NAME : VIKRAM RADHAKRISHNAN SMDM ASSIGNMENT – 23rd FEB,2020

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 Customer.csv) 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).

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?

SUMMARY BY REGION

Region	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	sum_of_expenditure
Lisbon	18095	854833	422454	570037	231026	204136	104327	2404908
Oporto	14899	464721	239144	433274	190132	173311	54506	1569987
Other	64026	3960577	1888759	2495251	930492	890410	512110	10741625

SUMMARY BY CHANNEL

Channel	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	sum_of_expenditure
Hotel	71034	4015717	1028614	1180717	1116979	235587	421955	8070603
Retail	25986	1264414	1521743	2317845	234671	1032270	248988	6645917

The Region with maximum expenditure is : Other

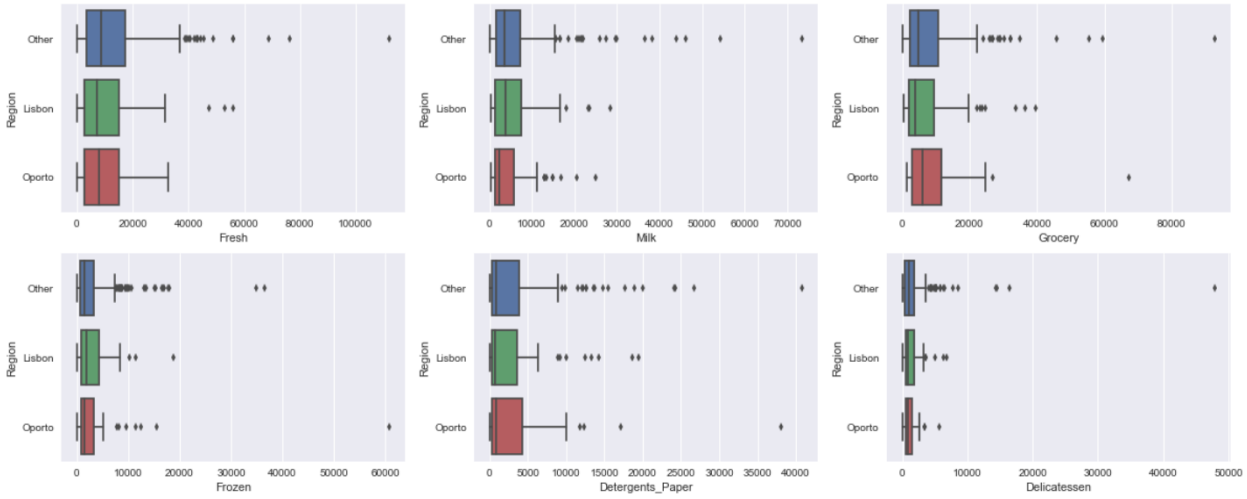
The Channel with maximum expenditure is : Hotel

The Region with minimum expenditure is : Oporto

The Channel with minimum expenditure is : Retail

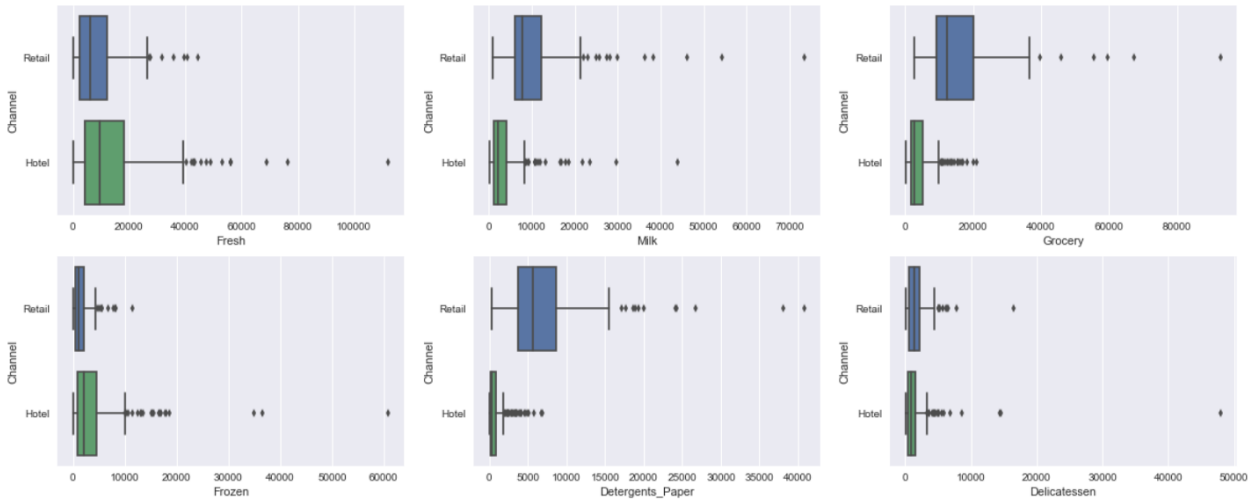
1.2. There are 6 different varieties of items are considered.
Do all varieties show similar behaviour across Region and Channel?

ITEMS ACROSS REGION



The above picture shows the behavior of the 6 items considered by Region. By looking at the plots, it can be summarized that the behavioral of the items across region is similar across region for different items.

ITEMS ACROSS CHANNEL



The above picture shows the behavior of the 6 items considered by Channel. By looking at the plots, it can be summarized that the behavioral of the items across region is not similar across channel for different items.

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?

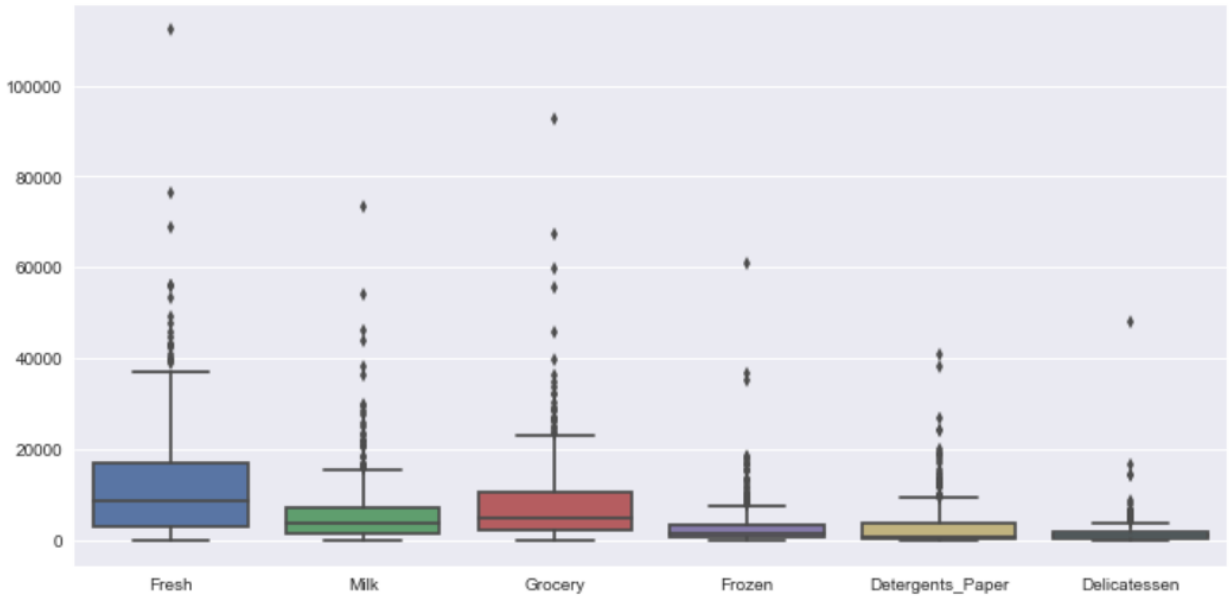
```
Coefficient of Variation values are

 {'Delicatessen': 1.8494068981158382,
 'Detergents_Paper': 1.6546471385005155,
 'Fresh': 1.0539179237473149,
 'Frozen': 1.5803323836352914,
 'Grocery': 1.1951743730016824,
 'Milk': 1.2732985840065414}
```

The **coefficient of variation** of the **Delicatessen** category is the **highest** with a value of : 1.8494068981158382 and therefore **exhibits most inconsitent behavior**

1.4. Are there any outliers in the data?

Yes! There are outliers in the data, and can be seen in the boxplots below.

There are outliers in all the variables as seen in boxplot below



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

By looking at this report, we are able to understand patterns of spending across different regions and channels for different products. In order to increase sales across these various categories, sales and marketing campaigns can be conducted based on the data presented to increase sales depending on the nature of the region/channel. Further, one can deep dive and try to understand the reason for these trends, and take remedial/improvisations to improve sales.

There are some categories like "Delicattesen" that could perform better as seen from the report, and more attention needs to be given to this item and understand how it can be made better.

It is also seen that customers are spending more money on basic necessities like Milk, Fresh products and Grocery, while spending less on processed consumer products like Frozen, Detergents, Paper and Delicatessen. Its therefore useful to pay more attention to these items to increase sales.

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).

Part I

- 2.1. For this data, construct the following contingency tables (Keep Gender as row variable)

2.1.1. Gender and Major

	Other	Management	CIS	Economics/Finance	Undecided	International Business	Retailing/Marketing	Accounting	Total
Female	3.0	4.0	3.0	7.0	0.0	4.0	9.0	3.0	33.0
Male	4.0	6.0	1.0	4.0	3.0	2.0	5.0	4.0	29.0
m_f_tot	7.0	10.0	4.0	11.0	3.0	6.0	14.0	7.0	62.0

2.1.2. Gender and Grad Intention

	Yes	Undecided	No	Total
Female	11.0	13.0	9.0	33.0
Male	17.0	9.0	3.0	29.0
m_f_tot	28.0	22.0	12.0	62.0

2.1.3. Gender and Employment

	Full-Time	Part-Time	Unemployed	Total
Female	3.0	24.0	6.0	33.0
Male	7.0	19.0	3.0	29.0
m_f_tot	10.0	43.0	9.0	62.0

2.1.4. Gender and Computer

	Laptop	Tablet	Desktop	Total
Female	29.0	2.0	2.0	33.0
Male	26.0	0.0	3.0	29.0
m_f_tot	55.0	2.0	5.0	62.0

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?

Prob that randomly selected student is a male

= (Number of male students/ Total number of Students)

=29/ (29+33)

=46.77%

Probability that randomly selected student is female

= 100% - 46.77%

=53.23%

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.

Among Males, the conditional probabilities of different majors are given in the picture below :

```
contigency_tab1.loc['Cond_prob_maj_males']
```

Other	0.137931
Management	0.206897
CIS	0.034483
Economics/Finance	0.137931
Undecided	0.103448
International Business	0.068966
Retailing/Marketing	0.172414
Accounting	0.137931
Total	1.000000

Among Females, the conditional probabilities are given in the picture below :

```
contigency_tab1.loc['Cond_prob_maj_fem']

Other                          0.090909
Management                     0.121212
CIS                            0.090909
Economics/Finance              0.212121
Undecided                      0.000000
International Business         0.121212
Retailing/Marketing            0.272727
Accounting                     0.090909
Total                          1.000000
Name: Cond_prob_maj_fem, dtype: float64
```

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.

	Yes	Undecided	No	Total
Female	11.0	13.0	9.0	33.0
Male	17.0	9.0	3.0	29.0
m_f_tot	28.0	22.0	12.0	62.0

	Yes	Undecided	No	Total
Female	11.000000	13.000000	9.000000	33.0
Male	17.000000	9.000000	3.000000	29.0
m_f_tot	28.000000	22.000000	12.000000	62.0
Intent_grad_mal_prob	0.586207	0.310345	0.103448	1.0
Intent_grad_fem_prob	0.333333	0.393939	0.272727	1.0

Intent to graduate given student is male = 17/33 = 0.586

Intent to graduate given student is female=11/33= 0.333

2.2.4. Find the conditional probability of employment status for the male students as well as for the female students.

	Full-Time	Part-Time	Unemployed	Total
Female	3.0	24.0	6.0	33.0
Male	7.0	19.0	3.0	29.0
m_f_tot	10.0	43.0	9.0	62.0

Employment status for male (full time) = 7/(29) = 0.24

Employment status for male (Part-time) = 19/(total males) = 19/24=0.79

Employment status for male (Unemployed) = 3/(total males) = 3/24=0.125

Employment status for female (full time) = 3/(33) = 0.10

Employment status for female (Part-time) = 24/(total females) = 19/33=0.58

Employment status for female (Unemployed) = 6/(total females) = 6/33=0.18

2.2.5. Find the conditional probability of laptop preference among the male students as well as among the female students.

	Laptop	Tablet	Desktop	Total
Female	29.0	2.0	2.0	33.0
Male	26.0	0.0	3.0	29.0
m_f_tot	55.0	2.0	5.0	62.0

Conditional prob for laptop (Male) = 26/(total males)=26/29 =0.9

Conditional prob for laptop (Female)=29/(total females)=29/33=0.88

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

Case - Gender and Major

In this case, its observed that a few majors like Marketing/Retailing & Economics/Finance are preferred by female candidates. It is also worthy to note that there are no female candidates with undecided majors! In males, management appears to be the major of choice. In comparison with females, a larger number of males appear to be taking the 'other' majors. A large fraction of males also appear to be undecided on the major they would like to pursue.

Case- Gender and Grad Intention

	Yes	Undecided	No	Total
Female	11.000000	13.000000	9.000000	33.0
Male	17.000000	9.000000	3.000000	29.0
m_f_tot	28.000000	22.000000	12.000000	62.0
Intent_grad_mal_prob	0.586207	0.310345	0.103448	1.0
Intent_grad_fem_prob	0.333333	0.393939	0.272727	1.0

It is clear that the intent to graduate is more in males than females, and therefore not independent of gender.

Case-Gender and Employment

	Full-Time	Part-Time	Unemployed	Total
Female	3.0	24.0	6.0	33.0
Male	7.0	19.0	3.0	29.0
m_f_tot	10.0	43.0	9.0	62.0

It is observed that the employment rate in males is slightly more than females. This could be due to various reasons beyond the scope of this study.

Case-Gender and Computer

	Laptop	Tablet	Desktop	Total
Female	29.0	2.0	2.0	33.0
Male	26.0	0.0	3.0	29.0
m_f_tot	55.0	2.0	5.0	62.0

Gender does not play are big role in the choice of computers as seen in the data presented. A majority of the students opt for laptops. While we see that a few females go for Tablets and desktops, the numbers are too small to draw a significant conclusion on gender related preference.

Overall, certain observations are found to be related to student gender, and others are independent.

Part II

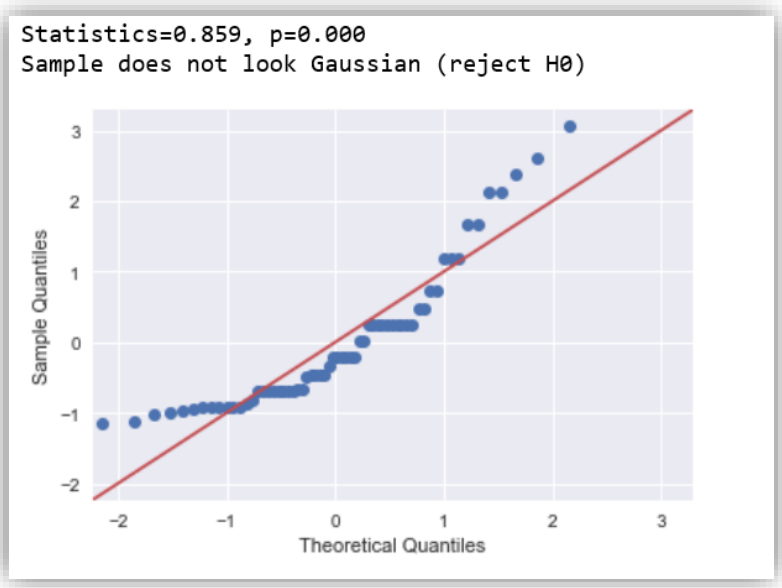
- 2.4. Note that there are three numerical (continuous) variables in the data set, Salary, Spending and Text Messages. For each of them comment whether they follow a normal distribution.
  Write a note summarizing your conclusions.
  [Recall that symmetric histogram does not necessarily mean that the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

To determine whether the column in question is normal

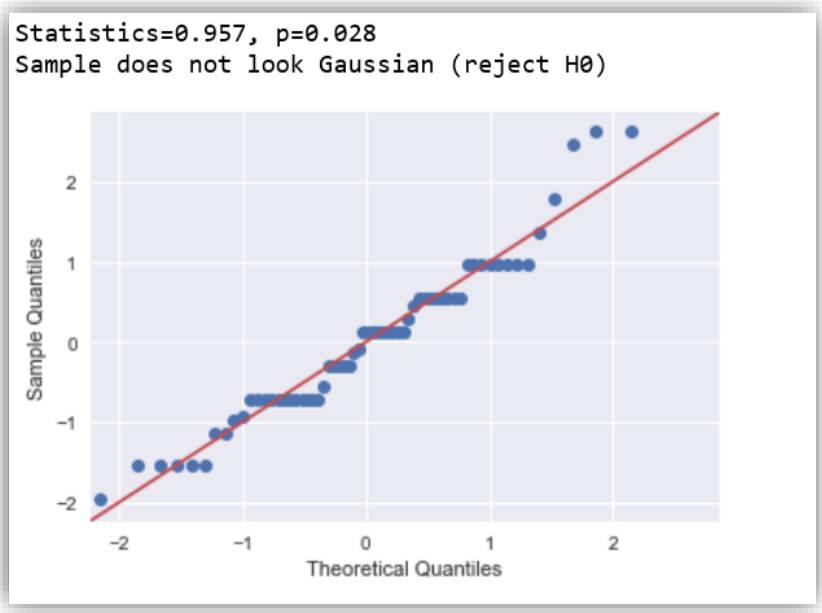
a) ***We plot a qq-plot to check if the values fall close to the '45' degree line.***

b) ***Alternatively, the SHAPIRO-WILK normality test can be run (Hypothesis approach), and a 'P' value can be determined. If P is less than alpha, the null hypothesis is rejected i.e. the sample will not be Gaussian/normal***

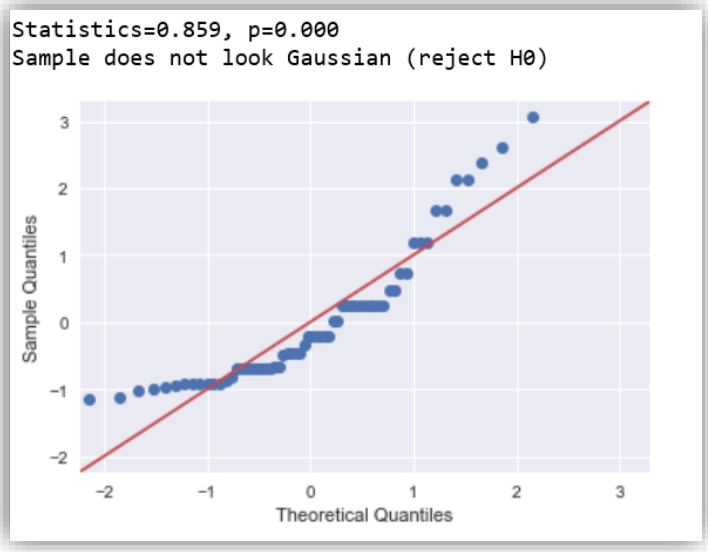
SALARY: - The sample does not look normal (in QQ Plot), and the P value (also in figure) is close to zero. Therefore is not Normal distribution



SPENDING : ***The sample does not look normal, and from P value is less than 0.05 (alpha). We therefore conclude its not a normal distribution. However, if we relax our alpha value to 0.01, the distribution would be normal by those modified standards.***



TEXT MESSAGES : ***- The sample does not look normal according to qqplot, and from P value is less than 0.05 (alpha). We therefore conclude its not a normal distribution.***



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 claims that the mean moisture content cannot be greater 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.

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.

HYPOTHESIS TEST :

Ho : Mean moisture content of A shingles = or < 0.35 pound per 100 sq ft

Ha : Mean moisture content of A shingles > 0.35 pound per 100 sq ft

****

```
df_shingles['A'].mean()
```

```
0.3166666666666666
```

From calculation, we observe that the mean moisture value of 'A' is 0.31666 units which is less that 0.35 units. Therefore, the Null Hypothsis (Ho) holds good. The Mean moisture content is less than 0.35 pounds per 100 sq ft.

3.2. 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.

HYPOTHESIS TEST :

Ho : Mean moisture content of B shingles = or < 0.35 pound per 100 sq ft

Ha : Mean moisture content of B shingles > 0.35 pound per 100 sq ft

From calculation, we observe that the mean moisture value of column 'B' is 0.27 units which is less that 0.35 units. Therefore, the Null Hypothesis (Ho) holds good. This means that B shingles under consideration have mean moisture content below 0.35 pounds per 100 sq ft.

3.3. 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?

For Hypothesis testing, we would assume that that the mean difference between the two samples would be zero:

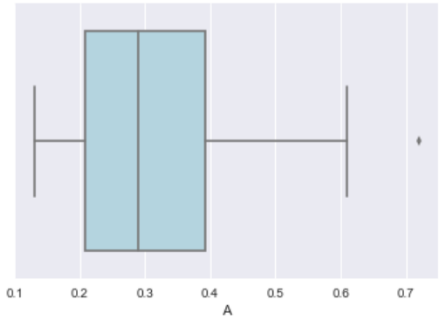
i.e. $\mu 1 = \mu 2 \Rightarrow \mu 1 - \mu 2 = 0$ (Null hypothesis)

$\mu 1 \neq \mu 2$ (Alternate hypothesis)

An assumption is that the data, when plotted, results in a normal distribution, bell-shaped distribution curve. When a normal distribution is assumed, one can specify a level of probability (alpha level, level of significance, p) as a criterion for acceptance. In most cases, a 5% value can be assumed of the data is normal. In our case, data in group2 does not appear to be strongly normally distributed when normal distribution test (like Shapiro-Wick) are run with alpha values of 5% . Therefore, we reduce the alpha level to 1%.

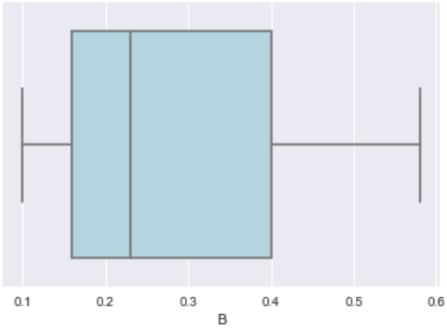
```
sns.boxplot((group1), color='lightblue')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1df9e576cc0>
```




```
sns.boxplot(group2, color='lightblue')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1df9ff64588>
```



To understand if the population means of the two sample sets are equal, we perform the Two Sample T-test with groups of Shingles 'A' and 'B'

Reject Null Hypothesis if |T| is large. Also check for value of P.

"If P value is high (greater than type one error coefficient, alpha=0.01), null will fly"

```
t_statistic, p_value=ttest_ind(group1,group2.dropna(),equal_var=False)
```

```
t_statistic,p_value
```

```
(1.2885080295255027, 0.20225822050217818)
```

We observe that the P value is greater than 0.01 (alpha). Therefore, the null hypothesis holds good at this criterion of acceptance (i.e) The population means of the two samples are equal at alpha=0.01. However, for alpha=0.05, this does not hold good as the samples will no longer be normal.

The assumption to check before equality of variance is performed is homogeneity of variance. Homogeneous, or equal, variance exists when the standard deviations of samples are approximately equal such as the case below where the standard deviations are almost same for samples A and B. The variances are also similar.

```
# Comparing Variances and standard deviation of two colums A and B
```

```
print("Column A has a variance of ",group1.var()," and standard deviation of ",group1.std()) # Column A
```

```
Column A has a variance of  0.018422857142857133  and standard deviation of  0.13573082605973166
```

```
print("Column B has a variance of ",group2.var()," and standard deviation of ",group2.std()) # Column B
```

```
Column B has a variance of  0.018850322580645163  and standard deviation of  0.13729647694185443
```

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

The assumption would be that a reasonably large sample size with continuous data type is used (usually more than 30 data points in practice). A larger sample size means the distribution of results should approach a normal bell-shaped curve.