Advanced Statistics (AS)

A Project Report

greatlearning
Learning for Life

BY

ABHISHEK K HIREMATH

List of Tables:

List Of Figures:

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**

Table: -1

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

Solution: Probability that a randomly chosen player would suffer an injury

= Total Number of Players Injured/ Total Number of Players

= 145/235

= **0.62.**

Probability that a randomly chosen player would suffer an injury is **0.617**

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

Solution: Probability that a player is a forward or a winger

= (Number of Forwards + Number of Wingers) */ Total Number of Players

= (94+29)/235

= **0.52.**

Probability that player is a forward or a winger **0.5234**

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

Solution: Probability that a randomly chosen player plays in a striker position a foot injury

= Number of Injured Strikers/ Total Number of Players

= 45/235

= **0.19.**

Probability that randomly chosen player plays in a striker position and has a foot injury **= 0.191**

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

Solution: Probability that a randomly chosen injured player is a striker

= Number of Injured Strikers/ Total Number of Injured Players

= 45/145

= **0.31.**

Probability that randomly chosen injured player is a striker **= 0.310**

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.

Mean = **5 kg per sq. centimeter = mu**

Standard deviation = **1.5 kg per sq. centimeter = sigma**

The breaking strength of gunny bags used for packaging cement is **normally distributed**

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

Solution: Here as per the problem statement we have

X = 3.17, μ = 5, σ = 1.5

Let us calculate the Z score & p value corresponds to it:

Z score = x - μ/ σ

Z-score= (3.17 - 5)/ 1.5 = -1.22

P-value from Z-Table:

P(x<3.17) = 0.11123

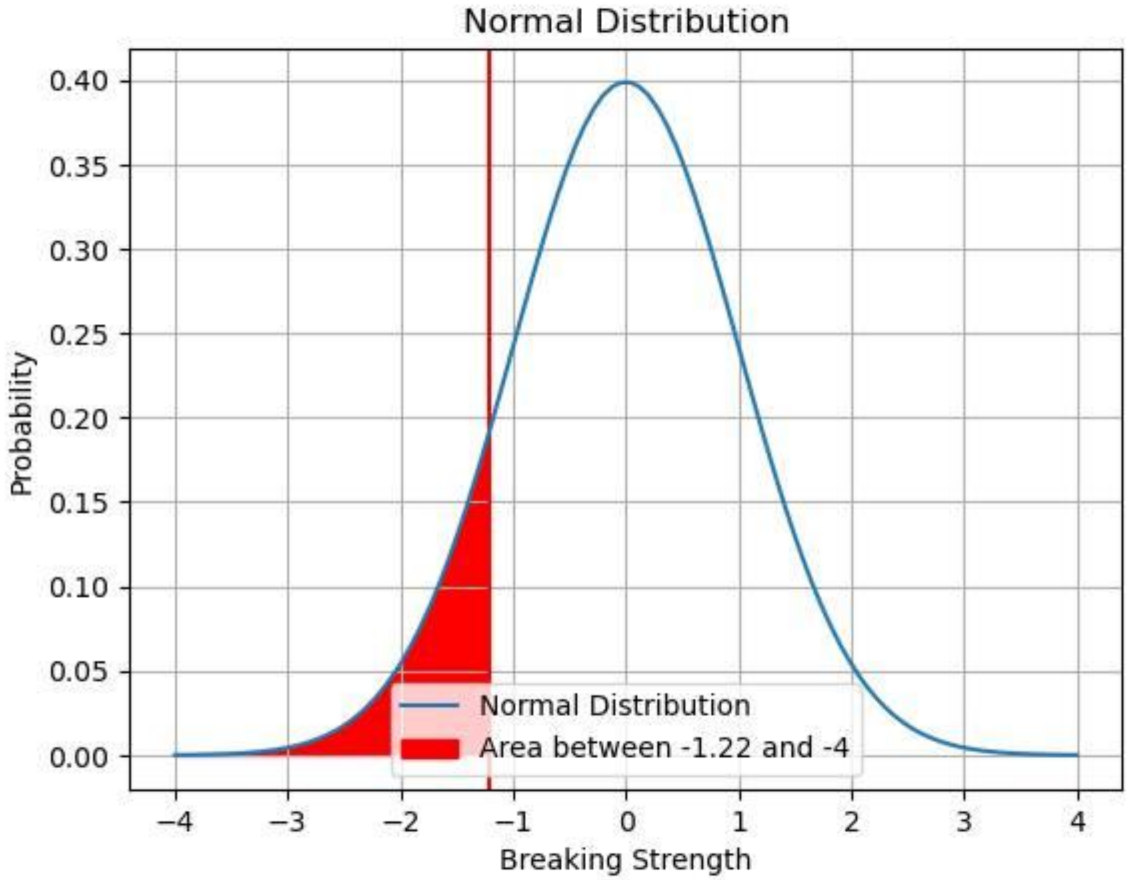

Fig: -1

Hence as per the p value, 11.12 % of the gunny bags have a breaking strength less than 3.17 kg per sq cm.

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

Solution: Here as per the problem statement we have

X = 3.6, $\mu = 5$, $\sigma = 1.5$

Let us calculate the Z score & p value corresponds to it:

Z score = x - μ/ σ

Z-score = (3.6 - 5)/ 1.5

= -0.9333

P-value from Z-Table:

P(x>3.6) = 1 - P(x<3.6) = 0.8247

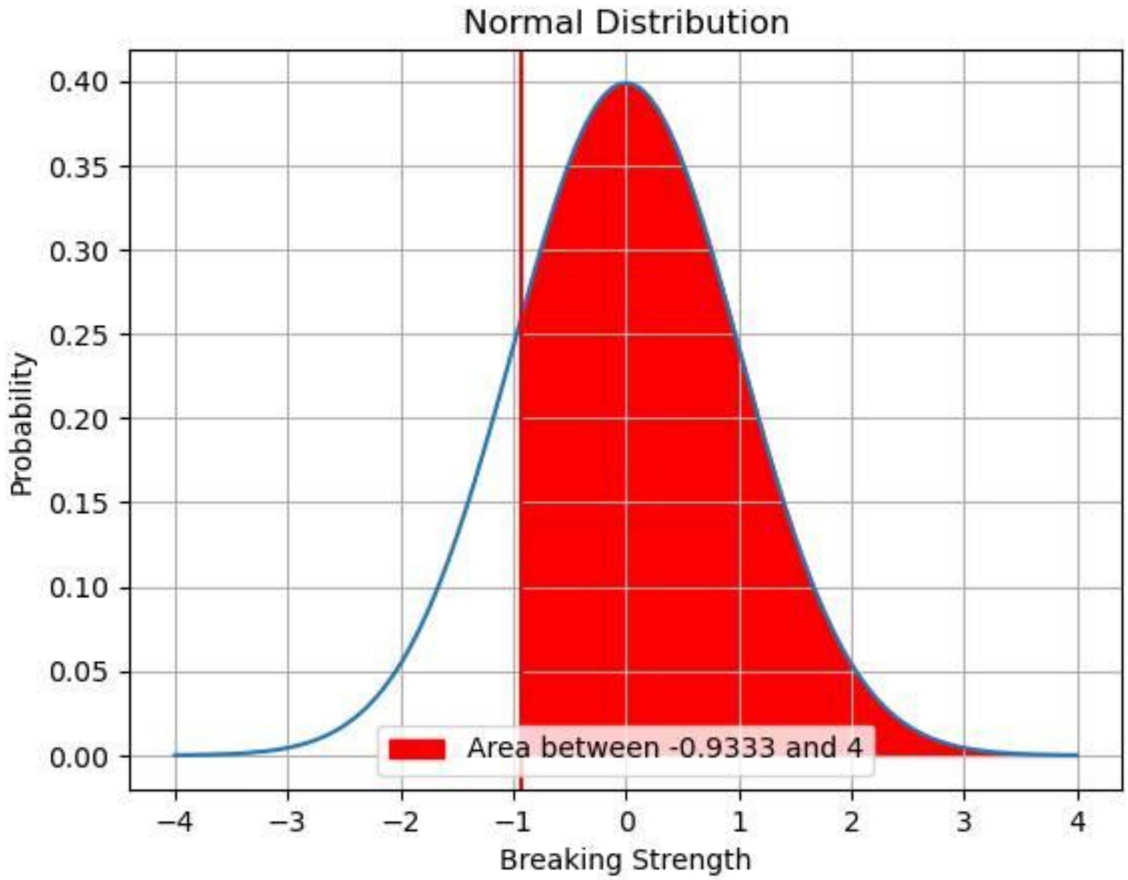

Fig: -2

Hence as per the p value, 82.47% of the 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.?

Solution: Here as per the problem statement we have

X = 5.5, μ = 5, σ = 1.5

Let us calculate the Z score & p value corresponds to it:

Z score = $x - \mu / \sigma$

Z-score = (5.5 - 5)/ 1.5

= 0.33333

P-value from Z-Table:

P(5<x<5.5) = P(x<5.5) - 0.5 = 0.13056

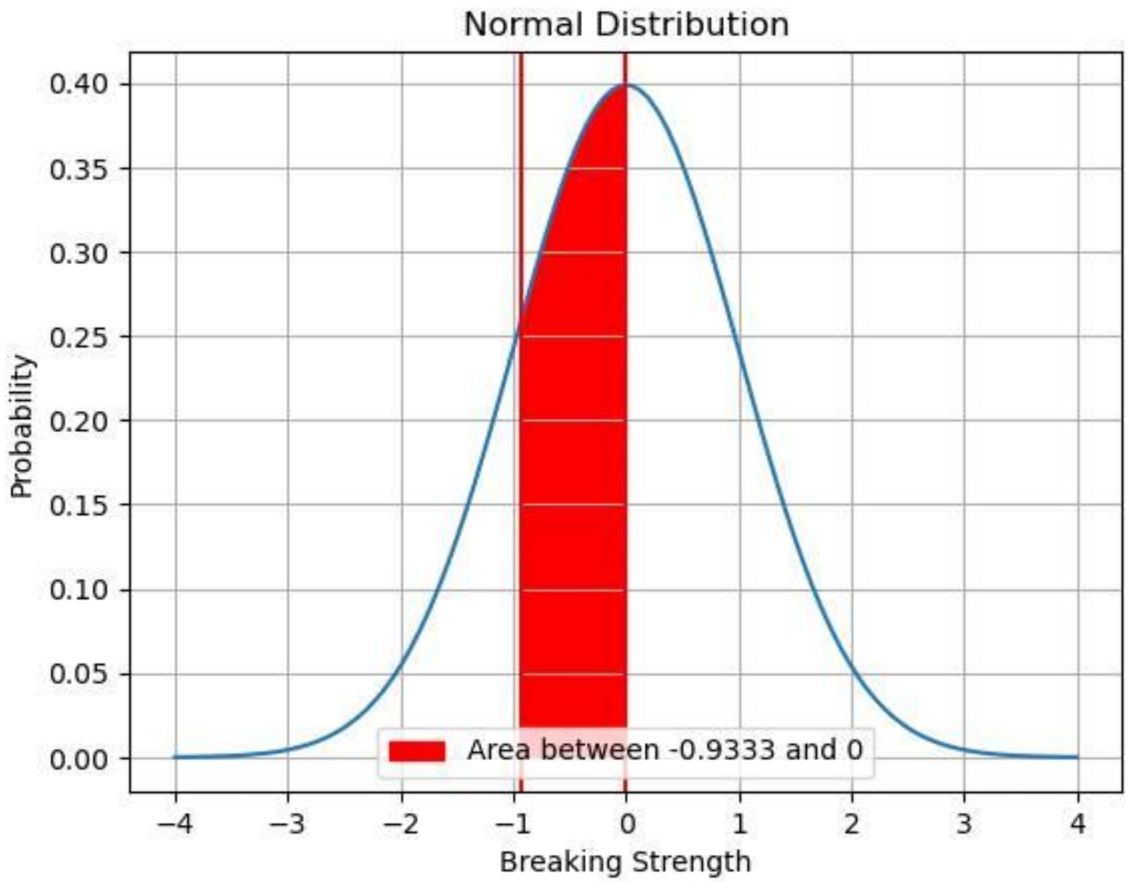

Fig: -3

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

Answer: Here as per the problem statement we have

X = 3 & 7.5, $\mu = 5$, $\sigma = 1.5$

Let us calculate the Z score & p value corresponds to it:

Z score when x is 3

Z score = $x - \mu/\sigma$

=3 - 5 / 1.5

= -1.33333

P-value from Z-Table:

P(x>3) = 1 - P(x<3) = 0.90879

Z score when x is 7.5

Z score = x - μ/σ

=7.5 - 5 / 1.5

= 1.66667

P-value from Z-Table:

P(x>7.5) = 1 - P(x<7.5) = 0.04779

P-value = 0.04779+ 0.091211

= 0.139001

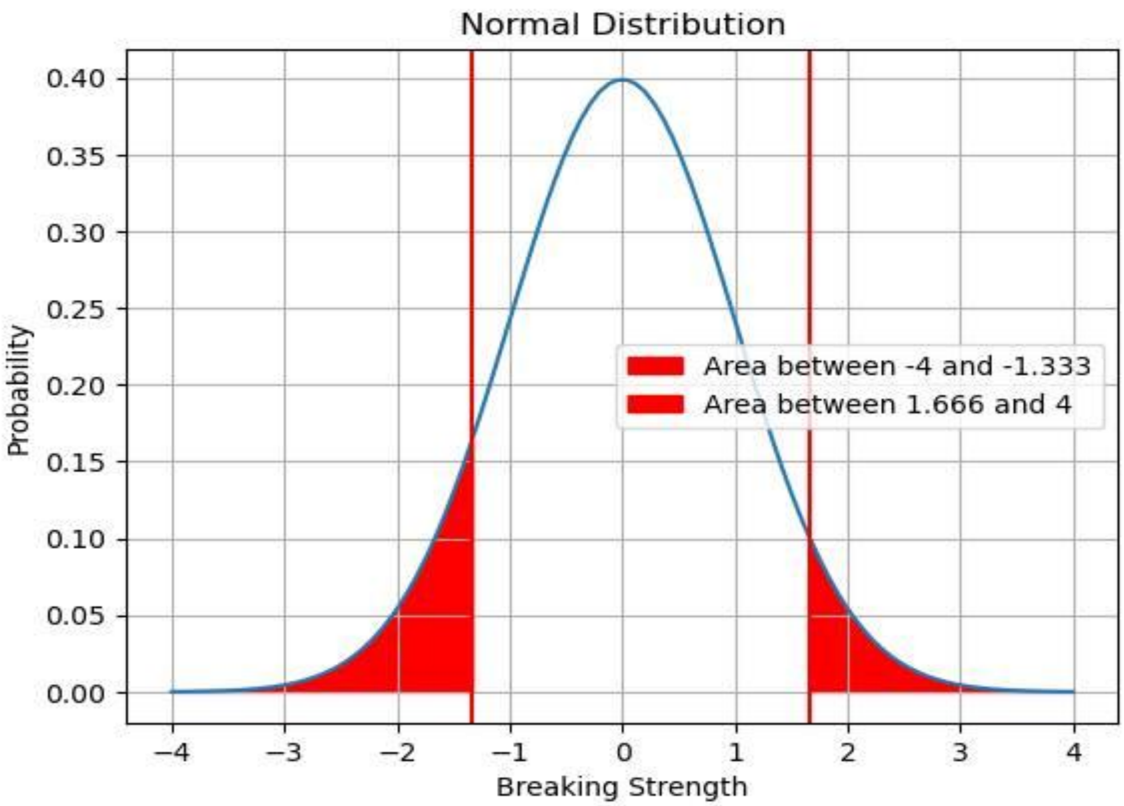

Fig: -4

Hence as per the p value, 13.90% of the gunny bags have a breaking strength NOT between 3 and 7.5 kg per sq cm.

Problem-3

Dataset - [Link]

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);

	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

	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

Fig: -5

```
<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
```

Fig: -6

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?

Solution: First, hypothesizing for unpolished stones, we have:

H0= Zingaro believes now that the unpolished stones may suitable for printing

H1= Zingaro believes now that the unpolished stones may suitable for printing

Step1: Ho (Null Hypothesis) -> Sample Mean >=150

H1 (Alternate Hypothesis) -> Sample Mean <150

Step2: Given, alpha = 0.05

Step 3: 1 sample z-test as this is a one-sided sample with **sample size is 75>30**
Step 4: mu = 150 and n=75(Sample size).

Degrees of Freedom =75-1=74, since the sample size
for both samples are the same.

Also, **mu** unpolished= 134.11, **Sigma** unpolished = 33.04, **mu** polished= 147.79, **Sigma** polished = 15.59.

One sample t test
t statistic: -4.171286997419652e-05, p value: 8.342573994839304e-05
Step 5: p-value for unpolished~0 < 0.05 (i.e. alpha) and hence Null Hypothesis is rejected in that case. **So, unpolished stones do not have a Brinell's hardness index of at least 150.** Whilst, in case of treated and polished p-value is 0.11 > 0.05 (i.e. alpha) and hence Null Hypothesis cannot be rejected in that case. **So, treated, and polished stones have a Brinell's hardness index of at least 150.**

<u>INSIGHTS:</u> <u>Since, the null hypothesis is rejected. At 5% significance level, there is enough evidence to make conclusion that the unpolished stones are not suitable for printing and his claim is justified.</u>

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

Solution: H0= Mean hardness of the polished and unpolished stones is same

H1= Mean hardness of the polished and unpolished stones is not same

Test: Two-tailed t-test, Also, **mu** unpolished= 134.11, **Sigma** unpolished = 33.04, **mu** polished= 147.79, **Sigma** polished = 15.59 and n =75.

The p-value of the two-tailed test is 0.0014655150194628353 and is significantly less than 0.05 (i.e. alpha), so alternate hypothesis prevails.

<u>**INSIGHTS:**</u> <u>Hence, the hardness of polishes and unpolished stones are significantly different. Since, the null hypothesis is rejected. At 5% significance level, there is insufficient evidence to conclude that the mean hardness of the polished and unpolished stones are not the same.</u>

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 favorite method. The response is the variable of interest.

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

Solution: The Hypothesis for the One Way ANOVA are:

H0: There is difference among Dentists on implant hardness

Ha: There is no difference among Dentists on implant hardness

```
ShapiroResult(statistic=0.9288938045501709, pvalue=0.1856756955385208)
ShapiroResult(statistic=0.9665580987930298, pvalue=0.7310513854026794)
ShapiroResult(statistic=0.9573881030082703, pvalue=0.5520960092544556)
ShapiroResult(statistic=0.8089418411254883, pvalue=0.002037237398326397)
ShapiroResult(statistic=0.9546374082565308, pvalue=0.5022943019866943)
```

Fig: -7

From the **Wilk's test**: The Test Statistic value is 0.808 and P-value is 0.00203

Conclusion: We can conclude that P-value is 0.002037 (<<0.05), so we have enough evidence to Reject Null Hypothesis and consider that the sample drawn for Alloy 1 does not follow a normal distribution

Hypothesis 1

Ho: The mean hardness of dental implant is same for dentists provided the alloy 1 is used.

Ha: for at least 1 dentist the mean Hardness of Dental implant is different when using Alloy1.

	df	sum_sq	mean_sq	F	PR(>F)
C(Dentist)	4.0	106683.688889	26670.922222	1.977112	0.116567
Residual	40.0	539593.555556	13489.838889	NaN	NaN

Fig: -8

P-value is 0.11 which is greater than alpha i.e., 0.05. Hence, we Fail to reject null hypothesis and consider there is no difference in means among the dentists in terms of implant hardness for Alloy 1.

Hypothesis 2

Ho: The mean Hardness of Dental implant is same dentists provided the Alloy 2 is used.
Ha: for at least 1 dentist the mean Hardness of Dental implant is different when using Alloy 2.

	df	sum_sq	mean_sq	F	PR(>F)
C(Dentist)	4.0	5.679791e+04	14199.477778	0.524835	0.718031
Residual	40.0	1.082205e+06	27055.122222	NaN	NaN

Fig: -9

P-value is 0.71 which is greater than alpha i.e., 0.05. Hence, we Fail to reject null hypothesis and consider there is no difference in means among the dentists in terms of implant hardness for Alloy 2

After performing one-way Anova on each Alloy, we found p value:

For Alloy 1: 0.116567, For Alloy 2: 0.718031

Hence at p - value > 0.05, we fail to Reject H0. i.e., at 95 % confidence we have statistical evidence to state that there is no difference among the Dentists on implant hardness.

For Alloy 1 - Anderson-Darling Normality Test gives following results-

```
AndersonResult(statistic=2.561066309273194, critical_values=array([0.535, 0.609, 0.731, 0.853, 1.014]),
significance_level=array([15. , 10. ,  5. ,  2.5,  1. ]))
```

Fig: -10

The test statistic is 2.56 We can compare this value to each critical value that corresponds to each significance level to see if the test results are significant.

The critical value for $\alpha = 0.025$ is 0.853. Because the test statistic (2.56) is greater than this critical value, the results are significant at a significance level of 0.025.

Same is the case with all the other values-

We can see that the test results are significant at every significance level, which means we would reject the null hypothesis of the test no matter which significance level we choose to use. Thus, we have sufficient evidence to say that the sample data is not normally distributed.

For Alloy 2 - Anderson-Darling Normality Test gives following results-

```
AndersonResult(statistic=1.8931311726356412, critical_values=array([0.535, 0.609, 0.731, 0.853, 1.014]),
significance_level=array([15. , 10. ,  5. ,  2.5,  1. ]))
```

Fig: -11

The test statistic is 1.89 We can compare this value to each critical value that corresponds to each significance level to see if the test results are significant.

The critical value for $\alpha = 0.025$ is 0.853. Because the test statistic (1.89) is greater than this critical value, the results are significant at a significance level of 0.025.

Same is the case with all the other values: -

We can see that the test results are significant at every significance level, which means we would reject the null hypothesis of the test no matter which significance level we choose to use. Thus, we have sufficient evidence to say that the sample data is not normally distributed.

- Both the distributions are not normally distributed, which violates the assumptions.
- Considerable presence of out liars could also be seen for both alloys, and ANOVA is sensitive to Outliers.

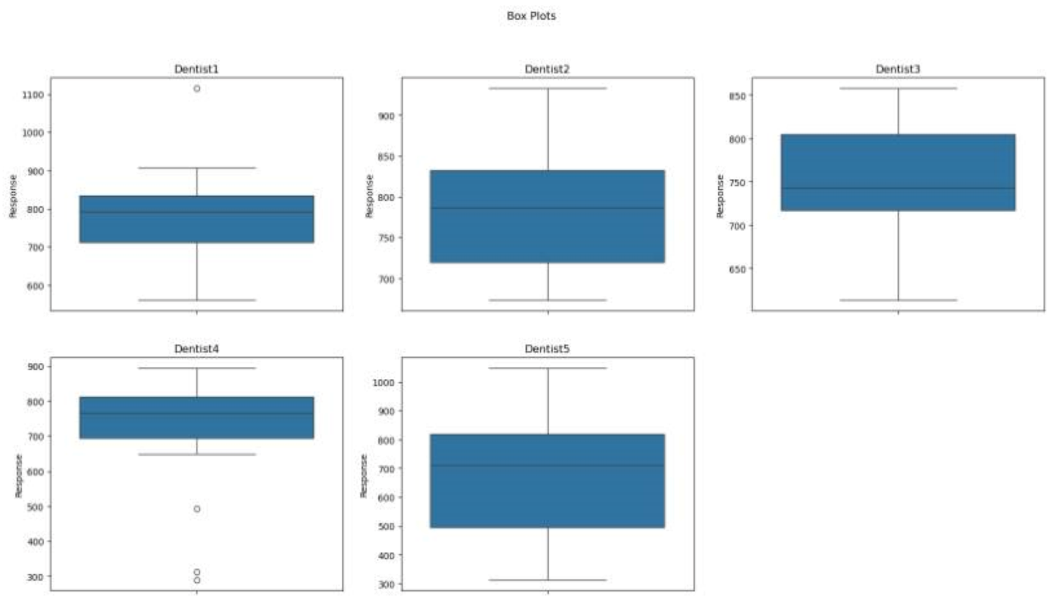

Fig: -12

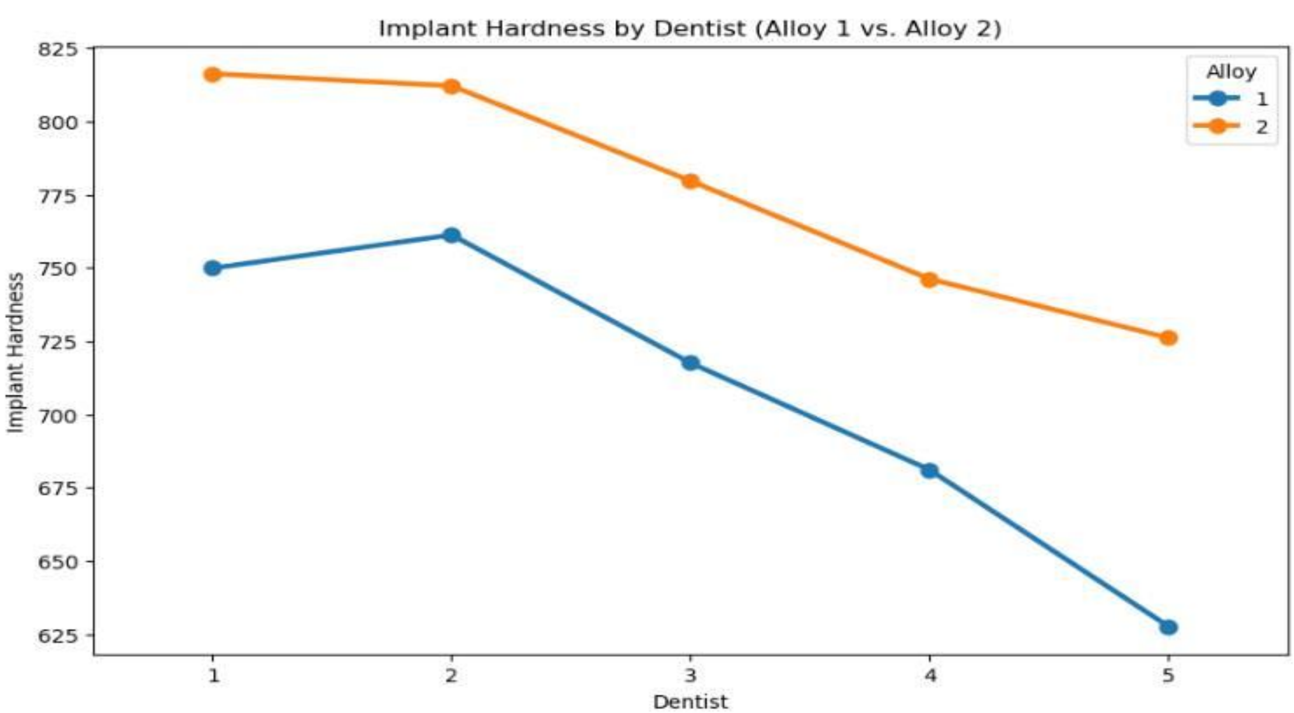

Fig: -13

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

Solution: The Hypothesis for the One Way ANOVA are:

H0: There is difference with Method on implant hardness

Ha: There is no difference with Method on implant hardness

```
ShapiroResult(statistic=0.9652422666549683, pvalue=0.41838711500167847)
ShapiroResult(statistic=0.9371030926704407, pvalue=0.07601878046989441)
ShapiroResult(statistic=0.9091948866844177, pvalue=0.014201466925442219)
```

Fig: -14

From the Wilk test: The Test Statistic value is 0.937 and P-value is 0.0760

Hypothesis Test for Alloy 1 data

- Null hypothesis (for Alloy 1): There is no difference in means among the Methods in terms of implant hardness for Alloy1.
- Alternative hypothesis (for Alloy 1): There is a difference in means among the Methods in terms of implant hardness for Alloy 1

	df	sum_sq	mean_sq	F	PR(>F)
C(Method)	2.0	148472.177778	74236.088889	6.263327	0.004163
Residual	42.0	497805.066667	11852.501587	NaN	NaN

Fig: -15

P-value is 0.004 which is lesser than alpha i.e., 0.05. Hence, we have enough evidence to reject null hypothesis and consider there is a difference in means among the Methods in terms of implant hardness for Alloy 1.

Hypothesis Test for Alloy 2 data

- Null hypothesis (for Alloy 2): There is no difference in means among the Methods in terms of implant hardness for Alloy2.
- Alternative hypothesis (for Alloy2): There is a difference in means among the Methods in terms of implant hardness for Alloy2.

	df	sum_sq	mean_sq	F	PR(>F)
C(Method)	2.0	499640.4	249820.200000	16.4108	0.000005
Residual	42.0	639362.4	15222.914286	NaN	NaN

Fig: -16

P-value is 0.000005 which is lesser than alpha i.e., 0.05. Hence, we have enough evidence to reject null hypothesis and consider there is a difference in means among the Methods in terms of implant hardness for Alloy 2.

As we have rejected the null hypothesis for both the alloys, let us check for which methods the mean implant hardness is significantly different.

We can use Tukey's multiple comparison test for this-

For Alloy 1 –

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
1	2	-6.1333	0.987	-102.714	90.4473	False
1	3	-124.8	0.0085	-221.3807	-28.2193	True
2	3	-118.6667	0.0128	-215.2473	-22.086	True

Fig: -17

Mean difference for method 3 is quite high when compared with both methods 1 & 2**.**

Tukey's HSD Test for multiple comparisons found that the mean implant hardness was significantly different between Method 1 and Method 3 it is different for Method 2 and Method 3 as well.

For Alloy 2-

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
1	2	27.0	0.8046	-82.4506	136.4506	False
1	3	-208.8	0.001	-318.2506	-99.3494	True
2	3	-235.8	0.001	-345.2506	-126.3494	True

Fig: -18

Mean difference for method 3 is quite high when compared with both methods.

Tukey's HSD Test for multiple comparisons found that the mean implant hardness was significantly different between Method 1 and Method 3 it is different for Method 2 and Method 3 as well.

After performing one-way Anova on each Alloy, we found p value:

For Alloy 1: 0.004163, For Alloy 2: 0.000005

Hence at p - value < 0.05, we accept H0, i.e., At 95 % confidence we have statistical evidence to state that there is difference among the method used on implant hardness

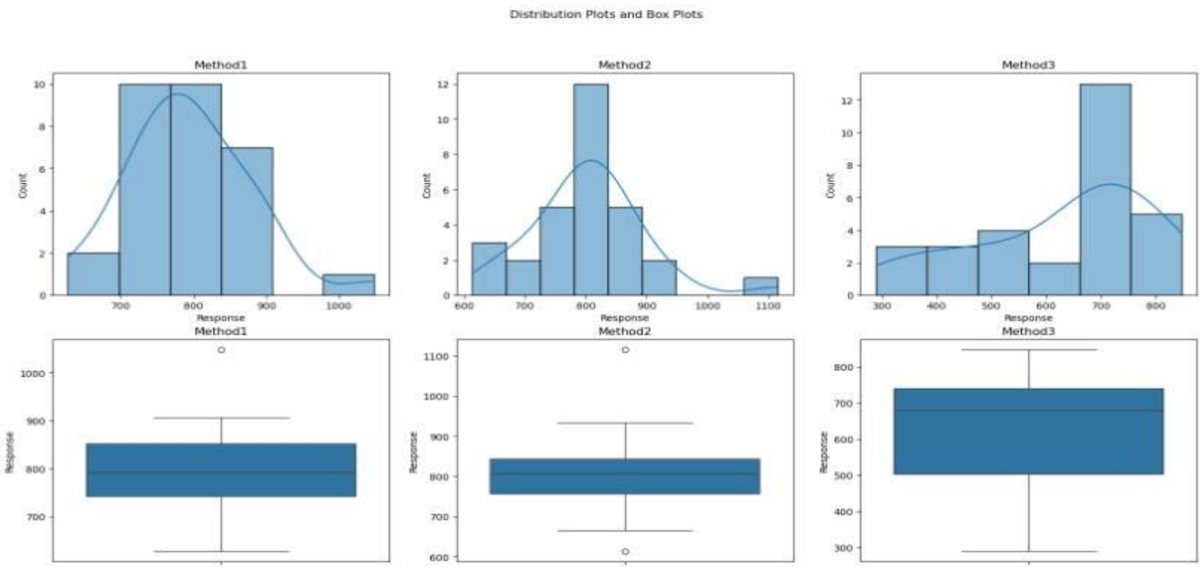

Fig: -19

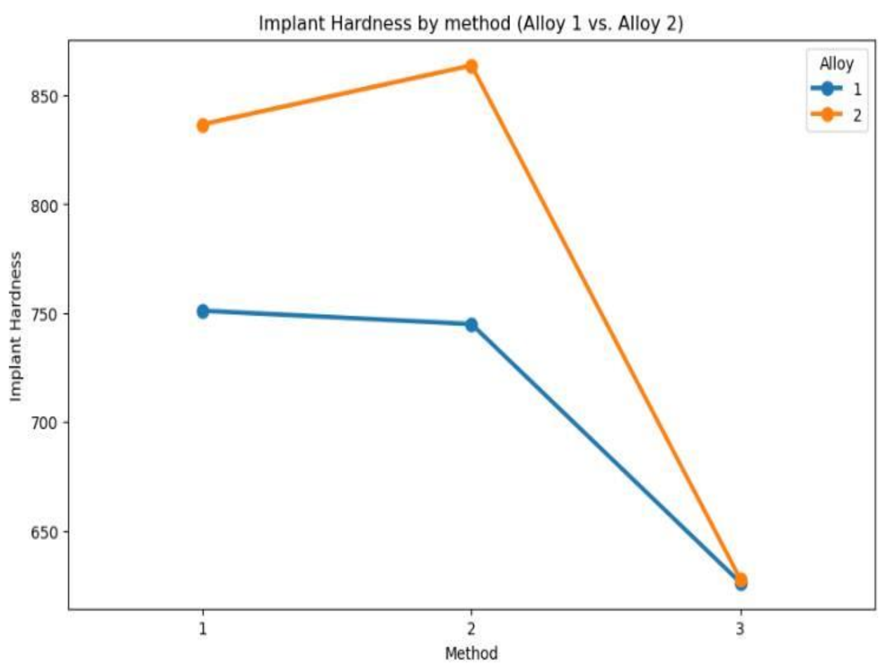

Fig: -20

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

Solution: Define null and alternative hypothesis

- Null hypothesis (H0): There is no difference among the Interaction effect between Dentist and Method levels in terms of implant hardness
- Alternative hypothesis (Ha): There is a difference among the Interaction effect between Dentist and Method levels in terms of implant hardness

	df	sum_sq	mean_sq	F	PR(>F)
C(Method)	2.0	148472.177778	74236.088889	10.854287	0.000284
C(Dentist)	4.0	106683.688889	26670.922222	3.899638	0.011484
C(Method):C(Dentist)	8.0	185941.377778	23242.672222	3.398383	0.006793
Residual	30.0	205180.000000	6839.333333	NaN	NaN

Fig: -21

For, the method and dentist P-value is 0.006793 which is lesser than alpha i.e., 0.05. Hence, we have enough evidence to reject null hypothesis and consider there is a difference in means among the Interaction effect between Dentist and Method levels in terms of implant hardness for Alloy 1.

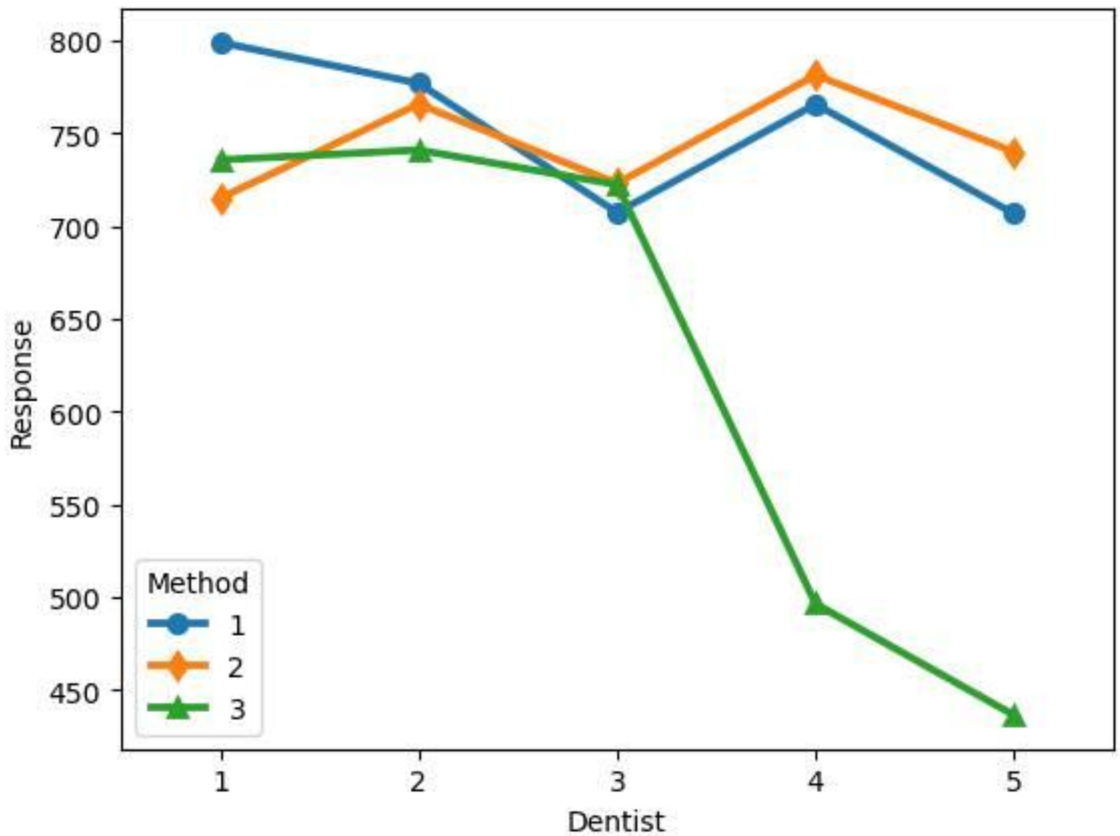

Fig: -22

Note: It could be noticed that there is interaction between dentist and method feature.

	df	sum_sq	mean_sq	F	PR(>F)
C(Method)	2.0	499640.400000	249820.200000	19.461218	0.000004
C(Dentist)	4.0	56797.911111	14199.477778	1.106152	0.371833
C(Method):C(Dentist)	8.0	197459.822222	24682.477778	1.922787	0.093234
Residual	30.0	385104.666667	12836.822222	NaN	NaN

Fig: -23

As p-value (0.09323) > alpha (0.05)

We failed to reject the null hypothesis. As there is no significant interaction between Dentist and Method considering only alloy 2.

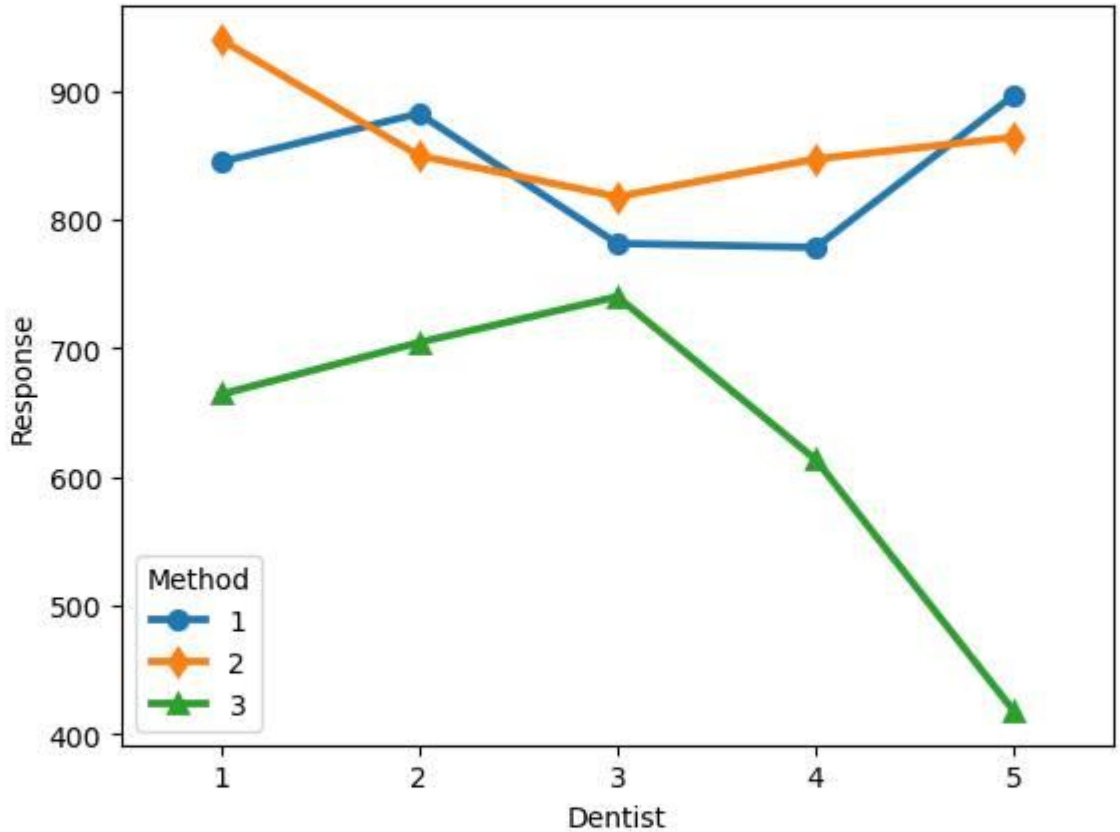

Fig: -24

Note: There is no as such significant interaction between method 3 and dentist feature for alloy 2. As the method 1 and method 2 has the interaction in between.

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

Solution: Hypothesis Test alloy1

- Null hypothesis (H0): There is no difference among the factors Dentist and Method levels in terms of implant hardness.
- Alternative hypothesis (Ha): There is a difference among the factors Dentist and Method levels in terms of implant hardness

	df	sum_sq	mean_sq	F	PR(>F)
C(Dentist)	4.0	106683.688889	26670.922222	2.591255	0.051875
C(Method)	2.0	148472.177778	74236.088889	7.212522	0.002211
Residual	38.0	391121.377778	10292.667836	NaN	NaN

Fig: -25

P value for Method is 0.0022 which is less than 0.05. Hence, we have enough evidence to reject null hypothesis and consider that at least one pair of Method means is different for Alloy 1.

Calculate two-way anova for interaction between dentist and method

For alloy 1:

	df	sum_sq	mean_sq	F	PR(>F)
C(Dentist)	4.0	106683.688889	26670.922222	3.899638	0.011484
C(Method)	2.0	148472.177778	74236.088889	10.854287	0.000284
C(Dentist):C(Method)	8.0	185941.377778	23242.672222	3.398383	0.006793
Residual	30.0	205180.000000	6839.333333	NaN	NaN

Fig: -26

P value for Dentist, Method, and Interaction variable (Dentist: Method) are all less than 0.05. Hence, we can conclude that to reject null hypothesis and consider there is at least one pair of variables means are different for Alloy 1.

Considering alpha= 0.05

Interaction	p-value
Dentist- Method	0.006793
Dentist- Temp	0.862862
Method –Temp	0.898357

Table: -2

Only the interaction between Dentist and Method is significant. Other then there is no significant interaction between variables for alloy 1.

Hypothesis Test for Alloy 2

- Null hypothesis H0 (for Alloy 2): There is no difference among the factors Dentist and Method levels in terms of implant hardness for Alloy 2.
- Alternative hypothesis Ha (for Alloy 2): There is a difference among the factors Dentist and Method levels in terms of implant hardness for Alloy 2.

	df	sum_sq	mean_sq	F	PR(>F)
C(Dentist)	4.0	56797.911111	14199.477778	0.926215	0.458933
C(Method)	2.0	499640.400000	249820.200000	16.295479	0.000008
Residual	38.0	582564.488889	15330.644444	NaN	NaN

Fig: -27

P value for Method is 0.000008 which is less than 0.05. Hence, we have enough evidence to reject null hypothesis and consider that at least one pair of Method means is different for Alloy 2

Calculate two-way Anova for interaction between dentist and method

For alloy 2:

	df	sum_sq	mean_sq	F	PR(>F)
C(Dentist)	4.0	56797.911111	14199.477778	1.106152	0.371833
C(Method)	2.0	499640.400000	249820.200000	19.461218	0.000004
C(Dentist):C(Method)	8.0	197459.822222	24682.477778	1.922787	0.093234
Residual	30.0	385104.666667	12836.822222	NaN	NaN

Fig: -28

Considering alpha= 0.05

Interaction	p-value
Dentist- Method	0.093234
Dentist- Temp	0.825318
Method –Temp	0.675983

Table: -3

There is no significant interaction between variables for alloy 2.

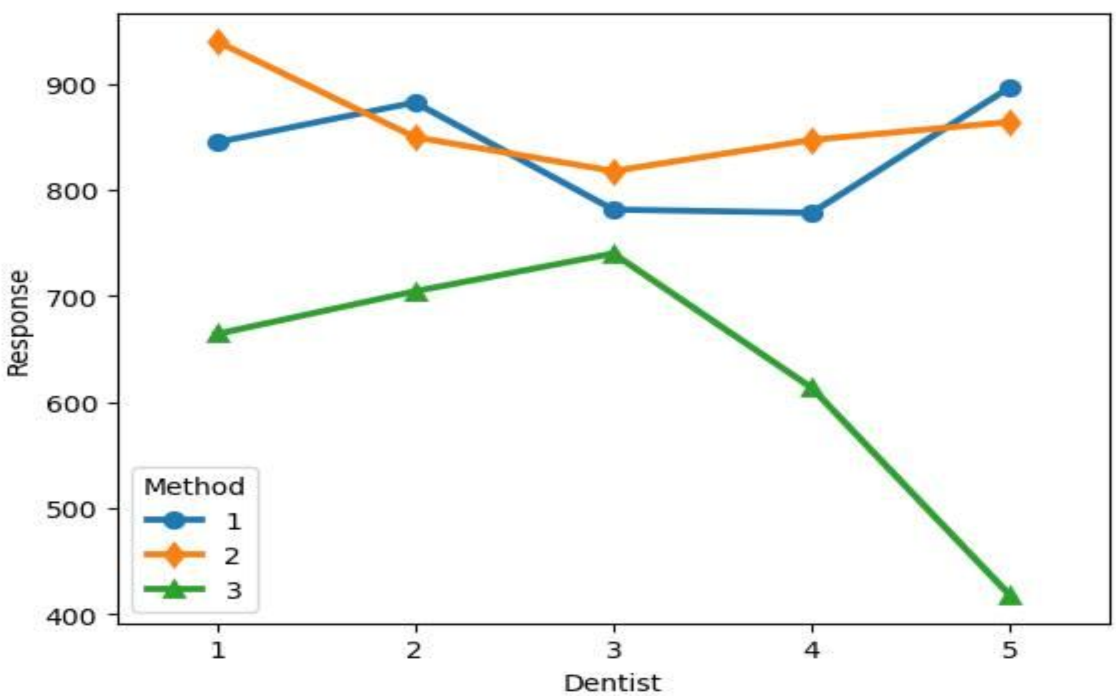

Fig: -29

With above interaction plot, we can conclude that:

- Method 1 & Method 2 has interaction among them. While Method 3 has no interactions with others.

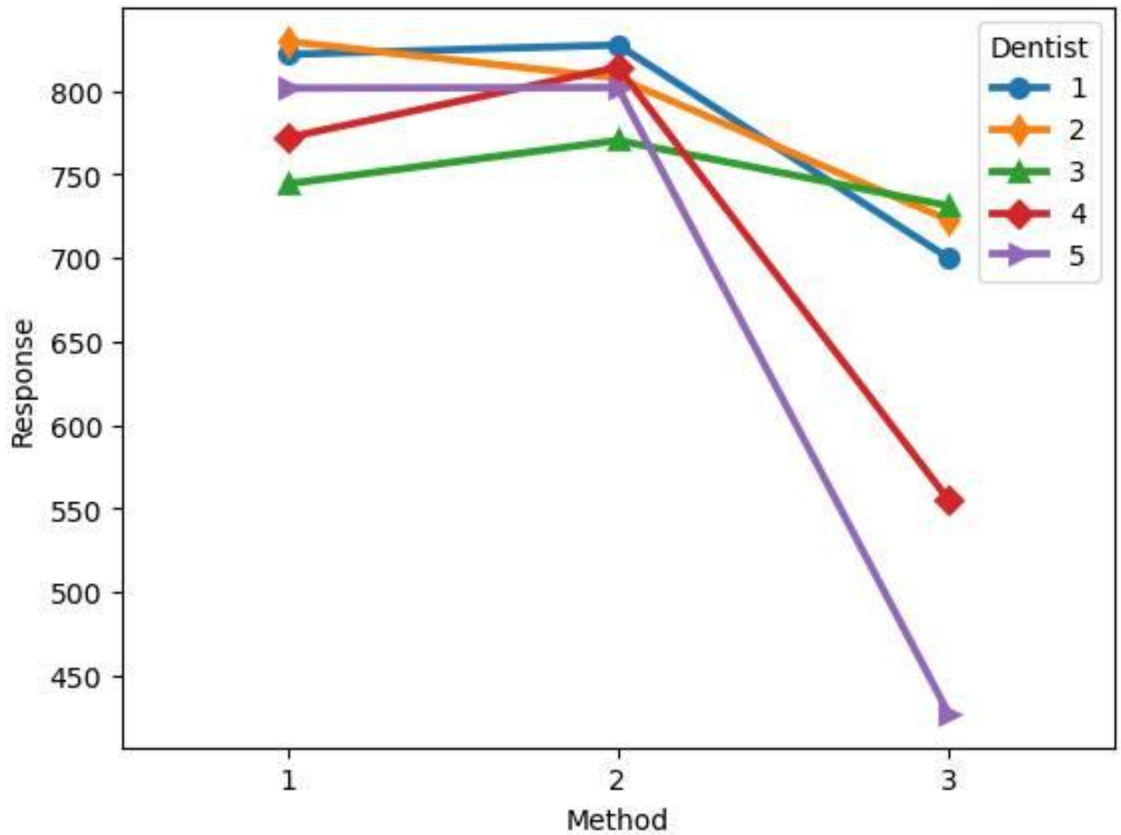

Fig: -30

We can see that there is some sort of interaction between the Methods used.

END