Estimations of Obesity Levels

Sangwook Park # 10176139 14sp74@queensu.ca
Serena Liu #20118930 17jl245@queensu.ca
Bofan Sun #20108143 17bs48@queensu.ca
Chenhao Zhu #20112538 17cz43@queensu.ca
Jingyu Xu #20091627 17jx17@queensu.ca
Chengfeng Jiang #20111648 17cj15@queensu.ca

List of Contributions

❿Sangwook Park: Introduction and description of dataset
❿Chengfeng Jiang: visualization of dataset, linear regression, normalization of the data
❿Chenhao Zhu: perform classification KNN method
❿Bofan Sun: perform classification C5.0 method
❿Serena Liu: regression on BMI
❿Jingyu Xu: hypothesis test, conclusion
❿All group members have discussed every part of the project together

Introduction

In a modern society, obesity is a complex health issue that could arise from individual factors such as their behaviour and genetics. Behaviours can include eating habits such as number of main meals and frequent consumption of high caloric food, whether they are physically active, and their sleeping patterns. Besides one’s eating and physical activity environment, genetics also play a big role in obesity as some specific disorders can directly cause obesity. Having a poorer mental health and a lower standard of living are associated with this issue as well.

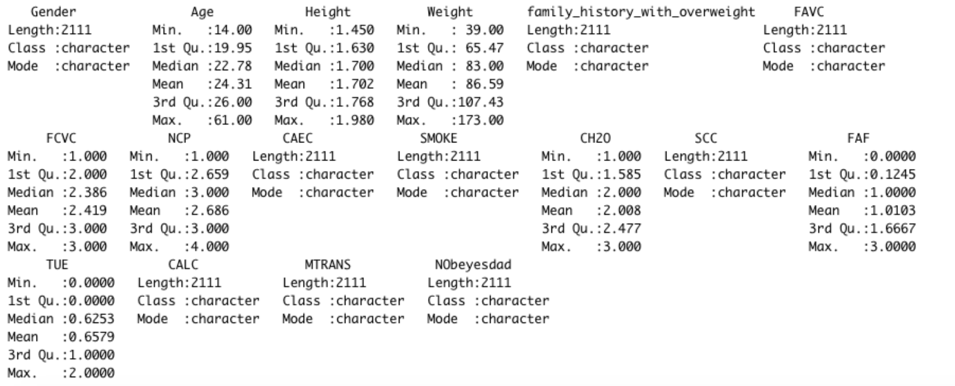
As said above, the prevalence of obesity has been a worldwide issue that should not be ignored as it could affect the economy badly. For those who are obese, they would have a larger economic burden than those who are not as they need to spend a larger expenditure on health care. Besides excess health care expenditure, having obesity would decrease productivity in the workplace which could inhibit economic growth. Combining all these aspects together, it is worth studying this problem because it will observe which factor closely relates to obesity. This way we can prevent obesity beforehand and be more aware of how obesity affects our daily life.

According to the study done by Vadera and three other authors, they tested the effect of dietary factors on the weight status of an adult population in Jamnagar city of Gujarat (Vadera, Yadav, Yadav, Parmar, & Unadkat, 2010). They used clustering method to select study samples and data were collected by interviewing people through house-to-house visits. They found the prevalence of overweight, and obesity was 22.04% and 5.20%, respectively. There were more overweight females than males and the prevalence rate had an increasing trend up to age 60. They found that the total calorie intake and habit of snacking were associated with weight gain. Also, the overweighted people tended to have more oily food and less vegetables. The result was quite obvious and not surprising at all, but it reminds us again that eating habits are a huge contributor to weight gain.

The fact that there are many other studies analyzing obesity proves that this is an important public health issue. Considering its seriousness that affects our health and economy, there should be an increased need for closely monitoring factors that contribute to obesity. By choosing this dataset as our final project, we can realize whether we should modify our lifestyle to prevent obesity beforehand.

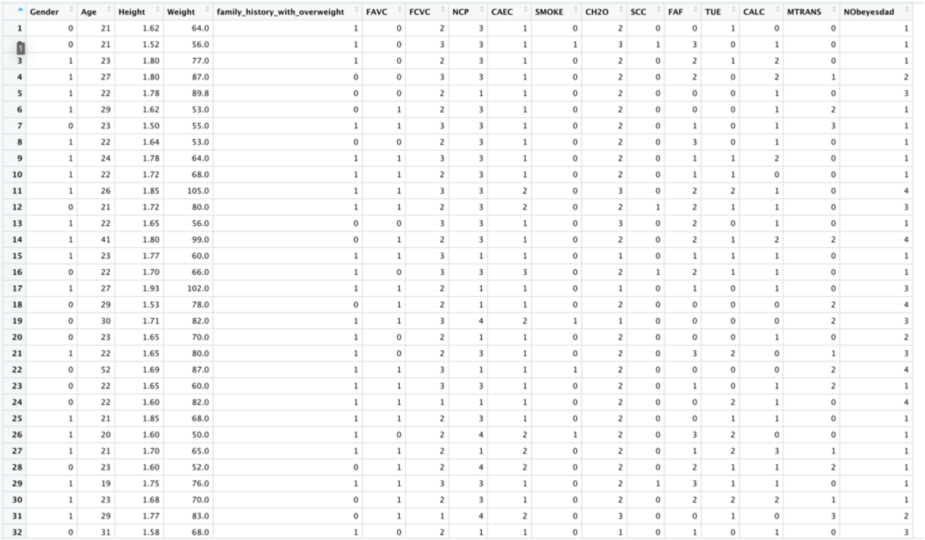
Description of Dataset

The dataset we decided to use is called “Estimation of obesity levels based on eating habits and physical condition data set” found on UCI Machine Learning Repository (Estimation of Obesity Levels Based on Eating Habits and Physical Condition Data Set, 2019). To briefly talk about variables in this dataset, there are two types of variables: numerical and categorical.



```
    Gender               Age            Height          Weight       family_history_with_overweight     FAVC
 Length:2111        Min.   :14.00   Min.   :1.450   Min.   : 39.00   Length:2111                    Length:2111
 Class :character   1st Qu.:19.95   1st Qu.:1.630   1st Qu.: 65.47   Class :character               Class :character
 Mode  :character   Median :22.78   Median :1.700   Median : 83.00   Mode  :character               Mode  :character
                    Mean   :24.31   Mean   :1.702   Mean   : 86.59
                    3rd Qu.:26.00   3rd Qu.:1.768   3rd Qu.:107.43
                    Max.   :61.00   Max.   :1.980   Max.   :173.00
      FCVC            NCP            CAEC              SMOKE                CH2O           SCC                 FAF
 Min.   :1.000   Min.   :1.000   Length:2111        Length:2111         Min.   :1.000   Length:2111        Min.   :0.0000
 1st Qu.:2.000   1st Qu.:2.659   Class :character   Class :character    1st Qu.:1.585   Class :character   1st Qu.:0.1245
 Median :2.386   Median :3.000   Mode  :character   Mode  :character    Median :2.000   Mode  :character   Median :1.0000
 Mean   :2.419   Mean   :2.686                                          Mean   :2.008                      Mean   :1.0103
 3rd Qu.:3.000   3rd Qu.:3.000                                          3rd Qu.:2.477                      3rd Qu.:1.6667
 Max.   :3.000   Max.   :4.000                                          Max.   :3.000                      Max.   :3.0000
      TUE              CALC              MTRANS            NObeyesdad
 Min.   :0.0000   Length:2111        Length:2111        Length:2111
 1st Qu.:0.0000   Class :character   Class :character   Class :character
 Median :0.6253   Mode  :character   Mode  :character   Mode  :character
 Mean   :0.6579
 3rd Qu.:1.0000
 Max.   :2.0000
```

According to the summary table above, it can be observed that “Age”, “Height”, “Weight”, “FCVC (Frequency of Consumption of Vegetables)”, “NCP (Number of main meals)”, “CH2O (Consumption of daily water)”, “FAF (Physical Activity Frequency)” and “TUE (Time using technology devices)” are classified as numerical variables that causes obesity level. Furthermore, we can break down categorical variables once again into binary and non-binary variables. “Gender”, “Family_history_with_overweight”, “FAVC (Frequent consumption of high caloric food", "SMOKE" and "SCC" (Calorie's consumption monitoring)" are classified as binary variables as there are only two responses to each data for these variables. On the other hand, "CAEC (Consumption of food between meals)," "CALC (Consumption of Alcohol)", and "MTRANS (Transportation used)" are represented as non-binary variables as there are multiple responses. For example, MTRANS has five responses which include "walking", "automobile", "bike," "motorbike" and "public transportation." Moreover "NobeyesDad" have 7 different responses, such as "insufficient weight", "normal weight", "Overweight_Level_I", "Overweight_Level II", "Obesity_Type I", "Obesity_Type_II and "Obesity_Type_III". Since there are three different types of data, which is numerical, binary, and categorical, converting these non-numerical types of data to numerical will be much more convenient to analyze the data by R. For those of data, which is binary, we could simply transfer them become yes equal to 1, and no becomes to 0, also for gender we also could transfer male to 1, and female equal to 0. For those of data, which is categorical, taking NobeyesDad as example. Simple transfering "insufficient to 0", "normal weight equals to 1", "Overweight_Level_I to 2", "Overweight_Level_II to 3", "Obesity_Type_II to 4", "Obesity_Type_II to 5", and "Obesity_Type_III to 6", and here is the data looks like, after normalizing the data.

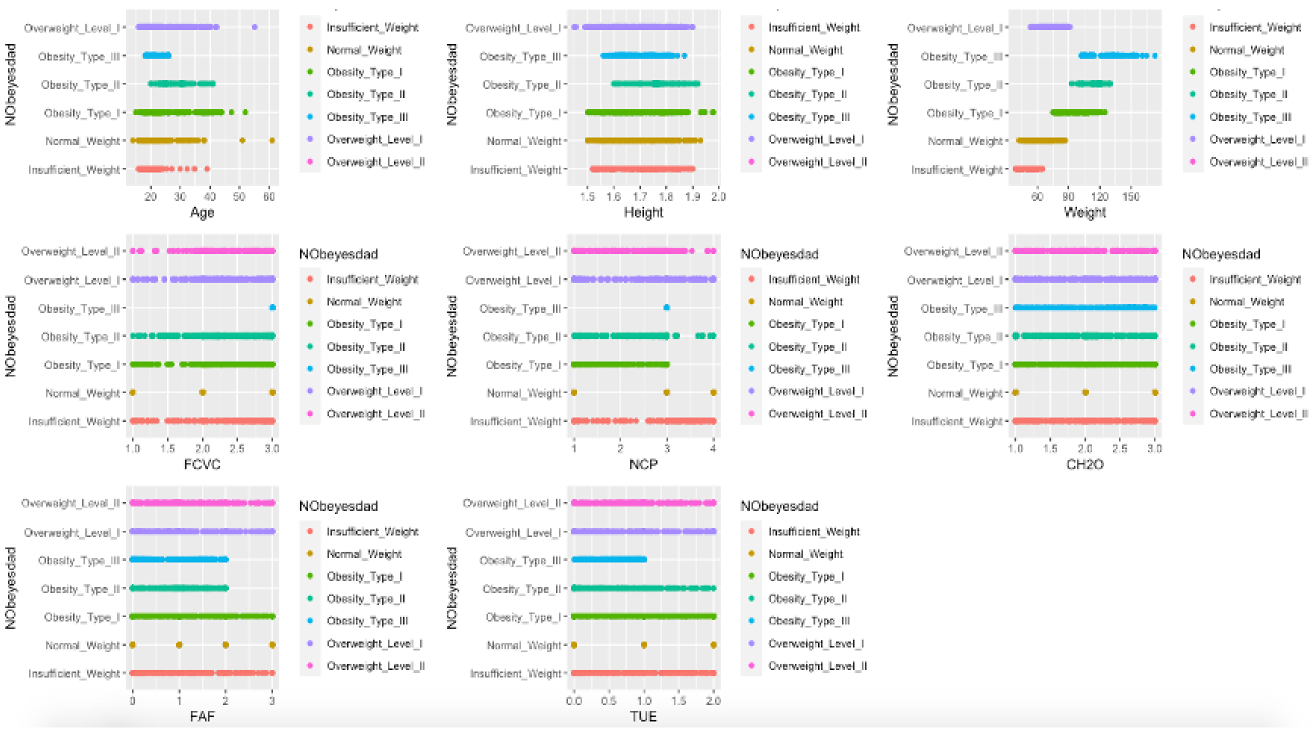


	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	NObeyesdad
1	0	21	1.62	64.0	1	0	2	3	1	0	2	0	0	1	0	0	1
2	0	21	1.52	56.0	1	0	3	3	1	1	3	1	3	0	1	0	1
3	1	23	1.80	77.0	1	0	2	3	1	0	2	0	2	1	2	0	1
4	1	27	1.80	87.0	0	0	3	3	1	0	2	0	2	0	2	1	2
5	1	22	1.78	89.8	0	0	2	1	1	0	2	0	0	0	1	0	3
6	1	29	1.62	53.0	0	1	2	3	1	0	2	0	0	0	1	2	1
7	0	23	1.50	55.0	1	1	3	3	1	0	2	0	1	0	1	3	1
8	1	22	1.64	53.0	0	0	2	3	1	0	2	0	3	0	1	0	1
9	1	24	1.78	64.0	1	1	3	3	1	0	2	0	1	1	2	0	1
10	1	22	1.72	68.0	1	1	2	3	1	0	2	0	1	1	0	0	1
11	1	26	1.85	105.0	1	1	3	3	2	0	3	0	2	2	1	0	4
12	0	21	1.72	80.0	1	1	2	3	2	0	2	1	2	1	1	0	3
13	1	22	1.65	56.0	0	0	3	3	1	0	3	0	2	0	1	0	1
14	1	41	1.80	99.0	0	1	2	3	1	0	2	0	2	1	2	2	4
15	1	23	1.77	60.0	1	1	3	1	1	0	1	0	1	1	1	0	1
16	0	22	1.70	66.0	1	0	3	3	3	0	2	1	2	1	1	0	1
17	1	27	1.93	102.0	1	1	2	1	1	0	1	0	1	0	1	0	3
18	0	29	1.53	78.0	0	1	2	1	1	0	2	0	0	0	0	2	4
19	0	30	1.71	82.0	1	1	3	4	2	1	1	0	0	0	0	2	3
20	0	23	1.65	70.0	1	0	2	1	1	0	2	0	0	0	1	0	2
21	1	22	1.65	80.0	1	0	2	3	1	0	2	0	3	2	0	1	3
22	0	52	1.69	87.0	1	1	3	1	1	1	2	0	0	0	0	2	4
23	0	22	1.65	60.0	1	1	3	3	1	0	2	0	1	0	1	2	1
24	0	22	1.60	82.0	1	1	1	1	1	0	2	0	0	2	1	0	4
25	1	21	1.85	68.0	1	1	2	3	1	0	2	0	0	1	1	0	1
26	1	20	1.60	50.0	1	0	2	4	2	1	2	0	3	2	0	0	1
27	1	21	1.70	65.0	1	1	2	1	2	0	2	0	1	2	3	1	1
28	0	23	1.60	52.0	0	1	2	4	2	0	2	0	2	1	1	2	1
29	1	19	1.75	76.0	1	1	3	3	1	0	2	1	3	1	1	0	1
30	1	23	1.68	70.0	0	1	2	3	1	0	2	0	2	2	2	1	1
31	1	29	1.77	83.0	0	1	1	4	2	0	3	0	0	1	0	3	2
32	0	31	1.58	68.0	1	0	2	1	1	0	1	0	1	0	1	0	3

Methods and Results

Data Visualization

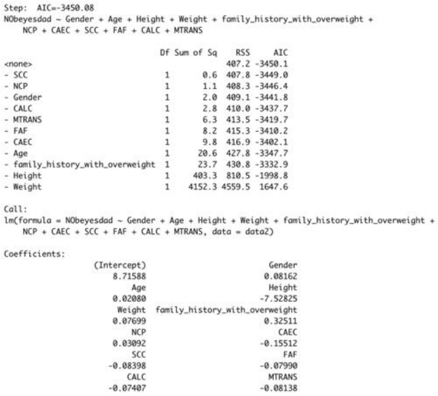
The graph below shows the relationship between obesity level and each numerical variable. Other than physical factors such as weight, height or age, it can be observed that normal weighted people tend to have a similar lifestyle, whereas people in other obesity levels tend to have inconsistent patterns. With this graph, I can conclude these graphs don’t have enough information to prove a certain factor contributes to obesity. However, we will explore more deeply which factor closely relates to obesity level using regression and classification method.



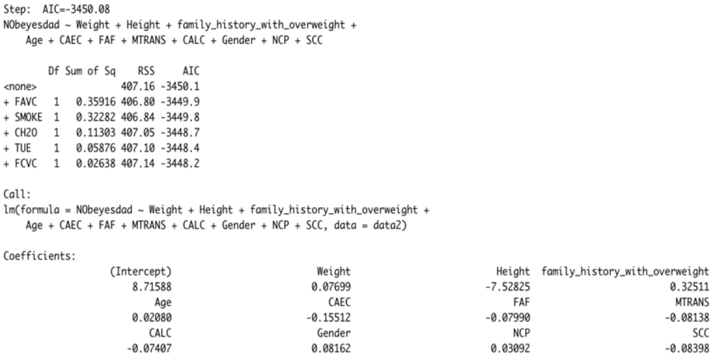
Method 1: Regression

After normalizing the data, there is 7 different obesity level which related to “NobeyesDad”, and there are also have 16 different variables in the data set. Do these 16 different variables all going to influence the “NobeyesDad”? With that question, let us use linear regression to check the answers, setting the “NobeyesDad” as the dependent variable, and setting other 16 different variables as independent variable. Using forward and backward selection to find out what is the best linear regression model. After the backward and forward selection, both AIC value is equal to –3450.08. Therefore, Nobeyesdad=Weight+Height+Family_history_with_overweight+Age+CAEC+FAF+MTRANS+CALC+Gender+NCP+SCC is the best fitted linear relationship with NobeyesDad.

However, numerical variable is much more stable than the binary variable and categorical variable. Therefore, although “Family_history_with_overweight,” “CAEC,” “FAF,” “MTRANS,” “CALC,” “Gender,” “NCP” and “SCC” will influence the value of Nobeyesdad, but compared to other numerical variable which is height, weight, and age. Numerical variables will have a more significance affect compared to binary and numerical variables.



```
Step:  AIC=-3450.08
NObeyesdad ~ Gender + Age + Height + Weight + family_history_with_overweight +
    NCP + CAEC + SCC + FAF + CALC + MTRANS

                                 Df Sum of Sq    RSS     AIC
<none>                                         407.2 -3450.1
- SCC                             1       0.6  407.8 -3449.0
- NCP                             1       1.1  408.3 -3446.4
- Gender                          1       2.0  409.1 -3441.8
- CALC                            1       2.8  410.0 -3437.7
- MTRANS                          1       6.3  413.5 -3419.7
- FAF                             1       8.2  415.3 -3410.2
- CAEC                            1       9.8  416.9 -3402.1
- Age                             1      20.6  427.8 -3347.7
- family_history_with_overweight  1      23.7  430.8 -3332.9
- Height                          1     403.3  810.5 -1998.8
- Weight                          1    4152.3 4559.5  1647.6

Call:
lm(formula = NObeyesdad ~ Gender + Age + Height + Weight + family_history_with_overweight +
    NCP + CAEC + SCC + FAF + CALC + MTRANS, data = data2)

Coefficients:
                   (Intercept)                          Gender
                       8.71588                         0.08162
                           Age                          Height
                       0.02080                        -7.52825
                        Weight  family_history_with_overweight
                       0.07699                         0.32511
                           NCP                            CAEC
                       0.03092                        -0.15512
                           SCC                             FAF
                      -0.08398                        -0.07990
                          CALC                          MTRANS
                      -0.07407                        -0.08138
```



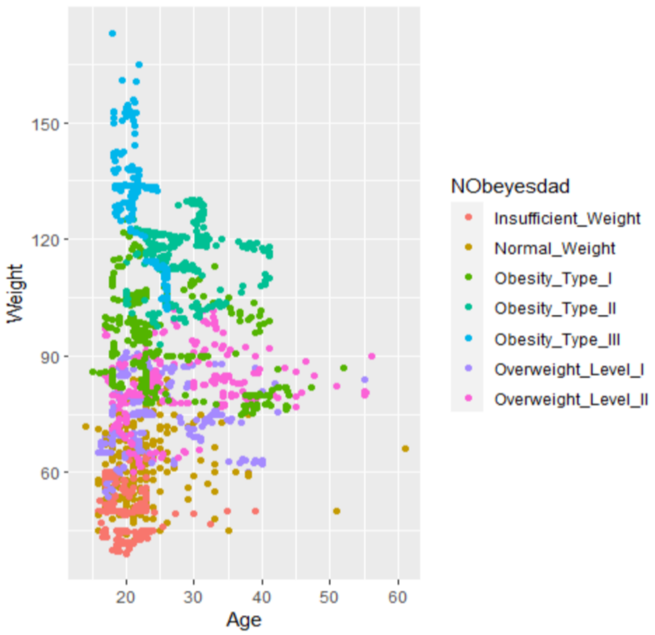
```
Step:  AIC=-3450.08
NObeyesdad ~ Weight + Height + family_history_with_overweight +
    Age + CAEC + FAF + MTRANS + CALC + Gender + NCP + SCC

        Df Sum of Sq    RSS     AIC
<none>                407.16 -3450.1
+ FAVC   1   0.35916 406.80 -3449.9
+ SMOKE  1   0.32282 406.84 -3449.8
+ CH2O   1   0.11303 407.05 -3448.7
+ TUE    1   0.05876 407.10 -3448.4
+ FCVC   1   0.02638 407.14 -3448.2

Call:
lm(formula = NObeyesdad ~ Weight + Height + family_history_with_overweight +
    Age + CAEC + FAF + MTRANS + CALC + Gender + NCP + SCC, data = data2)

Coefficients:
                   (Intercept)                          Weight                          Height  family_history_with_overweight
                       8.71588                         0.07699                        -7.52825                         0.32511
                           Age                            CAEC                             FAF                          MTRANS
                       0.02080                        -0.15512                        -0.07990                        -0.08138
                          CALC                          Gender                             NCP                             SCC
                      -0.07407                         0.08162                         0.03092                        -0.08398
```

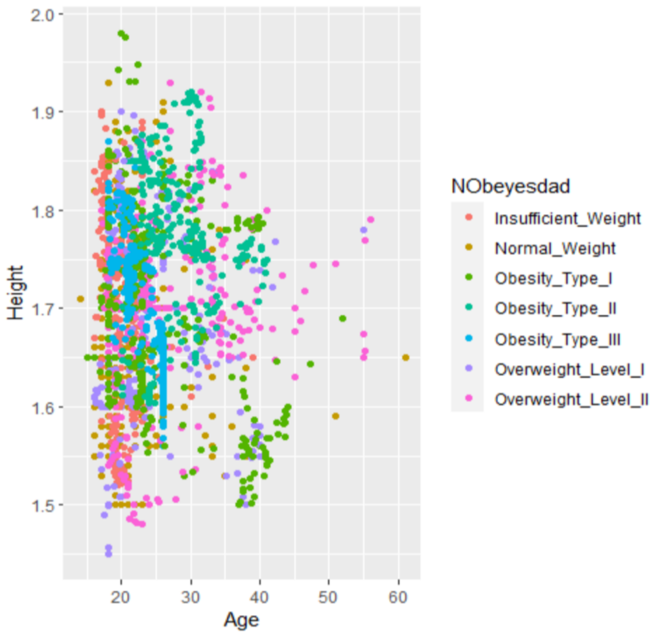
Method 2: Classification

After using Linear regression to approximate the obesity level, we found that Weight, Height and Age have a strong linear relationship with obesity level if numbering the obesity level. However, as the ‘NobyesDad’ variable in the data is originally categorical, with seven levels, then linear regression is not very convincing. For better predicting obesity level, it is worth considering using classification methods KNN and C5.0.

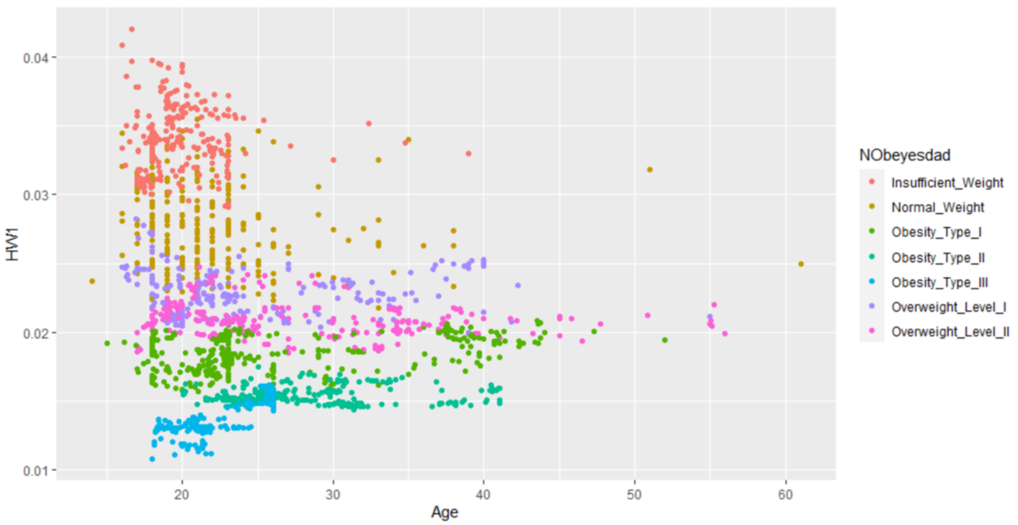
The first part of classification is the KNN method with all variables included. First, we put all the data into KNN and see how the classification method works. In order to increase the accuracy of the KNN method, some data transformation and normalization are necessary. After the data transformation (which is in description of the data), we apply the KNN method to the dataset to make a prediction. Then we used the table () function to test the accuracy. After several tests, the accuracy was around 90 percent.

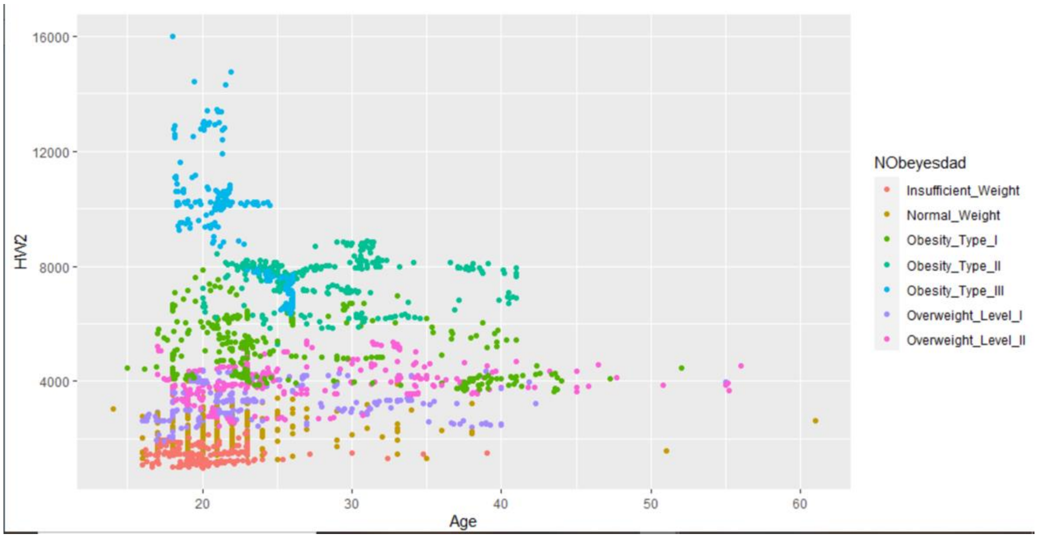
Moreover, it is still not clear that which variables contribute most to the prediction. It seems height, weight, and Age have linear relationship with obesity level. Then, other than KNN, it is important to plot height, weight versus age by Obesity level to observe their relationship in Classification.

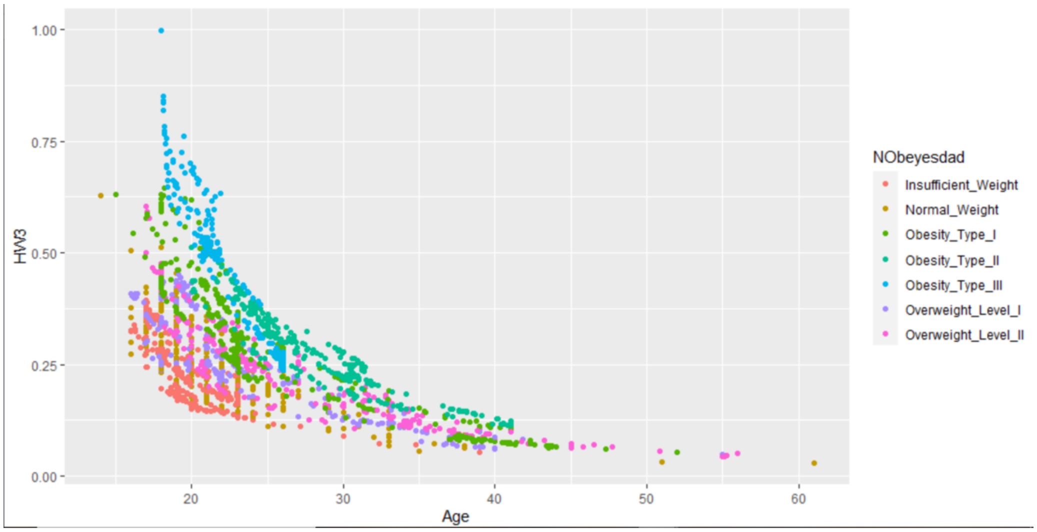




From the plot, it shows that simple height, weight, and age does not classify obesity level well. Then we applied some simple combinations, plotted them by obesity level, and tested their accuracy. The accuracy is improving, this implies that combinations of these variables may be an excellent way included in the data transformation before classification models. For example, I applied these combinations, Height/Weight, Weight^2/Height, Height*Weight/Age^2. They are called ’HW1’, ’HW2’, ’HW3’.



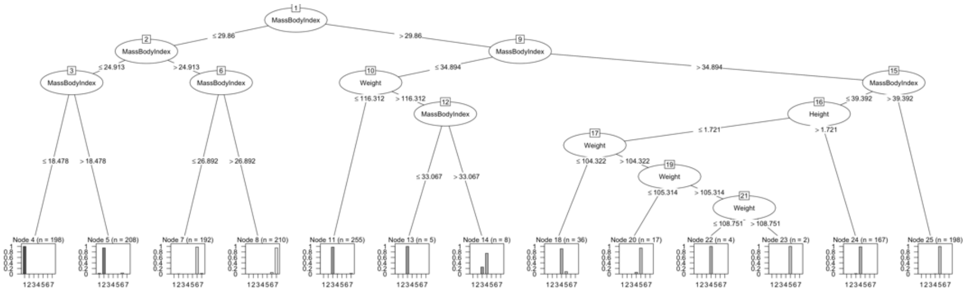




From these three combinations, the classification boundary on their graphs to classify obesity level is clearer than previous. Besides, we need to find out whether Age is an important feature in classifying obesity levels. From my observation, every obesity level has almost the same amount in percentage for every age. So probably the combination of Height and Weight will be more important.

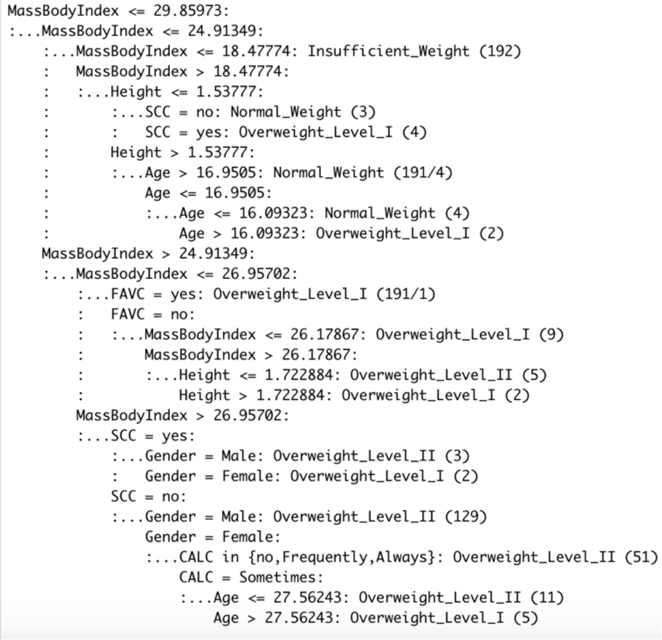
In short, a combination of height and weight is important in classifying obesity level. Furthermore, there is one vital weakness in KNN that it does not produce an actual model, meaning that it does not indicate the relationship between features and prediction. Thus, it is worth trying some other methods other than KNN to do the classification.

Next, apply C5.0 method to the dataset to examine how height and weight influence obesity levels together. The C5.0 method can generate a classification tree automatically and help better observe the relationship between obesity level and the combination of height and weight. Since most of the variables in this dataset are categorical, it is easy to be split into a classification tree. We created a new variable called “CHW”, which means the measure of how height and weight influence obesity level together. Then, we choose only height, weight, and CHW with the response ‘NObeyesdad’ which represent obesity levels to fit the model. We apply three combinations of height and weight (CHW) and fit each of them into a model to test their performance. The combinations that we tested are weight/height, weight / height^2, and weight / height^3. Then we applied C5.0 method to each model and compared their accuracy and model performance by confusion matrix with the function cross Table(). All of the three models have very high accuracy, which is around 98 percent, but the second model with CHW = weight / height^2 has the best model performance. As a result, we choose this model to analyze the relationship between CHW and obesity level. By drawing the decision tree of this model, we can get a detailed range of CHW for each obesity level. (See the plot below). Thus, obesity levels can be determined by CHW.

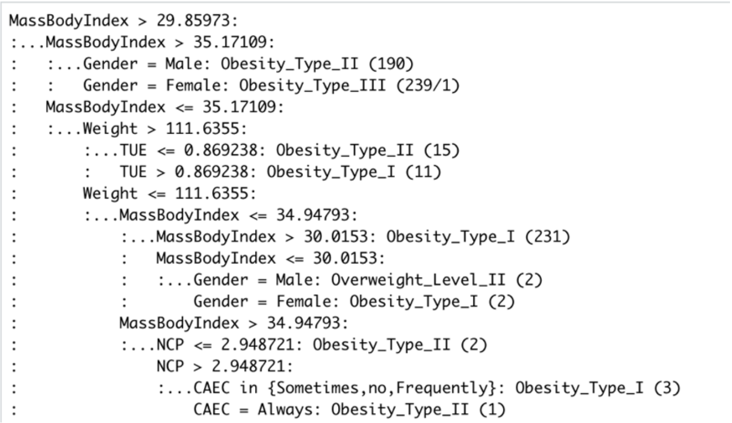


❿ "1" -
- Insufficient_Weight
❿ "2" -- Normal_Weight
❿ "3" -- Obesity_Type_I
❿ "4" -
- Obesity_Type_II
❿ "5" -

However, this model only contains three variables, considering that other features may also influence the CHW value, so we fit another model with all variables included. Since C5.0 method can handle both numerical and nominal features, this model uses the original dataset without dummy variables. Comparing the new model with the previous model, we can observe higher accuracy and better model performance. By analyzing the decision tree, the overall structure of the trees of the two models are similar, however, the decision tree obtained by this model is more detailed with more branch nodes.



```
MassBodyIndex <= 29.85973:
:...MassBodyIndex <= 24.91349:
    :...MassBodyIndex <= 18.47774: Insufficient_Weight (192)
    :   MassBodyIndex > 18.47774:
    :   :...Height <= 1.53777:
    :       :...SCC = no: Normal_Weight (3)
    :       :   SCC = yes: Overweight_Level_I (4)
    :       Height > 1.53777:
    :       :...Age > 16.9505: Normal_Weight (191/4)
    :           Age <= 16.9505:
    :           :...Age <= 16.09323: Normal_Weight (4)
    :               Age > 16.09323: Overweight_Level_I (2)
    MassBodyIndex > 24.91349:
    :...MassBodyIndex <= 26.95702:
        :...FAVC = yes: Overweight_Level_I (191/1)
        :   FAVC = no:
        :   :...MassBodyIndex <= 26.17867: Overweight_Level_I (9)
        :       MassBodyIndex > 26.17867:
        :       :...Height <= 1.722884: Overweight_Level_II (5)
        :           Height > 1.722884: Overweight_Level_I (2)
        MassBodyIndex > 26.95702:
        :...SCC = yes:
            :...Gender = Male: Overweight_Level_II (3)
            :   Gender = Female: Overweight_Level_I (2)
            SCC = no:
            :...Gender = Male: Overweight_Level_II (129)
                Gender = Female:
                :...CALC in {no,Frequently,Always}: Overweight_Level_II (51)
                    CALC = Sometimes:
                    :...Age <= 27.56243: Overweight_Level_II (11)
                        Age > 27.56243: Overweight_Level_I (5)
```



```
MassBodyIndex > 29.85973:
:...MassBodyIndex > 35.17109:
:   :...Gender = Male: Obesity_Type_II (190)
:   :   Gender = Female: Obesity_Type_III (239/1)
:   MassBodyIndex <= 35.17109:
:   :...Weight > 111.6355:
:       :...TUE <= 0.869238: Obesity_Type_II (15)
:       :   TUE > 0.869238: Obesity_Type_I (11)
:       Weight <= 111.6355:
:       :...MassBodyIndex <= 34.94793:
:           :...MassBodyIndex > 30.0153: Obesity_Type_I (231)
:           :   MassBodyIndex <= 30.0153:
:           :   :...Gender = Male: Overweight_Level_II (2)
:           :       Gender = Female: Obesity_Type_I (2)
:           MassBodyIndex > 34.94793:
:           :...NCP <= 2.948721: Obesity_Type_II (2)
:               NCP > 2.948721:
:               :...CAEC in {Sometimes,no,Frequently}: Obesity_Type_I (3)
:                   CAEC = Always: Obesity_Type_II (1)
```

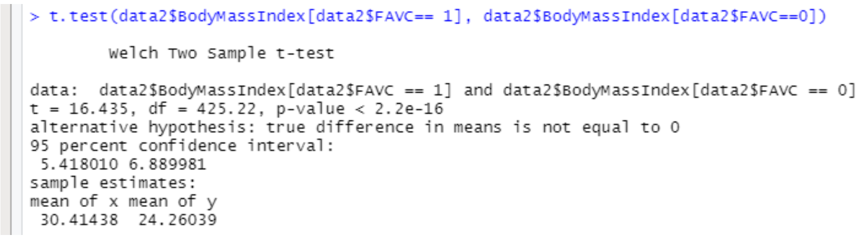
In summary, by fitting models with C5.0 method and analyzing the decision tree, we got the best combination of height and weight which is weight/height ^ 2 to illustrate the relationship with obesity levels, and a detailed range of CHW for each obesity level. (See table below)

Obesity Level	CHW
Insufficient Weight	CHW <= 18.5
Normal Weight	18.5 < CHW <= 24.9
Overweight Level I	24.9 < CHW <= 26.9
Overweight Level II	26.9 < CHW <= 29.9
Obesity Type I	29.9 < CHW <=34.9

Obesity Type II	24.9 < CHW <= 39.9
Obesity Type III	CHW > 39.9

Hypothesis Test

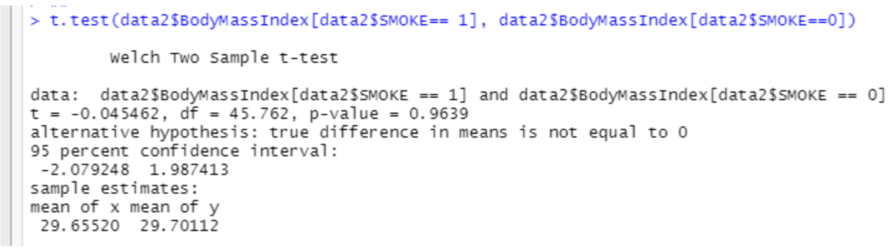
Since we have found out that height and weight have sufficient relationship with CHW, also, CHW(BMI) affects the Obesity level as well. So, we might now consider what the relationships between the other covariates with CHW(BMI) are. Then, we could use hypothesis test to figure them out. For example, to test the relationship between attribute FAVC（frequencies of eating high caloric food) with CHW, let set the µ1 represents the mean of CHW in eating high caloric group, and µ2 is the mean of CHW in without eating high caloric group. Set the null hypothesis H0: µ1= µ2, and its alternative hypothesis H1: µ1 ≠ µ2. Then, by using 2 sample t-test function, we could get the p value is extremely small than α = 0.05, which reject the null hypothesis.



```
> t.test(data2$BodyMassIndex[data2$FAVC== 1], data2$BodyMassIndex[data2$FAVC==0])

	Welch Two Sample t-test

data:  data2$BodyMassIndex[data2$FAVC == 1] and data2$BodyMassIndex[data2$FAVC == 0]
t = 16.435, df = 425.22, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 5.418010 6.889981
sample estimates:
mean of x mean of y 
 30.41438  24.26039 
```

This implies that their means are different, and this attribute FAVC will influence CHW. After we tried all attributes, we found that most of them have extremely small p values, but there are some different ones. First one is attribute SMOKE. Similarly, we set µ1 represents the mean of CHW in smoking group, and µ2 is the mean of CHW in non-smoking group. Set the null hypothesis H0: µ1= µ2, and its alternative hypothesis H1: µ1 ≠ µ2.

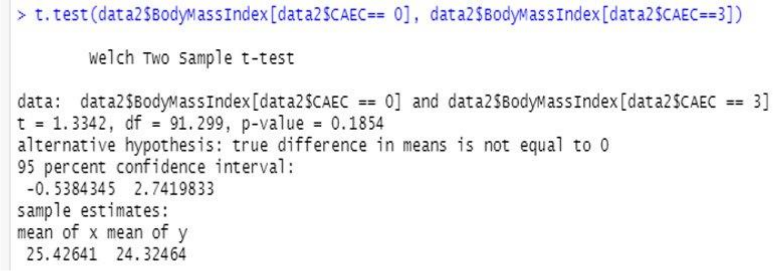
Then, by using 2 sample t-test function, we could get the p value 0.9639, which is larger than α = 0.05, so we fail to reject the null hypothesis.



```
> t.test(data2$BodyMassIndex[data2$SMOKE== 1], data2$BodyMassIndex[data2$SMOKE==0])

	Welch Two Sample t-test

data:  data2$BodyMassIndex[data2$SMOKE == 1] and data2$BodyMassIndex[data2$SMOKE == 0]
t = -0.045462, df = 45.762, p-value = 0.9639
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.079248  1.987413
sample estimates:
mean of x mean of y 
 29.65520  29.70112 
```

That implies that they have the same meaning, and it is does not matter to find its correlations. Therefore, we could conclude that SOMKE will not influence CHW.

Besides that, we could also test the categorical attributes, since the above ones are binary. There is attribute called CAEC (frequencies of eating any food between meals). In before data analysis showed that, we use 0 = “no”, 1 = “sometimes”, 2 = “frequently”, and 3 = “always”. By testing each of them, we found that the p value is larger than α only in {no, always} situation.



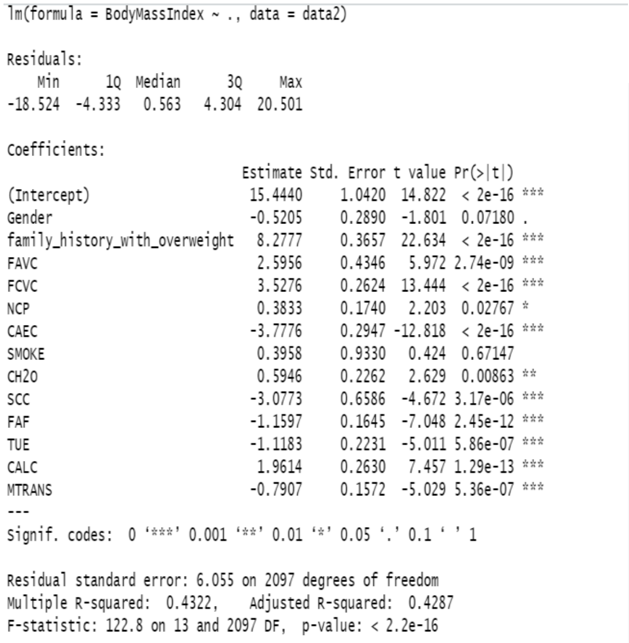
```
> t.test(data2$BodyMassIndex[data2$CAEC== 0], data2$BodyMassIndex[data2$CAEC==3])

	Welch Two Sample t-test

data:  data2$BodyMassIndex[data2$CAEC == 0] and data2$BodyMassIndex[data2$CAEC == 3]
t = 1.3342, df = 91.299, p-value = 0.1854
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5384345  2.7419833
sample estimates:
mean of x mean of y 
 25.42641  24.32464 
```

This implies that CAEC won’t influence CHW only when either person never eats between meals, or always eats between meals. Otherwise, CAEC will still influence CHW.

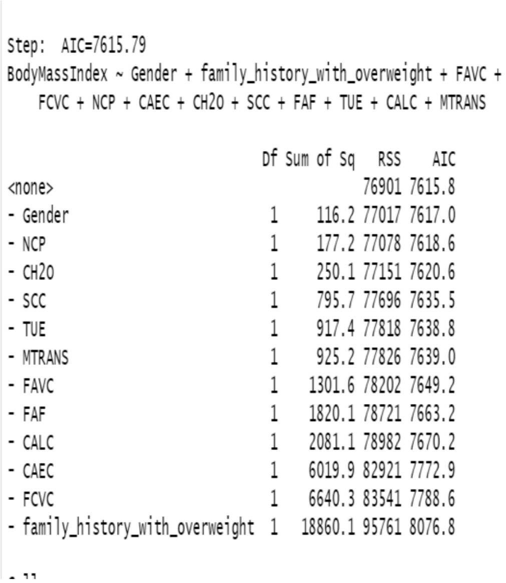
However, hypothesis test could summarize which attribute will influence CHW or not, but it won’t show that how important influences they are. Therefore, to figure this out, we need to go back to using linear regression model to fix it.

Method 1: Regression

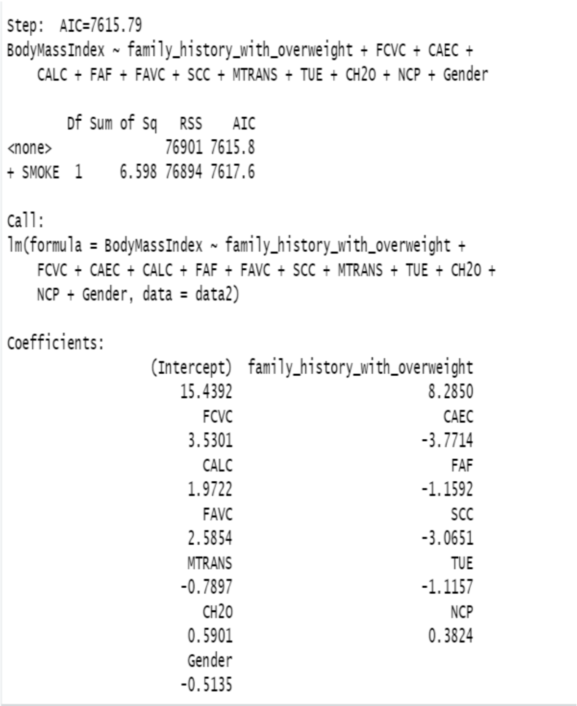
As we know from hypothesis test, we get that SMOKE won't influence CHW, but we still need to figure out the levels of importance attributes for CHW. Since CHW is continuous variables and Obesity levels are discrete. So, we consider using CHW as response y for the model is better to use Obesity levels. Hence, we continue the linear regression by using CHW as response y instead of Obesity level. Remodeled the stepwise method. It will show which variables are keeping the model which means those variables have influence with CHW.



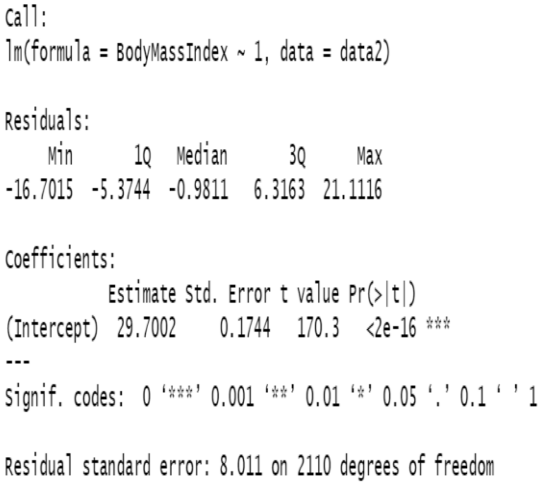
```
lm(formula = BodyMassIndex ~ ., data = data2)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.524  -4.333   0.563   4.304  20.501 

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                   15.4440     1.0420  14.822  < 2e-16 ***
Gender                        -0.5205     0.2890  -1.801  0.07180 .  
family_history_with_overweight  8.2777     0.3657  22.634  < 2e-16 ***
FAVC                           2.5956     0.4346   5.972 2.74e-09 ***
FCVC                           3.5276     0.2624  13.444  < 2e-16 ***
NCP                            0.3833     0.1740   2.203  0.02767 *  
CAEC                          -3.7776     0.2947 -12.818  < 2e-16 ***
SMOKE                          0.3958     0.9330   0.424  0.67147    
CH2O                           0.5946     0.2262   2.629  0.00863 ** 
SCC                           -3.0773     0.6586  -4.672 3.17e-06 ***
FAF                           -1.1597     0.1645  -7.048 2.45e-12 ***
TUE                           -1.1183     0.2231  -5.011 5.86e-07 ***
CALC                           1.9614     0.2630   7.457 1.29e-13 ***
MTRANS                        -0.7907     0.1572  -5.029 5.36e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.055 on 2097 degrees of freedom
Multiple R-squared:  0.4322,	Adjusted R-squared:  0.4287 
F-statistic: 122.8 on 13 and 2097 DF,  p-value: < 2.2e-16
```



```
Step:  AIC=7615.79
BodyMassIndex ~ Gender + family_history_with_overweight + FAVC + 
    FCVC + NCP + CAEC + CH2O + SCC + FAF + TUE + CALC + MTRANS

                                 Df Sum of Sq   RSS    AIC
<none>                                        76901 7615.8
- Gender                          1     116.2 77017 7617.0
- NCP                             1     177.2 77078 7618.6
- CH2O                            1     250.1 77151 7620.6
- SCC                             1     795.7 77696 7635.5
- TUE                             1     917.4 77818 7638.8
- MTRANS                          1     925.2 77826 7639.0
- FAVC                            1    1301.6 78202 7649.2
- FAF                             1    1820.1 78721 7663.2
- CALC                            1    2081.1 78982 7670.2
- CAEC                            1    6019.9 82921 7772.9
- FCVC                            1    6640.3 83541 7788.6
- family_history_with_overweight  1   18860.1 95761 8076.8
```

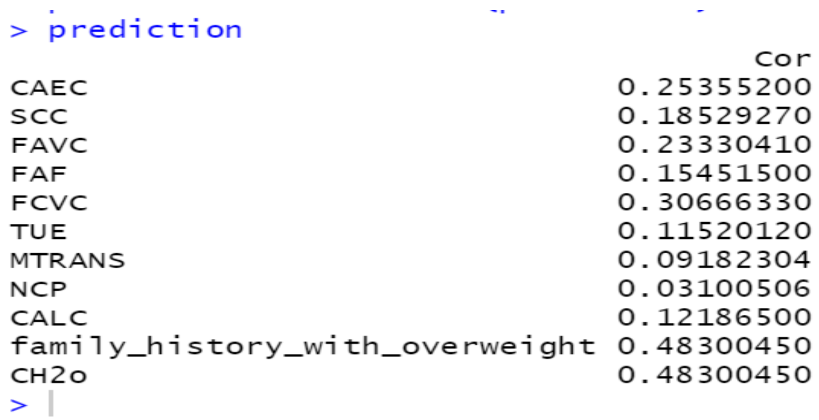


```
Step:  AIC=7615.79
BodyMassIndex ~ family_history_with_overweight + FCVC + CAEC + 
    CALC + FAF + FAVC + SCC + MTRANS + TUE + CH2O + NCP + Gender

        Df Sum of Sq   RSS    AIC
<none>                76901 7615.8
+ SMOKE  1     6.598 76894 7617.6

Call:
lm(formula = BodyMassIndex ~ family_history_with_overweight + 
    FCVC + CAEC + CALC + FAF + FAVC + SCC + MTRANS + TUE + CH2O + 
    NCP + Gender, data = data2)

Coefficients:
                   (Intercept)  family_history_with_overweight  
                       15.4392                          8.2850  
                          FCVC                            CAEC  
                        3.5301                         -3.7714  
                          CALC                             FAF  
                        1.9722                         -1.1592  
                          FAVC                             SCC  
                        2.5854                         -3.0651  
                        MTRANS                             TUE  
                       -0.7897                         -1.1157  
                          CH2O                             NCP  
                        0.5901                          0.3824  
                        Gender  
                       -0.5135  
```



```
Call:
lm(formula = BodyMassIndex ~ 1, data = data2)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.7015  -5.3744  -0.9811   6.3163  21.1116 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  29.7002     0.1744   170.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.011 on 2110 degrees of freedom
```

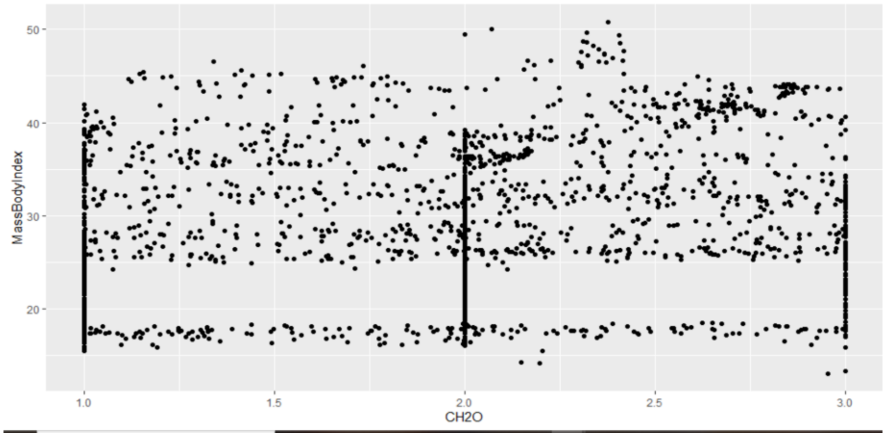
Unfortunately, we still do not know which variables affect more with CHW. Using correlation methods might be the way to perform the proportion of each variable related to CHW since we know the higher correlation, the better the model is. We can calculate the correlation individually first.

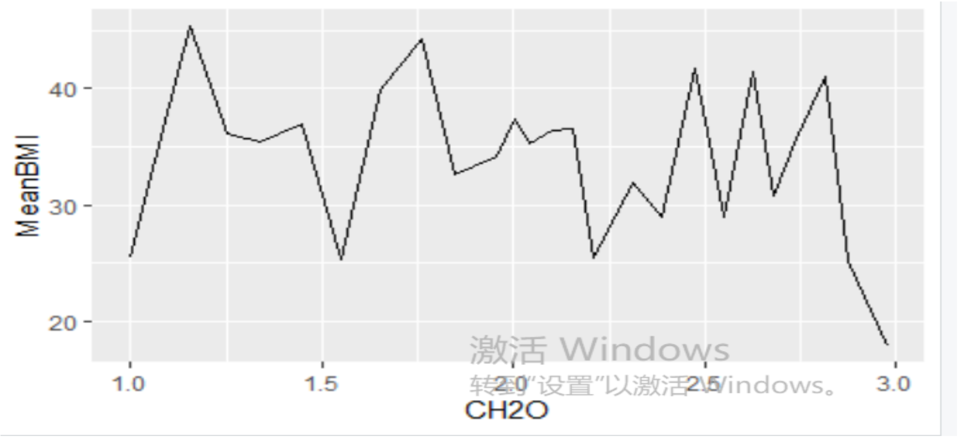


```
> prediction
                                      cor
CAEC                           0.25355200
SCC                            0.18529270
FAVC                           0.23330410
FAF                            0.15451500
FCVC                           0.30666330
TUE                            0.11520120
MTRANS                         0.09182304
NCP                            0.03100506
CALC                           0.12186500
family_history_with_overweight 0.48300450
CH2o                           0.48300450
> 
```

From the result, it shows CH2o, family_history_with_overweight,CAEC, FCVC have much higher correlation than other covariates. We can also add variables to the model to get a new correlation to the model with CH2o and family_history_with_overweight. The result of correlation is also higher than the model only contains CH2o or family_history_with_overweight.   Also, if we add new variables with Ch2o, the correlation is much smaller than the model with family_history_with_overweight. . Based on those, we could be able to conclude that those four covariates affect CHW more especially CH2o and family_history_with_overweight and family_history_with_overweight maybe affects more than CH2o on CHW.

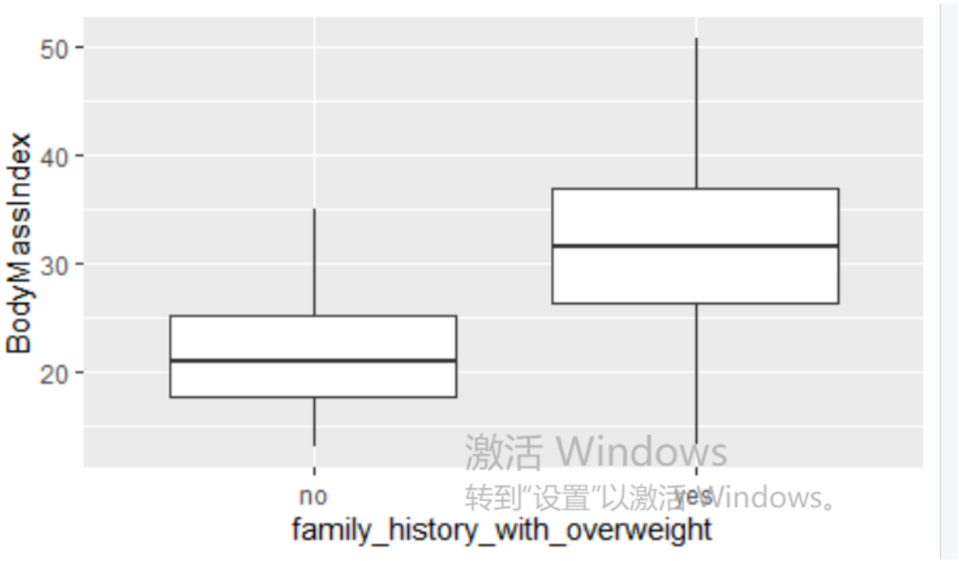
We further analyze CH2o and family_history_with_overweightby. ggplot is the best way to visualize and improve it. We have tried three different methods which are boxplot, lineplot and scatter plot. Different types of variables suitable for specific ggplot such as boxplot are best to describe the model with discrete vs continuous. Hence boxplot is used for family_history_with_overweight, and lineplot for CH2o. I have also tried to use scatter plots for CH2o. Unfortunately, it will not give a visual variation trend since there are so many approximate values in the dataset. The lineplot might be better. The figure is generated by defining the mean CHW as response y and CH2o are x. It has a clearer and visual tread than using scatter plot. Unlucky, it will not give us a clear relationship since the results are not stable.





We still can summarize from the graph that the highest value of CH2o has a much lower value of mean CHW than the lowest value of CH20. Hence, we conclude that people who have a higher consumption of the CH2o have a lower mean CHW.

Secondly, for the boxplot of family_history_with_overweight, people who have family_history_with_overweight have a higher minimum/median /maximum also the outliers than people who have no family_history_with_overweight.



We can assume that people who have family_history_with_overweight have a higher CHW than people who have no family_history_with_overweight. Comparing the two plots, the boxplot has a more visual and clear relationship than the line plot. So, we assume family_history_with_overweight might have more influence than CH2o with CHW.

Since we have used so many binary variables in the model. It might cause a lack of accuracy or weakness of the model; we could improve this weakness by transforming the variables to categorical variables such as using “no or yes” instead of “0 or 1”. Also, we can support the interval between 0 and 1 being low consumption, 1 and 2 being median consumption and 2 and 3 being high consumption for the CH2o for improving the model. It is tedious and we must pay more attention to transforming the variables for different methods. We have discussed so many about CHW and found serious variables that affect more on people’s CHW.

Conclusion and Discussion

Overall, to summarize our report, is that we choose the dataset about Obesity, to figure out which attributes could influence it and their levels of importance. The methods we used are regression and classification which includes the linear regression model, k_nn model C.50

model, and decision trees. In this process, we find that weight and height have a substantial impact on Obesity, so we define another attribute ratio called CHW. After we tried multiple times, we realized that if its formula is weight/height ^ 2 (also the formula of Body Mass Index), is the best one to determine the obesity level. After using linear regression model again, we find that attributes CH2o, (amount of water people drink daily), family_history_with_overweight, CAEC (frequencies of eating any food between meals) and FCVC (frequencies of eating vegetables daily in meals) have the most influences of CHW by their correlations. Therefore, to conclude our results that if we want to avoid obesity, besides factors of height, weight and family history, we need to drink more water, eat more vegetables, and regular our meals in our lives. However, CHW is not a perfect measure, because it does not directly evaluate body fat. For example, muscle and bone have higher density than fat, so a muscular person may have a high CHW, but not have too much fat. If we want to measure our body fat more accuracy, we could use some other ways, such as skinfold thicknesses, bioelectrical impedance, underwater weighing, and dual energy x-ray absorption. They are more expensive, and not used in general. Therefore, we could say that CHW is a reasonable indicator to measure the obesity level, but it couldn’t calculate the body fact exactly, and it could regard as a tool to identify potential weight problems in our live.

References

Estimation of obesity levels based on eating habits and physical condition Data Set. (2019, August 27). [Dataset]. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+

Vadera, B., Yadav, S., Yadav, B., Parmar, D., &amp; Unadkat, S. (2010). Study on obesity and influence of dietary factors on the Weight status of an adult population IN Jamnagar city of Gujarat: A CROSS-SECTIONAL analytical study. Indian Journal of Community Medicine, 35(4), 482-486. doi:10.4103/0970-0218.74346

Department of Health and Human Services Centers for Disease Control and Prevention.
Body Mass Index: Considerations for Practitioners
https://www.cdc.gov/obesity/downloads/bmiforpactitioners.pdf