In [14]:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
from scipy.stats import ttest_1samp, ttest_ind
from scipy.stats import variation
```

In [15]:

```
wh=pd.read_excel('D:\\ANALYTICS\\GREAT LEARNING\\7.Statistical Method for Decisoin Making-Week-4\\
Wholesale customers data-1.xlsx',
                 sheet_name='Wholesale customers data')
```

Problem 1:

A wholesale distributor operating in different regions of Portugal has information on annual spending of several items in their stores across different regions and channels. The data (Wholesale customers data.xlsx) consists of 440 large retailers' annual spending on 6 different varieties of products in 3 different regions (Lisbon, Oporto, Other) and across different sales channel (Hotel/Restaurant/Café HoReCa, Retail).

In [11]:

```
#1.1. Use methods of descriptive statistics to summarize data.
#Which Region and which Channel seems to spend more?
#Which Region and which Channel seems to spend less?
wh.head()
```

Out[11]:

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
0	1	Retail	Other	12669	9656	7561	214	2674	1338
1	2	Retail	Other	7057	9810	9568	1762	3293	1776
2	3	Retail	Other	6353	8808	7684	2405	3516	7844
3	4	Hotel	Other	13265	1196	4221	6404	507	1788
4	5	Retail	Other	22615	5410	7198	3915	1777	5185

We could see 2 Categorical(Channel and Region) and 7 Conitnuous variables

In [10]:

```
wh.tail()
```

Out[10]:

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
435	436	Hotel	Other	29703	12051	16027	13135	182	2204
436	437	Hotel	Other	39228	1431	764	4510	93	2346
437	438	Retail	Other	14531	15488	30243	437	14841	1867
438	439	Hotel	Other	10290	1981	2232	1038	168	2125
439	440	Hotel	Other	2787	1698	2510	65	477	52

....

Descriptive Statistics :

In [11]:

```
wh.describe(include='all')
```

Out[11]:

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	D
count	440.000000	440	440	440.000000	440.000000	440.000000	440.000000	440.000000	44
unique	NaN	2	3	NaN	NaN	NaN	NaN	NaN	Na
top	NaN	Hotel	Other	NaN	NaN	NaN	NaN	NaN	Na
freq	NaN	298	316	NaN	NaN	NaN	NaN	NaN	Na
mean	220.500000	NaN	NaN	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	15
std	127.161315	NaN	NaN	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	28
min	1.000000	NaN	NaN	3.000000	55.000000	3.000000	25.000000	3.000000	3.(
25%	110.750000	NaN	NaN	3127.750000	1533.000000	2153.000000	742.250000	256.750000	40
50%	220.500000	NaN	NaN	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	96
75%	330.250000	NaN	NaN	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	18
max	440.000000	NaN	NaN	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47

There are 2 and 3 unique values for Channel and Region respectively.

We could see Fresh shows highest standard deviation among the continuous variables, value being 12647.32

Mean for Fresh=12000.29

Mean for Milk=5796.26.

Mean for Grocery=7951.27.

Mean for Frozen=3071.93.

Mean for Detergents paper=2881.49.

Mean for Delicatessen=1524.87.

....

In [13]:

```
wh.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 440 entries, 0 to 439
Data columns (total 9 columns):
Buyer/Spender       440 non-null int64
Channel             440 non-null object
Region              440 non-null object
Fresh               440 non-null int64
Milk                440 non-null int64
Grocery             440 non-null int64
Frozen              440 non-null int64
Detergents_Paper    440 non-null int64
Delicatessen        440 non-null int64
dtypes: int64(7), object(2)
memory usage: 31.0+ KB
```

There are no null value across the variables in the dataset

....

In [44]:

```
wh1=wh
wh1['Total']=wh['Fresh']+wh['Milk']+wh['Grocery']+wh['Frozen']+wh['Detergents_Paper']+wh['Delicates
sen']
```

In [45]:

```
wh1.head()
```

Out[45]:

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
0	1	Retail	Other	12669	9656	7561	214	2674	1338	34112
1	2	Retail	Other	7057	9810	9568	1762	3293	1776	33266
2	3	Retail	Other	6353	8808	7684	2405	3516	7844	36610
3	4	Hotel	Other	13265	1196	4221	6404	507	1788	27381
4	5	Retail	Other	22615	5410	7198	3915	1777	5185	46100

In [46]:

```
#Region and channel spending more
wh1.groupby(['Region','Channel'])['Total'].sum().reset_index(name='Total').sort_values(by='Total',
ascending=False).head(1)
```

Out[46]:

	Region	Channel	Total
4	Other	Hotel	5742077

Other Region of Channel Hotel spends more

....

In [47]:

```
#Region and channel spending less
wh1.groupby(['Region','Channel'])['Total'].sum().reset_index(name='Total').sort_values(by='Total')
.head(1)
```

Out[47]:

	Region	Channel	Total
2	Oporto	Hotel	719150

Whereas Oporto Region via Channel Hotel spends less

....

In [37]:

```
#1.2. There are 6 different varieties of items are considered.
#Do all varieties show similar behaviour across Region and Channel?
```

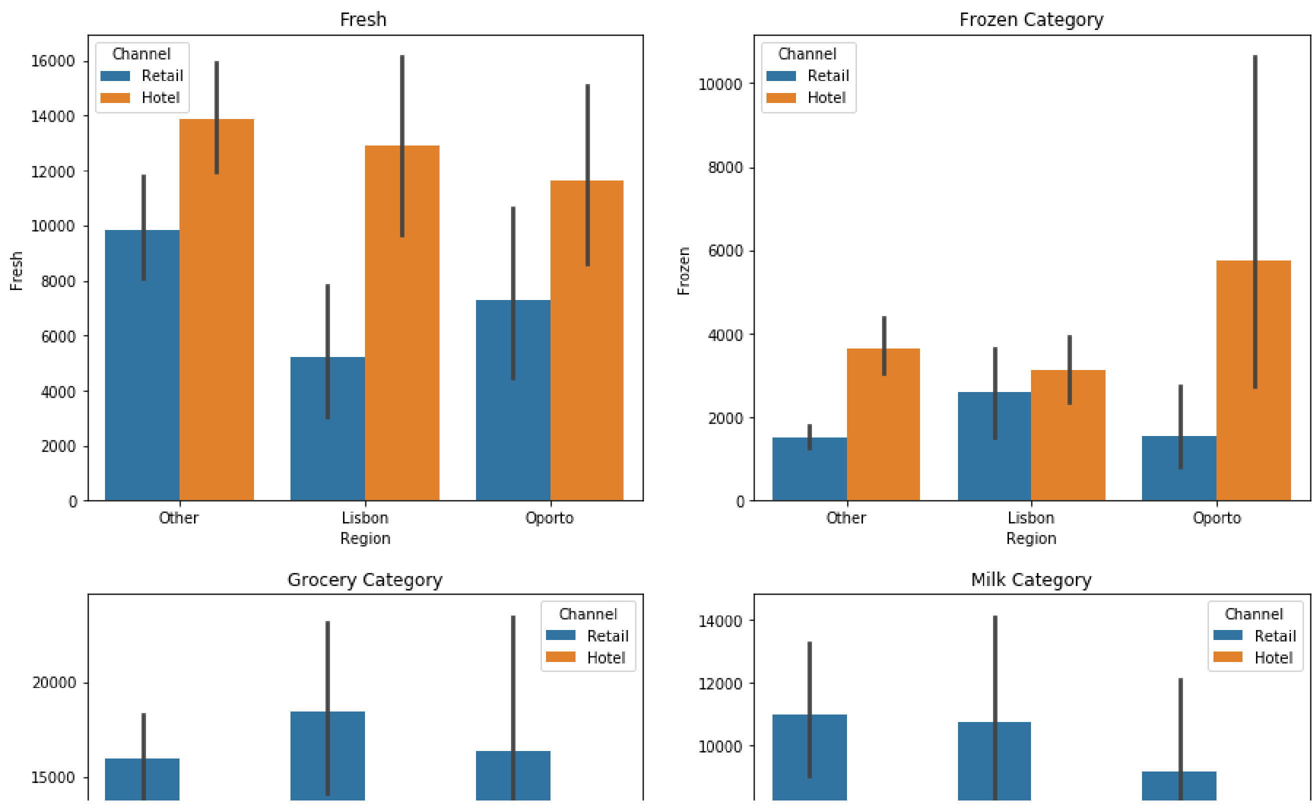
```
#Do all varieties show similar behaviour across Region and Channel?
pd.pivot_table(wh,index=['Region','Channel'])
```

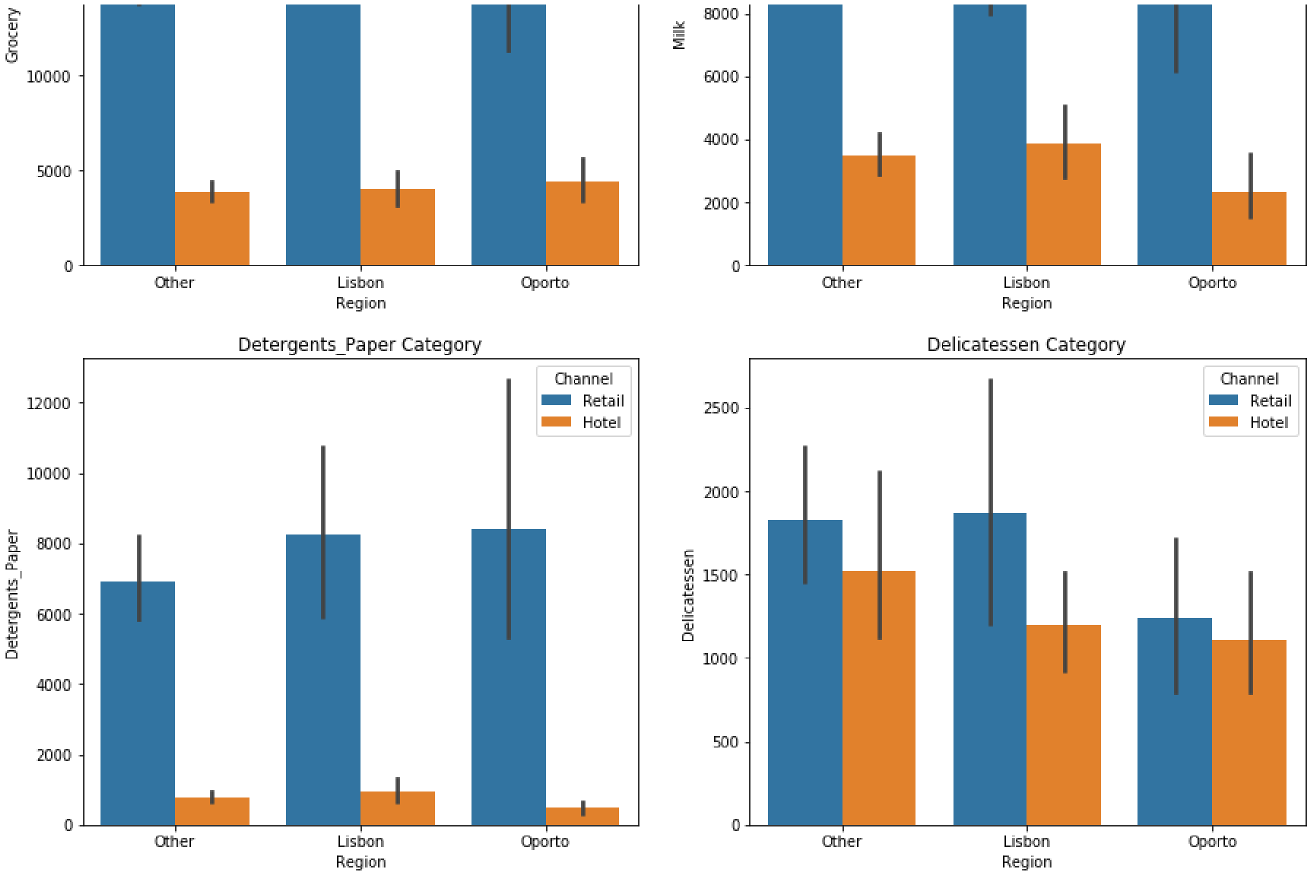
Out[37]:

		Buyer/Spender	Delicatessen	Detergents_Paper	Fresh	Frozen	Grocery	Milk
Region	**Channel**							
Lisbon	**Hotel**	237.728814	1197.152542	950.525424	12902.254237	3127.322034	4026.135593	3870.203390
	Retail	226.055556	1871.944444	8225.277778	5200.000000	2584.111111	18471.944444	10784.000000
Oporto	**Hotel**	321.000000	1105.892857	482.714286	11650.535714	5745.035714	4395.500000	2304.250000
	Retail	311.105263	1239.000000	8410.263158	7289.789474	1540.578947	16326.315789	9190.789474
Other	**Hotel**	227.582938	1518.284360	786.682464	13878.052133	3656.900474	3886.734597	3486.981043
	Retail	152.438095	1826.209524	6899.238095	9831.504762	1513.200000	15953.809524	10981.009524

In [87]:

```
fig,ax=plt.subplots(nrows=3,ncols=2,sharey=False,sharex=False,figsize=(15,20))
sns.barplot(x='Region',y='Fresh',data=wh,hue='Channel',ax=ax[0,0])
ax[0,0].title.set_text('Fresh')


sns.barplot(x='Region',y='Frozen',data=wh,hue='Channel',ax=ax[0,1])
ax[0,1].title.set_text('Frozen Category')


sns.barplot(x='Region',y='Grocery',data=wh,hue='Channel',ax=ax[1,0])
ax[1,0].title.set_text('Grocery Category')


sns.barplot(x='Region',y='Milk',data=wh,hue='Channel',ax=ax[1,1])
ax[1,1].title.set_text('Milk Category')


sns.barplot(x='Region',y='Detergents_Paper',data=wh,hue='Channel',ax=ax[2,0])
ax[2,0].title.set_text('Detergents_Paper Category')


sns.barplot(x='Region',y='Delicatessen',data=wh,hue='Channel',ax=ax[2,1])
ax[2,1].title.set_text('Delicatessen Category')
```





We could see Detergents_Paper,Delicatessen,Grocery and Milk categories has highest spending on Retail Channel on all Regions

while Fresh and Frozen categories has highest spending Hotel Channel on all Regions

There are variation in spending on each varieties across Region and Channel

....

In [62]:

```
#1.3. On the basis of the descriptive measure of variability, which item shows the most
inconsistent behaviour?
#Which items shows the least inconsistent behaviour?
wh.describe()
```

Out[62]:

	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
mean	220.500000	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	127.161315	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	1.000000	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	110.750000	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000
50%	220.500000	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000
75%	330.250000	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000
max	440.000000	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000

In [88]:

```
#standard deviation
wh.std()
```

Out[88]:

```
Buyer/Spender         127.161315
Fresh               12647.328865
Milk                 7380.377175
Grocery              9503.162829
Frozen               4854.673333
Detergents_Paper     4767.854448
Delicatessen         2820.105937
dtype: float64
```

In [89]:

```
#IQR
Q3=wh.quantile(0.75)
Q1=wh.quantile(0.25)
Q3-Q1
```

Out[89]:

```
Buyer/Spender         219.50
Fresh               13806.00
Milk                 5657.25
Grocery              8502.75
Frozen               2812.00
Detergents_Paper     3665.25
Delicatessen         1412.00
dtype: float64
```

In [90]:

```
#Range
mx=wh.max(numeric_only=True)
mn=wh.min(numeric_only=True)
mx-mn
```

Out[90]:

```
Buyer/Spender         439
Fresh               112148
Milk                 73443
Grocery              92777
Frozen               60844
Detergents_Paper     40824
Delicatessen         47940
dtype: int64
```

From the above measure of variablity we could see Fresh Category showing highest amount of variation

and Delicatessen shows least amount of variation

....

In [92]:

```
#1.4. Are there any outliers in the data?
IQR=Q3-Q1
out=((wh1.iloc[:,3:]<(Q1-1.5*IQR)) | (wh1.iloc[:,3:]>(Q3+1.5*IQR))).sum()
out
```

Out[92]:

```
Buyer/Spender        0
Delicatessen        27
Detergents_Paper    30
Fresh               20
Frozen              43
Grocery             24
Milk                28
Total               20
dtype: int64
```

All categories show various number of outliers,yes outliers are present in data

In [91]:

```
#1.5. On the basis of this report, what are the recommendations?
```

We were able to see Oporto region of Channel Hotel has less customer spending, we can improve by increasing quality sales on

Fresh and Frozen type Category which are the 2 categories dominating Hotel spending

Fresh Category shows highest variation,variability can be reduced on concentrating on increasing the sales on Retail channel as well

In [3]:

```
import pandas as pd
import numpy as np
import scipy.stats as stats
import statsmodels.api as sm
```

In [4]:

```
sv=pd.read_csv('D:\\ANALYTICS\\GREAT LEARNING\\7.Statistical Method for Decisoin Making-Week-4\\Su
rvey-1.csv')
```

Problem 2

The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students

that attend CMSU. CMSU creates and distributes a survey of 14 questions and receives responses from 62 undergraduates (stored in

the Survey.csv file).

In [5]:

```
sv.head()
```

Out[5]:

	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Comput
0	1	Female	20	Junior	Other	Yes	2.9	Full-Time	50.0	1	3	350	Laptop
1	2	Male	23	Senior	Management	Yes	3.6	Part-Time	25.0	1	4	360	Laptop
2	3	Male	21	Junior	Other	Yes	2.5	Part-Time	45.0	2	4	600	Laptop
3	4	Male	21	Junior	CIS	Yes	2.5	Full-Time	40.0	4	6	600	Laptop
4	5	Male	23	Senior	Other	Undecided	2.8	Unemployed	40.0	2	4	500	Laptop

In [32]:

```
sv.describe(include='all')
```

Out[32]:

	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking
count	62.000000	62	62.000000	62	62	62	62.000000	62	62.000000	62.000000
unique	NaN	2	NaN	3	8	3	NaN	3	NaN	NaN
top	NaN	Female	NaN	Senior	Retailing/Marketing	Yes	NaN	Part-Time	NaN	NaN
freq	NaN	33	NaN	31	14	28	NaN	43	NaN	NaN
mean	31.500000	NaN	21.129032	NaN	NaN	NaN	3.129032	NaN	48.548387	1.516129
std	18.041619	NaN	1.431311	NaN	NaN	NaN	0.377388	NaN	12.080912	0.844305
min	1.000000	NaN	18.000000	NaN	NaN	NaN	2.300000	NaN	25.000000	0.000000
25%	16.250000	NaN	20.000000	NaN	NaN	NaN	2.900000	NaN	40.000000	1.000000
50%	31.500000	NaN	21.000000	NaN	NaN	NaN	3.150000	NaN	50.000000	1.000000
75%	46.750000	NaN	22.000000	NaN	NaN	NaN	3.400000	NaN	55.000000	2.000000

	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking
max	62.000000	NaN	26.000000	NaN	NaN	NaN	3.900000	NaN	80.000000	4.000000

Descriptive statistics:

There are 6 categorical variables while rest are continuous

In [7]:

```
sv.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62 entries, 0 to 61
Data columns (total 14 columns):
ID                   62 non-null int64
Gender               62 non-null object
Age                  62 non-null int64
Class                62 non-null object
Major                62 non-null object
Grad Intention       62 non-null object
GPA                  62 non-null float64
Employment           62 non-null object
Salary               62 non-null float64
Social Networking    62 non-null int64
Satisfaction         62 non-null int64
Spending             62 non-null int64
Computer             62 non-null object
Text Messages        62 non-null int64
dtypes: float64(2), int64(6), object(6)
memory usage: 6.9+ KB
```

We could see there are no null values

In [8]:

```
#• 2.1. For this data, construct the following contingency tables (Keep Gender as row variable)
#2.1.1. Gender and Major
#2.1.2. Gender and Grad Intention
#2.1.3. Gender and Employment
#2.1.4. Gender and Computer

pd.set_option('display.max_columns',15)
gmt=pd.crosstab(sv['Gender'],sv['Major'],margins=True,margins_name='Total')
gmgi=pd.crosstab(sv['Gender'],sv['Grad Intention'],margins=True,margins_name='Total')
gme=pd.crosstab(sv['Gender'],sv['Employment'],margins=True,margins_name='Total')
gmc=pd.crosstab(sv['Gender'],sv['Computer'],margins=True,margins_name='Total')
print('Contingency table Gender vs Major')
print(gmt)
print(' ')
print('Contingency table Gender vs Grad intention')
print(gmgi)
print(' ')
print('Contingency table Gender vs Employment')
print(gme)
print(' ')
print('Contingency table Gender vs Computer')
print(gmc)
print(' ')
```

```
Contingency table Gender vs Major
Major   Accounting  CIS  Economics/Finance  International Business  \
Gender
Female           3    3                  7                       4
Male             4    1                  4                       2
Total            7    4                 11                       6

Major   Management  Other  Retailing/Marketing  Undecided  Total
Gender
Female           4      3                    9          0     33
Male             6      4                    5          3     29
Total           10      7                   14          3     62
```

```
Contingency table Gender vs Grad intention
Grad Intention  No  Undecided  Yes  Total
Gender
Female           9         13   11     33
Male             3          9   17     29
Total           12         22   28     62

Contingency table Gender vs Employment
Employment  Full-Time  Part-Time  Unemployed  Total
Gender
Female              3         24           6     33
Male                7         19           3     29
Total              10         43           9     62

Contingency table Gender vs Computer
Computer  Desktop  Laptop  Tablet  Total
Gender
Female          2      29       2     33
Male            3      26       0     29
Total           5      55       2     62
```

In [24]:

```
#2.2.1. What is the probability that a randomly selected CMSU student will be male?
#What is the probability that a randomly selected CMSU student will be female?
gmt
```

Out[24]:

Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	Total
Gender									
Female	3	3	7	4	4	3	9	0	33
Male	4	1	4	2	6	4	5	3	29
Total	7	4	11	6	10	7	14	3	62

In [25]:

```
#From above table we can observe
t_no_of_m_f=62
t_no_of_m=29
t_no_of_f=33
P_m=29/62
P_f=33/62
print('Probability of being male :',P_m)
print('Probability of being female :',P_f)
```

```
Probability of being male : 0.46774193548387094
Probability of being female : 0.532258064516129
```

In [31]:

```
#2.2.2. Find the conditional probability of different majors among the male students in CMSU.
#Find the conditional probability of different majors among the female students of CMSU.
#Accounting
print('Contingency table Gender vs Major :')
print(gmt)
print(' ')
print('Probability accounting given male: ',4/29)
print('Probability accounting given female: ',3/33)
print(' ')
#CIS
print('Probability CIS given male: ',1/29)
print('Probability CIS given female: ',3/33)
print(' ')
#Eco/Finance
print('Probability E/F given male: ',4/29)
print('Probability E/F given female: ',7/33)
```

```
print('Probability E/F given female: ',7/33)
print(' ')
#International business
print('Probability IB given male: ',2/29)
print('Probability IB given female: ',4/33)
print(' ')
#Management
print('Probability Mgmt given male: ',6/29)
print('Probability Mgmt given female: ',4/33)
print(' ')
#other
print('Probability Other given male: ',4/29)
print('Probability Other given female: ',3/33)
print(' ')
#Retailing and Markerting
print('Probability R/M given male: ',5/29)
print('Probability R/M given female: ',9/33)
print(' ')
#Undecided
print('Probability Und given male: ',3/29)
print('Probability Und given female: ',0/33)
print(' ')
```

```
Contingency table Gender vs Major :
Major   Accounting  CIS  Economics/Finance  International Business  \
Gender
Female           3    3                  7                       4
Male             4    1                  4                       2
Total            7    4                 11                       6

Major   Management  Other  Retailing/Marketing  Undecided  Total
Gender
Female           4      3                    9          0     33
Male             6      4                    5          3     29
Total           10      7                   14          3     62

Probability accounting given male:  0.13793103448275862
Probability accounting given female:  0.09090909090909091

Probability CIS given male:  0.034482758620689655
Probability CIS given female:  0.09090909090909091

Probability E/F given male:  0.13793103448275862
Probability E/F given female:  0.21212121212121213

Probability IB given male:  0.06896551724137931
Probability IB given female:  0.12121212121212122

Probability Mgmt given male:  0.20689655172413793
Probability Mgmt given female:  0.12121212121212122

Probability Other given male:  0.13793103448275862
Probability Other given female:  0.09090909090909091

Probability R/M given male:  0.1724137931034483
Probability R/M given female:  0.2727272727272727

Probability Und given male:  0.10344827586206896
Probability Und given female:  0.0
```

In [32]:

```
#2.2.3. Find the conditional probability of intent to graduate, given that the student is a male.
#Find the conditional probability of intent to graduate, given that the student is a female.
#Cont table grad intention
gmgi
```

Out[32]:

Grad Intention	No	Undecided	Yes	Total
Gender				
Female	9	13	11	33

Grad Intention	No	Undecided	Yes	Total
Gender				
Male	3	9	17	29
Total	12	22	28	62

In [34]:

```
print('Probability intent to graduate given male: ', 17/29)
print('Probability intent to graduate given female: ',11/33)
```

```
Probability intent to graduate given male:  0.5862068965517241
Probability intent to graduate given female:  0.3333333333333333
```

In [35]:

```
#2.2.4. Find the conditional probability of employment status for the male students as well as for the female students.
#Cont table Gender vs employment
gme
```

Out[35]:

Employment	Full-Time	Part-Time	Unemployed	Total
Gender				
Female	3	24	6	33
Male	7	19	3	29
Total	10	43	9	62

In [36]:

```
print('Prob Full time given female: ',3/33)
print('Prob Full time given male: ',7/29)
print(' ')
print('Prob Part time given female: ',24/33)
print('Prob Part time given male: ',19/29)
print(' ')
print('Prob Unemployed  given female: ',6/33)
print('Prob Unemployed  given male: ',3/29)
```

```
Prob Full time given female:  0.09090909090909091
Prob Full time given male:  0.2413793103448276

Prob Part time given female:  0.7272727272727273
Prob Part time given male:  0.6551724137931034

Prob Unemployed  given female:  0.18181818181818182
Prob Unemployed  given male:  0.10344827586206896
```

In [37]:

```
#2.2.5. Find the conditional probability of laptop preference among the male students as well as among the female students.
#Cont table Gender vs Computer
gmc
```

Out[37]:

Computer	Desktop	Laptop	Tablet	Total
Gender				
Female	2	29	2	33
Male	3	26	0	29
Total	5	55	2	62

In [38]:

```
print('Prob laptop given female: ',29/33)
print('Prob laptop given male: ',26/29)
```

```
Prob laptop given female:  0.8787878787878788
Prob laptop given male:  0.896551724137931
```

In [2]:

```
#2.3. Based on the above probabilities, do you think that the column variable in each case is inde
pendent of Gender?
#Justify your comment in each case.
```

A:1)Independent events are situation in which one event does not affect the probability of occurence of another event.Here

we are able to see indepedent cases.For example

ex:Probability(Accounting given female)=3/33

Probability(Accounting given male)=4/29

Probability(Accounting given CMSU student)=7/62

We could see that when we consider students in general, prob of selecting accounting is 7/62

When we consider male the probability of taking Accounting is 4/29 does not influences the probability of

taking accounting when student is female 3/33.That is Prob of taking X Major given one gender does not affect the prob of

of occurence of taking X subject given another gender

2)Similarly for Gender vs Grad intention,when we consider one column variable for example intent to graduate given student is male

is 17/29 whereas for female is 11/33 , even if another female intent to Graduate it wont affect the occurence of intent to

graduate given male ie even if prob intent to graduate given female is 12/33 the prob of occurence intent to graduate given male

will be 17/29

3)Similarly for Gender Vs Employment status prob of female being part timed(24/33) is independent of male being Part timed(19/29)

4)Also prob of laptop being picked given male is independent of laptop being picked given female

In [9]:

```
#Part II
#• 2.4. Note that there are three numerical (continuous) variables in the data set, Salary, Spendi
ng and Text Messages.
#For each of them comment whether they follow a normal distribution.
#Write a note summarizing your conclusions.
#[Recall that symmetric histogram does not necessarily mean that the underlying distribution is sy
mmetric]
```

Out[9]:

```
48.54838709677419
```

To check whether a particular continuous variable follows normal distribution or not we have to verify the empirical rule

i.e whether 68,95 or 99 percent of data lie within 1,2 or 3 standard deviation from mean respectively i.e whether 68,95 or 99 percent of data lie within 1,2 or 3 standard deviation from mean respectively

In [26]:

```
#Calculate percentage of data within 1 standard deviation away from mean
#Salary:We could see from below Salary follows empirical rule we strongly assume it follows normal
distribution
no_of_obs_sal=62
sal_mean=sv['Salary'].mean()
sal_std=sv['Salary'].std()
one_time_std_right=sal_mean+sal_std
one_time_std_left=sal_mean-sal_std
count_1std= ((sv['Salary'] < one_time_std_right) & (sv['Salary'] > one_time_std_left)).sum()
Per_1std= (count_1std/no_of_obs_sal) * 100
```

In [27]:

```
Per_1std
```

Out[27]:

```
79.03225806451613
```

In [28]:

```
#Spending:We could see 80 percent of data lies within 1 standard deviation so it follows normal di
stribution
no_of_obs_sal=62
sal_mean=sv['Spending'].mean()
sal_std=sv['Spending'].std()
one_time_std_right=sal_mean+sal_std
one_time_std_left=sal_mean-sal_std
count_1std= ((sv['Spending'] < one_time_std_right) & (sv['Spending'] > one_time_std_left)).sum()
Per_1std= (count_1std/no_of_obs_sal) * 100
```

In [29]:

```
Per_1std
```

Out[29]:

```
80.64516129032258
```

In [30]:

```
#Text Messages:We could see 79 percent of data lies within 1 standard deviation following the empi
rical rule
#thus it follows normal Distribution
no_of_obs_sal=62
sal_mean=sv['Text Messages'].mean()
sal_std=sv['Text Messages'].std()
one_time_std_right=sal_mean+sal_std
one_time_std_left=sal_mean-sal_std
count_1std= ((sv['Text Messages'] < one_time_std_right) & (sv['Text Messages'] > one_time_std_left)
).sum()
Per_1std= (count_1std/no_of_obs_sal) * 100
```

In [31]:

```
Per_1std
```

Out[31]:

```
79.03225806451613
```

In [1]:

```
import pandas as pd
import numpy as np
from scipy.stats import ttest_1samp, ttest_ind
```

In [2]:

```
shg=pd.read_csv('D:\\ANALYTICS\\GREAT LEARNING\\7.Statistical Method for Decisoin Making-Week-4\\A
& B shingles-1.csv')
```

Problem 3

An important quality characteristic used by the manufacturers of ABC asphalt shingles is the amount of moisture the shingles contain when they are packaged. Customers may feel that they have purchased a product lacking in quality if they find moisture and wet shingles inside the packaging. In some cases, excessive moisture can cause the granules attached to the shingles for texture and colouring purposes to fall off the shingles resulting in appearance problems. To monitor the amount of moisture present, the company conducts moisture tests. A shingle is weighed and then dried. The shingle is then reweighed, and based on the amount of moisture taken out of the product, the pounds of moisture per 100 square feet is calculated. The company would like to show that the mean moisture content is less than 0.35 pound per 100 square feet. The file (A & B shingles.csv) includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.

In [3]:

```
shg.head()
```

Out[3]:

	A	B
0	0.44	0.14
1	0.61	0.15
2	0.47	0.31
3	0.30	0.16
4	0.15	0.37

In [4]:

```
shg.tail(n=10)
```

Out[4]:

	A	B
26	0.49	0.16
27	0.34	0.52
28	0.36	0.36
29	0.29	0.22
30	0.27	0.39
31	0.40	NaN
32	0.29	NaN
33	0.43	NaN
34	0.34	NaN
35	0.37	NaN

In [5]:

```
shg.describe()
```

Out[5]:

	A	B
count	36.000000	31.000000
mean	0.316667	0.273548
std	0.135731	0.137296
min	0.130000	0.100000
25%	0.207500	0.160000
50%	0.290000	0.230000
75%	0.392500	0.400000
max	0.720000	0.580000

In [8]:

```
#3.1. For the A shingles, form the null and alternative hypothesis to test whether the
#population mean moisture content is less than 0.35 pound per 100 square feet.
```

A.H0(Null Hypothesis)=> Population mean=0.35 pound per 100 square feet,

Ha(Alternate Hypothesis=>Population mean < 0.35 pound per 100 square feet

In [7]:

```
#3.2. For the A shingles, conduct the test of hypothesis and find the p-value. Interpret the p-val
ue.
#Is there evidence at the 0.05 level of significance that the population mean moisture
#content is less than 0.35 pound per 100 square feet?
t_stat,p_val=ttest_1samp(shg['A'],0.35)
print('Statistic :',t_stat)
print('P-Value :',p_val)
```

```
Statistic : -1.4735046253382782
P-Value : 0.14955266289815025
```

At .05 significance level there is no clear evidence to reject the null hypothesis

Here we fail to reject the null hypothesis since P-val(0.14)>alpha-val(0.05)

In [9]:

```
#3.3. For the B shingles, form the null and alternative hypothesis to test
#whether the population mean moisture content is less than 0.35 pound per 100 square feet.
```

A. H0(Null Hypothesis)= Population mean = 0.35 pound per 100 square feet,

Ha(Alternate Hypothesis)= Population mean < 0.35 pound per 100 square feet

In [14]:

```
#3.4. For the B shingles, conduct the test of the hypothesis and find the p-value. Interpret the p
-value.
#Is there evidence at the 0.05 level of significance that the population mean moisture
#content is less than 0.35 pound per 100 square feet?
t_stats,p_value=ttest_1samp(shg['B'][:31],0.35)
print('Statistic :',t_stats)
print('P-Value :',p_value)
```

```
Statistic : -3.1003313069986995
P-Value : 0.004180954800638363
```

We reject the null hypothesis at 0.05 significance level since P-val(0.004)<Alpha_val(.05)

In [15]:

```
#3.5. Do you think that the population means for shingles A and B are equal?
#Form the hypothesis and conduct the test of the hypothesis.
#What assumption do you need to check before the test for equality of means is performed?
```

A.H0(Null Hypothesis)=>Mean of A equal to Mean of B

Ha(Alternate Hypothesis)=>Mean of A not equal to Mean of B

Since we are comparing 2 samples for equality of variance we assume equal variance

In [16]:

```
t_2samp,p_val_2samp=ttest_ind(shg['A'],shg['B'][:31])
print('Statistic :',t_2samp)
print('P-Value :',p_val_2samp)
```

```
Statistic : 1.289628271966112
P-Value : 0.2017496571835328
```

There is no evidence at 0.05 significance level to reject the null hypothesis,here we fail to

reject the null hypothesis

In [17]:

```
#3.6. What assumption about the population distribution is needed in order to conduct the
hypothesis tests above?
```

1.Sample data is a representative of population data and hypothesis made on sample data

2.Sample data taken is significantly larger so that it will follow normal distribution(bell-curve)

3.While testing 2 samples from same population for comparing means equal variance is assumed

In [23]:

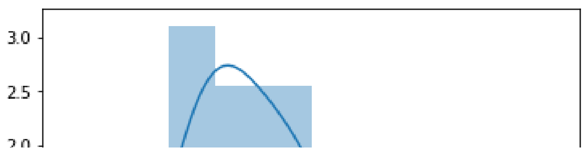
```
#3.7 Check the assumptions made with histograms, boxplots, normal probability plots or empirical r
ule
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
```

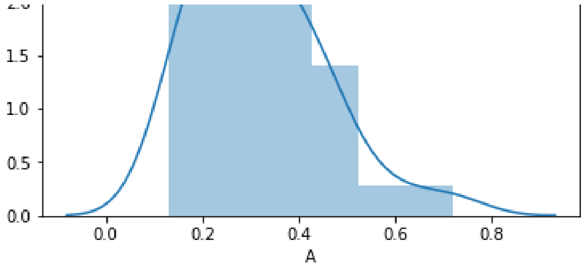
In [24]:

```
sns.distplot(shg['A'])
```

Out[24]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x262b2574b38>
```



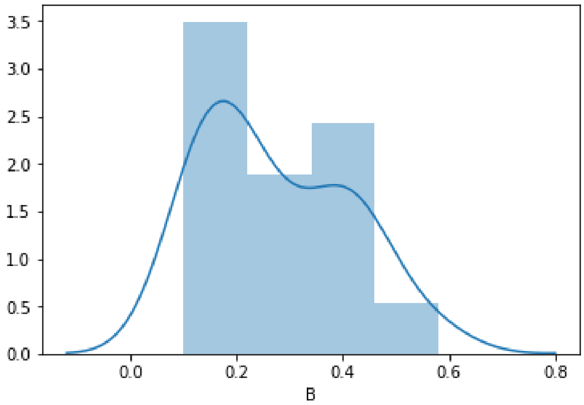


In [25]:

```
sns.distplot((shg['B'][:31]))
```

Out[25]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x262b2522898>
```



From above we could see there is slight distortion in bell-shaped curve it is not perfectly normally distributed.

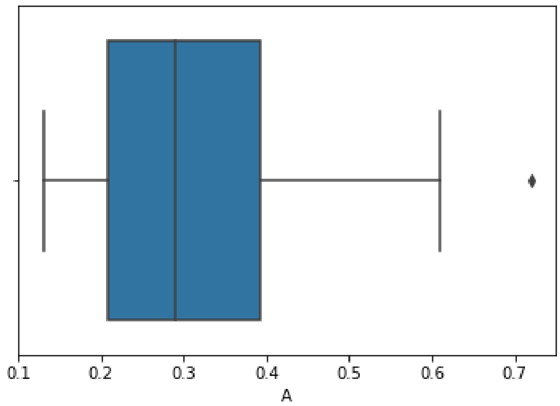
If we are able to add more samples for testing curve will approximate to bell-curve

In [26]:

```
sns.boxplot(shg['A'],orient='h')
```

Out[26]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x262b26489b0>
```



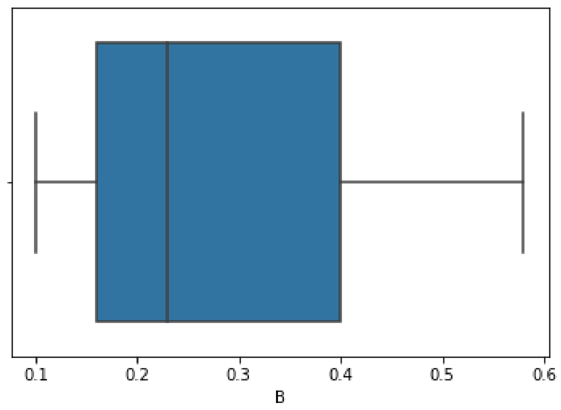
We could see an outlier which which makes the bell curve skewed and distorted

In [27]:

```
sns.boxplot(shg['B'][:31])
```

Out[27]:

```
&lt;matplotlib.axes._subplots.AxesSubplot at 0x262b26a0ac8&gt;
```



We could see for shingles B Median is towards left that is positively skewed

In [31]:

```
#Empirical rule
mean_A=shg['A'].mean()
std_A=shg['A'].std()
mean_1std_right=mean_A+std_A
mean_1std_left=mean_A-std_A
cnt=((shg['A'] < mean_1std_right) & (shg['A'] > mean_1std_left)).sum()
pect=cnt/36
```

In [32]:

```
pect
```

Out[32]:

```
0.7222222222222222
```

Empirical rule defines 68 percent of data lies between 1 standard deviation away from mean

but shingles A has 72 percentage of data between 1 standard deviation away from mean

In [33]:

```
mean_B=shg['B'][:31].mean()
std_B=shg['B'][:31].std()
mean_1std_right=mean_B+std_B
mean_1std_left=mean_B-std_B
cnt=((shg['B'][:31] < mean_1std_right) & (shg['B'][:31] > mean_1std_left)).sum()
pect=cnt/36
```

In [34]:

```
pect
```

Out[34]:

```
0.5277777777777778
```

We could see whil Shingles B only contains only 52 percentage of data within 1 standard deviation away from mean

which fails the empirical rule

In [35]:

```
#3.8. Do you think that the assumption needed in order to conduct the hypothesis tests above is va
lid? Explain
```

Shingles A has 36 samples and Shinlges B has 31 samples, given the population mean and no population standard deviation

it is safe to assume T-test for the samples. 30 samples is enough to assume that the sample data is large enough to follow

normal distribution, given the sample size of 31 and 36 it is safe to assume it follows normal distribution.