SMDM PROJECT REPORT

DSBA

Done By: Hariharan Manickam

Wholesale Customers Analysis

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 seems to spend more? Which Region and which Channel seems to spend less.

```
Out[7]: Region
        Lisbon     2386813
        Oporto     1555088
        Other     10677599
        Name: Total_Spend, dtype: int64

Out[8]: Channel
        Hotel     7999569
        Retail    6619931
        Name: Total_Spend, dtype: int64
```

Answer for 1.1

Of the 2 different channels:

- expense is more through the "Hotel" channel.
- expense is less through the "Retail" channel.

We do not know how many regions are present in the other category. Of the 3 different categories of regions:

- "Other" has the highest expense.
- "Oporto" has the lowest expense.

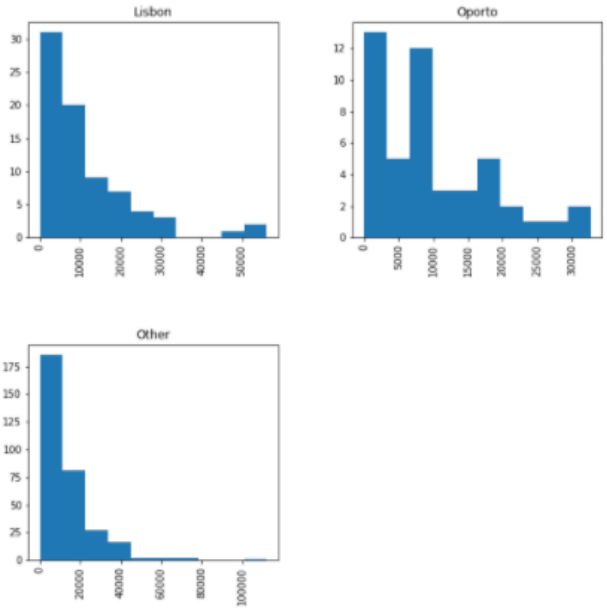
1.2 There are 6 different varieties of items are considered. Do all varieties show similar behaviour across Region and Channel? Provide justification for your answer

	Region	Lisbon	Oporto	Other
Fresh	count	77.000000	47.000000	316.000000
	mean	11101.727273	9887.680851	12533.471519
	std	11557.438575	8387.899211	13389.213115
	min	18.000000	3.000000	3.000000
	25%	2806.000000	2751.500000	3350.750000
	50%	7363.000000	8090.000000	8752.500000
	75%	15218.000000	14925.500000	17406.500000
	max	56083.000000	32717.000000	112151.000000
Milk	count	77.000000	47.000000	316.000000
	mean	5486.415584	5088.170213	5977.085443
	std	5704.856079	5826.343145	7935.463443
	min	258.000000	333.000000	55.000000
	25%	1372.000000	1430.500000	1634.000000
	50%	3748.000000	2374.000000	3684.500000
	75%	7503.000000	5772.500000	7198.750000
	max	28326.000000	25071.000000	73498.000000
Grocery	count	77.000000	47.000000	316.000000
	mean	7403.077922	9218.595745	7896.363924
	std	8496.287728	10842.745314	9537.287778
	min	489.000000	1330.000000	3.000000
	25%	2046.000000	2792.500000	2141.500000
	50%	3838.000000	6114.000000	4732.000000
	75%	9490.000000	11758.500000	10559.750000
	max	39694.000000	67298.000000	92780.000000
Frozen	count	77.000000	47.000000	316.000000
	mean	3000.337662	4045.361702	2944.594937
	std	3092.143894	9151.784954	4260.126243
	min	61.000000	131.000000	25.000000
	25%	950.000000	811.500000	664.750000
	50%	1801.000000	1455.000000	1498.000000
	75%	4324.000000	3272.000000	3354.750000
	max	18711.000000	60869.000000	36534.000000
Detergents_Paper	count	77.000000	47.000000	316.000000
	mean	2651.116883	3687.468085	2817.753165
	std	4208.462708	6514.717668	4593.051613
	min	5.000000	15.000000	3.000000
	25%	284.000000	282.500000	251.250000
	50%	737.000000	811.000000	856.000000
	75%	3593.000000	4324.500000	3875.750000
	max	19410.000000	38102.000000	40827.000000
Delicatessen	count	77.000000	47.000000	316.000000
	mean	1354.896104	1159.702128	1620.601266
	std	1345.423340	1050.739841	3232.581660
	min	7.000000	51.000000	3.000000
	25%	548.000000	540.500000	402.000000
	50%	806.000000	898.000000	994.000000
	75%	1775.000000	1538.500000	1832.750000
	max	6854.000000	5609.000000	47943.000000

	Channel	Hotel	Retail
Fresh	count	298.000000	142.000000
	mean	13475.560403	8904.323944
	std	13831.687502	8987.714750
	min	3.000000	18.000000
	25%	4070.250000	2347.750000
	50%	9581.500000	5993.500000
	75%	18274.750000	12229.750000
	max	112151.000000	44466.000000
Milk	count	298.000000	142.000000
	mean	3451.724832	10716.500000
	std	4352.165571	9679.631351
	min	55.000000	928.000000
	25%	1164.500000	5938.000000
	50%	2157.000000	7812.000000
	75%	4029.500000	12162.750000
	max	43950.000000	73498.000000
Grocery	count	298.000000	142.000000
	mean	3962.137584	16322.852113
	std	3545.513391	12267.318094
	min	3.000000	2743.000000
	25%	1703.750000	9245.250000
	50%	2684.000000	12390.000000
	75%	5076.750000	20183.500000
	max	21042.000000	92780.000000
Frozen	count	298.000000	142.000000
	mean	3748.251678	1652.612676
	std	5643.912500	1812.803662
	min	25.000000	33.000000
	25%	830.000000	534.250000
	50%	2057.500000	1081.000000
	75%	4558.750000	2146.750000
	max	60869.000000	11559.000000
Detergents_Paper	count	298.000000	142.000000
	mean	790.560403	7269.507042
	std	1104.093673	6291.089697
	min	3.000000	332.000000
	25%	183.250000	3683.500000
	50%	385.500000	5614.500000
	75%	899.500000	8662.500000
	max	6907.000000	40827.000000
Delicatessen	count	298.000000	142.000000
	mean	1415.956376	1753.436620
	std	3147.426922	1953.797047
	min	3.000000	3.000000
	25%	379.000000	566.750000
	50%	821.000000	1350.000000
	75%	1548.000000	2156.000000
	max	47943.000000	16523.000000

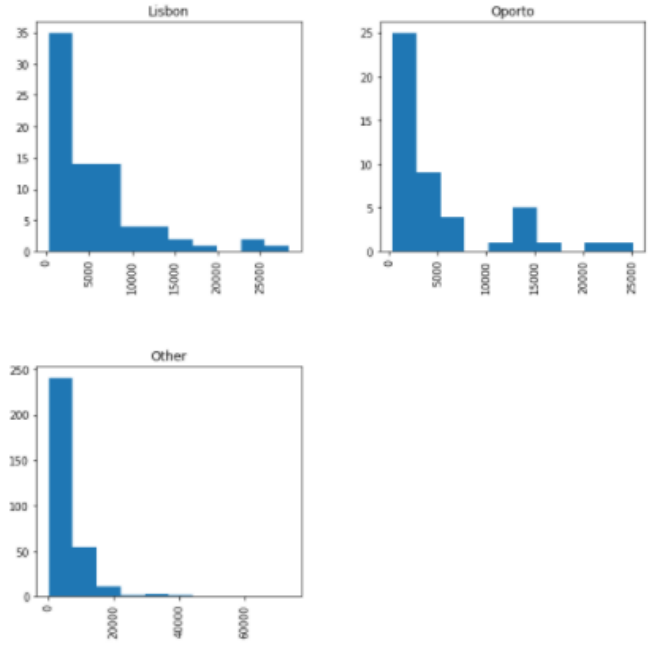
Histogram for Fresh in different regions

```
Out[13]: array([[<matplotlib.axes._subplots.AxesSubplot object at 0x000001DC051B2C48>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000001DC0699E188>],
                [<matplotlib.axes._subplots.AxesSubplot object at 0x000001DC069D90C8>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000001DC06A111C8>]],
               dtype=object)
```

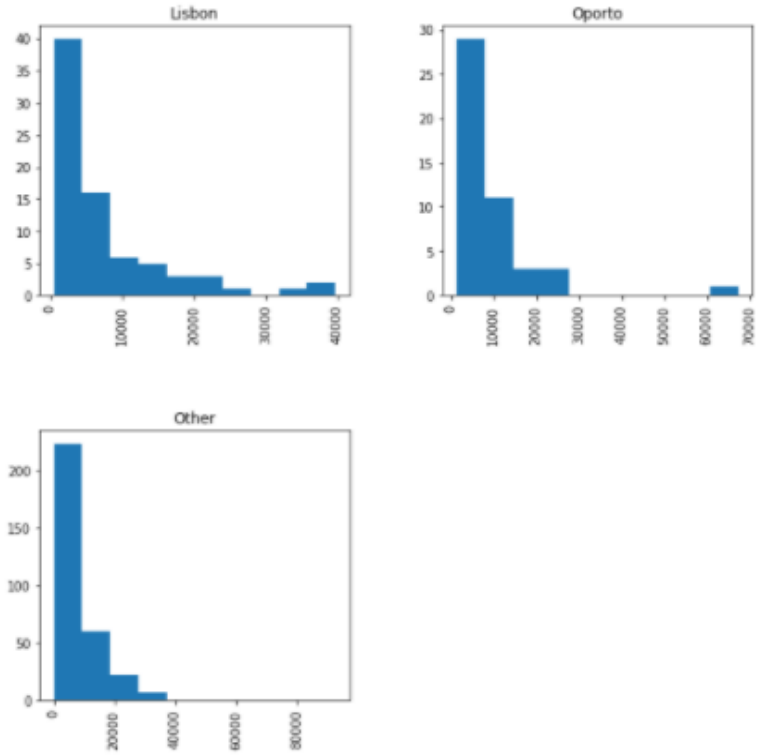


Histogram for Milk in different regions

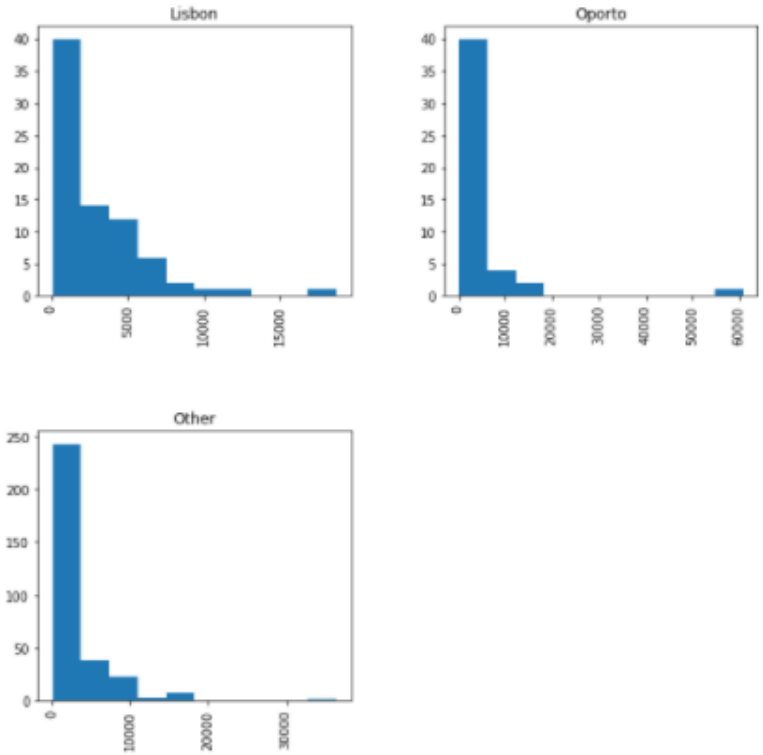
```
Out[14]: array([[<matplotlib.axes._subplots.AxesSubplot object at 0x000001DC06B2CF08>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000001DC06DCCD48>],
                [<matplotlib.axes._subplots.AxesSubplot object at 0x000001DC06B734C8>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000001DC06BA8EC8>]],
               dtype=object)
```



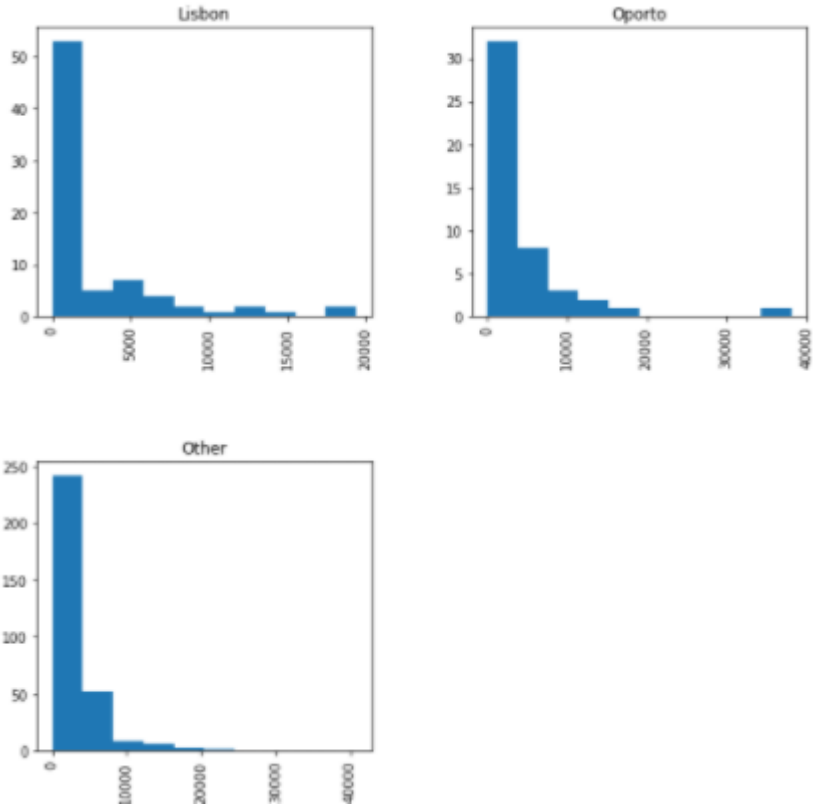
Histogram for Grocery in different regions



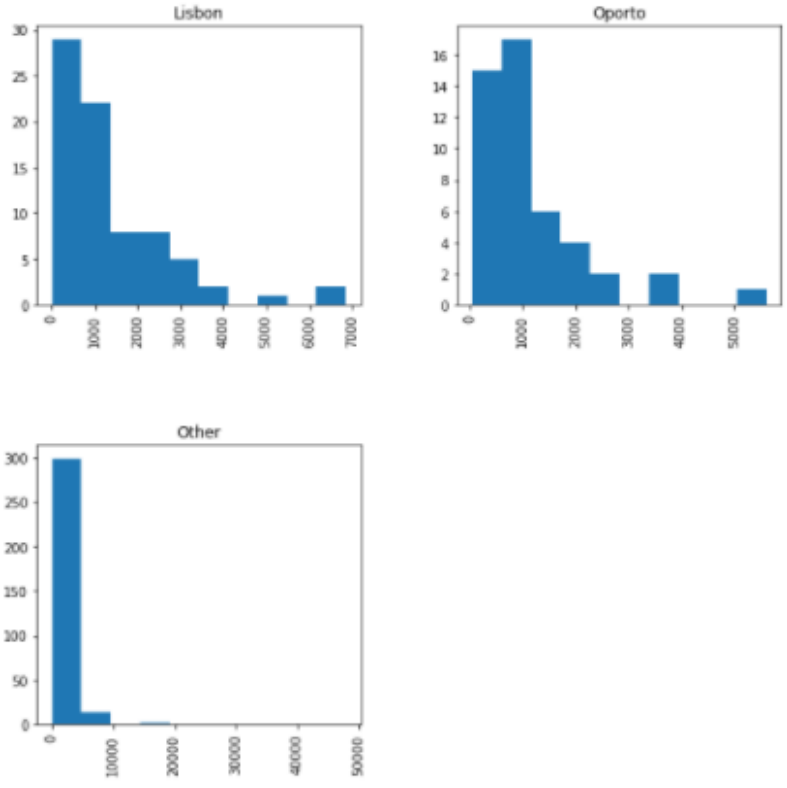
Histogram for Frozen in different regions



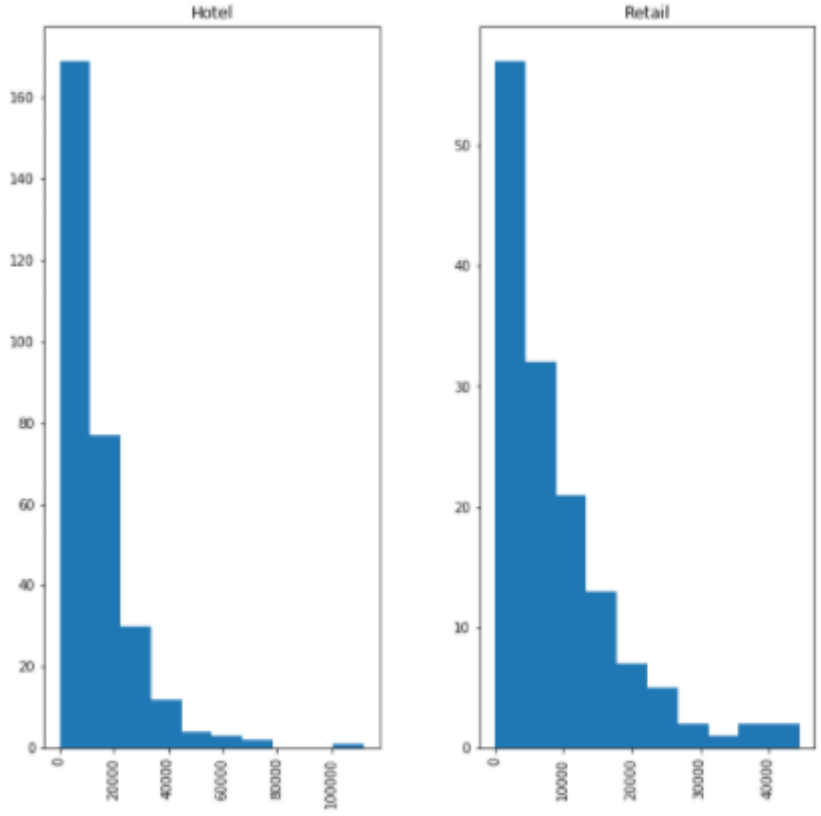
Histogram for Detergents Paper in different regions



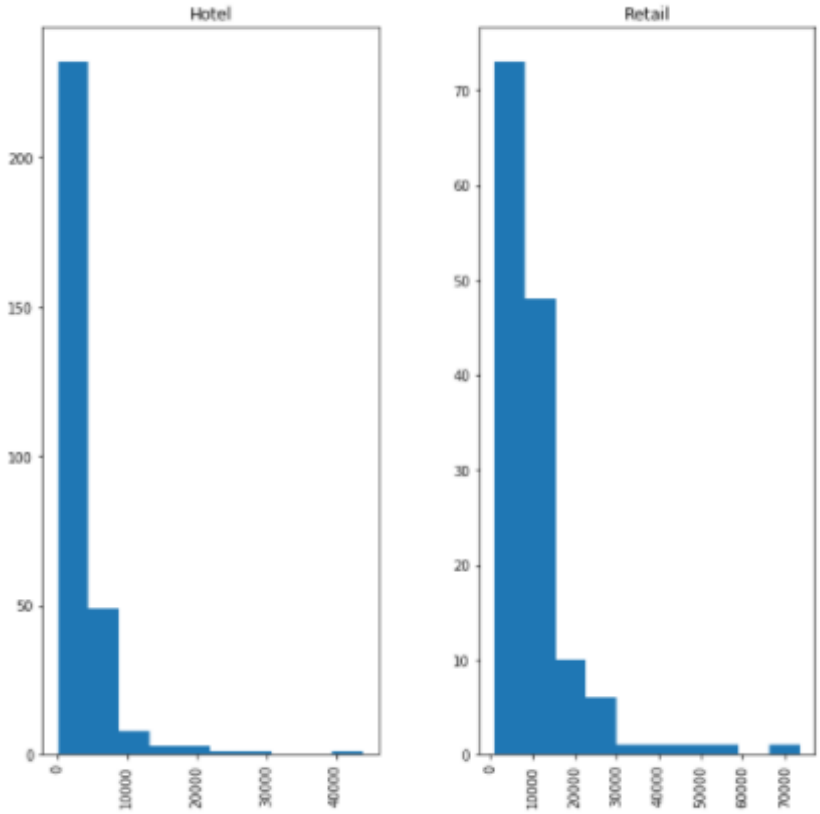
Histogram for Delicatessen in different regions



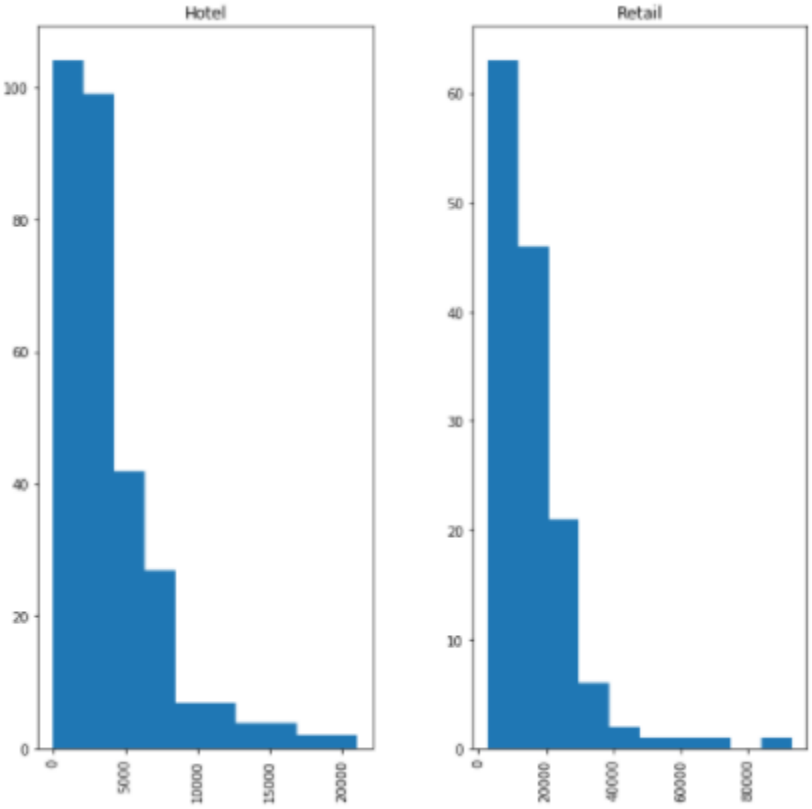
Histogram for Fresh in different sales channels



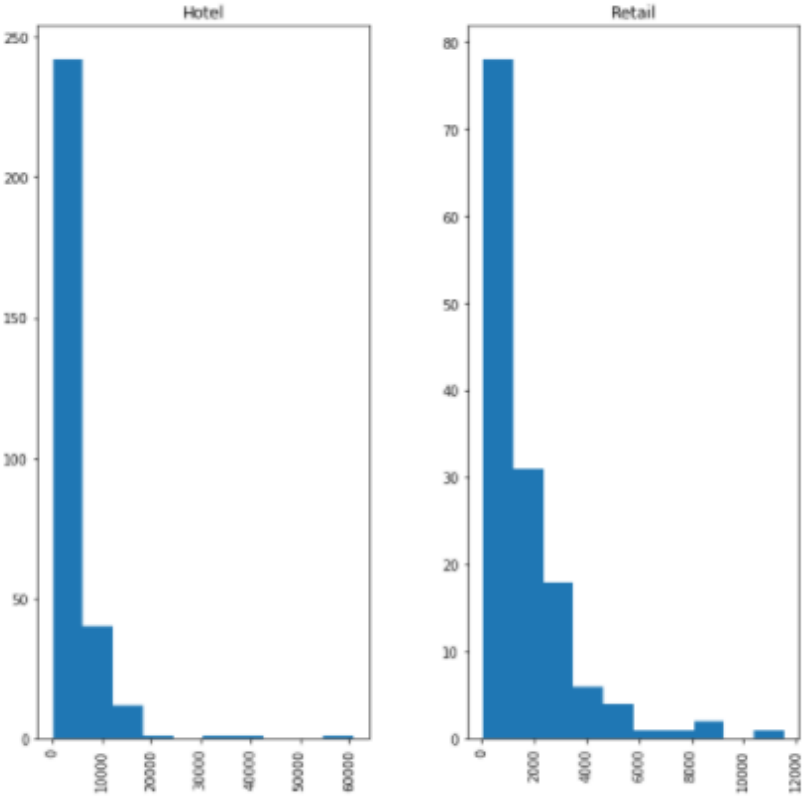
Histogram for Milk in different sales channels



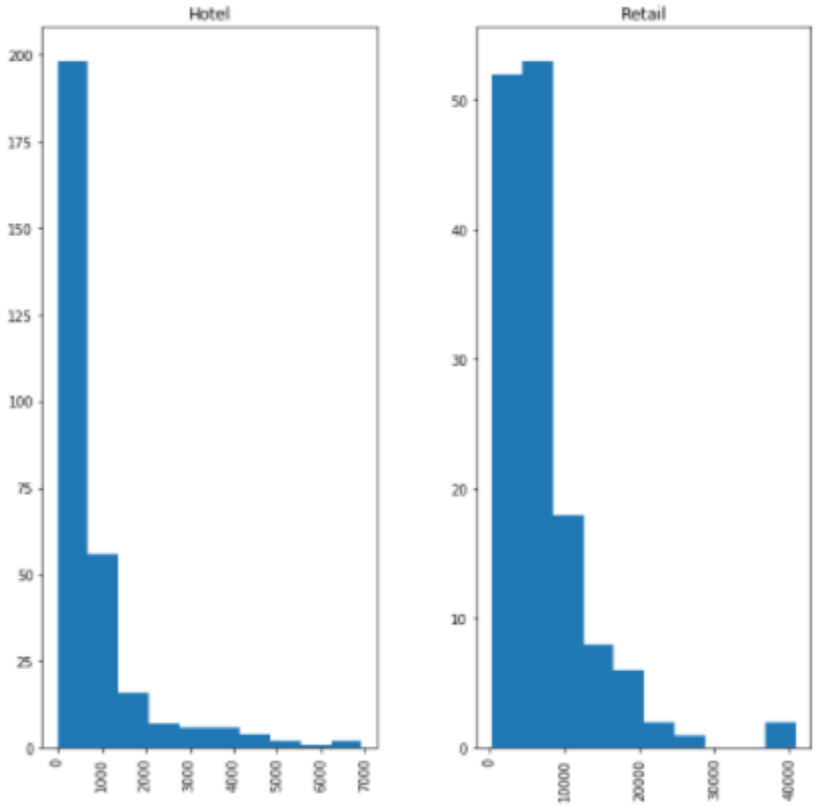
Histogram for Groceries in different sales channels



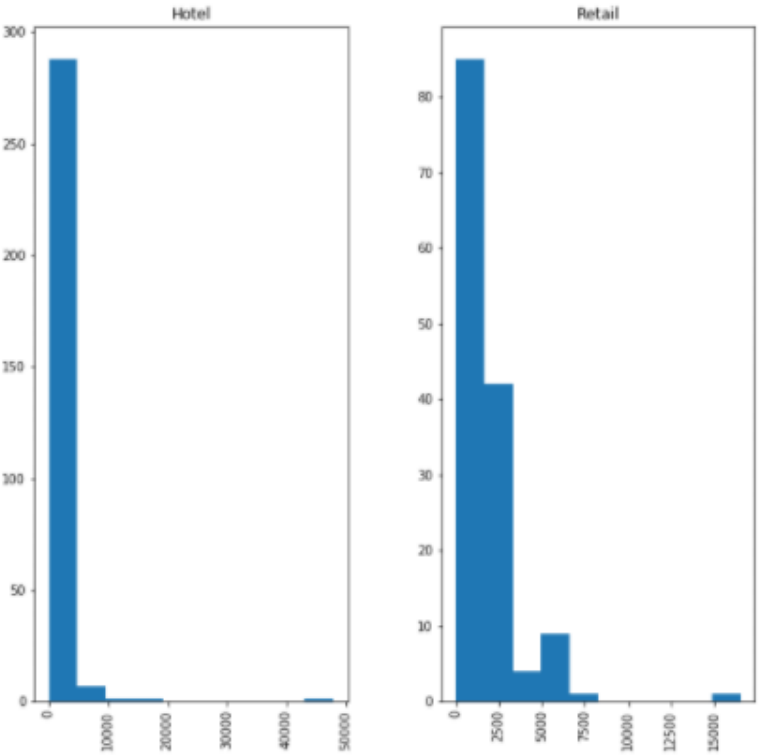
Histogram for Frozen in different sales channels



Histogram for Detergents Paper in different sales channels



Histogram for Delicatessen in different sales channels



Answer for 1.2

When classified by region and when we use the describe funtions, we observed the following

- Fresh has similar mean
- Milk has similar mean
- Delicatessen has similar mean

When classified by Channel and when we use the describe funtions, we observed the following

- Fresh mean differs alot
- Milk mean differs alot
- Grocery mean differs alot
- Frozen mean differs alot
- Detergents paper mean differs alot
- Delicatessen has a similar mean

From the histograms, we observed the following

- Almost all histograms are right skwewed

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?

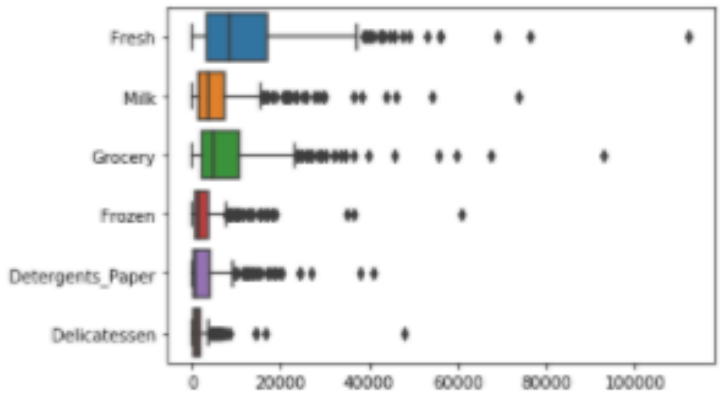
```
Coefficient of Variation for Fresh is 105.392
Coefficient of Variation for Milk is 127.33
Coefficient of Variation for Frozen is 168.478
Coefficient of Variation for Grocery is 119.517
Coefficient of Variation for Detergents_Paper is 165.465
Coefficient of Variation for Delicatessen is 184.941
```

Answer for 1.3

We can conculde that Delicatessen shows the most inconsistent behaviour.

1.4 Are there any outliers in the data?

```
Out[27]: <matplotlib.axes._subplots.AxesSubplot at 0x1dc087bf248>
```

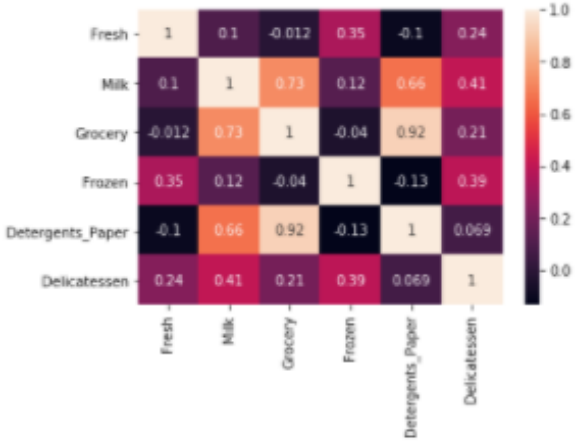


Answer for 1.4

From the above boxplots, we can say that the data has a lot of outliers.

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.

```
Out[28]: <matplotlib.axes._subplots.AxesSubplot at 0x1dc087c4048>
```



Answer for 1.5

Using the heatmap, we can identify the products that have a high correlation to each other when customers buy them.

They are:

- Grocery & Milk
- Detergents Paper & Milk
- Detergents Paper & Grocery

We can place the above products next to each other to increase total 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* data set).

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

2.1.1. Gender and Major

Out[5]:

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

Out[6]:

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

2.1.3. Gender and Employment

Out[7]:

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

2.1.4. Gender and Computer

Out[8]:

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?

```
Probability that a randomly selected CMSU student will be male is  0.468 .
```

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

```
Probability that a randomly selected CMSU student will be female is  0.532 .
```

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 probabilities of different majors among the male students of CMSU are as follows.

P(major as Accounting)= 0.091
P(major as CIS)= 0.091
P(major as Economics/Finance)= 0.212
P(major as International Business)= 0.121
P(major as Management)= 0.121
P(major as Other)= 0.091
P(major as Retailing/Marketing)= 0.273
P(major as Undecided)= 0.0
```

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

```
The conditional probabilities of different majors among the female students of CMSU are as follows.

P(major as Accounting)= 0.138
P(major as CIS)= 0.034
P(major as Economics/Finance)= 0.138
P(major as International Business)= 0.069
P(major as Management)= 0.207
P(major as Other)= 0.138
P(major as Retailing/Marketing)= 0.172
P(major as Undecided)= 0.103
```

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.

Out[16]:

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

Earlier, we found

- Probability of Female = 0.532
- Probability of Male = 0.468

```
Probability that a randomly chosen student is a male and intends to graduate is 0.156
```

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

Out[18]:

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

```
Probability that a randomly chosen student is a female and does not have laptop is 0.055
```

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?

```
Probability that a randomly chosen student is either a male or has full-time employment is 0.629
```

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

```
Probability that given a female student is randomly chosen, she is majoring in international business or management is 0.276
```

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?

Out[22]:

Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

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?

```
Out[24]: 0.3773883926969118

Out[25]: 3.129032258064516
```

Mean is 3.129

Standard Deviation is 0.377

Now we use the formula to find Z.

```
Z = -0.3421750663129974

If a student is chosen randomly, the probability that his/her GPA is less than 3 is 0.366
```

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.

```
Out[28]: 10.79317427068786

Out[29]: 48.275862068965516
```

Mean is 48.276

Standard Deviation is 10.793

Now we use the formula to find Z.

```
Z = 0.15973316038172866

The conditional probability that a randomly selected male earns 50 or more is 0.43654563903418264
```

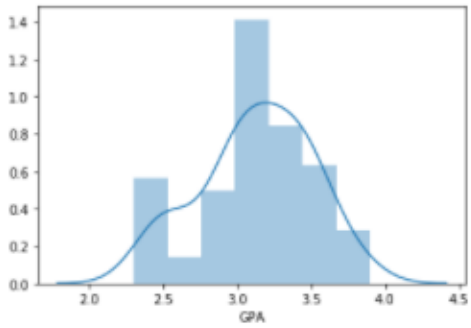
2.8. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarizing your conclusions.

Out[32]:

	count	mean	std	min	25%	50%	75%	max
ID	62.0	31.500000	18.041619	1.0	16.25	31.50	46.75	62.0
Age	62.0	21.129032	1.431311	18.0	20.00	21.00	22.00	26.0
GPA	62.0	3.129032	0.377388	2.3	2.90	3.15	3.40	3.9
Salary	62.0	48.548387	12.080912	25.0	40.00	50.00	55.00	80.0
Social Networking	62.0	1.516129	0.844305	0.0	1.00	1.00	2.00	4.0
Satisfaction	62.0	3.741935	1.213793	1.0	3.00	4.00	4.00	6.0
Spending	62.0	482.016129	221.953805	100.0	312.50	500.00	600.00	1400.0
Text Messages	62.0	246.209677	214.465950	0.0	100.00	200.00	300.00	900.0

Histogram for GPA

Out[33]: <matplotlib.axes._subplots.AxesSubplot at 0x21f3d491148>

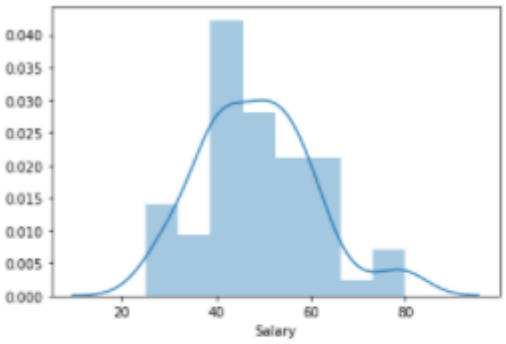


- Q1= 2.9
- Q3= 3.5
- IQR = 0.5
- 1.5*IQR = 0.75
- Lower Extreme = 2.9 - 0.75 = 2.15
- Upper Extreme = 3.5 + 0.75 = 4.25
- Length of lower Whisker = Q1 - Lower Extreme = 2.9 - 2.15 = 0.75
- Length of Upper Whisker = Upper Extreme - Q3 = 4.25 - 3.5 = 0.75

Since the length of both whiskers are equal, we can say that "GPA" follows a normal distribution.

Histogram for Salary

```
Out[34]: <matplotlib.axes._subplots.AxesSubplot at 0x21f3fefc3c8>
```

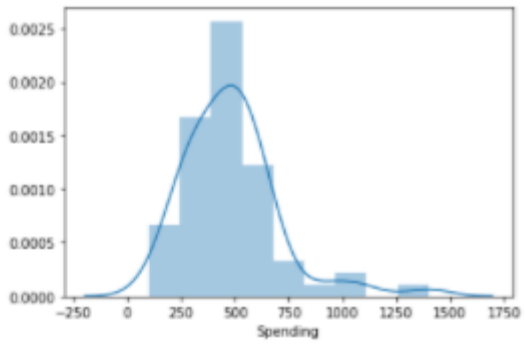


- Q1= 40
- Q3= 55
- IQR = 15
- 1.5*IQR = 22.5
- Lower Extreme = 40 - 22.5= 17.5
- Upper Extreme = 55 + 22.5 = 72.5
- Length of lower Whisker = Q1 - Lower Extreme = 40 - 17.5 = 22.5
- Length of Upper Whisker = Upper Extreme - Q3 = 72.5 - 55 = 17.5

Since the length of both whiskers are unequal we can say that "Salary" does not follow a normal distribution.

Histogram for Spending

```
Out[38]: <matplotlib.axes._subplots.AxesSubplot at 0x16482d24bc8>
```

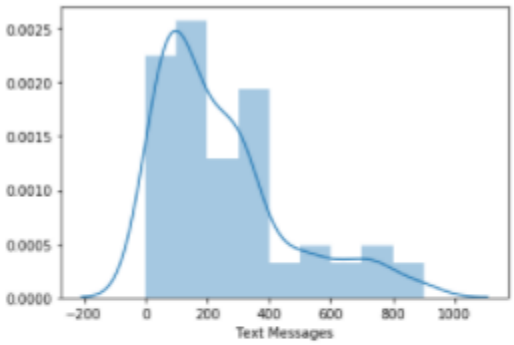


- Q3 = 312.5
- Q1 = 600
- IQR = 287.5
- 1.5*IQR = 431.25
- Lower Extreme = 312.5 - 431.25= -118.75 < 100 (minimum)
- Upper Extreme = 600 + 312.5 = 912.5 < 1400 (maximum)
- Length of lower Whisker = Q1 - Lower Extreme = 312.5 - 100 = 212.5
- Length of Upper Whisker = Upper Extreme - Q3 = 912.5 - 600 = 312.5

Since the length of both whiskers are unequal we can say that "Spending" does not follow a normal distribution.

Histogram for Text Messages

```
Out[39]: <matplotlib.axes._subplots.AxesSubplot at 0x16482d045c8>
```



- Q1 = 100
- Q3 = 300
- IQR = 200
- 1.5*IQR = 300
- Lower Extreme = 100 - 300 = -200 < 0 (minimum)
- Upper Extreme = 300 + 200 = 500 < 900 (maximum)
- Length of lower Whisker = Q1 - Lower Extreme = 100 - 0 = 100.
- Length of Upper Whisker = Upper Extreme - Q3 = 500 - 300 = 200

Since the Length of both whiskers are unequal we can say that Text Messages does not follow 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 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.

Step 1: Define null and alternative hypotheses

- H_0: μ <= 0.35
- H_1: μ > 0.35

Step 2: Decide the significance level

Here we select α = 0.05 and the population standard deviation is not known.

Step 3: Identify the test statistic

The test statistic is the value of data in column A & B.

Step 4: Calculate the p - value and test statistic

We use the scipy.stats.ttest_ind to calculate the t-test for the means of TWO INDEPENDENT samples of scores given the two sample observations. This function returns t statistic and two-tailed p value.

This is a one-sided test for the null hypothesis that the value is greater than the permissible value

Step 5: Decide to reject or accept null hypothesis

For a right tailed test, we take p/2.

0.149/2 = 0.0745 > 0.05

Answer for 3.1

Hence, we fail to reject the null hypothesis, Shingles A are within the permissible limits.

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?

Step 1 Define null and alternative hypothesis.

We Assume that the distribution of two population is Normal.

- Mu(A)- Population mean of Shingles A.
- Mu(B) – Population mean of Shingles B.
- Ho: Mu(A) – Mu(B) = 0
- H1: Mu(A) – Mu(B)! = 0

Step 2 Decide significance level

We conduct the paired t-test at 95% Significance. From the Two tail test we determine that- P-value= 0.20 α = 0.05

Step 3: Identify the test statistic

The test statistic is the value of data in column A & B.

Step 4: Calculate the p - value and test statistic

We use the scipy.stats.ttest_ind to calculate the t-test for the means of TWO INDEPENDENT samples of scores given the two sample observations. This function returns t statistic and two-tailed p value.

```
Out[11]: (1.2896282719661123, 0.2017496571835306)
```

- P-value= 0.20
- α = 0.05
- p-value > α

Step 5

Answer for 3.2

- Hence we fail to reject H0
- We can say that population mean for shingles A and B are equal.