설문조사를 통한 지지정당 예측

행복 범주형자료분석팀

김찬영 이혜인 김서윤 심은주 진수정

CONTENTS

주제 소개

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!

- 설문조사 데이터 (출처: Show of Hands)

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

⋮

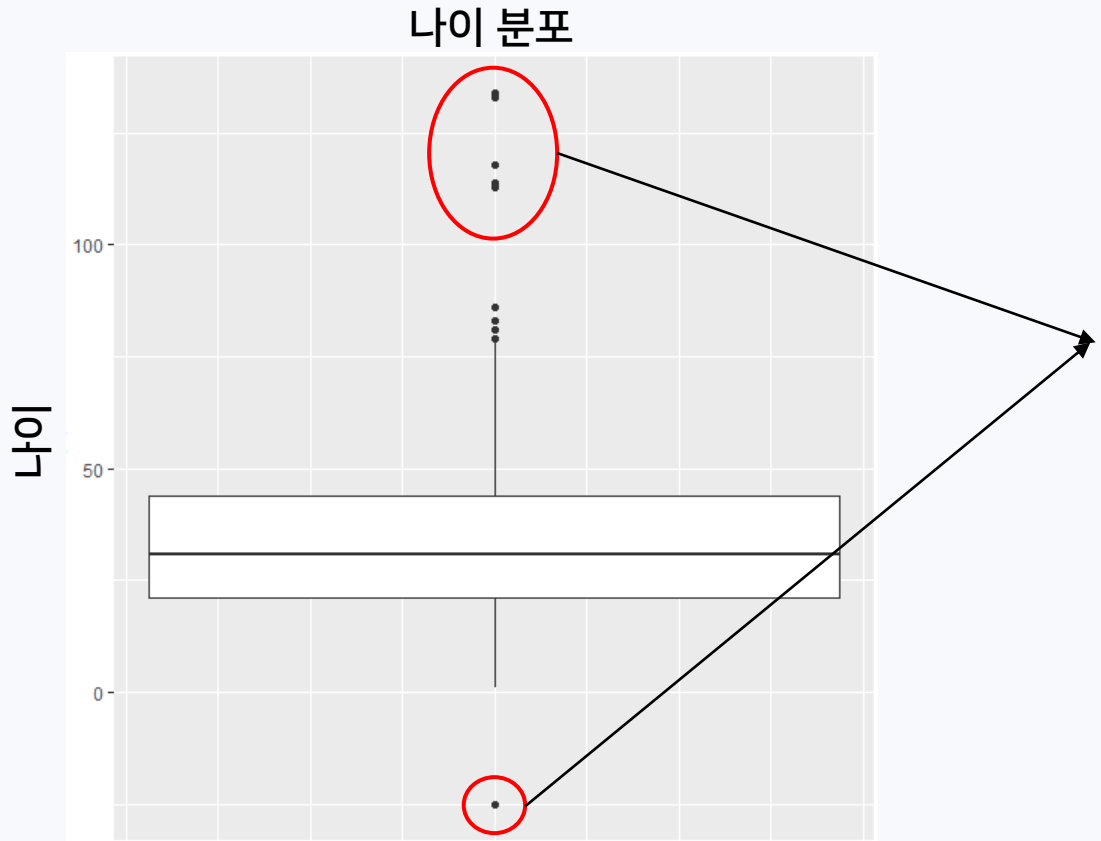
- **코드북 작성**

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

Ori_Q_ID	Fin_Q_ID	문제	Category	가능한 답변	바뀐 답변
Q98059	re_Q_havesibling	형제 자매 있음?	Relationships	Y/Only-Child	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



<나이 box plot>

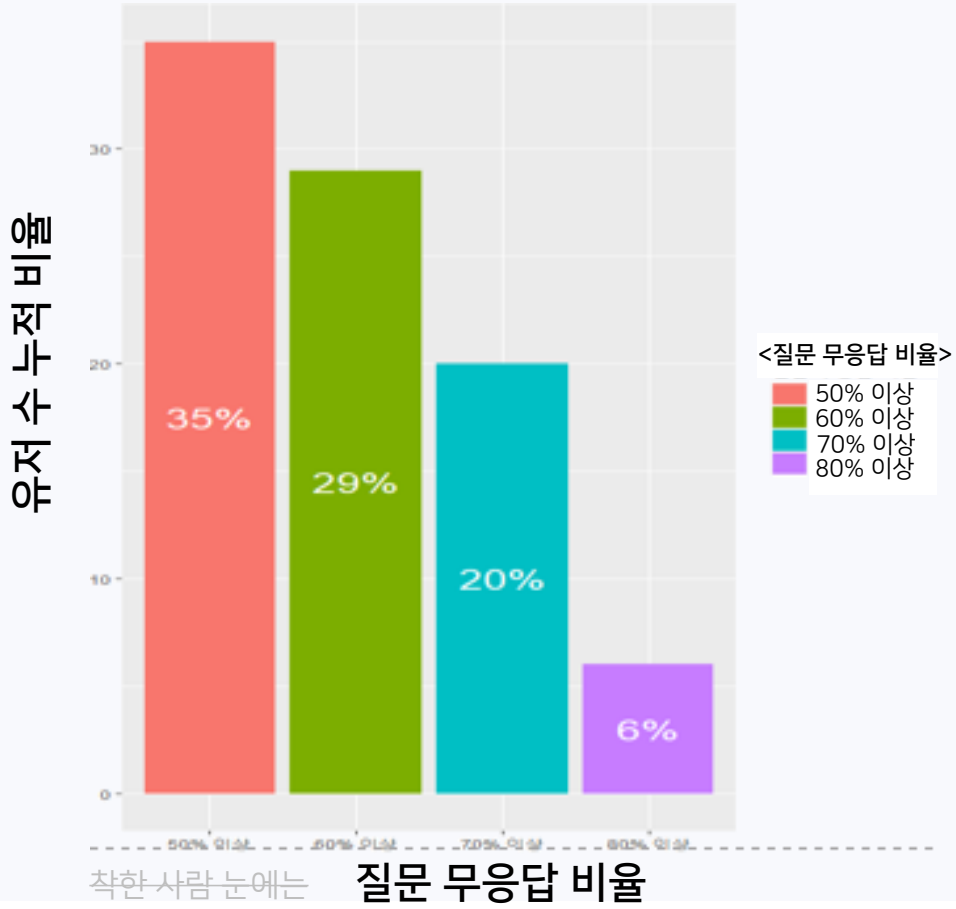
Plot 확인 결과, 이상치가 확인되어 Tukey방법을 이용해 구간을 확인하고 3세 이하나 100세 이상은 이상치로 판단, 삭제

➡ 12개 obs 삭제

잘가 소듕한 obs…

- 변수 제거

<User별 무응답 비율>



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

➡ 358개의 행 삭제

~~안그래도 obs적은데..^^~~
~~눈물이 납니다...~~

~~착한 사람 눈에는~~
~~보이는 변수명 2~~

DATA 설명 전처리 **파생변수 생성**

- *인간 도덕성의 5가지 기반*

Jonathan David Haidt

(TED에서 만나요!)

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

설문조사를 통해

5가지 도덕적 기반과 정치적 견해의

연관성을 발견

https://www.ted.com/talks/jonathan_haidt_the_moral_roots_of_liberals_and_conservatives

모든 과정을 거쳐 만들어진 통합 설문 데이터

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

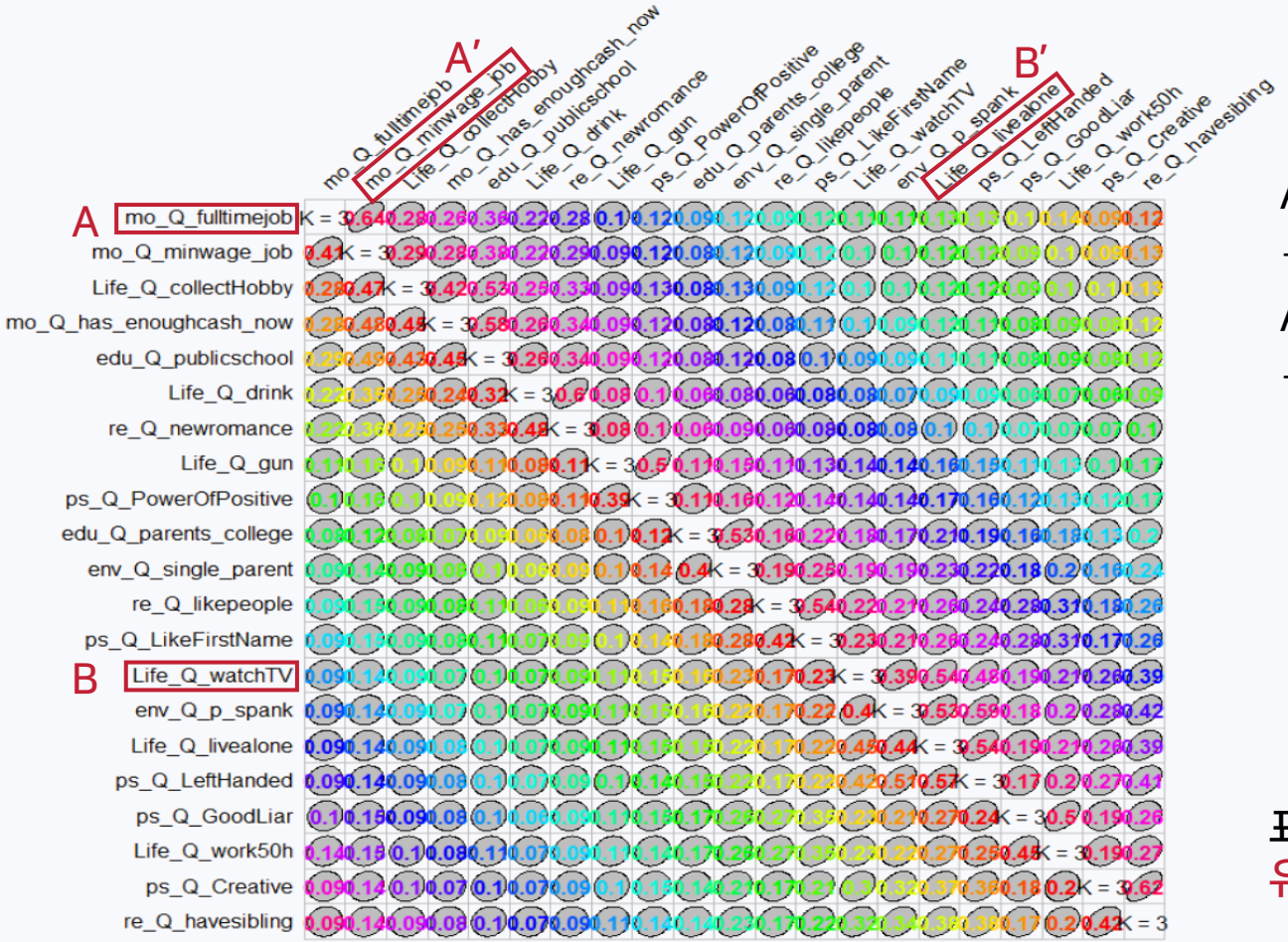
5912 obs & 112 variables

⋮

시각화

개인정보 데이터 설문조사 데이터

- 질문 간 연관성 알아보기(GoodmanKruskal package)



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!

결측치 처리 NA확인 MICE MCA

- NA의 종류

그렇다면 우리가 가진 설문지의 NA는 어떤 종류에 속할까?

MCAR

결측이 랜덤으로 발생
↓
변수에 영향 받지 않음

설문지 구성 방식

하나의 설문지가 아닌
개별 응답 설문들을
USER_ID 기준으로 병합

질문 간 연관성

대체로 질문 간의
연관성이 낮음

결측값 제외해도
분석 결과 편향 X

의도적인 무응답 가능성↓
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!

결측치 처리 NA확인 MICE MCA

MICE with NA

STEP1

설문지 구성 방식을 고려,
Missing Value를 무응답으로
가정하고 진행

STEP2

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

STEP3

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

STEP4

M = 5로 설정,
5개의 대치 셋의 평균을
대치 값으로 선택!

결측치 처리

MICE with Mean Imputation

STEP1

Ordinal Encoding 진행한
변수의 평균값 반올림하여 대치

STEP2

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

STEP3

위의 MICE with NA와
동일한 로지스틱 모델 적용

STEP4

M = 5로 설정,
5개의 대치 셋의 평균을
대치 값으로 선택!

Imputation 변수 선택 알고리즘

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..!

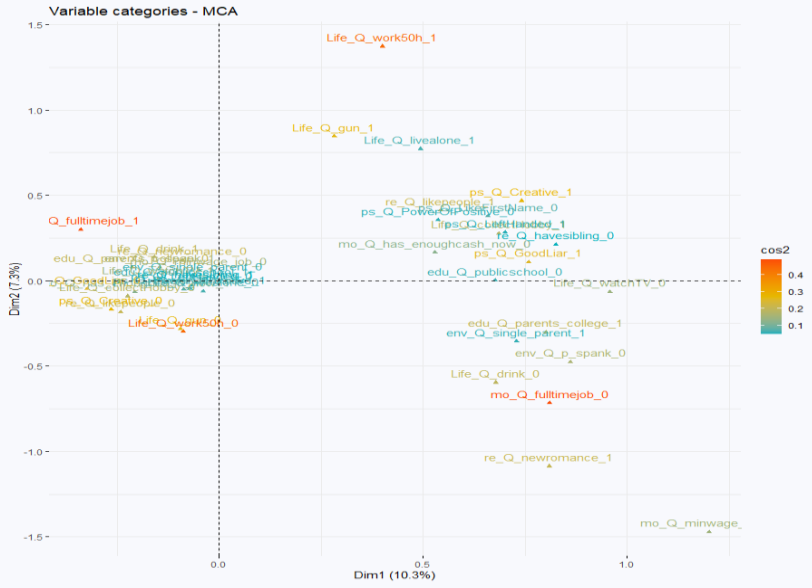
결측치 처리 NA확인 MICE MCA

Imputation 변수 선택 알고리즘

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

STEP2

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



	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

결측치 처리

Regularized Iterative MCA

Overfitting 문제로부터 자유롭다 !

설문 문항
(Binary Data)

ImputeMCA

+

인적사항 변수
(ex. Income, Age)

MICE

결측치 처리

NA확인 MICE MCA

"최종 데이터 분포 비교"