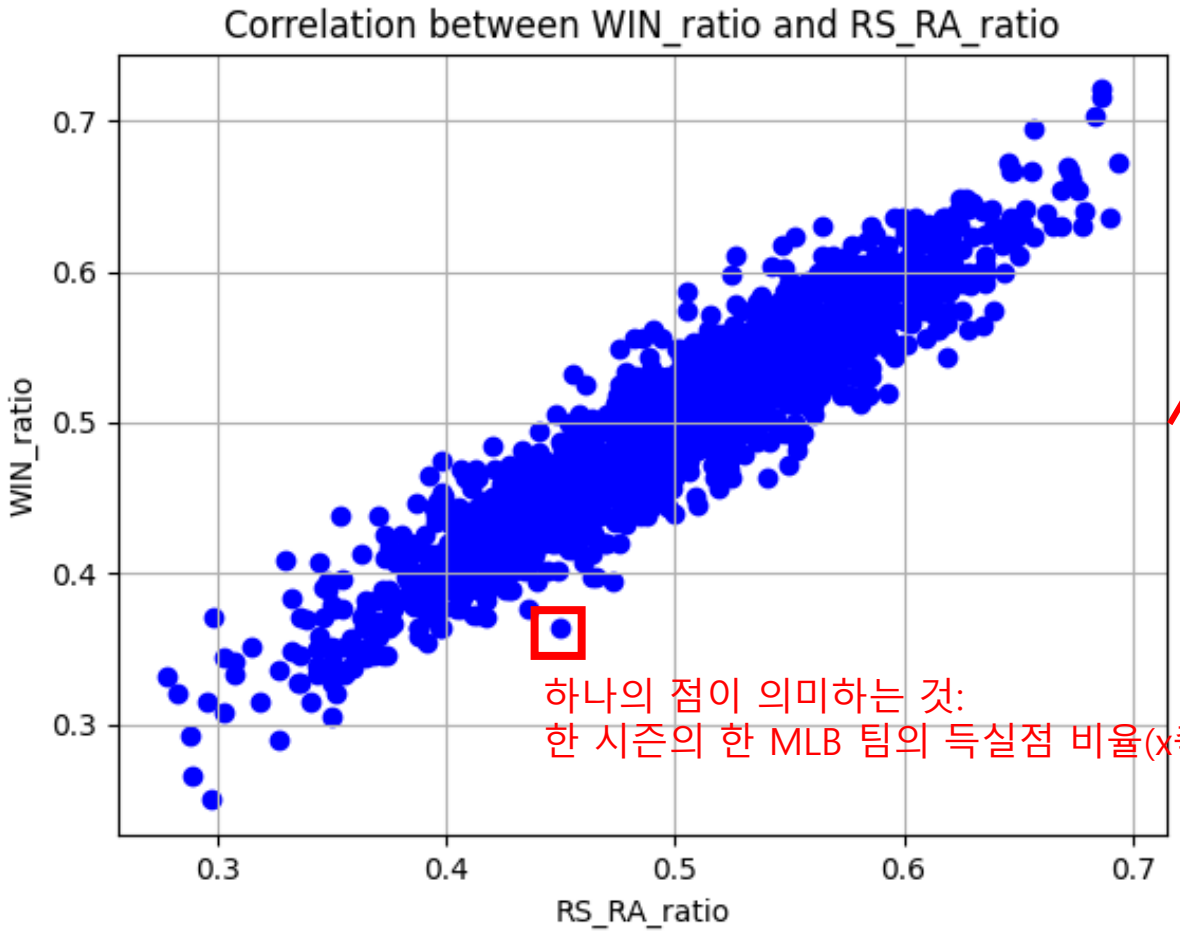
python ex1.py

ex1.py

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

with sqlite3.connect("lahmansbaseballdb.sqlite") as con: # 데이터베이스 연결
    cur = con.cursor() # 커서 객체 생성
    cur.execute('''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 >= 1954;''') # 쿼리문 실행
    result = cur.fetchall() # 쿼리 결과

# 컬럼명 가져오기
cols = []
for column in cur.description:
    cols.append(column[0])

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

# 산점도 그리기
plt.scatter(df['RS_RA_ratio'], df['WIN_ratio'], c='b')
plt.title('Correlation between WIN_ratio and RS_RA_ratio')
plt.xlabel('RS_RA_ratio')
plt.ylabel('WIN_ratio')
plt.grid(True)
plt.savefig('ex1_img.png')

correlation_coefficient = df.corr(method="pearson") # 피어슨 상관계수 계산
print("<상관계수>\n", correlation_coefficient)
```

다음과 같은 파이썬 코드면 그 모든 과정이 한번에 실행됨

산점도와 상관계수가 한번에!



파이썬으로 그린 플롯(plot)

<상관계수>

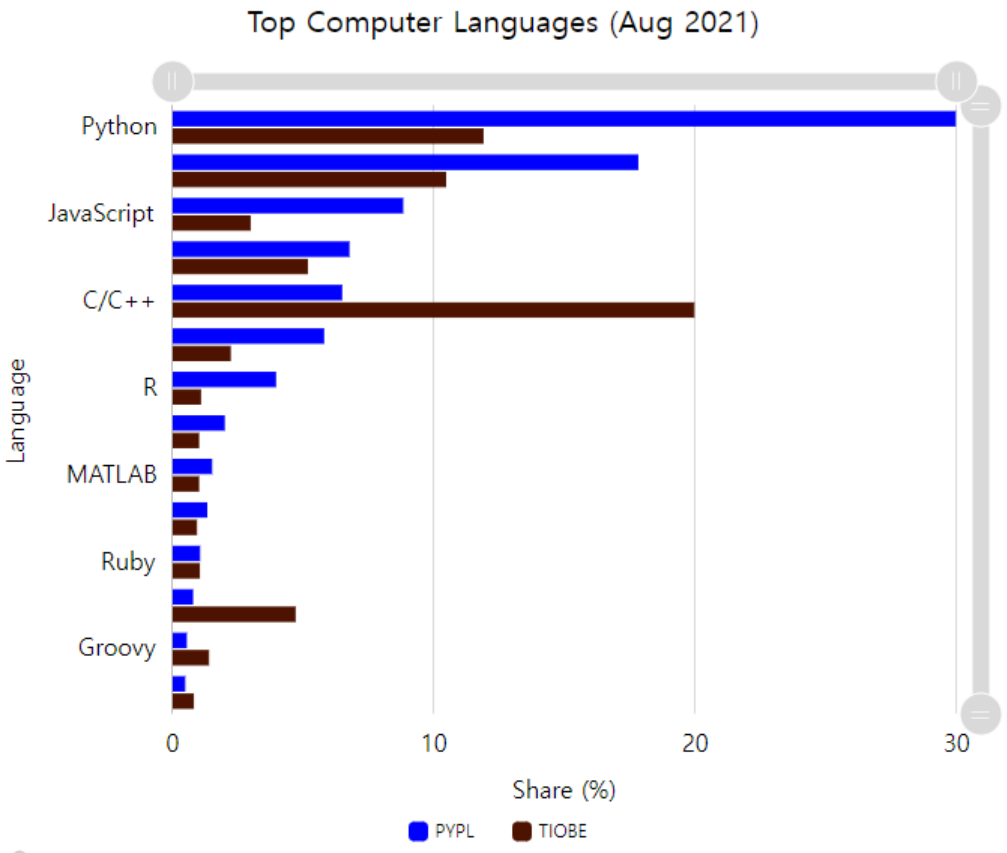
	WIN_ratio	RS_RA_ratio
WIN_ratio	1.000000	0.937159
RS_RA_ratio	0.937159	1.000000

매우 강한 상관관계

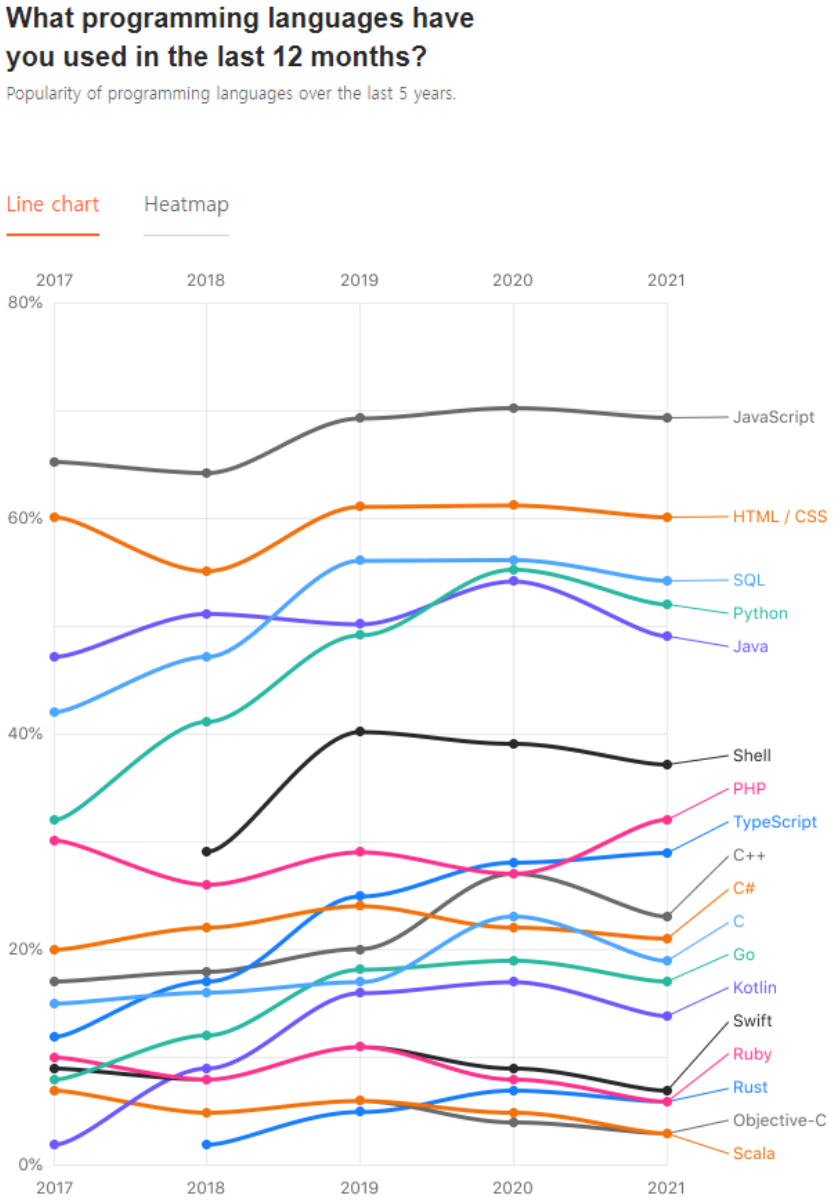
하나의 점이 의미하는 것:
한 시즌의 한 MLB 팀의 득실점 비율(x축)과 팀 승률(y축)

1954년~2019년 MLB 팀들의 승률과 득실점 비율을 산점도(scatter plot)로 나타낸 것

프로그래밍 언어 - 파이썬



출처: statisticstimes



출처: JETBRAINS

파이썬: 현재 가장 인기 있으면서도 배우기 쉬운 언어

프로그래밍 언어도 언어다

영문법 책으로
10년 독학

VS

1년 해외유학

문법을 아는 것도 중요하지만, 직접 써보는 것이 가장 좋음

코드 설명1

ex1.py

필요한 패키지 또는 모듈 불러오기

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

import sqlite3
- DBMS 중 하나인 sqlite3를 사용할 거야

import pandas as pd
- 데이터 분석 도구인 pandas를 사용할 거야
- pandas를 짧게 pd라고 지칭할 거야

import matplotlib.pyplot as plt
- 데이터 시각화 도구인 matplotlib의 pyplot을 사용할 거야
- 짧게 plt라고 부를 거야

코드 설명2

ex1.py

레먼 데이터베이스 연결 및 쿼리문 실행

```
with sqlite3.connect("lahmansbaseballdb.sqlite") as con: # 데이터베이스 연결
    cur = con.cursor() # 커서 객체 생성
    cur.execute('''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 >= 1954;''') # 쿼리문 실행
    result = cur.fetchall() # 쿼리 결과
```

with sqlite3.connect(“데이터베이스파일명”) as con:
sqlite3 데이터베이스 연결

cur = con.cursor()
쿼리문을 실행할 수 있는 능력을 지닌 cursor 객체 생성

cur.execute(‘”쿼리문”’)
쿼리문 실행

result = cur.fetchall()
조회 결과 result라는 변수에 저장

코드 설명3

ex1.py

조회 결과 pandas 데이터프레임에 넣기

```
# 컬럼명 가져오기
cols = []
for column in cur.description:
    cols.append(column[0])

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

cols = []
cols라는 이름의 빈 리스트 생성

for column in cur.description:
**   cols.append(column[0])**
빈 리스트에 조회 결과 컬럼명들 담기

df = pd.DataFrame.from_records(data=result, columns=cols)
조회 결과 pandas 데이터프레임에 담기

코드 설명4

ex1.py

산점도 그리기

```
# 산점도 그리기
plt.scatter(df['RS_RA_ratio'], df['WIN_ratio'], c='b')
plt.title('Correlation between WIN_ratio and RS_RA_ratio')
plt.xlabel('RS_RA_ratio')
plt.ylabel('WIN_ratio')
plt.grid(True)
plt.savefig('ex1_img.png')
```

plt.scatter(df['RS_RA_ratio'], df['WIN_ratio'], c='b')
x축은 득실점 비율, y축은 승률인 산점도 그리기

plt.title('그래프제목')
그래프의 제목 설정

plt.xlabel('x축 라벨'), plt.ylabel('y축 라벨'),
x축, y축 라벨 설정

plt.grid(True)
그래프에 격자 넣기

plt.savefig('이미지파일이름')
그래프 이미지로 저장

코드 설명5

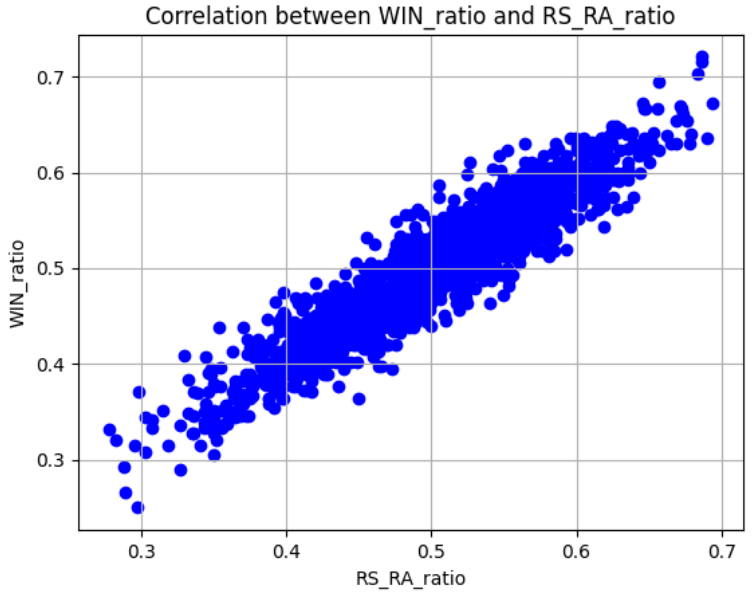
ex1.py

상관계수 구하기

```
correlation_coefficient = df.corr(method="pearson") # 피어슨 상관계수 계산
print("<상관계수>\n", correlation_coefficient)
```

correlation.coefficient = df.corr(method=“pearson”)
피어슨 상관계수 구하기

print(“<상관계수>₩n”, correlation_coefficient)
상관계수 화면에 출력하기



정말 득점을 많이 하고 실점은 적게 하는 팀의 승률이 좋을까? YES

그렇다면 팀 승리를 위해서는 득점에 기여를 많이 하는 선수를 뽑아야 한다는 결론이 나옴

타자의 대표적인 스탯인 타율과 득점은 어느 정도의 상관성을 갖고 있을까?

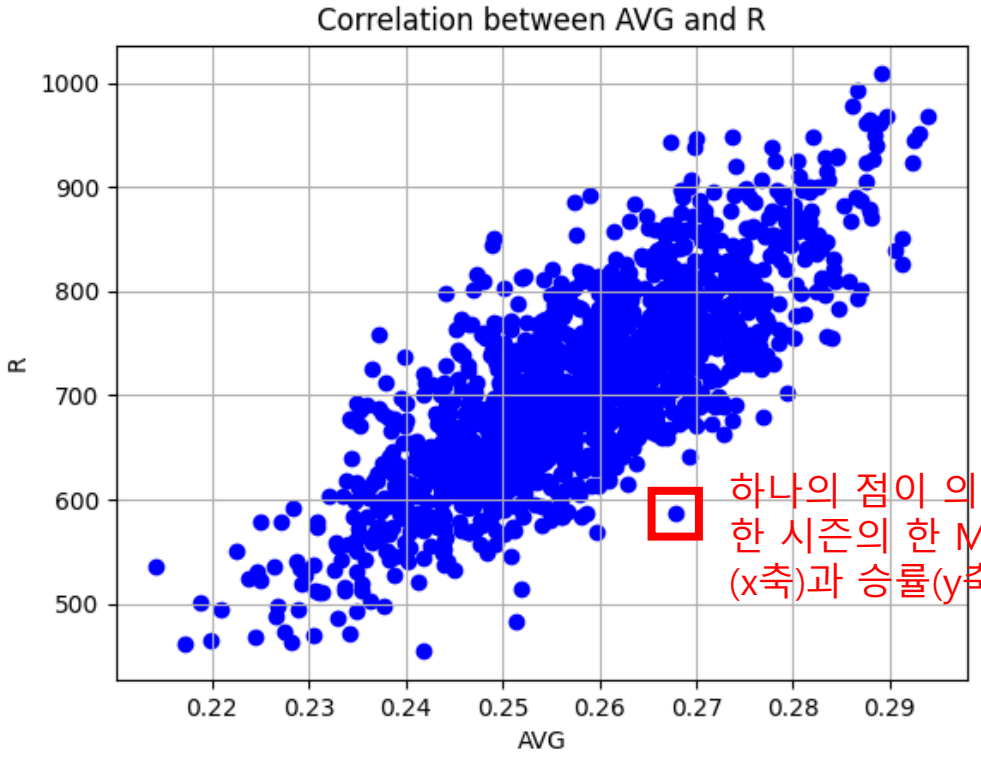
팀 타율과 팀 득점 사이의 상관계수를 구해보자.

파이썬으로 팀 타율과 팀 득점 사이의 상관관계를 파악해보자

팀 타율과 팀 득점 사이의 상관관계

ex2.py

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

with sqlite3.connect("lahmansbaseballdb.sqlite") as con:
    cur = con.cursor()
    cur.execute('''SELECT CAST(H as REAL)/AB AS AVG, R from teams where
yearID >= 1954 and AB > 5000;''')
    result = cur.fetchall()

# 컬럼명 가져오기
cols = []
for column in cur.description:
    cols.append(column[0])

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

plt.scatter(df['AVG'], df['R'], c='b')
plt.title('Correlation between AVG and R')
plt.xlabel('AVG')
plt.ylabel('R')
plt.grid(True)
plt.savefig('ex2_img.png')

correlation_coefficient = df.corr(method="pearson")
print("<상관계수>\n", correlation_coefficient)
```



하나의 점이 의미하는 것:
한 시즌의 한 MLB 팀의 타율(x축)과 승률(y축)

```
<상관계수>
          AVG         R
AVG  1.000000  0.787075
R    0.787075  1.000000
```

강한 상관관계

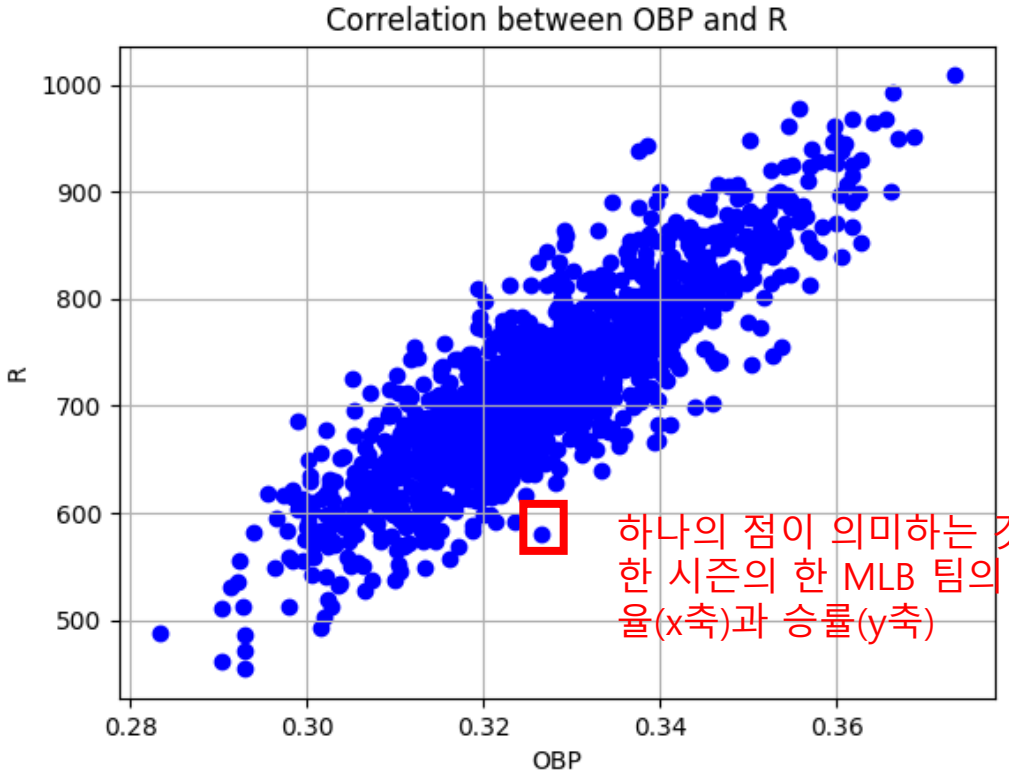
타율이 좋은 선수를 보유하고 있는 것은 득점력을 높일 수 있기 때문에
결과적으로 승리에 도움이 된다는 결론

타율보다 더 득점력과 상관성이 큰 것은 없을까? 출루율은 어떨까?

파이썬으로 팀 출루율과 팀 득점 사이의 상관관계를 파악해보자

팀 출루율과 팀 득점 사이의 상관관계

조건:
1954년 이후
5000 타수 초과



하나의 점이 의미하는 것:
한 시즌의 한 MLB 팀의 출루율(x축)과 승률(y축)

<상관계수>

	OBP	R
OBP	1.000000	0.863649
R	0.863649	1.000000

강한 상관관계

타율보다 출루율이 좋은 선수를 보유하는 것이 팀 득점력을 높이는 데 있어서 낫다는 결론을 내릴 수 있음

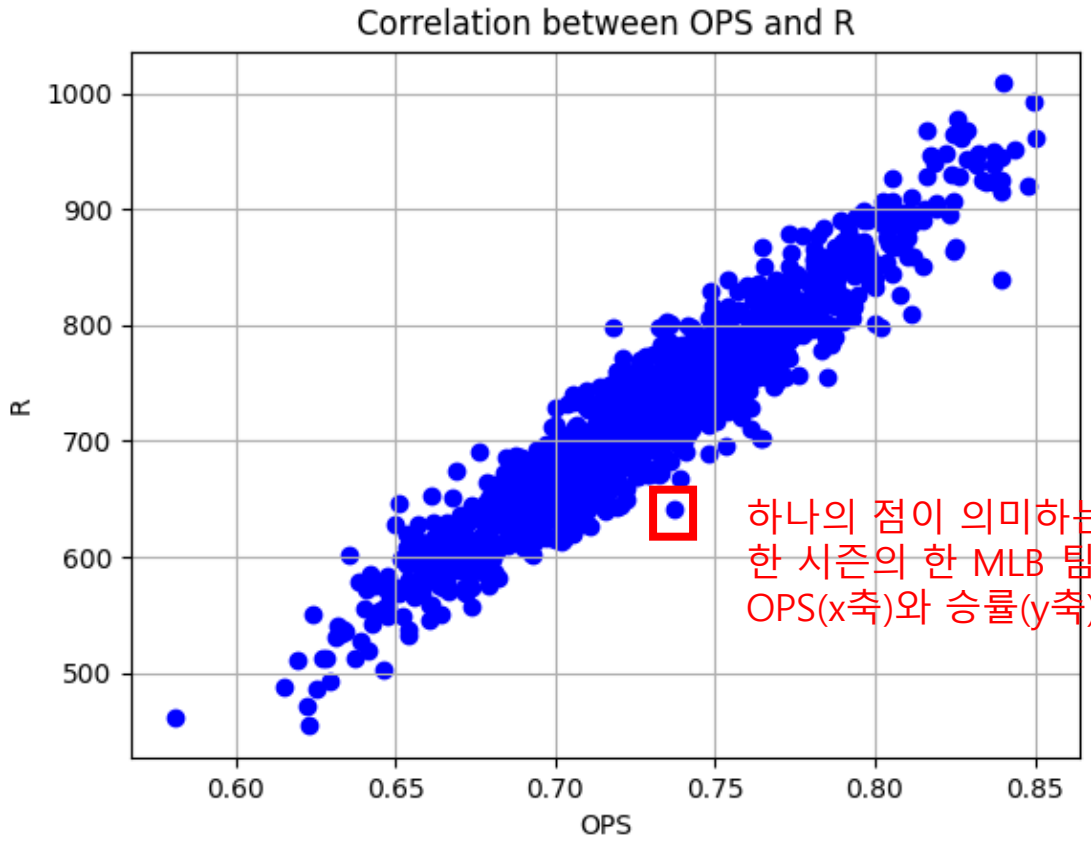
타율보단 출루율, 그렇다면 OPS는 어떨까?

파이썬으로 팀 OPS와 팀 득점 사이의 상관관계를 파악해보자

팀 OPS와 팀 득점 사이의 상관관계

ex4.py

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

with sqlite3.connect("lahmansbaseballdb.sqlite") as con:
    cur = con.cursor()
    cur.execute('''SELECT CAST((H + BB + HBP) AS REAL)/(AB + BB + HBP + SF)
+ CAST(((H - "2B" - "3B" - HR) + 2*"2B" + 3*"3B" + 4*HR) AS REAL)/AB AS
OPS, R from teams where yearID >= 1954 and AB > 5000;''')
    result = cur.fetchall()

cols = []
for column in cur.description:
    cols.append(column[0])

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

plt.scatter(df['OPS'], df['R'], c='b')
plt.title('Correlation between OPS and R')
plt.xlabel('OPS')
plt.ylabel('R')
plt.grid(True)
plt.savefig('ex4_img.png')

correlation_coefficient = df.corr(method="pearson")
print("<상관계수>\n", correlation_coefficient)
```



하나의 점이 의미하는 것:
한 시즌의 한 MLB 팀의
OPS(x축)와 승률(y축)

```
<상관계수>
          OPS         R
OPS  1.000000  0.951253
R    0.951253  1.000000
```

매우 강한 상관관계

타율 < 출루율 < OPS

정리

1. 팀이 승리하기 위해서는 득점을 많이 해야함
2. 득점을 많이 하려면 타율이 좋은 선수가 있어야함
3. 타율보다는 출루율이 좋은 선수가 좀 더 도움이 됨
4. 출루율보다는 OPS가 좋은 선수가 좀 더 도움이 됨

OPS가 좋은 선수를 영입하자!

단장님, OPS 좋은 ___ 사주세요~!!!

다소 위험한 결론. 이 이유는 다음 시간에.

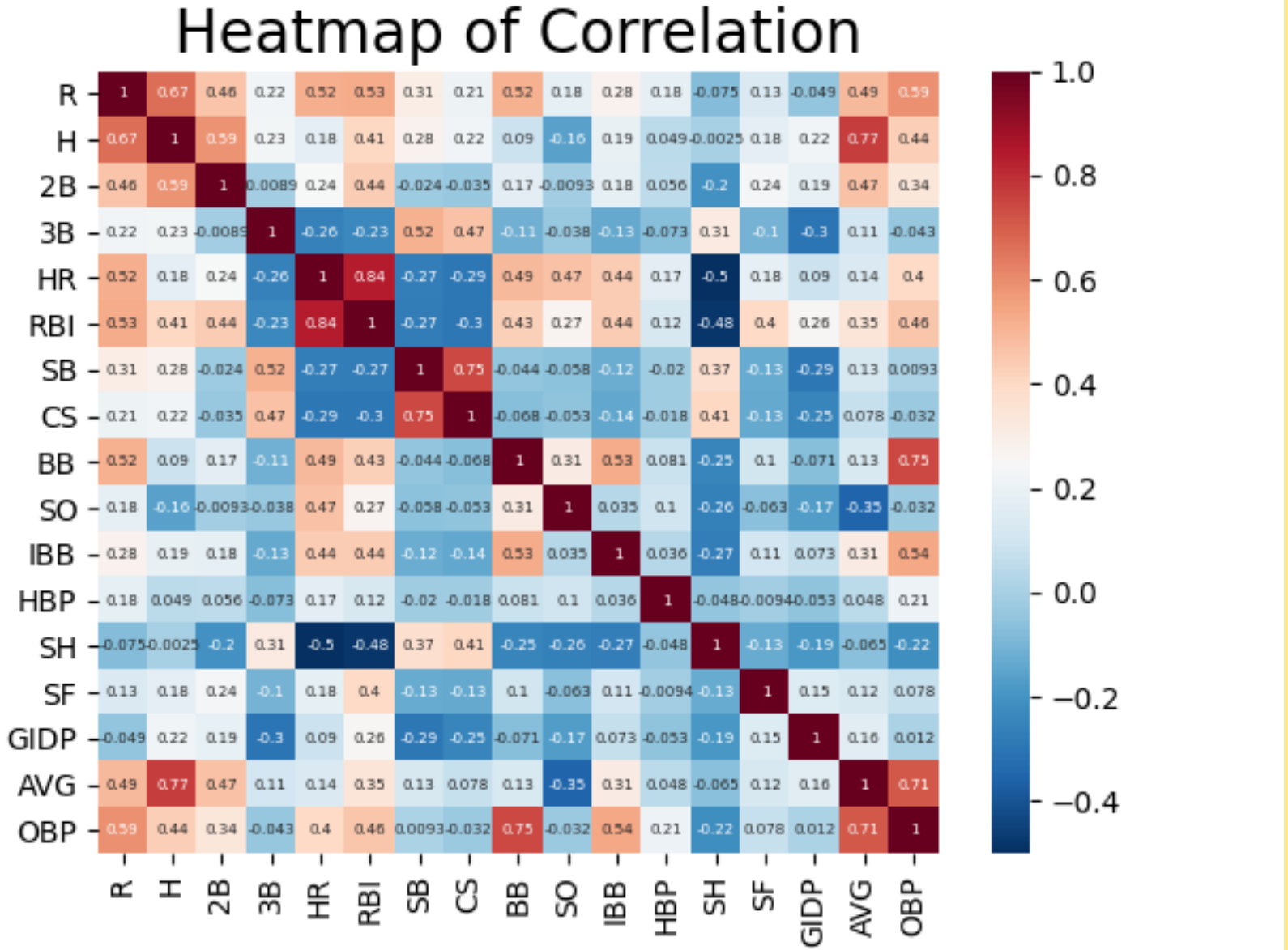
TRY

1. 팀이 승리하기 위해서는 득점을 많이 해야함
2. 득점을 많이 하려면 타율이 좋은 선수가 있어야함
3. 타율보다는 출루율이 좋은 선수가 좀 더 도움이 됨
4. 출루율보다는 OPS가 좋은 선수가 좀 더 도움이 됨
5. **OPS보다 선수를 평가하기에 더 좋은(또는 비슷한) 타자 스탯은?**

타자 기본 스탯들 중 상관성이 큰 것은?

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

with sqlite3.connect("lahmansbaseballdb.sqlite") as con: # 데이터베이스 연결
    cur = con.cursor() # 커서 객체 생성
    cur.execute('''
    SELECT
    R, H, "2B", "3B", HR, RBI, SB, CS, BB, SO, IBB, HBP, SH, SF,
    GIDP, (H+0.0)/AB AS AVG,
    (H + BB + HBP + 0.0)/(AB + BB + HBP + SF) AS OBP
    FROM batting WHERE yearID >= 2000 AND (AB + BB + HBP + SF) >= 502;
    ''') # 쿼리문 실행
    result = cur.fetchall() # 쿼리 결과

cols = []
for column in cur.description:
    cols.append(column[0])
df = pd.DataFrame.from_records(data=result, columns=cols)

print(df)

correlation_coefficient = df.corr(method="pearson") # 피어슨 상관계수 계산
print(correlation_coefficient)

ax = sns.heatmap(correlation_coefficient, annot=True, cmap="RdBu_r", annot_kws={"fontsize":5})

cbar = ax.collections[0].colorbar

plt.title('Heatmap of Correlation', fontsize=20)
plt.savefig('heatmap of correlation.png')
```

타자 기본 스탯들 중 상관성이 큰 것은?



홈런과 타점: 0.84
안타와 타율: 0.77
도루실패와 도루성공: 0.75
볼넷과 출루율: 0.75
안타와 득점: 0.67
득점과 출루율: 0.59
안타와 2루타: 0.59