Problem Statement - IS Project

Index

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**

Based on the above data, answer the following questions.

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

P(Injured)=145
P(players)=235
Prob_injured = (145/235) * 100

Probability of players injured. is 61.7%

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

P(Forward)= 94
P(Winger) = 29
P(Total_players)=235
prob_fwd = (Forward/Total_players,4)*100 =(94/235)*100 = 40%
prob_wng = (Winger/Total_players,4)*100 = (29/235) * 100 = 12.3%
total = prob_fwd + prob_wng = 40 + 12.3 = 52.3%
Probability of player being 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?

P(injured_striker) = 45
P(Total_players)=235
prob_strk_inj = injured_striker/Total_players,4)*100 = 45/235 * 100 = 19.1%
Probability of striker players injured. is 19.1%

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

P(injured_player) = 145
P(injured_striker) = 45
P(prob_strk_inj)= injured_striker/injured_player,4)*100 =
45/145 * 100 = 31%
Probability of striker players injured is 31%

Problem 2

The breaking strength of gunny bags used for packaging cement is normally distributed with a mean of 5 kg per sq. centimeter and a standard deviation of 1.5 kg per sq. centimeter. The quality team of the cement company wants to know the following about the packaging material to better understand wastage or pilferage within the supply chain; Answer the questions below based on the given information; **(Provide an appropriate visual representation of your answers, without which marks will be deducted)**

Let μ be the mean breaking strength of gunny bags .

The manager will test the null hypothesis

>$H_0: \mu = 5$

against the alternate hypothesis

> $H_a: \mu > 5$

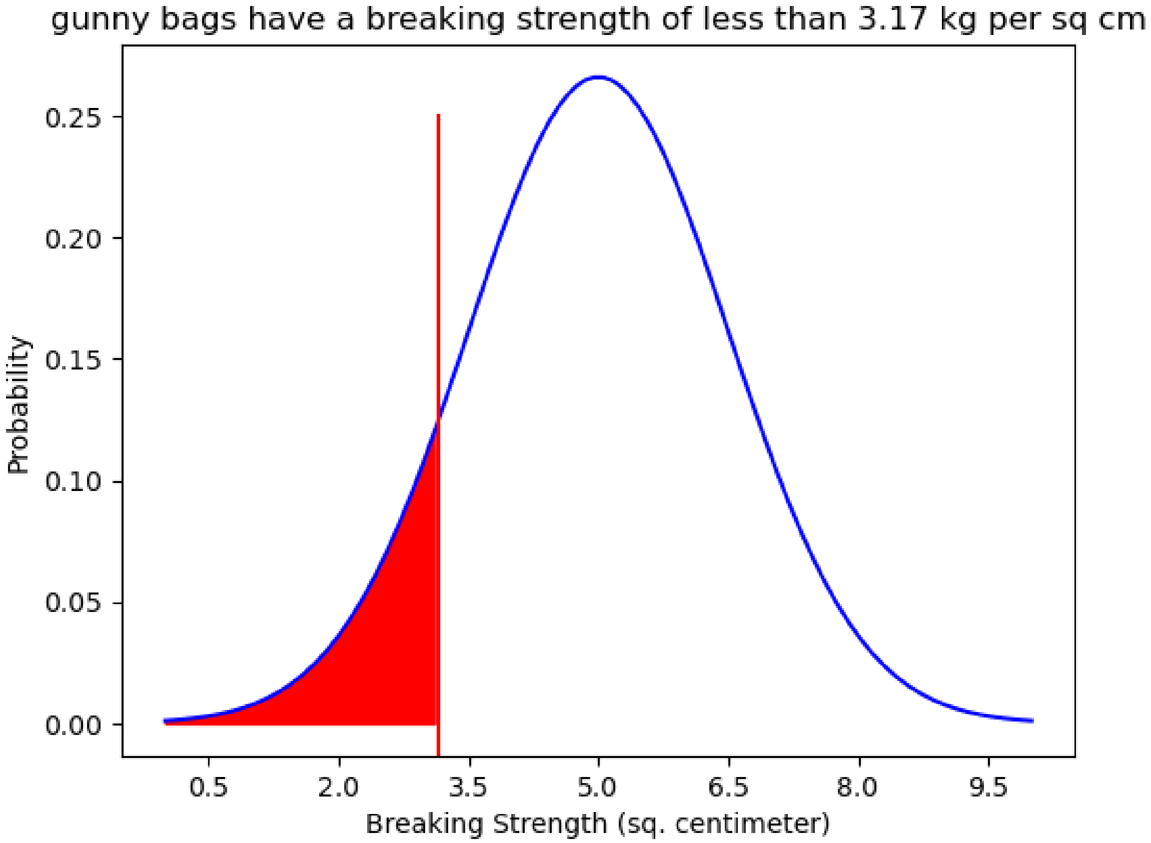
* Samples are drawn from a normal distribution - Since the sample size is 45(which is > 30), Central Limit Theorem states that the distribution of sample means will be normal. If the sample size was less than 30, we would have been able to apply z test on if we knew that the population distribution was normal.

* Observations are from a simple random sample - we are informed that the manager collected a simple random sample

* Standard deviation is known - Yes

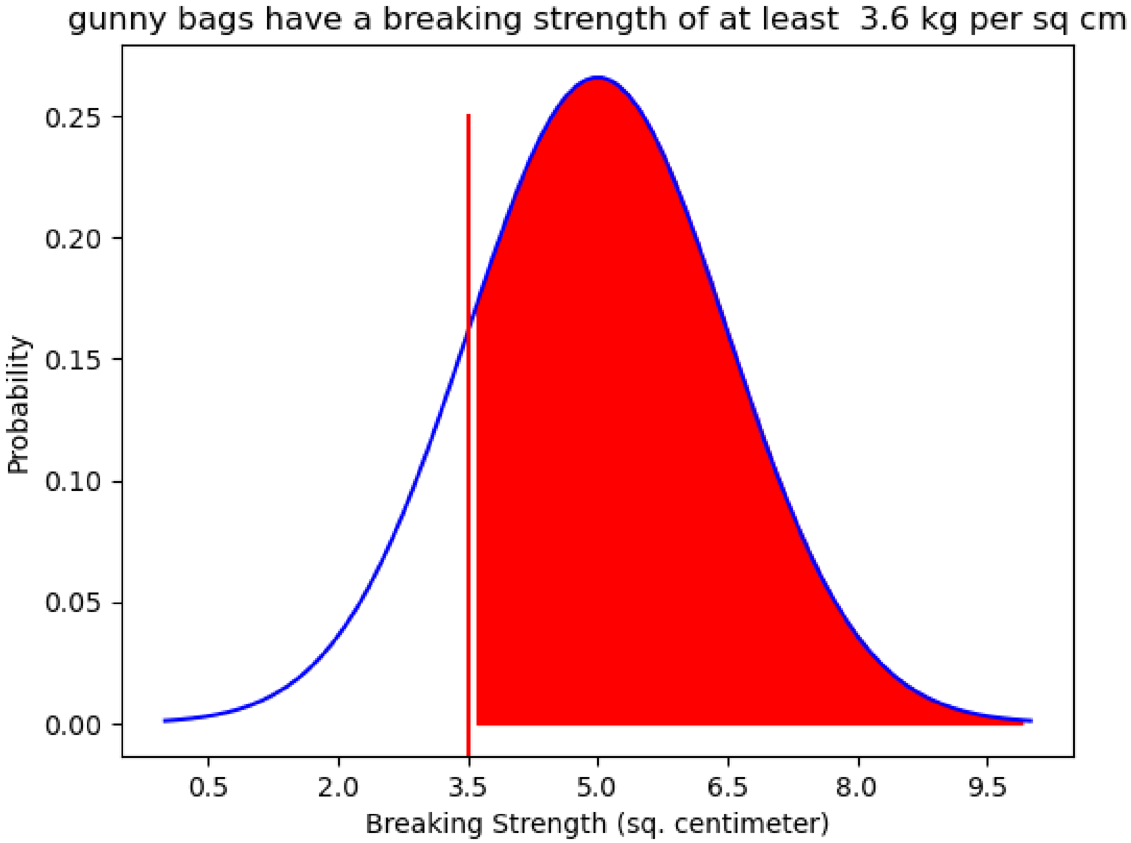
Voila! We can use Z-test for this problem.

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



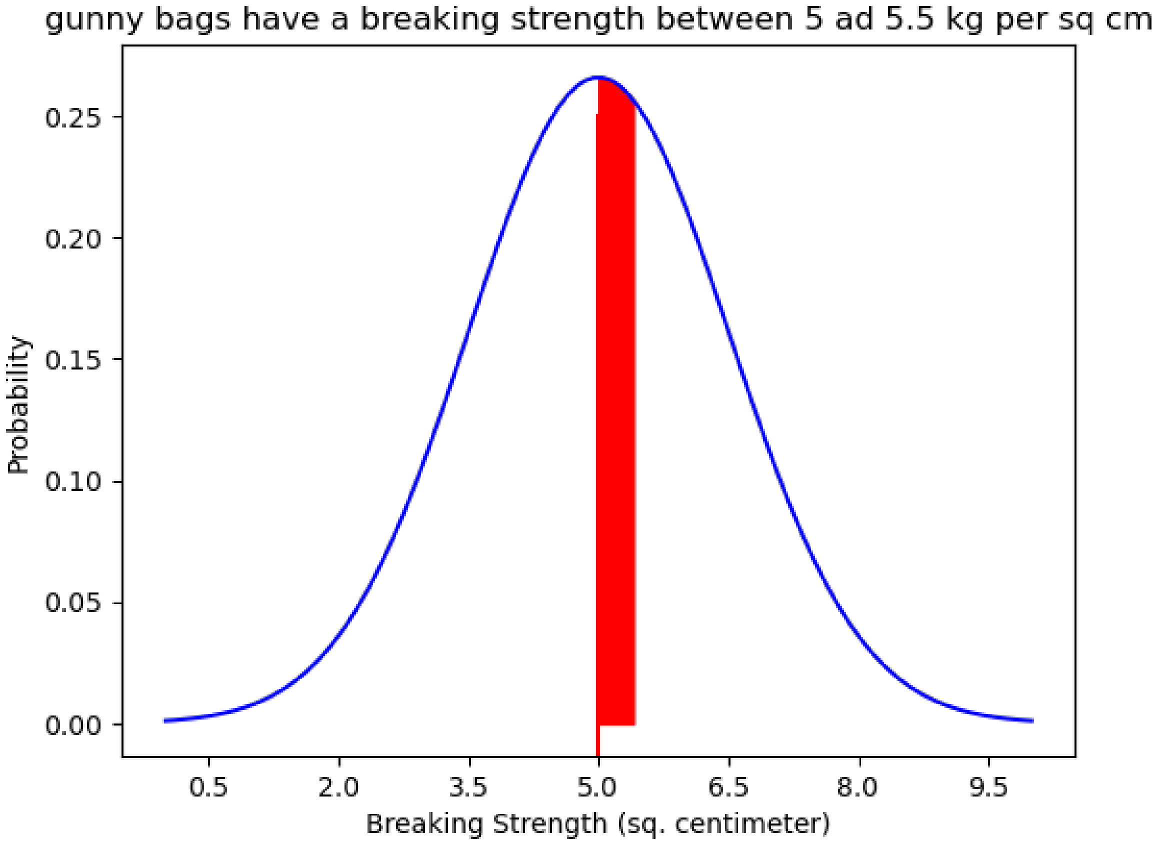
The proportion of the gunny bags have a breaking strength less than 3.17 kg per sq cm is : 0.11123243744783456

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



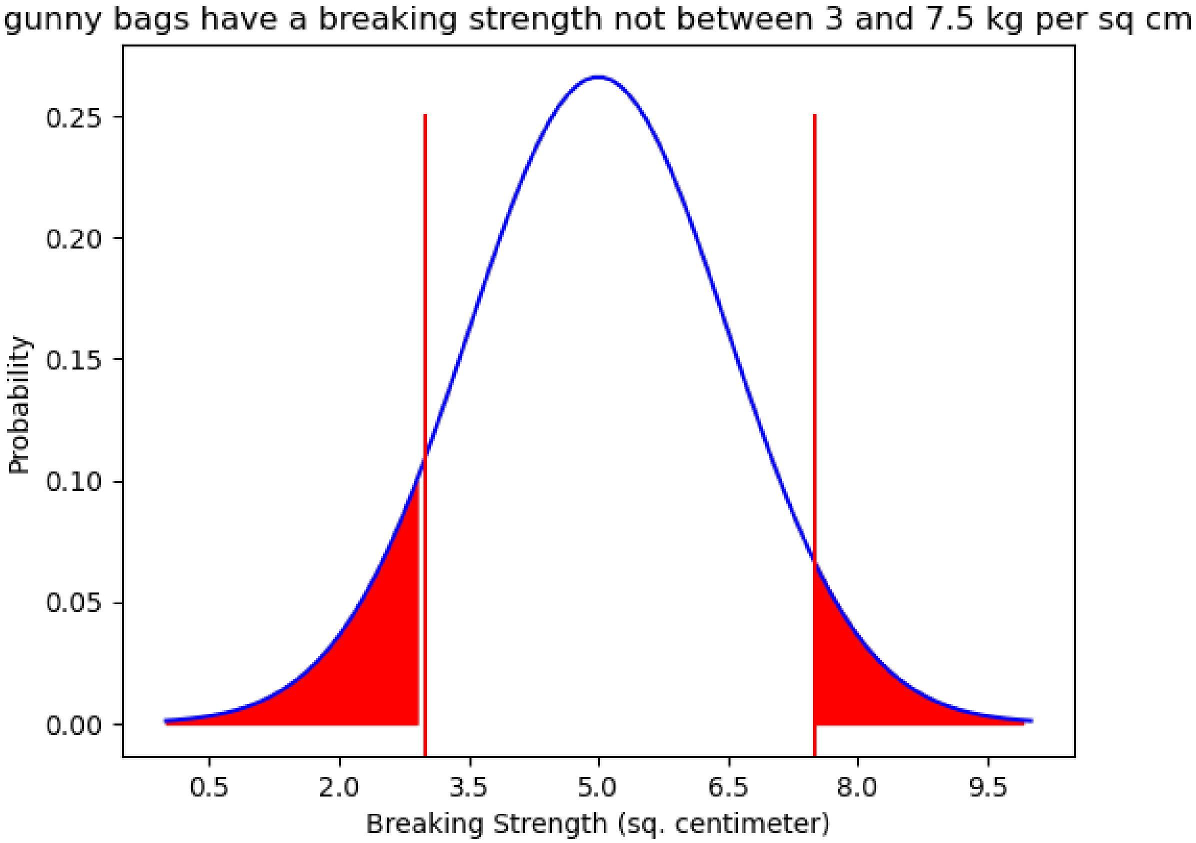
The proportion of the gunny bags have a breaking strength at least 3.6 kg per sq cm is: 0.8246760551477705

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 have a breaking strength between 5 and 5.5 kg per sq cm is 0.13055865981823633

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

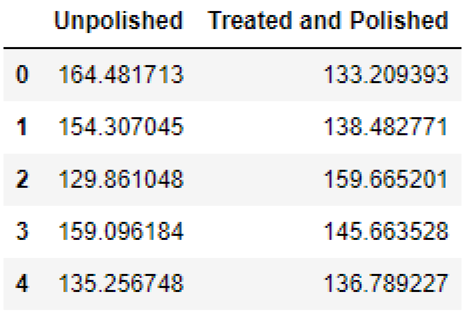


The proportion of the gunny bags have a breaking strength not between 3 and 7.5 kg per sq cm is 0.13900157199868257

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);

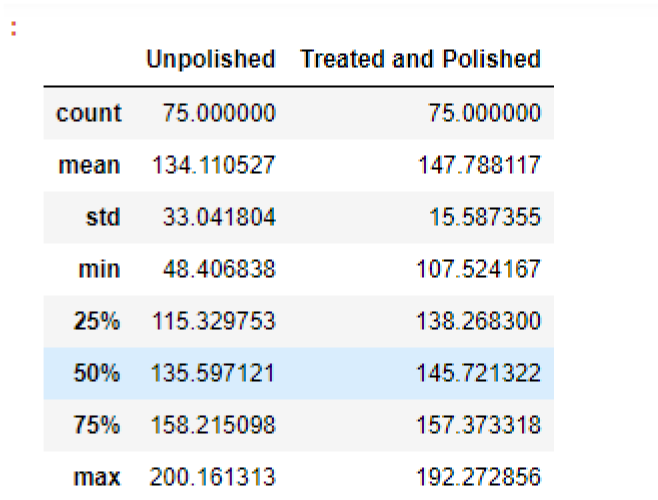
Top 5 Records of Zingaro Database:



	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

Dataframe has 75 records and 2 float columns so it is appropriate for solving our queries.

Dataframe has no Null values and negative values so it is appropriate for solving our hypothesis.



	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?

Step 1: Define null and alternative hypotheses

In testing the hardness index of stones.

Null hypothesis states that hardness index, μ is not less than 150.

Alternative hypothesis states that the mean hardness index, μ is unequal to 150.

* H_0: $\mu_{Unpolished}$ $\geq$ 150

* H_A: $\mu_{Unpolished}$ < 150

Here mu_{Unpolished} denotes hardness index of unpolished stones

Step 2: Decide the significance level

Here we select α= 0.05.

Step 3: Identify the test statistic

We do not know the population standard deviation and n = 30. So we use the t distribution and the t_{STAT} test statistic.

Step 4: Calculate the p - value and test statistic

scipy.stats.ttest_1samp calculates the t test for the mean of one sample given the sample observations and the expected value in the null hypothesis. This function returns t statistic and the two-tailed p value.

Step 5: Decide to reject or accept null hypothesis

Level of significance: 0.05

We have evidence to reject the null hypothesis since p value < Level of significance

Our one-sample t-test p-value= 8.342573994839304e-05

Zingaro is right to consider Unpolished stones are not right for printing.

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

Step 1: Define null and alternative hypotheses

In testing the hardness index of stones.

Null hypothesis states that hardness index, μ is not less than 150.

Alternative hypothesis states that the mean hardness index, μ is unequal to 150.

* H_0: $\mu_{Unpolished}$ = $\mu_{Polished}$

* H_A: $\mu_{Unpolished}$$\neq$ $\mu_{Polished}$

Here mu_{Unpolished} denotes hardness index of unpolished stones and mu_{Polished} denotes hardness index of polished stones.

Step 2: Decide the significance level

Here we select α= 0.05.

Step 3: Identify the test statistic

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

* Sample sizes for both samples are same.

* The sample is not a large sample, n < 30. So you use the t distribution and the t_{STAT} test statistic for two sample unpaired test.

Step 4: Calculate the p - value and test statistic

** We use the scipy.stats.ttest_ind to calculate the t-test for the means of TWO INDEPENDENT samples of scores given the two sample observations. This function returns t statistic and two-tailed p value.**

** This is a two-sided test for the null hypothesis that 2 independent samples have identical average (expected) values. This test assumes that the populations have identical variances.**

For this exercise, we are going to first assume that the variance is equal and then compute the necessary statistical values.

tstat -3.242

p-value for two-tail: 0.001588379295584306

Step 5: Decide to reject or accept null hypothesis

Level of significance: 0.05

We have evidence to reject the null hypothesis since p value < Level of significance

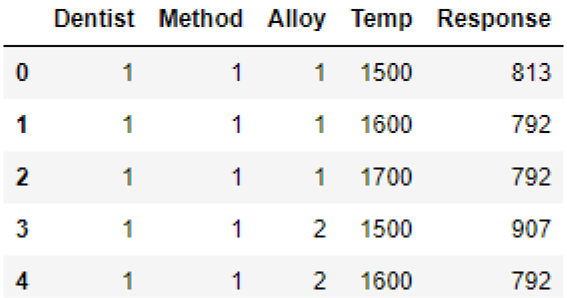
Our one-sample t-test p-value= 0.001588379295584306

As per the T-Test, We found the P value it is less than Level of Significance. Hence, We can reject Null Hypothesis. And conclude that Mean Hardness of “Unpolished Stones” and “Polished Stones” are not same

Problem 4

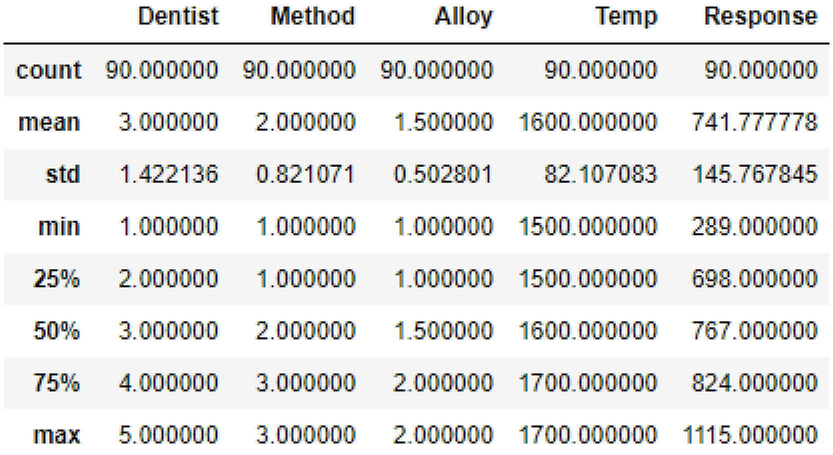
Dental implant data: The hardness of metal implants in dental cavities depends on multiple factors, such as the method of implant, the temperature at which the metal is treated, the alloy used as well as the dentists who may favor one method above another and may work better in his/her favorite method. The response is the variable of interest.

Top 5 records of Dental Database:



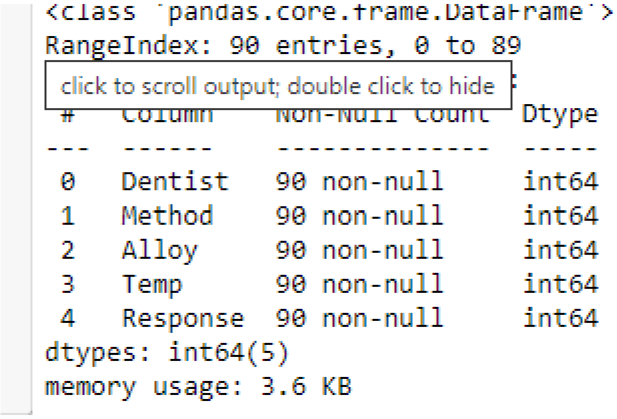
	Dentist	Method	Alloy	Temp	Response
0	1	1	1	1500	813
1	1	1	1	1600	792
2	1	1	1	1700	792
3	1	1	2	1500	907
4	1	1	2	1600	792

Statistical summary of variable looks perfect for solving our problems. Minimum, maximum and mean look normally distributed.



	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

There are no null values in database



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 90 entries, 0 to 89
click to scroll output; double click to hide
 #   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)
memory usage: 3.6 KB
```

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

There are 5 categories of Dentists.

Let's write the null and alternative hypothesis

Let $\mu_1, \mu_2, \mu_3,\mu_4,\mu_5$ be the means of Hardness of Implants for Dentists 1,2,3,4,5 respectively.

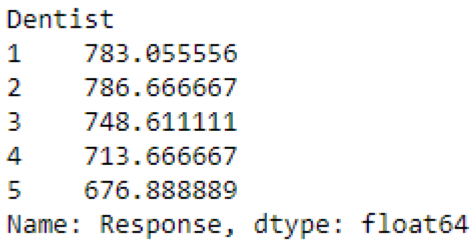
We will test the null hypothesis

>$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 =\mu_5$

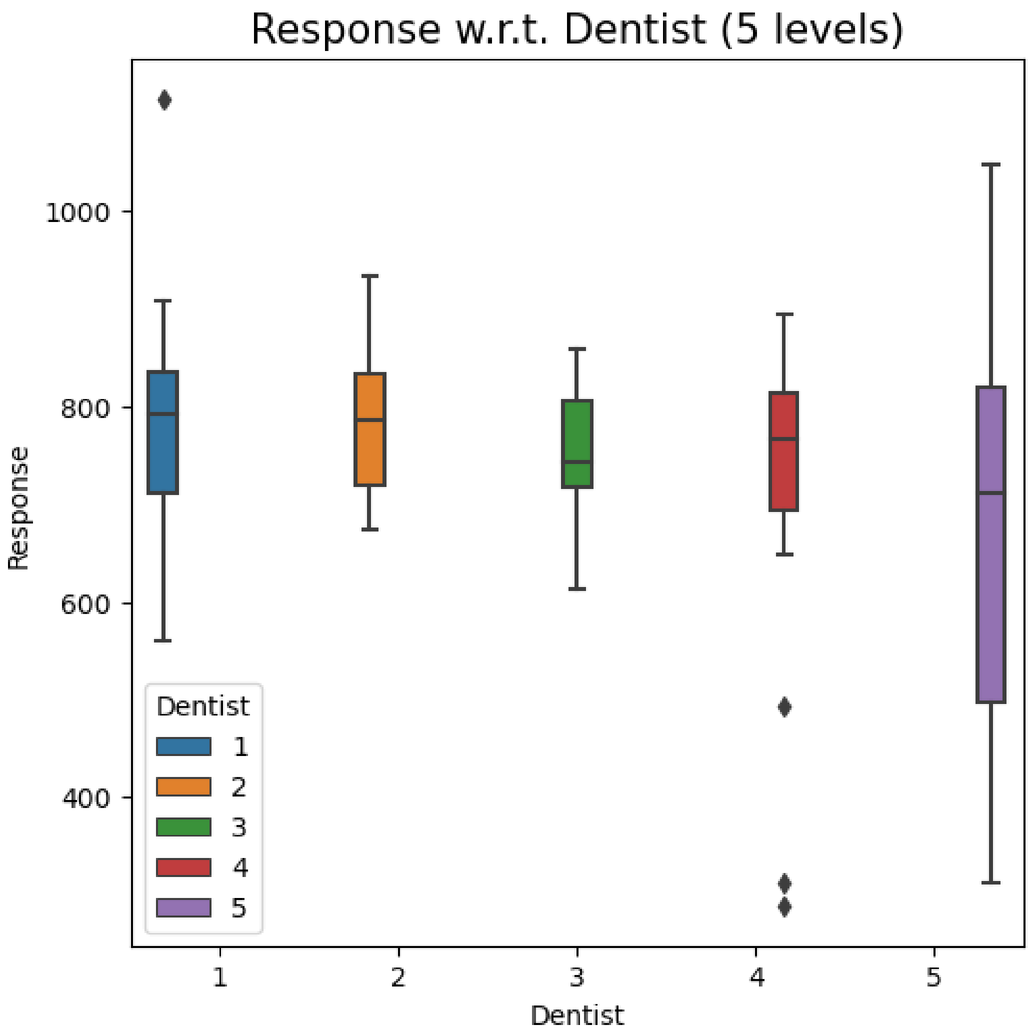
against the alternative hypothesis

>$H_a: $ At least one Hardness of Implants level is different from the rest.

Mean Response(Hardness of Implants) by Dentist:



```
Dentist
1    783.055556
2    786.666667
3    748.611111
4    713.666667
5    676.888889
Name: Response, dtype: float64
```



Shapiro-Wilk's test

We will test the null hypothesis

>$H_0:$ Dentist follows a normal distribution

against the alternative hypothesis

>$H_a:$ Dentist does not follow a normal distribution

Level of significance: 0.05

p-value= 1.1794428473876906e-06

We have evidence to reject the null hypothesis since p value < Level of significance

Since p-value of the test is small, we reject the null hypothesis that the Dentist follows the normal distribution.

Levene’s test

We will test the null hypothesis

>H_0: All the population variances are equal

against the alternative hypothesis

>H_a: At least one variance is different from the rest

The [`levene()’](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html) function of Scipy will be used to compute the test statistic and p-value.

Level of significance: 0.05

p-value= 0.007858817382355401

We have evidence to reject the null hypothesis since p value < Level of significance

One-Way tail-test

Level of significance: 0.05

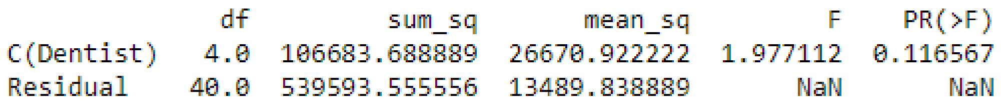
p-value= 0.11206595023098852

We have no evidence to reject the null hypothesis since p value > Level of significance

Insight

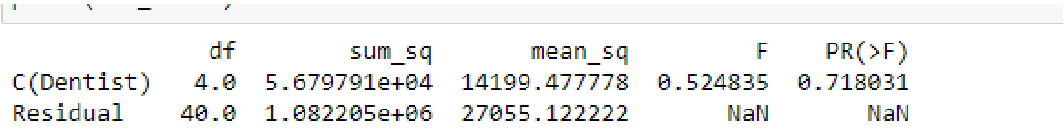
As the p-value is much less than the significance level, we fail to reject the null hypothesis. Hence, we have have enough statistical significance to conclude that all Dentist have same response at 5% significance level.

Alloy1:



```
                df         sum_sq       mean_sq         F    PR(>F)
C(Dentist)     4.0  106683.688889  26670.922222  1.977112  0.116567
Residual      40.0  539593.555556  13489.838889       NaN       NaN
```

Alloy 2:



```
                df        sum_sq       mean_sq         F    PR(>F)
C(Dentist)     4.0  5.679791e+04  14199.477778  0.524835  0.718031
Residual      40.0  1.082205e+06  27055.122222       NaN       NaN
```

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

There are 3 categories of Methods.

Let's write the null and alternative hypothesis

Let μ_1, μ_2, μ_3 be the means of Hardness of Implants for Methods 1,2,3 respectively.

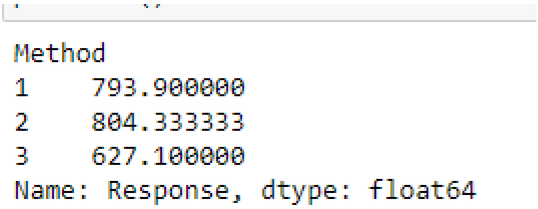
We will test the null hypothesis

>$H_0: \mu_1 = \mu_2 = \mu_3 $

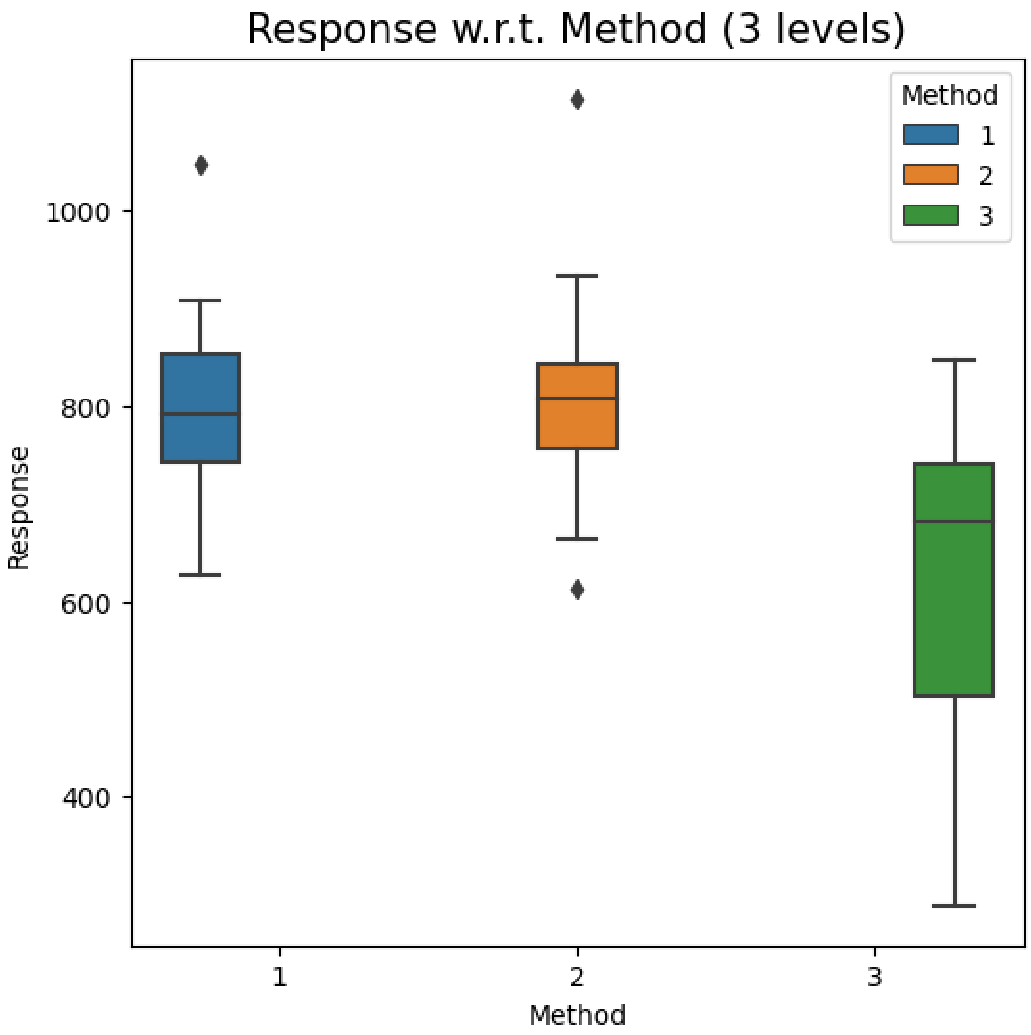
against the alternative hypothesis

>$H_a: $ At least one Hardness of Implants level is different from the rest.

Mean Response by Methods:



```
Method
1    793.900000
2    804.333333
3    627.100000
Name: Response, dtype: float64
```



Shapiro-Wilk’s test

We will test the null hypothesis

>$H_0:$ Method follows a normal distribution

against the alternative hypothesis

>$H_a:$ Method does not follow a normal distribution

Level of significance: 0.05

p-value= 6.475901481728386e-10

We have evidence to reject the null hypothesis since p value < Level of significance

Since p-value of the test is small, we reject the null hypothesis that the response follows the normal distribution.

Levene’s test

We will test the null hypothesis

>H_0: All the population variances are equal

against the alternative hypothesis

>H_a: At least one variance is different from the rest

The [`levene()`](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html) function of Scipy will be used to compute the test statistic and p-value.

Level of significance: 0.05

p-value= 0.004138452940152019

We have evidence to reject the null hypothesis since p value < Level of significance

Since the p-value is small, we reject the null hypothesis of homogeneity of variances.

Let's test whether the assumptions are satisfied or not

* The populations are normally distributed - No, the normality assumption can not be verified using the Shapiro-Wilk’s test.

* Samples are independent simple random samples - Yes, we are informed that the collected sample is a simple random sample.

* Population variances are equal - No, the homogeneity of variance assumption can not be verified using the Levene's test.

One-way test Results:

Level of significance: 0.05

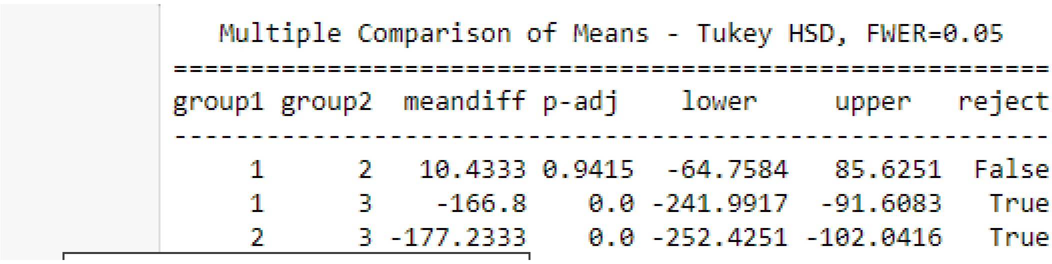
p-value= 7.683891892977992e-08

We have evidence to reject the null hypothesis since p value < Level of significance

Insight

As the p-value is much less than the significance level, we can reject the null hypothesis. Hence, we do have enough statistical significance to conclude that at least one Method is different from the rest at 5% significance level.

However, we don't know which mean is different from the rest or whether all pairs of means are different. Multiple comparison tests are used to test the differences between all pairs of means.



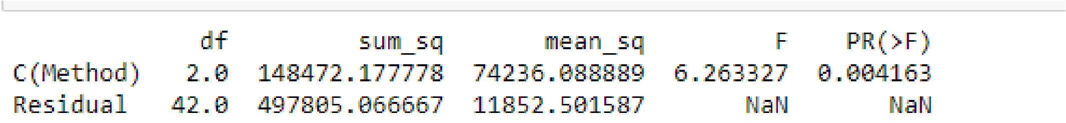
```
   Multiple Comparison of Means - Tukey HSD, FWER=0.05
=========================================================
group1 group2  meandiff p-adj    lower     upper   reject
---------------------------------------------------------
     1      2   10.4333 0.9415  -64.7584   85.6251  False
     1      3    -166.8    0.0 -241.9917  -91.6083   True
     2      3 -177.2333    0.0 -252.4251 -102.0416   True
```

Insight

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

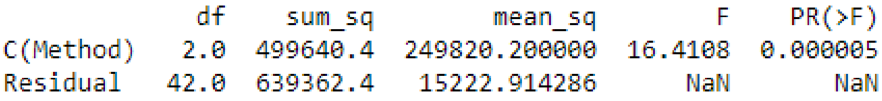
Thus, we can say that the mean Hardness of Implants for Methods 1 and 2 is similar but Hardness of Implants for Method 3 is significantly different from 1 and 2.

Alloy1:



```
             df         sum_sq        mean_sq         F    PR(>F)
C(Method)   2.0  148472.177778   74236.088889  6.263327  0.004163
Residual   42.0  497805.066667   11852.501587       NaN       NaN
```

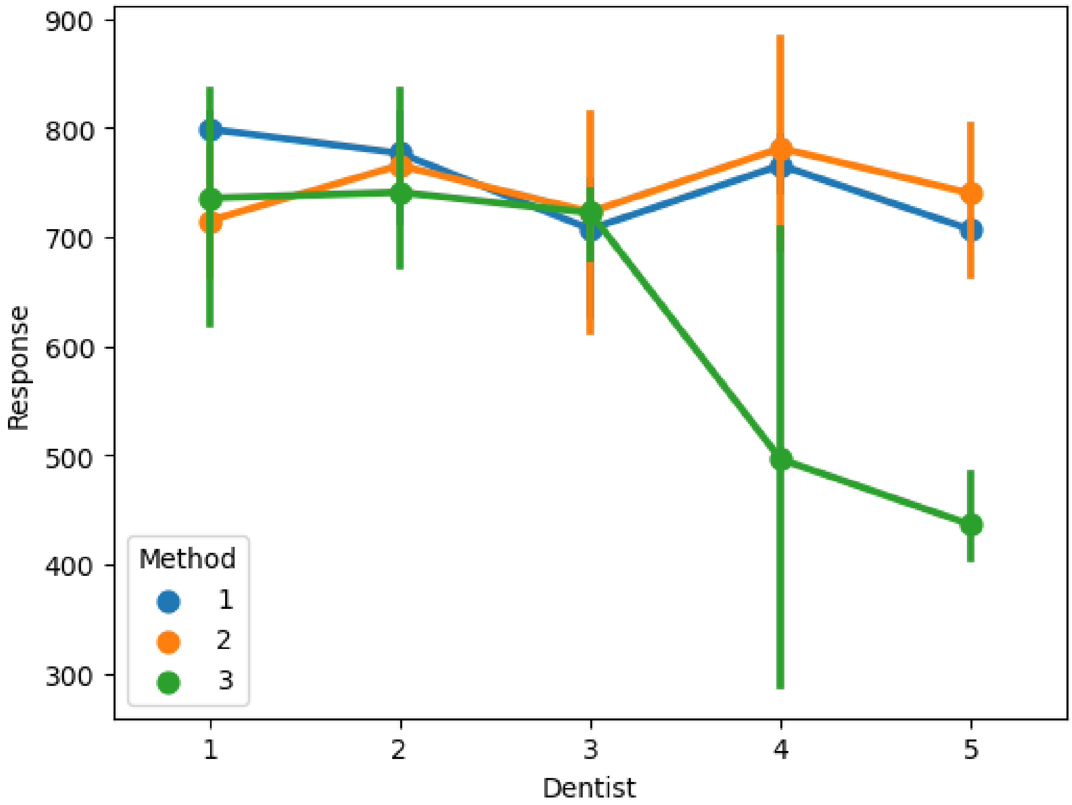
Alloy2:



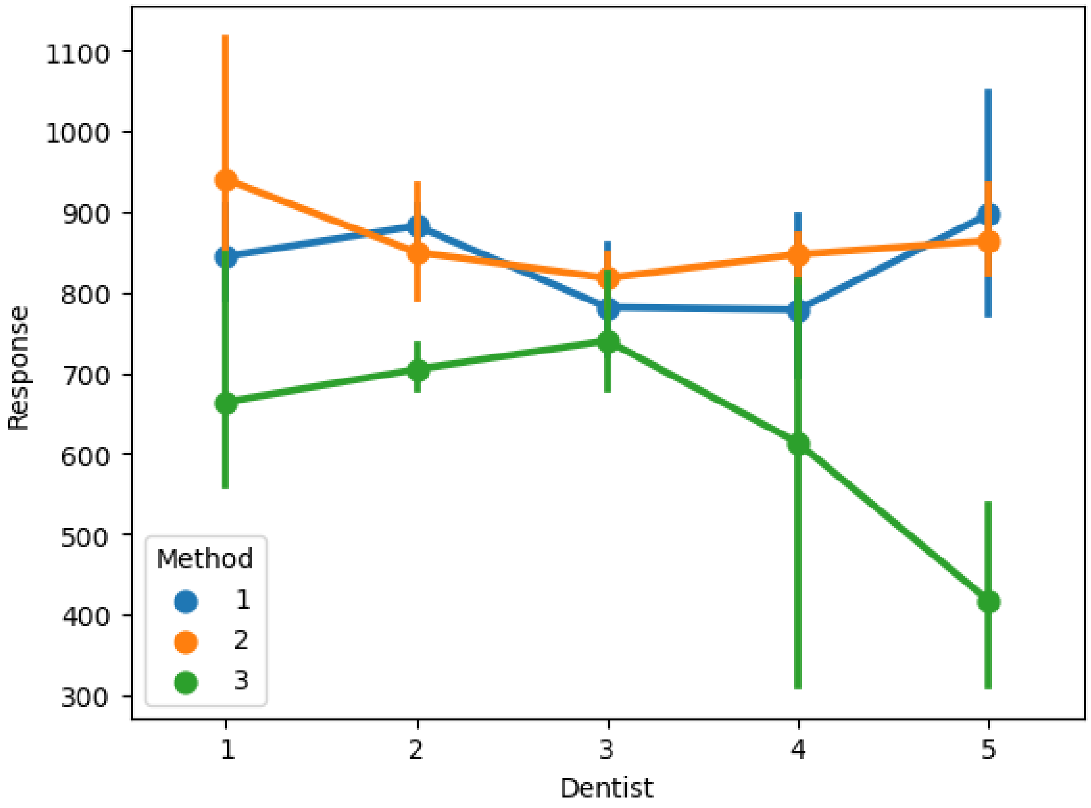
```
             df     sum_sq         mean_sq          F    PR(>F)
C(Method)   2.0   499640.4   249820.200000   16.4108  0.000005
Residual   42.0   639362.4    15222.914286       NaN       NaN
```

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

Alloy1:

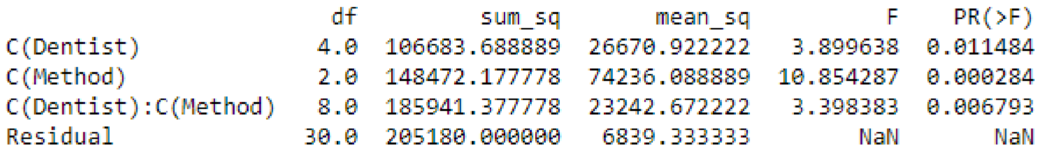

900
800
700
600
500
400
300
Response
1
2
3
4
5
Dentist
Method
● 1
● 2
● 3

Alloy2:

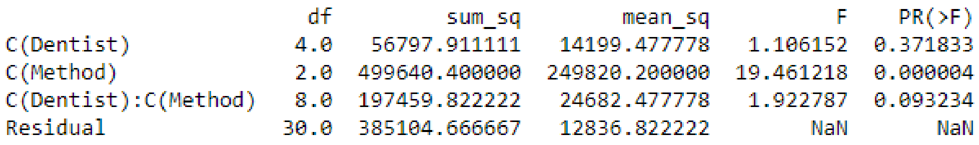


4.4 How does the hardness of implants vary depending on dentists and methods together?

Alloy1:


```
                          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
```

Alloy2:


```
                          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
```

My conclusions based on anova test are

1. **Dentists have effect on the hardness of implants**
2. **Methods have effect on hardness of implants.**
3. **we can say that the mean Hardness of Implants for Methods 1 and 2 is similar but Hardness of Implants for Method 3 is significantly different from 1 and 2.**