경쟁률 분석

- 경쟁률에 영향을 미치는 요인 분석

- 목차

＊변수 설명 및 분석대상 설정

0. 사전 작업

1. EDA  
 └ 기초 통계량 등

2. 가설 설정 및 검정

└ 직전 2시점 경쟁률 차이는 최종 경쟁률에 영향을 미친다.

└ 작년 경쟁률이 올해 경쟁률에 영향을 미친다.

└ 모집시기(군)가 변하면 경쟁률에 영향을 미친다.

└ 최초 2시점 경쟁률 차이가 최종 경쟁률에 영향을 미친다.

3. PCA

4. 범위 지정 후 재분석

5. 기타 변수 분석

└ 인문계열 중 수학의 반영 비율이 높은 학과들과 기타 학과들의 경쟁률이 차이가 있는가?

└ 자연계열 중 탐구의 반영 비율이 높은 학과들과 기타 학과들의 경쟁률이 차이가 있는가?

6. 결론

＊ 변수 설명

※ Dataset : 2023\_uni (2023학년도 1.정보 DB 중 일부)

•school\_cd (Num, 5) : 학교코드. 학과코드 앞 5자리 / •ue\_cd (Num, 9) : 학과코드

•ue\_cd\_2022 (Num, 9) : 2022학년도 학과코드

•area (Char, 2) : 학교 소속 지역 / •est\_type(Char, 2) : 대학 설립 형태(국립, 공립, 사립)

•uni\_name(Char) : 대학명 / •aply\_type(Char) : 모집전형유형 / •gun(Char, 1) : 모집시기 (군)

•line (Char, 3) : 계열 (인문, 자연, 공통, 예체능) / •line\_detail (Char) : 모집단위 상세계열

•class (Char) : 모집단위명(학과(부)명)

•where (Char) : 학과의 소재지가 다를 경우 해당 학과 소재지

•num (Num, 3) : 모집인원(이월 반영 O) / pre\_num (Num, 3) : 수시 이월 반영 전 모집인원

•aply\_per (float, 3) : 이투스 예상 누적백분위

•pre\_num\_2022(2021) (Num, 3) : 2022(2021)학년도 수시 이월 반영 전 모집인원

•num\_2022(2021) (Num, 3) : 2022(2021)학년도 모집인원(이월 반영 O)

•chuhap\_2022(2021) : 2022(2021)학년도 추가합격 예비번호

•comp\_d3\_2022 ~ comp\_d0\_pm\_2022(2021) : 2022(2021)학년도 사흘 전 ~ 당일 오후 경쟁률

•comp\_2022(2021) : 2022(2021)학년도 최종 경쟁률

•gun\_2022(2021) : 2022(2021)학년도 모집시기 (군)

＊분석 대상

- 경쟁률 추이는 전체 학교를 대상으로 할 경우, 유의미한 정보를 찾을 수 없다고 판단. (중하위권 대학의 경우 경쟁률이 낮아 오히려 분석에 악영향을 끼칠 수 있음)

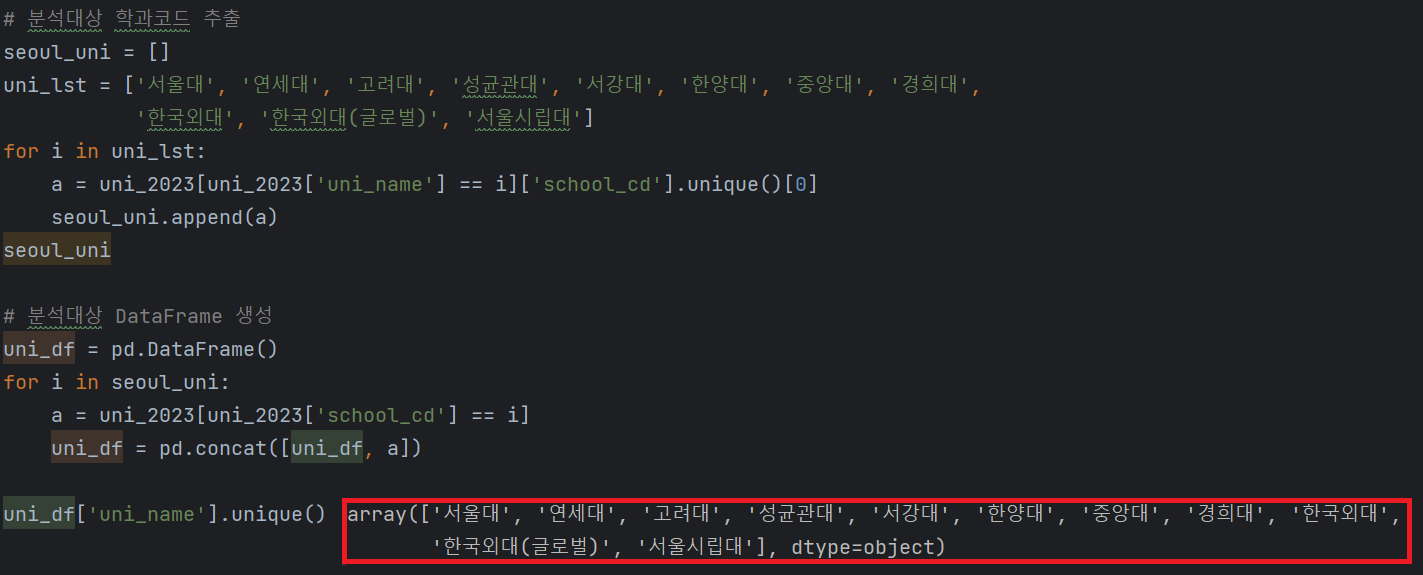
- 특수 학과의 경우, 일반적인 경쟁률 형태와 상이 (의·치·한·약·수의·교대 등)

- 여대의 경우 역시 일반적인 경쟁률 형태와 상이할 것으로 판단. (이화여대, 숙명여대 등)

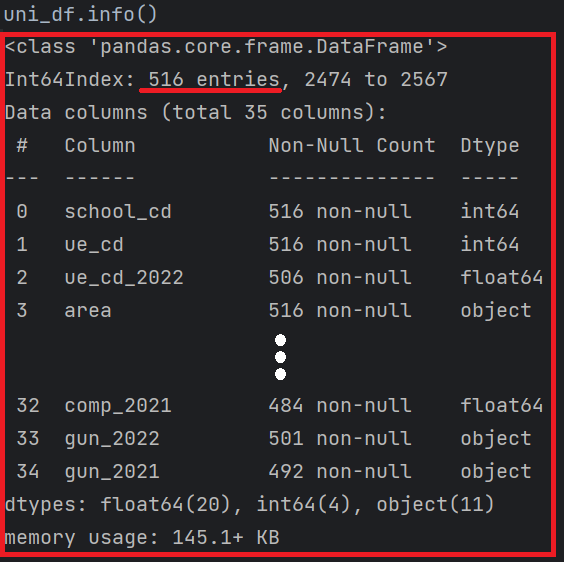
- 따라서, 서울권 상위 대학을 분석대상으로 두고 분석. (서울대, 연세대, 고려대, 성균관대, 서강대, 한양대, 중앙대, 경희대, 한국외대, 서울시립대)

- 분석 이후 결과가 유의미하다면 대학의 범위를 넓혀 확인.

0. 분석 대상 DataFrame 생성



- 분석대상에 해당하는 학교의 전체 학과를 uni\_df로 저장



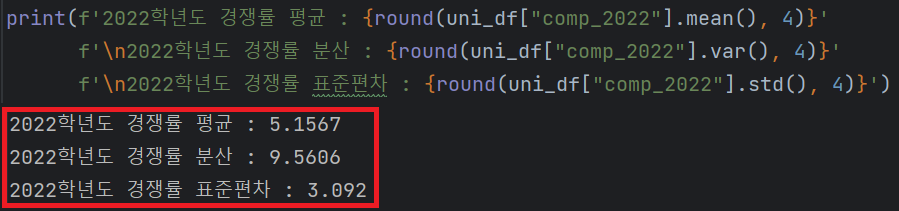
- 총 516개의 학과 존재.

- 신설 학과 10개(ue\_cd 기준, ue\_cd columns count - ue\_cd\_2022 columns count)

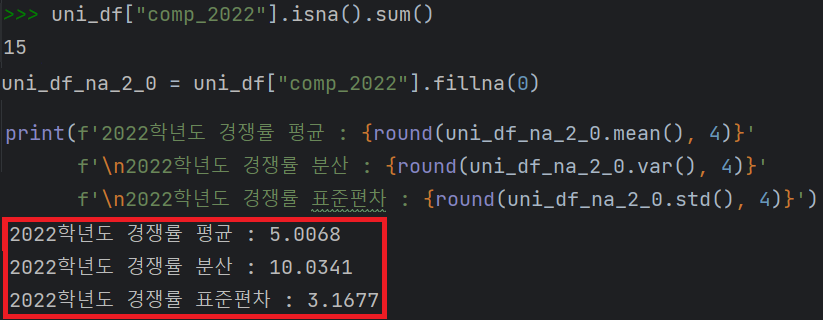
1. EDA (Exploratory Data Analysis)

1) 2022학년도 경쟁률

- DB 기준 2022학년도의 경쟁률의 평균, 분산, 표준편차



- 평균 : 5.1567 / 분산 : 9.5606 / 표준편차 : 3.092



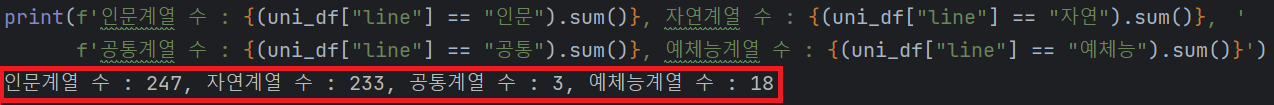
- 2022학년도 경쟁률의 값이 입력되지 않은 경우(NaN)는 총 15건.

- 경쟁률 값이 없는 경우 0으로 대체하였을 때의 평균 : 5.0068 / 분산 : 10.0341 / 표준편차 : 3.1677

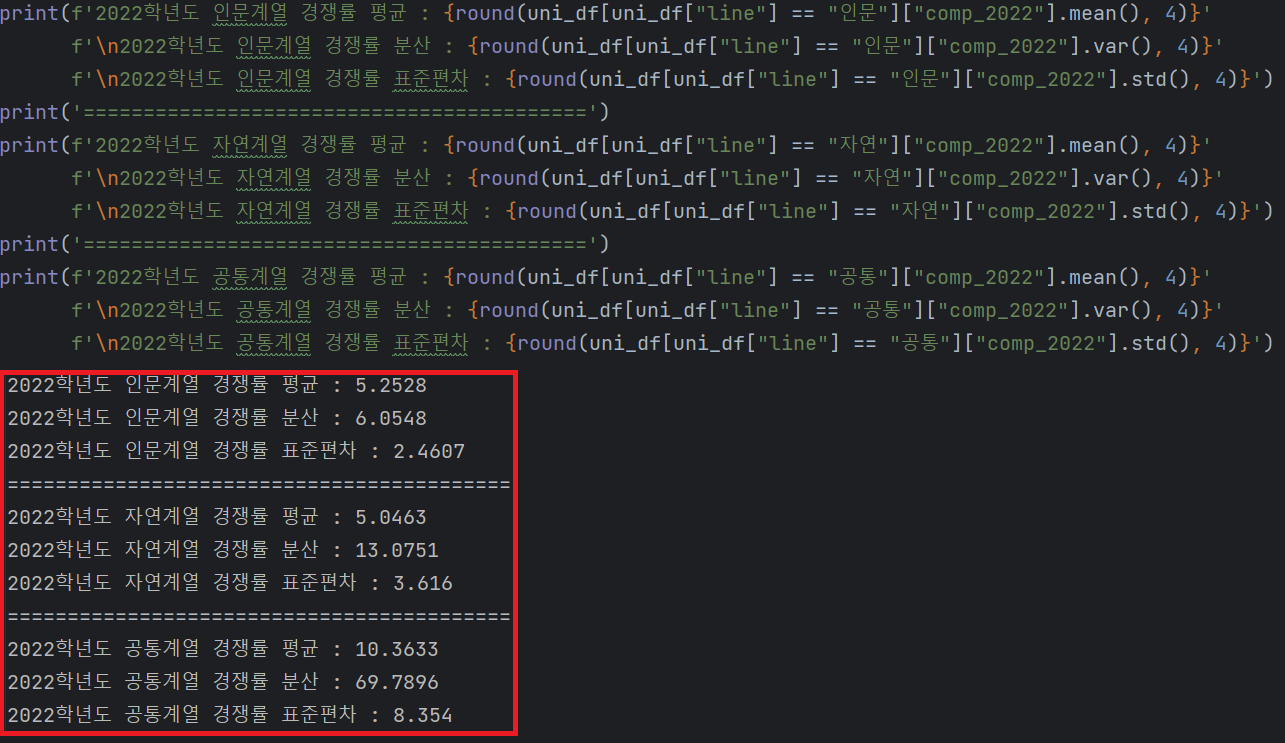
- 결측치를 모두 제거한 경우의 표준편차가 더 작으므로 결측치를 제거한 값을 분석대상으로 선택



1-1) 2022학년도 계열별 경쟁률



- 인문계열 : 247개 / 자연계열 : 233개 / 공통계열 : 3개 / 예체능계열 : 18개 학과로 구성됨.



- 인문계열 평균 : 5.2528 / 분산 : 6.0548 / 표준편차 : 2.4607 (247개 학과)

- 자연계열 평균 : 5.0463 / 분산 : 13.0751 / 표준편차 : 3.616 (233개 학과)

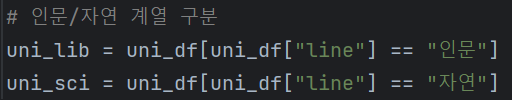
- 공통계열 평균 : 10.3633 / 분산 : 69.7896 / 표준편차 : 8.354 (3개 학과)

- **공통계열**은 표본 수가 상당히 적으므로 (3개 학과) **분석대상에서 제외.**

- **예체능계열**의 경우, 평균 : 4.3989 / 분산 : 1.236 / 표준편차 : 1.1118 (18개 학과)가 나왔으나, 지원자들의 일반 지원 형태와는 상이하므로 **분석대상에서 제외.**

1-2) 계열별 경쟁률의 평균 차이 유무 검정

- T-test로 계열별 평균의 차이가 있는지 검정



-

•계열별 분포 확인



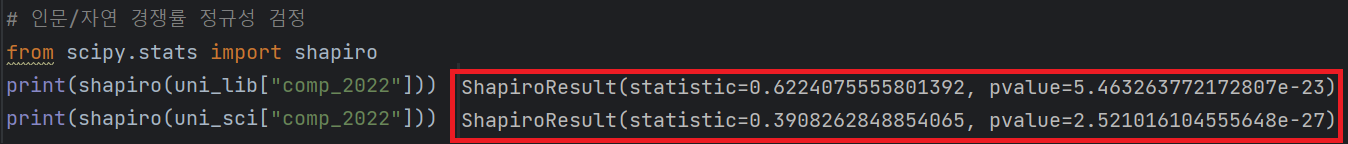
- 인문계열, 자연계열 경쟁률 분포 확인

- 대부분의 데이터가 2~6 사이에 분포.

•정규성 검정

- Shapiro-Wilk Test : : 데이터가 정규분포를 따른다.

- Shapiro-Wilk Test 결과 **P-value가 유의수준(0.05)보다 작으면 데이터는 정규성을 가지지 않음**.



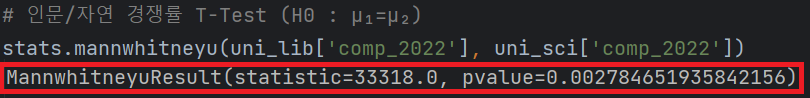
- 검정 결과, **인문계열과 자연계열 경쟁률 모두 정규성을 가지지 않음**.

•등분산성 검정



- 검정 결과, **계열 경쟁률은 등분산성을 지닌다**고 볼 수 있음(정규성을 만족하지 않는 경우, Levene와 Fligner로 등분산성을 검정).

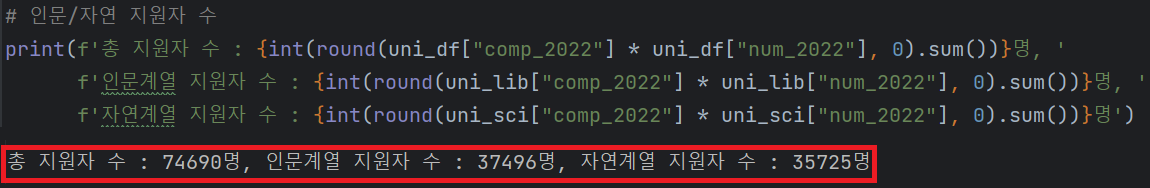
•계열 간 경쟁률에 대한 차이 검정



- 데이터가 비정규성이므로 Mann-whitney 검정 사용.

- 검정 결과, **계열 간 경쟁률의 차이는 존재한다**고 볼 수 있음.

※ 지원자 수 총합

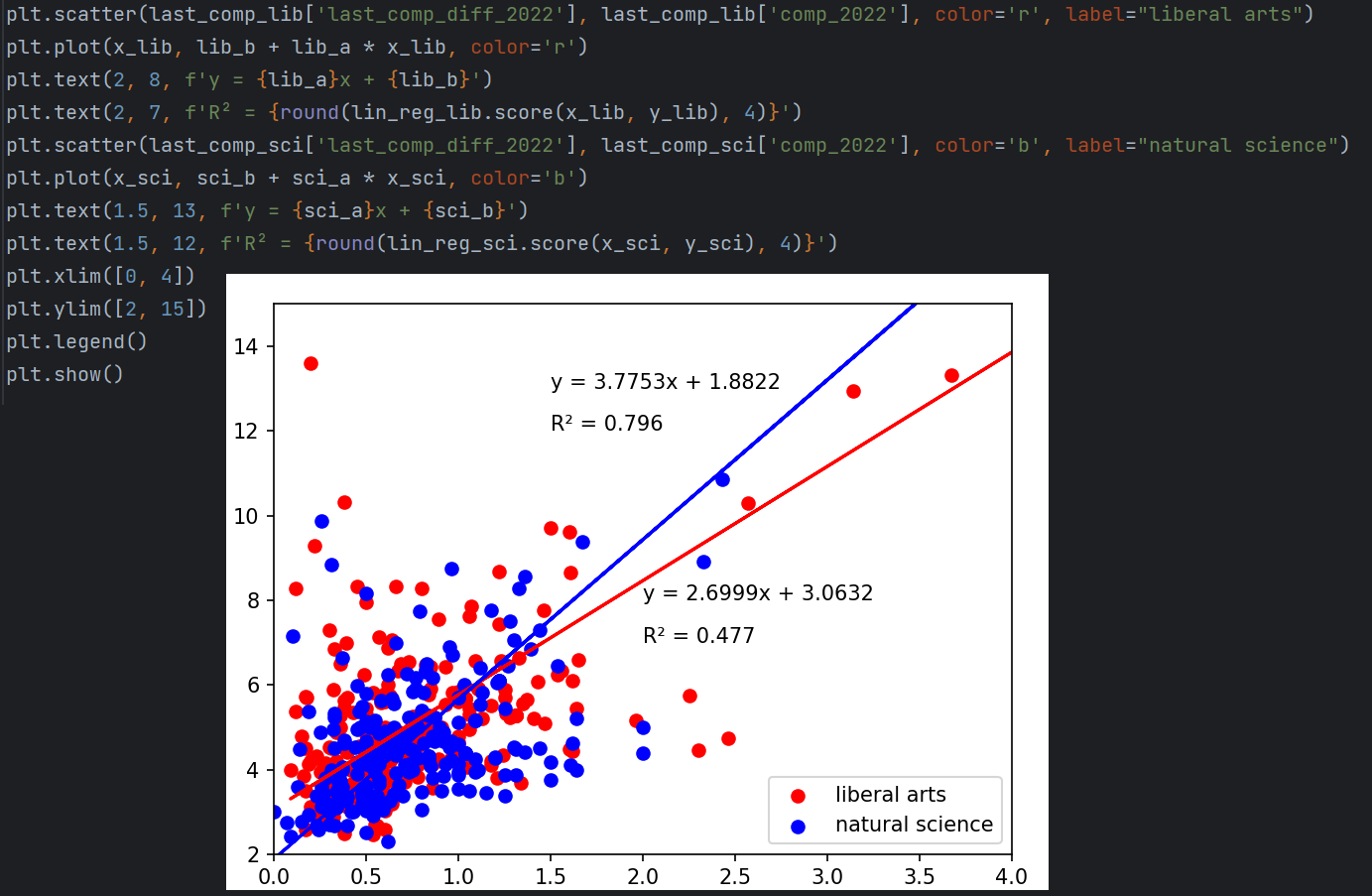


2. 가설 설정 및 검정

1) 직전 2시점 경쟁률 차이는 최종 경쟁률에 영향을 미친다.

- 이전 EDA에서 계열별 경쟁률은 유의하게 차이가 있다는 결론이 나왔으므로 계열을 나누어서 분석.

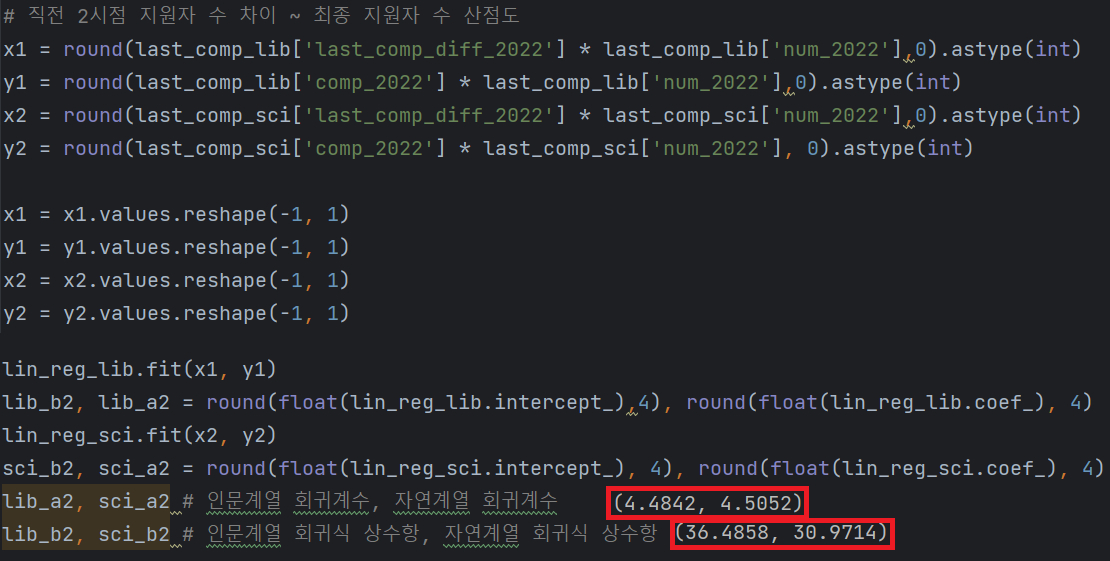




- 경쟁률을 기반으로 1차 선형회귀식을 만든 결과,  
**인문계열은 ‘’**, **자연계열은 ‘’**의 1차 회귀식이 나옴.

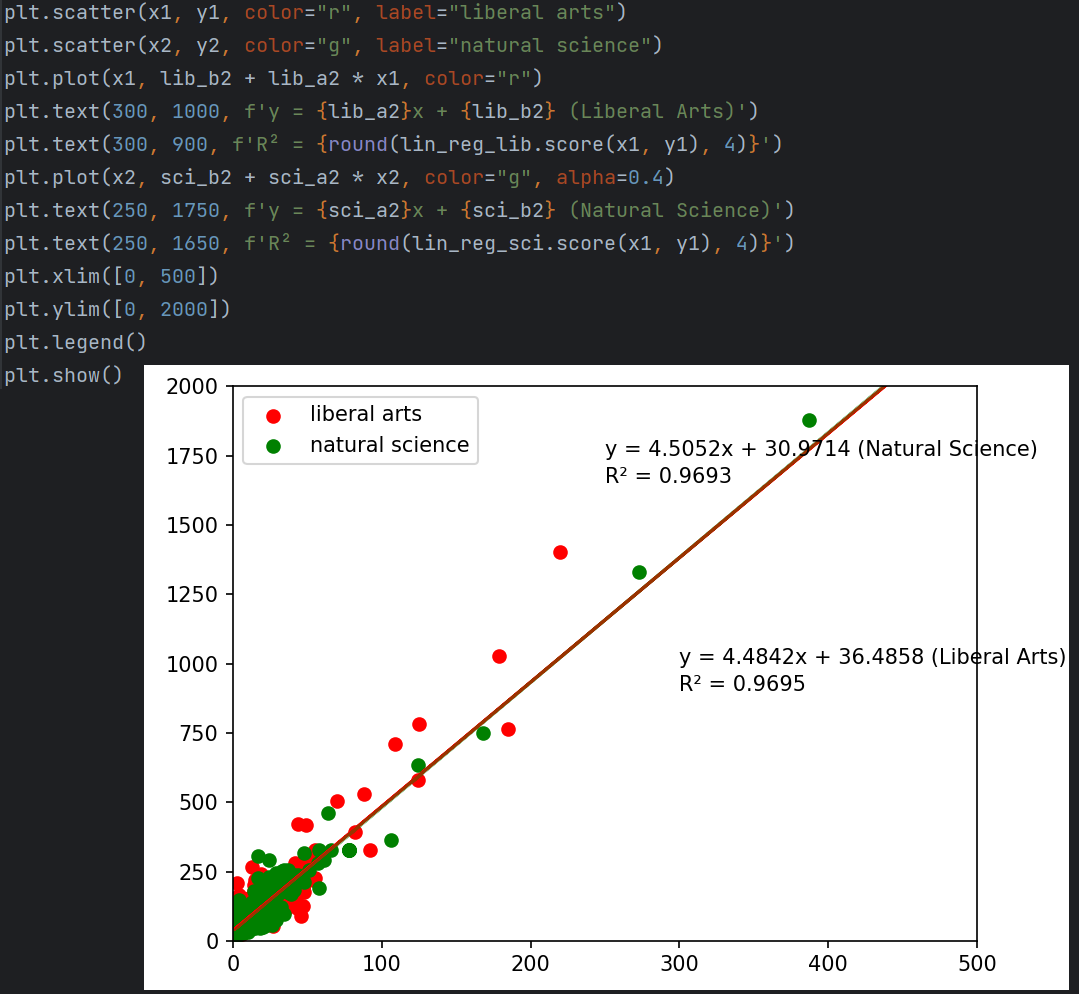
- 자연계열의 경우 회귀선의 결정계수가 80%에 가깝게 나왔기 때문에 회귀선의 정확도가 높은 편이었으나, 인문계열의 회귀선은 48% 가량 밖에 설명하지 못함.

1-1) 경쟁률 대신 지원자 수로 확인



- 경쟁률의 경우, **정확한 예측이 어려움**. (∵ 경쟁률의 변동 폭으로 수치 예상은 가능하나, 최종적으로 몇 명이 해당 학과에 지원했는지를 확인하기 위해서는 지원 인원의 변동 추이를 살펴보는 것이 맞음.)

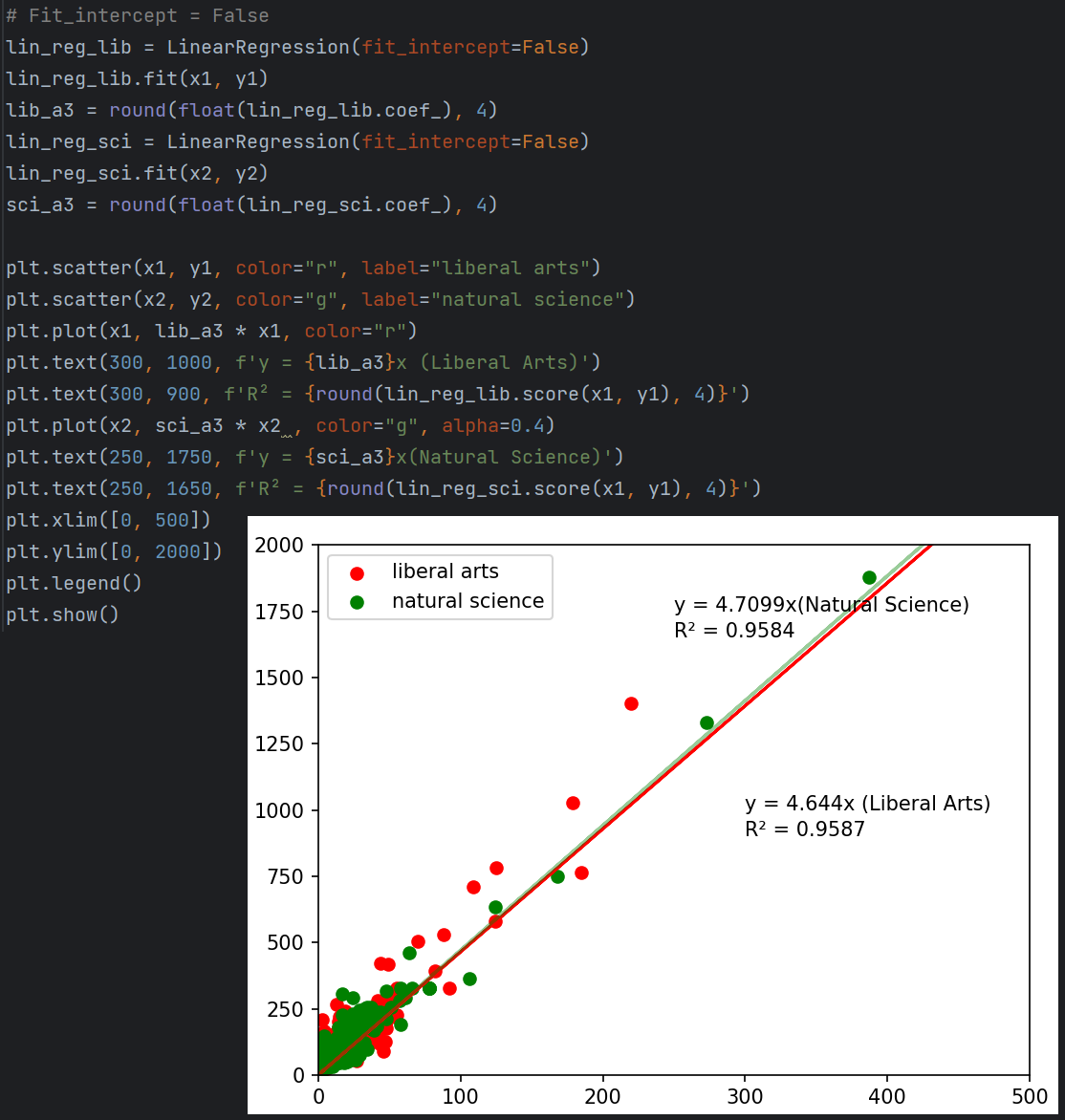
- 지원자 수는 해당 시점의 경쟁률 차이(last\_comp\_diff\_2022)에 모집인원(num\_2022)을 곱하여 계산. (last\_comp\_diff\_2022 × num\_2022)



- 지원자 수를 기반으로 1차 선형회귀식을 만든 결과,  
**인문계열은 ‘’, 자연계열은 ‘’**의 1차 회귀식이 나옴.

- 각 회귀식의 회귀계수 및 상수항이 비슷하므로 회귀선이 거의 일치.

※ 상수항이 없는 1차 선형회귀식 (fit\_intercept = False)

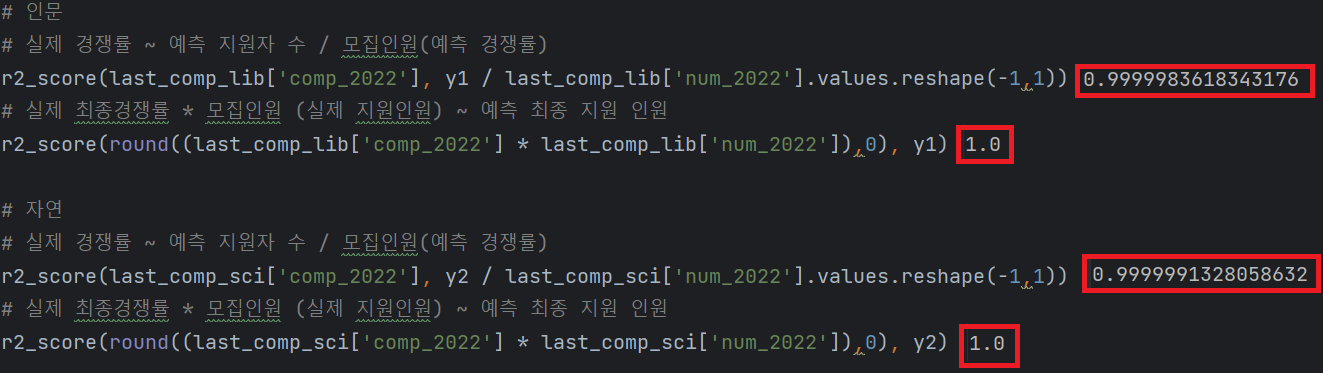


- Intercept를 제외한 회귀식은   
인문계열 : ‘’, 자연계열 : ‘’로 나옴.

- 두 결과 모두 결정계수의 값이 상당히 높으므로 **직전 2시점의 지원자 수의 차이는 최종 지원자 수에 영향을 미친다**고 볼 수 있음.

- 직전 시점에서의 지원자 차이는 다른 요인에 의한 변동이 적어 유의미한 변수로 활용이 가능하다고 판단되나, **전년도 기준 현재 년도에서의 경쟁률(지원자 수)을 예측하는 데에는 의미가 없음.**

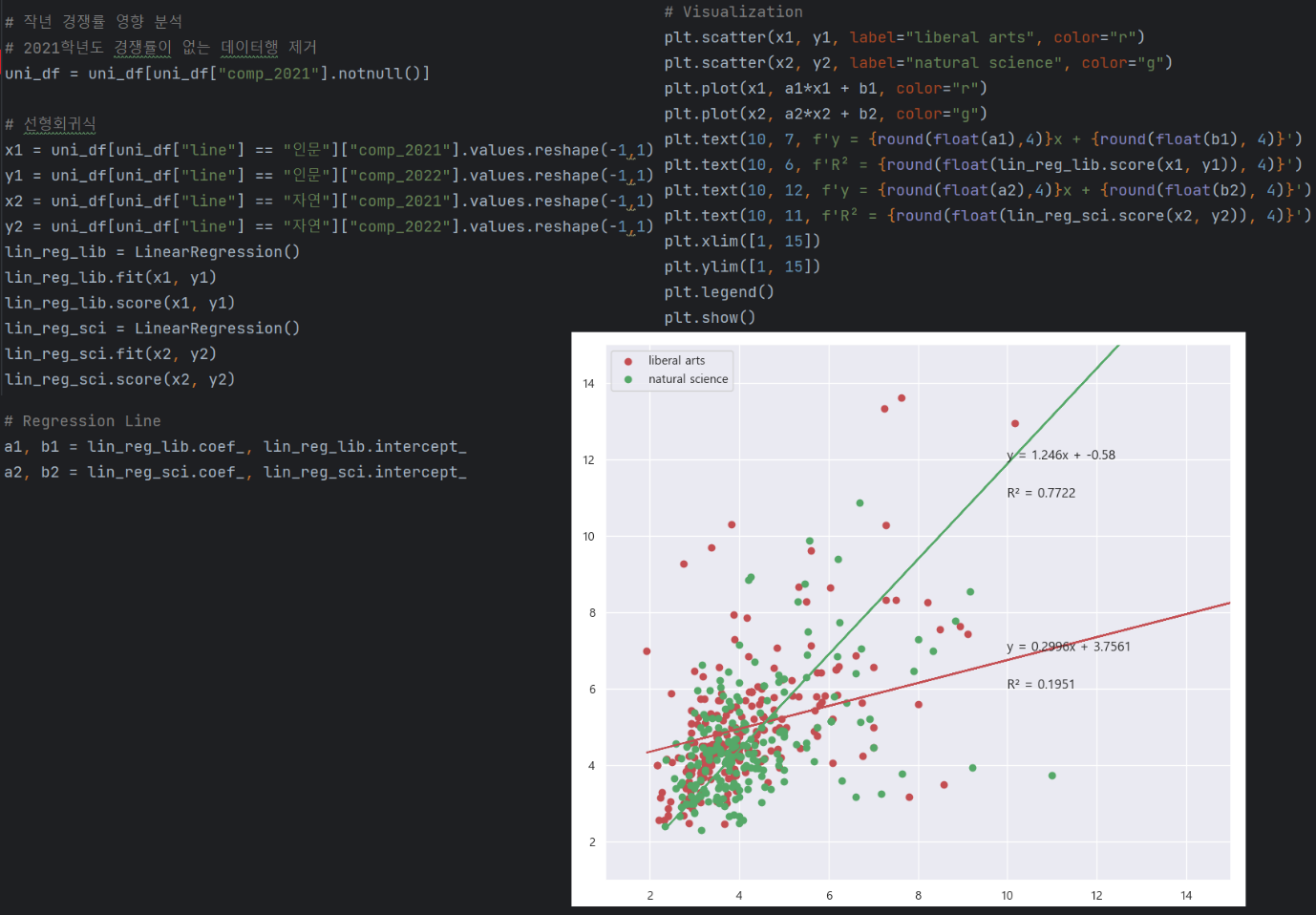
- **최종 경쟁률을 발표하기 직전, 예측하는 용도로 활용이 가능**할 것으로 판단.



- (상수항 포함)회귀식으로 예측한 최종 경쟁률(최종 지원 인원)에 대한 R2-Score가 1에 가까우므로, **예측에 대한 정확성이 아주 높다**고 할 수 있음(회귀선의 정확도가 높음).

2) 작년 경쟁률이 올해 경쟁률에 영향을 미친다.

- 2021학년도 최종 경쟁률이 2022학년도 최종 경쟁률에 영향을 미친다고 가정하고 둘의 상관관계를 분석.



- 경쟁률을 기반으로 1차 선형회귀식을 만든 결과,  
**인문계열은 ‘’, 자연계열은 ‘’**의 1차 회귀식이 나옴.

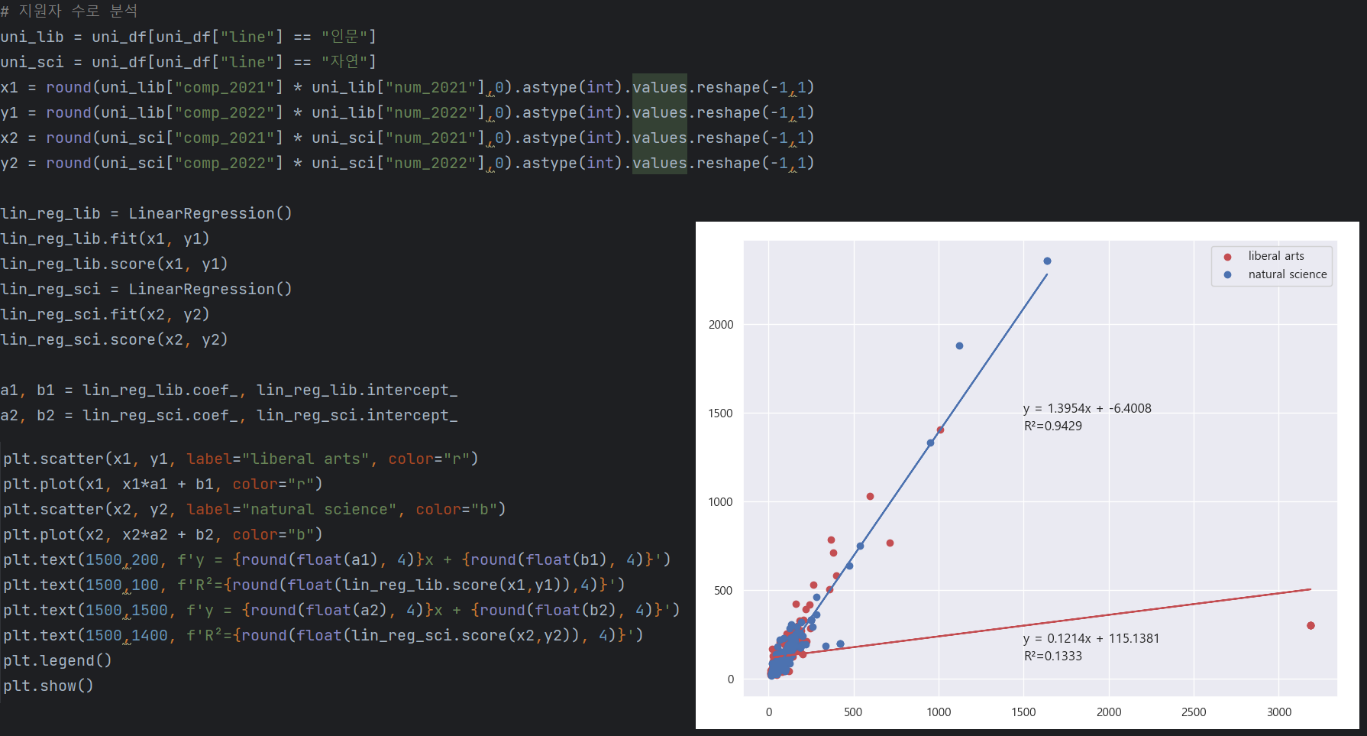
※ 상수항이 없는 1차 선형회귀식 (fit\_intercept = False)



- Intercept를 제외한 회귀식은  
**인문계열 : ‘’, 자연계열 : ‘’**로 나옴.

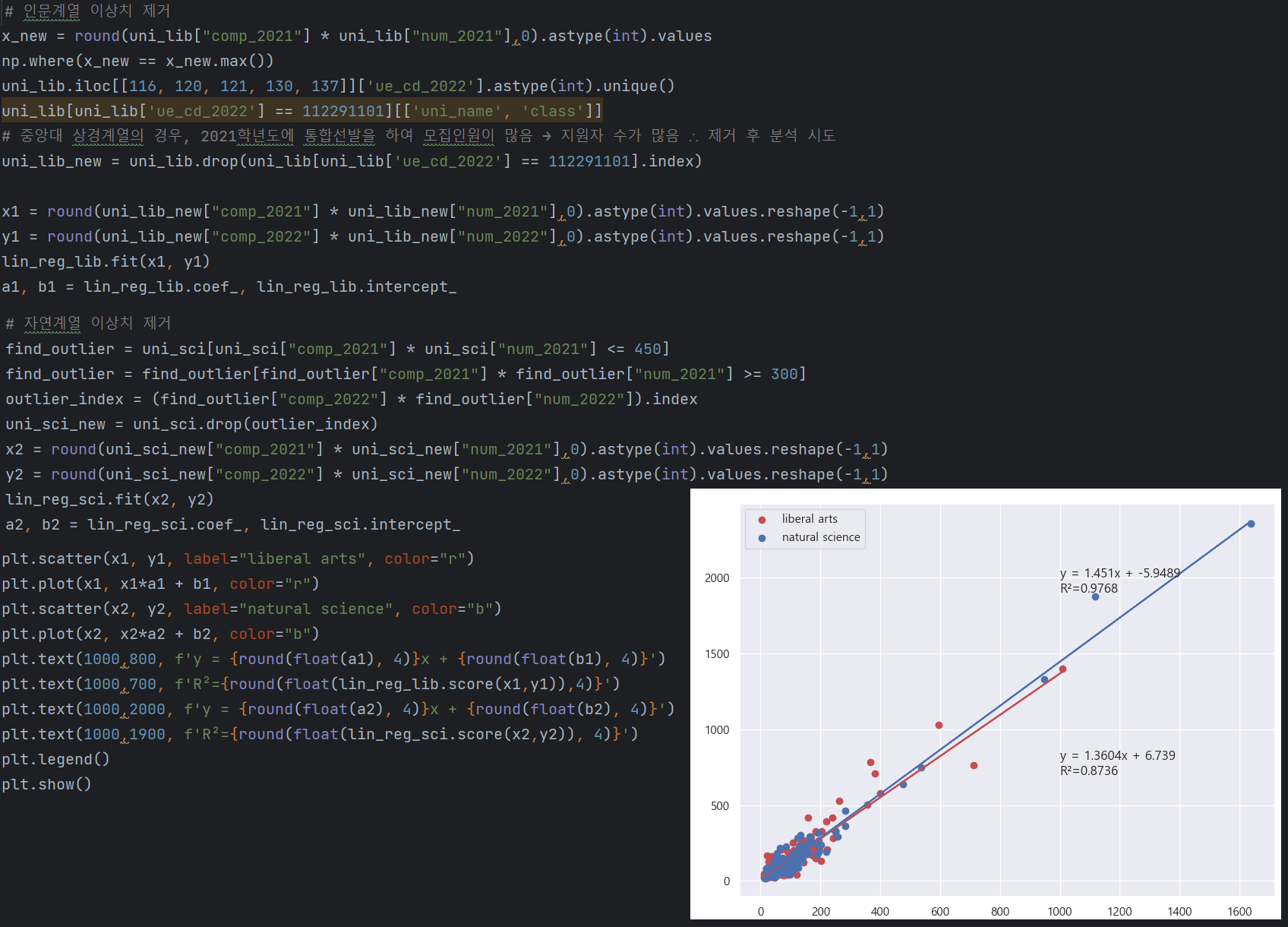
- 두 결과 모두 인문계열의 회귀식 결정계수가 너무 낮으므로 위와 동일 방식으로 경쟁률이 아닌, 지원자 수를 통한 분석을 시도.

2-1) 경쟁률이 아닌 지원자 수로 확인



- 산점도 및 회귀선을 확인한 결과, 인문계열의 이상치(outlier)로 인해 회귀계수가 큰 영향을 미치고 결정계수도 낮게 나오는 것을 볼 수 있음.

- 위 이상치(x1=3188)와 자연계열의 직선에서 살짝 벗어난 값(300<x1<450)을 제거하고 분석 시도.



- 인문계열의 결정계수가 0.1333에서 0.8736으로 상승함을 확인. (이상치는 중앙대 상경계열로, 통합선발을 하면서 지원자 수의 차이가 급격히 난 것으로 확인됨.)

- 자연계열의 결정계수가 0.9429에서 0.9768로 상승함을 확인. (각 학과는 성균관대 소프트웨어학, 중앙대 공과대학(2022학년도 기준)인 것으로 확인됨.)

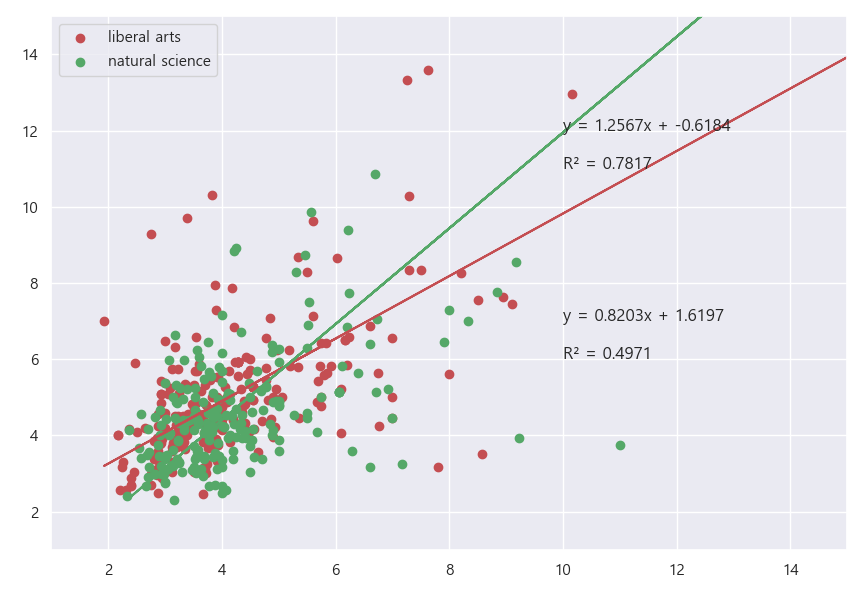
- 확인 결과,  
**인문계열 : ‘’, 자연계열 : ‘’**로 나옴.

- 작년 지원자 수가 적으면 올해 지원자 수도 적고, 작년 지원자 수가 많으면 올해 지원자 수도 많다고 볼 수 있으나, 이는 당연한 사실 중 하나에 불과함(∵ **요강이 크게 변동하는 것이 아니라면 모집 인원의 수는 예년과 비슷하므로 당연히 선형관계에 있을 수밖에 없음**).

- 선형성이 없는 경우 예외케이스를 확인하고자 하였으나, **‘중앙대 상경계열’의 통합선발-학과별 선발 변동 사항 변동 외의 특이점은 발견할 수 없었음.**

- 따라서, **모집단위의 형태가 바뀌지 않는 이상 예년도 지원자 수와 유사할 것으로 판단됨.**

※ 이상치 제거 후 경쟁률로 재분석



- 변화는 있었으나 여전히 결정계수가 낮아 경쟁률만으로는 예측이 힘들다고 볼 수 있음.

3) 모집시기(군)가 변하면 경쟁률에 영향을 미친다.

- 모집시기가 변동된 경우의 경쟁률 차이(μ₁)와 모집시기가 동일한 경우의 경쟁률 차이(μ₂)가 있으면 모집시기의 변화가 경쟁률에 영향을 미치는 것으로 판단.

- T-test를 통해 두 집단 간 경쟁률의 차이 유무를 분석.

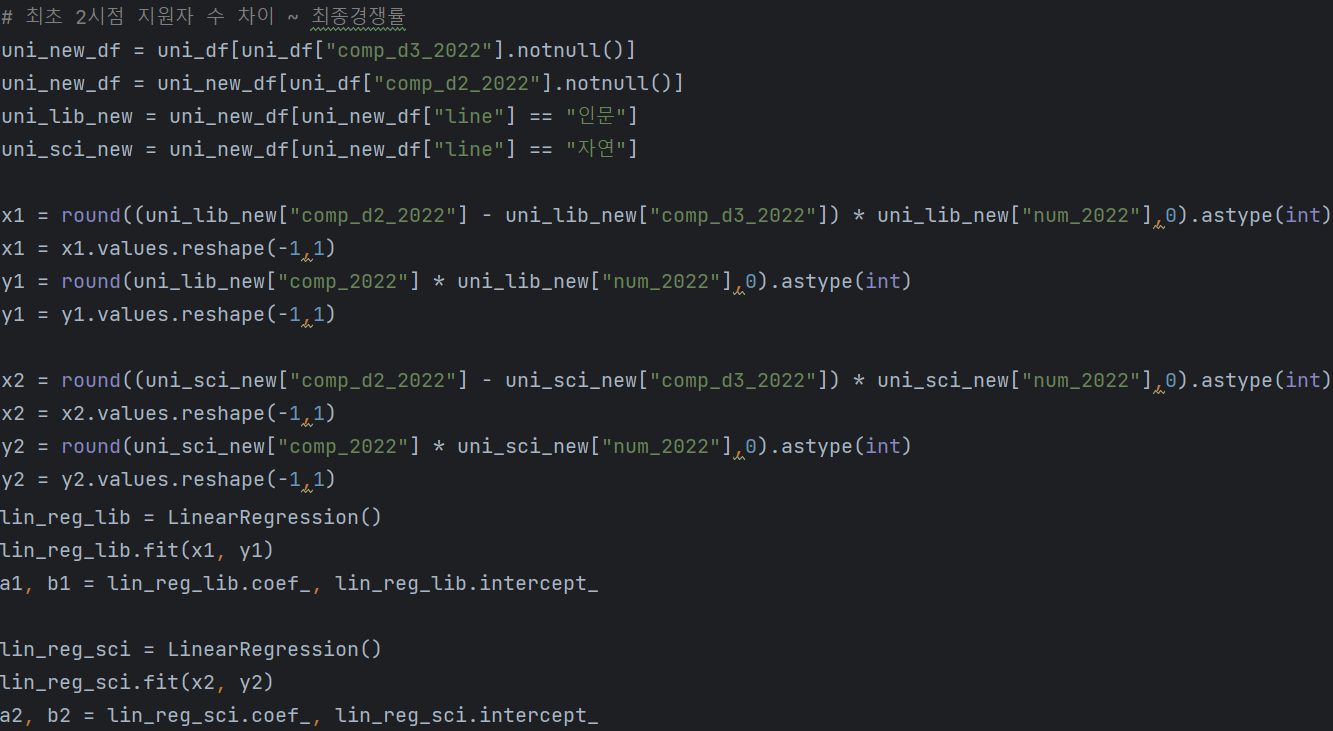


- 계열별 T-test 및 Mannwhitneyu-test 결과, P-value의 값이 크므로 귀무가설(군이 바뀌었을 때 경쟁률의 차이가 없다)을 기각할 수 없음.

- 따라서, **군 변동이 경쟁률에 영향을 미치지 않는다**고 볼 수 있음.  
(지원자 수로 볼 경우 둘의 차이가 난다고 나왔지만 이는 변동이 되지 않은 모집인원과 변동이 된 모집인원의 차이 때문.)

4) 최초 2시점 경쟁률 차이가 최종 경쟁률에 영향을 미친다.

- 작전 2시점 지원자 수 차이와 최종 지원자 수 간 선형성을 파악할 수 있었기 때문에 최초 2시점 지원자 수 차이에 대해서도 최종 지원자수와의 선형성이 존재하는지 확인.

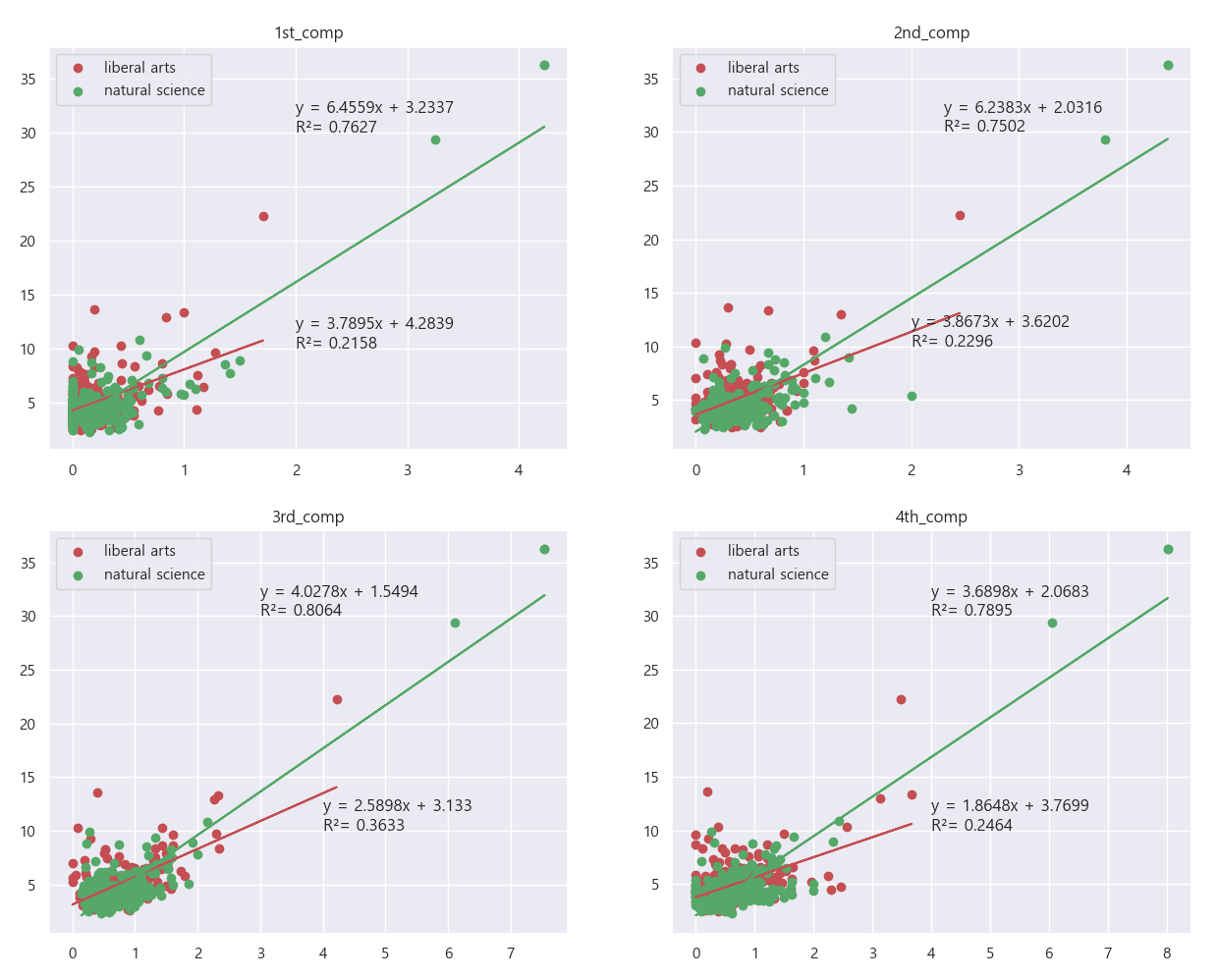




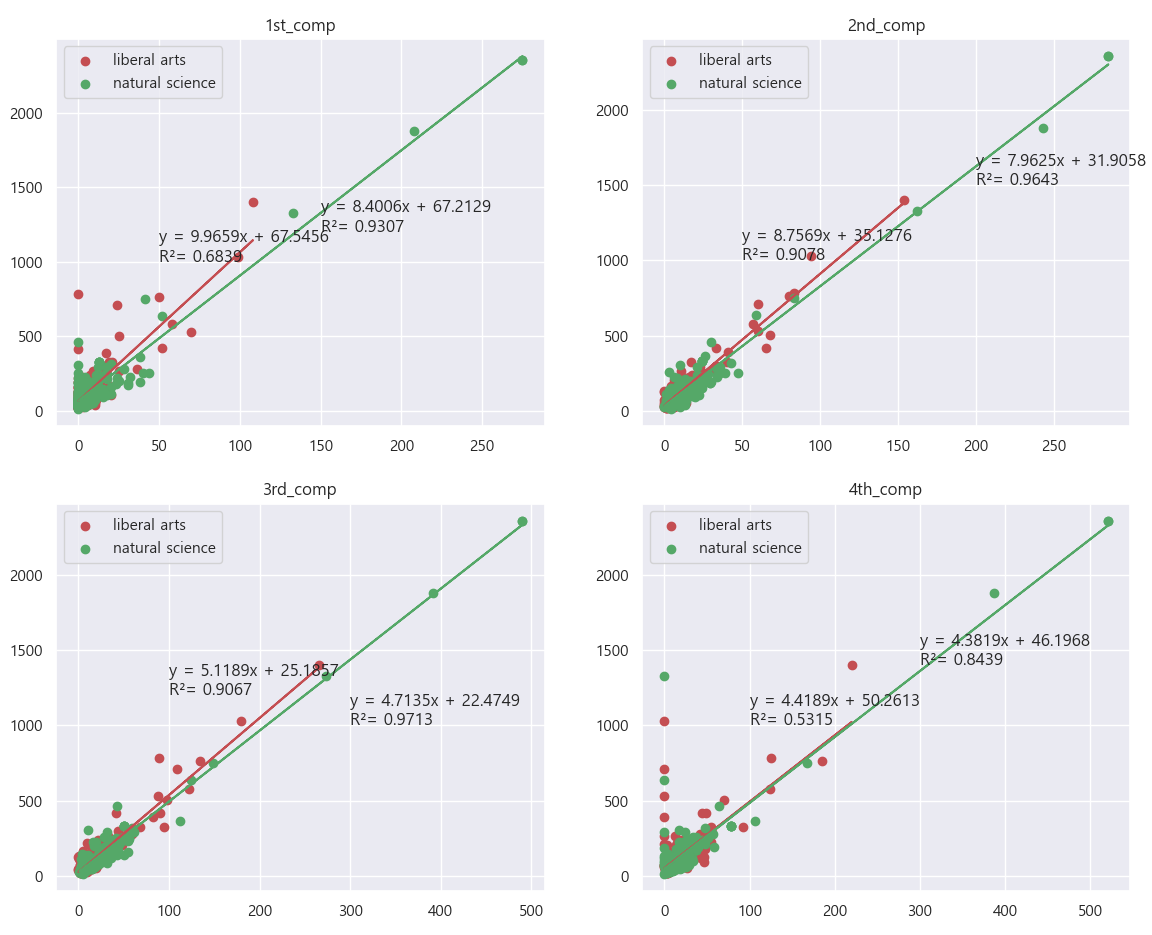
- 확인 결과,  
**인문계열은 ‘’,  
자연계열은 ‘’**로 나왔음.

- 최초 2시점 지원자 수 차이와 최종 지원자도 선형성이 있다는 결론이 나왔으므로 **시점에서의 지원자 수 차이가 최종 지원자 수를 예측하는 데에 큰 영향을 미친다**고 볼 수 있음.

+) 시점 차이별 경쟁률과 최종 경쟁률 산점도, 선형회귀직선 그래프



+) 시점 차이별 지원자 수와 최종 지원자 수 산점도, 선형회귀직선 그래프



- 보통 최종시점에서의 경쟁률 예측이 가장 높지만(위 그래프에서도 시점이 지날수록 정확성이 높아지고 있음), 최종시점에서의 결정계수가 많이 떨어진 원인은 서울대 등 **일부 학교(성균관대)의 경우, 당일 오후 경쟁률을 기록하지 않아 해당 column을 0으로 두었기 때문**.

3. PCA

위의 EDA 과정을 통해 최종 경쟁률에 큰 영향을 미치는 부분은 “작년 최종 지원자 수”와 “당년도 시점 간 지원자 수 차이”로 나옴.

따라서, 위 변수들을 활용하여 주성분 분석을 하여 어떤 변수를 사용하면 더 정확한 결과를 예측할 수 있는지 확인.



- 확인 결과, 당년도 지원자 수는 작년 지원자 수로 **인문계열은 90.91%, 자연계열은 98.22%**까지 예측할 수 있었고, 처음 지원자 수 격차(1st\_comp) 변수까지 포함하면 **인문계열은 99.82%, 자연계열은 99.43%**까지 예측이 가능.

- 당년도 지원자 수를 예측할 때에는 **작년 지원자 수와 처음 지원자 수 격차(이틀 전 지원자 수와 3일 전 지원자 수의 차이)만 보아도 최종 지원자 수를 짐작할 수 있다**는 결론이 나옴.

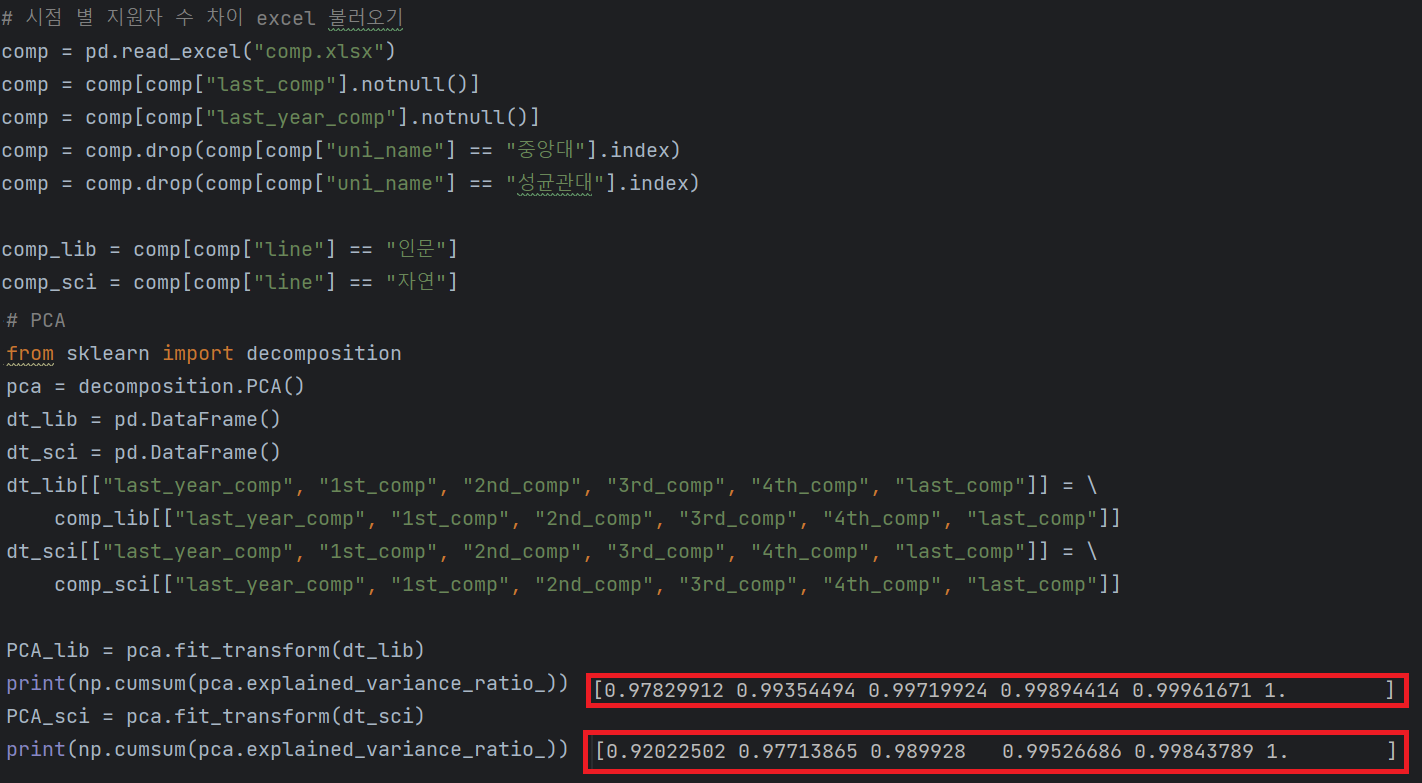
4. 범위 지정 후 재분석

- 지원자 수가 과도하게 많은 Case들이 존재하여(중앙대의 경우, 2022학년도에 통합선발을 했으므로 타 학교들에 비해 지원자 수가 과도하게 많음.) 그러한 경우들을 대상에서 제외 후 재분석.

- 직전 시점 경쟁률에서 당일 오후에 발표하지 않는 학교들이 존재하여(성균관대) 제외 후 재분석.

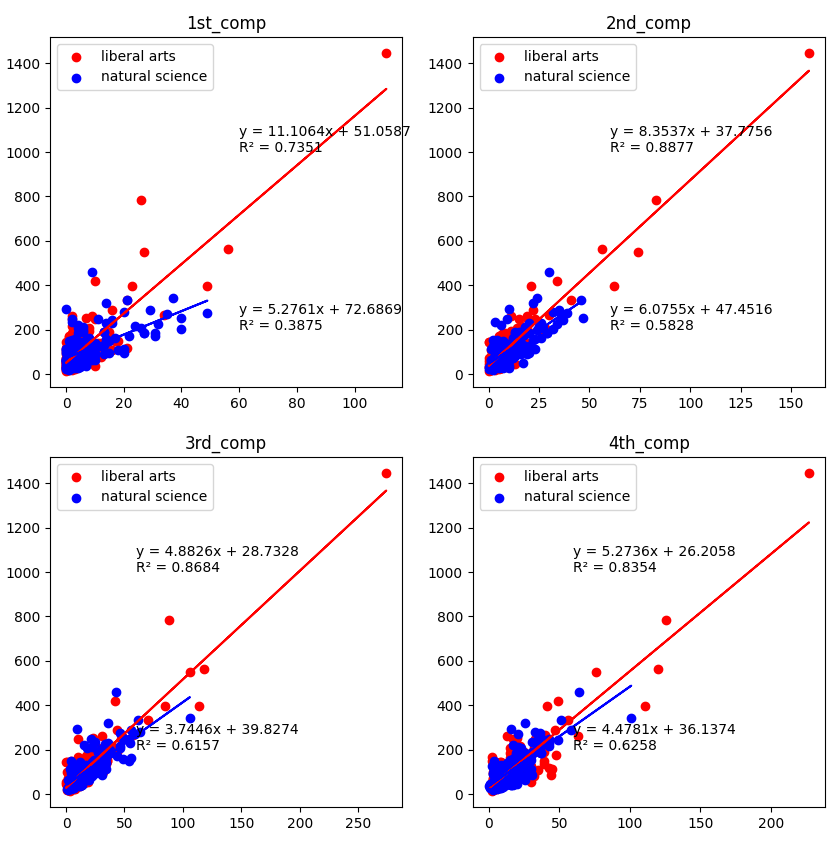
- 결론적으로 **중앙대와 성균관대를 제외하고 재분석** 실시.

1) PCA



- 제외 후 PCA 결과는 **작년 지원자 수와 처음 지원자 수 격차(1st\_comp) 변수까지 포함하면 인문계열은 99.35%, 자연계열은 97.71%까지 예측이 가능하다**는 결론이 나옴.

2) 시점 별 지원자 수 차이 산점도 분석



- 설명력이 오히려 감소하였음. (4th\_comp(직전 시점) - 인문계열 제외)

- 선형성은 여전히 존재하고, 제외한 데이터들은 결국 **지원자 수 차이가 최종 지원자 수에 영향을 끼친다는 결과를 뒷받침한다**고 볼 수 있음.

5. 기타 변수 분석

- 경쟁률에 직접적으로 영향을 미치는 변수들이 아닌, 부가적인 요소들을 통해 경쟁률과 연관이 있는지 분석.

- 1번 DB에 있는 자료 중 영역별 반영 비율 등을 분석에 활용.

1) 인문계열 중 수학의 반영 비율이 높은 학과들과 기타 학과들의 경쟁률이 차이가 있는가?

- 인문계열을 선택한 학생들의 경우 수학 과목의 비중이 높은 학과를 비선호할 것이라고 판단, 수학 영역 비중이 33%보다 큰 학과들의 경쟁률과 33% 이하 학과들의 경쟁률에 차이가 있는지 분석. (국어, 수학, 탐구과목을 보는 경우 1:1:1 비율인 33%보다 초과하면 수학을 중점적으로 보는 학과라고 판단하였기 때문)

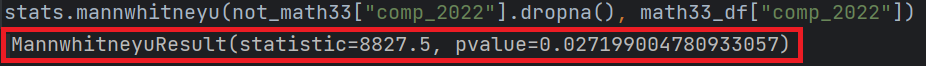




- 간단하게 통계량을 확인해보았을 때, 수학 반영비율이 높은 학과들(33% 초과)은 인문계열 총 247개 학과 중 115개의 학과에 해당되었으며, 평균 경쟁률은 5.0523(표준편차 2.669)이 나옴.

- 전체 인문계열의 평균 경쟁률은 5.2528(표준편차 2.4607), 수학 반영비율 33% 이하 학과들의 평균 경쟁률은 5.4276(표준편차 2.2594)으로 **수학 반영비율이 높은 학과들의 경쟁률이 더 낮게 나온다**고 짐작할 수 있음.

- 정확한 통계를 위해 T-test(비정규성이므로 Mannwhitney) 사용



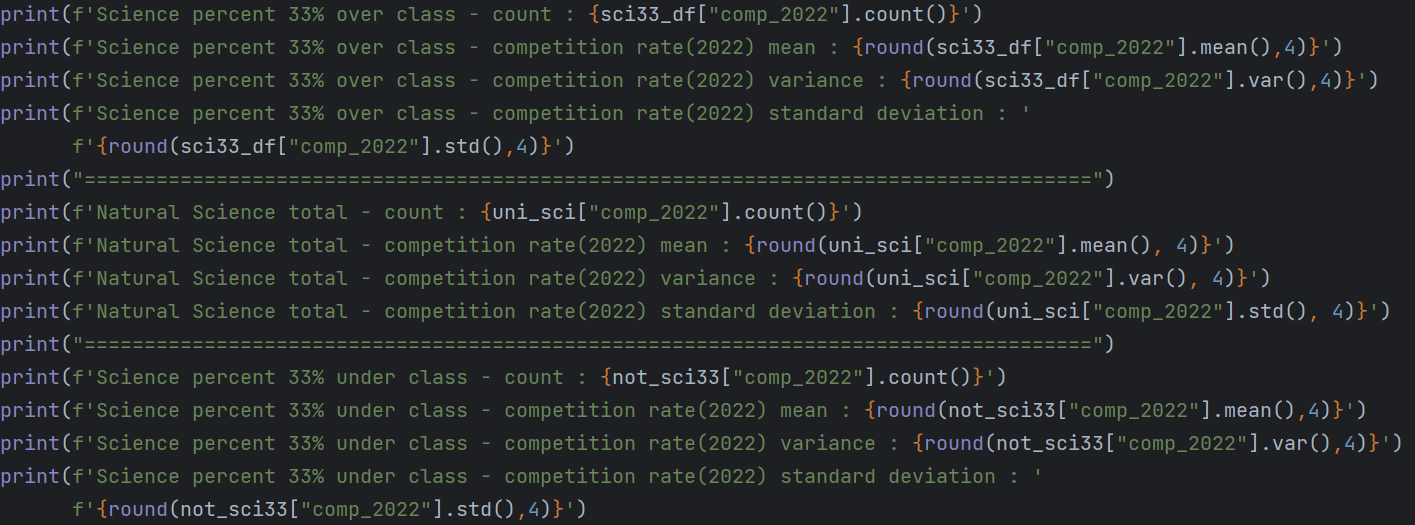
- 확인 결과, **수학 반영비율 33% 이하 학과들의 경쟁률 평균과 수학 반영비율 33% 초과 학과들의 경쟁률 평균은 다르다**고 볼 수 있었음.

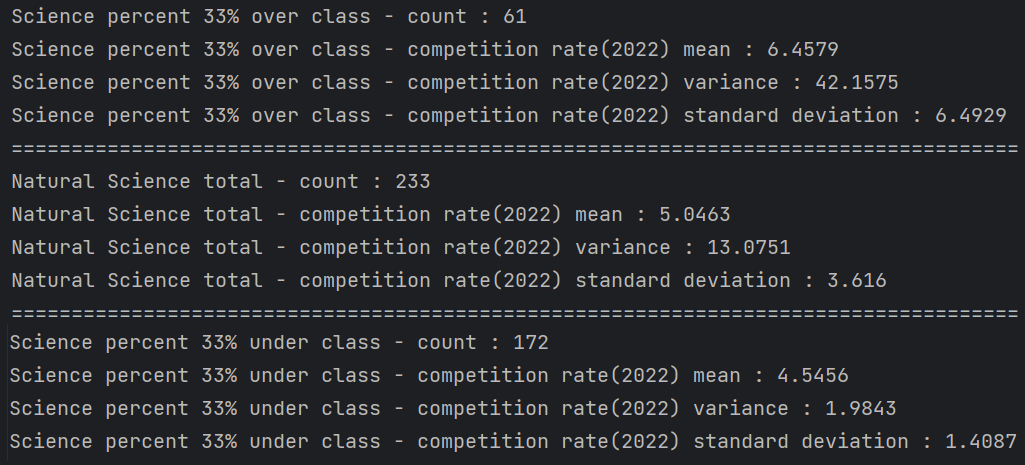
- 앞서 구한 기초 통계량을 기반으로, **수학 반영비율이 높은 학과들은 그렇지 않은 학과들보다 경쟁률이 낮다**고 볼 수 있음.

2) 자연계열 중 탐구의 반영 비율이 높은 학과들과 기타 학과들의 경쟁률이 차이가 있는가?

- 자연계열의 경우 탐구의 비중이 높을 때 경쟁률에 차이가 있을 것이라고 판단하여 탐구 비율이 33%를 초과하는 학과들과 아닌 학과들의 경쟁률을 비교.



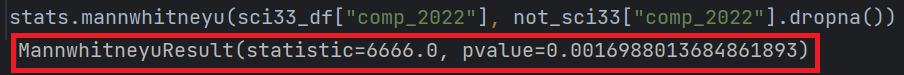




- 간단하게 통계량을 확인해보았을 때, 탐구 반영비율이 높은 학과들(33% 초과)은 자연계열 총 233개 학과 중 61개의 학과에 해당되었으며, 평균 경쟁률은 6.4579(표준편차 6.4929)이 나옴.

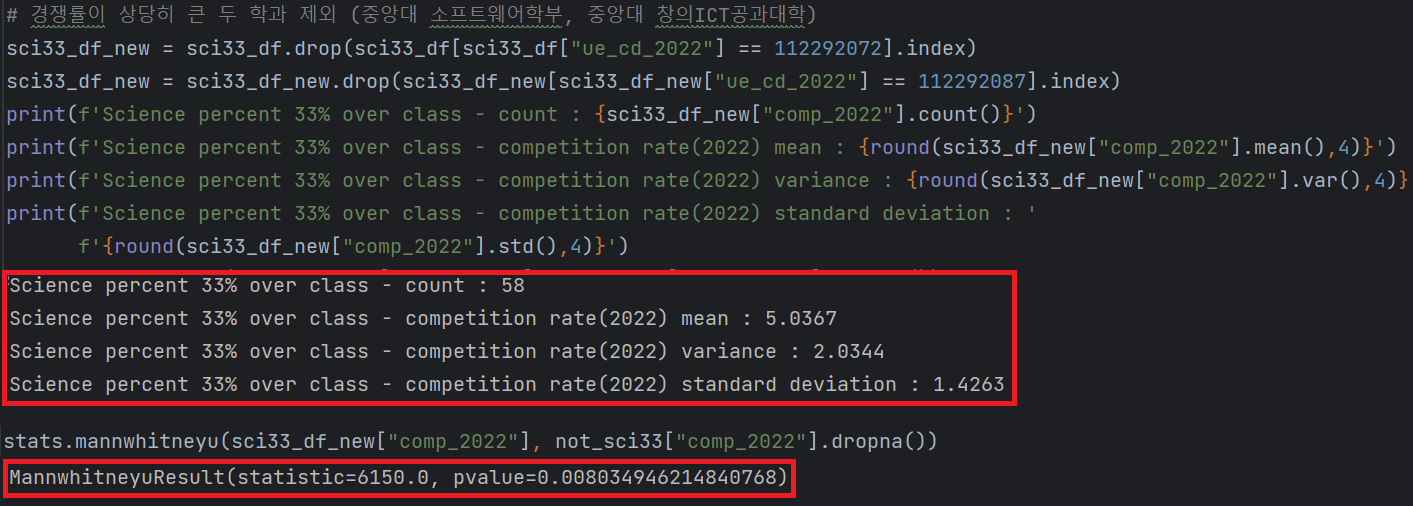
- 전체 인문계열의 평균 경쟁률은 5.0463(표준편차 3.616), 탐구 반영비율 33% 이하 학과들의 평균 경쟁률은 4.5456(표준편차 1.4087)으로 **탐구 반영비율이 높은 학과들의 경쟁률이 더 높게 나온다**고 짐작할 수 있음.

- 위와 마찬가지로 Mannwhitney를 사용하여 두 집단 간 차이가 존재하는지 확인.



- 확인 결과, 탐구 반영비율 33% 초과 학과들의 경쟁률 평균과 이하 학과들의 경쟁률 평균은 다르다고 볼 수 있었음.

- 하지만 분석대상 중 중앙대-소프트웨어학부와 창의ICT공과대학의 경쟁률이 타 학과에 비해 상당히 높은 값을 차지하고 있어 이 두 학과를 제외하고 재분석 실시.



- 경쟁률 이상치로 판단되는 두 값을 제외하고 기초 통계량 및Mannwhitney로 분석한 결과, 경쟁률은 평균 5.0367(표준편차 1.4263)이 나왔고, **탐구 반영비율이 높은 학과들의 경쟁률과 탐구 반영비율이 낮은 학과들의 경쟁률이 차이가 있다**는 결과가 나옴.

- **탐구 반영비율이 높은 학과들의 경쟁률이 여전히 더 높게 나온다**고 볼 수 있음.

6. 결론

- 다각도로 분석한 결과, 최종 경쟁률(최종 지원자 수)에 영향을 미치는 변수는 여러 변수가 존재하지만 특히 **작년 지원자 수 및 당년도 시점 별 지원자 수 차이에 의해 결정이 난다**고 볼 수 있음.

- 경쟁률로 볼 경우, 예측이 어렵기 때문에(모집인원의 변동에 대한 예측이 어려움, 경쟁률은 대개 20 이하의 작은 수로 수치 변동에 크게 영향을 받음) **지원자 수를 이용하여 예측에 활용하는 것이 좋음**.

- **모집단위의 변경(학과 통폐합 등)으로 모집인원이 크게 변동되는 경우, 예측이 어려우므로 이 경우에는 당년도 지원자 수의 차이로 추정**하는 것이 좋음.

- 기타 변수를 활용하여 경쟁률에 유의미한 차이가 있는지 확인한 결과, **인문계열에서는 수학 반영비율이 높은 학과들은 상대적으로 더 낮은 경쟁률**을 보였고, **자연계열에서는 탐구 반영비율이 높은 학과들이 상대적으로 더 높은 경쟁률**을 보였음.

- 경쟁률에 영향을 미치는 요소는 분석한 변수들 외의 다른 부분들도 상당히 많으므로 정확한 분석을 하기 위해서는 보다 다양한 변수가 있는 데이터 셋이 필요.

끝맺음말)

경쟁률에 유의미한 영향을 미치는 변수는 현재 분석한 변수들 외에도 더 많은 변수들이 존재한다. 모집단위명 변동에 따른 변동이 존재할 수 있고, 학교의 위치 변경 등도 주요 사유가 될 수 있다(모집단위명 변경에 따른 경쟁률 변동의 경우 자체적으로 확인하였을 때에는 변경의 정도를 자체적으로 지정하기 어려운 점과 임의로 설정하여 분석하였으나 차이가 있다고 볼 수 없었음). 그리고 경쟁률은 다양한 데이터들이 복합적으로 상호작용하고 전혀 예측하지 못한 제3변수 등으로 인해 쉽게 예측하기가 어렵다. 이러한 연유로 “당년도 마감 전 공개되는 데이터를 토대로 최종 예상 경쟁률을 예측해보자”고 생각하였고, 이를 경쟁률이 아닌 **지원자 수로 변경하여 확인했을 때 유의미한 결론**을 도출할 수 있었다.

하지만 이러한 예측은 **이상치에 큰 영향**을 받는다. 일례로 위의 분석 중 중앙대의 경우 2022학년도에는 통합선발을 하면서 경쟁률과 지원자 수가 평균치를 크게 벗어난 값이 나왔는데 이를 어떻게 활용하느냐에 따라 결과가 크게 바뀐다. 따라서, 올해의 최종 경쟁률(지원자 수)을 예측하고 싶다면 당년도의 입시 요강을 확인하여 모집사항을 꾸준히 파악하여야 한다.

또한 이 레포트는 서울 중상위권 이상 대학으로만 분석한 것으로, 일반화를 위해서는 더 많은 표본이 필요하다. **즉시 상용화에는 어려움**이 있으나, 현 레포트에서의 결론들을 토대로 어떠한 변수를 조금 더 중점적으로 볼 필요성이 있는지를 파악하는 자료로 사용하면 될 것이다.

- Python code –

<https://github.com/zaegeon/Workspace/blob/main/Competition/comp_rate_1>

<https://github.com/zaegeon/Workspace/blob/main/Competition/comp_rate_2>

<https://github.com/zaegeon/Workspace/blob/main/Competition/comp_rate_3>