Mini Project - Golf

Table of Contents

1. Project Objective

The objective of the report is to explore the Golf ball data set (“Golf.csv”) in R and generate insights about the dataset. This exploration report consists of the following:

- Importing the data set in R
- Understanding the structure of the data set
- Graphical exploration
- Descriptive statistics
- Insights from the dataset

2. Assumptions

Since, we are going to measure the difference between two means of the two groups in the given data set, we assume, two sample t test would be appropriate to test the data. The assumptions of the two sample t test are mentioned below.

- The data is continuous
- The data follow the normal distribution
- The variances of two samples are equal
- The two samples are independent
- Both the samples are simple random samples from their respective populations meaning, each individual in the population has an equal probability of being selected in the sample.

3. Exploratory Data Analysis – Step by step approach

A typical data exploration activity consists of the following steps:

1. Environment set up and Data import
2. Variable identification
3. Univariate Analysis
4. Bi-Variate Analysis
5. Missing value treatment (Not in scope of our project)
6. Outlier Treatment (Not in scope of our project)
7. Variable Transformation / Feature Creation
8. Feature Exploration

We shall follow these steps in exploring the given data set.

Although Steps 5 and 6 are not in scope for this project, a brief about these steps (and other steps as well) is given, as these are important steps for Data Exploration journey.

1. **Environment Set up and Data Import**

   i. **Install necessary packages and invoke libraries**
   Use this section to install necessary packages and invoke associated libraries. Having all the packages at the same places increases code readability.

   ii. **Set up Working Directory**

   Setting a working directory on starting of the R session makes importing and exporting data files and code files easier. Basically, working directory is the location/ folder on the PC where you have the data, codes etc. related to the project.

   Please refer Appendix A for Source Code.

   iii. **Import and Read the Dataset**

   The given dataset is in .csv format. Hence, the command 'read.csv' is used for importing the file. Please refer Appendix A for Source Code.

```
#Read the input file

golf = read.csv("Golf.csv")

attach(golf)
```

2. **Variable Identification**

```
#Find the internal structure of the data
str(golf)
#Find the descriptive statistics of the data
summary(golf)
#Test if the two samples are normally distributed
#Draw a box plot for comparing Current Ball and New Ball samples
boxplot(CurrentBall, NewBall, horizontal = TRUE,names = c("Current Ball", "New Ball"), main="Comparative box plot of New Ball and Current ball Data")
# Draw historgram for Current Ball group
hist(CurrentBall)
# Draw historgram for New Ball group
hist(NewBall)
```

# Use var.test function to check the equality of variances.

var.test(CurrentBall, NewBall, alternative = "two.sided")

# Calculate the p-value using 2 sample t test to compare the driving distances of

# Current and new golf ball data in the given data set

t.test(CurrentBall, NewBall, alternative = "two.sided", mu=0, paired=FALSE, var.equal=TRUE, conf.level = 0.95)

#Calculate the 95% confidence Interval for the population of Current Ball

t.test(CurrentBall)

#Calculate the 95% confidence Interval for the population of New Ball

t.test(NewBall)

#Calculate the Standard Deviation of Current Ball Data

sd(CurrentBall)

#Calculate the Standard Deviation of New Ball Data

sd(NewBall)

#Find the Standard Deviation of difference of the two means of the data

SD=sd(NewBall - CurrentBall)

# Find the true difference in the means of two samples

delta = mean(CurrentBall) - mean(NewBall)

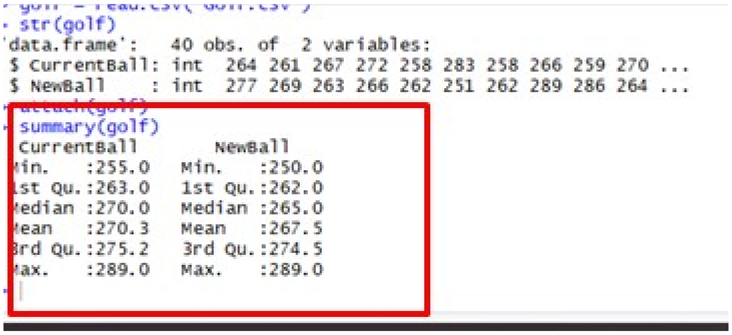
#Calculate the power of t test with the default value of beta=0.1(probability of type II error)

power.t.test(power = 0.9, delta = delta, sd=SD, sig.level = 0.05, type = "two.sample", alternative = "two.sided")

i. **Variable Identification – inferences**

#Find the descriptive statistics of the data

summary(golf)

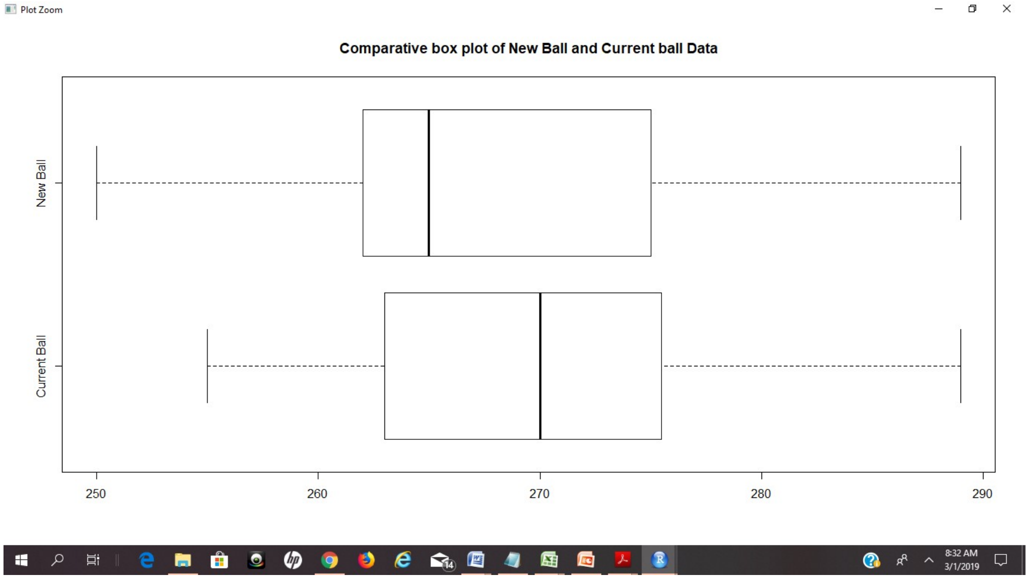


```
> golf = read.csv("Golf.csv")
> str(golf)
'data.frame':	40 obs. of  2 variables:
 $ CurrentBall: int  264 261 267 272 258 283 258 266 259 270 ...
 $ NewBall    : int  277 269 263 266 262 251 262 289 286 264 ...
> attach(golf)
> summary(golf)
  CurrentBall        NewBall     
 in.   :255.0   Min.   :250.0  
 st Qu.:263.0   1st Qu.:262.0  
 edian :270.0   Median :265.0  
 ean   :270.3   Mean   :267.5  
 rd Qu.:275.2   3rd Qu.:274.5  
 ax.   :289.0   Max.   :289.0  
> 
```

boxplot(CurrentBall, NewBall, horizontal = TRUE,names = c("Current Ball", "New Ball"), main="Comparative box plot of New Ball and Current ball Data")

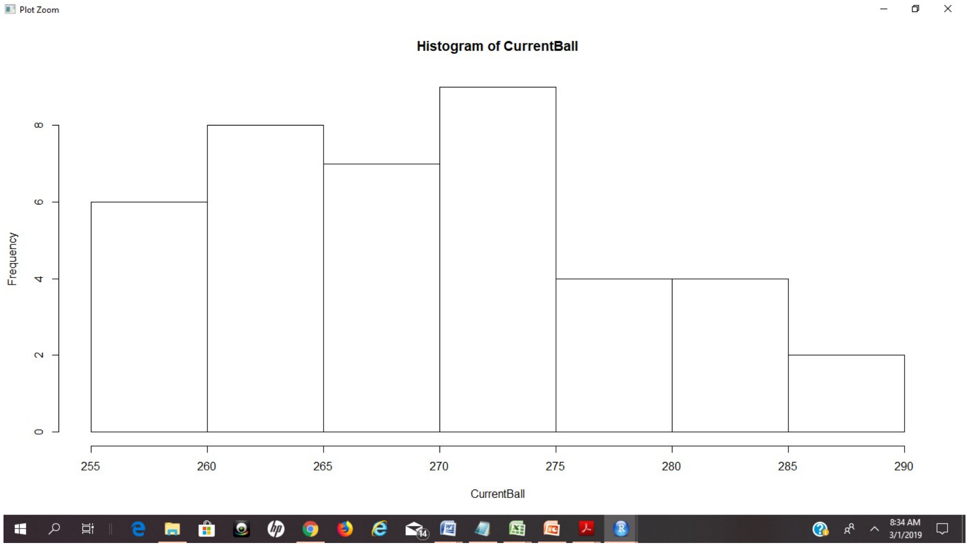
- ➢ As per the box plot, both the samples are independent of each other

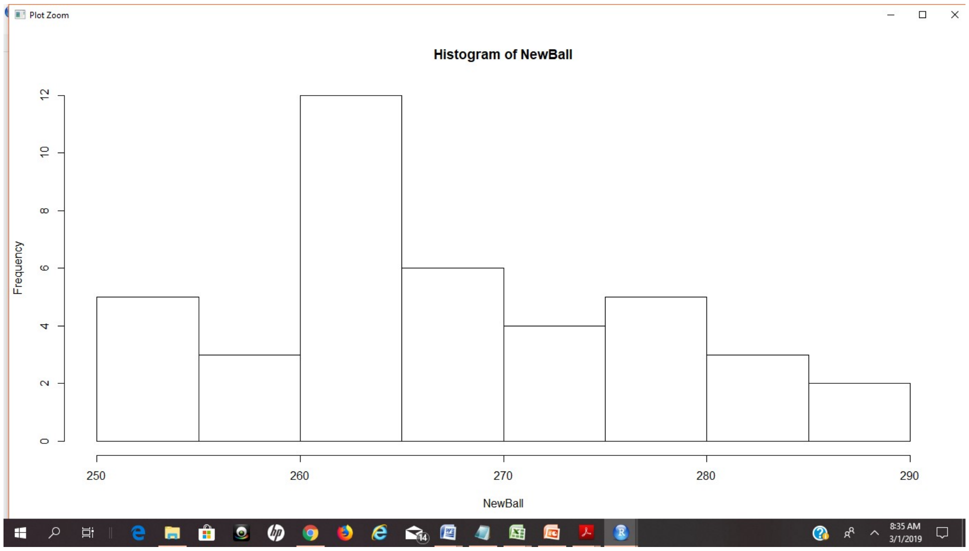
- As per the box plot, the mean and median are almost same for both the groups
- Hence, we can assume that each group data is normally distributed



```
hist(CurrentBall)
hist(NewBall)
```

- As per the histograms, data seems to be normally distributed



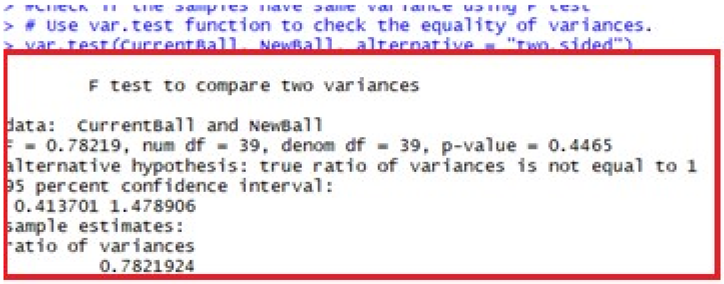


Use var.test function to check the equality of variances.

var.test(CurrentBall, NewBall, alternative = "two.sided")

Null Hypothesis, H0 = Variance of Current Ball = Variance of New Ball

Alternate Hypothesis Ha = Variance of Current Ball != Variance of New Ball



```
> #Check if the samples have same variance using F test
> # Use var.test function to check the equality of variances.
> var.test(CurrentBall, NewBall, alternative = "two.sided")

        F test to compare two variances

data:  CurrentBall and NewBall
F = 0.78219, num df = 39, denom df = 39, p-value = 0.4465
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.413701 1.478906
sample estimates:
ratio of variances
         0.7821924
```

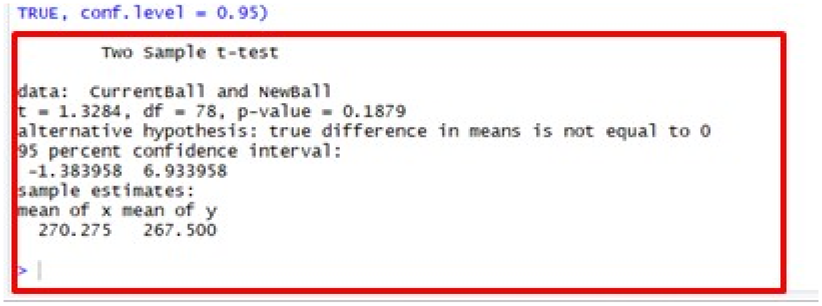
- Since p-Value of 0.4465 is > 0.05, we fail to reject the null hypothesis Hence both the variances are equal.

# Calculate the p-value using 2 sample t test to compare the driving distances of Current and new golf ball data in the given data set

Null Hypothesis H0= Mu New - Mu Curr = 0

Alternate Hypothesis Ha = Mu New - Mu Curr != 0

t.test(CurrentBall, NewBall, alternative = "two.sided", mu=0, paired=FALSE, var.equal=TRUE, conf.level = 0.95)

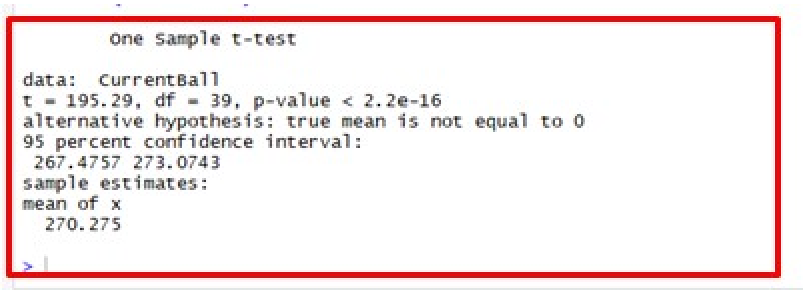


```
TRUE, conf.level = 0.95)

        Two Sample t-test

data:  CurrentBall and NewBall
t = 1.3284, df = 78, p-value = 0.1879
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.383958  6.933958
sample estimates:
mean of x mean of y
  270.275   267.500

> |
```

- ➢ Since p-value 0.1879 > 0.05, we fail to reject the null hypothesis Hence, we conclude that, there is no considerable difference between the two means of data to prove that the new coating has effect on the driving distances of golf balls.
- ➢ 95 percent confidence interval for the difference between the means of the population is: -1.383958 6.933958

#Calculate the 95% confidence Interval for the population of Current Ball

t.test(CurrentBall)

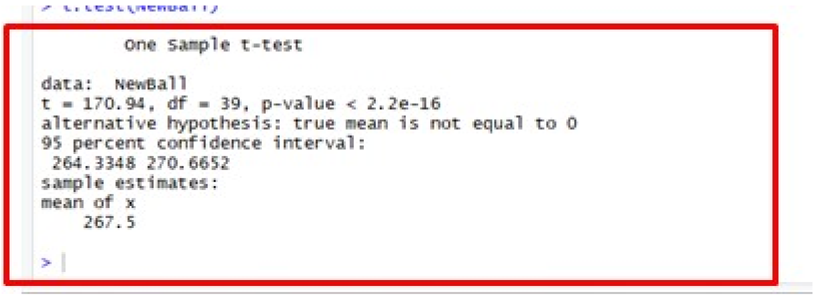


```
        One Sample t-test

data:  CurrentBall
t = 195.29, df = 39, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 267.4757 273.0743
sample estimates:
mean of x
  270.275

> |
```

- ➢ 95 percent confidence interval:267.4757 273.0743

#Calculate the 95% confidence Interval for the population of New Ball

t.test(NewBall)

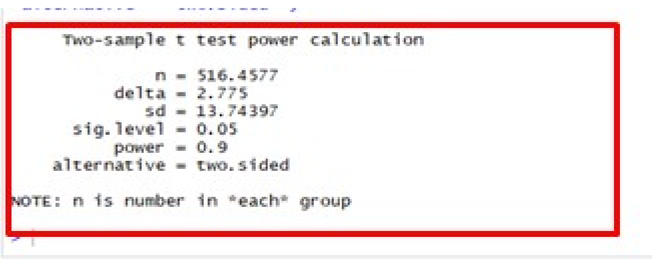


```
> t.test(NewBall)

        One Sample t-test

data:  NewBall
t = 170.94, df = 39, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 264.3348 270.6652
sample estimates:
mean of x
    267.5

> |
```

- ➢ 95 percent confidence interval:264.3348 270.6652

#Calculate the power of t test with the default value of beta=0.1(probability of type II error)

power.t.test(power = 0.9, delta = delta, sd=SD, sig.level = 0.05, type = "two.sample", alternative = "two.sided")



```
     Two-sample t test power calculation

              n = 516.4577
          delta = 2.775
             sd = 13.74397
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number in *each* group
```

- ➢ The power of the test is the probability that we make the right decision when the null is not correct.
- ➢ The above result shows that, we need a sample size of 517 (rounded to the next whole number) to make the right decision about the golf balls.

3. Univariate Analysis

<Explore individual variables one by one>
<Present your findings in tabular format>
<Summarise key observations about each variable>

4. Bi-Variate Analysis

<To Explore relationship between two variables>
<Interpret the findings>

5. Missing Value Identification

<See if any missing values>

6. Outlier Identification

<See if any outliers>

7. Variable Transformation / Feature Creation

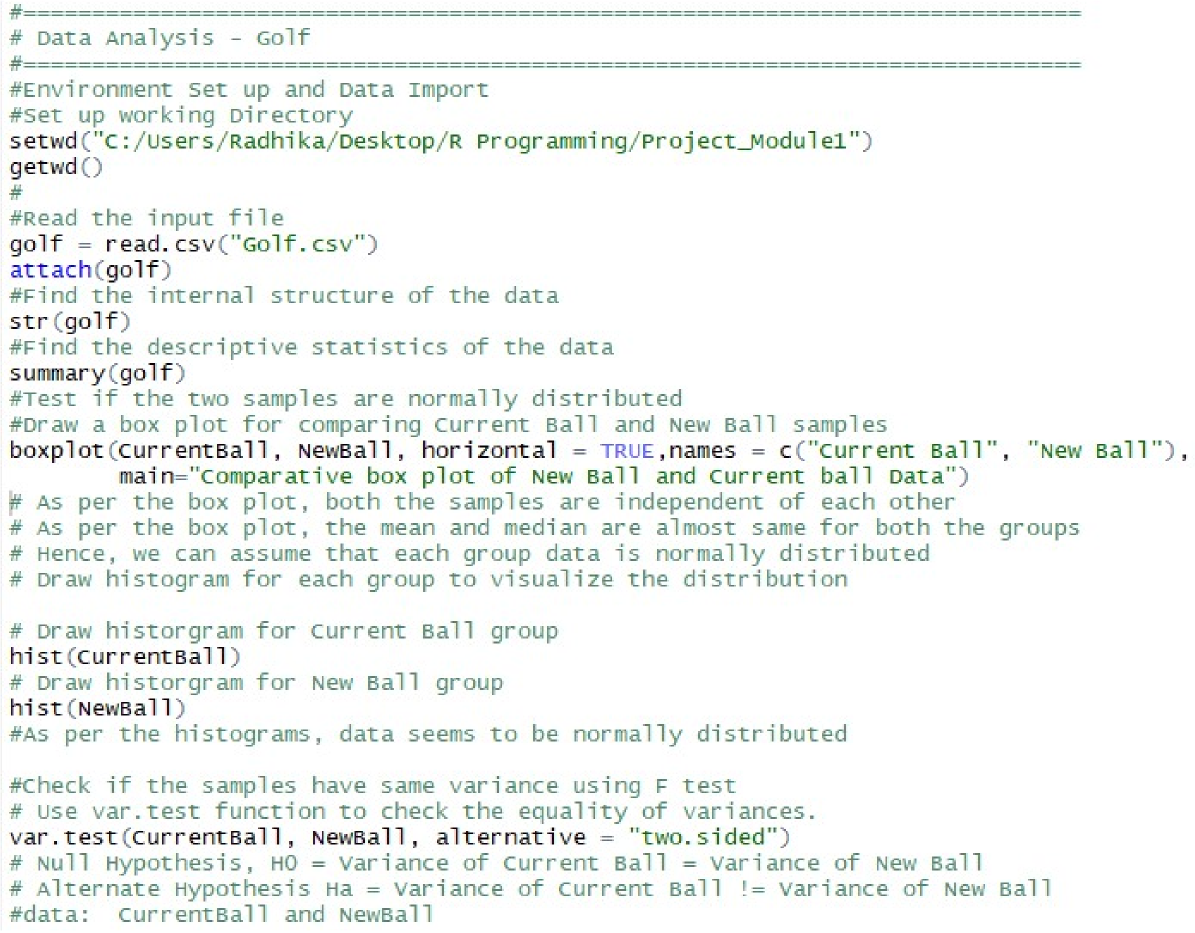
<Do you see a need of transforming a variable/ creating new variables for better understanding of the data, or presenting the results to the customer?>
<Act accordingly>

4. Conclusion

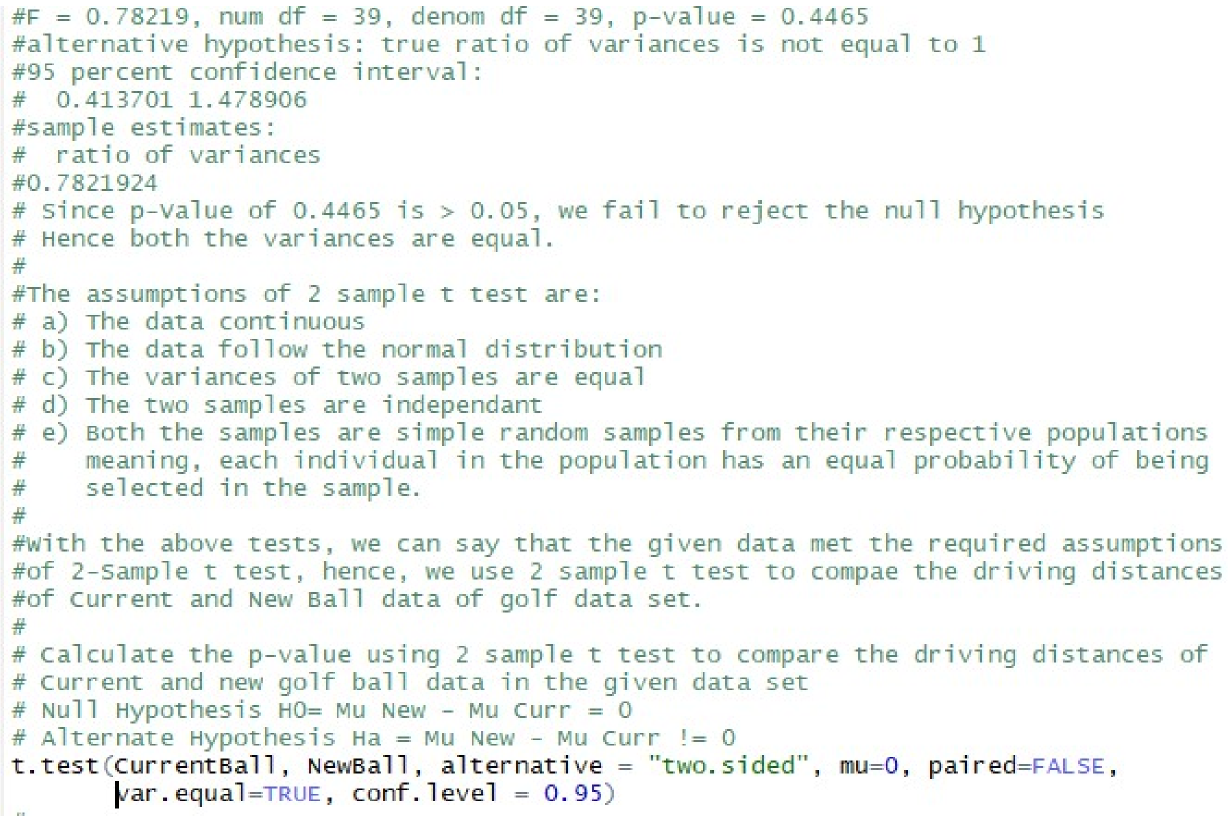
- ➢ Since p-value 0.1879 > 0.05, we fail to reject the null hypothesis. Hence, we conclude that, there is no considerable difference between the two means of data to prove that the new coating has effect on the driving distances of golf balls.

- The power of the test with the default probability of 0.1(Default probability of type II error) shows that, we need a sample size of 517 balls in each group to make the right decision about the golf balls.

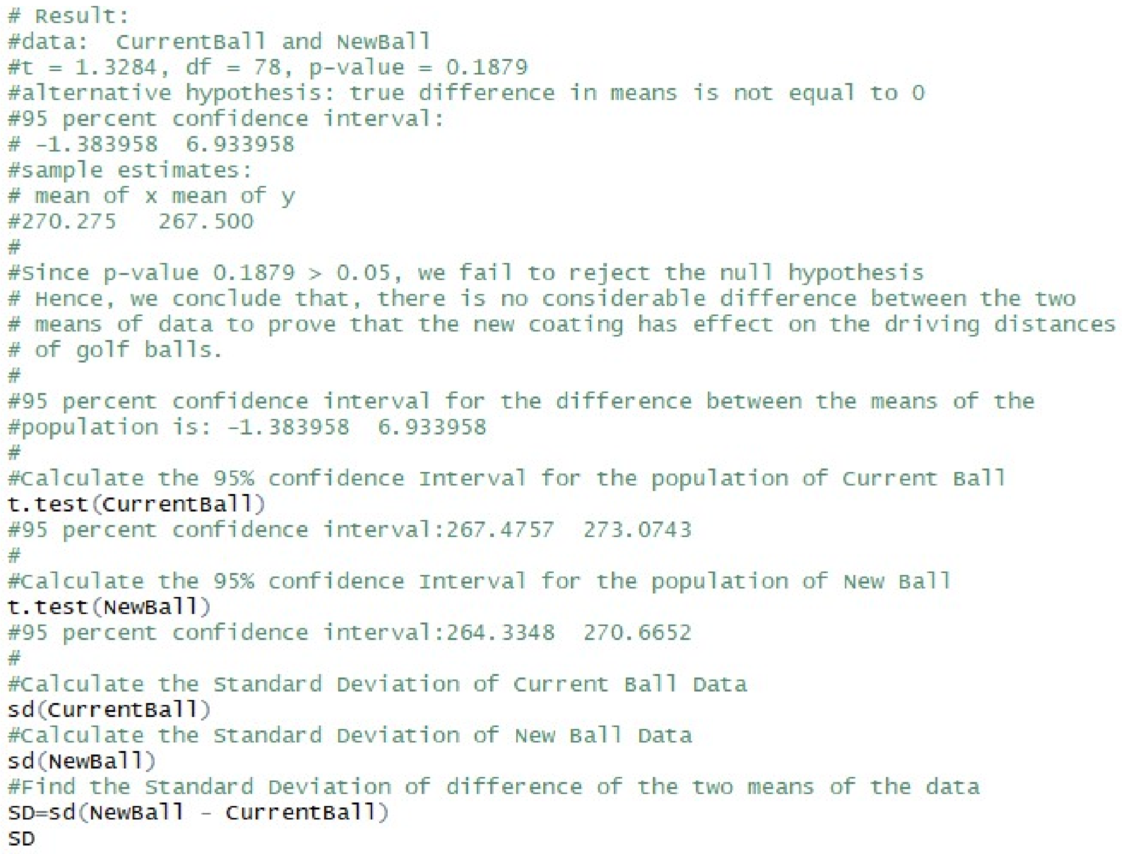
5. Appendix A – Source Code



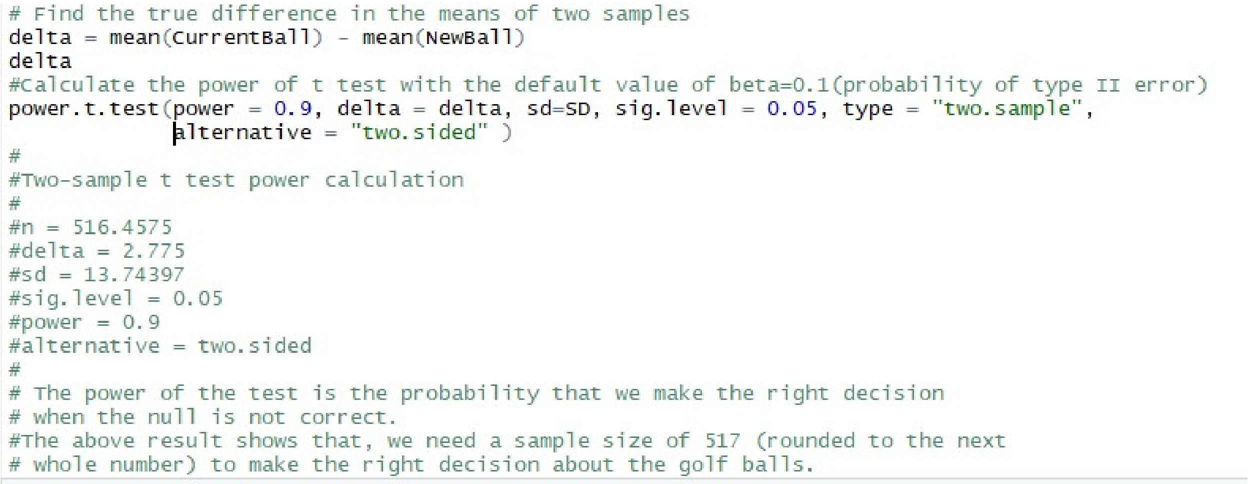
```
#=============================================================================
# Data Analysis - Golf
#=============================================================================
#Environment Set up and Data Import
#Set up working Directory
setwd("C:/Users/Radhika/Desktop/R Programming/Project_Module1")
getwd()
#
#Read the input file
golf = read.csv("Golf.csv")
attach(golf)
#Find the internal structure of the data
str(golf)
#Find the descriptive statistics of the data
summary(golf)
#Test if the two samples are normally distributed
#Draw a box plot for comparing Current Ball and New Ball samples
boxplot(CurrentBall, NewBall, horizontal = TRUE,names = c("Current Ball", "New Ball"),
        main="Comparative box plot of New Ball and Current ball Data")
# As per the box plot, both the samples are independent of each other
# As per the box plot, the mean and median are almost same for both the groups
# Hence, we can assume that each group data is normally distributed
# Draw histogram for each group to visualize the distribution

# Draw historgram for Current Ball group
hist(CurrentBall)
# Draw historgram for New Ball group
hist(NewBall)
#As per the histograms, data seems to be normally distributed

#Check if the samples have same variance using F test
# Use var.test function to check the equality of variances.
var.test(CurrentBall, NewBall, alternative = "two.sided")
# Null Hypothesis, H0 = Variance of Current Ball = Variance of New Ball
# Alternate Hypothesis Ha = Variance of Current Ball != Variance of New Ball
#data:  CurrentBall and NewBall
```



```
#F = 0.78219, num df = 39, denom df = 39, p-value = 0.4465
#alternative hypothesis: true ratio of variances is not equal to 1
#95 percent confidence interval:
#  0.413701 1.478906
#sample estimates:
#  ratio of variances
#0.7821924
# Since p-Value of 0.4465 is > 0.05, we fail to reject the null hypothesis
# Hence both the variances are equal.
#
#The assumptions of 2 sample t test are:
# a) The data continuous
# b) The data follow the normal distribution
# c) The variances of two samples are equal
# d) The two samples are independant
# e) Both the samples are simple random samples from their respective populations
#    meaning, each individual in the population has an equal probability of being
#    selected in the sample.
#
#With the above tests, we can say that the given data met the required assumptions
#of 2-Sample t test, hence, we use 2 sample t test to compae the driving distances
#of Current and New Ball data of golf data set.
#
# Calculate the p-value using 2 sample t test to compare the driving distances of
# Current and new golf ball data in the given data set
# Null Hypothesis H0= Mu New - Mu Curr = 0
# Alternate Hypothesis Ha = Mu New - Mu Curr != 0
t.test(CurrentBall, NewBall, alternative = "two.sided", mu=0, paired=FALSE,
       var.equal=TRUE, conf.level = 0.95)
```



```
# Result:
#data:  CurrentBall and NewBall
#t = 1.3284, df = 78, p-value = 0.1879
#alternative hypothesis: true difference in means is not equal to 0
#95 percent confidence interval:
# -1.383958  6.933958
#sample estimates:
# mean of x mean of y
#270.275   267.500
#
#Since p-value 0.1879 > 0.05, we fail to reject the null hypothesis
# Hence, we conclude that, there is no considerable difference between the two
# means of data to prove that the new coating has effect on the driving distances
# of golf balls.
#
#95 percent confidence interval for the difference between the means of the
#population is: -1.383958  6.933958
#
#Calculate the 95% confidence Interval for the population of Current Ball
t.test(CurrentBall)
#95 percent confidence interval:267.4757  273.0743
#
#Calculate the 95% confidence Interval for the population of New Ball
t.test(NewBall)
#95 percent confidence interval:264.3348  270.6652
#
#Calculate the Standard Deviation of Current Ball Data
sd(CurrentBall)
#Calculate the Standard Deviation of New Ball Data
sd(NewBall)
#Find the Standard Deviation of difference of the two means of the data
SD=sd(NewBall - CurrentBall)
SD
```



```
# Find the true difference in the means of two samples
delta = mean(CurrentBall) - mean(NewBall)
delta
#Calculate the power of t test with the default value of beta=0.1(probability of type II error)
power.t.test(power = 0.9, delta = delta, sd=SD, sig.level = 0.05, type = "two.sample",
             alternative = "two.sided" )
#
#Two-sample t test power calculation
#
#n = 516.4575
#delta = 2.775
#sd = 13.74397
#sig.level = 0.05
#power = 0.9
#alternative = two.sided
#
# The power of the test is the probability that we make the right decision
# when the null is not correct.
#The above result shows that, we need a sample size of 517 (rounded to the next
# whole number) to make the right decision about the golf balls.
```