PGP-DSBa project report

IS – Coded Project

**BY**

**Ishaan Shakti Jayaraman**

**PGPDSBA.O.JULY24.A**

Contents

[LIST OF FIGURES 2](#_Toc177302283)

[PROBLEM 1 3](#_Toc177302284)

[1.1 What is the probability that a randomly chosen player would suffer an injury? 3](#_Toc177302285)

[1.2 What is the probability that a player is a forward or a winger? 3](#_Toc177302286)

[1.3 What is the probability that a randomly chosen player plays in a striker position and has a foot injury? 4](#_Toc177302287)

[1.4 What is the probability that a randomly chosen injured player is a striker? 4](#_Toc177302288)

[PROBLEM 2 5](#_Toc177302289)

[2.1 What proportion of the gunny bags have a breaking strength of less than 3.17 kg per sq cm? 5](#_Toc177302290)

[2.2 What proportion of the gunny bags have a breaking strength of at least 3.6 kg per sq cm.? 6](#_Toc177302291)

[2.3 What proportion of the gunny bags have a breaking strength between 5 and 5.5 kg per sq cm.? 7](#_Toc177302292)

[2.4 What proportion of the gunny bags have a breaking strength NOT between 3 and 7.5 kg per sq cm.? 8](#_Toc177302293)

[PROBLEM 3 10](#_Toc177302294)

[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? 10](#_Toc177302295)

[3.2 Is the mean hardness of the polished and unpolished stones the same? 11](#_Toc177302296)

[PROBLEM 4 12](#_Toc177302297)

[4.1 How does the hardness of implants vary depending on dentists? 12](#_Toc177302298)

[4.2 How does the hardness of implants vary depending on methods? 14](#_Toc177302299)

[4.3 What is the interaction effect between the dentist and method on the hardness of dental implants for each type of alloy? 16](#_Toc177302300)

[4.4How does the hardness of implants vary depending on dentists and methods together? 17](#_Toc177302301)

# LIST OF FIGURES

Figure – Distribution Chart of Breaking Strength < 3.17 kg/cm²

Figure – Distribution chart of Breaking Strength > 3.6 kg/cm²

Figure – Distribution Chart of Breaking Strength between 5 and 5.5 kg/cm²

Figure – Distribution Chart of Breaking Strength not between 3 and 7.5 kg/cm²

Figure – Histogram of Unpolished Stone Hardness

Figure – Interaction plot of dentist and method for Alloy 1

Figure - Interaction plot of dentist and method for Alloy 2

# 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** |

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

The probability that a randomly chosen player suffers an injury can be calculated by dividing the number of injured players by the total number of players.

Total Players Injured = 145

Total Players = 235

P(Injury) = = ≈ 0.617

The probability that a randomly chosen player will suffer an injury is approximately 61.7%

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

The probability that a player is either a Forward or a Winger can be calculated by dividing the total number of forward players and total number of winger players by the total number of players.

Total Forward Players = 94

Total Winger Players = 29

Total Players = 235

P(Forward or Winger) =

P(Forward or Winger) = = ≈ 0.523

The probability that a player is either a forward or a winger is approximately 52.3%

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

The probability that a randomly chosen player who plays in a striker position and has a foot injury can be calculated by dividing the total injured striker players by the total number of players.

Total Injured Striker Players = 45

Total Players = 235

P(Striker and Injured) = = ≈ 0.191

The probability that a randomly chosen player who plays in a striker position and has a foot injury is approximately 19.1%

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

The probability that a randomly chosen injured player is a striker can be calculated by dividing the total number of injured striker players by the total number of injured players.

Total Injured Striker Players = 45

Total Injured Players = 145

P(Injured | Striker) = = ≈ 0.310

The probability that a randomly chosen injured player is a striker is approximately 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

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

The proportion of gunny bags that have breaking strength less than 3.17 kg/cm² can be found using the cumulative distribution function of the normal distribution.

Calculating the Z Score:

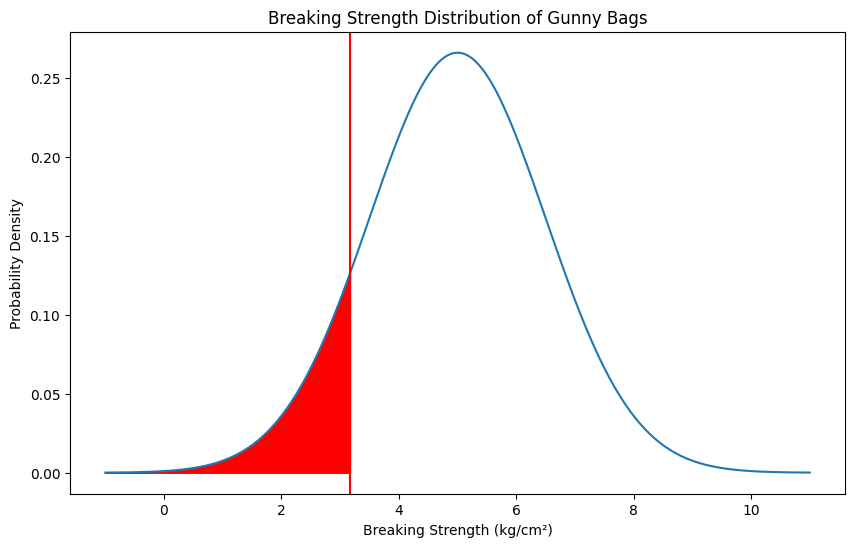
Mean (µ) = 5 kg/cm²  
Standard Deviation (σ) = 1.5 kg/cm²  
Variable (X) = 3.17 kg/cm²

Z = = = = -1.22

The CDF Value for Z = -1.22 is approximately 0.1112

Therefore, the proportion of gunny bags that have a breaking strength less than 3.17 kg/cm² is 0.1112 or 11.12%

Figure 8 – Distribution Chart of Breaking Strength < 3.17 kg/cm²



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

The proportion of gunny bags that have breaking strength of at least 3.6 kg/cm² can be found using the cumulative distribution function of the normal distribution.

Calculating the Z Score:

Mean (µ) = 5 kg/cm²  
Standard Deviation (σ) = 1.5 kg/cm²  
Variable (X) = 3.6 kg/cm²

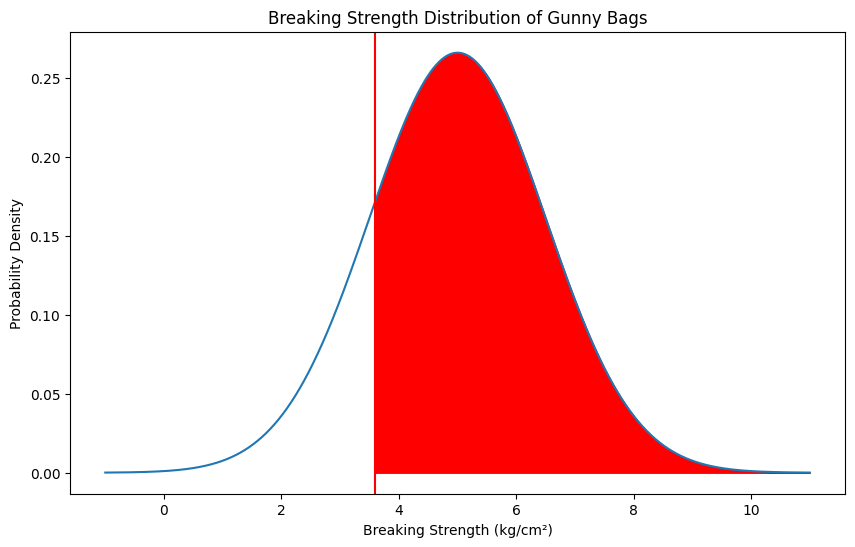
Z = = = = -0.9333

The CDF Value for Z = -0.9333 is approximately 0.1753

P(X ≥ 3.6) = 1 – CDF(-0.9333) = 1 – 0.1753 ≈ 0.8247

Therefore, the proportion of gunny bags that have a breaking strength of at least 3.6kg/cm² is 0.8247 or 82.47%

Figure 9 – Distribution chart of Breaking Strength > 3.6 kg/cm²



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

The proportion of gunny bags that have breaking strength between 5 kg/cm² and 5.5 kg/cm² can be found using the cumulative distribution function of the normal distribution.

Calculating the Z Score for X = 5 kg/cm²:

Mean (µ) = 5 kg/cm²  
Standard Deviation (σ) = 1.5 kg/cm²  
Variable (X) = 5 kg/cm²

Z = = = = 0

Calculating the Z Score for X = 5.5 kg/cm²:

Mean (µ) = 5 kg/cm²  
Standard Deviation (σ) = 1.5 kg/cm²  
Variable (X) = 5.5 kg/cm²

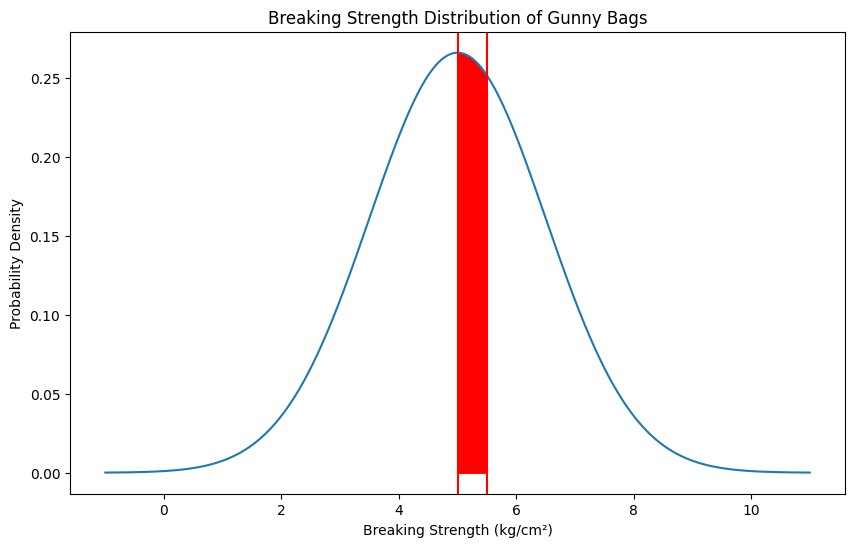
Z = = = = 0.3333

The CDF Value for Z = 0 is approximately 0.5  
The CDF Value for Z = 0.3333 is approximately 0.6306

P(5 < X < 5.5) = CDF(0.3333) – CDF(0) = 0.6306 – 0.5 = 0.1306

Therefore, the proportion of the gunny bags that have a breaking strength between 5 and 5.5 kg/cm² is 0.1306 or 13.06%

Figure 10 – Distribution Chart of Breaking Strength between 5 and 5.5 kg/cm²



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

The proportion of gunny bags that have breaking strength not between 3 kg/cm² and 7.5 kg/cm² can be found using the cumulative distribution function of the normal distribution.

Calculating the Z Score for X = 3 kg/cm²:

Mean (µ) = 5 kg/cm²  
Standard Deviation (σ) = 1.5 kg/cm²  
Variable (X) = 3 kg/cm²

Z = = = = -1.3333

Calculating the Z Score for X = 7.5 kg/cm²:

Mean (µ) = 5 kg/cm²  
Standard Deviation (σ) = 1.5 kg/cm²  
Variable (X) = 7.5 kg/cm²

Z = = = =1.6667

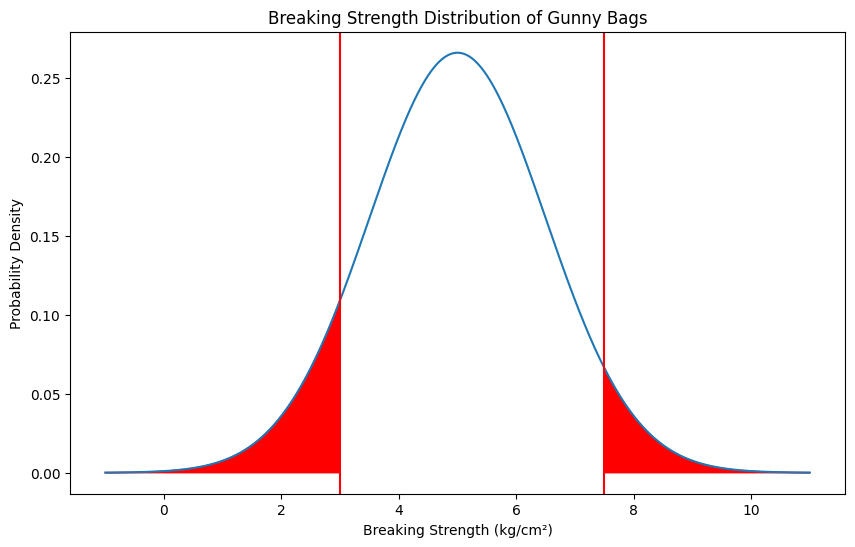
The CDF Value for Z = -1.3333 is approximately 0.0912  
The CDF Value for Z = 1.6667 is approximately 0.9522

P(3 < X < 7.5) = CDF(1.6667) – CDF(-1.3333) = 0.09522 – 0.0912 = 0.861

P(X not between 3 and 7.5) = 1 – P(3 < X < 7.5) = 1 – 0.861 = 0.139

Therefore, the proportion of gunny bags that have a breaking strength not between 3 and 7.5 kg/cm² is 0.139 or 13.9%

Figure 11 – Distribution Chart of Breaking Strength not between 3 and 7.5 kg/cm²



# 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. (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?

**Defining the Null & Alternative Hypothesis:**

Null Hypothesis (H0) = The average Brinell's hardness index of the unpolished stones is at least 150. (μ ≥ 150)

Alternative Hypothesis (Ha) = The average Brinell's hardness index of the unpolished stones is less than 150. (μ < 150)

Level of Significance (α) = 0.05

As the population standard deviation is unknown, to check if unpolished stones may be suitable for printing we can use the one sample t-test

Using stats.ttest\_1samp function of Scipy, we can determine the following

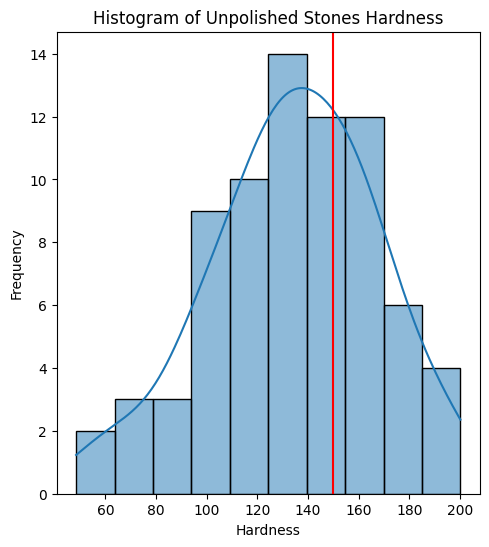
The value of the t\_statistic = -4.1646  
The value of p\_value is = 4.171286997419652e-05

**Conclusion:**We reject the null hypothesis as p value is less than the level of significance (0.05)

The mean hardness of unpolished stones is significantly less than 150.

Zingaro's concern is justified.

Figure 12 – Histogram of Unpolished Stone Hardness



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

**Defining the Null & Alternative Hypothesis:**

Null Hypothesis (H0) = The average Brinell's hardness index of the polished stones and unpolished stones is the same (μ1 = µ2)

Alternative Hypothesis (Ha) = The average Brinell's hardness index of the polished stones and unpolished stones are not same (µ1 ≠ µ2)

Level of Significance (α) = 0.05

The two sample t-test can be used to test for equality of means

Using stats.ttest\_ind function from Scipy, We can determine that

The value of t\_statistic = -3.2422  
The value of p\_value = 0.0016

**Conclusion:**

We reject the null hypothesis as the p\_value is lesser than the level of significance (0.05)

The mean hardness of polished and unpolished stones is significantly different.

# 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?

**Defining the Null & Alternative Hypothesis:**

**Alloy 1:**

Null Hypothesis (H0) = The average hardness of implants is same across all 5 dentists for Alloy 1 (μ1 = µ2 = μ3 = µ4 = μ5)

Alternative Hypothesis (Ha) = The average hardness of implants is not same for at least 1 dentist for Alloy 1

**Alloy 2:**

Null Hypothesis (H0) = The average hardness of implants is same across all 5 dentists for Alloy 2 (μ1 = µ2 = μ3 = µ4 = μ5)

Alternative Hypothesis (Ha) = The average hardness of implants is not same for at least 1 dentist for Alloy 2

**Assumptions:**

For an ANOVA hypothesis test, the assumptions are

1. **Normality:** The distribution of residuals is normal.
2. **Homogeneity of Variances:** The variance among dentists is approximately equal.
3. **Independence:** Observations are independent of each other.

**Checking the assumptions:**

We can check normality of the distribution by using the Shapiro-Wilk’s test

**Shapiro Wilk’s Test:**  
Null Hypothesis (H0) = The population from which the sample is drawn follows a normal distribution for alloy 1 or alloy 2.

Alternative Hypothesis (Ha) = The population from which the sample is drawn does not follow a normal distribution for alloy 1 or alloy 2.

**For Alloy 1:**

The p\_value for each dentist is more than the level of significance (0.05), We fail to reject the Null Hypothesis

There is evidence to suggest the sample comes from a normal distribution

**For Alloy 2:**

The p\_value for dentists 1, 2, 3 and 5 is more than the level of significance (0.05) and the p\_value for dentist 4 is less than the level of significance (0.05). As majority of the dataset is above the level of significance, we can say that we fail to reject the Null Hypothesis.

There is evidence to suggest the sample comes from a normal distribution.

We can check equality of variance assumptions by using the Levene test

**Levene Test:**

Null Hypothesis (H0) = The variances of the sample being compared are equal for alloy 1 or alloy 2

Alternative Hypothesis (Ha) = The variances of the sample being compared are not equal for alloy 1 or alloy 2

**For Alloy 1:**

The p\_value = 0.256 is greater than the level of significance (0.05), We fail to reject the Null Hypothesis.

There is evidence to suggest the variances of the samples are equal.

**For Alloy 2:**

The p\_value = 0.237 is greater than the level of significance (0.05), We fail to reject the Null Hypothesis.

There is evidence to suggest the variances of the samples are equal

**ANOVA TEST:**

**For Alloy 1:**

The p\_value for alloy 1 is 0.116 after performing the ANOVA test which is greater than the level of significance (0.05), We fail to reject the null hypothesis.

Evidence suggests that the average hardness of implants is same across all 5 dentists for Alloy 1.

**For Alloy 2:**

The p\_value for alloy 2 is 0.718 after performing the ANOVA test which is greater than the level of significance (0.05), We fail to reject the null hypothesis.

Evidence suggests that the average hardness of implants is same across all 5 dentists for Alloy 2.

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

**Defining the Null & Alternative Hypothesis:**

Null Hypothesis (H0) = There average implant hardness does not vary depending on methods for alloy 1 or alloy 2

Alternative Hypothesis (Ha) = There average implant hardness varies depending on methods for alloy 1 or alloy 2.

**Checking the assumptions:**

We can check normality of the distribution by using the Shapiro-Wilk’s test

**Shapiro Wilk’s Test:**  
Null Hypothesis (H0) = The population from which the sample is drawn follows a normal distribution for alloy 1 or alloy 2

Alternative Hypothesis (Ha) = The population from which the sample is drawn does not follow a normal distribution for alloy 1 or alloy 2

**For Alloy 1:**

The p\_value for each method is more than the level of significance (0.05), We fail to reject the Null Hypothesis

There is evidence to suggest the sample comes from a normal distribution

**For Alloy 2:**

The p\_value for methods 1 and 3 is more than the level of significance (0.05) and the p\_value for method 2 is less than the level of significance (0.05). As majority of the dataset is above the level of significance, we can say that we fail to reject the Null Hypothesis.

There is evidence to suggest the sample comes from a normal distribution.

We can check equality of variance assumptions by using the Levene test

**Levene Test:**

Null Hypothesis (H0) = The variances of the sample being compared are equal for alloy 1 and alloy 2

Alternative Hypothesis (Ha) = The variances of the sample being compared are not equal for alloy 1 and alloy 2

**For Alloy 1:**

The p\_value = 0.034 is lesser than the level of significance (0.05), We reject the Null Hypothesis.

There is no evidence to suggest the variances of the samples are equal.

**For Alloy 2:**

The p\_value = 0.044 is lesser than the level of significance (0.05), We reject the Null Hypothesis.

There is no evidence to suggest the variances of the samples are equal.

**ANOVA TEST:**

**For Alloy 1:**

The p\_value for alloy 1 is 0.004 after performing the ANOVA test which is lesser than the level of significance (0.05), We reject the null hypothesis.

Evidence suggests that there is variance in average implant hardness depending on methods.

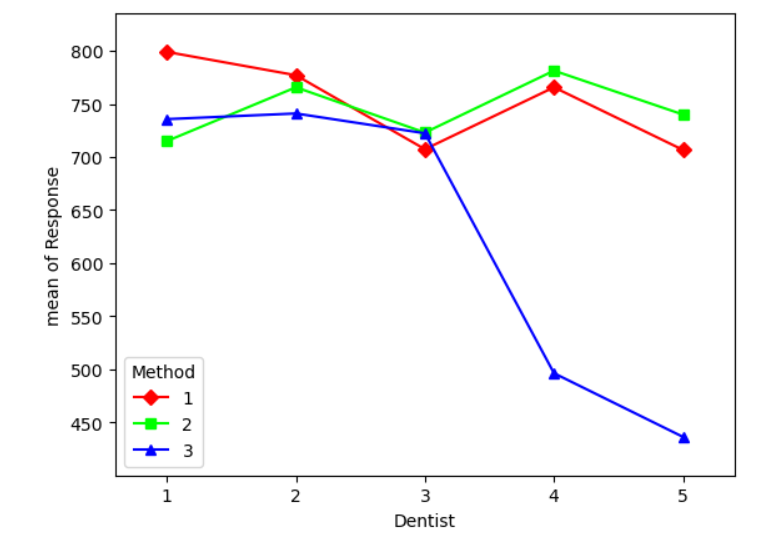
**For Alloy 2:**

The p\_value for alloy 2 is 0.000005 after performing the ANOVA test which is lesser than the level of significance (0.05), We reject the null hypothesis.

Evidence suggests that there is variance in average implant hardness depending on methods.

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

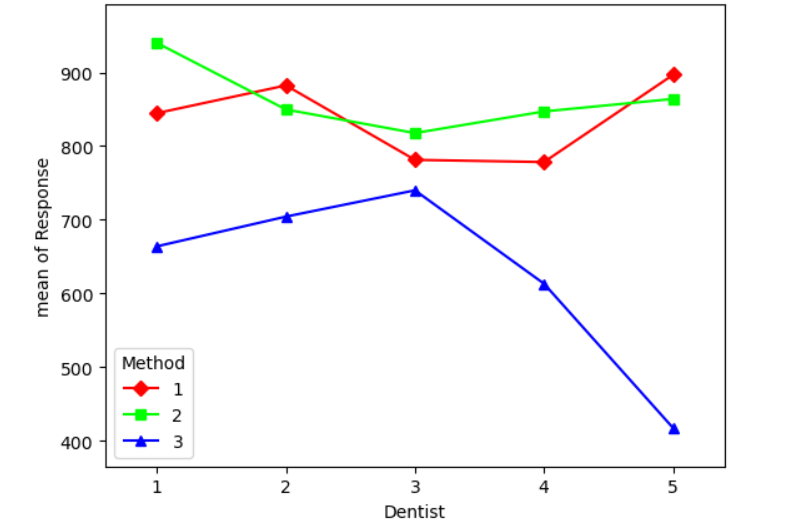
Figure 13 – Interaction plot of dentist and method for Alloy 1



**Observations for Alloy 1:**

There is interaction between Method 1, 2 and 3 for the first 3 dentists and

Figure 14 - Interaction plot of dentist and method for Alloy 2



**Observations for Alloy 2:**

There is some interaction between method 1 and 2 but there is no interaction between 1 and 3 and 2 and 3.

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

**Defining the Null & Alternative Hypothesis:**

Null Hypothesis (H0) = There is no significant interaction effect between dentists and methods on the hardness of implants for alloy 1 or alloy 2.

Alternative Hypothesis (Ha) = There is a significant interaction effect between dentists and methods on the hardness of implants for alloy 1 or alloy 2.

**ANOVA TEST:**

**For Alloy 1:**

The p\_value for alloy 1 is 0.0067 after performing the ANOVA test which is lesser than the level of significance (0.05), We reject the null hypothesis.

Evidence suggests that there is some interaction between Dentist and Methods on the hardness of implants for Alloy 1.

**For Alloy 2:**

The p\_value for alloy 2 is 0.0932 after performing the ANOVA test which is greater than the level of significance (0.05), We fail to reject the null hypothesis.

Evidence suggests that there is no interaction between Dentists and Methods on the hardness if implants for Alloy 2.