INFERENTIAL STATISTICS PROJECT

BUSINESS REPORT

Prepared By : Sahid
COURSE : DSBA
DATE : 02-06-2024

Table of Contents

1. Introduction

2. Problem (1) : Foot Injuries and Player Positions

- **2.1 Probability of Injury**
- **2.2 Probability of Being a Forward or Winger**
- **2.3 Probability of Being a Striker and Injured**
- **2.4 Probability of an Injured Player Being a Striker**

3. Problem (2) : Breaking Strength of Gunny Bags

- **3.1 Proportion Less than 3.17 kg/sq cm**
- **3.2 Proportion at Least 3.6 kg/sq cm**
- **3.3 Proportion Between 5 and 5.5 kg/sq cm**
- **3.4 Proportion Not Between 3 and 7.5 kg/sq cm**

4. Problem (3) : Brinell Hardness Index of Stones

- **4.1 Suitability of Unpolished Stones for Printing**
- **4.2 Comparison of Mean Hardness Between Polished and Unpolished Stones**

5. Problem (4) : Hardness of Dental Implants

- **5.1 Variation by Dentists**
- **5.2 Variation by Methods**
- **5.3 Interaction Effect Between Dentist and Method**
- **5.4 Combined Effect of Dentists and Methods**

6. References

List of Tables

No.	Name of the Table	Page No.
1	Top five rows of the dataset for Problem1	6
2	Table 2: Descriptive statistics of the dataset for Problem1)	6
3	Top five rows of the dataset Zingaro_Company	**12**
4	Basic Information of the Dataset Zingaro_Company	**13**
5	Numerical summarization of the dataset Zingaro_Company	**13**
6	Numerical summarization of the dataset after treating outliers	**16**
7	Top 5 rows of the dataset Dental_Hardness_data	**19**
8	Basic info of the dataset Dental_Hardness_data	**19**
9	new df after removing the irrelevant columns and Nan Values	**20**
10	Statistical Summary of the df after cleanup	**20**
11	ANOVA table for Alloy 1 (dentist)	**23**
12	ANOVA table for Alloy 2 (dentist)	**23**
13	ANOVA table for Alloy 1 (Methods)	**24**
14	ANOVA table for Alloy 2 (Methods)	**24**
15	ANOVA table for Alloy 1	**26**
16	ANOVA table for Alloy 2	**27**

List of Figures

No.	Name of Figure	Page No.
1	Plot showing the distribution of gunny bags with a breaking strength less than 3.17 kg/cm²	**8**
2	Plot showing the distribution gunny bag breaking strength of at least 3.6 kg/cm²	**9**
3	Plot showing the distribution gunny bags with a breaking strength between 5 and 5.5 kg/cm²	**10**
4	Plot showing the distribution breaking strength not between 3 and 7.5 kg/cm²	**11**
5	Plot showing the boxplot and distribution for Unpolished	**14**
6	Plot showing the boxplot and distribution for Treated and Polished Stones	**15**
7	The boxplot and distribution for Unpolished Stones after outlier treatment	**16**
8	The boxplot and distribution for Treated and Polished Stones after outlier treatment	**16**
9	Histogram showing the distribution for Unpolished Stones hardness	**17**
10	Boxplot showing the comparison of Unpolished vs Treated and Polished Stones hardness	**18**
11	Boxplot showing the distribution of Dentist variable	**21**
12	Boxplot showing the distribution of Method variable	**21**
13	Boxplot showing the distribution of Alloy variable	**21**
14	Boxplot showing the distribution of Temp variable	**22**
15	Boxplot showing the distribution of Response variable	**22**
16	Boxplot showing the distribution of Response variable after treating outliers	**22**
17	Pointplot showing the interaction effect between the dentist and method on the hardness of dental implants for Alloy 1	**25**
18	Pointplot showing the interaction effect between the dentist and method on the hardness of dental implants for Alloy 2	**25**

1. Introduction

The main goal of this report is to examine a series of distinct yet related problems through the application of various statistical methods to extract valuable insights and provide actionable recommendations. These problems include analyzing injury patterns among football players, assessing the breaking strength of gunny bags used for packaging, evaluating the suitability of stones for printing based on their hardness, and investigating the factors that influence the hardness of dental implants.

- Problem 1 focuses on the physiotherapist's interest in the relationship between foot injuries and the positions played by football team members. By analyzing data from a male football team, we aim to determine the probabilities of injuries among players in different positions and understand the distribution of these injuries.

- Problem 2 addresses the quality control concerns of a cement company's packaging material. The breaking strength of gunny bags used for packaging cement is examined to identify the proportion of bags falling within specific strength ranges. This analysis helps in understanding potential wastage or pilferage within the supply chain and ensures the durability of the packaging.

- Problem 3 delves into the hardness of stones received by Zingaro Stone Printing, a company specializing in printing images on stones. The suitability of unpolished stones for printing is questioned, and a hypothesis test is conducted to determine if they meet the required hardness index. Additionally, a comparison between the hardness of polished and unpolished stones is performed to ensure they meet quality standards.

- Problem 4 explores the various factors influencing the hardness of metal implants used in dental cavities. Factors such as the method of implant, the temperature of treatment, the alloy used, and the dentist's preference are analyzed to understand their impact on the hardness of implants. This section also examines the interaction effects between different factors to provide a comprehensive understanding of the variables affecting implant hardness.

This report encompasses thorough statistical analyses, hypothesis testing, and visual data representations to substantiate the findings. By methodically addressing each issue, the report aims to offer valuable insights and recommendations for enhancing processes and outcomes in their respective fields. The limitations of each analysis are recognized, and the scope is restricted to the data and methodologies utilized in this study.

2. Problem 1: Foot Injuries and Player Positions

The dataset has 4 rows and 2 columns. It is always a good practice to view a sample of the rows. A simple way to do that is to use head() function

	Position	Players Injured	Players Not Injured
0	Striker	45	32
1	Forward	56	38
2	Attacking Midfielder	24	11
3	Winger	20	9

(Table 1: Top five rows of the dataset for Problem1)

- ## Descriptive Statistics

	Players Injured	Players Not Injured
count	4.000000	4.000000
mean	36.250000	22.500000
std	17.134274	14.662878
min	20.000000	9.000000
25%	23.000000	10.500000
50%	34.500000	21.500000
75%	47.750000	33.500000
max	56.000000	38.000000

(Table 2: Descriptive statistics of the dataset for Problem1)

Observations:

The average number of injured players (36.25) is significantly higher than the average number of non-injured players (22.5), indicating a high incidence of injuries among the team.

The wide range and higher standard deviation in the injured players suggest that injuries are unevenly distributed across different positions.

The distribution of non-injured players is more consistent but still shows significant variability.

These statistics highlight the importance of understanding positional factors contributing to injuries, which could inform targeted interventions by the physiotherapist.

Including these descriptive statistics helps provide a clear understanding of the data set and sets the stage for further probability and risk analysis related to player injuries and positions.

"We decided not to treat or remove any outliers in the data set. Given the small size of the data frame and the fair distribution of values, we determined that proceeding with the analysis without altering the data would provide a more accurate reflection of the real-world scenario.

2.1 Probability of Injury

Analysis:

Total number of players injured: 145

Total number of players: 235

Probability = Total injured players / Total players

Calculation: 145/235 = 0.617

The probability that a randomly chosen player would suffer an injury is 0.617

2.2 Probability of Being a Forward or Winger

Analysis:

Number of forwards: 94

Number of wingers: 29

Total players in these positions: 123

Probability = Players in forward or winger positions / Total players

Calculation: 123/235 = 0.523

The probability that a player is a forward or a winger is: 0.523

2.3 Probability of Being a Striker and Injured

Analysis:

Number of injured strikers: 45

Total number of players: 235

Probability = Injured strikers / Total players

Calculation: 45/235 = 0.191

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

2.4 Probability of an Injured Player Being a Striker

Analysis: Number of injured strikers: 45

Total number of injured players: 145

Probability = Injured strikers / Total injured players

Calculation: 45/145 = 0.31

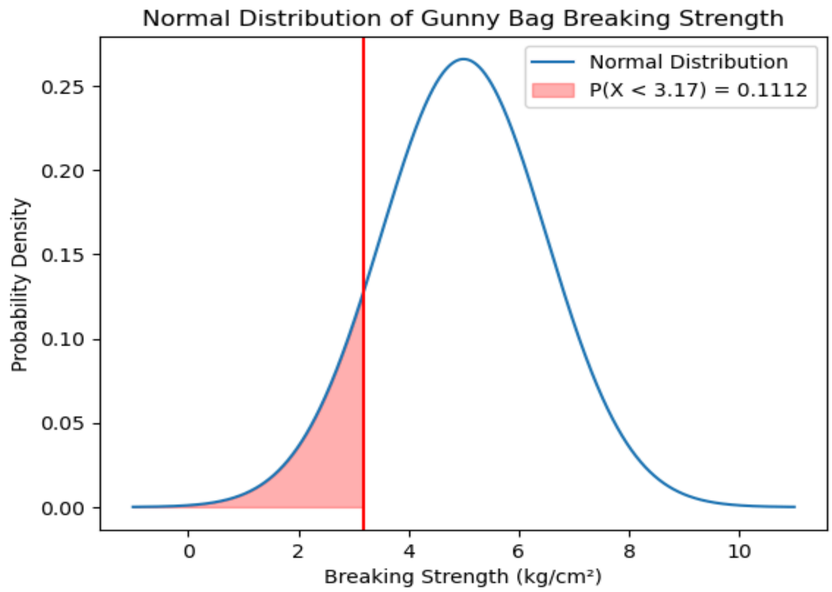
Probability that a randomly chosen injured player is a striker: 0.310

<u>3. Problem (2) : Breaking Strength of Gunny Bags</u>

3.1 Proportion Less than 3.17 kg/sq cm

To determine the proportion of gunny bags with a breaking strength less than 3.17 kg/cm², we utilized the cumulative distribution function (CDF) of the normal distribution. The CDF provides the probability that a random variable (in this case, breaking strength) is less than or equal to a specified value.

The formula used for this calculation is as follows: $P(X \leq x) = CDF(x; \mu, \sigma)$



(Fig1 : Plot showing the distribution of gunny bags with a breaking strength less than 3.17 kg/cm²)

Where:

$P(X \leq x)$ is the proportion of gunny bags with a breaking strength less than or equal to x kg/cm².

μ is the mean breaking strength.

σ is the standard deviation of breaking strength.

x is the specified breaking strength threshold.

Given the parameters:

Mean breaking strength (μ): 5 kg/cm²

Standard deviation (σ): 1.5 kg/cm²

Specified breaking strength (x): 3.17 kg/cm²

We substituted these values into the formula to calculate the proportion: P(X≤3.17)=CDF(3.17;5,1.5)

After computation, the proportion was found to be approximately (0.1112)

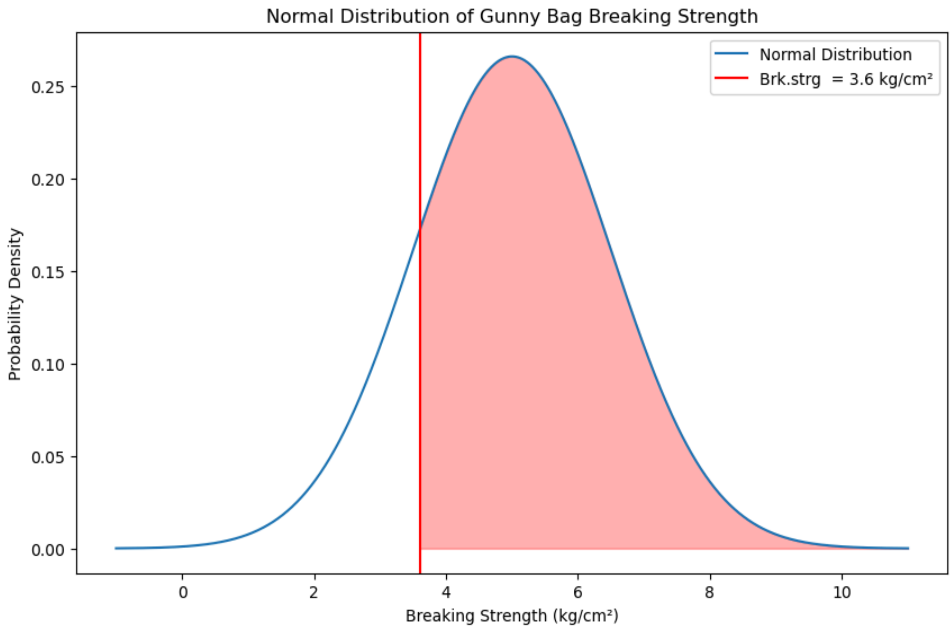
This calculation indicates that about 11.2 percent of gunny bags in our dataset have a breaking strength below 3.17 kg/cm².

3.2 Proportion at Least 3.6 kg/sq cm

To determine the proportion of gunny bags with a breaking strength of at least 3.6 kg/cm², we employed the complementary cumulative distribution function (CCDF) of the normal distribution. The CCDF provides the probability that a random variable (breaking strength) is greater than a specified value.

The formula used for this calculation is: P(X≥x)=1−CDF(x;μ,σ)

(Fig 2 : Plot showing the distribution gunny bag breaking strength of at least 3.6 kg/cm²)



Where:

P(X≥x) is the proportion of gunny bags with a breaking strength greater than or equal to x kg/cm².

μ is the mean breaking strength.

σ is the standard deviation of breaking strength.

x is the specified breaking strength threshold.

Given the parameters:

Mean breaking strength (μ): 5 kg/cm²

Standard deviation (σ): 1.5 kg/cm²

Specified breaking strength (x): 3.6 kg/cm²

We substituted these values into the formula to calculate the proportion:
$P(X \geq 3.6) = 1 - CDF(3.6; 5, 1.5)$

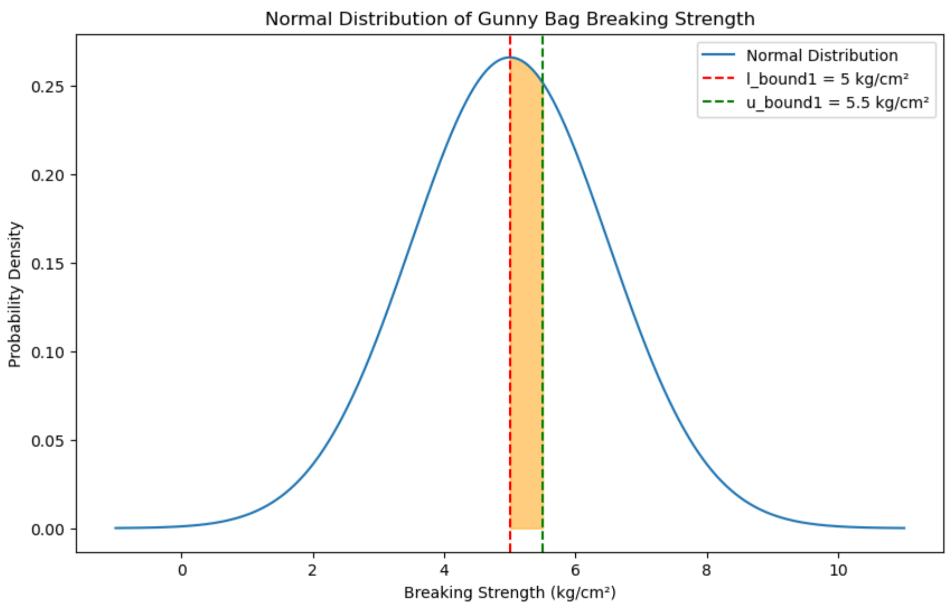
After computation, the proportion was found to be approximately (0.8247)

This calculation indicates that approximately 82.5 percent of gunny bags in our dataset have a breaking strength of at least 3.6 kg/cm².

3.3 Proportion with Breaking Strength Between 5 and 5.5 kg/cm²:

To determine the proportion of gunny bags with a breaking strength between 5 and 5.5 kg/cm², we utilized the cumulative distribution function (CDF) of the normal distribution. The CDF provides the probability that a random variable (in this case, breaking strength) falls below a specified value.

The formula used for this calculation is: $P(l \leq X \leq u) = CDF(u; \mu, \sigma) - CDF(l; \mu, \sigma)$



(Fig 3 : Plot showing the distribution gunny bags with a breaking strength between 5 and 5.5 kg/cm²)

Where:

$P(l \leq X \leq u)$ is the proportion of gunny bags with a breaking strength between l and u kg/cm².

μ is the mean breaking strength.

σ is the standard deviation of breaking strength.

l is the lower bound of breaking strength.

u is the upper bound of breaking strength.

Given the parameters:

Mean breaking strength (μ): 5 kg/cm²

Standard deviation (σ): 1.5 kg/cm²

Lower bound (l): 5 kg/cm²

Upper bound (u): 5.5 kg/cm²

We substituted these values into the formula to calculate the proportion:
$P(5 \le X \le 5.5) = CDF(5.5;5,1.5) - CDF(5;5,1.5)$

After computation, the proportion was found to be approximately (0.1306)

This calculation indicates that approximately 13 percent of gunny bags in our dataset have a breaking strength between 5 and 5.5 kg/cm².

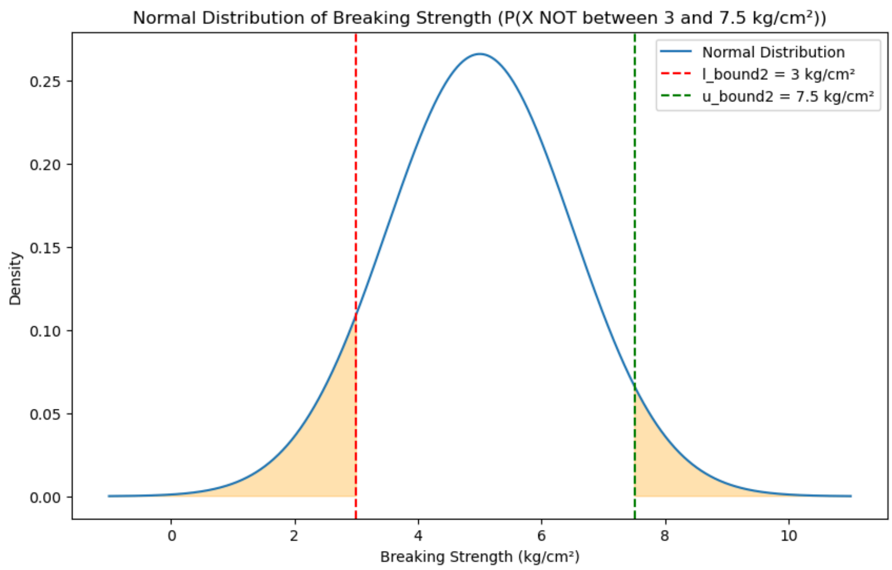
3.4 Proportion with Breaking Strength Not Between 3 and 7.5 kg/cm²:

To determine the proportion of gunny bags with a breaking strength not between 3 and 7.5 kg/cm² , we used the cumulative distribution function (CDF) of the normal distribution. The CDF provides the probability that a random variable (in this case, breaking strength) falls below a specified value.

The formula used for this calculation is:

$$P(X<l \text{ or } X>u) = CDF(l;\mu,\sigma) + (1 - CDF(u;\mu,\sigma))$$

(Fig 4 : Plot showing the distribution breaking strength not between 3 and 7.5 kg/cm²)



Where

$P(X<l \text{ or } X>u)$ is the proportion of gunny bags with a breaking strength not between l and u kg/cm².

μ is the mean breaking strength.

σ is the standard deviation of breaking strength.

l is the lower bound of breaking strength.

u is the upper bound of breaking strength.

Given the parameters:

Mean breaking strength (μ): 5 kg/cm²

Standard deviation (σ): 1.5 kg/cm²

Lower bound (l): 3 kg/cm²

Upper bound (u): 7.5 kg/cm²

We substituted these values into the formula to calculate the proportion:

$P(X<3 \text{ or } X>7.5)=CDF(3;5,1.5)+(1-CDF(7.5;5,1.5))$

After computation, the proportion was found to be approximately (0.1390)

This calculation indicates that approximately 13.9 percent of gunny bags in our dataset have a breaking strength not between 3 and 7.5 kg/cm².

4. Problem (3) : Brinell Hardness Index of Stones

Problem Definition

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.

Data Description

Zingaro_Company.csv dataset contains client's data with a batch of polished and unpolished stones

Data Overview

Load the required packages, set the working directory, and load the data file.

The dataset has 75 rows and 2 columns.

It is always a good practice to view a sample of the rows. A simple way to do that is to use head() function.

	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

(Table 3: Top five rows of the dataset Zingaro_Company)

A quick look at the dataset information tells us that there are 2 numerical and 0 categorical variables. There are no Null records present in the database.

```
<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
```

(Table 4 : Basic Information of the Dataset Zingaro_Company)

Statistical Summary

Inspecting the Summary Statistics of the Dataset (Numerical fields)

	Unpolished	Treated and Polished
count	75.000000	75.000000
mean	134.110527	147.788117
std	33.041804	15.587355
min	48.406838	107.524167
25%	115.329753	138.268300
50%	135.597121	145.721322
75%	158.215098	157.373318
max	200.161313	192.272856

(Table 5 : Numerical summarization of the dataset Zingaro_Company)

Observations:

Higher Hardness in Treated and Polished Stones:

Treated and polished stones have a higher mean and median hardness compared to unpolished stones. This indicates that the treatment and polishing process significantly enhances the hardness of the stones.

Lower Variability in Treated and Polished Stones:

The lower standard deviation in treated and polished stones (15.59) compared to unpolished stones (33.04) suggests that the treatment process results in more consistent hardness levels. This reduced variability is beneficial for applications requiring uniform hardness.

Consistent Hardness Levels:

The interquartile range (IQR), which is the difference between the 75th and 25th percentiles, is narrower for treated and polished stones (157.37 - 138.27 = 19.10) compared to unpolished stones (158.22 - 115.33 = 42.89). This further highlights the consistency in hardness achieved through the treatment process.

Wider Range in Unpolished Stones:

Unpolished stones exhibit a wider range of hardness values (48.41 to 200.16) compared to treated and polished stones (107.52 to 192.27). The presence of extremely low hardness values in unpolished stones suggests that some stones may be unsuitable for applications requiring high hardness.

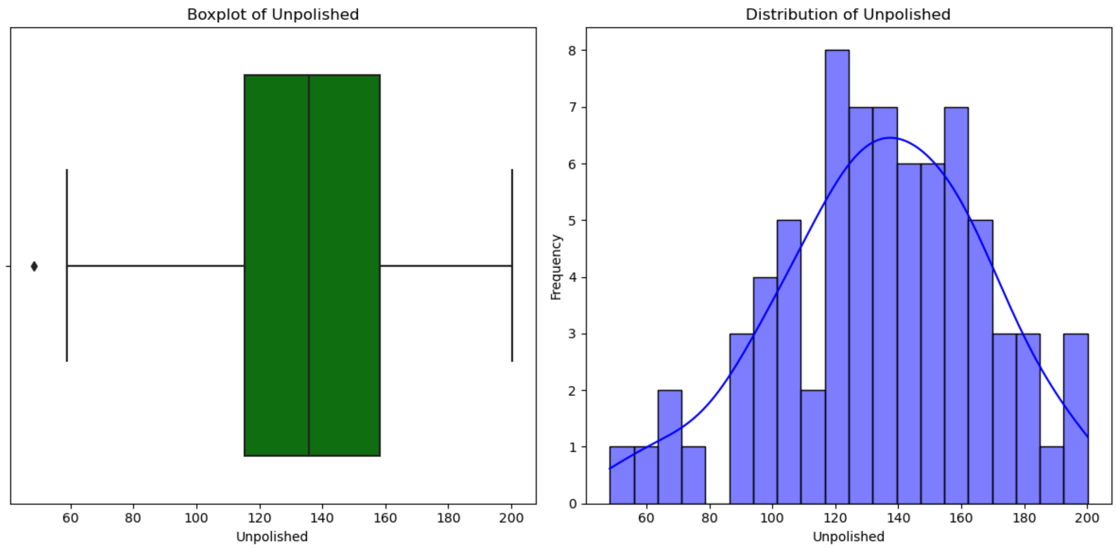
Quality Improvement: The significant improvement in mean hardness from 134.11 (unpolished) to 147.79 (treated and polished) demonstrates the effectiveness of the polishing and treatment process in enhancing the quality and performance of the stones.

Univariate Analysis

For performing Univariate analysis we will take a look at the Boxplots and Histograms to get a better understanding of the distributions.

Numerical variables

- **Observations on Unpolished**



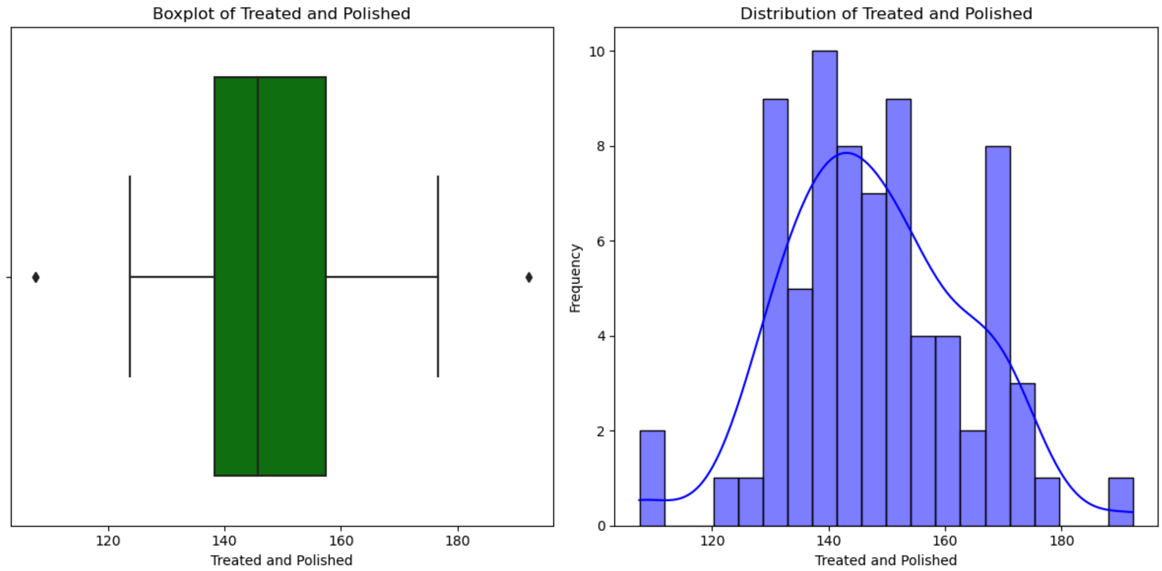
(Fig 5 : Plot showing the boxplot and distribution for Unpolished)

Observations:

● The Distribution of **Unpolished** is slightly left skewed.

● There is an outlier in this variable, We will treat the outlier.

Observations on Treated and Polished Stones



(Fig 6 : Plot showing the boxplot and distribution for Treated and Polished Stones)

Observations:

● The Distribution of **Treated and Polished Stones** is slightly left skewed.

● There are a few outliers in this variable, We will treat the outliers.

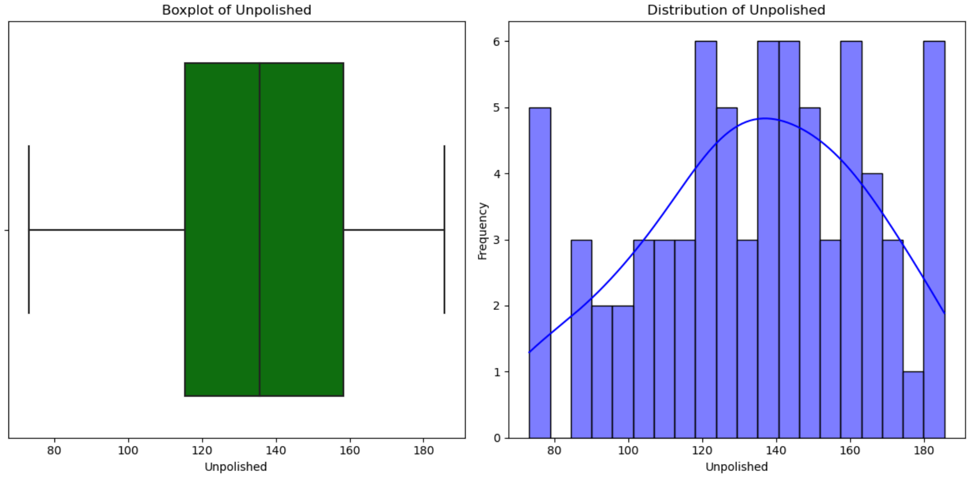
Referring to the above box plots it seems that the two variables Unpolished and Treated and Polished have outliers present in the variables.

These outlier values needs to be treated and there are several ways of treating them:**

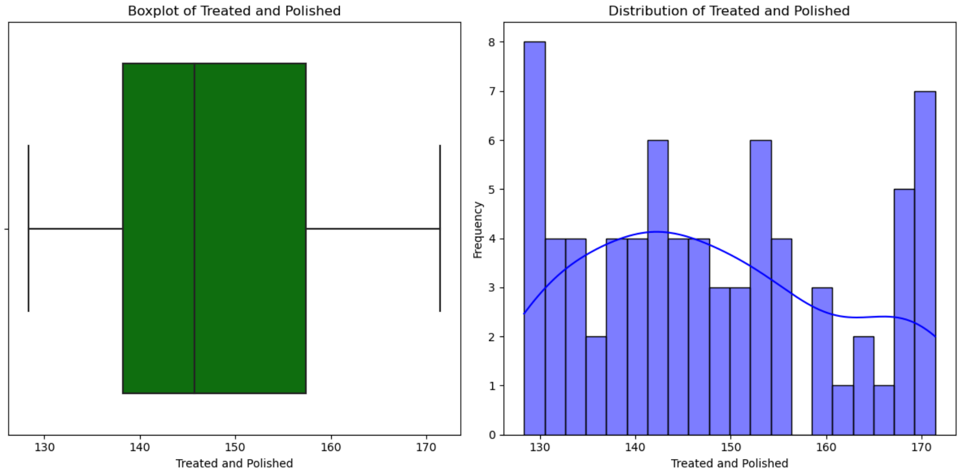
- Drop the outlier value
- Replace the outlier value using the IQR
- We can treat the outliers using percentile or with Inter quartile Range

Here, we are using percentile to replace outliers

Boxplots and Histograms after outliers treatment



(Fig 7 : The boxplot and distribution for Unpolished Stones after outlier treatment)



(Fig 8 : The boxplot and distribution for Treated and Polished Stones after outlier treatment)

	Unpolished	Treated and Polished
count	75.000000	75.000000
mean	134.334853	148.064394
std	30.623973	13.459406
min	73.178331	128.357069
25%	115.329753	138.268300
50%	135.597121	145.721322
75%	158.215098	157.373318
max	185.582012	171.416767

(Table 6 : Numerical summarization of the dataset after treating outliers)

4.1 Suitability of Unpolished Stones for Printing

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

To determine whether Zingaro is justified in thinking that the unpolished stones may not be suitable for printing, we can perform a hypothesis test on the mean Brinell hardness index of the unpolished stones. We'll assume a significance level of 5% ($\alpha = 0.05$). The null hypothesis (H_0) and the alternative hypothesis (H_1) can be stated as follows:

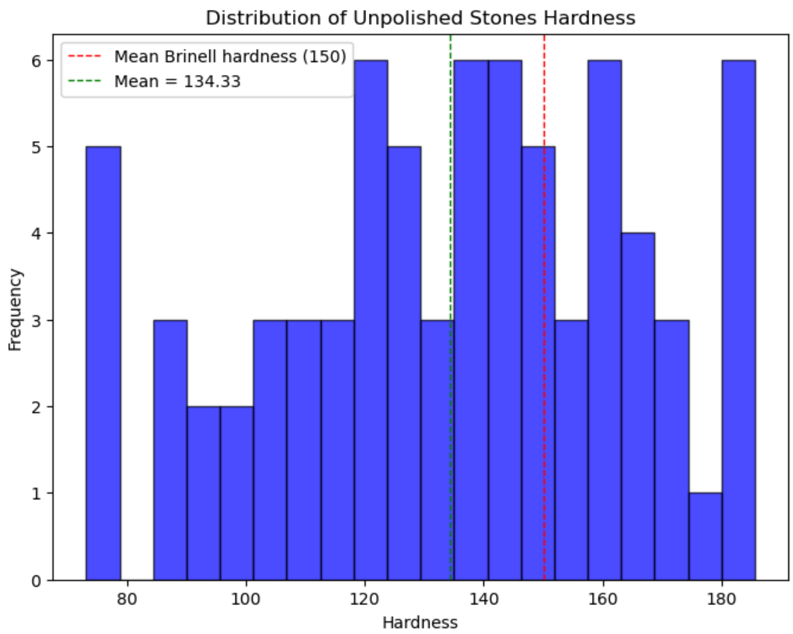
Null Hypothesis (H 0): The mean Brinell hardness of unpolished stones is at least 150. ($\mu \geq 150$)

Alternative Hypothesis (H 1): The mean Brinell hardness of unpolished stones is less than 150. ($\mu < 150$)

Sample Data:

- Sample Mean (x): 134.33
- Sample Standard Deviation (s): 30.62
- Sample Size (n): 75
- Population Mean (μ): 150

The t-statistic for a one-sample t-test is calculated as follows: $t = \frac{134.33-150}{30.62/\sqrt{75}} = \frac{-15.67}{3.54} \approx -4.43$



The one-tailed p-value is approximately 0.00002

Decision Rule:

At the 5% significance level (α=0.05), if the p-value is less than 0.05 and the t-statistic is negative, we reject the null hypothesis.

(Fig 9 : Histogram showing the distribution for Unpolished Stones hardness)

Conclusion:

Since the one-tailed p-value is approximately

0.00002

0.00002, which is much less than 0.05, and the t-statistic is -4.43, which is negative, we reject the null hypothesis. This provides sufficient evidence to conclude that the mean Brinell hardness index of unpolished stones is significantly less than 150.

There is strong evidence to support the claim that the unpolished stones have a mean Brinell hardness index less than 150. Therefore, Zingaro is justified in thinking that the unpolished stones may not be suitable for printing.

4.2 Comparison of Mean Hardness Between Polished and Unpolished Stones

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

Null Hypothesis : The mean hardness of polished stones is equal to the mean hardness of unpolished stones

Alternative Hypothesis : The mean hardness of polished stones is not equal to the mean hardness of unpolished stones**.**

Polished Stones:

Sample size: 75

Mean hardness: 148.06

Standard deviation: 13.46

Unpolished Stones:

Sample size: 75

Mean hardness: 134.33

Standard deviation: 30.62

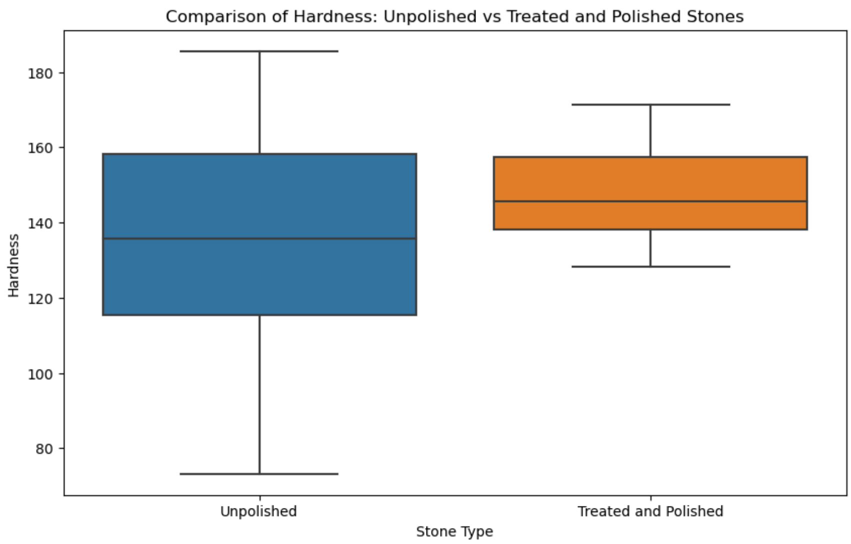
Test Statistic (t-value): $t \approx 3.55$

Degrees of Freedom (df): 148

Critical t-value (at $\alpha=0.05$): ±1.976

Results:

The calculated t-value (3.49) exceeds the critical t-value (±1.976), leading to the rejection of the null hypothesis. Therefore, there is a statistically significant difference in the mean hardness of polished and unpolished stones.



(**Fig 10 : Boxplot showing the comparison of Unpolished vs Treated and Polished Stones hardness**)

Conclusion:The findings suggest that the level of hardness varies between polished and unpolished stones, which could have implications for the printing process at Zingaro Stone Printing. Further investigation into the factors affecting stone hardness may be warranted to ensure optimal printing outcomes.

5. Problem (4) : Hardness of Dental Implants

Problem Definition :

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.

Data Description

Dental_Hardness_data.csv dataset contains Dental implant data.

Data Overview

Load the required packages, set the working directory, and load the data file.

The dataset has 107 rows and 14 columns. It is always a good practice to view a sample of the rows. A simple way to do that is to use head() function.

	Dentist	Method	Alloy	Temp	Response	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	Unnamed: 10	Unnamed: 11	Unnamed: 12	Unnamed: 13
0	1.0	1.0	1.0	1500.0	813.0	NaN	NaN	Anova: Two-Factor Without Replication	NaN	NaN	NaN	NaN	NaN	NaN
1	1.0	1.0	1.0	1600.0	792.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	1.0	1.0	1.0	1700.0	792.0	NaN	NaN	SUMMARY	Count	Sum	Average	Variance	NaN	NaN
3	1.0	1.0	2.0	1500.0	907.0	NaN	NaN	1	4	2315	578.75	523721.583333	NaN	NaN
4	1.0	1.0	2.0	1600.0	792.0	NaN	NaN	1	4	2394	598.5	584819	NaN	NaN

(Table 7 : Top 5 rows of the dataset Dental_Hardness_data)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 107 entries, 0 to 106
Data columns (total 14 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Dentist      90 non-null     float64
 1   Method       90 non-null     float64
 2   Alloy        90 non-null     float64
 3   Temp         90 non-null     float64
 4   Response     90 non-null     float64
 5   Unnamed: 5   0 non-null      float64
 6   Unnamed: 6   0 non-null      float64
 7   Unnamed: 7   102 non-null    object
 8   Unnamed: 8   100 non-null    object
 9   Unnamed: 9   100 non-null    object
 10  Unnamed: 10  99 non-null     object
 11  Unnamed: 11  98 non-null     object
 12  Unnamed: 12  3 non-null      object
 13  Unnamed: 13  3 non-null      object
dtypes: float64(7), object(7)
memory usage: 11.8+ KB
```

(Table 8 : Basic info of the dataset Dental_Hardness_data)

Removing the columns named , 'Unnamed (5-13)' , as it is not relevant with the problem statements , and creating a new temp df

	Dentist	Method	Alloy	Temp	Response
0	1.0	1.0	1.0	1500.0	813.0
1	1.0	1.0	1.0	1600.0	792.0
2	1.0	1.0	1.0	1700.0	792.0
3	1.0	1.0	2.0	1500.0	907.0
4	1.0	1.0	2.0	1600.0	792.0
...	...	...	...	...	...
85	5.0	3.0	1.0	1600.0	483.0
86	5.0	3.0	1.0	1700.0	405.0
87	5.0	3.0	2.0	1500.0	536.0
88	5.0	3.0	2.0	1600.0	405.0
89	5.0	3.0	2.0	1700.0	312.0

Removing the NaN (missing) values using the dropna() method.

(Table 9 : new df after removing the irrelevant columns and Nan Values)

Inspecting the Summary Statistics of the Dataset

	Dentist	Method	Alloy	Temp	Response
count	90.000000	90.000000	90.000000	90.000000	90.000000
mean	3.000000	2.000000	1.500000	1600.000000	741.777778
std	1.422136	0.821071	0.502801	82.107083	145.767845
min	1.000000	1.000000	1.000000	1500.000000	289.000000
25%	2.000000	1.000000	1.000000	1500.000000	698.000000
50%	3.000000	2.000000	1.500000	1600.000000	767.000000
75%	4.000000	3.000000	2.000000	1700.000000	824.000000
max	5.000000	3.000000	2.000000	1700.000000	1115.000000

(Table 10 : Statistical Summary of the df after cleanup)

Observations

Dentist: The responses range from 1 to 5, with a mean of 3. This indicates that most responses are centered around the higher end of the scale.

Method: The responses range from 1 to 3, with a mean of 2. It seems that there's a relatively even distribution among the three methods, with a slight preference for method 2.

Alloy: The responses range from 1 to 2, with a mean of 1.5. It appears that both alloys are commonly used, with no significant preference for one over the other.

Temperature: The temperatures range from 1500 to 1700 degrees, with a mean of 1600 degrees. This suggests that the majority of responses fall within this temperature range, with 1600 degrees being the most common.

Response: The response times range from 289 to 1115 units, with a mean of approximately 741.77 units. There seems to be some variability in response times, with a notable standard deviation of 145.77 units. We'll be treating the outliers if any.

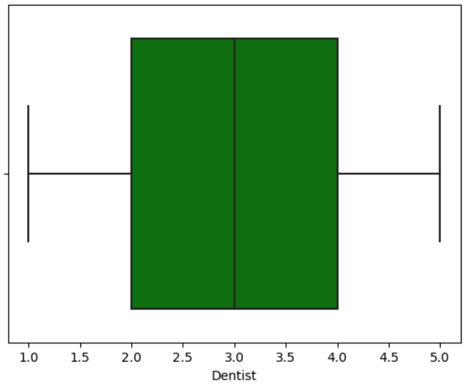
Overall, the data indicates that there's diversity in the methods and alloys used by dentists, but there's some consistency in the temperature and response time of their processes.

Univariate Analysis

For performing Univariate analysis we will take a look at the Boxplots to get better understanding of the distributions.

Observations on Dentist

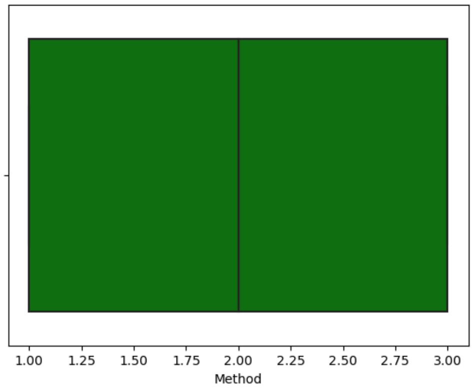
- **Follows a normal dist**
- **No outliers**



(Fig 11 : Boxplot showing the distribution of Dentist variable)

Observations on Method

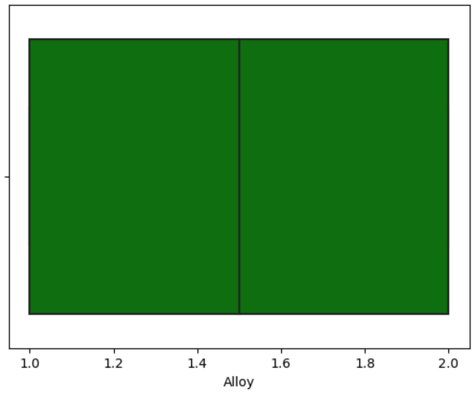
- **Follows a normal dist**
- **No Outliers**



(Fig 12 : Boxplot showing the distribution of Method variable)

Observations on Alloy

- **Follows a normal dist**
- **No Outliers**

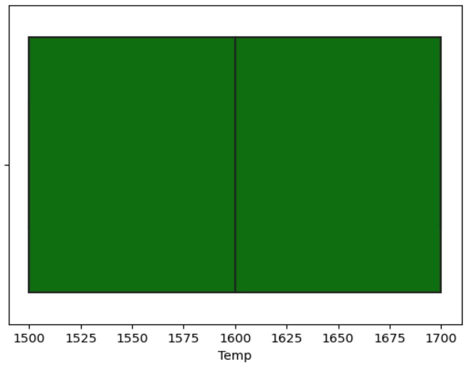


(Fig 13 : Boxplot showing the distribution of Alloy variable)

Observations on Temp

- **Follows a normal dist**
- **No Outliers**

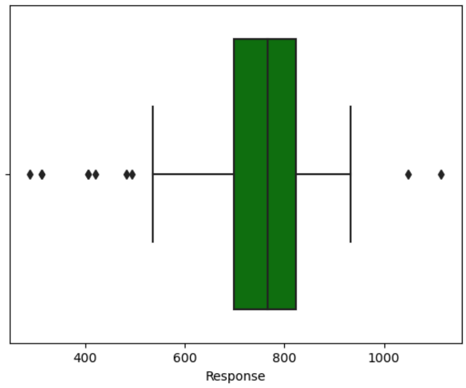
(Fig 14 : Boxplot showing the distribution of Temp variable)



Observations on Response

- **Skewed to the right**
- **Few outliers**

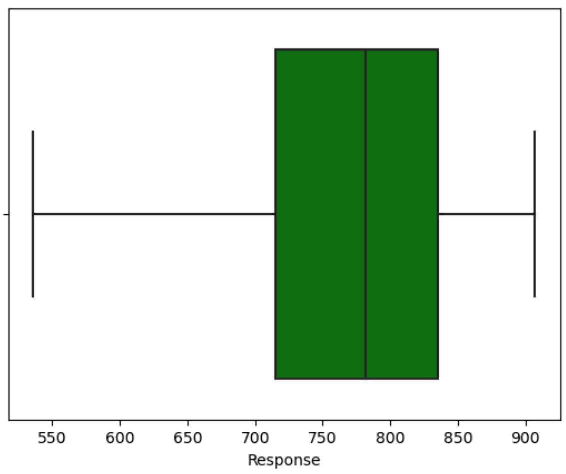
(Fig 15 : Boxplot showing the distribution of Response variable)



Referring to the above box plots it seems that the variables 'Response' have outliers present in the variables.

These outlier values needs to be treated and there are several ways of treating them:

- Drop the outlier value
- Replace the outlier value using the IQR
- We can treat the outliers using percentile or with Inter quartile Range



Here, we are using IQR to replace outliers

(Fig 16 : Boxplot showing the distribution of Response variable after treating outliers)

5.1 Variation by Dentists

How does the hardness of implants vary depending on dentists?

Hypotheses: For each alloy type, we can set up the following hypotheses:

Null Hypothesis (H0): There is no significant difference in the mean hardness of implants among dentists for the given alloy type.

Alternative Hypothesis (H1): There is a significant difference in the mean hardness of implants among dentists for the given alloy type.

ANOVA table for Alloy 1:

	df	sum_sq	mean_sq	F	PR(>F)
C(Dentist)	4.0	14633.1	3658.275000	0.953882	0.444837
Residual	35.0	134230.0	3835.142857	NaN	NaN

(Table 11 : ANOVA table for Alloy 1 (dentist))

The ANOVA test examines whether there are significant differences in mean implant hardness among different dentists for Alloy 1.

The p-value (PR(>F)) associated with the Dentist factor is 0.444837, which is greater than the typical significance level of 0.05. Therefore, we fail to reject the null hypothesis.

The F-statistic is 0.953882, indicating that there is not enough evidence to suggest significant differences in mean implant hardness among dentists for Alloy 1.

In conclusion, based on this ANOVA analysis, there is no statistically significant variation in mean implant hardness among different dentists for Alloy 1.

ANOVA table for Alloy 2:

	df	sum_sq	mean_sq	F	PR(>F)
C(Dentist)	4.0	5014.531746	1253.632937	0.123079	0.973296
Residual	37.0	376867.873016	10185.618190	NaN	NaN

(Table 12 : ANOVA table for Alloy 2 (dentist))

The ANOVA test examines whether there are significant differences in mean implant hardness among different dentists for Alloy 2.

The p-value (PR(>F)) associated with the Dentist factor is 0.973296, which is much greater than the typical significance level of 0.05. Therefore, we fail to reject the null hypothesis.

The F-statistic is 0.123079, indicating that there is not enough evidence to suggest significant differences in mean implant hardness among dentists for Alloy 2.

In conclusion, based on this ANOVA analysis, there is no statistically significant variation in mean implant hardness among different dentists for Alloy 2.

5.2 Variation by Methods

How does the hardness of implants vary depending on methods?

Null Hypothesis (H0): *There is no significant difference in the mean hardness of implants among different methods for the given alloy type.*

Alternative Hypothesis (H1)*: There is a significant difference in the mean hardness of implants among different methods for the given alloy type.*

ANOVA table for Alloy 1:

	df	sum_sq	mean_sq	F	PR(>F)
C(Method)	2.0	2622.966667	1311.483333	0.331816	0.719732
Residual	37.0	146240.133333	3952.436036	NaN	NaN

(Table 13 : ANOVA table for Alloy 1 (Methods))

The ANOVA test examines whether there are significant differences in mean implant hardness among different methods for Alloy 1.

The p-value (PR(>F)) associated with the Method factor is 0.719732, which is much greater than the typical significance level of 0.05. Therefore, we fail to reject the null hypothesis.

The F-statistic is 0.331816, indicating that there is not enough evidence to suggest significant differences in mean implant hardness among methods for Alloy 1.

In conclusion, based on this ANOVA analysis, there is no statistically significant variation in mean implant hardness among different methods for Alloy 1.

ANOVA table for Alloy 2:

	df	sum_sq	mean_sq	F	PR(>F)
C(Method)	2.0	165291.221429	82645.610714	14.881394	0.000016
Residual	39.0	216591.183333	5553.620085	NaN	NaN

(Table 14 : ANOVA table for Alloy 2 (Methods))

The ANOVA test examines whether there are significant differences in mean implant hardness among different methods for Alloy 2.

The p-value (PR(>F)) associated with the Method factor is very small (0.000016), which is much less than the typical significance level of 0.05. Therefore, we reject the null hypothesis.

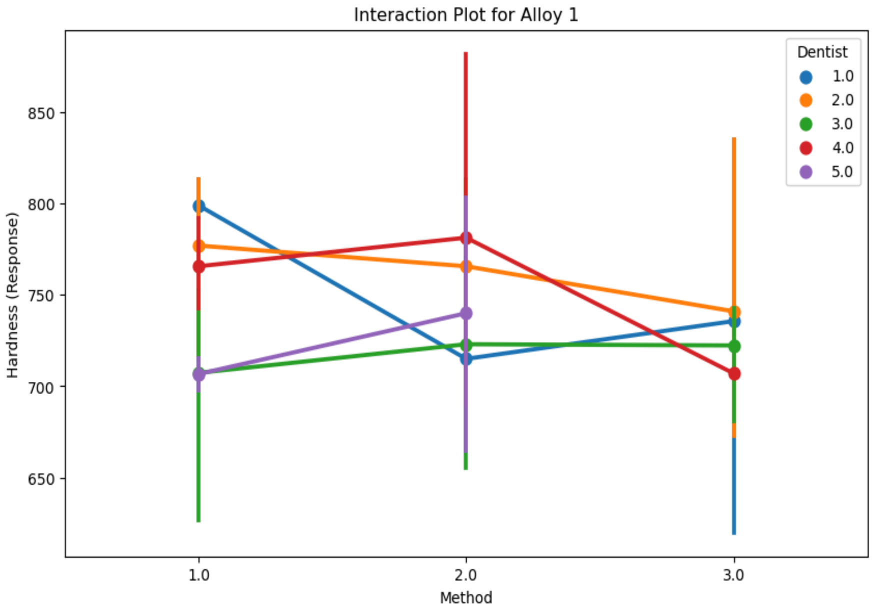
The F-statistic is 14.881394, indicating that there is evidence to suggest significant differences in mean implant hardness among methods for Alloy 2.

In conclusion, based on this ANOVA analysis, there is statistically significant variation in mean implant hardness among different methods for Alloy 2.

5.3 Interaction Effect Between Dentist and Method

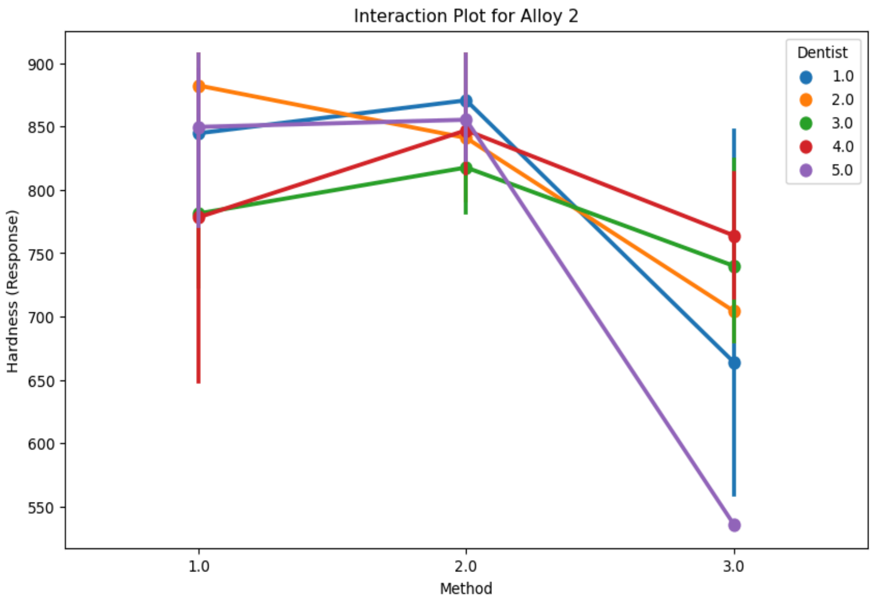
What is the interaction effect between the dentist and method on the hardness of dental implants for each type of alloy?

Create Interaction Plot



(Fig 17 : Pointplot showing the interaction effect between the dentist and method on the hardness of dental implants for Alloy 1)

Conclusion: There is a significant interaction between dentist and method for Alloy 1. The method's effect on hardness depends on which dentist is performing the implant.



(Fig 18 : Pointplot showing the interaction effect between the dentist and method on the hardness of dental implants for Alloy 2)

Conclusion: There is no significant interaction between dentist and method for Alloy 2. The method's effect on hardness is similar for all dentists.

5.4 Combined Effect of Dentists and Methods

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

Null Hypotheses (H0):

There is no difference in hardness of implants among different dentists.

There is no difference in hardness of implants among different methods.

There is no interaction effect between dentists and methods on the hardness of implants.

Alternative Hypotheses (H1):

There is a difference in hardness of implants among different dentists.

There is a difference in hardness of implants among different methods.

There is an interaction effect between dentists and methods on the hardness of implants.

	df	sum_sq	mean_sq	F	PR(>F)
C(Dentist)	4.0	14633.100000	3658.275000	0.831378	0.517482
C(Method)	2.0	3390.716667	1695.358333	0.385286	0.684075
C(Dentist):C(Method)	8.0	17024.950117	2128.118765	0.483635	0.856556
Residual	26.0	114406.666667	4400.256410	NaN	NaN

(Table 15 :ANOVA table for Alloy 1)

Dentist

The p-value (0.517482) is greater than the significance level (usually 0.05), indicating that we fail to reject the null hypothesis.

Conclusion: There is no significant difference in the hardness of implants among different dentists.

Method

The p-value (0.684075) is greater than the significance level, indicating that we fail to reject the null hypothesis.

Conclusion: There is no significant difference in the hardness of implants among different methods.

Dentist (Interaction between Dentist and Method):

The p-value (0.856556) is greater than the significance level, indicating that we fail to reject the null hypothesis.

Conclusion: There is no significant interaction effect between dentists and methods on the hardness of implants.

	df	sum_sq	mean_sq	F	PR(>F)
C(Dentist)	4.0	5014.531746	1253.632937	0.236412	0.915327
C(Method)	2.0	165209.192674	82604.596337	15.577717	0.000032
C(Dentist):C(Method)	8.0	68484.680342	8560.585043	1.614370	0.167194
Residual	27.0	143174.000000	5302.740741	NaN	NaN

(Table 16 :ANOVA table for Alloy 2)

Dentist: The p-value (0.915327) is much greater than the significance level (usually 0.05), indicating that we fail to reject the null hypothesis.

Conclusion: There is no significant difference in the hardness of implants among different dentists.

Method : The p-value (0.000032) is much smaller than the significance level, indicating that we reject the null hypothesis**.**

Conclusion: There is a significant difference in the hardness of implants among different methods.

Dentist(Interaction between Dentist and Method): The p-value (0.167194) is greater than the significance level, indicating that we fail to reject the null hypothesis.

Conclusion: There is no significant interaction effect between dentists and methods on the hardness of implants.

6. References

- **Previous jupyter notebooks**
- **Google**
- **Course Resources**