Business report

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

Based on the above data, answer the following questions.

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

P (Injured) = 145/235 = 0.6170

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

P (Forward OR Winger) = 94+29/235 = 0.523

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

P (Striker AND Injured) = 45/235 = 0.191

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

P (Striker|Injured) = 45/145 = 0.310

Problem 2

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

```
from scipy.stats import norm
mean = 5
std = 1.5
```

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

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

```
X = 3.17
z = (X-mean)/std
print("The value of z is", z)
p = round(stats.norm.cdf(z),4)
print('Proportion of the gunny bags have a breaking strength of less than 3.17 kg per sq cm is', p)
```

```
The value of z is -1.22
Proportion of the gunny bags have a breaking strength of less than 3.17 kg per sq cm is 0.1112
```

P (Breaking strength < 3.17) = 0.1112

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

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

```
X = 3.6
z = (X-mean)/std
print("The value of z is", z)
p = round(1 - stats.norm.cdf(z),4)
print('Proportion of the gunny bags have a breaking strength of at least 3.6 kg per sq cm is', p)
```

```
The value of z is -0.9333333333333332
Proportion of the gunny bags have a breaking strength of at least 3.6 kg per sq cm is 0.8247
```

P (Breaking strength > 3.6) = 0.8247

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

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

```
X1 = 5.5
z1 = (X1-mean)/std
print("The value of z1 is", z1)
X2 = 5
z2 = (X2-mean)/std
print("The value of z2 is", z2)
p = round(stats.norm.cdf(z1) - stats.norm.cdf(z2), 4)
print ('Proportion of the gunny bags have a breaking strength between 5 and 5.5 kg per sq cm', p)
```

```
The value of z1 is 0.3333333333333333
The value of z2 is 0.0
Proportion of the gunny bags have a breaking strength between 5 and 5.5 kg per sq cm 0.1306
```

P (5 < Breaking strength < 5.5) = 0.1306

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

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

```python
X1 = 7.5
z1 = (X1-mean)/std
print("The value of z1 is", z1)
X2 = 3
z2 = (X2-mean)/std
print("The value of z2 is", z2)
p = round((1- stats.norm.cdf(z1)) + stats.norm.cdf(z2), 4)
print ('Proportion of the gunny bags have a breaking strength NOT between 3 and 7.5 kg per sq cm', p)
```

```
The value of z1 is 1.6666666666666667
The value of z2 is -1.3333333333333333
Proportion of the gunny bags have a breaking strength NOT between 3 and 7.5 kg per sq cm 0.139
```

P ([Breaking strength < 3] or [Breaking strength > 7.5) = 0.139

Problem 3

Zingaro stone printing is a company that specializes in printing images or patterns on polished or unpolished stones. However, for the optimum level of printing of the image, the stone surface has to have a Brinell's hardness index of at least 150. Recently, Zingaro has received a batch of polished and unpolished stones from its clients. Use the data provided to answer the following (assuming a 5% significance level);

3.1 Zingaro has reason to believe that the unpolished stones may not be suitable for printing. Do you think Zingaro is justified in thinking so?

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

Answer:

ZingaroData = pd.read_csv('Zingaro_Company.csv')

In [33]:
```
ZingaroData = pd.read_csv('Zingaro_Company.csv')
ZingaroData
```
Out[33]:

	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
...	...	...
70	123.067611	142.293544
71	171.822218	140.124092
72	88.135994	141.393091
73	145.150397	131.370530
74	170.854823	144.502647

75 rows × 2 columns

In [4]:
```
ZingaroData.head()
```
Out[4]:

	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

In [5]:
```
ZingaroData.describe()
```
Out[5]:

	Unpolished	Treated and Polished
count	75.000000	75.000000
mean	134.110527	147.788117
std	33.041804	15.587355
min	48.406838	107.524167
25%	115.329753	138.268300
50%	135.597121	145.721322
75%	158.215098	157.373318
max	200.161313	192.272856

Step 1: Define null and alternative hypotheses

Step 1: Define null and alternative hypotheses

3.1 One-Sample t-Test for Unpolished Stones:

Hypotheses

Null Hypothesis (H0): The mean hardness of unpolished stones is at least 150. i.e >= 150.

Alternative Hypothesis (Ha): The mean hardness of unpolished stones is less than 150. i.e <150

Step 2: Decide the significance level¶

Here we select α = 0.05 and the population standard deviation is not known.

Step 3: Identify the test statistics

We have 1 samples and we do not know the population standard deviation.

The sample is a large sample, n > 30. So you use the $tSTAT$ test statistic for one sample test

```
import scipy.stats as stats

# Perform one-sample t-test
t_stat, p_value = stats.ttest_1samp(ZingaroData['Unpolished '], 150, alternative='less')

# Print the p-value
print("p-value:", p_value)
print("tstat-value:",t_stat)
```

```
p-value: 4.171286997419652e-05
tstat-value: -4.164629601426757
```

Since p_value is less than 0.05 level of significance we reject null hypothesis. i.e The mean hardness of unpolished stones is less than 150 not exactly or at least 150.Zingero is justified in concluding that the unpolished stones are not suitable for optimal level of printing.

3.2 Two-Sample t-Test for Mean Hardness (Polished vs. Unpolished Stones):

Hypotheses:

Null Hypothesis(H0): The mean hardness of polished stones is equal to the mean hardness of unpolished stones.

Alternative Hypothesis (Ha): The mean hardness of polished stones is not equal to the mean hardness of unpolished stones.

```
# Perform two-sample t-test
t_stat, p_value = stats.ttest_ind(ZingaroData['Treated and Polished'], ZingaroData['Unpolished '], equal_var=False)

# Print the p-value
print("p-value:", p_value)
print("tstat-value:",t_stat)
```

```
p-value: 0.001588379295584306
tstat-value: 3.242232050141406
```

Since p_value is less than 0.05 level of significance we reject null hypothesis. We accept that the mean hardness of polished stones is not equal to the mean hardness of unpolished stones.

Problem 4

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

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

Hypothesis test for Alloy 1

H0: The mean implant hardness is the same across different dentists with type 1 alloy.
Ha: Mean implant hardness is different for at least one pair of the dentists with type 1 alloy.

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

```
formula = 'Response ~ C(Dentist)'
model = ols(formula, data=Dental_Data[Dental_Data['Alloy'] == 1]).fit()
aov_table = anova_lm(model)
print(aov_table)
```

```
              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
```

Since p-value is greater than 0.05, we fail to reject the null hypothesis of equality. i.e The mean implant hardness is the same across different dentists with type 1 alloy.

Hypothesis test for Alloy 2

H0: The mean implant hardness is the same across different dentists with type 2 alloy.
Ha: Mean implant hardness is different for at least one pair of the dentists with type 2 alloy.

```
In [183]:
formula = 'Response ~ C(Dentist)'
model = ols(formula, data=Dental_Data[Dental_Data['Alloy'] == 2]).fit()
aov_table = anova_lm(model)
print(aov_table)
```

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

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

Since p-value is greater than 0.05, we fail to reject the null hypothesis of equality. i.e The mean implant hardness is the same across different dentists with type 2 alloy.

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

Hypothesis test for Alloy 1

H0: The mean implant hardness is the same across different methods with type 1 alloy.

Ha: Mean implant hardness is different for at least one pair of the methods with type 1 alloy.

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

```
formula = 'Response ~ C(Method)'
model = ols(formula,  data=Dental_Data[Dental_Data['Alloy'] == 1]).fit()
aov_table = anova_lm(model)
print(aov_table)
```

```
             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
```

```
formula = 'Response ~ C(Method)'
```

Since p-value is smaller than 0.05, null hypothesis is rejected. At least one method is different from the rest.

H0: The mean implant hardness is the same across different methods with type 2 alloy.

Ha: Mean implant hardness is different for at least one pair of the methods with type 2 alloy.

```
formula = 'Response ~ C(Method)'
model = ols(formula,  data=Dental_Data[Dental_Data['Alloy'] == 2]).fit()
aov_table = anova_lm(model)
print(aov_table)
```

```
             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
```

Since p-value is smaller than 0.05, null hypothesis is rejected. At least one method is different from the rest.

Comparsion

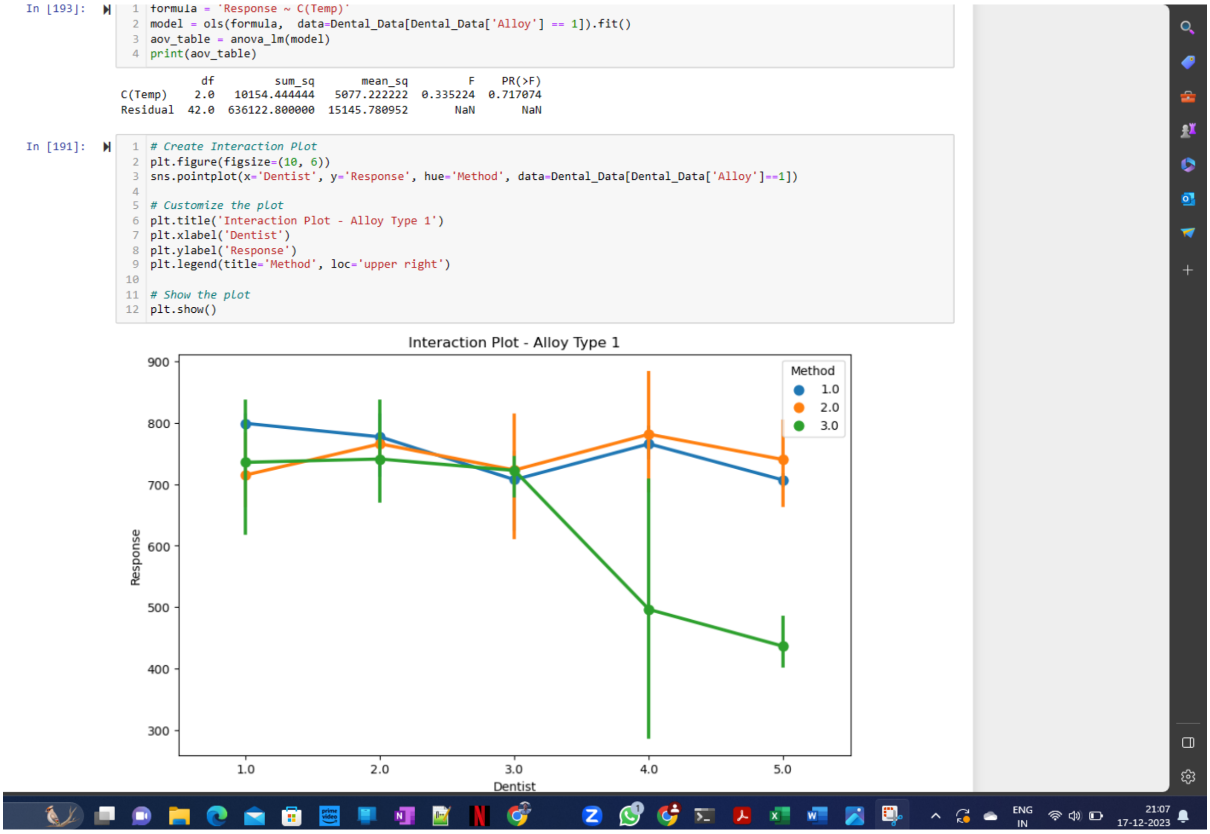
```
#import the required function
from statsmodels.stats.multicomp import pairwise_tukeyhsd

# perform multiple pairwise comparison (Tukey HSD)
m_comp = pairwise_tukeyhsd(endog = Dental_Data['Alloy'], groups = Dental_Data['Method'], alpha = 0.05)
print(m_comp)
```

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=============================================
group1 group2 meandiff p-adj lower upper reject
---------------------------------------------
   1.0    2.0      0.0   nan   nan   nan  False
   1.0    3.0      0.0   nan   nan   nan  False
   1.0    nan      nan   nan   nan   nan  False
   2.0    3.0      0.0   nan   nan   nan  False
   2.0    nan      nan   nan   nan   nan  False
   3.0    nan      nan   nan   nan   nan  False
---------------------------------------------
```

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

For type1 alloy

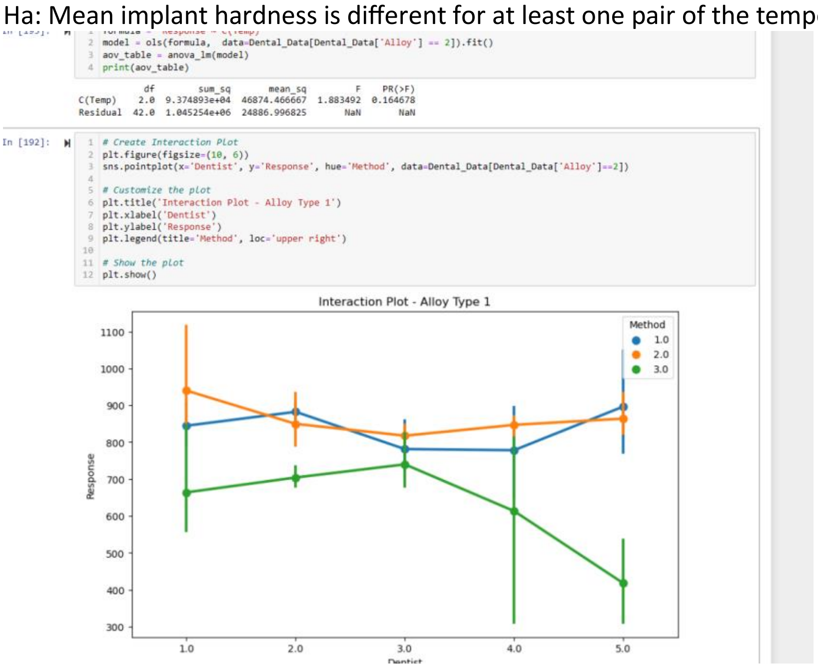
H0: The mean implant hardness is the same across different temperature with type 1 alloy.
Ha: Mean implant hardness is different for at least one pair of the temperature with type 1 alloy.



Since p-value is greater than 0.05, we fail to reject the null hypothesis. i.e The mean implant hardness is the same across different temperature with type 1 alloy

Type 2 alloy

H0: The mean implant hardness is the same across different temperature with type 2 alloy.
Ha: Mean implant hardness is different for at least one pair of the temperature with type 2 alloy.



Since p-value is greater than 0.05, we fail to reject the null hypothesis. i.e The mean implant hardness is the same across different temperature with type 2 alloy

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

Analysis for Alloy Type 1:

Null and Alternative Hypotheses:

H0 : There is no significant interaction between dentists and methods for Alloy Type 1.

Ha: There is a significant interaction between dentists and methods for Alloy Type 1.

Assumptions:

Independence: Observations within each cell (dentist-method combination) are independent.

Normality: The residuals are approximately normally distributed.

Homogeneity of Variances: Variances of the residuals are approximately equal across all combinations.

Hypothesis test:

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

```
formula = 'Response ~ C(Dentist) + C(Method)'
model = ols(formula, data=Dental_Data[Dental_Data['Alloy'] == 1]).fit()
anova_table = sm.stats.anova_lm(model)
print(anova_table)
```

```
               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
```

```
formula = 'Response ~ C(Dentist) * C(Method)'
model = ols(formula, data=Dental_Data[Dental_Data['Alloy'] == 1]).fit()
anova_table = sm.stats.anova_lm(model)
print(anova_table)
```

```
                        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 p_value is lesser for dentist and method interactions = 0.0067. That means there is a significant interaction between dentists and methods for Alloy Type 1.

Analysis for Alloy Type 2:

Null and Alternative Hypotheses:

H0 : There is no significant interaction between dentists and methods for Alloy Type 2.

Ha: There is a significant interaction between dentists and methods for Alloy Type 2.

```python
#type 2 alloy
formula = 'Response ~ C(Dentist) + C(Method)'
model = ols(formula, data=Dental_Data[Dental_Data['Alloy'] == 2]).fit()
anova_table = sm.stats.anova_lm(model)
print(anova_table)
```

```
              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
```

```python
formula = 'Response ~ C(Dentist) * C(Method)'
model = ols(formula, data=Dental_Data[Dental_Data['Alloy'] == 2]).fit()
anova_table = sm.stats.anova_lm(model)
print(anova_table)
```

```
                        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
```

The p_value is lesser for dentist and method interactions = 0.093. There is no significant interaction between dentists and methods for Alloy Type 2.