Class 6. 평가 vs 예측

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 TOR 승률	2010년 TOR 득실점비율
2010년 CIN 승률	2010년 CIN 득실점비율
2010년 NYN 승률	2010년 NYN 득실점비율
2010년 WAS 승률	2010년 WAS 득실점비율
…	…
2015년 TOR 승률	2015년 TOR 득실점비율
2015년 CIN 승률	2015년 CIN 득실점비율
2015년 NYN 승률	2015년 NYN 득실점비율
2015년 WAS 승률	2015년 WAS 득실점비율
…	…

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
…	…

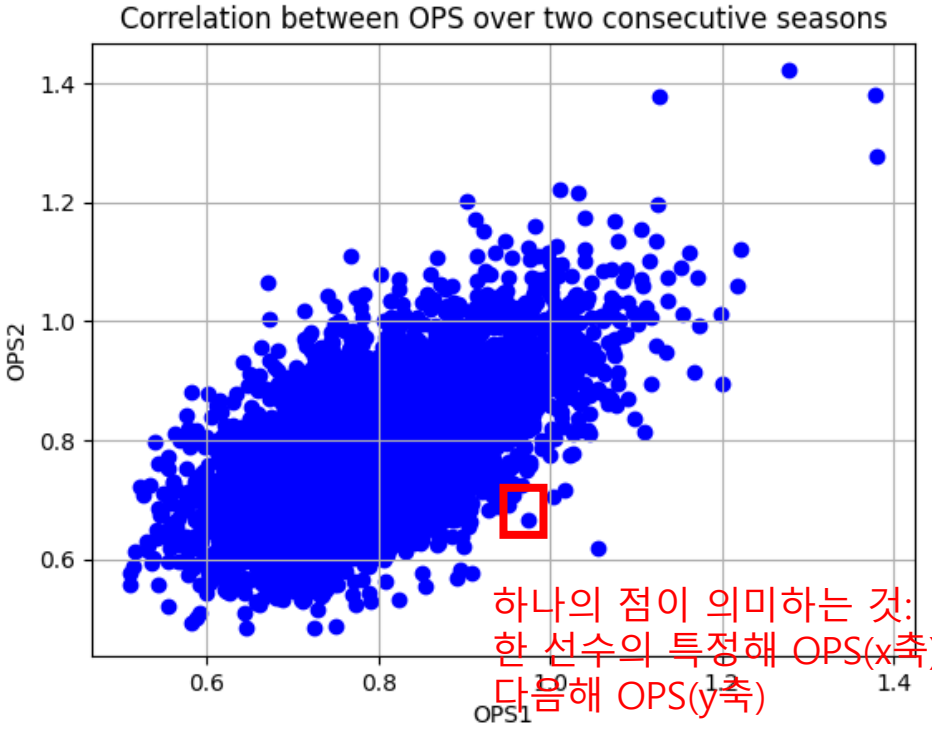
SQL 만으로 이렇게 조회하는 것은 매우 어렵다

파이썬으로 가지고 와서 이러한 형태가 되도록 처리를 해줘야 함

지난해 OPS와 올해 OPS 간의 상관관계

ex5.py

```python
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 > 250
        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])
```



하나의 점이 의미하는 것:
한 선수의 특정해 OPS(x축)와
다음해 OPS(y축)

```
상관계수: 0.5917067453629462
```

다소 강한 상관관계

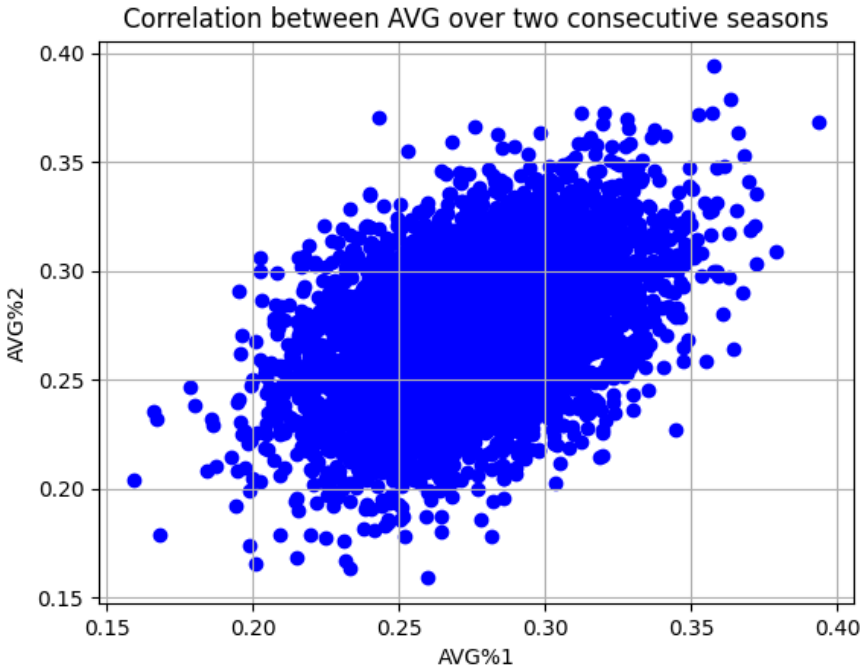
특정해에 OPS가 좋았던 선수는 다음 해에도 OPS가 좋을 가능성이 있다

OPS보다 좀 더 다음 해 성적을 예측하기 좋은 스탯은 무엇일까?

ex6.py

```python
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 AS REAL)/AB AS AVG
        FROM batting
        WHERE yearID >= 1990 and AB > 250
        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 AVG over two consecutive seasons')
plt.xlabel('AVG%1')
plt.ylabel('AVG%2')
plt.grid(True)
plt.savefig('ex6_img.png')

correlation_coefficient = stats.pearsonr(before, after)
print("상관계수:", correlation_coefficient[0])
```

타율




```
상관계수: 0.4431415430371724
```

다소 강한 상관관계

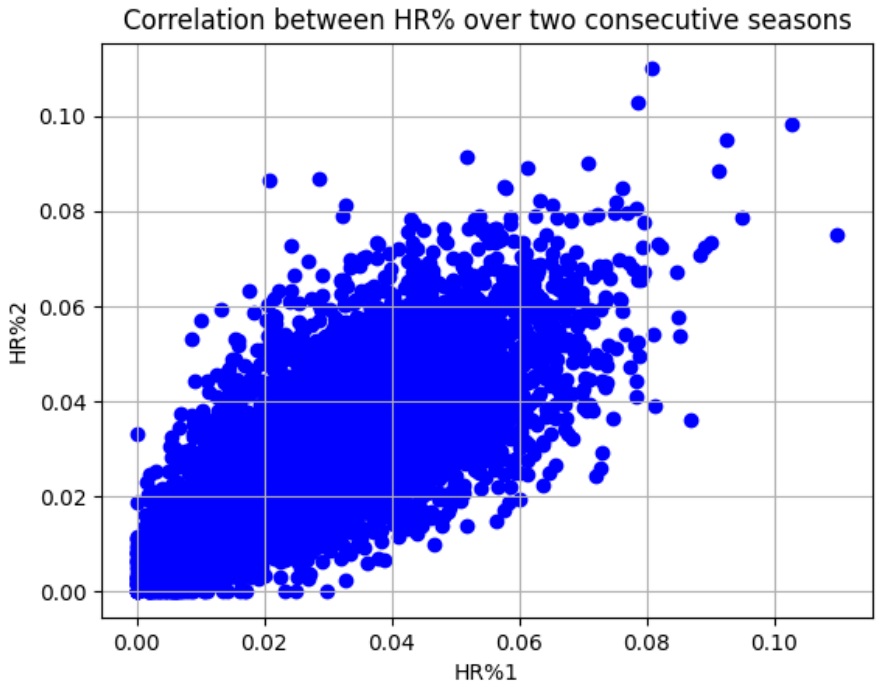
타율은 다음 시즌의 성적을 예측하는데 있어서 OPS보다 신뢰도가 떨어진다

OPS보다 좀 더 다음 해 성적을 예측하기 좋은 스탯은 무엇일까?

ex7.py

```python
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(HR AS REAL)/CAST((AB + BB + HBP + SH + SF) AS REAL) AS HRpercentage
        FROM batting
        WHERE yearID >= 1990 and AB > 250
        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)

OPS보다는 HR%가 다음 해 성적을 예측하는데 더 신뢰할 만하다

키움 히어로즈

선수명: 박병호
생년월일: 1986년 07월 10일
신장/체중: 185cm/107kg
입단 계약금: 33000만원
지명순위: 05 LG 1차

등번호: No.52
포지션: 내야수(우투우타)
경력: 영일초(광명리틀)-영남중-성남고-LG-상무-LG-넥센
연봉: 150000만원
입단년도: 05LG

타자 | 투수 | 기록용어

기본기록 | 통산기록 | 일자별기록 | 경기별기록 | 상황별기록 | 등록일수

KBO 정규시즌

KBO 정규시즌

연도	팀명	AVG	G	PA	AB	R	H	2B	3B	HR	TB	RBI	SB	CS	BB	HBP	SO	GDP	SLG	OBP	E
2005	LG	0.190	79	185	163	22	31	11	0	3	51	21	1	0	12	6	48	3	0.313	0.265	3
2006	LG	0.162	48	142	130	7	21	2	0	5	38	13	1	3	9	2	42	4	0.292	0.227	3
2009	LG	0.218	68	213	188	28	41	7	0	9	75	25	2	1	20	4	70	3	0.399	0.305	1
2010	LG	0.188	78	192	160	25	30	4	0	7	55	22	5	1	26	2	55	5	0.344	0.305	0
2011	넥센	0.254	66	230	201	31	51	11	2	13	105	31	2	0	26	2	76	5	0.522	0.343	5
2012	넥센	0.290	133	560	469	76	136	34	0	31	263	105	20	9	73	11	111	6	0.561	0.393	7
2013	넥센	0.318	128	556	450	91	143	17	0	37	271	117	10	2	92	8	96	7	0.602	0.437	5
2014	넥센	0.303	128	571	459	126	139	16	2	52	315	124	8	3	96	12	142	13	0.686	0.433	4
2015	넥센	0.343	140	622	528	129	181	35	1	53	377	146	10	3	78	12	161	10	0.714	0.436	12
2018	넥센	0.345	113	488	400	88	138	20	0	43	287	112	0	1	68	17	114	9	0.718	0.457	11
2019	키움	0.280	122	532	432	92	121	22	0	33	242	98	0	1	78	13	117	7	0.560	0.398	6
2020	키움	0.223	93	383	309	56	69	7	0	21	139	66	0	0	57	9	114	8	0.450	0.352	6
2021	키움	0.226	114	460	394	46	89	21	0	20	170	73	0	1	45	14	135	9	0.431	0.322	6
통산		0.278	1310	5134	4283	817	1190	207	5	327	2388	953	59	25	680	112	1281	89	0.558	0.386	69

대표적인 예) 올해 KBO의 박병호 선수, 타율은 2할2푼대로 매우 저조하지만, 그래도 20개의 홈런을 때렸음

총 정리

OPS와 같이 성적을 평가하기에 좋은 스탯이 있고

HR%와 같이 내년 성적을 예측하기에 좋은 스탯이 있다

과제#6

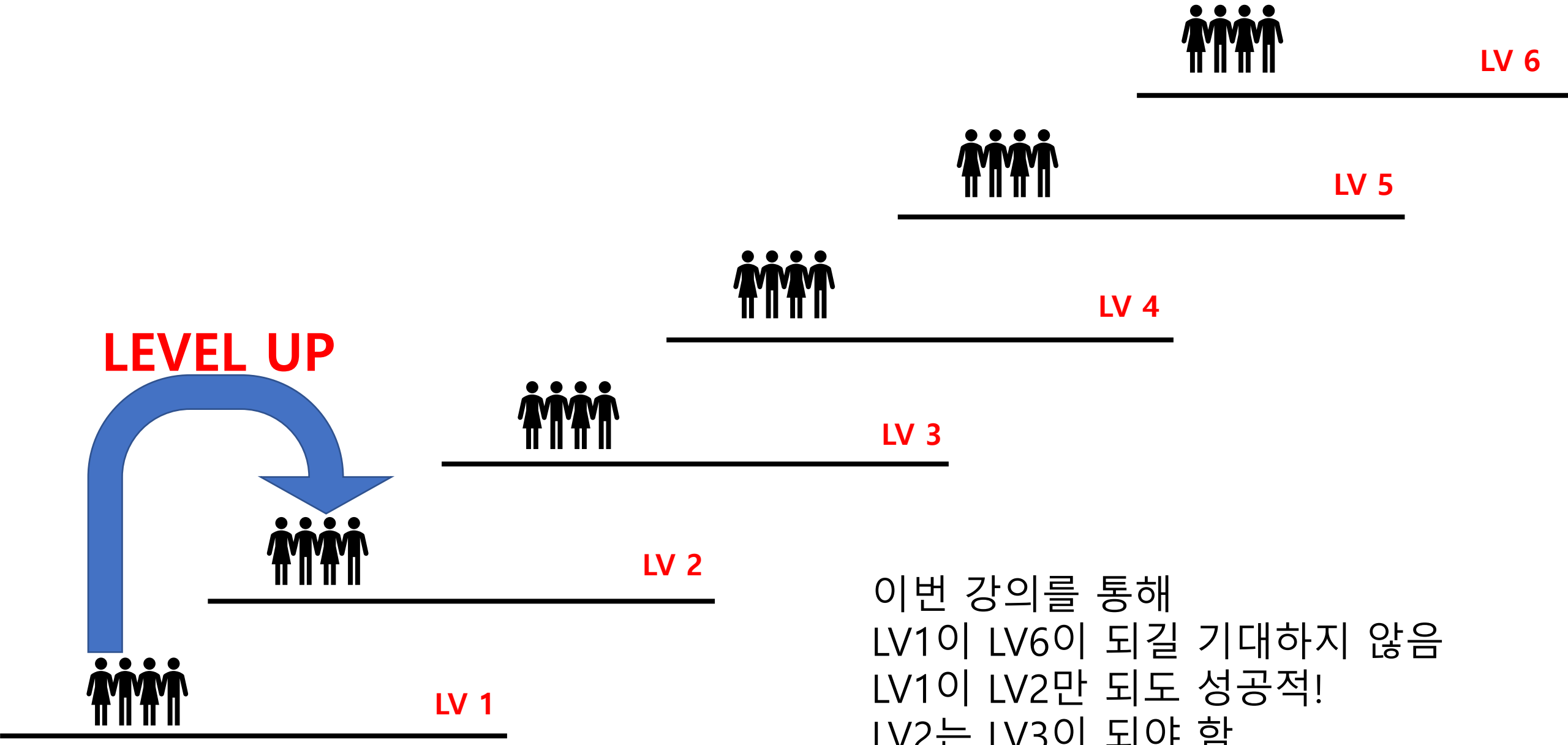
HR%과 같이 연속된 시즌간 상관관계가 강한 스탯을 찾아보라.

코드와 실행 결과 캡쳐 화면을 word로 정리해서 kyohoonsim@gmail.com 으로 보내주세요~!

문서 제목 양식:
KUSF데이터분석_과제6_이름.docx

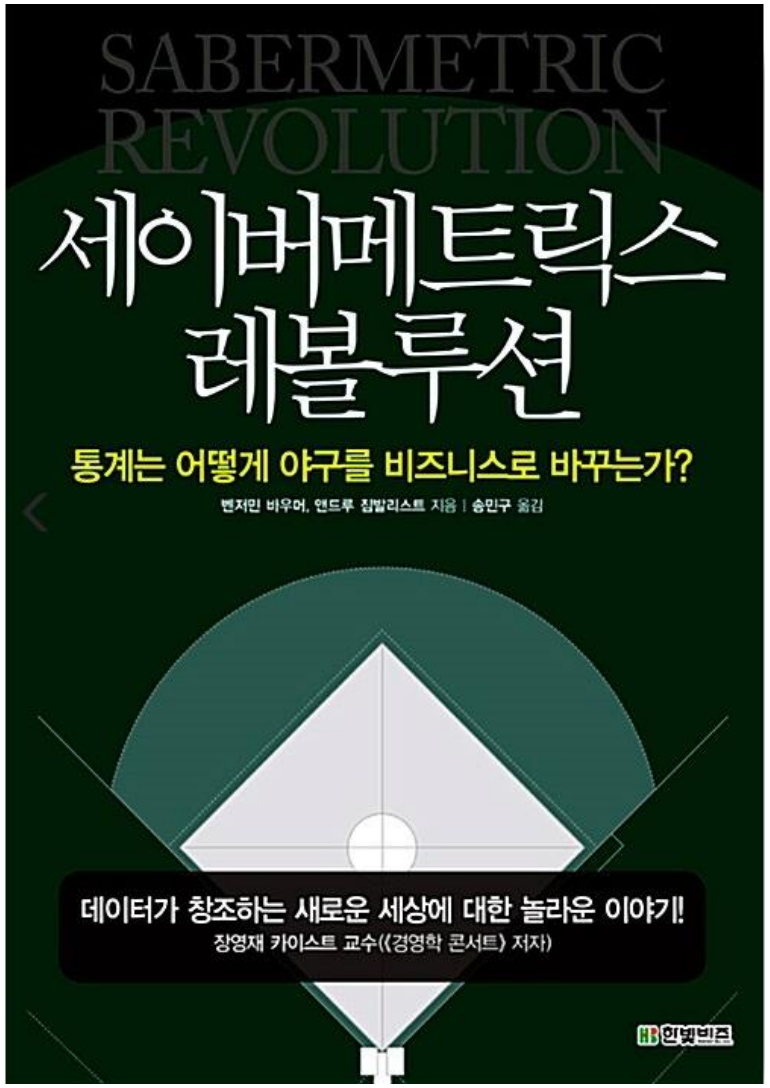
ex) KUSF데이터분석_과제6_심교훈.docx

이제 절반 왔습니다:D
수고하셨습니다

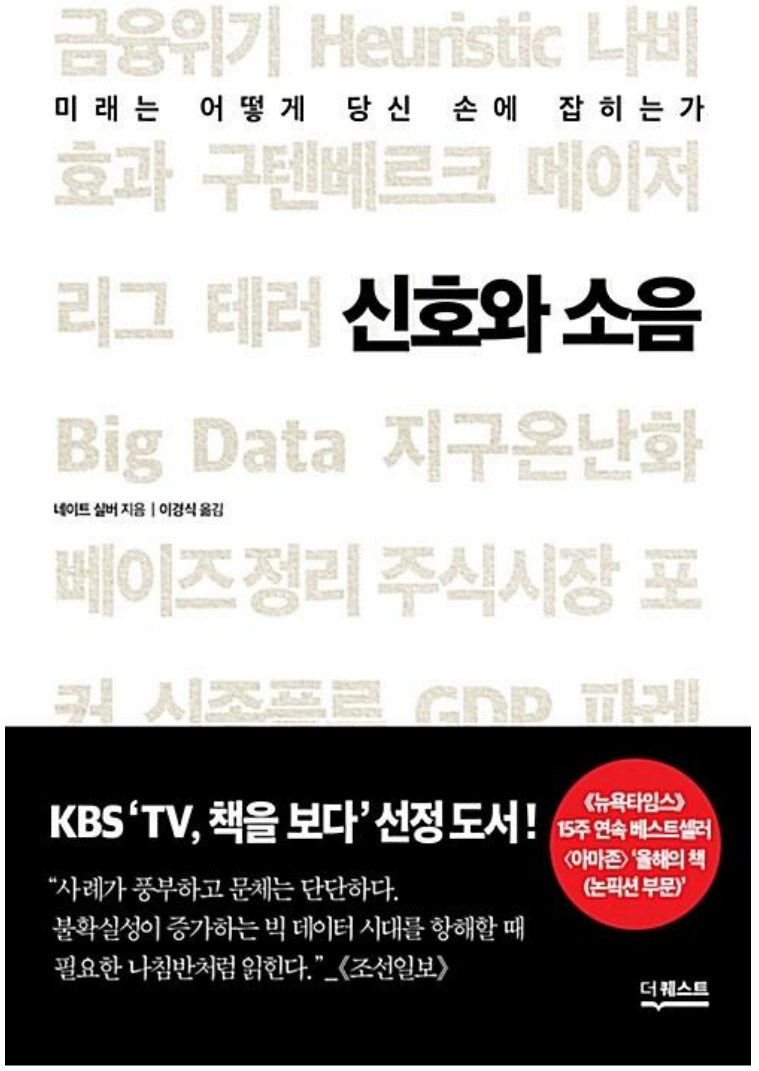


이번 강의를 통해
LV1이 LV6이 되길 기대하지 않음
LV1이 LV2만 되도 성공적!
LV2는 LV3이 되야 함

추천 도서

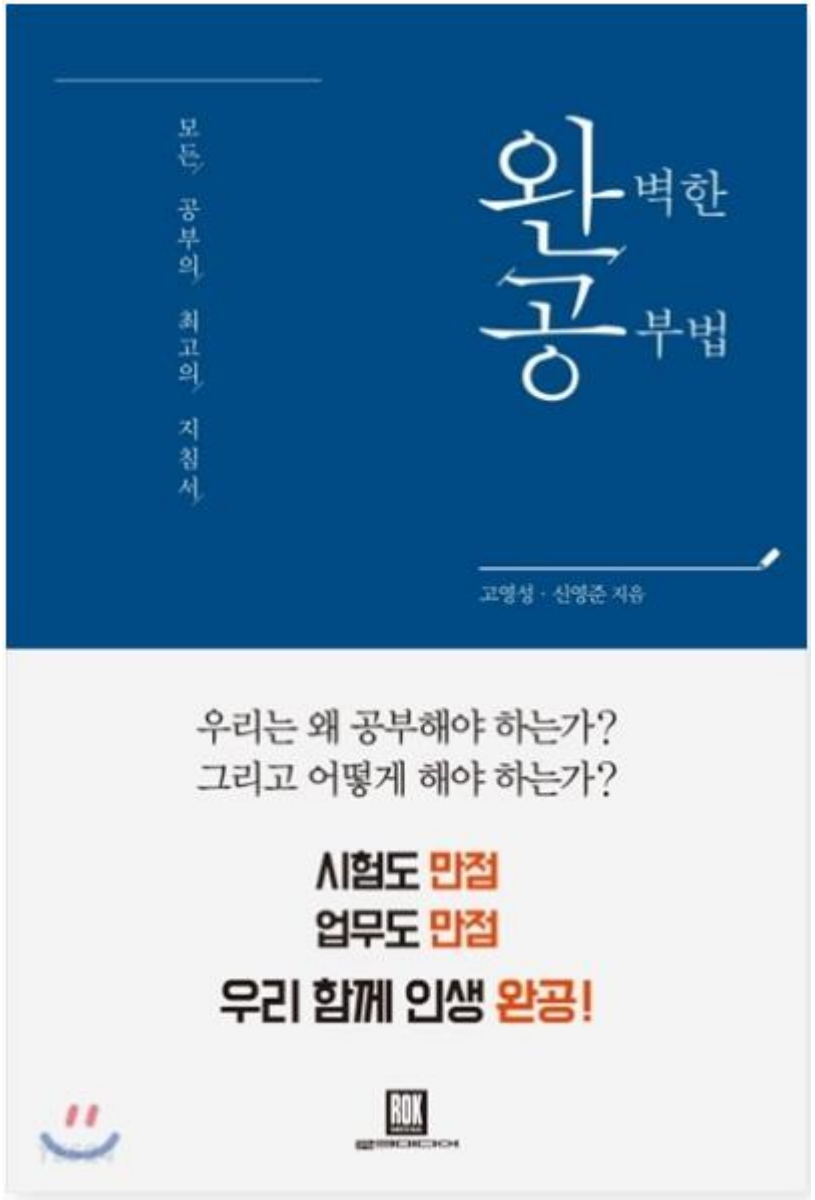


머니볼에서 지나치게 세이버메트릭스를 강조했던 것을 비평



야구선수 분석 및 예측 시스템인 PECOTA를 개발한 세이버메트리션 네이트 실버의 책

선거 예측 웹사이트인 FiveThirtyEight도 개발
https://fivethirtyeight.com/



"강의 듣기와 반복 읽기의 허상"

고영성, 신영준의 <완벽한 공부법>, Chapter 3 기억 중 소챕터 제목

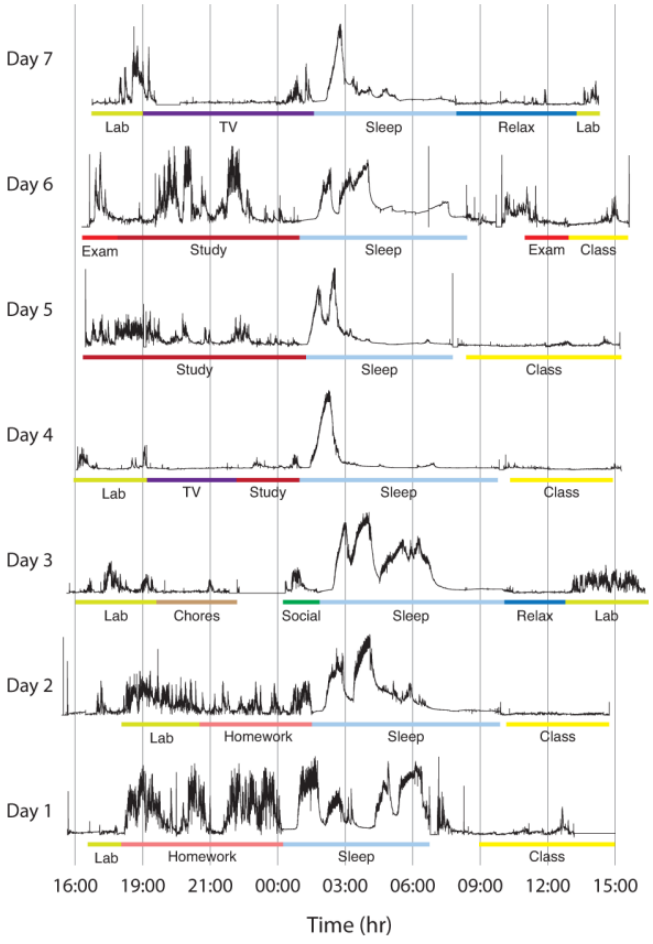


Fig. 9. Long-term *in situ* EDA recordings. Continuous skin conductance measurements were recorded for seven days in a natural home environment. Daily EDA waveforms displayed are normalized.

MIT 미디어랩 연구 결과(2010)

"수업을 들을 때나 TV를 볼 때 뇌의 교감신경계는 그다지 활성화되지 않는다. 심지어 잠자는 시간보다도!"

과제하고 퀴즈 풀고 복습할 때 우리의 뇌는 활동한다는 점!

야구 주요 스탯 공식 정리

1. 타수 = 타석수 – 볼넷 – 몸에맞는공 – 희생번트 – 희생플라이 – 타격방해 - 주루방해

2. 타율 = 안타/타수

3. 장타율 = (1루타 + 2*2루타 + 3*3루타 + 4*홈런)/타수

4. 순수장타율 = (1*2루타 + 2*2루타 + 3*홈런)/타수

5. 출루율 = (안타 + 볼넷 + 몸에맞는공)/(타수 + 볼넷 + 몸에맞는공 + 희생플라이)

6. OPS = 장타율 + 출루율

7. BABIP = (안타 – 홈런)/(타수 – 삼진 – 홈런 – 희생플라이)

8. K%, BB%, HR%
K% (타석당 삼진 비율) = 삼진 / 타석
BB% (타석당 볼넷 비율) = 볼넷 / 타석
HR% (타석당 홈런 비율) = 홈런 / 타석

야구 주요 스탯 공식 정리

9. FIP = (13*홈런 + 3*(볼넷 + 몸에맞는공) – 2*탈삼진)/이닝 + FIP상수

*FIP 상수는 리그마다 매년 다름, https://www.fangraphs.com/guts.aspx?type=cn

10. WHIP = (피안타 + 볼넷)/이닝

11. wOBA(2013 팬그래프버전)
= (0.69*고의사구제외한볼넷 + 0.72*몸에맞는공 + 0.89*1루타 + 1.27*2루타 + 1.62*3루타 + 2.10*홈런)
/(타수 + 고의사구제외한볼넷 + 희생플라이 + 몸에맞는공)

12. wOBA(2021 스탯티즈 버전)
= (0.72*고의사구제외한볼넷 + 0.75*몸에맞는공 + 0.9*1루타 + 0.92*실책출루 + 1.24*2루타 + 1.56*3루타 + 1.95*홈런)
/(타석수 –고의4구)

13. K/9, BB/9, HR/9
K/9 (9이닝당 삼진) = 삼진 / 이닝 * 9
BB/9 (9이닝당 볼넷) = 볼넷 / 이닝 * 9
HR/9 (9이닝당 피홈런) = 홈런 / 이닝 * 9

14. K/BB (볼삼비) = 삼진 / 볼넷