Diabetes & Health indicators

이화여자대학교 데이터사이언스대학원
빅데이터시각화 기말고사 PROJECT
4조(힘내조)

목차

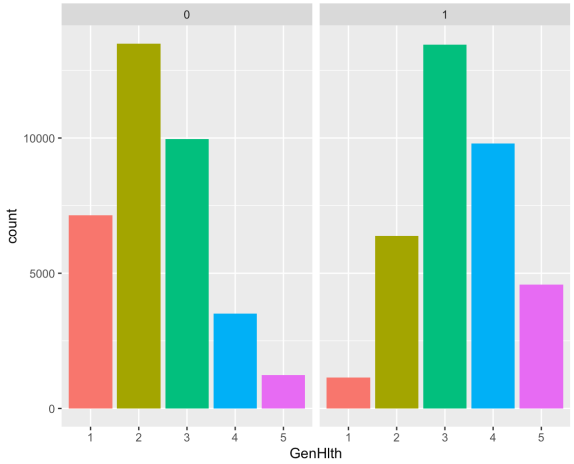
01 데이터셋 - 주제

주제	당뇨병 건강지표
목적	사람들의 생활방식과 당뇨 사이의 관계
출처	UC Irvine Machine Learning Repository CDC Diabetes Health Indicators

01 데이터셋 - 변수

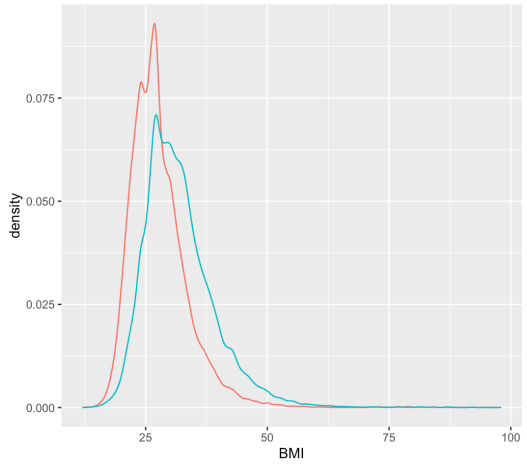
변수			변수 형태	변수			변수 형태
1	ID	환자	정수형	13	Hvy Alcohol Consump	과음 여부	범주형
2	Diabetes_binary	당뇨 여부	범주형	14	AnyHealthcare	의료서비스 이용여부	범주형
3	HighBP	고혈압 여부	범주형	15	No Docbc Cost	의료비 부담 여부	범주형
4	HighChol	고콜레스테롤 여부	범주형	16	GenHlth	일반 건강상태	범주형
5	CholCheck	콜레스테롤 체크	범주형	17	MentHlth	정신 건강상태	정수형
6	BMI	체질량지수	연속형	18	PhysHlth	신체 건강상태	정수형
7	Smoker	흡연 여부	범주형	19	DiffWalk	걷기에 어려움 여부	범주형
8	Stroke	뇌졸중 여부	범주형	20	Sex	성별	범주형
9	HeartDisease or Attack	심장병 혹은 심근경색	범주형	21	Age	나이	범주형
10	PhysActivity	신체활동	범주형	22	Education	교육수준	범주형
11	Fruits	과일 소비량	범주형	23	Income	소득	범주형
12	Veggies	채소 소비량	범주형				

02 EDA



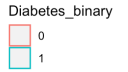
 # 당뇨 여부에 따른 일반 건강상태 분포

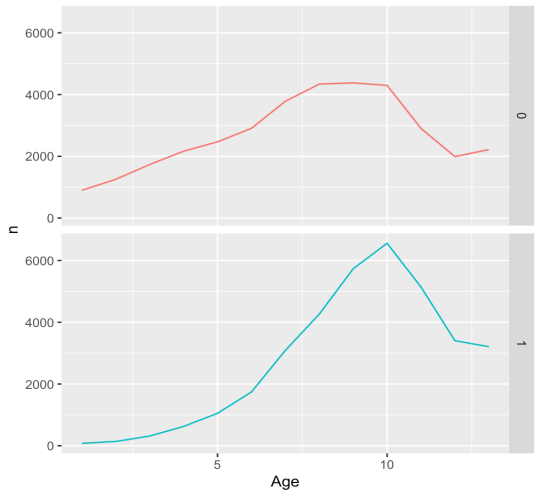
```
ggplot(team_data) +
  geom_bar(aes(x = factor(GenHlth), fill =factor(GenHlth))) +
  labs(x='GenHlth',fill='GenHlth') +
  facet_wrap(~Diabetes_binary)
```



 # 당뇨 여부에 따른 BMI 체질량 지수분포

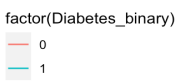
```
ggplot(team_data, aes(BMI,
colour=factor(Diabetes_binary)))+
  geom_density()+
  labs(colour='Diabetes_binary')
```

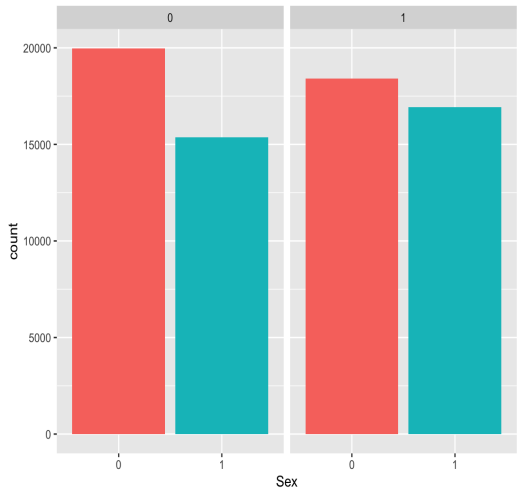




 # 당뇨 여부에 따른 연령 분포

```
team_data %>% count(Age, Diabetes_binary) %>%
  ggplot()+
  geom_line(aes(Age,n,colour=factor(Diabetes_binary)))+
  facet_grid(Diabetes_binary~.)
```





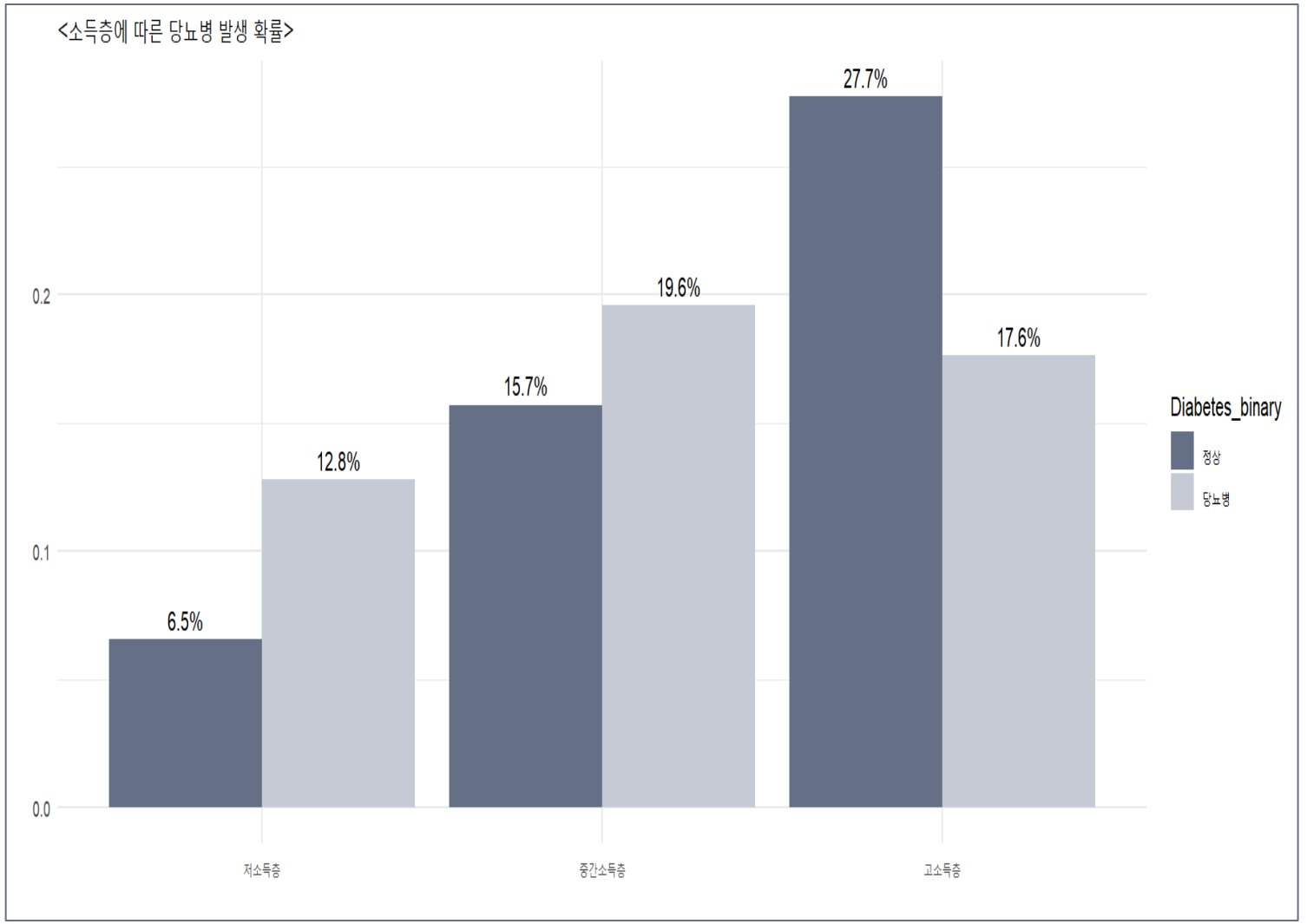
 # 당뇨 여부에 따른 성별 분포

```
ggplot(team_data) +
  geom_bar(aes(x = factor(Sex), fill = factor(Sex)))+
  labs(x='Sex',fill='Sex')+
  facet_wrap(~Diabetes_binary)
```

02 EDA - 시각화1

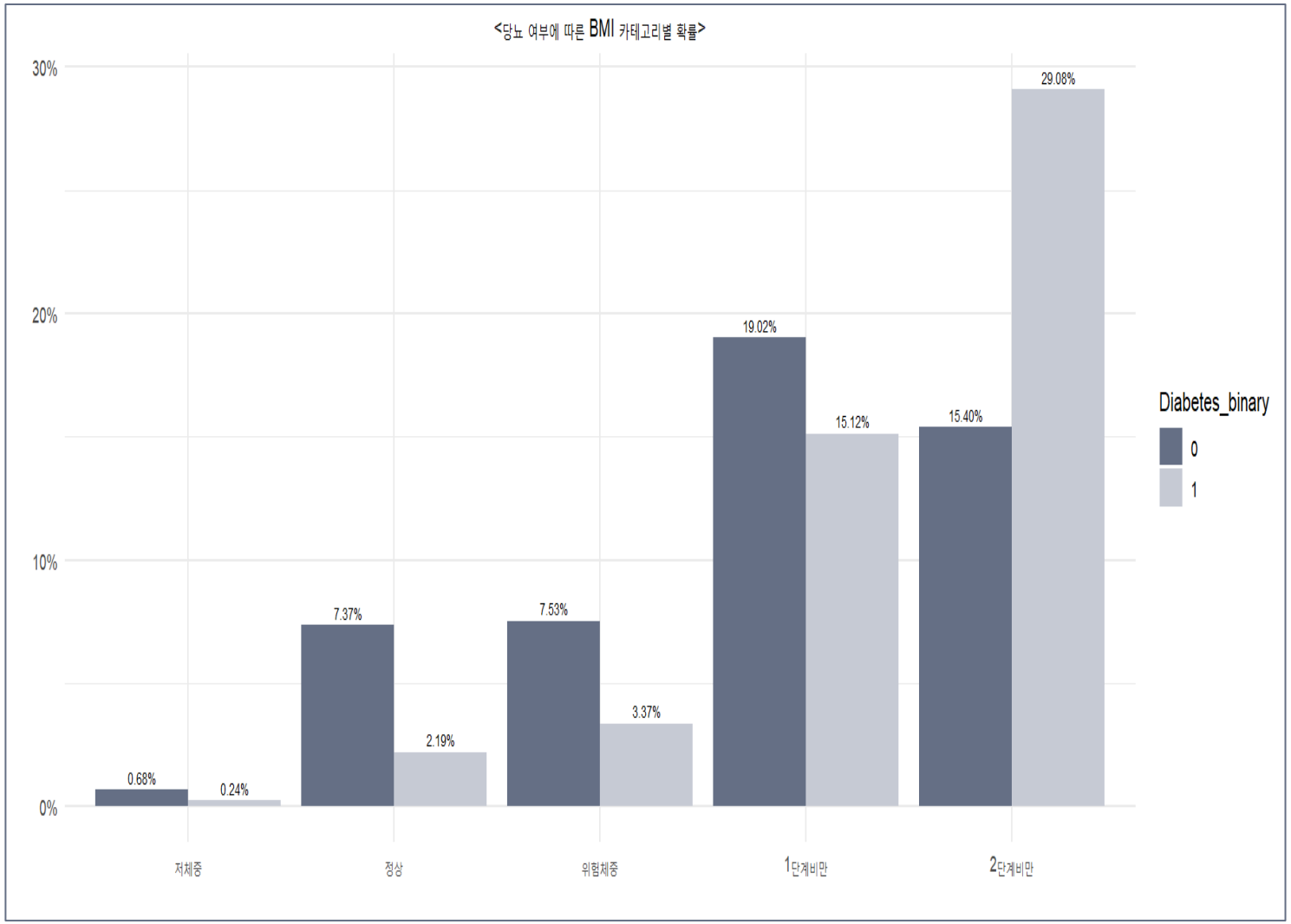
"소득층에 따른 당뇨병 발생 확률"

```
# 소득 기준에 따라 그룹 구분
team_data$Income_category
<- cut(team_data$Income,
breaks = c(1, 3, 6, 8),
labels = c('저소득층', '중간소득층', '고소득층
'),
include.lowest = TRUE)

# 소득 기준에 따라 당뇨병 발생 확률 계산
result_income = team_data %>%
count(Income_category,
Diabetes_binary) %>% mutate(probability =
n / sum(n))
```



02 EDA - 시각화2

"BMI 지수에 따른 당뇨병 발생 확률"



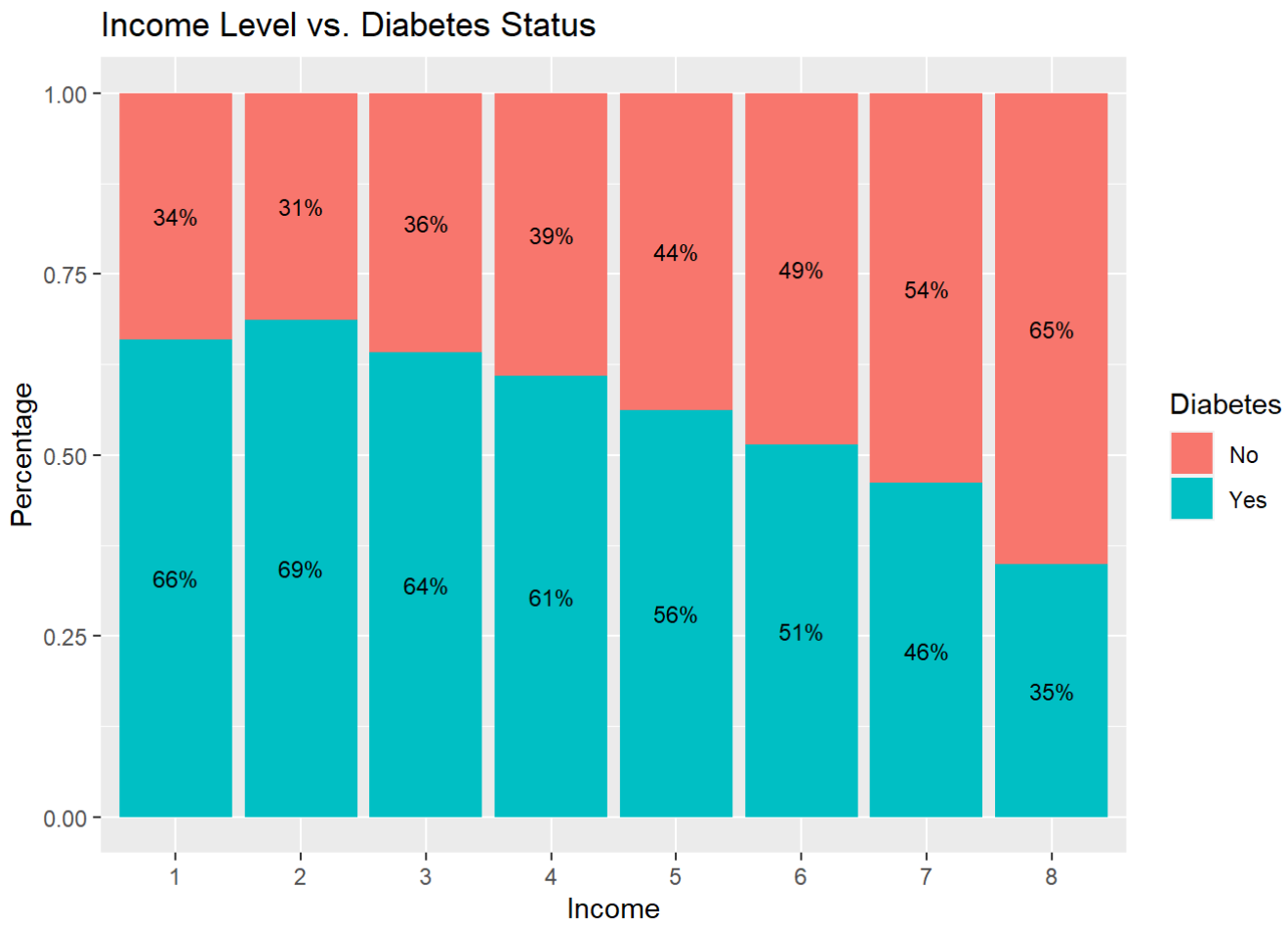

```
# BMI 지수에 따라 그룹 구분
team_data$BMI_category <-
cut(team_data$BMI,
breaks = c(-Inf, 18.5, 22.9, 24.9, 29.9, Inf),
labels = c('저체중', '정상', '위험체중', '1단계
비만', '2단계비만'),
include.lowest = TRUE)

# BMI 지수에 따라 당뇨병 발생 확률 계산
result = team_data %>%
count(BMI_category, Diabetes_binary) %>%
mutate(probability = n / sum(n))
```

03 가설

"소득이 낮은 사람과 BMI가 높은 사람이 당뇨에 더 취약할 것이다."

03 가설 - 시각화1

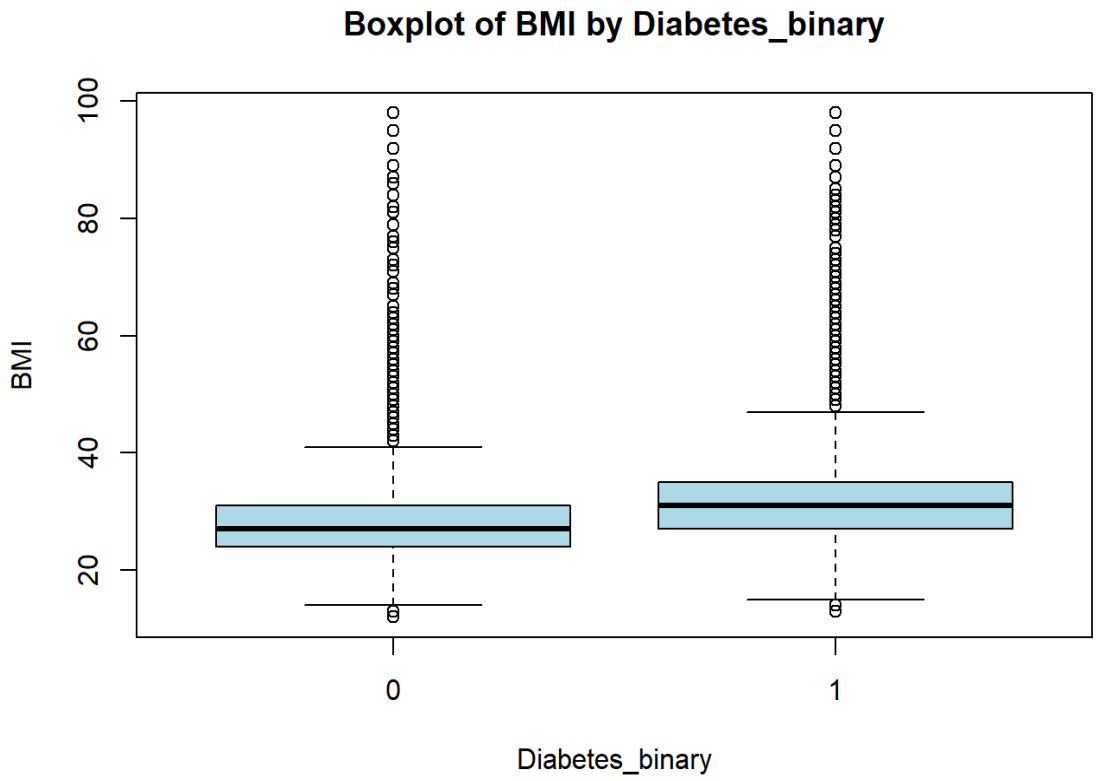
"소득수준과 당뇨여부 관계 – 카이제곱 검정 및 시각화"




```
# 카이제곱 검정

chisq_result <- chisq.test(table(income, diabetes))
chisq_result
##
##  Pearson's Chi-squared test
##
## data:  table(income, diabetes)
## X-squared = 3855.5, df = 7, p-value < 2.2e-16
```

03 가설 - 시각화2

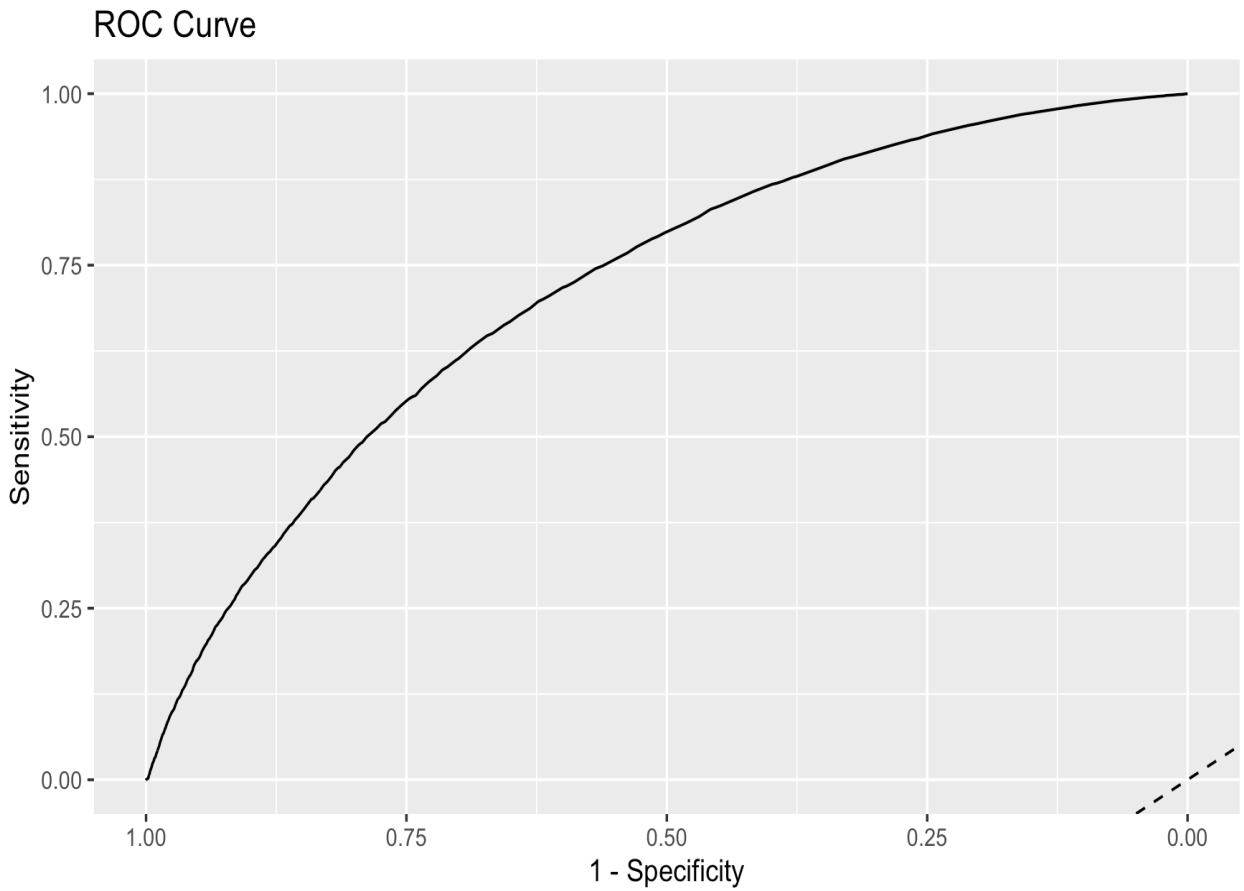
"당뇨여부에 따른 BMI의 평균 차이– T-test 및 시각화"




```
# t-test
t_test_result <- t.test(BMI ~ Diabetes_binary, data = team_data)
print(t_test_result)

##
##  Welch Two Sample t-test
##
## data:  BMI by Diabetes_binary
## t = -81.591, df = 68653, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -4.274321 -4.073781
## sample estimates:
## mean in group 0 mean in group 1
##        27.76996        31.94401
```

03 가설 – 시각화3

"로지스틱 회귀분석 및 ROC Curve 시각화"




```
# 로지스틱 회귀 모델 훈련
logistic_model <- glm(Diabetes_binary ~ Income * BMI, data = team_data, family = "binomial")

# summary(logistic_model)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.9438896  0.1116295  -8.456  < 2e-16 ***
Income      -0.3458025  0.0188515 -18.343  < 2e-16 ***
BMI          0.0697755  0.0037255  18.729  < 2e-16 ***
Income:BMI   0.0050866  0.0006315   8.054 7.98e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

# AUC 출력
auc_value <- auc(roc_curve)
cat("AUC:", auc_value, "₩n")
AUC : 0.7164124
```

04 결론

가설

"소득이 낮은 사람과 BMI가 높은 사람이 당뇨에 더 취약할 것이다."

EDA

1. 소득 기준에 따라 그룹 구분
2. BMI 지수에 따른 그룹 구분

카이제곱 검정

p-value : 매우 작음

► 소득수준 및 당뇨 간에 유의미한 차이가 있음

t-test

p-value : 매우 작음

► BMI 및 당뇨 간에 유의미한 차이가 있음

로지스틱 회귀

Intercept, Income, BMI, Income:BMI p-value : 매우 작음

AUC = 0.7164
: 어느 정도 예측 성능을 가지고 있음
► 유의미한 관련성 있음

E.O.D