AS PROJECT REPORT

DSBA

Contents

Problem 1

A physiotherapist with a male football team is interested in studying the relationship between foot injuries and the positions at which the players play from the data collected.

	Striker	Forward	Attacking Midfielder	Winger	**Total**
Players Injured	45	56	24	20	**145**
Players Not Injured	32	38	11	9	**90**
Total	**77**	**94**	**35**	**29**	**235**

Based on the above data, answer the following questions.

1.1 What is the probability that a randomly chosen player would suffer an injury?

Probability means number of favorable outcome /Total number of outcome of the experiment i.e.

Probability = Number of favorable outcomes
Total number of outcomes

Here, in this case as we can see from the above table that the total injured person is 145
And the total number of players are 235.

So basically,

Probability= Total injured person/Total number of players
=145/235
=0.61

1.2 What is the probability that a player is a forward or a winger?

As shown in data given

The players which are forward is 94
The players which are winger is 29

So

Total number of players that are forward and winger will be= 94+29 = 123

Total number of players are = 235

Probability = Total number of players that are forward and winger / Total number of players
=123/235
=0.52

1.3 What is the probability that a randomly chosen player plays in a striker position and has a foot injury?

Player that is injured and striker=45
Total number of striker=77

Probability= Player that is striker and injured/Total number of striker
=45/77
=0.58

1.4 What is the probability that a randomly chosen injured player is a striker?

Total number of striker which are injured= 45
Total number of injured players=145

Probability= Total number of striker which are injured/ Total number of injured players
=45/145
=0.31

Problem 2

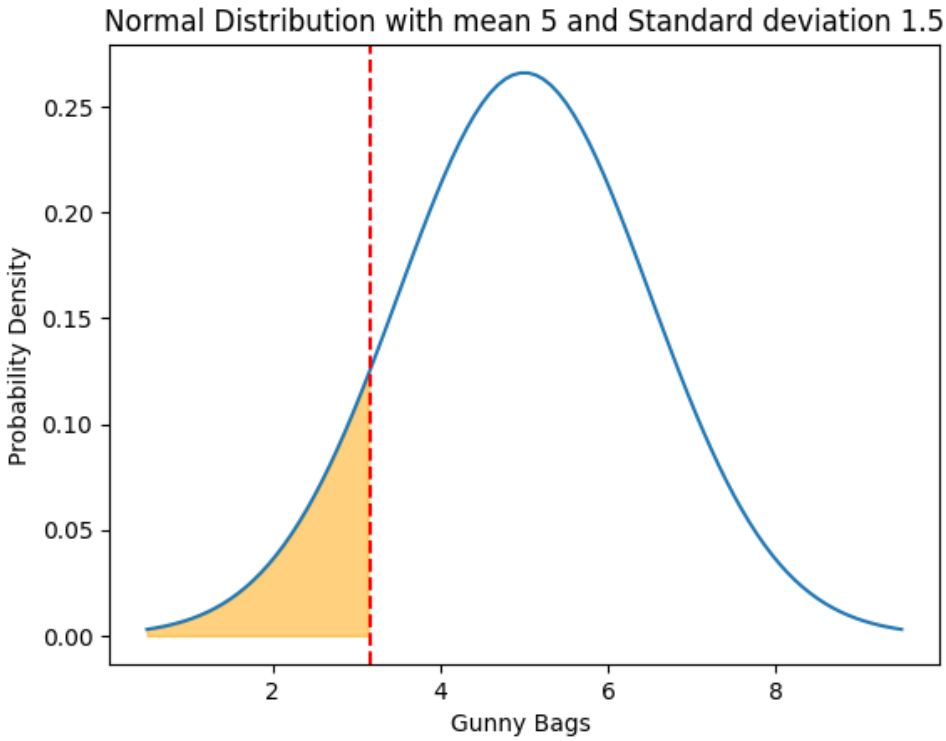
The breaking strength of gunny bags used for packaging cement is normally distributed with a mean of 5 kg per sq. centimeter and a standard deviation of 1.5 kg per sq. centimeter. The quality team of the cement company wants to know the following about the packaging material to better understand wastage or pilferage within the supply chain; Answer the questions below based on the given information; **(Provide an appropriate visual representation of your answers, without which marks will be deducted)**

2.1 What proportion of the gunny bags have a breaking strength of less than 3.17 kg per sq cm?

The mean given is 5 and the standard deviation is 1.5

We get,

- Proportion of gunny bags have a breaking strength of less than 3.17 kg per sq. cm: 0.11
- In another words, we can say that the 11% gunny bags have a breaking strength less than 3.17kg per sq. cm.

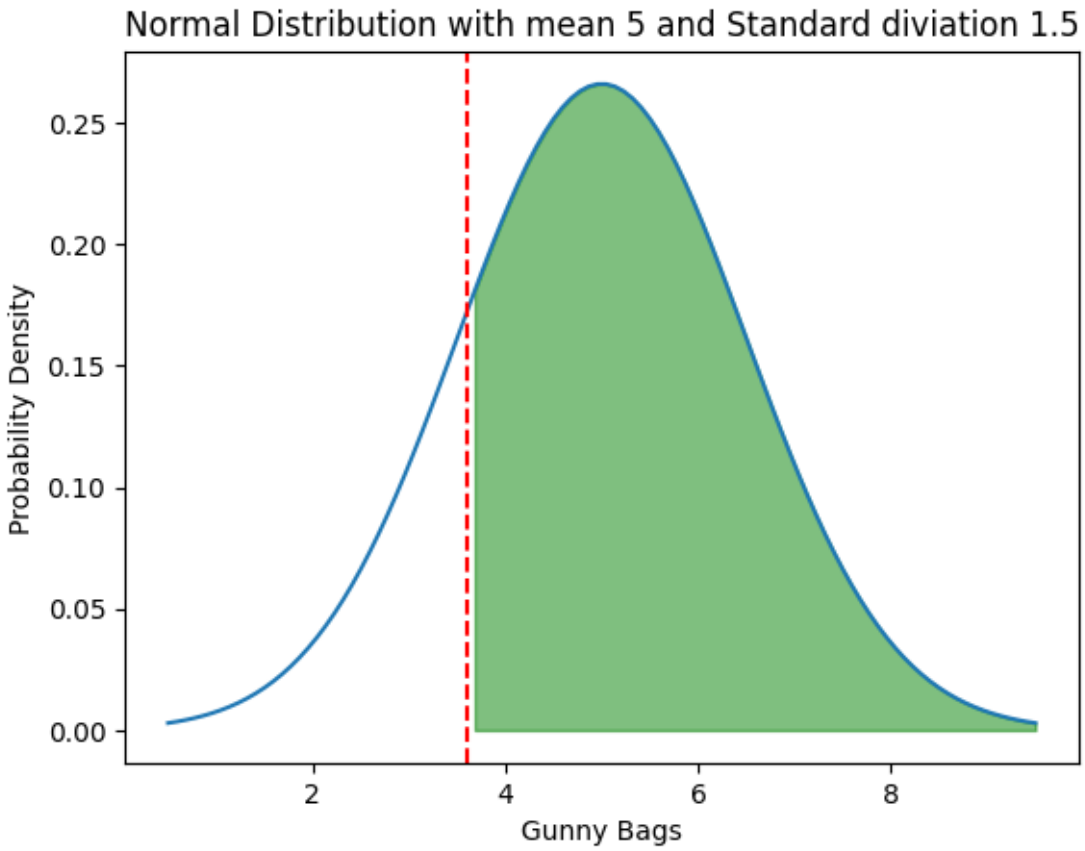


2.2 What proportion of the gunny bags have a breaking strength of at least 3.6 kg per sq cm.?

The mean given is 5 and the standard deviation is 1.5

We get,

- Proportion of gunny bags have a breaking strength of at least 3.6 kg per sq. cm: 0.82
- In another words, we can say that the 82% gunny bags have a breaking strength at least 3.6 kg per sq. cm.

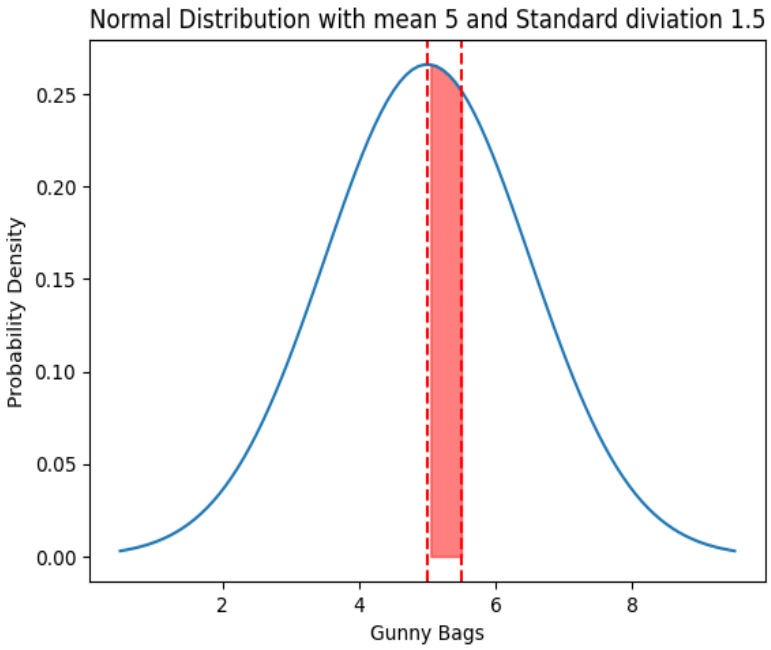


2.3 What proportion of the gunny bags have a breaking strength between 5 and 5.5 kg per sq cm.?

The mean given is 5 and the standard deviation is 1.5

We get,

- Proportion of gunny bags have a breaking strength between 5 and 5.5 kg per sq. cm: 0.13
- In another words, we can say that the 13% gunny bags have a breaking strength between 5 and 5.5 kg per sq. cm.

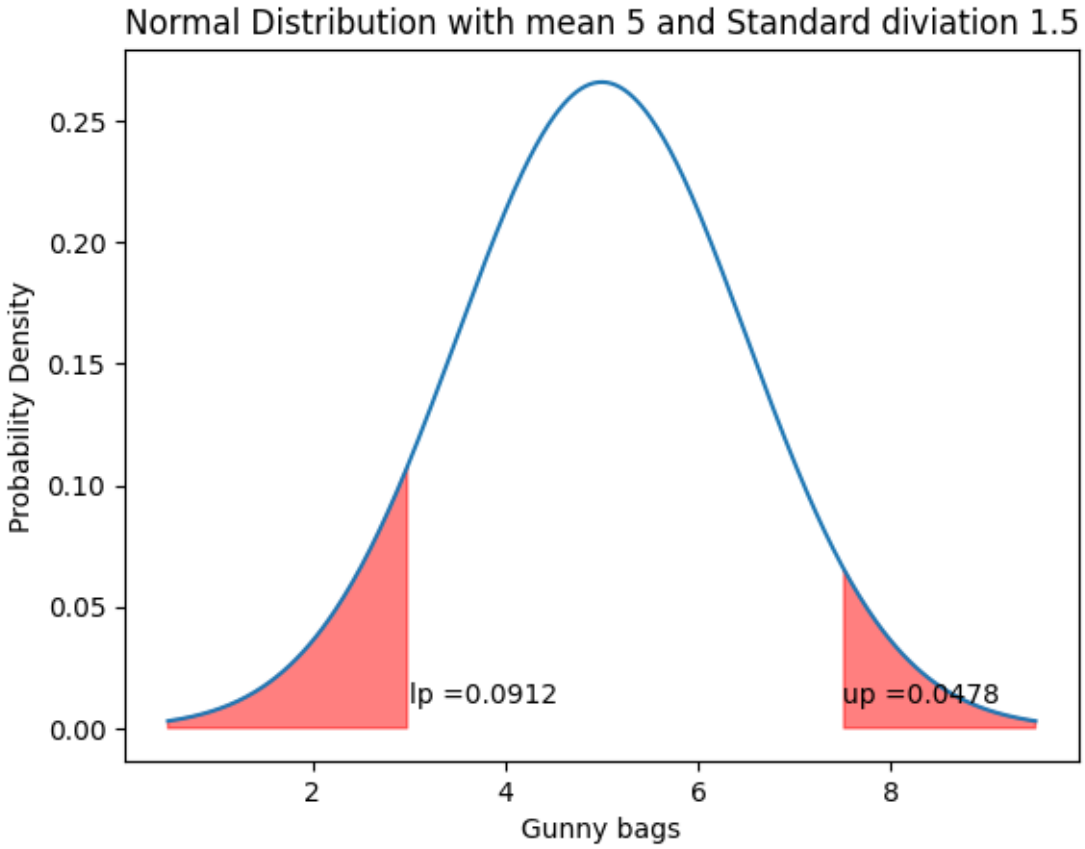

Normal Distribution with mean 5 and Standard diviation 1.5
Probability Density
0.25
0.20
0.15
0.10
0.05
0.00
2
4
6
8
Gunny Bags

2.4 What proportion of the gunny bags have a breaking strength NOT between 3 and 7.5 kg per sq cm.?

The mean given is 5 and the standard deviation is 1.5

We get,

- Proportion of gunny bags have a breaking strength NOT between 3 and 7.5 kg per sq. cm: 0.319
- In another words, we can say that the 31.9% gunny bags have a breaking strength NOT between 3 and 3.7 kg per sq. cm.



Problem 3

Zingaro stone printing is a company that specializes in printing images or patterns on polished or unpolished stones. However, for the optimum level of printing of the image, the stone surface has to have a Brinell's hardness index of at least 150. Recently, Zingaro has received a batch of polished and unpolished stones from its clients. Use the data provided to answer the following (assuming a 5% significance level);

3.1 Zingaro has reason to believe that the unpolished stones may not be suitable for printing. Do you think Zingaro is justified in thinking so?

- Loaded the dataset given of Zingaro
- To find the datatype i used .info() It tells how many datatype does it have like is it integer, float etc

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 75 entries, 0 to 74
Data columns (total 2 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Unpolished            75 non-null     float64
 1   Treated and Polished  75 non-null     float64
dtypes: float64(2)
memory usage: 1.3 KB
```

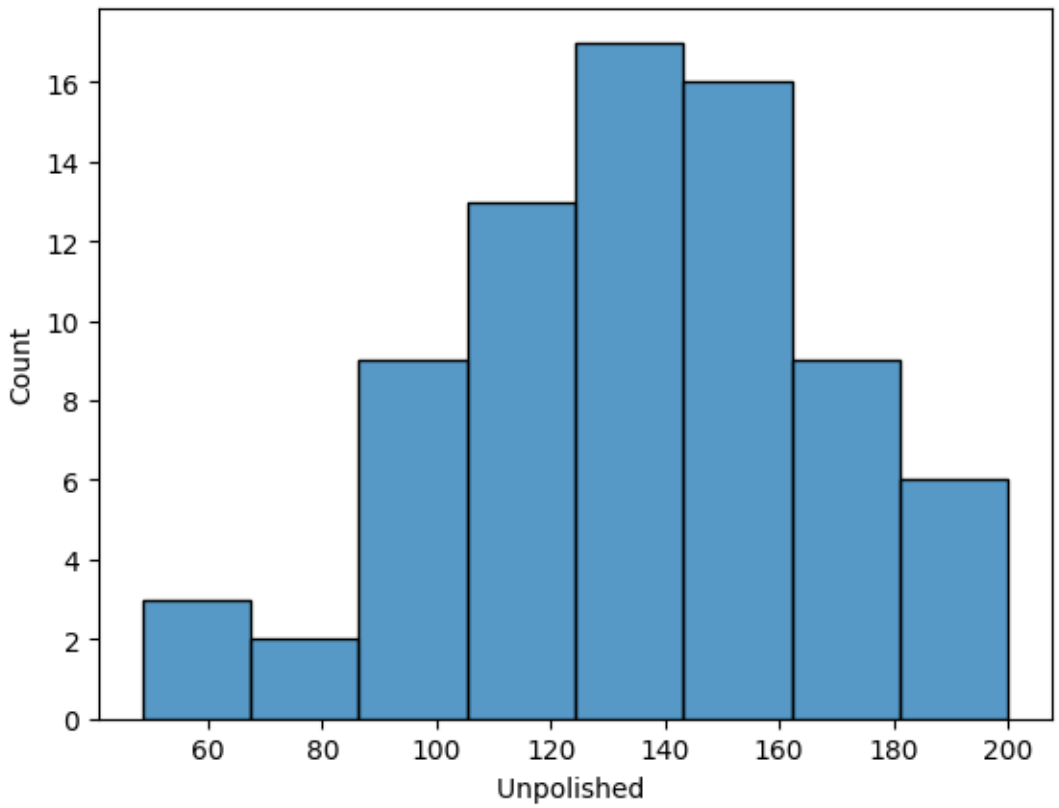
- Head the data == To check the first 5 columns

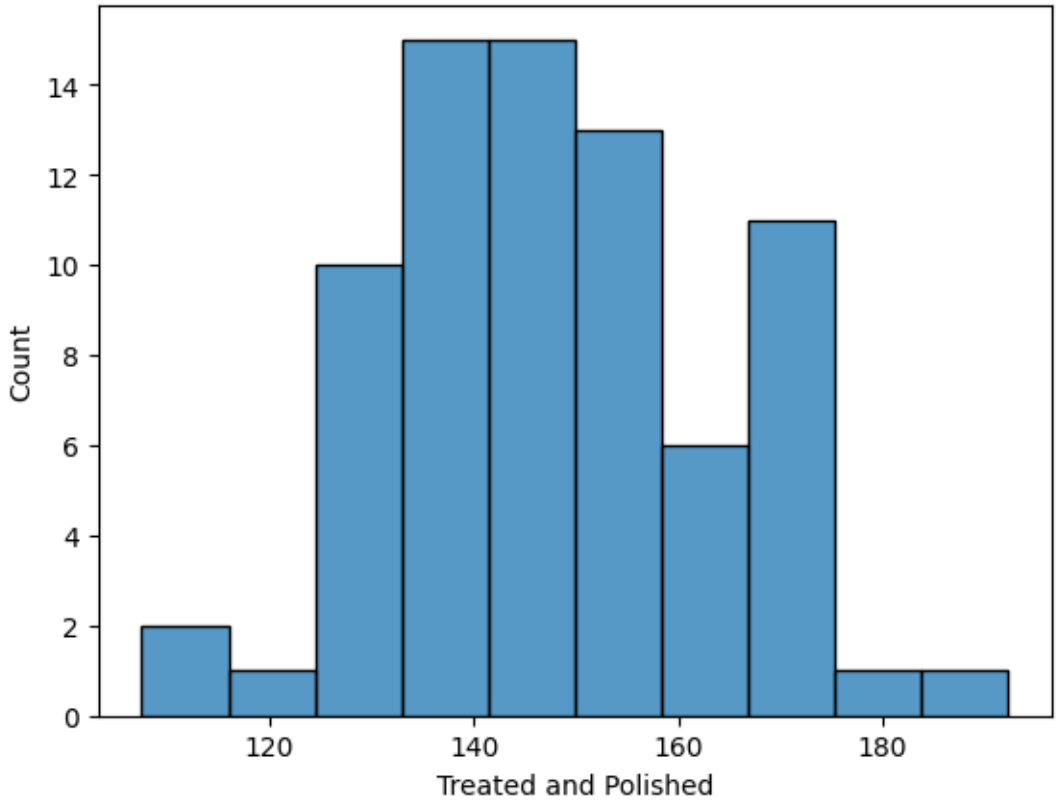
	Unpolished	Treated and Polished
0	164.481713	133.209393
1	154.307045	138.482771
2	129.861048	159.665201
3	159.096184	145.663528
4	135.256748	136.789227

- Tail the data == To check the last 5 columns

	Unpolished	Treated and Polished
70	123.067611	142.293544
71	171.822218	140.124092
72	88.135994	141.393091
73	145.150397	131.370530
74	170.854823	144.502647

- To check the normal distribution, we have plotted the histogram for both the parameters i.e. Unpolished and 'Treated and Polished'





- We have checked for null values and there is no null values in both the columns

```
Unpolished  0
Treated and Polished 0
```

According to the question we have to take H0 (null hypothesis) and H1: Alternative hypothesis.

Defining both below:

H0: stone found to be of adequate hardness >=150
H1: Unpolished stone hardness is not suitable for printing <150

We have performed one t test and found that the p value is less than the significance value (0.05) so we reject the null hypothesis

So Unpolished stone hardness is not suitable for printing

Now ,Zingaro has reason to believe that the unpolished stones may not be suitable for printing.

3.2 Is the mean hardness of the polished and unpolished stones the same?

According to the question we have to take H0 (null hypothesis) and H1: Alternative hypothesis.

Defining both below:

H0: mean hardness of the polished = mean hardness of the unpolished

H1: mean hardness of the polished != mean hardness of the unpolished

We performed two t test and found that the p value is less than the significance value that is 0.05

So, we rejected the null hypothesis

moreover, we have compared mean also for both the columns and found that the

Mean of unpolished

```
134.11052653373332
```

Mean of Treated and polished

```
147.78811718133335
```

So by this also we can say that the mean hardness of polished and unpolished stones are not the same.

Problem 4

Dental implant data: The hardness of metal implants in dental cavities depends on multiple factors, such as the method of implant, the temperature at which the metal is treated, the alloy used as well as the dentists who may favor one method above another and may work better in his/her favourite method. The response is the variable of interest.

4.1 How does the hardness of implants vary depending on dentists?

Problem 4

Dental implant data: The hardness of metal implants in dental cavities depends on multiple factors, such as the method of implant, the temperature at which the metal is treated, the alloy used as well as the dentists who may favor one method above another and may work better in his/her favourite method. The response is the variable of interest.

4.1 How does the hardness of implants vary depending on dentists?

- Loaded the data se given of dental
- Checked the head of the dataset

	Dentist	Method	Alloy	Temp	Response
0	1	1	1	1500	813
1	1	1	1	1600	792
2	1	1	1	1700	792
3	1	1	2	1500	907
4	1	1	2	1600	792

- Checked the tail of dataset

	Dentist	Method	Alloy	Temp	Response
85	5	3	1	1600	483
86	5	3	1	1700	405
87	5	3	2	1500	536
88	5	3	2	1600	405
89	5	3	2	1700	312

- Get the info of the data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 90 entries, 0 to 89
Data columns (total 5 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Dentist   90 non-null     int64
 1   Method    90 non-null     int64
 2   Alloy     90 non-null     int64
 3   Temp      90 non-null     int64
 4   Response  90 non-null     int64
dtypes: int64(5)
memory usage: 3.6 KB
```

- Get the details and summary of the dataset given

	count	mean	std	min	25%	50%	75%	max
Dentist	90.0	3.000000	1.422136	1.0	2.0	3.0	4.0	5.0
Method	90.0	2.000000	0.821071	1.0	1.0	2.0	3.0	3.0
Alloy	90.0	1.500000	0.502801	1.0	1.0	1.5	2.0	2.0
Temp	90.0	1600.000000	82.107083	1500.0	1500.0	1600.0	1700.0	1700.0
Response	90.0	741.777778	145.767845	289.0	698.0	767.0	824.0	1115.0

- Checked for null values

```
print(i,df_excel[i].is
```

```
Dentist 0
Method 0
Alloy 0
Temp 0
Response 0
```

Step 1::

Test for normality

- Null hypothesis:: The given observation is normally distributed
- Alternative hypothesis: The given observation is not normally distributed

We found that Shapiro test fails for dentist==4 so we need to do Anderson test

```
ShapiroResult(statistic=0.7741794586181641, pvalue=3.6831273586557245e-09)
ShapiroResult(statistic=0.7735344767570496, pvalue=3.549490479315409e-09)
ShapiroResult(statistic=0.7705162763595581, pvalue=2.988440606088716e-09)
ShapiroResult(statistic=0.7672054767608643, pvalue=2.4786037666757466e-09)
```

After this we performed levene test that also got failed

We have also split the dataframe based on Alloys as Alloy 1 and Alloy2

After this we have done the analysis of each alloys that is Alloy 1 and Alloy2

After this we have done anova analysis of each alloy as shown below:

Alloy 1 and Alloy2 respectively

	sum_sq	df	F	PR(>F)
C(Dentist)	30337.866667	4.0	1.016625	0.417813
Residual	186510.833333	25.0	NaN	NaN

	sum_sq	df	F	PR(>F)
C(Dentist)	106683.688889	4.0	2.591255	0.051875
C(Method)	148472.177778	2.0	7.212522	0.002211
Residual	391121.377778	38.0	NaN	NaN

H_0: The mean response is the same for all three dentists.
H_a: For at least one pair of dentists the mean response will be different

the mean response for all the three types of dentist is same

4.2 How does the hardness of implants vary depending on methods?

Shapiro test has been performed for methods:

```
ShapiroResult(statistic=0.7740979194641113, pvalue=2.6304073936361805e-12)
ShapiroResult(statistic=0.7745263576507568, pvalue=2.7103042032905478e-12)
ShapiroResult(statistic=0.7585681676864624, pvalue=9.120459379050039e-13)
```

P value is less than 0.05 and that's what we have failed the Shapiro test

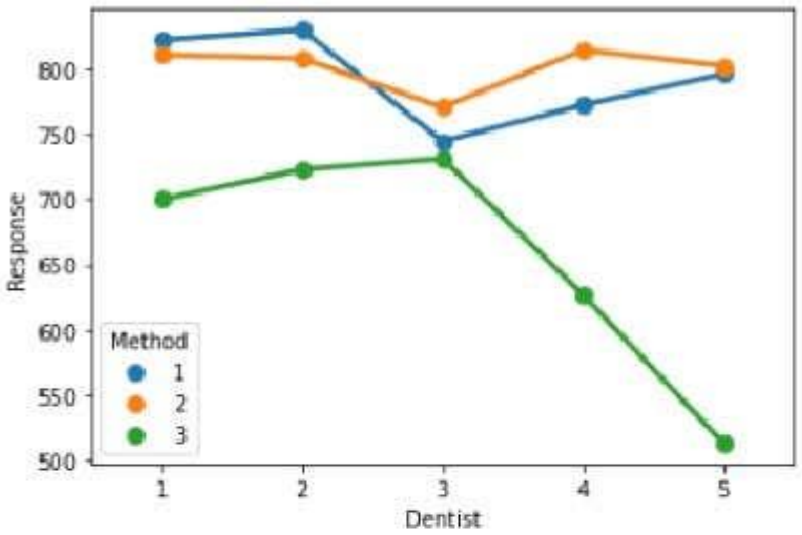
Tukey test has been performed

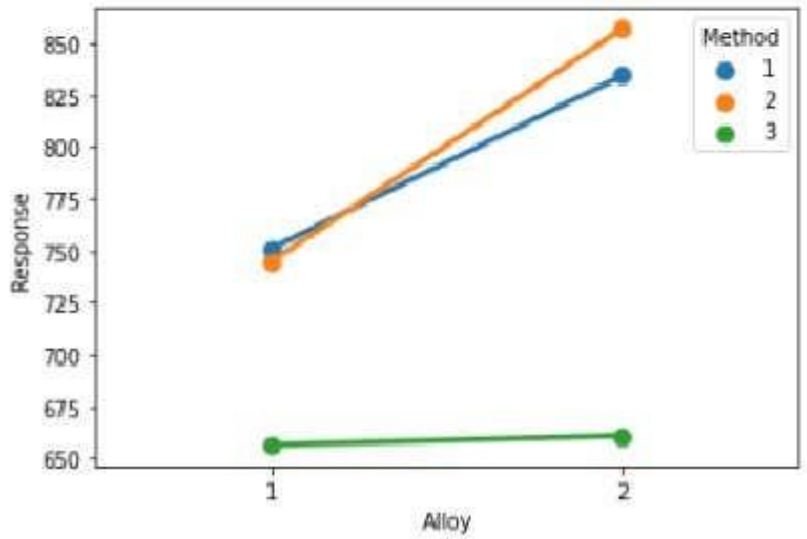
After this we have done anova analysis of each method as shown below:

	sum_sq	df	F	PR(>F)
C(Dentist)	30337.866667	4.0	1.016625	0.417813
Residual	186510.833333	25.0	NaN	NaN

4.3 What is the interaction effect between the dentist and method on the hardness of dental implants for each type of alloy?

P value for interaction effect of dentist and method is 1.657388e-02Since the p value for interaction effect is way less than alpha,
we can conclude that there is noeffect of interaction effect on our response variable.





Since the lines are not parallel to each other and clearly two lines are intersecting each other itmeans that there is significant interaction between the dentist and method use

4.4 How does the hardness of implants vary depending on dentists and methods together?

Testing of the null HypothesisAfter performing one way Anovaon ‘Dentist’ with respect to ‘method’ we get p value as 0.000003 Sine the p value is greater than alpha (0.05) we fail to reject null hypothesis.

```
ShapiroResult(statistic=0.7740979194641113, pvalue=2.6304073936361805e-12)
ShapiroResult(statistic=0.7745263576507568, pvalue=2.7103042032905478e-12)
ShapiroResult(statistic=0.7585681676864624, pvalue=9.120459379050039e-13)
```

	sum_sq	df	F	PR(>F)
C(Dentist)	106683.688889	4.0	2.591255	0.051875
C(Method)	148472.177778	2.0	7.212522	0.002211
Residual	391121.377778	38.0	NaN	NaN