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 could be attributed to a difference in the design.

Prepare a Managerial Report

Questions:

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. What is the p-value for your test? What is your recommendation for Par Inc.?
3. Do you see a need for larger sample sizes and more testing with the golf balls? Discuss

Managerial Report

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 the data of the Current and New distance of the golf ball are provided in the table provided below

Sl No	Current	New
1	264	277
2	261	269
3	267	263
4	272	266
5	258	262
6	283	251
7	258	262
8	266	289
9	259	286
10	270	264
11	263	274
12	264	266
13	284	262
14	263	271
15	260	260
16	283	281
17	255	250
18	272	263
19	266	278
20	268	264
21	270	272
22	287	259
23	289	264
24	280	280
25	272	274
26	275	281
27	265	276
28	260	269
29	278	268
30	275	262
31	281	283
32	274	250
33	273	253
34	263	260
35	275	270
36	267	263
37	279	261
38	274	255
39	276	263
40	262	279

Current: contains driving distances of golf balls without coating.
New: contains driving distances of golf balls with new coating.

Steps followed to arrive at the report

Importing necessary libraries

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
from scipy.stats import ttest_1samp,wilcoxon,mannwhitneyu,levene,shapiro,ttest_ind,iqr
from statsmodels.stats.power import ttest_power
import scipy.stats as stats
import math
```

Reading the file in python

Step 2 : Reading the SM4-Golf data (convert into CSV format)

```
In [2]: df= pd.read_csv('SM4-Golf.csv')
```

	Current	New
0	264	277
1	261	269
2	267	263
3	272	266
4	258	262
5	283	251
6	258	262
7	266	289
8	259	286
9	270	264
10	263	274
11	264	266
12	284	262
13	263	271
14	260	260
15	283	281
16	255	250
17	272	263
18	266	278
19	268	264
20	270	272
21	287	259
22	289	264
23	280	280
24	272	274
25	275	281
26	265	276
27	260	269
28	278	268
29	275	262
30	281	283
31	274	250
32	273	253
33	263	260
34	275	270
35	267	263
36	279	261
37	274	255
38	276	263
39	262	279

Observation: - There are 40 reading for of both the New and Current models

checking the information of headings

```
In [4]: df.head()
```

Out[4]:

	Current	New
0	264	277
1	261	269
2	267	263
3	272	266
4	258	262

Observation: - The reading starts with Sl No “0”

checking the information of last few reading (Tail)

```
In [5]: df.tail()
```

Out[5]:

	Current	New
35	267	263
36	279	261
37	274	255
38	276	263
39	262	279

checking the information of the data sets

```
In [6]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40 entries, 0 to 39
Data columns (total 2 columns):
Current    40 non-null int64
New        40 non-null int64
dtypes: int64(2)
memory usage: 720.0 bytes
```

Observation: - There are no null values in the dataset

checking the descriptive statistics of the both the population

```
In [7]: df.describe()
```

Out[7]:

	Current	New
count	40.000000	40.000000
mean	270.275000	267.500000
std	8.752985	9.896904
min	255.000000	250.000000
25%	263.000000	262.000000
50%	270.000000	265.000000
75%	275.250000	274.500000
max	289.000000	289.000000

- The Mean_Current = 270.275 and Mean_New =267.5
- Median values Median_Values_Current = 270.0 and Median_Values_New= 265.0
- Std_Current = 8.75 and Std_New = 9.89

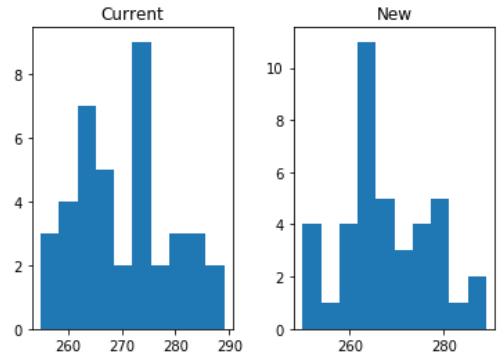
Observation:-

1) The Sample size:40 for both Current and New data set
2) Number of sample population : 2
3) These are unpaired variables.
4) There are no missing values.
5) The Mean_Current = 270.275 and Mean_New =267.5 and Median values Median_Values_Current = 270.0 and Median_Values_New= 265.0 of the two populations are not much different.
6) There is dip in the performance in Mean, Min, Median and Max Value for New golf balls as compared to Current golf balls data set.
7) The standard deviation for New golf balls **(9.89)** is much more than Current golf ball data set **(8.75).**

checking the histogram of both the population (Current and New)

```
df.hist(bins=10,grid=False)
```

```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x0000027F5DCA1940>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000027F5DEECEF0>]], dtype=object)
```

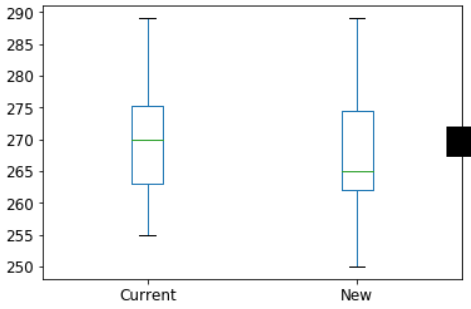


Observation:-

- The mean distance for Current data set = 270.27 , Min distance = 255 , Max distance value = 289.0 and stdev = 8.75
- The Mean distance for New data set = 267.50,Min distance= 250 , Max distance value= 289.0 and stdev = 9.89 ,

checking for outliers using Box plot

```
In [9]: df.boxplot(grid=False,fontsize=12)
Out[9]: <matplotlib.axes._subplots.AxesSubplot at 0x2323cdc3d68>
```



Inference:
There are **no outliers** in both the data for the Current and New golf balls .

Since the same test of driving distance is applied on two populations of Current and the new golf ball, the samples are classified under unpaired samples

Now we need to split the data set -Split the data into two samples of Current and New golf balls . **First we do for Current golf ball data set**

```
In [10]: Current = df.iloc[:,0]

In [11]: Current
Out[11]: 0     264
         1     261
         2     267
         3     272
         4     258
         5     283
         6     258
         7     266
         8     259
         9     270
         10    263
         11    264
         12    284
         13    263
         14    260
         15    283
         16    255
         17    272
         18    266
         19    268
         20    270
         21    287
         22    289
         23    280
         24    272
         25    275
         26    265
         27    260
         28    278
         29    275
         30    281
         31    274
         32    273
         33    263
         34    275
         35    267
         36    279
         37    274
         38    276
         39    262
         Name: Current, dtype: int64
```

Observation:-The data set for Current golf ball and # of data points =40

Check for interquartile range (IQR) for Current golf data set.

```
In [12]: iqr(Current, rng = (25,75))

Out[12]: 12.25
```

Observation : The interquartile range (IQR) is the difference between the 75th and 25th percentile of the data. It is a measure of the dispersion similar to standard deviation or variance, but is much more robust against outliers . **IQR is 12.25**

Check the Mean value for Current golf data set.

```
In [13]: meanC = Current.mean()
         meanC

Out[13]: 270.275
```

Observation: The Mean value of Current golf ball data set is 270.275

Check for Variance for Current golf data set.

```
In [14]: varC = Current.var()
         varC

Out[14]: 76.61474358974361
```

Observation: The variance is a measure of how far each value in the data set is from the mean in this case its 76.61

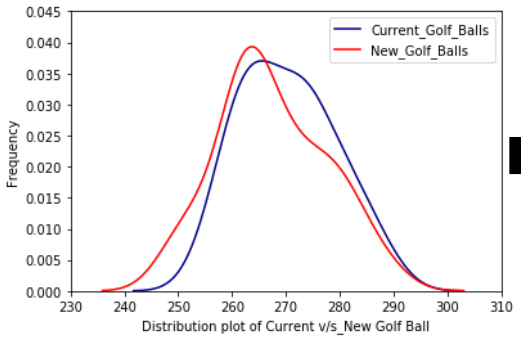
Now we need to split the data set -Split the data for New golf balls

```
In [15]: New = df.iloc[:,1]

In [16]: New

Out[16]: 0     277
         1     269
         2     263
         3     266
         4     262
         5     251
         6     262
         7     289
         8     286
         9     264
         10    274
         11    266
         12    262
         13    271
         14    260
         15    281
         16    250
         17    263
         18    278
         19    264
         20    272
         21    259
         22    264
         23    280
         24    274
         25    281
         26    276
         27    269
         28    268
         29    262
         30    283
         31    250
         32    253
         33    260
         34    270
         35    263
         36    261
         37    255
         38    263
         39    279
         Name: New, dtype: int64
```

Observation:-The data set for New golf ball and # of data points = 40

Comparing the distribution plot of Current and New distance of Golf Balls

```
In [17]: sns.distplot(Current,color='darkblue',hist=False,label='Current_Golf_Balls')
         sns.distplot(New,color='red',hist=False,label='New_Golf_Balls')
         plt.legend(loc='upper right')
         ax=plt.gca()
         ax.set_ylim([0,.045])
         ax.set_xlim([230,310])
         plt.xlabel('Distribution plot of Current v/s_New Golf Ball')
         plt.ylabel('Frequency')
         plt.show()
```



The distribution plot of Current and New distance of golf ball indicates that there is not much difference between the means of both the population

If we also look at the standard deviation (stdev) or the spread of data,the stdev of New is slightly more than the stdev of the Current golf ball data

Check the Mean value for New golf data set.

```
In [18]: meanN = New.mean()
         meanN
Out[18]: 267.5
```

Inference: The Mean value of New golf ball data set is 267.5

Check for interquartile range (IQR) for New golf data set.

```
In [19]: iqr(New, rng = (25,75))
Out[19]: 12.5
```

Observation: The interquartile range (IQR) is the difference between the 75th and 25th percentile of the data. It is a measure of the dispersion similar to standard deviation or variance, but is much more robust against outliers. **IQR is 12.5**

Check for Variance for New golf data set.

```
In [20]: varN= New.var()
         varN
Out[20]: 97.94871794871794
```

Observation The variance is a measure of how far each value in the data set is from the mean in this case its 97.94

 # Check for Normality using Shapiro-Wilk normality test for Current golf Ball data set
Testing whether samples are parametric or non parametric by Shapiro Normality Test

In the SciPy implementation of these tests, you can interpret the p value as follows
p <= alpha: reject H0, not normal.
p > alpha: fail to reject H0, normal.

```
In [21]: shapiro(Current)

Out[21]: (0.9707046747207642, 0.378787100315094)
```

Observation:-As the P value of Shapiro test (0.378) is Pvalue> 0.05, the null hypothesis of Shapiro test that the sample is drawn from the population following normal distribution cannot be rejected i.e it follows a normal distribution.

 # Check for Shapiro Normality using Shapiro-Wilk normality test for New golf Ball data set
Testing whether samples are parametric or non parametric by Shapiro Normality Test

In the SciPy implementation of these tests, the hypothesis for Normality test (the p value is as follows)
p <= alpha: reject H0, not normal.
p > alpha: fail to reject H0, normal.

```
In [22]: shapiro(New)

Out[22]: (0.9678263664245605, 0.3064655363559723)
```

Observation:-As the P value of Shapiro test (0.306) is greater than 0.05, the null hypothesis of Shapiro test that the sample is drawn from the population following normal distribution cannot be rejected. It follows a normal distribution.
Other Observations

- Sample size:40
- Number of samples: 2
- Unpaired variables.
- DOF = 40+40-2 = 78
- There are no outliers in given data, nor missing values.
- Both the samples are normally distributed.
- Mean and median values are not much different.
- The Current driving distance data looks more normally distributed, whereas the driving distances data for New balls looks right skewed.
- There is dip in the performance of Current and New balls driving force as mean, median, min, max values differ.

Q1.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) Hypothesis Formulation and Testing

Defining Null and Alternate Hypothesis

Use two tailed independent sample T test for means

Null Hypothesis:

H0: µold−µnew =0 (New coating does not have effect on driving distances)

Alternate Hypothesis:

H1: µold−µnew ≠0 (New coating does have significant effect on driving distances)

- µ = Mean
- Since the sole purpose of the test is to check whether there is any effect on driving Distance due to the new coating, we could prefer a Two Tailed T Test.

#Welch Two Sample t-test of the hypothesis

 # Check for T_statistics and P_value for both Current and New golf ball data set

```
In [23]: t_statistic,p_value = ttest_ind(Current,New)
         print(t_statistic,p_value)

         1.32836159352 0.187932284919
```

Inference:-The Pvalue = 0.188

The Null hypothesis is not rejected as p>0.05.When the population of golf balls is assumed to be parametric,i.e. it follows normal distribution, the ttest_ind tells us that the null hypothesis cannot be rejected as P value is 0.188, which is > 0.05 (based on Shapiro test) . We can say with **95% confidence that there is no significance difference between the means(driving distance) of the two population** i.e for Current and New golf ball.

 # Check for variances of population using Levene's Test

```
In [24]: levene(Current,New)
Out[24]: LeveneResult(statistic=0.25532382917657409, pvalue=0.61477595233313087)
```

Inference:- Since, the Pvalue(0.6147) for the levene's test > 0.05, we cannot reject the null hypothesis of levene's test, **which means that the population variances of both the samples are statistically the same.**

As the samples pass levene's test, Pooled standard deviation can be used for calculating the POWER OF TEST and also calculating the DELTA value.

Calculation of POWER OF TEST with pooled standard deviation
First we need to find the pooledstd and then we calculated the delta value

```
In [25]: Pooledstd = np.sqrt(((40-1)*varC+ (40-1)*varN)/(40+40-2))
         Pooledstd

Out[25]: 9.3424692008714043
```

Observation:- The pooled std is 9.3424

calculating the delta Value

```
In [26]: delta = (meanC - meanN)/Pooledstd
         delta

Out[26]: 0.29703068218208772
```

Observation:- The delta = 0.297

Calculation of POWER OF TEST

```
In [27]: print(ttest_power(delta, nobs = 40, alpha = 0.05, alternative = 'two-sided'))

         0.449274188539
```

Observation - The power of the test was calculated to be only 44.9%, with population of equal variances.

sample size required and re-run ttest_power

```
print(ttest_power(delta, nobs = 198, alpha = 0.05, alternative = 'two-sided'))

0.986067051943
```

Observation The power of the test was calculated to be only 44.9%, with population of equal variances. If the sample size above 198, it will increase the power of the test and it will approach the required industrial standards of 98% (levene's test pass case)

Calculating the Confidence Interval for Current Golf Ball

```
In [29]: z_critical_Current= stats.norm.ppf(q=.975)
         z_critical_Current
Out[29]: 1.959963984540054

In [30]: stderror_Current= 8.752/math.sqrt(40)
         stderror_Current
Out[30]: 1.3838127040896828

In [31]: CI_Lower_Current = meanC -z_critical_Current*stderror_Current
         CI_Upper_Current = meanC + z_critical_Current*stderror_Current

In [32]: CI_Lower_Current,CI_Upper_Current
Out[32]: (267.56277693863524, 272.98722306136472)
```

1. First calculate Z_critical =1.95
2. Std error formulae =

$$e = 1.96\frac{\sigma}{\sqrt{n}}$$

3. The calculate CI_Upper and CI_Lower
4. CI_Upper_Current = 272.98
5. CI_Lower_Current = 267.56

Calculating the Confidence Interval for Current Golf Ball

```
In [33]: z_critical_New= stats.norm.ppf(q=.975)
         z_critical_New
Out[33]: 1.959963984540054

In [34]: stderror_New= 9.89/math.sqrt(40)
         stderror_New
Out[34]: 1.5637463029532637

In [35]: CI_Lower_New = meanC -z_critical_New*stderror_New
         CI_Upper_New = meanC + z_critical_New*stderror_New

In [36]: CI_Lower_New,CI_Upper_New
Out[36]: (267.21011356525395, 273.33988643474601)
```

1. First calculate Z_critical
2. Std error formulae =

$$e = 1.96\frac{\sigma}{\sqrt{n}}$$

3.
4. The calculate CI_Upper and CI_Lower
5. CI_Upper_New= 273.339
6. CI_Lower_New = 267.210

Conclusions:

Q2 :-Analyze the data to provide the hypothesis testing conclusion. What is the p-value for your test? What is your recommendation for Par Inc.?

Conclusion:- The population of golf balls is assumed to be parametric,i.e. It follows normal distribution, the ttest_ind for the unpaired data, since the Pvalue =0.098 the null hypothesis cannot be rejected since Pvalue > 0.05 (based on Shapiro test) . We can say that with 95% confidence that there is no significance difference between the means(driving distance) of the two population i.e for Current and New golf ball) . We can also suggested to check the effect on the weights and other characteristics like size and shape of the new balls.

The 95% confidence interval for the population mean of the current model is 272.98 to 267.56 and of the new model is 273.3 to 267.21. It means that the estimated population mean for Par, Inc. should lie within this range for consistent result.

Q3:-Do you see a need for larger sample sizes and more testing with the golf balls? Discuss

Conclusion:- The power of the test was calculated to be only 44.9%, with population of equal variances. Any sample size above 198, will increase the power of the test and it will approach the required standards of 98% (levene's test pass case), also the given sample is from only one golf course, It is advisable that test should perform on different kind of golf courses to take care of the differences in grounds.