

설문조사를 통한 지지정당 예측



행복 범주형자료분석팀

김찬영 이혜인 김서윤 심은주 진수정





CONTENTS

01. 주제 선정 배경
02. DATA
03. 시각화
04. 결측치 처리
05. 3주차 예고



01

주제 선정 배경



주제 소개



01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

What

정치성향과 관련 없는 개인적인 질문들로 지지 정당을 파악할 수는 없을까?

Question	Answer
Do you have any siblings?	Yes/No
Does life have a purpose?	Yes/No
Do you have more than one pet?	Yes/No
Are you good good/effective liar?	Yes/No
Do you personally own gun?	Yes/No

:

응답에 따르면...

공화당 VS 민주당

지지자겠구나!

주제 선정 배경



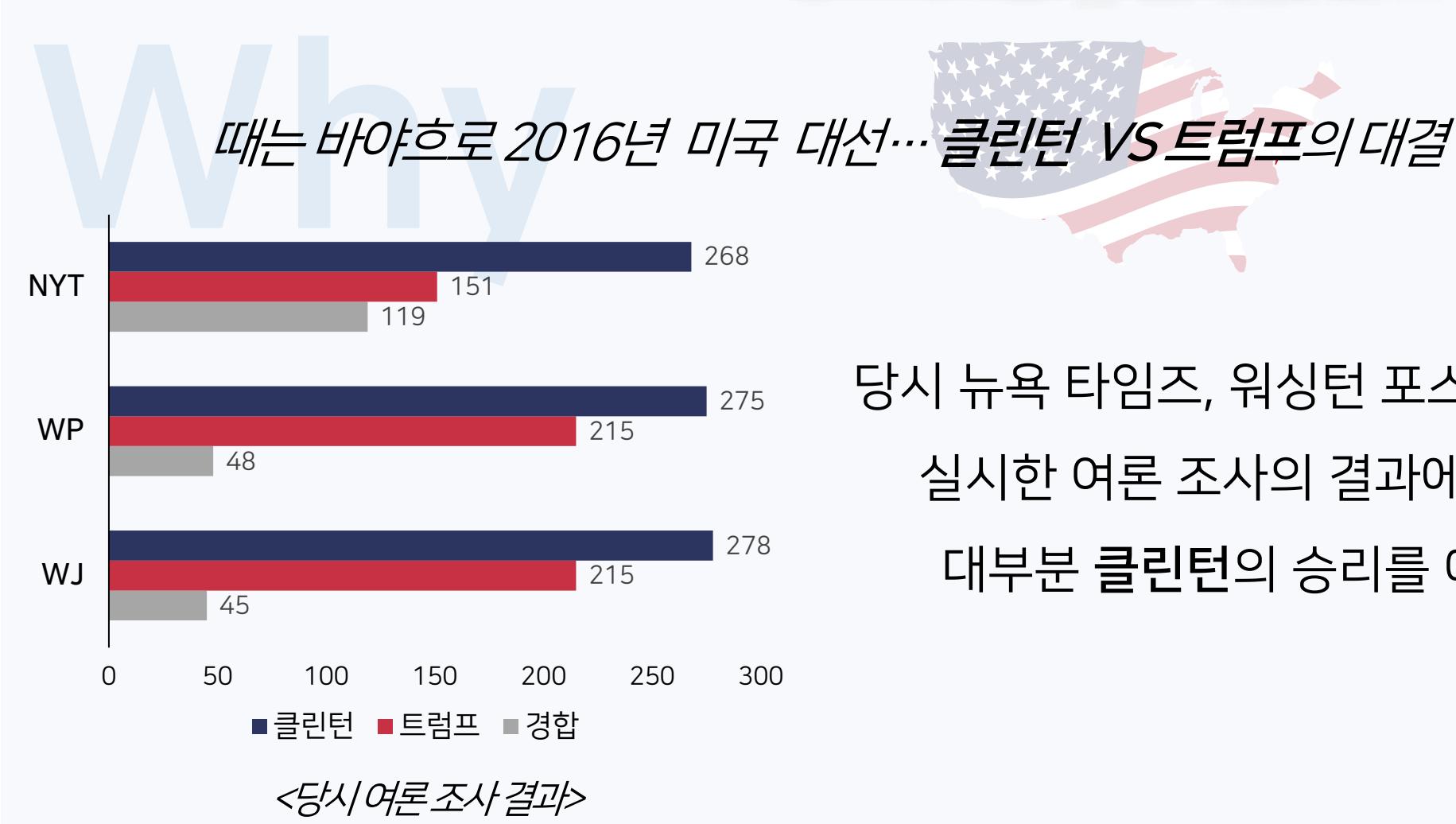
01. 주제 선정 배경

02. DATA

03. 시각화

04. 결측치 처리

05. 3주차 예고



주제 선정 배경

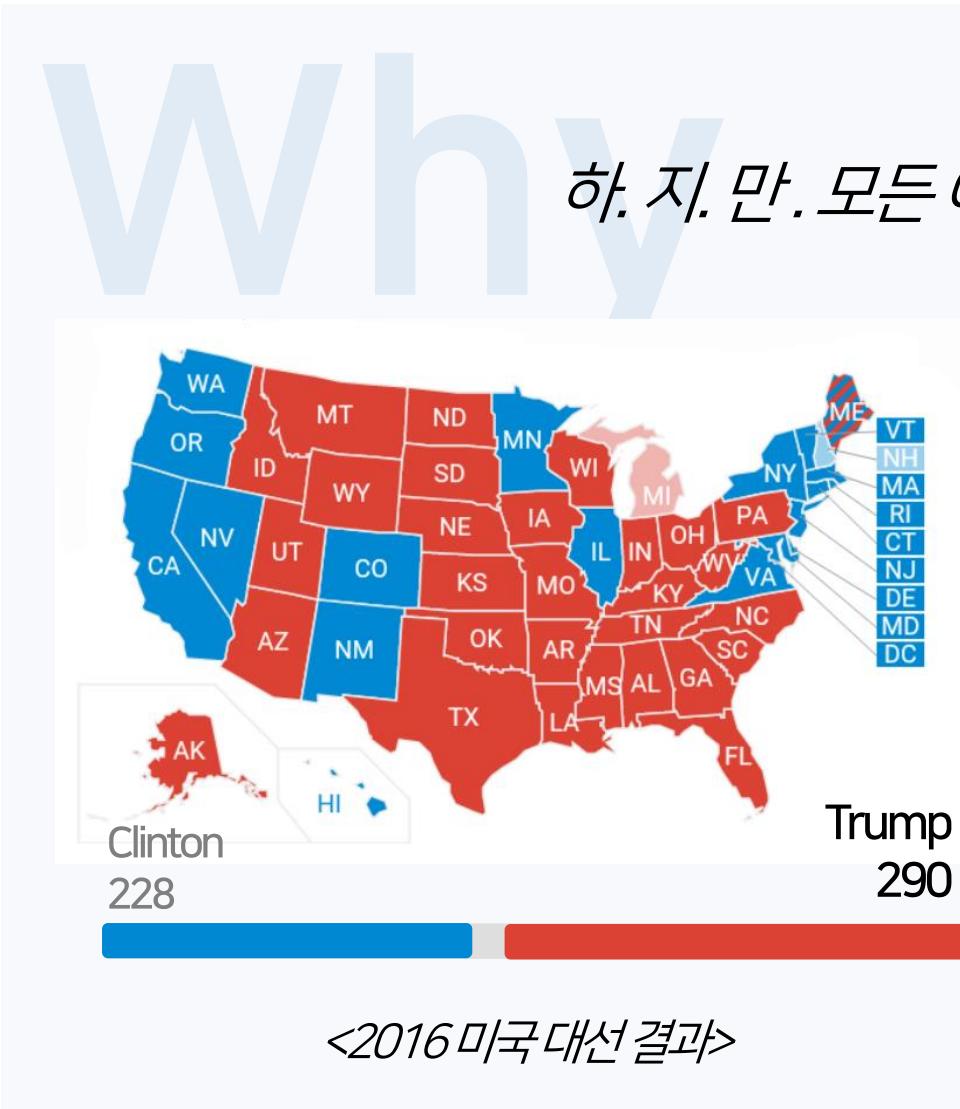
01. 주제 선정 배경

02. DATA

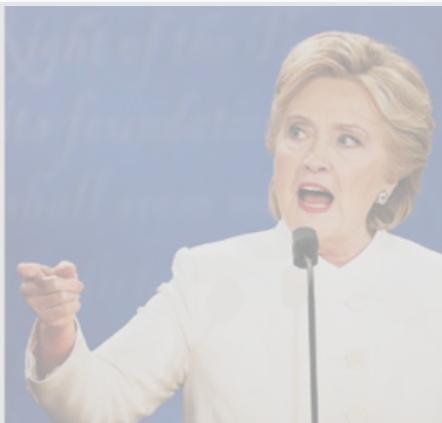
03. 시각화

04. 결측치 처리

05. 3주차 예고



WIN!



내가 뭐랬어!
여론조사 믿을거 아니랬지?

주제 선정 배경



01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

Why

이러한 결과에 따라, 당시 여론조사의 문제점이 드러남

문제점 1. 여론조사는 대졸 이상 학력의 유권자가 응답률이 더 높음.

즉, 조사의 특성상 조사의 대상이 편향될 수밖에 없음.

문제점 2. 'Shy Trump 효과'

사람들이 여론조사에서는 자신이 트럼프를 지지한다는 의사를 드러내지 않는 것

자신의 정치 성향을 드러내는 것을 꺼렸던 것

주제 선정 배경



01. 주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

이러한 결과에 따라, 당시 여론조사의 문제점이 드러남

문제점 1. 여론조사 ~~는 개인들의 특성을 반영할 수 있음~~ 따라서 개인들의 특성을 반영할 수 있으면.

이를 기반으로 그들이 직접적으로 성향을 드러내지 않아도

문제점 2. 정치 성향을 예측할 수 있는 방법에 대한 관심이 높아짐

사람들이 여론조사에서는 자신이 트럼프를 지지한다는 의사를 드러내지 않는 것

자신의 정치 성향을 드러내는 것을 꺼렸던 것



02 DATA





- Kaggle Competition (2016) : Can we predict voting outcomes?

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

Can we predict voting outcomes?

Can we accurately predict voting outcomes by using informal polling questions?

2,874 teams · 4 years ago

무려 2,874팀이 참여했던 공모전!

~~그냥 우리나라 대화 데이터 이용하는 것 아니라고 말하고 싶어서 넣어봄...^^~~



- 설문조사 데이터 (출처: Show of Hands)

01.
주제 선정 배경02.
DATA03.
시각화04.
결측치 처리05.
3주차 예고

User_ID	YOB	Gender	Income	...	Q1	Q2	Q3	Q4	Q5
1	1938	Male	NA		No	NA	No	No	No
4	1970	Female	over \$150,000		NA	Yes	No	No	No
5	1997	Male	\$75,000 - \$100,000		NA	Yes	Yes	No	NA
8	1983	Male	\$100,001 - \$150,000		No	Yes	No	Yes	No
9	1984	Female	\$50,000 - \$74,999		No	Yes	No	No	No
10	1997	Female	over \$150,000		NA	NA	NA	NA	No
:									

6542 obs & 108 variables

...



- 설문조사 데이터 (출처: Show of Hands)

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고



SOH

미국 내 700만명 이상이 사용하는 어플로
주로 설문조사를 하는데 사용됨

<설문조사 방식>

Show of Hands asks...
Do you own a gun?
DISCUSS - 155 RESULTS
Last year
Lifestyle
37

한 질문에 대해
여러 사용자들의 응답을 받는 방식



- 설문조사 데이터 (출처: Show of Hands)

The screenshot shows four survey questions from the Show of Hands platform:

- Do you own a gun?** (Last year Lifestyle) - DISCUSS - 155 RESULTS
- Are you generally more of an optimist or a pessimist?** (Last year Lifestyle) - DISCUSS - 36 RESULTS
- Are you a feminist?** (37 of Hands) - DISCUSS - 26 RESULTS
- Are you adventurous?** (4 years ago Lifestyle) - DISCUSS - 46 RESULTS
- Does the "power of positive thinking" actually work?** (6 years ago Lifestyle) - DISCUSS - 46 RESULTS
- Are you more of a night person or a morning person?** (4 years ago Lifestyle) - DISCUSS - 46 RESULTS

이렇게 개별 응답한 설문을 User_ID를 기준으로 합친 데이터

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고



• 개인정보 데이터

01.
주제 선정 배경02.
DATA03.
시각화04.
결측치 처리05.
3주차 예고

5가지 개인정보

User_ID	YOB	Gender	Income	Household Status	Education	Party
1	1938	Male	NA	Married (w/kids)	NA	Democrat
4	1970	Female	over \$150,000	Domestic Partners (w/kids)	Bachelor's Degree	Democrat
5	1997	Male	\$75,000 - \$100,000	Single (no kids)	High School Diploma	Republican
8	1983	Male	\$100,001 - \$150,000	Married (w/kids)	Bachelor's Degree	Democrat
9	1984	Female	\$50,000 - \$74,999	Married (w/kids)	High School Diploma	Republican
10	1997	Female	over \$150,000	Single (no kids)	Current K-12	Democrat

:



- 개인정보 데이터

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

5가지 개인정보

Y 변수

User_ID	YOB	Gender	Income	Household Status	Education	Party
1	1938	Male	NA	Married (w/kids)	NA	Democrat
4	1970	Female	over \$150,000	Domestic Partners (w/kids)	Bachelor's Degree	Democrat
5	1997	Male	\$75,000 - \$100,000	Single (no kids)	High School Diploma	Republican
8	1983	Male	\$100,001 - \$150,000	Married (w/kids)	Bachelor's Degree	Democrat
9	1984	Female	\$50,000 - \$74,999	Married (w/kids)	High School Diploma	Republican
10	1997	Female	over \$150,000	Single (no kids)	Current K-12	Democrat

:



- Y변수: Democrat / Republican

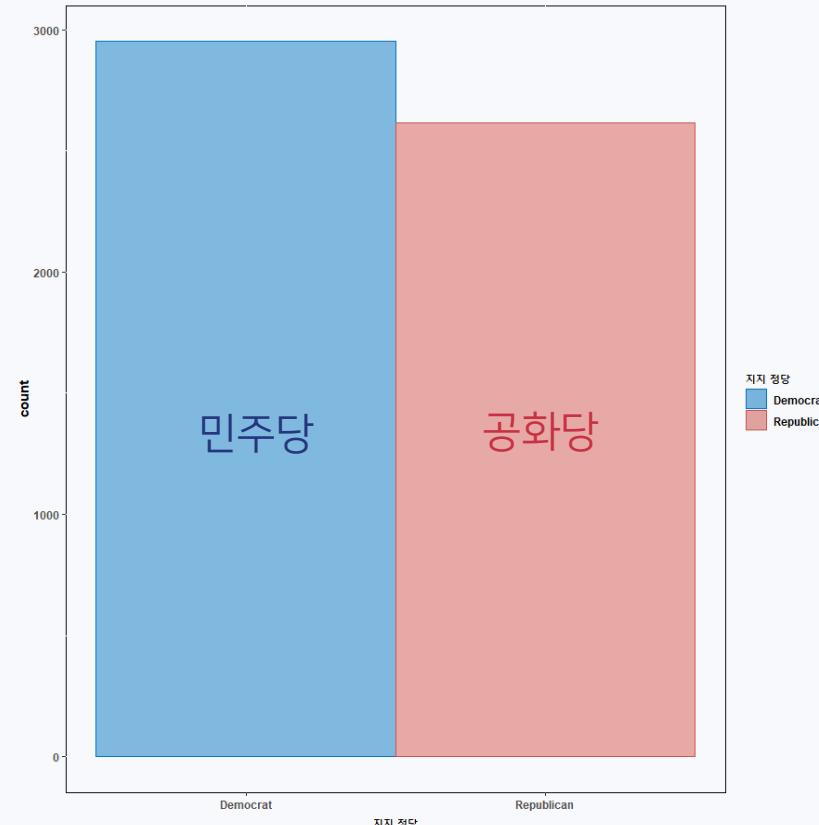
01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고



응답자의 **공화당** 지지자 비율과
민주당 지지자 비율이 비슷하여
샘플링을 따로 진행하지 않음



• 설문조사 응답 데이터

101개의 질문에 대한 설문 응답 데이터

User_ID	Q1	Q2	Q3	Q4	Q5	Q6	...	Q100	Q101
1	No	NA	No	No	No	Yes		No	Yes
4	NA	Yes	No	No	No	Yes		Yes	No
5	NA	Yes	Yes	No	NA	Yes		No	No
8	No	Yes	No	Yes	No	No		No	Yes
9	No	Yes	No	No	No	Yes		No	Yes
10	NA	NA	NA	NA	No	Yes		NA	NA

:

질문에 대한 대답은 모두 Yes/No 이거나

질문에 따라 두 가지 선택지 중 하나를 선택하는 것으로 이분화 되어 있음

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고



- 데이터 재범주화

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

User_ID	YOB	Income	Education
1	76	NA	NA
4	44	over \$150,000	Bachelor's Degree
5	17	\$75,000 - \$100,000	High School Diploma
8	31	\$100,001 - \$150,000	Bachelor's Degree
9	30	\$50,000 - \$74,999	High School Diploma
10	17	over \$150,000	Current K-12

YOB는 Year of Birth로 출생연도

설문조사 시행 연도인 2014년을 기준으로 나이 계산 후(미국 기준) age 변수 새롭게 만듦

- 데이터 재범주화

나이를 구간별로 재범주화

AGE

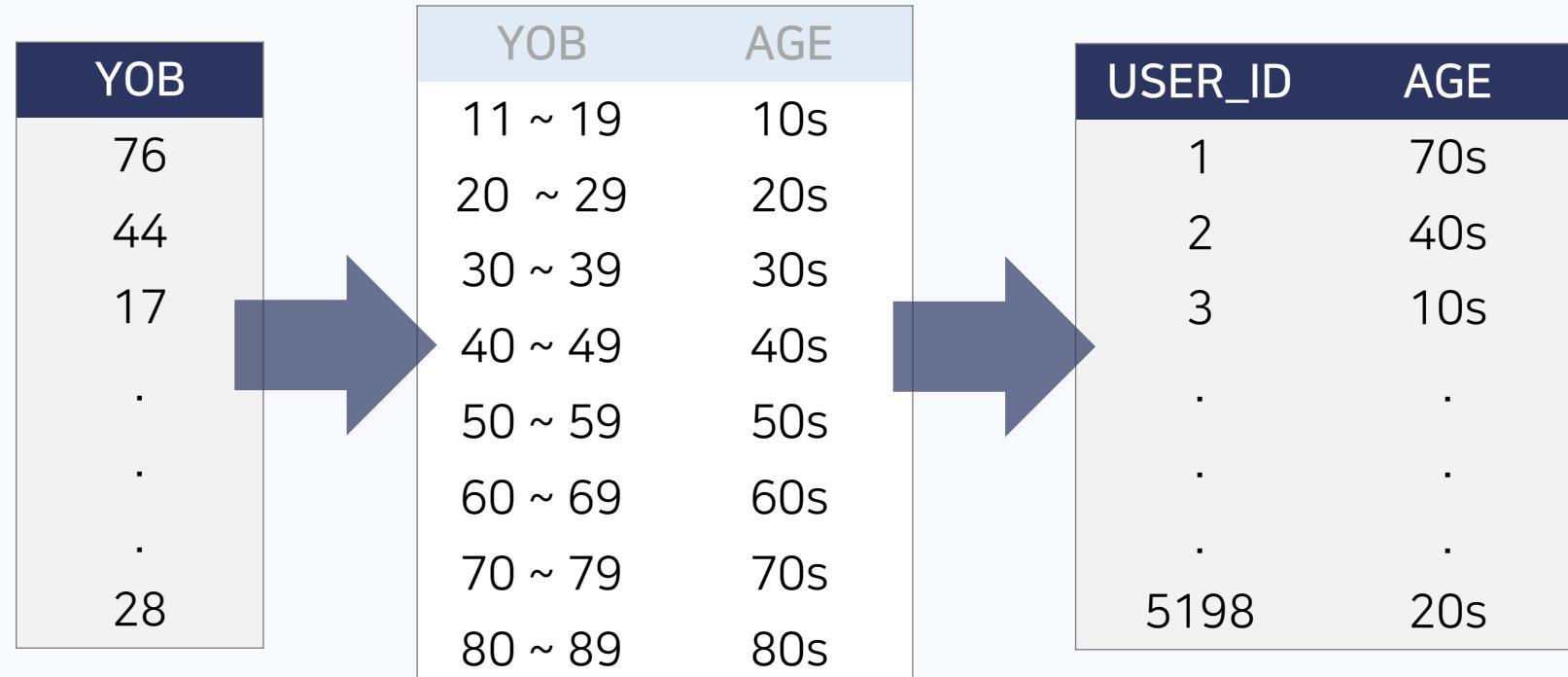
01.
주제 선정 배경

02.
DATA

03.
시각화

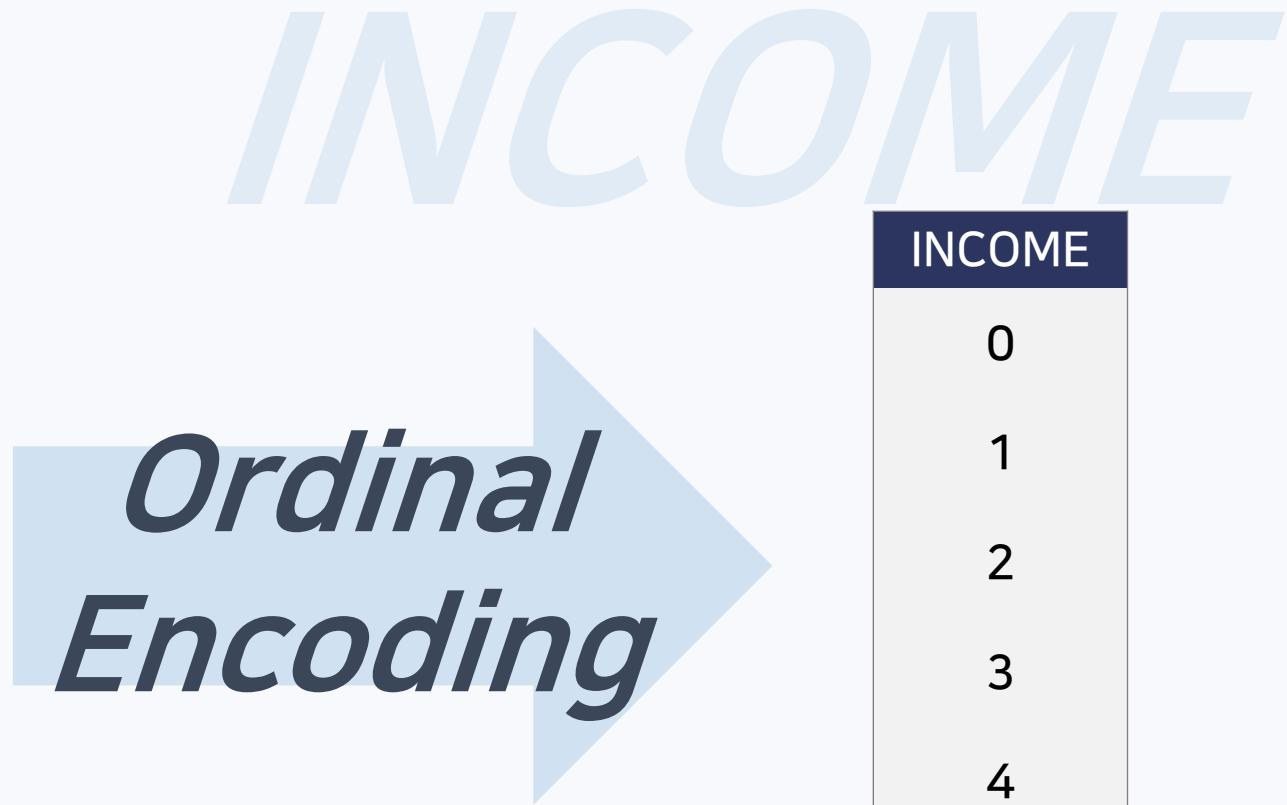
04.
결측치 처리

05.
3주차 예고



- 데이터 재범주화

INCOME
under \$25,000
\$25,001 - \$50,000
\$50,001 - \$75,000
\$75,001 - \$100,000
\$100,001 - \$150,000
over \$150,000



01.
주제 선정 배경

02.
DATA

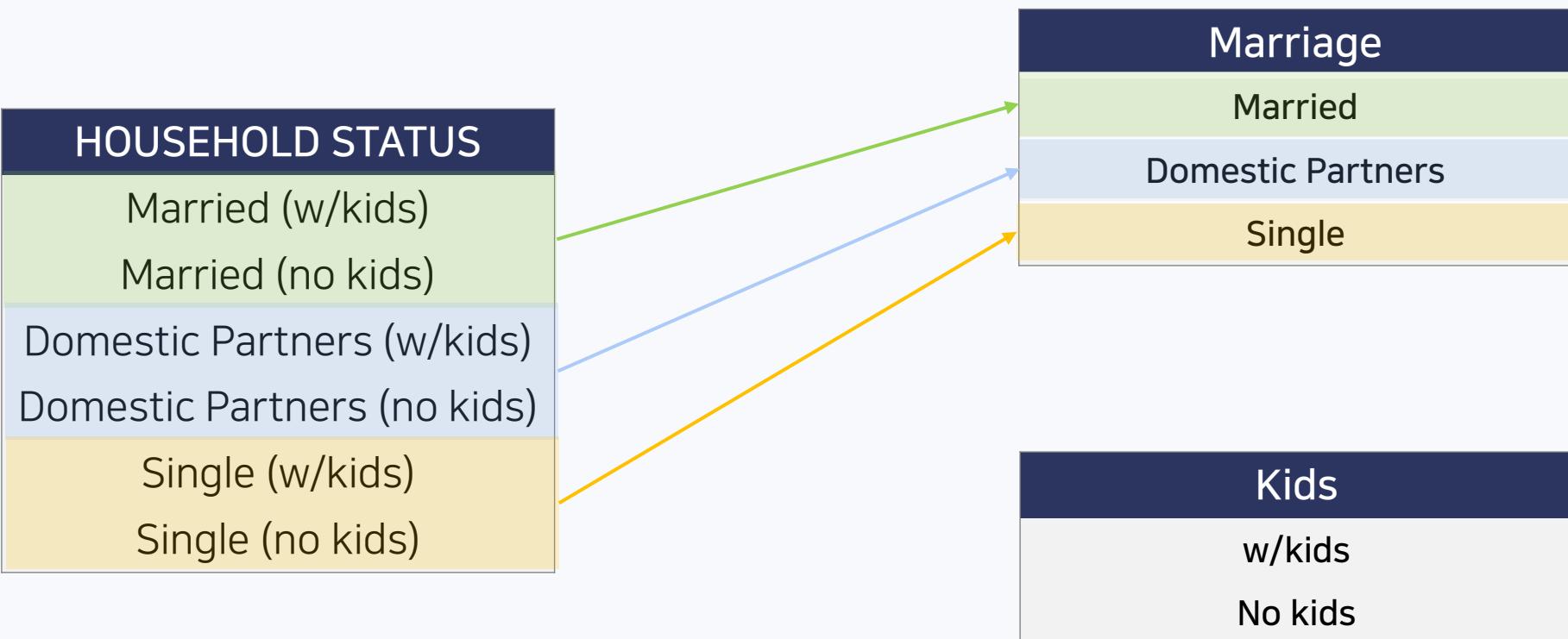
03.
시각화

04.
결측치 처리

05.
3주차 예고



- 데이터 재범주화



01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고



- 데이터 재범주화

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

HOUSEHOLD STATUS
Married (w/kids)
Married (no kids)
Domestic Partners (w/kids)
Domestic Partners (no kids)
Single (w/kids)
Single (no kids)





- 데이터 재범주화

EDUCATION LEVEL

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

재범주화 후에
Ordinal Encoding 진행

EDUCATION LEVEL
Current K-12
Highschool diploma
Current undergraduate
Associate's degree
Bachelor's degree
Master's degree
Doctoral degree

EDUCATION LEVEL
0
1
2



- 변수명 변경

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

Q_ID	Original Question	Answer
Q98059	Do/did you have any siblings? Which parent "wore the pants" in your household?	Yes/No
Q101163	Are you currently carrying a grudge against anyone in your personal life?	Mom/Dad
Q102906	Do you generally like people, or do most of them tend to get on your nerves pretty easily?	Yes/No Yay people/ Grrr People
Q106997	Do you punctuate text messages?	Yes/No
Q108342	Do you enjoy getting together with your extended family?	Yes/No
Q108855		

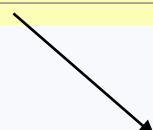
<원래 질문 데이터>



- 변수명 변경

01.
주제 선정 배경02.
DATA03.
시각화04.
결측치 처리05.
3주차 예고

Q_ID	Original Question	Answer
Q98059	Do/did you have any siblings? Which parent "wore the pants" in your household?	Yes/No
Q101163	Are you currently carrying a grudge against anyone in your personal life?	Mom/Dad
Q102906	Do you generally like people, or do most of them tend to get on your nerves pretty easily?	Yes/No
Q106997	Do you punctuate text messages?	Yay people/ Grrr People
Q108342	Do you enjoy getting together with your extended family?	Yes/No
Q108855		Yes/No



Question_ID가 숫자로 되어있어 분석 과정 중 질문 파악이 어려움
직관적 확인 위해 질문 핵심 내용 담아서 변수명 변경



- 변수명 변경

01.
주제 선정 배경02.
DATA03.
시각화04.
결측치 처리05.
3주차 예고

Q_ID	Original Question	Answer
re_Q_havesibling	Do/did you have any siblings ?	Yes/No
re_Q_Dad_household power	Which parent " wore the pants " in your household ?	Mom/Dad
re_Q_carrygrudge	Are you currently carrying a grudge against anyone in your personal life?	Yes/No
re_Q_likepeople	Do you generally like people , or do most of them tend to get your nerves pretty easily?	Yay people/ Grrr People
re_Q_meetoffline	Do you spend more time with friends online or in-person ?	Yes/No
re_Q_extendedfamily	Do you enjoy getting together with your extended family ?	Yes/No

Question_ID가 숫자로 되어있어 분석 과정 중 질문 파악이 어려움
직관적 확인 위해 **질문 핵심 내용** 담아서 변수명 변경



- 변수명 변경

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

Q_ID	Original Question	Answer
re_Q_havesibling	Do/did you have any siblings?	Yes/No
re_Q_Dad_householdpower	Which parent "wore the pants" in your household?	Mom/Dad
re_Q_carrygrudge	Are you currently carrying a grudge against anyone in your personal life?	Yes/No
re_Q_likepeople	Do you generally like people, or do most of them tend to get on your nerves pretty easily?	Yay people/ Grrr People
re_Q_meetoffline	Do you punctuate text messages?	Yes/No
re_Q_extendedfamily	Do you enjoy getting together with your extended family?	Yes/No

향후 진행할 인코딩 과정에서의 편의를 위해 답변을 Yes/No로 통일 해줌

Ex) Mom = Yes / Dad=No, Yay people = Yes / Grrr people = No



- 변수명 변경

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

Q_ID	Answer
Q98059	Yes/No
Q101163	Mom/Dad
Q102906	Yes/No Yay people/ Grrr People
Q106997	
Q108342	Yes/No
Q108855	Yes/No

Q_ID	Answer
re_Q_havesibling	Yes/No
re_Q_Dad_householdpower	Yes/No
re_Q_carrygrudge	Yes/No
re_Q_likepeople	Yes/No
re_Q_meetoffline	Yes/No
re_Q_extendedfamily	Yes/No

Yes/No는 한 번 더 1/0으로 바꿔줌!



- 변수명 변경

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

Q_ID	Answer
Q98059	Yes/No
Q101163	Mom/Dad
Q102906	Yes/No Yay people/ Grrr People
Q106997	
Q108342	Yes/No
Q108855	Yes/No

<기존 변수명>

Q_ID	Answer
re_Q_havesibling	1/0
re_Q_Dad_householdpower	1/0
re_Q_carrygrudge	1/0
re_Q_likepeople	1/0
re_Q_meetoffline	1/0
re_Q_extendedfamilly	1/0

<바뀐 변수명>



- 코드북 작성

바뀐 변수명을 raw data와 비교를 간편히 하기 위해 **코드북 작성**

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

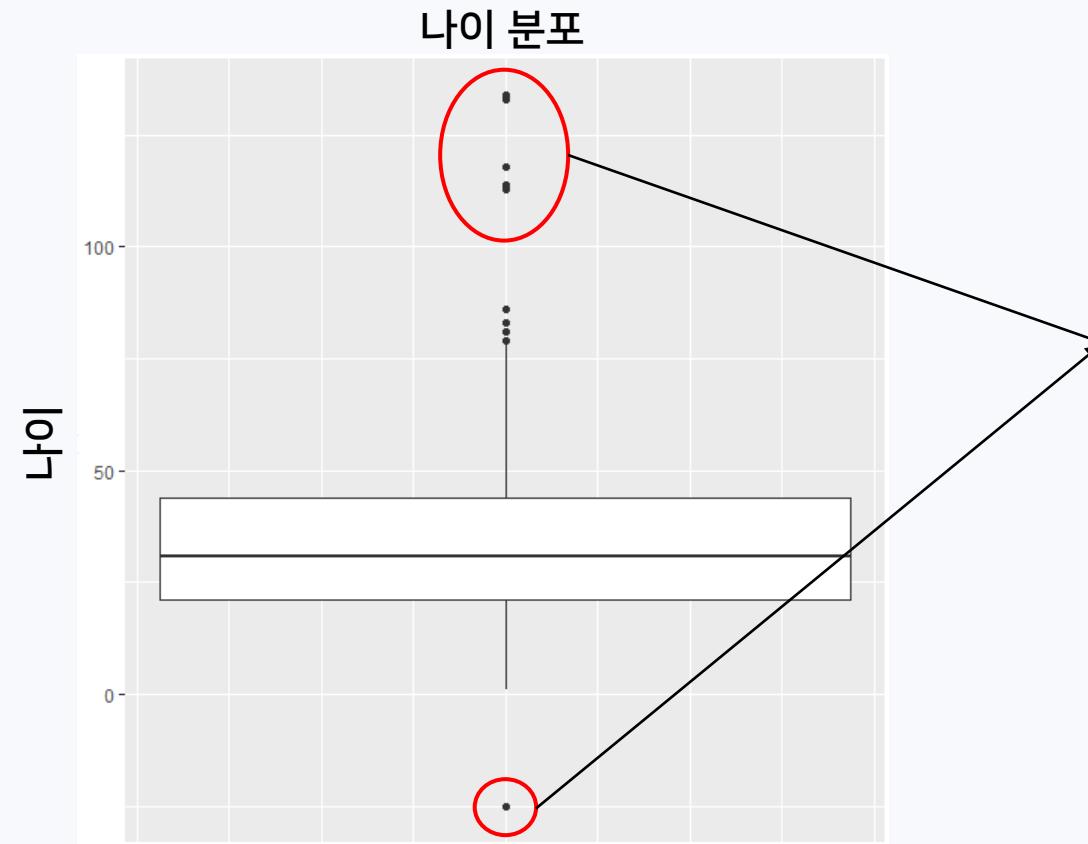
05.
3주차 예고

Ori_Q_ID	Fin_Q_ID	문제	Category	가능한 답변	바뀐 답변
Q98059	re_Q_havesibling	형제 자매 있음?	Relationships	Y/Only-Chil d	1/0
Q101163	re_Q_Dad_householdpower	부모님 중에 집안의 주도권을 가진 사람?	Relationships	Mom/ Dad	0/1
Q102906	re_Q_carrygrudge	인생에서 원한을 산적이 있는지?	Relationships	Y/N	1/0
Q106997	re_Q_likepeople	사람들 좋아하는지? Or 사람들로 인해 쉽게 짜증내는 타입?	Relationships	Yay people!/ Grrr people	1/0
Q109367	mo_Q_poor	가난했던적이 있는가?	Money	Y/N	1/0
Q115195	mo_Q_live_metro	주요 대도시 20마일 이내에 사는지	Money	Y/N	1/0
Q102687	Life_Q_breakfast	매일 아침 먹는지	Life style	Y/N	1/0
Q103293	Life_Q_pet	한마리 이상의 반려동물 있음?	Life style	Y/N	1/0
Q104996	Life_Q_brushT2	매일 양치 두 번 이상 하는지?	Life style	Y/N	1/0
Q102089	mo_Q_own_residence	지금 주거지 렌트인지 자가인지?	Money	Rent/Own	0/1

category는 원래 설문조사 당시 어플 자체에서 분류했던 기준



- 이상치 (Outlier) 제거



Plot 확인 결과, 이상치가
확인되어 Tukey방법을
이용해 구간을 확인하고
3세 이하나 100세 이상
은 이상치로 판단, 삭제
→ 12개 obs 삭제

잘가 소동한 obs...



01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고



- 변수 제거

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

<질문별 무응답 비율>

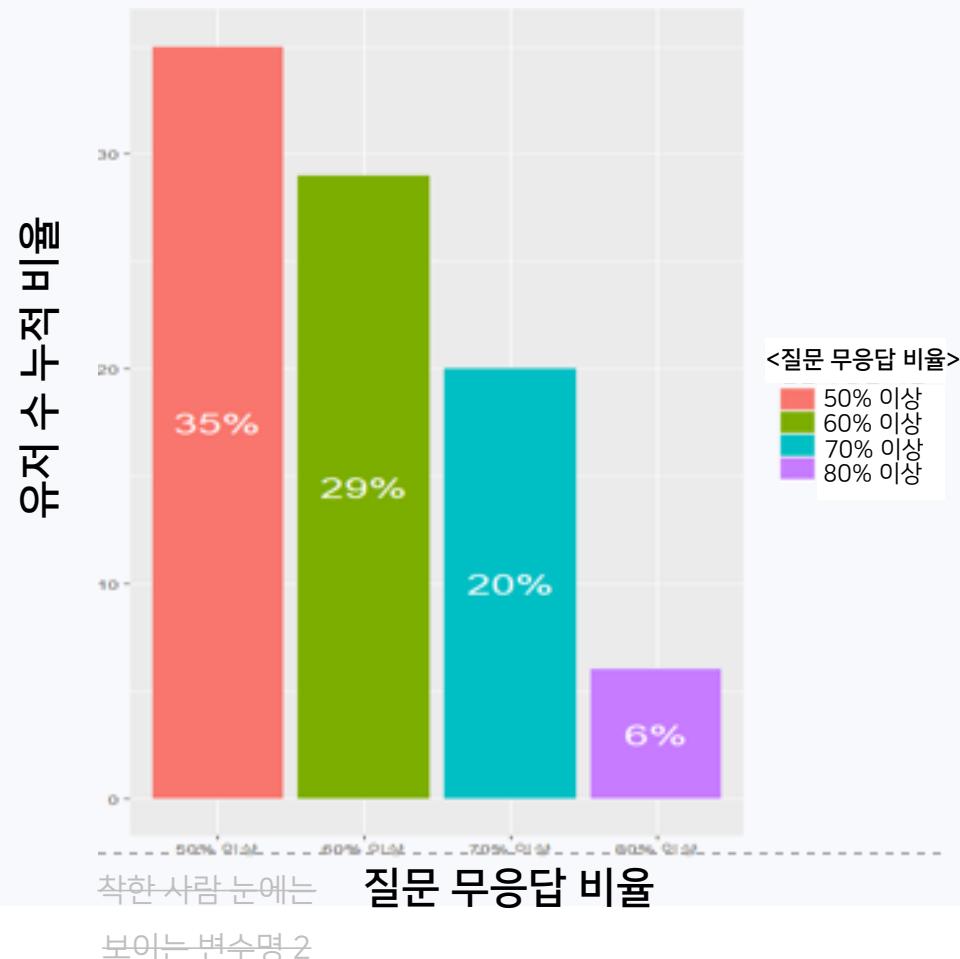


질문별 무응답 비율이
50% 이상일 경우,
무의미하다고 판단,
→ re_Q_interact_dislike
변수 삭제



- 변수 제거

<User별 무응답 비율>



User별 무응답 비율이
80% 이상일 경우,
무의미하다고 판단, 제거

→ 358개의 행 삭제

안그래도 obs 적은데.. ^^
눈물이 납니다....





- 인간 도덕성의 5가지 기반

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고



Jonathan David Haidt
(TED에서 만나요!)

배려 (Care) 공평성 (Fairness) 권위 (Authority) 충성심 (Loyalty) 고귀함 (Sanctity)

설문조사를 통해
5가지 도덕적 기반과 정치적 견해의
연관성을 발견

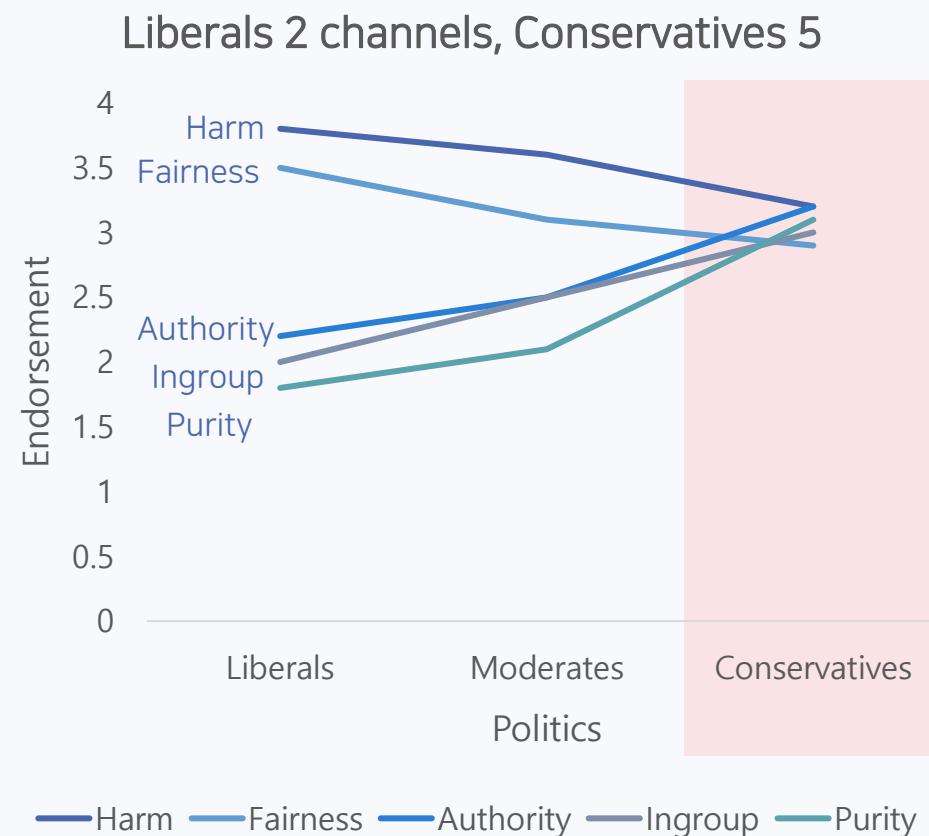
01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고



보수파

배려
(Care) 공평성
(Fairness)

권위
(Authority) 충성심
(Loyalty) 고귀함
(Sanctity)

보수파는 5가지 기반을
모두 잘 이용

01.
주제 선정 배경

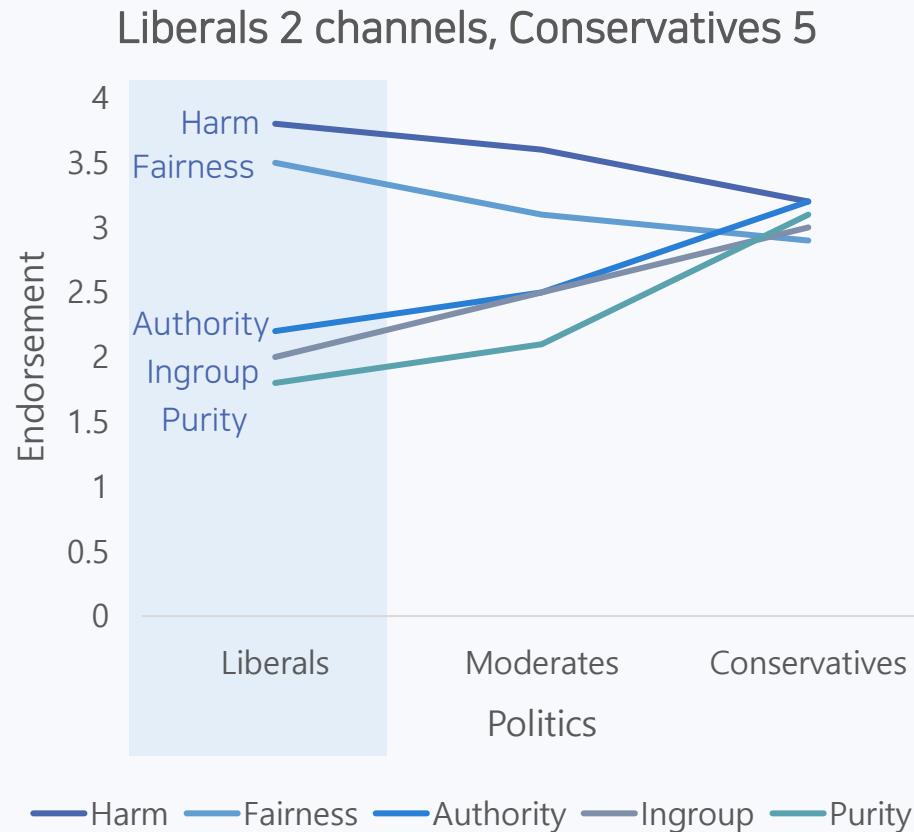
02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

- 인간 도덕성의 5가지 기반



진보파

배려
(Care)

공평성
(Fairness)

진보파는 충성심, 권위, 고귀함을
잘 받아들이지 못함



어떻게 설문 데이터로부터 5가지 기반을 나타내는 변수를 생성할까?

01.
주제 선정 배경

02.
DATA

03.
시각화

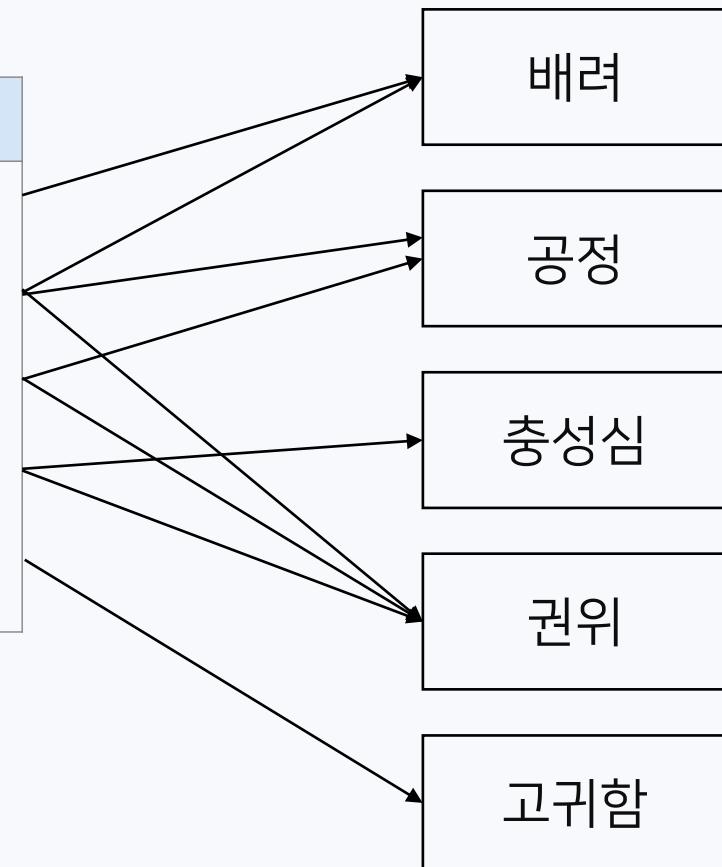
04.
결측치 처리

05.
3주차 예고

Question	Answer
Are you a good/effective liar?	Yes/No
Are you a feminist?	Yes/No
Can money buy happiness?	Yes/No
Were you an obedient child?	Yes/No
Do you pray or meditate on a regular basis?	Yes/No

:

5가지 범주와 관련 있는 설문 문항 추출





추출된 설문 문항들에 대하여 답이 특정 범주에 부합하면 점수를 1 더함

충성심

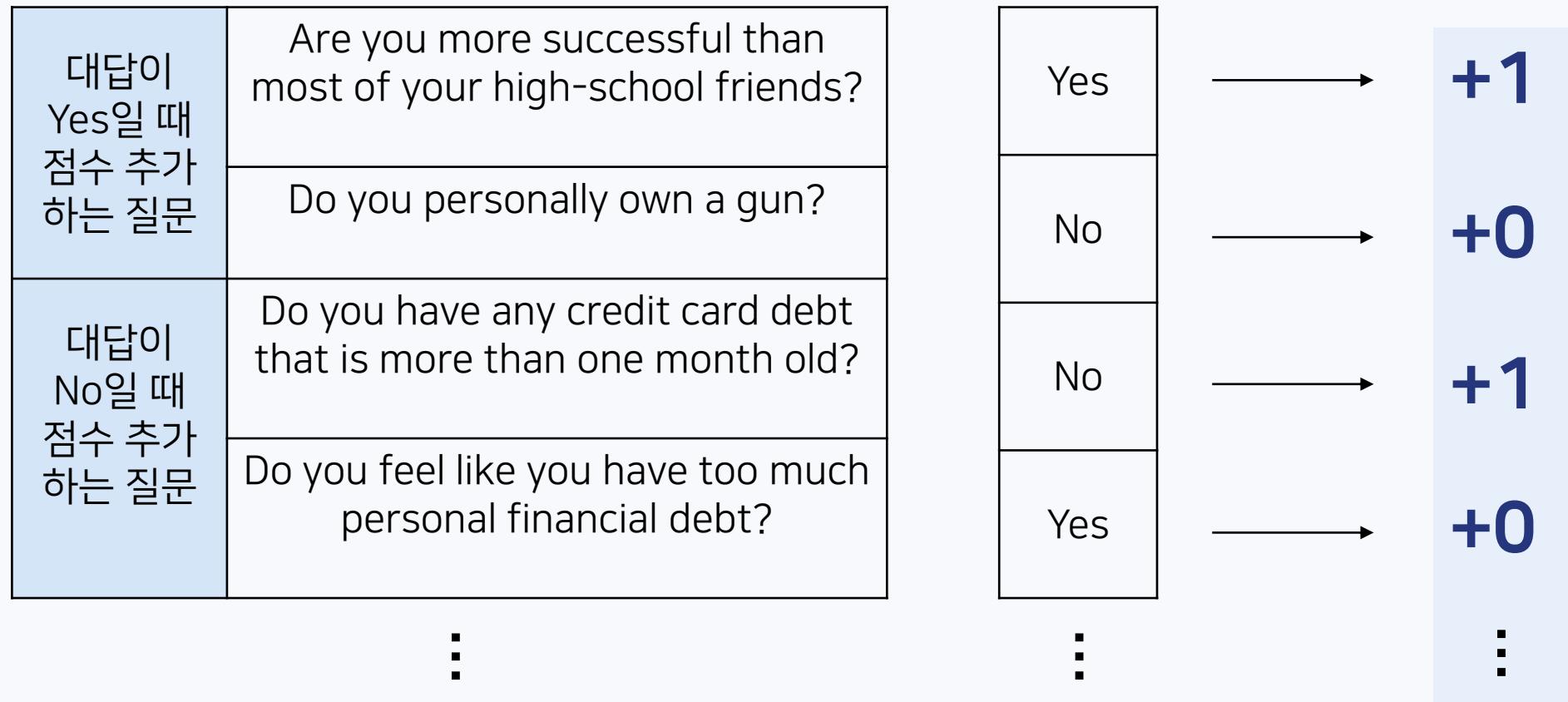
01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

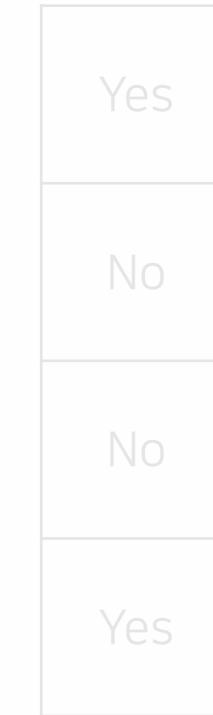




• ~~추출된 5개의 기본 변수에 대하여 답이 특정 범주에 부합하면 점수를 1 더함~~

충성심

	careness	fairness	authority	loyalty	sanctity
1	3	6	11	4	2
2	1	4	10	0	2
3	1	3	15	3	1
4	2	7	13	4	3
5	1	3	9	2	2
6	1	3	7	1	1
7	2	2	9	1	1
8	2	4	13	3	1
9	1	4	9	3	2
10	1	5	7	4	3
11	0	2	1	0	0
12	0	5	8	2	1
13	3	7	10	2	3
14	2	5	10	2	2
15	0	1	10	2	1
16	0	3	15	3	3
17	1	2	4	2	0
18	4	7	12	2	1
19	3	5	11	2	2
20	0	3	15	2	2



+1

+0

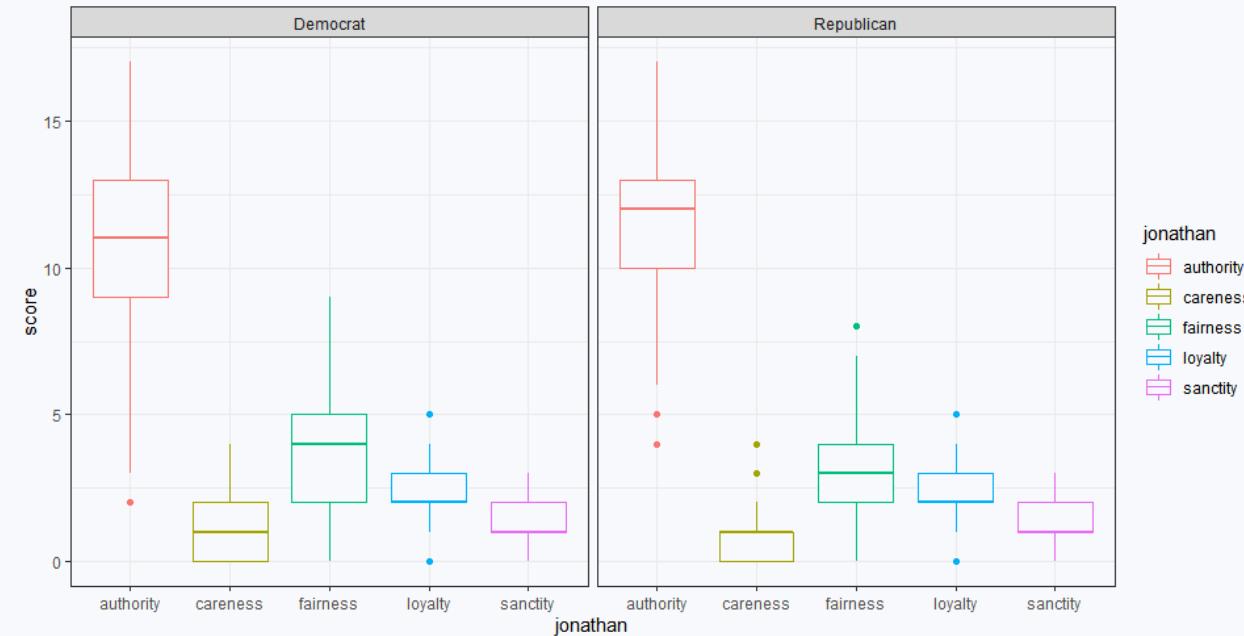
+1

+0

:

:

- 파생변수의 유의성 확인



01.
주제 선정 배경

02.
DATA

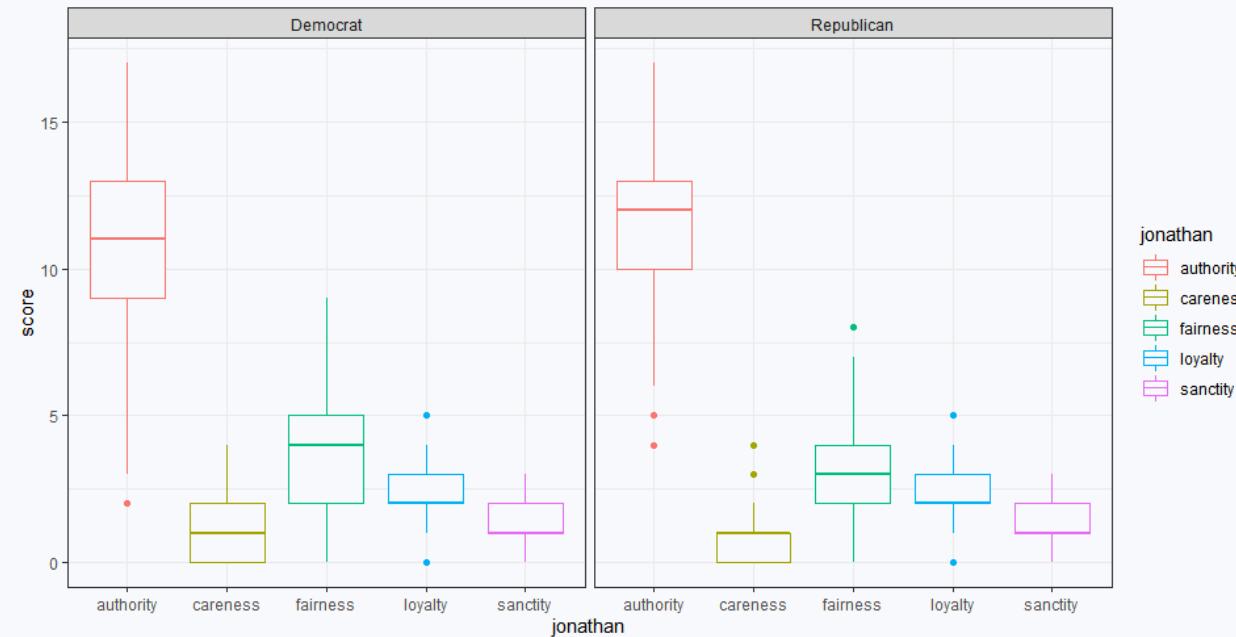
03.
시각화

04.
결측치 처리

05.
3주차 예고

	Care	Fairness	Authority	Loyalty	Sanctity
Democrat	0.3	0.33	0.43	0.54	0.54
Republican	0.23	0.31	0.43	0.57	0.57

- 파생변수의 유의성 확인



Care와 Fairness의 값은
Democrat에서 높게 나타남

Authority, Loyalty, Sanctity는
비슷하게 나타남

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

	Care	Fairness	Authority	Loyalty	Sanctity
Democrat	0.3	0.33	0.43	0.54	0.54
Republican	0.23	0.31	0.43	0.57	0.57



모든 과정을 거쳐 만들어진 통합 설문 데이터

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

USER_ID	Gender	Income	Level	Education							
				env_Q_p_fight	ifoU	mo_Q_minwage_job	mo_Q_minwage_job	care	sanctity
1	Male	2	1		1		1	1	0	0	0
4	Female	5	1		NA		1	1	0	0	0
5	Male	3	0		1		0	NA	1	0	0
8	Male	4	1		NA		NA	1	0	0	0
9	Female	2	0		1		0	0	0	0	0
10	Female	5	0	NA		1	1	1	1	0	0
11	Male	1	1		1		1	NA	1	1	1
12	Male	3	0		0		NA	0	0	1	0
13	Female	3	0		NA		1	1	1	0	0
14	Male	1	1		1		1	NA	1	1	1
15	Male	3	0		0		NA	0	0	1	0

:



03

시각화



01.
주제 선정 배경

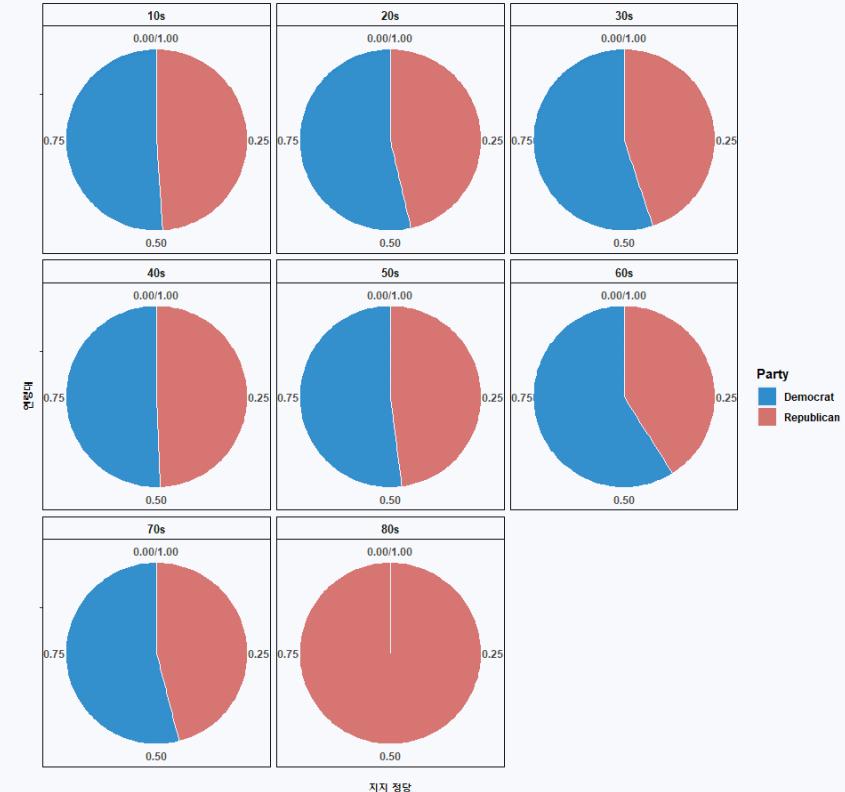
02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

1. Age 변수



<연령대 별 지지 정당비율>

01.
주제 선정 배경

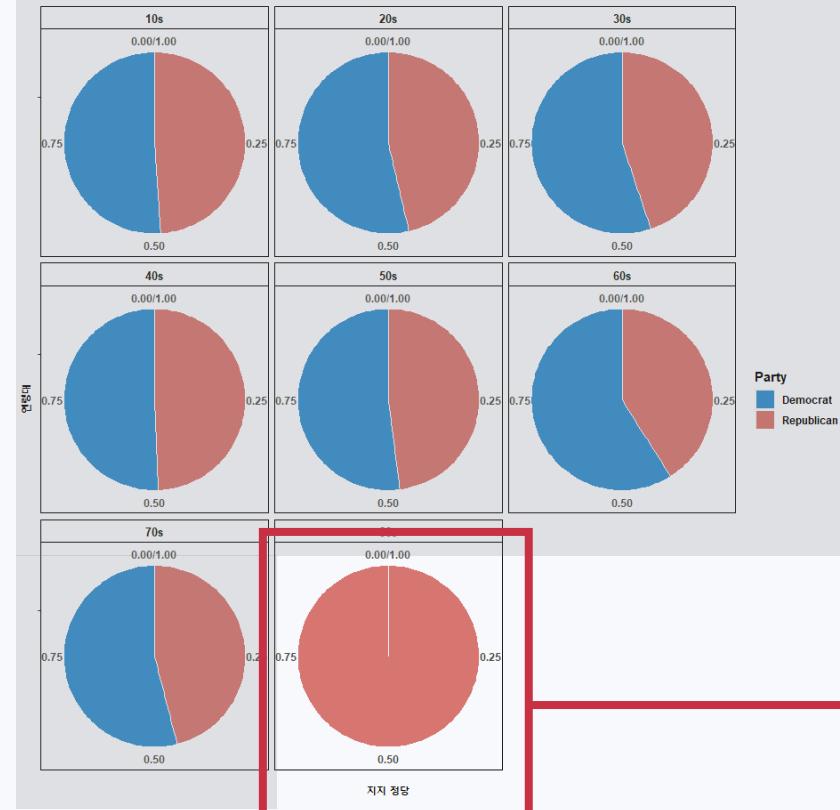
02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

1. Age 변수



<연령대 별 지지 정당비율>

대부분의 연령대

비슷 or 민주당 ↑

BUT 80대

공화당 ↑

* 하지만 80세 이상은
표본이 매우 적음.
편향된 해석에 유의!

01.
주제 선정 배경

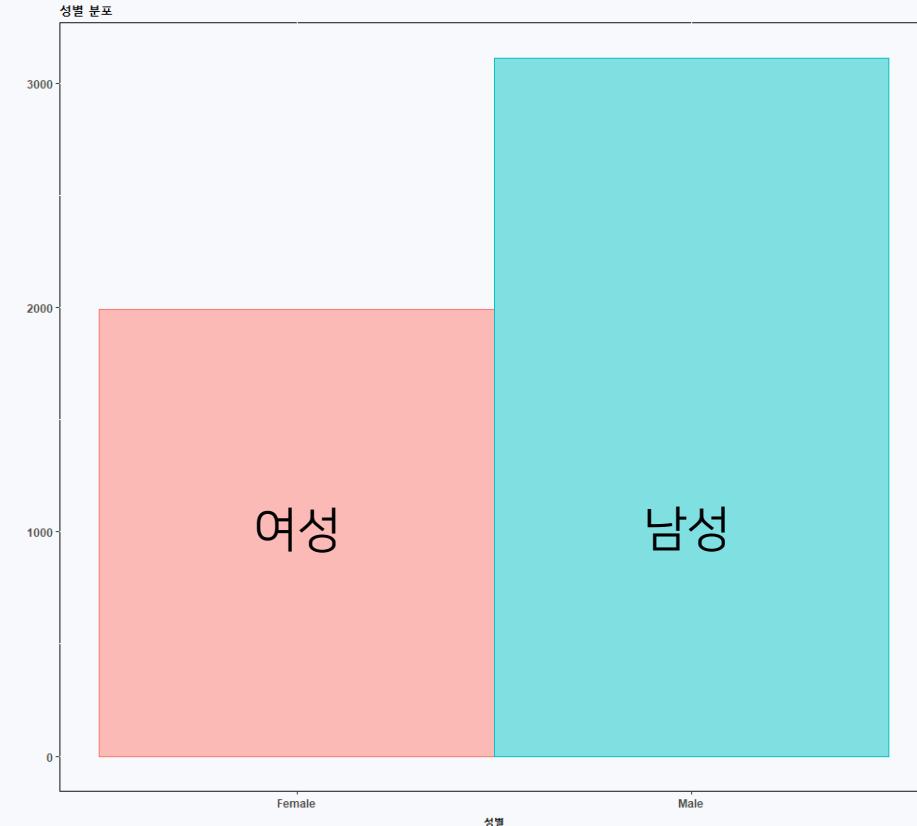
02.
DATA

03.
시각화

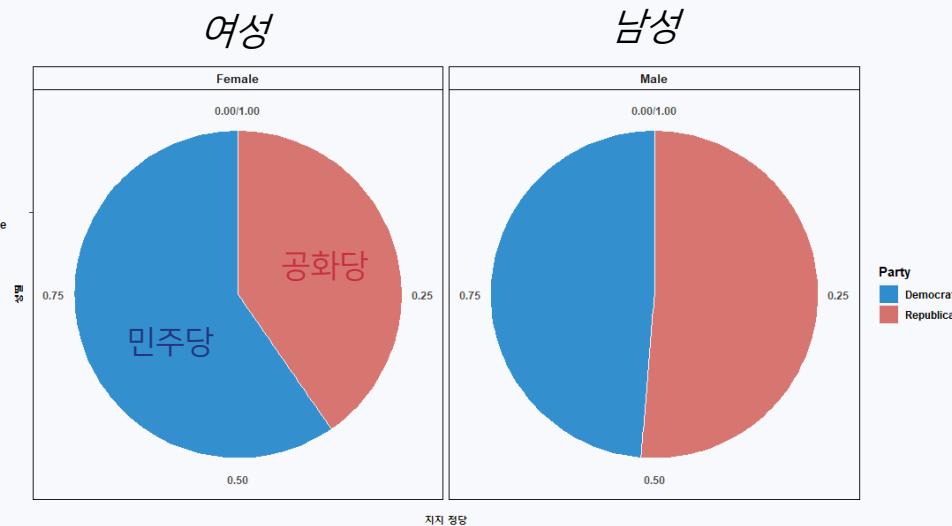
04.
결측치 처리

05.
3주차 예고

2. Gender 변수



<성별 분포 및 지지 정당 비율>



여성은 민주당의 비율이,
남성은 공화당의 비율이 높은 편

01.
주제 선정 배경

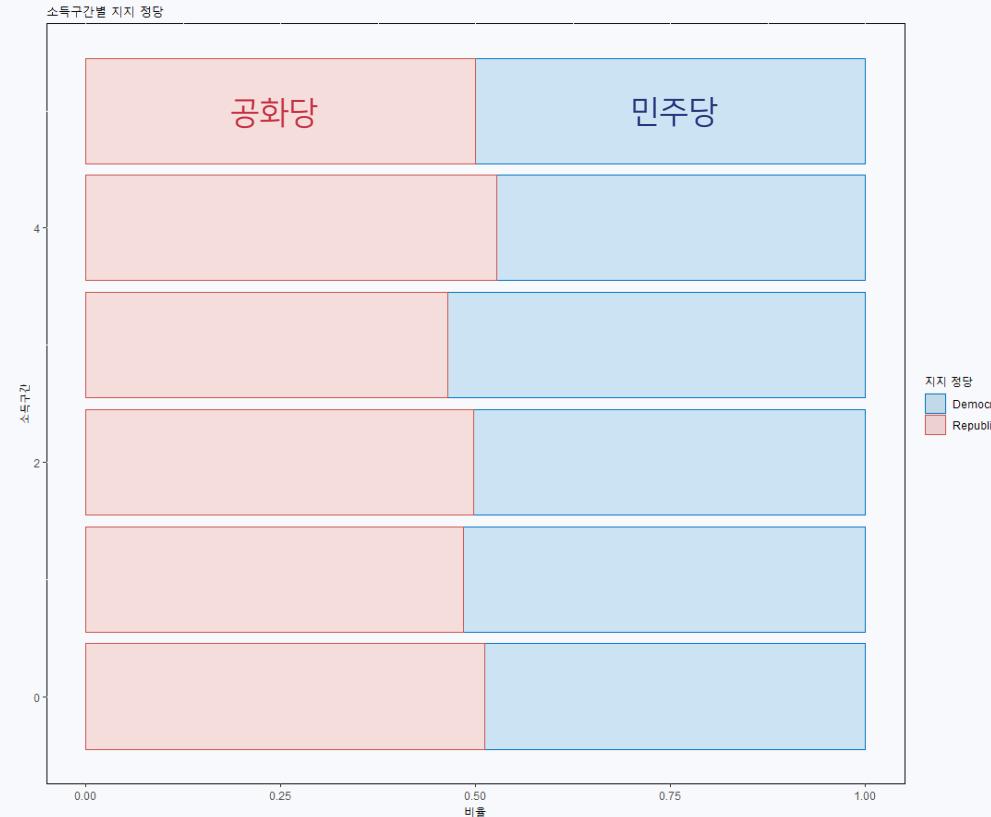
02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

3. Income 변수



소득구간 별 지지 정당 비율
거의 비슷!

<소득분위 별 지지 정당 비율>

01.
주제 선정 배경

02.
DATA

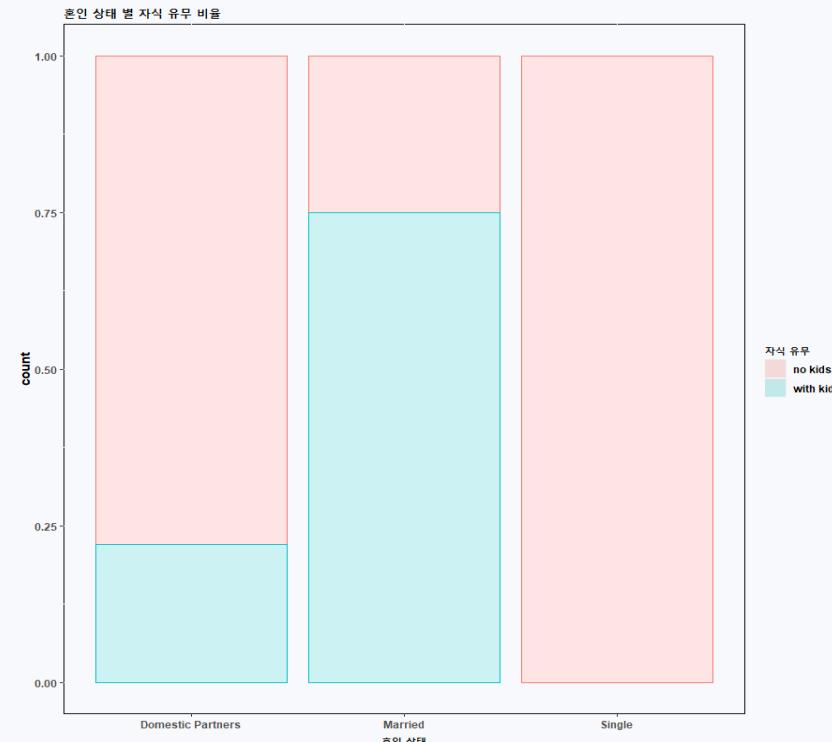
03.
시각화

04.
결측치 처리

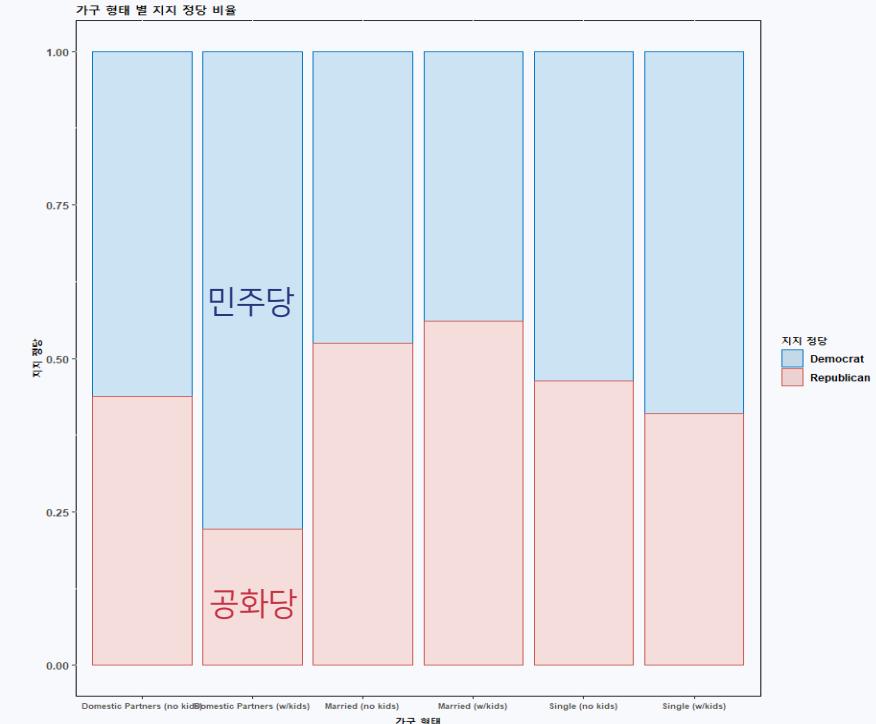
05.
3주차 예고

4. Marriage & kids 변수

<혼인 상태 별 자식 유무 비율>



<가구 형태별 지지 정당 비율>



아이가 있는 동거 커플 가구는
민주당 비율이 조금 높은 편

01.
주제 선정 배경

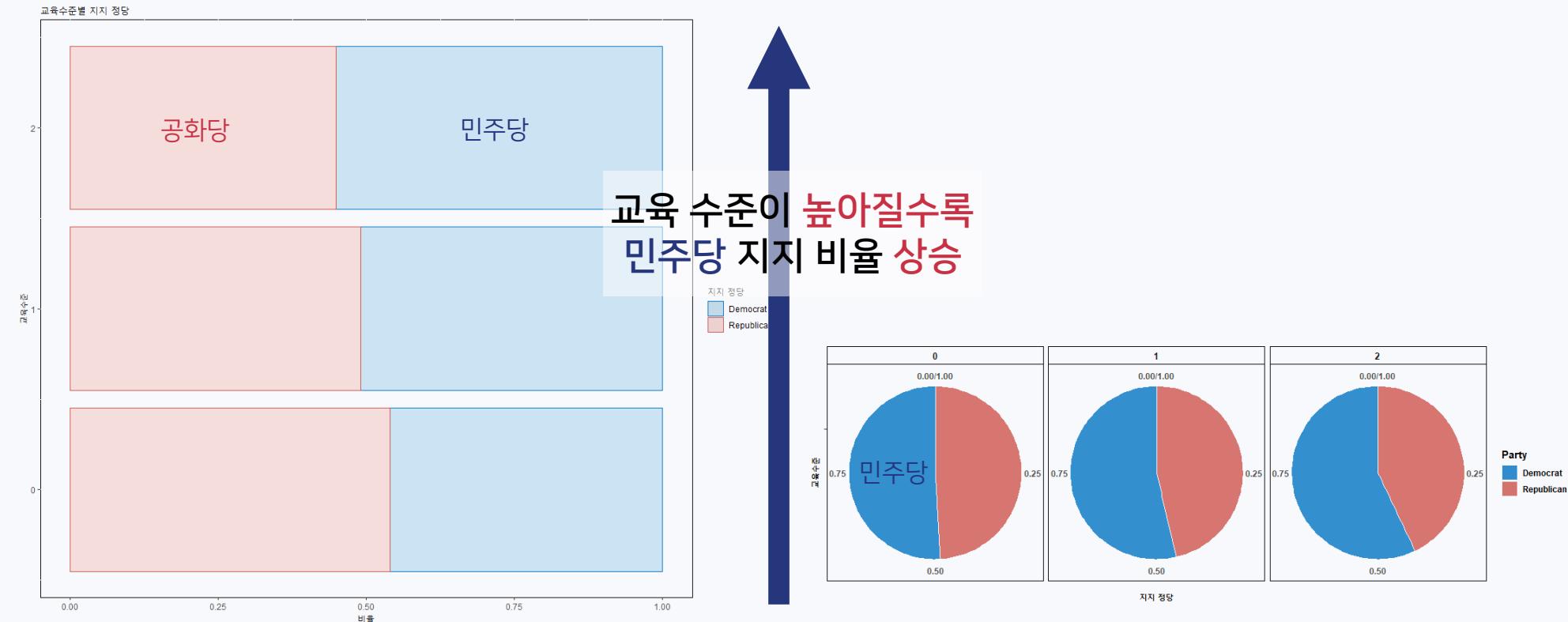
02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

5. Education Level 변수





- 우리의 인식은 실제로 유의할까?

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

Republicans

- 백인 남성/노인
- 자유경제
- 높은 소득 수준
- 총기 규제 완화
- 기업과 개인의 자유
- 현실적

Democrats

- 여성/성 소수자
- 사회경제적 평등
- 낮은 소득 수준
- 총기 규제 강화
- 노조 권리 보장
- 이상적

<미국 공화당과 민주당에 대한 일반적 인식>

- 우리의 인식은 실제로 유의할까?

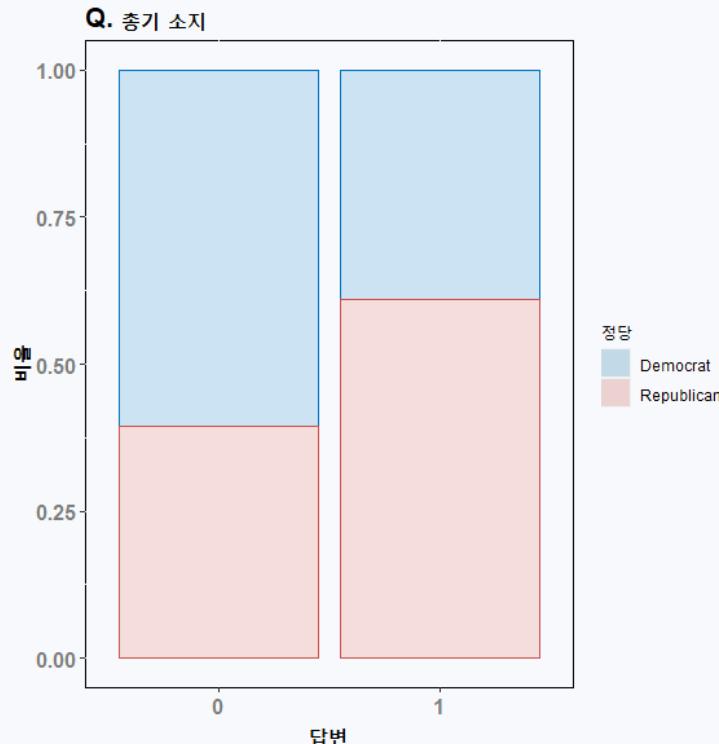
01.
주제 선정 배경

02.
DATA

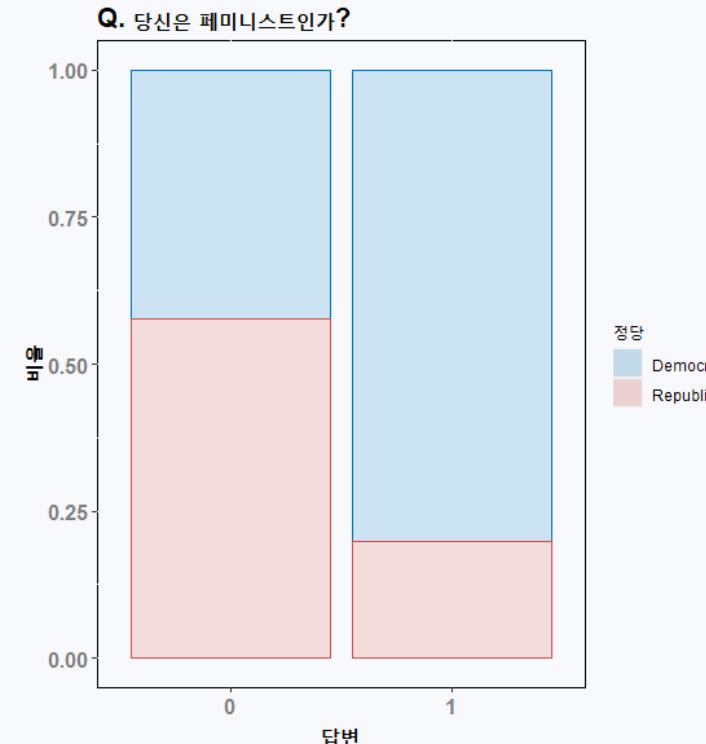
03.
시각화

04.
결측치 처리

05.
3주차 예고



총기를 소지한 응답자 중 약 70%가 **공화당**



페미니스트 중 약 80%가 **민주당**



- 우리의 인식은 실제로 유의할까?

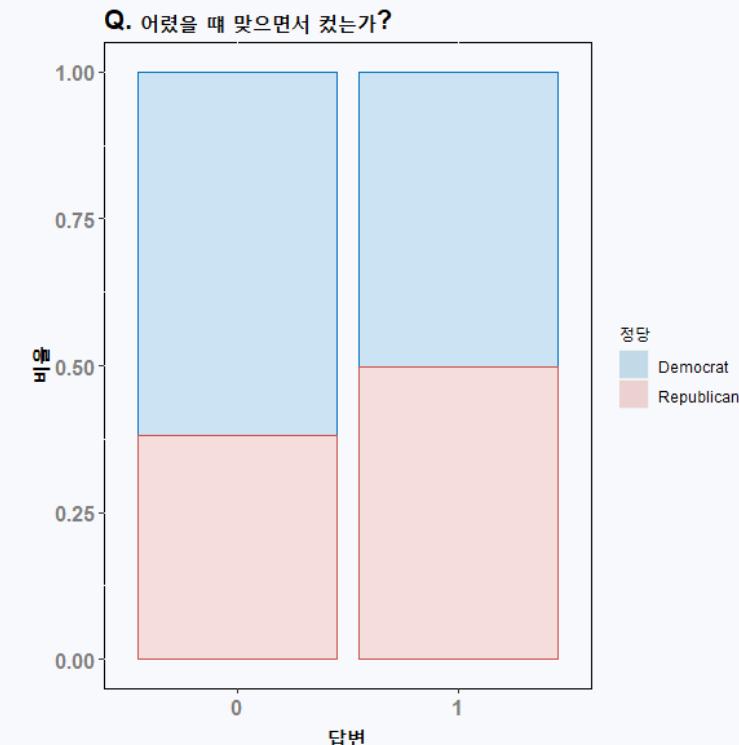
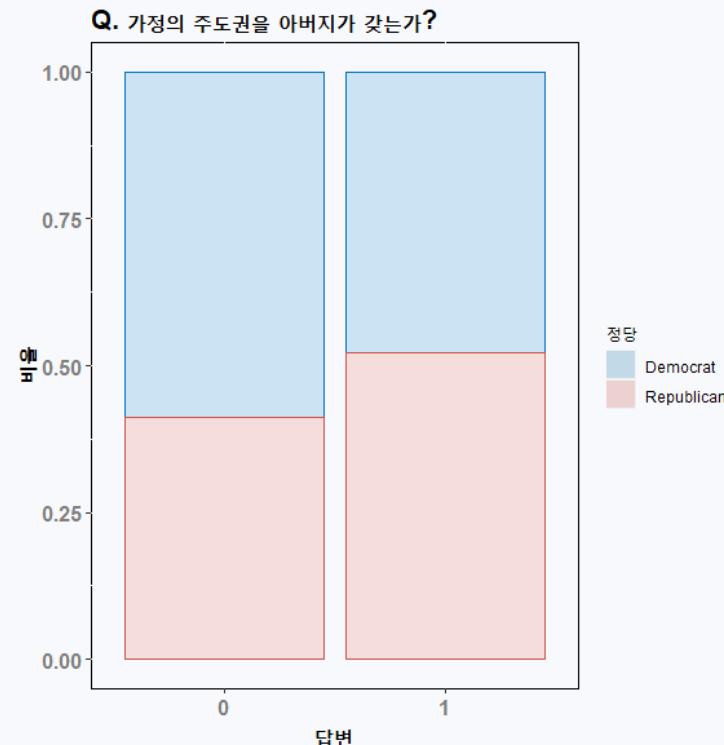
01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고



가정이 가부장적이지 않고 엄격하지 않은 응답자 중 민주당 비율이 더 높음



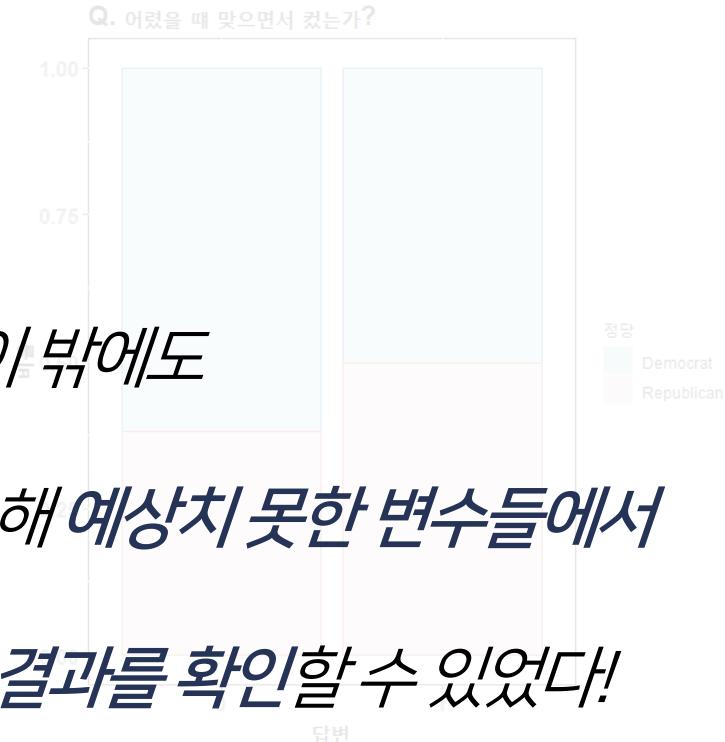
01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고



이 밖에도 문항별 시각화를 통해 예상치 못한 변수들에서

상당히 흥미로운 결과를 확인할 수 있었다!

가정이 가부장적/엄격하지 않은 응답자 중 민주당 비율이 더 높음

- 그 외에 흥미로웠던 결과들..

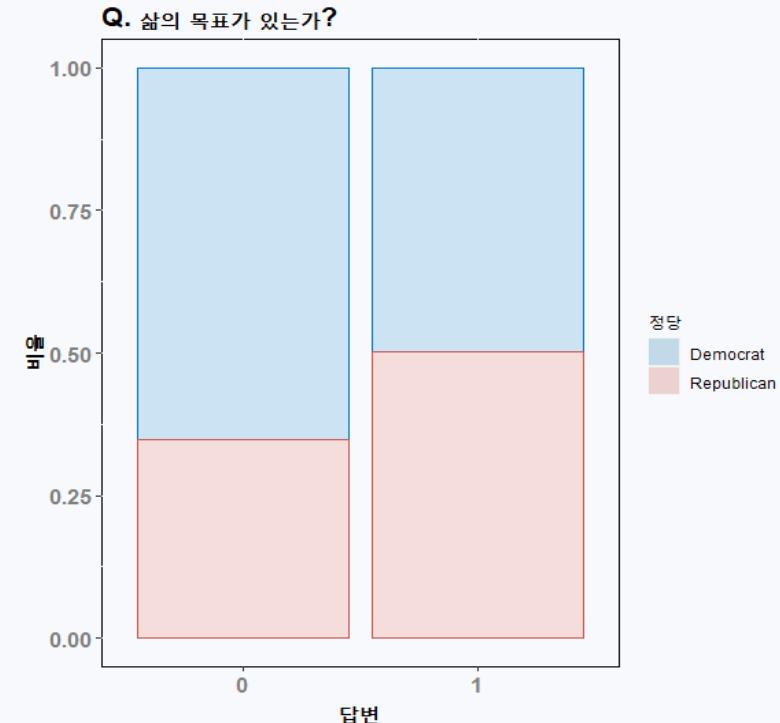
01.
주제 선정 배경

02.
DATA

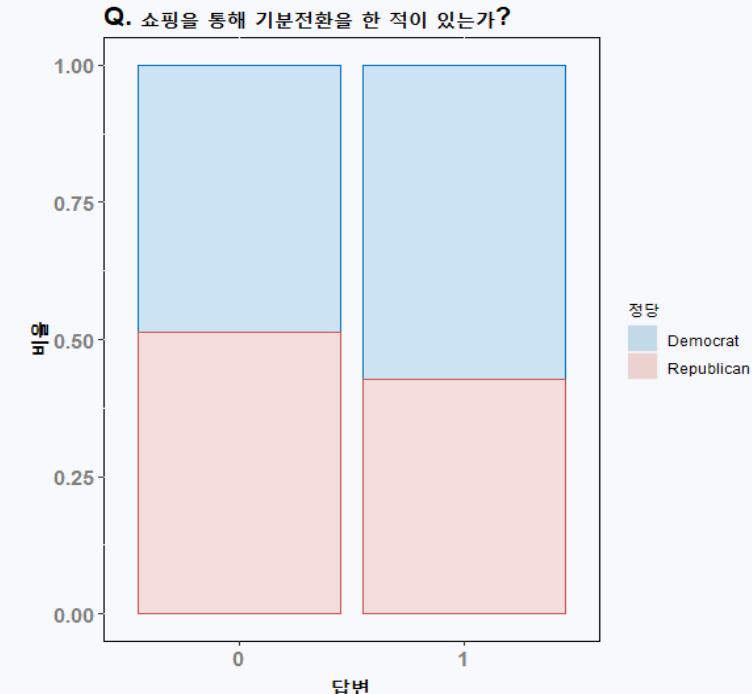
03.
시각화

04.
결측치 처리

05.
3주차 예고



삶에 목적이 없다고 응답한 참여자
-> 70% 이상 민주당



쇼핑을 통해 기분전환을 한 적이 있는
-> 민주당 비율 높음

- 분할표 분석

01.
주제 선정 배경

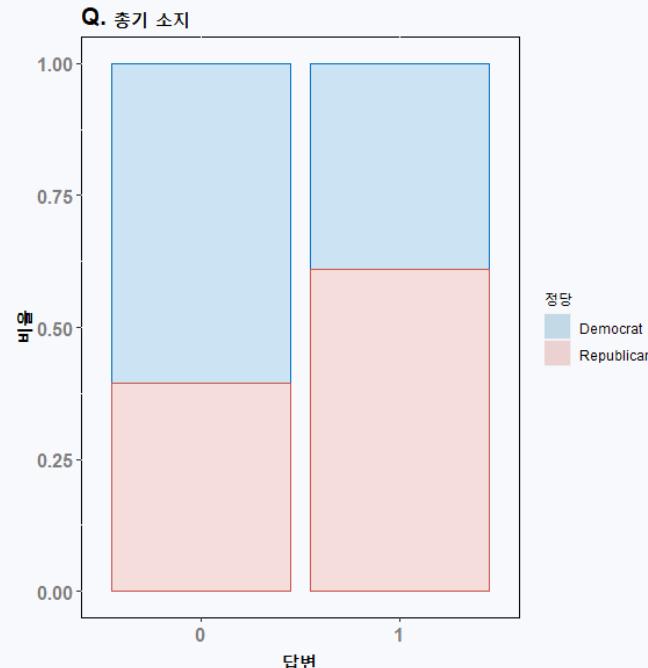
02.
DATA

03.
시각화

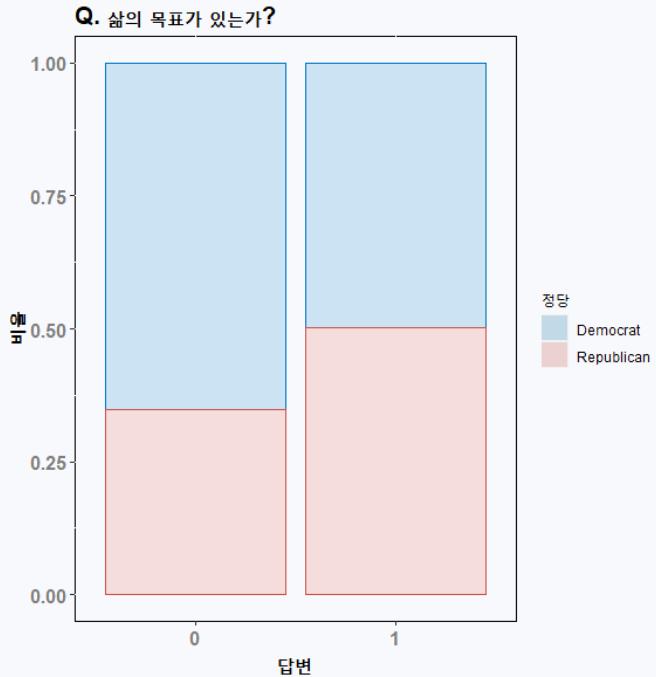
04.
결측치 처리

05.
3주차 예고

총기 소지 여부 응답 별 지지정당



삶의 목적 유무 별 지지정당



정당 별 유의미한 차이를 보이는 요인과 각 정당과의 관계를 파악해보자!



- “총기소지” 여부에 따른 지지 정당 분포 (crosstable)

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

<총기 소지 X vs 총기 소지 O>

result\$party	result\$have_gun		Row Total
	0	1	
Democrat	1394	506	1900
	25.937	46.150	
	0.734	0.266	0.527
	0.605	0.390	
	0.387	0.140	
Republican	912	790	1702
	28.954	51.519	
	0.536	0.464	0.473
	0.395	0.610	
	0.253	0.219	
Column Total		2306	1296
		0.640	0.360

정당	NO	YES
민주당	2252	1650
공화당	506	790



- “총기소지” 여부에 따른 지지 정당 분포 (crosstable)

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

		Outcome +	Outcome -	Total	Inc risk *	Odds
Exposed +	1394	912	2306	60.5	1.529	
Exposed -	506	790	1296	39.0	0.641	
Total	1900	1702	3602	52.7	1.116	

Point estimates and 95% CIs:

Inc risk ratio	1.55 (1.44, 1.67)
Odds ratio	2.39 (2.08, 2.74) → 오즈비 2.39
Attrib risk *	21.41 (18.09, 24.73)
Attrib risk in population *	13.71 (10.59, 16.82)
Attrib fraction in exposed (%)	35.41 (30.34, 40.12)
Attrib fraction in population (%)	25.98 (21.73, 30.00)

Test that OR = 1: chi2(1) = 152.559 Pr>chi2 = <0.001

민주당 총기소지X 오즈보가
공화당 총기소지X 오즈보다
139% 더 높다

총기소지하지 않을수록 민주당일 확률이 높다!



- “총기소지” 여부에 따른 지지 정당 분포 (crosstable)

총기소지하지 않을수록 **민주당일 확률이 높다!**

Democrats

- 여성/성소수자
- 사회경제적 평등
- 낮은 소득 수준
- **총기 규제 강화**
- 노조 권리 보장
- 이상적

앞서 봤던 민주당에 대한 인식에
부합하는 듯한 결론!



- “삶의 목적” 유무에 따른 지지 정당 분포 (crosstable)

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

<삶의 목표 X vs 삶의 목표 O>

result\$party	result\$have_life_purpose		Row Total
	0	1	
Democrat	474	1251	1725
	19.086	5.538	
	0.275	0.725	0.533
	0.651	0.499	
	0.146	0.386	
Republican	254	1258	1512
	21.774	6.318	
	0.168	0.832	0.467
	0.349	0.501	
	0.078	0.389	
Column Total		2509	3237
		0.775	
		0.225	

정당	NO	YES
민주당	474	1251
공화당	254	1258



- “삶의 목적” 유무에 따른 지지 정당 분포 (crosstable)

01.
주제 선정 배경02.
DATA03.
시각화04.
결측치 처리05.
3주차 예고

```
> epi.2by2(data, method="cohort.count", conf.level=0.95)
      Outcome +    Outcome -    Total    Inc risk *    odds
Exposed +        474        254     728       65.1   1.866 → 민주당 삶의 목표 X 오즈 추정값
Exposed -       1251       1258    2509       49.9   0.994 → 공화당 삶의 목표 X 오즈 추정값
Total          1725       1512    3237       53.3   1.141

Point estimates and 95% CIs:
-----
Inc risk ratio           1.31 (1.22, 1.40)
Odds ratio               1.88 (1.58, 2.23) → 오즈비 1.88
Attrib risk *             15.25 (11.27, 19.23)
Attrib risk in population * 3.43 (0.83, 6.03)
Attrib fraction in exposed (%) 23.42 (18.19, 28.32)
Attrib fraction in population (%) 6.44 (4.68, 8.16)
-----
Test that OR = 1: chi2(1) = 52.716 Pr>chi2 = <0.001
```

민주당 삶의 목적X 오즈가
공화당 삶의 목적X 오즈보다
88% 더 높다

삶의 목적이 없다고 응답했을 때 민주당일 확률이 높다!

- 질문 간 연관성 알아보기(GoodmanKruskal package)

- ## 01. 주제 선정 배경

- ## 02. DATA

- ## 03. 시각화

- ## 04. 결측치 처리

- ## 05. 3주차 예고

A & A' (Gktau measure = 0.64)

A': mo_Q_fulltimejob

→ 정규직에 종사하고 있는가?

A: mo_Q_minwage_job

→ 월급이 최저임금 이상인가?

B & B' (Gktau measure = 0.54)

B: Life_Q_watchTV

→ 하루에 TV를 일정시간 이상 시청하는가?

B':Life_Q_livealone

→ 혼자 살고 있는가?

표시된 몇 개의 질문 이외에,
유의한 연관성은 대체로 나타나지 않는다!

01. 주제 선정 배경

02. DATA

03. 시각화

04. 결측치 처리

05. 3주차 예고

GoodmanKruskal..? 좋은..사람? (Gktau measure = 0.64)

A mo_Q_fulltimejob
mo_Q_minwage_job

a.k.a. Gktau는 명목형(Nominal)변수 간의 연관성을 파악하는데 사용되는 측도이다!

A': mo_Q_fulltimejob

→ 정규직에 종사하고 있는가?

A: mo_Q_minwage_job

→ 월급이 최저임금 이상인가?

$$\tau(x, y) = \frac{\sum_{i=1}^K \sum_{j=1}^L \left(\frac{\pi_{ij}^2 - \pi_{i+}^2 \pi_{+j}^2}{\pi_{+j}} \right)}{1 - \sum_{j=1}^L \pi_{+j}^2}$$

B & B' (Gktau measure = 0.54)

B: Life_Q_watchTV
→ 하루 몇 시간을 보는가?

where $0 \leq \tau(x, y) \leq 1$

R': life_Q_livealone
→ 혼자 살고 있는가?

연속형 변수에 사용되는 피어슨 상관계수와 다르게
비대칭(asymmetric)이다! 즉, $\tau(x, y) \neq \tau(y, x)$

표시된 몇
유의한 상
관련을
나타나지
않는다.

GoodmanKruskal package로 질문 간 상관관계 표현
(With Gktau, CDA GREAT AGAIN!)





04

결속지 처리





- NA의 종류

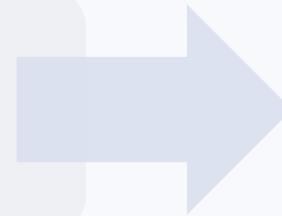
MCAR

Missing Completely at Random

결측이 **랜덤**으로 발생



해당 변수나 다른 변수에 영향 받지 않음



결측값 제외해도
분석 결과 편향 X

MAR

Missing at Random

결측 여부가 **다른 변수와 연관** 있음

ex) 교육 수준이 낮은 사람들이 소득 수준에 무응답



결측값 제외하면
분석 결과 **편향**

MNAR

Missing Not at Random

결측 여부가 **해당 변수의 값에 의해** 결정

ex) 교육 수준이 낮은 사람들이 교육 수준 무응답

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고



- NA의 종류
그렇다면 우리가 가진 설문지의 NA는 어떤 종류에 속할까?

MCAR Missing Completely at Random

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

설문지 구성 방식
결측이 랜덤으로 발생

하나의 설문지가 아닌
개별 응답 설문들을
USER_ID 기준으로 병합

결측 여부가 다른 변수와 연관 있음

질문 간 연관성

대체로 질문 간의
연관성이 낮음

결측 여부가 해당 변수의 값에 의해 결

ex) 교육 수준이 낮은 사람들이 교육 수준 무응답

결측값 제외해도
의도적인 무응답 가능성 ↓ 결과 편향 X
OR

질문 간의 연계로 인한 무응답 가능성 ↓
Ex) 문항 3을 예라고 대답했을 경우, 문항 4-1로 가시오.

결측값 제외하면
분석 결과 편향

MCAR에 속한다고 가정하고 진행!



• NA 처리방법

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

Likewise Deletion
(Compete case
analysis)

Pairwise Deletion
(available case
analysis)

Imputation
(intent to treat
analysis)

- 결측 값 행 삭제
- MCAR일 때만 사용 가능
- 표본수가 지나치게 감소
- > 분석의 정확도가 급감
- 결측자료 짹 별로 삭제
- MCAR일 때만 사용 가능
- 표본수 감소 비교적 적음
- 결측치 특정 값 대체 후
분석 진행



• NA 처리방법

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

Likewise Deletion
(Compete case analysis)

- 결측 값 행 삭제
- MCAR일 때만 사용 가능
- **표본수가 지나치게 감소**
-> 안그래도 데이터 적은데
너무나 치명적... 절대 안돼

Pairwise Deletion
(available case analysis)

- 결측자료 짹 별로 **삭제**
- MCAR일 때만 사용 가능
- 표본수 감소 비교적 적음
-> 감소 적어도 삭제는 삭제
데이터 하나 하나 소홀해...



Imputation
(intent to treat analysis)

- 결측치 특정 값 대체 후
분석 진행



- 결측치 비율 별 NA 처리방법

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

결측치 비율	처리방법
10%미만	제거 or 원하는 Imputation
10%이상 20% 미만	Hot deck, regression, Model based method
20% 이상	Model based method, Regression

Multivariate Data Analysis, hair et al.(2006)



• 데이터 NA확인

01.
주제 선정 배경02.
DATA03.
시각화04.
결측치 처리05.
3주차 예고

USER_ID	Gender	Income	Education Level	...	env_Q_p_fight.ifoU	mo_Q_minwage_job	mo_Q_minwage_job	...	care	...	sanctity
1	Male	2	1		1	1	1		0		0
4	Female	5	1		NA	1	1		0		0
5	Male	3	0		1	0		NA	1		0
8	Male	4	1		NA	NA	1		0		0
9	Female	2	0		1	0		0	0		0
10	Female	5	0		NA	1	1		1		0
11	Male	1	1		1	1		NA	1		1
12	Male	3	0		0	NA		0	0		1
13	Female	3	0		NA	1	1		1		0
14	Male	1	1		1	1		NA	1		1
15	Male	3	0		0	NA		0	0		1

설문조사 데이터 특성상 수많은 NA값이 존재, NA의 비율이 30% 이상



- NA 비율 별 처리방법

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

결측치 비율	처리방법
10%미만	제거 or 원하는 Imputation
10%이상 20% 미만	Hot deck, regression, Model based method
20% 이상	Model based method, regression





- MICE란?

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

MICE

Multiple Imputation by Chained Equations

변수가 여러 개인 data set에 대하여 한 변수 내의 결측치를 **다른 변수들과의 관계를 고려**한 방정식에 의해 대입하는 결측치 처리 방식으로,
결측치에 대한 **대입이 여러 번** 이루어져 상대적으로 bias가 감소한다



- MICE란?

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

MICE

Multiple Imputation by Chained Equations

변수가 여러 개인 data set에 대하여 한 변수 내의 결측치를 **다른 변수들과의 관계를 고려**한 방정식에 의해 대입하는 결측치 처리 방식으로,
결측치에 대한 **대입이 여러 번** 이루어져 상대적으로 bias가 감소한다



- MICE란?

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

MICE

Multiple Imputation by Chained Equations

변수가 여러 개인 data set에 대하여 한 변수 내의 결측치를 다른 변수들과의 관계를 고려한 방정식에 의해 대입하는 결측치 처리 방식으로,
결측치에 대한 대입이 여러 번 이루어져 상대적으로 bias가 감소한다

01.
주제 선정 배경

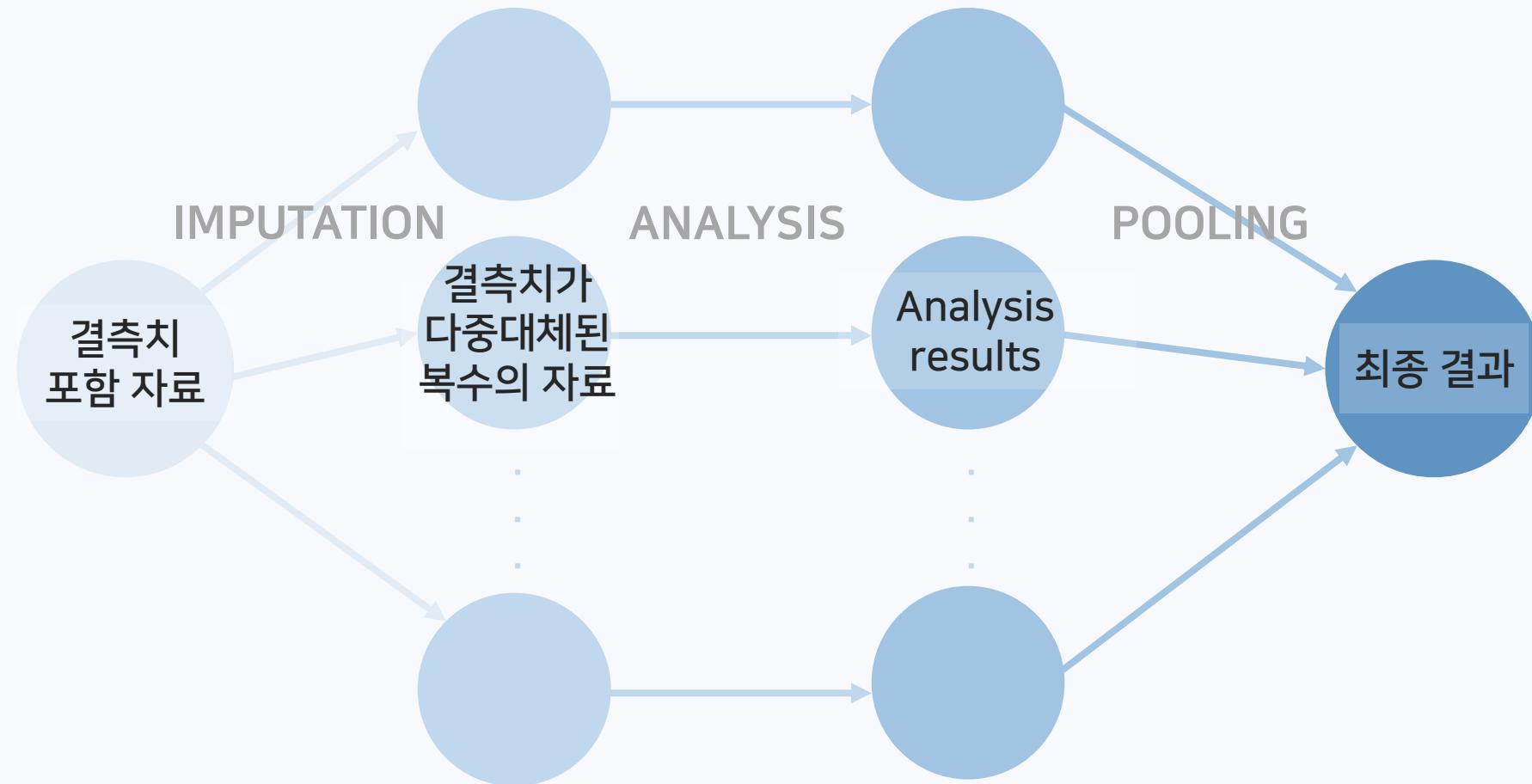
02.
DATA

03.
시각화

04.
결측치 처리

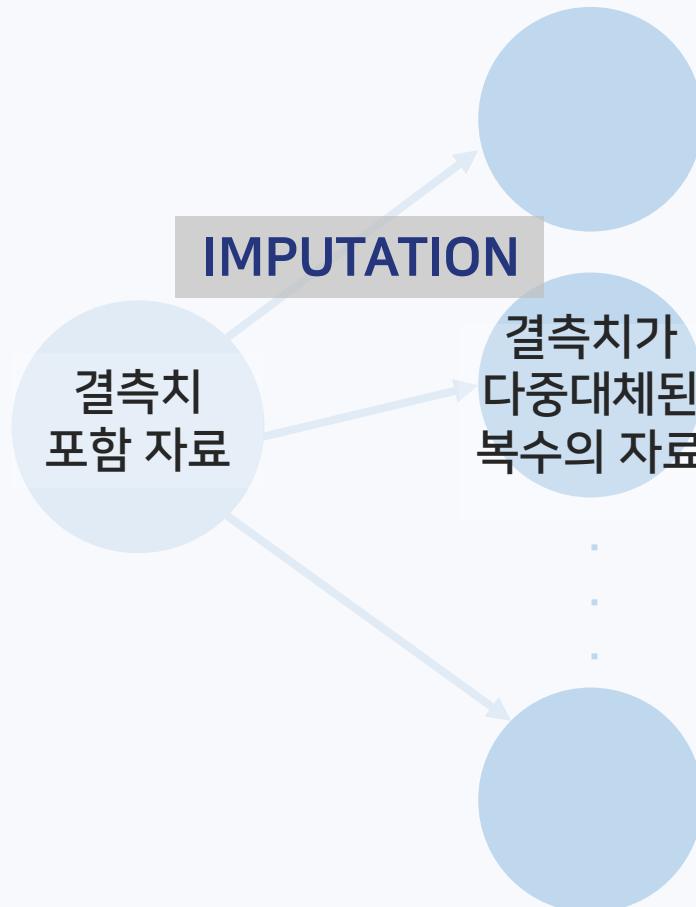
05.
3주차 예고

- MICE의 알고리즘





- MICE의 알고리즘



STEP1. IMPUTATION

다양한 통계적 기법을 이용하여
결측치 대치세트 m개 생성
default = 5

Method	Description	Scale type	Default
pmm	Predictive mean matching	numeric	Y
norm	Bayesian linear regression	numeric	
norm.nob	Linear regression, non-Bayesian	numeric	
mean	Unconditional mean imputation	numeric	
2L.norm	Two-level linear model	numeric	
logreg	Logistic regression	factor, 2 levels	Y
polyreg	Multinomial logit model	factor, >2 levels	Y
polr	Ordered logit model	ordered, >2 levels	Y
lda	Linear discriminant analysis	factor	
sample	Random sample from the observed data	any	

01.
주제 선정 배경

02.
DATA

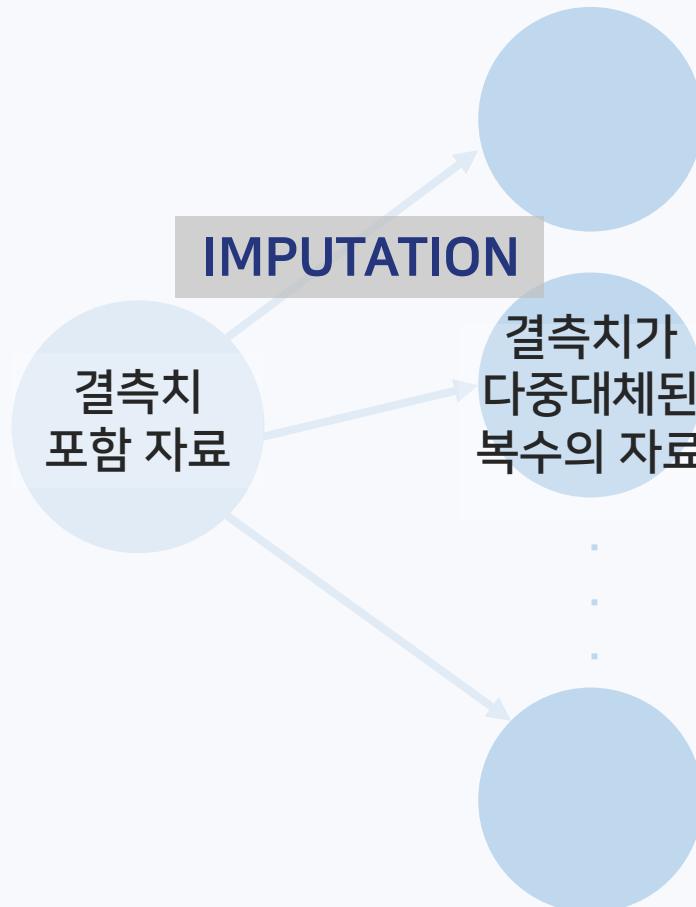
03.
시각화

04.
결측치 처리

05.
3주차 예고



- MICE의 알고리즘



STEP1. IMPUTATION

다양한 통계적 기법을 이용하여
결측치 대치세트 m개 생성
default = 5

Method	Description	Scale type	Default
pmm	Predictive mean matching	numeric	Y
norm	Bayesian linear regression	numeric	
norm.nob	Linear regression, non-Bayesian	numeric	
mean	Unconditional mean imputation	numeric	
2L.norm	Two-level linear model	numeric	
logreg	Logistic regression	factor, 2 levels	Y
polyreg	Multinomial logit model	factor, >2 levels	Y
polr	Ordered logit model	ordered, >2 levels	Y
lda	Linear discriminant analysis	factor	
sample	Random sample from the observed data	any	

연속형/범주형 변수를 **자동으로 인식**하여
적절한 imputation method를 적용해 줌!
method 변수를 통해 원하는 방법 직접 지정 가능!



- MICE의 알고리즘



STEP2. ANALYSIS

각각의 imputed dataset에서
별도의 통계분석 진행

01.
주제 선정 배경

02.
DATA

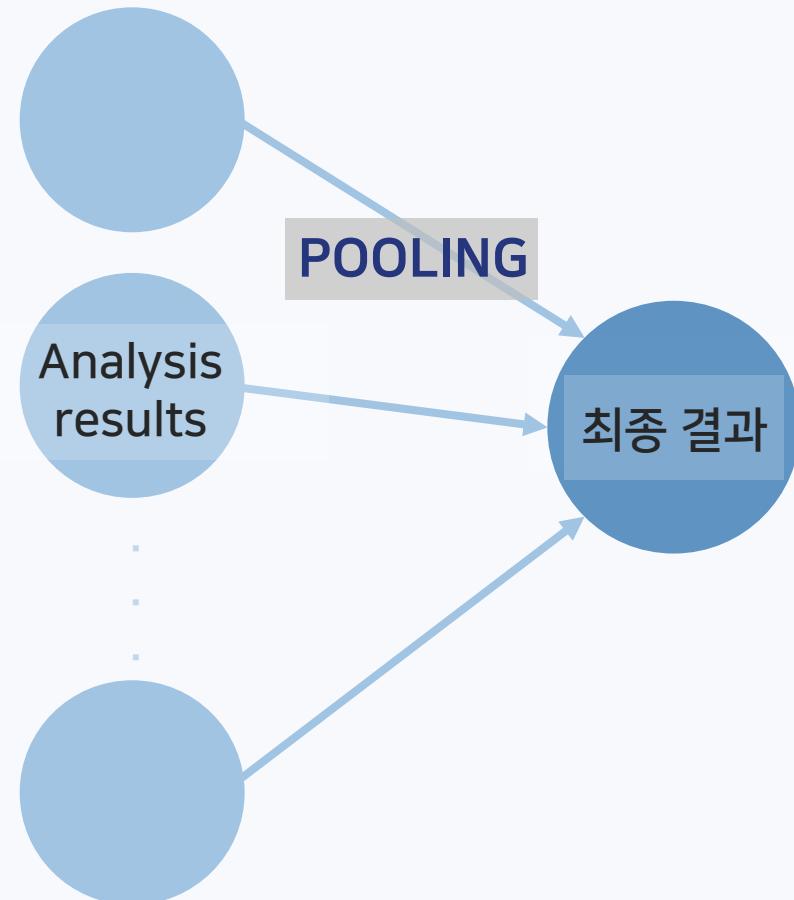
03.
시각화

04.
결측치 처리

05.
3주차 예고



- MICE의 알고리즘



STEP2. ANALYSIS

각각의 imputed dataset에서
별도의 통계분석 진행

STEP3. POOLING

Analysis를 통한 결과로
within/between variance를 통합하여
최종 결측값 도출

(2주차 패키지에 나온다는 소-문이..)

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고



MICE with NA

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

STEP1

설문지 구성 방식을 고려,
Missing Value를 무응답으로
가정하고 진행

STEP4

$M = 5$ 로 설정,
5개의 대치 셋의 평균을
대치 값으로 선택!

STEP2

Ordinal Encoding 진행한
Income 및 EducationLevel
순서형 로짓 모형(polar) 적용

STEP3

그 외의 인적변수와
모든 설문 문항에 대해
로지스틱 회귀 모형(logreg) 적용



MICE with Mean Imputation

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

STEP1

Ordinal Encoding 진행한
변수의 평균값 반올림하여 대치



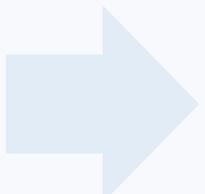
STEP4

$M = 5$ 로 설정,
5개의 대치 셋의 평균을
대치 값으로 선택!



STEP2

나머지 인적변수 및
모든 설문 문항의 평균값을
 $cutoff\ point = 0.5$ 를
기준으로 대치



STEP3

위의 MICE with NA와
동일한 로지스틱 모델 적용



Imputation 변수 선택 알고리즘

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

STEP1

모든 변수들 간의 Gktau measure 계산하여 0.5가 넘는 15개의 조합 추출

	v1	v2	v3	v4
Comb1	Life_Q_collectHobby	mo_Q_has_enoughcash_now	edu_Q_publicschool	ps_Q_Optimist
Comb2	Life_Q_watchTV	env_Q_p_spank	Life_Q_livealone	ps_Q_LeftHanded
Comb3	ps_Q_GoodLiar	mo_Q_has_enoughcash_now	Life_Q_collectHobby	
Comb4	mo_Q_fulltimejob	mo_Q_minwage_job		
		...		
Comb14	ps_Q_PowerOfPositive	edu_Q_parents_college	env_Q_single_parent	
Comb15	ps_Q_Creative	re_Q_havesibling		

그러나.. 각 조합의 변수의 수가 너무 적다..!



01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

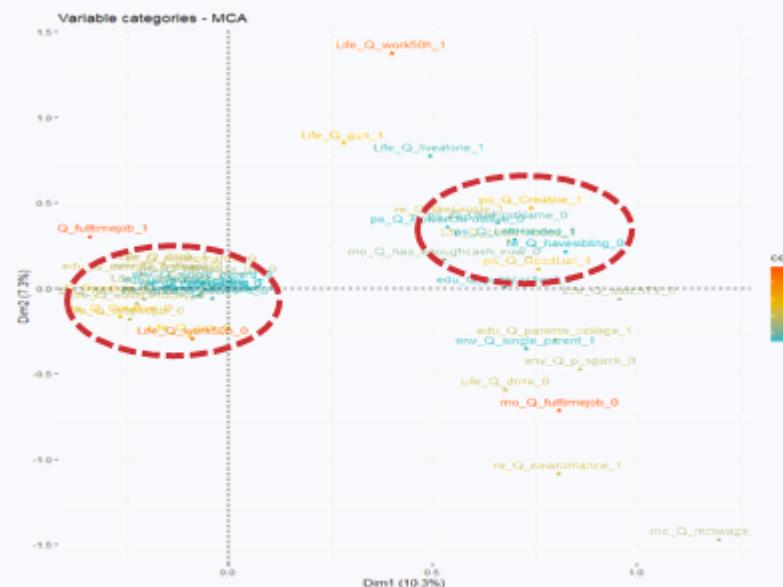
05.
3주차 예고

Imputation 변수 선택 알고리즘

STEP2

위에서 선택되지 않은 변수 중, MCA Biplot(상관도) 참고하여
각 조합이 15개의 변수가 되도록 선택

MICE의 제작자 Van Buren said, 변수의 개수 is 15개~25개가 적-절..



	V1	V15(추가)
Comb1	Life_Q_collectHobby	re_Q_havesibling
Comb2	Life_Q_watchTV	edu_Q_publicschool
Comb3	ps_Q_GoodLiar	re_Q_likepeople
		...
Comb14	env_Q_single_parent	re_Q_newromance
Comb15	ps_Q_Creative	mo_Q_carpayment



- MCA란?

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

MCA

Multiple Correspondence Analysis

3개 이상의 범주형 자료를 가진 data 내
모든 category 간의 관계를 확인하는 분석 방법으로,
Mapping을 통해 연관성의 내용을 시각화해 알아보기 쉽다!



- MCA의 알고리즘

STEP1. Burt Table

X 변수들의 각 변량들의 수준에 대한 교차빈도표

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

	Burt Table																		
	'America'	'EU'	'Japan'	'Large'	'Medium'	'Small'	'Family'	'Sporty'	'Work'	'Own'	'Rent'	'High'	'Low'	'Married'	'SingleWithKids'	'Single'	'Female'	'Male'	
'America'	125	0	0	36	60	29	81	24	20	93	32	67	58	37	56	32	58	67	
'EU'	0	44	0	4	20	20	17	23	4	38	6	26	18	13	16	15	21	23	
'Japan'	0	0	165	2	61	102	76	59	30	111	54	91	74	51	52	62	70	95	
'Large'	36	4	2	42	0	0	30	1	11	35	7	22	20	9	22	11	17	25	
'Medium'	60	20	61	0	141	0	89	39	13	106	35	84	57	42	59	40	70	71	
'Small'	29	20	102	0	0	151	55	66	30	101	50	78	73	50	43	58	62	89	
'Family'	81	17	76	30	89	55	174	0	0	130	44	105	69	50	89	35	83	91	
'Sporty'	24	23	59	1	39	66	0	106	0	71	35	51	55	35	14	57	44	62	
'Work'	20	4	30	11	13	30	0	0	54	41	13	28	26	16	21	17	22	32	
'Own'	93	38	111	35	106	101	130	71	41	242	0	162	80	76	114	52	114	128	
'Rent'	32	6	54	7	35	50	44	35	13	0	92	22	70	25	10	57	35	57	
'High'	67	26	91	22	84	78	105	51	28	162	22	184	0	91	83	10	102	82	
'Low'	58	18	74	20	57	73	69	55	26	80	70	0	150	10	41	99	47	103	
'Married'	37	13	51	9	42	50	50	35	16	76	25	91	10	101	0	0	53	48	
'SingleWithKids'	56	16	52	22	59	43	89	14	21	114	10	83	41	0	124	0	61	63	
'Single'	32	15	62	11	40	58	35	57	17	52	57	10	99	0	0	109	35	74	
'Female'	58	21	70	17	70	62	83	44	22	114	35	102	47	53	61	35	149	0	
'Male'	67	23	95	25	71	89	91	62	32	128	57	82	103	48	63	74	0	185	

- MCA의 알고리즘

STEP2. Inertia and Chi-Square Decomposition

Inertia and Chi-Square Decomposition									
Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	4	8	12	16	20
0.56510	0.31933	954.80	20.32	20.32	*****				
0.48291	0.23320	697.27	14.84	35.16	*****				
0.41452	0.17183	513.76	10.93	46.10	*****				
0.39856	0.15885	474.97	10.11	56.20	*****				
0.38745	0.15012	448.86	9.55	65.76	*****				
0.35866	0.12864	384.63	8.19	73.94	*****				
0.34063	0.11603	346.92	7.38	81.33	*****				
0.32080	0.10291	307.70	6.55	87.88	*****				
0.28274	0.07994	239.03	5.09	92.96	*****				
0.26127	0.06826	204.11	4.34	97.31	****				
0.20569	0.04231	126.51	2.69	100.00	***				
Total	1.57143	4698.56	100.00						
Degrees of Freedom = 289									

Burt table에서
카이제곱 값을
metric으로 하여
카테고리 간의 variation
(= inertia)을
설명할 수 있다!

Category Coordinates		
	Dim1	Dim2
'EU'	0.1490	0.8040
'Large'	-0.0263	-0.5329
'Med'	0.3244	-0.4570
'Small'	-0.6890	1.5879
'Family'	-0.2745	0.0740
'Sporty'	0.4479	-0.5108
'Work'	-0.4359	0.3418
'Own'	0.6873	-0.6439
'Rent'	0.0556	0.1625
'High'	-0.3767	-0.0917
'Low'	0.9909	0.2412
'Married'	-0.6513	-0.4504
'SingleWithKids'	0.7989	0.5525
'Single'	-0.4176	-0.8147
'Female'	-0.6879	0.3696
'Male'	1.1695	0.3344
	-0.3821	-0.2551
	0.3077	0.2055



(혹시가 역사가 되어버린 순간..)

Biplot(상관도)를 그려
각 변량들의 상관도를
상대적 거리로 판단 가능!



Regularized Iterative MCA using imputeMCA function

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

범주형 자료의 결측치를 처리하는 데 사용 가능하다.

Algorithm

	V1_a	V1_b	V1_c	V2_e	V2_f
Ind1	1	0	0	0.41	0.59
Ind2	0.2	0.3	0.5	0	1
Ind3	1	0	1	1	0
Ind4	0	1	1	0	0
Ind5	0	1	1	0	1
Ind6	0	1	1	1	1

...

Initialization

카테고리별 결측치를
비율로 초기화



Regularized Iterative MCA using imputeMCA function

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

범주형 자료의 결측치를 처리하는 데 사용 가능

Algorithm

	V1_a	V1_b	V1_c	V2_e	V2_f
Ind1	1	0	0	0.65	0.35
Ind2	0.11	0.2	0.69	0	1
Ind3	1	0	1	1	0
Ind4	0	1	1	0	0
Ind5	0	1	1	0	1
Ind6	0	1	1	1	1

...

Estimation
& Imputation

Until convergence



01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

Regularized Iterative MCA

Overfitting 문제로부터 자유롭다 !

설문 문항
(Binary Data) + 인적사항 변수
(ex. Income, Age)

ImputeMCA

MICE



“최종 데이터 분포 비교”

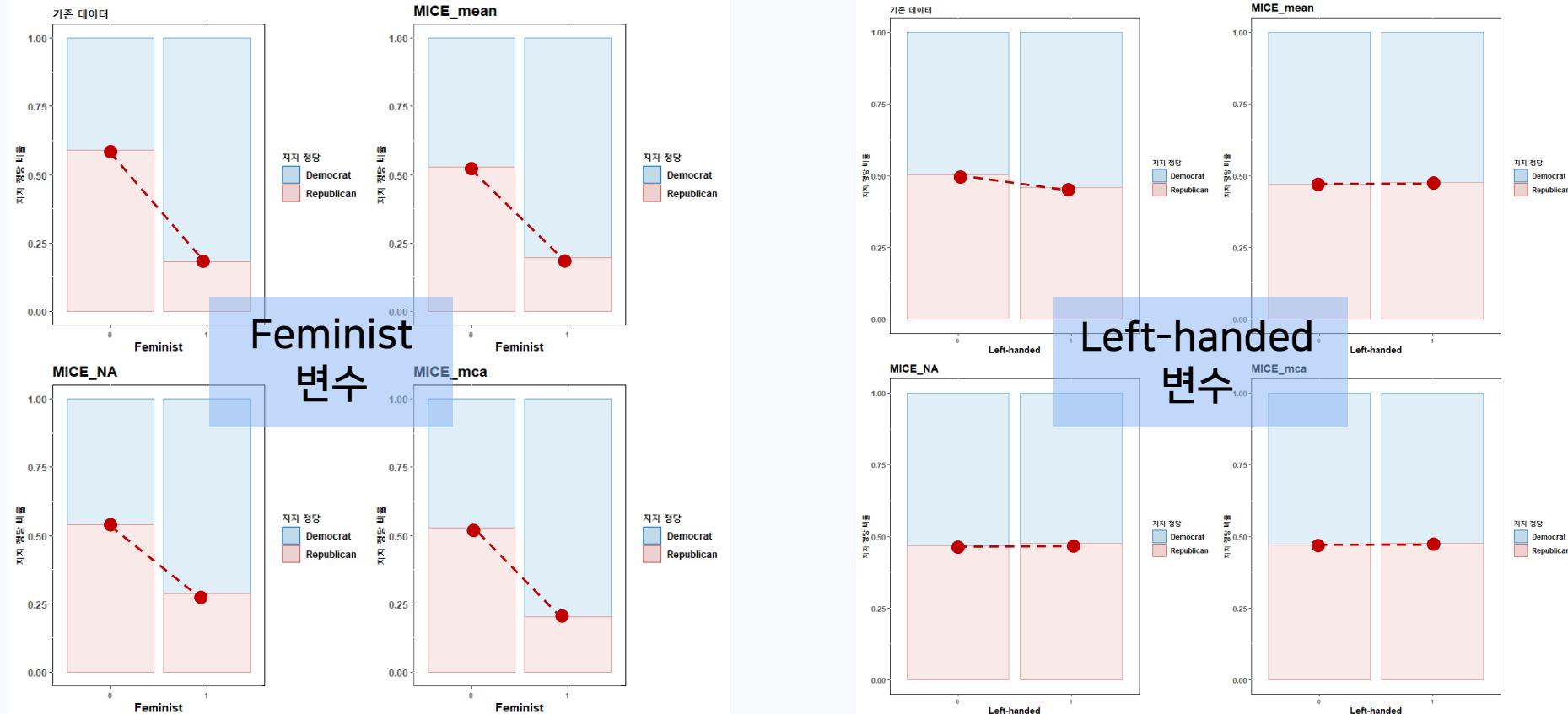
01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고





“최종 데이터 분포 비교”

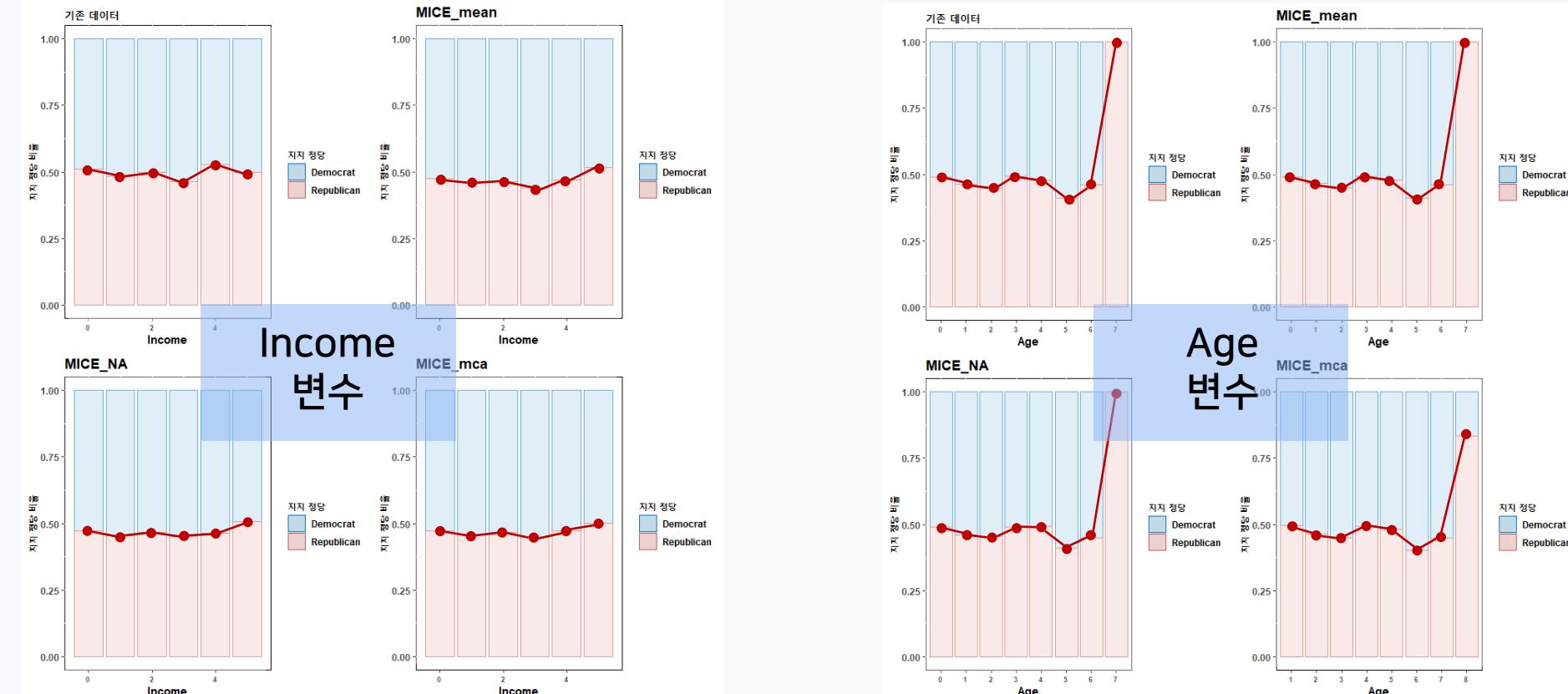
01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고





모두 비슷한 분포를 띤다!

The figure consists of two side-by-side line graphs. The left graph shows '지지 정당' (Supported Political Party) on the y-axis (0.0 to 1.0) against 'Age' on the x-axis (18 to 65). It compares 'Democrat' (blue line) and 'Republican' (red line) support. Both show a similar trend: low support in youth, peaking around 25-30, dipping slightly, then rising sharply after 55. The right graph shows 'N_E_NA' on the y-axis (0.0 to 0.5) against 'Age' on the x-axis (18 to 65). It also compares Democrat (blue line) and Republican (red line) support, showing a very similar pattern to the first graph.

- 01. 주제 선정 배경
 - 02. DATA
 - 03. 시각화
 - 04. 결측치 처리
 - 05. 3주차 예고



05

3주자 예고



데이터 박살 예고



3주차에는...

1. 새로운 파생 변수



씨와 함께 한 파생변수 만들기 외

설문 문항과 다양한
도메인 지식 활용해서
추가적인 파생변수 생성

2. 추가 NA Imputation

MICE

~~하다가 찬영 수정 힘들어 죽을뻔했지만^^~~
~~우리는 멈추지 않는다~~

MCA

두 가지 방법 이외에도
새로운 NA Imputation
방법 논의

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

데이터 박살 예고



3. MCA 해석 및 차원 축소

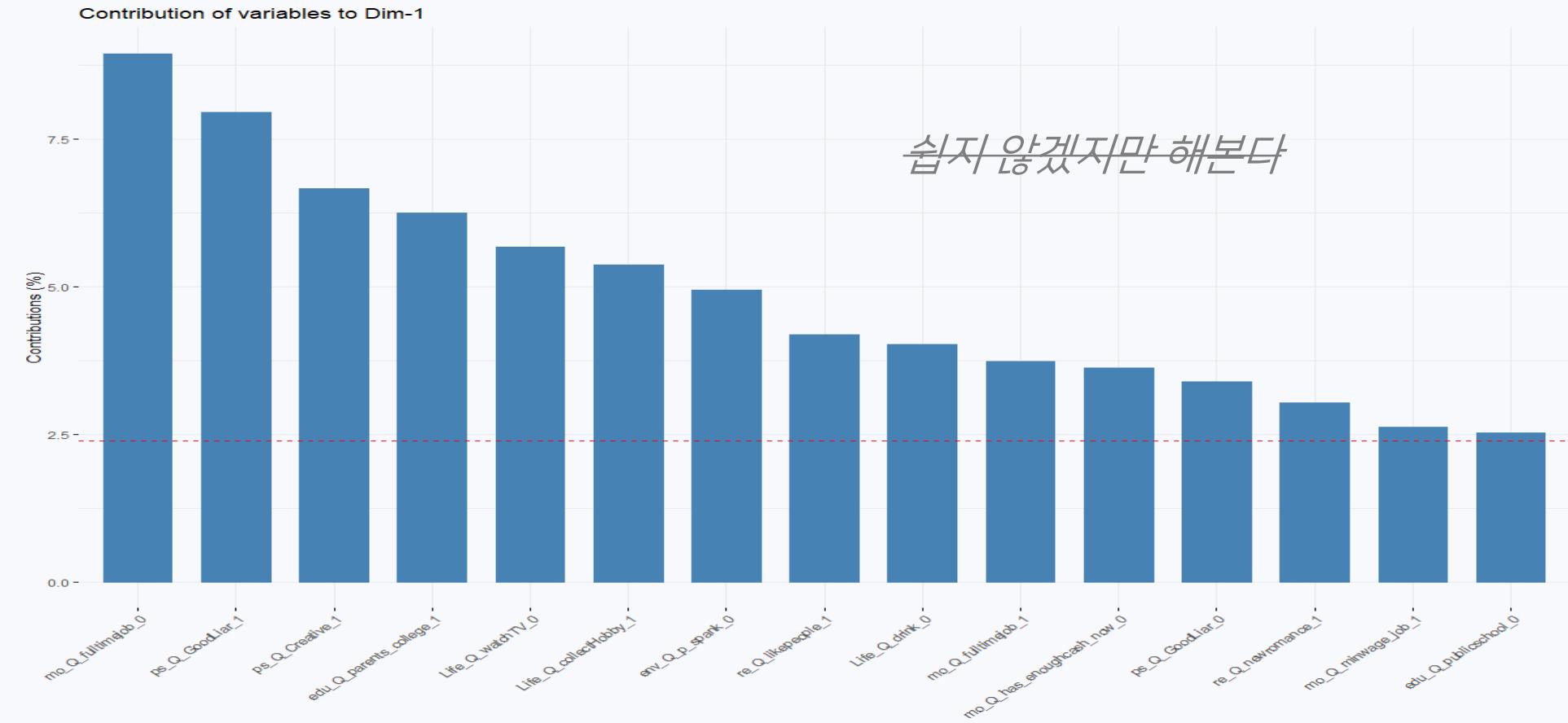
01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고



모델 박살 예고



예측도 해볼거야!

01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

NA → 무응답
으로 처리한
기준 데이터

NA → Mean
→ MICE
로 채운
데이터

NA → MICE
로 채운
데이터

NA →
MCA 이용한
MICE로 채운
데이터

4개의 데이터에 대해

*GLM · Random Forest
XGBoost · CATBoost*

등의 모델을 사용해
예측력 높이기!

모델 박살 예고(MIT 어서오고)



01.
주제 선정 배경

02.
DATA

03.
시각화

04.
결측치 처리

05.
3주차 예고

예측드 해보거니와

Accuracy 90% 실-화..? MIT는 다르다 이건가..

#	△pub	Team Name	Notebook	Team Members	Score
1	▲ 7	HugoSilveiradaCunha			0.92241
2	—	@mos			0.92097
3	▲ 1	Elie			0.91522
4	▲ 1	Keks			0.91379
5	▲ 1	쩝쩝 (쩝~쩝)▶			0.90948



다 이겨보자..!!!!!!

끄-을