
ADVANCED STATISTICS PROJECT

BUSINESS REPORT

JAYA PREETHI R M

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**

The probability of an event to occur as an outcome of an experiment can be enumerated as:

P (A) = m / n

m= number of ways that are favourable for event A to occur ; n= the total possible outcomes from the experiment.

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

P = m / n = 145 / 235 = 0.617

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

P = m / n = 123 / 235 = 0.523

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

P = m / n = 45 / 145 = 0.31

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

P = m / n = 45 / 145 = 0.31

1.5 What is the probability that a randomly chosen injured player is either a forward or an attacking midfielder?

P = m / n = 80 / 145 = 0.552

PROBLEM 2

An independent research organization is trying to estimate the probability that an accident at a nuclear power plant will result in radiation leakage. The types of accidents possible at the plant are fire hazards, mechanical failure, or human error. The research organization also knows that two or more types of accidents cannot occur simultaneously.

According to the studies carried out by the organization, the probability of a radiation leak in case of a fire is 20%, the probability of a radiation leak in case of a mechanical 50%, and the probability of a radiation leak in case of a human error is 10%. The studies also showed the following;

- The probability of a radiation leak occurring simultaneously with a fire is 0.1%.
- The probability of a radiation leak occurring simultaneously with a mechanical failure is 0.15%.
- The probability of a radiation leak occurring simultaneously with a human error is 0.12%.

On the basis of the information available, answer the questions below:

2.1 What are the probabilities of a fire, a mechanical failure, and a human error respectively?

Given the probabilities of radiation leak due to fire, mechanical error and human error are 20%, 50% and 10%. Thus the total possible outcomes can be calculated as:

n= 20% + 50% + 10% = 80%

Thus, the probability of a fire is

P = m / n = 20% / 80% = 0.25

The probability of mechanical error is

P = m / n = 50% + 80% = 0.625

The probability of human error is

P = m / n = 10% / 80% = 0.125

2.2 What is the probability of a radiation leak?

Probability of a radiation leak is: 0.1% + 0.15% + 0.12% = 0.37%

2.3 Suppose there has been a radiation leak in the reactor for which the definite cause is not known. What is the probability that it has been caused by:

- A Fire.

P = 0.1% / 0.37% = 0.27%

- Mechanical Failure.

P = 0.15% / 0.37% = 0.40%

- A Human Error.

P = 0.12% / 0.37% = 0.32%

PROBLEM 3

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)**

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

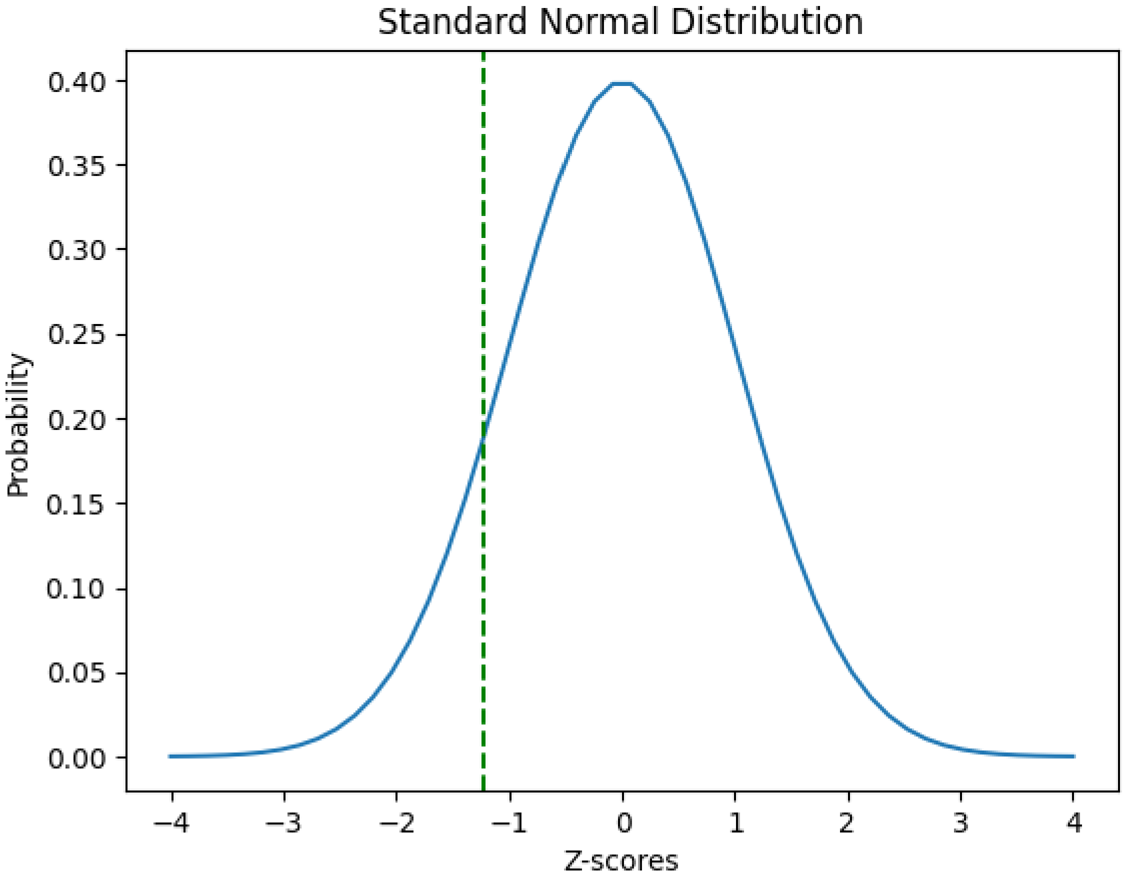
To find the proportion of gunny bags we must find the z-score of the data point 3.17 kg. Z-score is the measure of the total number of standard deviations from above or below the mean. The Z-score can be calculated as:

Z = (OBSERVED DATA POINT - MEAN) / STANDARD DEVIATION

Therefore, the Z-score for 3.17 kg is

Z = (3.17 - 5) / 1.5 = -1.22

From the Z-score, we can find out the probability of the observed data point lying in the given range. Thus, we can say that there is 11.12% chance of a gunny bag to have breaking strength less than 3.17 kg.



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

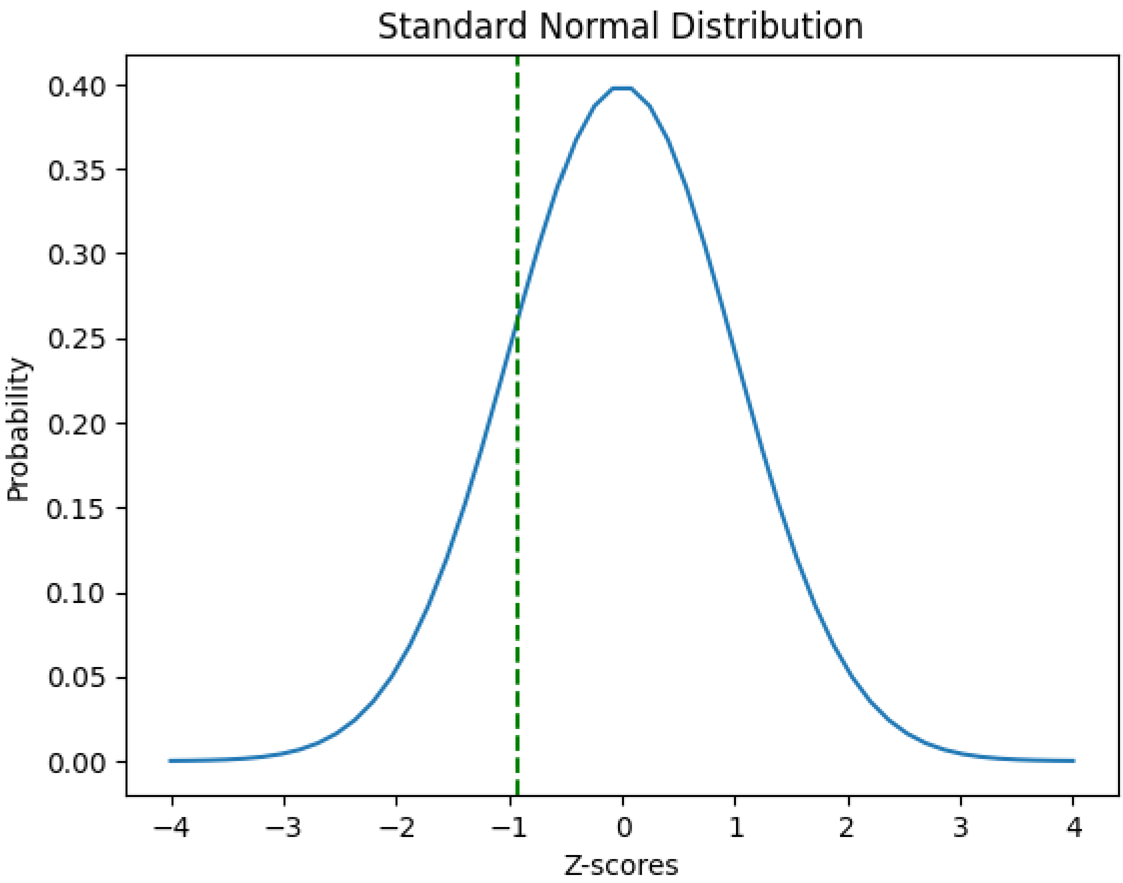
To find the proportion of gunny bags we must find the z-score of the data point 3.6 kg.

Z = (OBSERVED DATA POINT - MEAN) / STANDARD DEVIATION

Therefore, the Z-score for 3.17 kg is

Z = (3.6 - 5) /1.5 = -0.93

From the Z-score, we can say that there is a 17.61% chance of a gunny bag to have a breaking strength of at least 3.6 kg per sq cm.



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

To find the proportion of gunny bags we must find the z-score of the data points 5 and 5.5 kg per sq cm..

Z = (OBSERVED DATA POINT - MEAN) / STANDARD DEVIATION

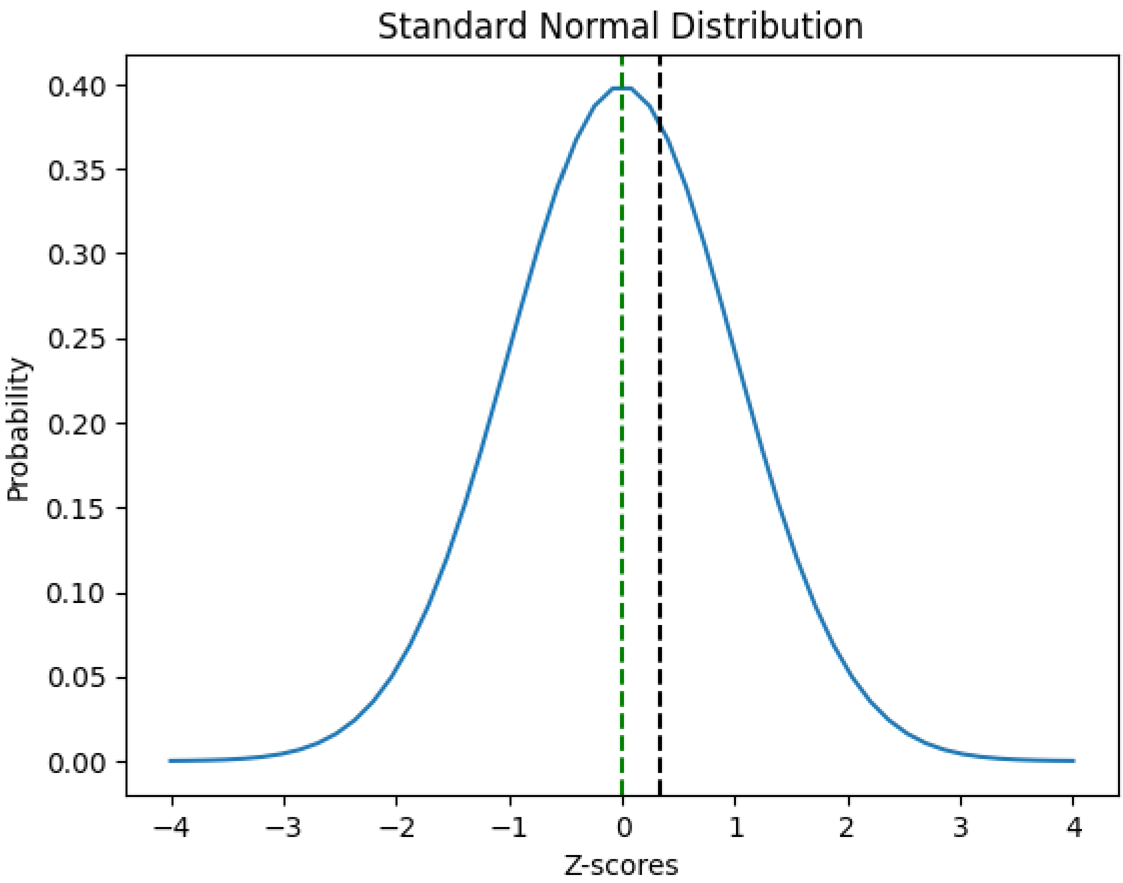
Therefore, the Z-score for 5 and 5.5 kg per sq cm. is

Z = (5 - 5)/1.5 = -0

Z1 = (5.5 - 5)/1.5 = 0.33

From the Z-score, we can say that there is a 13.05% chance of a gunny bag to have a breaking strength between 5 and 5.5 kg per sq cm.

cont.



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

To find the proportion of gunny bags we must find the z-score of the data points 3 and 7.5 kg per sq cm..

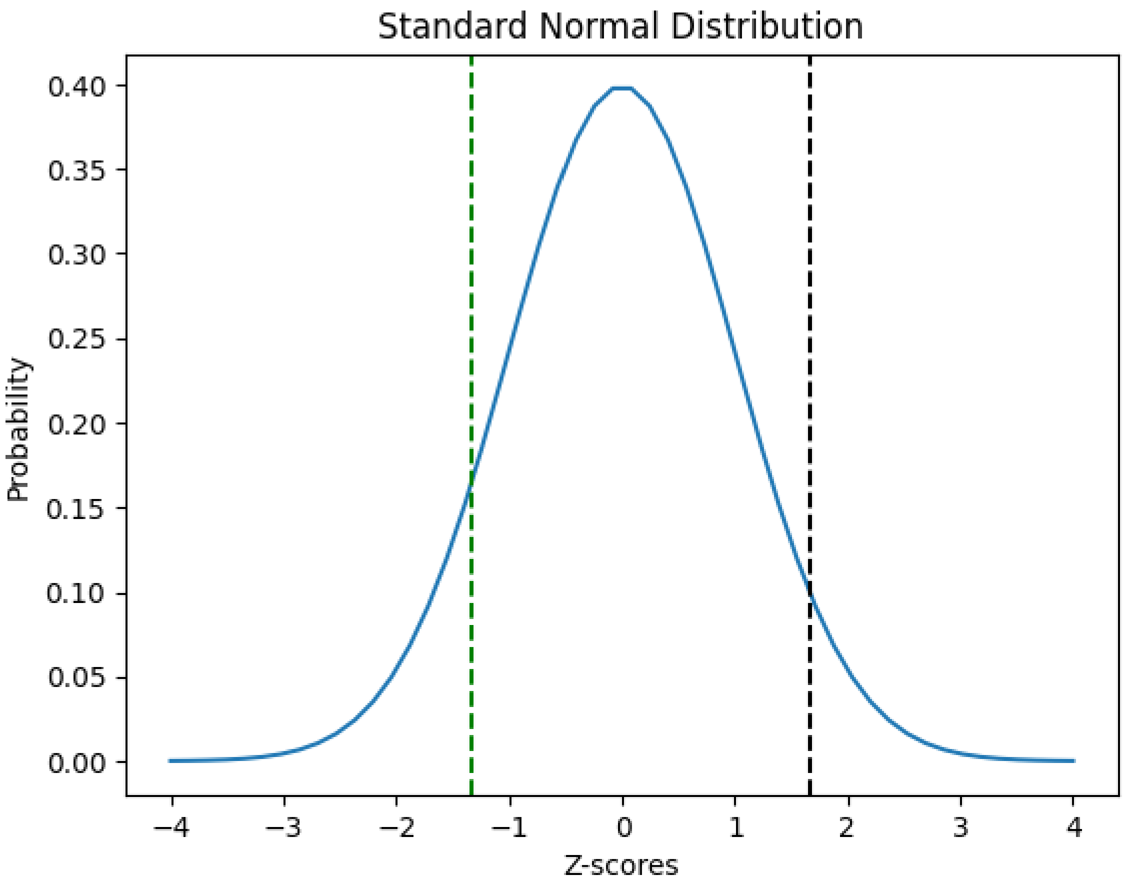
Z = (OBSERVED DATA POINT - MEAN) / STANDARD DEVIATION

Therefore, the Z-score for 3 and 7.5 kg per sq cm. is

Z = (3-5)/1.5 = -1.33

Z1 = (7.5-5)/1.5 = 1.66

From the Z-score, we can say that there is a 86.09% chance of a gunny bag to not have a breaking strength between 3 and 7.5 kg per sq cm.



PROBLEM 4:

Grades of the final examination in a training course are found to be normally distributed, with a mean of 77 and a standard deviation of 8.5. Based on the given information, answer the questions below.

4.1 What is the probability that a randomly chosen student gets a grade below 85 on this exam?

To find the probability of a randomly chosen student getting a grade below 85 on this exam, we must find the Z-score.

Z = (OBSERVED DATA POINT - MEAN) / STANDARD DEVIATION

Z = (85 - 77) / 8.5 = 0.94

With the Z-score, we can certainly say that there is a 54.37% probability of a randomly chosen student getting a grade below 85 on this exam.

4.2 What is the probability that a randomly selected student scores between 65 and 87?

To find the probability of a randomly selected student scores between 65 and 87 on this exam, we must find the Z-score.

Z = (OBSERVED DATA POINT - MEAN) / STANDARD DEVIATION

Z1 = (65 - 77) / 8.5 = -1.41

Z2 = (87 - 77) / 8.5 = 1.17

With the Z-score, we can certainly say that there is a 80.13% probability of a randomly selected student scores between 65 and 87 on this exam.

4.3 What should be the passing cut-off so that 75% of the students clear the exam?

To find the passing cut-off so that 75% of the students clear the exam, we should be able to find out the observed data point. Since, we already know the Z-score, mean, and standard deviation, with the Z-score formula we can find out the observed data point.

Z = (OBSERVED DATA POINT - MEAN) / STANDARD DEVIATION

0.75 = (OBSERVED DATA POINT - 8.5) / 8.5

OBSERVED DATA POINT = 83.37

Therefore, we can say that the passing cut-off should be 83.37 so that 75% of the students should be able to clear the exam.

PROBLEM 5:

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);

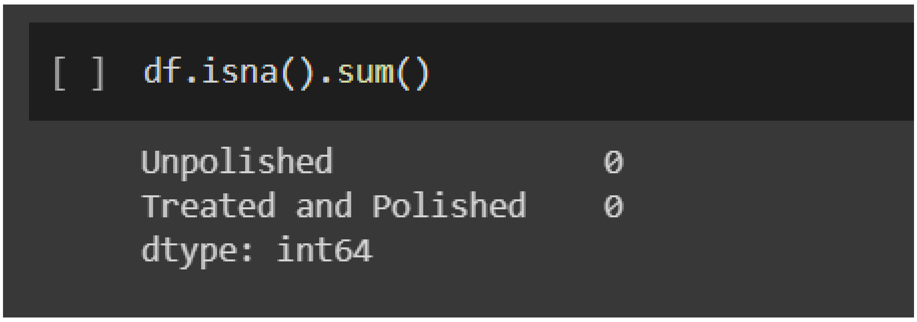
5.1 Earlier experience of Zingaro with this particular client is favorable as the stone surface was found to be of adequate hardness. However, Zingaro has reason to believe now that the unpolished stones may not be suitable for printing. Do you think Zingaro is justified in thinking so?

To find out whether Zingaro is justified in thinking that the unpolished stones may not be suitable for printing, we shall do the Independent T-test for hypothesis testing. Let us state the hypothesis:

Ho: the mean hardness of unpolished stones is more than or equal to 150.
Ha: the mean hardness of unpolished stones is less than 150.

Prior to testing the hypothesis, some preliminary analysis was done to the given dataset. Following are the results:

- ➔ To check if there are any null variables

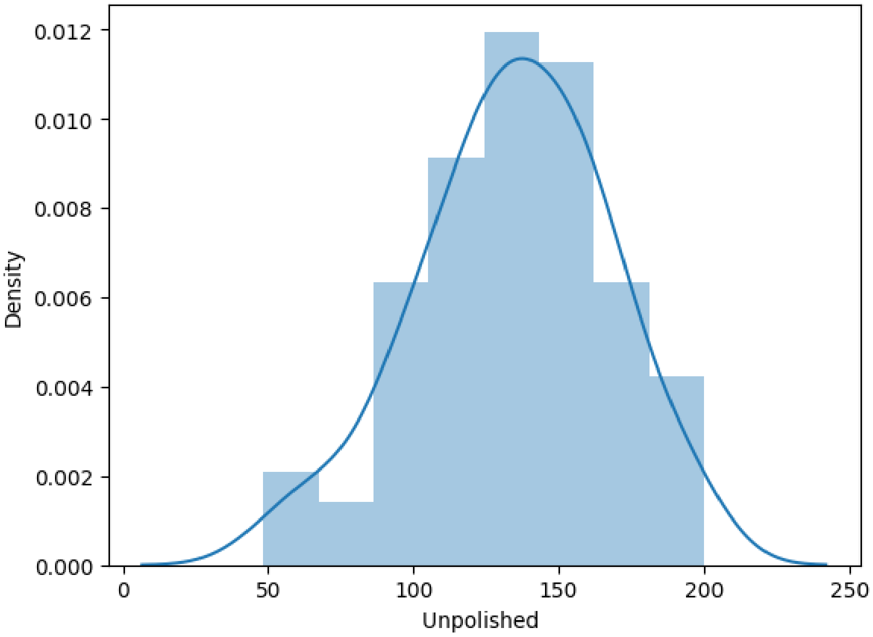


```
[ ]  df.isna().sum()

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

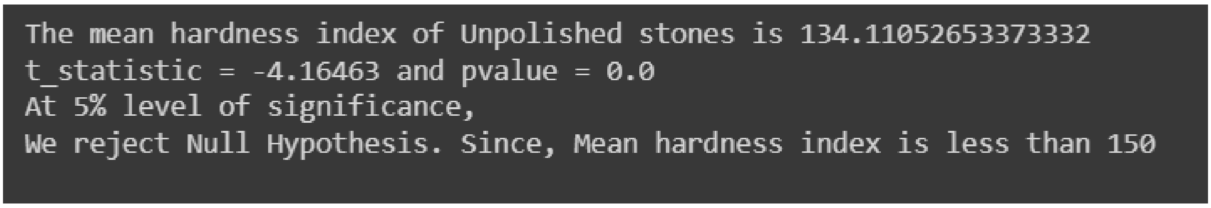
As stated in the above output there are no null variables.

- ➔ To check the distribution of the variables



The above given distribution plot is of the Unpolished variable. We can see that the unpolished variable is closer to a normal distribution, with distribution range within 0 to 0.012 density level.

Following the preliminary analysis, we got the following output from the hypothesis testing.



```
The mean hardness index of Unpolished stones is 134.11052653373332
t_statistic = -4.16463 and pvalue = 0.0
At 5% level of significance,
We reject Null Hypothesis. Since, Mean hardness index is less than 150
```

From the given output, we can learn that the mean hardness index of unpolished stones is 134.11, which is lesser than the desired mean index. To prove this point we shall also notice that the p-value is 0, which is lesser than 0.05. Therefore, we can conclude by saying that, at 5% level of significance, we failed to accept the null hypothesis. Thus, we can certainly say that the Zingaro co. was correct with their assumption.

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

To prove this statement, we shall state the hypothesis,

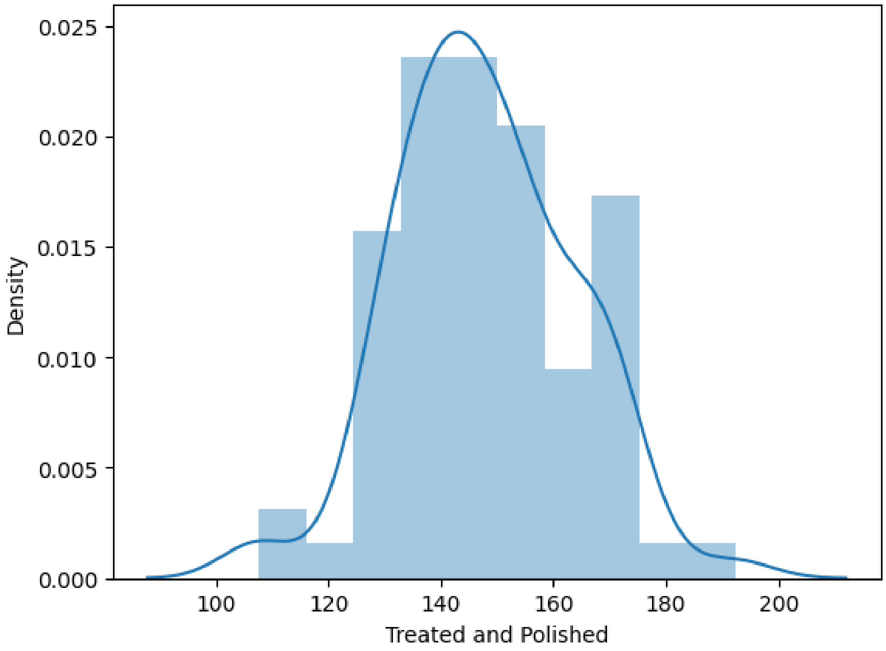
Ho: μ of treated and polished stones = μ of unpolished stones

Ha: μ of treated and polished stones ≠ μ of unpolished stones

Prior to testing the hypothesis, some preliminary analysis was done to the given dataset. Following are the results:

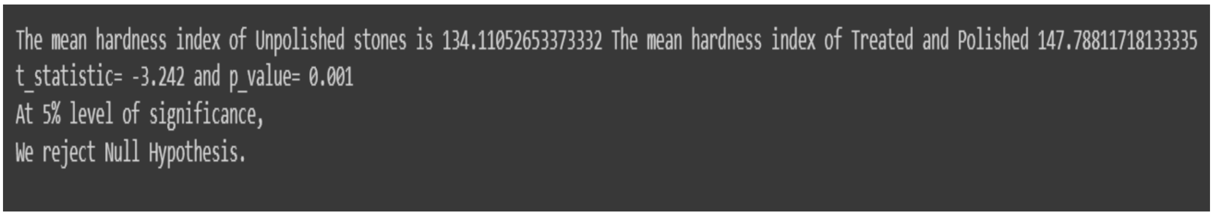
- ➔ Distribution plot for treated and polished stones

cont.



Unlike Unpolished stores, Treated and Unpolished stones are skewed towards the right side. The treated and polished stones vary between 0 to 0.025 hardness density index.

Following the preliminary analysis, we got the following output from the hypothesis testing.



```
The mean hardness index of Unpolished stones is 134.11052653373332 The mean hardness index of Treated and Polished 147.78811718133335
t_statistic= -3.242 and p_value= 0.001
At 5% level of significance,
We reject Null Hypothesis.
```

From the given output, we can learn that the mean hardness index of treated and polished stones is 147.78 and mean hardness index of unpolished stones is 134.11, which are not the same. To prove this point we shall also notice that the p-value is 0.001, which is lesser than 0.05. Therefore, we can conclude by saying that, at 5% level of significance, we failed to accept the null hypothesis. Thus, we can certainly say that µ of treated and polished stones is not equal to µ of unpolished stones.

PROBLEM 6:

Aquarius health club, one of the largest and most popular cross-fit gyms in the country, has been advertising a rigorous program for body conditioning. The program is considered successful if the candidate is able to do more than 5 push-ups, as compared to when he/she

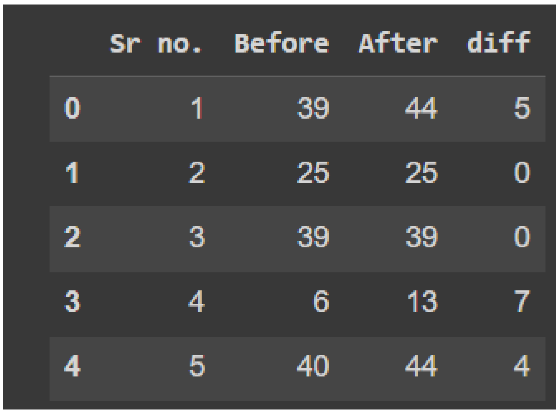
enrolled in the program. Using the sample data provided can you conclude whether the program is successful? (Consider the level of Significance as 5%)
Note that this is a problem of the paired-t-test. Since the claim is that the training will make a difference of more than 5, the null and alternative hypotheses must be formed accordingly.

To ascertain whether the program is successful, we must state the hypothesis.

Ho: µ of difference <= 5

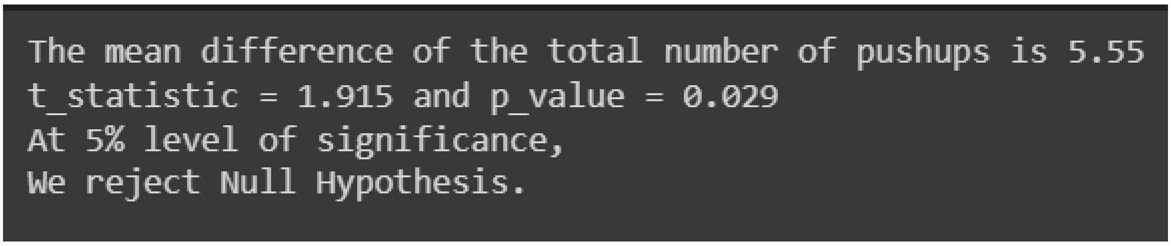
Ha: µ of difference > 5

Prior to calculating the mean difference, we must input a new column with the difference of push-ups done before and after the training.



	Sr no.	Before	After	diff
0	1	39	44	5
1	2	25	25	0
2	3	39	39	0
3	4	6	13	7
4	5	40	44	4

Above given table shows us the changes made to the dataset. Below given, is the hypothesis testing using paired t-test from python libraries.



```
The mean difference of the total number of pushups is 5.55
t_statistic = 1.915 and p_value = 0.029
At 5% level of significance,
We reject Null Hypothesis.
```

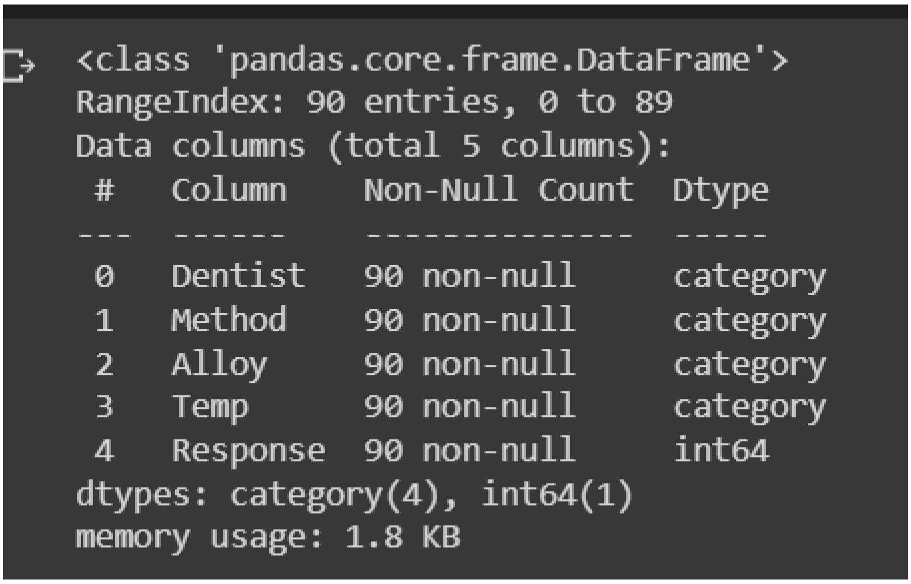
From the output, we shall notice that the mean difference is 5.55, which is higher than the expected mean difference of 5. To further prove this, we shall notice that p-value is 0.03, which is lesser than the significance level of 0.05. Therefore, we can certainly say that at a 5% level of significance, we fail to accept the null hypothesis and there is a difference in the push-up count from the training.

PROBLEM 7:

Dental implant data: The hardness of metal implant 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 on 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.

1)Test whether there is any difference among the dentists on the implant hardness. State the null and alternative hypotheses. Note that both types of alloys cannot be considered together. You must state the null and alternative hypotheses separately for the two types of alloys.?

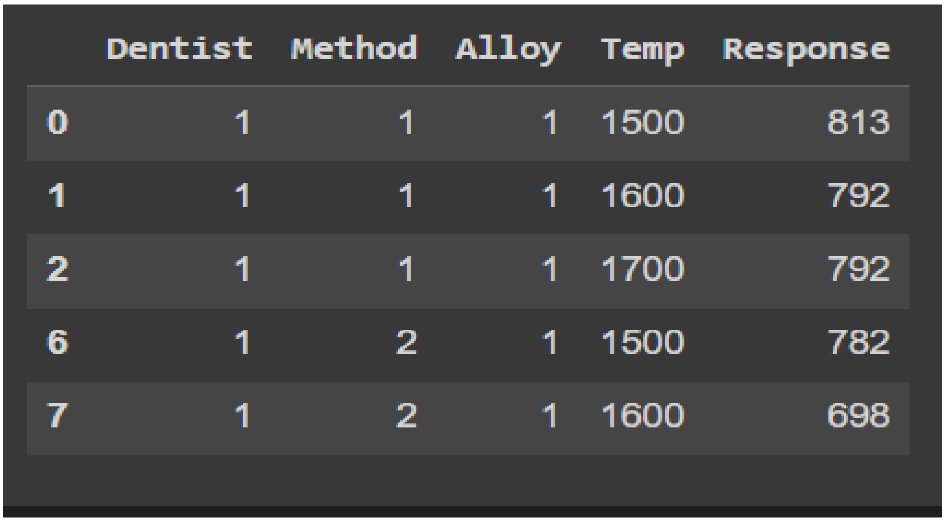
Prior to hypothesis testing I noticed that all the variables are integer type. Therefore I had converted all the independent variables to category type to perform ANOVA test, Since it is mandatory to use categorical tests.



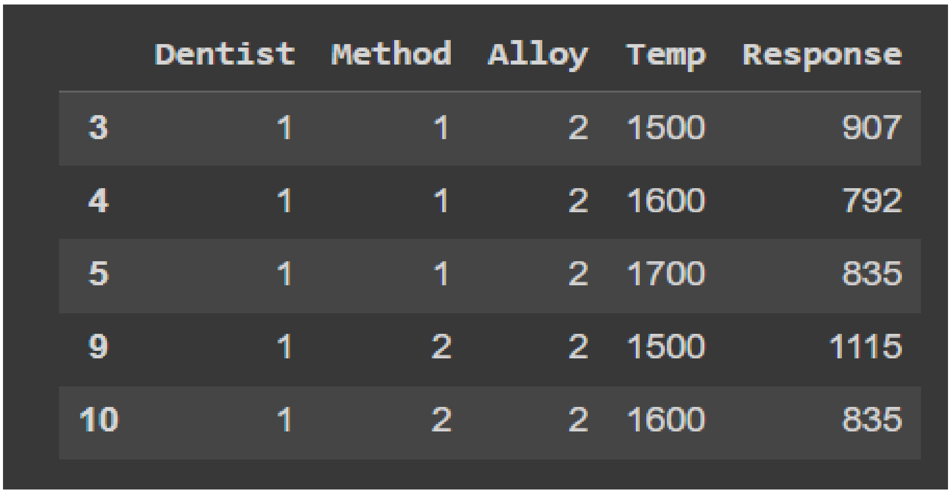
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 90 entries, 0 to 89
Data columns (total 5 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Dentist   90 non-null     category
 1   Method    90 non-null     category
 2   Alloy     90 non-null     category
 3   Temp      90 non-null     category
 4   Response  90 non-null     int64
dtypes: category(4), int64(1)
memory usage: 1.8 KB
```

I also subdivided the dataset into two subsets only consisting of either Alloy 1 or Alloy 2.

cont.



	Dentist	Method	Alloy	Temp	Response
0	1	1	1	1500	813
1	1	1	1	1600	792
2	1	1	1	1700	792
6	1	2	1	1500	782
7	1	2	1	1600	698



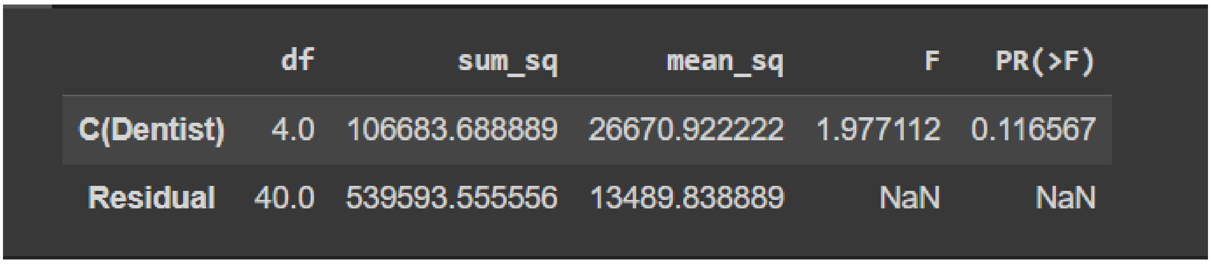
	Dentist	Method	Alloy	Temp	Response
3	1	1	2	1500	907
4	1	1	2	1600	792
5	1	1	2	1700	835
9	1	2	2	1500	1115
10	1	2	2	1600	835

ALLOY 1

The hypothesis for the one-way ANOVA test is as follow:

Ho: Mean impact hardness of Alloy 1 among all dentists are the same

Ha: Mean impact hardness of Alloy 1 is different for at least one dentist



	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

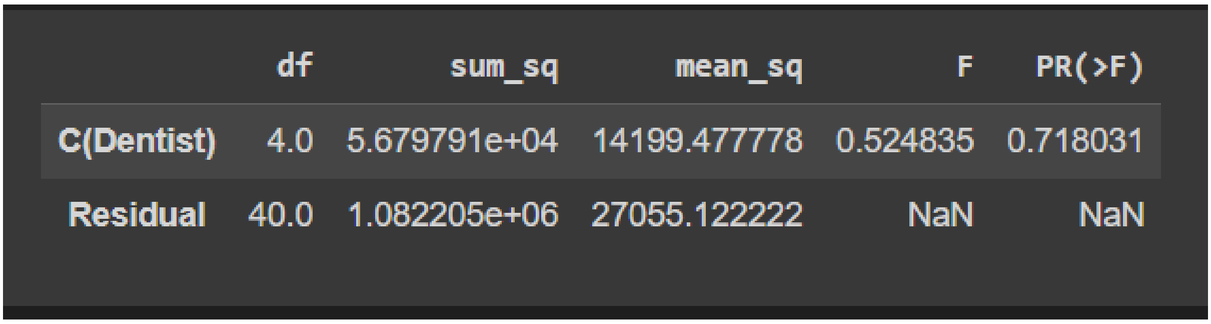
From the above given ANOVA table, we can see that the P-value is higher than the 0.05 level of significance. Therefore, we accept the null hypothesis at 5% level of significance. The mean impact hardness of Alloy 1 among all dentists are the same.

ALLOY 2

The hypothesis for the one-way ANOVA test is as follow:

Ho: Mean impact hardness of Alloy 2 among all dentists are the same

Ha: Mean impact hardness of Alloy 2 is different for at least one dentist



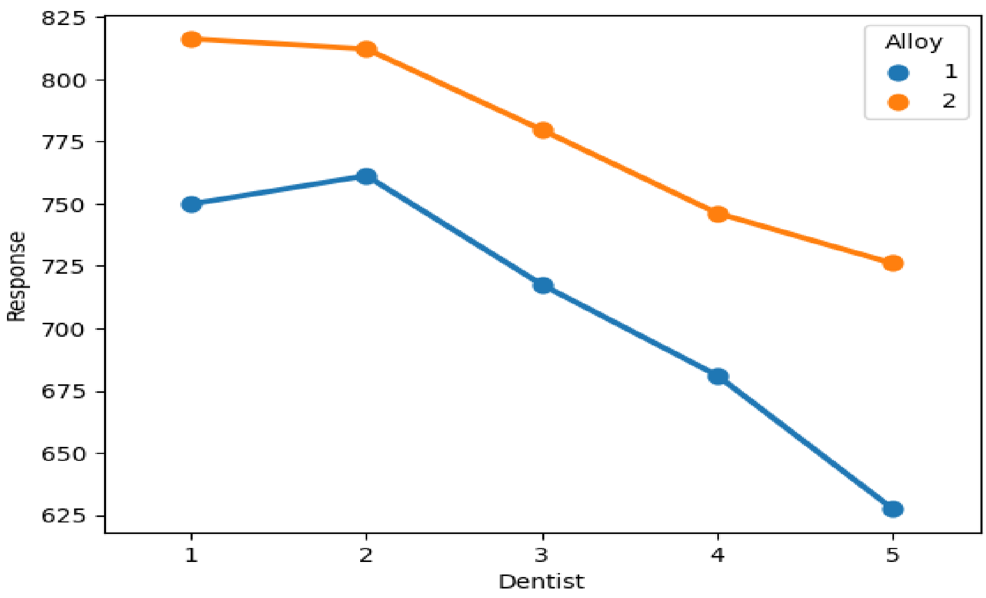
	df	sum_sq	mean_sq	F	PR(>F)
C(Dentist)	4.0	5.679791e+04	14199.477778	0.524835	0.718031
Residual	40.0	1.082205e+06	27055.122222	NaN	NaN

From the above given ANOVA table, we can see that the P-value is higher than the 0.05 level of significance. Therefore, we accept the null hypothesis at 5% level of significance. The mean impact hardness of Alloy 2 among all dentists are the same.

3)Irrespective of your conclusion in 2, we will continue with the testing procedure. What do you conclude regarding whether implant hardness depends on dentists? Clearly state your conclusion. If the null hypothesis is rejected, is it possible to identify which pairs of dentists differ?

To identify whether the implant hardness depends on the dentists let us look at a pointplot picturising the dataset.

cont.



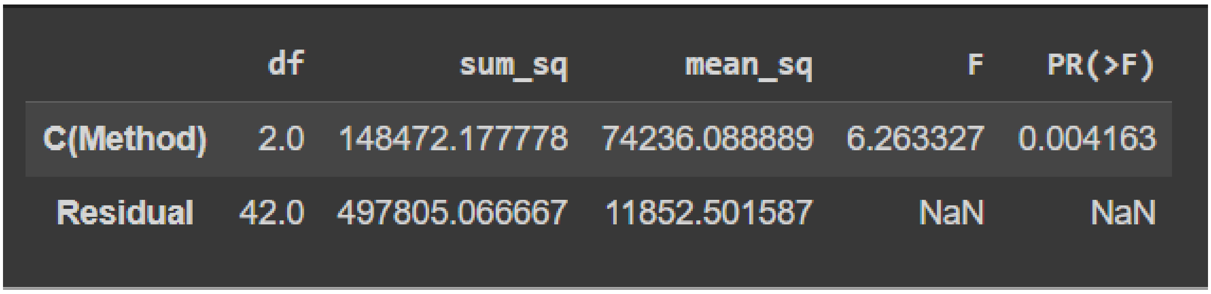
Since, the null hypothesis was accepted at 5% level of significance we may not be able to analyse the differences in the dentists in terms of their implant hardness. On the given pointplot, we can see that there are certain differences in dental hardness in between the dentists based on the alloy. Alloy 2 tends to have more hardness than Alloy 1 and Dentist 5 has the highest contrast in the hardness. Other than that, there are no noticeable differences between the dentists in terms of dental hardness in both alloys.

4)Now test whether there is any difference among the methods on the hardness of the dental implant, separately for the two types of alloys. What are your conclusions? If the null hypothesis is rejected, is it possible to identify which pairs of methods differ?

ALLOY 1

Ho: Mean impact hardness of Alloy 1 at different methods are the same

Ha: Mean impact hardness of Alloy 1 is different for at least one method



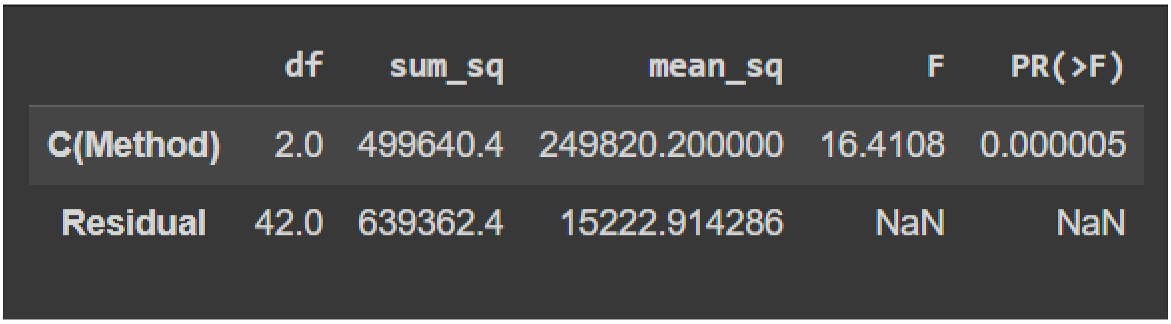
	df	sum_sq	mean_sq	F	PR(>F)
C(Method)	2.0	148472.177778	74236.088889	6.263327	0.004163
Residual	42.0	497805.066667	11852.501587	NaN	NaN

From the given ANOVA table, we can notice that the P-value is lesser than the 0.05 level of significance. Therefore, we fail to accept the null hypothesis at 5% level of significance. The mean impact hardness is different for at least one method.

ALLOY 2

Ho: Mean impact hardness of Alloy 2 at different methods are the same

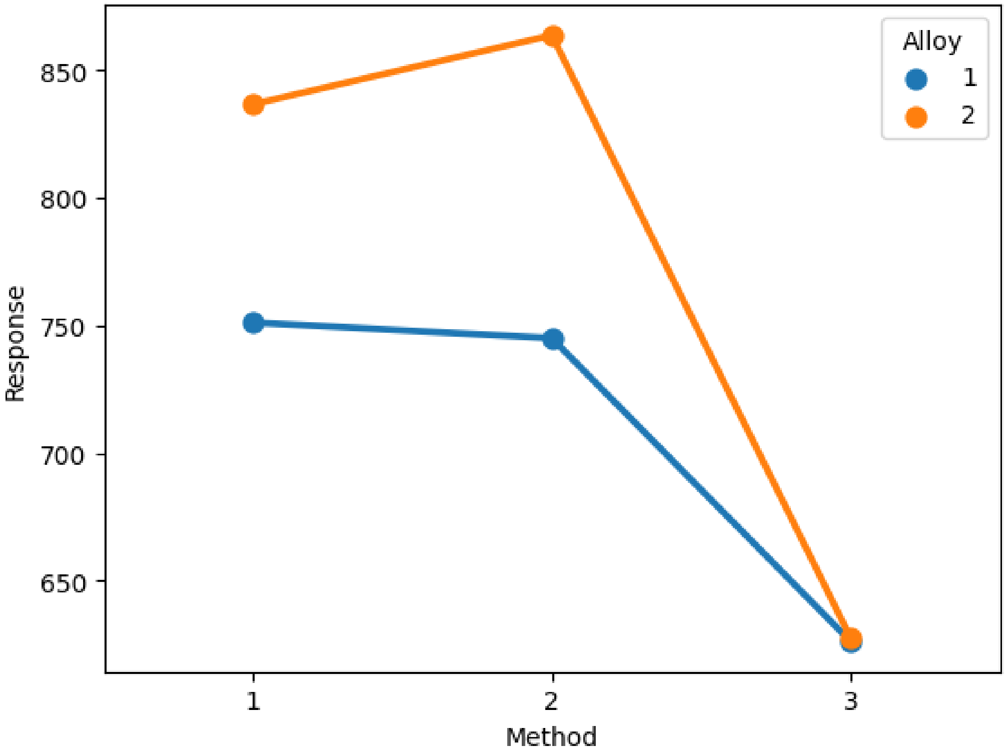
Ha: Mean impact hardness of Alloy 2 is different for at least one method



	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

We can notice from the given table that the P-value is lesser than the 0.05 level of significance. Therefore, we fail to accept the null hypothesis at 5% level of significance. The mean impact hardness is different for at least one method.

Since the null hypothesis is rejected for both the Alloys, we shall look into the pointplot diagram to ascertain the differences.



The Alloy 1 is much lesser than the alloy 2 in dental hardness based on the method. However, the 3rd method seems to have the lowest dental hardness for both the alloys. Therefore, there is an interaction between the alloys in the 3rd method. The dental hardness tends to vary the most between both alloys at 2nd method than the 1st method.

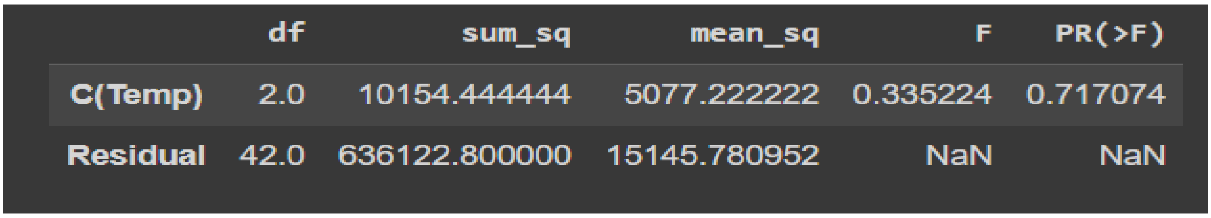
5)Now test whether there is any difference among the temperature levels on the hardness of the dental implant, separately for the two types of alloys. What are your conclusions? If the null hypothesis is rejected, is it possible to identify which levels of temperatures differ?

ALLOY 1

The hypothesis for the one-way ANOVA test is as follow:

Ho: Mean impact hardness of Alloy 1 among all the temperature levels are the same

Ha: Mean impact hardness of Alloy 1 is different for at least one the temperature levels



	df	sum_sq	mean_sq	F	PR(>F)
C(Temp)	2.0	10154.444444	5077.222222	0.335224	0.717074
Residual	42.0	636122.800000	15145.780952	NaN	NaN

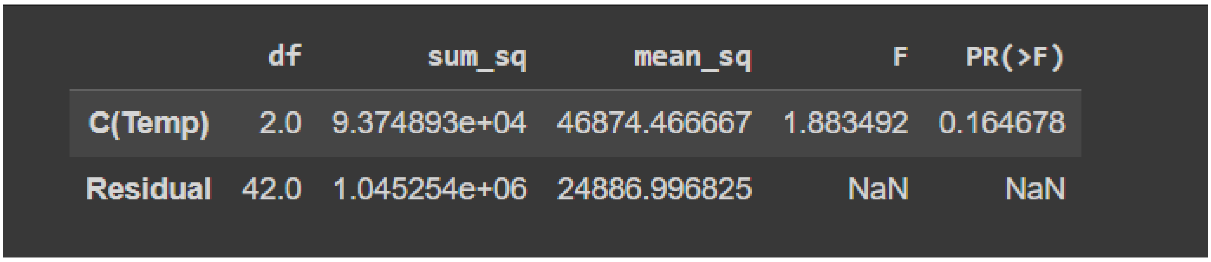
From the given ANOVA table we can see that the P-value is higher than the 0.05 level of significance. Therefore, we accept the null hypothesis at 5% level of significance. The mean impact hardness of Alloy 1 among all the temperature levels are the same.

ALLOY 2

The hypothesis for the one-way ANOVA test is as follow:

Ho: Mean impact hardness of Alloy 2 among all the temperature levels are the same

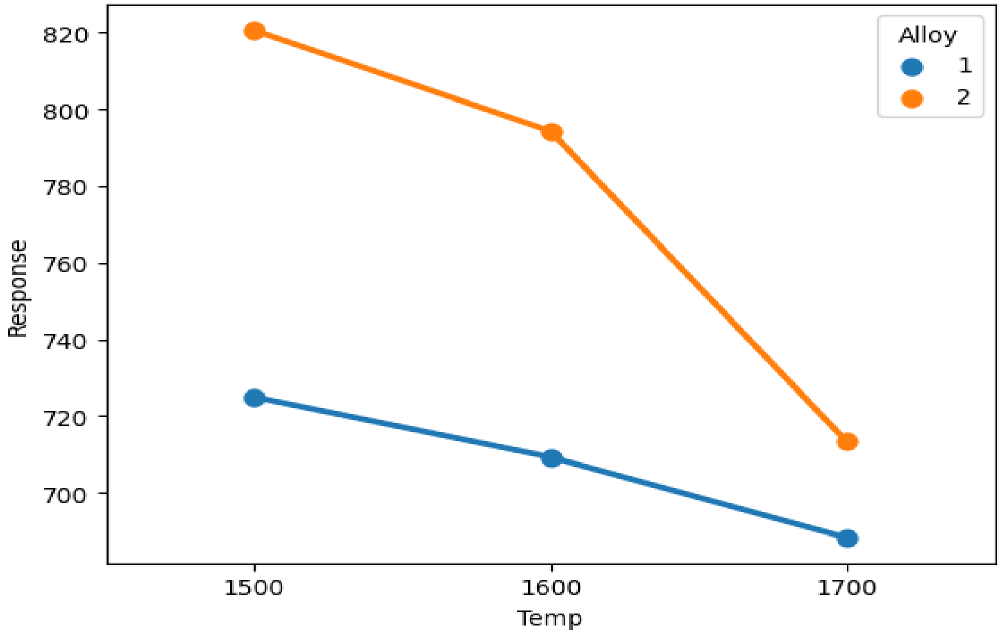
Ha: Mean impact hardness of Alloy 2 is different for at least one the temperature levels



	df	sum_sq	mean_sq	F	PR(>F)
C(Temp)	2.0	9.374893e+04	46874.466667	1.883492	0.164678
Residual	42.0	1.045254e+06	24886.996825	NaN	NaN

From the given ANOVA table we can see that the P-value is higher than the 0.05 level of significance. Therefore, we accept the null hypothesis at 5% level of significance. The mean impact hardness of Alloy 2 among all the temperature levels are the same.

Since the null hypothesis were accepted at 5% level of significance, let us try to find interactions or differences through a point plot diagram.



Generally, The dental hardness for Alloy 2 is higher than Alloy 1. Something to notice with the 3 temperature level is that at 1500 degrees the dental hardness is higher in both the alloys. It gets gradually lower as the temperature gets higher from 1600 degrees to 1700 degrees. Therefore the temperature seems to have a similar effect in both the alloys.

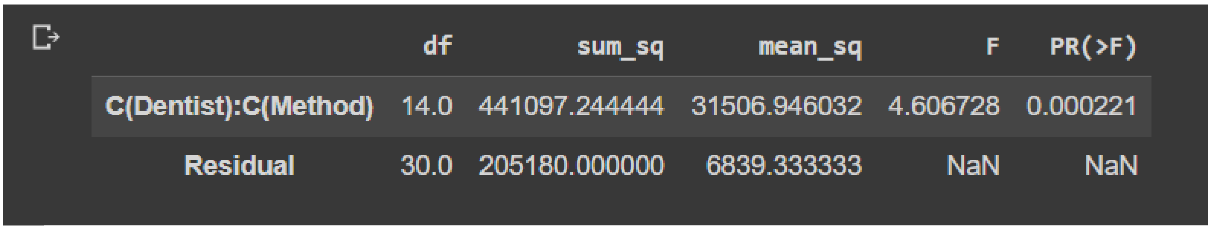
6)Consider the interaction effect of the dentist and method and comment on the interaction plot, separately for the two types of alloys?

ALLOY 1

The hypothesis for the ANOVA test is as follow:

Ho: There seems to be no interaction between method used and the different types of dentists in alloy 1

Ha: There seems to be an interaction between at least one method used and the different types of dentists in alloy 1



	df	sum_sq	mean_sq	F	PR(>F)
C(Dentist):C(Method)	14.0	441097.244444	31506.946032	4.606728	0.000221
Residual	30.0	205180.000000	6839.333333	NaN	NaN

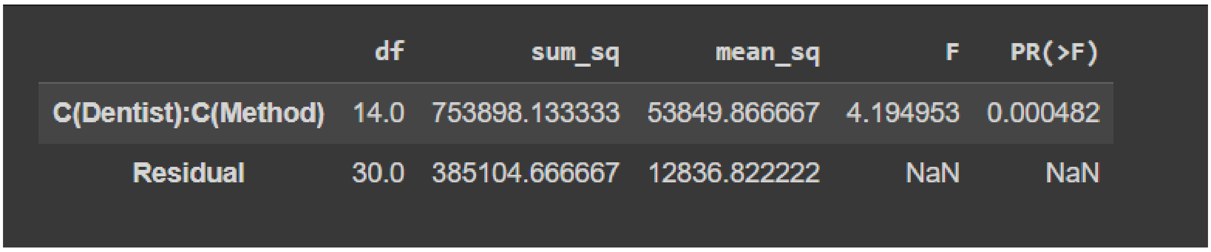
From the given ANOVA table we can see that the P-value is lesser than the 0.05 level of significance. Therefore, we fail to accept the null hypothesis at 5% level of significance. There seems to be an interaction between at least one method used and different types of dentists in alloy 1.

ALLOY 2

The hypothesis for the ANOVA test is as follow:

Ho: There seems to be no interaction between method used and the different types of dentists in alloy 2

Ha: There seems to be an interaction between at least one method used and the different types of dentists in alloy 2



	df	sum_sq	mean_sq	F	PR(>F)
C(Dentist):C(Method)	14.0	753898.133333	53849.866667	4.194953	0.000482
Residual	30.0	385104.666667	12836.822222	NaN	NaN

From the given ANOVA table we can see that the P-value is lesser than the 0.05 level of significance. Therefore, we fail to accept the null hypothesis at 5% level of significance. There seems to be an interaction between at least one method used and different types of dentists in alloy 2.

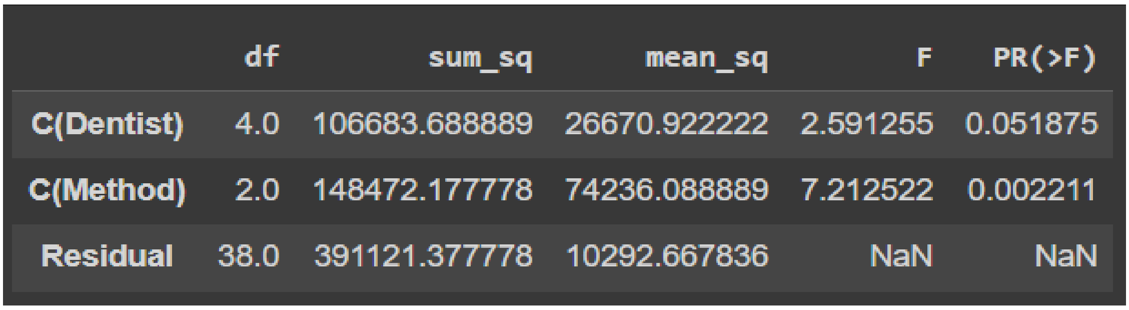
7)Now consider the effect of both factors, dentist, and method, separately on each alloy. What do you conclude? Is it possible to identify which dentists are different, which methods are different, and which interaction levels are different?

ALLOY 1

The hypothesis for the two-way ANOVA test is as follow:

Ho: Mean impact hardness is the same among all dentists and at all methods for Alloy 1.

Ha: Mean impact hardness is different at atleast one dentist or one method Alloy 1.



	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

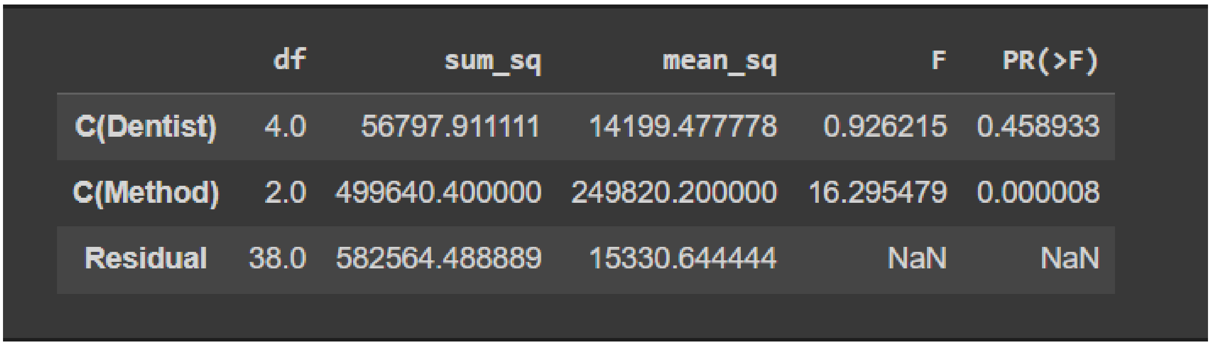
From the above given ANOVA table, we can see that the P-value of Dentist is slightly higher than the 0.05 level of significance. Whereas, the P-value of Method is lesser than 0.05 level of significance. Therefore, we fail to accept the null hypothesis at 5% level of significance. Mean impact hardness is different at atleast one dentist or one method Alloy 1.

ALLOY 2

The hypothesis for the two-way ANOVA test is as follow:

Ho: Mean impact hardness is the same among all dentists and at all methods for Alloy 2.

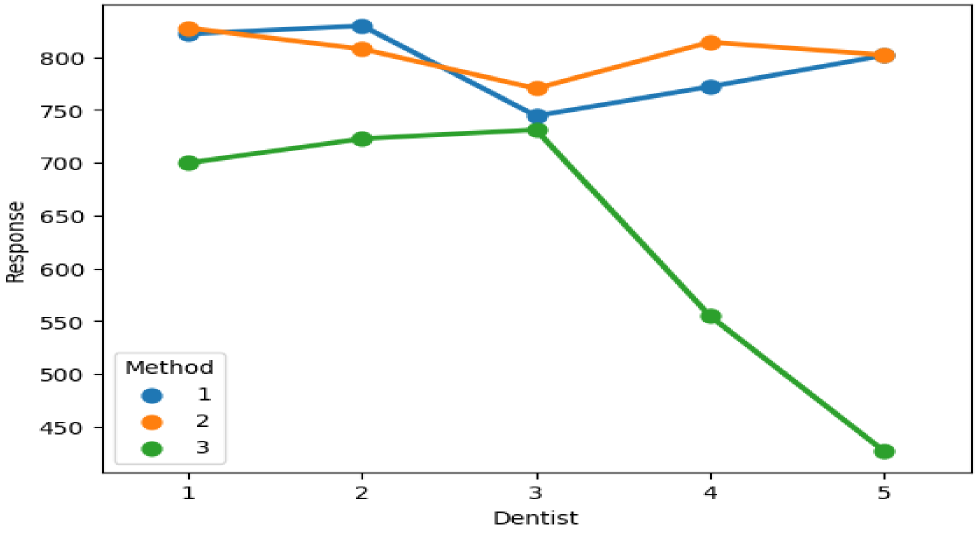
Ha: Mean impact hardness is different at atleast one dentist or one method Alloy 2.



	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

From the above given ANOVA table, we can see that the P-value of Dentist is higher than the 0.05 level of significance. Whereas, the P-value of Method is lesser than 0.05 level of significance. Therefore, we fail to accept the null hypothesis at 5% level of significance. Mean impact hardness is different at atleast one dentist or one method Alloy 2.

To study the difference and interaction effects in between the variables let us look at the point plot diagram.



The line representing the third method is not coinciding with the other lines. Thus, we can say that the influence of the third method on impact hardness is different from the others. The line representing the first method and the second method coincide at Dentist 1 and Dentist 5, i.e. there is an interaction between both variables at the aforementioned points. Another point to notice is that the first method line and second method line overlaps at a point in between 2 and 3 (Dentist) and the line lies at close proximity to each other.