SMDM PROJECT

Statistical Methods for Decision Making

Ayyalasomayajula Aditya

Project - SMDM

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?

Solution:

□ The Region that spent the most is: **Other Total Spending: 1,06,77,599**

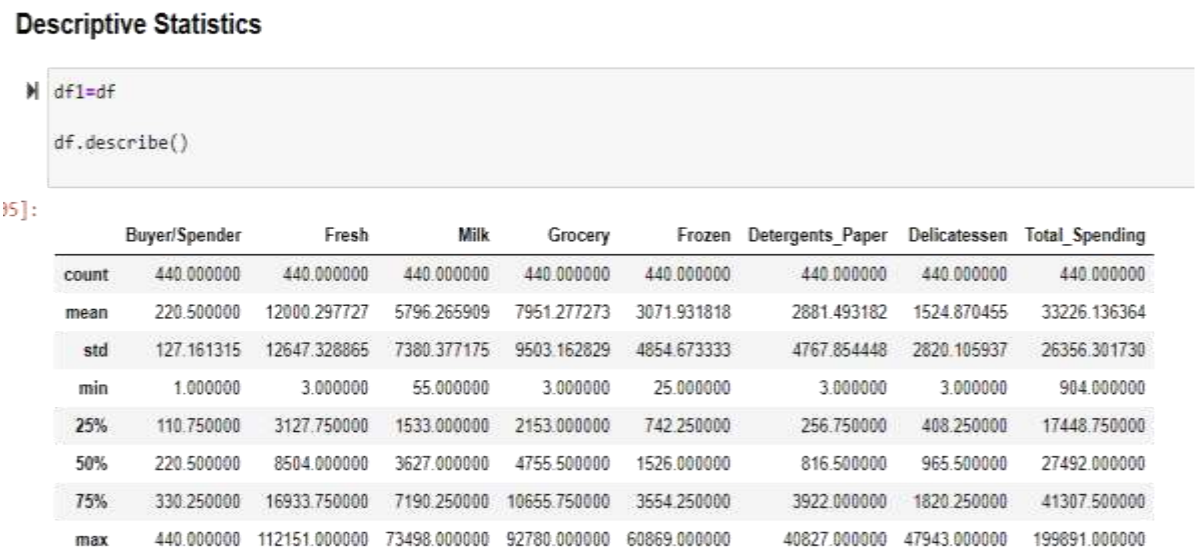
□ The Channel that spent the most is: **Retail Total Spending: 79,99,569**

□ The Region that spent the least is: **Other Total Spending: 15,55,088**

□ The Channel that spent the least is: **Hotel Total Spending: 66,19,931**

Supporting Code:

Descriptive Statistics:

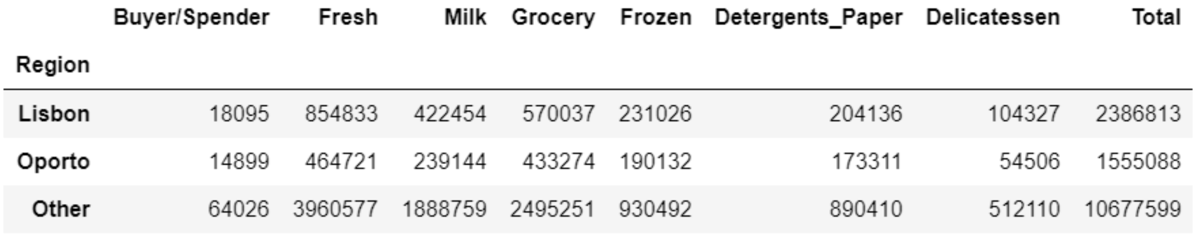


Descriptive Statistics

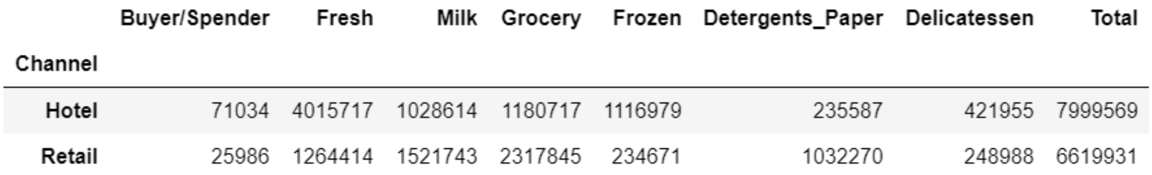
```
df1=df

df.describe()
```

```
]5]:
```

	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total_Spending
count	440.000000	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	33226.136364
std	127.161315	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937	26356.301730
min	1.000000	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000	904.000000
25%	110.750000	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000	17448.750000
50%	220.500000	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000	27492.000000
75%	330.250000	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000	41307.500000
max	440.000000	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000	199891.000000



	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
Region								
Lisbon	18095	854833	422454	570037	231026	204136	104327	2386813
Oporto	14899	464721	239144	433274	190132	173311	54506	1555088
Other	64026	3960577	1888759	2495251	930492	890410	512110	10677599

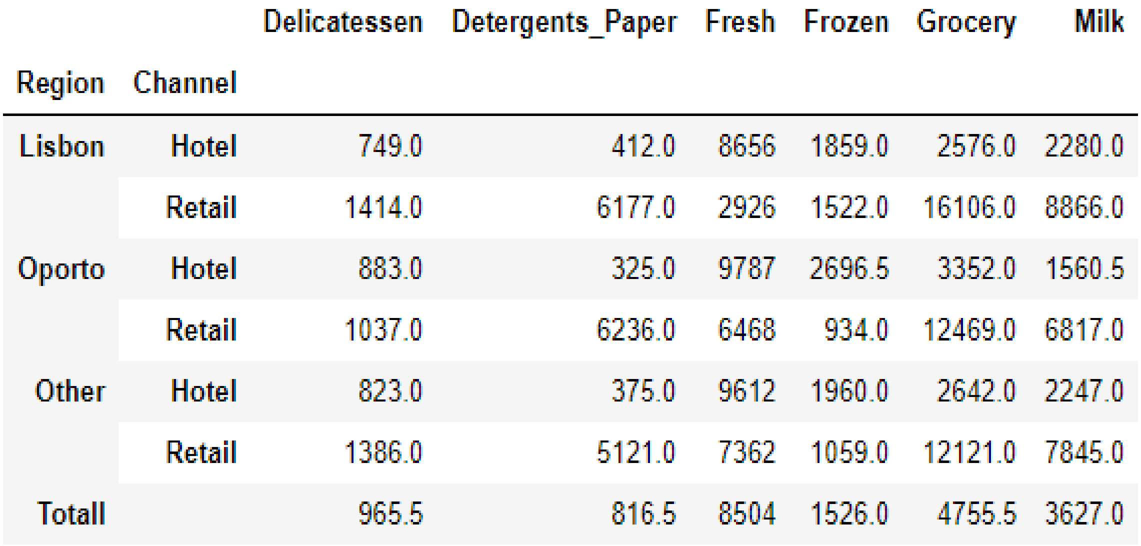


	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
Channel								
Hotel	71034	4015717	1028614	1180717	1116979	235587	421955	7999569
Retail	25986	1264414	1521743	2317845	234671	1032270	248988	6619931

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

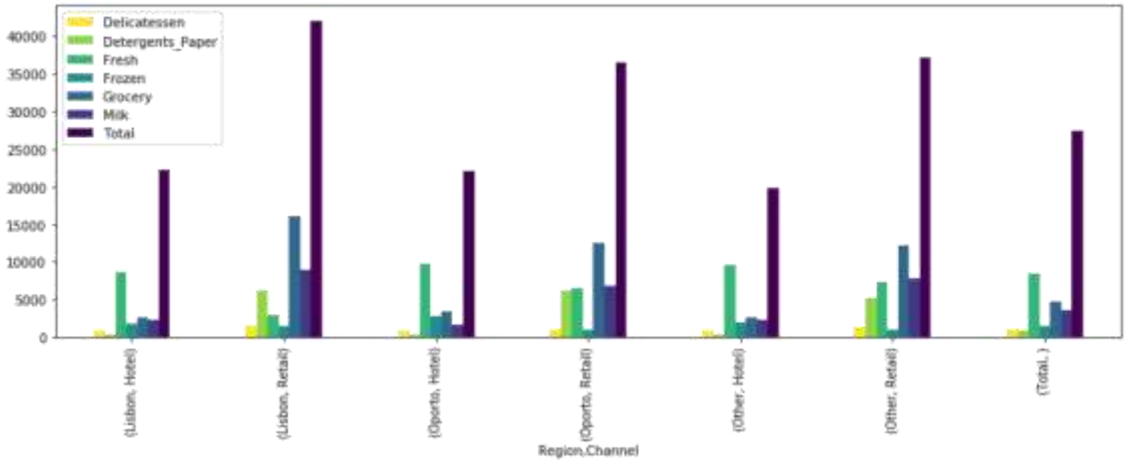
Solution:

Calculating Median across Region and Channel:

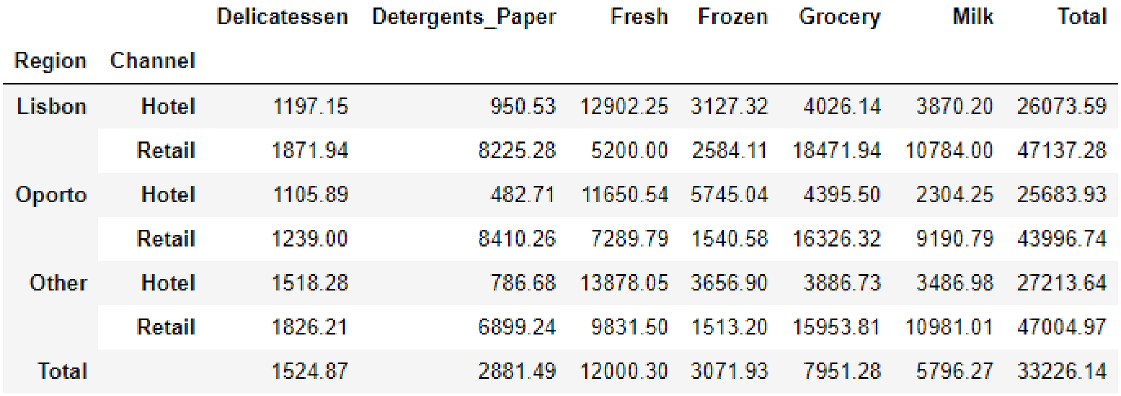


		Delicatessen	Detergents_Paper	Fresh	Frozen	Grocery	Milk
Region	Channel						
Lisbon	Hotel	749.0	412.0	8656	1859.0	2576.0	2280.0
	Retail	1414.0	6177.0	2926	1522.0	16106.0	8866.0
Oporto	Hotel	883.0	325.0	9787	2696.5	3352.0	1560.5
	Retail	1037.0	6236.0	6468	934.0	12469.0	6817.0
Other	Hotel	823.0	375.0	9612	1960.0	2642.0	2247.0
	Retail	1386.0	5121.0	7362	1059.0	12121.0	7845.0
Totall		965.5	816.5	8504	1526.0	4755.5	3627.0

Bar Plot:

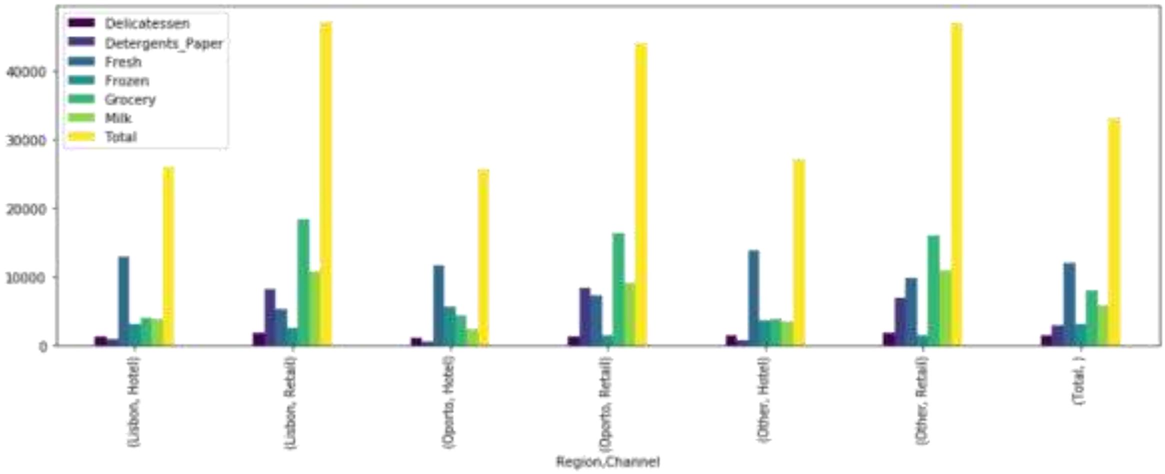


Calculating Mean across region and channel

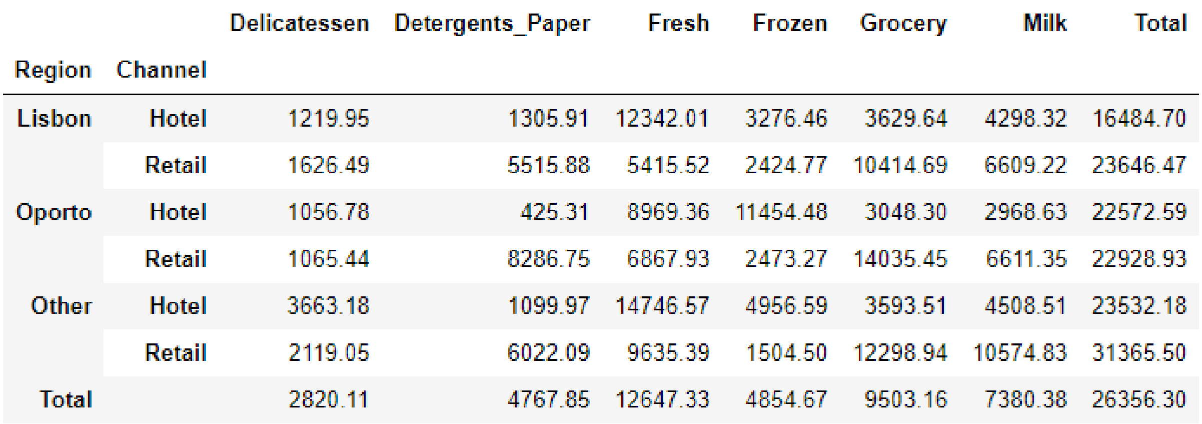


		Delicatessen	Detergents_Paper	Fresh	Frozen	Grocery	Milk	Total
Region	**Channel**							
Lisbon	**Hotel**	1197.15	950.53	12902.25	3127.32	4026.14	3870.20	26073.59
	Retail	1871.94	8225.28	5200.00	2584.11	18471.94	10784.00	47137.28
Oporto	**Hotel**	1105.89	482.71	11650.54	5745.04	4395.50	2304.25	25683.93
	Retail	1239.00	8410.26	7289.79	1540.58	16326.32	9190.79	43996.74
Other	**Hotel**	1518.28	786.68	13878.05	3656.90	3886.73	3486.98	27213.64
	Retail	1826.21	6899.24	9831.50	1513.20	15953.81	10981.01	47004.97
Total		1524.87	2881.49	12000.30	3071.93	7951.28	5796.27	33226.14

Bar plot:

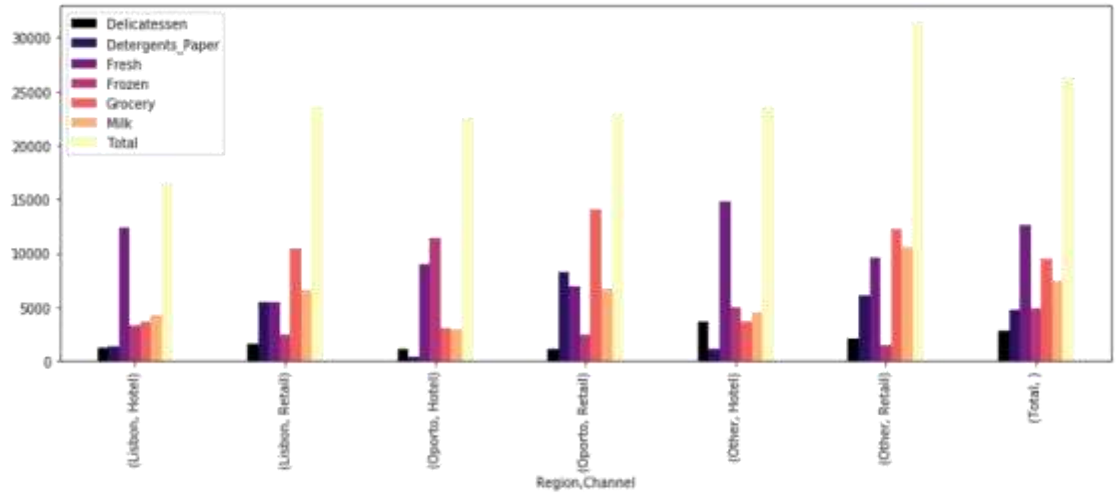


Calculating Standard Deviation across region and channel:

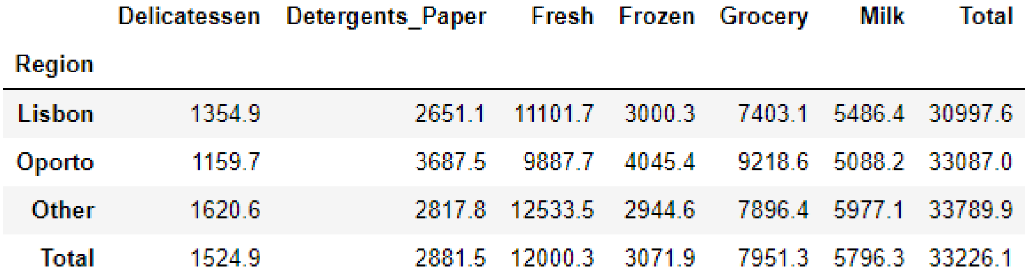


		Delicatessen	Detergents_Paper	Fresh	Frozen	Grocery	Milk	Total
Region	**Channel**							
Lisbon	**Hotel**	1219.95	1305.91	12342.01	3276.46	3629.64	4298.32	16484.70
	Retail	1626.49	5515.88	5415.52	2424.77	10414.69	6609.22	23646.47
Oporto	**Hotel**	1056.78	425.31	8969.36	11454.48	3048.30	2968.63	22572.59
	Retail	1065.44	8286.75	6867.93	2473.27	14035.45	6611.35	22928.93
Other	**Hotel**	3663.18	1099.97	14746.57	4956.59	3593.51	4508.51	23532.18
	Retail	2119.05	6022.09	9635.39	1504.50	12298.94	10574.83	31365.50
Total		2820.11	4767.85	12647.33	4854.67	9503.16	7380.38	26356.30

Bar plot:

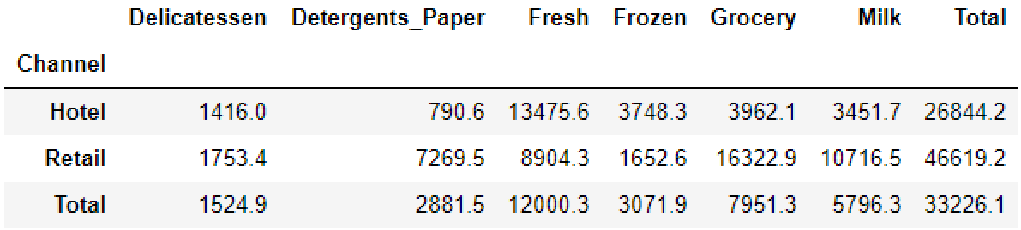


Calculating Mean across region:



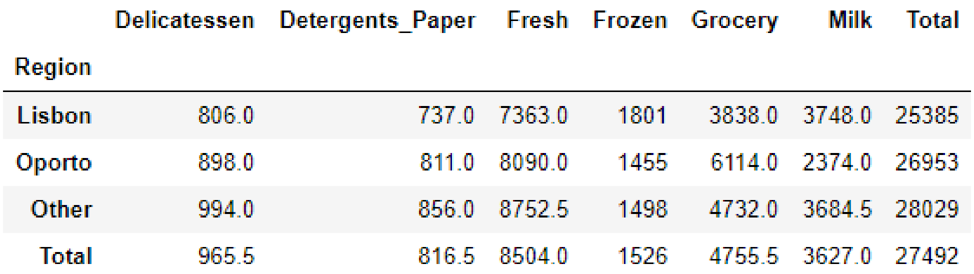
	Delicatessen	Detergents_Paper	Fresh	Frozen	Grocery	Milk	Total
Region							
Lisbon	1354.9	2651.1	11101.7	3000.3	7403.1	5486.4	30997.6
Oporto	1159.7	3687.5	9887.7	4045.4	9218.6	5088.2	33087.0
Other	1620.6	2817.8	12533.5	2944.6	7896.4	5977.1	33789.9
Total	1524.9	2881.5	12000.3	3071.9	7951.3	5796.3	33226.1

Calculating Mean across Channel:



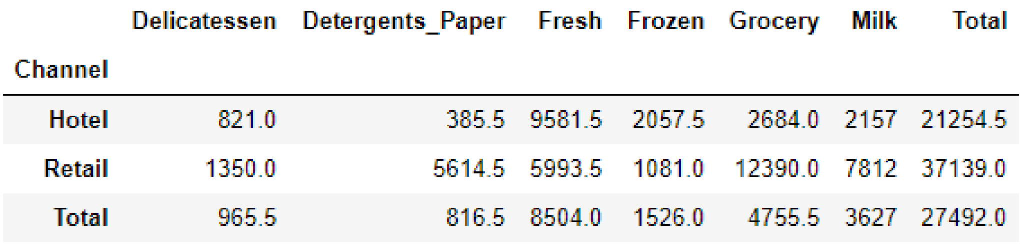
	Delicatessen	Detergents_Paper	Fresh	Frozen	Grocery	Milk	Total
Channel							
Hotel	1416.0	790.6	13475.6	3748.3	3962.1	3451.7	26844.2
Retail	1753.4	7269.5	8904.3	1652.6	16322.9	10716.5	46619.2
Total	1524.9	2881.5	12000.3	3071.9	7951.3	5796.3	33226.1

Calculating Median across region:



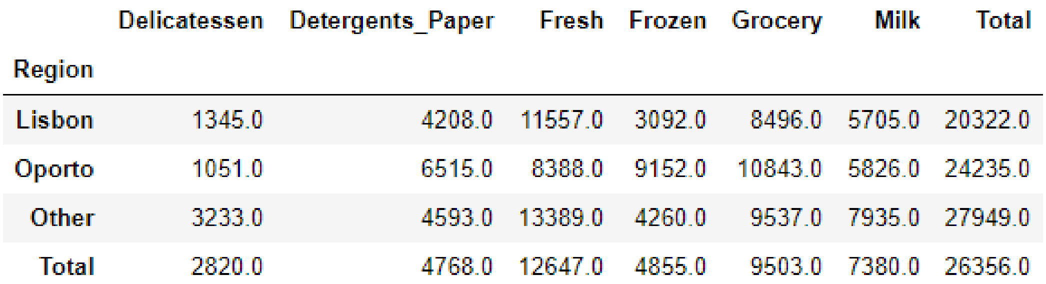
	Delicatessen	Detergents_Paper	Fresh	Frozen	Grocery	Milk	Total
Region							
Lisbon	806.0	737.0	7363.0	1801	3838.0	3748.0	25385
Oporto	898.0	811.0	8090.0	1455	6114.0	2374.0	26953
Other	994.0	856.0	8752.5	1498	4732.0	3684.5	28029
Total	965.5	816.5	8504.0	1526	4755.5	3627.0	27492

Calculating Median across Channel:



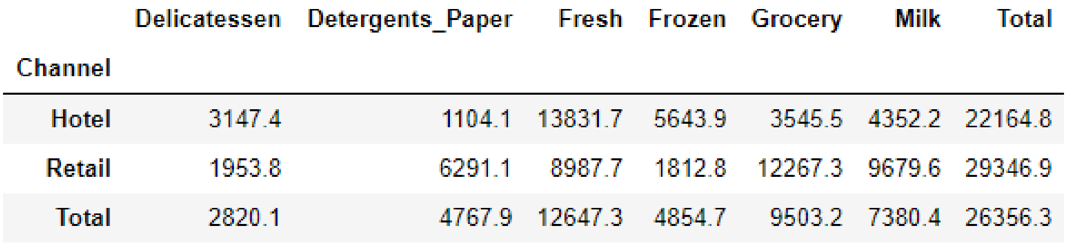
	Delicatessen	Detergents_Paper	Fresh	Frozen	Grocery	Milk	Total
Channel							
Hotel	821.0	385.5	9581.5	2057.5	2684.0	2157	21254.5
Retail	1350.0	5614.5	5993.5	1081.0	12390.0	7812	37139.0
Total	965.5	816.5	8504.0	1526.0	4755.5	3627	27492.0

Calculating Standard Deviation across region:



	Delicatessen	Detergents_Paper	Fresh	Frozen	Grocery	Milk	Total
Region							
Lisbon	1345.0	4208.0	11557.0	3092.0	8496.0	5705.0	20322.0
Oporto	1051.0	6515.0	8388.0	9152.0	10843.0	5826.0	24235.0
Other	3233.0	4593.0	13389.0	4260.0	9537.0	7935.0	27949.0
Total	2820.0	4768.0	12647.0	4855.0	9503.0	7380.0	26356.0

Calculating Standard Deviation across Channel:



	Delicatessen	Detergents_Paper	Fresh	Frozen	Grocery	Milk	Total
Channel							
Hotel	3147.4	1104.1	13831.7	5643.9	3545.5	4352.2	22164.8
Retail	1953.8	6291.1	8987.7	1812.8	12267.3	9679.6	29346.9
Total	2820.1	4767.9	12647.3	4854.7	9503.2	7380.4	26356.3

Conclusion:

On the basis of above descriptive Statistics we can say that:

Region wise analysis show for each category different mean and median values showing the central tendency of the data region wise is different Channel wise analysis show for each category different mean and median values showing the central tendency of the data channel wise is different Region wise analysis of standard deviation for each category show differing SDs region wise indicating heterogeneity in the data by region.

Similarly, channel wise analysis of standard deviation for each category show differing SDs channel wise indicating heterogeneity in the data by channel. Box plots by region and channel indicate similar observations as mentioned above. Many outliers are seen under each product at different degrees

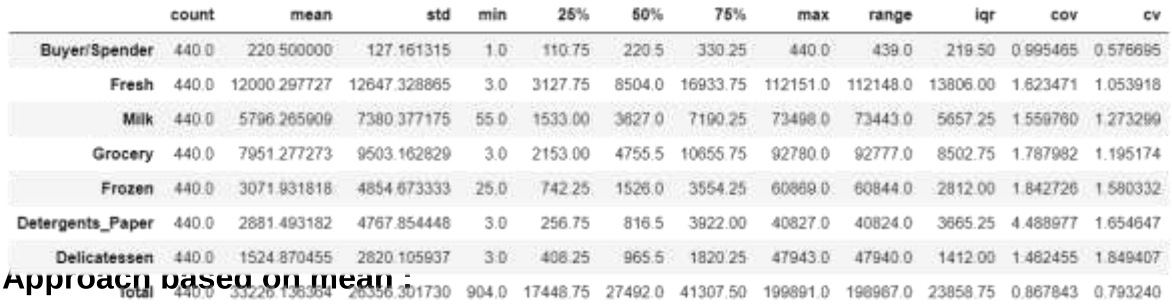
1.3. On the basis of the descriptive measure of variability, which item shows the most inconsistent behavior? Which items shows the least inconsistent behavior?

Solution

Approach based on median :

Detergents_Paper item shows most inconsistent behavior (CV =4.49)

Delicatessen item shows least inconsistent behavior (CV =1.46)

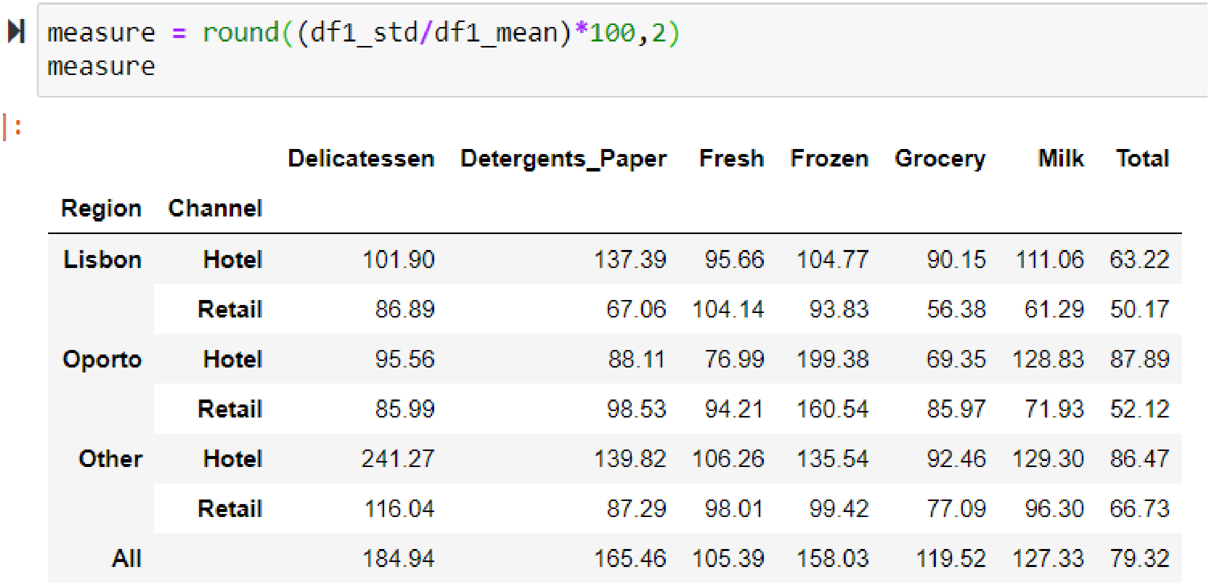


	count	mean	std	min	25%	50%	75%	max	range	iqr	cov	cv
Buyer/Spender	440.0	220.500000	127.161315	1.0	110.75	220.5	330.25	440.0	439.0	219.50	0.995465	0.576695
Fresh	440.0	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0	112148.0	13806.00	1.623471	1.053918
Milk	440.0	5796.265909	7380.377175	55.0	1533.00	3627.0	7190.25	73498.0	73443.0	5657.25	1.559760	1.273299
Grocery	440.0	7951.277273	9503.162829	3.0	2153.00	4755.5	10655.75	92780.0	92777.0	8502.75	1.787982	1.195174
Frozen	440.0	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0	60844.0	2812.00	1.842726	1.580332
Detergents_Paper	440.0	2881.493182	4767.854448	3.0	256.75	816.5	3922.00	40827.0	40824.0	3665.25	4.488977	1.654647
Delicatessen	440.0	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0	47940.0	1412.00	1.462455	1.849407
Total	440.0	33226.136364	26356.301730	904.0	17448.75	27492.0	41307.50	199891.0	198967.0	23858.75	0.867843	0.793240

Approach based on mean :

Fresh item shows least inconsistent behavior (CV = 105.39%)

Delicatessen item shows most inconsistent behavior (CV =184.94 %)

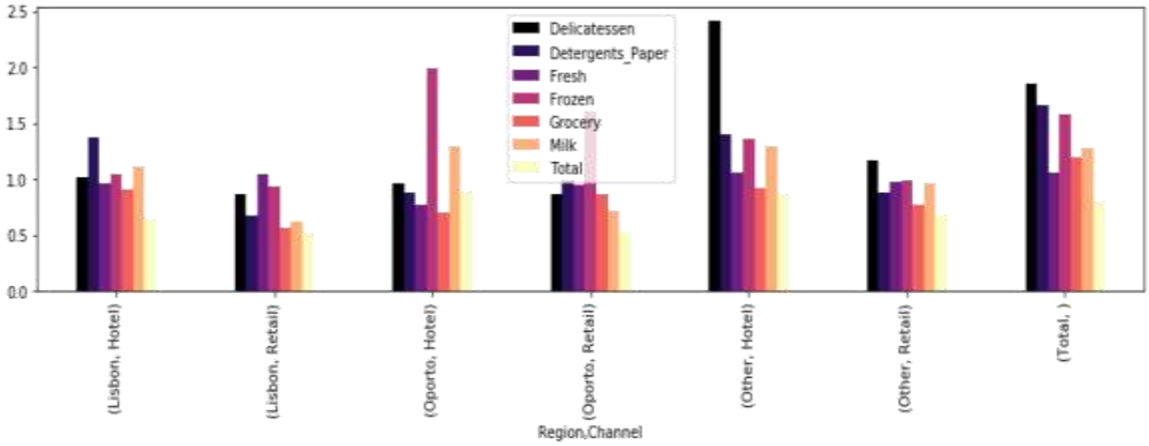


```
measure = round((df1_std/df1_mean)*100,2)
measure
```

]:

		Delicatessen	Detergents_Paper	Fresh	Frozen	Grocery	Milk	Total
Region	**Channel**							
Lisbon	**Hotel**	101.90	137.39	95.66	104.77	90.15	111.06	63.22
	Retail	86.89	67.06	104.14	93.83	56.38	61.29	50.17
Oporto	**Hotel**	95.56	88.11	76.99	199.38	69.35	128.83	87.89
	Retail	85.99	98.53	94.21	160.54	85.97	71.93	52.12
Other	**Hotel**	241.27	139.82	106.26	135.54	92.46	129.30	86.47
	Retail	116.04	87.29	98.01	99.42	77.09	96.30	66.73
All		184.94	165.46	105.39	158.03	119.52	127.33	79.32

Bar Plot

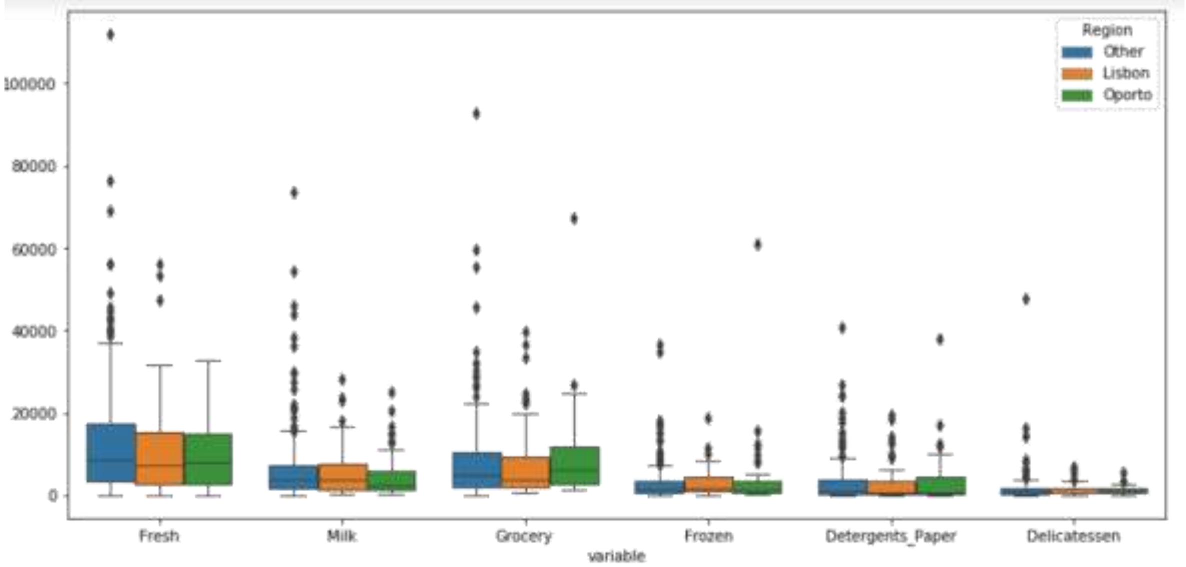


1.4. Are there any outliers in the data?

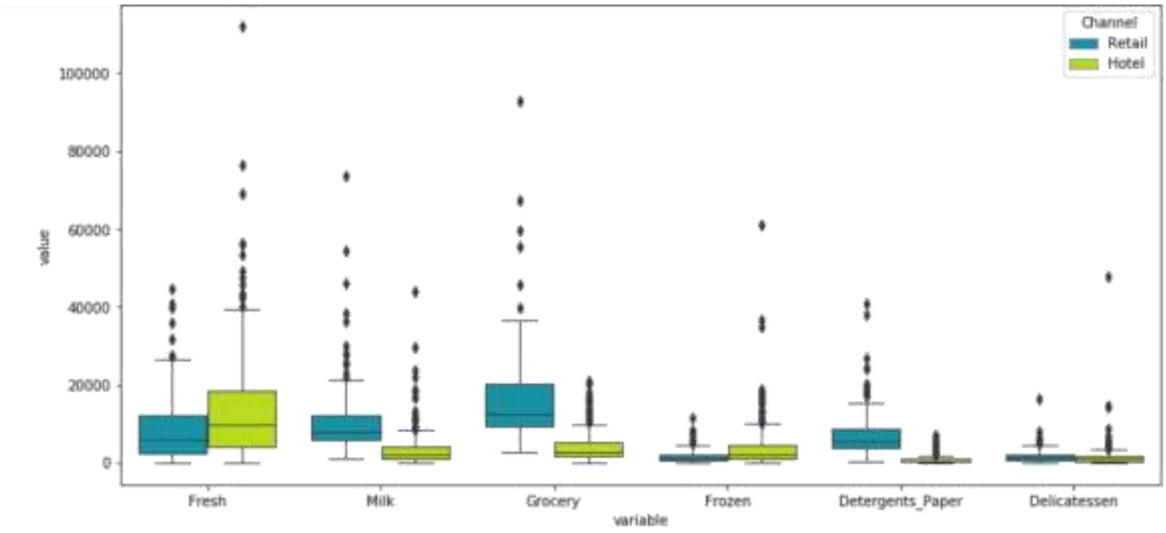
Solution

Yes . The boxplot below shows outliers at different regions for almost all the products. As the data pertains to sales/purchase of products by different channels, namely Hotel and Retail. The high variation in the data might be one of the possible reasons for the appearance of the so called outliers which in fact maybe the true nature of the data (Small and Large purchases).

Across Region



Across Channel



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

Solution

1. If data is available by month instead of for the entire year seasonal variations can be understood.
2. If in the Region Others can be given more details better insight is possible.
3. It is observed that the variability of each of the 6 products is less in Retail compared to Hotel perhaps due to size of the hotels in the group.
4. Further categorization of the channel and time in the data will help in getting better insights in to the behavior thus helping in developing strategies for the wholesale business. Will give insight in to volume of business by more detailed customer types.
5. Plans to expand the products in channels and regions they are weak by providing attractive discounts and conducting survey .
6. Launch a sales incentive program for the retailers to encourage higher sales.
7. Sale of fresh in Oporto- Retail should be focused to increase, along with Milk across hotels in Oporto.

Problem 2

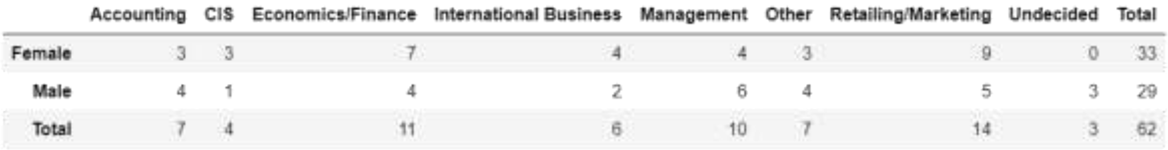
The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. CMSU creates and distributes a survey of 14 questions and receives responses from 62 undergraduates (stored in the Survey.csv file).

Part I

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

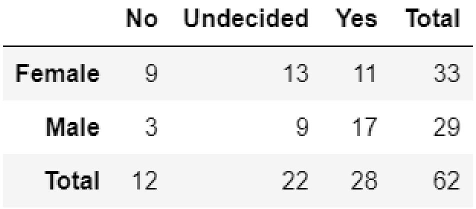
Solution

2.1.1. Gender and Major



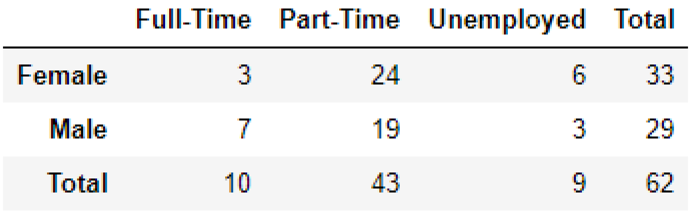
	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	Total
Female	3	3	7	4	4	3	9	0	33
Male	4	1	4	2	6	4	5	3	29
Total	7	4	11	6	10	7	14	3	62

2.1.2. Gender and Grad Intention



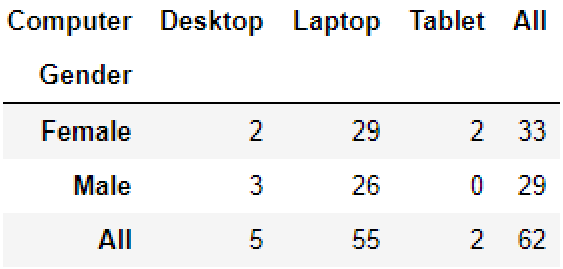
	No	Undecided	Yes	Total
Female	9	13	11	33
Male	3	9	17	29
Total	12	22	28	62

2.1.3. Gender and Employment



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

2.1.4. Gender and Computer



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

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

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?

Solution

Probability of Female	0.532258
Probability of Male	0.467742

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.

Conditional Probability of Male Students

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

Conditional Probability of Female Students

Major	
Accounting	0.090909
CIS	0.090909
Economics/Finance	0.212121
International Business	0.121212
Management	0.121212
Other	0.090909
Retailing/Marketing	0.272727

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.

Conditional Probability of Male Students

Yes	0.586207

Conditional Probability of Female Students

Yes	0.333333

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

Conditional Probability of Male Students

Employment	
Full-Time	0.241379
Part-Time	0.655172
Unemployed	0.103448

Conditional Probability of Female Students

Employment	
Full-Time	0.090909
Part-Time	0.727273
Unemployed	0.181818

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

Conditional Probability of Male Students

Computer	
Laptop	0.896552

Conditional Probability of Female Students

Computer	
Laptop	0.787879

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.

Solution

Tests: Chi-Square test of Independence was performed to determine the independence of column variables viz. Major, Grad Intention, Employment, and Computer.

Following assumptions are taken along with given parameters:

Null Hypothesis: The Gender and Column variables are Independent of each other.

Alternative Hypothesis: The Gender and Column variables are dependent of each other.

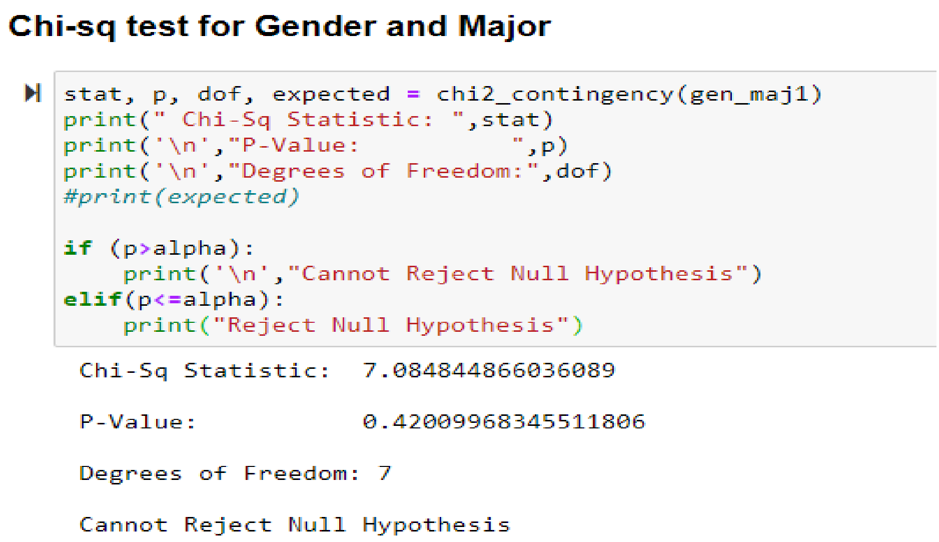
Given Parameters:

Alpha value: 0.05

Critical value obtained: 5.991464547107979

Following Results were obtained for each Column:

Chi-Square test for Gender and Major

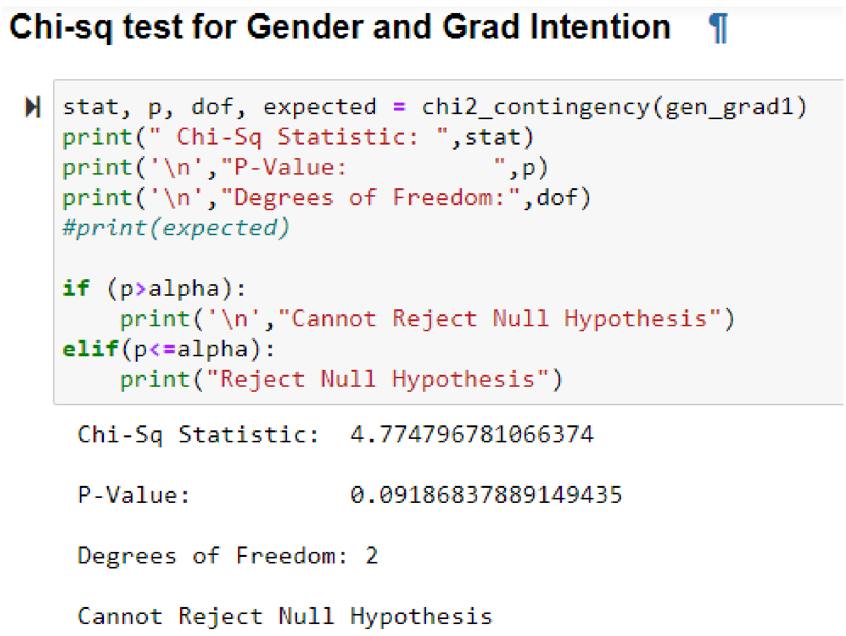


Chi-sq test for Gender and Major

```
stat, p, dof, expected = chi2_contingency(gen_maj1)
print(" Chi-Sq Statistic: ",stat)
print('\n',"P-Value:          ",p)
print('\n',"Degrees of Freedom:",dof)
#print(expected)

if (p>alpha):
    print('\n',"Cannot Reject Null Hypothesis")
elif(p<=alpha):
    print("Reject Null Hypothesis")
```

```
 Chi-Sq Statistic:  7.084844866036089

 P-Value:           0.42009968345511806

 Degrees of Freedom: 7

 Cannot Reject Null Hypothesis
```

Chi-Square test for Gender and Major

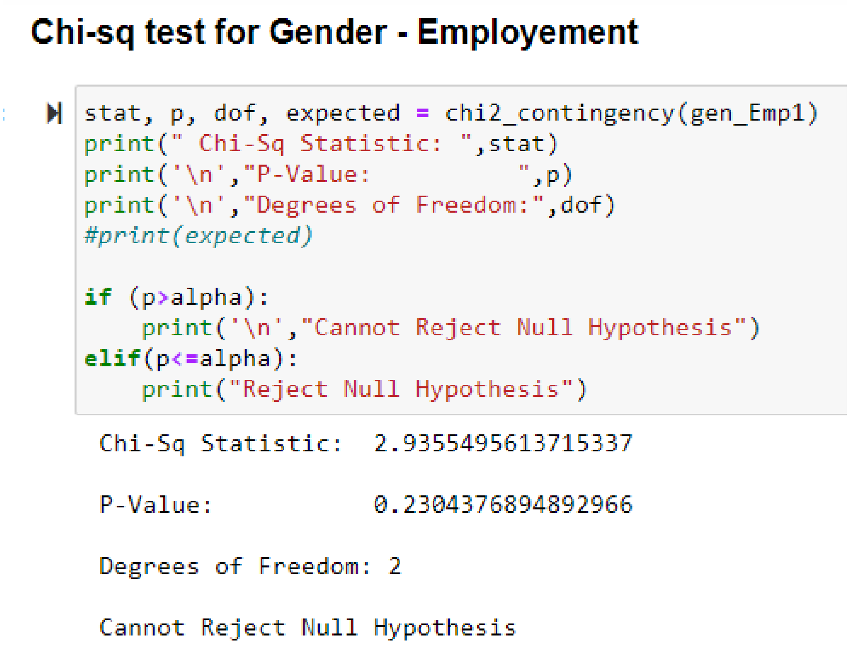


Chi-sq test for Gender and Grad Intention ¶

```
stat, p, dof, expected = chi2_contingency(gen_grad1)
print(" Chi-Sq Statistic: ",stat)
print('\n',"P-Value:          ",p)
print('\n',"Degrees of Freedom:",dof)
#print(expected)

if (p>alpha):
    print('\n',"Cannot Reject Null Hypothesis")
elif(p<=alpha):
    print("Reject Null Hypothesis")
```

```
 Chi-Sq Statistic:  4.774796781066374

 P-Value:           0.09186837889149435

 Degrees of Freedom: 2

 Cannot Reject Null Hypothesis
```

Chi-square test for Gender - Employment

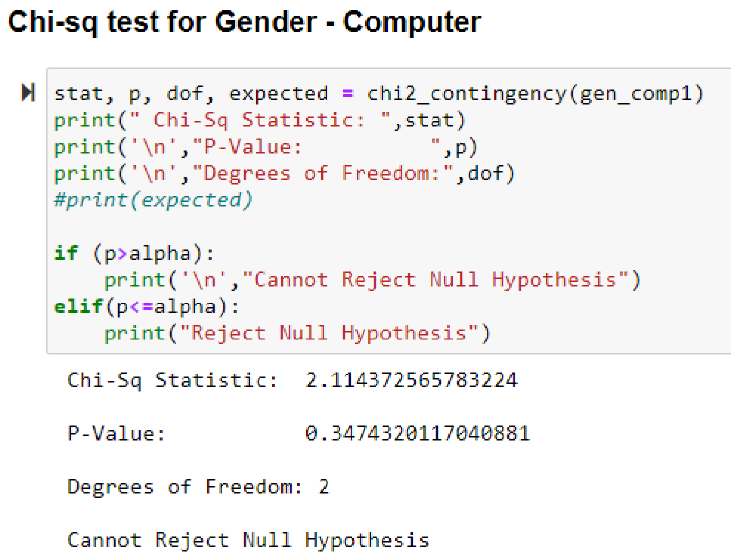


Chi-sq test for Gender - Employement

```
stat, p, dof, expected = chi2_contingency(gen_Emp1)
print(" Chi-Sq Statistic: ",stat)
print('\n',"P-Value:           ",p)
print('\n',"Degrees of Freedom:",dof)
#print(expected)

if (p>alpha):
    print('\n',"Cannot Reject Null Hypothesis")
elif(p<=alpha):
    print("Reject Null Hypothesis")
```

```
 Chi-Sq Statistic:  2.9355495613715337

 P-Value:            0.2304376894892966

 Degrees of Freedom: 2

 Cannot Reject Null Hypothesis
```

Chi-square test for Gender – Computer



Chi-sq test for Gender - Computer

```
stat, p, dof, expected = chi2_contingency(gen_comp1)
print(" Chi-Sq Statistic: ",stat)
print('\n',"P-Value:           ",p)
print('\n',"Degrees of Freedom:",dof)
#print(expected)

if (p>alpha):
    print('\n',"Cannot Reject Null Hypothesis")
elif(p<=alpha):
    print("Reject Null Hypothesis")
```

```
 Chi-Sq Statistic:  2.114372565783224

 P-Value:            0.3474320117040881

 Degrees of Freedom: 2

 Cannot Reject Null Hypothesis
```

Conclusion:

As per the above results the mentioned columns are independent of Gender as in every case the P-value is greater than alpha value which says of not rejecting **null Hypothesis**.

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]

Solution

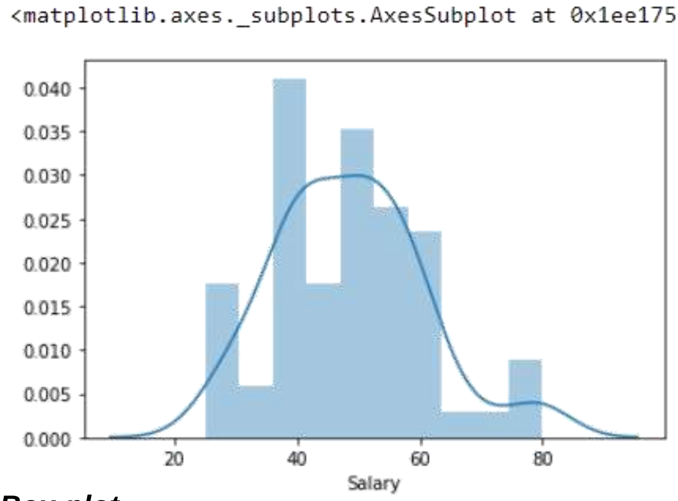
To know whether the given sample/Population is from normal distribution or not we conduct following tests which can prove the normality of the sample.

Statistical Tests: Normality test, Shapiro-Wilk test and many more **(**here we are considering the mentioned tests **.)**

Graphical Representation: Histogram and Box-Plot

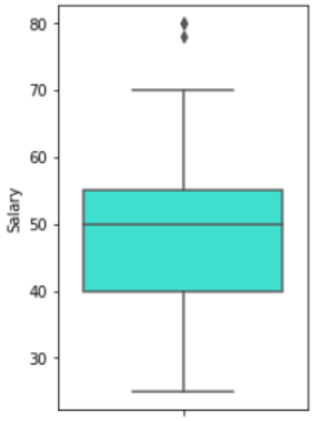
For Salary:

Histogram:



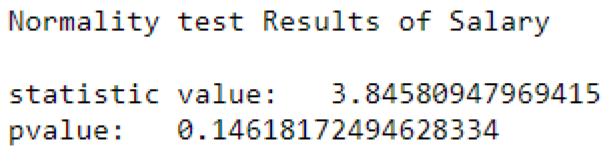
```
<matplotlib.axes._subplots.AxesSubplot at 0x1ee175f3ec8>
```

Box-plot:



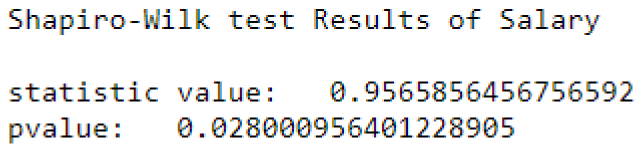
Statistical Tests:

Normality Test:



```
Normality test Results of Salary

statistic value:   3.84580947969415
pvalue:   0.14618172494628334
```

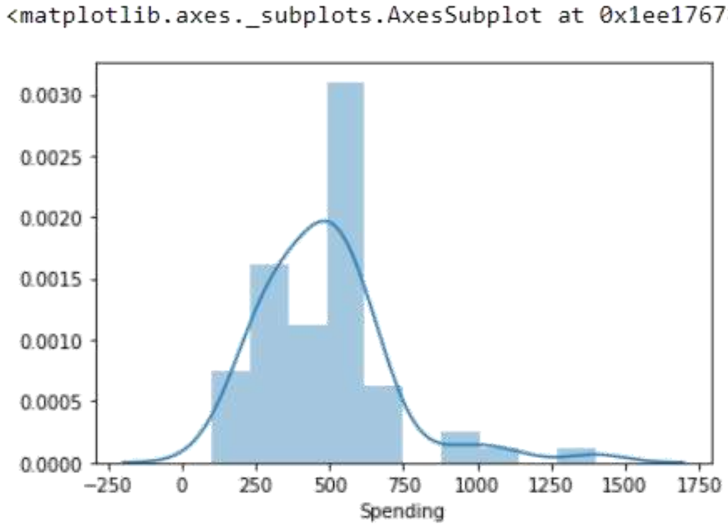
Shapiro-Wilk Test:



```
Shapiro-Wilk test Results of Salary

statistic value:   0.9565856456756592
pvalue:   0.028000956401228905
```

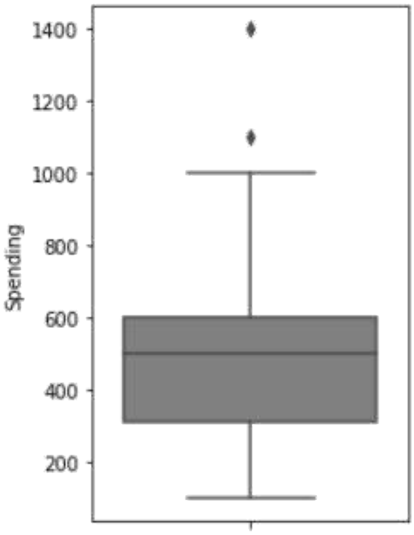
For Spending:

Histogram:



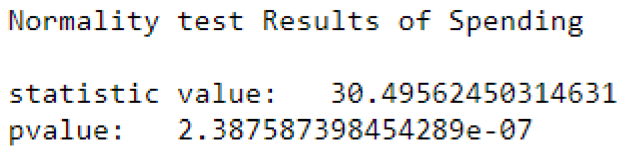
```
<matplotlib.axes._subplots.AxesSubplot at 0x1ee176782c8>
```

Box-Plot



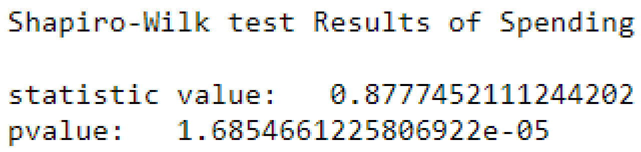
Statistical Tests:

Normality Test:



```
Normality test Results of Spending

statistic value:   30.49562450314631
pvalue:   2.387587398454289e-07
```

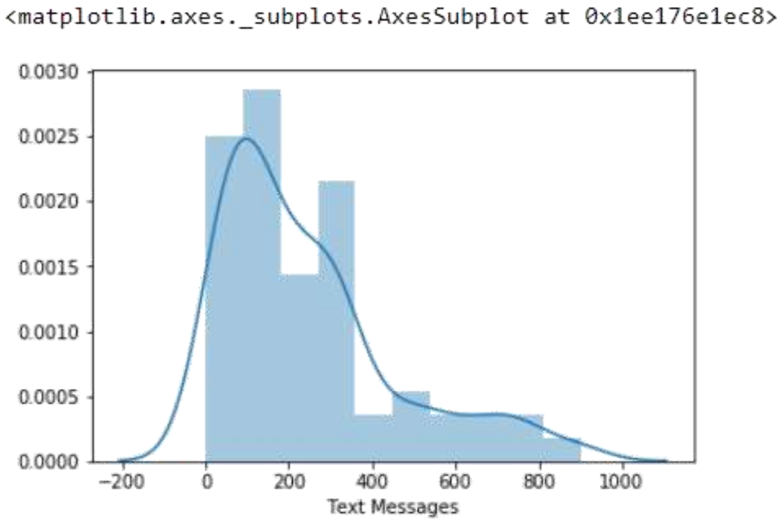
Shapiro-Wilk Test:



```
Shapiro-Wilk test Results of Spending

statistic value:   0.8777452111244202
pvalue:   1.6854661225806922e-05
```

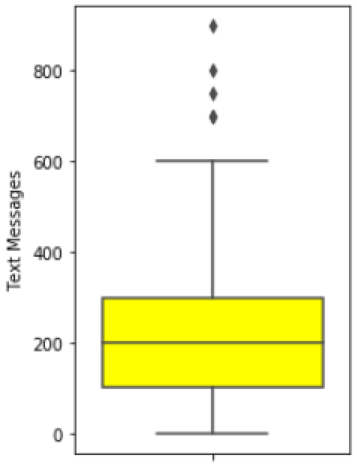
For Text Messages:

Histogram:



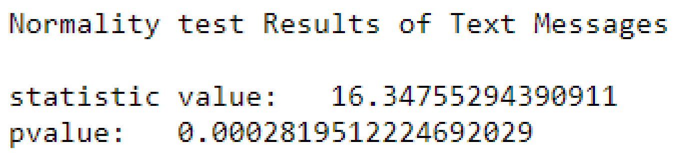
```
<matplotlib.axes._subplots.AxesSubplot at 0x1ee176e1ec8>
```

Box-Plot:



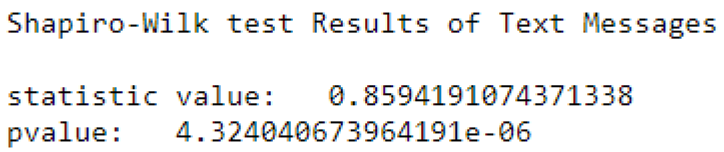
Statistical Tests:

Normality Test:



```
Normality test Results of Text Messages

statistic value:   16.34755294390911
pvalue:   0.0002819512224692029
```

Shapiro-Wilk Test:



```
Shapiro-Wilk test Results of Text Messages

statistic value:   0.8594191074371338
pvalue:   4.324040673964191e-06
```

Conclusion:

Salary: The sample taken from the population is not 100% normally distributed population but it is very nearer to the normal distribution.

Spending: The given population is not normally distributed - Moderately skewed

Text Messages: The given population is not normally distributed.- Highly skewed

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.**

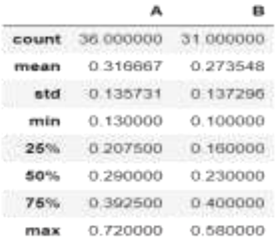
3.1. For the A shingles, form the null and alternative hypothesis to test whether the population mean moisture content is less than 0.35 pound per 100 square feet

Solution

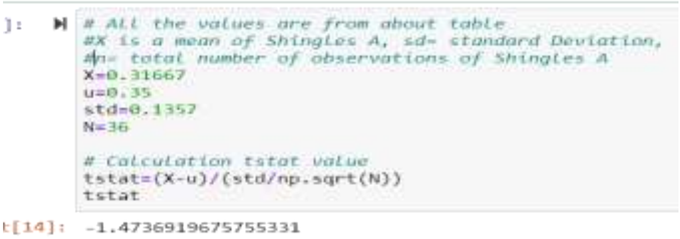
Ho : μ ≥ 0.35 H1 : μ < 0.35

Decision rule: Reject Ho if t-STAT < −1.690 (95% confidence) df=35

t-STAT value = -1.4736



	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

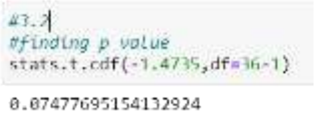


```
]:     # All the values are from about table
       #X is a mean of Shingles A, sd= standard Deviation,
       #n= total number of observations of Shingles A
       X=0.31667
       u=0.35
       std=0.1357
       N=36

       # Calculation tstat value
       tstat=(X-u)/(std/np.sqrt(N))
       tstat
t[14]: -1.4736919675755331
```

Decision: Since t-STAT > −1.690, do not reject Ho . There is not enough evidence to conclude that the mean moisture content for A shingles is less than 0.35 pounds per 100 square feet .

3.2. For the A shingles, conduct the test of 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?

p-value is 0.07477 and is greater than LOF (0.05) so we reject Null Hypothesis and no evidence to prove that the population mean moisture content is less than 0.35 pound per 100 square feet .



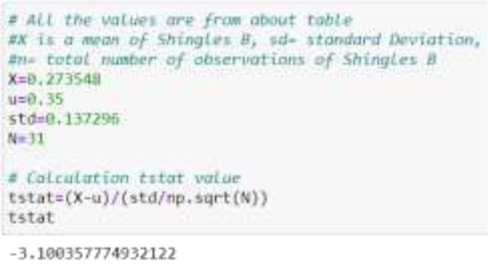
```
#3.2
#finding p value
stats.t.cdf(-1.4735,df=36-1)

0.07477695154132924
```

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

Ho : µ ≥ 0.35   H1 : µ < 0.35

Decision rule: Reject Ho if T-stat < −1.69 (95% confidence) , df=30

T-stat value = -3.1003

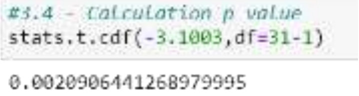


```
# All the values are from about table
#X is a mean of Shingles B, sd= standard Deviation,
#n= total number of observations of Shingles B
X=0.273548
u=0.35
std=0.137296
N=31

# Calculation tstat value
tstat=(X-u)/(std/np.sqrt(N))
tstat

-3.100357774932122
```

Decision: Since T-stat < -1.69, reject H0 . There is enough evidence to conclude that the mean moisture content for B shingles is less than 0.35 pounds per 100 square feet.

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?

P value is 0.00209 and less than alpha (0.05) so we reject Ho and there evidence is enough evidence at the 0.05 level of significance that the population mean moisture content is less than 0.35 pound per 100 square feet

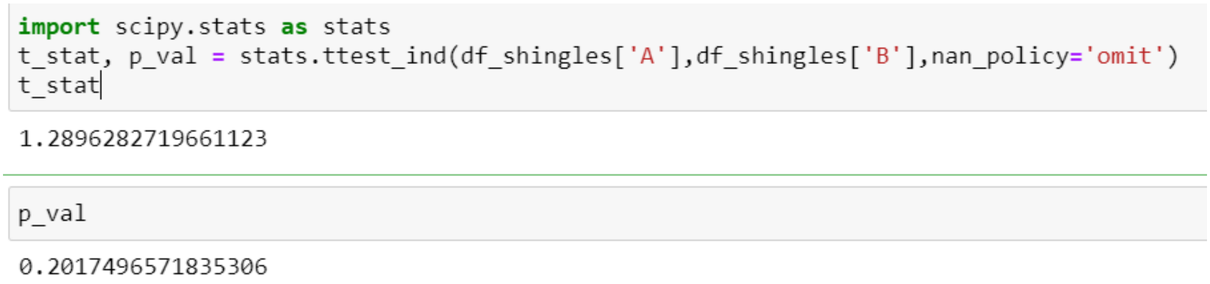


```
#3.4 - Calculation p value
stats.t.cdf(-3.1003,df=31-1)

0.0020906441268979995
```

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?

Ho: mean of A = mean of B, H1 : mean of A ≠ mean B

Based on the independent t test the p value is greater than alpha, so rejecting Ho .



```
import scipy.stats as stats
t_stat, p_val = stats.ttest_ind(df_shingles['A'],df_shingles['B'],nan_policy='omit')
t_stat
```

```
1.2896282719661123
```

```
p_val
```

```
0.2017496571835306
```

Assumptions : Below are the assumptions

1. Independent samples/groups (i.e., independence of observations)
2. Random sample of data from the population
3. Normal distribution (approximately) of the dependent variable for each group
   - Non-normal population distributions, especially those that are thick-tailed or heavily skewed, considerably reduce the power of the test
   - Among moderate or large samples, a violation of normality may still yield accurate p-values
4. No outliers

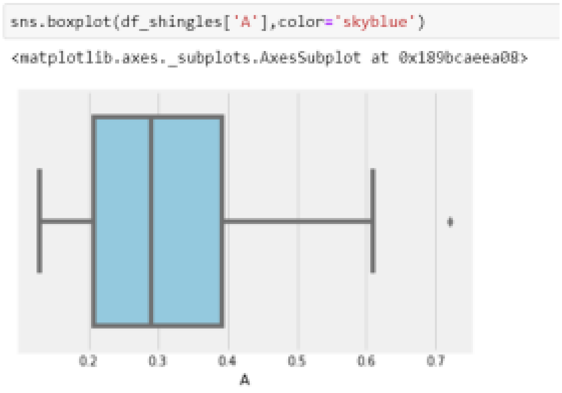
3.6. What assumption about the population distribution is needed in order to conduct the hypothesis tests above?

□ In order for the t test to be valid, the data are assumed to be independently drawn from a population that is normally distributed.

3.7. Check the assumptions made with histograms, boxplots, normal probability plots or empirical rule.

Boxplots

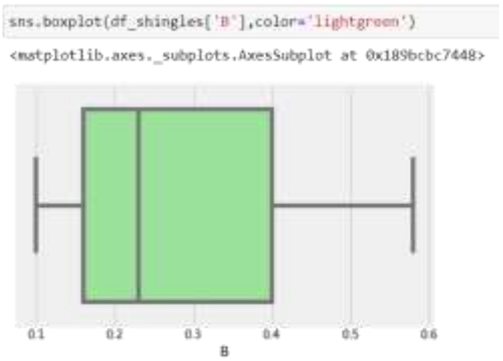
Shingles A



```
sns.boxplot(df_shingles['A'],color='skyblue')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x189bcaeea08>
```

Shingles B

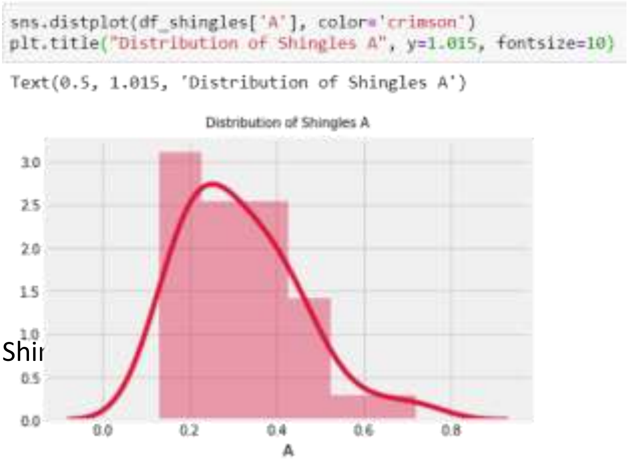


```
sns.boxplot(df_shingles['B'],color='lightgreen')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x189bcbc7448>
```

□ Both boxplots suggest that the data are skewed slightly to the right, more for the A shingles with outliers. However, the very large sample sizes mean that the results of the t test are relatively insensitive to the departure from normality .

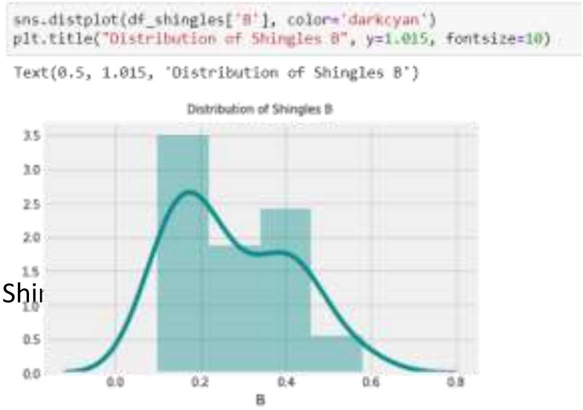
Histograms :



```
sns.distplot(df_shingles['A'], color='crimson')
plt.title("Distribution of Shingles A", y=1.015, fontsize=10)
```

```
Text(0.5, 1.015, 'Distribution of Shingles A')
```

Shi



```
sns.distplot(df_shingles['B'], color='darkcyan')
plt.title("Distribution of Shingles B", y=1.015, fontsize=10)
```

```
Text(0.5, 1.015, 'Distribution of Shingles B')
```

Shi

3.8. Do you think that the assumption needed in order to conduct the hypothesis tests above is valid? Explain.

□ Yes, the assumptions needed to conduct the hypo. are valid since the sample sizes are 36 and 31 respectively, which are quite large so the t distribution will provide a good approximation to the sampling distribution.

- For a t-test to be valid on a sample of smaller size, the population distribution would have to be approximately normal.
- The t-test is invalid for small samples from non-normal distributions, but it is valid for large samples from non-normal distributions

-------------------------End-------------------------