# Survey Link: https://goo.gl/forms/ElUp30eEAgCIj6fh1

# Load packages that will be used for the analysis

library(ggplot2)

library(plyr)

library(dplyr)

library(reshape2)

library(plotly)

# Library to install statsr library

library(devtools)

# Install Statsr from github repository

install\_github("StatsWithR/statsr")

library(statsr)

# Change the working directory

setwd('C:\\Users\\Abhay\\Desktop\\Business Statistics Project Files\\')

# Load data set

load("193\_project\_dataset.RData")

#covert the data to tbl\_df as they are easier to analyze

shopping\_tbl <- tbl\_df(shopping)

head(shopping\_tbl)

## # A tibble: 6 x 61

## Timestamp Age Gender Marital\_Status Education

## <fctr> <fctr> <fctr> <fctr> <fctr>

## 1 9/16/2016 14:20:30 26 to 30 Years Male Single Post Graduate

## 2 9/16/2016 14:21:39 26 to 30 Years Male Single Graduate

## 3 9/16/2016 14:32:17 20 to 25 Years Male Single Graduate

## 4 9/16/2016 15:28:16 20 to 25 Years Male Single Graduate

## 5 9/17/2016 13:49:09 20 to 25 Years Male Single Graduate

## 6 9/22/2016 7:37:54 26 to 30 Years Male Single Graduate

## # ... with 56 more variables: Internet\_Access\_Device <fctr>,

## # Internet\_Browsing\_Time\_Zone <fctr>, Browsing\_Time <fctr>,

## # Working\_Time\_Zone <fctr>, Working\_Hours <fctr>, Felt\_Unhappy <fctr>,

## # Felt\_Unsatisfied <fctr>, Felt\_Cheated <fctr>, Felt\_Uneasy <fctr>,

## # Felt\_Furious <fctr>, Felt\_Made\_A\_Mistake <fctr>,

## # Felt\_Irritated <fctr>, Felt\_Disappointed <fctr>,

## # Felt\_Frustrated <fctr>, Felt\_Concerned <fctr>,

## # Felt\_Done\_A\_Wastage <fctr>, Felt\_Anxious <fctr>, Felt\_Angry <fctr>,

## # Felt\_Negative <fctr>, Thought\_Of\_Corrective\_Action <fctr>,

## # Though\_Of\_Other\_Choice <fctr>, Anxiety\_Than\_Satisfaction <fctr>,

## # No\_Trust <fctr>, Not\_Sure\_About\_Quality <fctr>,

## # Delivery\_Of\_Same\_Product <fctr>, Good\_Deal <fctr>,

## # Not\_Good\_Bargain <fctr>, Cash\_Payment\_Than\_Online\_Payment <fctr>,

## # Answer\_To\_Quesries <fctr>, Info\_About\_Product <fctr>,

## # Negotiation\_For\_Better\_Deal <fctr>, Customer\_Reviews <fctr>,

## # Authenticity\_Of\_Online\_Review <fctr>, Made\_A\_Wrong\_Decision <fctr>,

## # Expectation\_Not\_Met <fctr>, Buying\_From\_Same\_Retailer <fctr>,

## # Fit\_Of\_Products <fctr>, Not\_Sure\_Of\_Performance <fctr>,

## # Genuine\_Discounts\_Or\_Not <fctr>, Increase\_In\_Spam\_Mails <fctr>,

## # Selling\_Personal\_Data <fctr>, Selling\_Financial\_Data <fctr>,

## # Accurate\_Info <fctr>, Keeping\_Promise <fctr>,

## # Missing\_Personal\_Touch <fctr>, Need\_Of\_Sales\_Person <fctr>,

## # Smile\_Of\_Sales\_Person <fctr>, Return\_And\_Exchange\_Policy <fctr>,

## # Painful\_Product\_Replacemnt <fctr>, Same\_Replaced\_Product <fctr>,

## # Product\_Information <fctr>, Authenticity\_Of\_Retailers <fctr>,

## # Decoding\_Shopping\_Pattern <fctr>, Why\_Purchase\_Online <fctr>,

## # Trail\_Of\_Product <fctr>, Satisfactory\_Experience <fctr>

# Check the summary of dataset for different variables

Summary(shopping\_tbl$age)

## 20 to 25 Years 26 to 30 Years

## 0 71 74

## Less than 20 Years More than 30 Years

## 6 11

age\_summary <- tbl\_df(summary(shopping\_tbl$Age))

# Plot a pie chart to represent the various age groups

labels = c('', '20 to 25 Years', '26 to 30 Years', 'Less than 20 Years', 'More than 30 Years')

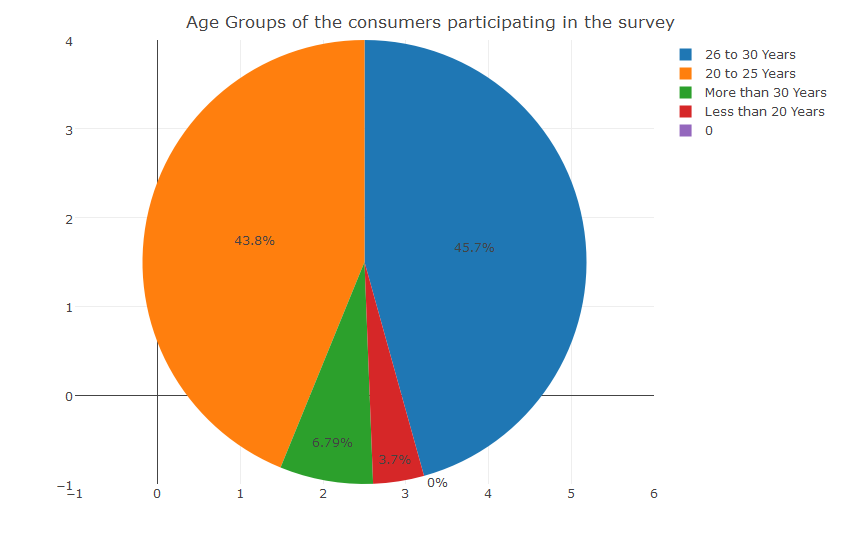
values = age\_summary$value

ds <- data.frame(labels = labels,

values = values)

plot\_ly(ds, labels = labels, values = values, type = "pie") %>%

layout(title = "Age Groups of the consumers participating in the survey")



summary(shopping\_tbl$Felt\_Furious)

# Agree Disagree Neutral Strongly Agree Strongly Disagree

# 0 32 47 64 8 11

# Check for the missing values in the dataset

sum(is.na(shopping\_tbl))

# [1] 0

# No missing values in the data set, so we can safely proceed

# This will give us the structure of the categorical variables

str(shopping\_tbl$Felt\_Unhappy)

# Factor w/ 6 levels "","Agree","Disagree",..: 3 4 2 3 4 3 4 2 3 2 ...

# This gives 6 levels, whereas we have only 5 levels, so we need to remove the extra "" level

for(i in names(shopping\_tbl[,3:60])){

shopping\_tbl[[i]] <- droplevels(shopping\_tbl[[i]])

shopping\_tbl[[i]] <- factor(shopping\_tbl[[i]], ordered=TRUE)

}

str(shopping\_tbl$Felt\_Unhappy)

# Assign the correct value to each of the ordinal level

new\_levels <- c("1"="Strongly Disagree", "2"="Disagree", "3"="Neutral", "4"="Agree", "5"="StronglyAgree”)

for(i in names(shopping\_tbl[,11:60])){

shopping\_tbl[[i]] <- revalue(shopping\_tbl[[i]], new\_levels)

shopping\_tbl[[i]] <- ordered(shopping\_tbl[[i]], levels=c("Strongly Disagree", "Disagree","Neutral", "Agree","Strongly Agree"))

}

# Check for the structure after removing the "" level

str(shopping\_tbl$Felt\_Unhappy)

## Ord.factor w/ 5 levels "Strongly Disagree"<..: 2 3 4 2 3 2 3 4 2 4 ...

# Assign the following codes to the ordinal values:

# "1"="Strongly Disagree", "2"="Disagree", "3"="Neutral", "4"="Agree", "5"="Strongly Agree"

for(i in names(shopping\_tbl[,11:60])) {

shopping\_tbl[[i]] = as.numeric(shopping\_tbl[[i]])

}

# Calculate the Online dissonance score

shopping\_tbl <- shopping\_tbl %>% mutate(Online\_Dissonance\_Score = Felt\_Unhappy + Felt\_Made\_A\_Mistake +

Felt\_Done\_A\_Wastage + Thought\_Of\_Corrective\_Action + Expectation\_Not\_Met +

Genuine\_Discounts\_Or\_Not + Return\_And\_Exchange\_Policy +

Painful\_Product\_Replacemnt + Keeping\_Promise + No\_Trust)

# Calculate the Human interface score

shopping\_tbl <- shopping\_tbl %>% mutate(Human\_Interface\_Score = Cash\_Payment\_Than\_Online\_Payment +

Negotiation\_For\_Better\_Deal + Missing\_Personal\_Touch +

Need\_Of\_Sales\_Person + Smile\_Of\_Sales\_Person + Trail\_Of\_Product )

#Calculate the privacy score

shopping\_tbl <- shopping\_tbl %>% mutate(Privacy\_Score = Increase\_In\_Spam\_Mails + Selling\_Personal\_Data + Selling\_Financial\_Data +

Decoding\_Shopping\_Pattern)

# After adding the three scores, check the table with the new values

head(shopping\_tbl[,62:64])

## # A tibble: 6 x 3

## Online\_Dissonance\_Score Human\_Interface\_Score Privacy\_Score

## <dbl> <dbl> <dbl>

## 1 31 22 14

## 2 30 20 12

## 3 40 24 16

## 4 19 13 9

## 5 29 19 12

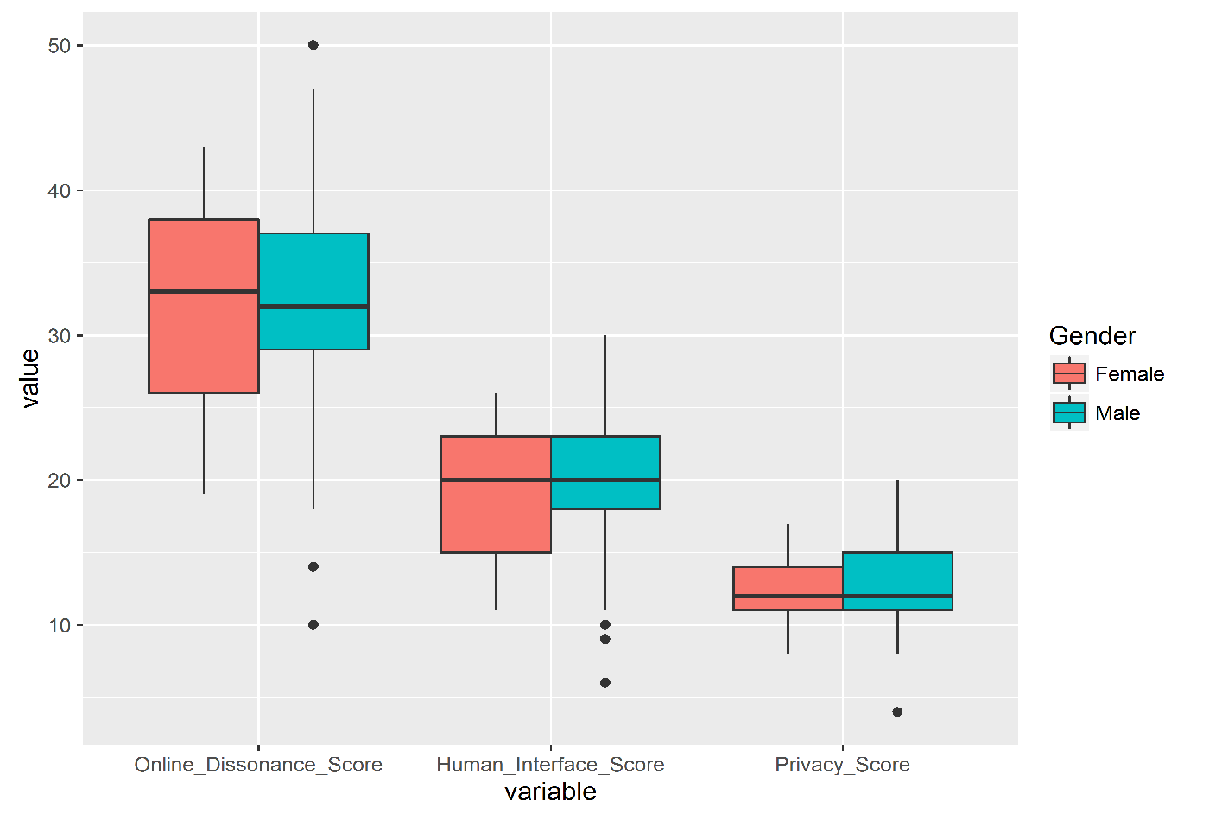
## 6 40 23 15

# Check the variation in diffent scores by gender

shopping\_tbl\_gender <- melt(shopping\_tbl, id.vars = "Gender", measure.vars = c("Online\_Dissonance\_Score","Human\_Interface\_Score","Privacy\_Score"))

# Plot the variation using ggplot

ggplot(data = shopping\_tbl\_gender, aes(x=variable, y=value)) + geom\_boxplot(aes(fill = Gender))



# Take the sample of 80 random observations, setting seed as the last three digits of the person in the group

set.seed(193)

n <- 80

samp <- sample\_n(shopping\_tbl, n)

## Hypothesis Testing Based on GENDER ##

# Testing hypotheis for Online Dissonance Score based on Gender

inference(y = Online\_Dissonance\_Score, x = Gender, data = samp, statistic = "mean", type = "ht", null = 0, alternative = "twosided", method = "theoretical")

## Response variable: numerical

## Explanatory variable: categorical (2 levels)

## n\_Female = 19, y\_bar\_Female = 31.7895, s\_Female = 7.5098

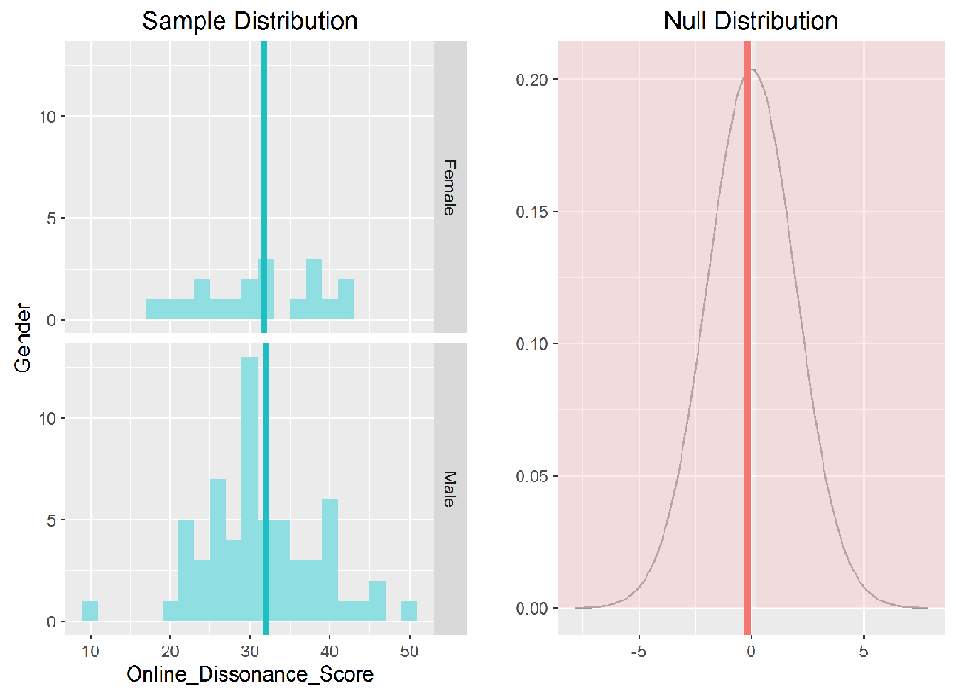
## n\_Male = 61, y\_bar\_Male = 31.9836, s\_Male = 7.2284

## H0: mu\_Female = mu\_Male

## HA: mu\_Female != mu\_Male

## t = -0.0993, df = 18

## p\_value = 0.922



# Testing hypotheis for Human Interface Score based on Gender

inference(y = Human\_Interface\_Score, x = Gender, data = samp, statistic = "mean", type = "ht", null = 0, alternative = "twosided", method = "theoretical")

## Response variable: numerical

## Explanatory variable: categorical (2 levels)

## n\_Female = 19, y\_bar\_Female = 18.4737, s\_Female = 4.8117

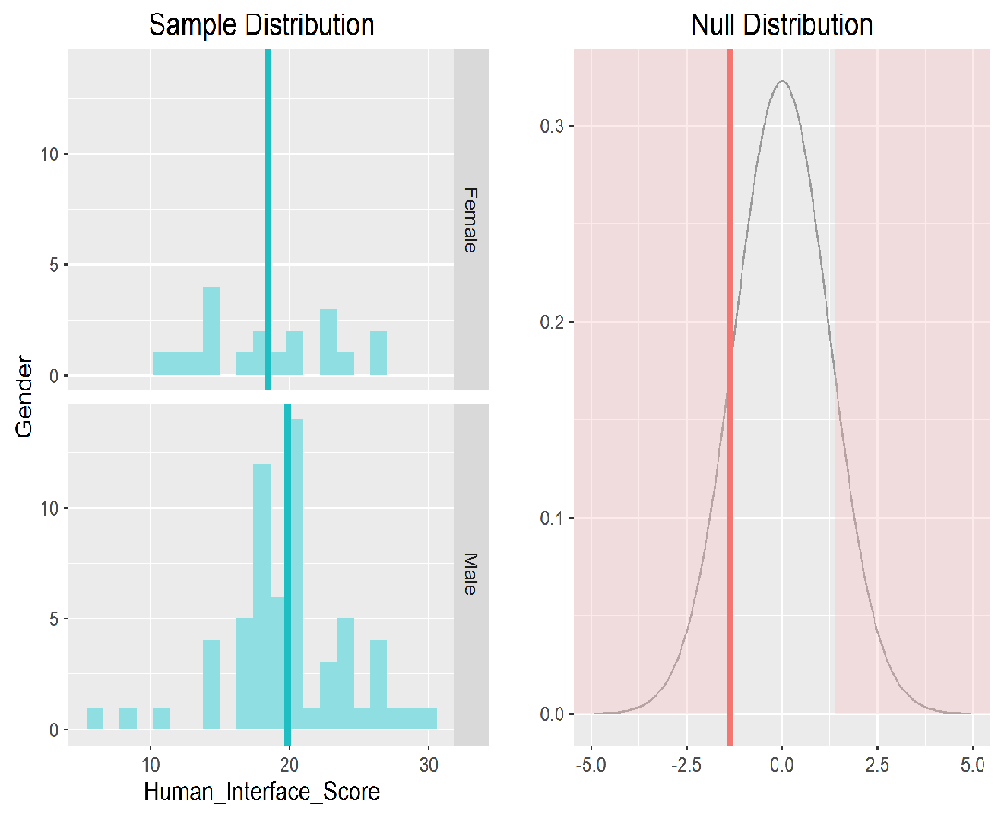
## n\_Male = 61, y\_bar\_Male = 19.8525, s\_Male = 4.3468

## H0: mu\_Female = mu\_Male

## HA: mu\_Female != mu\_Male

## t = -1.1153, df = 18

## p\_value = 0.2794



# Testing hypotheis for Privacy Score based on Gender

inference(y = Privacy\_Score, x = Gender, data = samp, statistic = "mean", type = "ht", null = 0, alternative = "twosided", method = "theoretical")

## Response variable: numerical

## Explanatory variable: categorical (2 levels)

## n\_Female = 19, y\_bar\_Female = 12.6842, s\_Female = 2.4732

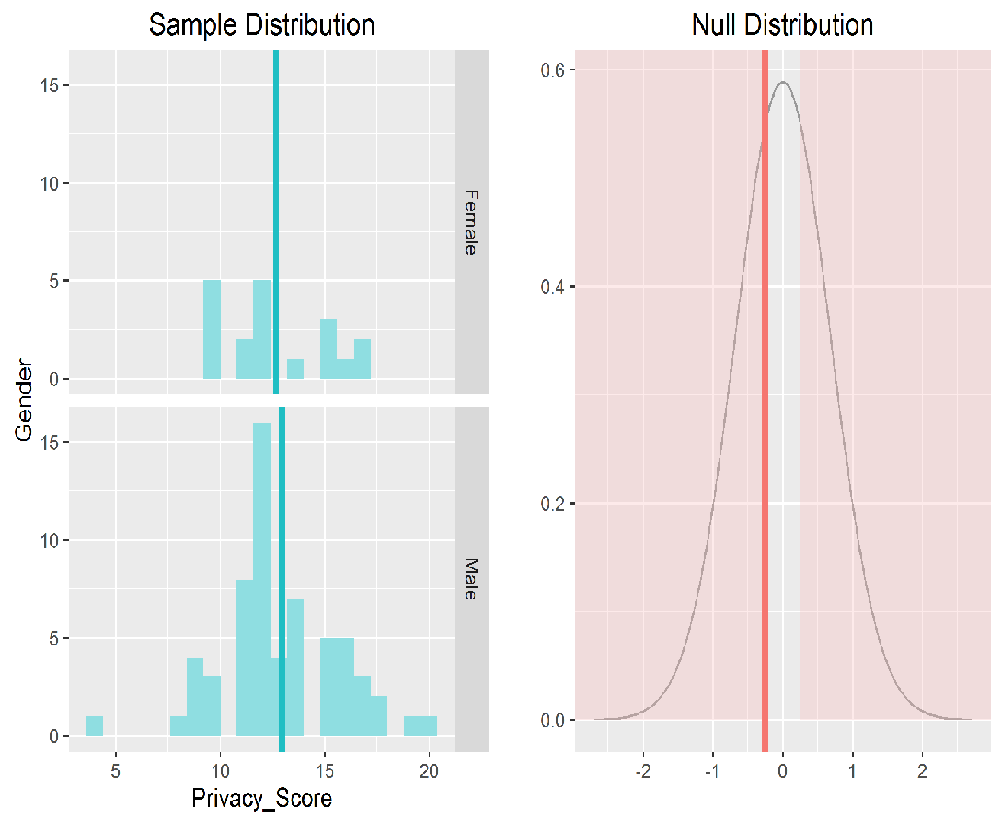
## n\_Male = 61, y\_bar\_Male = 12.9344, s\_Male = 2.886

## H0: mu\_Female = mu\_Male

## HA: mu\_Female != mu\_Male

## t = -0.3695, df = 18

## p\_value = 0.716

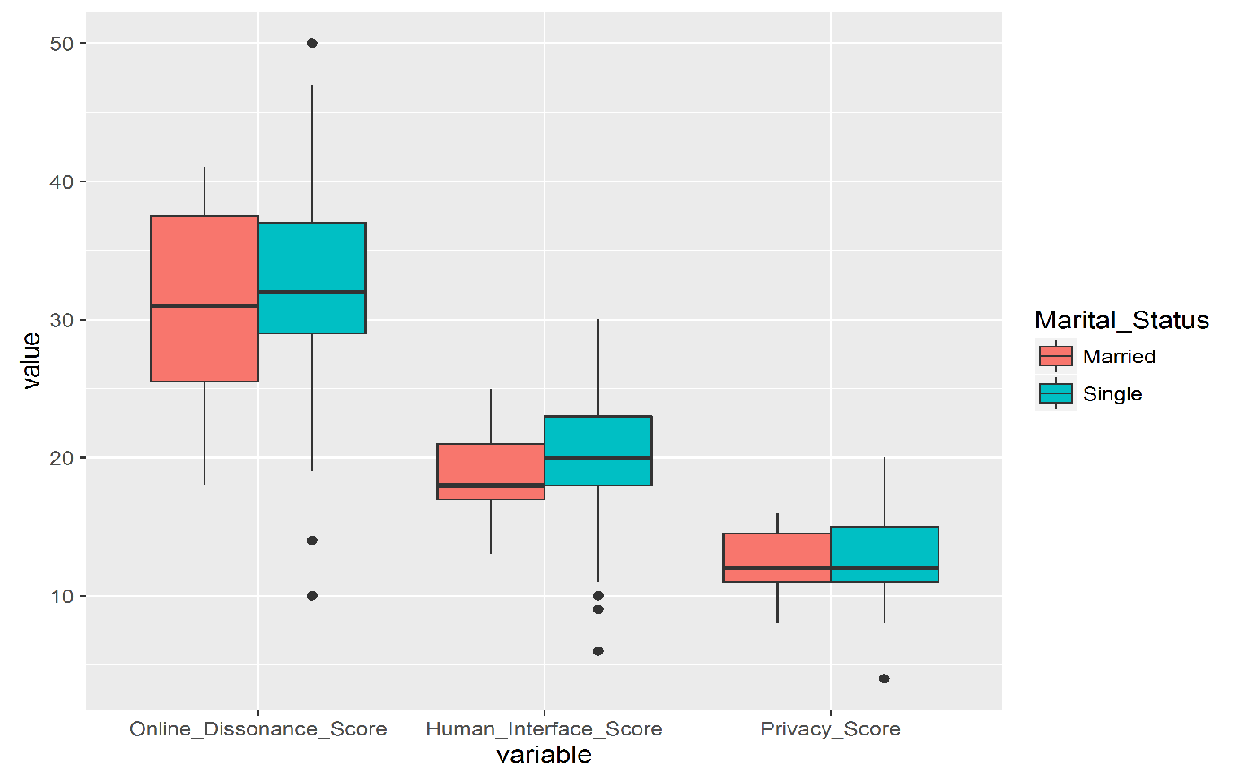


## Hypotheis Testing based on MARITAL STATUS ##

removing\_single\_again <- filter(shopping\_tbl, Marital\_Status != "Single Again")

shopping\_tbl\_marital\_status <- melt(removing\_single\_again, id.vars = "Marital\_Status", measure.vars = c("Online\_Dissonance\_Score","Human\_Interface\_Score","Privacy\_Score"))

ggplot(data = shopping\_tbl\_marital\_status, aes(x=variable, y=value)) + geom\_boxplot(aes(fill = Marital\_Status))



# Removing 'Single Again' level from the dataset

removing\_single\_again[["Marital\_Status"]] <- droplevels(removing\_single\_again[["Marital\_Status"]])

removing\_single\_again[["Marital\_Status"]] <- factor(removing\_single\_again[["Marital\_Status"]], ordered = TRUE)

# Taking a different seed here than our previous sample because that might contain data for `single again` level of Marital\_Status variable,

# which we are to exclude from our analysis.

set.seed(194)

n <- 80

samp\_marital\_status <- sample\_n(removing\_single\_again, n)

# Testing for Online Dissonance based on Marital Status

inference(y = Online\_Dissonance\_Score, x = Marital\_Status, data = samp\_marital\_status, statistic = "mean", type = "ht", null = 0, alternative = "twosided", method = "theoretical")

## Response variable: numerical

## Explanatory variable: categorical (2 levels)

## n\_Married = 14, y\_bar\_Married = 33, s\_Married = 6.4807

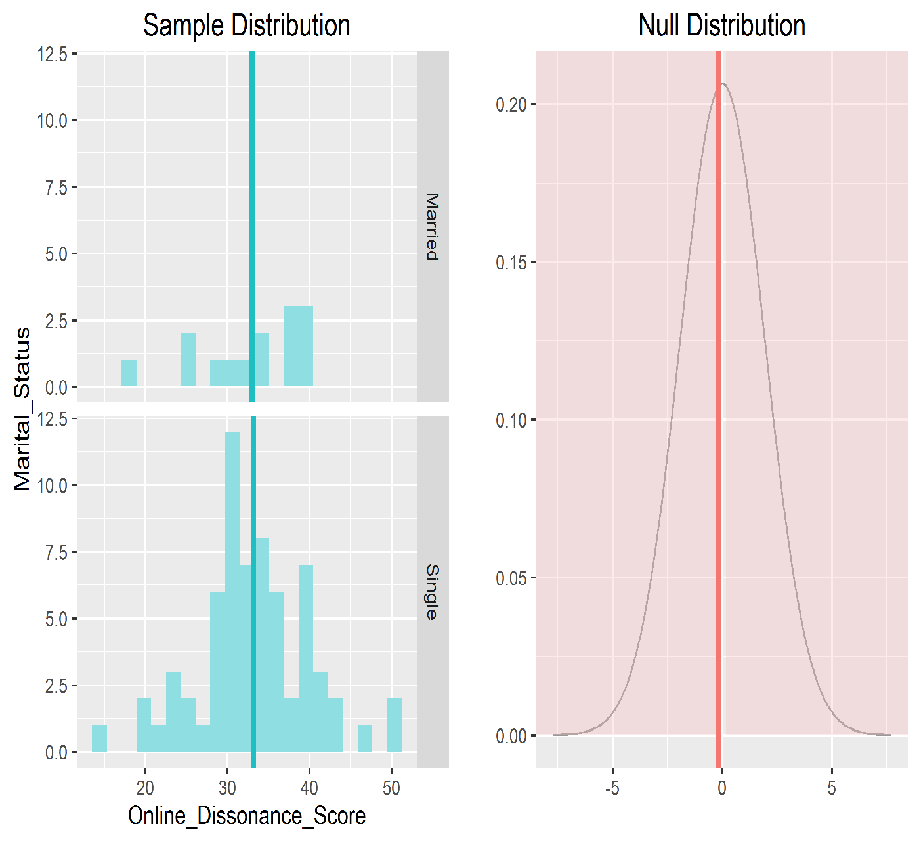
## n\_Single = 66, y\_bar\_Single = 33.1667, s\_Single = 6.9362

## H0: mu\_Married = mu\_Single

## HA: mu\_Married != mu\_Single

## t = -0.0863, df = 13

## p\_value = 0.9325



# Testing for Human Interface based on Marital Status

inference(y = Human\_Interface\_Score, x = Marital\_Status, data = samp\_marital\_status, statistic = "mean", type = "ht", null = 0, alternative = "twosided", method = "theoretical")

## Response variable: numerical

## Explanatory variable: categorical (2 levels)

## n\_Married = 14, y\_bar\_Married = 19.4286, s\_Married = 2.7932

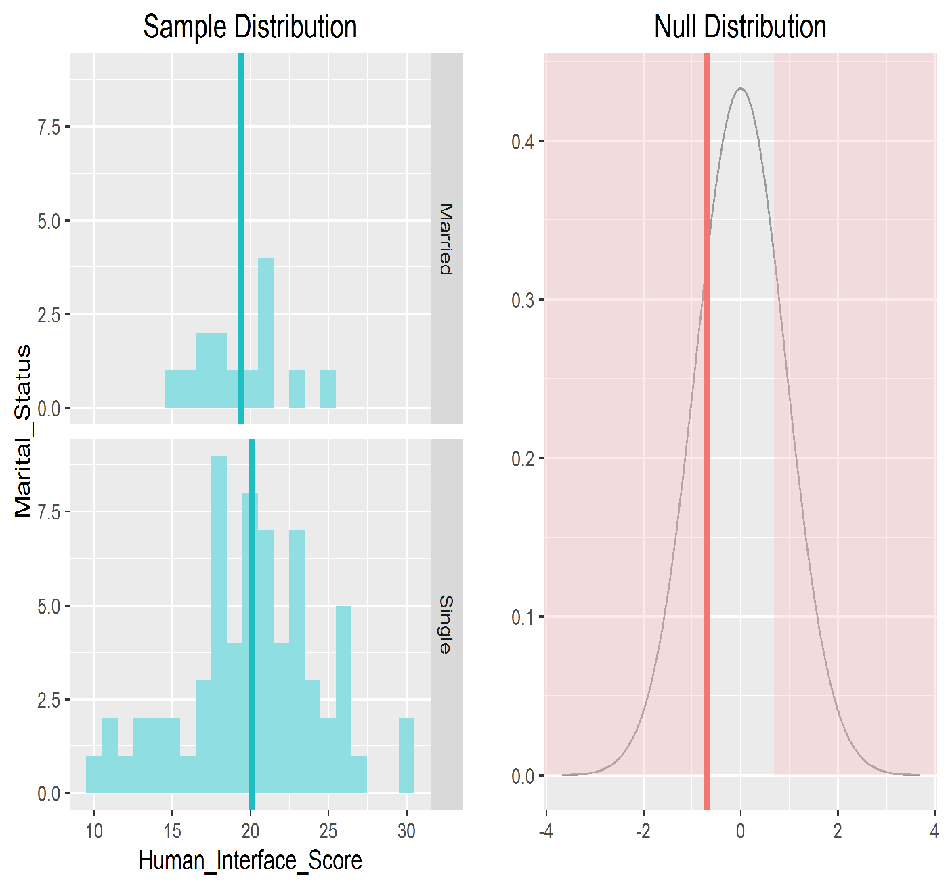
## n\_Single = 66, y\_bar\_Single = 20.1212, s\_Single = 4.3695

## H0: mu\_Married = mu\_Single

## HA: mu\_Married != mu\_Single

## t = -0.7528, df = 13

## p\_value = 0.465



# Testing for Privacy based on Marital Status

inference(y = Privacy\_Score, x = Marital\_Status, data = samp\_marital\_status, statistic = "mean", type = "ht", null = 0, alternative = "twosided", method = "theoretical")

## Response variable: numerical

## Explanatory variable: categorical (2 levels)

## n\_Married = 14, y\_bar\_Married = 12.7857, s\_Married = 2.1187

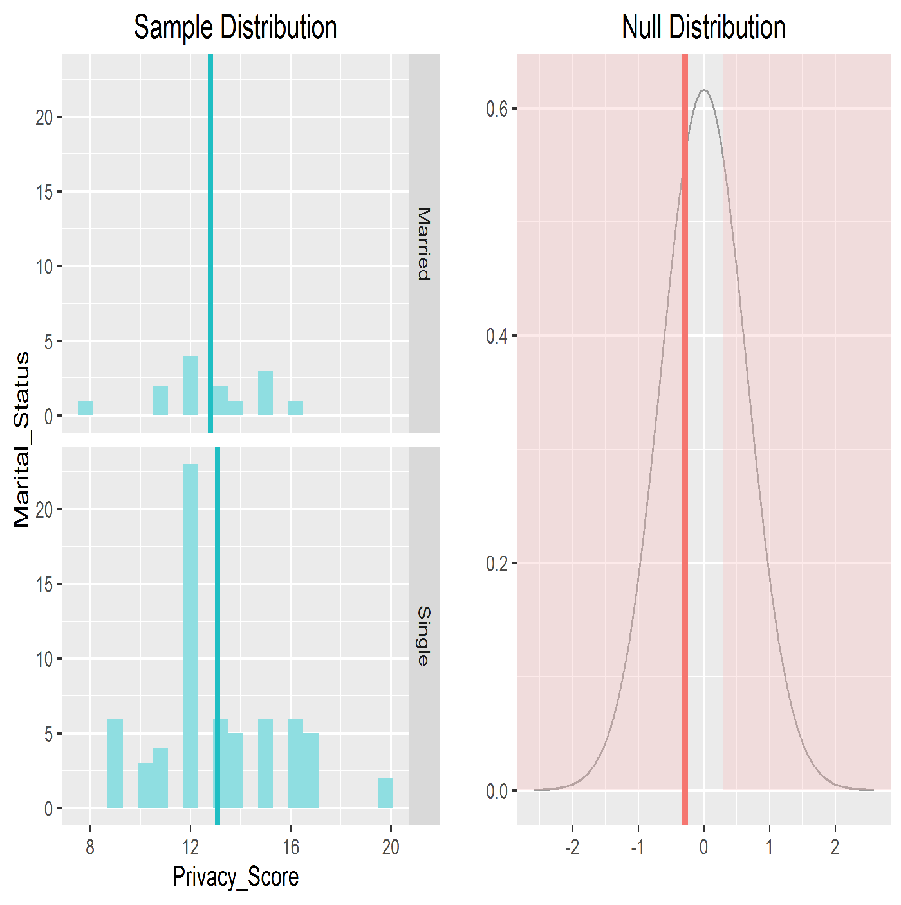
## n\_Single = 66, y\_bar\_Single = 13.0758, s\_Single = 2.5499

## H0: mu\_Married = mu\_Single

## HA: mu\_Married != mu\_Single

## t = -0.448, df = 13

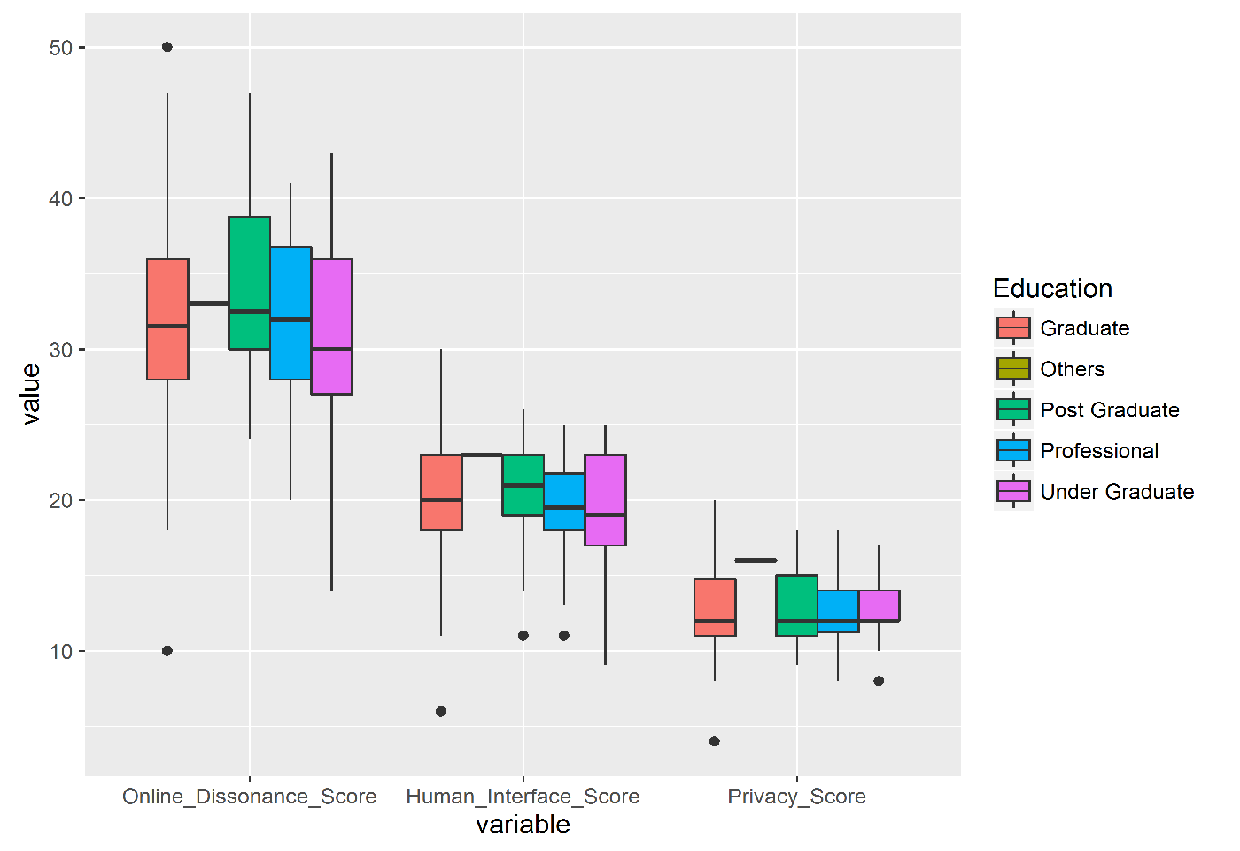
## p\_value = 0.6615



## Hypothesis testing based on EDUCATION ##

shopping\_tbl\_education <- melt(shopping\_tbl, id.vars = "Education", measure.vars = c("Online\_Dissonance\_Score","Human\_Interface\_Score","Privacy\_Score"))

ggplot(data = shopping\_tbl\_education, aes(x=variable, y=value)) + geom\_boxplot(aes(fill = Education))



# Testing for online Dissonance based on Education

inference(y = Online\_Dissonance\_Score, x = Education, data = samp, statistic = "mean", type = "ht", null = 0, alternative = "greater", method = "theoretical")

## Response variable: numerical

## Explanatory variable: categorical (5 levels)

## n\_Graduate = 25, y\_bar\_Graduate = 31.52, s\_Graduate = 8.8182

## n\_Others = NA, y\_bar\_Others = NA, s\_Others = NA

## n\_Post Graduate = 23, y\_bar\_Post Graduate = 33.6957, s\_Post Graduate = 6.4062

## n\_Professional = 19, y\_bar\_Professional = 31.9474, s\_Professional = 6.0962

## n\_Under Graduate = 13, y\_bar\_Under Graduate = 29.6154, s\_Under Graduate = 6.8743

##

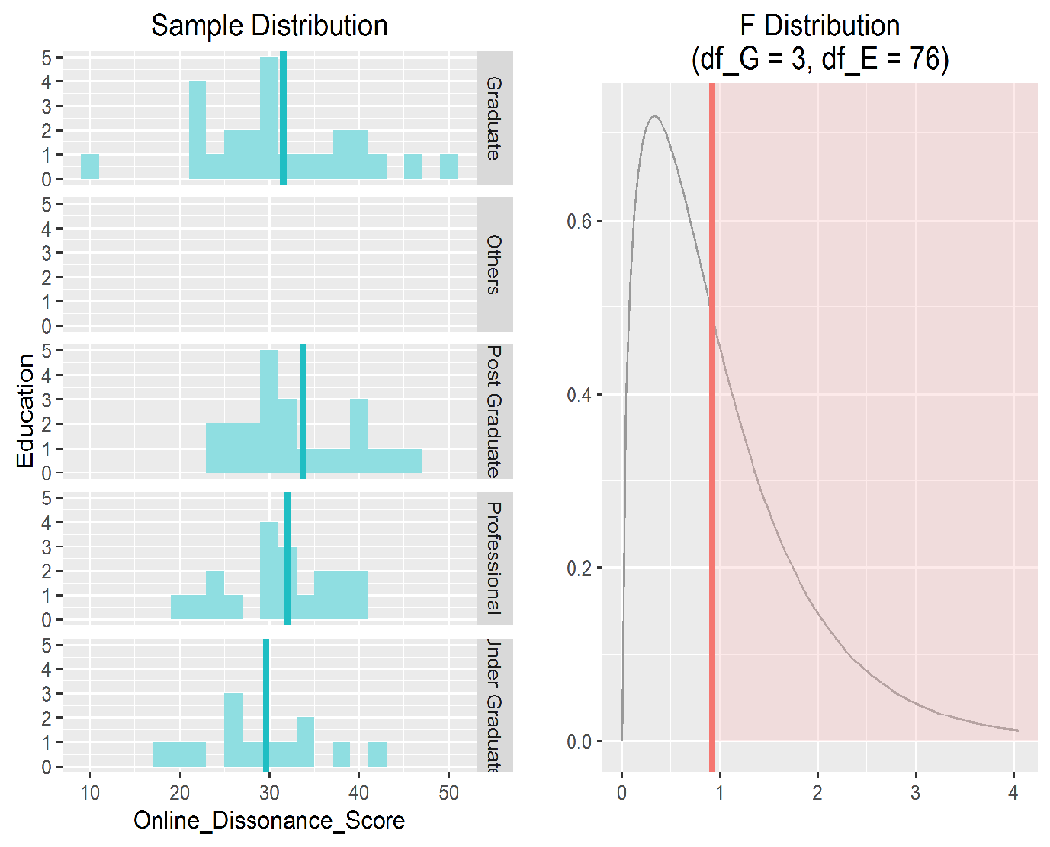
## ANOVA:

## df Sum\_Sq Mean\_Sq F p\_value

## Education 3 145.5536 48.5179 0.9207 0.435

## Residuals 76 4005.1339 52.6991

## Total 79 4150.6875



# Testing for Human Interface based on Education

inference(y = Human\_Interface\_Score, x = Education, data = samp, statistic = "mean", type = "ht", null = 0, alternative = "greater", method = "theoretical")

## Response variable: numerical

## Explanatory variable: categorical (5 levels)

## n\_Graduate = 25, y\_bar\_Graduate = 19.72, s\_Graduate = 5.6163

## n\_Others = NA, y\_bar\_Others = NA, s\_Others = NA

## n\_Post Graduate = 23, y\_bar\_Post Graduate = 20.3913, s\_Post Graduate = 3.5128

## n\_Professional = 19, y\_bar\_Professional = 19.7895, s\_Professional = 3.7354

## n\_Under Graduate = 13, y\_bar\_Under Graduate = 17.2308, s\_Under Graduate = 4.1864

##

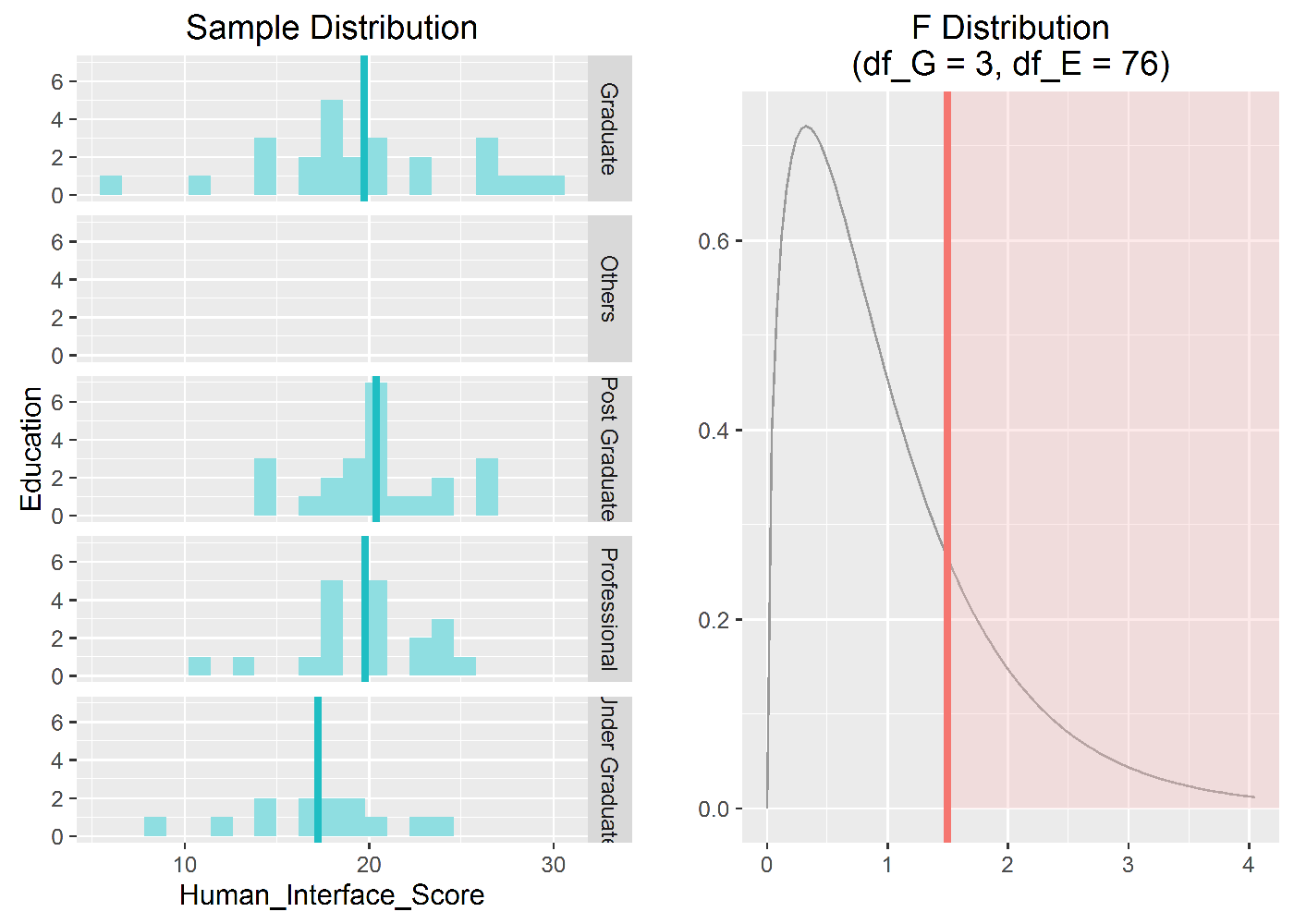
## ANOVA:

## df Sum\_Sq Mean\_Sq F p\_value

## Education 3 87.9662 29.3221 1.4956 0.2225

## Residuals 76 1489.9838 19.6051

## Total 79 1577.95



# Testing for Privacy based on Education

inference(y = Privacy\_Score, x = Education, data = samp, statistic = "mean", type = "ht", null = 0, alternative = "greater", method = "theoretical")

## Response variable: numerical

## Explanatory variable: categorical (5 levels)

## n\_Graduate = 25, y\_bar\_Graduate = 12.96, s\_Graduate = 3.4215

## n\_Others = NA, y\_bar\_Others = NA, s\_Others = NA

## n\_Post Graduate = 23, y\_bar\_Post Graduate = 12.913, s\_Post Graduate = 2.8109

## n\_Professional = 19, y\_bar\_Professional = 13.1579, s\_Professional = 2.3157

## n\_Under Graduate = 13, y\_bar\_Under Graduate = 12.2308, s\_Under Graduate = 2.0878

##

## ANOVA:

## df Sum\_Sq Mean\_Sq F p\_value

## Education 3 7.1299 2.3766 0.2992 0.8258

## Residuals 76 603.6201 7.9424

## Total 79 610.75

