2022

Project AS

YARESH VIJAYASUNDARAM

GREAT LEARNING | POST GRADUATE PROGRAM IN DATA SCIENCE AND BUSINESS ANALYTICS

Table of Contents

List of Figures:

List of Tables:

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**

From the above table, it is evident that there are total of 235 football players. Out of which, 145 are injured and 90 are not injured.

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

P(suffered with Injury)=Total players injured / Total
=145/235
=0.617

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

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

P(Forward U Winger) =P(Forward) + P(Winger) / P(Forward n Winger)
=94/235 + 29/235
=0.5234

The probability that a player is forward or a winger is **0.5234**

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

P(Striker player has a foot injury)/ P(Randomly chosen player)
=45/235
=0.1914

The probability of randomly chosen player plays in striker position and has a foot injury is **0.1914**

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

Total Sticker with foot injury =45
Total player injured = 145
Probability (Randomly selected player with foot injury) = Total Sticker with foot injury/ Total player injured
=45/145
=0.3103

The probability that a randomly chosen injured player is a striker is **0.3103**

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

Total player injured = 145
Total forward or attacking midfielder injured = 56 + 24
P(chosen injured player is either a forward or an attacking midfielder) = 80/145=0.5517

The probability that a randomly chosen injured player is either a forward or an attacking midfielder is **0.5517**

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:

The probability provided in the question:

P(R n F) = 0.001 P(R|F)=0.2

P(R n M) = 0.0015 P(R|M)=0.5
P(R n H) = 0.0012 P(R|H)=0.1

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

Probability of fire:

P(F) = P (R n F) / P (R|F)
= 0.001/0.2
= 0.005

The probability of the fire is **0.005.**

Probability of Mechanical failure:

P(M) = P (R n M) / P (R|M)
= 0.0015/0.5
= 0.003

The probability of Mechanical failure is **0.003.**

Probability of a Human Error:

P(H) = P(R n H) / P (R|H)
=0.0012/0.1
=**0.012**

The probability of a Human Error is 0.012.

2.2 What is the probability of a radiation leak?

P(R) = P(R n F) + P (R n M) + P (R n H)
= 0.001 + 0.0015 + 0.0012
= 0.0037

The probability of a radiation leak is **0.0037**

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.
A Mechanical Failure.
A Human Error.

P(F|R) = P (R n F) / P (R)
= 0.001 / 0.0037
= 0. 0.2702702702702703
Probability of radiation leak caused by fire is **0. 0.2703**

P(M|R) = P (R n M) / P (R)
= 0.0015 / 0.0037
= 0.4054054054054054

Probability of radiation leak caused by mechanical failure is **0.4054.**

P(H|R) = P (R n H) / P (R)
= 0.0012/0.0037
= **0.3243243243243243**

Probability of radiation leak caused by Human error is **0.3243.**

Problem 3:

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

Population Mean = 5
Population Standard deviation is 1.5kg is given the dataset.

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

Mu = 5
Sigma = 1.5

CDF (Cumulative distribution function) is used to find the proportion of gunny bags having a breaking strength less than 3.17 kg per sq. cm which is **11.12%**

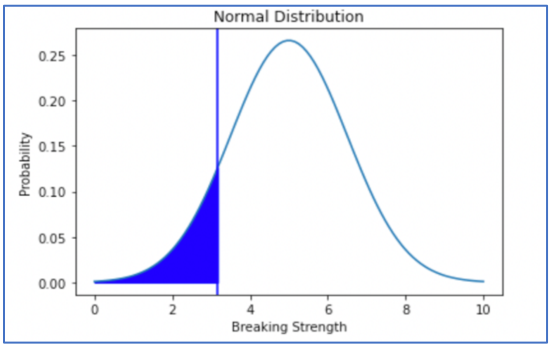

Figure 1: Probability density of breaking strength less than 3.17kg

The shaded area in blue shows the proportion of gunny bags having a breaking strength less than 3.17.

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

CDF (Cumulative distribution function) is used to find the proportion of gunny bags having a breaking strength at least 3.6kg per sq. cm which is **82.46%**

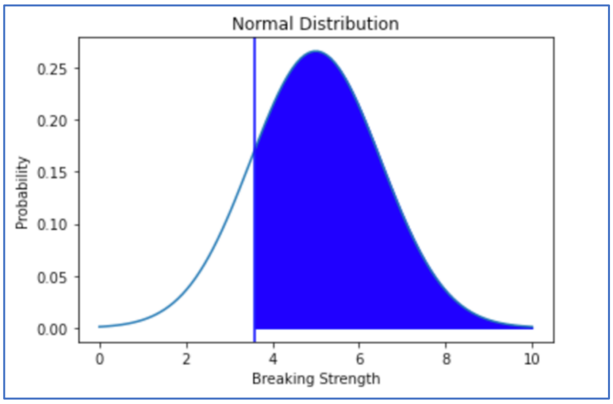

Figure 2: Probability density of breaking strength at least 3.6kg

The shaded area in blue denotes the proportion of gunny bags having a breaking strength at least 3.6kg per sq. cm.

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

CDF (Cumulative distribution function) is used to find the proportion of gunny bags having a breaking strength between 5 kg and 5.5 kg per sq. cm which is **13.06%**

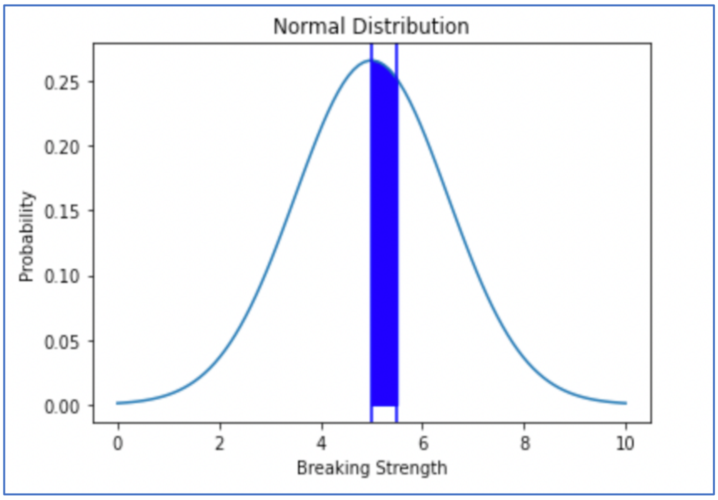

Figure 3:Probability density of breaking strength between 5 and 5.5 kg

The shaded area in blue depicts the proportion of gunny bags having a breaking strength between 5 and 5.5kg per sq. cm.

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

CDF (Cumulative distribution function) is used to find the proportion of gunny bags having a breaking strength not between 3 kg and 7.5 kg per sq. cm which is **13.90%**

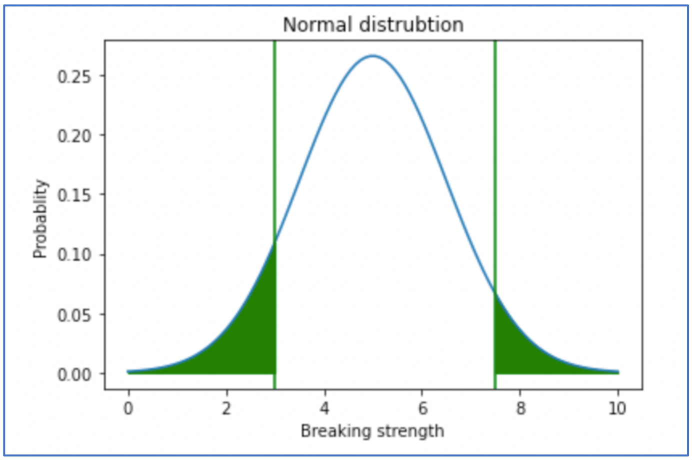

Figure 4:Probability density of breaking strength not between 3 and 7.5 kg

The two shaded area in green depicts the proportion of gunny bags having a breaking strength not between 3 and 7.5kg 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.

Mean is 7.7
Standard deviation is 8.5

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

CDF function is used to find the probability that a randomly chosen student who gets a grade below 85.

The probability is 0.8266

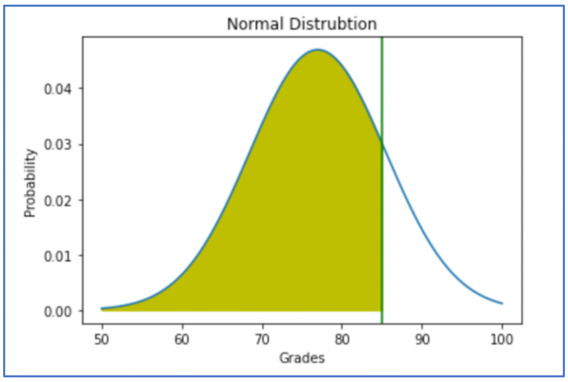

Figure 5:Probability density of a student getting grade below 85 on exams

The above yellow shade shows the probability of a randomly chosen student who gets a grade below 85.

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

CDF function is used to find the probability that a randomly selected student who scores between 65 and 87 and the probability is **0.8012**

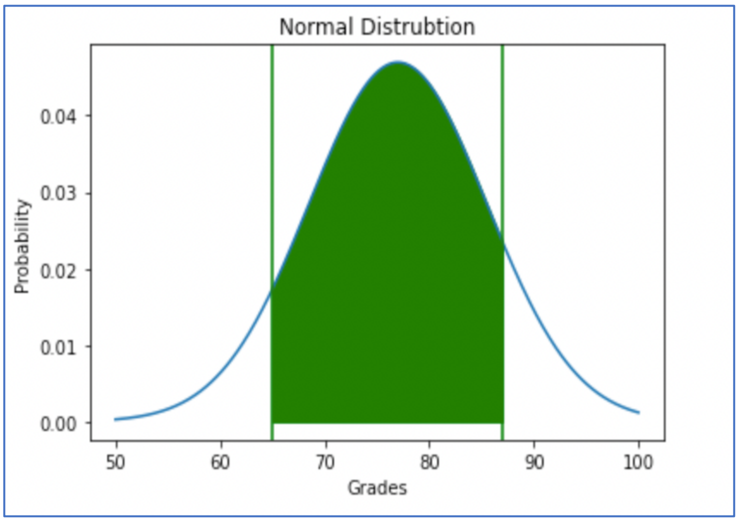

Figure 6: Probability density of a randomly selected student scores between 65 and 87

The above fig 6 shaded in green shows the probability of a randomly chosen student who gets a grade between 65 and 87.

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

norm.ppf **(percent point function)** function is used to find out the passing cut off so that 75% of the student clear the exam and **the cut off is 71.26**

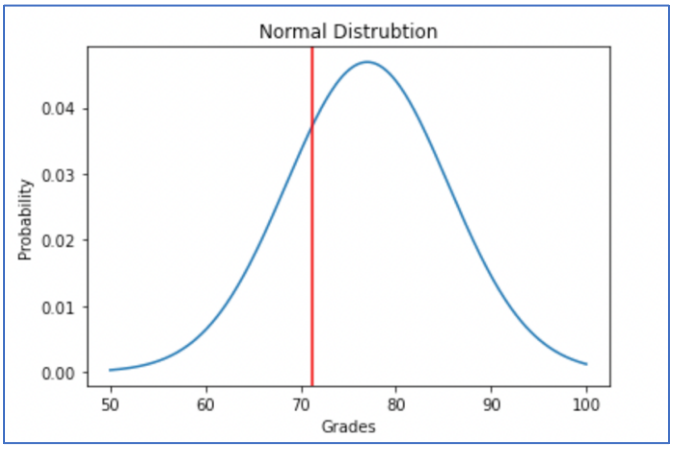

Figure 7:Percentage point function of the passing cut-off

The red line showing in the above graphs is the passing cut off.

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 favourable 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?

Null Hypothesis: The mean hardness of unpolished stone is at least equal to 150.
Alternative Hypothesis: The mean hardness of unpolished stone is less than 150.

Step 1: We define the null and alternative hypothesis

H0: $\mu>=150$
Ha: $\mu<150$

The sample size of the dataset is 75.
There is no null value present in the dataset.

Step 2:We Decide the significance level
Here we select α = 0.05.

Step 3: Identify the test statistic
We do not know the population standard deviation. So we use the t distribution and the t-test statistic.

Step 4: We use one sample t test to find the t – statistic and p values.

```
One sample t test
t statistic: -4.164629601426758 p value: 8.342573994839285e-05
```

For the one tail t -test we divide the p value by 2 which is: 4.1712869974196425e-05

Step 5: we decide whether to reject or accept the 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= 4.171287e-05
```

P value is 4.171287e-05 and it is less than 5% level of significance

So the statistical decision is to reject the null hypothesis at 5% level of significance. Therefore, at 95% confidence interval there is evidence that unpolished stones are not suitable for printing.

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

Hypothesis Defined:

Null Hypothesis: The mean hardness of polished and unpolished stones are the same.

Alternative Hypothesis: The mean hardness of polished and unpolished stones are not the same

H0: $\mu1= \mu2$
Ha: $\mu1\neq \mu2$

We Decide the significance level - Here we select α = 0.05.

Identify the test statistic - We do not know the population standard deviation. So we use the t distribution and the t-test statistic

We use one sample t test to find the t – statistic and p values.

```
two sample t test
t statistic: -3.242232050141406 p value: 0.001465515019462831
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= 0.001465515019462831
```

P value is 0.001465515019462831 and it is less than 5% level of significance. So the statistical decision is to reject the null hypothesis at 5% level of significance. Therefore, there is evidence that the mean hardness of unpolished and polished are not same.

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.

Defining Hypothesis:

Null Hypothesis – The mean difference of push ups is equal to 5
Alternative Hypothesis – The mean difference of push ups is greater than 5

H0: $\mu = 5$
Ha: $\mu > 5$

The Mu value given in the dataset is 5. The mean difference and standard deviation is calculated to find out the t stat by using the formula.

$$t_{stat} = \frac{(\bar{d} - \mu_D)}{S_d / \sqrt{n}}$$

And T stat is calculated to be 1.9148542155126758, and p value is calculated using the T stats which is 0.029198872141011245.

```
Level of significance: 0.05
We have evidence to reject the null hypothesis since p value < Level of significance
t-test P-value= 0.029198872141011245
```

P value = 0.029 is lesser than the 5% level of significance. So, the null hypothesis is rejected. Therefore, there is enough evidence to prove the claim that the training will make a difference of more than 5 push-ups.

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.

There are total of 90 values present in the dataset. There is no null value present in the dataset. However there 4 float variables which is supposed to be categorical variable. Therefore, the 4 variables (Dentist, Method, Alloy & Temp) is converted to a categorical variable using pd. Categorical() method.

7.1 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.?

Hypothesis for alloy 1:

H0: The Mean dental hardness among all the dentist is same for alloy 1

H1: The Mean dental hardness among all the dentist is not the same for alloy 1

	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

Table 1: ANOVA Table – Difference among the dentist on implant hardness for alloy 1

The significance level - Here we select **α = 0.05.**

Since the P value is greater than significance level 0.05. Therefore, we fail to reject the null hypothesis. We can conclude by saying the mean dental hardness among all the dentist is same for alloy 1.

Hypothesis for alloy 2:

H0: The Mean dental hardness among all the dentist is same for alloy 2

H1: The Mean dental hardness among all the dentist is not the same for alloy 2

Table 2: ANOVA Table – Difference among the dentist on implant hardness for 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

The significance level - Here we select $\alpha = 0.05$.

As we could see from the above table that the P value (0.718031) is > significance level of 0.05. we fail to reject the null hypothesis. Therefore, we conclude by saying that the mean dental hardness among all the dentist is same for the alloy 2.

7.2 Before the hypotheses may be tested, state the required assumptions. Are the assumptions fulfilled? Comment separately on both alloy types.?

Assumptions are:

The data is drawn from a normal distribution.

The variance are equal.

After the Shapiro result. We can conclude that, the alloy 1 is normally distributed and the alloy 2 is not normally distributed.

From the Levene test, it is evident that the both the alloys variance are equal.

7.3 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?

Define Hypothesis:

H0: The mean implant hardness is same with all five dentist.

H1: The mean implant hardness is not the same with at least one dentist.

	df	sum_sq	mean_sq	F	PR(>F)
C(Dentist)	4.0	1.577946e+05	39448.638889	1.934537	0.112066
Residual	85.0	1.733301e+06	20391.776471	NaN	NaN

Table 3:ANOVA Table – Difference among the dentist on implant hardness

The significance level - Here we select $\alpha = 0.05$.

As we could see from the above Table.3 that the P value (0.112066) is greater than the significance level, we fail to reject the null hypothesis, the mean implant hardness is same for all the dentist.

7.4 Now test whether there is any difference among the methods on the hardness of 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?

Defining Hypothesis for alloy 1:

H0:The mean dental hardness implant is same across all methods for alloy1

Ha: The mean hardness of dental implant is different for at least one method for alloy 1

Table 4:ANOVA Table – Difference among the methods on implant hardness for alloy 1

	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

The significance level is α = 0.05.

From the above Table 4 it is evident that the P value is (0.004163) is < (lesser than) the significance level. Therefore, we reject the null hypothesis. And conclude by saying the mean hardness of dental implant is different for at least one method for alloy 1

Though the null hypothesis is rejected, it is not possible for ANOVA to tell which particular groups were statistically significant from one another, since ANOVA can only compare only the mean of groups and exhibits the statistical difference between the means. However, the Tukeys HSD and Tukey Kramer procedure can be used identify which pair methods differ.

Defining Hypothesis for alloy 2:

H0:The mean dental hardness implant is same across all methods for alloy2

Ha: The mean hardness of dental implant is different for at least one method for alloy2

Table 5:ANOVA Table – Difference among the methods on implant hardness for alloy 2

	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

The significance level is α = 0.05.

From the above fig it is evident that the P value is (0.00005) is < (lesser than) the significance level. Therefore, we reject the null hypothesis. And conclude by saying the mean hardness of dental implant is different for at least one method for alloy2.

Though the null hypothesis is rejected, it is not possible for ANOVA tell to which particular groups were statistically significant from one another, since ANOVA can only compare the mean of groups and exhibits the statistical difference between the means. However, the Tukeys HSD and Tukey Kramer procedure can be used identify which pair methods differ

7.5 Now test whether there is any difference among the temperature levels on the hardness of 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?

Defining Hypothesis for alloy 1:

H0:The mean hardness of dental implant is the same for different level of temperature for alloy 1

Ha: The mean hardness of dental implant is different for at least one level of temperature for alloy 1

	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

Table 6:ANOVA Table – Difference among the temperature level implant hardness for alloy 1

The significance level is α = 0.05.

From the above Table 6, it is evident that the P value is (0.717074) is > (greater than) the significance level. Therefore, we fail to reject the null hypothesis. And conclude by saying the mean hardness of dental implant is same for the different level temperature for alloy 1.

Defining Hypothesis for alloy 2:

H0:The mean hardness of dental implant is the same for different level of temp for alloy 2

Ha: The mean hardness of dental implant is different for at least one level of temperature for alloy 2

	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

Table 7:ANOVA Table – Difference among the temperature level implant hardness for alloy 2

The significance level is α = 0.05.

From the above Table 7, it is evident that the P value is (0.164678) is > (greater than) the significance level. Therefore, we fail to reject the null hypothesis. And conclude by saying the mean hardness of dental implant is same for the different level temperature for alloy 2.

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

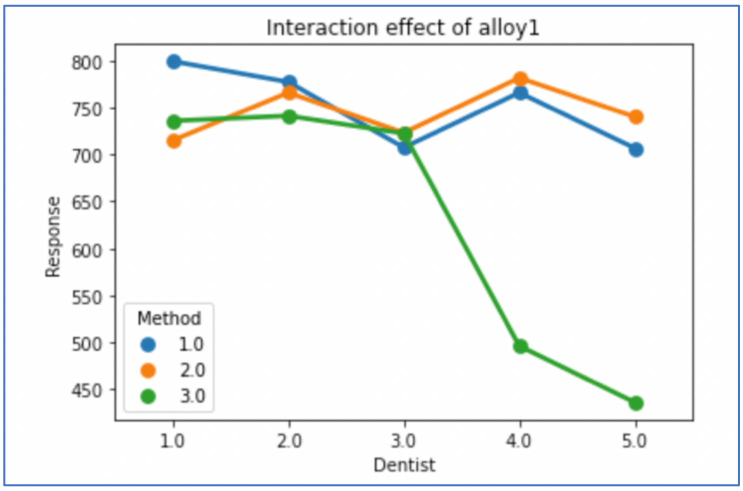

Figure 8:Interaction effect of Alloy 1

From the above Fig.8, we could see an interaction between the Method 1, 2 and 3 and it is also clear that the mean hardness among method 1 and 2 follows a similar pattern. Furthermore, we can also see the sharp decline in the method 3 for the dentist 3 and 4.

2

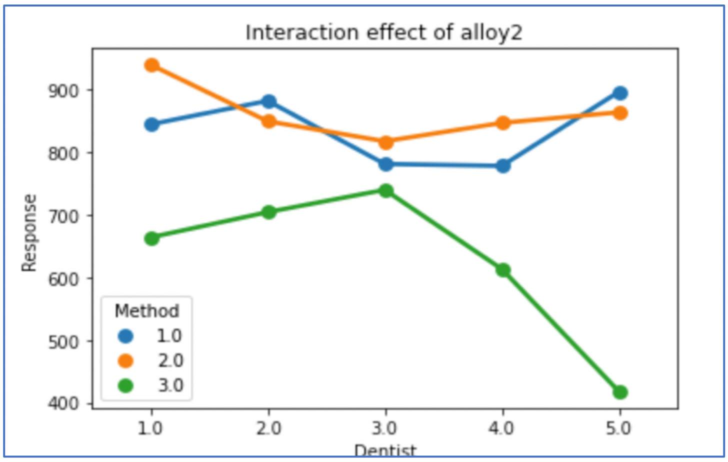

Figure 9:Interaction effect of Alloy 2

From the above fig.9, we can see there is an interaction between the method 1 and method 2. And the mean hardness remains high for the method 1 and 2 among all the dentist compared to the method 3.

7.7 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?

Interaction effect for dentist and method for alloy 1:

H0: There is interaction among the dentist and method for alloy 1
H1: There is no interaction among the dentist and methods for alloy 1

Table 8:ANOVA Table – Interaction effect of dentist and method for alloy 1

	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

We take the significance level as 0.05.

From the above table 8, it is clear that the P value (0.006793) is lesser than significance level. Therefore, we reject the null hypothesis. And we conclude by saying there is no interaction among the dentist and method for alloy 1.

Interaction effect for dentist and method for alloy 2:

H0: There is interaction among the dentist and method for alloy 2
H1: There is no interaction among the dentist and methods for alloy 2

	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

Table 9:ANOVA Table – Interaction effect of dentist and method for alloy 2

We take the significance level as 0.05.

From the above table 9, it is clear that the P value (0.093234) is greater than significance level. Therefore, we fail to reject the null hypothesis. And we conclude by saying there is no interaction among the dentist and method for alloy 2.

It is not possible for ANOVA to identify which dentists are different, which methods are different, and which interaction levels are different. Since ANOVA can only compare the mean of groups and exhibits the statistical difference between the means. However, the Tukeys HSD and Tukey Kramer procedure can be used identify which pairs differ.