AS Project Report

Contents

List of Figures

List of Tables

AS PROJECT

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**

Table 1: Physiotherapist Dataset

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

Total Injured Players = 145
Total Players = 235

Probability (Suffered Injury) = 145/235 = 0.61702

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

Total forward players = 94
Total winger players = 29
P(Forward U Winger) = P(Forward) + P(Winger) – P(Forward ∩ Winger)
= (29/235)+(94/235)+0

= 0.52340

Probability that a player is a forward or a winger = 0.52340

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

Total number of players plays striker position and has a foot injury = 45
Total number for players = 235
P(Striker ∩ Foot injury) = 45/234

= 0.19148

Probability that a randomly chosen player plays in a striker position and has a foot injury is 0.19148

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

Total strikers who were injured = 45

Total injured player = 145

Probability (Injured player is striker) = 45/145 = 0.3103

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

Total injured player = 145

Total forward or attacking midfielder who were injured = 56 + 24 = 80

Probability (Injured player is either a forward or an attacking midfielder)= 80/145
= 0.55172

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:

Assume,

F = Fire

M = Mechanical Failure

H = Human Error

R = Radiation Leak

N = No Accident

Given Probabilities:

$P(R|F) = 20\% = 0.2$

$P(R|M) = 50\% = 0.5$

$P(R|F) = 10\% = 0.1$

$P(R \cap F) = 0.1\% = 0.0010$

$P(R \cap M) = 0.15\% = 0.0015$

$P(R \cap H) = 0.12\% = 0.0012$

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

$P(F) = P(R \cap F) / P(R|F) = 0.001/0.2 = 0.005$
Probability of fire is = 0.005

$P(M) = P(R \cap M) / P(R|M) = 0.0015/0.5 = 0.003$
Probability of mechanical failure is = 0.003

$P(H) = P(R \cap M) / P(R|M) = 0.0012/0.1 = 0.012$
Probability of Human error is 0.012

2.2 What is the probability of a radiation leak?

$P(N) = 1 – P(F) + P(M) + P(H) = 0.98$

$P(R/N) = 0$

$P(R \cap N) = P(R/N)\, P(N) = 0$

$P(R) = P(R \cap F) + P(R \cap M) + P(R \cap H) + P(R \cap N)$

$= 0.005+0.003+0.012+0$

$= 0.0037$

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:

Using conditional probability rule (Bayes' theorem),

- A Fire.

$P(F/R) = P(R \cap F)/P(R)$

$= 0.001/0.0037$

$P(F/R) = 0.2702$

Probability of Radiation leak by Fire is 0.2702

- A Mechanical Failure.

P(M/R) = P(R∩M)/P(M)

= 0.0015/0.0037

P(M/R) = 0.4054

Probability of Radiation leak by Mechanical Failure is 0.2702

- A Human Error.

P(H/R) = P(R∩H)/P(R)

= 0.0012/.0037

P(H/R) = 0.3243

Probability of Radiation leak 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. 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?

Given Values:

Mean => mu = 5
Standard Deviation => sigma = 1.5
x=3.17

By calculating Z score,
z = (x-mu)/sigma

Calculating proportion
p = stats.norm.cdf(z) = 0.1112

Therefore, 0.1112 proportion of the gunny bags have a breaking strength less than 3.17 kg per sq cm.

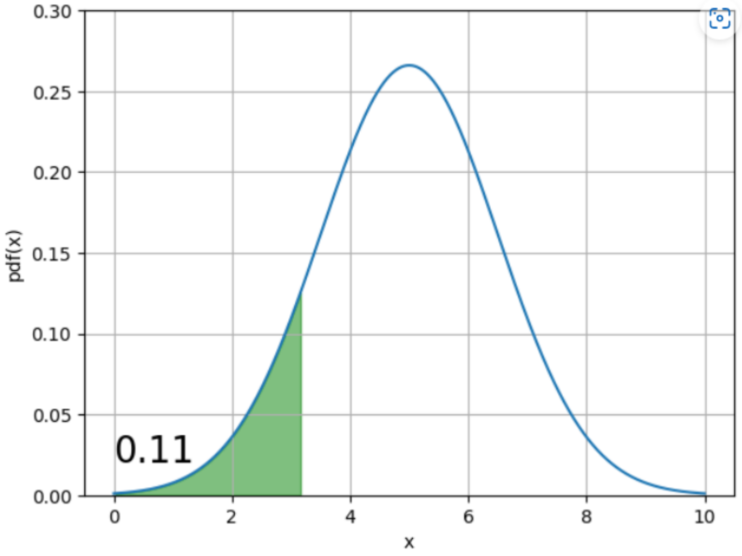

Figure 1: Plot for breaking strength less than 3.17 kg per sq cm.

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

Given Values:
Mean => mu = 5
Standard Deviation => sigma = 1.5
x=3.6

By calculating Z score,
z = (x-mu)/sigma

Calculating proportion
p = 1 - stats.norm.cdf(z) = 0.824

Therefore, 0.824 proportion of the gunny bags have a breaking strength at least 3.6 kg per sq cm.

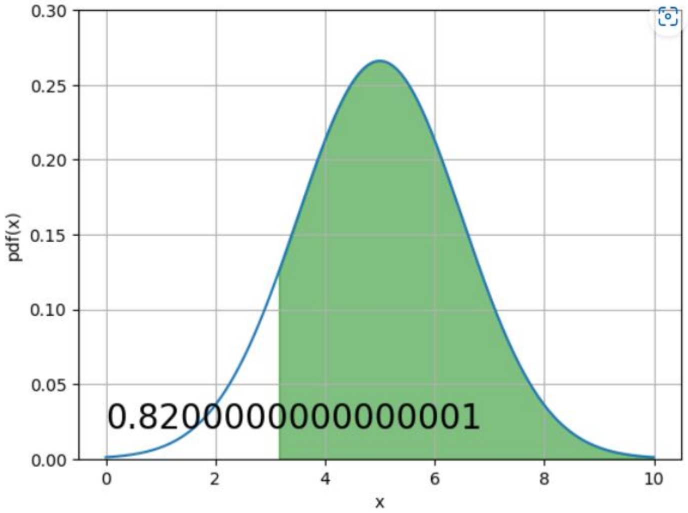

Figure 2: Plot for gunny bags have a breaking strength at least 3.6 kg per sq cm.

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

Given Values:
Mean => mu = 5
Standard Deviation => sigma = 1.5
x1=5
x2=5.5

Calculating proportion
p1 = norm.cdf(x1,mu,sigma)
p2 = norm.cdf(x2,mu,sigma)

p = p2 – p1 = 0.13

Therefore, 0.13 proportion of the gunny bags have a breaking strength between 5 and 5.5 kg per sq cm.

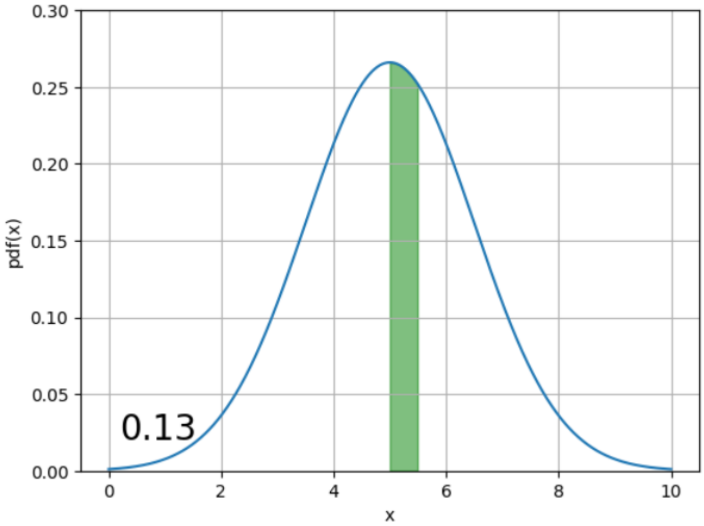

Figure 3: Plot for gunny bags have a breaking strength at least 3.6 kg per sq cm.

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

Given Values:
Mean => mu = 5
Standard Deviation => sigma = 1.5
x1=3
x2=7.5

Calculating proportion

p1 = norm.cdf(x1,mu,sigma)
p2 = norm.cdf(x2,mu,sigma)

p = 1 + p2 – p1 = 0.14

Therefore, 0.14 proportion of the gunny bags have a breaking strength NOT between 3 and 7.5 kg per sq cm

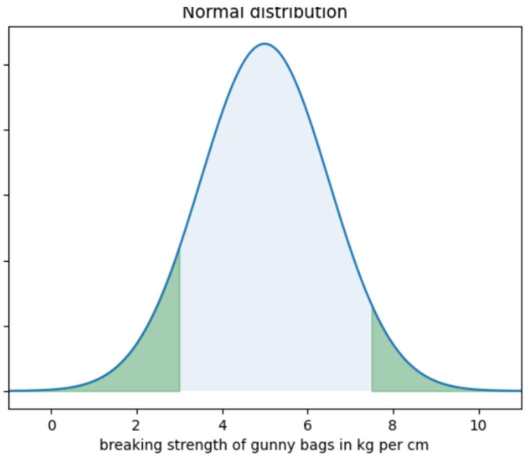

Figure 4: Plot for gunny bags have a breaking strength NOT between 3 and 7.5 kg per sq cm

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?

Given Values:

Mean => mu = 77
Standard Deviation => sigma = 8.5
x=85

By calculating Z score,

z = (x-mu)/sigma

Calculating Probability

p = stats.norm.cdf(z) = 0.8267

Therefore, 0.1112 probability that a randomly chosen student gets a grade below 85 on this exam is 0.8267.

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

Given Values:
Mean => mu = 77
Standard Deviation => sigma = 8.5
x1=65
x2=87

Calculating proportion
p1 = norm.cdf(x1,mu,sigma)
p2 = norm.cdf(x2,mu,sigma)

p = p2 – p1 = 0.80

Therefore, probability that a randomly selected student scores between 65 and 87 is 0.80

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

Given Value,

Z = 75%=0.75

75% of Z-score is 0.68

Z = (x − mu) / sigma
0.68 = (x – 77) / 8.5
(0.68) * (8.5) = x - 77
5.78 + 77 = x
x=82.78
The passing cut-off so that 75% of the students clear is 82.78

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?

H0 = Null Hypothesis

Ha = Alternative Hypothesis

H0 = Zingaro has reason to believe now that the unpolished stones may suitable for printing

Ha= Zingaro has reason to believe now that the unpolished stones may not be suitable for printing

```
u_mean = df['Unpolished '].mean()

u_std = df['Unpolished '].std()

n = df['Unpolished '].count()

t_stat = (u_mean - 150) / (u_std / (n**0.5))

p_value = stats.t.sf(abs(t_stat), n-1)

alpha_value = 0.05 # Level of significance
```

P value 4.171286997419643e-05 is less than Significance level 0.05 hence we reject the null hypothesis

we can conclude that the unpolished stones may not be suitable for printing.

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

H0 = Mean hardness of the polished and unpolished stones the same

Ha= Mean hardness of the polished and unpolished stones aren't same

```
p_mean = df['Treated and Polished'].mean()

u_mean = df['Unpolished '].mean()

p_std = df['Treated and Polished'].std()

u_std = df['Unpolished '].std()

n1 = df['Treated and Polished'].count()
```

n2 = df['Unpolished '].count()

sp = ((n1-1)*p_std**2 + (n2-1)*u_std**2) / (n1+n2-2)

t_stat = (p_mean - u_mean) / (sp * ((1/n1) + (1/n2)))**0.5

p_value = stats.t.sf(abs(t_stat), n1+n2-2)

P value 0.0007 is less than Significance level 0.05 hence we reject the null hypothesis

We can conclude that Mean hardness of the polished and unpolished stones aren't 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.

H0: difference = 5 (The mean result for before and after is same or the mean difference is 5)

Ha difference 5 (The mean result for after is greater than before or the mean difference is less than 5)

df1= pd.read_csv("Aquarius_gym.csv")

#Change to series

x=df1["Before"]

y=df1["After"]+5

The reason y has After+5 is because the hypotheses difference is 5. Python runs the test for mean difference to be zero. By adding 5 in each observation of After we are changing its mean by +5. So, the difference will be 0.

b=before, a=after, d=difference
Originally the mean difference is 5. µd=µb−µa=5

Increase the mean of after by 5,
µb – (ua+5) = 5

μb – μa = 0

#Running paired t-test

stats.ttest_rel(x, y, alternative="less")

TtestResult(statistic=-36.73038540665224, pvalue=8.7328623006375e-60, df=99)

From the test result output, the p-value is less than 0.05. So, we reject the null hypothesis at 5% level of significance.

We can conclude that the training will make a difference of more than 5

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.?

Factor A
H0 (Null hypothesis) = There is no difference in hardness of metal implant in dental cavities due to different methods.
Ha (Alternative Hypothesis) = There is a difference in hardness of metal implant due to temperature, method , alloy and work.

Factor B:
H0 : There is no difference in the means of alloy wheels

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

Yes, below assumptions are fulfilled.

The samples are drawn from different populations (Dentist 1 to 5)
The response variables of all population are continuous and normally distributed.
The variances all of the populations are almost 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?

H0 (Null Hypothesis) = There is no difference in hardness of metal implant in dental cavities due to different dentist.
Ha (Alternative Hypothesis) = There is a difference in hardness of metal implant due to dentist.

Formula = 'Response ~ C(Dentist) + C(Method)'
Model = ols(formula, df2).fit()
aov_table = anova_lm(model)

	df	sum_sq	mean_sq	F	PR(>F)
C(Dentist)	4.0	1.577946e+05	39448.638889	2.872456	2.784543e-02
C(Method)	2.0	5.934275e+05	296713.744444	21.605240	2.794351e-08
Residual	83.0	1.139874e+06	13733.415797	NaN	NaN

P value for Dentist (0.02784543) is less than 0.05 so, we can reject the null hypothesis here. Mean between different dentist is not same. **Implant hardness is depends on the Dentists.**

We don't have any methods to identify pairs from the Annova result.

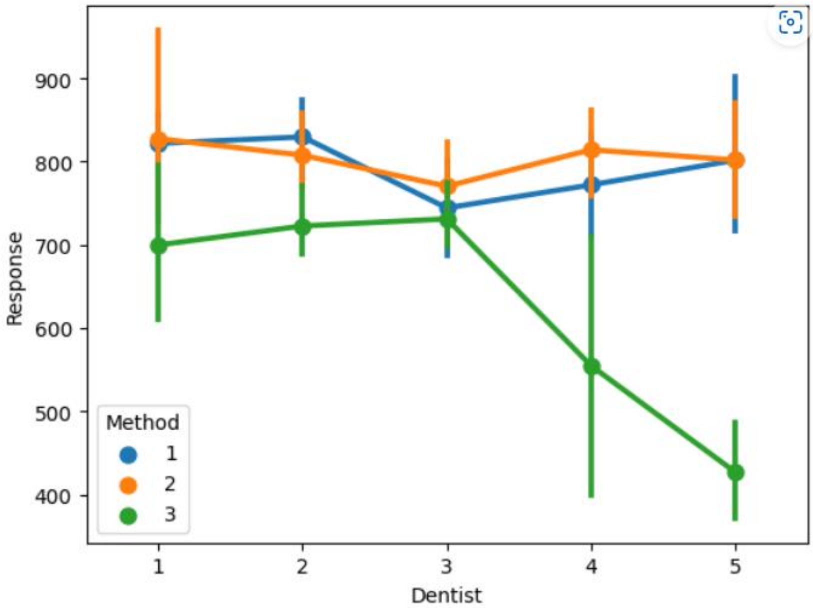

Figure 5: Point Plot Dentist vs Response with different Methods

But, from the above plot, we can say that below mentioned pair is having low number of response.

- Dentist 4 with Method 3
- Dentist 5 with Method 3

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?

H0 : There is no difference in the means of alloy wheels
Ha: There is a difference in the means between Alloy 1 and Alloy 2

formula='Response ~ C(Method) + C(Alloy)'
model=ols(formula, df2).fit()
aov_table=anova_lm(model)

	df	sum_sq	mean_sq	F	PR(>F)
C(Method)	2.0	5.934275e+05	296713.744444	21.409848	2.845744e-08
C(Alloy)	1.0	1.058155e+05	105815.511111	7.635285	6.999664e-03
Residual	86.0	1.191853e+06	13858.750646	NaN	NaN

P value is less than 0.05 so, we can reject the null hypothesis here.
There is a difference between Methods and Alloy types.

We don't have any methods to identify which pairs is causing issue from the Annova result.

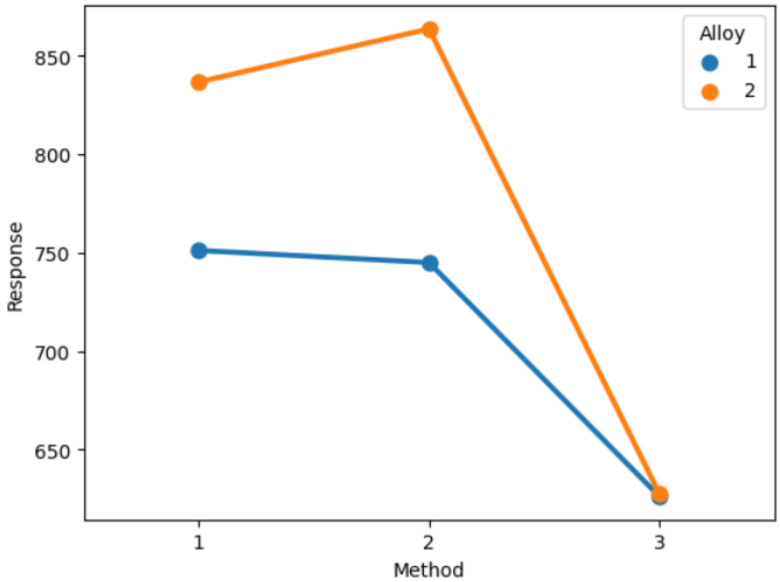

Figure 6: Point Plot Method vs Response with different Alloy Types

But, from the above plot, we might confirm that below mentioned pair is having low number of response.

- Alloy type 2 with Method 3
- Alloy type 1 with Method 3

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?

H0 : There is no difference in the means of Temperature
Ha: There is a difference in the means Temperature

formula='Response ~ C(Method) + C(Temp)'
model=ols(formula, df2).fit()

aov_table=anova_lm(model)
print(aov_table)

	df	sum_sq	mean_sq	F	PR(>F)
C(Method)	2.0	5.934275e+05	296713.744444	20.749383	4.588145e-08
C(Temp)	2.0	8.217802e+04	41089.011111	2.873381	6.201275e-02
Residual	85.0	1.215490e+06	14299.882876	NaN	NaN

P value is less than 0.05 so, we can reject the null hypothesis here.
There is a difference between Methods and Temperature types.

We don't have any methods to identify which pairs is causing issue from the Annova result.

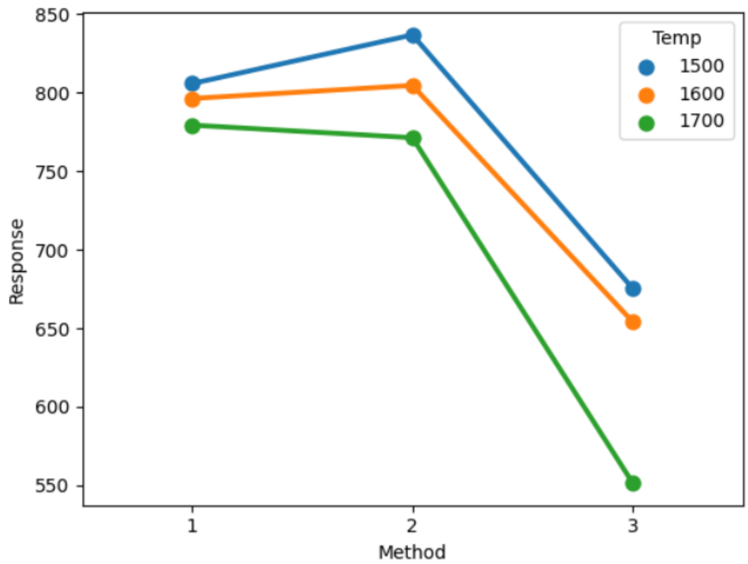

Figure 7: Point Plot Method vs Response with different Temperature

But, from the above plot, we might confirm that below mentioned pair is having low number of response.

- Method 3 with Temperature level 1700
- Method 3 with Temperature level 1600

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

Alloy Type 1:

Formula = 'Response ~ C(Dentist) + C(Method) + C(Dentist):C(Method)'
Model = ols(formula, df3).fit()
aov_table = anova_lm(model)

	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

- The relationship between Response and Dentist is being adjusted by method levels.
- There is strong evidence for Dentist is being a driver for Response.
- There is strong evidence for Method is being a driver for Response.
- And there is an interaction between Dentist and Method for alloy type 1.

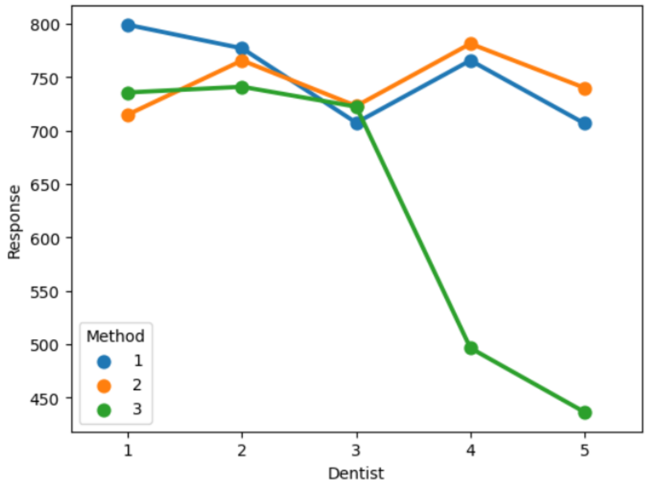

Figure 8: Point Plot Dentist vs Response with different Methods for Alloy type 1

Alloy Type 2:

Formula = 'Response ~ C(Dentist) + C(Method)'
model= ols(formula, df4).fit()
aov_table = anova_lm(model)

	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

- There is no strong evidence for Dentist is being a driver for Response.
- There is strong evidence for Method is being a driver for Response.
- There is no enough evidence to conclude that interaction between Dentist and Method.

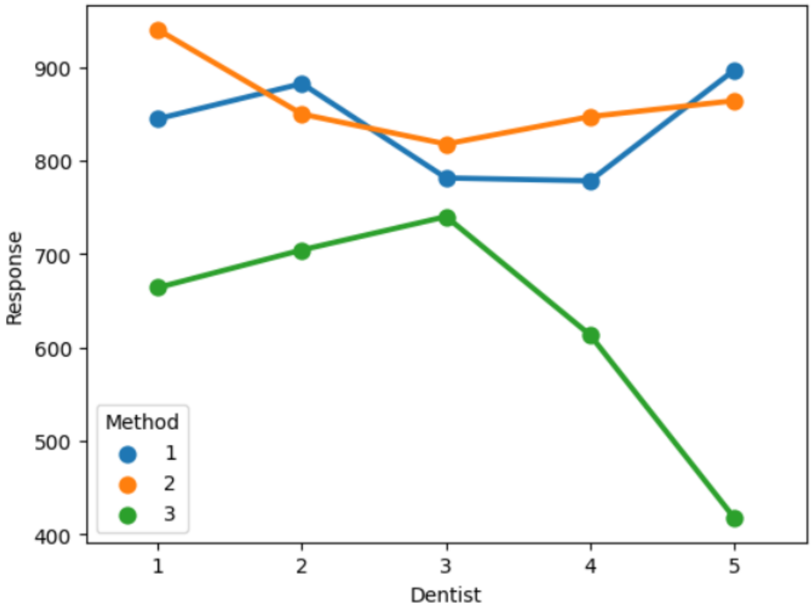

Figure 9: Point Plot Dentist vs Response with different Methods for Alloy type 2

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?

Alloy Type 1:

- There is an interaction between Dentist and Methods.
- Alloy type 1 having higher response when compare to Alloy type 2
- Method Type 1 and 2 having higher response when compared to Method 3.
- Dentist 4 with Method 3 are having low response rate.
- Dentist 5 with Method 3 are having low response rate.

Alloy Type 2:

- There is an interaction but not 95% between Dentist and Methods with respect to Alloy type 2.
- Dentist 4 with Method 3 are having low response rate.
- Dentist 5 with Method 3 are having low response rate.
- Method Type 1 and 2 having higher response when compared to Method 3