FINAL PROJECT IN STAT 362 PRESENTED BY

- JAY SANGWOOK PARK #10176139
- BOFAN SUN #20108143
- SERENA LIU # 20118930
- CHENHAO ZHU # 20112538
- CHENGFENG JIANG # 20111648
- JINGYU XU #20091627

Obesity

How serious is obesity?

- 5th leading risk for global deaths (EASO, 2020)
- 2.8 million adults die each year due to overweight and obesity

Main factors to cause obesity

- **Physical factors**
  - Lack of physical activity
  - Transportation
  - Unhealthy eating patterns
- **Social factors**
  - Lower standard of living
  - Financial or a stress from trauma
  - Lack of education regarding health or types of food choices

The impact of prevalence of obesity on economy

Larger expenditure on healthcare

- **Direct medical costs**
  - Preventive, diagnostic, and treatment services
- **Indirect costs**
  - Sickness and death, and a decrease in productivity

A lower productivity

- Inhibit economic growth
- Can this be prevented?

Knowing which factor closely relates to obesity

- Everyone has a different lifestyle
  - Meaning factors differ by person
- Find out the factor that affects obesity the most by using a dataset
- worth studying how obesity is caused because we can
  - Improve our health and lifestyle
  - **Prevent obesity beforehand**

Obesity levels based on eating habits and physical condition

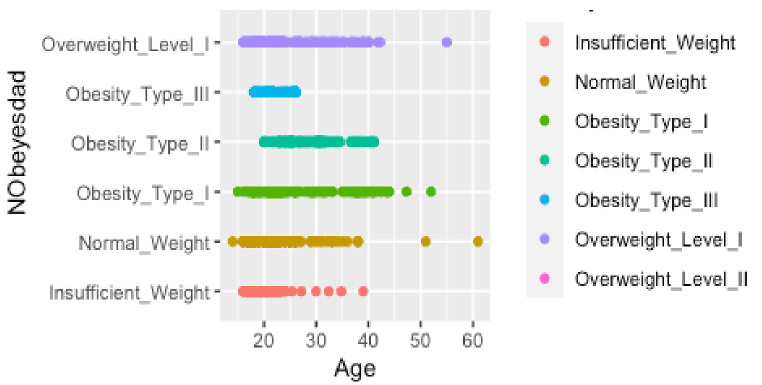
Categorical

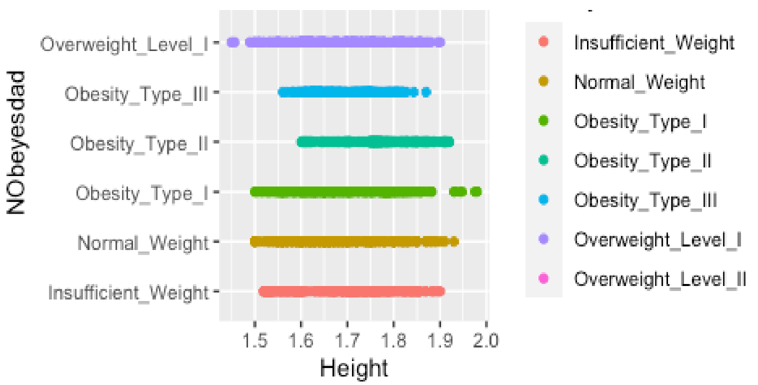
- **Binary**
  - Gender, family history with overweight, **frequent consumption of high caloric food**, smoke, calories consumption monitoring,
- **Non-binary**
  - Consumption of food between meals, consumption of alcohol, transportation used,

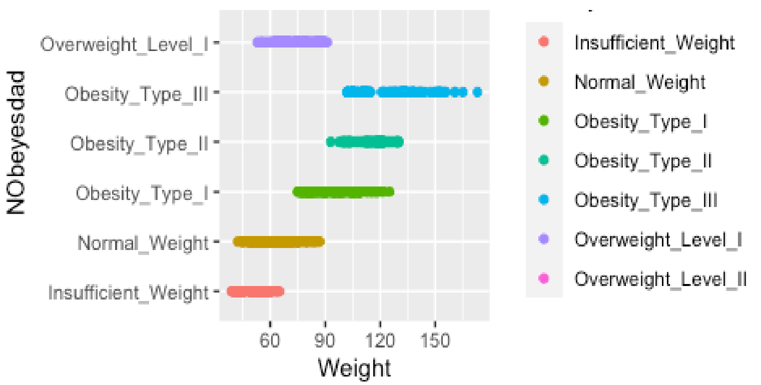
Numerical

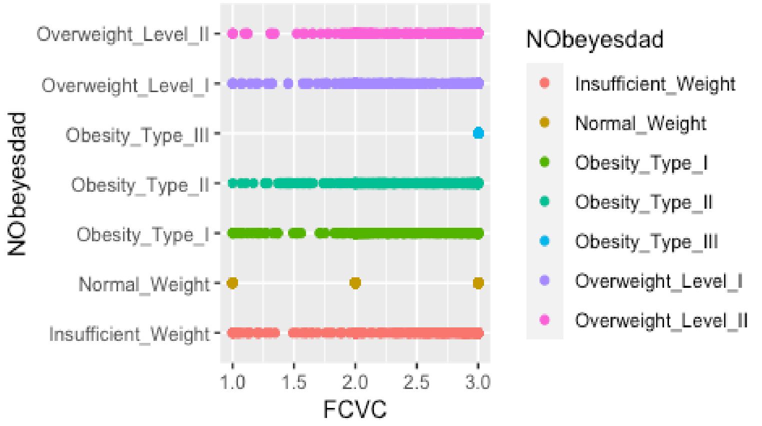
- age, height, weight, frequency of vegetable consumption, number of main meals, CH2O, **physical activity frequency**, time using technology devices

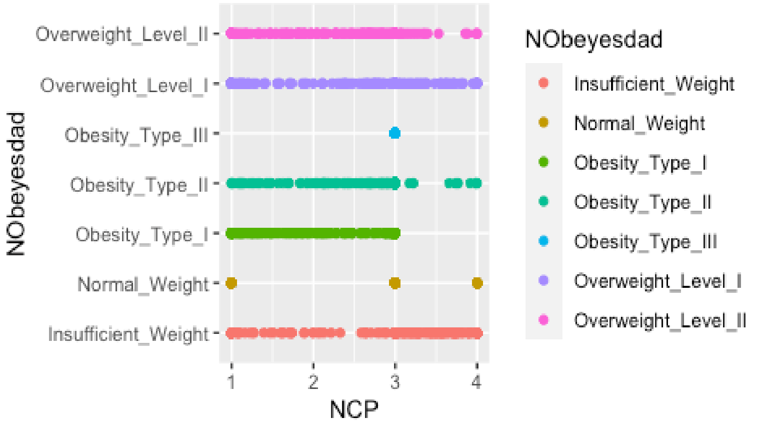
Any observations?

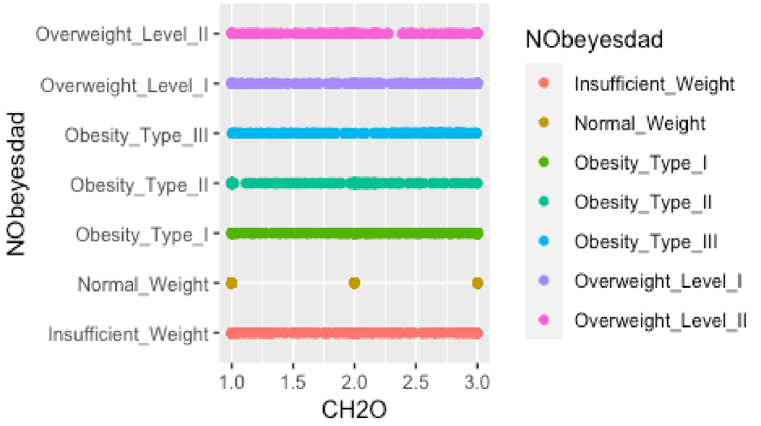


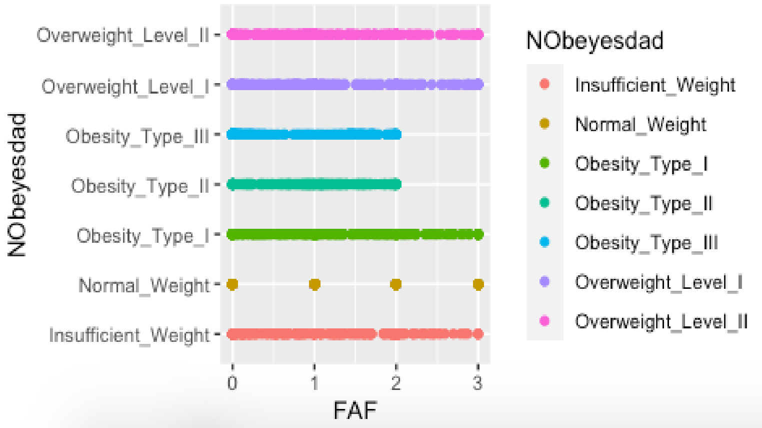


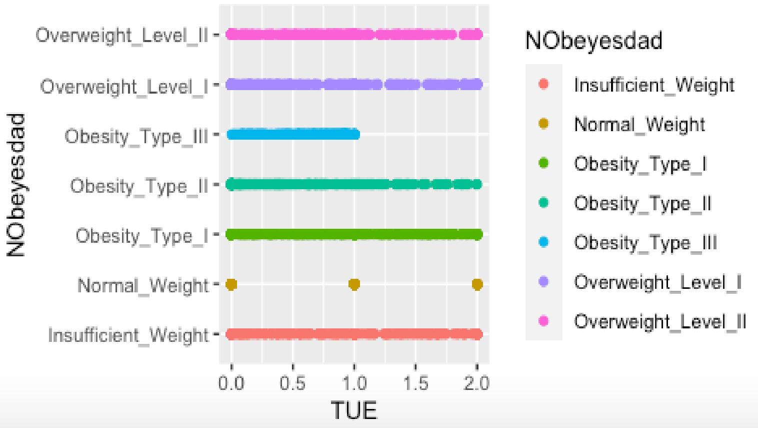




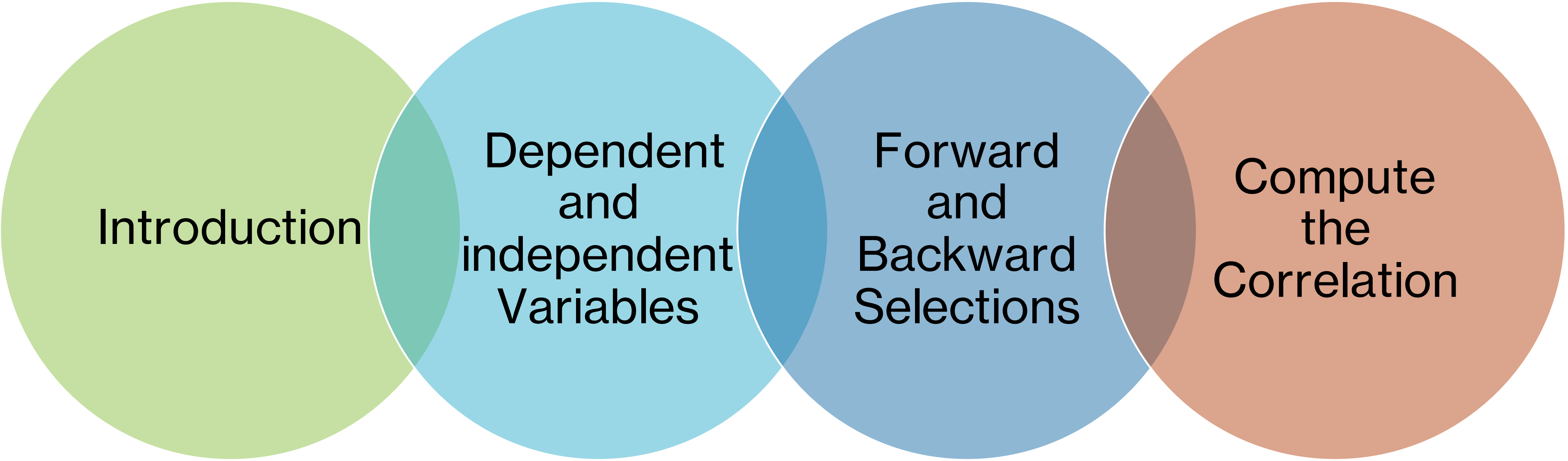








Regression



Introduction

Dependent and independent Variables

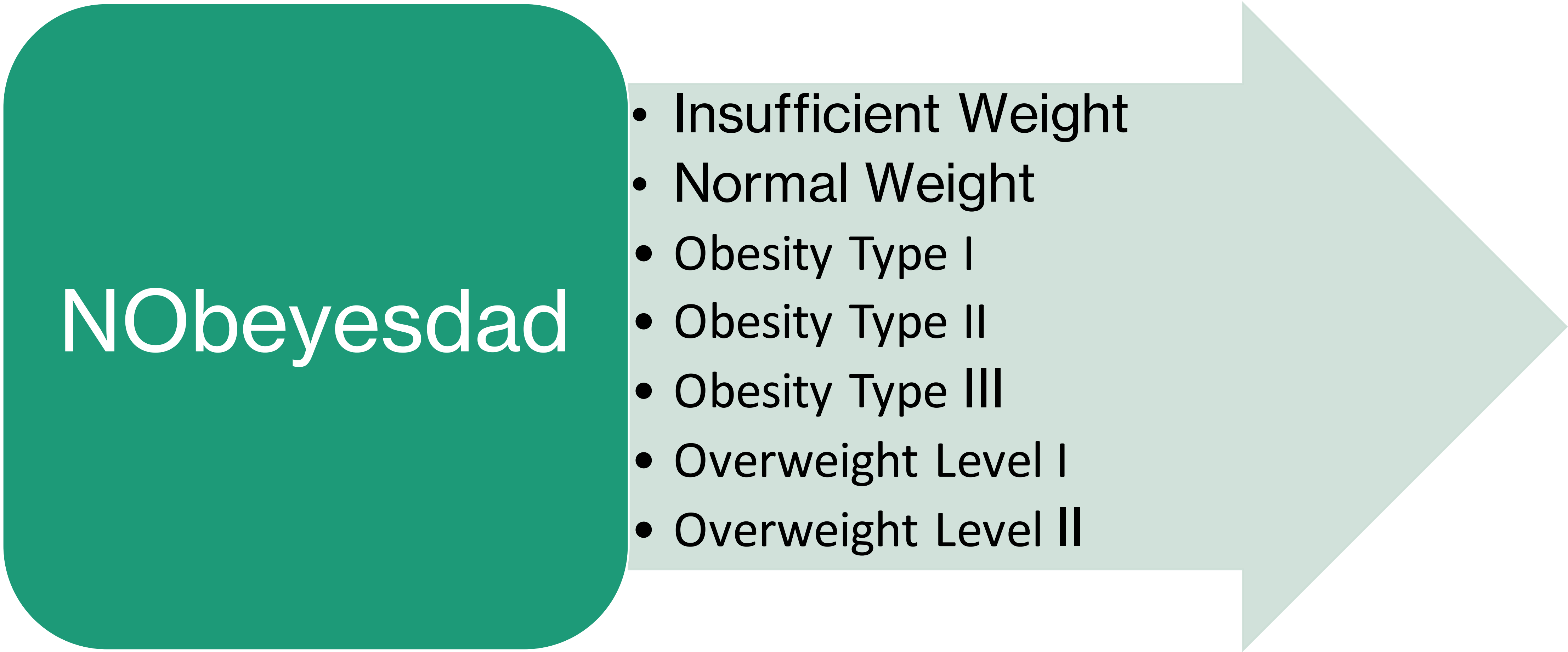
Forward and Backward Selections

Compute the Correlation

Transform of each variables

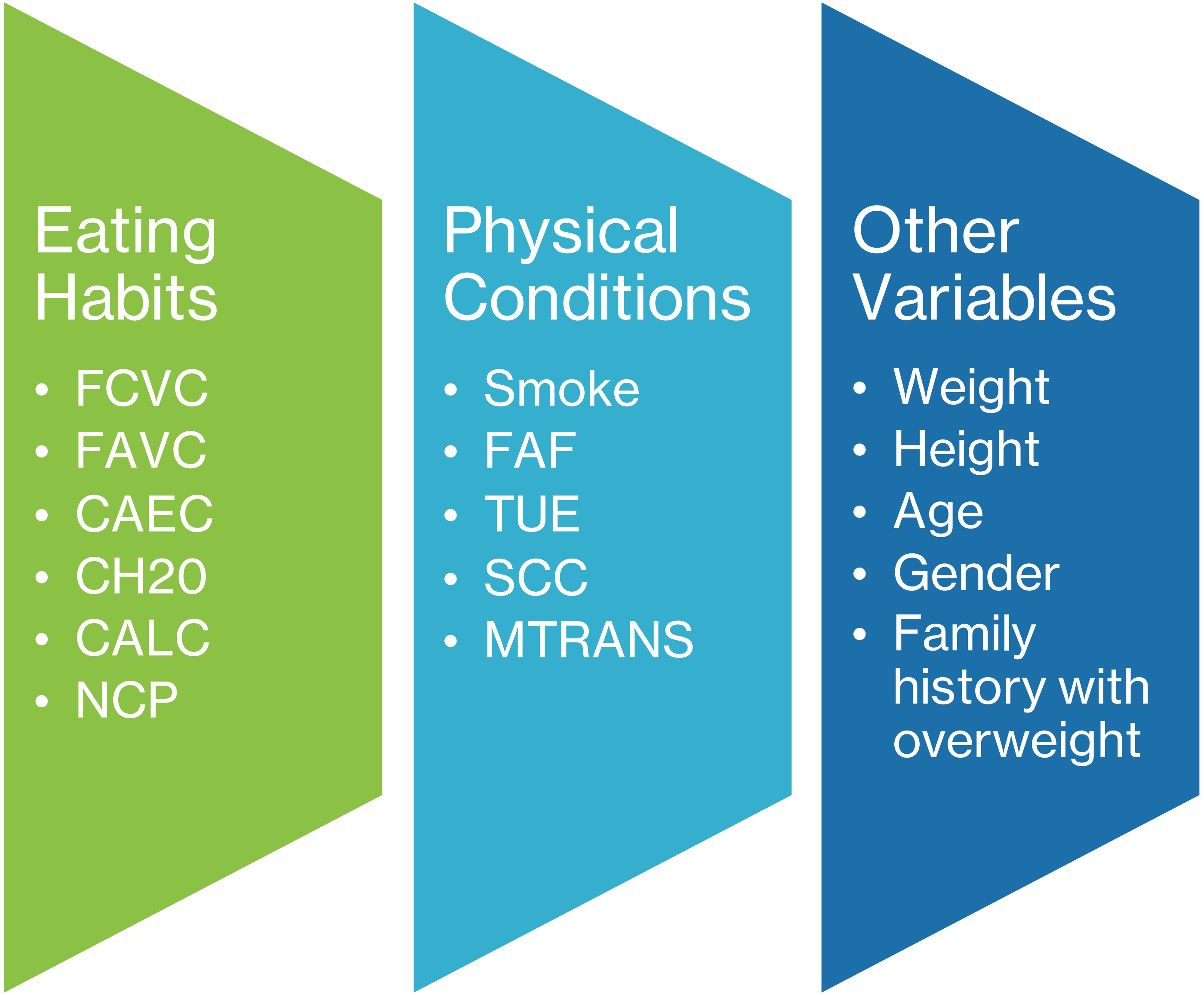
- Frequent consumption of high caloric food (FAVC)
- Frequency of consumption of vegetables (FCVC)
- Number of main meals (NCP)
- Consumption of food between meals (CAEC)
- Consumption of water daily (CH20)
- Consumption of alcohol (CALC)
- Calories consumption monitoring (SCC)
- Physical activity frequency (FAF)
- Time using technology devices (TUE)
- Transportation used (MTRANS)

Dependent Variable



NObeyesdad

- Insufficient Weight
- Normal Weight
- Obesity Type I
- Obesity Type II
- Obesity Type III
- Overweight Level I
- Overweight Level II



Eating Habits

- FCVC
- FAVC
- CAEC
- CH20
- CALC
- NCP

Physical Conditions

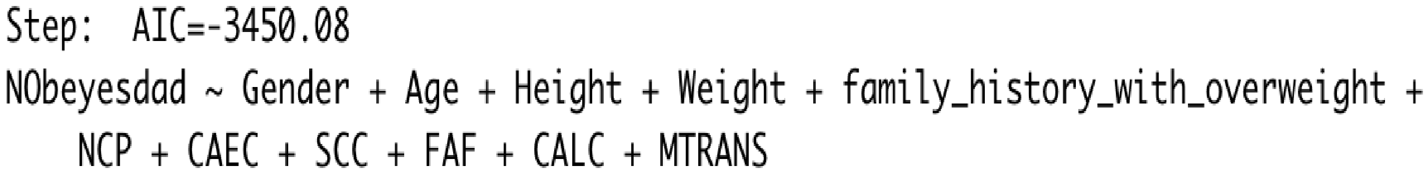
- Smoke
- FAF
- TUE
- SCC
- MTRANS

Other Variables

- Weight
- Height
- Age
- Gender
- Family history with overweight

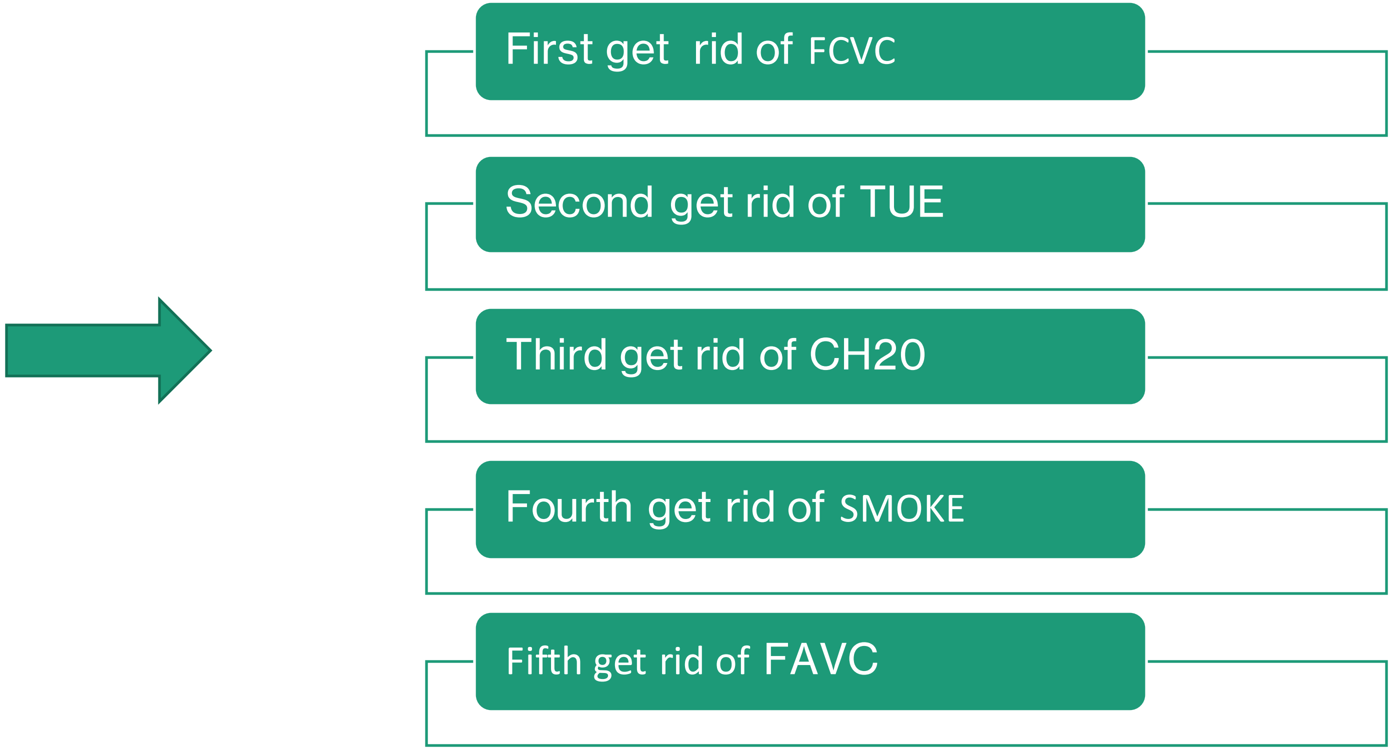
Independent variables

Backward Selection



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

But is it true?



First get rid of FCVC

Second get rid of TUE

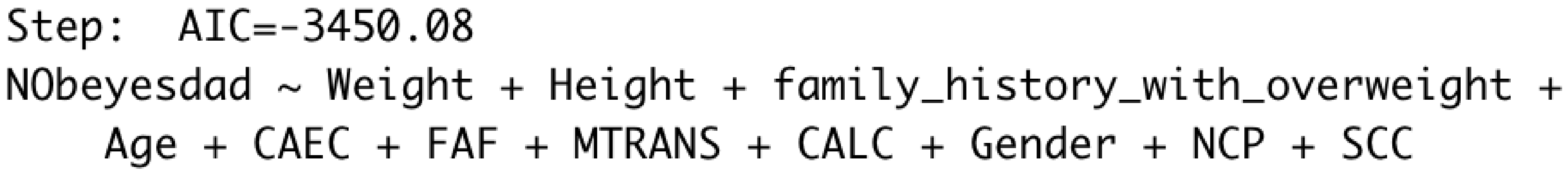
Third get rid of CH20

Fourth get rid of SMOKE

Fifth get rid of FAVC

Forward Selection

- Since forward and back ward selection are the same, and the AIC are all equal to the –3450.08
- But are the independent variables all have the siginifance influence?



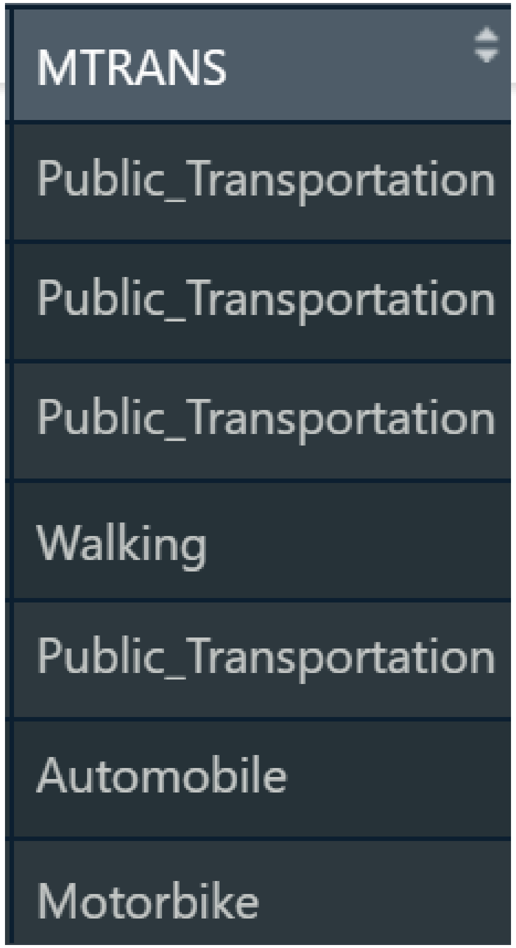
```
Step:  AIC=-3450.08
NObeyesdad ~ Weight + Height + family_history_with_overweight +
    Age + CAEC + FAF + MTRANS + CALC + Gender + NCP + SCC
```

Compute the correlation

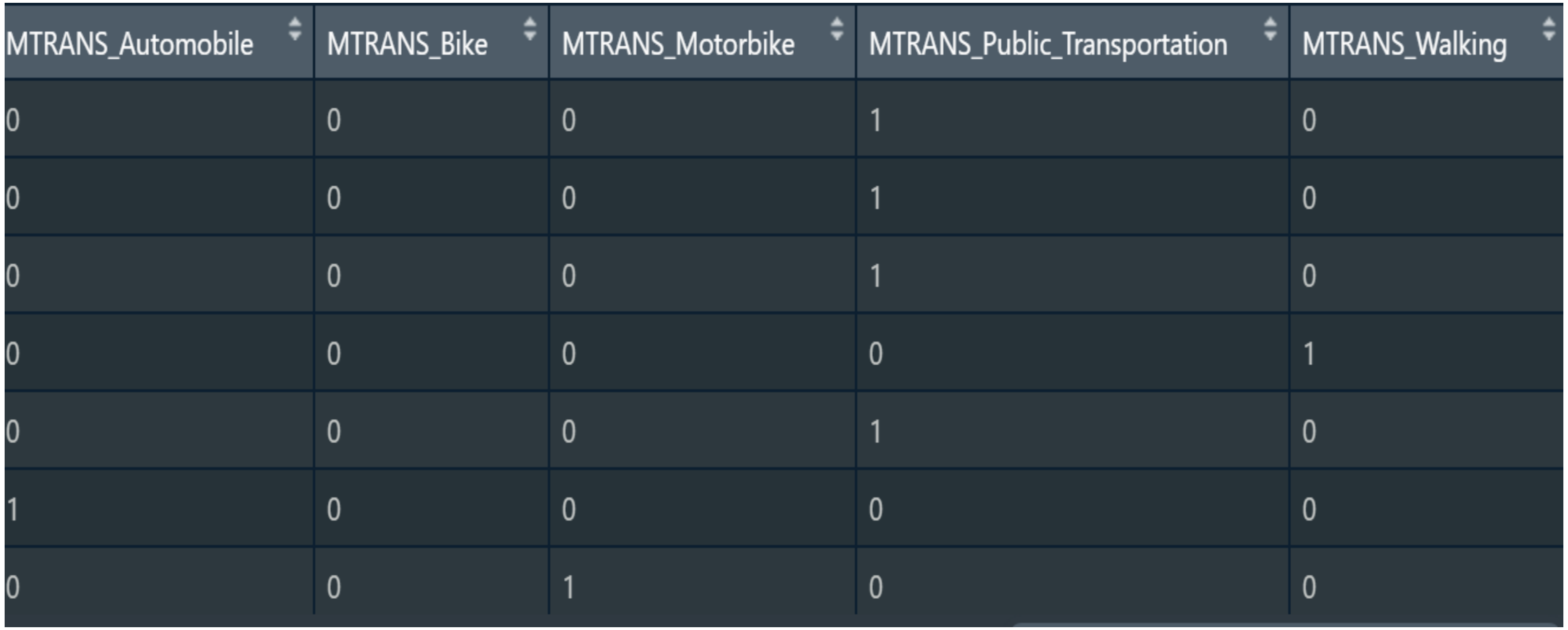
1. Age:0.2829
2. Height:0.13356
3. Weight:0.91325
4. Family_history_with_overweight:0.5051(binary)
5. NCP:0.02669
6. CAEC:0.3293(category)
7. CALC:0.15175(category)

Classification For Obesity Level -- KNN

- How can we determine which Obesity Level are you ?
- Applying Some Data Transformation --- Dummy Variables
- KNN --- Model to Classify Obesity Level with All Features included
- How About "Age", "Height", "Weight" ?

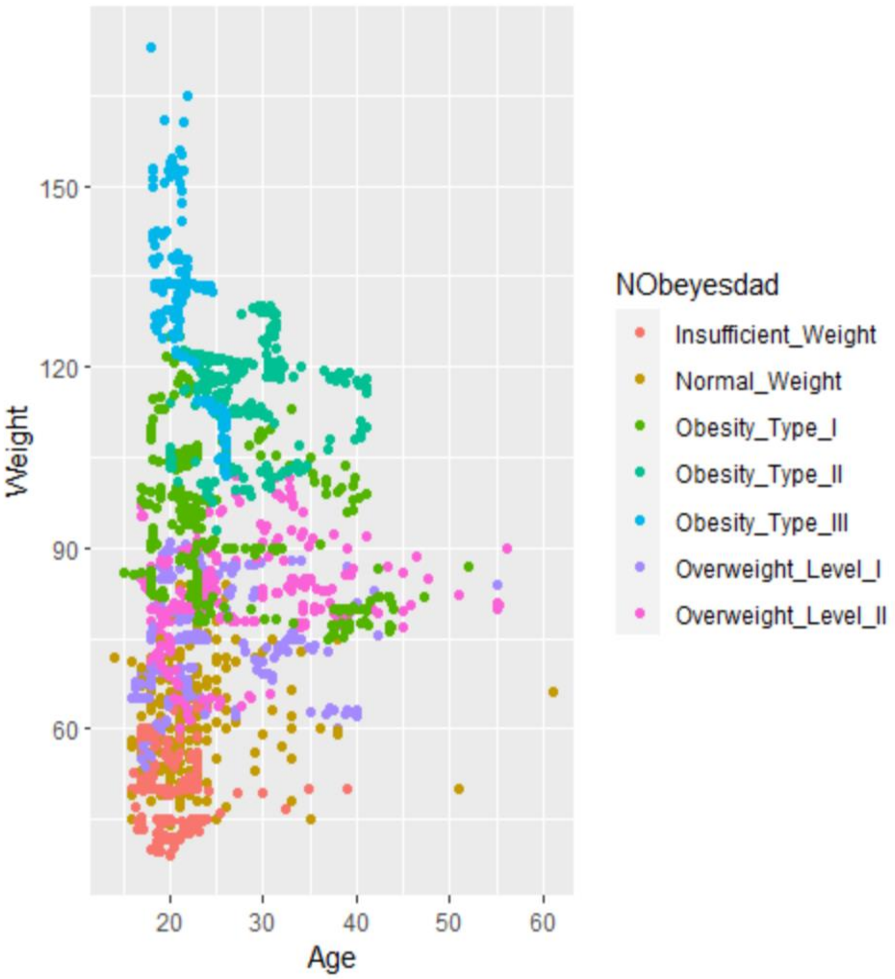


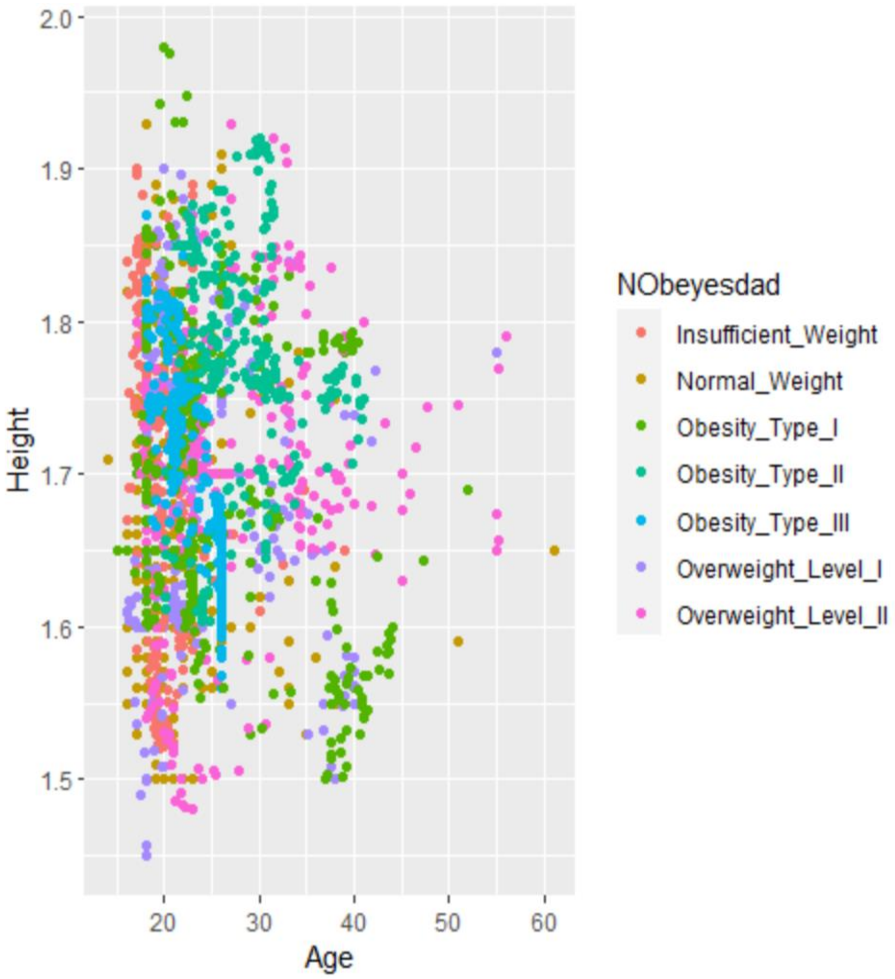
MTRANS
Public_Transportation
Public_Transportation
Public_Transportation
Walking
Public_Transportation
Automobile
Motorbike



MTRANS_Automobile	MTRANS_Bike	MTRANS_Motorbike	MTRANS_Public_Transportation	MTRANS_Walking
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	0	1
0	0	0	1	0
1	0	0	0	0
0	0	1	0	0

Graphic Display of "Height" and "Weight"





Apply Some Combinations For Height and Weight , In KNN

HW1 = Height / Weight

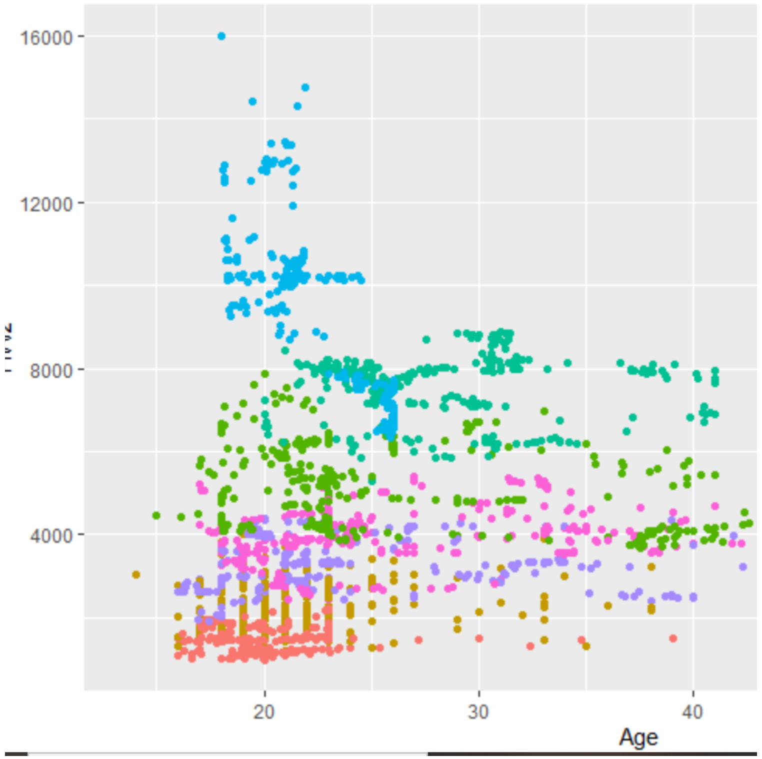
HW2 = Weight ^2 / Height

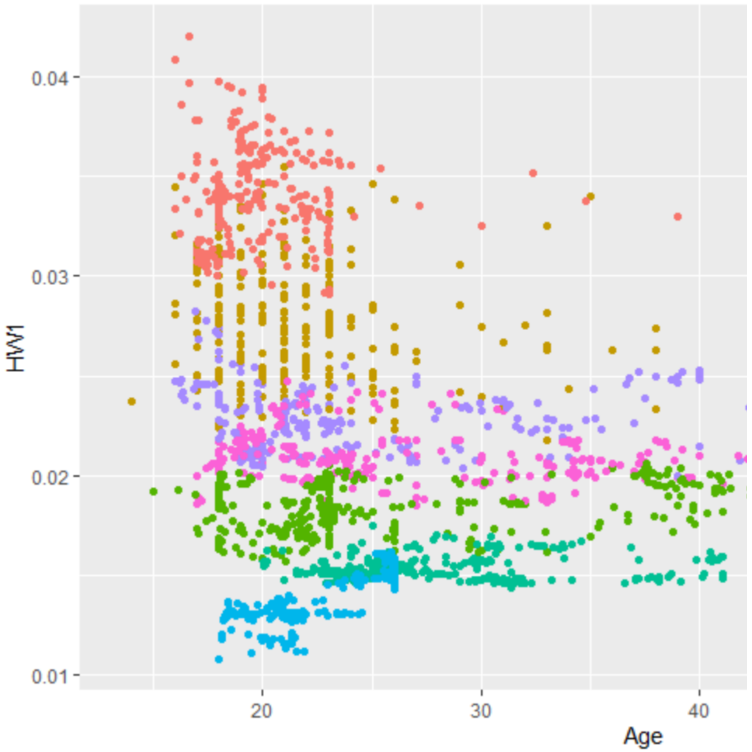
HW3 = Height * Weight / Age ^2

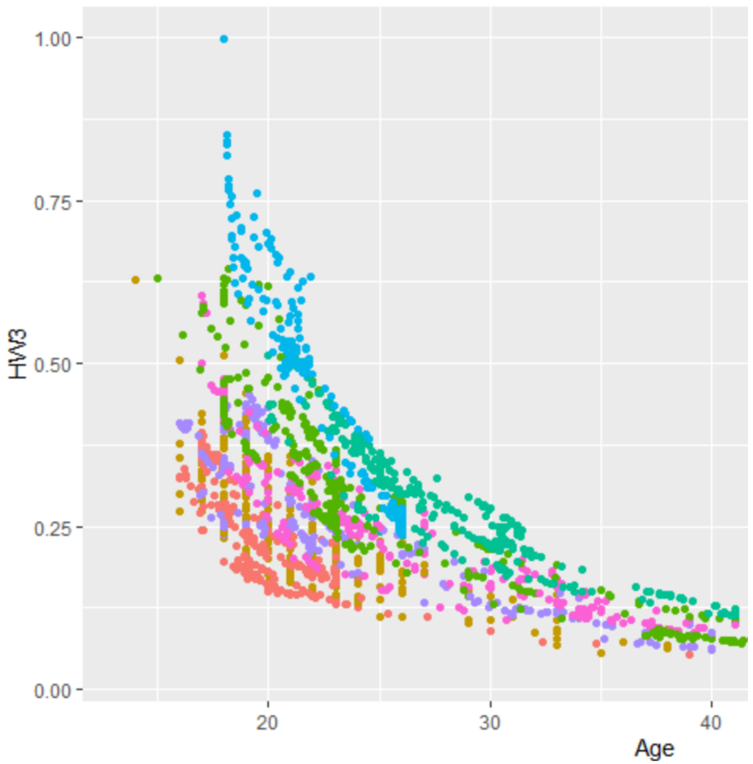
The Accuracy of KNN Model with these Combinations is Better than previous

HW1, HW2, HW3

- **IS "Age" Really Important ?**
- **Maybe Not ~**



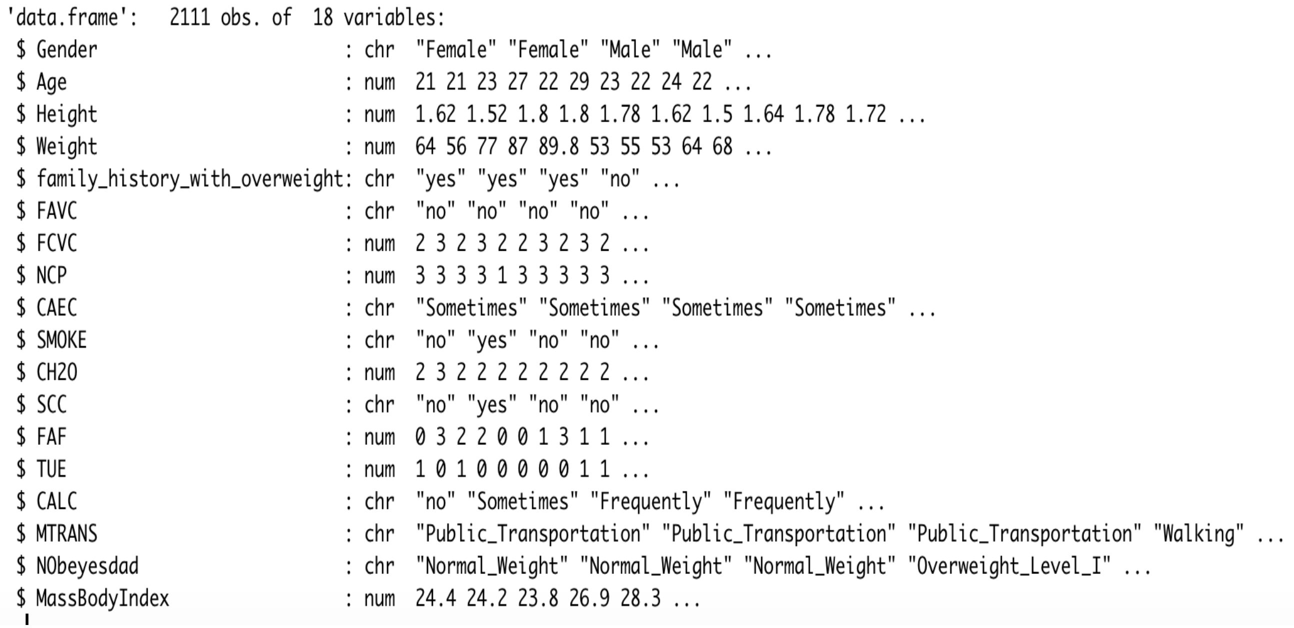




Classification For Obesity Level – C5.0

- **Why C5.0?**
- **Dataset Used**
- **How "height" and "weight" influence Obesity Level**
- **Model generated**

- Original dataset without dummy variables
- Add a new variable "BMI" --- measure of how height and weight influence obesity level together
- Model with only Height, Weight, BMI calculated by height and weight and response NObeyesdad.



```
'data.frame':	2111 obs. of  18 variables:
 $ Gender                        : chr  "Female" "Female" "Male" "Male" ...
 $ Age                           : num  21 21 23 27 22 29 23 22 24 22 ...
 $ Height                        : num  1.62 1.52 1.8 1.8 1.78 1.62 1.5 1.64 1.78 1.72 ...
 $ Weight                        : num  64 56 77 87 89.8 53 55 53 64 68 ...
 $ family_history_with_overweight: chr  "yes" "yes" "yes" "no" ...
 $ FAVC                          : chr  "no" "no" "no" "no" ...
 $ FCVC                          : num  2 3 2 3 2 2 3 2 3 2 ...
 $ NCP                           : num  3 3 3 3 1 3 3 3 3 3 ...
 $ CAEC                          : chr  "Sometimes" "Sometimes" "Sometimes" "Sometimes" ...
 $ SMOKE                         : chr  "no" "yes" "no" "no" ...
 $ CH2O                          : num  2 3 2 2 2 2 2 2 2 2 ...
 $ SCC                           : chr  "no" "yes" "no" "no" ...
 $ FAF                           : num  0 3 2 2 0 0 1 3 1 1 ...
 $ TUE                           : num  1 0 1 0 0 0 0 0 1 1 ...
 $ CALC                          : chr  "no" "Sometimes" "Frequently" "Frequently" ...
 $ MTRANS                        : chr  "Public_Transportation" "Public_Transportation" "Public_Transportation" "Walking" ...
 $ NObeyesdad                    : chr  "Normal_Weight" "Normal_Weight" "Normal_Weight" "Overweight_Level_I" ...
 $ MassBodyIndex                 : num  24.4 24.2 23.8 26.9 28.3 ...
```

Apply some combinations of Weight and Height to calculate BMI

Model Performance

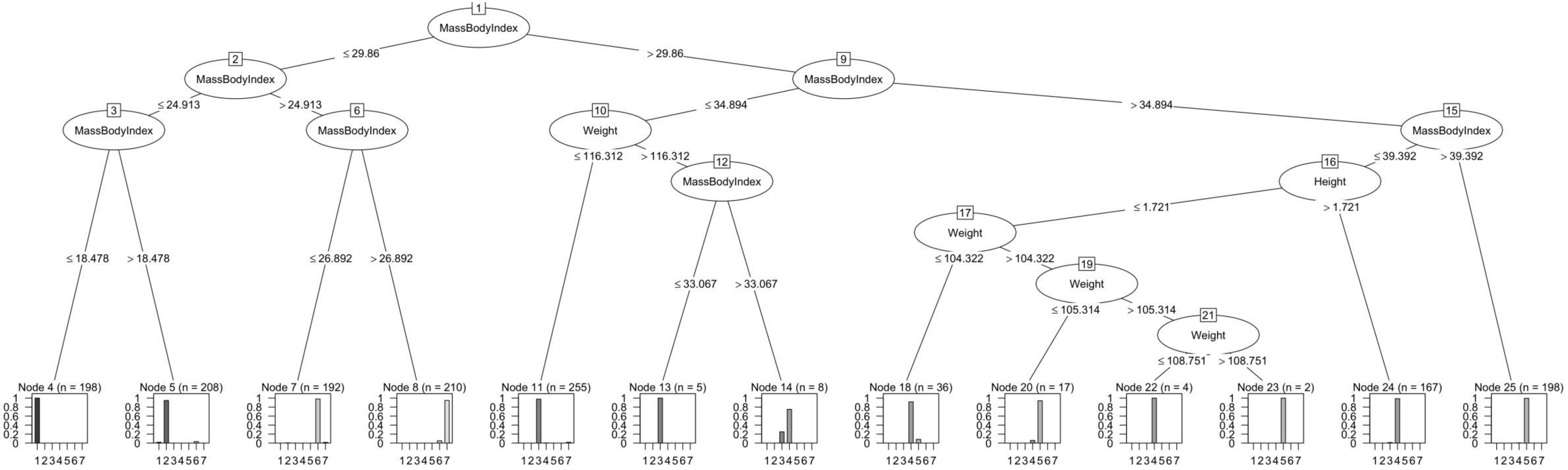
BMI = Weight / Height

BMI = Weight / Height^2

BMI = Weight / Height^3

Decision Tree

- "1" -- Insufficient_Weight
- "2" -- Normal_Weight
- "3" -- Obesity_Type_I
- "4" -- Obesity_Type_II
- "5" -- Obesity_Type_III
- "6" -- Overweight_Level_I
- "7" -- Overweight_Level_II



Model with All Attributes

- **Pros**

➢ Better performance

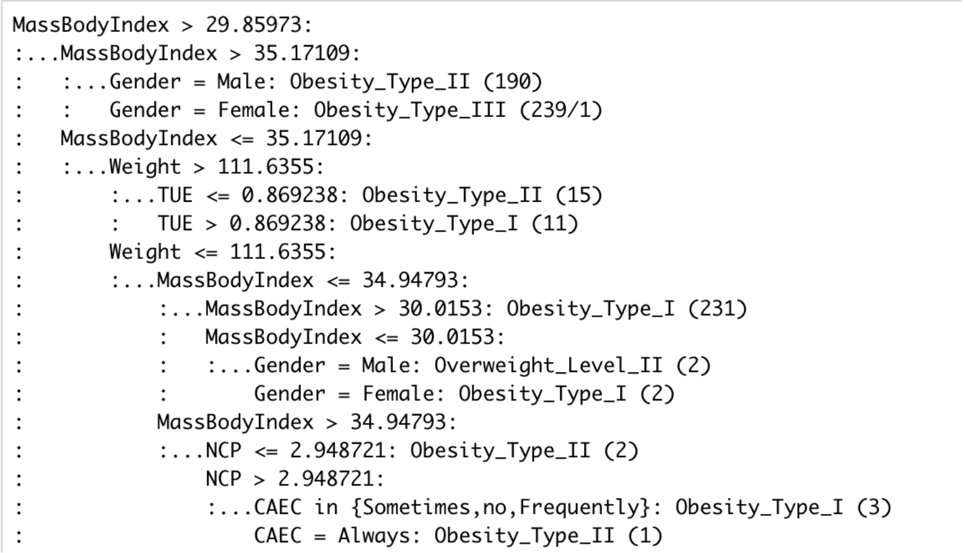
➢ Higher Accuracy

- **Cons**

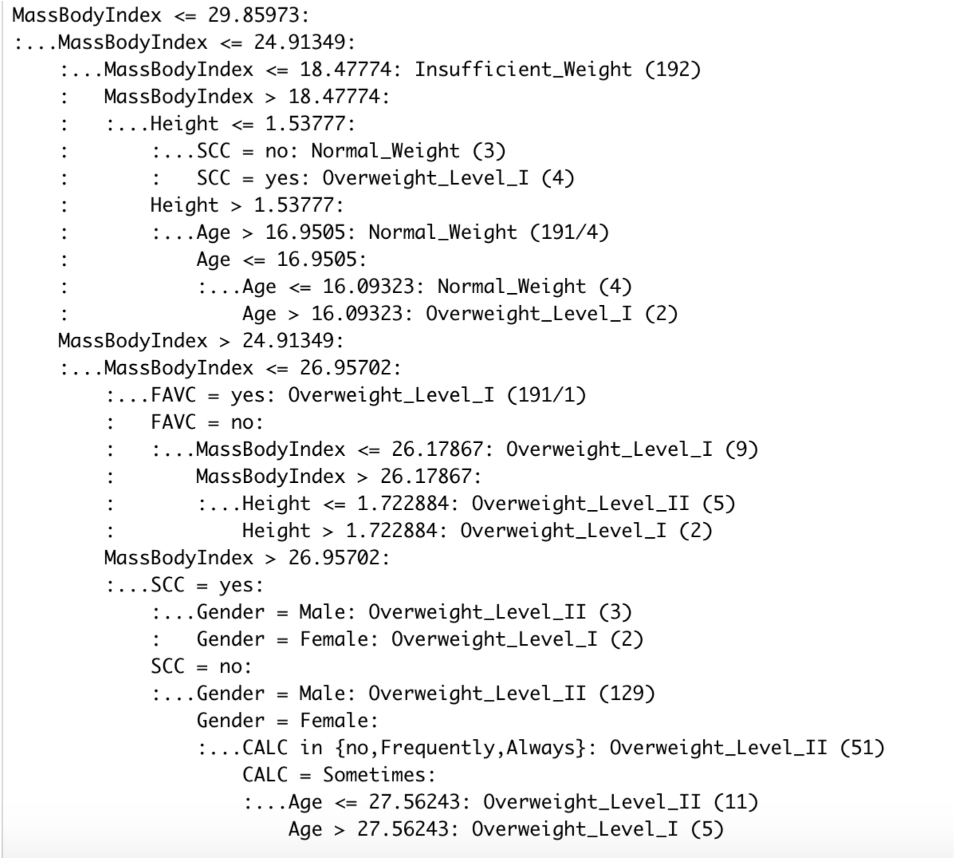
➢ Overfit

➢ Some Attributes are only used for a few times

Decision Tree:



```
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)
```



```
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)
```

BMI and Obesity Level

Best BMI Formula Found

- BMI = weight / height ^ 2

Insufficient_Weight	BMI <= 18.5
Normal_Weight	18.5 < BMI <= 24.9
Overweight_Level_I	24.9 < BMI <= 26.9
Overweight_Level_II	26.9 < BMI <= 29.9
Obesity_Type_I	29.9 < BMI <= 34.9
Obesity_Type_II	34.9 < BMI <= 39.9
Obesity_Type_III	BMI > 39.9

Hypothesis Test

- try to figure out which attributes influence BMI and which is not

- U1 = means of the BMI in the smoking group
- U2 = mean of the BMI in non-smoking group

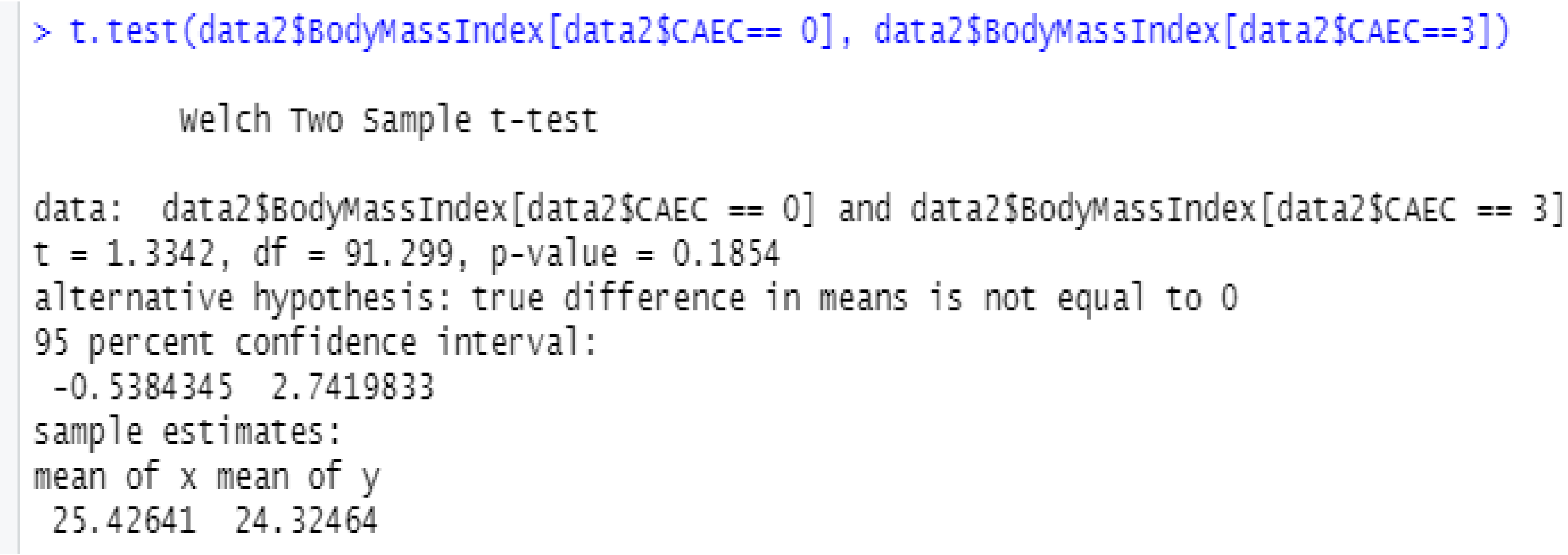
- H0 : U1 = U2
- H1: U1 is not equal to U2
- SMOKE: p- values of t-test is: 0.9639
- The p-value is > 0.05, we conclude that the means are same
- Won't influence BMI

- Could also test categorical variables:

- Recall: CAEC has responses with
- "no", "Sometimes", "Frequently", and "Always"

CAEC : Do you eat any food between meals?

- The p-value is large only when we compare {never , always}. This result shows that BMI will not be influenced by either never eat any food, or always eat between meals.



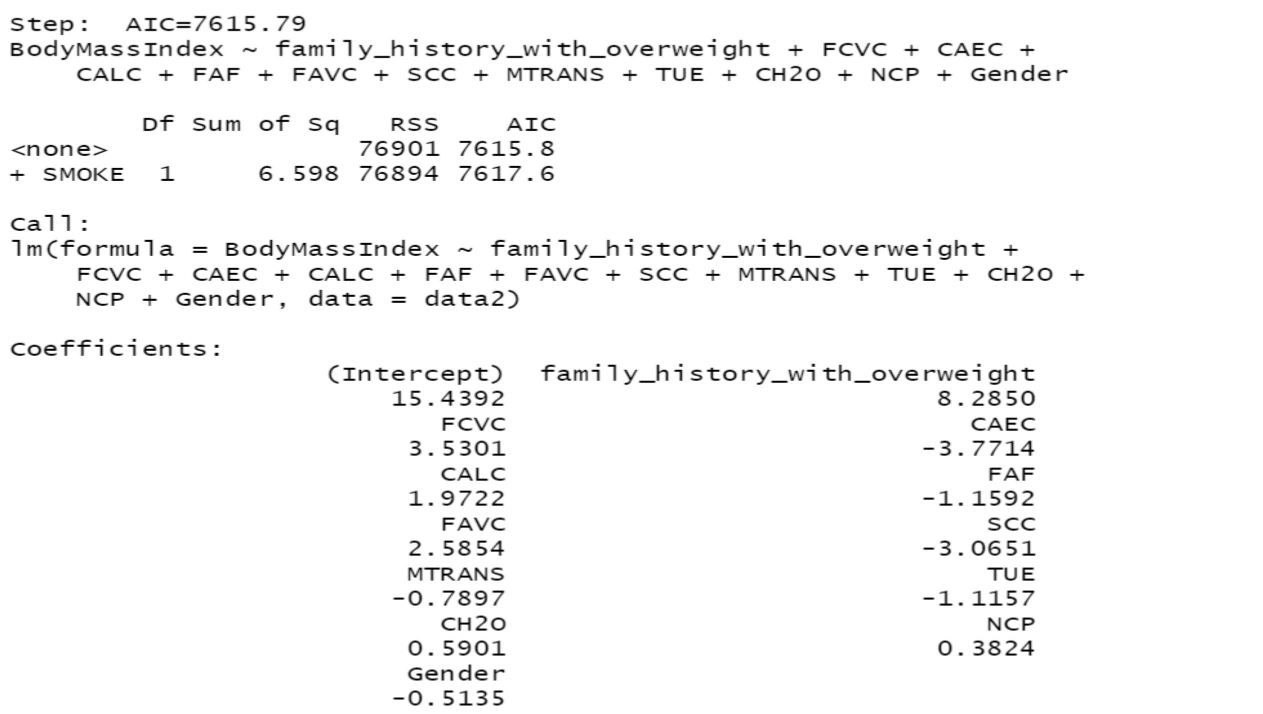
```
> t.test(data2$BodyMassIndex[data2$CAEC== 0], data2$BodyMassIndex[data2$CAEC==3])

	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 
```

- **Most of the test results have the smaller p values, which implies influence BMI. But we couldn't order the importance levels by the hypothesis test.**

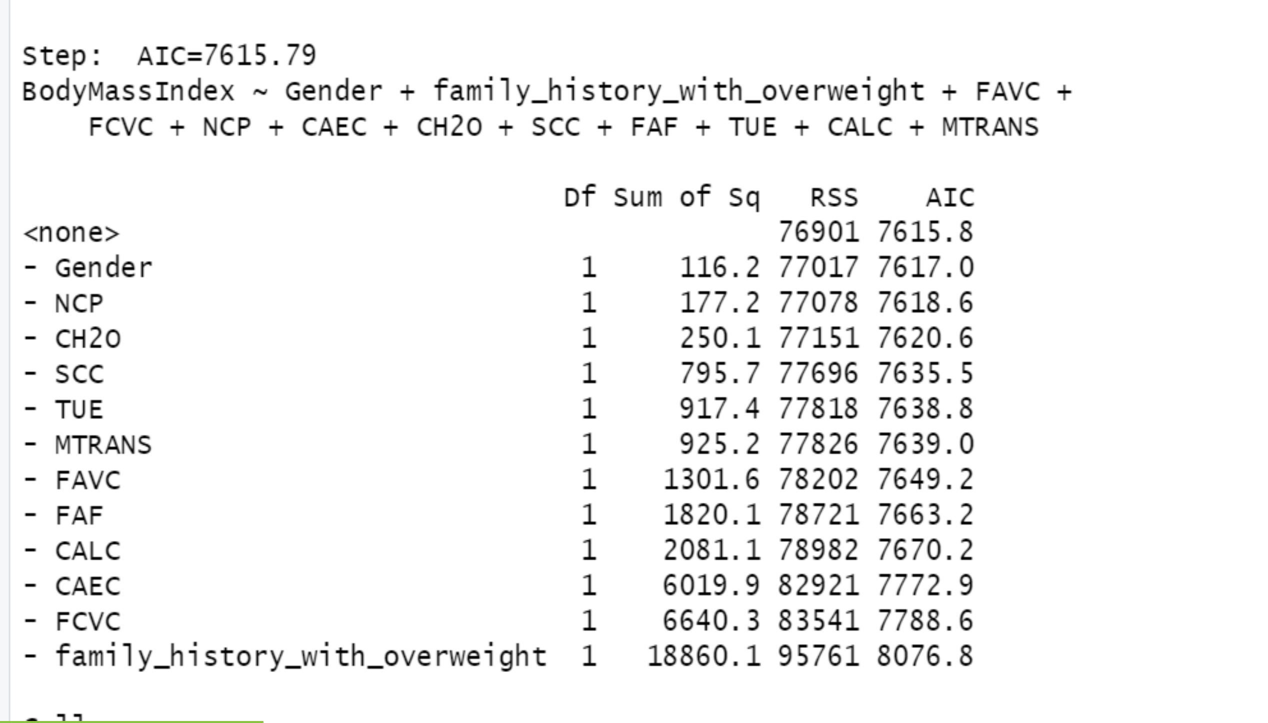
Continues regression based on BMI

1.why using BMI?
2.Which variables have more effect on BMI?

- Create BodyMassIndex(BMI) as new response y instead of NObeyesdad.
- Remodel by using stepwise method: Backward and forward
- Further improving
- correlation
- ggplot



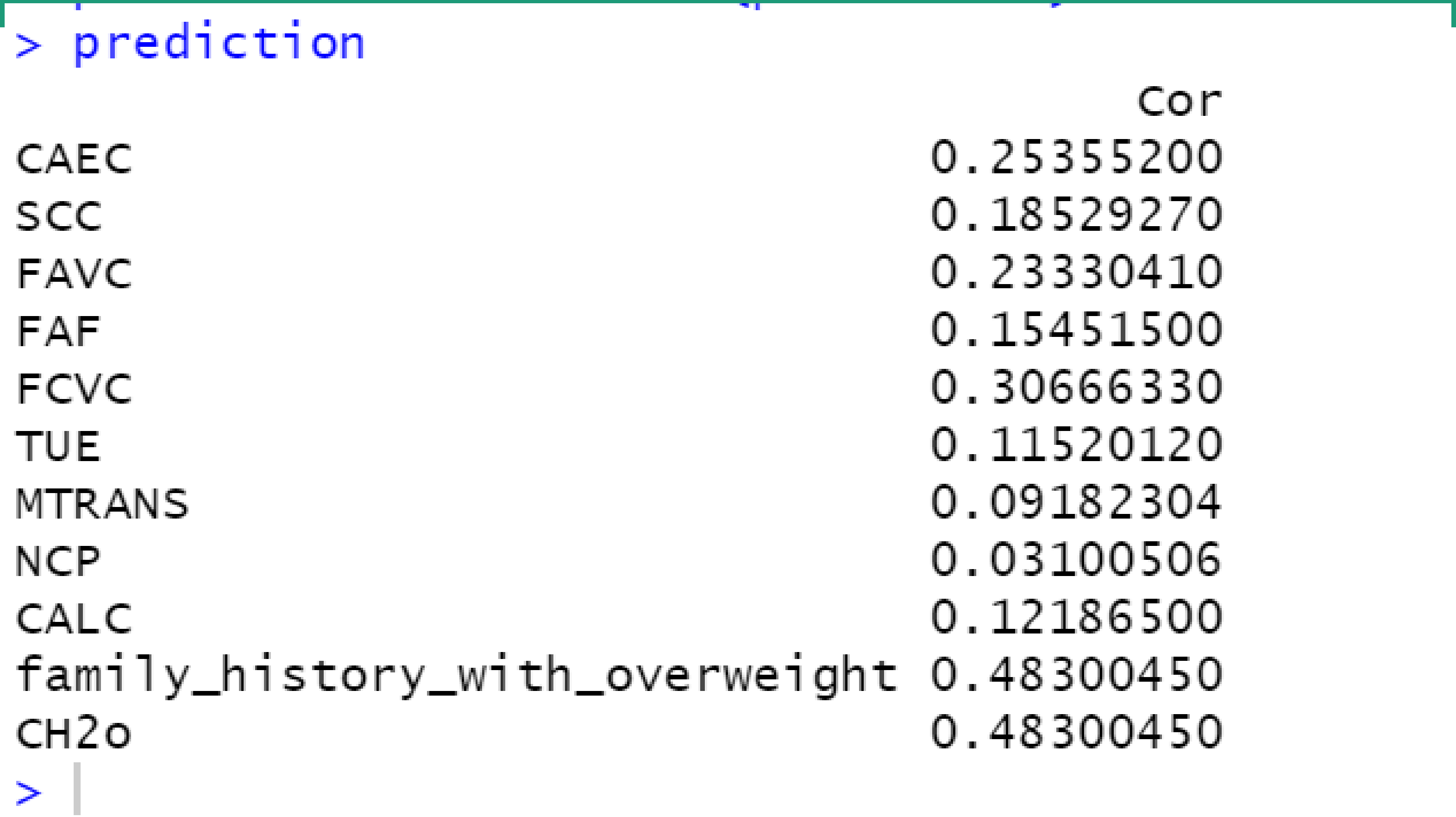
```
Step:  AIC=7615.79
BodyMassIndex ~ family_history_with_overweight + FCVC + CAEC +
    CALC + FAF + FAVC + SCC + MTRANS + TUE + CH2O + NCP + Gender

        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
```



```
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
```

Correlation

Find the higher correlation

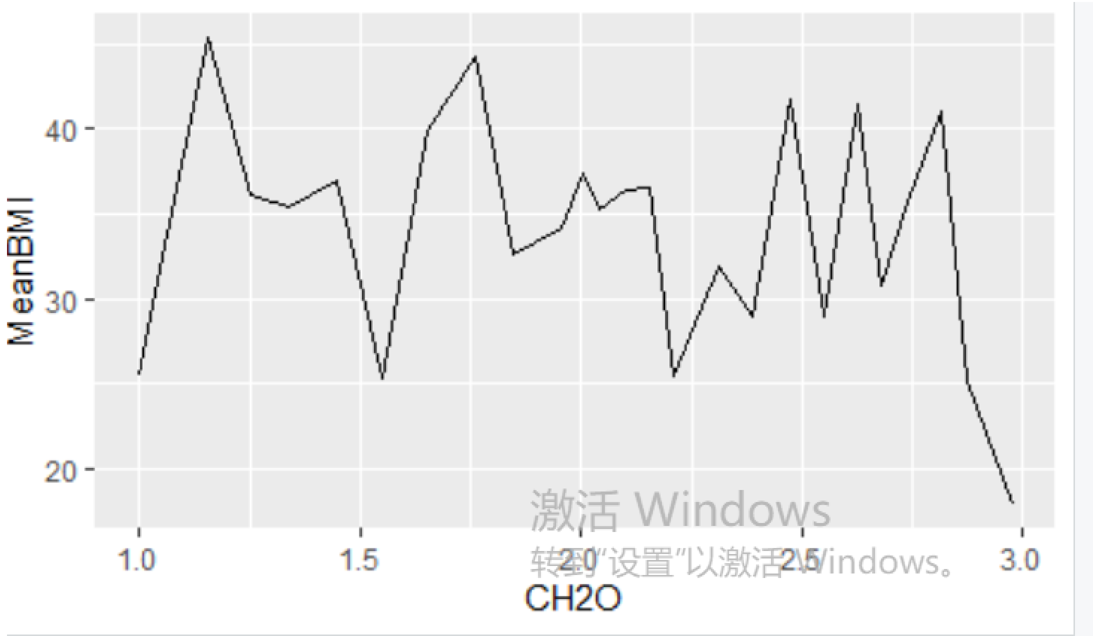


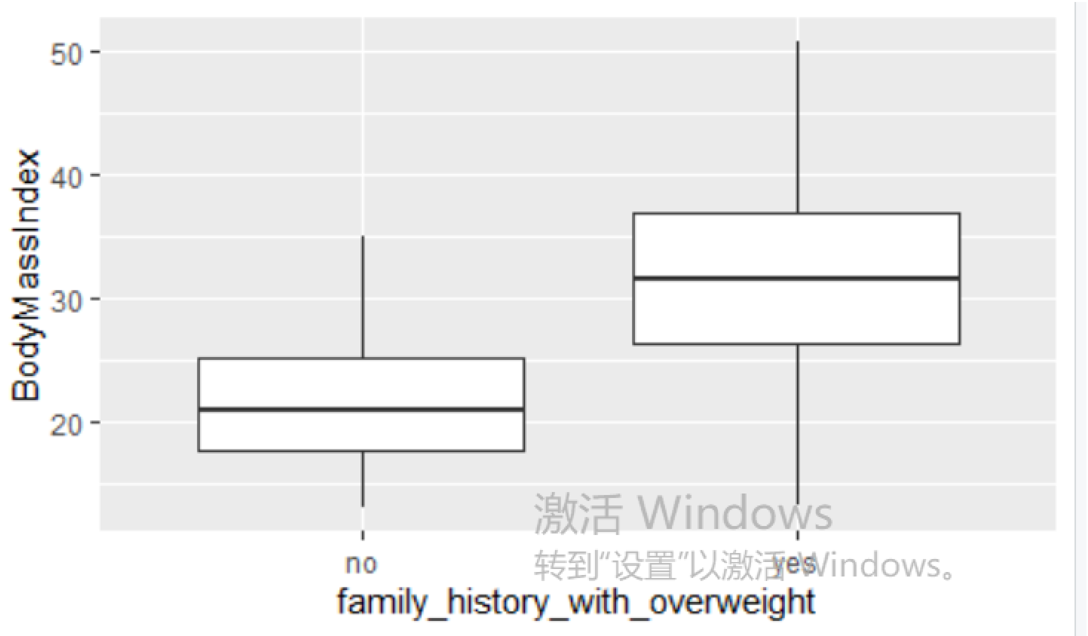
```
> prediction
                                        Cor
CAEC                             0.25355200
SCC                              0.18529270
FAVC                             0.23330410
FAF                              0.15451500
FCVC                             0.30666330
TUE                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

CH2o， family_history_with_overweight,FCVC,CAEC have a higher correlation=more influences on BMI?

Visualize the regression using ggplot()

- Boxplot and lineplot
- Covariates as continuous(1) or discrete(2) since BMI is continuous
- CH2o: y is mean of BMI, x=CH2o
- fmaily_history_with_overweight:y is BMI,x is family_history_with_overweight





Limitation and improvement

Limitation(potential problem):

- **Covariates: binary /categorical/continuous**
- **It will have lack of accuracy since doing the linear regression**

EX: family_history_overweight be binary(0,1) when doing stepwise, it will result at weakness

Improvement:

Use the ggplot() to improve

- **Change family_history_overweight to a discrete variables such as (no,yes)**
- **for CH2o . Assume x between(0,1) be low consumption ,(1,2) be median...**

covariates with different classification ,we have to define them first for different methods.

Conclusion

- Regression and Classification
- Height and Weight shows great influence on Obesity Level
- BMI is a measure to indicate Obesity Level, but not perfect.