K-Digital Training iM DiGital Banker Academy 5기

imagine More #, 건강한 금융

고객의 라이프를 읽다, 보험까지 잇다

이나경, 박민정, 박승욱, 전은진, 한민정 | 2조 투민정음

목차

01 프로젝트 개요

02 프로젝트 팀 구성 및 역할

03 프로젝트 수행 절차 및 방법

04 프로젝트 수행 경과

05 자체 평가 의견

1. 프로젝트 개요

기획 배경

건강에 대한 지속적 관심 증가

< 2019년 ~ 2023년 건강검진 종별 수검률 >

(단위: 천 명, %)

구 분		2019년	2020년	2021년	2022년	2023년	전년대비 증감률
일반 건강검진	대상인원(천 명)	21,717	21,446	22,837	22,862	**23,001**	0.6
	수검인원(천 명)	16,098	14,545	16,953	17,233	**17,462**	1.3
	수검률(%)	74.1	67.8	74.2	75.4	**75.9**	0.7
암검진	대상인원(천 명)	23,123	22,973	23,937	23,456	**23,313**	△0.6
	수검인원(천 명)	12,891	11,389	13,555	13,657	**13,932**	2.0
	수검률(%)	55.8	49.6	56.6	58.2	**59.8**	2.7
영유아 건강검진	대상인원(천 명)	2,712	2,488	2,468	2,558	**2,309**	△9.7
	수검인원(천 명)	2,103	2,064	2,149	2,064	**1,770**	△14.2
	수검률(%)	77.6	83.0	87.1	80.7	**76.7**	△5.0

주1) 수검률(%) = (수검인원/대상인원) × 100

2) (영유아수검률) '20~'22년 코로나로 인한 검진기간 연장(6개월)으로 일시적 증가, '23년은 코로나 시기 이전으로 회복

- 국민건강보험공단 보도자료 -「2023 건강검진 통계연보」

2023년도 건강검진 수검률은
일반검진 75.9%, 암검진 59.8%로
각각 전년대비 0.6%p, 2.7%p 증가

코로나 기간(2020년)을 제외하였을 때
건강검진의 수검률은 지속적으로 증가하는 추세
→ 건강에 대한 관심의 증가

기획 배경

질병과 빅데이터의 연결

- **목표**

  → 건강검진 데이터를 군집분석하여 MBTI처럼 건강검진 데이터를 즐길 수 있도록 서비스 제공

  → 카드 데이터 기반으로 생활패턴&소비패턴과 서비스 연결

질병과 건강보험 빅데이터 연결로 국민건강 향상된다!

- 국민건강보험공단-질병관리청, 건강정보 빅데이터 연계·활용과 협력 확대 방안 논의 -

- 감염병 대응부터 만성질환까지, 건강정보 통합 기반 과학적 정책 추진 -

□ 국민건강보험공단(이사장 정기석, 이하 '공단')과 질병관리청(청장 지영미, 이하 '질병청')은 21일 공단 본부에서 건강정보 빅데이터의 효과적인 연계와 활용 방안에 대해 논의한다.

□ 양 기관은 지난 2021년 4월 업무협약(MOU)을 체결한 이후, '코로나19 빅데이터(K-COV-N)*'를 공동으로 구축·개방하여 현재까지 총 36건의 연구성과**를 창출하는 등 국민 건강을 지키기 위한 긴밀한 협력을 지속해왔다.

* 질병청의 코로나19 확진자 및 예방접종 정보와 공단의 건강보험 자료를 결합

** 코로나19 예방접종 실시 기준 및 '25년 상반기 추가 접종 시행 근거 활용 등

○ 아울러, 2022년부터 협력해 온 결핵 빅데이터(K-TB-N)도 올해 9월 개방을 앞두고 있어, 결핵 퇴치 가속화를 위한 역학연구 활성화에 기여할 것으로 기대된다.

□ 이번 협의에서는 건강정보 데이터 결합으로, 감염병 외에도 만성질환까지 협력 분야를 확대하여 기존의 치료 중심 정책에서 예방 중심 정책으로 방향을 전환하기 위한 주요 과제를 논의한다.

① 국가 인플루엔자 정기보고서 공동발간 추진

매년 많은 환자가 발생하는 인플루엔자 유행에 체계적으로 대응하기 위해, 양 기관은 감염병 감시부터 백신 효과 평가, 의료이용 현황, 질병 부담까지 폭넓은 내용을 담은 정기보고서를 함께 발간할 예정이다. 이 보고서는 예방접종 정책과 건강보험 제도 개선에 중요한 근거 자료로 활용될 수 있다.

- 국민건강보험공단 보도자료 - 질병과 건강보험 빅데이터 연결로 국민건강 향상된다!

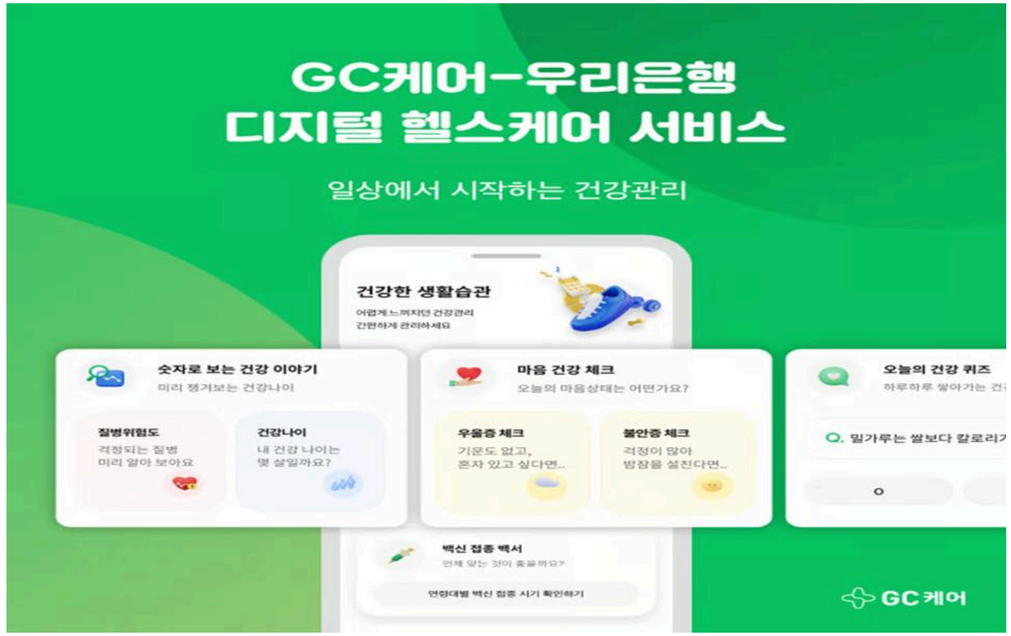
국민건강보험공단과 질병관리청의 공동 연구 등 빅데이터를 통해 질병에 대응하려는 연구가 지속적으로 논의됨

→ 국가건강검진 제도 개선을 위한 빅데이터 분석, 만성질환 데이터 통합 분석 등을 포함하여 다양한 분야에서의 연구가 진행

기획 배경

금융 X 헬스케어 서비스 현황

- 최근 금융앱은 자산 관리를 넘어, 고객 데이터를 분석하여 헬스케어 서비스를 제공
  → 장기적 고객 유치 및 리스크 관리 수단

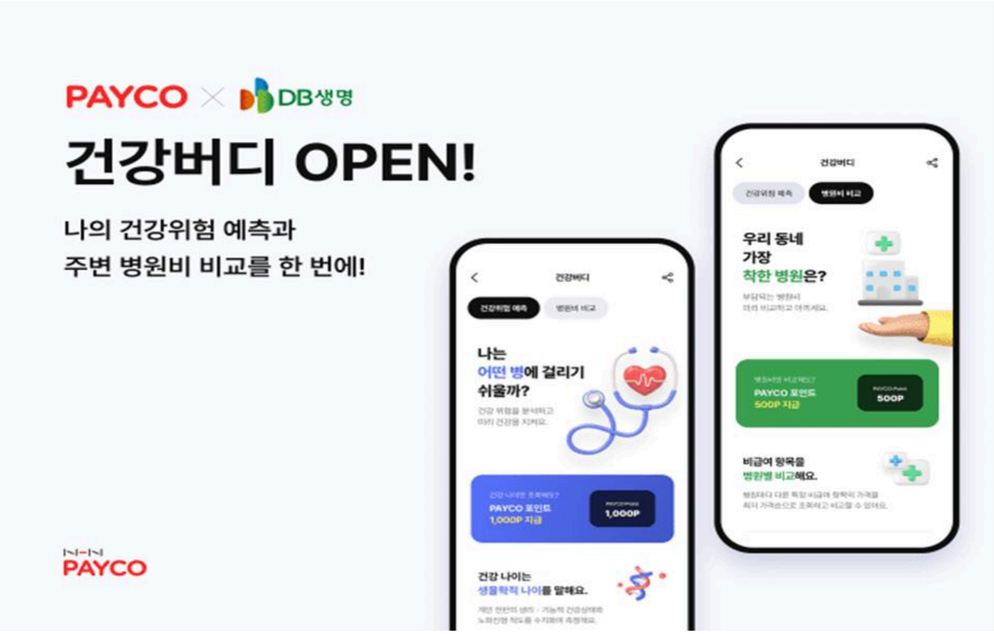


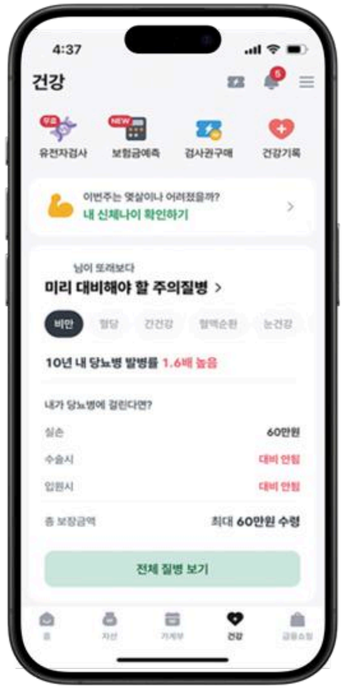
[KB라이프]
기존의 건강정보 분석 서비스에 더해
영양제 분석 서비스, 비급여 병원비 비교 서비스를
새롭게 도입
→ 비급여 진료 항목과 전문의 정보 비교 가능

[우리은행 - GC케어]
우리은행와 GC케어의 전략적 업무 협약
간단한 문진만으로도 고객의 질병 위험도,
건강 나이, 피부 나이 등을 확인 가능
→ 개인별 건강관리 가이드 제공

기획 배경

금융 X 헬스케어 서비스 현황





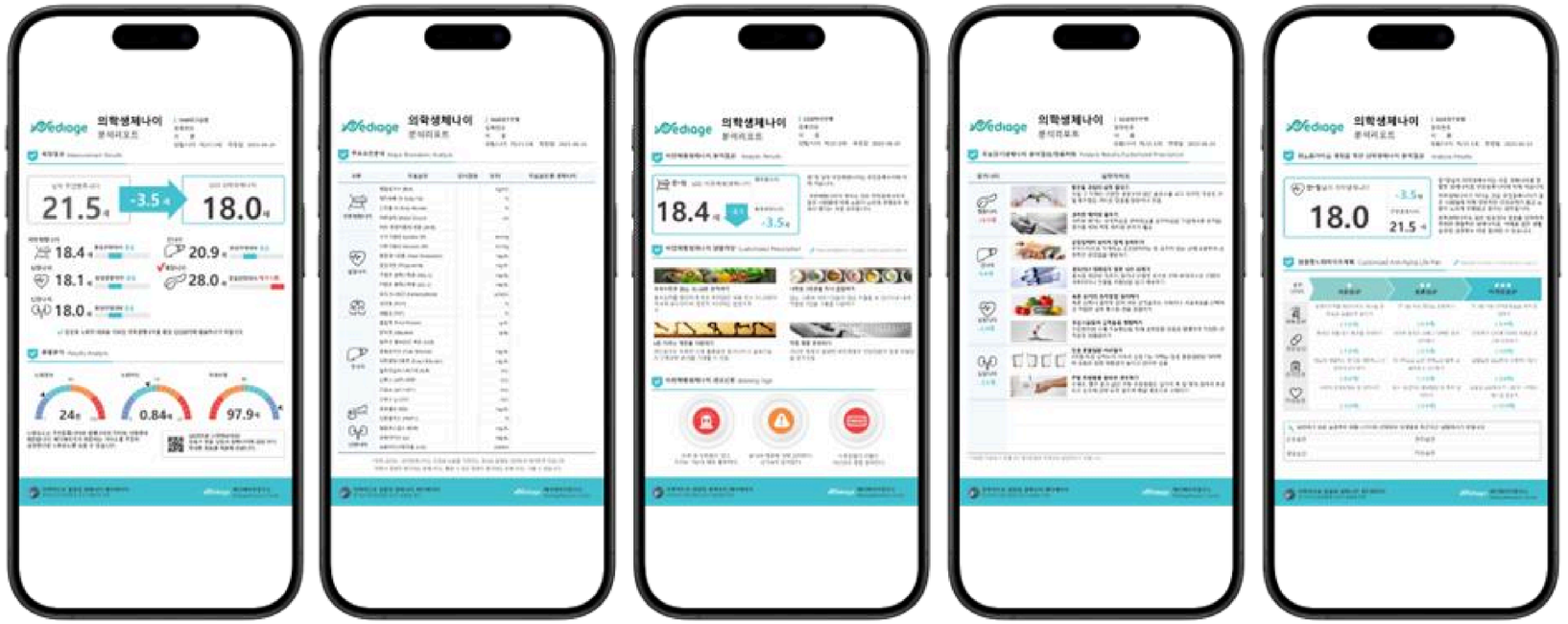
[건강버디 - PAYCO X DB생명]
앱을 통해 건강위험 예측과 비급여 진료비 비교
건강 나이, 7가지 주요 질환, 9가지 암 관련
건강 등급 , 맞춤형 영양성분 추천
→ 정상, 경계, 이상의 3단계로 시각화
비급여 항목별 진료비 비교 기능 제공

[뱅크샐러드]
금융, 건강 통합 분석을 통해 맞춤형 자산관리,
건강 컨설팅 서비스를 제공
업계 최초로 개인 건강 데이터를 활용
→ 맞춤형 보험 분석을 통해 효율적 보장 설계
건강검진 정보, 연령, 성별 가족력 등을 AI로 분석
→ 발병 위험 질병 예측

기획 배경

iM뱅크 헬스케어 서비스

- **생활금융 플랫폼 iM샵**
  자체 플랫폼 모바일 어플리케이션을 통한 생활 밀착형 서비스



[생체나이 분석 서비스]

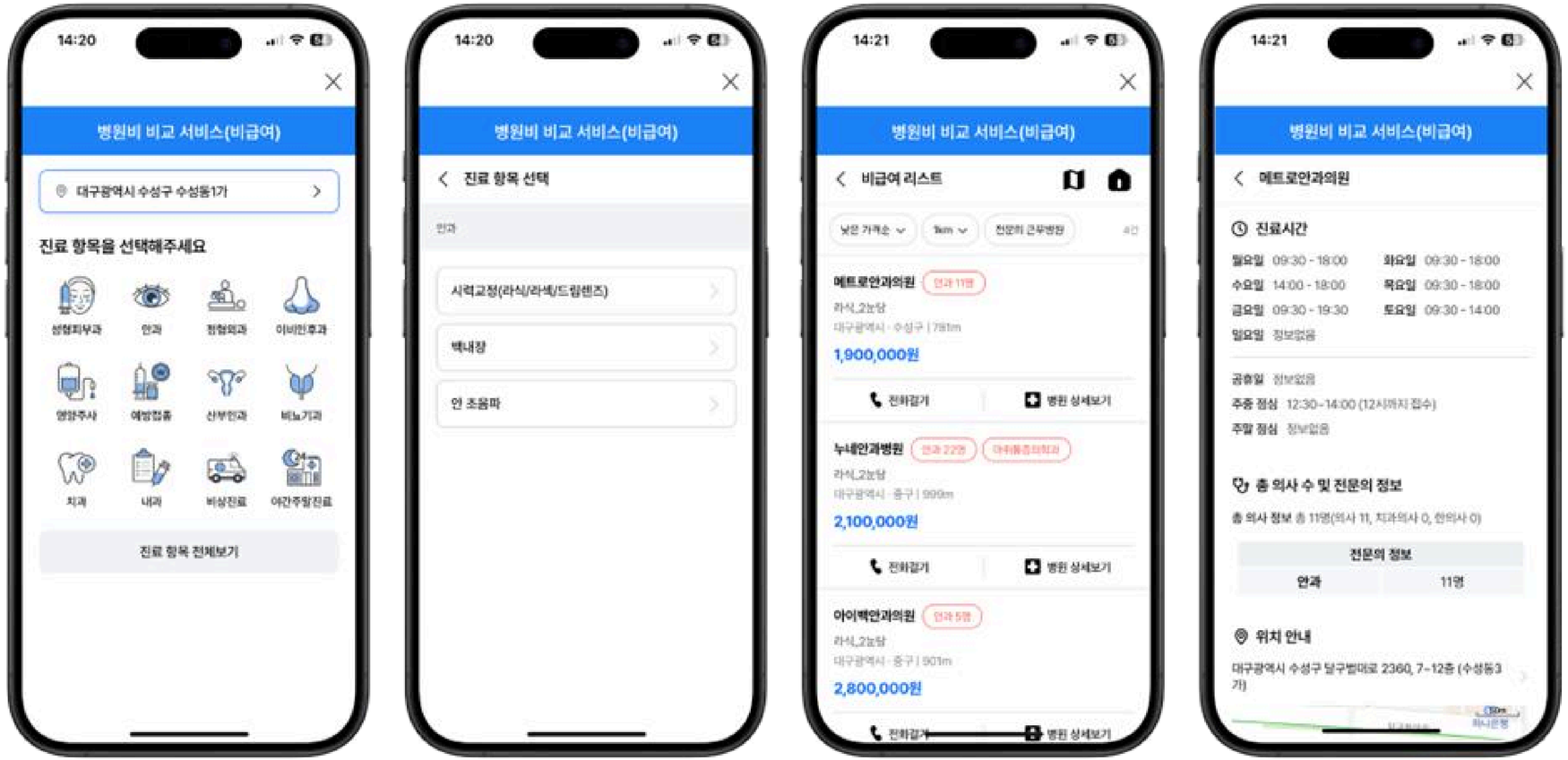
건강검진 정보를 기초로 생체 나이를 종합적 분석
비만 체형, 간, 심장, 췌장, 신장 기능을 나이로 측정
노화 및 기대 수명를 포함한 종합 생체 나이 측정

사용자 건강 수준에 대한 직관적 이해에 유용

하지만,
서비스의 활용 범위가 건강정보 조회에 국한
금융 정보와의 연계 부족

기획 배경

iM뱅크 헬스케어 서비스



[비급여 병원비 조회 서비스]

(주)헬스피디아와 연계하여 의료수가조회

비급여 진료항목별 가격 비교를 통해 병원비 절약

사용자 위치 기반 병원비 조회 가능

특정 비급여 진료 항목별 병원 정보 제공

→ 실제 소비자의 의료비 절감에 기여할 수 있는 실용적 기능 제공

기획 배경

iM뱅크 헬스케어 서비스

[데이터 기반 보험 연계 전략]
iM뱅크의 고객 데이터와 건강 데이터를 접목

→ 생활습관과 소비패턴을 입체적으로 파악
군집 분석을 통해 상호 연관성을 도출

나아가 고객 맞춤형 보험 추천 체계를 설계

→ 고객 건강과 재무 상태를 함께 관리하는
데이터 기반 라이프 스타일 금융 서비스로의 진화

→ iM뱅크 플랫폼 차별화 및 장기적인 수익 모델 다각화의 중요한 전략적 전환점

프로젝트 팀 구성 및 역할

투민정음 팀원

이나경

EDA
분류 모델

박민정

EDA
분류 모델

박승욱

EDA
분류 모델

전은진

데이터 전처리
군집분석

한민정

서비스 개발 및 구축

프로젝트 수행 절차 및 방법

프로젝트 수행 절차 및 방법

사전기획	데이터 수집	데이터 전처리	모델링	서비스 구축	발표
4/15 ~ 4/17 프로젝트 기획서 작성 제공된 데이터 파악	4/17 ~ 4/18 주제에 맞는 데이터 수집	4/18 ~ 4/19 데이터 전처리 EDA	4/20 ~ 4/23 모델링 및 분석	4/23 ~ 4/24 결과 도출 보고서 작성 발표 준비	4/25 최종 보고서 제출 발표

총 수행 기간: 4/15 ~ 4/25

사용한 데이터

국민건강보험

1. 국민건강보험_건강검진정보
https://www.data.go.kr/data/15007122/fileData.do

iM 뱅크

1. iMBANK_CUSTOMER_DATA_2023
2. iMBANK_CARD_DATA_2023

국립보건연구원

1. https://www.nih.go.kr/ko/main/contents.do?menuNo=300567

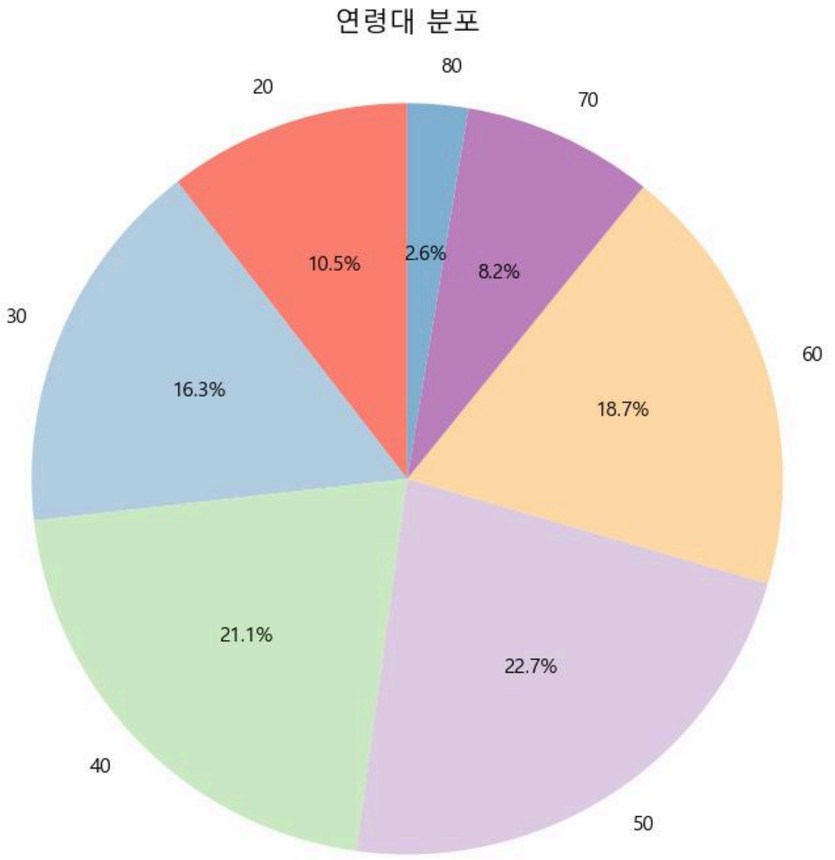
코호트 연구 데이터

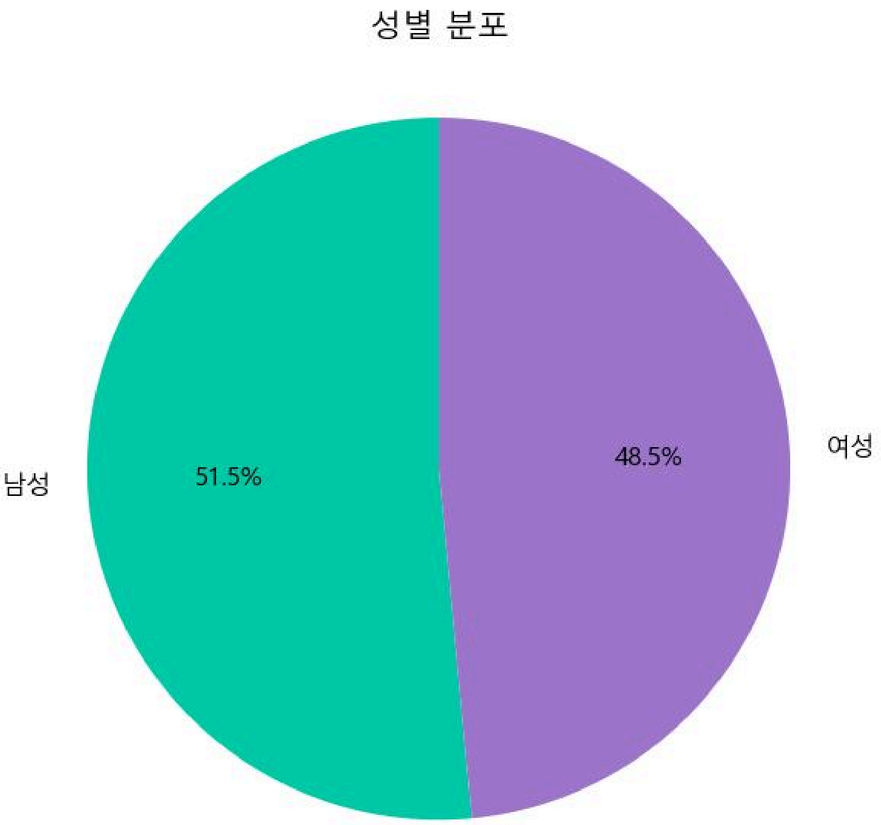
- 고혈압, 당뇨, 고지혈증에 대한 데이터

EDA

건강검진 데이터 분포 파악

연령대 및 성별 분포

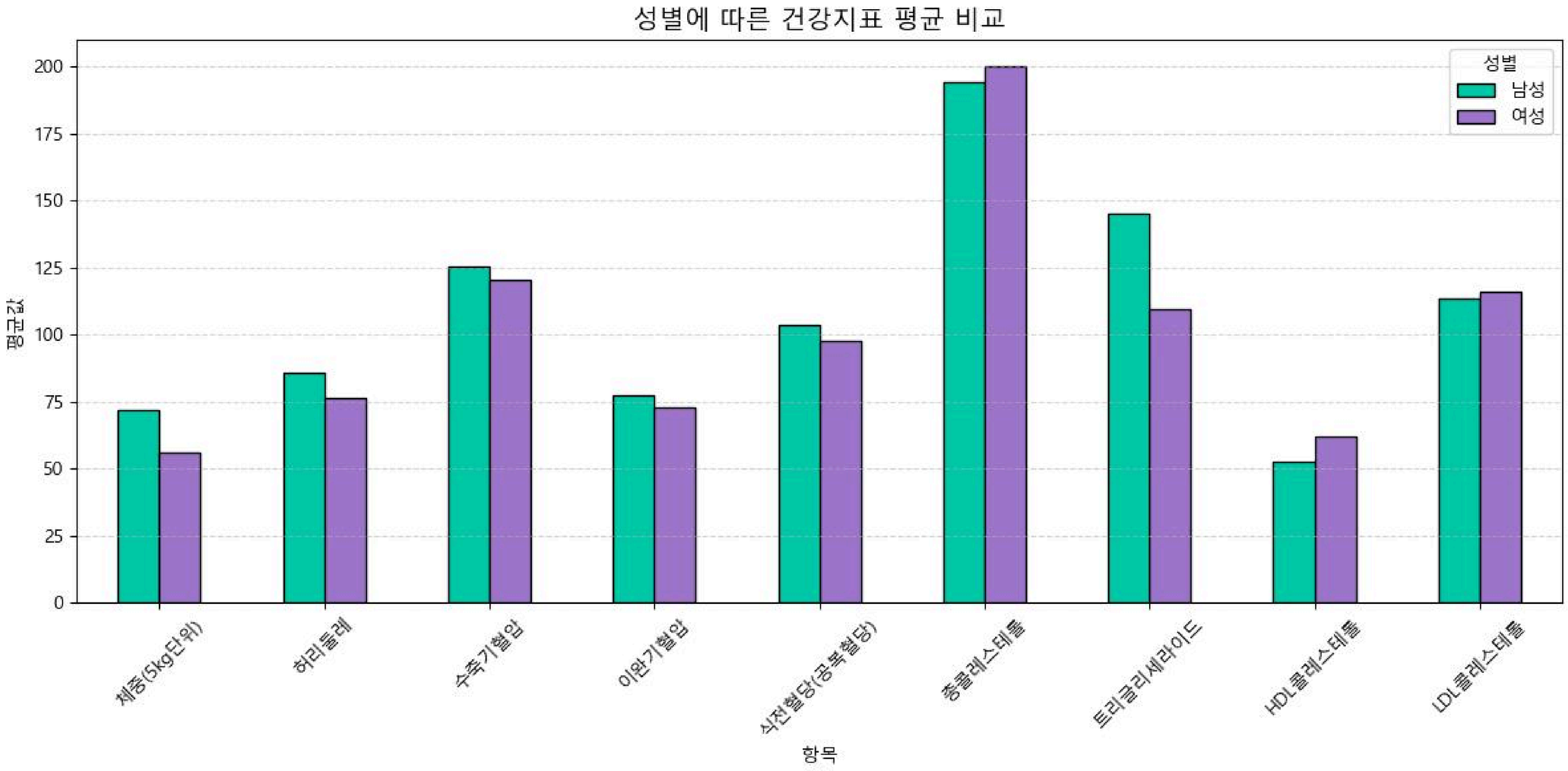




EDA

건강검진 데이터 분포 파악

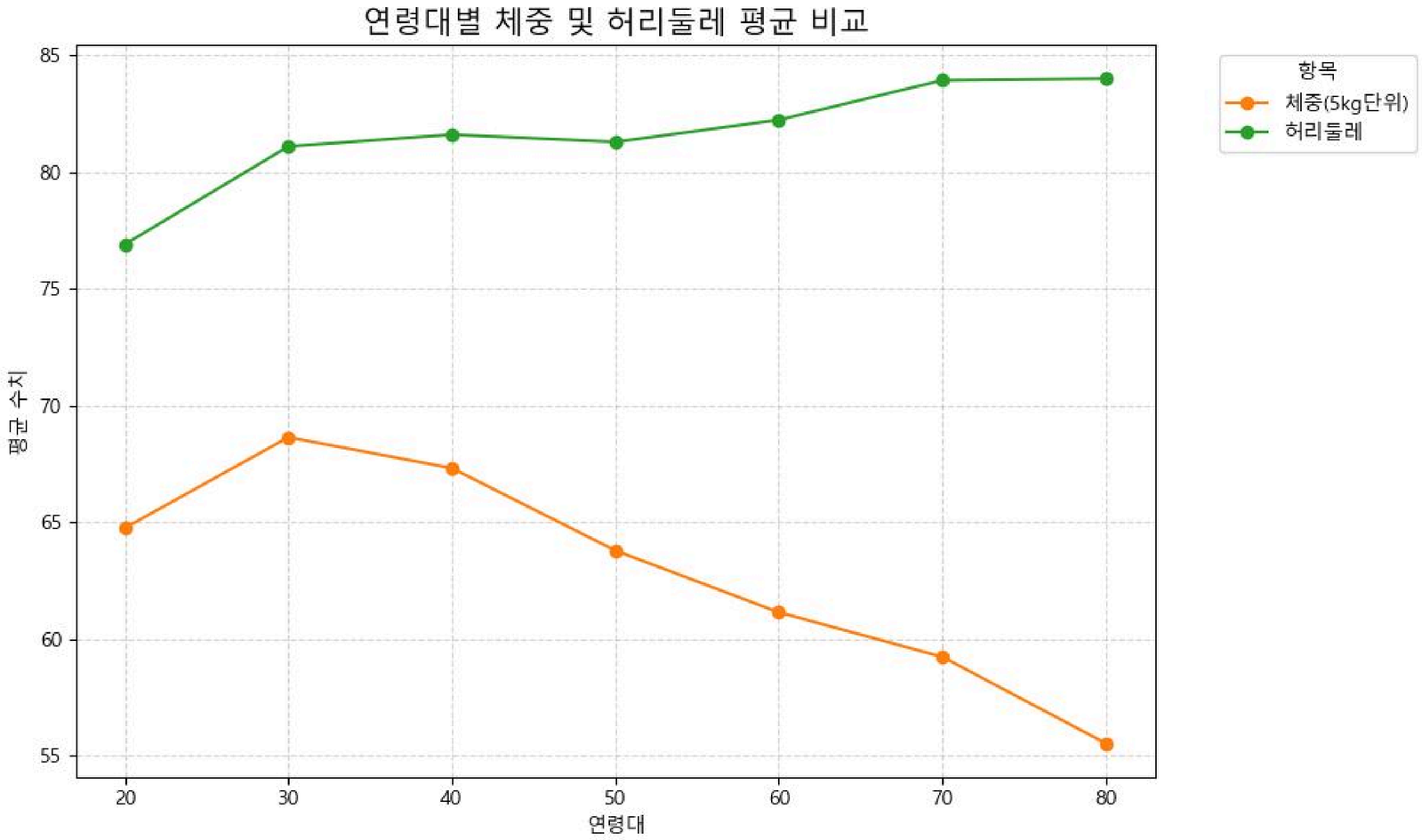
성별에 따른 건강 지표



EDA

건강검진 데이터 분포 파악

연령에 따른 체중 & 허리둘레



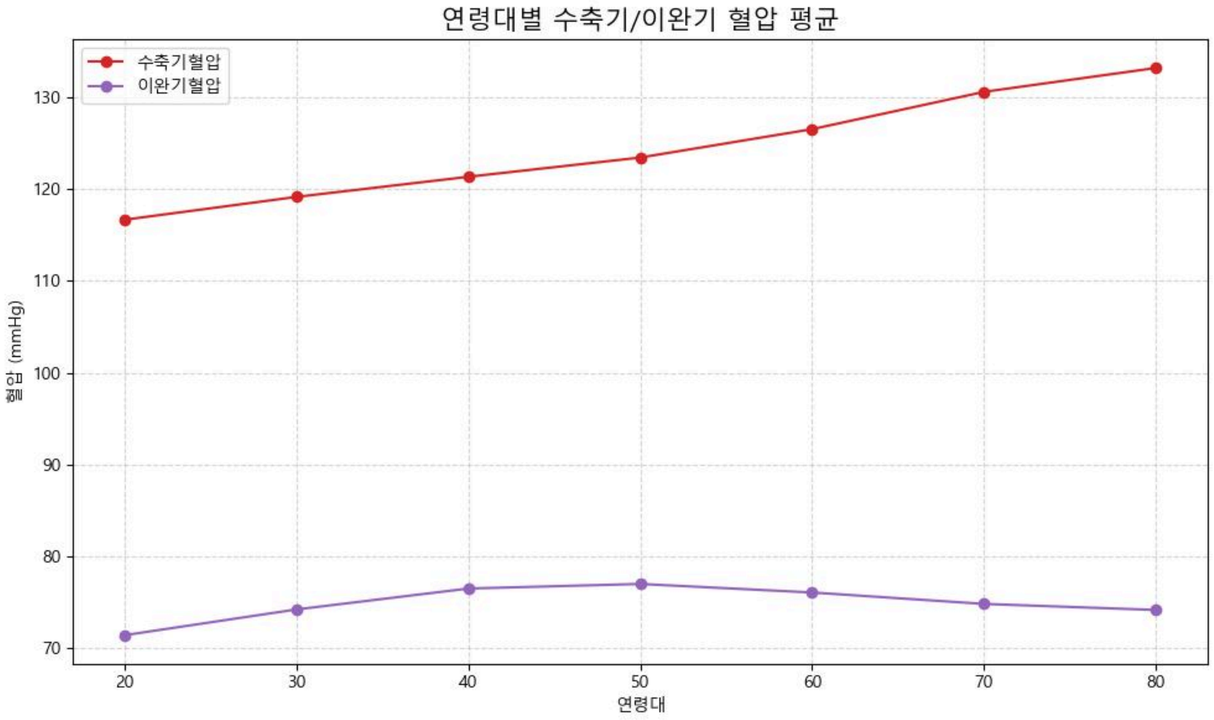
→ 연령대가 상승할수록 평균 체중은 줄어듦

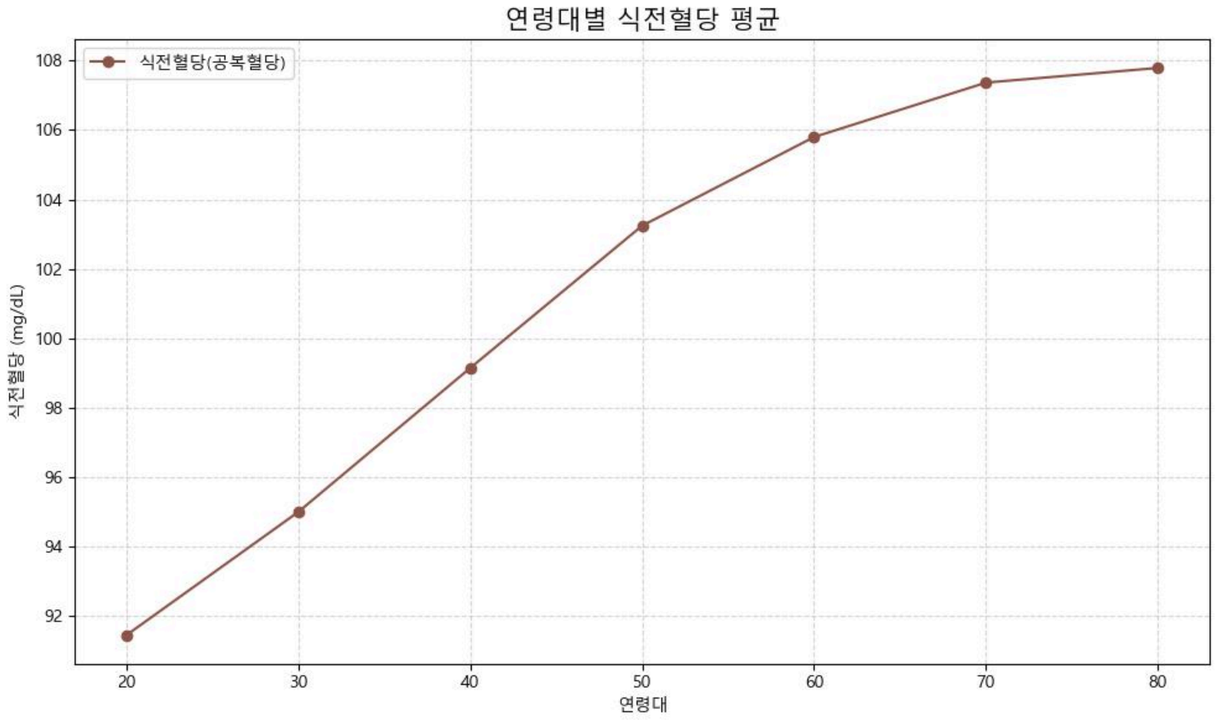
→ 함께 허리둘레도 증가하는 모습이지만 그 변화가 20~80대의 차이가 10cm에 못 미치는 수준

EDA

건강검진 데이터 분포 파악

연령에 따른 혈압 & 혈당



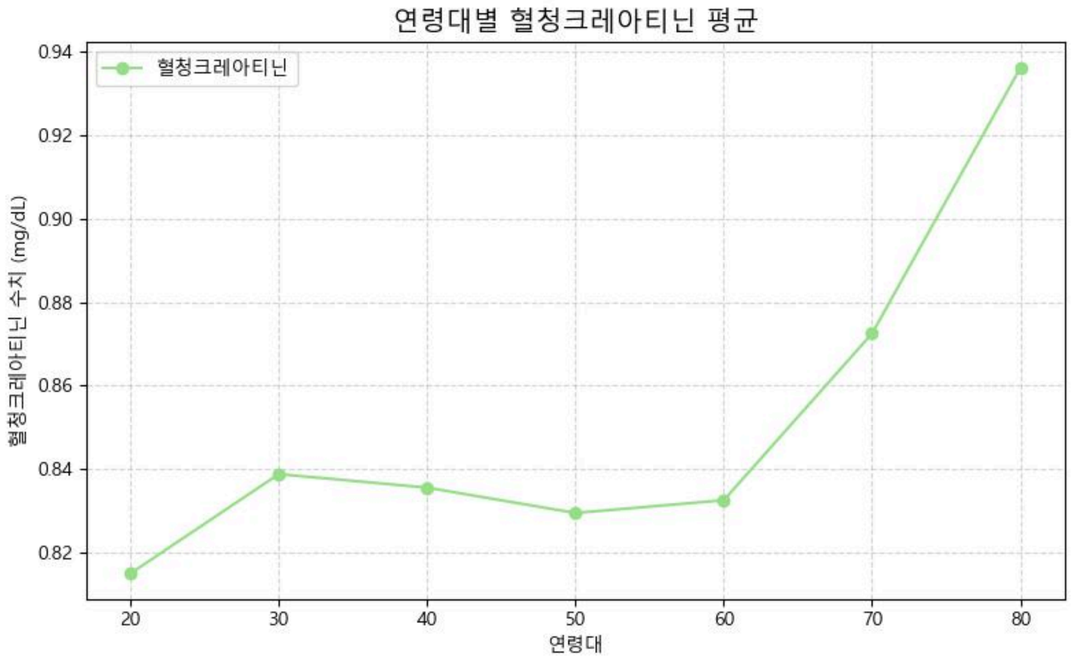


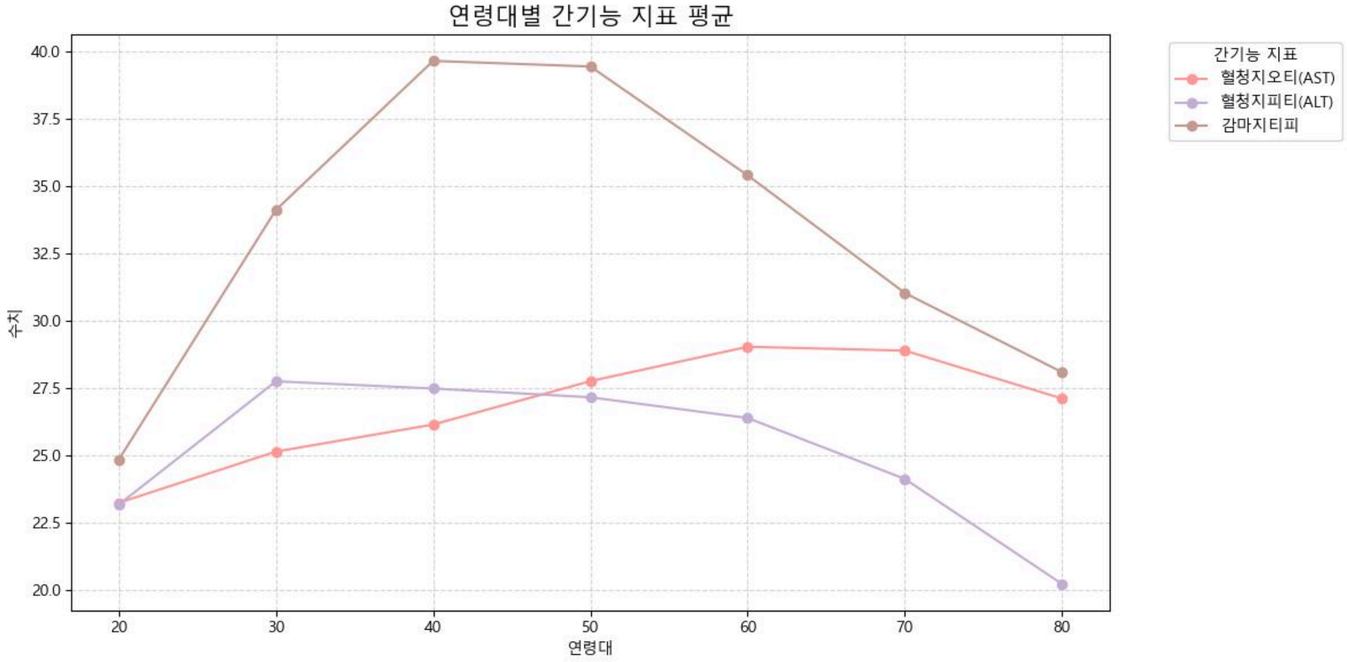
→ 연령대가 오를수록 수축기 혈압이 오르고 식전혈당이 오르는 양상
→ 연령대가 높아질수록 고혈압, 당뇨에 위험 증가

EDA

건강검진 데이터 분포 파악

연령에 따른 신장 및 간 기능





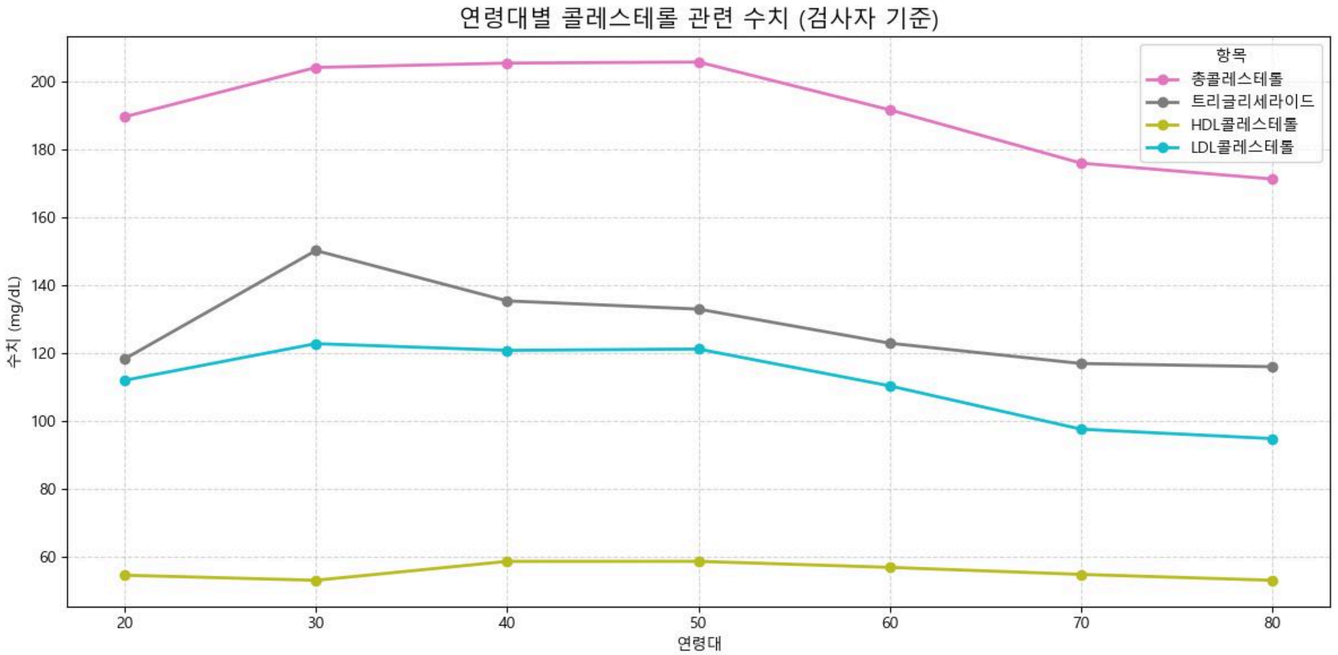
→ 두 지표 모두 구체적인 해석이 필요

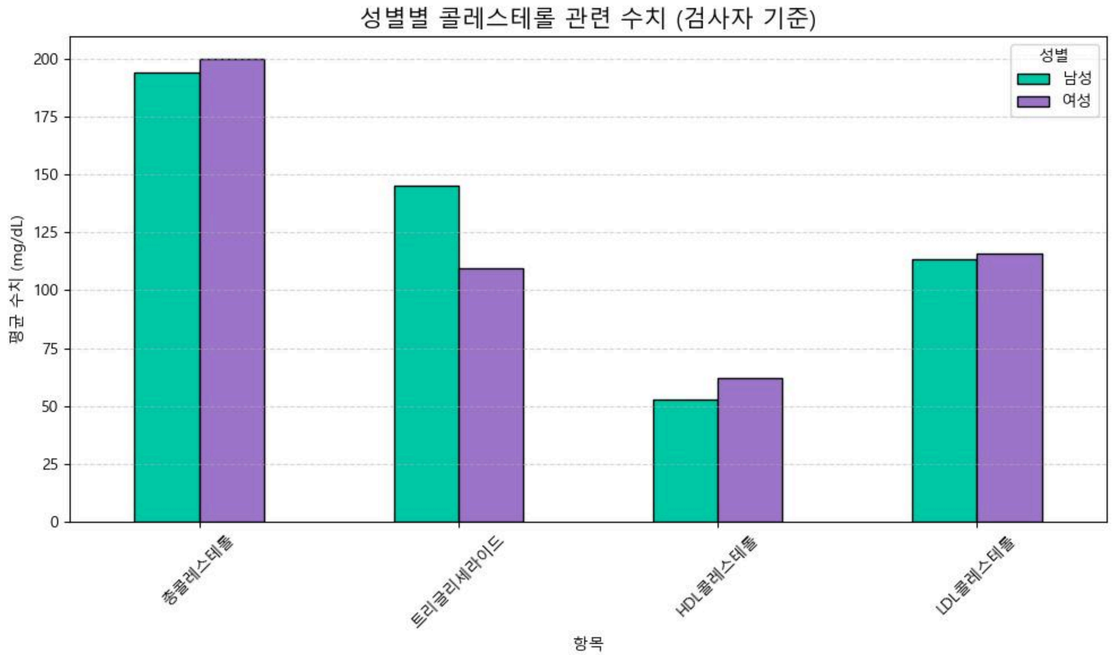
→ 지나치게 높을 경우 기능의 손상을 의미, 지나치게 낮을 경우 기능의 저하를 의미

EDA

건강검진 데이터 분포 파악

연령대 및 성별에 따른 콜레스테롤
(검사자 기준)



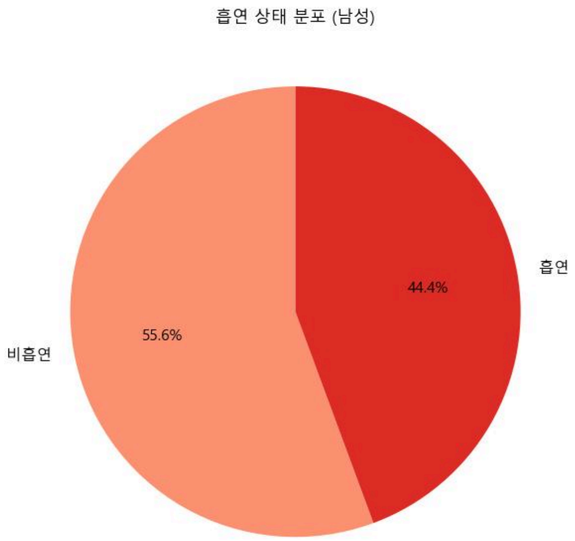


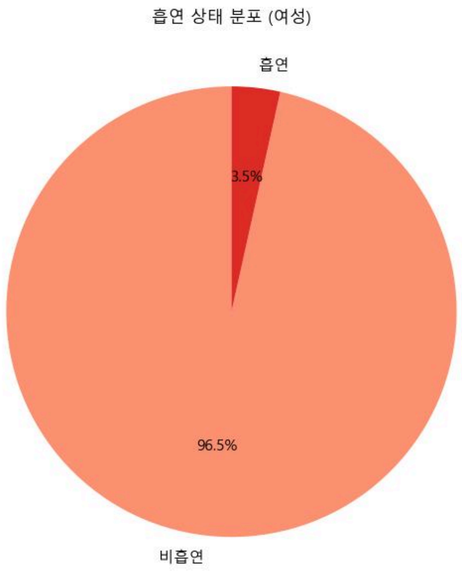
→ 트리글리세라이드, LDL콜레스테롤은 낮을수록 건강한 반면, HDL콜레스테롤은 높을수록 건강함을 의미
→ 다만 장년층의 콜레스테롤 감소는 노화로 인한 대사 기능 저하의 가능성도 존재

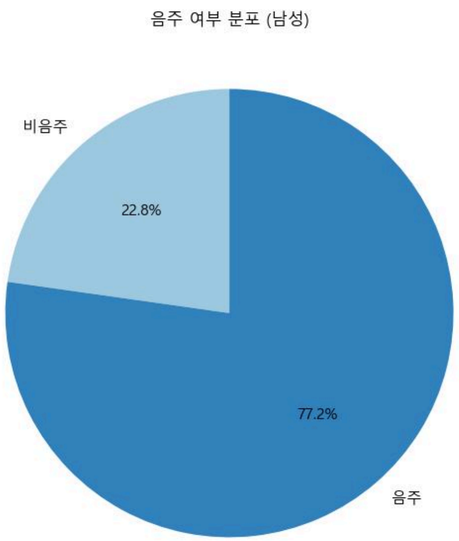
EDA

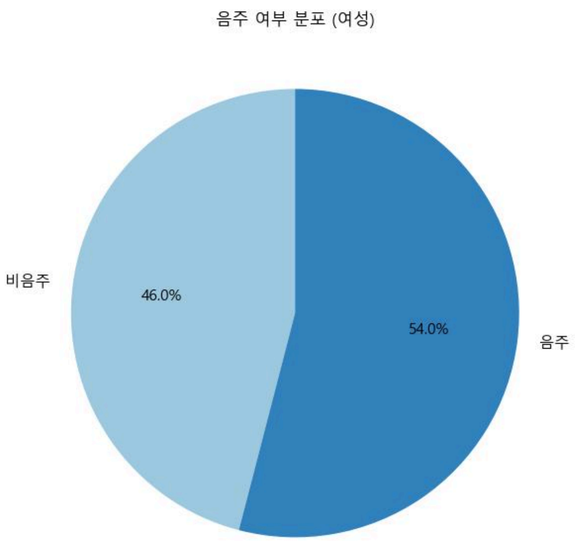
건강검진 데이터 분포 파악

성별 및 연령에 따른
흡연율 & 음주율







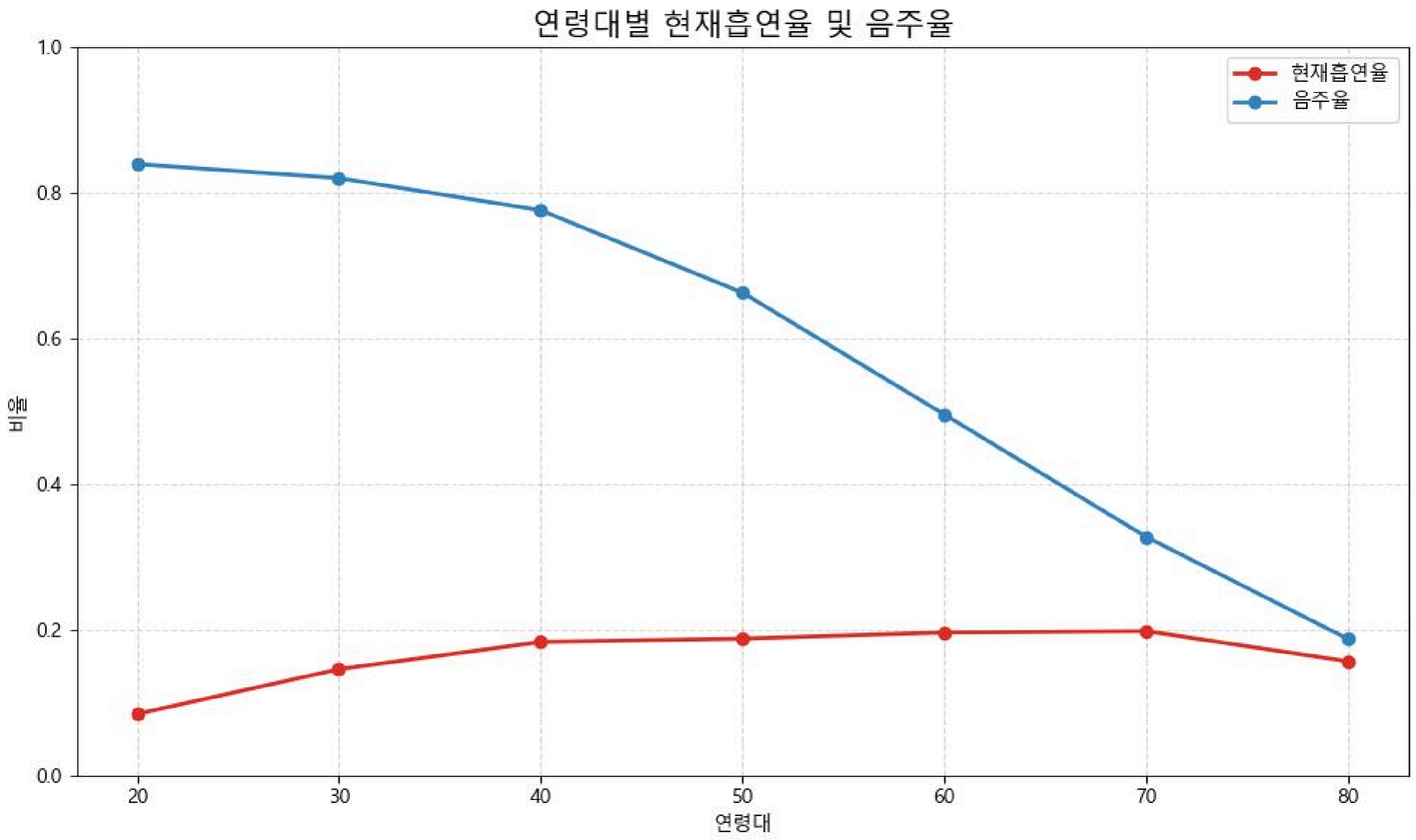


→ 흡연자가 남성은 46%에 육박하지만, 여성은 3.5%로 매우 적음
→ 음주는 남녀 모두 즐기는 모습이지만 남성의 음주율이 여성보다 약 23%p 더 높음

EDA

건강검진 데이터 분포 파악

연령에 따른
흡연율 & 음주율

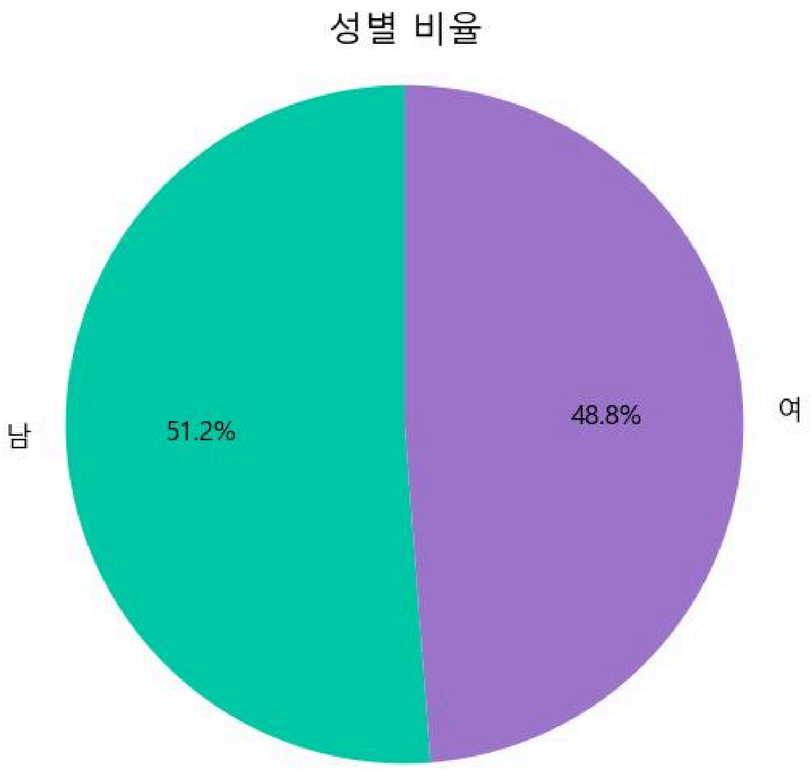


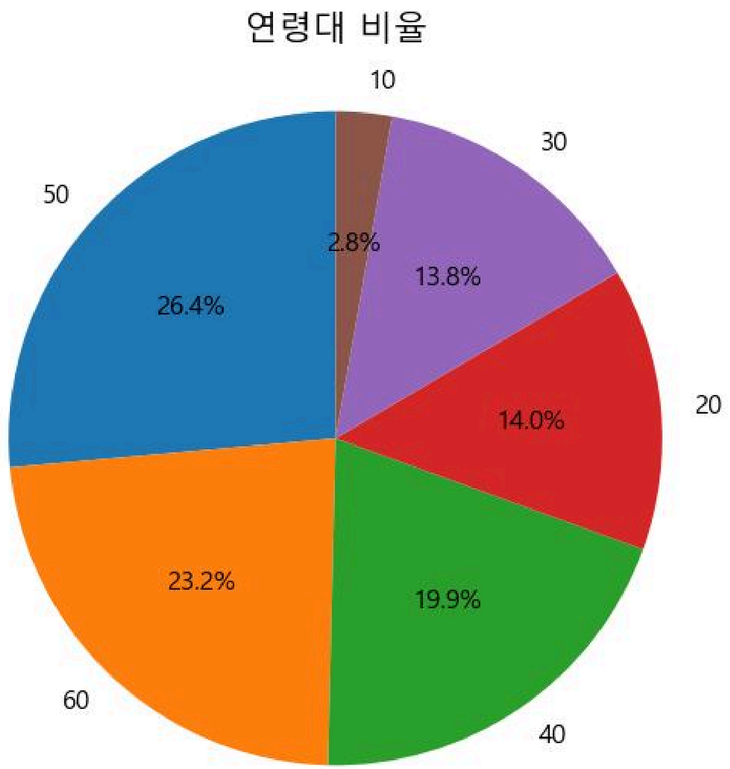
→ 연령대가 오를수록 음주율이 줄어들며 건강을 챙기는 모습이 관찰됨

EDA

iM뱅크 고객 카드 이용데이터 분포 파악

성별 및 연령대 비율



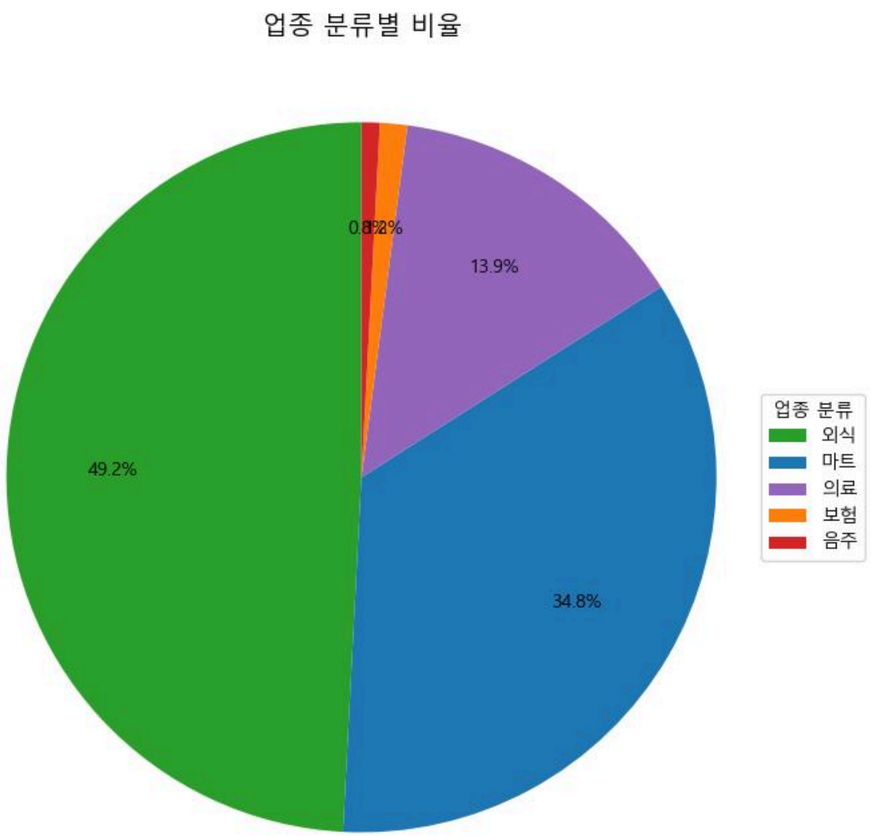


→ 남녀 고객이 비등하며, 40대 이상이 약 70%에 달함
→ 젊은 연령층을 계속해서 공략할 필요가 있음

EDA

iM뱅크 고객 카드 이용데이터 분포 파악

가맹점 업종 대분류별 비율

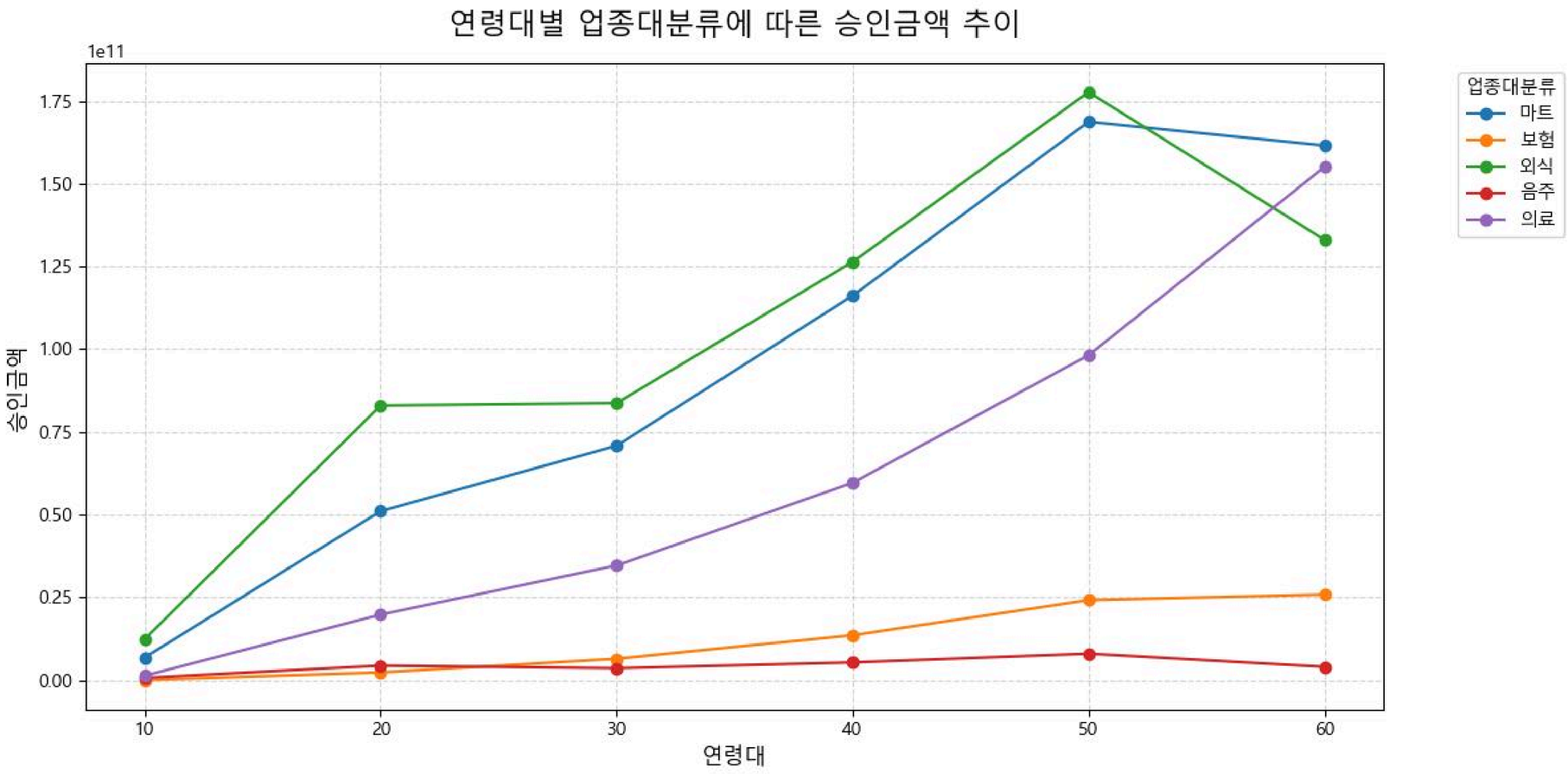


→ 외식, 마트 지출이 84%로 식비에 가장 많이 사용
→ 다음으로 의료 지출이 3순위로 큼

EDA

iM뱅크 고객 카드 이용데이터 분포 파악

연령대에 따른 승인금액



→ 3, 40대를 기점으로 마트, 보험, 외식의 승인금액이 큰 폭으로 상승함

프로젝트 수행 경과

프로젝트 수행 경과

1. 건강검진 데이터 군집분석 결과
   a. 이상지질검사 수검자 군집 분석
   b. 이상지질검사 미수검자 군집 분석
   c. 지질검사 비대상 2030 여성의 군집 분석
   d. 지질검사 비대상 20~23세 남성의 군집 분석
2. 당뇨병 예측 분류 모델 결과
3. iM 뱅크 고객의 이탈/비이탈 분석

1.2 데이터 전처리 [건강검진 데이터]

- 100만 인스턴스, 33개의 피처
- 이상지질혈증 검사 대상 : 남성 24세 이상, 여성 40세 이상(4년 마다)
  ex) 남성(24, 28, 32...) 여성(40, 44, 48...)
- 지질검사 항목 : 총콜레스테롤, HDL 콜레스테롤, LDL콜레스테롤, 트리글라세라이드

→ 데이터셋 2개로 분할하여 분석 진행

→ (이상지질검사 대상인 연령 , 이상지질검사 비대상 연령)

2) 성·연령별 검사 항목

구분		대상시기	비고
총콜레스테롤	이상지질혈증	남성 24세 이상, 여성 40세 이상 (4년 마다)	남성(24, 28, 32...) 여성(40, 44, 48...)
HDL 콜레스테롤			
중성지방			
LDL 콜레스테롤			

1.2 건강검진 데이터 : 결측치 처리

< 이상지질검사 비대상인 연령 데이터 >

- **전부 결측치인 컬럼 제거**

→ 총콜레스테롤, 트리글리세라이드, HDL콜레스테롤, LDL콜레스테롤, 결손치 유무, 치아마모증유무, 제3대 구치(사랑니)

컬럼명	결측치율	처리방법
허리둘레	0.041%	인스턴스 제거
시력(좌)	0.018%	인스턴스 제거
시력(우)	0.18%	인스턴스 제거
청력(좌)	0.014%	인스턴스 제거
청력(우)	0.014%	인스턴스 제거
감마지피티	0.588%	인스턴스 제거
혈청지피티(ALT)	0.589%	인스턴스 제거

컬럼명	결측치율	처리방법
수축기혈압	0.575%	인스턴스 제거
이완기혈압	0.578%	인스턴스 제거
식전혈당(공복혈당)	0.585%	인스턴스 제거
혈색소	0.586%	인스턴스 제거
요단백	1.123%	인스턴스 제거
혈청크레아티닌	0.59%	인스턴스 제거
혈청지오티(AST)	0.589%	인스턴스 제거

컬럼명	결측치율	처리방법
흡연상태	0.009%	인스턴스 제거
음주여부	0.006%	인스턴스 제거
치아우식증유무	65.315%	컬럼 제거
치석	65.315%	컬럼 제거

1.2 건강검진 데이터 : 결측치 처리

< 이상지질검사 대상인 연령 데이터 >

- **전부 결측치인 컬럼 제거 → 결손치 유무, 치아마모증유무, 제3대 구치(사랑니)**

컬럼명	결측치율	처리방법	컬럼명	결측치율	처리방법	컬럼명	결측치율	처리방법
허리둘레	0.041%	인스턴스 제거	수축기혈압	0.575%	인스턴스 제거	총콜레스테롤	66.139%	인스턴스 제거
시력(좌)	0.018%	인스턴스 제거	이완기혈압	0.578%	인스턴스 제거	트리글리세라이드	66.139%	인스턴스 제거
시력(우)	0.18%	인스턴스 제거	식전혈당(공복혈당)	0.585%	인스턴스 제거	HDL콜레스테롤	66.139%	인스턴스 제거
청력(좌)	0.014%	인스턴스 제거	혈색소	0.586%	인스턴스 제거	LDL콜레스테롤	66.139%	인스턴스 제거
청력(우)	0.014%	인스턴스 제거	요단백	1.123%	인스턴스 제거	치아우식증유무	65.315%	인스턴스 제거
감마지피티	0.588%	인스턴스 제거	혈청크레아티닌	0.59%	인스턴스 제거	치석	65.315%	인스턴스 제거
혈청지피티(ALT)	0.589%	인스턴스 제거	혈청지오티(AST)	0.589%	인스턴스 제거			

2. 파생변수 생성 - 일반

기존컬럼명	새로운변수명	예시	비고
키, 체중	BMI	체중/(키)^2	질병관리청 근거
혈청크레아티닌 연령대	사구체여과율	186*(혈청크레아티닌^2-1.154) *(연령^-0.203) *성별가중치(남 1 여 0.742)	
평균시력	시력(좌) 시력(우)	{시력(좌)+시력(우)}/2	

2. 파생변수 생성 - 건강검진 결과 기반의 질병판단 기준(1)

질병	판단 기준 변수	수치 기준/조건	판단 결과
신장질환	사구체여과율	>=90 60~89 30~59 15~29 <15	정상 신장기능 감소 시작 신장기능이 더욱 감소 생명만 유지할 수준 신장 이식 필요
간질환	AST, ALT, 감마지티피	AST>=40 또는 ALT>=40또는 GTP>63(남)/35(여)	간기능 이상 의심
고혈압	수축기혈압, 이완기혈압	수축기>=140 또는 이완기 >=90 수축기>=120 또는 이완기>=80 수축기<120 또는 이완기<80	고혈압 의심 고혈압 전단계 정상 혈압
빈혈	혈색소	16.5>=혈색소>=13(남) 15.5>=혈색소>=12(여)	정상
당뇨병	공복혈당	>=126 100~125 <100	당뇨병 의심 공복혈당장애 의심 정상혈당

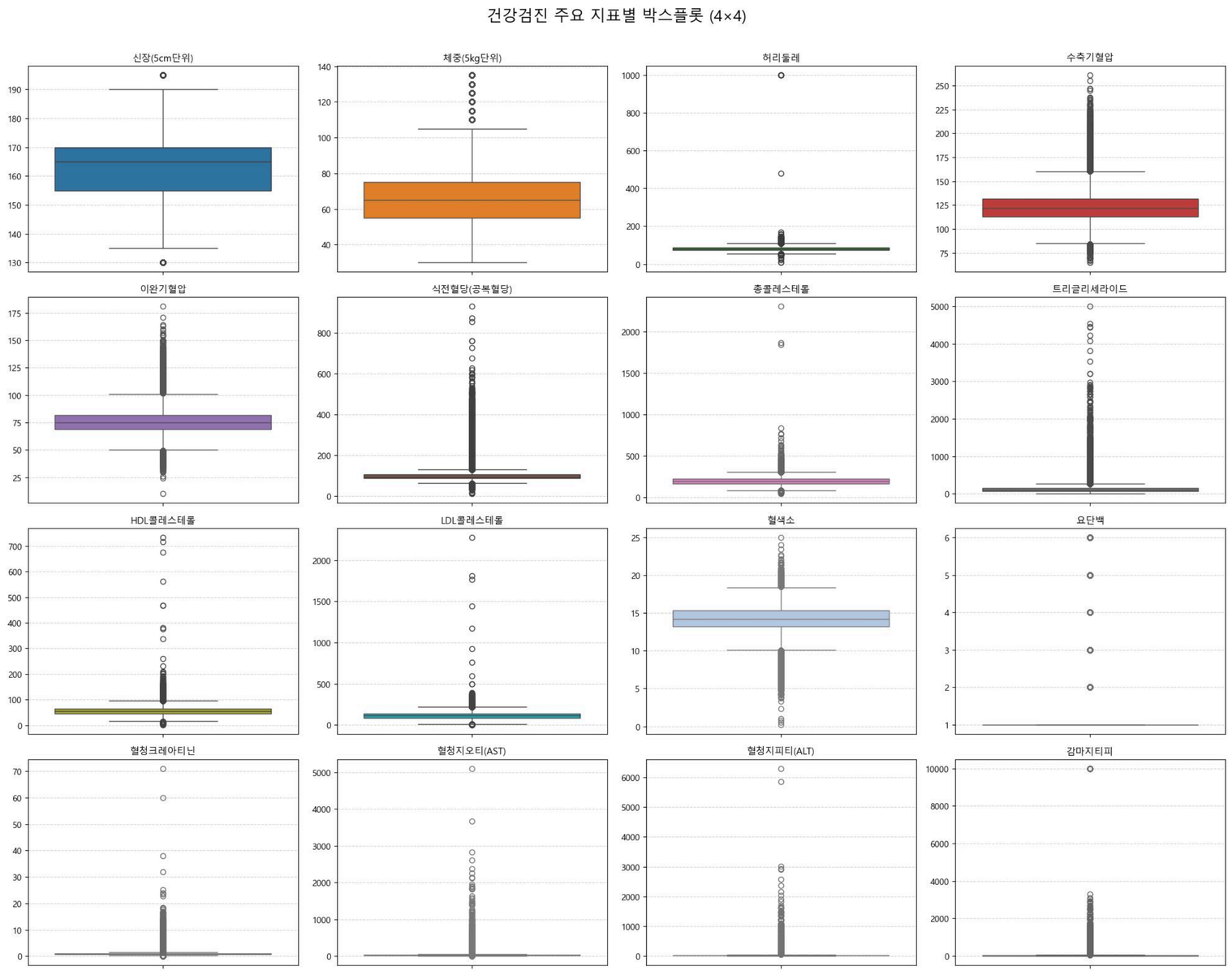
2. 파생변수 생성 - 건강검진 결과 기반의 질병판단 기준(2)

질병	판단 기준 변수	수치 기준/조건	판단 결과
이상지질혈증	총콜레스테롤	>250 또는 <150	이상지질혈증 의심
	HDL콜레스테롤	>65 또는 <30	이상지질혈증 의심
	LDL콜레스테롤	>=170	이상지질혈증 의심
	트리글리세라이드	>135 또는 <30	이상지질혈증 의심

2. 건강검진 데이터 : 이상치처리

- **오기입 항목** → 개별값 삭제
  ex) 청력(좌), 청력(우) 입력값 이외의 값인 3.00 삭제
- **분석 목적 외의 값**은 인스턴스 삭제
  ex) 시력(좌), 시력(우) 실명
- **분포상**
  - 허리둘레 200 이상 삭제
  - 허리둘레 480 제거
  - 공복혈당 700 초과 제거
  - 혈청크레아티닌 350 이상 제거
  - 혈청지오티 3000초과 제거
  - 혈청지피티 5000 초과 제거
  - 감마지티피 8000초과 제거

등의 **박스플롯에서 단일값이 떨어져있다면** 인스턴스 삭제



건강 군집 분석

1. 이상지질검사 수검자 건강군집 분석

- 건강 데이터는 대부분이 검사 수치로 연속형 수치라 KMeans로 모델 선정

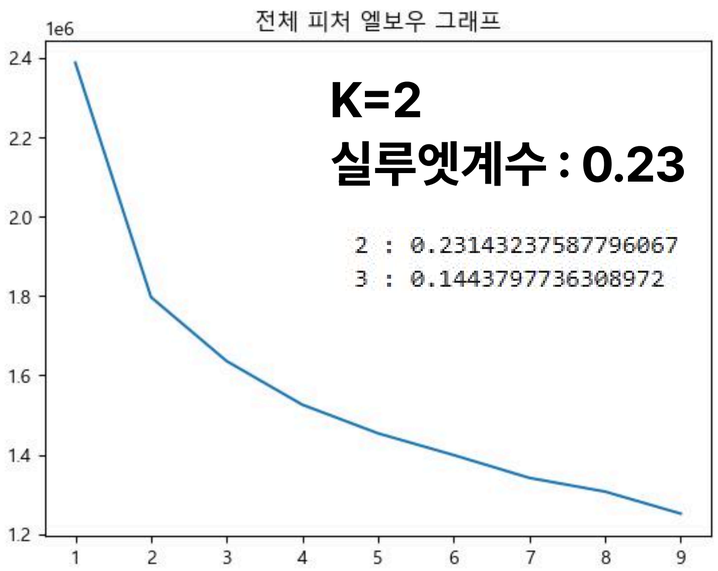
✅ 최종 선택 모델 – KMeans (수검자 군집 분석)

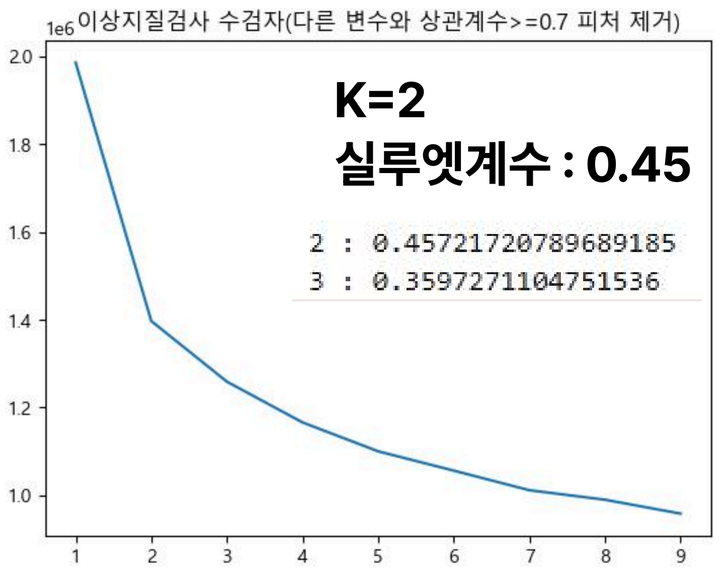
- 데이터 구성:
  - 약 10만건 인스턴스
  - 사용피처 : '성별', '연령대', '허리둘레', '수축기혈압', '이완기혈압', '식전혈당(공복혈당)', '트리글리세라이드', 'HDL콜레스테롤', '혈색소', '혈청크레아티닌', '혈청지오티(AST)', '혈청지피티(ALT)', '감마지티피', '흡연상태', '음주여부', 'BMI구분', '청력', '평균시력'

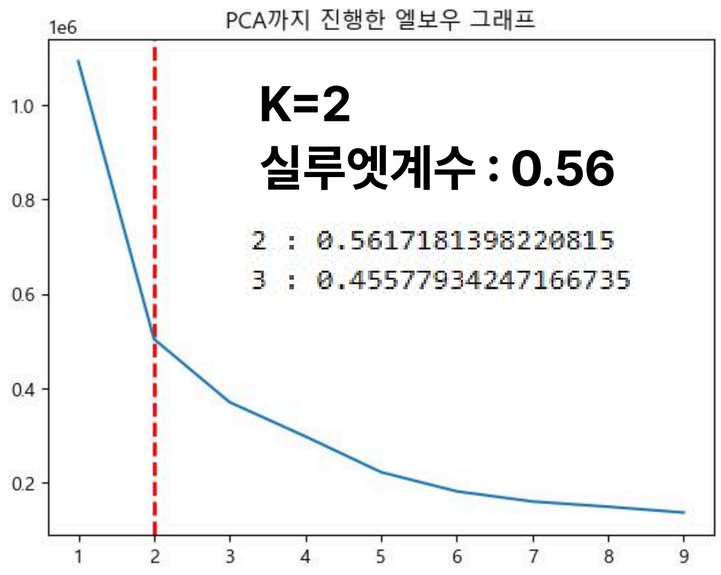
- 전처리 :
  - 범주형: OrdinalEncoder
  - 수치형: StandardScaler
  - 차원 축소: PCA(2D)

1. 이상지질검사 수검자 건강군집 분석

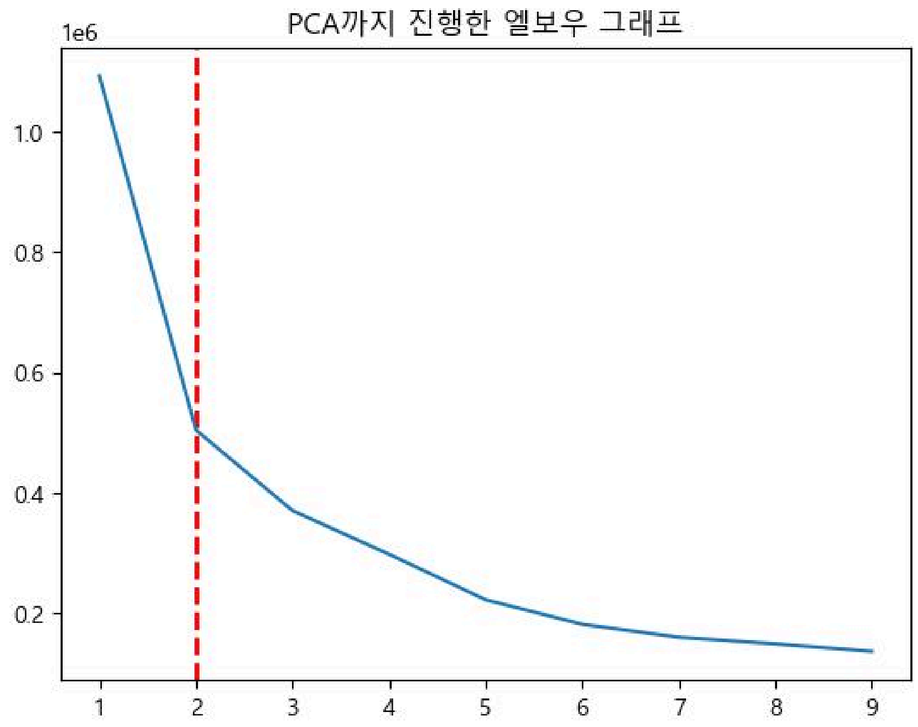
- **1차 : 전체 변수 다 넣었을 때**
- **사용 피처 :** '성별', '연령대', '허리둘레', '수축기혈압', '이완기혈압', '식전혈당(공복혈당)', '총콜레스테롤', '트리글리세라이드', 'HDL콜레스테롤', 'LDL콜레스테롤', '혈색소', '혈청크레아티닌', '혈청지오티(AST)', '혈청지피티(ALT)', '감마지티피', '흡연상태', '음주여부', '치아우식증유무', '치석', '청력', '평균시력', '사구체여과율'
- **2차 : 다른 변수와 상관계수 0.7 이상인 피처 제거**
  - 사용 피처 : → ex) 고혈압을 측정하는 지표 중 1개만 남김 (이완기혈압, 수축기혈압 중 수축기혈압만 채택)
    '혈청지피티(ALT)', '감마지티피', '흡연상태','음주여부', '치아우식증유무', '치석', '사구체여과율','청력', '평균시력', '식전혈당(공복혈당)', '총콜레스테롤', '트리글리세라이드'
    , 'HDL콜레스테롤' 성별', '연령대', '수축기혈압'
- **3차 : PCA 차원 축소 (2차원)**







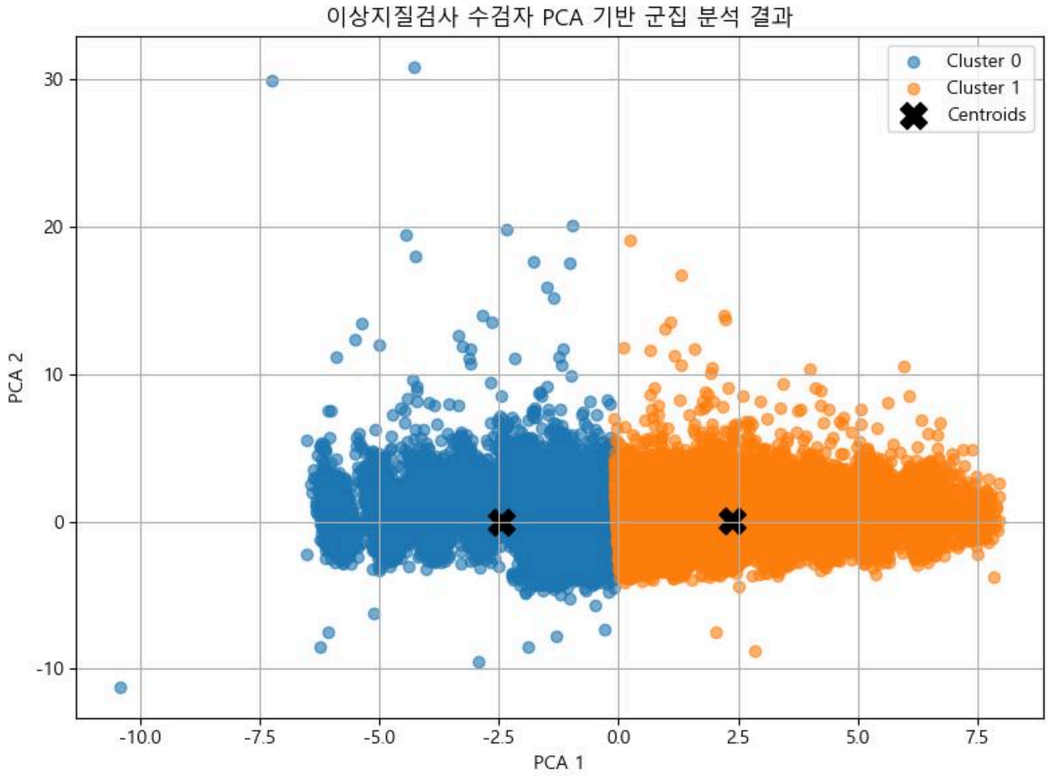
1. 이상지질검사 수검자 건강군집 분석



K=2

실루엣계수 : 0.56

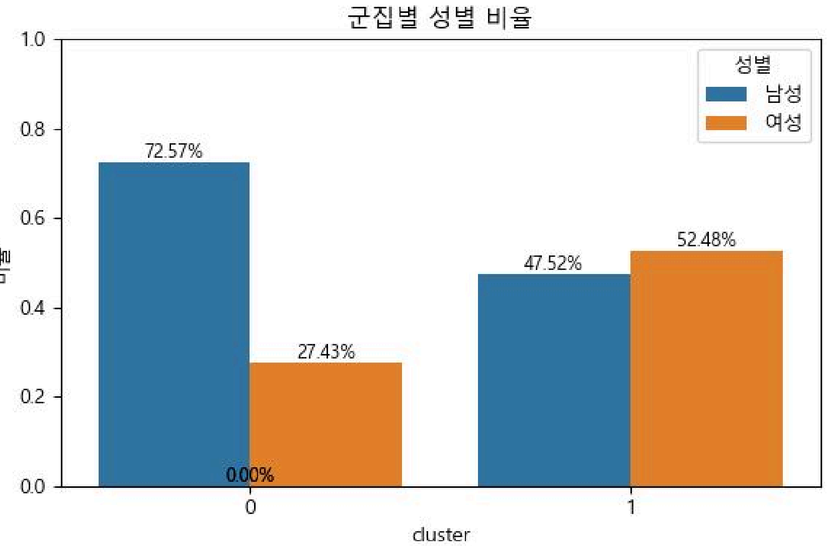
```
2 : 0.5617181398220815
3 : 0.45577934247166735
```

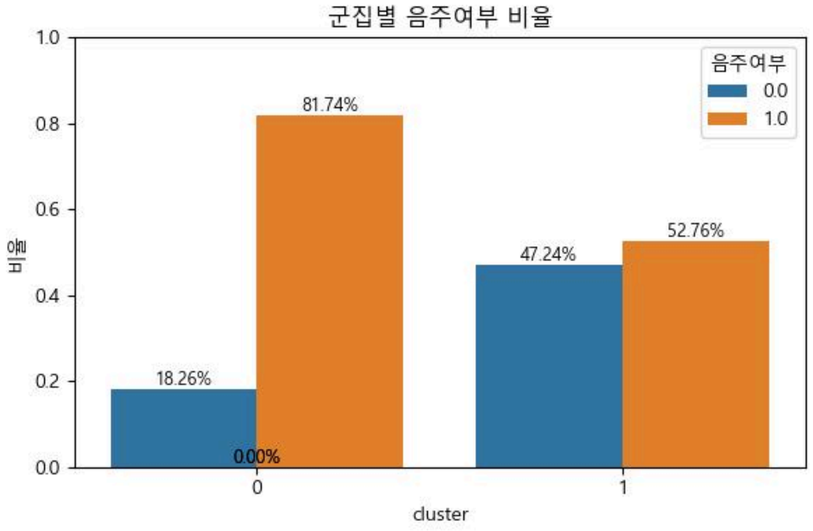


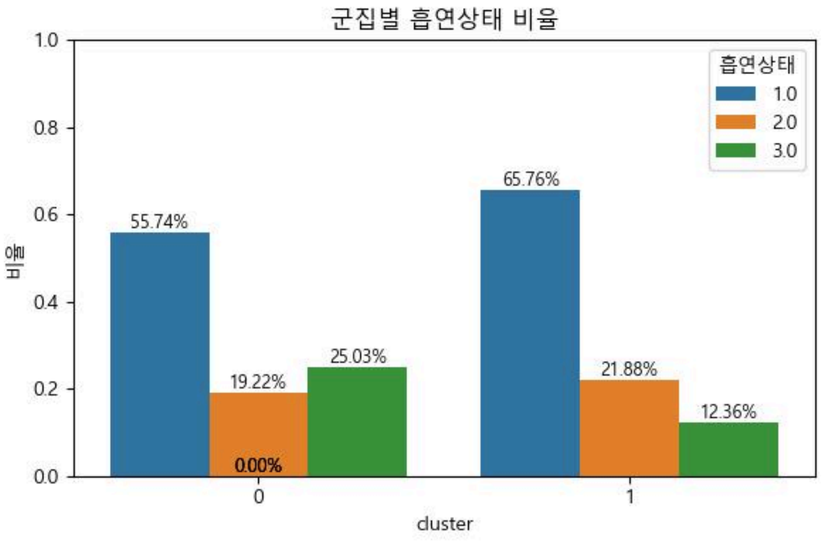
1. 이상지질검사 수검자 건강군집 특성 - 생활측면

- 비건강(군집0)
  - 남성7:여성3
  - 음주자가 80%에 달함
  - 비흡연자가 55.7%

- 건강(군집1)
  - 남5:여5
  - 음주자가 53%로 비건강집단에 비해 매우 낮음
  - 비흡연자가 65.7%로 비건강0에 비해 흡연율이 낮음

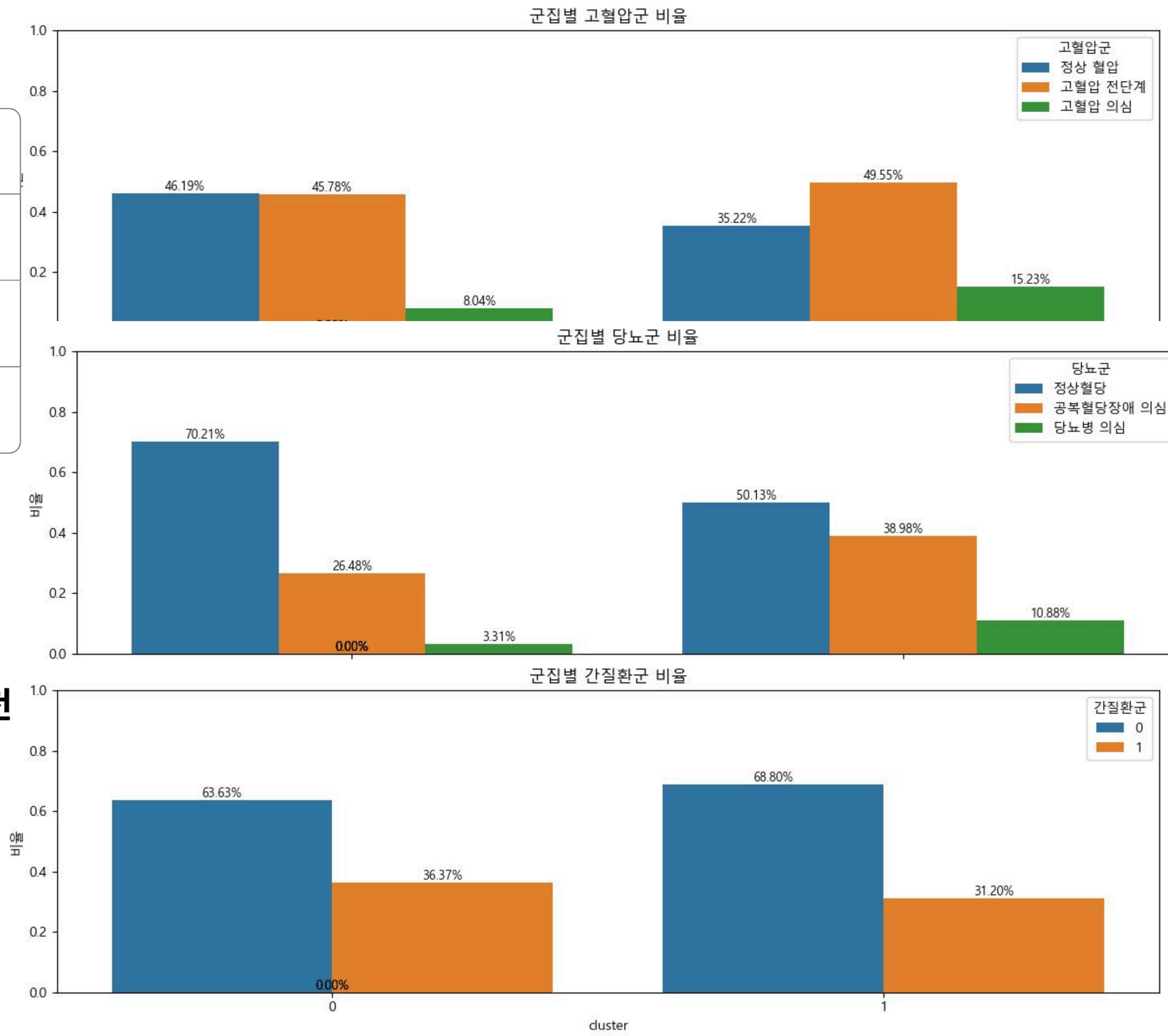






- **군집별 질병 발병 비율**

질병	군집0(비건강)	군집1(건강)
고혈압	53%	64%
당뇨	30%	50%
간질환	36%	31%



군집별 보험 전략 제안

건강 양호한 남성군 → 고혈압&당뇨 보험 우선 추천

고위험 중년 여성군 → 고혈압 보험 우선 추천

2. 이상지질검사 미수검자의 건강군집 분석

✅ 최종 선택 모델 – KMeans

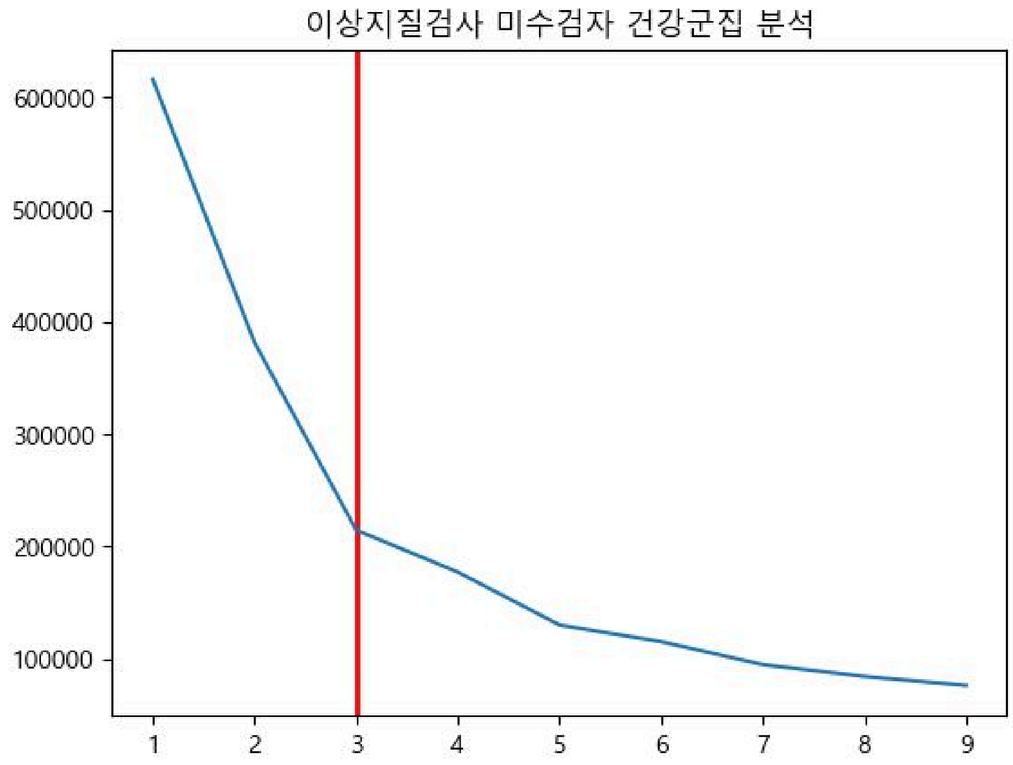
- 데이터 구성
  - 약 52만 건 인스턴스

사용피처:성별, 연령대, 허리둘레, 수축기혈압, 이완기혈압, 공복혈당, (AST), 혈청지피티(ALT), 감마지티피, 흡연상태, 혈색소, 요단백, 혈청크레아티닌, 혈청지오티음주여부, BMI, 사구체여과율

- 전처리 :
  - 범주형: OrdinalEncoder
  - 수치형: StandardScaler
  - 차원 축소: PCA(2D)

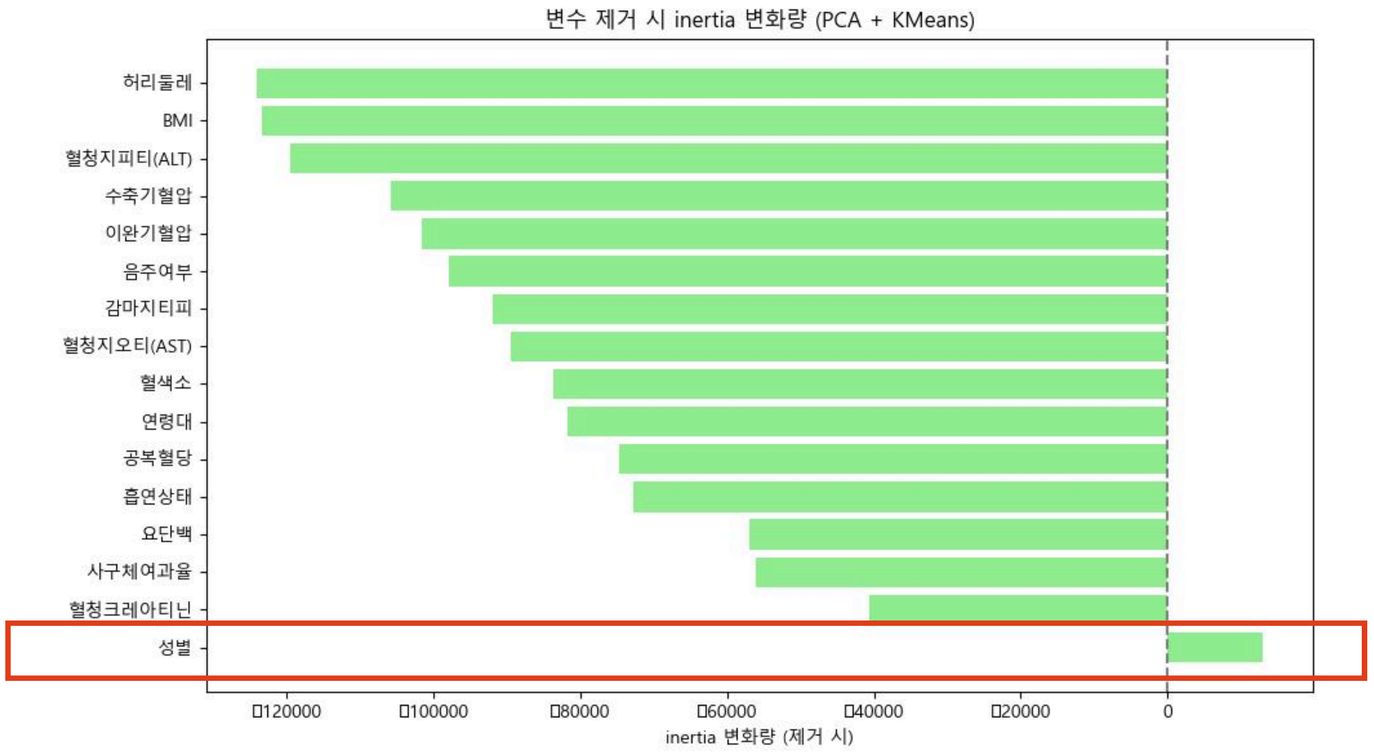
2. 이상지질검사 미수검자의 건강군집 분석

- 군집수 변화에 따른 Elbow 그래프 시각화




```
2 : 0.42578503030423415
3 : 0.5132965802796764
4 : 0.4917410993834699
5 : 0.4871047024973102
```

K=3에서 실루엣 계수 0.51로 최대

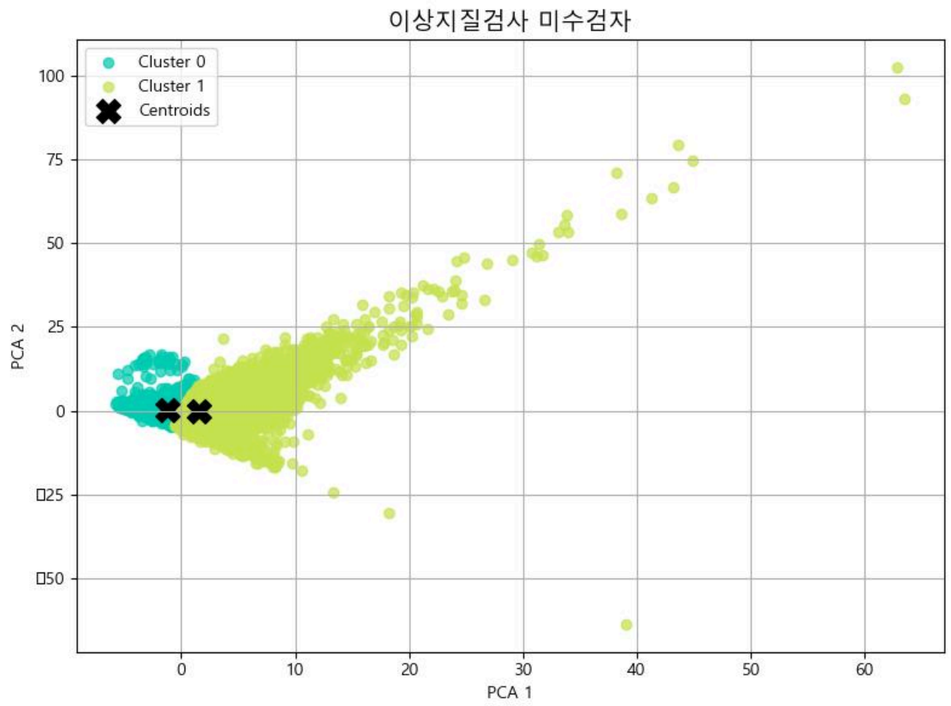


→ 가장 설명력 적은 변수가 성별인데 성별은 여러 연령대가 섞인 집단에서 해석에 중요한 피처이므로 기존 피처 유지

→ K=3, 기존에 사용한 변수 유지

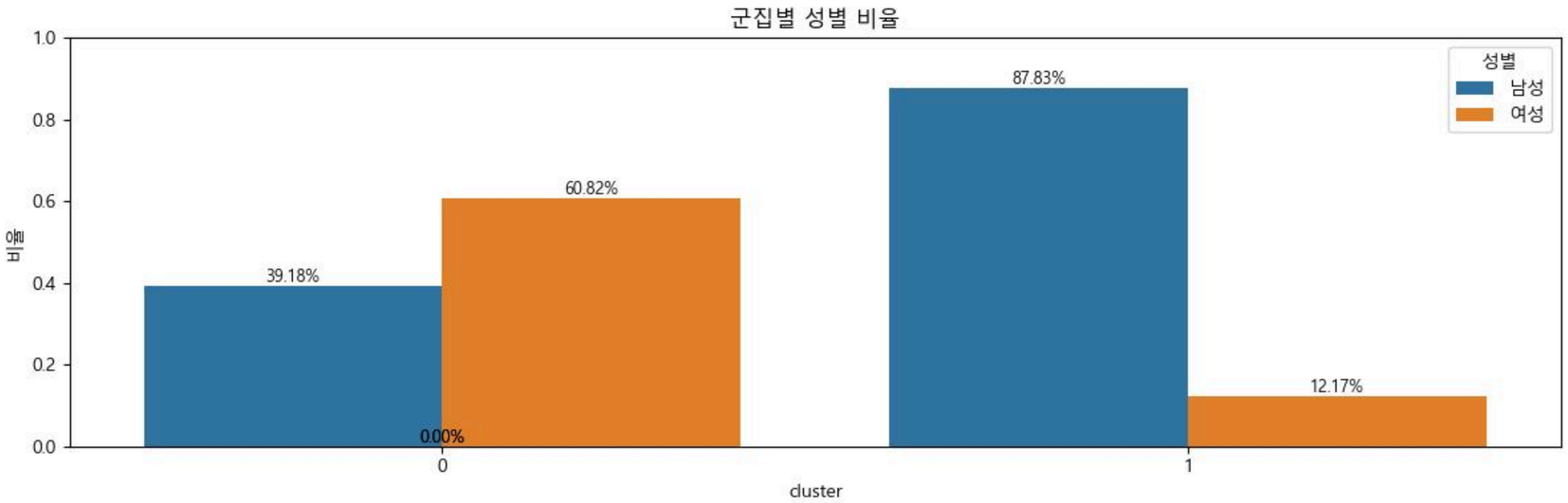
• 이상지질검사 미수검자 군집분석 결과

구분	Cluster 0	Cluster 1
주요 연령대	50~60대 여성	30~50대 남성
성별 비중	여성 40%	남성 88%
혈압	정상 다수	고혈압 의심 다수
당뇨	정상 많음	당뇨·공복혈당장애 다수
간질환	일부 존재 (25%)	100% 간질환
신장질환	경미한 기능 저하 일부 존재	경미한 기능 저하 + 요단백 다수
흡연	비흡연자 위주	흡연자 비율 매우 높음
음주	음주자 65%, 비교적 절제됨	음주자 89%, 과도한 음주 경향
빈혈	낮은 편 (13%)	높은 편 (20%)



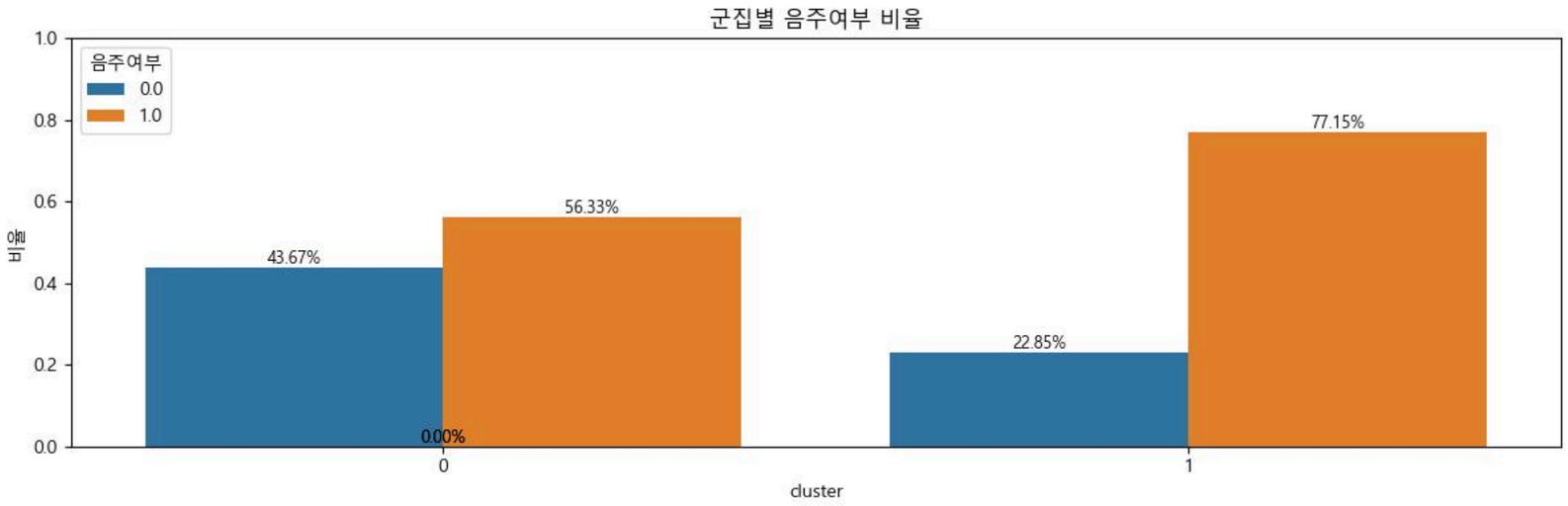
군집명명 : **절제미인** **건간한 연기중독러**

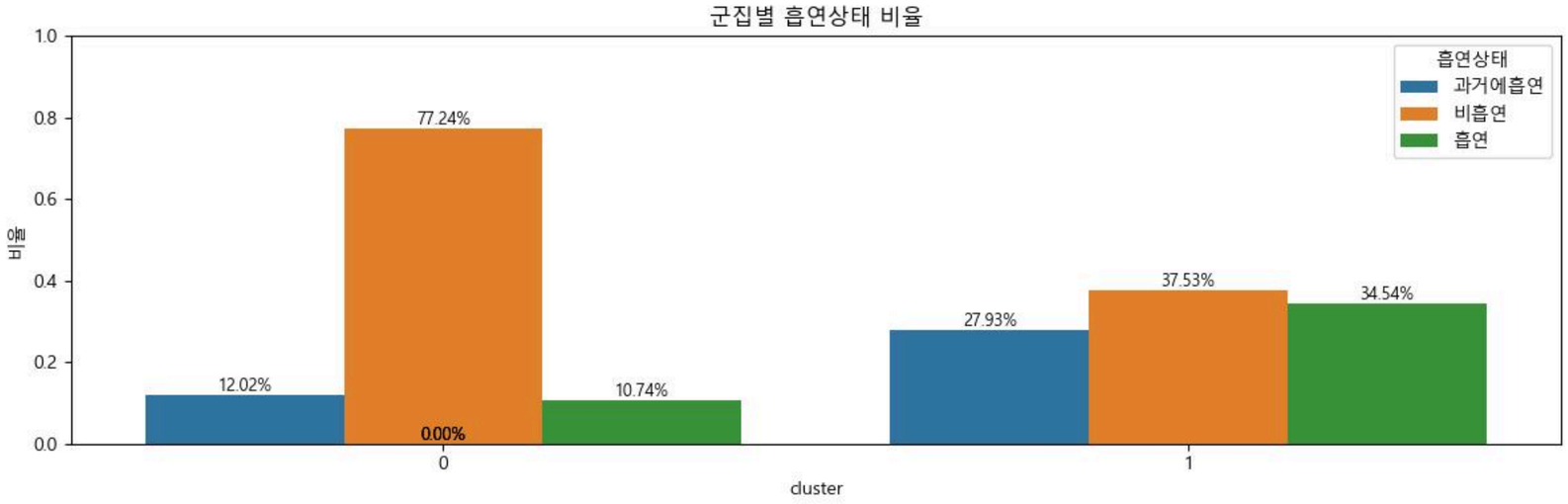
• 이상지질검사 미수검자 군집분석 결과



- 비건강(군집0)
  - 남성4:여성6
  - 전체적인 BMI분포가 낮은 편
- 건강(군집1)
  - 남 8.7:여1.2
  - 전체적인 BMI분포가 높은 편

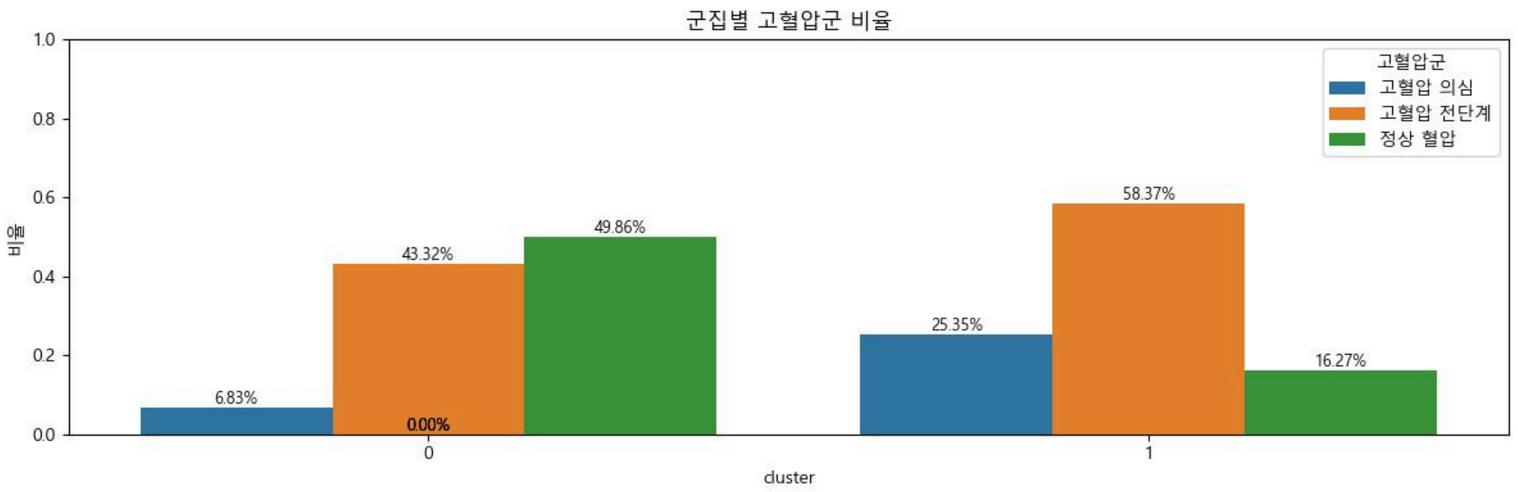
- **이상지질검사 미수검자 군집분석 결과**

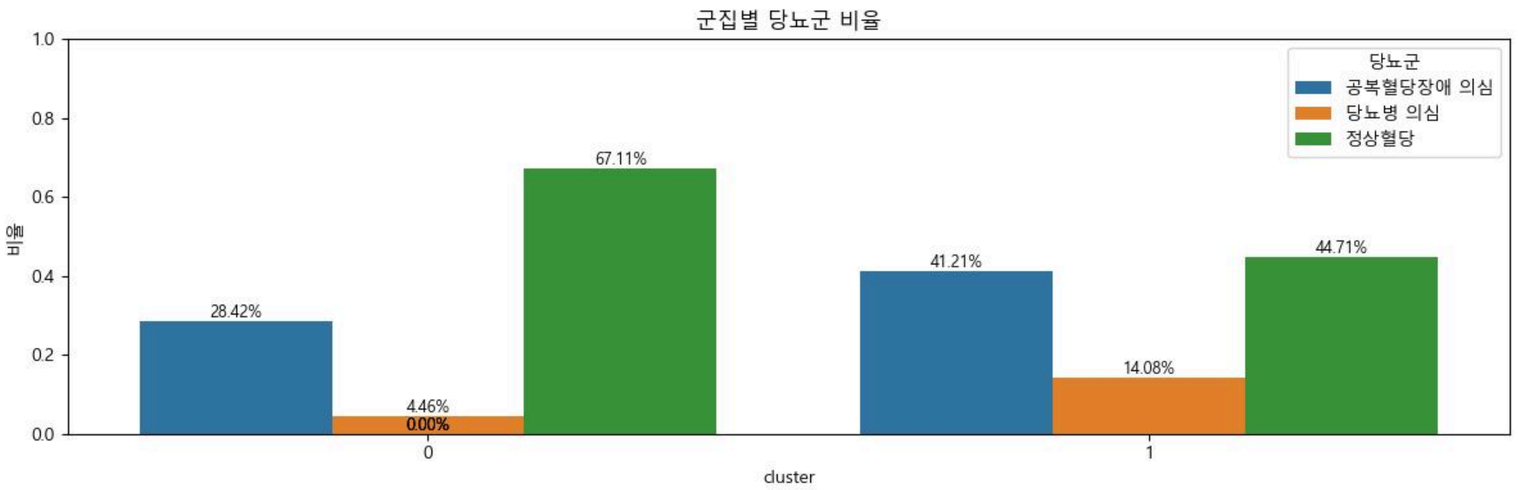




- **군집0, 1 전부 어느 정도 음주를 즐기는 편이나, 군집0은 비흡연자율이 77%로 군집1 비흡연자율 37%에 비해 매우 높음**

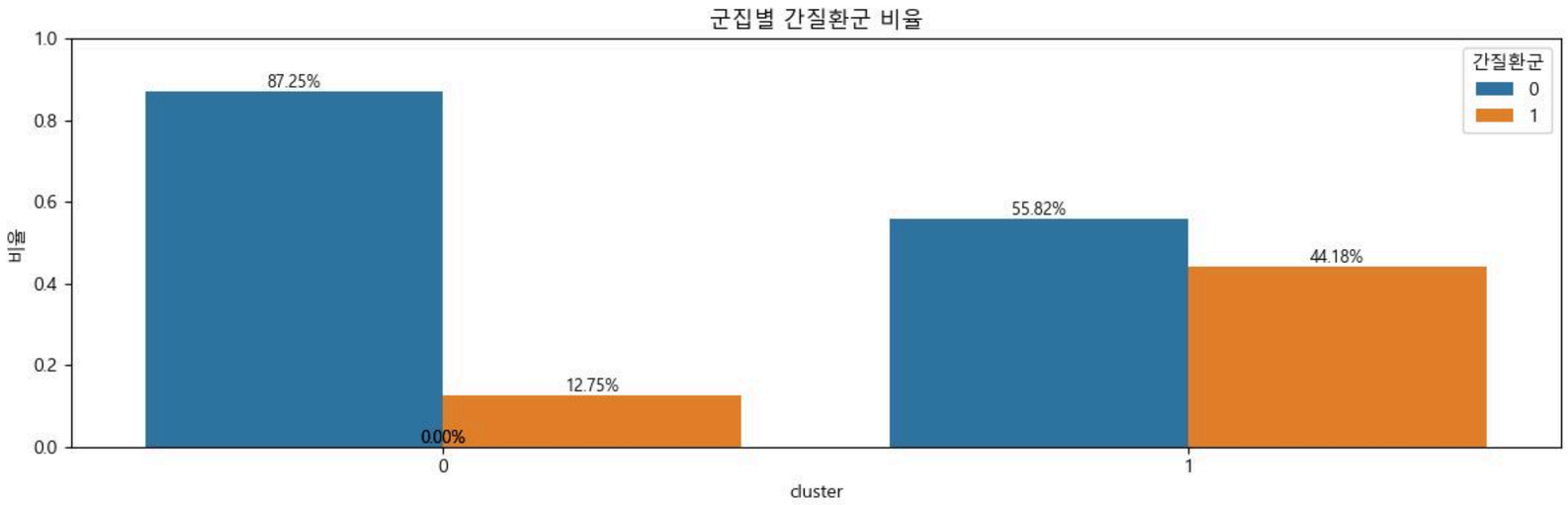
- **이상지질검사 미수검자 군집분석 결과**





- **군집0의 49%가 정상 혈압으로 군집 1에 비해 혈압 위험이 적음**
- **군집 0의 당뇨군은 67%가 정상 혈당으로 군집 1에 비해 당뇨군이 23% 적음**

- **이상지질검사 미수검자 군집분석 결과**
  - **군집0의 12.7%가 간질환이 의심되며 군집 1은 44%가 간질환이 의심됨**

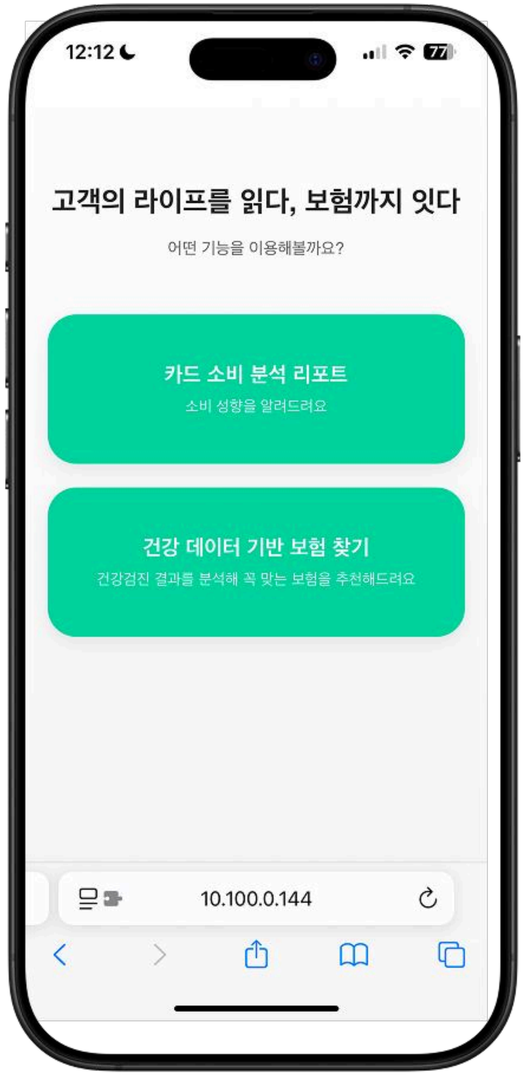


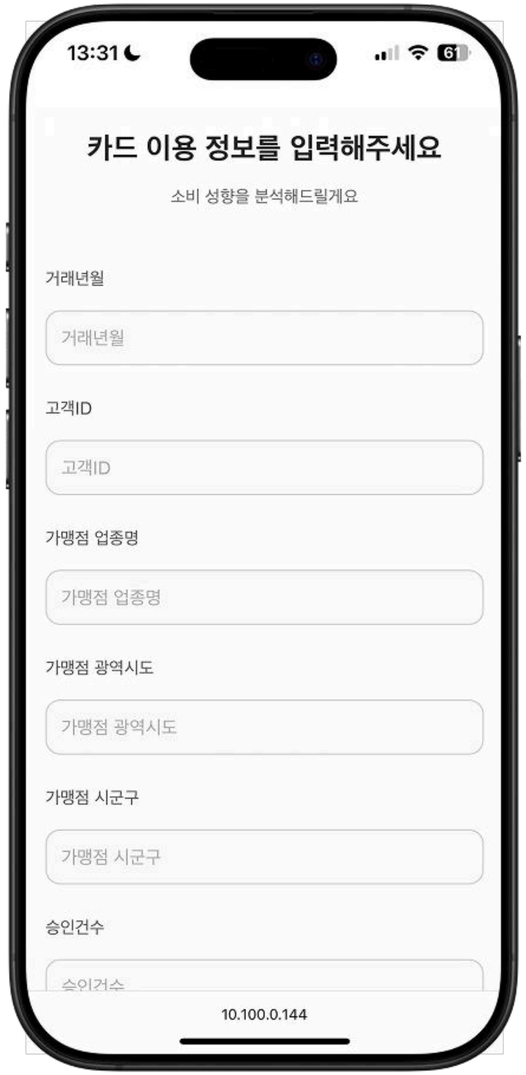
- **군집별 보험 전략 제안**
  - 군집0(고혈압군)→ 고혈압 관련 보험 우선 추천
  - 군집1(당뇨엔 간질환군) → 당뇨 및 간질환 결합 보험 우선 추천

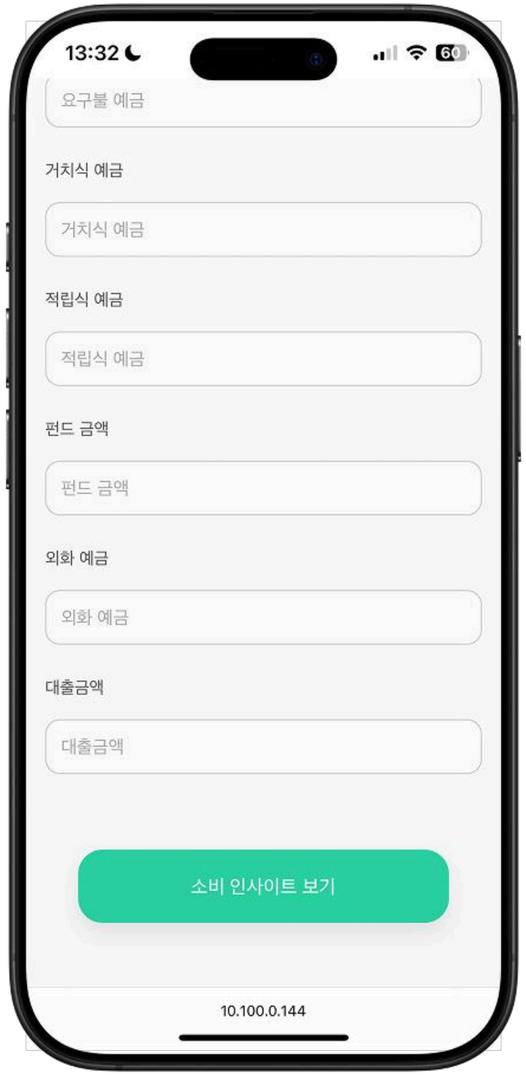
군집별 질병 발병 비율

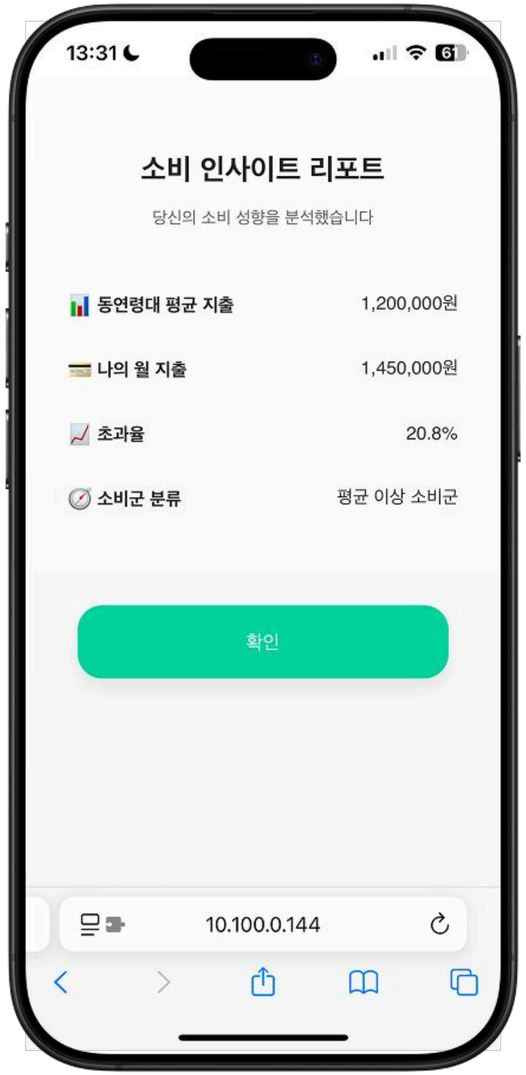
질병	군집0(비건강)	군집1(건강)
고혈압	43%	22.8%
당뇨	32%	55%
간질환	12.7%	44.1%

카드 소비 분석 리포트

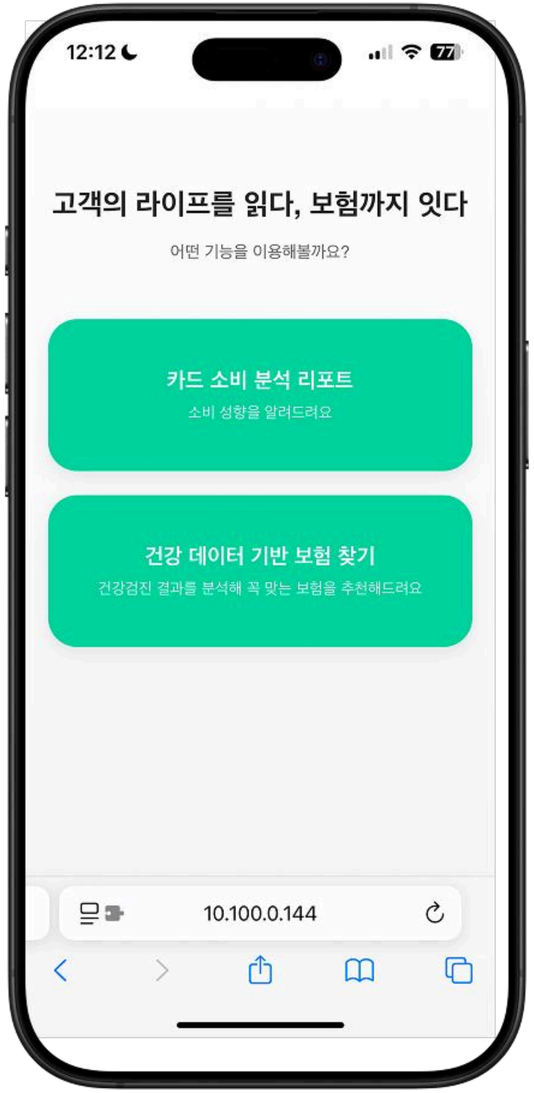

12:12
고객의 라이프를 읽다, 보험까지 잇다
어떤 기능을 이용해볼까요?
카드 소비 분석 리포트
소비 성향을 알려드려요
건강 데이터 기반 보험 찾기
건강검진 결과를 분석해 꼭 맞는 보험을 추천해드려요
10.100.0.144

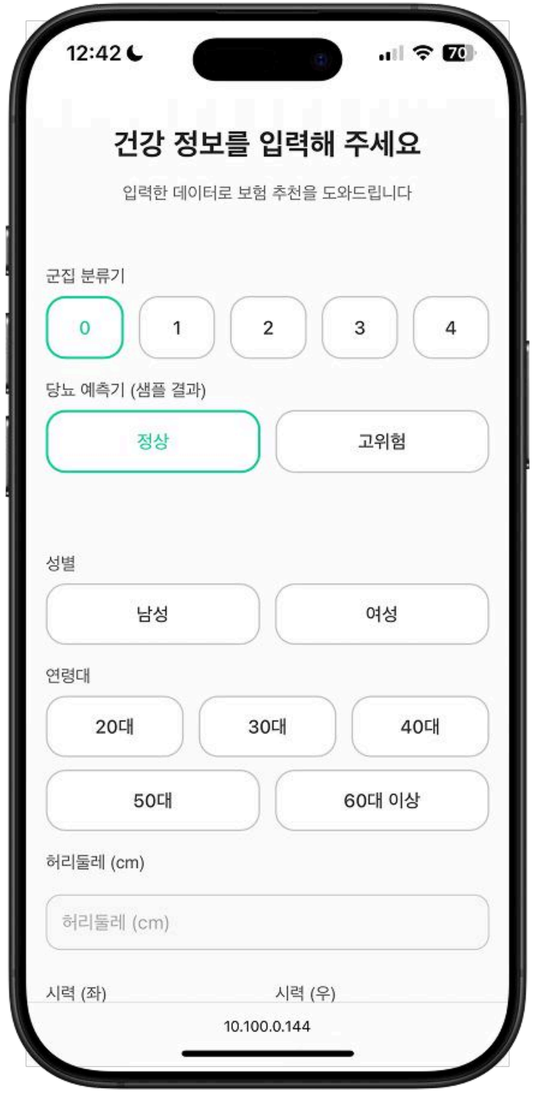

13:31
카드 이용 정보를 입력해주세요
소비 성향을 분석해드릴게요
거래년월
거래년월
고객ID
고객ID
가맹점 업종명
가맹점 업종명
가맹점 광역시도
가맹점 광역시도
가맹점 시군구
가맹점 시군구
승인건수
10.100.0.144


13:32
요구불 예금
거치식 예금
거치식 예금
적립식 예금
적립식 예금
펀드 금액
펀드 금액
외화 예금
외화 예금
대출금액
대출금액
소비 인사이트 보기
10.100.0.144

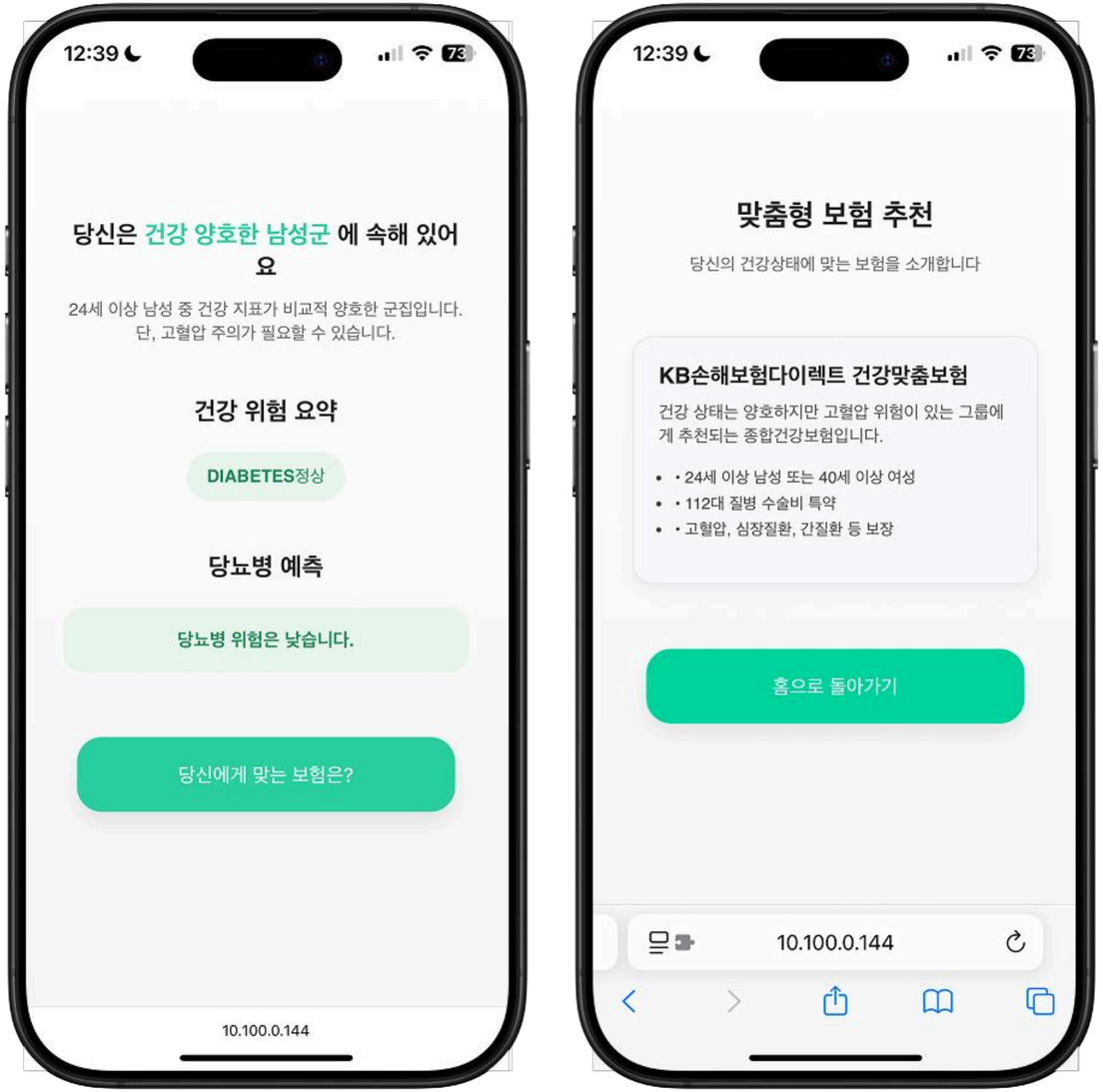

13:31
소비 인사이트 리포트
당신의 소비 성향을 분석했습니다
동연령대 평균 지출 1,200,000원
나의 월 지출 1,450,000원
초과율 20.8%
소비군 분류 평균 이상 소비군
확인
10.100.0.144

건강 데이터 기반 보험 찾기

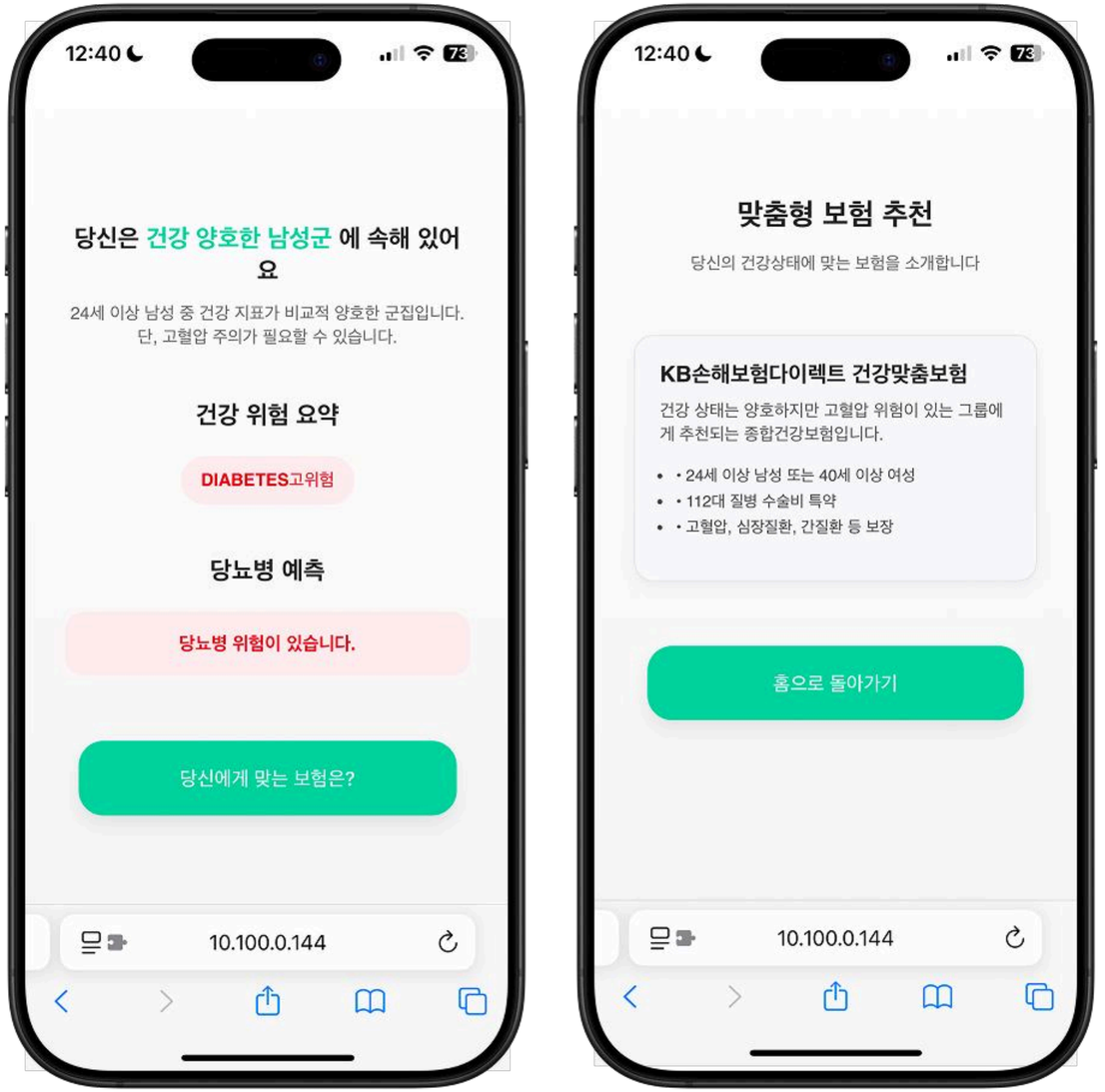




건강검진 분류기 0번

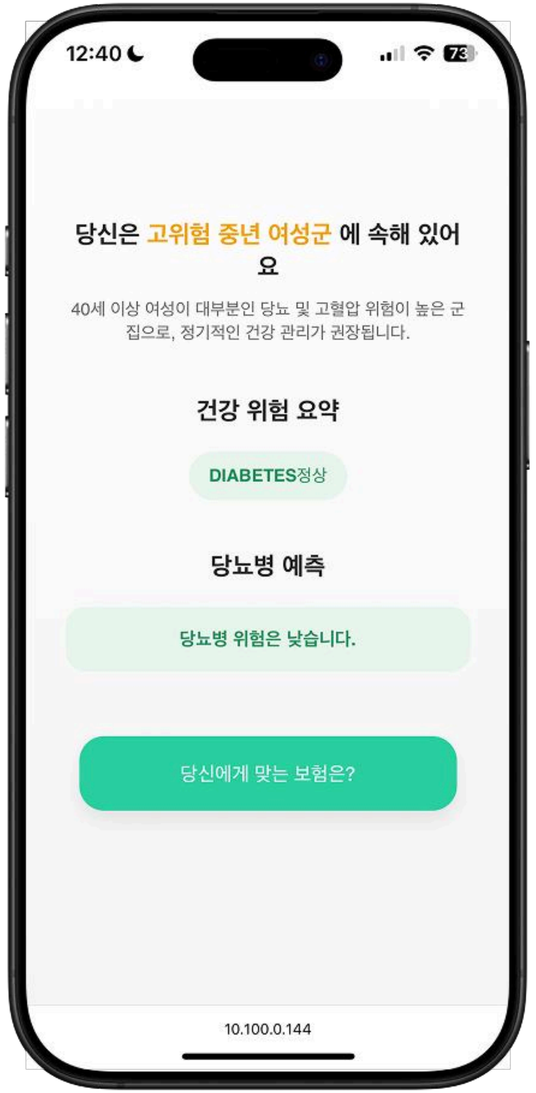


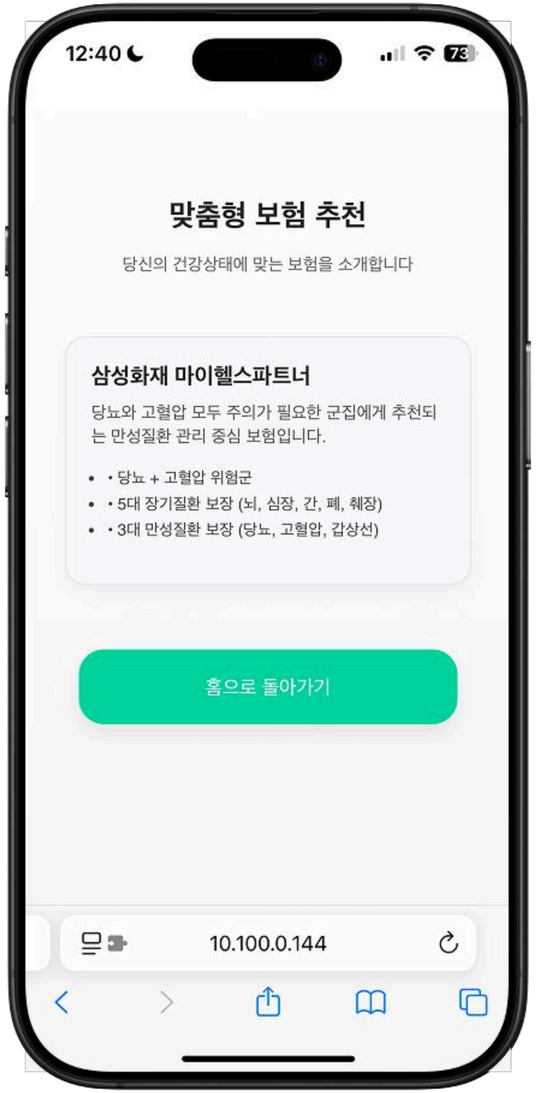
클러스터0-0 당뇨 0



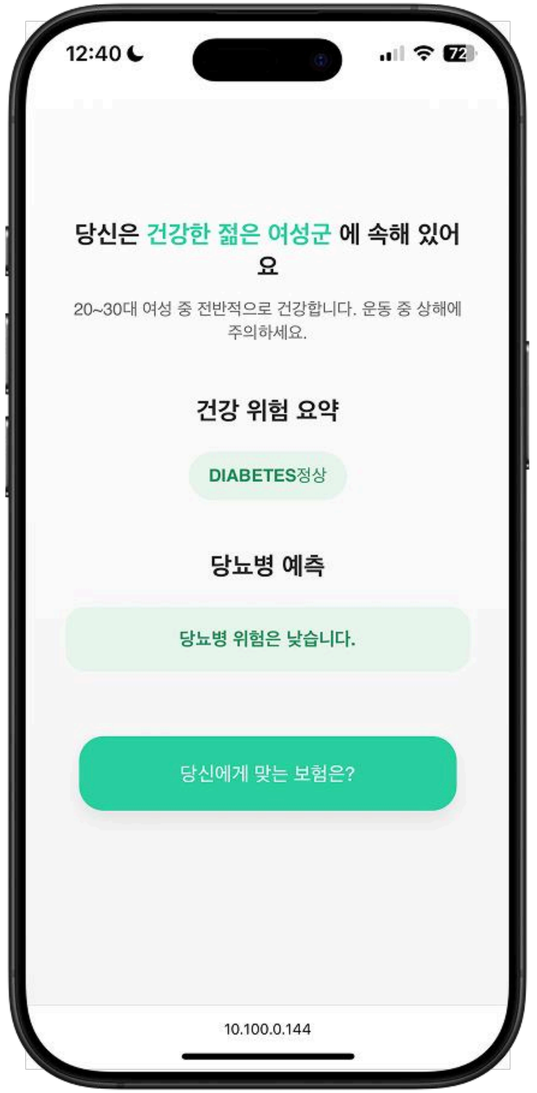
클러스터0-0 당뇨 1

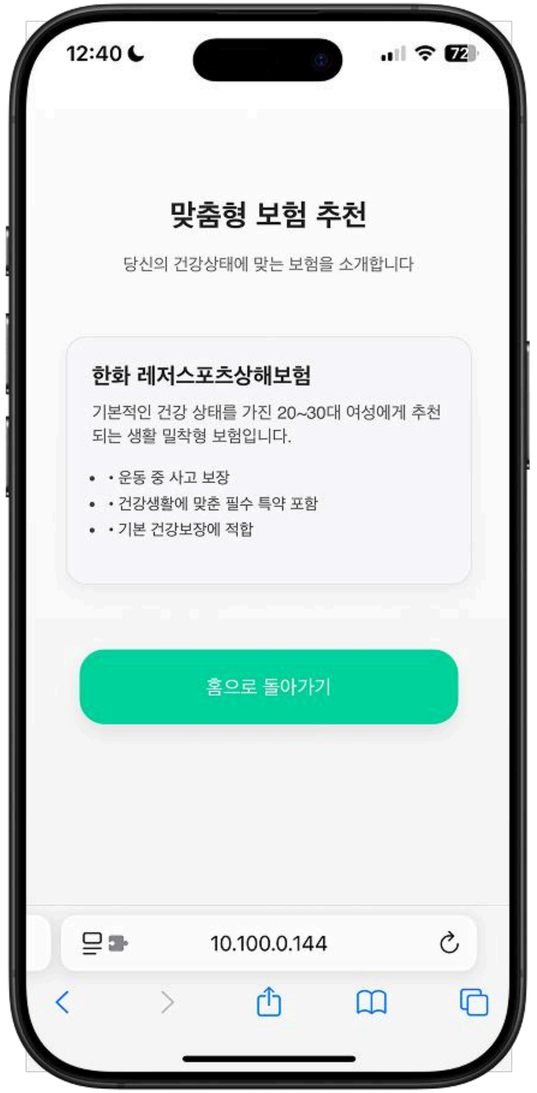
건강검진 분류기 1번





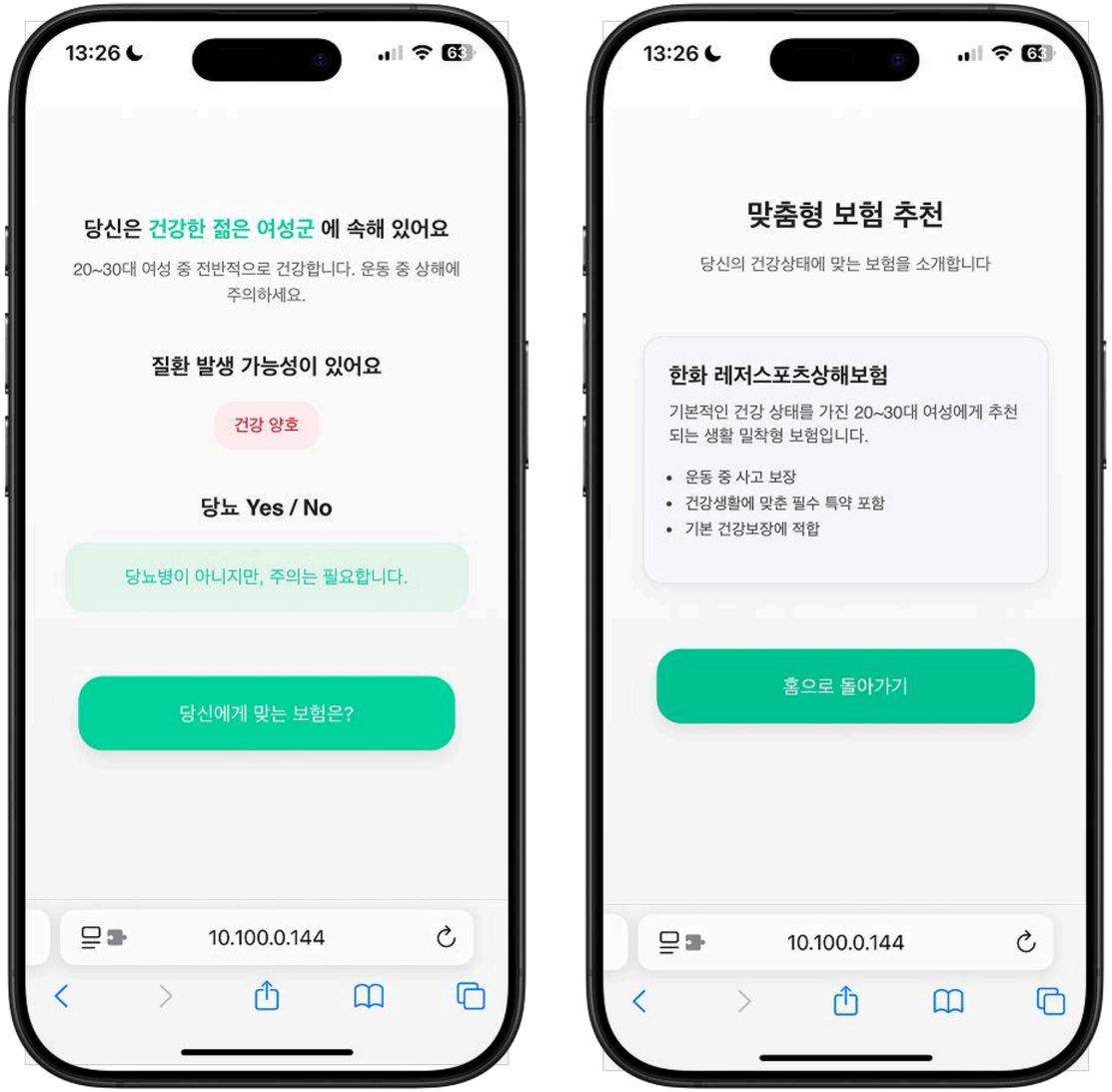
클러스터0-1 당뇨 0



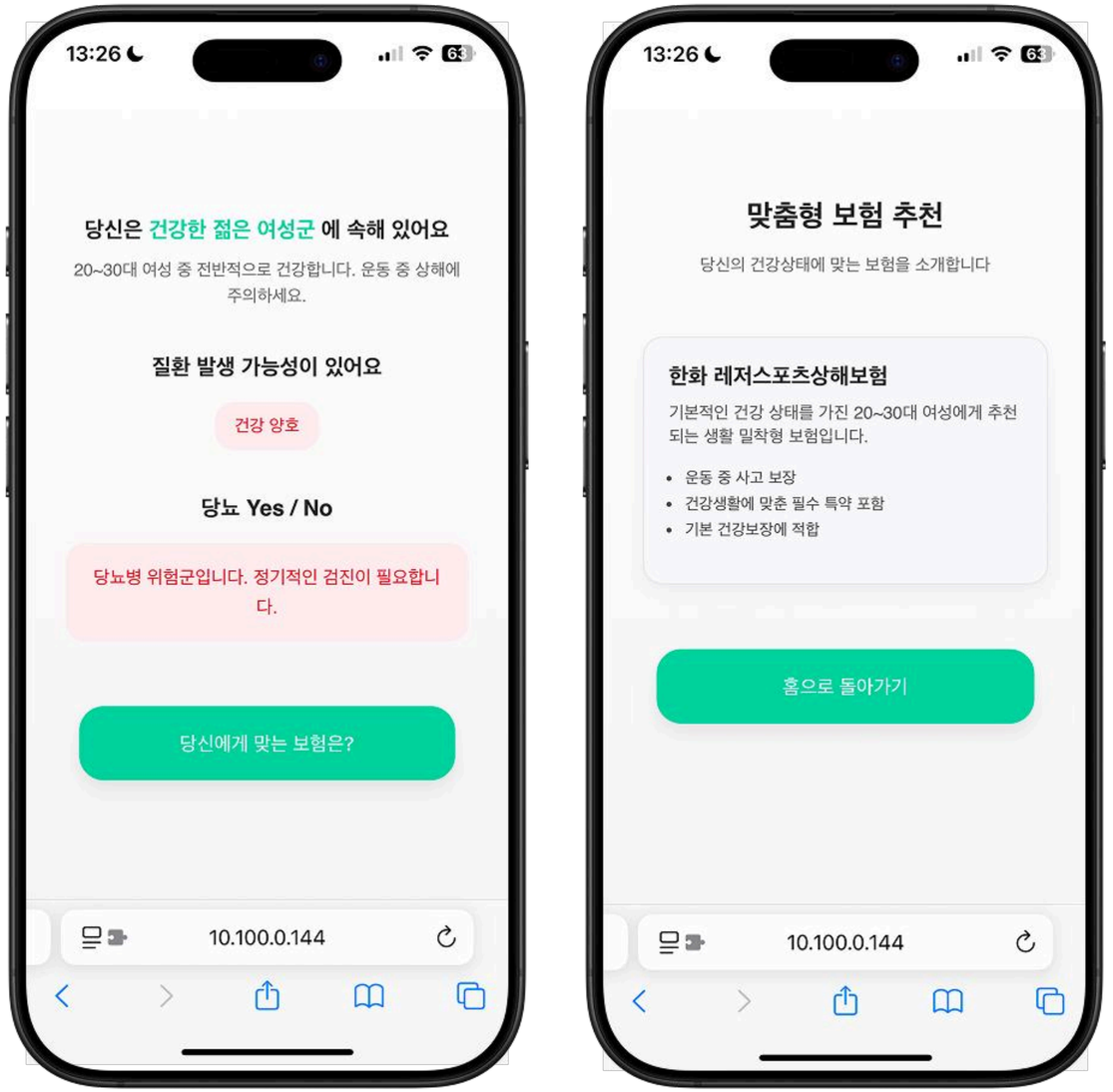


클러스터0-1 당뇨 1

건강검진 분류기 2번

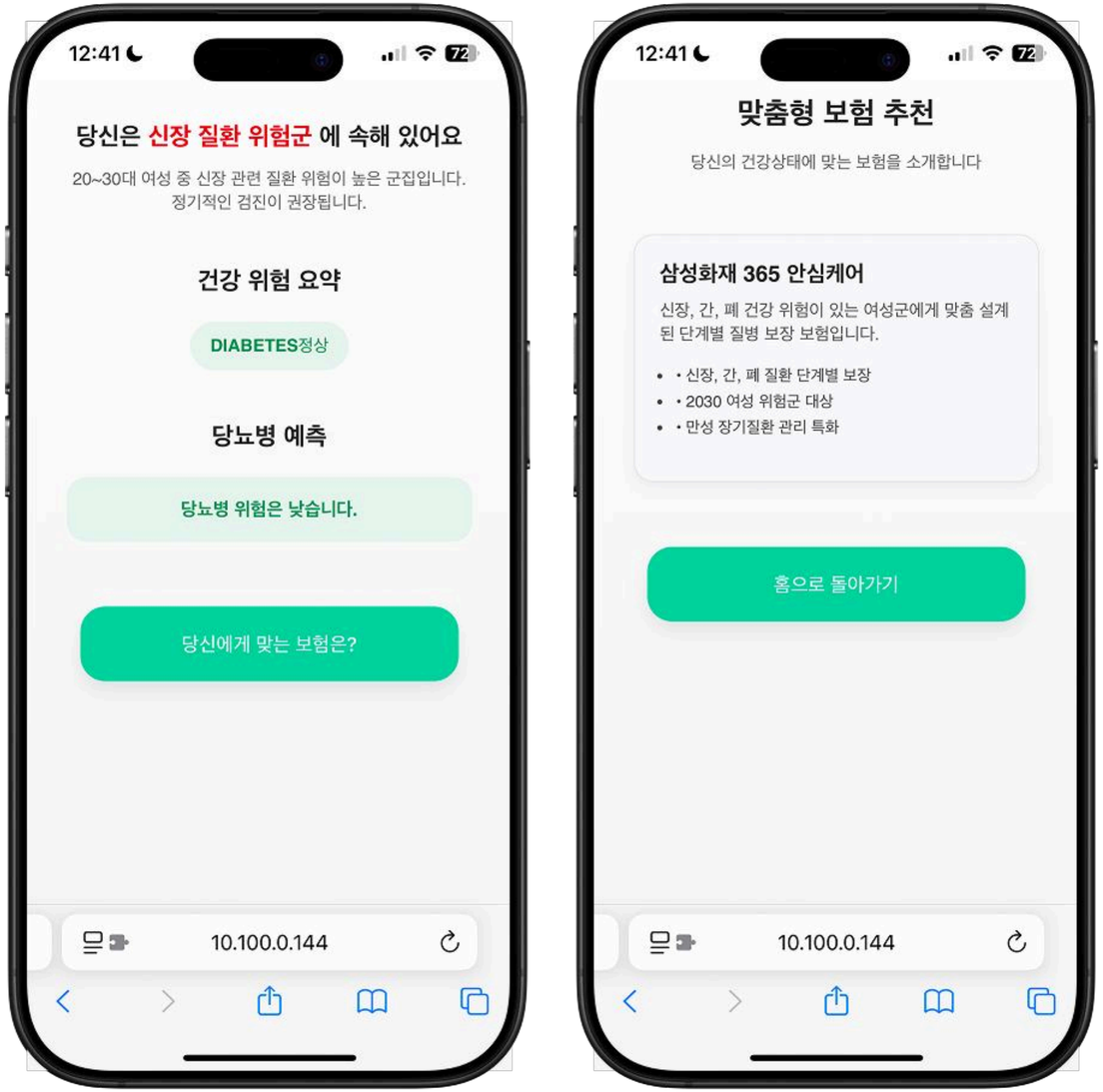


클러스터1-0 당뇨 0

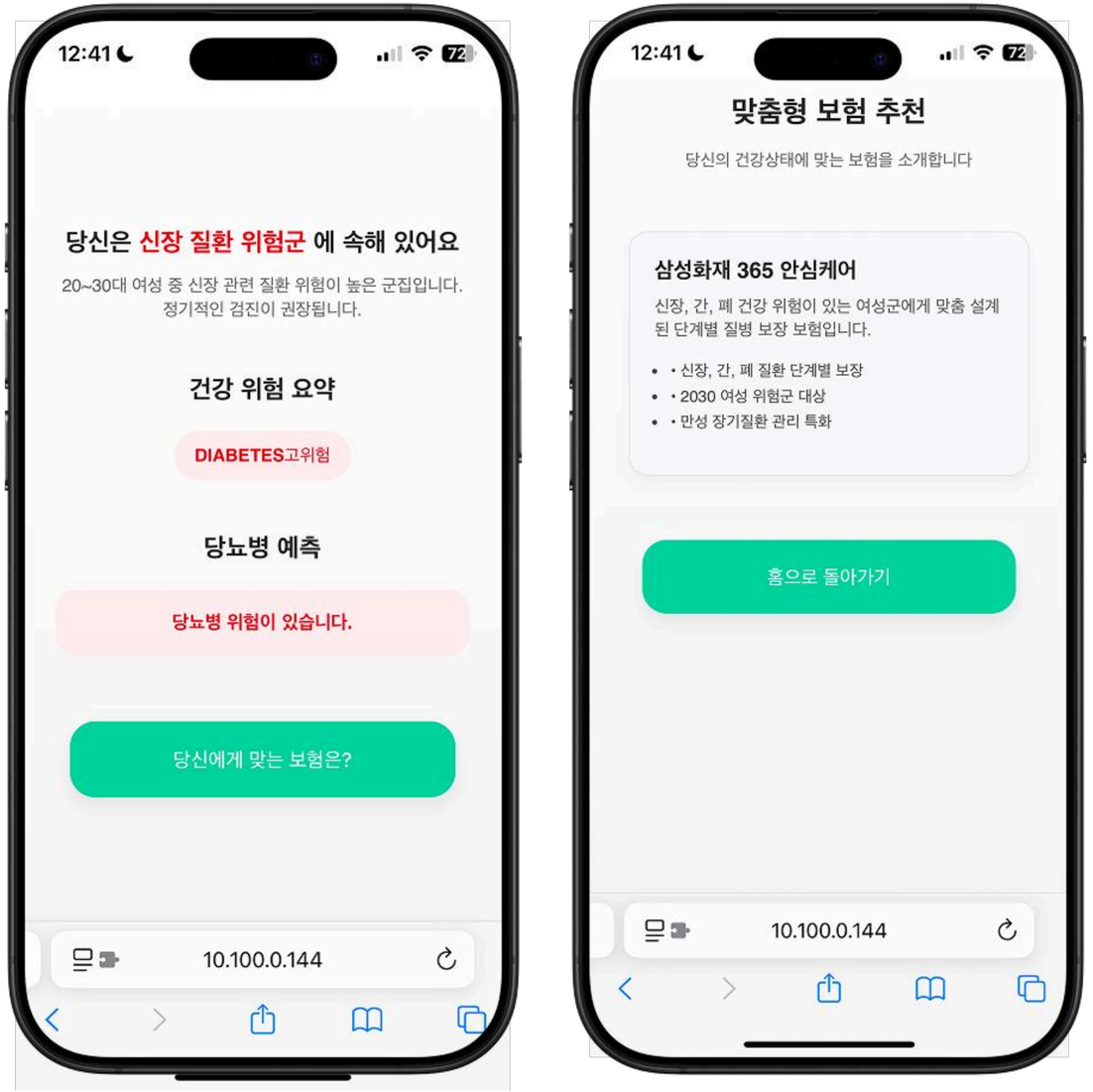


클러스터1-0 당뇨 1

건강검진 분류기 3번

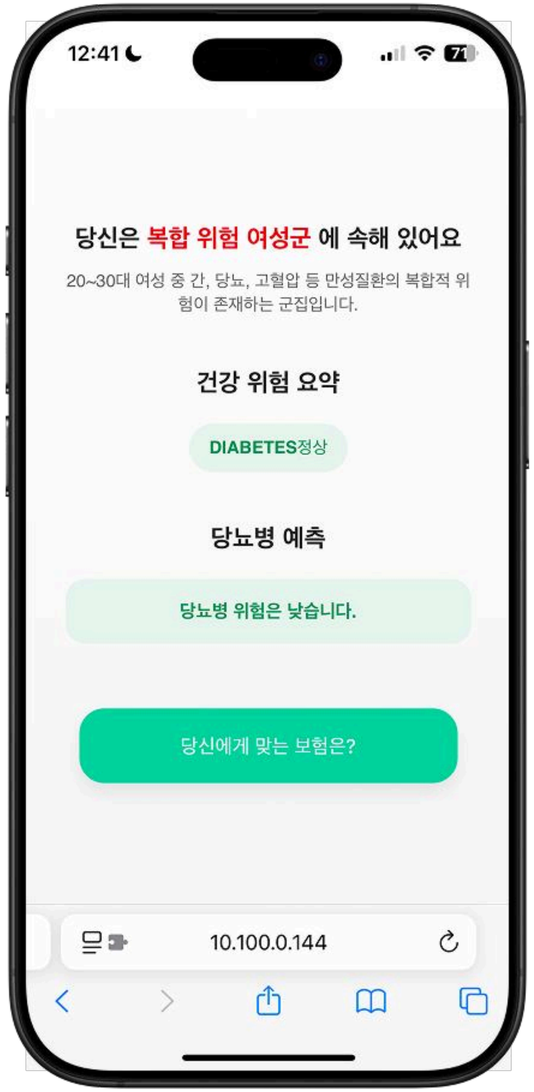


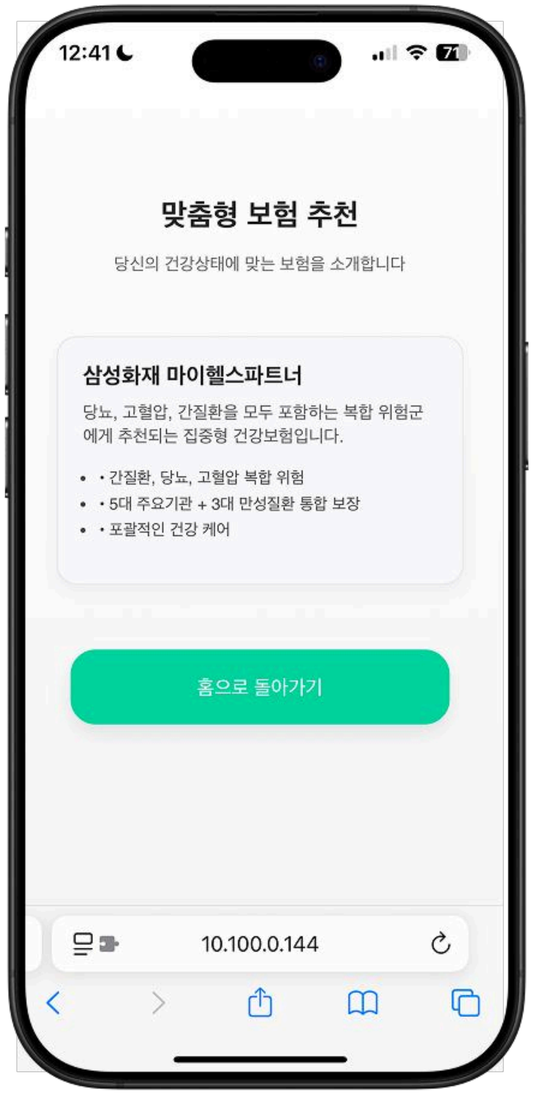
클러스터1-1 당뇨 0



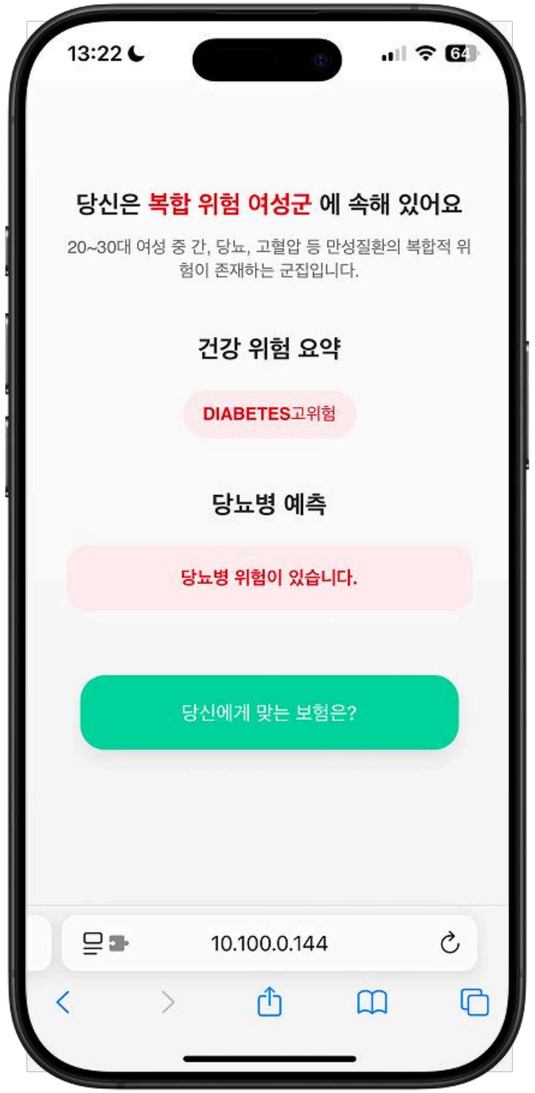
클러스터1-1 당뇨 1

건강검진 분류기 4번





클러스터1-2 당뇨 0



클러스터1-2 당뇨 1

3. 2030여성의 건강군집 분석

✅ 최종 선택 모델 – KMeans

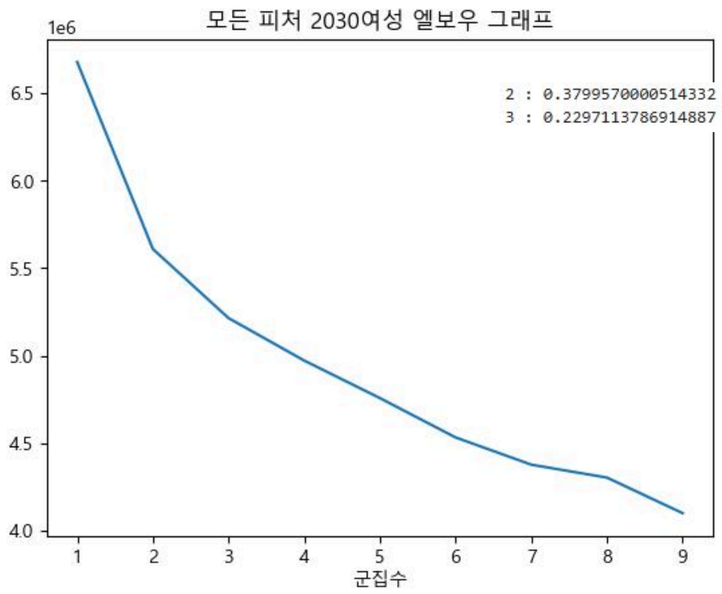
- 데이터 구성:
  - 33만 건
  - 주요변수 :

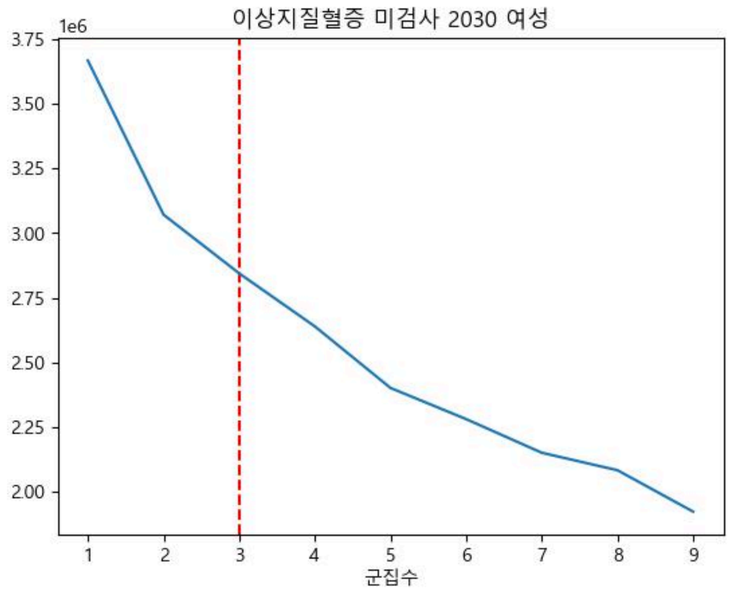
‘연령대', '허리둘레', '수축기혈압','공복혈당', '혈청크레아티닌', '혈청지오티(AST)', '혈청지피티(ALT)', '감마지티피', '흡연상태', '음주여부', 'BMI' , '사구체여과율'

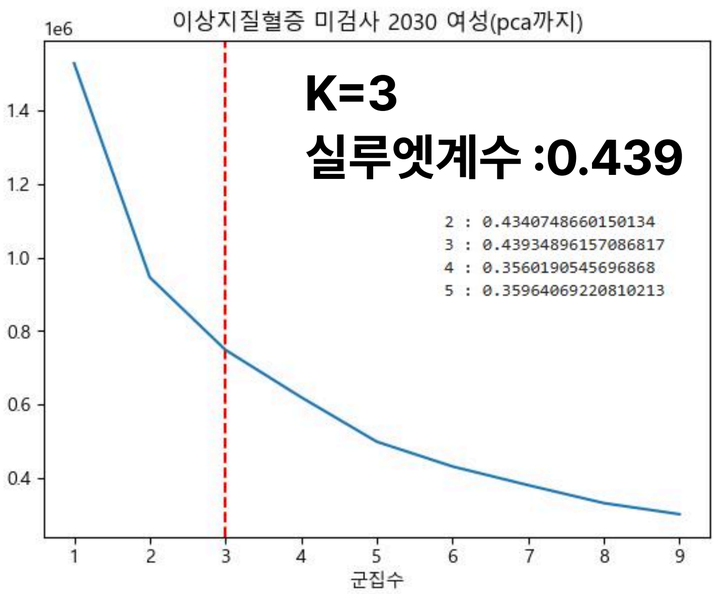
- 전처리 :
- 범주형: OrdinalEncoder
- 수치형: StandardScaler
- 차원 축소: PCA(2D)

3. 2030여성의 건강군집 분석

- **1차 : 전체 변수 다 넣었을 때**
- **사용 피처 :** '성별', '연령대', '허리둘레', '수축기혈압', '이완기혈압', '식전혈당(공복혈당)', '총콜레스테롤', '트리글리세라이드', 'HDL콜레스테롤', 'LDL콜레스테롤', '혈색소', '혈청크레아티닌', '혈청지오티(AST)', '혈청지피티(ALT)', '감마지티피', '흡연상태', '음주여부', '치아우식증유무', '치석', '청력', '평균시력', '사구체여과율'
- **2차 : 다른 변수와 상관계수 0.7 이상인 피처 제거**
  - 사용 피처 : → ex) 고혈압을 측정하는 지표 중 1개만 남김 (이완기혈압, 수축기혈압 중 수축기혈압만 채택)
    '연령대', '허리둘레', '수축기혈압','공복혈당', '혈청크레아티닌', '혈청지오티(AST)', '혈청지피티(ALT)', '감마지티피', '흡연상태', '음주여부', 'BMI' , '사구체여과율'
  - 제거 피처 : 요단백(전날 먹은 음식 영향 큼), 이완기혈압(수축기혈압으로 설명가능), 체중&키(BMI로 설명 가능), 혈색소(빈혈측정도구라 제거)
- **3차 : PCA 차원 축소 (2차원)**



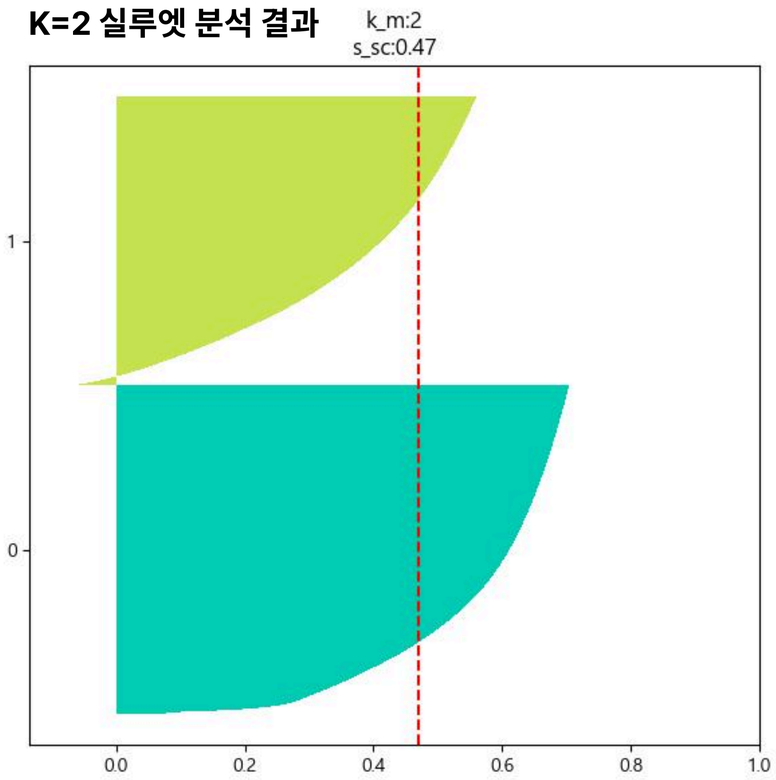




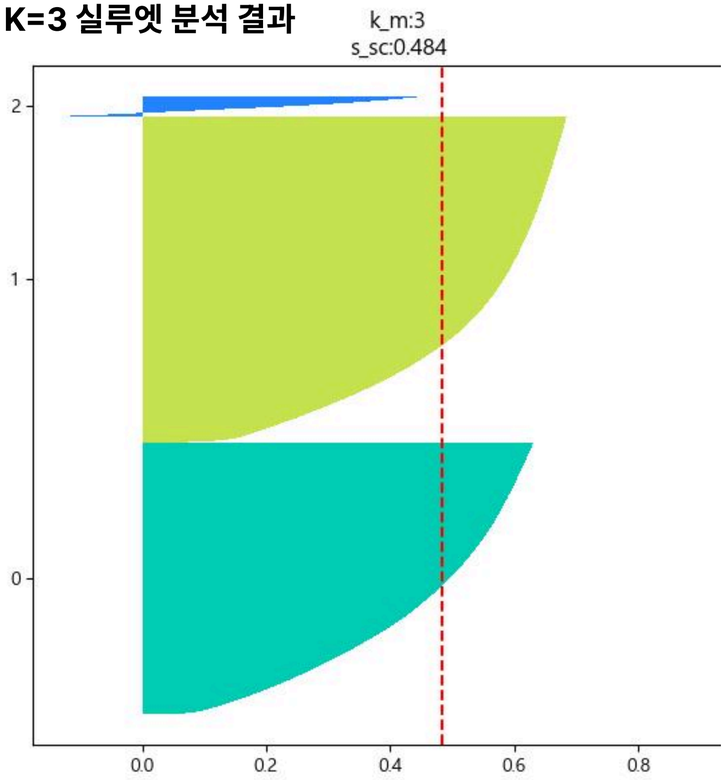
3. 2030여성의 건강군집 분석 과정

→k=2보다 k=3일 때 평균 실루엣 계수가 약간 더 높음 (0.47 → 0.484)
→ 다만, 실루엣 계수 차이가 크지 않음으로 해석과 병행

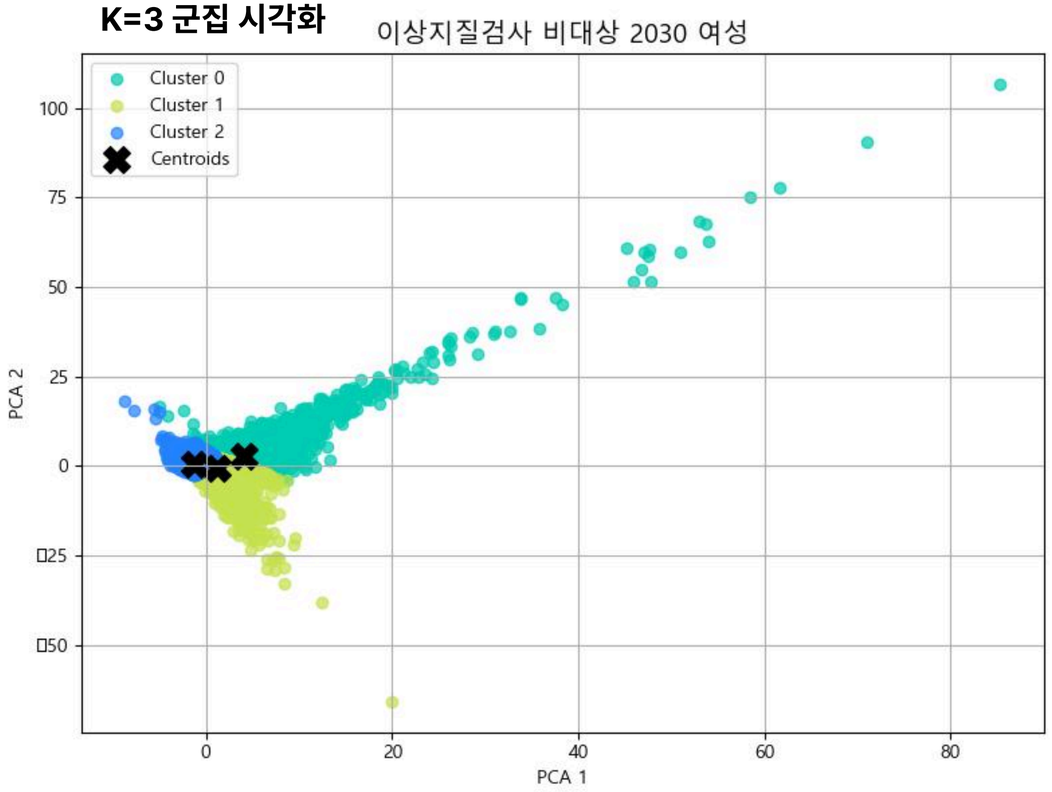
K=2 실루엣 분석 결과



K=3 실루엣 분석 결과



K=3 군집 시각화



→ K=3일 때, cluster 1은 건강, cluster1은 신장질환, cluster2는 다방면의 건강질환
→ 군집의 해석이 차이가 있었고, 2030여성을 세분화하여 더 잘 설명하므로 K=3으로 확정

3. 2030여성의 건강군집 해석

🟦 **Cluster 0 — 건강우수형 (20대 여성)** 👉 **기본 보험, 운동·식습관 유지 교육 중심**

- 대부분 지표에서 건강 상태 매우 양호

🟩 **Cluster 1 — 잠재위험 증가형 (30대 여성)**

위험 신호가 서서히 증가 중인 안정형 30대 여**성** 👉 **신장질환 보험, 대사증후군 예방형 보험 적합**

- 건강지표는 대체로 Cluster 0과 유사하지만, **신장기능 저하 비율 증가 (정상 70%)**
  - 공복혈당장애 증가 (10.7%)
  - 간질환 보유율 약간 상승

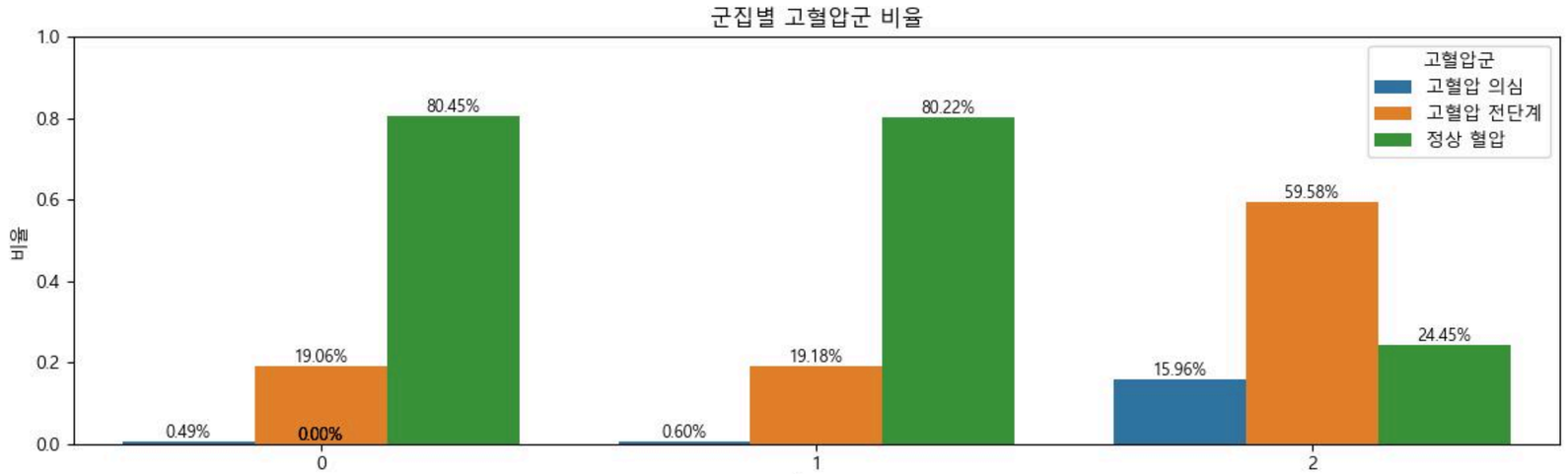
🟥 **Cluster 2 — 대사위험군 (혼합연령 여성)** 👉 **간질환/당뇨/고혈압 특화 보험**

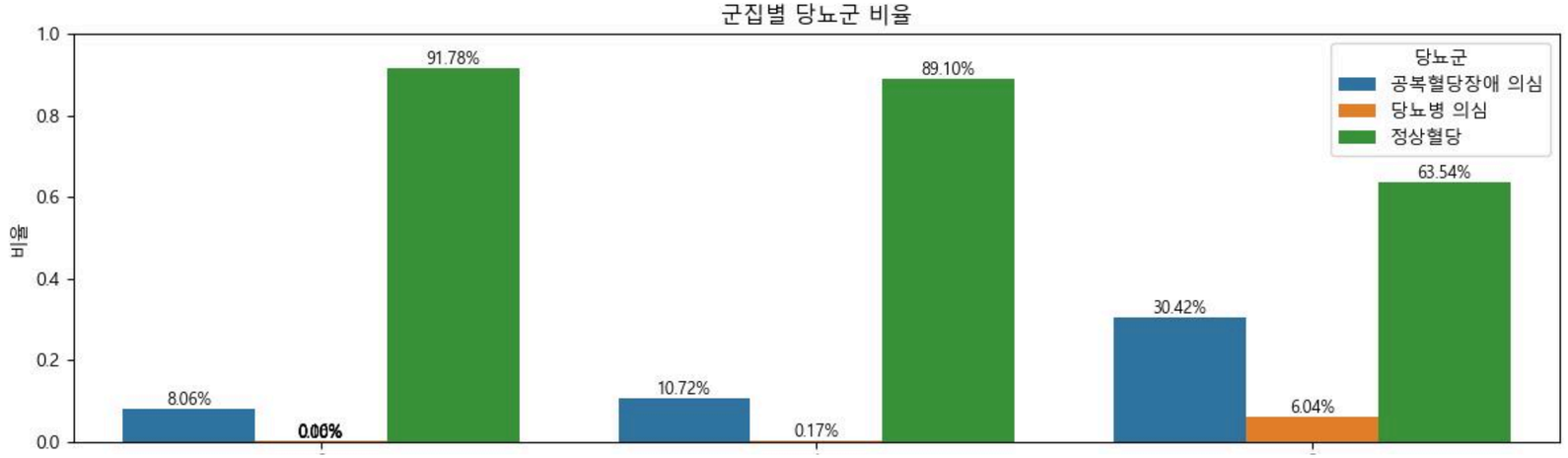
→ 대사증후군 및 만성질환 고위험군 여성

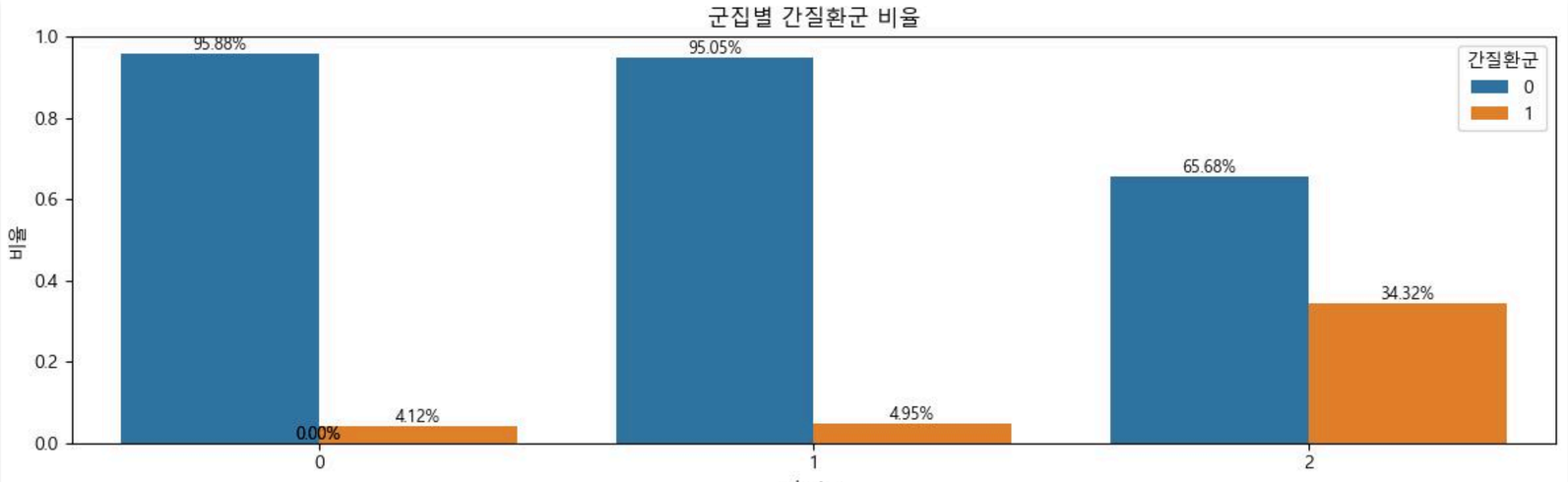
- 30대 비율 70%, 20대 30% 혼합
- **전반적인 건강 지표 악화:**
  - 고혈압 전단계 이상 75.6%
  - 정상혈당 비율 63.5%, 당뇨병 의심 6%
  - BMI 4~6: 고도 비만이 84.6%
  - 간질환 비율 34.3% → 급증

군집별 질병 발병률

- 군집 3은 고혈압, 당뇨, 간질환 위험이 있음
- 군집 0, 군집 1은 고혈압, 당뇨, 간질환의 지표가 거의 동일한 수준으로 건강함

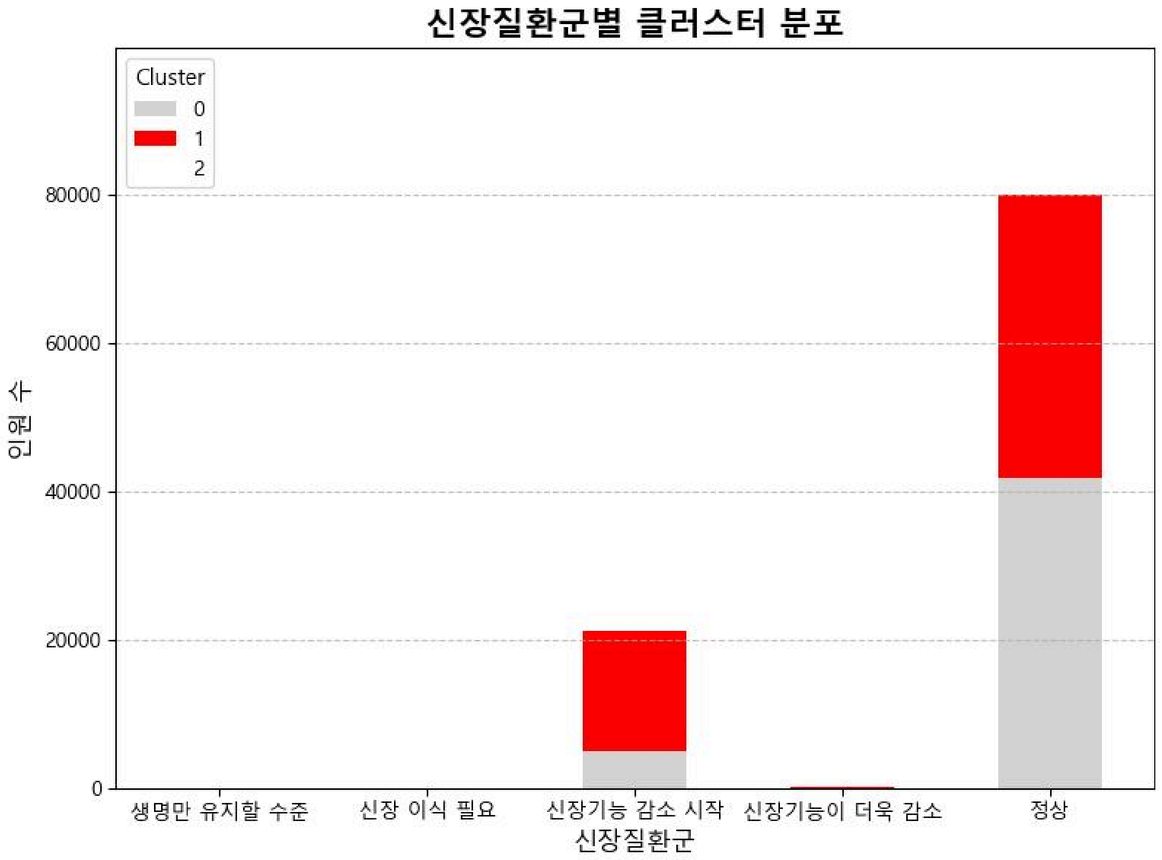






3. 2030여성의 건강군집 해석

- 건강(군집0)과 비교적건강(군집X)의 차이
  - 신장질환에서 차이가 있었음
  - 두 군집 다 건강해서 다른 지표에서는 차이가 없었느나 유일하게 신장질환에서 차이가 있었음
  - 따라서 군집0은 기본형 보험, 군집1은 신장관련 보험으로 추천할 수 있었음



3. 2030여성의 건강군집

군집별 질병 발병률

질병	군집0 (건강)	군집1 (비교적건강)	비건강 (군집)
고혈압	20%	20%	40%
당뇨	8%	11%	36%
간질환	4%	5%	34%

- **군집별 보험 전략 제안**
  - 건강한 젊은 여성군 → 운동 상해 보험 우선 추천
  - 신장 질환 위험군 → 신장 보험 우선 추천
  - 복합 위험 여성군 →

4. 20~23세 남성의 건강군집 분석

✅ 최종 선택 모델 – KMeans

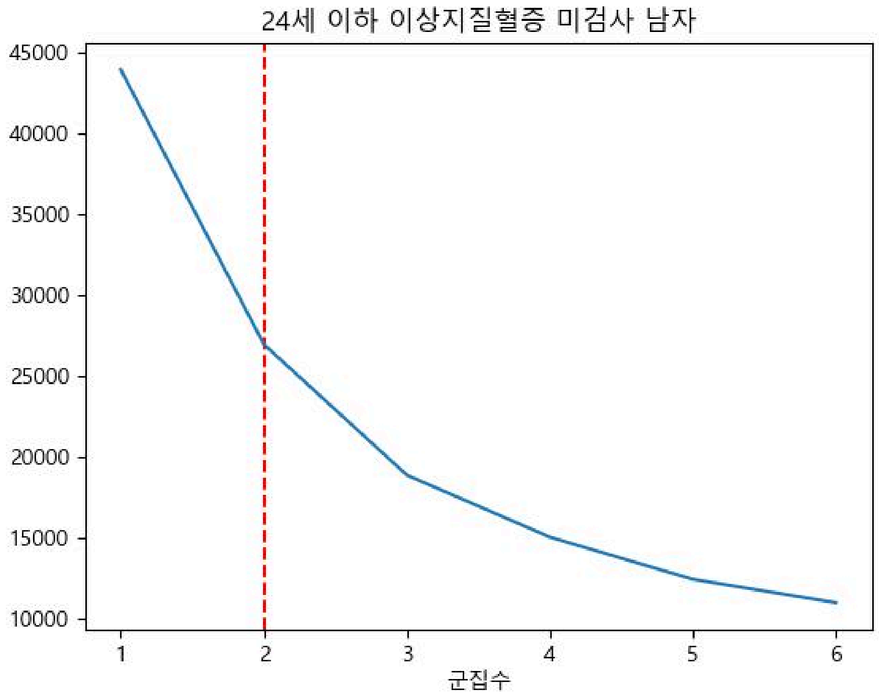
- 데이터 구성:
  - 약 8천 건의 데이터
  - 주요변수

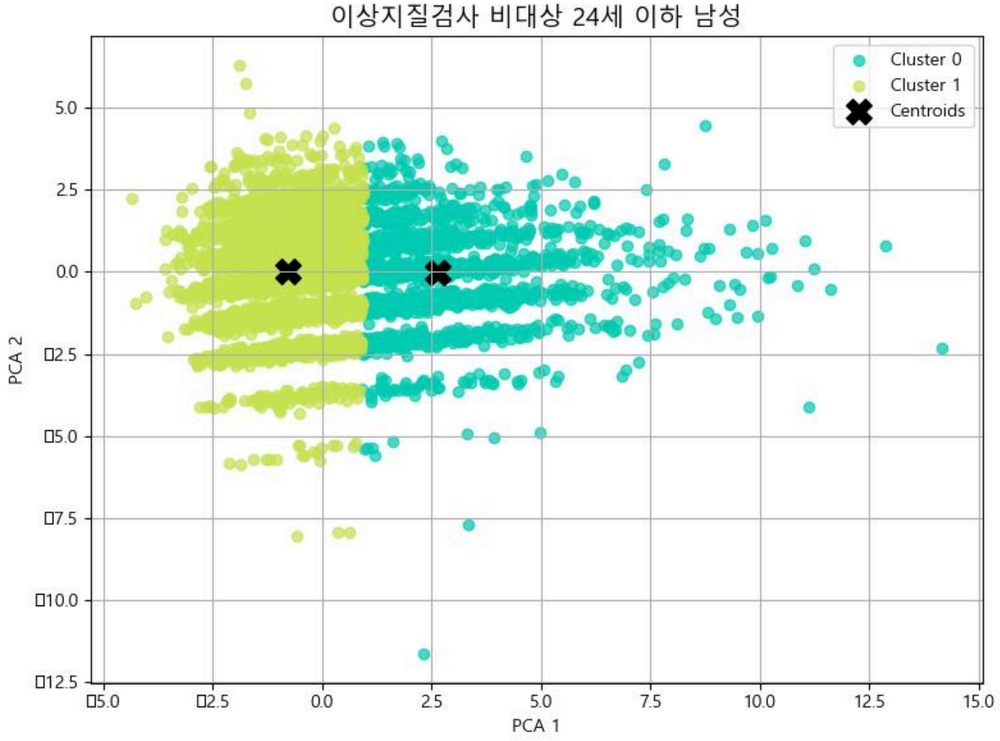
    →연령대, 허리둘레, 수축기혈압, 공복혈당, 혈청크레아티닌, 혈청지오티(AST), 혈청지피티(ALT)

    → 감마지티피, 흡연상태, 음주여부, BMI, 사구체여과율

- 전처리 :
- 범주형: OrdinalEncoder
- 수치형: StandardScaler
- 차원 축소: PCA(2D)

4. 20~23세 남성의 건강군집 분석 과정

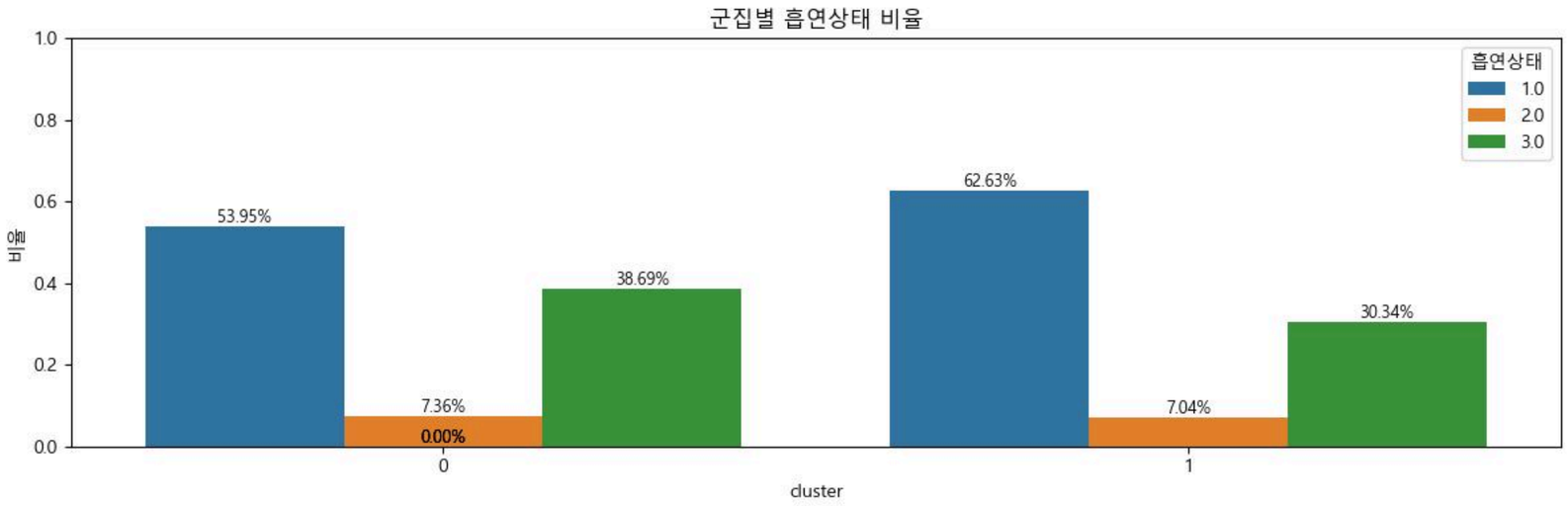


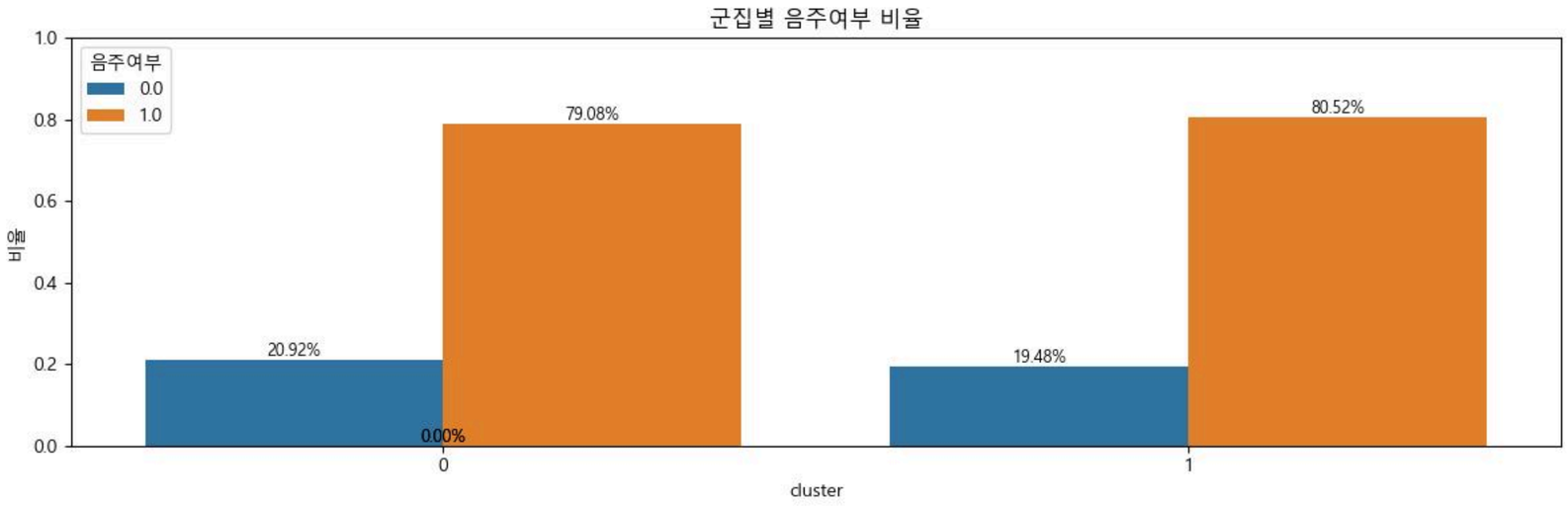



```
2 : 0.43568287161402686
3 : 0.353235269857784
4 : 0.34984924422337765
```

→ 제일 높은 실루엣계수를 보인 K=2 선택

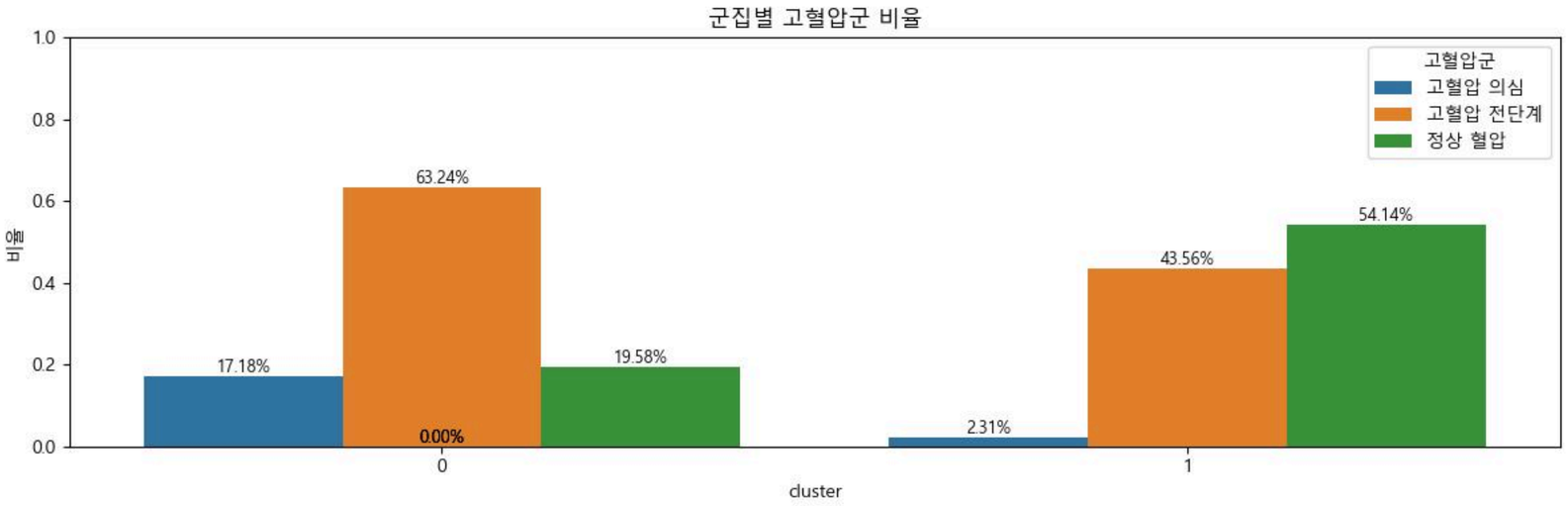
4. 20~23세 남성의 건강군집 해석

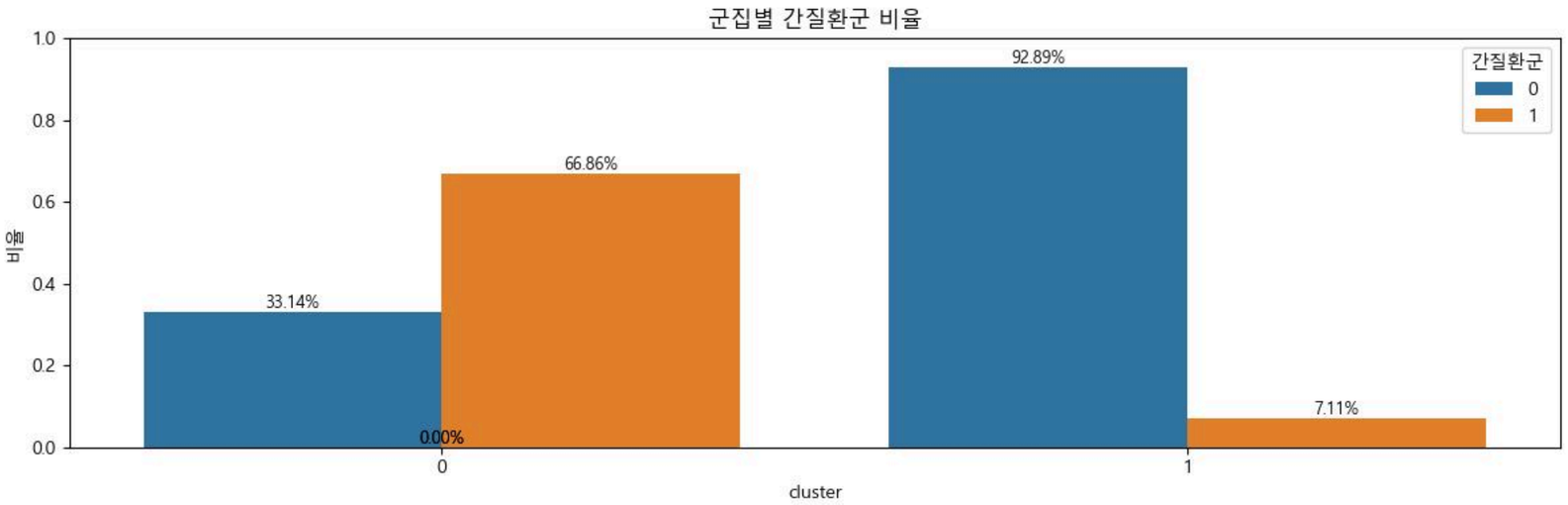




- **20대 남성은 군집별 흡연, 음주 비율에서 차이가 없었음**

- 군집별 보험 전략 제안
  - 비건"간"한 20대 초반 남성 → 간질환 보험 추천
  - 건강한 20대 초반 남성
    → 운동 상해&고혈압 보험 우선 추천
    → 20대 남성이 질병 보험을 잘 가입하지 않을 가능성 있으니 운동 보험도 함께 추천





- 군집별 질병 발병률

질병	군집0	군집1
간	66.8%	7.1%
고혈압	80.4%	45.8%

당뇨병 예측 모델

5. 당뇨 예측 모델

🔷 질병관리청 데이터 바탕의 당뇨 예측 모델 개발

→ 분석 목적 :

당뇨 예측 모델을 통해 건강검진 수검자의 검진 내역을 바탕으로 당뇨를 예측

1. 데이터셋

성별, 연령, 고혈압 진단 여부, 당뇨 진단 여부, 혈압 등의 건강 정보를 포함하여 총 30개의 컬럼으로 구성

→ 각 컬럼 별 결측치 존재 및 66666(조사안함), 77777(해당 없음), 99999(무응답/미측정) 등이 포함

해당 인스턴스 삭제 후 분석 진행

#	Column	Non-Null Count	Dtype
0	T_ID	10000 non-null	object
1	T_SEX	10000 non-null	int64
2	T_AGE	10000 non-null	int64
3	T_HTN	9982 non-null	float64
4	T_DM	9981 non-null	float64
5	T_LIP	9978 non-null	float64
6	T_FMFHT1	10000 non-null	int64
7	T_FMFHT2	10000 non-null	int64
8	T_FMFDM1	10000 non-null	int64
9	T_FMFDM2	10000 non-null	int64
10	T_DRINK	9939 non-null	float64
11	T_SMOKE	9932 non-null	float64
12	T_PSM	9859 non-null	float64
13	T_EXER	9944 non-null	float64
14	T_SBP	9959 non-null	float64
15	T_DBP	9959 non-null	float64
16	T_PULSE	9555 non-null	float64
17	T_WAIST	9868 non-null	float64
18	T_HIP	9861 non-null	float64
19	T_HEIGHT	9954 non-null	float64
20	T_WEIGHT	9953 non-null	float64
21	T_BMI	9949 non-null	float64
22	T_GLU0	9764 non-null	float64
23	T_CREATINE	9996 non-null	float64
24	T_AST	9763 non-null	float64
25	T_ALT	9744 non-null	float64
26	T_TCHL	9995 non-null	float64
27	T_HDL	9996 non-null	float64
28	T_LDL	9804 non-null	float64
29	T_TG	9974 non-null	float64

5. 당뇨 예측 모델

2. 이상치
   각 변수에 대해 IQR을 기준으로 1, 3분위수의 1.5*IQR값을 벗어난 값을 이상치로 간주하고, 이를 삭제 후 분석 진행하였으나 모델의 성능이 하락
   → 이상치를 제거하지 않고 분석 진행

3. 데이터 분할, 오버샘플링 및 스케일링
   Train_Test_Split을 통해 학습 데이터와 테스트 데이터를 분리한 후,
   학습 데이터에 SMOTE기법을 활용해 분류의 불균형성을 조정
   학습 데이터에 StandardScaler를 사용해 스케일링 후 학습 진행

#	Column	Non-Null Count	Dtype
0	T_ID	6473 non-null	object
1	T_SEX	6473 non-null	int64
2	T_AGE	6473 non-null	int64
3	T_HTN	6473 non-null	float64
4	T_DM	6473 non-null	float64
5	T_LIP	6473 non-null	float64
6	T_FMFHT1	6473 non-null	int64
7	T_FMFHT2	6473 non-null	int64
8	T_FMFDM1	6473 non-null	int64
9	T_FMFDM2	6473 non-null	int64
10	T_DRINK	6473 non-null	float64
11	T_SMOKE	6473 non-null	float64
12	T_PSM	6473 non-null	float64
13	T_EXER	6473 non-null	float64
14	T_SBP	6473 non-null	float64
15	T_DBP	6473 non-null	float64
16	T_PULSE	6473 non-null	float64
17	T_WAIST	6473 non-null	float64
18	T_HIP	6473 non-null	float64
19	T_HEIGHT	6473 non-null	float64
20	T_WEIGHT	6473 non-null	float64
21	T_BMI	6473 non-null	float64
22	T_GLU0	6473 non-null	float64
23	T_CREATINE	6473 non-null	float64
24	T_AST	6473 non-null	float64
25	T_ALT	6473 non-null	float64
26	T_TCHL	6473 non-null	float64
27	T_HDL	6473 non-null	float64
28	T_LDL	6473 non-null	float64
29	T_TG	6473 non-null	float64

5. 당뇨 예측 모델

4. 분석 진행 - 데이터의 모든 변수 학습

- 원본 데이터셋의 모든 변수를 모델 학습에 사용
  - → Decision Tree Classifier
  - → Random Forest Classifier
  - → Gradient Boosting Classifier
  - → Support Vector Classifier
  - → XG Boost Classifier

```
Decision Tree Classifier
0.3548387096774194

Random Forest Classifier
0.5045045045045045

Gradient Boosting Classifier
0.5677749360613811

Support Vector Classifier
0.4672897196261682

XG Booster Classifier
0.5
```

```
Decision Tree Classifier
              precision    recall  f1-score   support

         0.0       0.96      0.93      0.94      2492
         1.0       0.31      0.42      0.35       185

    accuracy                           0.90      2677
   macro avg       0.63      0.67      0.65      2677
weighted avg       0.91      0.90      0.90      2677
```

```
Random Forest Classifier
              precision    recall  f1-score   support

         0.0       0.96      0.97      0.97      2492
         1.0       0.57      0.45      0.50       185

    accuracy                           0.94      2677
   macro avg       0.76      0.71      0.74      2677
weighted avg       0.93      0.94      0.94      2677
```

```
Gradient Boosting Classifier
              precision    recall  f1-score   support

         0.0       0.97      0.96      0.97      2492
         1.0       0.54      0.60      0.57       185

    accuracy                           0.94      2677
   macro avg       0.75      0.78      0.77      2677
weighted avg       0.94      0.94      0.94      2677
```

```
Support Vector Classifier
              precision    recall  f1-score   support

         0.0       0.97      0.94      0.95      2492
         1.0       0.41      0.54      0.47       185

    accuracy                           0.91      2677
   macro avg       0.69      0.74      0.71      2677
weighted avg       0.93      0.91      0.92      2677
```

```
XG Booster Classifier
              precision    recall  f1-score   support

         0.0       0.98      0.90      0.94      2492
         1.0       0.37      0.78      0.50       185

    accuracy                           0.89      2677
   macro avg       0.68      0.84      0.72      2677
weighted avg       0.94      0.89      0.91      2677
```

→ Gradient Boosting Classifier의 성능이 F1-Score=0.5678로 가장 높게 나타남

5. 당뇨 예측 모델

5. 분석 진행 - 공통의 변수 학습

원본 데이터셋 중 건강검진 데이터셋의 변수에 존재하는 변수를 모델 학습에 사용

성별, 연령, 신장, 체중, 허리둘레, 수축기혈압, 이완기혈압, 공복혈당, 총 콜레스테롤, 트리글리세라이드, HDL 콜레스테롤, LDL콜레스테롤, 크레아티닌, 혈청지오티(AST), 혈청지피티(ALT), 흡연 여부, 음주 여부

```
Decision Tree Classifier
0.3755274261603376

Random Forest Classifier
0.5857519788918206

Gradient Boosting Classifier
0.5539906103286385

Support Vector Classifier
0.5062761506276151

XG Booster Classifier
0.5757575757575758
```

```
Decision Tree Classifier
              precision    recall  f1-score   support

         0.0       0.96      0.92      0.94      2492
         1.0       0.31      0.48      0.38       185

    accuracy                           0.89      2677
   macro avg       0.63      0.70      0.66      2677
weighted avg       0.91      0.89      0.90      2677
```

```
Random Forest Classifier
              precision    recall  f1-score   support

         0.0       0.97      0.97      0.97      2492
         1.0       0.57      0.60      0.59       185

    accuracy                           0.94      2677
   macro avg       0.77      0.78      0.78      2677
weighted avg       0.94      0.94      0.94      2677
```

```
Gradient Boosting Classifier
              precision    recall  f1-score   support

         0.0       0.97      0.95      0.96      2492
         1.0       0.49      0.64      0.55       185

    accuracy                           0.93      2677
   macro avg       0.73      0.79      0.76      2677
weighted avg       0.94      0.93      0.93      2677
```

```
Support Vector Classifier
              precision    recall  f1-score   support

         0.0       0.97      0.93      0.95      2492
         1.0       0.41      0.65      0.51       185

    accuracy                           0.91      2677
   macro avg       0.69      0.79      0.73      2677
weighted avg       0.93      0.91      0.92      2677
```

```
XG Booster Classifier
              precision    recall  f1-score   support

         0.0       0.96      0.98      0.97      2492
         1.0       0.66      0.51      0.58       185

    accuracy                           0.95      2677
   macro avg       0.81      0.75      0.77      2677
weighted avg       0.94      0.95      0.94      2677
```

→ Random Forest Classifier의 성능이 F1-Score=0.5858로 가장 높게 나타남

5. 당뇨 예측 모델

6. 분석 진행 - 선별된 변수 학습

원본 데이터셋 중 연관을 미칠 것이라 판단되는 변수를 모델 학습에 사용

공복혈당, 체중, 허리둘레, 연령, 트리글리세라이드, HDL혈당, 음주여부, 혈청지오티(ALT)

```
Decision Tree Classifier
0.38876889848812096

Random Forest Classifier
0.5833333333333334

Gradient Boosting Classifier
0.5759637188208617

Support Vector Classifier
0.48120300751879697

XG Booster Classifier
0.5525525525525525
```

```
Decision Tree Classifier
              precision    recall  f1-score   support

           0       0.96      0.92      0.94      2492
           1       0.32      0.49      0.39       185

    accuracy                           0.89      2677
   macro avg       0.64      0.71      0.67      2677
weighted avg       0.92      0.89      0.90      2677
```

```
Random Forest Classifier
              precision    recall  f1-score   support

           0       0.97      0.97      0.97      2492
           1       0.56      0.61      0.58       185

    accuracy                           0.94      2677
   macro avg       0.77      0.79      0.78      2677
weighted avg       0.94      0.94      0.94      2677
```

```
Gradient Boosting Classifier
              precision    recall  f1-score   support

           0       0.98      0.95      0.96      2492
           1       0.50      0.69      0.58       185

    accuracy                           0.93      2677
   macro avg       0.74      0.82      0.77      2677
weighted avg       0.94      0.93      0.94      2677
```

```
Support Vector Classifier
              precision    recall  f1-score   support

           0       0.98      0.91      0.94      2492
           1       0.37      0.69      0.48       185

    accuracy                           0.90      2677
   macro avg       0.67      0.80      0.71      2677
weighted avg       0.93      0.90      0.91      2677
```

```
XG Booster Classifier
              precision    recall  f1-score   support

           0       0.96      0.98      0.97      2492
           1       0.62      0.50      0.55       185

    accuracy                           0.94      2677
   macro avg       0.79      0.74      0.76      2677
weighted avg       0.94      0.94      0.94      2677
```

→ Random Forest Classifier의 성능이 F1-Score=0.5833으로 가장 높게 나타남

5. 당뇨 예측 모델

7. 분석 진행 - 최선의 모델 학습

Random Forest Classifier에
그리드서치를 통해 최선의 하이퍼파라미터 획득

max_depth=None, criterion='gini', max_features='log2',
min_samples_leaf=1, min_samples_split=2,
n_estimators=100, class_weight='balanced',
random_state=62

Random Forest Classifier

	precision	recall	f1-score	support
0	0.97	0.97	0.97	2492
1	0.58	0.61	0.59	185
accuracy			0.94	2677
macro avg	0.78	0.79	0.78	2677
weighted avg	0.94	0.94	0.94	2677

f1-score
0.5925925925925926

→ Random Forest Classifier의 성능이 F1-Score=0.5926으로 나타남

이탈 고객 분석

→ 추가적으로 이탈 고객 분석을 통해 이탈 고객을 대상으로 건강검진 분석 서비스를 기획하고자 함

이탈 고객 특징

휴면 계좌란?

출처: iM증권 홈페이지

- 최근 6개월간 매매거래 및 입출금·입출고 등이 발생 안한 예탁자산 평가액 10만원 이하인 계좌
- 최근 6개월간 매매거래 및 입출금·입출고 등이 발생 하지 않고 현금 및 금융투자상품 등 예탁자산 평가액이 10만원 초과 1,000만원 이하인 반송계좌

수신금액

- 수신상품은 예탁자산의 핵심
- 고객 자산 상태 직접 반영
- 이탈 위험 예측에 중요

대출금액

- 대출금액은 예탁자산에 간접 영향
- 고객 재무 상황 파악 도움

→ 휴면 계좌 기준에 직접적인 영향을 주지 않기 때문에 이탈 고객 판단시 대출금액을 제외

휴면계좌한계

반송 계좌 데이터 부재
→ 평가액 기준 설정의 어려움
→ 10만원 이하와 1000만원 이하 기준 혼재

절대적 휴면 기준

6개월 간 거래 내역 부재

이탈 고객 특징

휴면계좌와 이탈고객의 관계

수신금액의 잔액이 있는 휴면 계좌는
재활성화 가능성 존재
→ 모든 휴면 계좌를 이탈로 간주하면 안됨

6개월간 무거래와 수신 총합 0원이 이탈의 주요 지표
휴면 계좌 중 일부만 이탈 고객으로 분류해야 함
→ 휴면 고객 중 수신 금액의 총합이 0인 고객을 이탈 고객으로

→ 이탈 고객과 비이탈 고객 정의
이탈 : 2023년 7월 이전에 마지막 결제
&
수신 금액 총합이 0인 고객
비이탈 : 이탈 고객이 아닌 고객

수신금액 0원 계좌의 의미

고객의 계좌 사용 의사 추정

- 계좌 무관심 가능성 높음
- 타 금융사 이용 추정
- 고객 니즈 변화 시사

이탈 고객 판단 근거

- 거래 부재 + 수신 계좌 총합 잔액 0
- 계좌 폐쇄 의향 추정
- 고객 이탈 위험 높음

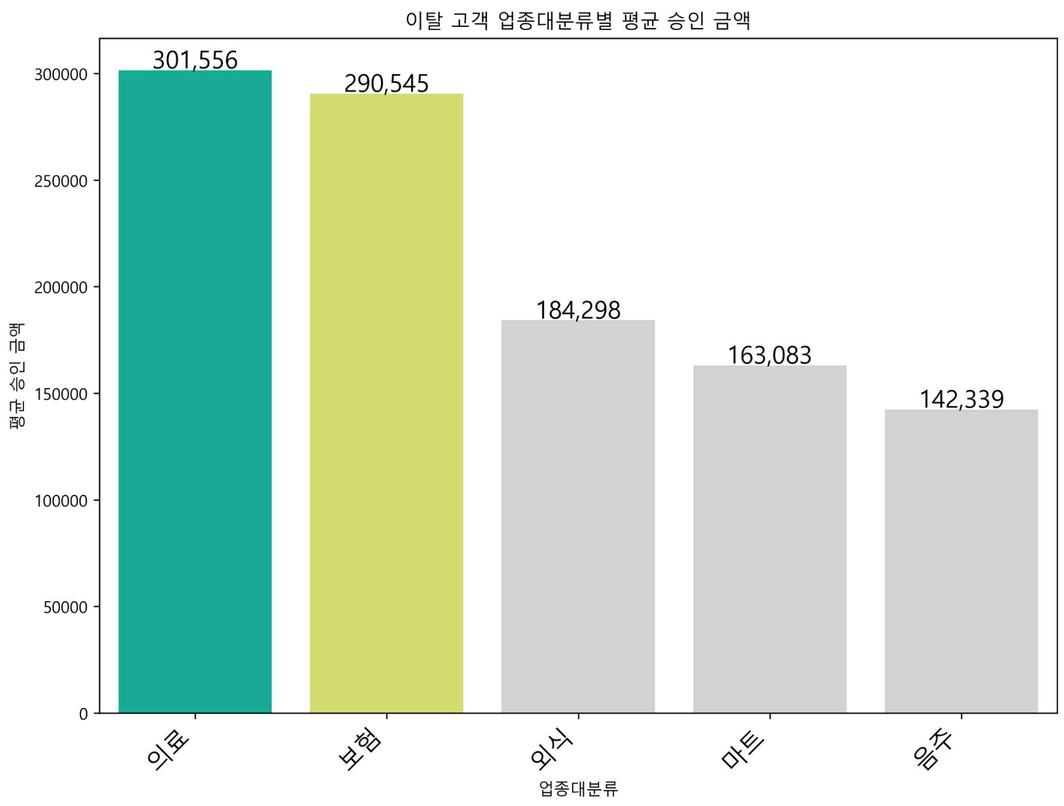
자금 이동 가능성

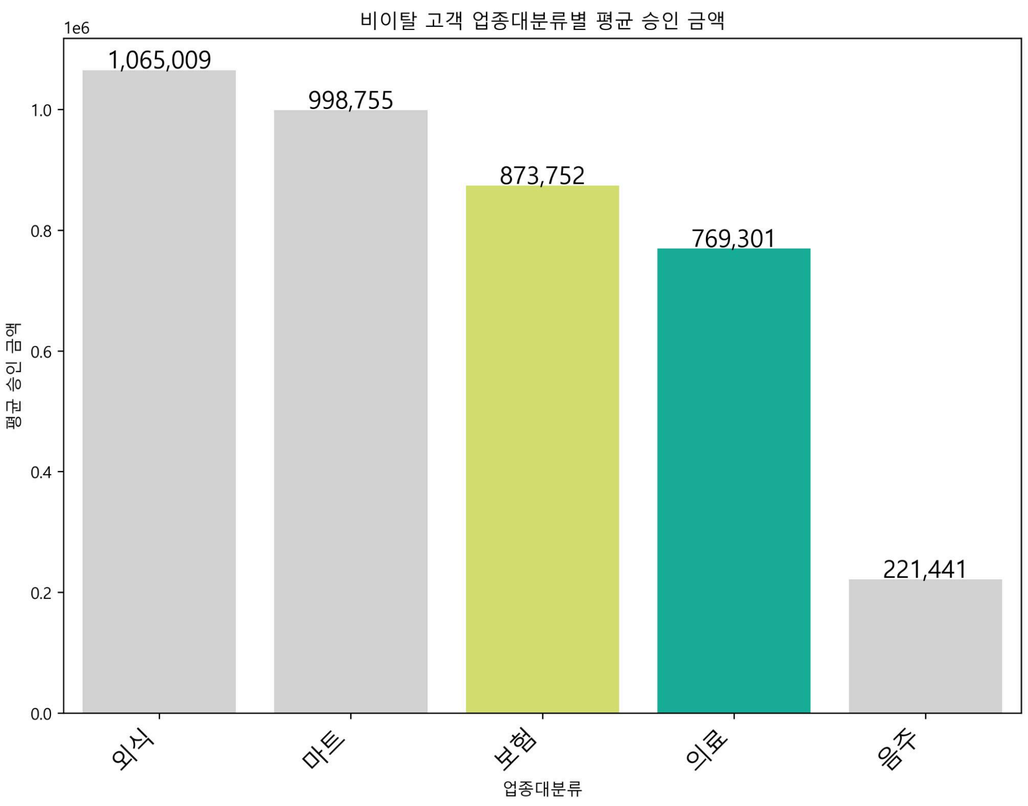
- 타 금융상품 선호 가능
- 경제 상황 변화 반영
- 투자 전략 변경 암시

재활성화 전략 필요성

- 맞춤형 상품 제안
- 고객 접촉 강화 필요
- 이탈 방지 전략 수립

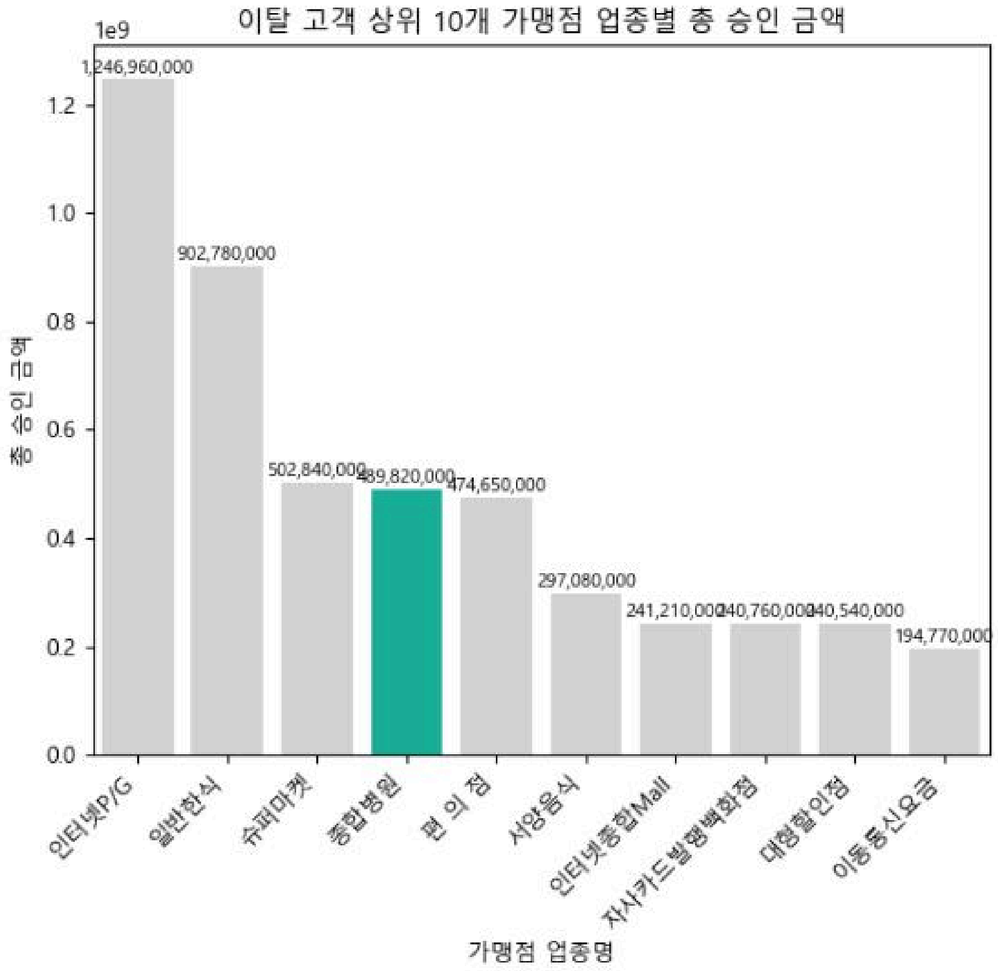
이탈 고객 특징

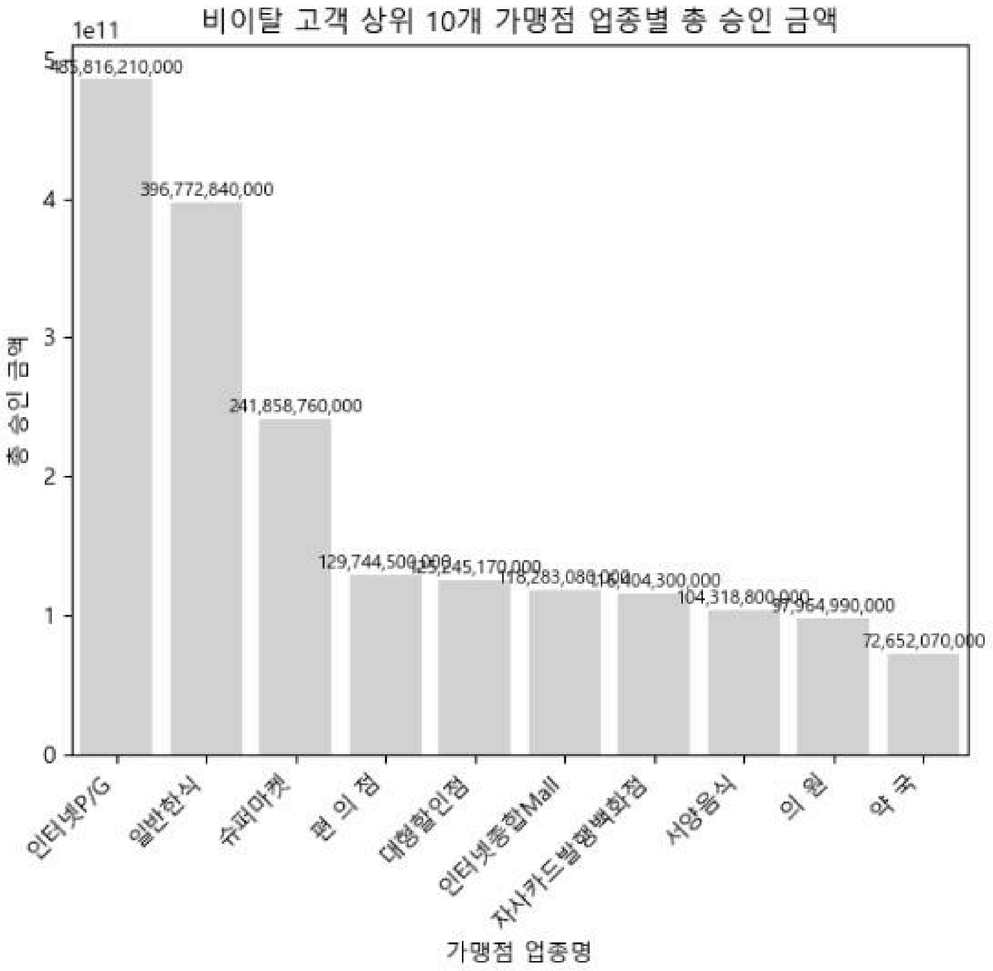




이탈 집단이 비이탈 집단보다 **의료와 보험**에 소비를 많이 하는 경향
→ 이탈 고객들은 비이탈 고객보다 건강에 관심이 많다고 해석 가능
→ 건강 서비스 또는 보험 서비스를 제공하면 고객 이탈 방지에 도움

이탈 고객 특징





특히, 이탈 집단이 비이탈 집단보다 **종합병원**에 소비를 많이 하는 경향

- 이탈 고객 대상 앱서비스 홍보 전략 제안

대상 고객 : 건강 관련 지출 비중이 높은 이탈 고객

- 알림 메시지 예시

  "건강검진 결과 업로드하면 커피 쿠폰 드려요!"

→ 건강 데이터 기반 맞춤 분석 리포트 제공 안내 포함

→ 군집별 캐릭터를 만들어서 MBTI 처럼 건강형 유형 제공

자체 평가 의견

자체 평가 의견

사전 기획의 관점에서 프로젝트 결과물에 대한 완성도 평가(8/10점)

1. 목표 달성도 (3/4점)
일부 미흡한 부분이 있어 추가적인 개선이 필요
2. 범위 충실도 (3/3점)
결과물이 사전 기획에서 정의했던 모든 범위와 기능을 포함하고 있음
3. 계획 준수도 (1/2점)
일부 계획 변경이 있었음
4. 결과물의 완성도 및 품질 (1점)
최종 결과물이 사용 목적에 적합하고, 기능, 성능, 디자인 등 모든 면에서 높은 수준의 완성도와 품질을 갖춤

잘한 부분과 아쉬운 점

잘한 부분: 실제 존재하는 어플의 발전과 경쟁력을 고민해보고 머신러닝 모델을 활용해 새로운 서비스를 구현함

아쉬운 점: 고객의 진료내역을 결합해 진료내역 기반의 보험 추천모델을 만들어 서비스를 구현해보고자 했으나, 진료내역 데이터를 심사받는 데 25일이라는 시간이 걸려 분석해보지 못했음

프로젝트 결과물의 추후 개선점이나 보완할 점

향후 프로젝트 개선을 위해
데이터 수집 전에 필요한 데이터와 확보 방안을 명확히 계획하고, 데이터 품질 기준을 미리 설정하는 것이 필요함

→ 이를 통해 데이터 관련 문제를 사전에 방지하고 프로젝트 완성도를 높일 수 있을 것이다.

프로젝트를 수행하면서 느낀 점이나 경험한 성과

1. 건강 검진 분류 모델 구축
건강검진 데이터 기반으로 특성을 분류하는 모델을 성공적으로 개발함
2. 개인 맞춤형 추천 시스템 구현
고객 데이터 분석을 통해 최적의 보험 상품을 추천하는 시스템을 구축
특정 목표에 부합하는 고객 그룹을 효과적으로 추천하는 분석 역량을 확보
군집 분석을 통해 고객을 세분화하고 각 그룹의 특징을 명확히 파악하여, 타겟 마케팅 전략 수립의 기반을 마련
3. 이탈 고객 특성 파악
이탈 가능성이 높은 고객을 예측하고 특징을 분석하여

프로젝트 방향성 피드백

기존 계획

- 국민건강보험공단의 건강검진 데이터와 진료내역 데이터 활용
- 성별, 연령대별 발병률이 높은 질병에 대한 보험 추천
- 건강/비건강 군집 분석 및 집단 비교

수정의 필요성

- 데이터 연결의 한계: 가입자 일련번호 익명화 문제
- nhiss 데이터 요청 시간 문제
- 프로젝트 기간 내 데이터 수집 불가능

수정 과정

- 질병 관련 데이터셋 재탐색
- 특정 질병 분석으로 방향 전환
- 암 데이터셋 검토 및 한계 파악

데이터셋 탐색 노력

- 암 데이터셋 및 파생 논문 검토
- 세부적인 지표 부족으로 인한 한계
- 암 분석 제외 결정

최종 수정 내용

- 질병 관련 데이터셋 활용
- 건강검진 데이터 기반 군집화특정 질병(예: 당뇨병)에 대한 심층 분석

수정 후 기대효과

- 특정 질병 예측 환자 대상 보험 추천
- 보험사의 리스크 감소
- 고객 맞춤형 건강 관리 서비스 제공 가능성

QnA

발표는 여기까지입니다. 감사합니다.

데이터 설명: iM뱅크 카드, 고객 데이터

- 2023년 데이터만 사용한 이유

제공받은 데이터는 2021년, 2022년, 2023년 데이터였으나 우리는 2023년 데이터만 사용

1. 2021년 2022년 2023년 데이터에 공통된 고유 고객ID가 많았고
2. 2021년 또는 2022년에 고객이었으나 2023년에는 더 이상 고객이 아닌 사람들은 1,997,337명 중 22,789명으로 약 1.1%에 불과

→ 가장 최신의 고객 정보로 분석을 하기 위해 2023년 데이터만 선택함

```
cus1['고객ID'].nunique(), cus2['고객ID'].nunique(),cus3['고객ID'].nunique(),customer['고객ID'].nunique()
```

```
(1781906, 1857530, 1974548, 1997337)
```

```
# 2021년 또는 2022년에 고객이었으나 2023년에는 더 이상 고객이 아닌 사람들
customer['고객ID'].nunique() - cus3['고객ID'].nunique()
```

```
22789
```

데이터 설명: iM뱅크 카드, 고객 데이터

```
customer.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1595714 entries, 0 to 1595713
Data columns (total 15 columns):
 #   Column     Non-Null Count    Dtype
---  ------     --------------    -----
 0   고객ID      1595714 non-null  category
 1   연령대       1595714 non-null  category
 2   성별        1595714 non-null  category
 3   고객등급      1595714 non-null  category
 4   자택_시도     1595714 non-null  category
 5   자택_시군구    1595714 non-null  category
 6   수신_요구불예금  1595714 non-null  int64
 7   수신_거치식예금  1595714 non-null  int64
 8   수신_적립식예금  1595714 non-null  int64
 9   수신_펀드     1595714 non-null  int64
 10  수신_외화예금   1595714 non-null  int64
 11  대출금액      1595714 non-null  int64
 12  수신_합계     1595714 non-null  int64
 13  이탈여부      1595714 non-null  category
 14  월         1595714 non-null  object
dtypes: category(7), int64(7), object(1)
memory usage: 158.6+ MB
```

- object형 데이터들을 category형으로 형변환
- '자택_시군구'의 null값들을 결측치 처리

```
card.info()
```

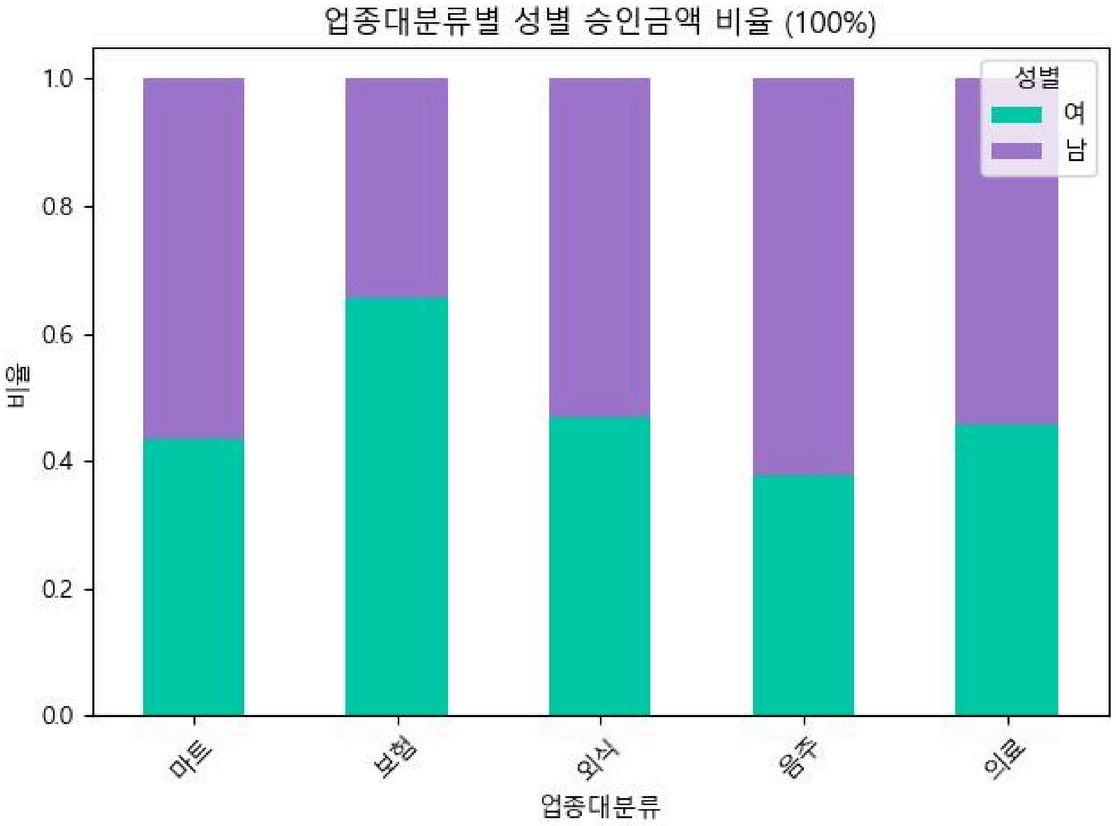
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 63313049 entries, 0 to 63313048
Data columns (total 8 columns):
 #   Column     Dtype
---  ------     -----
 0   고객ID      category
 1   가맹점업종명     category
 2   가맹점_광역시도   category
 3   가맹점_시군구    category
 4   승인건수       category
 5   승인금액       int64
 6   이탈여부       category
 7   월          object
dtypes: category(6), int64(1), object(1)
memory usage: 1.6+ GB
```

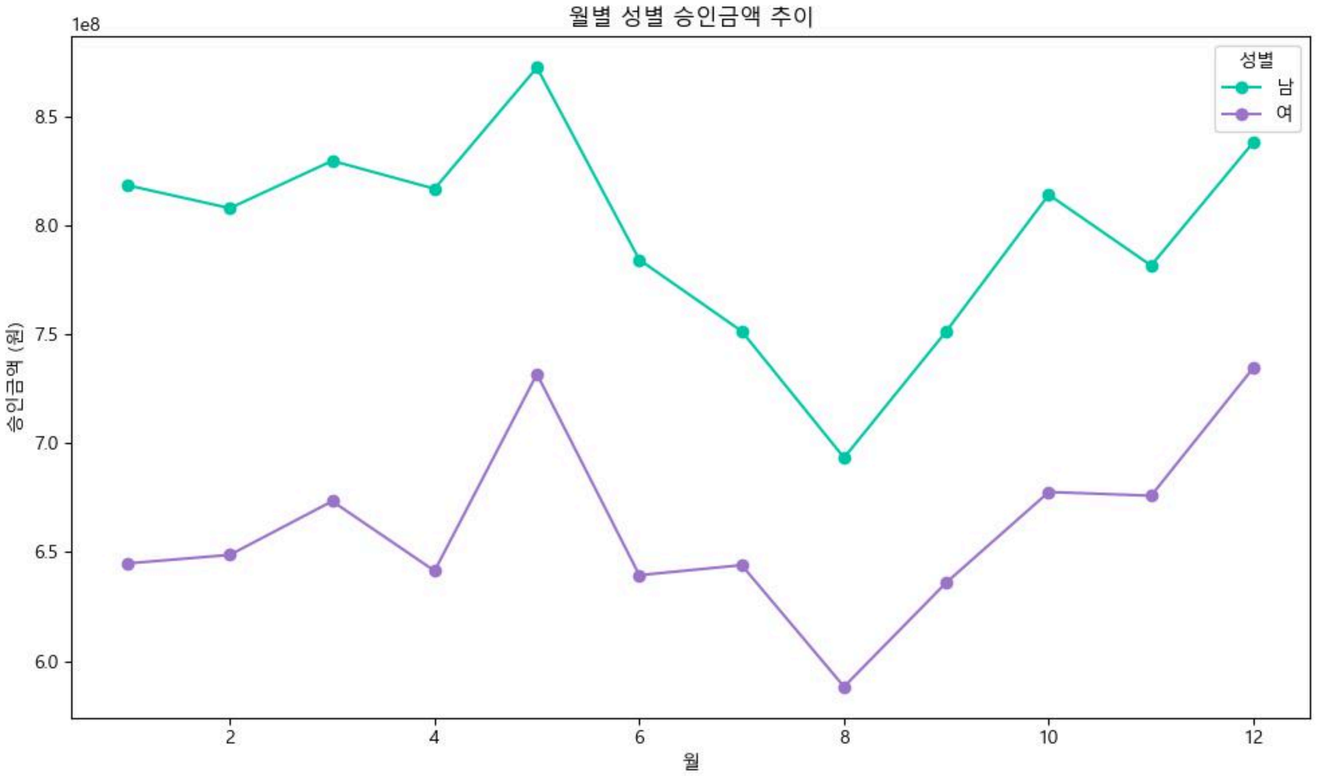
- object형 데이터들을 category형으로 형변환
- '가맹점_시군구'의 null값들을 결측치 처리

EDA

iM뱅크 고객 생활패턴 파악

10대 미만 고객





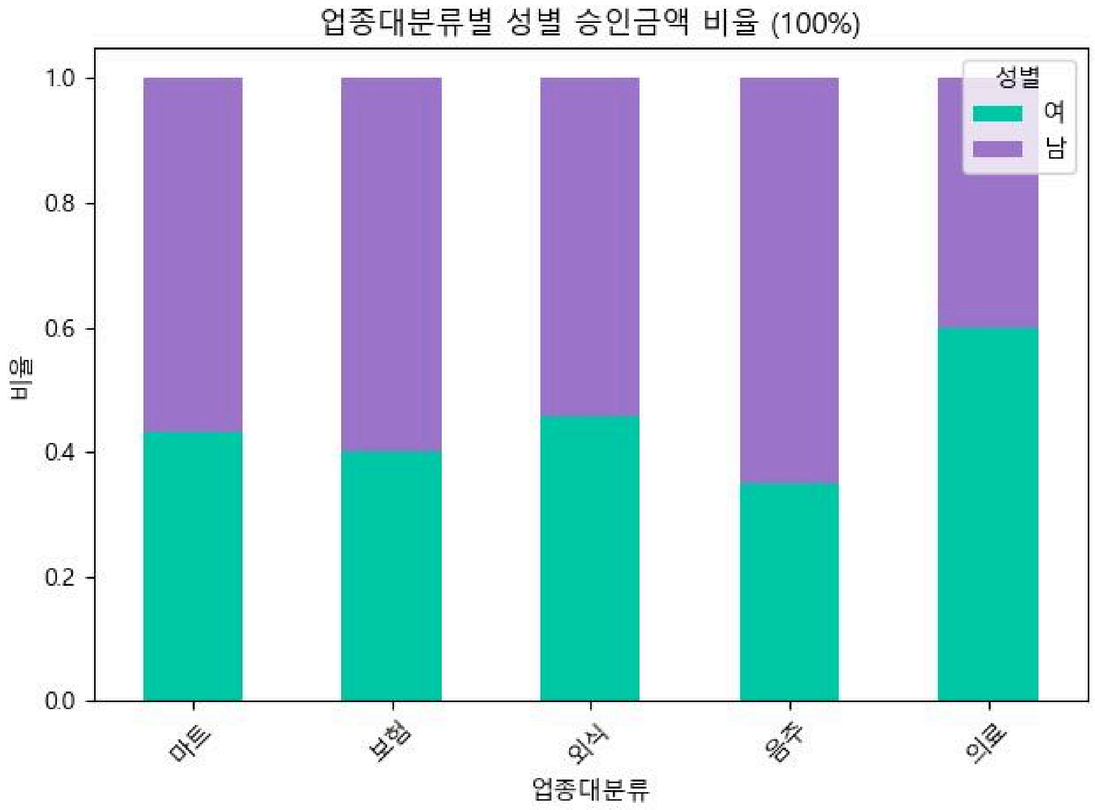
음주에서 남성의 비율이 가장 높음
보험에서 여성의 비율이 가장 높음

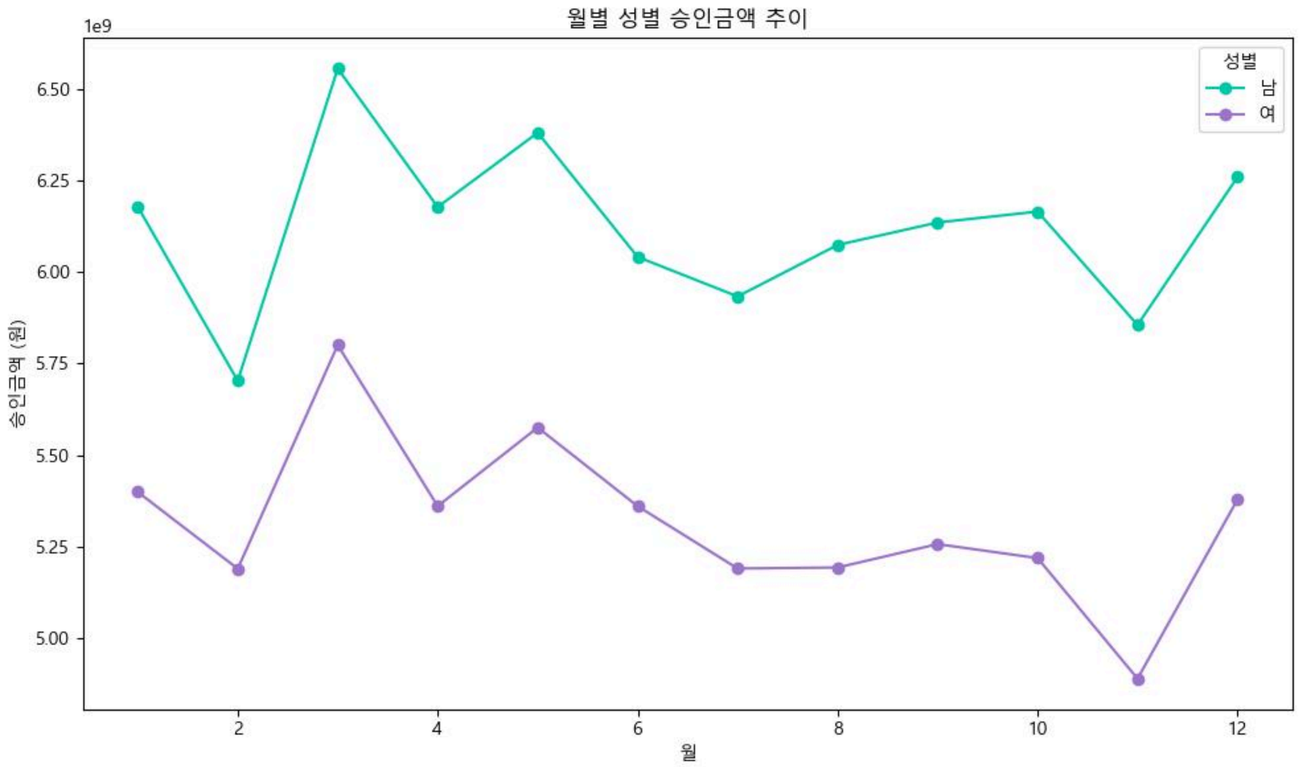
남성의 소비가 여성의 소비보다 높은 경향이 보임
둘다 8월에 소비가 가장 낮음

EDA

iM뱅크 고객 생활패턴 파악

20대 고객





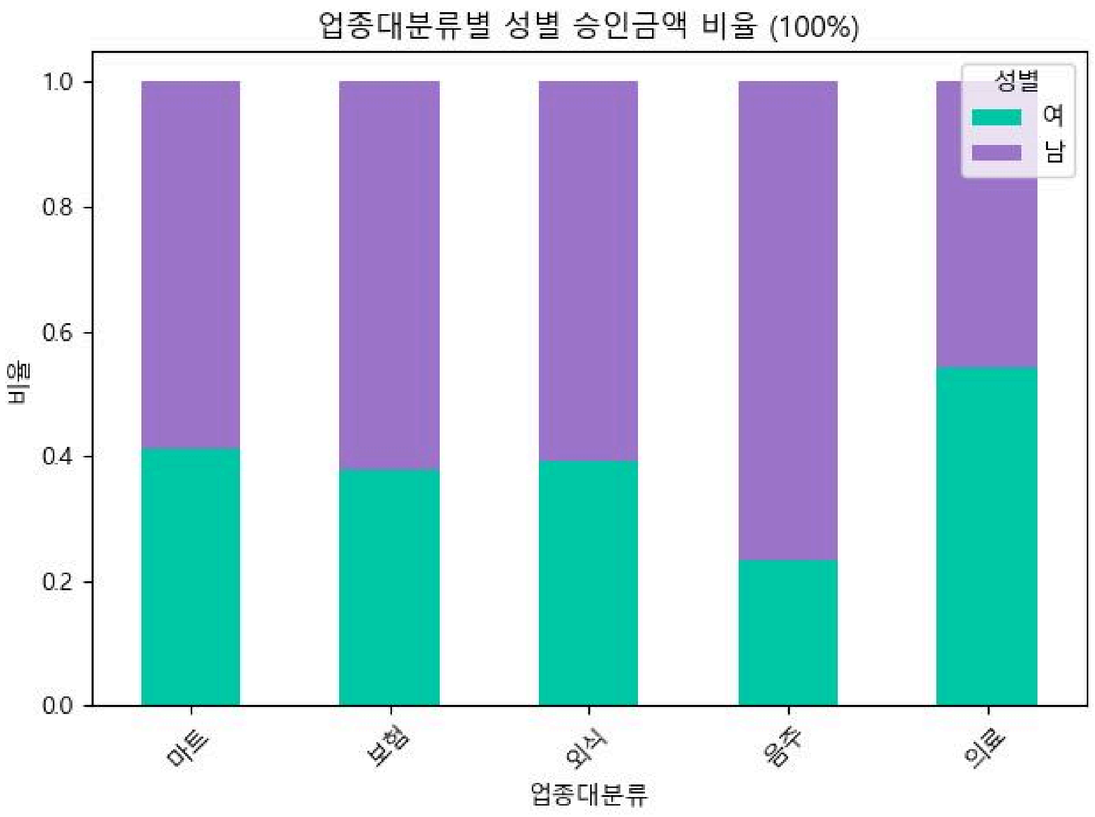
음주에서 남성의 비율이 가장 높음
의료에서 여성의 비율이 가장 높음

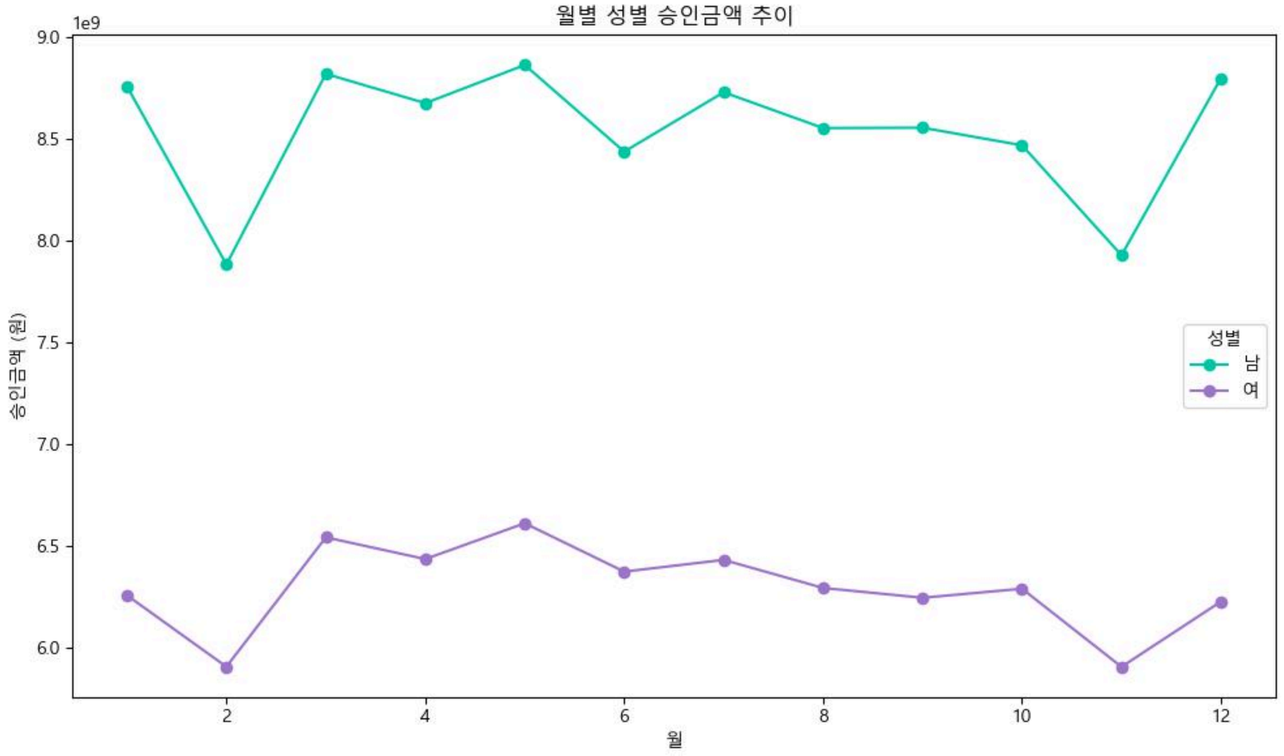
남성의 소비가 여성의 소비보다 높은 경향이 보임
둘다 3월에 소비가 가장 높고, 여성은 11월에 소비가 가장 낮음

EDA

iM뱅크 고객 생활패턴 파악

30대 고객





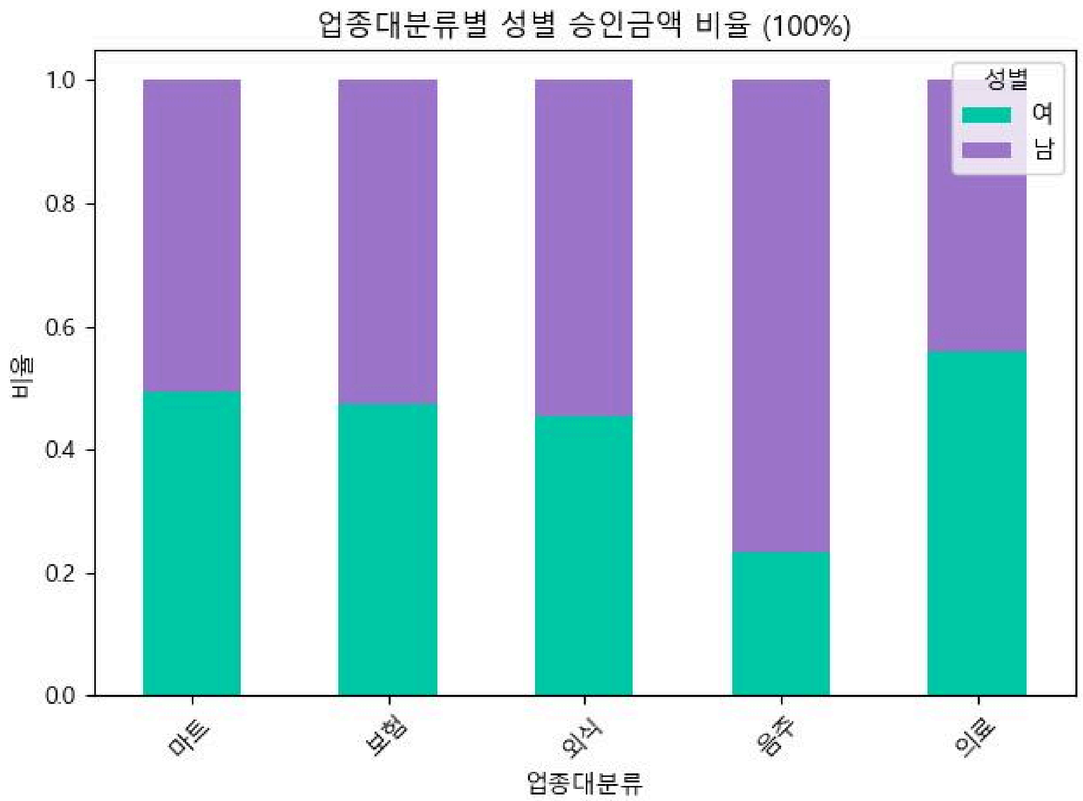
음주에서 남성의 비율이 가장 높음
의료에서 여성의 비율이 가장 높음

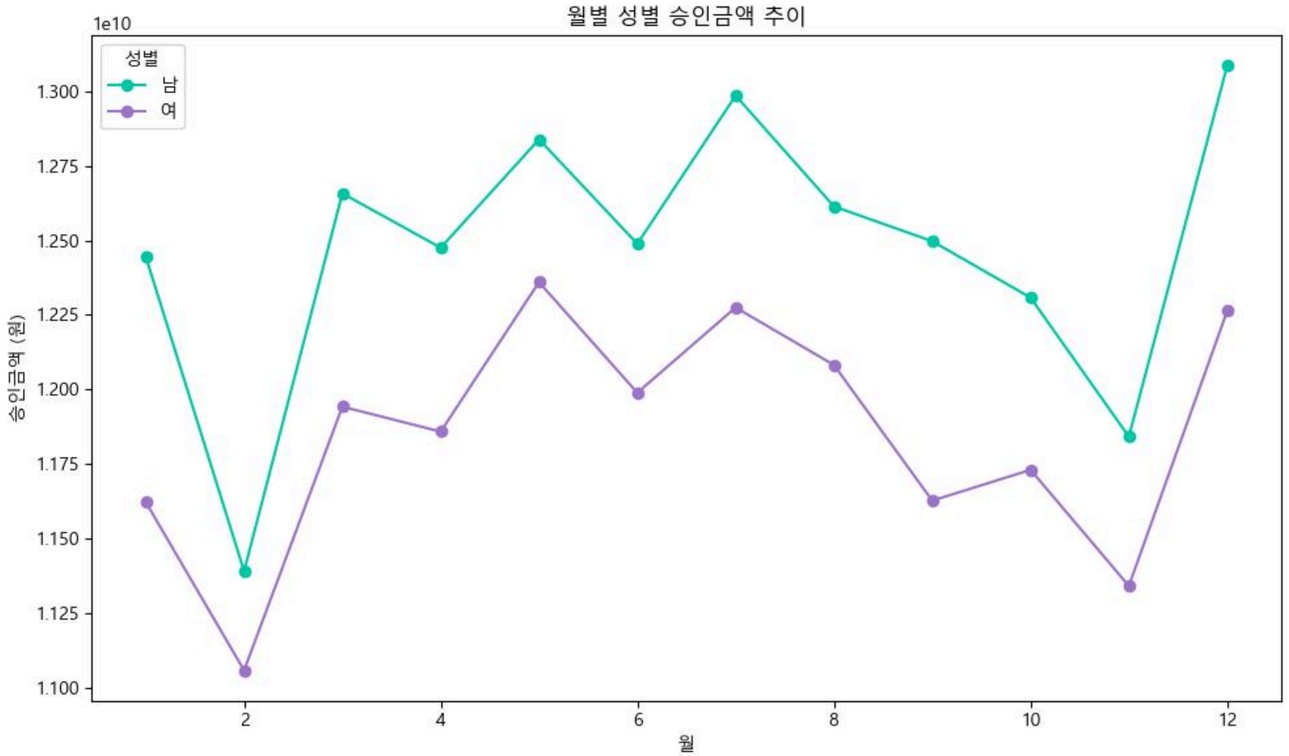
남성의 소비가 여성의 소비보다 높은 경향이 보임
둘다 2월, 11월에 소비가 눈에 띄게 감소함

EDA

iM뱅크 고객 생활패턴 파악

40대 고객





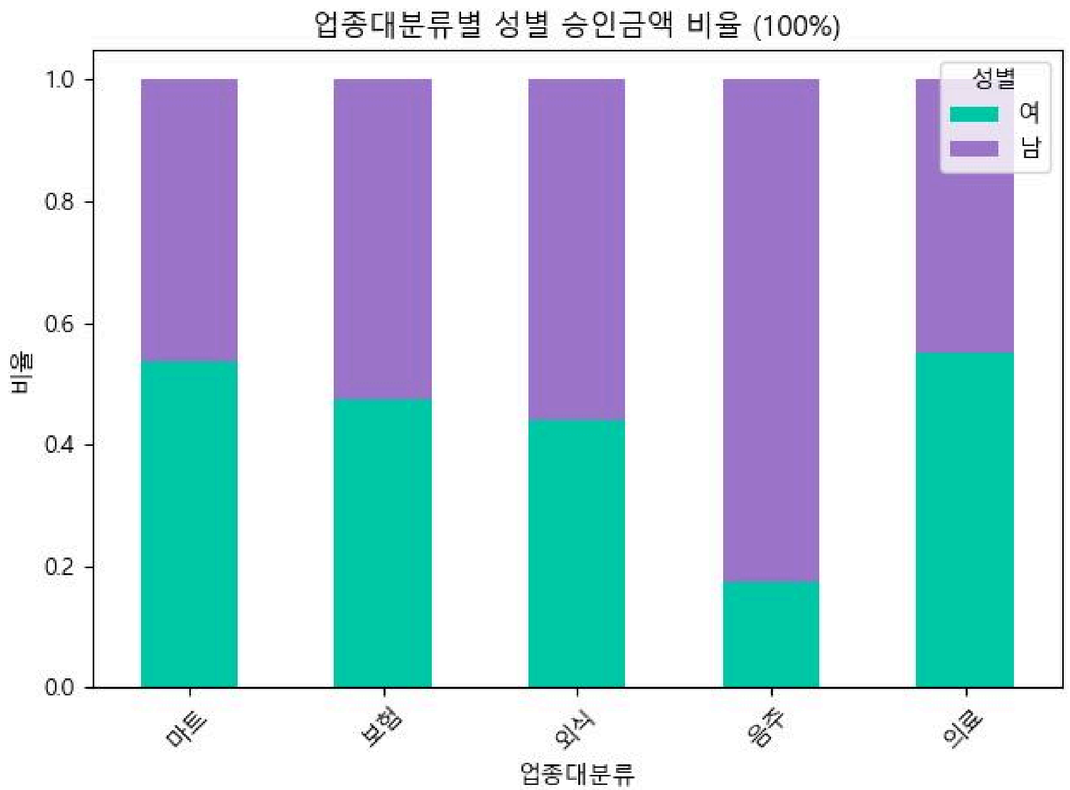
보험, 외식, 음주에서 남성의 비율이 높음
의료에서 여성의 비율이 가장 높음

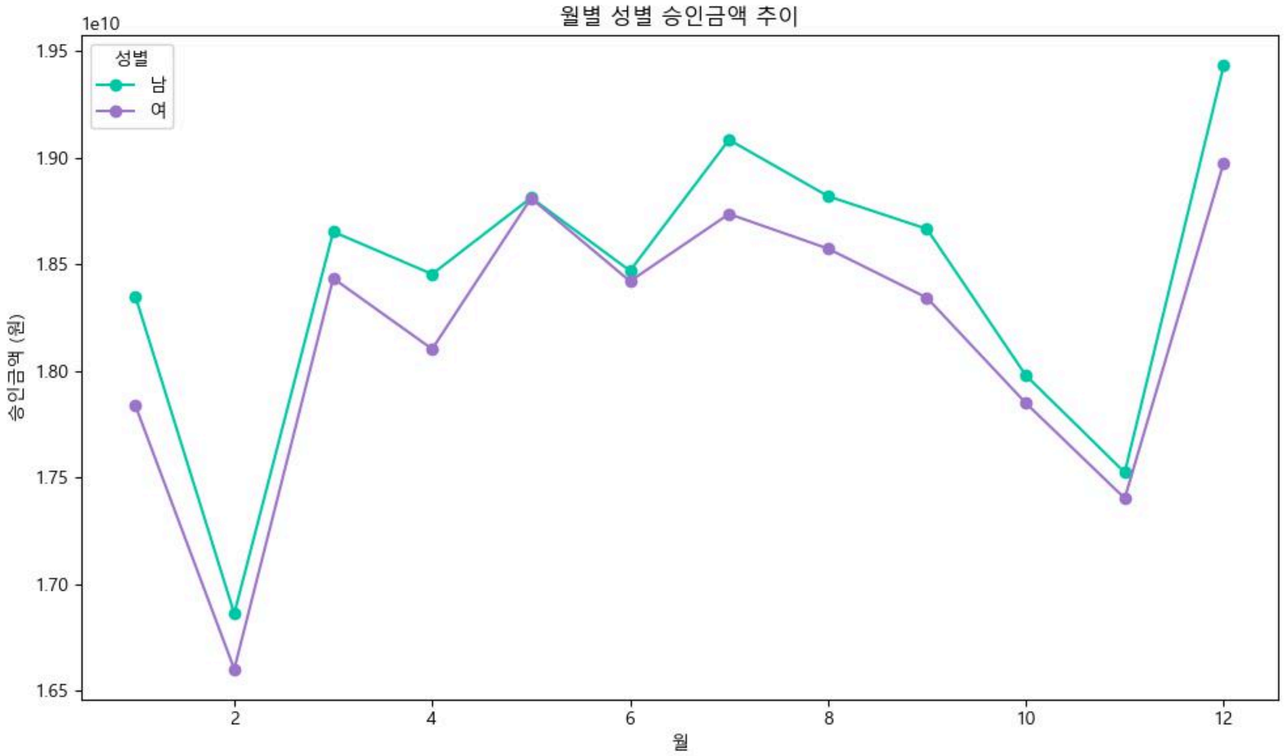
남성의 소비가 여성의 소비보다 높은 경향이 보임
둘다 2월에 소비가 가장 낮고, 12월에 다시 증가하는 경향이 보임

EDA

iM뱅크 고객 생활패턴 파악

50대 고객





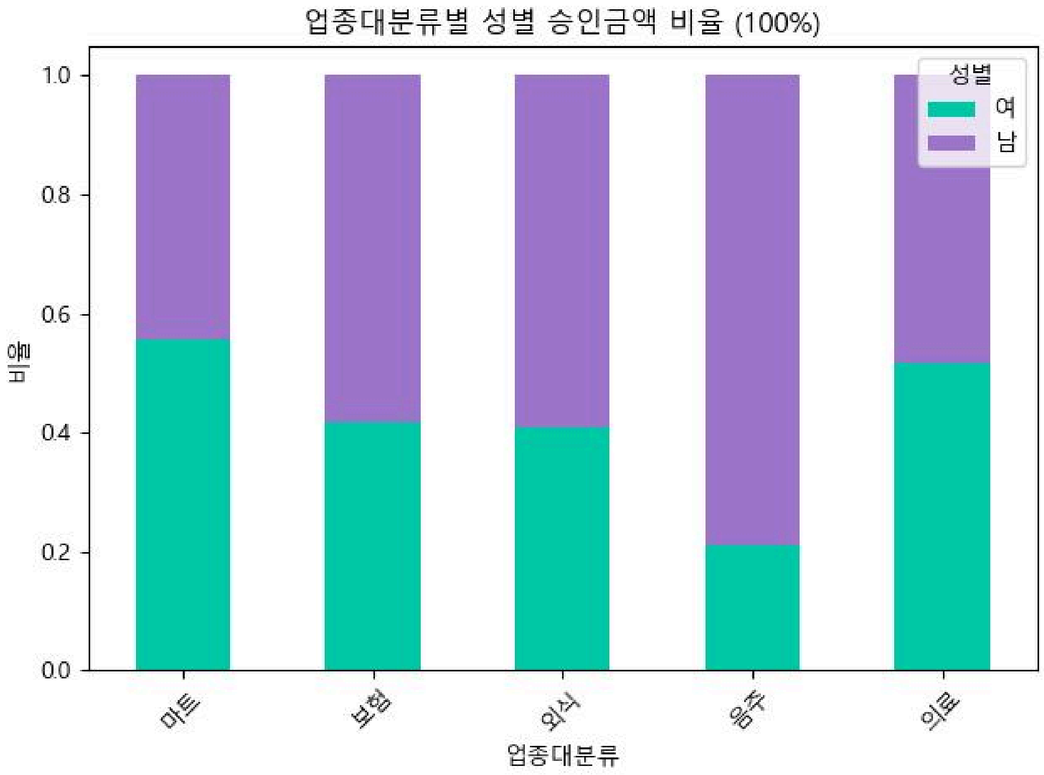
보험, 외식, 음주에서 남성의 비율이 높음
마트, 의료에서 여성의 비율이 높음

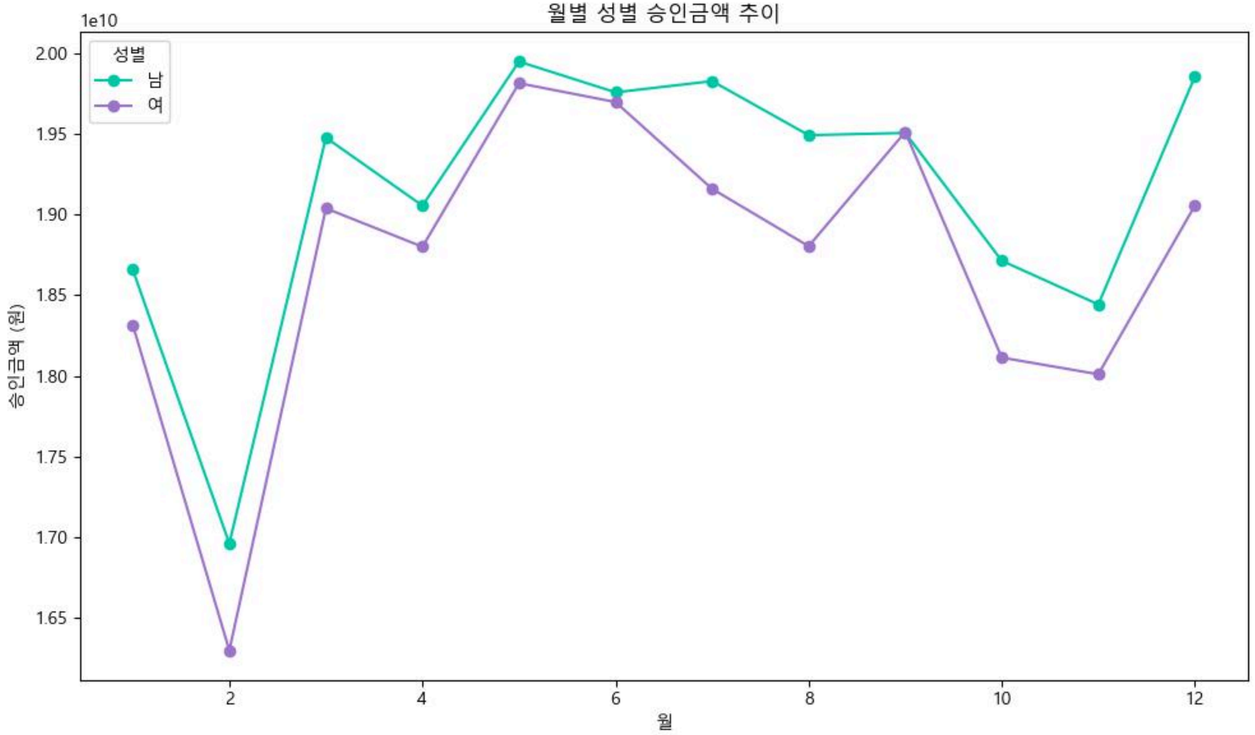
둘다 2월에 소비가 가장 낮음
5월, 12월에 특히 여성의 소비가 증가하는 경향이 보임

EDA

iM뱅크 고객 생활패턴 파악

60대 이상 고객





보험, 외식, 음주에서 남성의 비율이 높음
마트, 의료에서 여성의 비율이 높음

둘다 2월에 소비가 가장 낮고,
12월에 다시 증가하는 경향이 보임

보험 추천에 iM라이프 상품을 넣지 않은 이유

건강검진 데이터 군집 분석 결과

보험 추천에 iM라이프 상품을 넣지 않은 이유

건강검진 데이터 군집 분석 결과

2030 여성

- **Cluster 0** — **건강군** - 한화레저스포츠 상해보험
- **Cluster 1** — **신장질환** - 신장질환
- **Cluster 2** — **간질환, 당뇨, 고혈압** - 삼성화재 마이헬스파트너

20~23세 남성

- **Cluster 0** — **간질환** - 삼성화재 마이헬스파트너
- **Cluster 1** — **건강군** - 한화레저스포츠 상해보험

보험 추천에 iM라이프 상품을 넣지 않은 이유

iM라이프 보험

iM 프리미엄건강보험 무배당 2404

3대질병(암, 뇌출혈, 급성심근경색증)에 대해 **종합적으로 집중 보장!**

✓ 한번 가입으로 최대 110세까지보장!
✓ 만기시까지 갱신없이 동일보험료 적용

보험기간 / 보험료 납입기간 / 가입나이 / 납입주기

보험종목	보험기간	납입기간	가입나이		납입주기
			남자	여자	
1종 (암보장형)	80세	3년, 5년, 7년, 10년, 15년, 20년	0세~55세	0세~60세	월납, 연납
	100세, 110세			0세~62세	
2종 (2대질병보장형)	80세	3년, 5년, 7년, 10년, 15년, 20년	0세~60세	0세~60세	
	100세, 110세		0세~68세	0세~70세	
3종 (3대질병보장형)	80세	3년, 5년, 7년, 10년, 15년, 20년	0세~53세	0세~60세	
	100세, 110세			0세~61세	

든든한인생치매보험 무배당 2504 (해약환급금 일부지급형)

경도치매부터 **단계별 진단비**와 **보험금 증액**으로 **든든하게 보장!**

✓ 주계약 및 해당특약 가입시, 최초 1회에 한함, 치매보장개시일 이후 지급

보험기간 / 납입기간 / 납입주기 / 가입나이

구분	보험기간	납입기간	납입주기	가입나이	
				남자	여자
주계약	100세 만기	5년납	월납	35 ~ 69세	30 ~ 69세
		7년납			
경도이상치매진단특약		5년납		35 ~ 60세	30 ~ 62세
		7년납		35 ~ 58세	30 ~ 60세
중등도이상치매진단특약		5년납		35 ~ 69세	30 ~ 69세
		7년납		35 ~ 67세	
급여치매및뇌혈관질환검사비특약		5년납		35 ~ 69세	30 ~ 69세
		7년납			

iM 프리미엄안심보험 무배당 2504 (해약환급금 일부지급형(50%형))

완치가 없는 **오랜 투병**에도 **생활비 걱정없이!**

가벼운 경도치매부터 중등도, 중증치매까지 진단자금 최대 2,000만원 지급!

기준 : 1종(치매보장형), 주계약 보험가입금액 1,000만원

CDR척도 1점	• 경도치매 최종 진단확정 시 진단자금 300만원 (최초 1회한)
CDR척도 2점	• 중등도치매 최종 진단확정 시 진단자금 1,000만원 (최초 1회한) ※ 다만, "경도치매 진단자금"이 지급된 후에 "중등도치매상태"로 최종 진단확정 시 상기금액에서 "경도치매 진단자금"을 뺀 차액을 지급
CDR척도 3점이상	• 중증치매 최종 진단확정 시 진단자금 2,000만원 (최초 1회한) ※ 다만, "경도치매 진단자금" 또는 "중등도치매 진단자금"이 지급된 후에 "중증치매상태"로 최종 진단확정시 상기금액에서 이미 지급받은 진단자금의 총 합계액을 뺀 차액을 지급

참고문헌

Quick description about the section.

- 치주질환이 있는 사람의 폐암 위험이 1.5~2배 증가한다고 보고됨

예: "Periodontal disease and risk of lung cancer" (Annals of Oncology, 2008)

- 최영아, 「국민건강보험공단 건강검진 빅데이터를 활용한 머신러닝 기반 고비용 지출 예측 모델 개발 및 평가」, 경희대학교 대학원, 2022.08

앱 화면

Quick description about the section.

[부록]

3. 2030여성의 건강군집 특성표

항목	Cluster 0	Cluster 1	Cluster 2
연령대	100% 20대	100% 30대	30% 20대, 70% 30대
간질환 없음	**95.9%**	95.1%	**65.7%** (↓)
신장기능 정상	**89.0%**	70.0%	75.5%
빈혈 없음	90.3%	87.9%	**92.4%**
정상 혈압	**80.4%**	80.2%	**24.5%** (↓)
고혈압 전단계 이상	19.6%	19.8%	**75.6%**
정상 혈당	**91.8%**	89.1%	**63.5%**
당뇨 의심	0.2%	0.2%	**6.0%**
공복혈당장애	8.1%	10.7%	**30.4%**
비만 (BMI 4~6)	8.6%	6.8%	**84.6%**
흡연자(1번)	87.9%	90.5%	82.2%
음주율	84.5%	77.1%	79.8%

4. 20~23세 남성의 건강군집 특성표

항목	Cluster 0 (위험군)	Cluster 1 (건강군)
체중/복부비만	**과체중~고도비만**	**정상 체중**
혈압	**고혈압 경계~의심 많음**	대부분 정상
당뇨	당뇨 경계군 존재	대부분 정상 혈당
간 수치	**ALT, AST, 감마-GTP 높음**	정상 범위
흡연/음주	비율 유사하나 위험군은 흡연자 더 많음	상대적으로 덜함
BMI/허리둘레 분포	고도비만 집중	정상~과체중 중심
신장/빈혈 상태	신장기능 저하 및 빈혈 더 많음	안정적
의료개입 필요도	**조기 개입 시급**	**생활습관 유지 권장**

3. 2030여성의 건강군집 해석(군집 차이점 위주)

- **생활습관 부문**

→ 건강집단0은 비음주자가 85%에 달함
→ 건강집단&신장질환 집단은 음주율과 BMI가 건강집단0보다는 더하고 비건강집단보다는 덜함
→ 비건강집단은 전부 비만임

[음주여부 vs 성별연령대 교차비율 (%)]

음주여부	0.000000	1.000000
cluster		
0	15.5%	84.5%
1	22.9%	77.1%
2	20.2%	79.8%

[BMI구분 vs 성별연령대 교차비율 (%)]

BMI구분	1	2	3	4	5	6
cluster						
0	20.0%	59.3%	12.1%	8.3%	0.3%	0.0%
1	16.8%	63.1%	13.2%	6.7%	0.1%	0.0%
2	0.3%	5.2%	9.8%	48.3%	24.2%	12.1%

- **질병 부문**

→ 건강집단&신장딜환은 신장질환 외 질병X
→ 비건강집단은 당뇨, 간, 고혈압 다방면으로 다른 군집에 비해 건강이 좋지 못함

[고혈압군 vs 성별연령대 교차비율 (%)]

고혈압군	고혈압 의심	고혈압 전단계	정상 혈압
cluster			
0	0.5%	19.1%	80.4%
1	0.6%	19.2%	80.2%
2	16.0%	59.6%	24.5%

[당뇨군 vs 성별연령대 교차비율 (%)]

당뇨군	공복혈당장애 의심	당뇨병 의심	정상혈당
cluster			
0	8.1%	0.2%	91.8%
1	10.7%	0.2%	89.1%
2	30.4%	6.0%	63.5%

[간질환군 vs 성별연령대 교차비율 (%)]

간질환군	0	1
cluster		
0	95.9%	4.1%
1	95.1%	4.9%
2	65.7%	34.3%

데이터 설명: 건강검진 정보

```
허리둘레는 결측치가 411개 있습니다. 약 0.041%입니다.
시력(좌)는 결측치가 184개 있습니다. 약 0.018%입니다.
시력(우)는 결측치가 177개 있습니다. 약 0.018%입니다.
청력(좌)는 결측치가 137개 있습니다. 약 0.014%입니다.
청력(우)는 결측치가 138개 있습니다. 약 0.014%입니다.
수축기혈압는 결측치가 5747개 있습니다. 약 0.575%입니다.
이완기혈압는 결측치가 5747개 있습니다. 약 0.578%입니다.
식전혈당(공복혈당)는 결측치가 5814개 있습니다. 약 0.585%입니다.
혈색소는 결측치가 5817개 있습니다. 약 0.586%입니다.
요단백는 결측치가 11156개 있습니다. 약 1.123%입니다.
혈청크레아티닌는 결측치가 5814개 있습니다. 약 0.588%입니다.
혈청지오티(AST)는 결측치가 5816개 있습니다. 약 0.589%입니다.
혈청지피티(ALT)는 결측치가 5816개 있습니다. 약 0.589%입니다.
감마지티피는 결측치가 5813개 있습니다. 약 0.588%입니다.
흡연상태는 결측치가 89개 있습니다. 약 0.009%입니다.
음주여부는 결측치가 56개 있습니다. 약 0.006%입니다.

총콜레스테롤는 결측치가 651506개 있습니다. 약 65.941%입니다.
트리글리세라이드는 결측치가 651506개 있습니다. 약 65.941%입니다.
HDL콜레스테롤는 결측치가 651506개 있습니다. 약 65.941%입니다.
LDL콜레스테롤는 결측치가 657334개 있습니다. 약 66.531%입니다.
치아우식증유무는 결측치가 643197개 있습니다. 약 65.100%입니다.
치석는 결측치가 643197개 있습니다. 약 65.100%입니다.

치아우식증유무는 결측치가 225173개 있습니다. 약 68.095%입니다.
치석는 결측치가 225173개 있습니다. 약 68.095%입니다.
```

결측치와 이상치 처리

📌결손치 유무는 전부 결측치이므로 컬럼을 제거합니다.
📌치아마모증유무는 전부 결측치이므로 컬럼을 제거합니다.
📌제3대구치(사랑니) 이상는 전부 결측치이므로 컬럼을 제거합니다.

- 콜레스테롤 측정된 사람과 측정 안 된 사람의 차이 → 가설 : 콜레스테롤 측정한 사람들은 1차 건강검진에서 문제가 생겨서 2차 검진을 시행한 것이다.
- 치아우식증유무, 치석은 콜레스테롤 먼저 처리 후, 재검토

- 콜레스테롤 측정, 미측정 그룹이 다음의 이유로 차이없어서 결측치 전부 드롭
  ✅성별 외 거의 모든 변수에서 차이 없음
  ✅연령, 혈압, 혈당, 간수치, 비슷
  ✅지표로 유추할 수 있는 건강 수준 유사

→ 검진시설의 차이로 미측정 추정

- 치아우식증유무, 치석은 치주질환과 연관
- 치주질환이 있는 사람의 폐암 위험이 1.5~2배 증가한다고 보고됨

예: "Periodontal disease and risk of lung cancer"
(Annals of Oncology, 2008)

- 치주염 환자는 그렇지 않은 사람보다 폐암에 걸릴 확률이 유의하게 높았음

→ 치아우식증유무, 치석 정보가 있는 데이터만 선택

데이터 설명: 건강검진 정보

- 대한신장학회 근거
- 사구체여과율 계산식

단계	사구체여과율	특징	증상	자료
정상 혹은 1 단계	분당 90ml이상	신장 기능은 정상임. 이 경우 혈뇨, 단백뇨 등 소변검사 이상이 없을 경우 정상임. 하지만 혈뇨, 단백뇨 등 초기 신장손상의 증거가 있는 경우에는 사구체여과율이 정상이라도 만성신장질환 1단계에 해당될 수 있음	무증상	혈뇨, 단백뇨 여부를 체크하고, 이상이 있을 때는 원인을 찾아서 교정
2 단계	분당 60ml이상 ~ 90ml이하	신장기능이 감소하기 시작	무증상 BUN, 크레아티닌 등 혈액 검사수치 이상이 나타남	혈압조절 원인을 치료
3 단계	분당 30ml이상 ~ 59ml이하	신기능이 더욱 감소	피로 식욕감소,가려움증이 더욱악화	혈압 조절, 신장기능 악화를 늦추기 위한치료
4 단계	분당 15ml이상 ~ 29ml이하	생명유지에 필요한 신장의 기능을 겨우 유지	피로,식욕감소, 가려움증이 더욱 악화	투석 준비, 이식 가능성에 대해서도 준비
5 단계	분당 15ml미만	신장기능이 심각하게 손상되어 투석이나 이식없이는 생명을 유지하기 어려움	수면 장애,호흡곤란,가려움,구토	투석 또는 이식을 시행받아야 함

- 질병관리청 근거로 함수 만듦
https://health.kdca.go.kr/healthinfo/biz/health/ntcnInfo/healthSourc/thtimtCntnts/thtimtCntntsView.do?thtimt_cntnts_sn=7

건강검진으로 판단 가능한 질병

출처: 질병관리청 국가건강정보포털

신장 질환 판단하는 사구체여과율

질병관리청을 근거로 함수로 만든 대한신장학회사 구체여과율 계산식