CS4907/CS6444 Big Data and Analytics

Class project #2 Report

Spring 2025

Group 4

Tianfang Fang, Rohith Vijayakumar

1. For this data set, plot the data using pairwise plotting to get a sense of the relationships between the attributes.

a. Try plotting the data using several plotting functions to see what it looks like. Use pairs (e.g., 2D plots) or 3 variables (3D plots) based on the packages.

```
install.packages("GGally")
install.packages("plotly")
install.packages("reshape2")
install.packages("ggplot2")
library(GGally)
library(ggplot2)
library(reshape2)
library(plotly)

# Load the data
od <- read.csv('/Users/tianfangfang/Documents/Introduction to Big Data and Analytics_CSCI_6444/Project_2/ObesityDataSet_raw_and_data_sinthetic.csv', stringsAsFactors = FALSE)
od1 <- read.csv('/Users/tianfangfang/Documents/Introduction to Big Data and Analytics_CSCI_6444/Project_2/ObesityDataSet_raw_and_data_sinthetic.csv', stringsAsFactors = FALSE)
str(od)
str(od1)
```

Firstly, we installed the required packages and loaded them, then we loaded the data as od and od1.

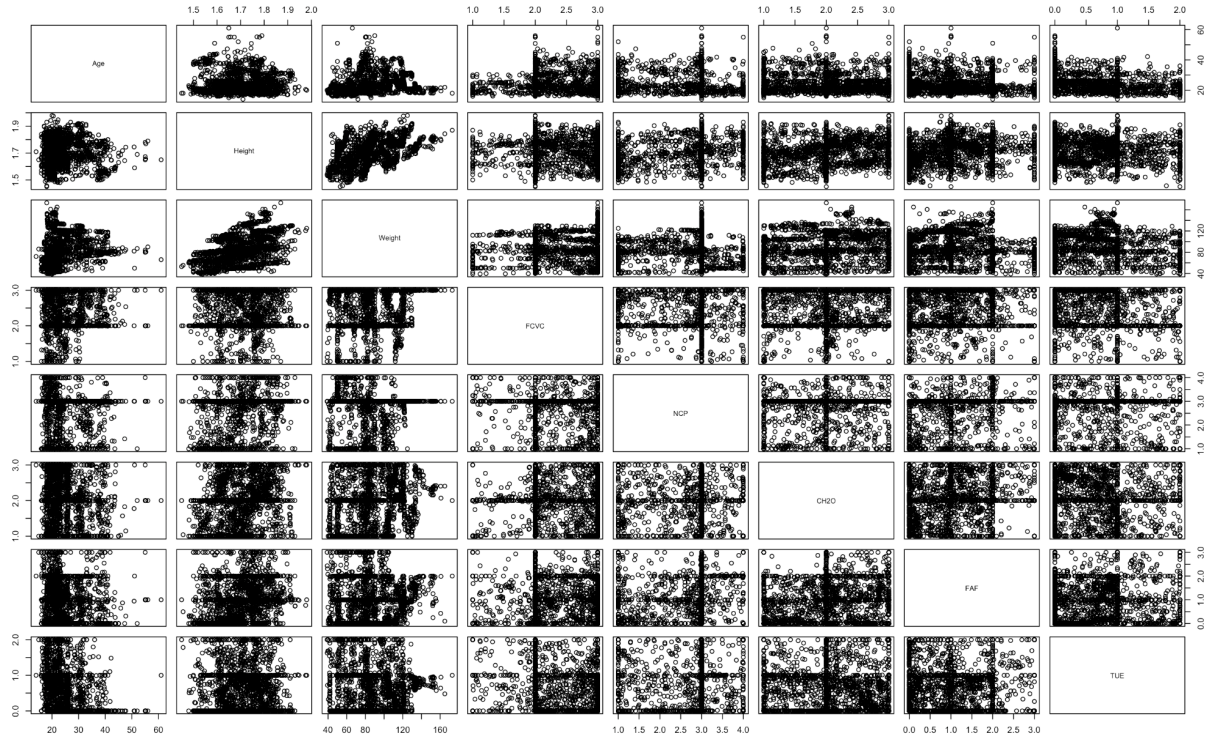
```
> str(od1)
'data.frame':   2111 obs. of  17 variables:
 $ Gender                        : chr  "Female" "Female" "Male" "Male" ...
 $ Age                           : num  21 21 23 27 22 29 23 22 24 22 ...
 $ Height                        : num  1.62 1.52 1.8 1.8 1.78 1.62 1.5 1.64 1.78 1.72 ...
 $ 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" "Public_Transportation" "Walking" ...
 $ NObeyesdad                    : Factor w/ 7 levels "Insufficient_Weight",..: 2 2 2 3 4 2 2 2 2 2 ...
```

The structure of the dataset includes **numeric attributes**: Age, Height, Weight, FCVC, NCP, CH2O, FAF and TUE, and **categorical attributes** such as Gender, family_history_with_overweight, FAVC, CAEC, SMOKE, SCC, CALC, MTRANS and NObeyesdad.

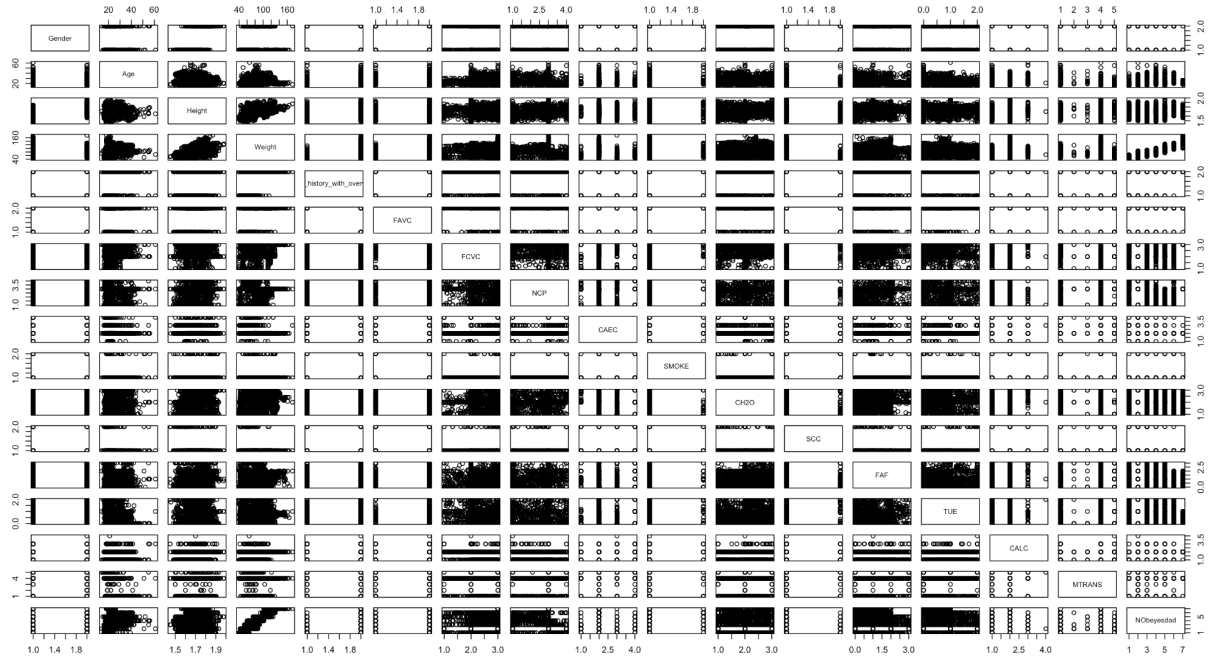
```
> str(od)
'data.frame':   2111 obs. of  17 variables:
 $ Gender                        : num  NA NA NA NA NA NA NA NA NA NA ...
 $ Age                           : num  21 21 23 27 22 29 23 22 24 22 ...
 $ Height                        : num  1.62 1.52 1.8 1.8 1.78 1.62 1.5 1.64 1.78 1.72 ...
 $ Weight                        : 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  NA NA NA NA NA NA NA NA NA NA ...
 $ 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  NA NA NA NA NA NA NA NA NA NA ...
 $ SMOKE                         : num  NA NA NA NA NA NA NA NA NA NA ...
 $ CH2O                          : num  2 3 2 2 2 2 2 2 2 2 ...
 $ SCC                           : num  NA NA NA NA NA NA NA NA NA NA ...
 $ 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  NA NA NA NA NA NA NA NA NA NA ...
 $ MTRANS                        : num  NA NA NA NA NA NA NA NA NA NA ...
 $ NObeyesdad                    : num  NA NA NA NA NA NA NA NA NA NA ...
```

```r
od_columnName <- names(od)
od_columnName
od_categorical <- od[,c(1,5,6,9,10,12,15,16,17)]
od_numerical <- od[,c(2,3,4,7,8,11,13,14)]

# 2D Pairwise Plot using pairs() function
#Pairwise plot for numeric attributes
pairs(od_numerical)

#Converting categorical data to numeric in od by labeling them
od$Gender = as.numeric(factor(od$Gender, levels = c("Female", "Male")))
od$family_history_with_overweight = as.numeric(factor(od$family_history_with_overweight, levels = c("no", "yes")))
od$FAVC = as.numeric(factor(od$FAVC, levels = c("no", "yes")))
od$CAEC = as.numeric(factor(od$CAEC, levels = c("no", "Sometimes","Frequently","Always")))
od$SMOKE = as.numeric(factor(od$SMOKE, levels = c("no", "yes")))
od$SCC = as.numeric(factor(od$SCC, levels = c("no", "yes")))
od$CALC = as.numeric(factor(od$CALC, levels = c("no", "Sometimes","Frequently","Always")))
od$MTRANS = as.numeric(factor(od$MTRANS, levels = c("Automobile", "Bike","Motorbike","Public_Transportation","Walking")))
od$NObeyesdad = as.numeric(factor(od$NObeyesdad, levels = c("Insufficient_Weight", "Normal_Weight","Overweight_Level_I","Overweight_Level_II","Obesity_Type_I","Obesity_Type_II","Obesity_Type_II

#Pairwise plot for all attributes
pairs(od)
```

We did pairwise plotting on the numerical dataset, and after converting categorical data to numeric by labeling them we did the pairwise plotting on the whole dataset.



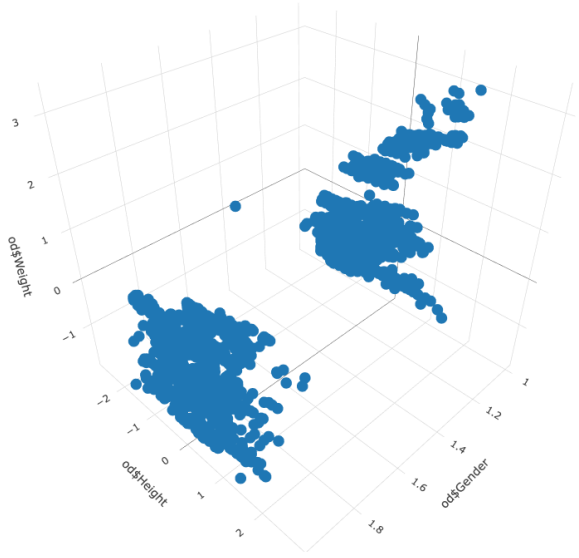
This shows the pairwise plot on numeric attributes.



This shows the pairwise plot on the whole dataset.

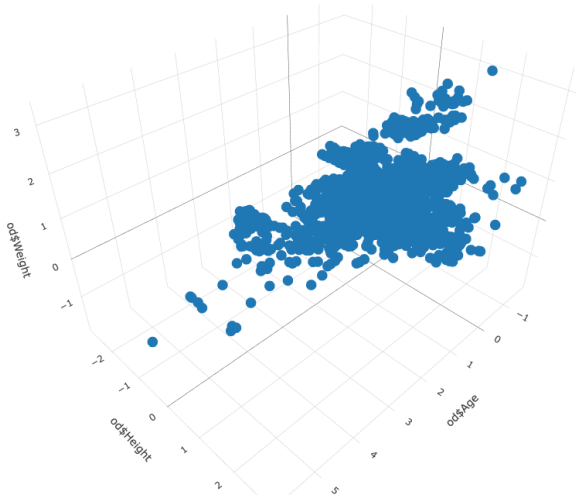
3D plotting on Height, Weight and Gender.

```
#3d Plotting on Height, Weight and Gender
fig <- plot_ly(data = od, x = ~od$Gender, y = ~od$Height, z = ~od$Weight, type = 'scatter3d', mode = 'markers')
fig
```



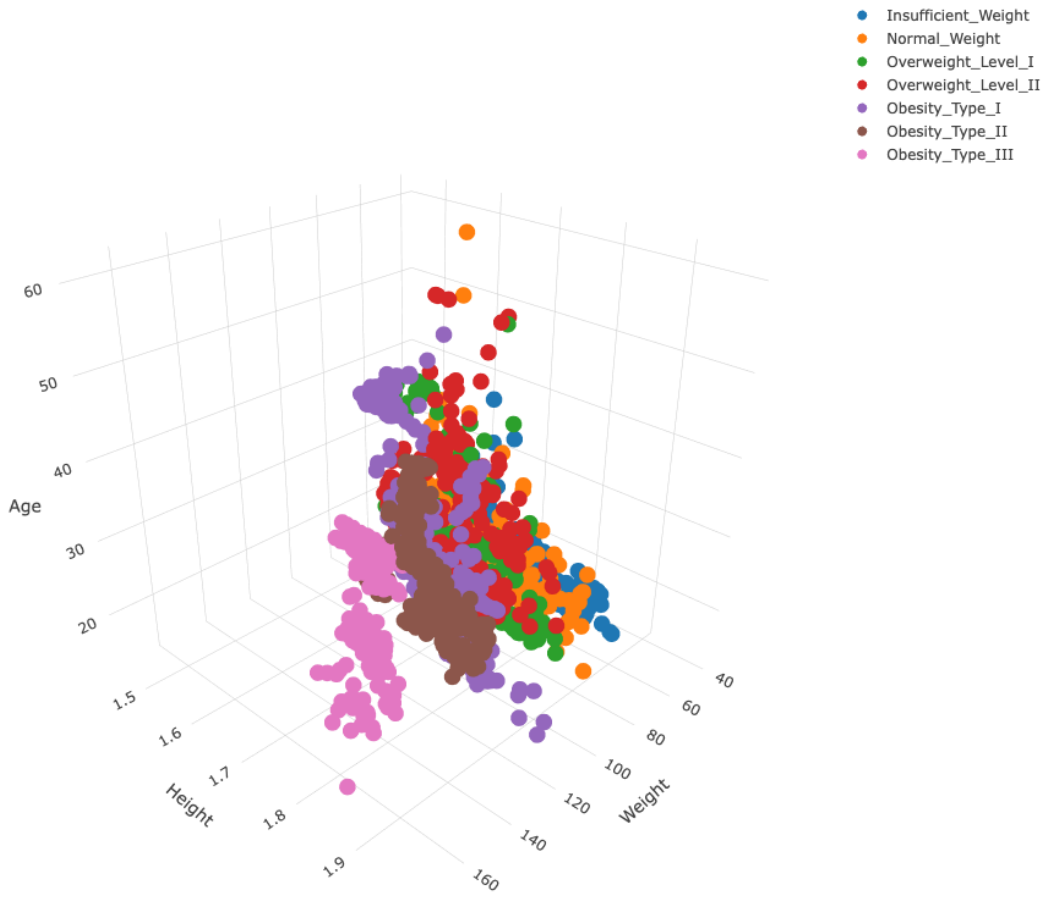
3D plotting on Height, Weight and Age.

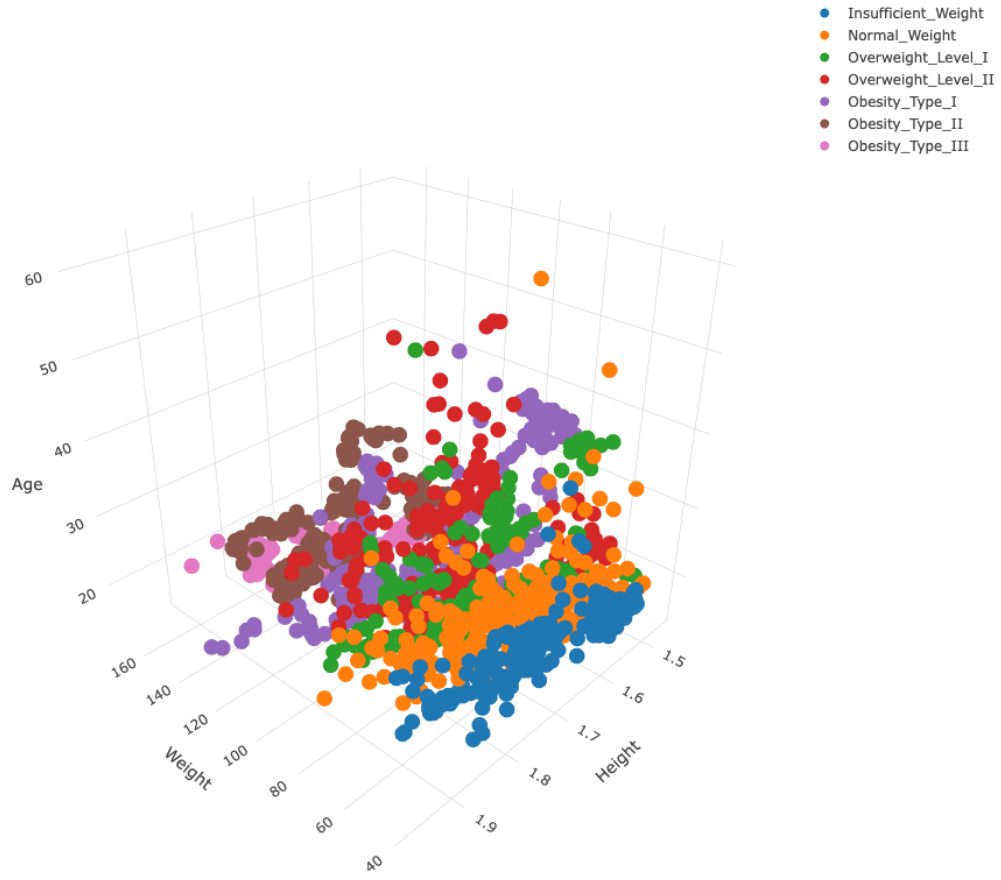
```
#3d Plotting on Height, Weight and Age
fig <- plot_ly(data = od, x = ~od$Age, y = ~od$Height, z = ~od$Weight, type = 'scatter3d', mode = 'markers')
fig
```



3D plotting on Weight, Height, Age and NObeyesdad

```
# Reorder factor levels to reflect increasing severity
od1$NObeyesdad <- factor(od1$NObeyesdad, levels = c(
  "Insufficient_Weight",
  "Normal_Weight",
  "Overweight_Level_I",
  "Overweight_Level_II",
  "Obesity_Type_I",
  "Obesity_Type_II",
  "Obesity_Type_III"
))

fig_mtrans <- plot_ly(od1,
                      x = ~Weight,
                      y = ~Height,
                      z = ~Age,
                      color = ~NObeyesdad,
                      colors = c('#1f77b4', '#ff7f0e', '#2ca02c',
                                 '#d62728', '#9467bd', '#8c564b', '#e377c2'))
fig_mtrans%>% add_markers()
fig_mtrans
```


Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
Age
60
50
40
30
20
Height
1.5
1.6
1.7
1.8
1.9
Weight
40
60
80
100
120
140
160



b. Which pairs of attributes seem to be correlated? How are they correlated?

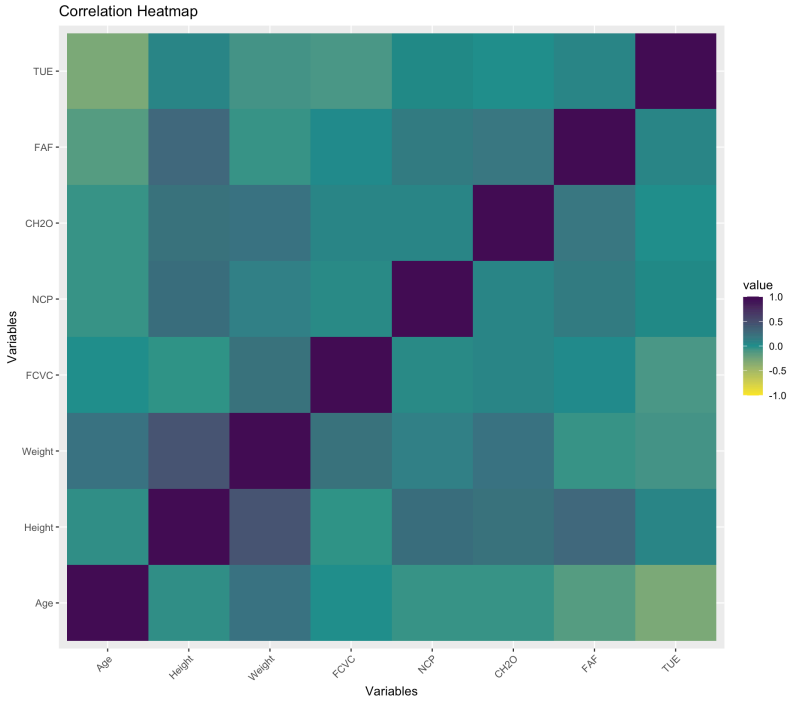
We set up a correlation matrix between numeric attributes.

```
#Correlation between numeric attributes
corrMatrix <- cor(od_numerical)
corrMatrix
```

```
> corrMatrix
                Age      Height      Weight        FCVC         NCP        CH2O         FAF         TUE
Age     1.00000000 -0.02595813  0.20256010  0.01629089 -0.04394373 -0.04530386 -0.14493833 -0.29693059
Height -0.02595813  1.00000000  0.46313612 -0.03812106  0.24367173  0.21337592  0.29470900  0.05191167
Weight  0.20256010  0.46313612  1.00000000  0.21612471  0.10746899  0.20057539 -0.05143627 -0.07156136
FCVC    0.01629089 -0.03812106  0.21612471  1.00000000  0.04221630  0.06846147  0.01993940 -0.10113485
NCP    -0.04394373  0.24367173  0.10746899  0.04221630  1.00000000  0.05708800  0.12950431  0.03632557
CH2O   -0.04530386  0.21337592  0.20057539  0.06846147  0.05708800  1.00000000  0.16723649  0.01196534
FAF    -0.14493833  0.29470900 -0.05143627  0.01993940  0.12950431  0.16723649  1.00000000  0.05856207
TUE    -0.29693059  0.05191167 -0.07156136 -0.10113485  0.03632557  0.01196534  0.05856207  1.00000000
```

After that we created a heatmap of the correlation matrix to have a better understanding of the correlations.

```
# Create a heatmap of the correlation matrix
ggplot(data = melted_corrMatrix, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "#FDE725", mid = "#21908C", high = "#440154", midpoint = 0, limits = c(-1,1)) +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)) +
  labs(title = "Correlation Heatmap", x = "Variables", y = "Variables")
```



We set up a correlation matrix between all the attributes.

```
> corrodMatrix <- cor(od)
> corrodMatrix
                                    Gender         Age      Height      Weight family_history_with_overweight        FAVC        FCVC         NCP        CAEC       SMOKE        CH2O
Gender                        1.000000000  0.04839420  0.61846630  0.161667575                    0.10251213  0.064933774 -0.27450478  0.06759988 -0.09154334  0.044698091  0.10792967
Age                           0.048394197  1.00000000 -0.02595813  0.202560104                    0.20572533  0.063901686  0.01629089 -0.043943727 -0.08373870  0.091987445 -0.04530385
Height                        0.618466297 -0.02595813  1.00000000  0.463136117                    0.24768389  0.178363778 -0.03812106  0.243671726 -0.04881820  0.055499384  0.21337591
Weight                        0.161667575  0.20256010  0.46313612  1.000000000                    0.49682038  0.272300490  0.21612471  0.107468988 -0.28749346  0.025746413  0.20057538
family_history_with_overweight 0.102512133 0.20572533  0.24768389  0.496820377                    1.00000000  0.208035507  0.04037225  0.071369697 -0.16978653  0.017385500  0.14743660
FAVC                          0.064933774  0.06390169  0.17836378  0.272300490                    0.20803551  1.000000000 -0.02728308 -0.006999943 -0.15006763 -0.050659965  0.00971913
FCVC                          -0.274504782 0.01629089 -0.03812106  0.216124705                    0.04037225 -0.027283080  1.00000000  0.042216296  0.05467024  0.014319529  0.06846147
NCP                           0.067599988 -0.04394373  0.24367173  0.107468988                    0.07136970 -0.006999943  0.04221630  1.000000000  0.09780134  0.007811192  0.05708799
CAEC                          -0.091543343 -0.08373870 -0.04881820 -0.287493463                   -0.16978653 -0.150067628  0.05467024  0.097801337  1.00000000  0.055281967 -0.14499514
SMOKE                         0.044698091  0.09198745  0.05549938  0.025746413                    0.01738550 -0.050659965  0.01431953  0.007811192  0.05528197  1.000000000 -0.03199453
CH2O                          0.107929676 -0.04530386  0.21337592  0.200575387                    0.14743661  0.009719131  0.06846147  0.057087996 -0.14499514 -0.031994530  1.00000000
SCC                           -0.102633482 -0.11628285 -0.13375278 -0.201906340                   -0.18542171 -0.190658309  0.07185219 -0.015623955  0.10917895  0.047731227  0.00803648
FAF                           0.189606957 -0.14493833  0.29470900 -0.051436270                    -0.05667320 -0.107995159  0.01993940  0.129504307  0.03011022  0.011216029  0.16723649
TUE                           0.017269473 -0.29693059  0.05191167 -0.071561359                    0.02294330  0.068416912 -0.10113485  0.036325572  0.04856704  0.017613134  0.01196533
CALC                          -0.007615872 0.04448711  0.12973186  0.206676696                    -0.03667591  0.089519515  0.06078109  0.071746765 -0.04753960  0.082471289  0.09138555
MTRANS                        -0.137537298 -0.60194519 -0.07360921  0.004609837                    -0.10153969 -0.069800209  0.06474348 -0.053858095  0.04853492 -0.010701669  0.04402831
NObeyesdad                    -0.031463618  0.28291341  0.13356456  0.913250802                    0.50514842  0.247793200  0.22775883  0.026690439 -0.32934972  0.003442179  0.13300843
                                     SCC         FAF         TUE         CALC       MTRANS   NObeyesdad
Gender                        -0.102633482  0.189606957  0.01726947 -0.007615872 -0.137537298 -0.031463618
Age                           -0.116282847 -0.144938327 -0.29693059  0.044487106 -0.601945191  0.282913410
Height                        -0.133752777  0.294708998  0.05191167  0.129731863 -0.073609212  0.133564558
Weight                        -0.201906340 -0.051436270 -0.07156136  0.206676696  0.004609837  0.913250802
family_history_with_overweight -0.185421707 -0.056673197  0.02294330 -0.036675907 -0.101539690  0.505148418
FAVC                          -0.190658309 -0.107995159  0.06841691  0.089519515 -0.069800209  0.247793200
FCVC                          0.071852192  0.019939398 -0.10113485  0.060781094  0.064743476  0.227758827
NCP                           -0.015623955  0.129504307  0.03632557  0.071746765 -0.053858095  0.026690439
CAEC                          0.109178952  0.030110223  0.04856704 -0.047539605  0.048534917 -0.329349716
SMOKE                         0.047731227  0.011216029  0.01761313  0.082471289 -0.010701669  0.003442179
CH2O                          0.008036485  0.167236492  0.01196534  0.091385557  0.044028311  0.133008436
SCC                           1.000000000  0.074220664 -0.01092798  0.003462876  0.043157450 -0.194507656
FAF                           0.074220664  1.000000000  0.05856207 -0.086798709  0.006393953 -0.199900835
TUE                           -0.010927978  0.058562066  1.00000000 -0.045864068  0.176944749 -0.107991495
CALC                          0.003462876 -0.086798709 -0.04586407  1.000000000  0.012452354  0.151752322
MTRANS                        0.043157450  0.006393953  0.17694475  0.012452354  1.000000000  0.011818148
NObeyesdad                    -0.194507656 -0.199900835 -0.10799149  0.151752322  0.011818148  1.000000000
```

Through these steps we can notice these visually identifiable pairs that are correlated attribute pairs:

1. Height and Weight(0.46)
   There is a positive correlation between height and weight for taller individuals generally weigh more.
2. FAF and Height(0.29)
   The positive correlation between FAF and height shows that taller individuals tend to be slightly more physically active.
3. Age and NObeyesdad(0.28)
   The positive correlation between age and obesity level shows that older individuals tend to have higher obesity levels.
4. CAEC and NObeyesdad(-0.33)
   The negative correlation between CAEC(eating between meals) and obesity level indicates that more frequent snacking is associated with higher obesity.

2. Prepare the Data

a. Investigate some of the statistics of the data set: summary(). Describe(). What do you glean from this data?

```
> summary(od1)
   Gender              Age            Height          Weight      family_history_with_overweight     FAVC               FCVC            NCP            CAEC              SMOKE
 Length:2111        Min.   :14.00   Min.   :1.450   Min.   : 39.00   Length:2111                    Length:2111        Min.   :1.000   Min.   :1.000   Length:2111        Length:2111
 Class :character   1st Qu.:19.95   1st Qu.:1.630   1st Qu.: 65.47   Class :character               Class :character   1st Qu.:2.000   1st Qu.:2.659   Class :character   Class :character
 Mode  :character   Median :22.78   Median :1.700   Median : 83.00   Mode  :character               Mode  :character   Median :2.386   Median :3.000   Mode  :character   Mode  :character
                    Mean   :24.31   Mean   :1.702   Mean   : 86.59                                                    Mean   :2.419   Mean   :2.686
                    3rd Qu.:26.00   3rd Qu.:1.768   3rd Qu.:107.43                                                    3rd Qu.:3.000   3rd Qu.:3.000
                    Max.   :61.00   Max.   :1.980   Max.   :173.00                                                    Max.   :3.000   Max.   :4.000

      CH2O            SCC                FAF              TUE             CALC              MTRANS                     NObeyesdad
 Min.   :1.000   Length:2111        Min.   :0.0000   Min.   :0.0000   Length:2111        Length:2111        Insufficient_Weight:272
 1st Qu.:1.585   Class :character   1st Qu.:0.1245   1st Qu.:0.0000   Class :character   Class :character   Normal_Weight      :287
 Median :2.000   Mode  :character   Median :1.0000   Median :0.6253   Mode  :character   Mode  :character   Overweight_Level_I :290
 Mean   :2.008                      Mean   :1.0103   Mean   :0.6579                                         Overweight_Level_II:290
 3rd Qu.:2.477                      3rd Qu.:1.6667   3rd Qu.:1.0000                                         Obesity_Type_I     :351
 Max.   :3.000                      Max.   :3.0000   Max.   :2.0000                                         Obesity_Type_II    :297
                                                                                                            Obesity_Type_III   :324
```

```
> describe(od1)
                                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  4.11  1.99   4.00    4.14  2.97  1.00   7.00   6.00 -0.08    -1.22 0.04
```

From the investigation on the statistics(using summary() and describe()), we found out that for demographics the age ranges from 14 to 61, with a mean of closing to 24, showing a young-skewed population. The gender appeared to be balanced(with a mean of 1.51 after encoded numerically). The weight of the demographics has a wide range from 39 to 173 kg, with a mean of 86.6 kg. The height is fairly normally distributed with a mean of 1.70 m with minimal skew.

On the side of Eating and Lifestyle Habits, FCVC(Vegetable consumption) is high with a mean of 2.42 on a 1-3 scale, and shows a strong left skew, meaning most people consume vegetables frequently in the set. CAEC and SCC also show skewed distributions, with most individuals reporting less frequent snacking and active calorie tracking. Water intake averages about 2 liters per day which is healthy, FAF averages around 1, suggesting moderate activity. SMOKE shows extremely right skew and high kurtosis, indicating that few people smoke. The obesity levels have 7 levels from 1 to 7. It has a nearly symmetric distribution(mean = 4.1, skew = -0.08), suggesting a fairly balanced sample acrossing from 'insufficient weight' to 'obesity type iii'.

In conclusion, the dataset appears well-distributed and contains a variety of obesity levels, making it suitable for classification.

b. To subset, pick the most correlated attributes to use – they may all be relevant, **so document your rationale for eliminating some attributes**. Use at least 5 attributes, but possibly more.

Based on correlation analysis and domain understanding, we selected 7 attributes that are most predictive of obesity level which are Weight, Age, FAF, FCVC, CAEC, CH2O, family_history_with_overweight. These 7 attributes include body metrics, lifestyle, behavior and age.

Features with high skew, low variability or redundancy are excluded and we eliminate these attributes: SMOKE, SCC, TUE, MTRANS, FAVC.

c. You may need to translate alphanumeric (e.g., character) values into numeric values.
Create a mapping for each field of values to integers.
Make sure you put these mapping tables into your report.

Below is the mapping table recording each field of values to integers.

```
> # Gender
> data.frame(Level = levels(factor(c("Female", "Male"))), MappedValue = 1:2)
   Level MappedValue
1 Female           1
2   Male           2
>
> # Family History
> data.frame(Level = levels(factor(c("no", "yes"))), MappedValue = 1:2)
  Level MappedValue
1    no           1
2   yes           2
>
> # FAVC
> data.frame(Level = levels(factor(c("no", "yes"))), MappedValue = 1:2)
  Level MappedValue
1    no           1
2   yes           2
>
> # CAEC
> data.frame(Level = levels(factor(c("no", "Sometimes", "Frequently", "Always"))), MappedValue = 1:4)
       Level MappedValue
1     Always           1
2 Frequently           2
3         no           3
4  Sometimes           4
>
> # SMOKE
> data.frame(Level = levels(factor(c("no", "yes"))), MappedValue = 1:2)
  Level MappedValue
1    no           1
2   yes           2
>
> # SCC
> data.frame(Level = levels(factor(c("no", "yes"))), MappedValue = 1:2)
  Level MappedValue
1    no           1
2   yes           2
>
> # CALC
> data.frame(Level = levels(factor(c("no", "Sometimes", "Frequently", "Always"))), MappedValue = 1:4)
       Level MappedValue
1     Always           1
2 Frequently           2
3         no           3
4  Sometimes           4
>
> # MTRANS
> data.frame(Level = levels(factor(c("Automobile", "Bike", "Motorbike", "Public_Transportation", "Walking"))), MappedValue = 1:5)
                 Level MappedValue
1           Automobile           1
2                 Bike           2
3            Motorbike           3
4 Public_Transportation           4
5              Walking           5
>
> # NObeyesdad
> data.frame(Level = levels(factor(c("Insufficient_Weight", "Normal_Weight", "Overweight_Level_I",
+                                    "Overweight_Level_II", "Obesity_Type_I", "Obesity_Type_II", "Obesity_Type_III"))),
+            MappedValue = 1:7)
                Level MappedValue
1 Insufficient_Weight           1
2       Normal_Weight           2
3     Obesity_Type_I            3
4    Obesity_Type_II            4
5   Obesity_Type_III            5
6  Overweight_Level_I           6
7 Overweight_Level_II           7
```

d. Split the original data set into Training and Test Sets. Use 70-30%, 60-40%, 50-50%

Dataset od1 has been successfully splitted into training and test sets with proportion 70-30%, 60-40%, and 50-50%.

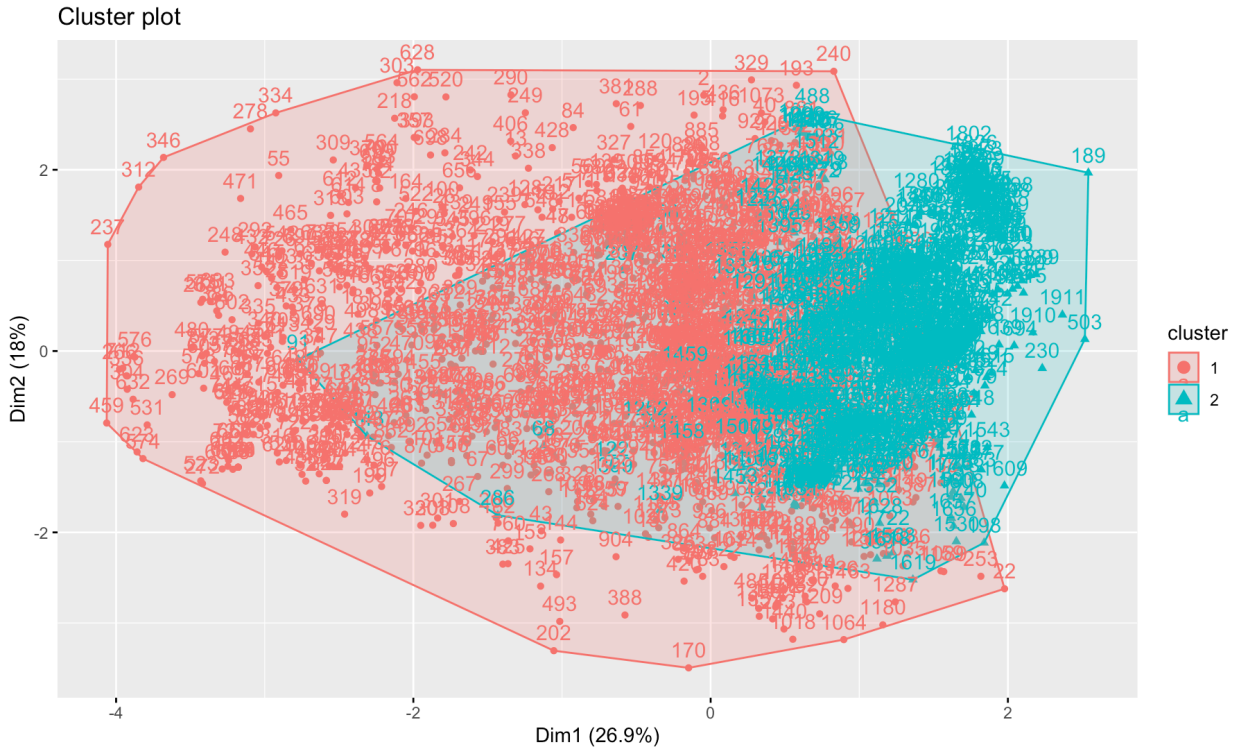
```
> set.seed(123)  # Ensures reproducibility
>
> # 70-30 split
> sample_70 <- sample(nrow(od1), 0.7 * nrow(od1))
> train_70 <- od1[sample_70, ]
> test_30 <- od1[-sample_70, ]
>
> # 60-40 split
> sample_60 <- sample(nrow(od1), 0.6 * nrow(od1))
> train_60 <- od1[sample_60, ]
> test_40 <- od1[-sample_60, ]
>
> # 50-50 split
> sample_50 <- sample(nrow(od1), 0.5 * nrow(od1))
> train_50 <- od1[sample_50, ]
> test_50 <- od1[-sample_50, ]
>
> nrow(train_70); nrow(test_30)
[1] 1477
[1] 634
> nrow(train_60); nrow(test_40)
[1] 1266
[1] 845
> nrow(train_50); nrow(test_50)
[1] 1055
[1] 1056
```

3. Clustering on the Whole Data Set! Do all the functions in the rubric!

a. Use the whole date set when clustering (the first part of the rubric).

We used k values of 2, 3, 5, 7, 9, 11 for the kmeans clustering.

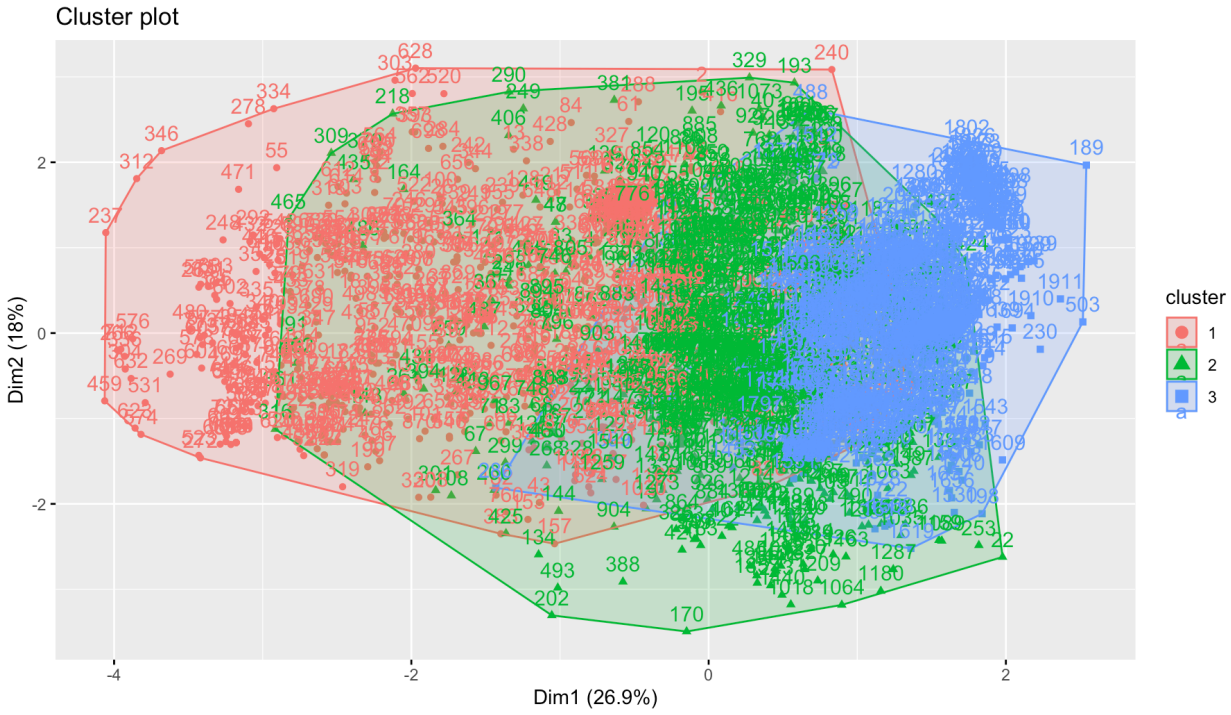
Case k = 2:




```
> kmeans_result
K-means clustering with 2 clusters of sizes 1286, 825

Cluster means:
     Weight      Age       FAF     FCVC     CAEC     CH2O
1  68.84224 23.57320 1.0711102 2.322428 2.216952 1.943098
2 114.24491 25.46517 0.9155039 2.569646 2.021818 2.109198
  family_history_with_overweight
1                       1.704510
2                       1.993939

Clustering vector:
   [1] 1 1 1 1 1 1 1 1 1 2 1 1 2 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
  [39] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 1
  [77] 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 1 1 1 1 2 1 1 1 1 1
 [115] 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1
 [153] 1 1 2 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 1
 [191] 1 1 1 1 1 1 1 2 1 1 1 1 2 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1
 [229] 1 2 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1
 [267] 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1
 [305] 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [343] 2 1 2 1 1 2 1 2 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1
 [381] 1 1 1 2 1 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1 1 1 2 2 1 1 1 1 1 1 1 1 1 2 1 1 1
 [419] 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1
 [457] 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1
 [495] 1 1 1 1 2 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 1 1 1
 [533] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [571] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [609] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [647] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [685] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [723] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [761] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [799] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [837] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [875] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [913] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [951] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1
 [989] 2 1 1 1 1 1 1 1 1 1 1
 [ reached getOption("max.print") -- omitted 1111 entries ]
```

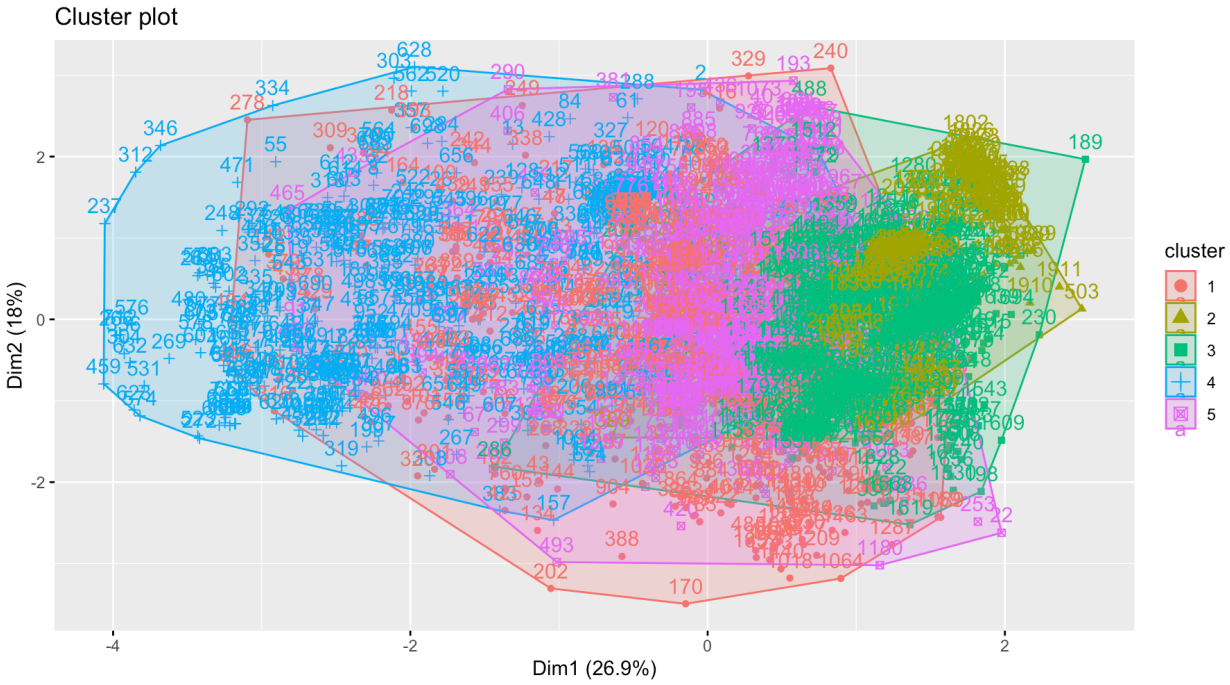
```
Within cluster sum of squares by cluster:
[1] 325108.0 173035.4
 (between_SS / total_SS =  67.6 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

Case k = 3:



```
> kmeans_result
K-means clustering with 3 clusters of sizes 655, 749, 707

Cluster means:
     Weight      Age       FAF     FCVC     CAEC     CH2O
1  56.78896 20.92175 1.1067993 2.357123 2.309924 1.898788
2  83.79290 26.23433 1.0775596 2.264029 2.113485 2.000130
3 117.15065 25.41816 0.8496362 2.640632 2.012730 2.117551
  family_history_with_overweight
1                       1.509924
2                       1.917223
3                       1.997171

Clustering vector:
   [1] 1 1 2 2 2 1 1 1 1 1 3 2 1 2 1 1 3 2 2 1 2 2 1 2 1 1 1 1 2 1 2 1 2 2 1 1 1 1
  [39] 1 2 2 1 1 1 2 1 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 3 1 2 1 1 1 2 1
  [77] 1 2 2 1 1 2 2 1 2 1 2 2 1 2 2 1 2 1 1 2 1 1 1 1 1 1 1 1 2 1 1 2 2 2 2 2 1 1 1
 [115] 1 1 2 2 1 2 2 2 1 1 2 2 1 1 1 2 1 1 1 2 3 2 2 2 2 1 1 1 2 2 1 1 1 2 1 2 2 1
 [153] 1 2 2 1 1 1 1 2 2 2 1 2 2 3 2 2 2 2 2 1 1 1 1 1 1 2 2 1 1 1 1 1 1 2 2 3 1
 [191] 1 2 2 1 2 2 2 3 1 1 2 2 3 2 2 2 2 2 1 1 3 1 1 1 1 1 1 2 1 1 2 2 1 1 2 3 1 1
 [229] 2 3 2 1 1 2 1 1 1 1 1 1 1 1 2 1 1 1 1 2 2 2 2 2 1 1 2 2 3 2 2 2 2 1 1 1 1
 [267] 1 2 1 1 3 1 1 2 1 1 1 1 1 1 1 1 1 1 3 1 1 1 2 1 1 1 2 3 1 1 2 2 2 2 1 1 1
 [305] 2 3 1 1 2 1 1 1 1 1 1 2 2 1 1 1 1 2 1 2 2 1 1 2 2 2 1 1 2 1 1 1 1 1 1 1 2 1
 [343] 3 1 3 1 1 2 2 3 1 1 1 1 1 1 1 1 3 1 2 2 2 2 2 1 2 1 2 2 1 1 1 1 1 2 1 1 1 1
 [381] 2 2 1 2 1 1 1 2 3 1 1 1 1 2 1 1 1 1 3 1 1 2 2 3 2 2 1 1 2 1 1 1 2 2 3 1 2 2
 [419] 2 2 1 1 1 2 2 1 1 1 1 2 2 1 1 1 2 2 1 1 1 1 1 1 1 1 1 2 2 3 1 1 1 1 1 1 1 1
 [457] 1 1 1 2 1 2 2 1 2 1 2 2 1 1 1 2 1 1 1 1 2 1 1 1 1 1 1 1 2 2 2 3 2 1 2 1 2 1
 [495] 1 1 1 1 1 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [533] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [571] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [609] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [647] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [685] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [723] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 2 1 2 2 2 2 2 2 2 1 1 1 1
 [761] 1 2 1 1 1 1 2 2 2 2 1 2 2 1 2 2 2 2 2 2 2 1 2 2 1 1 1 2 2 1 1 1 1 1 1 2 2 2
 [799] 2 2 1 2 1 2 2 1 1 2 2 1 1 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 2 2 1 1 1 1 1 2
 [837] 2 2 2 2 2 2 2 2 1 1 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 1 2 2 2 1 1 1 1 1 2 2 2 1
 [875] 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 1 2 1 2 2 1 1 2 2 2 1 2 2 2 2 2 2 2 2 2 2 1
 [913] 1 1 1 1 1 1 2 2 1 1 1 1 2 2 2 2 2 2 2 1 1 2 2 2 1 1 2 2 2 2 2 2 2 2 2 1 2 2
 [951] 2 1 1 1 1 2 2 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 1 2 2 1 2 2 2 2 3 2 2 2 2 1 2
 [989] 2 1 1 2 1 1 2 2 2 2 2 2
 [ reached getOption("max.print") -- omitted 1111 entries ]


 Within cluster sum of squares by cluster:
 [1]  63116.57  88015.11 125696.58
  (between_SS / total_SS =  82.0 %)

 Available components:

 [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
 [6] "betweenss"    "size"         "iter"         "ifault"
```

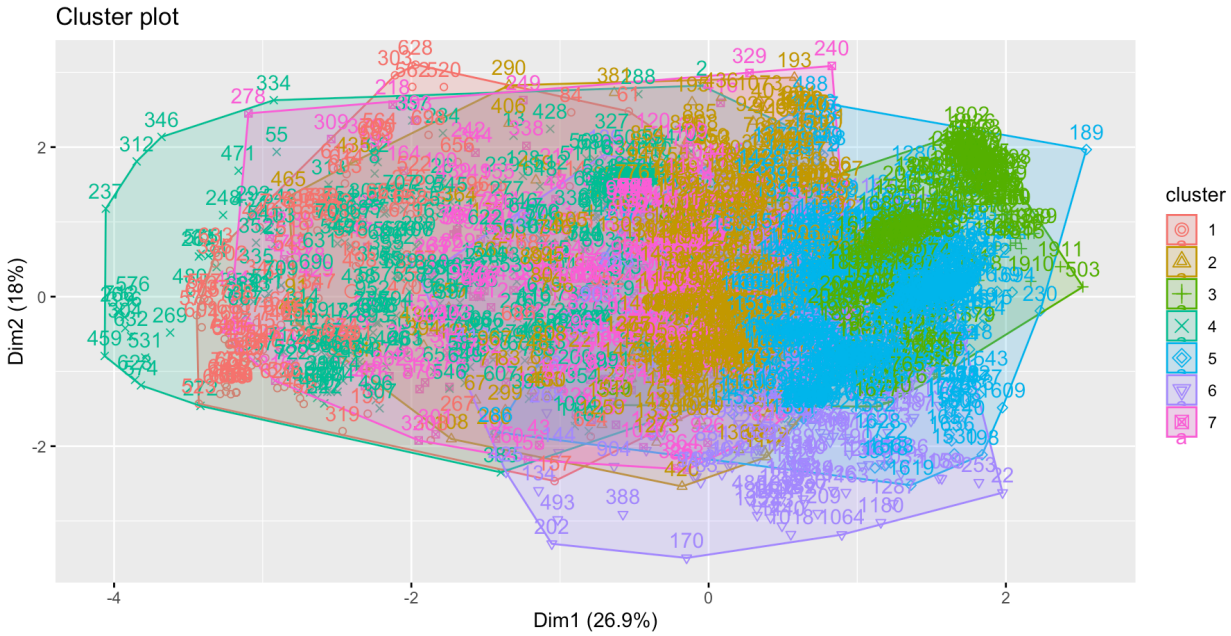
Case k = 5:




```
> kmeans_result
K-means clustering with 5 clusters of sizes 470, 166, 576, 435, 464

Cluster means:
     Weight      Age       FAF     FCVC     CAEC     CH2O family_history_with_overweight
1  71.83835 26.56350 0.9237007 2.329745 2.108511 1.956217                        1.770213
2 135.56135 22.07817 1.3811557 2.900642 2.030120 2.112241                        2.000000
3 110.77274 26.62796 0.7258409 2.556930 2.010417 2.098681                        1.994792
4  51.84119 20.11495 1.1827468 2.394158 2.439080 1.811285                        1.457471
5  86.55161 23.89303 1.1567846 2.189360 2.094828 2.095063                        1.918103

Clustering vector:
   [1] 1 4 1 5 5 4 4 4 1 1 3 5 4 3 4 1 3 1 5 1 5 5 4 5 1 4 1 4 1 1 5 1 1 5 4 1 4 4 1 5 1 4 1
  [44] 4 5 1 1 1 4 4 4 4 4 4 4 4 1 1 1 1 4 4 4 4 4 4 1 5 5 3 4 5 4 4 1 5 4 4 5 5 4 4 5 5 4 5 4
  [87] 5 5 1 1 5 1 1 4 4 1 4 4 4 1 1 1 4 4 1 4 4 5 5 5 1 4 4 4 4 4 5 5 4 1 1 5 4 1 5 5 1 4 1
 [130] 1 4 4 4 1 3 1 1 1 1 1 1 5 1 1 1 4 1 1 5 1 4 1 5 5 1 4 4 1 1 1 1 4 1 5 2 1 5 5 1 5 4
 [173] 4 1 1 1 4 5 5 1 4 1 4 4 1 1 5 5 3 1 4 5 5 1 5 5 1 3 4 4 5 1 3 5 1 1 3 1 1 1 3 4 1 4 4
 [216] 1 1 1 4 1 1 5 1 1 5 3 4 4 1 3 5 1 4 1 4 1 4 4 4 1 1 1 1 3 4 4 4 4 1 1 1 1 5 4 4 5 5 3
 [259] 5 5 1 1 1 4 1 4 4 1 4 1 3 4 4 1 4 4 4 1 4 4 4 4 4 4 1 3 4 4 4 5 1 4 1 1 3 4 4 5 5 1 1
 [302] 4 4 4 5 3 4 4 1 4 4 4 4 4 1 1 1 1 4 1 1 4 4 1 1 5 4 4 1 4 4 4 1 1 1 4 1 4 3 3 1
 [345] 2 4 4 5 5 2 1 4 4 4 4 1 4 4 3 4 1 1 1 5 5 1 5 1 5 1 4 1 4 4 1 5 1 1 1 1 5 1 4 5 4 4 4
 [388] 1 3 1 4 4 1 5 4 4 1 4 2 4 1 5 3 3 1 5 1 4 5 4 1 1 5 5 3 1 5 5 5 5 4 4 4 5 1 4 1 4 1 5
 [431] 5 4 1 4 5 5 4 4 1 4 4 4 1 4 4 4 5 5 3 4 1 1 1 4 1 1 1 4 4 5 4 3 5 1 5 4 1 1 4 4 4 5 1
 [474] 4 1 1 1 4 1 4 4 1 4 4 1 1 1 3 5 4 5 1 5 1 1 4 4 4 3 2 3 2 2 2 3 3 2 4 4 4 4 4 4 4 4 4
 [517] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 1 1 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
 [560] 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
 [603] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 1 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
 [646] 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
 [689] 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
 [732] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 5 1 5 1 1 5 5 5 1 5 1 1 1 1 1 1 1 1 1 5 5 5 1 1 1 5 1
 [775] 1 5 1 1 5 1 5 1 1 1 4 1 1 1 5 1 1 1 1 4 1 5 1 5 5 5 1 1 1 5 5 1 1 5 1 1 1 1 5 5 1 5 5 1
 [818] 1 5 1 1 1 1 1 1 1 1 1 1 5 1 1 1 1 1 1 1 5 5 5 5 5 5 1 1 1 1 1 5 5 1 1 1 5 1 1 1 1 5 5 1
 [861] 5 1 1 1 1 4 4 1 1 1 1 5 5 1 1 1 1 1 1 4 4 1 5 1 5 5 5 5 5 5 1 1 4 5 5 1 1 5 1 1 1 5 5
 [904] 1 5 5 1 5 1 1 1 1 1 1 1 1 1 1 5 1 1 1 1 1 1 1 5 5 5 1 1 1 1 1 5 1 1 1 5 5 1 1 1 5 5 1
 [947] 5 1 1 1 1 4 4 1 1 1 5 1 1 1 1 1 4 1 1 1 5 1 1 5 5 5 5 1 1 1 1 1 5 5 5 5 3 5 5 5 5 1 5 5
 [990] 1 4 5 1 1 1 5 5 5 5 5 5
 [ reached getOption("max.print") -- omitted 1111 entries ]

Within cluster sum of squares by cluster:
[1] 50653.17 15934.05 40303.52 21596.46 30733.05
 (between_SS / total_SS =  89.6 %)
```

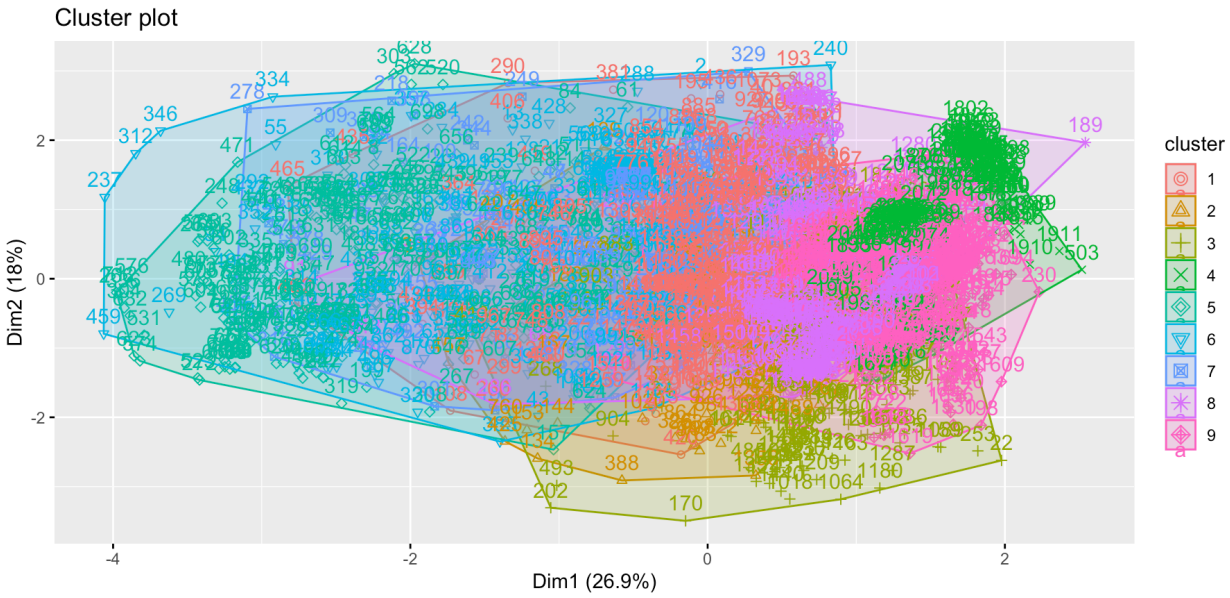
```
Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

Case k = 7:



```
> kmeans_result
K-means clustering with 7 clusters of sizes 122, 423, 218, 306, 548, 178, 316

Cluster means:
     Weight      Age       FAF     FCVC     CAEC     CH2O
1  44.15995 21.09103 1.0114101 2.611220 2.565574 1.756248
2  85.22573 22.21363 1.0990733 2.178975 2.094563 2.082738
3 132.20728 22.50609 1.3198027 2.857205 2.022936 2.130764
4  54.66710 19.70613 1.2574449 2.313365 2.395425 1.828927
5 109.19981 26.76374 0.7322587 2.524239 2.010949 2.099539
6  79.75396 38.02578 0.8135093 2.387369 2.101124 1.858006
7  68.85492 22.09780 1.0312052 2.301676 2.120253 2.019687
  family_history_with_overweight
1                       1.188525
2                       1.926714
3                       2.000000
4                       1.562092
5                       1.994526
6                       1.893258
7                       1.686709

Clustering vector:
   [1] 7 4 7 2 2 4 4 4 7 7 5 2 4 5 4 7 5 6 6 7 2 6 4 2 7 4 7 4 7 7 2 7 7 6 7 7 1 1
  [39] 7 2 7 4 7 7 2 7 7 7 4 4 4 1 4 4 4 7 7 7 7 4 1 4 7 4 4 7 2 2 5 4 2 4 4 7 2 1
  [77] 1 2 2 4 4 2 2 1 2 4 2 2 7 6 2 7 6 4 1 7 4 1 1 7 7 7 4 4 6 4 4 2 2 2 7 4 4 1
 [115] 4 4 2 2 4 7 7 2 4 7 2 2 7 4 7 2 4 4 4 6 5 7 7 6 6 7 7 7 2 6 7 7 4 6 7 2 2 4
 [153] 7 6 2 7 1 4 6 2 7 6 4 7 6 3 7 2 2 6 2 4 4 7 7 7 1 2 2 7 1 7 4 4 7 7 6 5 5 7
 [191] 4 2 2 7 2 2 7 5 1 4 2 6 5 2 7 2 5 7 7 7 5 4 7 7 4 7 7 7 1 7 2 2 7 7 2 3 4 4
 [229] 6 5 2 7 1 6 1 7 4 4 4 7 7 7 7 5 1 1 4 4 7 2 7 2 6 1 4 2 2 5 2 2 7 7 7 4 7 4
 [267] 1 6 4 7 5 4 4 7 4 4 4 7 4 4 7 4 4 4 7 5 4 4 4 2 7 4 7 7 5 4 4 2 2 7 7 4 1 4
 [305] 2 5 4 1 7 4 4 4 4 4 7 7 6 7 1 7 7 6 4 7 7 4 4 7 7 2 4 4 7 4 4 4 7 7 7 1 7 4
 [343] 5 7 3 4 1 2 2 3 7 4 4 4 4 7 4 4 5 1 7 6 6 2 2 7 6 6 6 7 4 7 4 4 7 6 7 7 7 7
 [381] 2 7 4 2 4 1 4 6 5 7 4 4 7 2 4 1 7 4 3 4 7 2 5 5 6 2 7 4 2 4 7 7 6 6 5 7 6 2
 [419] 2 2 4 4 4 2 6 4 7 4 7 2 2 4 7 4 2 2 4 4 7 4 4 4 7 4 4 1 2 2 5 7 7 7 7 4 7 7
 [457] 7 1 4 2 4 5 2 7 2 4 7 7 1 1 4 2 7 4 7 7 7 4 7 4 4 7 4 4 6 7 7 5 2 4 2 7 6 7
 [495] 7 4 4 1 5 3 5 3 3 3 5 5 3 4 4 4 4 4 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 4 4 4 7
 [533] 7 4 4 4 4 1 1 1 4 4 1 4 4 4 4 4 4 1 1 1 4 4 4 4 4 4 4 4 4 1 1 1 4 4 4 4 4 4
 [571] 4 4 4 4 4 4 1 1 1 4 4 4 4 4 4 4 4 4 1 1 1 4 4 4 4 4 4 4 4 4 1 1 1 1 1 1 4 4
 [609] 4 4 4 1 1 1 1 1 4 7 4 1 1 4 4 1 4 4 4 1 4 4 4 4 1 4 4 4 1 4 4 4 1 1 4 4 4 4
 [647] 4 4 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 4 4 4 4 4 4 4 4 4 4 1 1 1 4 1 1 4 4 4 4 4
 [685] 4 1 1 4 4 4 4 4 4 4 4 4 4 1 1 1 4 4 4 4 4 4 4 4 4 4 4 4 1 1 1 4 4 4 4 4 4 4
 [723] 4 4 1 1 1 4 4 4 4 4 4 4 4 4 1 1 1 1 1 1 4 4 4 2 7 2 7 7 2 6 2 7 2 7 6 7 7 7
 [761] 7 2 7 7 7 7 2 2 2 7 7 7 2 7 2 2 7 7 2 7 2 7 6 7 4 7 7 6 2 7 7 7 7 4 7 2 6 2
 [799] 2 2 7 7 7 2 2 7 7 2 7 7 7 2 2 6 2 2 7 7 2 7 6 6 7 7 7 7 6 6 2 2 7 7 7 7 7 7
 [837] 7 2 2 2 2 2 2 7 7 7 2 2 2 2 7 2 2 2 7 7 7 2 2 7 2 7 7 7 7 4 4 7 7 7 6 2 2 6
 [875] 7 7 7 7 7 4 4 7 6 6 2 2 2 2 2 2 7 7 4 2 2 7 7 2 7 7 7 2 6 6 2 2 7 2 7 7 6 7
 [913] 7 7 7 6 6 7 2 2 7 7 7 7 7 7 2 2 2 7 7 7 7 7 2 7 7 2 2 2 7 7 7 2 2 7 2 7 6 7
 [951] 7 4 4 7 7 6 2 6 7 7 7 7 4 7 7 7 2 6 6 2 2 2 2 7 7 7 7 2 2 2 2 5 6 2 2 2 7 6
 [989] 6 7 4 2 7 7 2 2 2 2 2 2
 [ reached getOption("max.print") -- omitted 1111 entries ]

Within cluster sum of squares by cluster:
[1]  3712.674 17157.686 24241.203  7266.490 36790.542 12189.613 11951.973
 (between_SS / total_SS =  92.6 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

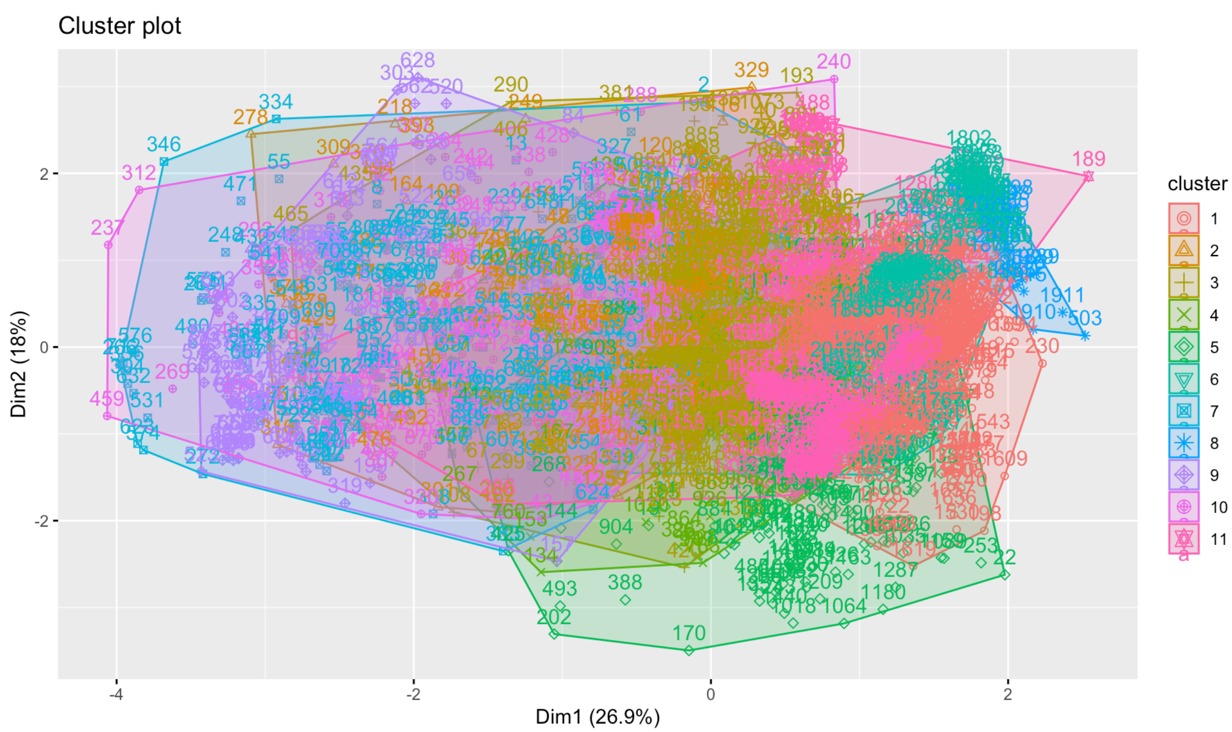
Case k = 9:




```
> kmeans_result
K-means clustering with 9 clusters of sizes 369, 79, 140, 151, 264, 240, 210, 307, 351

Cluster means:
     Weight      Age       FAF     FCVC     CAEC     CH2O
1  83.74737 22.01703 1.0907027 2.216672 2.094851 2.066970
2  70.20916 34.63825 0.8579478 2.434044 2.177215 1.724456
3  83.36474 38.64422 0.8552762 2.306776 2.064286 1.939838
4 136.57521 21.62098 1.4323386 2.928513 2.006623 2.159938
5  47.83422 20.12521 1.0539551 2.386019 2.515152 1.713230
6  59.87953 20.27319 1.2218963 2.395551 2.295833 1.924228
7  71.02897 20.91590 1.1060828 2.284616 2.071429 2.129084
8 102.25274 25.12564 0.9141093 2.510599 2.032573 1.932883
9 116.04829 27.07608 0.6896332 2.495271 2.017094 2.243960
  family_history_with_overweight
1                        1.918699
2                        1.886076
3                        1.900000
4                        2.000000
5                        1.329545
6                        1.654167
7                        1.676190
8                        1.990228
9                        1.997151

Clustering vector:
   [1] 6 6 7 1 1 5 6 5 6 7 8 1 6 3 6 7 8 1 1 7 1 3 6 1 7 5 6 5 7 7 1 2 7 3 6 6 5 5
  [39] 7 1 7 6 6 6 1 6 7 7 6 5 6 5 5 6 6 6 7 6 7 6 5 6 6 6 6 7 1 8 9 6 1 5 6 6 1 5
  [77] 5 1 1 6 6 1 1 5 1 6 1 1 6 1 8 6 3 6 5 7 6 5 5 7 6 7 6 6 3 6 6 1 8 1 7 6 5 5
 [115] 6 6 1 1 5 7 7 8 5 7 1 1 7 6 7 1 5 6 6 2 9 7 7 3 2 7 6 7 8 2 7 7 5 3 6 3 1 6
 [153] 2 3 8 2 5 6 2 1 7 3 6 7 3 4 2 1 1 3 1 6 6 6 7 2 5 1 1 7 5 7 5 6 7 2 3 8 8 2
 [191] 6 1 1 7 1 1 7 9 5 6 1 3 8 1 7 1 8 2 6 6 9 5 6 6 5 7 6 7 5 7 1 1 7 6 1 9 5 6
 [229] 3 9 1 7 2 2 5 7 6 5 5 6 6 7 6 8 5 5 6 5 7 1 7 1 3 5 6 1 1 9 1 1 7 2 7 5 7 5
 [267] 5 3 6 7 8 5 6 7 6 5 5 7 6 5 6 6 6 6 7 8 6 6 5 1 7 6 6 7 8 6 5 1 1 7 7 6 5 5
 [305] 1 9 6 5 7 6 6 6 6 6 6 7 2 7 5 6 7 2 6 7 2 6 6 7 7 1 5 6 7 6 5 5 6 6 6 5 7 5
 [343] 8 7 4 6 5 8 1 9 7 6 5 5 6 6 6 6 9 5 7 3 2 1 1 7 3 2 3 7 6 7 6 6 7 3 2 7 7 7
 [381] 1 7 6 8 6 5 5 2 8 6 5 5 7 1 6 5 2 6 9 6 7 1 8 9 2 1 2 6 1 6 7 6 3 3 9 7 3 1
 [419] 1 1 6 6 6 3 2 5 7 6 7 1 1 6 7 6 1 1 6 6 6 6 6 6 5 6 5 6 5 1 1 8 6 7 6 7 5 7 7
 [457] 2 5 6 1 6 8 1 7 1 5 7 7 5 5 5 1 7 5 6 7 7 6 6 5 5 6 5 6 3 7 7 8 1 5 1 7 3 6
 [495] 7 6 5 5 8 4 9 4 4 4 9 8 4 6 6 6 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 6
 [533] 6 6 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 6 6 6 5 5 5 6 6 6 6 6 6
 [571] 5 5 5 5 5 5 5 5 5 6 6 6 5 5 5 5 5 5 5 5 5 6 6 6 6 6 6 5 5 5 5 5 5 5 5 5 6
```

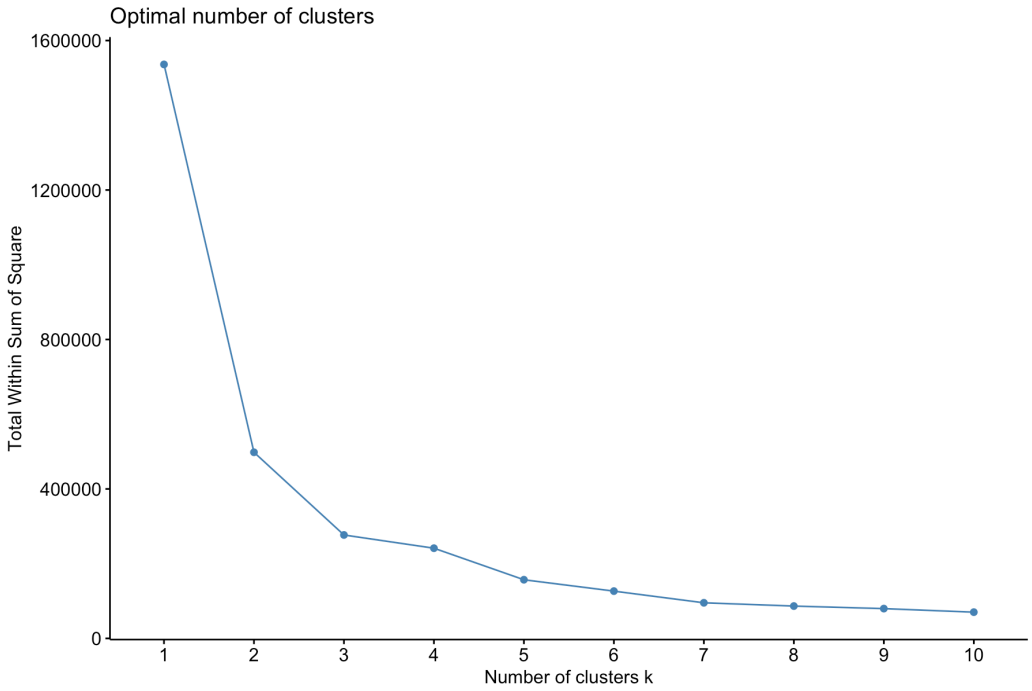
```
 [609] 5 6 5 5 5 5 5 5 5 6 5 5 5 5 5 5 5 5 5 6 5 6 6 5 5 5 6 5 5 5 6 6 5 5 5 5 6 6 6
 [647] 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 6 6 6 5 5 5 5 5 5 5 5 5 5 6 5 5 5
 [685] 5 5 5 5 5 5 5 5 5 5 6 6 6 5 5 5 6 6 6 6 6 6 5 5 5 5 5 5 5 5 5 6 6 6 5 5 5 5
 [723] 5 5 5 5 5 6 6 6 6 6 6 5 5 5 5 5 5 5 5 5 5 5 5 1 7 1 7 7 1 3 1 7 1 7 2 7 7 2
 [761] 2 1 7 7 7 2 1 1 1 7 7 7 1 7 1 1 7 2 1 7 1 6 2 7 6 6 6 2 1 2 7 7 7 6 6 1 2 1
 [799] 1 1 7 7 6 1 1 7 7 1 7 7 7 1 1 3 1 1 7 7 1 7 2 2 7 7 7 2 2 2 1 1 7 7 7 7 7 2
 [837] 2 1 1 1 1 1 1 7 7 7 1 1 1 1 7 1 1 1 7 7 7 1 1 7 1 6 2 2 7 6 6 6 6 6 2 1 1 2
 [875] 7 7 6 7 7 6 6 6 3 2 1 1 1 1 1 1 1 7 7 6 1 1 7 7 1 7 7 7 1 3 3 1 1 7 1 7 7 2 7
 [913] 7 7 7 2 2 2 1 1 7 7 7 7 2 2 1 1 1 7 7 7 7 7 1 7 7 1 1 1 7 7 2 1 1 7 1 7 2 7
 [951] 7 5 6 6 6 2 1 2 7 7 7 7 6 6 6 6 1 2 2 1 1 1 1 7 7 7 6 1 1 1 1 8 3 1 1 1 2 3
 [989] 3 6 6 1 7 7 1 1 1 1 1 1 1
 [ reached getOption("max.print") -- omitted 1111 entries ]

Within cluster sum of squares by cluster:
[1] 10205.727  5064.451  8104.030 13419.176  7151.176  5399.388  3674.201
[8] 12309.109 14884.522
 (between_SS / total_SS =  94.8 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

Case k = 11:



```
> kmeans_result
K-means clustering with 11 clusters of sizes 343, 195, 366, 56, 165, 134, 219, 26, 101, 195, 311

Cluster means:
      Weight      Age       FAF     FCVC     CAEC     CH2O
1  115.77077 27.13424 0.6893914 2.487764 2.005831 2.259612
2   71.44717 20.91599 1.1035114 2.262628 2.041026 2.131343
3   83.74481 21.97245 1.0996522 2.215679 2.092896 2.070861
4   66.08613 34.48497 0.9001337 2.446845 2.196429 1.873542
5   81.81380 37.90705 0.8079847 2.347290 2.078788 1.851395
6  132.51412 22.13471 1.3979771 2.908526 2.029851 2.091661
7   51.95107 19.90935 1.2099456 2.247146 2.447489 1.807616
8  153.77467 20.18516 1.4007776 3.000000 2.038462 2.292863
9   43.15151 20.38647 0.9283811 2.643272 2.603960 1.722619
10  61.72167 20.20206 1.1683396 2.422064 2.282051 1.884474
11 102.13680 25.27669 0.9150316 2.500577 2.032154 1.937935
   family_history_with_overweight
1                        1.997085
2                        1.687179
3                        1.918033
4                        1.857143
5                        1.909091
6                        2.000000
7                        1.525114
8                        2.000000
9                        1.118812
10                       1.625641
11                       1.987138

Clustering vector:
   [1] 10  7  2  3  3  7  7  7 10  2 11  3  7 11 10 10 11  3  5  2  3  5 10  3  2
  [26]  7 10  7  2  2  3  4  2  5 10 10  7  7  2  3  2 10 10 10  3 10  2  2 10  7
  [51]  7  9  7  7  7 10  2 10  2  7  7 10 10  7 10  2  3 11  1 10  3  7 10 10  3
  [76]  9  9  3  3 10 10  3  3  9  3 10  3  3 10  3 11 10  5 10  7  2 10  9  9  2
 [101] 10 10 10 10  5  7  7  3 11  3  2 10  7  7 10 10  3  3  7  2  2 11  7  2  3
 [126]  3  2 10  2  3  7 10  7  4  1  2  2  5  4  2 10 10 11  5  2  2  7  5 10  5
 [151]  3  4  4  5 11  4  9 10  4  3  2  5 10  2  5  6  4  3  3  5  3 10 10 10  2
 [176]  4  7  3  3  2  7  2  7 10  2  4  5 11 11  4  7  3  3  2  3  3  2  1  9 10
 [201]  3  5 11  3  2  3 11  4 10 10  1  7 10 10  7  2 10  2  9  2  3  3 10 10  3
 [226]  1  7 10  5  1  3 10  4  5  7  2 10  7  7 10 10 10 10 11  7  7 10  7  2  3
 [251]  2  3  5  9  7  3  3  1  3  3  2  2  2  7  2  7  4  5 10 10 11  7 10  2 10
 [276]  7  7  2  4  7 10  7  7 10  2 11 10 10  7  3  2 10 10  2 11 10  7  3  3  2
 [301]  2  7  9  7  3  1 10  7  2 10 10 10 10  7 10  2  5  2  9 10  2  5  7  2  4
 [326] 10  7  2  2  3  7  7  2  7  7  7 10 10 10  9  2  7 11 10  8  7  9 11  3  6
 [351]  2 10  7  7 10 10 10  7  1  9  2  5  4  3  3  2  5  4  5  2 10  2 10  7  2
 [376]  5  4  2  2  2  3  2  7 11 10  9  7  5 11 10  7  7  2  3  7  9  4 10  6 10
 [401]  2  3 11  1  4  3  4 10  3  4 10 10  5  5  1  2  5  3  3  3 10 10 10 11  5
 [426]  7  2 10  2  3  3  7  2 10  3  3  7  7 10 10  7  7 10  7 10  9  3  3 11 10
 [451]  2 10  2  7  2 10  4  9 10  3  7 11  3  2  3  7  2  2  9  9  7  3  2  7 10
 [476]  2  2  7 10  7  7 10  7 10  5  2  2 11  3  7  3  2  5 10 10 10  7  9 11  6
 [501]  1  6  8  6  1 11  6  7  7  7  7  7  9  9  9  9  9  9  9  9  9  9  9  9
 [526]  9  9  9  7  7  7 10 10 10  7  7  7  9  9  9  7  7  7  7  7  7  7  7  7
 [551]  9  7  7  7  7  7  7  7 10 10 10  9  9  9 10 10  7  7  7  7  7  7  7  7
 [576]  7  9  9  9 10 10 10  7  7  7  7  7  7  9  9  9  7 10 10 10 10 10  7  7  7
 [601]  9  9  9  9  9  9  7  7  7  7  7  9  9  9  9  9  7 10  7  9  7  7  7  7  7
 [626]  7 10  9  7  7  7  7  9 10  7  7  9 10 10  7  9  9  7  7  7  7  7  7  7  9
 [651]  9  9  9  9  9  9  9  9  9  9  9  9  9  9  7  7  7 10 10 10  7  7  7  9  9
 [676]  9  7  7  7  7  7  7  7  7  7  7  7  9  7  7  7  7  7  7  7 10 10 10  9  9  9
 [701] 10 10  7  7  7  7  7  7  7  7  7  7  7  9  9  9 10 10 10  7  7  7  7  7  7  9
 [726]  9  9  7 10 10 10 10 10  7  7  7  9  9  9  9  9  7  7  7  3  2  3  2  2
 [751]  3  5  3  2  3  2  5  2  2  4  4  3  2  2  2  4  3  3  3  2  2  2  3 10  3
 [776]  3  2  4  3  2  3 10  4  2  7 10 10  4  3  4  2  2  2 10 10  3  4  3  3  3
 [801]  2  2 10  3  3  2  2  3  2  2  2  3  3  5  3  3  2  2  3  2  5  5  2  2  2
 [826]  4  4  4  3  3  2  2  2  2  2  4  4  3  3  3  3  3  2  2  2  3  3  3  3
 [851]  2  3  3  3  2  2  2  3  3  2  3 10  4  4  2  7  7 10 10 10  4  3  3  4  2
 [876]  2 10  2  2 10 10 10  5  5  3  3  3  3  3  2  2 10  3  3  2  2  3  2  2
 [901]  2  3  5  5  3  3  2  3  2  2  5  2  2  2  2  4  4  4  3  3  2  2  2  2  4
 [926]  4  3  3  3  2  2  2  2  2  3  2 10  3  3  3  2  2  4  3  3  2  3 10  4  2
 [951]  2  7  7 10 10  4  3  4  2 10  2  2 10 10 10 10  3  4  4  3  3  3  3  2  2
 [976]  2 10  3  3  3  3 11  5  5  3  3  4  5  5 10 10  3  2  2  3  3  3  3  3  3
 [ reached getOption("max.print") -- omitted 1111 entries ]

Within cluster sum of squares by cluster:
 [1] 13738.3915  3173.3419 10008.8446  3971.0907  8921.3209  4006.8775  3166.8148
 [8]   869.8336  1432.6799  3314.8605 13217.5284
 (between_SS / total_SS =  95.7 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

After processing the kmeans clustering, we used Elbow method to find optimal number of clusters, and the graph was given below.



Based on the graph, the ideal selection for k would be 5, and below it shows cluster summary.

```
> kmeans_result <- kmeans(Clustering_DS[, -ncol(Clustering_DS)], centers = 5, iter.max = 20)
> Clustering_DS$Cluster <- kmeans_result$cluster
> cluster_summary <- aggregate(. ~ Cluster, data = Clustering_DS, mean)
> print(cluster_summary)
  Cluster    Weight      Age       FAF     FCVC     CAEC     CH2O family_history_with_overweight
1       1  54.11670 20.54970 1.1354803 2.396157 2.411654 1.821834                        1.494361
2       2  80.08771 36.89550 0.8029410 2.329432 2.083721 1.854996                        1.897674
3       3  79.37518 21.50617 1.1056309 2.237526 2.082437 2.086181                        1.836918
4       4 107.39703 26.31897 0.7535678 2.477804 2.016423 2.091085                        1.994526
5       5 130.34615 23.39414 1.2640829 2.808684 2.019380 2.173910                        2.000000
```

b. Build a table with your results succinctly displayed.

```
> print(ss_table)
  Centers BetweenSS_TotalSS_Percent
1       2                      67.6
2       3                      82.0
3       5                      89.8
4       7                      93.8
5       9                      94.8
6      11                      95.5
```

This is the result between_SS/total_SS (%) table showing clustering performance between different k values.

4. Prediction (using Training and Test Sets)

Following the rubric, try to predict the values of the dependent variable in the test sets.

a. Use the glm(…) method to get linear fits for the data.

Try clusters of 5,7,9,11
Remember to generate labels using kmeans.
Do a cross-tabulation of the results to see how good you did by comparing kmeans vs prediction results.

For analysis, we consider three cases for the training-test split: 70-30%, 60-40%, 50-50%.
Performing Using K-means clustering on the training set to create training labels for kNN, keeping center as 7

Cross Tabulation:
Train. Test - 70%-30%:
Comparing the results using CrossTable gmodels package.

Centers: 5

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

Total Observations in Table: 845

test.kmeans$cluster	kmeans(predicted_probs, centers)$cluster 1	2	3	4	5	Row Total
1	2	158	125	0	0	285
	64.863	205.746	104.393	55.651	55.651	
	0.007	0.554	0.439	0.000	0.000	0.337
	0.010	1.000	0.817	0.000	0.000	
	0.002	0.187	0.148	0.000	0.000	
2	0	0	0	7	130	137
	33.075	25.617	24.806	14.583	398.492	
	0.000	0.000	0.000	0.051	0.949	0.162
	0.000	0.000	0.000	0.042	0.788	
	0.000	0.000	0.000	0.008	0.154	
3	9	0	0	120	35	164
	23.639	30.665	29.695	241.691	0.277	
	0.055	0.000	0.000	0.732	0.213	0.194
	0.044	0.000	0.000	0.727	0.212	
	0.011	0.000	0.000	0.142	0.041	
4	59	0	9	1	0	69
	107.627	12.902	0.977	11.548	13.473	
	0.855	0.000	0.130	0.014	0.000	0.082
	0.289	0.000	0.059	0.006	0.000	
	0.070	0.000	0.011	0.001	0.000	
5	134	0	19	37	0	190
	169.325	35.527	6.896	0.000	37.101	
	0.705	0.000	0.100	0.195	0.000	0.225
	0.657	0.000	0.124	0.224	0.000	
	0.159	0.000	0.022	0.044	0.000	
Column Total	204	158	153	165	165	845
	0.241	0.187	0.181	0.195	0.195	

Centers: 7

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

Total Observations in Table: 634

test.kmeans$cluster	kmeans(predicted_probs, centers)$cluster 1	2	3	4	5	6	7	Row Total
1	5	0	0	85	0	5	0	95
	13.874	8.241	1.498	223.108	10.789	8.409	15.434	
	0.053	0.000	0.000	0.895	0.000	0.053	0.000	0.150
	0.033	0.000	0.000	0.659	0.000	0.044	0.000	
	0.008	0.000	0.000	0.134	0.000	0.008	0.000	
2	45	4	0	10	0	1	0	60
	65.158	0.279	0.946	0.399	6.814	8.788	9.748	
	0.750	0.067	0.000	0.167	0.000	0.017	0.000	0.095
	0.296	0.073	0.000	0.078	0.000	0.009	0.000	
	0.071	0.006	0.000	0.016	0.000	0.002	0.000	
3	0	0	0	12	0	107	0	119
	28.530	10.323	1.877	6.160	13.514	347.008	19.333	
	0.000	0.000	0.000	0.101	0.000	0.899	0.000	0.188
	0.000	0.000	0.000	0.093	0.000	0.947	0.000	
	0.000	0.000	0.000	0.019	0.000	0.169	0.000	
4	0	2	0	0	53	0	55	110
	26.372	5.962	1.735	22.382	131.354	19.606	77.143	
	0.000	0.018	0.000	0.000	0.482	0.000	0.500	0.174
	0.000	0.036	0.000	0.000	0.736	0.000	0.534	
	0.000	0.003	0.000	0.000	0.084	0.000	0.087	
5	12	44	0	0	0	0	48	104
	6.709	135.607	1.640	21.161	11.811	18.536	57.260	
	0.115	0.423	0.000	0.000	0.000	0.000	0.462	0.164
	0.079	0.800	0.000	0.000	0.000	0.000	0.466	
	0.019	0.069	0.000	0.000	0.000	0.000	0.076	
6	90	5	0	22	0	0	0	117
	136.816	2.613	1.845	0.137	13.287	20.853	19.008	
	0.769	0.043	0.000	0.188	0.000	0.000	0.000	0.185
	0.592	0.091	0.000	0.171	0.000	0.000	0.000	
	0.142	0.008	0.000	0.035	0.000	0.000	0.000	
7	0	0	10	0	19	0	0	29
	6.953	2.516	199.078	5.901	74.907	5.169	4.711	
	0.000	0.000	0.345	0.000	0.655	0.000	0.000	0.046
	0.000	0.000	1.000	0.000	0.264	0.000	0.000	
	0.000	0.000	0.016	0.000	0.030	0.000	0.000	
Column Total	152	55	10	129	72	113	103	634
	0.240	0.087	0.016	0.203	0.114	0.178	0.162	

Centers: 9

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

Total Observations in Table: 634

	kmeans(predicted_probs, centers)$cluster									
test.kmeans$cluster	1	2	3	4	5	6	7	8	9	Row Total
1	0	0	0	0	0	0	19	10	0	29
	3.797	2.790	2.379	3.888	3.110	5.215	105.494	199.078	4.803	
	0.000	0.000	0.000	0.000	0.000	0.000	0.655	0.345	0.000	0.046
	0.000	0.000	0.000	0.000	0.000	0.000	0.339	1.000	0.000	
	0.000	0.000	0.000	0.000	0.000	0.000	0.030	0.016	0.000	
2	0	0	0	18	51	0	10	0	0	79
	10.342	7.601	6.479	5.182	213.442	14.205	1.309	1.246	13.084	
	0.000	0.000	0.000	0.228	0.646	0.000	0.127	0.000	0.000	0.125
	0.000	0.000	0.000	0.212	0.750	0.000	0.179	0.000	0.000	
	0.000	0.000	0.000	0.028	0.080	0.000	0.016	0.000	0.000	
3	0	0	0	0	9	0	27	0	0	36
	4.713	3.464	2.953	4.826	6.839	6.473	178.439	0.568	5.962	
	0.000	0.000	0.000	0.000	0.250	0.000	0.750	0.000	0.000	0.057
	0.000	0.000	0.000	0.000	0.132	0.000	0.482	0.000	0.000	
	0.000	0.000	0.000	0.000	0.014	0.000	0.043	0.000	0.000	
4	4	14	0	1	0	31	0	0	8	58
	1.700	12.703	4.757	5.905	6.221	40.576	5.123	0.915	0.268	
	0.069	0.241	0.000	0.017	0.000	0.534	0.000	0.000	0.138	0.091
	0.048	0.230	0.000	0.012	0.000	0.272	0.000	0.000	0.076	
	0.006	0.022	0.000	0.002	0.000	0.049	0.000	0.000	0.013	
5	0	6	0	58	8	0	0	0	0	72
	9.426	0.124	5.905	242.146	0.010	12.946	6.360	1.136	11.924	
	0.000	0.083	0.000	0.806	0.111	0.000	0.000	0.000	0.000	0.114
	0.000	0.098	0.000	0.682	0.118	0.000	0.000	0.000	0.000	
	0.000	0.009	0.000	0.091	0.013	0.000	0.000	0.000	0.000	
6	0	22	0	8	0	8	0	0	0	38
	4.975	92.036	3.117	1.657	4.076	0.199	3.356	0.599	6.293	
	0.000	0.579	0.000	0.211	0.000	0.211	0.000	0.000	0.000	0.060
	0.000	0.361	0.000	0.094	0.000	0.070	0.000	0.000	0.000	
	0.000	0.035	0.000	0.013	0.000	0.013	0.000	0.000	0.000	
7	17	0	0	0	0	6	0	0	72	95
	1.674	9.140	7.792	12.737	10.189	7.189	8.391	1.498	201.223	
	0.179	0.000	0.000	0.000	0.000	0.063	0.000	0.000	0.758	0.150
	0.205	0.000	0.000	0.000	0.000	0.053	0.000	0.000	0.686	
	0.027	0.000	0.000	0.000	0.000	0.009	0.000	0.000	0.114	
8	62	0	52	0	0	0	0	0	3	117
	142.279	11.257	187.374	15.686	12.549	21.038	10.334	1.845	13.841	
	0.530	0.000	0.444	0.000	0.000	0.000	0.000	0.000	0.026	0.185
	0.747	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.029	
	0.098	0.000	0.082	0.000	0.000	0.000	0.000	0.000	0.005	
9	0	19	0	0	0	69	0	0	22	110
	14.401	6.693	9.022	14.748	11.798	122.487	9.716	1.735	0.785	
	0.000	0.173	0.000	0.000	0.000	0.627	0.000	0.000	0.200	0.174
	0.000	0.311	0.000	0.000	0.000	0.605	0.000	0.000	0.210	
	0.000	0.030	0.000	0.000	0.000	0.109	0.000	0.000	0.035	
Column Total	83	61	52	85	68	114	56	10	105	634
	0.131	0.096	0.082	0.134	0.107	0.180	0.088	0.016	0.166	

Centers: 11

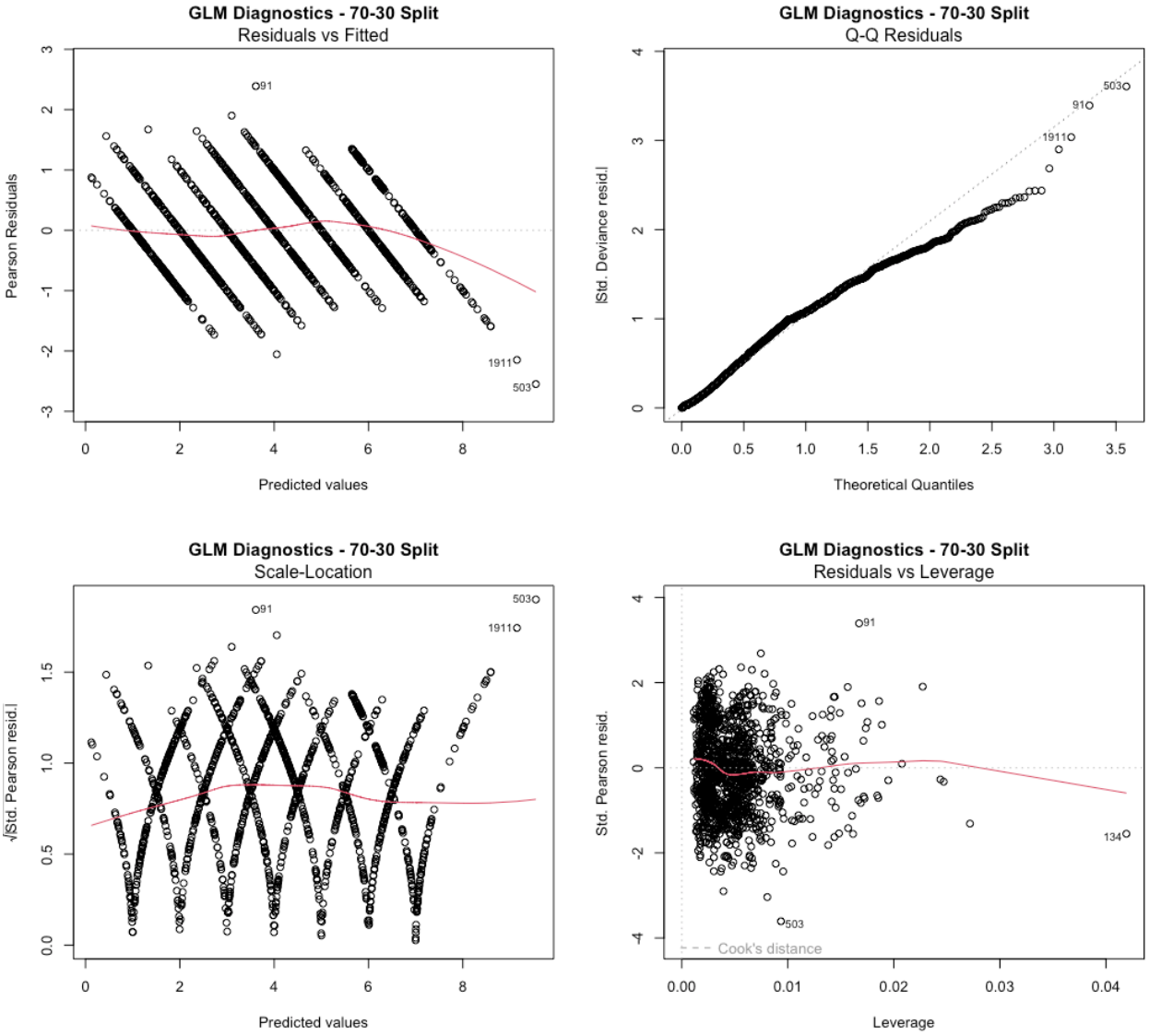
```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|
```

Total Observations in Table: 634

Total Observations in Table: 634

	kmeans(predicted_probs, centers)$cluster											
test.kmeans$cluster	1	2	3	4	5	6	7	8	9	10	11	Row Total
1	0	1	0	0	1	0	10	0	0	0	16	28
	1.634	1.065	1.590	1.413	2.392	2.650	7.244	0.574	4.902	1.457	70.502	
	0.000	0.036	0.000	0.000	0.036	0.000	0.357	0.000	0.000	0.000	0.571	0.044
	0.000	0.016	0.000	0.000	0.011	0.000	0.101	0.000	0.000	0.000	0.276	
	0.000	0.002	0.000	0.000	0.002	0.000	0.016	0.000	0.000	0.000	0.025	
2	1	11	0	0	0	2	18	0	0	4	11	47
	1.107	9.280	2.669	2.372	6.968	1.347	15.486	0.964	8.229	0.987	10.441	
	0.021	0.234	0.000	0.000	0.000	0.043	0.383	0.000	0.000	0.085	0.234	0.074
	0.027	0.180	0.000	0.000	0.000	0.033	0.182	0.000	0.000	0.121	0.190	
	0.002	0.017	0.000	0.000	0.000	0.003	0.028	0.000	0.000	0.006	0.017	
3	0	0	0	0	8	0	0	0	62	0	0	70
	4.085	6.735	3.975	3.533	0.545	6.625	10.931	1.435	201.910	3.644	6.404	
	0.000	0.000	0.000	0.000	0.114	0.000	0.000	0.000	0.886	0.000	0.000	0.110
	0.000	0.000	0.000	0.000	0.085	0.000	0.000	0.000	0.559	0.000	0.000	
	0.000	0.000	0.000	0.000	0.013	0.000	0.000	0.000	0.098	0.000	0.000	
4	0	0	0	0	62	0	0	0	34	0	0	96
	5.603	9.237	5.451	4.845	160.302	9.085	14.991	1.968	17.586	4.997	8.782	
	0.000	0.000	0.000	0.000	0.646	0.000	0.000	0.000	0.354	0.000	0.000	0.151
	0.000	0.000	0.000	0.000	0.660	0.000	0.000	0.000	0.306	0.000	0.000	
	0.000	0.000	0.000	0.000	0.098	0.000	0.000	0.000	0.054	0.000	0.000	
5	0	0	0	0	0	0	0	13	15	0	0	28
	1.634	2.694	1.590	1.413	4.151	2.650	4.372	268.931	20.800	1.457	2.562	
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.464	0.536	0.000	0.000	0.044
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.135	0.000	0.000	
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.021	0.024	0.000	0.000	
6	0	34	0	0	0	11	28	0	0	2	1	76
	4.435	97.402	4.315	3.836	11.268	2.016	21.930	1.558	13.306	0.967	5.097	
	0.000	0.447	0.000	0.000	0.000	0.145	0.368	0.000	0.000	0.026	0.013	0.120
	0.000	0.557	0.000	0.000	0.000	0.183	0.283	0.000	0.000	0.061	0.017	
	0.000	0.054	0.000	0.000	0.000	0.017	0.044	0.000	0.000	0.003	0.002	
7	18	0	19	1	0	1	0	0	0	15	0	54
	69.962	5.196	82.800	1.092	8.006	3.306	8.432	1.107	9.454	52.861	4.940	
	0.333	0.000	0.352	0.019	0.000	0.019	0.000	0.000	0.000	0.278	0.000	0.085
	0.486	0.000	0.528	0.031	0.000	0.017	0.000	0.000	0.000	0.455	0.000	
	0.028	0.000	0.030	0.002	0.000	0.002	0.000	0.000	0.000	0.024	0.000	
8	0	5	0	0	0	0	37	0	0	0	16	58
	3.385	0.060	3.293	2.927	8.599	5.489	86.214	1.189	10.155	3.019	21.553	
	0.000	0.086	0.000	0.000	0.000	0.000	0.638	0.000	0.000	0.000	0.276	0.091
	0.000	0.082	0.000	0.000	0.000	0.000	0.374	0.000	0.000	0.000	0.276	
	0.000	0.008	0.000	0.000	0.000	0.000	0.058	0.000	0.000	0.000	0.025	
9	0	0	0	0	23	0	6	0	0	0	14	43
	2.509	4.137	2.442	2.170	43.351	4.069	0.076	0.882	7.528	2.238	25.759	
	0.000	0.000	0.000	0.000	0.535	0.000	0.140	0.000	0.000	0.000	0.326	0.068
	0.000	0.000	0.000	0.000	0.245	0.000	0.061	0.000	0.000	0.000	0.241	
	0.000	0.000	0.000	0.000	0.036	0.000	0.009	0.000	0.000	0.000	0.022	
10	16	0	17	31	0	0	0	0	0	1	0	65
	39.279	6.254	47.993	234.201	9.637	6.151	10.150	1.333	11.380	1.679	5.946	
	0.246	0.000	0.262	0.477	0.000	0.000	0.000	0.000	0.000	0.015	0.000	0.103
	0.432	0.000	0.472	0.969	0.000	0.000	0.000	0.000	0.000	0.030	0.000	
	0.025	0.000	0.027	0.049	0.000	0.000	0.000	0.000	0.000	0.002	0.000	
11	2	10	0	0	0	46	0	0	0	11	0	69
	1.020	1.702	3.918	3.483	10.230	238.574	10.774	1.415	12.080	15.282	6.312	
	0.029	0.145	0.000	0.000	0.000	0.667	0.000	0.000	0.000	0.159	0.000	0.109
	0.054	0.164	0.000	0.000	0.000	0.767	0.000	0.000	0.000	0.333	0.000	
	0.003	0.016	0.000	0.000	0.000	0.073	0.000	0.000	0.000	0.017	0.000	
Column Total	37	61	36	32	94	60	99	13	111	33	58	634
	0.058	0.096	0.057	0.050	0.148	0.095	0.156	0.021	0.175	0.052	0.091	

Linear Modelling: using glm() function in stats package.
anova – Analysis of Variance on the model result



GLM Diagnostics (70-30 Split)

The residuals vs fitted plot shows a striped, curved pattern, which makes the model isn't handle non-linear patterns well. The Q-Q plot has a few points deviating in the upper tail, especially outliers like 5030 and 19110, so the residuals don't look fully normal.

In the scale-location plot, the spread isn't perfectly even, suggesting there's some variance inconsistency. The residuals vs leverage plot points out a couple of influential values, like 91 and 134, which could be affecting the model results.

Summary and Confidence Intervals:

```
===== GLM 70-30 Split =====
> summary(glm_70)

Call:
glm(formula = lastcol ~ Gender + Age + Height + Weight + SMOKE +
    MTRANS, family = gaussian(), data = trainData1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.8165608  0.3218038  30.505  < 2e-16 ***
Gender       0.1040694  0.0315819   3.295  0.00101 **
Age          0.0219855  0.0025629   8.579  < 2e-16 ***
Height      -7.8879651  0.1920019 -41.083  < 2e-16 ***
Weight       0.0807084  0.0005629 143.386  < 2e-16 ***
SMOKE       -0.0686357  0.0895823  -0.766  0.44370
MTRANS       0.0313520  0.0124635   2.516  0.01199 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.2193262)

    Null deviance: 5914.66  on 1476  degrees of freedom
Residual deviance:  322.41  on 1470  degrees of freedom
AIC: 1959.6

Number of Fisher Scoring iterations: 2

> ci_70 <- confint(glm_70)
Waiting for profiling to be done...
> print(ci_70)
                   2.5 %      97.5 %
(Intercept)  9.185837018 10.44728460
Gender       0.042170032  0.16596869
Age          0.016962379  0.02700858
Height      -8.264281902 -7.51164837
Weight       0.079605172  0.08181160
SMOKE       -0.244213727  0.10694233
MTRANS       0.006924029  0.05577994
> cat("Intercept 95% CI (70-30):\n")
Intercept 95% CI (70-30):
> print(ci_70["(Intercept)", ])
    2.5 %    97.5 %
 9.185837 10.447285
```

Train, Test - 60%-40%:
Run for 60-40 split

Comparing the results using CrossTable gmodels package.

Centers: 5

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


Total Observations in Table:  845
```

test.kmeans$cluster	kmeans(predicted_probs, centers)$cluster 1	2	3	4	5	Row Total
1	2	158	125	0	0	285
	64.863	205.746	104.393	55.651	55.651	
	0.007	0.554	0.439	0.000	0.000	0.337
	0.010	1.000	0.817	0.000	0.000	
	0.002	0.187	0.148	0.000	0.000	
2	0	0	0	7	130	137
	33.075	25.617	24.806	14.583	398.492	
	0.000	0.000	0.000	0.051	0.949	0.162
	0.000	0.000	0.000	0.042	0.788	
	0.000	0.000	0.000	0.008	0.154	
3	9	0	0	120	35	164
	23.639	30.665	29.695	241.691	0.277	
	0.055	0.000	0.000	0.732	0.213	0.194
	0.044	0.000	0.000	0.727	0.212	
	0.011	0.000	0.000	0.142	0.041	
4	59	0	9	1	0	69
	107.627	12.902	0.977	11.548	13.473	
	0.855	0.000	0.130	0.014	0.000	0.082
	0.289	0.000	0.059	0.006	0.000	
	0.070	0.000	0.011	0.001	0.000	
5	134	0	19	37	0	190
	169.325	35.527	6.896	0.000	37.101	
	0.705	0.000	0.100	0.195	0.000	0.225
	0.657	0.000	0.124	0.224	0.000	
	0.159	0.000	0.022	0.044	0.000	
Column Total	204	158	153	165	165	845
	0.241	0.187	0.181	0.195	0.195	

Centers: 7

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

Total Observations in Table: 845

test.kmeans$cluster	kmeans(predicted_probs, centers)$cluster							
	1	2	3	4	5	6	7	Row Total
1	0	0	0	4	80	0	65	149
	19.220	33.327	12.520	7.729	115.880	23.276	97.011	
	0.000	0.000	0.000	0.027	0.537	0.000	0.436	0.176
	0.000	0.000	0.000	0.048	0.552	0.000	0.560	
	0.000	0.000	0.000	0.005	0.095	0.000	0.077	
2	0	0	56	0	0	3	0	59
	7.611	13.196	525.548	5.795	10.124	4.193	8.099	
	0.000	0.000	0.949	0.000	0.000	0.051	0.000	0.070
	0.000	0.000	0.789	0.000	0.000	0.023	0.000	
	0.000	0.000	0.066	0.000	0.000	0.004	0.000	
3	95	23	0	0	1	21	0	140
	327.805	2.207	11.763	13.751	22.065	0.035	19.219	
	0.679	0.164	0.000	0.000	0.007	0.150	0.000	0.166
	0.872	0.122	0.000	0.000	0.007	0.159	0.000	
	0.112	0.027	0.000	0.000	0.001	0.025	0.000	
4	11	0	15	0	0	108	0	134
	2.285	29.972	1.243	13.162	22.994	362.151	18.395	
	0.082	0.000	0.112	0.000	0.000	0.806	0.000	0.159
	0.101	0.000	0.211	0.000	0.000	0.818	0.000	
	0.013	0.000	0.018	0.000	0.000	0.128	0.000	
5	0	0	0	79	1	0	50	130
	16.769	29.077	10.923	343.522	20.353	20.308	57.932	
	0.000	0.000	0.000	0.608	0.008	0.000	0.385	0.154
	0.000	0.000	0.000	0.952	0.007	0.000	0.431	
	0.000	0.000	0.000	0.093	0.001	0.000	0.059	
6	3	53	0	0	3	0	1	60
	2.902	116.733	5.041	5.893	5.170	9.373	6.358	
	0.050	0.883	0.000	0.000	0.050	0.000	0.017	0.071
	0.028	0.280	0.000	0.000	0.021	0.000	0.009	
	0.004	0.063	0.000	0.000	0.004	0.000	0.001	
7	0	113	0	0	60	0	0	173
	22.316	142.688	14.536	16.993	30.954	27.025	23.749	
	0.000	0.653	0.000	0.000	0.347	0.000	0.000	0.205
	0.000	0.598	0.000	0.000	0.414	0.000	0.000	
	0.000	0.134	0.000	0.000	0.071	0.000	0.000	
Column Total	109	189	71	83	145	132	116	845
	0.129	0.224	0.084	0.098	0.172	0.156	0.137	

Centers: 9

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

Total Observations in Table: 845

	kmeans(predicted_probs, centers)$cluster									
test.kmeans$cluster	1	2	3	4	5	6	7	8	9	Row Total
1	5	0	0	5	0	4	0	0	0	14
	10.872	1.359	1.972	4.142	1.011	7.934	2.369	0.845	2.021	
	0.357	0.000	0.000	0.357	0.000	0.286	0.000	0.000	0.000	0.017
	0.065	0.000	0.000	0.040	0.000	0.062	0.000	0.000	0.000	
	0.006	0.000	0.000	0.006	0.000	0.005	0.000	0.000	0.000	
2	0	6	0	44	0	0	91	0	0	141
	12.849	4.314	19.857	25.676	10.179	10.846	188.905	8.510	20.357	
	0.000	0.043	0.000	0.312	0.000	0.000	0.645	0.000	0.000	0.167
	0.000	0.073	0.000	0.352	0.000	0.000	0.636	0.000	0.000	
	0.000	0.007	0.000	0.052	0.000	0.000	0.108	0.000	0.000	
3	0	20	0	1	0	0	31	0	1	53
	4.830	42.916	7.464	5.968	3.826	4.077	54.113	3.199	5.783	
	0.000	0.377	0.000	0.019	0.000	0.000	0.585	0.000	0.019	0.063
	0.000	0.244	0.000	0.008	0.000	0.000	0.217	0.000	0.008	
	0.000	0.024	0.000	0.001	0.000	0.000	0.037	0.000	0.001	
4	0	17	14	0	0	0	0	0	93	124
	11.299	2.050	0.687	18.343	8.951	9.538	20.985	7.484	315.007	
	0.000	0.137	0.113	0.000	0.000	0.000	0.000	0.000	0.750	0.147
	0.000	0.207	0.118	0.000	0.000	0.000	0.000	0.000	0.762	
	0.000	0.020	0.017	0.000	0.000	0.000	0.000	0.000	0.110	
5	0	39	0	1	0	0	21	0	4	65
	5.923	169.442	9.154	7.719	4.692	5.000	9.091	3.923	3.090	
	0.000	0.600	0.000	0.015	0.000	0.000	0.323	0.000	0.062	0.077
	0.000	0.476	0.000	0.008	0.000	0.000	0.147	0.000	0.033	
	0.000	0.046	0.000	0.001	0.000	0.000	0.025	0.000	0.005	
6	0	0	95	0	0	0	0	2	24	121
	11.026	11.742	356.669	17.899	8.735	9.308	20.477	3.851	2.441	
	0.000	0.000	0.785	0.000	0.000	0.000	0.000	0.017	0.198	0.143
	0.000	0.000	0.798	0.000	0.000	0.000	0.000	0.039	0.197	
	0.000	0.000	0.112	0.000	0.000	0.000	0.000	0.002	0.028	
7	57	0	0	73	3	10	0	0	0	143
	148.364	13.877	20.138	127.070	5.195	0.091	24.200	8.631	20.646	
	0.399	0.000	0.000	0.510	0.021	0.070	0.000	0.000	0.000	0.169
	0.740	0.000	0.000	0.584	0.049	0.154	0.000	0.000	0.000	
	0.067	0.000	0.000	0.086	0.004	0.012	0.000	0.000	0.000	
8	15	0	0	1	58	51	0	0	0	125
	1.144	12.130	17.604	16.545	265.821	178.119	21.154	7.544	18.047	
	0.120	0.000	0.000	0.008	0.464	0.408	0.000	0.000	0.000	0.148
	0.195	0.000	0.000	0.008	0.951	0.785	0.000	0.000	0.000	
	0.018	0.000	0.000	0.001	0.069	0.060	0.000	0.000	0.000	
9	0	0	10	0	0	0	0	49	0	59
	5.376	5.725	0.344	8.728	4.259	4.538	9.985	579.820	8.518	
	0.000	0.000	0.169	0.000	0.000	0.000	0.000	0.831	0.000	0.070
	0.000	0.000	0.084	0.000	0.000	0.000	0.000	0.961	0.000	
	0.000	0.000	0.012	0.000	0.000	0.000	0.000	0.058	0.000	
Column Total	77	82	119	125	61	65	143	51	122	845
	0.091	0.097	0.141	0.148	0.072	0.077	0.169	0.060	0.144	

Centers: 11

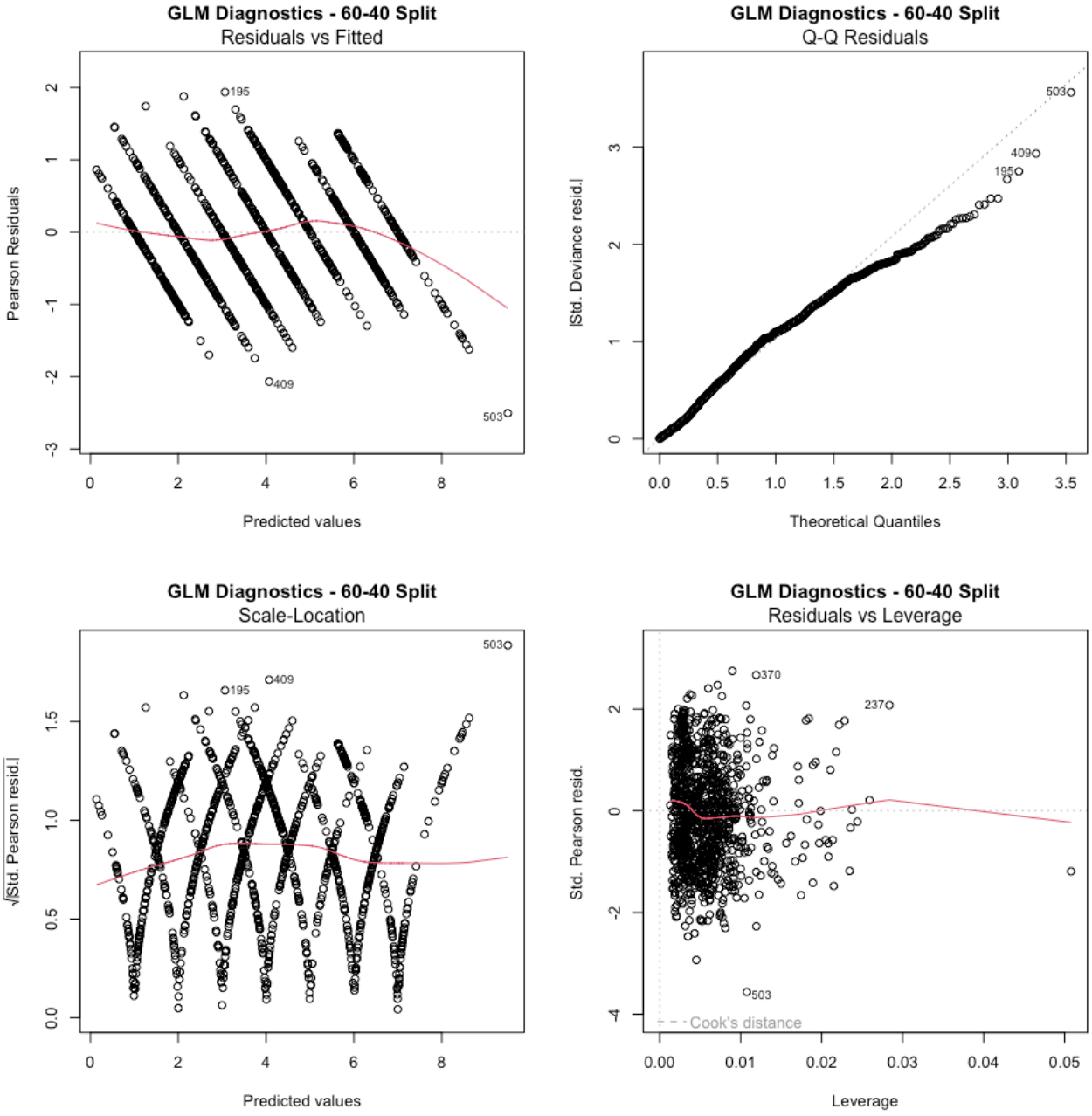
```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|
```

Total Observations in Table: 845

test.kmeans$cluster	kmeans(predicted_probs, centers)$cluster 1	2	3	4	5	6	7	8	9	10	11	Row Total
1	0	0	20	0	16	1	0	0	0	6	0	43
	3.104	4.682	39.549	0.509	53.575	1.381	5.598	4.071	4.376	1.577	5.496	
	0.000	0.000	0.465	0.000	0.372	0.023	0.000	0.000	0.000	0.140	0.000	0.051
	0.000	0.000	0.189	0.000	0.262	0.017	0.000	0.000	0.000	0.085	0.000	
	0.000	0.000	0.024	0.000	0.019	0.001	0.000	0.000	0.000	0.007	0.000	
2	0	0	0	0	0	0	9	15	0	0	50	74
	5.342	8.057	9.283	0.876	5.342	5.254	0.042	9.122	7.531	6.218	173.785	
	0.000	0.000	0.000	0.000	0.000	0.000	0.122	0.203	0.000	0.000	0.676	0.088
	0.000	0.000	0.000	0.000	0.000	0.000	0.082	0.188	0.000	0.000	0.463	
	0.000	0.000	0.000	0.000	0.000	0.000	0.011	0.018	0.000	0.000	0.059	
3	46	3	0	10	0	0	0	0	0	0	0	59
	409.069	1.825	7.401	123.919	4.259	4.189	7.680	5.586	6.005	4.957	7.541	
	0.780	0.051	0.000	0.169	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.070
	0.754	0.033	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
	0.054	0.004	0.000	0.012	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
4	0	0	0	0	0	0	0	64	0	0	29	93
	6.714	10.125	11.666	1.101	6.714	6.604	12.107	346.009	9.465	7.814	24.640	
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.688	0.000	0.000	0.312	0.110
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.800	0.000	0.000	0.269	
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.076	0.000	0.000	0.034	
5	0	0	15	0	23	9	0	0	0	3	0	50
	3.609	5.444	12.145	0.592	104.168	8.365	6.509	4.734	5.089	0.343	6.391	
	0.000	0.000	0.300	0.000	0.460	0.180	0.000	0.000	0.000	0.060	0.000	0.059
	0.000	0.000	0.142	0.000	0.377	0.150	0.000	0.000	0.000	0.042	0.000	
	0.000	0.000	0.018	0.000	0.027	0.011	0.000	0.000	0.000	0.004	0.000	
6	0	0	0	0	0	0	68	1	0	8	24	101
	7.291	10.996	12.670	1.195	7.291	7.172	228.838	7.667	10.279	0.028	9.529	
	0.000	0.000	0.000	0.000	0.000	0.000	0.673	0.010	0.000	0.079	0.238	0.120
	0.000	0.000	0.000	0.000	0.000	0.000	0.618	0.013	0.000	0.113	0.222	
	0.000	0.000	0.000	0.000	0.000	0.000	0.080	0.001	0.000	0.009	0.028	
7	0	11	0	0	6	48	0	0	59	0	0	124
	8.951	0.463	15.555	1.467	0.973	174.482	16.142	11.740	170.450	10.419	15.849	
	0.000	0.089	0.000	0.000	0.048	0.387	0.000	0.000	0.476	0.000	0.000	0.147
	0.000	0.120	0.000	0.000	0.098	0.800	0.000	0.000	0.686	0.000	0.000	
	0.000	0.013	0.000	0.000	0.007	0.057	0.000	0.000	0.070	0.000	0.000	
8	0	0	59	0	4	0	25	0	0	47	0	135
	9.746	14.698	104.487	1.598	3.387	9.586	3.138	12.781	13.740	112.086	17.254	
	0.000	0.000	0.437	0.000	0.030	0.000	0.185	0.000	0.000	0.348	0.000	0.160
	0.000	0.000	0.557	0.000	0.066	0.000	0.227	0.000	0.000	0.662	0.000	
	0.000	0.000	0.070	0.000	0.005	0.000	0.030	0.000	0.000	0.056	0.000	
9	15	78	0	0	0	1	0	0	27	0	0	121
	4.494	318.994	15.179	1.432	8.735	6.708	15.751	11.456	17.512	10.167	15.465	
	0.124	0.645	0.000	0.000	0.000	0.008	0.000	0.000	0.223	0.000	0.000	0.143
	0.246	0.848	0.000	0.000	0.000	0.017	0.000	0.000	0.314	0.000	0.000	
	0.018	0.092	0.000	0.000	0.000	0.001	0.000	0.000	0.032	0.000	0.000	
10	0	0	0	0	0	0	8	0	0	3	5	16
	1.155	1.742	2.007	0.189	1.155	1.136	16.810	1.515	1.628	2.039	4.270	
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.000	0.000	0.188	0.312	0.019
	0.000	0.000	0.000	0.000	0.000	0.000	0.073	0.000	0.000	0.042	0.046	
	0.000	0.000	0.000	0.000	0.000	0.000	0.009	0.000	0.000	0.004	0.006	
11	0	0	12	0	12	1	0	0	0	4	0	29
	2.093	3.157	19.221	0.343	46.878	0.545	3.775	2.746	2.951	1.003	3.707	
	0.000	0.000	0.414	0.000	0.414	0.034	0.000	0.000	0.000	0.138	0.000	0.034
	0.000	0.000	0.113	0.000	0.197	0.017	0.000	0.000	0.000	0.056	0.000	
	0.000	0.000	0.014	0.000	0.014	0.001	0.000	0.000	0.000	0.005	0.000	
Column Total	61	92	106	10	61	60	110	80	86	71	108	845
	0.072	0.109	0.125	0.012	0.072	0.071	0.130	0.095	0.102	0.084	0.128	

Linear Modelling: using glm() function in stats package.

anova – Analysis of Variance on the model result



GLM Diagnostics (60-40 Split)

The residuals vs fitted plot shows a clear striped and curved pattern again, which suggests the model is still missing non-linear trends. The Q-Q plot shows that residuals deviate in the upper tail, especially with outliers like 5030, 4090, and 195.

Points like 503 and 2370 in the leverage plot stand out, meaning they might be influencing the model more than usual.

Summary and Confidence Intervals:

```
> summary(glm_60)

Call:
glm(formula = lastcol ~ Gender + Age + Height + Weight + SMOKE +
    MTRANS, family = gaussian(), data = trainData2)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.0836856  0.3374952  29.878  < 2e-16 ***
Gender       0.1495397  0.0337166   4.435 1.00e-05 ***
Age          0.0218155  0.0027847   7.834 9.99e-15 ***
Height      -8.0211694  0.2050402 -39.120  < 2e-16 ***
Weight       0.0810528  0.0006062 133.715  < 2e-16 ***
SMOKE       -0.2314676  0.0913148  -2.535  0.01137 *
MTRANS       0.0390709  0.0133391   2.929  0.00346 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.2155336)

    Null deviance: 5011.74  on 1265  degrees of freedom
Residual deviance:  271.36  on 1259  degrees of freedom
AIC: 1658.9

Number of Fisher Scoring iterations: 2
> ci_60 <- confint(glm_60)
Waiting for profiling to be done...
> print(ci_60)
                  2.5 %      97.5 %
(Intercept)  9.42220715 10.74516397
Gender       0.08345637  0.21562303
Age          0.01635753  0.02727354
Height      -8.42304072 -7.61929801
Weight       0.07986473  0.08224083
SMOKE       -0.41044134 -0.05249387
MTRANS       0.01292672  0.06521503
> cat("Intercept 95% CI (60-40):\n")
Intercept 95% CI (60-40):
> print(ci_60["(Intercept)", ])
    2.5 %    97.5 %
 9.422207 10.745164
```

Train, Test - 50%-50%:

Centers: 5

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

Total Observations in Table: 1056

	kmeans(predicted_probs, centers)$cluster					
test.kmeans$cluster	1	2	3	4	5	Row Total
1	0	0	120	126	0	246
	47.756	45.892	137.230	136.823	62.898	
	0.000	0.000	0.488	0.512	0.000	0.233
	0.000	0.000	0.649	0.633	0.000	
	0.000	0.000	0.114	0.119	0.000	
2	1	0	65	0	79	145
	26.184	27.050	61.725	27.325	47.413	
	0.007	0.000	0.448	0.000	0.545	0.137
	0.005	0.000	0.351	0.000	0.293	
	0.001	0.000	0.062	0.000	0.075	
3	0	0	0	73	0	73
	14.171	13.618	12.789	255.134	18.665	
	0.000	0.000	0.000	1.000	0.000	0.069
	0.000	0.000	0.000	0.367	0.000	
	0.000	0.000	0.000	0.069	0.000	
4	137	6	0	0	191	334
	80.310	50.886	58.513	62.941	130.587	
	0.410	0.018	0.000	0.000	0.572	0.316
	0.668	0.030	0.000	0.000	0.707	
	0.130	0.006	0.000	0.000	0.181	
5	67	191	0	0	0	258
	5.712	424.088	45.199	48.619	65.966	
	0.260	0.740	0.000	0.000	0.000	0.244
	0.327	0.970	0.000	0.000	0.000	
	0.063	0.181	0.000	0.000	0.000	
Column Total	205	197	185	199	270	1056
	0.194	0.187	0.175	0.188	0.256	

Centers: 7

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

Total Observations in Table: 1056

test.kmeans$cluster	kmeans(predicted_probs, centers)$cluster 1	2	3	4	5	6	7	Row Total
1	0	27	143	0	0	0	6	176
	31.500	0.255	526.500	29.167	23.500	30.167	2.391	
	0.000	0.153	0.812	0.000	0.000	0.000	0.034	0.167
	0.000	0.184	0.917	0.000	0.000	0.000	0.090	
	0.000	0.026	0.135	0.000	0.000	0.000	0.006	
2	0	0	0	30	53	5	0	88
	15.750	12.250	13.000	16.298	144.814	6.741	5.583	
	0.000	0.000	0.000	0.341	0.602	0.057	0.000	0.083
	0.000	0.000	0.000	0.171	0.376	0.028	0.000	
	0.000	0.000	0.000	0.028	0.050	0.005	0.000	
3	17	0	0	29	0	136	0	182
	7.446	25.335	26.886	0.045	24.301	352.109	11.547	
	0.093	0.000	0.000	0.159	0.000	0.747	0.000	0.172
	0.090	0.000	0.000	0.166	0.000	0.751	0.000	
	0.016	0.000	0.000	0.027	0.000	0.129	0.000	
4	0	0	6	0	0	0	61	67
	11.991	9.327	1.535	11.103	8.946	11.484	757.585	
	0.000	0.000	0.090	0.000	0.000	0.000	0.910	0.063
	0.000	0.000	0.038	0.000	0.000	0.000	0.910	
	0.000	0.000	0.006	0.000	0.000	0.000	0.058	
5	0	120	7	3	33	0	0	163
	29.173	417.322	12.114	21.345	5.800	27.938	10.342	
	0.000	0.736	0.043	0.018	0.202	0.000	0.000	0.154
	0.000	0.816	0.045	0.017	0.234	0.000	0.000	
	0.000	0.114	0.007	0.003	0.031	0.000	0.000	
6	0	0	0	113	55	24	0	192
	34.364	26.727	28.364	207.130	33.633	2.412	12.182	
	0.000	0.000	0.000	0.589	0.286	0.125	0.000	0.182
	0.000	0.000	0.000	0.646	0.390	0.133	0.000	
	0.000	0.000	0.000	0.107	0.052	0.023	0.000	
7	172	0	0	0	0	16	0	188
	568.875	26.170	27.773	31.155	25.102	8.168	11.928	
	0.915	0.000	0.000	0.000	0.000	0.085	0.000	0.178
	0.910	0.000	0.000	0.000	0.000	0.088	0.000	
	0.163	0.000	0.000	0.000	0.000	0.015	0.000	
Column Total	189	147	156	175	141	181	67	1056
	0.179	0.139	0.148	0.166	0.134	0.171	0.063	

Centers: 9

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

Total Observations in Table: 1056

test.kmeans$cluster	kmeans(predicted_probs, centers)$cluster 1	2	3	4	5	6	7	8	9	Row Total
1	116	5	3	0	37	0	0	0	1	162
	403.316	14.683	19.366	14.574	20.725	13.193	15.494	10.278	18.146	
	0.716	0.031	0.019	0.000	0.228	0.000	0.000	0.000	0.006	0.153
	0.811	0.032	0.018	0.000	0.319	0.000	0.000	0.000	0.008	
	0.110	0.005	0.003	0.000	0.035	0.000	0.000	0.000	0.001	
2	27	143	0	0	0	0	0	6	0	176
	0.421	536.381	27.167	15.833	19.333	14.333	16.833	2.391	21.833	
	0.153	0.812	0.000	0.000	0.000	0.000	0.000	0.034	0.000	0.167
	0.189	0.929	0.000	0.000	0.000	0.000	0.000	0.090	0.000	
	0.026	0.135	0.000	0.000	0.000	0.000	0.000	0.006	0.000	
3	0	0	0	82	0	0	45	0	0	127
	17.198	18.521	19.603	435.949	13.951	10.343	88.858	8.058	15.755	
	0.000	0.000	0.000	0.646	0.000	0.000	0.354	0.000	0.000	0.120
	0.000	0.000	0.000	0.863	0.000	0.000	0.446	0.000	0.000	
	0.000	0.000	0.000	0.078	0.000	0.000	0.043	0.000	0.000	
4	0	0	46	0	40	1	0	0	1	88
	11.917	12.833	77.362	7.917	95.184	5.306	8.417	5.583	9.008	
	0.000	0.000	0.523	0.000	0.455	0.011	0.000	0.000	0.011	0.083
	0.000	0.000	0.282	0.000	0.345	0.012	0.000	0.000	0.008	
	0.000	0.000	0.044	0.000	0.038	0.001	0.000	0.000	0.001	
5	0	6	0	0	0	0	0	61	0	67
	9.073	1.455	10.342	6.027	7.360	5.456	6.408	757.585	8.312	
	0.000	0.090	0.000	0.000	0.000	0.000	0.000	0.910	0.000	0.063
	0.000	0.039	0.000	0.000	0.000	0.000	0.000	0.910	0.000	
	0.000	0.006	0.000	0.000	0.000	0.000	0.000	0.058	0.000	
6	0	0	11	0	0	35	7	0	82	135
	18.281	19.688	4.645	12.145	14.830	52.416	2.707	8.565	254.248	
	0.000	0.000	0.081	0.000	0.000	0.259	0.052	0.000	0.607	0.128
	0.000	0.000	0.067	0.000	0.000	0.407	0.069	0.000	0.626	
	0.000	0.000	0.010	0.000	0.000	0.033	0.007	0.000	0.078	
7	0	0	0	13	0	38	46	0	0	97
	13.135	14.146	14.973	2.093	10.655	114.693	145.357	6.154	12.033	
	0.000	0.000	0.000	0.134	0.000	0.392	0.474	0.000	0.000	0.092
	0.000	0.000	0.000	0.137	0.000	0.442	0.455	0.000	0.000	
	0.000	0.000	0.000	0.012	0.000	0.036	0.044	0.000	0.000	
8	0	0	0	0	0	11	3	0	7	21
	2.844	3.062	3.241	1.889	2.307	50.461	0.489	1.332	7.414	
	0.000	0.000	0.000	0.000	0.000	0.524	0.143	0.000	0.333	0.020
	0.000	0.000	0.000	0.000	0.000	0.128	0.030	0.000	0.053	
	0.000	0.000	0.000	0.000	0.000	0.010	0.003	0.000	0.007	
9	0	0	103	0	39	1	0	0	40	183
	24.781	26.688	197.825	16.463	17.765	12.971	17.503	11.611	13.181	
	0.000	0.000	0.563	0.000	0.213	0.005	0.000	0.000	0.219	0.173
	0.000	0.000	0.632	0.000	0.336	0.012	0.000	0.000	0.305	
	0.000	0.000	0.098	0.000	0.037	0.001	0.000	0.000	0.038	
Column Total	143	154	163	95	116	86	101	67	131	1056
	0.135	0.146	0.154	0.090	0.110	0.081	0.096	0.063	0.124	

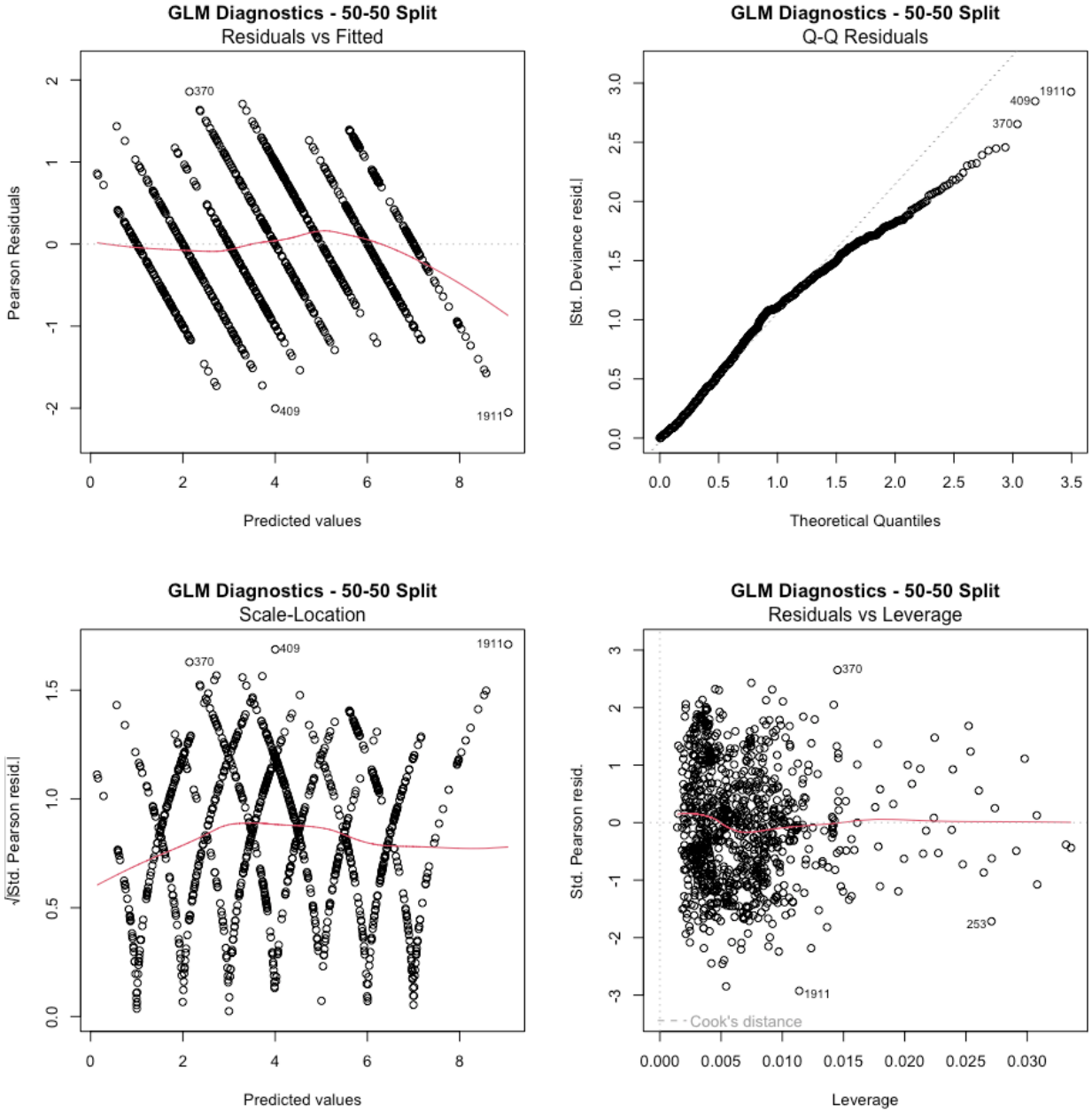
Centers: 11

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

Total Observations in Table: 1056

test.kmeans$cluster	kmeans(predicted_probs, centers)$cluster 1	2	3	4	5	6	7	8	9	10	11	Row Total
1	0	0	0	0	0	0	0	0	0	77	27	104
	9.356	8.470	7.583	1.280	13.000	12.015	11.227	12.803	9.159	457.870	34.004	
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.740	0.260	0.098
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.770	0.287	
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.073	0.026	
2	0	0	29	0	0	0	65	0	9	0	7	110
	9.896	8.958	54.873	1.354	13.750	12.708	237.664	13.542	0.049	10.417	0.796	
	0.000	0.000	0.264	0.000	0.000	0.000	0.591	0.000	0.082	0.000	0.064	0.104
	0.000	0.000	0.377	0.000	0.000	0.000	0.570	0.000	0.097	0.000	0.074	
	0.000	0.000	0.027	0.000	0.000	0.000	0.062	0.000	0.009	0.000	0.007	
3	0	0	31	0	0	0	2	0	0	12	42	87
	7.827	7.085	95.831	1.071	10.875	10.051	5.818	10.710	7.662	1.717	151.524	
	0.000	0.000	0.356	0.000	0.000	0.000	0.023	0.000	0.000	0.138	0.483	0.082
	0.000	0.000	0.403	0.000	0.000	0.000	0.018	0.000	0.000	0.120	0.447	
	0.000	0.000	0.029	0.000	0.000	0.000	0.002	0.000	0.000	0.011	0.040	
4	41	0	0	0	0	38	2	0	15	0	0	96
	121.278	7.818	7.000	1.182	12.000	65.288	6.750	11.818	5.067	9.091	8.545	
	0.427	0.000	0.000	0.000	0.000	0.396	0.021	0.000	0.156	0.000	0.000	0.091
	0.432	0.000	0.000	0.000	0.000	0.311	0.018	0.000	0.161	0.000	0.000	
	0.039	0.000	0.000	0.000	0.000	0.036	0.002	0.000	0.014	0.000	0.000	
5	32	0	1	0	0	40	1	0	12	0	0	86
	76.092	7.004	4.430	1.059	10.750	90.973	7.392	10.587	2.587	8.144	7.655	
	0.372	0.000	0.012	0.000	0.000	0.465	0.012	0.000	0.140	0.000	0.000	0.081
	0.337	0.000	0.013	0.000	0.000	0.328	0.009	0.000	0.129	0.000	0.000	
	0.030	0.000	0.001	0.000	0.000	0.038	0.001	0.000	0.011	0.000	0.000	
6	22	0	0	0	112	4	0	16	0	0	0	154
	4.790	12.542	11.229	1.896	446.886	10.691	16.625	0.462	13.562	14.583	13.708	
	0.143	0.000	0.000	0.000	0.727	0.026	0.000	0.104	0.000	0.000	0.000	0.146
	0.232	0.000	0.000	0.000	0.848	0.033	0.000	0.123	0.000	0.000	0.000	
	0.021	0.000	0.000	0.000	0.106	0.004	0.000	0.015	0.000	0.000	0.000	
7	0	53	0	13	0	0	0	1	0	0	0	67
	6.027	414.261	4.885	179.720	8.375	7.741	7.233	6.369	5.901	6.345	5.964	
	0.000	0.791	0.000	0.194	0.000	0.000	0.000	0.015	0.000	0.000	0.000	0.063
	0.000	0.616	0.000	1.000	0.000	0.000	0.000	0.008	0.000	0.000	0.000	
	0.000	0.050	0.000	0.012	0.000	0.000	0.000	0.001	0.000	0.000	0.000	
8	0	0	4	0	0	40	38	0	56	0	0	138
	12.415	11.239	3.653	1.699	17.250	36.300	35.825	16.989	158.188	13.068	12.284	
	0.000	0.000	0.029	0.000	0.000	0.290	0.275	0.000	0.406	0.000	0.000	0.131
	0.000	0.000	0.052	0.000	0.000	0.328	0.333	0.000	0.602	0.000	0.000	
	0.000	0.000	0.004	0.000	0.000	0.038	0.036	0.000	0.053	0.000	0.000	
9	0	33	0	0	20	0	0	113	0	0	0	166
	14.934	28.073	12.104	2.044	0.027	19.178	17.920	419.276	14.619	15.720	14.777	
	0.000	0.199	0.000	0.000	0.120	0.000	0.000	0.681	0.000	0.000	0.000	0.157
	0.000	0.384	0.000	0.000	0.152	0.000	0.000	0.869	0.000	0.000	0.000	
	0.000	0.031	0.000	0.000	0.019	0.000	0.000	0.107	0.000	0.000	0.000	
10	0	0	3	0	0	0	0	0	0	11	18	32
	2.879	2.606	0.190	0.394	4.000	3.697	3.455	3.939	2.818	20.960	80.593	
	0.000	0.000	0.094	0.000	0.000	0.000	0.000	0.000	0.000	0.344	0.562	0.030
	0.000	0.000	0.039	0.000	0.000	0.000	0.000	0.000	0.000	0.110	0.191	
	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.010	0.017	
11	0	0	9	0	0	0	6	0	1	0	0	16
	1.439	1.303	52.595	0.197	2.000	1.848	10.569	1.970	0.119	1.515	1.424	
	0.000	0.000	0.562	0.000	0.000	0.000	0.375	0.000	0.062	0.000	0.000	0.015
	0.000	0.000	0.117	0.000	0.000	0.000	0.053	0.000	0.011	0.000	0.000	
	0.000	0.000	0.009	0.000	0.000	0.000	0.006	0.000	0.001	0.000	0.000	
Column Total	95	86	77	13	132	122	114	130	93	100	94	1056
	0.090	0.081	0.073	0.012	0.125	0.116	0.108	0.123	0.088	0.095	0.089	

anova – Analysis of Variance on the model result



GLM Diagnostics (50-50 Split)

The residuals vs fitted plot shows a repeating striped pattern again, indicating the model still isn't capturing some non-linear behavior. The Q-Q plot shows upper tail deviations with standout outliers like 19110, 4090, and 3700, suggesting non-normal residuals.

Influential points like 1911 and 2530 appear in the leverage plot, which could be affecting model stability.

Summary and Confidence Interval:

```
===== GLM 50-50 Split =====
> summary(glm_50)

Call:
glm(formula = lastcol ~ Gender + Age + Height + Weight + SMOKE +
    MTRANS, family = gaussian(), data = trainData3)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.3556409  0.3718153  25.162  < 2e-16 ***
Gender       0.1191015  0.0371413   3.207  0.00138 **
Age          0.0267302  0.0031658   8.443  < 2e-16 ***
Height      -7.6549392  0.2256475 -33.924  < 2e-16 ***
Weight       0.0797352  0.0006747 118.171  < 2e-16 ***
SMOKE       -0.1220903  0.0935265  -1.305  0.19204
MTRANS       0.0484963  0.0154119   3.147  0.00170 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.2166708)

    Null deviance: 4100.66  on 1054  degrees of freedom
Residual deviance:  227.07  on 1048  degrees of freedom
AIC: 1389.4

Number of Fisher Scoring iterations: 2

> ci_50 <- confint(glm_50)
Waiting for profiling to be done...
> print(ci_50)
                  2.5 %      97.5 %
(Intercept)  8.62689636 10.08438553
Gender       0.04630581  0.19189720
Age          0.02052541  0.03293506
Height      -8.09720004 -7.21267828
Weight       0.07841273  0.08105766
SMOKE       -0.30539882  0.06121815
MTRANS       0.01828966  0.07870303
> cat("Intercept 95% CI (50-50):\n")
Intercept 95% CI (50-50):
> print(ci_50["(Intercept)", ])
    2.5 %    97.5 %
 8.626896 10.084386
```

b. Compute the performance measurements for each table. Calculate the accuracy, false positive, etc. and display in a table.

```
> cat("\nPerformance Metrics - 70-30 Split:\n")

Performance Metrics - 70-30 Split:
> print(metrics_70)
  Centers  Accuracy Precision    Recall  F1_Score False_Positive_Rate False_Negative_Rate
1       5 0.9069401  0.913486 0.9348958 0.9240669                0.136          0.06510417
2       7 0.9069401  0.913486 0.9348958 0.9240669                0.136          0.06510417
3       9 0.9069401  0.913486 0.9348958 0.9240669                0.136          0.06510417
4      11 0.9069401  0.913486 0.9348958 0.9240669                0.136          0.06510417
>
> cat("\nPerformance Metrics - 60-40 Split:\n")

Performance Metrics - 60-40 Split:
> print(metrics_60)
  Centers  Accuracy Precision    Recall  F1_Score False_Positive_Rate False_Negative_Rate
1       5 0.9076923 0.9184466 0.9292731 0.9238281                0.125          0.07072692
2       7 0.9076923 0.9184466 0.9292731 0.9238281                0.125          0.07072692
3       9 0.9076923 0.9184466 0.9292731 0.9238281                0.125          0.07072692
4      11 0.9076923 0.9184466 0.9292731 0.9238281                0.125          0.07072692
>
> cat("\nPerformance Metrics - 50-50 Split:\n")

Performance Metrics - 50-50 Split:
> print(metrics_50)
  Centers  Accuracy Precision Recall  F1_Score False_Positive_Rate False_Negative_Rate
1       5 0.9128788 0.9144635 0.9408 0.9274448           0.1276102              0.0592
2       7 0.9128788 0.9144635 0.9408 0.9274448           0.1276102              0.0592
3       9 0.9128788 0.9144635 0.9408 0.9274448           0.1276102              0.0592
4      11 0.9128788 0.9144635 0.9408 0.9274448           0.1276102              0.0592
```

5. Analysis of what this project helped you learn about data science.

This project has deepened our understanding of the data science process, especially the value of plotting and preprocessing. At first glance, the Obesity dataset appeared clean, but as we dove deeper, we discovered numerous categorical variables that required mapping, missing values, and the need for normalization which are all critical steps before meaningful analysis or modeling. Through pairwise plots, correlation matrices, and distribution analysis, we learned how to uncover hidden patterns and relationships, helping us select relevant features for modeling.

The part of Implementing k-means clustering taught us how unsupervised learning can reveal natural groupings within data—even without labels—and how choosing the right number of clusters (k) requires careful evaluation using methods like the elbow method and between-cluster variance. Moreover, the transition into GLM prediction and classification helped bridge the gap between clustering and supervised learning. We were able to assess performance through confusion matrices, precision, recall, F1 scores, and saw how model accuracy can be impacted by both data splits and feature selection.

In conclusion, we realized that in the world of data science, the quality of our insights directly ties to the rigor of exploration. This project has equipped us with both the technical tools and the analytical mindset to approach datasets in the real world.