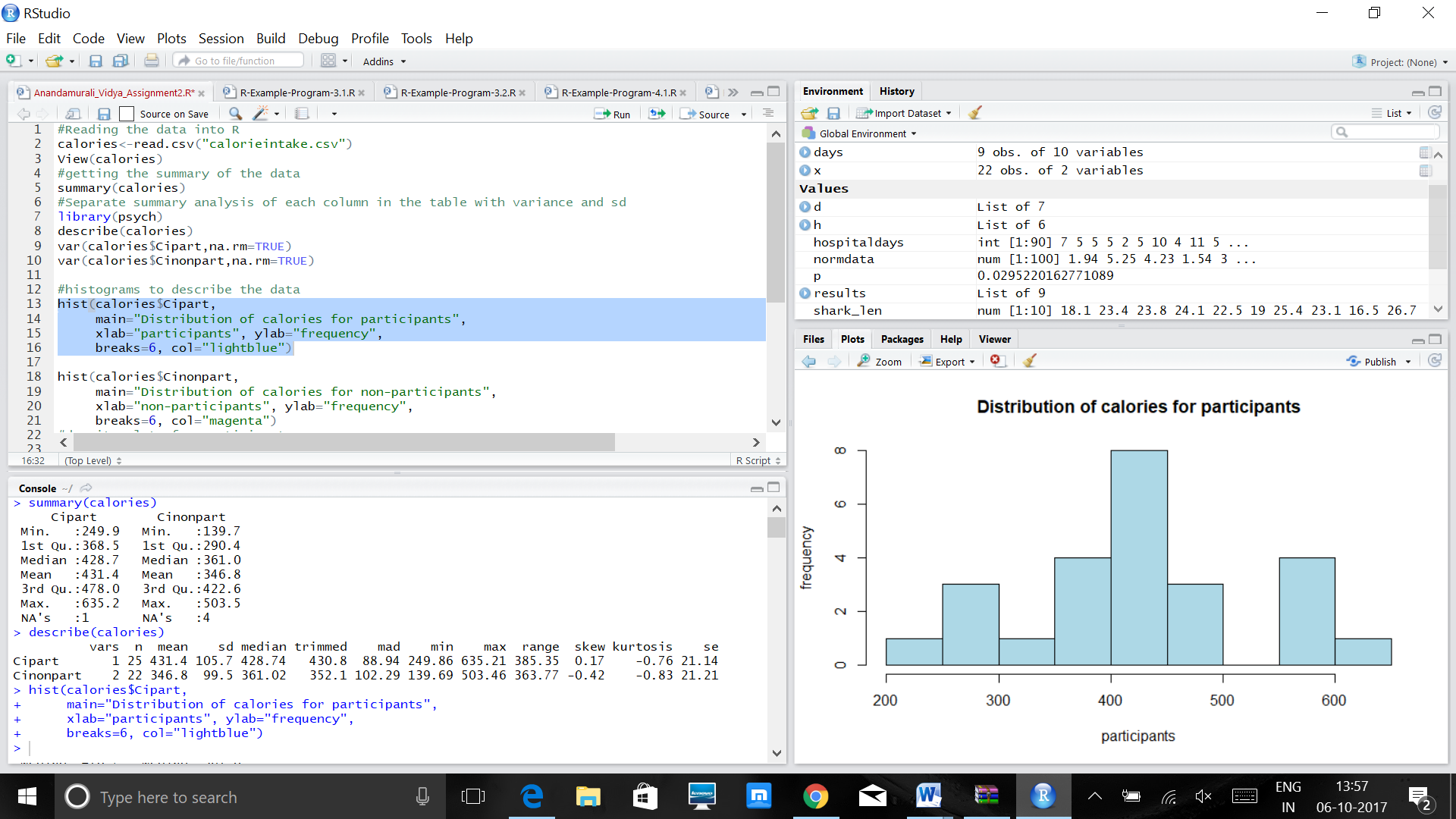
CS 555- Assignment 2  
BU ID – U55-32-1699

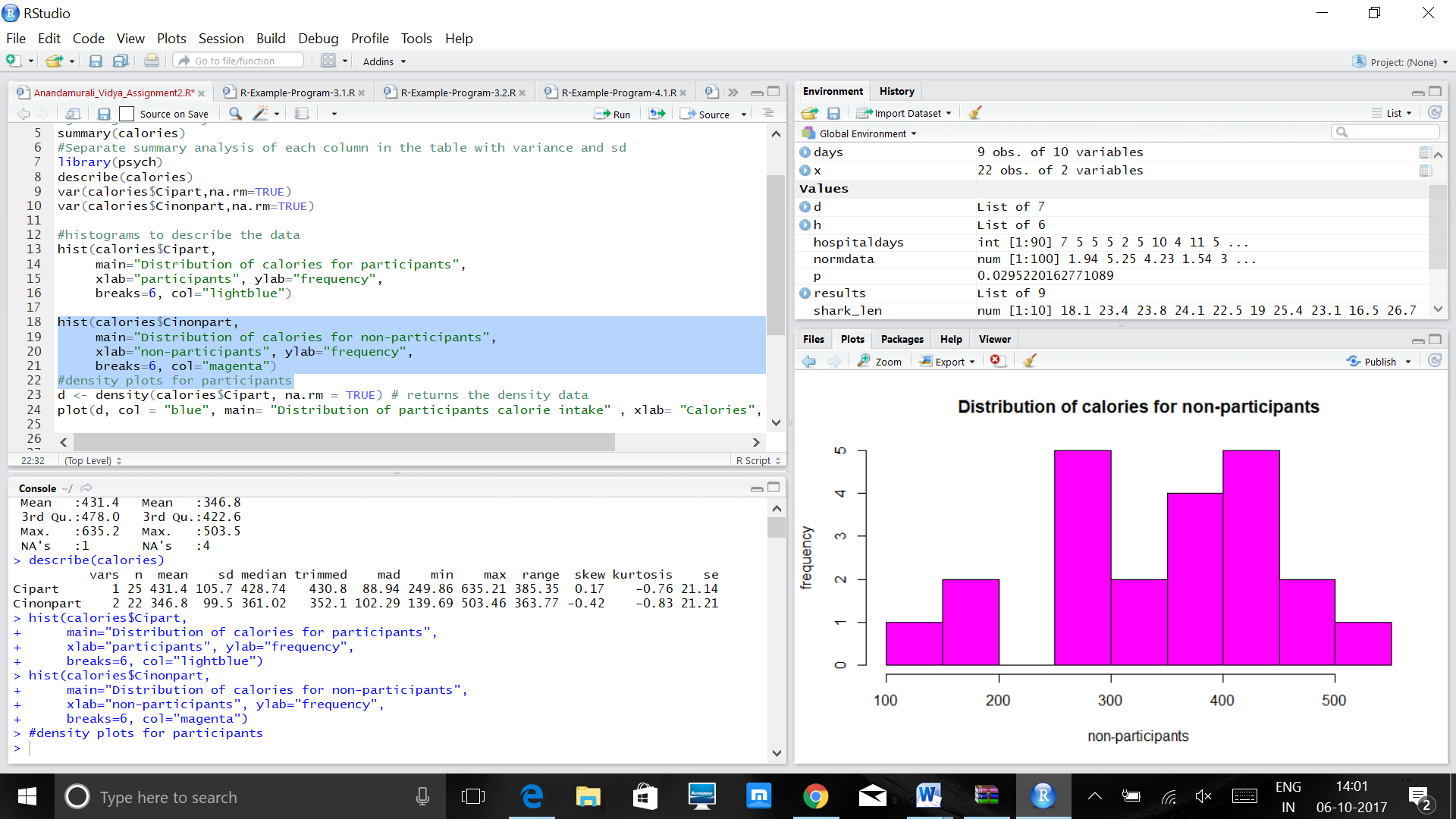
**(1)Data regarding the participants and non participants in the particular meal with respect to the calorie intake.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mean | SD | Median | 1st quartile | 3rd quartile | Min | Max |
| 431.4 | 105.7 | 428.7 | 368.5 | 478.0 | 249.9 | 635.2 |

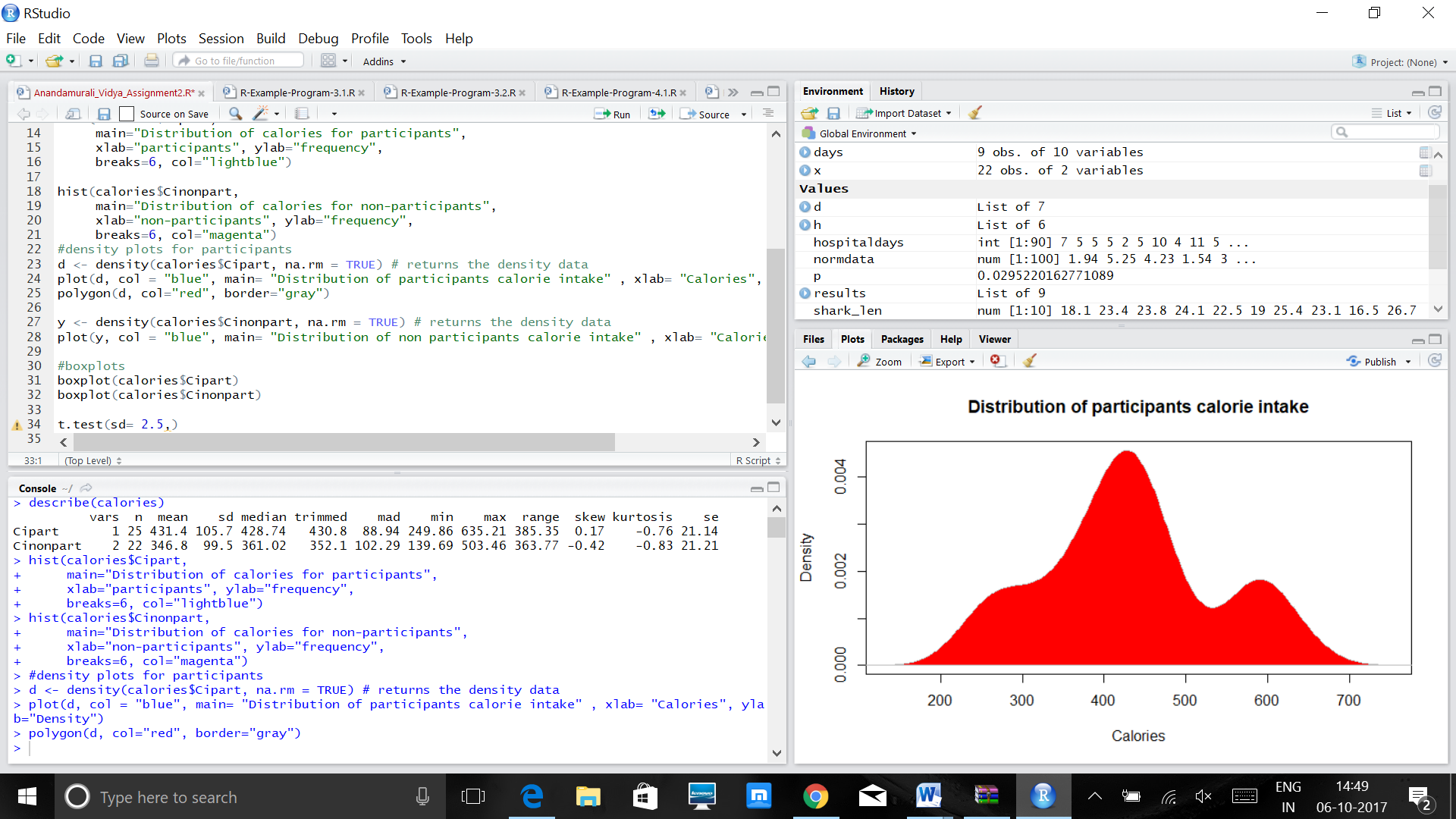
Summary of the complete set of people who participated :  
  
Summary of the set of people who did not participate :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mean | SD | Median | 1st quartile | 3rd quartile | Min | Max |
| 346.8 | 99.5 | 361.0 | 290.4 | 422.6 | 139.7 | 503.5 |

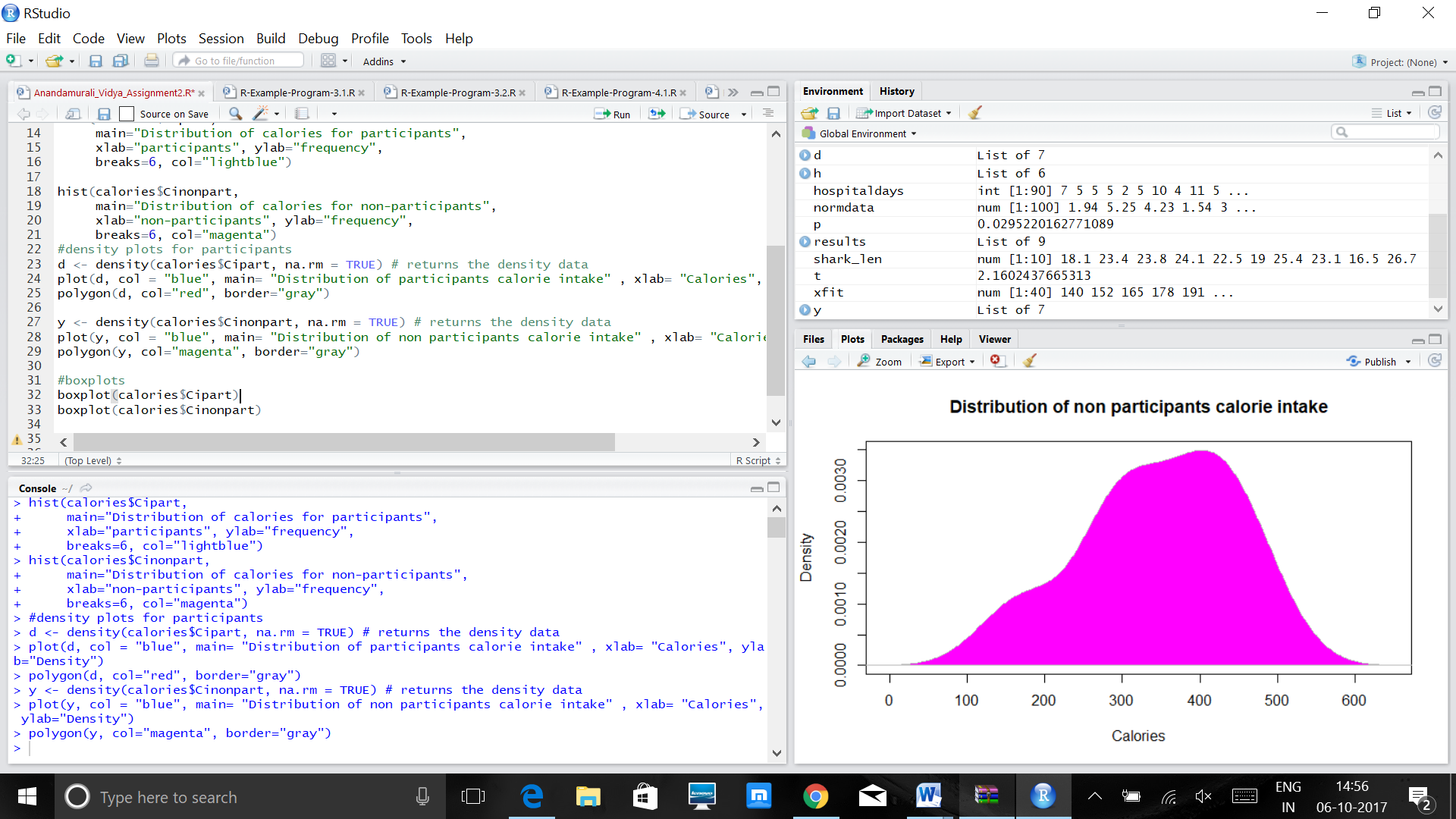
Visual representation of the data:  
In the form of an histogram:  
  
The above graph is for those who participated in the program, the histogram here , has been made to find out the calories that have been taken are around a maximum between the range of 400-450 calories. Also, from the graph, we do find that gaps do exist in the above histogram, which tells us that there are no values that are found within the range of 500-550 calories. Secondly, There is no specific skew present for this data, although outliers are present. Also, there is symmetry in the above representation of the data, as the mean and the median are almost equal to each other.



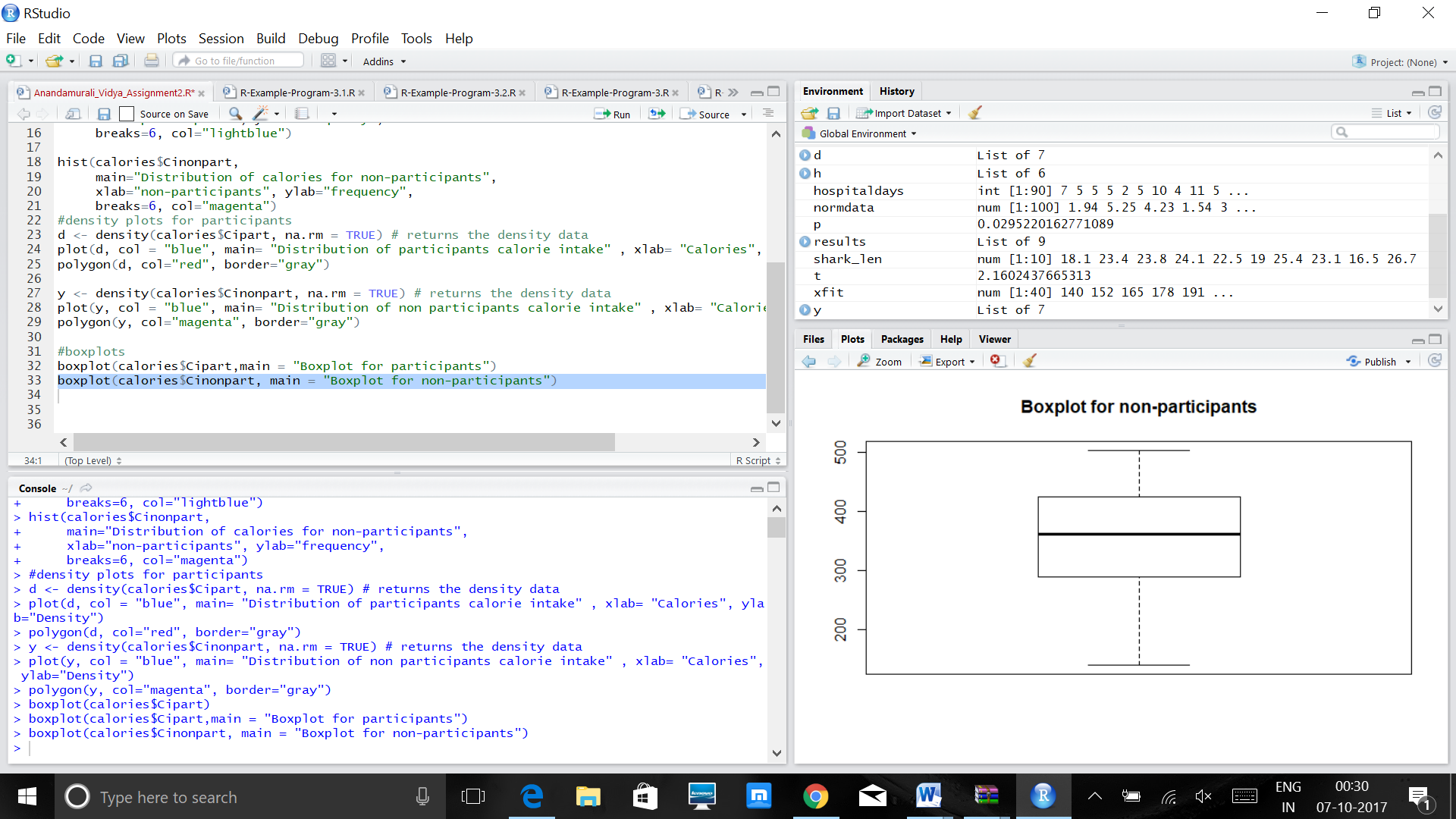
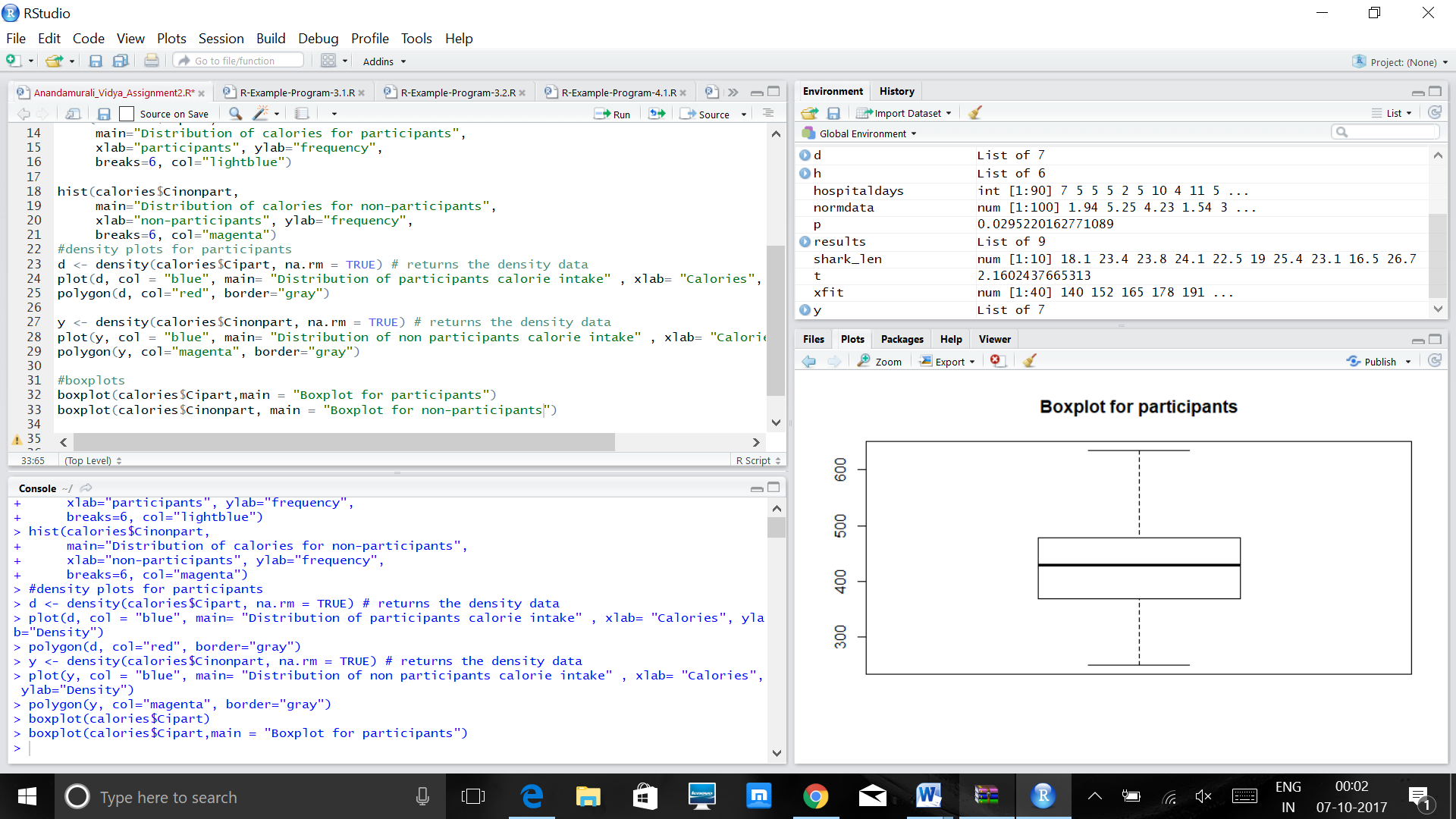
This set of data gives us an idea that the data has a gap between 200-250 calories and that no one had an intake of that value. Secondly, ranges from 250-300 and 400-450 have an equal frequency of 5 occurrences in the data set. The data here tells us that the range of the data is same for the two types of samples of the population. Also, the non participants have a maximum of 550 and the shape here is almost symmetrical with the median is slightly greater than the mean of the data.

Density Plots:  


From this graph, we have three peaks and this shows that the data is largely spread out. Secondly, the density is shown with respect to the statistics which almost represent symmetry and no such skew is present in the data.



The data here is represented not as a proper bell shaped curve, it looks like it is slightly skewed to the left and we can find out the density pertaining to the data with the plot made.

Boxplots:  


From the above boxplots it is evident that there are no outliers in the data presented and that the set of data followed a particular trend in them which didn’t involve very absurd values in the data set, which was one of the major reasons for the mean and the median of the data set to be almost similar to each other.

(2) **Does the mean calorie consumption for those who participated in the meal preparation differ from 425? Formally test at the level using the 5 steps outlined in the module.**

Step 1: Setting up the hypothesis:  
H0: μ = 425 – The mean of those who participated in the meal preparation do not have a value different than 425.

H1: μ ≠ 425 – The mean of those who participated in the meal preparation have a value different than 425.

Step 2: t = (x¯−μ / (s/√n)) since the n is small (less than 30) and the population standard deviation is not known. Here we use a two sided test as we will be looking at the difference present in the data. Here, it can be either greater or less than the desired mean.

Step 3: The total n for the data is 25. The degree of freedom value is (n-1) = 25-1 =24. Here, the tail probability is α = 0.05/2 = 0.025. Using the table, we get the appropriate critical value = 2.064.  
Decision Rule: Reject H0 :H0 if t ≥ 2.064  
  
Step 4: Computing the t statistic: t = 431.4 -425/(105.7/5) = 0.30272

One Sample t-test

data: calories$Cipart

t = 0.30272, df = 24, p-value = 0.7647

alternative hypothesis: true mean is not equal to 425

95 percent confidence interval:

387.7683 475.0309

sample estimates:

mean of x

431.3996

Step 5: Conclusion :  
Fail to reject H0 since p-value is greater than α.

We do not have significant evidence at α = 0.05 to prove that the mean is very different from the value given, 425.

We do not reject the null hypothesis that the mean differs from 425.

**(3) Calculate a 90% confidence interval for the mean calorie intake for participants in the meal preparation. Interpret the confidence interval.**

One Sample t-test

data: calories$Cipart

t = -1.8921e-05, df = 24, p-value = 1

alternative hypothesis: true mean is not equal to 431.4

90 percent confidence interval:

395.2311 467.5681

sample estimates:

mean of x

431.3996

The 90% **confidence interval** defines a range of values that can be 90% certain contains the population mean. The lower limit of the confidence interval is 395.3211 and the higher limit is 467.5681. The p-value is 1 which is not smaller than the alpha value and hence it is difficult to reject the null hypothesis.

**(4) Formally test whether or not participants consumed more calories than non-participants at the level using the 5 steps outlined in the module.**

Step 1: Set up the hypotheses and select the alpha level  
H0:μ1=μ2 (the mean calorie intake by participants is same as those who did not participate in the program.)  
H1:μ1> μ2 - (the mean calorie intake by participants is the greater than those who did not participate in the program.)  
α=0.05

Step 2: selecting the appropriate T – test:  
using the t statistics formula for two samples –

Step 3: using the formula, we get the t value as 2.8248. and here since our n1 and n2 values are different, we use the minimum degree of freedom from the two, which is n=22 for non participants which is lesser, therefore the df is 21. Therefore, the critical value is 1.721. In this case the t value is greater than the critical and the p value is less than the alpha value, therefore, here we can reject the null hypothesis.

Step 4:Using the code we arrive at:

Welch Two Sample t-test

data: calories$Cipart and calories$Cinonpart

t = 2.8248, df = 44.779, p-value = 0.00352

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

34.29842 Inf

sample estimates:

mean of x mean of y

431.3996 346.7991

Step 5: Conclusion:  
We can reject the null hypothesis for this particular alpha level which proves to be significant for this set of data. Therefore, here we have the alternate hypothesis that proves to be true, which states that the mean valve of the participants is greater than the mean calorie intake value of the non participants.

**(5).** The assumptions were made with respect to two hypothesis, one is that the mean of the two samples of the population are the same, secondly, the alternate hypothesis was that the mean of the first sample is bigger than that of the second sample, which proves to be true. Therefore, the null hypothesis is rejected here. The names command we used, helps us to know the parameters that are calculated in the process. Therefore, we have arrived at t test values and confidence intervals that we calculated manually.

**R CODE:**#Reading the data into R  
calories<-read.csv("calorieintake.csv")  
View(calories)  
#getting the summary of the data  
summary(calories)  
#Separate summary analysis of each column in the table with variance and sd  
library(psych)  
describe(calories)  
var(calories$Cipart,na.rm=TRUE)  
var(calories$Cinonpart,na.rm=TRUE)  
#histograms to describe the data  
hist(calories$Cipart,   
main="Distribution of calories for participants", xlab="participants", ylab="frequency", breaks=6, col="lightblue")  
hist(calories$Cinonpart,

main="Distribution of calories for non-participants",

xlab="non-participants", ylab="frequency",

breaks=6, col="magenta")

#density plots for participants  
d <- density(calories$Cipart, na.rm = TRUE) # returns the density data   
plot(d, col = "blue", main= "Distribution of participants calorie intake" , xlab= "Calories", ylab="Density")  
polygon(d, col="red", border="gray")

y <- density(calories$Cinonpart, na.rm = TRUE) # returns the density data   
plot(y, col = "blue", main= "Distribution of non participants calorie intake" , xlab= "Calories", ylab="Density")  
polygon(y, col="magenta", border="gray")

#boxplots

boxplot(calories$Cipart,main = "Boxplot for participants")  
boxplot(calories$Cinonpart, main = "Boxplot for non-participants")

# t tests

t.test(calories$Cipart, mu=425, alternative="two.sided", conf.level=0.95,na.rm =TRUE)

t.test(calories$Cipart,mu = 431.4, alternative="two.sided", conf.level=0.90)

#two samples test

t.test(calories$Cipart, calories$Cinonpart, alternative="greater", conf.level=0.95)

ttest <- t.test(calories$Cipart, calories$Cinonpart, alternative="greater", conf.level=0.95)

names(ttest)