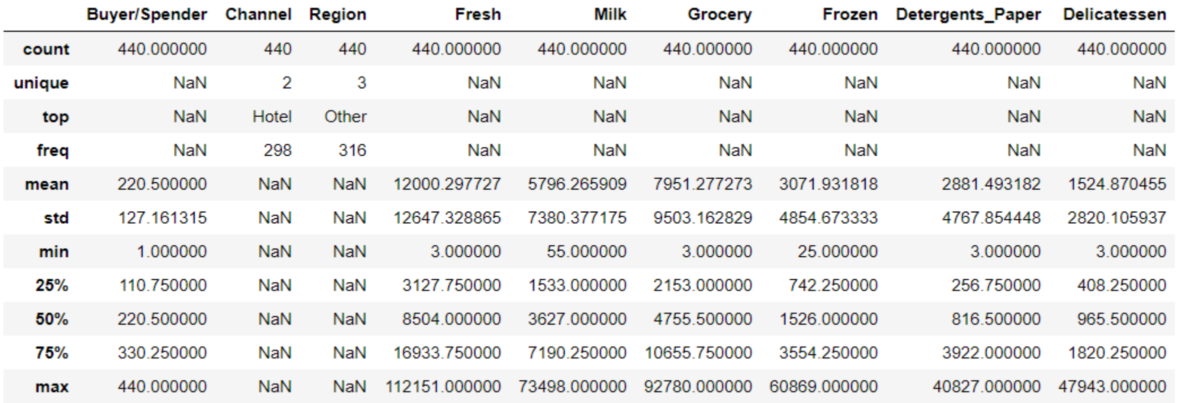
Contents

Problem 1

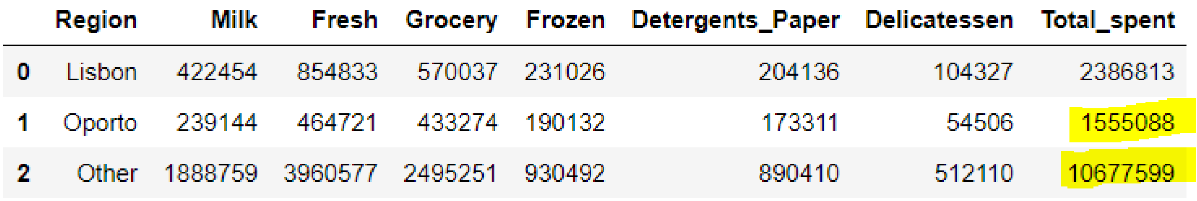
A wholesale distributor operating in different regions of Portugal has information on annual spending of several items in their stores across different regions and channels. The data consists of 440 large retailers' annual spending on 6 different varieties of products in 3 different regions (Lisbon,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

1.1 Use methods of descriptive statistics to summarize data. Which Region and which Channel spent the most? Which Region and which Channel spent the least?



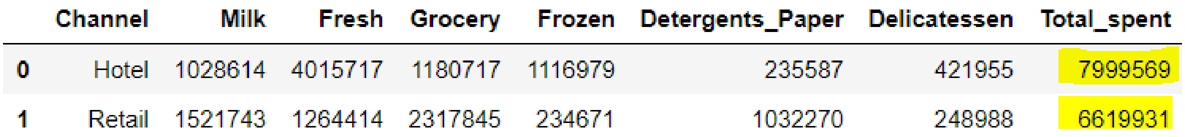
	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440.000000	440	440	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
unique	NaN	2	3	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	Hotel	Other	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	298	316	NaN	NaN	NaN	NaN	NaN	NaN
mean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

The descriptive statistics of the data is shown above.



	Region	Milk	Fresh	Grocery	Frozen	Detergents_Paper	Delicatessen	Total_spent
0	Lisbon	422454	854833	570037	231026	204136	104327	2386813
1	Oporto	239144	464721	433274	190132	173311	54506	1555088
2	Other	1888759	3960577	2495251	930492	890410	512110	10677599

From the above data, we can see that the Region "Other" spends the most and the Region "Oporto" spends the least.



	Channel	Milk	Fresh	Grocery	Frozen	Detergents_Paper	Delicatessen	Total_spent
0	Hotel	1028614	4015717	1180717	1116979	235587	421955	7999569
1	Retail	1521743	1264414	2317845	234671	1032270	248988	6619931

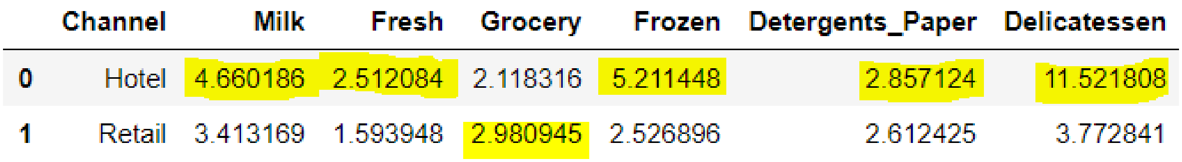
From the above data, we can see that the Channel "Hotel" spends more than the Channel "Retail".

So, we can conclude that "Other" region and "Hotel" channel spent the most and "Oporto" region and "Retail" channel spent the least.

1.2 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.

We will look upon skewness and use describe function to analyse and explain the data.

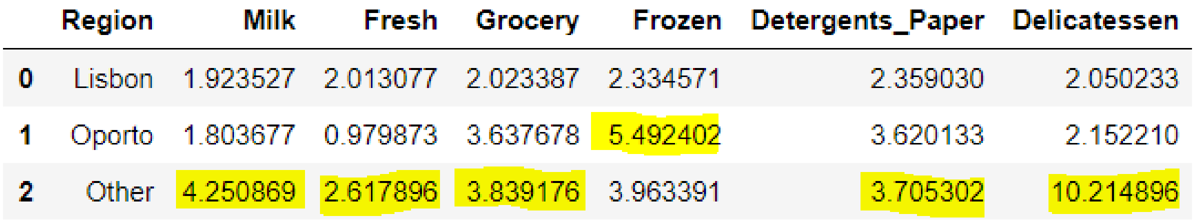
- Channel Specific:



	Channel	Milk	Fresh	Grocery	Frozen	Detergents_Paper	Delicatessen
0	Hotel	4.660186	2.512084	2.118316	5.211448	2.857124	11.521808
1	Retail	3.413169	1.593948	2.980945	2.526896	2.612425	3.772841

From the above data, we can see that all are right skewed. We can see the maximum skewness in Delicatessen from channel called “Hotel” and the minimum skewness in Fresh from a channel called “Retail”

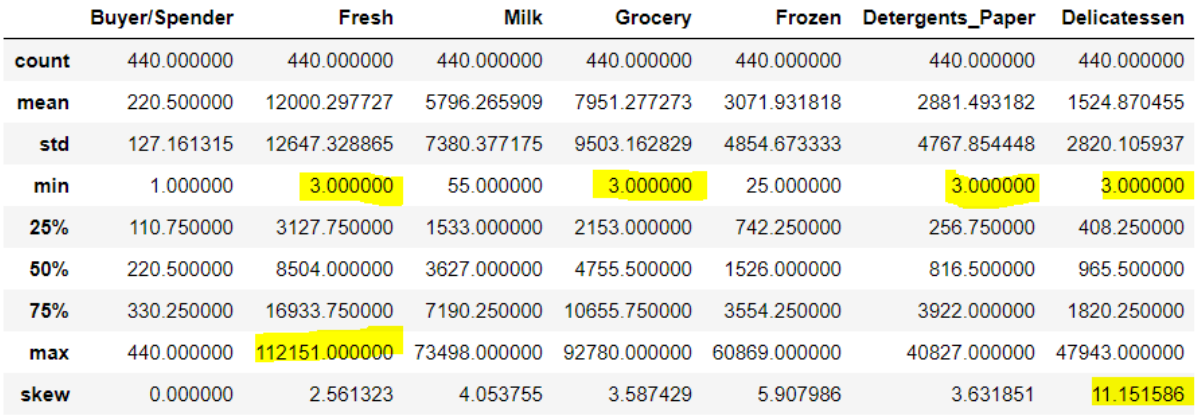
- Region Specific:



	Region	Milk	Fresh	Grocery	Frozen	Detergents_Paper	Delicatessen
0	Lisbon	1.923527	2.013077	2.023387	2.334571	2.359030	2.050233
1	Oporto	1.803677	0.979873	3.637678	5.492402	3.620133	2.152210
2	Other	4.250869	2.617896	3.839176	3.963391	3.705302	10.214896

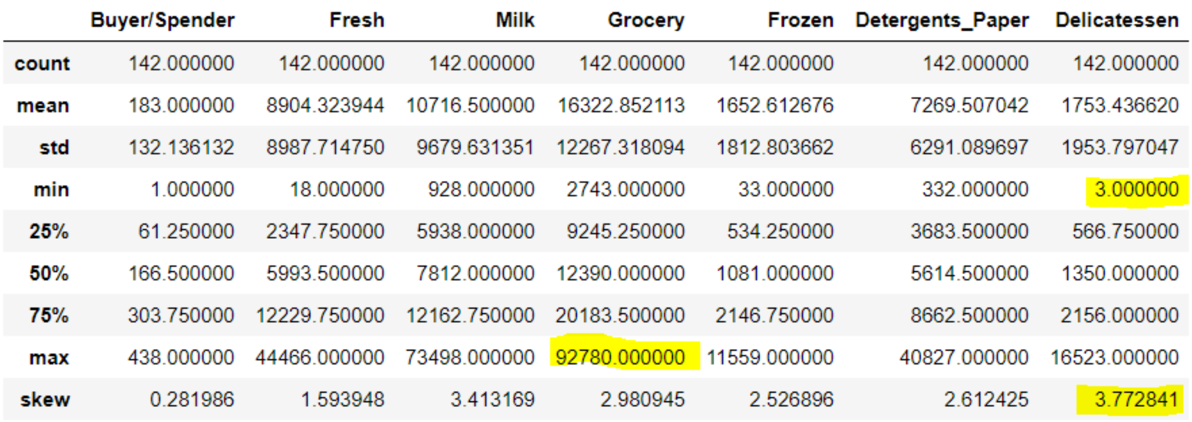
From the above data, we can see that all are right skewed. We can see the maximum skewness in Delicatessen from region called “Other” and the minimum skewness in fresh from a region called “Oporto”.

- Channel Hotel:



	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
skew	0.000000	2.561323	4.053755	3.587429	5.907986	3.631851	11.151586

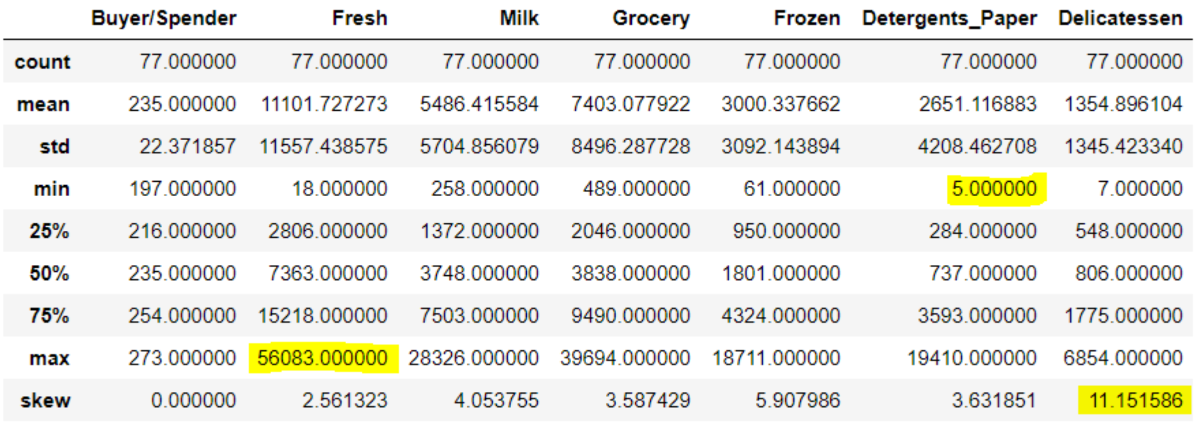
- Channel Retail:



	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	142.000000	142.000000	142.000000	142.000000	142.000000	142.000000	142.000000
mean	183.000000	8904.323944	10716.500000	16322.852113	1652.612676	7269.507042	1753.436620
std	132.136132	8987.714750	9679.631351	12267.318094	1812.803662	6291.089697	1953.797047
min	1.000000	18.000000	928.000000	2743.000000	33.000000	332.000000	3.000000
25%	61.250000	2347.750000	5938.000000	9245.250000	534.250000	3683.500000	566.750000
50%	166.500000	5993.500000	7812.000000	12390.000000	1081.000000	5614.500000	1350.000000
75%	303.750000	12229.750000	12162.750000	20183.500000	2146.750000	8662.500000	2156.000000
max	438.000000	44466.000000	73498.000000	92780.000000	11559.000000	40827.000000	16523.000000
skew	0.281986	1.593948	3.413169	2.980945	2.526896	2.612425	3.772841

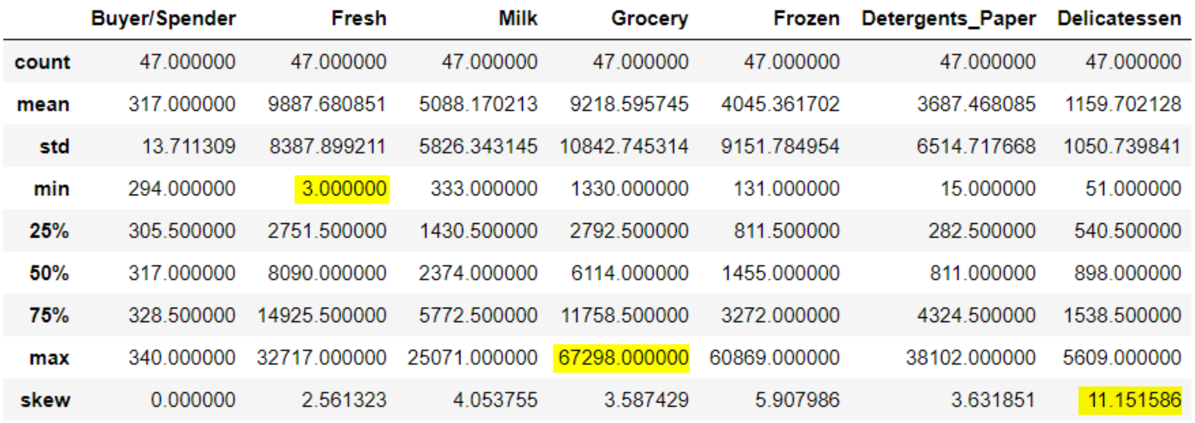
Based on the analysis of the individual channels, we can see that Hotel spends most on Fresh and least in Grocery, Detergents_paper and Delicatessen. Retail spends most on Grocery and least in Delicatessen. We can also see that, Delicatessen has most skewness in both the channels.

- Region Lisbon:



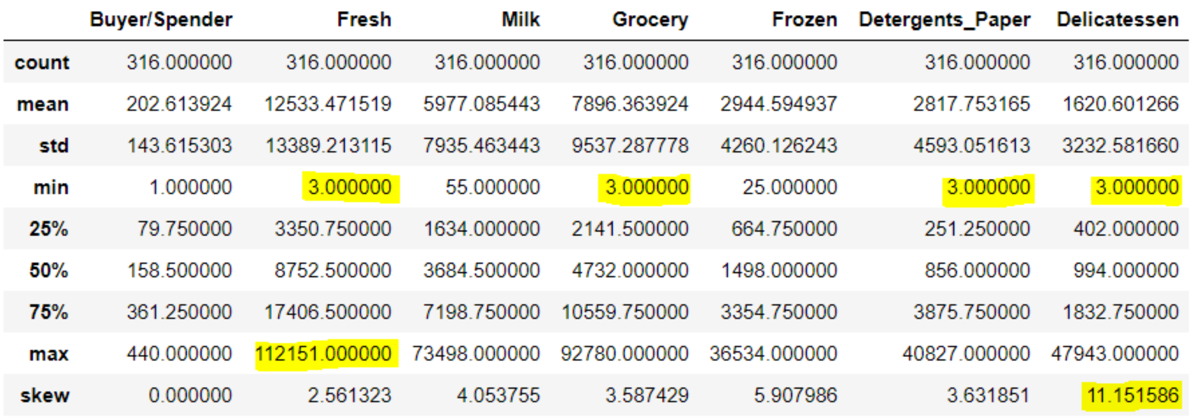
	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000
mean	235.000000	11101.727273	5486.415584	7403.077922	3000.337662	2651.116883	1354.896104
std	22.371857	11557.438575	5704.856079	8496.287728	3092.143894	4208.462708	1345.423340
min	197.000000	18.000000	258.000000	489.000000	61.000000	5.000000	7.000000
25%	216.000000	2806.000000	1372.000000	2046.000000	950.000000	284.000000	548.000000
50%	235.000000	7363.000000	3748.000000	3838.000000	1801.000000	737.000000	806.000000
75%	254.000000	15218.000000	7503.000000	9490.000000	4324.000000	3593.000000	1775.000000
max	273.000000	56083.000000	28326.000000	39694.000000	18711.000000	19410.000000	6854.000000
skew	0.000000	2.561323	4.053755	3.587429	5.907986	3.631851	11.151586

- Region Oporto:



	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	47.000000	47.000000	47.000000	47.000000	47.000000	47.000000	47.000000
mean	317.000000	9887.680851	5088.170213	9218.595745	4045.361702	3687.468085	1159.702128
std	13.711309	8387.899211	5826.343145	10842.745314	9151.784954	6514.717668	1050.739841
min	294.000000	3.000000	333.000000	1330.000000	131.000000	15.000000	51.000000
25%	305.500000	2751.500000	1430.500000	2792.500000	811.500000	282.500000	540.500000
50%	317.000000	8090.000000	2374.000000	6114.000000	1455.000000	811.000000	898.000000
75%	328.500000	14925.500000	5772.500000	11758.500000	3272.000000	4324.500000	1538.500000
max	340.000000	32717.000000	25071.000000	67298.000000	60869.000000	38102.000000	5609.000000
skew	0.000000	2.561323	4.053755	3.587429	5.907986	3.631851	11.151586

- Region Other:



	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	316.000000	316.000000	316.000000	316.000000	316.000000	316.000000	316.000000
mean	202.613924	12533.471519	5977.085443	7896.363924	2944.594937	2817.753165	1620.601266
std	143.615303	13389.213115	7935.463443	9537.287778	4260.126243	4593.051613	3232.581660
min	1.000000	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	79.750000	3350.750000	1634.000000	2141.500000	664.750000	251.250000	402.000000
50%	158.500000	8752.500000	3684.500000	4732.000000	1498.000000	856.000000	994.000000
75%	361.250000	17406.500000	7198.750000	10559.750000	3354.750000	3875.750000	1832.750000
max	440.000000	112151.000000	73498.000000	92780.000000	36534.000000	40827.000000	47943.000000
skew	0.000000	2.561323	4.053755	3.587429	5.907986	3.631851	11.151586

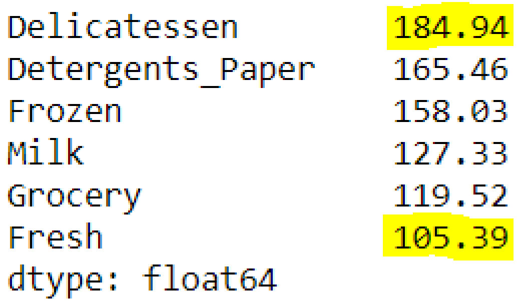
We can see from the data that Region Other and Region Lisbon spends more on fresh and Region Oporto spends more on Grocery.

We also know that Channel Hotel spends more on fresh and Channel Retail spends more on Grocery.

Hence, it is safe to assume that Region Other and Region Lisbon consists of a greater number of channel Hotel. Region Oporto consists of a greater number of Channel Retail.

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

Coefficient of variation can give us information regarding the consistency. So, lets find out the coefficient of variation of each item.

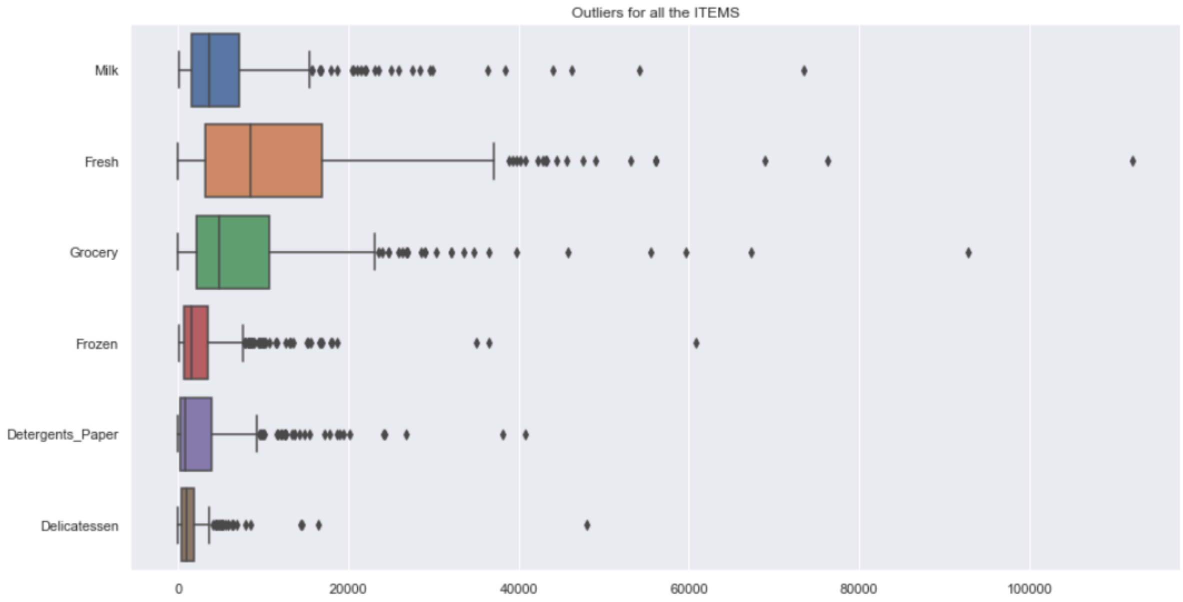


```
Delicatessen        184.94
Detergents_Paper    165.46
Frozen              158.03
Milk                127.33
Grocery             119.52
Fresh               105.39
dtype: float64
```

From the above data, we can say that Delicatessen is the most Consistent and Fresh is the least Consistent.

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

A box plot can help us in showing if the data has any outliers or not. So, lets plot a box plot to show the outliers present in the data.



From the above box plot, we can see that all the 6 items contain outliers. The outliers are however under the excessive category. Only a few are outside the limit.

1.5 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

From Problem 1.2, we can infer that Region Other and Region Lisbon consists of a greater number of channel Hotel. Region Oporto consists of a greater number of Channel Retail.

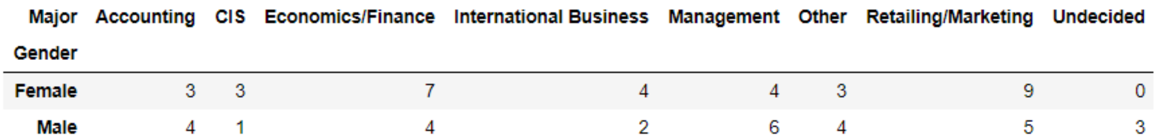
We also know that channel Hotel spends more on Fresh and Channel Retail spends more on Grocery. So, it would be better if Region other and Region Lisbon also sells Fresh because there is a greater number of channel hotels in those 2 regions. And it would be better if Region Oporto sells grocery because there is a greater number of Channel Retail present.

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.

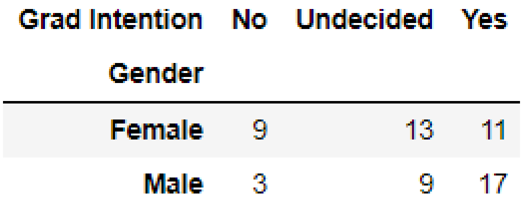
2.1. For this data, construct the following contingency tables (Keep Gender as row variable)

2.1.1. Gender and Major



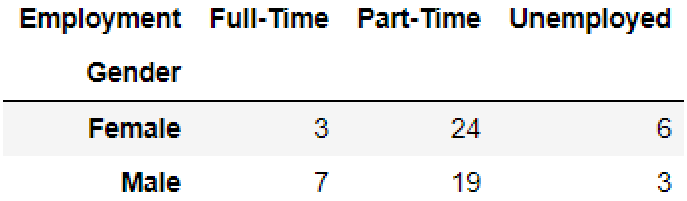
Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Gender								
Female	3	3	7	4	4	3	9	0
Male	4	1	4	2	6	4	5	3

2.1.2 Gender and Grad Intention



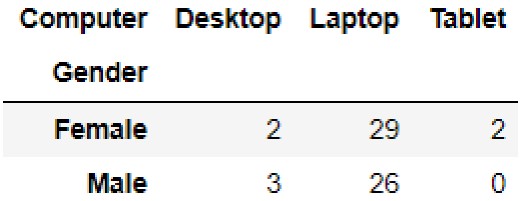
Grad Intention	No	Undecided	Yes
Gender			
Female	9	13	11
Male	3	9	17

2.1.3 Gender and Employment



Employment	Full-Time	Part-Time	Unemployed
Gender			
Female	3	24	6
Male	7	19	3

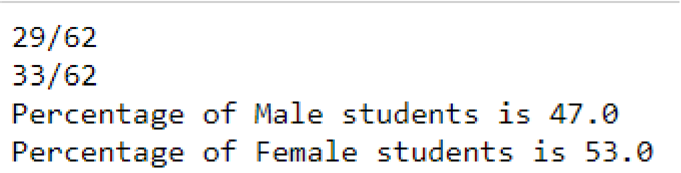
2.1.4 Gender and Computer



Computer	Desktop	Laptop	Tablet
Gender			
Female	2	29	2
Male	3	26	0

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?

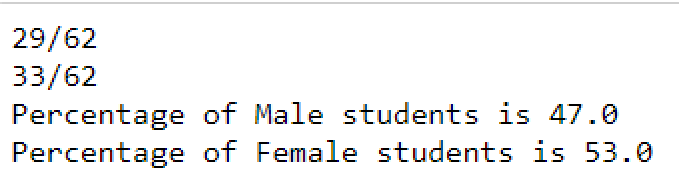


```
29/62
33/62
Percentage of Male students is 47.0
Percentage of Female students is 53.0
```

The Total Male gender count is 29. Dividing that with the total number gives us the probability.

Male = 29 divided by 62 gives 0.46 and that multiplied with 100 gives 47 which is the probability that a randomly selected CMSU student will be male.

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



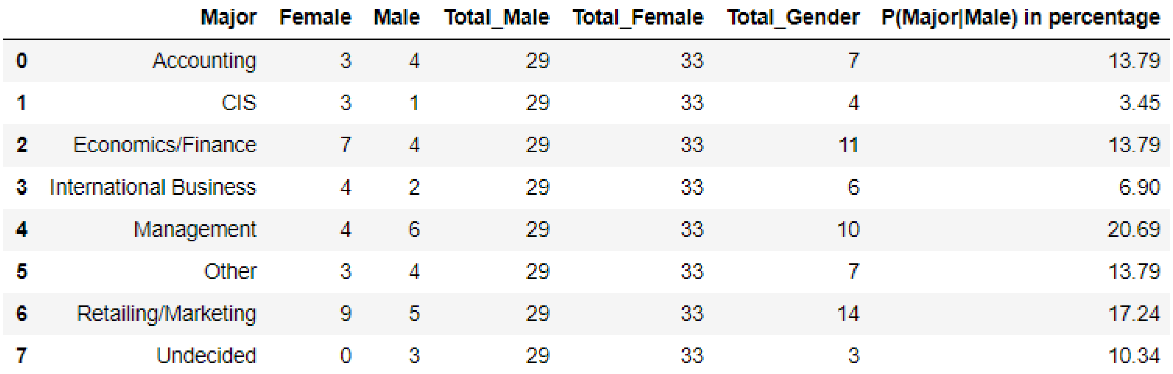
```
29/62
33/62
Percentage of Male students is 47.0
Percentage of Female students is 53.0
```

The total Female gender count is 33. Dividing that with the total number gives us the probability.

Female = 33 divided by 62 gives 0.53 and that multiplied with 100 gives 53 which is the probability that a randomly selected CMSU student will be Female.

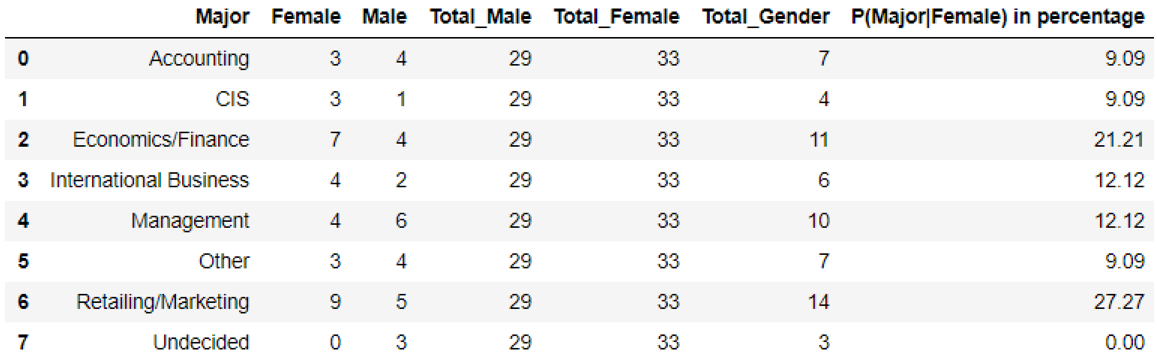
2.3. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

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



	Major	Female	Male	Total_Male	Total_Female	Total_Gender	P(Major|Male) in percentage
0	Accounting	3	4	29	33	7	13.79
1	CIS	3	1	29	33	4	3.45
2	Economics/Finance	7	4	29	33	11	13.79
3	International Business	4	2	29	33	6	6.90
4	Management	4	6	29	33	10	20.69
5	Other	3	4	29	33	7	13.79
6	Retailing/Marketing	9	5	29	33	14	17.24
7	Undecided	0	3	29	33	3	10.34

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



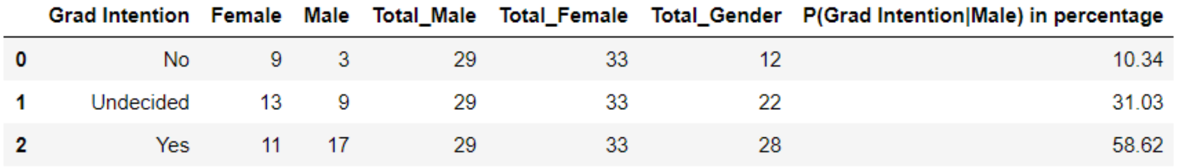
	Major	Female	Male	Total_Male	Total_Female	Total_Gender	P(Major|Female) in percentage
0	Accounting	3	4	29	33	7	9.09
1	CIS	3	1	29	33	4	9.09
2	Economics/Finance	7	4	29	33	11	21.21
3	International Business	4	2	29	33	6	12.12
4	Management	4	6	29	33	10	12.12
5	Other	3	4	29	33	7	9.09
6	Retailing/Marketing	9	5	29	33	14	27.27
7	Undecided	0	3	29	33	3	0.00

To find the conditional probability of different majors among the student of CMSU:

Divide the Number of students (Male/Female) in a particular Major by Total number of students (Male/Female).

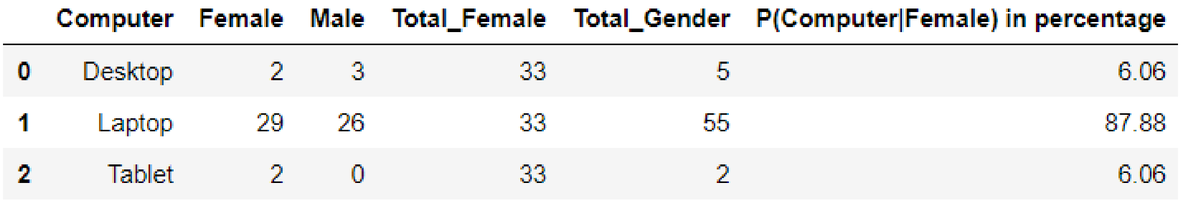
2.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:

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



	Grad Intention	Female	Male	Total_Male	Total_Female	Total_Gender	P(Grad Intention|Male) in percentage
0	No	9	3	29	33	12	10.34
1	Undecided	13	9	29	33	22	31.03
2	Yes	11	17	29	33	28	58.62

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



	Computer	Female	Male	Total_Female	Total_Gender	P(Computer|Female) in percentage
0	Desktop	2	3	33	5	6.06
1	Laptop	29	26	33	55	87.88
2	Tablet	2	0	33	2	6.06

The probability that a randomly selected student is female and does not have a laptop is 100-87 which is 13%.

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 either a male or has full-time employment?

Using contingency tables for Gender and Employment, we get the total number of males who are full time employed.

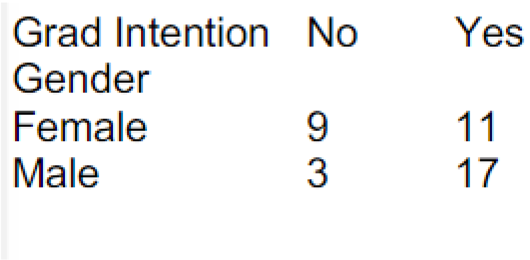
And after calculating we get the probability is 74%.

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

Using contingency tables for Gender and Major, we get the total number of females and number of females majoring in international business or management.

And after calculating we get the probability is 24%.

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?



```
Grad Intention  No    Yes
Gender
Female          9     11
Male            3     17
```

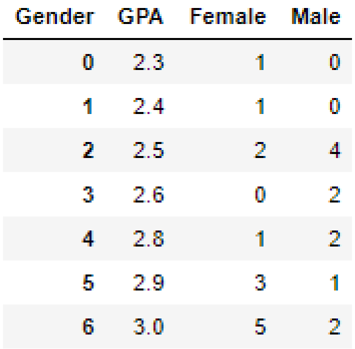
P(Grad Intention Yes) = 28/40 = 0.7

P(Grad Intention Yes | female) = 11 / 20 = 0.55

The probabilities are not equal. Hence, these 2 are independent events.

2.7. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages.

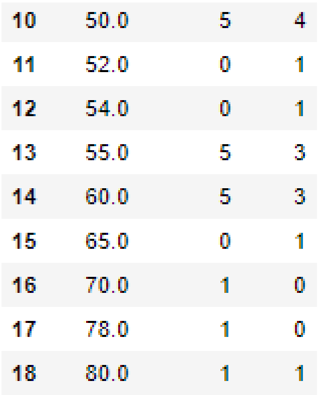
2.7.1. If a student is chosen randomly, what is the probability that his/her GPA is less than 3?



Gender	GPA	Female	Male
0	2.3	1	0
1	2.4	1	0
2	2.5	2	4
3	2.6	0	2
4	2.8	1	2
5	2.9	3	1
6	3.0	5	2

Using contingency tables of Gender and GPA we got the total numbers of students and number of students GPA less than 3. And after calculation we find out that - Probability that student is chosen randomly and that his/her GPA is less than 3 is 22.58%

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.

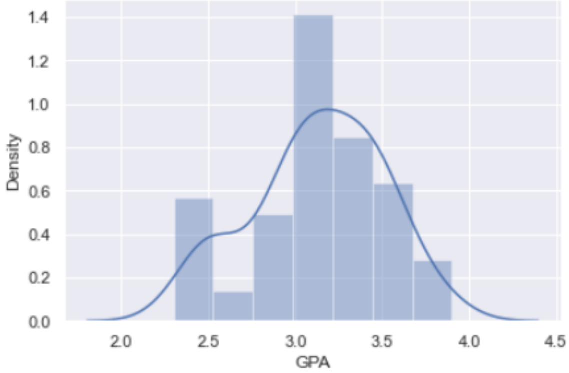


10	50.0	5	4
11	52.0	0	1
12	54.0	0	1
13	55.0	5	3
14	60.0	5	3
15	65.0	0	1
16	70.0	1	0
17	78.0	1	0
18	80.0	1	1

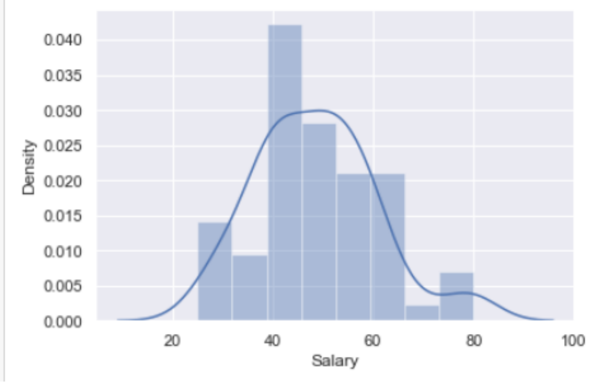
Using contingency tables of gender and Salary, we got the totalnumber of students and number of students whose salary is greater than 50. After calculation we find out that - Probability that randomly selected male earns 50 or more is 34.48% And Probability that a randomly selected female earns 50 or more is 30.3%

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 for this whole Problem 2.

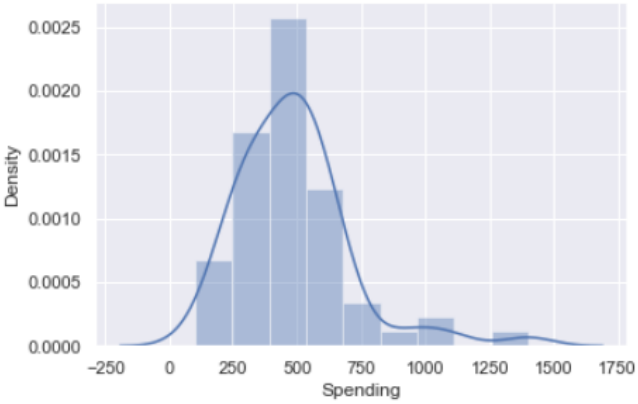
- GPA



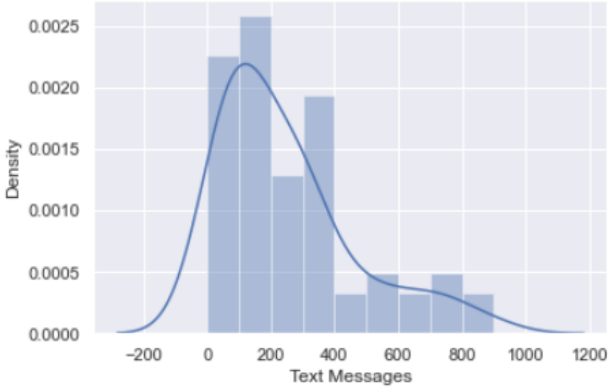
- Salary



- Spending



- Text messages



From the above graphs we can conclude that, GPA and salary are following Normal distribution. And Spending and Text messages are not following normal distribution.

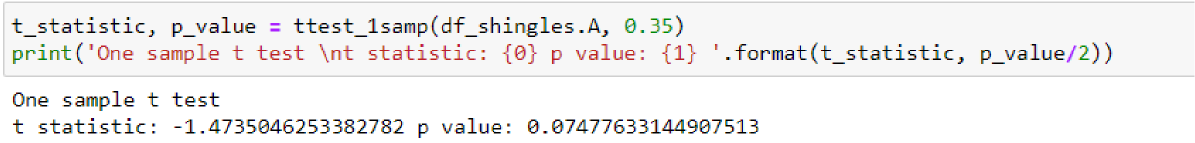
Although, none of these are perfectly normally distributed, but they show signs of normal distribution.

Conclusion: We applied various methods of conditional probabilities to get various insights into the data. We also got insights on how gender influences various parameters for example, the number of female students are more in the high GPA category and also number of female students getting salary more than 50 is more than male. Which shows, higher GPA is directly proportional to higher salary. There can be many more such insights derived from the data.

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.

3.1 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.

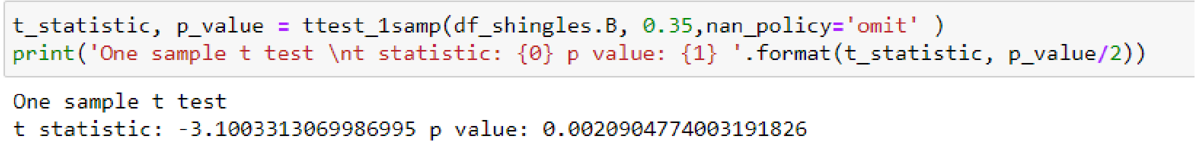


```
t_statistic, p_value = ttest_1samp(df_shingles.A, 0.35)
print('One sample t test \nt statistic: {0} p value: {1} '.format(t_statistic, p_value/2))
```

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

From the above data, we can see that the p value is more than 0.05. Hence, we do not reject H0.

The probability of observing a sample of 36 shingles that will result in a sample mean moisture content of 0.3167 pounds per 100 square feet or less is .0748.



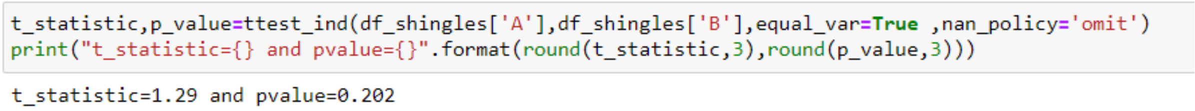
```
t_statistic, p_value = ttest_1samp(df_shingles.B, 0.35,nan_policy='omit' )
print('One sample t test \nt statistic: {0} p value: {1} '.format(t_statistic, p_value/2))
```

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

From the above data, we can see that the p value is less than 0.05. Hence, we reject H0.

The probability of observing 31 shingles that will result in a sample mean moisture content of 0.273 pounds per 100 square feet or less is .0021.

3.2 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?



```
t_statistic,p_value=ttest_ind(df_shingles['A'],df_shingles['B'],equal_var=True ,nan_policy='omit')
print("t_statistic={} and pvalue={}".format(round(t_statistic,3),round(p_value,3)))
```

```
t_statistic=1.29 and pvalue=0.202
```

Here we can see that the p value is greater than 0.05 and hence we do not reject H0.

The assumptions to be checked are that both the populations have normal distribution and variance of both the distributions are the same.