

Class 10. 성적 예측에 좋은 스탓

질문!

그런데 과연 당해 OPS가 좋았던 선수는 다음 시즌 OPS도 좋을까?

비싼 돈 주고 영입했는데 OPS가 확 나빠진다면??

“선수들의 지난해 OPS와 올해 OPS 간의 상관관계”를 살펴보자

타자들의 지난해 OPS와 올해 OPS 간의 상관관계

x	y
2010년 추신수 OPS	2011년 추신수 OPS
2011년 추신수 OPS	2012년 추신수 OPS
2012년 추신수 OPS	2013년 추신수 OPS
2013년 추신수 OPS	2014년 추신수 OPS
...	...
2010년 트라웃 OPS	2011년 트라웃 OPS
2011년 트라웃 OPS	2012년 트라웃 OPS
2012년 트라웃 OPS	2013년 트라웃 OPS
2013년 트라웃 OPS	2014년 트라웃 OPS
...	...

SQL 만으로 이렇게 조회하는 것은 매우 어렵다

파이썬으로 가지고 와서 이러한 형태가 되도록 처리를 해줘야 함

지난해 OPS와 올해 OPS 간의 상관관계

ex5.py

코드 상세 설명!

```
import sqlite3
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib import font_manager, rc
from scipy import stats
with sqlite3.connect("lahmansbaseballdb.sqlite") as con:
    cur = con.cursor()
    cur.execute('''
        SELECT playerID, yearID, 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
        FROM batting WHERE yearID >= 1990 and (AB + BB + HBP + SH + SF) >= 502 ORDER BY playerID;
    ''')
    result = cur.fetchall()

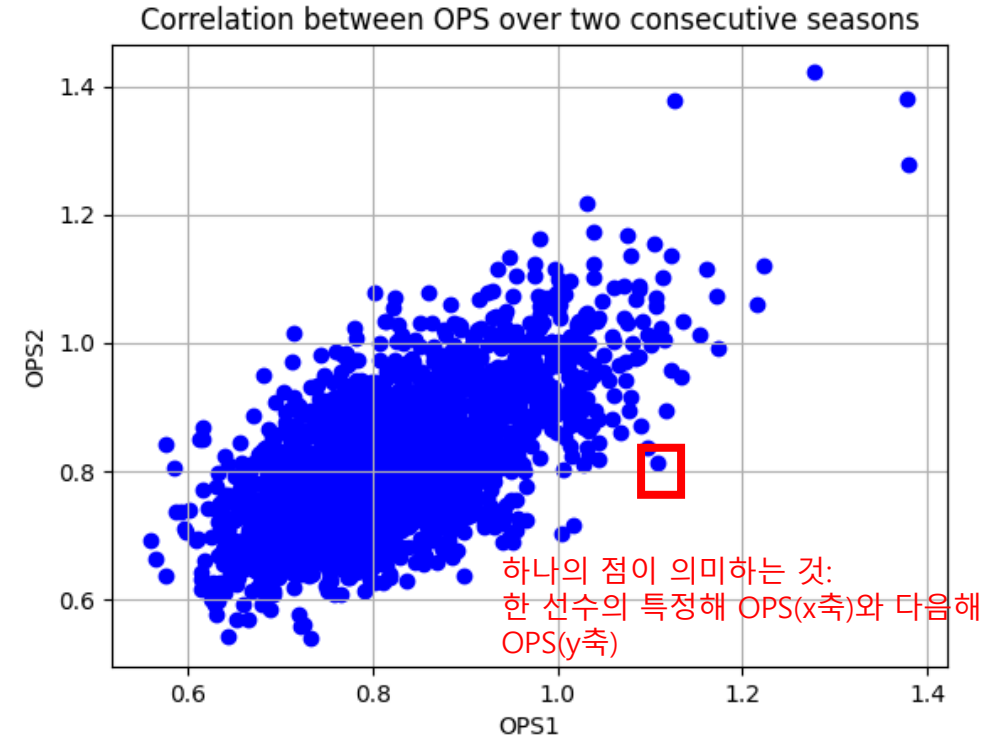
cols = [column[0] for column in cur.description] # 컬럼명 가져오기

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

before = []
after = []

for i in range(len(df)-1):
    if df.iloc[i+1, 0] == df.iloc[i, 0]:
        if df.iloc[i+1, 1] == df.iloc[i, 1] + 1:
            before.append(df.iloc[i, 2])
            after.append(df.iloc[i+1, 2])

plt.scatter(before, after, c='b')
plt.title('Correlation between OPS over two consecutive seasons')
plt.xlabel('OPS1')
plt.ylabel('OPS2')
plt.grid(True)
plt.savefig('ex5_img.png')
correlation_coefficient = stats.pearsonr(before, after)
print("상관계수:", correlation_coefficient[0])
```



상 관계 수 : 0.6525293122819918

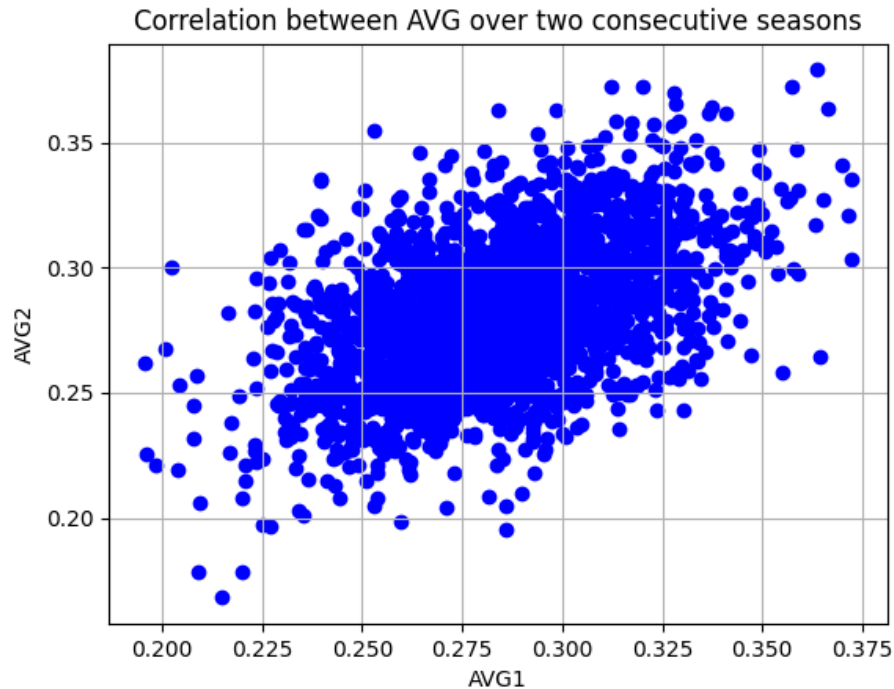
다소 강한 상관관계

특정해에 OPS가 좋았던 선수는 다음 해에도 OPS가 좋을 가능성이 있다

OPS보다 좀 더 다음 해 성적을 예측하기 좋은 스탯은 무엇일까?

조건:
1990년 이후

타율



상관계수 : 0.4905995833790807

다소 강한 상관관계

타율은 다음 시즌의 성적을 예측하는데 있어서 OPS보다 신뢰도가 떨어진다

OPS보다 좀 더 다음 해 성적을 예측하기 좋은 스탯은 무엇일까?

ex7.py

```
import sqlite3
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib import font_manager, rc
from scipy import stats

with sqlite3.connect("lahmansbaseballdb.sqlite") as con:
    cur = con.cursor()
    cur.execute('''
        SELECT playerID, yearID, (HR + 0.0)/(AB + BB + HBP + SH + SF) AS "HR%"
        FROM batting WHERE yearID >= 1990 and (AB + BB + HBP + SH + SF) >= 502 ORDER BY playerID;
    ''')
    result = cur.fetchall()

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

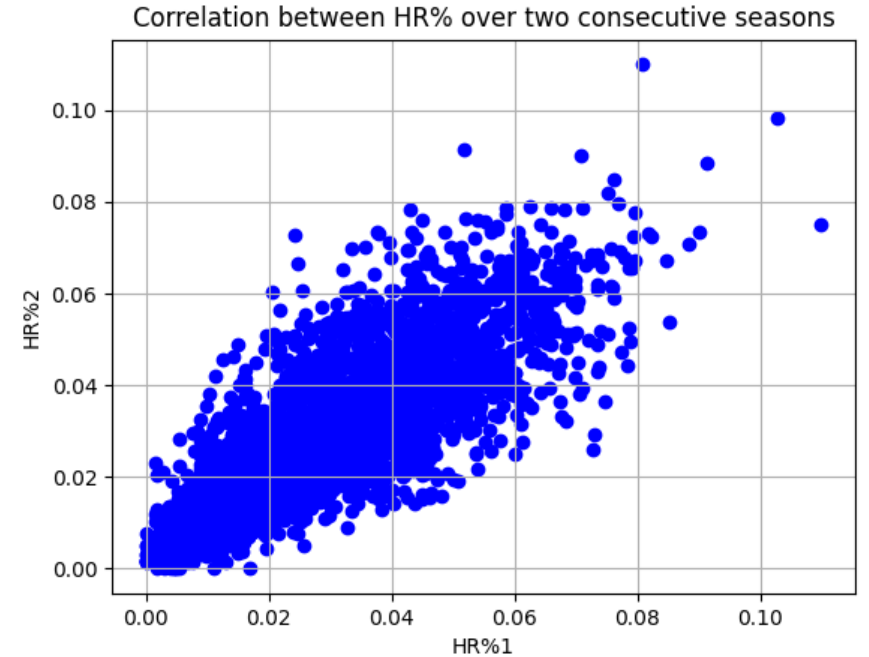
df = pd.DataFrame.from_records(data=result, columns=cols)

before = []
after = []

for i in range(len(df)-1):
    if df.iloc[i+1, 0] == df.iloc[i, 0]:
        if df.iloc[i+1, 1] == df.iloc[i, 1] + 1:
            before.append(df.iloc[i, 2])
            after.append(df.iloc[i+1, 2])

plt.scatter(before, after, c='b')
plt.title('Correlation between HR% over two consecutive seasons')
plt.xlabel('HR%1')
plt.ylabel('HR%2')
plt.grid(True)
plt.savefig('ex7_img.png')
correlation_coefficient = stats.pearsonr(before, after)
print("상관계수:", correlation_coefficient[0])
```

타석당 홈런 비율(HR%)



상관계수 : 0.7314803303478029

강한 상관관계

이번 시즌에 홈런을 잘 친 타자는 다음 시즌에도 홈런을 잘 칠 확률이 높다 (HR% > OPS > AVG)

정리

OPS와 같이 성적을 평가하기에 좋은 스탯이 있고

HR%와 같이 내년 성적을 예측하기에 좋은 스탯이 있다

TRY

1. 타자 스탯 중 HR%보다 다음 해 성적을 예측하는데 더 신뢰할 만한 스탯은?
2. 투수 스탯 중 다음 해 성적을 예측하는데 신뢰할 만한 스탯은?
3개 이상의 스탯을 시도해보라