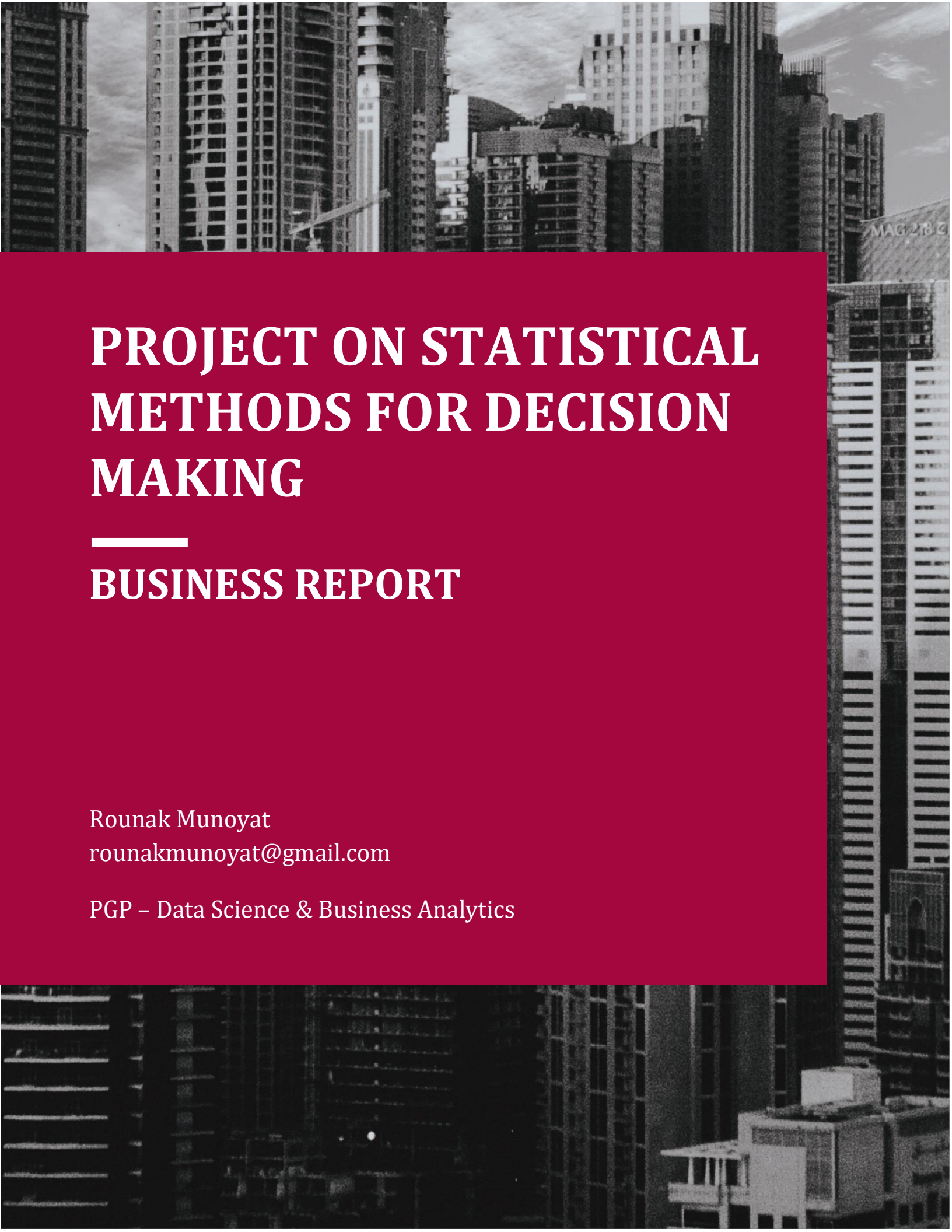

PROJECT ON STATISTICAL METHODS FOR DECISION MAKING

BUSINESS REPORT

Rounak Munoyat
rounakmunoyat@gmail.com

PGP – Data Science & Business Analytics

TABLE OF CONTENTS

Case Study-1: Wholesale Customer Analysis

Problem Statement:

A wholesale distributor operating in different regions of Portugal has information on annual spending of several items in their stores across different regions and channels. The data consists of 440 large retailers' annual spending on 6 different varieties of products in 3 different regions (Lisbon, Oporto, Other) and across different sales channel (Hotel, Retail).

Summary:

This business report provides detailed explanation on the approach to each problem definition, solution to those the problems provides some key insights/recommendations to the business.

Correlation Plot:

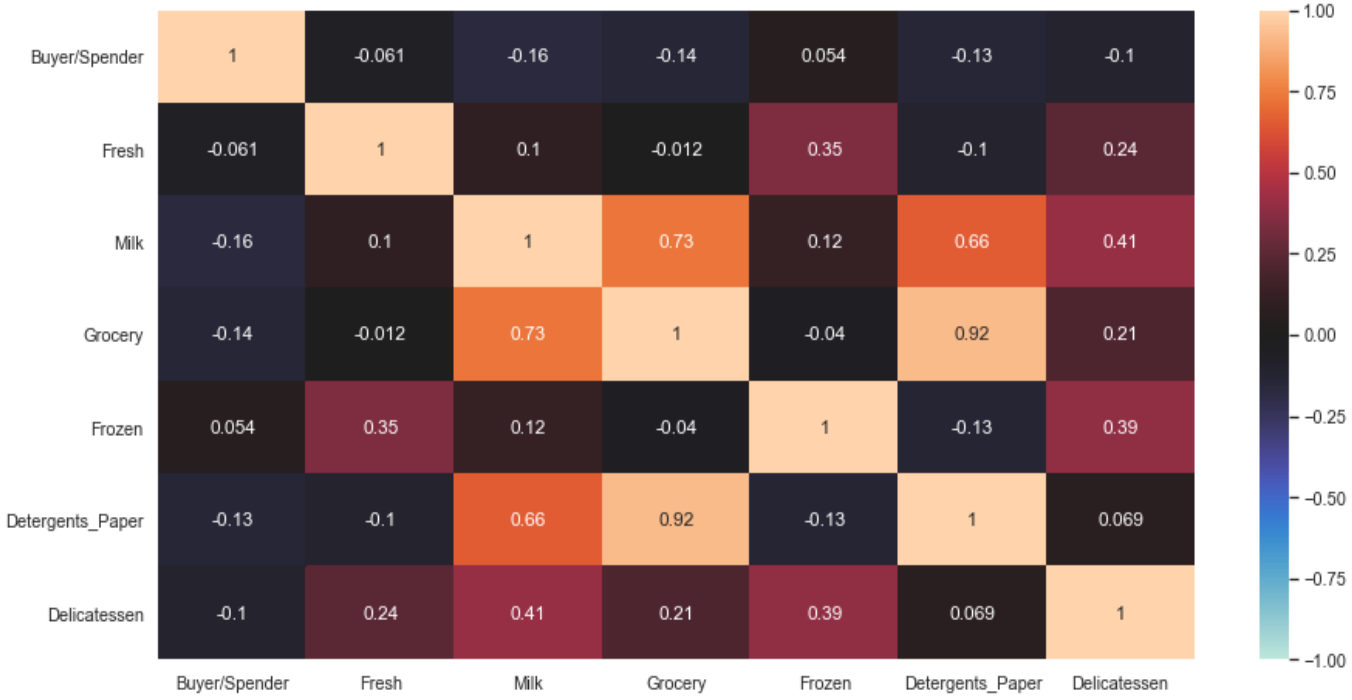

Figure 1: Correlation Plot of all the variables of the dataset

Q1.1.1) Use methods of descriptive statistics to summarize data.

Describe function helped us summarise the dataset by giving information about basic statistical parameters of the sample and measures of the data.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Buyer/Spender	440	NaN	NaN	NaN	220.5	127.161	1	110.75	220.5	330.25	440
Channel	440	2	Hotel	298	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Region	440	3	Other	316	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Fresh	440	NaN	NaN	NaN	12000.3	12647.3	3	3127.75	8504	16933.8	112151
Milk	440	NaN	NaN	NaN	5796.27	7380.38	55	1533	3627	7190.25	73498
Grocery	440	NaN	NaN	NaN	7951.28	9503.16	3	2153	4755.5	10655.8	92780
Frozen	440	NaN	NaN	NaN	3071.93	4854.67	25	742.25	1526	3554.25	60869
Detergents_Paper	440	NaN	NaN	NaN	2881.49	4767.85	3	256.75	816.5	3922	40827
Delicatessen	440	NaN	NaN	NaN	1524.87	2820.11	3	408.25	965.5	1820.25	47943

Table 1: Summary of Descriptive Statistics

Q1.1.2 & Q1.1.3) Which Region and which Channel spent the most? Which Region and which Channel spent the least?

In order to find which region and channel spent the most we first need to calculate total spending for all 6 items for all the 440 retailers. Once this is done, we can plot the graph for the same.

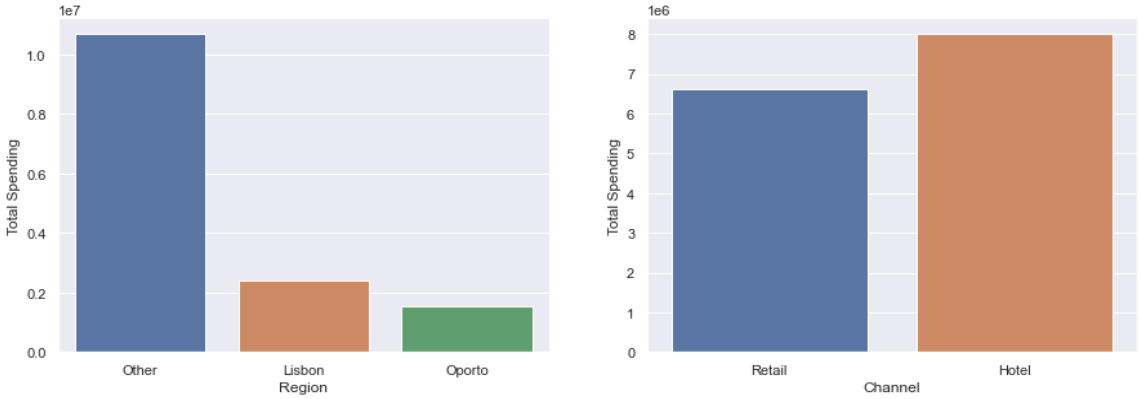

Figure 2: Graph of Region & Channel vs Total Spending respectively

From figure 2 we understand that:

- The Region that spent the most is Other and that spent the least is Oporto.
- The Channel that spent the most is Hotel and that spent the least is Retail.

Q1.2) 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.

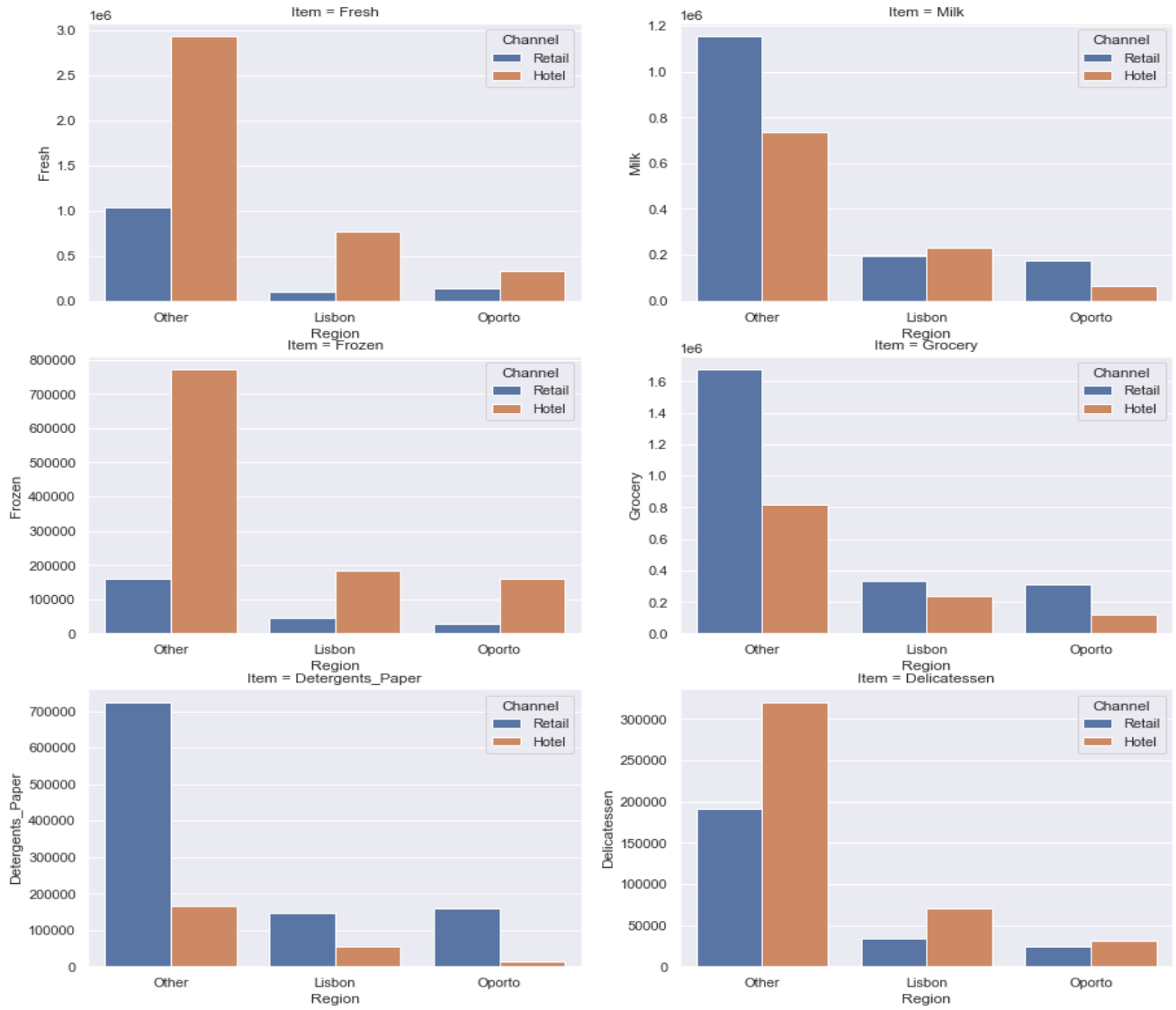

Figure 3: Graph of all 6 items vs Region (Channel being the comparing parameter for all regions)

From figure 3 we understand that:

	Lisbon		Other		Oporto	
	Hotel	**Retail**	**Hotel**	**Retail**	**Hotel**	**Retail**
Fresh	↑	↓	↑	↓	↑	↓
Milk	↑	↓	↓	↑	↓	↑
Frozen	↑	↓	↑	↓	↑	↓
Grocery	↓	↑	↓	↑	↓	↑
Detergents_Paper	↓	↑	↓	↑	↓	↑
Delicatessen	↑	↓	↑	↓	↑	↓

Table 2: Table Showing which Channel has highest & lowest spending in each Region for all Items

↑ indicates highest spending, ↓ indicates lowest spending

Q1.3) On the basis of the descriptive measure of variability, which item shows the most inconsistent behaviour? Which items shows the least inconsistent behaviour?

Variability describes how far apart data points lie from the centre of a distribution. It helps better predict behavioural consistency of variables. It is measured using standard deviation, variance, range, coefficient of variation & Interquartile range (IQR). However, except for IQR, all other methods are affected by the presence of outliers. Hence, they are not quite reliable to measure behavioural consistency of variables. Therefore, we use IQR to measure variability as it is not affected by outliers and it truly reflects dataset behaviour.

ITEM	INTERQUARTILE RANGE
Fresh	13806.0
Milk	5657.25
Frozen	2812.0
Grocery	8502.75
Detergents_Paper	3665.25
Delicatessen	1412.0

Table 3: Table of Items & corresponding Interquartile range values

From table 3 we understand that:

- Item 'Fresh' has the highest IQR, hence, it shows the most inconsistent behaviour.
- Item 'Delicatessen' has the lowest IQR, it shows the least inconsistent behaviour.

Q1.4) Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.

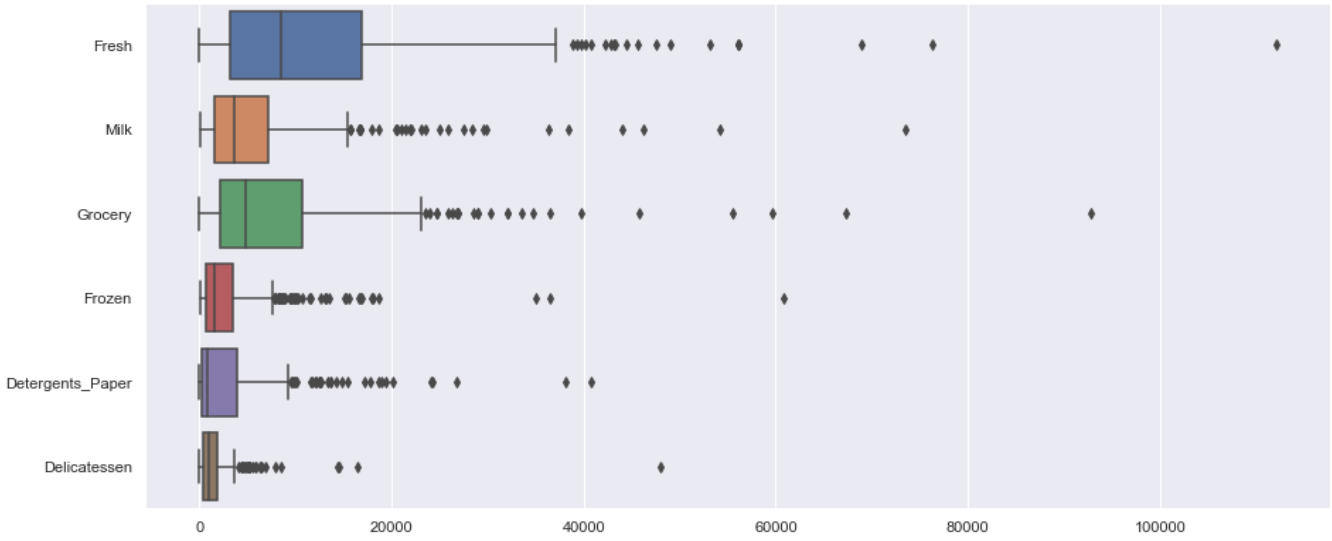

Figure 4: Graph showing outliers present for each item

From figure 4 we understand that:

- There are outliers present in all the 6 items. Hence, we need to further investigate the outlier's presence in order to take better business decisions. If ignored, can skew the information towards outliers leading to outlier biased decision making as they can affect business significantly.

Q1.5) On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem? Answer from the business perspective

As per the analysis done, we can say that:

- From the correlation plot, we saw that Grocery & Detergents_Paper is positively correlated to each other with highest strength of 0.92. We can use this correlation to improve sales of either of the two across channels/regions. We can provide incentives or discounts in Oporto region for hotel channel on Detergents_Paper when bought along with grocery to bump its spend in the region.
- Milk is strongly correlated with grocery and Detergents_Paper. For Lisbon region, milk has lower spend in retail channel compared to the hotel though. Both grocery and detergents have higher spend in retail for the same region. We can leverage these two items and provide incentives or discounts on milk alongside these two correlated items to gain bigger share in retail channel.
- Also, in order to increase sales of item Delicatessen, it could be provided for free on certain amount of purchase of grocery so that it could further be bought and tried by more & more customers down the purchasing line thereby increasing its popularity and future purchases.
- Higher the number of outliers, greater is the inconsistency. We should investigate why there are so many outliers and then figure out how can we reduce those. Looking at the correlation is also important while gauging the impact to other items. So, any decisions to improve consistency should not directly impact spend on other items.
- The spending of Hotel and Retail channel are different in almost every region, which should be more or less equal for better outcomes.

Case Study-2: CMSU Student Survey Analysis

Problem Statement:

The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. CMSU creates and distributes a survey of 14 questions and receives responses from 62 undergraduates (stored in the Survey data set).

Summary:

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

Q2.1) For this data, construct the following contingency tables (Keep Gender as row variable)

Q2.1.1) Gender and Major

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

Table 4: Contingency table of Gender & Major

Q2.1.2) Gender and Grad Intention

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

Table 5: Contingency table of Gender & Grad Intention

Q2.1.3) Gender and Employment

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

Table 6: Contingency table of Gender & Employment

Q2.1.4) Gender and Computer

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

Table 7: Contingency table of Gender & Computer

The contingency tables above help us find various conditional probabilities quite easily as the data is displayed in a manner which makes it easier to interpret & read data and perform calculations.

Q2.2) Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:

Q2.2.1) What is the probability that a randomly selected CMSU student will be Male?

We first find the total number of males & females in the dataset.
Following is the output for the number of males & females.

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

P (Randomly chosen student being male) = 29/ (29+33) = 0.4677

The probability of a randomly selected CMSU student being male is 46.77%.

Q2.2.2) What is the probability that a randomly selected CMSU student will be Female?

Similar to the question above (Q2.2.1)

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

P (Randomly chosen student being female) = 33/ (29+33) = 0.5323

The probability of a randomly selected CMSU student being female is 53.23%.

Q2.3) Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question

Q2.3.1) Find the conditional probability of different majors among the Male students in CMSU.

We will use the contingency table of Gender & Major (Table 4)

P (Accounting | Male) = 4/29 = 0.1379

P (CIS | Male) = 1/29 = 0.0345

P (Economics/Finance | Male) = 4/29 = 0.1379

P (International Business | Male) = 2/29 = 0.068

P (Management | Male) = 6/29 = 0.2068

P (Other | Male) = 4/29 = 0.1379

P (Retailing/Marketing | Male) = 5/29 = 0.1724

P (Undecided | Male) = 3/29 = 0.1034

From above we can say that:

```
The probability of males opting for the Major field 'Accounting' is 13.79 %
The probability of males opting for the Major field 'CIS' is 3.45 %
The probability of males opting for the Major field 'Economics/Finance' is 13.79 %
The probability of males opting for the Major field 'International Business' is 6.9 %
The probability of males opting for the Major field 'Management' is 20.69 %
The probability of males opting for the Major field 'Other' is 13.79 %
The probability of males opting for the Major field 'Retailing/Marketing' is 17.24 %
The probability of males who are 'Undecided' for a Major field is 10.34 %
```

Q2.3.2) Find the conditional probability of different majors among the Female students in CMSU.

We will use the contingency table of Gender & Major (Table 4)

P (Accounting | Female) = 3/33 = 0.0909

P (CIS | Female) = 3/33 = 0.0909

P (Economics/Finance | Female) = 7/33 = 0.2121

P (International Business | Female) = 4/33 = 0.1212

P (Management | Female) = 4/33 = 0.1212

P (Other | Female) = 3/33 = 0.0909

P (Retailing/Marketing | Female) = 9/33 = 0.2727

P (Undecided | Female) = 0/33 = 0.0

From above we can say that:

```
The probability of females opting for the Major field 'Accounting' is 9.09 %
The probability of females opting for the Major field 'CIS' is 9.09 %
The probability of females opting for the Major field 'Economics/Finance' is 21.21 %
The probability of females opting for the Major field 'International Business' is 12.12 %
The probability of females opting for the Major field 'Management' is 12.12 %
The probability of females opting for the Major field 'Other' is 9.09 %
The probability of females opting for the Major field 'Retailing/Marketing' is 27.27 %
The probability of females who are 'Undecided' for a Major field is 0.0 %
```

Q2.4) Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:

Q2.4.1) Find the probability that a randomly chosen student is a male and intends to graduate.

We will use the contingency table of Gender & Grad Intent (Table 5)

P (Student being male and intends to graduate) = 17/62 = 0.2742

From above we can say that:
The probability of a randomly chosen student being a male and intends to graduate is 27.42 %.

Q2.4.2) Find the probability that a randomly selected student is a female and does NOT have a laptop.

We will use the contingency table of Gender & Computer (Table 7)

P (Student being female and does not have laptop) = (2+2)/62 = 0.0645

From above we can say that:
The probability of a randomly chosen student being a female and does not have laptop is 6.45 %.

Q2.5) Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:

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

We will use the contingency table of Gender & Employment (Table 6)

P (Student being a male or has full time employment) = (29+10-7)/62 = 0.5161

From above we can say that:
The probability of a randomly chosen student being a male or has a full-time employment is 51.61 %.

Q2.5.2) Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management

We will use the contingency table of Gender & Major (Table 4)

P (Student majoring in international business or management | Female) = (4+4)/33 = 0.2424

From above we can say that:
Given a female student is randomly chosen, the probability of she majoring in international business or management is 24.24 %.

Q2.6) 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 graduate intention and being female are independent events?

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

Table 8: Contingency table of Gender & Grad Intention (after removing the 'undecided' column)

P (Randomly chosen student being female) = 20/40 = 0.50
P (Randomly chosen student having intention to graduate) = 28/40 = 0.70
P (Randomly chosen student being a female and intends to graduate) = 11/40=0.275
P (Being Female ∩ Intent to graduate) = 0.50 x 0.70 = 0.35

From above we can say that:
Since, P (Being Female ∩ Intent to graduate) ≠ P (Being Female) * P (Intent to graduate), we can say that the two events Graduate Intention & Being Female are not independent Events.

Q2.7) 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¶

Q2.7.1) If a student is chosen randomly, what is the probability that his/her GPA is less than 3?

In order to answer the question above we will create a contingency table of 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	Total
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
Total	1	1	6	2	3	4	7	7	6	5	7	6	3	2	1	1	62

Table 9: Contingency table of Gender & GPA

P (Randomly chosen student having GPA < 3) = (1+1+6+2+3+4)/62 = 0.2742

From above we can say that:
The probability of a randomly chosen student having GPA less than 3 is 27.42 %

Q2.7.2) Find conditional probability that a randomly selected male earns 50 or more. Find conditional probability that a randomly selected female earns 50 or more.

In order to answer the question above we will create a contingency table of 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	Total
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
Total	1	5	2	1	1	12	1	5	1	1	9	1	1	8	8	1	1	1	2	62

Table 10: Contingency table of Gender & Salary

P (Randomly selected male earning => 50) = (4+1+1+3+3+1+0+0+1)/29 = 0.4828
P (Randomly selected female earning => 50) = (5+0+0+5+5+0+1+1+1)/33= 0.5455

From above we can say that:
The probability of a randomly selected male earning 50 or more is 48.28 %
The probability of a randomly selected female earning 50 or more is 54.55 %

Q2.8.1) Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending and Text Messages. For each of them comment whether they follow a normal distribution.

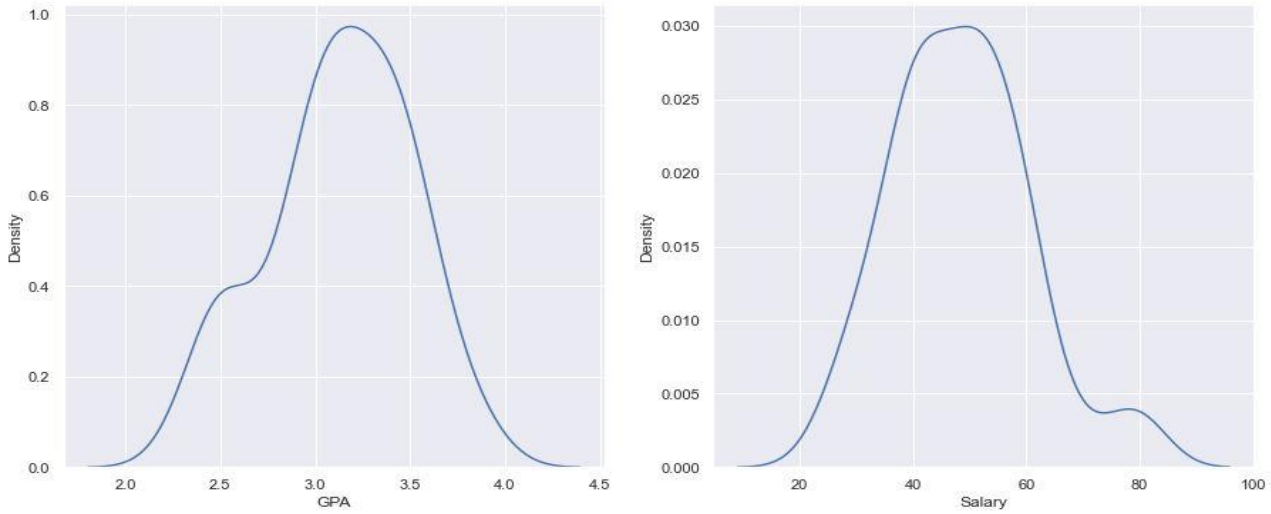

Figure 5: Distribution graphs for GPA & Salary

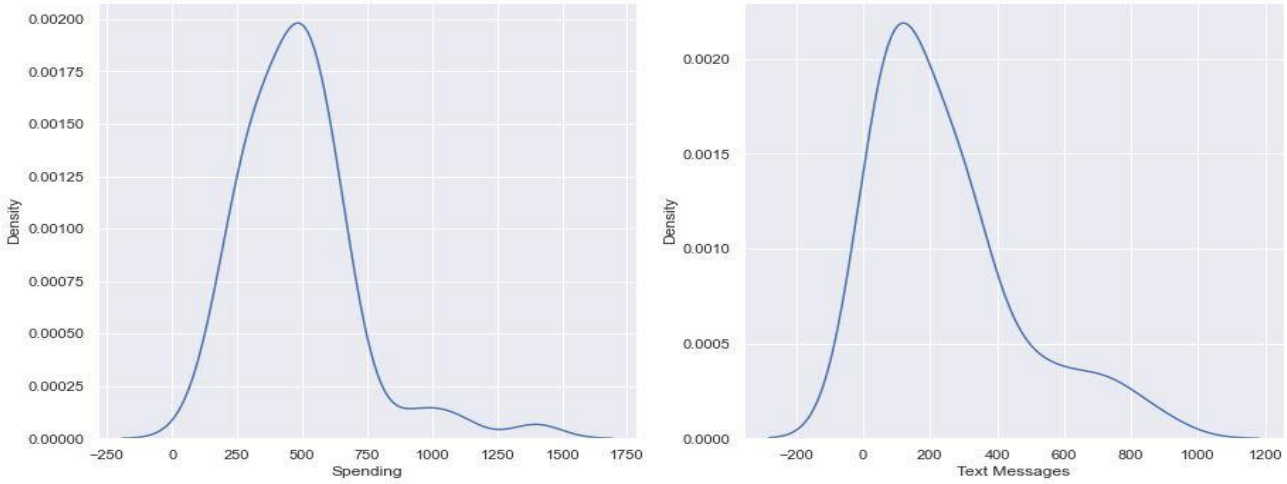

Figure 6: Distribution graphs for GPA & Salary

However, in order to check if the above variables follow normal distribution or not, we use Shapiro Wilk Test for normality. If p-value>0.05 it follows normal distribution, if not, it doesn't follow normal distribution.

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

From above we can say that:
Only GPA shows normal distribution as it has p-value > 0.05.
Salary, Spending and Text Messages do not show normal distribution as they have p-value < 0.05.

Q2.8.2) Write a note summarizing your conclusions

From all the above information we analyzed, we can say that:

1. Since GPA shows normal distribution,
   - It is symmetric.
   - The mean, median, and mode are all equal.
   - Half of the population is less than the mean and half is greater than the mean.
   - The Empirical Rule allows you to determine the proportion of values that fall within certain distances from the mean.
2. Since Salary, Spending & Text messages do not show normal distribution,
   - They are skewed.
   - The mean, median and mode are all different.
   - Population distribution is uneven about the mean.
   - It does not follow empirical rule hence we cannot determine the proportion of values that fall within certain distances from the mean.

Case Study-3: Shingle Type Analysis

Problem Statement:

An important quality characteristic used by the manufacturers of ABC asphalt shingles is the amount of moisture the shingles contain when they are packaged. Customers may feel that they have purchased a product lacking in quality if they find moisture and wet shingles inside the packaging. In some cases, excessive moisture can cause the granules attached to the shingles for texture and coloring purposes to fall off the shingles resulting in appearance problems. To monitor the amount of moisture present, the company conducts moisture tests. A shingle is weighed and then dried. The shingle is then reweighed, and based on the amount of moisture taken out of the product, the pounds of moisture per 100 square feet are calculated. The company would like to show that the mean moisture content is less than 0.35 pounds per 100 square feet.
The file (A & B shingles.csv) includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.

Summary:

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

Q3.1) Do you think there is evidence that mean moisture contents in both types of shingles are within the permissible limits? State your conclusions clearly showing all steps.

- **For Shingle A**

Step 1: Define null and alternative hypotheses

$H0$: $\mu A \geq 0.35$

HA: $\mu A < 0.35$

Step 2: Decide the significance level

Here we select $\alpha = 0.05$.

Step 3: Identify the test statistic

We do not know the population standard deviation and n = 36. So, we use the t distribution and the t_{STAT} test statistic.

Step 4: Calculate the p - value and test statistic

```
One sample t test
t statistic: -1.4735046253382782 p value: 0.14955266289815025
```

Step 5: Decide to reject or accept null hypothesis

Since p-value > 0.05, there is not enough evidence to reject the null hypothesis.

Hence, we can say that mean moisture content of Shingle type A is not within permissible limits.

- **For Shingle B**

Step 1: Define null and alternative hypotheses

$H0$: $\mu A \geq 0.35$

HA: $\mu A < 0.35$

Step 2: Decide the significance level

Here we select $\alpha = 0.05$.

Step 3: Identify the test statistic

We do not know the population standard deviation and n = 31. So, we use the t distribution and the t_{STAT} test statistic.

Step 4: Calculate the p - value and test statistic

```
One sample t test
t statistic: -3.1003313069986995 p value: 0.004180954800638365
```

Step 5: Decide to reject or accept null hypothesis

Since p-value < 0.05, there is enough evidence to reject null hypothesis.

Hence, we can say that mean moisture content of Shingle type B is within permissible limits.

Q3.2) Do you think that the population means for shingles A and B are equal? Form the hypothesis and conduct the test of the hypothesis. What assumption do you need to check before the test for equality of means is performed?

Step 1: Define null and alternative hypothesis

$H0$: $\mu A = \mu B$

HA: $\mu A \neq \mu B$

Step 2: Decide the significance level

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

Step 3: Identify the test statistic

We have two samples and sample sizes for both samples are not same.

The sample is not a large sample, n1 & n2 < 40. So, we use the t distribution and the $tSTAT$ test statistic for two sample unpaired test.

Step 4: Calculate the p - value and test statistic

The assumption that we need to check before the test for equality of means is performed is that: The variance is equal and then we compute the necessary statistical values.

```
tstat 1.2896282719661123
P Value 0.2017496571835306
```

Step 5: Decide to reject or accept null hypothesis

Since p-value > 0.05, there is not enough evidence to reject the null hypothesis

Hence, we can say that:

- The basic assumptions are that, the distributions of the two populations are normal and that the variances of the two distributions are the same.
- The population means for Shingles A & B are equal.