STAT 40001/STAT 50001 **Statistical Computing** **Fall 2024**

Homework 2- Solution

Name:

Instruction: Please submit your R code along with a brief write-up of the solutions (do not submit raw output with ERRORS:). Some of the questions below can be answered with very little or no programming. However, write R code that outputs the final answer and does not require any additional paper calculations.

Q.N. 1) Table 1 and Table 2 below are the test scores of 11 students in Test 1 and Test 2 in addition to their major field of study and class standing

Name	Major	Test 1
Aaron	MA	56
Alex	STAT	87
Brian	MA	78
Cathy	CS	87
Dough	CS	89
John	STAT	95
Lucas	STAT	98
Marcus	STAT	59
Nabin	MA	78
William	CS	87
Zoe	STAT	98

Table 1: Test 1 Scores

Name	Class	Test 2
Aaron	Junior	86
Alex	Senior	88
Brian	Junior	67
Cathy	Senior	78
Dough	Junior	89
John	Senior	87
Lucas	Senior	67
Marcus	Junior	94
Nabin	Senior	78
William	Senior	81
Zoe	Junior	83

Table 2: Test 2 scores

a) Use ***merge(.,.)*** function to create a single table containing the student's test 1 and test 2 scores.

Solution: We can save both tables and import them individually and merge them.
Alternatively, we can simply use the following codes to merge them

```
> Name=c("Aaron","Alex", "Brian","Cathy","Dough","John","Lucas","Marcus","Nabin","William","Zoe")
> Major=c("MA","STAT","MA","CS","CS", "STAT","STAT","STAT","MA","CS","STAT")
> Class=c("Junior","Senior", "Junior","Senior","Junior","Senior","Senior","Junior","Senior","Senior","Junior")
> Test1=c(56,87,78,87,89,95,98,59,78,87,98)
> Test2=c(86,88, 67,78,89,87,67,94,78,81,83)
> Table1=data.frame(Name, Major, Test1)
> Table2=data.frame(Name, Class, Test2)
> merge(Table1, Table2, by="Name")
      Name Major Test1  Class Test2
1    Aaron    MA    56 Junior    86
2     Alex  STAT    87 Senior    88
3    Brian    MA    78 Junior    67
4    Cathy    CS    87 Senior    78
5    Dough    CS    89 Junior    89
6     John  STAT    95 Senior    87
7    Lucas  STAT    98 Senior    67
8   Marcus  STAT    59 Junior    94
9    Nabin    MA    78 Senior    78
10 William    CS    87 Senior    81
11     Zoe  STAT    98 Junior    83
```

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

```
> sum(Test2>Test1)
[1] 3
```

There are 3 students who did better on the second test than the first test.

c) How many students did better in the first test?
Solution:

```
> sum(Test1>Test2)
[1] 6
```

There are 6 students that did better on the first test.

d) How many students have the same test score in both tests?
Solution:

```
> sum(Test1==Test2)
[1] 2
```

There are 2 students who have the same score in both tests.

e) Calculate the mean and standard deviation of each tests.
Solution: below are the mean and standard deviation of test 1 and test 2

```
>mean(Test1)
 [1] 82.91
> mean(Test2)
  [1] 81.64
> sd(Test1)
  [1] 14.258
> sd(Test2)
 [1] 8.675
```

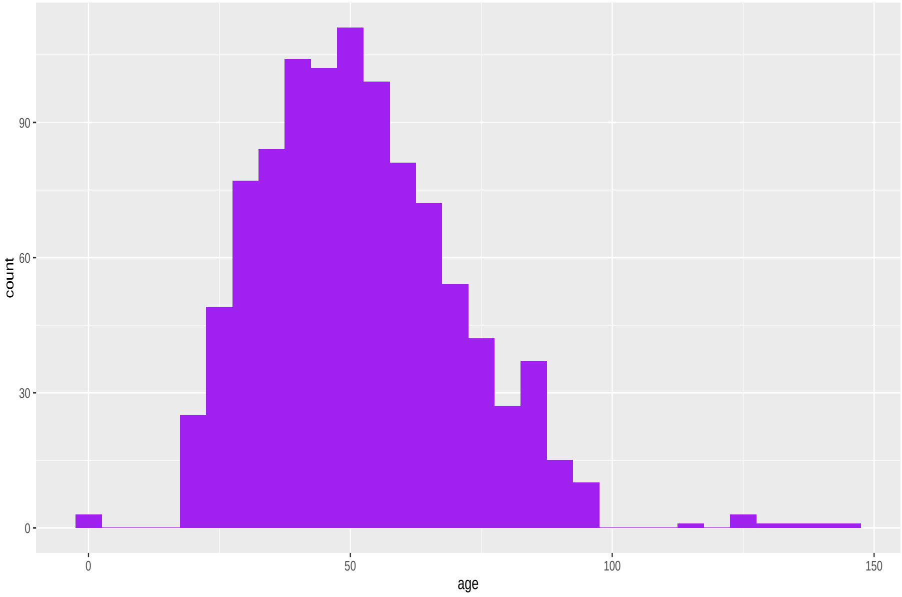
Q.N. 2) The dataset related to health insurance customers is provided in custdata.tsv. Here, “tsv”stands for tab-separated values.
a) Import the data in R
b) Display the age distributions of the customers using a histogram.
c) Display the marital status of all the customers using bar graph.
d) How many customers are from the state of Indiana?

Solution: a) We saved data in “C:” drive and import in R using R code below

```
> data=read.csv("C:\\aryalg\\Desktop\\custdata.tsv", sep="\t", header=T)
> head(data)
  custid sex is.employed income  marital.stat health.ins                 housing.type recent.move num.vehicles age state.of.res
1   2068   F          NA  11300       Married       TRUE     Homeowner free and clear       FALSE            2  49     Michigan
2   2073   F          NA      0       Married       TRUE                       Rented        TRUE            3  40      Florida
3   2848   M        TRUE   4500 Never Married      FALSE                       Rented        TRUE            3  22      Georgia
4   5641   M        TRUE  20000 Never Married      FALSE        Occupied with no rent       FALSE            0  22   New Mexico
5   6369   F        TRUE  12000 Never Married       TRUE                       Rented        TRUE            1  31      Florida
6   8322   F        TRUE 180000 Never Married       TRUE Homeowner with mortgage/loan       FALSE            1  40     New York
```

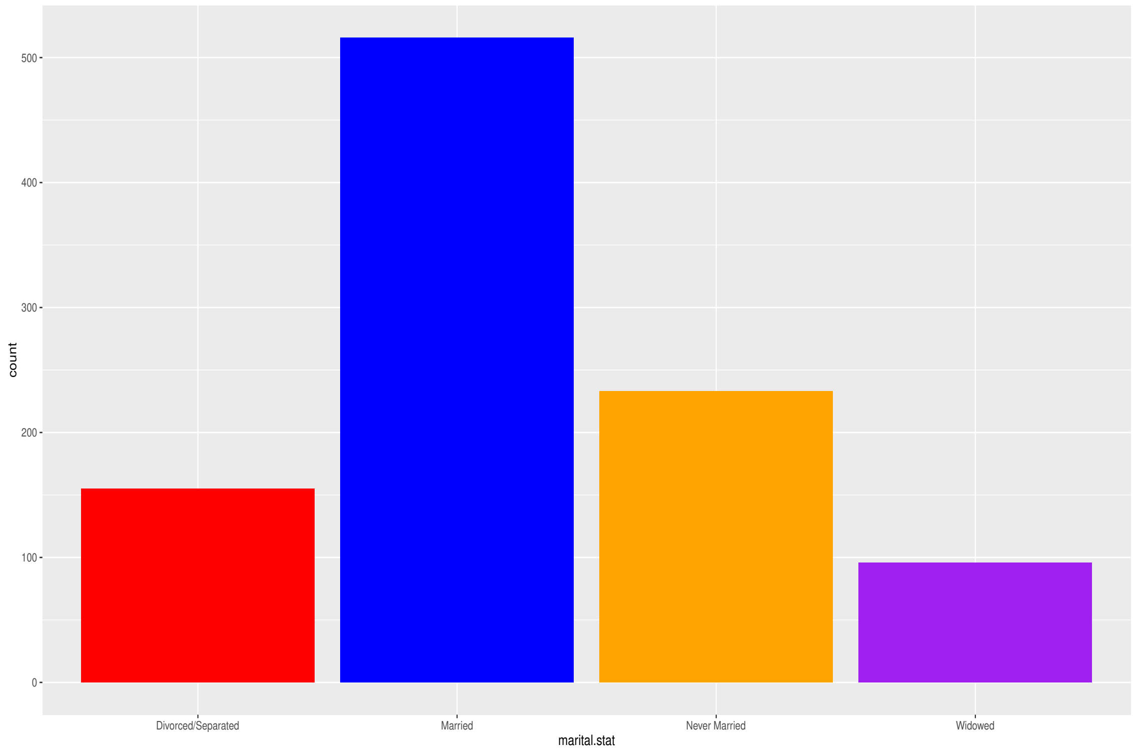
b) We used R code below to display the age distribution in ggplot2 package

```
> attach(data)
> library(ggplot2)
> ggplot(data)+geom_histogram(aes(x=age), binwidth=5,fill="purple")
```



c) The distribution of the marital status of the health insurance customer data can be displayed using R code below

```
> attach(data)
> colors=c("red","blue", "orange","purple")
> ggplot(data)+geom_bar(aes(x=marital.stat),fill=colors)
```



d) We will create the subset of the data by choosing the state of residence Indian using R code below. It appears that there are 29 customers from the state of Indiana.

```
> Indiana=subset(data, state.of.res=="Indiana")
> dim(Indiana)
[1] 29 11
```

Q.N. 3) Access the data from url `http://www.stat.berkeley.edu/users/statlabs/data/vote.data` and store the information in an object named **vote** using the function *read.table()*. This includes the 1988 Stockton Primary Exit Poll Survey:
a) How many variables are included in the survey? Please print the variables.

Solution: We can then use the following R command to access the data

```
>vote<-read.table(’http://www.stat.berkeley.edu/~statlabs/data/vote.data’, header=T)
> names(vote)
[1] "precinct" "candidate" "race" "income"
> dim(vote)
[1] 1867    4
```

There are four variables included in the data set.

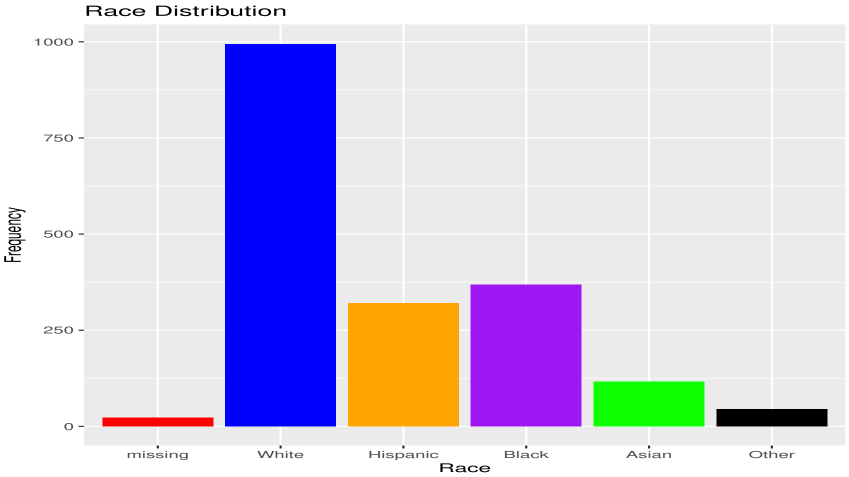
b) One of the variable included is the voter’s information is race. Note that following code are used.

$$\boldsymbol{0 = missing, 1 = White, 2 = Hispanic, 3 = Black, 4 = Asian, 5 = Other}$$

Display the distribution of the voter's race graphically using an appropriate graph. (please use ggplot2 package to create the graph.)

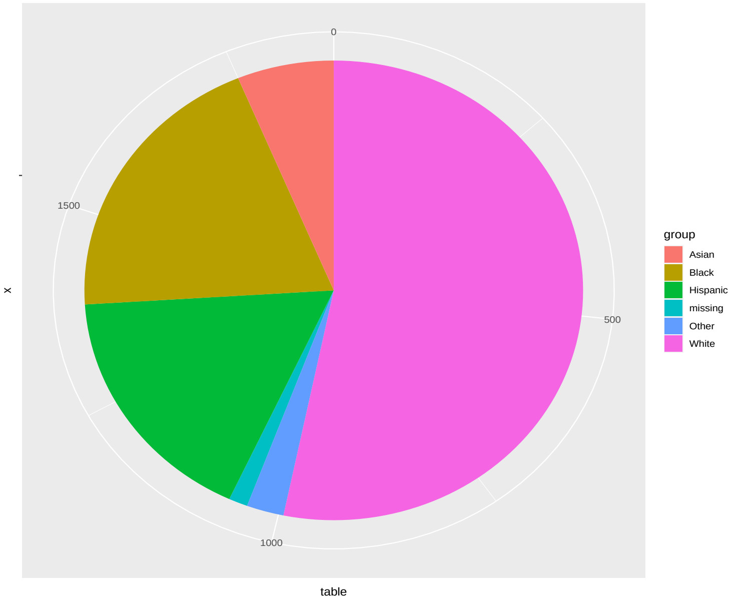
Solution: Since race is a categorical data we can display it graphically either using bargraph or piechart. R code below can be used to draw a bar chart :

```
> library(ggplot2)
> vote<-read.table("http://www.stat.berkeley.edu/~statlabs/data/vote.data", header=T)
> vote$race=factor(vote$race,levels=0:5,labels=c("missing","White","Hispanic","Black","Asian", "Other"))
> colors=c("red","blue", "orange","purple", "green", "black")
> ggplot(vote, aes(race))+geom_bar(fill=colors)+labs(title="Race Distribution", x="Race", y="Frequency")
```



Using ggplot2 package and R code below we can create the pie chart

```
> library(ggplot2)
> group=c("missing", "White","Hispanic","Black","Asian", "Other")
> table=table(vote$race)
> df=data.frame(group,table)
> ggplot(df, aes(x="",y=table,fill=group))
  +geom_bar(width=1, stat="identity")+coord_polar("y", start=0)
```



Q.N. 4) This dataset (YouthRisk, provided with this assignment) is derived from the 2007 Youth Risk Behavior Surveillance System (YRBSS), which is an annual survey conducted by the Centers for Disease Control and Prevention (CDC) to monitor the prevalence of health-risk youth behaviors. This dataset focuses on whether or not youths have recently (in past 30 days) ridden with a drunk driver. The description of the variables is provided at
`https://vincentarelbundock.github.io/Rdatasets/doc/Stat2Data/YouthRisk.html`
a) Import the data in R and determine its dimension.
b) Is there any missing value? If so please remove the missing values from the data set.
c) Display the age distribution of the individuals based on gender using Parallel boxplot.
d) Display the grade(Year in high school) distribution using a pie chart.

Solution:
a) We use R code below to import the data and determine its dimension

```
> data=read.csv("C:\\Users\\aryalg\\STAT 40001\\Assignmants\\YouthRisk.csv")
> attach(data)
> dim(data)
[1] 13387     7
```

b) We use R code below to check whether there is a missing value

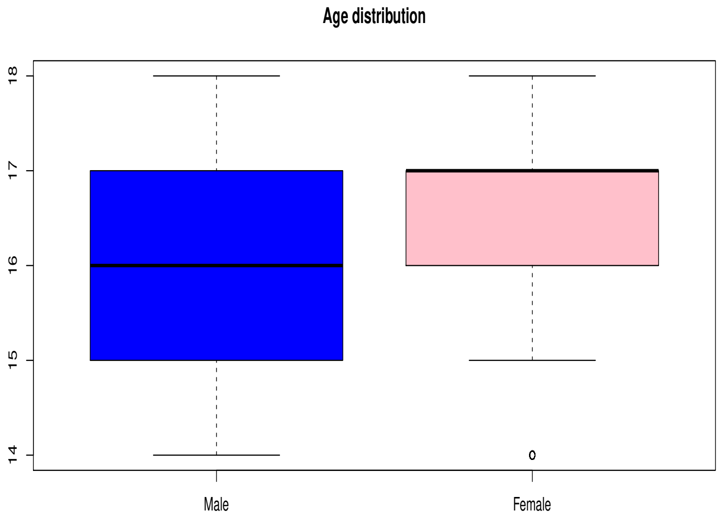
```
> any(is.na(data))
[1] TRUE
```

In order to remove the missing values we use R code below

```
> newdata=na.omit(data)
> dim(newdata)
[1] 12282     7
```

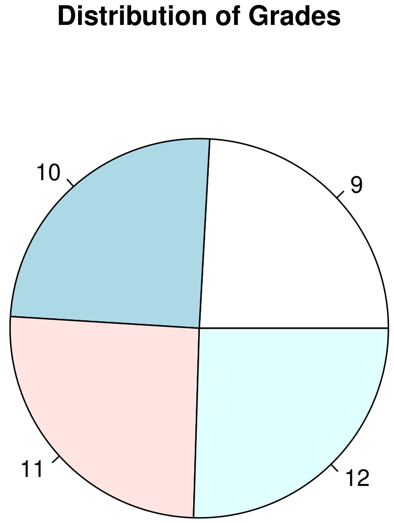
It appears that thee were 13387-12282=1,105 observations with missing value(s).
c) In order to display the age distribution we use R code below

```
> boxplot(newdata$age4~newdata$female, col=c("blue", "pink"),
 names=c("Male","Female"), main="Age distribution")
```



d) We use R code below to display the gender distribution

```
> pie(table(newdata$grade),main='Distribution of Grades')
```



Q.N. 5) Generate 500 random numbers from normal distribution with mean 7 and variance 16. How many observations are within one, two and there standard deviations from the mean? Compare your findings with the empirical rule.
According to the empirical rule **68%**, **95%** *and* **99.7%** *data reside within one, two and three standard deviation of the mean. Does your data meet this rule?*
Solution: We use the code below in R to generate 500 random numbers from normal random variable with mean 50 and standard deviation 5 and count the number of observations in the given range

```
> x=rnorm(500,mean=7,sd=4)
> sum(3<x&x<11)/500*100
[1] 66.2
> sum(-1<x&x<15)/500*100
[1] 95.6
> sum(-5<x&x<19)/500*100
[1] 99.4
```

For this random sample, we observed that **66.2%, 95.6%** *and* **99.4%** *data are within one, two and three standard deviation of the mean respectively. Very close to the results as expected by the empirical rule.*

Q.N. 6) FEV (forced expiratory volume) is an index of pulmonary function that measures the volume of air expelled after one second of constant effort. The data provided in the link below contains determinations of FEV on children ages 6-22 who were seen in the Childhood Respiratory Disease Study in 1980 in East Boston, Massachusetts. The data are part of a larger study to follow the change in pulmonary function over time in children.
ID - ID number
Age - years FEV - litres
Height - inches
Sex - Male or Female

`http://www.statsci.org/data/general/fev.txt`
a) Import the data in R. How many children are included in this study?
b) How many female smokers are included in the database?
c) Display the FEV of Male and Female children using parallel boxplot.
d) Test the hypothesis whether there is a difference in FEV for male and female.
e) Construct a 95% confidence interval for the difference in the mean FEV for male and female students.

Solution:
a) We use R code below to import the data in R

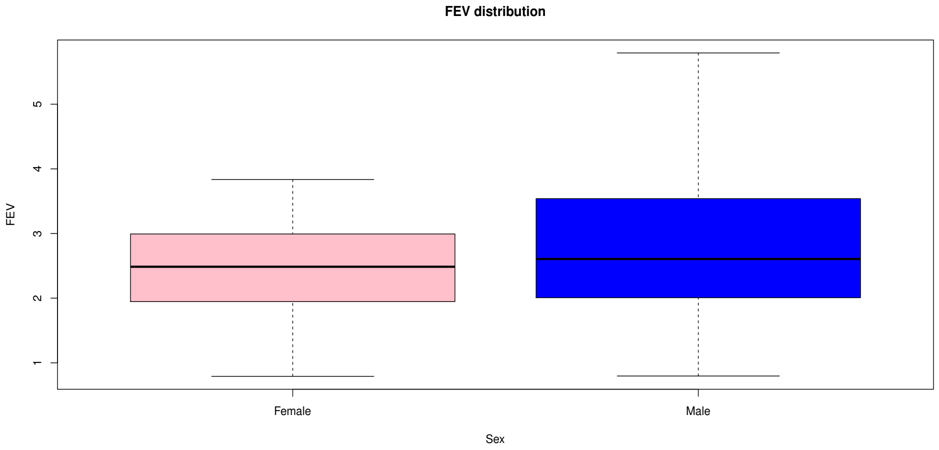
```
> data=read.table("http://www.statsci.org/data/general/fev.txt", header=T)
> head(data,3)
   ID Age   FEV Height    Sex Smoker
1 301   9 1.708   57.0 Female    Non
2 451   8 1.724   67.5 Female    Non
3 501   7 1.720   54.5 Female    Non
> dim(data)
[1] 654   6
```

There are 654 children included in this study.
b) Based on the R output below there are 39 female smokers.

```
> attach(data)
> xtabs(~Sex+Smoker)
        Smoker
Sex      Current Non
  Female      39 279
  Male        26 310
```

c) We use R code below to display the FEV distribution for male and female.

```
> boxplot(FEV~Sex, col=c("pink", "blue"), main="FEV distribution")
```



d) Let $\boldsymbol{\mu_M}$ *and* $\boldsymbol{\mu_F}$ *be the average FEV for male and female respectively. We would like to test the following hypothesis*

$$\begin{aligned} \boldsymbol{H_0} &: \boldsymbol{\mu_F = \mu_M} \\ \boldsymbol{H_a} &: \boldsymbol{\mu_F \neq \mu_M} \end{aligned}$$

Based on the R output below, since p-value is less than 0.05 we reject the null hypothesis and conclude that there is overwhelming evidence that the FEV differ based on gender.

```
> t.test(FEV~Sex)

        Welch Two Sample t-test

data:  FEV by Sex
t = -5.5037, df = 575.75, p-value = 5.604e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4902038 -0.2323494
sample estimates:
mean in group Female   mean in group Male
            2.451170             2.812446
```

e) 95% confidence interval for the difference in FEV (Female-Male) is $\boldsymbol{(-0.4902038 - 0.2323494)}$*and for Male-Female is* $\boldsymbol{(0.2323494, 0.4902038)}$.

```
> t.test(FEV~Sex)$conf.int
[1] -0.4902038 -0.2323494
attr(,"conf.level")
[1] 0.95
```

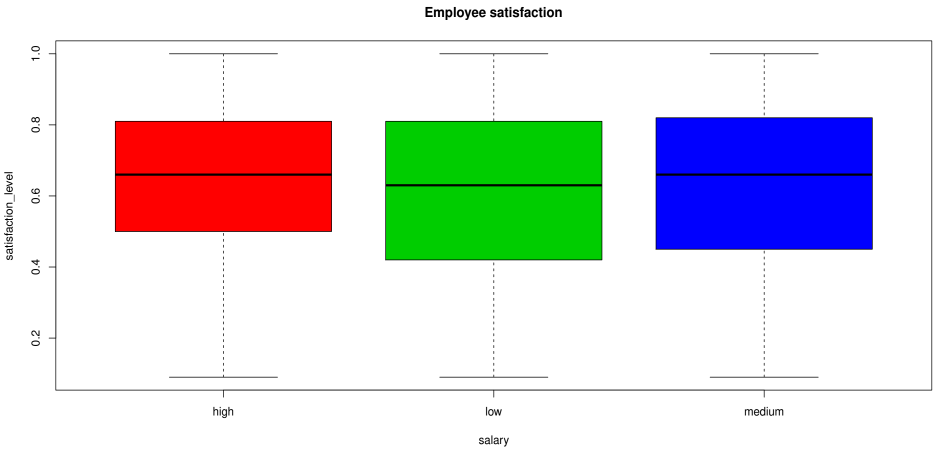
Q.N. 7) The employee satisfaction in any job depends on several factors including the salary. The attached data (Employee Satisfaction) provides information about 15000 employees.
a) Import the data in R
b) Display the satisfaction scores for low, medium and high salary employees.
c) Test whether the job satisfaction level for high earning employees is significantly different from the low earning employees.

Solution:
a) Note that first three lines of the data should be remove as it only contains the description of the data so we use R code below to import the data without description

```
> data=read.csv("C:\\Users\\aryal\\Purdue- Northwest\\STAT 40001\\Assignment\\employee.csv", skip=3)
> head(data)
  employee_id satisfaction_level last_evaluation_score number_of_projects average_monthly_hours years_spent_at_company work_accident left_company
1           1               0.38                  0.53                  2                   157                      3             0            1
2           2               0.80                  0.86                  5                   262                      6             0            1
3           3               0.11                  0.88                  7                   272                      4             0            1
4           4               0.72                  0.87                  5                   223                      5             0            1
5           5               0.37                  0.52                  2                   159                      3             0            1
6           6               0.41                  0.50                  2                   153                      3             0            1
  promotion_in_last_5years department salary      salary_range
1                        0      sales    low Less than $45,000
2                        0      sales medium $45,000 - $74,999
3                        0      sales medium $45,000 - $74,999
4                        0      sales    low Less than $45,000
5                        0      sales    low Less than $45,000
6                        0      sales    low Less than $45,000
> attach(data)
>
```

b) We can use R code below to display the employee satisfaction based on their salary.

```
> boxplot(satisfaction_level~salary,col=c(2,3,4), main="Employee satisfaction")
```



c) Let $\boldsymbol{\mu_H}$ and $\boldsymbol{\mu_L}$ be the average satisfaction level for high earning and low earning employees respectively. We would like to test the following hypothesis

$$\begin{aligned} \boldsymbol{H_0} &: \boldsymbol{\mu_H = \mu_L} \\ \boldsymbol{H_a} &: \boldsymbol{\mu_H \neq \mu_L} \end{aligned}$$

Based on the R output below, since p-value is less than 0.05 we reject the null hypothesis and conclude that there is a significance difference on the job satisfaction based on the earning level.

```
> Low=subset(data, salary=="low")
> High=subset(data, salary=="high")
> dim(Low);dim(High)
[1] 7317   12
[1] 1237   12
> t.test(High$satisfaction_level, Low$satisfaction_level)

        Welch Two Sample t-test

data:  High$satisfaction_level and Low$satisfaction_level
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
```

Q.N. 8) Results from an experiment to compare yields (as measured by dried weight of plants) obtained under a control and two different treatment conditions is provided in the data frame `PlantGrowth` in the R dataset.
a) How many observations are recorded in the data set?
b) What is the mean of each of the control and treatment conditions?
c) Test the hypothesis whether there is a significance difference between the treatment 1 and treatment 2.

Solution:
a) Note that the data set `PlantGrowth` *in the R dataset so we can simply use the R code below to read the data and identify the number of observations*

```
> data(PlantGrowth)
> dim(PlantGrowth)
[1] 30  2
```

Therefore, there are 30 observations measured in two variables.
b) We can use the R code below to find the mean of each groups:

```
> tapply(PlantGrowth$weight,PlantGrowth$group,mean)
 ctrl  trt1  trt2
5.032 4.661 5.526
```

OR

```
> x1=subset(PlantGrowth, group=="ctrl")
> x2=subset(PlantGrowth, group=="trt1")
> x3=subset(PlantGrowth, group=="trt2")
> mean(x1$weight)
[1] 5.032
> mean(x2$weight)
[1] 4.661
```

```
> mean(x3$weight)
[1] 5.526
```

c) Let $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ denote the mean dried weight of plants under treatment 1 and treatment 2 respectively. We would like to test the following hypothesis

$$\begin{aligned} \boldsymbol{H_0} &: \boldsymbol{\mu_1 = \mu_2} \\ \boldsymbol{H_a} &: \boldsymbol{\mu_1 \neq \mu_2} \end{aligned}$$

In order to test whether there is a significance difference between the treatment 1 and treatment 2 we can use R code below

```
> t.test(x2$weight, x3$weight)
        Welch Two Sample t-test
data:  x2$weight and x3$weight
t = -3.0101, df = 14.104, p-value = 0.009298
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.4809144 -0.2490856
sample estimates:
mean of x mean of y
    4.661     5.526
```

Note that $\boldsymbol{p - valve < 0.05}$ *so we reject the null hypothesis and conclude that there is a significance difference in the weight based on the treatment type.*

Q.N. 9) The babies data frame in the `UsingR` packages has a collection of variables taken for each new mother in a Child and Health Development Study. The variable age contains the mom's age and the variable `dage` contains the dad's age for several babies. Do a significance test of the null hypothesis of equal ages against a one-sided alternative that dads are older.

Solution: Suppose $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ be the mean of of dad's age and mom's age respectively. We would like to test the following

$$\begin{aligned} \boldsymbol{H_0} &: \boldsymbol{\mu_1 \leq \mu_2} \\ \boldsymbol{H_a} &: \boldsymbol{\mu_1 > \mu_2} \end{aligned}$$

We use the R code below to access the data and perform the test

```
>library(UsingR)
> data(babies)
>attach(babies)
> t.test(dage,age,alternative="greater")

        Welch Two Sample t-test

data:  dage and 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
```

Decision: Since $\mathbf{p} < \boldsymbol{\alpha}$ *for typical value of* $\boldsymbol{\alpha} = \mathbf{0.05}$*, we reject the null hypothesis and conclude that there is evidence that fathers are older than mothers.*

Q.N. 10) A person makes a doctor appointment, receives all the instructions and doesn't show up for appointment, Who to blame? Data set containing some information including the age, gender are provided in the data set (Noshow). (Other variable names are self-explanatory).
a) Import the data in R and identify its dimension
b) Print the variables included in the dataset.
c) Display the Age distribution by gender creating parallel box plot.
d) Test whether female are more likely to miss the appointment than male.
e) Are female older than male? Perform the test.
(Hint: xtabs function in R will be useful to create Cross-Tabulation in part(d))

Solution: a) We use R code below to import and find its dimension

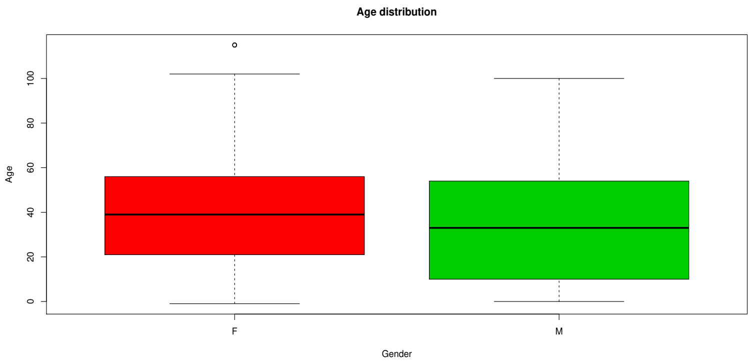
```
> data=read.csv("C:\\Users\\aryalg\\STAT 40001\\Assignmants\\noshow.csv")
> dim(data)
[1] 110527     14
```

b) The variables of the dataset are printed using R code below

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

c) We use R code below to create a parallel boxplot

```
> attach(data)
> plot(Age~Gender, col=c(2,3), main="Age distribution")
```



d) First we will create a cross tabulation of the data by Gender and No.show and then perform the test using R code below

```
> xtabs(~Gender+No.show)
      No.show
Gender    No   Yes
     F 57246 14594
     M 30962  7725

> prop.test(x=c(14594,7725), n=c(57246+14594, 30962+7725), alt="greater")

        2-sample test for equality of proportions with continuity correction

data:  c(14594, 7725) out of c(57246 + 14594, 30962 + 7725)
X-squared = 1.8534, df = 1, p-value = 0.08669
alternative hypothesis: greater
95 percent confidence interval:
 -0.0007094951  1.0000000000
sample estimates:
   prop 1    prop 2
0.2031459 0.1996795
```

Since $\boldsymbol{p - value > \alpha}$ *for typical value of* $\boldsymbol{\alpha = 0.05}$*, we fail to reject the null hypothesis and conclude that there is no evidence that female are more likely to miss the appointment than male.*

e) We use R code below to perform the test

```
> t.test(Age~Gender, alt="greater")

        Welch Two Sample t-test

data:  Age by Gender
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 in group F mean in group M
       38.89399        33.73686
```

Since $\boldsymbol{p < \alpha}$*, we reject the null hypothesis and conclude that on average Female are older than Male.*