SMDM -
Business Report

Rohan R. Khade

Table of Contents

Back to Top

Back to Top

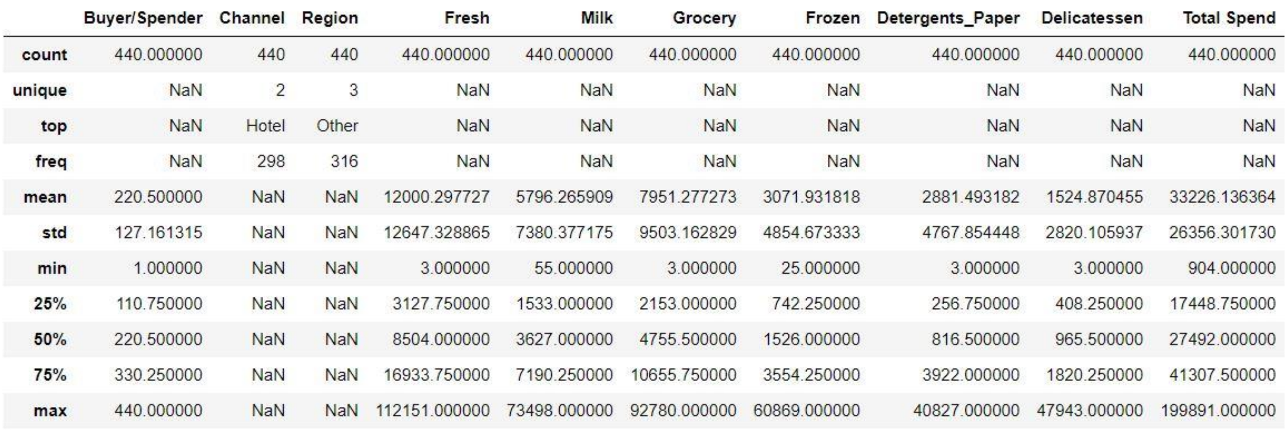
Back to Top

Chapter 1. Problem 1: Wholesale Customers Analysis

1.1 Problem Statement

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 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, Retail).

1.2 Use methods of descriptive statistics to summarize data.



	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total Spend
count	440.000000	440	440	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
unique	NaN	2	3	NaN	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	Hotel	Other	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	298	316	NaN	NaN	NaN	NaN	NaN	NaN	NaN
mean	220.500000	NaN	NaN	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455	33226.136364
std	127.161315	NaN	NaN	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937	26356.301730
min	1.000000	NaN	NaN	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000	904.000000
25%	110.750000	NaN	NaN	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000	17448.750000
50%	220.500000	NaN	NaN	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000	27492.000000
75%	330.250000	NaN	NaN	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000	41307.500000
max	440.000000	NaN	NaN	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000	199891.000000

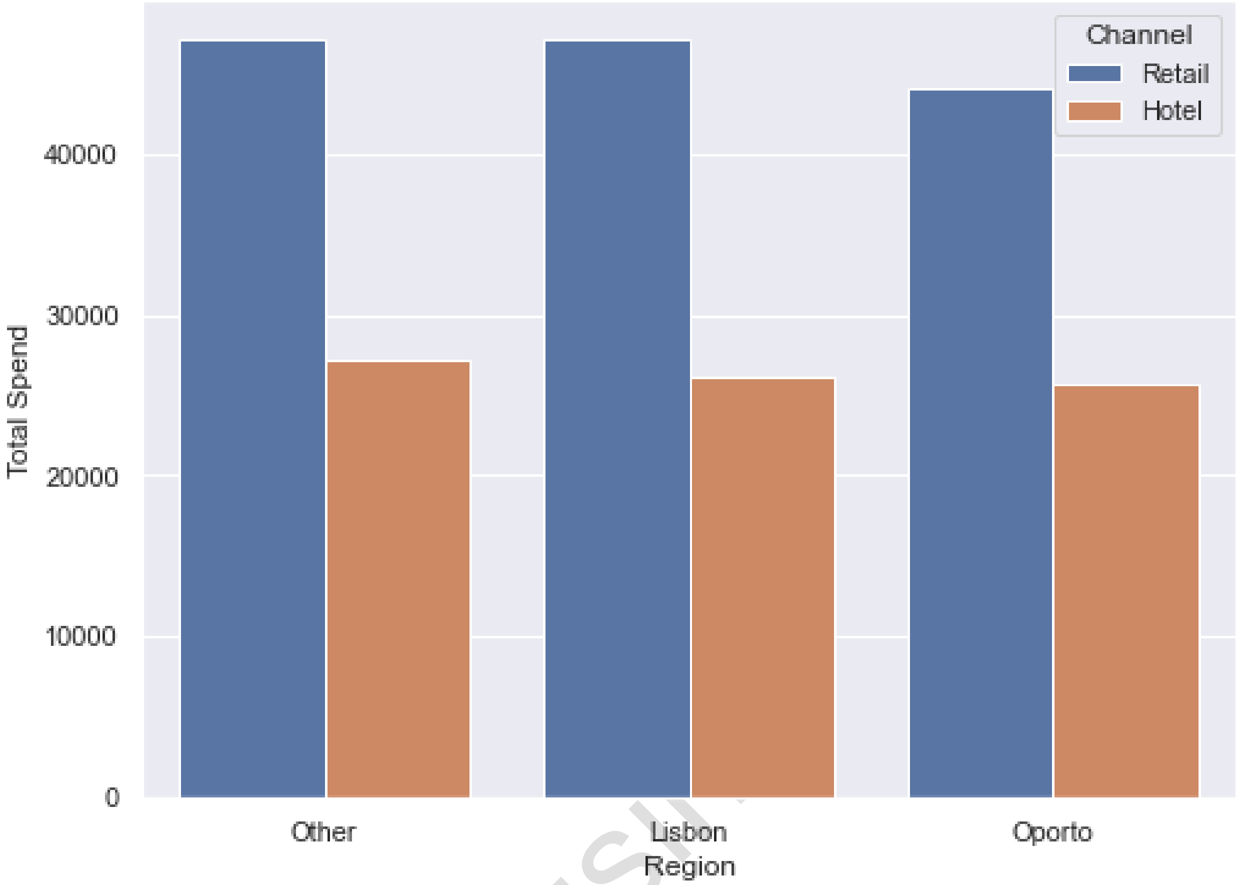
By loading the dataframe we identify the continuous variables and add a new column with total spend by the buyer/spender then we use the describe function (with include all) to analyze the data.

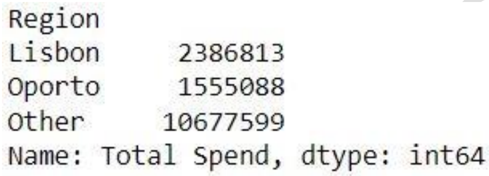
1.2.1 Which Region and which Channel spent the most? Which Region and which Channel spent the least?

By plotting a bar graph and pivot table for region, channel, and total spend, we can understand which region and channel spent most across the products given in the dataframe. From the below python outputs, we can conclude that among the given channels, hotel segment dominated the spending, whereas other segment was the largest spender in the given regions. Furthermore, the Oporto and retail segments were the lowest spenders in region and channel segment respectively.

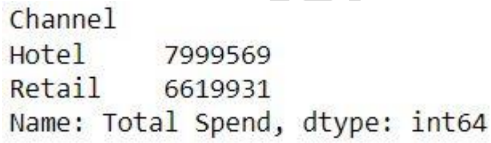
Back to Top

Python Output:





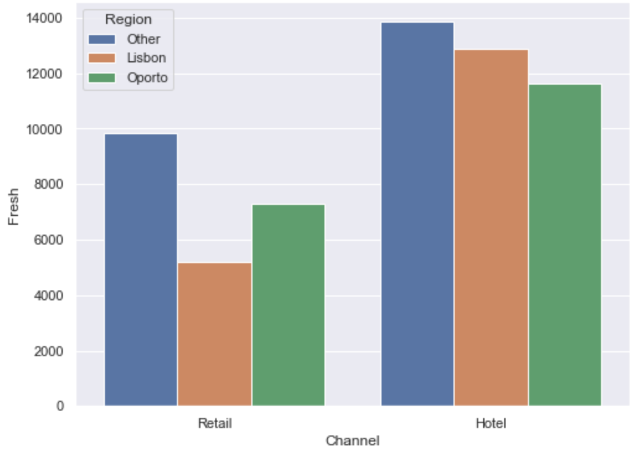
```
Region
Lisbon      2386813
Oporto      1555088
Other      10677599
Name: Total Spend, dtype: int64
```

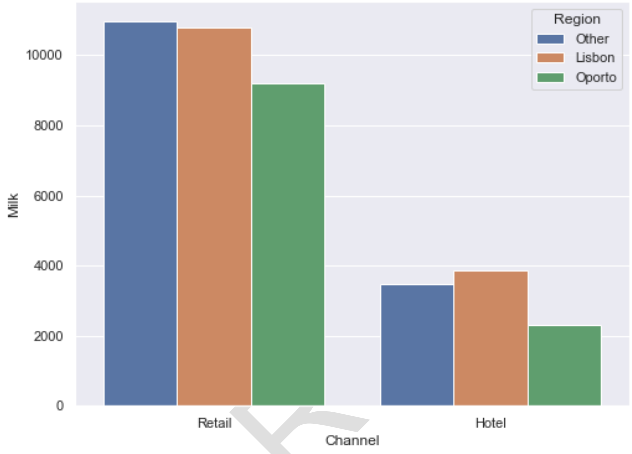


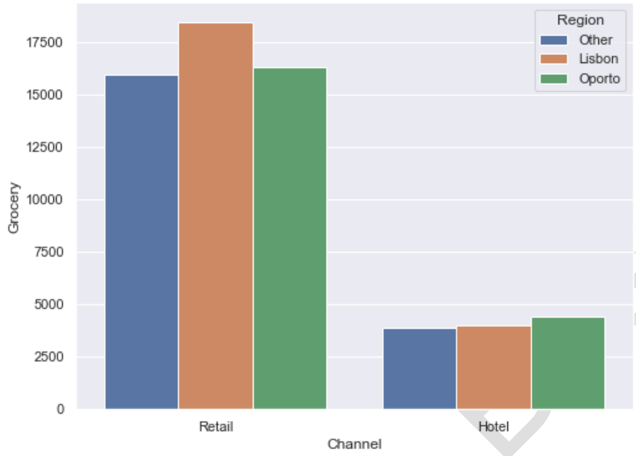
```
Channel
Hotel     7999569
Retail    6619931
Name: Total Spend, dtype: int64
```

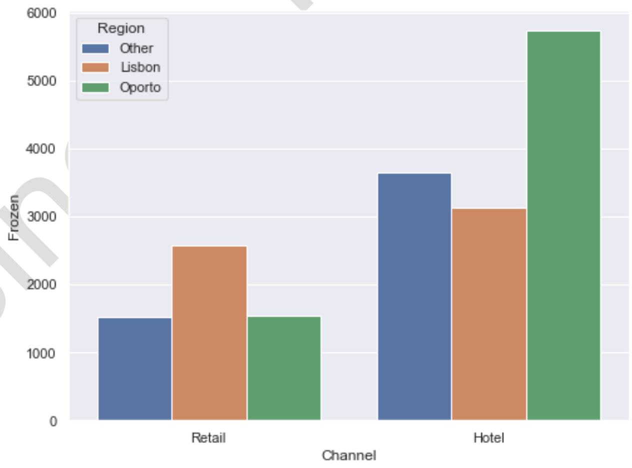
Back to Top

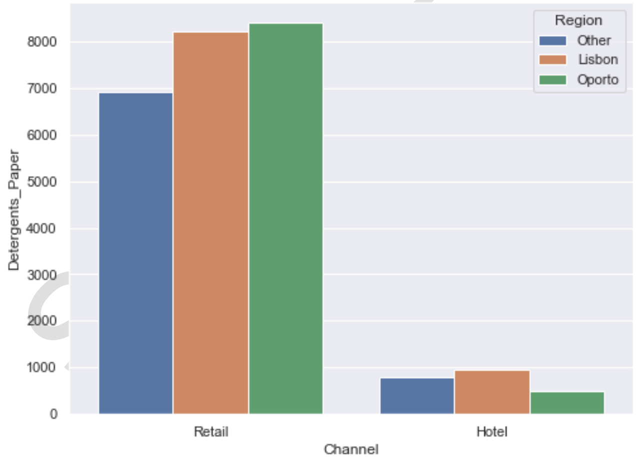
1.3 There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.

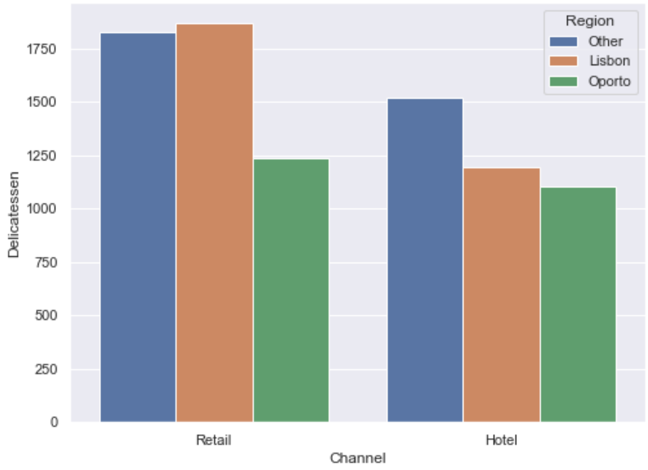








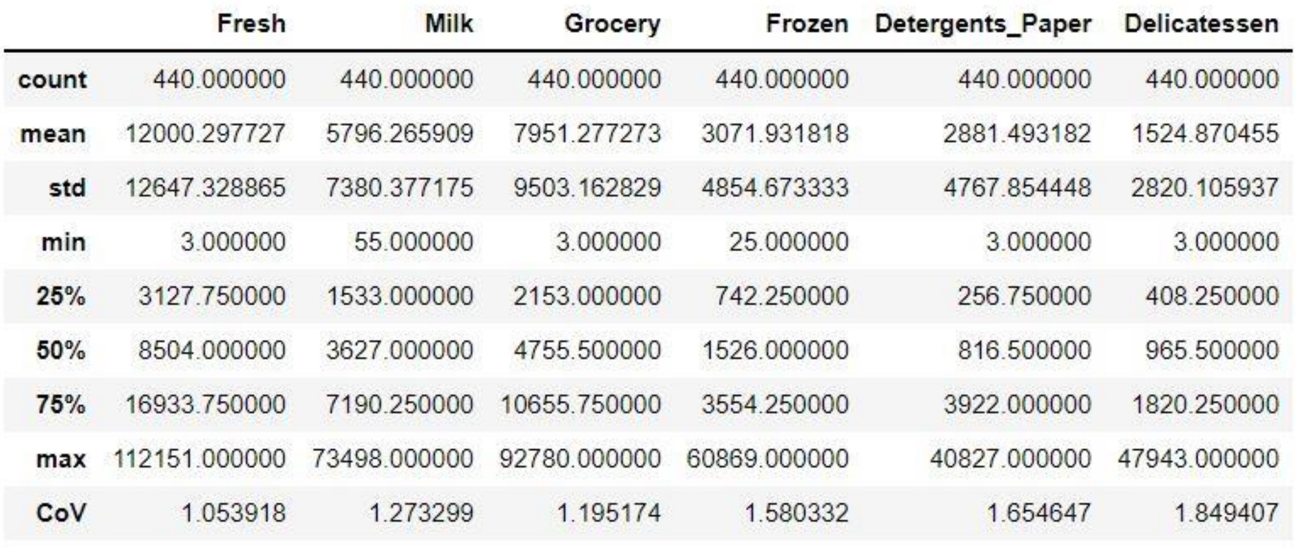




The spending on milk, grocery, detergents_paper, and delicatessen through retail channel is higher than through hotels across the three regions. However, the spending on fresh and frozen products is higher through the hotel channel across the regions.

Furthermore, the other region spent more than Lisbon and Oporto on fresh, milk, grocery, detergents_paper, and delicatessen. In addition to the aforementioned data points, we can also understand that except for delicatessen, spending through one channel is significantly higher as compared to other channel on each of the given products.

1.4 On the basis of a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?



	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
mean	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000
50%	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000
75%	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000
max	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000
CoV	1.053918	1.273299	1.195174	1.580332	1.654647	1.849407

The aforementioned table was derived by,

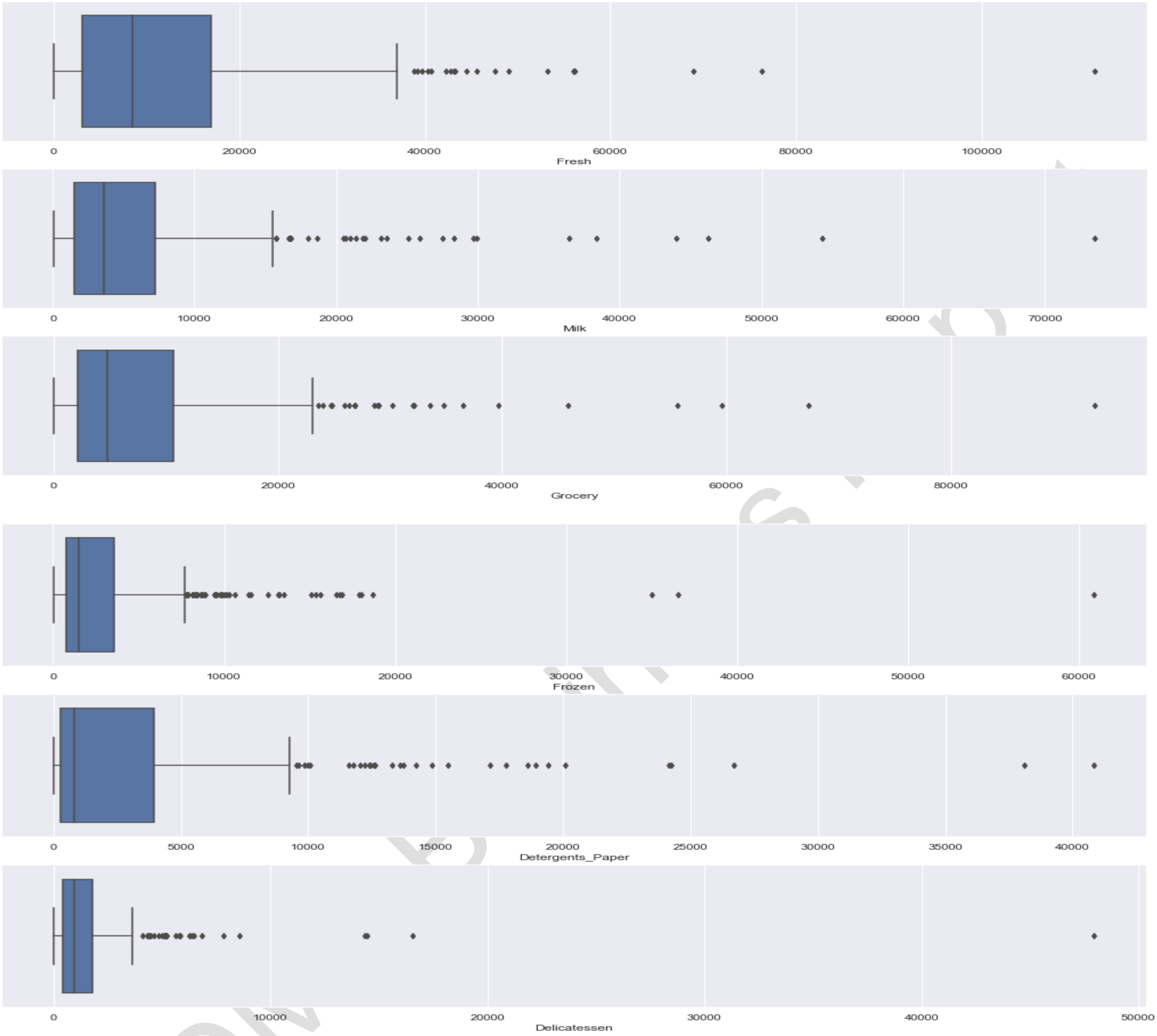
a) making new dataframe (df1 in the jupyter notebook) dropping the variables with data type as string and using the describe function in the same code.

b) Then calculating the coefficient of variation by using the .loc function and dividing the standard deviation by mean of each product.

By looking at the aforementioned output we can conclude that the Fresh product category shows least inconsistency with 1.05 as coefficient of variation, whereas Delicatessen category shows most inconsistency with 1.85 as coefficient of variation.

Back to Top

1.5 Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.



By plotting boxplots for all the product categories, we can be definitive that all the categories have outliers and they are positively skewed as none of the outliers are on the left side of the whisker.

Back to Top

1.6 On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem? Answer from the business perspective

Considering one series at a time in the dataframe, we can see that the spending through the two channels is irregular, the spending on few product categories is high through retail such as milk, grocery, detergents_paper, and delicatessen. However, the spending through hotel is high for fresh and frozen product categories. There need to consistency in spending across the channel in all the regions and for all the products.

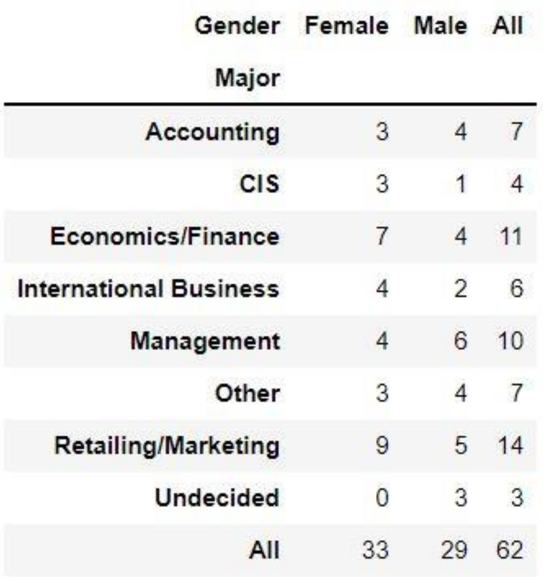
The business should find a solution to balance the spending across the regions and channels for all products. The spending on milk, grocery, and detergents_paper is extremely low through hotel channel across the regions. The business needs to identify the factors responsible for the same and take necessary steps to improve the sales or focus on products which are star performers.

Back to Top

Chapter 2. Problem 2: Student News Service Survey at Clear Mountain State University

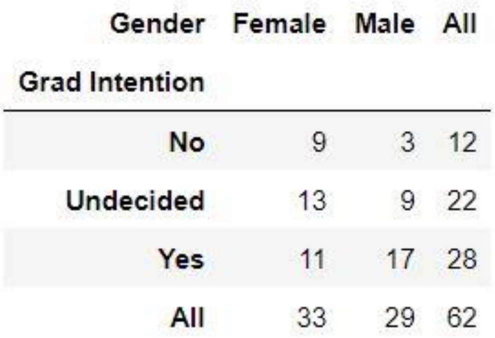
2.1 For this data, construct the following contingency tables (Keep Gender as row variable)

2.1.1 Gender and Major



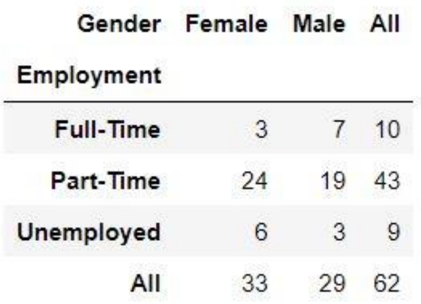
Gender	Female	Male	All
Major			
Accounting	3	4	7
CIS	3	1	4
Economics/Finance	7	4	11
International Business	4	2	6
Management	4	6	10
Other	3	4	7
Retailing/Marketing	9	5	14
Undecided	0	3	3
All	33	29	62

2.1.2 Gender and Grad Intention



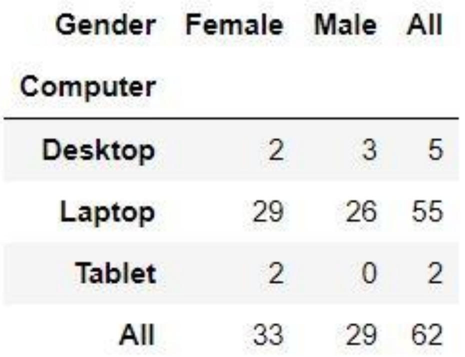
Gender	Female	Male	All
Grad Intention			
No	9	3	12
Undecided	13	9	22
Yes	11	17	28
All	33	29	62

2.1.3 Gender and Employment



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

2.1.4 Gender and Computer



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

Back to Top

2.2 Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

2.2.1 What is the probability that a randomly selected CMSU student will be male?

The probability of a randomly selected student will be male is 0.46774193548387094 based on calculation in jupyter notebook.

2.2.2 What is the probability that a randomly selected CMSU student will be female?

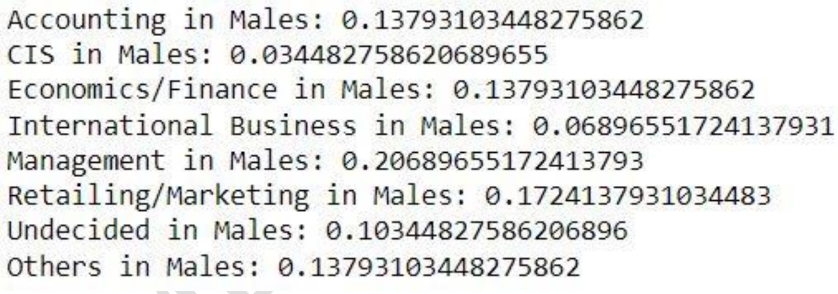
The probability of a randomly selected student will be female is 0.532258064516129 based on calculation in jupyter notebook

2.3 Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

2.3.1 Find the conditional probability of different majors among the male students in CMSU.

The conditional probability of different majors among the male students in CMSU based on calculation in jupyter notebook is:

Python Output:



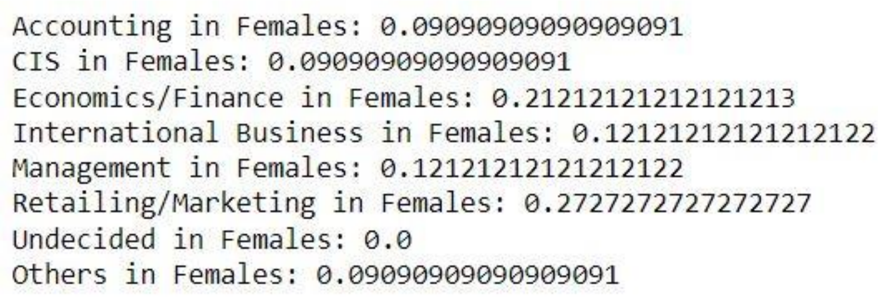
```
Accounting in Males: 0.13793103448275862
CIS in Males: 0.034482758620689655
Economics/Finance in Males: 0.13793103448275862
International Business in Males: 0.06896551724137931
Management in Males: 0.20689655172413793
Retailing/Marketing in Males: 0.1724137931034483
Undecided in Males: 0.10344827586206896
Others in Males: 0.13793103448275862
```

Majors	Male Probability
Accounting	0.13793103448275862
CIS	0.034482758620689655
Economics/Finance	0.13793103448275862
International Business	0.06896551724137931
Management	0.20689655172413793
Retailing/Marketing	0.1724137931034483
Undecided	0.10344827586206896
Others	0.13793103448275862

2.3.2 Find the conditional probability of different majors among the female students of CMSU.

The conditional probability of different majors among the female students in CMSU based on calculation in jupyter notebook is:

Python Output:



```
Accounting in Females: 0.09090909090909091
CIS in Females: 0.09090909090909091
Economics/Finance in Females: 0.21212121212121213
International Business in Females: 0.12121212121212122
Management in Females: 0.12121212121212122
Retailing/Marketing in Females: 0.2727272727272727
Undecided in Females: 0.0
Others in Females: 0.09090909090909091
```

Majors	Male Probability
Accounting	0.09090909090909091
CIS	0.09090909090909091
Economics/Finance	0.21212121212121213
International Business	0.12121212121212122
Management	0.12121212121212122
Retailing/Marketing	0.2727272727272727
Undecided	0.0
Others	0.09090909090909091

2.4 Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:

2.4.1 Find the probability That a randomly chosen student is a male and intends to graduate.

The probability of a randomly chosen student to be male who intend to graduate is 0.5862068965517241 based on calculation in jupyter notebook.

2.4.2 Find the probability that a randomly selected student is a female and does NOT have a laptop.

The probability of a randomly chosen student to be female and does not have a laptop is 0.12121212121212122 based on calculation in jupyter notebook.

Back to Top

2.5 Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

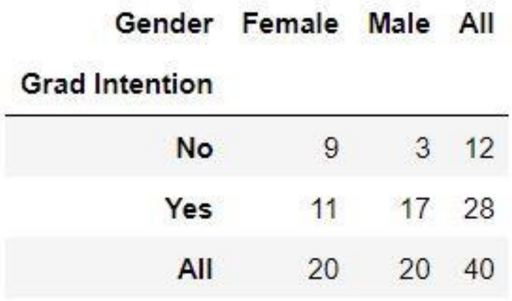
2.5.1 Find the probability that a randomly chosen student is a male or has full-time employment?

The probability of a randomly chosen student is a male is 0.46774193548387094. The probability of a randomly chosen student is a full-time employee is 0.16129032258064516. As a result, the probability of a randomly chosen student is a male or has full-time employment is 0.5161290322580645 based on calculations in jupyter notebook.

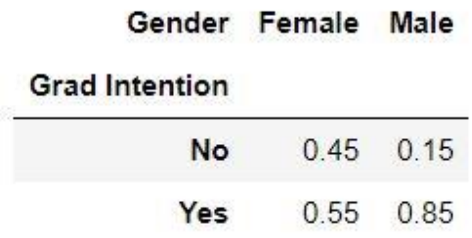
2.5.2 Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management.

The probability of a randomly chosen female student is majoring in International Business or Management is 0.24242424242424243.

2.6 Construct a contingency table of Gender and Intent to Graduate at 2 levels (Yes/No). The Undecided students are not considered now and the table is a 2x2 table. Do you think the graduate intention and being female are independent events?



Gender	Female	Male	All
Grad Intention			
No	9	3	12
Yes	11	17	28
All	20	20	40



Gender	Female	Male
Grad Intention		
No	0.45	0.15
Yes	0.55	0.85

The probability of a randomly selected student is a female is 0.5 and the probability of a randomly selected female student intends to graduate is 0.55. The probabilities of the two events are not equal owing to which conclude that they are independent.

Back to Top

2.7 Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. Answer the following questions based on the data:

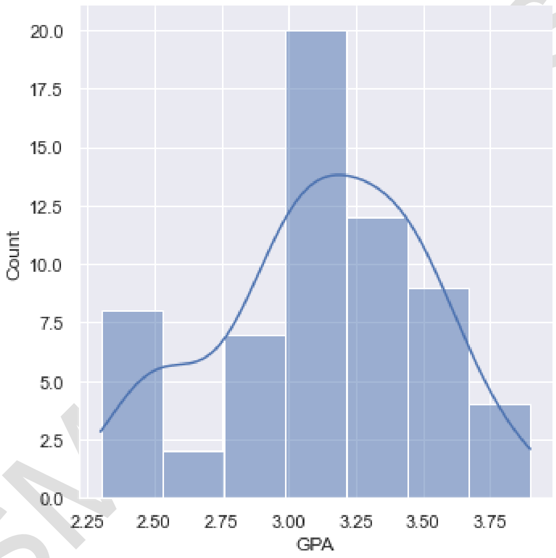
2.7.1 If a student is chosen randomly, what is the probability that his/her GPA is less than 3?

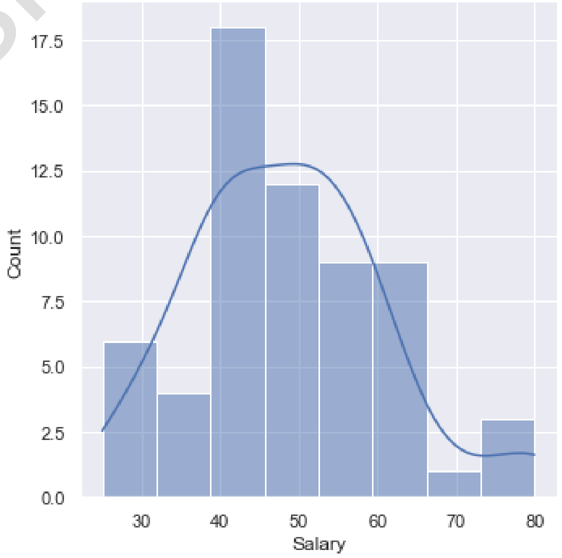
The probability of a randomly selected student has less than 3.0 GPA is 0.27419354838709675

2.7.2 Find the conditional probability that a randomly selected male earns 50 or more. Find the conditional probability that a randomly selected female earns 50 or more.

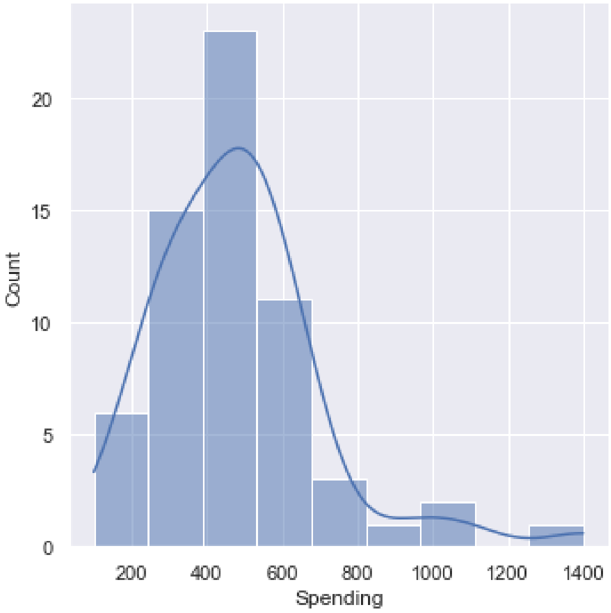
The probability of a randomly selected male student earns 50.0 or more is 0.4827586206896552 and the probability of a selected female student earns 50.0 or more 0.5454545454545454.

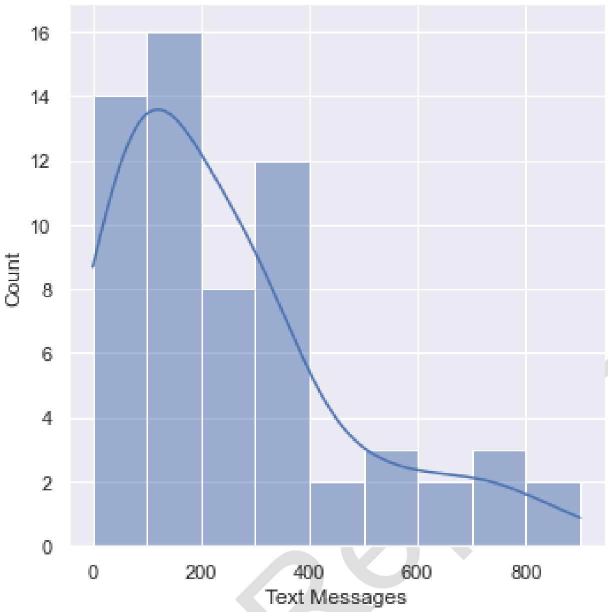
2.8 Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarizing your conclusions.





Back to Top





In line with the above histograms, we can conclude that GPA and Salary do not follow normal distribution; however, they are closer to normal distribution whereas Spending and Text Messages are right-skewed or positively skewed and do not show patterns of normal distribution.

Chapter 3. Problem 3: ABC Asphalt Shingles Analysis

3.1 Problem Statement

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 coloring 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 are calculated. The company would like to show that the mean moisture content is less than 0.35 pounds per 100 square feet.

3.2 Do you think there is evidence that means moisture contents in both types of shingles are within the permissible limits? State your conclusions clearly showing all steps.

3.2.1 Shingle Type A

3.2.1.1 Define Hypothesis

- Null Hypothesis or H0: Shingle Type A is within the permissible limit that is 0.35
- Alternate Hypothesis or H1: Shingle Type A is not within the permissible limit that is 0.35

3.2.1.2 Define Value of Alpha or Level of Significance

- 0.05

3.2.1.3 Determine the Test

- One sample t test

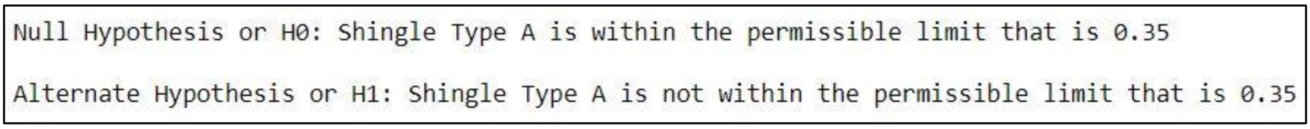
3.2.1.4 Calculate the p Value

- p Value: 0.07477633144907513

3.2.1.5 Conclusion

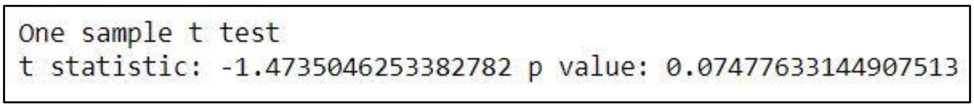
- Level of significance: 0.05
- Our one-sample t-test p-value = 0.07477633144907513
- We have no evidence to reject the null hypothesis since p value is more than level of significance

Python Output:

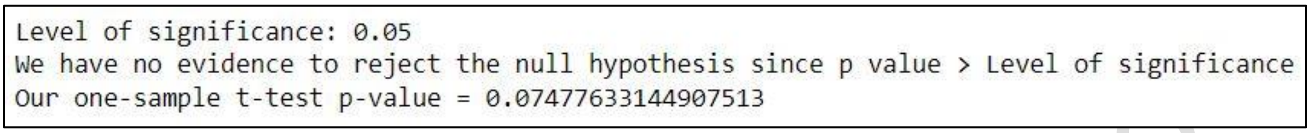


```
Null Hypothesis or H0: Shingle Type A is within the permissible limit that is 0.35

Alternate Hypothesis or H1: Shingle Type A is not within the permissible limit that is 0.35
```



```
One sample t test
t statistic: -1.4735046253382782 p value: 0.07477633144907513
```



```
Level of significance: 0.05
We have no evidence to reject the null hypothesis since p value > Level of significance
Our one-sample t-test p-value = 0.07477633144907513
```

3.2.2 Shingle Type B

3.2.2.1 Define Hypothesis

- Null Hypothesis or H0: Shingle Type B is within the permissible limit that is 0.35
- Alternate Hypothesis or H1: Shingle Type B is not within the permissible limit that is 0.35

3.2.2.2 Define Value of Alpha or Level of Significance

- 0.05

3.2.2.3 Determine the Test

- One sample t test

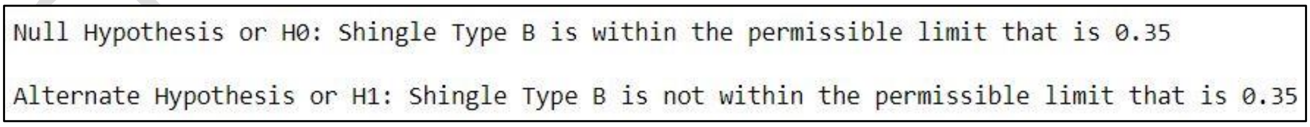
3.2.2.4 Calculate the p Value

- p Value: 0.0020904774003191826

3.2.2.5 Conclusion

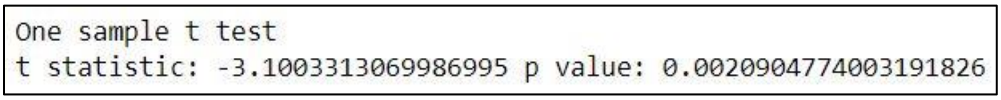
- Level of significance: 0.05
- Our one-sample t-test p-value = 0.0020904774003191826
- We have evidence to reject the null hypothesis since p value is less than level of significance

Python Output:

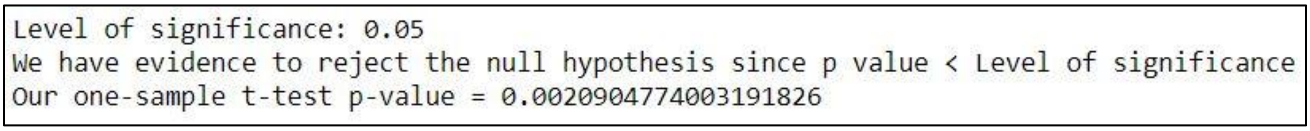


```
Null Hypothesis or H0: Shingle Type B is within the permissible limit that is 0.35

Alternate Hypothesis or H1: Shingle Type B is not within the permissible limit that is 0.35
```



```
One sample t test
t statistic: -3.1003313069986995 p value: 0.0020904774003191826
```



```
Level of significance: 0.05
We have evidence to reject the null hypothesis since p value < Level of significance
Our one-sample t-test p-value = 0.0020904774003191826
```

3.2.3 Analysis

When we calculate the p value for both the type of shingles A & B, we can conclude that since the p value for A shingle is less than the level of significance or alpha value, population mean moisture content is within the permissible limit. Whereas, p value for shingle type B is lower than the alpha value, as a result, population mean moisture content is not within the permissible limit which is 0.35 pound per 100 square feet.

3.3 Do you think that the population mean 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?

3.3.1 Define Hypothesis

- Null Hypothesis or H0: Population Mean of Type A Shingle = Population Mean of Type B Shingle
- Alternate Hypothesis or H1: Population Mean of Type A Shingle != Population Mean of Type B Shingle

3.3.2 Assumptions

- The datasets are normally distributed
- The variances of both the datasets are equal
- Both the datasets are independent from each other

3.3.3 Define Value of Alpha or Level of Significance

- 0.05

3.3.4 Determine the Test

- Two independent sample t test

3.3.5 Calculate the p Value

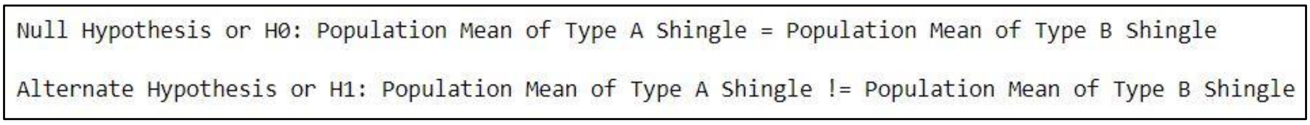
- p Value: 0.2017496571835306

3.3.6 Conclusion

- Level of significance: 0.05
- Our one-sample t-test p-value = 0.2017496571835306
- We do not have enough evidence to reject the null hypothesis in favour of alternative hypothesis

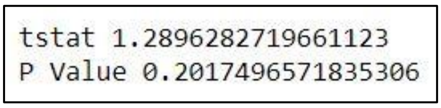
Back to Top

Python Output:

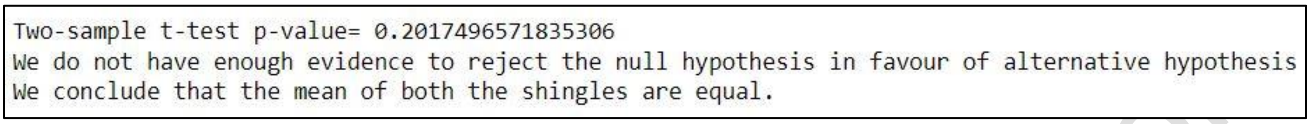


```
Null Hypothesis or H0: Population Mean of Type A Shingle = Population Mean of Type B Shingle

Alternate Hypothesis or H1: Population Mean of Type A Shingle != Population Mean of Type B Shingle
```



```
tstat 1.2896282719661123
P Value 0.2017496571835306
```



```
Two-sample t-test p-value= 0.2017496571835306
We do not have enough evidence to reject the null hypothesis in favour of alternative hypothesis
We conclude that the mean of both the shingles are equal.
```

3.3.7 Analysis

While using the two-sample t test for independent samples, we can arrive at a p value of 0.20175 which is more than the alpha value. As a result, we can conclude that the population mean of shingle type A is equal to the population mean of shingle type B.

Back to Top