패스트캠퍼스 데이터사이언스 스쿨 14기
머신러닝 프로젝트 3조
박경원 전진경 정성용

마키아벨리즘 성향 테스트 결과 데이터 기반
국가선거 투표참여 여부 예측

마키아벨리즘성향테스트결과 데이터 기반

국가선거 투표참여 여부 예측

1. 머신러닝 프로젝트 **목표**
2. **데이터** 소개
3. 모델링 **과정**
4. 모델 **시연**
5. 분류 **결과 해석**
6. **한계** 및 과제

#1. GOAL

목표 = "어떤사람이 투표했나 알아맞히기"

마키아벨리즘 성향 테스트 답변 데이터

머신러닝 분류모델 활용

투표참여 여부 예측

#2. DATA INFO

EDA

DATA TOPIC

마키아벨리즘이란 ?

국가의 발전과 인민의 복리증진을 위해서는

어떠한 수단이나 방법도 허용된다는

국가 지상주의적인 정치 이념

출처 | 표준국어대사전

#2. DATA INFO

그래서, 마키아벨리즘 성향이면 어떻다는 거?

+ ▲

마키아벨리즘 성향이 높은 사람

낮은 친화력, 높은 성실성

사람들과 소통할 때
계산적이고 **신중**하게 접근하는 경향

평균 약70점

마키아벨리즘 성향이 낮은 사람

- ▼

비교적 수동적, 순응적

사람들과 소통할 때
개인적이며 **감정을 이입**하여 접근하는 경향

추측해보자

마키아벨리즘 성향이 강한 사람은
계산적이니까
투표에 더 적극적으로 참여하려나요?

#2. DATA INFO

DATASET = 테스트에 대한 각 참여자의 답변들

< DATASET 기본 정보 >		
1	데이터 출처	심리학 공공데이터 사이트 (미국) https://openpsychometrics.org/
2	데이터 내용 (73,489 , 105)	2017년 7월부터 2019년 3월까지의 기간 중 **온라인**에서 진행된 마키아벨리즘 테스트의 **각국 참여자 답변 데이터 (영어 사용자)**
3	데이터 구성	① **마키아벨리즘 성향 판단** 테스트 답변 ② 성격, 연령 등 **인적사항 설문조사** 답변

마키아벨리즘 테스트 질문지 구성

< 마키아벨리즘 성향 판단 테스트 예시 >

Q8 마키아벨리즘 +▲

대부분의 사람들은 강제하지 않으면 열일하지 않는다

1점 ---------- 매우 반대
2점 ---------- 조금 반대
3점 ---------- 보통
4점 ---------- 조금 동의
5점 ---------- 매우 동의

Q13 마키아벨리즘 +▲

일반인과 범죄자의 가장 큰 차이는 발각될 만큼 멍청한지 여부에 있다

1점 ---------- 매우 반대
2점 ---------- 조금 반대
3점 ---------- 보통
4점 ---------- 조금 동의
5점 ---------- 매우 동의

Q11 마키아벨리즘 -▼

세계적인 리더들은 청렴하고, 도덕적이다

1점 ---------- 매우 반대
2점 ---------- 조금 반대
3점 ---------- 보통
4점 ---------- 조금 동의
5점 ---------- 매우 동의

마키아벨리즘 테스트 질문지 구성

< 마키아벨리즘 성향 측정 척도에 따른 질문지 구분 >

	+ ▲ 마키아벨리즘 지지 성향 판단 (10)	- ▼ 마키아벨리즘 반대 성향 판단 (10)
T	Q12. 아무나 믿는 사람은 문제에 잘 휘말린다 Q15. 중요한 사람에게는 아첨을 해야 한다 Q2. 사람을 잘 다루려면 듣기 좋은 소리를 해야 한다 Q1. 일에 대한 진짜 목적은 도움이 되지 않는 한, 남들에게 말하지 않아야 한다	Q3. 도덕적으로 옳은 일만 행해야 한다 Q6. 정직함이 최선의 방책이다 Q10. 무엇인가를 부탁할 때는 그 일이 중요한 듯이 말하기보다 솔직한 이유를 말해야 한다 Q16. 모든 면에서 선하게 행하는 것이 가능하다 Q7. 거짓말을 했다면, 어떤 상황이라도 변명의 여지는 없다
V	Q5. 사람은 기본적으로 욕망이 있으며, 기회가 있으면 그 본성이 드러난다 Q8. 대부분의 사람들은 강제하지 않으면 열심히 일하지 않는다 Q18. 남들보다 앞서기 위해서는 꼼수를 써야 한다 Q13. 일반인과 범죄자의 가장 큰 차이는 범죄사실이 발각될 만큼 멍청한 지 여부에 있다 Q20. 대부분의 사람들은 자산을 잃은 사실보다 부모를 잃은 사실을 더 쉽게 잊는다	Q11. 세계를 리드하는 사람은 청렴하고 도덕적이다 Q14. 대부분의 사람들은 용감하다 Q17. "매 초마다 나쁜 놈이 태어난다"는 말은 틀렸다 Q4. 대부분의 사람들은 기본적으로 착하고 친절하다
M	Q19. 안락사는 인정되어야 한다	Q9. 부도덕하고 유명한 것보다 겸손하고 정직한 것이 모든 면에서 더 낫다

T : tactics 전략적 사고 **V** : views 관점 **M** : morality 도덕성

인적사항 설문지 구성

< 인적사항 설문지 예시 >

10개 성격판단 설문지

나는 외향적이고, 열정적인 편이다

1점 ---------- 매우 반대
2점 ---------- 조금 반대
3점 ---------- 쬐끔 반대
4점 ---------- 뭣도 아님
5점 ---------- 쬐끔 동의
6점 ---------- 조금 동의
7점 ---------- 매우 동의

16개 신뢰도판단 설문지

나는 다음 단어의 뜻을 알고있다

< pastiche >

1 ---------- 알고 있다
0 ---------- 모른다

13개 인적사항 설문지

대학에 다닌다면, (다녔다면) 전공은 무엇인가?

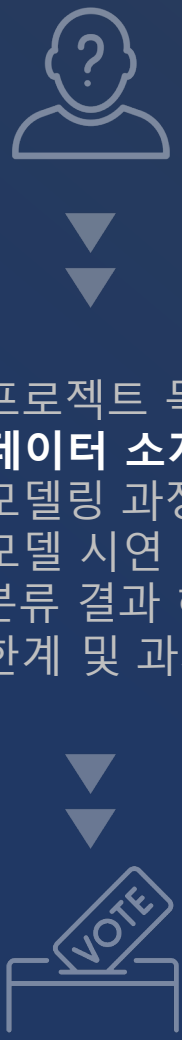
주관식 답변 방식
ex) '심리학', '토목공학' 등

#2. DATA INFO

EDA

COLUMNS in DATASET

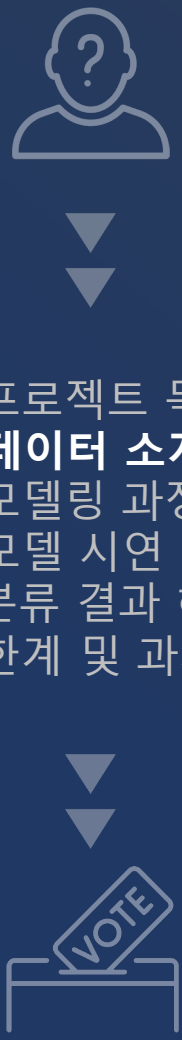
EDA – 컬럼 구성



1	Q1A~Q20A	**마키아벨리즘 테스트 질문지에 대한 답변** (int) 1~5 : 매우 반대 … 매우 동의
2	Q1I ~ Q20I	**마키아벨리즘 테스트 질문지 순서** (int) 1~20 : 테스트 과정에서의 각 질문지 순서
3	Q1E ~ Q20E	**마키아벨리즘 테스트 질문지에 대한 답변 시간** (float) 0~??? : 밀리초
4	TIPI1 ~ TIPI10	**참여자의 성격 판단 질문지에 대한 답변** (int) 1~7 : 매우 반대 … 매우 동의

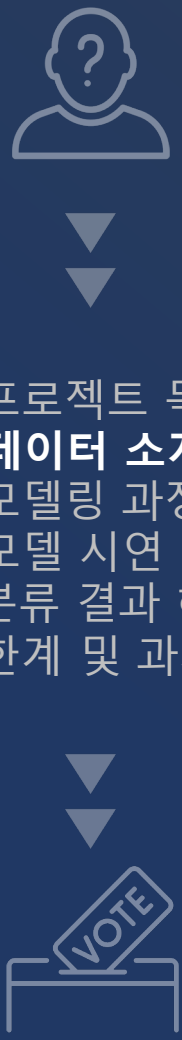
이상치 존재

EDA – 컬럼 구성



5	VCL1 ~ VCL16	**참여자 답변 신뢰도 측정 질문지 답변** (int) 0~1 : 단어 뜻을 안다, 모른다
6	education	**참여자의 학력** (int) 1~4 : 중졸, 고졸, 학사, 석박사
7	urban	**참여자의 고향 설문지 답변** (int) 1~3 : 농어촌, 중소도시, 대도시
8	gender	**참여자의 성별** (int) 1~3 : 남, 여, 기타

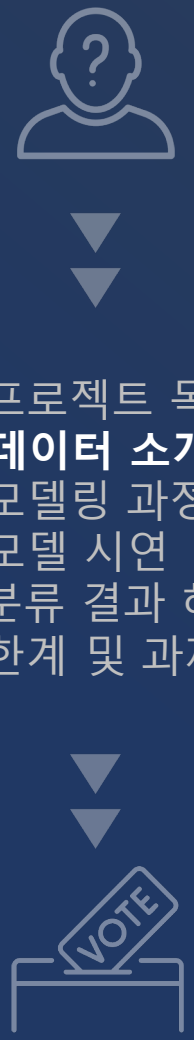
EDA – 컬럼 구성



9	engnat	**참여자의 모국어** (int) 1~2 : 영어, 비영어
10	age	**참여자의 나이** (int) 13~??? : 주관식 답변
11	hand	**참여자의 주사용 손** (int) 1~3 : 오른손, 왼손, 양손
12	religion	**참여자의 종교** (int) 1~12 : 불신론자, 크리스찬, 불교도..

이상치 존재

EDA – 컬럼 구성



13	orientation	**참여자의 성정체성** (int) 1~5 : 이성애자, 동성애자, 양성애자..
14	race	**참여자의 인종** (int) 10~70 : 10단위 수 : 백인, 흑인, 아시아인..
15	voted	**참여자의 과거 국가 선거 투표 참여 여부** (int) 1~2 : 참여, 불참
16	married	**참여자의 결혼 상태** (int) 1~3 : 미혼, 최근 결혼, 과거 결혼

label column

EDA – 컬럼 구성

17	familysize	**참여자의 형제자매 수** (int) 0~??? : 주관식 답변	이상치 존재
18	major	**참여자의 학부 전공** (str) ??? : 주관식 답변	군집화 불가
19	country	**참여자의 인터넷 서버 지역** (str) US~KR : 알파벳 2자리 국가 코드	*자동수집 정보*
20	screenw~h	**참여자의 스크린 사이즈** (float) 0~??? : 스크린 가로/세로 사이즈	*자동수집 정보*
21	intro~...elapse	**참여자의 테스트 구간별 답변 시간** (float) 0~??? : 밀리초	*자동수집 정보* 이상치 존재

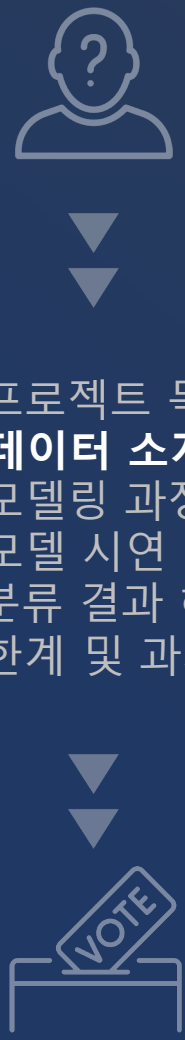
#2. DATA INFO

EDA

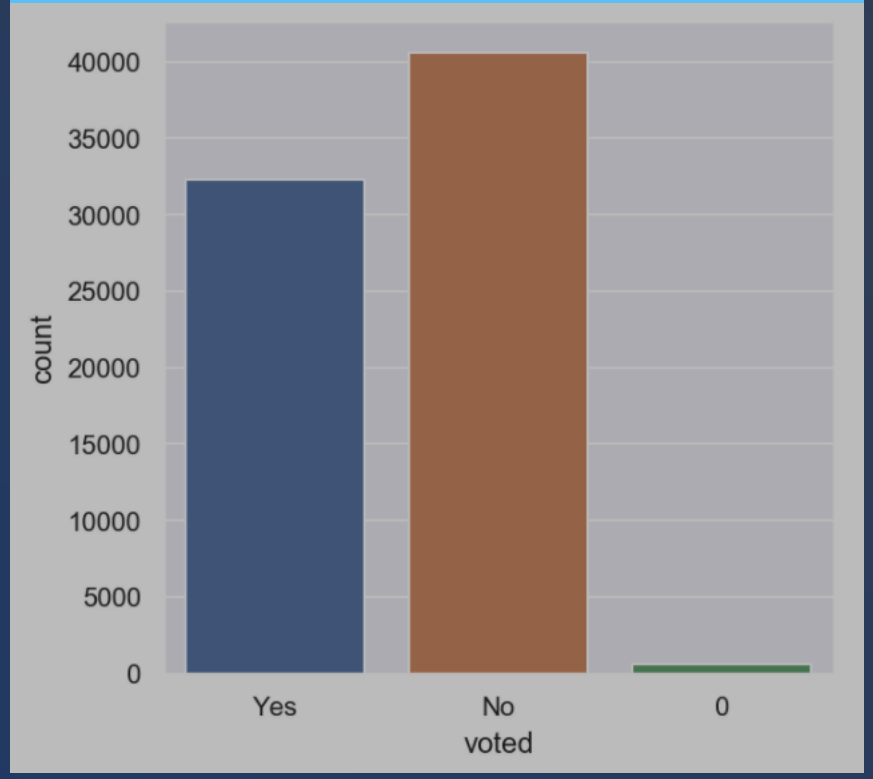
DATA ASSOCIATION

EDA

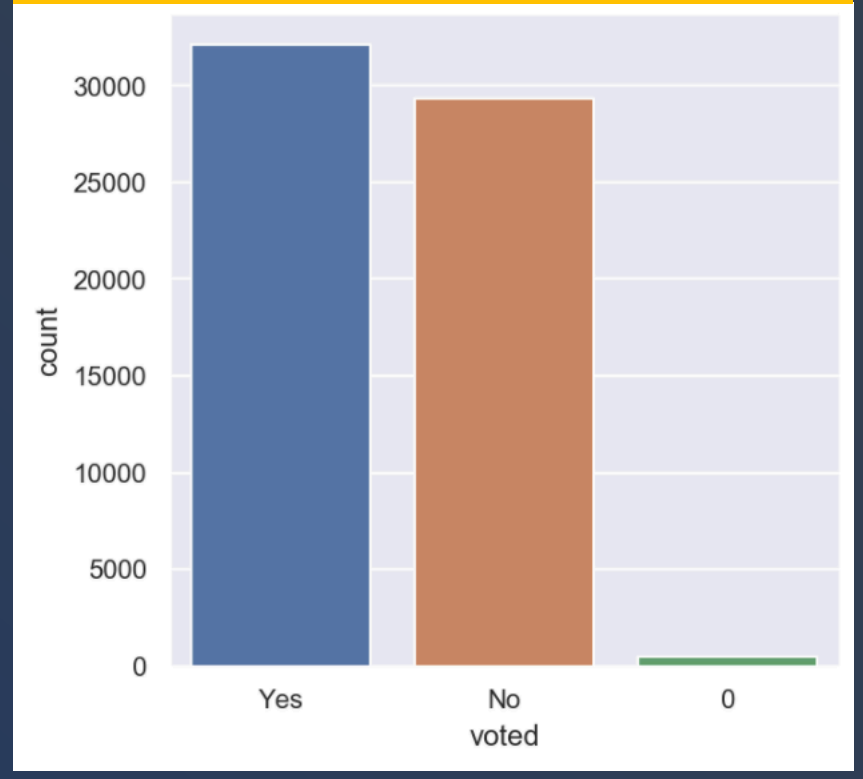
voted & VOTED 전체 투표참여 현황



투표 불가능 연령 포함



ONLY 투표가능 연령



'voted' 분류를 위해 18세 미만, 선거 불가능연령대 참여자 데이터 제거

#2. DATA INFO

모든 사람은 어느정도 마키아벨리안이다

평균 약 70점

\+ ▲

마키아벨리즘 성향이 높은 사람

낮은 친화력, 높은 성실성

사람들과 소통할 때
계산적이고 **신중**하게 접근하는 경향

마키아벨리즘 성향이 낮은 사람

\- ▼

비교적 수동적, 순응적

사람들과 소통할 때
개인적이며 **감정을 이입**하여 접근하는 경향

#2. DATA INFO

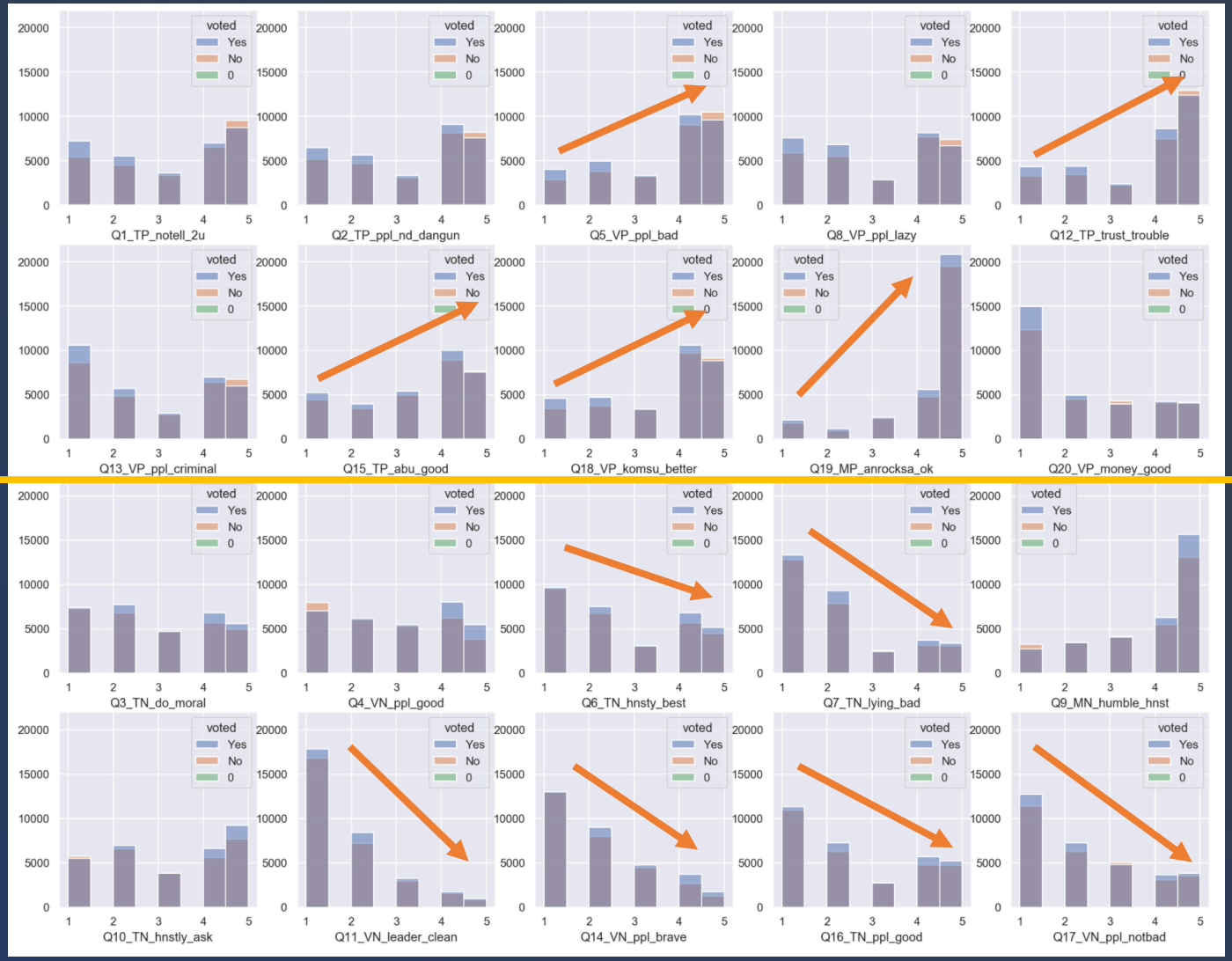
EDA

voted &

Q1A~Q20A 마키아벨리즘 답변

+ 마키아벨리즘 지지성향 질문

- 마키아벨리즘 반대성향 질문



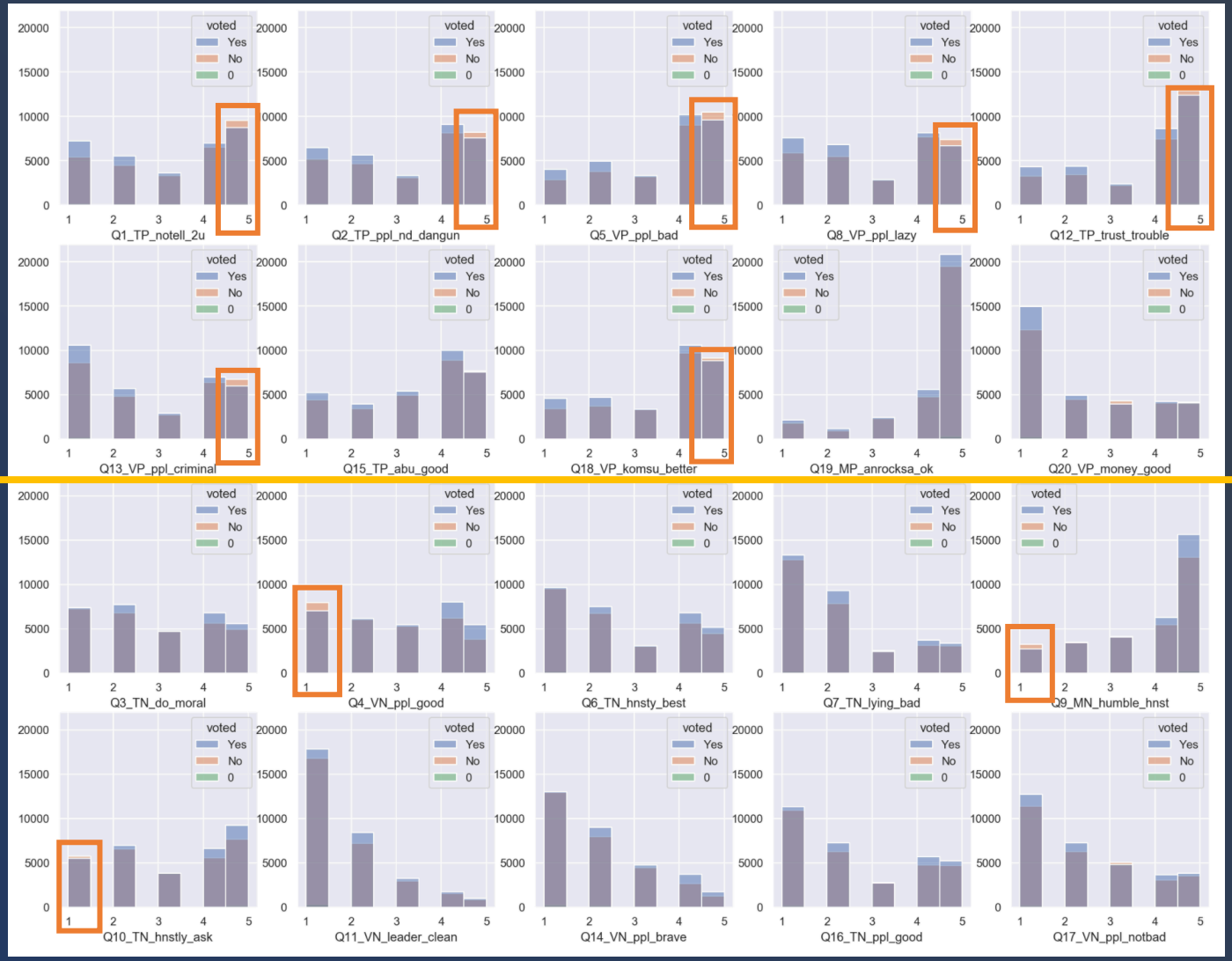
#2. DATA INFO

EDA voted &

Q1A~Q20A 마키아벨리즘 답변

+ 마키아벨리즘 지지성향 질문

- 마키아벨리즘 반대성향 질문



#2. DATA INFO

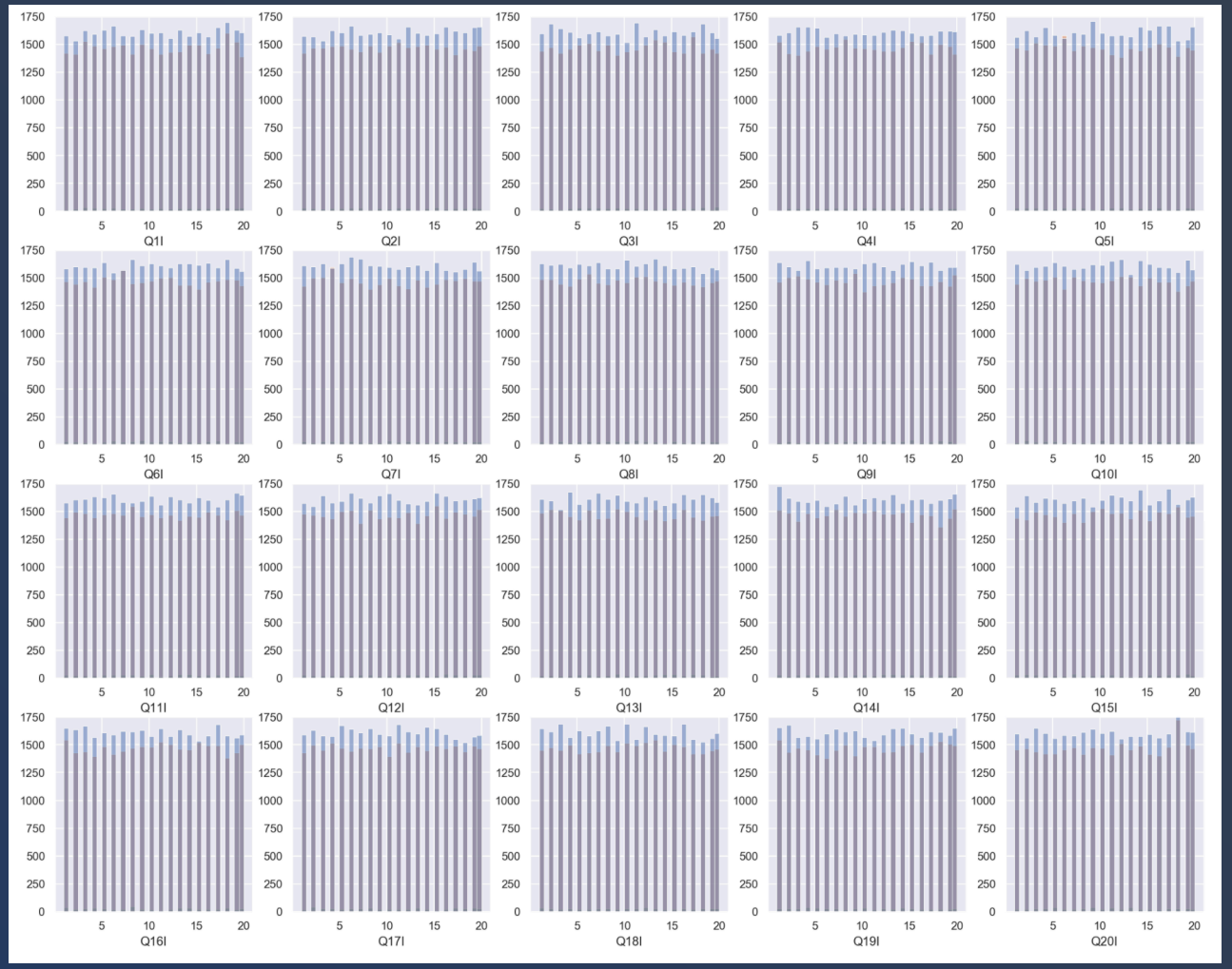
EDA
voted &

Q1I~Q20I 마키아벨리즘 질문지 순서

QI
: the position of that item in the survey

II

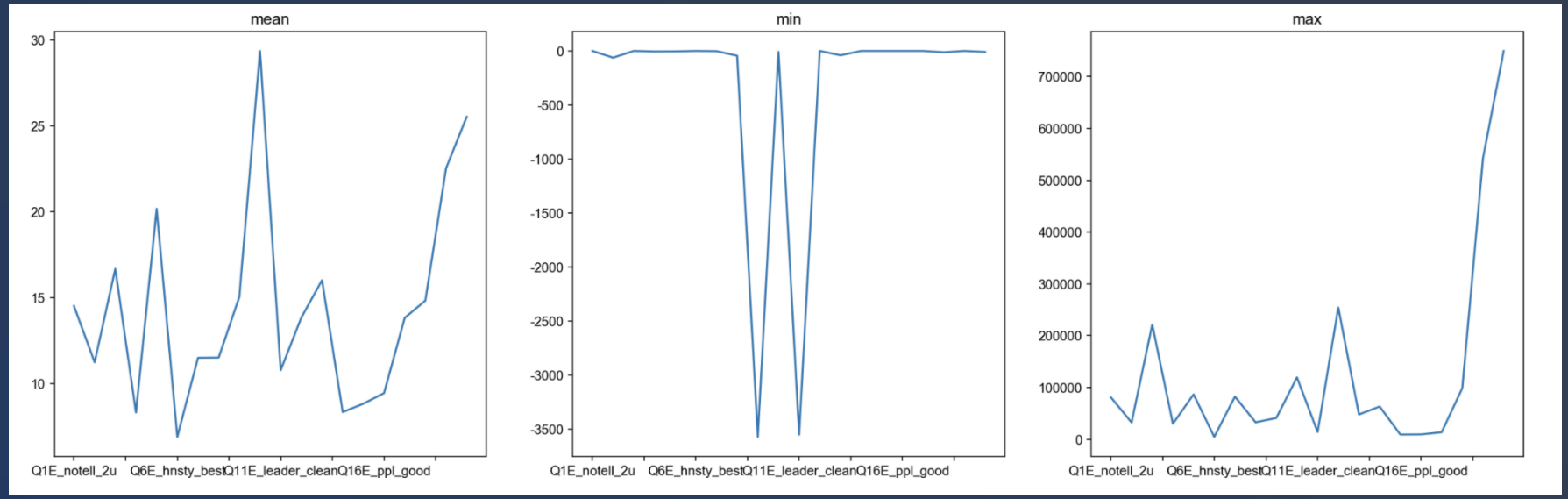
성향 테스트 결과 데이터의 일관성 유지를 위한 심리검사 방법론적 장치이므로 **연관 없는 정보**



#2. DATA INFO

EDA voted &

Q1E~Q20E 마키아벨리즘 답변 시간

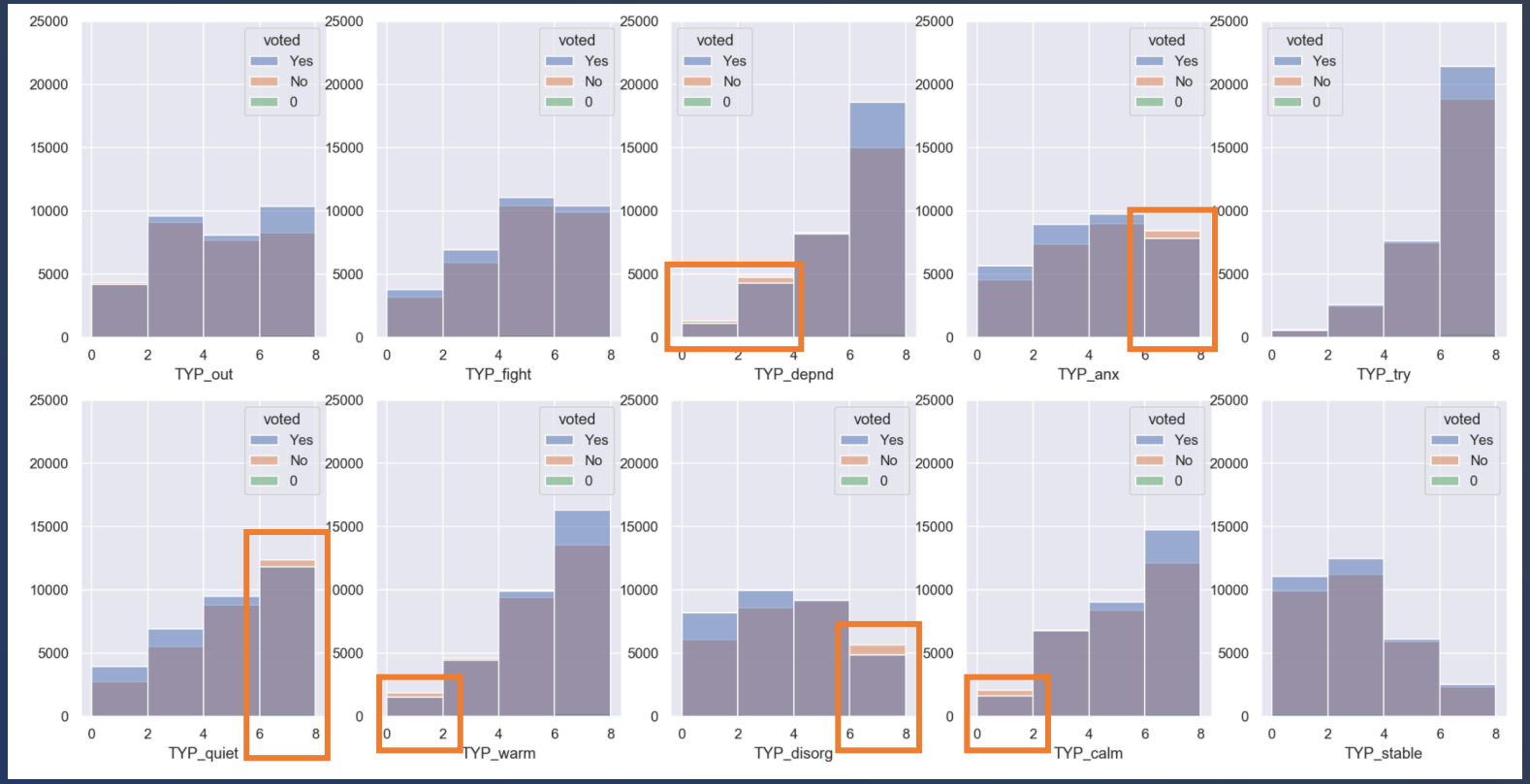


평균 답변 시간	약 14초		
평균 이상	+ 5개, - 3개	평균 이하	+ 5개, -7개
최장 시간	Q10 (가장 긴 문장)	최단 시간	Q6 (가장 짧은 문장)

#2. DATA INFO

EDA

voted & TIPI1~ TIPI20 참여자의 성격



독립적, 다혈질, 과묵한, 냉혈한인, 부주의한, 불안정한 성격

투표 불참 비율 높음

#2. DATA INFO

EDA voted &

VCL1~VCL16 단어 인지 여부 확인 질문

<VCL 문제>
" 아래 단어의 뜻을 아십니까?"

16개 질문지 중 3개는 허구로 만들어낸 가상의 단어가 포함됨
-> cuivocal, florted, verdid

||

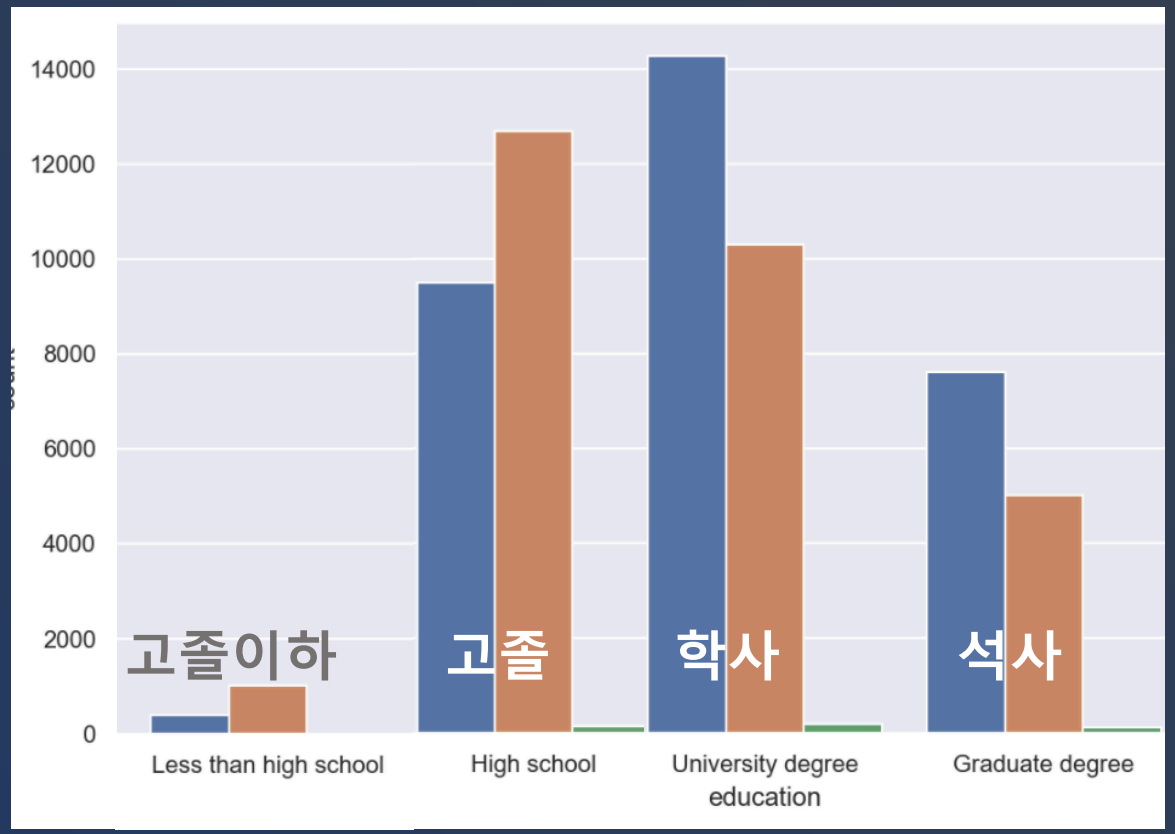
지어낸 단어마저 "알고있다"라고 **거짓 답변하는 참여자 확인 목적**

거짓답변을 한 참여자의 투표여부에는 차이가 미미. **BUT 대다수가 모르는 단어를 안다고 답변한 참여자의 투표 참여율이 더 높음**

#2. DATA INFO

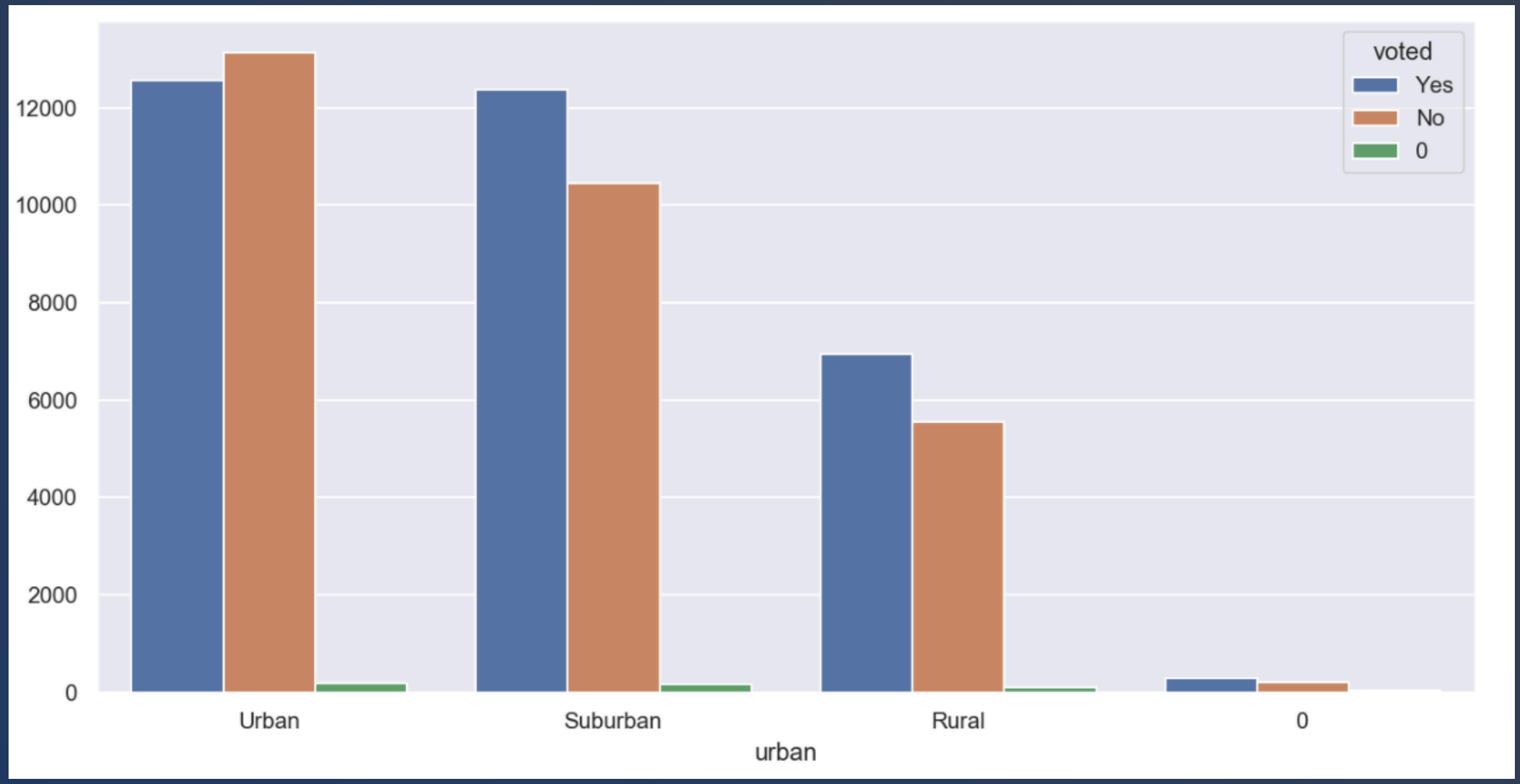
EDA

voted & education 참여자의 학력



고학력자의 투표참여 비율이 높음

EDA voted & urban 참여자의 고향

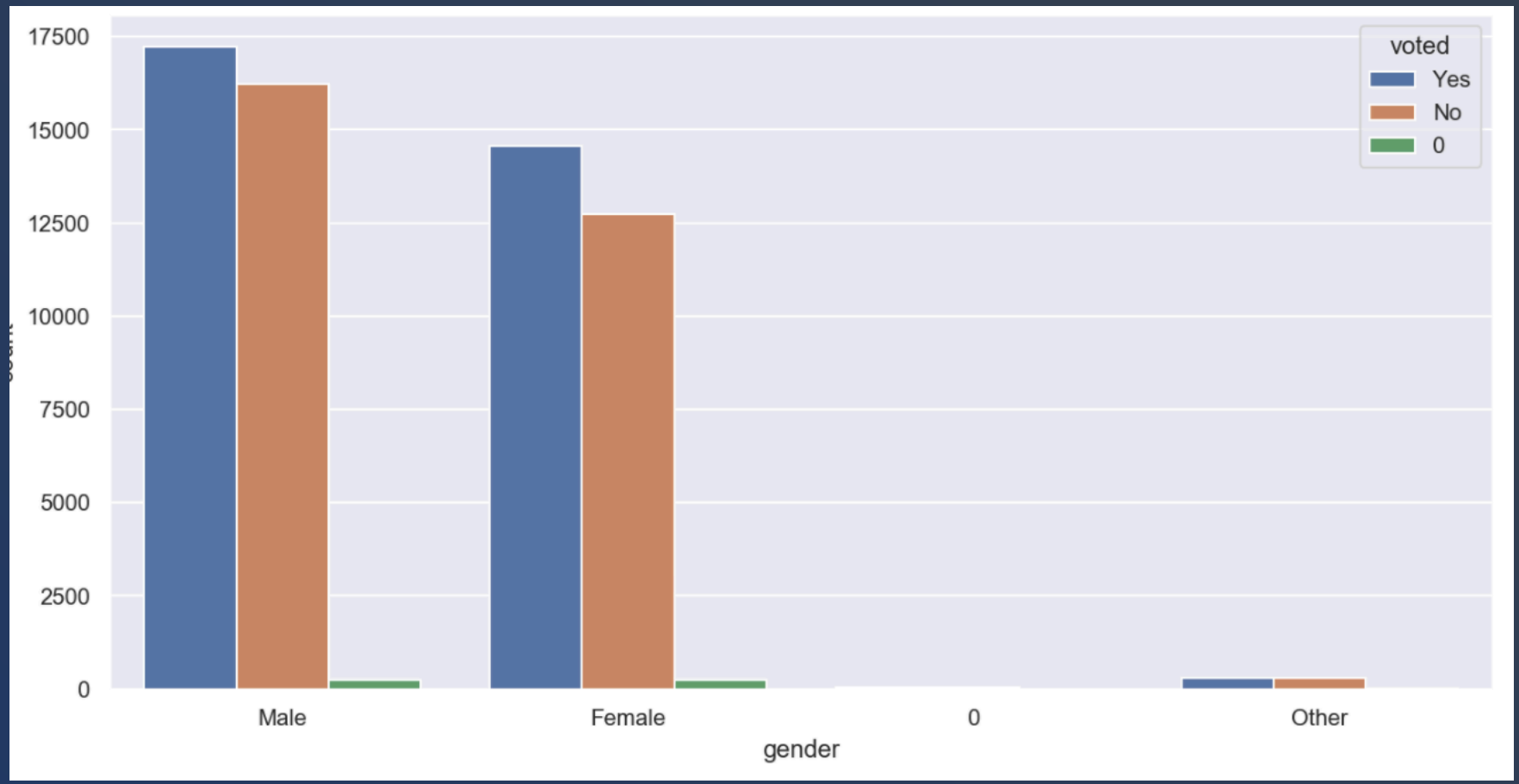


비도심지 출신일수록 투표참여 비율이 높음

#2. DATA INFO

EDA

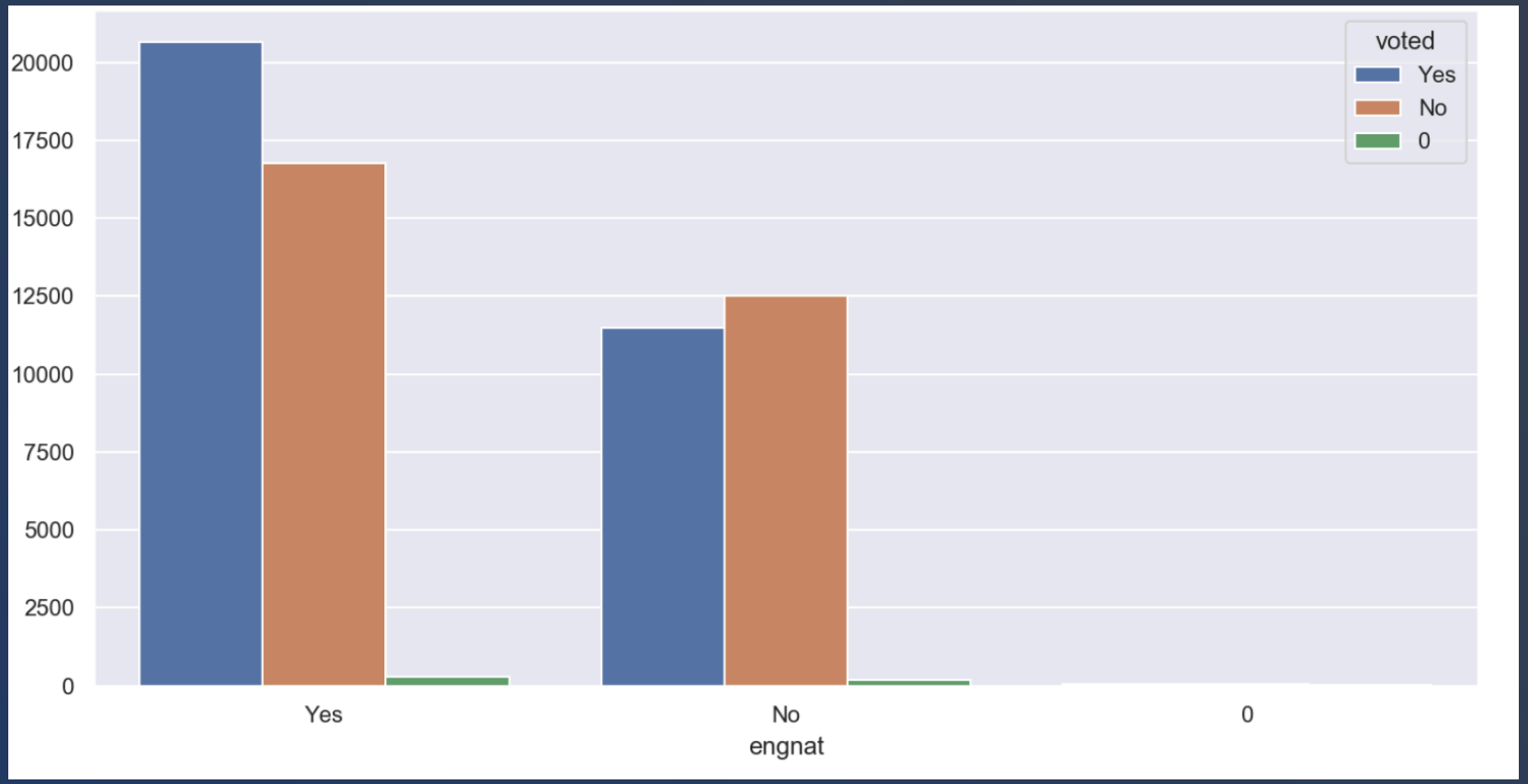
voted & gender 참여자의 성별



여성의 투표참여 비율이 높음

#2. DATA INFO

EDA voted & engnat 참여자의 모국어

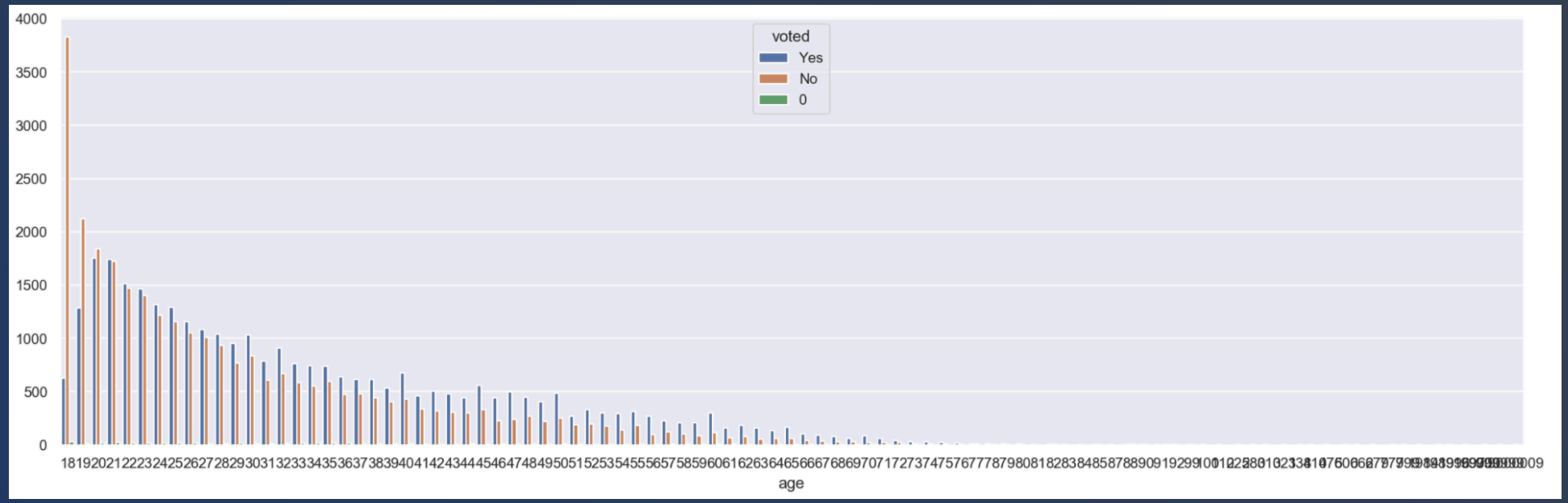


영어 원어민의 투표참여 비율이 높음

#2. DATA INFO

EDA

voted & age 참여자의 나이

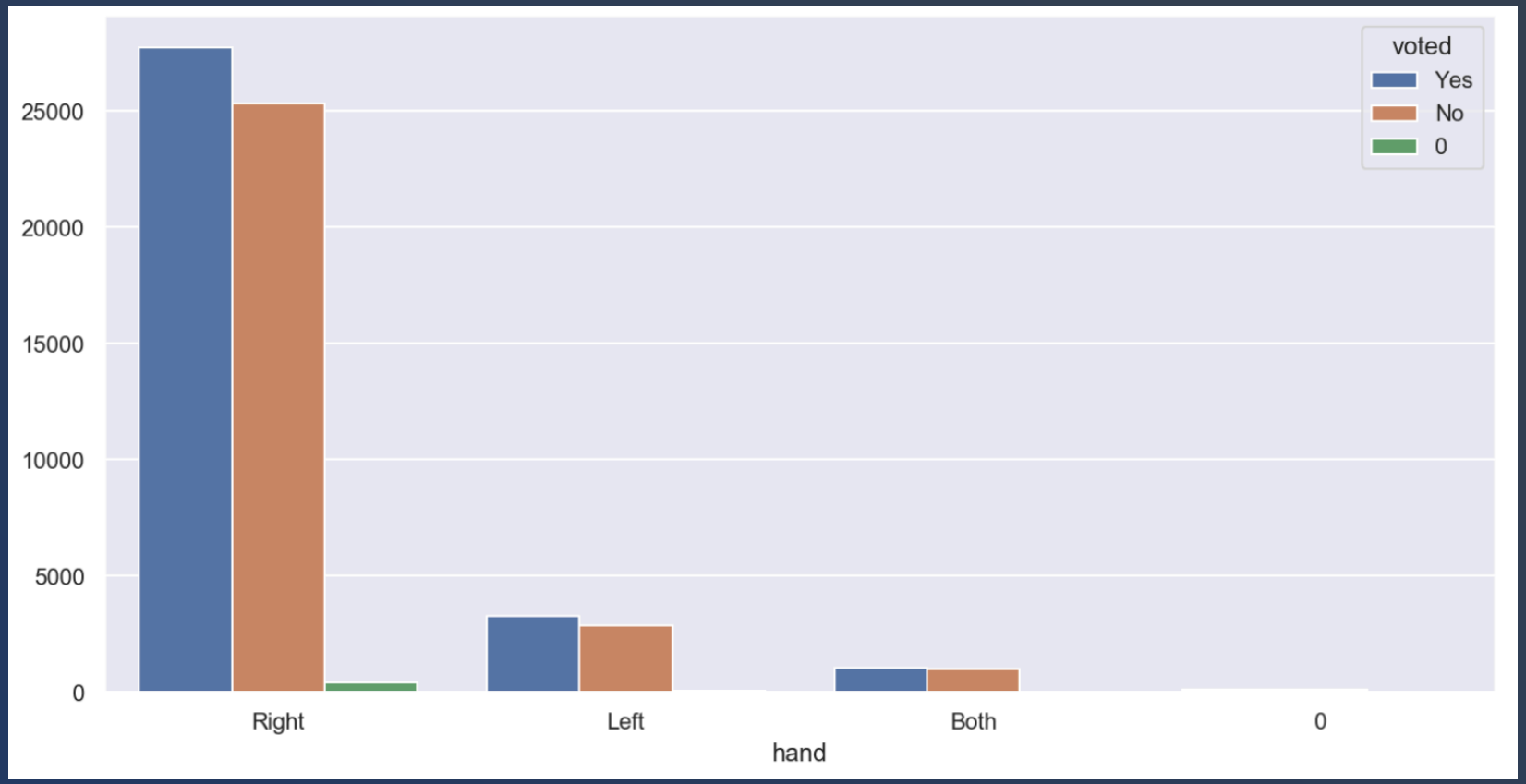


21세 이후 투표참여 비율이 상승하기 시작

#2. DATA INFO

EDA

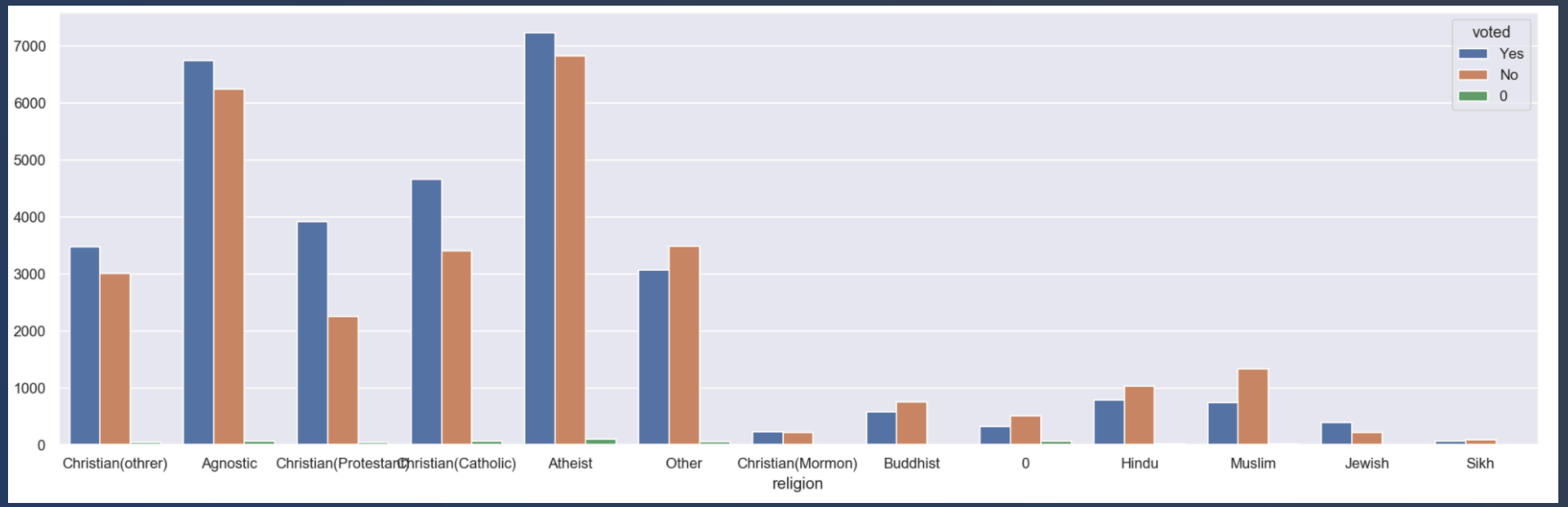
voted & hand 참여자의 주사용 손



왼손잡이 투표참여 비율이 좀더 높음

#2. DATA INFO

EDA voted & religion 참여자의 종교

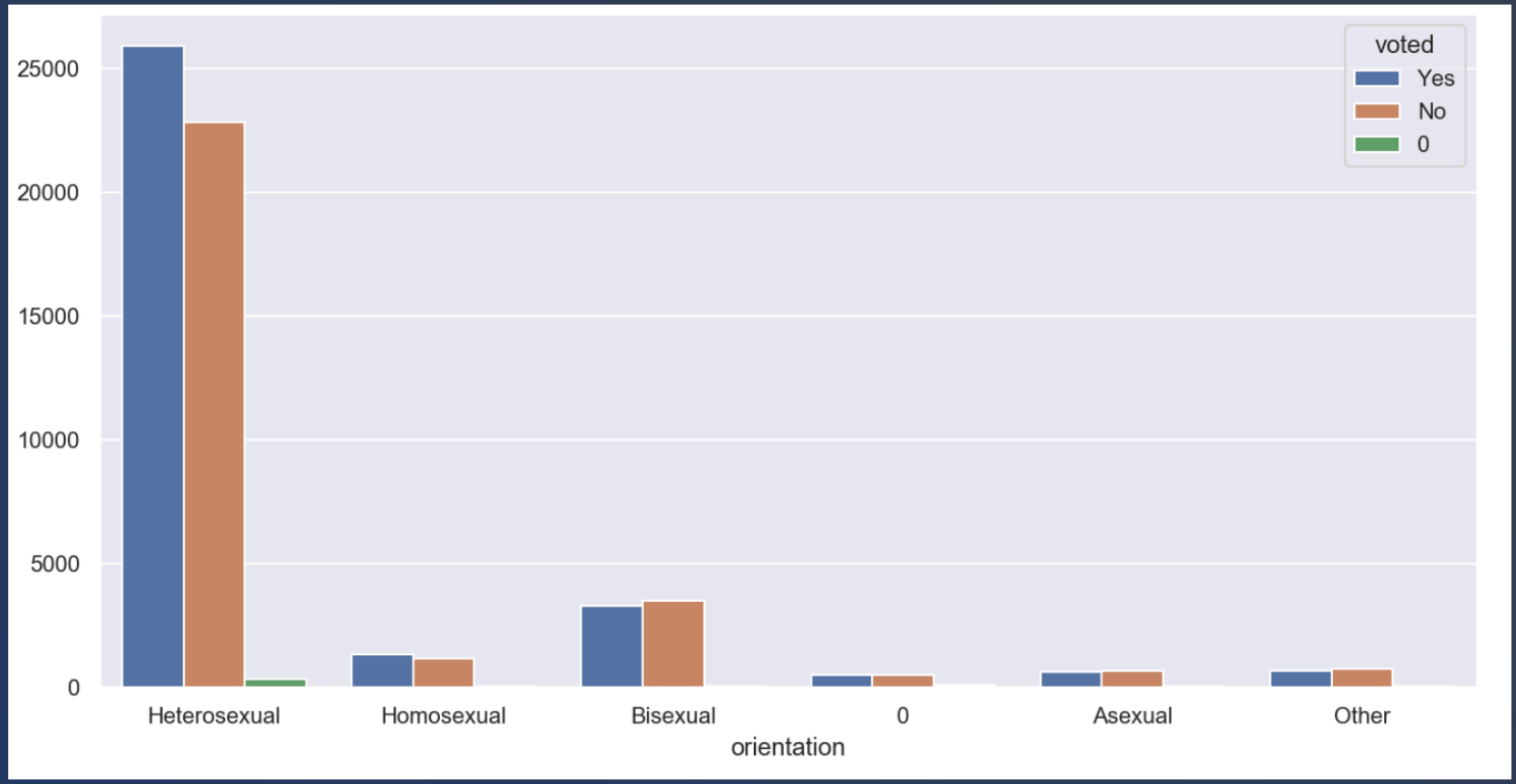


높은 투표참여 비율 기독교 정교, 카톨릭교

낮은 투표참여 비율 무슬림, 힌두교, 불교 등

#2. DATA INFO

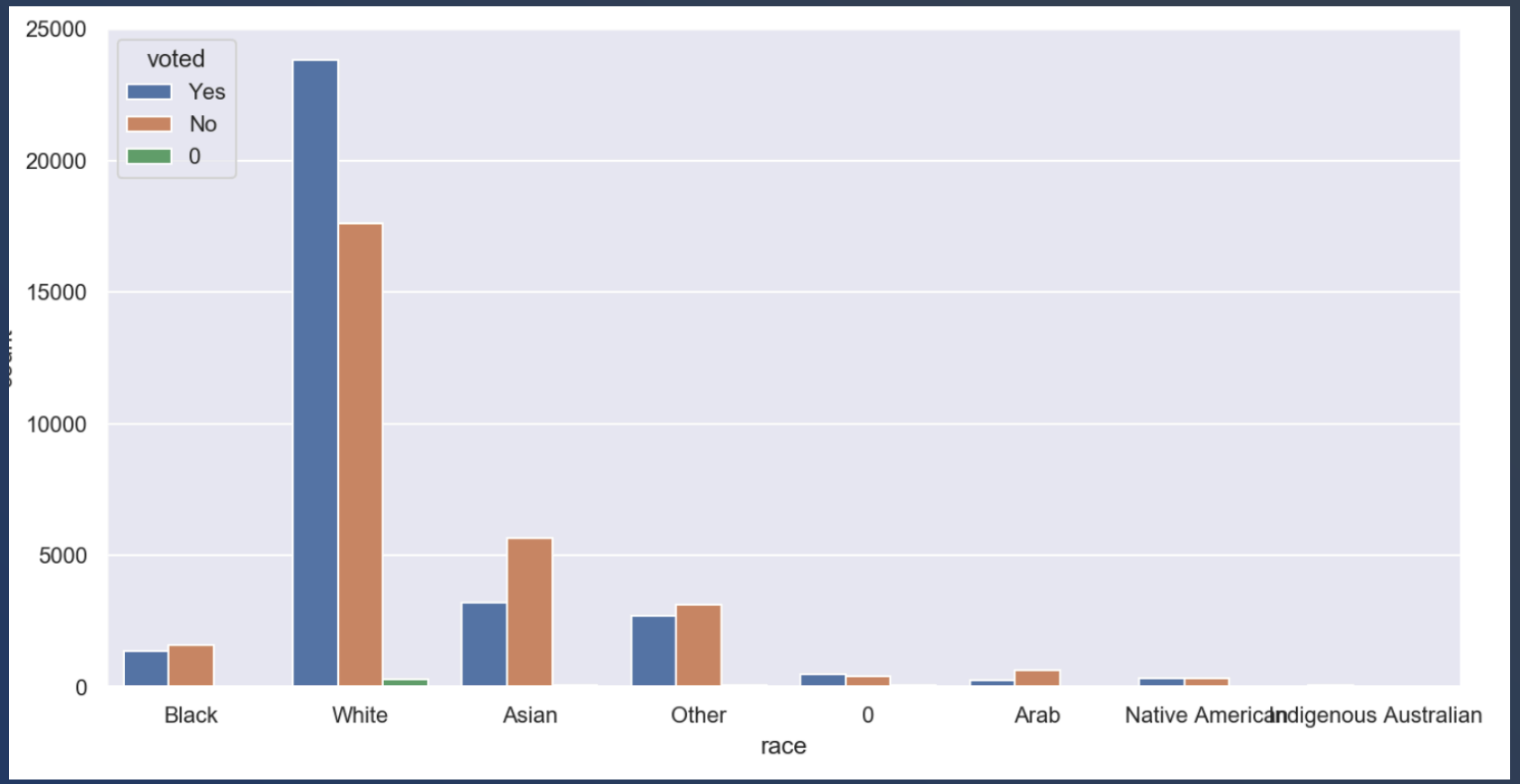
EDA voted & orientation 참여자의 성정체성



이성애자의 투표참여 비율이 더 높음

EDA voted & race 참여자의 인종

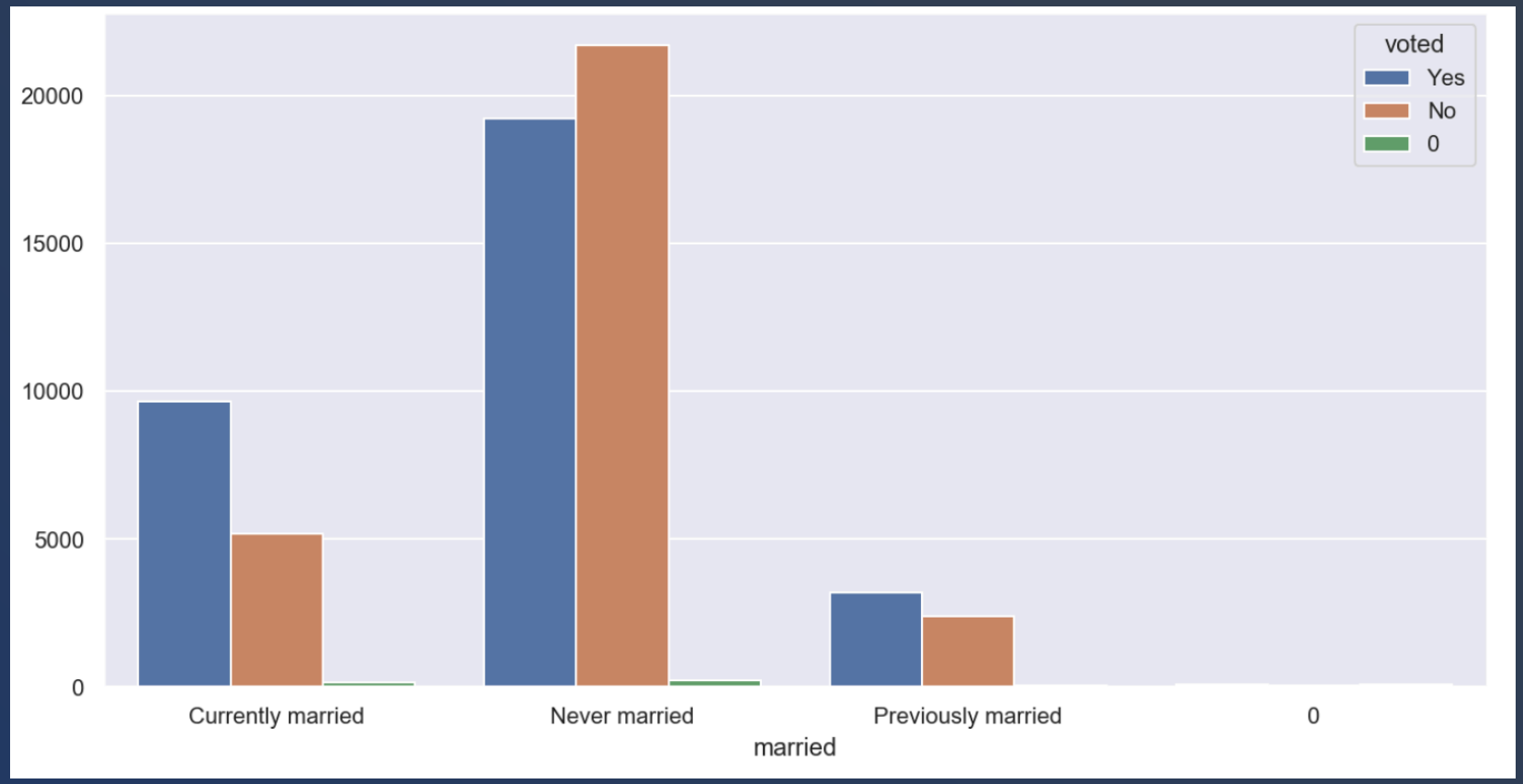
1. 프로젝트 목표
2. **데이터 소개**
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



타 인종에 비해 투표참여 비율이 높은 백인

#2. DATA INFO

EDA voted & married 참여자의 결혼상태



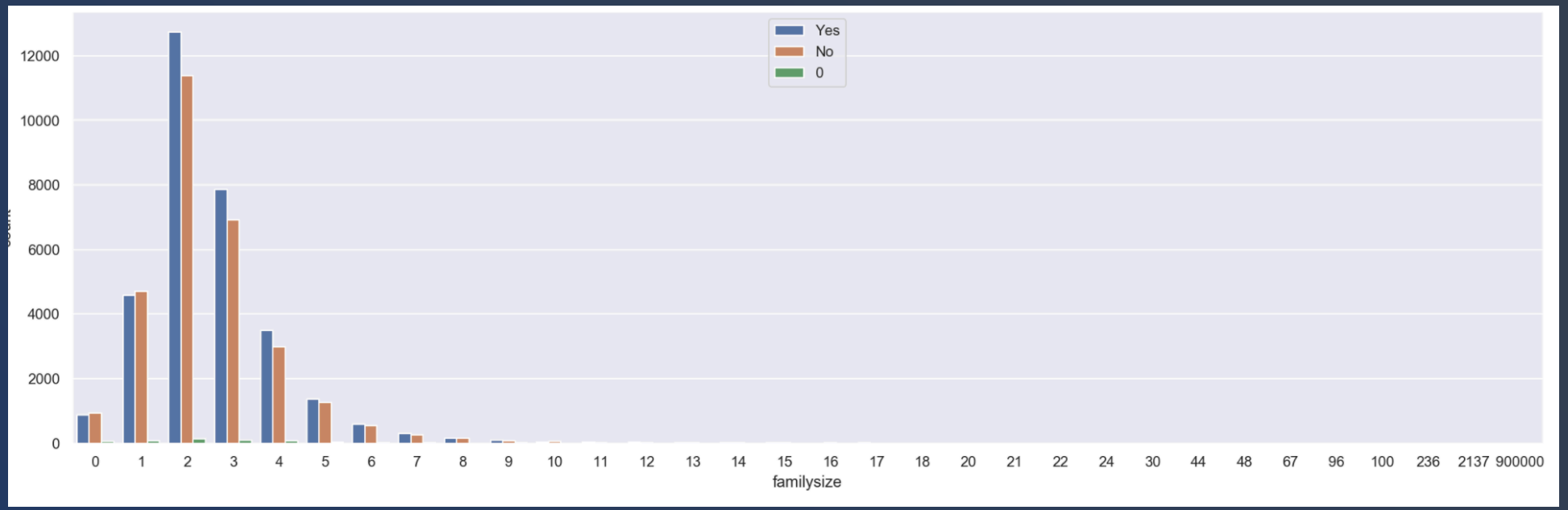
투표참여 비율이 낮은 미혼 참여자

#2. DATA INFO

EDA

voted &

familysize 참여자의 형제자매수



외동이거나 형제자매가 1명인 경우
투표참여 비율이 낮음

EDA voted & major 참여자의 학부 전공


```
psychology                                      3367
business                                        1988
engineering                                     1734
english                                         1501
computer science                                1435
                                                ...
politic scienze                                    1
chinese language & literature program;             1
radio television film                              1
training/education                                 1
journalism and mass media studies                  1
Name: major, Length: 6380, dtype: int64
```

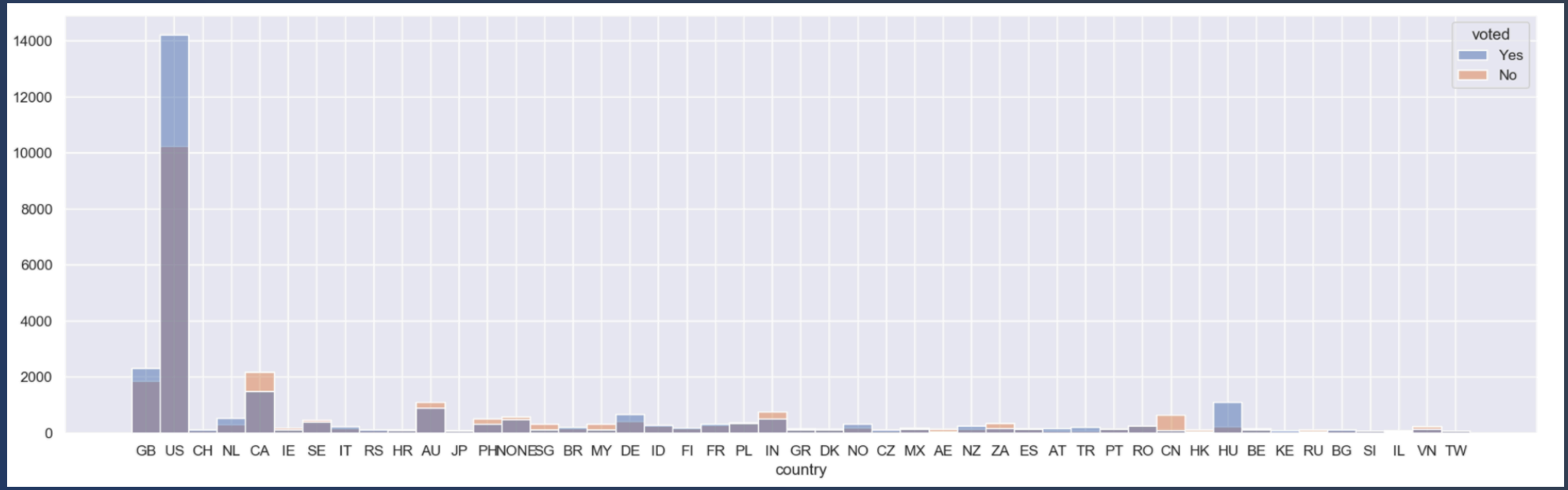
6380 종류의 전공, 너무나 다양한 데이터

#2. DATA INFO

EDA

voted & country 참여자의 서버 지역

184개 국가 데이터



높은 투표율 미국, 영국, 헝가리, 덴마크, 네덜란드, 노르웨이...

낮은 투표율 중국, 캐나다, 호주, 인도, 필리핀, 말레이시아, 싱가폴...

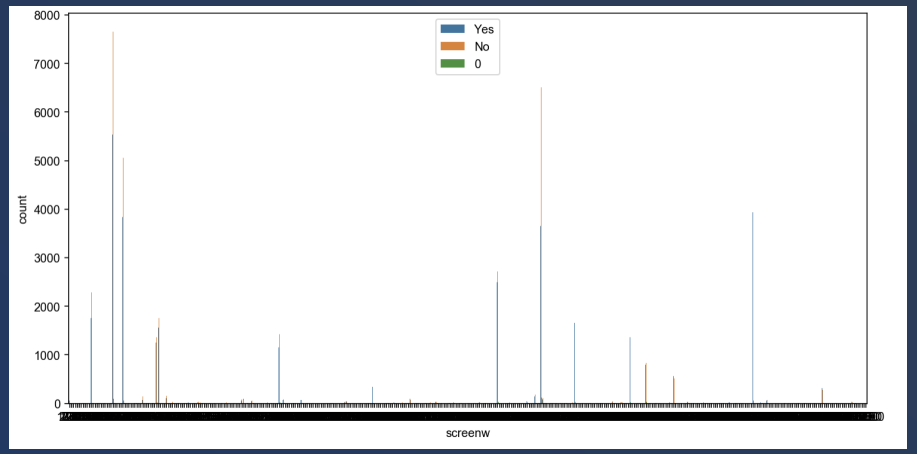
#2. DATA INFO

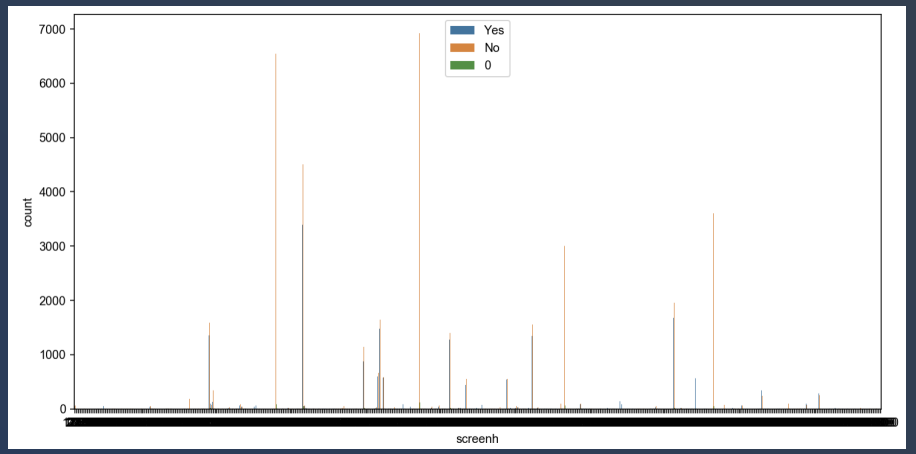
EDA

voted &

screenw~h 참여자의 스크린 사이즈

참여자 대부분 모바일 이용자





스크린 사이즈가 커질수록
투표참여 비율이 높아짐

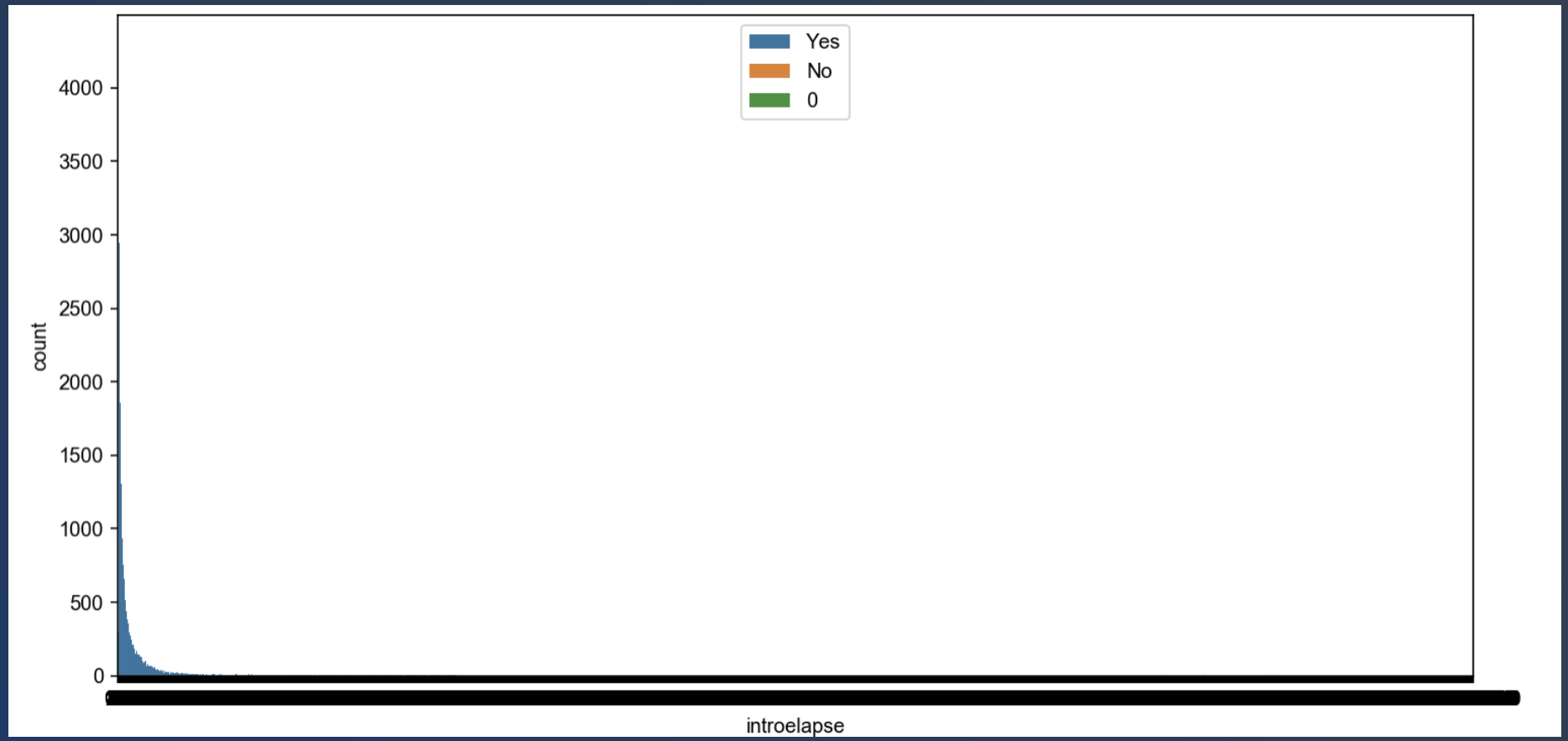
#2. DATA INFO

EDA

voted &

introelapse 참여자의 답변 시간

1. 프로젝트 목표
2. **데이터 소개**
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



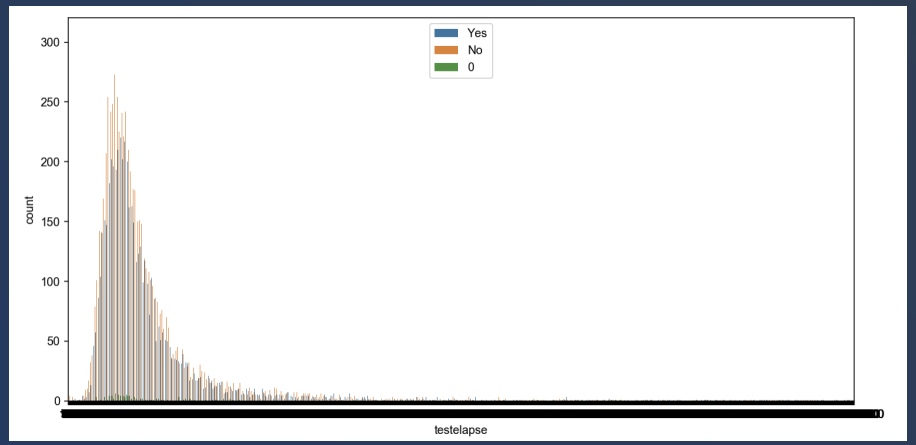
테스트 시작 전에서 머문시간 : 11초 (중위값)

#2. DATA INFO

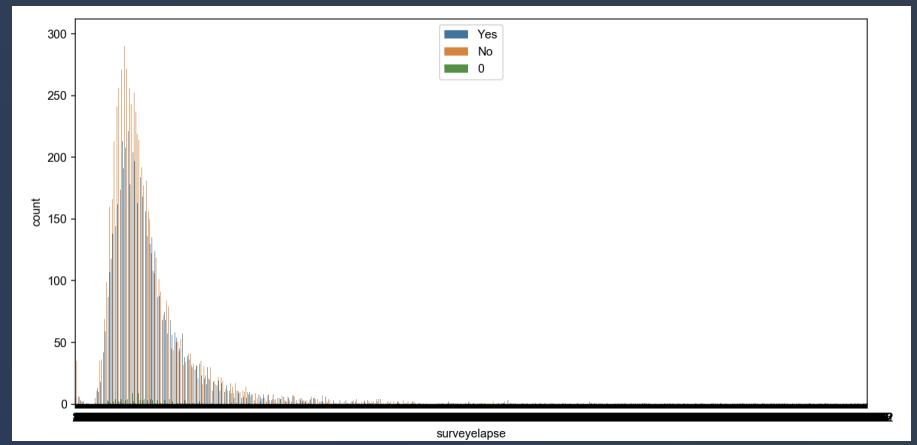
EDA

voted &

test~survey elapse 참여자의 답변 시간



'test' 파트에서 머문시간
3분 (중위값)

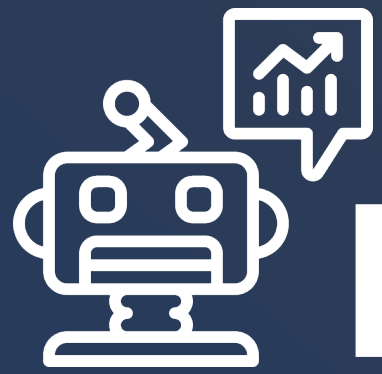


'survey' 파트에서 머문시간
3분 (중위값)

소요시간이 짧을수록
투표참여 비율이 낮아짐

Insights from EDA

어떤 컬럼이
분류 성능 향상에 도움이 될까요?



MODELING

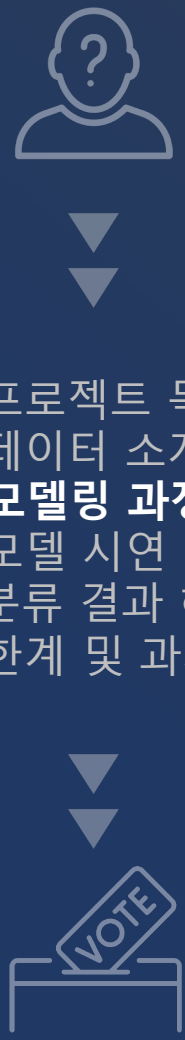
Review, Refine, Repeat

Insights from EDA

분류모델 개발을 위한 핵심과제

1	수치 데이터 이상치 처리	**age, familysize** (참여자 직접입력 데이터) **QEn, intro~surveyelapse** (자동수집 시간 데이터)
2	유의미한 범주데이터 선별	**country, race, TIPI1~10 등** (특정 범주에서의 voted 값이 특이한 데이터)
3	마키아벨리즘 관련 데이터에 가중치 부여	**Q1~20, QE1~20, + score, T, V, M** (마키아벨리즘 테스트 답변 및 답변시간 데이터)

Insights from EDA



분류모델 개발을 위한 핵심과제

1	수치 데이터 이상치 처리	scaling
2	유의미한 범주데이터 선별	feature selection
3	마키아벨리즘 관련 데이터에 가중치 부여	feature addition

Insights from EDA

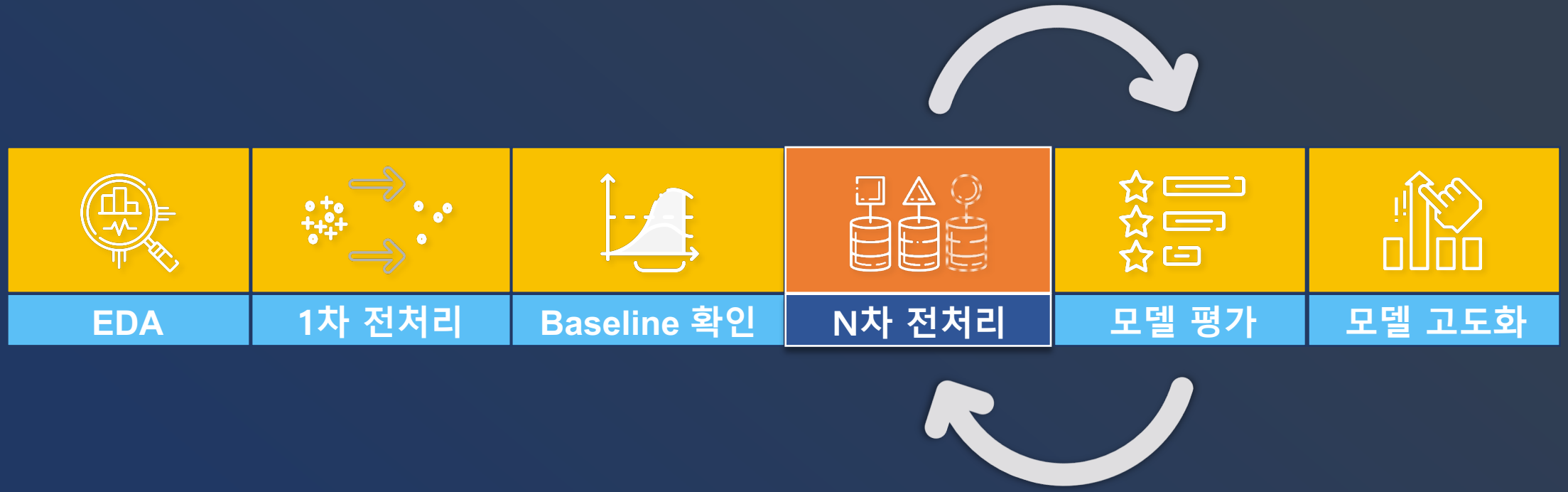
분류모델 개발을 위한 핵심과제

1	수치 데이터 이상치 처리	age, familysize (참여자 직접입력 데이터) QEn, intro~surveyelapse (자동수집 시간 데이터)
2	유의미한 범주데이터 선별	[illegible]try, [illegible]c, TIP[illegible]10 등 (특정 범주[illegible]서 [illegible]ed [illegible] 특이한 데이터)
3	마키아벨리즘 관련 데이터에 가중치 부여	Q1~20, QE1~20 (마키아벨리즘 테스트 답변 및 답변시간 데이터)

전처리

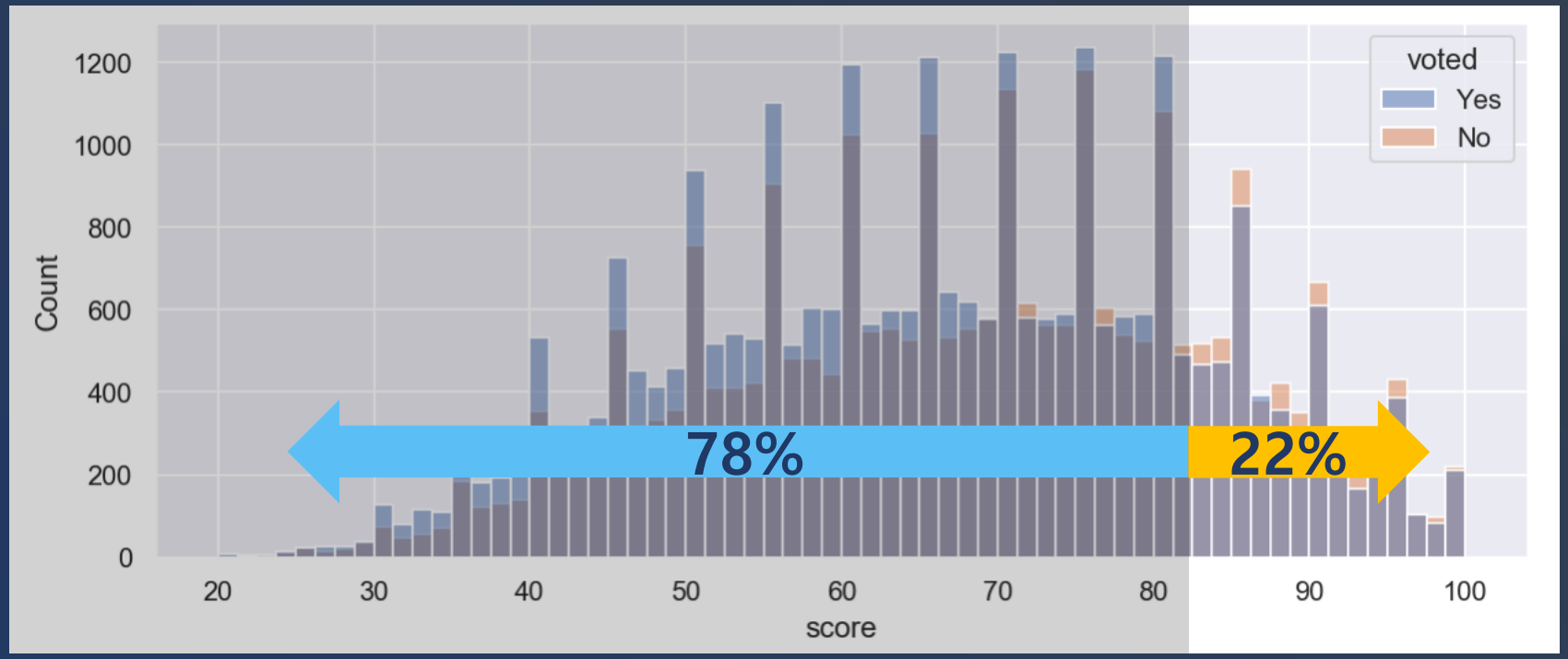
#3. MODELING

분류모델 개발 과정



1	0, null 제거	len(0, null data) == 8,920
2	선거가능연령 기준 이하 제거	len(age<18) == 11,521 - 전세계 기준 선거가능연령 : 18세
3	마키아벨리즘 점수 합산 데이터 추가	score = 지지 경향 답변 점수의 합 + 반대 경향 답변 점수의 역변환 값 (6 - 점수)의 합
	마키아척도별 점수 합산 데이터 추가	T, V, M = 각 척도별 점수를 score 산출 방식으로 합산
4	밀리초 단위 환산	round(밀리초*0.001)

score & voted 관계



마키아벨리즘 성향이 강할수록 투표에 참여하지 않음

1차 전처리 ▶ 성과 미미 . .

(최대 성능 모델의 AUC값 기준)

Before (RAW)

	accuracy	AUC	precision	recall	f1
Ada	0.648592	0.644708	0.642712	0.734684	0.685628
GBC	0.647220	0.643093	0.640252	0.738707	0.685965
XGB	0.632212	0.629347	0.634453	0.695699	0.663666
LGBM	0.655209	0.651630	0.649966	0.734530	0.689665

After (1차 전처리)

	accuracy	AUC	precision	recall	f1
Ada	0.650402	0.644314	0.641925	0.759590	0.695818
GBC	0.656159	0.649482	0.643907	0.775907	0.703771
XGB	0.648209	0.644031	0.648809	0.723138	0.683960
LGBM	0.657255	0.651714	0.649821	0.756639	0.699174

0.6516 → +▲ **0.0001** → 0.6517

N차 전처리

1차 Feature Importance 기준 Feature Selection

2차 수치형데이터 Robust Scaling

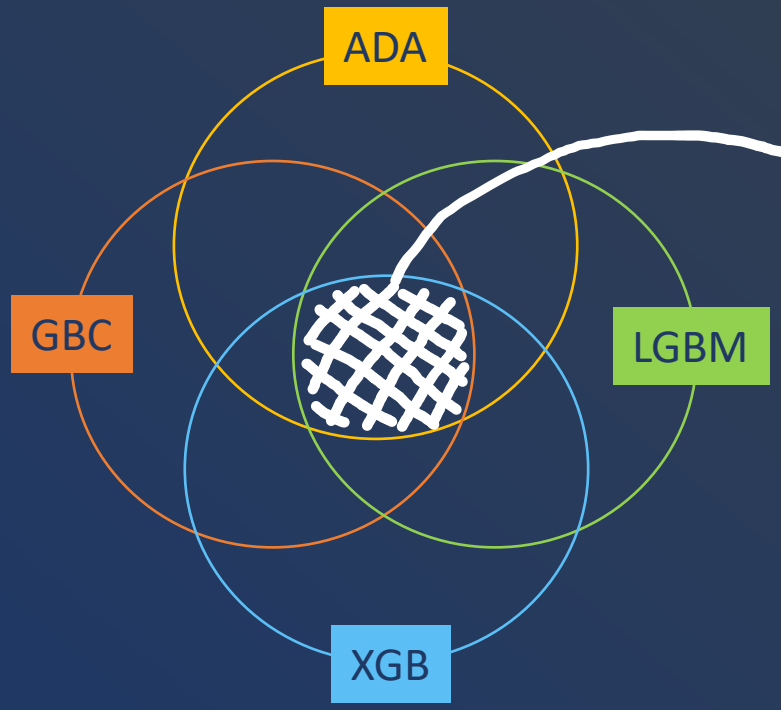
3차 'score_voted' score 연관 Feature Addition

4차 Feature Selection DATASET + Feature Addition DATASET

① Feature Importance 기준 컬럼 선별 ② 수치형데이터 Scaling ③ 'score_voted' 컬럼생성 ④ Feature 선별 + 추가

목표 주요 모델 4개에서 공통으로 Feature Importance가 0인 컬럼 제거

model.feature_importance == 0



'country_XX'

'country_AF',
'country_SD',
'country_PS',
'country_PL',
'country_BQ',
'country_TL',
'country_GP',
'country_GU',
⋮

'VCL_XX'

'VCL3_n_know',
'VCL2_n_know',
'VCL4_n_know',
'VCL16_n_know'
'VCL5_n_know',
'VCL14_n_know'
'VCL15_n_know'
'VCL13_n_know'
⋮

'race_XX'

'race_Indigenous

'hand_XX'

'hand_Both'

'religion_XX'

'religion_Sikh'

len(ada_gbc_xgb_lgbm)

>> **118**

① Feature Importance 기준 컬럼 선별 ② 수치형데이터 Scaling ③ 'score_voted' 컬럼생성 ④ Feature 선별 + 추가

feature 정리 ▶ 성능 향상 !

(feature importance 값이 0인 컬럼 제거)

Before

	accuracy	AUC	precision	recall	f1
Ada	0.650402	0.644314	0.641925	0.759590	0.695818
GBC	0.656159	0.649482	0.643907	0.775907	0.703771
XGB	0.648209	0.644031	0.648809	0.723138	0.683960
LGBM	0.657255	0.651714	0.649821	0.756639	0.699174

After

	accuracy	AUC	precision	recall	f1
Ada	0.650402	0.644314	0.641925	0.759590	0.695818
GBC	0.656250	0.649578	0.643999	0.775907	0.703826
XGB	0.648209	0.644031	0.648809	0.723138	0.683960
LGBM	0.657255	0.651714	0.649821	0.756639	0.699174

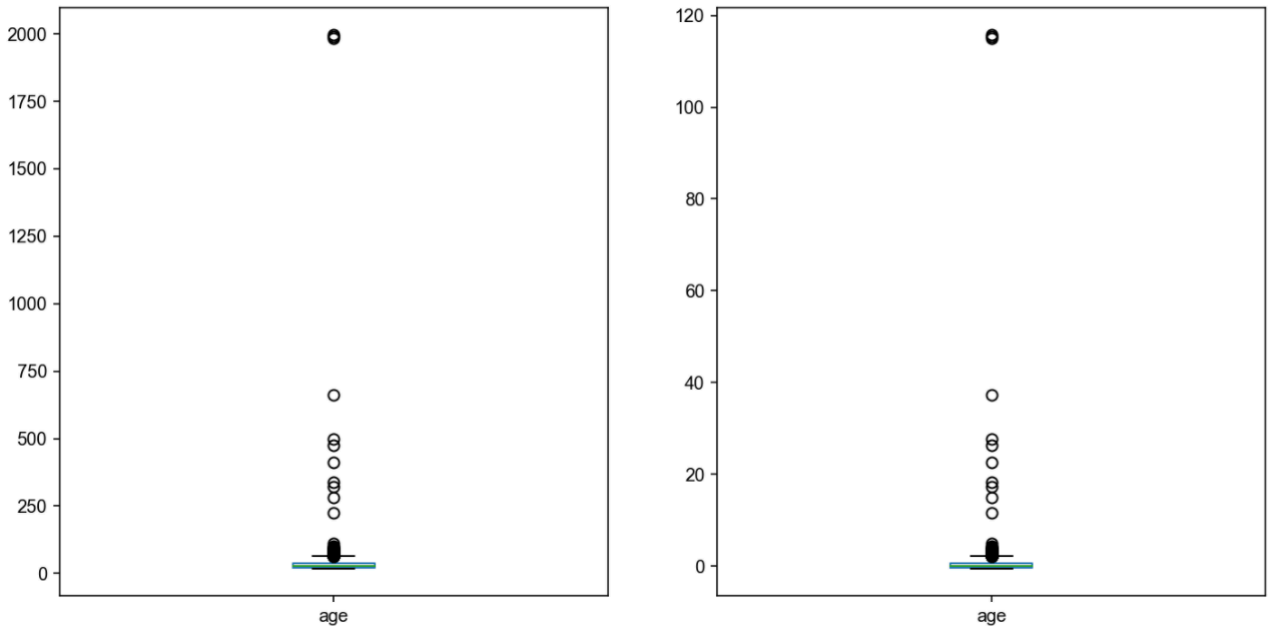
0.6494 ——— +▲ **0.0001** ———→ 0.6495

① Feature Importance 기준 컬럼 선별 ② 수치형데이터 Scaling ③ 'score_voted' 컬럼생성 ④ Feature 선별 + 추가

이상치를 포함한 수치형데이터에 Robust Scaling

```
df_18['age'].describe(), df_18_cp['age'].describe()
```

```
(count    54718.000000
 mean        31.739537
 std         19.666469
 min         18.000000
 25%         22.000000
 50%         28.000000
 75%         39.000000
 max       1997.000000
 Name: age, dtype: float64,
 count    54718.000000
 mean         0.219973
 std          1.156851
 min         -0.588235
 25%         -0.352941
 50%          0.000000
 75%          0.647059
 max        115.823529
 Name: age, dtype: float64)
```



Robust Scaling : 평균 ▶ 중위값 I 분산 ▶ IQR

① Feature Importance 기준 컬럼 선별 ② 수치형데이터 Scaling ③ 'score_voted' 컬럼생성 ④ Feature 선별 + 추가

수치 Scaling ▶ 성능 하락

(이상치가 있는 수치형 데이터 Robust Scaling)

Before

	accuracy	AUC	precision	recall	f1
Ada	0.650402	0.644314	0.641925	0.759590	0.695818
GBC	0.656159	0.649482	0.643907	0.775907	0.703771
XGB	0.648209	0.644031	0.648809	0.723138	0.683960
LGBM	0.657255	0.651714	0.649821	0.756639	0.699174

After

	accuracy	AUC	precision	recall	f1
Ada	0.648940	0.642983	0.641332	0.755772	0.693865
GBC	0.653417	0.647129	0.643440	0.766186	0.699469
XGB	0.645285	0.641486	0.648163	0.713418	0.679227
LGBM	0.652778	0.648680	0.653036	0.726263	0.687705

0.6494 — **-▼0.0023** → 0.6471

① Feature Importance 기준 컬럼 선별 ② 수치형데이터 Scaling ③ 'score_voted' 컬럼생성 ④ Feature 선별 + 추가

마키아벨리즘 성향 'score' 연관 컬럼 데이터 생성 공식

$$\text{'score\_voted'} = \frac{\text{(점수별) 투표}\mathbf{참여}\text{ 인원} - \text{(점수별) 투표}\mathbf{불참}\text{ 인원}}{\text{점수별 }\mathbf{전체}\text{ 인원}}$$

'score_voted'

score	voted
20.0	1.000000
21.0	0.000000
22.0	0.200000
23.0	-0.333333
24.0	0.130435

=

len(yes)

score	voted
20.0	5
21.0	1
22.0	3
23.0	2
24.0	13

-

len(no)

score	voted
20.0	0
21.0	1
22.0	2
23.0	4
24.0	10

/

len(yes+no)

score	voted
20.0	5
21.0	2
22.0	5
23.0	6
24.0	23

① Feature Importance 기준 컬럼 선별 ② 수치형데이터 Scaling ③ 'score_voted' 컬럼생성 ④ Feature 선별 + 추가

Label 연관 컬럼 추가 ▶ 성능 향상 !

('score_voted' 컬럼 추가)

Before

	accuracy	AUC	precision	recall	f1
Ada	0.537646	0.527545	0.546234	0.718799	0.620747
GBC	0.539382	0.527113	0.544832	0.759417	0.634472
XGB	0.529788	0.522482	0.543935	0.660823	0.596708
LGBM	0.527412	0.518493	0.540172	0.687381	0.604950

After

	accuracy	AUC	precision	recall	f1
Ada	0.537738	0.527680	0.546355	0.718104	0.620566
GBC	0.542032	0.531256	0.548492	0.735289	0.628300
XGB	0.527321	0.519422	0.541292	0.668981	0.598401
LGBM	0.537829	0.528377	0.547200	0.707342	0.617050

0.5184 ——— + ▲ **0.01** ———▶ 0.5283

#3. MODELING

① Feature Importance 기준 컬럼 선별 ② 수치형데이터 Scaling ③ 'score_voted' 컬럼생성 ④ Feature 선별 + 추가

+▲ ① Feature Importance 기준
Feature Selection

② 수치형데이터
Robust Scaling
gg...

+▲ ③ 'score_voted' score 연관
Feature Addition

Let's MIX!

④ Feature Selection DATASET
+
Feature Addition DATASET

① Feature Importance 기준 컬럼 선별 ② 수치형데이터 Scaling ③ 'score_voted' 컬럼생성 ④ Feature 선별 + 추가

Feature 선별 + 추가 ▶ 성능 향상 !

(Feature Importance 0 컬럼제거 + Label 연관 컬럼 추가)

	Before		
Feature 선별	GBC	0.656250	0.649578
	LGBM	0.657255	0.651714
Feature 추가	GBC	0.539382	0.527113
	LGBM	0.527412	0.518493

(AUC)

+▲ 0.0021

	After		
MIX	GBC	0.654697	0.647744
	LGBM	0.659448	0.653855

N차 전처리 4가지 DATASET 성능 비교

① Feature Importance 기준 컬럼 선별

	accuracy	AUC	precision	recall	f1
Ada	0.648592	0.644708	0.642712	0.734684	0.685628
GBC	0.647220	0.643093	0.640252	0.738707	0.685965
XGB	0.632212	0.629347	0.634453	0.695699	0.663666
LGBM	0.655209	0.651630	0.649966	0.734530	0.689665

② 수치형데이터 Scaling

	accuracy	AUC	precision	recall	f1
Ada	0.650402	0.644314	0.641925	0.759590	0.695818
GBC	0.656159	0.649482	0.643907	0.775907	0.703771
XGB	0.648209	0.644031	0.648809	0.723138	0.683960
LGBM	0.657255	0.651714	0.649821	0.756639	0.699174

③ 'score_voted' 컬럼 추가

	accuracy	AUC	precision	recall	f1
Ada	0.537738	0.527680	0.546355	0.718104	0.620566
GBC	0.542032	0.531256	0.548492	0.735289	0.628300
XGB	0.527321	0.519422	0.541292	0.668981	0.598401
LGBM	0.537829	0.528377	0.547200	0.707342	0.617050

④ Feature 선별 + 추가

	accuracy	AUC	precision	recall	f1
Ada	0.649214	0.642673	0.639074	0.766534	0.697025
GBC	0.654697	0.647744	0.641612	0.779379	0.703817
XGB	0.647752	0.643345	0.647341	0.726784	0.684766
LGBM	0.659448	0.653855	0.651339	0.759764	0.701386

BEST OUTCOME

LGBM +

① feature importance 기준 컬럼 선별

	accuracy	AUC	precision	recall	f1
Ada	0.643392	0.644708	0.642712	0.734684	0.685628
GBC	0.647220	0.643093	0.640252	0.738707	0.685965
XGB	0.632212	0.629347	0.634453	0.695699	0.663951
LGBM	0.655209	0.651630	0.649966	0.734530	0.689665

Feature Importance가 0인 컬럼 제거

+

‘score’ 연관 컬럼 추가

② 수치형데이터 scaling

	accuracy	AUC	precision	recall	f1
Ada	0.650402	0.644314	0.641925	0.759590	0.695818
GBC	0.650159	0.649402	0.643907	0.775907	0.703771
XGB	0.643209	0.641241	0.648809	0.723138	0.683960
LGBM	0.657255	0.651714	0.649821	0.756639	0.699174

③ SCORE_VOTE 컬럼 추가

	accuracy	AUC	precision	recall	f1
Ada	0.537738	0.527680	0.546355	0.718104	0.620566
GBC	0.542032	0.531256	0.548492	0.735289	0.628300
XGB	0.527321	0.519422	0.541292	0.668981	0.598401
LGBM	0.537829	0.528377	0.547200	0.707342	0.617050

④ Feature 정리 + 컬럼 추가

	accuracy	AUC	precision	recall	f1
Ada	0.649214	0.642673	0.639074	0.766534	0.697025
GBC	0.654697	0.647744	0.641612	0.779379	0.703817
XGB	0.647752	0.645345	0.647341	0.726784	0.684766
LGBM	0.659443	0.653855	0.651339	0.759764	0.701386

하이퍼파라미터 튜닝

Tuned Parameters

```
{'min_child_weight': [1, 5, 10],
'gamma': [0.5, 1, 1.5, 2, 5],
'subsample': [0.6, 0.8, 1.0],
'colsample_bytree': [0.6, 0.8, 1.0],
'max_depth': [2, 4, 7, 10]}
```

BEST PARAMS

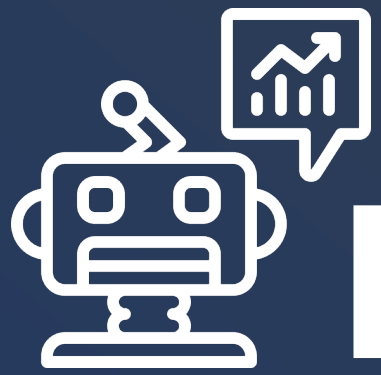
```
{'colsample_bytree': 0.6,
 'gamma': 0.5,
 'max_depth': 10,
 'min_child_weight': 10,
 'subsample': 0.6}
```

Best Outcome : LGBM & DATASET & Tuned Hyper Parameters

GBC & DATA		
Confusion Matrix	2707	2476
	1312	4449
Accuracy	0.6539	

LGBM & DATA		
Confusion Matrix	2881	2302
	1406	4355
Accuracy	0.6612	

#4. PREDICTION



PREDICTION

Did She Vote?

#4. PREDICTION

미국시민권자 홍콩거주 여성은 투표했을까?

- 여성, 50대, 아시안
- 미국국적, 홍콩거주
- 도시출신, 영어원어민
- 마키아벨리즘 37점
- 조용한 성격, 고학력자

#4. PREDICTION

미국시민권자 홍콩거주 여성은 투표했을까?

Low MACH-IV score,
Over 50s , Female, Asian...

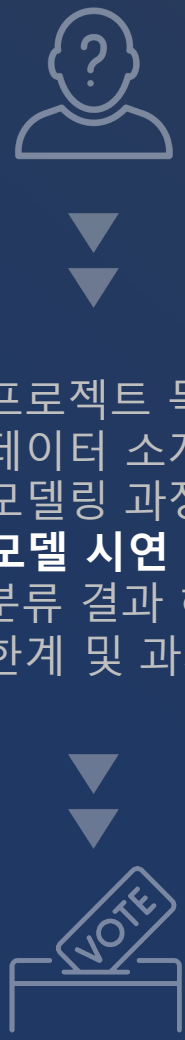
Did She Vote?

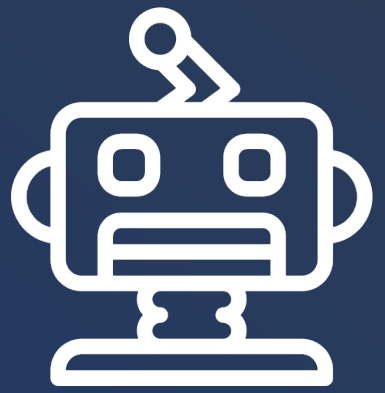
▸▸▸

YES

#4. PREDICTION

우리 분류모델이 판단한 투표참여자 성향





LGBM Model

Predict Probability

NO	0.3631
YES	0.6368

Feature Importance

	0	1
	age	137
	score_voted	52
	TYP_quiet	44
	introelapse	42
	surveyelapse	42
	screenh	41
	testelapse	41
	Q9I	40
	Q1I	38
	country_US	38

⋮

ANALYSIS

Is our Machine LEARNING?

#5. ANALYSIS

머신러닝 모델 고도화 == 영혼 끌어모으기

전처리 단계별 AUC value	
RAW_DATASET	1차 전처리_DATASET
0.6516	0.6517
N차 전처리_DATASET	하이퍼파라미터_최종DATASET
0.6538	0.6559

0.6516 ——— + ▲0.043 ———→ 0.6559

#5. ANALYSIS

머신러닝 모델 고도화 == 영혼 끌어모으기

전처리 단계별 confusion matrix

RAW_DATASET		
Confusion Matrix	3372	2557
	1716	4748
Accuracy	0.6552	

1차 전처리_DATASET		
Confusion Matrix	2834	2349
	1402	4359
Accuracy	0.6573	

N차 전처리_DATASET		
Confusion Matrix	2840	2343
	1384	4377
Accuracy	0.6594	

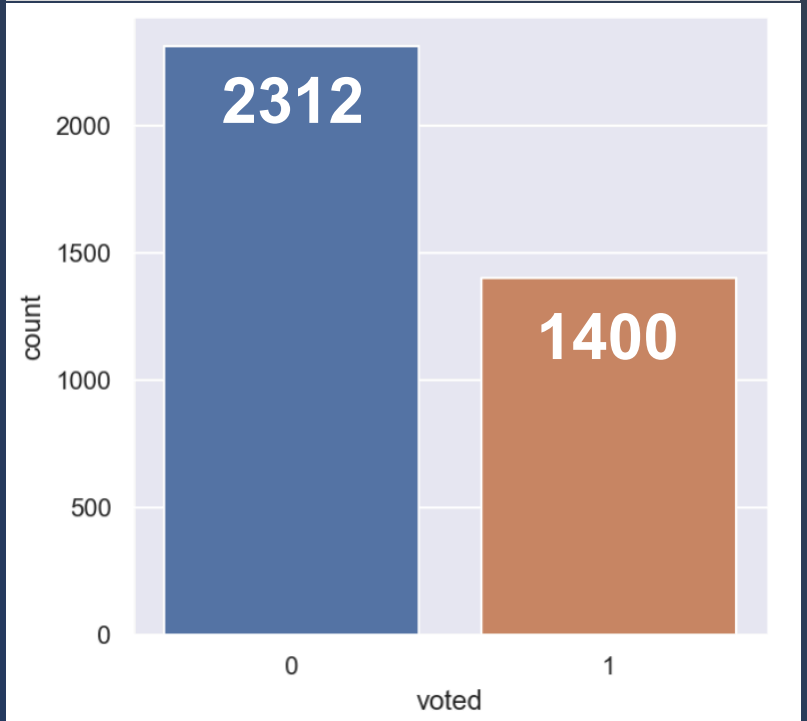
하이퍼파라미터_최종DATASET		
Confusion Matrix	2881	2302
	1406	4355
Accuracy	**0.6612**	

0.6552 ——— + ▲**0.006** ———→ 0.6612

#5. ANALYSIS

우리 머신은 'voted' == 0 분류를 잘 못해요..

df.predict != df.test

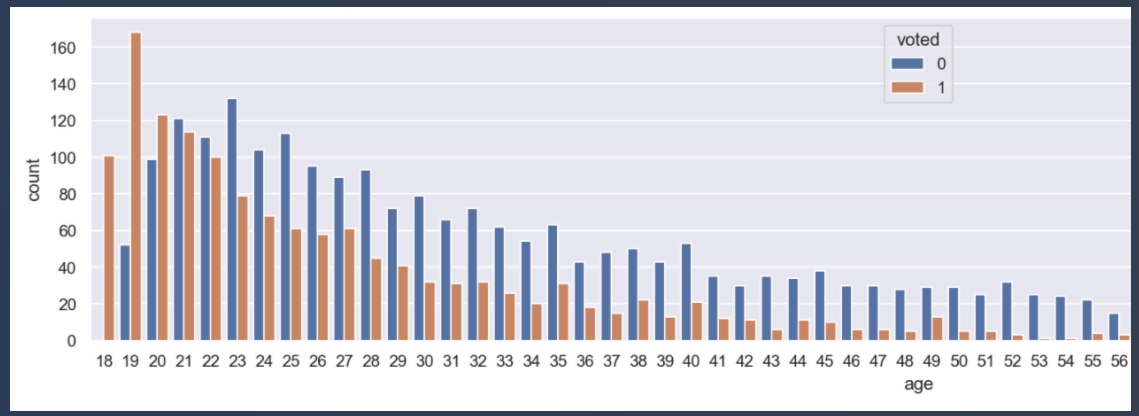


#5. ANALYSIS

우리 머신은 'voted' == 0 분류를 잘 못해요..

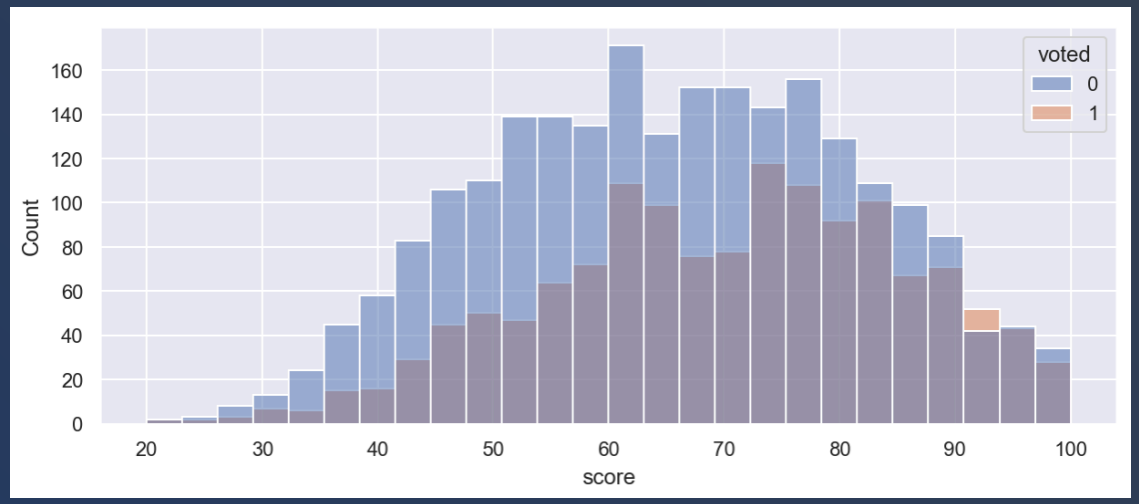
age

어른이라고
다 투표하는 건
아닌데...



score

점수 낮다고
다 투표하는 건
아닌데...



#6. REVIEW

REVIEW

Moving Forward

#6. REVIEW

머신 잘 키우는 방법 알려주세요... (내공0.6659)

1	수치 데이터 이상치 처리	한 컬럼에서만 이상치를 보이는 데이터.. 제거하기엔 너무 아까울 때, Scaling이 최선의 방법일까?
2	유의미한 범주데이터 선별	어떤 Feature가 머신러닝 모델 개발에 유의미한지 판단하는 기본적 기준이 있을까?
3	마키아벨리즘 관련 데이터에 가중치 부여	가중치, 그 수치는 어떻게 구하며 연산 방법은 어떻게 결정해야 할까?
4	하이퍼파라미터 튜닝	튜닝하면 성능이 하락하는 모델.. 무엇이 문제인 걸까?