

심리 성향과 성격 유형에 따른 투표 참여 예측



“ AM 11:40

이동재(팀장)

노용철

정성훈

홍세준

”

Contents

01 개요

02 데이터 변수 설명

03 데이터 전처리

04 데이터 분석 (PoC)

05 머신러닝 모델링

06 평가 및 개선사항

01 개요

심리학 테스트 분석 알고리즘 개발

심리학 테스트의 범주가 넓어짐에 따라 심리학을 통한 다른 분야의 데이터를 해석하려는 연구가 활발히 진행



DAICON

마키아벨리즘 심리테스트를 활용하여

참가자의 국가 선거 투표 여부 예측

02 데이터 변수 설명

Data Shape

```
train.shape
```

```
(45532, 78)
```

```
test.shape
```

```
(11383, 77)
```

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 56915 entries, 0 to 56914  
Data columns (total 77 columns):
```

Train set : 78개 속성의 45532 행

Test set : 타겟 데이터 제외 77개 속성의 11383 행

02 데이터 변수 설명

Data Summary

마키아벨리즘 설문 답변 내용 (20개) : QaA ~ QtA

마키아벨리즘 설문 답변 시간 (20개) : QaE ~ QtE

설문자 개인정보 (10개) : Age_group , gender , education, ..., etc

TIPi 성격 유형 설문 답변 내용 (10개) : tp01 ~ tp10

설문자 어휘능력 (16개) : wr_01~13 , wf_01~03

투표 여부 [타겟] (1개) : voted

02 데이터 변수 설명

마키아벨리즘 설문 답변 내용 : QaA ~ QtA

마키아벨리즘이란 ? (Machiavellism)

개인적인 욕구의 충족을 위해 남을 속이거나 조종하려는 욕구를 가리키는

단어로 성격심리학과 사회심리학에서 사용하는 용어이다.

60점 <

계산적

이기적

02 데이터 변수 설명

마키아벨리즘 설문 답변 내용 : QaA ~ QtA

총 20개의 문항으로 이루어짐

예시 : 사람을 다루는 가장 좋은 방법은 듣기 원하는 말을 해주는 것이다.

① ② ③ ④ ⑤

설문자는 1 (강한 부정) ~ 5 (강한 긍정) 으로 평가

특징 : 8개의 문항 (A번 , D번 , N번 , ...) 은 비식별을 위해 문항정보가 가려져 있다.

02 데이터 변수 설명

마키아벨리즘 설문 답변 시간 : QaE ~ QtE

20개의 문항에 대한 답변 시간 데이터

연속형 데이터로 상대적인 시간을 나타낸다

특징 : 이상치와 정상데이터의 구분이 어려울 만큼 넓은 범위의 데이터 $10^0 \sim 10^7$

02 데이터 변수 설명

설문자 개인정보 (10개)

age_group (연령대) : '10s', '20s', '30s', '40s', '50s', '60s', '+70s'

education (교육수준) :

- 1=Less than high school
- 2=High school
- 3=University degree
- 4=Graduate degree

02 데이터 변수 설명

설문자 개인정보 (10개)

Engnat (모국어 영어 여부) : 1=Yes 2=No

Urban (유년기 거주 지역) : 1=Rural (시골) , 2=Suburban (도심 주변) , 3=Urban (도심)

Familysize (형제자매 수) : 연속형 데이터

Gender (성별) : 'Female' , 'Male'

02 데이터 변수 설명

설문자 개인정보 (10개)

Race (인종) : 6개의 인종 + 기타

Asian, Arab, Black, Indigenous Australian,
Native American, White, Other

Religion (종교) : 11개의 종교 + 기타

Agnostic, Atheist, Buddhist, Christian_Catholic,
Christian_Mormon, Christian_Protestant,
Christian_Other, Hindu, Jewish, Muslim, Sikh, Other

02 데이터 변수 설명

설문자 개인정보 (10개)

Hand (손잡이) : 1=오른손 , 2=왼손 , 3=양손

Married (결혼 여부) : 1=미혼 , 2=기혼 , 3=이혼 혹은 사별

02 데이터 변수 설명

TIPi 성격 유형 설문 답변 내용 (10개)

TIPi란 ? (Ten Item Personality Inventory)

인간의 성격을 5가지의 상호 독립적인 요인들로 검사하는 Big 5 검사의 약식 버전 설문

외향성

정서안정성

친화성

개방성

외향성

02 데이터 변수 설명

TIP1 성격 유형 설문 답변 내용 (10개)

총 10개의 문항으로 이루어짐

예시 : 나는 다음 단어와 관련이 깊다 { 내성적이다 , 조용하다 }

① ② ③ ④ ⑤ ⑥ ⑦



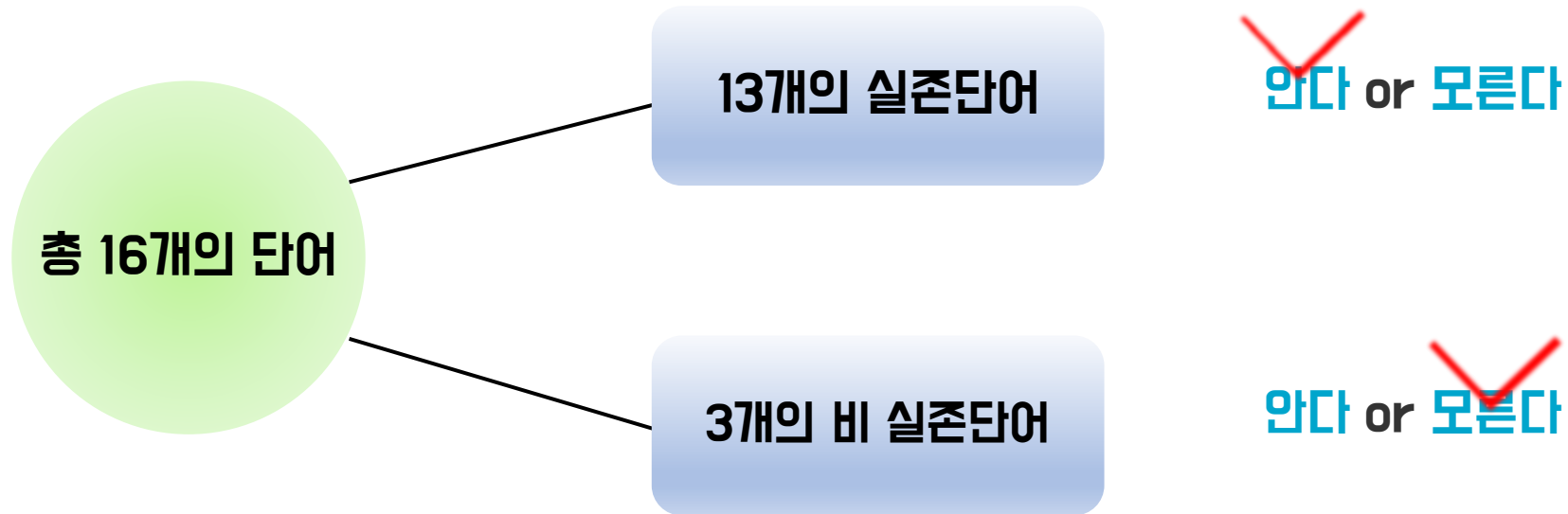
설문자는 1 (강한 긍정) ~ 7 (강한 부정) 으로 평가

특징 : 2 개 문항 씩 짝지어 하나의 성격 유형에 대한 점수로 계산

02 데이터 변수 설명

설문자의 어휘 능력 (16개)

- 명목형 이진 데이터로 1 = (안다)/ 0 = (모른다) 2가지 대답으로 나뉜다.



02 데이터 변수 설명

투표 여부 : 타겟 데이터

voted

지난 해 국가 선거 투표 여부 : 1 = Yes , 2 = No

03 데이터 전처리

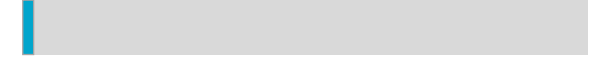


각 feature 결측값 비율

Hand(0.3%)



Familysize
(0.005%)



Education(1.1%)



Urban(0.7%)



Engnat(0.01%)



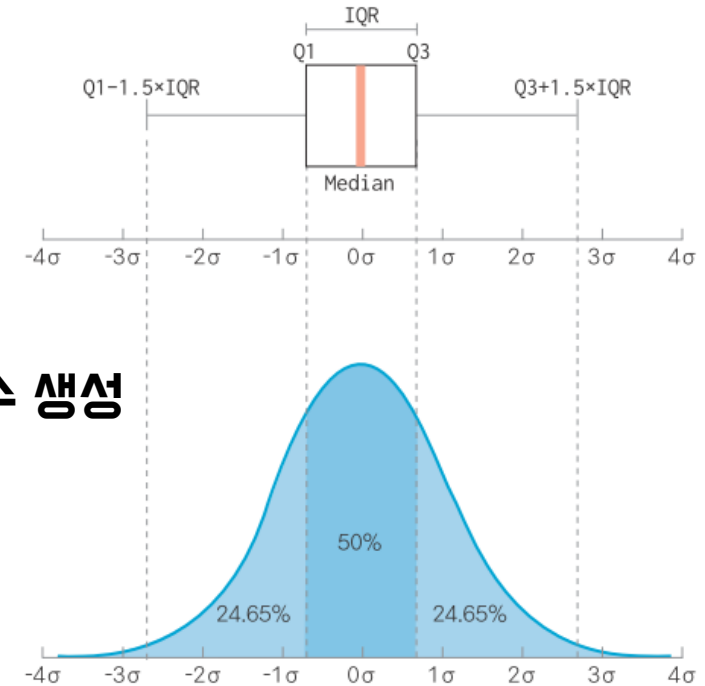
Married(0.02%)



03 데이터 전처리

4가지 처리 방식

1. 상관계수 기반 그룹화 후 속한 그룹의 최빈값으로 대체
2. 상관계수 기반 그룹화 후 속한 그룹의 데이터 비율을 전달하여 난수 생성
3. 1% 미만의 이상치는 처리하지 않고 버림
4. KNN 기반의 라벨링을 통해 해당 레이블의 과반수 데이터를 활용



03 데이터 전처리

Familysize

```
data.familysize.sort_values(ascending=False)[:10]
```

24598	2147483647
379	999
25661	100
21567	44
34847	44
12056	44
28111	34
41326	30
48605	23
34749	21

Name: familysize, dtype: int64

drop

```
outlier_idx = data.familysize[data.familysize>99].index
for idx in outlier_idx:
    if idx < split_point:
        data = data.drop(idx,axis=0)
        split_point-=1
data.shape # 3개 행 모두 제거
(56912, 76)
```

상식적으로 납득하기 어려운 형제자매 수

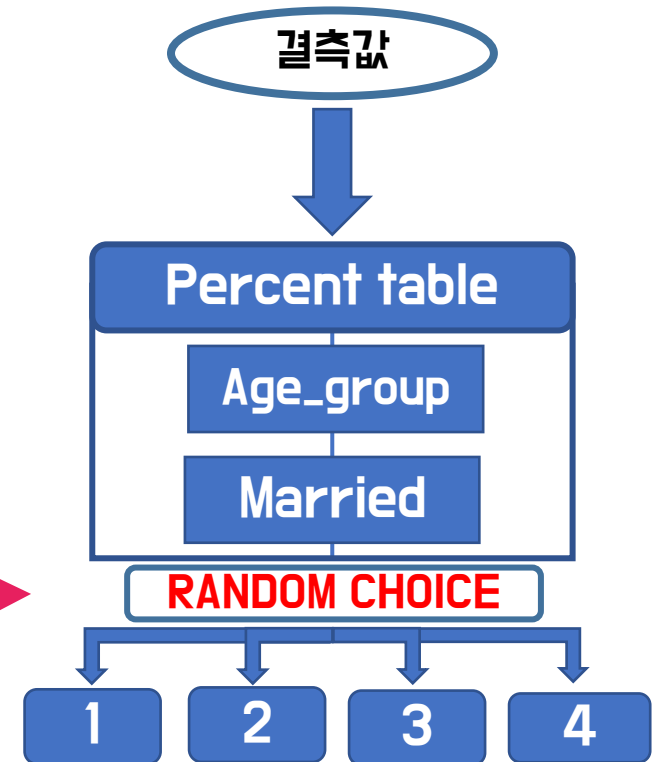
2147483647 . 999 . 100

03 데이터 전처리

Education

1. 상관계수 기준 선형관계가 있는 연령대, 결혼 여부 정보로 그룹화
2. 그룹화한 테이블을 기준으로 난수의 비율을 결정하여 결측값 대체

	education	1	2	3	4
age_group	married				
+70s	1	0.111111	0.333333	0.222222	0.333333
	2	0.028249	0.214689	0.322034	0.435028
	3	0.038835	0.407767	0.271845	0.281553



03 데이터 전처리

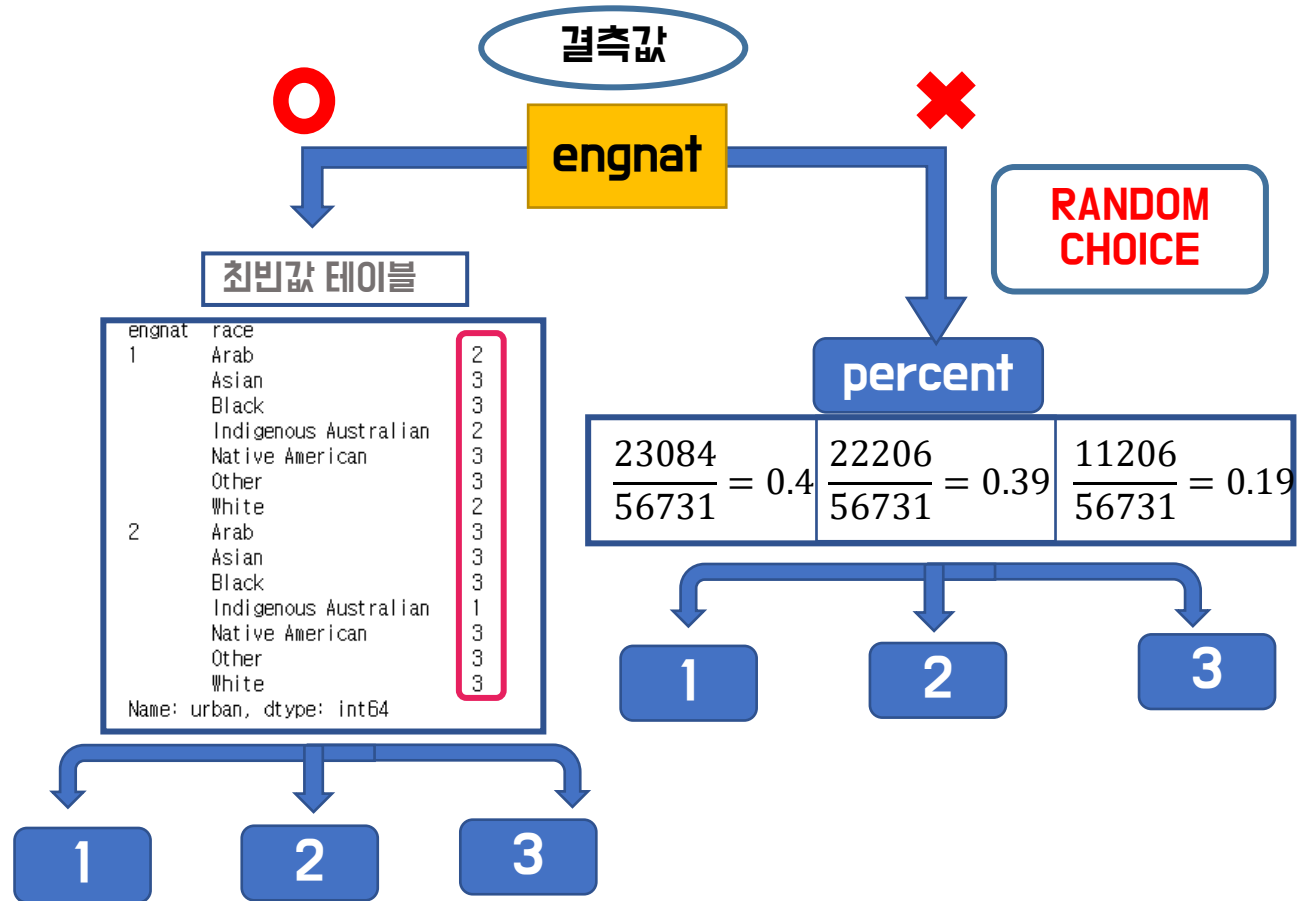
Urban

1. 상관계수 기준 모국어 영어 여부와 인종

데이터와 선형 관계

2. 두 컬럼을 기준으로 그룹화

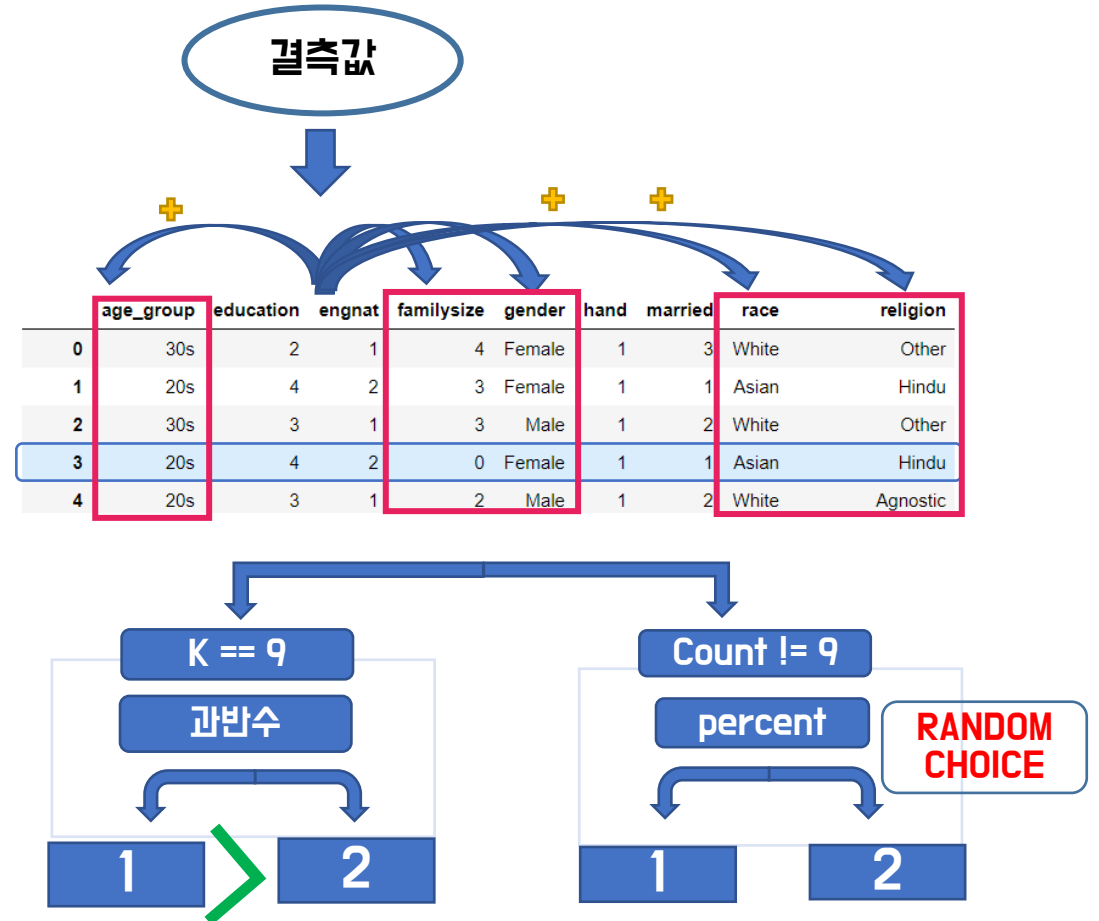
3. 각 그룹의 최빈값을 활용



03 데이터 전처리

Engnat

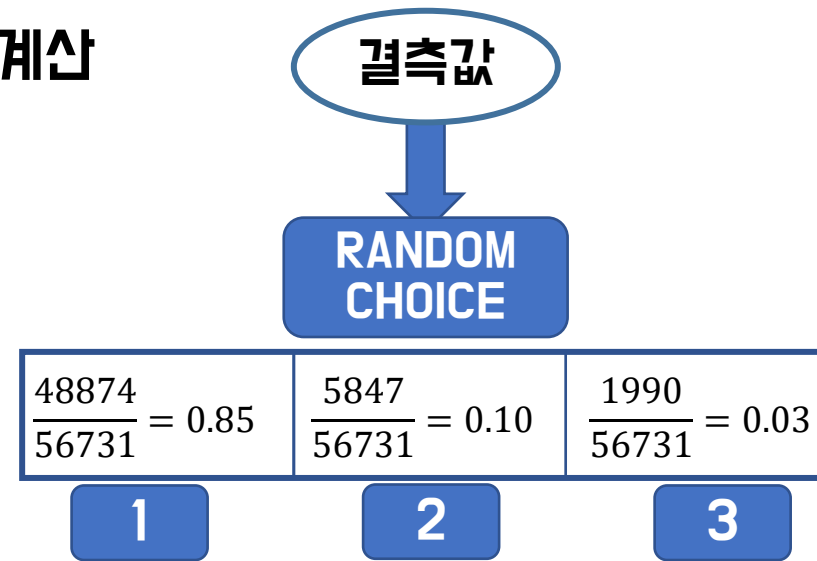
1. KNN 적용을 위해 결측 위험이 없는 컬럼들만 활용
2. 결측을 포함한 행의 데이터를 기준으로 KNN
3. 추출된 그룹의 과반수 데이터 활용



03 데이터 전처리

Hand

1. 결측치를 제외한 hand 데이터의 전체 비율을 계산
2. 비율 기반 난수를 생성하여 활용



03 데이터 전처리

특성 추가 (Feature Engineering)

1. 20개의 마키아벨리즘 테스트 문항을 통해 성향 점수를 계산하고 MACH_score 컬럼 추가
2. 불균형이 심한 답변 시간 데이터로부터 이분화를 통해 총 답변시간 컬럼 추가
3. TIPI 계산방법에 따라 5 유형의 성격에 대한 점수 컬럼 추가

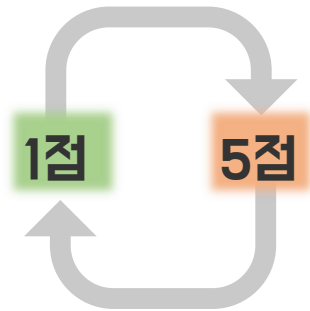


03 데이터 전처리

마키아벨리즘 답변내용 : MACH_score

마키아 스코어 : 20개의 답변 수치의 평균

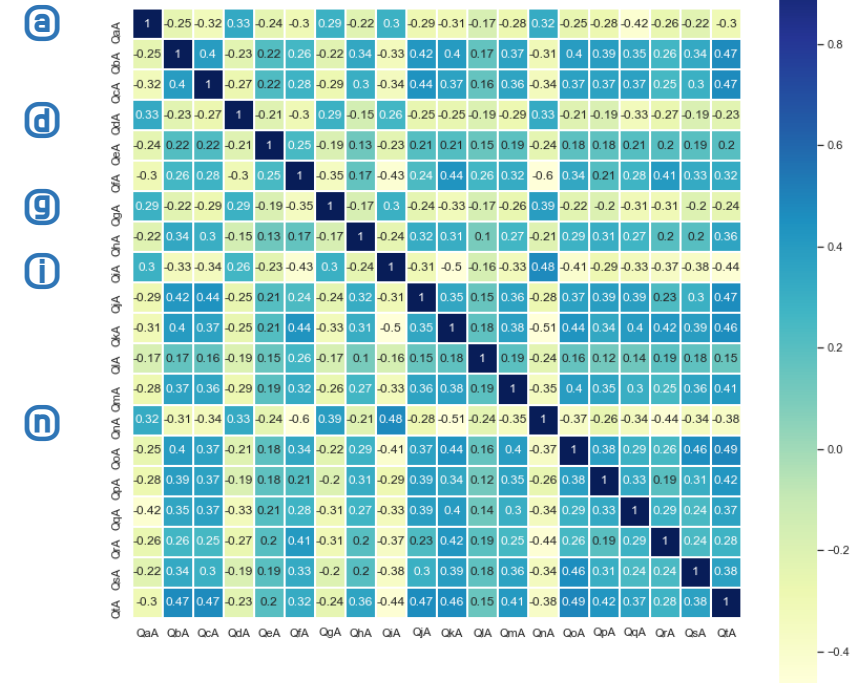
문항별로 부정, 긍정의 성격이 달라 스코어 계산 시 점수를 뒤집어야 하는 문항 존재



8개의 비공개 문항의 경우 부호를 알아내 적용 해야함

마키아벨리즘 답변내용 : MACH_score

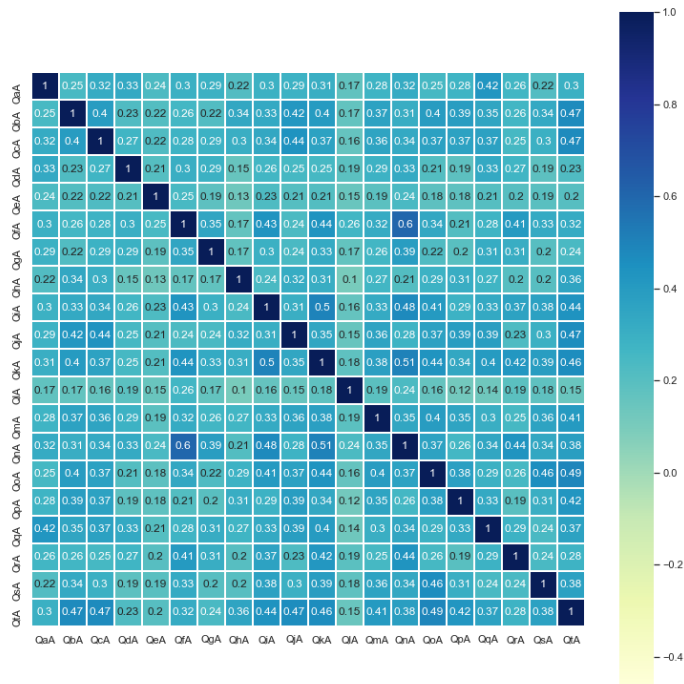
공개된 질의응답 보호 적용 후 상관계수 확인



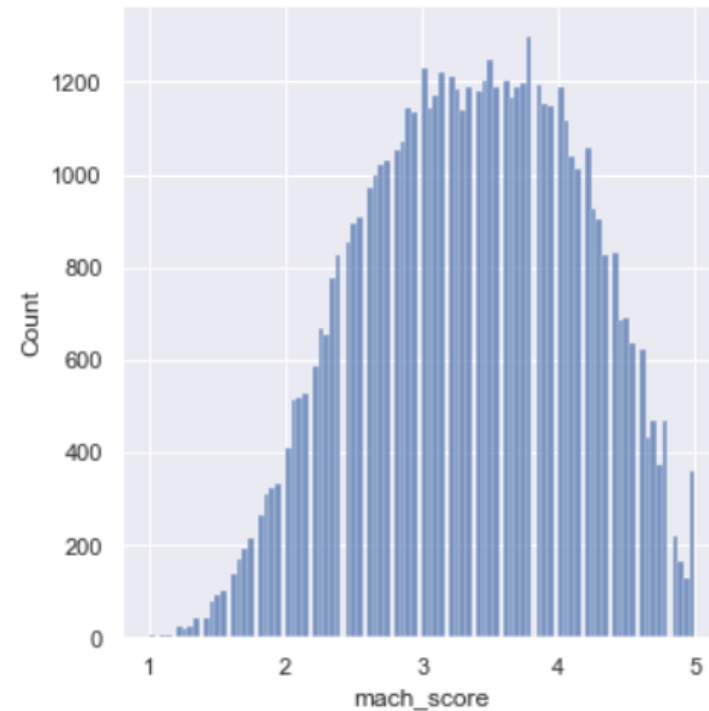
03 데이터 전처리

마키아벨리즘 답변내용 : MACH_score

5 항목의 점수를 뒤집은 후



추가한 MACH_score 데이터의 분포

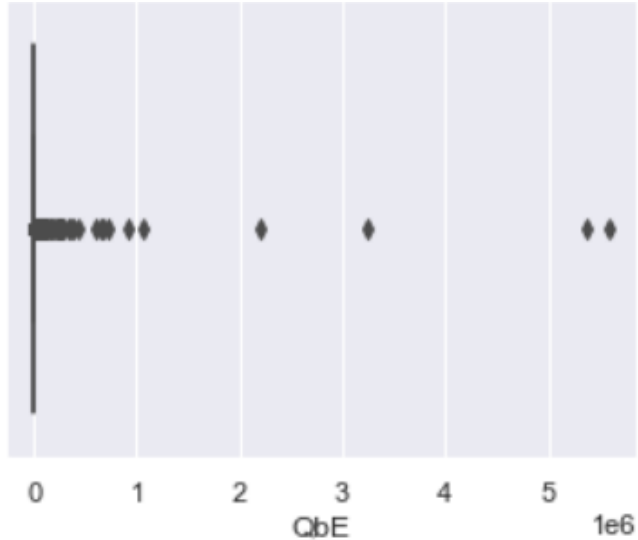


03 데이터 전처리

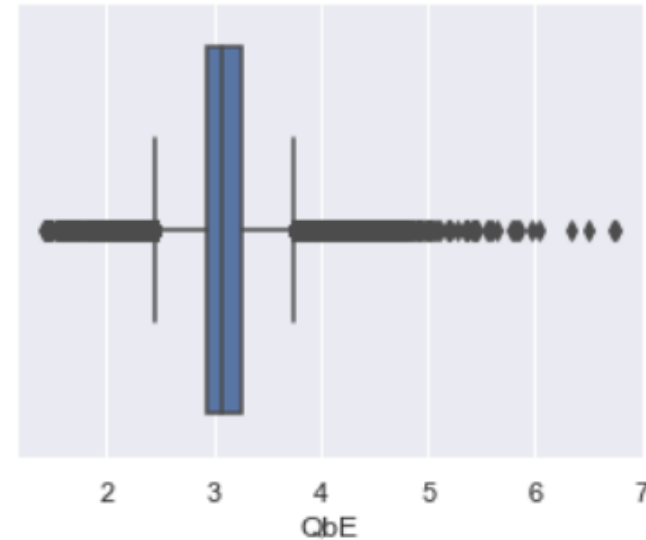
마키아벨리즘 답변시간

특징 : 극단값이 매우 많고 큰 데이터

기존 데이터 boxplot



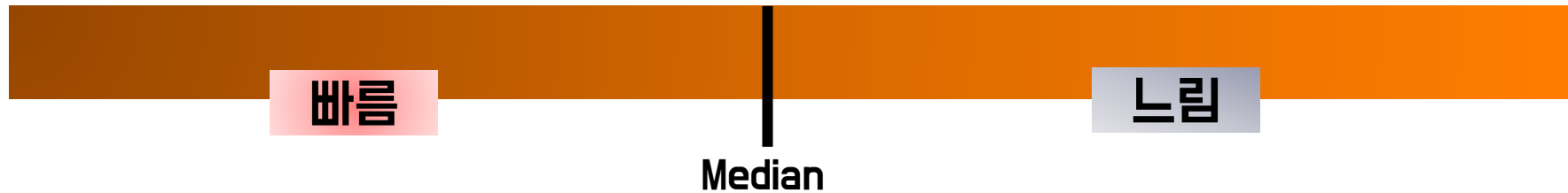
로그 스케일 후 boxplot



03 데이터 전처리

마키아벨리즘 답변시간

1. 각 컬럼을 중위값 기준으로 답변이 빠름(0)과 느림(1)으로 변환



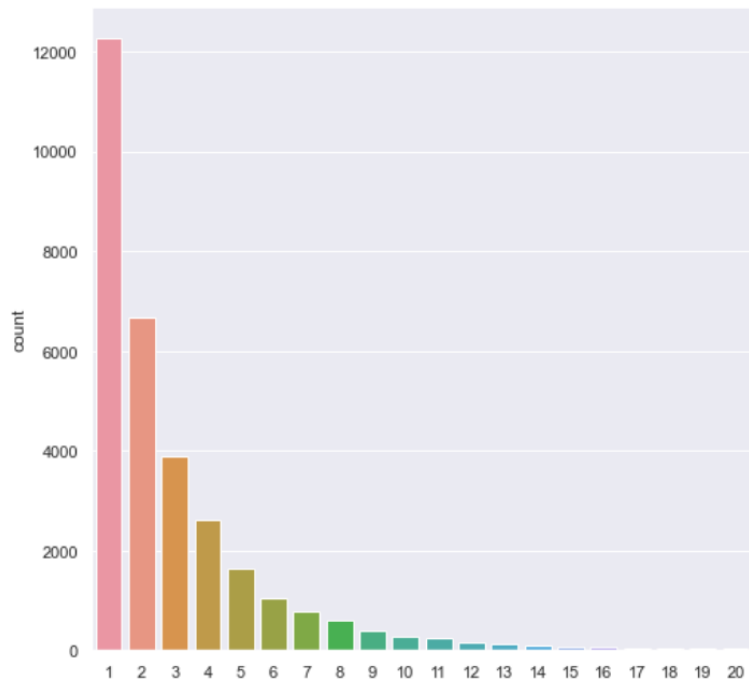
2. 변환된 데이터를 바탕으로 총 답변 시간 점수 컬럼을 추가 (0~20 사이 값)

A번	B번	C번	...	S번	T번	
1	0	0		0	1	= 12

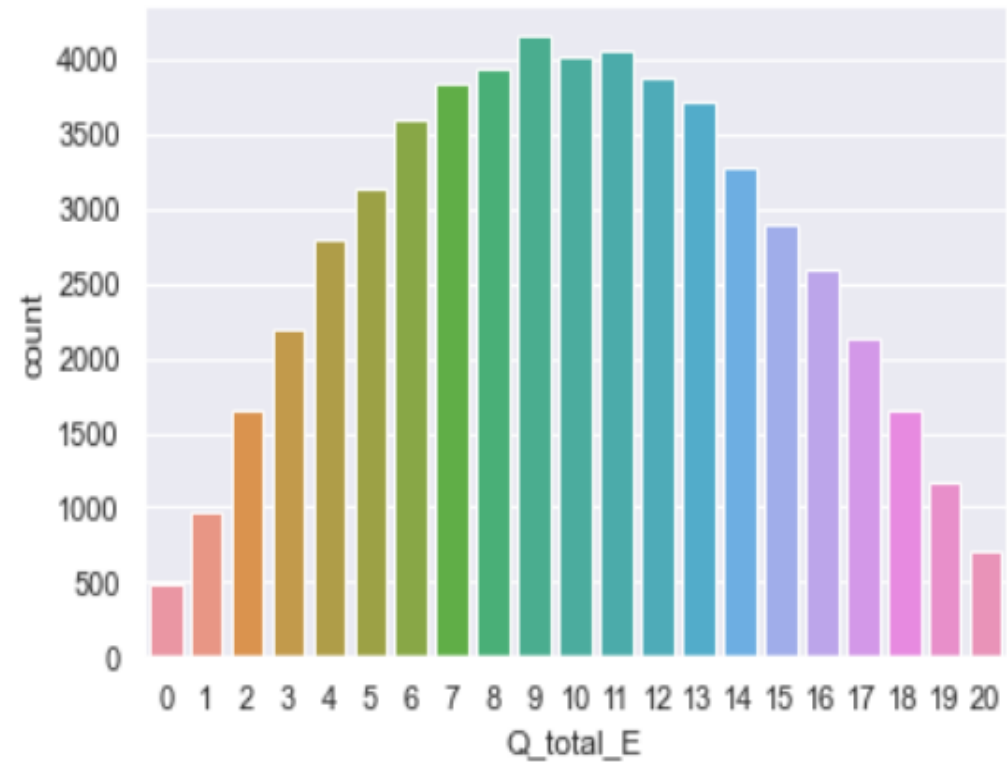
03 데이터 전처리

마키아벨리즘 답변시간

전처리 이전 이상치 데이터 분포



전처리 후 추가된 칼럼 데이터 분포



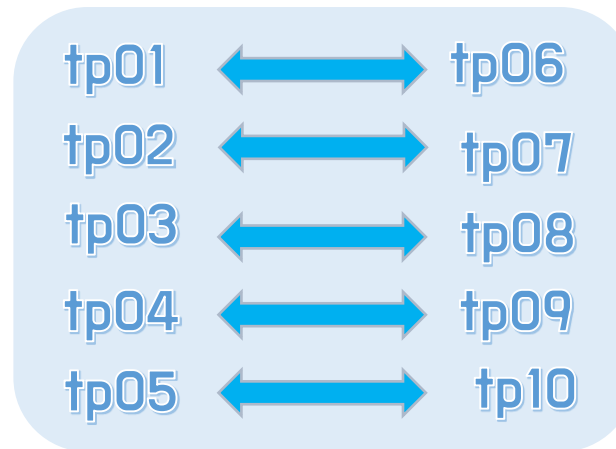
03 데이터 전처리

TIPi 5 유형 성격 점수 계산

각 질문사이 상관관계수 heatmap



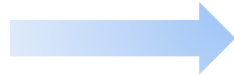
질문 쌍



03 데이터 전처리

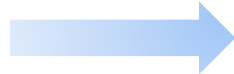
TIP1 5 유형 성격 점수 계산

$$\frac{tp01+tp06}{2}$$



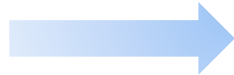
Extraversion (외향성)

$$\frac{tp02+tp07}{2}$$



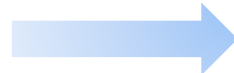
Agreeableness (친화성)

$$\frac{tp03+tp08}{2}$$



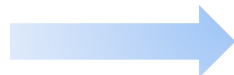
Conscientiousness (성실성)

$$\frac{tp04+tp09}{2}$$



Emotional Stability (심리적 안정성)

$$\frac{tp05+tp10}{2}$$



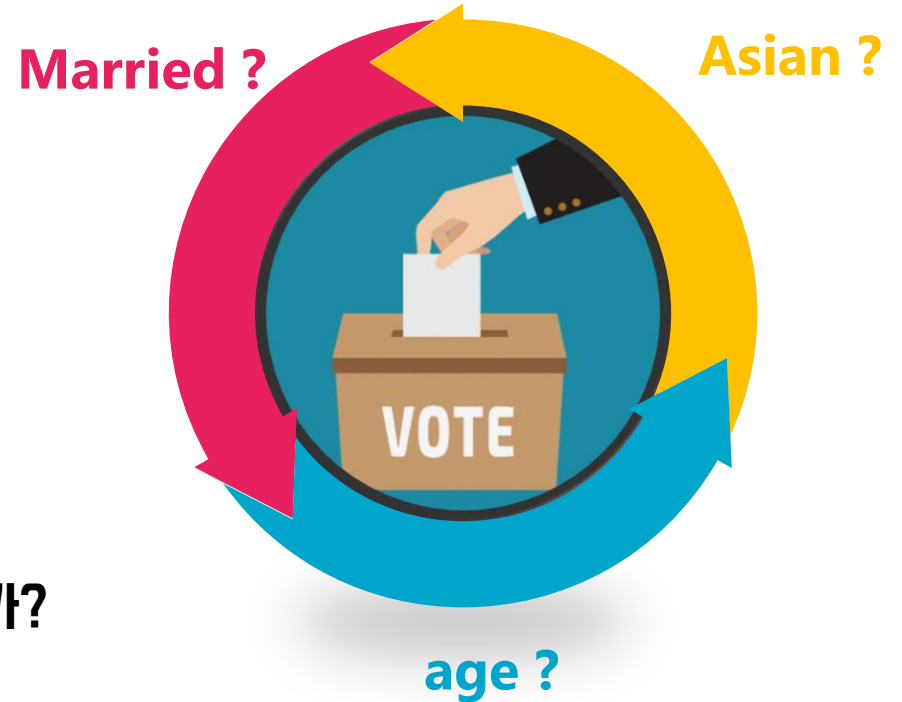
Openness to Experiences (개방성)

5개의 성격 칼럼 추가

04 데이터 분석 (PoC)

테이터들 사이의 관계를 추측하고 검증

1. 교육 수준이 설문자의 투표 참여에 영향을 미칠까?
2. 나이와 마키아벨리즘 성향은 얼마나 관련이 있을까?
3. 손잡이와 성별의 차이가 흔히 예상하는 특징들과 맞게 나올까?
4. 군집화를 통한 유권자 분리



04 데이터 분석 (PoC)

1. 교육 수준이 설문자의 투표 참여에 영향을 미칠까?

결혼 상태

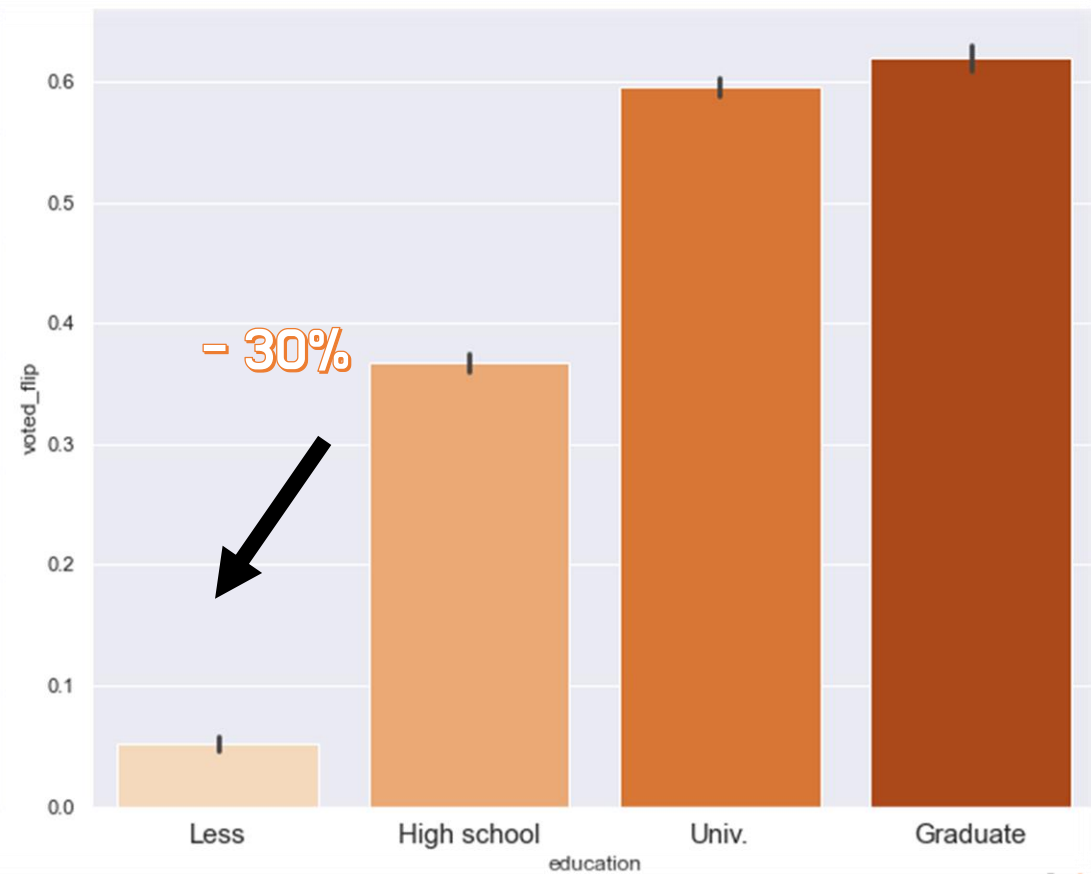
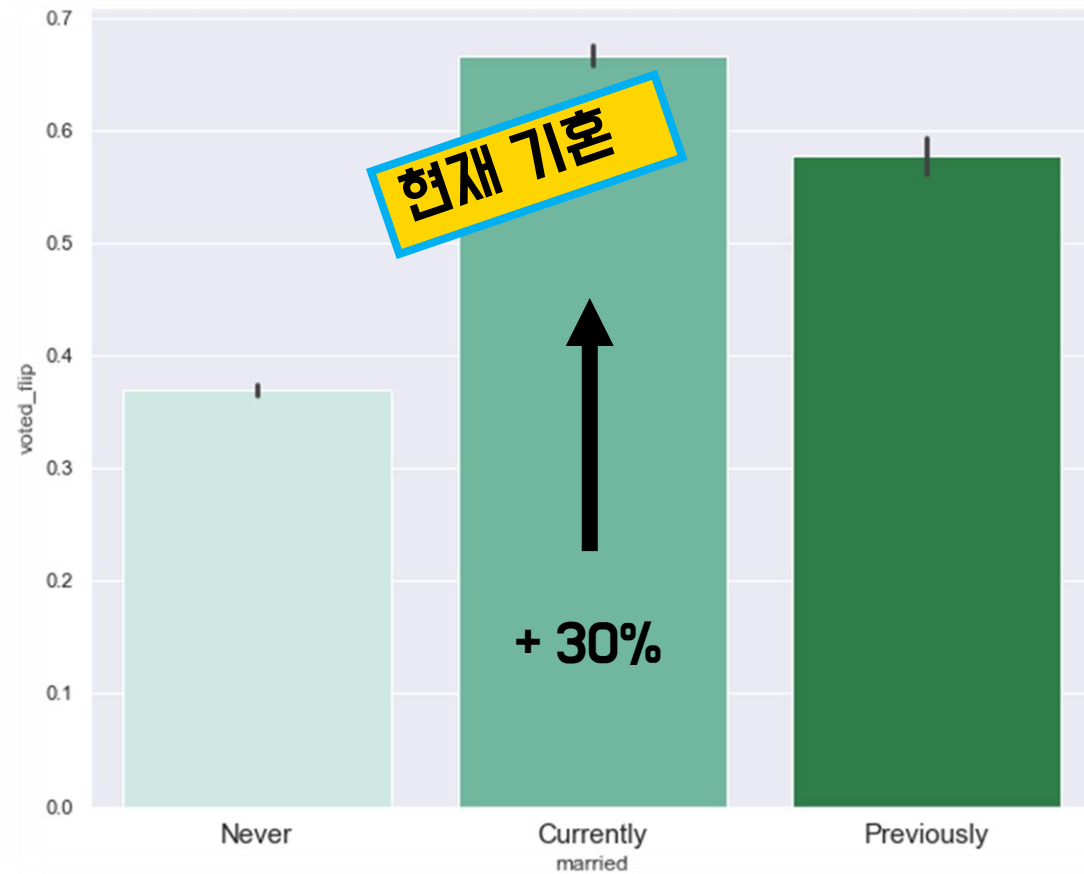
최종 학력

```
train.corr(method='spearman')['voted'].abs().sort_values(ascending=False)
```

voted	1.000000
education	0.337424
married	0.241011
QqA	0.123367
wr_11	0.116715
...	...
QkE	0.018619
QeA	0.013556
hand	0.006725
QdA	0.002807
wf_01	0.000531

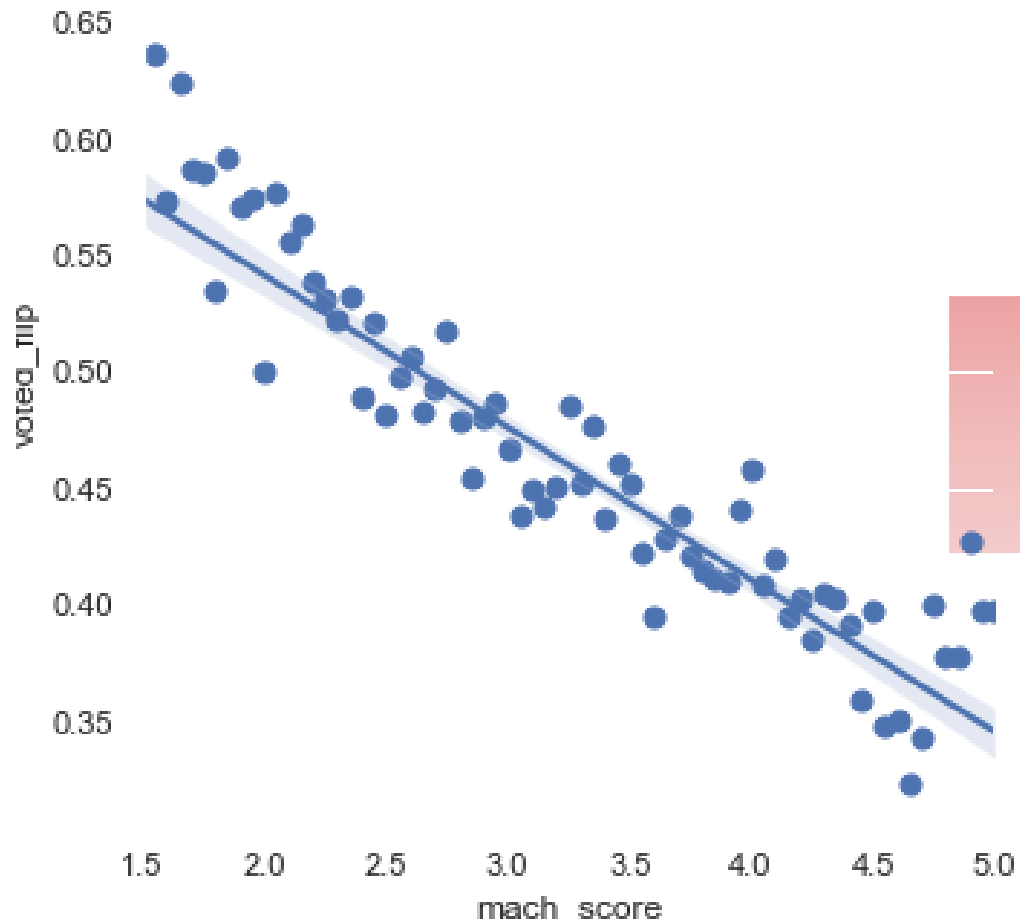
04 데이터 분석 (PoC)

1. 교육 수준이 설문자의 투표 참여에 영향을 미칠까?



04 데이터 분석 (PoC)

1. 마키아벨리즘이 설문자의 투표 참여에 영향을 미칠까?

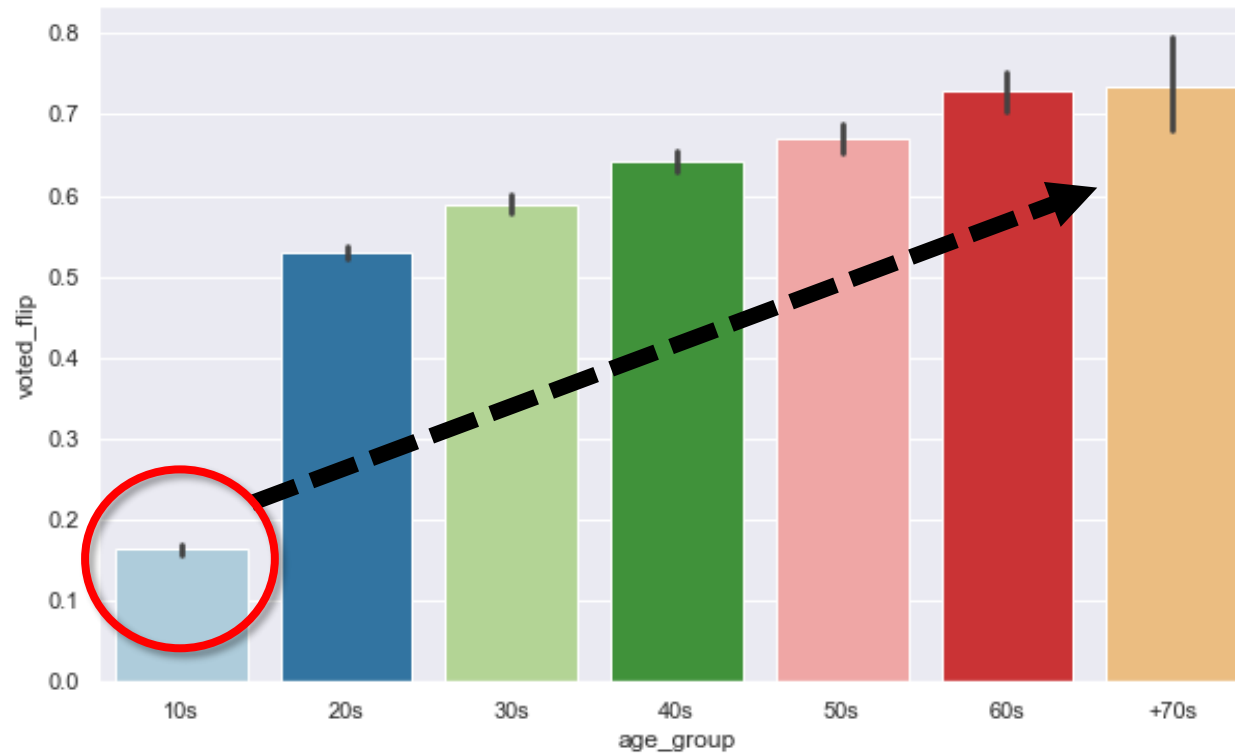


마키아벨리즘 성향과 투표는 반비례 관계

04 데이터 분석 (PoC)

1. 설문자의 나이가 투표 참여에 영향을 미칠까?

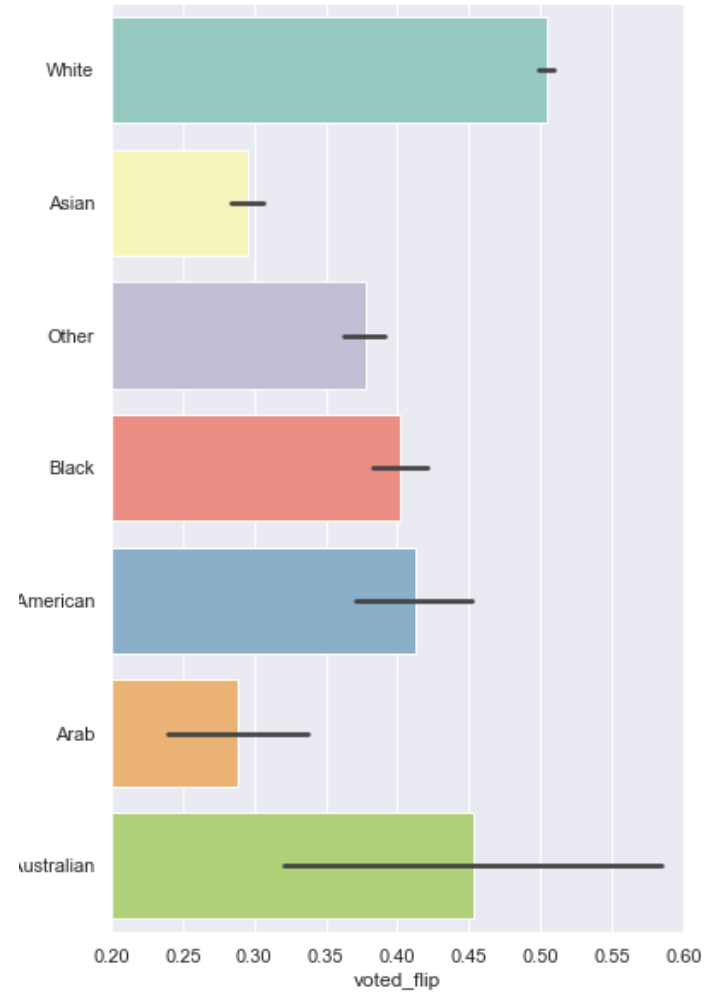
그 밖 : age , race



04 데이터 분석 (PoC)

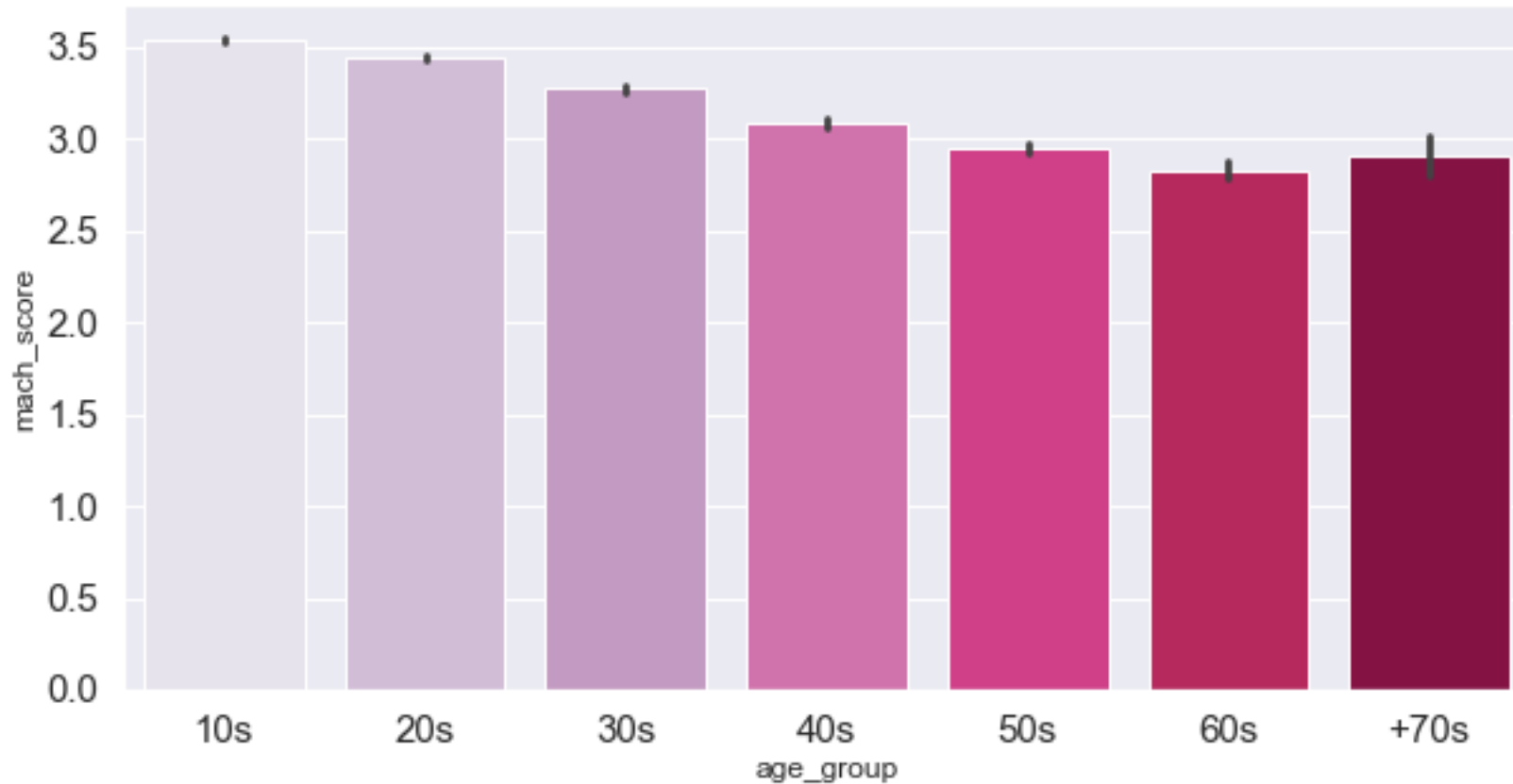
1. 설문자의 인종이 투표 참여에 영향을 미칠까?

Arab , Asian 25%



04 데이터 분석 (PoC)

2. 나이와 마키아벨리즘 성향은 얼마나 관련이 있을까?



연령대가 높은 그룹일수록
마키아벨리즘 성향이 낮아짐

04 데이터 분석 (PoC)

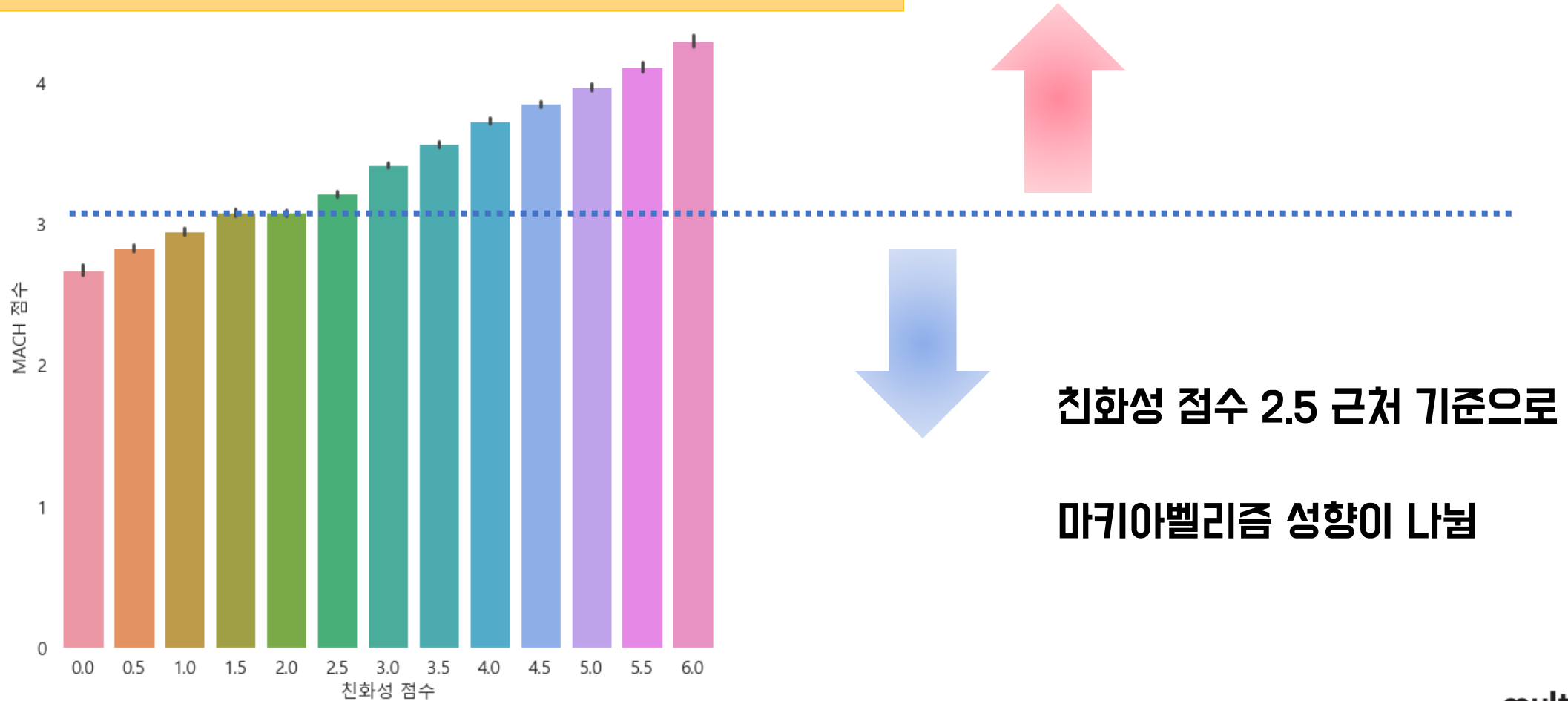
2. TIPI 성향은 마키아벨리즘 성향은 얼마나 관련이 있을까?

	mach_score
mach_score	1.000000
Extraversion	0.084890
Agreeableness	0.478096
Conscientiousness	0.107476
Emotional Stability	-0.015156
Openness to Experiences	-0.008133

친화성

04 데이터 분석 (PoC)

2. TIP1 성향은 마키아벨리즘 성향은 얼마나 관련이 있을까?



04 데이터 분석 (PoC)

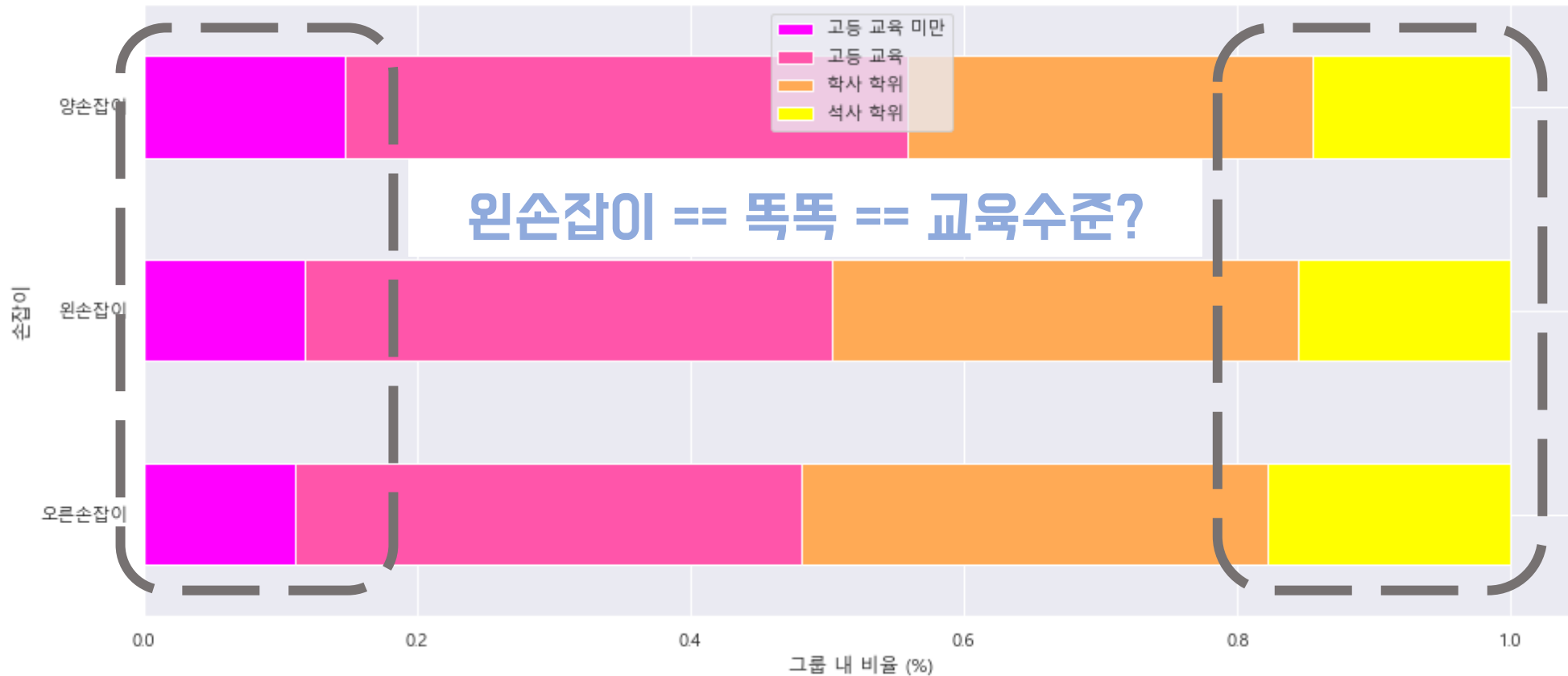
3. 손잡이와 성별의 차이가 흔히 예상하는 특징들과 맞게 나올까?

예상1. 왼손잡이 == 똑똑 == 교육수준?

예상2. 남녀의 성격 유형 차이는 결혼 여부에 나타날까?

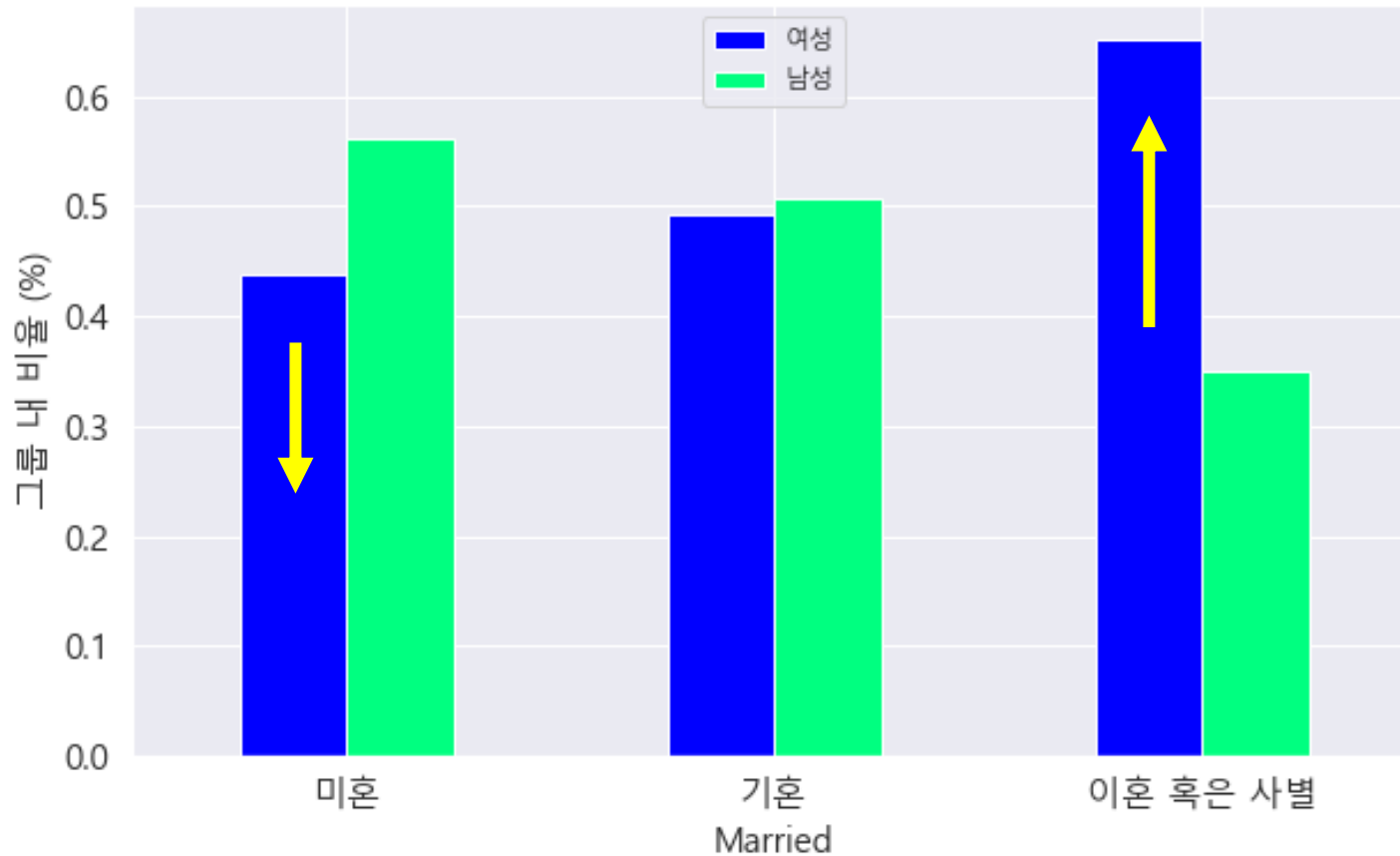
04 데이터 분석 (PoC)

3. 손잡이와 성별의 차이가 흔히 예상하는 특징들과 맞게 나올까?



04 데이터 분석 (PoC)

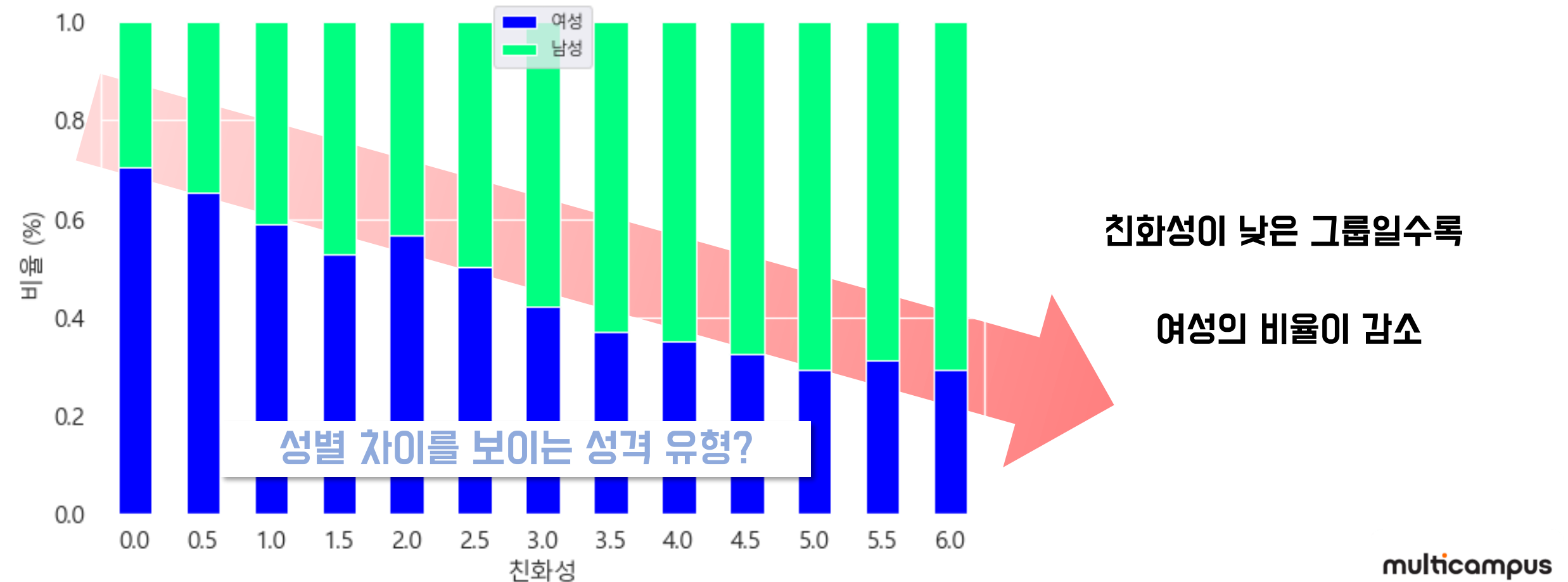
3. 남녀의 성격 유형 차이는 결혼 여부에 나타날까?



미혼과 이혼 그룹에서 남녀의
분포 비율이 상이함

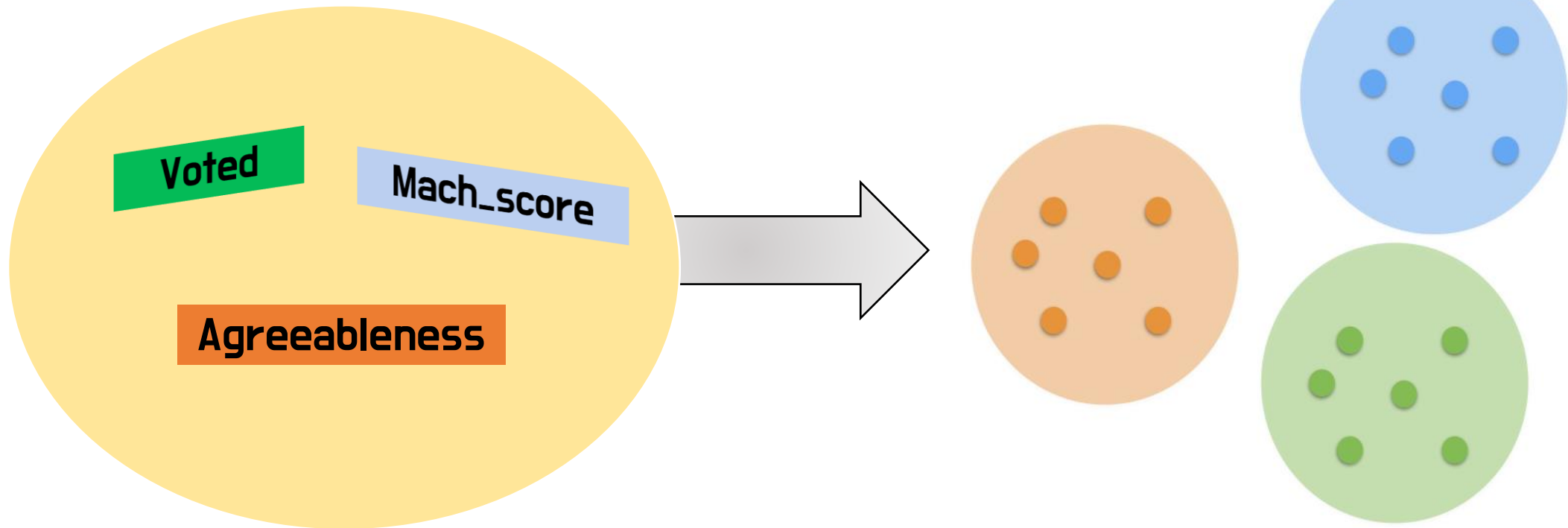
04 데이터 분석 (PoC)

3. 남녀의 성격 유형 차이는 결혼 여부에 나타날까?



04 데이터 분석 (PoC)

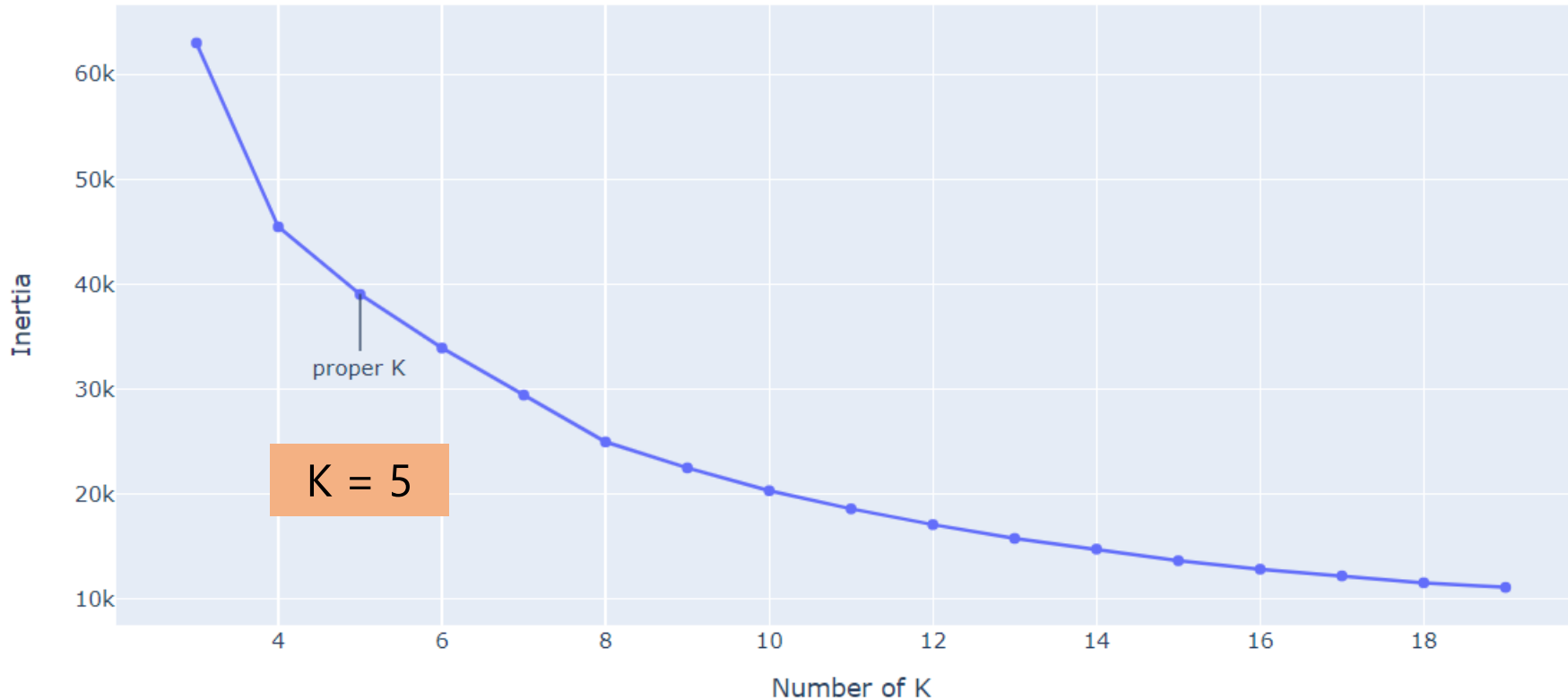
4. 군집화를 통한 유권자 분리



04 데이터 분석 (PoC)

4. 군집화를 통한 유권자 분리

Elbow graph of K-means

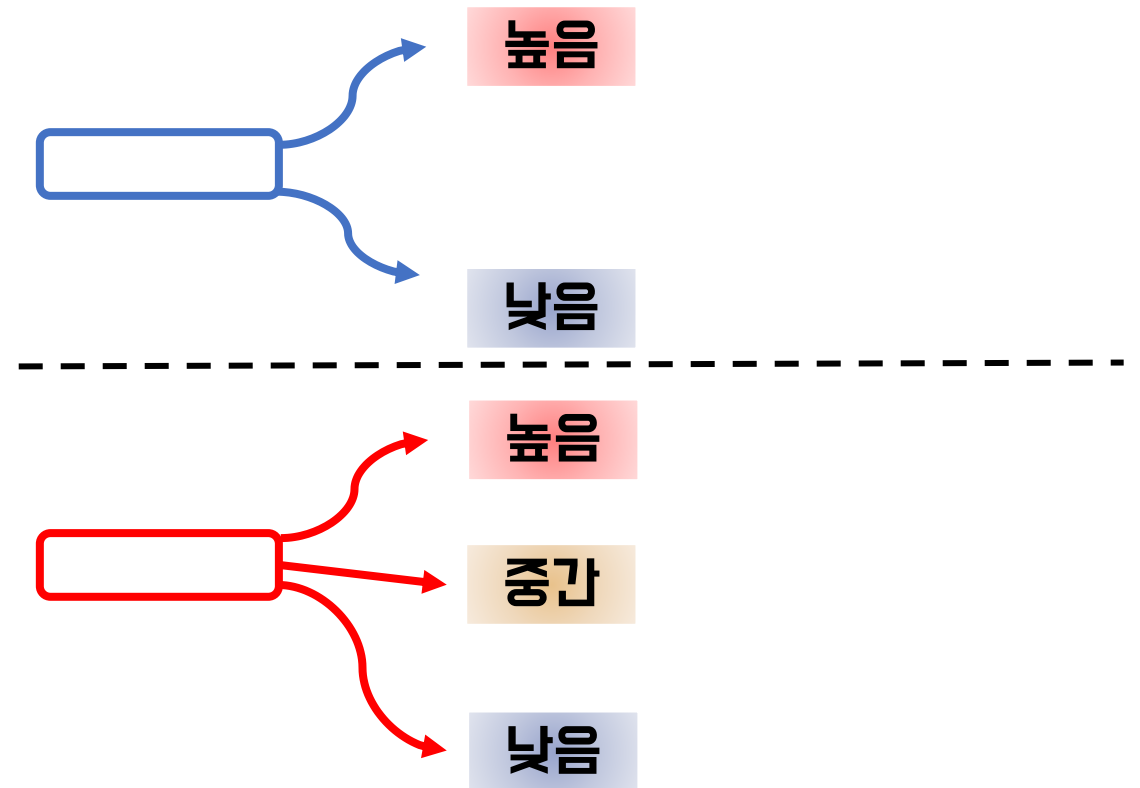


04 데이터 분석 (PoC)

4. 군집화를 통한 유권자 분리

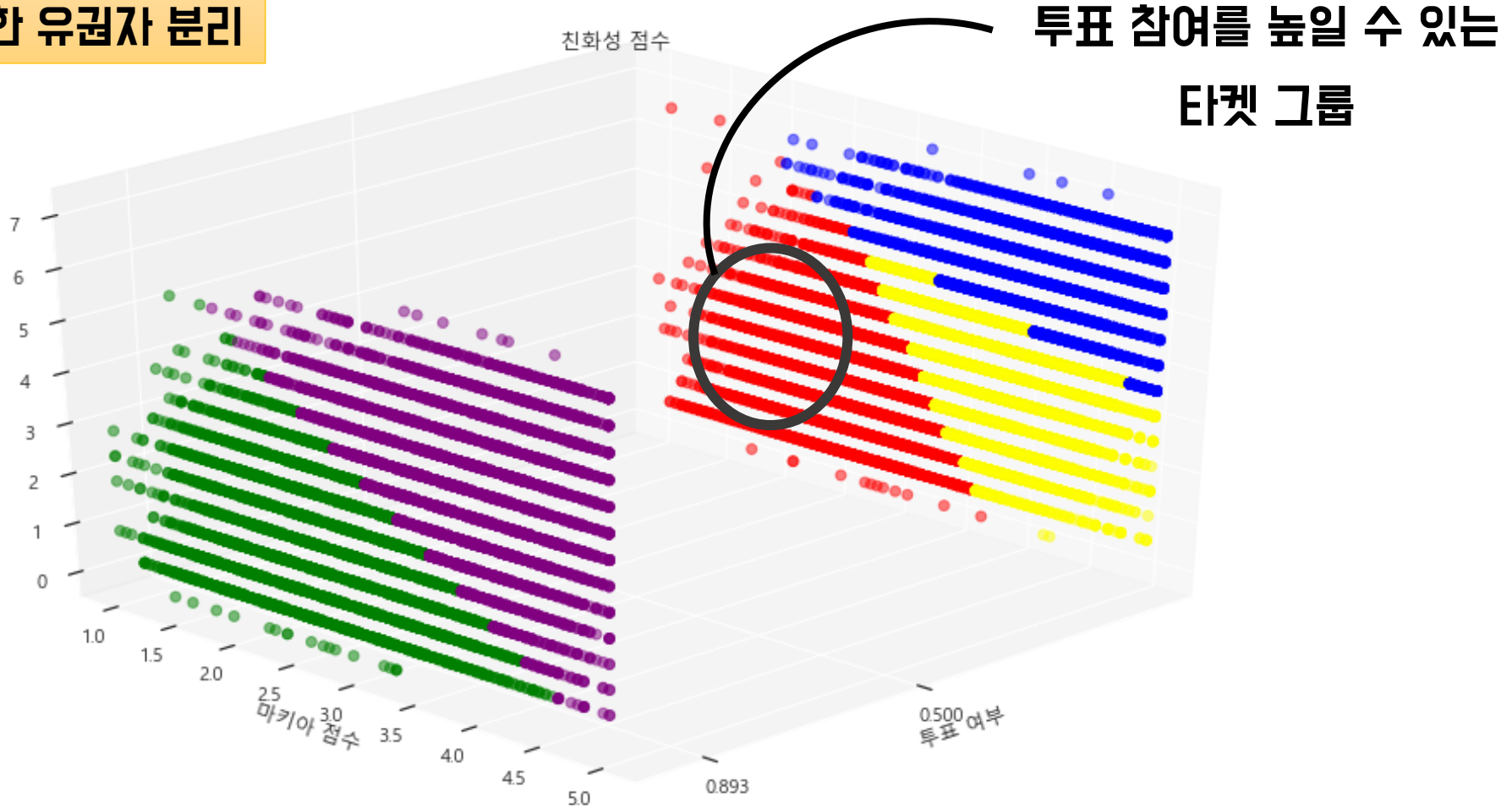
	voted	MachScore	Agreeableness
0	6.494805e-15	4.099582	5.089110
1	1.193490e-14	2.576245	2.209265
2	1.000000e+00	2.698080	2.116366
3	7.438494e-15	3.695750	3.055440
4	1.000000e+00	3.863547	4.061313

마키아벨리즘 성향



04 데이터 분석 (PoC)

4. 군집화를 통한 유권자 분리



05 머신러닝 모델링

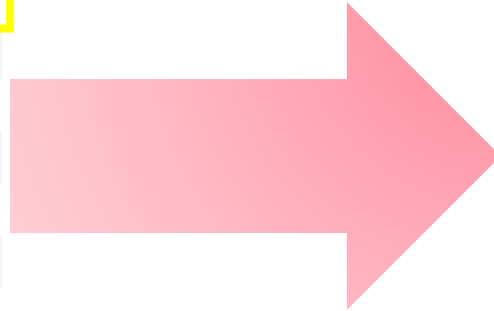


기초적인 데이터 전처리부터 베이스 모델 구축까지 가능한

파이썬의 Auto ML 패키지

05 머신러닝 모델링

	Description	Value
0	session_id	7258
1	Target Type	Binary
2	Label Encoded	1: 0, 2: 1
3	Original Data	(45529, 102)
4	Missing Values	False
5	Numeric Features	60
6	Categorical Features	41
7	Ordinal Features	False
8	High Cardinality Features	False
9	High Cardinality Method	None
10	Sampled Data	(31870, 102)
11	Transformed Train Set	(22308, 142)
12	Transformed Test Set	(9562, 142)
13	Numeric Imputer	mean
14	Categorical Imputer	constant
15	Normalize	False



〈Columns〉

76 : original

102 : preprocessed

142 : Auto ML

05 머신러닝 모델링

from PyCaret,

Gradient Boosting Classifier

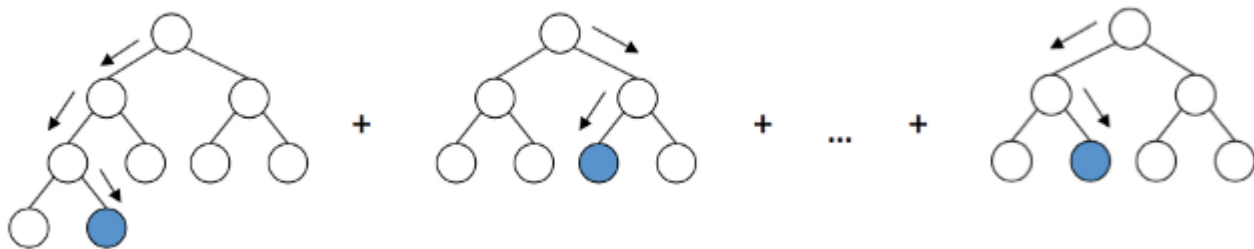
CatBoost Classifier

Light Gradient Boosting Machine

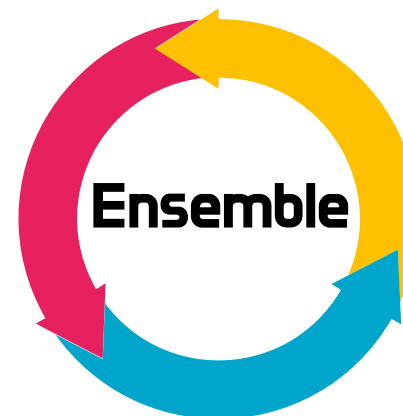
```
best_3 = compare_models(sort = 'AUC', n_select = 3)
```

	Model	Accuracy	AUC	Recall	Prec.	F1
0	Gradient Boosting Classifier	0.6932	0.7650	0.6406	0.7608	0.6955
1	CatBoost Classifier	0.6927	0.7640	0.6533	0.7523	0.6992
2	Light Gradient Boosting Machine	0.6931	0.7630	0.6440	0.7586	0.6965
3	Ada Boost Classifier	0.6890	0.7572	0.6529	0.7467	0.6966
4	Linear Discriminant Analysis	0.6786	0.7555	0.7281	0.6975	0.7124
5	Logistic Regression	0.6787	0.7550	0.7292	0.6973	0.7128
6	Extra Trees Classifier	0.6816	0.7518	0.6854	0.7193	0.7019
7	Extreme Gradient Boosting	0.6741	0.7418	0.6646	0.7184	0.6904
8	Random Forest Classifier	0.6585	0.7136	0.6127	0.7209	0.6623
9	Naive Bayes	0.6212	0.6768	0.5275	0.7054	0.6035
10	K Neighbors Classifier	0.6129	0.6420	0.6628	0.6414	0.6518
11	Decision Tree Classifier	0.6141	0.6109	0.6445	0.6480	0.6462
12	Quadratic Discriminant Analysis	0.5217	0.5391	0.4499	0.5840	0.4939
13	SVM - Linear Kernel	0.6779	0.0000	0.6884	0.7200	0.6991
14	Ridge Classifier	0.6789	0.0000	0.7285	0.6977	0.7127



05 머신러닝 모델링



```
blended = blend_models(estimator_list = best_3, fold = 5, method = 'soft')
```



	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Voting Classifier	0.6898	0.7638	0.6393	0.7558	0.6927	0.384	0.3894

64	씨룻메		0.77674	1	2분 전
1	문성민		0.7829	40	2일 전

06 평가 및 개선사항

1. 사용하지 못한 데이터(Word) 로 모델의 성능을 더 발전시킬 수 있지 않았을까
2. 설문조사를 영어권 국가 대신 **국내로 한정하여 진행했다면** 결과가 바뀌었을까
3. 생각보다 **새로운 인사이트**를 도출해내는 PoC를 해내지는 못했음
4. AutoML 패키지를 이용하지 않고, **직접 파라미터 조정**을 통해 모델링을 했으면 하는 아쉬움
5. 난수 생성의 결측 처리 방식 대신 조금 **더 통계학으로 접근하면** 더 좋은 성능을 보일지



Q&A
THANKS