SMDM Project

By: THILAK RAJ

18 FEB 2022

Executive Summary (PROBLEM-1)

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).

Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. The data consists of 440 different channel details with 3 different regions called Lisbon, Oporto and Others. Analyze the expenditure of each of the channels. This assignment should help the student in exploring the summary statistics, contingency tables, conditional probabilities & hypothesis testing.

Data Description

1. Buyer/Spender: Serial number of the channel
2. Channel: Type of the channel (Retail or Hotel)
3. Region: Region of the channel (Lisbon, Oporto or Others).
4. Fresh, Milk, Grocery, Frozen, Detergents_Paper, Delicatessen: items in their stores.

Sample of the 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

Dataset has 9 variables with 2 Channels and 3 Regions. The data consists of 440 large retailers' annual spending on 6 different varieties of products in 3 different regions and across different sales channel.

Exploratory Data Analysis:

Let us check the types of variables and Missing Values in the data frame.

```
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.

We found that there are no Null values in the entire data frame.

If we group the data frame by Region and Channels, we will get the below data.

Region	Channel	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Lisbon	Hotel	761233	228342	237542	184512	56081	70632
	Retail	93600	194112	332495	46514	148055	33695
Oporto	Hotel	326215	64519	123074	160861	13516	30965
	Retail	138506	174625	310200	29271	159795	23541
Other	Hotel	2928269	735753	820101	771606	165990	320358
	Retail	1032308	1153006	1675150	158886	724420	191752

All three Regions Lisbon, Oporto and others contains Hotel and Retails channels. All the items are sold in all three regions and two types of channels. There are no items which is not sold in any region in any channel.

Q 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 help describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data. The most recognized types of descriptive statistics are measures of center: the mean, median, and mode, which are used at almost all levels of math and statistics.

Descriptive Statistics of the Data:

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

	count	unique	top	freq	mean	std	min	25%	50%	75%	max	IQR
Channel	440	2	Hotel	298	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Region	440	3	Other	316	NaN	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	13806.0
Milk	440.0	NaN	NaN	NaN	5796.265909	7380.377175	55.0	1533.0	3627.0	7190.25	73498.0	5657.25
Grocery	440.0	NaN	NaN	NaN	7951.277273	9503.162829	3.0	2153.0	4755.5	10655.75	92780.0	8502.75
Frozen	440.0	NaN	NaN	NaN	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0	2812.0
Detergents_Paper	440.0	NaN	NaN	NaN	2881.493182	4767.854448	3.0	256.75	816.5	3922.0	40827.0	3665.25
Delicatessen	440.0	NaN	NaN	NaN	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0	1412.0

From the above describe function,

Channel has two unique values, with "Hotel" as most frequent with 298 out of 440 transactions. The 67.7 % of spending is from "Hotel" channel.

Region has three unique values, with "Other" as most frequent with 316 out of 440 transactions. The 71.8 % of spending is from "Other" region.

All the items are having 440 records without any null values.

Fresh Item:
It has a mean of 12000.3, standard deviation of 12647.3, with min value of 3 and max value of 112151.
Q1(25%) is 3127.75, **Q3**(75%) is 16933.75 with **Q2**(50%) 8504
range = max-min =112148, **IQR** = Q3-Q1 = 13806.0

Milk item:
It has mean of 5796.27, standard deviation of 7380.38, with min value of 55 and max value of 73498.
Q1(25%) is 1533, **Q3**(75%) is 7190.25, with **Q2**(50%) 3627
range = max-min =73443, IQR = Q3-Q1 = 5657.25.

Grocery item:
It has a mean of 7951.28, standard deviation of 9503.16, with min value of 3 and max value of 92780.
Q1(25%) is 2153, **Q3**(75%) is 10655.8, with **Q2**(50%) 4755.5
range = max-min =92777, **IQR** = Q3-Q1 = 8502.8

Frozen:
It has a mean of 3071.93, standard deviation of 4854.67, with min value of 25 and max value of 60869.
Q1(25%) is 742.25, **Q3**(75%) is 3554.25, with **Q2**(50%) 1526
range = max-min =60844 & **IQR** = Q3-Q1 = 2812

Detergents_Paper:
It has a mean of 2881.49, standard deviation of 4767.85, with min value of 3 and max value of 40827.
Q1(25%) is 256.75, **Q3**(75%) is 3922, with **Q2**(50%) 816.5
range = max-min =40824 & **IQR** = Q3-Q1 = 3665.25.

Delicatessen:
It has a mean of 1524.87, standard deviation of 2820.11, with min value of 3 and max value of 47943.
Q1(25%) is 408.25, **Q3**(75%) is 1820.25, with **Q2(**50%) 965.5
range = max-min =47940 & **IQR** = Q3-Q1 = 1412.

When we group the data frame by the Regions,

```
Region
Lisbon     2386813
Oporto     1555088
Other     10677599
dtype: int64
```

Highest spend in the Region is from 'Others' and lowest spend in the region is from 'Oporto'

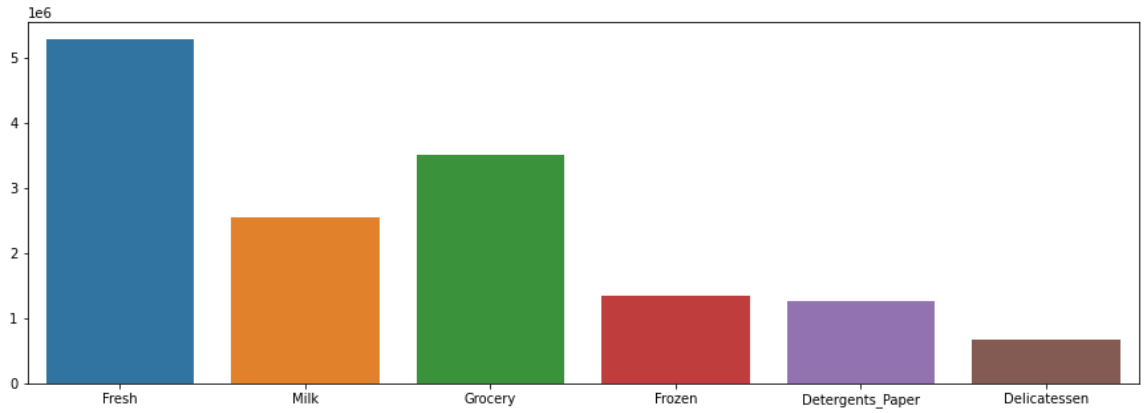
When we group the data frame by the Channels,

```
Channel
Hotel     7999569
Retail    6619931
dtype: int64
```

Highest spend in the Channel is from 'Hotel' and lowest spend in the Channel is from 'Retail'.

If we consider all items irrespective of the Region and channel, we find the below data.

```
Fresh              5280131
Milk               2550357
Grocery            3498562
Frozen             1351650
Detergents_Paper   1267857
Delicatessen        670943
dtype: object
```



The Highest is spent on Fresh and lowest is spent on Delicatessen.

If we consider for each region and each channel, below are the total spendings

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

Highest spend in the Region/Channel is from Others/Hotel and lowest spend in the Region/Channel is from Oporto/Hotel

If we consider Regions and Channel separately,

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

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
Lisbon	11101.73	5486.42	7403.08	3000.34	2651.12	1354.9
Oporto	9887.68	5088.17	9218.60	4045.36	3687.47	1159.7
Other	12533.47	5977.09	7896.36	2944.59	2817.75	1620.6

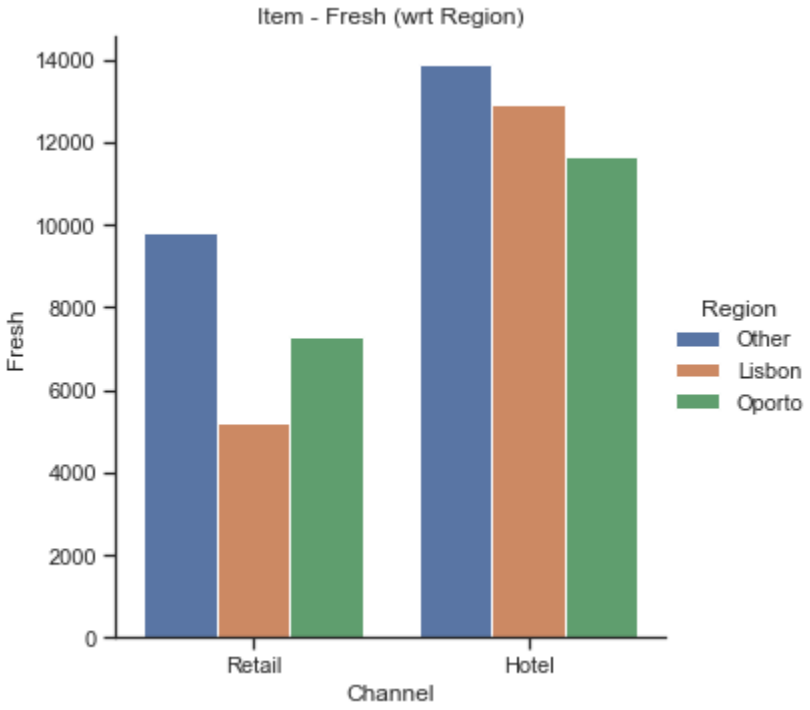
In Region "Lisbon" Average Highest Spending in Fresh and Lowest in Delicatessen items.
In Region "Oporto" Average Highest Spending in Fresh and Lowest in Delicatessen items.
In Region "Other" Average Highest Spending in Fresh and Lowest in Delicatessen items.

Q 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.

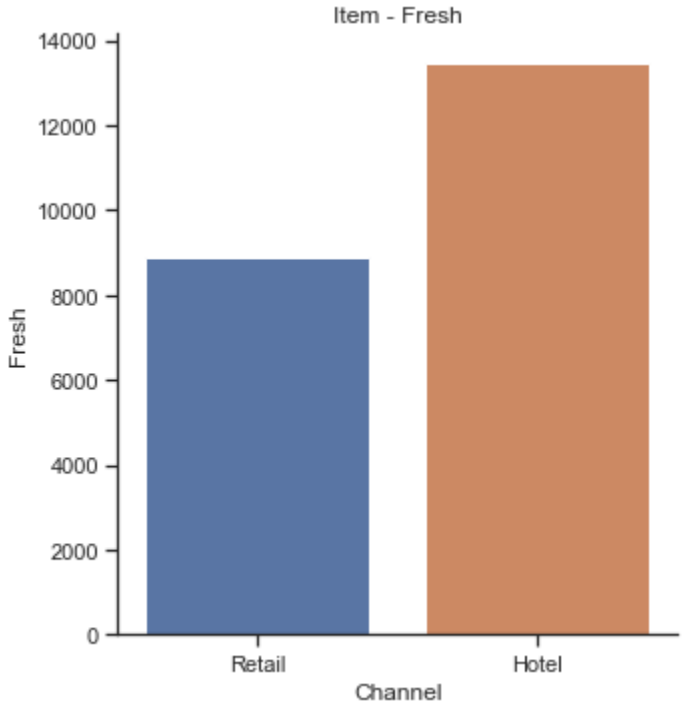
Here we see behavior in all items across Channel and Region use Bar Plot (Mean). They are different in Channel and Region.

FRESH:

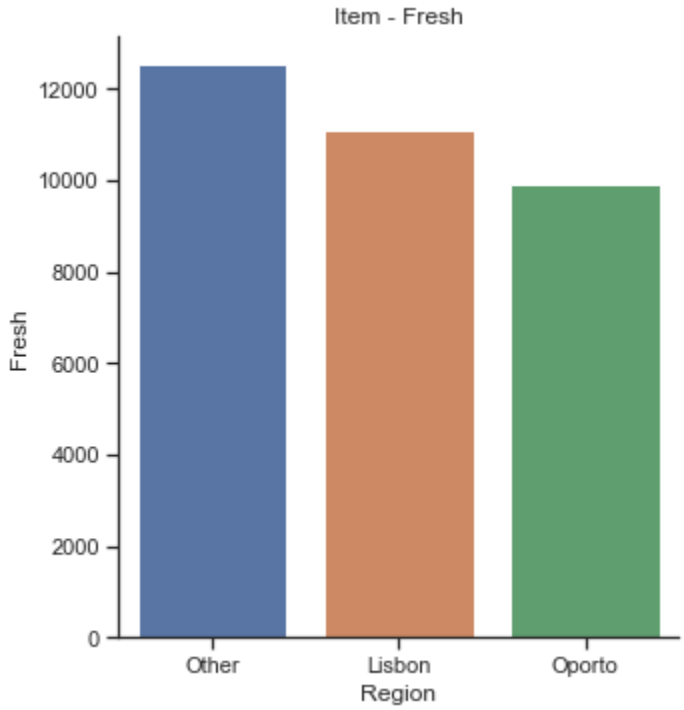
1. Barplot against Channels and Regions.



2. Barplot against Channels.



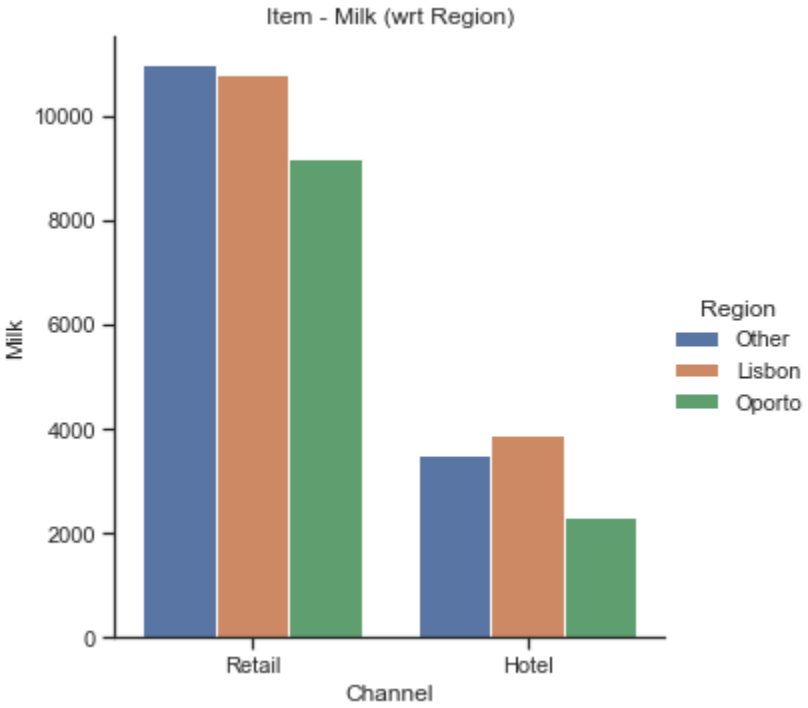
3. Barplot against Regions.



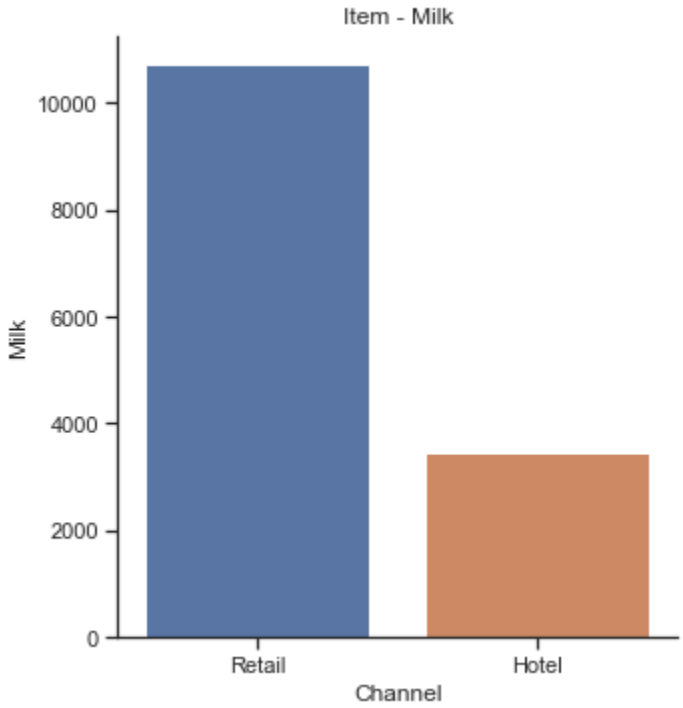
From the above data we can see that 'Other' region is spent more on Fresh.
Out of 2 channels, Hotel is spent more on Fresh.
If we combine channels and Regions together, the hotels in the 'Other' region spend more on Fresh and Retail channel in the Lisbon region is spent less.

MILK:

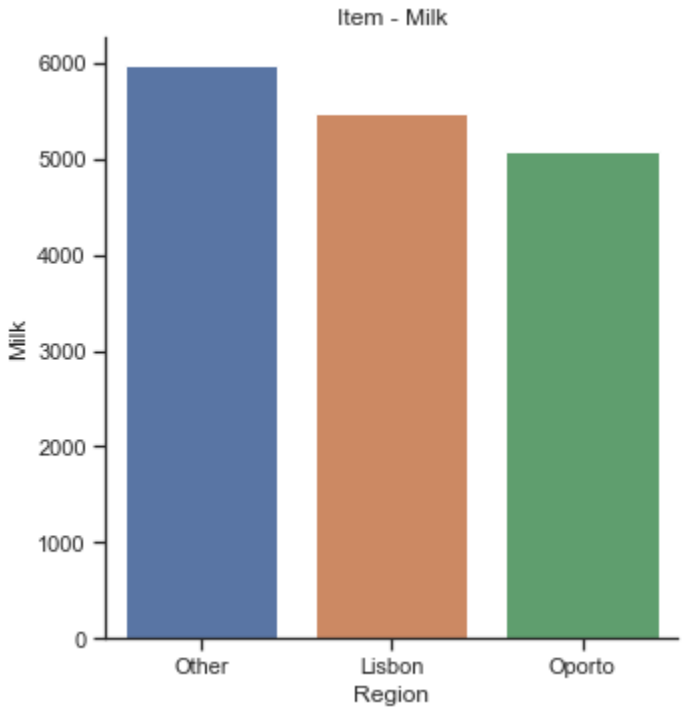
1. Barplot against Channels and Regions.



2. Barplot against Channels.



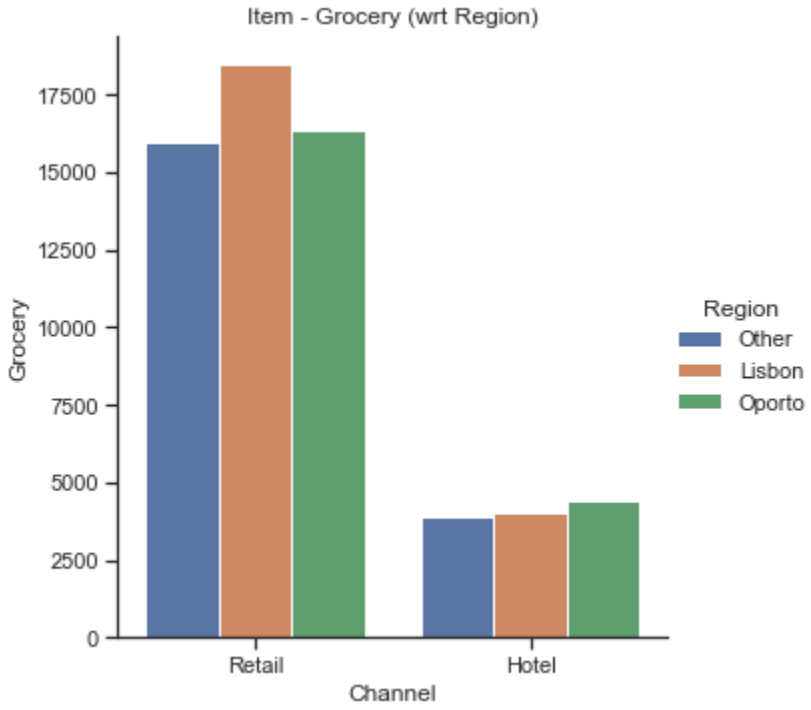
3. Barplot against Regions.



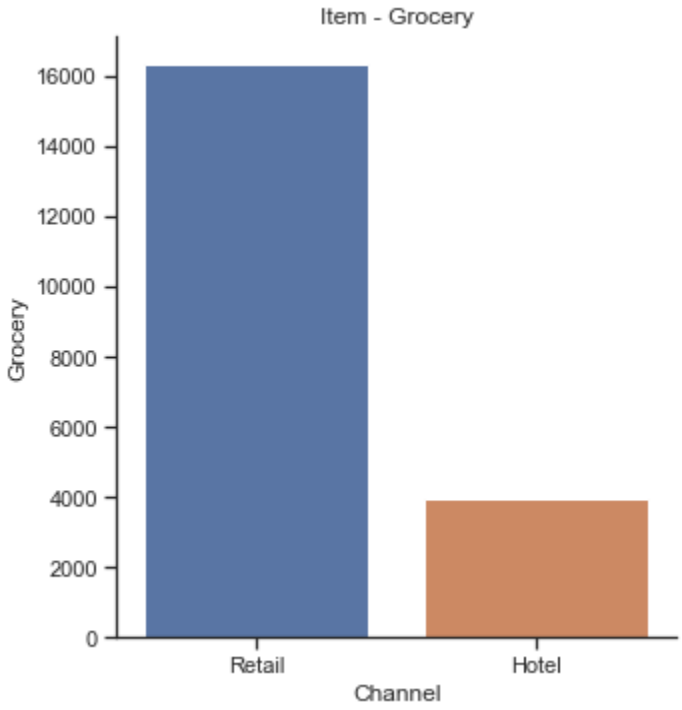
From the above data we can see that 'Other' region is spent more on Milk.
Out of 2 channels, Retail is spent more on Milk.
If we combine channels and Regions together, the retail in the 'Other' region spends more on Milk and Hotel channel in the Oporto region is spent less.

GROCERY:

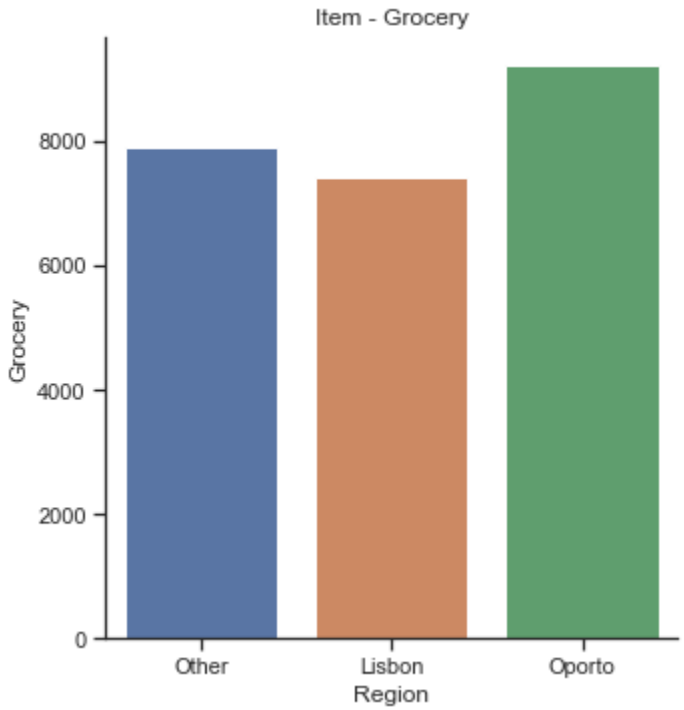
1. Barplot against Channels and Regions.



2. Barplot against Channels.



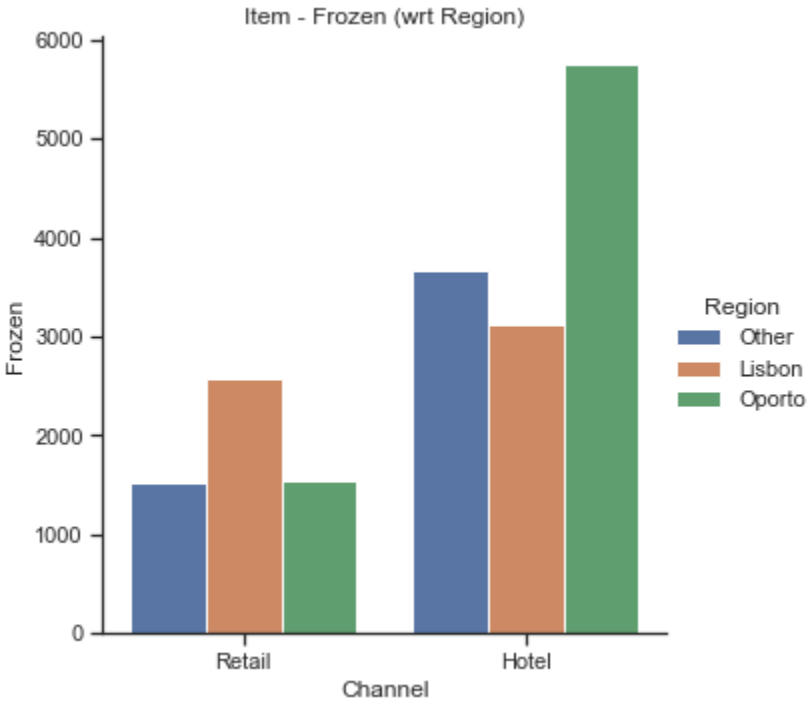
3. Barplot against Regions.



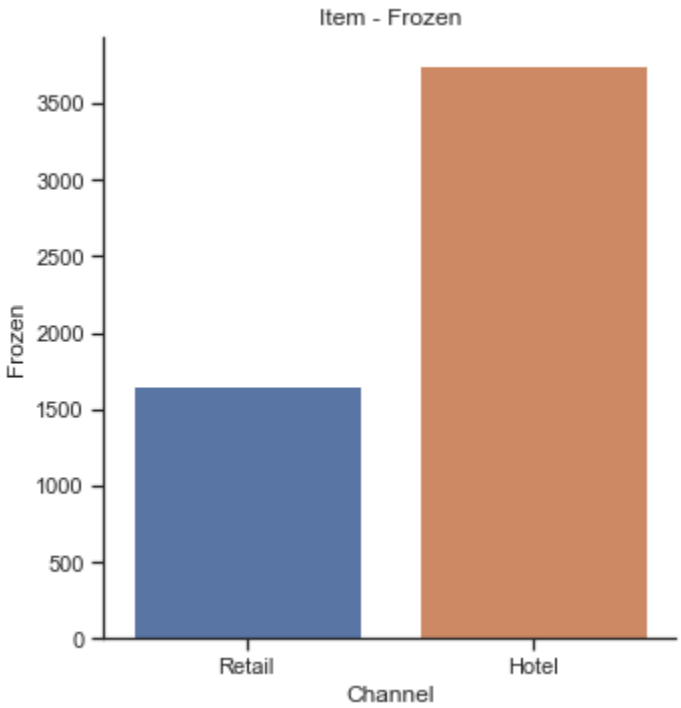
From the above data we can see that 'Oporto' region is spent more on Grocery.
Out of 2 channels, Retail is spent more on Grocery.
If we combine channels and Regions together, the Retail in the 'Lisbon' region spent more on Grocery and Hotel channel in the 'other' region is spent less.

FROZEN:

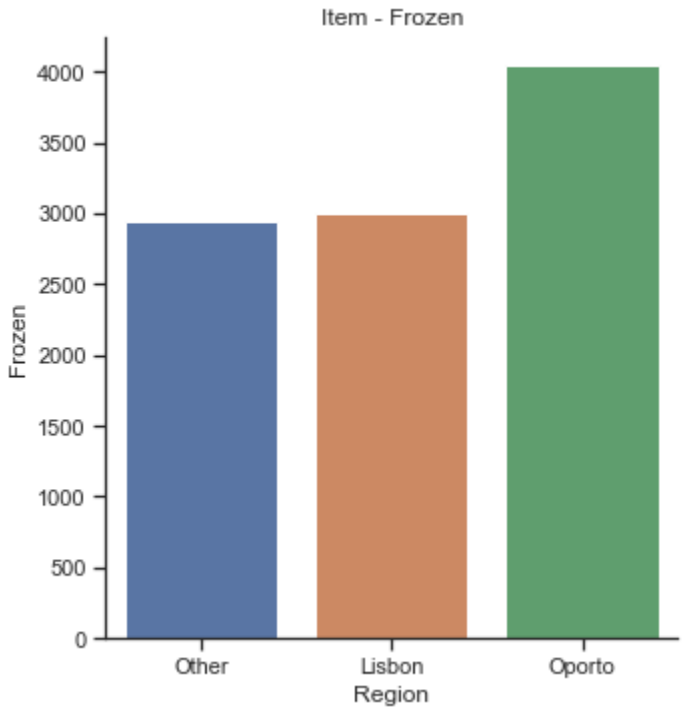
1. Barplot against Channels and Regions.



2. Barplot against Channels.



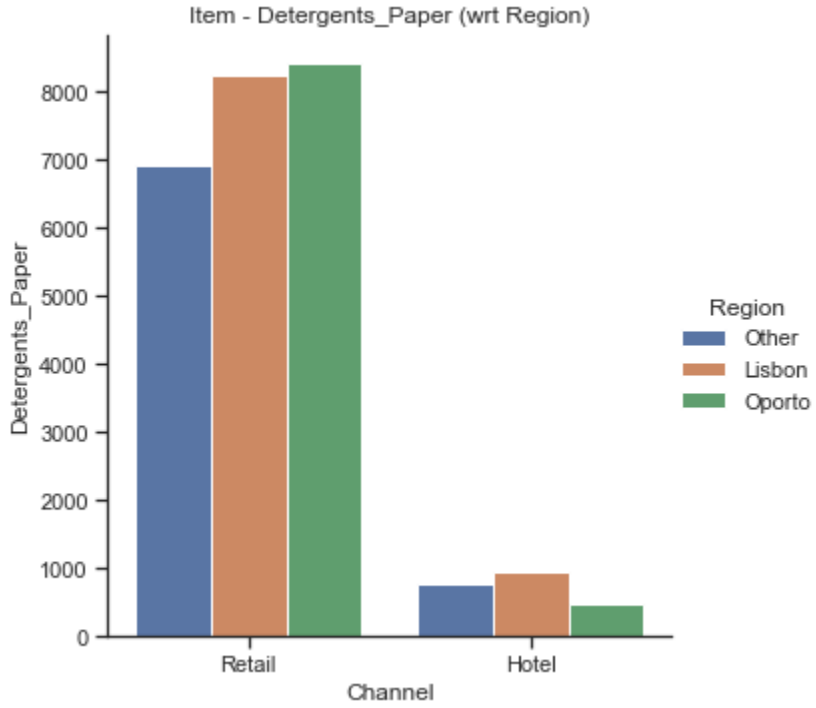
3. Barplot against Regions.



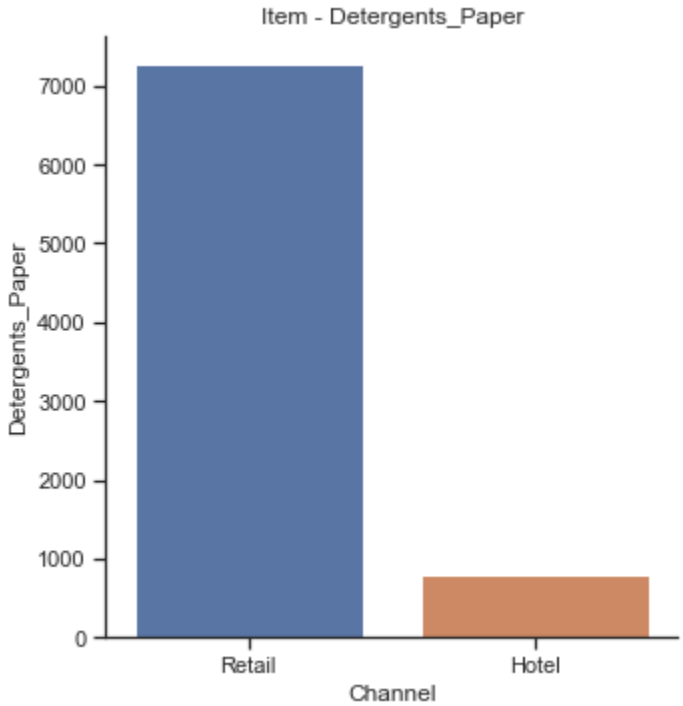
From the above data we can see that 'Oporto' region is spent more on Grocery.
Out of 2 channels, Hotel is spent more on Grocery.
If we combine channels and Regions together, the Hotel in the 'Other' region spent more on Grocery and Retail channel in the 'other' region is spent less.

DETERGENT PAPER:

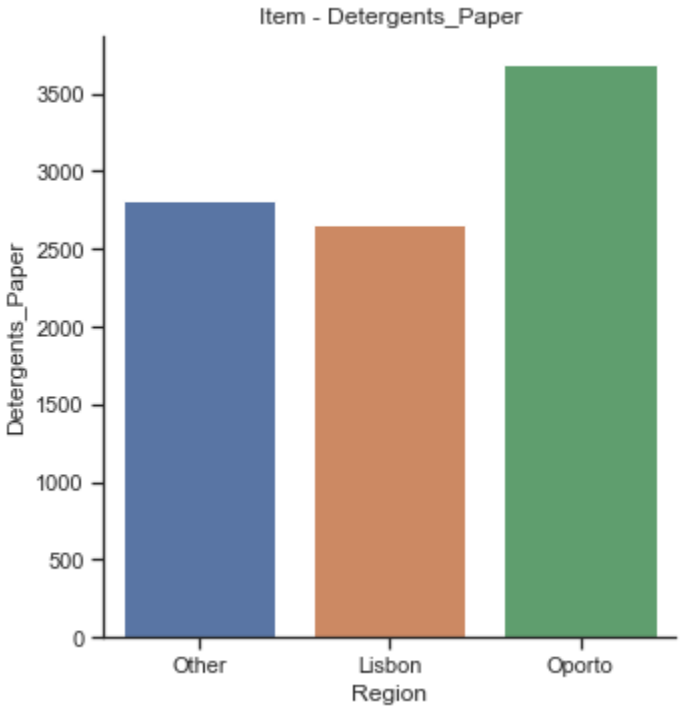
1. Barplot against Channels and Regions.



2. Barplot against Channels.



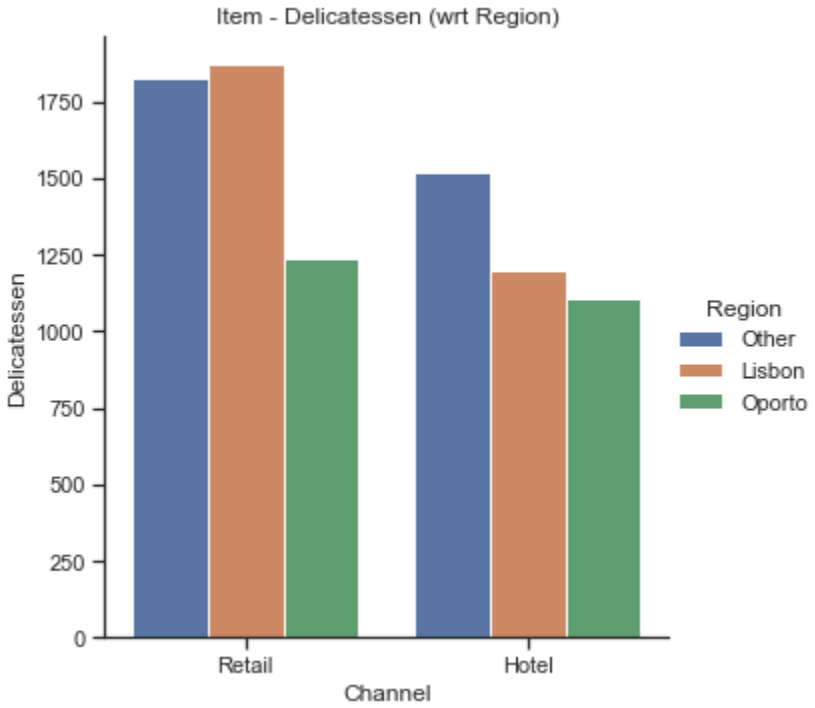
3. Barplot against Regions.



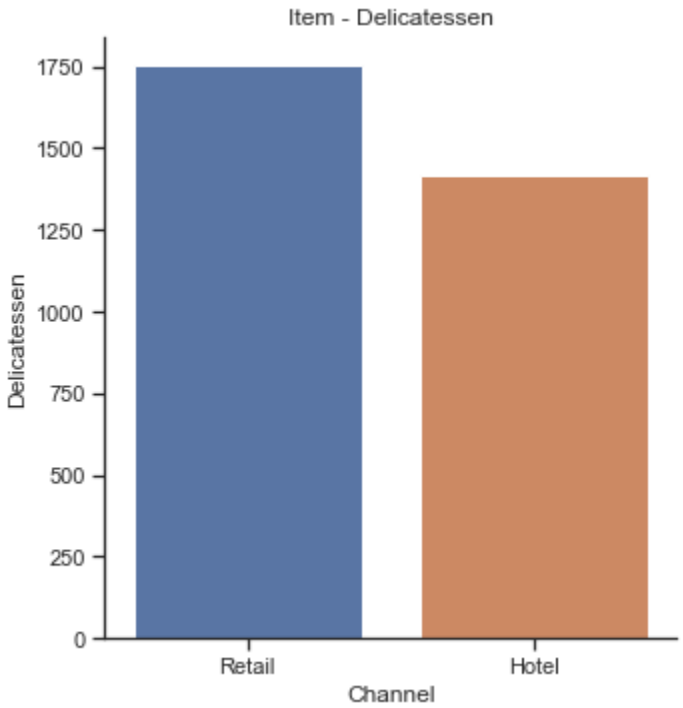
From the above data we can see that 'Oporto' region is spent more on Detergents_Paper.
Out of 2 channels, Retail is spent more on Detergents_Paper.
If we combine channels and Regions together, the Retail in the 'Oporto' region spent more on Detergents_Paper and Hotel channel in the 'Oporto' region is spent less.

DETERGENT PAPER:

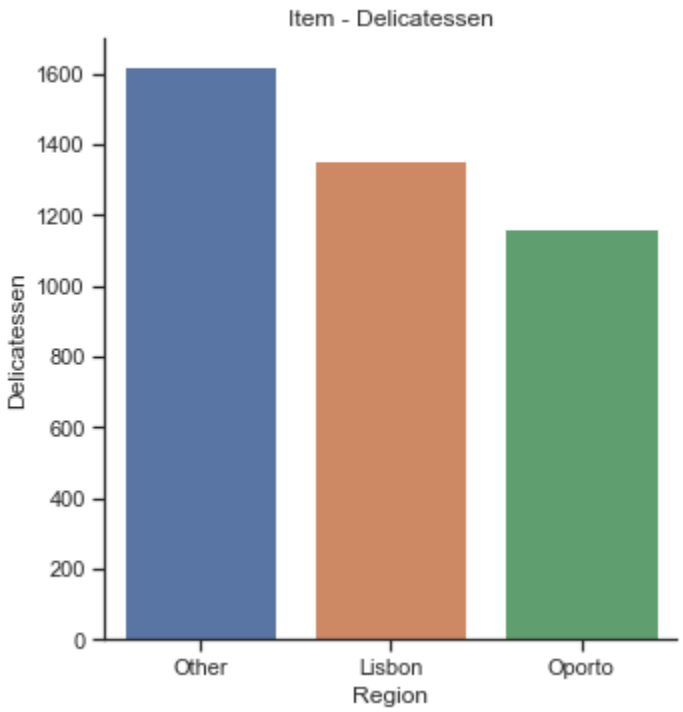
1. Barplot against Channels and Regions.



2. Barplot against Channels.



3. Barplot against Regions.



From the above data we can see that 'Other' region is spent more on Delicatessen.
Out of 2 channels, Retail is spent more on Delicatessen.
If we combine channels and Regions together, the Retail in the 'Lisbon' region spent more on Delicatessen and Hotel channel in the 'Oporto' region is spent less.

Q 1.3. On the basis of the descriptive measure of variability, which item shows the most inconsistent behavior? Which items shows the least inconsistent behavior?

	count	mean	std	min	25%	50%	75%	max	CV
Fresh	440.0	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0	1.053918
Milk	440.0	5796.265909	7380.377175	55.0	1533.00	3627.0	7190.25	73498.0	1.273299
Grocery	440.0	7951.277273	9503.162829	3.0	2153.00	4755.5	10655.75	92780.0	1.195174
Frozen	440.0	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0	1.580332
Detergents_Paper	440.0	2881.493182	4767.854448	3.0	256.75	816.5	3922.00	40827.0	1.654647
Delicatessen	440.0	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0	1.849407

If we consider the standard deviation of the items, we get the below data.

```
Fresh              12647.329
Milk                7380.377
Grocery             9503.163
Frozen              4854.673
Detergents_Paper    4767.854
Delicatessen        2820.106
dtype: float64
```

From the above data,
Fresh item have highest Standard deviation So that is Inconsistent.
Delicatessen item have smallest Standard deviation, So that is consistent.

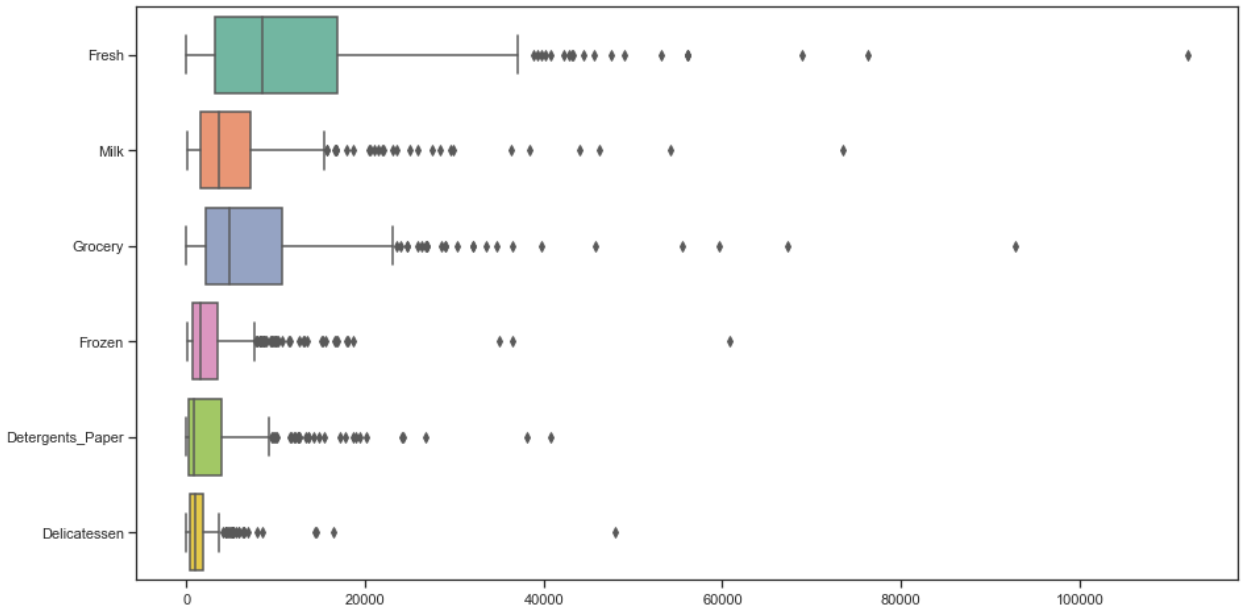
If we Calculate the Coefficient of variables, we get the below data.

```
Fresh              1.053918
Milk               1.273299
Grocery            1.195174
Frozen             1.580332
Detergents_Paper   1.654647
Delicatessen       1.849407
Name: CV, dtype: float64
```

From the above data,
Fresh item have lowest coefficient of Variation So that is consistent.
Delicatessen item have highest coefficient of Variation, So that is Inconsistent.

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

Let us use the box plot to identify the outliers in the data.



We could see that there are outliers in all the items across the product range. i.e. All the products data are having outliers. Perhaps the most common or familiar type of outlier is the observations that are far from the rest of the observations or the center of mass of observations. Removing outliers from data prior to modeling can result in a better fit of the data and, in turn, more skillful predictions.

Q 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 I found out that:
1. There are inconsistencies in spending of different items by calculating CV.
2. The spending of Hotel and Retail channel are different which can be equal.
3. Also the spend can be equal for different Regions if business team focus on other items than 'Fresh' and 'Grocery'.
4. Since the Delicatessen show the least inconsistent behavior, the business should invest more in this food item because it is less risky
5. Fresh products require more spending, to cut cost the wholesale distributor can concentrate more on other food items like Milk, Grocery, Frozen, Detergent paper and Delicatessen

Executive Summary (PROBLEM-2)

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).

Data Description

1. ID: Serial number of the student
2. Gender: Gender of the student
3. Age: age of the student
4. Class: class of the student (senior or junior)
5. Major: major of the student (ex: Management, Economics/Finance or any other)
6. Grad Intention: Graduation intention of the student (Yes, No or undecided)
7. GPA: GPA of the student
8. Employment: type of employee (Fulltime or parttime)
9. Salary: salary of the employed student
10. Social Networking: No of social network usage.
11. Satisfaction: satisfaction
12. Spending: Spending in education
13. Computer: Type of computer using (Laptop or desktop)
14. Text Messages: Count of text messages

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

2.1.1. Gender and Major

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

2.1.2. Gender and Grad Intention

Grad Intention	No	Undecided	Yes
Gender			
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 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?

Here we have total 62 students' data. Out of which 33 are female and 29 are male.

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

The probability that a randomly selected CMSU student will be Male is 46.77%.

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 is 53.22%.

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.

The following table shows the conditional probability of different majors among the male students in CMSU which is calculated by number of male students in accounting, CIS, economics/finance, international business, management, other, Retailing/ Marketing, undecided divided by total number of male students.

Gender Major	Female	Male	Male_percentage	Female_percentage
Accounting	3	4	13.793103	9.090909
CIS	3	1	3.448276	9.090909
Economics/Finance	7	4	13.793103	21.212121
International Business	4	2	6.896552	12.121212
Management	4	6	20.689655	12.121212
Other	3	4	13.793103	9.090909
Retailing/Marketing	9	5	17.241379	27.272727
Undecided	0	3	10.344828	0.000000

From the above data table:
The conditional probability of Accounting among the male students is 13.8%
The conditional probability of CIS among the male students is 3.44%
The conditional probability of Economics/Finance among the male students is 13.8%
The conditional probability of International Business among the male students is 6.9%
The conditional probability of Management among the male students is 20.7%
The conditional probability of Other among the male students is 13.8%
The conditional probability of Retailing/Marketing among the male students is 17.24%
The conditional probability of Undecided among the male students is 10.34%

\female students of CMSU.

From the above table,

The conditional probability of Accounting among the female students is 9.09%
The conditional probability of CIS among the female students is 9.09%
The conditional probability of Economics/Finance among the female students is 21.21%
The conditional probability of International Business among the female students is 12.12%
The conditional probability of Management among the female students is 12.12%
The conditional probability of Other among the female students is 9.09%
The conditional probability of Retailing/Marketing among the female students is 27.27%
The conditional probability of Undecided among the female students is 0%

2. 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.

probability That a randomly chosen student is a male and intends to graduate = number of male students who intends to graduate/total number of students=17/62

Probability that a randomly chosen student is a male and intends to graduate is 0.27419354838709675 i.e., 27.42%

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

probability that a randomly selected student is a female and does NOT have a laptop = number of female students without laptop/total number of students=4/62.

Probability that a randomly chosen student is a female and does not have a laptop is 0.06451612903225806 i.e., 6.45%

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?

probability that a randomly chosen student is a male or has full-time employment = number of male students or students who have full time employment/total number of students=32/62.

Probability that a randomly chosen student is a male or has full time employment is 0.5161290322580645 i.e., 51.62%

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

There is total 33 female students out of which 8 are majoring in international or management. conditional probability that given a female student is randomly chosen; she is majoring in international

business or management= number of female student from international business or management/total number of female students=8/33.

Probability that a randomly chosen student is female and has Major in Management or International Business 0.2424 i.e., 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?

Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

Let us consider event A=graduate intention, B= Being a female student.
These two events A and B are said to be independent if the fact that one event has occurred does not affect the probability that the other event will occur. We can see out of 29 male, 17 intent to graduate and out of 33 females only 11 intents to graduate.
Events A and B are independent if the equation $P(A \cap B) = P(A) \cdot P(B)$ holds true.

P(female) = 33/62 = 0.5322
P(intent to graduate) =28/62 = 0.45
P(female & intent to graduate) =11/62 = 0.17

Here,
P(female) * P(intent to graduate) =0.53*0.45 =0.24

Also we found that, P(female) * P(intent to graduate) not equal to P(female & intent to graduate)

Hence, the graduate intention and being female are dependent 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?

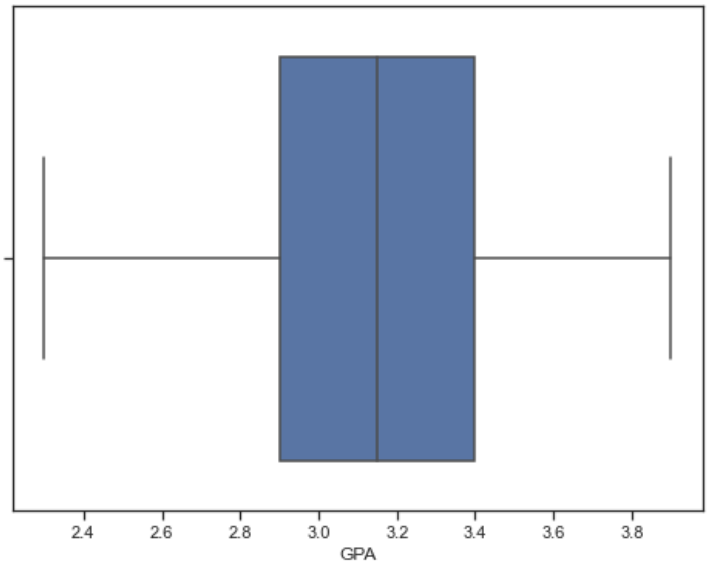
There are 17 students with GPA less than 3. Therefore, the probability that his/her GPA is less than 3 is 27.42%

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.

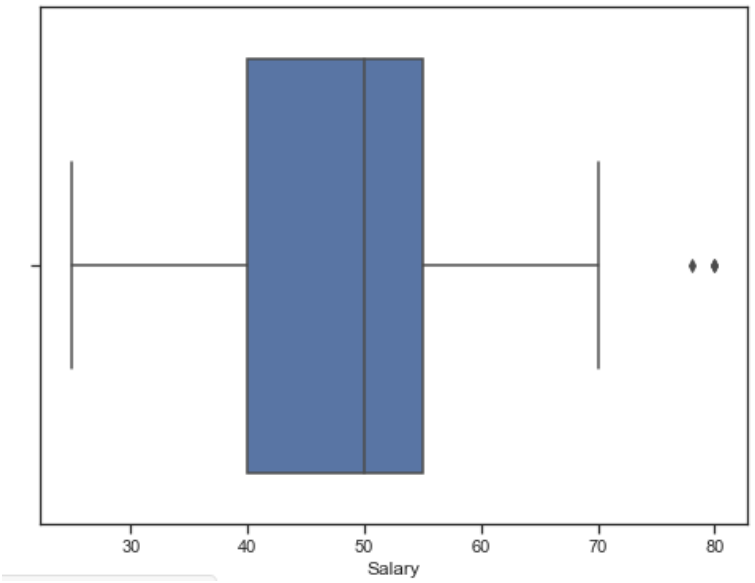
Probability that a randomly selected male earns more than 50 = 14/29=0.4827=48.27%
Probability that a randomly selected Female earns more than 50 = 18/33=0.5454=54.54%

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.

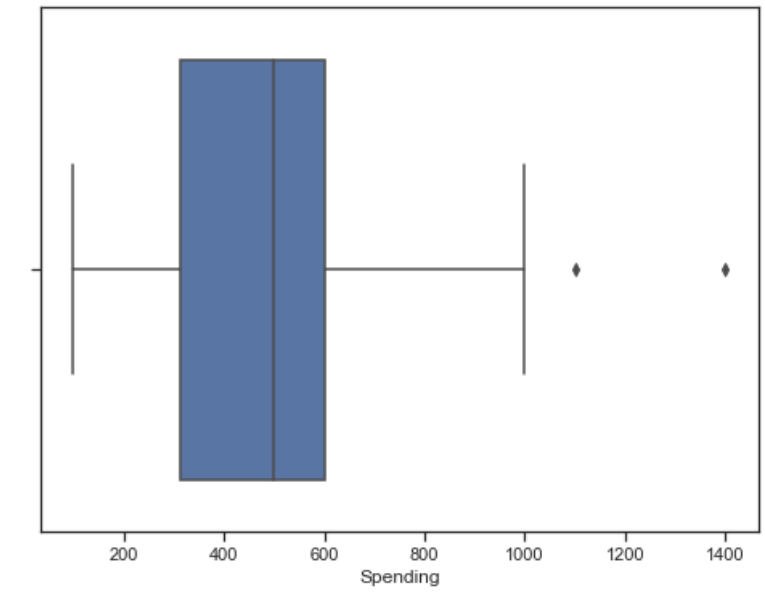
Boxplot of GPA:



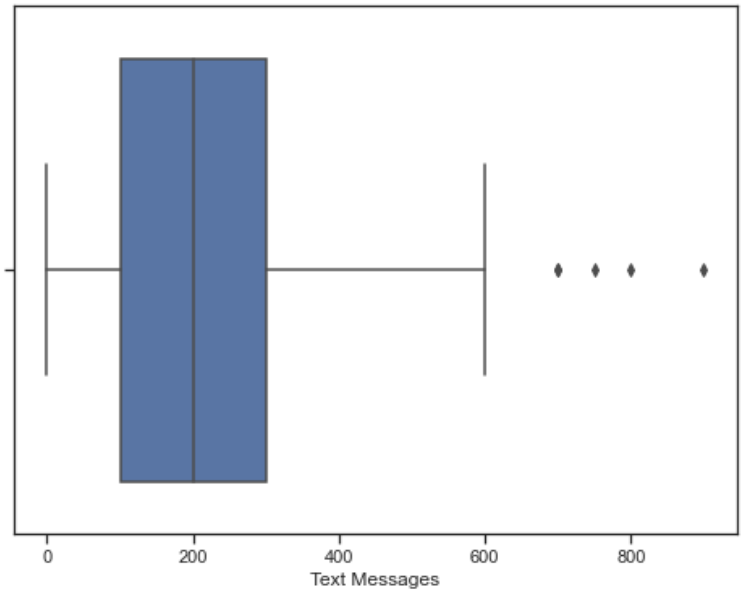
Boxplot of Salary:



Boxplot of Spending:



Boxplot of Text messages:



From the above boxplots, we could say that only GPA is normally distributed.

And to confirm this, we did the Shapiro–Wilk test and the output from Python we got below results.

GPA:

```
ShapiroResult(statistic=0.9685361981391907, pvalue=0.11204058676958084)
```

P-value is more than 0.05

Salary:

```
ShapiroResult(statistic=0.9565856456756592, pvalue=0.028000956401228905)
```

p-value is less than 0.05

Spending:

```
ShapiroResult(statistic=0.8777452111244202, pvalue=1.6854661225806922e-05)
```

p-value is less than 0.05

Text Messages:

```
ShapiroResult(statistic=0.8594191074371338, pvalue=4.324040673964191e-06)
```

p-value is less than 0.05

Therefore, we could confirm that only GPA is normally distributed among four data sets.

From this analysis, we can conclude that the data shows that there are multiple factors that affect the graduation of a student. The survey conducted by Student News Service 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. We can conclude that in order to help students graduate and find suitable employment the university can work on improving the infrastructure by providing easy access to computers and conducting social networking events. The probabilities of male students graduating is more than that of female students, so female students need more support and choice of major.

Executive Summary (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 includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.

Q. 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 sample A:
One sample t test statistic: -1.4735046253382782
p-value: 0.07477633144907513
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.

F**or sample B**
One sample t test statistic: -3.1003313069986995
p-value: 0.0020904774003191826
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.

Q.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?

H0 : mu(A)= mu(B)
Ha : mu(A)!= mu(B)
alpha = 0.05

t_statistic=1.29
pvalue=0.202
As the p-value > alpha, **do not reject H0**;

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.

----------END----------