SMDM Project Report

Deepti Yadav
April'22
Date : 24/04/2022

Table of Contents

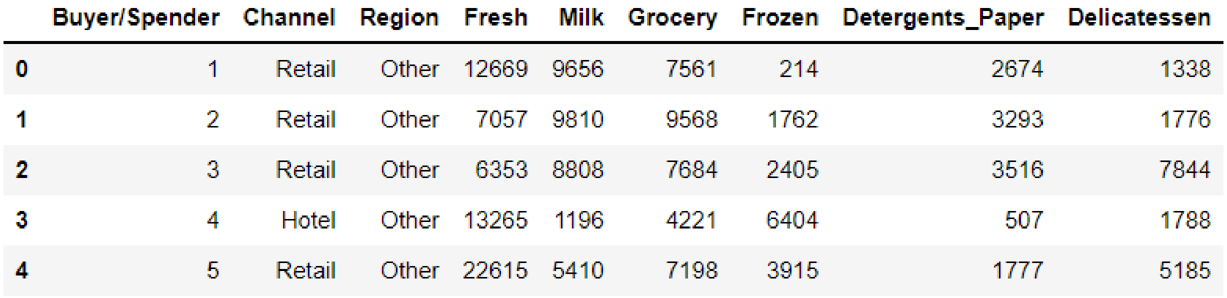
List of Figures

List of Tables

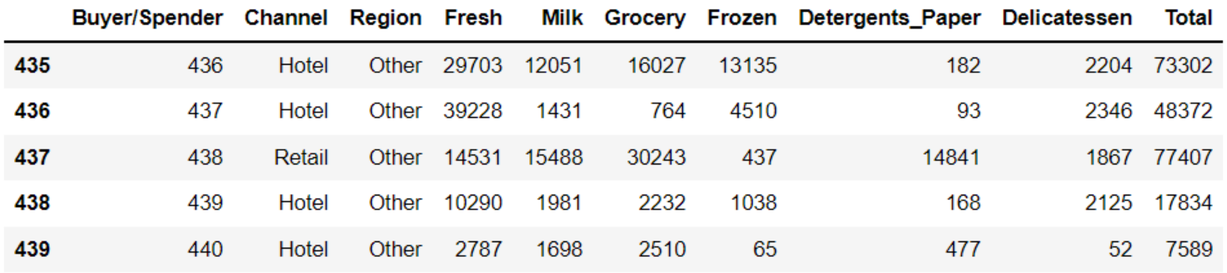
1 Problem 1 : Wholesale Customers Analysis

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).

Table 1.1 - Dataset Description



	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

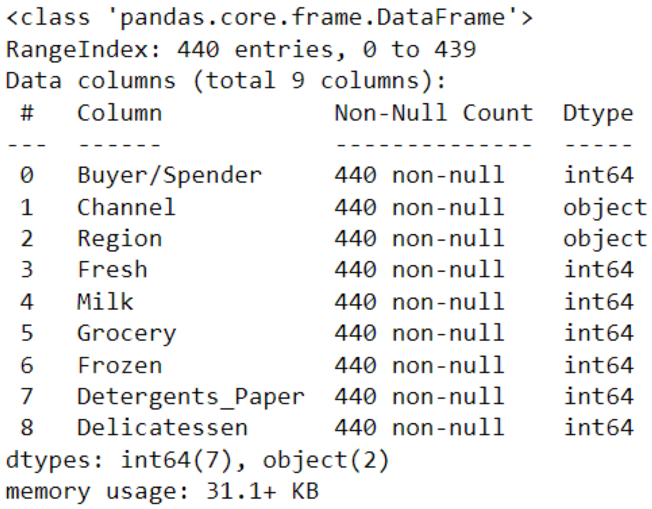


	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
435	436	Hotel	Other	29703	12051	16027	13135	182	2204	73302
436	437	Hotel	Other	39228	1431	764	4510	93	2346	48372
437	438	Retail	Other	14531	15488	30243	437	14841	1867	77407
438	439	Hotel	Other	10290	1981	2232	1038	168	2125	17834
439	440	Hotel	Other	2787	1698	2510	65	477	52	7589

Exploratory Data Analysis

Let us check the types of variables in the data frame.

Table 1.2 - Dataset Information

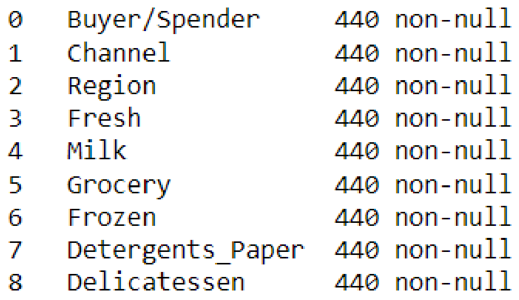


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 440 entries, 0 to 439
Data columns (total 9 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   Buyer/Spender     440 non-null    int64 
 1   Channel           440 non-null    object
 2   Region            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
```

There is total 440 rows and 9 columns in the dataset. Out of 9, 2 columns are of object type and rest 7 are of integer.

Check for missing

Table 1.3 - Missing values Check



```
0   Buyer/Spender     440 non-null
1   Channel           440 non-null
2   Region            440 non-null
3   Fresh             440 non-null
4   Milk              440 non-null
5   Grocery           440 non-null
6   Frozen            440 non-null
7   Detergents_Paper  440 non-null
8   Delicatessen      440 non-null
```

From the above results we can see that there is no missing value present in the dataset.

Correlation Plot

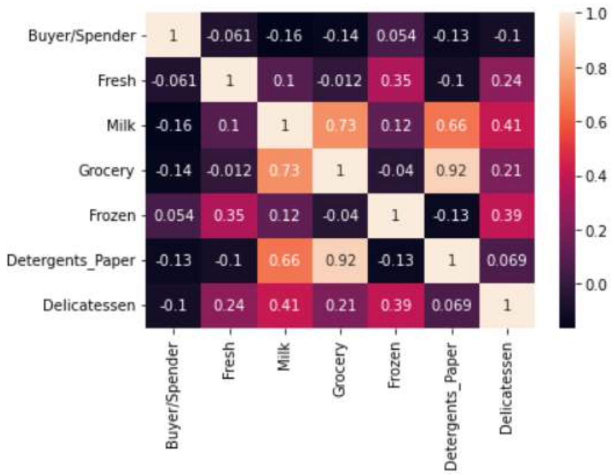


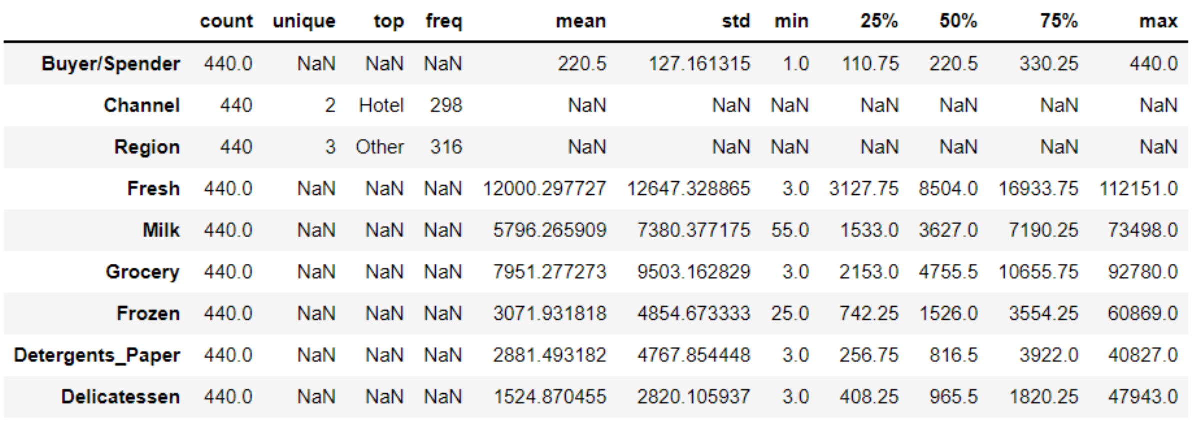
Figure 1.1 - Heatmap Correlation

From the correlation plot, we can see that annual spending of several items across different regions and channels are majorly positively Correlated.

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?

Descriptive statistics is concerned with Data Summarization in the form of Graphs/Charts and tables. Arithmetic Mean, Median and Mode are the most widely used measures of central tendency.

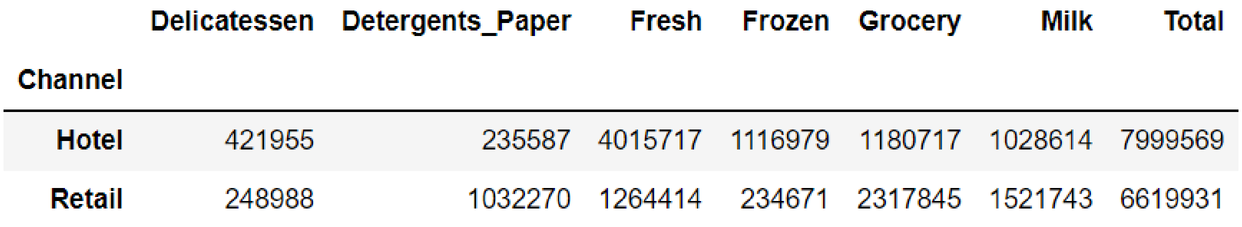
Table 1.4 - Summary of the data



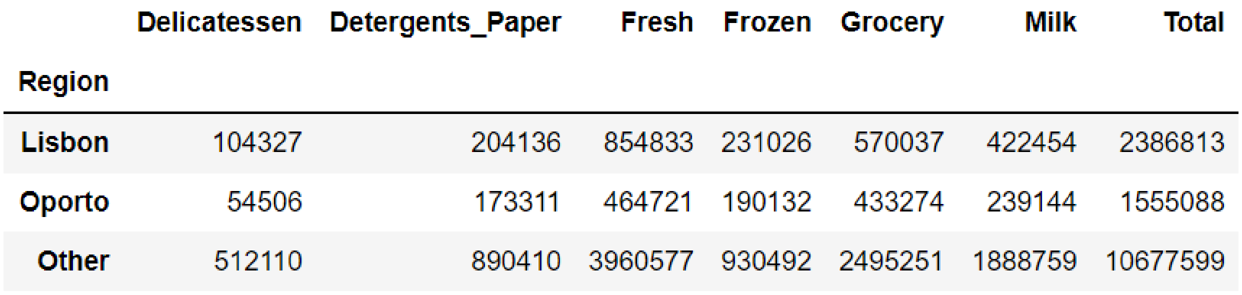
	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Buyer/Spender	440.0	NaN	NaN	NaN	220.5	127.161315	1.0	110.75	220.5	330.25	440.0
Channel	440	2	Hotel	298	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Region	440	3	Other	316	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Fresh	440.0	NaN	NaN	NaN	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0
Milk	440.0	NaN	NaN	NaN	5796.265909	7380.377175	55.0	1533.0	3627.0	7190.25	73498.0
Grocery	440.0	NaN	NaN	NaN	7951.277273	9503.162829	3.0	2153.0	4755.5	10655.75	92780.0
Frozen	440.0	NaN	NaN	NaN	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0
Detergents_Paper	440.0	NaN	NaN	NaN	2881.493182	4767.854448	3.0	256.75	816.5	3922.0	40827.0
Delicatessen	440.0	NaN	NaN	NaN	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0

From the descriptive statistics, we can see that there are 2 Channel and 3 regions.
On an average Fresh has maximum spending and Delicatessen has the least.

Table 1.5 – Spending across Channel and Region



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



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

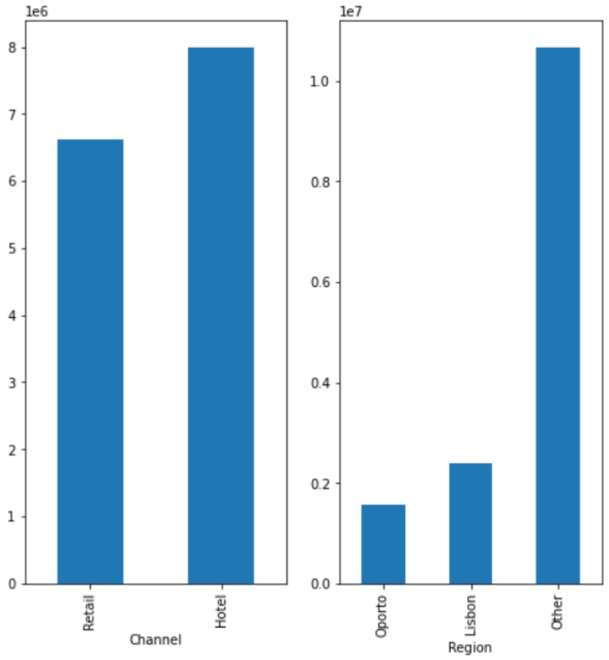


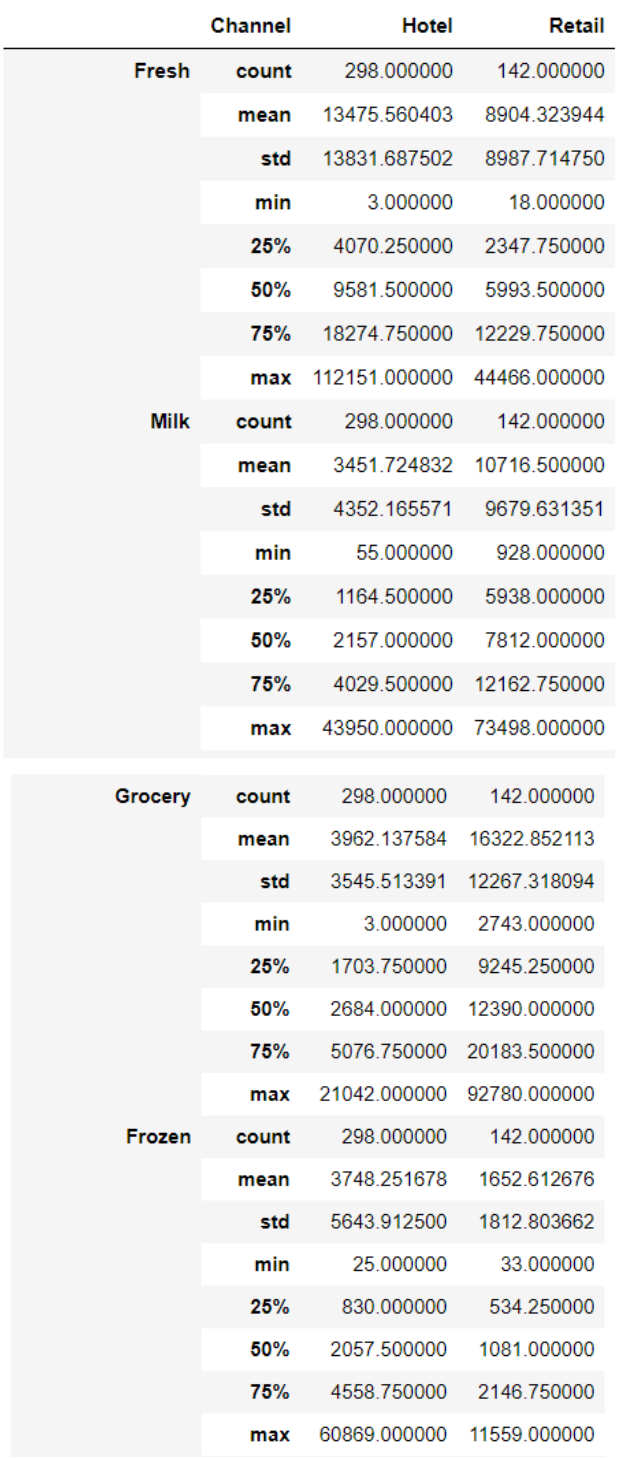
Figure 1.2 - Spending across Channel and Region

Out of 2 channels, **Hotel** spends most while **Retail** spends least.

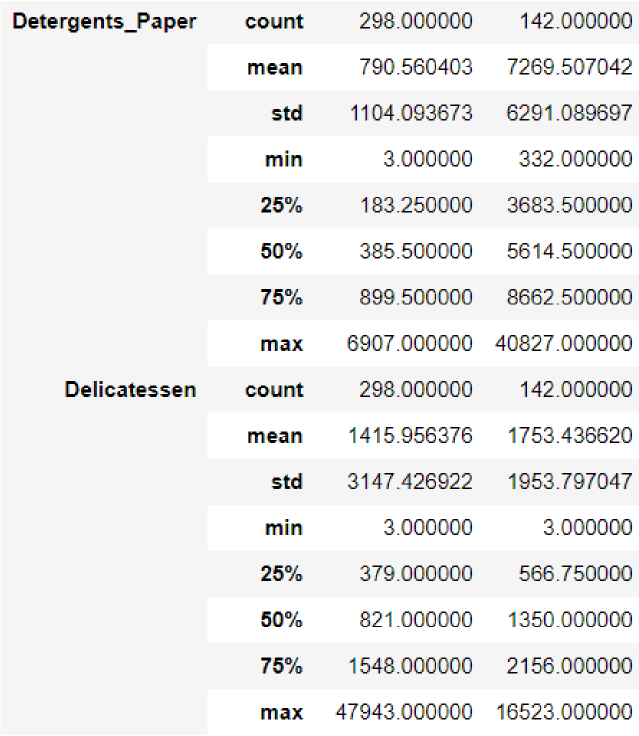
Out of 3 Regions, **Other** spends more while **Oporto** spends least.

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.

Table 1.6 – Varieties across Channel



	Channel	Hotel	Retail
Fresh	count	298.000000	142.000000
	mean	13475.560403	8904.323944
	std	13831.687502	8987.714750
	min	3.000000	18.000000
	25%	4070.250000	2347.750000
	50%	9581.500000	5993.500000
	75%	18274.750000	12229.750000
	max	112151.000000	44466.000000
Milk	count	298.000000	142.000000
	mean	3451.724832	10716.500000
	std	4352.165571	9679.631351
	min	55.000000	928.000000
	25%	1164.500000	5938.000000
	50%	2157.000000	7812.000000
	75%	4029.500000	12162.750000
	max	43950.000000	73498.000000
Grocery	count	298.000000	142.000000
	mean	3962.137584	16322.852113
	std	3545.513391	12267.318094
	min	3.000000	2743.000000
	25%	1703.750000	9245.250000
	50%	2684.000000	12390.000000
	75%	5076.750000	20183.500000
	max	21042.000000	92780.000000
Frozen	count	298.000000	142.000000
	mean	3748.251678	1652.612676
	std	5643.912500	1812.803662
	min	25.000000	33.000000
	25%	830.000000	534.250000
	50%	2057.500000	1081.000000
	75%	4558.750000	2146.750000
	max	60869.000000	11559.000000



Detergents_Paper	**count**	298.000000	142.000000
	mean	790.560403	7269.507042
	std	1104.093673	6291.089697
	min	3.000000	332.000000
	25%	183.250000	3683.500000
	50%	385.500000	5614.500000
	75%	899.500000	8662.500000
	max	6907.000000	40827.000000
Delicatessen	**count**	298.000000	142.000000
	mean	1415.956376	1753.436620
	std	3147.426922	1953.797047
	min	3.000000	3.000000
	25%	379.000000	566.750000
	50%	821.000000	1350.000000
	75%	1548.000000	2156.000000
	max	47943.000000	16523.000000

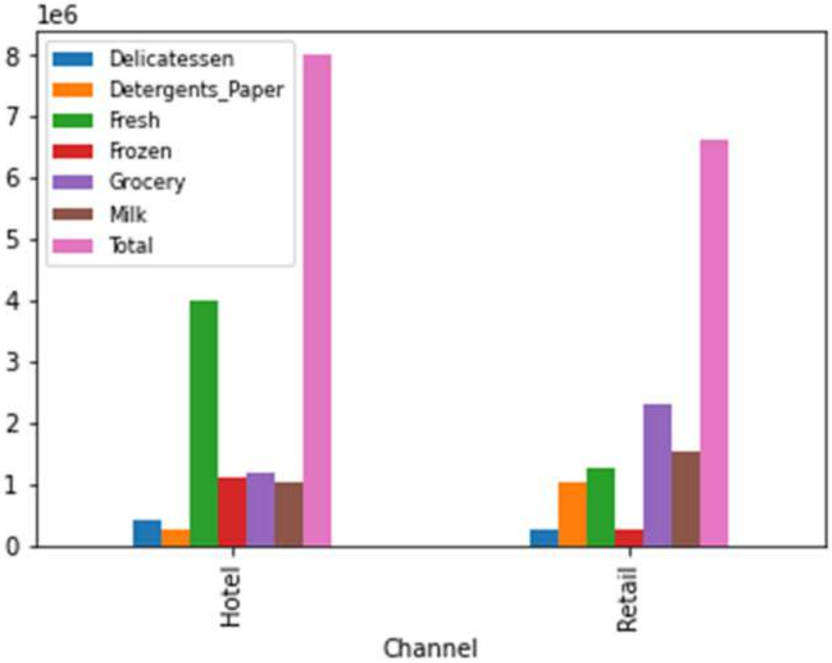
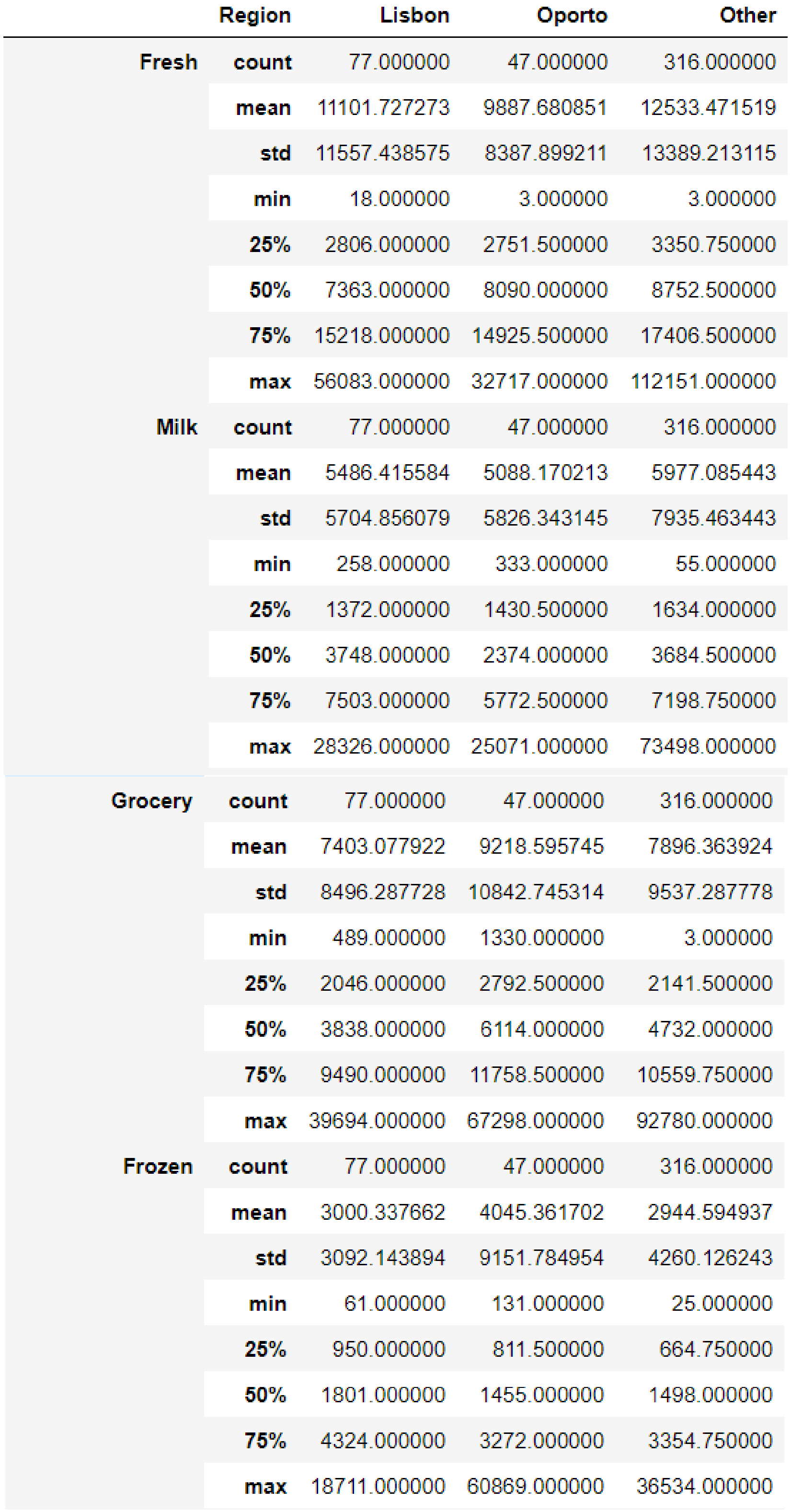


Figure 1.3 - Varieties across Channel

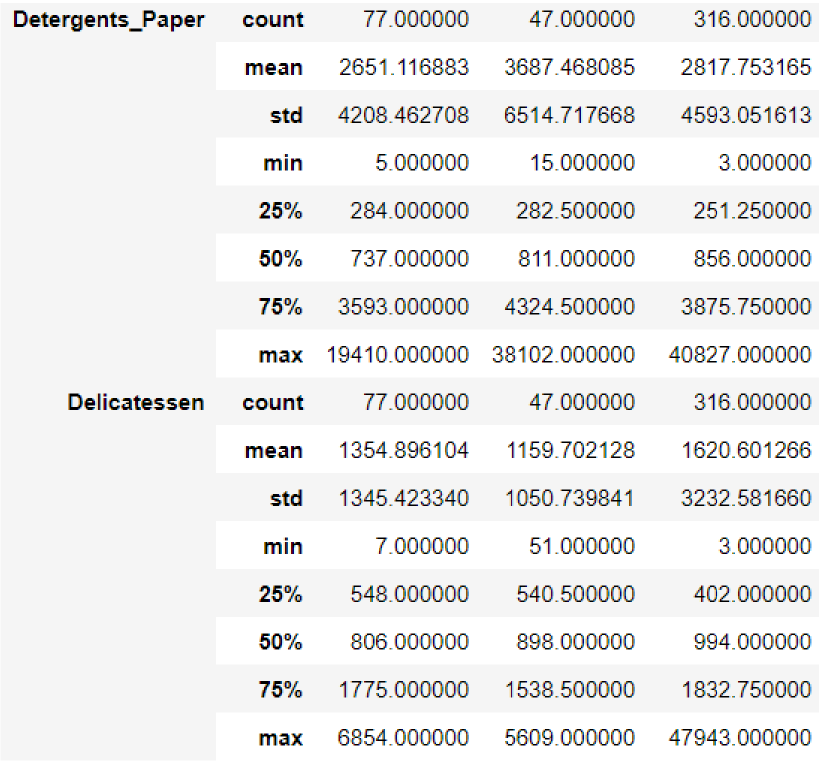
On an average, the Spending on **Fresh by a Hotel** channel is the **highest**, whereas, on an average, the Spending on **Detergents_Paper by a Hotel** channel is the **lowest**.

On an average, the Spending on **Grocery by a Retail** channel is the **highest**, whereas, on an average, the Spending on **Frozen by a Retail** channel are the **lowest**.

Table 1.7 – Varieties across Region



	Region	Lisbon	Oporto	Other
Fresh	**count**	77.000000	47.000000	316.000000
	mean	11101.727273	9887.680851	12533.471519
	std	11557.438575	8387.899211	13389.213115
	min	18.000000	3.000000	3.000000
	25%	2806.000000	2751.500000	3350.750000
	50%	7363.000000	8090.000000	8752.500000
	75%	15218.000000	14925.500000	17406.500000
	max	56083.000000	32717.000000	112151.000000
Milk	**count**	77.000000	47.000000	316.000000
	mean	5486.415584	5088.170213	5977.085443
	std	5704.856079	5826.343145	7935.463443
	min	258.000000	333.000000	55.000000
	25%	1372.000000	1430.500000	1634.000000
	50%	3748.000000	2374.000000	3684.500000
	75%	7503.000000	5772.500000	7198.750000
	max	28326.000000	25071.000000	73498.000000
Grocery	**count**	77.000000	47.000000	316.000000
	mean	7403.077922	9218.595745	7896.363924
	std	8496.287728	10842.745314	9537.287778
	min	489.000000	1330.000000	3.000000
	25%	2046.000000	2792.500000	2141.500000
	50%	3838.000000	6114.000000	4732.000000
	75%	9490.000000	11758.500000	10559.750000
	max	39694.000000	67298.000000	92780.000000
Frozen	**count**	77.000000	47.000000	316.000000
	mean	3000.337662	4045.361702	2944.594937
	std	3092.143894	9151.784954	4260.126243
	min	61.000000	131.000000	25.000000
	25%	950.000000	811.500000	664.750000
	50%	1801.000000	1455.000000	1498.000000
	75%	4324.000000	3272.000000	3354.750000
	max	18711.000000	60869.000000	36534.000000



Detergents_Paper	**count**	77.000000	47.000000	316.000000
	mean	2651.116883	3687.468085	2817.753165
	std	4208.462708	6514.717668	4593.051613
	min	5.000000	15.000000	3.000000
	25%	284.000000	282.500000	251.250000
	50%	737.000000	811.000000	856.000000
	75%	3593.000000	4324.500000	3875.750000
	max	19410.000000	38102.000000	40827.000000
Delicatessen	**count**	77.000000	47.000000	316.000000
	mean	1354.896104	1159.702128	1620.601266
	std	1345.423340	1050.739841	3232.581660
	min	7.000000	51.000000	3.000000
	25%	548.000000	540.500000	402.000000
	50%	806.000000	898.000000	994.000000
	75%	1775.000000	1538.500000	1832.750000
	max	6854.000000	5609.000000	47943.000000

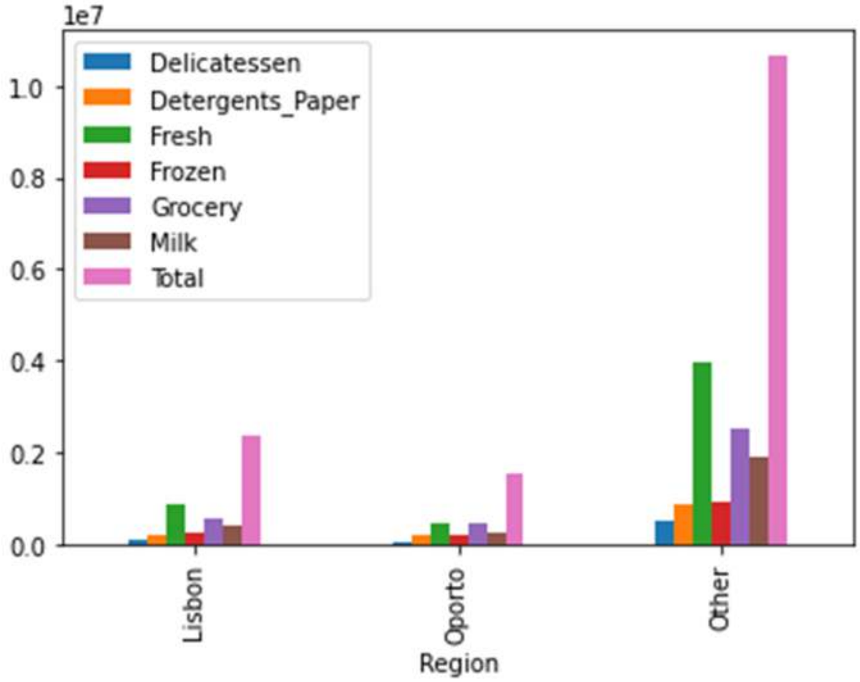


Figure 1.4 - Varieties across Region

On an average, **Lisbon** has **highest** Spending on **Fresh** and **lowest** spending on **Delicatessen.**

On an average, **Oporto** has **highest** Spending on **Fresh** and **lowest** spending on Spending on **Delicatessen.**

On an average, **Other** has **highest** Spending on **Fresh** and **lowest** spending on Spending on **Delicatessen.**

Overall conclusion is that all the three region has highest spending on fresh and lowest spending on Delicatessen.

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?

Table 1.8 – Coefficient of variation across Channel

	Hotel	**Retail**
CV_Fresh	1.026428	1.009365
CV_Milk	1.260867	0.903246
CV_Grocery	0.894849	0.751543
CV_Frozen	1.505745	1.096932
CV_Detergents_Paper	1.396596	0.865408
CV_Delicatessen	2.222828	1.114267

Delicatessen has most inconsistent behavior in Hotel and Retail with CV = 2.22 and CV = 1.11 respectively.

Grocery has least inconsistent behavior in Hotel and Retail with CV = 0.89 and CV = 0.75 respectively.

Table 1.9 – Coefficient of variation across Region

	Lisbon	Oporto	Other
CV_Fresh	1.041049	0.848318	1.068277
CV_Milk	1.039815	1.145076	1.327648
CV_Grocery	1.147670	1.176182	1.207808
CV_Frozen	1.030599	2.262291	1.446761
CV_Detergents_Paper	1.587430	1.766718	1.630040
CV_Delicatessen	0.993008	0.906043	1.994680

In Lisbon, Detergents_Paper has most inconsistent behavior with CV = 1.587 and Delicatessen has least inconsistent behavior with CV = 0.993.

In Oporto, Frozen has most inconsistent behavior with CV = 2.26 and Fresh has least inconsistent behavior with CV = 0.848

In Other, Delicatessen has most inconsistent behavior with CV = 1.99 and Fresh has least inconsistent behavior with CV = 1.068

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

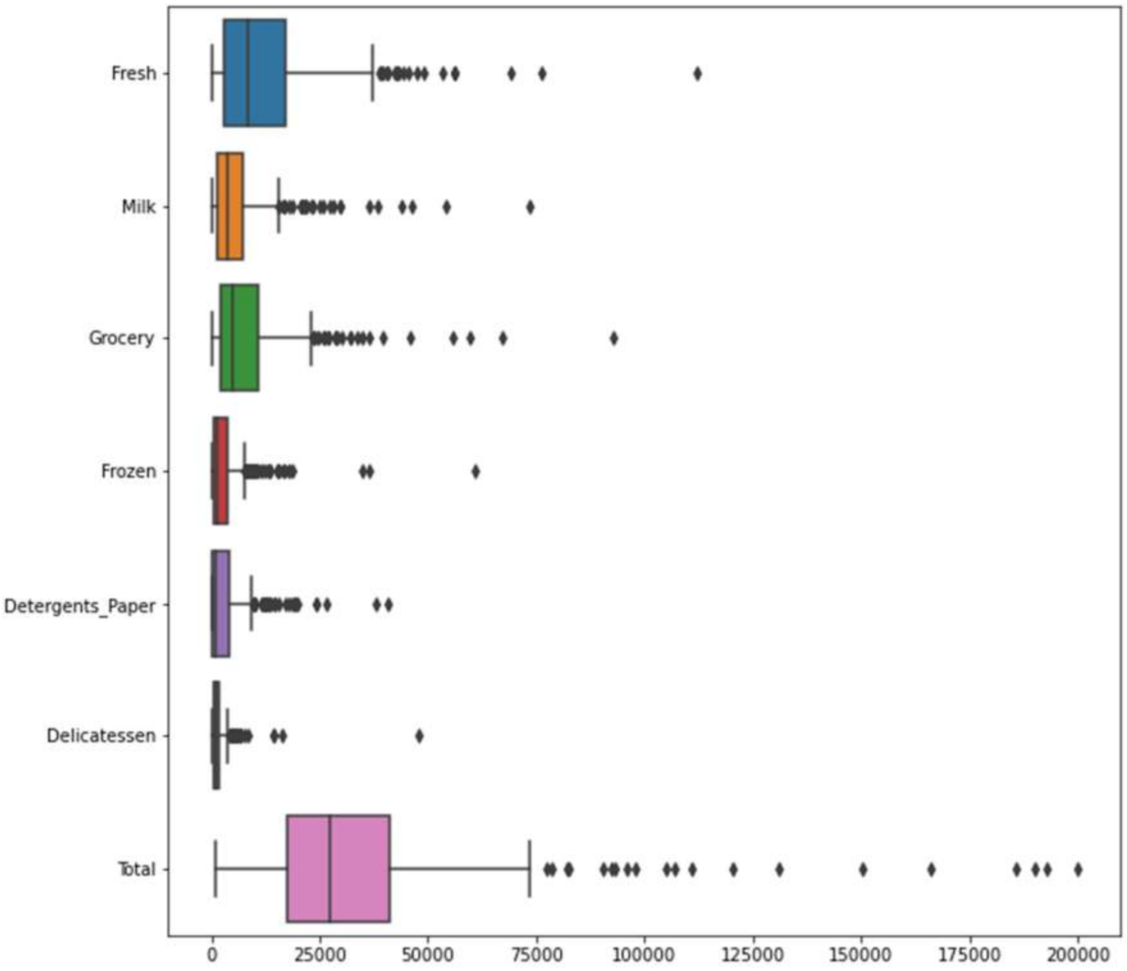


Figure 1.5 – Boxplot for Data

From the above boxplot, every variety has outliers.

Outliers are extreme values that stand out from the pattern of other values in a dataset.

This can potentially help in discovering inconsistencies and detect any errors in the data.

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

From this analysis recommendations for the business are:

(1) The total spending on fresh products across different regions is maximum and Spending on Grocery by a Retail channel is the highest so the company needs to ensure that it is driving the most profit from these food items and accordingly inventory is maintained.

(2) Grocery has least inconsistent behavior in Hotel and Retail and Fresh has least inconsistent in Oporto and other regions, So, the business should invest more in these food item because it is less risky.

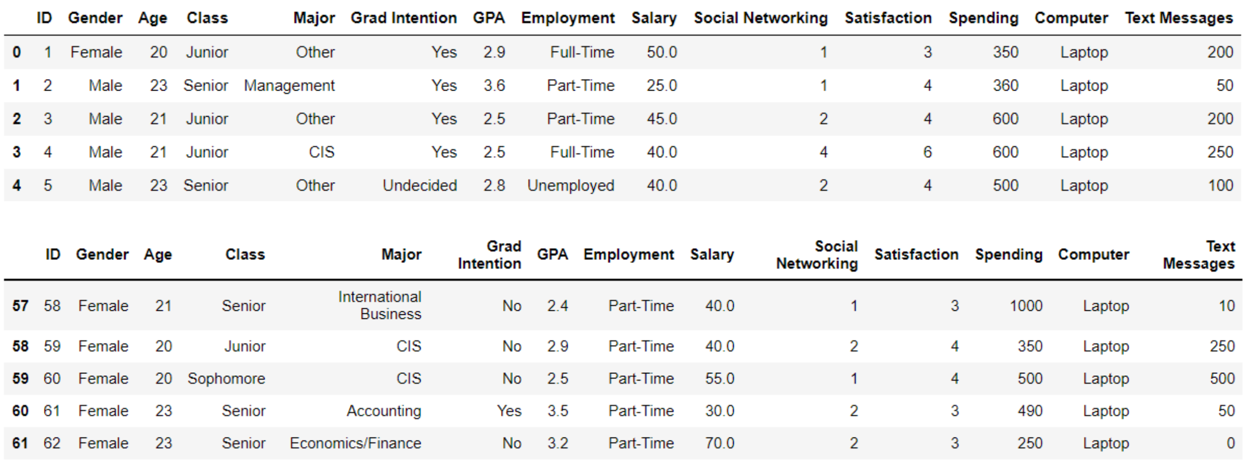
(3) Delicatessen has most inconsistent behavior in Hotel and Retail while Detergents_Paper, Delicatessen and Frozen has most inconsistent behavior in different regions. Distributor must look for the reasons of these inconsistencies and try to minimize them.

(4) Fresh item has highest standard deviation which should be minimized.

2 Problem 2 : Survey

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).

Table 2.1 - Dataset Description



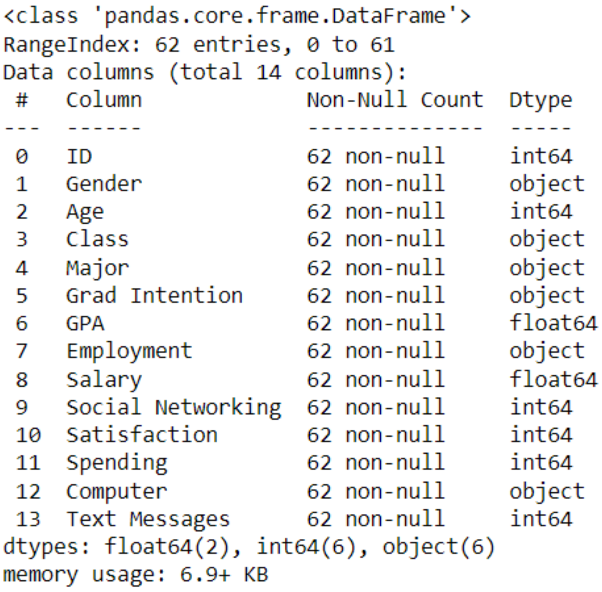
	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
0	1	Female	20	Junior	Other	Yes	2.9	Full-Time	50.0	1	3	350	Laptop	200
1	2	Male	23	Senior	Management	Yes	3.6	Part-Time	25.0	1	4	360	Laptop	50
2	3	Male	21	Junior	Other	Yes	2.5	Part-Time	45.0	2	4	600	Laptop	200
3	4	Male	21	Junior	CIS	Yes	2.5	Full-Time	40.0	4	6	600	Laptop	250
4	5	Male	23	Senior	Other	Undecided	2.8	Unemployed	40.0	2	4	500	Laptop	100

	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
57	58	Female	21	Senior	International Business	No	2.4	Part-Time	40.0	1	3	1000	Laptop	10
58	59	Female	20	Junior	CIS	No	2.9	Part-Time	40.0	2	4	350	Laptop	250
59	60	Female	20	Sophomore	CIS	No	2.5	Part-Time	55.0	1	4	500	Laptop	500
60	61	Female	23	Senior	Accounting	Yes	3.5	Part-Time	30.0	2	3	490	Laptop	50
61	62	Female	23	Senior	Economics/Finance	No	3.2	Part-Time	70.0	2	3	250	Laptop	0

Exploratory Data Analysis

Let us check the types of variables in the data frame.

Table 2.2 - Dataset Information

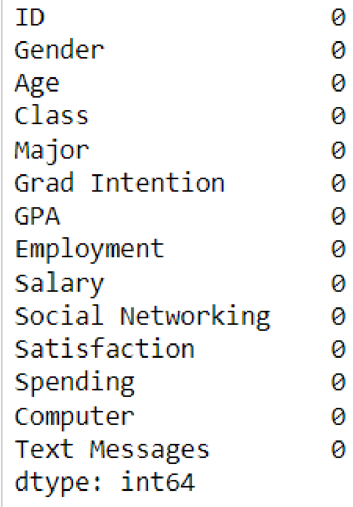


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62 entries, 0 to 61
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   ID                 62 non-null     int64  
 1   Gender             62 non-null     object 
 2   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
```

There is total 62 rows and 14 columns in the dataset. Out of 14, 6 columns are of object type , 6 are of integer type and 2 are float type.

Check for missing

Table 2.3 - Missing values Check



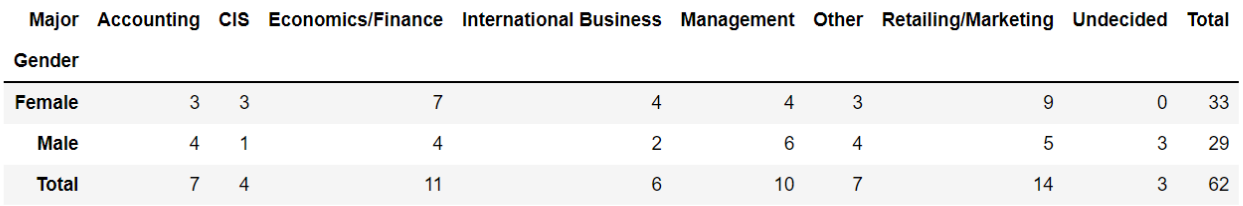
```
ID                  0
Gender              0
Age                 0
Class               0
Major               0
Grad Intention      0
GPA                 0
Employment          0
Salary              0
Social Networking   0
Satisfaction        0
Spending            0
Computer            0
Text Messages       0
dtype: int64
```

From the above results we can see that there is no missing value present in the dataset.

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

2.1.1 Gender and Major

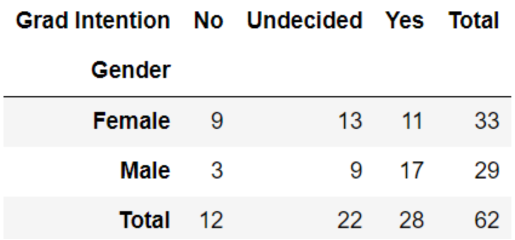
Table 2.4 – Contingency Table (Gender & Major)



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

2.1.2 Gender and Grad Intention

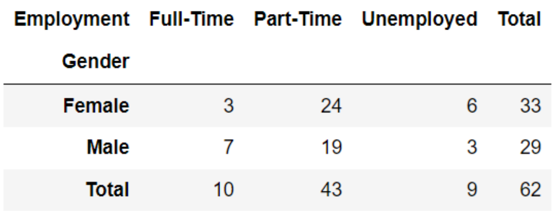
Table 2.5 – Contingency Table (Gender & Grad Intention)



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

2.1.3 Gender and Employment

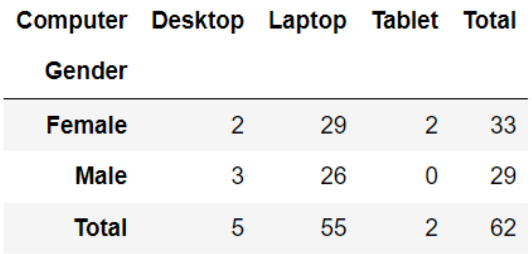
Table 2.6 – Contingency Table (Gender & Employment)



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

2.1.4 Gender and Computer

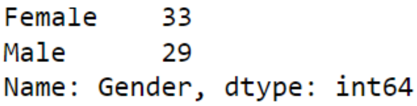
Table 2.7 – Contingency Table (Gender & Computer)



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

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
Name: Gender, dtype: int64
```

The probability that a randomly selected CMSU student will be male = 29/62 = 0.4677

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 female = 33/62 = 0.532

2.3 Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

2.3.1. Find the conditional probability of different majors among the male students in CMSU.

From Table 2.1 (Contingency Table (Gender & Major)

Among MALE candidates:
Probability of Accounting as major = 4/29 = **0.138**

Probability of CIS as major= 1/29 = **0.0345**
Probability of Economics/Finance as major = 4/29 = **0.138**
Probability of International Business = 2/29 = **0.069**
Probability of Management as Major = 6/29 = **0.207**
Probability of Other as Major = 4/29 = **0.138**
Probability of Retailing/Marketing as Major = 5/29 = **0.172**
Probability of Undecided as Major = 3/29 = **0.103**

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

From Table 2.1 (Contingency Table (Gender & Major)

Among FEMALE candidates:
Probability of Accounting as major = 3/33 = **0.091**
Probability of CIS as major= 3/33 = **0.091**
Probability of Economics/Finance as major = 7/33 = **0.212**
Probability of International Business = 4/33 = **0.121**
Probability of Management as Major = 4/33 = **0.121**
Probability of Other as Major = 3/33 = **0.091**
Probability of Retailing/Marketing as Major = 9/33= 0.272
Probability of Undecided as Major = 0/33 = 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.

From Table 2.2 (Contingency Table (Gender & Grad Intention)

Probability That a randomly chosen student is a male = 29/62
Probability of Male intends to graduate = 17/29
Probability that a randomly chosen student is a male and intends to graduate = 17/29) * (29/62) = **0.274**

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

From Table 2.4 (Contingency Table (Gender & **Computer)**

Probability That a randomly chosen student is a Female = 33/62
Probability of Female with No Laptop = 4/33
Probability that a randomly selected student is a female and does NOT have a laptop = (4/33) * (33/62) = **.0645**

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

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

From Table 2.3 (Contingency Table (Gender & Employment)

Probability of a Student being Male = 29/62
Probability of a student having Full Time Employment = 10/62
Probability of a Male having Full Time Employment = 7/62

Probability that a randomly chosen student is a male or has full-time employment=
Probability of a Student being Male + Probability of a student having Full Time Employment
-Probability of a Male having Full Time Employment
= (29/62) + (10/62) -(7/62)
= 0.516

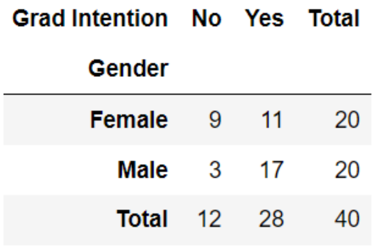
2.5.2. Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management.

From Table 2.1 (Contingency Table (Gender & Major)

Probability of Female in international Business = 4/33
Probability of Female in Management = 4/33
Probability that given a female student is randomly chosen, she is majoring in international business or management = Probability of Female in international Business + Probability of Female in Management
= 4/33 + 4/33 = **0.242**

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?

Table 2.8 – Contingency Table without undecided (Gender & Grad Intention)



Grad Intention	No	Yes	Total
Gender			
Female	9	11	20
Male	3	17	20
Total	12	28	40

For 2 events to be independent, following condition is to be satisfied

$P(A \cap B) = P(A) * P(B)$

So, $P\ (\text{Yes} \cap \text{Female}) = P(\text{Yes}) * P(\text{Female})$

$P(\text{Female}) = 33/62 = 0.532258064516129$

$P(\text{Yes}) = 28/62 = 0.45161290322580644$

$P(\text{Yes}) * P(\text{Female}) = 0.532258064516129 * 0.45161290322580644 = 0.24037460978147762$

$P\ (\text{Yes} \cap \text{Female}) = 11/62 = 0.177$

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

No of students with GPA is less than 3 = 17

The probability that if a student is chosen randomly, his/her GPA is less than 3 = 17/62 = 0.274

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.

The probability that a randomly selected male earns 50 or more = 14/29 = 0.483
The probability that a randomly selected female earns 50 or more is : 0.545

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.

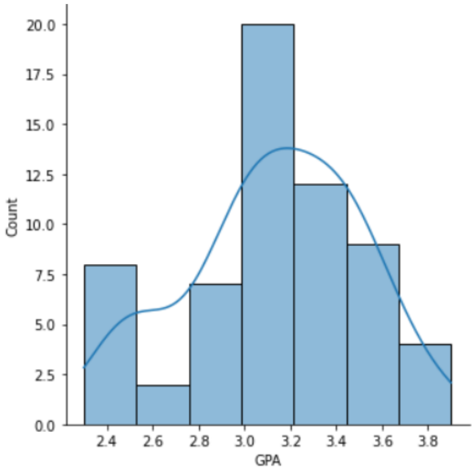


Figure 2.1 – Histogram for GPA

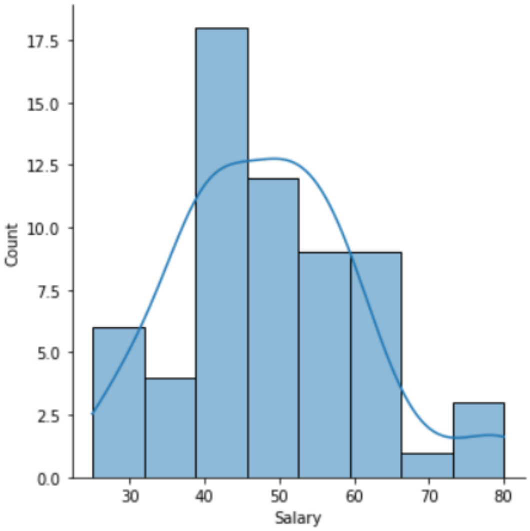


Figure 2.2 – Histogram for Salary

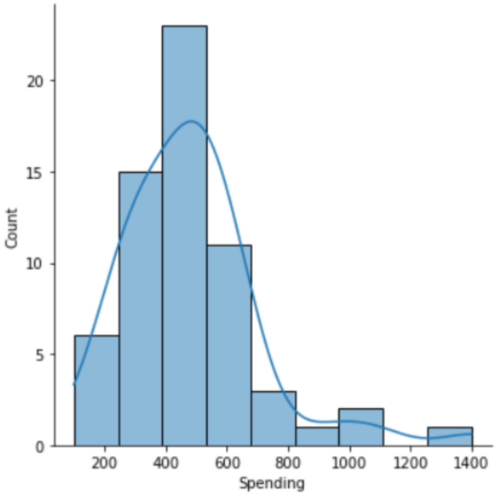


Figure 2.3 – Histogram for Spending

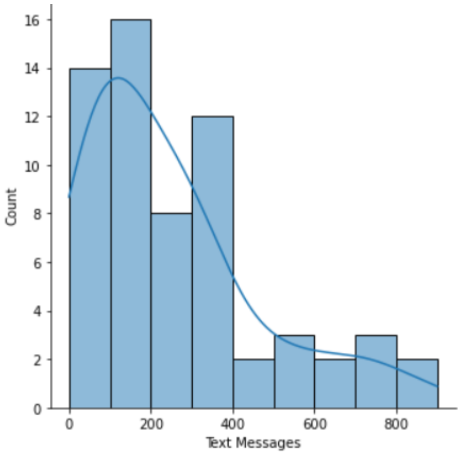


Figure 2.4 – Histogram for Text Messages

For GPA, ShapiroResult(statistic=0.9685361981391907, pvalue=0.11204058676958084)
For Salary, ShapiroResult(statistic=0.9565856456756592, pvalue=0.028000956401228905)
For Spending, hapiroResult(statistic=0.8777452111244202, pvalue=1.6854661225806922e-05)
For Text Messages, hapiroResult(statistic=0.8594191074371338, pvalue=4.324040673964191e-06)

From Shapiro Wilk Test, only for GPA p value > 0.05, so only GPA is normally distributed.

Conclusion : From this analysis, we can conclude that the sample survey conducted for the students from CMSU shows that there are multiple factors that affect the graduation of a student.

The survey conducted by has information about what major the undergrad students are pursuing, whether they intent to graduate, what is their GPA, nature of their employment and their salary, social networking, spending, satisfaction, computer, and text messages.

Using our analysis, we have constructed contingency tables and calculated probabilities between these variables, investigated about outliers, checked the distribution about some variables.

We can conclude that

Retailing/Marketing

1) Retail/Marketing is opted by 14 students which is maximum among all other majors.
2) Probability that a randomly selected female earns 50 is higher than that of a male.
3) Probabilities of male students graduating is more than that of female students, so female students need more support and choice of major.

3 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.

The file (A & B shingles.csv) includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.

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

H0_A : mean moisture content <=0.35

Ha_A : mean moisture content > 0.35

For the B shingles, the null and alternative hypothesis to test whether the population mean moisture content is less than 0.35 pound per 100 square feet is given:

H0_B : mean moisture content <=0.35

Ha_B : mean moisture content > 0.35

α = 0.05

For the A shingles

One sample t test
t_stats: -1.4735046253382782 p value: 0.07477633144907513

Since pvalue_A > 0.05, do not reject H0 . **So, the statistical decision is to fail to reject the null hypothesis at 5% level of significance level**. So, there is no sufficient evidence to prove that mean moisture content for Sample A shingles is greater than 0.35 pounds per 100 square feet.

For the B shingles

One sample t test
t_stats: -3.1003313069986995 p value: 0.0020904774003191826
Since pvalue < 0.05 at 5% level of significance level . **So, the statistical decision is to reject the null hypothesis at 5% level of significance.** So, there is sufficient evidence to prove that mean moisture content for Sample B shingles is greater than 0.35 pounds per 100 square feet.

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?

For the A & B shingles, the null and alternative hypothesis to test whether population mean for shingles A and B are equal given:

H0 : mean moisture content A = mean moisture content B

Ha : mean moisture content A ≠ mean moisture content B

α = 0.05

t_statistic = 1.2896282719661123
p_value = 0.2017496571835306

As the pvalue = .2017 > α So the statistical decision is do not reject H0. We can say that mean moisture content for shingles A and B are equal.