Business Report

IS Coded Project

PGPDSBA

Chithira Raj

Table of Contents

List of Tables

List of Figures

Problem 1

1.1. Objective

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.

1.2. Data Overview

	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**

Problem 1 - Data

1.3. Questions

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

P (Injury) = Total Injured Players / Total Players = (145/235) = 0.617

So, the probability that a randomly chosen player would suffer an injury is approximately 0.617 or 61.7%.

1.3.2. What is the probability that a player is a forward or a winger?

P (Forward or Winger) = (Total Forwards + Total Wingers) / Total Players = (94+29) / 235 = 0.523

So, the probability that a player is a forward or a winger is approximately 0.523 or 52.3%.

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

P (Striker and Injured) = Striker Injured / Total Players = 45 / 235 = 0.191

So, the probability that a randomly chosen player is a striker and has a foot injury is approximately 0.191 or 19.1%.

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

P (Striker given Injury) = Striker Injured / Total Injured Players = 45 / 145 = 0.310

So, the probability that a randomly chosen injured player is a striker is approximately 0.310 or 31.0%.

Problem 2

2.1. Objective

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.

2.2. Questions

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

Proportion of bags with breaking strength less than 3.17 kg per sq cm: 0.111

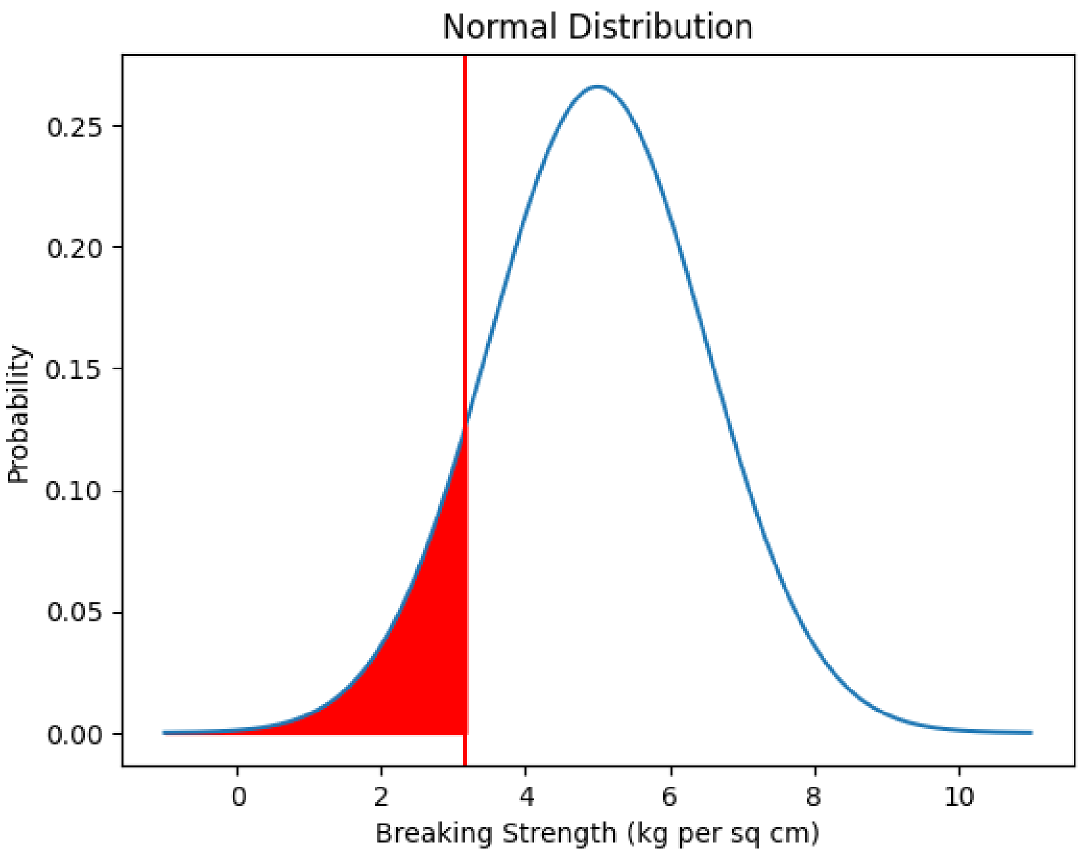

Figure 1: Distribution Plot

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

Proportion of bags with breaking strength at least 3.6 kg per sq cm: 0.824

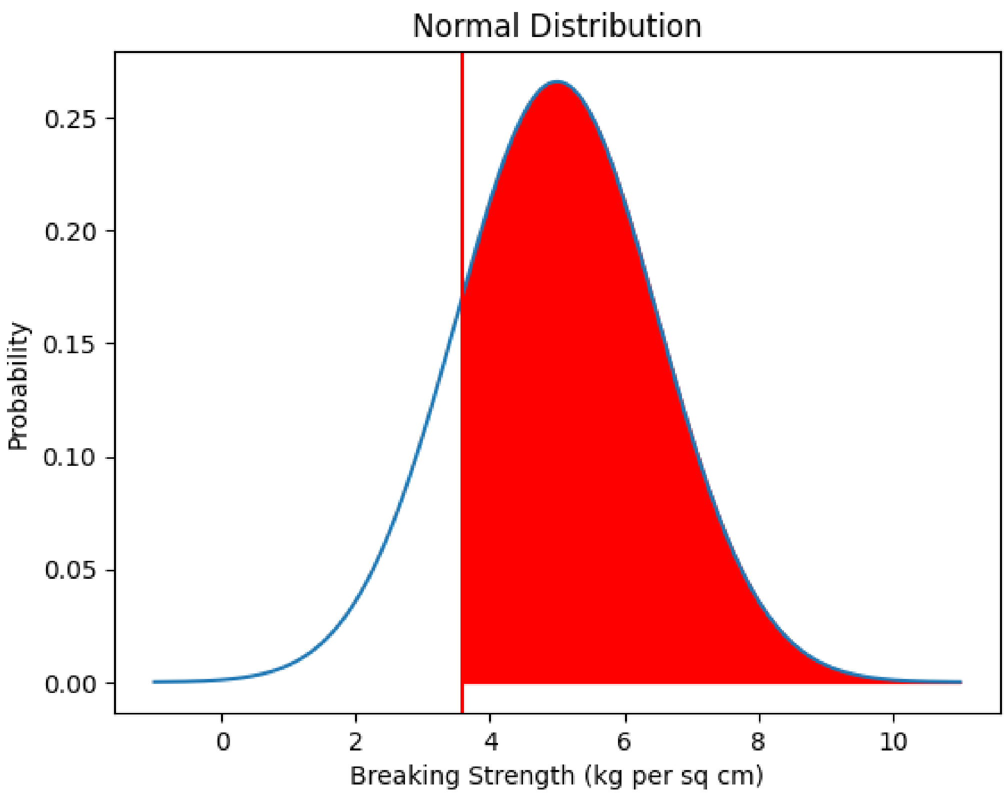

Figure 2: Distribution Plot

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

Proportion of bags with breaking strength between 5 and 5.5 kg per sq cm: 0.130

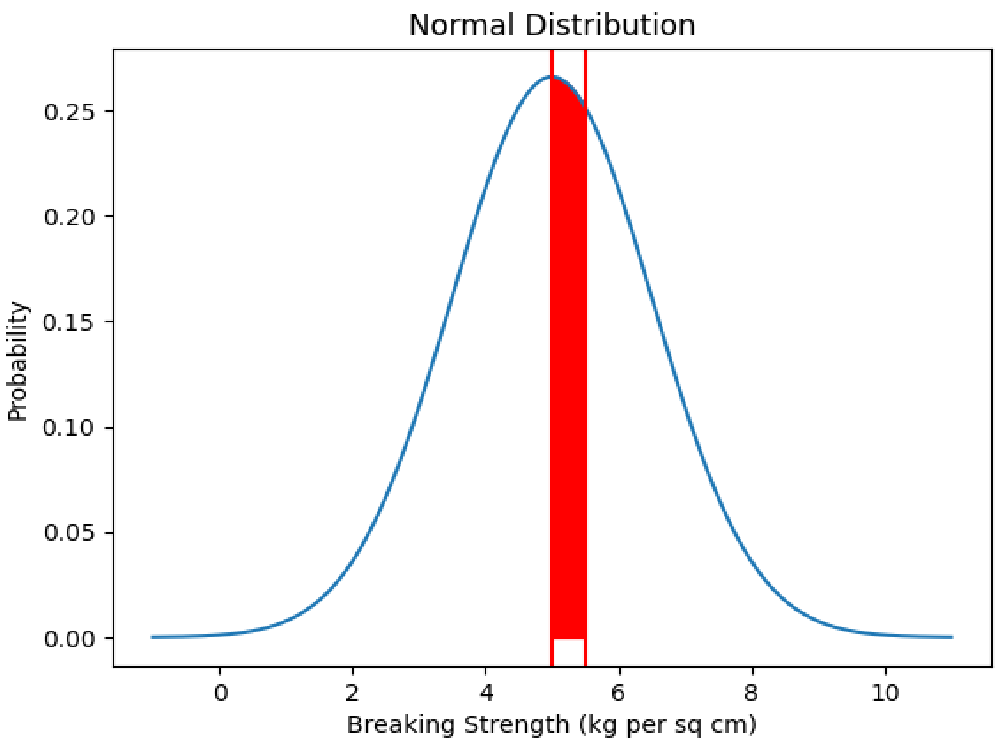

Figure 3: Distribution Plot

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

Proportion of bags with breaking strength NOT between 3 and 7.5 kg per sq cm: 0.139

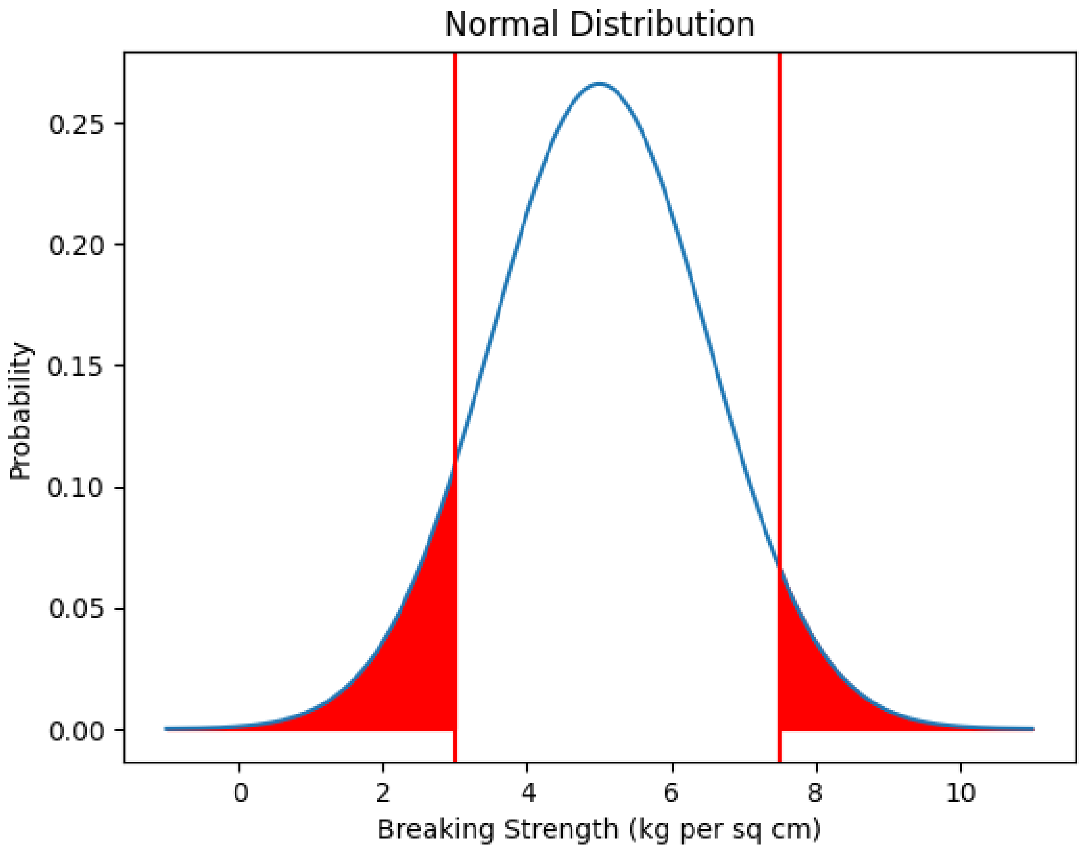

Figure 4: Distribution Plot

Problem 3

3.1. Objective

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).

3.2. Data Dictionary

S.No.	Variables	Description
1	Unpolished	Hardness of Unpolished Stones
2	Treated and Polished	Hardness of Polished stones

Table 1: Data Dictionary

3.3. Data Overview

3.3.1. Import libraries and load the data

Unpolished	Treated and Polished
164.482	133.209
154.307	138.483
129.861	159.665
159.096	145.664
135.257	136.789

Figure 5: Data Overview

3.3.2. Check the structure of data

Shape of the dataset: 75 rows and 2 columns

3.3.3. Check the types of the data

```
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Unpolished            75 non-null     float64
 1   Treated and Polished  75 non-null     float64
dtypes: float64(2)
```

Figure 6: Datatypes

3.3.4. Check for and treat (if needed) missing values

```
Unpolished              0
Treated and Polished    0
dtype: int64
```

Figure 7: Missing Value Check

3.3.5. Check the statistical summary

	count	mean	std	min	25%	50%	75%	max
Unpolished	75.000	134.111	33.042	48.407	115.330	135.597	158.215	200.161
Treated and Polished	75.000	147.788	15.587	107.524	138.268	145.721	157.373	192.273

Figure 8: Statistical Summary

3.3.6. Univariate Analysis

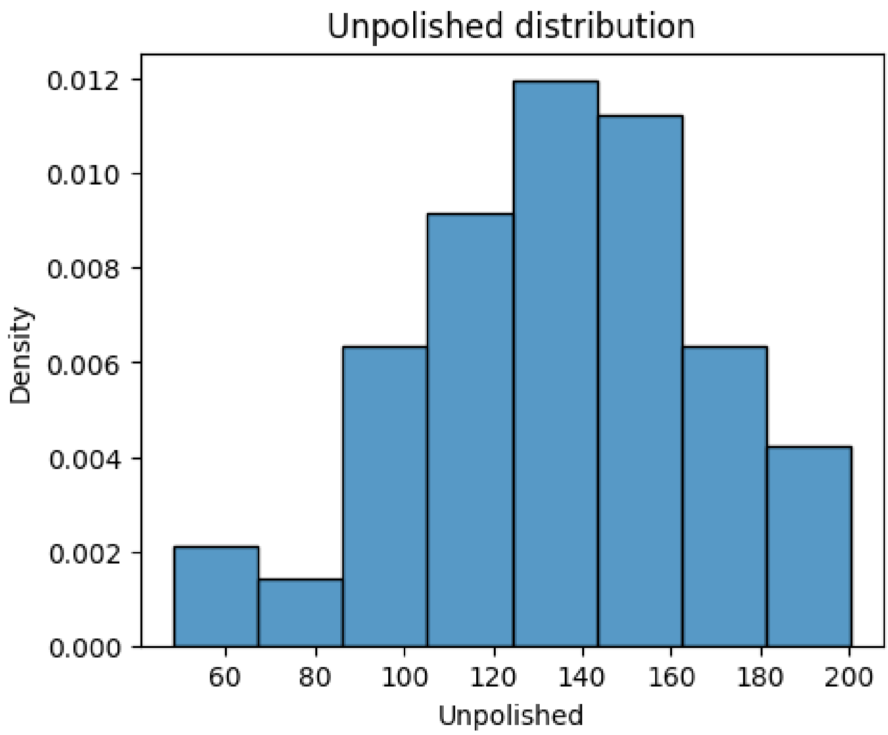

Figure 9: Histogram – Unpolished data

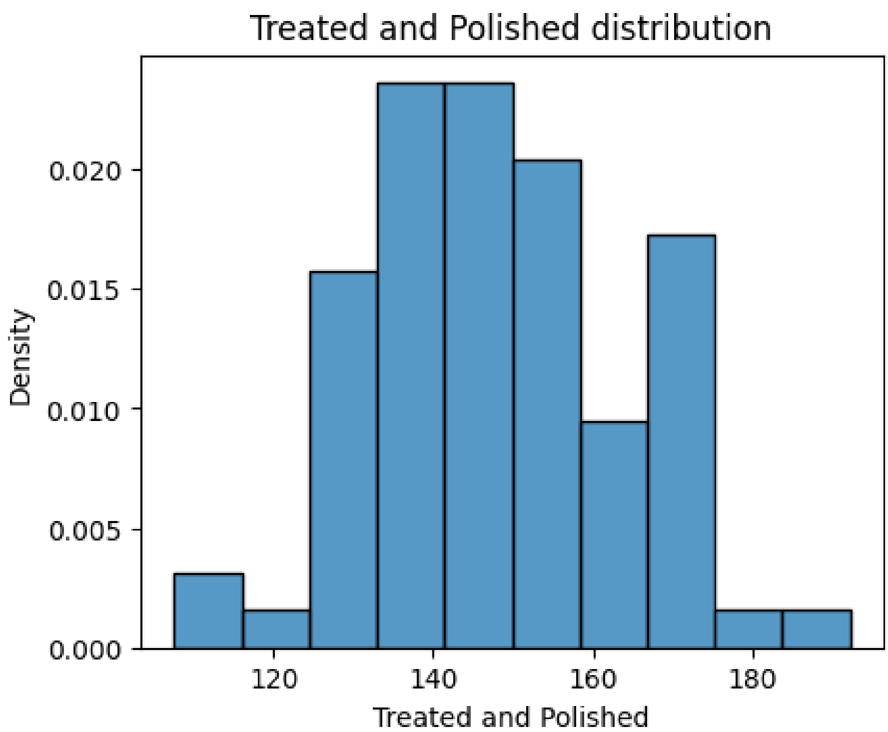

Figure 10: Histogram – Polished data

3.3.7. Bivariate Analysis

Heatmap

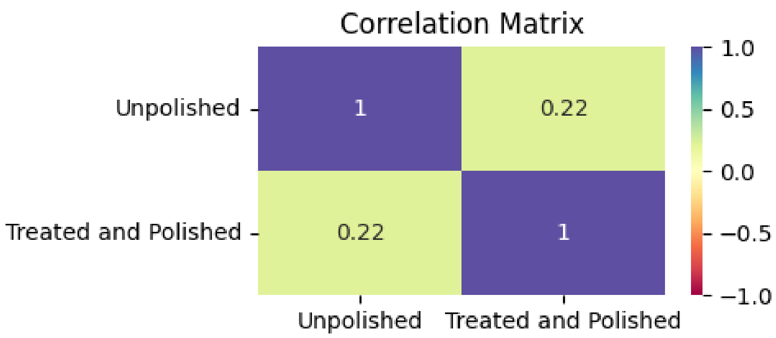

Figure 11: Heatmap

3.4. Questions

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

3.4.1.1. Assumptions

Shapiro-Wilk's test: To check whether hardness of unpolished stones follows normal distribution.

We will test the
Null hypothesis: Hardness of Unpolished stones follow a normal distribution against
against
Alternative hypothesis: Hardness of Unpolished stones does not follow a normal distribution.

p-value = 0.677
Since p-value of the test is much larger than level of significance (0.05), we fail to reject the null hypothesis that the response follows the normal distribution.

Let's check if the assumptions for the T-test are satisfied:

- Continuous data: Yes, the hardness of stones is measured on a continuous scale.
- Normally distributed population: Yes, the population is assumed to be normal based on the Shapiro-Wilk test.
- Observations are from a simple random sample: Yes, we are informed that the collected sample is a simple random sample.
- Population standard deviation is known: No.

  Based on these points, we can use the T-test for this problem.

3.4.1.2. One-Sample T-test

We will test the null hypothesis
Mean = 150
against the alternate hypothesis
Mean ≠ 150

The p-value is 8.342573994839304e-05

3.4.1.3. Insights

Given that the p-value is much smaller than the threshold of 0.05, we reject the null hypothesis. This statistical evidence strongly suggests that the true mean hardness of the unpolished stones is different from 150.

Zingaro can confidently conclude that the unpolished stones do not meet the hardness requirement for their printing processes. This finding justifies Zingaro's concerns regarding the suitability of these stones for printing. The deviation in hardness could lead to issues in the printing quality or durability, indicating that alternative materials or further processing may be necessary to meet the required standards.

3.4.2. Is the mean hardness of the polished and unpolished stones the same?

3.4.2.1. Assumptions

Levene's test: To check Homogeneity of variances

We will test the
Null hypothesis: All the population variances are equal
against the
Alternative hypothesis: At least one variance is different from the rest

The p-value is 3.4771056592659467e-07

Since the p-value is much less than the significance level of 0.05, we reject the null hypothesis of homogeneity of variances.

Standard Deviation:

The standard deviation of hardness of Unpolished stone is 33.04

The standard deviation of hardness of Polished stone is 15.59

Let's check if the assumptions for the T-test are satisfied:

- Continuous data: Yes, the hardness is measured on a continuous scale.
- Normally distributed populations: Yes, we are informed that the populations are assumed to be normal.
- Independent populations: Yes, since we are taking random samples from two different groups, the two samples are from two independent populations.
- Unequal population standard deviations: Yes, since the sample standard deviations are different, we can assume the population standard deviations may also be different.
- Random sampling from the population: Yes, we are informed that the collected sample is a simple random sample.

Given these points, we can use the two-sample T-test for this problem.

3.4.2.2. Two Independent Sample T-test for Equality of Means- Unequal Std Dev

Let $\mu1$, $\mu2$ be the mean hardness of Unpolished stones and Polished stones respectively.

We will test the

Null hypothesis: $\mu1=\mu2$

against the

Alternate hypothesis: $\mu1\neq\mu2$

The p-value is 0.001.

3.4.2.3. Insights

Given that the p-value is much smaller than the threshold of 0.05, we reject the null hypothesis. This statistical evidence strongly supports the claim that there is a significant difference in the mean hardness between polished and unpolished stones.

Problem 4

4.1. Objective

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.

4.2. Data Dictionary

S.No.	Variables	Description
1	Dentist	Dentists who may favor one method above another and may work better in his/her favorite method
2	Method	Method of implant
3	Alloy	Alloy used for implant
4	Temp	Temperature at which the metal is treated
5	Response	The hardness of metal implants in dental cavities

Table 2: Data Dictionary

4.3. Data Overview

4.3.1. Import libraries and load the data

	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

Figure 12: Data Overview

4.3.2. Check the structure of data

Shape of the dataset: 90 rows and 5 columns

4.3.3. Check the types of the data

```
 #   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)
```

Figure 13: Datatypes

4.3.4. Check for and treat (if needed) missing values

```
Dentist     0
Method      0
Alloy       0
Temp        0
Response    0
dtype: int64
```

Figure 14: Missing value check

4.3.5. Check the statistical summary

	count	mean	std	min	25%	50%	75%	max
Dentist	90.000	3.000	1.422	1.000	2.000	3.000	4.000	5.000
Method	90.000	2.000	0.821	1.000	1.000	2.000	3.000	3.000
Alloy	90.000	1.500	0.503	1.000	1.000	1.500	2.000	2.000
Temp	90.000	1600.000	82.107	1500.000	1500.000	1600.000	1700.000	1700.000
Response	90.000	741.778	145.768	289.000	698.000	767.000	824.000	1115.000

Figure 15: Statistical Summary

4.3.6. Univariate Analysis

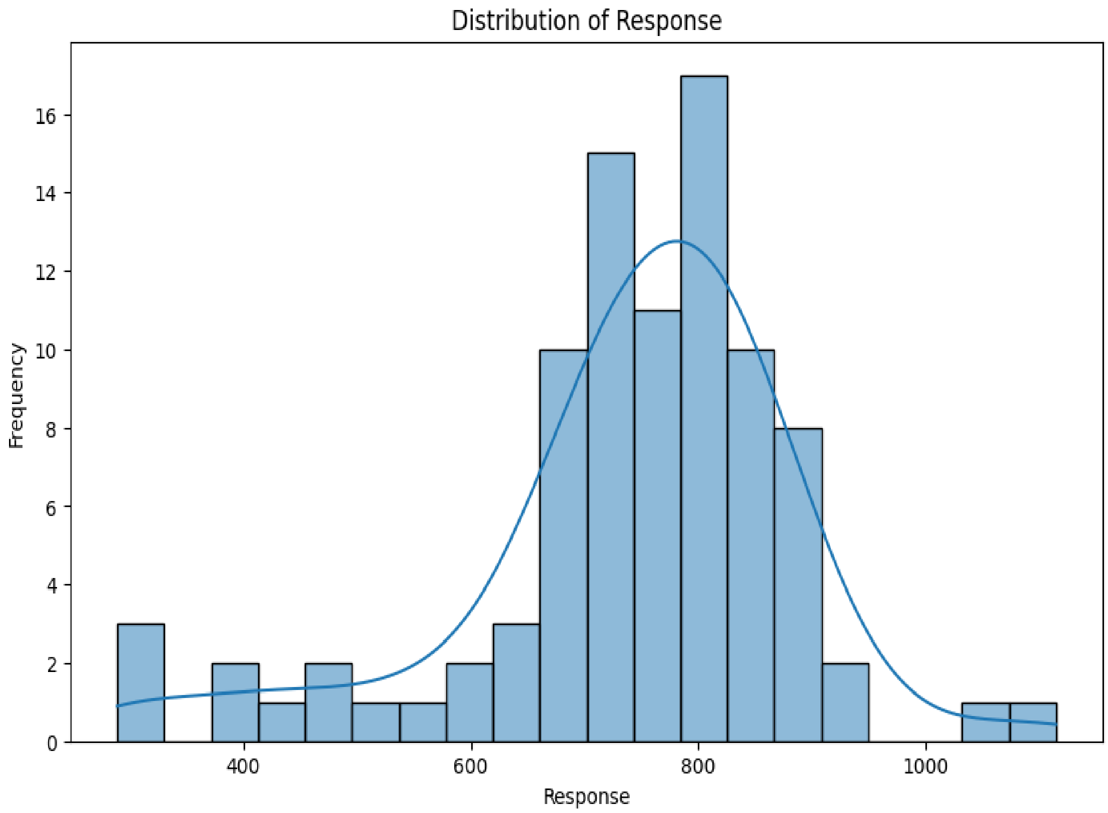

Figure 16: Distribution of Response Variable

4.3.7. Bivariate Analysis

4.3.7.1. Heatmap

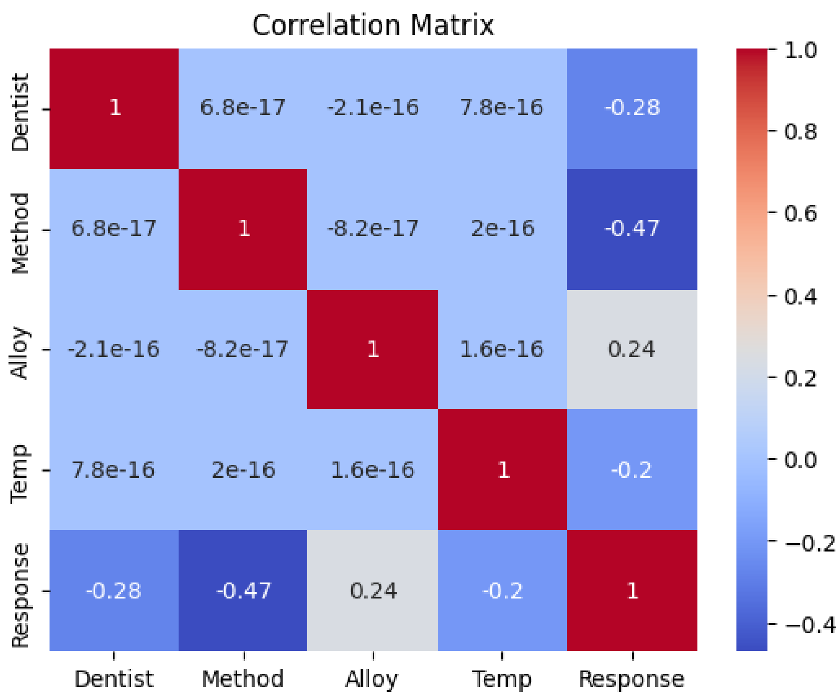

Figure 17: Heatmap

4.3.7.2. Boxplots

Dentist vs Response

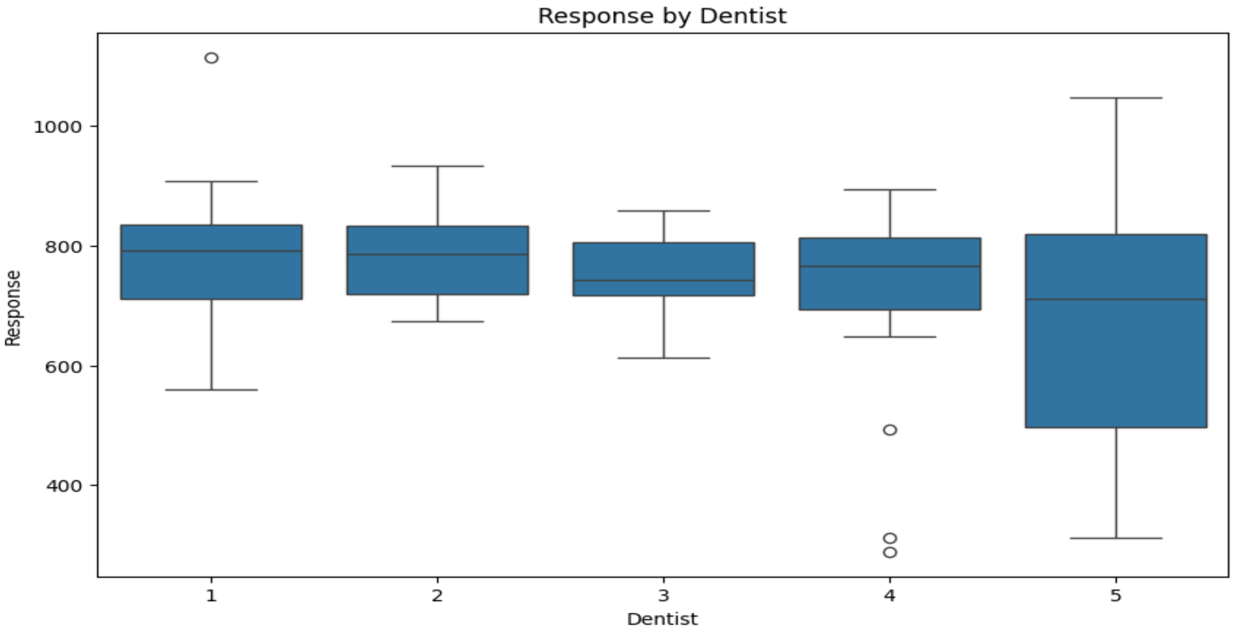

Figure 18: Dentist vs Response

Method vs Response

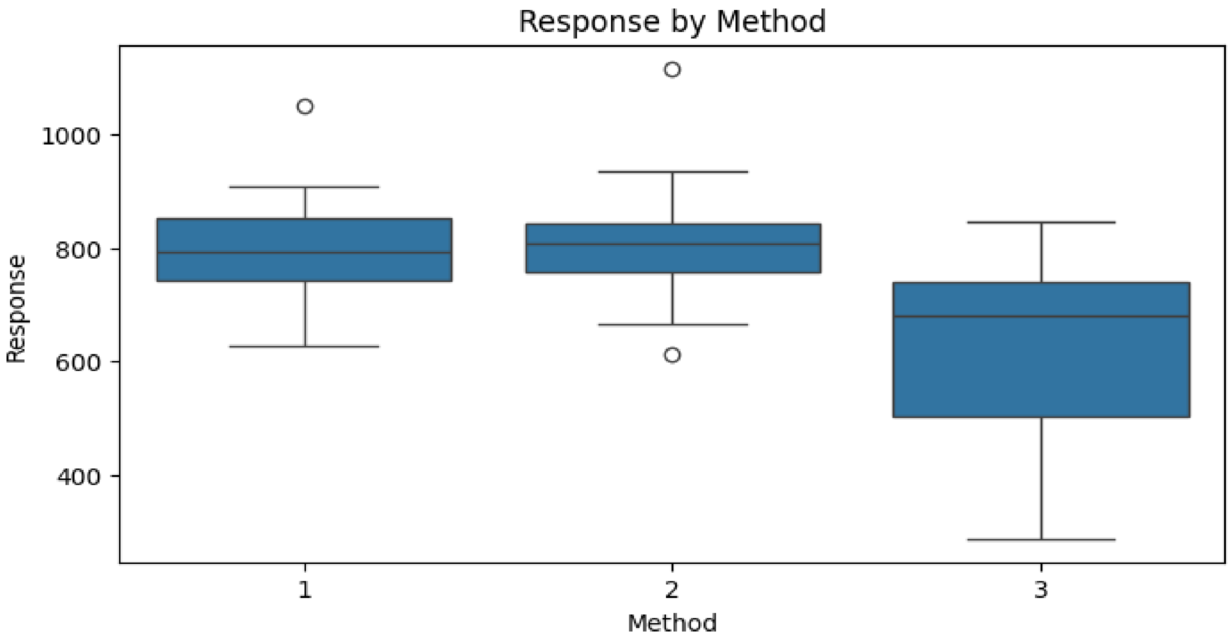

Figure 19: Method vs Response

Alloy vs Response

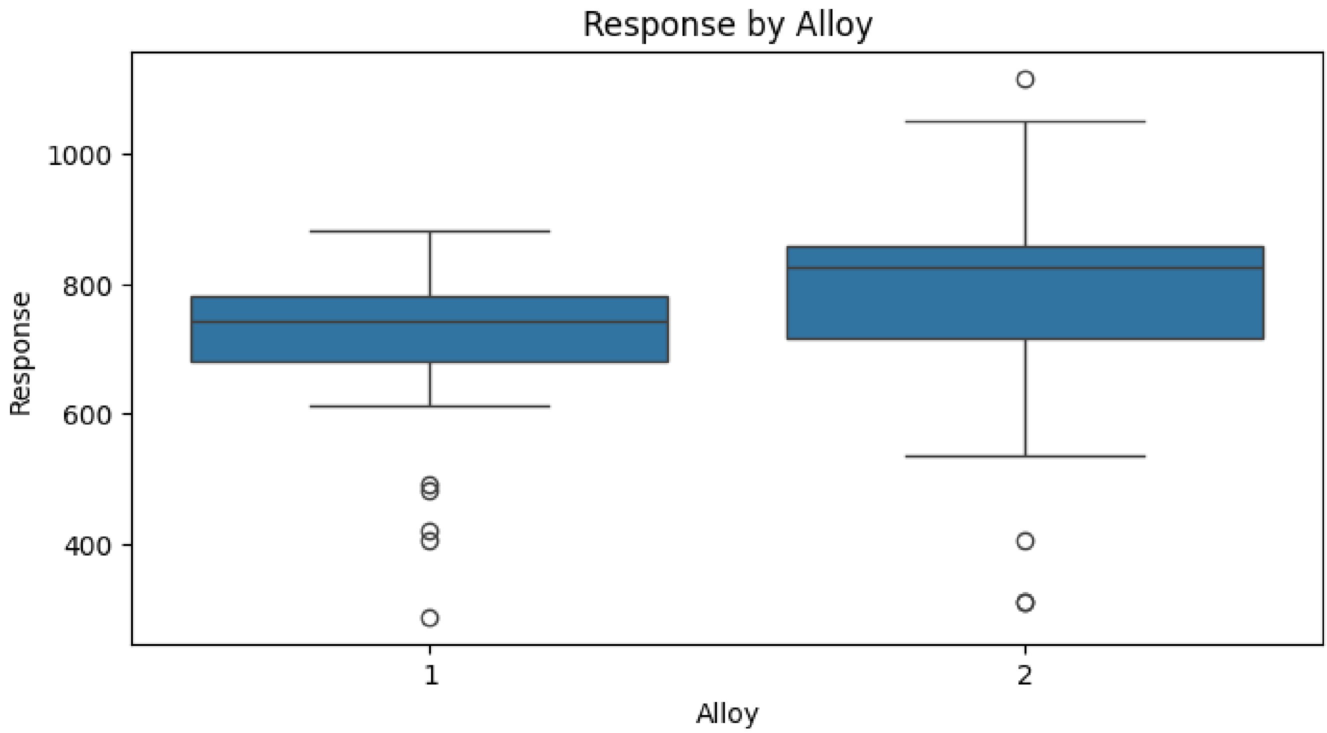

Figure 20: Alloy vs Response

Temperature vs Response

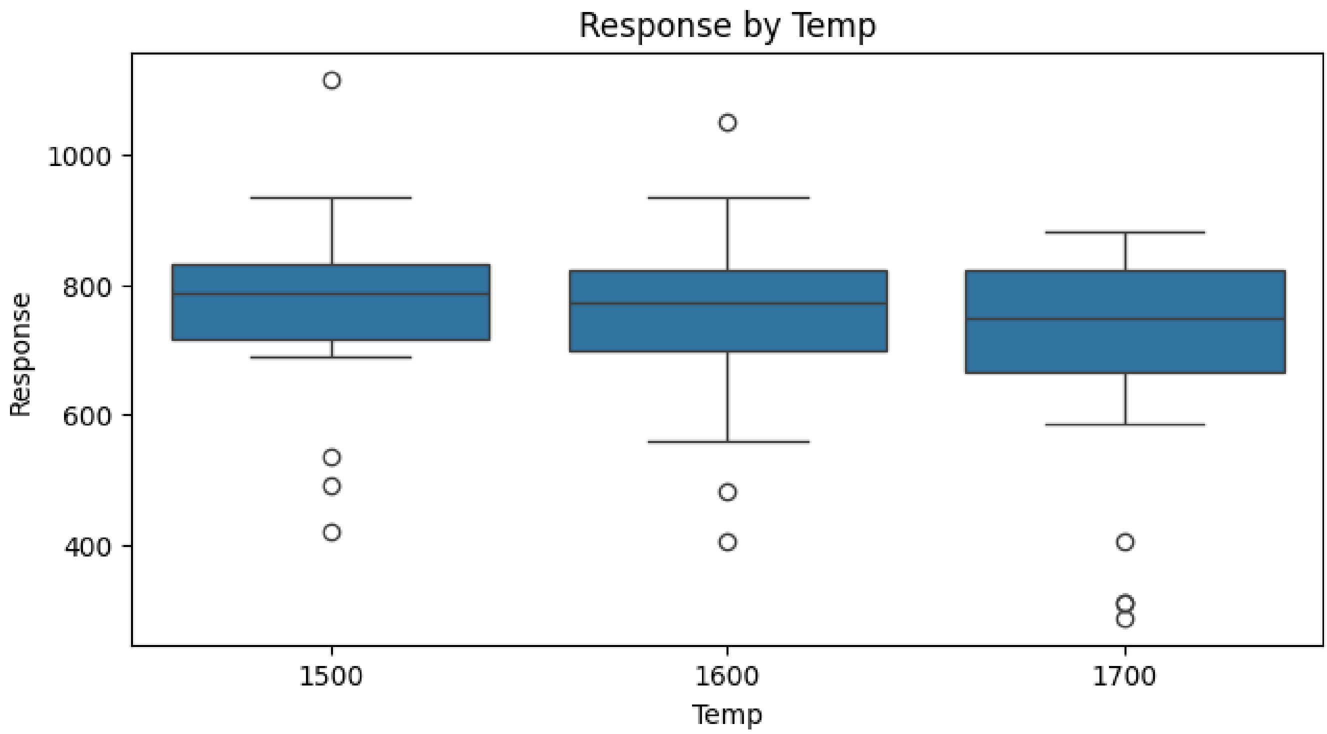

Figure 21: Temperature vs Response

4.4. Questions

4.4.1. How does the hardness of implants vary depending on dentists?

One Way ANOVA test

Alloy 1

Let's write the null and alternative hypothesis

Let µ1, µ2, µ3, µ4, µ5 be the means of hardness of implants for each dentist respectively.
We will test the
Null hypothesis: µ1=µ2=µ3=µ4=µ5
against the
Alternative hypothesis: At least one hardness of implants is different from the rest.

```
Means of hardness of implants for each dentist:
 Dentist
1    749.889
2    761.222
3    717.556
4    681.111
5    627.667
Name: Response, dtype: float64
```

Figure 22: Mean Values

Now, the normality and equality of variance assumptions need to be checked.

For testing of normality, Shapiro-Wilk's test is applied to the response variable.
For equality of variance, Levene test is applied to the response variable.

Shapiro-Wilk's test
We will test the
Null hypothesis: Hardness of implants follows a normal distribution
against the
Alternative hypothesis: Hardness of implants does not follow a normal distribution

The p-value is 1.1945070582441986e-05

Since the p-value is less than the significance level of 0.05, we reject the null hypothesis. Hence, we have enough evidence to support the claim that the hardness of implants does not follow a normal distribution.

Levene's test
We will test the
Null hypothesis: All the population variances are equal
against the
Alternative hypothesis: At least one variance is different from the rest

The p-value is 0.2565537418543795

Since the p-value is large, we fail to reject the null hypothesis of homogeneity of variances.

Now, the One-way ANOVA test gives,

p-value = 0.116

Since the p-value is larger than the significance level of 0.05, we fail to reject the null hypothesis that the mean hardness of implants for each dentist is equal.

Alloy 2

Let's write the null and alternative hypothesis

Let µ1, µ2, µ3, µ4, µ5 be the means of hardness of implants for each dentist respectively.
We will test the
Null hypothesis: µ1=µ2=µ3=µ4=µ5
against the
Alternative hypothesis: At least one hardness of implants is different from the rest.

```
Means of hardness of implants for each dentist:
 Dentist
1    816.222
2    812.111
3    779.667
4    746.222
5    726.111
Name: Response, dtype: float64
```

Figure 23: Mean Values

Now, the normality and equality of variance assumptions need to be checked.

For testing of normality, Shapiro-Wilk's test is applied to the response variable.
For equality of variance, Levene test is applied to the response variable.

<u>Shapiro-Wilk's test</u>
We will test the
Null hypothesis: Hardness of implants follows a normal distribution
against the
Alternative hypothesis: Hardness of implants does not follow a normal distribution

The p-value is 0.00040293222991749644

Since the p-value is less than the significance level of 0.05, we reject the null hypothesis. Hence, we have enough evidence to support the claim that the hardness of implants does not follow a normal distribution.

<u>Levene's test</u>
We will test the
Null hypothesis: All the population variances are equal
against the
Alternative hypothesis: At least one variance is different from the rest

The p-value is 0.23686777576324952

Since the p-value is large, we fail to reject the null hypothesis of homogeneity of variances.

Now, the One-way ANOVA test gives,

p-value = 0.718

Since the p-value is larger than the significance level of 0.05, we fail to reject the null hypothesis that the mean hardness of implants for each dentist is equal.

4.4.2. How does the hardness of implants vary depending on methods?

One Way ANOVA test

Alloy 1

Let's write the null and alternative hypothesis

Let μ1, μ2, μ3, μ4, μ5 be the means of hardness of implants for each method respectively.
We will test the
Null hypothesis: μ1=μ2=μ3
against the
Alternative hypothesis: At least one hardness of implants is different from the rest.

```
Means of hardness of implants for each method:
 Method
1    751.133
2    745.000
3    626.333
Name: Response, dtype: float64
```

Figure 24: Mean Values

Now, the normality and equality of variance assumptions need to be checked.

For testing of normality, Shapiro-Wilk's test is applied to the response variable.
For equality of variance, Levene test is applied to the response variable.

Shapiro-Wilk's test
We will test the
Null hypothesis: Hardness of implants follows a normal distribution
against the
Alternative hypothesis: Hardness of implants does not follow a normal distribution

The p-value is 1.1945070582441986e-05

Since the p-value is less than the significance level of 0.05, we reject the null hypothesis. Hence, we have enough evidence to support the claim that the hardness of implants does not follow a normal distribution.

Levene's test
We will test the
Null hypothesis: All the population variances are equal
against the
Alternative hypothesis: At least one variance is different from the rest
The p-value is 0.003

Since the p-value is less than level of significance, we can reject the null hypothesis of homogeneity of variances.

Now, the One-way ANOVA test gives,

p-value = 0.004

Since the p-value is less than the significance level of 0.05, we reject the null hypothesis. Hence, we have enough evidence to support the claim that the mean hardness of implants for each method is not equal.

Multiple Comparison test (Tukey HSD):

In order to identify for which method, mean hardness of implant is different from other groups, the

null hypothesis is

$\mu1=\mu2=\mu3$

against the

Alternative hypothesis: At least one pair of means is significantly different.

```
  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
```

Figure 25: Multiple comparison of means

Insights

As the p-values (refer to the p-adj column) for comparing the mean hardness of implant for the pair Method(1,3) and Method(2,3) is less than the significance level, the null hypothesis of equality of all population means can be rejected.

Thus, we can say that the mean hardness of implant for Method 1 and Method 2 is similar but hardness for Method 3 is significantly different from Method 1 and Method 2.

Alloy 2

Let's write the null and alternative hypothesis

Let $\mu1$, $\mu2$, $\mu3$, $\mu4$, $\mu5$ be the means of hardness of implants for each method respectively.
We will test the
Null hypothesis: $\mu1=\mu2=\mu3$
against the
Alternative hypothesis: At least one hardness of implants is different from the rest.

```
Means of hardness of implants for each method:
 Method
1    836.667
2    863.667
3    627.867
Name: Response, dtype: float64
```

Figure 26: Mean values

Now, the normality and equality of variance assumptions need to be checked.

For testing of normality, Shapiro-Wilk's test is applied to the response variable.
For equality of variance, Levene test is applied to the response variable.

Shapiro-Wilk's test

We will test the
Null hypothesis: Hardness of implants follows a normal distribution
against the
Alternative hypothesis: Hardness of implants does not follow a normal distribution

The p-value is 0.0004

Since the p-value is less than the significance level of 0.05, we reject the null hypothesis. Hence, we have enough evidence to support the claim that the hardness of implants does not follow a normal distribution.

Levene's test

We will test the
Null hypothesis: All the population variances are equal
against the
Alternative hypothesis: At least one variance is different from the rest

The p-value is 0.04

Since the p-value is less than level of significance, we can reject the null hypothesis of homogeneity of variances.

Now, the One-way ANOVA test gives,

p-value = 5.415871051443187e-06

Since the p-value is less than the significance level of 0.05, we reject the null hypothesis. Hence, we have enough evidence to support the claim that the mean hardness of implants for each method is not equal.

Multiple Comparison test (Tukey HSD):

In order to identify for which method, mean hardness of implant is different from other groups, the

null hypothesis is

μ1=μ2=μ3

against the

Alternative hypothesis: At least one pair of means is significantly different.

```
  Multiple Comparison of Means - Tukey HSD, FWER=0.05
=======================================================
group1 group2 meandiff p-adj    lower     upper   reject
-------------------------------------------------------
     1      2     27.0 0.8212  -82.4546  136.4546  False
     1      3   -208.8 0.0001 -318.2546  -99.3454   True
     2      3   -235.8    0.0 -345.2546 -126.3454   True
-------------------------------------------------------
```

Figure 27: Multiple comparison of means

Insights

As the p-values (refer to the p-adj column) for comparing the mean hardness of implant for the pair Method (1,3) and Method (2,3) is less than the significance level, the null hypothesis of equality of all population means can be rejected.

Thus, we can say that the mean hardness of implant for Method 1 and Method 2 is similar but hardness for Method 3 is significantly different from Method 1 and Method 2.

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

4.4.2.1. Interaction Plot for each Alloy type

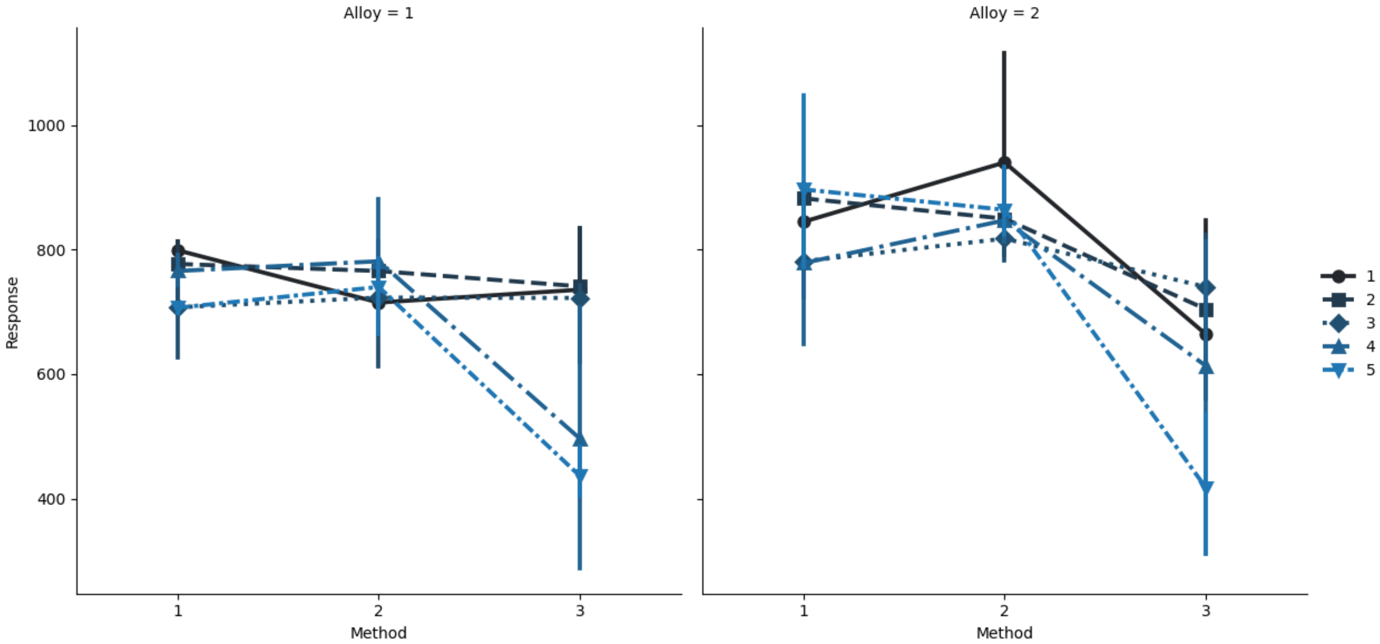

Figure 28: Interaction Plot

4.4.2.2. Insights

- We can notice some interaction effect between the dentist and method on the hardness of dental implants for each type of alloy.
- Dentist 1 prefers using method 2 with alloy 2 to achieve a higher hardness level on dental implants.
- Dentist 5 obtains a low hardness level with method 3 for both alloy implants.
- Method 2 consistently results in a higher hardness level for alloy 2.

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

Two Way ANOVA test

Alloy 1

Let's write the null and alternative hypothesis

We will test the
Null hypothesis: There is no interaction effect between dentists and methods on the hardness of implants.
against the
Alternative hypothesis: There is an interaction effect between dentists and methods on the hardness of implants.

Now, the normality and equality of variance assumptions need to be checked.

For testing of normality, Shapiro-Wilk's test is applied to the response variable.
For equality of variance, Levene test is applied to the response variable.

Shapiro-Wilk's test
We will test the
Null hypothesis: The residuals (error terms) from the two-way ANOVA model are normally distributed in the population,
against the
Alternative hypothesis: The residuals (error terms) from the two-way ANOVA model are not normally distributed in the population.

The p-value is 0.474

Since the p-value is larger than the significance level of 0.05, we fail to reject the null hypothesis. Hence, the residuals (error terms) from the two-way ANOVA model are normally distributed in the population.

Levene's test
We will test the
Null hypothesis: All the population variances are equal across groups formed by combination of Dentist and Method.
against the
Alternative hypothesis: At least one variance is different from the rest

The p-value is 0.312

Since the p-value is larger than the significance level of 0.05, we fail to reject the null hypothesis. Hence, we do not have enough evidence to conclude that the population variances are different across groups formed by the combination of Dentist and Method.

Now, the Two-way ANOVA test gives,

p-value = 0.011

Since the p-value is less than the chosen significance level (0.05), we reject the null hypothesis and conclude that there is a significant interaction effect between dentists and methods on the hardness of implants.

Multiple Comparison test (Tukey HSD):

Null hypothesis: There is no interaction effect dentist and method combinations on the hardness of implant.

Alternate Hypothesis: There is at least one dentist-method combination where the interaction effect on the hardness of implants differs significantly.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
1-1	1-2	-84.0	0.9933	-332.8283	164.8283	False
1-1	1-3	-63.3333	0.9996	-312.1617	185.495	False
1-1	2-1	-22.0	1.0	-270.8283	226.8283	False
1-1	2-2	-33.3333	1.0	-282.1617	215.495	False
1-1	2-3	-58.0	0.9999	-306.8283	190.8283	False
1-1	3-1	-91.6667	0.9853	-340.495	157.1617	False
1-1	3-2	-76.0	0.9975	-324.8283	172.8283	False
1-1	3-3	-76.6667	0.9972	-325.495	172.1617	False
1-1	4-1	-33.3333	1.0	-282.1617	215.495	False
1-1	4-2	-17.6667	1.0	-266.495	231.1617	False
1-1	4-3	-302.6667	0.007	-551.495	-53.8383	True
1-1	5-1	-92.3333	0.9844	-341.1617	156.495	False
1-1	5-2	-59.0	0.9998	-307.8283	189.8283	False
1-1	5-3	-362.6667	0.0007	-611.495	-113.8383	True
1-2	1-3	20.6667	1.0	-228.1617	269.495	False
1-2	2-1	62.0	0.9997	-186.8283	310.8283	False
1-2	2-2	50.6667	1.0	-198.1617	299.495	False
1-2	2-3	26.0	1.0	-222.8283	274.8283	False

Figure 29: Multiple comparison of means – Sample result

Insights

Since the p-values for some comparisons are less than the chosen significance level of 0.05, we reject the null hypothesis. Therefore, we conclude that those specific dentist-method combinations exhibit significantly different interaction levels with the hardness of implants.

Alloy 2

Let's write the null and alternative hypothesis

We will test the
Null hypothesis: There is no interaction effect between dentists and methods on the hardness of implants.
against the
Alternative hypothesis: There is an interaction effect between dentists and methods on the hardness of implants.

Now, the normality and equality of variance assumptions need to be checked.

For testing of normality, Shapiro-Wilk’s test is applied to the response variable.
For equality of variance, Levene test is applied to the response variable.

Shapiro-Wilk's test

We will test the

Null hypothesis: The residuals (error terms) from the two-way ANOVA model are normally distributed in the population,

against the

Alternative hypothesis: The residuals (error terms) from the two-way ANOVA model are not normally distributed in the population.

The p-value is 0.485

Since the p-value is larger than the significance level of 0.05, we fail to reject the null hypothesis. Hence, the residuals (error terms) from the two-way ANOVA model are normally distributed in the population.

Levene's test

We will test the

Null hypothesis: All the population variances are equal across groups formed by combination of Dentist and Method.

against the

Alternative hypothesis: At least one variance is different from the rest

The p-value is 0.783

Since the p-value is larger than the significance level of 0.05, we fail to reject the null hypothesis. Hence, we do not have enough evidence to conclude that the population variances are different across groups formed by the combination of Dentist and Method.

Now, the Two-way ANOVA test gives,

p-value = 0.09

Since the p-value is greater than the chosen significance level of 0.05, we fail to reject the null hypothesis. Therefore, we conclude that there is no significant interaction effect between dentists and methods on the hardness of implants.