# Business Report for the

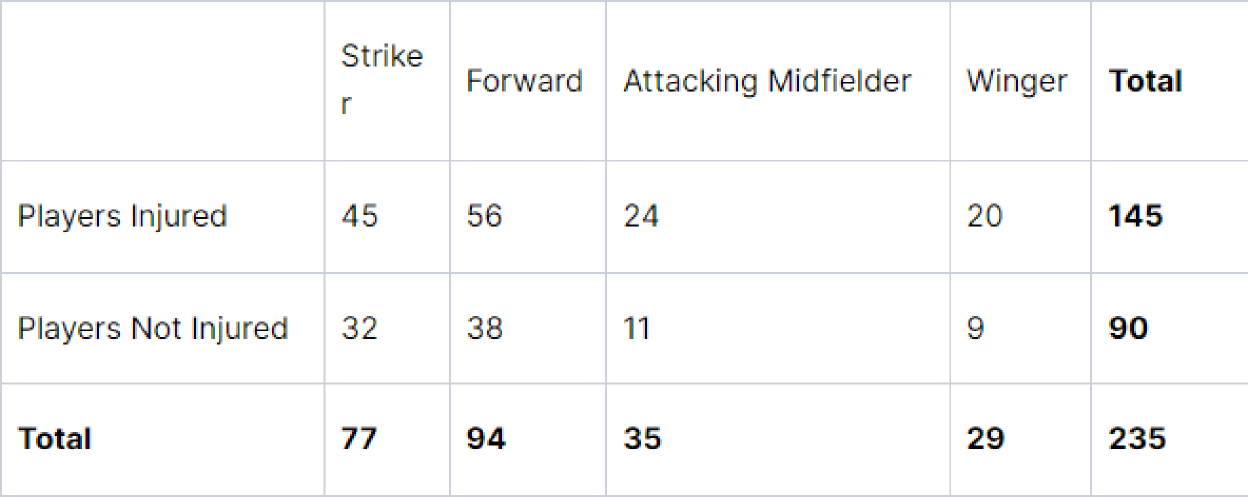
# Inferential Statistics Project

Name: Aishwariya Hariharan
PGP-DSBA Online September’ 23
Date: 10/12/2023

# Contents

# 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?**

Code:

```
total_injured = 145
total_players = 235
prob_injury = (total_injured/total_players)*100
print('Probability that a randomly chosen player would suffer an injury is %1.1f' % prob_injury +'%')
```

Answer:

**The probability that a randomly chosen player would suffer an injury is 61.7%.**

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

Code:

```
total_forward = 94
```

```
total_winger = 29
prob_forward = total_forward/total_players
prob_winger = total_winger/total_players
#Mutually exclusive events
prob_forward_or_winger = (prob_forward + prob_winger)*100
print('Probability that a player is a forward or a winger is %1.1f' % prob_forward_or_winger +'%')
```

Answer:

**The probability that a player is a forward or a winger is 52.3%.**

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

Code:

```
No_of_injured_strikers = 45
prob_that_random_person_is_injured_striker = (No_of_injured_strikers/total_players)*100
print('Probability that a randomly chosen player plays in a striker position and has a foot injury is %1.1f' % prob_that_random_person_is_injured_striker +'%')
```

Answer:

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

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

Code:

```
total_injured = 145
prob_that_random_injured_player_is_striker = (No_of_injured_strikers/total_injured)*100
print('Probability that a randomly chosen injured player is a striker is %1.1f' % prob_that_random_injured_player_is_striker +'%')
```

Answer:

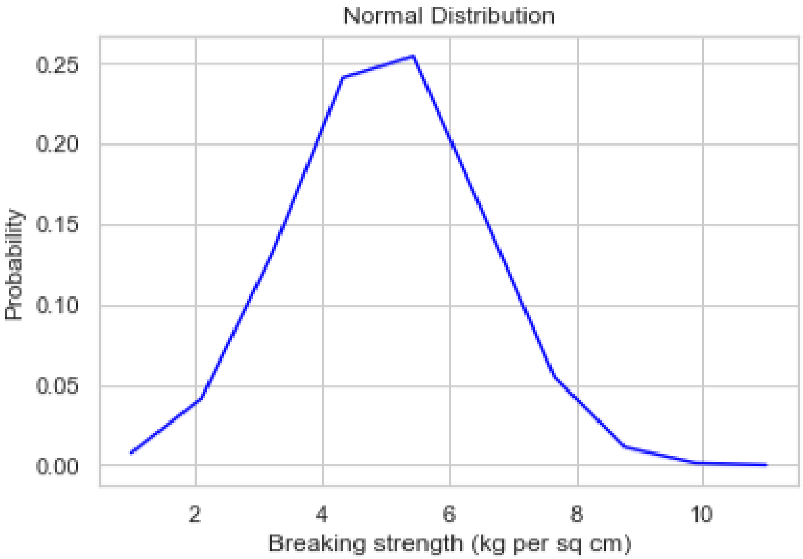
**The probability that a randomly chosen injured player is a striker is 31.0%.**

# 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.

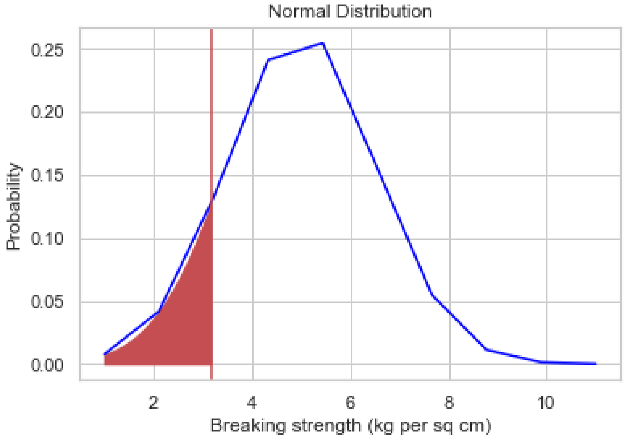
**About data:**

- We have created an array of 10 numbers in the range of 1 to 10 using linespace for the breaking strength parameter.
- After calculating the probability distribution function (pdf) for the breaking strength, we plot the distribution to observe that the dataset is approximately normally distributed.



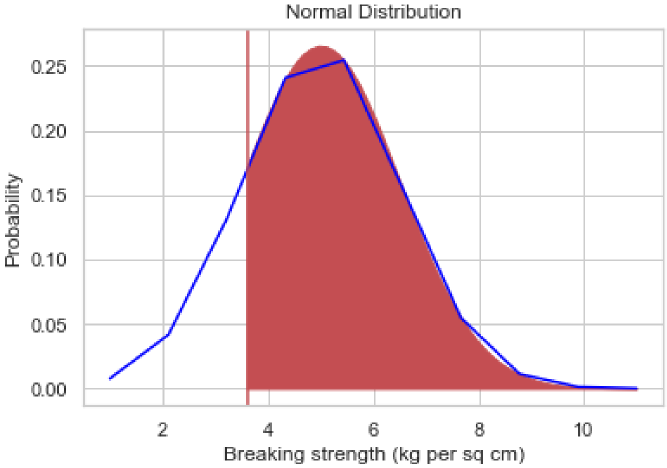
### 2.1 What proportion of the gunny bags have a breaking strength of less than 3.17 kg per sq cm?

**The proportion of the gunny bags that have a breaking strength of less than 3.17 kg per sq cm is 11.1%.**



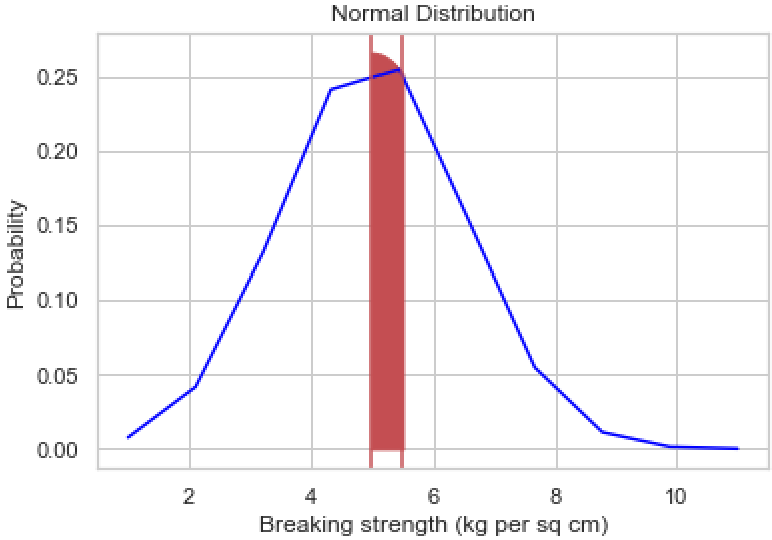
### 2.2 What proportion of the gunny bags have a breaking strength of at least 3.6 kg per sq cm.?

**The proportion of the gunny bags that have a breaking strength of at least 3.6 kg per sq cm. is 82.5%.**



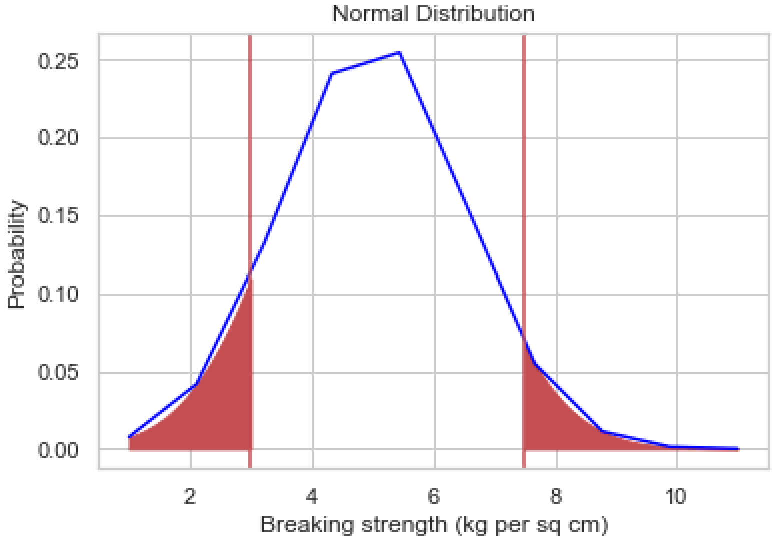
**2.3 What proportion of the gunny bags have a breaking strength between 5 and 5.5 kg per sq cm.?**

**The proportion of the gunny bags that have a breaking strength between 5 and 5.5 kg per sq cm. is 13.1%.**



**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 the gunny bags that have a breaking strength NOT between 3 and 7.5 kg per sq cm.is 13.9%.**

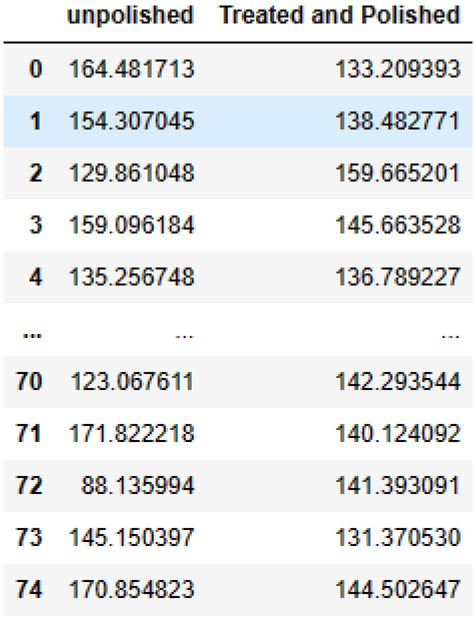


# Problem 3

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. Use the data provided to answer the following (assuming a 5% significance level);

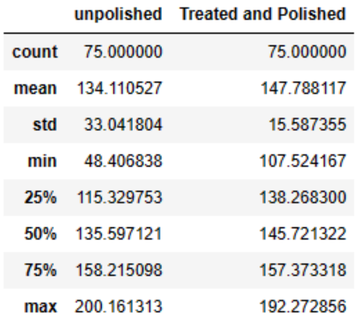
**About data:**

The dataset has 75 rows (Brinell's hardness index) and 2 columns (Unpolished and Treated & Polished).



| | 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 |
| ... | ... | ... |
| 70 | 123.067611 | 142.293544 |
| 71 | 171.822218 | 140.124092 |
| 72 | 88.135994 | 141.393091 |
| 73 | 145.150397 | 131.370530 |
| 74 | 170.854823 | 144.502647 |

The summary of data is as follows:



| | 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 |

### 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?

We need to check whether unpolished stones have an average Brinell's hardness index of at least 150 (i.e. >= 150). We are not comparing between unpolished and polished stones.
Therefore, we use the one-sample t-test to solve the problem.

- Null hypothesis: mean hardness >= 150 (unpolished stones are suitable for printing)
- Alternate hypothesis: mean hardness < 150 (i.e. unpolished stones are unsuitable for printing)
- Significance level (α) = 0.05

We get:

- T-statistic = -4.1646296
- P-value = $8.34257399 \times 10^{-5}$

Since p-value < level of significance, we reject the null hypothesis. There is sufficient evidence to suggest that the mean Brinell's hardness index of the unpolished stones is less than 150. Unpolished stones are unsuitable for printing.

**Zingaro is justified in thinking that unpolished stones may not be suitable for printing.**

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

We have two samples and we do not know the population standard deviation.

Sample sizes for both samples are the same.

We can use the two-sample t-test.

- Null hypothesis: The mean of Brinell's hardness index of both is the same.
- Alternate hypothesis: The mean of Brinell's hardness index of both is not the same.

We get:

- t-statistic = -3.242232050141406
- p-value = 0.001465515019462831

Since p-value < level of significance, we have enough evidence to reject the null hypothesis in favour of the alternative hypothesis.

**The mean of the Brinell hardness index of both are not the same.**

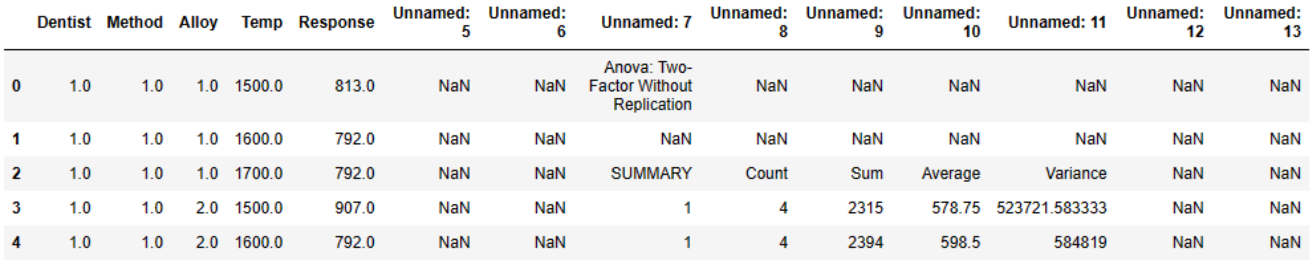
# Problem 4

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 favour one method above another and may work better in his/her favourite method. The response is the variable of interest.

**About data:**

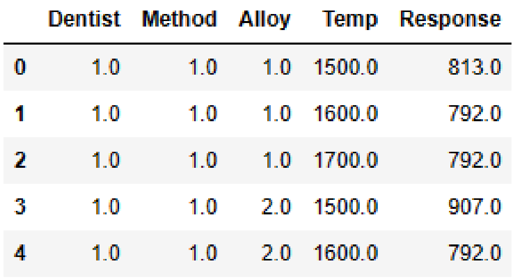
We are informed that the collected sample is a simple random sample.

The initial dataset (top 5 rows) looked like this:



| | 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 |

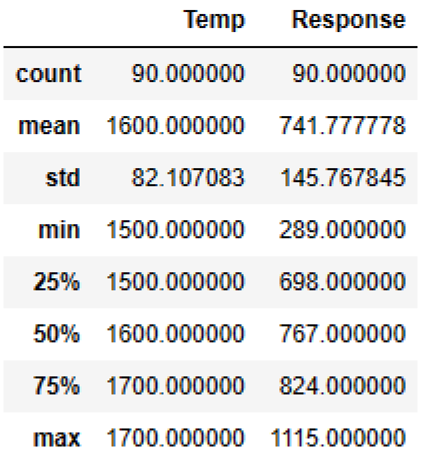
After the data preprocessing, i.e. treating duplicate data and null values, the dataset (top 5 rows) looks like this:



| | 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 |

The dataset has 90 rows and 5 columns.

The summary of data is as follows:

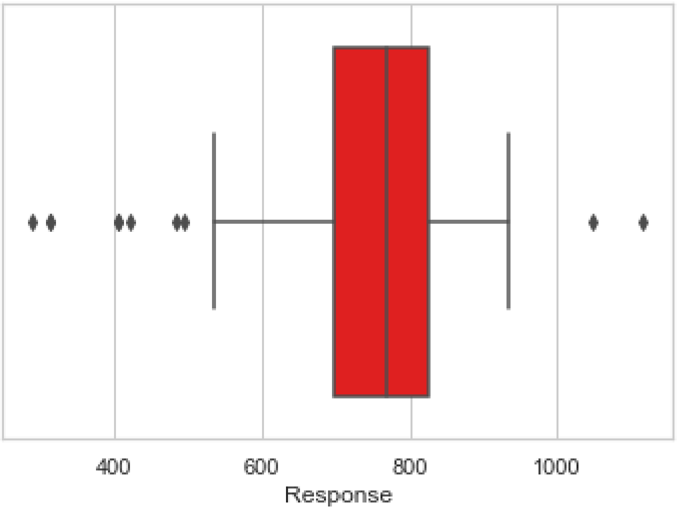


| | Temp | Response |
|---|---|---|
| **count** | 90.000000 | 90.000000 |
| **mean** | 1600.000000 | 741.777778 |
| **std** | 82.107083 | 145.767845 |
| **min** | 1500.000000 | 289.000000 |
| **25%** | 1500.000000 | 698.000000 |
| **50%** | 1600.000000 | 767.000000 |
| **75%** | 1700.000000 | 824.000000 |
| **max** | 1700.000000 | 1115.000000 |

The following columns are **categorical** with unique values as mentioned below:

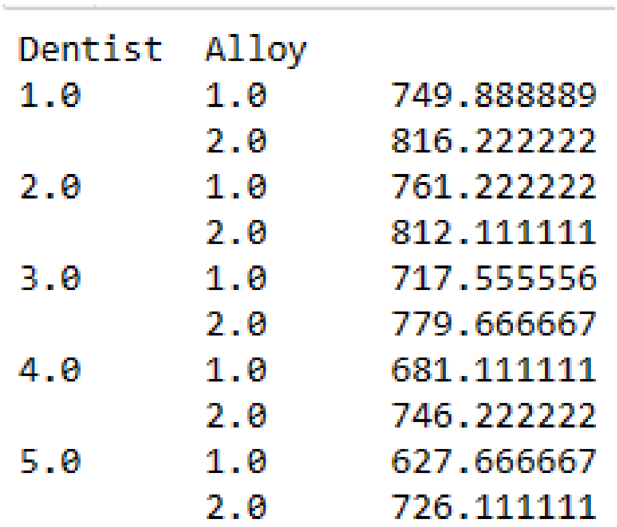
- Dentist: 1, 2, 3, 4, 5 (the sample data has observed 5 different dentists)
- Method: 1, 2, 3 (there are 3 different methods used for dental implantation)
- Alloy: 1, 2 (a choice of 2 different alloys are available)

Response is the hardness of implants, the dependent variable of interest. It has quite some outliers which could affect the analysis.

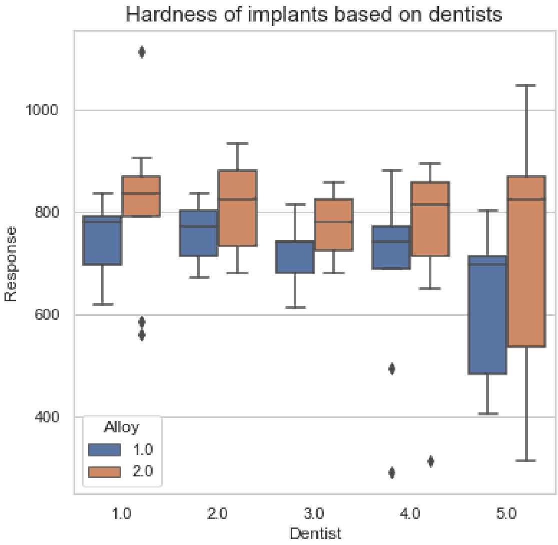


### 4.1 How does the hardness of implants vary depending on dentists?

The mean hardness of implants for the given sample of data based on the dentists who treat and the alloys preferred is as follows:



```
Dentist  Alloy
1.0      1.0      749.888889
         2.0      816.222222
2.0      1.0      761.222222
         2.0      812.111111
3.0      1.0      717.555556
         2.0      779.666667
4.0      1.0      681.111111
         2.0      746.222222
5.0      1.0      627.666667
         2.0      726.111111
```



**We can observe that Alloy 2 has a higher mean hardness of implants for the given sample.**

We will test how the hardness of implants varies depending on dentists for Alloy 1 and Alloy 2 separately by creating subsets of the data for Alloy 1 and Alloy 2 each.

**Alloy 1:**

We first check if the data is normally distributed using **Shapiro-Wilk's test.**

- Null hypothesis H0: Hardness of implants follows a normal distribution.
- Alternate hypothesis Ha: Hardness of implants does not follow a normal distribution.

We get the p-value as 1.1945070582441986 x $10^{-5}$ which is very small compared to our significance level of 0.05.

Therefore, we reject the null hypothesis that the hardness of implants follows a normal distribution.

Next, we use **Levene's test** to check the homogeneity of variances.

- H0: All the population variances are equal.

- Ha: At least one variance is different from the rest.

We get the p-value as 0.2565537418543795, which is greater than 0.05
Hence, we fail to reject the null hypothesis of homogeneity of variances.

We use **one-way Anova** to test our hypothesis.

Let μ1, μ2, μ3, μ4, and μ5 be the means of hardness of implants based on the dentist, respectively.

- Null Hypothesis H0: μ1 = μ2 = μ3 = μ4 = μ5
- Alternate hypothesis Ha: atleast one of the means of hardness of implants is different from the rest.

We get a p-value of 0.11656712140267628, which is greater than 0.05
**Therefore, we fail to reject the null hypothesis that the means of hardness of implants based on the dentist are equal for Alloy 1.**

**The hardness of implants is not affected by who the dentist is for Alloy 1.**

**Alloy 2:**

The hypothesis assumptions are the same as we did for Alloy 1. The results of the validations of the assumptions are as follows:

**Shapiro-Wilk's test:**

P-value = 0.00040292771882377565 < 0.05
Since the p-value is very small, we reject the null hypothesis that the hardness of implants follows a normal distribution.

**Levene's test:**

P-value = 0.23686777576324952
Since the p-value is large, we fail to reject the null hypothesis of homogeneity of variances.

**One-way Anova test:**

Let μ1, μ2, μ3, μ4, and μ5 be the means of hardness of implants based on dentists, respectively.

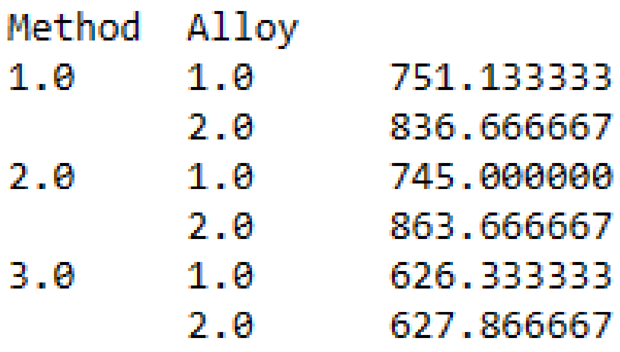
- Null Hypothesis: H0: μ1 = μ2 = μ3 = μ4 = μ5
- Alternate hypothesis: Ha: atleast one of the means of hardness of implants is different from the rest.

P-value = 0.7180309510793431
The p-value is greater than 0.05.
**Therefore, we fail to reject the null hypothesis that the means of the hardness of implants based on the dentists are equal for Alloy 2.**

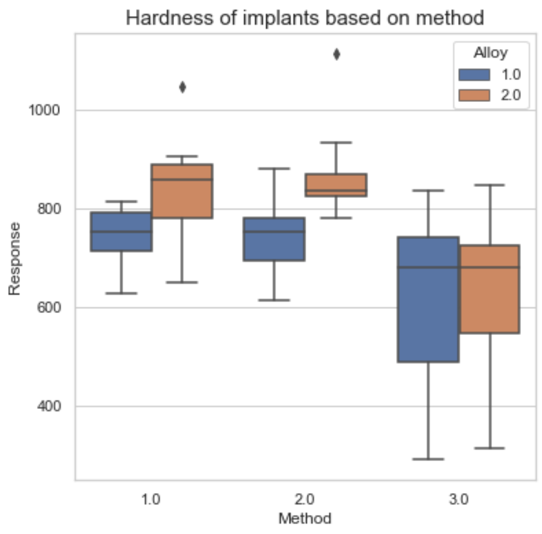
**The hardness of implants is not affected by who the dentist is for Alloy 2.**

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

The mean hardness of implants for the given sample of data based on the methods and the alloys preferred is as follows:



```
Method  Alloy
1.0     1.0      751.133333
        2.0      836.666667
2.0     1.0      745.000000
        2.0      863.666667
3.0     1.0      626.333333
        2.0      627.866667
```



**We can observe that Alloy 2 has a higher mean hardness of implants for the given sample.**

We will test how the hardness of implants varies depending on methods for Alloy 1 and Alloy 2 separately by creating subsets of the data for Alloy 1 and Alloy 2 each.

We know that the data, i.e. the response values are not normally distributed

## Alloy 1

We perform Levene's test since we are checking the variance of the response variable based on the method.

**Levene's test:**

p-value = 0.0034160381460233975
Since the p-value is less than 0.05, we reject the null hypothesis of homogeneity of variances.

**One-way Anova test:**

Let $\mu1$, $\mu2$, $\mu3$ be the means of hardness of implants based on the methods used, respectively.

- Null Hypothesis H0: $\mu1 = \mu2 = \mu3$
- Alternate hypothesis Ha: atleast one of the means of hardness of implants is different from the rest.

The p-value is 0.004163412167505543
The p-value is less than 0.05.
**Therefore, we reject the null hypothesis that the means of hardness of implants based on the methods are equal for Alloy 1.**

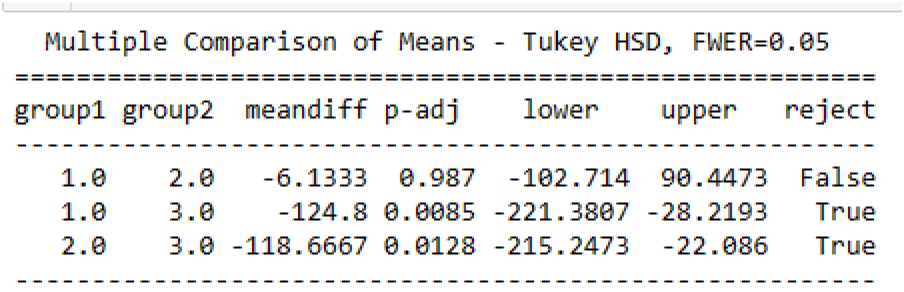
**The type of method used influences the hardness of implants for Alloy 1.**

We don't know which of the means is different from the rest or whether all pairs of means are different.

Multiple comparison tests (Tukey HSD) are used to test the differences between all pairs of means to identify which method the mean hardness of implants is different from other groups.

**Tukey HSD test:**

- Null hypothesis H0: $\mu1 = \mu2$ and $\mu1 = \mu3$ and $\mu2 = \mu3$
- Alternate hypothesis Ha: $\mu1 \neq \mu2$ or $\mu1 \neq \mu3$ or $\mu2 \neq \mu3$



```
  Multiple Comparison of Means - Tukey HSD, FWER=0.05
==========================================================
group1 group2  meandiff p-adj    lower    upper   reject
----------------------------------------------------------
   1.0    2.0   -6.1333  0.987  -102.714  90.4473  False
   1.0    3.0    -124.8 0.0085 -221.3807 -28.2193   True
   2.0    3.0 -118.6667 0.0128 -215.2473  -22.086   True
----------------------------------------------------------
```

As the p-values (refer to the p-adj column) for comparing the mean hardness of implants for pairs of methods 1 & 3 and 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 implants for method pair 1 & 2 is similar but that for method 3 is significantly different from methods 1 and 2.**

## Alloy 2

### Levene's test

- H0: All the population variances are equal.
- Ha: At least one variance is different from the rest.

The p-value is 0.04469269939158668
Since the p-value is less than 0.05, we reject the null hypothesis of homogeneity of variances.

### One-way Anova test:

The p-value is 5.415871051443187 x $10^{-6}$

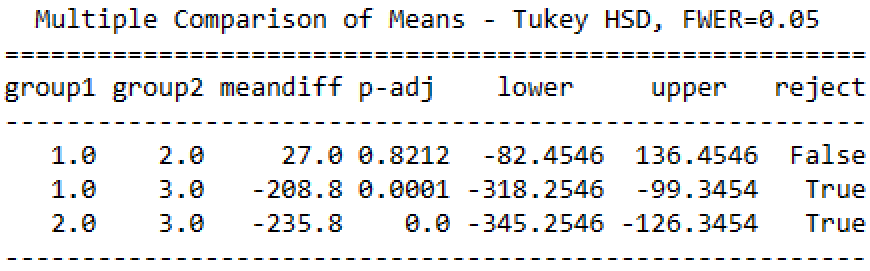
The p-value is less than 0.05.
**Therefore, we reject the null hypothesis that the means of hardness of implants based on the methods are equal for Alloy 2.**

**The type of method used influences the hardness of implants for Alloy 2 as well.**

We don't know which of the means is different from the rest or whether all pairs of means are different. Multiple comparison tests (Tukey HSD) are used to test the differences between all pairs of means.

**Tukey HSD test:**

- Null hypothesis H0: μ1 = μ2 and μ1 = μ3 and μ2 = μ3
- Alternate hypothesis Ha: μ1 ≠ μ2 or μ1 ≠ μ3 or μ2 ≠ μ3

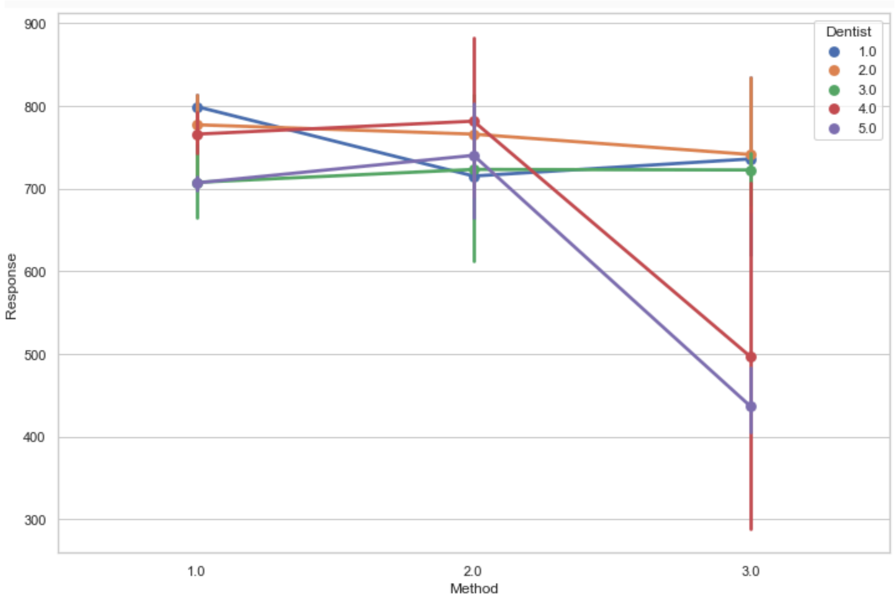


```
  Multiple Comparison of Means - Tukey HSD, FWER=0.05
========================================================
group1 group2 meandiff p-adj    lower     upper   reject
--------------------------------------------------------
   1.0    2.0     27.0 0.8212  -82.4546  136.4546  False
   1.0    3.0   -208.8 0.0001 -318.2546  -99.3454   True
   2.0    3.0   -235.8    0.0 -345.2546 -126.3454   True
--------------------------------------------------------
```

**Similar to the results we got for Alloy 1, we can say that the mean hardness of implants for method pair 1 & 2 is similar but that for method 3 is significantly different from methods 1 and 2.**

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

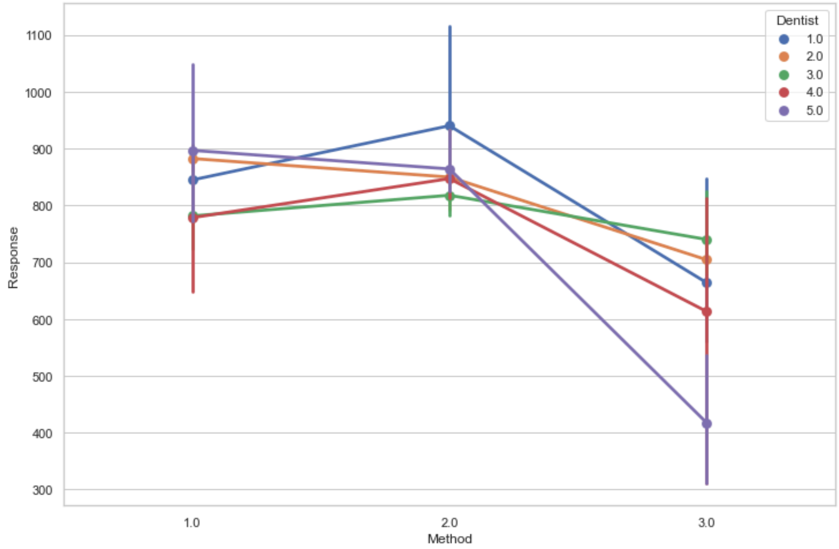
**Alloy 1**



- The error level for method 3 is the highest.
- For methods 1 and 2, the hardness lies between 700 and 800.
- For method 3, there is a huge variation.

- We know that the variation in dentists does not have a significant impact on the hardness of implants, but they moderate the value for different methods.

## Alloy 2

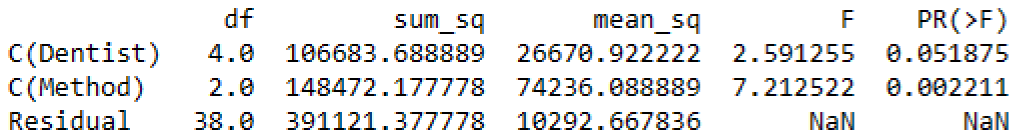


- The error level for method 3 is the highest.
- For methods 1 and 2, the hardness lies between 800 and 900.
- For method 3, there is a huge variation.
- We know that the variation in dentists does not have a significant impact on the hardness of implants, but they moderate the value for different methods.

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

## Alloy 1

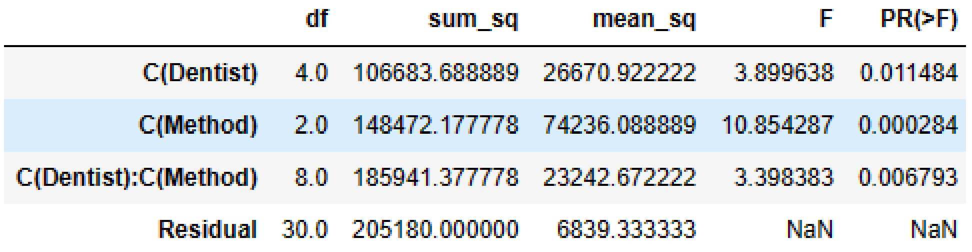
- Null Hypothesis H0: The mean hardness of implants for Alloy 1 is the same for all methods used by all dentists, i.e., there is no interaction effect on the hardness.
- Alternate hypothesis Ha: At least one of the means of hardness of implants is different from the rest. There is an interaction effect.



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

Previously we saw there is an interaction effect between the treatments. So, let's introduce a new term while performing the Two Way ANOVA to confirm the hypothesis statistically.

**Two-way Anova:**



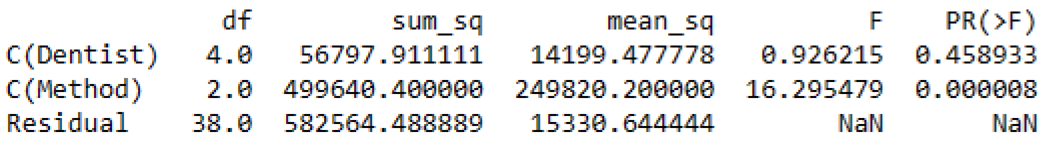
| | df | sum_sq | mean_sq | F | PR(>F) |
|---|---|---|---|---|---|
| **C(Dentist)** | 4.0 | 106683.688889 | 26670.922222 | 3.899638 | 0.011484 |
| **C(Method)** | 2.0 | 148472.177778 | 74236.088889 | 10.854287 | 0.000284 |
| **C(Dentist):C(Method)** | 8.0 | 185941.377778 | 23242.672222 | 3.398383 | 0.006793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

Due to the inclusion of the interaction effect term, we can see a significant change in the p-value of the first two treatments as compared to the Two-Way ANOVA.

**We see that the p-value (0.006 < 0.05) of the interaction effect term of 'Dentist' and 'Method' suggests that the null hypothesis is rejected in this case, i.e. for Alloy 1, confirming the existence of interaction effect on the hardness.**

**For Alloy 2:**

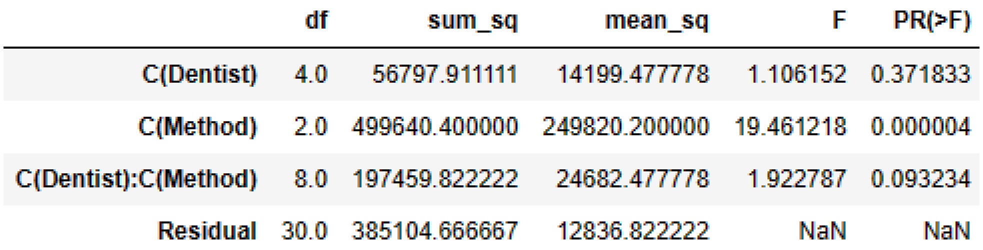
- Null Hypothesis H0: The mean hardness of implants for Alloy 2 is the same for all methods used by all dentists, i.e., there is no interaction effect on the hardness
- Alternate hypothesis Ha: At least one of the means of hardness of implants is different from the rest. There is an interaction effect.



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

Previously we saw there is an interaction effect between the treatments. So, let's introduce a new term while performing the Two Way ANOVA to confirm the hypothesis statistically.

**Two-way Anova test:**



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

By the inclusion of the interaction effect term, we can only see a slight change in the p-value of the first two treatments as compared to the Two-Way ANOVA.

**We see that the p-value (0.09 > 0.05) of the interaction effect term of 'Dentist' and 'Method' suggests that we fail to reject the null hypothesis in this case, i.e. for Alloy 2. The calibration effect is not significant for Alloy 2.**

.