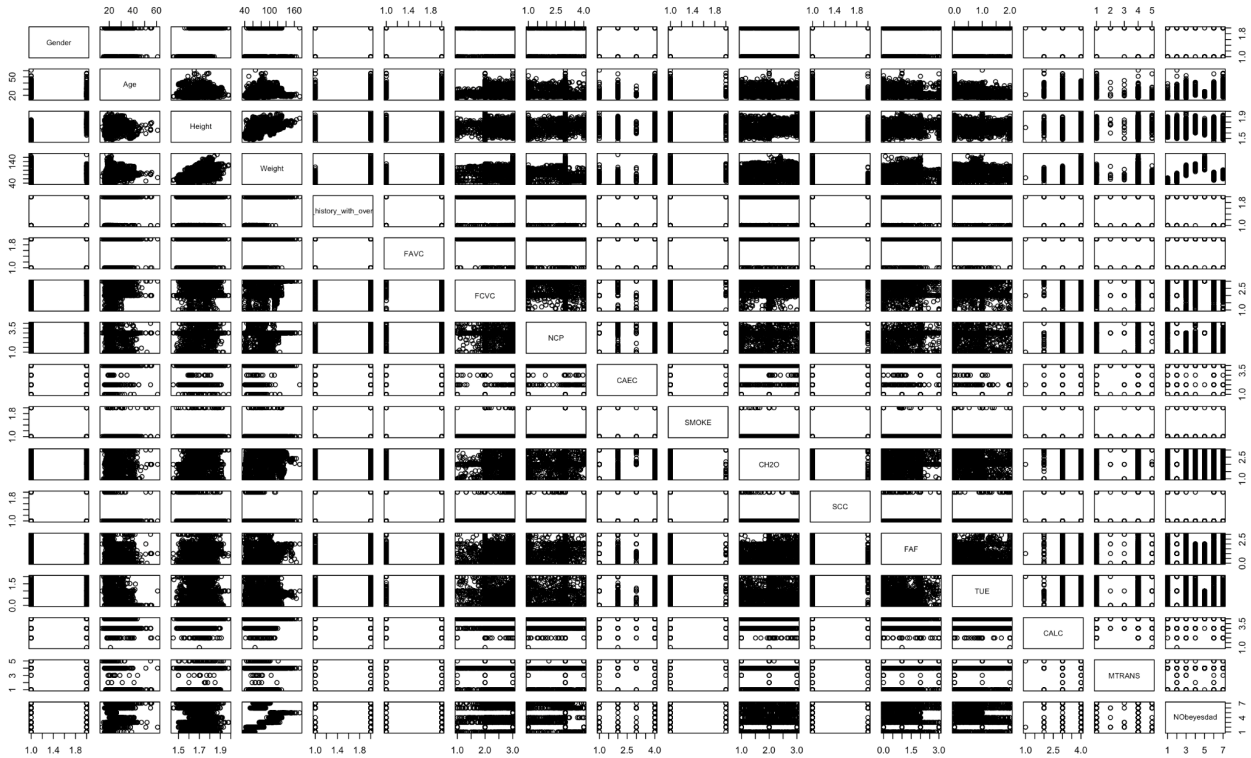
CSCI 6444: Big Data and Analytics

Class Project-2

Exploring Variations in Clustering and Predictive Analysis

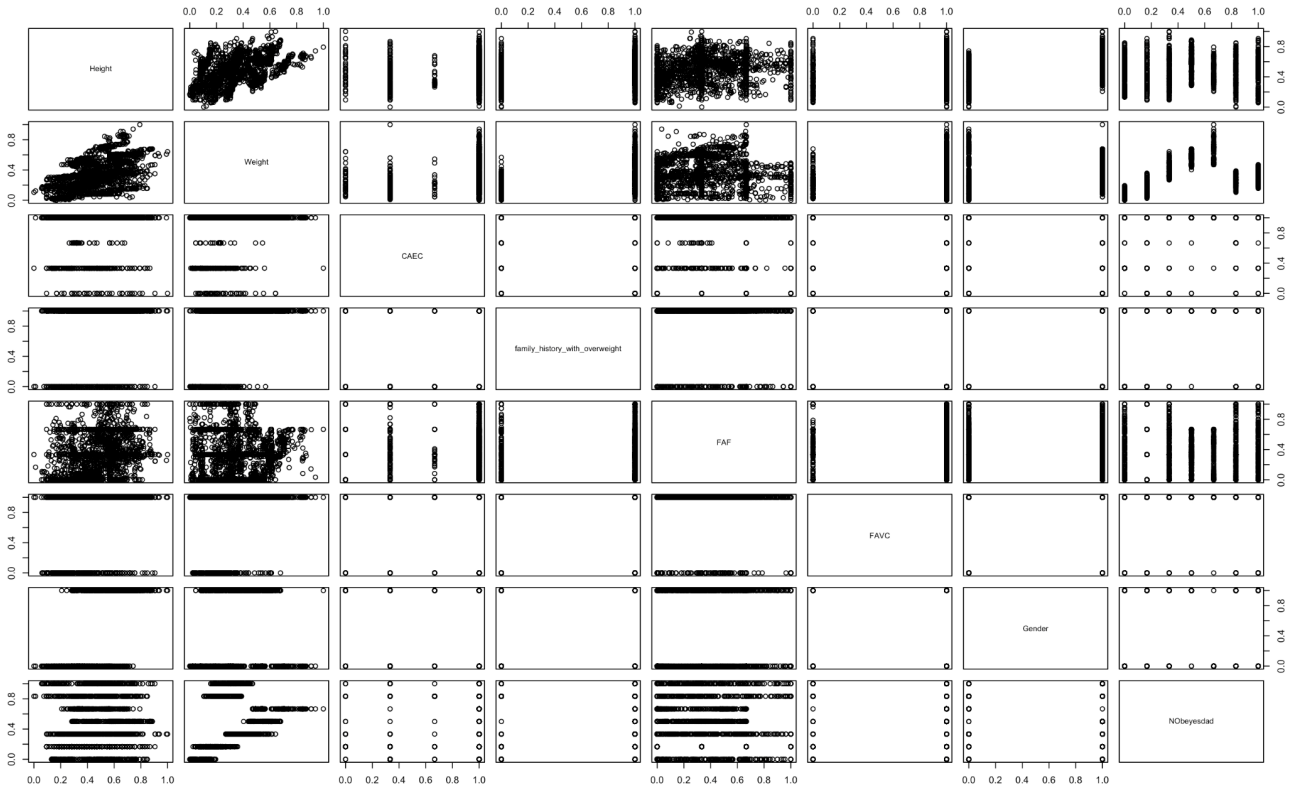
1. Plotting the data to draw relationships between the attributes

a. (i) Pairwise plot for the whole dataset

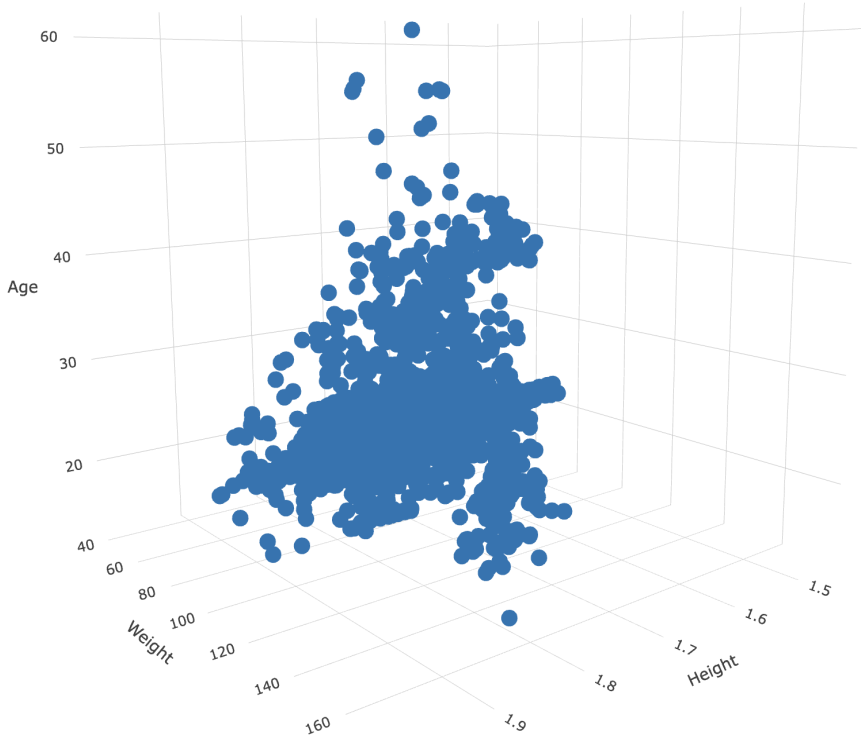


The pairwise plot reveals a few vaguely linear relationships (Height v/s. Weight), however since there are over 2000 data points, most of the plots are very convoluted.

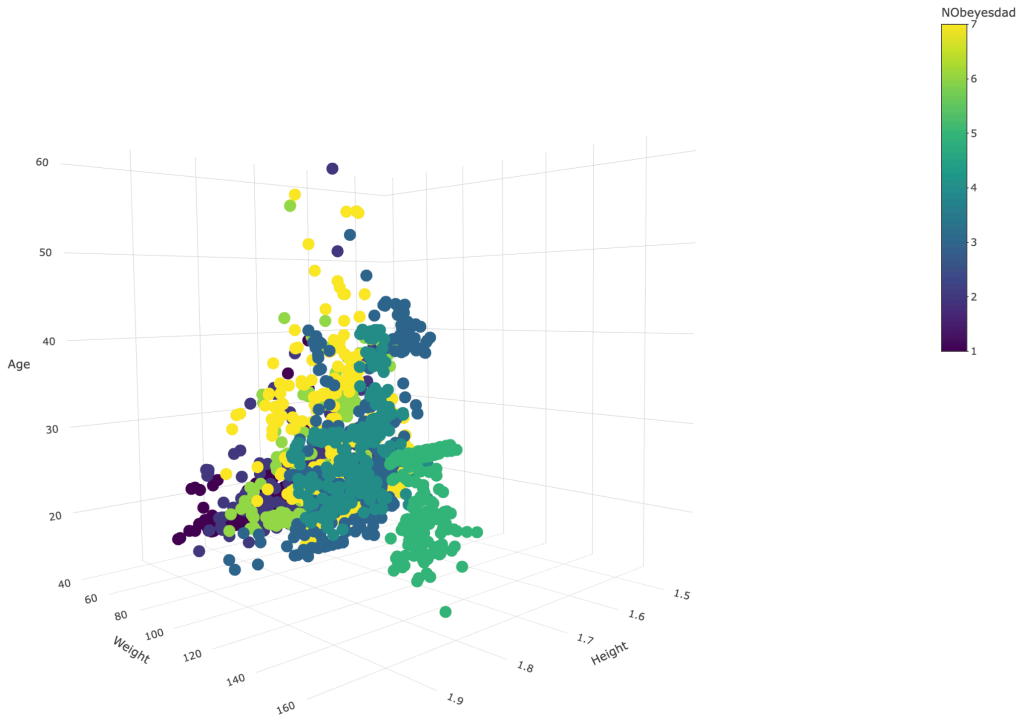
(ii) Pairwise plot on the 7 most relevant variables



(iii) 3D plot (Height v/s. Weight v/s. Age)



(iv) 3D plot (Height v/s. Weight v/s. Age v/s. NObeyesdad)



The above plot (Height v/s. Weight v/s. Age v/s. NObeyesdad) gives us a good idea of how the obesity levels are closely related to the height, weight and age of a person, the levels are concentrated and exist in a cluster.

b. Pairs of correlated attributes

Top Correlations	Negative Correlations	Target Variable Correlations
Height - Gender: 0.6184663	MTRANS - Age: -0.6019452	NObeyesdad - Weight: 0.3876425
FHWO - Weight: 0.4968204	TUE - Age: -0.2969306	NObeyesdad - CAEC: 0.3360195
Height - Weight: 0.4631361	FCVC - Gender: -0.2745048	NObeyesdad - FHWO: 0.3136670
CAEC - Weight: 0.3915363		NObeyesdad - FAF: -0.129564308
CAEC - FHWO: 0.3030188		NObeyesdad - Age: 0.236170353
FAF - Height: 0.2947090		
FAVC - Weight: 0.2723005		

The pairs of attributes that look to be correlated are:

Height, weight, and gender are strongly related. Weight is influenced by family history, food in between meals (snacking), and obesity level. Age affects lifestyle choices, such as transportation and technology use. Obesity is weakly related to physical activity but more influenced by weight, snacking, and family history.

2. Preparing the data

a. Statistics of the dataset

```
> summary(obesity)
    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
```

```
> str(obesity)
'data.frame':   2111 obs. of  17 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" ...
```

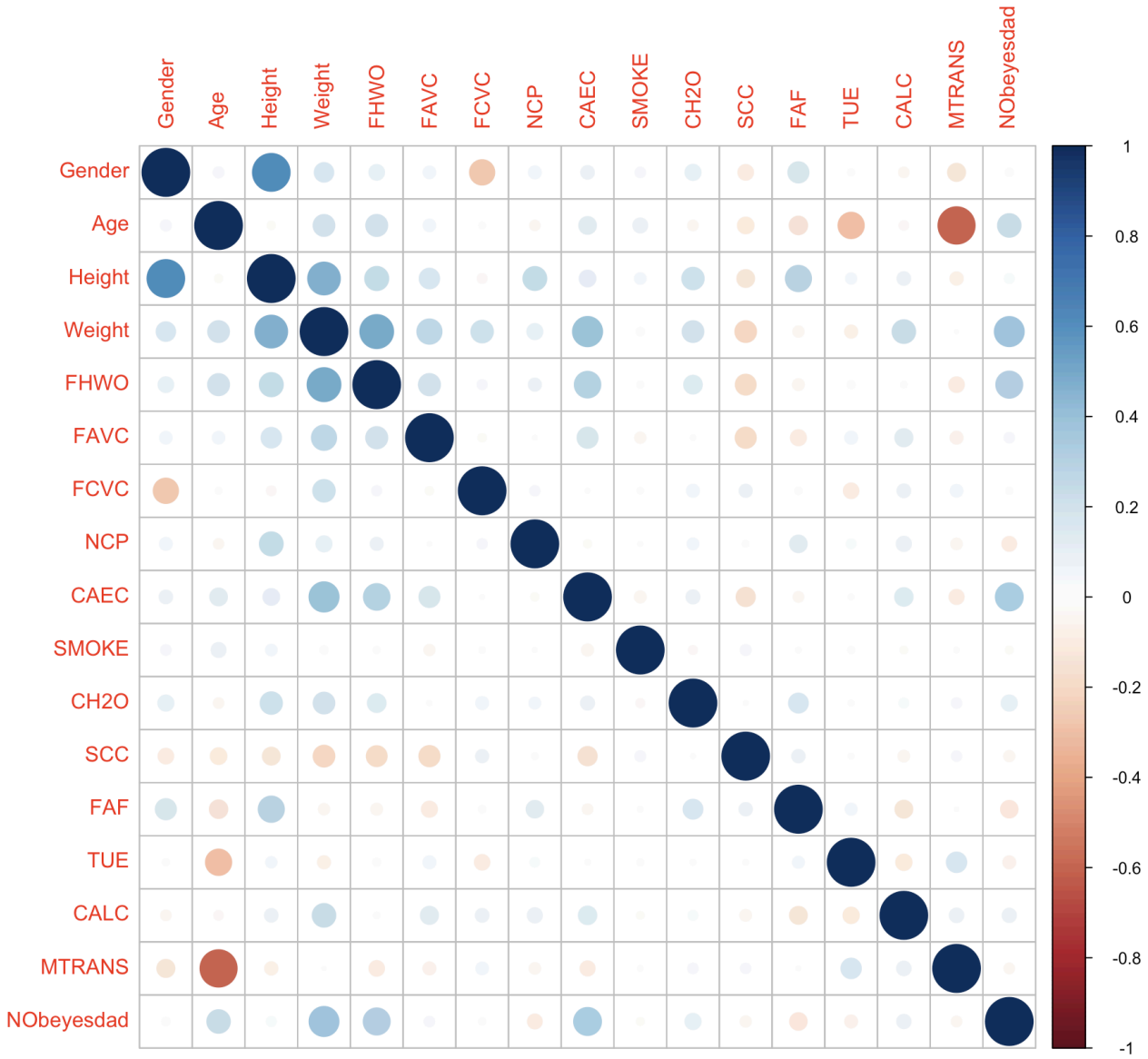
```
> describe(obesity)
                                vars    n  mean    sd median trimmed   mad   min    max  range  skew kurtosis   se
Gender*                            1 2111  1.51  0.50   2.00    1.51  0.00  1.00   2.00   1.00 -0.02    -2.00 0.01
Age                                2 2111 24.31  6.35  22.78   23.34  4.78 14.00  61.00  47.00  1.53     2.81 0.14
Height                             3 2111  1.70  0.09   1.70    1.70  0.10  1.45   1.98   0.53 -0.01    -0.57 0.00
Weight                             4 2111 86.59 26.19  83.00   85.82 32.22 39.00 173.00 134.00  0.26    -0.70 0.57
family_history_with_overweight*    5 2111  1.82  0.39   2.00    1.90  0.00  1.00   2.00   1.00 -1.64     0.70 0.01
FAVC*                              6 2111  1.88  0.32   2.00    1.98  0.00  1.00   2.00   1.00 -2.40     3.74 0.01
FCVC                               7 2111  2.42  0.53   2.39    2.46  0.57  1.00   3.00   2.00 -0.43    -0.64 0.01
NCP                                8 2111  2.69  0.78   3.00    2.77  0.00  1.00   4.00   3.00 -1.11     0.38 0.02
CAEC*                              9 2111  3.67  0.78   4.00    3.87  0.00  1.00   4.00   3.00 -2.13     3.06 0.02
SMOKE*                            10 2111  1.02  0.14   1.00    1.00  0.00  1.00   2.00   1.00  6.70    42.95 0.00
CH2O                              11 2111  2.01  0.61   2.00    2.01  0.67  1.00   3.00   2.00 -0.10    -0.88 0.01
SCC*                              12 2111  1.05  0.21   1.00    1.00  0.00  1.00   2.00   1.00  4.36    17.02 0.00
FAF                               13 2111  1.01  0.85   1.00    0.94  1.19  0.00   3.00   3.00  0.50    -0.62 0.02
TUE                               14 2111  0.66  0.61   0.63    0.59  0.72  0.00   2.00   2.00  0.62    -0.55 0.01
CALC*                             15 2111  3.63  0.55   4.00    3.70  0.00  1.00   4.00   3.00 -1.17     0.46 0.01
MTRANS*                           16 2111  3.37  1.26   4.00    3.55  0.00  1.00   5.00   4.00 -1.28    -0.20 0.03
NObeyesdad*                       17 2111  4.02  1.95   4.00    4.02  2.97  1.00   7.00   6.00  0.01    -1.19 0.04
```

The obesity dataset contains 2,111 observations and 17 attributes, which are a mix of numerical and categorical variables. The dataset represents a broad spectrum of individuals in terms of age (14 - 61), height (1.45m - 1.98m) , and weight (39 kg - 173 kg). The dataset includes key lifestyle and hereditary factors influencing obesity (family_history_with_overwieght, FAVC, CAEC). There are diverse exercise levels and sedentary behaviors, making it possible to analyze how lifestyle affects obesity. The ‘NObeyesdad’ variable serves as the classification label for obesity prediction.

b. Subsetting the dataset by picking the most relevant variables

Note: 'FHWO' is 'family_history_with_overweight', the name was shortened for better visualization in the correlation plot.

```
> obesity2cor<-cor(obesity2)
> obesity2cor
                                  Gender         Age      Height       Weight family_history_with_overweight
Gender                         1.00000000  0.04839420  0.61846630  0.161667575                     0.10251213
Age                            0.04839420  1.00000000 -0.02595813  0.202560104                     0.20572533
Height                         0.61846630 -0.02595813  1.00000000  0.463136117                     0.24768389
Weight                         0.16166757  0.20256010  0.46313612  1.000000000                     0.49682038
family_history_with_overweight 0.10251213  0.20572533  0.24768389  0.496820377                     1.00000000
FAVC                           0.06493377  0.06390169  0.17836378  0.272300490                     0.20803551
FCVC                          -0.27450478  0.01629089 -0.03812106  0.216124705                     0.04037225
NCP                            0.06759999 -0.04394373  0.24367173  0.107468988                     0.07136970
CAEC                           0.07715739  0.13164367  0.11677807  0.391536253                     0.30301879
SMOKE                          0.04469809  0.09198745  0.05549938  0.025746413                     0.01738550
CH2O                           0.10792968 -0.04530386  0.21337592  0.200575387                     0.14743661
SCC                           -0.10263348 -0.11628285 -0.13375278 -0.201906340                    -0.18542171
FAF                            0.18960696 -0.14493833  0.29470900 -0.051436270                    -0.05667320
TUE                            0.01726947 -0.29693059  0.05191167 -0.071561359                     0.02294330
CALC                          -0.04543604 -0.03671108  0.07720265  0.238067172                    -0.01476754
MTRANS                        -0.13753730 -0.60194519 -0.07360921  0.004609837                    -0.10153969
NObeyesdad                     0.02490758  0.23617035  0.03898583  0.387642534                     0.31366704
                                      FAVC         FCVC          NCP         CAEC         SMOKE         CH2O
Gender                         0.064933774 -0.274504782  0.067599988  0.077157395  0.044698091  0.107929676
Age                            0.063901686  0.016290886 -0.043943727  0.131643668  0.091987445 -0.045303858
Height                         0.178363778 -0.038121058  0.243671726  0.116778067  0.055499384  0.213375917
Weight                         0.272300490  0.216124705  0.107468988  0.391536253  0.025746413  0.200575387
family_history_with_overweight 0.208035507  0.040372247  0.071369697  0.303018786  0.017385500  0.147436606
FAVC                           1.000000000 -0.027283080 -0.006999943  0.187893194 -0.050659965  0.009719131
FCVC                          -0.027283080  1.000000000  0.042216296 -0.008008518  0.014319529  0.068461472
NCP                           -0.006999943  0.042216296  1.000000000 -0.026022447  0.007811192  0.057087996
CAEC                           0.187893194 -0.008008518 -0.026022447  1.000000000 -0.057911782  0.081346709
SMOKE                         -0.050659965  0.014319529  0.007811192 -0.057911782  1.000000000 -0.031994530
CH2O                           0.009719131  0.068461472  0.057087996  0.081346709 -0.031994530  1.000000000
SCC                           -0.190658309  0.071852192 -0.015623955 -0.156783796  0.047731227  0.008036485
FAF                           -0.107995159  0.019939398  0.129504307 -0.046203176  0.011216029  0.167236492
TUE                            0.068416912 -0.101134846  0.036325572  0.011336000  0.017613134  0.011965338
CALC                           0.137820617  0.078570668  0.095261657  0.144412940 -0.022315358  0.037408541
MTRANS                        -0.069800209  0.064743476 -0.053858095 -0.095256586 -0.010701669  0.044028311
NObeyesdad                     0.044582308  0.018522160 -0.092616323  0.336019467 -0.023256315  0.108868258
                                       SCC          FAF         TUE        CALC       MTRANS  NObeyesdad
Gender                        -0.102633482  0.189606957  0.01726947 -0.04543604 -0.137537298  0.02490758
Age                           -0.116282847 -0.144938327 -0.29693059 -0.03671108 -0.601945191  0.23617035
Height                        -0.133752777  0.294708998  0.05191167  0.07720265 -0.073609212  0.03898583
Weight                        -0.201906340 -0.051436270 -0.07156136  0.23806717  0.004609837  0.38764253
family_history_with_overweight -0.185421707 -0.056673197  0.02294330 -0.01476754 -0.101539690  0.31366704
FAVC                          -0.190658309 -0.107995159  0.06841691  0.13782062 -0.069800209  0.04458231
FCVC                           0.071852192  0.019939398 -0.10113485  0.07857067  0.064743476  0.01852216
NCP                           -0.015623955  0.129504307  0.03632557  0.09526166 -0.053858095 -0.09261632

CAEC                          -0.156783796 -0.046203176  0.01133600  0.14441294 -0.095256586  0.33601947
SMOKE                          0.047731227  0.011216029  0.01761313 -0.02231536 -0.010701669 -0.02325632
CH2O                           0.008036485  0.167236492  0.01196534  0.03740854  0.044028311  0.10886826
SCC                            1.000000000  0.074220664 -0.01092798 -0.05556208  0.043157450 -0.05067879
FAF                            0.074220664  1.000000000  0.05856207 -0.13477105  0.006393953 -0.12956431
TUE                           -0.010927978  0.058562066  1.00000000 -0.11203443  0.176944749 -0.06944760
CALC                          -0.055562075 -0.134771048 -0.11203443  1.00000000  0.087169353  0.08440824
MTRANS                         0.043157450  0.006393953  0.17694475  0.08716935  1.000000000 -0.04620226
NObeyesdad                    -0.050678789 -0.129564308 -0.06944760  0.08440824 -0.046202261  1.00000000
```




```
> obesity_subset <- obesity2.norm[, c("Height", "Weight", "CAEC", "family_history_with_overweight", "FAF", "FAVC", "Gender",
"NObeyesdad")]
> obesity_subset.norm.rows=nrow(obesity_subset)
```

Rationale for picking the variables:

We chose the attributes Height, Weight, CAEC, family_history_with_overweight, FAF, FAVC, Gender, and the target variable NObeyesdad, this is based on the correlation matrix values that suggest these are the set of attributes that are closely related to each other, and their potential influence on the obesity levels. The chosen features capture key aspects of physical characteristics, lifestyle habits, etc. and provide a well-rounded view of obesity-related factors.

c. Translating alphanumeric values into numeric values (executed before subsetting)

```
> obesity2$Gender<-as.numeric(factor(obesity2$Gender))
> obesity2$family_history_with_overweight<-as.numeric(factor(obesity2$family_history_with_overweight))
> obesity2$FAVC<-as.numeric(factor(obesity2$FAVC))
> obesity2$CAEC<-as.numeric(factor(obesity2$CAEC))
> obesity2$SMOKE<-as.numeric(factor(obesity2$SMOKE))
> obesity2$SCC<-as.numeric(factor(obesity2$SCC))
> obesity2$CALC<-as.numeric(factor(obesity2$CALC))
> obesity2$MTRANS<-as.numeric(factor(obesity2$MTRANS))
> obesity2$NObeyesdad<-as.numeric(factor(obesity2$NObeyesdad))
```

```
> str(obesity2)
'data.frame':   2111 obs. of  17 variables:
 $ Gender                         : num  1 1 2 2 2 2 1 2 2 2 ...
 $ 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: num  2 2 2 1 1 1 2 1 2 2 ...
 $ FAVC                           : num  1 1 1 1 1 2 2 1 2 2 ...
 $ 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                           : num  4 4 4 4 4 4 4 4 4 4 ...
 $ SMOKE                          : num  1 2 1 1 1 1 1 1 1 1 ...
 $ CH2O                           : num  2 3 2 2 2 2 2 2 2 2 ...
 $ SCC                            : num  1 2 1 1 1 1 1 1 1 1 ...
 $ 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                           : num  3 4 2 2 4 4 4 4 2 3 ...
 $ MTRANS                         : num  4 4 4 5 4 1 3 4 4 4 ...
 $ NObeyesdad                     : num  2 2 2 6 7 2 2 2 2 2 ...
```

```
> summary(obesity2)
     Gender           Age            Height          Weight       family_history_with_overweight      FAVC      
 Min.   :1.000   Min.   :14.00   Min.   :1.450   Min.   : 39.00   Min.   :1.000                  Min.   :1.000  
 1st Qu.:1.000   1st Qu.:19.95   1st Qu.:1.630   1st Qu.: 65.47   1st Qu.:2.000                  1st Qu.:2.000  
 Median :2.000   Median :22.78   Median :1.700   Median : 83.00   Median :2.000                  Median :2.000  
 Mean   :1.506   Mean   :24.31   Mean   :1.702   Mean   : 86.59   Mean   :1.818                  Mean   :1.884  
 3rd Qu.:2.000   3rd Qu.:26.00   3rd Qu.:1.768   3rd Qu.:107.43   3rd Qu.:2.000                  3rd Qu.:2.000  
 Max.   :2.000   Max.   :61.00   Max.   :1.980   Max.   :173.00   Max.   :2.000                  Max.   :2.000  
      FCVC            NCP             CAEC           SMOKE            CH2O            SCC             FAF        
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :0.0000  
 1st Qu.:2.000   1st Qu.:2.659   1st Qu.:4.000   1st Qu.:1.000   1st Qu.:1.585   1st Qu.:1.000   1st Qu.:0.1245  
 Median :2.386   Median :3.000   Median :4.000   Median :1.000   Median :2.000   Median :1.000   Median :1.0000  
 Mean   :2.419   Mean   :2.686   Mean   :3.671   Mean   :1.021   Mean   :2.008   Mean   :1.045   Mean   :1.0103  
 3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:4.000   3rd Qu.:1.000   3rd Qu.:2.477   3rd Qu.:1.000   3rd Qu.:1.6667  
 Max.   :3.000   Max.   :4.000   Max.   :4.000   Max.   :2.000   Max.   :3.000   Max.   :2.000   Max.   :3.0000  
      TUE             CALC          MTRANS        NObeyesdad   
 Min.   :0.0000   Min.   :1.00   Min.   :1.000   Min.   :1.000  
 1st Qu.:0.0000   1st Qu.:3.00   1st Qu.:4.000   1st Qu.:2.000  
 Median :0.6253   Median :4.00   Median :4.000   Median :4.000  
 Mean   :0.6579   Mean   :3.63   Mean   :3.365   Mean   :4.016  
 3rd Qu.:1.0000   3rd Qu.:4.00   3rd Qu.:4.000   3rd Qu.:6.000  
 Max.   :2.0000   Max.   :4.00   Max.   :5.000   Max.   :7.000  
```

Mapping table for all variables:

Variable	Categories & Mappings
Gender	Female = 1, Male = 2
family_history_with_overweight	No = 1, Yes = 2
FAVC (Frequent high-calorie food)	No = 1, Yes = 2
CAEC (Eating between meals)	No = 1, Sometimes = 2, Frequently = 3, Always = 4
SMOKE	No = 1, Yes = 2
SCC (Calorie monitoring)	No = 1, Yes = 2
CALC (Alcohol consumption)	No = 1, Sometimes = 2, Frequently = 3, Always = 4
MTRANS (Mode of transport)	Public Transport = 1, Walking = 2, Automobile = 3, Motorbike = 4, Bike = 5
NObeyesdad (Obesity level)	Insufficient_Weight = 1, Normal_Weight = 2, Overweight_Level_I = 3, Overweight_Level_II = 4, Obesity_Type_I = 5, Obesity_Type_II = 6, Obesity_Type_III = 7

We translate alphanumeric values into numeric values because categorical variables are harder to process and analyze by models.

d. Splitting the dataset into train and test.

Note: The below split is for 70%-30%, the same is done for both 60%-40% and 50%-50% splits, the results to which are included in 4. a. (viii) and 4. a. (ix).

```
> obesity_subset.sample=0.7
> obesity_subset.rows=obesity_subset.sample*obesity_subset.norm.rows
> obesity_subset.rows
[1] 1477.7
> obesity_subset.train.index=sample(obesity_subset.norm.rows,obesity_subset.rows)
> length(obesity_subset.train.index)
[1] 1477
```

```
> obesity_subset.train=obesity_subset[obesity_subset.train.index,]
> obesity_subset.train[1:10,]
        Height     Weight      CAEC family_history_with_overweight        FAF FAVC Gender NObeyesdad
603  0.2869472 0.04477612 0.3333333                                0 0.41950133    0      0  0.0000000
1587 0.6079057 0.56043027 1.0000000                                1 0.45366467    1      1  0.5000000
1062 0.5199755 0.32447921 1.0000000                                1 0.33333333    1      1  1.0000000
819  0.5667151 0.30469198 1.0000000                                1 0.43122167    1      1  0.8333333
2012 0.5553189 0.70486933 1.0000000                                1 0.41034367    1      0  0.6666667
1467 0.6955491 0.49624608 1.0000000                                1 0.02091667    1      1  0.3333333
910  0.4293830 0.26865672 1.0000000                                1 0.11152633    1      1  0.8333333
986  0.3773585 0.30597015 1.0000000                                1 0.15035933    1      1  1.0000000
1841 0.5722000 0.70785397 1.0000000                                1 0.48171600    1      0  0.6666667
680  0.4719057 0.08338377 0.3333333                                1 0.44054133    1      0  0.0000000
```

```
> obesity_subset.test=obesity_subset[-obesity_subset.train.index,]
> obesity_subset.test[1:10,]
       Height    Weight      CAEC family_history_with_overweight       FAF FAVC Gender NObeyesdad
2  0.13207547 0.1268657 1.0000000                                1 1.0000000    0      0  0.1666667
4  0.66037736 0.3582090 1.0000000                                0 0.6666667    0      1  0.8333333
6  0.32075472 0.1044776 1.0000000                                0 0.0000000    1      1  0.1666667
7  0.09433962 0.1194030 1.0000000                                1 0.3333333    1      0  0.1666667
8  0.35849057 0.1044776 1.0000000                                0 1.0000000    0      1  0.1666667
11 0.75471698 0.4925373 0.3333333                                1 0.6666667    1      1  0.3333333
12 0.50943396 0.3059701 0.3333333                                1 0.6666667    1      0  1.0000000
21 0.37735849 0.3059701 1.0000000                                1 1.0000000    0      1  1.0000000
23 0.37735849 0.1567164 1.0000000                                1 0.3333333    1      0  0.1666667
24 0.28301887 0.3208955 1.0000000                                1 0.0000000    1      0  0.3333333
```

Splitting the dataset into training and test sets enables us to accurately judge the performance of the fitted model.

e. Normalizing values across all attributes in dataset (executed before splitting into train and test sets)

```
> normalize<-function(x){((x-min(x))/(max(x)-min(x)))}
> normalize()
Error in normalize() : argument "x" is missing, with no default
> normalize
function(x){((x-min(x))/(max(x)-min(x)))}
> obesity2.norm<-as.data.frame(lapply(obesity2,normalize))
> obesity2.norm[1:10,]
   Gender       Age     Height    Weight family_history_with_overweight FAVC FCVC       NCP CAEC SMOKE CH2O SCC
1       0 0.1489362 0.32075472 0.1865672                                1    0  0.5 0.6666667    1     0  0.5   0
2       0 0.1489362 0.13207547 0.1268657                                1    0  1.0 0.6666667    1     1  1.0   1
3       1 0.1914894 0.66037736 0.2835821                                1    0  0.5 0.6666667    1     0  0.5   0
4       1 0.2765957 0.66037736 0.3582090                                0    0  1.0 0.6666667    1     0  0.5   0
5       1 0.1702128 0.62264151 0.3791045                                0    0  0.5 0.0000000    1     0  0.5   0
6       1 0.3191489 0.32075472 0.1044776                                0    1  0.5 0.6666667    1     0  0.5   0
7       0 0.1914894 0.09433962 0.1194030                                1    1  1.0 0.6666667    1     0  0.5   0
8       1 0.1702128 0.35849057 0.1044776                                0    0  0.5 0.6666667    1     0  0.5   0
9       1 0.2127660 0.62264151 0.1865672                                1    1  1.0 0.6666667    1     0  0.5   0
10      1 0.1702128 0.50943396 0.2164179                                1    1  0.5 0.6666667    1     0  0.5   0
         FAF TUE      CALC MTRANS NObeyesdad
1  0.0000000 0.5 0.6666667   0.75  0.1666667
2  1.0000000 0.0 1.0000000   0.75  0.1666667
3  0.6666667 0.5 0.3333333   0.75  0.1666667
4  0.6666667 0.0 0.3333333   1.00  0.8333333
5  0.0000000 0.0 1.0000000   0.75  1.0000000
6  0.0000000 0.0 1.0000000   0.00  0.1666667
7  0.3333333 0.0 1.0000000   0.50  0.1666667
8  1.0000000 0.0 1.0000000   0.75  0.1666667
9  0.3333333 0.5 0.3333333   0.75  0.1666667
10 0.3333333 0.5 0.6666667   0.75  0.1666667
```

```
> obesity2.znorm<-as.data.frame(lapply(obesity2,scale))
> obesity2.znorm[1:10,]
       Gender         Age      Height      Weight family_history_with_overweight       FAVC       FCVC       NCP
1  -1.0116740 -0.52200070 -0.8753819 -0.86235386                     0.4721794 -2.7591155 -0.7848327  0.404057
2  -1.0116740 -0.52200070 -1.9471379 -1.16780030                     0.4721794 -2.7591155  1.0880839  0.404057
3   0.9879925 -0.20683997  1.0537789 -0.36600341                     0.4721794 -2.7591155 -0.7848327  0.404057
4   0.9879925  0.42348149  1.0537789  0.01580463                    -2.1168357 -2.7591155  1.0880839  0.404057
5   0.9879925 -0.36442034  0.8394277  0.12271089                    -2.1168357 -2.7591155 -0.7848327 -2.166509
6   0.9879925  0.73864222 -0.8753819 -1.28234271                    -2.1168357  0.3622633 -0.7848327  0.404057
7  -1.0116740 -0.20683997 -2.1614891 -1.20598110                     0.4721794  0.3622633  1.0880839  0.404057
8   0.9879925 -0.36442034 -0.6610307 -1.28234271                    -2.1168357 -2.7591155 -0.7848327  0.404057
9   0.9879925 -0.04925961  0.8394277 -0.86235386                     0.4721794  0.3622633  1.0880839  0.404057
10  0.9879925 -0.36442034  0.1963741 -0.70963065                     0.4721794  0.3622633 -0.7848327  0.404057
        CAEC      SMOKE        CH2O        SCC        FAF        TUE       CALC     MTRANS NObeyesdad
1  0.4241109 -0.1458657 -0.01307017 -0.2182203 -1.1877577  0.5618636 -1.1446172  0.5032175  -1.032551
2  0.4241109  6.8523733  1.61837509  4.5803537  2.3391959 -1.0803687  0.6735069  0.5032175  -1.032551
3  0.4241109 -0.1458657 -0.01307017 -0.2182203  1.1635447  0.5618636 -2.9627414  0.5032175  -1.032551
4  0.4241109 -0.1458657 -0.01307017 -0.2182203  1.1635447 -1.0803687 -2.9627414  1.2959728   1.016535
5  0.4241109 -0.1458657 -0.01307017 -0.2182203 -1.1877577 -1.0803687  0.6735069  0.5032175   1.528806
6  0.4241109 -0.1458657 -0.01307017 -0.2182203 -1.1877577 -1.0803687  0.6735069 -1.8750485  -1.032551
7  0.4241109 -0.1458657 -0.01307017 -0.2182203 -0.0121065 -1.0803687  0.6735069 -0.2895378  -1.032551
8  0.4241109 -0.1458657 -0.01307017 -0.2182203  2.3391959 -1.0803687  0.6735069  0.5032175  -1.032551
9  0.4241109 -0.1458657 -0.01307017 -0.2182203 -0.0121065  0.5618636 -2.9627414  0.5032175  -1.032551
10 0.4241109 -0.1458657 -0.01307017 -0.2182203 -0.0121065  0.5618636 -1.1446172  0.5032175  -1.032551
```

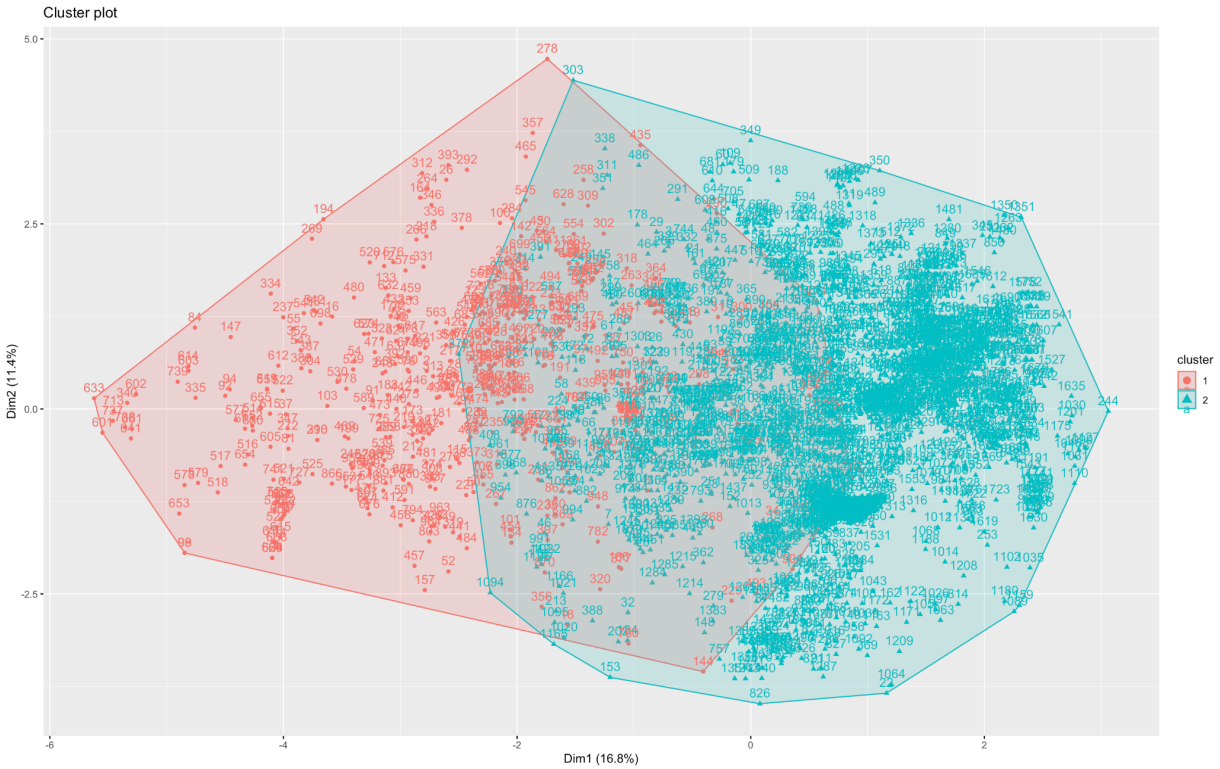
Normalizing values across all attributes in the dataset ensures that each feature contributes equally to the analysis. It helps avoid bias toward attributes with larger scales.

3. K-means clustering on the whole dataset

a. (i) k=2

```
> obesity2.k2<-kmeans(obesity2.norm,centers = 2)
> str(obesity2.k2)
List of 9
 $ cluster     : int [1:2111] 1 1 2 1 1 1 2 1 2 2 ...
 $ centers     : num [1:2, 1:17] 0.312 0.562 0.161 0.236 0.379 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:2] "1" "2"
  .. ..$ : chr [1:17] "Gender" "Age" "Height" "Weight" ...
 $ totss       : num 2879
 $ withinss    : num [1:2] 755 1727
 $ tot.withinss: num 2482
 $ betweenss   : num 397
 $ size        : int [1:2] 475 1636
 $ iter        : int 1
 $ ifault      : int 0
 - attr(*, "class")= chr "kmeans"
```

```
> obesity2.k2
K-means clustering with 2 clusters of sizes 475, 1636

Cluster means:
     Gender       Age    Height    Weight family_history_with_overweight      FAVC      FCVC       NCP      CAEC
1 0.3115789 0.1610592 0.3787399 0.1389783                     0.2084211 0.7263158 0.7049009 0.5444621 0.6540351
2 0.5623472 0.2363608 0.5027715 0.4178749                     0.9944988 0.9297066 0.7108631 0.5669320 0.9590465
       SMOKE      CH2O       SCC       FAF       TUE      CALC    MTRANS NObeyesdad
1 0.02315789 0.4132652 0.14105263 0.3799215 0.3271019 0.8554386 0.6815789  0.2375439
2 0.02017115 0.5303515 0.01772616 0.3242360 0.3294646 0.8826406 0.5650978  0.5795640

Clustering vector:
   [1] 1 1 2 1 1 1 2 1 2 2 2 2 1 1 2 1 2 1 2 2 2 2 2 2 2 1 2 1 2 1 1 2 1 1 2 2 2 1 2 2 1 2 2 2 1 1 2 1 1 1 1 1
  [57] 2 2 2 2 2 1 1 2 1 2 2 2 2 2 2 1 1 2 2 1 1 2 2 2 1 2 2 1 2 2 2 2 1 2 1 1 2 1 1 2 1 1 1 1 1 2 1 1 2 1 1 2
 [113] 1 1 1 1 2 2 2 2 2 2 1 2 1 2 1 1 2 2 1 1 1 1 2 2 2 2 1 1 2 1 1 1 2 1 1 2 1 2 2 2 2 2 2 2 1 1 2 2 2 2 2 1
 [169] 2 2 1 2 1 1 1 2 1 2 2 1 1 1 1 2 1 1 2 2 2 2 1 2 2 1 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 1 1 2 2 2 2 1 1 1 2 1
 [225] 2 2 1 1 1 2 2 2 1 2 1 1 1 2 1 1 1 1 1 2 1 1 1 1 2 2 2 2 2 2 1 1 2 2 1 1 2 1 1 1 2 1 1 1 2 1 1 2 1 2 1 1
 [281] 1 2 1 1 2 2 1 2 1 1 2 2 2 1 1 1 2 1 1 1 1 2 1 1 2 2 1 1 1 1 2 1 2 1 1 2 1 1 2 1 1 2 2 2 1 2 2 1 2 1 1 1
 [337] 1 2 1 1 1 1 2 1 2 1 1 2 2 2 2 1 1 2 1 1 1 2 2 1 1 2 2 1 2 1 2 1 2 1 2 1 1 2 1 1 2 2 1 1 1 2 1 2 1 2 1 1
 [393] 1 1 2 2 1 1 2 1 2 1 2 2 1 1 1 2 2 2 1 1 1 1 2 2 1 2 1 2 2 2 1 2 1 1 1 1 1 2 1 1 2 1 2 1 1 1 1 1 1 2 2 2
 [449] 2 2 1 2 1 1 1 1 1 1 1 2 1 2 2 2 1 1 2 2 1 1 1 2 1 1 1 1 2 1 1 1 2 1 1 2 2 1 2 2 1 1 1 1 1 1 1 2 1 2 2 2
 [505] 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2
 [561] 2 1 1 1 2 2 2 1 1 1 1 1 1 1 1 1 1 1 2 2 2 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 2 2 2 2 1 1
 [617] 1 2 1 1 1 1 1 1 1 1 2 1 2 1 1 1 1 2 1 1 1 2 2 2 1 1 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2
 [673] 2 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 1 1 1 2 2 2 1 2 1 2 1 1 1 1 1 1 1 1 1 1 2 2 2 1 1 1 1 1
 [729] 2 2 2 2 2 2 2 2 1 1 1 1 1 1 2 2 2 2 1 2 2 1 2 2 1 2 2 2 2 2 2 1 1 2 2 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2
 [785] 1 2 1 2 2 2 1 2 2 1 2 1 2 2 2 2 2 2 2 1 1 2 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 1 1 1
 [841] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 1 1 1 1 1
 [897] 1 2 2 2 2 2 2 1 1 2 2 2 2 2 2 2 2 2 2 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1
 [953] 1 2 1 2 2 2 2 1 2 2 2 1 1 2 2 1 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [ reached getOption("max.print") -- omitted 1111 entries ]

Within cluster sum of squares by cluster:
[1]  755.4651 1726.6334
 (between_SS / total_SS =  13.8 %)
```

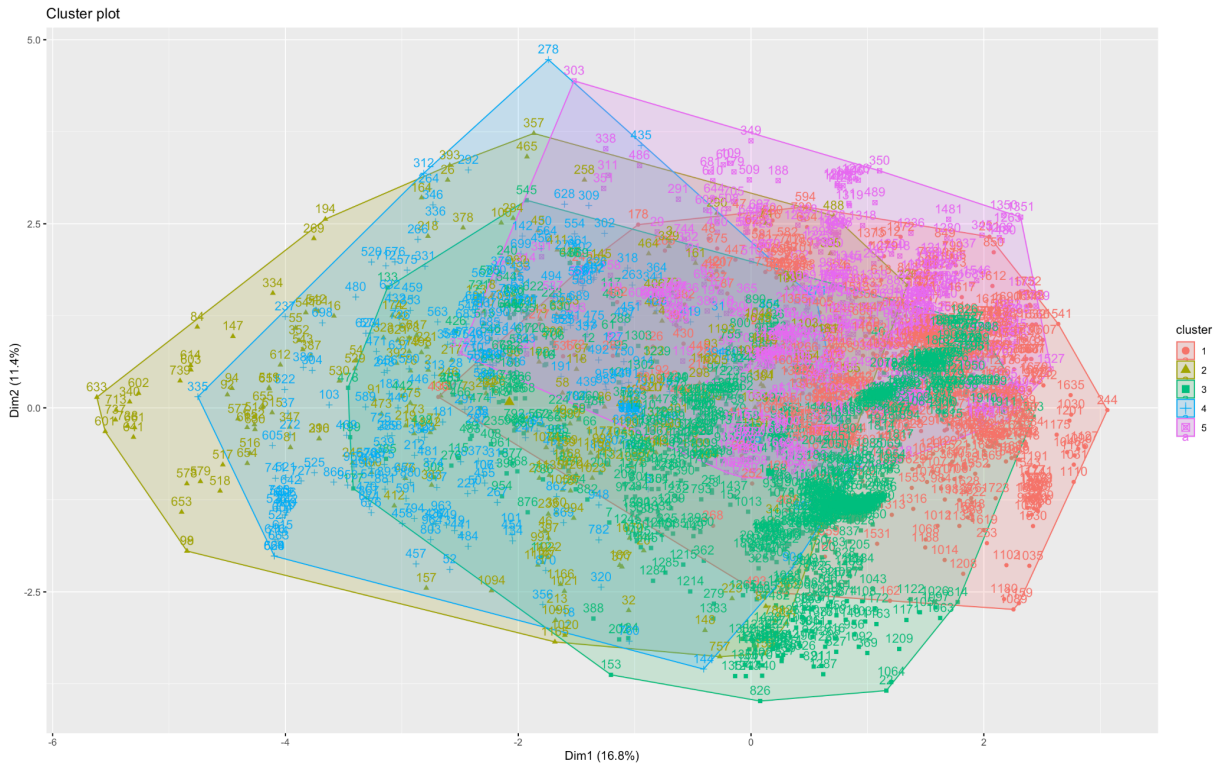


Only having 2 clusters had a lot of overlap between them, meaning there were no discernable characteristics for each cluster and therefore not a great model with a Between SS / Total SS of 13.8%.

(ii) k=5

```
> factoextra::fviz_cluster(obesity2.k2,obesity2.norm)
> obesity2.k5<-kmeans(obesity2.norm,centers = 5)
> str(obesity2.k5)
List of 9
 $ cluster     : int [1:2111] 2 2 2 2 2 4 3 2 5 5 ...
 $ centers     : num [1:5, 1:17] 1 0.373 0 0.361 1 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:5] "1" "2" "3" "4" ...
  .. ..$ : chr [1:17] "Gender" "Age" "Height" "Weight" ...
 $ totss       : num 2879
 $ withinss    : num [1:5] 373 355 526 314 249
 $ tot.withinss: num 1817
 $ betweenss   : num 1062
 $ size        : int [1:5] 511 228 730 266 376
 $ iter        : int 4
 $ ifault      : int 0
 - attr(*, "class")= chr "kmeans"
```

```
> obesity2.k5
K-means clustering with 5 clusters of sizes 511, 228, 730, 266, 376

Cluster means:
     Gender       Age    Height    Weight family_history_with_overweight      FAVC      FCVC       NCP      CAEC
1 1.0000000 0.2802180 0.6157442 0.4365659                     0.9647750 0.9804305 0.6625629 0.6032959 0.9523810
2 0.3728070 0.1867655 0.3753988 0.1926634                     0.5657895 0.0000000 0.7442253 0.5652930 0.7383041
3 0.0000000 0.2334113 0.3980729 0.4100443                     1.0000000 1.0000000 0.8044800 0.5571337 0.9191781
4 0.3609023 0.1435432 0.3605885 0.1379134                     0.0000000 1.0000000 0.6423418 0.4933816 0.7568922
5 1.0000000 0.1830922 0.5736429 0.3899686                     0.9946809 0.9813830 0.6154619 0.5611759 0.9370567
       SMOKE      CH2O         SCC       FAF       TUE      CALC    MTRANS NObeyesdad
1 0.02152642 0.5172355 0.023483366 0.4067609 0.1276076 0.8825832 0.3909002  0.5469667
2 0.03508772 0.4857130 0.166666667 0.4149692 0.2609979 0.8143275 0.6743421  0.4532164
3 0.01369863 0.5210013 0.026027397 0.2541426 0.3186629 0.8885845 0.6181507  0.5726027
4 0.01879699 0.3842688 0.093984962 0.3406034 0.3525457 0.9235589 0.6832707  0.2725564
5 0.02659574 0.5488291 0.005319149 0.3519160 0.6469719 0.8492908 0.6961436  0.4991135

Clustering vector:
   [1] 2 2 2 2 2 4 3 2 5 5 5 3 2 1 5 2 1 4 3 2 5 3 3 3 5 2 5 4 5 4 4 2 2 2 4 2 4 4 5 3 3 2 3 5 2 2 1 1 2 4 3 4 4 2 2 2
  [57] 1 2 3 5 3 4 4 5 3 3 5 1 1 2 3 2 2 2 5 2 2 3 3 2 2 1 5 2 5 2 1 1 3 3 2 2 2 2 4 2 2 2 2 2 4 2 4 3 3 4 4 1 5 3 5 3
 [113] 3 2 3 3 3 3 2 5 2 3 5 2 1 4 1 4 4 1 5 3 2 3 4 1 1 5 2 2 4 5 4 4 4 2 4 2 2 4 3 5 3 3 1 1 3 2 5 2 1 2 1 3 2 2 1 2 5
 [169] 3 3 4 1 2 2 4 3 3 1 5 4 4 3 4 3 4 2 1 5 1 3 4 1 2 2 3 5 5 1 3 3 2 3 3 1 5 3 1 2 3 3 5 4 2 1 1 5 2 2 2 3 2 5 3 3
 [225] 3 2 4 4 2 5 1 3 2 2 3 2 4 3 4 3 4 1 3 1 2 2 4 4 1 1 3 1 1 3 2 1 2 2 2 3 4 2 4 4 1 4 4 1 2 5 4 4 4 1 2 3 1 4 3 2
 [281] 4 1 3 2 5 1 4 3 4 2 5 4 5 2 5 2 4 2 5 5 4 4 5 4 2 5 3 3 4 2 5 4 4 5 4 4 2 4 4 4 5 2 2 1 3 4 1 2 2 3 4 5 1 2 4 4
 [337] 4 5 4 2 4 2 1 2 5 4 2 5 5 5 5 2 4 3 4 4 2 3 1 2 4 3 2 4 5 4 1 4 3 2 4 5 4 4 1 1 2 2 4 5 1 1 2 2 3 4 2 3 2 3 5 2
 [393] 2 4 3 3 2 4 1 2 2 4 5 2 4 2 2 3 2 3 2 2 4 2 5 5 1 5 4 1 1 1 4 5 2 3 2 4 4 1 4 4 4 3 4 3 4 3 4 4 4 3 5 5 1 4 1 5
 [449] 3 3 4 2 2 4 4 4 4 4 4 5 3 5 1 2 2 3 1 1 4 4 4 3 2 3 4 2 3 3 4 4 4 1 3 4 3 5 4 2 5 4 4 4 1 4 4 1 5 2 3 3 3 3 3 3
 [505] 3 3 3 5 5 5 3 3 3 2 2 2 2 2 2 4 4 4 4 4 4 4 4 4 4 2 2 2 1 1 1 5 1 5 4 4 4 2 2 2 3 3 3 4 4 4 3 3 3 4 4 4 4 4 4 1 5
 [561] 5 4 4 4 4 1 1 5 3 3 3 4 4 4 4 4 4 4 2 2 2 1 1 1 3 3 3 4 4 4 4 4 4 4 1 1 1 1 1 1 5 5 5 2 2 2 4 4 4 5 5 5 5 5 3 2 2 2 4 4
 [617] 2 1 3 4 2 3 4 3 4 4 1 4 1 3 4 4 2 1 3 4 4 1 1 5 2 4 5 5 5 3 3 3 3 2 2 2 2 2 2 4 4 4 4 4 2 4 4 4 3 3 4 1 1 5 5 3
 [673] 5 4 4 4 4 2 2 3 5 3 4 4 4 3 3 3 4 4 4 4 5 5 1 5 5 4 4 4 1 1 5 3 5 3 4 4 4 4 4 4 2 4 4 1 1 1 3 3 3 4 4 4 4 4 4 5
 [729] 1 1 1 1 1 5 5 5 2 2 2 4 4 4 5 5 5 1 4 1 1 4 1 4 5 1 3 5 2 3 4 4 3 3 4 4 4 3 1 1 1 1 1 3 1 3 3 1 1 3 5 5 1 4 3 5
 [785] 4 3 4 2 1 3 4 3 3 4 3 1 3 1 1 3 3 3 4 1 1 4 4 1 1 3 3 1 1 3 5 5 5 5 5 1 3 3 3 3 4 3 3 3 3 3 4 4 4 4 4 3 3 1 1 1
 [841] 1 5 1 5 3 3 3 3 3 1 1 1 3 3 3 1 1 3 5 1 5 1 4 3 3 5 4 4 3 4 4 2 1 1 3 3 3 3 3 3 4 4 3 1 3 3 3 5 1 3 3 3 3 4 1 1 4
 [897] 4 1 1 1 3 1 4 4 5 5 1 3 5 5 3 3 3 4 4 3 3 3 3 3 4 4 4 4 3 3 3 1 1 1 3 3 5 5 1 3 3 3 3 5 1 1 3 5 5 5 1 4 3 5 5 4
 [953] 4 3 4 3 1 3 4 3 3 1 4 4 3 3 1 3 3 1 1 1 3 3 3 3 4 5 1 5 5 1 1 1 5 5 2 1 1 2 2 5 2 2 5 5 3 3 5 5
 [ reached getOption("max.print") -- omitted 1111 entries ]

Within cluster sum of squares by cluster:
[1] 373.2014 354.6265 525.9417 313.9613 248.7872
 (between_SS / total_SS =  36.9 %)
```



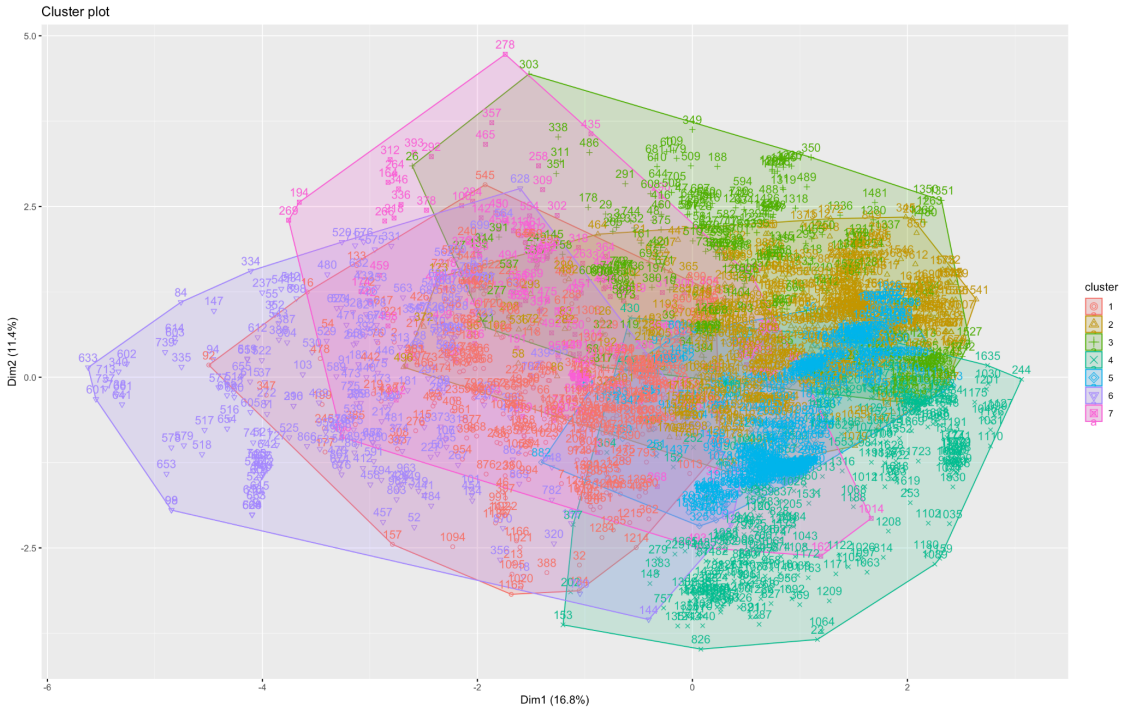
The 5 clusters model is much better than the 2 clusters model considering the within sum of squares improving to 36.9% and clusters are slightly more interpretable.

(iii) k=7

```
> obesity2.k7<-kmeans(obesity2.norm,centers = 7)
> str(obesity2.k7)
List of 9
 $ cluster     : int [1:2111] 1 1 3 7 7 7 1 7 3 3 ...
 $ centers     : num [1:7, 1:17] 0 1 1 0.523 0 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:7] "1" "2" "3" "4" ...
  .. ..$ : chr [1:17] "Gender" "Age" "Height" "Weight" ...
 $ totss       : num 2879
 $ withinss    : num [1:7] 330 308 222 239 102 ...
 $ tot.withinss: num 1641
 $ betweenss   : num 1238
 $ size        : int [1:7] 304 499 266 287 372 230 153
 $ iter        : int 4
 $ ifault      : int 0
 - attr(*, "class")= chr "kmeans"
```

```
> obesity2.k7
K-means clustering with 7 clusters of sizes 304, 499, 266, 287, 372, 230, 153

Cluster means:
     Gender       Age    Height     Weight family_history_with_overweight      FAVC      FCVC       NCP      CAEC
1 0.0000000 0.1664660 0.3371317 0.22207926                     1.0000000 0.7697368 0.6381700 0.4566319 0.7850877
2 1.0000000 0.2023717 0.5907137 0.43685954                     1.0000000 0.9238477 0.6363052 0.5632743 0.9478958
3 1.0000000 0.1955540 0.6169636 0.34038472                     1.0000000 0.9285714 0.6796041 0.6483070 0.9072682
4 0.5226481 0.4395931 0.4495646 0.36996784                     0.9930314 0.9372822 0.6416139 0.5290587 0.9709640
5 0.0000000 0.1955030 0.4454441 0.56581372                     1.0000000 0.9946237 0.9647537 0.6509618 0.9973118
6 0.0000000 0.1418697 0.2965644 0.09148167                     0.0000000 0.7347826 0.7285077 0.5100935 0.7144928
7 1.0000000 0.1834147 0.5106476 0.23468220                     0.0000000 0.7581699 0.6203726 0.5389641 0.7363834
       SMOKE      CH2O         SCC       FAF       TUE      CALC       MTRANS NObeyesdad
1 0.032894737 0.4653628 0.095394737 0.3357659 0.4231293 0.7598684 0.7368421053  0.4309211
2 0.028056112 0.4781513 0.012024048 0.3519675 0.2677982 0.8937876 0.7555110220  0.6315965
3 0.026315789 0.6768525 0.033834586 0.4901027 0.6240195 0.8258145 0.4285714286  0.2500000
4 0.020905923 0.4311353 0.006968641 0.2501629 0.1237854 0.8710801 0.0008710801  0.6463415
5 0.002688172 0.6041779 0.002688172 0.2400367 0.2951235 0.9829749 0.7506720430  0.6836918
6 0.004347826 0.3157507 0.173913043 0.3316943 0.3499521 0.9043478 0.6923913043  0.1746377
7 0.032679739 0.5407375 0.058823529 0.4278483 0.2635578 0.8496732 0.6176470588  0.4466231

Clustering vector:
   [1] 1 1 3 7 7 7 1 7 3 3 3 1 7 7 2 1 2 6 4 1 2 4 4 1 3 3 3 6 3 7 7 1 6 7 7 1 6 6 3 5 5 7 1 3 7 1 3 3 1 6 5 6 6 1 6 7
  [57] 4 2 1 3 1 7 7 3 1 1 2 3 4 3 1 1 1 3 2 1 1 5 1 2 6 2 2 6 2 2 2 2 1 4 6 1 2 6 6 2 1 6 6 7 6 2 6 1 4 6 6 4 3 1 3 1
 [113] 1 1 1 1 1 1 3 1 1 2 2 2 7 2 7 6 2 7 1 7 1 6 4 4 3 2 3 7 3 7 7 6 3 7 6 4 6 4 3 4 4 4 2 1 1 3 4 4 3 7 1 7 3 2 1 2
 [169] 1 4 7 2 6 7 7 1 1 3 3 7 6 1 6 1 7 6 4 3 3 1 7 2 3 7 1 2 3 4 1 1 1 4 5 4 3 1 2 2 1 1 2 6 1 2 2 3 1 7 1 5 1 2 1 1
 [225] 1 2 6 6 4 3 2 1 1 4 1 1 6 1 6 1 6 7 1 4 1 6 6 6 3 2 5 4 4 1 1 2 2 7 7 1 7 1 7 7 3 7 6 7 7 2 7 6 7 2 6 1 3 7 4 7
 [281] 7 2 1 7 2 2 6 1 6 7 3 7 2 2 3 6 6 7 2 7 7 7 3 6 2 2 1 1 7 6 3 7 6 3 6 7 3 7 6 6 3 3 6 3 5 6 2 2 3 5 6 3 2 6 6 7
 [337] 7 3 6 6 7 7 2 7 2 7 1 2 3 3 3 6 6 4 6 6 7 1 4 1 7 1 4 7 2 7 4 6 4 7 7 2 6 6 3 4 4 7 7 3 7 2 1 3 1 6 6 1 2 1 3 6
 [393] 7 7 1 1 1 7 2 7 2 7 2 4 7 7 7 1 2 1 7 6 7 7 3 3 7 2 7 2 3 3 7 2 7 1 7 6 7 4 7 6 7 1 7 1 6 1 6 6 6 1 2 3 2 6 2 2
 [449] 1 1 7 2 7 6 6 7 6 6 7 3 1 3 4 2 7 1 4 4 6 6 6 1 6 1 7 6 5 1 7 6 6 2 1 6 4 3 7 3 3 6 7 7 7 7 6 2 3 7 5 5 5 5 5 5
 [505] 5 5 5 3 3 3 3 1 1 1 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 3 3 3 3 2 3 6 6 6 6 6 6 1 1 1 6 6 6 1 1 1 7 7 7 7 7 7 3 3
 [561] 3 6 6 6 3 3 3 3 1 1 1 6 6 6 6 6 6 6 6 6 3 3 3 1 1 1 6 6 6 6 6 6 3 3 3 3 3 3 3 3 3 3 6 6 6 6 6 6 3 3 3 3 1 6 6 6 6 6
 [617] 1 3 1 6 6 1 6 1 7 7 3 6 3 1 6 6 6 3 1 6 6 3 3 3 6 6 3 3 3 3 1 1 1 1 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 1 1 6 3 3 3 3 1
 [673] 3 6 6 6 6 6 6 1 3 1 6 6 6 1 1 1 7 6 7 7 3 3 3 3 3 6 6 6 3 3 3 1 3 1 6 6 6 6 6 6 6 6 6 6 3 3 3 1 1 1 6 6 6 6 6 6 3
 [729] 3 3 3 3 3 3 3 3 6 6 6 6 6 6 3 3 3 7 7 7 2 7 2 4 2 2 1 2 4 1 7 6 4 5 7 7 7 4 4 2 2 2 2 1 2 5 5 2 2 4 2 3 2 6 4 2
 [785] 6 1 6 4 4 4 7 1 1 6 1 7 4 2 2 1 1 5 6 7 7 7 7 7 2 1 1 2 2 4 2 2 2 2 2 2 4 4 5 1 7 4 4 4 1 5 7 7 7 7 7 4 4 2 2 2
 [841] 2 2 2 2 5 1 5 5 2 2 2 5 5 1 2 2 4 2 2 3 2 6 4 4 2 6 6 1 6 6 4 4 4 4 1 1 1 1 5 6 6 5 7 4 1 1 2 2 1 1 1 5 6 7 7 7
 [897] 7 4 2 2 1 2 6 6 2 2 2 1 2 2 4 5 1 7 7 4 4 4 5 1 7 7 7 7 4 4 5 2 2 2 5 5 3 2 2 5 5 5 5 2 2 2 4 2 2 2 2 6 4 3 2 6
 [953] 6 1 6 4 4 4 7 1 1 2 6 6 1 1 7 4 4 2 2 2 1 1 5 5 6 2 2 2 2 2 4 4 2 2 1 4 4 1 1 2 1 1 2 2 1 1 2 2
 [ reached getOption("max.print") -- omitted 1111 entries ]
```

```
Within cluster sum of squares by cluster:
[1] 329.8939 308.3783 222.3904 238.5390 101.9429 264.1608 175.9317
 (between_SS / total_SS =  43.0 %)
```



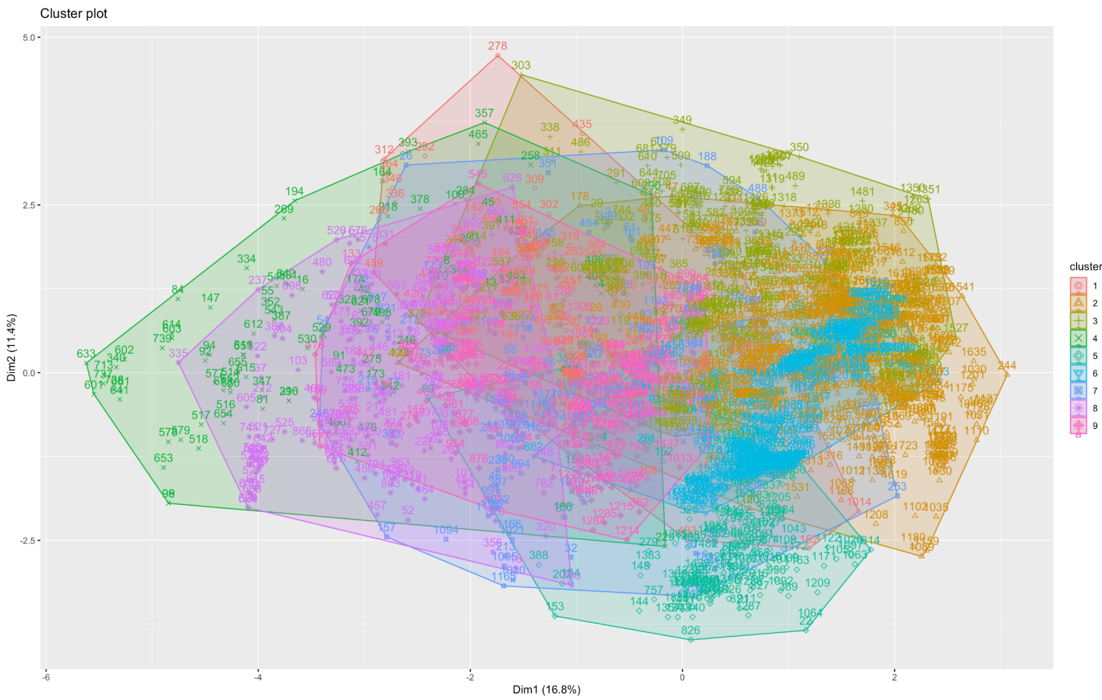
SSB/SST improves to 43% with 7 clusters which is trending in the right direction.

(iv) k=9

```
> obesity2.k9<-kmeans(obesity2.norm,centers = 9)
> str(obesity2.k9)
List of 9
 $ cluster     : int [1:2111] 7 7 7 4 1 1 9 4 3 3 ...
 $ centers     : num [1:9, 1:17] 1 1 1 0.33 0 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:9] "1" "2" "3" "4" ...
  .. ..$ : chr [1:17] "Gender" "Age" "Height" "Weight" ...
 $ totss       : num 2879
 $ withinss    : num [1:9] 117.3 344.5 207.2 119.2 81.3 ...
 $ tot.withinss: num 1492
 $ betweenss   : num 1387
 $ size        : int [1:9] 121 556 294 97 138 363 141 166 235
 $ iter        : int 4
 $ ifault      : int 0
 - attr(*, "class")= chr "kmeans"
```

```
> obesity2.k9
K-means clustering with 9 clusters of sizes 121, 556, 294, 97, 138, 363, 141, 166, 235

Cluster means:
     Gender       Age    Height     Weight family_history_with_overweight      FAVC      FCVC       NCP      CAEC
1 1.0000000 0.1907746 0.5027041 0.23834060                     0.00000000 0.9586777 0.6026198 0.5024367 0.7382920
2 1.0000000 0.2633370 0.6136708 0.44609725                     1.00000000 1.0000000 0.6445670 0.5917089 0.9622302
3 1.0000000 0.1787153 0.5707327 0.36688742                     1.00000000 1.0000000 0.6371969 0.5693344 0.9126984
4 0.3298969 0.1538250 0.3883321 0.13378155                     0.03092784 0.0000000 0.8074626 0.5953887 0.5567010
5 0.0000000 0.4744008 0.3221575 0.27451527                     0.97101449 0.9855072 0.6850593 0.4986625 0.9613527
6 0.0000000 0.1970158 0.4465228 0.57284103                     1.00000000 1.0000000 0.9653231 0.6508804 0.9963269
7 0.4609929 0.2175154 0.3823629 0.25410522                     1.00000000 0.0000000 0.6768261 0.5608895 0.8817967
8 0.0000000 0.1336092 0.2873316 0.09003432                     0.00000000 1.0000000 0.6783314 0.4968912 0.7931727
9 0.0000000 0.1548684 0.3693111 0.23552360                     1.00000000 1.0000000 0.6291830 0.4448676 0.7773050
       SMOKE      CH2O         SCC       FAF       TUE      CALC     MTRANS NObeyesdad
1 0.041322314 0.5308499 0.024793388 0.3922380 0.2506144 0.8650138 0.60537190  0.4903581
2 0.017985612 0.5099396 0.016187050 0.3835214 0.1599218 0.9136691 0.48785971  0.5887290
3 0.030612245 0.5688456 0.010204082 0.3694096 0.7077741 0.8083900 0.61904762  0.3832200
4 0.010309278 0.4672809 0.247422680 0.4406770 0.2340437 0.8144330 0.70876289  0.1701031
5 0.028985507 0.3519001 0.007246377 0.1968534 0.1462798 0.8647343 0.01449275  0.6461353
6 0.002754821 0.6072876 0.002754821 0.2412966 0.2918011 0.9871442 0.75000000  0.6864096
7 0.063829787 0.5291105 0.099290780 0.4200986 0.3035836 0.7966903 0.62234043  0.6453901
8 0.000000000 0.2864762 0.144578313 0.3125483 0.4109815 0.9277108 0.68674699  0.1907631
9 0.021276596 0.4785647 0.072340426 0.3105900 0.4562136 0.7531915 0.75319149  0.3581560

Clustering vector:
   [1] 7 7 7 4 1 1 9 4 3 3 3 9 4 1 3 4 2 8 5 7 7 5 5 9 3 7 3 8 3 1 1 7 4 1 1 7 8 8 3 6 6 4 9 3 4 7 2 2 7 8 9 8 8 7 4 4
  [57] 3 7 9 3 9 1 1 3 9 9 3 3 2 7 9 7 7 7 3 7 7 6 9 7 4 2 3 4 2 7 2 2 5 5 4 4 7 4 8 7 7 4 4 4 8 7 8 9 5 8 8 2 7 9 3 9
 [113] 9 7 9 9 9 7 3 7 9 3 7 2 1 2 1 8 2 1 9 4 9 8 2 2 3 7 7 1 3 1 1 5 7 1 4 5 8 5 3 5 5 7 2 5 7 3 7 2 7 1 9 4 7 7 7 7
 [169] 9 5 1 2 4 4 1 9 9 2 3 1 8 9 8 5 1 4 7 7 3 9 1 2 7 4 9 3 7 2 9 9 7 5 6 7 3 9 2 7 9 9 3 8 7 2 2 3 7 4 7 6 7 2 9 9
 [225] 9 7 8 8 4 7 2 9 7 7 9 7 8 9 8 9 8 1 9 2 7 4 8 8 2 2 6 7 7 9 7 7 7 4 1 9 1 7 1 1 3 1 8 1 4 3 1 8 1 2 4 9 3 1 5 4
 [281] 1 2 9 4 3 2 8 9 8 4 3 1 3 7 3 4 8 1 3 1 1 1 3 8 7 3 9 9 1 4 3 1 8 3 8 1 7 1 8 8 3 7 4 2 5 8 2 7 7 9 8 3 2 4 8 1
 [337] 1 3 8 4 1 4 2 4 2 1 4 3 3 3 7 4 8 5 8 8 4 5 7 7 1 9 7 1 3 1 2 8 5 4 1 3 8 8 3 2 7 4 1 3 1 2 7 7 9 8 4 5 7 9 3 4
 [393] 4 1 9 9 7 1 2 4 7 1 2 7 1 4 4 9 7 9 4 4 1 4 3 3 1 3 1 2 2 2 1 2 1 9 4 8 1 2 1 8 1 9 1 9 8 9 8 8 8 9 3 3 2 8 2 2
 [449] 9 9 1 7 4 8 8 1 8 8 1 3 9 3 2 7 4 9 2 2 8 8 8 9 4 9 1 4 6 9 1 8 8 2 9 8 5 3 1 7 3 8 1 1 1 1 8 2 3 4 6 6 6 6 6 6
 [505] 6 6 6 3 3 3 9 9 9 4 4 4 4 4 4 8 8 8 8 8 8 8 8 8 8 8 4 4 4 2 3 2 3 3 3 8 8 8 4 4 4 9 9 9 8 8 8 8 9 9 9 1 1 1 1 1 1 2 3
 [561] 3 8 8 8 2 2 3 9 9 9 8 8 8 8 8 8 4 4 4 3 3 3 9 9 9 8 8 8 8 8 8 3 2 3 2 2 2 3 3 3 4 4 4 8 8 8 3 3 3 3 9 4 4 4 8 8
 [617] 7 2 9 8 4 9 8 9 1 1 3 8 3 9 8 8 4 2 9 8 8 2 2 3 4 8 3 3 3 9 9 9 9 4 4 4 4 4 4 4 8 8 8 8 8 4 8 8 8 9 9 8 2 2 3 3 9
 [673] 3 8 8 8 8 4 4 9 3 9 8 8 8 9 9 9 1 8 1 1 3 3 2 3 3 8 8 8 2 2 3 9 3 9 8 8 8 8 8 8 4 8 8 3 3 2 9 9 9 8 8 8 8 8 8 3
 [729] 2 3 2 2 2 3 3 3 4 4 4 8 8 8 3 3 3 1 1 1 2 1 2 5 2 2 9 3 5 9 1 8 5 6 1 1 1 5 2 2 2 2 2 9 2 6 6 2 2 5 2 3 2 8 5 3
 [785] 8 9 8 7 2 5 1 9 9 8 9 1 5 2 2 9 9 9 8 1 1 1 1 1 2 9 9 2 2 5 2 2 3 3 3 2 5 5 6 9 1 5 5 5 9 6 1 1 1 1 1 5 5 2 2 2
```

```
 [785] 8 9 8 7 2 5 1 9 9 8 9 1 5 2 2 9 9 9 8 1 1 1 1 1 2 9 9 2 2 5 2 2 3 3 3 2 5 5 6 9 1 5 5 5 9 6 1 1 1 1 1 5 5 2 2 2
 [841] 2 2 2 2 6 6 6 6 2 2 2 6 6 9 2 2 5 2 2 3 2 8 5 5 2 8 8 9 8 8 7 2 2 5 9 9 9 9 6 8 8 6 1 5 9 9 2 2 9 9 9 6 8 1 1 1
 [897] 1 2 2 2 9 2 5 5 2 2 2 9 3 3 5 6 9 1 1 5 5 5 6 9 1 1 1 1 5 5 6 2 2 2 6 6 3 3 2 6 6 6 6 2 2 2 5 2 2 3 2 8 5 3 3 8
 [953] 8 9 8 5 2 5 1 9 9 2 8 8 9 9 1 5 5 2 2 2 9 9 9 6 8 2 2 3 2 2 2 2 3 3 7 2 2 7 7 2 7 7 3 2 9 9 2 2
 [ reached getOption("max.print") -- omitted 1111 entries ]

Within cluster sum of squares by cluster:
[1] 117.26901 344.51334 207.24802 119.20761  81.32901  94.63646 178.86635 152.18349 196.47640
 (between_SS / total_SS =  48.2 %)
```

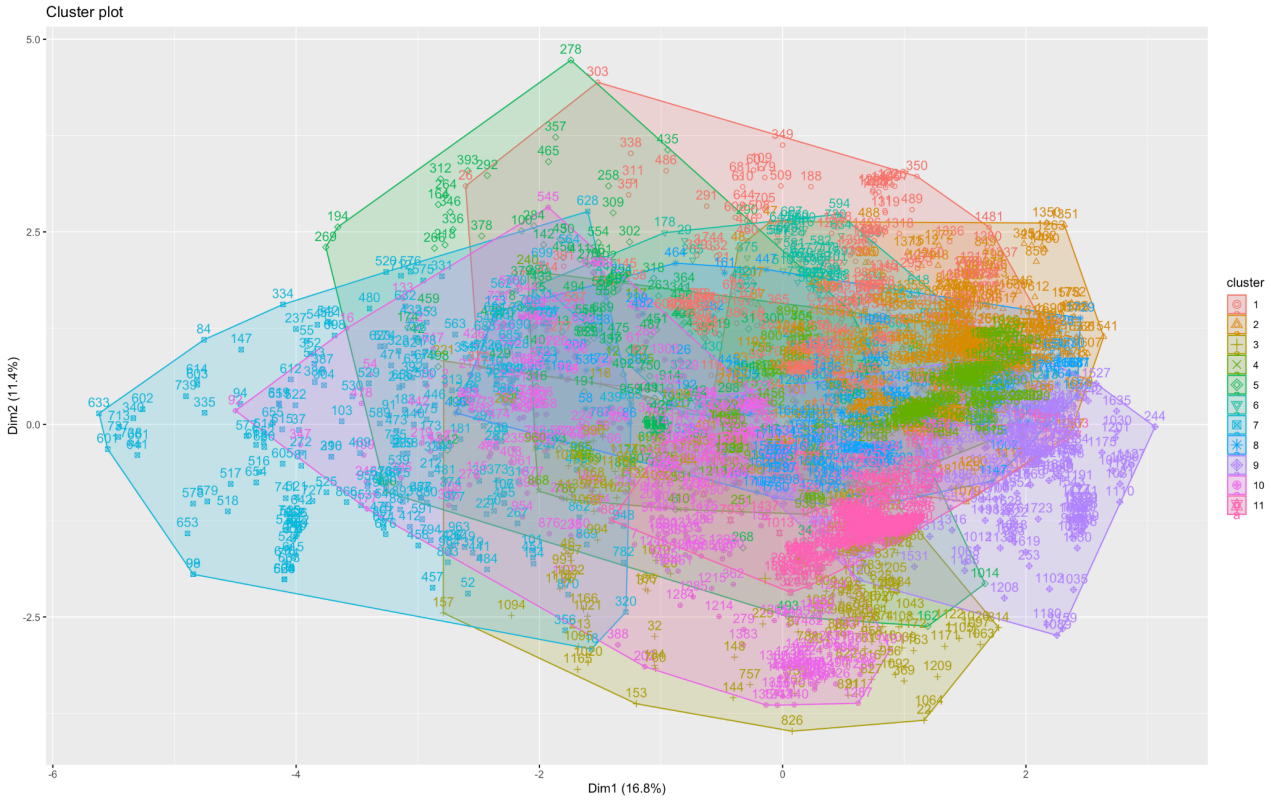


As the number of clusters increases, SSB/SST only improves a small percentage (improved to 48.3%).

(v) k=11

```
> obesity2.k11<-kmeans(obesity2.norm,centers = 11)
> str(obesity2.k11)
List of 9
 $ cluster     : int [1:2111] 10 10 1 5 5 5 10 5 1 1 ...
 $ centers     : num [1:11, 1:17] 1 1 0 0 1 1 0 1 1 0 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:11] "1" "2" "3" "4" ...
  .. ..$ : chr [1:17] "Gender" "Age" "Height" "Weight" ...
 $ totss       : num 2879
 $ withinss    : num [1:11] 127.1 124.7 133.4 76.9 175.9 ...
 $ tot.withinss: num 1459
 $ betweenss   : num 1420
 $ size        : int [1:11] 188 301 137 188 153 71 225 156 199 264 ...
 $ iter        : int 5
 $ ifault      : int 0
 - attr(*, "class")= chr "kmeans"
```

```
> obesity2.k11
K-means clustering with 11 clusters of sizes 188, 301, 137, 188, 153, 71, 225, 156, 199, 264, 229

Cluster means:
   Gender       Age    Height     Weight family_history_with_overweight      FAVC      FCVC       NCP      CAEC
1       1 0.1564257 0.5629504 0.32931202                     1.0000000 0.9414894 0.5604628 0.5859061 0.8882979
2       1 0.1898563 0.6308127 0.45750951                     1.0000000 0.9302326 0.5931666 0.6211043 0.9811739
3       0 0.3544557 0.3252074 0.25016153                     0.9489051 0.5620438 0.7462209 0.5174152 0.9513382
4       0 0.1408720 0.5293342 0.58232674                     1.0000000 0.9946809 0.9190202 0.5298517 0.9911348
5       1 0.1834147 0.5106476 0.23468220                     0.0000000 0.7581699 0.6203726 0.5389641 0.7363834
6       1 0.1693023 0.6799645 0.21433897                     1.0000000 0.9295775 0.8074473 0.7885515 0.8920188
7       0 0.1347122 0.2942873 0.08758581                     0.0000000 0.7333333 0.7263417 0.5105249 0.7081481
8       1 0.2324914 0.5501671 0.42929016                     1.0000000 0.8974359 0.7758902 0.4538917 0.8931624
9       1 0.3935969 0.5733580 0.47686300                     1.0000000 0.9396985 0.6193371 0.5639408 0.9765494
10      0 0.2412725 0.3234081 0.22644141                     1.0000000 0.9166667 0.6153638 0.4910742 0.7739899
11      0 0.2310334 0.3700126 0.49698956                     1.0000000 1.0000000 0.9452668 0.6660606 0.9927220
        SMOKE      CH2O         SCC        FAF       TUE      CALC      MTRANS NObeyesdad
1  0.021276596 0.6264660 0.010638298 0.40967929 0.7708163 0.8102837 0.757978723 0.38918440
2  0.003322259 0.6469994 0.003322259 0.39944738 0.2001074 0.9545958 0.753322259 0.66943522
3  0.036496350 0.3846700 0.014598540 0.25346415 0.2919451 0.8029197 0.259124088 0.90754258
4  0.005319149 0.6655701 0.042553191 0.51288587 0.4408067 0.9255319 0.751329787 0.67375887
5  0.032679739 0.5407375 0.058823529 0.42784834 0.2635578 0.8496732 0.617647059 0.44662309
6  0.000000000 0.5954794 0.126760563 0.70797555 0.3404257 0.7417840 0.049295775 0.08920188
7  0.004444444 0.3184578 0.177777778 0.33316454 0.3530066 0.9022222 0.707777778 0.16000000
8  0.083333333 0.1757091 0.025641026 0.28500467 0.2474903 0.8162393 0.754807692 0.49038462
9  0.030150754 0.5439005 0.005025126 0.29623349 0.2442192 0.8927973 0.003768844 0.60050251
10 0.026515152 0.4070137 0.071969697 0.28474580 0.3004471 0.8017677 0.585227273 0.26767677
11 0.004366812 0.5844686 0.004366812 0.05781286 0.2441764 0.9796215 0.751091703 0.68995633

Clustering vector:
   [1] 10 10  1  5  5  5 10  5  1  1  1  4  5  5  8 10  8  7  3  3  1  3 10 10  1  1  1  7  6  5  5  3  7  5  5  3  7
  [38]  7  1  4 11  5 10  1  5  3  2  2 10  7 11  7  7 10  7  5  9  8 10  1 10  5  5  1 10 10  1  9  9  1 10 10 10  6
  [75]  1 10 10 11 10  8  7  2  1  7  2  8  2  2 10  3  7 10  2  7  7  8 10  7  7  5  7  8  7 10 10  7  7  9  1  4  8
 [112] 10 10 10 10 10  4  3  1  3 10  1  8  8  5  8  5  7  8  5 10  5 10  7  9  9  1  9  6  5  1  5  5  3  1  5  7  3
 [149]  7 10  1 10  3  9  2 10  3  1  3  9  8  5 10  5  9  8  3  1  4  3  5  8  7  5  5 10 10  6  1  5  7 10  7  3  5
 [186]  3  9  1  6 10  5  8  6  5  4  2  1  9 10 10  4 10 11  9  6 10  2  8 10 10  1  7  3  8  2  1 10  5 10 11  3  8
 [223] 10 10 10  2  7  7  3  9  2 10 10  3 10 10  7 10  7  4  7  5 10  9 10  7  7  7  6  2  4  9  9 10 10  2  2  5  5
 [260] 10  5  3  5  5  6  5  7  5  5  1  5  7  5  8  7 10  1  5 10  5  5  8 10  5  1  8  7 10  7  5  1  5  8  8  1  7
 [297]  7  5  8  5  5  5  1  7  2  2 10 10  5  7  1  5  7  1  7  5  6  5  7  7  8  6  7  6 11  7  6  8  1 11  7  1  8
 [334]  7  7  5  5  1  7  7  5  5  8  5  2  5 10  2  1  1  1  7  7 10  7  7  5 10  9 10  5 10  3  5  1  5  9  7  3  5
 [371]  5  1  7  7  6  9  3  5  5  1  5  8 10  9 10  7  7 10  2 10  1  7  5  5 10 10  3  5  8  5  2  5  8  3  5  5  5
 [371]  5  1  7  7  6  9  3  5  5  1  5  8 10  9 10  7  7 10  2 10  1  7  5  5 10 10  3  5  8  5  2  5  8  3  5  5  5
 [408] 10  8  4  5  7  5  5  1  1  5  2  5  8  6  6  5  8  5 10  5  7  5  6  5  7  5 10  5  4  7 10  7  7  7 10  1  1
 [445]  8  7  8  2 10 10  5  8  5  7  7  5  7  7  5  1 10  1  9  8  5 10  6  6  7  7  7 10  7 10  5  7 11 10  5  7  7
 [482]  8 10  7  3  1  5  2  1  7  5  5  5  5  7  8  1  5 11  4 11  4 11  4 11 11  4  1  1  6 10 10 10  7  7  7  7  7
 [519]  7  7  7  7  7  7  7  7  7  7  7  7  6  6  6  1  8  1  7  7  7  7  7  7 10 10 10  7  7  7 10 10 10  5  5  5
 [556]  5  5  5  6  6  6  7  7  7  6  6  6 10 10 10  7  7  7  7  7  7  7  7  7  6  6  6 10 10 10  7  7  7  7  7  7  6
 [593]  6  6  6  6  6  1  1  1  7  7  7  7  7  7  1  1  1  1 10  7  7  7  7  7 10  6 10  7  7 10  7 10  5  5  6  7  6
 [630] 10  7  7  7  6 10  7  7  6  6  1  7  7  1  1  6 10 10 10 10  7  7  7  7  7  7  7  7  7  7  7  7  7  7 10 10
 [667]  7  6  6  6  1 10  1  7  7  7  7  7  7 10  1 10  7  7  7 10 10 10  5  7  5  5  1  1  6  6  6  7  7  7  6  6  6
 [704] 10  1 10  7  7  7  7  7  7  7  7  7  6  6  6 10 10 10  7  7  7  7  7  7  6  6  6  6  6  6  1  1  1  7  7  7  7
 [741]  7  7  1  1  1  5  5  5  2  5  8  3  2  2  4  1  3 11  5  3  3  4  5  5  5  3  9  2  2  2  2 10  2  4  4  2  2
 [778]  3  2  1  2  7  3  8  7  4  7  3  9  3  5 10 11  7  4  5  3  2  2  4 10  4  7  5  5  5  5  5  2 11 10  2  2  3
 [815]  8  8  1  1  1  8  3  3  4  4  5  3  3  3  4  4  5  5  5  5  5  3  3  2  2  2  2  2  2 11 11 11  4  2  2  2
 [852]  4  4  4  2  8  3  2  2  1  2  7  3  3  8  7  7  4  7  7  3  9  9  3  4 10 10 10 11  7  7 11  5  3  4  4  2  2
 [889]  4  4  4 11  7  5  5  5  5  9  2  2 10  2  3  3  2  2  2  4  1  1  3 11 10  5  5  3  3  3  4  4  5  5  5  5  3
 [926]  3  4  2  2  2 11 11  1  1  2 11  4  4  4  2  8  2  3  2  2  1  2  7  3  1  1  7  7 10  7  3  9  3  5  4 10  2
 [963]  7  7  4  4  5  3  3  2  2  2  4 10  4  4  7  2  2  1  2  2  9  9  1  1  3  9  9  3  3  2  3  3  2  2  4  3  2
[1000]  2
 [ reached getOption("max.print") -- omitted 1111 entries ]

Within cluster sum of squares by cluster:
 [1] 127.10684 124.67212 133.40818  76.94804 175.93174  45.67771 255.21749 108.52762 109.50133 252.91790  48.99289
 (between_SS / total_SS =  49.3 %)
```



As the number of clusters increases, clusters become more interpretable, leading to a higher BSS/SST ratio, which indicates better-defined groupings. However, the rate of decrease in the within-cluster sum of squares decreases as more clusters are added. This suggests that while additional clusters improve the model, the returns are diminishing.

(vi) Cluster centers

```
> obesity2.k5$centers
     Gender       Age    Height    Weight family_history_with_overweight      FAVC      FCVC       NCP      CAEC
1 1.0000000 0.2802180 0.6157442 0.4365659                      0.9647750 0.9804305 0.6625629 0.6032959 0.9523810
2 0.3728070 0.1867655 0.3753988 0.1926634                      0.5657895 0.0000000 0.7442253 0.5652930 0.7383041
3 0.0000000 0.2334113 0.3980729 0.4100443                      1.0000000 1.0000000 0.8044800 0.5571337 0.9191781
4 0.3609023 0.1435432 0.3605885 0.1379134                      0.0000000 1.0000000 0.6423418 0.4933816 0.7568922
5 1.0000000 0.1830922 0.5736429 0.3899686                      0.9946809 0.9813830 0.6154619 0.5611759 0.9370567
       SMOKE      CH2O         SCC       FAF       TUE      CALC    MTRANS NObeyesdad
1 0.02152642 0.5172355 0.023483366 0.4067609 0.1276076 0.8825832 0.3909002  0.5469667
2 0.03508772 0.4857130 0.166666667 0.4149692 0.2609979 0.8143275 0.6743421  0.4532164
3 0.01369863 0.5210013 0.026027397 0.2541426 0.3186629 0.8885845 0.6181507  0.5726027
4 0.01879699 0.3842688 0.093984962 0.3406034 0.3525457 0.9235589 0.6832707  0.2725564
5 0.02659574 0.5488291 0.005319149 0.3519160 0.6469719 0.8492908 0.6961436  0.4991135
> obesity2.k7$centers
     Gender       Age    Height     Weight family_history_with_overweight      FAVC      FCVC       NCP      CAEC
1 0.0000000 0.1664660 0.3371317 0.22207926                      1.0000000 0.7697368 0.6381700 0.4566319 0.7850877
2 1.0000000 0.2023717 0.5907137 0.43685954                      1.0000000 0.9238477 0.6363052 0.5632743 0.9478958
3 1.0000000 0.1955540 0.6169636 0.34038472                      1.0000000 0.9285714 0.6796041 0.6483070 0.9072682
4 0.5226481 0.4395931 0.4495646 0.36996784                      0.9930314 0.9372822 0.6416139 0.5290587 0.9709640
5 0.0000000 0.1955030 0.4454441 0.56581372                      1.0000000 0.9946237 0.9647537 0.6509618 0.9973118
6 0.0000000 0.1418697 0.2965644 0.09148167                      0.0000000 0.7347826 0.7285077 0.5100935 0.7144928
7 1.0000000 0.1834147 0.5106476 0.23468220                      0.0000000 0.7581699 0.6203726 0.5389641 0.7363834
       SMOKE      CH2O         SCC       FAF       TUE      CALC       MTRANS NObeyesdad
1 0.032894737 0.4653628 0.095394737 0.3357659 0.4231293 0.7598684 0.7368421053  0.4309211
2 0.028056112 0.4781513 0.012024048 0.3519675 0.2677982 0.8937876 0.7555110220  0.6315965
3 0.026315789 0.6768525 0.033834586 0.4901027 0.6240195 0.8258145 0.4285714286  0.2500000
4 0.020905923 0.4311353 0.006968641 0.2501629 0.1237854 0.8710801 0.0008710801  0.6463415
5 0.002688172 0.6041779 0.002688172 0.2400367 0.2951235 0.9829749 0.7506720430  0.6836918
6 0.004347826 0.3157507 0.173913043 0.3316943 0.3499521 0.9043478 0.6923913043  0.1746377
7 0.032679739 0.5407375 0.058823529 0.4278483 0.2635578 0.8496732 0.6176470588  0.4466231
```

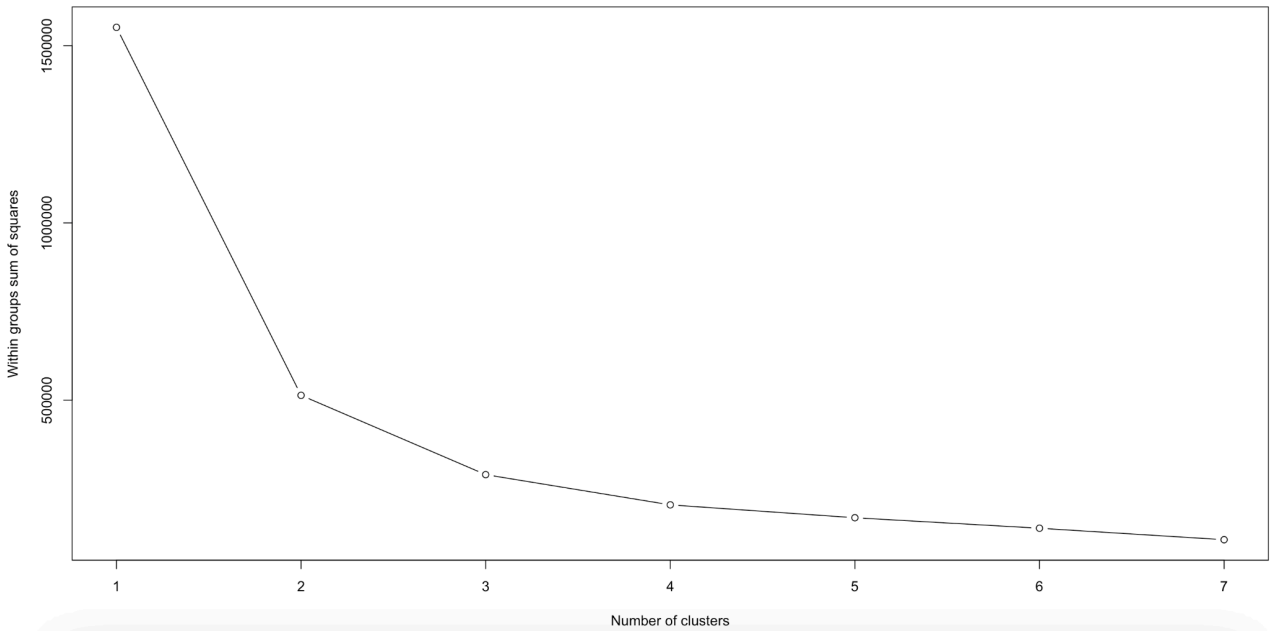
```
> obesity2.k9$centers
     Gender       Age    Height     Weight family_history_with_overweight      FAVC      FCVC       NCP      CAEC
1 1.0000000 0.1907746 0.5027041 0.23834060                     0.00000000 0.9586777 0.6026198 0.5024367 0.7382920
2 1.0000000 0.2633370 0.6136708 0.44609725                     1.00000000 1.0000000 0.6445670 0.5917089 0.9622302
3 1.0000000 0.1787153 0.5707327 0.36688742                     1.00000000 1.0000000 0.6371969 0.5693344 0.9126984
4 0.3298969 0.1538250 0.3883321 0.13378155                     0.03092784 0.0000000 0.8074626 0.5953887 0.5567010
5 0.0000000 0.4744008 0.3221575 0.27451527                     0.97101449 0.9855072 0.6850593 0.4986625 0.9613527
6 0.0000000 0.1970158 0.4465228 0.57284103                     1.00000000 1.0000000 0.9653231 0.6508804 0.9963269
7 0.4609929 0.2175154 0.3823629 0.25410522                     1.00000000 0.0000000 0.6768261 0.5608895 0.8817967
8 0.0000000 0.1336092 0.2873316 0.09003432                     0.00000000 1.0000000 0.6783314 0.4968912 0.7931727
9 0.0000000 0.1548684 0.3693111 0.23552360                     1.00000000 1.0000000 0.6291830 0.4448676 0.7773050
       SMOKE      CH2O         SCC       FAF       TUE      CALC     MTRANS NObeyesdad
1 0.041322314 0.5308499 0.024793388 0.3922380 0.2506144 0.8650138 0.60537190  0.4903581
2 0.017985612 0.5099396 0.016187050 0.3835214 0.1599218 0.9136691 0.48785971  0.5887290
3 0.030612245 0.5688456 0.010204082 0.3694096 0.7077741 0.8083900 0.61904762  0.3832200
4 0.010309278 0.4672809 0.247422680 0.4406770 0.2340437 0.8144330 0.70876289  0.1701031
5 0.028985507 0.3519001 0.007246377 0.1968534 0.1462798 0.8647343 0.01449275  0.6461353
6 0.002754821 0.6072876 0.002754821 0.2412966 0.2918011 0.9871442 0.75000000  0.6864096
7 0.063829787 0.5291105 0.099290780 0.4200986 0.3035836 0.7966903 0.62234043  0.6453901
8 0.000000000 0.2864762 0.144578313 0.3125483 0.4109815 0.9277108 0.68674699  0.1907631
9 0.021276596 0.4785647 0.072340426 0.3105900 0.4562136 0.7531915 0.75319149  0.3581560
> obesity2.k11$centers
   Gender       Age    Height     Weight family_history_with_overweight      FAVC      FCVC       NCP      CAEC
1       1 0.1564257 0.5629504 0.32931202                      1.0000000 0.9414894 0.5604628 0.5859061 0.8882979
2       1 0.1898563 0.6308127 0.45750951                      1.0000000 0.9302326 0.5931666 0.6211043 0.9811739
3       0 0.3544557 0.3252074 0.25016153                      0.9489051 0.5620438 0.7462209 0.5174152 0.9513382
4       0 0.1408720 0.5293342 0.58232674                      1.0000000 0.9946809 0.9190202 0.5298517 0.9911348
5       1 0.1834147 0.5106476 0.23468220                      0.0000000 0.7581699 0.6203726 0.5389641 0.7363834
6       1 0.1693023 0.6799645 0.21433897                      1.0000000 0.9295775 0.8074473 0.7885515 0.8920188
7       0 0.1347122 0.2942873 0.08758581                      0.0000000 0.7333333 0.7263417 0.5105249 0.7081481
8       1 0.2324914 0.5501671 0.42929016                      1.0000000 0.8974359 0.7758902 0.4538917 0.8931624
9       1 0.3935969 0.5733580 0.47686300                      1.0000000 0.9396985 0.6193371 0.5639408 0.9765494
10      0 0.2412725 0.3234081 0.22644141                      1.0000000 0.9166667 0.6153638 0.4910742 0.7739899
11      0 0.2310334 0.3700126 0.49698956                      1.0000000 1.0000000 0.9452668 0.6660606 0.9927220
        SMOKE      CH2O         SCC       FAF       TUE      CALC      MTRANS NObeyesdad
1  0.021276596 0.6264660 0.010638298 0.40967929 0.7708163 0.8102837 0.757978723 0.38918440
2  0.003322259 0.6469994 0.003322259 0.39944738 0.2001074 0.9545958 0.753322259 0.66943522
3  0.036496350 0.3846700 0.014598540 0.25346415 0.2919451 0.8029197 0.259124088 0.90754258
4  0.005319149 0.6655701 0.042553191 0.51288587 0.4408067 0.9255319 0.751329787 0.67375887
5  0.032679739 0.5407375 0.058823529 0.42784834 0.2635578 0.8496732 0.617647059 0.44662309
6  0.000000000 0.5954794 0.126760563 0.70797555 0.3404257 0.7417840 0.049295775 0.08920188
```

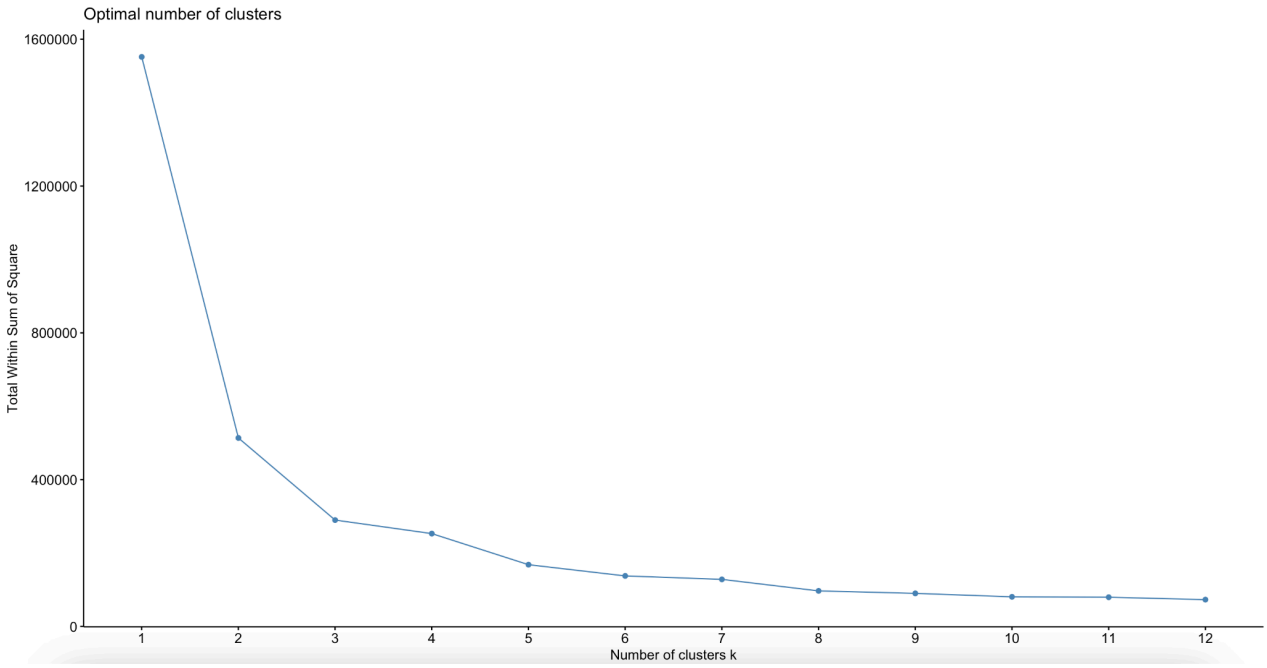
```
7  0.004444444 0.3184578 0.177777778 0.33316454 0.3530066 0.9022222 0.707777778 0.16000000
8  0.083333333 0.1757091 0.025641026 0.28500467 0.2474903 0.8162393 0.754807692 0.49038462
9  0.030150754 0.5439005 0.005025126 0.29623349 0.2442192 0.8927973 0.003768844 0.60050251
10 0.026515152 0.4070137 0.071969697 0.28474580 0.3004471 0.8017677 0.585227273 0.26767677
11 0.004366812 0.5844686 0.004366812 0.05781286 0.2441764 0.9796215 0.751091703 0.68995633
```

The cluster centers (centroids) give us a mean of all the data points in that cluster, summarizing the characteristics of each cluster.

(vii) Finding optimal number of clusters using function and factoextra::fviz_nbclust()

```
> optk<-function(data,nc=15,seed=1234)
+ {
+     opt<-(nrow(data)-1)*sum(apply(data,2,var))
+     for(i in 2:nc)
+     {
+         set.seed(seed)
+         opt[i]<-sum(kmeans(data,centers=i)$withinss)
+     }
+     plot(1:nc,opt,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
+ }
> optk(obesity2,nc=7,seed=12324)
```





b. Results table for clustering

No. Of Clusters	between_SS/total_SS (%)
2	13.8
5	36.9
7	43.0
9	48.2
11	49.3

As seen in the plot, the SSB / SST ratio improves as more clusters are added, but the rate at which they improve decreases. We came to a conclusion that 7 clusters (5 or 6 is a good number of clusters as well) would be optimal as moving to 9 or 11 doesn't improve the model too much and a simpler model usually accounts for more variation in the testing data.

4. Prediction

Note: The below results are for 70%-30% split, the results for the other splits are listed later in this section.

(i) Training labels for KNN using kmeans to compare with glm results

```
> obesity_subset.train.k7=kmeans(obesity_subset.train,centers = 7)
> obesity_subset.train.k7
K-means clustering with 7 clusters of sizes 90, 188, 65, 469, 54, 551, 60

Cluster means:
     Height    Weight      CAEC family_history_with_overweight       FAF      FAVC    Gender NObeyesdad
1 0.3881653 0.2829004 0.9666667                      0.9555556 0.3935491 0.0000000 0.4888889  0.7666667
2 0.3503844 0.1494438 0.8173759                      0.0000000 0.3132159 1.0000000 0.3776596  0.3546099
3 0.3601687 0.1185774 0.5487179                      0.0000000 0.4167030 0.0000000 0.2769231  0.1410256
4 0.4007573 0.4467286 0.9914712                      1.0000000 0.2405537 1.0000000 0.0000000  0.6265103
5 0.5924852 0.2372165 0.2222222                      0.6851852 0.4466942 0.9259259 1.0000000  0.2129630
6 0.5978028 0.4273893 0.9939504                      1.0000000 0.3726879 1.0000000 1.0000000  0.5251059
7 0.3921336 0.1159044 0.3277778                      1.0000000 0.4022895 0.8666667 0.0000000  0.1000000

Clustering vector:
 603 1587 1062  819 2012 1467  910  986 1841  680 1019 1412  972 1005  444  666 1980  673  162  938 1658  472 1526
   3    6    6    6    4    6    6    6    4    7    6    6    6    1    6    7    4    6    2    4    6    4    6
2093  469 1962  843 1438 1501 1468 1358  132 1785 1700  125  435 2091   65 2105 1733  333 1615  988  956  873 1684
   4    2    4    6    4    6    6    6    3    6    6    2    5    4    7    4    6    6    6    6    4    6    6
2035 1842   36  299  517 2052 1762  599 1482  567 1224 1618 1542  694 1193 1783  398  217 1784 1357  999  740 1289
   4    4    1    5    3    4    6    6    4    6    4    6    6    6    1    6    5    7    6    4    6    2    4
1750 1641 1954 1612  809 1161 2101 1106  415  905 1128  250 1753  124  490  482  994  442  165  362  706 1300  614

1748 1338  338  523 1382 2094 1165 1393 1731 1140 1983 1860  844 1974 1267 1498  580   83  448  319 1650 1058 1000
   6    4    5    2    4    4    1    4    6    6    4    4    6    4    4    4    6    5    6    2    6    6    6
1437 1465 1773  534  516  864 1657 2048  246  171 1064  244  546 1520 1027  642 1912  127 1433  402 1364 1016 1898
   4    6    6    6    3    4    6    4    3    2    4    6    7    1    4    2    4    2    4    5    6    6    4
 266 1926 1727  385 1325 1320  163   94 2003 1098 2004 1711  572  665 1293 1359 1154  930  805  758  750 1095 1874
   5    4    6    4    4    6    7    3    4    4    4    6    2    7    6    6    6    6    2    4    2    1    4
2086 1194  292 1504  915  609 1915  342 1039  267 1453 1533  637  536  801  397 1350  309  607 1094 1984  316  511
   4    1    5    4    2    6    4    3    4    2    6    6    2    5    4    1    6    5    6    1    4    5    7
 149 1113  340  261 1119  560  417 1578  271 1075  373  382  559  629  351 1814   53 1957 1557 1032  881   35   28
   2    6    3    2    1    6    2    6    2    6    2    5    6    6    1    4    2    4    6    1    2    5    2
1258  608  627  900 1478   16 1054 1946  281 1947  531  985  792  230 1997   19  483 2070 1348 1833  652  633 1838
   6    6    6    6    6    7    1    4    5    4    3    6    4    1    4    4    7    4    6    4    3    3    4
1880 1222 1497 1844  460   29 1148  856 1809  995  593
   4    4    6    4    5    6    4    6    4    6    6
 [ reached getOption("max.print") -- omitted 477 entries ]

Within cluster sum of squares by cluster:
[1]  51.23794 108.68290  32.86772  78.73912  27.13874 112.27801  17.53264
 (between_SS / total_SS =  64.7 %)
```

We use the kmeans function to generate the training labels for the training set, each data point is assigned a cluster based on its characteristics.

(ii) KNN on test set

```
> obesity_subset.test.k7=knn(obesity_subset.train,obesity_subset.test,obesity_subset.train.k7$cluster,k=7)
> obesity_subset.test.k7
  [1] 1 1 2 4 3 5 4 1 4 4 5 2 2 3 3 1 5 7 2 6 5 4 1 5 1 4 1 3 5 1 6 7 3 1 1 2 1 2 7 4 2 6 1 1 4 5 5 2 2 7 7 2 1 5 5 5 3
 [58] 6 7 7 5 1 3 4 7 5 5 7 3 4 5 3 6 7 7 1 6 5 4 6 3 1 7 4 2 1 7 2 1 2 2 4 1 1 1 7 3 5 7 6 5 2 3 1 1 2 2 1 5 2 5 5 5 6
[115] 1 4 6 2 2 7 6 6 3 1 6 1 3 2 7 1 5 3 4 4 5 2 3 1 3 6 6 2 6 3 7 5 5 5 2 2 6 2 4 2 5 7 5 1 5 2 2 4 5 2 5 2 2 5 2 4 4
[172] 6 6 3 3 3 3 5 2 2 7 2 2 6 2 2 6 6 2 3 6 6 2 6 6 6 2 7 2 6 7 3 2 2 2 6 6 7 7 7 2 2 2 3 2 2 6 6 7 2 3 2 2 7 6 2 2 2
[229] 2 2 2 6 6 6 3 2 6 6 2 2 4 6 6 6 6 4 6 4 4 4 2 4 4 6 6 6 4 4 2 4 4 4 2 4 4 4 4 6 4 4 6 4 2 6 4 4 4 2 6 6 4 6 4 2 4
[286] 2 4 4 6 6 6 6 6 6 2 4 2 4 6 6 6 6 1 6 6 1 1 4 6 6 6 1 6 6 6 6 4 1 4 6 6 6 4 6 6 4 4 6 4 6 1 1 6 6 6 4 6 6 6 6 6 1
[343] 6 1 6 6 6 6 6 6 6 6 4 6 4 4 4 4 1 1 1 1 6 6 6 6 6 6 6 6 6 6 4 6 4 4 4 4 4 4 6 6 4 4 4 6 6 4 6 6 6 6 6 6 4 4 6 4 6 4 4
[400] 6 6 4 4 6 4 4 6 6 6 6 4 6 6 4 6 6 6 6 4 4 4 4 4 4 4 4 6 6 4 4 6 4 6 6 6 4 4 6 6 6 4 6 4 4 4 4 6 6 6 6 6 6 6 4 4 6 6
[457] 6 6 4 6 6 6 6 6 1 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6
[514] 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 4
[571] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
[628] 4 4 4 4 4 4 4
Levels: 1 2 3 4 5 6 7
```

(iii) kmeans on test set

```
> obesity_subset.test.kmeans.k7=kmeans(obesity_subset.test,centers = 7)
> obesity_subset.test.kmeans.k7
K-means clustering with 7 clusters of sizes 209, 28, 36, 52, 91, 134, 84

Cluster means:
     Height    Weight      CAEC family_history_with_overweight       FAF       FAVC    Gender NObeyesdad
1 0.3933724 0.4019630 0.9027113                      1.0000000 0.2476079 1.0000000 0.0000000  0.5693780
2 0.2572891 0.1750994 0.8571429                      0.9642857 0.4285905 0.0000000 0.0000000  0.6190476
3 0.5449552 0.2704395 0.8055556                      0.5833333 0.4867856 0.0000000 1.0000000  0.4351852
4 0.6292368 0.2568552 0.8141026                      1.0000000 0.6406986 1.0000000 1.0000000  0.1442308
5 0.3777034 0.1303953 0.7179487                      0.0000000 0.4076570 0.8461538 0.2637363  0.1978022
6 0.5950639 0.5277417 0.9776119                      1.0000000 0.3118258 1.0000000 1.0000000  0.4378109
7 0.5890910 0.3456437 0.9365079                      0.9523810 0.3526298 1.0000000 1.0000000  0.9444444

Clustering vector:
   2    4    6    7    8   11   12   21   23   24   26   31   37   42   45   46   48   54   62   64   68   71   73
   2    3    5    1    3    4    1    3    1    1    3    5    5    3    3    2    4    2    5    6    6    1    2
  75   76   78   80   84   85   86   88   89   91   93   97  101  102  103  104  105  107  111  118  120  121  122
   7    2    1    3    5    7    3    7    1    5    3    2    5    3    5    1    1    5    6    2    2    1    7
 126  128  130  131  133  134  138  140  142  143  147  155  156  157  158  159  174  176  177  178  179  182  186
   7    5    7    1    1    5    3    5    5    5    5    6    1    2    4    2    3    1    1    4    4    1    2
 190  192  194  198  199  202  204  205  207  209  216  218  221  223  225  228  233  235  237  245  247  248  251

1755 1756 1759 1763 1766 1767 1769 1770 1771 1772 1774 1778 1779 1780 1782 1786 1787 1790 1791 1796 1799 1802 1807
   6    6    6    6    6    6    6    6    6    6    6    6    6    6    6    6    6    6    6    6    6    1    1
1810 1819 1823 1825 1828 1839 1848 1859 1865 1867 1869 1870 1877 1879 1881 1882 1884 1887 1888 1892 1893 1899 1905
   1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
1913 1914 1918 1920 1928 1932 1935 1936 1941 1949 1951 1953 1961 1968 1971 1973 1975 1976 1977 1982 1986 1988 1989
   1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
1990 1991 1994 2001 2005 2007 2008 2016 2020 2024 2025 2026 2028 2033 2040 2042 2045 2046 2049 2051 2053 2057 2059
   1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
2060 2061 2063 2068 2073 2074 2076 2077 2078 2082 2083 2104 2109
   1    1    1    1    1    1    1    1    1    1    1    1    1

Within cluster sum of squares by cluster:
[1] 53.57883 10.79093 22.28325 11.23564 60.27978 12.63846 14.60733
 (between_SS / total_SS =  64.7 %)
```

(iv) Test labels

```
> obesity_subset.test.k7.labels=obesity_subset.test.kmeans.k7$cluster
> length(obesity_subset.test.k7.labels)
[1] 634
> obesity_subset.test.k7.labels
  2   4   6   7   8  11  12  21  23  24  26  31  37  42  45  46  48  54  62  64  68  71  73
  2   3   5   1   3   4   1   3   1   1   3   5   5   3   3   2   4   2   5   6   6   1   2
 75  76  78  80  84  85  86  88  89  91  93  97 101 102 103 104 105 107 111 118 120 121 122
  7   2   1   3   5   7   3   7   1   5   3   2   5   3   5   1   1   5   6   2   2   1   7
126 128 130 131 133 134 138 140 142 143 147 155 156 157 158 159 174 176 177 178 179 182 186
  7   5   7   1   1   5   3   5   5   5   5   6   1   2   4   2   3   1   1   4   4   1   2
190 192 194 198 199 202 204 205 207 209 216 218 221 223 225 228 233 235 237 245 247 248 251
  1   7   3   6   1   1   3   4   4   1   6   3   2   1   1   5   2   1   5   2   5   5   1
255 257 259 260 280 282 283 285 286 289 290 294 298 302 304 305 312 313 314 317 322 324 329
  2   3   3   1   3   4   1   4   6   5   3   3   3   5   5   3   5   5   4   3   3   4   3
330 332 339 341 347 348 349 352 360 367 377 378 386 388 389 391 392 395 396 399 405 407 409
```

a. (i) Using glm() to get linear fits for the data

```
> obesity_subset.train.glm=glm(formula = obesity_subset.train$NObeyesdad ~ .,family = gaussian, data=obesity_subset.train)
> summary(obesity_subset.train.glm)

Call:
glm(formula = obesity_subset.train$NObeyesdad ~ ., family = gaussian,
    data = obesity_subset.train)

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                     0.22934    0.03662   6.262 4.98e-10 ***
Height                         -0.37596    0.06344  -5.927 3.85e-09 ***
Weight                          0.56572    0.05133  11.022  < 2e-16 ***
CAEC                            0.28127    0.03289   8.552  < 2e-16 ***
family_history_with_overweight  0.10466    0.02229   4.695 2.91e-06 ***
FAF                            -0.05394    0.02824  -1.910   0.0563 .
FAVC                           -0.10359    0.02456  -4.218 2.61e-05 ***
Gender                          0.04864    0.01905   2.554   0.0108 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.08005764)

    Null deviance: 156.24  on 1476  degrees of freedom
Residual deviance: 117.60  on 1469  degrees of freedom
AIC: 472.09

Number of Fisher Scoring iterations: 2
```

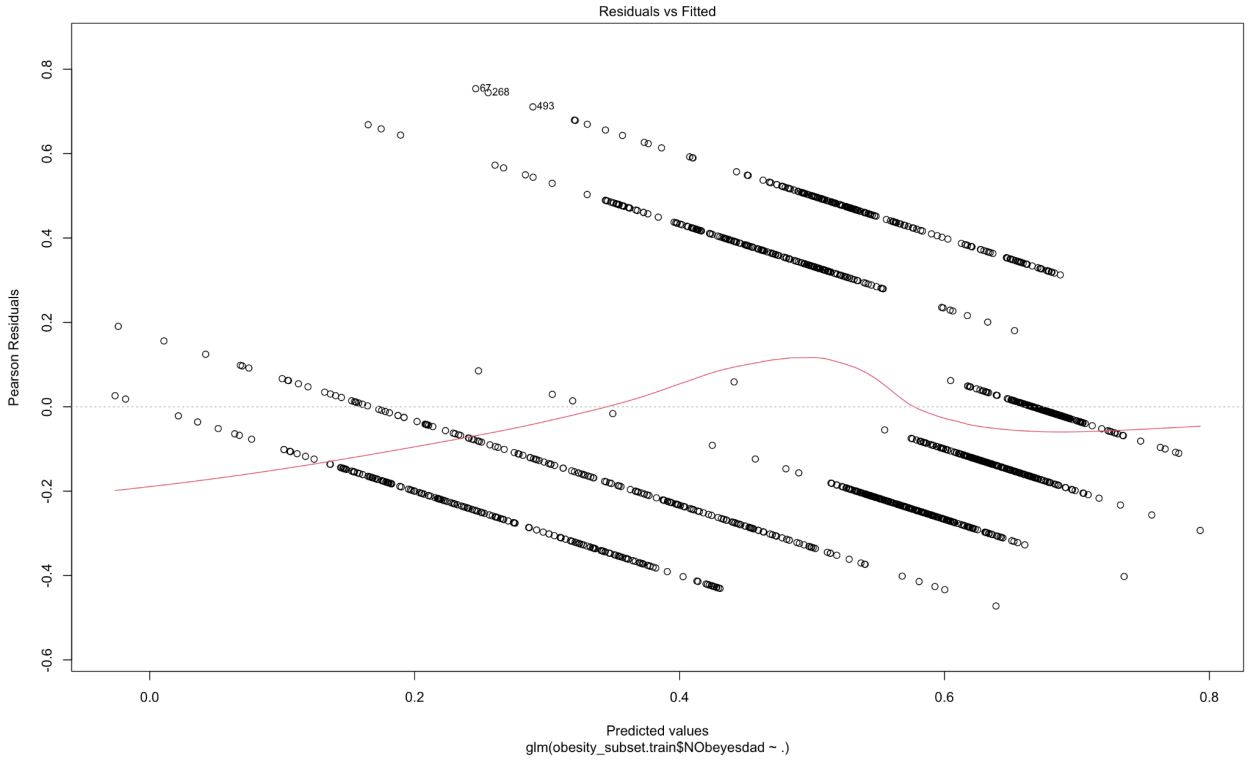
Using the glm function to train the model, with ‘NObeyesdad’ being the response variable and every other column in the dataset as the independent variable.

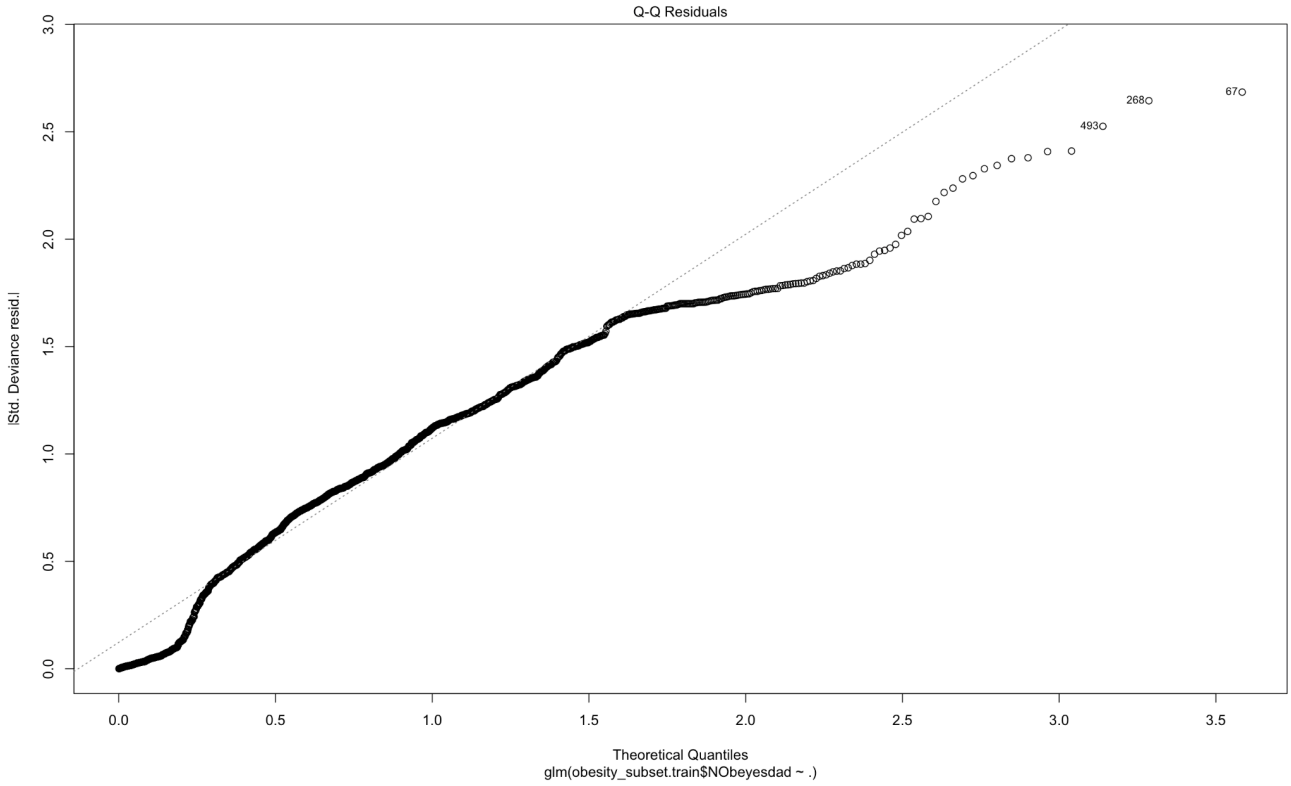
(ii) ANOVA

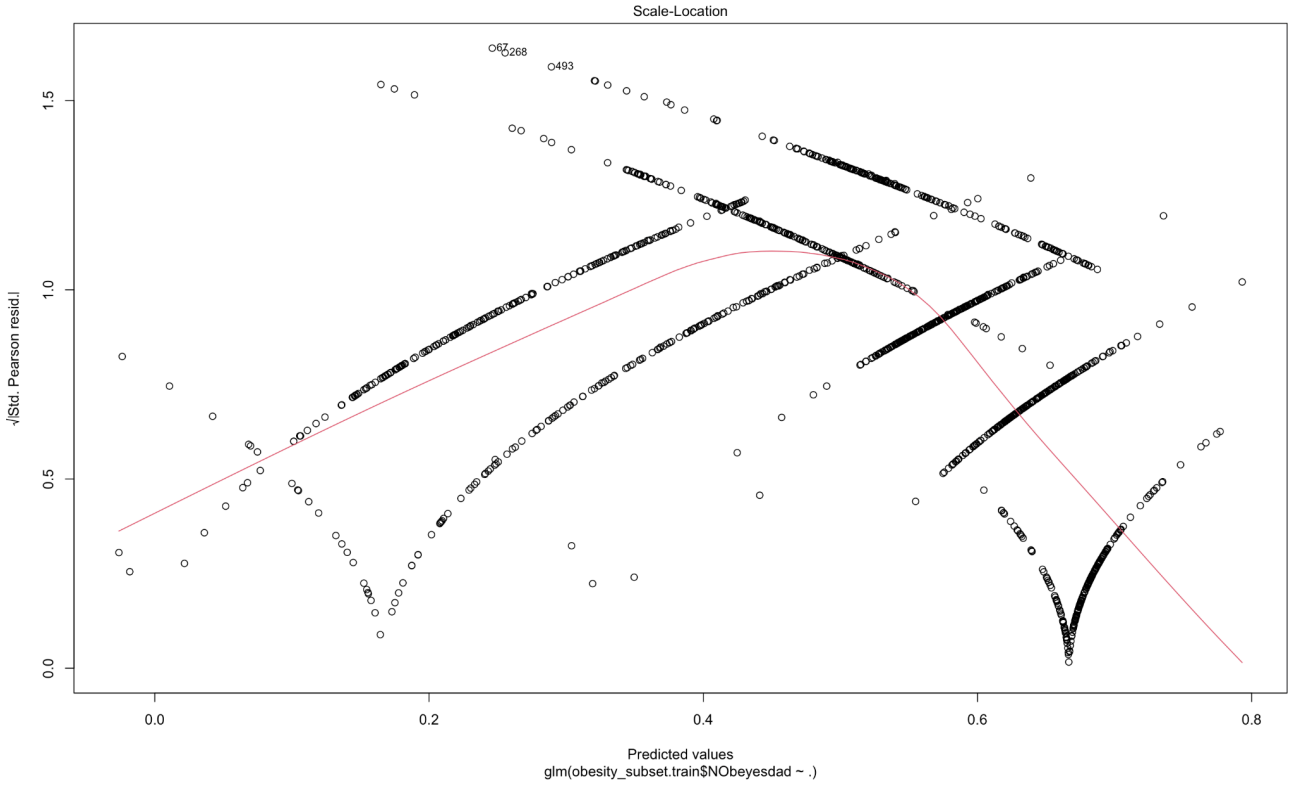
```
> obesity_subset.train.glm.anova=anova(obesity_subset.train.glm,test="Chisq")
> obesity_subset.train.glm.anova
Analysis of Deviance Table

Model: gaussian, link: identity

Response: obesity_subset.train$NObeyesdad

Terms added sequentially (first to last)


                               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                            1476     156.25
Height                          1   0.1327      1475     156.11   0.19793
Weight                          1  27.7663      1474     128.35 < 2.2e-16 ***
CAEC                            1   7.0192      1473     121.33 < 2.2e-16 ***
family_history_with_overweight  1   1.5608      1472     119.77 1.008e-05 ***
FAF                             1   0.1769      1471     119.59   0.13714
FAVC                            1   1.4620      1470     118.13 1.926e-05 ***
Gender                          1   0.5221      1469     117.61   0.01066 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

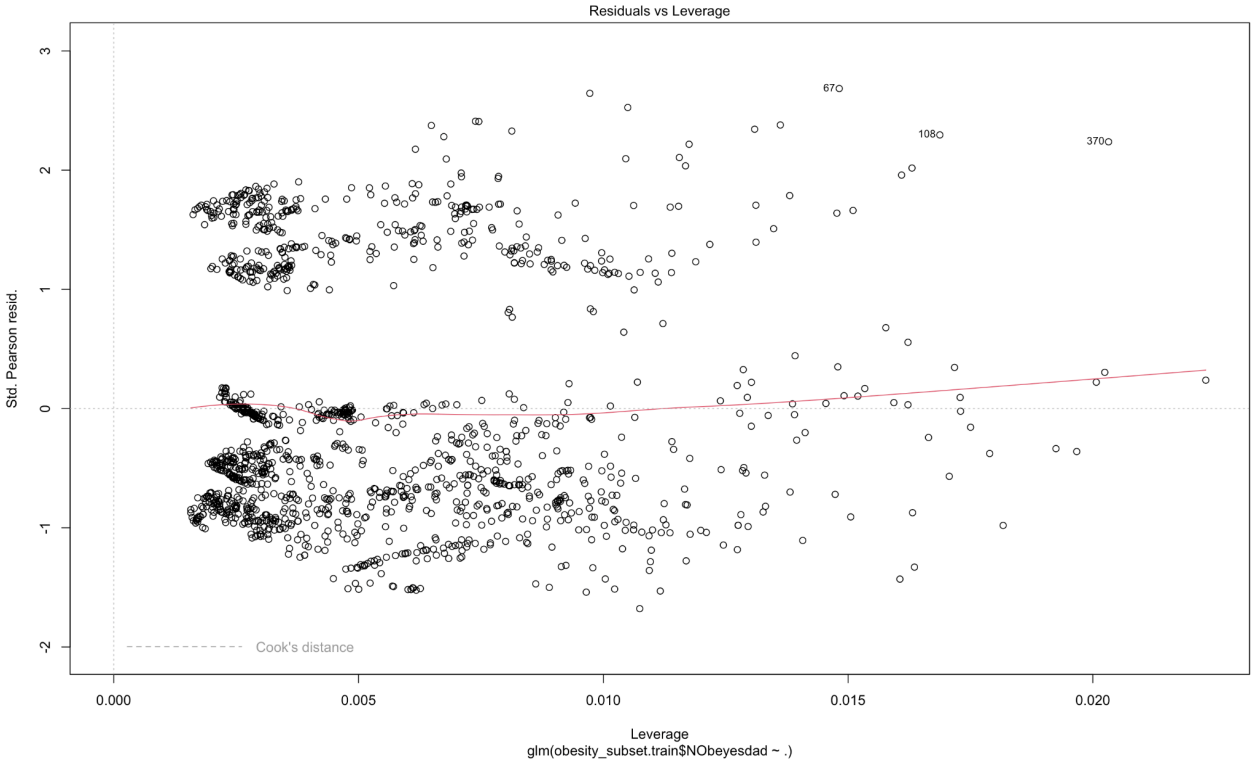
The deviance and residual deviance values have a decent separation between them.

(iii) Plots for glm




Q-Q Residuals
|Std. Deviance resid.|
Theoretical Quantiles
glm(obesity_subset.train$NObeyesdad ~ .)
67
268
493


Scale-Location
√|Std. Pearson resid.|
Predicted values
glm(obesity_subset.train$NObeyesdad ~ .)
67
268
493



The residuals vs fitted plot checks for linearity and homoscedasticity, the QQ plot tests normality of residuals, the scale-location plot checks for constant variance, and the residuals vs leverage plot identifies influential data points that could distort the model.

(iv) Predicting the values using the fitted glm model

```
> obesity_subset.test.pred <- predict(obesity_subset.train.glm,newdata = obesity_subset.test)
> obesity_subset.test.pred
          2           4           6           7           8          11          12          21          23          24
0.583444762 0.477663009 0.394175711 0.525779676 0.429639725 0.331747998 0.269777173 0.641193820 0.440485930 0.586811625
         26          31          37          42          45          46          48          54          62          64
0.362498103 0.226915883 0.335248838 0.297245739 0.347262015 0.668285070 0.263364754 0.425418620 0.340066113 0.464718115
         68          71          73          75          76          78          80          84          85          86
0.373603924 0.261635197 0.544074870 0.389673855 0.476956077 0.507433039 0.610696403 0.299821567 0.381230322 0.604363753
         88          89          91          93          97         101         102         103         104         105
0.500505571 0.274084517 0.293675600 0.378355072 0.501022755 0.338096611 0.414737937 0.188606229 0.244532151 0.606742113
        107         111         118         120         121         122         126         128         130         131
0.353485932 0.499498909 0.426641261 0.420320594 0.320094476 0.311301931 0.239342373 0.340047794 0.424992528 0.273077855
        133         134         138         140         142         143         147         155         156         157
0.236088618 0.079891258 0.730599260 0.213243921 -0.045502934 0.274545642 0.259771065 0.586805979 0.244360468 0.396339493
        158         159         174         176         177         178         179         182         186         190
0.269451597 0.641432760 0.459082020 0.428755777 0.165307404 0.232205100 0.200529868 0.279226839 0.494469615 0.463116473
        192         194         198         199         202         204         205         207         209         216
0.297138886 0.224126209 0.681034871 0.229080689 0.339663754 0.690996074 0.371925886 0.381574217 0.433220741 0.528044613
        218         221         223         225         228         233         235         237         245         247
0.285353807 0.379600674 0.154986812 0.565531109 0.348663010 0.526438931 0.324873101 0.007286777 0.557684690 0.410553374
        248         251         255         257         259         260         280         282         283         285
0.099863555 0.503812428 0.552260613 0.627326210 0.537025001 0.387642741 0.354973990 0.263020858 0.253147367 0.446223761
        286         289         290         294         298         302         304         305         312         313
0.316512516 0.379650980 0.369161530 0.347943761 0.490671675 0.295725582 0.082848573 0.573879379 0.026033274 0.322669644
        314         317         322         324         329         330         332         339         341         347
0.183882218 0.379618993 0.361639689 0.397428169 0.552021673 0.551344099 0.390420240 0.392402386 0.328726876 0.267975788
        348         349         352         360         367         377         378         386         388         389
0.538930942 0.417015291 0.140170293 0.557255218 0.500615643 0.575750101 0.256292525 0.155924336 0.440684066 0.703146325
```

```
      1649       1665       1666       1667       1669       1670       1671       1672       1679       1681
0.659013288 0.656216718 0.712267549 0.696752852 0.649807627 0.627868358 0.619545596 0.628885198 0.608328140 0.671576309
      1686       1690       1691       1693       1695       1696       1701       1702       1707       1709
0.617825130 0.607784390 0.610880642 0.663091909 0.669391330 0.648858303 0.607216311 0.628981883 0.593419115 0.705946923
      1712       1717       1718       1719       1726       1737       1741       1742       1743       1755
0.655930278 0.617456641 0.626143525 0.629462424 0.641565356 0.610402414 0.655491585 0.631922226 0.630950538 0.631782989
      1756       1759       1763       1766       1767       1769       1770       1771       1772       1774
0.596028149 0.641036003 0.656442711 0.642534833 0.620911089 0.626315836 0.621128234 0.621320089 0.628021448 0.624725528
      1778       1779       1780       1782       1786       1787       1790       1791       1796       1799
0.653625533 0.654842534 0.712031861 0.615155393 0.639468260 0.635271382 0.615991453 0.622948117 0.693946326 0.654229150
      1802       1807       1810       1819       1823       1825       1828       1839       1848       1859
0.650390443 0.646540773 0.688698389 0.682709591 0.685394727 0.667611720 0.673354155 0.734476285 0.665260115 0.678155657
      1865       1867       1869       1870       1877       1879       1881       1882       1884       1887
0.654099522 0.687702535 0.681569887 0.695301325 0.671280980 0.733445781 0.675344832 0.643494018 0.690521369 0.678218633
      1888       1892       1893       1899       1905       1913       1914       1918       1920       1928
0.681145077 0.693538795 0.687528940 0.746850579 0.697056977 0.672283868 0.689350011 0.663503921 0.664885418 0.677653731
      1932       1935       1936       1941       1949       1951       1953       1961       1968       1971
0.630937593 0.673775218 0.687213897 0.690775683 0.672998029 0.703437628 0.675421600 0.644911198 0.654559679 0.665868318
      1973       1975       1976       1977       1982       1986       1988       1989       1990       1991
0.652759773 0.687469581 0.673199938 0.679103748 0.625649814 0.685809096 0.691565330 0.711432588 0.689677772 0.690978662
      1994       2001       2005       2007       2008       2016       2020       2024       2025       2026
0.673800158 0.666890631 0.689093149 0.683619335 0.654755412 0.664033621 0.669882859 0.675419424 0.672000082 0.661955836
      2028       2033       2040       2042       2045       2046       2049       2051       2053       2057
0.656223803 0.685198537 0.661686466 0.622636197 0.652347715 0.687845815 0.707168672 0.675338979 0.670068723 0.687835268
      2059       2060       2061       2063       2068       2073       2074       2076       2077       2078
0.680001132 0.689319705 0.694366025 0.710460635 0.662010903 0.652078312 0.666890818 0.732036335 0.652363943 0.697095289
      2082       2083       2104       2109
0.672872793 0.674983762 0.686467994 0.671638102
```

```
> summary(obesity_subset.test.pred)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.0455  0.3826  0.5315  0.4976  0.6273  0.7565
```

(v) Confidence intervals of the prediction

```
> confint(obesity_subset.train.glm)
Waiting for profiling to be done...
                                    2.5 %       97.5 %
(Intercept)                     0.15755927  0.301124549
Height                         -0.50028733 -0.251624222
Weight                          0.46511996  0.666313477
CAEC                            0.21680638  0.345724363
family_history_with_overweight  0.06096942  0.148350387
FAF                            -0.10928036  0.001404535
FAVC                           -0.15171899 -0.055458890
Gender                          0.01131046  0.085973994
```

(vi) Kmeans prediction

```
> obesity_subset.test.pred.k7=kmeans(obesity_subset.test.pred,centers = 7)
> obesity_subset.test.pred.k7
K-means clustering with 7 clusters of sizes 111, 114, 104, 111, 99, 35, 60

Cluster means:
       [,1]
1 0.6794626
2 0.6222878
3 0.4891358
4 0.5563291
5 0.3710486
6 0.1293808
7 0.2542218

Clustering vector:
   2    4    6    7    8   11   12   21   23   24   26   31   37   42   45   46   48   54   62   64   68   71   73   75   76   78
   4    3    5    4    5    5    7    2    3    4    5    7    5    7    5    1    7    5    5    3    5    7    4    5    3    3
  80   84   85   86   88   89   91   93   97  101  102  103  104  105  107  111  118  120  121  122  126  128  130  131  133  134
   2    7    5    2    3    7    7    5    3    5    5    6    7    2    5    3    5    5    5    7    7    5    5    7    7    6
 138  140  142  143  147  155  156  157  158  159  174  176  177  178  179  182  186  190  192  194  198  199  202  204  205  207
   1    7    6    7    7    4    7    5    7    2    3    5    6    7    7    7    3    3    7    7    1    7    5    1    5    5
 209  216  218  221  223  225  228  233  235  237  245  247  248  251  255  257  259  260  280  282  283  285  286  289  290  294
   3    4    7    5    6    4    5    4    5    6    4    5    6    3    4    2    4    5    5    7    7    3    5    5    5    5
 298  302  304  305  312  313  314  317  322  324  329  330  332  339  341  347  348  349  352  360  367  377  378  386  388  389
   3    7    6    4    6    5    6    5    5    5    4    4    5    5    5    7    4    5    6    4    3    4    7    6    3    1
 391  392  395  396  399  405  407  409  411  416  418  419  421  425  426  429  430  433  437  439  443  446  449  455  456  461
   6    5    3    5    5    7    5    4    5    3    3    3    5    3    6    6    7    6    5    5    3    5    1    5    6    7
 463  464  468  470  475  477  479  484  486  491  492  494  495  499  501  508  510  518  519  529  530  535  538  540  544  549
   7    3    7    5    7    3    6    3    6    5    5    6    5    2    1    5    5    7    7    6    6    7    5    5    6    7

   3    3    4    3    3    2    3    4    4    4    3    3    4    3    3    4    3    3    2    1    3    4    4    4    4    4
1126 1129 1130 1132 1134 1138 1143 1145 1152 1153 1155 1157 1159 1160 1163 1170 1171 1172 1174 1177 1178 1179 1181 1182 1183 1185
   3    3    3    2    3    2    4    4    3    4    4    4    4    3    3    4    3    3    4    2    2    2    2    3    3    4
1187 1191 1195 1197 1201 1203 1208 1209 1211 1214 1216 1217 1218 1223 1231 1238 1239 1241 1245 1246 1247 1248 1249 1269 1277 1280
   3    3    3    3    3    4    4    3    4    2    4    2    4    4    4    4    4    4    4    4    4    4    2    4    2    4
1282 1283 1284 1287 1290 1295 1298 1302 1303 1304 1305 1310 1311 1317 1324 1326 1328 1329 1333 1340 1352 1360 1361 1370 1375 1376
   4    4    2    4    2    4    4    4    4    2    2    4    4    4    4    2    4    4    2    2    4    4    4    4    4    4
1378 1379 1384 1386 1387 1388 1391 1397 1402 1404 1406 1408 1409 1413 1414 1420 1422 1428 1440 1441 1443 1446 1449 1451 1454 1457
   4    4    4    4    4    4    4    4    2    2    3    4    4    4    2    4    2    4    2    2    2    4    4    4    4    4
1458 1459 1460 1461 1462 1463 1464 1466 1469 1470 1473 1475 1476 1477 1486 1487 1489 1495 1496 1500 1506 1518 1521 1532 1541 1545
   4    4    2    2    4    2    2    2    4    4    4    4    2    4    2    2    2    4    4    2    2    4    1    2    4    2
1546 1551 1559 1560 1561 1562 1563 1564 1574 1576 1581 1584 1589 1591 1597 1599 1600 1601 1605 1606 1610 1613 1617 1621 1626 1628
   2    1    2    2    2    1    2    2    2    2    2    2    2    2    1    2    1    1    2    2    2    2    2    2    2    2
1630 1633 1634 1635 1643 1648 1649 1665 1666 1667 1669 1670 1671 1672 1679 1681 1686 1690 1691 1693 1695 1696 1701 1702 1707 1709
   2    2    2    2    2    1    1    1    1    1    2    2    2    2    2    1    2    2    2    1    1    2    2    2    2    1
1712 1717 1718 1719 1726 1737 1741 1742 1743 1755 1756 1759 1763 1766 1767 1769 1770 1771 1772 1774 1778 1779 1780 1782 1786 1787
   1    2    2    2    2    2    1    2    2    2    2    2    1    2    2    2    2    2    2    2    1    1    1    2    2    2
1790 1791 1796 1799 1802 1807 1810 1819 1823 1825 1828 1839 1848 1859 1865 1867 1869 1870 1877 1879 1881 1882 1884 1887 1888 1892
   2    2    1    1    2    2    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    2    1    1    1    1
1893 1899 1905 1913 1914 1918 1920 1928 1932 1935 1936 1941 1949 1951 1953 1961 1968 1971 1973 1975 1976 1977 1982 1986 1988 1989
   1    1    1    1    1    1    1    1    2    1    1    1    1    1    1    2    1    1    1    1    1    1    2    1    1    1
1990 1991 1994 2001 2005 2007 2008 2016 2020 2024 2025 2026 2028 2033 2040 2042 2045 2046 2049 2051 2053 2057 2059 2060 2061 2063
   1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    2    1    1    1    1    1    1    1    1    1    1
2068 2073 2074 2076 2077 2078 2082 2083 2104 2109
   1    1    1    1    1    1    1    1    1    1

Within cluster sum of squares by cluster:
[1] 0.05044118 0.03177519 0.06744124 0.04656531 0.10386705 0.10734625 0.06674264
 (between_SS / total_SS =  97.1 %)
```

(vii) Comparing the predictions between kmeans and glm using CrossTable

```
> obesity_subset.test.ct.k7 = CrossTable(obesity_subset.test.pred.k7$cluster,obesity_subset.test.kmeans.k7$cluster,prop.chisq = TRUE)
```

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  634
```

obesity_subset.test.pred.k7$cluster	obesity_subset.test.kmeans.k7$cluster 1	2	3	4	5	6	7	Row Total
1	6	66	0	1	0	8	42	123
	0.054	4.528	14.550	7.228	4.850	0.668	9.241	
	0.049	0.537	0.000	0.008	0.000	0.065	0.341	0.194
	0.176	0.252	0.000	0.021	0.000	0.145	0.309	
	0.009	0.104	0.000	0.002	0.000	0.013	0.066	
2	12	108	0	0	0	0	25	145
	2.294	38.577	17.153	10.749	5.718	12.579	1.198	
	0.083	0.745	0.000	0.000	0.000	0.000	0.172	0.229
	0.353	0.412	0.000	0.000	0.000	0.000	0.184	
	0.019	0.170	0.000	0.000	0.000	0.000	0.039	
3	0	5	30	9	11	0	1	56
	3.003	14.222	82.482	5.663	35.004	4.858	10.096	
	0.000	0.089	0.536	0.161	0.196	0.000	0.018	0.088
	0.000	0.019	0.400	0.191	0.440	0.000	0.007	
	0.000	0.008	0.047	0.014	0.017	0.000	0.002	
4	2	40	4	7	4	12	9	78
	1.139	1.871	2.961	0.256	0.278	4.048	3.573	
	0.026	0.513	0.051	0.090	0.051	0.154	0.115	0.123
	0.059	0.153	0.053	0.149	0.160	0.218	0.066	
	0.059	0.153	0.053	0.149	0.160	0.218	0.066	
	0.003	0.063	0.006	0.011	0.006	0.019	0.014	
5	0	0	25	9	7	0	0	41
	2.199	16.943	83.712	11.689	17.925	3.557	8.795	
	0.000	0.000	0.610	0.220	0.171	0.000	0.000	0.065
	0.000	0.000	0.333	0.191	0.280	0.000	0.000	
	0.000	0.000	0.039	0.014	0.011	0.000	0.000	
6	13	22	0	0	0	35	56	126
	5.768	17.365	14.905	9.341	4.968	53.001	31.055	
	0.103	0.175	0.000	0.000	0.000	0.278	0.444	0.199
	0.382	0.084	0.000	0.000	0.000	0.636	0.412	
	0.021	0.035	0.000	0.000	0.000	0.055	0.088	
7	1	21	16	21	3	0	3	65
	1.773	1.279	8.982	54.339	0.074	5.639	8.589	
	0.015	0.323	0.246	0.323	0.046	0.000	0.046	0.103
	0.029	0.080	0.213	0.447	0.120	0.000	0.022	
	0.002	0.033	0.025	0.033	0.005	0.000	0.005	
Column Total	34	262	75	47	25	55	136	634
	0.054	0.413	0.118	0.074	0.039	0.087	0.215	

(viii) Results for 60%-40% split

```
> obesity_subset2.train.glm=glm(formula = obesity_subset2.train$NObeyesdad ~ .,family = gaussian, data=obesity_subset2.train)
> summary(obesity_subset2.train.glm)

Call:
glm(formula = obesity_subset2.train$NObeyesdad ~ ., family = gaussian,
    data = obesity_subset2.train)

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                     0.23263    0.04013   5.797 8.53e-09 ***
Height                         -0.24726    0.07110  -3.478 0.000523 ***
Weight                          0.54557    0.05792   9.420  < 2e-16 ***
CAEC                            0.25245    0.03451   7.314 4.59e-13 ***
family_history_with_overweight  0.11632    0.02462   4.724 2.57e-06 ***
FAF                            -0.08040    0.03095  -2.597 0.009502 **
FAVC                           -0.11985    0.02703  -4.433 1.01e-05 ***
Gender                          0.02562    0.02147   1.193 0.232918
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.0836835)

    Null deviance: 137.11  on 1265  degrees of freedom
Residual deviance: 105.27  on 1258  degrees of freedom
AIC: 462.14

Number of Fisher Scoring iterations: 2
```

```
> obesity_subset2.train.glm.anova=anova(obesity_subset2.train.glm,test="Chisq")
> obesity_subset2.train.glm.anova
Analysis of Deviance Table

Model: gaussian, link: identity

Response: obesity_subset2.train$NObeyesdad

Terms added sequentially (first to last)


                                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                             1265     137.11
Height                           1   0.4500      1264     136.66   0.02040 *
Weight                           1  22.6008      1263     114.06 < 2.2e-16 ***
CAEC                             1   4.9759      1262     109.08 1.247e-14 ***
family_history_with_overweight   1   1.6915      1261     107.39 6.927e-06 ***
FAF                              1   0.3138      1260     107.08   0.05283 .
FAVC                             1   1.6822      1259     105.39 7.340e-06 ***
Gender                           1   0.1192      1258     105.27   0.23269
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> confint(obesity_subset2.train.glm)
Waiting for profiling to be done...
                                        2.5 %      97.5 %
(Intercept)                          0.15397631  0.31128433
Height                              -0.38660935 -0.10791862
Weight                               0.43205068  0.65908273
CAEC                                 0.18480679  0.32010173
family_history_with_overweight       0.06806205  0.16458352
FAF                                 -0.14106391 -0.01973155
FAVC                                -0.17283932 -0.06686803
Gender                              -0.01645599  0.06769940
```

```
> summary(obesity_subset2.test.pred)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.006509 0.400432 0.547540 0.511719 0.639128 0.816851
```

```
> obesity_subset2.test.ct.k7 = CrossTable(obesity_subset2.test.pred.k7$cluster,obesity_subset2.test.kmeans.k7$cluster,prop.chisq = TRUE)

   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  845
```

obesity_subset2.test.pred.k7$cluster	obesity_subset2.test.kmeans.k7$cluster 1	2	3	4	5	6	7	Row Total
1	36	5	39	6	12	4	3	105
	47.576	0.035	0.975	0.329	10.236	0.717	6.800	
	0.343	0.048	0.371	0.057	0.114	0.038	0.029	0.124
	0.371	0.135	0.146	0.098	0.051	0.082	0.031	
	0.043	0.006	0.046	0.007	0.014	0.005	0.004	
2	0	0	24	22	116	0	0	162
	18.596	7.093	14.590	9.081	110.648	9.394	18.596	
	0.000	0.000	0.148	0.136	0.716	0.000	0.000	0.192
	0.000	0.000	0.090	0.361	0.492	0.000	0.000	
	0.000	0.000	0.028	0.026	0.137	0.000	0.000	
3	0	0	134	17	35	0	0	186
	21.351	8.144	95.373	0.951	5.529	10.786	21.351	
	0.000	0.000	0.720	0.091	0.188	0.000	0.000	0.220
	0.000	0.000	0.500	0.279	0.148	0.000	0.000	
	0.000	0.000	0.159	0.020	0.041	0.000	0.000	
4	53	2	64	7	32	0	0	158
	67.012	3.497	3.849	1.702	3.333	9.162	18.137	
	0.335	0.013	0.405	0.044	0.203	0.000	0.000	0.187
	0.546	0.054	0.239	0.115	0.136	0.000	0.000	
	0.546	0.054	0.239	0.115	0.136	0.000	0.000	
	0.063	0.002	0.076	0.008	0.038	0.000	0.000	
5	1	5	0	0	3	7	35	51
	4.025	3.428	16.175	3.682	8.876	5.526	145.097	
	0.020	0.098	0.000	0.000	0.059	0.137	0.686	0.060
	0.010	0.135	0.000	0.000	0.013	0.143	0.361	
	0.001	0.006	0.000	0.000	0.004	0.008	0.041	
6	5	8	6	8	30	19	21	97
	3.380	3.316	19.935	0.142	0.312	31.804	8.740	
	0.052	0.082	0.062	0.082	0.309	0.196	0.216	0.115
	0.052	0.216	0.022	0.131	0.127	0.388	0.216	
	0.006	0.009	0.007	0.009	0.036	0.022	0.025	
7	2	17	1	1	8	19	38	86
	6.277	46.511	25.312	4.369	10.683	39.375	80.142	
	0.023	0.198	0.012	0.012	0.093	0.221	0.442	0.102
	0.021	0.459	0.004	0.016	0.034	0.388	0.392	
	0.002	0.020	0.001	0.001	0.009	0.022	0.045	
Column Total	97	37	268	61	236	49	97	845
	0.115	0.044	0.317	0.072	0.279	0.058	0.115	

(ix) Results for 50%-50% split

```
> obesity_subset3.train.glm=glm(formula = obesity_subset3.train$NObeyesdad ~ .,family = gaussian, data=obesity_subset3.train)
> summary(obesity_subset3.train.glm)

Call:
glm(formula = obesity_subset3.train$NObeyesdad ~ ., family = gaussian,
    data = obesity_subset3.train)

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                     0.23903    0.04361   5.481 5.29e-08 ***
Height                         -0.37447    0.07737  -4.840 1.49e-06 ***
Weight                          0.59420    0.06115   9.717  < 2e-16 ***
CAEC                            0.23108    0.03760   6.146 1.13e-09 ***
family_history_with_overweight  0.11304    0.02666   4.240 2.44e-05 ***
FAF                            -0.08192    0.03480  -2.354  0.01876 *
FAVC                           -0.08758    0.02838  -3.086  0.00208 **
Gender                          0.06116    0.02275   2.688  0.00729 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.08088854)

    Null deviance: 111.66  on 1054  degrees of freedom
Residual deviance:  84.69  on 1047  degrees of freedom
AIC: 350.94

Number of Fisher Scoring iterations: 2
```

```
> obesity_subset3.train.glm.anova=anova(obesity_subset3.train.glm,test="Chisq")
> obesity_subset3.train.glm.anova
Analysis of Deviance Table

Model: gaussian, link: identity

Response: obesity_subset3.train$NObeyesdad

Terms added sequentially (first to last)


                                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                              1054    111.665
Height                           1   0.1925      1053    111.472  0.122874
Weight                           1  20.0962      1052     91.376 < 2.2e-16 ***
CAEC                             1   3.7586      1051     87.618 9.318e-12 ***
family_history_with_overweight   1   1.2397      1050     86.378 9.047e-05 ***
FAF                              1   0.3121      1049     86.066  0.049491 *
FAVC                             1   0.7909      1048     85.275  0.001767 **
Gender                           1   0.5846      1047     84.690  0.007178 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> confint(obesity_subset3.train.glm)
Waiting for profiling to be done...
                                        2.5 %      97.5 %
(Intercept)                          0.15355489  0.32449525
Height                              -0.52610952 -0.22282988
Weight                               0.47434926  0.71406004
CAEC                                 0.15738512  0.30476596
family_history_with_overweight       0.06078202  0.16529968
FAF                                 -0.15011849 -0.01371225
FAVC                                -0.14319407 -0.03196386
Gender                               0.01657203  0.10574403
```

```
> summary(obesity_subset3.test.pred)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.006571  0.376000  0.529531  0.493937  0.630322  0.791519
```

```
> obesity_subset3.test.ct.k7 = CrossTable(obesity_subset3.test.pred.k7$cluster,obesity_subset3.test.kmeans.k7$cluster,prop.chisq = TRUE)

   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  1056

                                 | obesity_subset3.test.kmeans.k7$cluster
obesity_subset3.test.pred.k7$cluster |         1 |         2 |         3 |         4 |         5 |         6 |         7 | Row Total |
-------------------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
                                   1 |        93 |         0 |         0 |        60 |        17 |        90 |         0 |       260 |
                                     |    21.342 |    29.299 |    18.958 |    32.964 |     0.030 |     2.450 |    30.777 |           |
                                     |     0.358 |     0.000 |     0.000 |     0.231 |     0.065 |     0.346 |     0.000 |     0.246 |
                                     |     0.396 |     0.000 |     0.000 |     0.508 |     0.236 |     0.290 |     0.000 |           |
                                     |     0.088 |     0.000 |     0.000 |     0.057 |     0.016 |     0.085 |     0.000 |           |
-------------------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
                                   2 |         8 |        35 |        37 |         5 |         6 |        38 |         3 |       132 |
                                     |    15.554 |    27.228 |    77.859 |     6.445 |     1.000 |     0.015 |    10.201 |           |
                                     |     0.061 |     0.265 |     0.280 |     0.038 |     0.045 |     0.288 |     0.023 |     0.125 |
                                     |     0.034 |     0.294 |     0.481 |     0.042 |     0.083 |     0.123 |     0.024 |           |
                                     |     0.008 |     0.033 |     0.035 |     0.005 |     0.006 |     0.036 |     0.003 |           |
-------------------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
                                   3 |        44 |         0 |         1 |         5 |        29 |       119 |        14 |       212 |
                                     |     0.214 |    23.890 |    13.523 |    14.745 |    14.637 |    51.776 |     4.905 |           |
                                     |     0.208 |     0.000 |     0.005 |     0.024 |     0.137 |     0.561 |     0.066 |     0.201 |
                                     |     0.187 |     0.000 |     0.013 |     0.042 |     0.403 |     0.384 |     0.112 |           |
                                     |     0.042 |     0.000 |     0.001 |     0.005 |     0.027 |     0.113 |     0.013 |           |
-------------------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
                                   4 |        15 |        55 |        10 |         7 |         1 |         8 |         0 |        96 |
                                     |     1.896 |   180.440 |     1.286 |     1.295 |     4.698 |    14.453 |    11.364 |           |
                                     |     0.156 |     0.573 |     0.104 |     0.073 |     0.010 |     0.083 |     0.000 |     0.091 |
                                     |     0.064 |     0.462 |     0.130 |     0.059 |     0.014 |     0.026 |     0.000 |           |
```

	0.064	0.462	0.130	0.059	0.014	0.026	0.000	
	0.014	0.052	0.009	0.007	0.001	0.008	0.000	
5	57	2	4	15	8	33	97	216
	1.660	20.505	8.766	3.458	3.073	14.583	199.565	
	0.264	0.009	0.019	0.069	0.037	0.153	0.449	0.205
	0.243	0.017	0.052	0.127	0.111	0.106	0.776	
	0.054	0.002	0.004	0.014	0.008	0.031	0.092	
6	18	13	15	22	11	20	11	110
	1.715	0.029	6.073	7.668	1.633	4.679	0.314	
	0.164	0.118	0.136	0.200	0.100	0.182	0.100	0.104
	0.077	0.109	0.195	0.186	0.153	0.065	0.088	
	0.017	0.012	0.014	0.021	0.010	0.019	0.010	
7	0	14	10	4	0	2	0	30
	6.676	33.357	27.902	0.125	2.045	5.261	3.551	
	0.000	0.467	0.333	0.133	0.000	0.067	0.000	0.028
	0.000	0.118	0.130	0.034	0.000	0.006	0.000	
	0.000	0.013	0.009	0.004	0.000	0.002	0.000	
Column Total	235	119	77	118	72	310	125	1056
	0.223	0.113	0.073	0.112	0.068	0.294	0.118	

(x) Results for 70%-30% split with ‘quasi’ (quasiprobability) distribution instead of ‘gaussian’

```
> summary(obesity_subset.train.glm)

Call:
glm(formula = obesity_subset.train$NObeyesdad ~ ., family = quasi,
    data = obesity_subset.train)

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                     0.23473    0.03720   6.309 3.70e-10 ***
Height                         -0.33441    0.06478  -5.162 2.77e-07 ***
Weight                          0.55056    0.05217  10.553  < 2e-16 ***
CAEC                            0.26111    0.03145   8.303 2.28e-16 ***
family_history_with_overweight  0.12111    0.02256   5.368 9.26e-08 ***
FAF                            -0.06782    0.02856  -2.375   0.0177 *
FAVC                           -0.10093    0.02422  -4.166 3.28e-05 ***
Gender                          0.03489    0.01929   1.809   0.0707 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasi family taken to be 0.08269117)

    Null deviance: 159.17  on 1476  degrees of freedom
Residual deviance: 121.47  on 1469  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 2
```

```
> obesity_subset.train.glm.anova=anova(obesity_subset.train.glm,test="Chisq")
> obesity_subset.train.glm.anova
Analysis of Deviance Table

Model: quasi, link: identity

Response: obesity_subset.train$NObeyesdad

Terms added sequentially (first to last)


                              Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                           1476     159.17
Height                         1   0.0799      1475     159.09   0.32561
Weight                         1  27.0420      1474     132.04 < 2.2e-16 ***
CAEC                           1   6.2910      1473     125.75 < 2.2e-16 ***
family_history_with_overweight 1   2.2235      1472     123.53 2.154e-07 ***
FAF                            1   0.3122      1471     123.22   0.05202 .
FAVC                           1   1.4731      1470     121.74 2.435e-05 ***
Gender                         1   0.2705      1469     121.47   0.07049 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> confint(obesity_subset.train.glm)
Waiting for profiling to be done...
                                    2.5 %      97.5 %
(Intercept)                     0.161810019  0.30764137
Height                         -0.461365991 -0.20744451
Weight                          0.448303826  0.65281806
CAEC                            0.199474341  0.32275118
family_history_with_overweight  0.076884958  0.16532725
FAF                            -0.123798972 -0.01183990
FAVC                           -0.148408838 -0.05344986
Gender                         -0.002916905  0.07269726
```

```
> summary(obesity_subset.test.pred)
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-0.0000174  0.4057617  0.5458519  0.5108846  0.6371660  0.7869863
```

```
> obesity_subset.test.ct.k7 = CrossTable(obesity_subset.test.pred.k7$cluster,obesity_subset.test.kmeans.k7$cluster,prop.chisq = TRUE)


   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  634


                                 | obesity_subset.test.kmeans.k7$cluster
obesity_subset.test.pred.k7$cluster |         1 |         2 |         3 |         4 |         5 |         6 |         7 | Row Total |
-------------------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
                                   1 |         0 |         0 |        25 |         9 |         7 |         0 |         0 |        41 |
                                     |     2.199 |    16.943 |    83.712 |    11.689 |    17.925 |     3.557 |     8.795 |           |
                                     |     0.000 |     0.000 |     0.610 |     0.220 |     0.171 |     0.000 |     0.000 |     0.065 |
                                     |     0.000 |     0.000 |     0.333 |     0.191 |     0.280 |     0.000 |     0.000 |           |
                                     |     0.000 |     0.000 |     0.039 |     0.014 |     0.011 |     0.000 |     0.000 |           |
-------------------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
                                   2 |        14 |       107 |         0 |         0 |         0 |         0 |        19 |       140 |
                                     |     5.614 |    41.747 |    16.562 |    10.379 |     5.521 |    12.145 |     4.052 |           |
                                     |     0.100 |     0.764 |     0.000 |     0.000 |     0.000 |     0.000 |     0.136 |     0.221 |
                                     |     0.412 |     0.408 |     0.000 |     0.000 |     0.000 |     0.000 |     0.140 |           |
                                     |     0.022 |     0.169 |     0.000 |     0.000 |     0.000 |     0.000 |     0.030 |           |
-------------------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
                                   3 |         2 |        42 |         4 |         7 |         4 |        13 |        10 |        82 |
                                     |     1.307 |     1.943 |     3.350 |     0.140 |     0.182 |     4.871 |     3.275 |           |
                                     |     0.024 |     0.512 |     0.049 |     0.085 |     0.049 |     0.159 |     0.122 |     0.129 |
                                     |     0.059 |     0.160 |     0.053 |     0.149 |     0.160 |     0.236 |     0.074 |           |
                                     |     0.003 |     0.066 |     0.006 |     0.011 |     0.006 |     0.021 |     0.016 |           |
-------------------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
                                   4 |         1 |        21 |        16 |        21 |         3 |         0 |         3 |        65 |
                                     |     1.773 |     1.279 |     8.982 |    54.339 |     0.074 |     5.639 |     8.589 |           |
                                     |     0.015 |     0.323 |     0.246 |     0.323 |     0.046 |     0.000 |     0.046 |     0.103 |
                                     |     0.029 |     0.080 |     0.213 |     0.447 |     0.120 |     0.000 |     0.022 |           |

                                     |     0.029 |     0.080 |     0.213 |     0.447 |     0.120 |     0.000 |     0.022 |           |
                                     |     0.002 |     0.033 |     0.025 |     0.033 |     0.005 |     0.000 |     0.005 |           |
-------------------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
                                   5 |         0 |         5 |        30 |         9 |        11 |         0 |         1 |        56 |
                                     |     3.003 |    14.222 |    82.482 |     5.663 |    35.004 |     4.858 |    10.096 |           |
                                     |     0.000 |     0.089 |     0.536 |     0.161 |     0.196 |     0.000 |     0.018 |     0.088 |
                                     |     0.000 |     0.019 |     0.400 |     0.191 |     0.440 |     0.000 |     0.007 |           |
                                     |     0.000 |     0.008 |     0.047 |     0.014 |     0.017 |     0.000 |     0.002 |           |
-------------------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
                                   6 |         6 |        67 |         0 |         1 |         0 |         7 |        47 |       128 |
                                     |     0.109 |     3.761 |    15.142 |     7.594 |     5.047 |     1.517 |    13.909 |           |
                                     |     0.047 |     0.523 |     0.000 |     0.008 |     0.000 |     0.055 |     0.367 |     0.202 |
                                     |     0.176 |     0.256 |     0.000 |     0.021 |     0.000 |     0.127 |     0.346 |           |
                                     |     0.009 |     0.106 |     0.000 |     0.002 |     0.000 |     0.011 |     0.074 |           |
-------------------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
                                   7 |        11 |        20 |         0 |         0 |         0 |        35 |        56 |       122 |
                                     |     3.037 |    18.350 |    14.432 |     9.044 |     4.811 |    56.329 |    34.001 |           |
                                     |     0.090 |     0.164 |     0.000 |     0.000 |     0.000 |     0.287 |     0.459 |     0.192 |
                                     |     0.324 |     0.076 |     0.000 |     0.000 |     0.000 |     0.636 |     0.412 |           |
                                     |     0.017 |     0.032 |     0.000 |     0.000 |     0.000 |     0.055 |     0.088 |           |
-------------------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
                        Column Total |        34 |       262 |        75 |        47 |        25 |        55 |       136 |       634 |
                                     |     0.054 |     0.413 |     0.118 |     0.074 |     0.039 |     0.087 |     0.215 |           |
-------------------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
```

b. Performance metrics

Note: These metrics are the average over all clusters.

Dataset Split (%)	Accuracy	Recall (Sensitivity)	Specificity	Precision	F-score
70-30 (Quasi)	0.3249211	0.2696537	0.8827498	0.2637537	0.2900850
70-30 (Gaussian)	0.3091483	0.2965776	0.8824303	0.2733899	0.2579819
60-40	0.2804734	0.2540154	0.8814517	0.2577359	0.2845252
50-50	0.1553030	0.1339712	0.8600643	0.1313332	0.1462961

These metrics indicate that the larger the split, the greater the accuracy of the model. This makes sense since a model trained on more of the dataset will (most likely) have better performance when looking at data it hasn't seen. However, this could introduce potential overfitting issues and account for noise that doesn't exist in the real world.

5. Miscellaneous functions from the rubric

(i) Extracting a single attribute from the entire dataset

```
> heights<-obesity[,3]
> heights
   [1] 1.620000 1.520000 1.800000 1.800000 1.780000 1.620000 1.500000 1.640000 1.780000 1.720000 1.850000 1.720000
  [13] 1.650000 1.800000 1.770000 1.700000 1.930000 1.530000 1.710000 1.650000 1.650000 1.690000 1.650000 1.600000
  [25] 1.850000 1.600000 1.700000 1.600000 1.750000 1.680000 1.770000 1.580000 1.770000 1.790000 1.650000 1.500000
  [37] 1.560000 1.600000 1.650000 1.750000 1.670000 1.680000 1.660000 1.660000 1.810000 1.530000 1.820000 1.750000
  [49] 1.660000 1.550000 1.610000 1.500000 1.640000 1.630000 1.600000 1.680000 1.700000 1.640000 1.650000 1.760000
  [61] 1.550000 1.650000 1.670000 1.680000 1.660000 1.620000 1.800000 1.650000 1.760000 1.800000 1.650000 1.670000
  [73] 1.650000 1.850000 1.700000 1.630000 1.600000 1.700000 1.650000 1.650000 1.630000 1.800000 1.670000 1.600000
  [85] 1.700000 1.650000 1.850000 1.820000 1.650000 1.700000 1.630000 1.610000 1.780000 1.600000 1.600000 1.700000
  [97] 1.660000 1.520000 1.520000 1.720000 1.690000 1.700000 1.550000 1.650000 1.560000 1.570000 1.570000 1.880000
 [109] 1.750000 1.650000 1.750000 1.580000 1.560000 1.500000 1.610000 1.750000 1.650000 1.700000 1.620000 1.630000
 [121] 1.670000 1.870000 1.750000 1.660000 1.760000 1.750000 1.670000 1.650000 1.720000 1.700000 1.580000 1.620000
 [133] 1.650000 1.650000 1.770000 1.700000 1.790000 1.600000 1.760000 1.700000 1.890000 1.870000 1.740000 1.680000
 [145] 1.610000 1.620000 1.560000 1.630000 1.600000 1.670000 1.780000 1.620000 1.500000 1.690000 1.740000 1.680000
 [157] 1.530000 1.670000 1.550000 1.640000 1.830000 1.650000 1.630000 1.890000 1.770000 1.920000 1.740000 1.650000
 [169] 1.730000 1.630000 1.720000 1.600000 1.650000 1.740000 1.620000 1.640000 1.570000 1.840000 1.910000 1.620000
 [181] 1.580000 1.680000 1.680000 1.480000 1.620000 1.620000 1.780000 1.780000 1.780000 1.630000 1.600000 1.750000
 [193] 1.830000 1.780000 1.600000 1.800000 1.750000 1.750000 1.590000 1.660000 1.630000 1.540000 1.560000 1.690000
```

(ii) Min and max of an attribute

```
> obesity2.min.height<-min(obesity2$Height)
> obesity2.min.height
[1] 1.45
> obesity2.min.age<-min(obesity2$Age)
> obesity2.min.age
[1] 14
```

```
> obesity2.max.weight<-max(obesity2$Weight)
> obesity2.max.weight
[1] 173
> obesity2.max.height<-max(obesity2$Height)
> obesity2.max.height
[1] 1.98
```

(iii) zscore

```
> zscore<-function(x){(x-mean(x))/sd(x)}
> zscore(c(110,120,130,140,150))
[1] -1.2649111 -0.6324555  0.0000000  0.6324555  1.2649111
```

6. Analysis and Discussion

Through the analysis of the obesity dataset, we have gained insights into how raw data can be transformed, normalized, and analyzed to extract patterns. Before applying any machine learning models, it was essential to understand the dataset through visualization, correlation analysis, and statistical summaries. By using techniques like normalization and dimensionality reduction, we saw how these preprocessing steps impact clustering and classification outcomes. The use of the Between-Cluster Sum of Squares (BSS) to Total Sum of Squares (SST) ratio demonstrated how model performance improves with more clusters but with diminishing returns. This highlighted a key concept in data science- finding a balance between model complexity and interpretability. It has shown us that data exploration is not just a preliminary step but a crucial phase that significantly impacts the success of predictive modeling.