Lab 11

Alexander Hernandez

09/29/2022

1) Test Scores, Major, and Class

a) merge() both into a single table

```
test_scores1 = read.csv("C:\\repos\\STAT 50001\\Homework 2\\test-scores1.tsv",
         sep=" ", header=TRUE)

test_scores2 = read.csv("C:\\repos\\STAT 50001\\Homework 2\\test-scores2.tsv",
                        sep=" ", header=TRUE)

test_scores_all = merge(test_scores1, test_scores2)
test_scores_all
```

```
##       Name Major Test1  Class Test2
## 1      Ana    MA    56 Junior    86
## 2    Brian    MA    78 Junior    67
## 3    Cathy    CS    87 Senior    78
## 4    Dough    CS    89 Junior    89
## 5     John  STAT    95 Senior    87
## 6    Lucas  STAT    98 Senior    67
## 7   Marcus  STAT    59 Junior    94
## 8    Nabin    MA    78 Senior    78
## 9  William    CS    87 Senior    81
## 10     Zoe  STAT    98 Junior    83
```

b) How many students did better in the second test?

```
nrow(test_scores_all[test_scores_all$Test2 > test_scores_all$Test1,])
```

```
## [1] 2
```

c) How many did better in the first test?

```
nrow(test_scores_all[test_scores_all$Test1 > test_scores_all$Test2,])
```

```
## [1] 6
```

d) how many have the same score in both tests?

```
nrow(test_scores_all[test_scores_all$Test2 == test_scores_all$Test1,])
```

```
## [1] 2
```

e) Calculate the average and SD of both tests

```
mean(test_scores_all$Test1)
```

```
## [1] 82.5
```

```
sd(test_scores_all$Test1)
```

```
## [1] 14.96106
```

```
mean(test_scores_all$Test2)
```

```
## [1] 81
```

```
sd(test_scores_all$Test2)
```

```
## [1] 8.869423
```

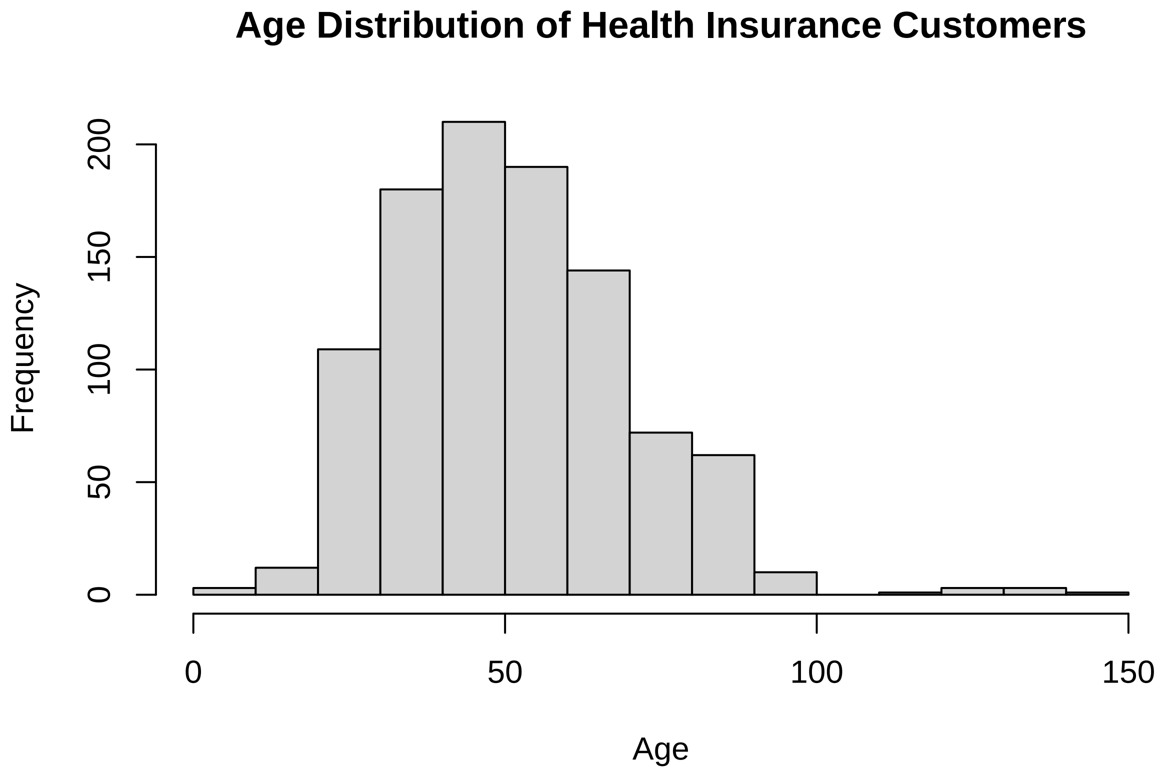
2) Health Insurance from 'custdata.tsv'

a) Import the data

```
cust_data = read.csv("C:\\repos\\STAT 50001\\Homework 2\\custdata.tsv",
                     sep='\t', header=TRUE)
```

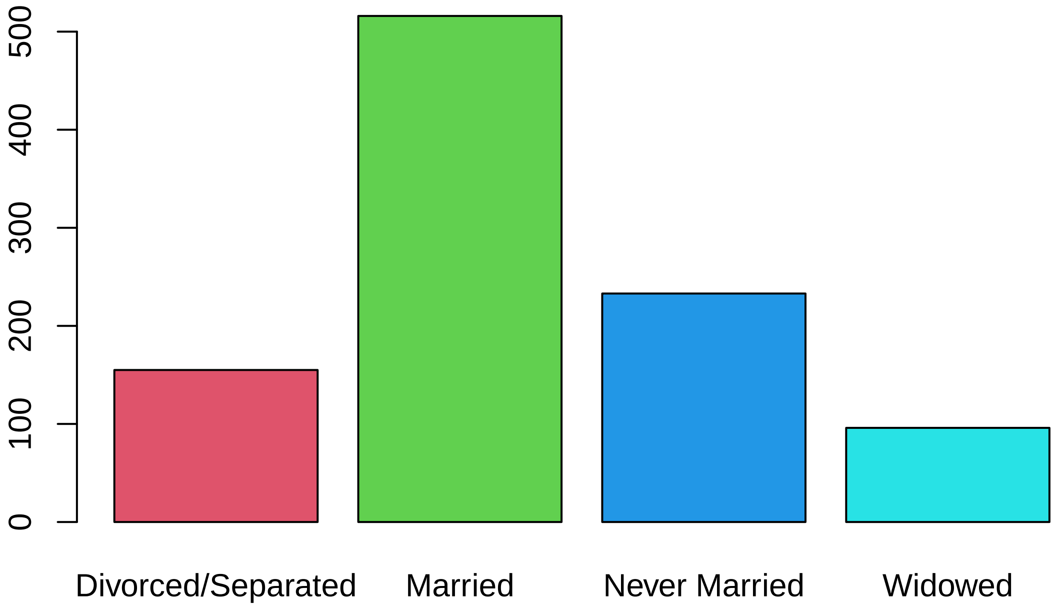
b) Display the age dist of customers using a histogram

```
hist(cust_data$age,
     main="Age Distribution of Health Insurance Customers",
     xlab="Age")
```



c) Display marital status using bar graph

```
barplot(table(cust_data$marital.stat), col=c(2,3,4,5))
```



d) How many customers are from Indiana?

```
table(cust_data$state.of.res)["Indiana"]
```

```
## Indiana
##      29
```

3) 1988 Stockton PRrimary Exit Poll

```
# http://www.stat.berkeley.edu/users/statlabs/data/vote.data
```

a) How many variables are included? Print them.

```
vote = read.table("http://www.stat.berkeley.edu/users/statlabs/data/vote.data",
                  header=TRUE)
length(names(vote))
```

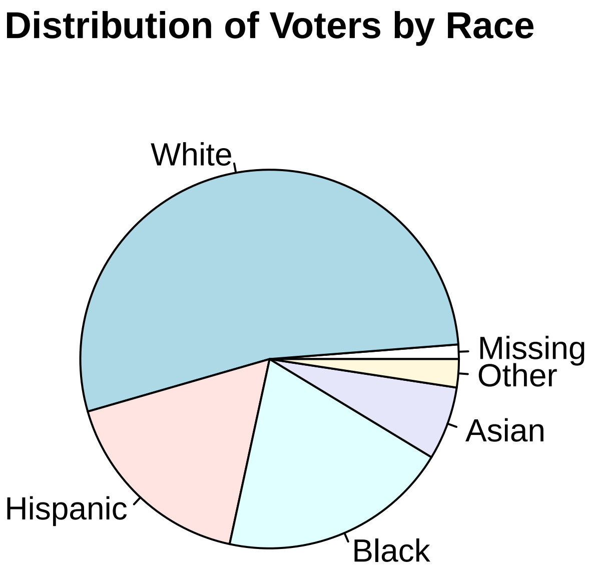
```
## [1] 4
```

```
names(vote)
```

```
## [1] "precinct"  "candidate" "race"      "income"
```

b) Display distribution of the voter's race

```
pie(table(vote$race), main="Distribution of Voters by Race",
    c("Missing", "White", "Hispanic", "Black", "Asian", "Other"))
```



4) YouthRisk

a) Import data and determine dimension

```
youth_risk = read.csv("C:\\repos\\STAT 50001\\Homework 2\\YouthRisk.csv", header=TRUE)
dim(youth_risk)
```

```
## [1] 13387     7
```

b) Is there any missing value? Remove if so

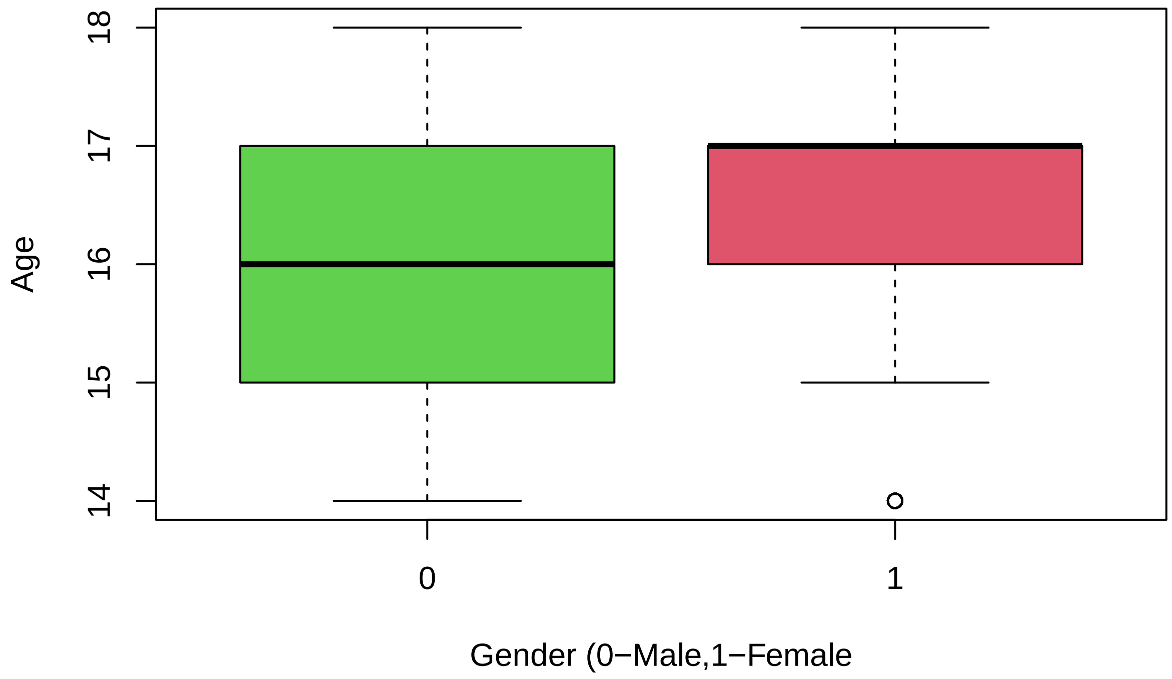
```
sum(is.na(youth_risk))
```

```
## [1] 1318
```

```
new_youth_risk = na.omit(youth_risk)
```

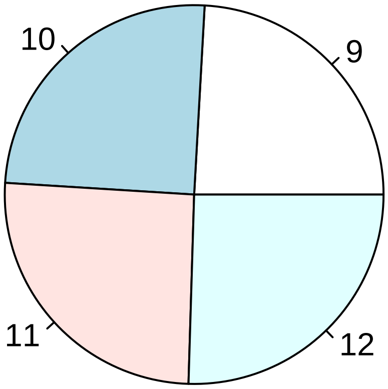
c) Display the age distribution based on gender using Parallel boxplot

```
boxplot(new_youth_risk$age4 ~ new_youth_risk$female,
        col= c(3,2),
        xlab="Gender (0-Male,1-Female", ylab="Age")
```



d) Display the grade distribution using a pie chart

```
pie(table(new_youth_risk$grade))
```



5) Generate 500 random numbers with rnorm with mean=10, var=25.

```
norm_dist = rnorm(500, mean=10, sd=sqrt(25))
```

a) How many observations are within one SD from the mean? (68%)

```
length(norm_dist[5 < norm_dist & norm_dist< 15])
```

```
## [1] 351
```

```
100 * length(norm_dist[5 < norm_dist & norm_dist< 15]) / 500
```

```
## [1] 70.2
```

```
# ~ 68%
```

b) Two?

```
length(norm_dist[0 < norm_dist & norm_dist< 20])
```

```
## [1] 481
```

```
100 * length(norm_dist[0 < norm_dist & norm_dist< 20]) / 500
```

```
## [1] 96.2
```

```
# ~ 95%
```

c) Three?

```
length(norm_dist[-5 < norm_dist & norm_dist< 25])
```

```
## [1] 499
```

```
100 * length(norm_dist[-5 < norm_dist & norm_dist< 25]) / 500
```

```
## [1] 99.8
```

```
# ~ 99.7%
```

6) FEV

a) Import data. How many children are included in the study?

```
FEV = read.table("http://www.statsci.org/data/general/fev.txt",
                 header=TRUE)
nrow(FEV)
```

```
## [1] 654
```

b) Display the FEV of male and female children

```
mean(FEV$Sex=="Male")
```

```
## [1] 0.5137615
```

```
mean(FEV$Sex=="Female")
```

```
## [1] 0.4862385
```

c) Test the hypothesis whether there is a difference in FEV of the sexes

```
# Null:         u(M) - u(F)  = 0
# Alternative:  u(M) - u(F) != 0
t.test(FEV$Sex == "Male", FEV$Sex=="Female")
```

```
##
##  Welch Two Sample t-test
##
## data:  FEV$Sex == "Male" and FEV$Sex == "Female"
## t = 0.99502, df = 1306, p-value = 0.3199
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.02674142  0.08178729
## sample estimates:
## mean of x mean of y
## 0.5137615 0.4862385
```

```
# With a p-value of 0.3199, we do not have enough evidence to reject the null.
```

d) Construct a 95% confidence interval for the difference in the mean for male and female students

```
t.test(FEV$Sex == "Male", FEV$Sex=="Female")$conf.int
```

```
## [1] -0.02674142  0.08178729
## attr(,"conf.level")
## [1] 0.95
```

7) Employee Satisfaction

a) Import the data in R

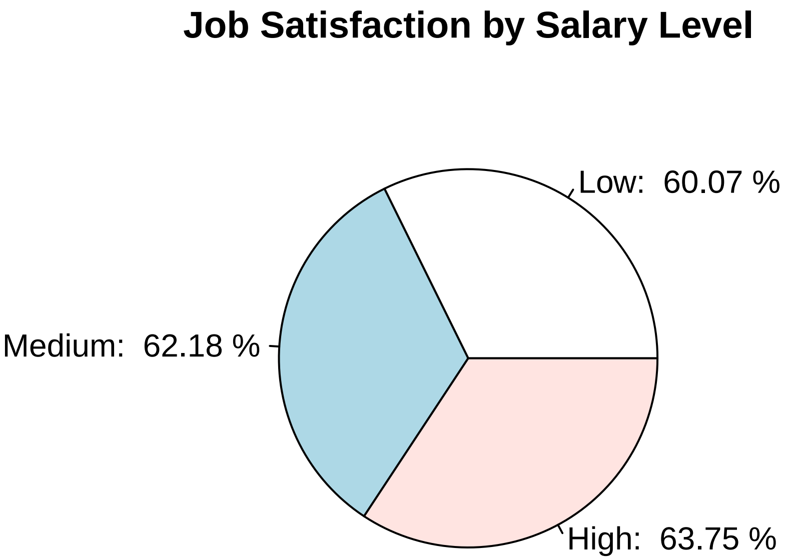
```
employee = read.csv("C:\\repos\\STAT 50001\\Homework 2\\employee.csv",
                    header=TRUE, skip=3)
```

b) Display the satisfaction scores for low, medium, and high salary employees

```
low = round(100 * mean(employee$satisfaction_level[employee$salary == "low"]),
            digits = 2)
med = round(100 * mean(employee$satisfaction_level[employee$salary == "medium"]),
            digits = 2)
high = round(100 * mean(employee$satisfaction_level[employee$salary == "high"]),
             digits = 2)

employee_labels = c(paste("Low: ", low, "%"),
                    paste("Medium: ", med, "%"),
                    paste("High: ", high, "%"))

pie(c(low,med,high),
    main = "Job Satisfaction by Salary Level",
    labels = employee_labels)
```



c) Test job satisfact level for high earners is different from low earners

```r
# Null:         u(high) - u(low)  = 0
# Alternative:  u(high) - u(low) != 0
t.test(employee$satisfaction_level[employee$salary == "high"],
       employee$satisfaction_level[employee$salary == "low"])
```

```
##
##  Welch Two Sample t-test
##
## data:  employee$satisfaction_level[employee$salary == "high"] and employee$satisfaction_level[employe
## t = 5.1713, df = 1804.8, p-value = 2.583e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.02279736 0.05065506
## sample estimates:
## mean of x mean of y
## 0.6374697 0.6007435
```

```r
# With a p-value of 2.58e-07,
# we have enough evidence to reject the null hypothesis.
```

8) PlantGrowth

a) How many observation are recorded in the data set?

```
nrow(PlantGrowth)
```

```
## [1] 30
```

b) What is the mean of each of the control and treatment conditions?

```
mean(PlantGrowth$weight[PlantGrowth$group=="ctrl"])
```

```
## [1] 5.032
```

```
mean(PlantGrowth$weight[PlantGrowth$group!="trt1"])
```

```
## [1] 5.279
```

```
mean(PlantGrowth$weight[PlantGrowth$group!="trt2"])
```

```
## [1] 4.8465
```

c) Test whether there is a significant difference between t1 and t2

```
# Null:         u(t1) - u(t2)  = 0
# Alternative:  u(t1) - u(t2) != 0
t.test(PlantGrowth$weight[PlantGrowth$group!="trt1"],
       PlantGrowth$weight[PlantGrowth$group!="trt2"])
```

```
##
##  Welch Two Sample t-test
##
## data:  PlantGrowth$weight[PlantGrowth$group != "trt1"] and PlantGrowth$weight[PlantGrowth$group != "
## t = 2.1442, df = 36.272, p-value = 0.03879
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.02352754 0.84147246
## sample estimates:
## mean of x mean of y
##    5.2790    4.8465
```

```
# With a p-value of 0.039,
# we have enough evidence to reject the null hypothesis.
```

9) Child and Health Development Study: babies

```
# Null:         u(dage) - u(age) = 0
# Alternative:  u(dage) - u(age) > 0
library(UsingR)
```

```
## Loading required package: MASS
```

```
## Loading required package: HistData
```

```
## Loading required package: Hmisc
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, units

##
## Attaching package: 'UsingR'

## The following object is masked from 'package:survival':
##
##     cancer
```

```
t.test(babies$dage, babies$age, alt="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  babies$dage and babies$age
## t = 11.067, df = 2301.5, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  2.865266      Inf
## sample estimates:
## mean of x mean of y
##  30.73706  27.37136
```

```
# With a p-value of 2.2e-16,
# we have enough evidence to reject the null hypothesis.
```

10) Doctor noshows

a) Import the data in R and identify its dims

```
noshows = read.csv("C:\\repos\\STAT 50001\\Homework 2\\noshow.csv", header=TRUE)
dim(noshows)
```

```
## [1] 110527     14
```

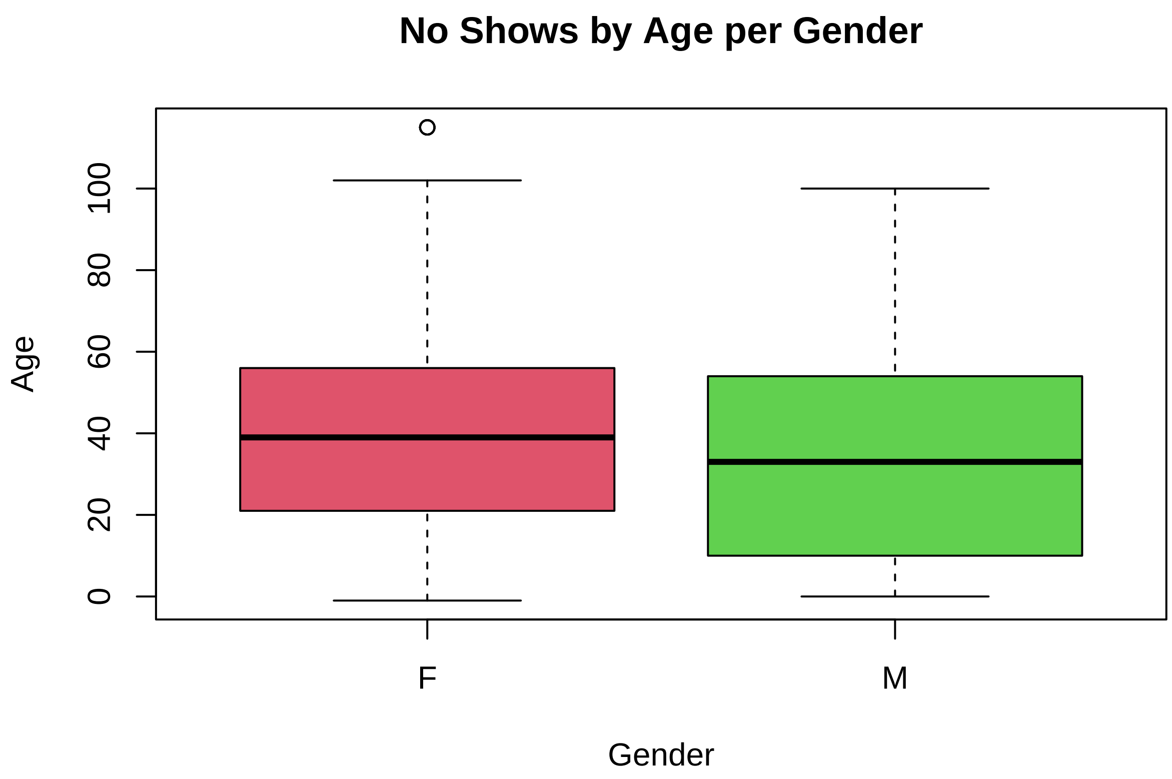
b) Print the variables included in the dataset

```
names(noshows)
```

```
##  [1] "PatientId"      "AppointmentID"  "Gender"         "ScheduledDay"
##  [5] "AppointmentDay" "Age"            "Neighbourhood"  "Scholarship"
##  [9] "Hipertension"   "Diabetes"       "Alcoholism"     "Handcap"
## [13] "SMS_received"   "No.show"
```

c) Display the Age distribution by gender creating parallel box plot

```
boxplot(noshows$Age ~ noshows$Gender, col=c(2,3),
        xlab="Gender", ylab="Age", main="No Shows by Age per Gender")
```



d) Test whether females are more likely to miss the appointment than males

```
# Null:         u(female) - u(male) = 0
# Alternative:  u(female) - u(male) > 0
t.test(noshows$Gender=="F", noshows$Gender=="M", alt="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  noshows$Gender == "F" and noshows$Gender == "M"
## t = 147.83, df = 221052, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.2966165       Inf
## sample estimates:
## mean of x mean of y
## 0.6499769 0.3500231
```

```
# With a p-value of 2.2e-16,
# we have enough evidence to reject the null hypothesis.
```

e)Are females older than males. Perform the test.

```
# Null:         u(fage) - u(dage) = 0
# Alternative:  u(fage) - u(dage) > 0
t.test(noshows$Age[noshows$Gender=="F"],
       noshows$Age[noshows$Gender=="M"], alt="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  noshows$Age[noshows$Gender == "F"] and noshows$Age[noshows$Gender == "M"]
## t = 34.561, df = 72834, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  4.911678      Inf
## sample estimates:
## mean of x mean of y
##  38.89399  33.73686
```

```
# With a p-value of 2.2e-16,
# we have enough evidence to reject the null hypothesis.
```