```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
import scipy.stats as stats
from statsmodels.formula.api import ols
from statsmodels.stats.anova import _get_covariance,anova_lm
from sklearn.linear_model import LinearRegression

from scipy.stats import f_oneway


data=pd.read_csv('/content/case_study_2.csv')
df=data.copy()


df.head()
```

	diet	preweight	weight6weeks	age
0	B	60	60.0	45
1	B	103	103.0	38
2	A	58	54.2	31
3	A	60	54.0	18
4	A	64	63.3	35

```
df['weight_loss']=df['weight6weeks']-df['preweight']

def categorize_age(age):
  if age>=18 and age<25:
    return "18-25"
  elif age>=25 and age<40:
    return "25-40"
  else:
    return "40+"

df['age_category']=df['age'].apply(categorize_age)
df.head()
```

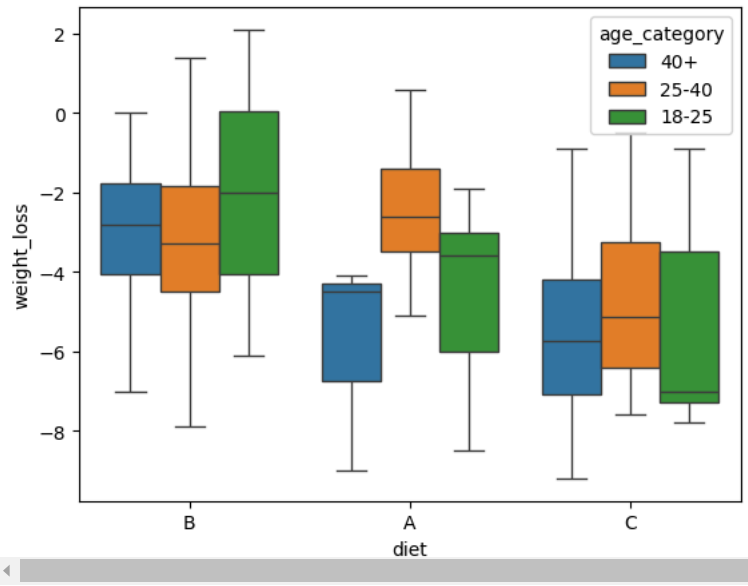
	diet	preweight	weight6weeks	age	weight_loss	age_category
0	B	60	60.0	45	0.0	40+
1	B	103	103.0	38	0.0	25-40
2	A	58	54.2	31	-3.8	25-40
3	A	60	54.0	18	-6.0	18-25
4	A	64	63.3	35	-0.7	25-40

```
df.age_category.value_counts()
```

age_category	count
25-40	47
40+	19
18-25	12

dtype: int64

```
sns.boxplot(x='diet',y='weight_loss',hue='age_category',data=df)
plt.show()
```



Hypothesis 1

```
#assuming Normality

w,p_value=stats.shapiro(df['weight_loss'])
print(round(p_value,3))
```

```
0.802
```

```
#Assuming homogeneity of variance

statistic,p_value =stats.levene(df[df['diet']=='A']['weight_loss'],
                                df[df['diet']=='B']['weight_loss'],
                                df[df['diet']=='C']['weight_loss'])

print(round(p_value,3))

#here we get the p_value greater than 5 so we are rejecting the null hypothesis
```

```
0.538
```

```
weightloss_diet_A=df[df['diet']=='A']['weight_loss']
weightloss_diet_B=df[df['diet']=='B']['weight_loss']
weightloss_diet_C=df[df['diet']=='C']['weight_loss']


test_stat,p_value=f_oneway(weightloss_diet_A,weightloss_diet_B,weightloss_diet_C)
print(round(p_value,3))
```

```
0.003
```

```
if p_value<0.05:
  print("Reject the null hypothesis")
else:
  print("Failed to reject Null Hypothesis")
```

```
Reject the null hypothesis
```

Hypothesis 2

```
#Assuming homogeneity of variance
```

```
statistic,p_value =stats.levene(df[df['age_category']=='18-25']['weight_loss'],
                                df[df['age_category']=='25-40']['weight_loss'],
                                df[df['age_category']=='40+']['weight_loss'])

print(round(p_value,3))
```

```
0.125
```

```
weightloss_Young=df[df['age_category']=='18-25']['weight_loss']
weightloss_Adult=df[df['age_category']=='25-40']['weight_loss']
weightloss_Elder=df[df['age_category']=='40+']['weight_loss']


test_stat,p_value=f_oneway(weightloss_Young,weightloss_Adult,weightloss_Elder)
print(round(p_value,3))
```

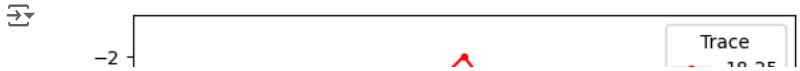
```
0.055
```

```
if p_value<0.05:
  print("Reject the null hypothesis")
else:
  print("Failed to reject Null Hypothesis")
```

```
Failed to reject Null Hypothesis
```

Hypothesis 3

```
from statsmodels.graphics.factorplots import interaction_plot
interaction_plot(np.array(df['diet']),np.array(df['age_category']),np.array(df['weight_loss']))
```




```
formula='weight_loss ~C(diet)+C(age_category)+C(diet):C(age_category)'
model=ols(formula,df).fit()
aov_table=anova_lm(model)
print(aov_table)
```

```
                          df      sum_sq    mean_sq         F    PR(>F)
C(diet)                  2.0   71.093689  35.546845  6.399140  0.002822
C(age_category)          2.0   17.498000   8.749000  1.574994  0.214359
C(diet):C(age_category)  4.0   29.390330   7.347582  1.322711  0.270226
Residual                69.0  383.290930   5.554941       NaN       NaN
```

```
row_name='C(diet):C(age_category)'
p_value=aov_table.loc[row_name,'PR(>F)']
print(round(p_value,3))

if p_value<0.05:
  print("p_value is less than the level of significance")
else:
  print("p_value is greater than the level of significance")
```

```
0.27
p_value is greater than the level of significance
```