

# 주제: Gamma-GTP 이상 수치 분석



Gamma-GTP 수치의 예측 모델을 통해

해당 예측값이 비정상 범주에 들어간 사람들을 선별하여 주의가 필요함을 강조



## 팀 소개

[통계마법사] 서준형, 이해승, 이시은, 정다연

## 팀 리소스

Health Checkup Result

→ 통최전 안내 페이지

# 목차

- 데이터 전처리
- 데이터 시각화
- 변수 재생성
- 특정 피처별 GAMMA\_GTP 비교분석
- GAMMA\_GTP 이상 수치 예측
- 결론

## 데이터 전처리

1. 2013 이전 데이터 AGE\_GROUP+4해서 2013 이후 데이터와 AGE\_GROUP 통일

2002~2013

AGE_GROUP	REAL_AGE	AGE_GROUP	REAL_AGE
1	20~24	8	55~59
2	25~29	9	60~64
3	30~34	10	65~69
4	35~39	11	70~74
5	40~44	12	75~79
6	45~49	13	80~84
7	50~54	14	85 +

2014~

AGE_GROUP	REAL_AGE	AGE_GROUP	REAL_AGE
1	0~4	10	45~49
2	5~9	11	50~54
3	10~14	12	55~59
4	15~19	13	60~64
5	20~24	14	65~69
6	25~29	15	70~74
7	30~34	16	75~79
8	35~39	17	80~84
9	40~44	18	85+

2. 중복된 IDV\_ID 제거

3. 이상치 처리 ⇒ 하단의 첨부한 표와 동일한 내용

- weight : 165 -> Nan
- waist : 999 -> Nan
- TOT\_CHOLE : 1000 이상 -> Nan
- TRIGLYCERIDE : 2000 이상 -> Nan
- HDL\_CHOLE : 8000 이상 -> Nan
- LDL\_CHOLE : 2000 이상 -> Nan
- HMG : 1.6 -> Nan
- CREATININE : 40 이상 -> Nan
- SGOT\_AST : 700 이상 -> Nan
- SGPT\_ALT : 700 이상 -> Nan
- GAMMA\_GTP : 999 -> Nan

4. 'SMK\_STAT', 'DRK\_YN', 'HCHK\_CE\_IN', 'CRS\_YN', 'TTR\_YN' (1,0으로 통일)

5. TTR\_YN의 2 값들 nan으로 대체

4. 결측치 처리

- HEIGHT, WEIGHT, WAIST, SIGHT\_LEFT, SIGHT\_RIGHT, TOT\_CHOLE, TRIGLYCERIDE, HDL\_CHOLE, LDL\_CHOLE : 성별, 연령별 중앙값으로 채우기
- HEAR\_LEFT, HEAR\_RIGHT, SMK\_STAT, DRK\_YN, HCHK\_CE\_IN, CRS\_YN, TTR\_YN : 성별, 연령별 최빈값으로 채우기
- BP\_HIGH, BP\_LWST, BLDS, HMG, OLIG\_PROTE\_CD, CREATININE, SGOT\_AST, SGPT\_ALT, GAMMA\_GTP : 성별, 연령별, 지역별 중앙값으로 채우기

→ 해당 표의 값들은 검사자의 측정치에 대한 내용.

Feature		value	전처리	결측값
GAMMA_GTP	간 기능	<정상 값> 남성: 11-64 IU/L 여성: 8-35 IU/L	[이상치] 999 → Nan	(중앙값) 성별&연 령별&지역별
SGPT_ALT	간 기능	<정상 값> 0-40 IU/L	[이상치] 700 이상 → Nan	(중앙값) 성별&연 령별&지역별
SGOT_AST	간 기능	<정상 값> 0-40 IU/L	[이상치] 700 이상 → Nan	(중앙값) 성별&연 령별&지역별

YEAR	정보의 기준 연도			
IDV_ID	개인 식별 번호			제거
SEX	성별	1 : 남성 2 : 여성		
AGE_GROUP	연령대 코드	나이 관련 정보는 위의 표 참고(5세 단위)		그룹핑 (하단 참 고)
AREA_CODE	거주지 코드			통일 후 그룹핑 (하단 참고)
HEIGHT	키	5cm 단위		(중앙값) 성별&연 령별
WEIGHT	몸무게	5kg 단위		(중앙값) 성별&연 령별
WAIST	허리 둘레			(중앙값) 성별&연 령별
SIGHT_LEFT	왼쪽 눈 시력	0.1-2.5, 시력 < 0.1 == 0.1, 시각 장애==9.9		(중앙값) 성별&연 령별
SIGHT_RIGHT	오른쪽 눈 시력	0.1-2.5, 시력 < 0.1 == 0.1, 시각 장애==9.9		(중앙값) 성별&연 령별
HEAR_LEFT	왼쪽 귀 청력	1 : 정상 2 : 이상		(최빈값) 성별&연 령별
HEAR_RIGHT	오른쪽 귀 청력	1 : 정상 2 : 이상		(최빈값) 성별&연 령별
BP_HIGH	수축기 혈압			(중앙값) 성별&연 령별&지역별
BP_LWST	이완기 혈압			(중앙값) 성별&연 령별&지역별
BLDS	공복 혈당	100ml 당 포도당 농도		(중앙값) 성별&연 령별&지역별
TOT_CHOLE	콜레스테롤	<정상 값> 150- 250 mg/dL	[이상치] 1000 이 상 -> Nan	(중앙값) 성별&연 령별
TRIGLYCERIDE	[지질] 간단한 지 질 or 중성 지질의 양	<정상 값> 30- 135 mg/dL	[이상치] 2000 이 상 -> Nan	(중앙값) 성별&연 령별
HDL_CHOLE	[콜레스테롤] HDL 에 포함된 콜레스 테롤 양	<정상 값> 30-65 mg/dL	[이상치] 8000 이 상 -> Nan	(중앙값) 성별&연 령별

LDL_CHOLE	[콜레스테롤] LDL 에 포함된 콜레스 테롤 양	170 mg/dL 이상 이면 고LDL콜레 스테롤증 진단	[이상치] 2000 이 상 -> Nan	(중앙값) 성별&연 령별
HMG	혈액 속의 피그먼 트 단백질 → 혈액 에서 산소 운반하 는 역할		[이상치] 1.6 -> Nan	(중앙값) 성별&연 령별&지역별
OLIG_PROTE_CD	요증의 단백질 배 설	1(-), 2(±), 3(+1), 4(+2), 5(+3), 6(+4)		(중앙값) 성별&연 령별&지역별
CREATININE	크레아티닌의 혈 중 농도	음식과 관련 X 근 육 발달과 운동과 관련 <정상 값> 0.8-1.7 mg/dL	[이상치] 40 이상 - > Nan	(중앙값) 성별&연 령별&지역별
SMK_STAT	흡연 여부	1 : 비흡연 2 : 이 전에 피웠지만 중 단 3 : 현재 흡연	(1,0으로 통일)	(최빈값) 성별&연 령별
DRK_YN	음주 여부	0 : 음주하지 않음 1 : 음주함	(1,0으로 통일)	(최빈값) 성별&연 령별
HCHK_CE_IN	구강 검진 여부	0 : 검사 안 함 1 : 검사항.	(1,0으로 통일)	(최빈값) 성별&연 령별
CRS_YN	치아 우식 여부	0 : 없음 1 : 있음	(1,0으로 통일)	제거
TTR_YN	치석 여부	0 : 없음 1 : 있음	(1,0으로 통일) ⇒ 2 인 값들 Nan으 로 대체	제거

### 간 기능 관련 피처들

- GAMMA\_GTP** : 간 기능을 나타내는 혈액 검사 수치. 감마 GTP는 주로 **간의 담관에 존재**하며 **담증이나 간세포 이상이 발생할 때 혈중 농도가 증가함**.
- SGPT\_ALT** : 간 기능을 나타내는 혈액 검사 수치. ALT는 주로 **간세포에만 존재**하며 간세포가 손상될 때 농도가 증가함.
- SGOT\_AST** : 간 기능을 나타내는 혈액 검사 수치. **간세포, 심장, 신장, 뇌 및 근육세포가 손상될 때 농도가 증가함**.

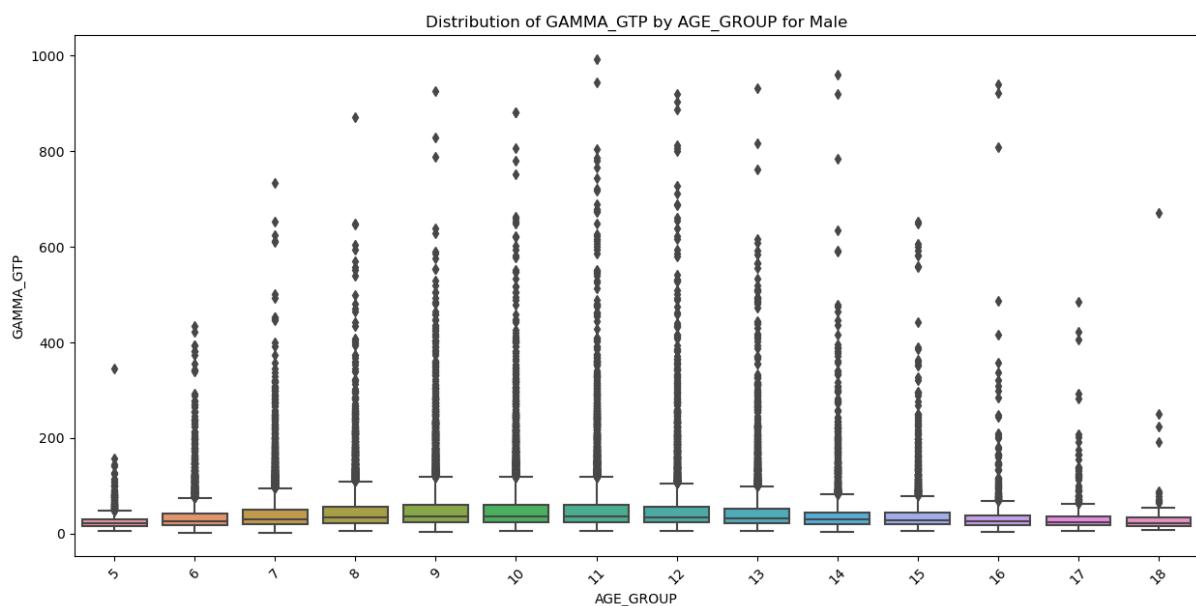
## 변수 재생성

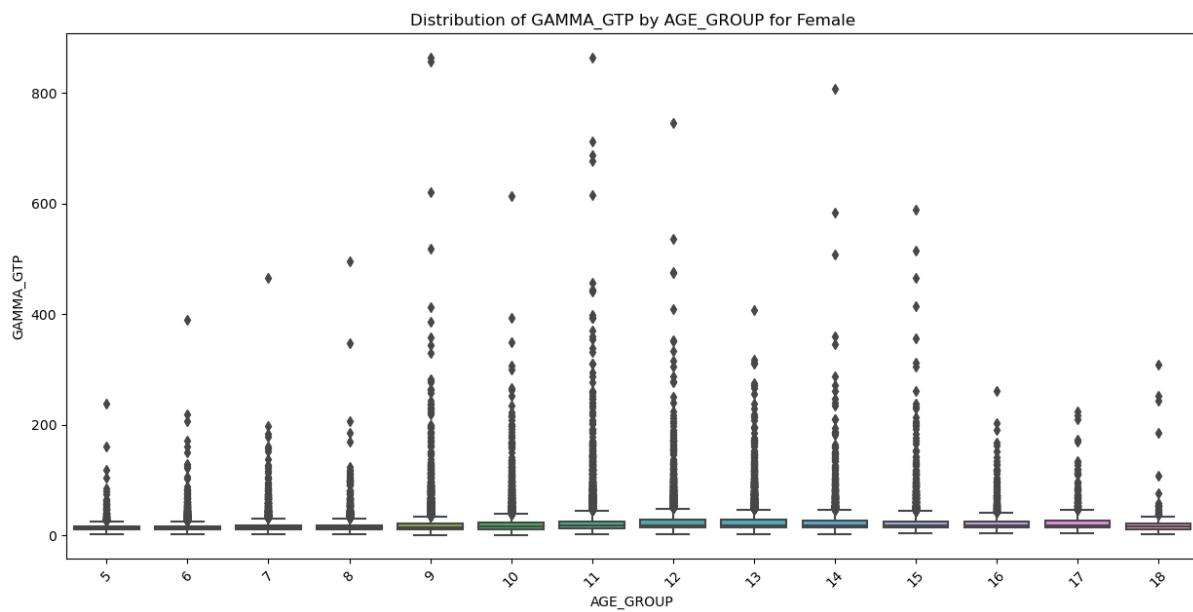
## 지역 그룹핑\_5개

AREA\_CODE 지역별로 나눔 5개의 그룹

- **Metropolitan** (대도시 그룹): 주요 대도시 지역 서울, 부산, 대구, 인천, 광주, 대전, 울산
- **Gyeonggi & Gangwon** (경기 & 강원 그룹): 경기도 및 강원도
- **Chungcheong & Sejong** (충청 & 세종 그룹): 충청남도, 충청북도 및 세종
- **Jeolla** (전라 그룹): 전라남도, 전라북도 및 제주도(제주도 데이터 적어서) -
- **Gyeongsang** (경상 그룹): 경상남도, 경상북도

## 나이 그룹핑\_5개 → GAMMA\_GTP 분포 확인





남성 (SEX=1):

20대 초반 - 60대 초반 : GAMMA\_GTP 값이 상승하는 경향을 보임

60대 중반부터는 GAMMA\_GTP 값이 감소하는 경향을 보임

- **YOUNG** 청년기: 20대 중반까지
- **Middle Age Early** 중년 전기: 20대 후반부터 40대 중반까지
- **Middle Age Late** 중년 후기: 40대 후반부터 50대 중반까지
- **Elderly Early** 노년 전기: 50대 후반부터 60대 중반까지
- **Elderly Late** 노년 후기: 60대 후반 이후

여성 (SEX=2):

20대 중반부터 60대 중반까지 GAMMA\_GTP 값이 상승하는 경향을 보임

70대부터는 GAMMA\_GTP 값이 감소하는 경향을 보임

- **YOUNG** 청년기: 20대까지 ~29
- **Middle Age Early** 중년 전기: 30대 와 40대 초반까지
- **Middle Age Late** 중년 후기: 40대 중반부터 50대까지
- **Elderly Early** 노년 전기: 60대까지
- **Elderly Late** 노년 후기: 70대 이후

## 원핫 인코딩

특정 피처들 범주형 데이터를 수치형 데이터로 변환

[SMK\_STAT, DRK\_YN, HCHK\_CE\_IN, AREA\_GROUP, AGE\_GROUP]

## GAMMA\_GTP 정상/비정상 비교분석

모델링 이전에 GAMMA\_GTP 데이터에 대한 분석 결과를 제시하며, 정상 범위와 비정상 범위 데이터 간의 비교 분석에 초점을 맞추고자 합니다. 이 분석은 데이터 자체의 특성을 이해하고, 정상과 비정상 범위 데이터의 차이를 확인하기 위한 것입니다.

GAMMA\_GTP는 간 기능을 나타내는 혈액 검사 수치입니다. GAMMA\_GTP는 주로 간의 담즙관에 존재하는 효소이며, 담즙 배출 장애나 간세포 장애가 발생할 때 혈중 농도가 증가합니다.

이러한 정보 기반으로 데이터를 분석하기 전에 생물학적접근으로 영향을 줄 것이라 예상 변수들을 나열해보았습니다.

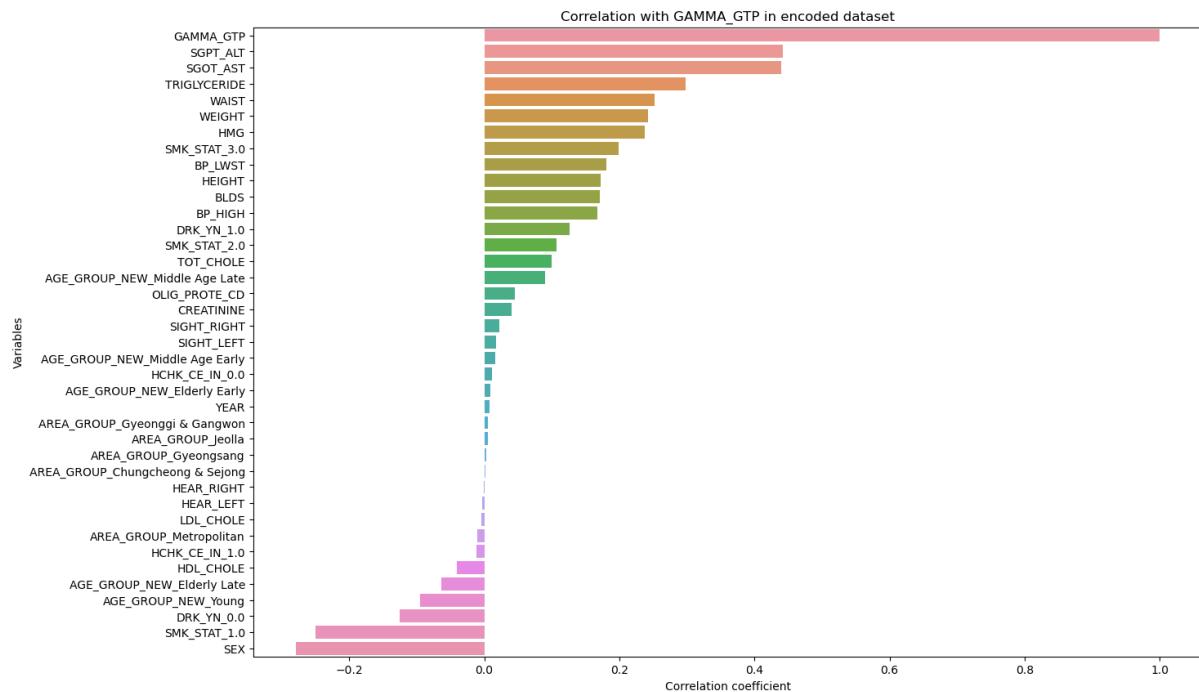
- SGOT\_AST - 간 기능을 나타내는 다른 중요한 표지자입니다. 간세포, 심장, 신장, 뇌 및 근육세포가 손상되면 농도가 증가합니다.
- SGPT\_ALT - ALT는 주로 간세포에만 존재하며, 간세포가 손상되면 농도가 증가합니다.
- TRIGLYCERIDE - 간은 중성지방의 주요 저장 및 생성 장소입니다. 중성지방의 수치는 간 기능과 관련이 있습니다.
- HDL\_CHOLE & LDL\_CHOLE & TOT\_CHOLE - 콜레스테롤은 간에서 생성되며, 간 기능의 상태에 따라 수치가 변동될 수 있습니다.
- CREATININE - 크레아티닌은 간에서 생성되며, 간과 신장 기능의 상태에 따라 수치가 변동될 수 있습니다.
- WAIST - 허리 둘레는 복부 비만의 지표로 간 기능 장애와 관련이 있을 수 있습니다.
- SMK\_STAT - 흡연은 간에 손상을 줄 수 있으므로, 흡연 상태는 GAMMA\_GTP 수치와 관련이 있을 수 있습니다.
- DRK\_YN - 음주는 간에 손상을 줄 수 있으므로, 음주 상태도 GAMMA\_GTP 수치와 관련이 있을 수 있습니다.

## GAMMA\_GTP 데이터 분석

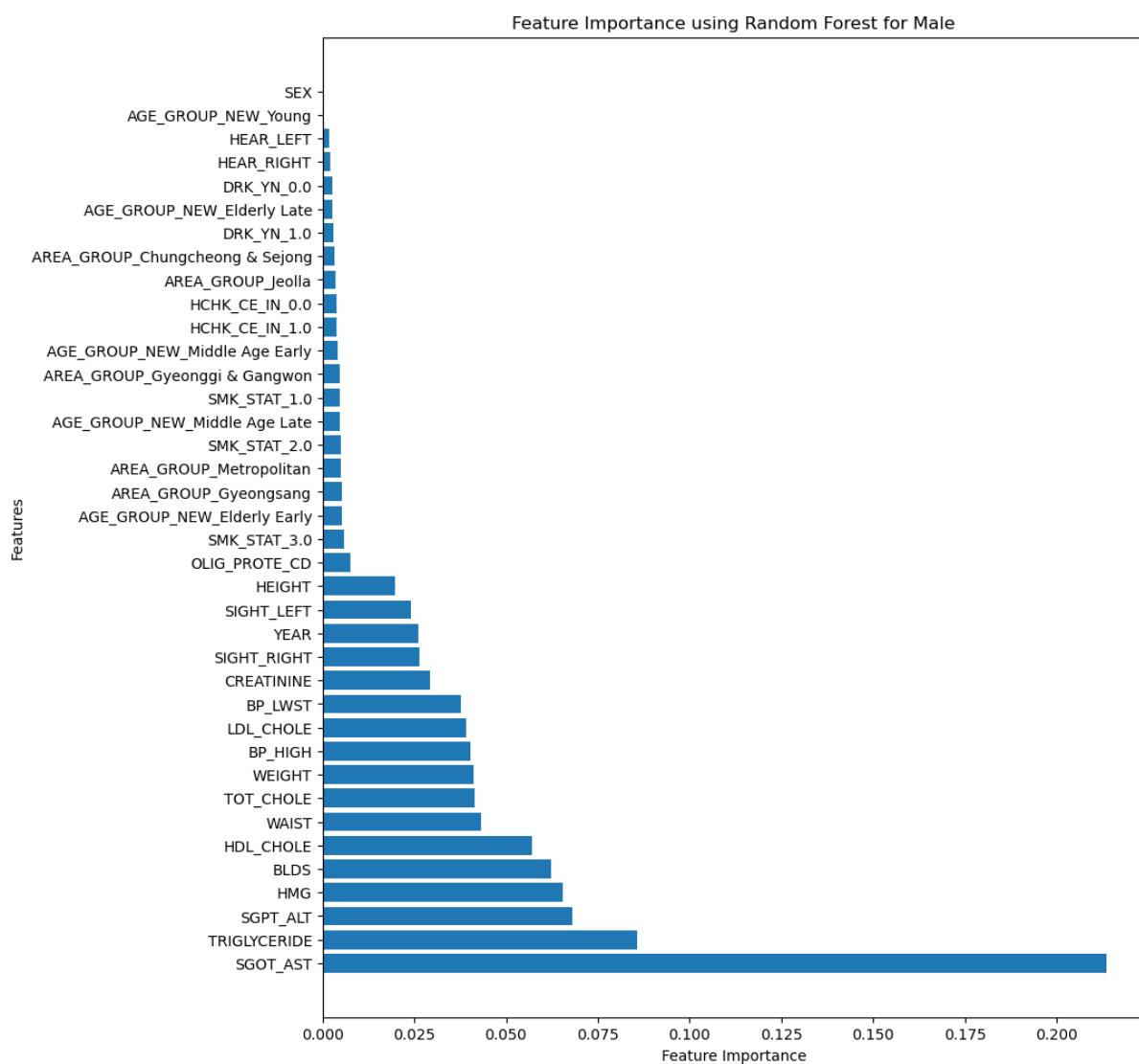
GAMMA\_GTP는 남,여에 따라 정상 수치 범위가 다르기에 남,여 데이터를 분리하여 비정상 범위에 속한 데이터를 추출하였습니다

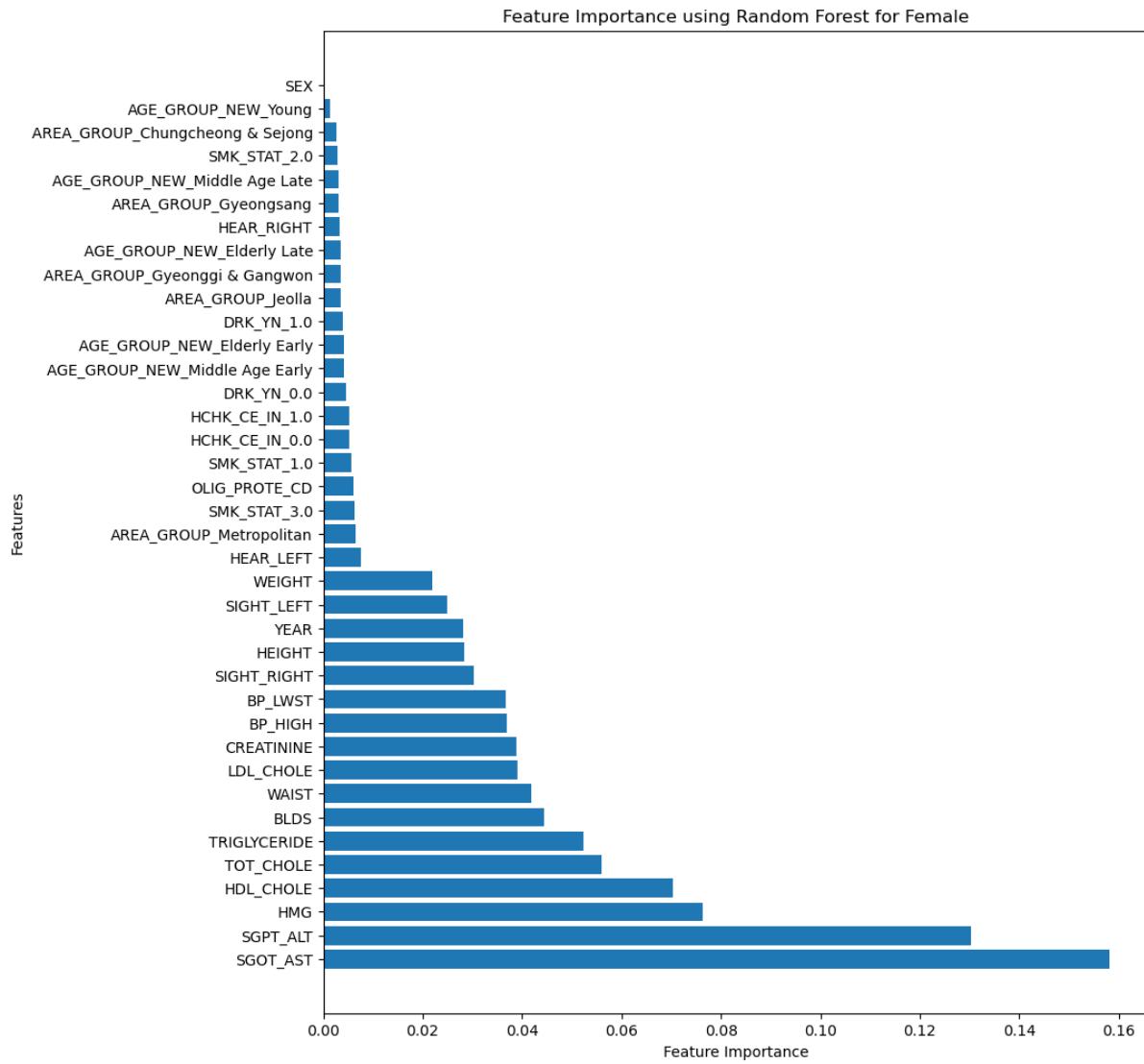
GAMMA\_GTP가 비정상인 데이터들의 다른 피처와의 상관관계 파악하고, RandomForestRegresoor를 사용하여 각 피처의 중요도를 확인하였습니다.

## 상관관계



## RandomForestRegresoor를 사용한 feature\_importance





GAMMA\_GTP가 비정상인 남,여 데이터 따로 분리하여서 구한 피처간의 Correlation 값과, RandomForestRegressor을 통해 구한 importance값을 토대로 둘의 합을 통해 변수들의 중요도의 순서를 매긴 남,여 각각 20개의 변수들만 선택하였습니다.

### 여성 데이터에서 GAMMA\_GTP가 비정상인 것에 영향을 가장 많이 주는 20개의 피처

- ['SGOT\_AST', 'SGPT\_ALT', 'TRIGLYCERIDE', 'BLDS', 'HMG', 'WAIST', 'TOT\_CHOLE', 'BP\_HIGH', 'BP\_LWST', 'AGE\_GROUP\_NEW\_Young', 'YEAR', 'SMK\_STAT\_1.0', 'SMK\_STAT\_3.0', 'HDL\_CHOLE', 'AGE\_GROUP\_NEW\_Elderly Early', 'HEIGHT', 'LDL\_CHOLE', 'WEIGHT', 'AGE\_GROUP\_NEW\_Middle Age Early', 'CREATININE']

### 남성 데이터에서 GAMMA\_GTP가 비정상인 것에 영향을 가장 많이 주는 20개의 피처

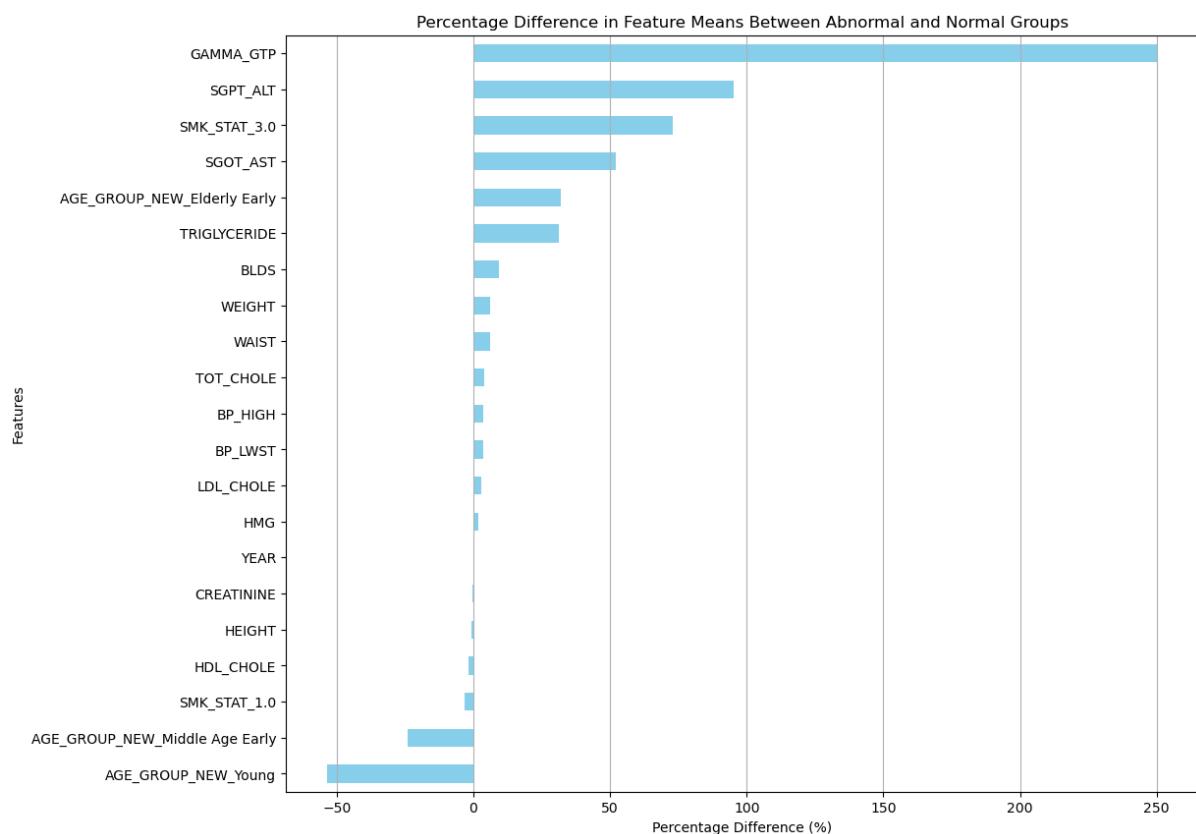
- ['SGOT\_AST', 'SGPT\_ALT', 'TRIGLYCERIDE', 'BLDS', 'BP\_HIGH', 'HDL\_CHOLE', 'BP\_LWST', 'HMG', 'TOT\_CHOLE', 'WEIGHT', 'LDL\_CHOLE', 'HEIGHT', 'SMK\_STAT\_1.0', 'SMK\_STAT\_3.0', 'AGE\_GROUP\_NEW\_Middle Age Early', 'AGE\_GROUP\_NEW\_Elderly Early']

'OLIG\_PROTE\_CD', 'WAIST', 'DRK\_YN\_1.0', 'DRK\_YN\_0.0',  
 'AGE\_GROUP\_NEW\_Young']

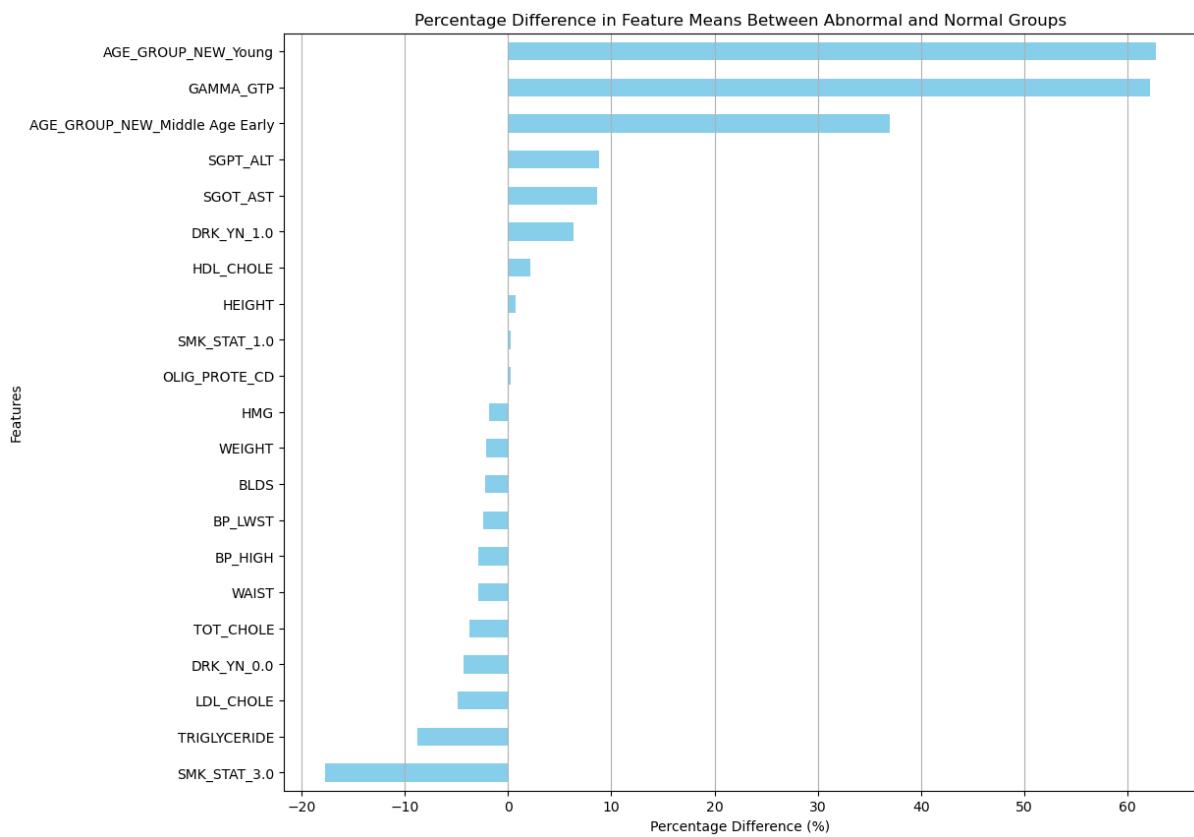
선택된 피처들은 GAMMA\_GTP가 비정상인 여성, 남성에게 영향을 많이 준다고 판단하였습니다

**GAMMAA\_GTP가 비정상인 그룹, 정상인 그룹으로 나눈 후 선택한 20개의 피쳐에 대하여  
 피처별 평균값의 차이를 계산**

여성



남성

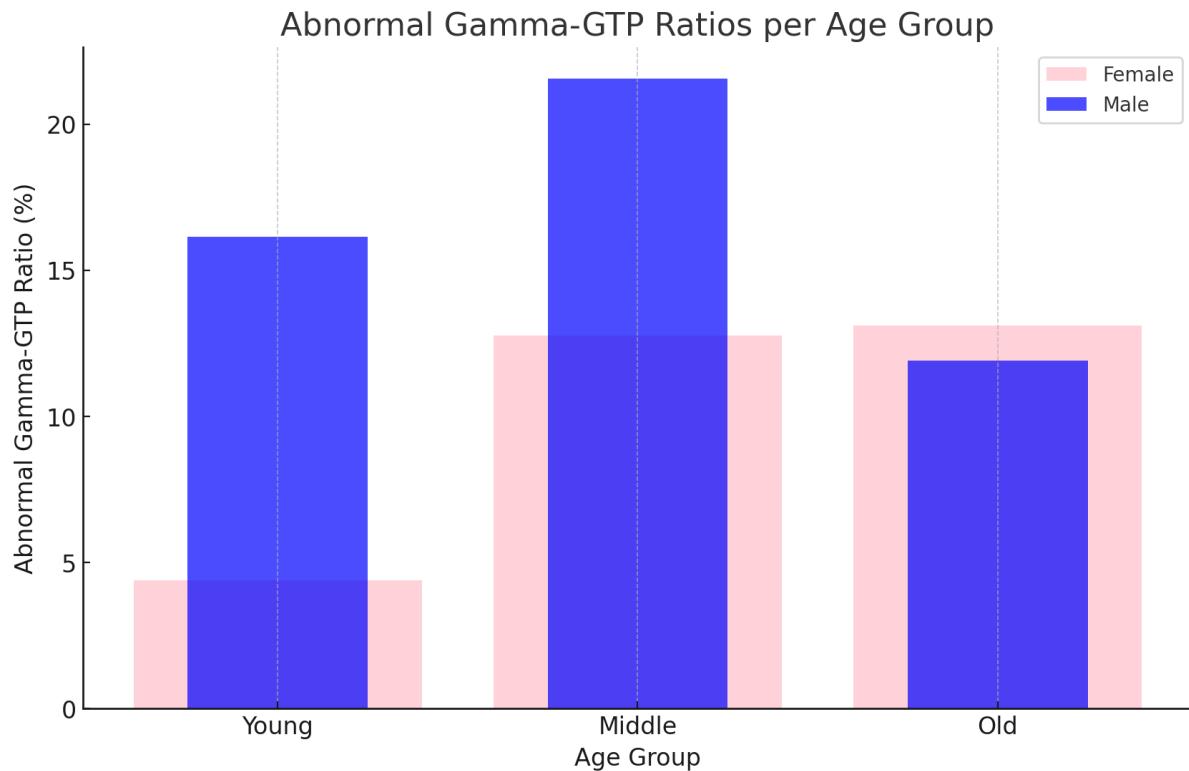


위 그래프의 수치가 양수일 경우 비정상 그룹의 평균이 정상 그룹의 평균보다 출력된 (%) 만큼 높은 것이고 음수일 경우 낮은 것입니다.

이러한 그래프를 토대로 비정상 그룹의 평균과 정상 그룹의 평균의 차이가 많이 나는 변수들을 참고하여 GAMMA\_GTP 비교 분석을 진행하였습니다.

## GAMMA\_GTP 비교 분석

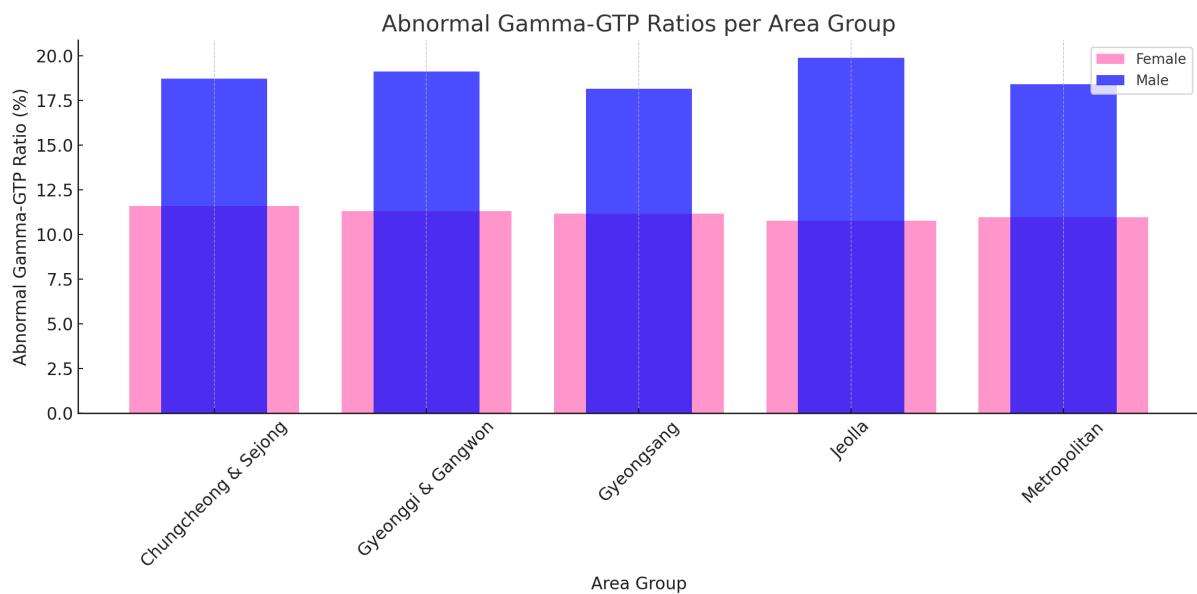
### 1. 연령대 별로 gamma-GTP 이상치의 비율 시각화



- 여성: 감마-GTP > 35 IU/L
- 남성: 감마-GTP > 64 IU/L

그림에서 볼 수 있듯이, 여성 데이터에서는 "Middle" 연령대에서 이상치 비율이 가장 높으며, 남성 데이터에서는 "Old" 연령대에서 이상치 비율이 가장 높습니다.

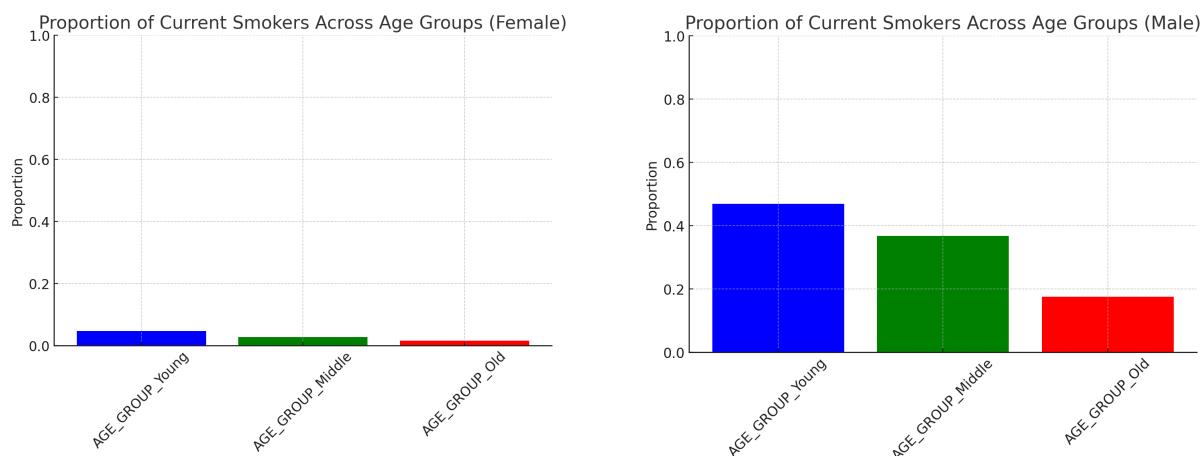
## 2. 지역별 gamma-GTP 이상치 비율 시각화

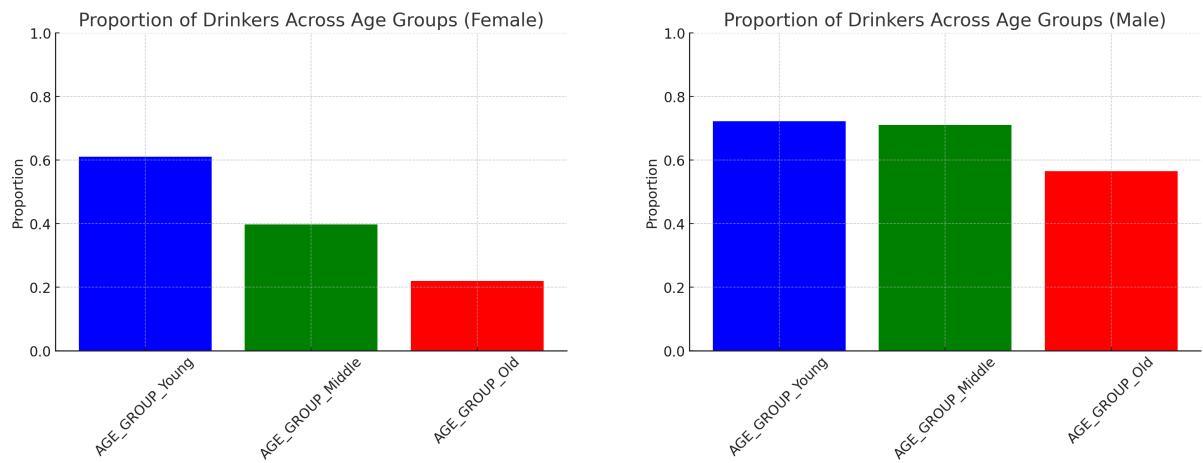


지역별로 큰 차이는 없는 것으로 판단.

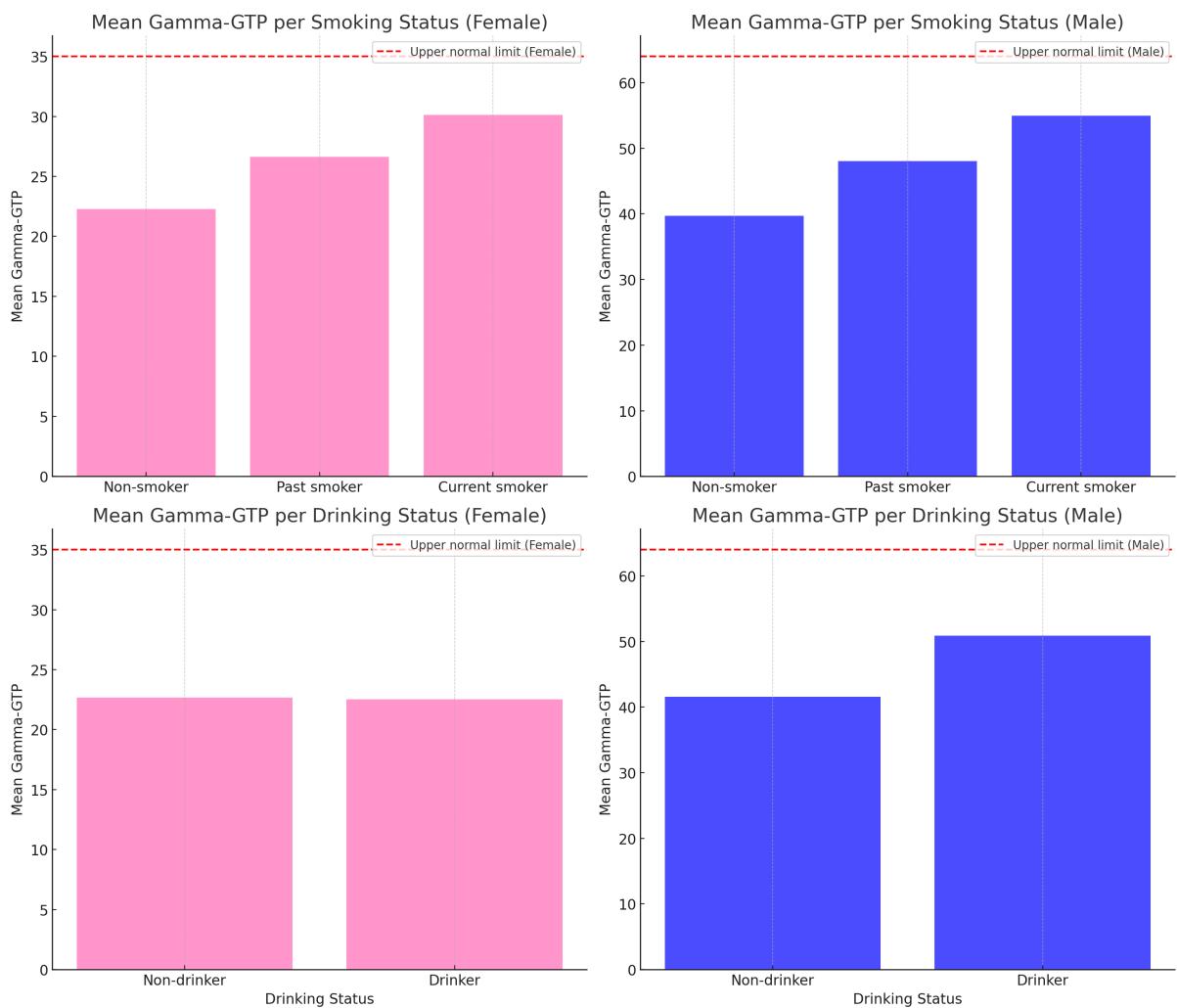
### 3. 흡연 & 음주와 gamma-GTP 연관성

#### 연령대 별 흡연비율과 음주비율





## 흡연 & 음주와 gamma-GTP 연관성



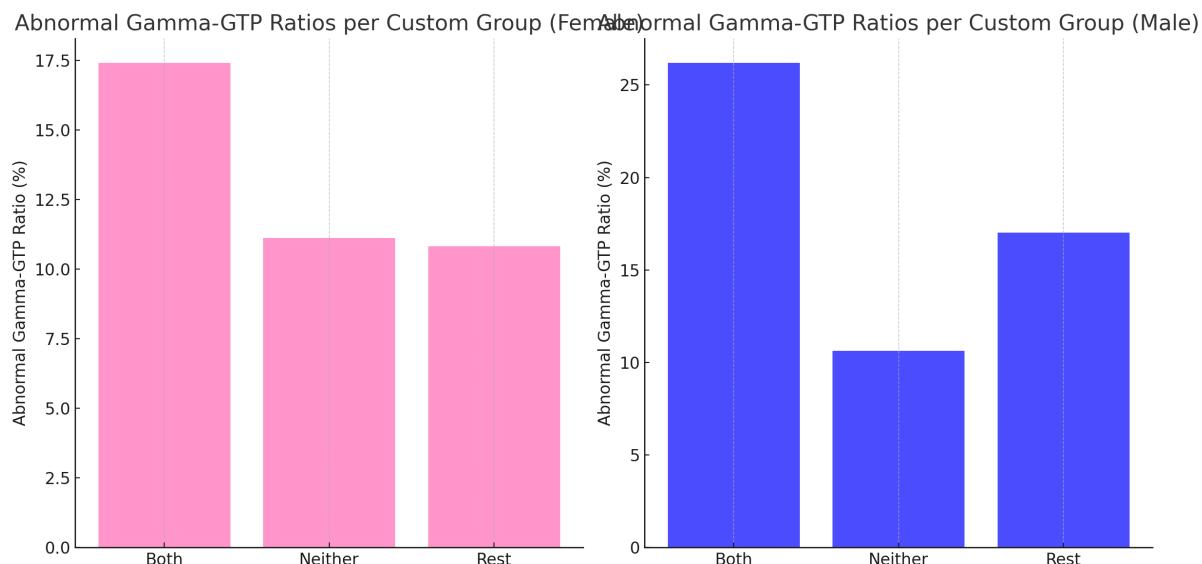
상단의 두 차트는 흡연 상태에 따른 감마-GTP 수치의 평균을 여성과 남성에 대해 나타냅니다.  
하단의 두 차트는 음주 여부에 따른 감마-GTP 수치의 평균을 여성과 남성에 대해 나타냅니다.

빨간색 점선은 각 성별의 감마-GTP 정상치 상한값을 나타냅니다:

- 여성: 35 IU/L
- 남성: 64 IU/L

이 분석은 감마-GTP의 평균 수치가 특정 행위 (흡연 또는 음주)와 관련이 있음을 시사합니다.  
특히, 현재 흡연자와 음주자 그룹에서 감마-GTP 수치가 더 높게 나타나는 경향이 있습니다.

## 흡연 음주 관련 추가 분석



위의 두 차트는 새롭게 정의한 세 그룹에 따른 감마-GTP 이상치 비율을 나타냅니다.

왼쪽 차트는 여성 데이터를, 오른쪽 차트는 남성 데이터를 나타냅니다.

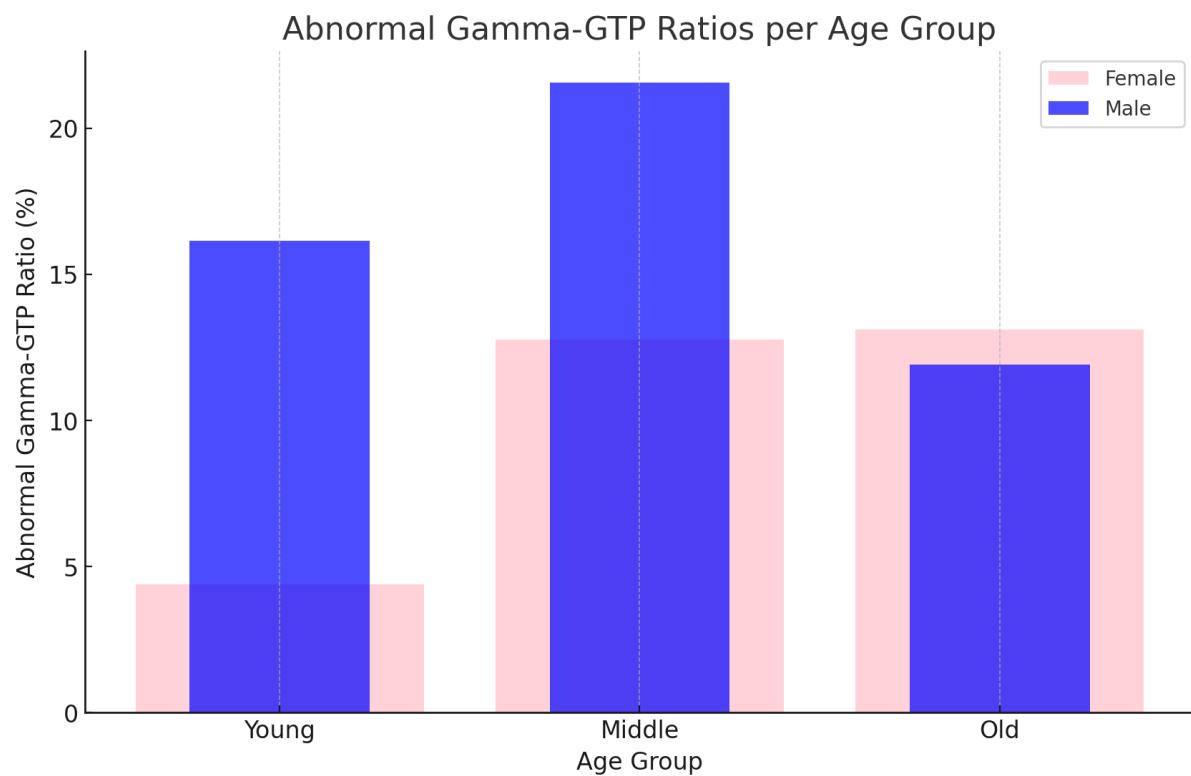
그룹 정의는 다음과 같습니다:

- **Both:** 현재 흡연하고 음주하는 사람들
- **Neither:** 흡연하거나 음주한 적이 없는 사람들
- **Rest:** 나머지 사람들 (과거 흡연자, 현재 흡연자이면서 음주하지 않는 사람 등)

"Both" 그룹에서 이상치 비율이 높게 나타나는 경향이 보입니다.

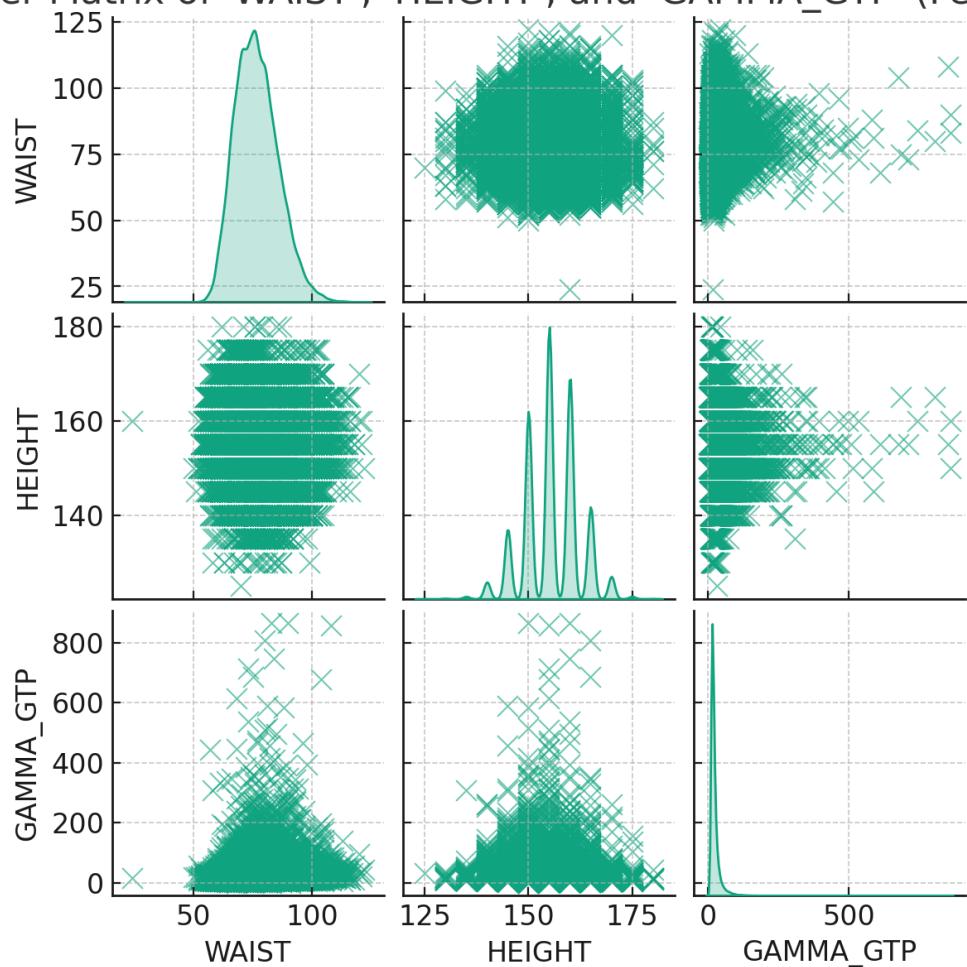
따라서 흡연과 음주가 감마-GTP 수치에 영향을 미칠 수 있음을 나타냅니다.

참고)

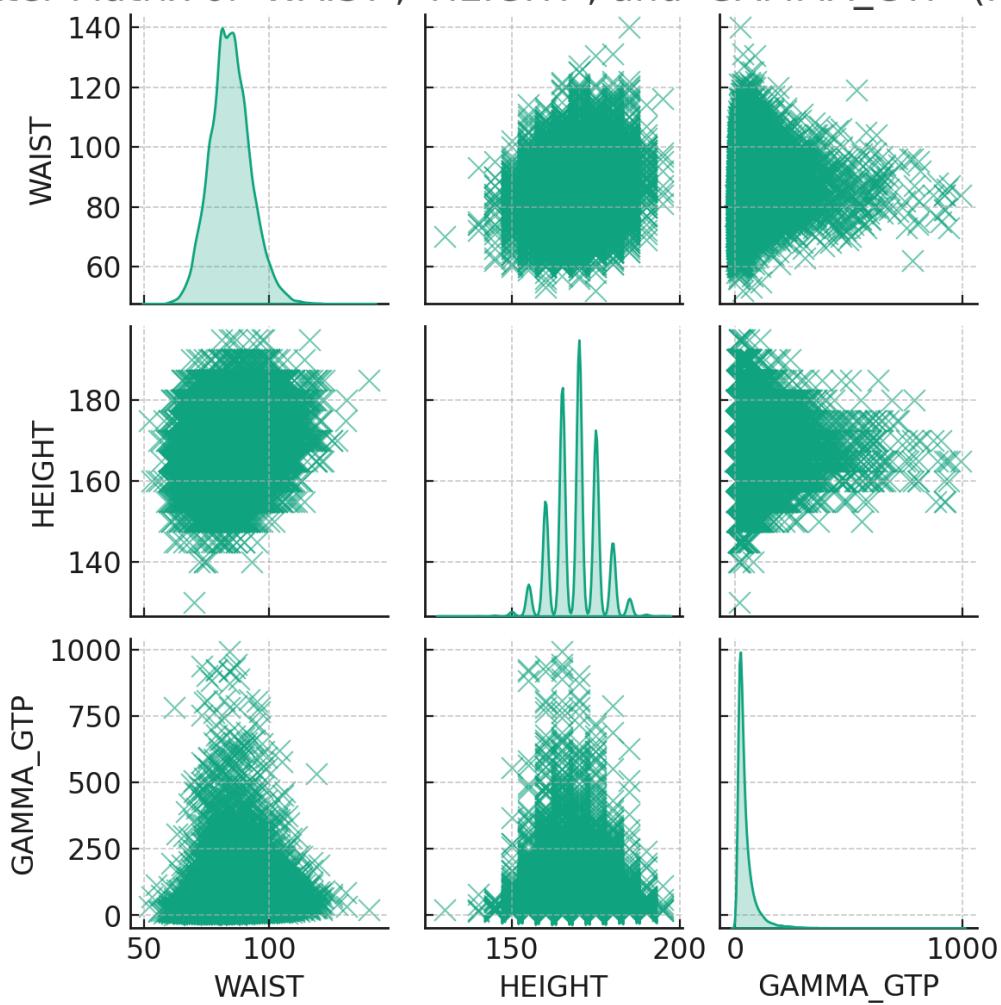


#### 4. 'WAIST', 'HEIGHT', 'GAMMA\_GTP' 간의 상관관계 분석

Scatter Matrix of 'WAIST', 'HEIGHT', and 'GAMMA\_GTP' (Female)



Scatter Matrix of 'WAIST', 'HEIGHT', and 'GAMMA\_GTP' (Male)



산점도 행렬은 'WAIST', 'HEIGHT', 그리고 'GAMMA\_GTP' 간의 관계를 여성과 남성에 대해 나타냅니다. 대각선의 그래프는 각 변수의 분포를 나타내는 Kernel Density Estimate (KDE) 그래프입니다.

상관 계수 행렬은 다음과 같습니다:

여성:

	WAIST	HEIGHT	GAMMA_GTP
WAIST	1.000	-0.132	0.183
HEIGHT	-0.132	1.000	-0.068
GAMMA_GTP	0.183	-0.068	1.000

남성:

	WAIST	HEIGHT	GAMMA_GTP
WAIST	1.000	0.149	0.170
HEIGHT	0.149	1.000	-0.032
GAMMA_GTP	0.170	-0.032	1.000

상관 계수는 -1에서 1 사이의 값을 가지며, 1에 가까울수록 양의 상관관계, -1에 가까울수록 음의 상관관계를 나타냅니다.

- 여성:

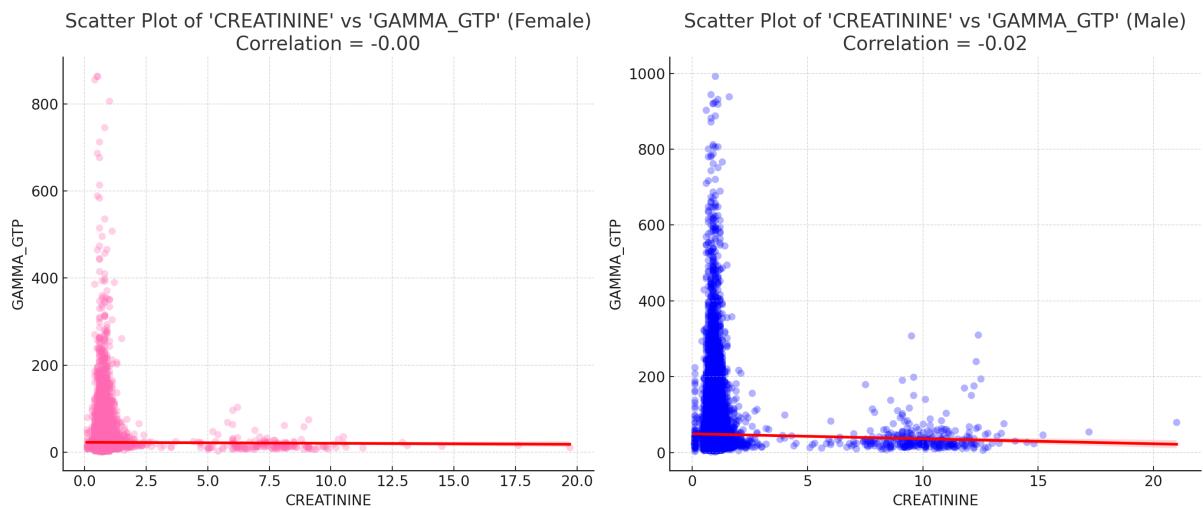
- 'WAIST'와 'GAMMA\_GTP'는 0.183의 상관 계수를 가지고 있어, 약한 양의 상관관계를 가지고 있습니다.
- 'HEIGHT'와 'GAMMA\_GTP'는 -0.068의 상관 계수를 가지고 있어, 약한 음의 상관관계를 가지고 있습니다.

- 남성:

- 'WAIST'와 'GAMMA\_GTP'는 0.170의 상관 계수를 가지고 있어, 약한 양의 상관관계를 가지고 있습니다.
- 'HEIGHT'와 'GAMMA\_GTP'는 -0.032의 상관 계수를 가지고 있어, 매우 약한 음의 상관관계를 가지고 있습니다.

이로써 'WAIST'가 'GAMMA\_GTP'와 약한 양의 상관관계를 가지고 있음을 확인할 수 있습니다. 그러나 'HEIGHT'와 'GAMMA\_GTP'의 상관관계는 매우 약해 의미 있는 관계로 보기 어렵습니다.

## 5. 'CREATININE' 수치와 'GAMMA\_GTP' 관계

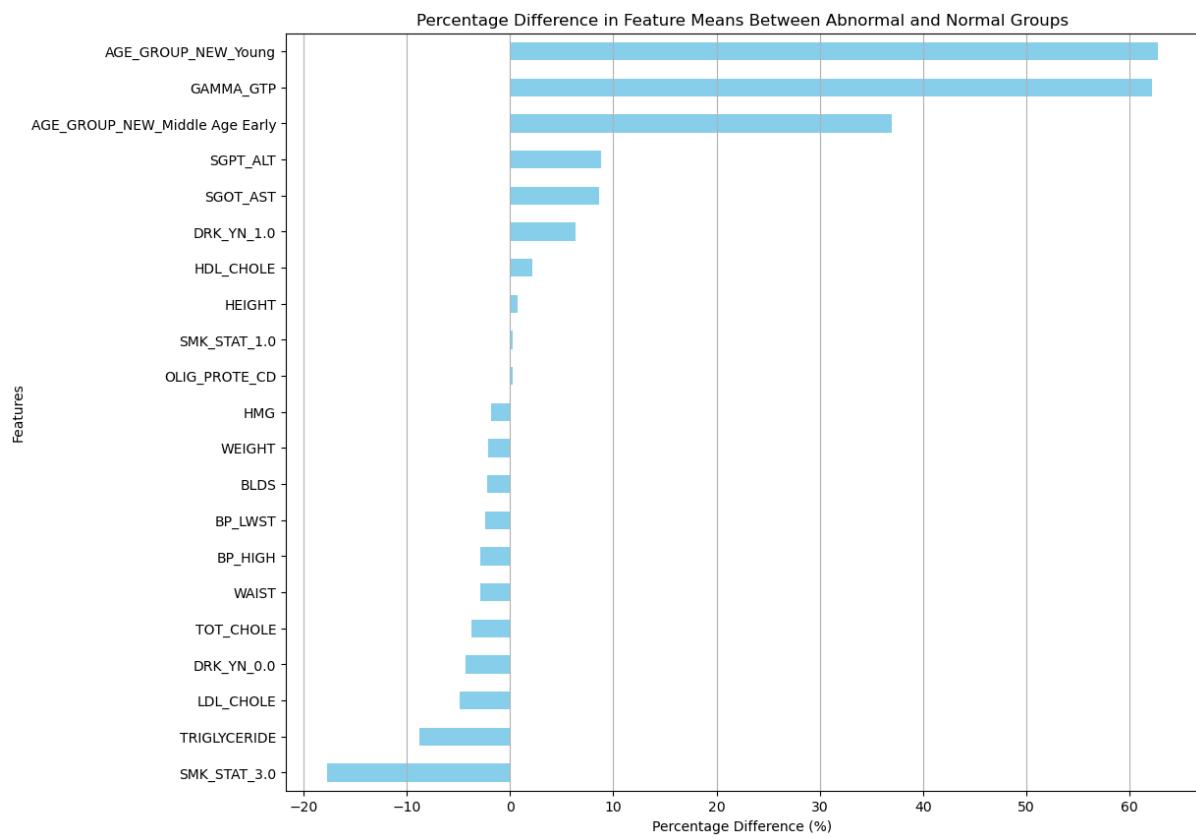


- 여성: CREATININE과 GAMMA\_GTP 간의 상관계수는 약 -0.004  
-0.004
- 남성: CREATININE과 GAMMA\_GTP 간의 상관계수는 약 -0.018  
-0.018

상관 계수의 절댓값이 매우 작아, 'CREATININE'과 'GAMMA\_GTP' 간에는 별다른 선형적인 관계가 없다고 판단됩니다. 산점도와 선형 회귀선 역시 이를 뒷받침합니다. 이는 이 두 변수 간에는 강한 관련이 없음을 의미합니다.

## 남성 데이터 추가 분석

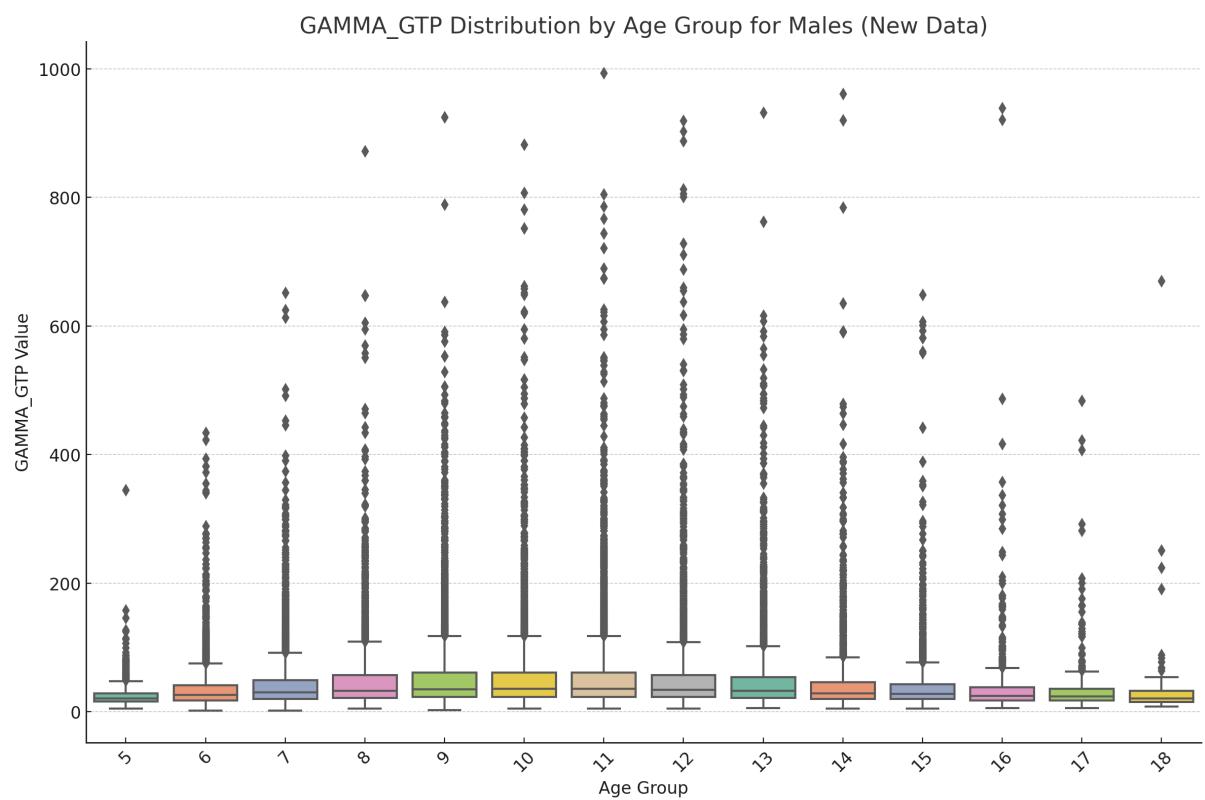
남성 데이터를 중심으로 특정 피처와의 관계를 분석해 보았습니다.



위에서 언급하였던 해당 그래프를 해석하기 위한 분석 진행

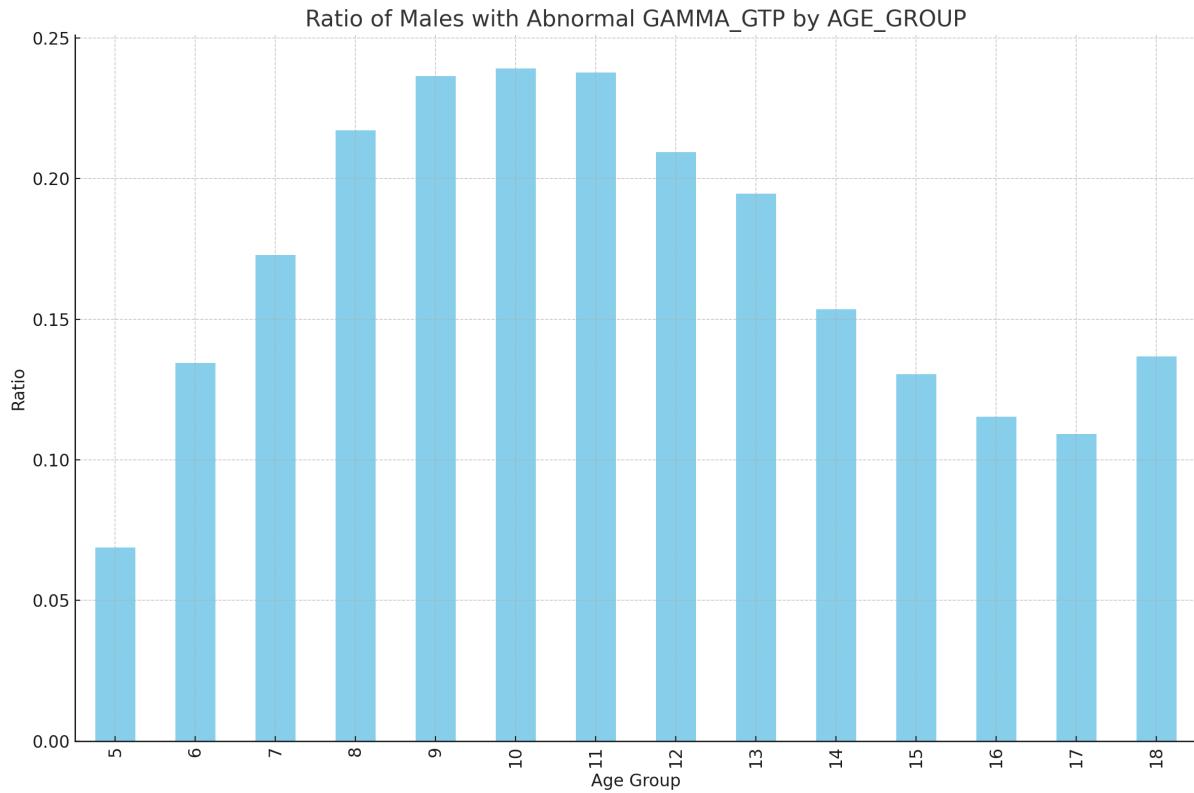
age\_group = 5 는 나이가 20~25를 의미하며, age\_group\_new\_youth 에 해당합니다.

남성의 나이가 **20대 초중반**인 그룹에서 큰 수치를 보이고 있습니다. 따라서 나이별 감마 분포를 확인하였습니다.



이를 통해 얻을 수 있는 정보가 적어 나이별 비정상값을 가지는 수치를 시각화 했습니다.

`age_group = n & gamma_gtp가 비정상 / age_group = n`



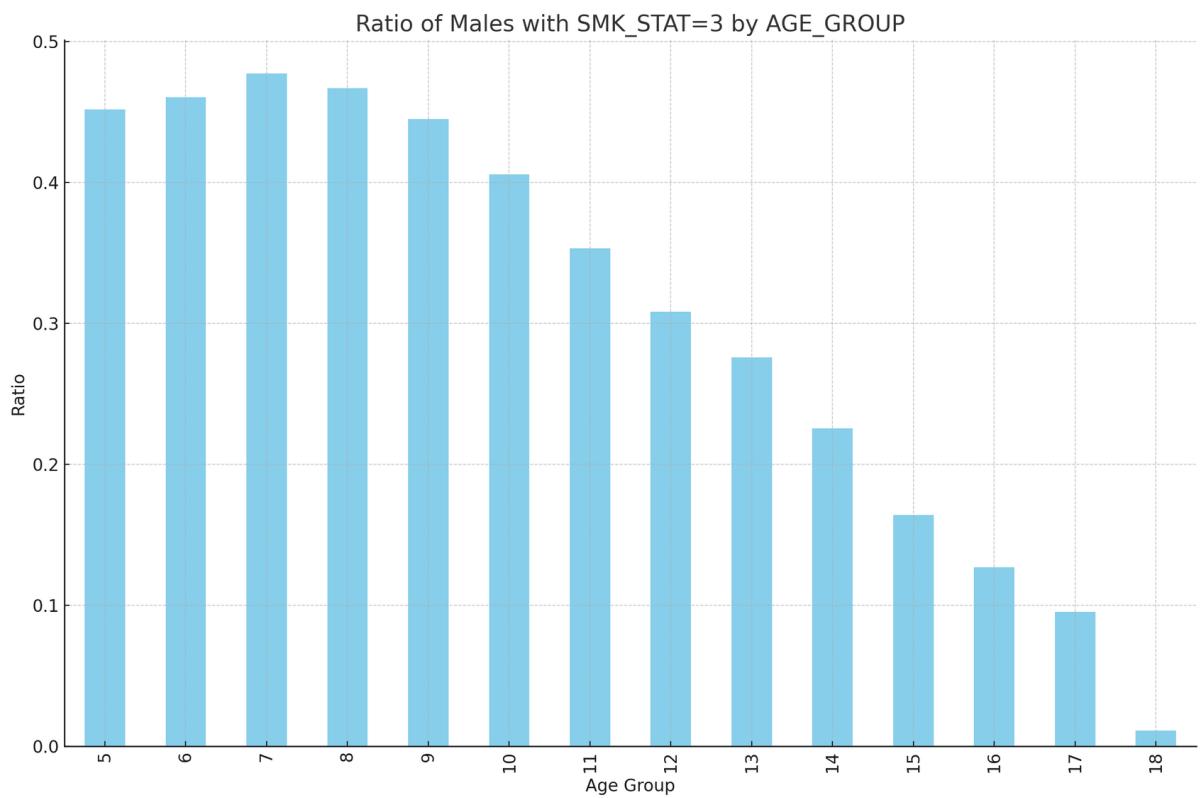
나이가 많아질 수록 감사 수치가 증가하였다가 정점을 찍고 다시 하강하는 모양을 띠고 있습니다. 어찌되었건 중요한 것은 Age\_Group이 5에 해당하는 데이터가 gamma\_gtp 수치에 어떠한 영향을 미치고 있다는 것임으로 **age\_group=5의 뚜렷한 특징**이 매우 중요한 요소가 될 것이라고 판단하였습니다.

따라서, 특정 피처 중에서 20대 중반의 뚜렷한 특징을 살펴보기로 하였습니다.

### 1. 흡연

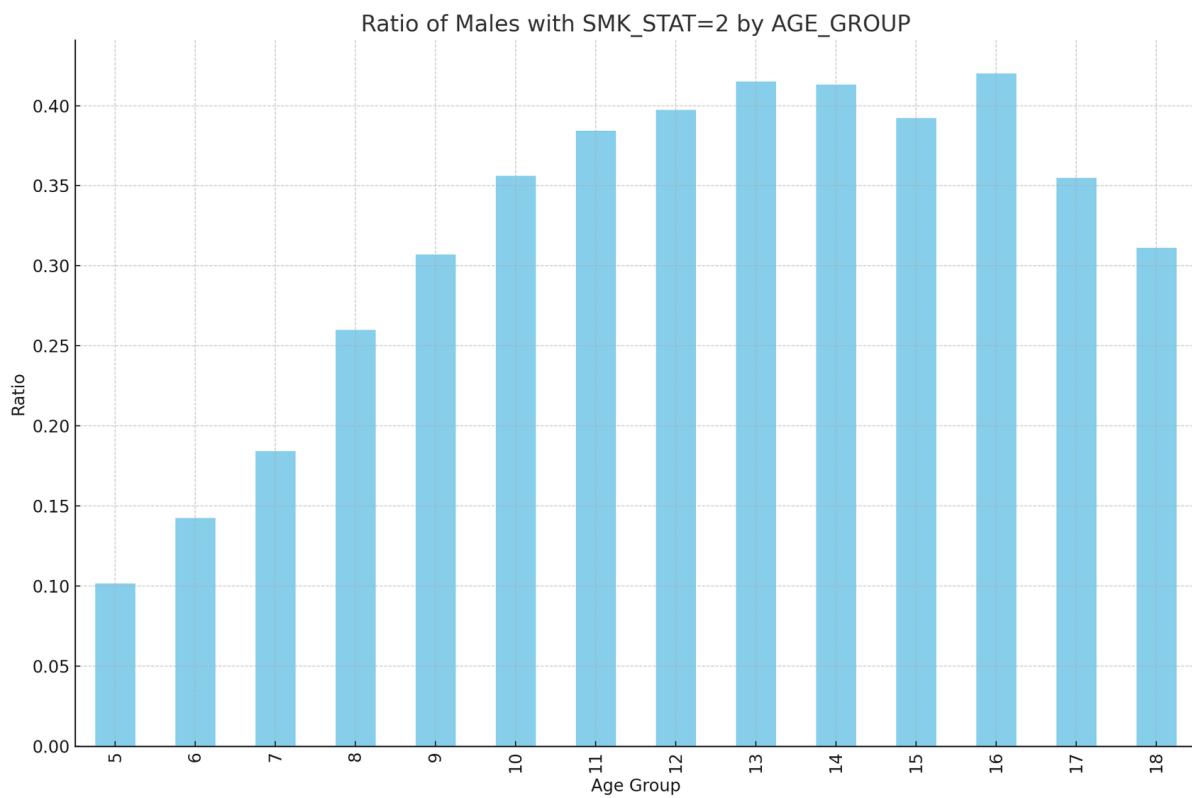
$$\frac{\text{Number of people with } SMK\_STAT = 3 \text{ in } age\_group = n}{\text{Total number of people in } age\_group = n}$$

나이 별로 흡연을 하는 사람의 비율을 수치화한 그래프입니다.

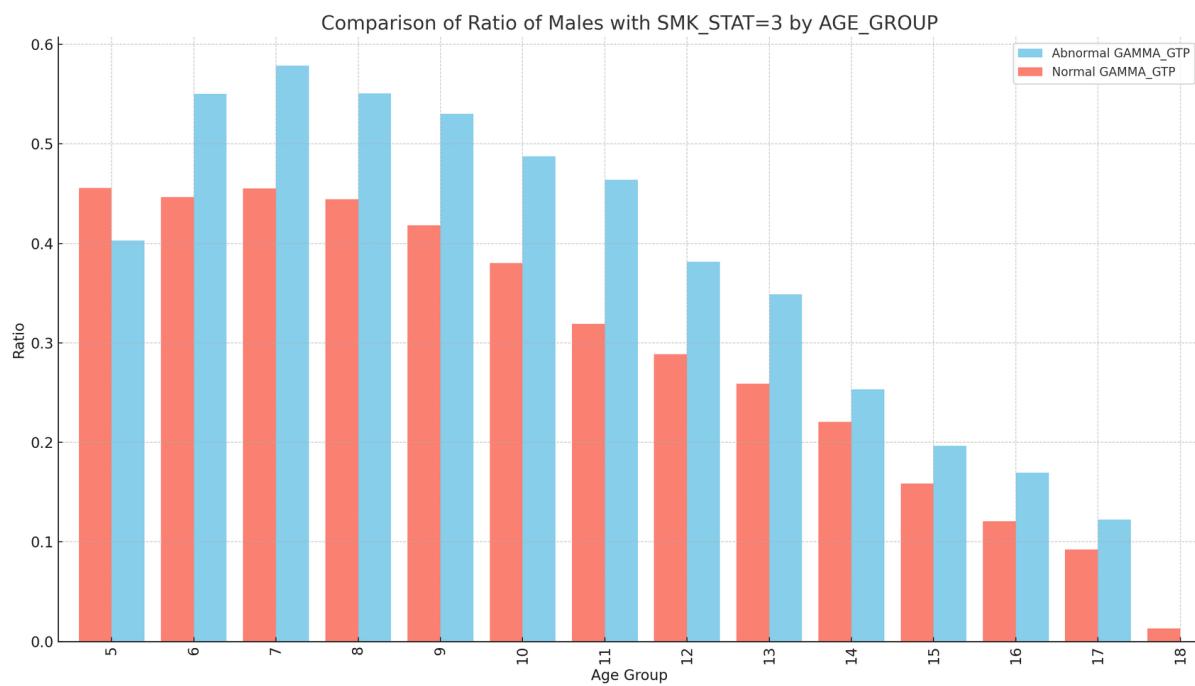


나이가 들 수록 흡연을 하는 사람이 적어진다는 것은 알았지만 Age\_Group=5,6,7,8에 해당하는 숫자는 높은 수치를 가질 뿐 특이사항이 없습니다.

따라서 SMK\_STAT=2의 ratio를 시각화하고 SMK\_STAT=3은 Age\_Group=5일 때부터 흡연을 한 사람들이 계속 한다는 가정을 만들어 줍니다.



흡연을 시작하고 끊은 사람은 나이가 많아질 수록 증가합니다. 그러나 흡연 자체가 감마 수치에 미치는 영향의 중요도가 상대적으로 낮은 편이고 20대 초반의 특이사항이 없다고 판단되어 다른 접근을 시도해 보았습니다.



이는 나이별로 감마 수치가 정상인 그룹과 비정상인 그룹을 나누어 흡연을 하는 사람의 수치를 비교한 그래프입니다. 유일하게 감사 수치가 정상인 그룹의 흡연자가 비흡연자보다 높게 나온 나이 그룹이 5였습니다. 이를 통해 특이사항을 발견하였으며, 이는 흡연을 오랫동안 지속하게 된다면 감마 수치가 증가할 확률이 높다는 접근을 할 수 있었습니다.

## 2. TRIGLYCERIDE (지질)

편의상 지질이라고 대체하여 언급하도록 하겠습니다. 이는 정상값이 30-135 mg/dL으로 여러 그래프를 그려보았습니다.

### (1) age\_group 별로 지질이 30-135을 벗어나는 수치



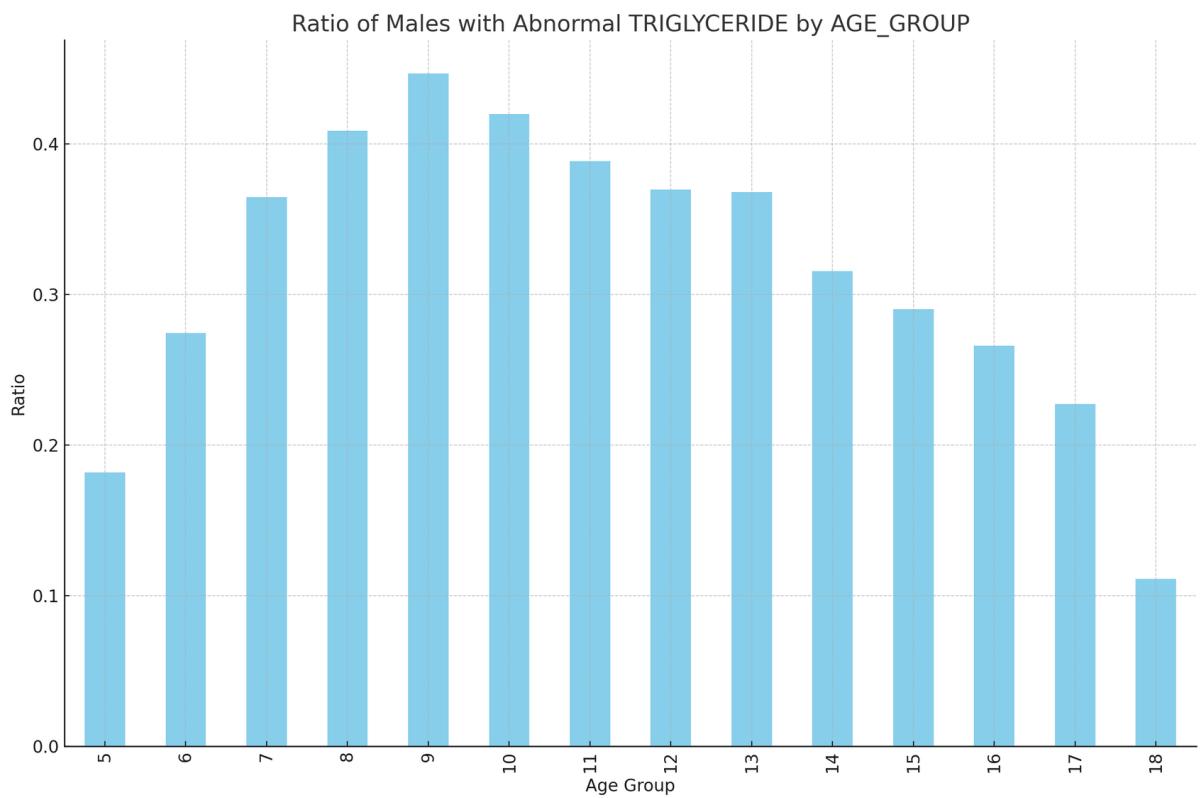
별다른 특이사항을 발견하지 못 하여서 다른 접근 시도.

### (2) 전체 데이터의 abnormal 의 비율

(여기서 abnormal 은 지질 수치가 정상값을 벗어나는 집단)

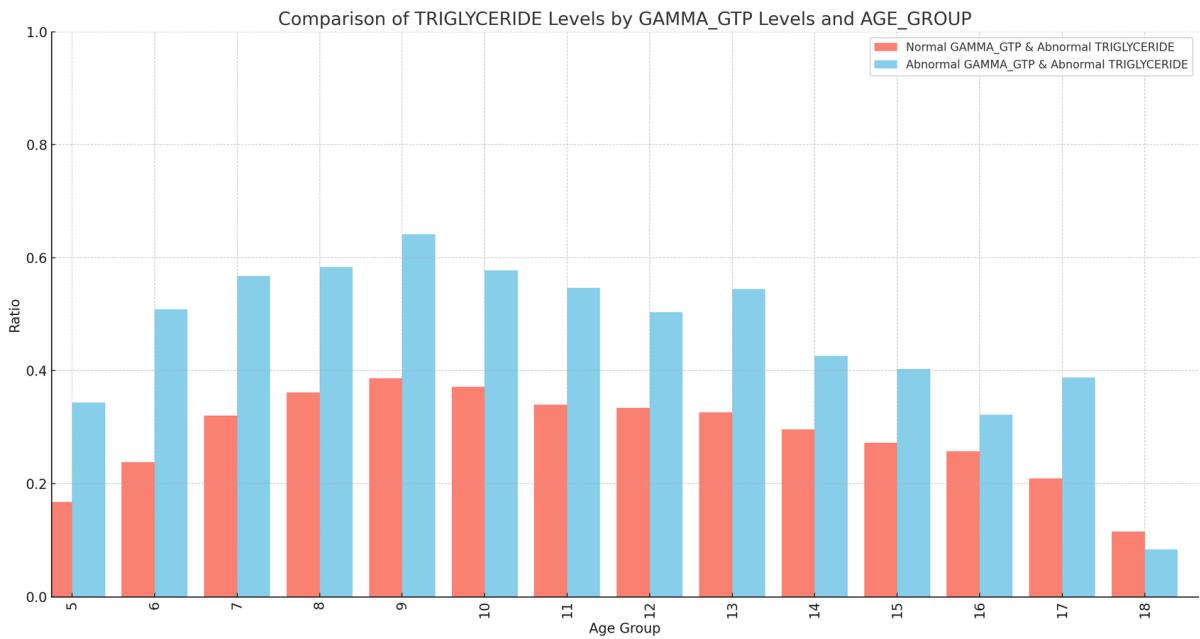
$$\frac{\text{Number of people with } \text{TRIGLYCERIDE} \text{ out of } 30-135 \text{ in } \text{age\_group} = n}{\text{Total number of people in } \text{age\_group} = n}$$

위의 식을 적용하여 계산하면



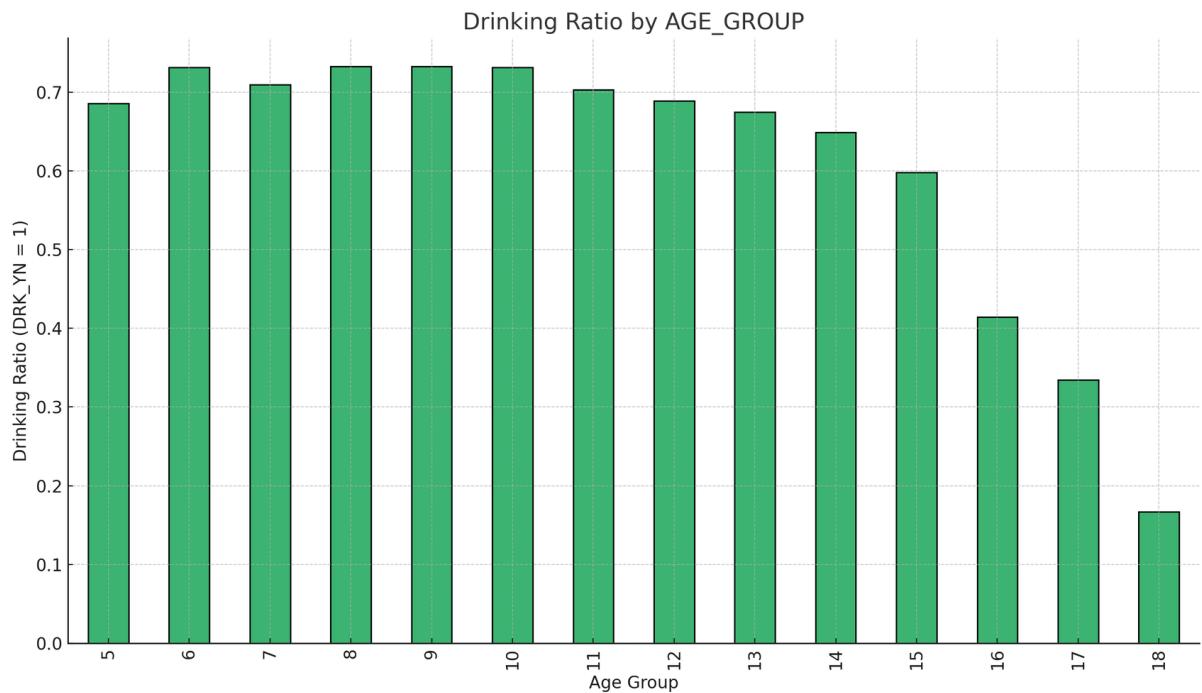
이렇게 그래프가 확인됩니다. 그래프의 양상이 ‘점점 증가하고 정점을 찍은 후 감소’하는 것을 보아 더 디테일한 그래프를 그려보아야겠다고 판단하였습니다.

### (3) 나이별로 감마 수치가 정상인 그룹과 비정상인 그룹을 나누어 abnormal 의 비율 비교.



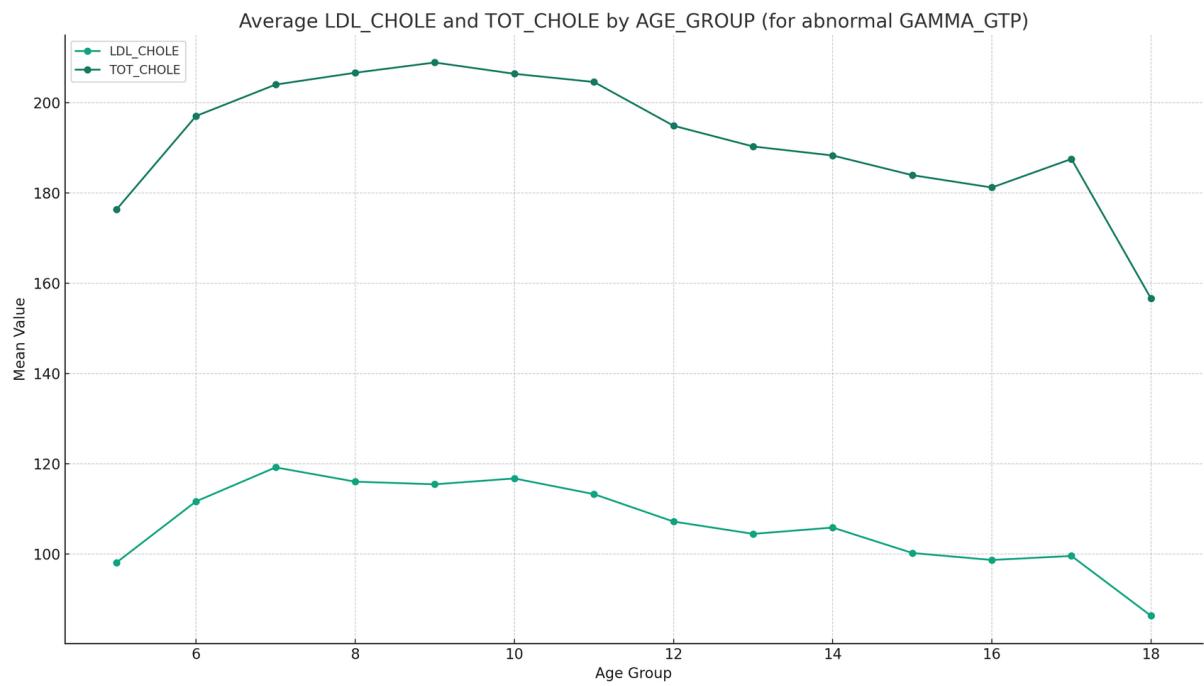
age\_group=5인 데이터의 특이사항을 발견하지는 못 하였지만 지질의 수치가 정상값을 벗어날 수록 감마수치도 벗어남을 확인할 수 있었습니다.

### 3. DRK\_YN\_1.0 (음주)



음주를 하는 사람의 비율 또한 대부분 높은 수치를 지니고 점점 감소하는 것을 보아 특이 사항이 존재하지 않습니다.

### 4. LDL\_CHOLE / TOT\_CHOLE (콜레스테롤)



이 또한 감마 수치와 비슷한 양상의 그래프를 그리며 관련이 있음을 알 수 있지만 `age_group=5`인 데이터의 특이사항을 발견하지 못 하였으며, 위의 분석과 합쳐 애초에 주어진 데이터가 20대의 감마수치가 높은 집단에서 추출되었을 수도 있다고 결론지었습니다.

## 감마 GTP 수치의 이상 여부를 예측

### → 주의 대상 식별

감마 GTP 수치가 비정상 범위는 아니지만 주의가 필요한 경우를 식별하는 것은 매우 중요할 수 있습니다. 이를 위해 다양한 접근 방법을 사용할 수 있습니다. 일반적인 접근 방법 중 하나는 사용 가능한 모든 변수를 고려하여 기계 학습 모델을 사용하여 감마 GTP 수치의 이상 여부를 예측하는 것입니다.

이 경우, 데이터의 다른 특성 (예: 연령, 성별, 음주/흡연 상태 등)을 기반으로 감마 GTP 수치를 예측하고, 예측된 값과 실제 값 간의 차이가 특정 임계값 이상인 경우를 "주의가 필요한" 경우로 식별할 수 있습니다.

# 모델링

## Random Forest → Gamma GTP 수치 예측

전처리가 완료된 데이터를 기반으로 GAMMA\_GTP를 예측하는 모델을 만들어보았습니다.

## 모델 평가

### 여성 데이터

- Root Mean Squared Error (RMSE): 23.14
- 특성 중요도:
  - SGOT\_AST: 37.2%
  - SGPT\_ALT: 49.5%
  - 현재 흡연자: 2.6%
  - 음주자: 4.7%
  - Young Age Group: 0.8%
  - Middle Age Group: 2.5%
  - Old Age Group: 2.8%

### 남성 데이터

- Root Mean Squared Error (RMSE): 50.23
- 특성 중요도:
  - SGOT\_AST: 49.0%
  - SGPT\_ALT: 38.1%
  - 현재 흡연자: 2.9%
  - 음주자: 3.4%
  - Young Age Group: 2.3%
  - Middle Age Group: 2.4%
  - Old Age Group: 1.8%

# 결론

Gamma-GTP 수치와 다른 데이터들 간의 상관관계에 대해 분석해보았습니다. 결론적으로 랜덤포레스트를 이용한 Gamma-GTP 수치 예측모델을 생성하였고 다음과 같은 결론을 도출하였습니다.

랜덤포레스트 예측 모델

실제로는 감마 GTP 수치가 정상이지만, 예측값이 비정상 범주에 들어간 사람의 **IDV\_ID**

## 여성 데이터

- 여성 데이터에서는 이러한 경우가 **1건** 확인되었습니다.
- 관련 **IDV\_ID** : [1]

## 남성 데이터

- 남성 데이터에서는 이러한 경우가 **9777건** 확인되었습니다.
- 관련 **IDV\_ID** : [621908, 259976, 257647, 784108, 907096, 413709, 39755, 48180, 445005, 112325, 65063, 314826, 312789, 867766, 157142, 575513, 77474, 688984, 572884, 320844, ...]

이러한 경우들은 특별한 주의가 필요할 수 있으며, 추가적인 진단이나 검사를 고려할 수 있습니다.