SMDM PROJECT

DSBA – 2021 BATCH

SAUMYA RAMANAN

APRIL 2021

TABLE OF CONTENTS

PROBLEM 1

SECTION 1.01 PROBLEM SUMMARY:

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). The requirement is to analyse the data and provide recommendations to solve problems and improve business.

SECTION 1.02 DATA QUALITY

The data for amount spent on 5 products ie., Milk, Grocery, Frozen, Detergents_Paper, Delicatessen for 2 channels and 3 Regions by wholesale retailer annually is provided.

Data was verified for data columns and observed to have no null data.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 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)
memory usage: 31.1+ KB
```

Output false for df.isnull().values.any()

SECTION 1.03 SOLUTIONS

(A) 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?

(I) 1.1.1 USE METHODS OF DESCRIPTIVE STATISTICS TO SUMMARIZE DATA.

The methology used is to Total the retailer expense, grouped by Channel and Region

(II) WHICH REGION AND WHICH CHANNEL SPENT THE MOST?

As evidenced in data(below) : **Hotel** spends the most in Channels and **Other** in region spends most.

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

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

(III) WHICH REGION AND WHICH CHANNEL SPENT THE LEAST?

It is found **Retail** spends the least in Channel, and **Oporto** spends least in Region**.**

(B) 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.

	Fresh	Milk	Grocery	Frozen	Detergent_Paper	Delicatessen
Channel	Higher ***mean, med, covariance and IQR*** in **Hotel** compared to Retail.	Mean & IQR higher in **Hotel** however cov is lower in Retail	Mean & iqr in Hotel almost 4 times that of Hotel. Lower cov in retail.	Double mean and higher iqr in Hotel but lesser cov in Retail	Almost comparable means in both channels but lower cov in Retail.	Higher mean and lower cov in Retail
Region	Highest mean on Other but highest cov in **Oporto**	Highest Iqr,mean & cov in Others.	Slightly higher mean in Oporto, but cov is min in Lisbon.	Highest mean in Oporto, lowest cov in Lisbon	Almost comparable means, with lowest cov in Lisbon	Lowest cov in Oporto, highest in Others.Almost comparable means.

Mean-STD-IQR for Products: Channel wise:

Fresh	Milk	Grocery	Frozen	Detergent_Paper	Delicatessen

FRESH:

count	mean	std	min	25%	50%	75%	max	
Channel								
Hotel	298	13475.5604	13831.6875	3	4070.25	9581.5	18274.75	112151
Retail	142	8904.32394	8987.71475	18	2347.75	5993.5	12229.75	44466

MILK:

	count	mean	std	min	25%	50%	75%	max
Channel								
Hotel	298	3451.72483	4352.16557	55	1164.5	2157	4029.5	43950
Retail	142	10716.5	9679.63135	928	5938	7812	12162.75	73498

GROCERY:

	count	mean	std	min	25%	50%	75%	max
Channel								
Hotel	298	3962.13758	3545.51339	3	1703.75	2684	5076.75	21042
Retail	142	16322.8521	12267.3181	2743	9245.25	12390	20183.5	92780

FROZEN:

	count	mean	std	min	25%	50%	75%	max
Channel								
Hotel	298	3748.25168	5643.9125	25	830	2057.5	4558.75	60869
Retail	142	1652.61268	1812.80366	33	534.25	1081	2146.75	11559

DETERGENTS PAPER:

	count	mean	std	min	25%	50%	75%	max
Channel								
Hotel	298	790.560403	1104.09367	3	183.25	385.5	899.5	6907
Retail	142	7269.50704	6291.0897	332	3683.5	5614.5	8662.5	40827

DELICATESSEN

	count	mean	std	min	25%	50%	75%	max
Channel								
Hotel	298	1415.95638	3147.42692	3	379	821	1548	47943
Retail	142	1753.43662	1953.79705	3	566.75	1350	2156	16523

Mean-STD-IQR for Products: Region-wise:

Fresh	Milk	Grocery	Frozen	Detergent_Paper	Delicatessen

FRESH:

count	mean	std	min	25%	50%	75%	max	
Region								
Lisbon	77	11101.7273	11557.4386	18	2806	7363	15218	56083
Oporto	47	9887.68085	8387.89921	3	2751.5	8090	14925.5	32717
Other	316	12533.4715	13389.2131	3	3350.75	8752.5	17406.5	112151

MILK:

	count	mean	std	min	25%	50%	75%	max
Region								
Lisbon	77	5486.41558	5704.85608	258	1372	3748	7503	28326
Oporto	47	5088.17021	5826.34315	333	1430.5	2374	5772.5	25071
Other	316	5977.08544	7935.46344	55	1634	3684.5	7198.75	73498

GROCERY:

	count	mean	std	min	25%	50%	75%	max
Region								
Lisbon	77	7403.07792	8496.28773	489	2046	3838	9490	39694
Oporto	47	9218.59575	10842.7453	1330	2792.5	6114	11758.5	67298
Other	316	7896.36392	9537.28778	3	2141.5	4732	10559.75	92780

FROZEN:

	count	mean	std	min	25%	50%	75%	max
Region								
Lisbon	77	3000.33766	3092.14389	61	950	1801	4324	18711
Oporto	47	4045.3617	9151.78495	131	811.5	1455	3272	60869
Other	316	2944.59494	4260.12624	25	664.75	1498	3354.75	36534

DETERGENTS_PAPER:

	count	mean	std	min	25%	50%	75%	max
Region								
Lisbon	77	2651.11688	4208.46271	5	284	737	3593	19410
Oporto	47	3687.46809	6514.71767	15	282.5	811	4324.5	38102
Other	316	2817.75317	4593.05161	3	251.25	856	3875.75	40827

DELICATESENNE:

	count	mean	std	min	25%	50%	75%	max
Region								
Lisbon	77	1354.8961	1345.42334	7	548	806	1775	6854
Oporto	47	1159.70213	1050.73984	51	540.5	898	1538.5	5609
Other	316	1620.60127	3232.58166	3	402	994	1832.75	47943

<u>Inferences based on observations:</u>

- Significantly higher mean in Fresh, Milk , Frozen and Grocery in Hotel
- Consistently covariance is better for Retail showing more reliability in data
- For Detergent_papers across Channels or across Regions there is no significant difference in mean expenditure.
- Higher expenditure in Delicatesenne in Retail
- Others – region has a higher mean among Regions
- Most of the times, Lisbon has lowest covariance

(REFER MULTIPLE TABLES IN CODE SUBMISSION)

(C) 1.3 ON THE BASIS OF THE DESCRIPTIVE MEASURE OF VARIABILITY, WHICH ITEM SHOWS THE MOST INCONSISTENT BEHAVIOUR? WHICH ITEMS SHOWS THE LEAST INCONSISTENT BEHAVIOUR?

- Most of the times, **Lisbon** has lowest covariance showing more reliable data(Others has highest covariance)
- Consistently covariance is better for **Retail** showing more reliability in data
- Also, Lisbon has the minimum expenditure value (2386813, and Other is max with 10677599). Expenditure data for Lisbon is just less than quarter of Others.
- Similarly in Channels, Retail has lower expenditure value (6619931 when compared to Hotel which is 7999569).
- However there is not much significant difference in expenditure between Hotel & Retail when compared to regions, say Lisbon and Others. That being said, overall the data in **Regions** seems more reliable than the data of Regions, as the covariances difference when compared to actual expenditure difference is not very high.

<u>TOTAL EXPENDITURE:</u>

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

Fresh Milk Grocery Frozen Detergents_Paper Delicatessen Row Total
Region
Lisbon 854833 422454 570037 231026 204136 104327 2386813

Oporto	464721	239144	433274	190132	173311	54506	1555088
Other	3960577	1888759	2495251	930492	890410	512110	10677599

<u>COEFFICIENT OF VARIANCE:</u>

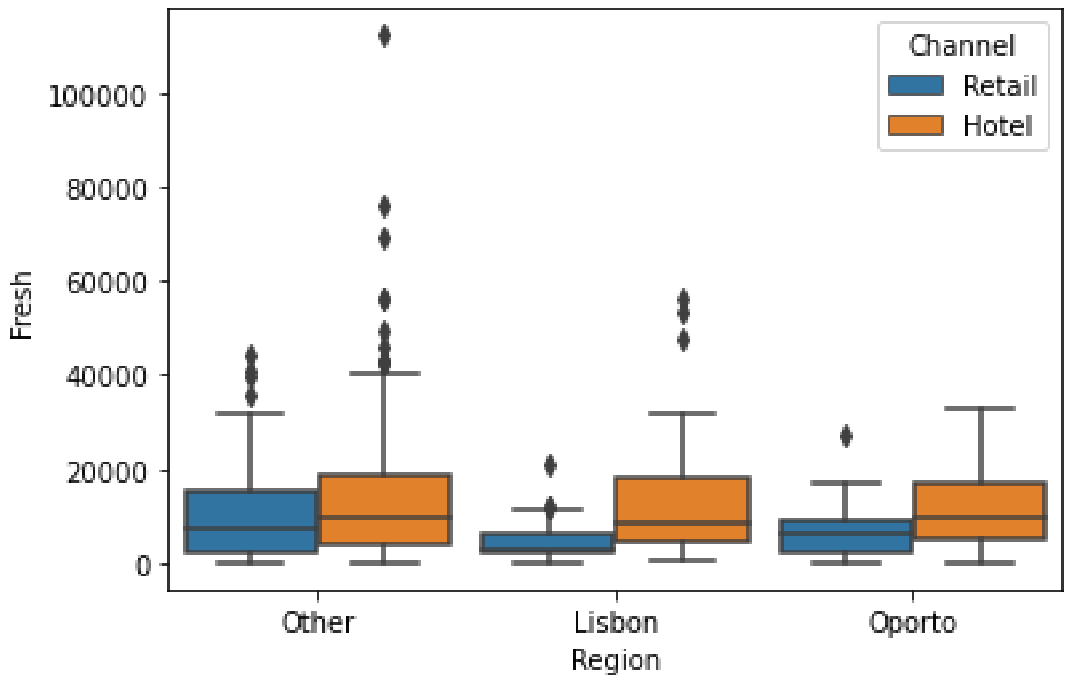
	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Row Total
Channel							
Hotel	102.642763	126.086689	89.484863	150.574534	139.659622	222.282761	82.568474
Retail	100.936520	90.324559	75.154256	109.693196	86.540802	111.426728	62.950128

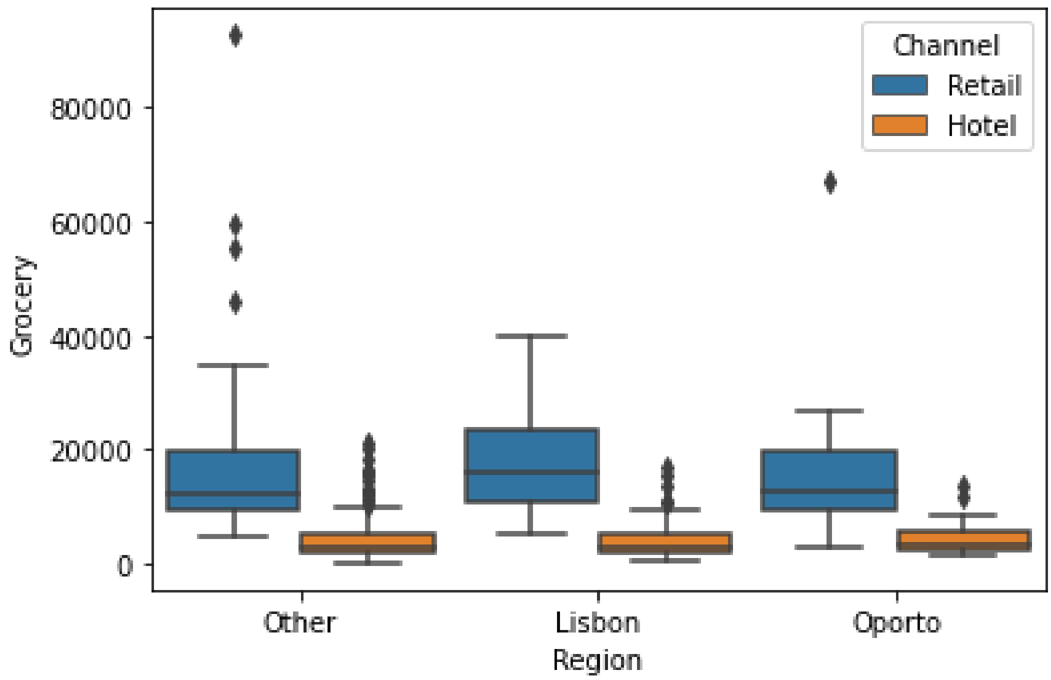
	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Row Total
Region							
Lisbon	104.104868	103.981479	114.766963	103.059863	158.743009	99.300849	65.559374
Oporto	84.831816	114.507630	117.618188	226.229090	176.671839	90.604287	73.244848
Other	106.827650	132.764765	120.780753	144.676138	163.004044	199.468045	82.715138

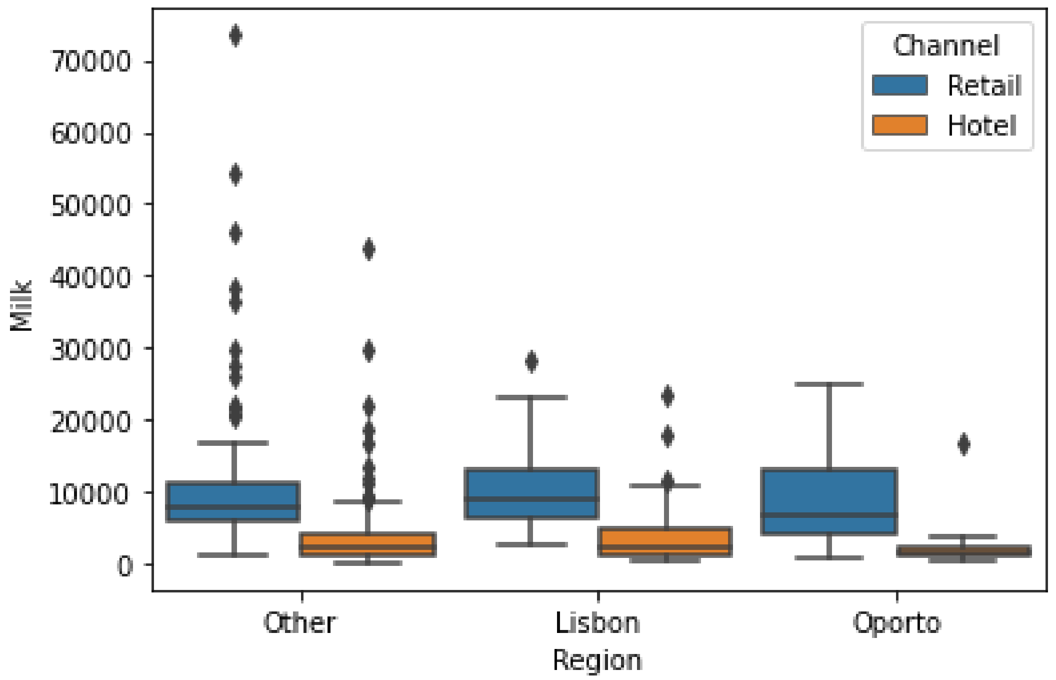
(D) 1.4 ARE THERE ANY OUTLIERS IN THE DATA? BACK UP YOUR ANSWER WITH A SUITABLE PLOT/TECHNIQUE WITH THE HELP OF DETAILED COMMENTS.

Observations:

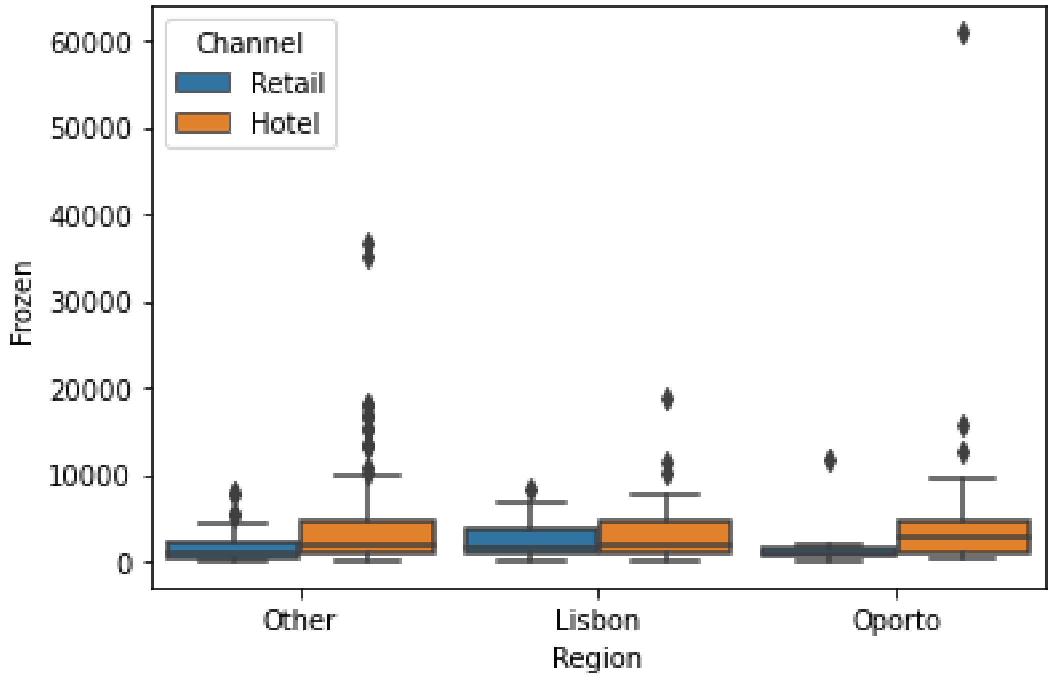
- Based on various Box plots, overall there are significant outliers in "Others" especially in "Hotel" Channel.
- Overall the least number of outliers are in"Oporto" region in the "Retail" channel.
- Product wise – Milk, Fresh , Frozen overall have more outliers when compared to say, detergents_Paper (which still has outliers in Hotel)

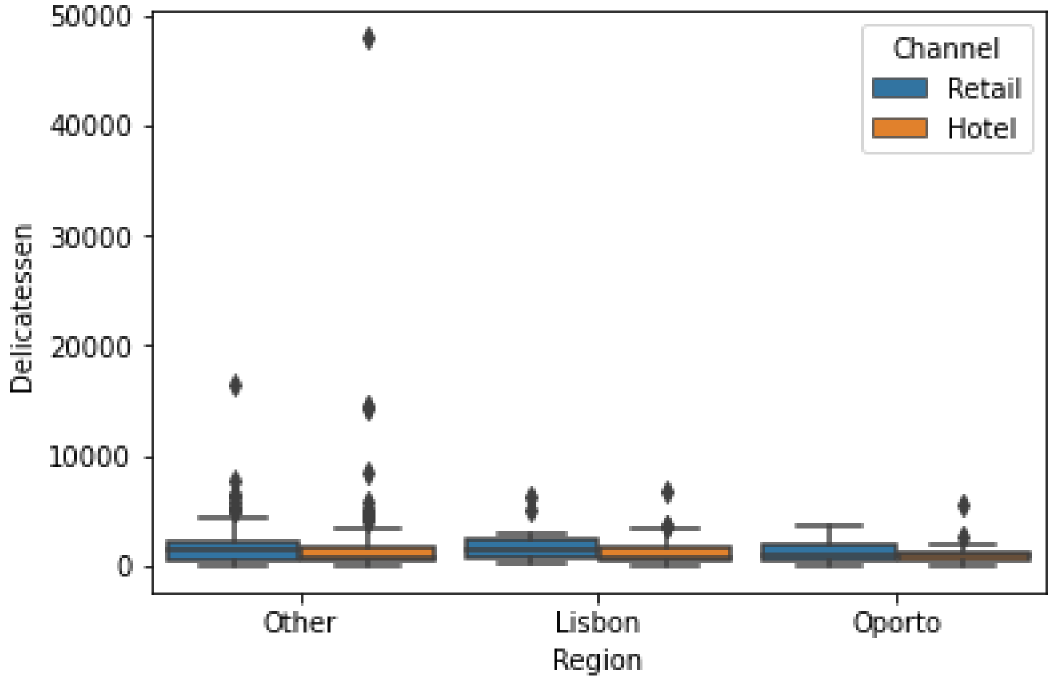

Channel
Retail
Hotel
Fresh
100000
80000
60000
40000
20000
0
Other
Lisbon
Oporto
Region

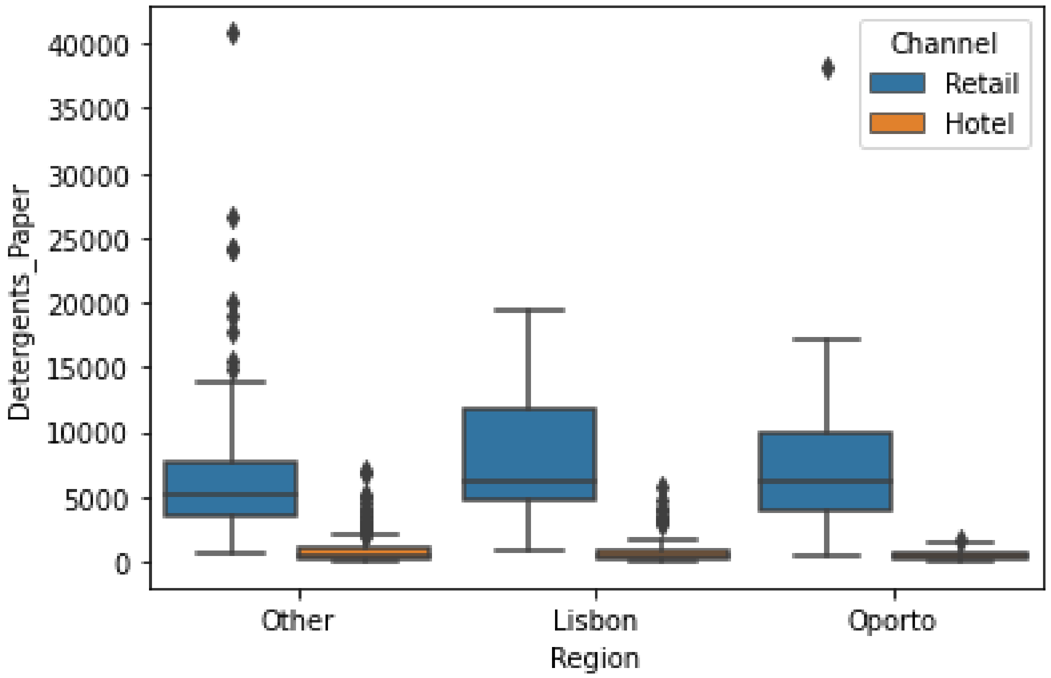

Channel
Retail
Hotel
Grocery
80000
60000
40000
20000
0
Other
Lisbon
Oporto
Region


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

In [126]:


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





In [243]:

(E) 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

Deductions:

- In addition to all the analysis done in answers above, below is the correlation between products Channel/Regions wise.
- Also, the skewness in data – Channel and Region wise.
- Strong correlation between – Milk/Grocery, Milk/Delicatessen and Detergents_Paper/Grocery
- Low or negative correlation between Frozen/Grocery and Detergent/Frozen

Observations:

- Based on coefficient of variation calculation : Retail- Grocery has min coeff of variation and Hotel : Delicatessen has max coeff of variation
- Based on covariance
  - Strong positive: Grocery& Delicatessen[Oporto], Fresh&Frozen[Lisbon], Detergents & Grocery[Lisbon & Oporto], Delicatessen & Detergents[Others]
  - Strong negative: Milk and Fresh [Lisbon], Detergents & Frozen [Oporto]
- Based on correlation:
  - Positive : Grocery & Milk | Grocery & Detergents | Milk & Detergents
  - Negative : Detergents & Frozen, Detergents & Fresh

Recommendations to Business:

- Increase Frozen expenses especially in Retail – will bring down the Grocery expense.
- Similarly increase Frozen expenses especially in Others – will bring down the Grocery expense.
-

COVARIANCES:

		Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Row Total
Channel								
Hotel	Fresh	1.913156e+08	1.484039e+07	1.155690e+07	2.621146e+07	-1.138710e+05	1.108078e+07	2.548912e+08
	Milk	1.484039e+07	1.894135e+07	9.298710e+06	1.004335e+07	1.200992e+06	8.624927e+06	6.294972e+07
	Grocery	1.155690e+07	9.298710e+06	1.257067e+07	5.176321e+06	2.145355e+06	4.986764e+06	4.573472e+07
	Frozen	2.621146e+07	1.004335e+07	5.176321e+06	3.185375e+07	-1.944282e+05	7.608778e+06	8.069924e+07

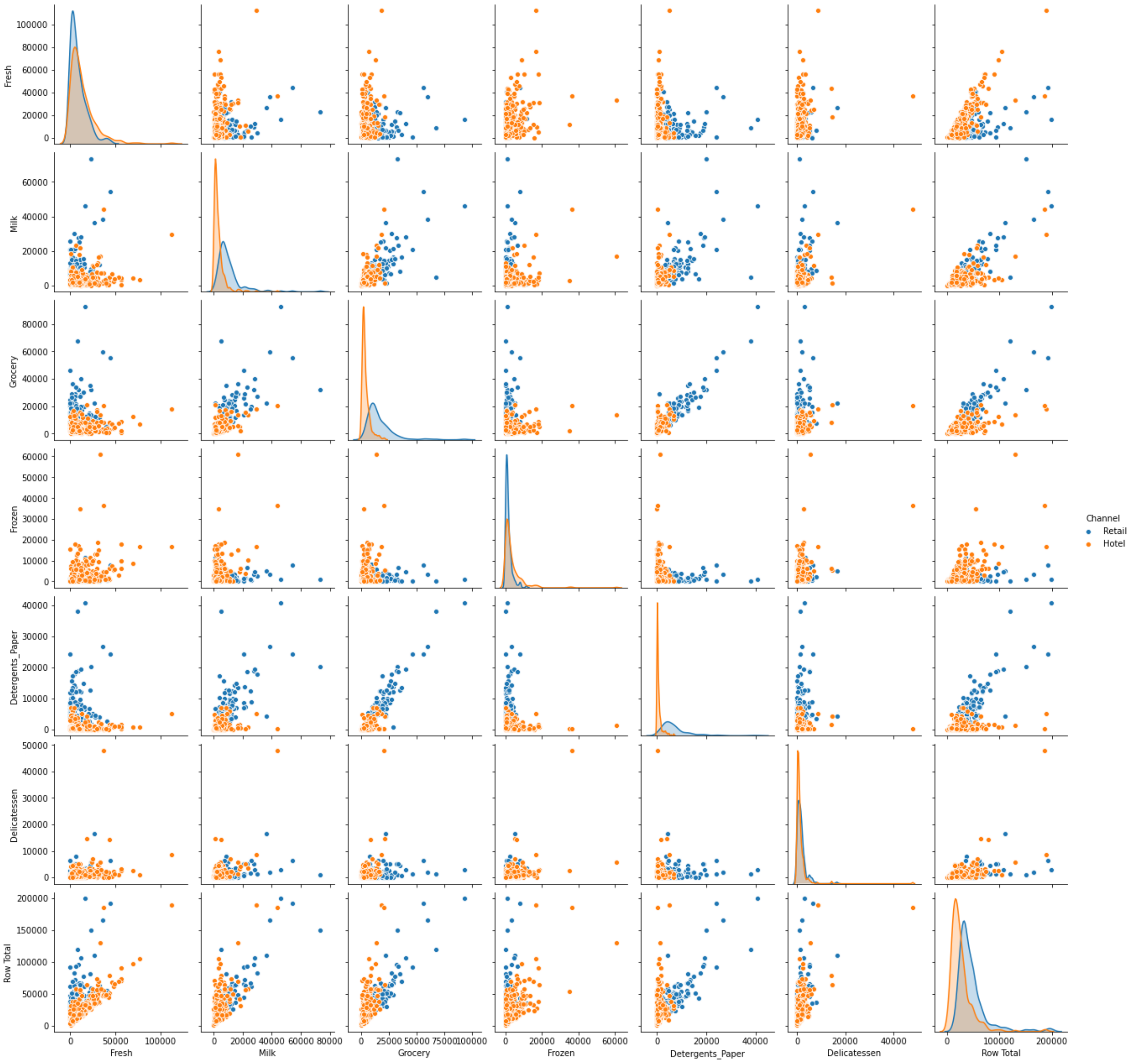
Detergents_Paper -1.138710e+05 1.200992e+06 2.145355e+06 -1.944282e+05 1.219023e+06 2.702777e+05 4.527349e+06
Delicatessen 1.108078e+07 8.624927e+06 4.986764e+06 7.608778e+06 2.702777e+05 9.906296e+06 4.247782e+07
Row Total 2.548912e+08 6.294972e+07 4.573472e+07 8.069924e+07 4.527349e+06 4.247782e+07 4.912801e+08
Retail Fresh 8.077902e+07 2.060149e+07 9.760788e+06 4.374357e+06 1.299662e+06 4.884133e+06 1.216994e+08
Milk 2.060149e+07 9.369526e+07 7.821065e+07 3.060973e+06 3.787391e+07 6.493880e+06 2.399362e+08
Grocery 9.760788e+06 7.821065e+07 1.504871e+08 9.916477e+05 7.129703e+07 3.797502e+06 3.145447e+08
Frozen 4.374357e+06 3.060973e+06 9.916477e+05 3.286257e+06 1.920210e+05 1.119771e+06 1.302503e+07
Detergents_Paper 1.299662e+06 3.787391e+07 7.129703e+07 1.920210e+05 3.957781e+07 8.400858e+05 1.510805e+08
Delicatessen 4.884133e+06 6.493880e+06 3.797502e+06 1.119771e+06 8.400858e+05 3.817323e+06 2.095270e+07
Row Total 1.216994e+08 2.399362e+08 3.145447e+08 1.302503e+07 1.510805e+08 2.095270e+07 8.612386e+08

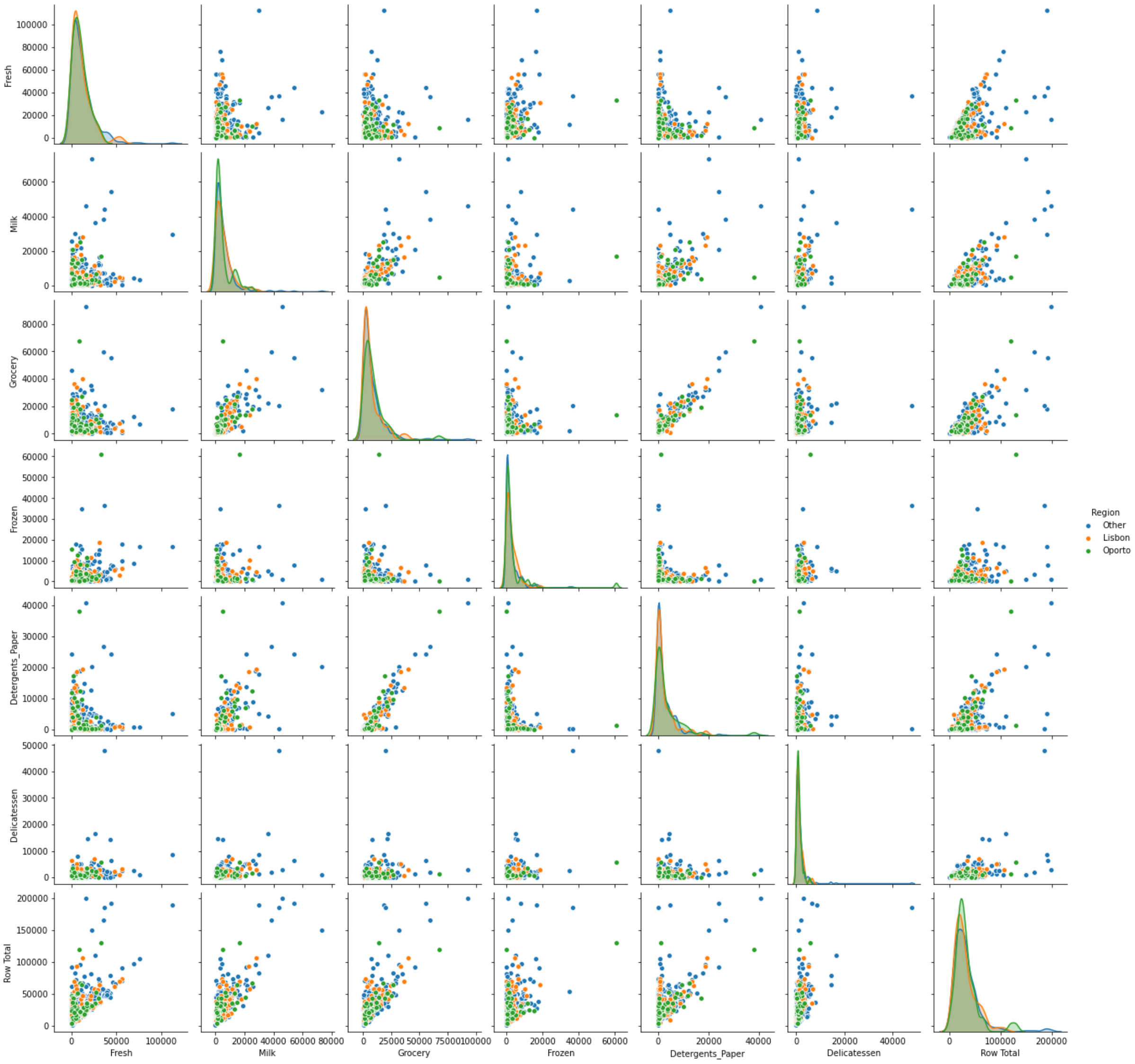
Fresh Milk Grocery Frozen Detergents_Paper Delicatessen Row Total
Region
Lisbon Fresh 1.335744e+08 -9.848517e+06 -1.868297e+07 1.035036e+07 -1.308769e+07 3.227098e+06 1.055327e+08
Milk -9.848517e+06 3.254538e+07 4.009660e+07 2.372360e+06 1.780135e+07 3.604031e+06 8.657120e+07
Grocery -1.868297e+07 4.009660e+07 7.218691e+07 -6.905125e+05 3.190504e+07 3.952371e+06 1.287674e+08
Frozen 1.035036e+07 2.372360e+06 -6.905125e+05 9.561354e+06 -9.147403e+05 1.409678e+06 2.208850e+07
Detergents_Paper -1.308769e+07 1.780135e+07 3.190504e+07 -9.147403e+05 1.771116e+07 1.298936e+06 5.471405e+07

Delicatessen 3.227098e+06 3.604031e+06 3.952371e+06 1.409678e+06 1.298936e+06 1.810164e+06 1.530228e+07
Row Total 1.055327e+08 8.657120e+07 1.287674e+08 2.208850e+07 5.471405e+07 1.530228e+07 4.129761e+08
Oporto Fresh 7.035685e+07 -2.641270e+06 -1.046219e+07 2.930388e+07 -1.126419e+07 3.486513e+06 7.877960e+07
Milk -2.641270e+06 3.394627e+07 2.611661e+07 8.785218e+06 1.386089e+07 1.988416e+06 8.205613e+07
Grocery -1.046219e+07 2.611661e+07 1.175651e+08 -4.183396e+06 6.695991e+07 9.989221e+05 1.969950e+08
Frozen 2.930388e+07 8.785218e+06 -4.183396e+06 8.375517e+07 -9.704490e+06 6.063431e+06 1.140198e+08
Detergents_Paper -1.126419e+07 1.386089e+07 6.695991e+07 -9.704490e+06 4.244155e+07 -2.370966e+05 1.020566e+08
Delicatessen 3.486513e+06 1.988416e+06 9.989221e+05 6.063431e+06 -2.370966e+05 1.104054e+06 1.340424e+07
Row Total 7.877960e+07 8.205613e+07 1.969950e+08 1.140198e+08 1.020566e+08 1.340424e+07 5.873113e+08
Other Fresh 1.792710e+08 1.544882e+07 4.358336e+06 2.317918e+07 -3.527783e+06 1.067143e+07 2.294010e+08
Milk 1.544882e+07 6.297158e+07 5.780650e+07 4.456671e+06 2.621603e+07 1.055866e+07 1.774583e+08
Grocery 4.358336e+06 5.780650e+07 9.095986e+07 -2.007389e+06 4.072448e+07 6.627325e+06 1.984691e+08
Frozen 2.317918e+07 4.456671e+06 -2.007389e+06 1.814868e+07 -2.734098e+06 6.296021e+06 4.733906e+07
Detergents_Paper -3.527783e+06 2.621603e+07 4.072448e+07 -2.734098e+06 2.109612e+07 1.060130e+06 8.283488e+07
Delicatessen 1.067143e+07 1.055866e+07 6.627325e+06 6.296021e+06 1.060130e+06 1.044958e+07 4.566315e+07
Row Total 2.294010e+08 1.774583e+08 1.984691e+08 4.733906e+07 8.283488e+07 4.566315e+07 7.811655e+08

CORRELATION TABLE:

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Row Total
Fresh	1.000000	0.100510	-0.011854	0.345881	-0.101953	0.244690	0.575178
Milk	0.100510	1.000000	0.728335	0.123994	0.661816	0.406368	0.776909
Grocery	-0.011854	0.728335	1.000000	-0.040193	0.924641	0.205497	0.740680
Frozen	0.345881	0.123994	-0.040193	1.000000	-0.131525	0.390947	0.388436
Detergents_Paper	-0.101953	0.661816	0.924641	-0.131525	1.000000	0.069291	0.633882
Delicatessen	0.244690	0.406368	0.205497	0.390947	0.069291	1.000000	0.496849
Row Total	0.575178	0.776909	0.740680	0.388436	0.633882	0.496849	1.000000


Fresh
Milk
Grocery
Frozen
Detergents_Paper
Delicatessen
Row Total
Channel
Retail
Hotel
0
20000
40000
60000
80000
100000
50000
150000
200000
25000
75000
0
50000
100000
0
20000
40000
60000
80000
0
25000
50000
75000
100000
0
20000
40000
60000
0
20000
40000
0
20000
40000
0
100000
200000


Fresh
100000
80000
60000
40000
20000
0
Milk
60000
40000
20000
0
Grocery
80000
60000
40000
20000
0
Frozen
60000
50000
40000
30000
20000
10000
0
Detergents_Paper
40000
30000
20000
10000
0
Delicatessen
50000
40000
30000
20000
10000
0
Row Total
200000
150000
100000
50000
0
0 50000 100000
Fresh
0 20000 40000 60000 80000
Milk
0 25000 50000 75000100000
Grocery
0 20000 40000 60000
Frozen
0 20000 40000
Detergents_Paper
0 20000 40000
Delicatessen
0 100000 200000
Row Total
Region
Other
Lisbon
Oporto

PROBLEM 2

SECTION 2.01 PROBLEM SUMMARY

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).

Section 2.02 DATA QUALITY

No null data.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62 entries, 0 to 61
Data columns (total 14 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   ID                  62 non-null     int64  
 1   Gender              62 non-null     object 
 2   Age                 62 non-null     int64  
 3   Class               62 non-null     object 
 4   Major               62 non-null     object 
 5   Grad Intention      62 non-null     object 
 6   GPA                 62 non-null     float64
 7   Employment          62 non-null     object 
 8   Salary              62 non-null     float64
 9   Social Networking   62 non-null     int64  
 10  Satisfaction        62 non-null     int64  
 11  Spending            62 non-null     int64  
 12  Computer            62 non-null     object 
 13  Text Messages       62 non-null     int64  
dtypes: float64(2), int64(6), object(6)
memory usage: 6.9+ KB
```

SECTION 2.03 SOLUTIONS

(A) 2.1. FOR THIS DATA, CONSTRUCT THE FOLLOWING CONTINGENCY TABLES (KEEP GENDER AS ROW VARIABLE)

(I) 2.1.1. GENDER AND MAJOR

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

(II) 2.1.2. GENDER AND GRAD INTENTION

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

(III) 2.1.3. GENDER AND EMPLOYMENT

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

(IV) 2.1.4. GENDER AND COMPUTER

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

(B) 2.2. ASSUME THAT THE SAMPLE IS A REPRESENTATIVE OF THE POPULATION OF CMSU. BASED ON THE DATA, ANSWER THE FOLLOWING QUESTION:

(I) 2.2.1 WHAT IS THE PROBABILITY THAT A RANDOMLY SELECTED CMSU STUDENT WILL BE MALE?

P_male = 29/62

0.468

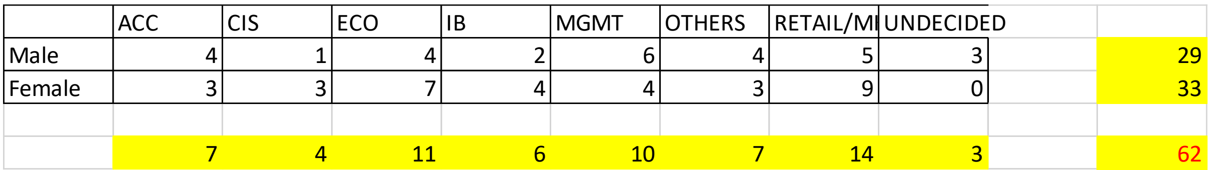
(II) 2.2.2 WHAT IS THE PROBABILITY THAT A RANDOMLY SELECTED CMSU STUDENT WILL BE FEMALE?

P_female = 33/62

0.532

(C) 2.3. ASSUME THAT THE SAMPLE IS A REPRESENTATIVE OF THE POPULATION OF CMSU. BASED ON THE DATA, ANSWER THE FOLLOWING QUESTION:

(I) 2.3.1 FIND THE CONDITIONAL PROBABILITY OF DIFFERENT MAJORS AMONG THE MALE STUDENTS IN CMSU.



	ACC	CIS	ECO	IB	MGMT	OTHERS	RETAIL/MI	UNDECIDED		
Male	4	1	4	2	6	4	5	3		29
Female	3	3	7	4	4	3	9	0		33
	7	4	11	6	10	7	14	3		62

P(ACC|M)[Accounts|Male]	0.137931
P(CIS|M)	0.034483
P(ECO|M)	0.137931
P(IB|M)	0.068966
P(MGM|M)	0.206897
P(OT|M)	0.137931
P(RE|M)	0.172414
P(UN|M)	0.103448

(II) 2.3.2 FIND THE CONDITIONAL PROBABILITY OF DIFFERENT MAJORS AMONG THE FEMALE STUDENTS OF CMSU.

P(ACC|F)	0.090909
P(CIS|F)	0.090909
P(ECO|F)	0.212121
P(IB|F)	0.121212
P(MGM|F)	0.121212
P(OT|F)	0.090909
P(RE|F)	0.272727
P(UN|F)	0

(D) 2.4. ASSUME THAT THE SAMPLE IS A REPRESENTATIVE OF THE POPULATION OF CMSU. BASED ON THE DATA, ANSWER THE FOLLOWING QUESTION:

(I) 2.4.1 FIND THE PROBABILITY THAT A RANDOMLY CHOSEN STUDENT IS A MALE AND INTENDS TO GRADUATE.

	Graduate	Not Graduate	Undecided

Male	17	3	9	29
Female	11	9	13	33
	28	12	22	

P(GR|M) 0.586207

(II) 2.4.2 FIND THE PROBABILITY THAT A RANDOMLY SELECTED STUDENT IS A FEMALE AND DOES NOT HAVE A LAPTOP.

	Laptop	Desktop	Tablet
Male			
Female	29	2	2

P(NOL|F) 0.121212

(E) 2.5. ASSUME THAT THE SAMPLE IS A REPRESENTATIVE OF THE POPULATION OF CMSU. BASED ON THE DATA, ANSWER THE FOLLOWING QUESTION:

(I) 2.5.1 FIND THE PROBABILITY THAT A RANDOMLY CHOSEN STUDENT IS EITHER A MALE OR HAS A FULL-TIME EMPLOYMENT

	FullTime	PartTime	Unemp	
Male	7	19	3	29
Female	3	24	6	33
	10	43	9	62

P(M) OR P(FT) 0.629032

(II) 2.5.2 FIND THE CONDITIONAL PROBABILITY THAT GIVEN A FEMALE STUDENT IS RANDOMLY CHOSEN, SHE IS MAJORING IN INTERNATIONAL BUSINESS OR MANAGEMENT.

P(Ibor MG | F) 0.242424242

(F) 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 GRADUATE INTENTION AND BEING FEMALE ARE INDEPENDENT EVENTS?

	Graduate	Not Graduate

Male	17	3
Female	11	9

28 12

Is Check if P(F (AND) P[G or NG]) = P(F) * P(G or NG).

	Graduate	Not Graduate	Undecided
Male	17	3	9
Female	11	9	13

28 12

P(F)	0.725	
P(G)	0.7	0.91
P(NG)	0.3	
P(G or NG)	1	

Is Check if P(F (AND) P[G or NG]) = P(F) * P(G or NG).

(11+9)/40 0.5

THEY ARE INDEPENDENT

(G) 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

(I) 2.7.1 IF A STUDENT IS CHOSEN RANDOMLY, WHAT IS THE PROBABILITY THAT HIS/HER GPA IS LESS THAN 3?

```
GPA     2.3  2.4  2.5  2.6  2.8  2.9  3.0  3.1  3.2  3.3  3.4  3.5  3.6
  3.7  \
Gender

Female    1    1    2    0    1    3    5    2    4    3    2    4    1
    2
Male      0    0    4    2    2    1    2    5    2    2    5    2    2
    0
All       1    1    6    2    3    4    7    7    6    5    7    6    3
    2

GPA     3.8  3.9  All
Gender
```

```
Female      1     1    33
Male        0     0    29
All         1     1    62

Salary   52.0   54.0   55.0   60.0   65.0   70.0   78.0   80.0   All
Gender
Female      0      0      5      5      0      1      1      1    33
Male        1      1      3      3      1      0      0      1    29
All         1      1      8      8      1      1      1      2    62
```

GPA	2.3	2.4	2.5	2.6	2.8	2.9
Male	0	0	4	2	2	1
Female	1	1	2	0	1	3

The probability that a randomly chose student has GPA<3 is 0.274194

(II) 2.7.2 FIND CONDITIONAL PROBABILITY THAT A RANDOMLY SELECTED MALE EARNS 50 OR MORE. FIND CONDITIONAL PROBABILITY THAT A RANDOMLY SELECTED FEMALE EARNS 50 OR MORE.

```
Salary   25.0   30.0   35.0   37.0   37.5   40.0   42.0   45.0   47.0   47.5   50.
0   \
Gender

Female      0      5      1      0      1      5      1      1      0      1
5
Male        1      0      1      1      0      7      0      4      1      0
4
All         1      5      2      1      1     12      1      5      1      1
9
```

p_female_above_fifty = (18)/33 =**0.5454**

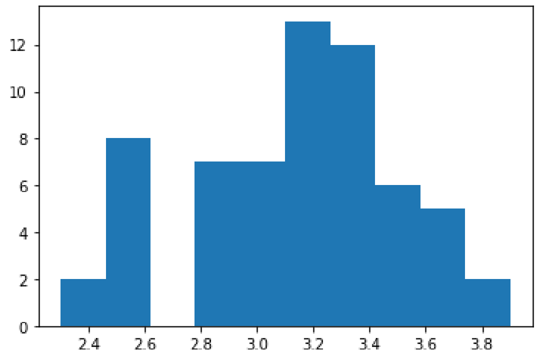
p_male_above_fifty = (14)/29 =**0.4828**

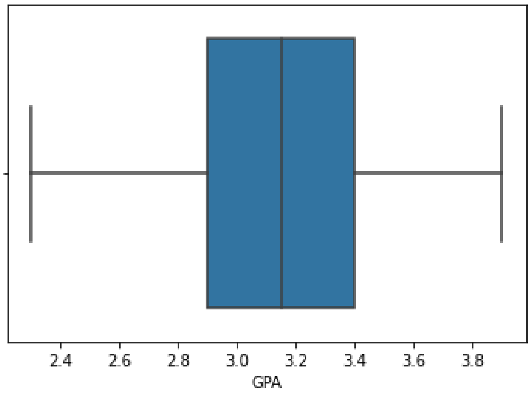
(III) 2.8.1 NOTE THAT THERE ARE FOUR NUMERICAL (CONTINUOUS) VARIABLES IN THE DATA SET, GPA, SALARY, SPENDING AND TEXT MESSAGES. FOR EACH OF THEM COMMENT WHETHER THEY FOLLOW A NORMAL DISTRIBUTION.

GPA –NORMAL

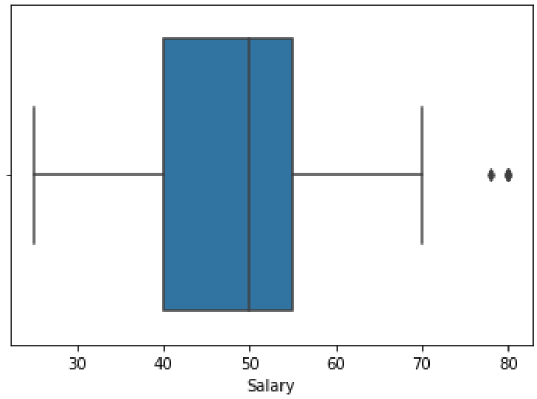
SALARY – NORMAL

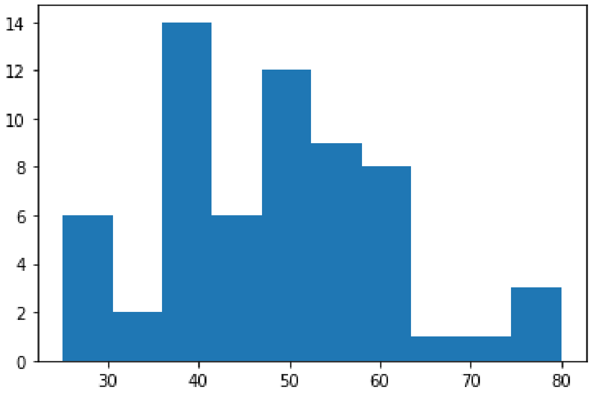
GPA:



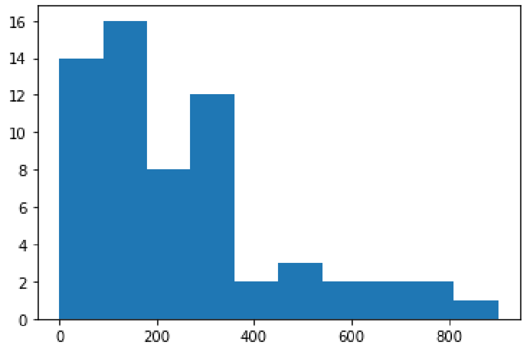


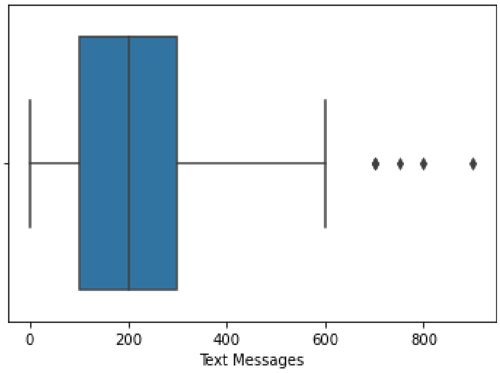
SALARY:

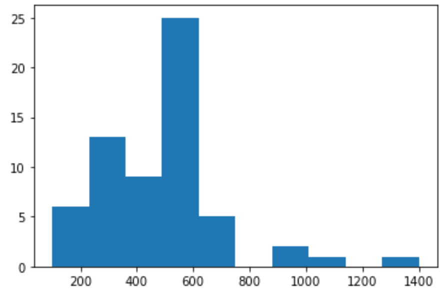


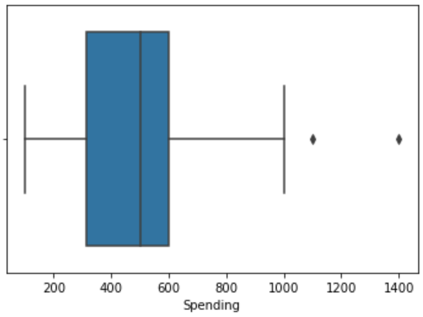


TEXT & SPEND – NOT NORMAL









(Refer Hist & Box plot)

(IV) 2.8.2 WRITE A NOTE SUMMARIZING YOUR CONCLUSIONS FOR THIS WHOLE PROBLEM 2.

- It is concluded that there is not much significance in Gender and intention to graduate.
- Not much significance of Gender in earning above 50
- Retail is opted by Female and Management by Male. The probability that a randomly chosen female opts Retail is better than a randomly chosen Male who opts Management
- Male students are least likely to opt International Business and Female students are least likely to opt Others
- The probability of “Undecided” for graduation is higher for Female than Male

PROBLEM 3

SECTION 3.01 PROBLEM STATEMENT

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.

SECTION 3.02 DATA QUALITY

When null checks is run – returns True

There are 5 na in Set B data.

Also the not-null data in set A & set B vary.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36 entries, 0 to 35
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   A       36 non-null     float64
 1   B       31 non-null     float64
dtypes: float64(2)
memory usage: 704.0 bytes
```

SECTION 3.03 SOLUTIONS

(A) 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.

Null Hypothesis:	H0	The mean moisture content in both sets less than equal to 0.35/100 sq ft
Alternate Hypothesis:	H1	The mean moisture content in both sets Greater than 0.35/100 sq ft

	One Tailed Hypothesis
Assume	Alpha is 0.05

SET A:

```
One sample t test for Set A
t statistic -1.4735046253382782 and p statistic 0.07477633144907513
```

Since P is not significantly < alpha, we fail to reject H0.

So, Set A mean moisture less/equal to 0.35/100 sq ft

SET B:

```
One sample t test for Set B
t statistic -3.1003313069986995 and p statistic 0.0020904774003191813
```

Since P is significantly < alpha, we reject H0.

So, Set B mean moisture is **NOT less/equal to 0.35/100 sq ft**

(B) 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?

Null Hypothesis:	H0	MuA equal to Mu B
Alternate Hypothesis:	H1	MuA not equal to MuB
		One Tailed Hypothesis
	Assume	Alpha is 0.05

```
Two sample t test for Set B
t statistic 1.289628271966112 and p statistic 0.2017496571835328
Accept null hypothesis
```

The mean of Set A is equal to mean of Set B