# Group project - Final

2024-03-31

Intro Explain why we used BMI as scoring system

## Introduction

## Methods

After viewing the correlation Matrix and computing BMI as the Gold standard, out first step in gathering information on our data was to make a linear regression model in order to find the significant predictors of BMI from the relevant variables.

Since BMI is calculated using weight/Height^2 and Obesity level is calculated based on BMI, we first needed to clean the data to build an appropriate model by removing these variables.

```
##
## Call:
## lm(formula = BMI ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.4414  -3.9418   0.3361   3.4788  23.7055
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          3.35885    6.13000   0.548 0.583823
## GenderMale                          -0.82875    0.33596  -2.467 0.013753
## Age                                  0.31405    0.03353   9.368  < 2e-16
## Family_History_w_Overweightyes       6.80735    0.44395  15.334  < 2e-16
## HiCal_Food_Consumpyes                2.32198    0.50496   4.598 4.65e-06
## Veggie_Consump                       3.00114    0.30679   9.782  < 2e-16
## Main_Meal_Consump                    0.47339    0.20460   2.314 0.020827
## Food_bw_MealsFrequently             -4.03993    1.05179  -3.841 0.000128
## Food_bw_Mealsno                      1.88574    1.36437   1.382 0.167153
## Food_bw_MealsSometimes               3.10074    0.97915   3.167 0.001575
## Does_Smokeyes                       -0.40957    1.05380  -0.389 0.697589
## Water_Consump                        0.57999    0.26596   2.181 0.029369
## Monitor_Caloriesyes                 -2.33645    0.74096  -3.153 0.001649
## Physical_Activ_Amt                  -0.68224    0.19496  -3.499 0.000481
## Tech_Time                           -0.63740    0.27248  -2.339 0.019463
## Alcohol_ConsumpFrequently           -3.36754    5.90954  -0.570 0.568873
## Alcohol_Consumpno                   -5.04959    5.84758  -0.864 0.387992
## Alcohol_ConsumpSometimes            -2.48026    5.85110  -0.424 0.671707
## Transportation_UseBike              -0.84260    4.12896  -0.204 0.838328
```

```
## Transportation_UseMotorbike              5.78189    1.86771   3.096 0.002003
## Transportation_UsePublic_Transportation  5.39302    0.50164  10.751  < 2e-16
## Transportation_UseWalking                2.59810    1.12729   2.305 0.021329
##
## (Intercept)
## GenderMale                              *
## Age                                     ***
## Family_History_w_Overweightyes          ***
## HiCal_Food_Consumpyes                   ***
## Veggie_Consump                          ***
## Main_Meal_Consump                       *
## Food_bw_MealsFrequently                 ***
## Food_bw_Mealsno
## Food_bw_MealsSometimes                  **
## Does_Smokeyes
## Water_Consump                           *
## Monitor_Caloriesyes                     **
## Physical_Activ_Amt                      ***
## Tech_Time                               *
## Alcohol_ConsumpFrequently
## Alcohol_Consumpno
## Alcohol_ConsumpSometimes
## Transportation_UseBike
## Transportation_UseMotorbike             **
## Transportation_UsePublic_Transportation ***
## Transportation_UseWalking               *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.705 on 1385 degrees of freedom
## Multiple R-squared:  0.5017, Adjusted R-squared:  0.4941
## F-statistic: 66.39 on 21 and 1385 DF,  p-value: < 2.2e-16
```

```
## [1] "RMSE:"
```

```
## [1] 5.729089
```

From the summary of our model, we noted that our R^2 is .49 which means only half the variability of BMI was captured by this data. This is normal for a dataset of this nature which deals with predicting humans. This low value is also justified because our data is missing major predictors of obesity like calorie intake.

Based on the model we were able to find the most significant predictors for BMI from the P-values provided

We were also able to find out how accurate these predictors were in predicting an individuals level of obesity by performing multiple logistic regression using Obesity level as the target.

```
accuracy
```

```
## [1] 0.4758523
```

The accuracy of our predictors ability to predict obesity group was around 45% which is standard for data of this nature. It also means that we need more information than the variables provided to fully predict the Obesity group for an individual.

Based on results of first model (regression) we tested the impact/dependency between the significant predictors BMI, and BMI using Kruskal-Wallis tests. If a predictor had a significant impact on BMI, we then performed a Chi-Square test to determine if this relationship was significant or random using the Obesity level, a health based grouping of BMI values. Based on these test results we were be able to determine the significant predictors of an individuals BMI.

Also, in order to test the effects of BMI on an individuals health we performed a Kruskal-Wallis test to measure the effect of BMI on an individuals age as well as the effect of Obesity level on age.

As a result of the aforementioned tests, every variable that passed the KW test also passed the Chi-Square test and as a result we were able to establish the significant predictors of obesity:

Family History of Overweight

Hi Calorie Food Consumption

Physical Activity

Food between Meals

Water Consumption

Monitor Calories

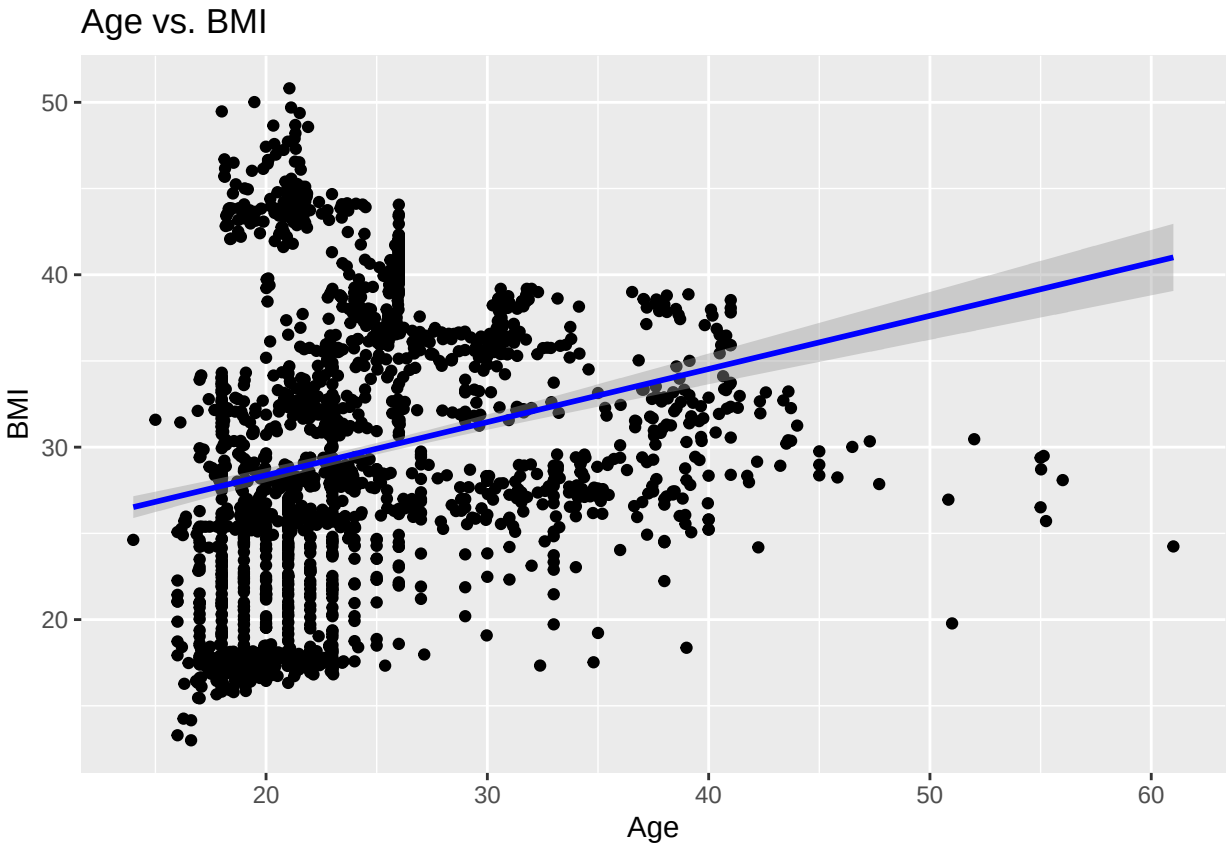
Transportation Use

Tech Time

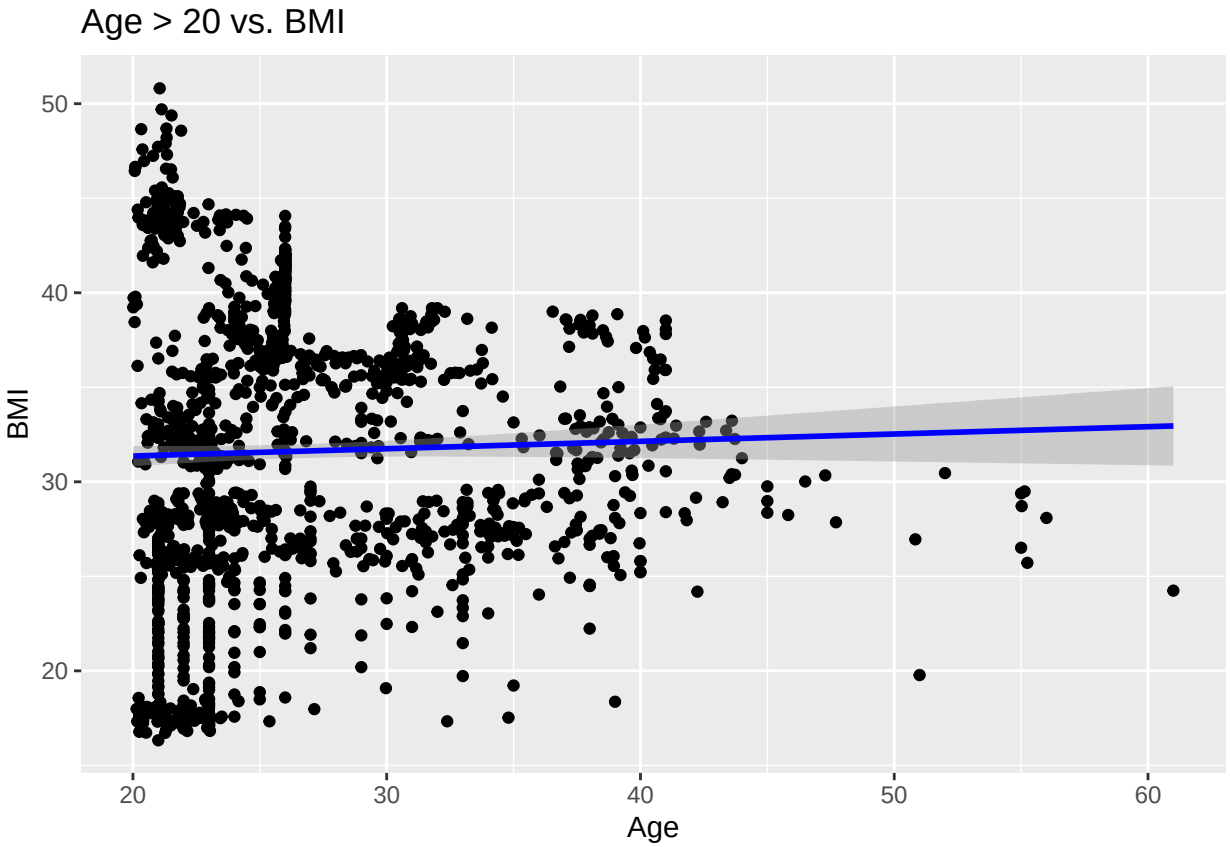
We also learned that while BMI itself doesn't have a significant impact on age, an individuals obesity level does.

Armed with these insights, we went on to determine how much obesity level impacts age by performing a logistic regression predicting age using BMI. Before we did this, we plotted and analyzed a regression line of BMI and age at different age ranges to visualize this relationship.

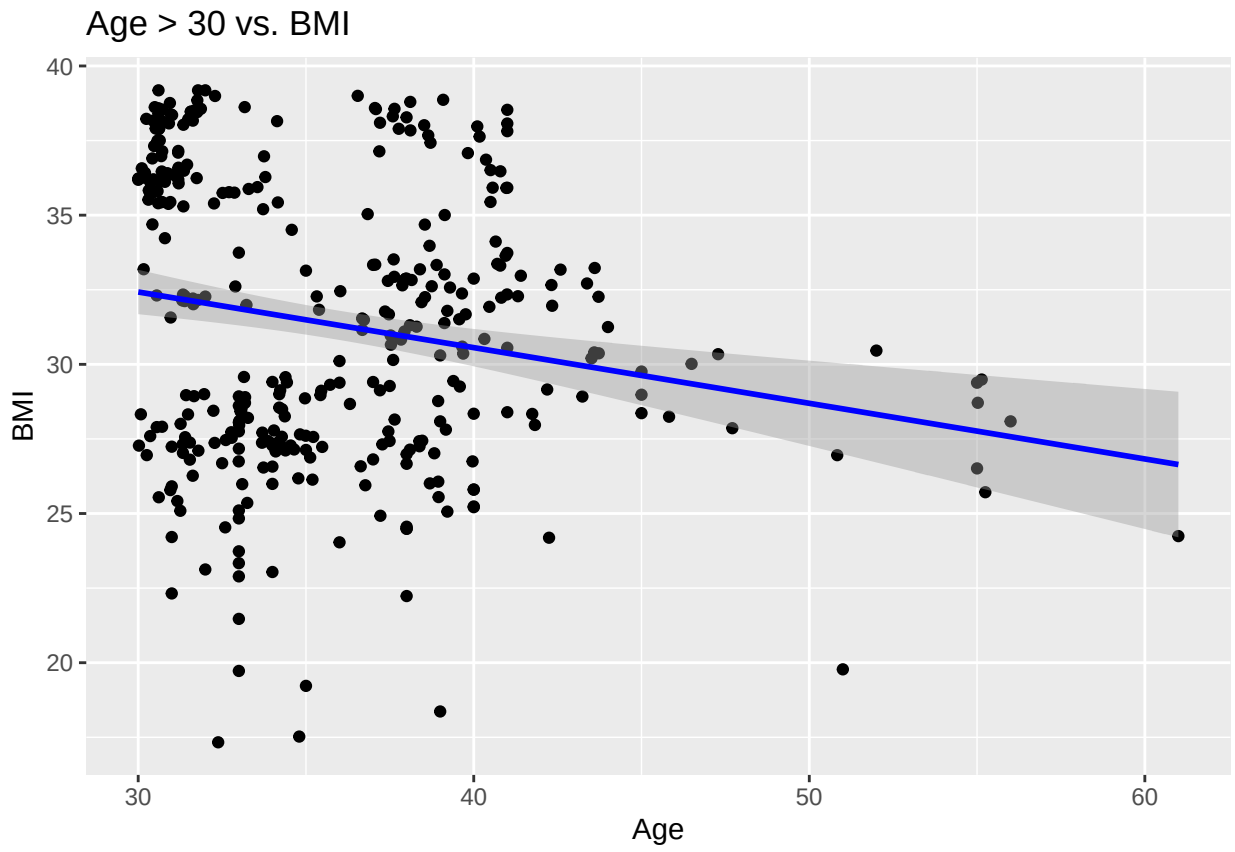
```
## `geom_smooth()` using formula = 'y ~ x'
```



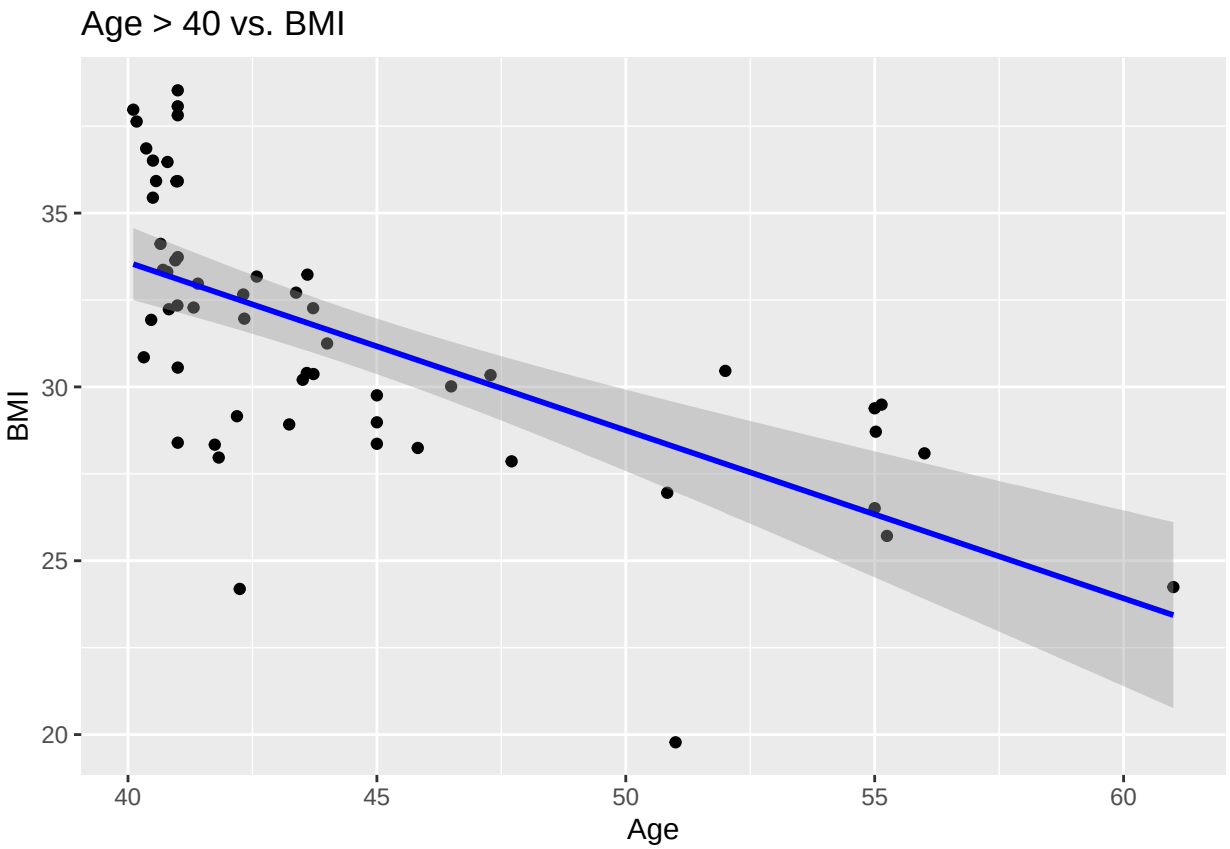

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
## `geom_smooth()` using formula = 'y ~ x'
```




```
## `geom_smooth()` using formula = 'y ~ x'
```




Based on the plots, it was apparent that as the age of the observations increase, the maximum BMI observed decreases. We then extrapolated that younger individuals tend to have higher BMIs as physically, their bodies are not stable or fully developed yet. We then inferred that as one ages, a high BMI can lead to health complications which could be the reason for a decrease in observations. Based on these result we then performed our logistic regression.

In order to avoid skewing the data, as BMI can often be misinterpreted for younger and developing individuals, we took a look at the effect of type 2 obesity (BMI $>$ 35) on whether an individual is older than 40 years old.

```
data2<-data
data2$Age40 <- ifelse(data2$Age > 40, 1, 0)
data2$BMI35 <- ifelse(data2$BMI > 35, 1, 0)
 # Fit logistic regression model
log_model4035 <- glm(Age40 ~ BMI35, data = data2, family = binomial)
summary(log_model4035)
```

```
##
## Call:
## glm(formula = Age40 ~ BMI35, family = binomial, data = data2)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.4795     0.1514 -22.990   <2e-16 ***
## BMI35        -0.4224     0.3285  -1.286    0.198
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 524.20  on 2110  degrees of freedom
## Residual deviance: 522.42  on 2109  degrees of freedom
## AIC: 526.42
##
## Number of Fisher Scoring iterations: 6
```

Overall, based on these results, there was no statistically significant evidence to suggest that having a BMI over 35 affects the likelihood of an individual being younger than 40 years old, as the coefficient for BMI35 was not significant. However we used these results to find the likelihood that an individual is over 40 if their BMI is greater than 35

intercept (Estimated Coefficient: -3.4795): The intercept represents the estimated log odds of an individual being younger than 40 years old when they do not have a BMI over 35. A negative coefficient suggests that individuals who do not have a BMI over 35 are less likely to be younger than 40 years old.

BMI35 (Estimated Coefficient: -0.4224): The coefficient for BMI35 represents the change in the log odds of an individual being younger than 40 years old when they have a BMI over 35 compared to when they do not have a BMI over 35. Here, however, we're interested in how it affects the likelihood of an individual being older than 40 years old. Given that the coefficient is negative, it implies that individuals with a BMI over 35 are less likely to be older than 40 years old.

Now from this we were able to find the probability an individual is over 40 years old if their BMI is greater than 35 (Obesity type 2)

```r
intercept <- -3.4795
BMI35_coefficient <- -0.4224

# BMI value indicating over 35
BMI_over_35 <- 1

# Calculate log odds
log_odds <- intercept + BMI35_coefficient * BMI_over_35

# Convert log odds to probability using logistic function
probability_over_40 <- exp(log_odds) / (1 + exp(log_odds))

# Print the result
probability_over_40
```

```
## [1] 0.01980339
```

As you can see, the probability that an individual is older than 40 if their BMI is >35 is around 1% which is very low likelihood. This may suggest some form of missing data which we believe is not random as our results show that the higher in age, the lower the instances of very high BMI. In fact the maximum BMI observed decreases inversely with age. We further extrapolated this data using this information by assuming that this is caused due to health complications that we will attribute to obesity. this conclusion was uses as the basis for our health scoring system

# Results

Our results showed what variables are predictors of BMI and Obesity group, we used this info to construct a scoring system based which assesses obesity risk. Our result also pointed to the fact that Age is significantly influenced by an individuals Obesity level especially if they were classified as Obesity Type 2 or 3.

We also gathered enough information to construct a health risk scoring system which is based upon the variables impact on an individuals ability to reach an old age.

### Scoring Systems

# Obesity Risk Score

Physical Activity Score: Lower scores are assigned for higher amounts of physical activity, reflecting its role in reducing obesity risk. This emphasizes the protective effect of physical activity against obesity.

Tech Time Score: Increased screen time, captured as Tech Time, is associated with higher scores, indicating its negative impact due to sedentary behavior.

Vegetable Consumption Score: Higher vegetable consumption is rewarded with lower scores, supporting the role of a plant-rich diet in maintaining a healthy weight.

Family History Score: A positive family history of overweight or obesity significantly increases the score, acknowledging the genetic and environmental influence on obesity risk.

High-Calorie Food Score and Food Between Meals Score: High scores for high calorie food consumption and frequent eating between meals highlight their contribution to caloric excess and obesity.

Water Consumption Score: Higher water consumption, indicative of healthier lifestyle choices, is assigned lower scores.

Monitor Calories Score: Engaging in monitoring calories is seen as a positive behavior and thus receives a lower score, reflecting its importance in weight management.

Transportation Use Score: Preference for active modes of transportation like walking or biking scores lower, aligning with the promotion of physical activity.

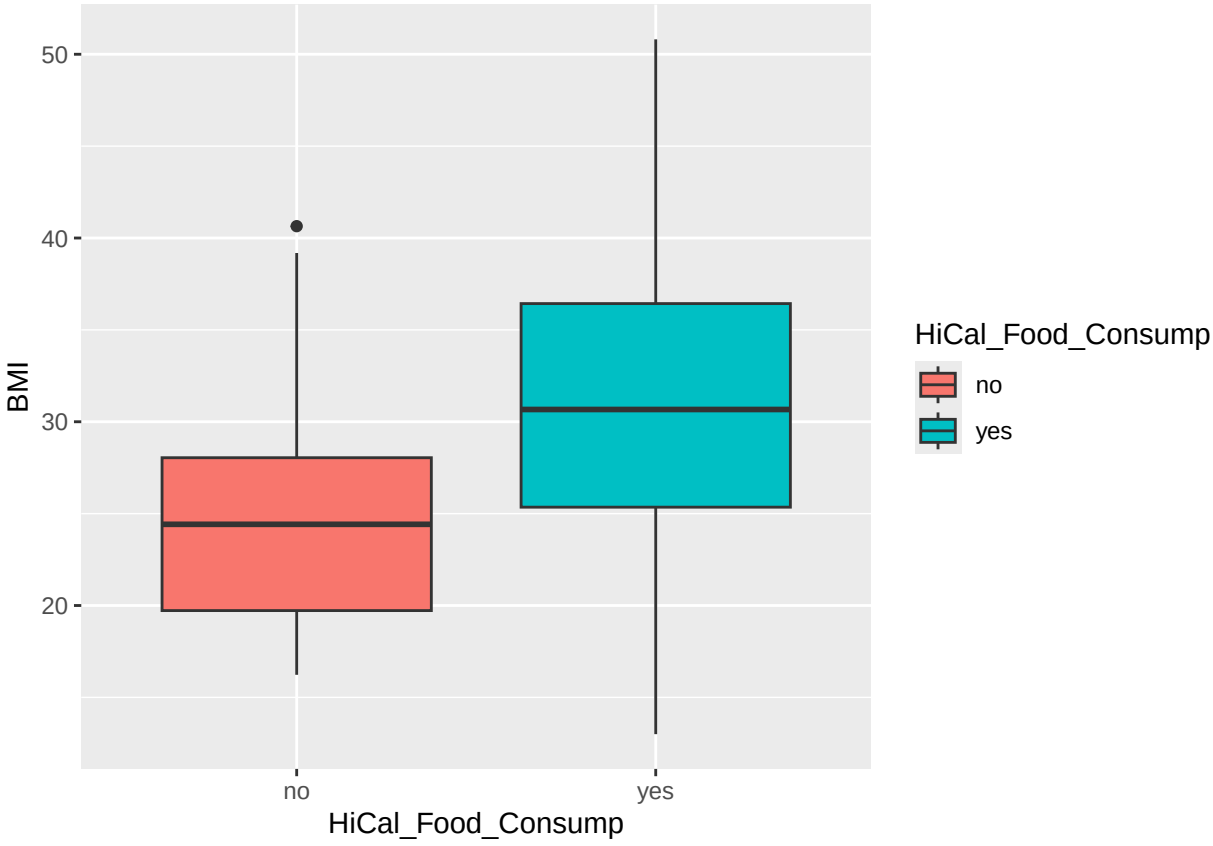
BMI Score: Directly incorporates the WHO classification of BMI into the scoring, with higher categories of BMI receiving higher scores to directly account for current weight status in the obesity risk assessment.

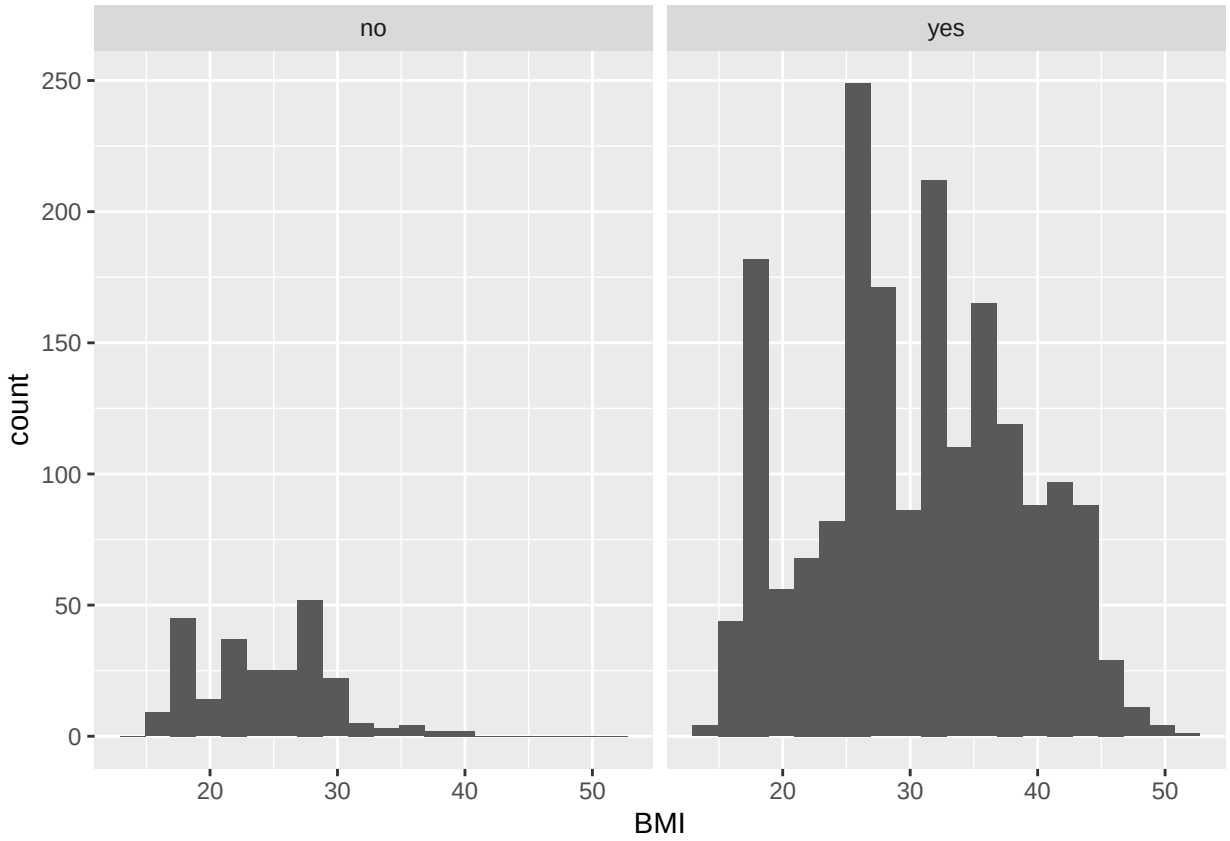
Obesity Risk Score: The aggregate score, with specific weights assigned to each factor, integrates these diverse elements into a comprehensive measure of obesity risk. The weights reflect the perceived impact of each factor on obesity risk, with significant factors like physical activity, family history, and BMI itself given slightly more influence

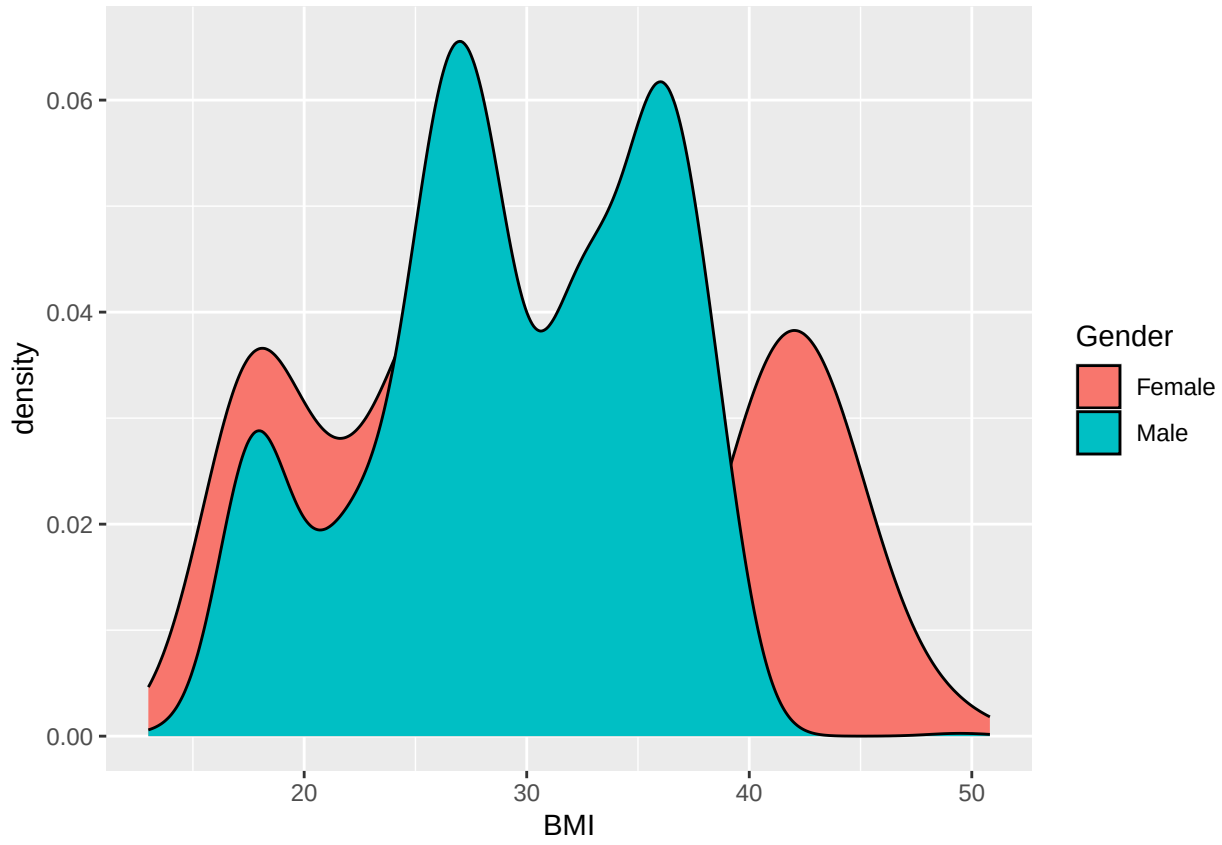
# Health Risk

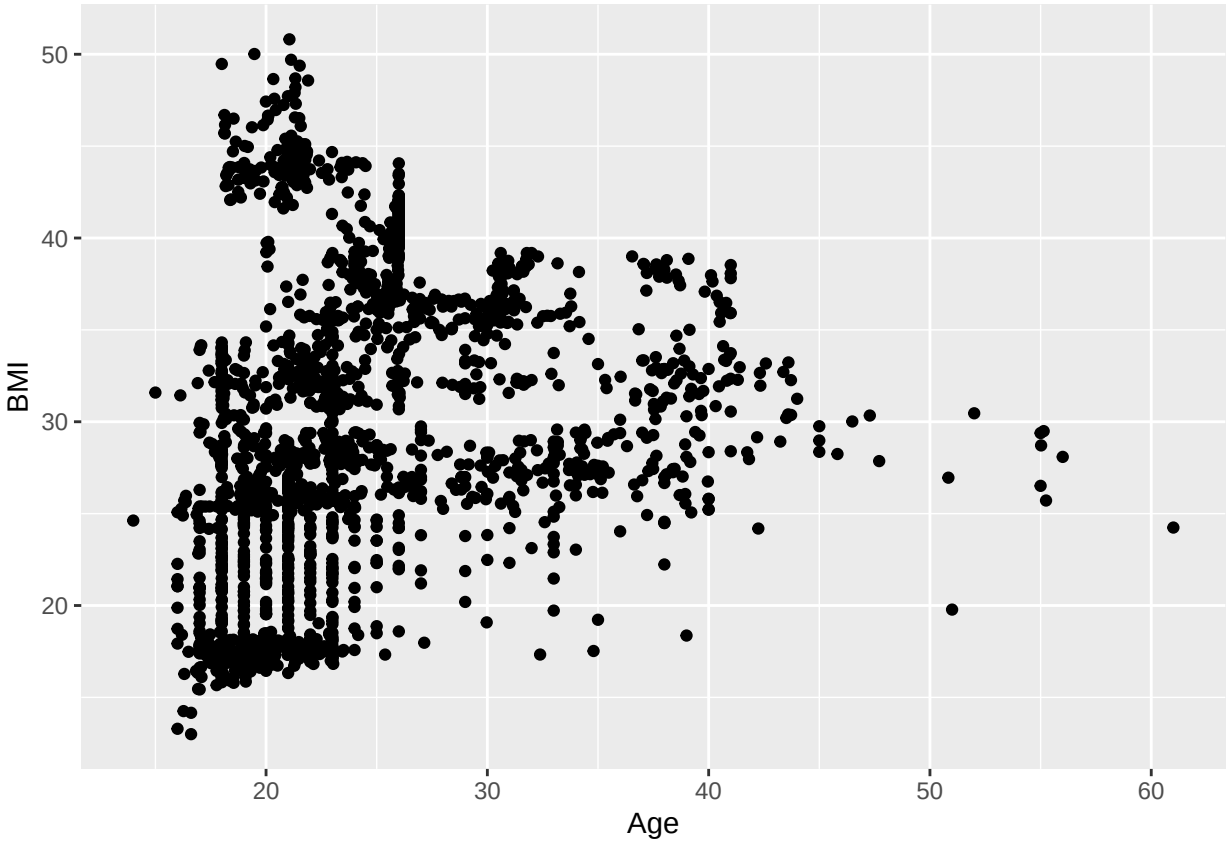
This system is specifically designed to assess health risks associated with high BMI, incorporating similar variables from the obesity risk system but adding alcohol consumption into the mix. The rationale behind each variable remains consistent, with a few adjustments in weighting to accommodate the new variable:

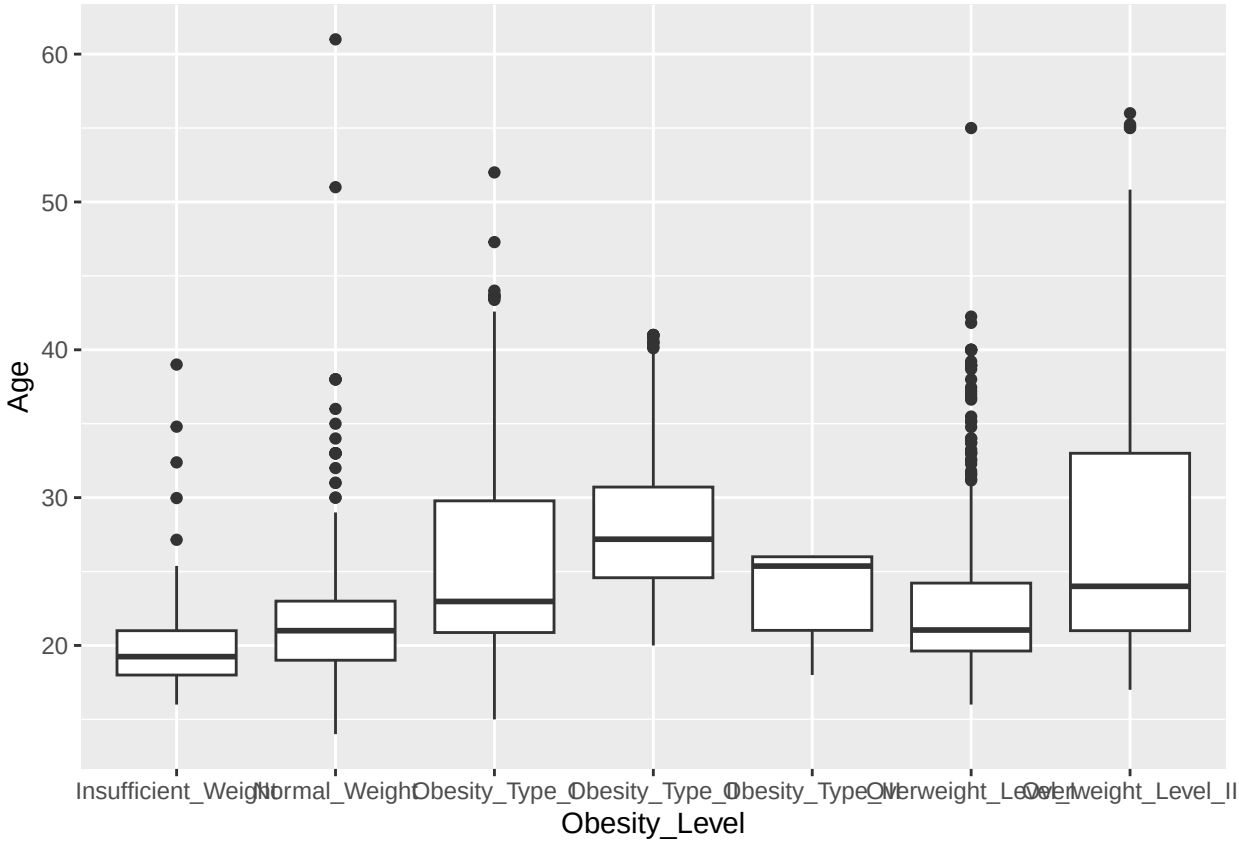
Alcohol Consumption is introduced as a factor with its score increasing with higher consumption levels due to the caloric intake and lifestyle impacts associated with alcohol use. The Health Risk Score similarly aggregates these scores, with adjustments in weights to balance the impact of each factor, including alcohol consumption. A higher cumulative score in this system also indicates a higher health risk related to high BMI.

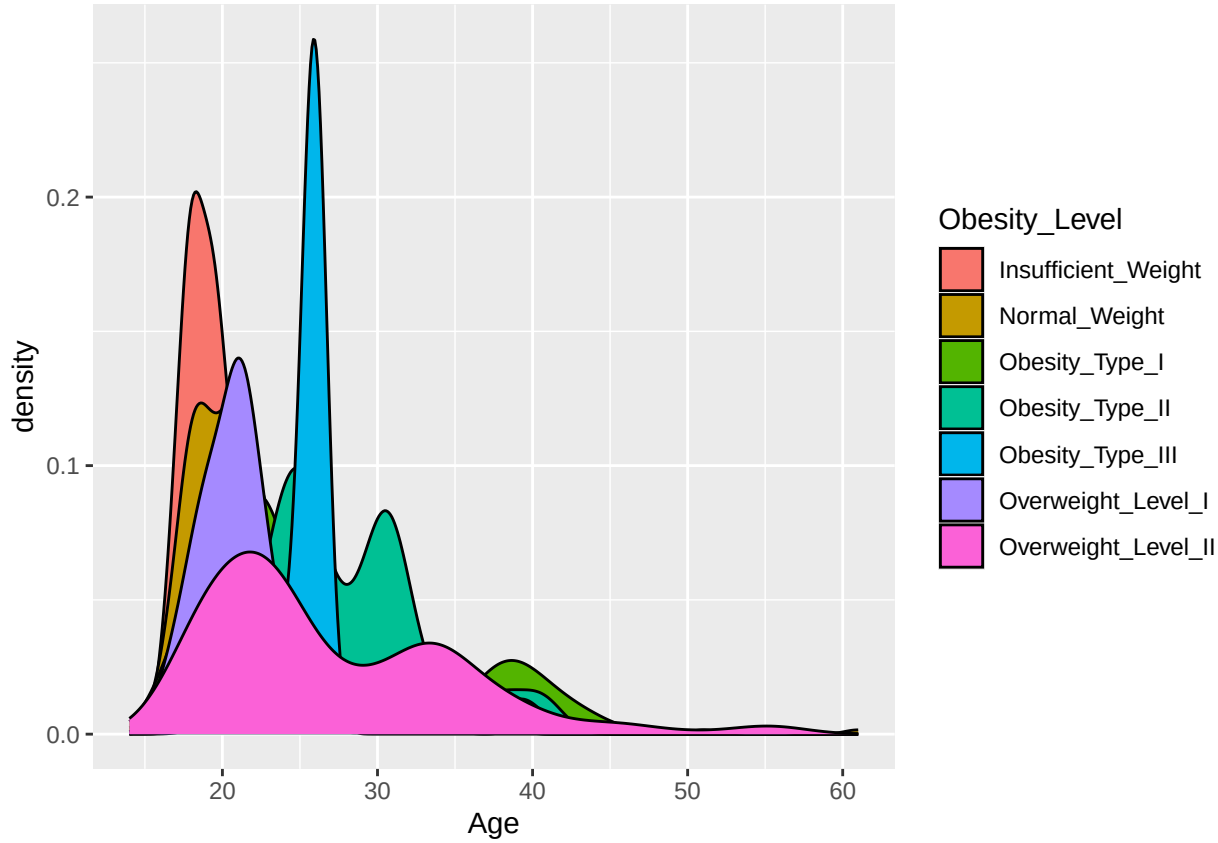

50
40
30
20
BMI
no
yes
HiCal_Food_Consump
HiCal_Food_Consump
no
yes

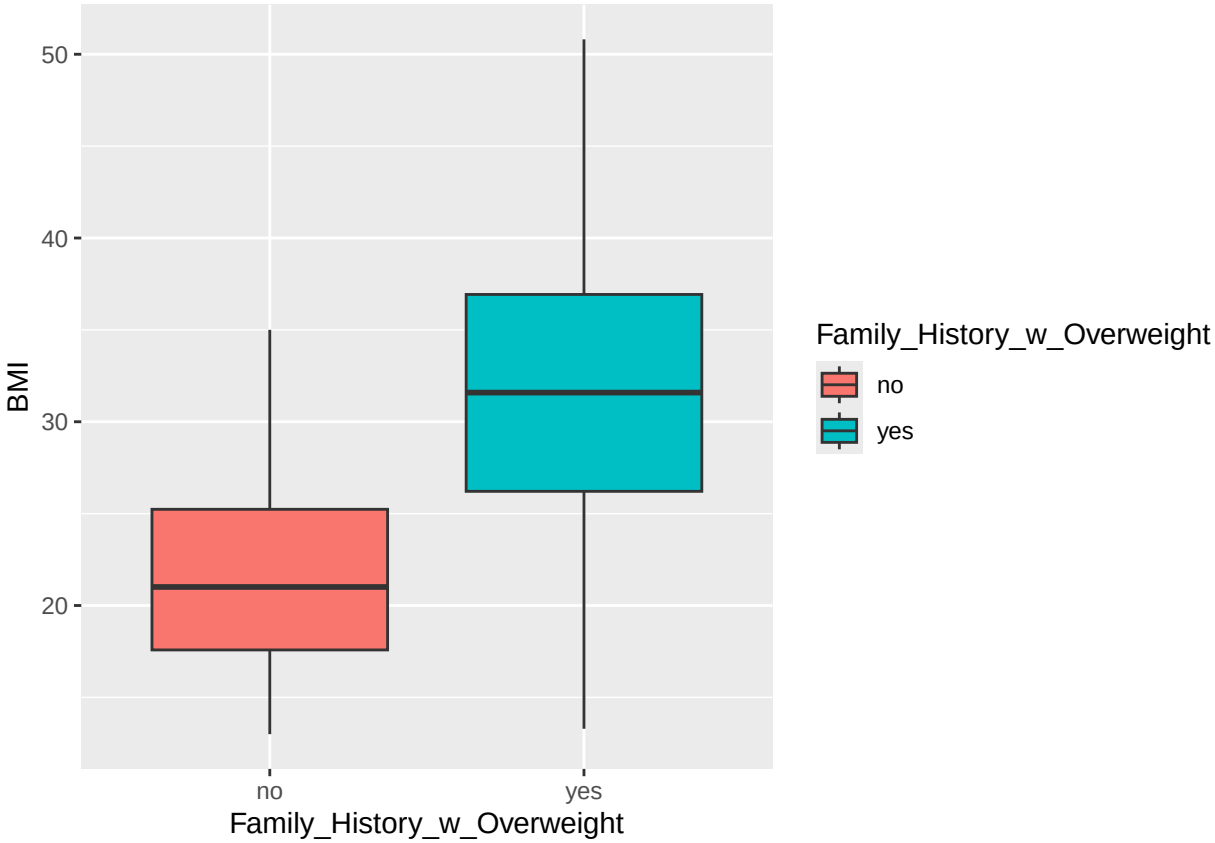

no
yes
250
200
150
100
50
0
count
20
30
40
50
BMI

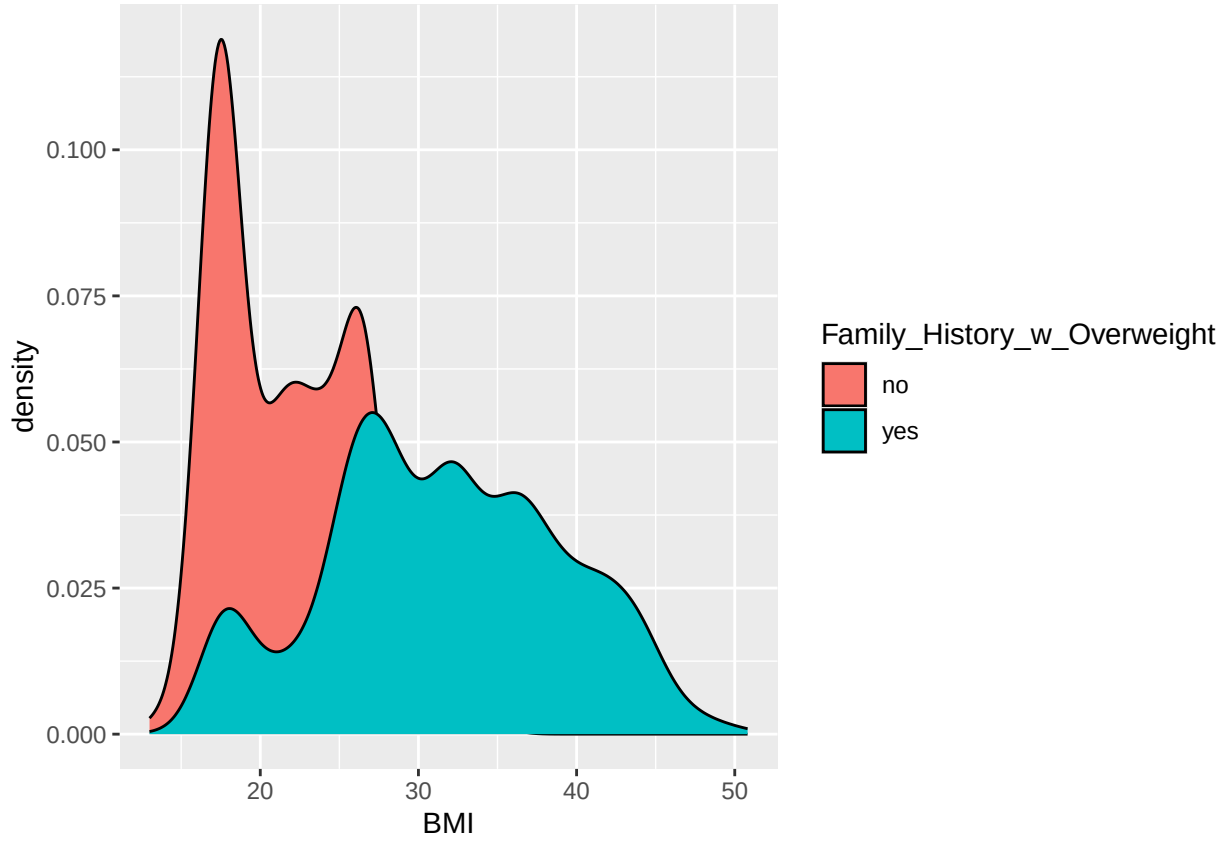

0.06
0.04
0.02
0.00
density
20
30
40
50
BMI
Gender
Female
Male


BMI
50
40
30
20
20
30
40
50
60
Age

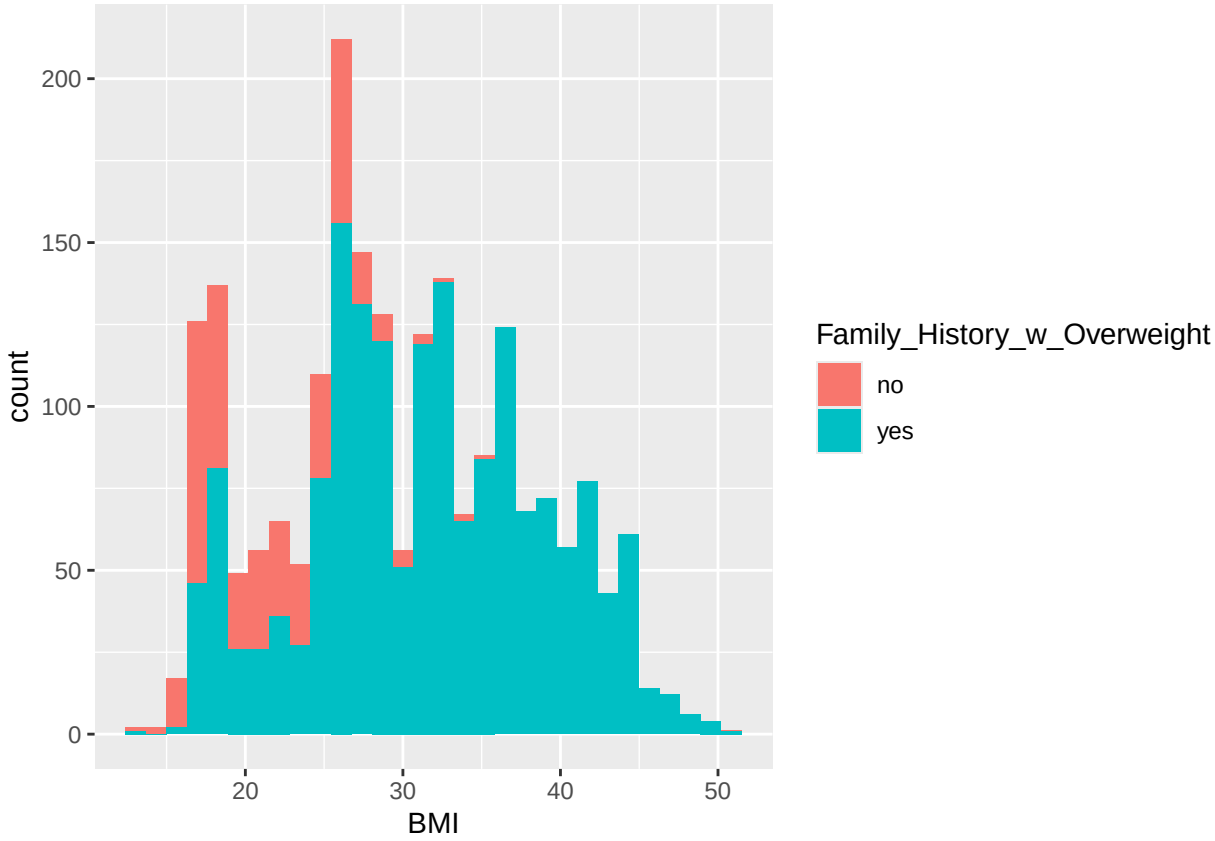

60
50
40
30
20
Age
Insufficient_Weight
Normal_Weight
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
Overweight_Level_I
Overweight_Level_II
Obesity_Level


density
0.2
0.1
0.0
20
30
40
50
60
Age
Obesity_Level
Insufficient_Weight
Normal_Weight
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
Overweight_Level_I
Overweight_Level_II


50
40
30
20
BMI
no
yes
Family_History_w_Overweight
Family_History_w_Overweight
no
yes




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




## Conclusions

While we were able to construct a good scoring system to rank the variables effect of BMI and overall health we also understand that the models we used were not fully accurate and thus our scoring systems contain flaws

First off, our BMI model was only at best 50% accurate as an individuals BMI involves much more variables and predictors than we had access to. In fact, the biggest predictor of weight, which is a key component in BMI, is calories consumed - burned, data we don't have access to.

Finally, the age range in out dataset wasn't large or diverse enough to display the impacts of BMI and the other habits on an individuals life expectancy. That being said, we understand that our Health rating score may include some inaccuracy as well.

## References