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?

Cross tab by region and channel to understand the spread of spend.

Channel	Hotel	Retail	All
Region			
Lisbon	1538342	848471	2386813
Oporto	719150	835938	1555088
Other	5742077	4935522	10677599
All	7999569	6619931	14619500

- ✓ Other regions are spending significantly more than Lisbon and Oporto put together.
- ✓ Hotels are spending more than retail going by total sales value across regions. Oporto alone has its retail spending slightly better than the hotel.

Additional insights as below:

- ✓ Among Lisbon and Opporto, Lisbon spends higher from total sales volume perspective.
- ✓ Lisbon's hotels have highest spend proportion against its retail channel by doing about 80% more than retail. Other region's hotel channel does around 16% more than its retail channel

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

Cross tab by region and channel by each of the 6 variety

Region	**Lisbon**	**Oporto**	**Other**	
	Channel			
Delicatessen	Hotel	70632	30965	320358
	Retail	33695	23541	191752
Detergents_Paper	Hotel	56081	13516	165990
	Retail	148055	159795	724420
Fresh	Hotel	761233	326215	2928269
	Retail	93600	138506	1032308
Frozen	Hotel	184512	160861	771606
	Retail	46514	29271	158886
Grocery	Hotel	237542	123074	820101

	Retail	332495	310200	1675150
Milk	Hotel	228342	64519	735753
	Retail	194112	174625	1153006

Crosstab purely by channel level alone for each of the 6 variety

Channel	**Hotel**	**Retail**	**All**
Delicatessen	421955	248988	670943
Detergents_Paper	235587	1032270	1267857
Fresh	4015717	1264414	5280131
Frozen	1116979	234671	1351650
Grocery	1180717	2317845	3498562
Milk	1028614	1521743	2550357

✓ Other regions show good number of outliers for every variety.

✓ Milk, grocery and detergents_paper have wider gap in spend pattern across channels with retail on the higher side.

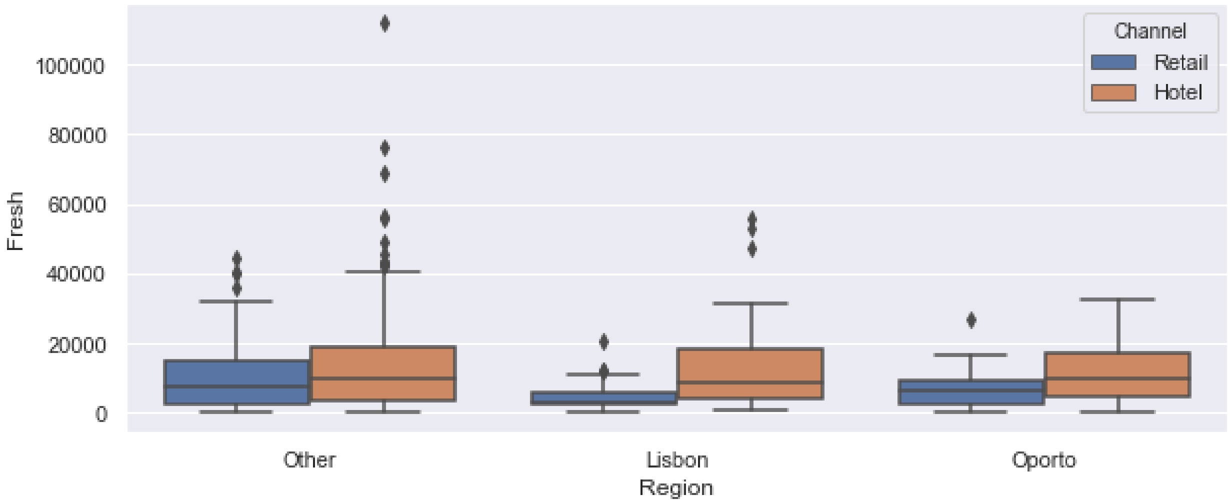
✓ Fresh, grocery, frozen and delicatessen show a common trend of higher sales value in hotel channel in comparison to retail channel across all regions.

✓ Grocery spend is considerably lower in hotels across the region compared to the retail.

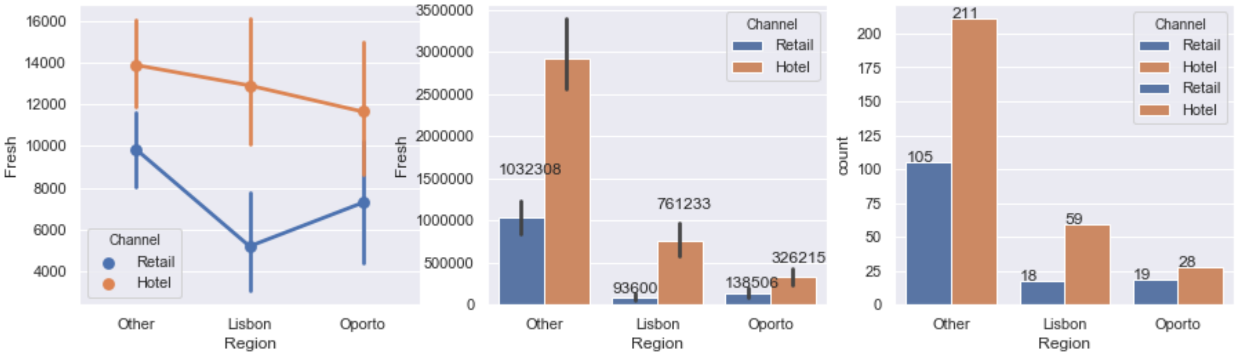
✓ Delicatessen's spend across channels are pretty close across the regions.

✓ Almost 4 varieties have their 3 quartiles within 10000 with 2 of them above that but less than 20000 leaving significant gaps with the maximum spend for that variety.

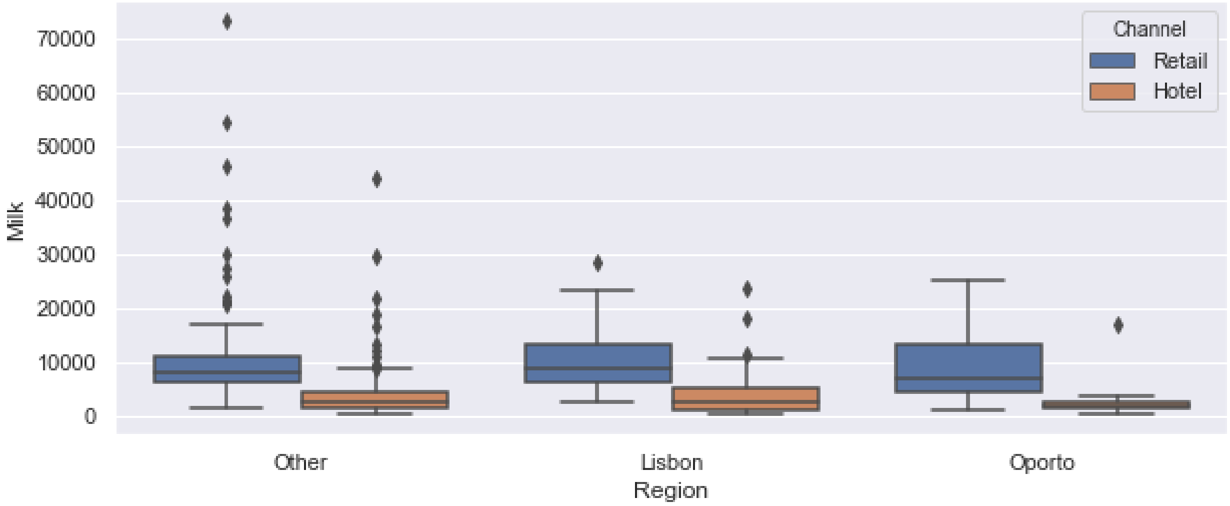
Please find below the box plot, point plot and bar plot by region and channel for each variety towards depicting the outlier trend, mean sales across region by channel and total sales value by region and channel. Also a side by side transaction count view by region and channel is extended.



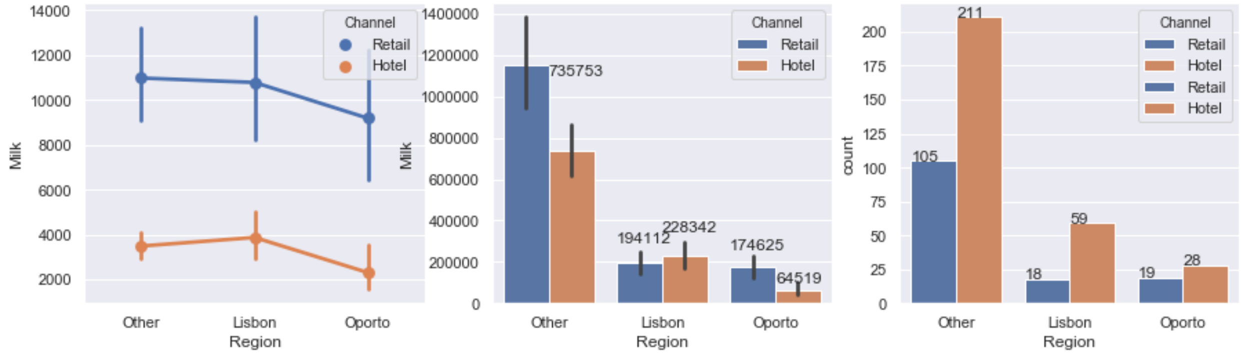
Channel
Retail
Hotel
Fresh
100000
80000
60000
40000
20000
0
Other
Lisbon
Oporto
Region

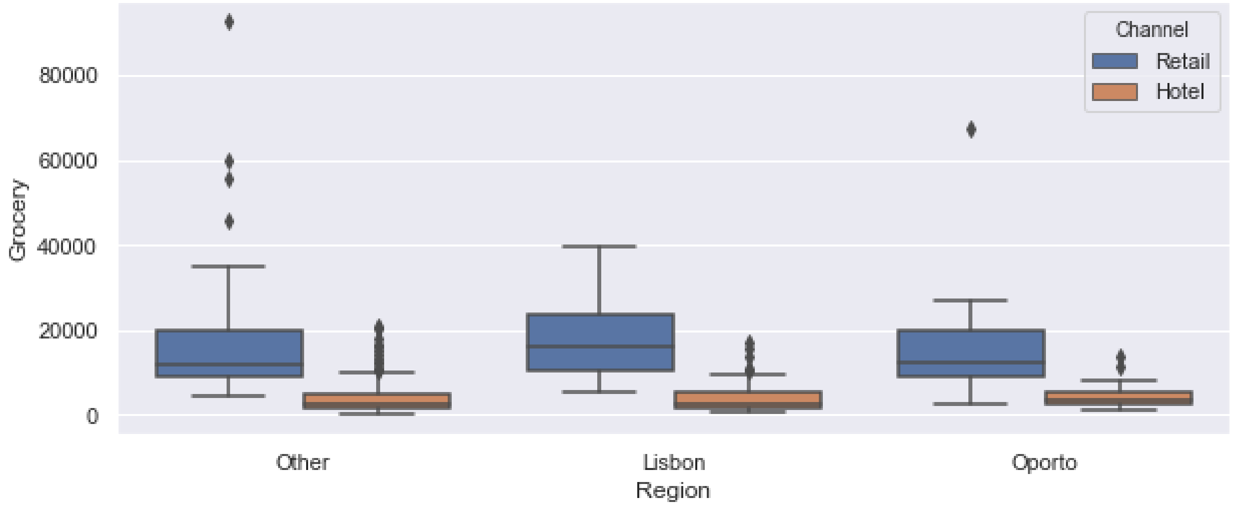


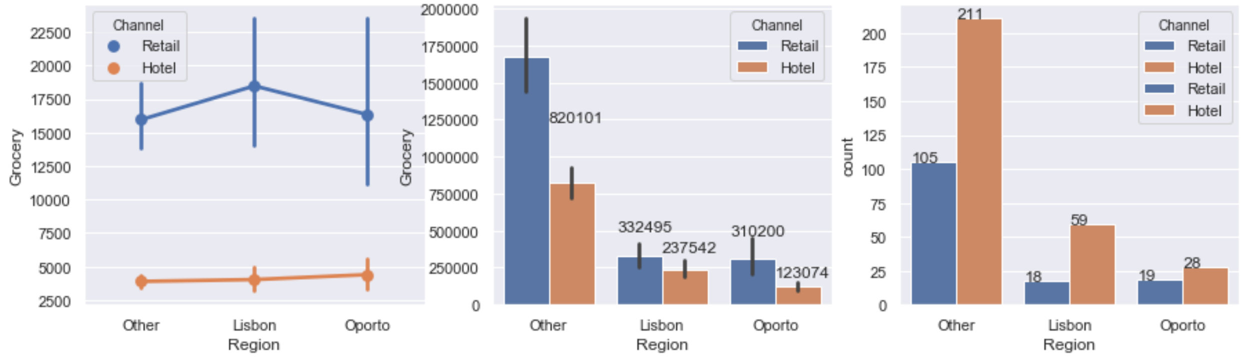
16000
14000
12000
10000
8000
6000
4000
Fresh
Channel
Retail
Hotel
Other
Lisbon
Oporto
Region
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
Fresh
Channel
Retail
Hotel
1032308
761233
326215
93600
138506
Other
Lisbon
Oporto
Region
200
175
150
125
100
75
50
25
0
count
Channel
Retail
Hotel
Retail
Hotel
211
105
59
18
19
28
Other
Lisbon
Oporto
Region

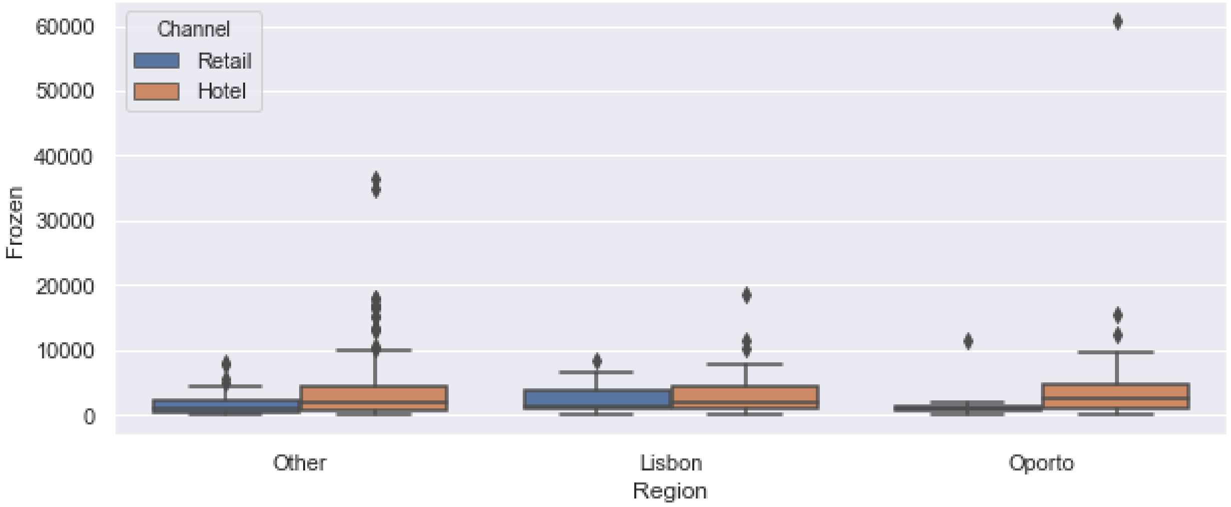


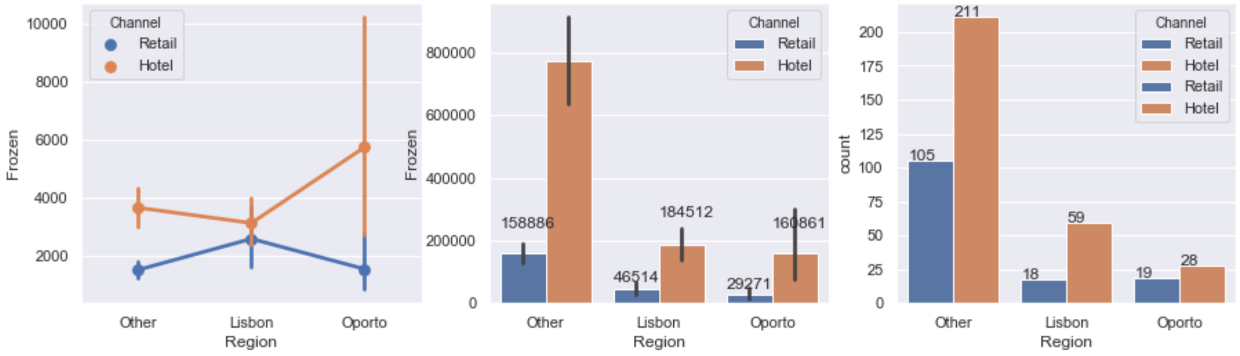
Channel
Retail
Hotel
Milk
70000
60000
50000
40000
30000
20000
10000
0
Other
Lisbon
Oporto
Region

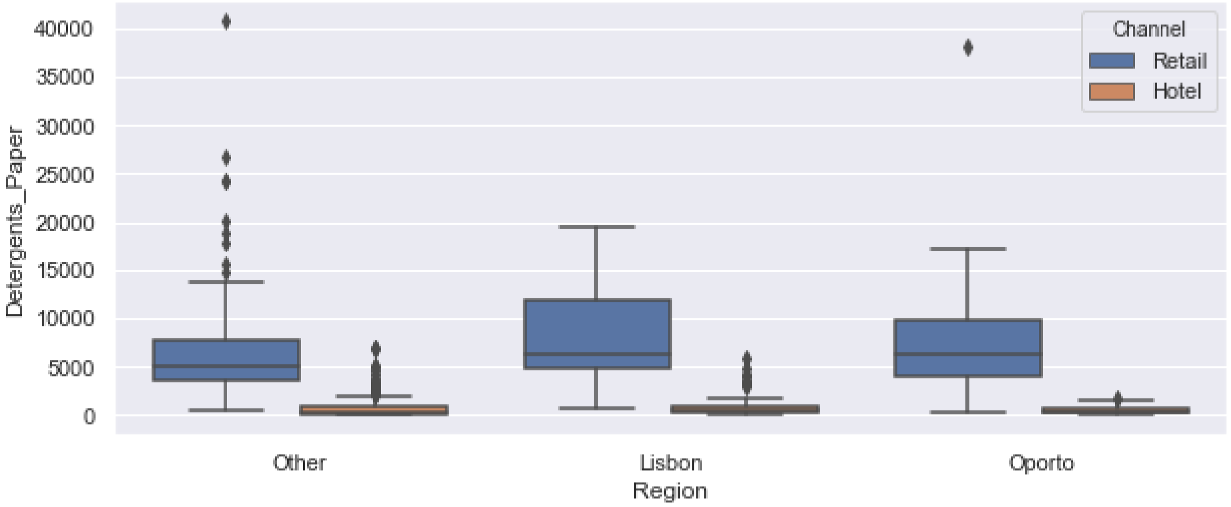

Milk
14000
12000
10000
8000
6000
4000
2000
Other
Lisbon
Oporto
Region
Channel
Retail
Hotel
Milk
1400000
1200000
1000000
800000
600000
400000
200000
0
735753
194112
228342
174625
64519
Other
Lisbon
Oporto
Region
Channel
Retail
Hotel
count
200
175
150
125
100
75
50
25
0
211
105
59
18
19
28
Other
Lisbon
Oporto
Region
Channel
Retail
Hotel
Retail
Hotel

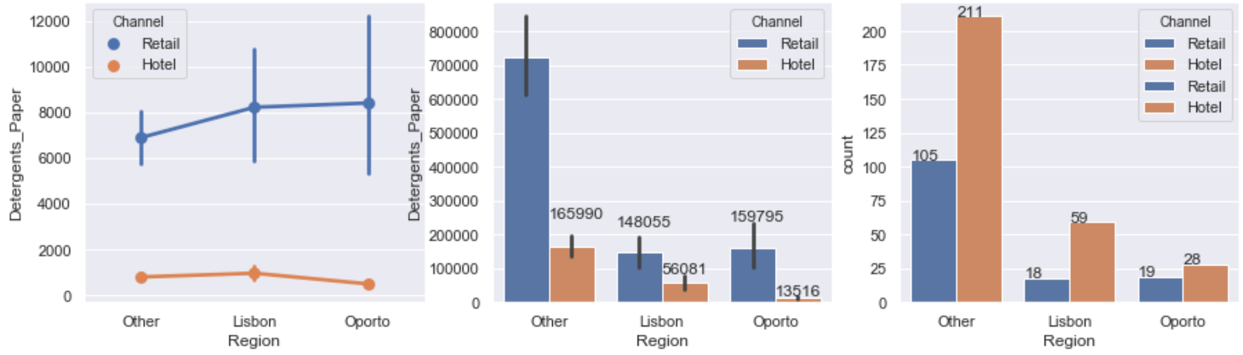

Grocery
80000
60000
40000
20000
0
Other
Lisbon
Oporto
Region
Channel
Retail
Hotel

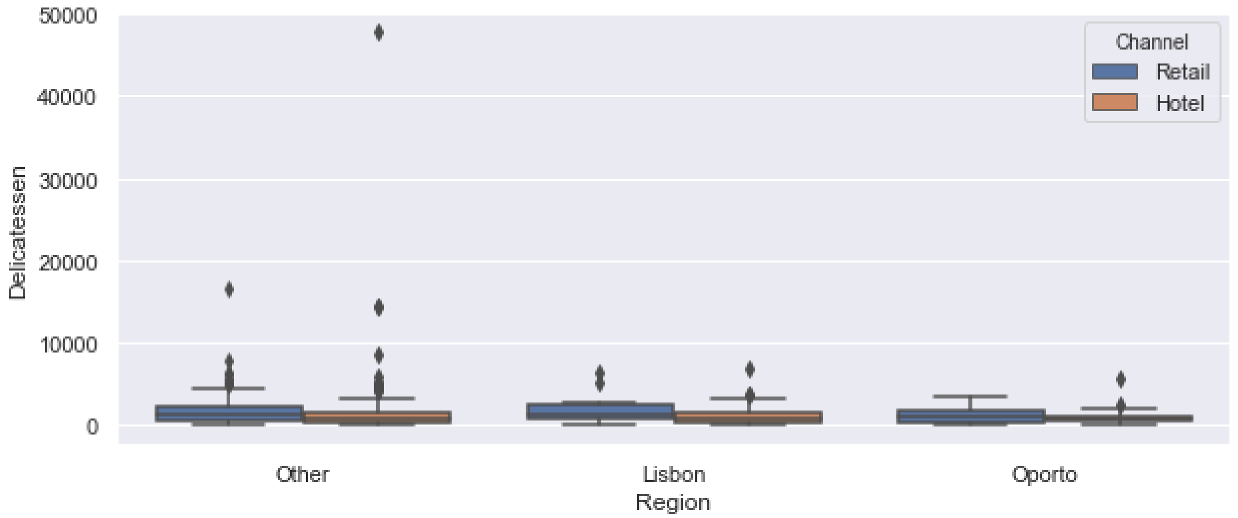

Grocery
22500
20000
17500
15000
12500
10000
7500
5000
2500
Other
Lisbon
Oporto
Region
Channel
Retail
Hotel
Grocery
2000000
1750000
1500000
1250000
1000000
750000
500000
250000
0
820101
332495
237542
310200
123074
Other
Lisbon
Oporto
Region
Channel
Retail
Hotel
count
200
175
150
125
100
75
50
25
0
211
105
59
18
19
28
Other
Lisbon
Oporto
Region
Channel
Retail
Hotel
Retail
Hotel

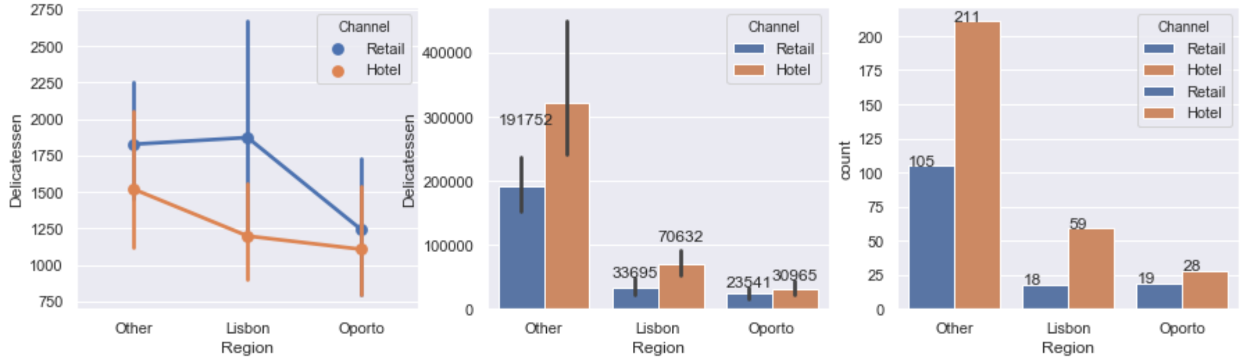

Channel
Retail
Hotel
60000
50000
40000
30000
20000
10000
0
Frozen
Other
Lisbon
Oporto
Region


Channel
Retail
Hotel
10000
8000
6000
4000
2000
Frozen
Other
Lisbon
Oporto
Region
Channel
Retail
Hotel
800000
600000
400000
200000
0
Frozen
158886
184512
160861
46514
29271
Other
Lisbon
Oporto
Region
Channel
Retail
Hotel
Retail
Hotel
200
175
150
125
100
75
50
25
0
count
211
105
59
18
19
28
Other
Lisbon
Oporto
Region


Channel
Retail
Hotel
40000
35000
30000
25000
20000
15000
10000
5000
0
Detergents_Paper
Other
Lisbon
Oporto
Region







1.3. On the basis of the descriptive measure of variability, which item shows the most inconsistent behaviour?

Which items shows the least inconsistent behavior?

Co efficient of variation (sample standard deviation divided by sample mean) would allow us to understand the relative dispersion which will enable measure of consistency here. Lesser the co efficient of variation higher the consistency and vice versa.

- ✓ Based on the coefficient of variation Fresh shows the most consistent behavior followed by Grocery
- ✓ Delicatessen shows the least inconsistent behavior followed by detergents_paper and frozen.

Below depicts the co efficient of variation for each item:

```
Items
Delicatessen        1.849407
Detergents_Paper    1.654647
Frozen              1.580332
Milk                1.273299
Grocery             1.195174
Fresh               1.053918
```

Note:
Going by relationship strength the Detergent paper and Grocery strongest positive co relation. This part is explained further in section 1.5.

1.4. Are there any outliers in the data?

- ✓ In general, good number of outliers are present in Other regions across all the items.
- ✓ Opporto has minimum outliers among the 3 regions.
- ✓ Retail channel depicts very minimal or no outliers across Lisbon and Opporto.

Note: Related charts are given as a response to question 1.2

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

Pre amble: Few analysis done as below towards final recommendation:

Region and channel level observation:

- ✓ Other regions top the chart in terms of total sales value and transaction count across the 3 regions followed by Lisbon and Opporto in that order.
- ✓ Looking at the aggregate sales value Hotels does more business as a channel for all regions put together. Among regions except for Opporto which does more in retail other regions align with that trend.
- ✓ From a channel wise transaction count perspective , Lisbon and Opporto does almost same in retail while for hotel channel Lisbon's count is double that of Opporto.
- ✓ From a channel wise sales value perspective, Opporto's retail sales value is greater than Lisbon by approximately 40% despite both of their transaction counts being congruent to each other. i.e Lisbon's mean sales value is higher than Opporto.
- ✓ Overall retail sales value is right skewed for other region and Opporto. Other region is right skewed in both the channels.
- ✓ Other regions have good number of outliers across the channels with Lisbon and Opporto has some outliers.

✓ Hotel sales for Lisbon and Opporto tend to be about normally distributed except for few outliers which Lisbon has more than Opporto while both have their sales distribution just about normal.

Observations at variety level:

✓ Except for frozen and fresh, hotels fall short of retail in terms of sales value and transaction count across the items.

✓ Sales value for Fresh item is spread more consistently followed by groceries while delicatessen is least consistent followed by detergents_paper.

✓ For Fresh, Lisbon hotels does much better mean sales value compared to Opporto while Opporto does better in retail.

✓ For milk, Lisbon's does better mean sales value in hotels across all the regions. Opporto's mean is lesser for hotel channel while Lisbon does much lower mean sales for retail across regions.

✓ For grocery, there is a wider gap in sales value between hotels and retail with retail doing significantly better. In retail channel Lisbon does better mean across regions, while for hotel channel Opporto is highest despite much lower transaction count.

✓ For frozen, in hotel channel Opporto has highest mean sales value despite lower transaction count while it does much lesser than Lisbon when it comes to retail.

✓ For detergents_paper, in hotel channel Lisbon is almost 4 times higher in sales value than Opporto while its transaction count is double that of Opporto. In retail Other regions is doing much lower sales value.

✓ For delicatessen, in retail Lisbon does higher mean sales value across region while Opporto is noticed to do much lesser mean despite similar transaction count with Lisbon.

Recommendations:

✓ Improving market share in hotels for Opporto can be focussed upon as it is comparatively much lesser in terms of penetration.

✓ Lisbon has rooms to improve when it comes to retail sales business looking at their performance in comparison to Opporto but has potential.

✓ Fresh items will need a focus in the retail market for Lisbon as it is below average comparing other regions.

✓ Sales performance for milk needs further investigation on the low business among Opporto hotels looking in comparison to Lisbon's. Fact that both regions have no outliers the mean sales value is disproportionate between two regions for this channel.

✓ Mean sales for Frozen is pretty close between hotel and retail channel and can be looked up for common promotional programs across these 2 channels to improve the market share.

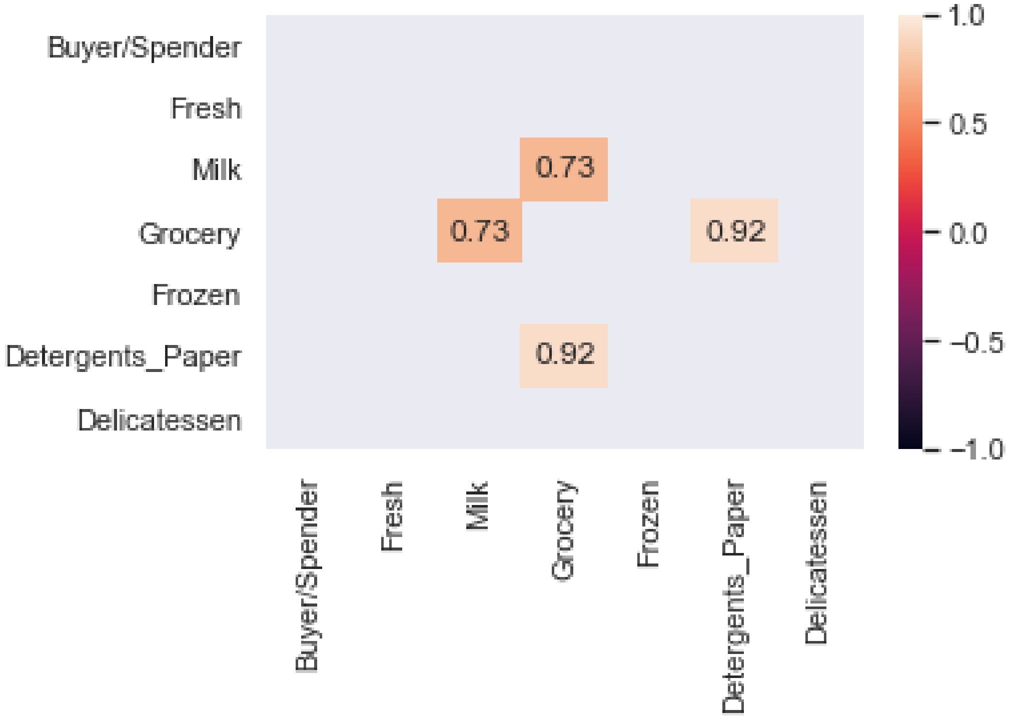
✓ Opporto hotel channel seems to be an attractive market for Frozen and hence more focus can be given in the retail sector too to take advantage towards improving revenue. Also there is a potential that retail business is being lost to a competitor and hence investments needs to be done to improve market share and has better return on investment potential.

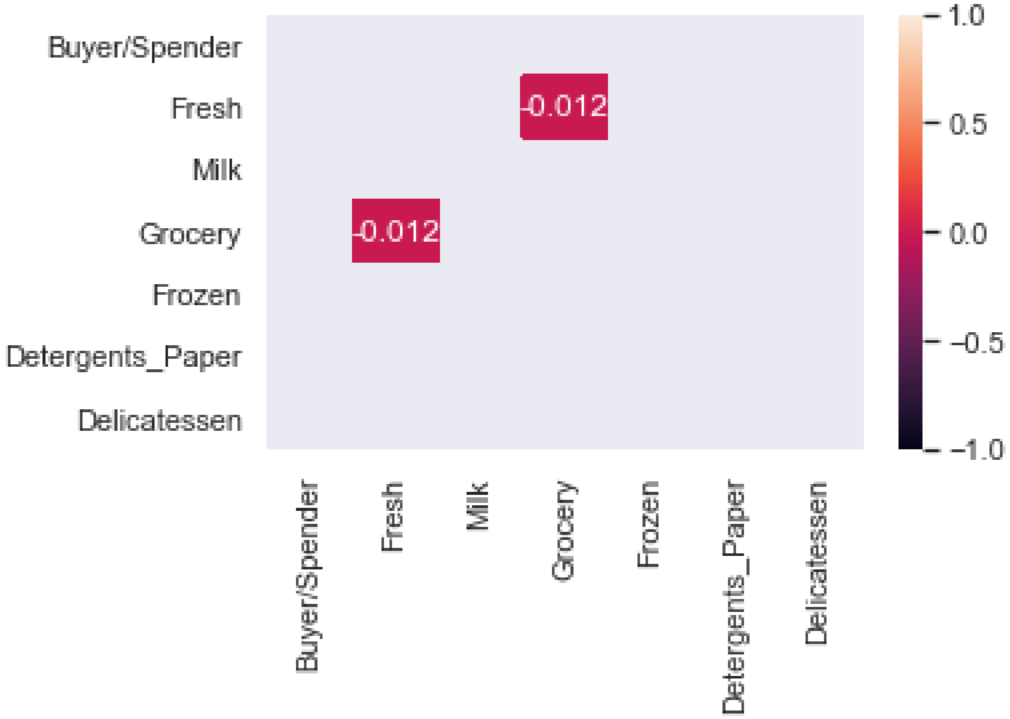
- ✓ Statistics indicate the detergents_paper business is not that great in Opporto hotel channel as it shows significantly low mean sales value comparatively. Also considering that their retail channel is having a much wider gap on their performance positively across regions, this opens up an option to investigate better product distribution into hotels and hence look at promotions accordingly.
- ✓ Delicatessen seems to have a similarities with overlaps on the error margin observed on the mean sales value across the channels and hence could look at how channels can elevate sales to take advantage with common promotional programs across channel especially in other regions and Opporto.
- ✓ Also going by the relationship strength, with detergents paper and grocery having strongest correlation suggests an opportunity to improve detergent paper sales by providing promotions through cross selling them along with Groceries. Recommendation is to explore new promotion models towards the same.
- ✓ With milk having better correlation to groceries, milk's sales can be looked upon in Lisbon region towards promotional activity as grocery is doing better than milk comparatively.

Supplemented charts:

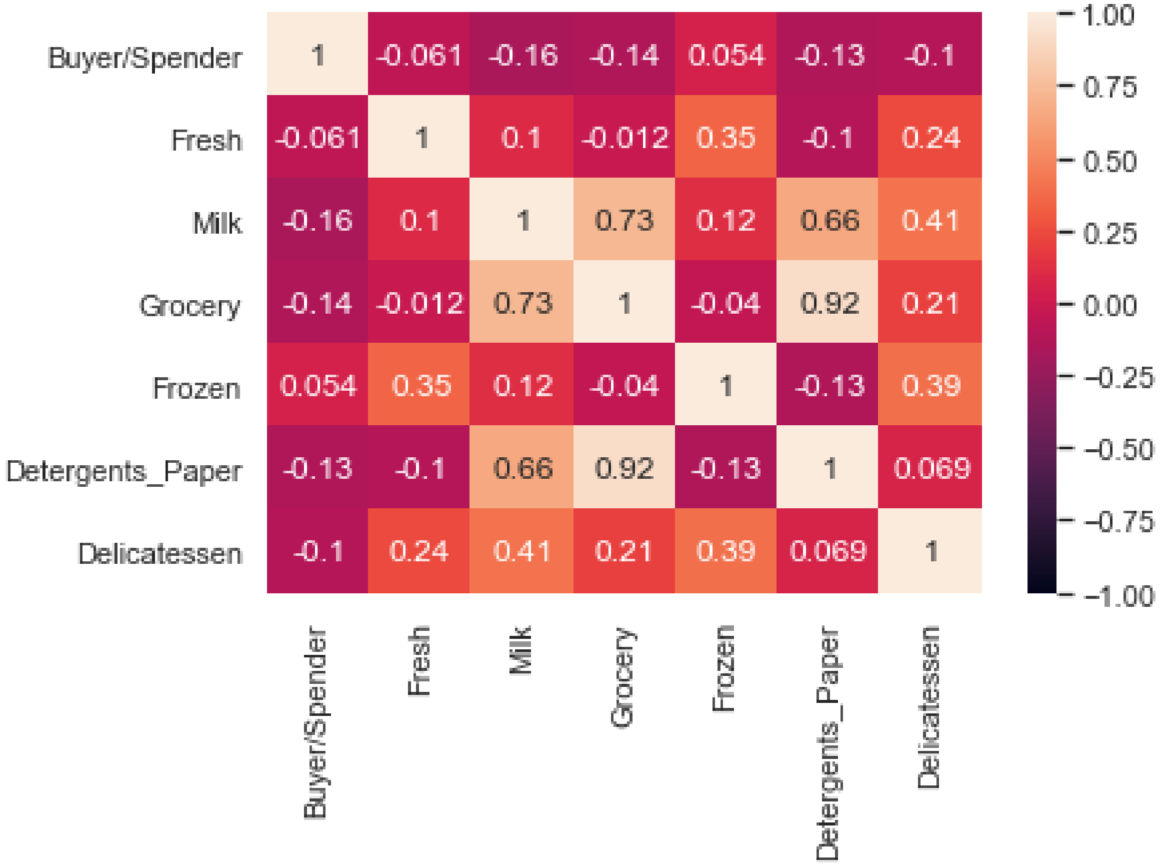
Below charts provide relationship strength among the items based on the concept of inferential statistics namely coefficient of correlation which is a measure of relative strength across multiple variables (sales values across items in this case).

Accordingly when we try to find which varieties have highest and lowest correlation among themselves in terms of sales we get to understand that Milk and Grocery are the ones that are with maximum correlation whereas detergents_paper and grocery are the least corelated. This indicates that there is a potential that milk sales could induce grocery sales and vice versa giving an insight for appropriate marketing steps. Accordingly any discounts that are based on bundled sales for grocery and detergents_paper can be deprioritized as such investments does not yield top line as well as impacting bottom line negatively.





Full view of coefficient of co relation across varieties as below:



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

Contingency tables as below for each sub question.

2.1.1. Gender and Major

Major	Undecided	CIS	International Business	Accounting	Other	Management	Economics/Finance	Retailin
Gender								
Female	0	3	4	3	3	4	7	
Male	3	1	2	4	4	6	4	
All	3	4	6	7	7	10	11	

The above contingency table depicts the count of number of students by their gender as well as total students across each of the major. The last row and last column depicts marginal total for the given major and the given gender respectively.

2.1.2. Gender and Grad Intention

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

The above contingency table depicts the count of number of students by their gender as well as total students based on their intention to graduate. The last row and last column depicts marginal total for the given graduation intention and the given gender respectively.

2.1.3. Gender and Employment

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

The above contingency table depicts the count of number of students by their gender as well as total students by various employment status. The last row and last column depicts marginal total for the given employment status and the given gender respectively.

2.1.4. Gender and Computer

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

The above contingency table depicts the count of number of students by their gender as well as total students by category of computers. The last row and last column depicts marginal total for the given employment status and the given Computer type respectively.

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?

For Male:
Going by one of the contingency table above , for example the one for question 2.1.1, we could get the marginal total for male whose proportion with marginal total for “All” students would give us the probability of a randomly selected CSMU student being a male.

Probability that a randomly selected student will be a male is: 46.77%

For Female:
Going by one of the contingency table above , for example the one for question 2.1.1, we could get the marginal total for female whose proportion with marginal total for “All” students would give us the probability of a randomly selected CSMU student being a female.

Probability that a randomly selected student will be a female is:53.23%

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.

For Male:
Going by the contingency table built for 2.1.1, we have male count by majors which sums to the marginal total of complete male count. Since we are finding the probability among male for doing a particular major we will have to take the proportion of each major level count for male vs the entire male count in the marginal total. Accordingly the derived probabilities are as below:

Conditional probability of Undecided major for male is: 10.34%
Conditional probability of CIS major for male is: 3.45%
Conditional probability of International Business major for male is: 6.90%
Conditional probability of Accounting major for male is: 13.79%
Conditional probability of Other major for male is: 13.79%
Conditional probability of Management major for male is: 20.69%
Conditional probability of Economics/Finance major for male is: 13.79%
Conditional probability of Retailing/Marketing major for male is: 17.24%

For Female:
Going by the contingency table built for 2.1.1, we have female count by majors which sums to the marginal total of complete female count. Since we are finding the probability of a female doing a particular major in this case we will have to take the proportion of major level count for

female vs the entire female count in the marginal total. Accordingly the derived probabilities are as below:

Conditional probability of Undecided major for female is: 0.00%
Conditional probability of CIS major for female is: 9.09%
Conditional probability of International Business major for female is: 12.12%
Conditional probability of Accounting major for female is: 9.09%
Conditional probability of Other major for female is: 9.09%
Conditional probability of Management major for female is: 12.12%
Conditional probability of Economics/Finance major for female is: 21.21%
Conditional probability of Retailing/Marketing major for female is: 27.27%

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.

For male:

Going by the contingency table built for 2.1.2, we have male count by graduation intent which sums to the marginal total of complete male count. Since we are finding the probability across each of the graduation intents among male we will have to take the proportion of graduation intent count for male vs the entire male count in the marginal total. Accordingly the derived probabilities are as below:
Conditional probability of graduate intention as No for a male is: 10.34%
Conditional probability of graduate intention as Undecided for a male is: 31.03%
Conditional probability of graduate intention as Yes for a male is: 58.62%

For female:

Going by the contingency table built for 2.1.2, we have female count by graduation intent which sums to the marginal total of complete female count. Since we are finding the probability across each of the graduation intents among female we will have to take the proportion of graduation intent count for female vs the entire female count in the marginal total. Accordingly the derived probabilities are as below:

Conditional probability of graduate intention as No for a female is: 27.27%
Conditional probability of graduate intention as Undecided for a female is: 39.39%
Conditional probability of graduate intention as Yes for a female is: 33.33%

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

For male:

Going by the contingency table built for 2.1.3, we have male count by employment status which sums to the marginal total of complete male count. Since we are finding the probability for a given employment status among the male we will have to take the proportion of employment status count for male vs the entire male count in the marginal total. Accordingly the derived probabilities are as below:

Conditional probability of Unemployed status for male is: 10.34%
Conditional probability of Full-Time status for male is: 24.14%
Conditional probability of Part-Time status for male is: 65.52%

For female:
Going by the contingency table built for 2.1.3, we have female count by employment status which sums to the marginal total of complete female count. Since we are finding the probability for a given employment status among the female we will have to take the proportion of employment status count for female vs the entire female count in the marginal total. Accordingly the derived probabilities are as below:

Conditional probability of Unemployed status for female is: 18.18%
Conditional probability of Full-Time status for female is: 9.09%
Conditional probability of Part-Time status for female is: 72.73%

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

For male:
Going by the contingency table built for 2.1.4, we have male count by laptop preference which sums to the marginal total of complete male count. Since we are finding the probability for a given laptop preference among the male we will have to take the proportion of laptop preference based count for male vs the entire male count in the marginal total. Accordingly the derived probabilities are as below:

Conditional probability of Tablet preference for male is: 0.00%
Conditional probability of Desktop preference for male is: 10.34%
Conditional probability of Laptop preference for male is: 89.66%

For female:
Going by the contingency table built for 2.1.4, we have female count by laptop preference which sums to the marginal total of complete female count. Since we are finding the probability for a given laptop preference among the female we will have to take the proportion of laptop preference based count for female vs the entire female count in the marginal total. Accordingly the derived probabilities are as below:

```
Conditional probability of Tablet preference for female is:     6.06%
Conditional probability of Desktop preference for female is:     6.06%
Conditional probability of Laptop preference for female is:     87.88%
```

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.

In order to find out whether the probabilities at the column level are dependent or independent of gender, we shall first find the probability for the overall students across each columns to then compare them with if there is a impact on the same due to gender which were listed in the previous set of questions.
In order to find the same, we will have to get the proportion of the students (male+female) for each of the values across employment status, graduation intent, Student major and computer preferences against the total students and then compare them to gender level probabilities found in the previous exercise to see if they are same or different.

```
Probabilities of employment status among status
------------------------------------------------------------------
Probability of Unemployed status among students is:         14.52%
Probability of Full-Time status among students is:        16.13%
Probability of Part-Time status among students is:         69.35%


Probabilities on graduation intent across students
---------------------------------------------------------------------
Probability of No for graduation intent among students is:      19.35%
Probability of Undecided for graduation intent among students is:   35.48%
Probability of Yes for graduation intent among students is:      45.16%


Probabilities for each of the major across students
-------------------------------------------------------------
Probability of Undecided intent among students is:          4.84%
Probability of CIS intent among students is:            6.45%
Probability of International Business intent among students is:    9.68%
Probability of Accounting intent among students is:          11.29%
Probability of Other intent among students is:          11.29%
Probability of Management intent among students is:          16.13%
Probability of Economics/Finance intent among students is:      17.74%
Probability of Retailing/Marketing intent among students is:     22.58%
```

```
Probabilities on computer preferences across students
-----------------------------------------------------------------------
Probability of Tablet preference among students is:              3.23%
Probability of Desktop preference among students is:              8.06%
Probability of Laptop preference among students is:              88.71%
```

Answer and justification as below:

Based on the marginal probabilities at the student level, the observations on conditional probabilities across the columns for male and female are not independent of the Gender as they are different from marginal probabilities are the student level

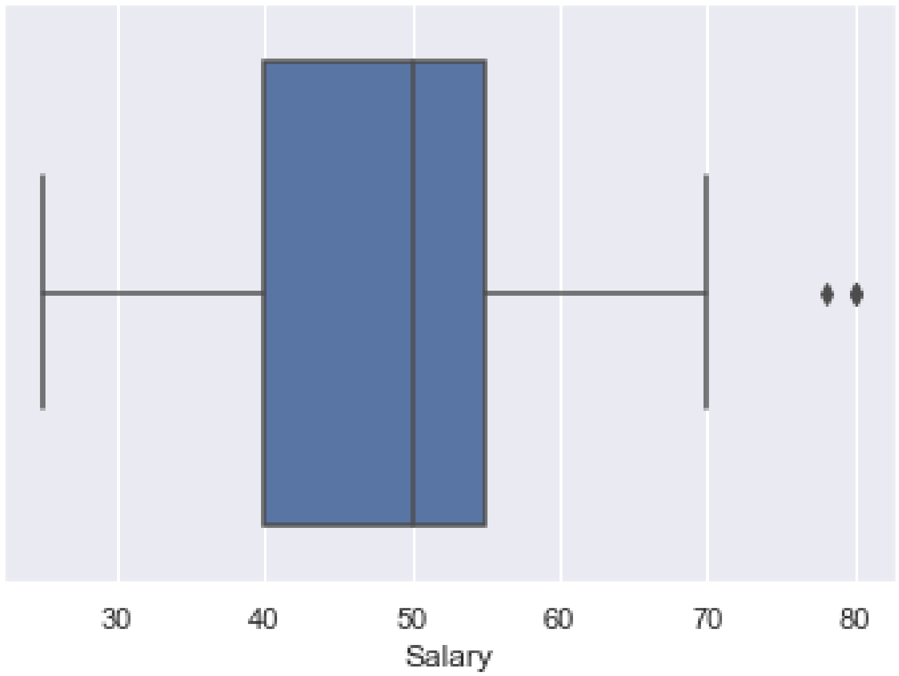
2.4. Note that there are three numerical (continuous) variables in the data set, Salary, Spending and Text Messages. For each of them comment whether they follow a normal distribution.

Write a note summarizing your conclusions.

[Recall that symmetric histogram does not necessarily mean that the underlying distribution is symmetric]

A. Is Salary normally distributed?

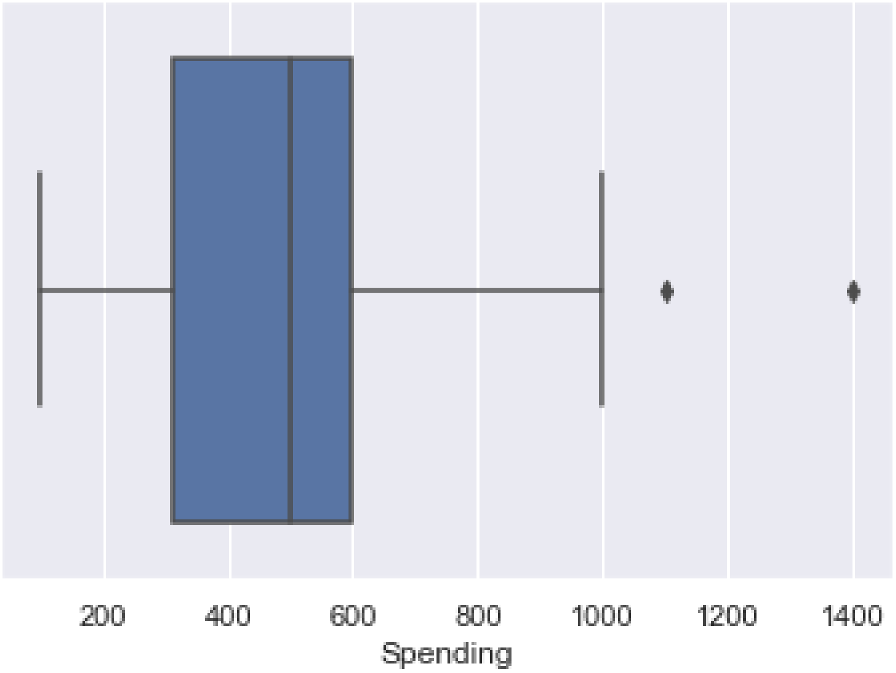
```
-------------------------------------------------
mean of Salary is 48.55
mode of Salary is 40.00
median of Salary is 50.00
Q1: 40.00, Mid:50.00, Q3:55.00, Skew_ind:-5.00
```



Answer and justification: Based on the above findings for Salary wherein the mean, mode and t he median are different along with box plot suggesting that Salary is not normally distributed a nd is left skewed. Also please note that the mean is lesser than median depicting comparatively larger coverage of distribution to the left of median despite outliers beyond the whisker for the maximum indicating the distribution is left skewed.

B. Is Spending normally distributed?

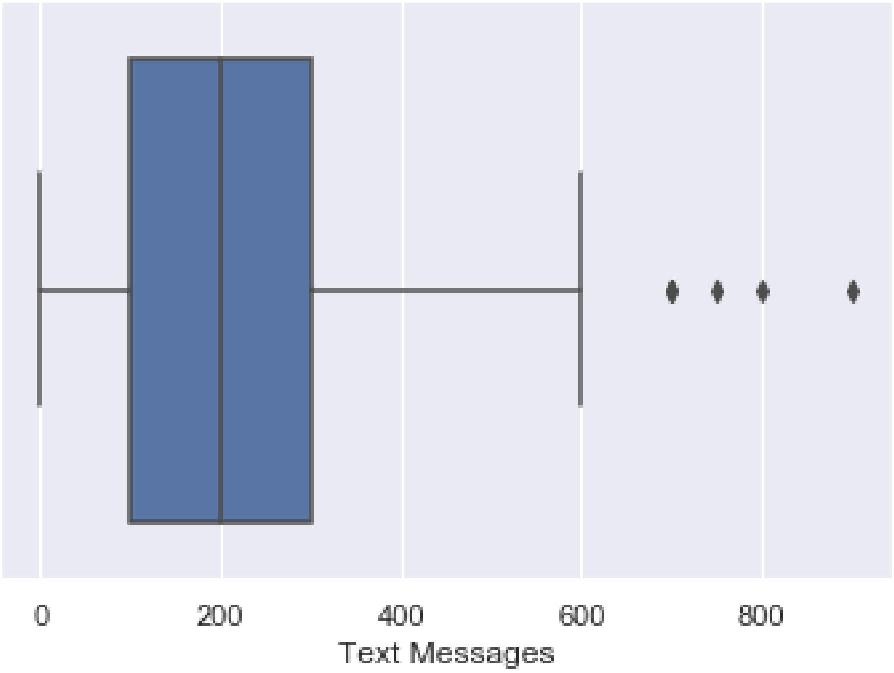
```
----------------------------------------
mean of Spending is 482.02
mode of Spending is 500.00
median of Spending is 500.00
Q1: 312.50, Mid:500.00, Q3:600.00, Skew_ind:-87.50
```



Answer and justification: Based on the above findings for Spending wherein the mean, mode an d the median are different along with the box plot suggesting that is not normally distributed a nd is left skewed. The fact the mean is to the left of median has to be noted as good number of lower values affect the mean despite the outlier beyond the max and hence that is the justificat ion for the left skew.

C. Is Text Messages normally distributed?

Mean of Text Messages is 246.21
Mode of Text Messages is 300.00
Median of Text Messages is 200.00
Q1: 100.00, Mid:200.00, Q3:300.00, Skew_ind:0.00



Answer and justification: Mean, mode and the median for the Text Messages are not equal to each other .

Based on the above, Text Messages is not normally distributed since the mean is not equal to the median and is to the right of the median with some outliers beyond the whisker for the maximum value. The relevant box plot is as shown depicting distribution skewed on the right side.

3.1 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?

Hypothesis to be considered as below:

- ✓ H0 is Sample mean for A Shingles <= Sample mean of B Shingles
- ✓ H1 is Sample mean for A Shingles > Sample mean of B Shingles

Assumptions being made before the test for equality of means is performed as below:

- o We have two samples but do not have population standard deviation. Sample standard deviation can be derived from the given samples.
- o Hence this needs to go through the 2 sample T test
- o Alpha value to be assumed as 0.05 since it is not explicitly stated
- o Since the manufacturer is claiming that the mean moisture weight per 100 square feet cannot be greater than 0.35 we are to prove the reverse of it through a null hypothesis to ascertain if the statistics alludes to the fact that the mean weight of the dried shingles would be greater than or equal to 0.35 per 100 sq. feet.

- Both samples are different in sizes and hence the paired 2 sample T test is ruled out.
- Since this is a comparative study on two samples before and after drying the Shingles we shall proceed with paired t test.
- Since the sample size is different across the samples (B Shingles) with the second sample containing lesser sample size than first one (A Shingles) we shall avoid NaN's on the second sample to enable the test.

Observations:

- Std deviation for A Shingles is: 0.135731 while mean is 0.316667
- Std deviation for B Shingles is: 0.137296 while mean is 0.273548
- One sample t test on A Shingles alone the outcome of which fails to reject the null hypothesis given in the question
    - t statistic: -1.4735046253382782 p value: 0.14955266289815025
- One sample t test on B Shingles alone the outcome of which rejects the null hypothesis given in the question.
    - t statistic: -3.1003313069986995 p value: 0.004180954800638363

Result:

✓ t_statistic is 0.844501

✓ p_value is 0.41

✓ We do not have enough evidence to reject the null hypothesis in favour of alternative hypothesis

✓ We conclude that the mean moisture content across A and B shingles are unequal favoring the company’s claim.

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

Assumptions required on the population distribution

✓ Population across samples for Shingles A and B are normally distributed

✓ Null hypothesis is a "fail to reject" based on samples for Shingles A that it leads to the population mean weight as <= 0.35

✓ Null hypothesis is a "reject" based on samples for on Shingles B that it leads to the population mean weight as <= 0.35