Shahjalal University of Science and Technology, Sylhet

Assignment Number :01

Course Title: Statistical Computing-II Lab

Course Code: STA356L

Submission Date:

Submitted To:

Dr. Mohammad Romel Bhuia

Professor,

Department of Statistics, Sust

Submitted By

Tanvir Hassan Ruhan

Regi:2019134115

Session 2019-2020

Department of Statistics, Sust

At first, we run the basic libraries

```
library(dplyr)

library(ggplot2)

library(psych)
```

Ans To the Question Number 01

Code of reading the csv file of (ObesityData.csv):

```
data<-read.csv(file.choose(),header = T)

View(data)

head(data)
```

I read the obesity data as data

Ans to the Question Number 02

The R code of Calculating frequencies, percentages and cumulative frequencies of all the categorical variables:

Code:

```
categorical_variables <- c("Gender", "family_history_with_overweight", "FAVC", "CAEC", "SCC", "CALC", "MTRANS","SMOKE")

# Calculate frequencies, percentages, and cumulative frequencies for categorical variables

for (var in categorical_variables ) {

 cat("\nSummary for:", var, "\n")

  freq_table <- table(data[[var]])

 percent_table <- prop.table(freq_table) * 100

 cum_freq_table <- cumsum(freq_table)

 print(data.frame(Frequency = freq_table, Percentage = percent_table, Cumulative = cum_freq_table))

}
```

Output:

Summary for: Gender

Frequency variable	Frequency	Percentage	Cumulative Frequency
Male	1043	49.40786	1043
Female	1068	50.59214	2111

Summary for: family_history_with_overweight

Frequency variable	Frequency	Percentage	Cumulative Frequency
No	385	18.2378	384
Yes	1726	81.7622	2111

Summary for: FAVC

Frequency variable	Frequency	Percentage	Cumulative Frequency
No	245	11.60587	245
Yes	1866	88.39413	2111

Summary for: CAEC

Frequency variable	Frequency	Percentage	Cumulative Frequency
Always	53	2.510658	53
Frequently	242	11.463761	295
no	51	2.415917	346
Sometimes	1765	83.609664	2111

Summary for: SMOKE

Frequency variable	Frequency	Percentage	Cumulative Frequency
no	2067	97.91568	2067
yes	44	2.08432	2111

Summary for: SCC

Frequency Variable	Frequency	Percentage	Cumulative Frequency
no	2015	95.452392	2015
yes	96	4.547608	2111

Summary for: CALC

Frequency Variable	Frequency	Percentage	Cumulative Frequency
Always	1	0.04737091	1
Frequently	70	3.31596400	71
no	639	30.27001421	710
Sometimes	1401	66.36665088	2111

Summary for: MTRANS

Frequency variable	Frequency	Percentage	Cumulative Frequency
Automobile	457	21.6485078	457
Bike	7	0.3315964	464
Motorbike	11	0.5210801	475
Public Transport	1580	74.8460445	2055
Walking	56	2.6527712	2111

Ans to the Question Number 03

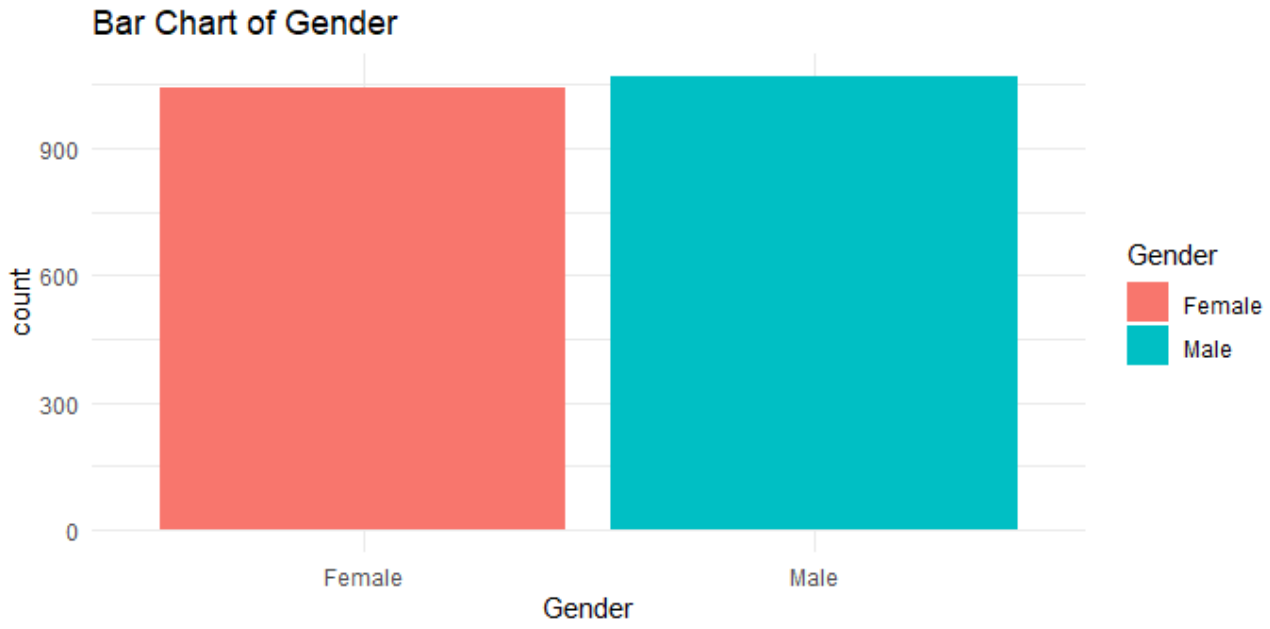
For categorical variables the suitable graph is horizontal and vertical bar diagram,

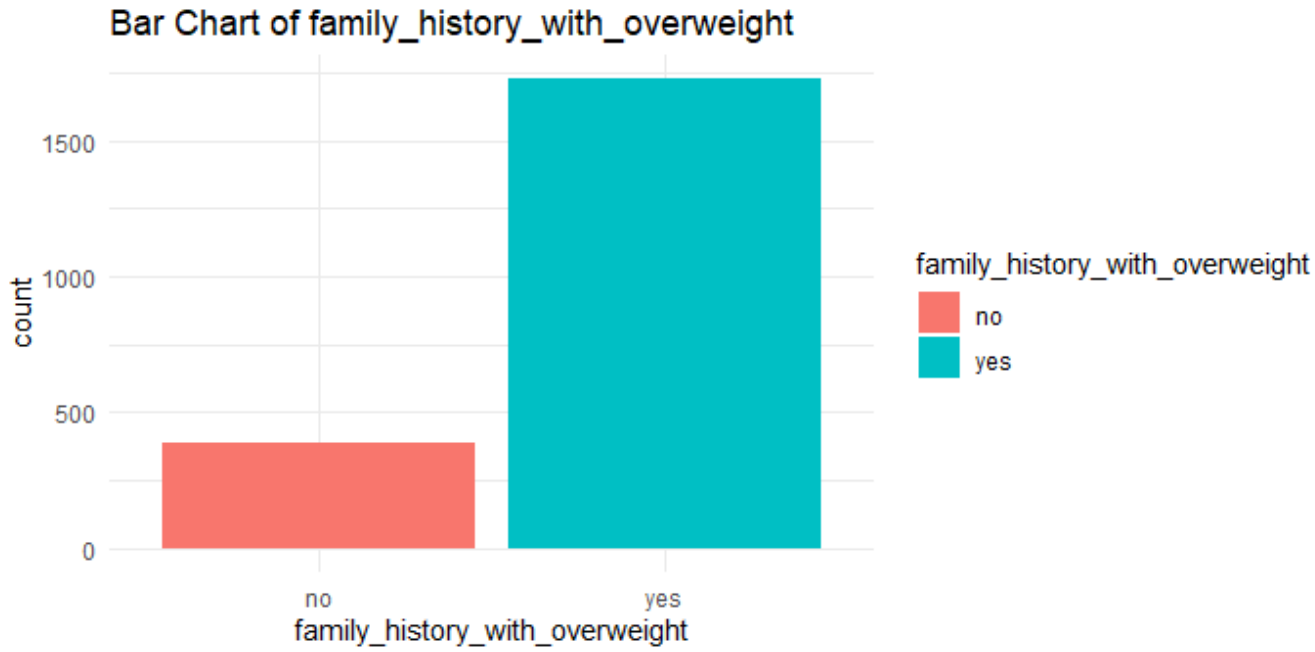
pie chart, etc. Here I select the horizontal and vertical bar diagram and pie chart

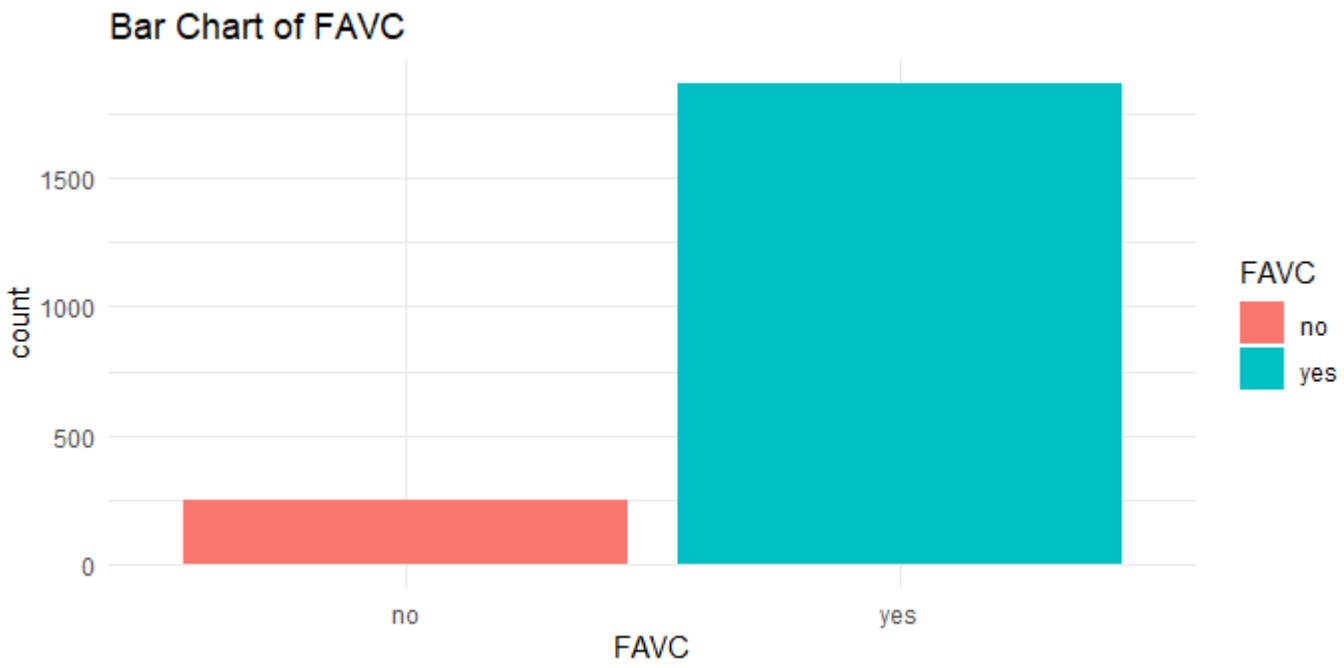
Code:

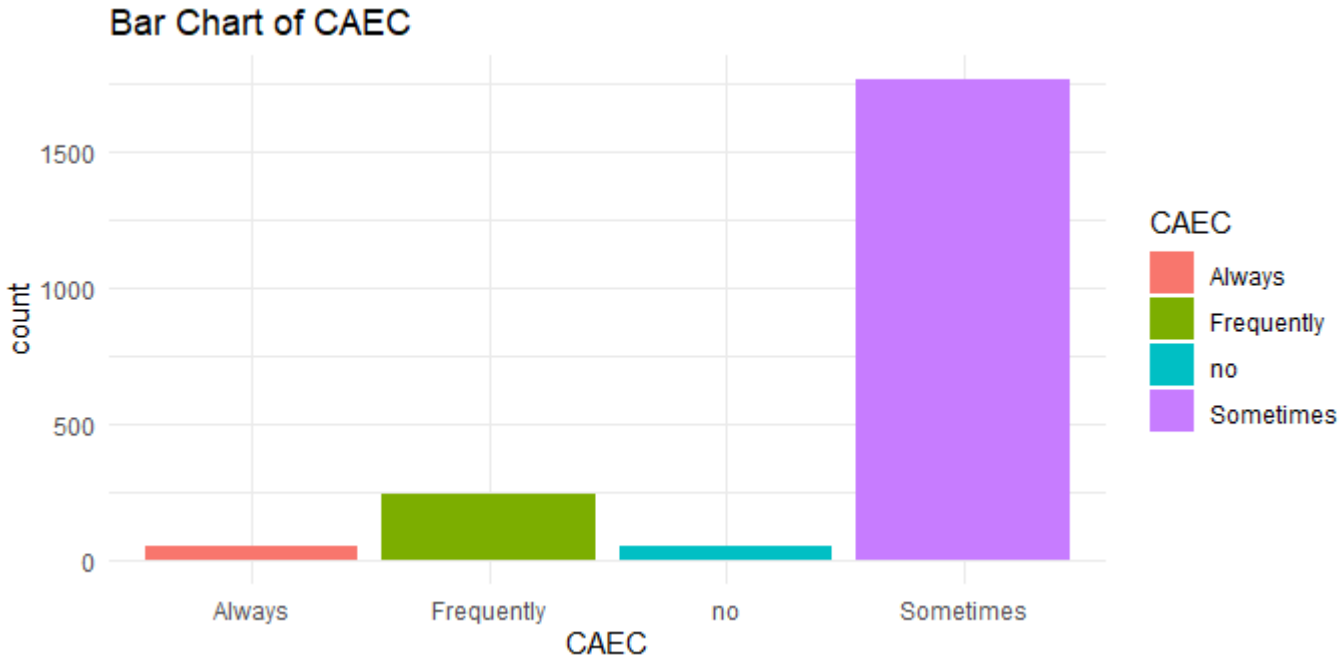
```
categorical_vars <- c("Gender", "family_history_with_overweight", "FAVC",
                      "CAEC", "SMOKE", "SCC", "CALC", "MTRANS")
for (var in categorical_vars) {
  p <- ggplot(data, aes_string(x = var, fill = var)) + geom_bar() +  ggtitle(paste("Bar Chart of", var)) + theme_minimal() +  theme(legend.position = "right")
  print(p)
}
```

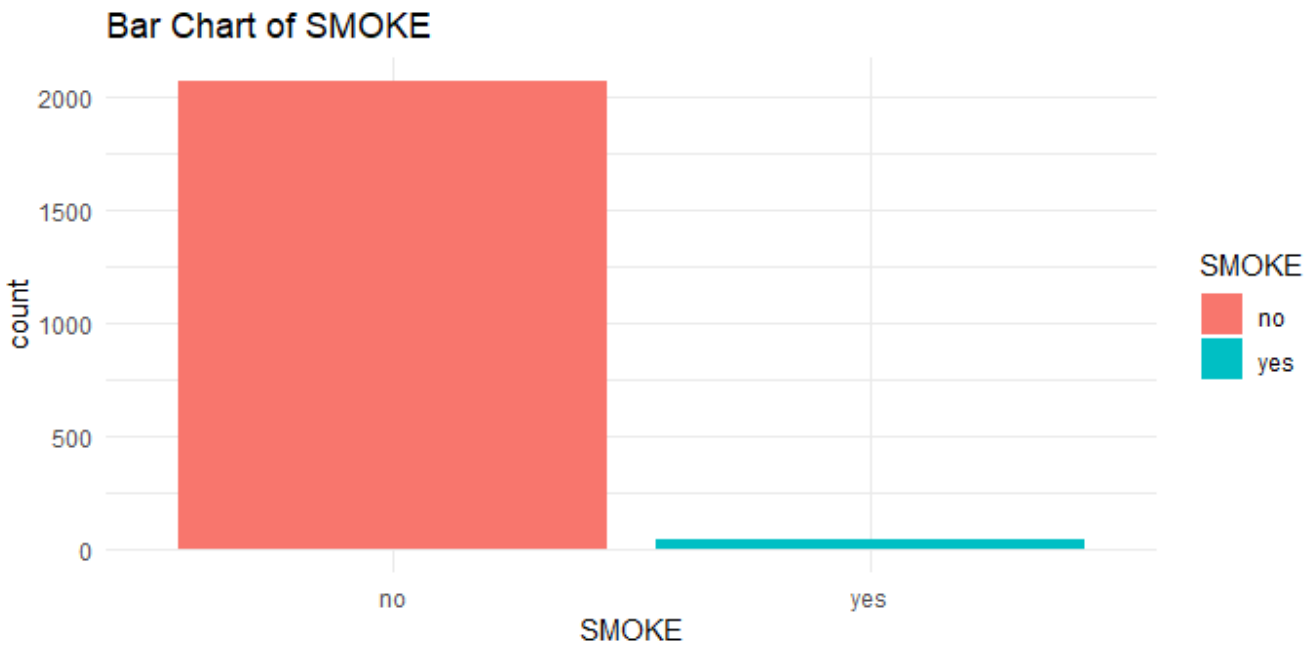
horizontal and vertical bar diagram

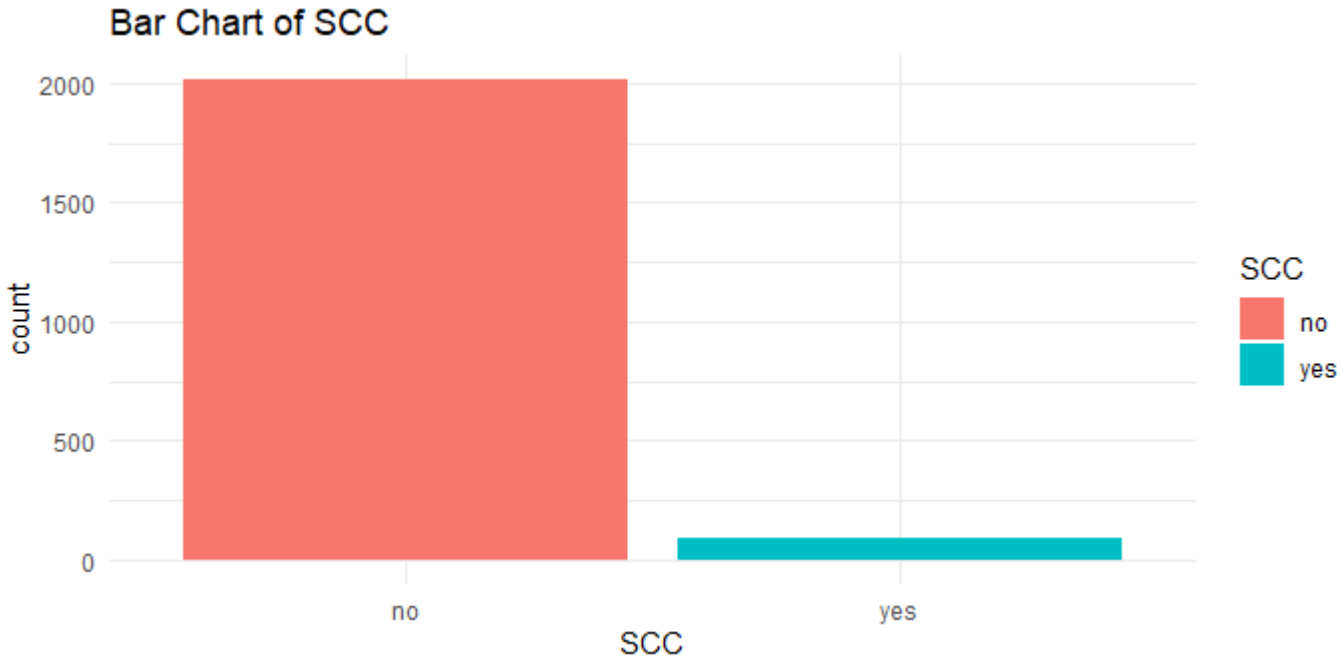


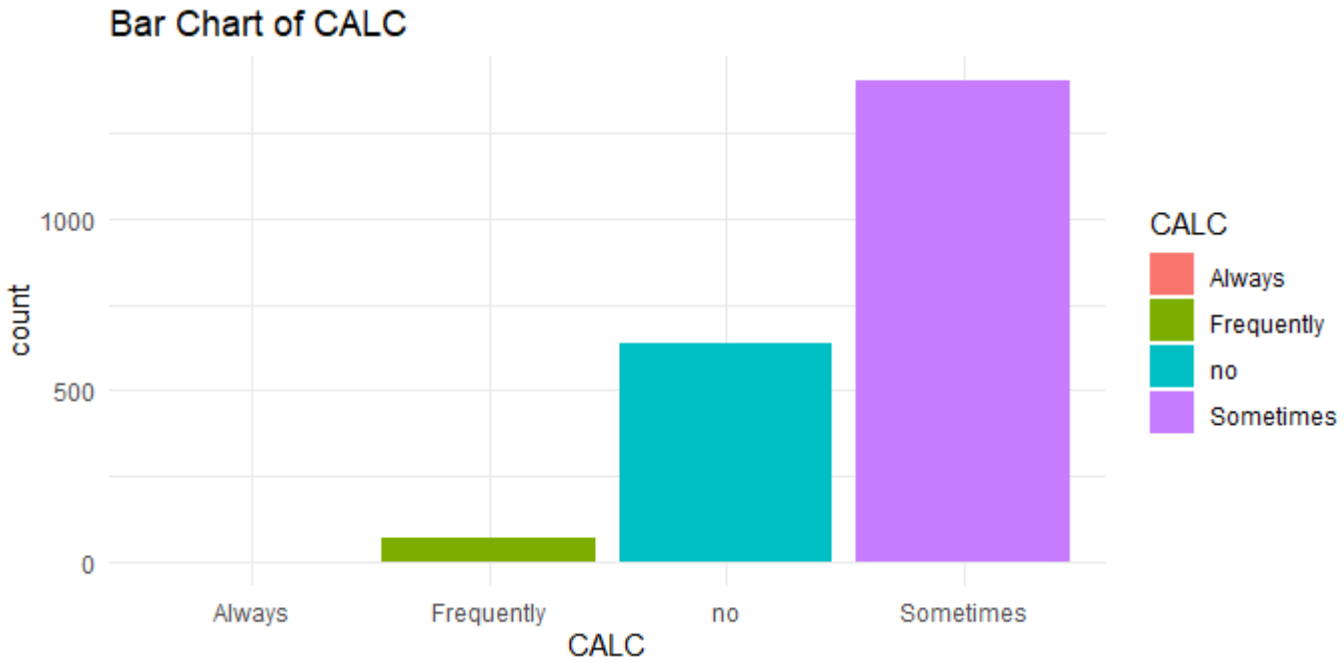


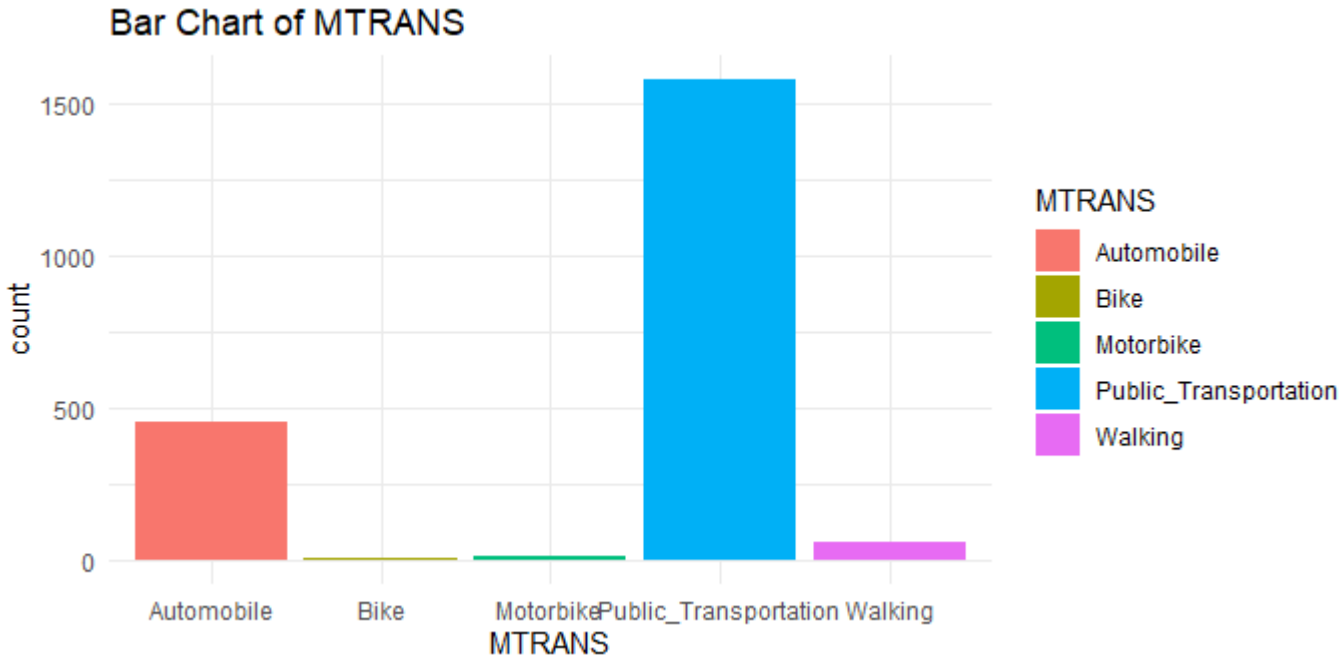

Bar Chart of FAVC
count
1500
1000
500
0
no
yes
FAVC
FAVC
no
yes


Bar Chart of CAEC
count
1500
1000
500
0
Always
Frequently
no
Sometimes
CAEC
CAEC
Always
Frequently
no
Sometimes


Bar Chart of SMOKE
2000
1500
1000
500
0
count
no
yes
SMOKE
SMOKE
no
yes


Bar Chart of SCC
2000
1500
1000
500
0
count
no
yes
SCC
SCC
no
yes


Bar Chart of CALC
count
1000
500
0
Always
Frequently
no
Sometimes
CALC
CALC
Always
Frequently
no
Sometimes

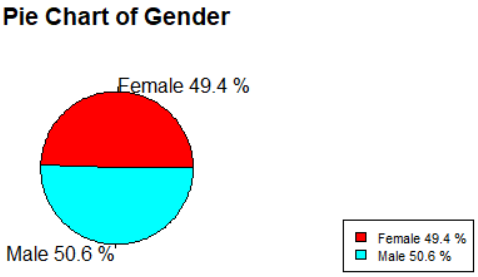

Bar Chart of MTRANS
count
1500
1000
500
0
Automobile
Bike
Motorbike
Public_Transportation
Walking
MTRANS
MTRANS
Automobile
Bike
Motorbike
Public_Transportation
Walking

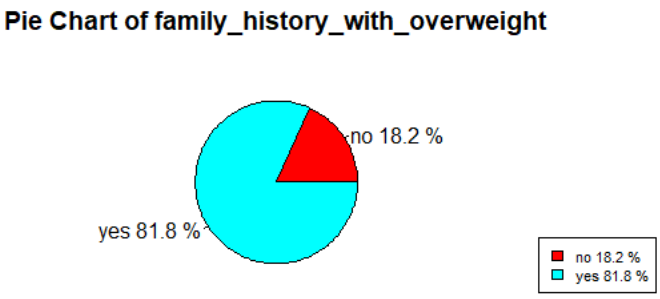
Pie chart

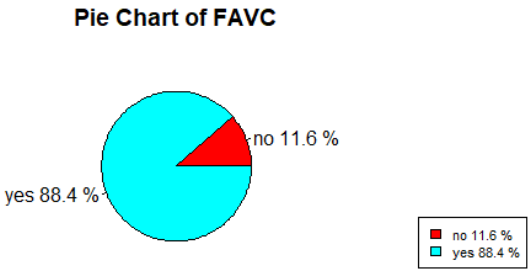
pie chart code:

```
categorical_vars <- c("Gender", "family_history_with_overweight", "FAVC",
                      "CAEC", "SMOKE", "SCC", "CALC", "MTRANS")


# Loop through each categorical variable
for (var in categorical_vars) {
  tab <- table(data[[var]])  # Frequency table
  percent <- round(prop.table(tab) * 100, 1)  # Convert to percentage
  labels <- paste(names(tab), percent, "%")  # Labels with category & percentage

  # Create Pie Chart
  pie(tab, main = paste("Pie Chart of", var), labels = labels,
      col = rainbow(length(tab)))

  # Add Legend
  legend("bottomright", legend = labels, fill = rainbow(length(tab)), cex = 0.7)
}
```

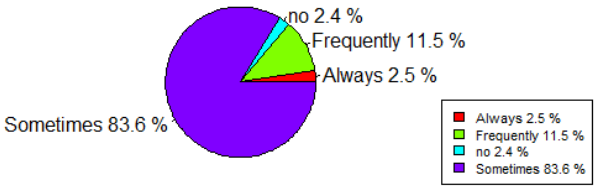
Pie charts of categorical variables





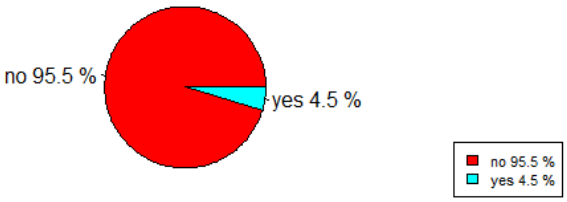


Pie Chart of CAEC

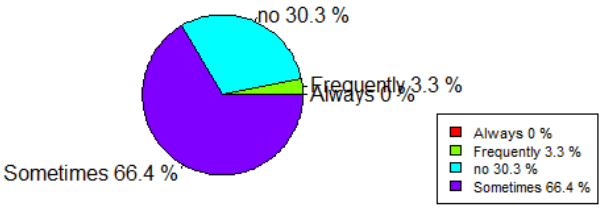


Pie Chart of SMOKE

Pie Chart of SCC



Pie Chart of CALC



Ans to the question no 04

Code:

```
numerical_variables <- c('Age','Height','Weight','FCVC','NCP','CH2O','FAF','TUE')

for (x in numerical_variables) {

 cat("\n\nSummary for variable:", x, "\n")

 z <-data[[x]]

  # Calculate statistics

 stats_df <- data.frame(

  min = min(z),

  max = max(z),

  mean = mean(z),

  median = median(z),

  mode = as.numeric(names(sort(table(z), decreasing = TRUE)[1])), # Corrected mode calculation

  q1 = quantile(z, 0.25),

  q3 = quantile(z, 0.75),

  variance = var(z),

  sd = sd(z),

  cv = sd(z) / mean(z) * 100,

  iqr = IQR(z),

  cqd = (quantile(z, 0.75) - quantile(z, 0.25)) / (quantile(z, 0.75) + quantile(z, 0.25)) # Corrected CQD calculation

 )


  print(stats_df) # Print as a data frame for better readability

}
```

Summary for variable: Age

Summary Statistics	Minimum	Maximum	Mean	Median	Mode	1st Quartile
	14	61	24.3126	22.77789	18	19.94719
Summary Statistics	3rd Quartile	Variance	Standard deviation	coefficient of variation	interquartile range	coefficient of quartile deviation,
	26	40.27131	6.345968	26.10156	6.052808	0.131734

Summary for variable: Height

Summary Statistics	Minimum	Maximum	Mean	Median	Mode	1st Quartile
	1.45	1.98	1.701677	1.700499	1.7	1.63
Summary Statistics	3rd Quartile	Variance	Standard deviation	coefficient of variation	interquartile range	coefficient of quartile deviation,
	1.768464	0.008705789	0.09330482	5.483109	0.138464	0.04074311

Summary for variable: Weight

Summary Statistics	Minimum	Maximum	Mean	Median	Mode	1st Quartile
	39	173	86.58606	83	80	65.47334
Summary Statistics	3rd Quartile	Variance	Standard deviation	coefficient of variation	interquartile range	coefficient of quartile deviation,
	107.4307	685.9775	26.19117	30.24872	41.95734	0.2426626

Summary for variable: FCVC

Summary Statistics	Minimum	Maximum	Mean	Median	Mode	1st Quartile
	1	3	2.419043	2.385502	3	2
Summary Statistics	3rd Quartile	Variance	Standard deviation	coefficient of variation	interquartile range	coefficient of quartile deviation,
	3	0.2850776	0.5339266	22.07181	1	0.2

Summary for variable: NCP

Summary Statistics	Minimum	Maximum	Mean	Median	Mode	1st Quartile
	1	4	2.685628	3	3	2.658738
Summary Statistics	3rd Quartile	Variance	Standard deviation	coefficient of variation	interquartile range	coefficient of quartile deviation,
	3	0.6053441	0.7780386	28.97045	0.341262	0.06030709

Summary for variable: CH2O

Summary Statistics	Minimum	Maximum	Mean	Median	Mode	1st Quartile
	1	3	2.008011	2	2	1.584812
Summary Statistics	3rd Quartile	Variance	Standard deviation	coefficient of variation	interquartile range	coefficient of quartile deviation,
	2.47742	0.3757119	0.6129535	30.5254	0.8926075	0.2197332

Summary for variable: FAF

Summary Statistics	Minimum	Maximum	Mean	Median	Mode	1st Quartile
	0	2	1.010298	1	0	0.124505
Summary Statistics	3rd Quartile	Variance	Standard deviation	coefficient of variation	interquartile range	coefficient of quartile deviation,
	1.666678	0.7235075	0.8505924	84.19226	1.542172	0.8609801

Summary for variable: TUE

Summary Statistics	Minimum	Maximum	Mean	Median	Mode	1st Quartile
	0	2	0.6578659	0.62535	0	0
Summary Statistics	3rd Quartile	Variance	Standard deviation	coefficient of variation	interquartile range	coefficient of quartile deviation,
	1	0.3707924	0.6089273	92.561	1	1

Ans to the Question Number 05

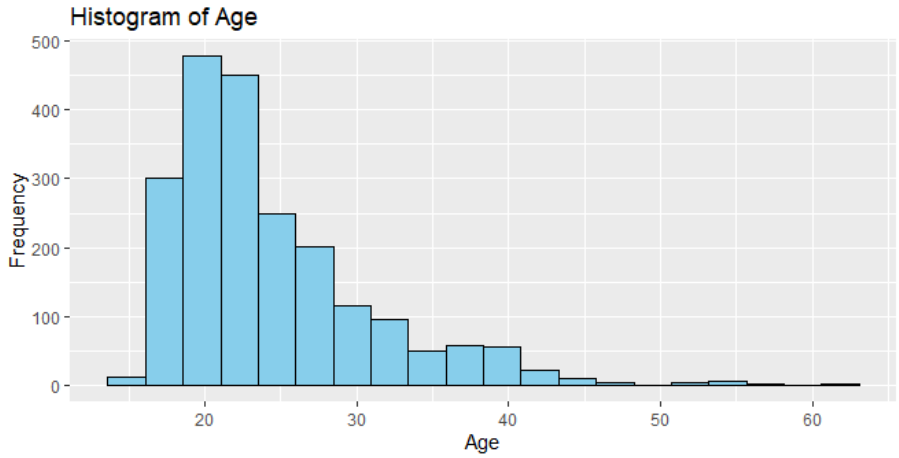
The code of Presenting the numerical variables using suitable graphs. Here we present histogram, frequency curve and histogram with normal probability curve.

Code:

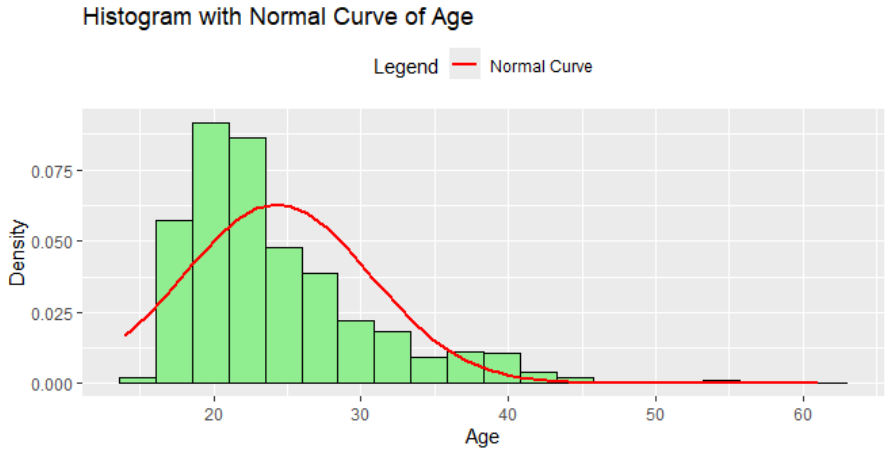
```
numerical variables <- c('Age','Height','Weight','FCVC','NCP','CH2O','FAF','TUE')  # Replace with actual numerical variables

# Loop through numerical variables

for (i in numerical_variables) {

 tab <-data[[i]]

 # Create a data frame for ggplot

 df <- data.frame(Value = tab)

 # Histogram

 p1 <- ggplot(df, aes(x = Value)) +

  geom_histogram(bins = 20, fill = "skyblue", color = "black") +

  ggtitle(paste("Histogram of", i)) +

  xlab(i) + ylab("Frequency") +

  theme(legend.position = "top")

 print(p1)

 # Histogram with normal curve

 p2 <- ggplot(df, aes(x = Value)) +

  geom_histogram(aes(y = ..density..), bins = 20, fill = "lightgreen", color = "black") +

  stat_function(fun = dnorm, args = list(mean = mean(tab, na.rm = TRUE),

                                         sd = sd(tab, na.rm = TRUE)),

                color = "red", size = 1, aes(linetype = "Normal Curve")) +

  ggtitle(paste("Histogram with Normal Curve of", i)) +

  xlab(i) + ylab("Density") +

  scale_linetype_manual(name = "Legend", values = c("Normal Curve" = "solid")) +

  theme(legend.position = "top")

 print(p2)

 # Histogram with frequency curve

 p3 <- ggplot(df, aes(x = Value)) +

  geom_histogram(bins = 20, fill = "orange", color = "black") +   geom_density(color = "blue", size = 1, aes(linetype = "Frequency Curve")) +  ggtitle(paste("Histogram with Frequency Curve of", i)) +

  xlab(i) + ylab("Frequency") +  scale_linetype_manual(name = "Legend", values = c("Frequency Curve" = "solid")) +

  theme(legend.position = "top")

 print(p3)

}
```

The histogram, frequency curve and histogram with normal probability curve is in below

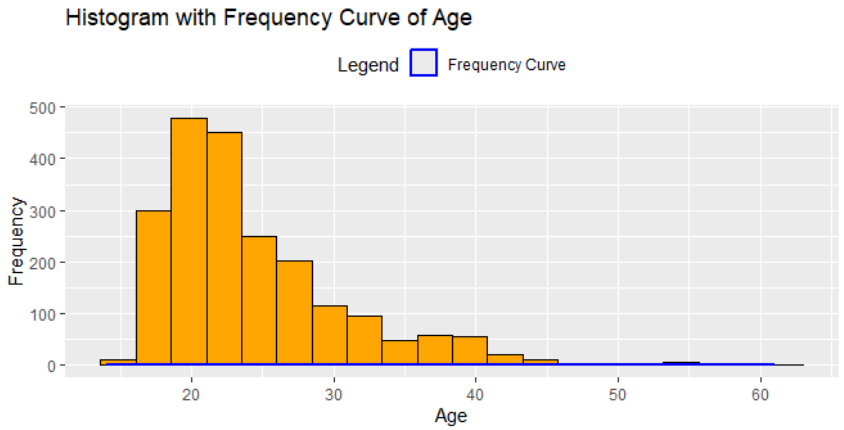
Histogram of Age Variable



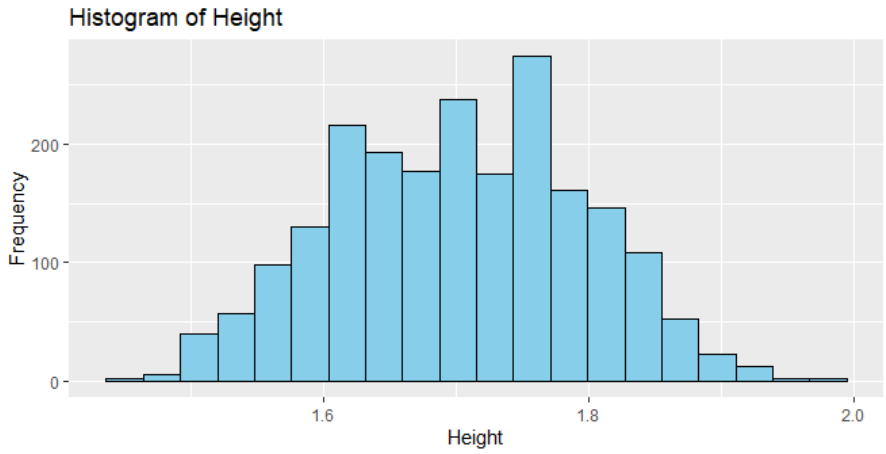
Histogram with normal probability curve of age Variable



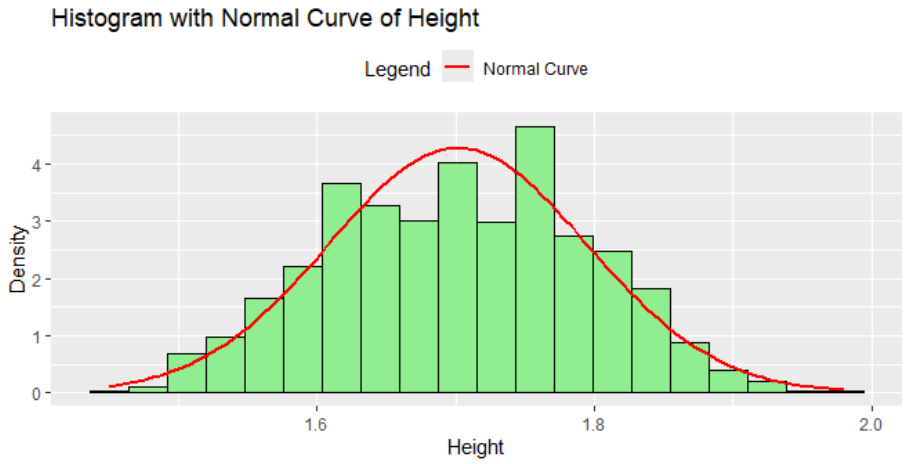
Histogram with Frequency Curve of Age



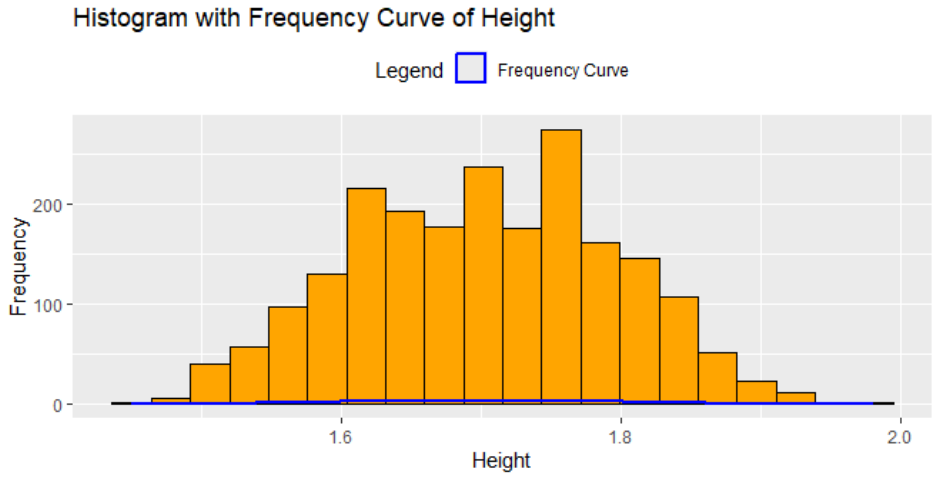
Histogram of Height



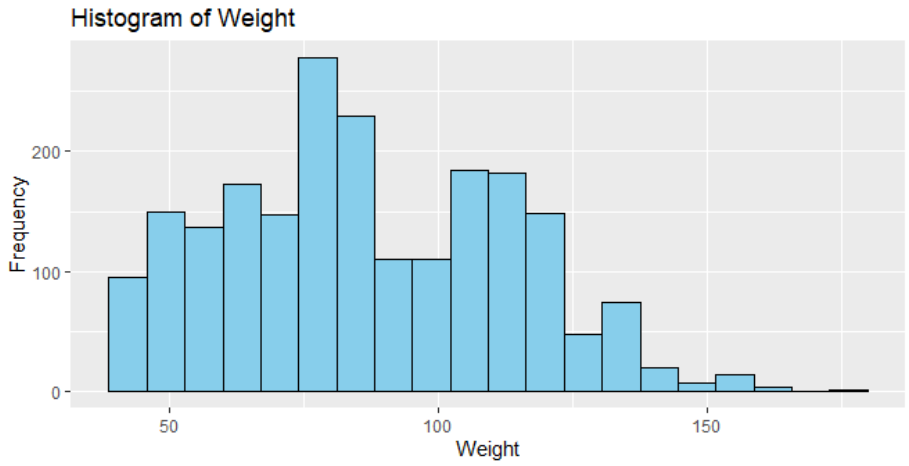
Histogram with Normal Curve of Height



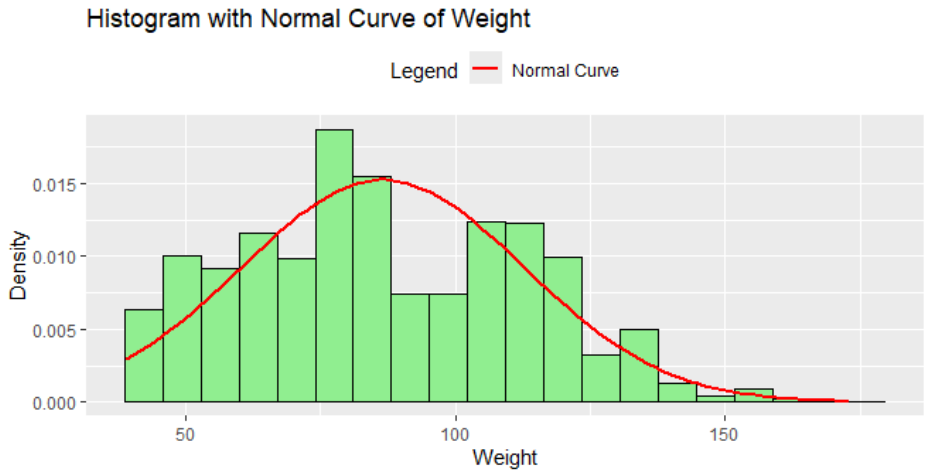
Histogram with Frequency Curve of Height



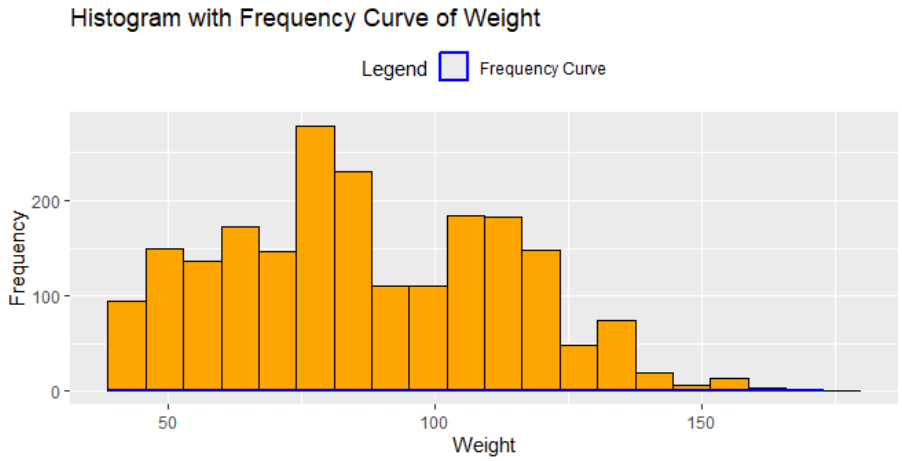
Histogram of Weight



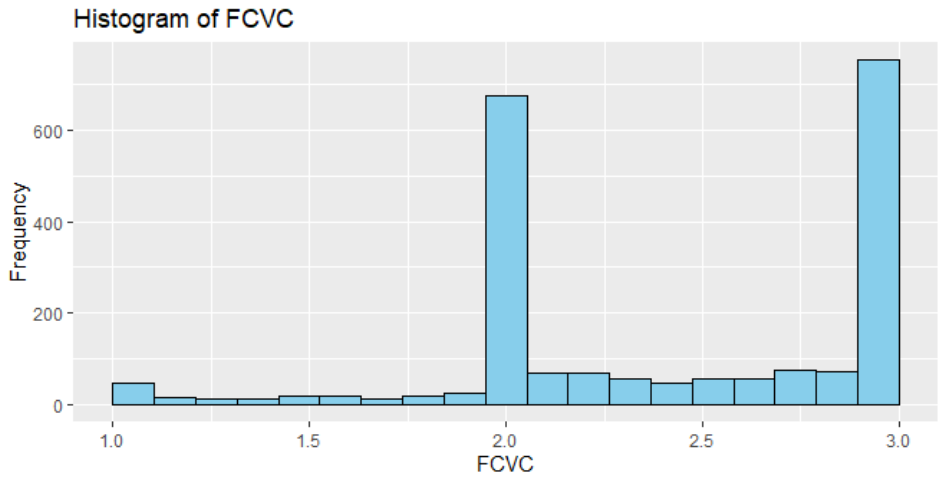
Histogram with Normal Curve of Weight



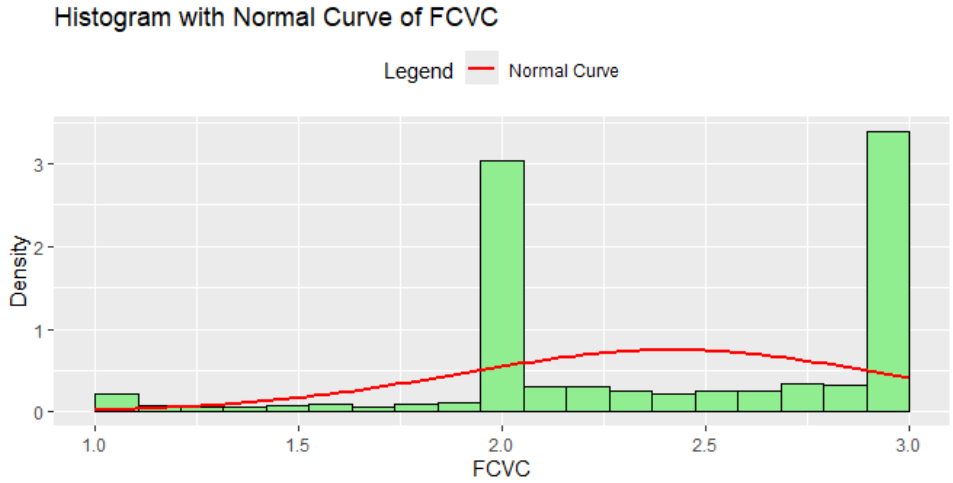
Histogram with Frequency Curve of Weight



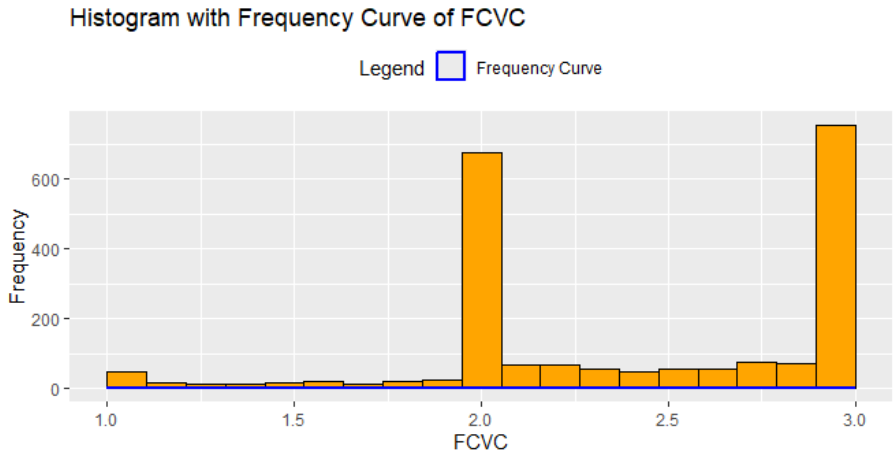
Histogram of FCVC



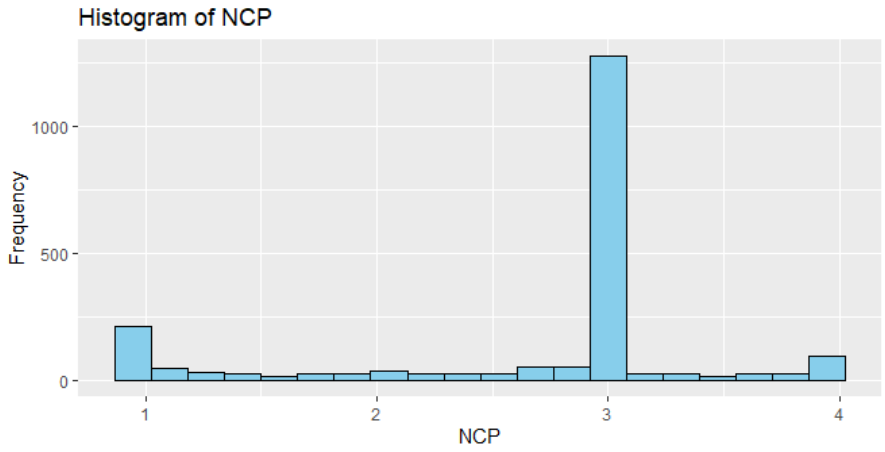
Histogram with Normal Curve of FCVC



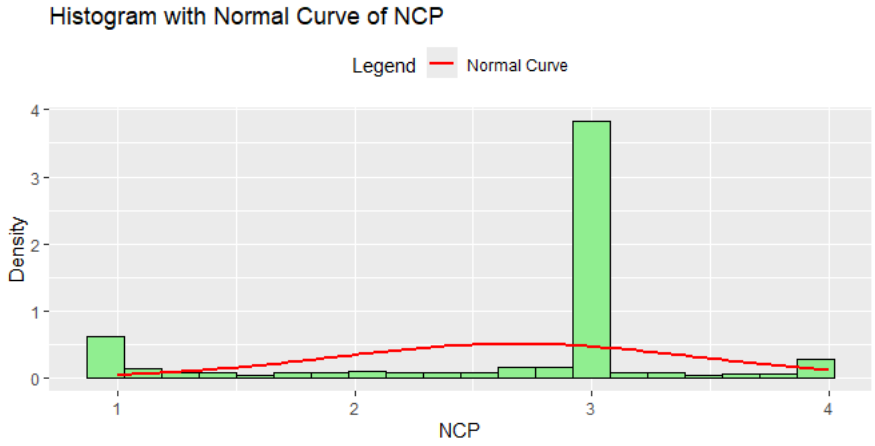
Histogram with Frequency Curve of FCVC



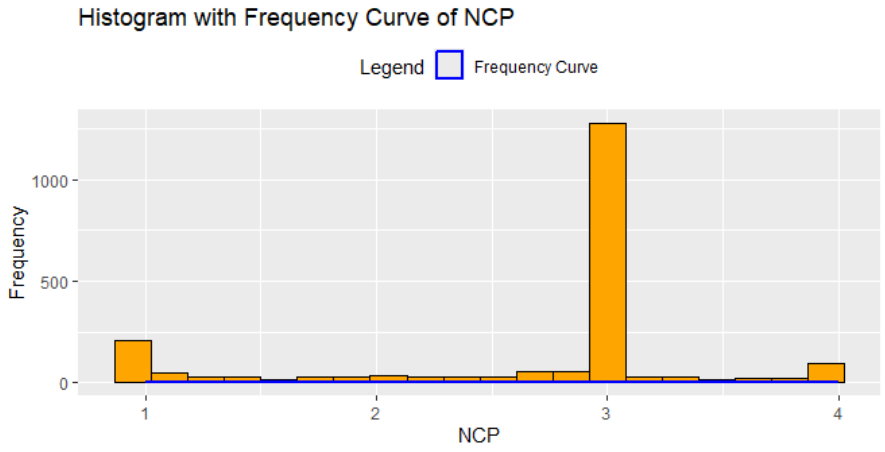
Histogram of NCP



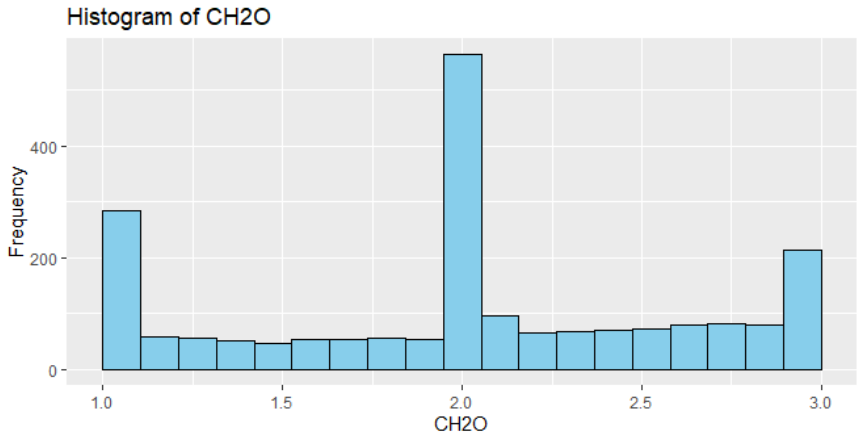
Histogram with Normal Curve of NCP



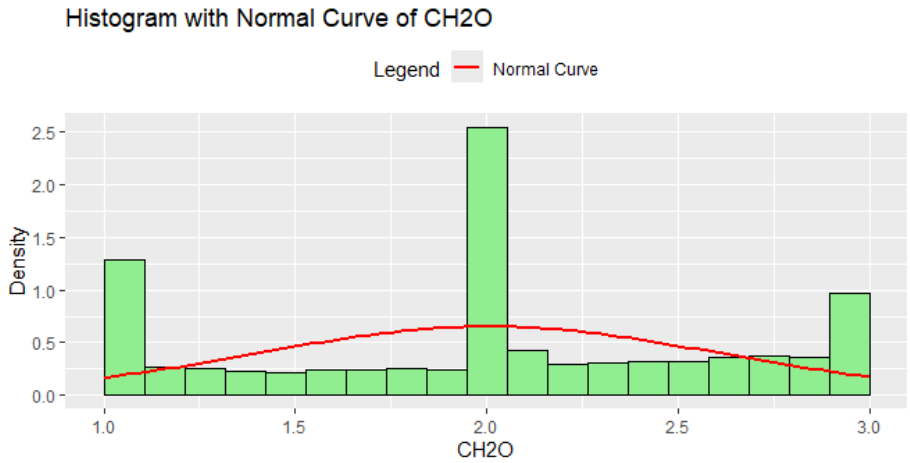
Histogram with Frequency Curve of NCP



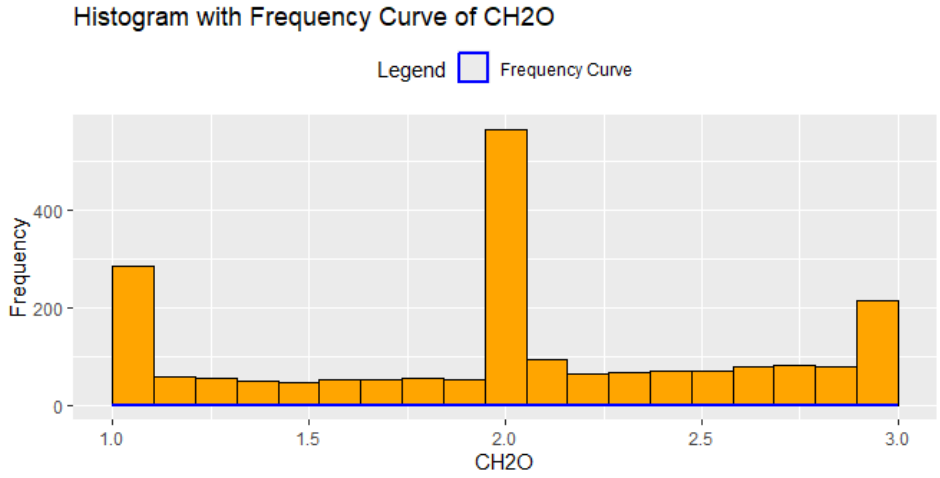
Histogram of CH2O



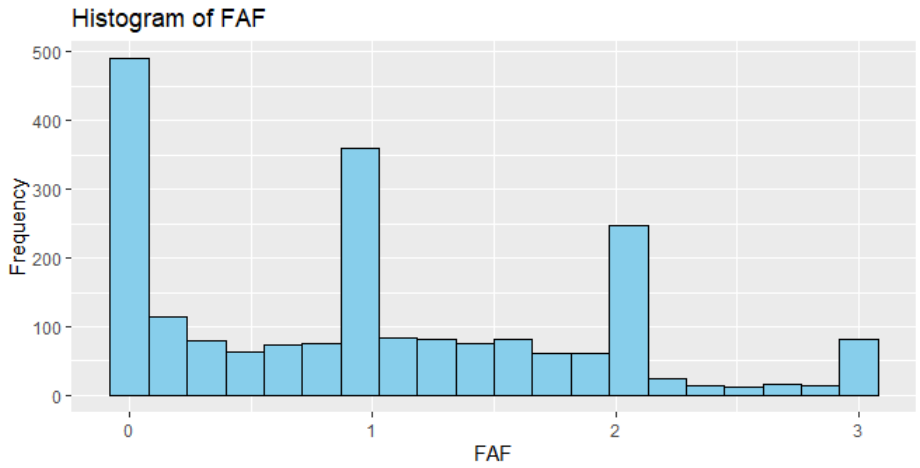
Histogram with Normal Curve of CH2O



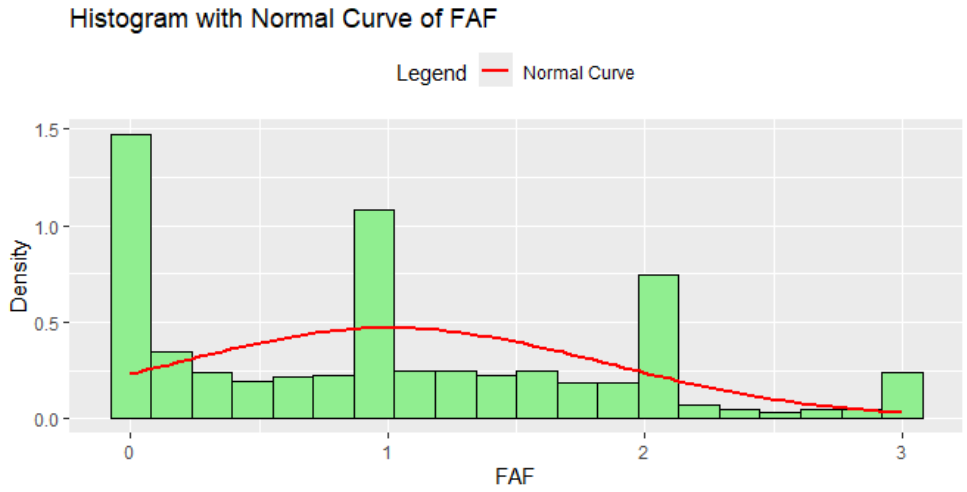
Histogram with Frequency Curve of CH2O



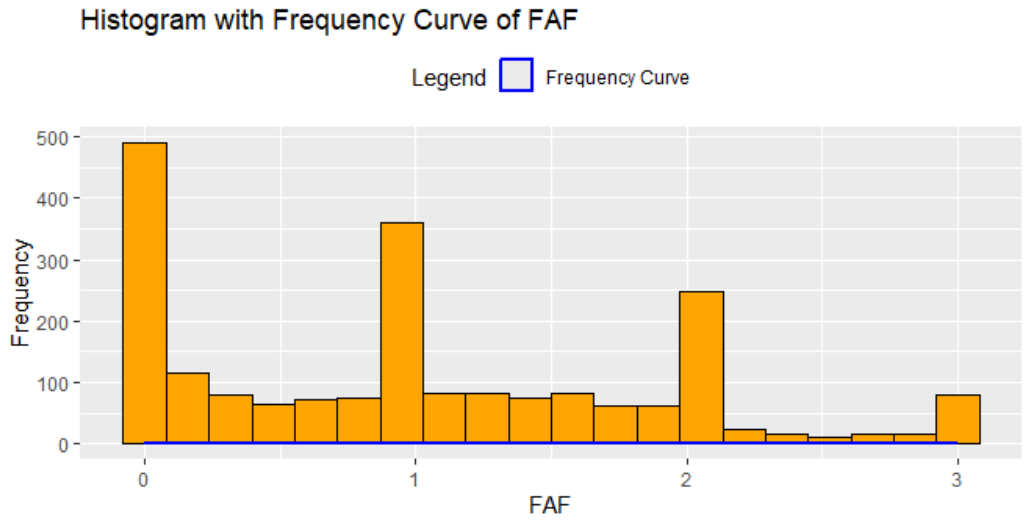
Histogram of FAF



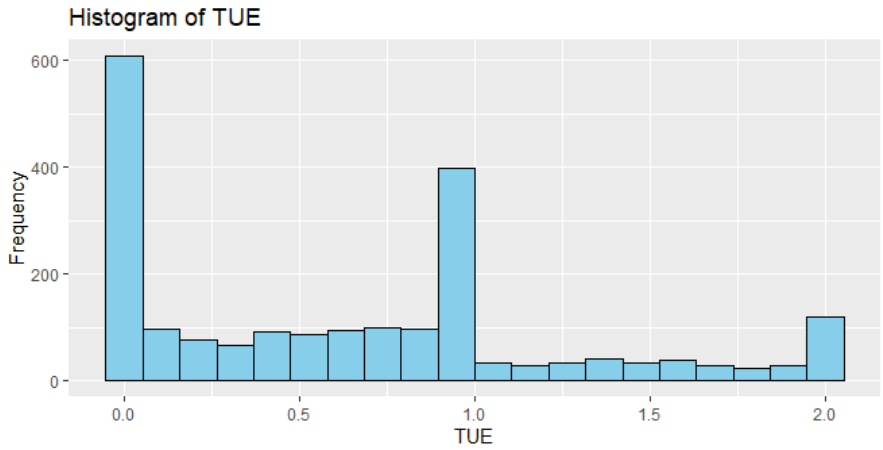
Histogram with Normal Curve of FAF



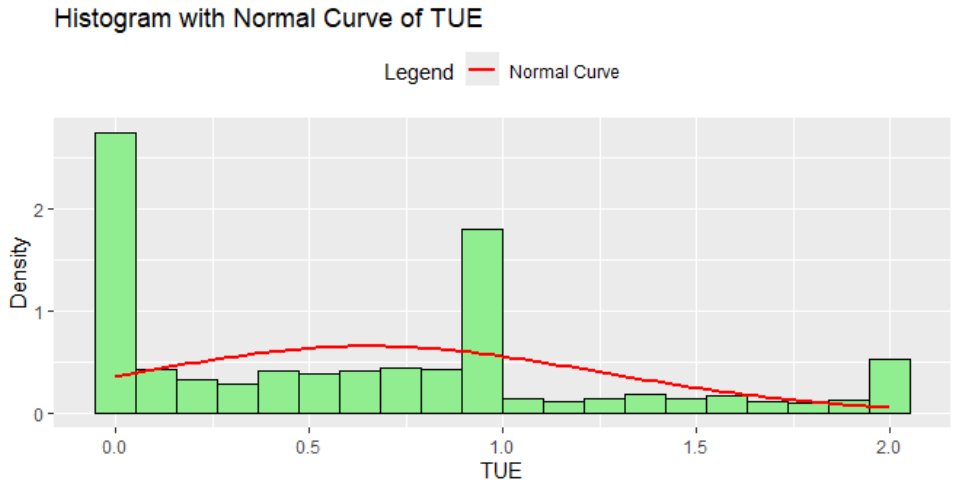
Histogram with Frequency Curve of FAF



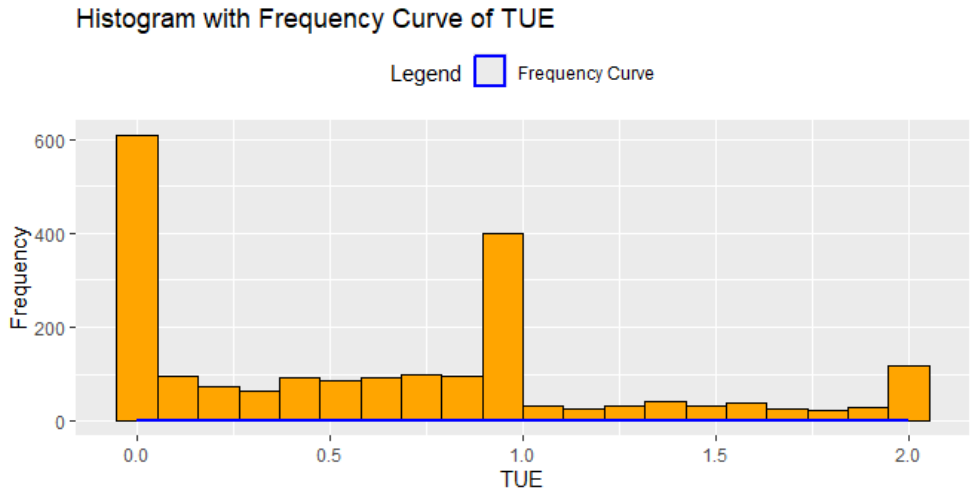
Histogram of TUE



Histogram with Normal Curve of TUE



Histogram with Frequency Curve of TUE



Ans to the Question Number 06

Code of Normality test by Shapiro wilk with 95% confidence interval

Code:

```
library(ggplot2)

install.packages("rstatix")

library(rstatix)  # For confidence interval calculation


# Normality test with 95% Confidence Interval

for (var in numerical_variables) {

 cat("\nShapiro-Wilk test for", var, "\n")

 shapiro_test_result <- shapiro.test(data[[var]])


 # Compute confidence interval

 n <- length(data[[var]])

 se <- sqrt((1.0 - shapiro_test_result$statistic^2) / (n - 1))

 lower_ci <- shapiro_test_result$statistic - 1.96 * se

 upper_ci <- shapiro_test_result$statistic + 1.96 * se


 print(shapiro_test_result)

 cat("95% CI for W-statistic:", round(lower_ci, 4), "to", round(upper_ci, 4), "\n\n")

}
```

Output:

Shapiro-Wilk normality test for Age

Test Statistic, W= 0.86606,

p-value < 2.2e-16

95% CI for W-statistic: 0.8447 to 0.8874

Comment: Here we see that for age Shapiro wilk test p value < 2.2e-16 which is also less than 0.05. So, we reject the null hypothesis. We accept alternative hypothesis. So, age is not normally distributed

Shapiro-Wilk normality test for Height

Test Statistic, W= 0.99323, p-value = 2.772e-08

95% CI for W-statistic: 0.9883 to 0.9982

Comment: Here we see that for age Shapiro wilk test p value < 2.2e-16 which is also less than 0.05. So, we reject the null hypothesis. We accept alternative hypothesis. So, Height is not normally distributed.

Shapiro-Wilk normality test for Weight

Test Statistic, W= 0.9765

p-value < 2.2e-16

95% CI for W-statistic: 0.9673 to 0.9857

Comment: Here we see that for age Shapiro wilk test p value < 2.2e-16 which is also less than 0.05. So, we reject the null hypothesis. We accept alternative hypothesis. So, Weight is not normally distributed.

Shapiro-Wilk normality test for FCVC

Test Statistic, W= 0.84491

p-value < 2.2e-16

95% CI for W-statistic: 0.8221 to 0.8677

Comment: Here we see that for age Shapiro wilk test p value < 2.2e-16 which is also less than 0.05. So, we reject the null hypothesis. We accept alternative hypothesis. So, FCVC is not normally distributed.

Shapiro-Wilk normality test for NCP

Test Statistic, W = 0.74095

p-value < 2.2e-16

95% CI for W-statistic: 0.7123 to 0.7696

Comment: Here we see that for age Shapiro wilk test p value < 2.2e-16 which is also less than 0.05. So, we reject the null hypothesis. We accept alternative hypothesis. So, NCP is not normally distributed.

Shapiro-Wilk normality test for CH2O

Test Statistic, W= 0.93362

p-value < 2.2e-16

95% CI for W-statistic: 0.9183 to 0.9489

Comment: Here we see that for age Shapiro wilk test p value < 2.2e-16 which is also less than 0.05. So, we reject the null hypothesis. We accept alternative hypothesis. So, CH2O is not normally distributed.

Shapiro-Wilk normality test for FAF

Test Statistic, W= 0.91518

p-value < 2.2e-16

95% CI for W-statistic: 0.898 to 0.9324

Comment: Here we see that for age Shapiro wilk test p value < 2.2e-16 which is also less than 0.05. So, we reject the null hypothesis. We accept alternative hypothesis. So, FAF is not normally distributed.

Shapiro-Wilk normality test for TUE

Test Statistic, W= 0.88615

p-value < 2.2e-16

95% CI for W-statistic: 0.8664 to 0.9059

Comment: Here we see that for age Shapiro wilk test p value < 2.2e-16 which is also less than 0.05. So, we reject the null hypothesis. We accept alternative hypothesis. So, TUE is not normally distributed.

SO, here we see that all numeric variables are not normally distributed. They all are non-normally distributed. So, we present it using median, interquartile range, coefficient of quartile deviation and box plot.

The code of using median, interquartile range, coefficient of quartile deviation and box plot is

Code:

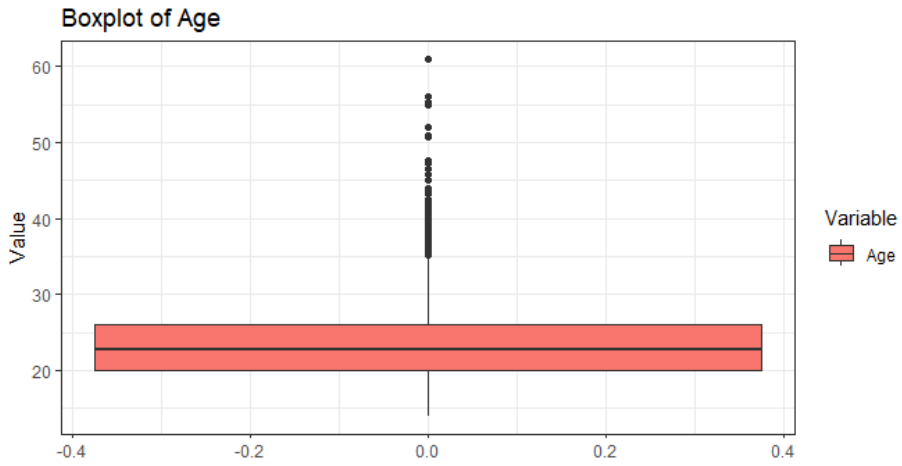
```
# Median, IQR, and CQD
for (var in numerical_variables) {
 z <- data[[var]]
 cat("\n\nSummary for variable:", var, "\n")
 print(cbind(
  median = median(z, na.rm = TRUE),
  iqr = IQR(z, na.rm = TRUE),
  cqd = IQR(z, na.rm = TRUE) / (quantile(z, 0.25, na.rm = TRUE) + quantile(z, 0.75, na.rm = TRUE))
 ))
}
# Boxplots using ggplot2 with legend
for (var in numerical_variables) {
 print(
  ggplot(data, aes(y = .data[[var]], fill = var)) +
   geom_boxplot() +
   labs(title = paste("Boxplot of", var),
      y = "Value",
      fill = "Variable") +
   theme_bw()
 )
}
```

Median, Interquartile range, Coefficient of quartile deviation of all numerical Variables

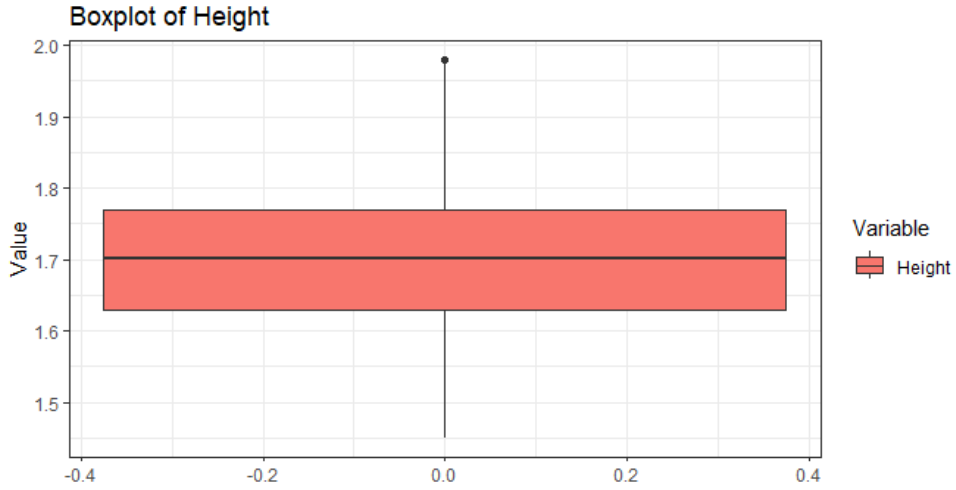
Summary of variables	Median	Interquartile range	Coefficient of quartile deviation
Age	22.77789	6.052808	0.131734
Height	1.700499	0.138464	0.04074311
Weight	83	41.95734	0.2426626
FCVC	2.385502	1	0.2
NCP	3	0.341262	0.06030709
CH2O	2	0.8926075	0.2197332
FAF	1	1.542172	0.8609801
TUE	0.62535	1	1

Output of Boxplot:

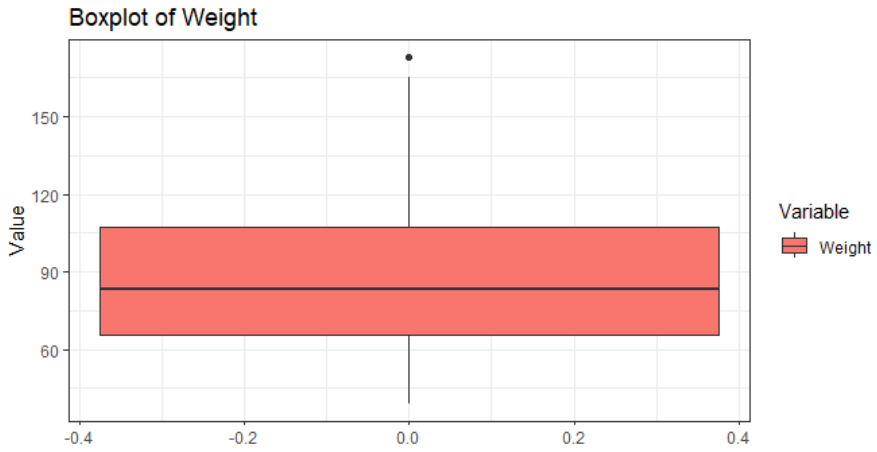
Boxplot of Age:



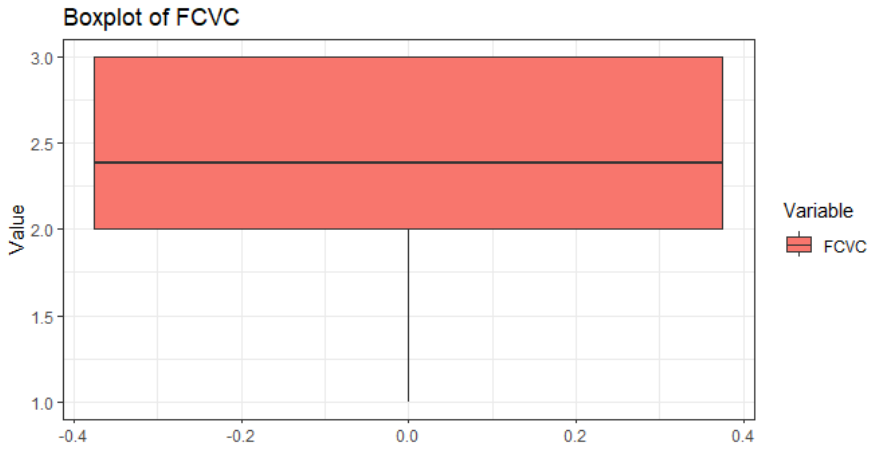
Boxplot of Height



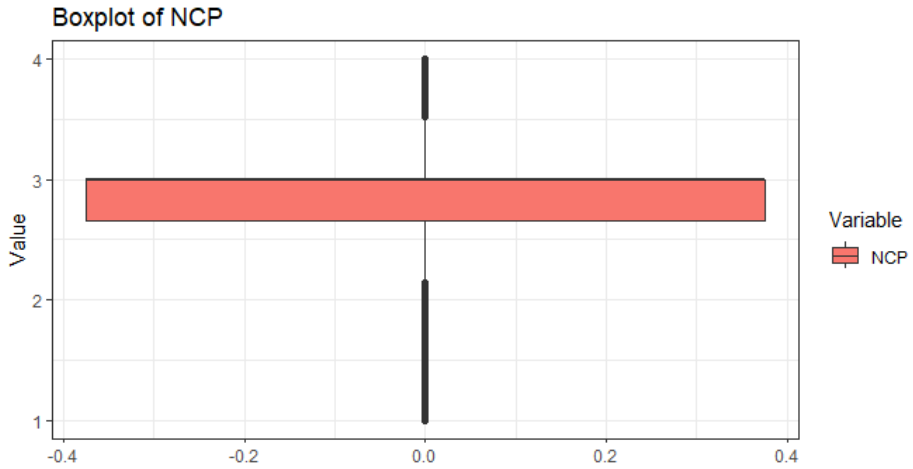
Boxplot of Weight



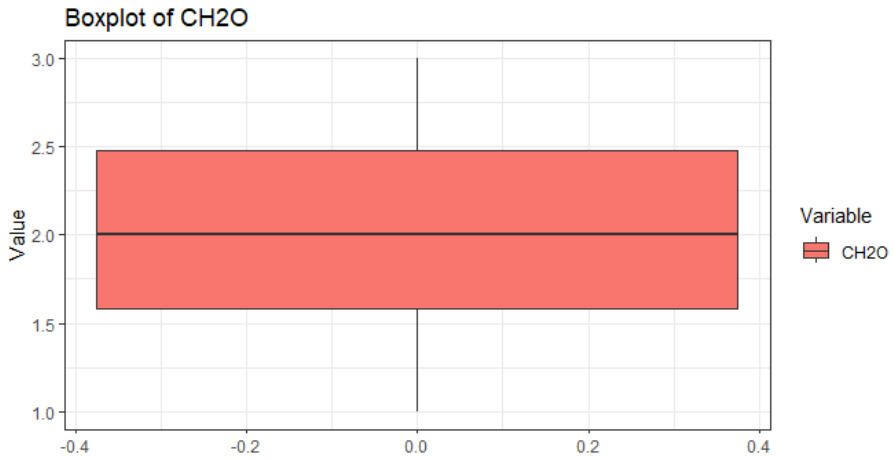
Boxplot of FCVC



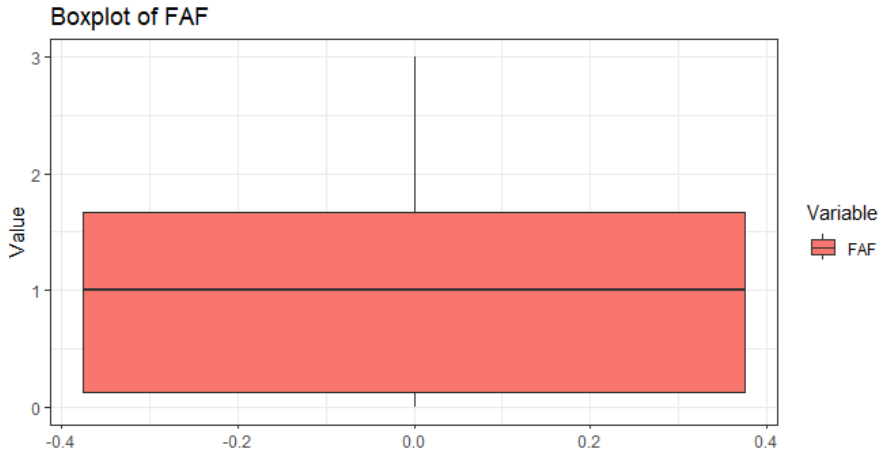
Boxplot of NCP



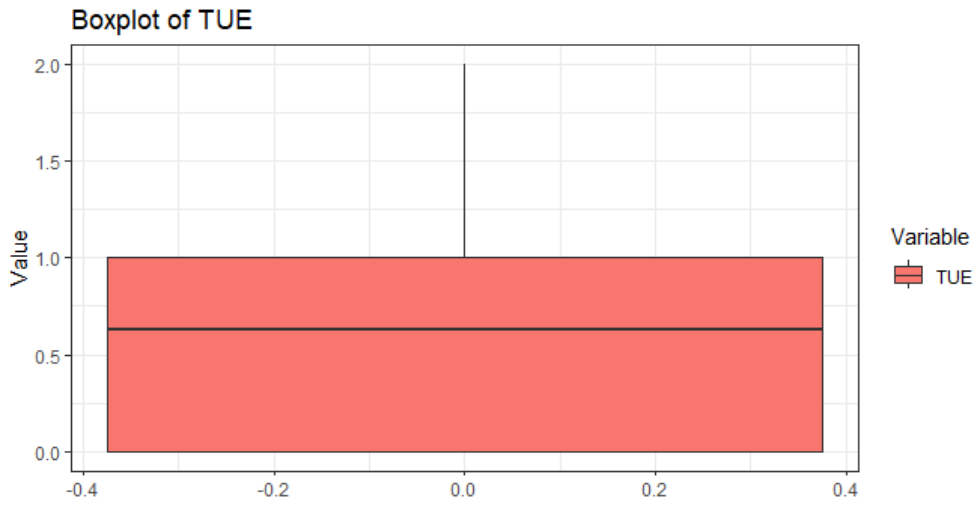
Boxplot of CH2O



Boxplot of FAF



Boxplot of TUE



Ans to the Question Number 07

The code of Performing exploratory subgroup analysis: calculate summary statistics and draw graphs for each numerical variable by every categorical variable. For example, calculate minimum, maximum, mean, median, mode, 1st quartile, 3rd quartile, standard deviation, variance, coefficient of variation, interquartile range, coefficient of quartile deviation of age,height, weight, FCVC, CH2O, FAF, TUE separately for every level of the categorical variables such as separately for male and female, smoker and non-smoker, ... so on.

Code:

```
install.packages("tidyverse")

library(tidyverse)

fun <- function(x,y)

{

 data %>%

  group_by(!!sym(x)) %>%

  summarise(min=min(!!sym(y)),

        max=max(!!sym(y)),

        mean=mean(!!sym(y)),

        median=median(!!sym(y)),

        mode=as.numeric(names(sort(table(!!sym(y)),decreasing=T)[1])),

        q1=as.numeric(quantile(!!sym(y),0.25)),

        q3=as.numeric(quantile(!!sym(y),0.75)),

        sd=sd(!!sym(y)),

        var=var(!!sym(y)),

        cv=sd(!!sym(y))/mean(!!sym(y))*100,

        iqr=IQR(!!sym(y)),

        cqd=IQR(!!sym(y))/(as.numeric(quantile(!!sym(y),0.25))+as.numeric(quantile(!!sym(y),0.75)))

  )

}

for(x in categorical_variables)

 for(y in numerical_variables)

 {

  print(paste(x,"X",y))

  print(fun(x,y))

 }
```

Output:

Gender X Age,Height ,Weight ,FCVC,NCP,CH2O,FAF,TUE

Variable	Gender	Min	Max	Mean	Median	Mode	Q1	Q3	SD	Var	CV	IQR
Age	Female	15	61	24	22	26	19.6	26	6.41	41.1	26.7	6.37
	Male	14	56	24.6	23	18	20	27.9	6.27	39.4	25.5	7.93
Height	Female	1.45	1.84	1.64	1.64	1.6	1.6	1.7	0.0745	0.00555	4.53	0.103
	Male	1.56	1.98	1.76	1.76	1.7	1.71	1.81	0.0722	0.00521	4.1	0.102
Weight	Female	39	165	82.3	78	50	58	105	29.7	884	36.1	47
	Male	45	173	90.8	89.9	70	75	108	21.4	458	23.6	33.5
FCVC	Female	1	3	2.57	2.96	3	2	3	0.548	0.3	21.3	1
	Male	1	3	2.27	2.03	2	2	2.74	0.477	0.228	21	0.737
NCP	Female	1	4	2.63	3	3	2.66	3	0.816	0.666	31	0.337
	Male	1	4	2.74	3	3	2.66	3	0.735	0.541	26.9	0.341
CH2O	Female	1	3	1.94	2	2	1.38	2.49	0.634	0.402	32.7	1.1
	Male	1	3	2.07	2	2	1.88	2.46	0.585	0.342	28.2	0.584
FAF	Female	0	3	0.847	0.742	0	0	1.51	0.84	0.706	99.2	1.51
	Male	0	3	1.17	1	1	0.583	1.93	0.83	0.69	71	1.34
TUE	Female	0	2	0.647	0.656	0	0	1	0.572	0.327	88.4	1
	Male	0	2	0.668	0.602	0	0	1	0.643	0.413	96.2	1

Family History of Overweight X Age, Height, Weight, FCVC,NCP,CH2O,FAF,TUE

Variable	Family History of Overweight	Min	Max	Mean	Median	Mode	Q1	Q3	SD
Age	No	16	61	21.5	20	21	18.8	22.1	5.59
	Yes	14	56	24.9	23	26	20.8	26.8	6.34
Height	No	1.45	1.93	1.65	1.64	1.62	1.58	1.72	0.0948
	Yes	1.48	1.98	1.71	1.71	1.65	1.64	1.78	0.0894
Weight	No	39.1	115	59	56	50	49	69.5	14.2
	Yes	39	173	92.7	90	80	75.9	112	24.2
FCVC	No	1	3	2.37	2.21	2	2	3	0.586
	Yes	1	3	2.43	2.4	3	2	3	0.521
NCP	No	1	4	2.57	3	3	1.63	3	0.971
	Yes	1	4	2.71	3	3	2.77	3	0.726
CH2O	No	1	3	1.82	2	2	1.06	2	0.664
	Yes	1	3	2.05	2	2	1.69	2.53	0.593
FAF	No	0	3	1.11	1	0	0.11	2	0.928
	Yes	0	3	0.988	1	0	0.129	1.6	0.831
TUE	No	0	2	0.628	0.715	0	0	1	0.636
	Yes	0	2	0.664	0.624	0	0.0226	1	0.603

Variable	Frequent Consumption of High Caloric Food (FAVC)	Min	Max	Mean	Median	Mode	Q1	Q3	SD
Age	No	16	56	23.2	21	23	19.3	23.5	6.45
	Yes	14	61	24.5	23	18	20	26	6.32
Height	No	1.48	1.93	1.66	1.65	1.75	1.59	1.73	0.0968
	Yes	1.45	1.98	1.71	1.71	1.7	1.64	1.77	0.0912
Weight	No	42	130	66.9	66	60	53	77	17.1
	Yes	39	173	89.2	86	80	69.5	110	26.1
FCVC	No	1	3	2.46	2.49	3	2	3	0.521
	Yes	1	3	2.41	2.37	3	2	3	0.535
NCP	No	1	4	2.7	3	3	2.73	3	0.831
	Yes	1	4	2.68	3	3	2.66	3	0.771
CH2O	No	1	3	1.99	2	2	1.62	2.31	0.655
	Yes	1	3	2.01	2	2	1.57	2.49	0.607
FAF	No	0	3	1.26	1	0	0.417	2	0.948
	Yes	0	3	0.977	1	0	0.112	1.61	0.832
TUE	No	0	2	0.543	0.122	0	0	1	0.637
	Yes	0	2	0.673	0.646	0	0.0201	1	0.604

Consumption of Food Between Meals (CAEC)	Min	Max	Mean	Median	Mode	Q1	Q3	SD
Always	16	61	23.1	21	21	19	24	7.31
Frequently	16	55	22.2	21	18	19	23	5.31
Sometimes	14	56	24.7	23	18	20	26.7	6.44
No	19	35	21.8	21	21	21	21.8	2.76
Always	1.5	1.98	1.7	1.7	1.6	1.63	1.8	0.107
Frequently	1.45	1.91	1.67	1.68	1.7	1.59	1.74	0.0976
Sometimes	1.46	1.98	1.71	1.71	1.65	1.64	1.77	0.0918
No	1.59	1.81	1.64	1.62	1.62	1.62	1.64	0.0558
Always	45	125	71.1	66	60	60	80	18.4

Frequently	40	173	58.9	52	42	49.2	69	16.6
Sometimes	39	165	91.4	88.1	80	75	111	25
No	45	112	68.9	70	70	68.1	70	10.8
Always	1	3	2.36	2	3	2	3	0.682
Frequently	1	3	2.49	2.86	3	2	3	0.594
Sometimes	1	3	2.42	2.37	3	2	3	0.518
No	1	3	2.07	2	2	2	2	0.481
Always	1	4	2.81	3	3	3	3	0.962
Frequently	1	4	2.79	3	3	3	3	0.923
Sometimes	1	4	2.69	3	3	2.68	3	0.728
No	1	3.95	1.96	1	1	1	3	1.1
Always	1	3	2.02	2	2	2	2	0.604
Frequently	1	3	1.77	2	2	1	2	0.654
Sometimes	1	3	2.02	2	2	1.62	2.49	0.596
No	2	3	2.63	2.94	3	2.06	3	0.449
Always	0	3	1.13	1	0	0	2	1.14
Frequently	0	3	1.09	1	0	0.0583	2	0.903
Sometimes	0	3	0.992	1	0	0.129	1.62	0.837
No	0	3	1.14	1	1	0.933	1.06	0.67
Always	0	2	0.736	1	0	0	1	0.788
Frequently	0	2	0.654	0.641	0	0	1	0.667
Sometimes	0	2	0.668	0.64	0	0.0266	1	0.595
No	0	1.19	0.229	0	0	0	0.381	0.386

And the result Is soo on

Ans to the Question Number 08

The code of scatter diagram between (i) age and weight, (ii) age and height and (iii) height and weight.And respective regression lines on the scatterplots is

Code:

```
# Age vs Weight
ggplot(data, aes(x = Age, y = Weight)) +
 geom_point() +
 geom_smooth(method = "lm", color = "red") +
 ggtitle("Age vs Weight") +
 xlab("Age") +  # Added x-axis label
 ylab("Weight") # Added y-axis label


# Height vs Weight
ggplot(data, aes(x = Weight, y = Height)) +
 geom_point() +
 geom_smooth(method = "lm", color = "red") +
 ggtitle("Weight vs Height") +
 xlab("Weight") +
 ylab("Height")

# Age vs Height
ggplot(data, aes(x = Age, y = Height)) +
 geom_point() +
 geom_smooth(method = "lm", color = "red") +
 ggtitle("Age vs Height") +
 xlab("Age") +
 ylab("Height")
```

And code of Scatter matrix all the numerical variables.

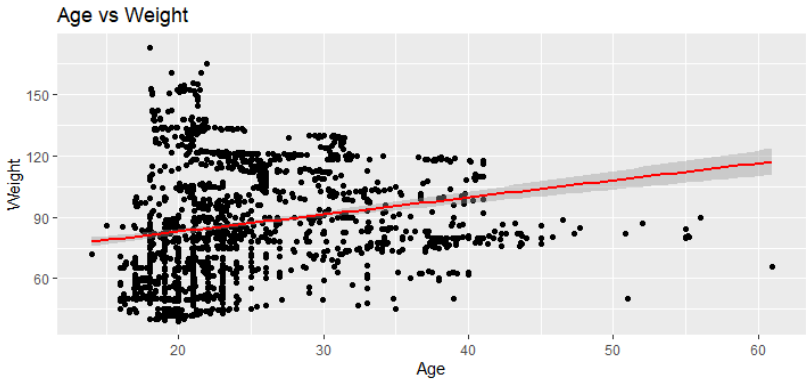
Code:

```
pairs(data[, c('Age','Height','Weight','FCVC','NCP','CH2O','FAF','TUE')], pch = 16)
```

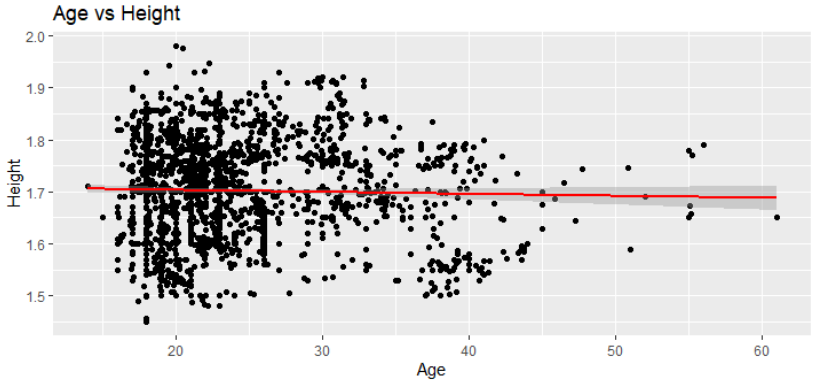
Output:

Scatter plot of between (i) age and weight, (ii) age and height and (iii) height and weight.And respective regression lines on the scatterplots

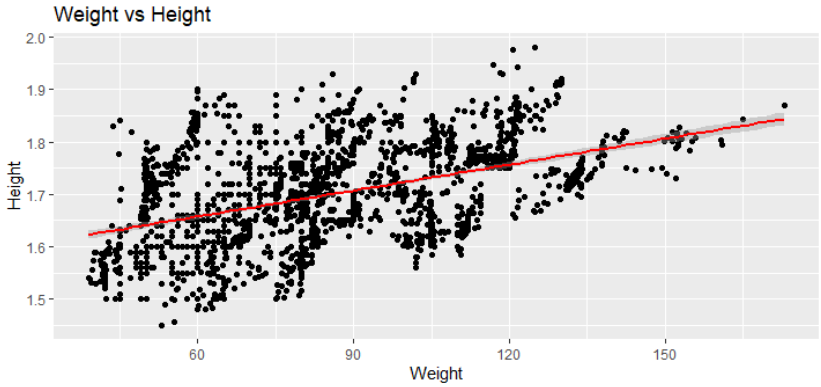
(i) age and weight



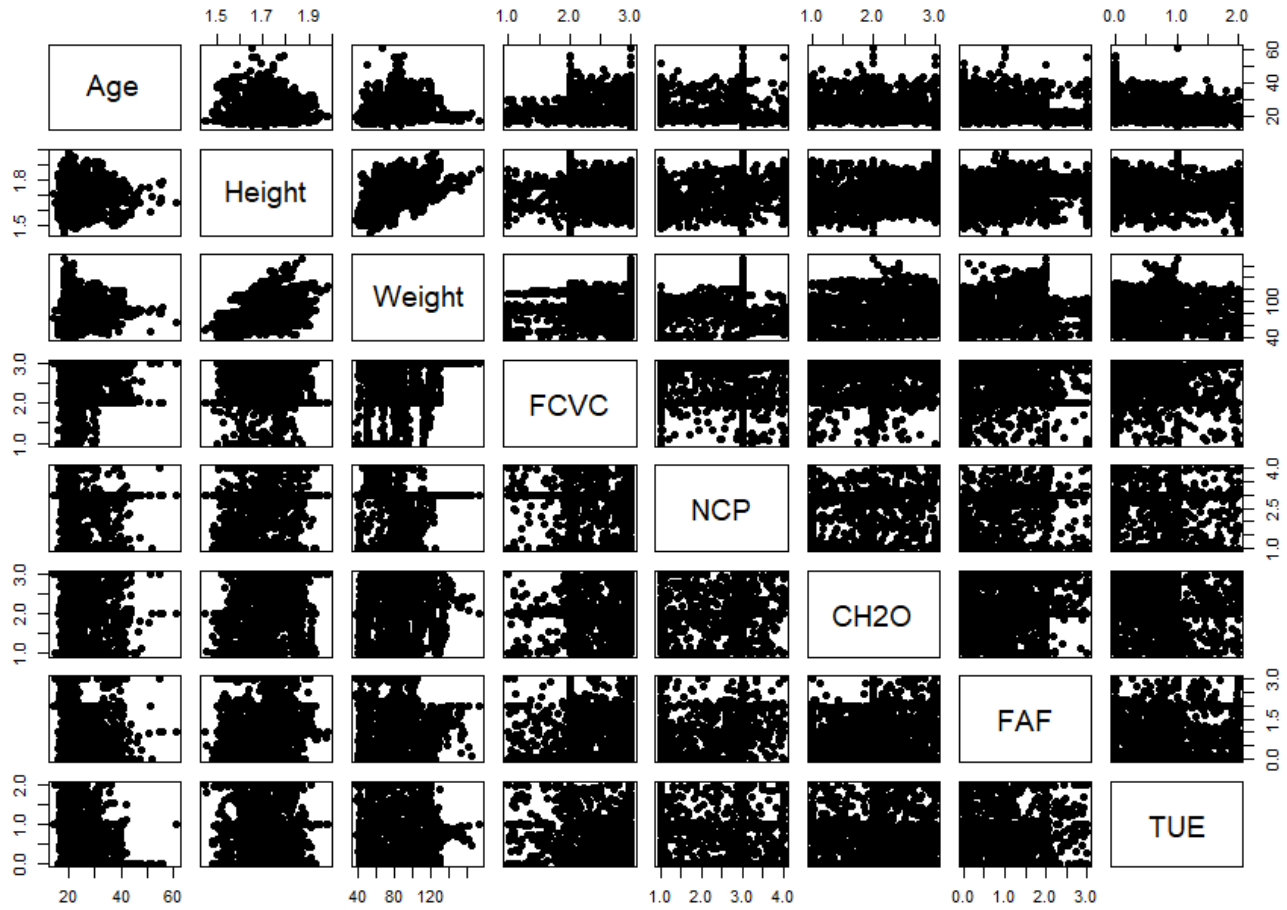
ii)age and height



(iii) height and weight



Scatter matrix



Ans to the Question number 09

Code of Recode the variable MTRANS into MTRANS_RC in which both the 'Walking' and 'Bike' categories will be 'Ownself' and all other categories will be 'Car'.and Also, convert the variable FCVC to a factor (name as FCVC_factor) by labeling 1 as 'Never', 2 as 'Sometimes' and 3 as 'Always'.

Code:

```
data$MTRANS_RC <- ifelse(data$MTRANS %in% c("Walking", "Bike"), "Ownself", "Car")

data$FCVC_factor <- factor(data$FCVC, levels = c(1, 2, 3), labels = c("Never", "Sometimes", "Always"))
```

My Recoded data At My RStudio

Ans to the Question number 10

The code of Calculate body mass index (BMI) by using the formula, BMI = Weight/(Height)2 is

Code:

```
data$BMI<-data$Weight/(data$Height^2)
```

The code of Make a categorical variable (name as BMI_cat) using the following categorization of BMI values: Less than 18.5 as 'Underweight'; 18.5 to 24.9 as 'Normal'; 25.0 to 29.9 as 'Overweight';Greater than 30 as 'Obesity'. Is

Code:

```
data$BMI_cat<-cut(data$BMI,breaks = c(0,18.5,24.9,29.9,Inf),labels =c("Underweight","Normal","Overweight","Obisity"))
```

My caltulated variable with category included in my Rstudio data

Ans to the Question number 11

The code of Calculate BMI for respondents (i) whose age > 30 years, (ii) who are non-smokers, have physical activity of 2 days and drink more than 1 liter of water daily.

Code:

```
subset_data <- data %>% filter(Age > 30, SMOKE == "no", FAF >= 2, CH2O >= 1)

View(subset_data)

subset_data$BMI

mean(subset_data$BMI)
```

we have got 43 samples of the ObesityData (i) whose age > 30 years, (ii) who are non-smokers, have physical activity of 2 days and drink more than 1 liter of water daily.

And there BMI are

30.55556 ,26.51180, 24.21229, 32.27079, 29.41073, 27.13141, 22.23379, 33.13976 ,24.48980, 24.48980 ,32.87311 ,28.05571,

25.23634 ,24.03461,24.55775 ,23.04002 ,21.46823, 22.89282, 25.99244 ,27.96820 ,24.92449,26.06826 ,25.98459, 28.76961, 28.53671, 28.20605, 28.86065, 28.42717, 28.40860 ,27.60695, 32.30919 ,31.67679 ,31.82920, 31.49210, 31.26181, 32.27684 ,32.60889 ,32.61778, 30.35989 ,31.51047 ,31.15444 ,31.53726, 31.10044

And their Mean is 28.32775

There create a new data set name subset_data

Full subset_data in my RStudio

Ans to the Question number 12

The code of create a new dataset (name as: obesity_sub) by taking the respondents whose height is more than 1.8 meter and who eat high caloric food frequently .And the code of Calculating mean and standard deviation of BMI using the obesity_sub dataset.

Code:

```
obesity_sub <- data %>% filter(Height > 1.8, FAVC == "yes")

View(obesity_sub)

mean(obesity_sub$BMI)

sd(obesity_sub$BMI)
```

The obesity_sub data in my RStudio

Mean of obesity_sub data BMI is 30.44498 and

Standard Deviation of obesity_sub data BMI is 7.630551

Ans to the Question number 13

The code of Calculating correlation between (i) age and weight, (ii) age and height and (iii) height and weight.

Code:

```
cor(data$Age, data$Weight, use = "complete.obs")

cor(data$Height, data$Weight, use = "complete.obs")

cor(data$Age, data$Height, use = "complete.obs")
```

Output:

The correlation coefficient i)age and weight is, r1= 0.2025601

There is a weak positive correlation between age and weight becouse this coefficient in interval 0.1 to 0.3

The correlation coefficient ii)age and height is, r2= -0.02595813 and

There is negative correlation between age and height

The correlation coefficient iii)height and weight is, r3= 0.4631361

There is Moderate Positive correlation between height and weight becouse this coefficient in interval 0.4 to 0.6

Ans to the Question number 14

The code of Calculating correlation between age and BMI. And test correlation significantly differ from zero.Calculating correlation matrix of all the numerical variables.

Code:

```
#Correlation between Age and BMI

correlation_value <- cor(data$Age, data$BMI, use = "complete.obs")

correlation_test<- cor.test(data$Age, data$BMI,conf.level = 0.95)

correlation_test$P_value

print(paste("Correlation:", correlation_value))

# Check significance

if (p_value <= 0.05) {

 print("Correlation is significant (not equal to 0)")

} else {

 print("Correlation is not significant (close to 0)")

}

# Correlation matrix for numerical variables

cor_matrix <- cor(data[numerical_vars], use = "complete.obs")

print(cor_matrix)
```

Output:

The correlation coefficient between the age and BMI is r = 0.244163116121791 which is indicate Weak Positive relationship because the correlation interval between 0.1 to 0.3 .

Hypothesis

Ho: Correlation is not significant

H1: Correlation is significant

P_valu = 2.2e-16

Test statistic = 11.563

95% confidence interval = 0.2036214 0.2838689

Correlation is significant because the p_value of correlation test is 5.00646724638642e-30 < 2.2e-16

so we can reject the null hypothesis.So it is significantly differs from zero

Calculate correlation matrix of all the numerical variables.

	Age	Height	Weight	FCVC	NCP	CH2O	FAF	TUE
Age	1.000000	-0.025958	0.202560	0.016291	-0.043944	-0.045304	-0.144938	-0.296931
Height	-0.025958	1.000000	0.463136	-0.038121	0.243672	0.213376	0.294709	0.051912
Weight	0.202560	0.463136	1.000000	0.216125	0.107469	0.200575	-0.051436	-0.071561
FCVC	0.016291	-0.038121	0.216125	1.000000	0.042216	0.068461	0.019939	-0.101135
NCP	-0.043944	0.243672	0.107469	0.042216	1.000000	0.057088	0.129504	0.036326
CH2O	-0.045304	0.213376	0.200575	0.068461	0.057088	1.000000	0.167236	0.011965
FAF	-0.144938	0.294709	-0.051436	0.019939	0.129504	0.167236	1.000000	0.058562
TUE	-0.296931	0.051912	-0.071561	-0.101135	0.036326	0.011965	0.058562	1.000000

Ans to the Question number 15

Here Data is not normally distributed ,so the appropriate method is non-parametric tests.Also here the groups are not independent(paired) .So in this question we apply the Wilcoxon signed rank test . The code of (i) average age is equal to 30 years, (ii) average height is greater than to 1.7 meters, (iii) average consumption of water daily (CH2O) is equal to 2 liters and (iv) average BMI is less than 30 is.

Code:

```
#question 15i

b<-wilcox.test(data$Age,alternative = "two.sided",mu = 30,conf.int = TRUE, conf.level = 0.95)

b

b$p.value


print(paste('P value =',b$p.value ))

ifelse(b$p.value<=0.05,"Mean of Age is not equal to 30","Mean of age is equal to 30")

#question 15ii

b<-wilcox.test(data$Height,alternative = 'greater',mu = 1,conf.int = TRUE, conf.level = 0.95)

b

b$p.value

print(paste('P value =',b$p.value))

ifelse(b$p.value<=0.05,"Mean of Height is greater than 1.7","Mean of Height is equal to 1.7")
```

```
#question 15iii

b<-wilcox.test(data$CH2O,alternative = "less",mu=2,conf.int = TRUE, conf.level = 0.95)

b

b$p.value

print(paste('P valur=',b$p.value))

ifelse(b$p.value<=0.05,"Mean of CH20 is not equal to 2","Mean of CH20 is equal to 2")

#question 15iv

b<-wilcox.test(data$BMI,alternative ="less",mu =30,conf.int = TRUE, conf.level = 0.95)

b

b$p.value

print(paste('P value=',b$p.value))

ifelse(b$p.value<=0.05,"Mean of BMI is less than 30","Mean of BMI is equal to 30")
```

Output:

Question 15(i)

Hypothesis

H0: Mean of Height is equal to 1.7

HA: Mean of Height is greater than 1.7

Test statistic,V = 258768

P value = 2.19675225529564e-203 < 2.2e-16

95% confidence interval 23.00001 and 23.50002

Comment: Here P value = 2.19675225529564e-203 which is <=0.05 .So we can reject the null hypothesis.Alternative Hypothesis is accepted here .So Mean of Age is not equal to 30

Question 15(ii)

Hypothesis

H0: Mean of Height is equal to 1.7

HA: Mean of Height is greater than 1.7

Test Statistic,V = 1075418

P value = 0.192988500307103

95 % confidence interval: 1.698264 and Inf

Comment: Here P value =0.192988500307103 which is not <=0.05 .So we cannot reject the null hypothesis. So Mean of Height is equal to 1.7

Question 15(iii)

Hypothesis

H0: Mean of CH20 is equal to 2

HA: Mean of CH20 is not equal to 2

Test Statistic,V = 700548

P value = 0.672515363171495

95% confidence interval= -Inf and 2.020164

Comment : Here P value =0.672515363171495 which is not <=0.05 .So we cannot reject the null hypothesis.So , Mean of CH20 is equal to 2

Question 15(iv)

Hypothesis

H0: Mean of BMI is equal to 30

HA: Mean of BMI is less than 30

Test Statistic ,V = 1053288

p-value = 0.01429

95% confidence interval: -Inf and 29.89906

Comment: Here P value =0.01429 which is <=0.05 .So we can reject the null hypothesis.So Mean of BMI is less than 30

Ans to the Question number 16

Code:

```
#Question 16i

b<-wilcox.test(data$Age~data$Gender,conf.int = TRUE, conf.level = 0.95)

b$p.value

print(paste('P valu =',b$p.value))

ifelse(b$p.value<=0.05 ,"Average age significantly differs between male and female","Average age doesn't significantly differ between male and female ")

b<-wilcox.test(data$Height~data$Gender,conf.int = TRUE, conf.level = 0.95)

b

b$p.value

print(paste('P value=',b$p.value))

ifelse(b$p.value<=0.05,"Average height significantly differs between male and female","Average height doesn't significantly differ between male and female")
```

```
b<-wilcox.test(data$BMI~data$Gender,conf.int = TRUE, conf.level = 0.95)

b

b$p.value

print(paste('P value =',b$p.value))

ifelse(b$p.value<=0.05,"Average BMI significantly differs between male and female","Average BMI doesn't significantly differ between male and female")
```

Output:

a)Between Age and Gender we see that

H0: Average age doesn't significantly differ between male and female

HA: Average age significantly differs between male and female

Test statistics , W = 513089,

p-value = 0.001723

95% Confidance Interval : -1.0063792 and -0.1429384

Comment: Here P Value = 0.001723 which is less than 0.05 .So it can reject the null hypothesis .So Average age significantly differs between male and female.

b)Between Height and Gender

H0: Average height doesn't significantly differ between male and female

HA: Average height significantly differs between male and female

Test Statistic , W = 153180

P Value = 7.009999e-183

95% Confidence Interval = -0.1236116 and -0.1103230

Comment: Here P Value =7.009999e-183 which is less than 0.05 .So we can reject the null hypothesis. So Average height significantly differs between male and female

C) Between BMI And Gender

Hypothesis

H0: Average BMI doesn't significantly differ between male and female

HA: Average BMI significantly differs between male and female

Test Statistics, W = 568467

p-value = 0.4113

95% Confidence Interval : -0.4114121 and 1.1658265

Comment: Here P value =0.4113 which is greater than 0.05 .So we can reject the null hypothesis . Average BMI doesn't significantly differ between male and female.

Code:

```
#Question 16ii

b<-wilcox.test(data$Age~data$SMOKE,conf.int = TRUE, conf.level = 0.95)

b

b$p.value

print(paste('P value =',b$p.value))

ifelse(b<=0.05,"Average age significantly differs between smoker and non smoker","Average age doesn't significantly differs between smoker and non smoker")

b<-wilcox.test(data$Height~data$SMOKE,conf.int = TRUE, conf.level = 0.95)

b

b$p.value

print(paste('P value=',b$p.value))

ifelse(b$p.value<=0.05,"Average height significantly differ between smoker and non smoker","Average height doestn't significantly differ between smoker and non smoker")

b<-wilcox.test(data$BMI~data$SMOKE,conf.int = TRUE, conf.level = 0.95)

b

b$p.value

print(paste('P value = ',b))

ifelse(b$p.value<=0.05,"Average BMI significantly differ between smoker and non smoker","Average BMI doesn't significantly differ between smoker and non smoker")
```

Output:

a)Between Age and Smoke

Hypothesis

H0: Average age doesn't significantly differs between smoker and non smoker

HA: Average age significantly differs between smoker and non smoker

Test statistic , W = 31130

p-value = 0.0003357

95% confidence interval: -5.128242 -1.343135

Comment: Here p-value = 0.0003357 which is less than 0.05 . So we can reject the null hypothesis. Average age significantly differs between smoker and non smoker.

b)Between Height and Smoke

Hypothesis:

H0: Average height doestn't significantly differ between smoker and non smoker

HA: Average height significantly differ between smoker and non smoker

Test Statistic , W = 37178

p-value = 0.03813

95 percent confidence interval: -0.078352292 -0.001022176

Comment: Here p-value = 0.03813 which is less than 0.05. So we can reject the null hypothesis.So Average height significantly differ between smoker and non smoker.

c)Between BMI and Smoke

Hypothesis

H0: Average BMI doesn't significantly differ between smoker and non smoker

HA: Average BMI significantly differ between smoker and non smoker

Test Statistic, W = 45057

p-value = 0.917

95% confidence interval: -2.573370 and 2.304262

Comment: Here p-value = 0.917 which is greater than 0.05 .So we can not reject the null hypothesis. So Average BMI doesn't significantly differ between smoker and non smoker

Ans to the Question number 17

Code:

```
b<-kruskal.test(Age~CAEC,data = data)

b

b$p.value

print(paste('P value =',b$p.value))

ifelse(b$p.value<=0.05,"Median of age significantly differs between different groups of CAEC","Median of age doesn't significantly differs between different groups of CAEC")

b<-kruskal.test(Height~CAEC,data = data)

b

b$p.value

print(paste('P value =',b$p.value))
```

```
ifelse(b$p.value<=0.05,"Median of height significantly differs between different groups of CAEC","Median of height doesn't significantly differs between different groups of CAEC")

b<-kruskal.test(BMI~CAEC,data = data)

b

b$p.value

print(paste('P value =',b$p.value))

ifelse(b$p.value<=0.05,"Median of bmi significantly differs between different groups of CAEC","Median of bmi doesn't significantly differs between different groups of CAEC")

b<-kruskal.test(Age~CALC,data = data)

b

b$p.value

print(paste('P value =',b$p.value))

ifelse(b$p.value<=0.05,"Median of age significantly differs between different groups of CALC","Median of age doesn't significantly differs between different groups of CALC")

b<-kruskal.test(Height~CALC,data = data)

b

b$p.value

print(paste('P value =',b$p.value))

ifelse(b$p.value<=0.05,"Median of height significantly differs between different groups of CALC","Median of height doesn't significantly differs between different groups of CALC")

b<-kruskal.test(BMI~CALC,data = data)

b

b$p.value

print(paste('P value =',b$p.value))

ifelse(b$p.value<=0.05,"Median of bmi significantly differs between different groups of CALC","Median of age doesn't significantly differs between different groups of CALC")

b<-kruskal.test(Age~MTRANS,data = data)

b

b$p.value

print(paste('P value =',b$p.value))

ifelse(b$p.value<=0.05,"Median of age significantly differs between different groups of MTRANS","Median of age doesn't significantly differs between different groups of MTRANS")

b<-kruskal.test(Height~MTRANS,data = data)

b

b$p.value

print(paste('P value =',b$p.value))

ifelse(b$p.value<=0.05,"Median of height significantly differs between different groups of MTRANS","Median of age doesn't significantly differs between different groups of MTRANS")

b<-kruskal.test(BMI~MTRANS,data = data)
```

```
b

b$p.value

print(paste('P value =',b$p.value))

ifelse(b$p.value<=0.05,"Median of BMI significantly differs between different groups of MTRANS","Median of BMI doesn't significantly differs between different groups of MTRANS")
```

code of multiple comparison

```
#Multiple comparison

install.packages("DescTools")

library(DescTools)

DunnTest(Age~CAEC,data = data,method = "bonferroni")

DunnTest(Age~CALC,data = data,method = "bonferroni",)

DunnTest(Age~MTRANS,data = data,method = "bonferroni")

DunnTest(Height~CAEC,data = data,method = "bonferroni")

DunnTest(Height~CALC,data = data,method = "bonferroni")

DunnTest(Height~MTRANS,data = data,method = "bonferroni")

DunnTest(BMI~CAEC,data = data,method = "bonferroni")

DunnTest(BMI~CALC,data = data,method = "bonferroni")

DunnTest(BMI~MTRANS,data = data,method = "bonferroni")
```

Output:

a) Age by CAEC

Hypothesis

H0: Median of age doesn't significantly differs between different groups of CAEC

HA: Median of age significantly differs between different groups of CAEC

Test Statistic, Kruskal-Wallis chi-squared = 59.711

p-value = 6.776e-13

In Kruskal-Wallis chi-squared test we cannot include the 95% confidence interval

Comment: Here p-value = 6.776e-13 which is less than 0.05 . So we cam reject the null hypothesis.we use here alternative hypothesis. So Median of age significantly differs between different groups of CAEC

b) Height by CAEC

Hypothesis

H0: Median of height doesn't significantly differs between different groups of CAEC

HA: Median of height significantly differs between different groups of CAEC

Test Statistic , Kruskal-Wallis chi-squared = 409.66

p-value =1.790187e-88

In Kruskal-Wallis chi-squared test we cannot include 95% confidence interval

Comment: Here we see that p-value =1.790187e-88 which is less than 0.05 ,so we can reject the null hypothesis .We accept alternatives hypothesis.So Median of height significantly differs between different groups of CAEC.

c) BMI by CAEC

Hypothesis

H0: Median of bmi doesn't significantly differs between different groups of CAEC

HA: Median of bmi significantly differs between different groups of CAEC

Test Statistic , Kruskal-Wallis chi-squared = 409.66

P value = 1.79018730009495e-88

In Kruskal-Wallis chi-squared test we cannot include 95% confidence interval

Comment: Here we see that P value = 1.79018730009495e-88 which is less than 0.05 .So we can reject the null hypothesis.We accept the alternative hypothesis.So Median of bmi significantly differs between different groups of CAEC

d) Age by CALC

Hypothesis:

H0: Median of age doesn't significantly differs between different groups of CALC

HA: Median of age significantly differs between different groups of CALC

Test Statistic , Kruskal-Wallis chi-squared = 20.683,

p-value = 0.0001225

In Kruskal-Wallis chi-squared test we cannot include 95% confidence interval

Comment: Here we see that p-value = 0.0001225 which is less than 0.05 .So we can reject the null hypothesis.We accept the alternative hypothesis.So Median of age significantly differs between different groups of CALC

e) Height by CALC

Hypothesis:

H0: Median of age doesn't significantly differs between different groups of CALC

HA: Median of age significantly differs between different groups of CALC

Test Statistic: Kruskal-Wallis chi-squared = 39.114

p-value = 1.642e-08

In Kruskal-Wallis chi-squared test we cannot include 95% confidence interval

Comment : Here we see p-value = 1.642e-08 which is less than 0.05 .So we can reject the null hypothesis.We accept the alternative hypothesis.So Median of height significantly differs between different groups of CALC

f) BMI by CALC

H0: Median of bmi doesn't significantly differs between different groups of CALC

HA: Median of bmi significantly differs between different groups of CALC

Test Statistic, Kruskal-Wallis chi-squared = 98.299

P value = 3.60798558751883e-21

In Kruskal-Wallis chi-squared test we cannot include 95% confidence interval

Comment : Here ,P value = 3.60798558751883e-21 which is less than 0.05 .So we can reject the null hypothesis.We accept the alternative hypothesis. So Median of bmi significantly differs between different groups of CALC .

g) Age by MTRANS

Hypothesis

H0: Median of age doesn't significantly differs between different groups of MTRANS

HA: Median of age significantly differs between different groups of MTRANS

Test Statistic, Kruskal-Wallis chi-squared = 491.19

P value = 5.39227819213896e-105

In Kruskal-Wallis chi-squared test we cannot include 95% confidence interval

Comment: Here , P value = 5.39227819213896e-105 , which is less than 0.05 .So we can reject the null hypothesis.We accept the alternative hypothesis. So Median of age significantly differs between different groups of MTRANS

h) Height by MTRANS

Hypothesis

H0: Median of age doesn't significantly differs between different groups of MTRANS

HA: Median of height significantly differs between different groups of MTRANS

Test Statistic, Kruskal-Wallis chi-squared = 24.094,

p-value = 7.65e-05

In Kruskal-Wallis chi-squared test we cannot include 95% confidence interval

Comment: Here , P value = 7.65e-05, which is less than 0.05 .So we can reject the null hypothesis.We accept the alternative hypothesis.So ,Median of height significantly differs between different groups of MTRANS.

i)BMI by MTRANS

Hypothesis:

H0: Median of BMI doesn't significantly differs between different groups of MTRANS

HA: Median of BMI significantly differs between different groups of MTRANS

Test Statistics , Kruskal-Wallis chi-squared = 46.479

p-value = 1.958e-09

In Kruskal-Wallis chi-squared test we cannot include 95% confidence interval

Comment: Here , P value = 1.958e-09 which is less than 0.05 .So we can reject the null hypothesis.We accept the alternative hypothesis.So Median of BMI significantly differs between different groups of MTRANS

Multiple comparison

Dunn's test of multiple comparisons using rank sums : bonferroni

a)Age by CAEC

Dunn's test of multiple comparisons using rank sums : bonferroni

Output: Here adjusted p-value for Frequently-Always is 1.0000 >0.05 which indicates that there is no significant differences between those groups. Adjusted p-value for no-Always is 1.0000>0.05 which indicates that there is no significant differences between those groups. Adjusted p-value for Sometimes-Always is 0.0620>0.05 which indicates that there is no significant differences between those groups. Adjusted p-value for no-Frequently 1.0000>0.05 which indicates that there is no significant differences between those groups .Adjusted p-value for Sometimes-Frequently is 1.7e-11<0.05 which indicates that there is no significant differences between those groups .Adjusted p-value for Sometimes-no is 0.0189 <0.05 05 which indicates that there is no significant differences between those groups

b) Age by CALC

Dunn's test of multiple comparisons using rank sums : bonferroni

Output: Here adjusted p-value for Frequently-Always is 1.0000 > 0.05 which indicates that there is no significant difference between those groups. Adjusted p-value for no-Always is 1.0000 > 0.05 which indicates that there is no significant difference between those groups. Adjusted p-value for Sometimes-Always is 1.0000 > 0.05 which indicates that there is no significant difference between those groups. Adjusted p-value for no-Frequently is 0.00919 < 0.05 which indicates that there is a significant difference between those groups. Adjusted p-value for Sometimes-Frequently is 0.50695 > 0.05 which indicates that there is no significant difference between those groups. Adjusted p-value for Sometimes-no is 0.00051 < 0.05 which indicates that there is a significant difference between those groups.

c) Age by MTRANS

Dunn's test of multiple comparisons using rank sums : bonferroni

Output: Here adjusted p-value for Bike-Automobile is 0.0780 > 0.05 which indicates that there is no significant difference between those groups. Adjusted p-value for Motorbike-Automobile is 0.2506 > 0.05 which indicates that there is no significant difference between those groups. Adjusted p-value for Public_Transportation-Automobile is < 2e-16 < 0.05 which indicates that there is a significant difference between those groups. Adjusted p-value for Walking-Automobile is < 2e-16 < 0.05 which indicates that there is a significant difference between those groups. Adjusted p-value for Motorbike-Bike is 1.0000 > 0.05 which indicates that there is no significant difference between those groups. Adjusted p-value for Public_Transportation-Bike is 1.0000 > 0.05 which indicates that there is no significant difference between those groups. Adjusted p-value for Walking-Bike is 1.0000 > 0.05 which indicates that there is no significant difference between those groups. Adjusted p-value for Public_Transportation-Motorbike is 1.0000 > 0.05 which indicates that there is no significant difference between those groups. Adjusted p-value for Walking-Motorbike is 0.3100 > 0.05 which indicates that there is no significant difference between those groups. Adjusted p-value for Walking-Public_Transportation is 0.8422 > 0.05 which indicates that there is no significant difference between those groups.

d)Height by CAEC

Dunn's test of multiple comparisons using rank sums : bonferroni

Output: Adjusted p-value for Frequently-Always is 0.2218 > 0.05 which indicates that there is no significant difference between those groups. Adjusted p-value for no-Always is 0.0029 < 0.05 which indicates that there is a significant difference between those groups. Adjusted p-value for Sometimes-Always is 1.0000 > 0.05 which indicates that there is no significant difference between those groups.Adjusted p-value for no-Frequently is 0.1023 > 0.05 which indicates that there is no significant difference between those groups. Adjusted p-value for Sometimes-Frequently is 2.9e-08 < 0.05 which indicates that there is a significant difference between those groups. Adjusted p-value for Sometimes-no is 3.7e-07 < 0.05 which indicates that there is a significant difference between those groups.

e) Height by CALC

Dunn's test of multiple comparisons using rank sums : bonferroni

Output: Adjusted p-value for Frequently-Always is 1.00000 > 0.05, indicating no significant difference. Adjusted p-value for no-Always is 1.00000 > 0.05, indicating no significant difference. Adjusted p-value for Sometimes-Always is 1.00000 > 0.05, indicating no significant difference. Adjusted p-value for no-Frequently is 0.00085 < 0.05, indicating a significant difference. Adjusted p-value for Sometimes-Frequently is 0.58397 > 0.05, indicating no significant difference. Adjusted p-value for Sometimes-no is 4.4e-08 < 0.05, indicating a significant difference.

f) Height by MTRANS

Dunn's test of multiple comparisons using rank sums : bonferroni

Output:Adjusted p-value for Bike-Automobile is 1.0000 > 0.05, indicating no significant difference. Adjusted p-value for Motorbike-Automobile is 1.0000 > 0.05, indicating no significant difference.Adjusted p-value for Public_Transportation-Automobile is 7.7e-05 < 0.05, indicating a significant difference.Adjusted p-value for Walking-Automobile is 1.0000 > 0.05, indicating no significant difference.Adjusted p-value for Motorbike-Bike is 1.0000 > 0.05, indicating no significant difference.Adjusted p-value for Public_Transportation-Bike is 1.0000 > 0.05, indicating no significant difference.Adjusted p-value for Walking-Bike is 1.0000 > 0.05, indicating no significant difference.Adjusted p-value for Public_Transportation-Motorbike is 1.0000 > 0.05, indicating no significant difference.Adjusted p-value for Walking-Motorbike is 1.0000 > 0.05, indicating no significant difference.Adjusted p-value for Walking-Public_Transportation is 0.6817 > 0.05, indicating no significant difference.

g) BMI BY CAEC

Dunn's test of multiple comparisons using rank sums : bonferroni

Output: Adjusted p-value for Frequently-Always is 0.0482 < 0.05, indicating a significant difference. Adjusted p-value for no-Always is 1.0000 > 0.05, indicating no significant difference. Adjusted p-value for Sometimes-Always is 4.9e-10 < 0.05, indicating a significant difference. Adjusted p-value for no-Frequently is 0.0021 < 0.05, indicating a significant difference. Adjusted p-value for Sometimes-Frequently is < 2e-16 < 0.05, indicating a significant difference. Adjusted p-value for Sometimes-no is 5.9e-07 < 0.05, indicating a significant difference.

h)BMI by CALC

Dunn's test of multiple comparisons using rank sums : bonferroni

Output: Adjusted p-value for Frequently-Always is 1.0000 > 0.05, indicating no significant difference. Adjusted p-value for no-Always is 1.0000 > 0.05, indicating no significant difference. Adjusted p-value for Sometimes-Always is 1.0000 > 0.05, indicating no significant difference. Adjusted p-value for no-Frequently is 1.0000 > 0.05, indicating no significant difference. Adjusted p-value for Sometimes-Frequently is 1e-04 < 0.05, indicating a significant difference. Adjusted p-value for Sometimes-no is <2e-16 < 0.05, indicating a significant difference.

i)BMI by MTRANS

Dunn's test of multiple comparisons using rank sums : bonferroni

Output: Adjusted p-value for Bike-Automobile is 1.0000 > 0.05, indicating no significant difference. Adjusted p-value for Motorbike-Automobile is 1.0000 > 0.05, indicating no significant difference. Adjusted p-value for Public_Transportation-Automobile is 1.0000 > 0.05, indicating no significant difference. Adjusted p-value for Walking-Automobile is 1.6e-07 < 0.05, indicating a significant difference. Adjusted p-value for Motorbike-Bike is 1.0000 > 0.05, indicating no significant difference. Adjusted p-value for Public_Transportation-Bike is 0.7154 > 0.05, indicating no significant difference. Adjusted p-value for Walking-Bike is 1.0000 > 0.05, indicating no significant difference. Adjusted p-value for Public_Transportation-Motorbike is 0.6929 > 0.05, indicating no significant difference. Adjusted p-value for Walking-Motorbike is 1.0000 > 0.05, indicating no significant difference. Adjusted p-value for Walking-Public_Transportation is 2.2e-09 < 0.05, indicating a significant difference.

Ans to the Question number 18

Code of test whether there is significant association between BMI and (i) gender, (ii) family history with overweight, (iii) smoking,

(iv) FAVC, (v) CAEC, (vi)SCC, (vii) CALC and (viii) MTRANS.

Code:

```
#Question 18 i

b<-kruskal.test(data$BMI~data$Gender)

b

b$p.value

print(paste('P value =',b$p.value))

ifelse(b$p.value<=0.05,"There is association between bmi and gender","There is no association between bmi and gender")

#Question 18 ii

b<-kruskal.test(data$BMI~data$family_history_with_overweight)

b

b$p.value

print(paste('P value =',b$p.value))

ifelse(b$p.value<=0.05,"There is association between bmi and family_history_with_overweight","There is no association between bmi and family_history_with_overweight")

#Question 18 iii

b<-kruskal.test(data$BMI~data$SMOKE)

b

b$p.value

print(paste('P value =',b$p.value))

ifelse(b$p.value<=0.05,"There is association between bmi and smoke","There is no association between bmi and smoke")

#Question 18 (iv)

b<-kruskal.test(data$BMI~data$FAVC)

b

b$p.value

print(paste('P value =',b$p.value))

ifelse(b$p.value<=0.05,"There is association between bmi and FAVC","There is no association between bmi and FAVC")

#Question 18(iv)

b<-kruskal.test(data$BMI~data$CAEC)
```

```
b

b$p.value

print(paste('P value =',b$p.value))

ifelse(b$p.value<=0.05,"There is association between bmi and CAEC","There is no association between bmi and CAEC")

#Question 18(v)

b<-kruskal.test(data$BMI~data$SCC)

b

b$p.value

print(paste('P value =',b$p.value))

ifelse(b$p.value<=0.05,"There is association between bmi and SCC","There is no association between bmi and SCC")

#Question 18(vi)

b<-kruskal.test(data$BMI~data$CALC)

b

b$p.value

print(paste('P value =',b$p.value))

ifelse(b$p.value<=0.05,"There is association between bmi and CALC","There is no association between bmi and CALC")

#Question18(vii)

b<-kruskal.test(data$BMI~data$MTRANS)

b

b$p.value

print(paste('P value =',b$p.value))

ifelse(b<=0.05,"There is association between bmi and MTRANS","There is no association between bmi and MTRANS")
```

Output:

a) BMI by Gender

Hypothesis

H0: There is no association between bmi and gender

HA: There is association between bmi and gender

Test Statistic, Kruskal-Wallis chi-squared = 0.67516

p-value = 0.4113

In Kruskal-Wallis chi-squared test we cannot include 95% confidence interval

Comment: Here p-value = 0.4113 which is greater than 0.05 ,So we cannot reject the null hypothesis, So we accept null hypothesis.

So, There is no association between bmi and gender

b) BMI and family history with overweight

Hypothesis

H0: There is no association between bmi and family history with overweight

HA: There is association between bmi and family history with overweight

Test Statistic, Kruskal-Wallis chi-squared = 524

P value =5.71162484725321e-116

In Kruskal-Wallis chi-squared test we cannot include 95% confidence interval

Comment: Here p value =5.71162484725321e-116 which is less than 0.05. So, we can reject the null hypothesis. We accept alternative hypothesis. So, there is association between bmi and family history with overweight

c) BMI and Smoke

Hypothesis

H0: There is no association between bmi and smoke

HA: There is association between bmi and smoke

Test Statistic, Kruskal-Wallis chi-squared = 0.010889

p-value = 0.9169

In Kruskal-Wallis chi-squared test we cannot include 95% confidence interval

Comment: Here P value = 0.9169, which is greater than 0.05. So we cannot reject the null hypothesis .We accept null hypothesis .So, There is no association between bmi and smoke

d) BMI and FAVC

Hypothesis

H0: There is no association between bmi and FAVC

HA: There is association between bmi and FAVC

Test Statistic, Kruskal-Wallis chi-squared = 131.06

P value = 2.40687632343042e-30

In Kruskal-Wallis chi-squared test we cannot include 95% confidence interval

Comment: Here P value = 2.40687632343042e-30 which is less than 0.05. So, we can reject the null hypothesis. We accept alternative hypothesis. So, there is association between bmi and FAVC

e) BMI and CAEC

Hypothesis

H0: There is no association between bmi and CAEC

HA: There is association between bmi and CAEC

Test Statistic, Kruskal-Wallis chi-squared = 409.66

P value = 1.79018730009495e-88

In Kruskal-Wallis chi-squared test we cannot include 95% confidence interval

Comment: Here P value =1.79018730009495e-88 which is less than 0.05. So, we can reject the null hypothesis. We accept alternative hypothesis. So, there is association between bmi and CAEC

f) BMI and SCC

Hypothesis

H0: There is no association between bmi and SCC

HA: There is association between bmi and SCC

Test Statistic, Kruskal-Wallis chi-squared = 85.414

P value = 2.41985461478364e-20

In Kruskal-Wallis chi-squared test we cannot include 95% confidence interval

Comment: Here, P value = 2.41985461478364e-20, which is less than 0.05. So, we can reject the null hypothesis. We accept alternative hypothesis. So, there is association between bmi and SCC.

g) BMI and CALC

Hypothesis

H0: There is no association between bmi and CALC

HA: There is association between bmi and CALC

Test Statistic, Kruskal-Wallis chi-squared = 98.299

P value = 3.60798558751883e-21

In Kruskal-Wallis chi-squared test we cannot include 95% confidence interval

Comment: Here we see that, P value = 3.60798558751883e-21 which is less than 0.05. So, we can reject the null hypothesis. We accept alternative hypothesis. So, there is association between bmi and CALC.

h) BMI and MTRANS

Hypothesis

H0: There is no association between bmi and MTRANS

HA: There is association between bmi and MTRANS

Test Statistic, Kruskal-Wallis chi-squared = 46.479

p-value = 1.958e-09

In Kruskal-Wallis chi-squared test we cannot include 95% confidence interval

Comment: Here we see that, p-value = 1.958e-09 which is less than 0.05, we can reject the null hypothesis. We accept alternative hypothesis. So, there is association between bmi and MTRANS.

Ans to the Question number 19

The code of Creating a new dataset (give name as: obesity_small) from the obesity dataset keeping the variables only: Gender, Age, Height, Weight, family history with overweight, FAVC, CH20, SCC, FAF, TUE and CALC.is

Code:

```
obesity_small<-select(data,'Gender','Age','Height','Weight','family_history_with_overweight','FAVC','CH2O', 'SCC','FAF','TUE','CALC')

View(obesity_small)
```

The code of Creating another dataset (namely: obesity_new) by dropping the variables FAF, TUE and CALC from the obesity_small dataset.

Code:

```
obesity_new<-select(obesity_small,-'FAF',-'TUE',-'CALC')

View(obesity_new)
```

Output: The 2 data name obesity_small and obesity_new in my RStudio

The code of fitting a multiple linear regression model of Weight on Gender, Age, Height, family history with overweight, FAVC, CH20 and SCC. usign

Code:

```
model1<-lm(Weight~Gender+Age+Height+family_history_with_overweight+FAVC+CH2O+SCC,data = obesity_new)

model1
```

The code of Interpreting the outputs of the model including estimates of the parameters.

Code:

```
modsum<-summary(model1)

modsum

confint(model1, level = 0.95)
```

Output:

Predicted Outcome=β0+β1(Gender)+β2(Age)+β3(Height)+β4(Family History)+β5(FAVC)+β6(CH2O)+β7(SCC)

The estimated regression coefficients

Coefficients:

	Estimate	Standard Error	t-value	Pr(>|t|)
Intercept	-172.05064	10.20741	-16.855	< 2e-16 ***
Gender(Male)	-9.47279	1.10618	-8.564	< 2e-16 ***
Age	0.59907	0.07032	8.519	< 2e-16 ***
Height	126.87986	6.23585	20.347	< 2e-16 ***
Family history with Overweight	22.13712	1.21548	18.213	< 2e-16 ***
FAVCyes	9.29533	1.40583	6.612	4.79e-11 ***
CH2O	3.48173	0.72584	4.797	1.72e-06 ***
SCCyes	-7.73917	2.14661	-3.605	0.000319 ***

Substituting Coefficients:

Weight =−172.0506−9.4728(Gender:Male)+0.5991(Age)+126.8799(Height)+22.1371(FamilyHistory:Yes)+9.2953(FAVC:Yes)+

3.4817(CH2O)−7.7392(SCC: Yes)

Output:

Residuals:

Minimum	1st Quartile	Median	3rd Quartile	Maximum
-58.398	-13.470	-0.041	15.244	69.079

Residual standard error=19.78, degrees of freedom=2103

Multiple coefficient of determination R-squared = 0.4314,

Comment: Here 43% of Weight is explained by the explanatory variables of my model

Adjusted R-squared= 0.4295

F-statistic= 228 on 7 and 2103 DF, p-value: < 2.2e-16

Comment: This is the overall model p value which is less than 0.05 ,so overall model is highly significant

95% confidence interval

	2.5%	97.5%
Intercept	-192.0683152	-152.0329584
GenderMale	-11.6421022	-7.3034763
Age	0.4611633	0.7369696
Height	114.6507856	139.1089310
family_history_with_overweightyes	19.7534549	24.5207903
FAVCyes	6.5383663	12.0522890
CH2O	2.0583046	4.9051650
SCCyes	-11.9488713	-3.5294674

Code of Prediction the weight of a 23 years old male respondent having family history with overweight, Height = 1.77, FAVC = yes, CH20 = 1 and SCC = no.

Code:

```
prediction<-data.frame(Gender = 'Male',Age = 23,family_history_with_overweight='yes',Height =1.77,FAVC='yes',CH2O=1,SCC='no')

predict(model1,newdata = prediction,interval = "confidence", level = 0.95)
```

Output:

Fit	Lower	Upper
91.74664	89.70848	93.78479

Multiple coefficient of determination R^2

Code:

```
 print(paste("R square =",modsum$r.squared))

print(paste("Adjusted R square=",modsum$adj.r.squared))
```

Output:

Multiple coefficient of determination,R square = 0.431430029693384

Comment: Here also 43% of Weight is explained by the explanatory variables of my model(predict)

Adjusted R square= 0.429537500072772

Ans to the Question number 20

20(i)

Code for again read the data ,calculating the BMI and fitting linear model of the variables

Code:

```
# Load the dataset

data1 <- read.csv(file.choose(),header = T)

View(data1)

# Compute BMI

data1$BMI <- data1$Weight / (data1$Height^2)

# Convert categorical variables to factors

categorical_vars <- c("Gender", "family_history_with_overweight", "FAVC", "CAEC", "SMOKE", "SCC", "CALC", "MTRANS")

data1[categorical_vars] <- lapply(data1[categorical_vars], as.factor)

# Initial full model

full_model <- lm(BMI ~ ., data = data1)

full_model
```

Output: The BMI in included my RStudio data1

The calculated BMI in My data1 datafile and the full model in the below

BMI = 56.019507 + 0.134163 * GenderMale + 0.002189 * Age - 33.210622 * Height + 0.331226 * Weight + 0.597874 * family_history_with_overweightyes + 0.258939 * FAVCyes + 0.282684 * FCVC + 0.190437 * NCP - 0.437410 * CAECFrequently + 0.320667 * CAECno + 0.102193 * CAECSometimes - 0.387064 * SMOKEyes + 0.007996 * CH2O - 0.294291 * SCCyes - 0.096685 * FAF - 0.021420 * TUE - 0.209971 * CALCFrequently - 0.278607 * CALCno - 0.478306 * CALCSometimes + 0.438963 * MTRANSBike + 0.423799 * MTRANSMotorbike - 0.059481 * MTRANSPublic_Transportation + 0.168390 * MTRANSWalking

Code of stepwise selection using AIC,Model Performance with confidence interval and calculating Predicted values and residuals is below

Code:

```
# Stepwise selection using AIC

stepwise_model <- stepAIC(full_model, direction = "both", trace = FALSE)

confint(stepwise_model, level = 0.95) summary(stepwise_model)

# Model performance

AIC(stepwise_model)

# Predicted values and residuals

data1$Predicted <- predict(stepwise_model)

data1$Residuals <- residuals(stepwise_model)
```

Output:

The stepwise model is below

BMI = 56.12 + 0.1368 * GenderMale - 33.25 * Height + 0.3315 * Weight + 0.6002 * family_history_with_overweightyes + 0.2531 * FAVCyes + 0.2855 * FCVC + 0.1889 * NCP - 0.4341 * CAECFrequently + 0.3364 * CAECno + 0.1019 * CAECSometimes - 0.3833 * SMOKEyes - 0.2975 * SCCyes - 0.09745 * FAF - 0.1816 * CALCFrequently - 0.2517 * CALCno - 0.4485 * CALCSometimes + 0.4303 * MTRANSBike + 0.4203 * MTRANSMotorbike - 0.08644 * MTRANSPublic_Transportation + 0.1404 * MTRANSWalking

Multiple R-squared= 0.9915 which is indicates that 99% variation of Weight is explained by the explanatory variables of stepwise model

F-statistic=1.215e+04 P value of F-statistic is < 2.2e-16 which is also less than 0.05 .It indicates that the overall model is significant

Adjusted R-squared: 0.9914

95% confidence interval of this stepwise model

Variable	Coefficient	2.50%	97.50%
(Intercept)	56.12	54.42	57.81
GenderMale	0.1368	0.0483	0.2252
Height	-33.25	-33.8	-32.71
Weight	0.3315	0.3297	0.3333
family_history_with_overweightyes	0.6002	0.5004	0.7
FAVCyes	0.2531	0.145	0.3612
FCVC	0.2855	0.2196	0.3513
NCP	0.1889	0.1457	0.2321
CAECFrequently	-0.4341	-0.6582	-0.2101
CAECno	0.3364	0.0425	0.6303
CAECSometimes	0.1019	-0.1061	0.3099
SMOKEyes	-0.3833	-0.6083	-0.1583
SCCyes	-0.2975	-0.4576	-0.1375
FAF	-0.09745	-0.1388	-0.0561
CALCFrequently	-0.1816	-1.67	1.307
CALCno	-0.2517	-1.73	1.227
CALCSometimes	-0.4485	-1.928	1.031
MTRANSBike	0.4303	-0.1297	0.9902
MTRANSMotorbike	0.4203	-0.0289	0.8696
MTRANSPublic_Transportation	-0.08644	-0.1665	-0.0064
MTRANSWalking	0.1404	-0.0738	0.3545

The AIC of the model is 4761.973

predicted values and the residuals in My data1

20(ii) Perform model adequacy checking (check model assumptions)

Code of checking Assumption

Code:

Normality check

```
b<-shapiro.test(data1$Residuals) #Shapiro wilk test

b

P_value<-b$p.value

P_value

ifelse(P_value>0.05,"Normal","Not Normal")


install.packages("lmtest")

library(lmtest)
```

Homoscedasticity check

```
b<-bptest(stepwise_model) # Breusch-Pagan Test (from lmtest package)

b

p_value<-b$p.value

p_value

ifelse(p_value>0.05,"Homoscedasticity present","Heteroscedasticity present")
```

Multicollinearity test

```
vif(stepwise_model)

plot(stepwise_model)
```

Output:

Assumption Testing:

Normality check: By shapiro wilk test we see that test statistic, W = 0.98977, p-value = 4.375e-11 for the residuals . Here p value is less than 0.05 . So it is not following the normal distribution .So the distribution is not normal .

Homoscedasticity check: By studentized Breusch-Pagan test ,we have test statistic, BP = 282.5 , P value 2.980107e-48 which is less than 0.05 .So it is not present the Homoscedasticity .There are Heteroscedasticity present.

Multicollinearity check: By checking Variance Inflation Factor(vif) for stepwise_model we have only for Height and Weights Variance Inflation Factor is greater than 2 which is not mejor Multicollinearity this is called mild Multicollinearity ,So we can say that there is no multicollinearity for avoid.

20(iii) **Outlier detection and check influential**

Code:

```
influence <- influence.measures(stepwise_model)

cooksD <- cooks.distance(stepwise_model)

cooksD

cooks_out<-which(cooksD >1)

cooks_out

influential <- which(cooksD > (4 / nrow(data)))

influential
```

Output: Here I see that by cook.distance method there is no cook distance is greater than 1 so there is no outlier in this model . By checking influential of the model we have there are few influential the elements numbers is

7 , 51, 74 , 77, 84, 114 ,142, 143, 157, 164, 177, 179, 192, 194, 198, 203, 218, 219, 235, 242, 265 ,278, 292, 303, 340, 7, 51, 74, 77,84, 114, 142 ,143, 157, 164, 177, 179, 192, 194, 198, 203, 218, 219, 235, 242, 265, 278, 292, 303, 340 ,345, 347, 357 , 396, 430, 465, 494, 503, 628

Ans to the Question number 21

The code of drawing a random sample of size n = 50 from the Obesity Dataset (ObesityData.csv) using the last three digits of my registration number as seed number.

Code:

```
main_data<-read.csv(file.choose(),header = T)

View(main_data)

set.seed(115)

sampled_data<-main_data[sample(nrow(main_data),50,replace =F),]

sampled_data

View(sampled_data)
```

Here I read ObesityData.csv as main_data and

I draw sample of 50 as sampled_data

Repeatation of the tasks 10

The code of Calculating body mass index (BMI) by using the formula, BMI = Weight/(Height)2. Make a categorical variable (name as BMI_cat) using the following categorization of BMI values: Less than 18.5 as ‘Underweight’; 18.5 to 24.9 as ‘Normal’; 25.0 to 29.9 as ‘Overweight’; Greater than 30 as ‘Obesity’.

Code:

```
sampled_data$BMI<-sampled_data$Weight/(sampled_data$Height^2)

sampled_data$BMI_cat<-cut(sampled_data$BMI,breaks = c(0,18.5,24.9,29.9,Inf),labels
=c("Underweight","Normal","Overweight","Obisity"))
```

The sampled_data in my Rstudio

Repeatation of the tasks 11

The code of Calculating BMI for respondents (i) whose age > 30 years, (ii) who are non-smokers, have physical activity of 2 days and drink more than 1 liter of water daily is ,

Code:

```
subset_data1 <- sampled_data %>% filter(Age > 30, SMOKE == "no", FAF >= 2, CH2O >= 1)

View(subset_data1)

mean(subset_data1$BMI)
```

Here there is no data in this data set

we read this data ase subset_data1,and this data included in my RStudio

Repeatation of the tasks 12

The code of Creating a new dataset (name as: obesity_sub1) by taking the respondents whose height is more than 1.8 meter and who eat high caloric food frequently. Calculating mean and standard deviation of BMI using the obesity_sub1 dataset.

Code:

```
obesity_sub1 <- sampled_data %>% filter(Height > 1.8, FAVC == "yes")

View(obesity_sub1)

mean(obesity_sub$BMI)

sd(obesity_sub$BMI)
```

The mean of BMI of the obesity_sub1 data set = 30.44498

The standard deviation of BMI of the obesity_sub1 data set= 7.630551

Repeatation of the tasks 13

The code of Calculating correlation between (i) age and weight, (ii) age and height and (iii) height and weight is

Code:

```
cor(sampled_data$Age, sampled_data$Weight, use = "complete.obs")

cor(sampled_data$Height, sampled_data$Weight, use = "complete.obs")

cor(sampled_data$Age,sampled_data$Height, use = "complete.obs")
```

Output:

The correlation coefficient of age and weight = 0.04950611

The correlation coefficient of Height and weight = 0.5374769

The correlation coefficient of age and Height = -0.1190911

Repeatation of the tasks 14

Code:

```
correlation_value <- cor(sampled_data$Age, sampled_data$BMI, use = "complete.obs")

correlation_test<- cor.test(sampled_data$Age, sampled_data$BMI,use ="complete.obs",conf.level = 0.95)

correlation_test

print(paste("Correlation:", correlation_value))

p_value<-correlation_test$p.value

print(paste("p_value:",p_value))
```

Output:

Correlation coefficient = 0.114757026489956

The code of significance test is

Code:

```
if (p_value <= 0.05) {
  print("Correlation is significant (not equal to 0)")
} else {
  print("Correlation is not significant (close to 0)")
}
```

Output:

Here ,

Hypothesis

H0: Correlation is not significant (close to 0)")

HA: Correlation is significant (not equal to 0)")

Test Statistic, t = 0.80035

P value = 0.4275

95% Confidence Interval = -0.1689886 and 0.3809367

Comment: Here we see that , P value = 0.4275 which is greater than 0.05 ,so we cannot reject the null hypothesis.We accept Null hypothesis .so , Correlation is not significant (close to 0)

Code of Calculating correlation matrix of all the numerical variables

Code:

```
numerical_variables <- c('Age','Height','Weight','FCVC','NCP','CH2O','FAF','TUE')
cor_matrix <- cor(sampled_data[numerical_variables], use = "complete.obs")
print(cor_matrix)
```

Output:

	Age	Height	Weight	FCVC	NCP	CH2O	FAF	TUE
Age	1.00000000	-0.1190911	0.04950611	-0.08221408	-0.01849564	-0.09681727	-0.21183567	-0.48907243
Height	-0.1190911	1.0000000	0.53747686	0.11215433	0.17513189	0.31597673	0.27827950	0.01014080
Weight	0.04950611	0.5374769	1.00000000	0.33468350	0.08577468	0.07899164	0.10179823	-0.16576368
FCVC	-0.08221408	0.1121543	0.33468350	1.00000000	-0.11587724	-0.21245055	0.16098642	-0.02090286
NCP	-0.01849564	0.1751319	0.08577468	-0.11587724	1.00000000	0.27959657	0.19808246	-0.03155993
CH2O	-0.09681727	0.3159767	0.07899164	-0.21245055	0.27959657	1.00000000	0.10921159	-0.01692943
FAF	-0.21183567	0.2782795	0.10179823	0.16098642	0.19808246	0.10921159	1.00000000	0.12668490
TUE	-0.48907243	0.0101408	-0.16576368	-0.02090286	-0.03155993	-0.01692943	0.12668490	1.00000000

Repeatation of the tasks 16i

Code:

```
b<-wilcox.test(sampled_data$Age~sampled_data$Gender,exact=F,conf.int = TRUE, conf.level = 0.95)

b

b$p.value

print(paste('P value =',b$p.value))

ifelse(b$p.value<=0.05 ,"Average age significantly differs between male and female","Average age doesn't significantly differ between male and female ")

b<-wilcox.test(sampled_data$Height~sampled_data$Gender,exact =F,conf.int = TRUE, conf.level = 0.95)

b

b$p.value

print(paste('P value=',b$p.value))

ifelse(b<=0.05,"Average height significantly differs between male and female","Average height doesn't significantly differ between male and female")

b<-wilcox.test(sampled_data$BMI~sampled_data$Gender,conf.int = TRUE, conf.level = 0.95)

b

b$p.value

print(paste('P valu =',b$p.value))

ifelse(b$p.value<=0.05,"Average BMI significantly differs between male and female","Average BMI doesn't significantly differ between male and female")
```

Output:

a)Age by Gender:

Hypothesis

H0: Average age doesn't significantly differ between male and female

HA: Average age significantly differs between male and female

Test Statistic, W = 307,

p-value = 0.9303

95% confidence interval = -4.021850 to 3.000011

Comment: Here we see that p-value = 0.9303 which is greater than 0.05 .So we cannot reject the null hypothesis .We accept null hypothesis .So, Average age doesn't significantly differ between male and female

b)Height by Gender:

Hypothesis

H0: Average height doesn't significantly differ between male and female

HA: Average height significantly differs between male and female

Test Statistic, W = 135.5

p-value = 0.0006309

95% confidence interval = -0.12923347 to -0.03950672

Comment: Here we see that p-value = 0.0006309 which is less than 0.05 . So we can reject the null hypothesis.We accept alternative hypothesis.So, Average height significantly differs between male and female

c) BMI by Gender:

Hypothesis

H0: Average BMI doesn't significantly differ between male and female

HA: Average BMI significantly differs between male and female

Test Statistic, W = 383.5

p-value = 0.168

95% confidence interval: -1.855854 to 9.006952

Comment: Here we see that p-value = 0.168 which is greater than 0.05 .So we cannot reject the null hypothesis.We accept the null hypothesis.So , Average BMI doesn't significantly differ between male and female.

Repeatation of the tasks 16ii

Code:

```
b<-wilcox.test(sampled_data$Age~sampled_data$SMOKE,conf.int = TRUE, conf.level = 0.95)

b

b$p.value

print(paste('P value =',b$p.value))

ifelse(b$p.value<=0.05,"Average age significantly differs between smoker and non smoker","Average age doesn't significantly differs between smoker and non smoker")

b<-wilcox.test(sampled_data$Height~sampled_data$SMOKE,conf.int = TRUE, conf.level = 0.95)

b

b$p.value

print(paste('P value=',b$p.value))

ifelse(b$p.value<=0.05,"Average height significantly differ between smoker and non smoker","Average height doestn't significantly differ between smoker and non smoker")

b<-wilcox.test(sampled_data$BMI~sampled_data$SMOKE,conf.int = TRUE, conf.level = 0.95)

b

b$p.value

print(paste('P value = ',b$p.value))

ifelse(b$p.value<=0.05,"Average BMI significantly differ between smoker and non smoker","Average BMI doesn't significantly differ between smoker and non smoker")
```

Output:

a)Age by smoke

Hypothesis

H0: Average age doesn't significantly differs between smoker and non smoker

HA: Average age significantly differs between smoker and non smoker

Test Statistic, W = 35.5

p-value = 0.4664

95% confidence interval = -3.071209 to 35.000000

Comment: Here we see that p-value = 0.4664 which is greater than 0.05. So, we cannot reject the null hypothesis. We accept null hypothesis. So, Average age doesn't significantly differ between smoker and non-smoker

b) Height by smoke:

Hypothesis

H0: Average height doesn't significantly differ between smoker and non-smoker

HA: Average BMI significantly differ between smoker and non-smoker

Test Statistic, W = 40.5

p-value = 0.2827

95% confidence interval= -0.067357 to 0.330000

Comment: Here we see that, p-value = 0.2827, which is greater than 0.05. So, we cannot reject the null hypothesis. We accept null hypothesis. So, Average height doesn't significantly differ between smoker and non-smoker.

c) BMI by Smoke:

Hypothesis

H0: Average BMI doesn't significantly differ between smoker and non smoker

HA: Average BMI significantly differ between smoker and non smoker

W = 47

p-value = 0.1274

95% confidence interval= -4.074016 to 30.482773

Comment: Here we see that , p-value = 0.1274,which is greater than 0.05 , So, we cannot reject the null hypothesis. We accept null hypothesis.So, Average BMI doesn't significantly differ between smoker and non smoker