2024 PROJECT REPORT

질병 진단 개수에 대한 알고리즘

나우리
권소연, 박혜인, 오채은, 최지민

목차

연구 목적

목적 1

꾸준한 질병 관리에 대한 통계적 지침 마련

목적 2

현재 생활 습관 상태 점검,
질병에 걸릴 위험이 있는지에 대한 여부 파악

목적 3

비감염성 질병 환자들의
체계적인 질병 관리 확대

변수 설명

<표1> 변수 설명

자료유형	변수명	설명		
반응변수	dg_sum	질병 진단빈도	0~6	
범주형	city	시도	0: 비수도권	1:수도권
	sex3	성별·임신여부	0. 남자 1. 여자&임신O	2.여자&임신X
	marri_2	결혼상태	1.유배우자, 동거 3.사별	2. 별거 및 이혼 4. 미혼
	npins	보험가입여부	1.가입	2.미가입
	BH1	건강검진 여부	1.검진	2.미검진
순서형	stress	스트레스	1. 거의 안 느낌 3. 많이 느낌	2. 조금 느낌 4. 대단히 많이 느낌
	smoke_second	간접흡연 노출 정도	0. 노출 안 됨 1. 조금 노출	2. 많이 노출
	BE3_31	걷기 운동 일수	1. 운동 안 함 2. 1일 3. 2일 4. 3일 8.7일(매일)	5. 4일 6. 5일 7. 6일
	BE5_1	근력 운동 일수	1. 운동 안 함 2. 1일 3. 2일 4. 3일	2.1일 3. 2일 5. 4일 6. 5일 이상
	fh_sum	부모 질병 진단율	0. 질병 없음 3. 질병 3개 4. 질병 4개	1. 질병 1개 2. 질병 2개 5. 질병 5개
	urlglu	요당	0. 음성 2. 양성 +,++,+++	1. 미량+- 3. 양성 ++++
	HE_Uket2	요케톤	0. 음성 1. 양성 + 2. 양성 ++	3. 양성 +++
	HE_Ubld	요잠혈	0. 음성	1. 미량+-
			2. 양성 + 3. 양성 ++	4. 양성 +++
연속형	eatout	외식횟수	1~7	
	ainc	소득	17~1500	
	EC_wht_23	주당 평균 근로시간	0~999	
	weightch	체중변화	-9~9	
	BD.total	1년간 음주량	0~999	
	sleep	수면시간	14~999	
	smoke_sum	흡연량	0~999	
	slt	착석 시간	30~1200	
	dls_bp	맥압	19~125.50	
	HE_BMI	체질량지수	13.54~46.72	
	HE_Uph	요산도	5~9	
	HE_Usg	요비중	1.001~1.050	
	N_SUGAR	당 섭취량	0.1897~23.9426	
	age	나이	10~80	
	foodfreq	일주일간 식사 빈도	0~2	
	N_EN	에너지 섭취량(Kcal)	2.683~82.201	
	fateat	지방 섭취량	0.4243~24.7967	
	N_NA	나트륨 섭취량	0~145.22	

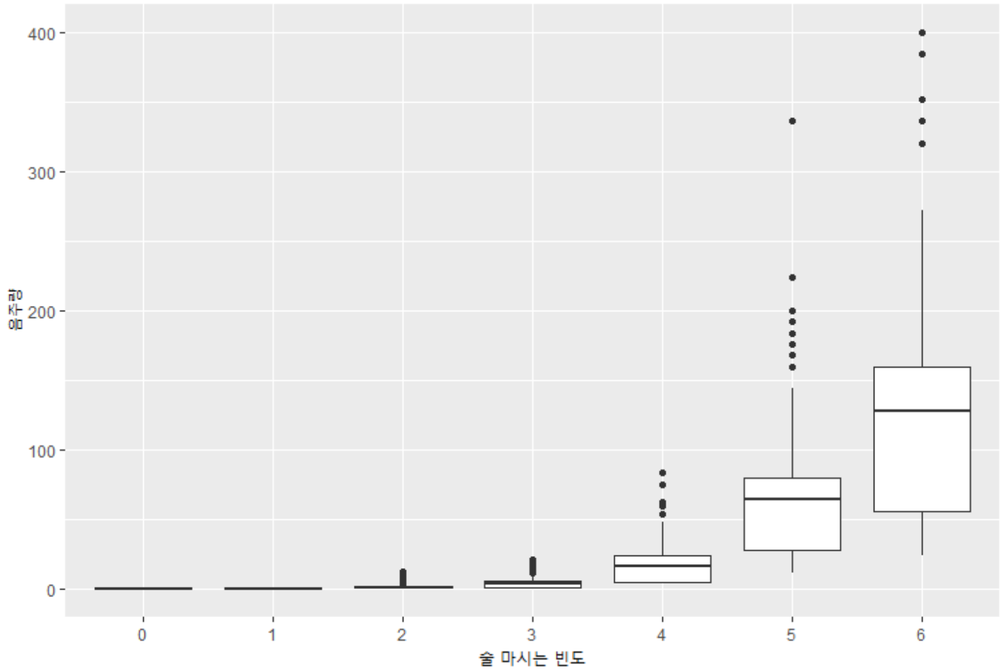
데이터 전처리

- 9개의 질병을 합쳐 '질병 진단 빈도를 반응변수로 형성

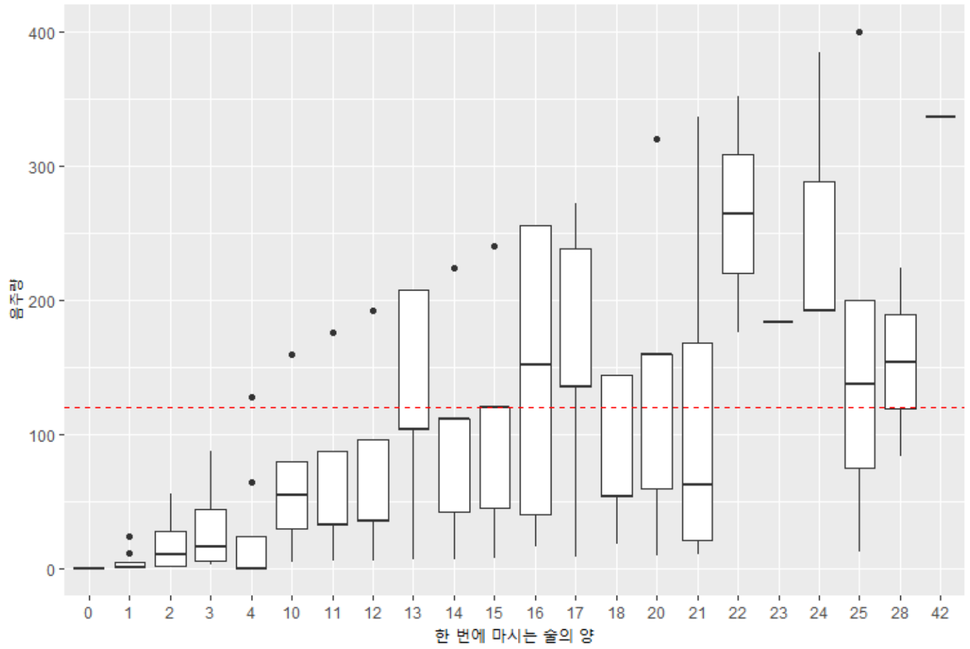
-범주 중 질병의 개수가 5~6개인 데이터가 적음 → 이를 질병 개수 4로 합쳐 0~4까지의 빈도를 가지도록 범주를 형성

데이터 전처리

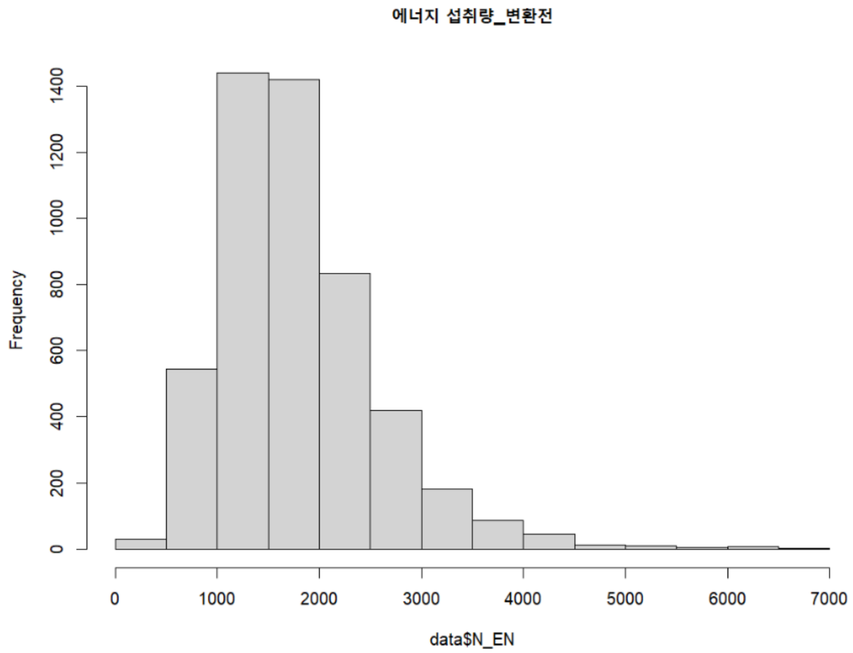
한달 음주량 : 평생 음주 경험 여부, 음주 빈도, 1회 음주량을 합쳐 생성

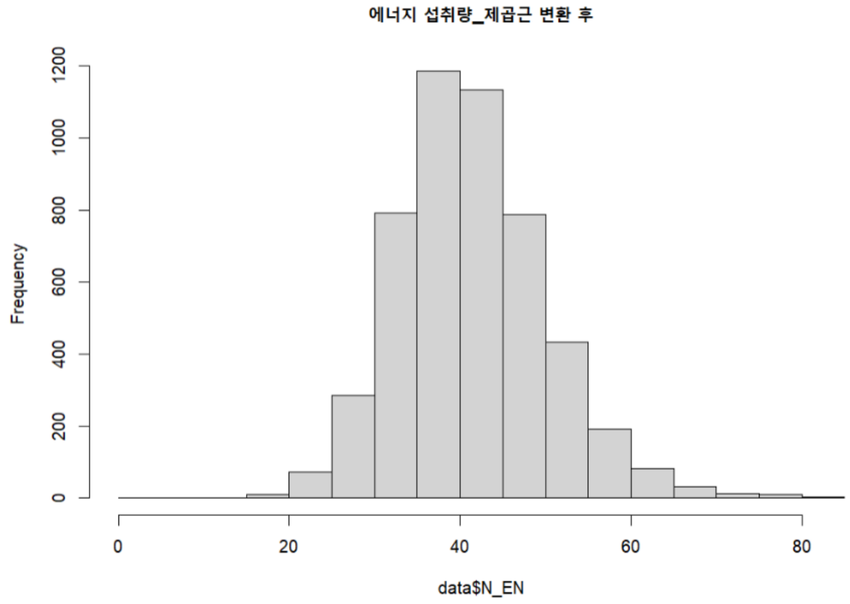


한 달 동안 섭취하는 음주량이 작더라도 한 번에 마시는 술의 양이 많을 수 있음을 주의하여 해석해야 함



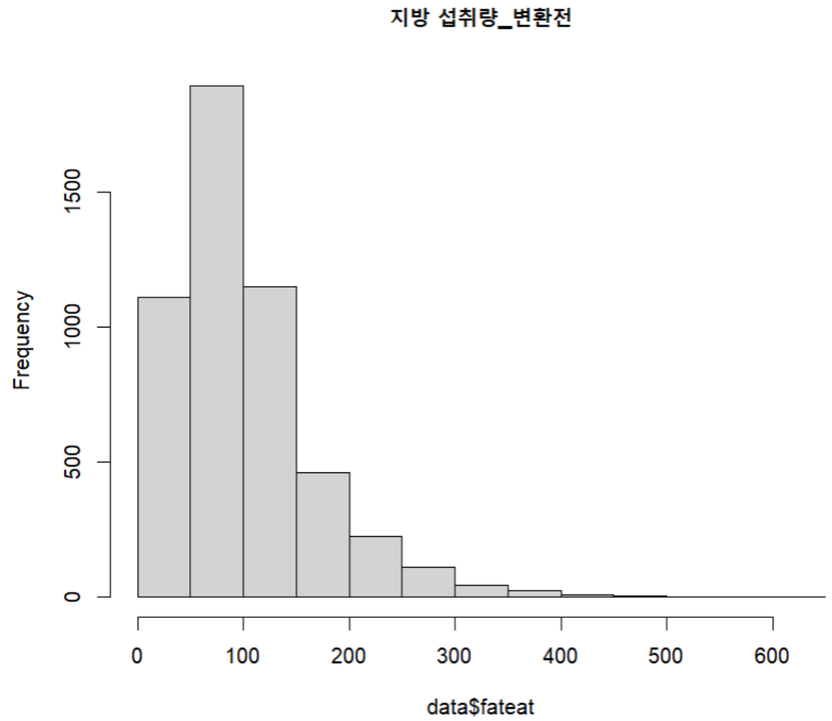
데이터 전처리

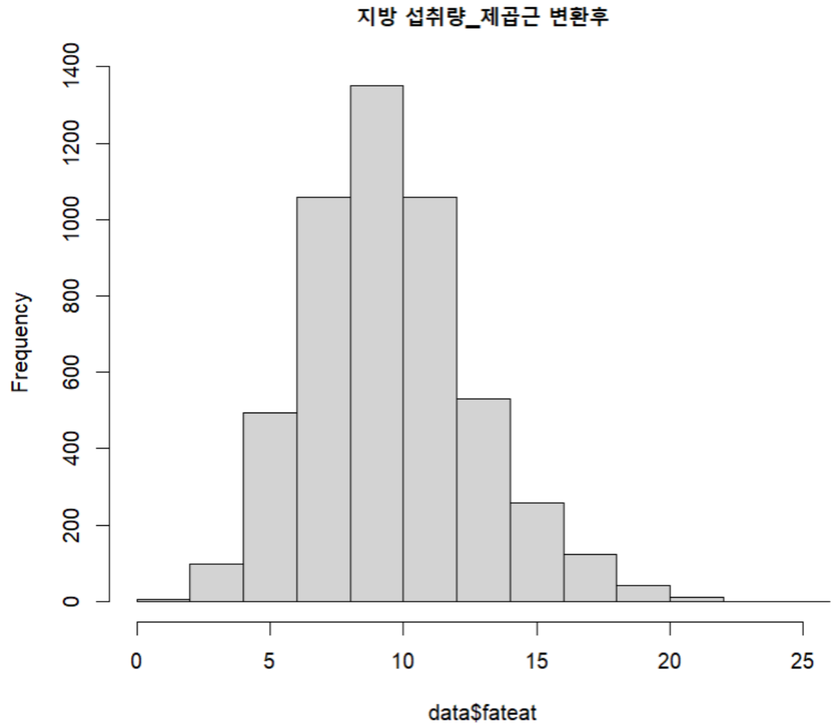




에너지 섭취량 : 왼 쪽으로 분포가 편향되어 sqrt 변환 진행

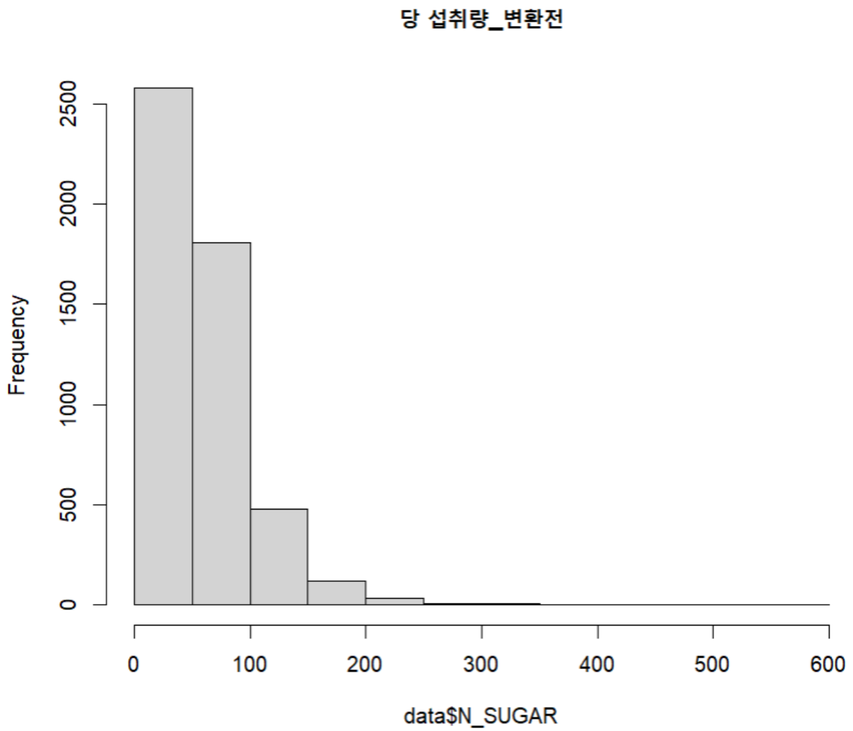
데이터 전처리

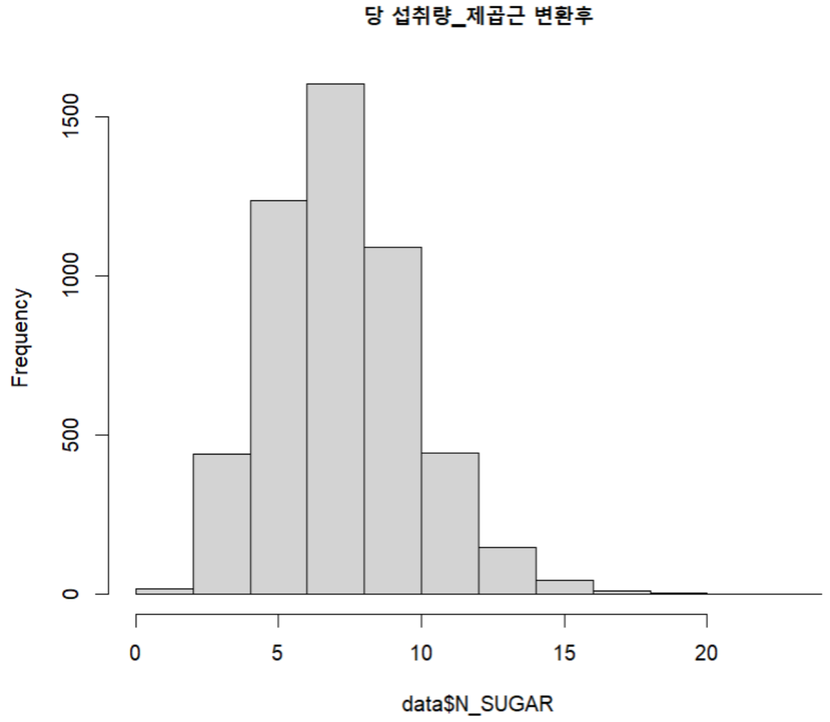




지방 섭취량 : 왼 쪽으로 분포가 편향되어 sqrt 변환 진행

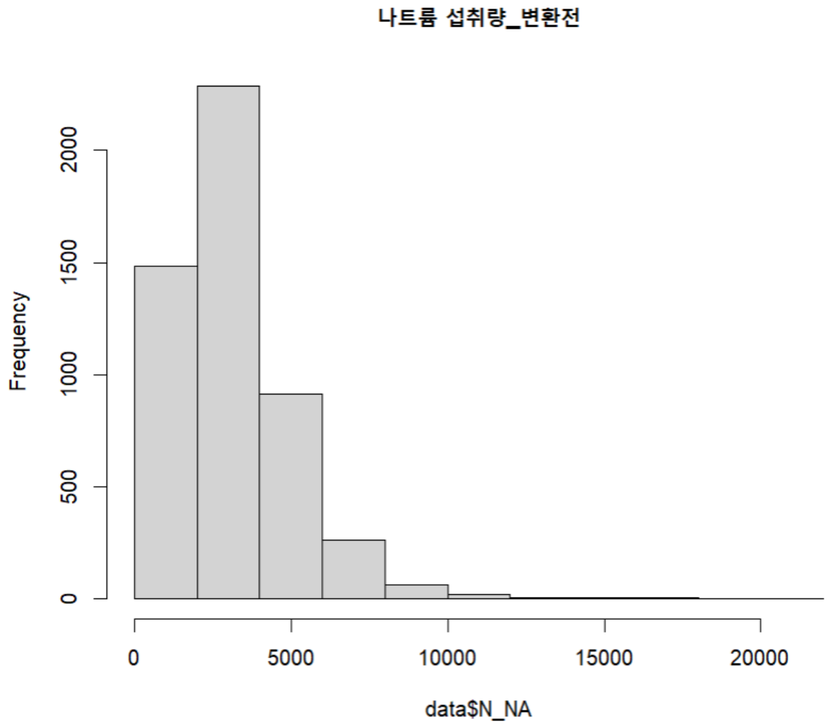
데이터 전처리

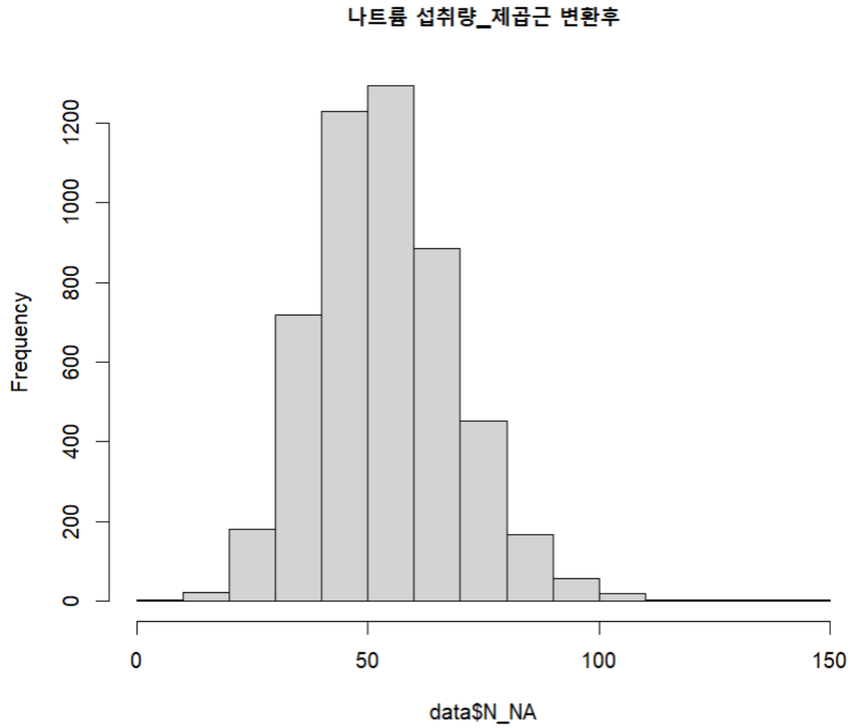




당 섭취량 : 왼 쪽으로 분포가 편향되어 sqrt 변환 진행

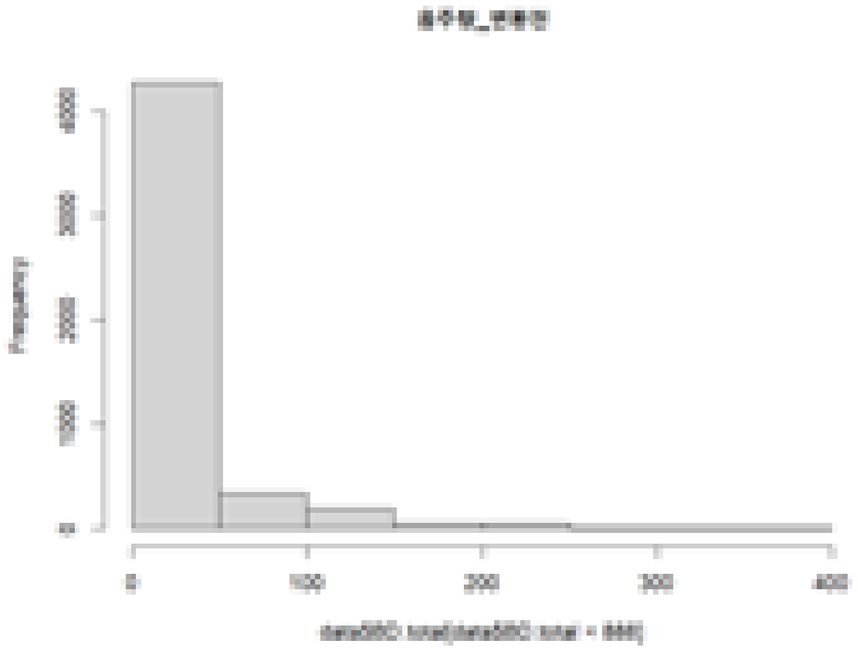
데이터 전처리

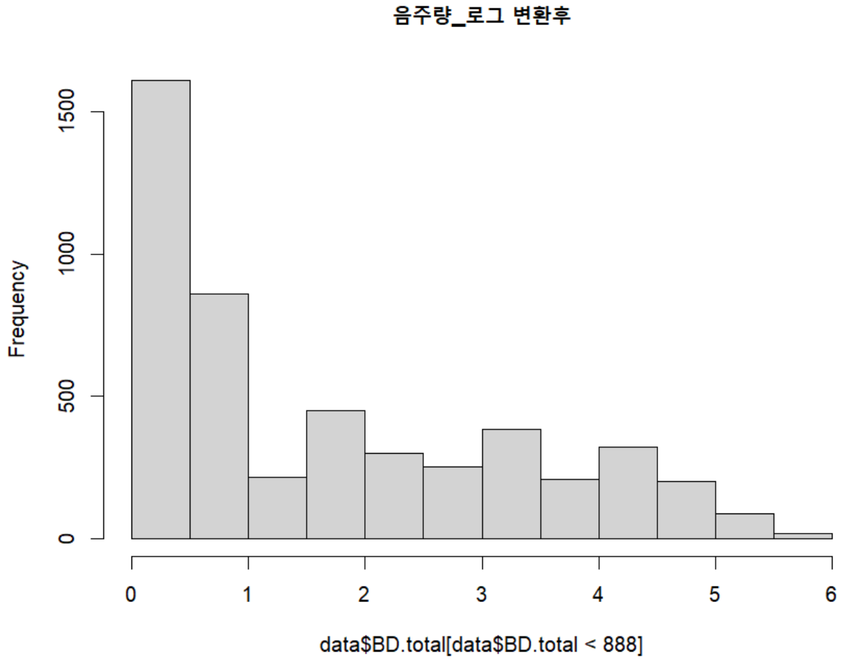




나트륨 섭취량 : 왼 쪽으로 분포가 편향되어 sqrt 변환 진행

데이터 전처리





음주량 변화량 : 분포에 0이 존재, 기존 데이터에 1을 더한 후 log 변환 진행

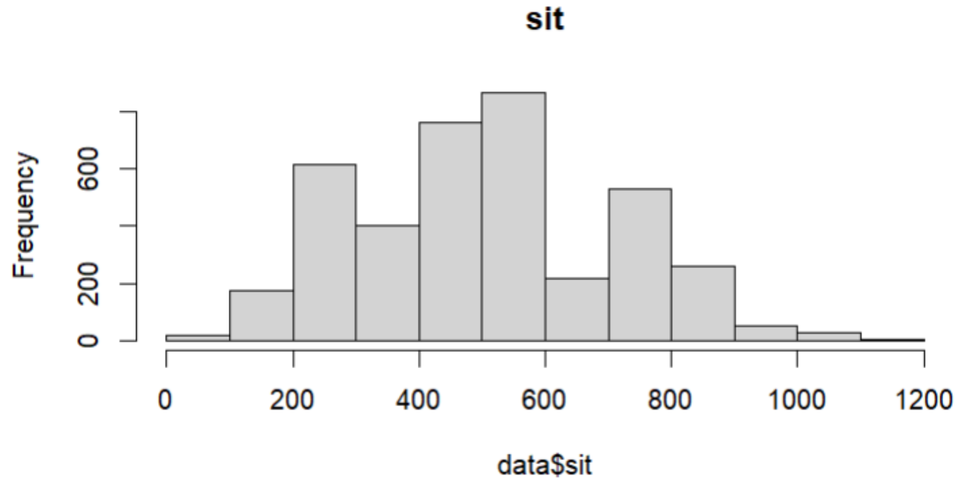
데이터 전처리

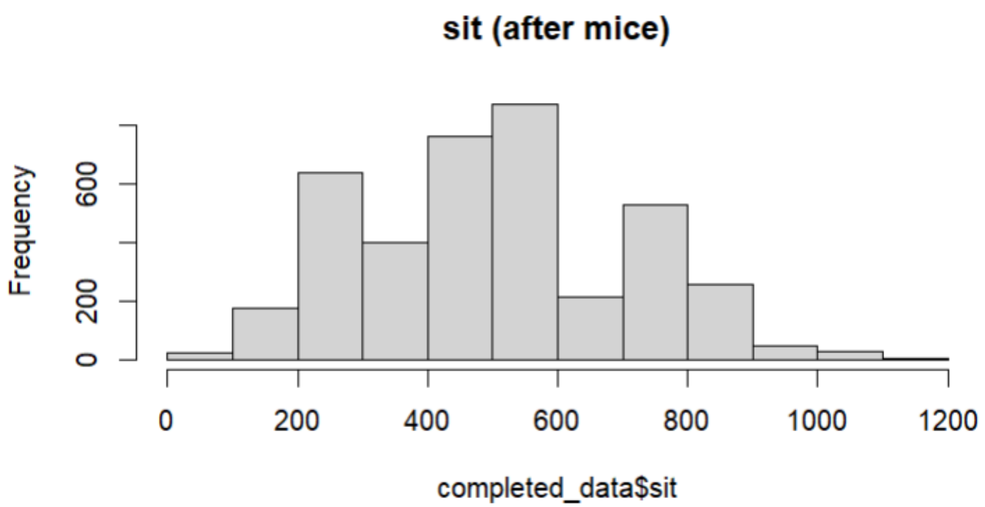
착석 시간	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
mice 전	30	360	480	517.9	660	1200
mice 후	30	360	500	519.1	660	1200

연속형 변수의 무응답 : MICE로 결측 처리

결측치 보간 전후 분포의 차이 크게 나타나지 않음

데이터 전처리

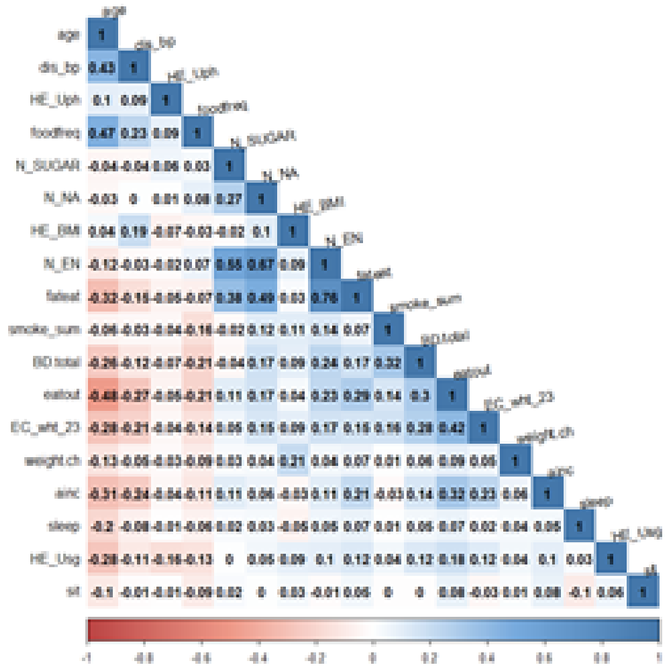


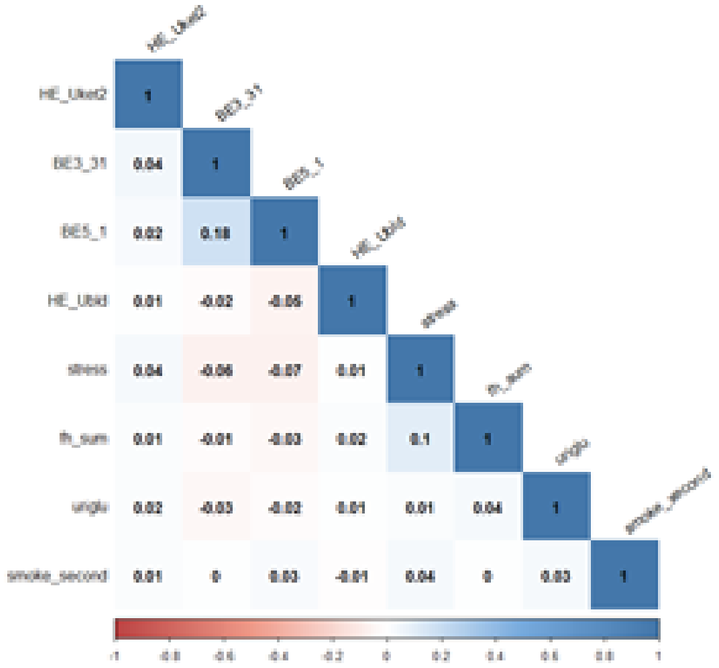


결측치의 보간 전후 분포의 차이 크게 존재하지 않음

착석 시간의 경우, 무응답 빈도 많아 PMM 기법 사용

연관성 분석





피어슨 상관분석 진행 : 다중공선성 의심 변수 발견

스피어만 상관분석 진행 : 모든 순서형 변수의 상관관계가 0.3 이하,

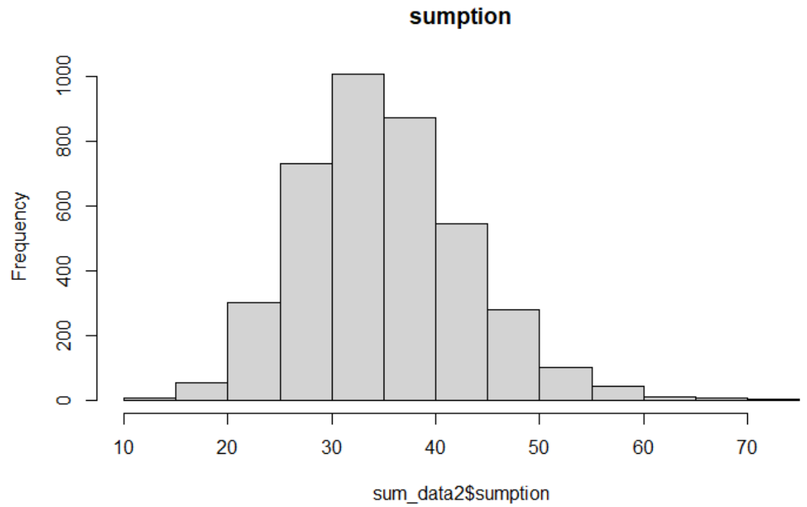
다중공선성에 미치지 않음

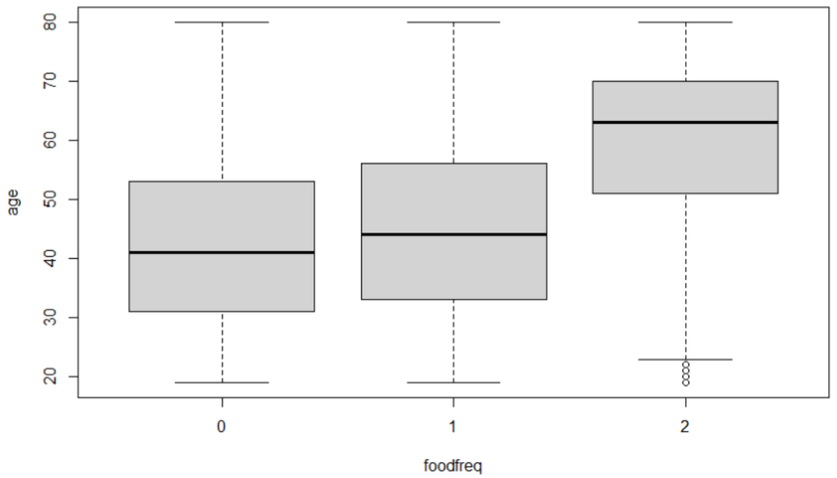
연관성 분석

p-value	수도권	성별·임신여부	결혼상태	보험가입여부	건강검진 여부
수도권					
성별·임신여부	4.72×10^{-3}				
결혼상태	1.38×10^{-15}	$< 2.23 \times 10^{-16}$			
보험가입여부	5.58×10^{-15}	0.0001736	$< 2.23 \times 10^{-16}$		
건강검진 여부	1	1.88×10^{-15}	$< 2.23 \times 10^{-16}$	1.90×10^{-6}	

범주형 설명변수 간 카이제곱 검정 결과,
유의수준 1%에서 다중공선성 의심 변수 발견

연관성 분석



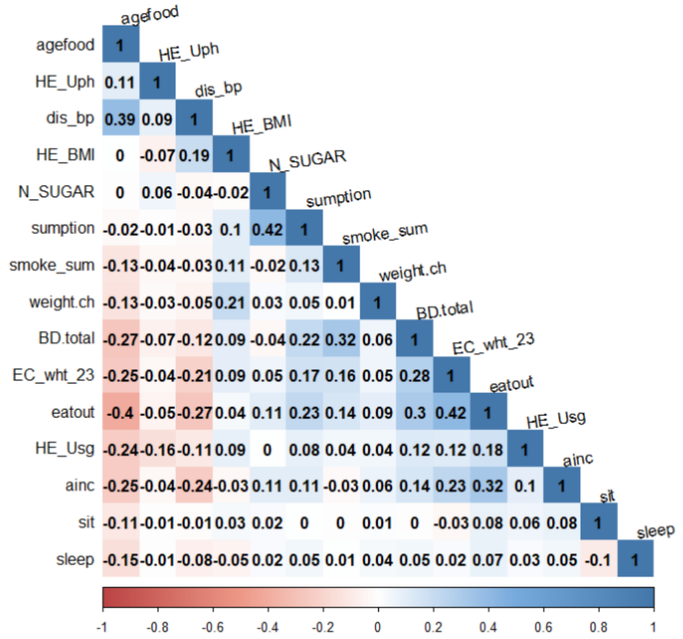


차원 축소로 다중공선성 문제 해결 위해 요인분석 진행

에너지 섭취량, 지방 섭취량, 나트륨 섭취량, 나이, 식사빈도 데이터 기반으로

두 가지 요인 설정하여 분석 진행

연관성 분석



요인분석 후 식습관별 연령대 변수와 외식횟수만 다중공선성 의심된다는 결과 도출

차원 축소를 통해 연속형 다중공선성 문제 감소 확인함

최종 변수 설명

자료 유형	변수명	설명		
반응변수	dg_sum	질병 진단빈도	0~4	
범주형	city	시도	0: 비수도권	1:수도권
	sex3	성별·임신여부	0. 남자 1. 여자&임신O	2.여자&임신X
	marri_2	결혼상태	1.유배우자. 동거 3.사별	2. 별거 및 이혼 4. 미혼
	npins	보험가입여부	1.가입	2.미가입
	BH1	건강검진 여부	1.검진	2.미검진
순서형	stress	스트레스	1. 거의 안 느낌 3. 많이 느낌	2. 조금 느낌 4. 대단히 많이 느낌
	smoke_second	간접흡연 노출 정도	0. 노출 안 됨 1. 조금 노출	2. 많이 노출
	BE3_31	걷기 운동 일수	1. 운동 안 함 2. 1일 3. 2일 4. 3일 8.7일(매일)	5. 4일 6. 5일 7. 6일
	BE5_1	근력 운동 일수	1. 운동 안 함 2. 1일 3. 2일 4. 3일	2.1일 3. 2일 5. 4일 6. 5일 이상
	fh_sum	부모 질병 진단율	0. 질병 없음 3. 질병 3개 4. 질병 4개	1. 질병 1개 2. 질병 2개 5. 질병 5개
	uriglu	요당	0. 음성 2. 양성 +.++.+++	1. 미량+- 3. 양성 ++++
	HE_Uket2	요케톤	0. 음성 1. 양성 + 2. 양성 ++	3. 양성 +++
	HE_Ubld	요잠혈	0. 음성	1. 미량+-
			2. 양성 + 3. 양성 ++	4. 양성 +++

연속형	eatout	외식횟수	1~7
	ainc	소득	17~1500
	EC_wht_23	주당 평균 근로시간	0~119
	weightch	체중변화	-3~3
	BD.total	1년간 음주량	0~5.953
	sleep	수면시간	14~98
	smoke_sum	흡연량	0~60
	sit	착석 시간	30~1200
	dis_bp	맥압	20~117.50
	HE_BMI	체질량지수	13.54~46.72
	HE_Uph	요산도	5~9
	HE_Usg	요비중	1.001~1.050
	N_SUGAR	당 섭취량	1.711~23.943
	agefood	연령대별 식습관	9.5~41.00
	sumption	영양소 섭취량	10.57~73.18

모형 적합

MSE	예측값과 실제값의 차이를 제곱한 값의 평균
AIC	모델의 적합성을 의미하는 값으로 AIC가 낮을수록 적합도가 높음
잔차이탈도	모형의 예측값과 실측값이 근접하게 일치하는지 나타내는 통계량으로, 잔차이탈도가 작을수록 모형이 관측치를 잘 예측함

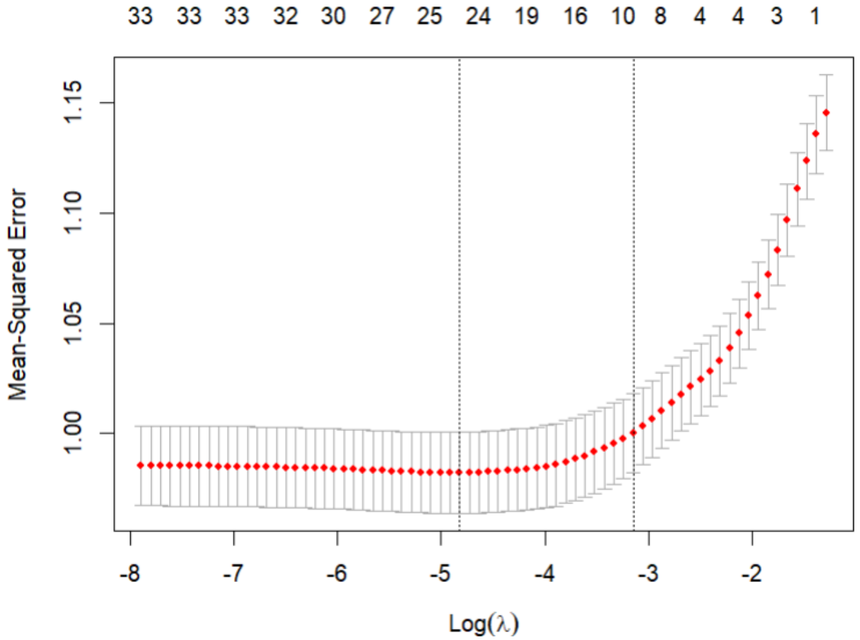
학습용 데이터와 평가용 데이터 세트를 7:3 비율로 분할

반응변수가 질병 진단 빈도임을 고려, 포아송 회귀 모형 선택

요인분석 진행하여 새로운 변수를 포함한 데이터로 모형 적합, 단계적 선택법 활용하여 적합한 축소모형과 결과 비교

비교모형으로는 요인분석 전 데이터를 기반으로 한 기존 포아송 모형과 Lasso를 이용한 축소 모형 설정

모형 적합



MSE가 가장 작을 때의 람다값은 0.008047023이고,
회귀계수가 24개 일 때 최적의 모델임을 알 수 있음

모형 적합

반응변수	항목	비표준화계수	항목	비표준화계수	항목	비표준화계수
질병 진단 빈도	임신경험있는여성	-0.05978	착석 시간	0.00026	체질량지수	0.03913
	체중변화	-0.04261	수면시간	0.003407	스트레스	0.06239
	1년간 음주량	-0.03254	맥압	0.007837	별거 및 이혼	0.08385
	사별	-0.02416	식사 빈도	0.008871	건강검진 여부	0.1074
	당섭취량	-0.0123	비수도권	0.01544	부모 질병 진단율	0.109
	간접흡연 노출정도	-0.00133	걷기 운동 일수	0.01653	미혼	0.1503
	나트륨 섭취량	-0.00104	나이	0.01778	요당	0.2059
	소득	1.37×10^{-5}	요산도	0.0387	임신경험없는여성	0.2412

회귀계수의 값이 0이 아닌 변수에 대한 설명변수들의 비표준화 계수값

모형 적합

	AIC	MSE	설명변수 개수	해석할 설명변수 개수
기본포아송	7124.7	0.9777828	31	34
lasso	7110.8	0.9704639	21	24
ANOVA	잔차이탈도 차이=6.1223, df=10, p=0.8049			

전체모형과 축소모형의 성능 평가 지표

모형간 MSE,AIC값이 크지 않았으며, 모형 간 차이를 비교한 ANOVA 결과도

유의수준 1%에서 전체모형과 축소모형이 다르지 않았음

최종모형으로 설명변수 개수가 적은 Lasso를 적합한 축소모형 선택

모형 적합

	AIC	MSE	설명변수 개수	해석할 설명변수 개수
lasso	7110.8	0.9704639	21	24
요인분석+단계	7140.41	0.9752307	15	16
ANOVA	잔차이탈도 차이=45.58, 자유도=8, $p=2.856\times10^{-7}$			

최종 모형의 성능 평가 지표

Lasso를 적용한 축소모형과 단계적 선택법을 적용한 요인분석 축소모형 비교

요인분석 데이터를 기반으로 단계적 선택법을 적용한 포아송 회귀모델을 최종모형으로 결정

최종 모델

반응변수	항목	비표준화계수	표준편차	검정통계량	p-값	유의성
질병 진단 빈도	상수	-1.8290	0.2443	-7.487	7.04×10^{-14}	
	식습관별 연령대	0.1130	0.0147	7.694	1.42×10^{-14}	O
	요당	0.2131	0.0278	7.677	1.63×10^{-14}	O
	체질량지수	0.0390	0.0051	7.653	1.96×10^{-14}	O
	맥압	0.0102	0.0017	5.847	5×10^{-9}	O
	부모 질병 진단율	0.0938	0.0209	4.487	7.23×10^{-6}	O
	착석 시간	0.0002	0.0001	2.605	0.0092	O
	음주량	-0.0348	0.0135	-2.571	0.0101	X
	근로시간	-0.0024	0.0010	-2.447	0.0144	X
	근력 운동 일수	0.1047	0.0452	2.318	0.0204	X
	임신 경험x 여자	0.1712	0.0759	2.256	0.0241	X
	체중변화	-0.0465	0.0212	-2.197	0.028	X
	스트레스	0.0554	0.0268	2.071	0.0383	X
	영양소섭취량	-0.0054	0.0026	-2.069	0.0385	X
	임신 경험o 여자	-0.0866	0.0468	-1.852	0.064	X
	걷기 운동일수	0.0122	0.0074	1.645	0.0999	X
	요산도	0.0399	0.0243	1.64	0.101	X
MSE= 0.9752307, AIC=7140.41, Residual deviance = 3196.4 on df 2754						

- 모형의 예측력은 96%, AIC는 7102.3으로 나타남
- Z통계량이 크고 추정된 회귀계수들의 p-value가 작은 변수가 반응변수에 큰 영향을 미친다고 판단

기대 효과

효과 1

개인의 질병 위험도 및 건강 상태를 파악

효과 2

질병 발생 확률이 높은 대상 분류하여 질병 조기 진단 및 예방 가능

효과 3

개인의 건강한 생활 습관 제안 가능

감사합니다