Problem - 1

Wholesale Customer Analysis

Sharjil Shah

PGP-DSBA Online January' 21

Date: 13/March/2022

CONTENTS

EXECUTIVE 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 annual spending on different varieties of products in different regions and across different sales channel. In this problem we will explore the different attributes of the given data and

INTRODUCTION

The purpose of this whole exercise is to explore data. We do the exploratory analysis. Explore the dataset using central tendency and other parameters. The data consists of 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). Analysed the different

DATA DESCRIPTION

```
Buyer/Spender
Channel - Retail/Hotel
Region - Other/Oporto/Lisbon
Fresh - continuous from 3.0 to 112151.0
Milk  - continuous from 55.0 to 73498.0
Grocery - continuous from 3.0 to 92780.0
Frozen - continuous from 25.0 to 60869.0
Detergents_Paper  - continuous from 3.0 to 40827.0
Delicatessen  -  continuous from 3.0 to 47943.0
```

- 6 continuous types of feature ('Fresh', 'Milk', 'Grocery', 'Frozen', 'Detergents Paper', 'Delicatessen')
- 2 categorical features ('Channel', 'Region')
- 1 continuous types of feature (Buyer/Spender) will be dropped as no use for our analysis

Sample dataset

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

EXPLORATORY DATA ANALYSIS

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

```
Buyer/Spender       int64
Channel             object
Region              object
Fresh               int64
Milk                int64
Grocery             int64
Frozen              int64
Detergents_Paper    int64
Delicatessen        int64
```

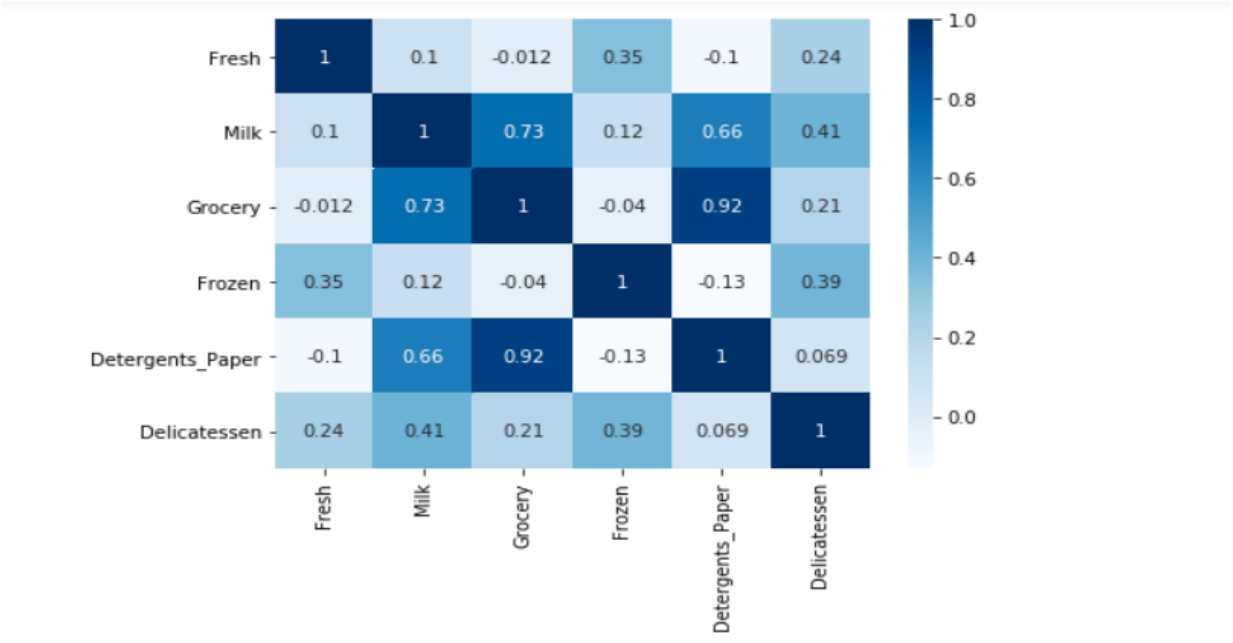
There are total 440 rows and 9 columns in the dataset. Out of 9, 7 columns are of object type and rest 2 are of either integer type or float type.

Check for missing values in the dataset:

```
RangeIndex: 440 entries, 0 to 439
Data columns (total 9 columns):
Buyer/Spender       440 non-null
Channel             440 non-null
Region              440 non-null
Fresh               440 non-null
Milk                440 non-null
Grocery             440 non-null
Frozen              440 non-null
Detergents_Paper    440 non-null
Delicatessen        440 non-null
```

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

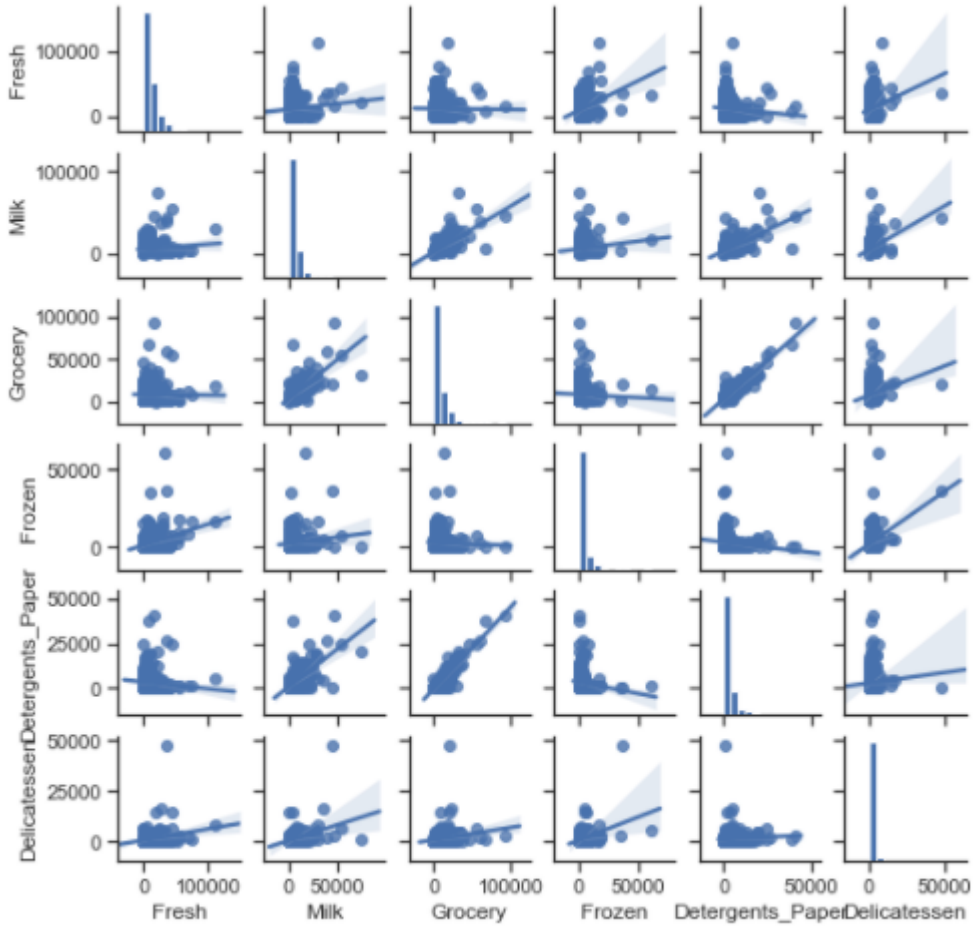
CORRELATION PLOT



From the correlation plot, we can see that few attributes are highly correlated to each other. Correlation values near to 1 or -1 are highly positively correlated and highly negatively correlated respectively. Correlation values near to 0 are not correlated to each other.

PAIRPLOT

Pairplot shows the relationship between the variables in the form of scatterplot and the distribution of the variable in the form of histogram.



QUESTIONS:

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 Statastics of our Data:

	count	mean	std	min	25%	50%	75%	max
Fresh	440.0	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0
Milk	440.0	5796.265909	7380.377175	55.0	1533.00	3627.0	7190.25	73498.0
Grocery	440.0	7951.277273	9503.162829	3.0	2153.00	4755.5	10655.75	92780.0
Frozen	440.0	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0
Detergents_Paper	440.0	2881.493182	4767.854448	3.0	256.75	816.5	3922.00	40827.0
Delicatessen	440.0	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0

Descriptive Statastics of our Data including Channel & Retail:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
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	NaN	NaN	NaN	12000.3	12647.3	3	3127.75	8504	16933.8	112151
Milk	440	NaN	NaN	NaN	5796.27	7380.38	55	1533	3627	7190.25	73498
Grocery	440	NaN	NaN	NaN	7951.28	9503.16	3	2153	4755.5	10655.8	92780
Frozen	440	NaN	NaN	NaN	3071.93	4854.67	25	742.25	1526	3554.25	60869
Detergents_Paper	440	NaN	NaN	NaN	2881.49	4767.85	3	256.75	816.5	3922	40827
Delicatessen	440	NaN	NaN	NaN	1524.87	2820.11	3	408.25	965.5	1820.25	47943

Measure of Central Tendency - Mean, Median, mode Measure of Dispersion - Range, IQR, Standard Deviation

Summary:

From the above two describe function, we can infer the following:

- Channel has two unique values, with "Hotel" as most frequent with 298 out of 440 transactions. Which means 67.7% of spending comes from "Hotel" channel.
- Retail has three unique values, with "Other" as most frequent with 316 out of 440 transactions. Which means 71.8% of spending comes from "Other" region.
- Fresh item (440 records),

  Has a mean of 12000.3, standard deviation of 12647.3, with min value of 3 and max value of 112151.

  The other aspect is Q1 (25%) is 3127.75, Q3 (75%) is 16933.8, with Q2 (50%) 8504

  Range = max-min =112151-3=112,148 & IQR = Q3-Q1 = 16933.8-3127.75 = 13,806.05 (this helpful in calculating the outlier (1.5 IQR Lower/Upper limit))

- Milk item (440 records),

  Has a mean of 5796.27, standard deviation of 7380.38, with min value of 55 and max value of 73498.

The other aspect is Q1 (25%) is 1533, Q3 (75%) is 7190.25, with Q2 (50%) 3627

Range = max-min =73498-55=73443 & IQR = Q3-Q1 = 7190.25-1533 = 5657.25 (this helpful in calculating the outlier (1.5 IQR Lower/Upper limit))

- Grocery item (440 records),

  Has a mean of 7951.28, standard deviation of 9503.16, with min value of 3 and max value of 92780.

  The other aspect is Q1 (25%) is 2153, Q3 (75%) is 10655.8, with Q2 (50%) 4755.5

  Range = max-min =92780-3=92777 & IQR = Q3-Q1 = 10655.8-2153 = 8502.8 (this helpful in calculating the outlier (1.5 IQR Lower/Upper limit))

- Frozen (440 records),

  Has a mean of 3071.93, standard deviation of 4854.67, with min value of 25 and max value of 60869.

  The other aspect is Q1 (25%) is 742.25, Q3 (75%) is 3554.25, with Q2 (50%) 1526

  Range = max-min =60869-25=60844 & IQR = Q3-Q1 = 3554.25-742.25 = 2812 (this helpful in calculating the outlier (1.5 IQR Lower/Upper limit))

- Detergents Paper (440 records),

  Has a mean of 2881.49, standard deviation of 4767.85, with min value of 3 and max value of 40827.

  The other aspect is Q1 (25%) is 256.75, Q3 (75%) is 3922, with Q2 (50%) 816.5

  Range = max-min =40827-3=40824 & IQR = Q3-Q1 = 3922-256.75 = 3665.25 (this helpful in calculating the outlier (1.5 IQR Lower/Upper limit))

- Delicatessen (440 records),

  Has a mean of 1524.87, standard deviation of 2820.11, with min value of 3 and max value of 47943.

  The other aspect is Q1 (25%) is 408.25, Q3 (75%) is 1820.25, with Q2 (50%) 965.5

  Range = max-min =47943-3=47940 & IQR = Q3-Q1 = 1820.25-408.25 = 1412 (this helpful in calculating the outlier (1.5 IQR Lower/Upper limit))

To find the which region and channel has spent the most and which region and channel has spent the least, we create a new column “Spending” and append it to the dataset.

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

```
Region
Lisbon      2386813
Oporto      1555088
Other      10677599
Name: Spending, dtype: int64

Channel
Hotel      7999569
Retail     6619931
Name: Spending, dtype: int64
```

- **The Highest spend in the Region/Channel is from Others/Hotel**
- **The lowest spend in the Region/Channel is from Oporto/Hotel**

Channel	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Spending
Hotel	13475.56	3451.72	3962.14	3748.25	790.56	1415.96	26844.19
Retail	8904.32	10716.50	16322.85	1652.61	7269.51	1753.44	46619.23

- **In Channel "Hotel" Average Highest Spending in Fresh items and Lowest Spending in Detergents_Paper.**
- **In Channel "Retail" Average Highest Spending in Grocery items and Lowest Spending in Frozen items.**

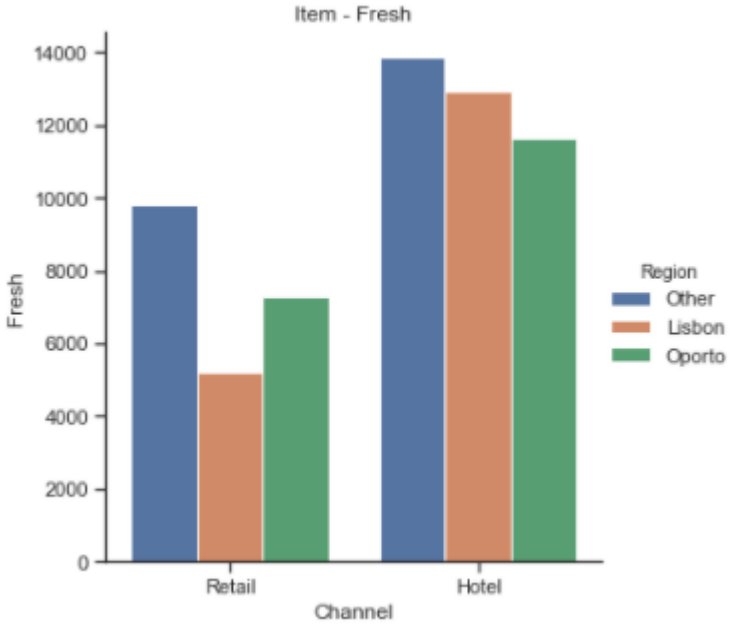
Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Spending
Lisbon	11101.73	5486.42	7403.08	3000.34	2651.12	1354.9	30997.57
Oporto	9887.68	5088.17	9218.60	4045.36	3687.47	1159.7	33086.98
Other	12533.47	5977.09	7896.36	2944.59	2817.75	1620.6	33789.87

- **In Region "Lisbon" Average Highest Spending in Fresh and Lowest in Delicatessen items.**
- **In Region "Oporto" Average Highest Spending in Fresh and Lowest in Delicatessen items.**
- **In Region "Other" Average Highest Spending in Fresh and Lowest in Delicatessen items.**

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.

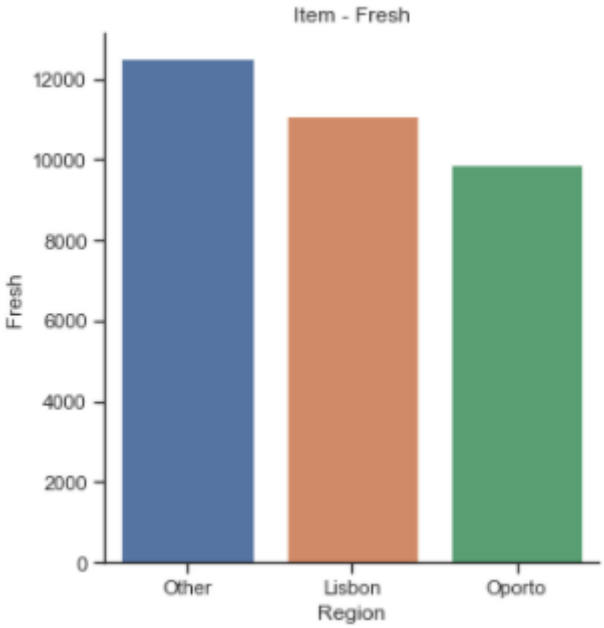
First let us compare the spending in each variety with respect to Region and with respect to Channel.

Item – Fresh v/s Channel:



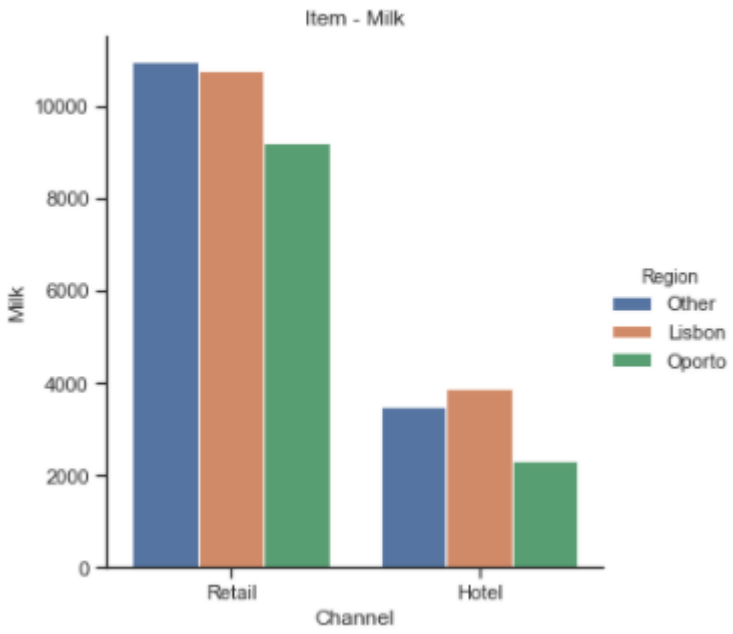
With the above graph, we can see the spending of Fresh Item in Hotel is higher than the spending of Fresh Item in Retail. The total spending of fresh region in Oporto is the least in the Hotel Channel where as the spending of fresh items in Lisbon is least in Retail Channel. Comparatively spending of Fresh Item in Other region in both Hotel and Retail channel is significantly higher.

Item- Fresh v/s Region:



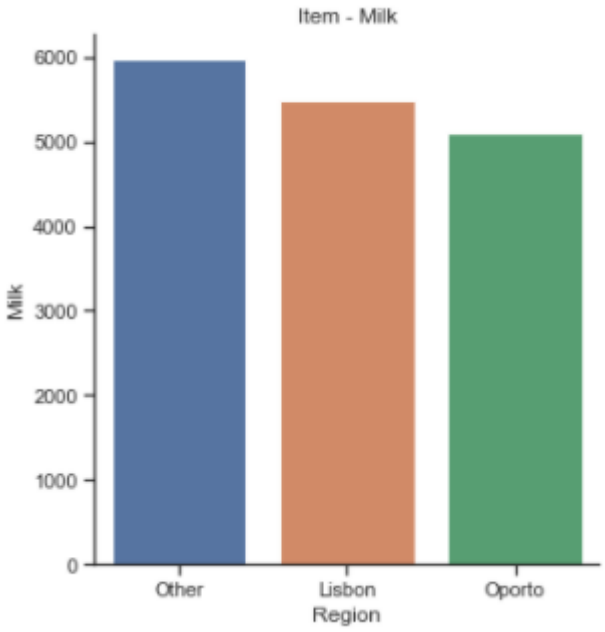
The spending of Fresh Items in other region is the highest. The spending of Fresh Items in other region is the lowest in Oporto.

Item Milk v/s Channel:



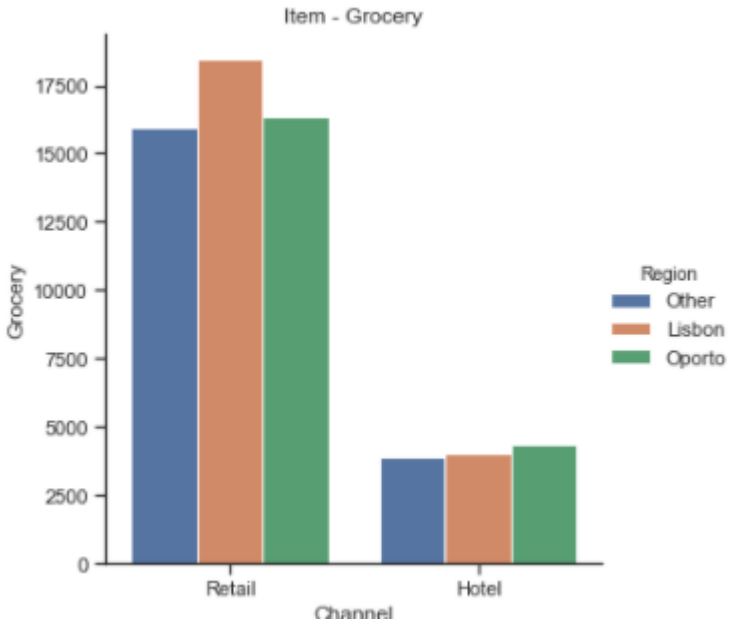
From the above plot, we can conclude that the spending of Item – Milk is higher in Retail than the same in Hotel. It can be interpreted that the Milk is essential for the retailers and they consume it in daily basis. When it comes for the sales, a large population of the different region purchases milk for their household. It is expected the spending in Retail channel is significantly higher than in Hotel channel.

Item Milk v/s Region:



From the above plot, it is evident that the spending of Item Milk in all the region is higher.

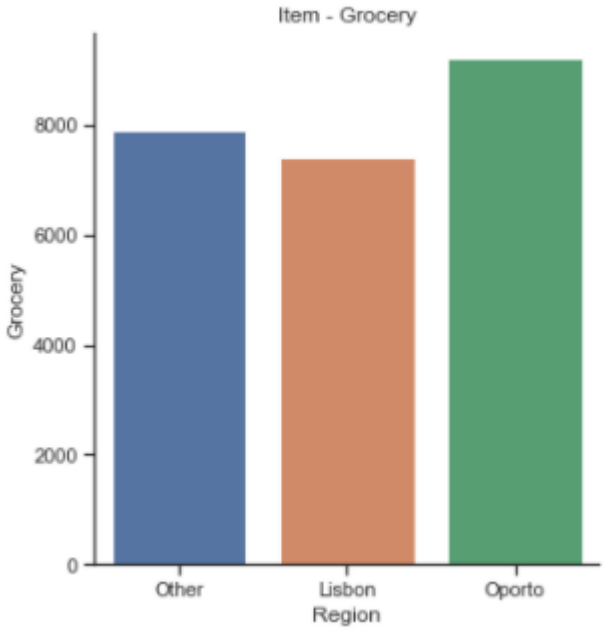
Item – Grocery v/s Channel:



From the above plot, it is observed that the spending on Grocery in Retail is much higher than in Hotel channel.

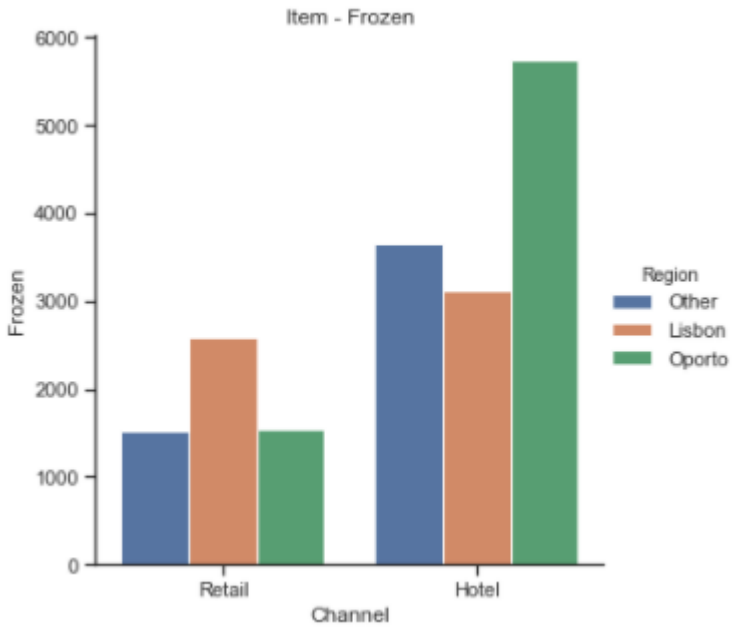
Spending of Grocery in Lisbon region in Retail is the highest whereas the spending in region Oporto in Hotel is the highest.

Item – Grocery v/s Region:



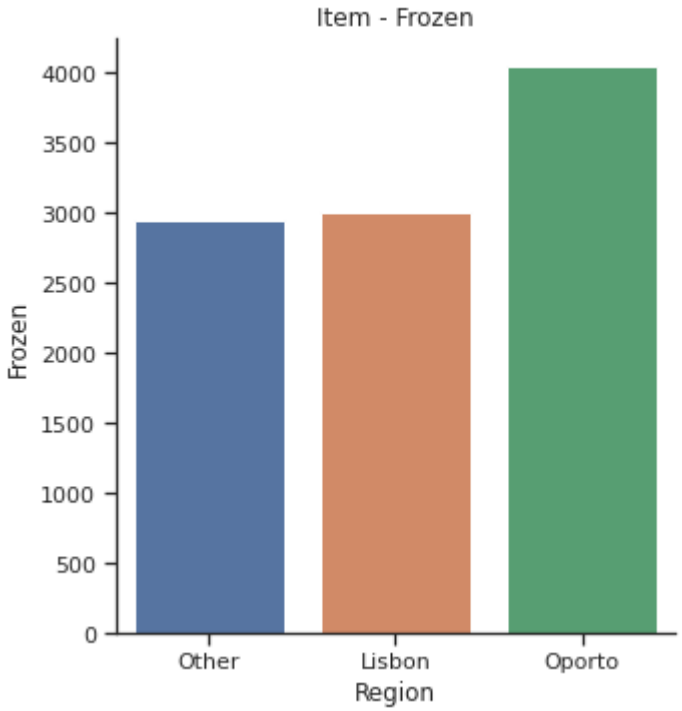
The spending of Groceries in Oporto region is observed to be higher when compared with Lisbon and Other region.

Item – Frozen v/s Channel:



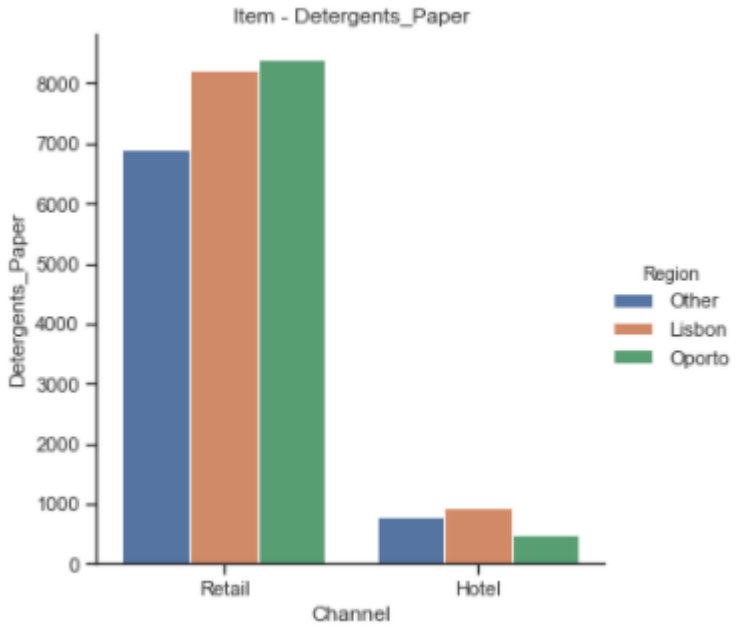
From the above plot, it is observed that the spending on frozen item sold more in Hotel region.

Item – Frozen v/s Region:



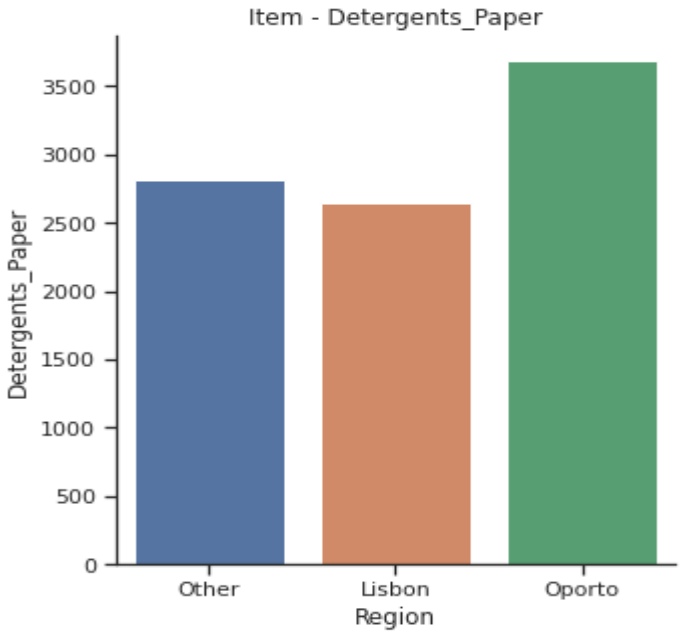
Based on the above plot, It is observed that the spending on Frozen Items in Oporto is higher than the spending on Frozen items in the Other and Lisbon regions.

Item – Detergents Paper vs Channel:



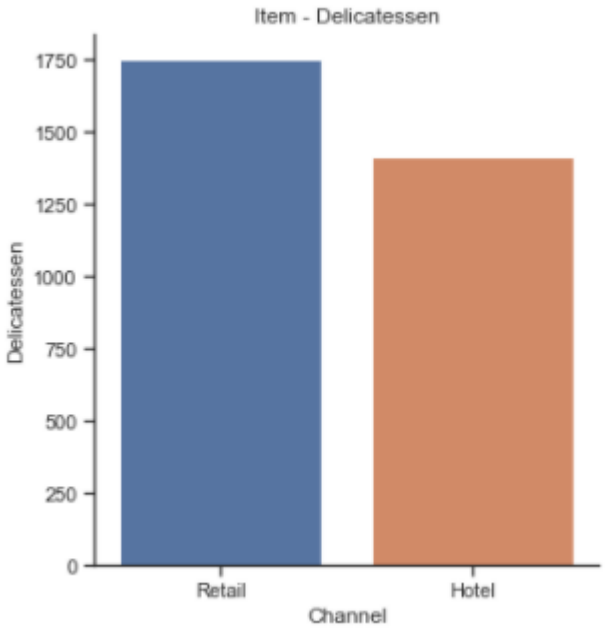
Based on the above plot, it is observed that the spending on Detergents Paper in Retail channel is more than the spending on Detergents Paper in Hotel channel.

Item – Detergents Paper v/s Region:



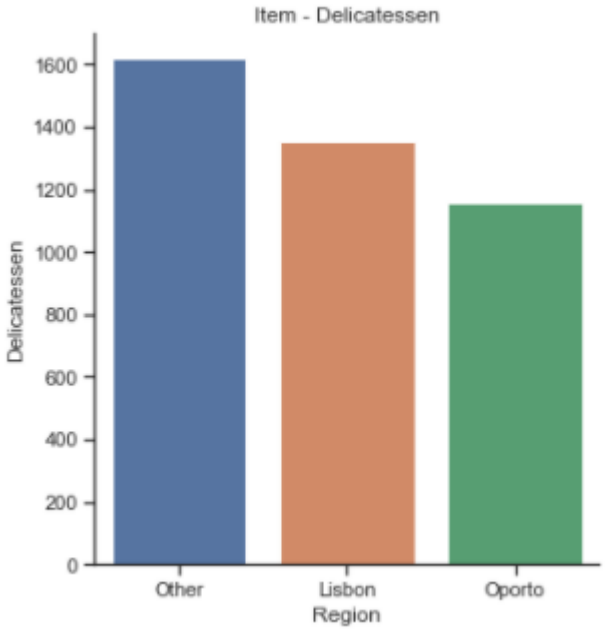
Based on the above plot, it is observed that the spending on Detergents Paper in Oporto region is higher than the spending on the same in other and Lisbon region.

Item – Delicatessen v/s Channel



Based on the above plot, it is observed that the spending on Delicatessen is more in Retail channel than the spending in Hotel channel

Item – Delicatessen v/s Region:



Based on the above plot, it is observed that the spending on Delicatessen items is the highest in Other region when compared with spending on the other two regions.

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?

Standard deviation is the best parameter to define consistent and inconsistent behaviour for the data set or each column. Let us determine the standard deviation of the items and observe

```
Fresh               12647.33
Milk                 7380.38
Grocery              9503.16
Frozen               4854.67
Detergents_Paper     4767.85
Delicatessen         2820.11
dtype: float64
```

Coefficient of Variation for Fresh is 1.0539179237473144

Coefficient of Variation for Milk is 1.2732985840065412

Coefficient of Variation for Grocery is 1.193815447749267

Coefficient of Variation for Frozen is 1.5785355298607762

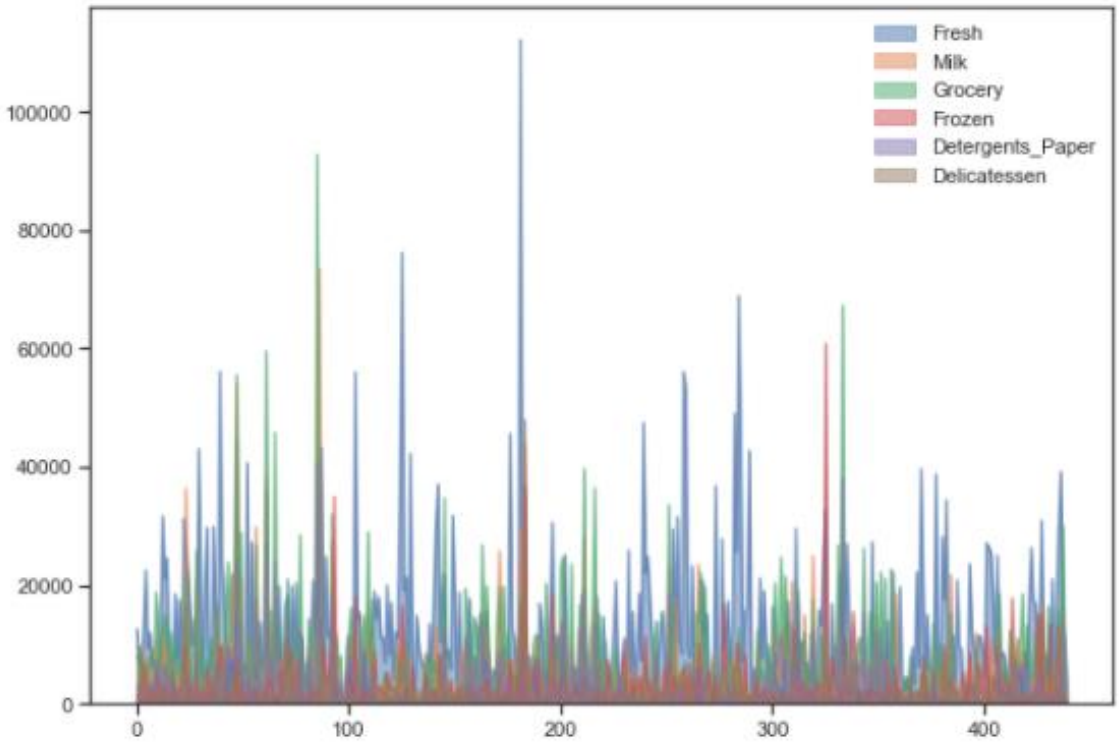
Coefficient of Variation for Detergents_Paper is 1.6527657881041729

Coefficient of Variation for Delicatessen is 1.8473041039189306

Variance:

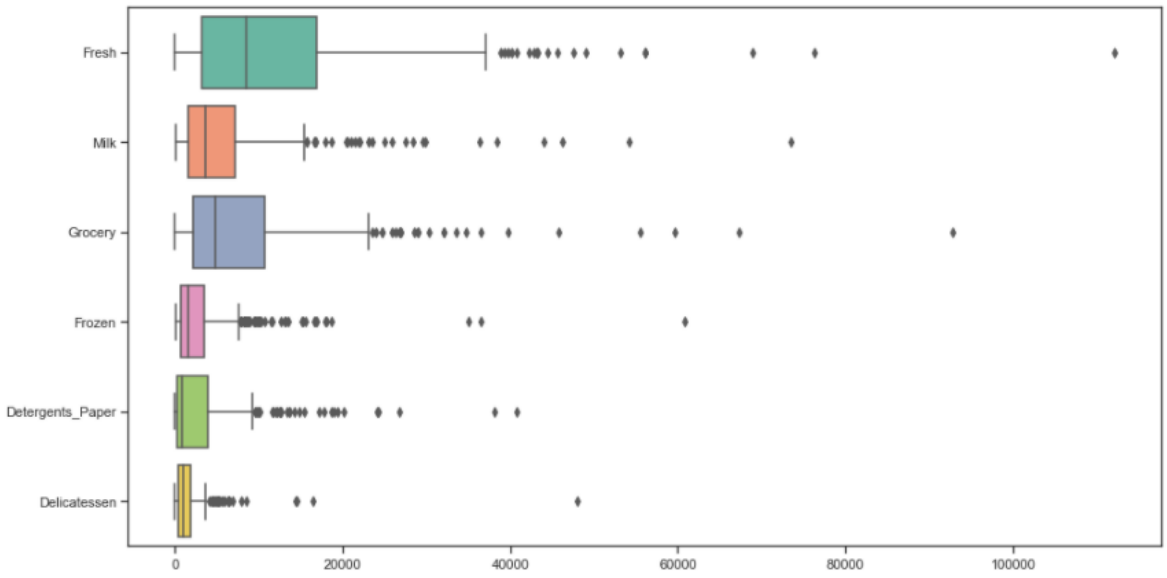
```
Fresh               1.599549e+08
Milk                5.446997e+07
Grocery             9.031010e+07
Frozen              2.356785e+07
Detergents_Paper    2.273244e+07
Delicatessen        7.952997e+06
```

From the above calculation, it is observed that Delicatessen items have smallest Standard deviation, so that is consistent. Fresh items have highest Standard deviation so that is Inconsistent.



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

Box plots are useful as they show outliers within a data set. An outlier is an observation that is numerically distant from the rest of the data. When reviewing a box plot, an outlier is defined as a data point that is located outside the whiskers of the box plot. These "too far away" points are called "outliers", because they lie outside the range. An outlier is any value that lies more than one and a half times the length of the box from either end of the box.



As we can see there are each verity has more than 1 outliers. There are outliers in all the items across the product range (Fresh, Milk, Grocery, Frozen, Detergents_Paper & Delicatessen).

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.

- As per the analysis, I find out that there are inconsistencies in spending of different items (by calculating Coefficient of Variation), which should be minimized.

- The spending of Hotel and Retail channel are different which should be more or less equal. And also spent should be equal for different regions.

- Not just “Fresh” and “Grocery”, but need focus on other Items as well.

Problem 2
CMSU Survey Data Analysis

INTRODUCTION

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 theSurvey.csvfile).

SUMMARY

This business report provides detailed explanation of approach to each problem given in the assignment and provides relative information with regards to solving the problem.

QUESTIONS:

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

2.1.1. Gender and Major:

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

2.1.2. Gender and Grad Intention:

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

2.1.3. Gender and Employment:

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

2.1.4. Gender and Computer:

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

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?

Solution:

The total number of students:

```
Female    33
Male      29
Name: Gender, dtype: int64
```

The total number of students = 33 + 29 = 62

Number of Male students = 29

Probability = Number of Male students / Total number of students

After calculation we got the result that probability of 46.77% student will be male in CMSU if randomly selected

```
#Probability of the randomly selected CMSU student will be male

print('The probability of the randomly selected CMSU student will be Male {} percent'.format(29/(29+33)*100))
```

```
The probability of the randomly selected CMSU student will be Male 46.774193548387096 percent
```

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

Solution:

The total number of students:

```
Female     33
Male       29
Name: Gender, dtype: int64
```

The total number of students = 33 + 29 = 62

Number of Female students = 33

Probability = Number of Female students / Total number of students

After calculation we got the result that probability of 53.23% student will be female in CMSU if randomly selected

```
#Probability of the randomly selected CMSU student will be Female

print('The probability of the randomly selected CMSU student will be Female {} percent'.format(33/(29+33)*100))
```

```
The probability of the randomly selected CMSU student will be male 53.2258064516129 percent
```

2.3. Assume that the sample is a 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.

Using contingency tables of Gender and Majors we got the total numbers of males and number of males opting for different majors. Below is the screen shot of the Python Code and Output.

```
#2.2.2. Find the conditional probability of different majors among the male students in CMSU.Find the conditional probability of d

# From Gender and major contingency table:

print("Probability percentage of Males opting for Accounting is {:.2f} %".format(4/29*100))
print("Probability percentage of Males opting for CIS is {:.2f} %".format(1/29*100))
print("Probability percentage of Males opting for Economics/Finance is {:.2f} %".format(4/29*100))
print("Probability percentage of Males opting for International Business is {:.2f} %".format(2/29*100))
print("Probability percentage of Males opting for Management is {:.2f} %".format(6/29*100))
print("Probability percentage of Males opting for Other is {:.2f} %".format(4/29*100))
print("Probability percentage of Males opting for Retailing/Marketing is {:.2f} %".format(5/29*100))
print("Probability percentage of Males opting for Undecided is {:.2f} %".format(3/29*100))
```

Output:

```
Probability percentage of Males opting for Accounting is 13.79 %
Probability percentage of Males opting for CIS is 3.45 %
Probability percentage of Males opting for Economics/Finance is 13.79 %
Probability percentage of Males opting for International Business is 6.
90 %
Probability percentage of Males opting for Management is 20.69 %
Probability percentage of Males opting for Other is 13.79 %
Probability percentage of Males opting for Retailing/Marketing is 17.24
 %
Probability percentage of Males opting for Undecided is 10.34 %
```

From the above output, it can be interpreted that most of the Males students prefer Management as Majors and CIS is the least preferred.

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

Using contingency tables of Gender and Majors we got the total numbers of males and number of Females opting for different majors.

Below is the screen shot of the python code and Output.

```
#Find the conditional probability of different majors among the female students in CMSU.

# From Gender and major contingency table:

print("Probability percentage of Females opting for Accounting is {:.2f} %".format(3/33*100))
print("Probability percentage of Females opting for CIS is {:.2f} %".format(3/33*100))
print("Probability percentage of Females opting for Economics/Finance is {:.2f} %".format(7/33*100))
print("Probability percentage of Females opting for International Business is {:.2f} %".format(4/33*100))
print("Probability percentage of Females opting for Management is {:.2f} %".format(4/33*100))
print("Probability percentage of Females opting for Other is {:.2f} %".format(3/33*100))
print("Probability percentage of Females opting for Retailing/Marketing is {:.2f} %".format(9/33*100))
print("Probability percentage of Females opting for Undecided is {:.2f} %".format(0/33*100))
```

Output:

```
Probability percentage of Females opting for Accounting is 9.09 %
Probability percentage of Females opting for CIS is 9.09 %
Probability percentage of Females opting for Economics/Finance is 21.21
 %
Probability percentage of Females opting for International Business is
12.12 %
Probability percentage of Females opting for Management is 12.12 %
Probability percentage of Females opting for Other is 9.09 %
Probability percentage of Females opting for Retailing/Marketing is 27.
27 %
Probability percentage of Females opting for Undecided is 0.00 %
```

From the above output, it is observed that female students prefer Retailing/Marketing as Majors.

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 conditional probability of intent to graduate, given that the student is a male.

Using contingency tables of Gender and Grad Intention, we get the total numbers of males and number of males intends to be graduate.

Below is the screen shot of the python code and Output.

```
#Find the conditional probability of intent to graduate, given that the student is a male.
# Using contingency tables of Gender and Grad Intention we got the total numbers of males and number of males intends to be gradu

print("Probability of intent to graduate, given that the student is a male {:.2f} %".format(17/29*100))
```

Output:

```
Probability of intent to graduate, given that the student is a male 58.
62 %
```

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

Using contingency tables of Gender and Computer, we get the total numbers of females and number of males prefer laptop.

Below is the screenshot of python code and Output

```
# 2.4.2 Find the probability that a randomly selected student is a female and does NOT have a laptop.
#Using contingency tables of Gender and Grad Computer, we get the total numbers of females and number of males prefer laptop.

print("The probability that a randomly selected student is a female and does NOT have a laptop. {:.2f} %".format(4/29*100))
```

```
The probability that a randomly selected student is a female and does NOT have a laptop. 13.79 %
```

Output:

```
The probability that a randomly selected student is a female and does N
OT have a laptop. 13.79 %
```

2.5. Assume that the sample is a 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 a full-time employment

Using contingency tables of Gender and Employment, we get the total numbers of males and number of males who are full time employed

And post calculation we find out that - Probability of randomly chosen student is either Male or has full time employment is 74.19%

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

Using contingency tables of Gender and Major we got the total numbers of females and number of females majoring in international business or management.

Upon calculation we find out that - Probability that given a female student is randomly chosen, she is majoring in international business or management is 24.24%

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?

Solution:

Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

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

The Probability that a randomly selected student the graduate intention and being female

P(Grad Intention Yes) = 28/40 = 0.7
P(Grad Intention Yes | female) = 11 / 20 = 0.55

These probabilities are not equal. This suggests that the two events are independent.

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?

Solution:

Using contingency tables of Gender and GPA we got the total numbers of students and number of students GPA less than 3

Upon calculation we find out that - Probability that student is chosen randomly and that his/her GPA is less than 3 is 27.42 %

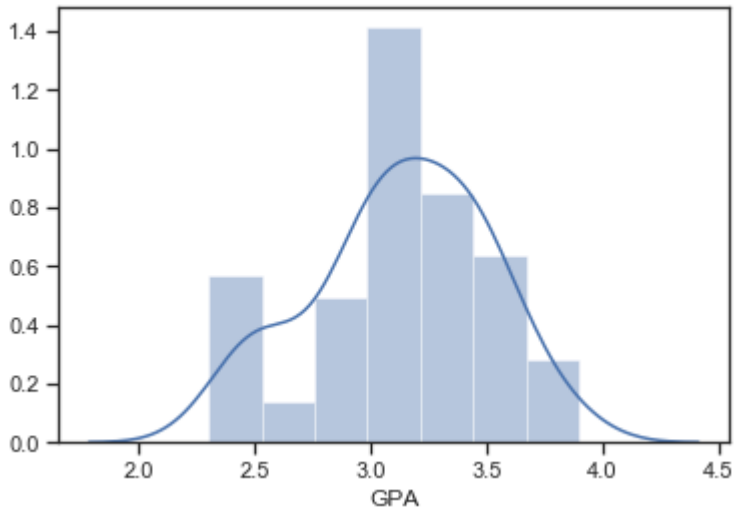
2.7.2 Find the conditional probability that a randomly selected male earns 50 or more. Find the conditional probability that a randomly selected female earns 50 or more.

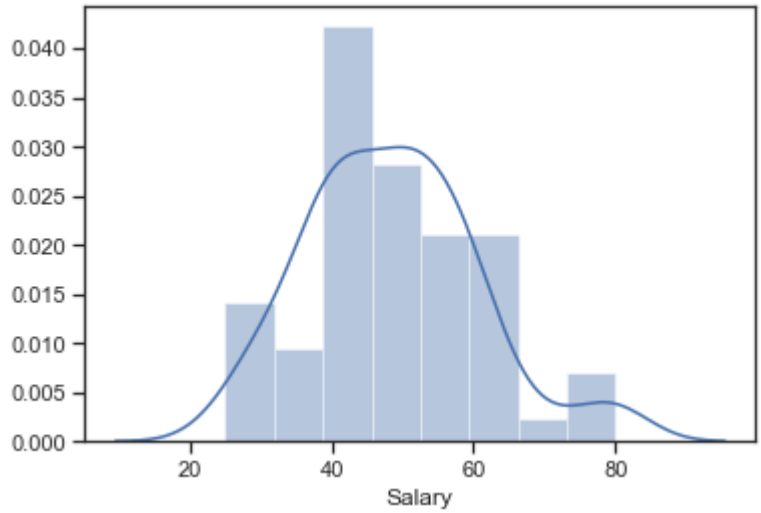
Using contingency tables of Gender and Salary we got the total numbers of Male and Female and number of male and female earning 50 or more

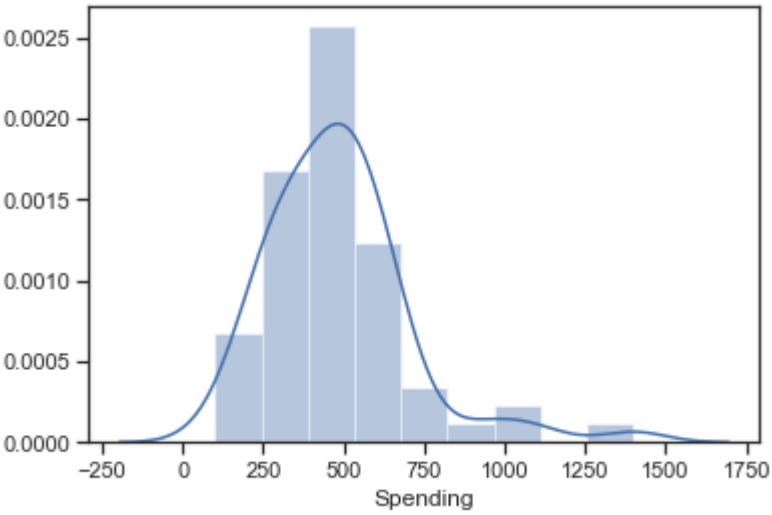
Post calculation we find out that - Probability that randomly selected male earns 50 or more is 22.58%
And Probability that randomly selected female earns 50 or more is 29.03%.

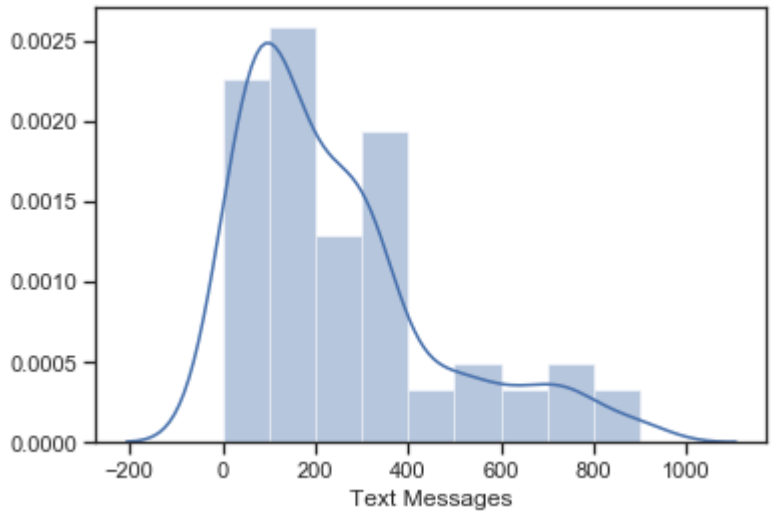
2.8.1 Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending and Text Messages. For each of them comment whether they follow a normal distribution.

Used distplot to know the normal distribution of these four numerical (continuous) variables in the data set – GPA, Salary, Spending and Text Messages









To confirm whether these four data sets are following normal distribution or not, we done the Shapiro Wilk test and the output from Python we got –

ShapiroResult(statistic=0.9685361981391907, pvalue= 0.11204058676958084)
ShapiroResult(statistic= 0.9565856456756592, pvalue=0.028000956401228905)
ShapiroResult(statistic=0.8777452111244202, pvalue=1.6854661225806922e-05)
ShapiroResult(statistic=0.8594191074371338, pvalue=4.324040673964191e-06)

By these details we confirm that out of the given four data sets 'GPA' and 'Salary' are following normal distribution whereas other two 'Spending' and 'Text Messages' are not following the normal distribution

2.8.2 Write a note summarizing your conclusions

We tested various probabilities among the different variables to determine very keen insights. Using the contingency table we derived the probability in terms of genders. It is observed that the graduate intent and gender are two independent events. From the distribution plot, we checked that 'GPA' and 'Salary' are following normal distribution whereas other two 'Spending' and 'Text Messages' are not following the normal distribution.

Problem 3

INTRODUCTION:

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 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.

SUMMARY:

This business report provides detailed explanation of approach to each problem given in the assignment and provides relative information with regards to solving the problem.

3 – Asphalt Shingles Data Analysis

We imported the 'A+&+B+shingles.csv' dataset in python to analyze the data about the Asphalt Shingles. Below is the detailed approach and answer.

3.1 Do you think there is evidence that mean moisture contents in both types of shingles are within the permissible limits? State your conclusions clearly showing all steps.

Since p-value > 0.05, do not reject H_0. There is not enough evidence to conclude that the mean moisture content for Sample A shingles is less than 0.35 pounds per 100 square feet. p-value = 0.0748.

If the population mean moisture content is in fact no less than 0.35 pounds per 100 square feet, the probability of observing a sample of 36 shingles that will result in a sample mean moisture content of 0.3167 pounds per 100 square feet or less is 0.0748.

Output from Python:

```
t_statistic, p_value = ttest_1samp(pp.A, 0.35)
print('One sample t test \nt statistic: {0} p value: {1} '.format(t_statistic, p_value/2))
```

```
One sample t test
t statistic: -1.4735046253382782 p value: 0.07477633144907513
```

Since p-value < 0.05, reject H_0 . There is enough evidence to conclude that the mean moisture content for Sample B shingles is not less than 0.35 pounds per 100 square feet. p-value = 0.0021.

If the population mean moisture content is in fact no less than 0.35 pounds / 100 square feet, the probability of observing a sample of 31 shingles that will result in a sample mean moisture content of 0.2735 pounds per 100 square feet or less is 0.0021.

Output from Python:

```
t_statistic, p_value = ttest_1samp(pp.B, 0.35,nan_policy='omit' )
print('One sample t test \nt statistic: {0} p value: {1} '.format(t_statistic, p_value/2))
```

```
One sample t test
t statistic: -3.1003313069986995 p value: 0.0020904774003191826
```

3.2 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?

H_0 : μ(A)= μ(B)

H_a : μ(A)!= μ(B)

α = 0.05

As the p-value > α , do not reject H0; and we can say that population mean for shingles A and B are equal Test Assumptions When running a two-sample t-test, the basic assumptions are that the distributions of the two populations are normal, and that the variances of the two distributions are the same. If those assumptions are not likely to be met, another testing procedure could be use.

```
t_statistic,p_value=scipy.stats.ttest_ind(pp['A'],pp['B'],equal_var=True ,nan_policy='omit')
print("t_statistic={} and pvalue={}".format(round(t_statistic,3),round(p_value,3)))
```

```
t_statistic=1.29 and pvalue=0.202
```

Output:

t_statistic=1.29 and pvalue=0.202

--*-*-*-*-*-*-*