SMDM Project – Business Report

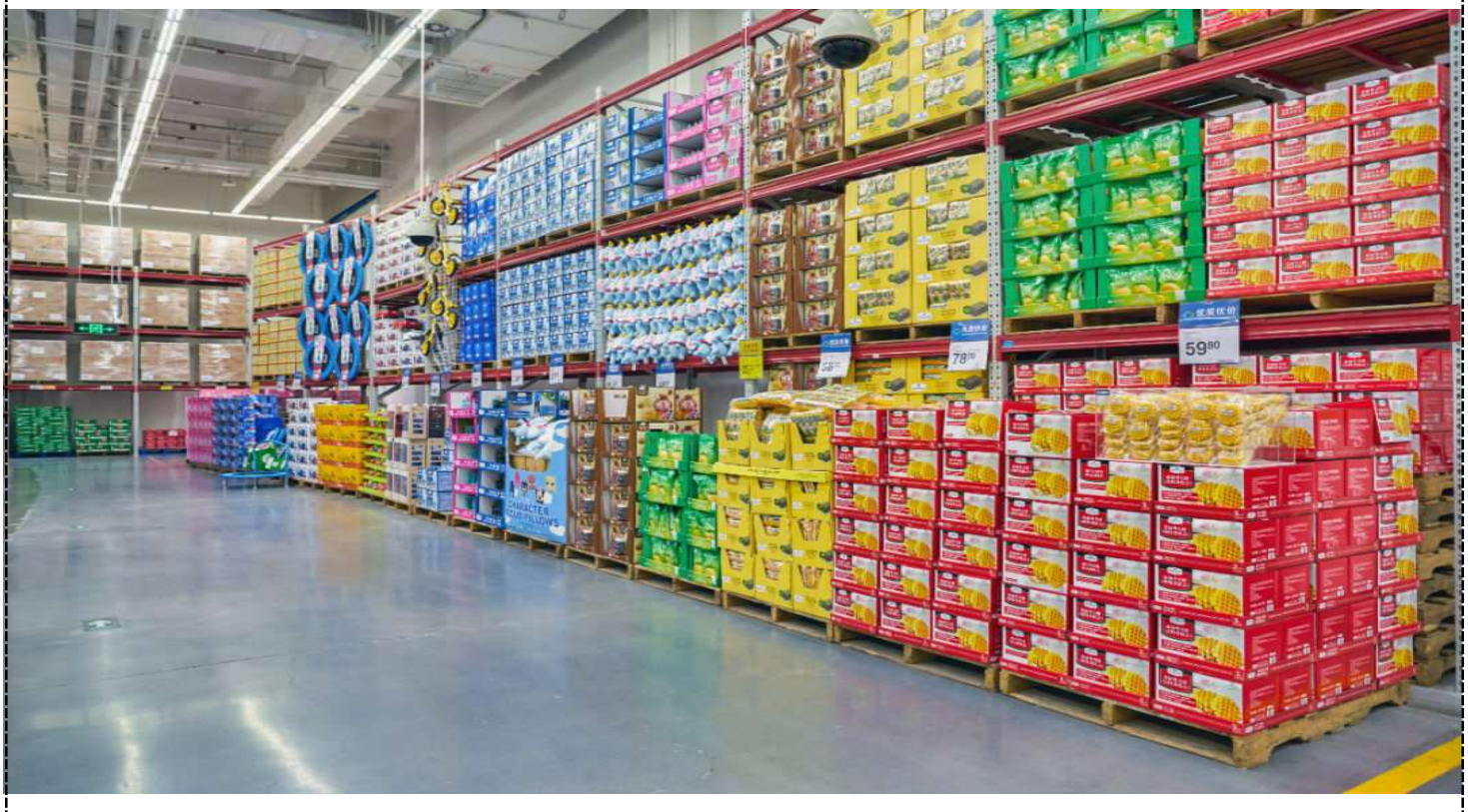
DSBA Program

February 2022

Submitted by:

N. Aishwarya

Contents Outline



Business problem 1: Wholesale Customers Analysis

Problem Statement:

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).

Summary:

This business report provides detailed explanation of approach to each problem and provides insights and relevant recommendations, based on the analysis.

As a first step, the base dataset on 'Wholesale Customer data' was imported in python to understand the problem in depth. Descriptive statistics provided insights across different dimensions.

1.1. Descriptive statistics to summarize data:

This data set has 440 rows and 9 columns. The data set refers to clients of a wholesale distributor. It includes the annual spending in monetary units (m.u.) on diverse product categories.

Description of variables are as below, to understand the data better:

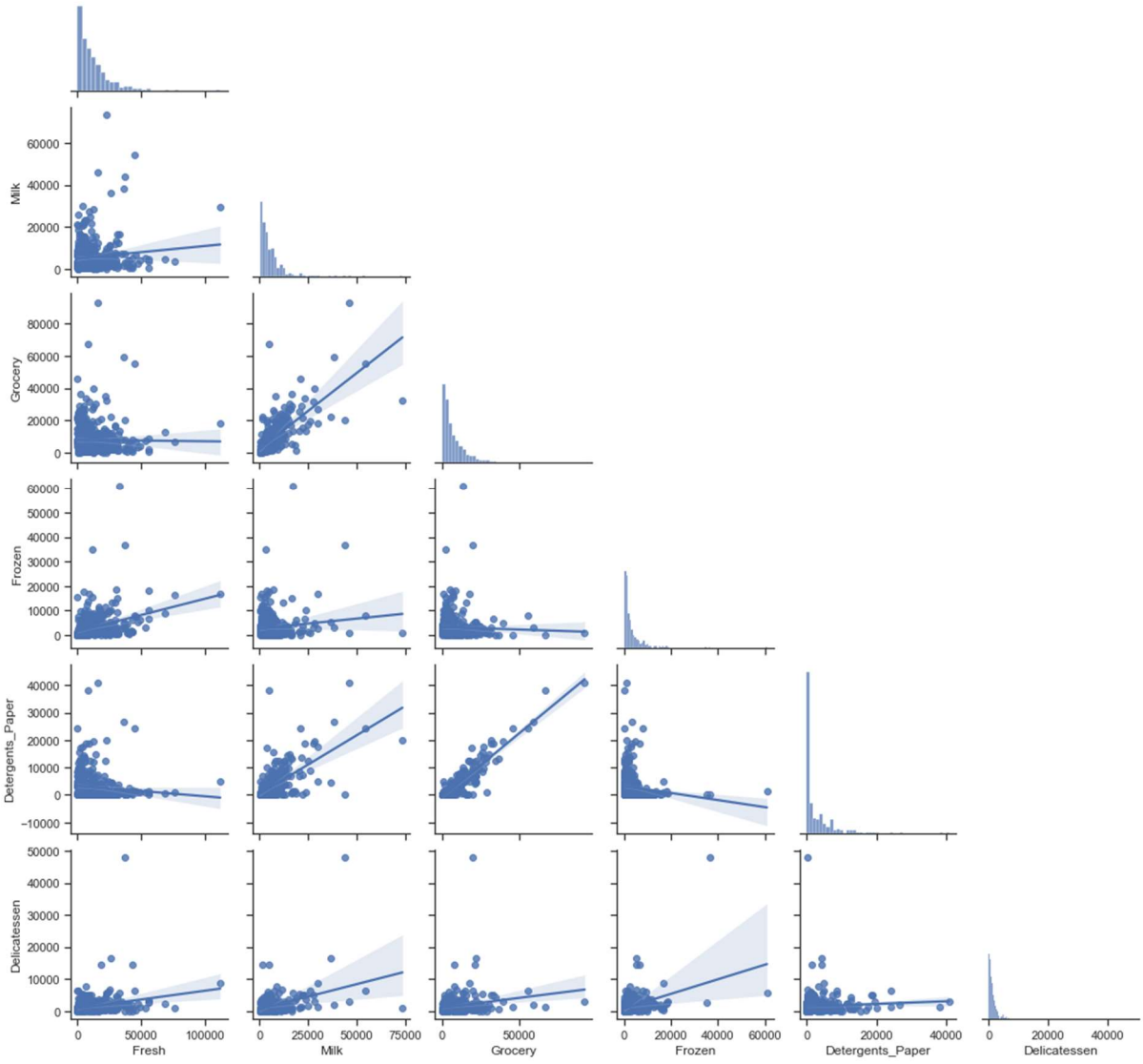
- **FRESH**: annual spending (m.u.) on fresh products (Continuous);
- **MILK:** annual spending (m.u.) on milk products (Continuous);
- **GROCERY:** annual spending (m.u.) on grocery products (Continuous);
- **FROZEN:** annual spending (m.u.) on frozen products (Continuous);
- **DETERGENTS_PAPER:** annual spending (m.u.) on detergents and paper products (Continuous);
- **DELICATESSEN:** annual spending (m.u.) on and delicatessen products (Continuous);
- **CHANNEL:** customers Channel - Hotel (Hotel/Restaurant/Cafe) or Retail channel (Nominal);
- **REGION:** customers Region Lisbon, Oporto or Other (Nominal);
- **BUYER/SPENDER:** it is showing running id number (assumption it is index) (Continuous);

Below is the snapshot of the Wholesale customer data:

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

Below is the summary of the data, providing descriptive statistical variables:

	count	mean	std	min	25%	50%	75%	max
Buyer/Spender	440.0	220.500000	127.161315	1.0	110.75	220.5	330.25	440.0
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



From the pair plot above, the correlation between the "detergents and paper products" and the "grocery products" seems to be pretty strong, meaning that consumers would often spe nd money on these two types of products.

1.1. 1. Which Region and which Channel spent the most? Which Region and which channel spent the least?

Solution:

Using describe function in python, we first looked at the basic descriptive statistics of the data set.

1. **Region specific spending:**

Below are the Python outputs that helps to derive the above spending pattern.

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

- **Highest spend in the Region is from Others and lowest spend in the region is from Oporto.**

2. **Channel specific spending:**

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

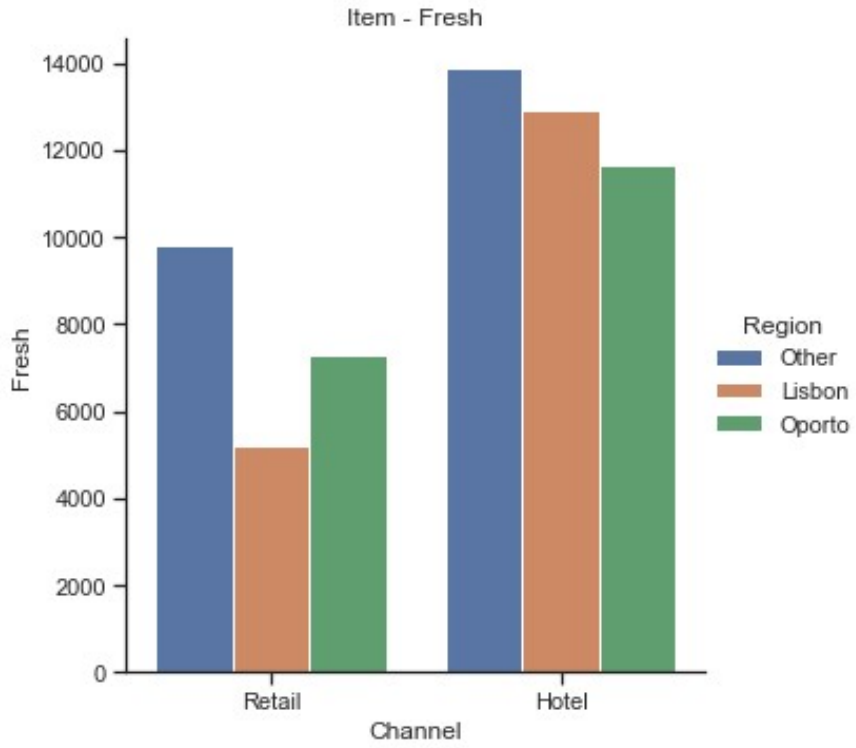
- **Highest spend in the Channel is from Hotel and lowest spend in the Channel is from Retail**

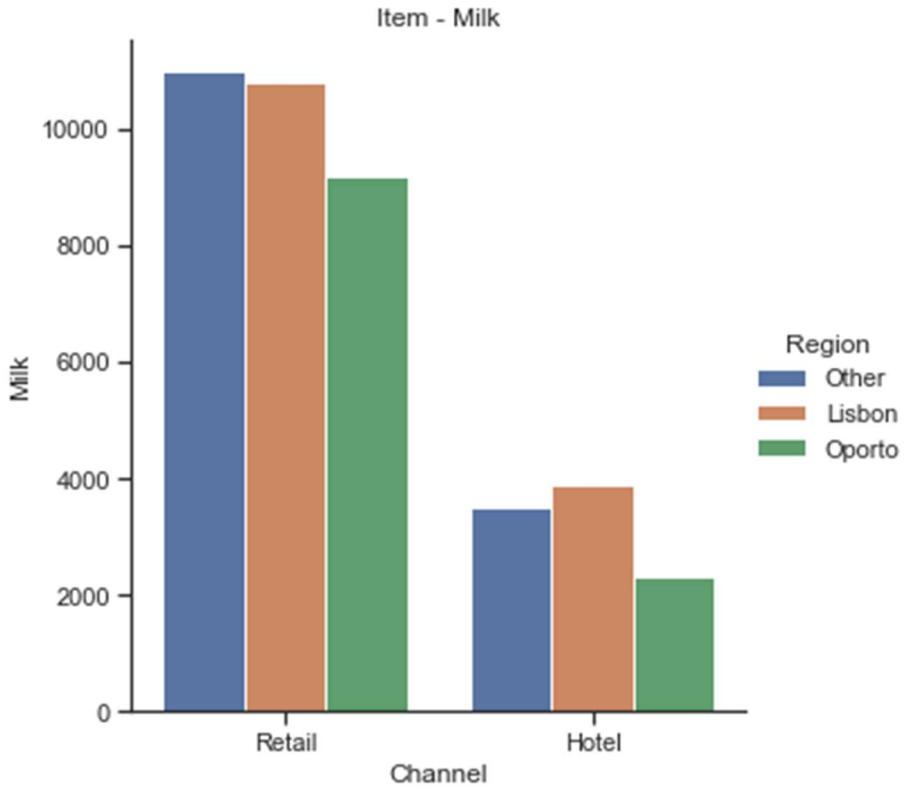
Region with Channel spending:

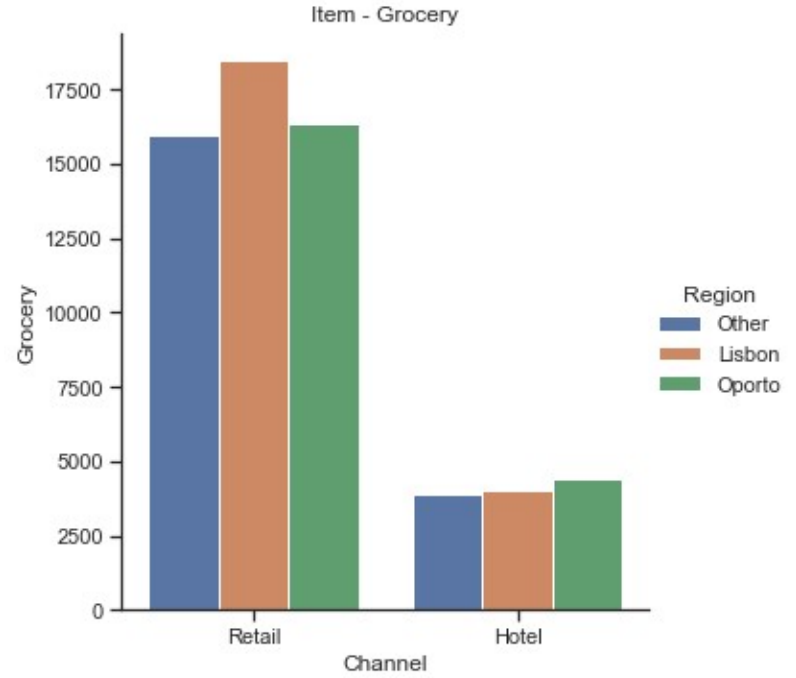
```
Region  Channel
Lisbon  Hotel      1538342
        Retail      848471
Oporto  Hotel       719150
        Retail      835938
Other   Hotel      5742077
        Retail     4935522
Name: Spending, dtype: int64
```

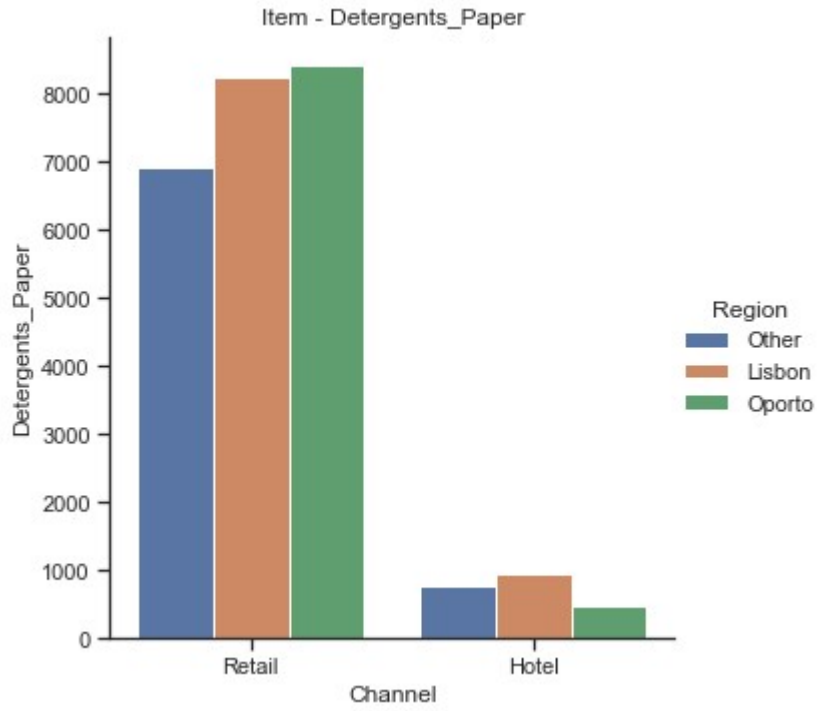
- **Key inference: Highest spend in the Region/Channel is from Others/Hotel and lowest spend in the Region/Channel is from Oporto/Hotel**

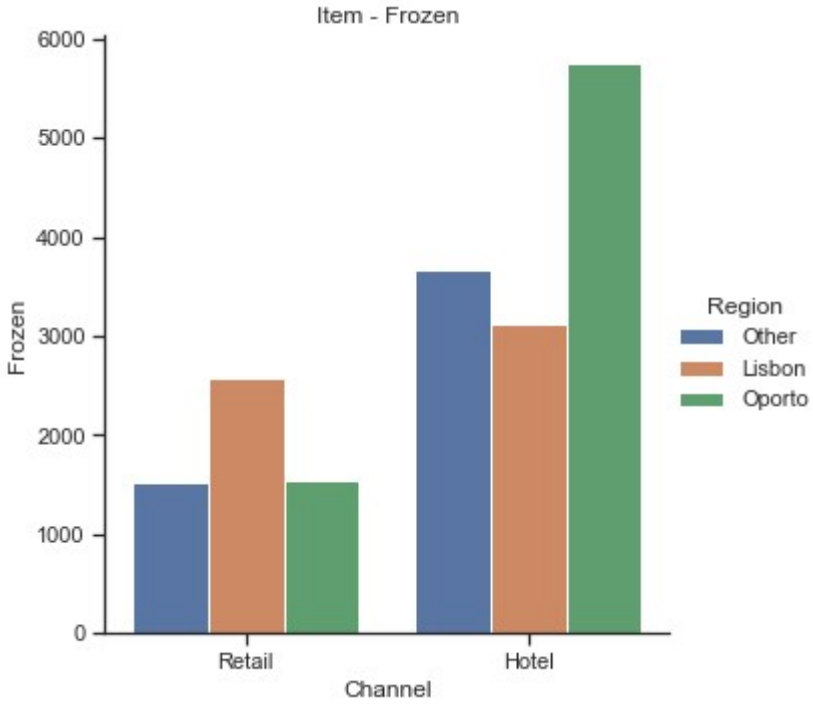
1.2. There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.

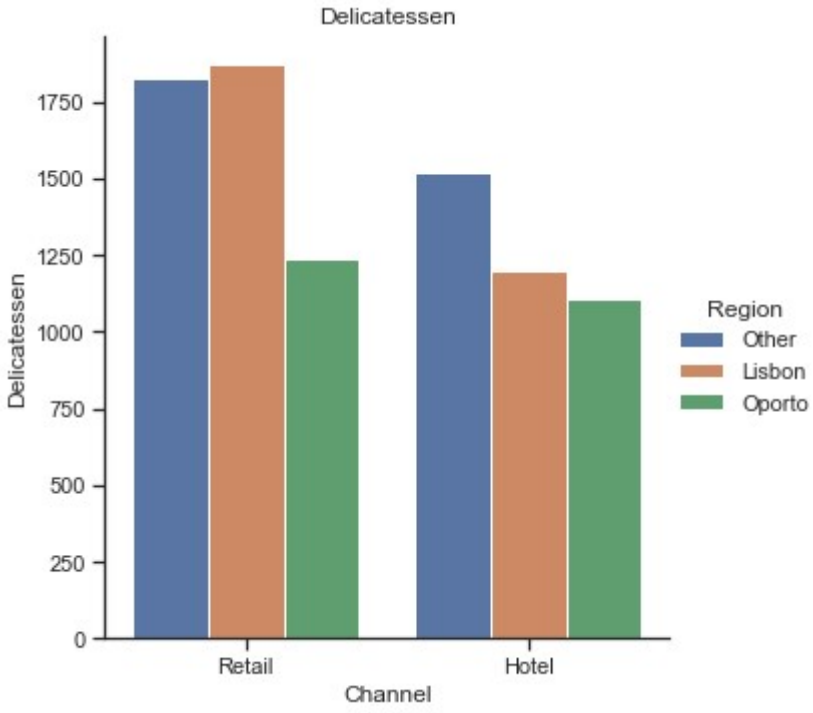





Item - Grocery
Grocery
17500
15000
12500
10000
7500
5000
2500
0
Retail
Hotel
Channel
Region
Other
Lisbon
Oporto


Item - Detergents_Paper
Detergents_Paper
8000
7000
6000
5000
4000
3000
2000
1000
0
Retail
Hotel
Channel
Region
Other
Lisbon
Oporto





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?

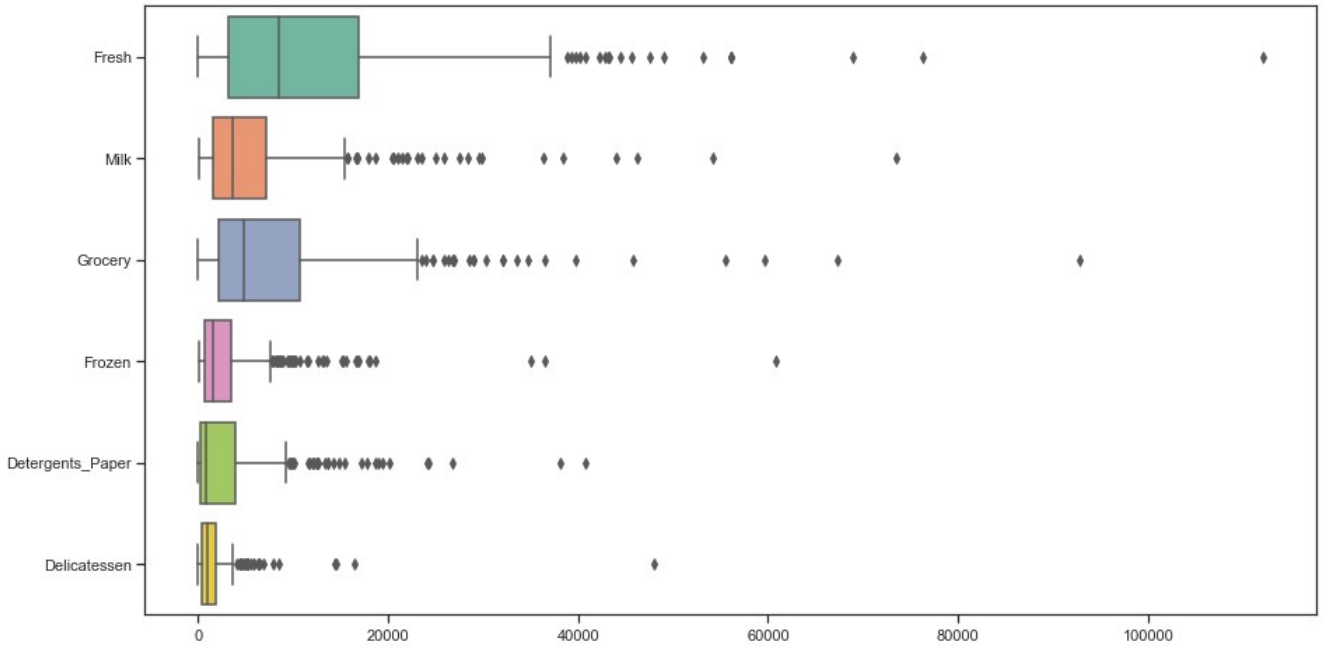
By using the metric Coefficient of Variation, the least value is of Category “Fresh” (1.05) and highest value is of Category “Delicatessen” (1.84). Hence, the most inconsistent behaviour shown by item – Delicatessen and least inconsistent behaviour is shown by item – Fresh

Below is the output from Python for Coefficient of Variation for:

- Fresh is 1.05
- Milk is 1.27
- Frozen is 1.58
- Grocery is 1.20
- Detergents_Paper is 1.65
- Delicatessen is 1.85

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

Boxplot below, provides the output regarding the outliers:



There are outliers across all the product range - Fresh, Milk, Grocery, Frozen, Detergents_Paper & Delicatessen. It seems to be more pronounced in Fresh, Milk and Grocery categories.

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?

As per the analysis, there are inconsistencies in spending of different items (based on Coefficient of Variation), which should be minimized. The spending of Hotel and Retail channels are different, which should be ideally in the same range. And also spent should equal for different regions. Need to focus on other items also than “Fresh” and “Grocery”

Business problem 2: Clear Mountain State University (CMSU) Survey

Problem Statement:

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.

Solution:

Imported the 'CMSU Survey' dataset in python to analyze the data about the undergraduate students who attend CMSU. Below is the detailed approach and analysis.

2.1. For this data, construct the following contingency tables

2.1.1. Gender and Major

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

2.1.2. Gender and Grad Intention

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

2.1.3. Gender and Employment

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

2.1.4. Gender and Computer

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

2.2. Assume that the sample is a representative of the population of CMSU. Based on the data, below are the observations:

2.2.1. What is the probability that a randomly selected CMSU student will be male?

46.8% of the student population, will be male in CMSU if randomly selected. This is based on the ratio of total male students out of total students from the given data.

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

53.2% of the student population, will be female in CMSU if randomly selected. This is based on the ratio of total female students out of total students from the given data.

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

Male category students prefer Management as Majors and CIS is the least preferred one. Below is the Python output, which validates the above statement:

- **P(Accounting|Male)=13.79%**
- **P(CIS|Male)=3.45%**
- **P(Economics/Finance|Male)=13.79%**
- **P(International Business|Male)=6.9%**
- **P(Management|Male)=20.69%**
- **P(Other|Male)=13.79%**
- **P(Retailing/Marketing|Male)=17.24%**
- **P(Undecided|Male)=10.34%**

2.3.2. Conditional probability of different majors among the female students of CMSU.

Female category students prefer Retailing/Marketing as Majors and Accounting, CIS being the least preferred one. Below is the Python output, which validates the above statement:

- **P(Accounting|Female)=9.09%**
- **P(CIS|Female)=9.09%**
- **P(Economics/Finance|Female)=21.21%**
- **P(International Business|Female)=12.12%**
- **P(Management|Female)=12.12%**
- **P(Other|Female)=9.09%**
- **P(Retailing/Marketing|Female)=27.27%**
- **P(Undecided|Female)=0.0%**

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

Probability of Males and intends to be Graduate is 58.62%. This is derived based on the contingency tables of Gender and Grad Intention.

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

12.1% of selected student is a Female and does NOT have a laptop. This is using contingency tables of Gender and Computer we got the total numbers of females and number of females does not have a laptop.

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

51.6% of selected student is a male and has full time employment. This is derived based on the below approach:

P(Male)=29/62

P(Full-Time Employment)=10/62

P(Male and Full-Time Employment)=7/62

P(Male or Full-Time Employment)=P(Male)+P(Full-Time Employment)-P(Male and Full-Time Employment) = 29/62+10/62-7/62 = 32/62

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

24.2% of selected student is a female and has been randomly chosen, she is majoring in international business or management. This is derived based on the below approach:

P(International_business or Management | Female) = ?

P(International_buinesss | Female) = 4/33

P(Management | Female) = 4/33

P(International_business or Management | Female) = P(International_buinesss | Female) + P(Management | Female) = 4/33 + 4/33 = 8/33

(Since choosing international business and Management are mutually exclusive events)

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.

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

Do you think graduate intention and being female are independent events?

To answer this question, we compare the probability that a randomly selected student intents to graduate with the probability that a randomly selected female student intends to graduate. If these two probabilities are the same (or very close), we say that the events are independent. In other words, independence means that being female does not affect the likelihood of having the intention to graduate.

To answer this question, we compare:

1. the unconditional probability: P (Intention to graduate)
2. the conditional probability: P (Intention to graduate | female)

If these probabilities are equal (or at least close to equal), then we can conclude that having the intention to graduate is independent of being a female. If the probabilities are substantially different, then we say the variables are dependent.

If P (A | B) = P(A), then the two events A and B are independent

2.7 Problem. 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?

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

And post 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 conditional probability that a randomly selected male earns 50 or more. Find 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

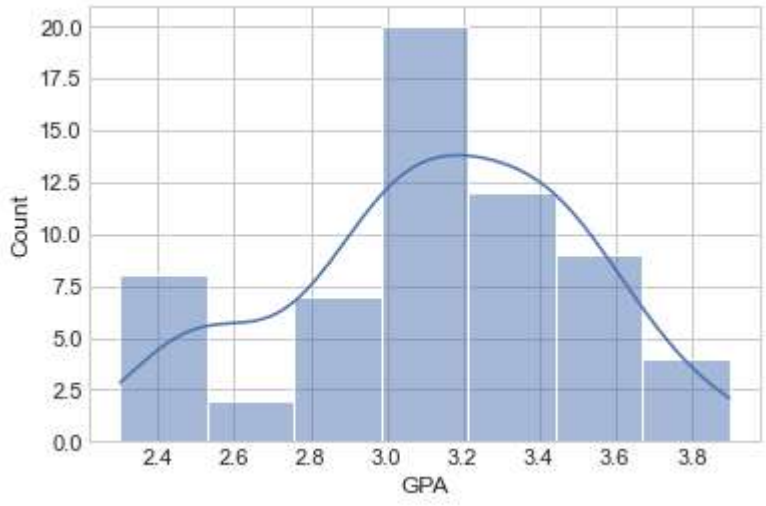
And basis the calculation, we can derive:

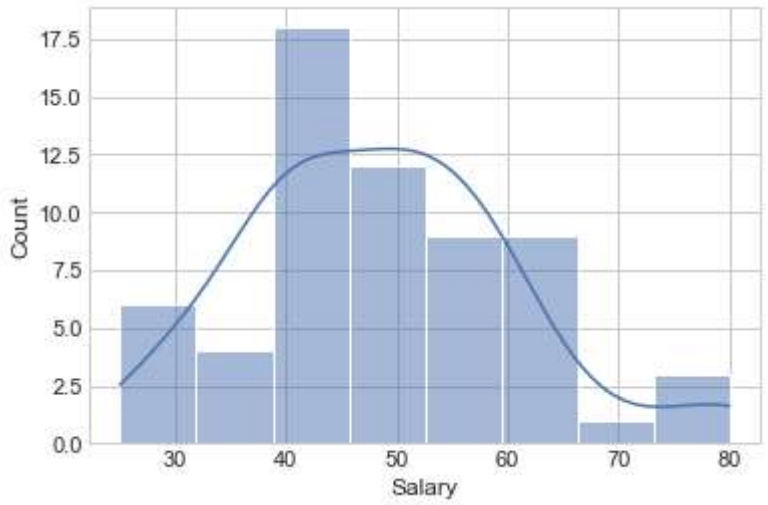
Probability that randomly selected male earns 50 or more is 48.28%

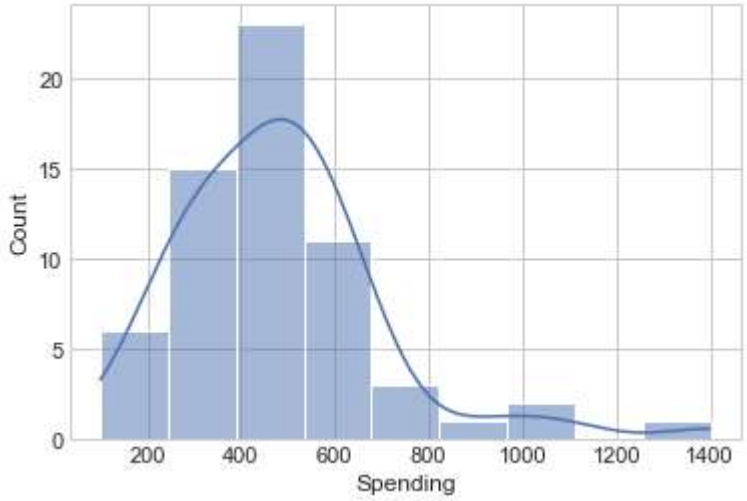
Probability that randomly selected female earns 50 or more is 54.55%

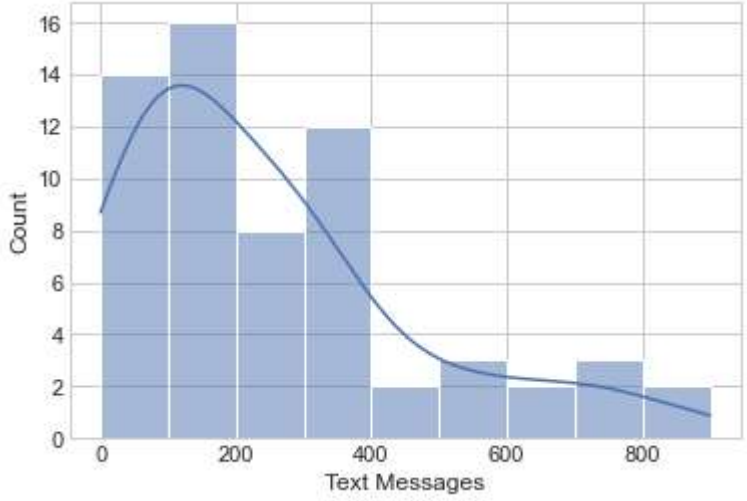
2.8. Note that there are three numerical (continuous) variables in the data set, Salary, Spending and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarizing your conclusions.

Based on the diagrams below, we can say that the variables 'GPA' and 'Salary' are normally distributed to an extent. Whereas 'Spending' and 'Text Messages' are right skewed and hence not normally distributed.




Count
Salary
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
30
40
50
60
70
80


Count
Spending
0
5
10
15
20
200
400
600
800
1000
1200
1400


Count
Text Messages
0
2
4
6
8
10
12
14
16
0
200
400
600
800

Business problem 3: Hypothesis Testing for Quality of Shingles

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 claims that the mean moisture content cannot be greater 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.

3.1 Do you think that the population means for shingles A and B are equal? Form the hypothesis and conduct the test of the hypothesis. What assumption do you need to check before the test for equality of means is performed? Write the hypothesis as in given format and paste the test results along with the interpretations from the results.

For the A shingles,

The null hypothesis states the population mean moisture content is less than 0.35 pound per 100 square feet.

The alternative hypothesis states that the population mean moisture content is greater than 0.35 pound per 100 square feet.

$H0 : \mu \leq 0.35$

$H1 : \mu > 0.35$

Input to the problem:

```
t_statistic, p_value = ttest_1samp(df3, 0.35)

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

Output of the problem:

One sample t test

t statistic: [-1.47350463 nan] p value: [0.07477633 nan]

Conclusion:

Since p value > 0.05, do not reject H0. 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 .0748

For the B shingles,

The null hypothesis states the population mean moisture content is less than 0.35 pound per 100 square feet.

The alternative hypothesis states that the population mean moisture content is greater than 0.35 pound per 100 square feet.

$H0 : \mu \leq 0.35$

$H1 : \mu > 0.35$

Input to the problem:

```
t_statistic, p_value = ttest_1samp(df3, 0.35,nan_policy='omit' )

print('One sample t test \nt statistic: {0} p value: {1} '.format(t_statistic, p_value))
```

Output of the problem:

```
One sample t test

t statistic: [-1.4735046253382782 -3.1003313069986995] p value: [0.07477633144907499
0.0020904774003191826]
```

Conclusion:

Since p value < 0.05, reject H0 . 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 0.35pounds per 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 .0021.

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

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.

Step 1: Define null and alternative hypotheses

In testing whether the population mean of shingles A and B are equal.

The Null Hypothesis: The population mean of both shingles A and B are equal. i.e $\mu A = \mu B$

The Alternative Hypothesis: The population mean of both shingles A and B are different. i.e. $\mu A \neq \mu B$

Step 2: Decide the significance level

Here we select α = 0.05 and the population standard deviation is not known.

Step 3: Calculate the p - value and test statistic

Two-sample t-test p-value= 0.20175

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

Conclusion:

As the p-value > α, do not reject H0; and we can say that population mean for shingles A and B are equal.