South Africa Heart Disease

데이터마이닝 SAHeart 8조 | 오병찬 김상현 윤수연 최솔

SOUTH AFRICA HEART DISEASE

INDEX

SOUTH AFRICA HEART DISEASE

1. 연구 배경 및 목적

Research Background and Purpose

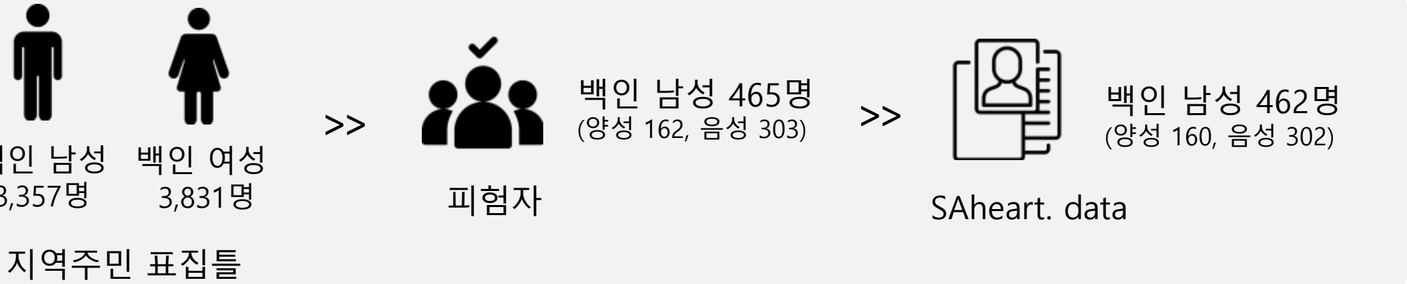
01 연구 소개

1. 연구 배경 및 목적

1.1 연구 배경

1970년대 후반, 남아프리카 공화국의 아프리칸스어 사용 백인 계층에서 이례적으로 높은 빈혈성 심장질환 발생률 관찰 (Wyndham, 1982)

남성 사망률은 여성의 약 2.5 배 (Rossouw et al., 1983)



1.2 연구 목적

주요 위험 요인과 그 강도를 조사 → CHD 관련 인사이트 도출

* CHD : Coronary Heart Disease의 약자로, 관상동맥질환을 의미

SOUTH AFRICA HEART DISEASE

2. 데이터 소개

Data Information

01 데이터 소개

row.names	Sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
1	160	12	5.73	23.11	Present	49	25.3	97.2	52	1
2	144	0.01	4.41	28.61	Absent	55	28.87	2.06	63	1
3	118	0.08	3.48	32.28	Present	52	29.14	3.81	46	0
4	170	7.5	6.41	38.03	Present	51	31.99	24.26	58	1
5	134	13.6	3.5	27.78	Present	60	25.99	57.34	49	1
6	132	6.2	6.47	36.21	Present	62	30.77	14.14	45	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
458	214	0.4	5.98	31.72	Absent	64	28.45	0	58	0
459	182	4.2	4.41	32.1	Absent	52	28.61	18.72	52	1
460	108	3	1.59	15.23	Absent	40	20.09	26.64	55	0
461	118	5.4	11.61	30.79	Absent	64	27.35	23.97	40	0
462	132	0	4.82	33.41	Present	62	14.7	0	46	1

SAHeart.csv

462 rows x 11 columns

CHD 양성 160 / 음성 302

타겟변수 : chd

설명변수 : 9개의 수치형 변수와 1개의 범주형 변수

01 데이터 소개

01 sbp 수치형 변수

| 수축기 혈압 (systolic blood pressure)

심장이 수축할 때 혈관에 가해지는 압력

02 tobacco 수치형 변수

| 누적된 담배 양(kg)

03 ldl 수치형 변수

| 저밀도 지단백 콜레스테롤

(low density lipoprotein cholesterol)

혈중 콜레스테롤을 운반
혈관벽에 과도한 콜레스테롤 침착 유발
심질환 발생시켜 나쁜 콜레스테롤로
분류

04 adiposity 수치형 변수

| Body Adiposity Index(BAI)

BMI의 한계 보완 위해 제안된 지표
키와 엉덩이 둘레로 체지방률 측정 (체중 고려 x)
모든 인종의 성인에게 적용 가능

05 famhist 범주형 변수

| 심장 질환의 가족력

(family history of heart disease)

가족이나 가까운 친척의 의학적 내력
Present : 192 / Absent : 270

06 typea 수치형 변수

| A 유형 성격

A 유형 성격의 정도를 나타내기 위한 점수
성공에 대한 강력한 욕구와 경쟁심
강박관념, 난폭, 분노

07 obesity 수치형 변수

| 체질량 지수 Body Mass Index(BMI)

비만도를 나타내는 수치
체중과 키의 관계로 계산(BMI=체중/키^2)

08 alcohol 수치형 변수

| 알코올 소비량

피실험자의 최근 알코올 소비량

09 age 수치형 변수

| 나이

10 chd (target) 범주형 변수

| 관상동맥질환

관상동맥이 좁아져 심장근육 일부에
혈액 공급이 부족해질 때 발생
1 : 160 / 0 : 302

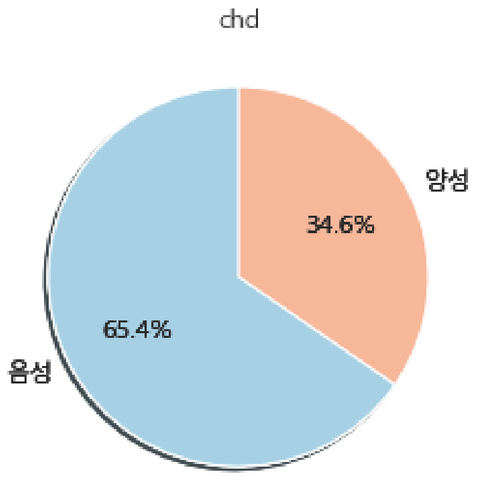
SOUTH AFRICA HEART DISEASE

3. 탐색적 데이터 분석

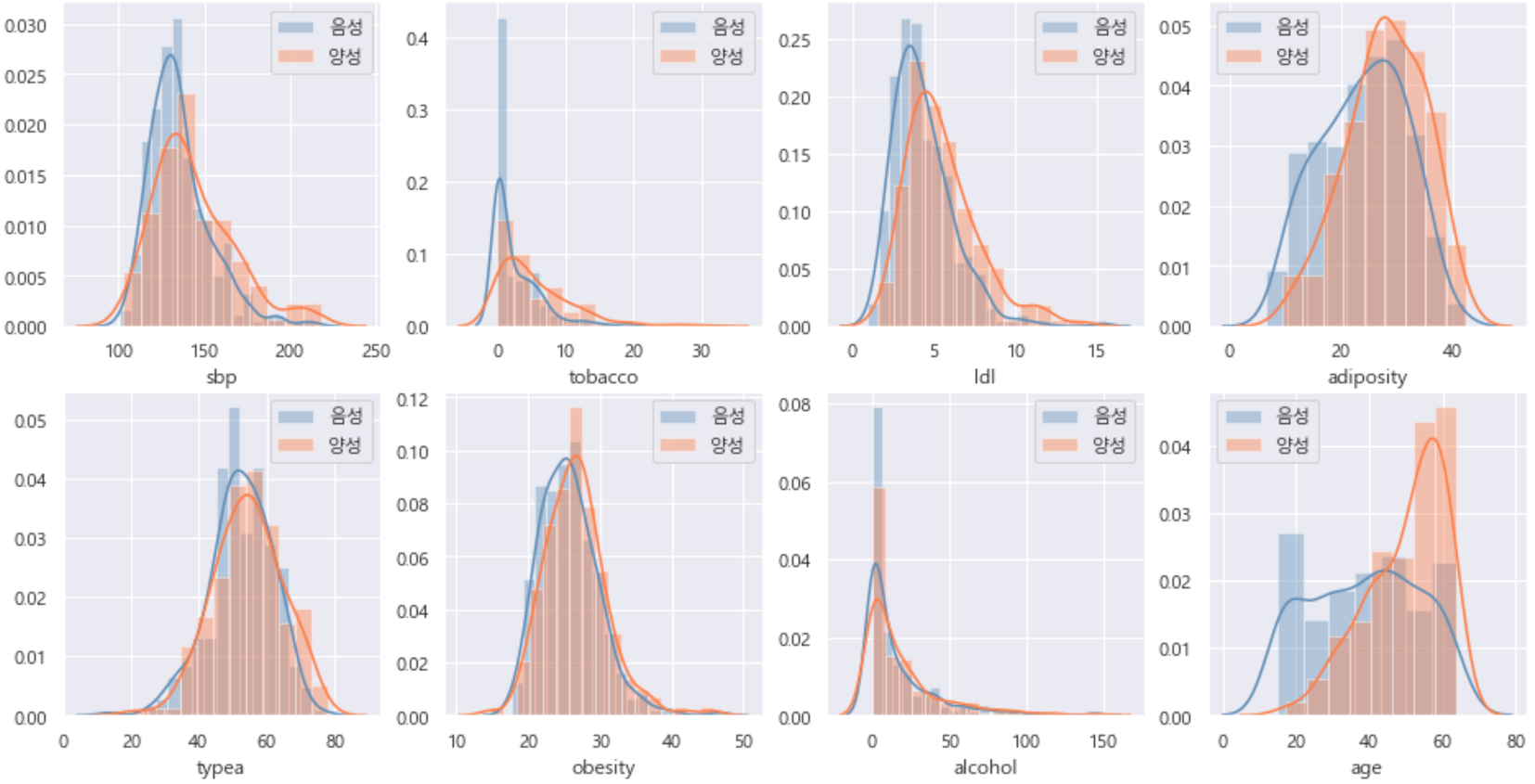
Exploratory Data Analysis

03 EDA

타겟 변수

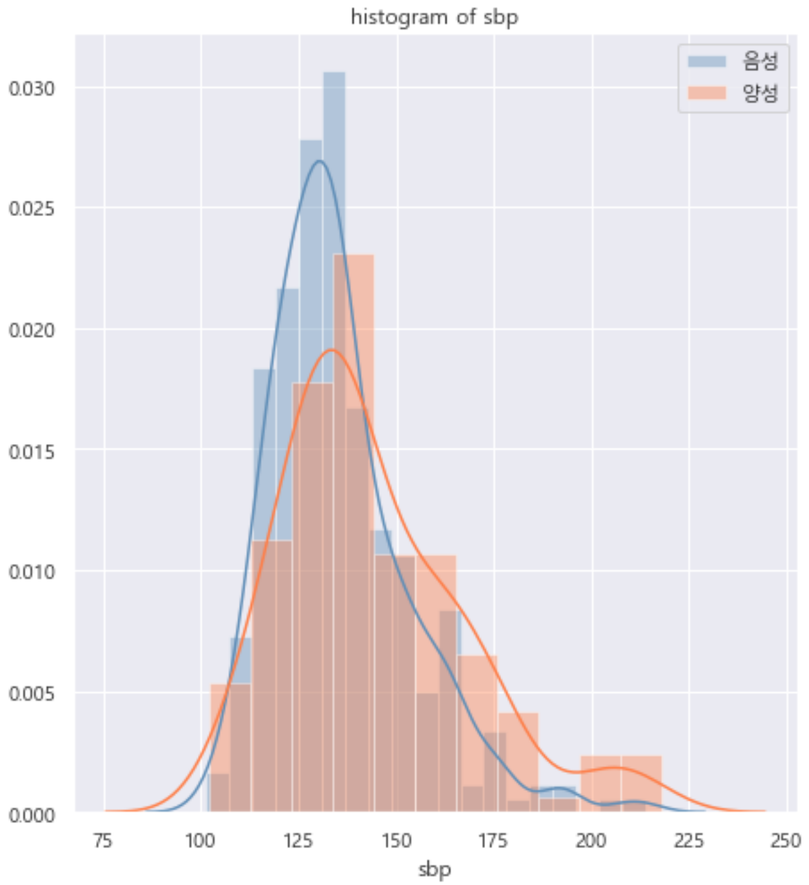


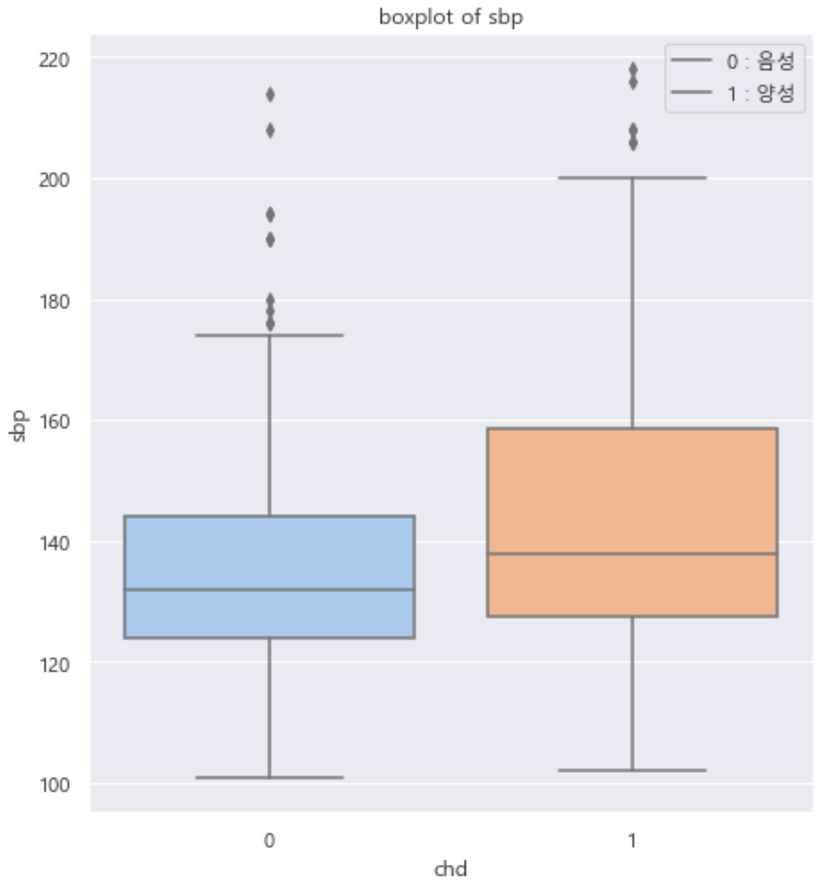
수치형 변수



03 EDA

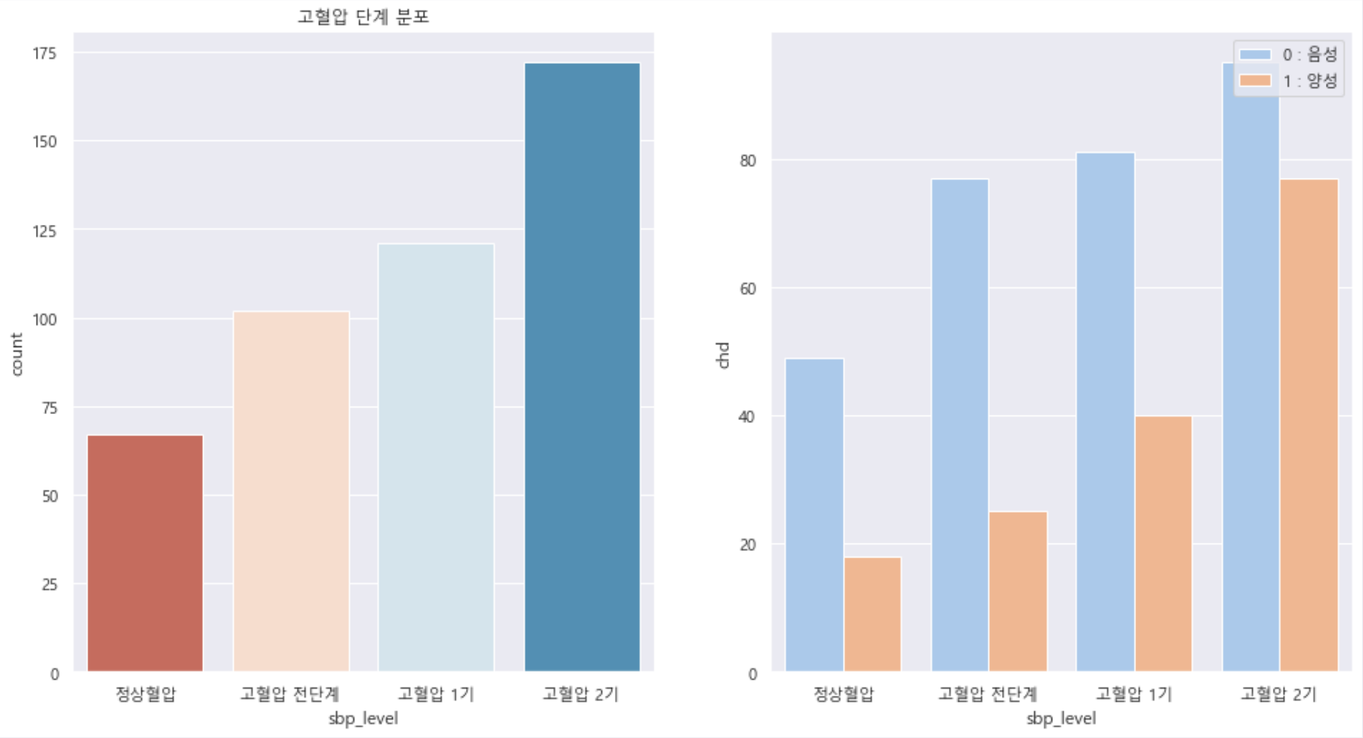
01 SBP


histogram of sbp
음성
양성
0.030
0.025
0.020
0.015
0.010
0.005
0.000
75
100
125
150
175
200
225
250
sbp


boxplot of sbp
0 : 음성
1 : 양성
220
200
180
160
140
120
100
sbp
0
1
chd

03 EDA

01 SBP



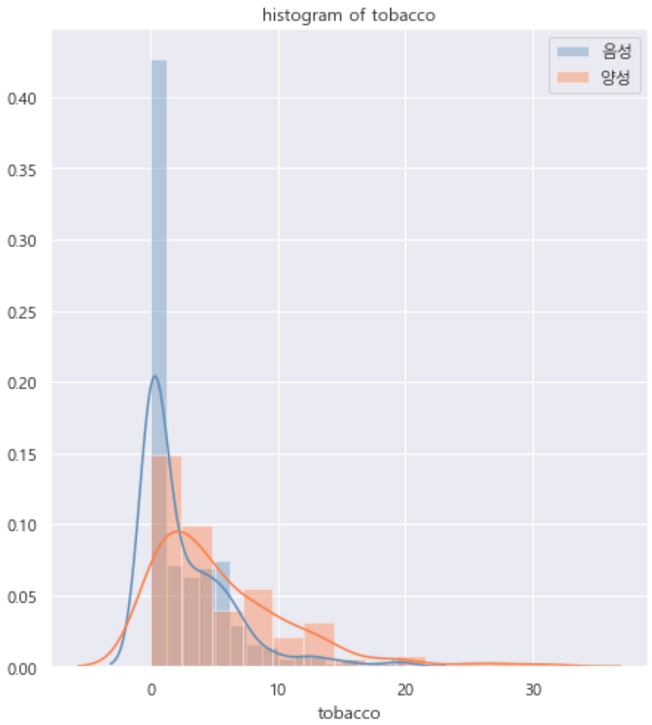
미국심장학회·
심장협회(ACC·AHA)

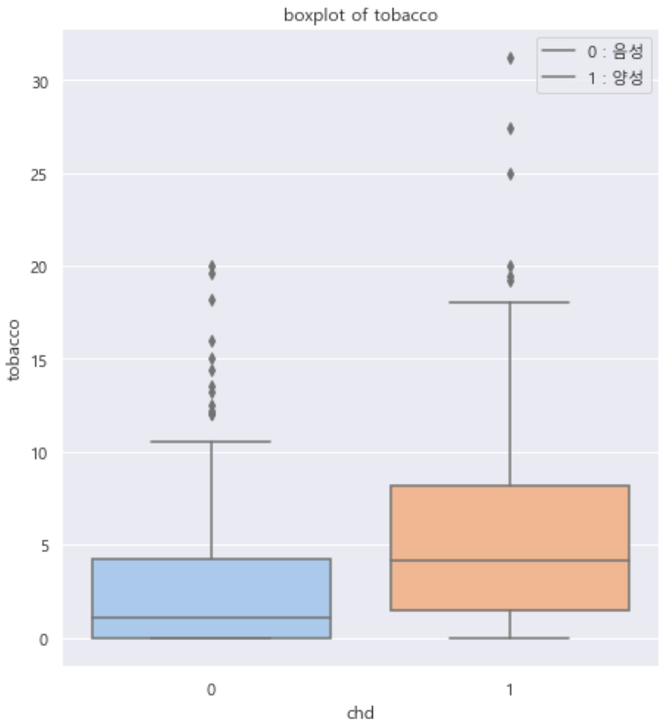
고혈압 가이드라인

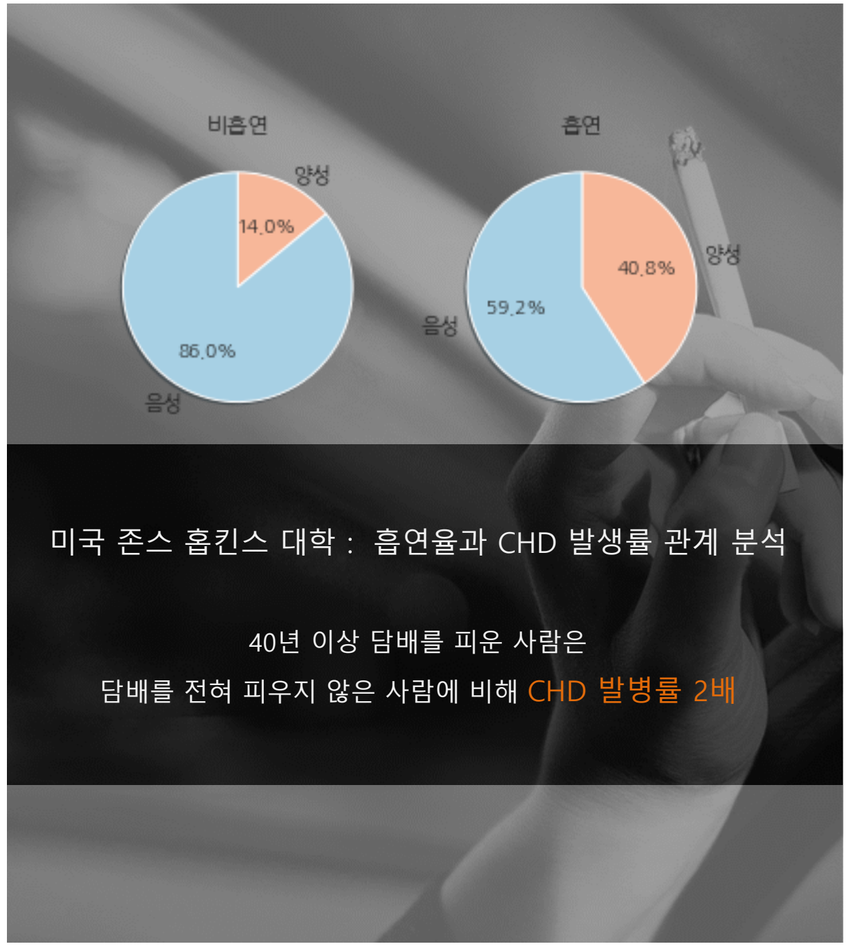
정상혈압 : 120mmHg 미만
고혈압 전 단계 : 120~129mmHg
1기 고혈압 : 130~139mmHg
2기 고혈압 : 140mmHg 이상

03 EDA

02 tobacco

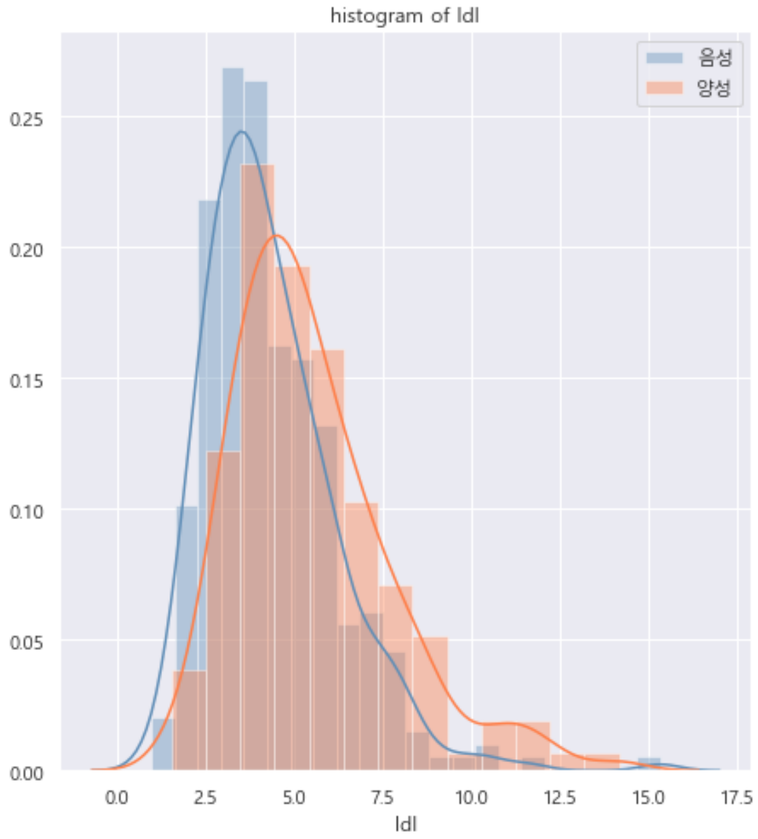


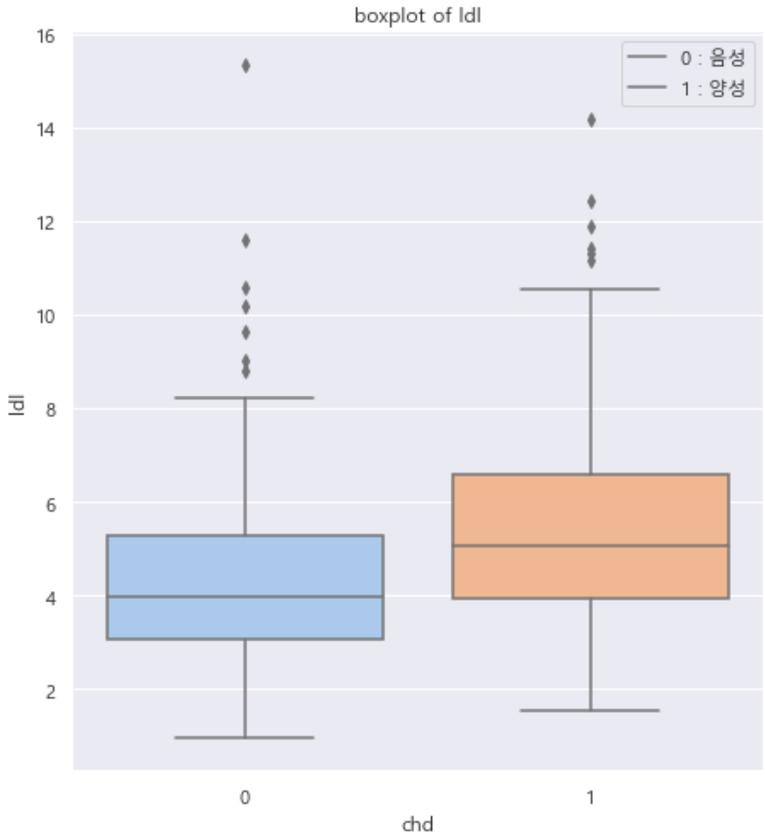




03 EDA

03 ldl


histogram of ldl
음성
양성
0.25
0.20
0.15
0.10
0.05
0.00
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
ldl


boxplot of ldl
0 : 음성
1 : 양성
16
14
12
10
8
6
4
2
ldl
0
1
chd

03 EDA

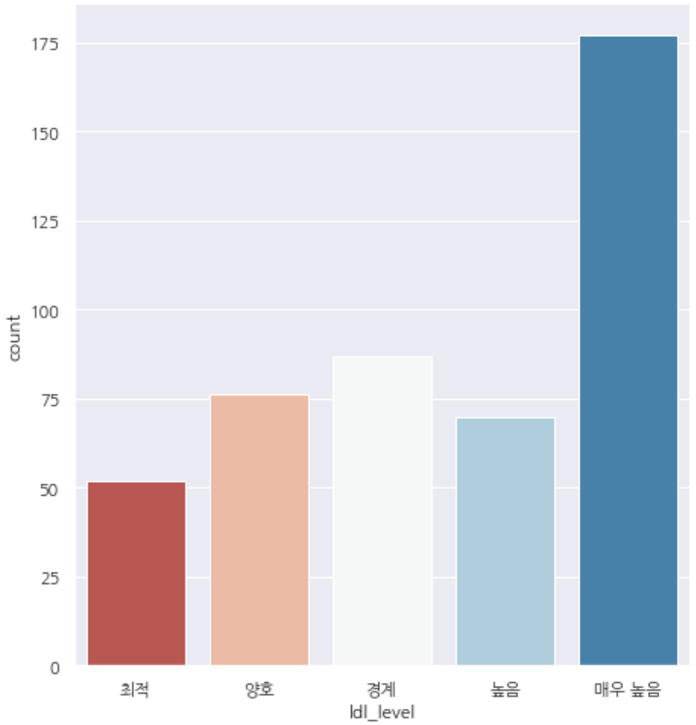
03 ldl

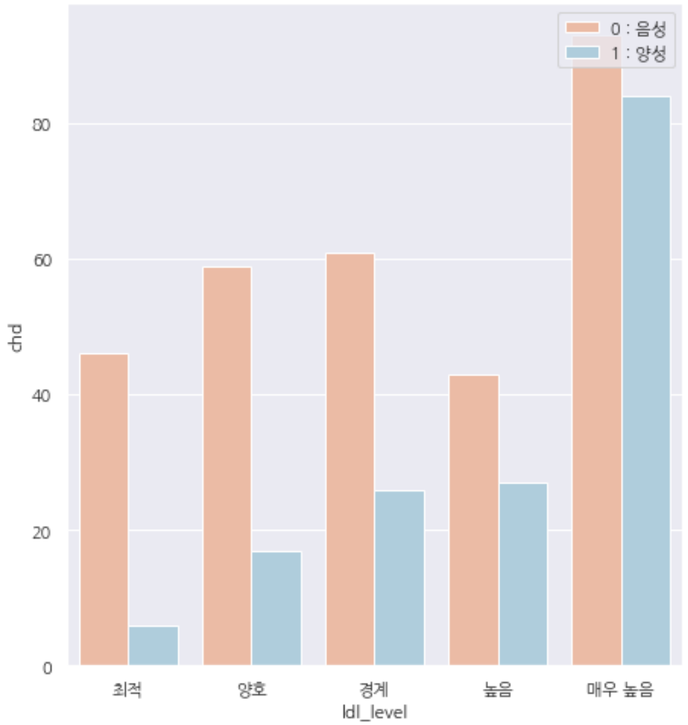
Mayo Clinic &
US National Library of Medicine

미국 성인 기준 LDL 가이드라인

최적 LDL : 100mg/dL 미만
양호 : 100~129mg/dL
경계 : 130~159mg/dL
높음 : 160~189mg/dL
매우 높음 : 190mg/dL 이상

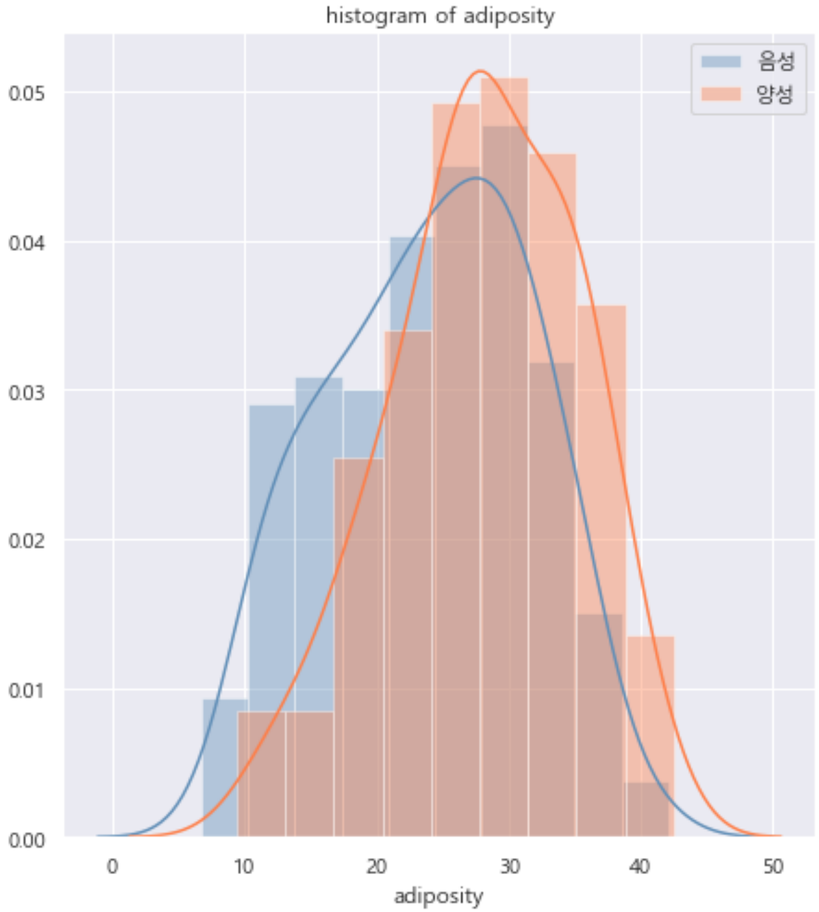
※ ldl 변수 단위는 mmol/L

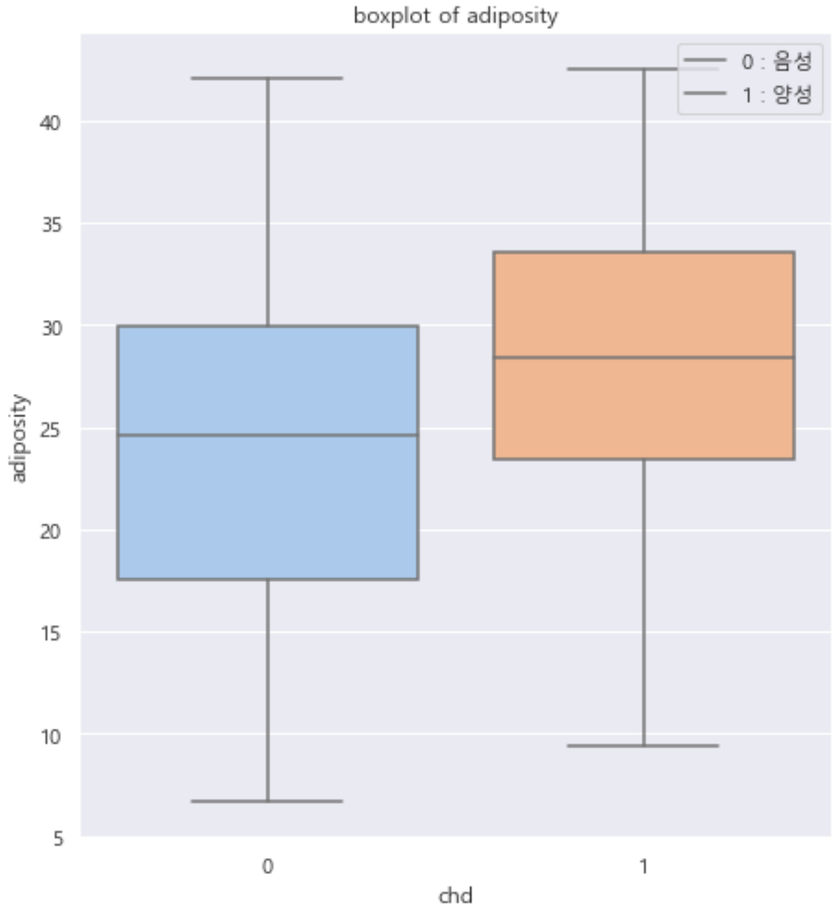




03 EDA

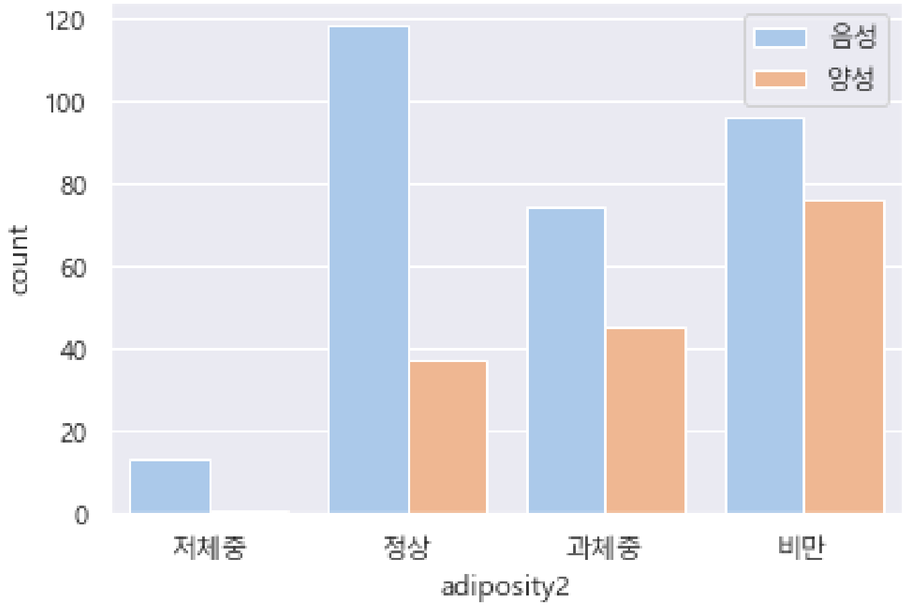
04 adiposity

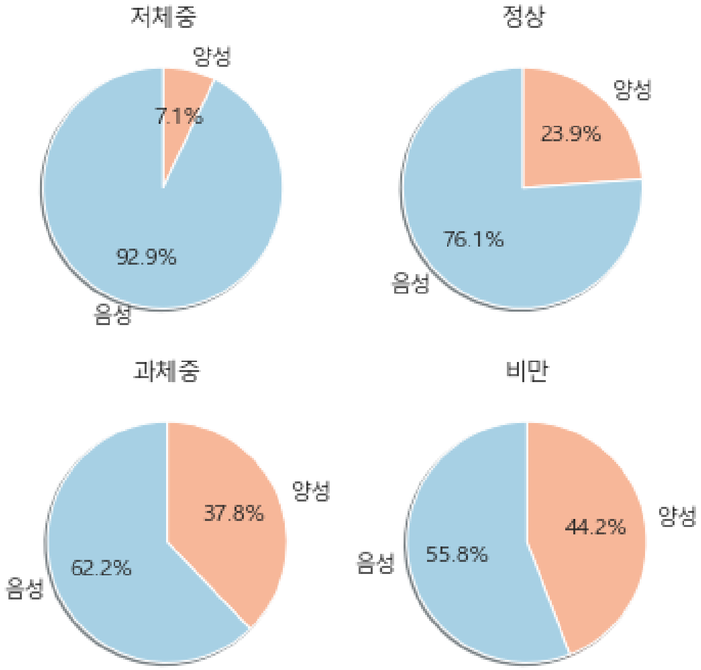




03 EDA

04 adiposity





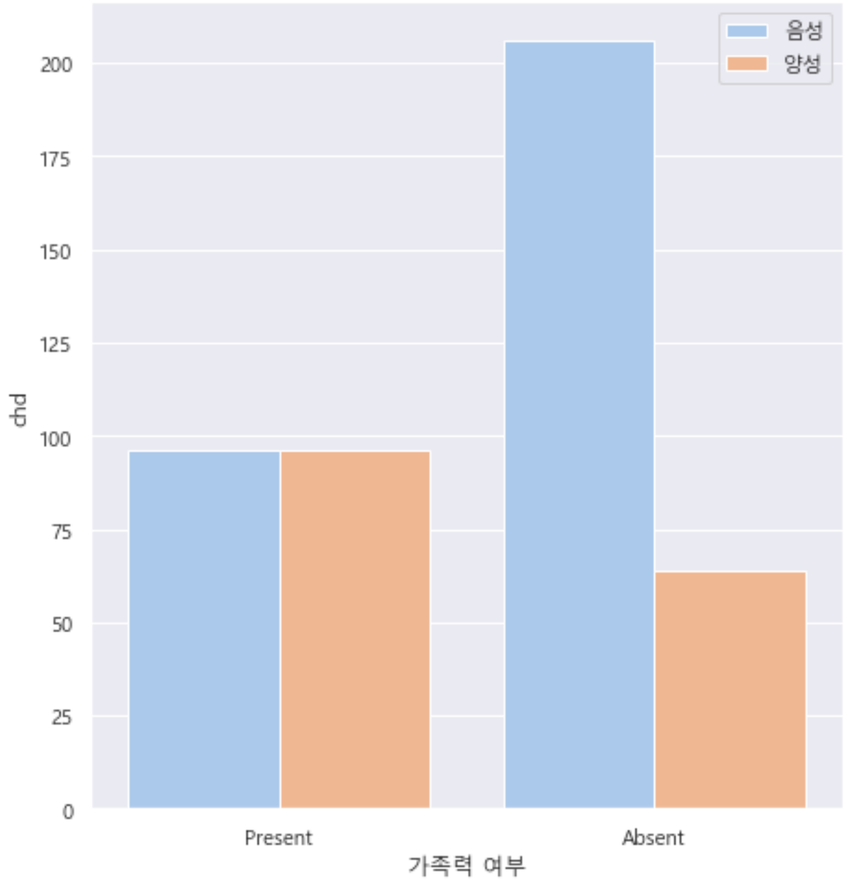
adiposity 변수 저체중/정상/과체중/비만으로 범주화

비만 그룹에서 가장 많은 CHD 발생

비만으로 갈수록 CHD 양성 비율 증가

03 EDA

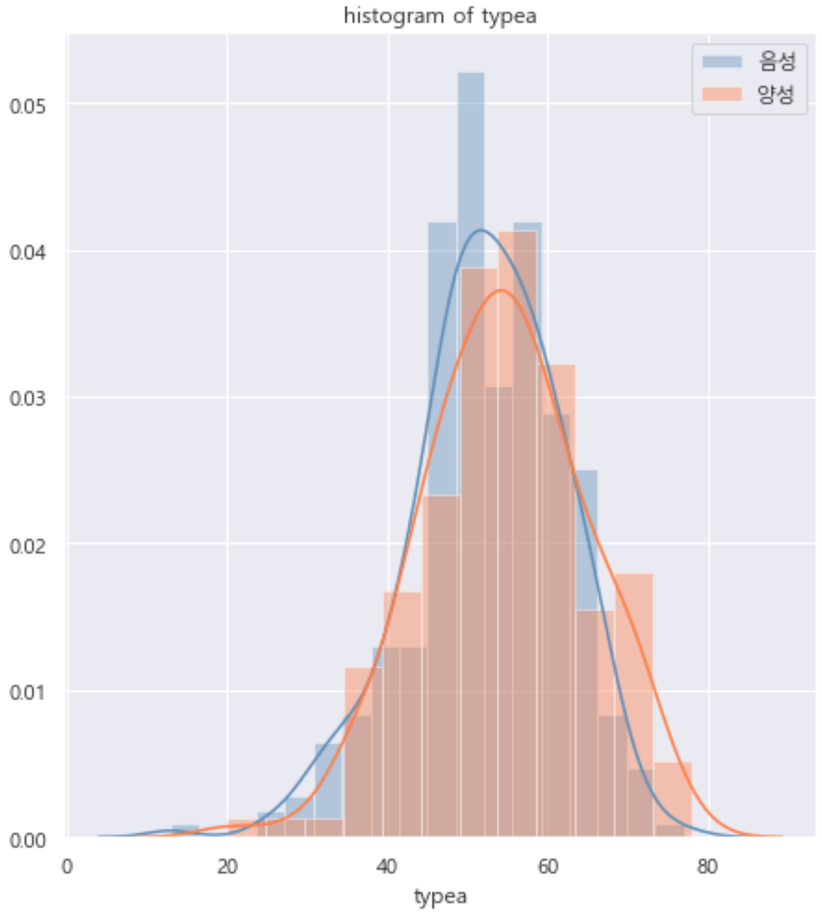
05 famhist

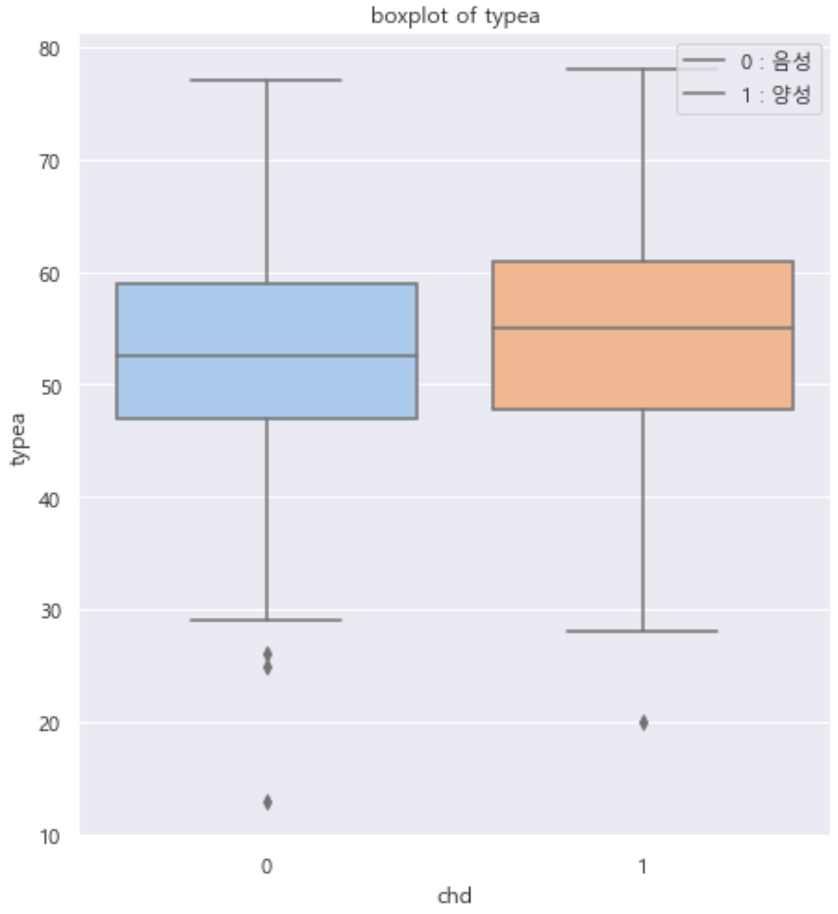


가족력이 있는 경우, 양성률이
가족력이 없는 경우 대비
약 22%p 더 높음

03 EDA

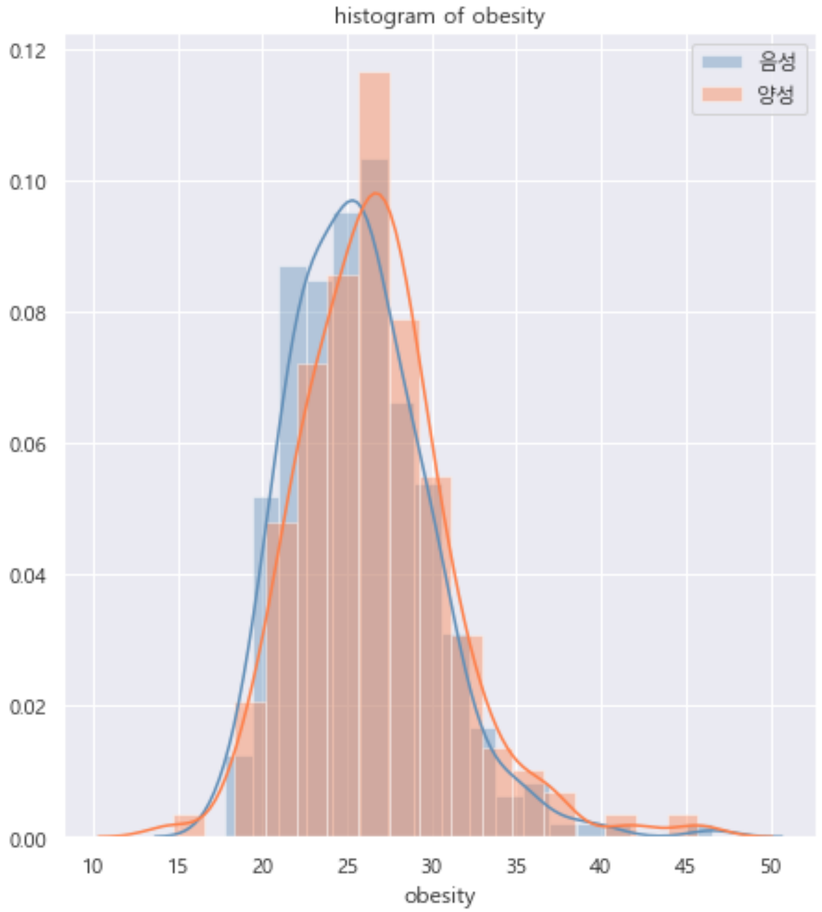
06 type A

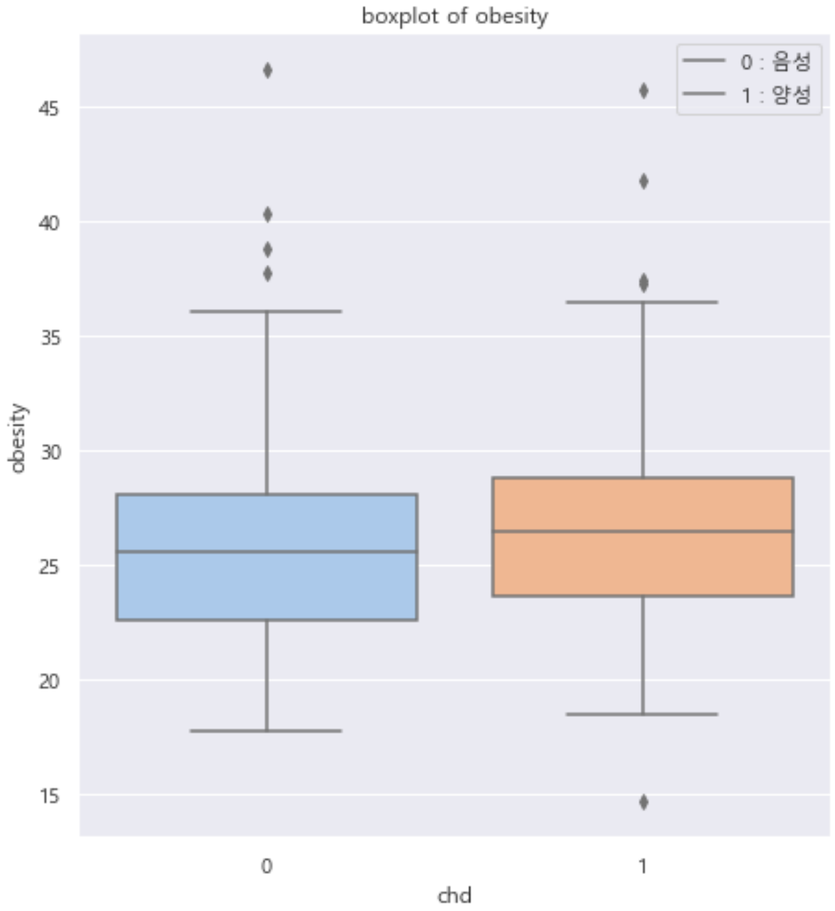




03 EDA

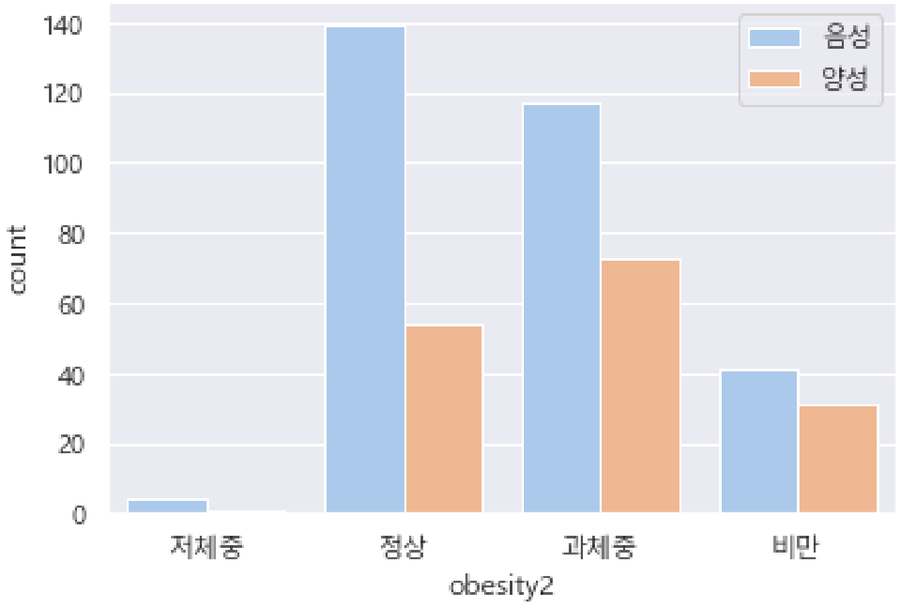
07 obesity

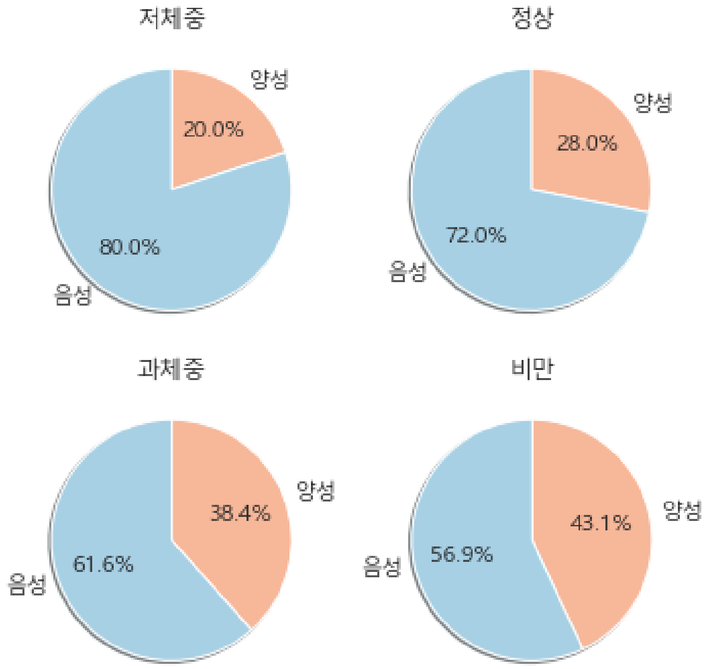




03 EDA

07 obesity





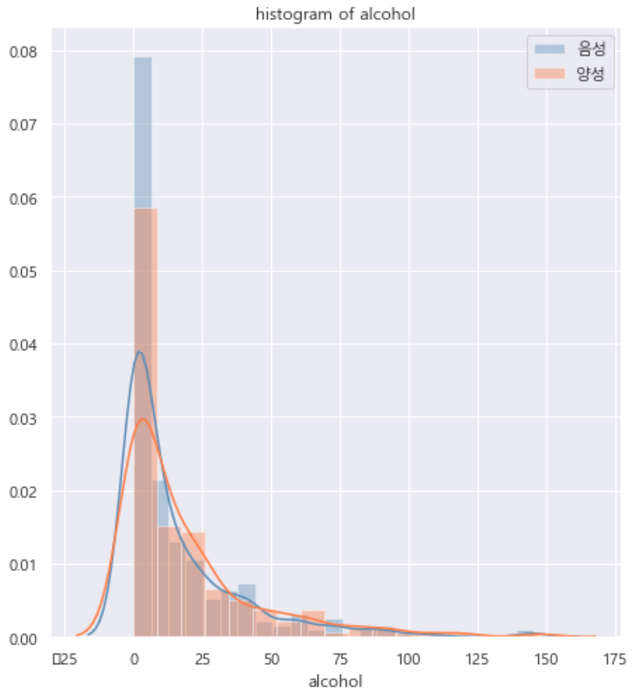
obesity 변수 저체중/정상/과체중/비만으로 범주화

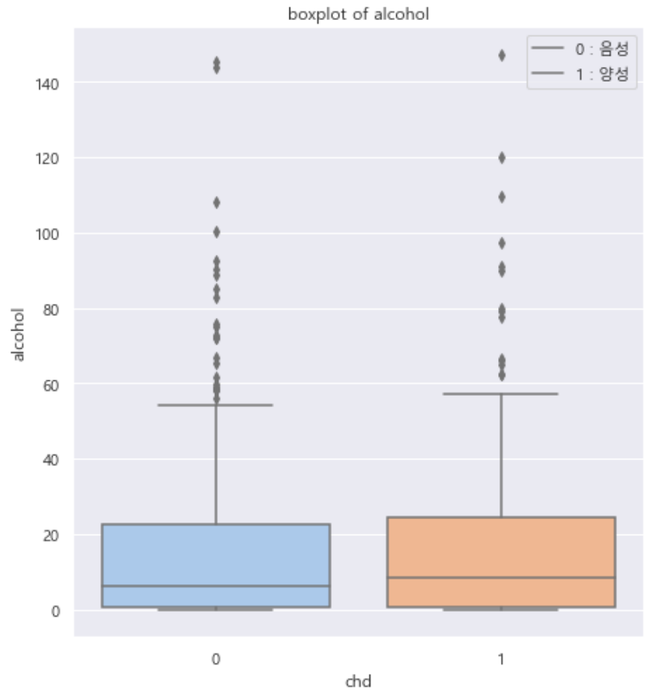
과체중 그룹에서 가장 많은 CHD 발생

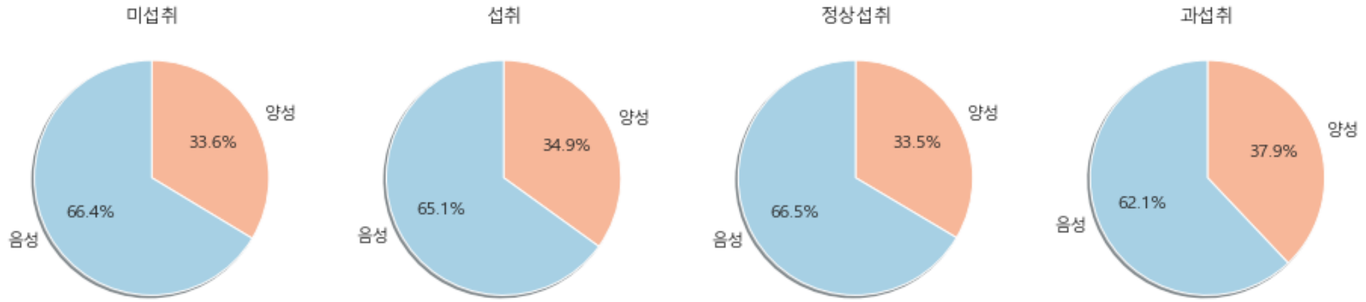
비만으로 갈수록 CHD 양성 비율 증가

03 EDA

08 alcohol







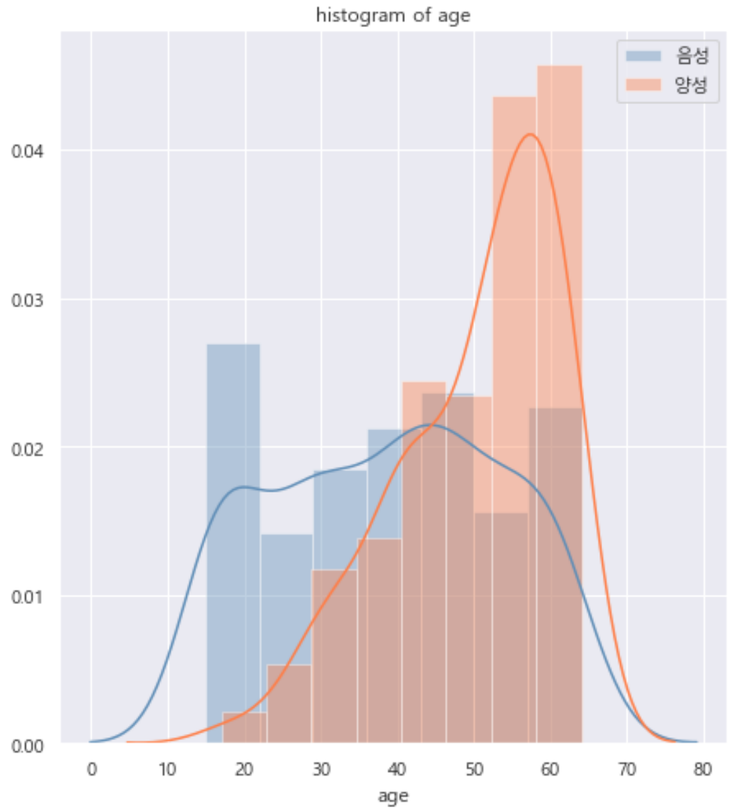
right-skewed 분포

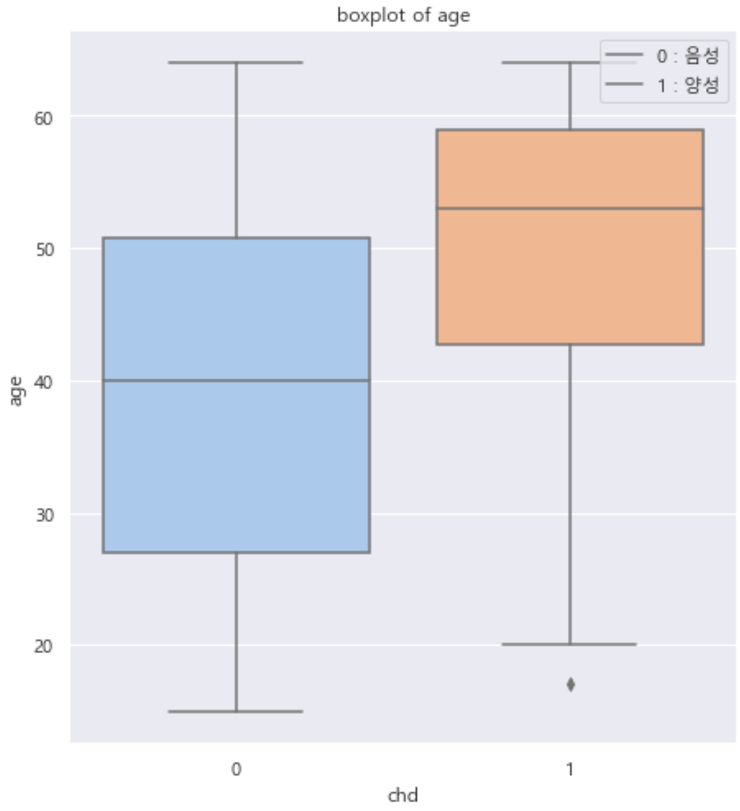
알코올 미섭취자/섭취자 그룹 비교
- 유의미한 차이 X

알코올 적정섭취자/과섭취자 그룹 비교
- 유의미한 차이 X

03 EDA

09 age





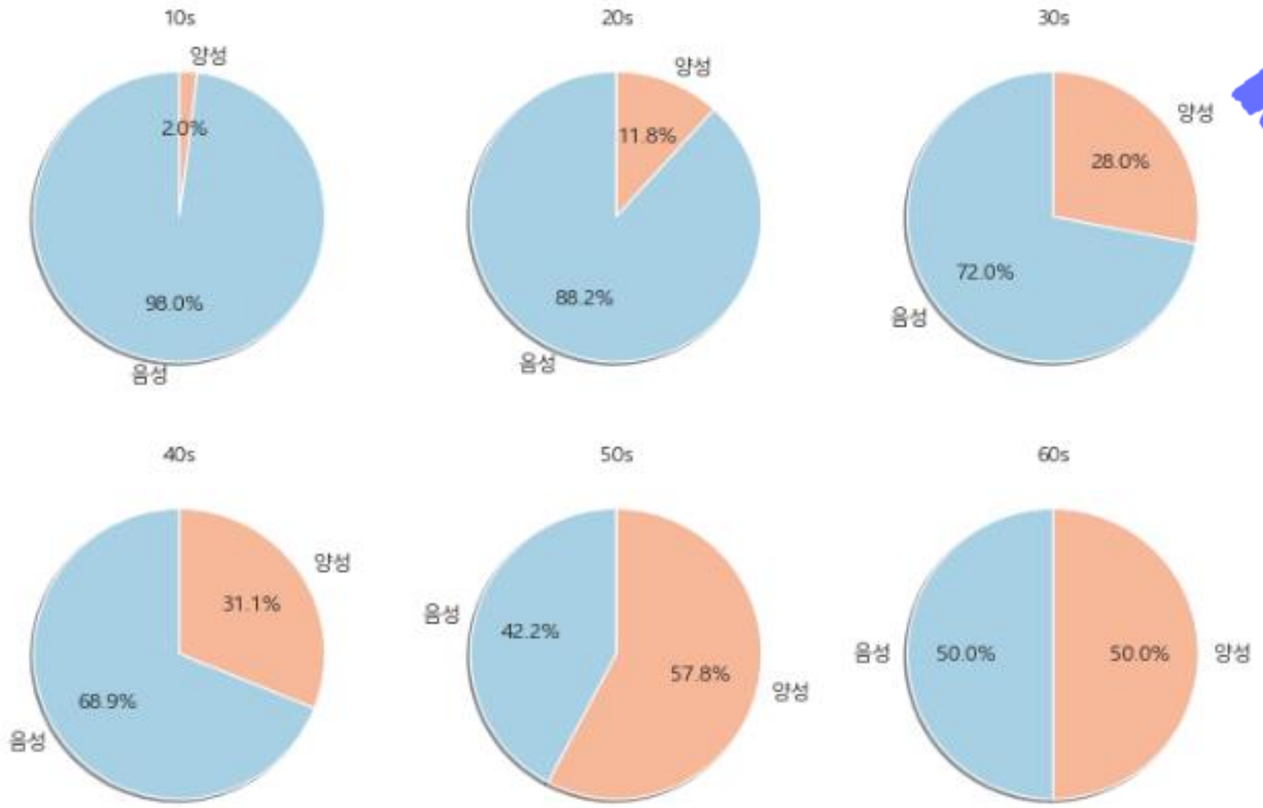
CHD 음성그룹 – 고르게 분포

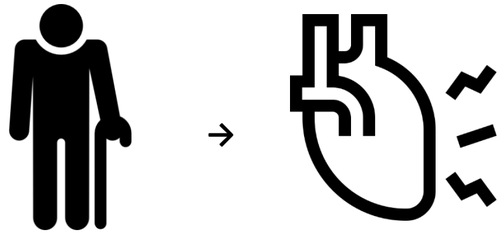
CHD 양성그룹 - 높은 나이대에 분포

CHD 발병 나이 평균↑

03 EDA

09 age





연령대가 높아질수록 급격히 CHD 양성 비율 증가

50대, 60대에서는 CHD 양성비율이 음성보다 더 많거나 같아짐

SOUTH AFRICA HEART DISEASE

4. 데이터 전처리

Data Preprocessing

04 데이터 전처리

4.1 변수 제거 및 결측치 처리

| 변수 제거

단순 행의 순서 나타내는 row.names 제거

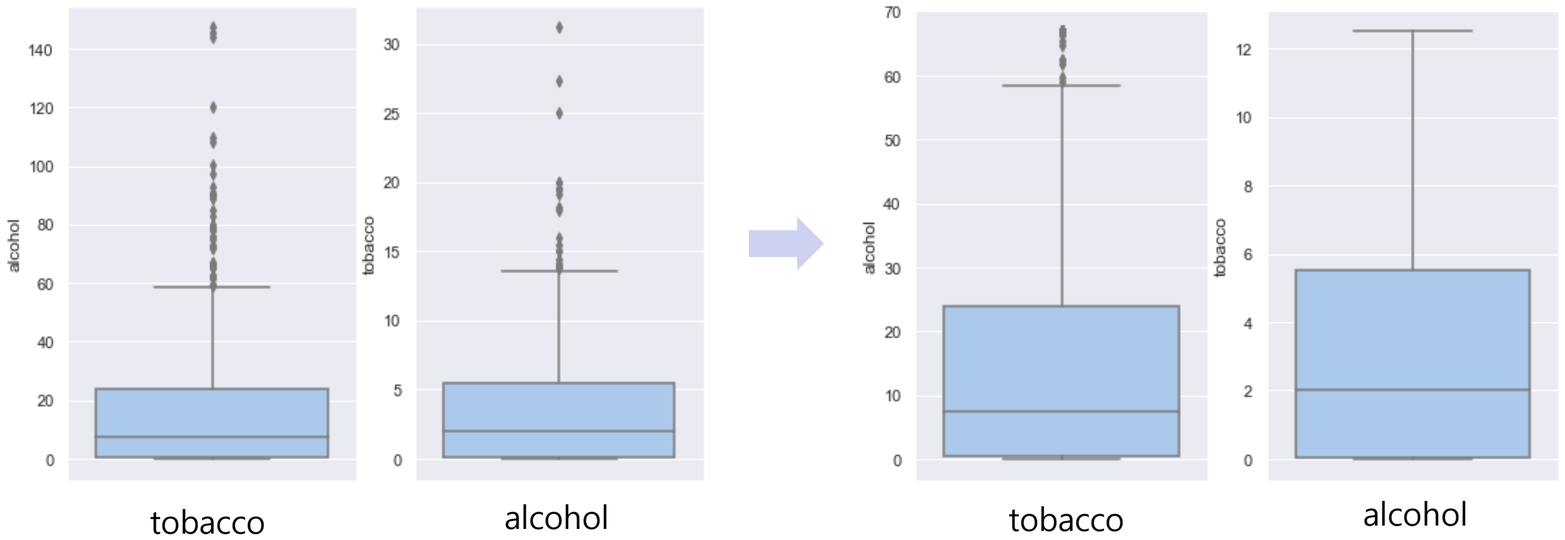
| 결측치 처리

데이터 상 결측치 없음

4.2 이상치

| 이상치 대체

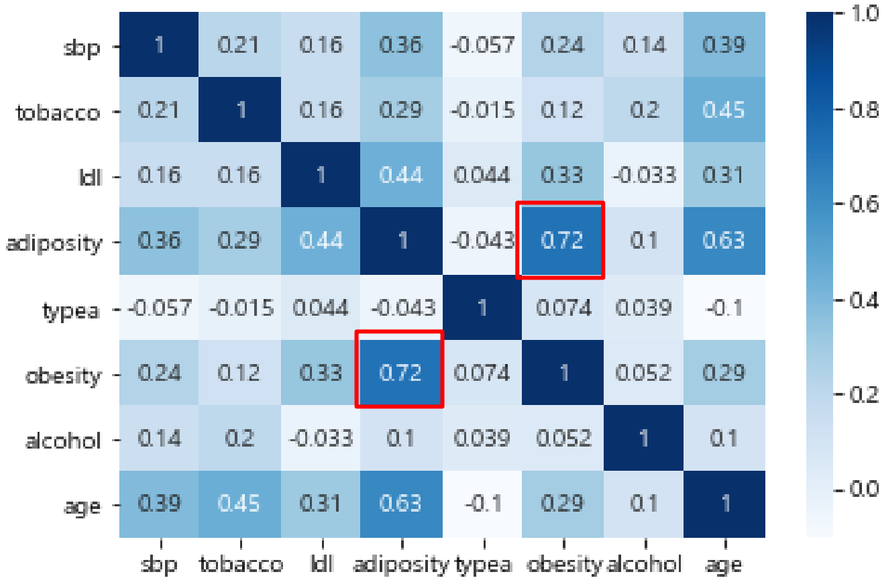
95 percentile 이상의 값을 95 percentile 값으로 대체



04 데이터 전처리

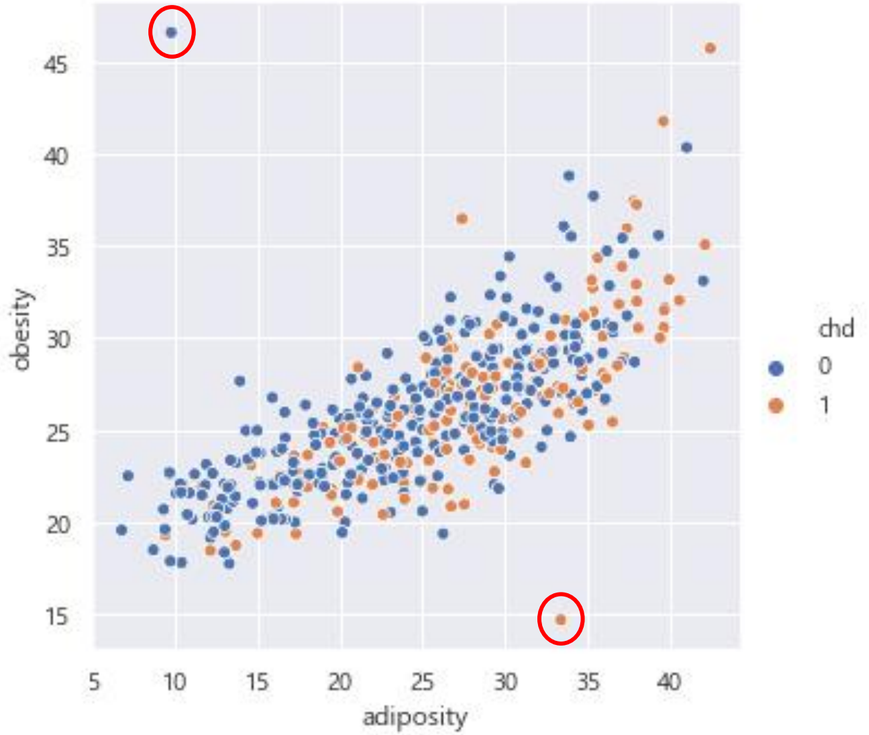
| 이상치 제거

adiposity와 obesity사이의 높은 상관관계



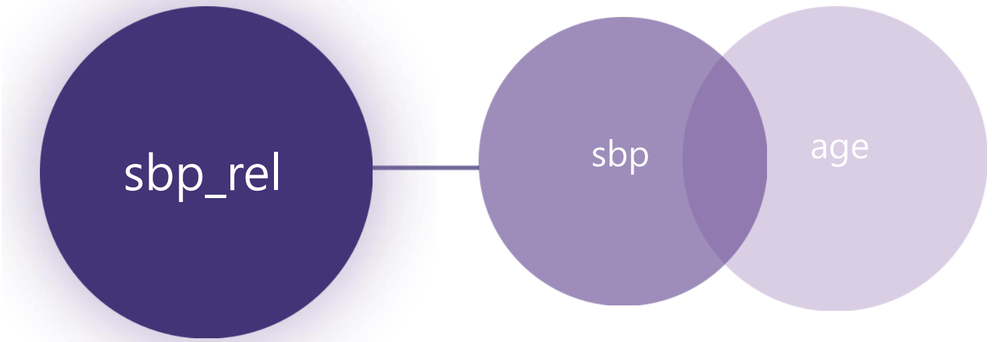
파생변수 생성 후 이상치 제거

weight_rate1 = adiposity/ obesity
weight_rate2 = obesity / adiposity



04 데이터 전처리

4.3 파생변수 생성 : sbp_rel



age	sbp	dbp
21-25	120.5	78.5
26-30	119.5	76.5
31-35	114.5	75.5
36-40	120.5	75.5
41-45	115.5	78.5
46-50	119.5	80.5
51-55	125.5	80.5
56-60	129.5	79.5
61-65	143.5	76.5

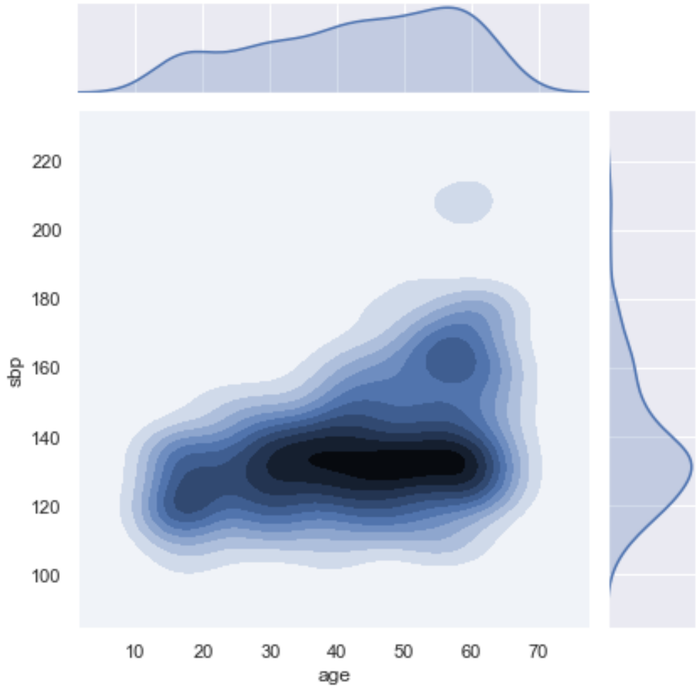
출처 : [OnHealth] 2021

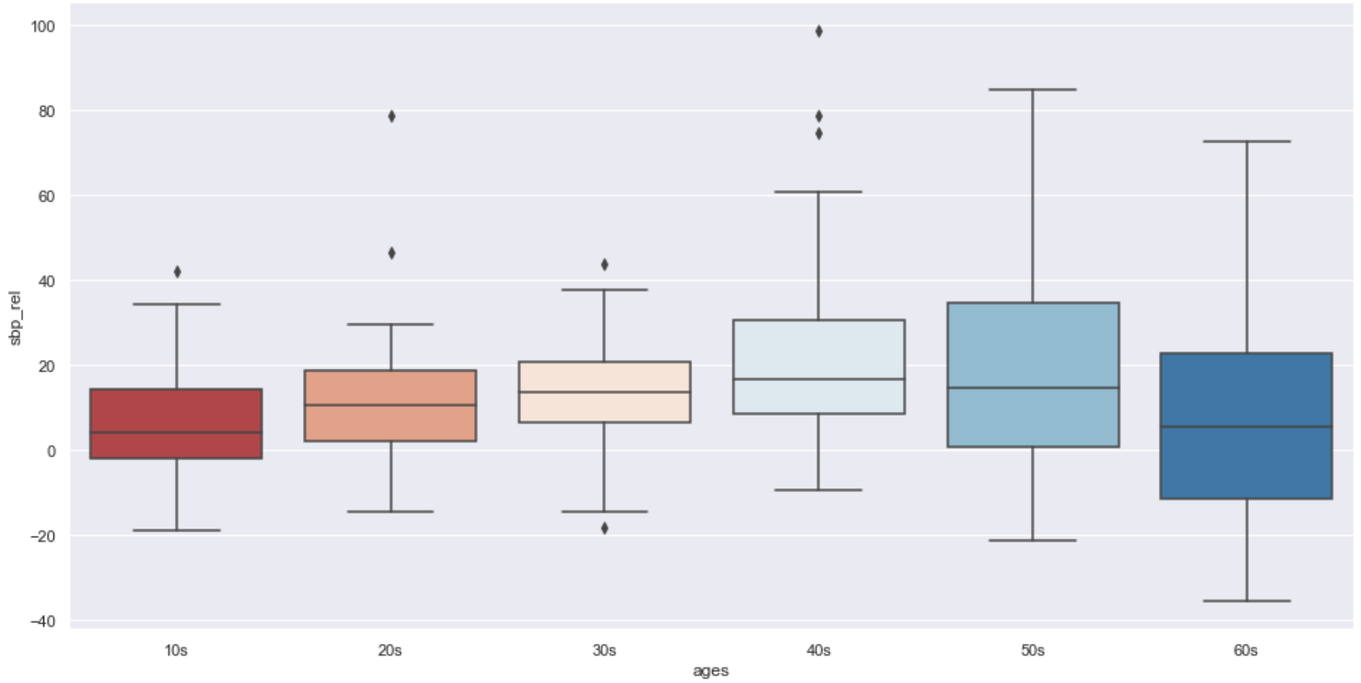
노화로 인한 혈압 상승은 자연스러운 현상 :: 연령에 따라 평균적 혈압 수치가 다름 고려 필요

나이(age) 변수에 따라 수축기 혈압(sbp) 변수와 상대적 수치를 나타내는 파생변수 sbp_rel을 생성

04 데이터 전처리

4.3 파생변수 생성 : sbp_rel





04 데이터 전처리

4.4 다중공선성 (VIF)

VIF Factor	Features
107.2	Obesity
101.1	SBP
50.5	Adiposity
29.3	TypeA
27.8	Age
7.9	Ldl
4.1	SBP_rel
2.5	Tobacco
1.8	Alcohol

VIF Factor	Features
21.8	Adiposity
17.5	Age
9.7	TypeA
7.9	Ldl
2.5	Tobacco
1.8	Alcohol
1.7	SBP_rel

SOUTH AFRICA HEART DISEASE

5. 모델링

Data Modelling

05 모델링

5.1 데이터 준비

tobacco	ldl	adiposity	famhist	typea	alcohol	age	chd	sbp_rel
12	5.73	23.11	1	49	97.2	52	1	34.5
0.01	4.41	28.61	0	55	2.06	63	1	0.5
0.08	3.48	32.28	1	52	3.81	46	0	-1.5
7.5	6.41	38.03	1	51	24.26	58	1	40.5
13.6	3.5	27.78	1	60	57.34	49	1	14.5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
0.4	5.98	31.72	0	64	0	58	0	40.5
4.2	4.41	32.1	0	52	18.72	52	1	84.5
3	1.59	15.23	0	40	26.64	55	0	56.5
5.4	11.61	30.79	0	64	23.97	40	0	-21.5
0	4.82	33.41	1	62	0	46	1	2.5

461 rows x 9 columns

05 모델링

5.1 데이터 준비

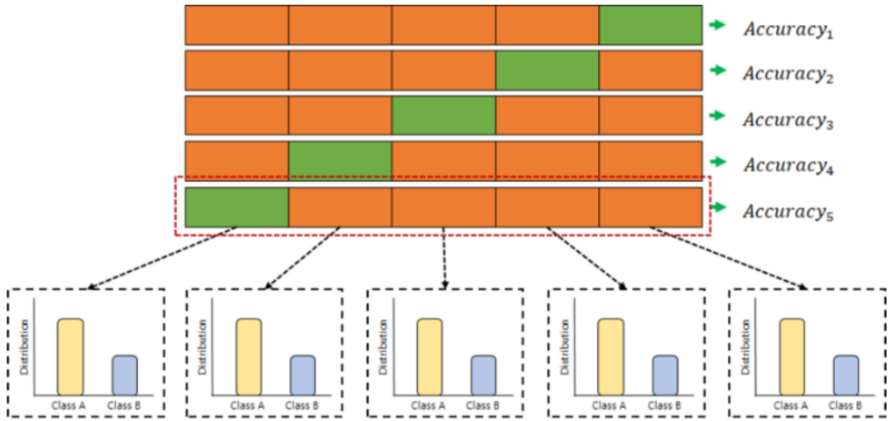
| 7:3 비율로 Train / Test 분리

| 계층별 k-겹 교차검증(Stratified k-fold cross validation)

데이터 양이 적어 별도의 검증 데이터를 학습에 재활용
→ 최대한 많은 데이터셋 활용

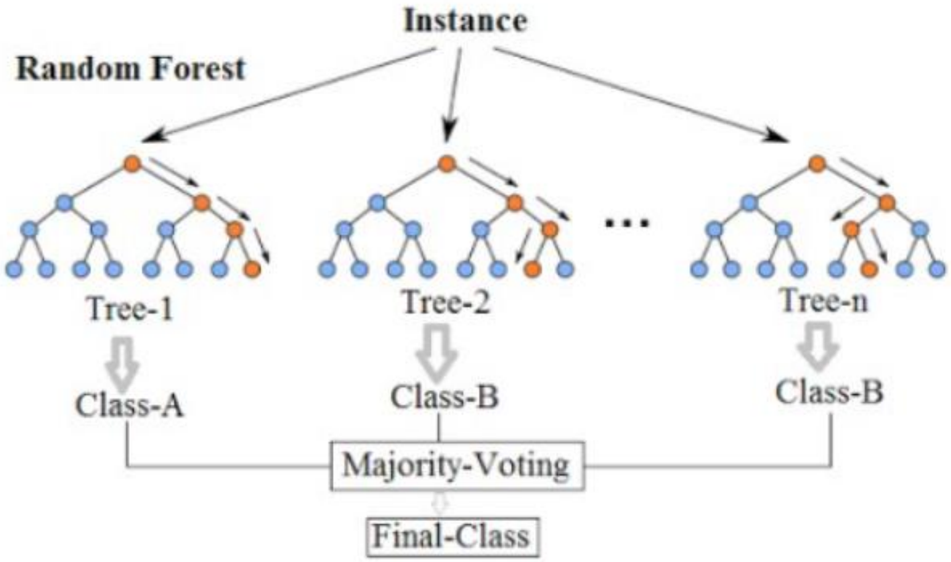
하나의 학습 / 검증 데이터로 이루어진 모델은 과적합(overfitting) 가능성이 높음
→ 교차 검증을 통해 일반화 된 모델을 생성

불균형 데이터
→ 양성과 음성 비율 유지해 외적타당도를 높여 최대한 일반화 가능하게 함



05 모델링

5.2 모델 후보 생성 1) RandomForest

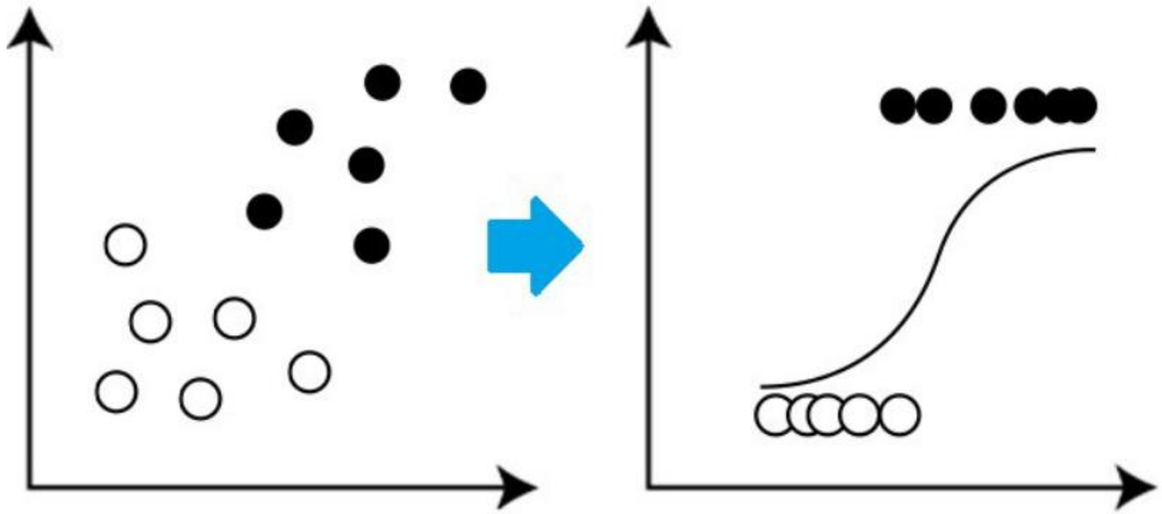


여러개의 decision tree를 형성하고 새로운 데이터 포인트를 각 트리에 동시에 통과시키며, 각 트리가 분류한 결과에서 투표를 실시하여 가장 많이 득표한 결과를 최종 분류 결과로 선택

랜덤 포레스트가 생성한 일부 트리는 overfitting될 수 있지만, 많은 수의 트리를 생성함으로써 overfitting 방지

05 모델링

5.2 모델 후보 생성 2) LogisticRegression

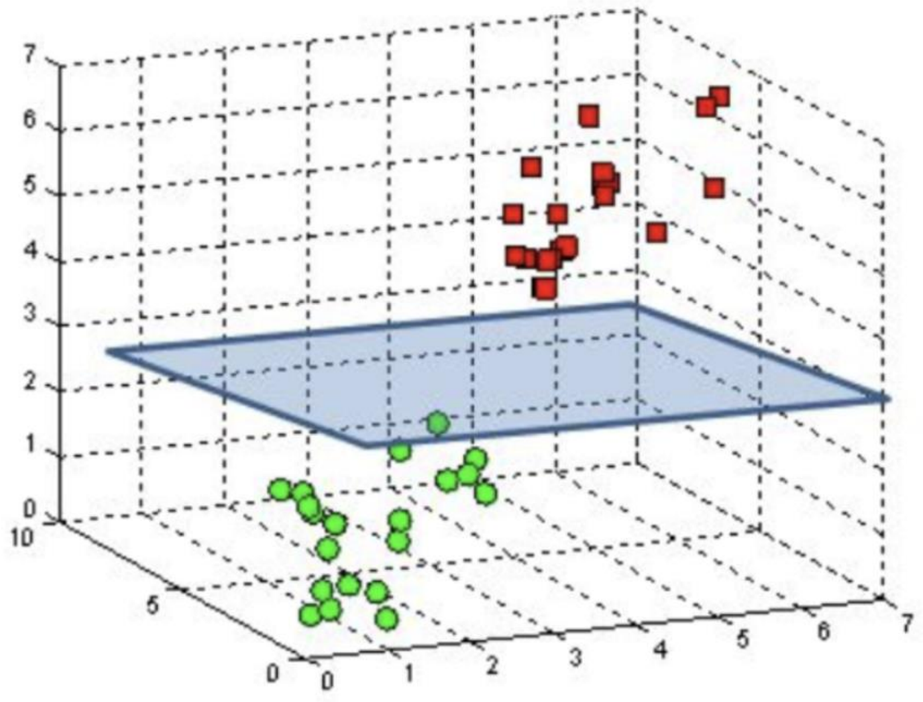


회귀를 사용하여 데이터가 어떤 범주에 속할 확률을 0에서 1 사이의 값으로 예측하고 그 확률에 따라 가능성이 더 높은 범주에 속하는 것으로 분류해주는 지도 학습 알고리즘

시그모이드 (Sigmoid) 함수에 속하며, 로짓 (로그+오즈)에 대한 해석 필요

05 모델링

5.2 모델 후보 생성 3) SVM



Support Vector Machine(SVM)은 학습 데이터를 비선형 매핑(Mapping)을 통해 고차원으로 변환, 이 새로운 차원에서 초평면(hyperplane)을 최적으로 분리하는 방법을 탐색

즉, 최적의 Decision Boundary(의사결정 영역)를 탐색하는 알고리즘

05 모델링

5.2 모델 후보 생성

| 그리드 서치(grid search) : 각각의 모델을 학습할 때 **최적 파라미터** 탐색

Model	Hyper-parameter Grid Search
LogisticRegression	'penalty' : ['l1', 'l2', 'elasticnet'] 'C' : [0.001, 0.01, 0.1, 1, 10, 100]
RandomForest	'max_depth' : [4, 6, 8] 'n_estimators' : [10, 50, 100] 'min_samples_split' : [10, 100]
SVM	'gamma' : [0.01, 0.1, 1, 10] 'C' : [0.01, 0.1, 1, 10]

5.3 모델 선정 기준

| F1 score $F_1 = 2 * \frac{precision * recall}{precision + recall}$

정밀도와 재현율의 조화평균

조화평균

: 차지하는 비중이 큰 경우의 bias를 줄여
불균형이 일어나도 모델의 성능을 정확하게 평가 가능

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

05 모델링

5.4 최종 모델 선정

| 최종 모델

로지스틱 회귀모형

- F1 score : 0.59
- C : 0.1
- 패널티 : 'l2 '

$$\boldsymbol{logit}(\boldsymbol{P}(\boldsymbol{chd}=\mathbf{1})) = -6.442 + 0.088(tobacco) + 0.114(ldl) - 0.017(adiposity) + 0.631(famhist) + 0.038(typea) - 0.003(alcohol) + 0.065(age) + 0.004(sbprel)$$

나이(age)가 한 단위 증가함에 따라 CHD 발병 대한 오즈 추정값은 exp(0.065) 즉, 1.07배 증가한다고 해석

| 변수중요도

Columns	Coef_scale
Intercept	-0.845
Age	0.666
Famhist	0.453
Tobacco	0.329
Typea	0.281
ldl	0.207
Sbp_rel	0.054
Alcohol	-0.053
Adiposity	-0.027

SOUTH AFRICA HEART DISEASE

6. 한계점 및 활용

Limiation & Utilization

06 한계점

| 사전조치

피실험자들이 CHD 양성 판정 이후 혈압 감소 치료와 위험요인을 줄이기 위한 조치를 받고 측정
특정 변수에서 CHD가 양성인 경우의 특징을 보여주지 못할 가능성 존재

| 단위와 측정기간의 부재

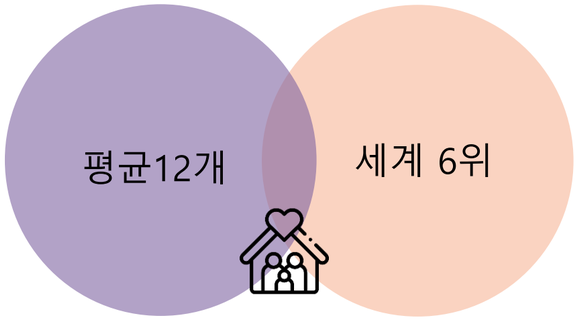
단위와 측정기간의 언급이 없는 경우 존재
ldl 변수 : 수치로 유추
alcohol 변수 : '최근의 알코올 소비량' → '최근'의 기준을 알 수 없어 단위의 유추 불가

| 결과 활용의 한계

남아프리카 공화국에서 백인남성을 대상으로 실시된 조사 → 인종•성별 편향
CHD 발병은 인종과 성별에 따라 위험요인의 영향과 그 정도가 다름
분석결과 전 지구적으로 활용 X

06 활용 방안

1) 효율적인 보험 관리



2018년 기준 대한민국의 가구당 평균 보험 상품의 개수

국민 1인당 보험료 지출 순위

경제력 대비 보험료 지출이 많은 편이라 평가

보험료 계산 기준

연령 >> 성별 >> 과거 입원 이력

표준화 회귀계수 중 age 변수 크기가 가장 높았던 결과와 동일

본 프로젝트에서 진행한 방식 차용하여

다양한 질병 데이터를 대상으로 모델링하고 예측

∨
∨

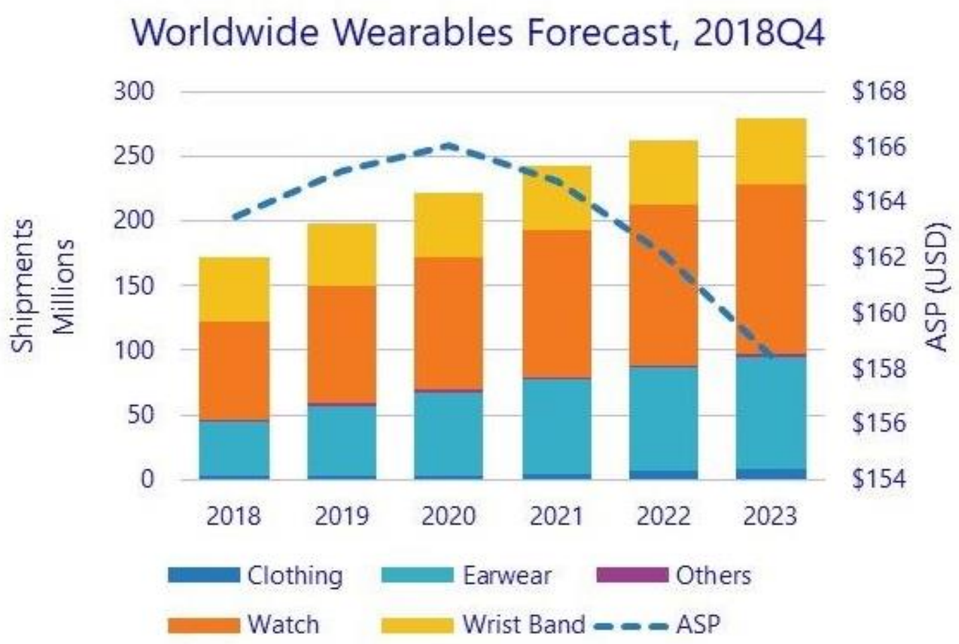
필요한 보험만 가입

현재 건강 상태에 필요한 보험을 판단

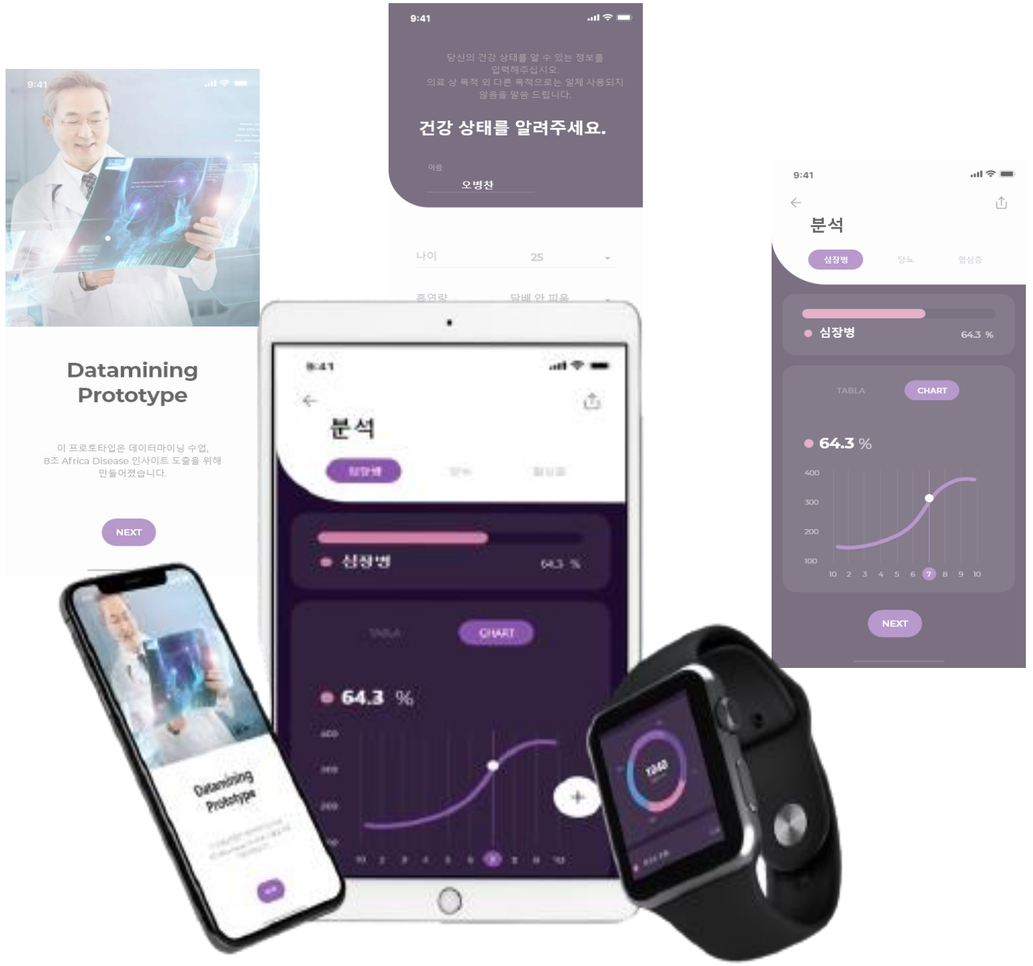
근거 없이 막연한 걱정만으로 들었던 보험의 수 줄이기

06 활용 방안

2. 웨어러블 기기 건강 정보 앱



Source: IDC 2019



Q & A

감사합니다

데이터마이닝 SAHeart 8조 | 오병찬 김상현 윤수연 최솔