ADVANCED STATISTICS

-Aisha Khan

Table of Contents

List of Tables

Table 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

Table 1. Player Details

	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

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

- **Total No of Players=P(A)=235**
- **No of injured players=P(B)=145**
- **Probability of a randomly chosen player who would suffer an injury is: P(B)/P(A)=0.62**

Hence, the probability of a random chosen player who would suffer an injury is 0.62

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

- **Sum of Forward & Winger player=P(B)=94+29=123**
- **Total No of Players=P(A)=235**
- **Probability that a player is a forward or a winger is: P(B)/P(A)= 0.52**

Hence, the probability that a player is a forward or a winger 0.52

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

- **Value of Striker position player= P(B)=45**
- **Total No of players =P(A)=235**
- **Probability that a randomly chosen player plays in a striker position and has a foot injury is: P(B)/P(A)=0.19**

Probability of randomly chosen player in a striker position with foot injury is 0.19

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

- **Count of injured players= P(A)=145**
- **Players at striker position=P(B)=45**
- **Probability that a randomly chosen injured player is a striker is: P(B)/P(A)=0.31**

Probability of randomly chosen injured player who is a striker is 0.31

1.5 What is the probability that a randomly chosen injured player is either a forward or an attacking midfielder?

- **Sum of randomly chosen player who is either forward on an attacking midfielder P(B)= 24+56=80**
- **Total no of players who are injured=P(A)=145**
- **Probability that a randomly chosen injured player is either a forward or an attacking midfielder is P(B)/P(A) is 0.55**

Probability of randomly chosen injured player who is either forward or attacking midfielder is 0.55

Problem 2:

An independent research organization is trying to estimate the probability that an accident at a nuclear power plant will result in radiation leakage. The types of accidents possible at the plant are, fire hazards, mechanical failure, or human error. The research organization also knows that two or more types of accidents cannot occur simultaneously.

According to the studies carried out by the organization, the probability of a radiation leak in case of a fire is 20%, the probability of a radiation leak in case of a mechanical 50%, and the probability of a radiation leak in case of a human error is 10%. The studies also showed the following;

- **The probability of a radiation leak occurring simultaneously with a fire is 0.1%.**
- **The probability of a radiation leak occurring simultaneously with a mechanical failure is 0.15%.**
- **The probability of a radiation leak occurring simultaneously with a human error is 0.12%.**

Based on the information available, answer the questions below:

2. **From the above data we can assume that:**
   - **Probability of radiation leak (R) given Fire (F) is P(R/F) =0.2**
   - **Probability of radiation leak (R) given Mechanical failure (F) is P(R/M) =0.5**
   - **Probability of radiation leak (R) given Human Error (F) is P(R/H) =0.1**
   - **Probability of radiation leak (R) with Fire (F) is P(R∩F) =0.001**
   - **Probability of radiation leak (R) with mechanical fire (F) is P(R∩M) =0.0015**
   - **Probability of radiation leak (R) with human error (F) is P(R∩H) =0.0012**

2.1 What are the probabilities of a fire, a mechanical failure, and a human error respectively?

The probabilities are as follows:

P(R/F)	**= P(R∩F)/P(F)**
P(F)	**= P(R∩F)/P(R/F)**
	= 0.005
P(M)	**= P(R∩M)/P(R/M)**
	= 0.003
P(H)	**= P(R∩H)/P(R/H)**
	= 0.012

From the above table we can conclude that:

Probability of Fire P(F) = 0.5% (0.005)

Probability of mechanical failure P(M) =0.03% (0.003)

Probability of human error P(H) = 1.2% (0.012)

2.2 What is the probability of a radiation leak?

As per the total probability theorem:
(Note: The probability values of fire, mechanical failure and human error are considered from the above (2.1) conclusions:

P(R)= P(F)*P(R/F)+P(M)*P(R/M)+P(H)*P(R/H)
P(R)= 0.0037

The probability of a radiation leak is 0.3% (0.0037)

2.3 Suppose there has been a radiation leak in the reactor for which the definite cause is not known. What is the probability that it has been caused by:

As per Bayes' Theorem the probabilities are as follows:

P(F/R) = (P(R/F)/P (R))*P(F)

0.27027

P(M/R) = (P(R/M)/P (R))*P(M)

0.405405

P(H/R) = (P(R/H)/P(R))*P(H)

0.324324

From the above table we can conclude that:

Probability of radiation leak caused by Fire is 27% (0.27027)

Probability of radiation leak caused by mechanical failure is 40.5% (0.405405)

Probability of radiation leak caused by human error is 32.4% (0.324324)

Problem 3:

The breaking strength of gunny bags used for packaging cement is normally distributed with a mean of 5 kg per sq. centimetre and a standard deviation of 1.5 kg per sq. centimetre. 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.

Given Data:

- **Mean value of the gunny bags =μ=5 kg per sq.cm**
- **Standard deviation of the gunny bags= σ= 1.5 kg per sq.cm**

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

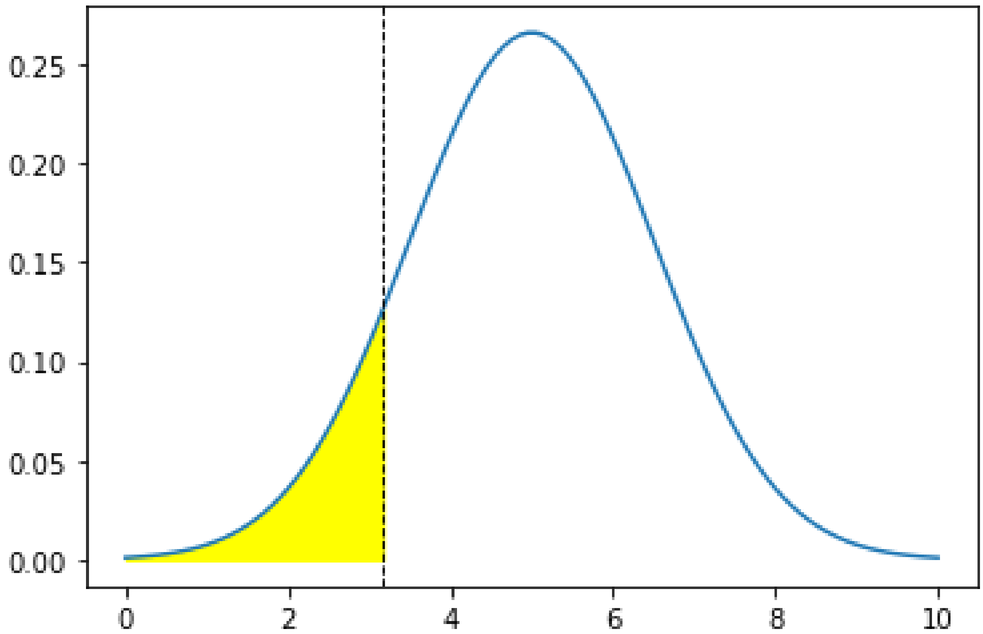

Fig 1. Probability of Gunny bags less than 3.17 kg

- **Considering the values of μ & σ as 5 & 1.5 respectively as mentioned above**
- **Using CDF(Cumulative distribution function) , the above graph indicates that there is a chance of 11.1%(~11.1232437) that the gunny bags have a breaking strength less than 3.17 kg per sq cm.**

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

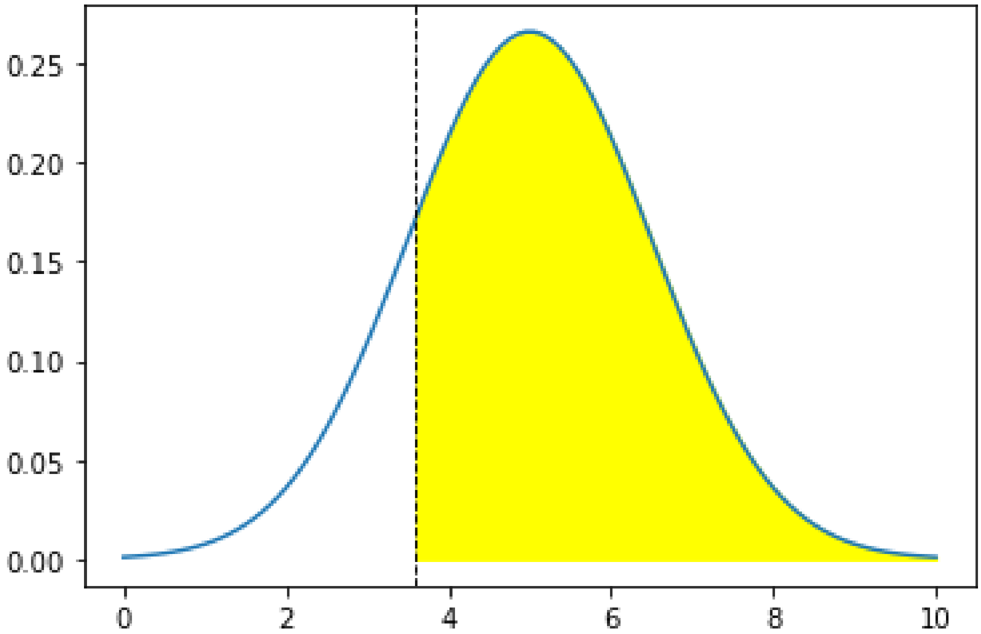

Fig 2. Probability of Gunny bags atleast 3.6

- **Considering the values of μ & σ as 5 & 1.5 respectively as mentioned above.**
- **Using the CDF (Cumulative distribution function), the above graph indicates that the probability of the gunny bags having a breaking strength of at least 3.6kg per sq. cm is 82.5%.**

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

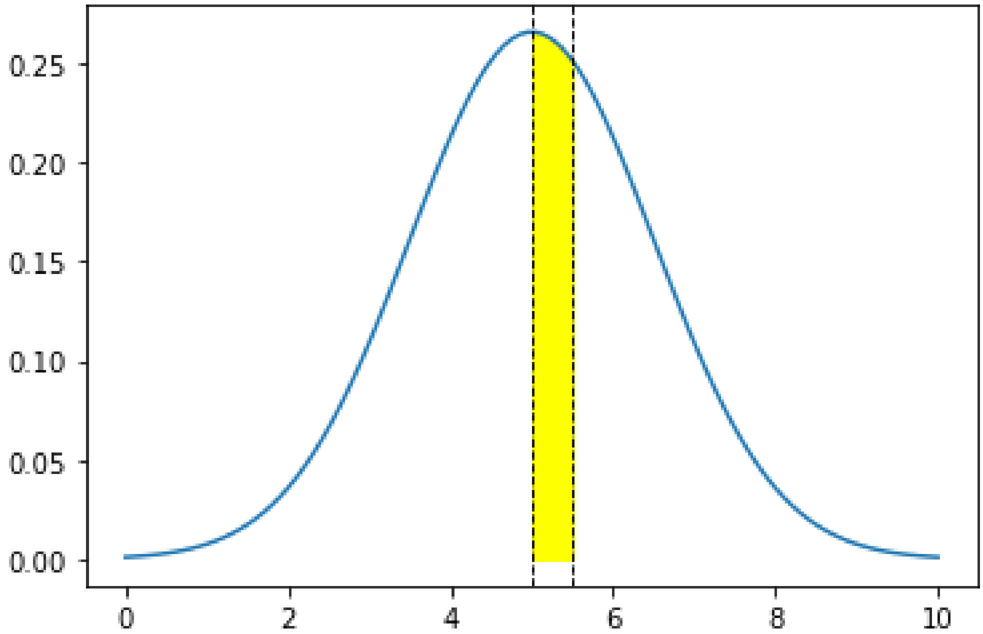

Fig 3. Probability of Gunny bags between 5 & 5.5

- **Considering the values of μ & σ as 5 & 1.5 respectively as mentioned above.**
- **Using the CDF (Cumulative Distribution Function),the above graph indicates that the proportion of gunny bags having a breaking strength between 5 and 5.5kg per sq cm is 13.05%**

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

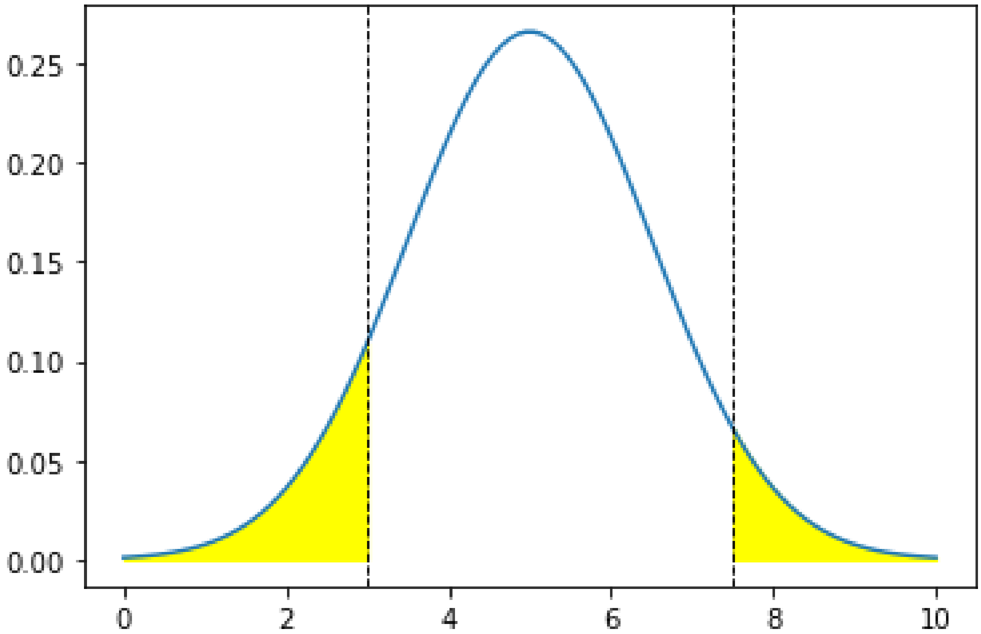

Fig 4. Probability of Gunny bags NOT between 3 and 7.5

- **Considering the values of μ & σ as 5 & 1.5 respectively as mentioned above.**
- **Using the CDF (Cumulative Distribution Function), the above graph indicates that the proportion of gunny bags having a breaking strength not between 3 and 7.5kg per sq. cm is 13.90%**

Problem 4:

Grades of the final examination in a training course are found to be normally distributed, with a mean of 77 and a standard deviation of 8.5. Based on the given information answer the questions below.

Given Data:

- Mean value of grades for the final examination in a training course=μ=77.
- Standard deviation of grades for the final examination in a training course= σ =8.5.

4.1 What is the probability that a randomly chosen student gets a grade below 85 on this exam?

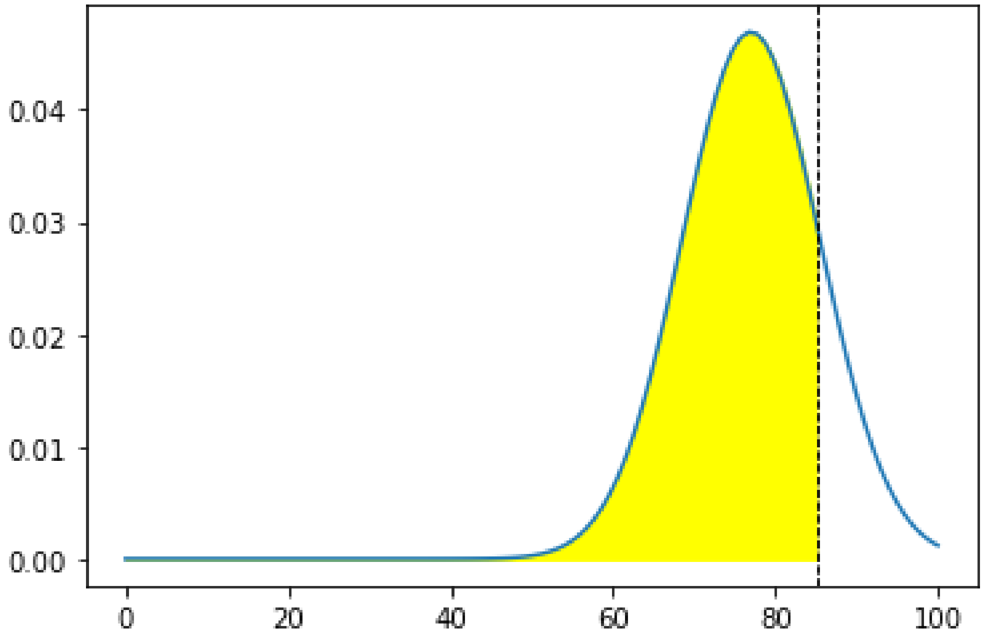

Fig 5. Students below 85 grade

- **Considering the values of μ & σ as 77 & 8.5 respectively.**
- **The above graph indicates that the probability of a randomly chosen student who gets a grade below 85 is 82.67%**

4.2 What is the probability that a randomly selected student score between 65 and 87?

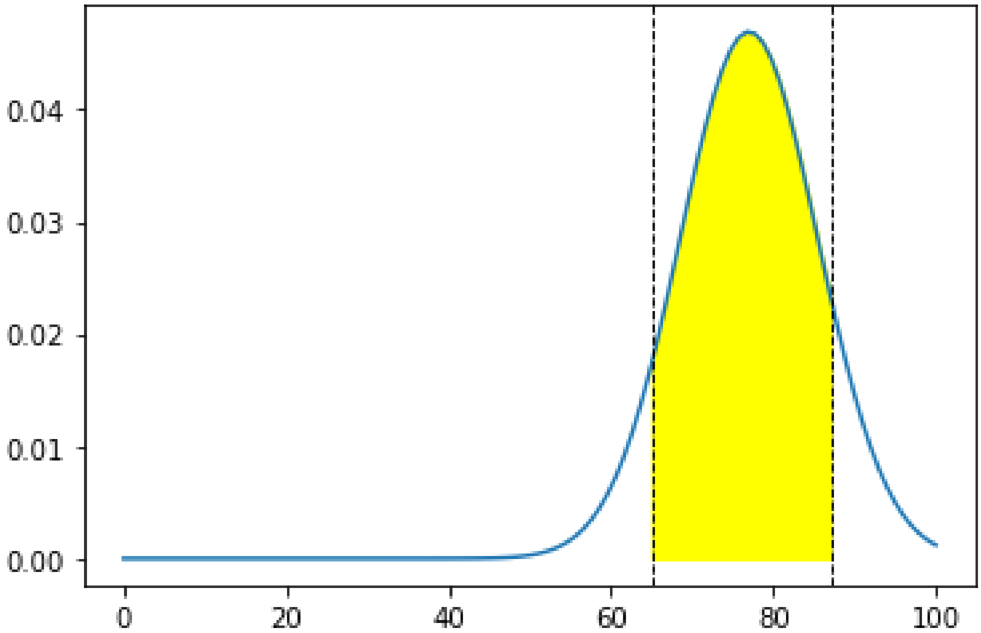

Fig 6. Students between 65 and 87grade

- **Considering the values of μ & σ as 77 & 8.5 respectively.**
- **The above graph indicates that the probability of a randomly selected student who scores between 65 & 87 is 80.13%.**

4.3 What should be the passing cut-off so that 75% of the students clear the exam?

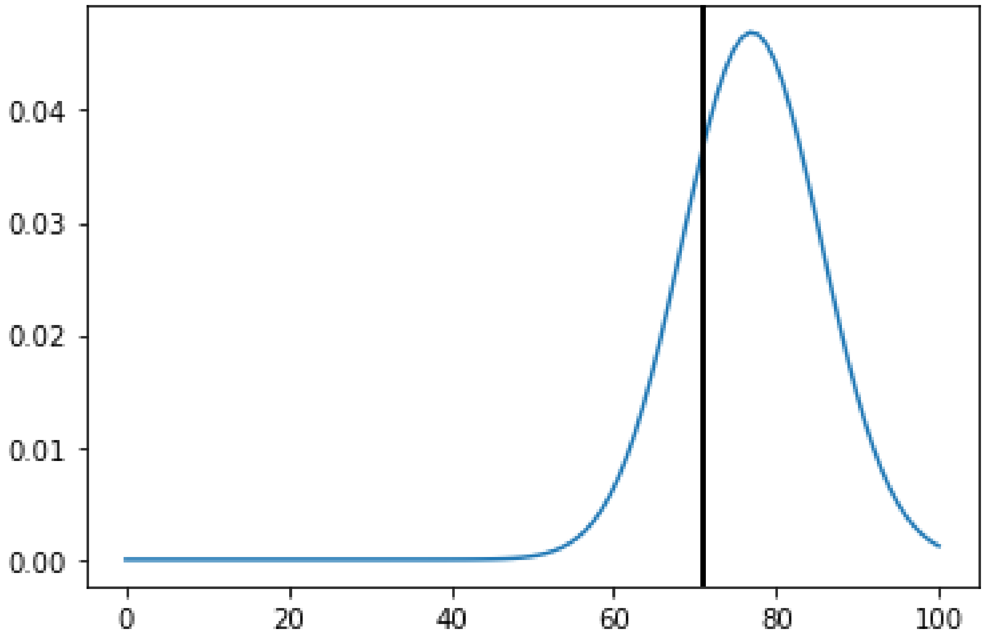

Fig 7. Passing cutoff with 75% passed students

- **Considering the values of μ & σ as 77 & 8.5 respectively.**
- **The above graph indicates that the passing cut off should be 71 in order to have 75% students clear the exam.**

Problem 5:

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);

Observations:

- There are 75 rows and 2 variables.
- No missing values are observed.
- The variables are of float64 datatype.
- The level of significance is defined as 5% (0.05%).

Table 2. Data info

Factors	Non - Null count	Data type
Unpolished	75 non-null	float64
Treated and Polished	75 non-null	float64

The above data can be described as:

Table 3. Data Description

Factors	count	mean	std	min	25%	50%	75%	max
Unpolished	75	134.1105	33.0418	48.40684	115.3298	135.5971	158.2151	200.1613
Treated and Polished	75	147.7881	15.58736	107.5242	138.2683	145.7213	157.3733	192.2729

5.1 Earlier experience of Zingaro with this particular client is favourable as the stone surface was found to be of adequate hardness. However, Zingaro has reason to believe now that the unpolished stones may not be suitable for printing. Do you think Zingaro is justified in thinking so?

Observations:

H_o: The Hardness of unpolished stones are 150

H_1: The Hardness of unpolished stones are not 150.

- Standard deviation is unknown which is why T-test of 1 sample is used to derive to conclusion.
- It is observed that the probability value is ~0.000083 which is less than the level of significance defined (5%).
- Hence, we reject the unpolished stones from the client.

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

- H_O: $\mu_{polished}=\mu_{unpolished}$
- H_A: $\mu_{polished} \neq \mu_{unpolished}$
- At 5% significance, it is observed that the mean values are not same.
- Using two-sample t-test, it is observed that p_value is ~0.0014655 which is less that the level of significance defined (5%).
- Since the defined p-value is less than the level of significance we reject the null hypothesis.
- Hence, we can say that mean hardness of the polished and unpolished stones is not the same.

Problem 6:

Aquarius health club, one of the largest and most popular cross-fit gyms in the country has been advertising a rigorous program for body conditioning. The program is considered successful if the candidate is able to do more than 5 push-ups, as compared to when he/she enrolled in the program. Using the sample data provided can you conclude whether the program is successful? (Consider the level of Significance as 5%)

Note that this is a problem of the paired-t-test. Since the claim is that the training will make a difference of more than 5, the null and alternative hypotheses must be formed accordingly.

Observations:

From the above dataset it is observed:

- It includes 100 rows and 3 variables.
- There are no missing values.
- The variables are integer datatype.

Table 4. Data Info

Factors	Count	Non-Null	Dtype
Sr No	100	non-null	int64
Before	100	non-null	int64
After	100	non-null	int64

The above data can be described as:

Table 5. Data Description

Factors	count	mean	std	min	25%	50%	75%	max
Sr no.	100	50.5	29.01149	1	25.75	50.5	75.25	100
Before	100	26.94	8.806357	3	21.75	28	32.25	47
After	100	32.49	8.779562	10	26	34	39	51

- **H_0: Training will make a difference of more than 5(Weight difference>5)**
- **H_1: Training will not make a difference of more than 5(Weight difference<=5).**
- **Using two paired sample t-test, it is observed that the p_value of the test is less than the level of significance defined (5%).**
- **Hence we have enough evidence to say that program for body conditioning is Failed, as the null hypothesis is rejected.**

Problem 7**:**

Dental implant data: The hardness of metal implant 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 on the dentists who may favour one method above another and may work better in his/her favourite method. The response is the variable of interest.

Observations:

From the data set it is observed that:

- **There are 90 rows and 5 variables.**
- **There are no missing values.**
- **The variables are integer datatype.**

Table 6. Data Info

Column	Data type	Non-Null count
Dentist	Int64	90 non-null
Method	Int64	90 non-null
Alloy	Int64	90 non-null
Temp	Int64	90 non-null
Response	Int64	90 non-null

The above data can be described as:

Table 7. Data Description

Variables	count	mean	std	min	25%	50%	75%	max
Dentist	90	3	1.422136	1	2	3	4	5
Method	90	2	0.821071	1	1	2	3	3
Alloy	90	1.5	0.502801	1	1	1.5	2	2
Temp	90	1600	82.10708	1500	1500	1600	1700	1700
Response	90	741.7778	145.7678	289	698	767	824	1115

7.1 Test whether there is any difference among the dentists on the implant hardness. State the null and alternative hypotheses. Note that both types of alloys cannot be considered together. You must state the null and alternative hypotheses separately for the two types of alloys?

<u>Observations:</u>

- **The data indicates 2 different types of alloys in alloy variable.**
- **The variable Dentist has 5 different types.**

Defining the types as Alloy1 & Alloy 2 we can assume the hypothesis as:

For Alloy 1:
H_0 : There is no difference among the dentist on the implant hardness.
The null hypothesis can be considered as the mean responses are equal among the dentist.

$$\boxed{\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5}$$

H_1 : There is difference among the dentist on the implant hardness. The alternate hypothesis can be considered as at least one of the mean responses are not equal among the dentist.

Table 8. ANOVA table

Factors	df	sum_sq	mean_sq	F	PR(>F)
C(Dentist)	4	106683.7	26670.92	1.977112	0.116567
Residual	40	539593.6	13489.84	NaN	NaN

From the above table, it is observed that:

- **The probability value (11.7%) is higher that the level of significance (5%) which confirms that we fail to reject null hypothesis.**
- **The mean responses of different dentists are same for Alloy1.**

<u>Conclusion:</u>

There is no difference among the dentists on implant hardness.

For Alloy 2:

H_0 : There is no difference among the dentist on the implant hardness.
The null hypothesis can be considered as the mean responses are equal among the dentist.

$$\boxed{\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5}$$

H_1 : There is difference among the dentist on the implant hardness. The alternate hypothesis can be considered as at least one of the mean responses are not equal among the dentist.

*Table 9. ANOVA Table_*

Factors	df	sum_sq	mean_sq	F	PR(>F)
C(Dentist)	4	5.68E+04	14199.48	0.524835	0.718031
Residual	40	1.08E+06	27055.12	NaN	NaN

From the above table, it is observed that:

- **The probability value (71.8%) is higher that the level of significance (5%) which confirms that we fail to reject null hypothesis.**
- **The mean responses of different dentists are same for Alloy2.**

Conclusion:

There is no difference among the dentists on implant hardness.

7.2 Before the hypotheses may be tested, state the required assumptions. Are the assumptions fulfilled? Comment separately on both alloy types.?

As per the dataset, the ANOVA assumptions are as follows:

1. **"Response" variable of the population is continuous and normally distributed.**

We have applied Shapiro-Wilke Test considering the dataset to be small.
H_0: The data is drawn from a normal distribution.
H_1: The data is not drawn from a normal distribution.

Shapiro-Wilke Test Result

```
ShapiroResult(statistic=0.8304629921913147,
pvalue=1.1945070582441986e-05)
ShapiroResult(statistic=0.8877691626548767,
pvalue=0.0004029277188237756)
```

2. Variances of all the populations are equal or approximately equal.

Considering Levene test to check homogeneity.

H_0: Variances of Alloy 1 and Alloy 2 are equal.
H_a: Variances of Alloy 1 and Alloy 2 are not equal.

```
LeveneResult(statistic=1.4194717470917784, pvalue=0.23669380462584474)
```

- **From the above snippet, it is observed that the p-value is greater than the level of significance (5%)**
- **We can conclude that the variances are equal. as, we fail to reject Null hypothesis.**

3. Alloy 1 & Alloy 2 randomly & independently selected populations.

7.3 Irrespective of your conclusion in 7.2, we will continue with the testing procedure. What do you conclude regarding whether implant hardness depends on dentists? Clearly state your conclusion. If the null hypothesis is rejected, is it possible to identify which pairs of dentists differ?

- **As the mean hardness responses of the dentists are same irrespective of the alloy, it is evident that implant hardness does not depend on dentists as the mean responses are equal.**
- **From the above, it is observed that we failed to reject null hypothesis and there is difference between the dentists regarding implantation hardness.**
- **Using ANOVA, we are unable to conclude on which pair of dentists differ.**

Assuming Null hypothesis is rejected:

- **Using the below plot we can assume that, using different alloys we have a slight dip in mean responses for 4th and 5th Doctors.**
- **Hence, we can conclude that there is a difference in the means which is inaccurate.**

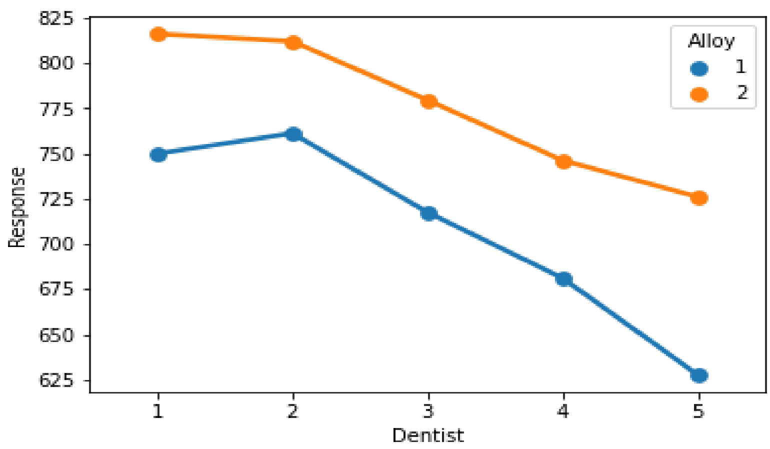

Fig 8 Response vs Dentist

7.4 Now test whether there is any difference among the methods on the hardness of dental implant, separately for the two types of alloys. What are your conclusions? If the null hypothesis is rejected, is it possible to identify which pairs of methods differ?

From the above condition we can derive the hypothesis as:

H_0: There is no difference among the methods on the hardness of dental impact.

$$\mu 1 = \mu 2 = \mu 3$$

H_1: There is difference among the methods on the hardness of dental impact.

<u>Observation for Alloy 1:</u>

Using ANOVA, we have derived to the below conclusion:

- **As the p- value is less than 0.05% null hypothesis is rejected.**
- **There is a mean difference among the methods on the hardness of dental implant.**

Table 10. ANOVA Table Method

Factors	df	sum_sq	mean_sq	F	PR(>F)
C(Method)	2	83943.64444	41971.82	5.693239	0.006489
Residual	42	309633.3333	7372.222	NaN	NaN

- **As the null hypothesis is rejected, we cannot say which pair of methods differ considering ANOVA.**
- **However, using the below graph, we can assume that the pair of method which differ:**

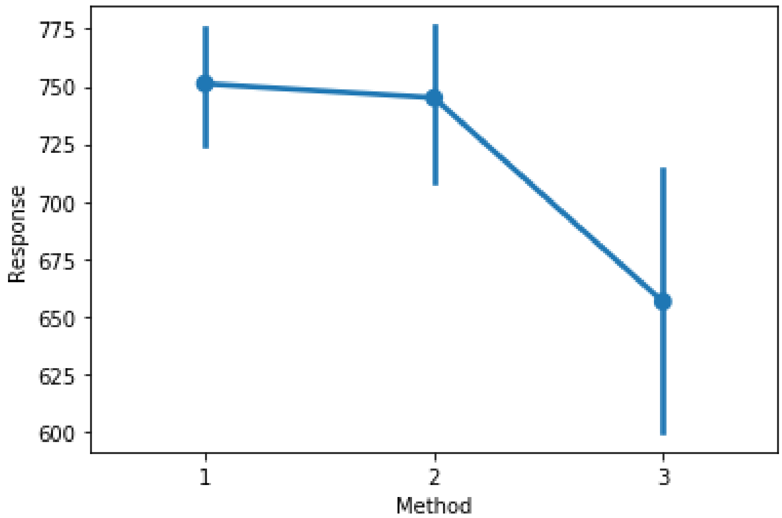

Fig 9. Alloy_1, Method vs Response

- There is a significant dip in between 2nd & 3rd pair of methods.

<u>Observation for Alloy 2:</u>

Using ANOVA, we have derived to the below conclusion:

- **As the p- value is less than 0.05% null hypothesis is rejected.**
- **There is a mean difference among the methods on the hardness of dental implant.**

Table 11. ANOVA Table Method

	df	sum_sq	mean_sq	F	PR(>F)
C(Method)	2	344333.6444	172166.8	20.04204	7.74E-07
Residual	42	360792	8590.286	NaN	NaN

- **As the null hypothesis is rejected, we cannot say which pair of methods differ considering ANOVA.**
- **However, using the below graph, we can assume that the pair of method which differ:**

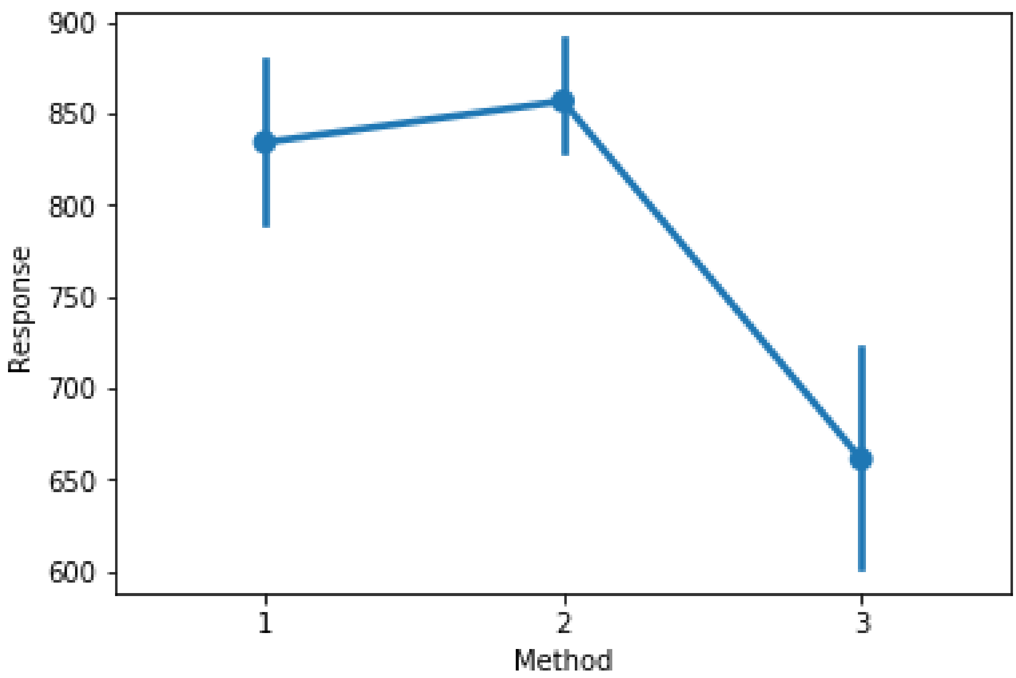

Fig 10. Alloy_2, Method vs Response

- **Significant dip can be observed in 2nd & 3rd pair of method.**

7.5 Now test whether there is any difference among the temperature levels on the hardness of dental implant, separately for the two types of alloys. What are your conclusions? If the null hypothesis is rejected, is it possible to identify which levels of temperatures differ?

H_0: There is no difference among the temperature on the hardness of dental impact.

$$\boxed{\mu 1 = \mu 2 = \mu 3}$$

H_1: There is difference among the temperature on the hardness of dental impact.

<u>Observation for Alloy 1:</u>

Using ANOVA, we have derived to the below conclusion:

- **As the p- value is greater than 0.05% we fail to reject the null hypothesis.**
- **There is no mean difference among the temperature on the hardness of dental implant.**

Table 12. ANOVA table Temp

	df	sum_sq	mean_sq	F	PR(>F)
C(Temp)	2	4603.51111	2301.756	0.248536	0.781081
Residual	42	388973.467	9261.273	NaN	NaN

- **We cannot say which pair of methods differ considering ANOVA, though we fail to reject the null hypothesis.**
- **However, using the below graph, we can assume that the pair of temperature which differs:**

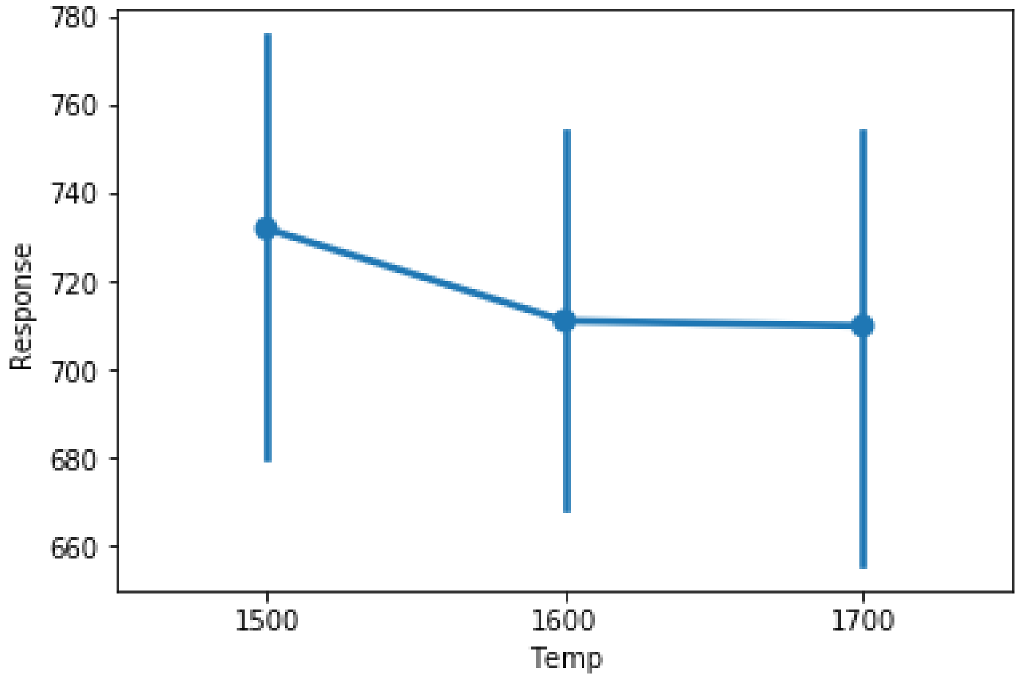

Fig 11. Alloy_1, Temp vs Response

- As we have failed to reject null hypothesis along with no difference between the mean values of temperatures in the above graph, there is no mean difference among the temperatures.

Observation for Alloy 2:

Using ANOVA, we have derived to the below conclusion:

- **As the p- value is greater than 0.05% we fail to reject the null hypothesis.**
- **There is no mean difference among the temperature on the hardness of dental implant.**

Table 13. ANOVA table Temp

	df	sum_sq	mean_sq	F	PR(>F)
C(Temp)	2	46205.9111	23102.96	1.472598	0.240927
Residual	42	658919.733	15688.57	NaN	NaN

- **We cannot say which pair of methods differ considering ANOVA, though we fail to reject the null hypothesis.**
- **However, using the below graph, we can assume that the pair of temperature which differs:**

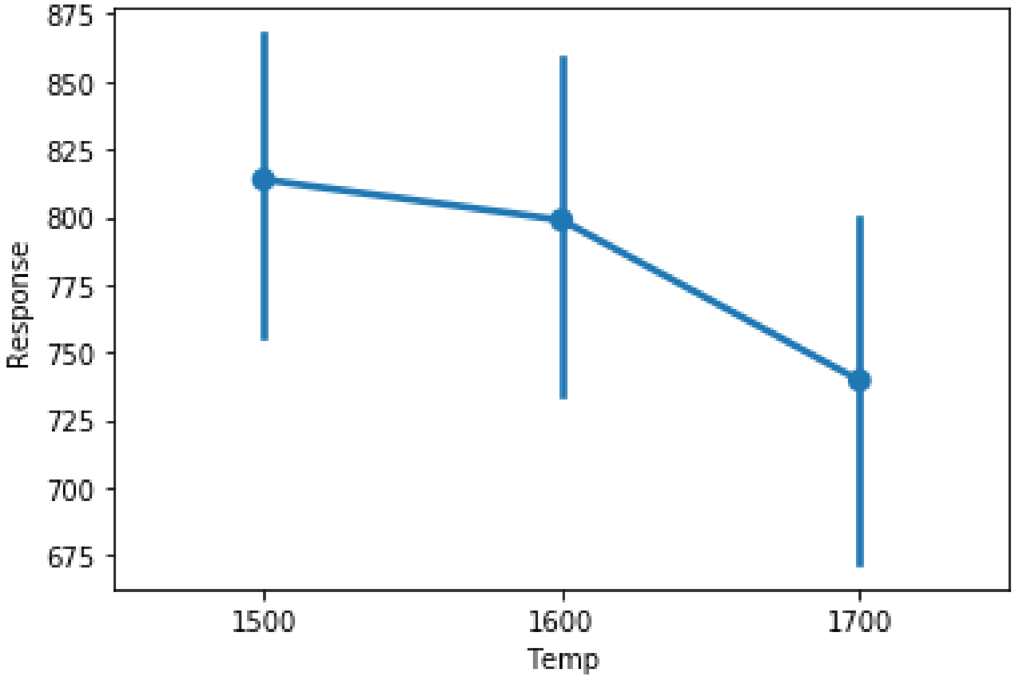

Fig 12. Alloy_2, Temp vs Response

- As we have failed to reject null hypothesis along with no difference between the mean values of temperatures in the above graph, there is no mean difference among the temperatures.

7.6 Consider the interaction effect of dentist and method and comment on the interaction plot, separately for the two types of alloys?

Let us observe the interaction plot of dentist and method based of different types of alloys:

<u>For Alloy 1:</u>

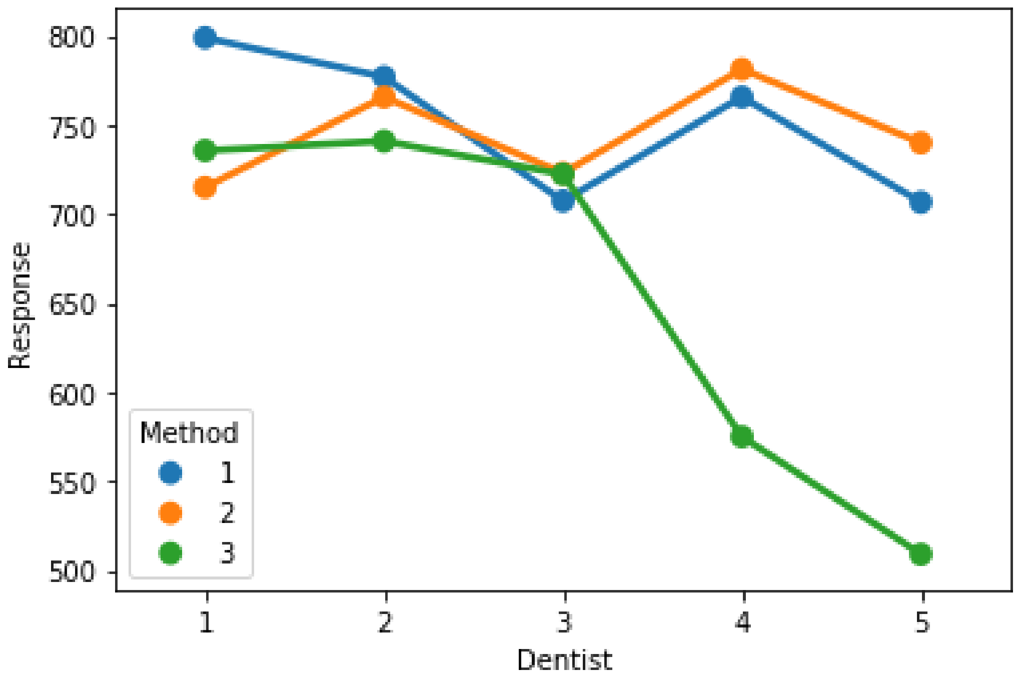

Fig 13. Alloy_1, Dentist vs Response

- **The above graph indicates an interaction between the variable's dentist and method.**

For Alloy 2:

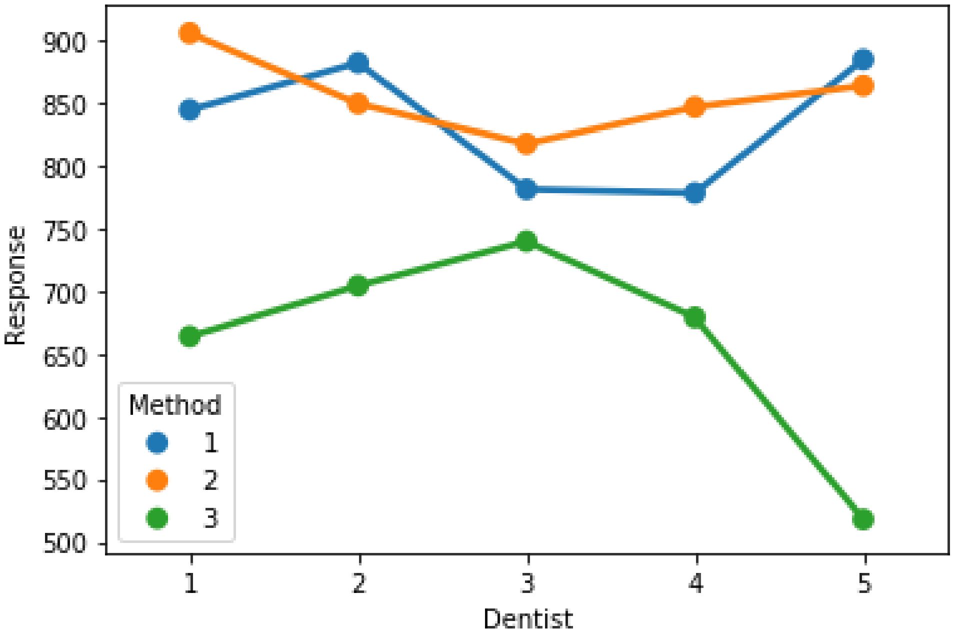

Fig 14. Alloy_2, Dentist vs Response

- **The above graph indicates no interaction between the 3rd method and the other methods.**

7.7 Now consider the effect of both factors, dentist, and method, separately on each alloy. What do you conclude? Is it possible to identify which dentists are different, which methods are different, and which interaction levels are different?

Considering both factors separately for the different types of alloy it is observed that:

For Alloy 1:

Table 14. ANOVA Table, Interaction b/w Dentist vs Method

	df	sum_sq	mean_sq	F	PR(>F)
C(Dentist)	4	66319.4222	16579.8556	3.539108	0.01762
C(Method)	2	83943.6444	41971.8222	8.959234	0.00089
C(Dentist):C(Method)	8	102771.244	12846.4056	2.742172	0.021263
Residual	30	140542.667	4684.75556	NaN	NaN

- **Above table indicates the effect of interaction between dentist and methods by using Two Way ANOVA.**
- **Due to the inclusion of interaction effect term, a slight change is observed in the p-value for the first two treatments as compared to the 2-way ANOVA without the interaction effect terms.**
- **It is also observed the p- value of interaction effect term of dentist and method suggest that the null hypothesis rejected in this case.**

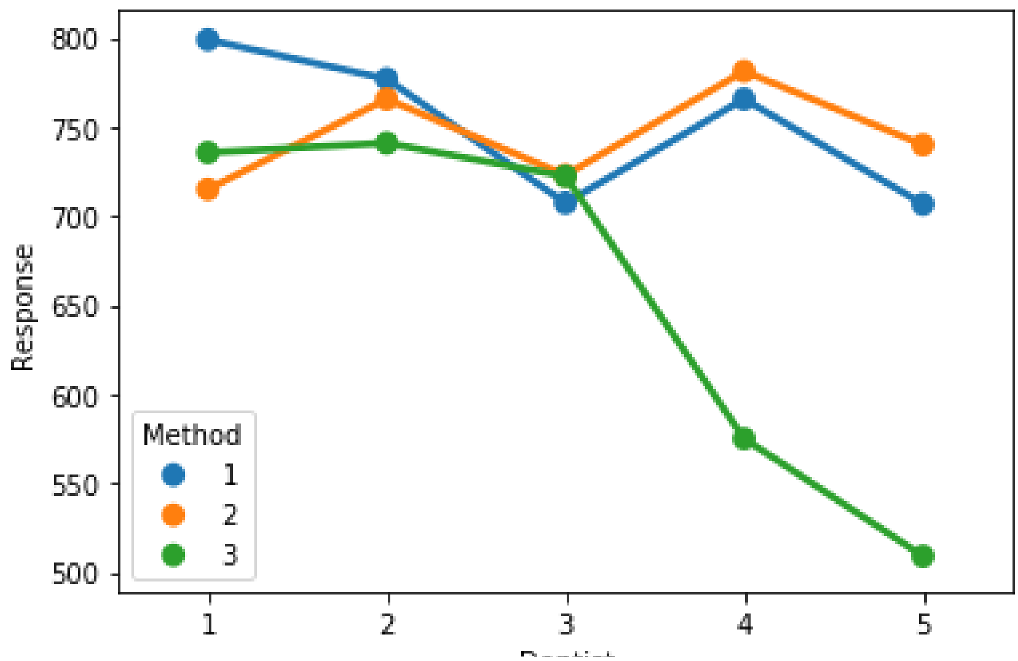

Fig 15. Alloy_1, Dentist vs Response

- ANOVA cannot conclude on the difference of pairs in dentists or methods or interaction between them.
- As per the graph above, we can assume that there is a difference between dentist 3 & dentist 4 for method 3 as there is a significant dip between them.
- There is no interaction difference between all the methods with dentists 3.

For Alloy 2:

Table 15. ANOVA Table, Interaction between Dentist vs Method

	df	sum_sq	mean_sq	F	PR(>F)
C(Dentist)	4	20704.9778	5176.24444	0.6785	0.612236
C(Method)	2	344333.644	172166.822	22.567548	0.000001
C(Dentist):C(Method)	8	111218.356	13902.2944	1.822306	0.111831
Residual	30	228868.667	7628.95556	NaN	NaN

- Above table indicates the effect of interaction between dentist and methods by using Two Way ANOVA.
- Due to the inclusion of interaction effect term, a change is observed in the p-value for the first two treatments as compared to the 2-way ANOVA without the interaction effect terms.
- It is also observed the p- value of interaction effect term of dentist and method suggest that we fail to reject the null hypothesis rejected in this case.

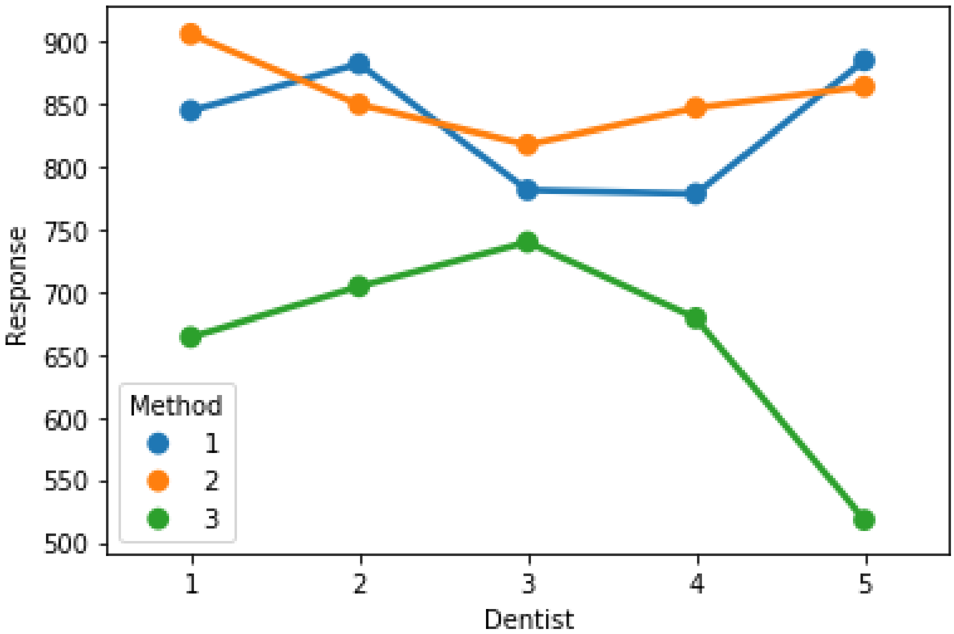

Fig 16. Alloy_2, Dentist vs Response

- **ANOVA cannot conclude on the different pairs in dentists or methods or interaction between them.**
- **As per the graph above, we can assume that there is a difference between dentist 4 & dentist 5 for method 3 as there is a significant dip between them.**
- **There is an interaction difference between method 3 with the other 2 methods.**