MET CS555: Homework 2

Patrick Ryan



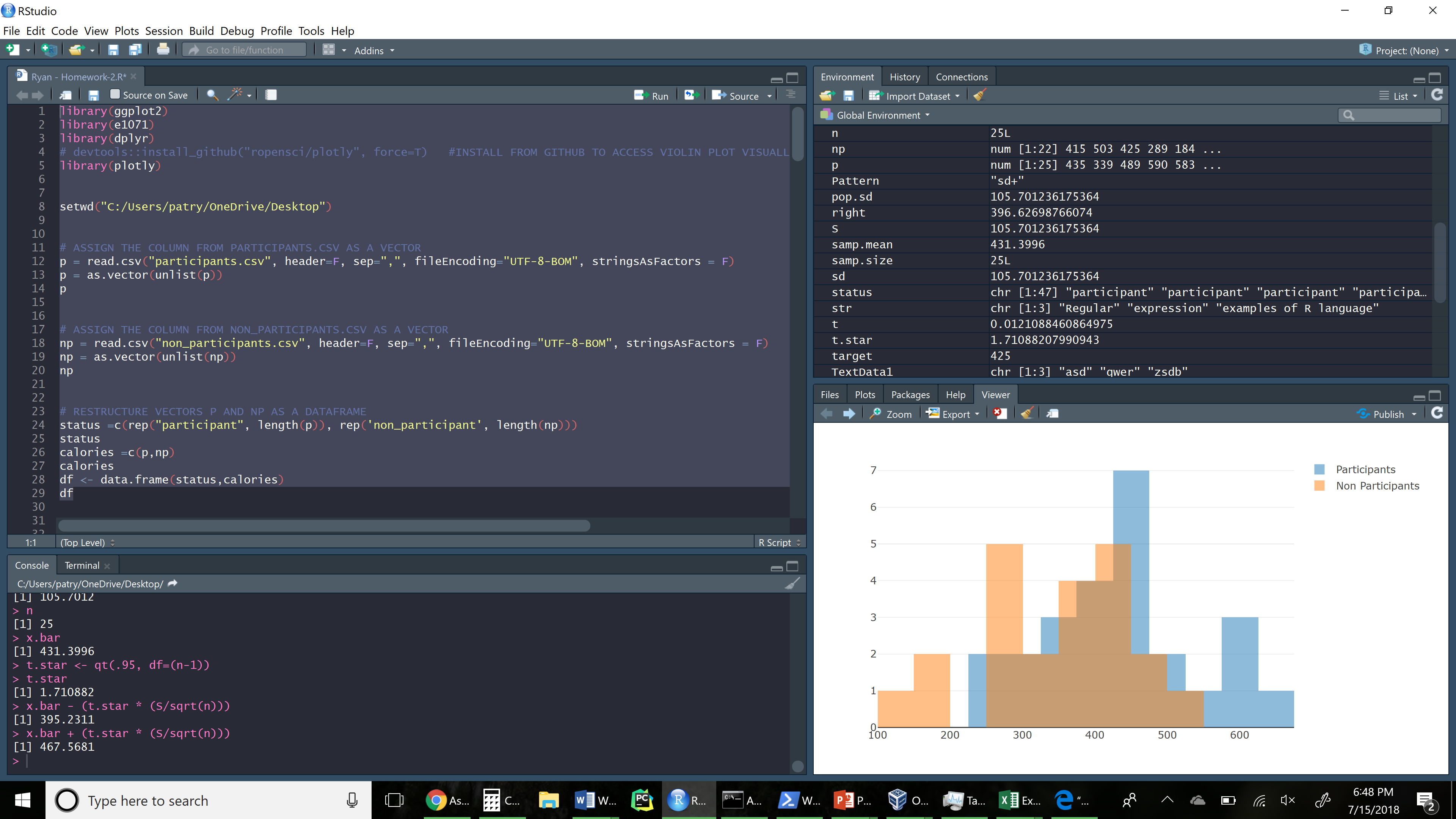
**(1) Summarize the data by whether children participated in the meal preparation or not. Use an appropriately labelled table to show the results. Also include a graphical presentation that shows the distribution of calories for participants vs. non-participants. Describe the shape of each distribution and comment on the similarity (or lack thereof) between the distributions in each population.**

|  |
| --- |
| library(ggplot2)  library(e1071)  library(dplyr)  # devtools::install\_github("ropensci/plotly", force=T) #INSTALL FROM GITHUB TO ACCESS VIOLIN PLOT VISUALLZATIONS  library(plotly)  setwd("C:/Users/patry/OneDrive/Desktop")  # ASSIGN THE COLUMN FROM PARTICIPANTS.CSV AS A VECTOR  p = read.csv("participants.csv", header=F, sep=",", fileEncoding="UTF-8-BOM", stringsAsFactors = F)  p = as.vector(unlist(p))  p  # ASSIGN THE COLUMN FROM NON\_PARTICIPANTS.CSV AS A VECTOR  np = read.csv("non\_participants.csv", header=F, sep=",", fileEncoding="UTF-8-BOM", stringsAsFactors = F)  np = as.vector(unlist(np))  np  # RESTRUCTURE VECTORS P AND NP AS A DATAFRAME  status =c(rep("participant", length(p)), rep('non\_participant', length(np)))  status  calories =c(p,np)  calories  df <- data.frame(status,calories)  df |

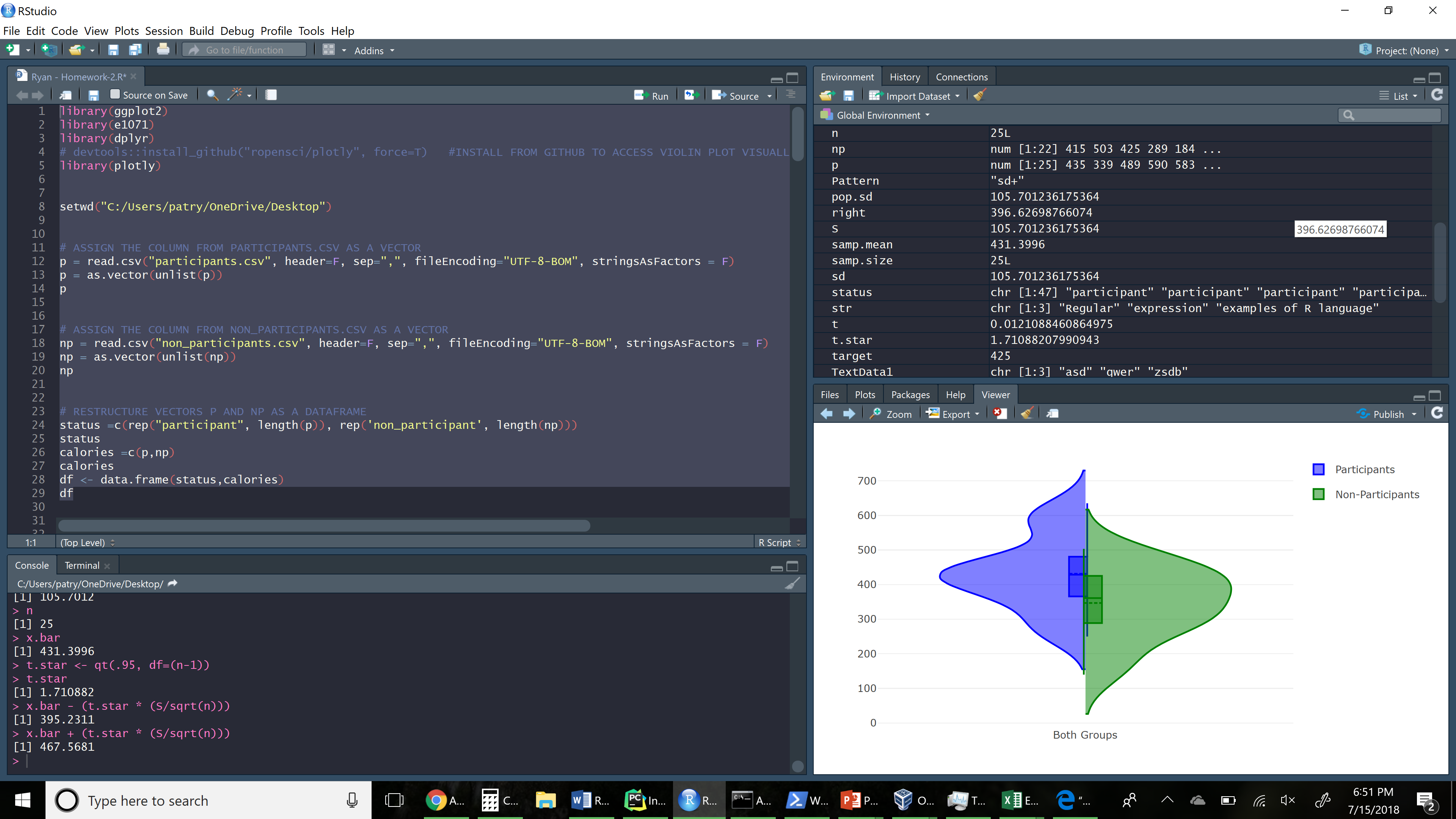
Description of the Data groups:

* Neither group features any statistical outliers (relative to their dataset)
* The participants groups features a notably higher mean (431.39) than the non participants group (346.79)
* Both groups feature similar standard deviations (105.7 for participants and 99.5 for non-participants)
* Both groups feature a similar spread (max value - min value): (385.35 for participants and 363.77 for non-participants)
* Each group appears to feature a near-normal distribution, which can be further verified using the built-in shapiro.test function(). (participants = W = 0.95473, p-value = 0.3195), (non-participants = W = 0.95806, p-value = 0.451)
* Images of each group’s distribution can be seen in [Figure 1] and [Figure 2].

[Figure 1: Histogram of Participant and Non-Participant Groups]



[Figure 2: Violin Plot of Participant and Non-Participant Groups]



**(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.**

2.1) Set up the hypotheses and select the alpha level:

H0: u=425 (The true population mean of participants is equal to 425)

H1: u!=425 (The true population mean of participants is not equal to 425)

Alpha = 0.05

2.2) Select the appropriate test-statistics

The appropriate test statistic is the t statistic (since the sample size is <30)

t = (x.bar - mu) / (S / sqrt(n))

2.3) State the decision rule:

Determine the appropriate critical value from the standard normal distribution table associated with a right-hand tail probability of alpha=0.05

A standard tstar table associate with a right hand tail probability of alpha/2 = 0.05/2 = 0.025

Using qt(0.975, df=24), the critical value is 2.063899

Decision Rule: Reject H0 if |t| is >= 2.063899; Otherwise do not reject H0

2.4) Compute the test statistics and the associated p-value:

x <- qt(.975, df=24)

x # 2.109816

x.bar <- 431.3996

mu <- 425

S <- 105.7012

n <- 25

t = ((x.bar - mu) / (S / sqrt(n)))

t # 0.3027213

(pt(t, df = 24, lower.tail = F) \* 2) # 0.7647106

**2.5) Conclusion:**

**We fail to reject H0 because |t| (0.3027) < 2.063. That is, we are not able to reject the null hypothesis that mean calories eaten is equal to 425. The corresponding p.value (0.7647) is significantly larger than our predetermined alpha value, thus providing very weak evidence against the null hypothesis.**

Alternatively we could have also used the built-in t.table function to achieve the same findings:

x <- (df$calories[df$status=='participant'])

x

t <- (mean(x) - 425) / (sd(x)/ sqrt(length(x)))

t

t.test(x, mu=425, alternative="two.sided", conf.level = 0.95)

pt(t, df = 24, lower.tail = F) \* 2 # 0.7647106

|  |
| --- |
| One Sample t-test  data: x  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 |

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

CI <- x.bar +- t \* (S/sqrt(n))

conf <- 0.90

x.bar <- mean(df$calories[df$status=='participant'])

S <- sd(df$calories[df$status=='participant'])

n <- length(df$calories[df$status=='participant'])

t.star <- qt(.95, df=(n-1))

t.star

x.bar - (t.star \* (S/sqrt(n)))

x.bar + (t.star \* (S/sqrt(n)))

**CI = (395.2311 , 467.5681)**

**Interpretation: If we took repeated samples, we would expect that this confidence interval that contains the true population mean 90% of the time.**

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

# 4.1) Set up the hypotheses and select the alpha level:

# H0: u1 = u2

# H1: u1 > u2

# alpha: 0.05

#4.2) Select the appropriate test statistic:

#Use the two sample t-statistics:

# t <- (x.bar1 - x.bar2) / (sqrt((var1/n1)+ (var2/n2)))

# 4.3) State the decision rule:

* Determine the appropriate critical value from the standard t-distribution table associated with a right-hand tail probability of alpha/2.
* Use the minimum of the two sample sizes as an input for degrees of freedom
* Critical value = 2.079614

crit <- qt(.975, df=min(n1,n2)-1)

crit

* Decision rule: Reject H0 if |t| >= 2.0796
* Otherwise, do not reject H0

4.4) Compute the test statistic and associated p-value:

x.bar1 <- mean(df$calories[df$status=='participant'])

n1 <- length(df$calories[df$status=='participant'])

var1 <- var(df$calories[df$status=='participant'])

x.bar2 <- mean(df$calories[df$status=='non\_participant'])

n2 <- length(df$calories[df$status=='non\_participant'])

var2 <- var(df$calories[df$status=='non\_participant'])

max(df$calories[df$status=='participant']) - min(df$calories[df$status=='participant'])

max(df$calories[df$status=='non\_participant']) - min(df$calories[df$status=='non\_participant'])

t <- (x.bar1 - x.bar2) / (sqrt((var1/n1)+ (var2/n2)))

t #2.824836

pt(t,df=44.779, lower.tail=F ) # 0.003519736

**4.5) Conclusion:**

**We reject H0, since 2.824 is greater than 2.0796, demonstrating that at alpha = 005, we have strong evidence to support that participants consumed more calories than non-participants. Furthermore the corresponding p.value (0.00325) provides strong evidence against**

Alternatively, we could have loaded each dataset-containing vector into R’s built in t.test function:

|  |
| --- |
| x <- (df$calories[df$status=='participant'])  y <- (df$calories[df$status=='non\_participant'])  t.test(x, y, alternative='greater', conf.level = 0.90)  >>>  Welch Two Sample t-test  data: x and y  t = 2.8248, df = 44.779, p-value = 0.00352  alternative hypothesis: true difference in means is greater than 0  90 percent confidence interval:  45.64473 Inf  sample estimates:  mean of x mean of y  431.3996 346.7991 |

**(5) Are the assumptions of the test used in (4) met? How do you know?**

The conditions assumed for a two-sample t-test are as follows :

1. Samples must be independent and randomly selected from the two distinct populations of interest
2. The variable of interest must be measured in the same way in each of the populations.
3. The parameter of interest should be normally distributed.
4. No existing heavy outliers

Although the assignment does not say explicitly, we assume that the samples have been independently and randomly selected.

Again, although it does not say explicitly, the variable of interest is assumed to have been measured in the same way, which would make sense considering the spread and shape of each dataset.

The degree to which each sample set is normal can be quantitively assessed using R’s built in shapiro.test function:

|  |
| --- |
| shapiro.test(df$calories[df$status=='participant'])  shapiro.test(df$calories[df$status=='non\_participant'])  > shapiro.test(df$calories[df$status=='participant'])  Shapiro-Wilk normality test  data: df$calories[df$status == "participant"]  W = 0.95473, p-value = 0.3195  > shapiro.test(df$calories[df$status=='non\_participant'])  Shapiro-Wilk normality test  data: df$calories[df$status == "non\_participant"]  W = 0.95806, p-value = 0.451 |

Alternatively, each dataset’s inherent normality can be visually be observed in [Figure 1] and [Figure2].

Each dataset also did not feature any significant statistical outlers, as demonstrated in the code below.

|  |
| --- |
| boxplot(df$calories[df$status=='participant'])$out  boxplot(df$calories[df$status=='non\_participant'])$out  > boxplot(df$calories[df$status=='participant'])$out  numeric(0)  > boxplot(df$calories[df$status=='non\_participant'])$out  numeric(0) |

[Figure 3: R Code to Generate the Solutions for Problems 1 - 5]

|  |
| --- |
| library(ggplot2)  library(e1071)  library(dplyr)  # devtools::install\_github("ropensci/plotly", force=T) #INSTALL FROM GITHUB TO ACCESS VIOLIN PLOT VISUALLZATIONS  library(plotly)  setwd("C:/Users/patry/OneDrive/Desktop")  # ASSIGN THE COLUMN FROM PARTICIPANTS.CSV AS A VECTOR  p = read.csv("participants.csv", header=F, sep=",", fileEncoding="UTF-8-BOM", stringsAsFactors = F)  p = as.vector(unlist(p))  p  # ASSIGN THE COLUMN FROM NON\_PARTICIPANTS.CSV AS A VECTOR  np = read.csv("non\_participants.csv", header=F, sep=",", fileEncoding="UTF-8-BOM", stringsAsFactors = F)  np = as.vector(unlist(np))  np  # RESTRUCTURE VECTORS P AND NP AS A DATAFRAME  status =c(rep("participant", length(p)), rep('non\_participant', length(np)))  status  calories =c(p,np)  calories  df <- data.frame(status,calories)  df  boxplot(df$calories[df$status=='non\_participant'])$out  ####################################################################################  #1)  # VISUALIZE PARTICIPANTS AND NON-PARTICIPANTS AS AN OVERLAPPING HISTOGRAM  plot1 <- plot\_ly(alpha = 0.5,nbinsx=15) %>%  add\_histogram(x = df$calories[df$status=="participant"], name = "Participants") %>%  add\_histogram(x = df$calories[df$status=="non\_participant"], name = "Non Participants") %>%  layout(barmode = "overlay")  plot1  plot2 <- df %>%  plot\_ly(type = 'violin') %>%  add\_trace(  x = "Both Groups",  y = p,  legendgroup = 'Yes',  scalegroup = 'Yes',  name = 'Participants',  side = 'negative',  box = list(  visible = T  ),  meanline = list(  visible = T  ),  line = list(  color = 'blue'  )  ) %>%  add\_trace(  x = "Both Groups",  y = np,  legendgroup = 'No',  scalegroup = 'No',  name = 'Non-Participants',  side = 'positive',  box = list(  visible = T  ),  meanline = list(  visible = T  ),  line = list(  color = 'green'  )  ) %>%  layout(  xaxis = list(  title = ""  ),  yaxis = list(  title = "",  zeroline = F  ),  violingap = 0,  violingroupgap = 80,  violinmode = 'overlay'  )  plot2  ####################################################################################  #2)  # Does the mean calorie consumption for those who participated in the meal preparation differ from 425?  # Formally test at the alpha=0.05 level using the 5 steps outlined in the module.  #MEAN CALORIE CONSUMPTION FOR THOSE WHO PARTICIPATED IN THE STUDY  mean(df$calories[df$status=='participant'])  length(df$calories[df$status=='participant'])  sd(df$calories[df$status=='participant'])  pnorm(2)  qnorm(.975)  #2.1)  # SET UP THE HYPOPTHESES AND SELECT THE ALPHAT LEVEL:  # H0: u=425 (TtHE MEAN OF PROGRAM PARTICIPANTS IS LESS THAN 425)  # H1: u!=425 (THE MEAN OF PROGRAM PARTICIPANTS IS GREATER THAN 425)  # Alpha = 0.05  #2.2)  # z <- (samp.mean - pop.mean) / (pop.sd / sqrt(samp.size))  # The appropriate test statistic is the t statistic (since the sample size is <30)  # t = (x.bar - mu) / (S / sqrt(n))  #2.3)  # Determine the appropriate critical value from the standard normal distribution table associated with a right hand tail probability of alpha=0.05  # A standard tstar table associate with a right hand tail probability of alpha/2 = 0.05/2 = 0.025  # Using qt(0.975, df=24), the critical value is 2.063899  # Decision Rule: Reject H0 if |t| is >= 2.063899; Otherwise do not reject H0  # 2.4)  x.bar <- 431.3996  mu <- 425  S <- 105.7012  n <- 25  t = ((x.bar - mu) / (S / sqrt(n)))  t # 0.3027213  x <- qt(.975, df=24)  x # 2.109816  # 2.5)  # We do not reject H0 because |t| (0.3027) < 2.063  #That is, we are not able to reject the nulll hypothesis that mean calories eaten is equal to 425  # or using the built in t.test function  x <- (df$calories[df$status=='participant'])  x  t <- (mean(x) - 425) / (sd(x)/ sqrt(length(x)))  t  t.test(x, mu=425, alternative="two.sided", conf.level = 0.95)  pt(t, df = 24, lower.tail = F) \* 2  ####################################################################################  #3)  # Calculate a 90% confidence interval for the mean calorie intake for participants in the meal preparation.  # Interpret the confidence interval.  # CI <- x.bar +- t\* \* sd  conf <- 0.90  x.bar <- mean(df$calories[df$status=='participant'])  S <- sd(df$calories[df$status=='participant'])  n <- length(df$calories[df$status=='participant'])  S  n  x.bar  t.star <- qt(.95, df=(n-1))  t.star  x.bar - (t.star \* (S/sqrt(n)))  x.bar + (t.star \* (S/sqrt(n)))  pnorm(-2.72)  # Interpretation: If 100 additional samples were collected in the same distribution, 90% would fall within this confidence margin.  qnorm(.95)  pnorm(mean=.9, q=.4)  ####################################################################################  # 4.1)  # H0: u1 = u2  # H1: u1 > u2  # alpha: 0.05  #4.2)  #Use the two sample t-statistics:  # t <- (x.bar1 - x.bar2) / (sqrt((var1/n1)+ (var2/n2)))  # 4.3)  # Determine the appropriate critical value from the standard t-distibution table associated with a right hand tail probability of A/2  # Use the miniumum of the two sample sizes as an input for degrees of freedom  # Critical value = 2.079614  crit <- qt(.975, df=min(n1,n2)-1)  crit  # Decision rule: Reject H0 if |t| >= 2.0796  # Otherwise, do not reject H0  #4.4) Compute the test statistic and assciated p-value  x.bar1 <- mean(df$calories[df$status=='participant'])  n1 <- length(df$calories[df$status=='participant'])  var1 <- var(df$calories[df$status=='participant'])  x.bar2 <- mean(df$calories[df$status=='non\_participant'])  n2 <- length(df$calories[df$status=='non\_participant'])  var2 <- var(df$calories[df$status=='non\_participant'])  x.bar1  n1  var1  x.bar2  n1  var2  t <- (x.bar1 - x.bar2) / (sqrt((var1/n1)+ (var2/n2)))  t #2.824836  pt(t,df=44.779, lower.tail=F )  # or we can load the vectors into the built in t.test function  x <- (df$calories[df$status=='participant'])  y <- (df$calories[df$status=='non\_participant'])  t.test(x, y, alternative='greater', conf.level = 0.90)  # 4.5)  # We reject H0, since 2.824 is greater than 2.0796, demonstrating that at alpha = 005, we have  # significant evidence to support that participants consumed more calories than non-participants.    #5) Are the assumptions of the test used in (4) met?  # Yes they have been met. Aside from visual comparisions (histogram/violin plot), the difference in the sample means  # supports the idea that participants in the plan consumped more calories than non-participants.  # furthermore the associated p.value of 0.00507 demonstrates that low likleyhood of this outcome occouring by chance.  pt(t, df=44.779, lower.tail = F) |