Group project - Final

2024-03-31

Intro Explain why we used BMI as scoring system

Introduction

Methods

We first made a linear regression model to find the significant predictors of BMI from the relevant 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
```

We see our R^2 is .49 which means only 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 because our data is missing major predictors of obesity like calorie intake

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

We can also find out what predictors have effect on level of obesity using multiple logistic regression

```
accuracy
```

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

the 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 for an individual

based on results of first model (regression) we can test the impact/dependency between predictors and target variables using Kruskal-Wallis test for BMI and Chi-Square for Obesity Level

we have now established 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

next we will determine how much obesity level impacts age by performing a logistic regression predicting age using BMI. to avoid skewing the data, as BMI can often be misinterpreted for younger and developing individuals, we will take 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 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 we 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
```

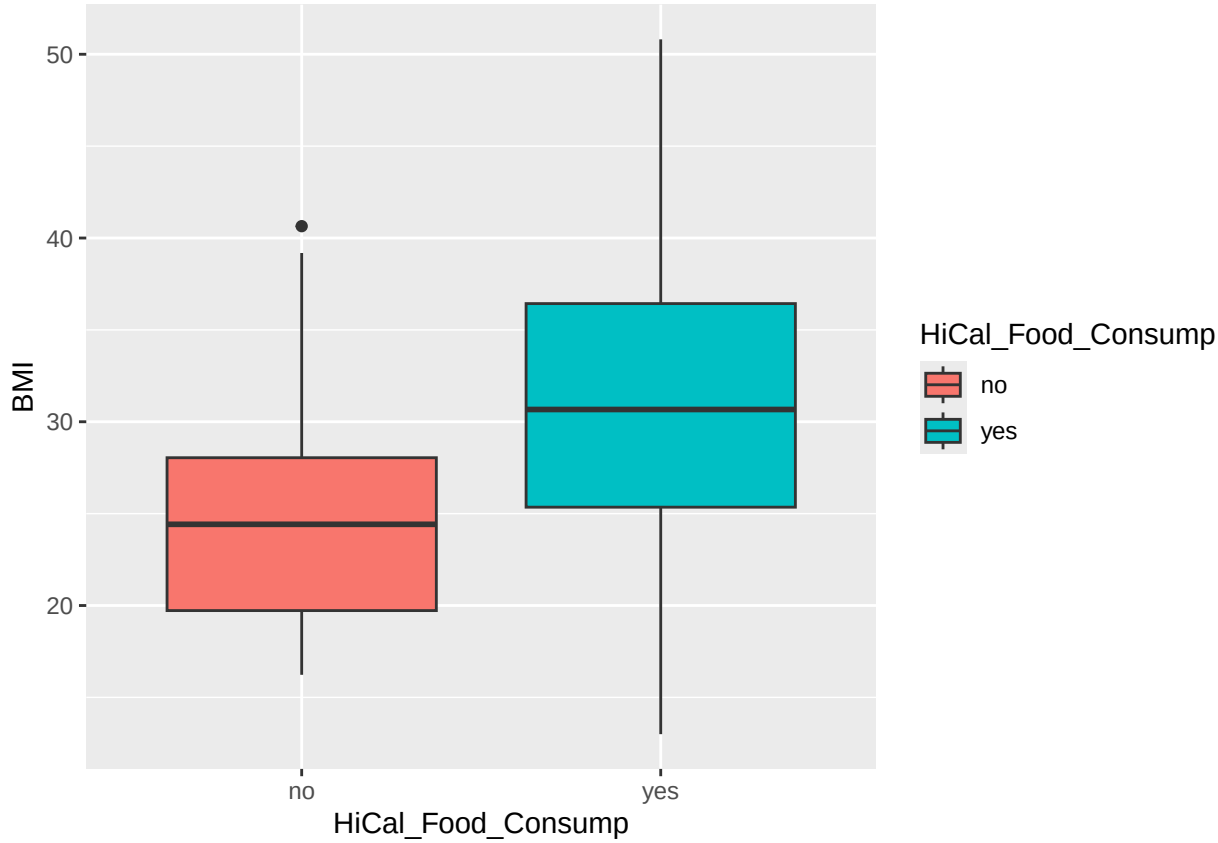
```r
# Calculate log odds
log_odds <- intercept + BMI35_coefficient * BMI_over_35

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

# Print the result
probability_over_40
```

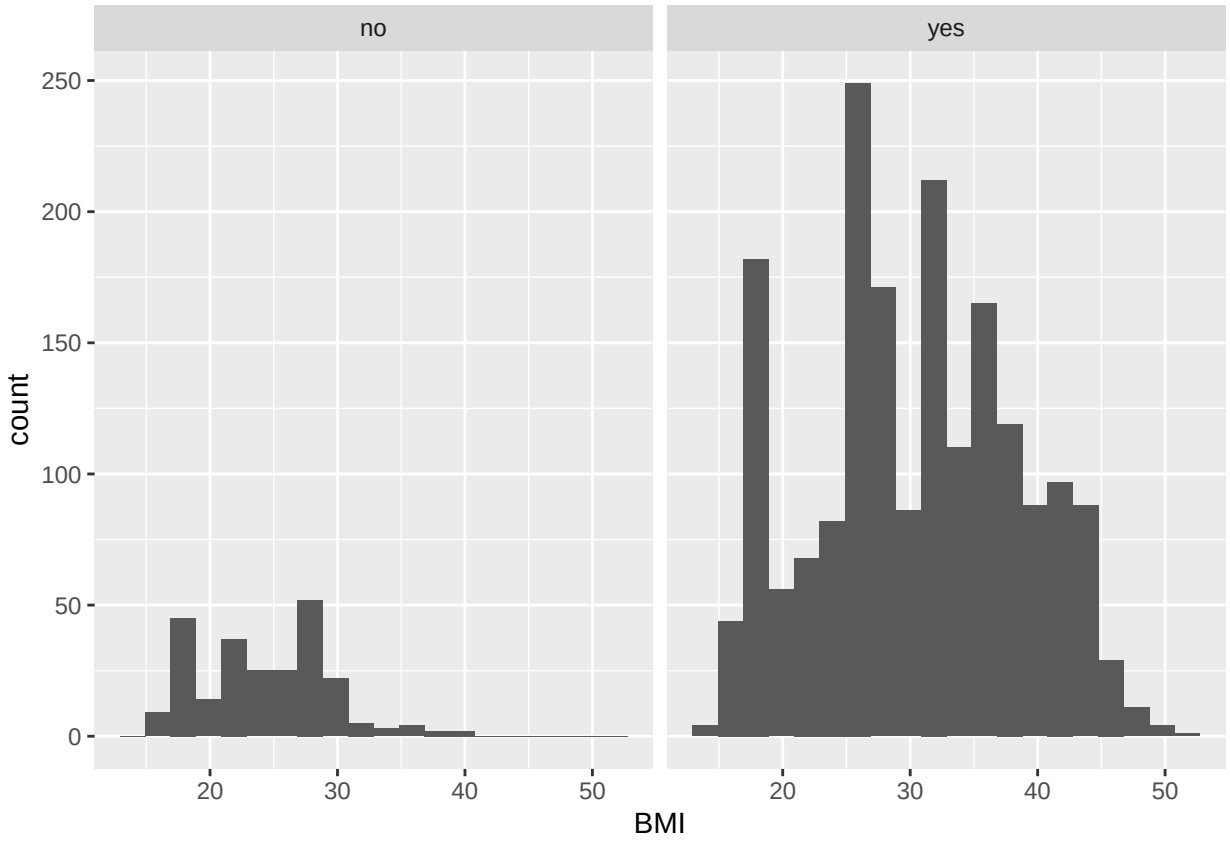
```
## [1] 0.01980339
```

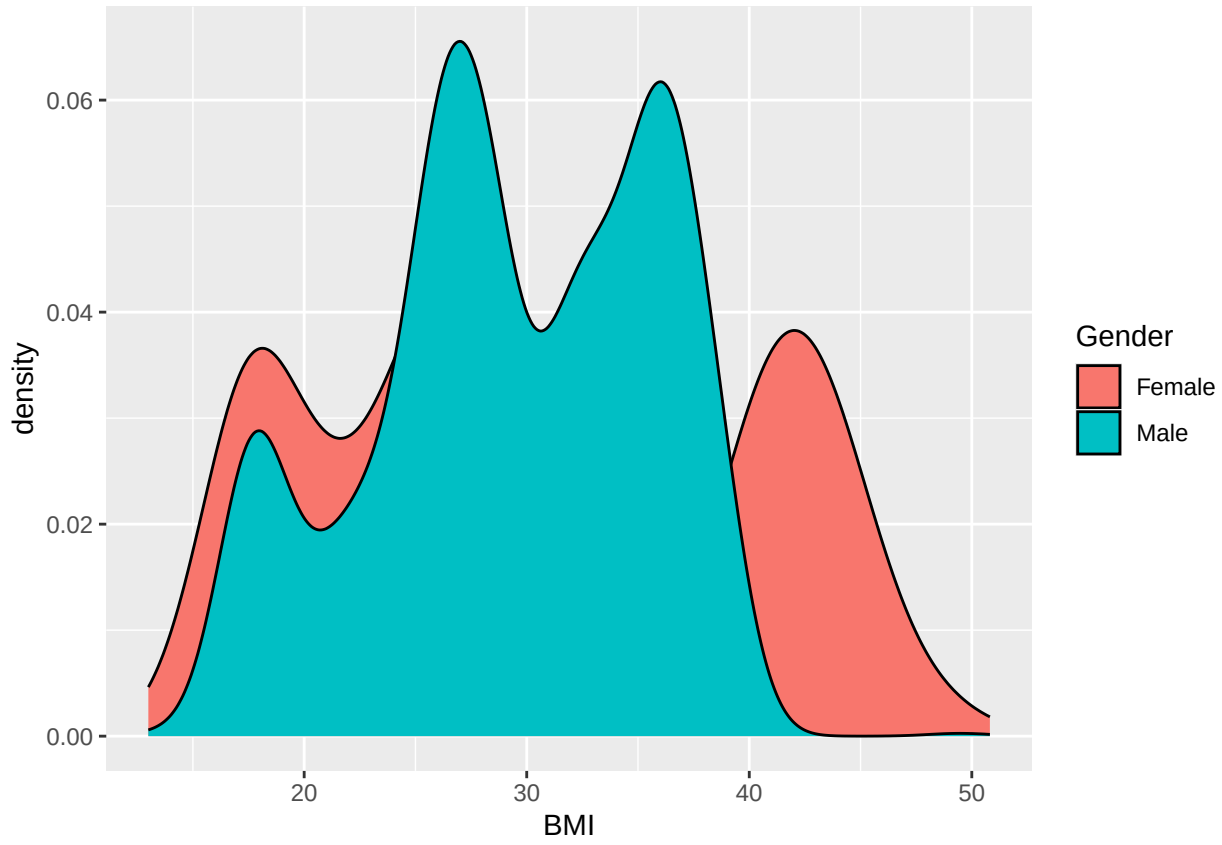
As you can see, the probability that an individual is older than 40 if their BMI is >35 is around 1% which is very low likelylihood. this suggests some form of missing data which we believe is not random as our results show that the higher in age, the lower the instances of very high BMI. In fact the maximum BMI observed decreases inversly with age. we exrapolate this data by assuming that this is caused due to health complications that we will attribute to obesity, and we will uses this as the basis for our health scoring system

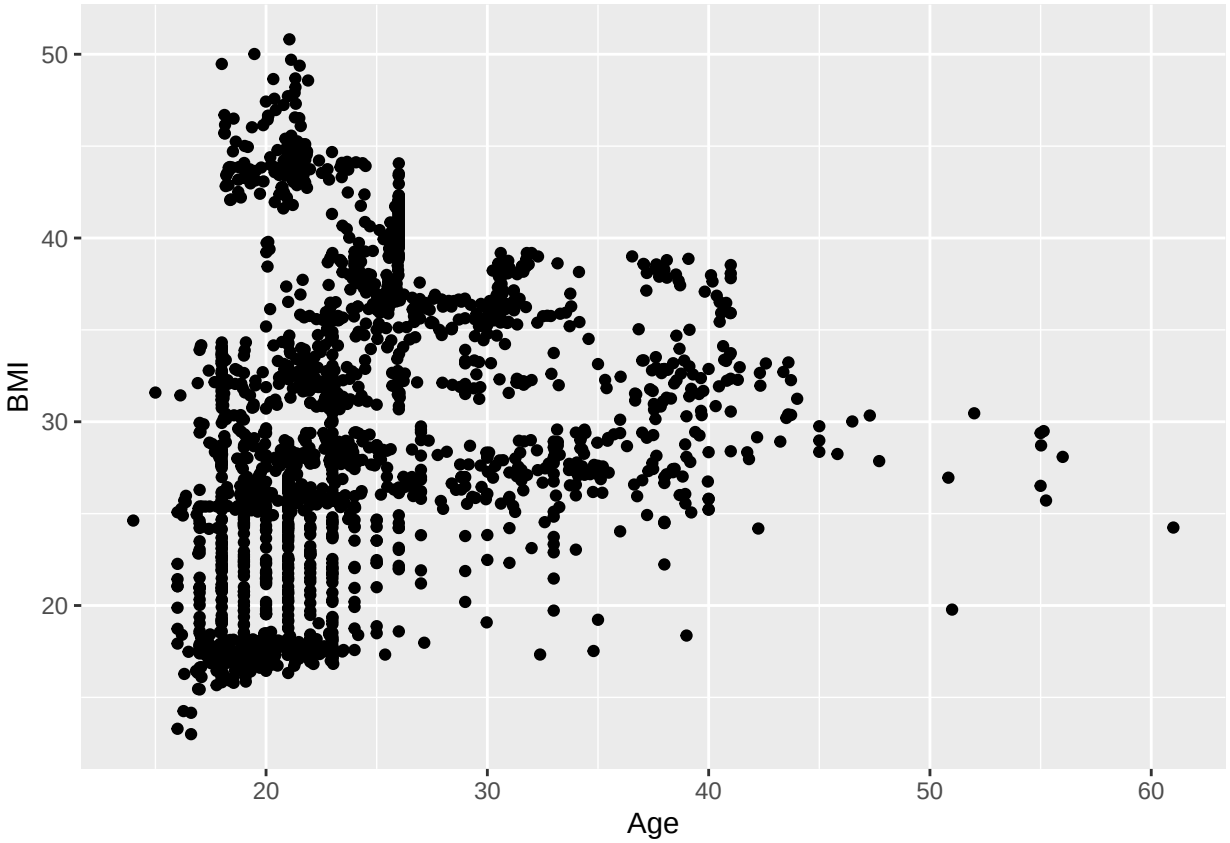
Results

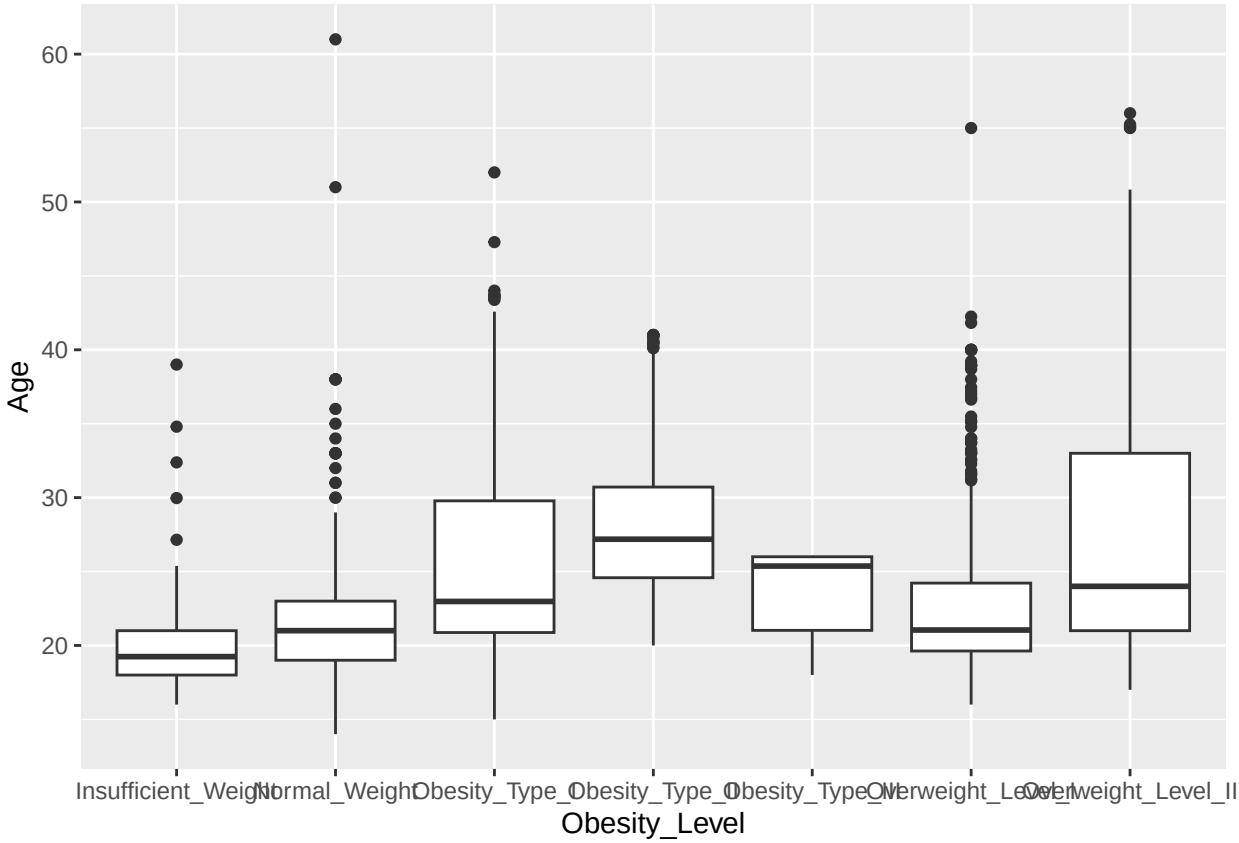
Our results show what variables are predictors of BMI and Obesity group, we will use this info to construct a scoring system based on obesity risk. Our result also point to the fact that Age is significantly influence by an individuals Obesity level especially Obesity Type 2 and 3 we will use this info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

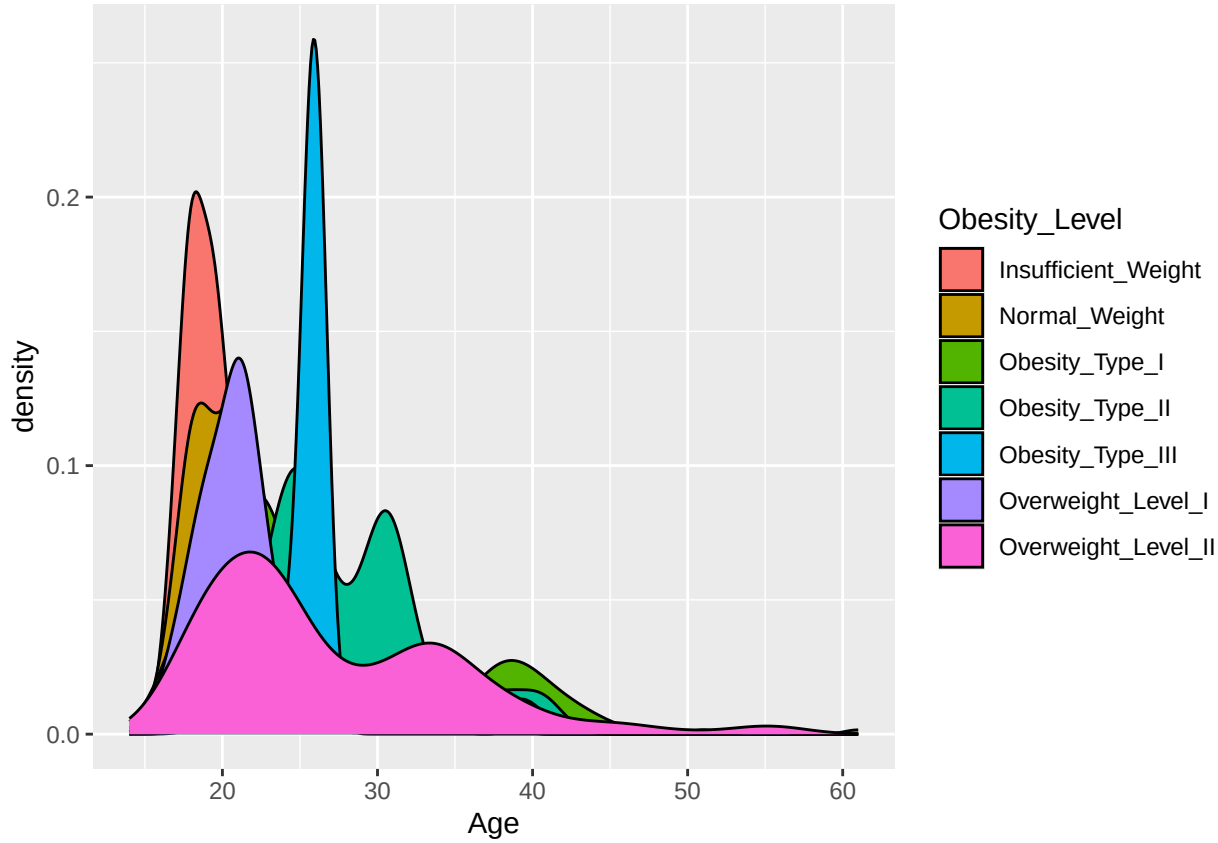



no
yes
250
200
150
100
50
0
count
20
30
40
50
BMI


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



“‘{, echo = FALSEr} # 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() “‘

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