Group project - Final

2024-03-31

Intro Explain why we used BMI as scoring system

```
# Initialization of data
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.3

## Warning: package 'ggplot2' was built under R version 4.3.3

## Warning: package 'tibble' was built under R version 4.3.3

## Warning: package 'tidyr' was built under R version 4.3.3

## Warning: package 'readr' was built under R version 4.3.3

## Warning: package 'purrr' was built under R version 4.3.3

## Warning: package 'dplyr' was built under R version 4.3.3

## Warning: package 'stringr' was built under R version 4.3.3

## Warning: package 'forcats' was built under R version 4.3.3

## Warning: package 'lubridate' was built under R version 4.3.3

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(dplyr)
library(readr)
library(nnet)
```

```
## Warning: package 'nnet' was built under R version 4.3.3
```

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.3.3
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```r
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.3.3
```

```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
library(Metrics)
```

```
## Warning: package 'Metrics' was built under R version 4.3.3
```

```
##
## Attaching package: 'Metrics'
##
## The following objects are masked from 'package:caret':
##
##     precision, recall
```

```r
library(infer)
```

```
## Warning: package 'infer' was built under R version 4.3.3
```

```r
data = read.csv("obesity.csv")

data <- data %>%
  rename(
    Obesity_Level = NObeyesdad, Transportation_Use = MTRANS, Alcohol_Consump = CALC,
    Tech_Time = TUE, Physical_Activ_Amt = FAF, Monitor_Calories = SCC, Water_Consump = CH2O,
    Does_Smoke = SMOKE, Food_bw_Meals = CAEC, Main_Meal_Consump = NCP, Veggie_Consump = FCVC,
    HiCal_Food_Consump = FAVC, Family_History_w_Overweight = family_history_with_overweight
  )

# Add BMI variable
data <- data %>%
  mutate(BMI = Weight/(Height^2))
```

Introduction

Methods

We first made a linear regrssion model to find the significatn predictors of BMI from the relevant variables

```
#clean data for linear model
BMI_data <- select(data, -Weight, -Height, -Obesity_Level)
set.seed(99)
rownum <- sample(1:nrow(BMI_data), 2/3*nrow(BMI_data))

train <- BMI_data[rownum,]
test <- BMI_data[-rownum,]

BMI_model <- lm(BMI ~., data = train)

predict_BMI <- predict(BMI_model, newdata = test)


summary(BMI_model)
```

```
##
## 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
```

```
Metrics::rmse(test$BMI, predict_BMI)
```

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

We see our R^2 is .49 which means olf half the variability of BMI is captured by this data, this is normal for a dataset of this nature that deals with predicting humans. this is also beacuse our data is missign major predictors of obesity

Based on this we can see the most significant predictors for BMI

we can also find out what predicotrs have effect on level of obesity using multiple logistic regression

```
#reclean
level_data <- select(data, -Weight, -Height, -BMI)
#factorize

level_data$Gender <- as.factor(level_data$Gender)
level_data$Family_History_w_Overweight <- as.factor(level_data$Family_History_w_Overweight)
level_data$Obesity_Level <- as.factor(level_data$Obesity_Level)
level_data$HiCal_Food_Consump <- as.factor(level_data$HiCal_Food_Consump)
level_data$Veggie_Consump <- as.factor(level_data$Veggie_Consump)
level_data$Main_Meal_Consump <- as.factor(level_data$Main_Meal_Consump)
level_data$Food_bw_Meals <- as.factor(level_data$Food_bw_Meals)
level_data$Does_Smoke <- as.factor(level_data$Does_Smoke)
```

```r
level_data$Water_Consump <- as.factor(level_data$Water_Consump)
level_data$Monitor_Calories <- as.factor(level_data$Monitor_Calories)
level_data$Physical_Activ_Amt <- as.factor(level_data$Physical_Activ_Amt)
level_data$Tech_Time <- as.factor(level_data$Tech_Time)
level_data$Alcohol_Consump <- as.factor(level_data$Alcohol_Consump)
level_data$Transportation_Use <- as.factor(level_data$Transportation_Use)
#train and test data
rownum <- sample(1:nrow(level_data), 2/3*nrow(level_data))

train <- level_data[rownum,]
test <- level_data[-rownum,]
#multiple logistic regression
nnet_model <- multinom(Obesity_Level ~ Family_History_w_Overweight + Food_bw_Meals + HiCal_Food_Consump
```

```
## # weights:  112 (90 variable)
## initial  value 2737.895580
## iter  10 value 2257.213442
## iter  20 value 2060.186606
## iter  30 value 2027.133071
## iter  40 value 2018.698339
## iter  50 value 2016.432324
## iter  60 value 2015.957457
## iter  70 value 2015.910367
## final  value 2015.908640
## converged
```

```r
# Make predictions on test data
nnet_pred <- predict(nnet_model, newdata = test, type = "class")

# Evaluate model performance
accuracy <- mean(nnet_pred == test$Obesity_Level)
accuracy
```

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

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

based on results of first model (regression) we can test the impact or dependacy betweeen predictors and target variables using KW test for BMI and Chi square for Obesity Level

```r
library(infer)
kruskal.test(BMI ~ Family_History_w_Overweight, data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  BMI by Family_History_w_Overweight
## Kruskal-Wallis chi-squared = 524, df = 1, p-value < 2.2e-16
```

```r
kruskal.test(BMI ~ Veggie_Consump, data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  BMI by Veggie_Consump
## Kruskal-Wallis chi-squared = 846.08, df = 809, p-value = 0.1776
```

```r
kruskal.test(BMI ~ HiCal_Food_Consump, data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  BMI by HiCal_Food_Consump
## Kruskal-Wallis chi-squared = 131.06, df = 1, p-value < 2.2e-16
```

```r
kruskal.test(BMI ~ Physical_Activ_Amt, data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  BMI by Physical_Activ_Amt
## Kruskal-Wallis chi-squared = 1504.1, df = 1189, p-value = 1.119e-09
```

```r
kruskal.test(BMI ~ Food_bw_Meals,data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  BMI by Food_bw_Meals
## Kruskal-Wallis chi-squared = 409.66, df = 3, p-value < 2.2e-16
```

```r
kruskal.test(BMI ~ Main_Meal_Consump, data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  BMI by Main_Meal_Consump
## Kruskal-Wallis chi-squared = 685.09, df = 634, p-value = 0.07831
```

```r
kruskal.test(BMI ~ Water_Consump, data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  BMI by Water_Consump
## Kruskal-Wallis chi-squared = 1602.4, df = 1267, p-value = 3.528e-10
```

```r
kruskal.test(BMI ~ Monitor_Calories, data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  BMI by Monitor_Calories
## Kruskal-Wallis chi-squared = 85.414, df = 1, p-value < 2.2e-16
```

```r
kruskal.test(BMI ~ Transportation_Use, data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  BMI by Transportation_Use
## Kruskal-Wallis chi-squared = 46.479, df = 4, p-value = 1.958e-09
```

```r
kruskal.test(BMI ~ Tech_Time, data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  BMI by Tech_Time
## Kruskal-Wallis chi-squared = 1504.3, df = 1128, p-value = 3.145e-13
```

```r
kruskal.test(BMI ~ Age, data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  BMI by Age
## Kruskal-Wallis chi-squared = 1809.3, df = 1401, p-value = 7.063e-13
```

```r
#we also test effect of bmi and obesity level on age
kruskal.test(Age ~ BMI, data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Age by BMI
## Kruskal-Wallis chi-squared = 2031.8, df = 1967, p-value = 0.1509
```

```r
kruskal.test(Age ~ Obesity_Level, data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Age by Obesity_Level
## Kruskal-Wallis chi-squared = 541.77, df = 6, p-value < 2.2e-16
```

```r
#we can see the while BMI itself doesnt affect age, the obesity level does

#if the tests indicate a significant differece in BMI across groups (low P) we move forward with Chisqu


chisq_test(level_data, formula = Obesity_Level ~ Family_History_w_Overweight)
```

```
## # A tibble: 1 x 3
##   statistic chisq_df   p_value
##       <dbl>    <int>     <dbl>
## 1      622.        6 4.23e-131
```

```r
chisq_test(level_data, formula = Obesity_Level ~ HiCal_Food_Consump)
```

```
## # A tibble: 1 x 3
##   statistic chisq_df  p_value
##       <dbl>    <int>    <dbl>
## 1      233.        6 1.48e-47
```

```r
chisq_test(level_data, formula = Obesity_Level ~ Physical_Activ_Amt)
```

```
## Warning in stats::chisq.test(table(x), ...): Chi-squared approximation may be
## incorrect
```

```
## # A tibble: 1 x 3
##   statistic chisq_df    p_value
##       <dbl>    <int>      <dbl>
## 1     7678.     7134 0.00000428
```

```r
chisq_test(level_data, formula = Obesity_Level ~ Food_bw_Meals)
```

```
## # A tibble: 1 x 3
##   statistic chisq_df   p_value
##       <dbl>    <int>     <dbl>
## 1      803.       18 7.38e-159
```

```r
chisq_test(level_data, formula = Obesity_Level ~ Water_Consump)
```

```
## Warning in stats::chisq.test(table(x), ...): Chi-squared approximation may be
## incorrect
```

```
## # A tibble: 1 x 3
##   statistic chisq_df p_value
##       <dbl>    <int>   <dbl>
## 1     7955.     7602 0.00236
```

```r
chisq_test(level_data, formula = Obesity_Level ~ Monitor_Calories)
```

```
## # A tibble: 1 x 3
##   statistic chisq_df  p_value
##       <dbl>    <int>    <dbl>
## 1      123.        6 3.77e-24
```

```r
chisq_test(level_data, formula = Obesity_Level ~ Transportation_Use)
```

```
## Warning in stats::chisq.test(table(x), ...): Chi-squared approximation may be
## incorrect
```

```
## # A tibble: 1 x 3
##   statistic chisq_df  p_value
##       <dbl>    <int>    <dbl>
## 1      293.       24 5.18e-48
```

```r
chisq_test(level_data, formula = Obesity_Level ~ Tech_Time)
```

```
## Warning in stats::chisq.test(table(x), ...): Chi-squared approximation may be
## incorrect
```

```
## # A tibble: 1 x 3
##   statistic chisq_df p_value
##       <dbl>    <int>   <dbl>
## 1     7119.     6768 0.00147
```

```r
##all passed chiquare test to indicate the depenacy or relationship between them and obesity level is n
```

```r
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 is 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 is not significant. However we can use 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 i will find the probability an individual is over 40 years old if their BMI is greater than 35 (Obesity type 2)

```
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
```

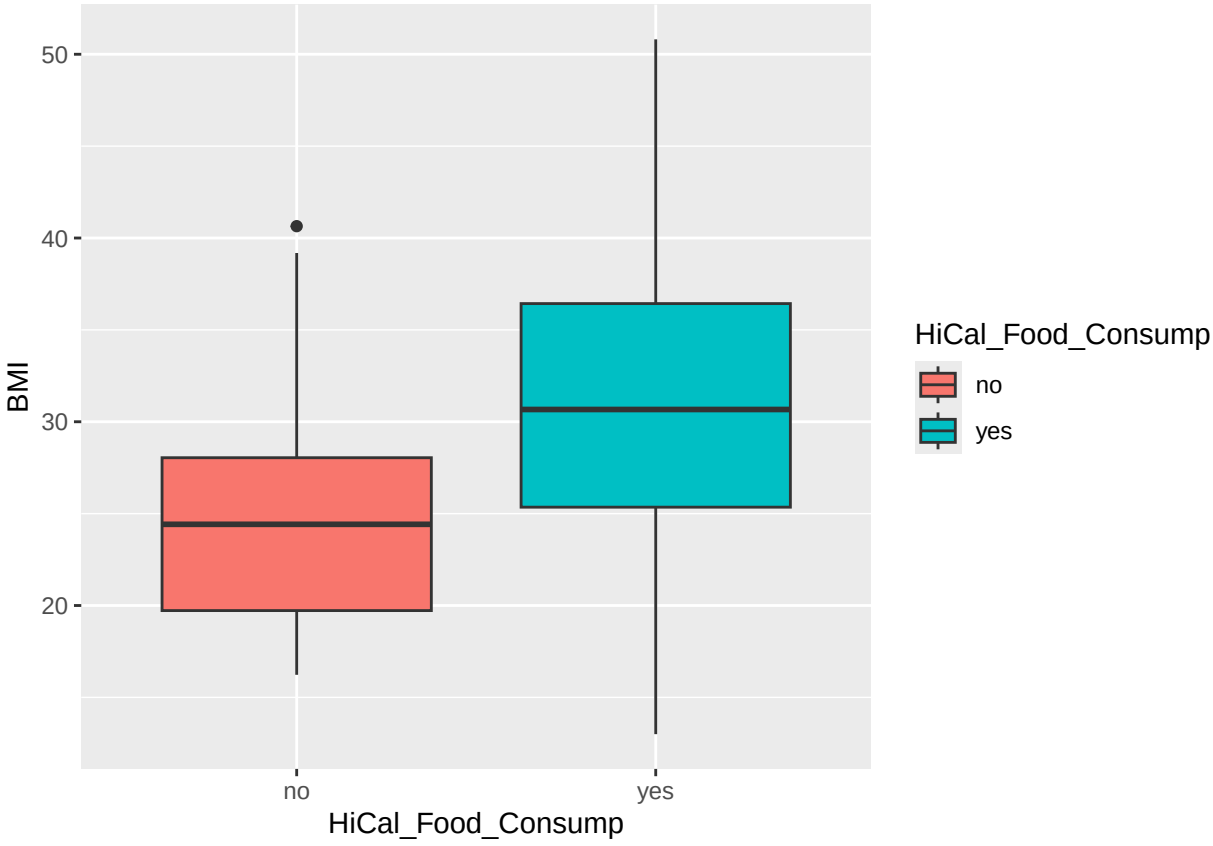
```
## [1] 0.01980339
```

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 likelylihood.

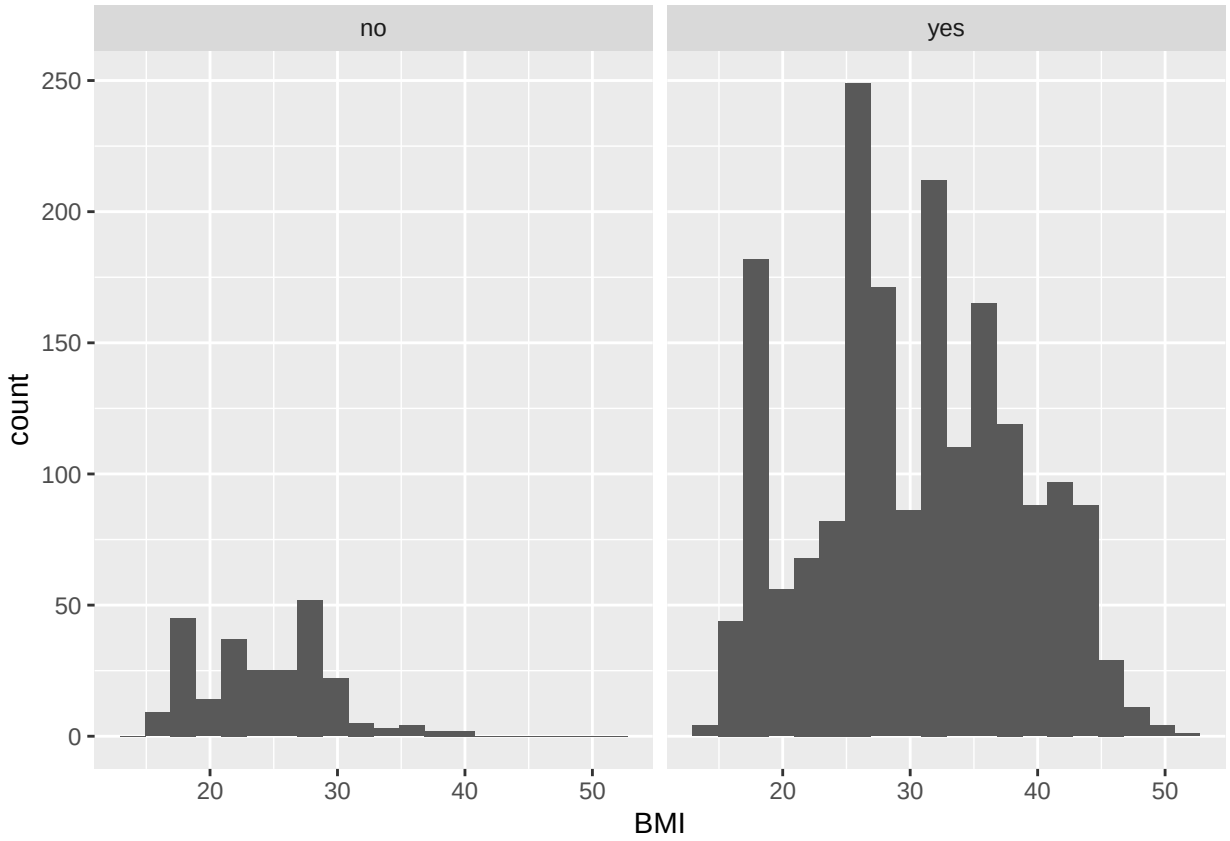
Results

Our results show what variables are predictors of BMI and Obesity group, we will use this info to construct a scoring system based on obesity risk. Our result also point to the fact that Age is significantly influence by an individuals Obesity level especially Obesity Type 2 and 3 we will use this info to constuct a health risk scoring system which will use the variables impact on an individuals age

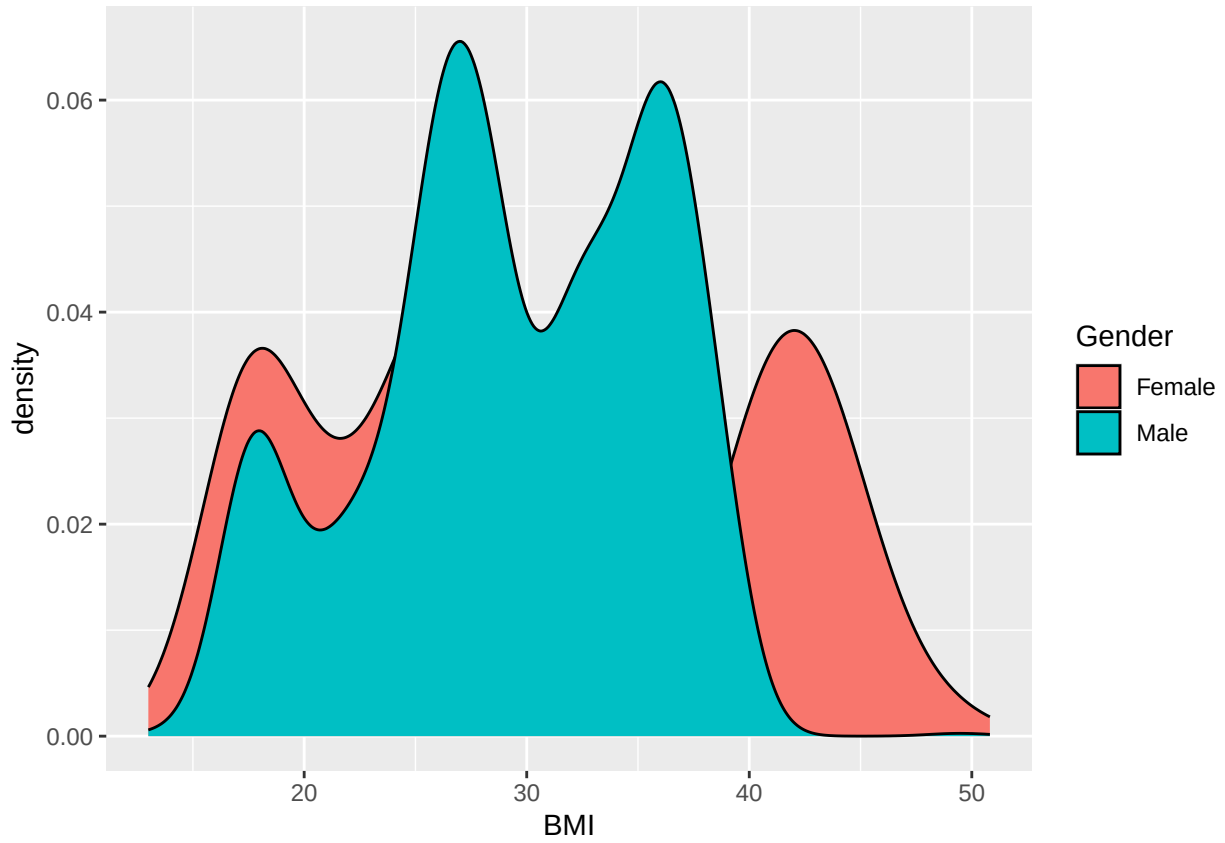
```
# BMI vs HiCal_Food_Consump
data %>%
  ggplot(aes(x = HiCal_Food_Consump, y = BMI, fill = HiCal_Food_Consump)) +
  geom_boxplot()
```



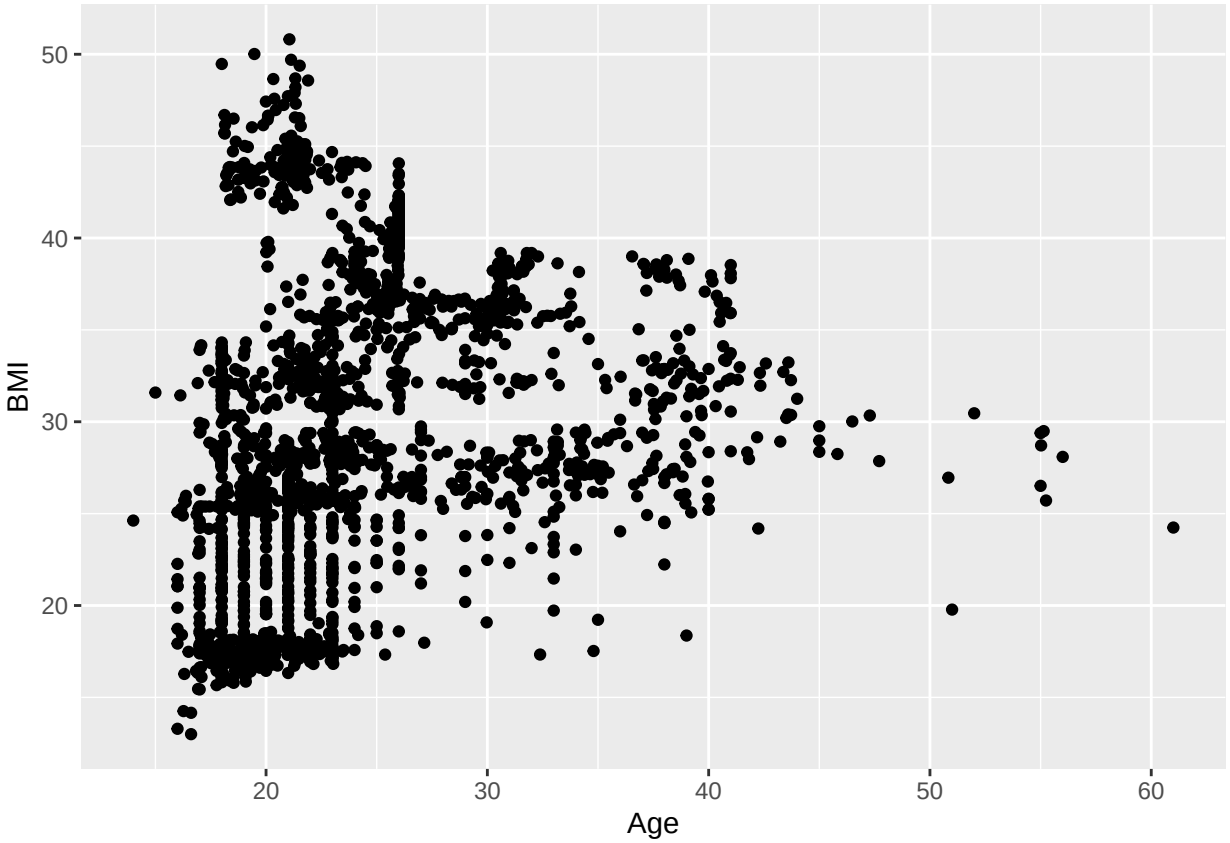

```
ggplot(data, aes(x = BMI)) + geom_histogram(bins=20) + facet_grid(~HiCal_Food_Consump)
```



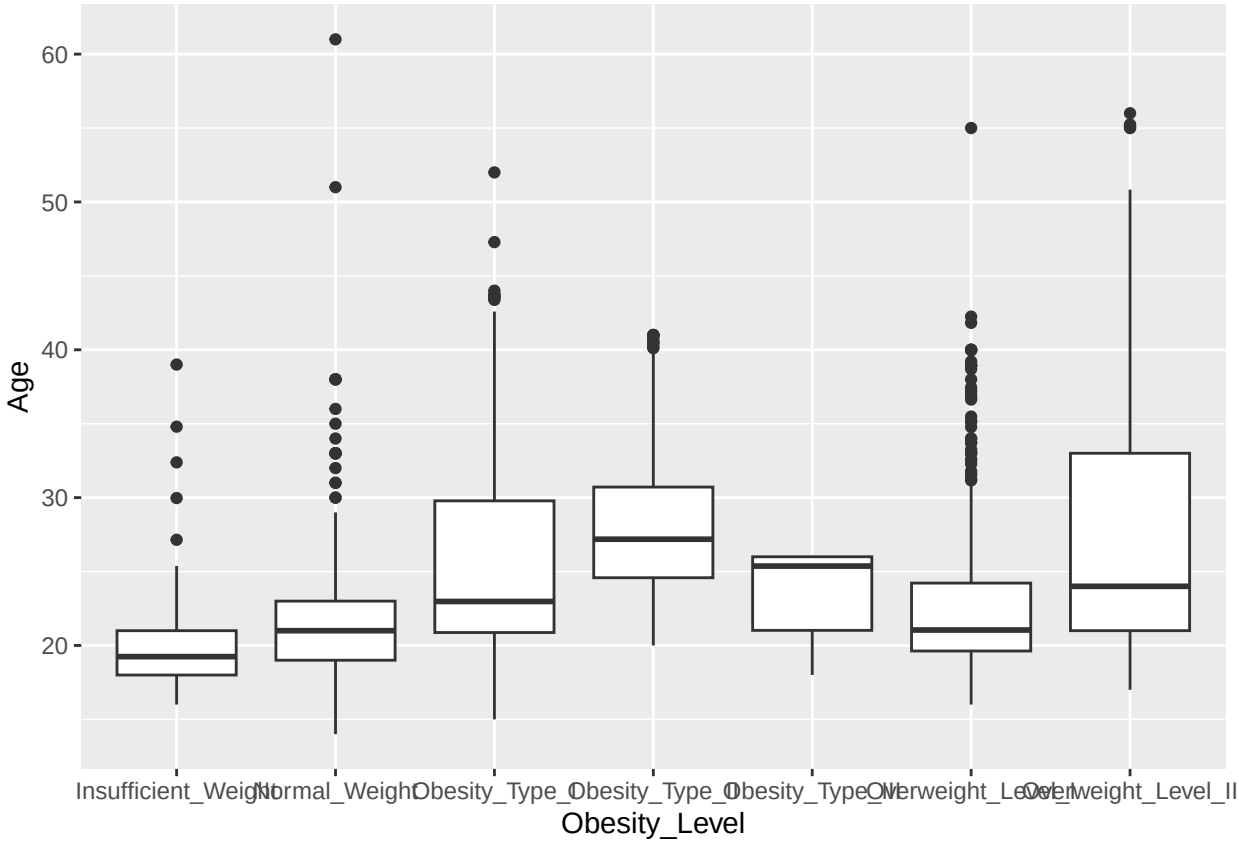

```
# BMI vs Gender
data %>% ggplot(aes(x=BMI, fill=Gender)) + geom_density()
```



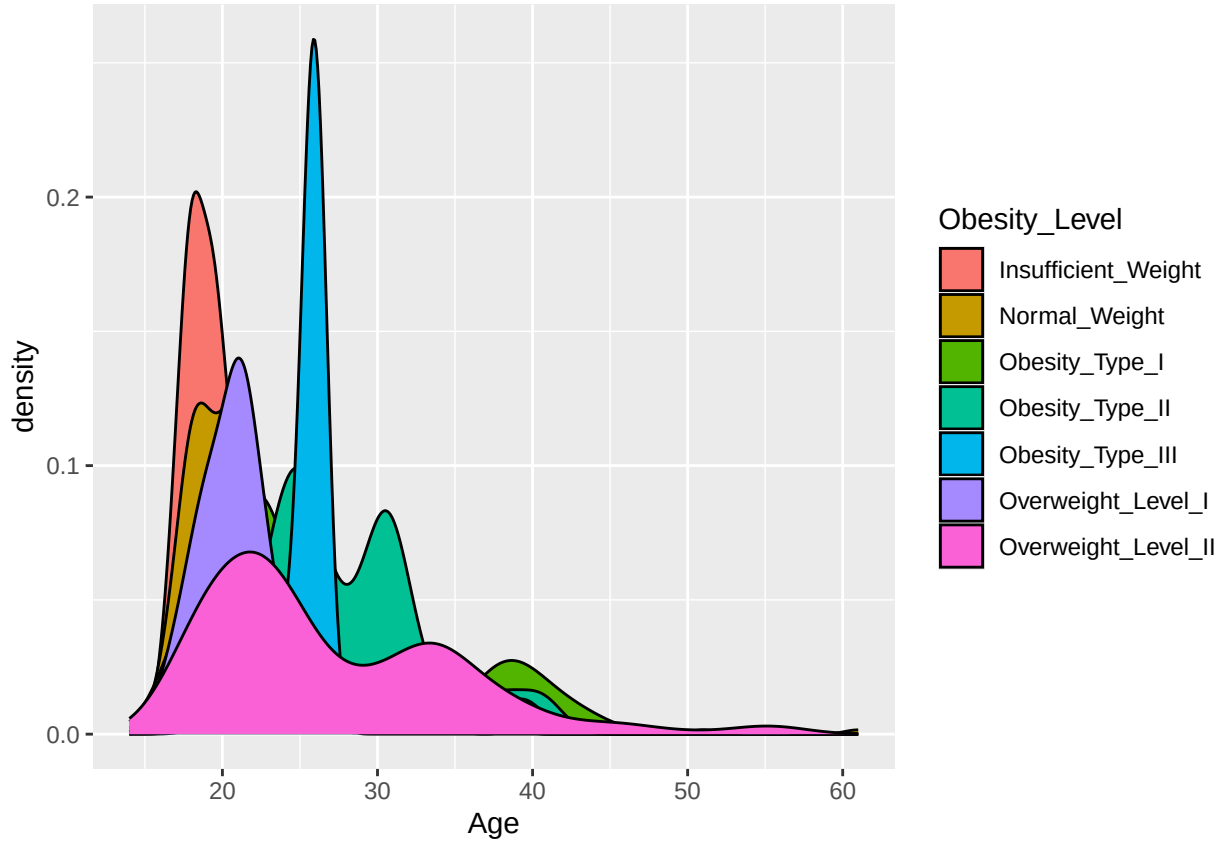

```
# BMI vs Age - Younger people tend to have the most variation in BMI
data %>% ggplot(aes(x=Age, y=BMI)) + geom_point()
```



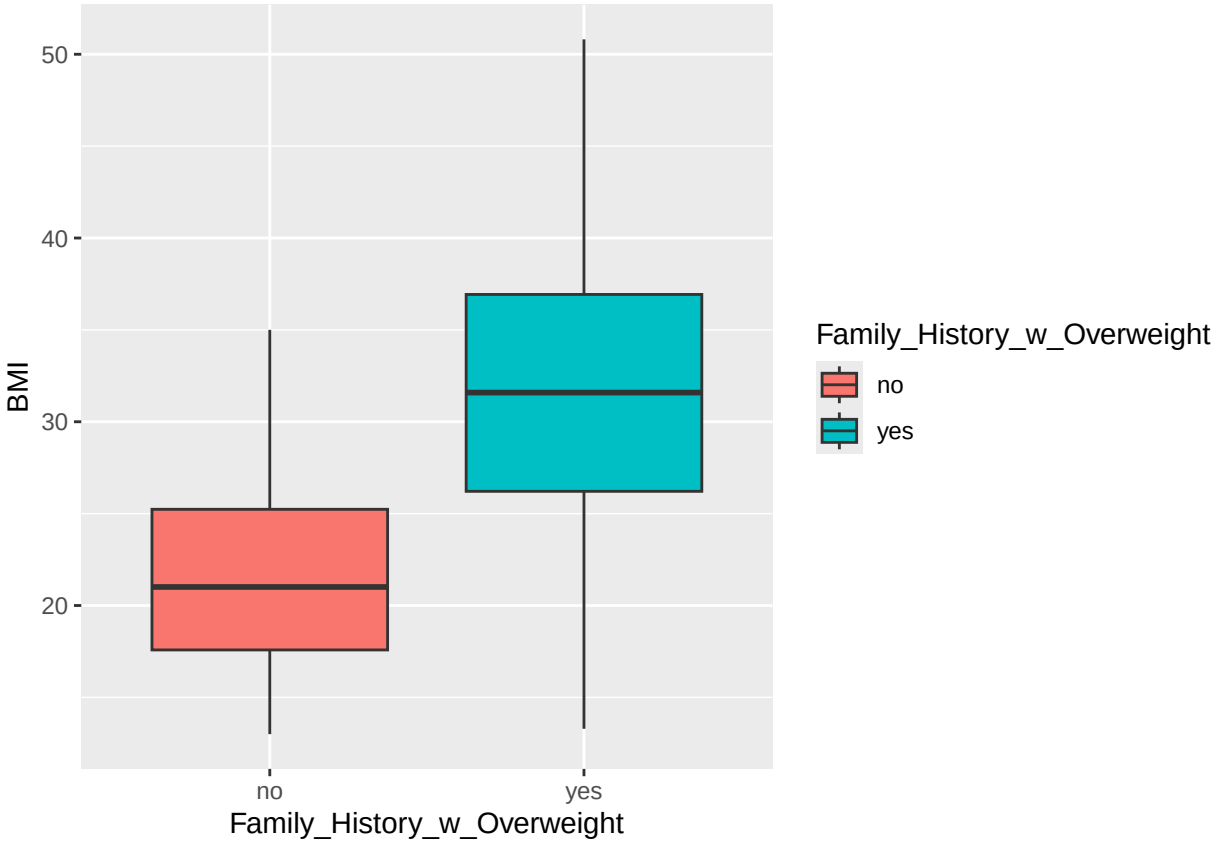

```
# Obesity_Level vs Age
data %>% ggplot(aes(x=Obesity_Level,y=Age)) + geom_boxplot()
```



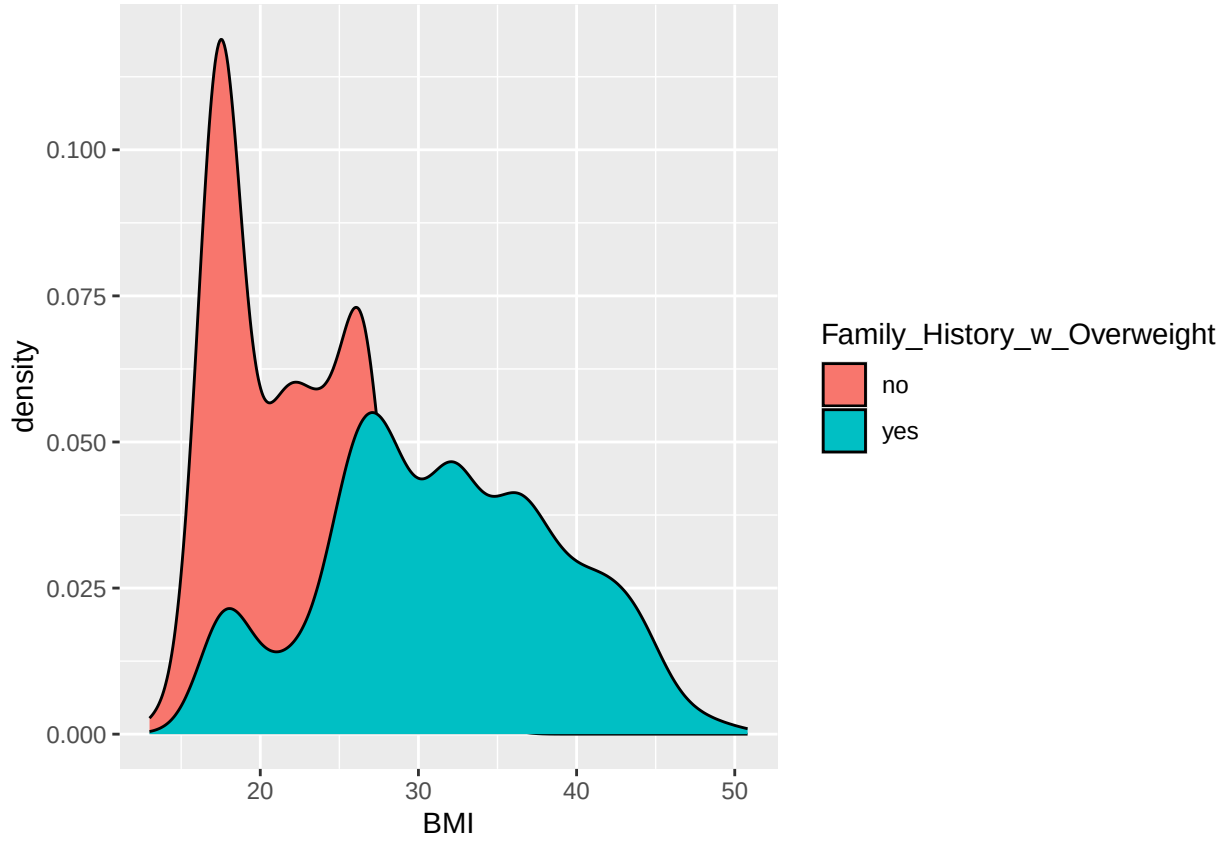

```
data %>% ggplot(aes(fill=Obesity_Level,x=Age)) + geom_density()
```




```
# BMI vs Family history
data %>%
  ggplot(aes(x = Family_History_w_Overweight, y = BMI, fill = Family_History_w_Overweight)) +
  geom_boxplot()
```

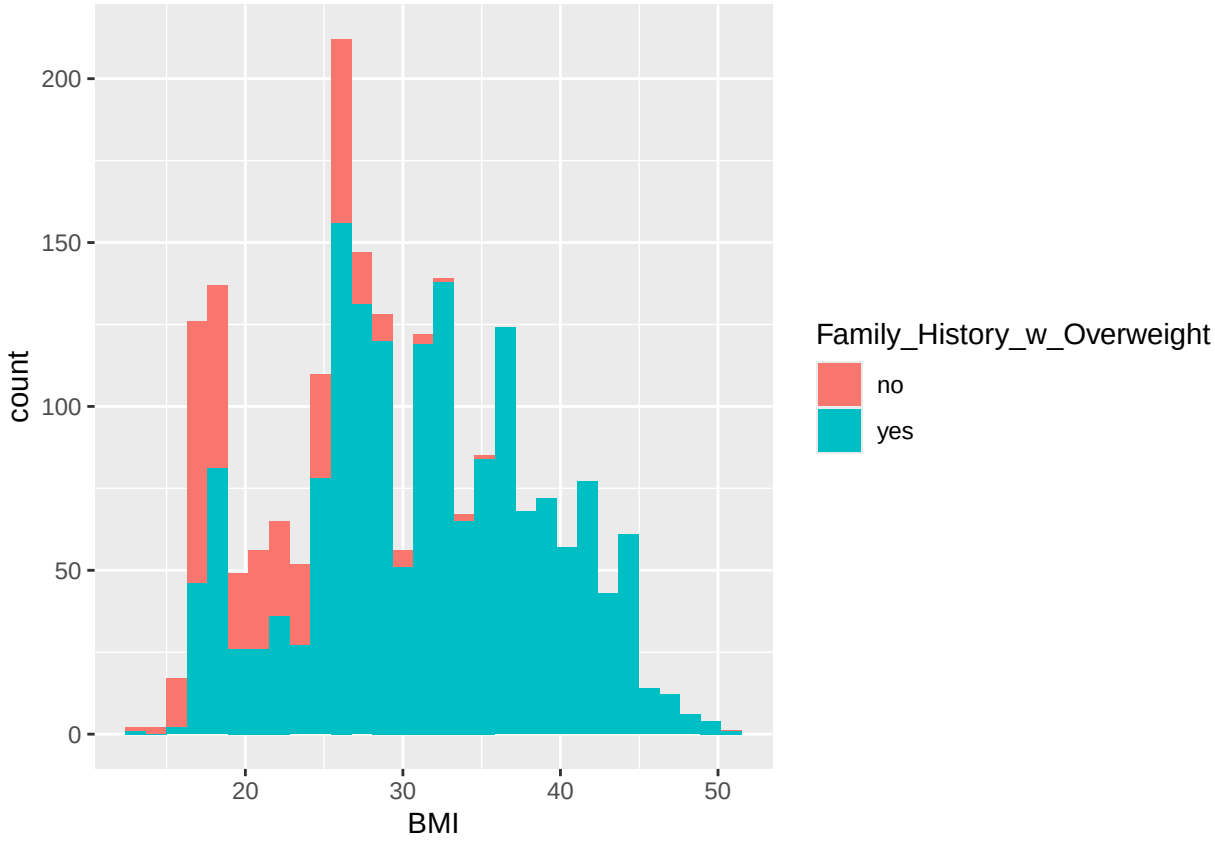



```
data %>% ggplot(aes(x=BMI,fill=Family_History_w_Overweight)) + geom_density()
```




```
data %>% ggplot(aes(x=BMI, fill=Family_History_w_Overweight)) + geom_histogram()
```

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



Conclusions

While we were able to constuct a good scoring system to rank the varaibles effect of BMI and overall health we also understand that the models we used were not fully acurate and thus our scoring systems contain flaws first our BMI model was only at best 50% accurate as an individuals BMI involves much more variables and predictors than we have access to, in fact, the biggest predictor of weight, which is a key component in BMI is calories consumed - burned, data we dont have access to.

secondly, the age range in out dataset wasnt large or diverse enough to display the impacts of BMI and the other habits on an individuals life expectancy. as such we understand that our Health rating score may include some inaccuracy as well.

References