# Fundamentals of Business Statistics

## Par Inc. Golf Ball Problem

## Project 2

### PGP - DSBA

KUDAKWASHE NYIKADZINO

# Problem Statement

Par Inc., is a major manufacturer of golf equipment. Management believes that Par's market share could be increased with the introduction of a cut-resistant, longer-lasting golf ball. Therefore, the research group at Par has been investigating a new golf ball coating designed to resist cuts and provide a more durable ball. The tests with the coating have been promising. One of the researchers voiced concern about the effect of the new coating on driving distances. Par would like the new cut-resistant ball to offer driving distances comparable to those of the current-model golf ball. To compare the driving distances for the two balls, 40 balls of both the new and current models were subjected to distance tests. The testing was performed with a mechanical hitting machine so that any difference between the mean distances for the two models of the ball could be attributed to a difference in the design.

The results of the tests, with distances measured to the nearest yard, are contained in the data set "Golf".

# Questions to be answered

**1.** Formulate and present the rationale for a hypothesis test that Par could use to compare the driving distances of the current and new golf balls.

**2.** Analyze the data to provide the hypothesis testing conclusion. Whether the test used to validate the hypothesis is one tail or two-tail? What is the p-value for your test? What is your recommendation for Par Inc.? Based on the observations whether t-test will be applicable or z test?

**3.** Provide observations from the driving distances of the 2 balls (descriptive statistics)

**4.** What is the 95% confidence interval for the population mean of each model, and what is the 95% confidence interval for the difference between the means of the two population?

**5.** Do you see a need for larger sample sizes and more testing with the golf balls? Discuss (This is an optional question, this will be not be counted towards the final grading, requires extra self-study on *Power of a Test*)

# Findings and Recommendations

1. The current golf ball model shows an average driving distance larger than the new golf ball. This is not statistically significant, and Par Inc. should proceed with the introduction of the new model into the market.
2. The two models have the same maximum driving distance of 289 yards which means they can equally compete on this factor.
3. 50% of the observations of driving distances between the two model fall in an almost similar range.
4. The new model has more variance in driving distances compare to the current model. The new coating could still undergo further tests to help manage the larger variance.
5. Most of the measure of central tendency and measures of dispersion favour the current model compared to the new model though the test confirmed there is no statistical difference between the average driving distances of the two golf ball models. I would then recommend determining our probability of making the correct decision to reject the null hypothesis when it is false (power of the test)
6. I further recommend that other factors be considered besides the average driving distance for example the cost of adding and manufacturing the new coat and the impact this will have on the ability of Par Inc. to get the market share with the new golf ball model.

**Question 1**. Formulate and present the rationale for a hypothesis test that Par could use to compare the driving distances of the current and new golf balls

## 1.1) Considerations for formulating a hypothesis test

Our objective for this test is to determine if adding the new golf ball coating will change the average driving distance of the golf balls. With the new coating, our aim is to make sure the driving distances of golf balls with the new coating are comparable to those of the current golf balls. To make sure we do not have unwanted bias in the outcome of the test, we assume all golf balls with and without the new coating are selected at random. At the same time a machine is used to negate the effect a person may have on the driving distances.

In determining whether there are differences between the types of golf balls, we look at whether their average differences are statistically significant to warrant a conclusion that there is a real difference between their driving distances. We develop two hypotheses, firstly a null hypothesis which is the status quo. This means we believe there is no difference between the average driving distances between golf balls with and those without the new coating.

The concern that was raised i.e. the new coating could affect driving distances, becomes the basis of the other hypothesis or alternative hypothesis.

If the outcome of the test requires us to reject the null hypothesis, then there is no evidence that the average driving distances between the two golf ball models are different.

## 1.2) Factors used to provide a rationale for appropriate hypothesis test

There are factors that can be used in combination or in some instance in isolation when determining the appropriate hypothesis test. Some of these factors are as follows:

### 1.2.1) Availability of population standard deviation

We do not have any information on the population standard deviation which means we are using sample statistics in this case the standard error. It is not always practical or feasible to obtain population parameters like the standard deviation which is why we use the sample statistics. Based on the above question, this information means a t-test is appropriate.

### 1.2.2) Sample size

The size of the two samples is N = 40 each. Since this considered a large sample, a Z-test can be considered for testing the hypothesis. However, sample size alone is not a definitive factor in determining the applicable hypothesis test. The other factors below will help clarify the appropriate test.

### 1.2.3) Independent vs Paired Samples

Independence of sample observations determines the appropriate t-test that can be used. Our two models of golf balls are independent of each other and are not paired. If they were paired, the sample observations would have been dependent on each other. This would have been the case if for example the current golf balls were then coated and their driving distances before and after the coatings were recorded.

### 1.2.4) Level of Significance

For us to reach a conclusion for the test, we use a level of significance of 5% which indicates the risk we are comfortable taking in case we conclude that there is a difference in the average driving distances when there is none.

## 1.3) Conclusions expected from the test

If we fail to reject the null hypothesis, we conclude there is no evidence that there is a difference in the average driving distance between current model golf balls and golf balls with the new coating.

If we reject the null hypothesis, we conclude that there is evidence that there is a difference in the average driving distance between current model golf balls and golf balls with the new coating

**Question 2**. Analyze the data to provide the hypothesis testing conclusion. Whether the test used to validate the hypothesis is one tail or two-tail? What is the p-value for your test? What is your recommendation for Par Inc.? Based on the observations whether t-test will be applicable or z test?

## 2.1) Selecting the appropriate hypothesis test

### 2.1.1) One-Tailed vs Two-Tailed Test

We need to determine whether the test is a one-tailed or two-tailed. Our problem states that we need to ensure the mean driving distances between the current golf ball model and the golf balls with the new coating are comparable. This means there should not be a significant difference between the mean driving distances. Since the mean driving distances should be comparable, it means the differences of the new golf ball model should not be significantly higher or significantly less than those of the current golf ball model. This means a two-tailed hypothesis test is appropriate.

### 2.1.2) Number of samples

The number of samples being tested also provide a guide to the appropriate hypothesis test to implement. Our case has two samples namely the current golf ball model and the new golf ball model with the cut-resistant and durable coating. Testing the differences between two groups means both the t-test and the z-test can be used. We further narrow down the most appropriate of the tests below.

## 2.2) Hypothesis formulation

Based on the above, the appropriate test is a **Two-Tailed Independent Samples T-Test**.

## 2.3) Assumptions for the Independent Samples t-test

### 2.3.1) Normality

We assume the data are normally distributed, particularly that the average driving distances for both golf ball models are normally distributed. Since the sample size is large i.e. N = 40 which is greater than 30, the t-distribution approximates a normal distribution. Knowing the shape of the distribution when we fail to reject the null hypothesis, helps us in determining how all possible mean differences behave by applying the central limit theorem.

### 2.3.2) Independence

Observations from the two samples are randomly selected helping ensure that golf ball mean driving distances are independent of each other.

### 2.3.3) Homogeneity of Variance (Homoscedasticity)

Variances of the two samples are assumed equal. This may not occur that often, so the approach is to use the Welch Two Sample T-Test which assumes variances are not equal. This is the default test used in R and for this problem.

## 2.4) Solution

- Sample size – **N = 40**
- Degrees of Freedom – **N – 1** which is **39 degrees of freedom**
- $\mu_{\text{current}}$ – mean driving distance of the current golf ball model
- $\mu_{\text{new}}$ – mean driving distance of the golf ball model with the new coat

The hypotheses can be stated as follows:

**$H_o$ - Null Hypothesis** (there is no difference in the average driving distance between current model golf balls and golf balls with the new coating)

**$H_a$** – Alternative Hypothesis (there is a difference in the average driving distance between current model golf balls and golf balls with the new coating)

**$H_o$** : $\mu_{\text{current}} - \mu_{\text{new}} = 0$ (the average driving distances are the same)

**$H_a$** : $\mu_{\text{current}} - \mu_{\text{new}} \neq 0$ (the average driving distances are not the same)

### 2.4.1) Output

***Fig 2.4 a***

```
        Welch Two Sample t-test

data:  Current and New
t = 1, df = 80, p-value = 0.2
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.38  6.93
sample estimates:
mean of x mean of y
      270       268
```

### 2.4.2) p-value

the p-value of the test is 0.2 (*Fig 2.4 a above*) which is larger than the level of significance we set initially at 0.05. This far exceeds the risk we are comfortable with in case when we incorrectly reject our null hypothesis when it's true.

### 2.4.3) Conclusion and Recommendation

We fail to reject the null hypothesis and conclude that there is no evidence there is a statistical difference between the driving distances of the two golf ball models. Par Inc. should continue with the introduction of the cut-resistant, longer-lasting golf ball to the market based on this test.

**Question 3**. Provide observations from the driving distances of the 2 balls (descriptive statistics)

## 3.1) Variable Identification

The driving distances data set has two numeric variable each with 40 observation. The data is contained in a tibble which is a subclass of a data frame. Chart 3.1 a below shows this information.

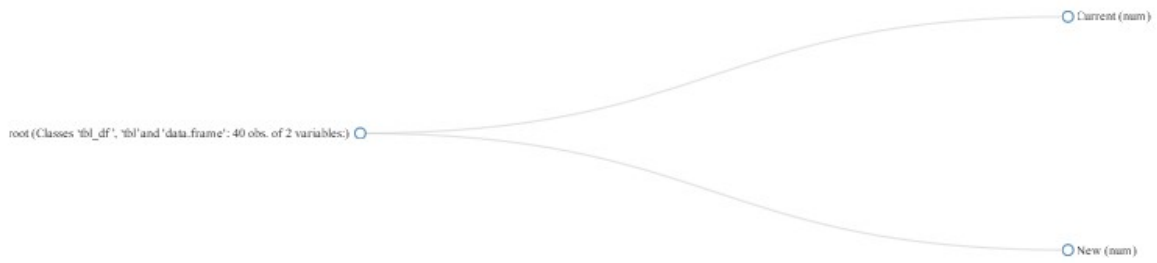



*Chart 3.1 a*

## 3.2) Missing Variable Identification

The data contains no missing observations and shown in Chart 3.2 a below. The two variables in addition to being numeric, they are also continuous variables. There are also no missing rows as confirmed by Chart 3.2 b below.

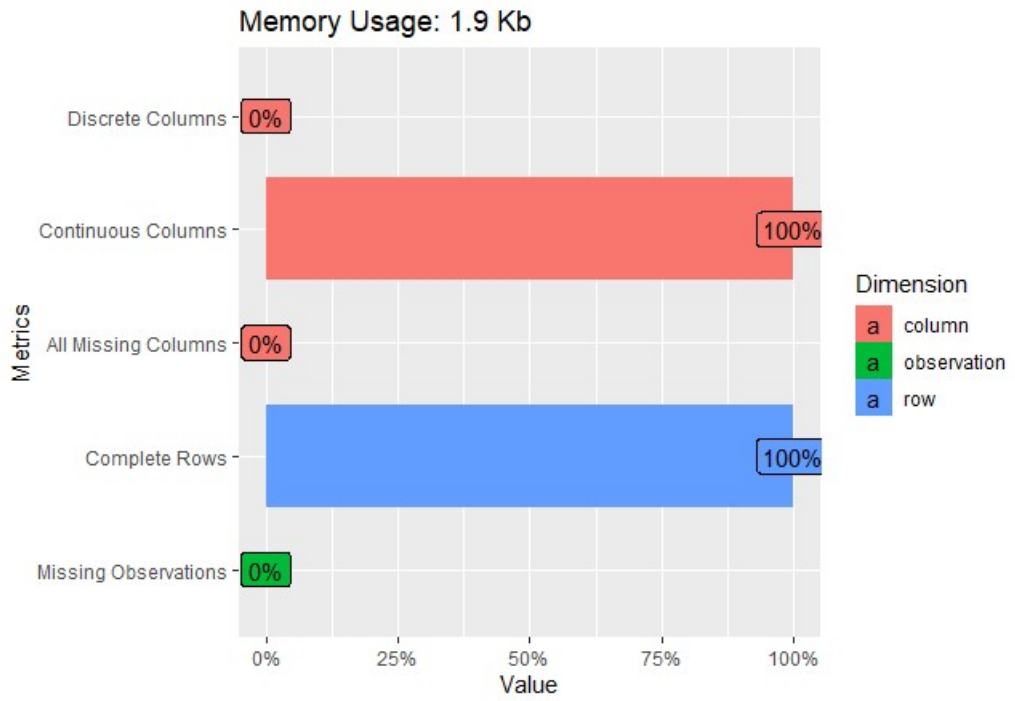



*Chart 3.2 a*

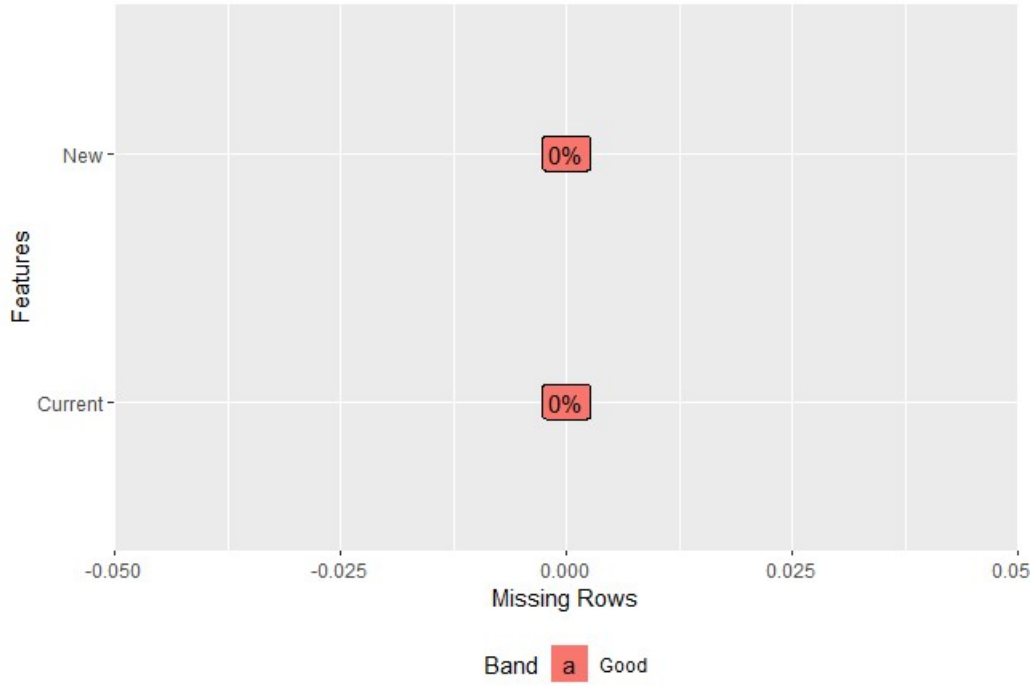



*Chart 3.2 b*

## 3.3) Variable Analysis

From the Table 3.3.1 a below, the average driving distance for the current golf ball model is about 270 yards and for the new golf ball model about 2 yards less at about 268 yards. The median value of the driving distance for new golf ball model is also less than that of the current golf ball model. The median driving distances are 265 yards and 270 yards respectively. The minimum driving distance of the new golf ball model is less than the current golf ball by 5 yards. The interesting part is that the maximum distances for the two golf ball models are the same at 289 yards. From the maximum driving distance alone, it seems that there is potentially no difference between the two models.

| | Current | New |
|---|---|---|
| nbr.val | 40.0000 | 40.000 |
| nbr.null | 0.0000 | 0.000 |
| nbr.na | 0.0000 | 0.000 |
| min | 255.0000 | 250.000 |
| max | 289.0000 | 289.000 |
| range | 34.0000 | 39.000 |
| sum | 10811.0000 | 10700.000 |
| median | 270.0000 | 265.000 |
| mean | 270.2750 | 267.500 |
| SE.mean | 1.3840 | 1.565 |
| CI.mean.0.95 | 2.7993 | 3.165 |
| var | 76.6147 | 97.949 |
| std.dev | 8.7530 | 9.897 |
| coef.var | 0.0324 | 0.037 |

*Table 3.3.1 a*

Both golf ball models show distributions that are right skewed though the current model show a slightly larger skewness than the new model (*below Chart 3.1.1 b*). The right whisker of the current model's box plot also is longer than the left which confirms the skewness. The new model has whiskers that are of almost equal length. Using this information from the boxplots, we confirm the larger skewness exhibited by the current model relative to the new model. The boxplots also show that 50% of the driving distance observations for the current model and the new model are 263 – 275 yards and 262 – 274 yards respectively. From this observation, there also isn't much of a difference between the two golf ball models.

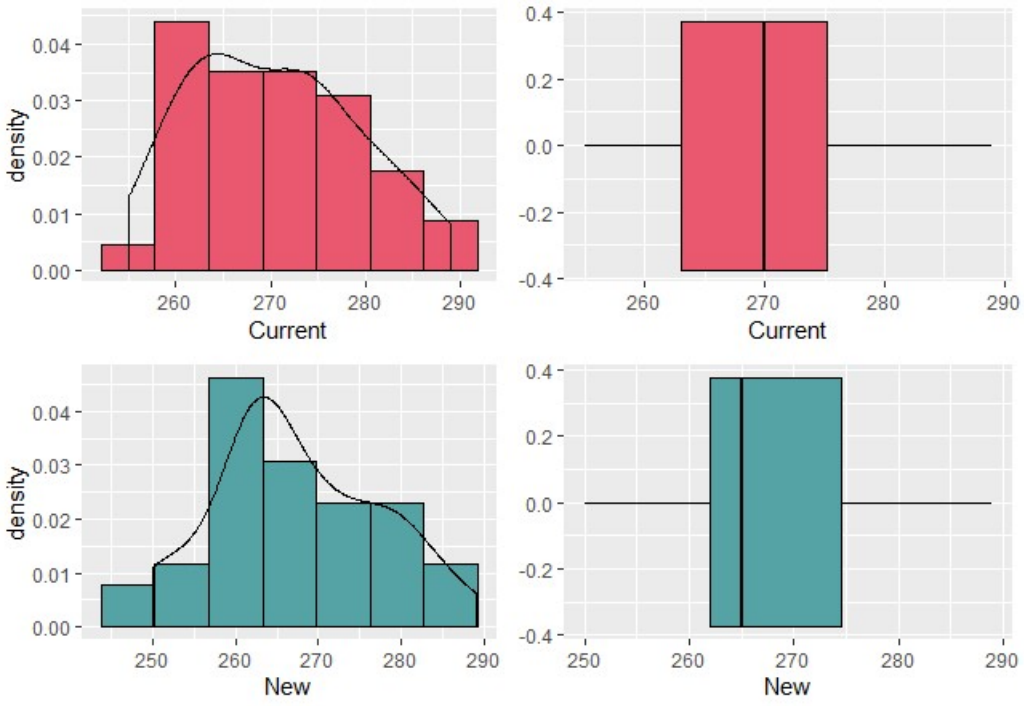



*Chart 3.1.1 b*

The density plots in Chart 3.1.1 b show distributions that are not too far from the normal distribution. The Q-Q plots in Chart 3.1.1 c also show that observations are close to the diagonal line which means from this we can assume the normal distribution assumption holds.

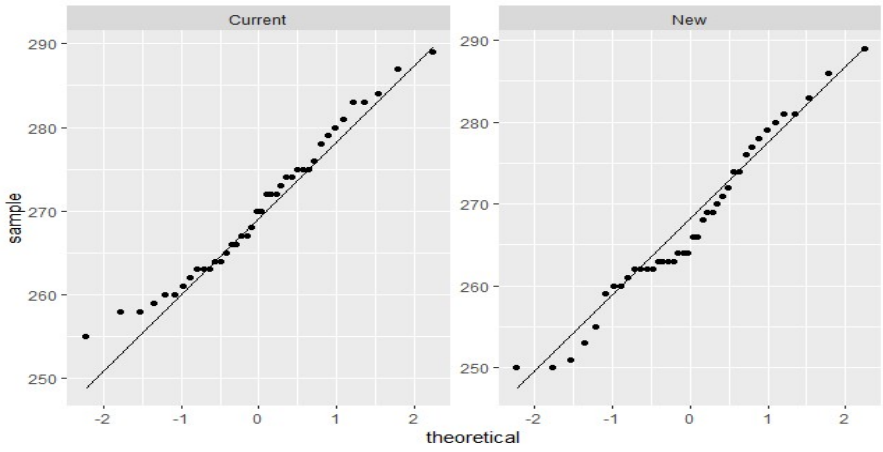



*Chart 3.1.1 c*

**Question 4**. What is the 95% confidence interval for the population mean of each model, and what is the 95% confidence interval for the difference between the means of the two population?

## 4.1) Confidence Intervals

### 4.1.1) Confidence Intervals for each model

The confidence intervals for each model at 95% confidence are given in the table below:

| | **Current Model** | **New Model** |
|---|---|---|
| **Lower CI** | 267 | 264 |
| **Upper CI** | 273 | 271 |

The lower and upper bounds of the confidence intervals for the two population means of each golf models indicate that we have 95% confidence that the true driving distance mean lies between the lower and upper bounds. For the current model, the true driving distance mean lies between 267 and 273 yards and for the new model it lies between 264 and 271 yards.

### 4.1.2) Confidence interval for the difference between means of the two populations

The confidence intervals for the differences in means between the two golf ball models is shown below:

| | **Differences between means** |
|---|---|
| **Lower CI** | -1.38 |
| **Upper CI** | 6.93 |

This means that the maximum difference between the two golf ball models ranges between -1.38 and 6.93 yards with a confidence level of 95%.

**Question 5.** Do you see a need for larger sample sizes and more testing with the golf balls? Discuss (This is an optional question, this will be not be counted towards the final grading, requires extra self-study on *Power of a Test*)

There is a close relationship between the following:

- Effect size
- Sample size
- Level of significance - P(Type I error)
- Power of a test – P(1 – Type II error)
-

When we have three of the above, we can determine the fourth. So, we can determine if a larger sample size of golf balls is required based on the other three factors.

Let's assume the following based on the output given in the code in Annexure A:

- Type I error is the same as our level of significance which is 5%.
- Null hypothesis is $\mathbf{H_o}$ : $\mu_{\text{current}} - \mu_{\text{new}} = 0$ (the average driving distances are the same)
- Alternative Hypothesis is $\mathbf{H_a}$ : $\mu_{\text{current}} - \mu_{\text{new}} = 2.77$ (the average difference of the driving distances between the models)
- Two tailed test
- T Critical values are -1.99 or 1.99
- Degrees of freedom is 80

Calculation of TStat:

- $tstat=(\bar{d}- \mu D)/Sd/\sqrt{n}$
- -/+1.99 = $(\bar{d}$ - 0)/(13.7)/ √40
- $\bar{d}$ = -1.99 * 2.17 and +1.99*2.17
- $\bar{d}$ = +/-4.31

Given

$\mu_{\text{current}} - \mu_{\text{new}}$ = 2.77

- $tstat$=(-4.31− 2.77)/ 2.17 or (4.31− 2.77)/ 2.17
- $tstat$=-3.26 and 0.71

Therefore β = 0.759 hence the probability of making a **Type II error is 75.9%** and the power of the test is **1 – 0.759 = 0.241 or 24.1%**

If we want the Type I and Type II errors to be equal at 0.05 then our sample size should be 637 for each golf ball model. The conclusion is that for the power of the test increase from 24.1% we need a larger sample size. If we are to match the Type I and Type II errors, we will need large sample size. From Chart 5 a below, the density plots overlap by a large area which makes it difficult to differentiate between the null and alternative hypothesis. This means there is weak power of the test. Increasing the sample sizes will assist with increasing the power of the test as seen above.

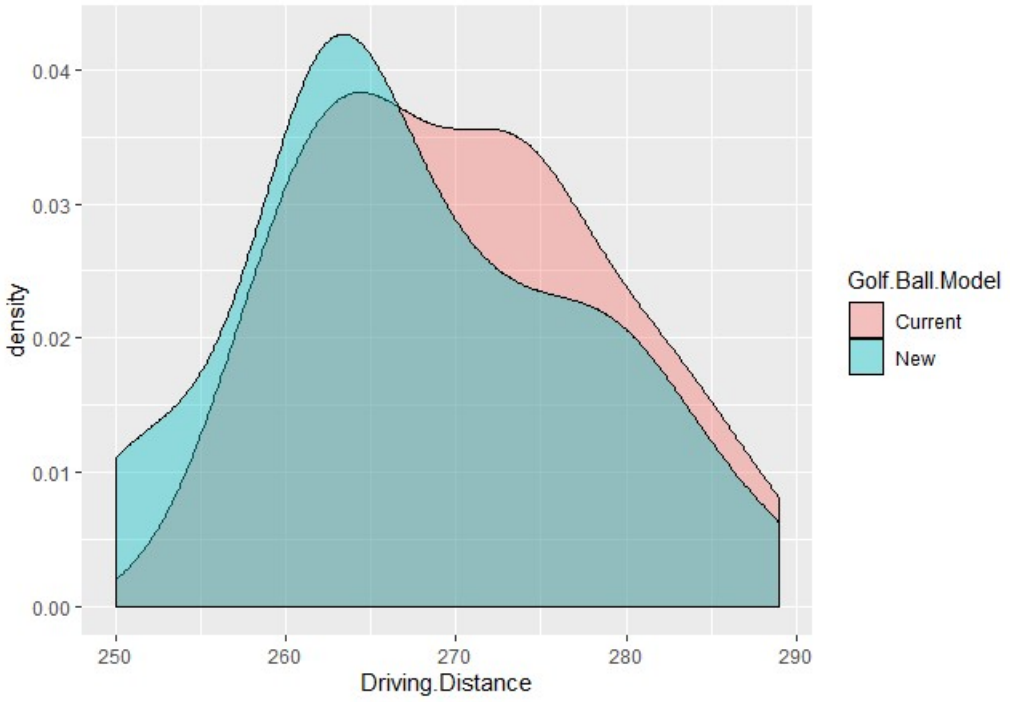

0.04
0.03
0.02
0.01
0.00
density
250
260
270
280
290
Driving.Distance
Golf.Ball.Model
Current
New

*Chart 5 a*

## Annexure A

# Golf_Ball_Project

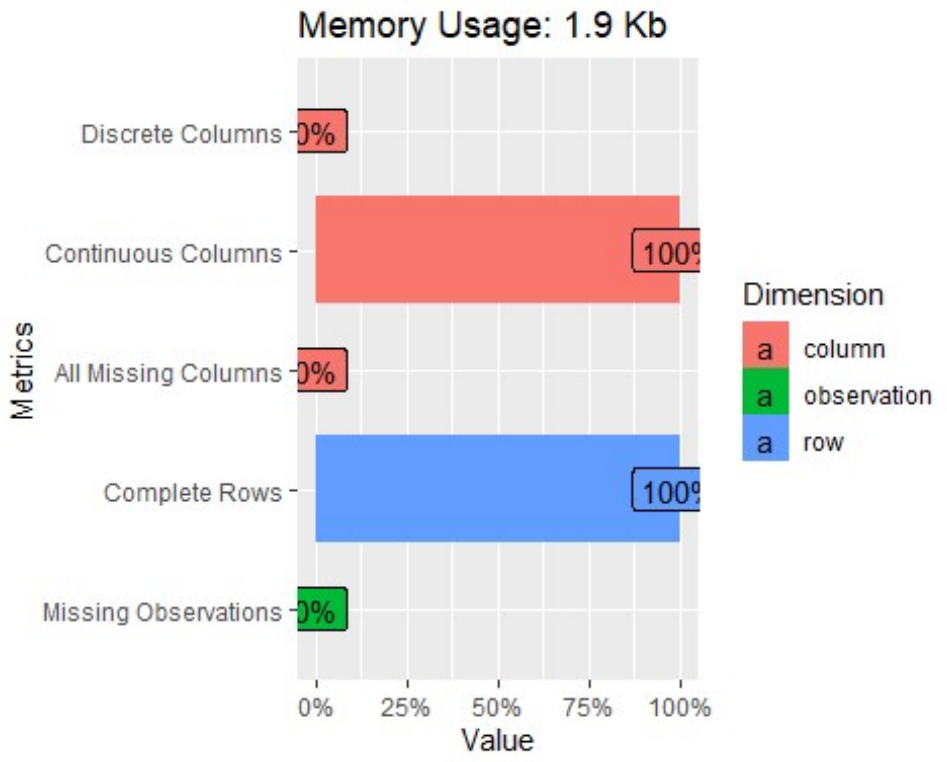
Kudakwashe Nyikadzino

03 August 2019

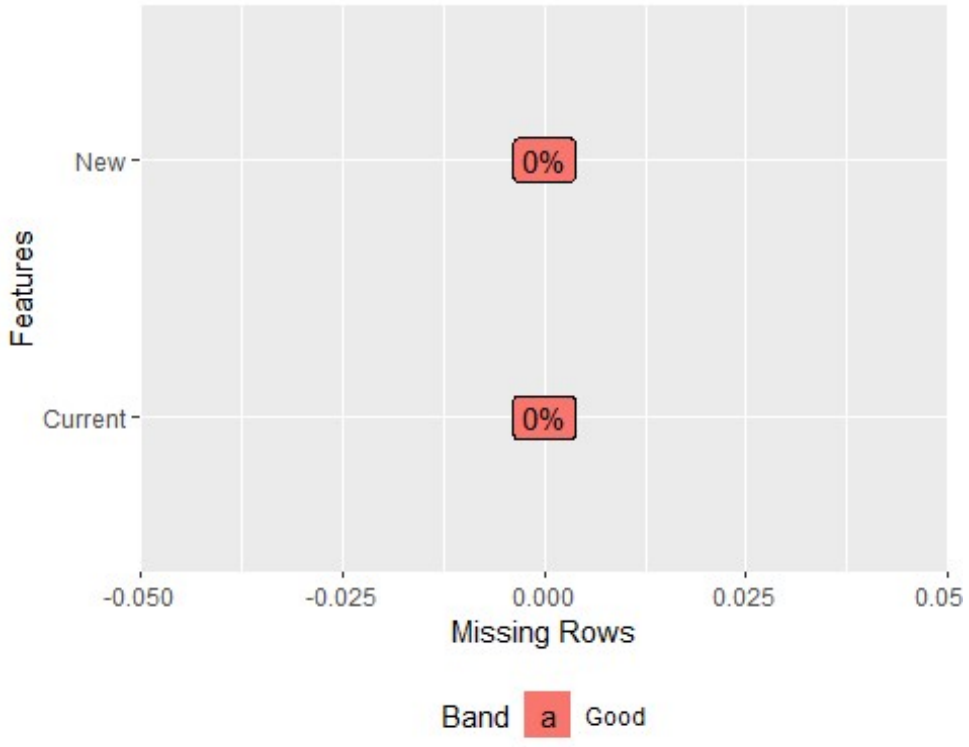
```
#Load libraries
library(readxl)

## Warning: package 'readxl' was built under R version 3.5.3

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.5.3

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.5.3

library(tidyr)

## Warning: package 'tidyr' was built under R version 3.5.3

library(DataExplorer)

## Warning: package 'DataExplorer' was built under R version 3.5.3

library(gridExtra)

## Warning: package 'gridExtra' was built under R version 3.5.3

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine

library(pastecs)

## Warning: package 'pastecs' was built under R version 3.5.3
```

```
##
## Attaching package: 'pastecs'

## The following object is masked from 'package:tidyr':
##
##     extract

## The following objects are masked from 'package:dplyr':
##
##     first, last

library(formattable)

## Warning: package 'formattable' was built under R version 3.5.3

library(moments)

## Warning: package 'moments' was built under R version 3.5.2

#Set working directory
setwd("C:/Users/kudaakwashe/Documents/Study/PGPDSBA/Fundamentals_of_Busine
ss_Statistics/Project")

golf_data = read_xls("Golf.xls")

attach(golf_data)

#Understanding our data set
head(golf_data)

## # A tibble: 6 x 2
##   Current   New
##     <dbl> <dbl>
## 1     264   277
## 2     261   269
## 3     267   263
## 4     272   266
## 5     258   262
## 6     283   251

str(golf_data)

## Classes 'tbl_df', 'tbl' and 'data.frame':    40 obs. of  2 variables:
##  $ Current: num  264 261 267 272 258 283 258 266 259 270 ...
##  $ New    : num  277 269 263 266 262 251 262 289 286 264 ...

summary(golf_data)

##     Current           New
##  Min.   :255.0   Min.   :250.0
##  1st Qu.:263.0   1st Qu.:262.0
##  Median :270.0   Median :265.0
##  Mean   :270.3   Mean   :267.5
##  3rd Qu.:275.2   3rd Qu.:274.5
##  Max.   :289.0   Max.   :289.0
```

```
plot_str(golf_data)
plot_intro(golf_data)
```




```
plot_missing(golf_data)
```




```
#remove scinetific notation
options(scipen=999, digits = 3)
```

```r
#Variable Analysis
stat_table = stat.desc(golf_data)
formattable(stat_table)
```

Current
New
nbr.val
40.0000
40.000
nbr.null
0.0000
0.000
nbr.na
0.0000
0.000
min
255.0000
250.000
max
289.0000
289.000
range
34.0000
39.000
sum
10811.0000
10700.000
median
270.0000
265.000
mean
270.2750
267.500
SE.mean
1.3840
1.565
CI.mean.0.95
2.7993
3.165
var
76.6147
97.949
std.dev

8.7530

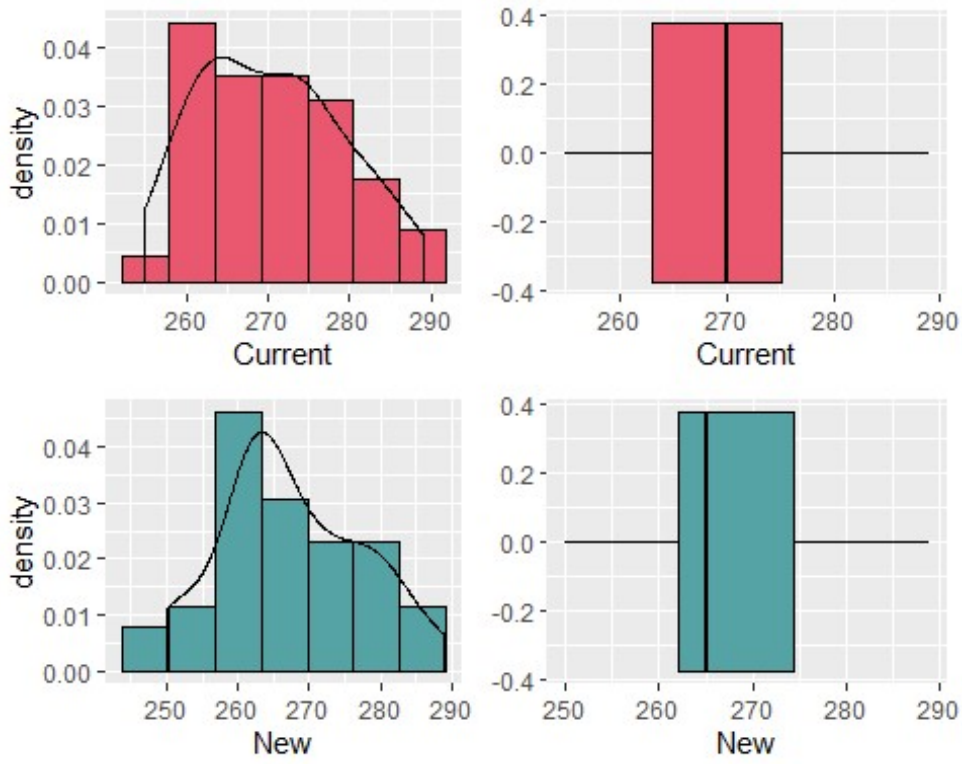
9.897

coef.var

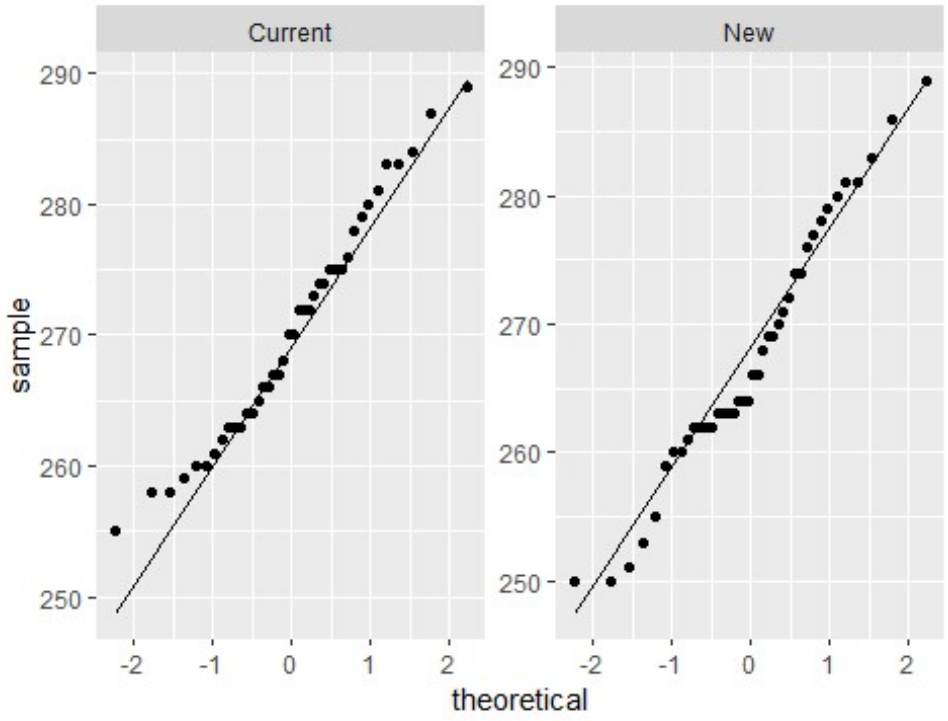
0.0324

0.037

```
#Plots Histogram, Box Plots and Density Plots
c.Hist = ggplot(golf_data, aes(Current)) +
         geom_histogram(aes(y = ..density..), col = "black", fill = "#EA58
6F",
                        position = "identity", bins = 7) +
         geom_density()

c.Box =  ggplot(golf_data) +
         geom_boxplot(aes(y = Current), color = "black", fill = "#EA586F")
+
         coord_flip()

n.Hist = ggplot(golf_data, aes(New)) +
         geom_histogram(aes(y = ..density..), col = "black", fill = "#56A3
A6",
                        position = "identity", bins = 7) +
         geom_density()

n.Box =  ggplot(golf_data) +
         geom_boxplot(aes(y = New), color = "black", fill = "#56A3A6") +
         coord_flip()

#graph layout
grid.arrange(c.Hist, c.Box, n.Hist, n.Box, nrow = 2)
```



```r
#Check normality
plot_qq(golf_data)
```




```r
skewness(Current)
```

```
## [1] 0.295
```

```r
skewness(New)
```

```
## [1] 0.231
```

```r
#t-tests
t.test(Current)
```

```
##
##  One Sample t-test
##
## data:  Current
## t = 200, df = 40, p-value <0.0000000000000002
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  267 273
## sample estimates:
## mean of x
##       270
```

```r
t.test(New)
```

```
##
##  One Sample t-test
##
## data:  New
```

```
## t = 200, df = 40, p-value <0.0000000000000002
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  264 271
## sample estimates:
## mean of x
##       268
```

```
t.test(Current, New,
       paired = FALSE,
       conf.level = 0.95,
       alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  Current and New
## t = 1, df = 80, p-value = 0.2
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.38  6.93
## sample estimates:
## mean of x mean of y
##       270       268
```

```
#Question 5
sd(Current - New)
```

```
## [1] 13.7
```

```
mean(Current - New)
```

```
## [1] 2.77
```

```
qt(0.025, 80)
```

```
## [1] -1.99
```

```
#Beta
pt(0.71,80, lower.tail = TRUE)-pt(-3.26,80, lower.tail = TRUE)
```

```
## [1] 0.759
```

```
#finding sample size
power.t.test(power = 0.95,
             delta = 2.77,
             sd = 13.7,
             sig.level = 0.05,
             type = "two.sample",
             alternative = "two.sided")
```

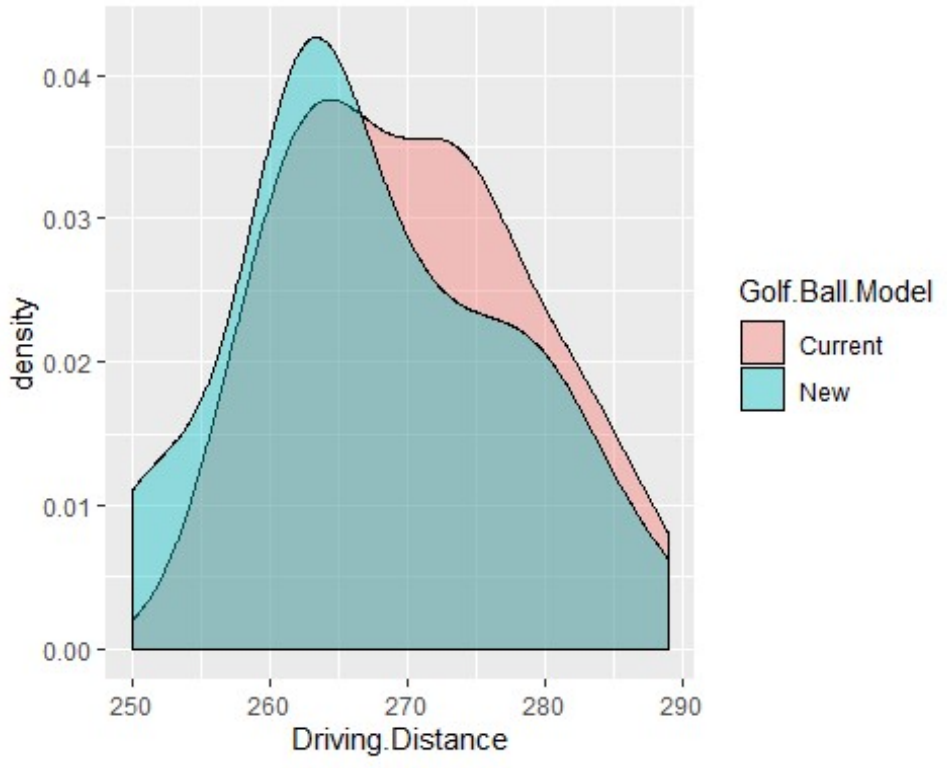
```
##
##      Two-sample t test power calculation
##
##               n = 637
##           delta = 2.77
```

```
##              sd = 13.7
##       sig.level = 0.05
##           power = 0.95
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
#Reshaping data set for easier visualisation
golf_data_vis = golf_data %>% gather(Golf.Ball.Model, Driving.Distance, Cu
rrent:New)
head(golf_data_vis)
```

```
## # A tibble: 6 x 2
##   Golf.Ball.Model Driving.Distance
##   <chr>                      <dbl>
## 1 Current                      264
## 2 Current                      261
## 3 Current                      267
## 4 Current                      272
## 5 Current                      258
## 6 Current                      283
```

```
#Plotting density plots
ggplot(golf_data_vis, aes(x = Driving.Distance)) + geom_density(aes(fill =
Golf.Ball.Model), alpha = 0.4)
```




Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.