```
#1

# Load the dataset
Cereal_Data <- read.csv("Cereal_Data.csv", header=TRUE, sep=",")  # Assuming the dataset is stored
in a file named "Cereal_Data.csv"


# Standard Deviation and Variance for Calories
calories_sd <- sd(Cereal_Data$Calories)
calories_var <- var(Cereal_Data$Calories)

# Standard Deviation and Variance for Sugar
sugar_sd <- sd(Cereal_Data$Sugar)
sugar_var <- var(Cereal_Data$Sugar)

# Geometric Mean for Calories and Sugar
library(MASS)
NROW(Cereal_Data)

library(psych) # Needed for the Geometric Mean function below

# Assign the values from the Units column to new data table called units
hours <- Cereal_Data$Calories

# calculate Geometric mean by hand using the long hand formula
exp(mean(log(hours)))
# use the library call to do the calculation
calories_geo_mean <- geometric.mean(Cereal_Data$Calories ,na.rm=TRUE)

sugar_geo_mean <- geometric.mean(Cereal_Data$Sugars, na.rm = TRUE)


#Weighted Mean
weighted_mean <- weighted.mean(Cereal_Data$Calories, Cereal_Data$Sodium, Cereal_Data$Fiber,
Cereal_Data$Carbs, Cereal_Data$Sugars)

# Print out the results
cat("Standard Deviation for Calories:", calories_sd, "\n")
cat("Variance for Calories:", calories_var, "\n")
cat("Standard Deviation for Sugar:", sugar_sd, "\n")
cat("Variance for Sugar:", sugar_var, "\n")
cat("Geometric Mean for Calories:", calories_geo_mean, "\n")
cat("Geometric Mean for Sugar:", sugar_geo_mean, "\n")
cat("Weighted Mean for everything:",weighted_mean, "\n" )


#2 Histogram of Calories & Sugar
library(ggplot2)

column <- Cereal_Data$Calories

ggplot(Cereal_Data, aes(x = column)) +
  geom_histogram(binwidth = 1, fill = 'skyblue', color = 'black')+
  labs(title = "Calories vs Sugar", x = "Calories", y = "Sugar") +
  theme_minimal()

#The histogram has a bell curve
#There is a clustering of values at 110 calories, meaning many cereals have the same calorie
content
#There is a normal distribution


#3 Plot a Box and Whisker Plot for Calories and a Box and Whisker Plot for Sugar
```

```
#box and whisker plot for calories:
boxplot(Cereal_Data$Calories)

#The median of calories is 110
#The outliers in calories is 160 and 40

#box and whisker for sugar:
boxplot(Cereal_Data$Sugars)

#The median of sugar is 6
#The outliers in sugars is 0 and 15

#4 Scatter plot for carbs vs calories


ggplot(Cereal_Data, aes(x = Carbs, y = Calories)) + geom_point()
  labs(title = "Carbs vs Calories",
       x = "Carbs",
       y = "Calories")
#There is a relationship between 15 carbs and 110 calories
#There is a strong linear relationship

#5 Use Regression Analysis to calculate a Regression Model

#Using the standard linear model and stepwise method to find the best model
  #that explains the relationship between calories and other variables
#Standard linear model
  library(dplyr)
  NROW(Cereal_Data$Calories)
  # show how many rows were read
  #
  #
  lmoutput <- lm(Calories ~  Sugars + Fiber + Carbs + Sodium, data = Cereal_Data)

  lmobj<-summary(lmoutput)
  summary(lmoutput)

  fstatistic = lmobj$fstatistic[1]
  #
  #Calculate the F-Critical value or look it up in the F-Distribution table
  #
  # This is testing the hypothesis that all Coefficients are zreo predictors
  #
  #df1data can be found from the df output in the 1st table entry (lmoutput)$df[1], subtract 1 for
n-1
  #df2data is n - (k-1) found from the residuals in the regression modeldf1data
  #n is the number of observations found from NROW
  #k is the number of independent variables in the regression equation
  #
  df2data <- df.residual(lmoutput)
  df1data <- summary(lmoutput)$df[1] - 1
  df1data
  df2data
  fcritical <- qf(.95, df1=df1data, df2=df2data)
  fcritical
  cat("F-Statistic is: ",lmobj$fstatistic[1],"F Critical is: ", fcritical, "\n")
  if_else(fstatistic > fcritical, "Reject The Null Hypothesis that coefficients are 0 predictors",
"Do Not Reject The Null Hypothesis")

  #Stepwise Method
  NROW(Cereal_Data$Calories)
  # show how many rows were read
  #
```

```
  step.model <- step(lm(Cereal_Data$Calories~Cereal_Data$Sugars + Cereal_Data$Fiber +
Cereal_Data$Sodium + Cereal_Data$Carbs,data=Cereal_Data),direction="both")
  summary(step.model)
  fstat <- summary(step.model)
  fstat$fstatistic[1]


 #The linear equation is Y=28.50359+3.75871X1+0.02682X2+3.14108X3+ε
#This is the best model for this data because sugars, sodium, and cards direct affect calories,
fiber does not.
#There is also an r2 value of 0.7373 which is close to one meaning it is a good model
#This model also determined that sugar is the biggest contributing coeffecient to calories, since
it has the highest t-value of 11.623.
 #The f-statistic is also very high at 58.94022 meaning that sugars, carbs, and sodiums have a
significant effect on calories
#The p value is very small which also justifies the rejection of the null hypothesis
```