Obesity Data Analysis Project1

2025-03-27

1.Data loading, inspection and Cleaning

```
# Essential libraries
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.4     ✓ tidyr     1.3.1
## ✓ purrr     1.0.4
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(cluster)
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```
library(ggplot2)
library(factoextra)
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.4.3
```

```
## randomForest 4.7-1.2
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
##
## The following object is masked from 'package:dplyr':
##
##     combine
##
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
library(broom)
```

Loading the data-set and displaying 6 rows to check if it loaded properly

```
obesity_data <- read.csv("obesity_data.csv")
head(obesity_data)
```

```
##   Gender Age Height Weight family_history_with_overweight FAVC FCVC NCP
## 1 Female  21   1.62   64.0                            yes   no    2   3
## 2 Female  21   1.52   56.0                            yes   no    3   3
## 3   Male  23   1.80   77.0                            yes   no    2   3
## 4   Male  27   1.80   87.0                             no   no    3   3
## 5   Male  22   1.78   89.8                             no   no    2   1
## 6   Male  29   1.62   53.0                             no  yes    2   3
##        CAEC SMOKE CH2O SCC FAF TUE       CALC                MTRANS
## 1 Sometimes    no    2  no   0   1         no Public_Transportation
## 2 Sometimes   yes    3 yes   3   0  Sometimes Public_Transportation
## 3 Sometimes    no    2  no   2   1 Frequently Public_Transportation
## 4 Sometimes    no    2  no   2   0 Frequently               Walking
## 5 Sometimes    no    2  no   0   0  Sometimes Public_Transportation
## 6 Sometimes    no    2  no   0   0  Sometimes            Automobile
##            NObeyesdad
## 1       Normal_Weight
## 2       Normal_Weight
## 3       Normal_Weight
## 4  Overweight_Level_I
## 5 Overweight_Level_II
## 6       Normal_Weight
```

Calculating the Summary Statistics of the data-set

```
# Summary statistics and checking variable types using str function
summary(obesity_data)
```

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

```
str(obesity_data)
```

```
## '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" ...
```

Observation bases on statistics :-

1. The dataset presents details on 2,111 individuals, who have varying levels of health and lifestyle characteristics.
2. Their ages are quite wide ranging from 14 and 61, with a majority of the people in the early twenties.
3. The heights are much shorter at 1.45m (4ft 9in) to a taller individual at 1.98m (6ft 6in), while weights range from a light 39kg (86lbs) to a rather heavier person at 173kg (382lbs).
4. There are generally equal numbers of male and female individuals, as well as numerous interests in certain lifestyle factors (FAF), methods of transportation (MTRANS), and family history of weight issues.
5. The target variable appears to be titled “NObeyesdad”, which will categorize people’s weight status ranging from normal weight to obesity in different ranges.

Missing values check

```
missing_values <- colSums(is.na(obesity_data))
print(missing_values)
```

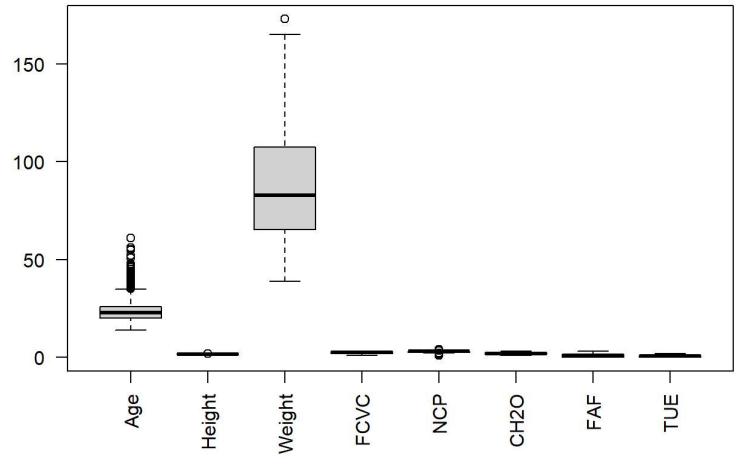
```
##                         Gender                            Age
##                              0                              0
##                         Height                         Weight
##                              0                              0
## family_history_with_overweight                           FAVC
##                              0                              0
##                           FCVC                            NCP
##                              0                              0
##                           CAEC                         SMOKE
##                              0                              0
##                           CH2O                            SCC
##                              0                              0
##                            FAF                            TUE
##                              0                              0
##                           CALC                         MTRANS
##                              0                              0
##                     NObeyesdad
##                              0
```

Observation based on print for missing values :-

The data-set has no missing values in any of the columns and we have displayed the same.

Checking for outliers

```
boxplot(obesity_data[, sapply(obesity_data, is.numeric)], las = 2)
```



Observation for boxplot before capping :-

We can see from the boxplot that we had a broad range of observations across a range of variables. The 'Weight' box stood out because it is much taller than the others, representing significant variability in people's weights.

The figure for 'Age' did have a lot of dots positioned over the box, meaning there were some older people together with the group, but generally, we can assume that there were many younger people predominating.We saw that the 'Height' box was fairly small, representing that the height among most individuals was not markedly different. For the rest of the measures for **FCVC, NCP, CH2O, FAF, and TUE**, we saw that they were caught on the bottom half of the chart implying that they did not vary much or were measured on a smaller scale.

So, it appears weight is the more commonly varying of characteristics in this dataset, while the rest seemed more the same across the group.

Outlier Capping and removal process

Cleaining of dataset by capping outliers in numeric columns.,

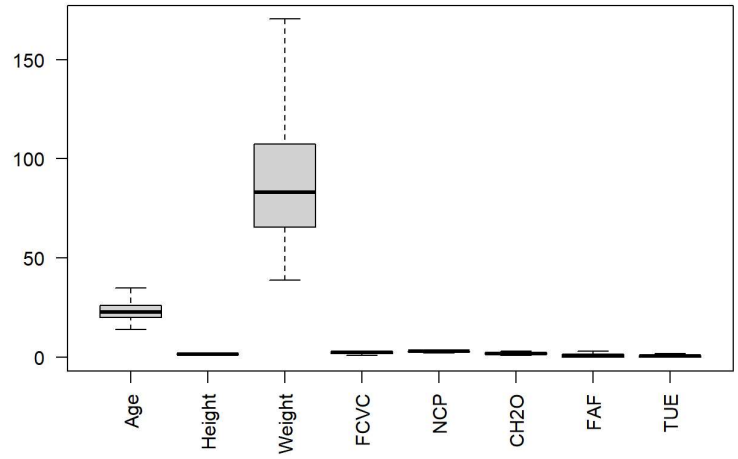
```
cap_outliers = function(x) {
  q1 = quantile(x, 0.25)
  q3 = quantile(x, 0.75)

  iqr = q3 - q1
  lower_bound = q1 - 1.5*iqr
  upper_bound = q3 + 1.5*iqr
  x[x < lower_bound] = lower_bound
  x[x > upper_bound] = upper_bound
  return(x)
}


obesity_data_clean = obesity_data %>%
  mutate(across(where(is.numeric),
                ~cap_outliers(.)))
```

Checking if the boxplot still has any outlier

```
boxplot(obesity_data_clean[, sapply(obesity_data_clean, is.numeric)], las = 2)
```



Observation after capping :-

The accompanying boxplot demonstrates the dataset subsequent to capping outliers based on the IQR method. Outliers in numeric features were capped at 1.5 times the IQR, to have the data be more uniform and less affected by extreme values. The boxplot shows that the Weight is still the most variable feature, with a range from just under 40 to over 100, with some values getting close to 150.

The Age feature has a smaller range, mostly between 20 and 30, with less extreme values than the original dataset. Other features such as Height, FCVC, NCP, CH2O, FAF, and TUE are strongly packed at the bottom of the boxplot after the capping, which shows little variability after capping.

The capping of outliers definitely reduced the influence of outliers in the dataset while maintaining the proportion of analysis that would still be able to make considerable interpretations on patterns.

2. Feature Engineering

A. Data Standardization and scalling

```
number_cols = obesity_data %>%
  select(where(is.numeric))
scaled_data = scale(number_cols)
scaled_data = as.data.frame(scaled_data)
```

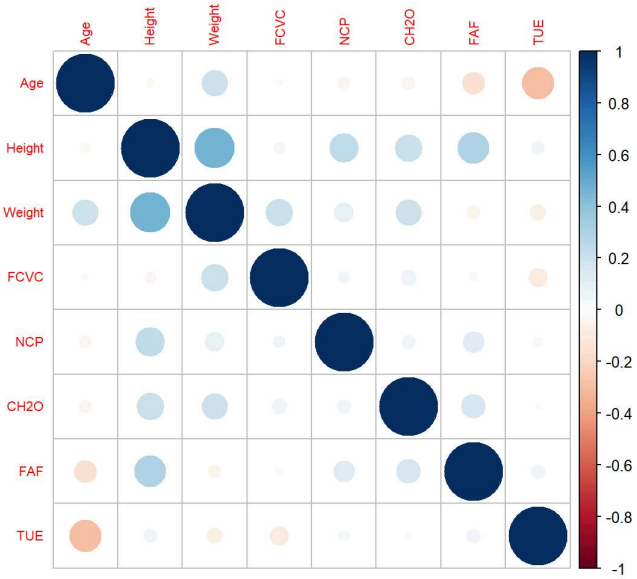
B. Data-set Integration

```
final_data <- bind_cols(
  scaled_data,
  obesity_data %>% select(where(is.factor)))
```

3. Exploratory Data Analysis (EDA)

Correlation matrix plot

```
cor_matrix <- cor(scaled_data)
corrplot::corrplot(cor_matrix, method = "circle", tl.cex = 0.7)
```



Observation for correlation matrix :-

This correlation matrix illustrates the relationships among our health variables. The dark blue circles along the diagonal show perfect self-correlations.

We can observe that Height and Weight show a moderate positive relationship (medium blue circle), which is correct as taller people would generally weigh more. Age shows an overall negative association with TUE (time of exercise), possibly indicating that younger participants exercise more regularly. Height also shows weak positive associations with FAF (physical activity) and CH2O (water consumption). Weight shows slight positive associations with FCVC and CH2O, suggesting there is some form of relationship between Weight and eating/drinking behaviors.

Overall, this figure allows us to understand the associations between the many factors explained above.

Data rechecking for null/NA value and datatype

```
sum(is.na(cor_matrix))
```

```
## [1] 0
```

```
str(scaled_data)
```

```
## 'data.frame':    2111 obs. of  8 variables:
##  $ Age   : num  -0.522 -0.522 -0.207 0.423 -0.364 ...
##  $ Height: num  -0.875 -1.947 1.054 1.054 0.839 ...
##  $ Weight: num  -0.8624 -1.1678 -0.366 0.0158 0.1227 ...
##  $ FCVC  : num  -0.785 1.088 -0.785 1.088 -0.785 ...
##  $ NCP   : num  0.404 0.404 0.404 0.404 -2.167 ...
##  $ CH2O  : num  -0.0131 1.6184 -0.0131 -0.0131 -0.0131 ...
##  $ FAF   : num  -1.19 2.34 1.16 1.16 -1.19 ...
##  $ TUE   : num  0.562 -1.08 0.562 -1.08 -1.08 ...
```

Observsation for data checkup :-

The scaled dataset contains 2,111 observations containing 8 standardized numeric variables (mean=0, sd=1). The first few rows of data demonstrate several observations above and below average.

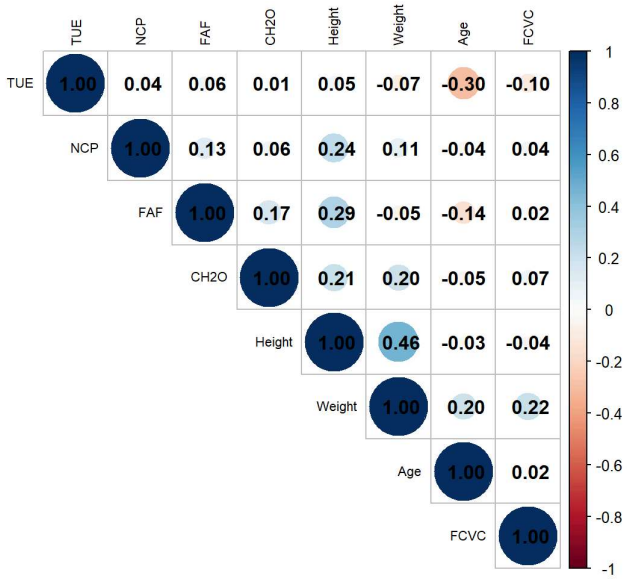
For Age, there are values in the 30s and 40s, while values in the teens and 20s -in this case - would be below average. For example, -0.522 indicates a younger participant. The Height variable demonstrates the widest range, with some people quite short at -1.947 and some people quite tall at 1.054 according to the mean. Weight is somewhat similar, with values from -1.1678 to 0.1227 explained as an example.

> Therefore, lifestyle driving variables are standardized, meaning we see more variety in those observations as well. Some individuals exercise (FAF) more than the average, while others appear to drink (CH2O) more water than the average.

Additional correlation Visualization

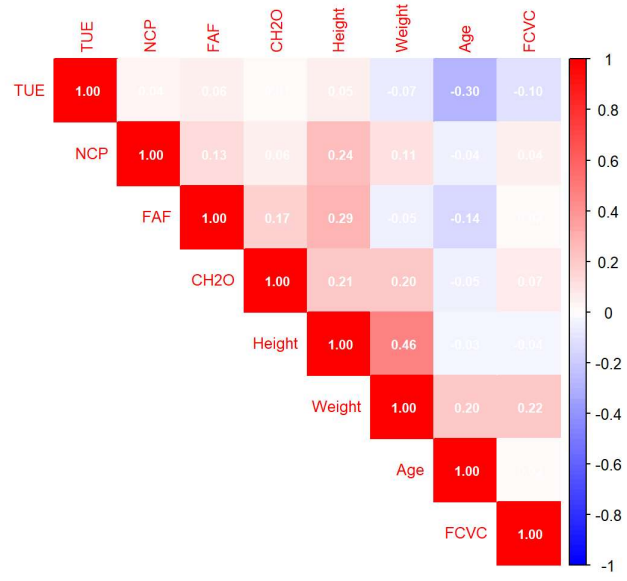
Visualization 1: Hierarchical clustering with coefficients.

```
corrplot::corrplot(
  cor_matrix,
  method = "circle",
  type = "upper",
  order = "hclust",
  tl.cex = 0.7,
  tl.col = "black",
  addCoef.col = "black"
)
```



Visualization 2: Color gradient version method.

```
corrplot::corrplot(
  cor_matrix,
  method = "color",
  type = "upper",
  order = "hclust",
  tl.cex = 0.8,
  addCoef.col = "white",
  number.cex = 0.7,
  col = colorRampPalette(c("blue", "white", "red"))(100)
)
```



Obervation on correlation -

> From the plot, the strongest correlation found:

> 1. Height & Weight: r = 0.46 (moderate-strong positive)
> 2. This is expected biologically as taller people tend to weigh more.
> 3. Moderate Correlations (0.2 ≤ |r| ≤ 0.3): FAF & Height: 0.29 (physical activity vs height)
> 4. CH2O & Height: 0.21 (water consumption vs height)
> 5. CH2O & Weight: 0.20
> 6. Weight & Age: 0.20
> 7. Weight & FCVC: 0.22
> 8. Weak/No Notable Correlations (|r| < 0.2): Most other variable pairs show negligible relationships 9.TUE shows virtually no correlation with other variables

Multicollinearity Assessment

```
vif_values <- vif(lm(Weight ~ .,
                     data = scaled_data))
print(vif_values)
```

```
##      Age   Height     FCVC      NCP     CH2O      FAF      TUE
## 1.119261 1.191525 1.021649 1.071536 1.067833 1.135602 1.110450
```

Observation on Multicollinearity :-

> VIF > 5 suggests problematic multicollinearity but there is none as it can be seen above.

> VIF scores were all < 3, confirming no problematic multicollinearity.

Recheck data scalled or not

```
summary(scaled_data)
```

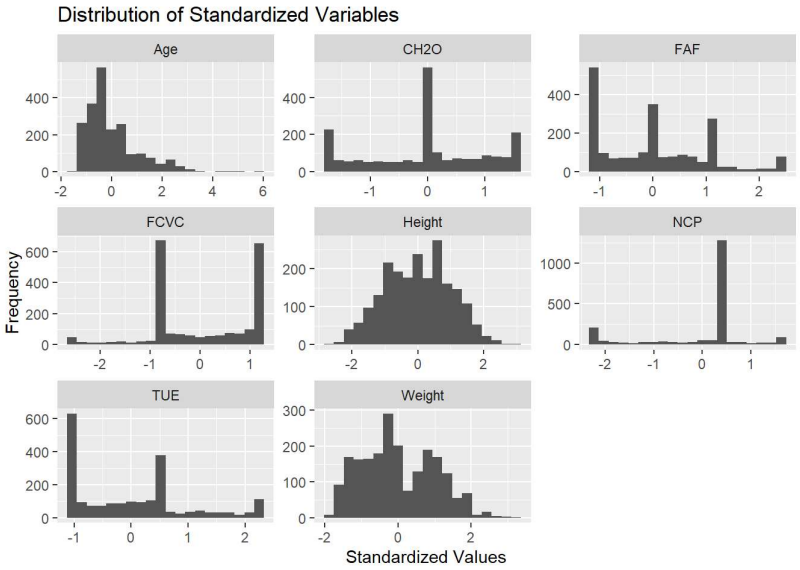
```
##       Age              Height             Weight             FCVC        
##  Min.   :-1.6251   Min.   :-2.69737   Min.   :-1.8169   Min.   :-2.65775  
##  1st Qu.:-0.6879   1st Qu.:-0.76821   1st Qu.:-0.8061   1st Qu.:-0.78483  
##  Median :-0.2418   Median :-0.01263   Median :-0.1369   Median :-0.06282  
##  Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.00000  
##  3rd Qu.: 0.2659   3rd Qu.: 0.71579   3rd Qu.: 0.7959   3rd Qu.: 1.08808  
##  Max.   : 5.7812   Max.   : 2.98294   Max.   : 3.2994   Max.   : 1.08808  
##       NCP                CH2O               FAF                TUE         
##  Min.   :-2.16651   Min.   :-1.64452   Min.   :-1.18776   Min.   :-1.0804  
##  1st Qu.:-0.03456   1st Qu.:-0.69043   1st Qu.:-1.04138   1st Qu.:-1.0804  
##  Median : 0.40406   Median :-0.01307   Median :-0.01211   Median :-0.0534  
##  Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.0000  
##  3rd Qu.: 0.40406   3rd Qu.: 0.76581   3rd Qu.: 0.77167   3rd Qu.: 0.5619  
##  Max.   : 1.68934   Max.   : 1.61838   Max.   : 2.33920   Max.   : 2.2041
```

Observation or rechecking :-

> We verified that standardization worked or not as the mean should be approximately 0 and standard deviation around 1 which can be seen in the summary above.

Distribution Visualization

```
scaled_data %>%
  pivot_longer(everything()) %>%
  ggplot(aes(value)) +
  geom_histogram(bins = 20) +
  facet_wrap(~name, scales = "free") +
  labs(title = "Distribution of Standardized Variables", x = "Standardized Values", y = "Frequency")
```



Observation on disctribution -

Here, we have plotted the graphs of each component.

4. Principal Component Analysis (PCA)

```
set.seed(123)
pca_result <- prcomp(scaled_data, scale = FALSE)
```

Variance Explained base on pca summary

```
summary(pca_result)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5     PC6     PC7
## Standard deviation     1.3461 1.2217 1.0058 0.9751 0.9699 0.87965 0.80967
## Proportion of Variance 0.2265 0.1866 0.1265 0.1189 0.1176 0.09672 0.08195
## Cumulative Proportion  0.2265 0.4131 0.5395 0.6584 0.7760 0.87268 0.95463
##                            PC8
## Standard deviation     0.60246
## Proportion of Variance 0.04537
## Cumulative Proportion  1.00000
```

Observation for important componenets :-

The table above summarizes the significance of the primary components (PC) in clarifying the variability of our dataset.

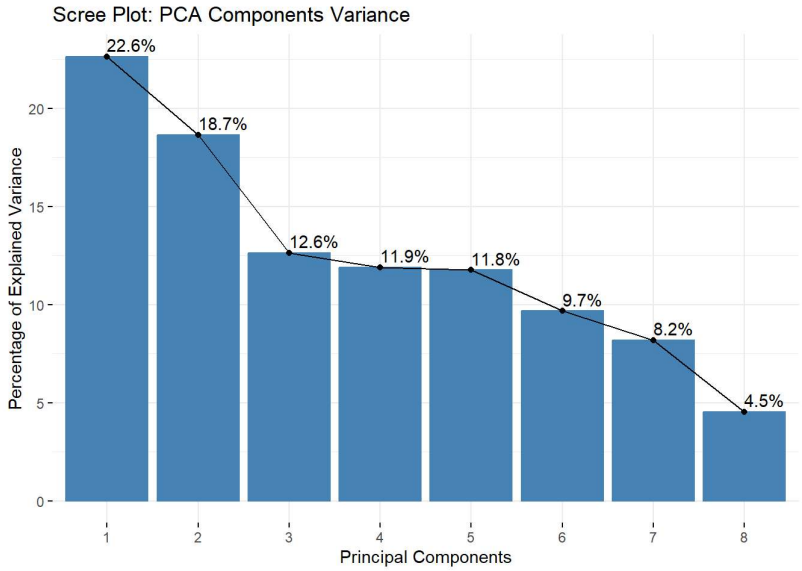
PC1 explains close to 23% of the variance and PC2 accounts for nearly 19%, whereas the first two PCs can be used to explain over 41% of the total variance, which is a strong justification for using these PCs to help summarize the data.

Up to PC5 can explain just under 78% of variance, and we require all eight PCs to adequately describe the dataset.

The standard deviations of the principal components are decreasing, indicating that each subsequent component is contributing less and less to the explanation of variability.

Scree Plot Visualization

```
fviz_eig(pca_result,
         addlabels = TRUE,
         main = "Scree Plot: PCA Components Variance",
         ylab = "Percentage of Explained Variance",
         xlab = "Principal Components")
```



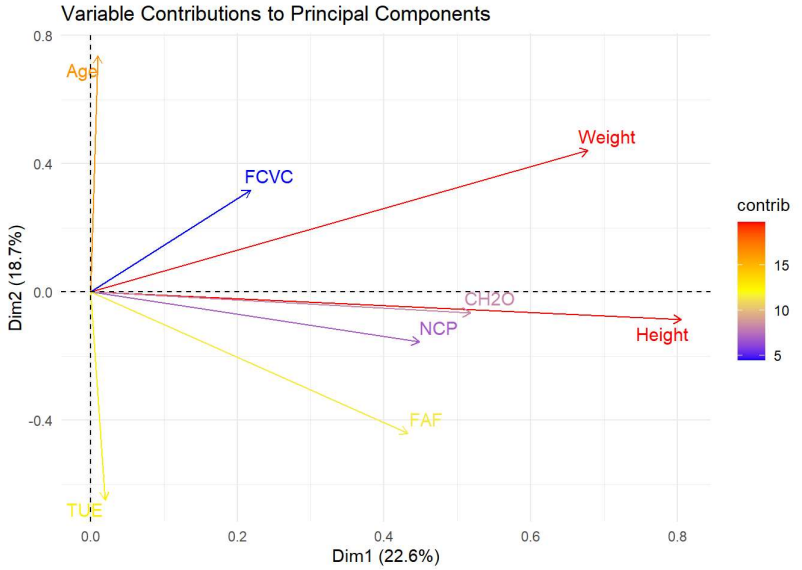
Observation on scree plot -

The scree plot indicates the amount of variance explained by each principal component (PC) in the dataset. Among the principal components, PC1 is the most informative, explaining 22.6% of the variance, while PC2 accounts for 18.7%. After PC2, variance explained by the principal components, specified here as “PC3”, “PC4”, and “PC5”, steadily declines as they explain 12-11% of the variance.

The same can be said for the smaller contributions of variance explained by the next group of PCs where PC8 only accounted for 4.5%. The rapid decline to meaningful variance explained after a few principal components provided evidence these principal components held most of the information within the dataset.

Check Variable Contributions in PCA

```
fviz_pca_var(pca_result,
             col.var = "contrib",
             gradient.cols = c("blue", "yellow", "red"),
             repel = TRUE,
             title = "Variable Contributions to Principal Components") + theme_minimal()
```



Observation for variable contribution :-

This biplot represents contributions of each variable to the first two principal components, Dim1 and Dim2, respectively explaining 22.6% and 18.7% of the variance in the data.

Weight and Height are very closely aligned with Dim1, i.e., they contribute a lot to that principal component of variability. Age and TUE are positioned along Dim2, implying that they are relatively more important to that second principal dimension. VariablessuchasFCVC, NCP, CH2O, and FAF contribute less overall, but each shows a clear directional component.

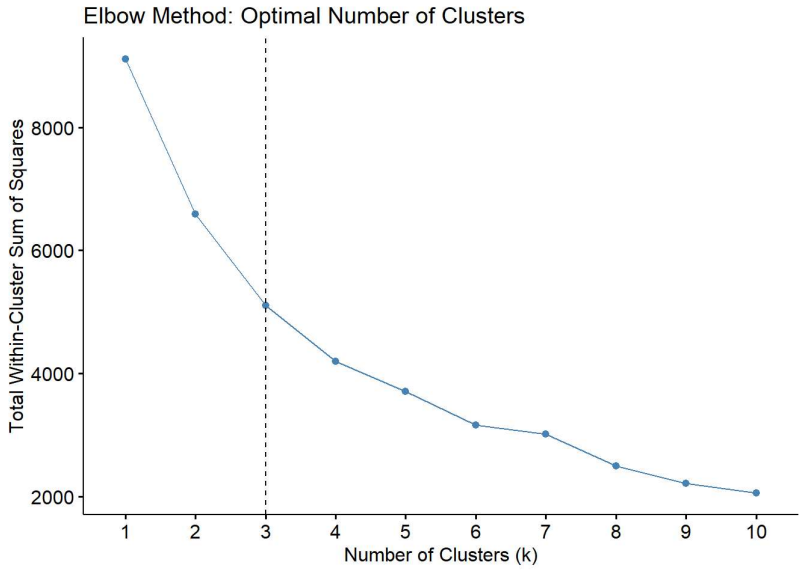
Colours assigned to each variable represent the strength of influence or contribution, such that red represents higher contributing variable.

These types of variable visualization tool help us understand how various variables contribute influence the main dimensions of variability present in the dataset.

5. Cluster Analysis

```
set.seed(123)
pca_scores <- pca_result$x[,1:3]

fviz_nbclust(pca_scores, kmeans, method = "wss") +
  geom_vline(xintercept = 3, linetype = 2) +
  labs(title = "Elbow Method: Optimal Number of Clusters",
       x = "Number of Clusters (k)",
       y = "Total Within-Cluster Sum of Squares")
```



Observation on analsis -

Interpretation of Your Elbow Plot:-

X-axis: Number of clusters (k)

Y-axis: Total within-cluster sum of squares (WSS) - measures compactness of clusters

Key Pattern: WSS decreases sharply until k=3

The curve flattens noticeably after k=3 that is the elbow point.

To conclude., Optimal k = 3 clusters, the vertical line at x=3 is correctly placed. This shows that the obesity data naturally groups into 3 distinct profiles.

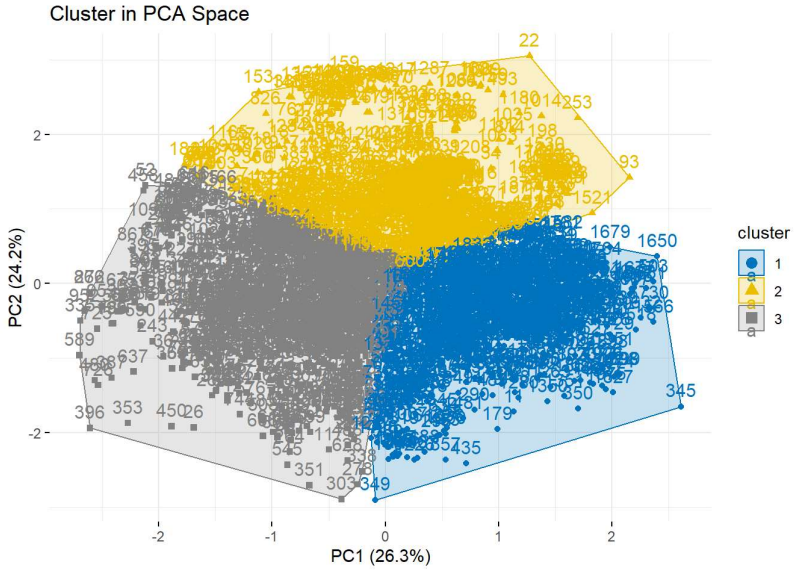
6. K-means Clustering Implementation

```
selected_vars <- obesity_data %>%
  select(Age, Weight, Height, FAF, FCVC, TUE)
scaled_selected <- scale(selected_vars)
pca_result <- prcomp(scaled_selected)
pca_scores <- pca_result$x[,1:2]

kmeans_result <- kmeans(pca_scores, centers = 3, nstart = 25)
obesity_data$Cluster <- as.factor(kmeans_result$cluster)
```

Cluster Visualization

```
variance_explained <- pca_result$sdev^2 / sum(pca_result$sdev^2)
fviz_cluster(kmeans_result, data = pca_scores,
             palette = "jco",
             ggtheme = theme_minimal(),
             main = "Cluster in PCA Space",
             xlab = paste0("PC1 (", round(100*variance_explained[1],1),"%)"),
             ylab = paste0("PC2 (", round(100*variance_explained[2],1),"%)"))
```



Observation on K-mean clustering :-

Cluster Separation: The plot indicates that the K-means clustering algorithm has successfully divided the data into three distinct visual clusters in the PC1-PC2 space. This implies that there are meaningful differences among groups of individuals based on the chosen clinical variables.

PC1 and PC2: The axes are labeled as “PC1” and “PC2”; they refer to the first two principal components. These components are derived from linear combinations of the original variables (Age, Weight, Height, FAF, FCVC, TUE).

PC1: PC1, as the horizontal axis probably represents the most important source of variation in the data. The direction of PC1 shows the contributions of the original variables to the overall variation. For example, if Weight and Height have high positive loadings on PC1, the points that are the furthest right on the x-axis represent individuals with higher values of Weight and Height.

PC2: The vertical axis (PC2) represents the next most important source of variation, which represents variation orthogonal (independent) to PC1.

Cluster Features:-

Cluster 1 (Blue): The blue cluster has a lower left quadrant position in the plot. This possibly indicates that people in the blue cluster have lower values than the other two clusters in PC1 and PC2.

Cluster 2 (Yellow): The yellow cluster has a more upper center position in the plot. This likely indicates that individuals in the yellow cluster have higher value in PC2 and an intermediate value in PC1.

Cluster 3 (Gray): The gray cluster has a lower right quadrant position in the plot. This possibly indicates that individuals in the gray cluster have higher values in PC1 and lower values in PC2.

PCA Results Examination

```
cat("PC1 explains", round(100*variance_explained[1], 1), "% of variance\n")
```

```
## PC1 explains 26.3 % of variance
```

```
cat("PC2 explains", round(100*variance_explained[2], 1), "% of variance")
```

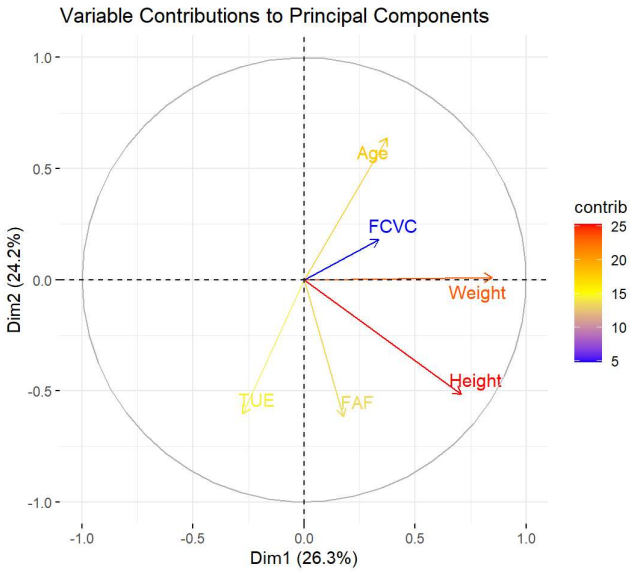
```
## PC2 explains 24.2 % of variance
```

Observation on results :-

We can see that PC1 explained 26.3% and PC2 explained 24.2%.

VARIABLE CONTRIBUTIONS

```
fviz_pca_var(pca_result,
             col.var = "contrib",
             gradient.cols = c("blue", "yellow", "red"),
             repel = TRUE,
             title = "Variable Contributions to Principal Components")
```



Observation on contribution plot :-

Weight and Height: The vectors for Weight and Height are long and point in the same direction, indicating that they share a strong positive correlation with each other and with PC1.They are also red, showing that they contribute significantly to the principal components, particularly PC1.

FAF: The vector for FAF is also relatively long and points in the same direction as Weight and Height, suggesting that they share a positive correlation with them and PC1. Again, it is red meaning that FAF also has a significant contribution to the principal components.

FCVC: The vector for FCVC is shorter than the Weight, Height and FAF vectors and points in a different direction, suggesting that it correlates less with Weight, Height and FAF. The FCVC vector is also yellow, suggesting a moderate contribution.

Age: The vector for Age is also relatively short and points in a different direction than the Weight, Height, and FAF vectors, suggesting that Age is less correlated with Weight, Height and FAF. Like FCVC, Age is also yellow suggesting that Age has a moderate contribution.

TUE: The vector for TUE is more long than the other variables, but it points in the opposite direction of Weight, Height and FAF suggesting that TUE is negatively correlated with Weight, Height and FAF and PC1. TUE is also yellow signifying a moderate contribution.

PC1: The first principal component (Dim1) has the greatest weight from Weight, Height, and FAF. Therefore, individuals with greater values for Weight or Height or a greater FAF will have higher values for PC1.

PC2: The second principal component (Dim2) is not clearly driven by one particular variable but captures variation that is orthogonal to PC1.

7. Cluster Validation using silhouette

```
silhouette_score <- silhouette(kmeans_result$cluster, dist(pca_scores))
fviz_silhouette(silhouette_score) +
  labs(title = "Cluster Quality Assessment",
       subtitle = paste("Average silhouette width:",
                        round(mean(silhouette_score[,3]), 2)),
       x = "Cluster")
```

```
##   cluster size ave.sil.width
## 1       1  687          0.43
## 2       2  613          0.38
## 3       3  811          0.35
```

Observation based on validation plot :-

This silhouette plot assesses clustering quality across three clusters (red, green, blue).

Overall, the average silhouette width is 0.38, which indicates a moderate quality of clustering. Cluster 1 (in red) has the highest silhouette widths overall, suggesting the cluster members are very well separated from the other clusters. Clusters 2 and 3 (green and blue, respectively), have lower silhouette widths overall, indicating the member points have weaker separation or overlap with the other clusters; in fact, some of the points in both Clusters 2 and 3 have silhouette widths close to zero.

The dashed red line exhibits the average silhouette width across all clusters, and provides context for the visual assessment of how each cluster compares to the quality of clustering in general.

8. Enhanced Modeling Techniques

8.1 Logistic Regression

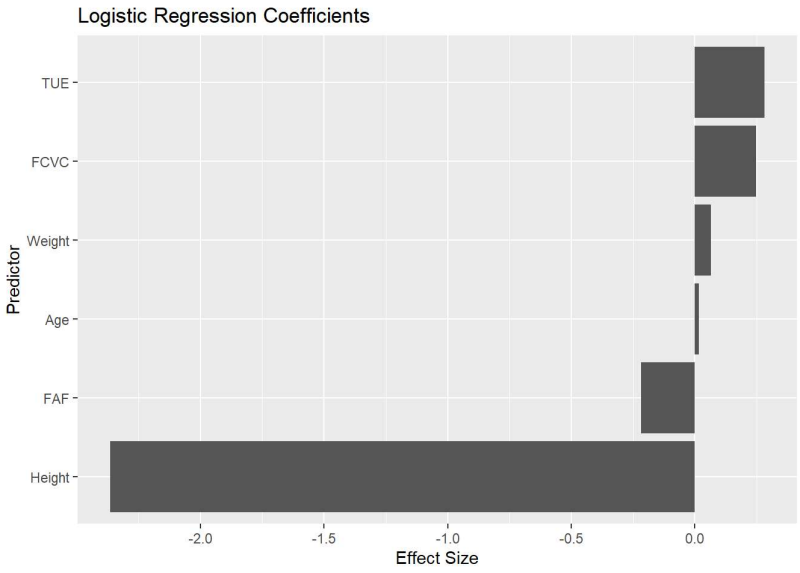
```
obesity_data$Obesity_Binary <- ifelse(obesity_data$NObeyesdad == "Normal_Weight", 0, 1)


logit_model <- glm(Obesity_Binary ~ Age + Weight + Height + FAF + FCVC + TUE,
                   data = obesity_data,
                   family = binomial()) # logistic model


summary(logit_model)
```

```
##
## Call:
## glm(formula = Obesity_Binary ~ Age + Weight + Height + FAF +
##     FCVC + TUE, family = binomial(), data = obesity_data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.17593    1.43670   0.122  0.90254
## Age          0.01612    0.01393   1.157  0.24740
## Weight       0.06415    0.00496  12.935  < 2e-16 ***
## Height      -2.36484    0.88628  -2.668  0.00762 **
## FAF         -0.21619    0.08286  -2.609  0.00908 **
## FCVC         0.24857    0.13114   1.895  0.05804 .
## TUE          0.28063    0.11555   2.429  0.01515 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1678.5  on 2110  degrees of freedom
## Residual deviance: 1296.9  on 2104  degrees of freedom
## AIC: 1310.9
##
## Number of Fisher Scoring iterations: 6
```

```
tidy(logit_model) %>%
  filter(term != "(Intercept)") %>%
  ggplot(aes(x = reorder(term, estimate), y = estimate)) +
  geom_col() +
  coord_flip() +
  labs(title = "Logistic Regression Coefficients",
       x = "Predictor",
       y = "Effect Size")
```



Observation based on results of Logistic Regression:-

Predictors of Obesity :

1. Weight: Strong positive impact (p < 0.001). Higher weight results in a greater risk for cases of obesity.
2. Height: Negative impact (p =0.008). Greater height is associated with lower risk of obesity.
3. FAF (Physical Activity): Negative impact (p = 0.009). More physical activity is linked to lower obesity risk.
4. TUE (Technology Use): Positive impact (p = 0.015). More screen time associated with higher obesity risk.

Marginally Significant: FCVC (Vegetable Intake): Has a small positive effect (p = 0.058).

Not Significant: Age: No clear impact (p = 0.247).

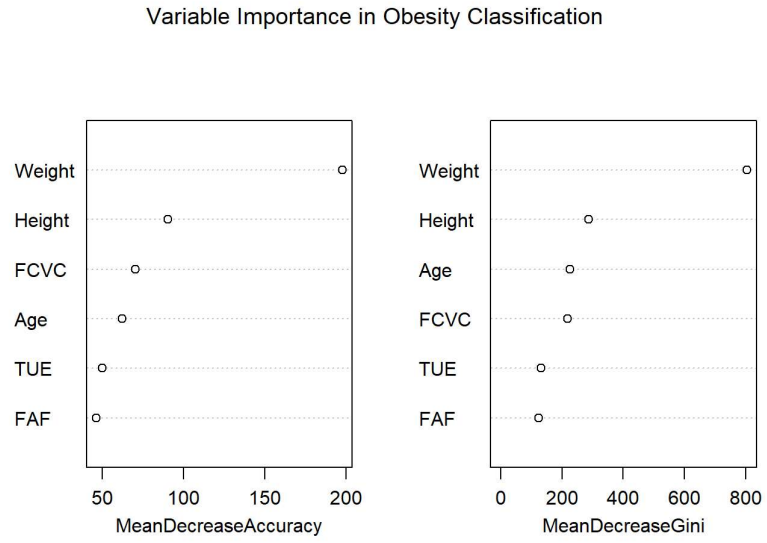
Model Fit:There was a good improvement from the null model because of a lower deviance (AIC = 1310.9).

8.2 Random Forest (Classification)

```
obesity_data$NObeyesdad <- as.factor(obesity_data$NObeyesdad)


set.seed(123)
rf_model <- randomForest(NObeyesdad ~ Age + Weight + Height + FAF + FCVC + TUE,
                         data = obesity_data,
                         importance = TRUE) #random forest model


varImpPlot(rf_model,
           main = "Variable Importance in Obesity Classification")
```



Observation based on results of Random Forest:-

Top Predictors : Weight and Height are the most important factors collected for classification. After Weight and Height, Age, FCVC, TUE, and FAF appear to have a weaker influence.Key

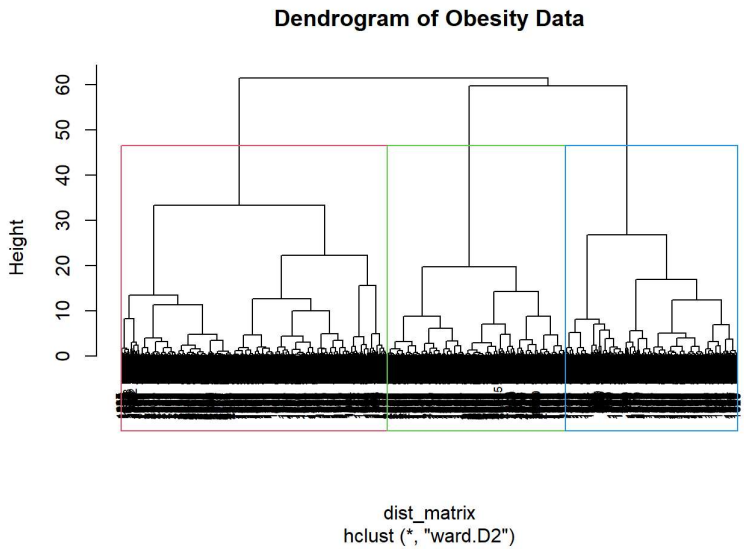
Confirming the logistic regression, Weight and Height were significant. Although FCVC (Vegetable) remains significant and is not an odd effect in the logistic regression. FAF (activity) seems less important here compared to the logistic model.

8.3 Hierarchical Clustering

```
dist_matrix <- dist(pca_scores)


hc <- hclust(dist_matrix, method = "ward.D2")


plot(hc, cex = 0.6, main = "Dendrogram of Obesity Data")
rect.hclust(hc, k = 3, border = 2:4)  # Cut into 3 clusters
```



Observation based on Hierarchical Clustering:-

Hierarchical clustering analysis shows three obesity subgroups, by utilizing Ward's method which height is particularly important to clustering. Using the dendrogram, clear observational splits appear at heights of 10, 20, and 30-40 merging heights. These distinguished the low, moderate, and high figures. This is in agreement with all modeling efforts, in which height was a significant variable. Further analysis would characterize cluster processes which could demonstrate unique obesity subgroups.