Statistical Methods For Decision Making Project

CIJTIH JOSE

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 (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, 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?*

Data Description:

- FRESH: annual spending on fresh products
- MILK: annual spending on milk products
- GROCERY: annual spending on grocery products
- FROZEN: annual spending on frozen products
- DETERGENTS_PAPER: annual spending on detergents and paper products
- DELICATESSEN: annual spending on and delicatessen products
- CHANNEL: Customers Channel - Hotel (Hotel/Restaurant) or Retail channel (stores/outlets);
- REGION: Customers location Lisnon, Oporto or Other
- BUYER/SPENDER: number indicating the buyers

Sample of the dataset:

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
0	1	Retail	Other	12669	9656	7561	214	2674	1338
1	2	Retail	Other	7057	9810	9568	1762	3293	1776
2	3	Retail	Other	6353	8808	7684	2405	3516	7844
3	4	Hotel	Other	13265	1196	4221	6404	507	1788
4	5	Retail	Other	22615	5410	7198	3915	1777	5185

The Dataset has 9 variables with the information about annual spending on six different items available in the stores which is sold through two different channels and across 3 different regions in Portugal.

Exploratory Data Analysis

Let us check for the type of variables and the missing values in the dataset:

```
Range Index: 440 entries, 0 to 439
Data columns (total 9 columns):
 #  Column          Non-Null Count  Dtype
 0  Buyer/Spender   440 non-null    int64
 1  Channel         440 non-null    object
 2  Region          440 non-null    object
```

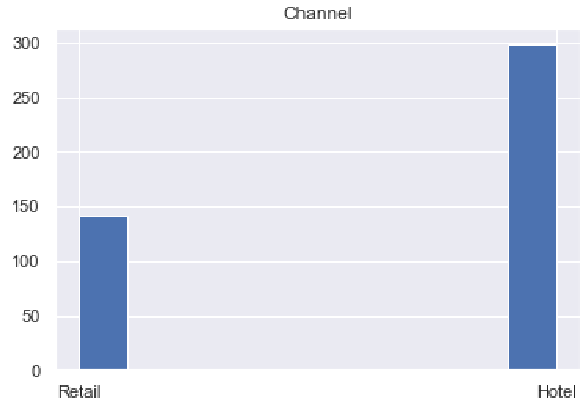
3 Fresh 440 non-null int64
4 Milk 440 non-null int64
5 Grocery 440 non-null int64
6 Frozen 440 non-null int64
7 Detergents_Paper 440 non-null int64
8 Delicatessen 440 non-null int64
dtypes: int64(7), object(2)

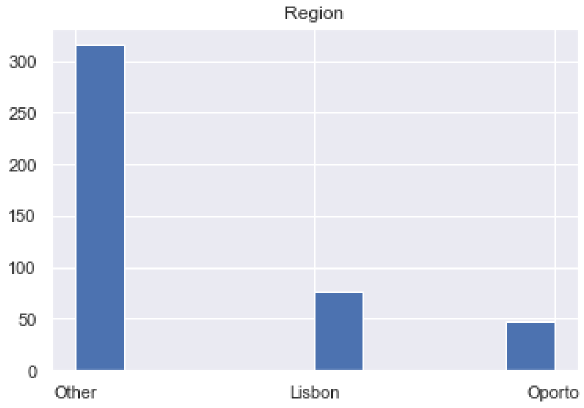
There are total 440 rows and 9 columns in the dataset. Out of 9, 2 columns are of object type and rest 7 are integer data type. We can also see that there are no missing values present in the dataset .

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Buyer/Spender	440				220.5	127.1613	1	110.75	220.5	330.25	440
Channel	440	2	Hotel	298							
Region	440	3	Other	316							
Fresh	440				12000.3	12647.33	3	3127.75	8504	16933.75	112151
Milk	440				5796.266	7380.377	55	1533	3627	7190.25	73498
Grocery	440				7951.277	9503.163	3	2153	4755.5	10655.75	92780
Frozen	440				3071.932	4854.673	25	742.25	1526	3554.25	60869
Detergents_Paper	440				2881.493	4767.854	3	256.75	816.5	3922	40827
Delicatessen	440				1524.87	2820.106	3	408.25	965.5	1820.25	47943

From the above data we can infer that there are two unique values under 'Channel' and three unique values under 'Region' & also see that more data is available on the annual expenditure of products which are sold through the Hotel Channel as they are total of 298 in number as compared to Retail which is 142 out of 440 . We can similarly see that the information on the annual expenditure of products sold from 'Other' region is higher. Which can also mean that the highest contributors are from the Channel 'Hotel' and from the region 'Other, which we will find out in our analysis ahead.

One more inference from the above data would be that there are number of outliers present in the dataset across all the variables and there is positive skewness.



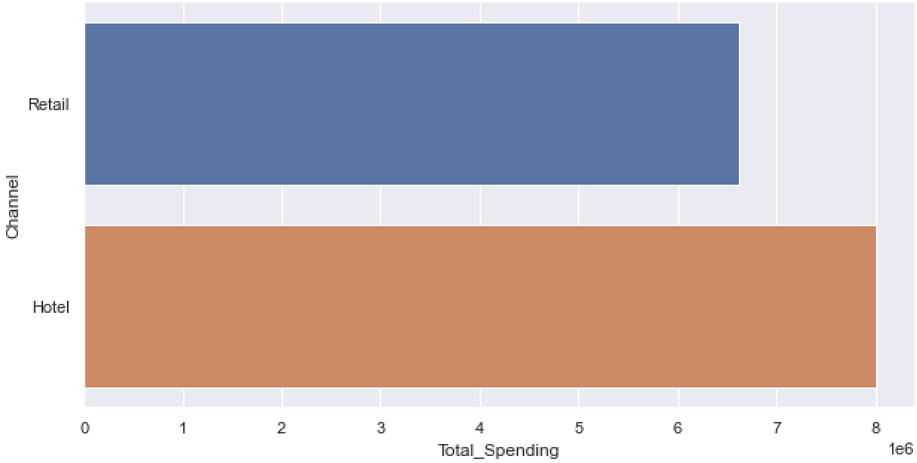


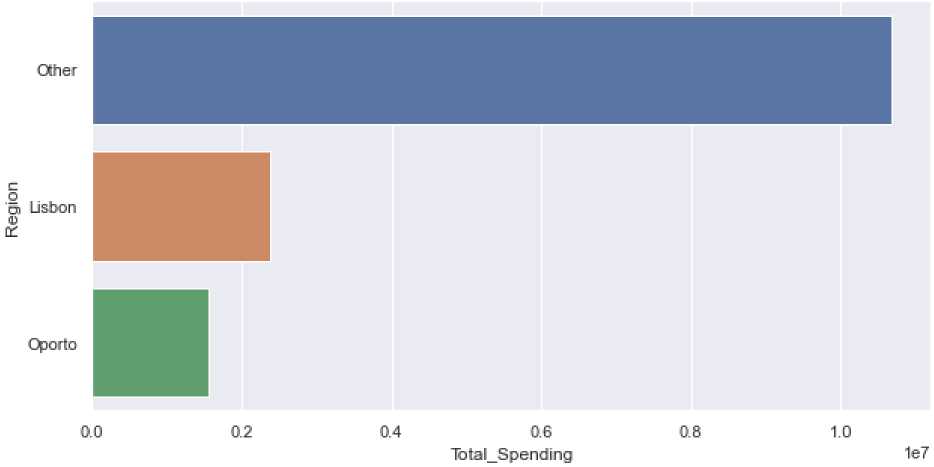
Which Region and which Channel spent the most? Which Region and which Channel spent the least?

As inferred let’s find out whether the Region ‘other’ & Channel ‘Hotel’ contribute most in the annual spending of the items mentioned .

In order to find that we first have to know the total spending of the items for each buyer/Spender as below:

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total_Spending
0	1	Retail	Other	12669	9656	7561	214	2674	1338	34112
1	2	Retail	Other	7057	9810	9568	1762	3293	1776	33266
2	3	Retail	Other	6353	8808	7684	2405	3516	7844	36610
3	4	Hotel	Other	13265	1196	4221	6404	507	1788	27381
4	5	Retail	Other	22615	5410	7198	3915	1777	5185	46100





From the above plot we can see that the Total Spending in the channel ‘Hotel’ and Region ‘Other’ has been the most . Whereas the Total Spending in the Channel ‘Retail ‘and Region ‘Oporto’ has been the least .

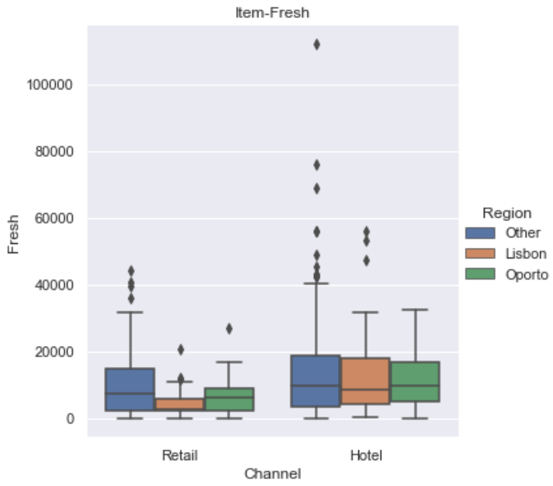
1.2 There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.

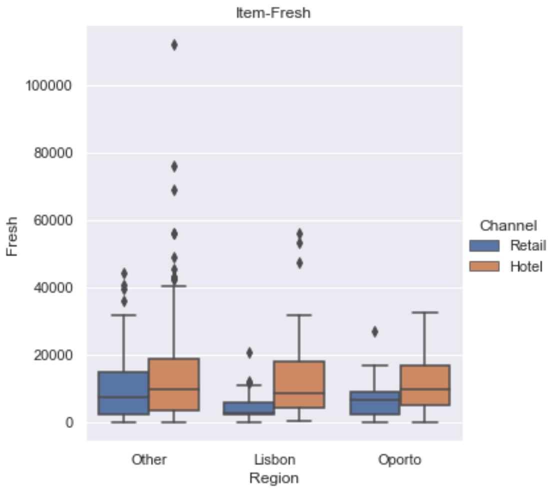
Let us now find out whether these 6 items available in the dataset show similar kind of behavior or do they vary across Region and Channel .

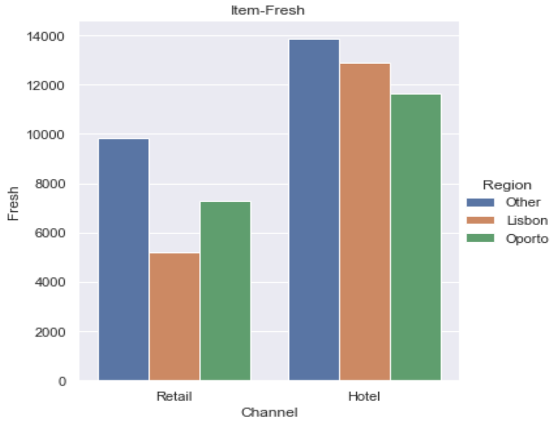
Lets check for Fresh items behavior as below :

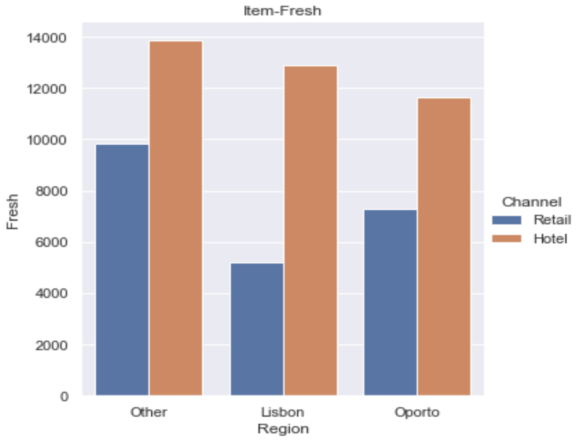
Region	Channel	count	mean	std	min	25%	50%	75%	max
Lisbon	Hotel	59	12902.25	12342.01	514	4437.5	8656	18135	56083
	Retail	18	5200	5415.521	18	2378.25	2926	5988	20782
Oporto	Hotel	28	11650.54	8969.363	3	4938.25	9787	17031.5	32717
	Retail	19	7289.789	6867.935	161	2368	6468	9162	27082
Other	Hotel	211	13878.05	14746.57	3	3702.5	9612	18821	112151
	Retail	105	9831.505	9635.394	23	2343	7362	15076	44466

We can see there is a positive skewness in the data , now lets visualize the above information with the below graphs.









From the above plot we can identify the outliers in different regions and channels, which seems to be the most in 'Other' location & 'Hotel' Channels. Whereas in Retail channel and the Oporto region seems to be significantly lesser comparatively in terms of expenditure on Fresh items.

We can also visualize from the bar graph that the over all annual expenditure on 'Fresh Item' is more through the 'Hotel' Channels in all the regions , but the most of it being from the 'Other region .

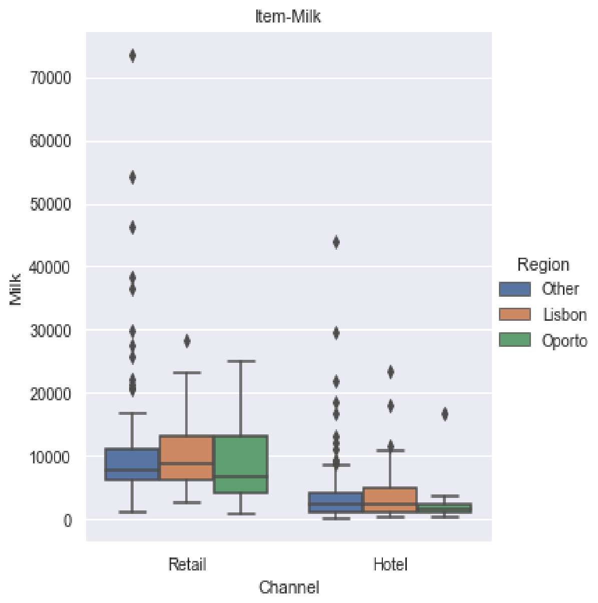
Let check for the variations the data for Fresh items across Regions and Channels :

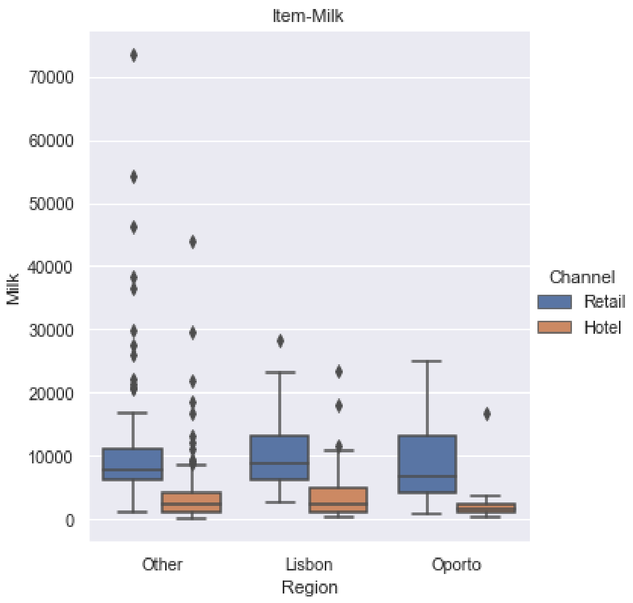
Item-Fresh							
count	mean	std	min	25%	50%	75%	max
440	12000.3	12647.33	3	3127.75	8504	16933.75	112151

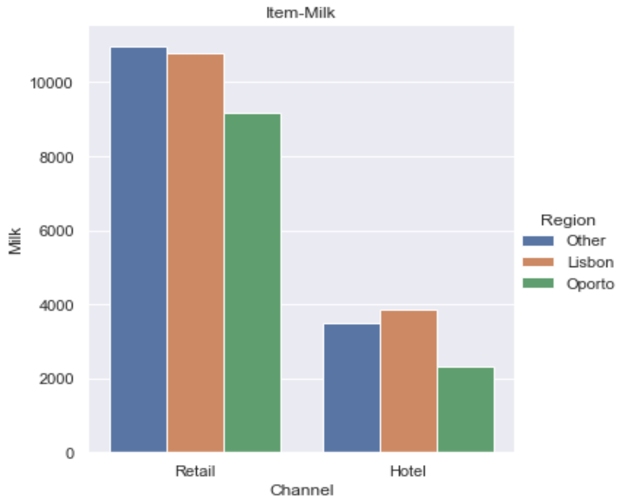
Range= max-min : 112148 , IQR=Q3-Q1= 13806, Q2= 8504

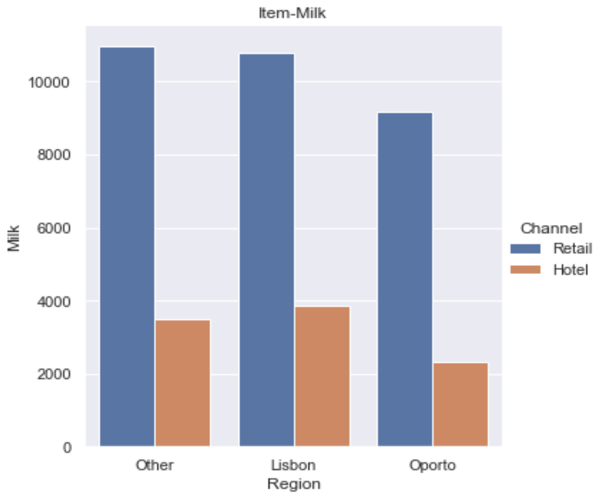
Lets check the behaviour of Milk Item now:

Region	Channel	count	mean	std	min	25%	50%	75%	max
Lisbon	Hotel	59	3870.203	4298.321	258	1071	2280	4995.5	23527
	Retail	18	10784	6609.221	2527	6253.25	8866	13112.25	28326
Oporto	Hotel	28	2304.25	2968.629	333	1146	1560.5	2344.75	16784
	Retail	19	9190.789	6611.354	928	4148.5	6817	13127.5	25071
Other	Hotel	211	3486.981	4508.505	55	1188.5	2247	4205	43950
	Retail	105	10981.01	10574.83	1124	6128	7845	11114	73498









From the above plot we can identify the outliers in different regions and channels, which seems to be the most in ‘Other’ location & ‘Retail’ Channels.

We can also visualize from the bar graph that the over all annual expenditure on ‘Milk Item’ is more through the ‘Retail’ Channels in all the regions , but the most of it being from the ‘Other region . Whereas the spending’s’ from the ‘Hotel’ channels are more in ‘Lisbon’ Region

The average expenditure through the retail channel is 10716.50 compared to the Hotel channel which is 3452.

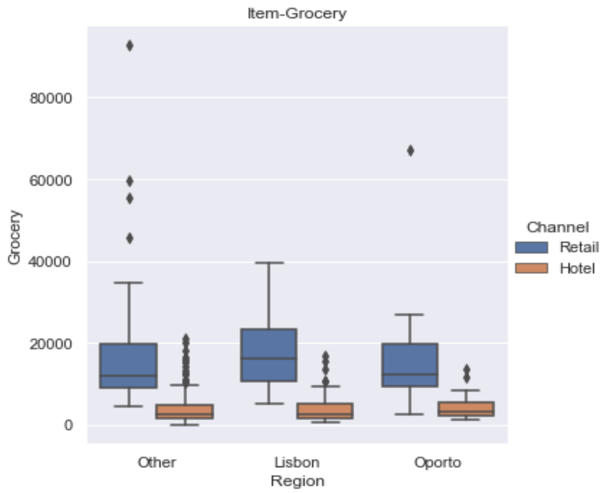
Lets check for the variations in the data for Milk item across Regions and Channels :

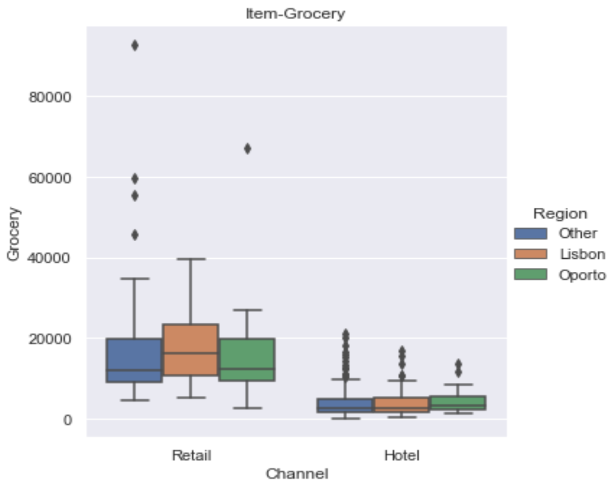
Milk							
Count	mean	std	min	25%	50%	75%	max
440	5796.266	7380.377	55	1533	3627	7190.25	73498

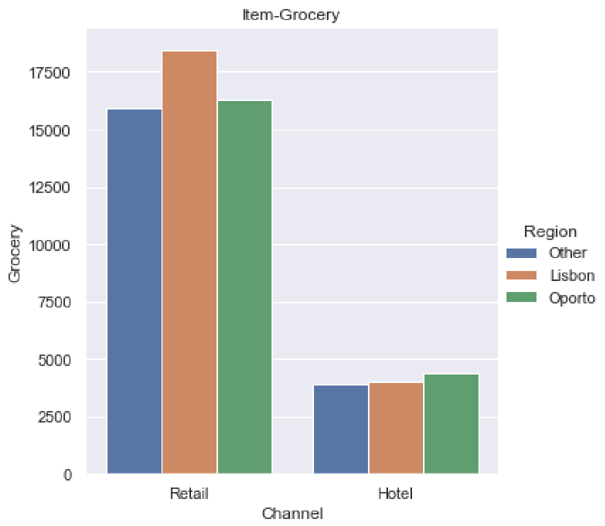
Range= max-min : 73443 , IQR=Q3-Q1= 5657.25, Q2= 3627

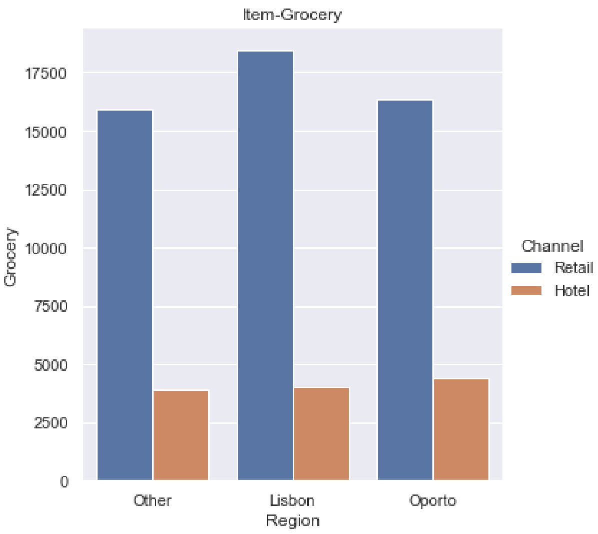
Let us now check the behaviour of Grocery items:

Region	Channel	count	mean	std	min	25%	50%	75%	max
Lisbon	Hotel	59	4026.136	3629.644	489	1620	2576	5172.5	16966
	Retail	18	18471.94	10414.69	5265	10634.25	16106	23478.75	39694
Oporto	Hotel	28	4395.5	3048.299	1330	2373.75	3352	5527.5	13626
	Retail	19	16326.32	14035.45	2743	9318.5	12469	19785.5	67298
Other	Hotel	211	3886.735	3593.506	3	1666	2642	4927.5	21042
	Retail	105	15953.81	12298.94	4523	9170	12121	19805	92780









From the above plot we can identify the outliers in different regions and channels, which seems to be the most in 'Other' location & 'Retail' Channels.

We can also visualize from the bar graph that the over all annual expenditure on 'Grocery Item' is more through the 'Retail' Channels in all the regions , but the most of it being from the 'Lisbon' region, then Oporto and lastly 'Other' regions . Whereas spending from the Hotel channels are almost equally distributed across all regions .

The average expenditure through the retail channel is 16322.85 compared to the Hotel channel which is 3962.14.

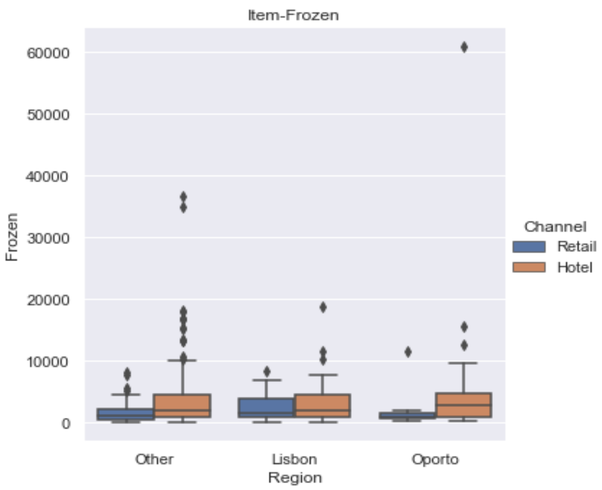
Lets check for the variations in the data for Grocery item across Regions and Channels :

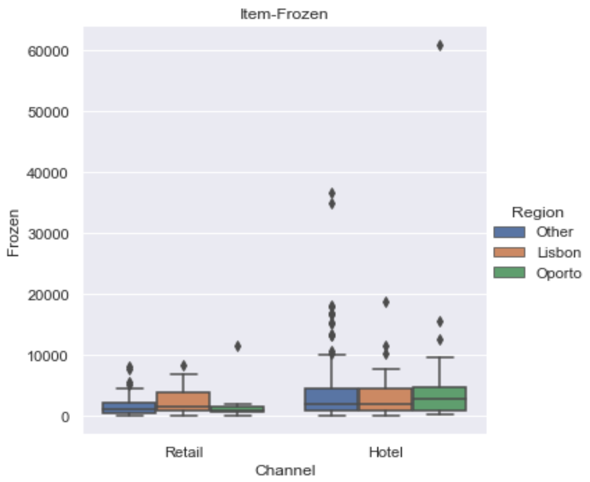
Grocery							
count	mean	std	min	25%	50%	75%	max
440	7951.277	9503.163	3	2153	4755.5	10655.75	92780

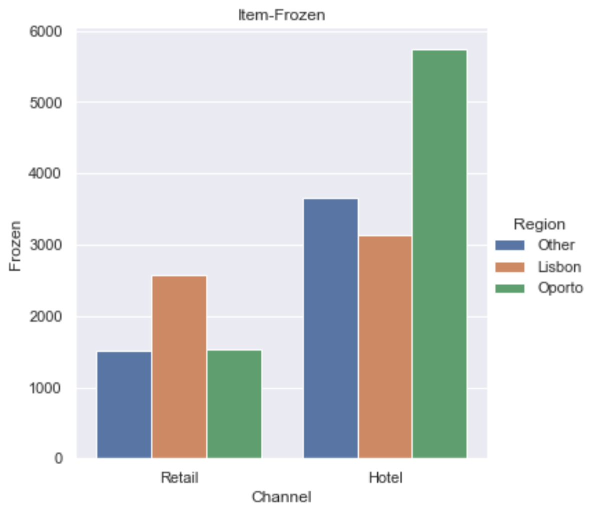
Range= max-min : 92777 , IQR=Q3-Q1= 8502.75, Q2= 4755.5

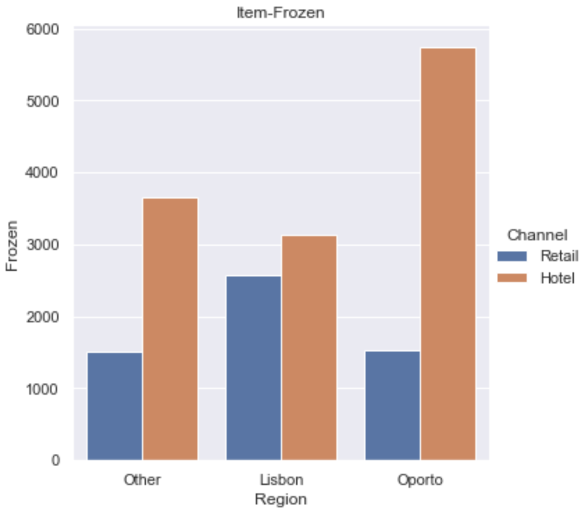
Lets check the behavior for Frozen Items:

Region	Channel	count	mean	std	min	25%	50%	75%	max
Lisbon	Hotel	59	3127.322	3276.46	91	966	1859	4479	18711
	Retail	18	2584.111	2424.775	61	923.5	1522	3843	8321
Oporto	Hotel	28	5745.036	11454.48	264	962.25	2696.5	4617	60869
	Retail	19	1540.579	2473.266	131	639.5	934	1410	11559
Other	Hotel	211	3656.9	4956.591	25	779	1960	4542.5	36534
	Retail	105	1513.2	1504.499	33	437	1059	2194	8132









From the above plot we can visualize that the overall annual expenditure on ‘Frozen Item’ is more through the ‘Hotel’ Channels in all the regions , but the most of it being from the ‘Oporto’ region, then Other and lastly ‘Lisbon’ regions . Whereas in the Retail Channel we can see more spending are from ‘Lisbon’ Region.

The average expenditure through the Hotel channel is 1652.61 compared to the Retail channel which is 3748.25.

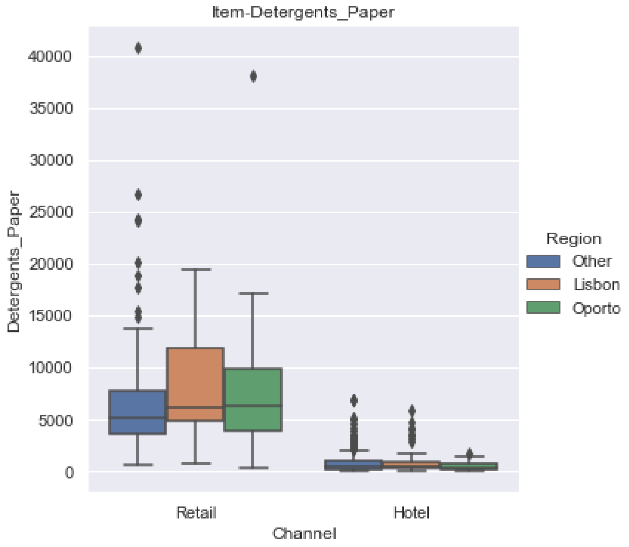
Lets check for the variations in the data for ‘Frozen’ item across Regions and Channels :

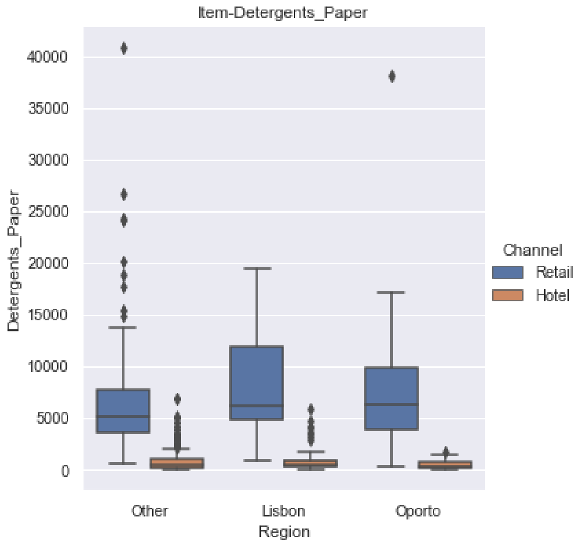
Frozen							
count	mean	std	min	25%	50%	75%	max
440	3071.932	4854.673	25	742.25	1526	3554.25	60869

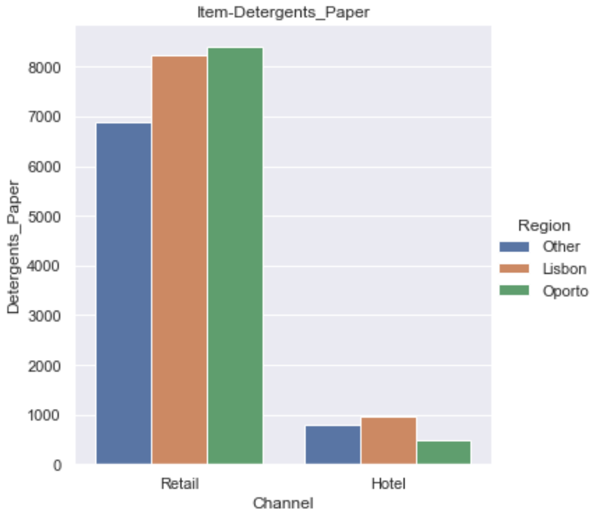
Range= max-min : 60844 , IQR=Q3-Q1= 2812, Q2= 1526

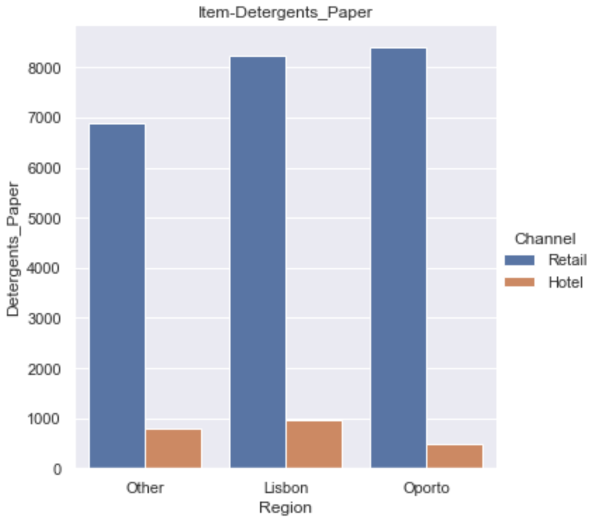
Let us now check the behaviour of Detergents Paper :

Region	Channel	count	mean	std	min	25%	50%	75%	max
Lisbon	Hotel	59	950.5254	1305.908	5	237	412	874	5828
	Retail	18	8225.278	5515.879	788	4818.25	6177	11804.75	19410
Oporto	Hotel	28	482.7143	425.3105	15	182.75	325	707	1679
	Retail	19	8410.263	8286.748	332	3900	6236	9837.5	38102
Other	Hotel	211	786.6825	1099.971	3	176.5	375	948.5	6907
	Retail	105	6899.238	6022.091	523	3537	5121	7677	40827









From the above plot we can visualize that the overall annual expenditure on 'Detergent Paper Item' is more through the 'Retail' Channels in all the regions , but the most of it being from the 'Oporto' region, then 'Lisbon and lastly 'Other' regions . Whereas in the Hotel Channel we can see spending are almost minimal for Detergents Paper .

The average expenditure through the Retail channel is 7269.51 compared to the Hotel channel which is 790.560.

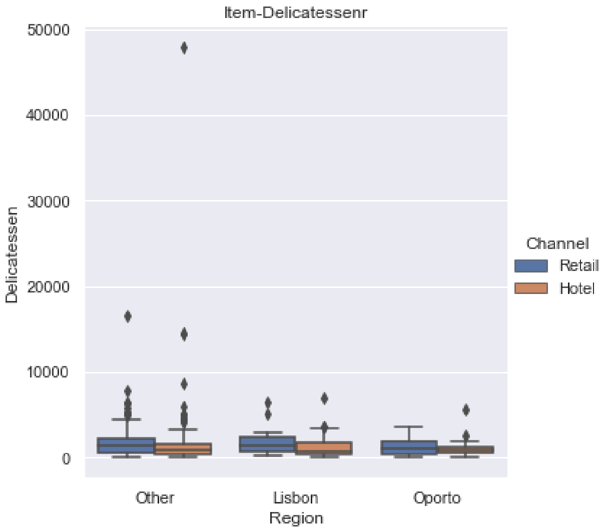
Lets check for the variations in the data for 'Detergents Paper' item across Regions and Channels :

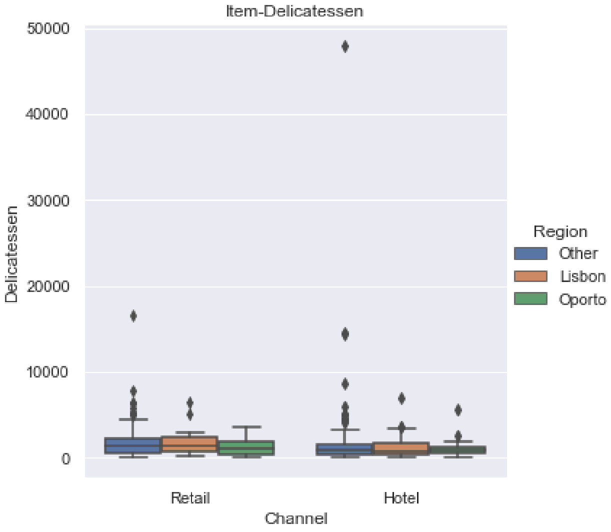
Detergents_Paper							
count	mean	std	min	25%	50%	75%	max
440	2881.493	4767.854	3	256.75	816.5	3922	40827

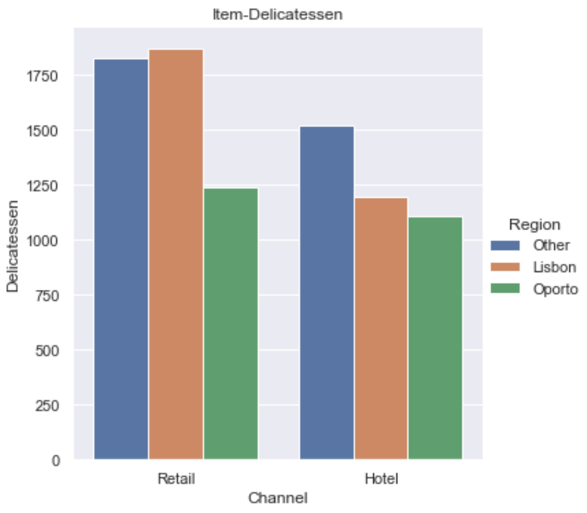
Range= max-min : 40824 , IQR=Q3-Q1= 3665.25, Q2= 816.5

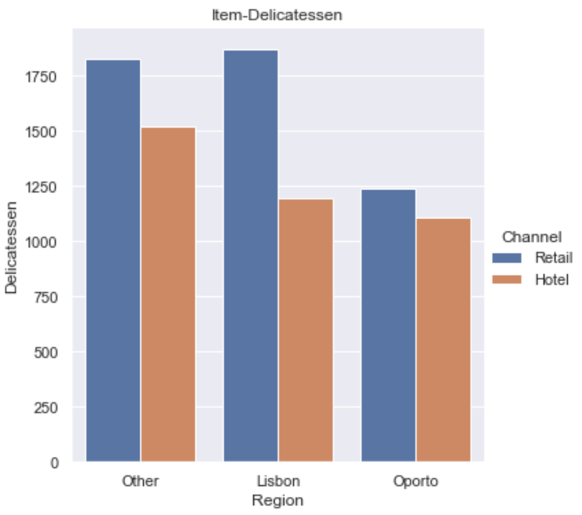
Lastly lets us now check the behaviour of Delicatessen:

Region	Channel	count	mean	std	min	25%	50%	75%	max
Lisbon	Hotel	59	1197.153	1219.945	7	374	749	1621.5	6854
	Retail	18	1871.944	1626.487	120	746	1414	2456.5	6372
Oporto	Hotel	28	1105.893	1056.779	51	567.25	883	1146	5609
	Retail	19	1239	1065.438	59	392.5	1037	1815	3508
Other	Hotel	211	1518.284	3663.183	3	378.5	823	1582	47943
	Retail	105	1826.21	2119.052	3	545	1386	2158	16523









From the above plot we can visualize that the overall annual expenditure on ‘Delicatessen’ is more through the ‘Retail’ Channels in all the regions , but the mostly being from the ‘Lisbon’ region, then ‘other and lastly ‘Oporto’ region . Whereas in the Hotel Channel we can see spending are more from other Regions , then Lisbon and lastly Oporto .

The average expenditure through the Retail channel is 1753.44 compared to the Hotel channel which is 1415.96.

Lets check for the variations in the data for ‘Delicatessen’ item across Regions and Channels :

Delicatessen							
count	mean	std	min	25%	50%	75%	max
440	1524.87	2820.106	3	408.25	965.5	1820.25	47943

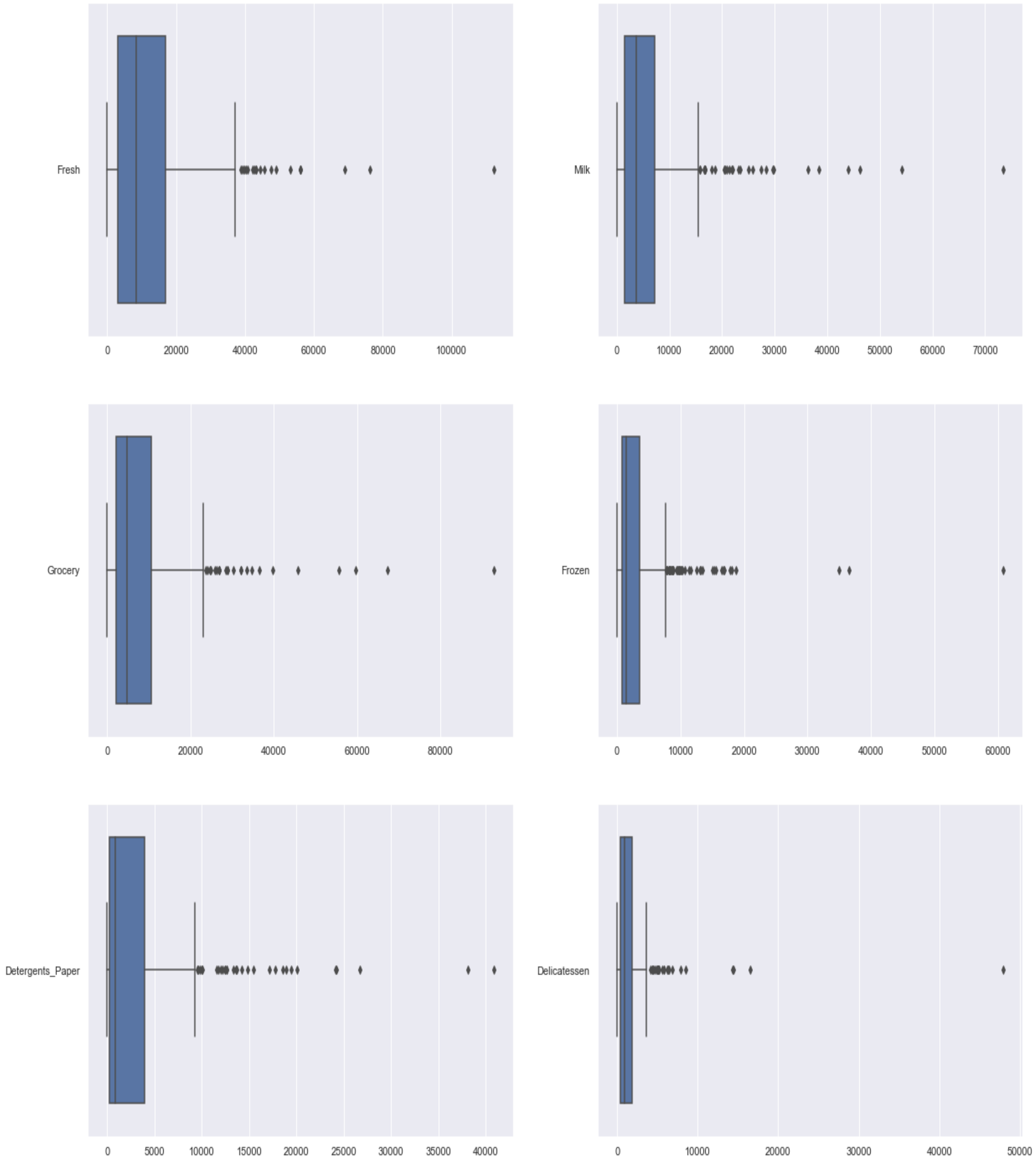
Range= max-min : 47940 , IQR=Q3-Q1= 1412, Q2= 965.5.

<u>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?</u>

CV for Fresh item expenditure is 1.0539179237473149
CV for Milk expenditure is 1.2732985840065414
CV for Grocery expenditure is 1.1951743730016824
CV for Frozen Item expenditure is 1.5803323836352914
CV for Detergents Paper expenditure is 1.6546471385005155
CV for Delicatessen expenditure is 1.8494068981158382

Since the Coefficient of Variation is the lowest for ‘Fresh’ item it is the least inconsistent, and since the c oefficient of variation is the highest for ‘Delicatessen’ it is the most inconsistent .

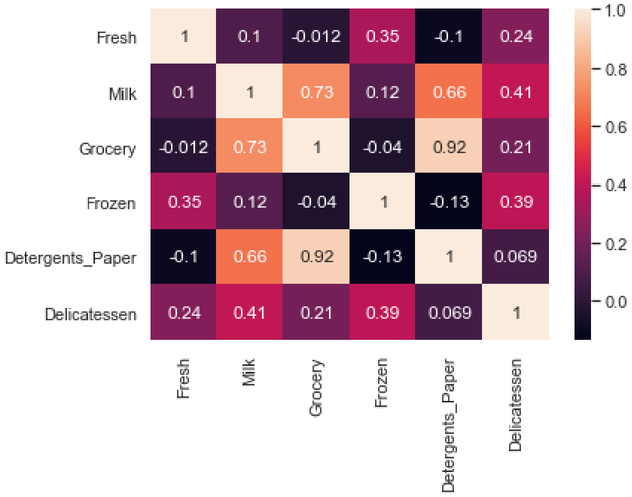
1.4 Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.



We can clearly see that there are outliers across the range of item Fresh, Milk, Grocery, Frozen, Detergents Paper & Delicatessen. The outliers show a pattern wherein we can conclude that all the variables show right or positive skewness.

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.

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Fresh	1	0.10051	-0.01185	0.345881	-0.101952938	0.24468997
Milk	0.10051	1	0.728335	0.123994	0.661815679	0.40636832
Grocery	-0.01185	0.728335	1	-0.04019	0.924640691	0.20549651
Frozen	0.345881	0.123994	-0.04019	1	-0.131524906	0.39094747
Detergents_Paper	-0.10195	0.661816	0.924641	-0.13152	1	0.0692913
Delicatessen	0.24469	0.406368	0.205497	0.390947	0.069291297	1



On the basis of the above analysis conducted we can infer that item such as Milk, Grocery, and detergent paper are highly corelated and spending on these items are mostly through the retail stores. Therefore I would suggest to increase the availability of these items and have it stocked together in the retail stores for increased profitability. We can also see that the population from Lisbon are the highest consumers of Grocery Items , hence the focus to increase the profitability in that area should be more by making sure that there is sufficient stock of the three corelated items all the items .

On the other hand, Fresh and Frozen have higher consumption in the Hotel channel in comparison with the Retail channel across all regions. We can also summarize that the expenditure for Fresh and groceries is the maximum across region and channel while for Delicatessen it is the least. Therefore more focus should be given on the items like Frozen, Delicatessen, Detergent Paper.

There should be specific measure taken to minimize the gap in spending pattern through Hotel and Retail Channels.

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

Let us have a look at the data

	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
0	1	Female	20	Junior	Other	Yes	2.9	Full-Time	50	1	3	350	Laptop	200
1	2	Male	23	Senior	Managemer	Yes	3.6	Part-Time	25	1	4	360	Laptop	50
2	3	Male	21	Junior	Other	Yes	2.5	Part-Time	45	2	4	600	Laptop	200
3	4	Male	21	Junior	CIS	Yes	2.5	Full-Time	40	4	6	600	Laptop	250
4	5	Male	23	Senior	Other	Undecided	2.8	Unemployed	40	2	4	500	Laptop	100

2.1. For this data, construct the following contingency tables

2.1.1. Gender and Major

Gender/ Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
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

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

2.1.3. Gender and Employment

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

2.1.4. Gender and Computer

Gender/Computer	Desktop	Laptop	Tablet
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?

Female	33
Male	29
Total	62

The probability that randomly selected CMSU student will be male is 29/62 = 46.77%

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

The probability that a randomly selected CMSU student will be a female is 33/62=53.23%

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.

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

The Probability that a male student will major in Accounting is= 4/29, 13.79 %
The Probability that a male student will major in CIS is =1/29, 3.45 %
The Probability that a male student will major in Economics/Finance is= 4/29, 13.79 %
The Probability that a male student will major in International Business is=2/29, 6.9 %
The Probability that a male student will major in Management is=6/29, 20.69 %
The Probability that a male student will major in Other courses is=4/29, 13.79 %
The Probability that a male student will major in Retailing/Marketing is= 5/29, 17.24 %
The Probability that a male student will major in undecided course is=3/29, 10.34 %

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

The Probability that a female student will major in Accounting is=3/33, 9.09 %
The Probability that a female student will major in CIS is=3/33, 9.09 %
The Probability that a female student will major in Economics/Finance is=7/33, 21.21 %
The Probability that a female student will major in International Business is=4/33, 12.12 %
The Probability that a female student will major in Management is=4/33, 12.12 %
The Probability that a female student will major in Other is=3/33, 9.09 %
The Probability that a female student will major in Retailing/Marketing is=9/33, 27.27 %
The Probability that a female student will major in undecided is=0/33, 0.0 %

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.

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

P(Intends to graduate ∩ Male) = P (Intends to graduate| Male) x P (male),(17/29)*(29/62) = 0.27

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

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

P(No Laptop ∩ Female) = P (No Laptop| Female) x P (Female),(4/33)*(33/62) = 0.06

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

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

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

P(Full Time Employment U Male)=P (Full Time Employment) + P (Male)-P(Full Time Employment ∩ Male) ,(10/62)+(29/62) –(7/29*29/62)=0.161+0.4677-(0.241*0.4677) =0.5161

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

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

The probability that given a female student is randomly chosen , she is majoring in international business or management is =
33/62 * 8/33= 12.90

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/Grad Intention	No	Yes
Female	9	11
Male	3	17

P(A ∩ B) = P(A) * P(B)

So, P (Grad Intention ∩ Female) = P(Grad intention) * P(Female)

P(Female) = 20/40 = 0.5

P(Grad intention) = 28/40 =0.7

P(Grad Intention) * P(Female) = 0.5*0.7 = 0.35

P (Grad Intention ∩ Female) = 11/40 = 0.275

This is not independent events as probability multiplication of both events is not equal to combined event, so graduate intention and being female student are not independent events.

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 total number of students who have GPA less than 3 are 17.

Therefore, the probability that if a student is chosen randomly, the probability that his/her GPA is less than 3 would be 17/62 = 27.41%

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.

Attributes	**Earns less than 50**	**Earns 50 or more**	**Marginal Total**
Male	15	14	29
Female	15	18	33
Marginal Total	30	32	62

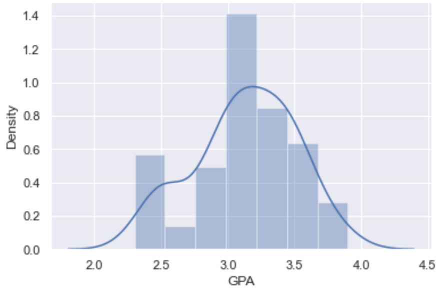
The conditional probability that a randomly selected male earns 50 or more is14/29 = 48.27%

The conditional probability that a randomly selected Female earns 50 or more is 18/33=54.54%

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 the variable 'GPA'

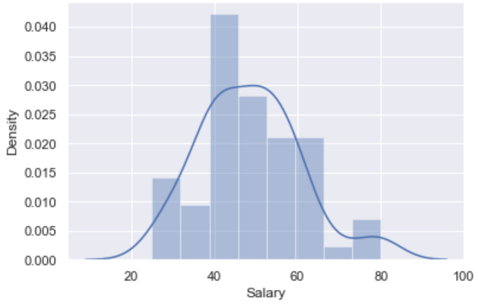
	count	mean	std	min	25%	50%	75%	max
GPA	62	3.129032	0.377388	2.3	2.9	3.15	3.4	3.9



The variable 'GPA' follows normal distribution , since it's a bell curve and the mean and median are almost the same .

For the variable 'Salary'

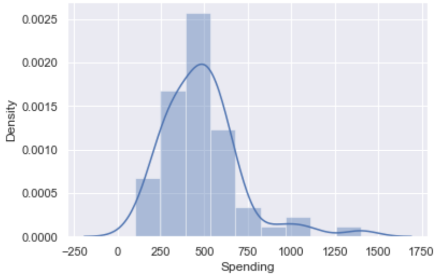
	count	mean	std	min	25%	50%	75%	max
Salary	62	48.54839	12.08091	25	40	50	55	80



The variable follows a normal distribution as the mean and median are close and it's a bell curve .

For variable 'Spending'

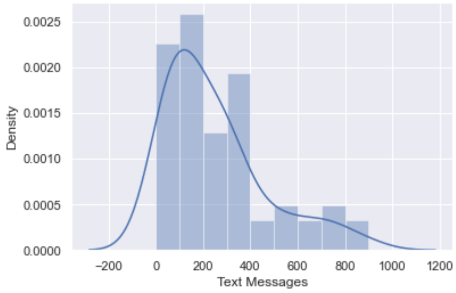
count	mean	std	min	25%	50%	75%	max
62	482.0161	221.9538	100	312.5	500	600	1400



The variable Spending follows a normal distribution as we can see a bell curve and the mean, median are almost equal.

For the variable 'Text Messages'

count	mean	std	min	25%	50%	75%	max
62	246.2097	214.466	0	100	200	300	900



We can see from the above plot here that the variable is right tailed or right skewed or has a positive skewness as there a large variation in the median and mean mentioned above. We can see there are potential outliers as well in this 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 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.

3.1 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.

For A Shingles :

Step 1: Defining the Null Hypothesis and Alternate Hypothesis

- Null Hypothesis = H0: mean moisture content <0.35
- Alternate Hypothesis=HA: mean moisture content >0.35

Step 2: Deciding the significance level

Here we select Alpha as 0.05 as nothing is specified in the question .

The sample size is 36 .

Step 3: Identifying the test statistic
We do not know the population SD and n= > 30 . So we use the t distribution test statistic.

Step 4: Calculating the p-value and test statistic

After computing the details in python we have got the following data

one sample t test
t statistic:-1.4735046253382782 , p value: 0.07477633144907513

p value > 0.05

Basis the hypothesis test performed for the given sample of 36 observations at 95% confidence level we fail to reject the null hypothesis. Hence conclude that the moisture content for A shingles is under permi ssible limits .

For B Shingles :

Step 1: Defining the Null Hypothesis and Alternate Hypothesis

- Null Hypothesis = H0: mean moisture content <0.35
- Alternate Hypothesis=HA: mean moisture content >0.35

Step 2: Deciding the significance level

Here we select Alpha as 0.05 as nothing is specified in the question .

The sample size is 30 .

Step 3: Identifying the test statistic
We do not know the population SD and n= 30 . So we use the t distribution test statistic.

Step 4: Calculating the p-value and test statistic.

After computing the details in python we have got the following data

one sample t test
t statistic:-3.1003313069986995, p value:0.0020904774003191813

p value < 0.05

Basis the hypothesis test performed for the given sample of 30 observations at 95% confidence level we reject the null hypothesis. Hence conclude that the moisture content for B shingles is not under permissi ble limits but higher than 0.35.

3.2 Do you think that the population mean for shingles A and B are equal? Form the hypothesis and co nduct the test of the hypothesis. What assumption do you need to check before the test for equality of means is performed?

Step 1: Defining the Null Hypothesis and Alternate Hypothesis

- Null Hypothesis = H0: Population mean for shingles A = Population mean for shingles B
- Alternate Hypothesis=HA: Population mean for shingles A ≠ Population mean for shingles B

Step 2: Deciding the significance level

Here we select Alpha as 0.05 as nothing is specified in the question .

Step 3: Identifying the test statistic

We do not know the population SD and we have two sample which are independent of each other . So we use the test statistic for two sample unpaired test.

Step 4: Calculating the p-value and test statistic.

After computing the details in python we have got the following data

two sample t test
tstat =1.289628271966112
P Value= 0.2017496571835328

p value > 0.05

Basis the hypothesis test performed for the given sample of A &B observations at 95% confidence level we fail to reject the null hypothesis. Hence conclude that Population mean for shingles A is equal to Population mean for shingles B.
The assumptions for conducting the above two sample test is that the population of both sample is normally distributed, and that the variances are the same.

****