Advance Statistics Project Report

Krishnameera K S

Table of 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

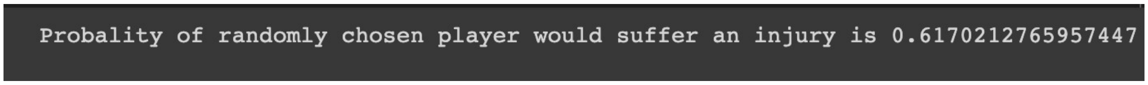
	Striker	Forward	Attacking Midfielder	Winger	**Total**
Players Injured	45	56	24	20	**145**
Players Not Injured	32	38	11	9	**90**
Total	**77**	**94**	**35**	**29**	**235**

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

Probability that a randomly chosen player would suffer an injury

= No of injured players/ Total no of players
= 145/ 235

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



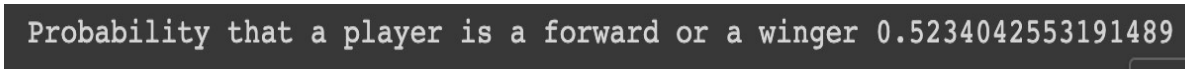
```
Probality of randomly chosen player would suffer an injury is 0.6170212765957447
```

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

Probability that a player is a forward or a winger

= (No of forward players+ No of Winger players) / Total no of players
= (94+29)/235
= 123/235

Probability that a player is a forward or a winger is **0.523**



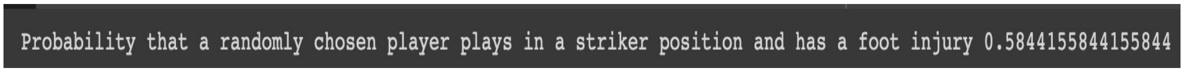
```
Probability that a player is a forward or a winger 0.5234042553191489
```

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

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

= No of Striker player and injured/ Total no of Striker players
= 45/77

Probability that a randomly chosen player plays in a striker position and has a foot injury is **0.584**



```
Probability that a randomly chosen player plays in a striker position and has a foot injury 0.5844155844155844
```

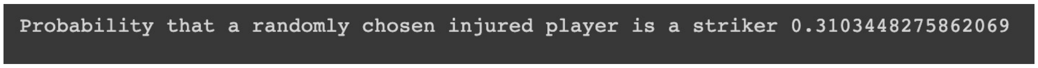
1.4 What is the probability that a randomly chosen injured player is a striker?

Probability that a randomly chosen injured player is a striker

= No of Striker player and injured/ Total no of Injured players

= 45/145

Probability that a randomly chosen injured player is a strike is **0.310**



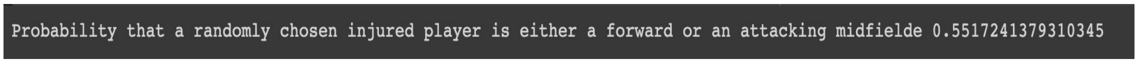
```
Probability that a randomly chosen injured player is a striker 0.3103448275862069
```

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

Probability that a randomly chosen injured player is either a forward or an attacking midfielder

= (No of injured forward players + No of injured attacking midfielder players) / Total no of injured players
= (56+24)/145

Probability that a randomly chosen injured player is either a forward or an attacking midfielder is **0.551**



```
Probability that a randomly chosen injured player is either a forward or an attacking midfielde 0.5517241379310345
```

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%.

Based on the information available, answer the questions below:

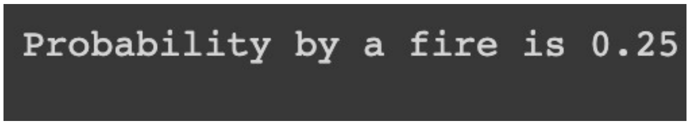
2.1 What are the probabilities of a fire, a mechanical failure, and a human error respectively?

2.1.1 Probability by a fire = Probability of Radiation leak in case of fire / Total no of radiation leak outcomes

= 20 / (20+50+10)

= ¼

Probability by a fire is **0.25**



```
Probability by a fire is 0.25
```

2.1.2 Probability by a Mechanical failure = Probability of Radiation leak in case of Mechanical fire / Total no of radiation leak outcomes

= 50/(20+50+10)

= 5/8

Probability by a Mechanical failure is **0.625**

```
Probability by a Mech is 0.625
```

2.1.3 Probability by a Human error = Probability of Radiation leak in case of Human error / Total no of radiation leak outcomes

= 10/(20+50+10)

= 1/8

Probability by a Human error is **0.125**

```
Probability by a Human error is 0.125
```

2.2 What is the probability of a radiation leak?

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%.

Probability of a radiation leak = Sum of probabilities of a radiation leak occurring simultaneously with a fire, mechanical failure, human error

= 0.1 + 0.15 + 0.12

= 0.37

Probability of a radiation leak is **0.37**

```
Probability of a Radiation Leak is 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.

A Mechanical Failure.

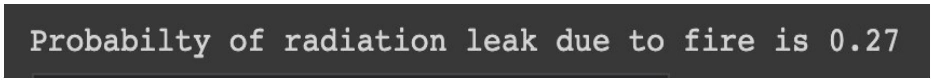
A Human Error.

2.3.1 Probability that it has been caused by a fire

= Probability of radiation leak due to fire/Probability of radiation leak

= 0.1/0.37

Probability that it has been caused by a fire is **0.27%**



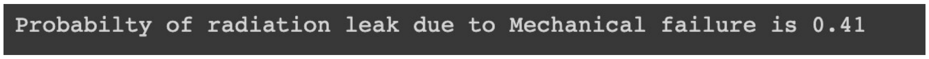
```
Probabilty of radiation leak due to fire is 0.27
```

2.3.2 Probability that it has been caused by a Mechanical Failure

= Probability of radiation leak due to Mechanical Failure/Probability of radiation leak

= 0.15/0.37

Probability that it has been caused by a Mechanical Failure is **0.41%**



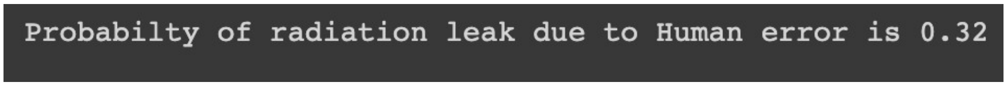
```
Probabilty of radiation leak due to Mechanical failure is 0.41
```

2.3.3 Probability that it has been caused by a Human error

= Probability of radiation leak due to Human error /Probability of radiation leak

= 0.15/0.37

Probability that it has been caused by a Human error is **0.32%**



```
Probabilty of radiation leak due to Human error is 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. 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)

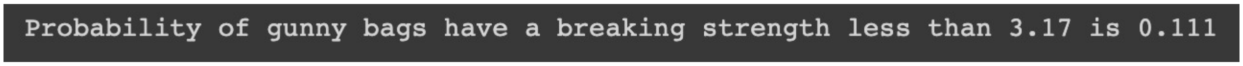
3.1 What proportion of the gunny bags have a breaking strength less than 3.17 kg per sq cm?

Mean = 5 kg per sq. centimetre

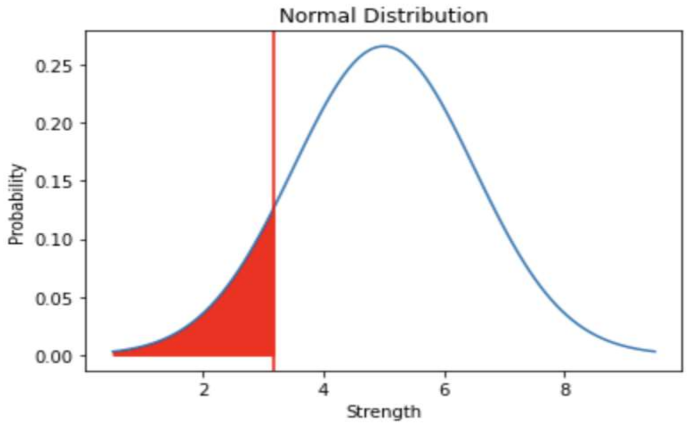
Standard Deviation = 1.5 kg per sq. centimetre

To find the probability of the gunny bags have a breaking strength less than 3.17 kg per sq cm, we have to find the cumulative probability

Probability of the gunny bags have a breaking strength less than 3.17 kg per sq cm is **0.111**



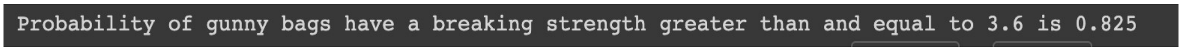
```
Probability of gunny bags have a breaking strength less than 3.17 is 0.111
```



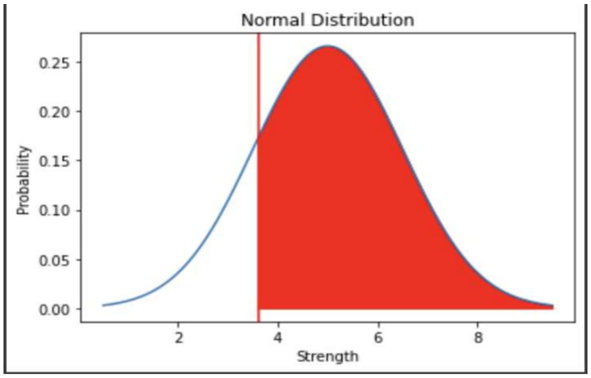
3.2 What proportion of the gunny bags have a breaking strength at least 3.6 kg per sq cm.?

To find the probability of the gunny bags have a breaking strength at least 3.6 kg per sq cm, which means more than or equal 3.6 kg per sq cm.Then we have to find the cumulative probability and subtract it from 1.

Probability of the gunny bags have a breaking strength at least 3.6 kg per sq cm is **0.825**



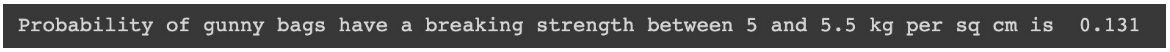
```
Probability of gunny bags have a breaking strength greater than and equal to 3.6 is 0.825
```



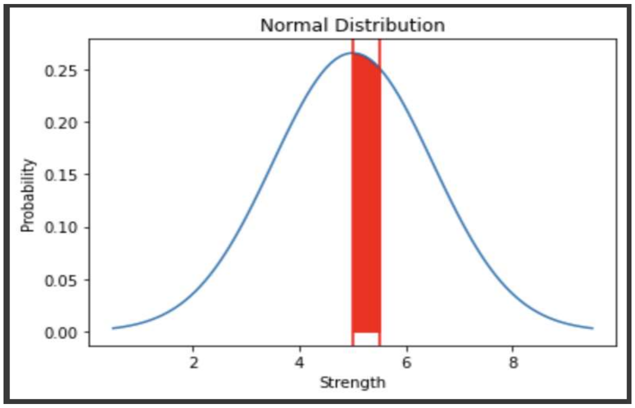
3.3 What proportion of the gunny bags have a breaking strength between 5 and 5.5 kg per sq cm.?

To find the probability of the gunny bags have a breaking strength between 5 and 5.5 kg per sq cm. Then we have to find the cumulative probability 5.5 and 5. After that subtract them.

Probability of the gunny bags have a breaking strength between 5 and 5.5 kg per sq cm is **0.131**



```
Probability of gunny bags have a breaking strength between 5 and 5.5 kg per sq cm is  0.131
```

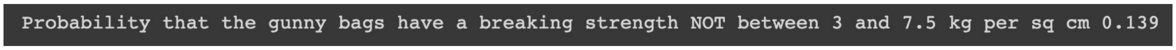


3.4 What proportion of the gunny bags have a breaking strength NOT between 3 and 7.5 kg per sq cm.?

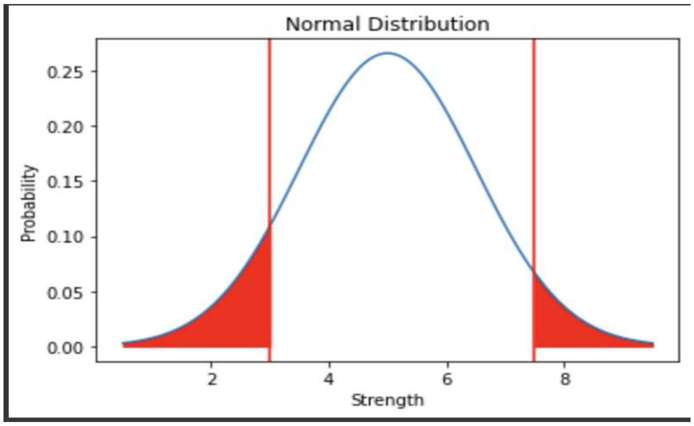
To find the probability of the gunny bags have a breaking strength not between 3and 7.5 kg per sq cm. Which means less that 3 and greater than 7.5.

Then we have to find the cumulative probability of 3 .Then cumulative probability of 7.5 and subtract it from 1. Add both these values to get the result.

Probability of the gunny bags have a breaking strength not between 3 and 7.5 kg per sq cm is **0.139**



```
Probability that the gunny bags have a breaking strength NOT between 3 and 7.5 kg per sq cm 0.139
```



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.

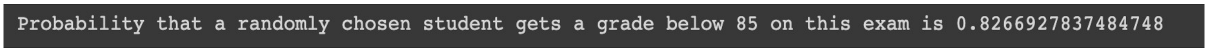
4.1 What is the probability that a randomly chosen student gets a grade below 85 on this exam?

Mean = 77

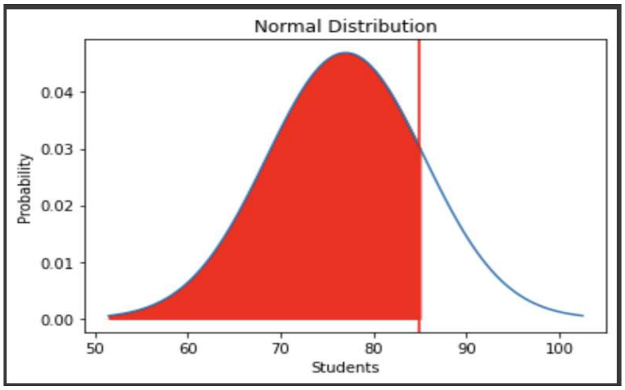
Standard Deviation = 8.5

To find the probability that a randomly chosen student gets a grade below 85 on this exam, we must find the cumulative probability.

Probability that a randomly chosen student gets a grade below 85 on this exam is **0.826**



```
Probability that a randomly chosen student gets a grade below 85 on this exam is 0.8266927837484748
```

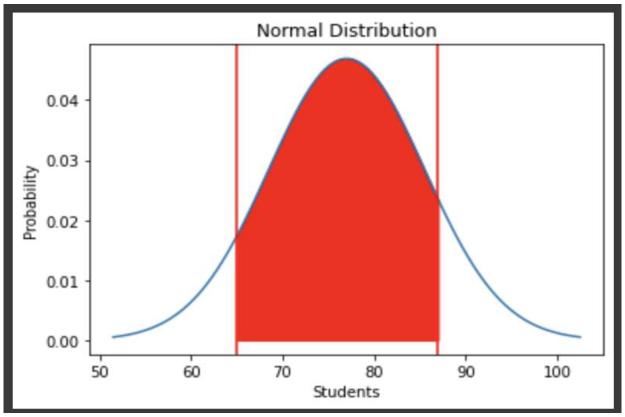


4.2 What is the probability that a randomly selected student score between 65 and 87?

To find the probability that a randomly chosen student score between 65 and 87on this exam, we must find the cumulative probability of 65 and same of 87. Then subtract both the values.

Probability that a randomly chosen student score between 65 and 87 on this exam is **0.801**

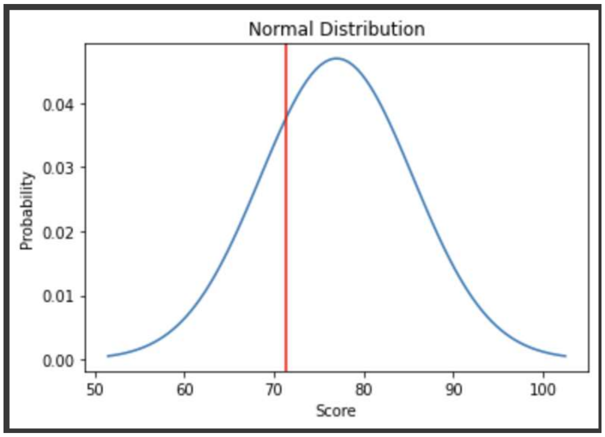
```
Probability that a randomly selected student scores between 65 and 87 is  0.801
```



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

Calculate the ppf of 25% percentile of students for passing cutoff

The passing cut-off mark so that 75% of the students clear the exam is **71.266**



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);

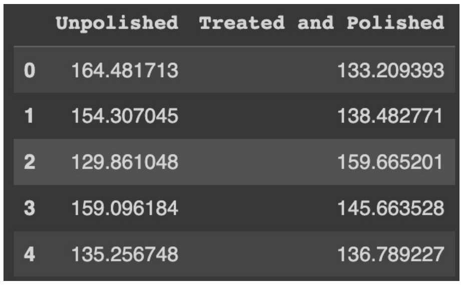
5.1 Earlier experience of Zingaro with this particular client is favourable as the stone surface was found to be of adequate hardness. However, Zingaro has reason to believe now that the unpolished stones may not be suitable for printing. Do you think Zingaro is justified in thinking so?

Step 1:

Let's define the Null (H0) and Alternate Hypothesis (H1)

H0: Unpolished stone has adequate hardness and it is suitable for printing >=150

H1: Unpolished stone does not have adequate hardness and it is not suitable for printing < 150



	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

Step 2:

Provided the details as level of significance (alpha) = 0.05

Step 3:

We don't know population sample deviation; hence we are using One Sample T test.

Size of data is 75 = N

Degree of freedom = N-1 = 75-1 = 74

Step 4:

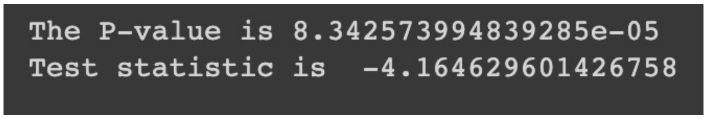
Calculate the Test statistic and P value.

Test statistic: -4.164

P-value: 0.000083

Step 5:

P-value is less than the alpha value (0.05). Hence we have enough evidence to reject the Null hypothesis. So



```
The P-value is 8.342573994839285e-05
Test statistic is  -4.164629601426758
```

Unpolished stone is not suitable for printing and hardness is less than 150

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

Step 1:

Let’s define the Null (H0) and Alternate Hypothesis (H1)

H0: Unpolished stone mean is equal to polished stone mean

H1: Unpolished stone mean is not equal to polished stone mean

Step 2:

Provided the details as level of significance (alpha) = 0.05

Step 3:

We don’t know population sample deviation and we have independent sample; hence we are using Two Sample T test.

N1=75

N2=75

Degree of freedom = N1 + N2 -2 = 75 + 75-2 = 148

Step 4:

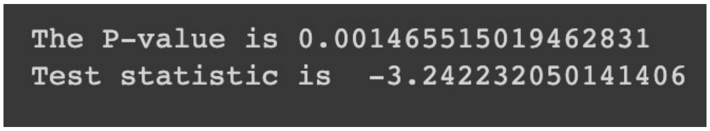
Calculate the Test statistic and P value.

Test statistic: -3.242

P-value: 0.00147

Step 5:

P-value is less than the alpha value (0.05). Hence we have enough evidence to reject the Null hypothesis. So



```
The P-value is 0.0014655150194628
Test statistic is  -3.242232050141406
```

Mean of Polished stone and Unpolished stone are not equal.

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 can 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.

Step 1:

Let’s define the Null (H0) and Alternate Hypothesis (H1)

H0: Mean push up of candidates before attending the program is equal t Mean push up of candidates after attending the program.

H1: Candidates able to do more than 5 push-ups after attending the program.

Step 2:

Provided the details as level of significance (alpha) = 0.05

Step 3:

Sample size of both sample is 100

Degree of freedom is 100-1 = 99

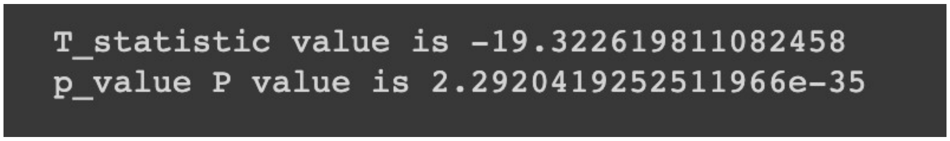
We have 2 paired sample; Hence we will proceed with Paired T-test.

Step 4:

Calculate the Test statistic and two- tailed P value.

We calculate the two related samples of the values.

We will validate if the mean of 2 related samples are equal.



```
T_statistic value is -19.322619811082458
p_value P value is 2.2920419252511966e-35
```

Test statistic: -19.323

P-value : 0.0000

Step 5:

P-value is less than the alpha value (0.05). Hence, we have enough evidence to reject the Null hypothesis. So we can conclude that

Candidates able to take 5 more push-ups after attending the program.

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.

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.?**

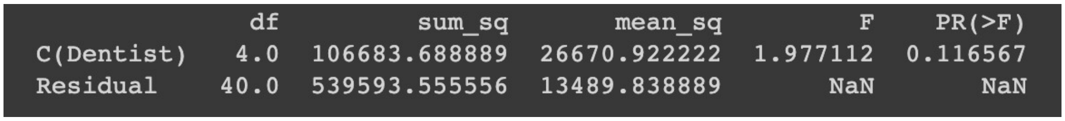
Let’s define the Null (H0) and Alternate Hypothesis (H1)

H0: The difference in dentist on different types on alloys does not affect the response

H1: The difference in dentist on different types on alloys affect the response

Let significance level is 0.05

Create data frames for 2 alloys. By creating the model with response and Dentist on different alloy types.



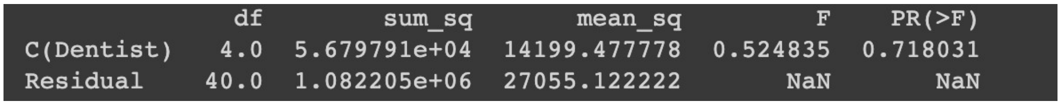
```
               df         sum_sq       mean_sq         F    PR(>F)
C(Dentist)    4.0  106683.688889  26670.922222  1.977112  0.116567
Residual     40.0  539593.555556  13489.838889       NaN       NaN
```

We could see the p value for dentist is 0.116 for First Alloy

P-value is greater than 0.05. Hence not able to reject null hypothesis.

Hence difference is dentist on Alloy1 does not affect the response.

Lets consider second Alloy2



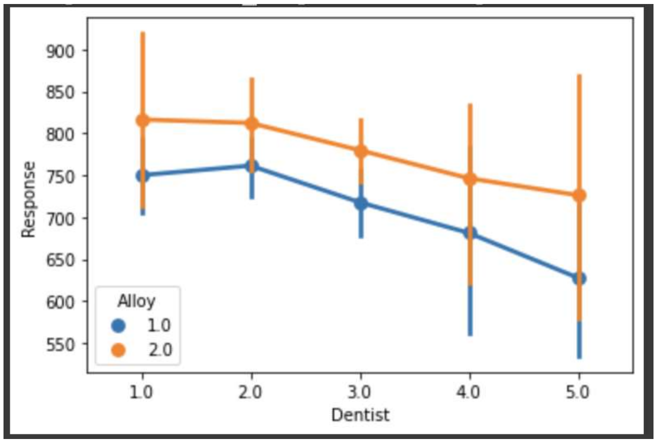
```
               df        sum_sq       mean_sq         F    PR(>F)
C(Dentist)    4.0  5.679791e+04  14199.477778  0.524835  0.718031
Residual     40.0  1.082205e+06  27055.122222       NaN       NaN
```

We could see the p value for dentist is 0.72 for Alloy2

P-value is greater than 0.05. Hence not able to reject null hypothesis.

Hence difference is dentist on Alloy2 does not affect the response

7.3 Irrespective of your conclusion in 7.2, 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?



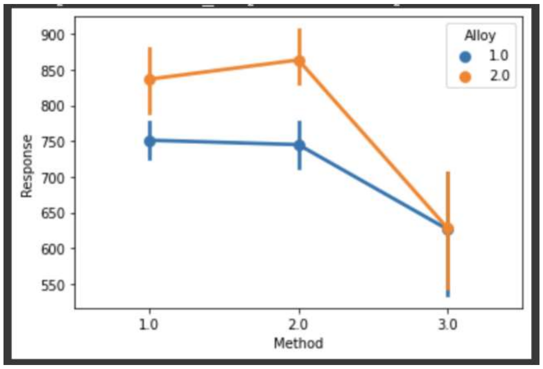
From the above graph we can conclude that Dentist 1 has high response and the response does not that much change between dentist if it alloy 1 is using. But Dentist response reduces from 2 to 5 with Alloy2 in more. Hence Alloy2 and Dentist make a difference compared to Alloy1 and Dentist

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?

Let's define the Null (H0) and Alternate Hypothesis (H1)

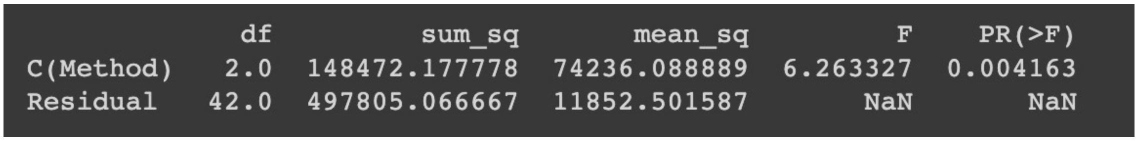
H0: The difference in Method on different types on alloys does not affect the response

H1: The difference in Method on different types on alloys affect the response



Let significance level is 0.05

Create data frames for 2 alloys. By creating the model with response and Method on different alloy types.



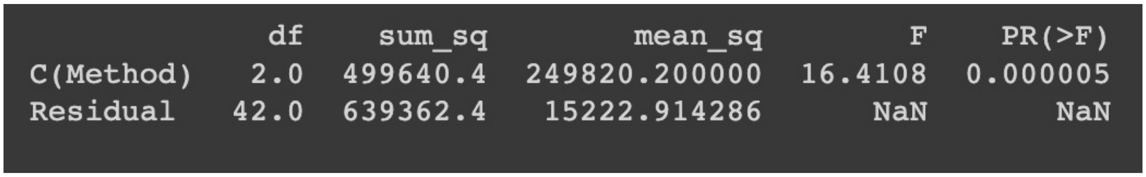
```
                df          sum_sq        mean_sq         F    PR(>F)
C(Method)     2.0   148472.177778   74236.088889  6.263327  0.004163
Residual     42.0   497805.066667   11852.501587       NaN       NaN
```

We could see the p value for Method is 0.00416 for First Alloy1

P-value is less than 0.05. Hence able to reject null hypothesis.

Hence difference is Method on Alloy1 affect the response.

Lets consider second 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
```

We could see the p value for Method is 0.000005 for Alloy2

P-value is less than 0.05. Hence we are able to reject null hypothesis.

Hence difference is dentist on Alloy2 affect the response

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?

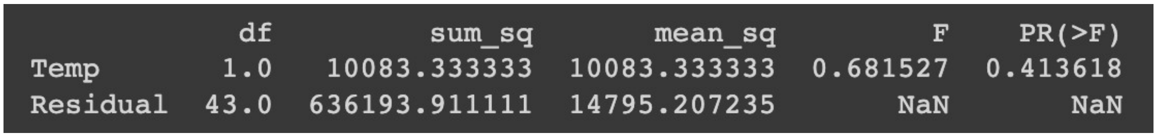
Let’s define the Null (H0) and Alternate Hypothesis (H1)

H0: The difference in Temperature on different types on alloys does not affect the response

H1: The difference in Temperature on different types on alloys affect the response

Let significance level is 0.05

Create data frames for 2 alloys. By creating the model with response and Temperature on different alloy types.



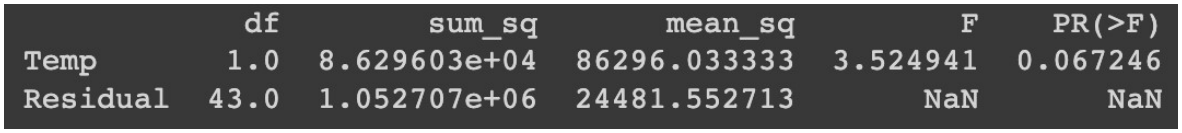
```
            df          sum_sq        mean_sq         F    PR(>F)
Temp       1.0    10083.333333   10083.333333  0.681527  0.413618
Residual  43.0   636193.911111   14795.207235       NaN       NaN
```

We could see the p value for Temperature is 0.413 for First Alloy

P-value is greater than 0.05. Hence not able to reject null hypothesis.

Hence difference is Temperature on Alloy1 does not affect the response.

Lets consider second Alloy2



```
            df        sum_sq        mean_sq         F    PR(>F)
Temp       1.0  8.629603e+04   86296.033333  3.524941  0.067246
Residual  43.0  1.052707e+06   24481.552713       NaN       NaN
```

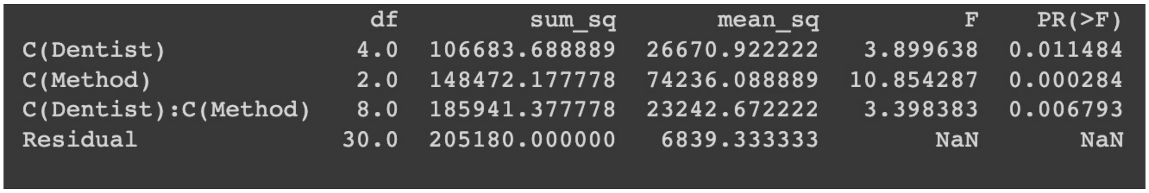
We could see the p value for Temperature is 0.067 for Alloy2

P-value is greater than 0.05. Hence not able to reject null hypothesis.

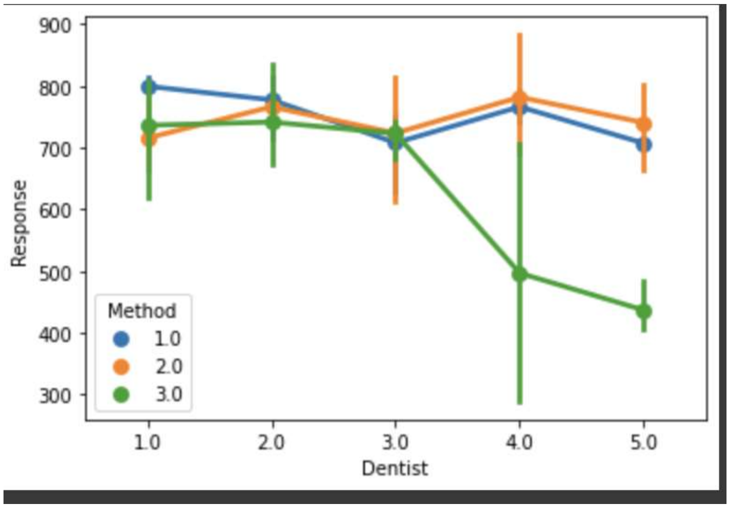
Hence difference is Temperature on Alloy2 does not affect the response

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

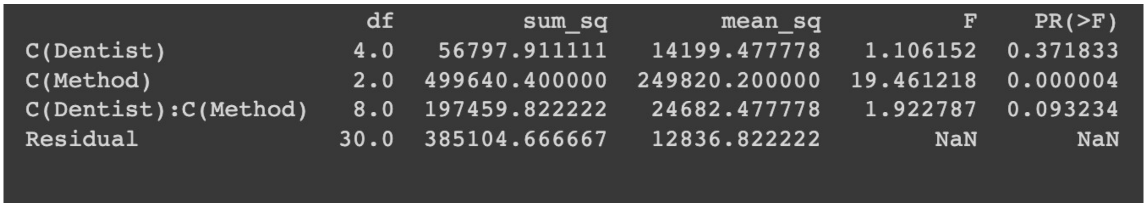
As per the Plot and table obtained, we can say that for Alloy1 Density and Method interaction is 0.006 which is less than 0.05. Hence there is statistical interaction



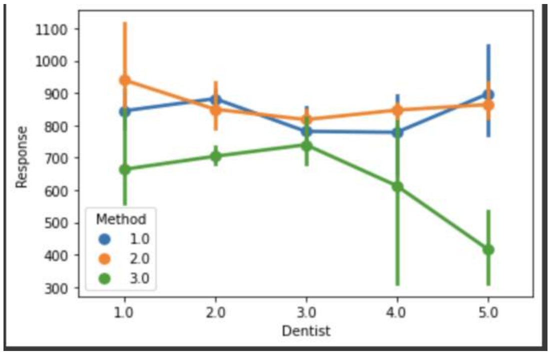
```
                         df         sum_sq       mean_sq          F    PR(>F)
C(Dentist)              4.0  106683.688889  26670.922222   3.899638  0.011484
C(Method)               2.0  148472.177778  74236.088889  10.854287  0.000284
C(Dentist):C(Method)    8.0  185941.377778  23242.672222   3.398383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



As per the Plot and table obtained, we can say that for Alloy2 Density and Method interaction is 0.09 which is greater than 0.05. Hence there is no statistical interaction



```
                          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
```



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?

From the graph above we could say that

Dentist 1, Method 1 and Alloy 1 could make high response

Dentist 5, Method 3 and Alloy 1 could make low response

Dentist 1, Method 2 and Alloy 2 could make high response

Dentist 5, Method 3 and Alloy 2 could also make low response