```python
import numpy as np
import pandas as pd
import math
from scipy import stats
import matplotlib.pyplot        as     plt
%matplotlib inline
import seaborn as sns
import statsmodels.stats.api as sm
from scipy.stats import   norm
import scipy.stats as stats
import os
import scipy.stats as stats
import pylab
from statsmodels.formula.api import ols      # For n-way ANOVA
from statsmodels.stats.anova import _get_covariance,anova_lm # For n-way ANOVA
from statsmodels.graphics.factorplots import interaction_plot
```

Problem 1 Statement:

1a) A popular Two-Wheelers company claims that its best-selling model averages 18 km per litre of petrol. But recently a government agency used a sample of 80 two-wheelers of this model and finds the sample mean to be 16.8 km/litre. We also know from previous studies that the population standard deviation is 3 km/litre. Can we expect (within 2 standard errors) that we could select such a sample if indeed the population mean is 18 km/litre?

Solution

Given The Sample A has mean mu1 = 16.8km/lt, sample Size = n1=80.

The Population mean = μ =18km/lt with Standard Deviation = σ = 3.

Can we expect a sample whose population mean is 18km/lt & within 2 standard error?

Formulation of Hypothesis

Null Hypothesis H_0 : m1 = 18km/lt.

Alternative Hypothesis H_A : mu1 ≠ 18km/lt.

In [2]:
```python
mu=18
sigma = 3
n1=80
mu1 =16.8
```

In [3]:
```python
# For A sample
zstat=(mu1-mu)/(sigma/np.sqrt(n1))
zstat
```

Out[3]: -3.577708763999661

In [4]:
```python
zcrit=stats.norm.ppf(.01)
zcrit
```

Out[4]: -2.3263478740408408

In [5]:
```python
zstat < zcrit
```

Out[5]: True

In [6]:
```python
std_err=sigma/np.sqrt(n1)
std_err
```

Out[6]: 0.33541019662496846

In [7]:
```python
# For B Sample
#condition given is 2 std err is there for the new sample is std_err1
#so first we find sample size for the same
```

In [8]:
```python
std_err1=2*std_err
```

In [9]:
```python
n2=(sigma/std_err1)**2
round(n2,2)
```

Out[9]: 20.0

In [10]:
```python
# now we need to find the mean (mu2) of this sample size(n2) & it should be in the range of 2 standard error
```

In [11]:
```python
zcrit2=stats.norm.ppf(.05/2)
zcrit2
```

Out[11]: -1.9599639845400545

In [12]:
```python
mu2=((zcrit2*sigma)/np.sqrt(n2))+mu
mu2
```

Out[12]: 16.685216189135126

In [13]:
```python
zcrit3=stats.norm.ppf(1-.05/2)
zcrit3
```

Out[13]: 1.959963984540054

In [14]:
```python
mu3=((zcrit3*sigma)/np.sqrt(n2))+mu
mu3
```

Out[14]: 19.31478381086487

so the range is 16.685 to 19.314. it is possible to have it with sample size 20

Problem 2 Statement:

In [15]:
```python
#2a) Marriott International is an American multinational diversified hospitality company that manages and fr
#Top Level       Middle Level     Bottom Level
#    8               8                   5
#    7               7                   6
#    6               6                   7
#    7               9                   6
#    9              10                   7
#                    9
#                                              10

#Assuming level of significance as 0.05, formulate the null and alternative hypotheses and determine which t
```

Solution

The scores given by Top Level, Middle Level & bottom Level are in ordinal type.

So Anova one way test can't be performed as the data is ordinal type but there is a test which has below conditions are followed also in the given question.

1.It's applicable on 3 independent samples

2.Must have minimum 5 or more observations

3.The data samples size can be different

The test that can be conducted in this case is Kruskal -Wallis H Test.

Formulation of Null Ho & Alternative Ha Hypothesis

Null Hypothesis H_0 : Evaluations are same at all levels.

Alternative Hypothesis H_A : One or more evaluations are different.

In [16]:
```python
from Kruskals import Kruskals
```

In [17]:
```python
Top=[8,7,6,7,9]
Top
```

Out[17]: [8, 7, 6, 7, 9]

In [18]:
```python
Middle=[8,7,6,9,10,9]
Middle
```

Out[18]: [8, 7, 6, 9, 10, 9]

In [19]:
```python
Bottom=[5,6,7,6,7,8,10]
Bottom
```

Out[19]: [5, 6, 7, 6, 7, 8, 10]

In [45]:
```python
H,pval=stats.kruskal(Top,Middle,Bottom)
```

In [46]:
```python
print ("Kruskal -Wallis H Test p-value=", pval)

alpha_level = 0.05

if pval < alpha_level:
    print('We have enough evidence to reject the null hypothesis in favour of alternative hypothesis')
    print('We conclude that One or more evaluations are different.')
else:
    print('We do not have enough evidence to reject the null hypothesis in favour of alternative hypothesis')
    print('We conclude that Evaluations are same at all levels.')
```

```
Kruskal -Wallis H Test p-value= 0.35270982873167683
We do not have enough evidence to reject the null hypothesis in favour of alternative hypothesis
We conclude that Evaluations are same at all levels.
```

Decision Rule–

Since pvalue is > 0.05. so at 95% confidence level we fail to reject Null Hypothesis and conclude that evaluation are same at all levels.

In [ ]:

In [ ]:

Problem 3 Statement:

3b) The table shows the quantity of soaps sold for different brands at different locations, collected over 20 days. Conduct a two-way ANOVA with interaction at α = 5% to test the effects of brands, locations and interaction on sales.

The file (soaps.csv) is there.

In [22]:
```python
data = pd.read_csv ('SOAPS- Q3.csv')
```

In [23]:
```python
data.head()
```

Out[23]:

	Loc	Brand	Qty
0	1	X	20
1	2	X	20
2	1	X	16
3	2	X	21
4	1	X	24

In [24]:
```python
data.shape
```

Out[24]: (120, 3)

In [25]:
```python
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 120 entries, 0 to 119
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Loc     120 non-null    int64
 1   Brand   120 non-null    object
 2   Qty     120 non-null    int64
dtypes: int64(2), object(1)
memory usage: 2.9+ KB
```

In [26]:
```python
data.describe()
```

Out[26]:

	Loc	Qty
count	120.000000	120.000000
mean	1.500000	24.908333
std	0.502096	5.130230
min	1.000000	10.000000
25%	1.000000	21.750000
50%	1.500000	25.000000
75%	2.000000	28.000000
max	2.000000	39.000000

In [27]:
```python
data.Loc = pd.Categorical(data.Loc)
```

In [28]:
```python
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 120 entries, 0 to 119
[illegible]
```

In [ ]:
```python
data["Loc"].value_counts()
```

[illegible]

In [ ]:
```python
data["Brand"].value_counts()
```

[illegible]

Formulation of Hypothesis

Null Hypothesis H_0 : The Qty sold variable with respect to Location is equal.

Alternative Hypothesis H_A : At least one of the Qty sold variable with respect to Location is unequal.

Null Hypothesis H_0 : The Qty sold variable with respect to each Brand is equal.

Alternative Hypothesis H_A : At least one of the Qty sold variable with respect to Brand is unequal.

Null Hypothesis H_0 : The Qty sold variable with respect to Location and Brand and their interaction is equal.

Alternative Hypothesis H_A : At least one of the Qty sold variable with respect to Location and Brand and their interaction is unequal.

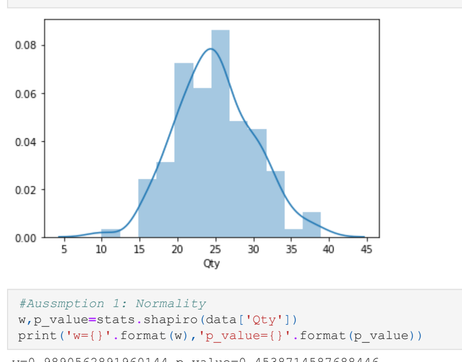
Assumptions for ANOVA

1. All populations under consideration have normal distribution

2. All populations under consideration have equal variances.

3. The sample is a random sample, i.e. the observations are collected independently of each other.

In [ ]:
```python
sns.distplot(data["Qty"])
plt.show()
```

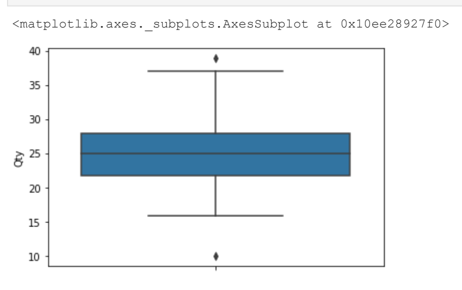


In [ ]:
```python
[illegible]
```

[illegible]

Since p value of the test is very large, so the response follows the normal distribution.

In [ ]:
```python
sns.boxplot(data["Qty"])
plt.show()
```



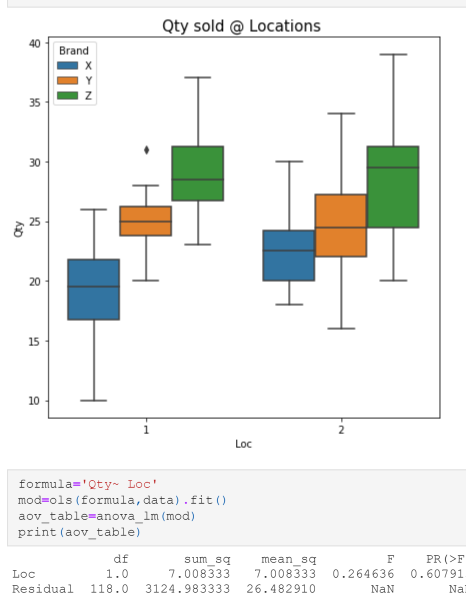
Checking the impact of Location on qty sold

Formulation of Hypothesis

Null Hypothesis H_0 : The Qty sold variable with respect to Location is equal.

Alternative Hypothesis H_A : At least one of the Qty sold variable with respect to Location is unequal.

In [ ]:
```python
[illegible]
```



In [ ]:
```python
[illegible]
```

[illegible]

Since F value is small and p is more than alpha we fail to reject null hypothesis. so qty sold is same at all locations

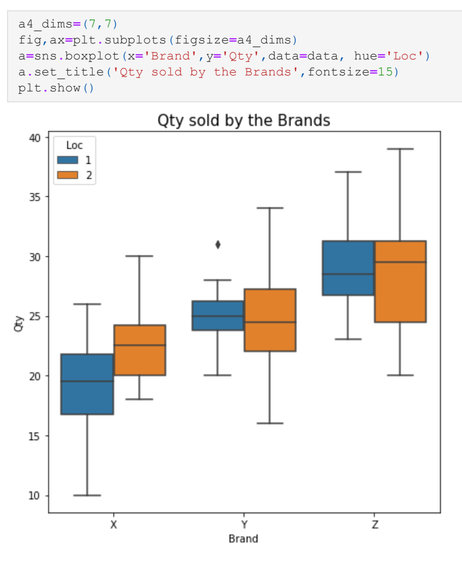
Checking the impact of brand on the qty sold

Formulation of Hypothesis

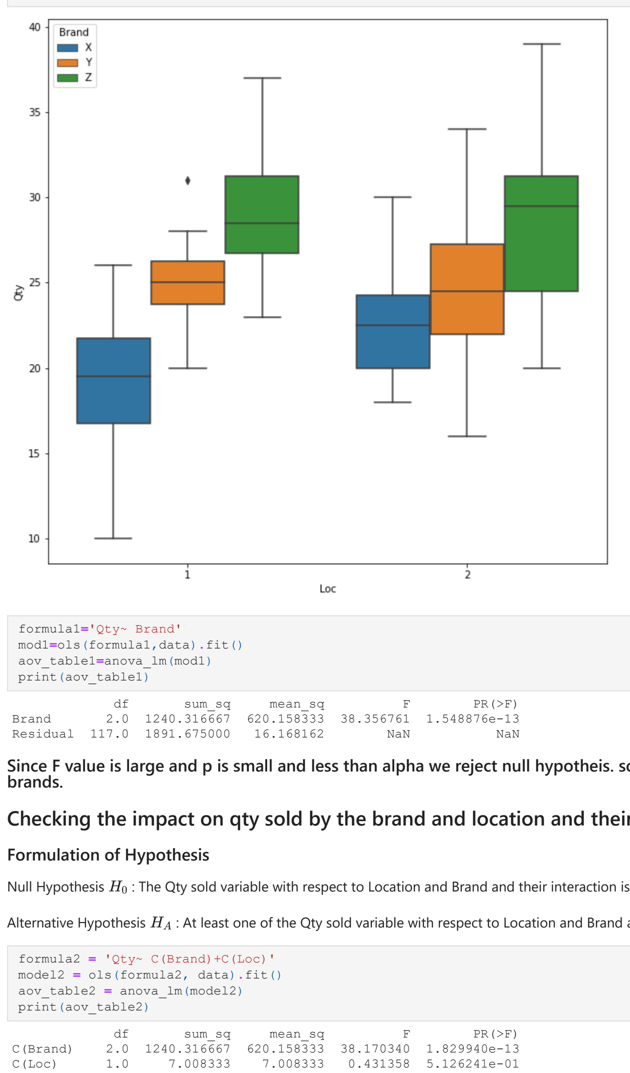
Null Hypothesis H_0 : The Qty sold variable with respect to each Brand is equal.

Alternative Hypothesis H_A : At least one of the Qty sold variable with respect to Brand is unequal.

In [ ]:
```python
[illegible]
```



In [ ]:
```python
[illegible]
```



In [38]:
```python
formula1='Qty~ Brand'
mod1=ols(formula1,data).fit()
aov_table1=anova_lm(mod1)
print(aov_table1)
```

```
            df       sum_sq     mean_sq          F        PR(>F)
Brand      2.0  1240.316667  620.158333  38.356761  1.548876e-13
Residual 117.0  1891.675000   16.168162        NaN           NaN
```

Since F value is large and p is small and less than alpha we reject null hypotheis. so qty sold is impacted by the brands.

Checking the impact on qty sold by the brand and location and their interaction

Formulation of Hypothesis

Null Hypothesis H_0 : The Qty sold variable with respect to Location and Brand and their interaction is equal.

Alternative Hypothesis H_A : At least one of the Qty sold variable with respect to Location and Brand and their interaction is unequal.

In [39]:
```python
formula2 = 'Qty~ C(Brand)+C(Loc)'
model2 = ols(formula2, data).fit()
aov_table2 = anova_lm(model2)
print(aov_table2)
```

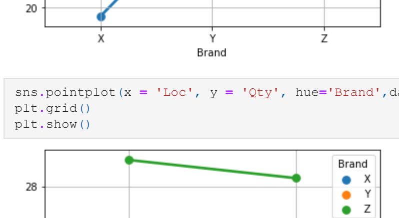
```
             df       sum_sq     mean_sq          F        PR(>F)
C(Brand)    2.0  1240.316667  620.158333  38.170340  1.829940e-13
C(Loc)      1.0     7.008333    7.008333   0.431358  5.126241e-01
Residual  116.0  1884.666667   16.247126        NaN           NaN
```

In [40]:
```python
formula3 = 'Qty~ C(Brand)+C(Loc)+ C(Brand):C(Loc)'
model3 = ols(formula3, data).fit()
aov_table3 = anova_lm(model3)
print(aov_table3)
```

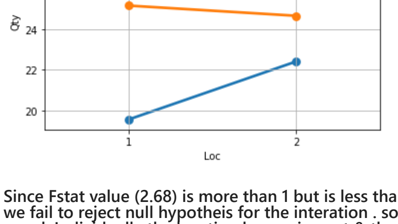
```
                    df       sum_sq     mean_sq          F        PR(>F)
C(Brand)           2.0  1240.316667  620.158333  39.279968  1.055160e-13
C(Loc)             1.0     7.008333    7.008333   0.443898  5.065930e-01
C(Brand):C(Loc)    2.0    84.816667   42.408333   2.686085  7.246036e-02
Residual         114.0  1799.850000   15.788158        NaN           NaN
```

Brand impact is visible in the below plot too.

In [43]:
```python
sns.pointplot(x = 'Brand', y = 'Qty', hue='Loc',data=data,ci= None)
plt.grid()
plt.show()
```



In [44]:
```python
sns.pointplot(x = 'Loc', y = 'Qty', hue='Brand',data=data,ci= None)
plt.grid()
plt.show()
```



Since Fstat value (2.68) is more than 1 but is less than F crit value(3.07)and p is 0.07 and more than alpha (0.05) we fail to reject null hypotheis for the interation . so qty sold with respect to location and their interation is equal. Individually the location has no impact & the brand has great impact on the qty sold..

In [ ]: