STATISTICAL METHODS FOR DECSION MAKING

PROJECT

BUSINESS REPORT

Name: Yashveer Kothari. A

Course: PGP-DSBA-2022

Project: Statistical Methods for Decision Making

Date:07th May 2022

TABLE OF CONTENTS:

LIST OF TABLES:

LIST OF FIGURES:

PROBLEM 1:

EXECUTIVE SUMMARY:

A Wholesale distributor operates in different regions of Portugal like Lisbon, Oporto and Other regions selling 6 different varieties of products namely Fresh, Milk, Grocery, Frozen, Detergent Paper, Delicatessen through different sales channels Hotel and Retail. The Dataset consists of 440 large retailers annual spending on the varieties of products in the different regions and through different channels.

INTRODUCTION:

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and other parameters. Analyse the Total spending made by the 3 regions (Lisbon, Oporto and others) through different channels (Hotel and Retail) on different varieties of products (Fresh, Milk, Grocery, Frozen, Detergent Paper and Delicatessen). This problem will help in exploring the summary statistics and descriptive statistics.

DATA DESCRIPTION:

- Buyer/spender: The no. of buyers.
- Channel: The distribution of products for sale happens through Hotel and Retail channels
- Region: The area of operations (Lisbon, Oporto, Others)
- Fresh: Variety of product continuous from 3 to 112151
- Milk: Variety of product continuous from 55 to 73498
- Grocery: Variety of product continuous from 30 to 92780
- Frozen: Variety of product continuous from 25 to 60869
- Detergent paper: Variety of product continuous from 3 to 40827
- Delicatessen: Variety of product continuous from 3 to 47943

SAMPLE DATASET:

Out[3]:

	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

Out[4]:

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

Table 1: Dataset Sample

Dataset has 9 variables with 6 different types of products and its spending value in different regions and Channels.

EXPLORATORY DATA ANALYSIS:

Checking the data types and shape of the Data set:

Python Output:

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

```
The total Number of rows and coloumns are: (440, 10) respectively
```

As above mentioned, 2 columns are object or categorical type and other 8 are integer or continuous type. There are 440 row and 10 columns in the dataset

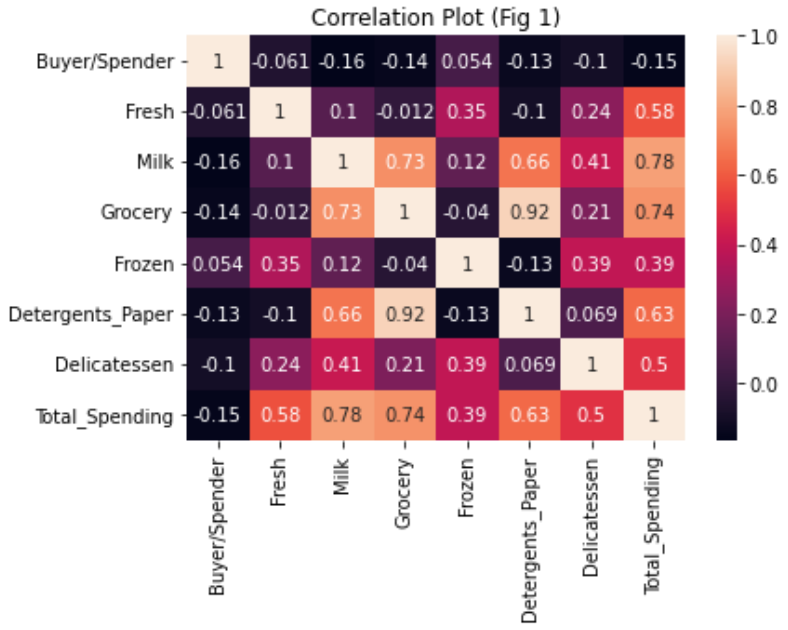
Checking the Missing value of the Dataset:

Python Output:

```
Buyer/Spender       0
Channel             0
Region              0
Fresh               0
Milk                0
Grocery             0
Frozen              0
Detergents_Paper    0
Delicatessen        0
Total_Spending      0
dtype: int64
```

The Dataset does not have any missing values.

Correlation Plot:



From the correlation plot, we can see that various attributes are highly correlated to each other. Correlation values near to 1 or 0 are highly positively correlated and highly negatively correlated respectively.

PAIRPLOT:

Pair plot shows the relationship between the variables in the form of scatterplot and the distribution of the variable in the form of histogram. From the graph, we can see that there is positive linear relationship between variables.

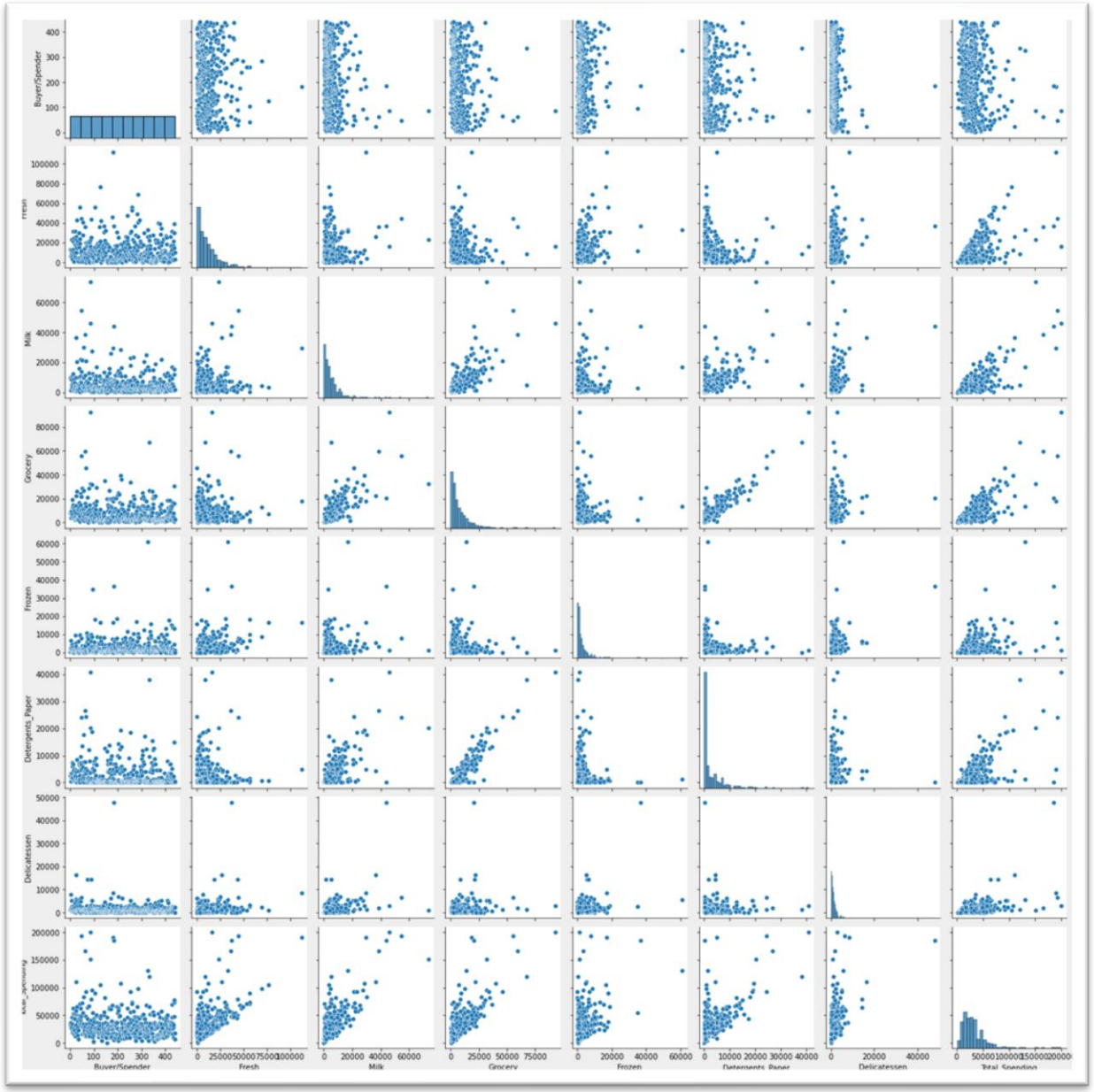


Fig 2: Pair Plot

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?

Answer: 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 centre: the mean, median, and mode, which are used at almost all levels.

The descriptive Statistics to summarize data:

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

Table 2: Descriptive Statistics

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. i.e., 67.7 percentage of spending comes from "Hotel" channel.

Retail has three unique values, with "Other" as most frequent with 316 out of 440 transactions. i.e.,71.8 percentage of spending comes from "Other" region.

Fresh item: 440 records, mean of 12000.3, Standard deviation of 12647.3, minimum value of 3, maximum value of 112151, Q1(25%) is 3127.75, Q3(75%) is 16933.8, with Q2(50%) 8504 range = max-min =112151-3=112148 & IQR = Q3-Q1 = 16933.8-3127.75 = 13,806.05

Milk item: 440 records have a mean of 5796.27, Standard deviation of 7380.38, with minimum value of 55, max value of 73498. Q1(25%) is 1533, Q3(75%) is 7190.25, Q2(50%) 3627. range = max-min =73498-55=73443 IQR = Q3-Q1 = 7190.25-1533 = 5657.25

Grocery item 440 records mean of 7951.28, Standard deviation of 9503.16, Min value of 3, Max value of 92780. Q1(25%) is 2153, Q3(75%) is 10655.8, Q2(50%) 4755.5 range = max-min =92780-3=92777 IQR = Q3-Q1 = 10655.8-2153 = 8502.8

Frozen 440 records mean of 3071.93, standard deviation of 4854.67, min value of 25, max value of 60869. Q1(25%) is 742.25, Q3(75%) is 3554.25, Q2(50%) 1526 range = max-min =60869-25=60844, IQR = Q3-Q1 = 3554.25-742.25 = 2812

Detergents Paper 440 records, mean of 2881.49, standard deviation of 4767.85, min value of 3, max value of 40827. Q1(25%) is 256.75, Q3(75%) is 3922, Q2(50%) 816.5 range = max-min =40827-3=40824 IQR = Q3-Q1 = 3922-256.75 = 3665.25

Delicatessen (440 records), mean of 1524.87, standard deviation of 2820.11, min value of 3 max value of 47943. 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

To calculate the Highest and the lowest spending region wise and channel wise

a. We calculate the total number of regions and channels

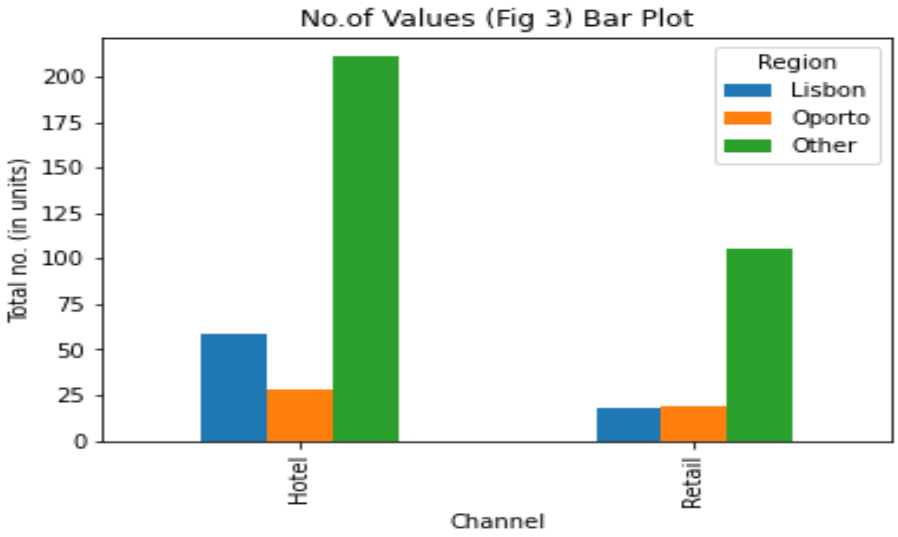
Python Output:

Region:

```
Out[6]: Other     316
        Lisbon     77
        Oporto     47
        Name: Region, dtype: int64
```

Channel:

```
Out[7]: Hotel     298
        Retail    142
        Name: Channel, dtype: int64
```



Python Output:

```
Region    Lisbon  Oporto  Other
Channel
Hotel         59      28    211
Retail        18      19    105
```

b. Adding a new column “Total Spending” to get the output:

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total_Spending
0	1	Retail	Other	12669	9656	7561	214	2674	1338	34112
1	2	Retail	Other	7057	9810	9568	1762	3293	1776	33266
2	3	Retail	Other	6353	8808	7684	2405	3516	7844	36610
3	4	Hotel	Other	13265	1196	4221	6404	507	1788	27381
4	5	Retail	Other	22615	5410	7198	3915	1777	5185	46100
...	...	...	...	...	...	...	...	...	...	...
435	436	Hotel	Other	29703	12051	16027	13135	182	2204	73302
436	437	Hotel	Other	39228	1431	764	4510	93	2346	48372
437	438	Retail	Other	14531	15488	30243	437	14841	1867	77407
438	439	Hotel	Other	10290	1981	2232	1038	168	2125	17834
439	440	Hotel	Other	2787	1698	2510	65	477	52	7589

440 rows × 10 columns

Table 3: Adding new column

c. Calculating the Highest and lowest spending regions and channels

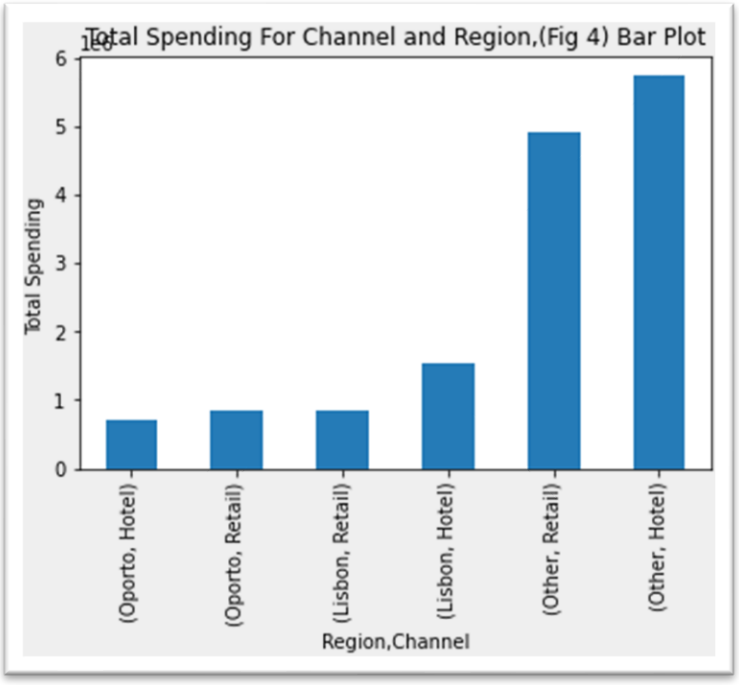
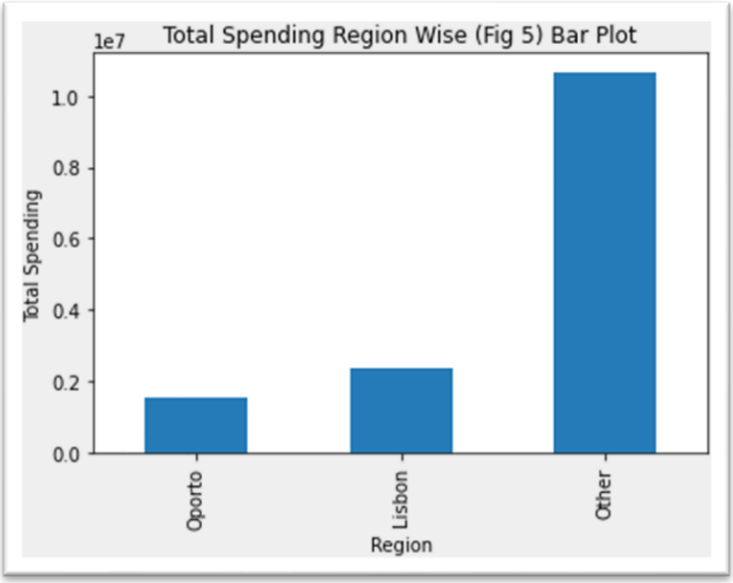


Table 9: Total Spending

From Fig 4 and from the Python Output we conclude that:

Highest Total Spending in Region/Channel is from Other/Hotel 5742077 respectively.

Lowest Total Spending in Region/Channel is from Oporto/Hotel 719150 respectively.

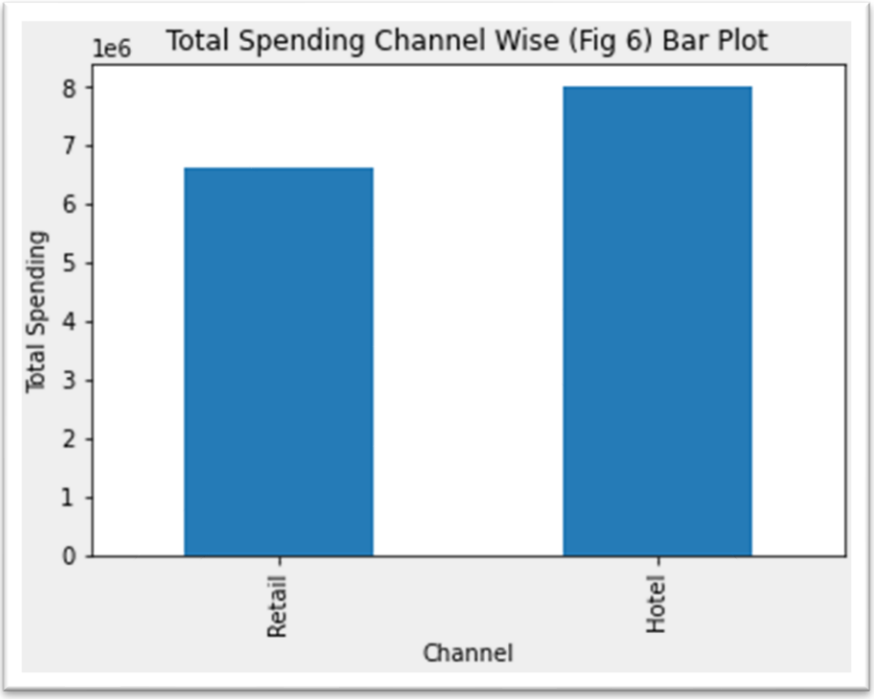


Python Output:

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

From Fig 5 and from the Python Output it can be inferred that individually:

The Other regions have spent more and Oporto has spent the least.



Python Output:

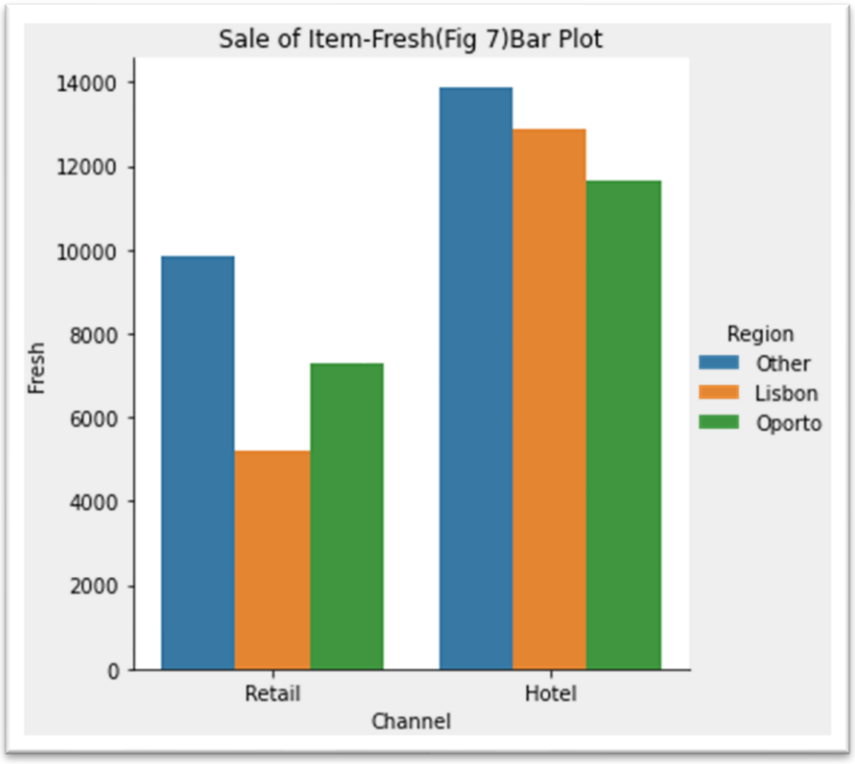
```
Channel
Hotel     7999569
Retail    6619931
Name: Total_Spending, dtype: int64
```

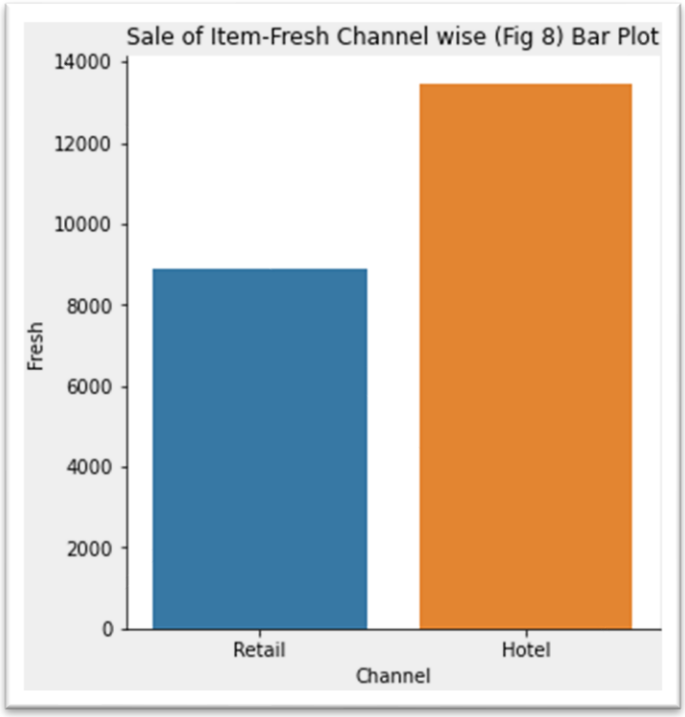
From Fig 6 and from the output it can inferred that individually:

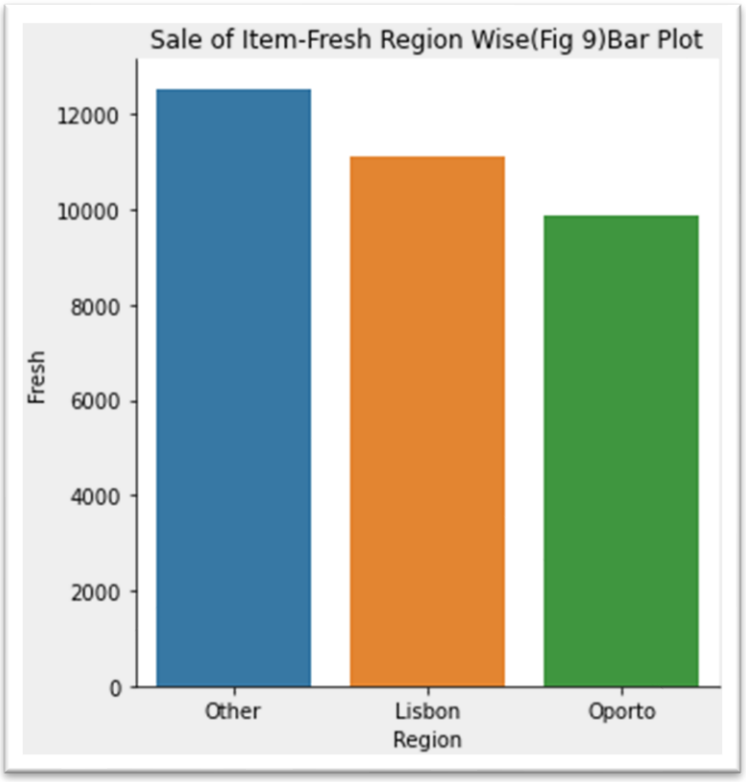
The Hotel Channels have spent more and Retail Channels have spent the least.

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

Item 1: Fresh





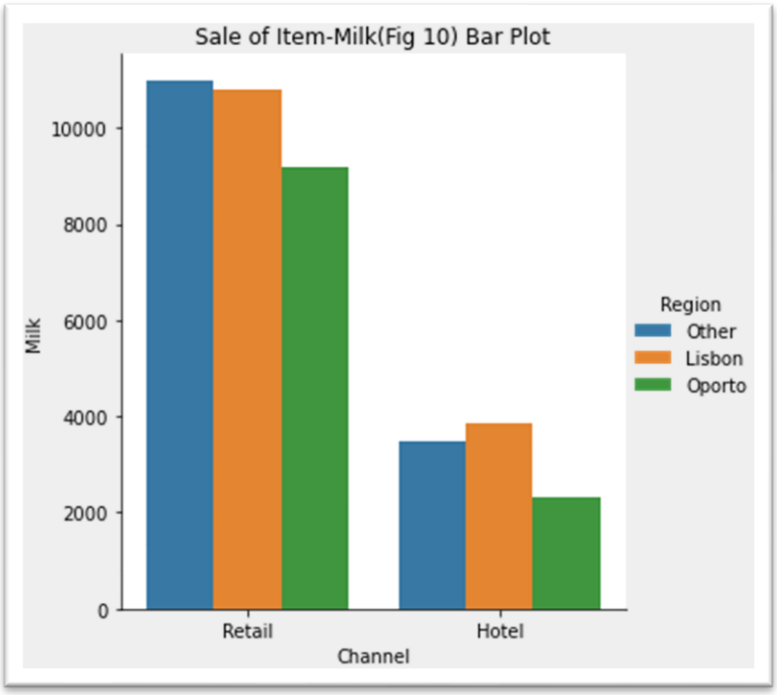


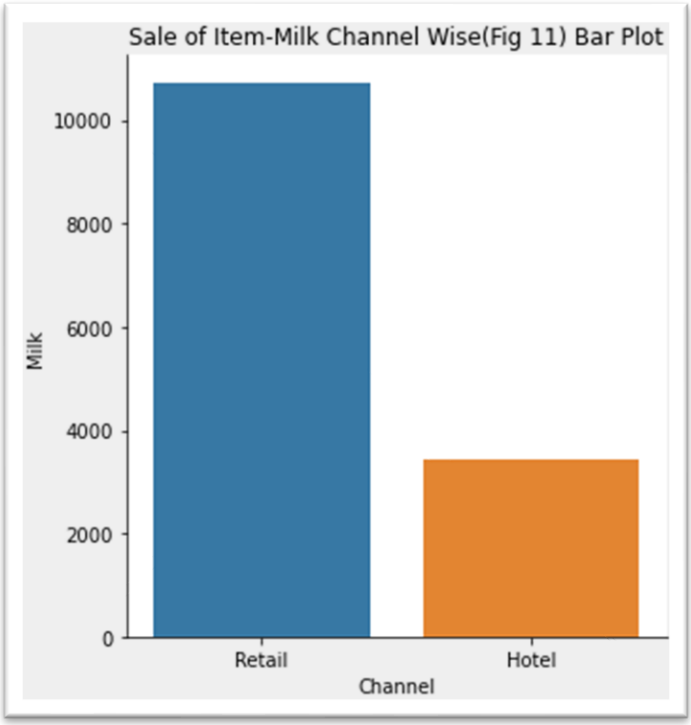
As per the above plots it is inferred that (Refer Fig 7)

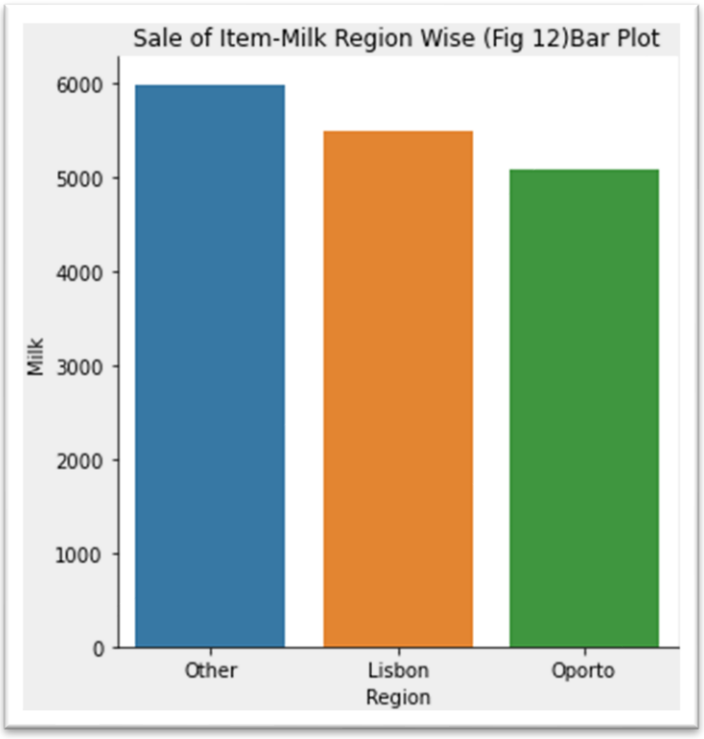
1. The Fresh Item is sold more in the Hotel Channel with Other Region being the highest and Oporto being the least in the same channel

2. Whereas the sale in the Retail Channel is less with Other Regions having the highest Retail Sale and Lisbon having least retail sale.

Item 2: Milk


Sale of Item-Milk(Fig 10) Bar Plot
Milk
10000
8000
6000
4000
2000
0
Retail
Hotel
Channel
Region
Other
Lisbon
Oporto


Sale of Item-Milk Channel Wise(Fig 11) Bar Plot
Milk
10000
8000
6000
4000
2000
0
Retail
Hotel
Channel



As per the above plots it is inferred that (Refer Fig 10)

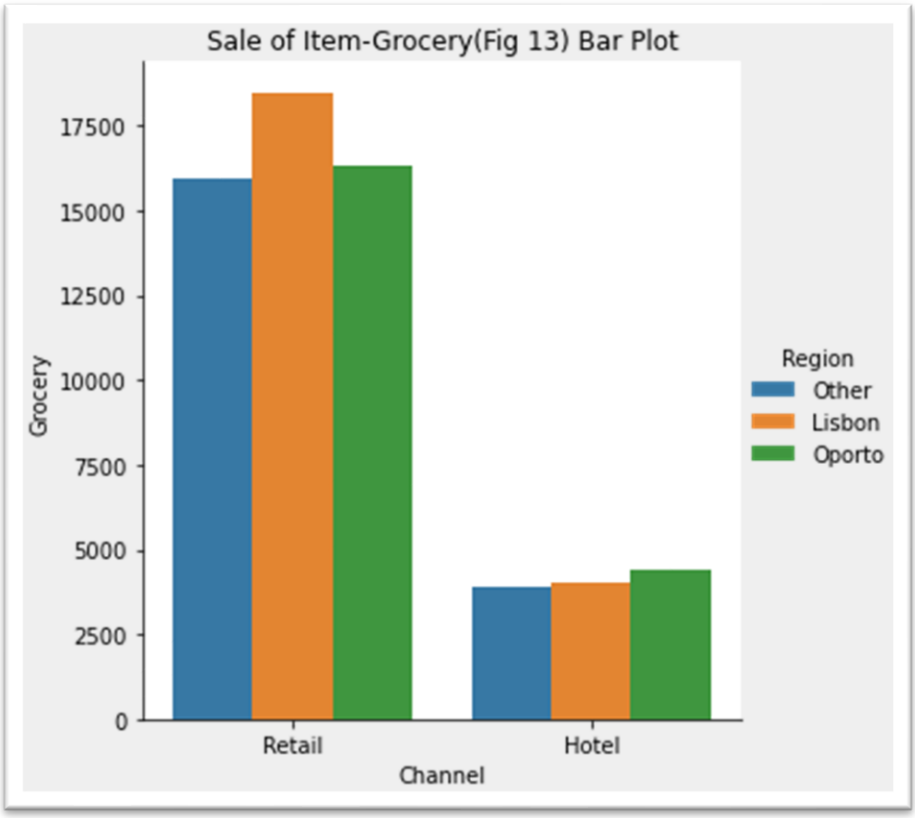
For Item Milk,

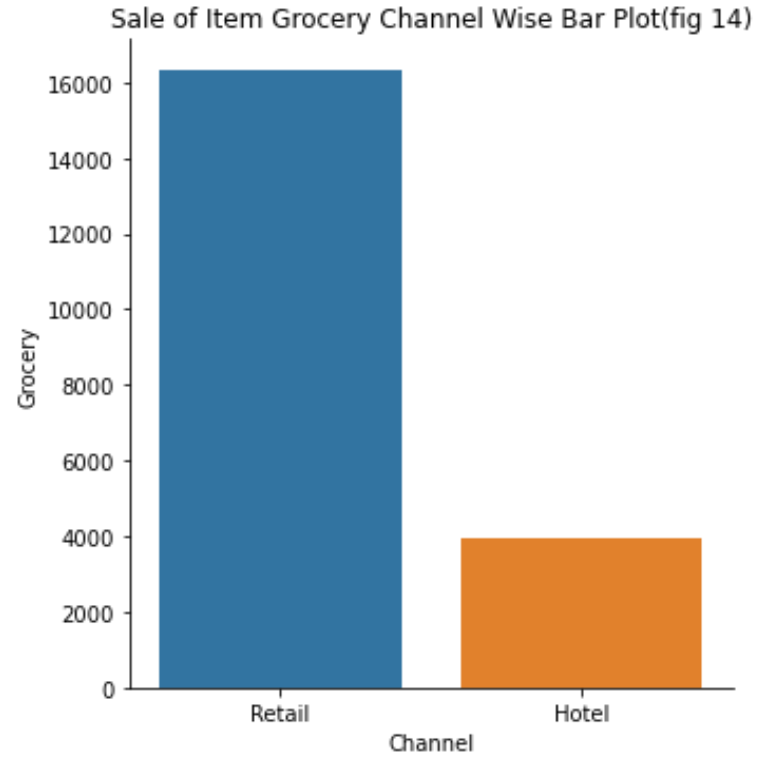
1.The Channel which has spent Highest is the Retail Channel with Other Regions spending High and Oporto spending less within the Retail channel

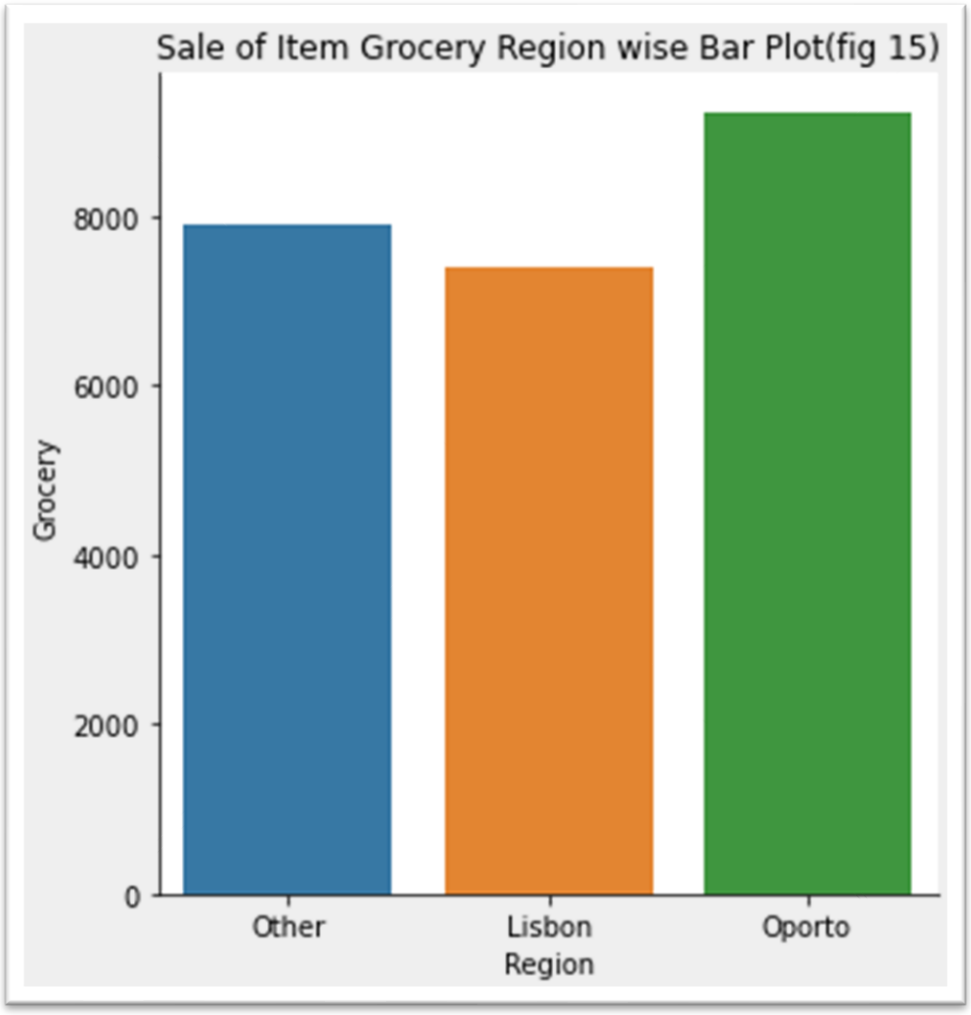
2.Whereas, the Hotel channel has spent less collectively with Lisbon spending more and Oporto spending less within the Hotel channel.

3.The same inference can be seen individually in Fig 11 & Fig 12

Item 3: Grocery







From the above plotting it is inferred that: Refer Fig 13

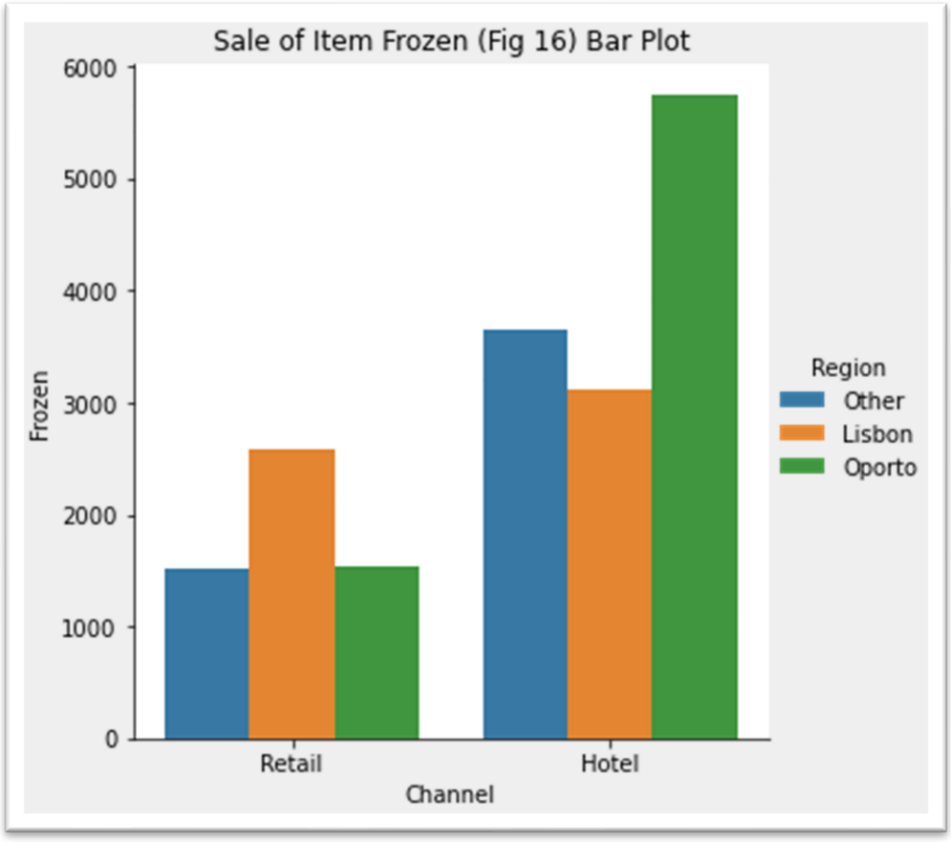
In Item Grocery:

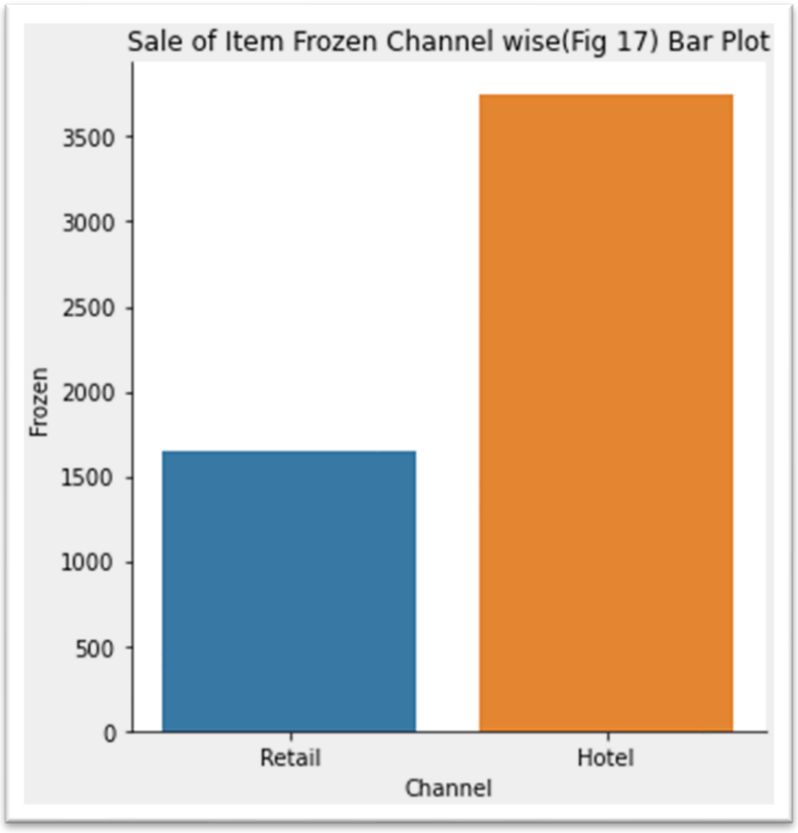
1.The Retail Channel Has Spent more on Groceries compared to Hotel Channels with Lisbon spending more and other regions spending less

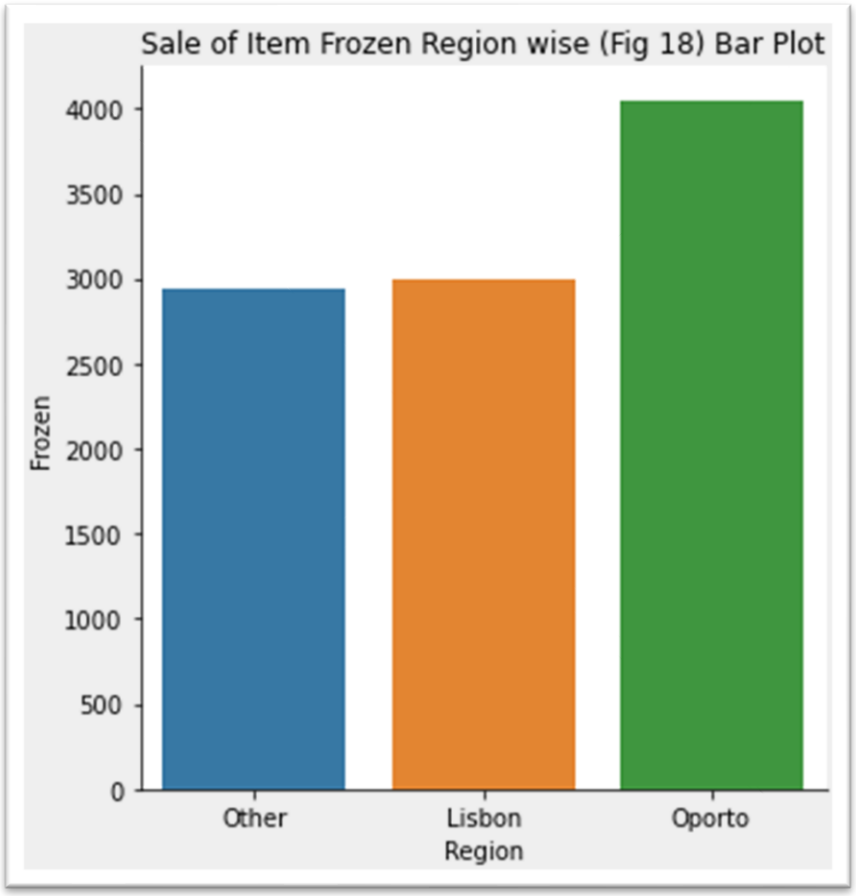
2.The Hotel channel has spent less with Oporto spending More and Other regions spending less.

3.The same can be inferred using Fig14 & Fig 15

Item 4: Frozen







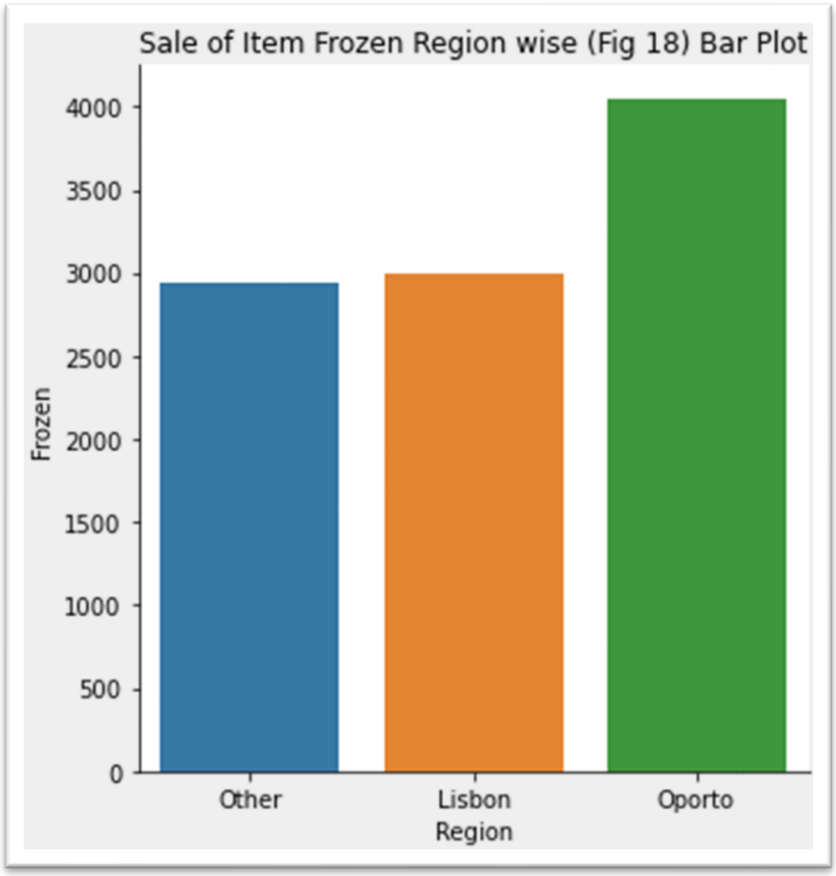
From the above plotting it is inferred that: Refer Fig 16

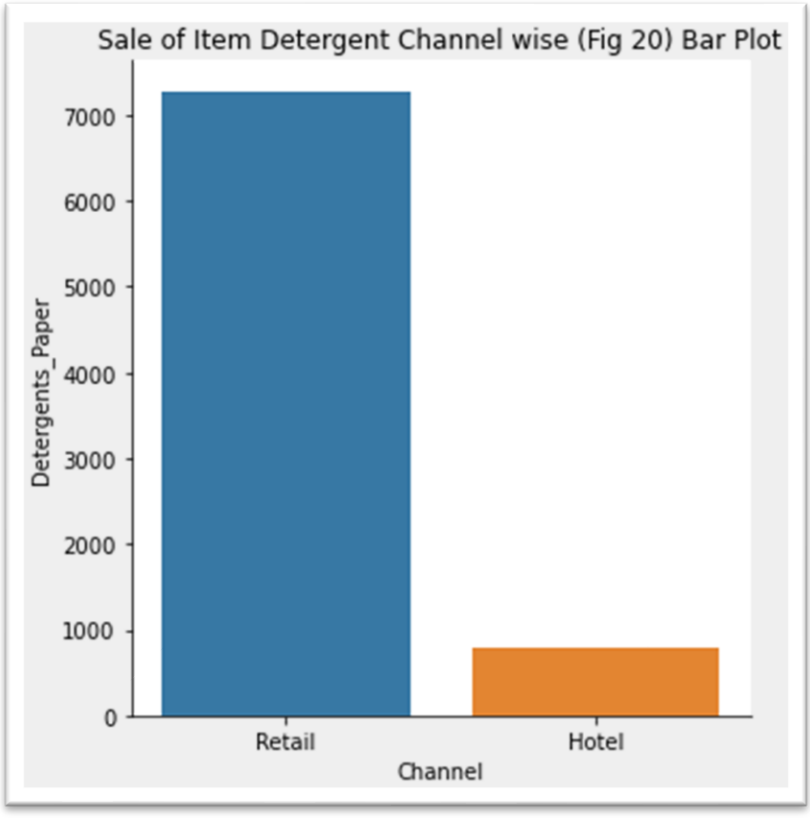
In Item Frozen:

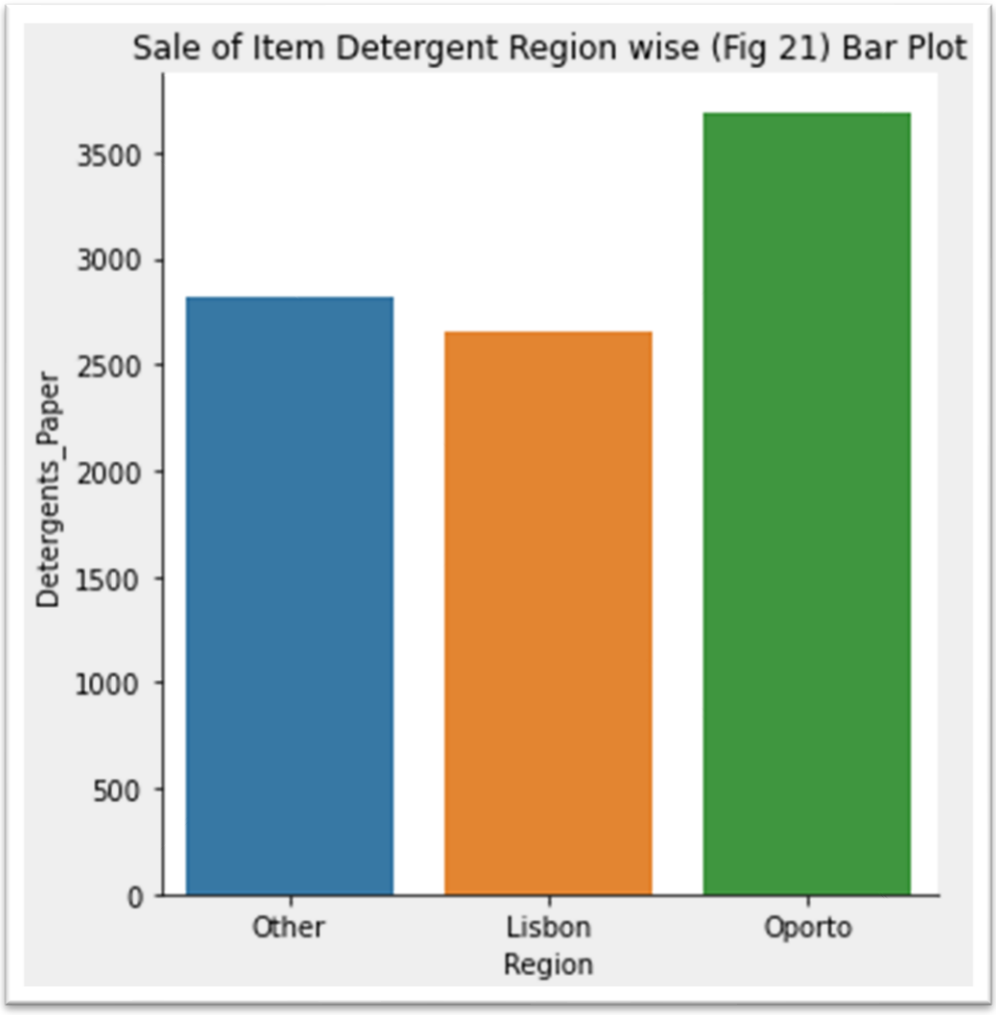
1.The Hotel Channel Has Spent more on Item Frozen compared to Retail Channels with Oporto spending more and Lisbon spending the least

2.The Retail channel has spent less with Lisbon spending More and Other regions and Oporto spending less (almost the same)

Item 5: Detergents Paper







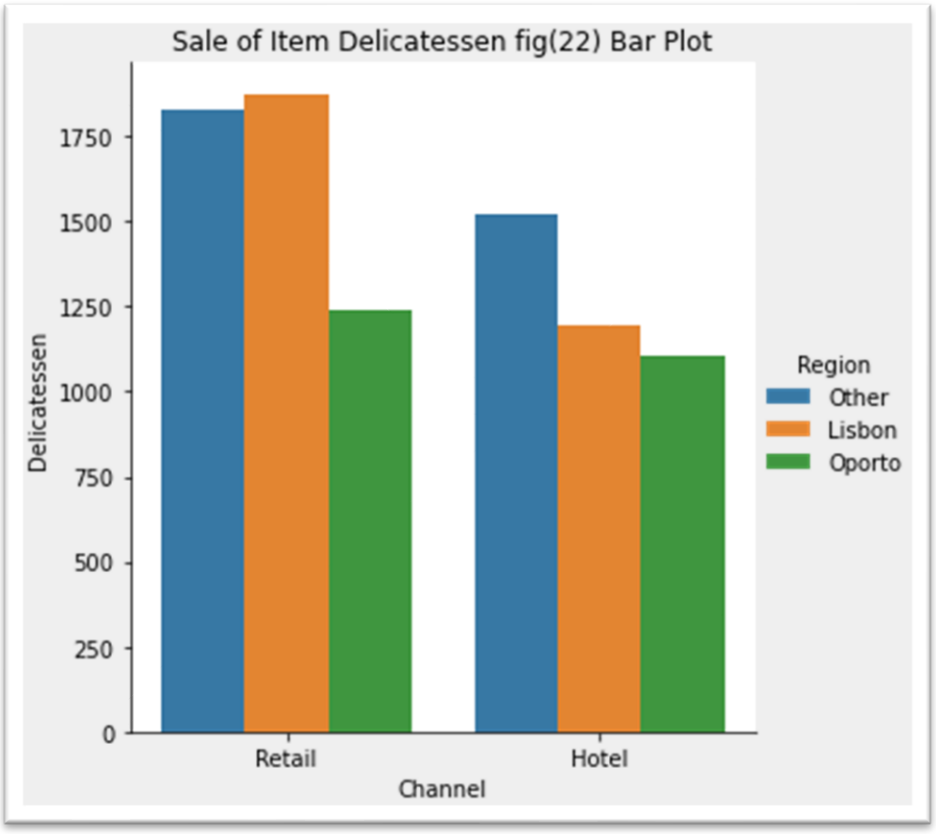
From the above plotting it is inferred that: Refer Fig 19

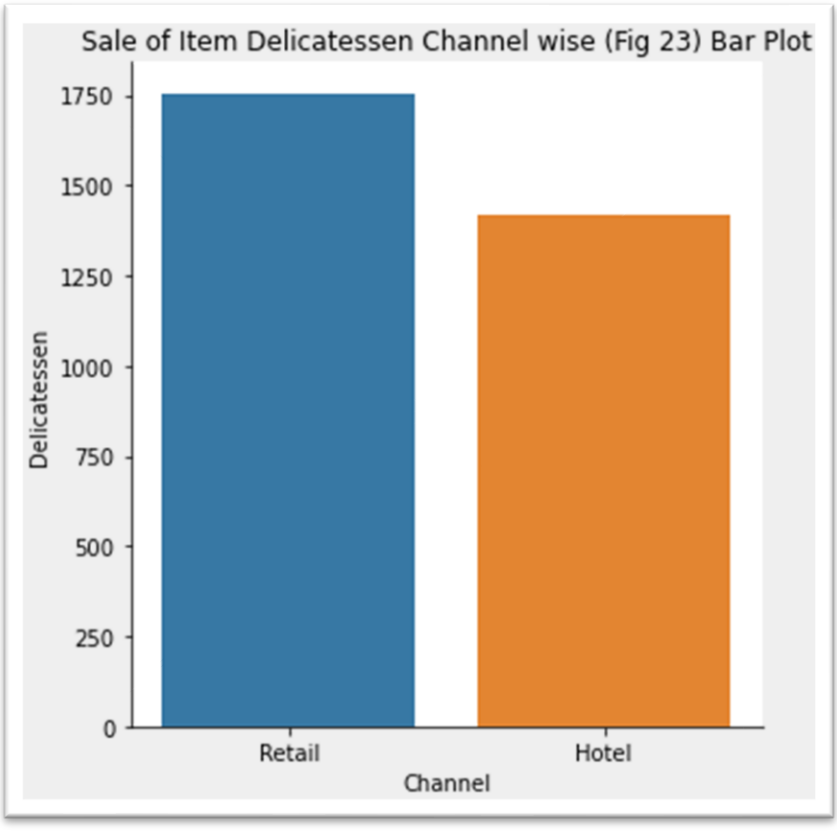
In Item Detergent Paper:

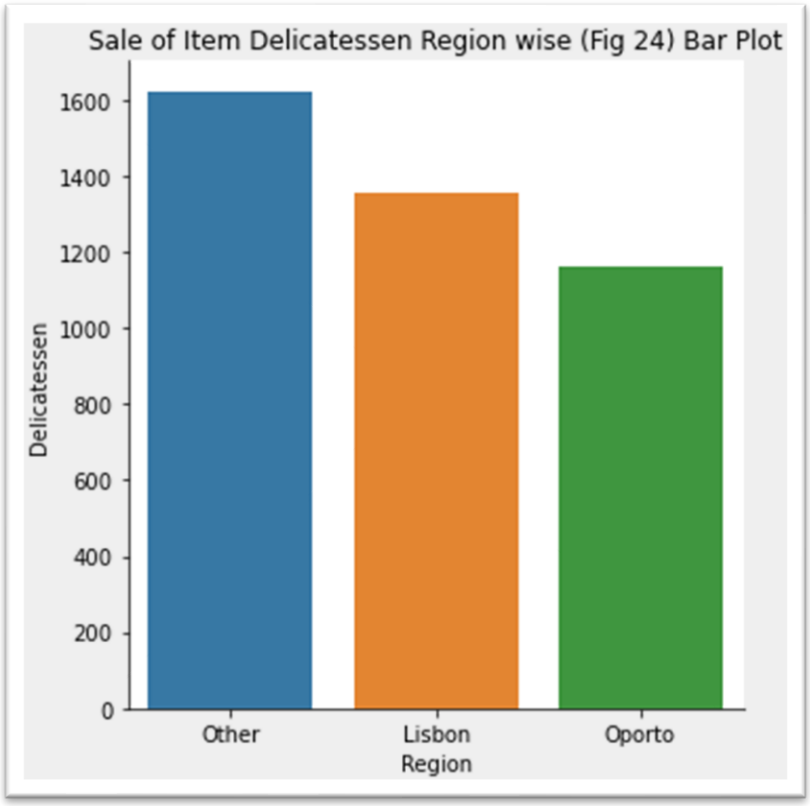
1.The Retail Channel Has Spent more compared to Hotel Channels with Oporto spending more and other regions spending the least

2.The Hotel channel has spent less with Lisbon spending More and Oporto spending least.

Item 6: Delicatessen







From the above plotting it is inferred that: Refer Fig 22

In Item Delicatessen:

1.The Retail Channel Has Spent more compared to Hotel Channels with Lisbon spending more and Oporto spending the least

2.The Hotel channel has spent less with other regions spending More and Oporto spending least.

3.The same can be inferred from Fig 23 and Fig 24

NOTE:

1. In this question it can be inferred, that Item Fresh and Frozen account for the majority sale in the Hotel channel making it the Highest spending channel.
2. Item Delicatessen has the lowest sale among all other items, probably because it is expensive and not many can afford it.
3. The Retail channel is having low sales than hotel on the whole, even though it has High sales when it comes to items like Milk, Grocery, Detergents and Delicatessen.

1.3 On the basis of a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?

a. Creating a subset called “Product” which includes the columns Region, Fresh, Milk, Grocery, Frozen, Detergents paper, Delicatessen and Total spending

	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total_Spending
0	Other	12669	9656	7561	214	2674	1338	34112
1	Other	7057	9810	9568	1762	3293	1776	33266
2	Other	6353	8808	7684	2405	3516	7844	36610
3	Other	13265	1196	4221	6404	507	1788	27381
4	Other	22615	5410	7198	3915	1777	5185	46100
...	...	...	...	...	...	...	...	...
435	Other	29703	12051	16027	13135	182	2204	73302
436	Other	39228	1431	764	4510	93	2346	48372
437	Other	14531	15488	30243	437	14841	1867	77407
438	Other	10290	1981	2232	1038	168	2125	17834
439	Other	2787	1698	2510	65	477	52	7589

Table 4: Product subset Data Frame

b. Calculating the Inconsistency among all the items using descriptive measures of variability using coefficient of variance.

c. To find the coefficient of variance we calculate the standard deviation of all the items in a subset called std.

```
The standard deviation of the items are as follows:
```

	index	0
0	Fresh	12647.33
1	Milk	7380.38
2	Grocery	9503.16
3	Frozen	4854.67
4	Detergents_Paper	4767.85
5	Delicatessen	2820.11

Table 5: Subset std Data Frame

d. Calculating the coefficient of variation by dividing the standard deviation and mean of each item individually.

- Item Fresh:

```
The coefficicent of variance of item Fresh is: 1.0527196084948245
```

- Item Milk:

```
The coefficicent of variance of item Milk is: 1.2718508307424503
```

- Item Grocery:

```
The coefficicent of variance of item Grocery is: 1.193815447749267
```

- Item Frozen:

```
The coefficicent of variance of item Frozen is: 1.5785355298607762
```

- Item Detergents Paper:

```
The coefficicent of variance of item Detergents_Paper is: 1.6527657881041729
```

- Item Delicatessen:

```
The coefficicent of variance of item Delicatessen is: 1.8473041039189306
```

As per the descriptive measures of Variability,

a. Higher the coefficient of variance (CV) Lower the inconsistency

b. Lower the coefficient of variance (CV) Higher the inconsistency

c. Higher the value of standard deviation means Higher the deviation or inconsistency

d. Lower the value of Standard deviation means Lower deviation or inconsistency.

Therefore, it can be inferred that:

1. Item Fresh has a CV of 1.05 which is low among all CVs, therefore it has inconsistent behaviour

2. Item Delicatessen has CV of 1.84 which is higher among all CVs, therefore it has a consistent behaviour

4.Item Fresh has a deviation of 12647.33 which is higher among all, meaning it has Higher Inconsistent Behaviour

5.Item Delicatessen has a deviation of 2820.11, having the least inconsistent behaviour among all.

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

a. Plotting Boxplots to find the outliers in the Product data set/ Items individually.

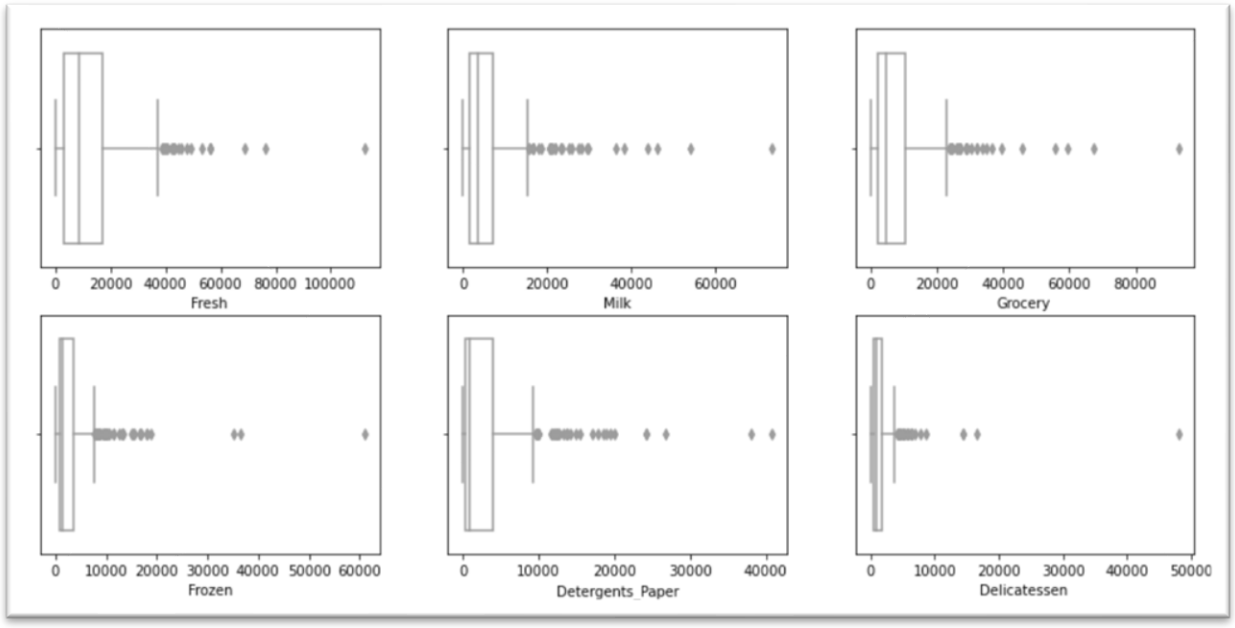


Fig 25: Boxplot of all Items for finding outliers

From the Above plotting it can be inferred that:

1. There are outliers in all the Items in the Data (Fresh, Milk, Grocery, Frozen, Detergents Paper & Delicatessen)

2. The points plotted after the whiskers represent the presence of outliers in the dataset of all the items.

3.As per the description above the outliers can be found out by calculating the inter-quartile range by using

*(IQR*1.5) +Q3=Any point above the result is an outlier.*

a. Fresh IQR = Q3-Q1 = 16933.8-3127.75 = 13,806.05, (13806.05*1.5) +16933.8 = 37642.88, *any point above this is an outlier the same is also graphically plotted in Fig 25*

b. Milk IQR = Q3-Q1 = 7190.25-1533 = 5657.25, (5657.25 * 1.5) +7190.25 = 15676.12

c. Grocery IQR = 10655.8-2153 = 8502.8, (8502.8 * 1.5) +10655.8 = 23410.00

d. Frozen IQR = 3554.25-742.25 = 2812, (2812 * 1.5) +3554.25 = 7772.25

e. Detergents Paper IQR = 3922-256.75 = 3665.25, (3665.25 * 1.5) + 3922 = 4471.38

f. Delicatessen = 1820.25-408.25 = 1412, (1412 * 1.5) + 1820.25 = 3938.25

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.

1. To improve the sales the Manufacture can increase the level of channels.
2. The Wholesaler can use the new features available in the market, such as E-commerce Technology, Online sales, D2C marketing, to see high sales and profit in the business
3. In the Hotel channel though the sales are less in other items apart from Frozen and Fresh, there is volume sale in these two items making Hotel channels sale higher than Retail channels. Therefore, it is recommended that the wholesaler reaches out more to hotel segment market more to improve the sale and cover this whole segment.
4. It is also recommended that the wholesaler should increase the product items so that the wholesaler can capture the market and increase the sales.
5. As per the analysis, though delicatessen has a consistent behaviour it has the lowest sales, comparatively, price reduction in this item can make it more saleable.

PROBLEM 2:

EXECUTIVE SUMMARY:

The student news service of Clear Mountain state University (CMSU) has gathered details about 62 undergraduate students, which includes the ID ,the gender of the students, there class which is either Junior or senior , Major which includes there subjects, their intention to graduate with Yes, No and Not Decided being there response, if they are Employed partly fully or unemployed, Their Salary, Social Networking, Satisfaction, Spending, Computer if they have a laptop, Desktop or a Tablet and Text messages

INTRODUCTION:

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using attributes of Inferential statistics i.e., using the contingency tables in different variables of the Dataset and knowing the possibility of a particular event. Contingency Tables are in matrix format displaying the Frequency Distribution of the variables.

DATA DESCRIPTION:

- ID: The index values / the no. of students
- Gender: Male/Female
- Class: Junior/Senior
- Major: Management, accountancy, CIS, others, International Business, Economics/Finance, Retailing/Marketing, Undecided.
- Grad Intention: Yes, No, Undecided
- GPA: Continuous Data starting from 2.3 to 3.9
- Employment: Full-Time, Part-Time, Unemployed.
- Salary: Continuous data starting from 25 to 80
- Social Networking: Continuous data starting from 0 to 4
- Spending: Continuous data starting from 100 to 1400
- Computer: Laptop, Desktop, Tablet
- Text Message: Continuous data from 0 to 900.

EXPLORATORY DATA ANALYSIS:

a. Importing the .csv file to Jupyter notebook
b. Printing the first 5 rows and last 5 rows of the dataset

Out[253]:

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

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

Table 6: Dataset Sample

c. Finding the missing values in the data set.

Python Output:

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

The Data set does not have any missing values.

d. Finding the Data type, shape of the data set.

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

The data set has 62 entries, with 62 rows and 14 columns, there are 6 columns which are Integer type, 6 are Object type and 2 are float type.

e. Summarizing the data:

:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
ID	62.0	NaN	NaN	NaN	31.5	18.041619	1.0	16.25	31.5	46.75	62.0
Gender	62	2	Female	33	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Age	62.0	NaN	NaN	NaN	21.129032	1.431311	18.0	20.0	21.0	22.0	26.0
Class	62	3	Senior	31	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Major	62	8	Retailing/Marketing	14	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Grad Intention	62	3	Yes	28	NaN	NaN	NaN	NaN	NaN	NaN	NaN
GPA	62.0	NaN	NaN	NaN	3.129032	0.377388	2.3	2.9	3.15	3.4	3.9
Employment	62	3	Part-Time	43	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Salary	62.0	NaN	NaN	NaN	48.548387	12.080912	25.0	40.0	50.0	55.0	80.0
Social Networking	62.0	NaN	NaN	NaN	1.516129	0.844305	0.0	1.0	1.0	2.0	4.0
Satisfaction	62.0	NaN	NaN	NaN	3.741935	1.213793	1.0	3.0	4.0	4.0	6.0
Spending	62.0	NaN	NaN	NaN	482.016129	221.953805	100.0	312.5	500.0	600.0	1400.0
Computer	62	3	Laptop	55	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Text Messages	62.0	NaN	NaN	NaN	246.209677	214.46595	0.0	100.0	200.0	300.0	900.0

Table 7: Dataset description

From Table 17 we can infer the following:

1. There are more females i.e., 33 Females and 29 males
2. The students age ranges between 18-26, where 50% of the students are of 21 years of age
3. There are 31 senior year students
4. Majority of students major in Retail/Marketing.
5. 28 students have Intentions to graduate
6. GPA of the students ranges from 2.3-3.9 and 50% having a GPA of 3.15
7. The student salary ranges from 25- 55.
8. Majority of the students have Laptops.
9. Majority of the students are there at least one social networking.

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

2.1.1 Gender and Major

Creating a subset called “Gender_Major”, which equates to a contingency table of Gender and Major

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

Table 8: Contingency Table of Gender and Major

2.1.2. Gender and Grad Intention:

Creating a subset called “Gender_GradIntention”, which equates to a contingency table of Gender and Grad Intention.

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

Table 9: Contingency table for Gender and Grad Intention

2.1.3. Gender and Employment

Creating a subset called “Gender_Employement”, which equates to a contingency table of Gender and Grad Employment.

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

Table 10: Contingency table for Gender and Employment

2.1.4 Gender and Computer

Creating a subset called “Gender_Computer”, which equates to a contingency table of Gender and Computer.

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

Table 11: Contingency Table for Gender and Computer

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?

a. Calculating the total no. of males and females

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

b. Printing the solution:

```
The probability that the selected candidate is a Male: 0.46774193548387094
```

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

a. Calculating the total no. of males and females

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

b. Printing the solution:

```
The probability that the selected candidate is a Female: 0.532258064516129
```

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

a. Printing the Table 8: Contingency table for Gender and Major for reference

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

b. Printing the total no. of males and females

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

c. Printing the output using the Outcomes expected divided by Total Outcomes with help of Table 8:

```
The conditional probability of different majors among male candidates is:

Probabilty of a Male Majoring in Accounting is 0.13793103448275862

Probabilty of a Male Majoring in CIS is  0.034482758620689655

Probabilty of a Male Majoring in Economic/Finance is 0.13793103448275862

Probabilty of a Male Majoring in Management is 0.20689655172413793

Probabilty of a Male Majoring in International Business is 0.06896551724137931

Probability of a Male Majoring in other is  0.13793103448275862

Probability of a Male Majoring in Retail/Marketing is 0.1724137931034483

Probabilty of a Male whose Majors are undecided is 0.10344827586206896
```

From the above, it can be inferred that

1. A male majoring in Management is more which is 20.7% approx.
2. Least is a male majoring in CIS which is approx. 3.45%

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

a. Printing the Table 8: Contingency table for Gender and Major for reference

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

b. Printing the total no. of males and females

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

c. Printing the output using the Outcomes expected divided by Total Outcomes with help of Table 8:

```
The Conditional Probability of different Majors among Female candidates is:

Probabilty of a Female Majoring in Accounting is 0.09090909090909091

Probabilty of a Female Majoring in CIS is  0.09090909090909091

Probabilty of a Female Majoring in Economic/Finance is 0.21212121212121213

Probabilty of a Female Majoring in Management is 0.12121212121212122

Probabilty of a Female Majoring in International Business is 0.12121212121212122

Probability of a Female Majoring in other is  0.09090909090909091

Probability of a Female Majoring in Retail/Marketing is 0.2727272727272727

Probabilty of Females whose Majors are undecided is 0.0
```

d. From the above, it can be inferred that
    1. The probability that a female majoring in Retail/Marketing is more which is approx. 27.27%
    2. Females majoring in Accounting, CIS and other subjects is the least with 9%.
    3. There is 0 probability that a female has not decided her majors.

2.4 Assume that the sample is 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

a. Printing table 9 Contingent table for Gender and Grad Intention for reference

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

b. Printing the output:

```
The probability that the randomly chosen student is male and Intends to Graduate is: 58.620689655172406
```

c. Inference:

1. The probability that a randomly chosen student is a male and intents to graduate is 17/29 or 58.62%

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

a. Printing table 11 Contingency table for Gender and Computer for reference

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

b. Using the formula,

P (No Laptop ∩ Female) = (Total Female- Female with Laptop)/Total Female

(33-29)/33*100) =12.12%

c. Output:

```
The probability that a randomly selected student is a female and does not have laptop is: 12.121212121212121
```

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?

a. Printing Table 10 contingency table For Gender and Employment for reference.

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

b. Since the given data for Gender and employee is not mutually exclusive,

we use P (A or B) =P(A)+P(B)-P(A∩B)

(Male Or Full Time Employment) =P(Male)+P (Full Time Employment)-P (Male and Full time Employed)

P(Male)=29/62

P (Full time employment) =10/62

P (Male and Full time employed) =7/29

((29/62+10/62) - (7/29)) *100) = 38.77%

c. Output:

```
The Probability that a randomly chosen student is a male or has full time employement is: 38.76529477196885
```

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

a. Printing Table 8 contingency table for Gender and Major for reference:

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

b. Since The given data for Gender and Major is mutually exclusive,

we use P (A or B) =P(A)+P(B)

Using the formula P (A|B) = P (A∩B)/P(B), where,

P (A)=Majoring in International business or Management =4/33 +4/33=0.24

P (B)=The student is a female =33/62=0.53

(0.24*0.53)/0.53) = 0.24/24%

Output:

```
The conditional probability that given a female student is randomly chosen, she is majoring in international business or manage
ment is: 0.24
```

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?

a. Constructing a new Contingency Table/subset data frame by dropping column Undecided and Gender_GraduationIntention1.

Grad Intention	No	Yes	All
Gender			
Female	9	11	33
Male	3	17	29
All	12	28	62

Table 12: Contingency table for Gender and Grad Intention without Undecided

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

P (A ∩ B) = P(A) * P(B)

So, P (Intent to graduate ∩ Female) = P (Intent to graduate) * P(Female)

P(Female) = 33/62 = 0.53225806

P (Intent to graduate) = 28/62 = 0.4516129

P (Intent to graduate) * P(Female) = 0.4516129 × 0.53225806 = 0.24037461

P (Intent to graduate ∩ Female) = 11/62 = 0.17741935

P (Intent to graduate) * P(Female)! = P (Intent to graduate ∩ Female) i.e., 0.24037461! = 0.17741935

This is not independent events as probability multiplication of both events is not equal to combined event, so having the intent to graduate and being a female are not independent events.

2.7. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages.

Answer the following questions based on the data:

2.7.1. If a student is chosen randomly, what is the probability that his/her GPA is less than 3?

a. Constructing a new contingency table or subset called Gender_GPA

GPA	2.3	2.4	2.5	2.6	2.8	2.9	3.0	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	All
Gender																	
Female	1	1	2	0	1	3	5	2	4	3	2	4	1	2	1	1	33
Male	0	0	4	2	2	1	2	5	2	2	5	2	2	0	0	0	29
All	1	1	6	2	3	4	7	7	6	5	7	6	3	2	1	1	62

Table 13: Contingency table for Gender and GPA

b. Using the above contingency table, we find out that:

The probability that his/her GPA is less than 3 is: 17/62*100% = 27.42%

c. Output:

```
The probability that the random student's GPA is less than 3 is: 27.419354838709676 %
```

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

a. Constructing a new contingency table/subset named Gender_Salary

Salary	25.0	30.0	35.0	37.0	37.5	40.0	42.0	45.0	47.0	47.5	50.0	52.0	54.0	55.0	60.0	65.0	70.0	78.0	80.0	All
Gender																				
Female	0	5	1	0	1	5	1	1	0	1	5	0	0	5	5	0	1	1	1	33
Male	1	0	1	1	0	7	0	4	1	0	4	1	1	3	3	1	0	0	1	29
All	1	5	2	1	1	12	1	5	1	1	9	1	1	8	8	1	1	1	2	62

Table 14: Contingency table for Gender and salary

b. From the above contingency Table, we infer:
   1. The probability that the randomly selected male earns 50 or more is 14/29*100 = 48.27%
   2. The probability that the randomly selected Female earns 50 or more is 18/33*100= 54.54%

c. Output:

```
The conditional probability of a random male selected earns 50 or more is: 48.275862068965516 %
```

```
The conditional probability of a random Female selected earns 50 or more is: 54.54545454545454 %
```

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

Using Table 7 Data set Description for reference for finding mean, median, IQR for the numerical variables.

:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
ID	62.0	NaN	NaN	NaN	31.5	18.041619	1.0	16.25	31.5	46.75	62.0
Gender	62	2	Female	33	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Age	62.0	NaN	NaN	NaN	21.129032	1.431311	18.0	20.0	21.0	22.0	26.0
Class	62	3	Senior	31	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Major	62	8	Retailing/Marketing	14	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Grad Intention	62	3	Yes	28	NaN	NaN	NaN	NaN	NaN	NaN	NaN
GPA	62.0	NaN	NaN	NaN	3.129032	0.377388	2.3	2.9	3.15	3.4	3.9
Employment	62	3	Part-Time	43	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Salary	62.0	NaN	NaN	NaN	48.548387	12.080912	25.0	40.0	50.0	55.0	80.0
Social Networking	62.0	NaN	NaN	NaN	1.516129	0.844305	0.0	1.0	1.0	2.0	4.0
Satisfaction	62.0	NaN	NaN	NaN	3.741935	1.213793	1.0	3.0	4.0	4.0	6.0
Spending	62.0	NaN	NaN	NaN	482.016129	221.953805	100.0	312.5	500.0	600.0	1400.0
Computer	62	3	Laptop	55	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Text Messages	62.0	NaN	NaN	NaN	246.209677	214.46595	0.0	100.0	200.0	300.0	900.0

Normal Distribution:

Normal distribution is a continuous Distribution where its values are symmetrical and situated around the mean.

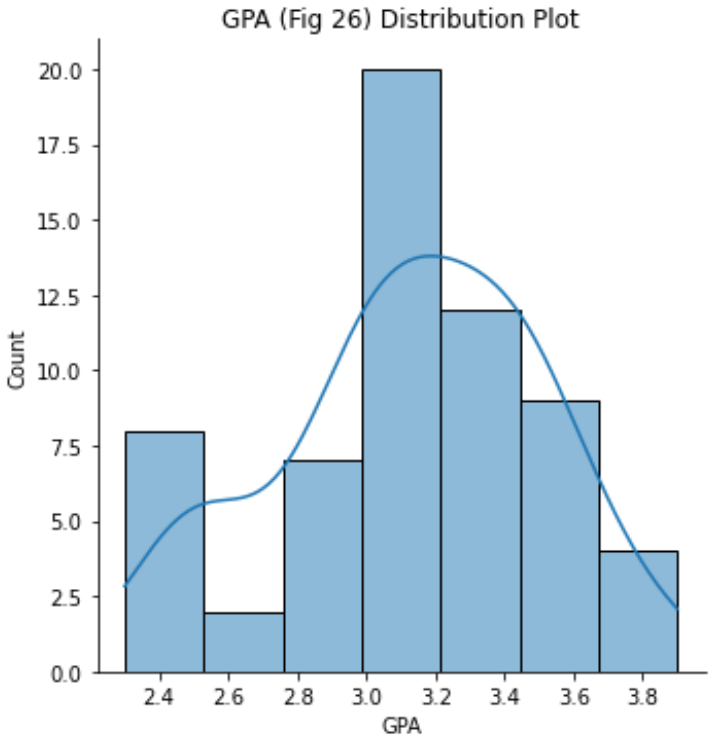
It can be found by plotting a distribution plot and check whether there is a bell curve shape

If the mean, median mode has equal values with small deviations also prove that the data is normally distributed.

But to have an accurate answer, we can use the Shapiro test where we can compare and the P-value and the Alpha

GPA:

Plotting a Distribution Plot for GPA.



Finding the mean using the descriptive table and running a code.

Mean = 3.13

Python Output:

```
The mean for GPA is 3.129032258064516
```

Finding the Standard Deviation using the descriptive Table and running a code

Standard deviation(std) = 0.37

Python Output:

```
The standard deviation for GPA is 0.37433256594525566
```

Finding the mode for GPA by running the code

```
The mode for GPA is 0    3.0
1    3.1
2    3.4
dtype: float64
```

Finding the median (Q2) from the descriptive table and by running the code

Median Q2 = 3.15

Python Output:

```
The Median for GPA is 3.1500000000000004
```

Calculating the Empirical Rule:

One standard deviation:

$\mu - \sigma$

$\mu + \sigma$

μ=GPA_mean

σ=GPA_std

$\mu - \sigma = 3.129032258064516 - 0.37433256594525566 = 2.75$

$\mu + \sigma = 3.129032258064516 + 0.37433256594525566 = 3.50$

68% of people have GPA between 2.75 and 3.50

Two Standard Deviation:

$\mu - 2\sigma = 3.129032258064516 - (2 * 0.37433256594525566) = 2.38$

$\mu + 2\sigma = 3.129032258064516 + (2 * 0.37433256594525566) = 3.88$

95% of people have GPA between 2.38 and 3.88

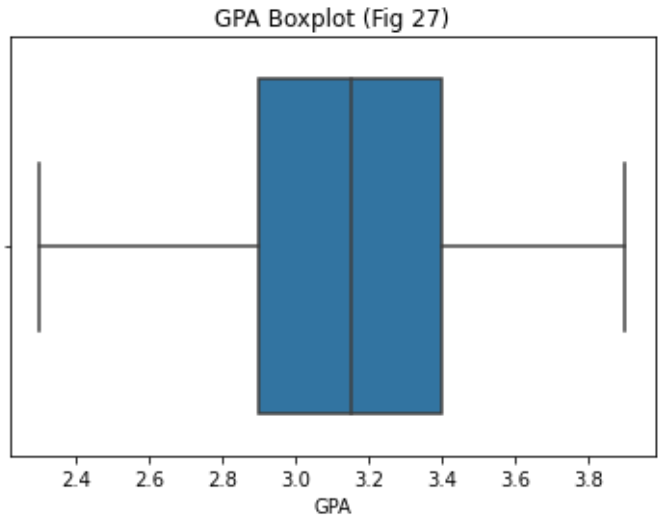
Three Standard Deviation:

$\mu - 3\sigma = 3.129032258064516 - (3 * 0.37433256594525566) = 2.006$

$\mu + 3\sigma = 3.129032258064516 + (3 * 0.37433256594525566) = 4.252$

99.7% of people have GPA between 2 and 4.25

Plotting a Boxplot to find out if there any outliers and the skewness of the variable GPA



Applying the Shapiro Test to find if GPA is normally Distributed:

H0: It is normally Distributed

HA: It is not normally distributed

Alpha=0.05

Output for Shapiro Test:

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

Conclusion:

In GPA,

1.GPA is normally distributed due to the following reasons

a. The graph (Fig 26) forms almost a bell curve with slight deviations in the graph

b. The Empirical rule is satisfied where the Mean=Median=Mode, though having very minor deviations

c. The Shapiro Test proves that the Null hypothesis is accepted with P-value being greater than alpha

2.The Boxplot (Fig 27) for GPA has no outliers

3.It is normally skewed.

4.The median is 3.15 as per the boxplot in Fig 27

5.GPA has a range of 3.9 - 2.3 = 1.6(Max-Min)

6.Q1(25%) =2.9, Q2(50%) =3.15, Q3(75%) =55.

7.IQR=Q3-Q1=55-2.9 = 52.1

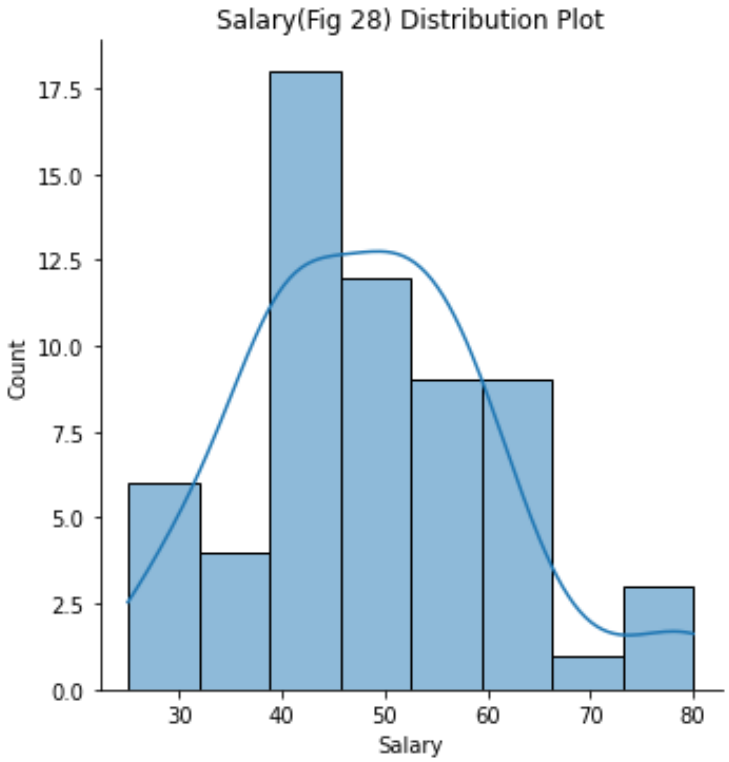
8.68% of people have GPA between 2.75 and 3.50

9.95% of people have GPA between 2.38 and 3.88

10.99.7% of people have GPA between 2 and 4.25

Salary:

Plotting a Distribution Plot to find the normal distribution by verifying if the histogram has a bell curve.



Finding the mean of salary using the descriptive table and by running the code

Mean = 48.54

Python Output:

```
The mean for Salary is 48.54838709677419
```

Finding the standard deviation using the descriptive table and running the code

Standard Deviation(std) = 12.08

Python Output:

```
The standard deviation for Salary is: 11.983089454828177
```

Finding the median using the descriptive table and running the code

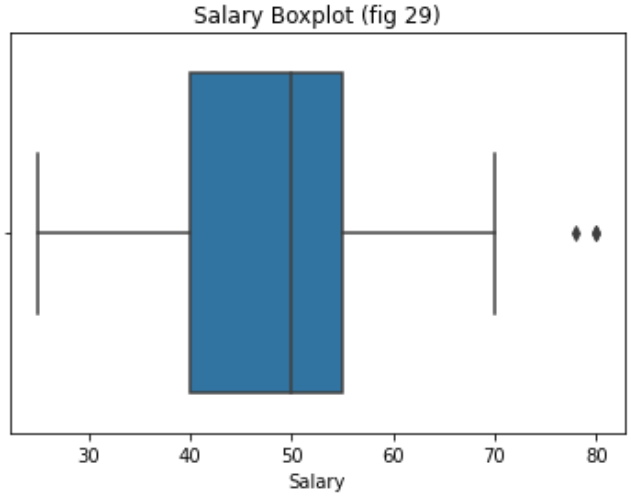
Median (Q2): 50

```
The median for salary is: 50.0
```

Finding the mode by running a code:

```
The mode for salary is 0    40.0
dtype: float64
```

Plotting a boxplot for salary to find the outliers and check whether it is normally distributed.



Applying the Shapiro Test:

H0: It is normally Distributed

HA: It is not normally distributed

Alpha=0.05

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

Calculating the Empirical Rule:

One standard deviation:

$\mu - \sigma$

$\mu + \sigma$

μ=Salary_mean

σ=Salary_std

$\mu - \sigma$ = 48.54838709677419 - 11.983089454828177 = 36.56

$\mu + \sigma$ = 3.129032258064516 + 0.37433256594525566 = 60.53

68% of people have Salary between 36.56 and 60.53

Two Standard Deviation:

$\mu - 2\sigma$ = 48.54838709677419 - (2 * 11.983089454828177) = 24.58

$\mu + 2\sigma$ = 48.54838709677419 + (2 * 11.983089454828177) = 72.51

95% of people have Salary between 24.58 and 72.51

Three Standard Deviation:

$\mu - 3\sigma$ = 48.54838709677419 - (3 * 11.983089454828177) = 12.60

$\mu + 3\sigma$ = 48.54838709677419 + (2 * 11.983089454828177) = 84.50

99.7% of people have Salary between 2 and 4.25

Conclusion:

In Salary,

1.The Salary variable is not normally distributed due to the following reasons:

a. The graph (Fig 28) forms almost a bell curve with slight deviations in the graph

b. The Empirical rule is not satisfied where the Mean! = Median! =Mode

c. The Shapiro Test proves that the Null hypothesis is rejected with P- value being lower than alpha

2.The Boxplot (Fig 29) Salary has outliers at 78 and 80

3.The salary boxplot is Left skewed.

4.Q1(25%) =40, Q2(50%) =50, Q3(75%) =55.

5.IQR=Q3-Q1=55-40 = 15

6.Range= Max-Min= 80-25=55

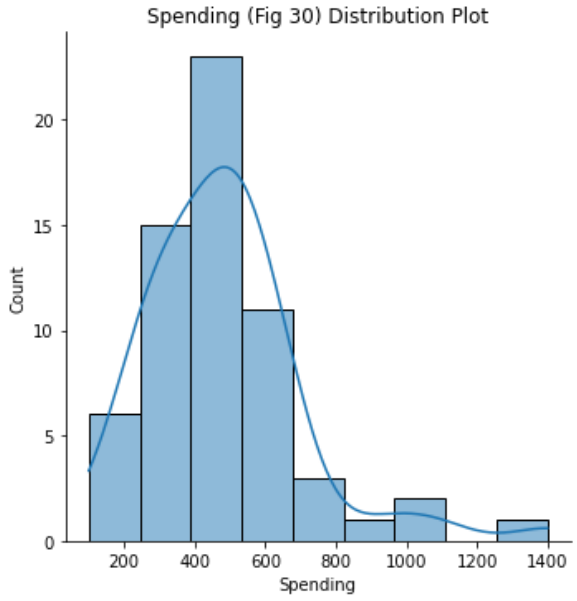
7.68% of people have Salary between 36.56 and 60.53

8.95% of people have Salary between 24.58 and 72.51

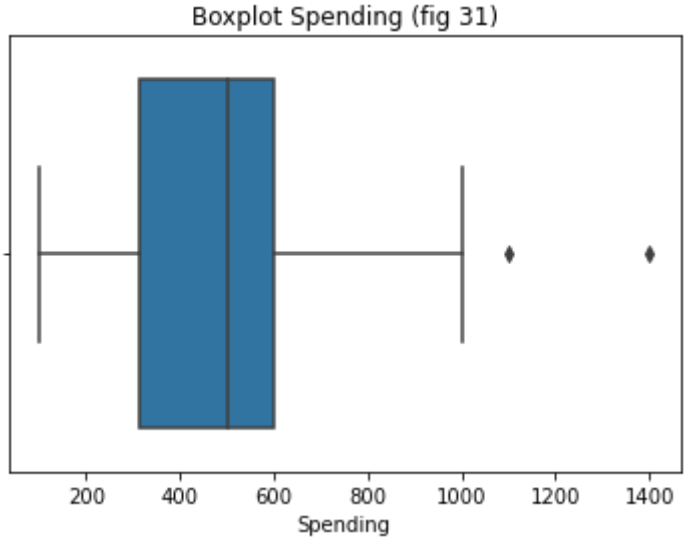
9.99.7% of people have Salary between 2 and 4.25

Spending:

Plotting a Distribution Plot to check if the shape forms a bell curve



Plotting a boxplot to find out outliers.



Finding the Mean of Spending using descriptive table and running the code:

Mean = 482.01

Python Output:

```
The mean for spending is 482.01612903225805
```

Finding the median for spending using descriptive table and running the code:

Median Q2 = 500

Python Output:

```
The median for spending is 500.0
```

Finding the standard deviation using the descriptive table and running the code:

Standard Deviation(std) = 220.15

Python Output:

```
The standard deviation for spending is: 220.1565785859987
```

Finding the mode by running the code:

```
The mode for spending is 0    500
dtype: int64
```

Calculating the Empirical Rule:

One standard deviation:

$\mu - \sigma$

$\mu + \sigma$

μ=Salary_mean

σ=Salary_std

$\mu - \sigma$ = 482.01612903225805 - 220.1565785859987 = 261.86

$\mu + \sigma$ = 3.129032258064516 + 0.37433256594525566 = 702.17

68% of people have Salary between 261.86 and 702.17

Two Standard Deviation:

$\mu - 2\sigma$ = 482.01612903225805 - (2*220.1565785859987) = 41.70

$\mu + 2\sigma$ = 482.01612903225805 + (2*220.1565785859987) = 922.33

95% of people have Salary between 41.70 and 922.33

Three Standard Deviation:

$\mu - 3\sigma$ = 482.01612903225805 - (3*220.1565785859987) = -178.45

$\mu + 3\sigma$ = 482.01612903225805 + (3*220.1565785859987) = 1142.48

99.7% of people have Salary between -178.45 and 1142.48

Applying the Shapiro Test:

H0: It is normally Distributed

HA: It is not normally distributed

Alpha=0.05

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

Conclusion:

In Spending,

1.The Spending variable is not normally distributed due to the following reasons:

a. The graph (Fig 30) does not form a bell curve bell curve, it is Right skewed

b. The Empirical rule is not satisfied where the Mean! = Median=Mode

c. The Shapiro Test proves that the Null hypothesis is rejected with P- value being lower than alpha

2.The Boxplot (Fig 31) Spending has outliers

3.The spending boxplot is Left skewed.

4.Q1(25%) =312.5, Q2(50%) =500, Q3(75%) =600.

5.IQR=Q3-Q1=600-312.50 = 287.5

6.Range= Max-Min= 1400-100=1300

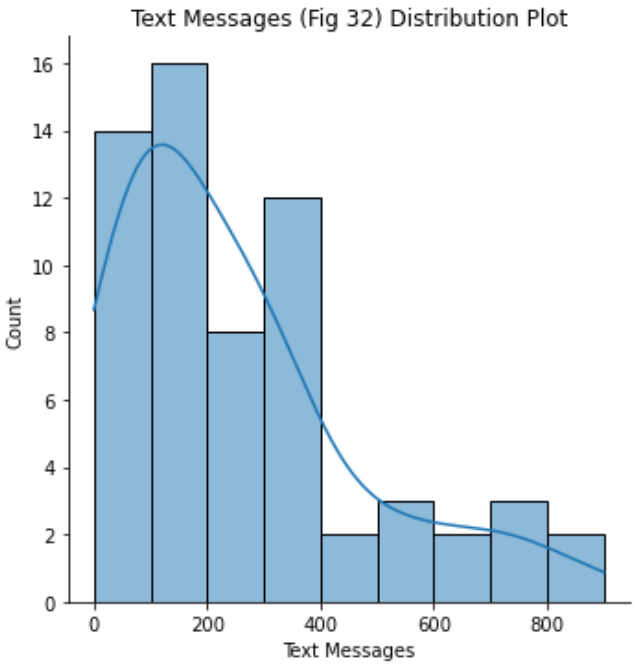
7.68% of people have Salary between 261.86 and 702.17

8.95% of people have Salary between 41.70 and 922.33

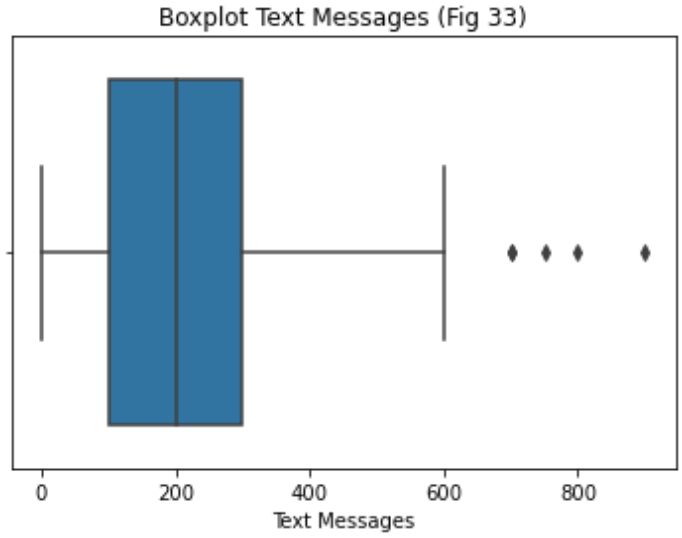
9. 99.7% of people have Salary between -178.45 and 1142.48

Text Messages:

Plotting the Distribution Plot to check if it forms a bell curve



Plotting a boxplot to find out if there any outliers



Finding the mean from the descriptive table and by running the code:

Mean: 246.20

Python Output:

```
The mean for text messages is: 246.20967741935485
```

Finding the median from the descriptive table and by running the code:

Median = 200

Python Output:

```
The Median for text messages is: 200.0
```

Finding the standard deviation from the descriptive table and by running the code

Standard deviation(std) = 212.72

Python Output:

```
The standard deviation for Text Messages is: 212.72935532273155
```

Finding the mode for spending by running the code:

```
The mode for Text Messages is: 0    300
dtype: int64
```

Calculating the Empirical Rule:

One standard deviation:

$\mu - \sigma$

$\mu + \sigma$

μ=Salary_mean

σ=Salary_std

$\mu - \sigma$ =246.20967741935485 - 212.72935532273155 = 33.48

$\mu + \sigma$ = 246.20967741935485 + 212.72935532273155 = 458.93

68% of people have Salary between 33.48 and 458.93

Two Standard Deviation:

$\mu - 2\sigma$ = 246.20967741935485 - (2*212.72935532273155) = 179.25

$\mu + 2\sigma$ = 246.20967741935485 + (2*212.72935532273155) = 671.67

95% of people have Salary between 179.25 and 671.67

Three Standard Deviation:

$\mu - 3\sigma$ = 246.20967741935485 - (3*212.72935532273155) = -391.98

$\mu + 3\sigma$ = 482.01612903225805 + (3*220.1565785859987) = 884.40

99.7% of people have Salary between -391.98 and 884.40

Applying the Shapiro Test:

H0: It is normally Distributed

HA: It is not normally distributed

Alpha=0.05

Output:

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

Conclusion:

In Text Message,

1.The Text Messages variable is not normally distributed due to the following reasons:

a. The graph (Fig 32) does not form a bell curve bell curve, it is Right skewed

b. The Empirical rule is not satisfied where the Mean! = Median! =Mode

c. The Shapiro Test proves that the Null hypothesis is rejected with P- value being lower than alpha

2.The Boxplot (Fig 33) Text message has outliers

3.The Text Messages boxplot is skewed normally.

4.Q1(25%) =100, Q2(50%) =200, Q3(75%) =300.

5.IQR=Q3-Q1=300-100 = 200

6.Range= Max-Min= 900-0=900

7.68% of people have Salary between 261.86 and 702.17

8.95% of people have Salary between 41.70 and 922.33

9.99.7% of people have Salary between -178.45 and 1142.48

PROBLEM 3

EXECUTIVE SUMMARY:

The Data set contains two columns A and B having the weight of Shingle A and Shingle B respectively. These Shingles are processed by going through the moisture test and then reweighed based upon the moisture taken out in pounds per 100 square feet. There a total of 36 weights of Shingle A and 31 weights of Shingle B. The mean moisture content is less than 0.35 pounds per 100 square feet.

INTRODUCTION:

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. using attributes of Hypothesis testing (One sample T-test, two sample T-test, Paired Test, Chi-square test) Hypothesis Testing helps to assess the reasonability or probability of the assumed hypothesis.

DATA DESCRIPTION:

A: The weight measurement of Shingle A, continuous Data from 0.13 to 0.72

B: The weight measurement of Shingle B, continuous Data from 0.10 to 0.58

1. Import the .csv file to Jupyter Note Book
2. Data Set sample:

	A	B
0	0.44	0.14
1	0.61	0.15
2	0.47	0.31
3	0.30	0.16
4	0.15	0.37

	A	B
31	0.40	NaN
32	0.29	NaN
33	0.43	NaN
34	0.34	NaN
35	0.37	NaN

Table 15: Data Set Sample

3. Checking The Missing Value:

```
df2.isnull().sum()
```

```
A    0
B    5
dtype: int64
```

There is no missing value in Shingle A and 5 missing Value in Shingle B

4. Checking the Data Type, Data Shape.

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

There are a total of 2 columns and 36 rows, Both the columns have a Data Type Float.

5. Summarizing the data set

	count	mean	std	min	25%	50%	75%	max
A	36.0	0.316667	0.135731	0.13	0.2075	0.29	0.3925	0.72
B	31.0	0.273548	0.137296	0.10	0.1600	0.23	0.4000	0.58

Table 15: Descriptive Table.

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.

1.Constructing a Hypothesis using Null and Alternate Hypothesis for Shingle A.

Null Hypothesis H0:mu<=0.35

Alternative Hypothesis HA: mu>0.35

Alpha=0.05

a Calculating the T Test and P value for Sample A using One sample T Test.

Python Output:

```
t statistic is: -1.4735046253382782 p value is: 0.07477633144907513
```

From The above calculation it can be Inferred that:

Since P value>0.05, do not reject H0. There is not enough evidence to conclude that 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 not less than 0.35 pounds per square feet, the probability of observing a sample of 36 Shingles that will result in a sample mean moisture of 0.3167 or 0.0748 pounds per 100 square feet.

3. Constructing a Hypothesis using Null and Alternate Hypothesis for Shingle B.

Null Hypothesis H0:mu<=0.35

Alternative Hypothesis HA: mu>0.35

Alpha=0.05

Python Output:

```
t statistic is: -3.1003313069986995 p value is: 0.0020904774003191826
```

From the above Calculation and Python Output it can be inferred that:

Since P value<0.05, reject Ho. 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 is 0. 0021.If the Population mean moisture content is not less than 0.35 pounds per square feet, the probability of observing a sample of 31 Shingles that will result in a sample mean moisture of 0.2735 or 0.0021 pounds per 100 square feet.

3.2 Do you think that the population mean for shingles A and B are equal? Form the hypothesis and conduct the test of the hypothesis. What assumption do you need to check before the test for equality of means is performed?

The assumptions made when doing a t-test include:

1. The scale of measurement,

2. Random sampling,

3. Normality of data distribution,

4. Adequacy of sample size and

5. Equality of variance in standard deviation.

a. Constructing a Hypothesis:

Ho: Mu(A)=Mu(B)

HA: Mu(A)! = Mu(B)

b. Calculating The T test and P value for the Sample A & B using Two sample T test

Python Output:

```
t statistic is: 1.2896282719661123 p value is: 0.1008748285917653
```

From the above Calculation and Python Output it can be inferred that:

Since P value>0.05, do not reject Ho, thus Null hypothesis where population mean of shingle A and shingle B are equal is true. There is enough evidence that the population mean of Shingle A and Shingle B is equal.

THE END