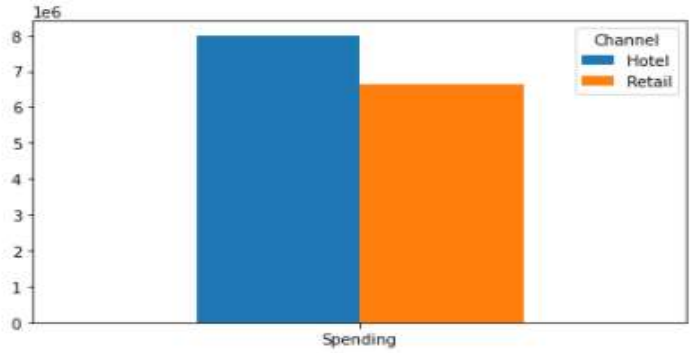
Business Report

Problem 1

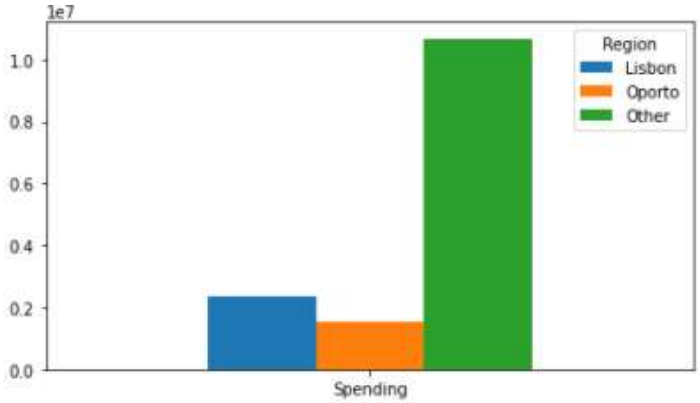
A wholesale distributor operating in different regions of Portugal has information on annual spending of several items in their stores across different regions and channels. The data consists of 440 large retailers' annual spending on 6 different varieties of products in 3 different regions (Lisbon,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

1.1 Use methods of descriptive statistics to summarize data. Which Region and which Channel seems to spend more? Which Region and which Channel seems to spend less?

Hotel channel spent $7999569 with the highest spent amount and retail channel spent $6619931 with the least spent amount based on channel.

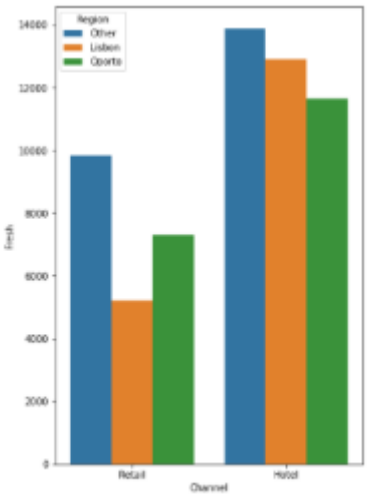


Other region spent $10677599 with the highest spent amount and Oporto region spent $1555088 with the least spent amount when grouped by region.

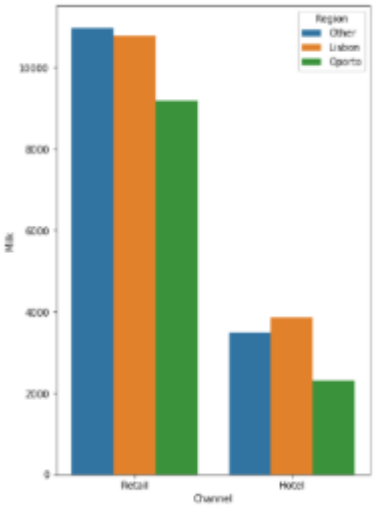


1.2 There are 6 different varieties of items are considered. Do all varieties show similar behaviour across Region and Channel? Provide justification for your answer

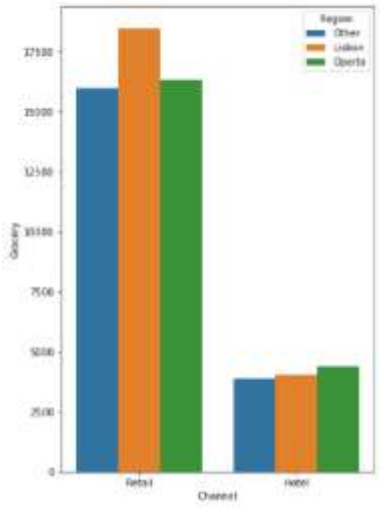
Fresh



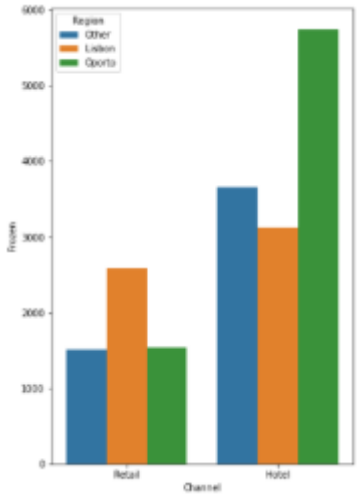
Milk



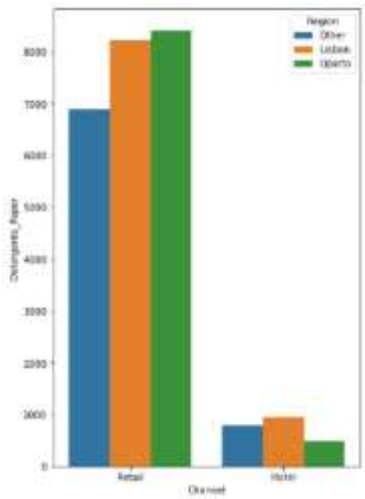
Grocery



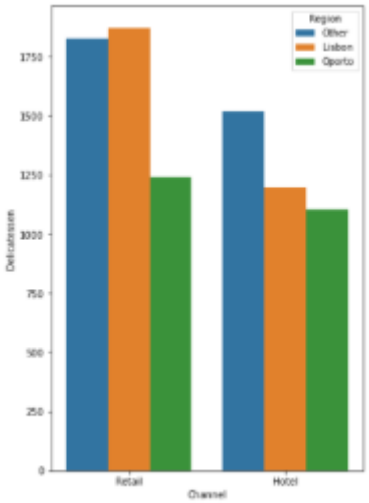
Frozen



Detergents paper



Delicatessen



From the above figures we can tell that categories like milk, grocery, delicatessen and detergent paper spend more in retail than hotel across all regions.

Frozen and fresh spend more in hotel than retail across all regions.

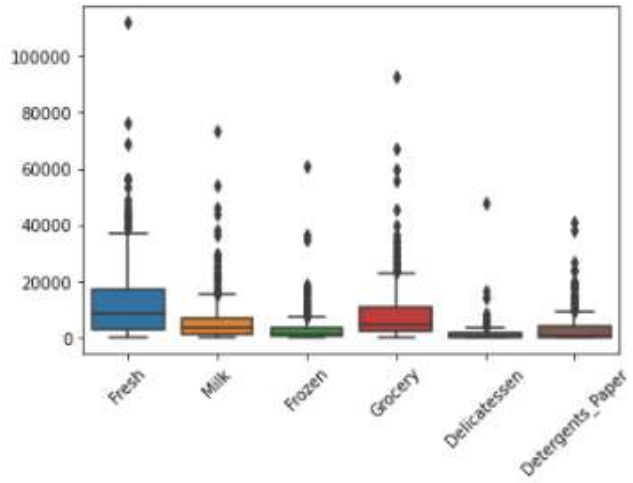
1.3 On the basis of a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?

```
Co-efficient of variation for Fresh is 1.05
Co-efficient of variation for Milk is 1.27
```

```
Co-efficient of variation for Frozen is 1.58
Co-efficient of variation for Grocery is 1.2
Co-efficient of variation for Delicatessen is 1.85
Co-efficient of variation for Detergents_Paper is 1.65
```

For comparison across different products (using covariance), the product “Fresh” has the least covariance (1.05) and the product “Delicatessen” has the highest covariance (1.85). From the above covariance we can infer that the fresh products show the highest inconsistency and delicatessen products show consistent behaviour.

1.4 Are there any outliers in the data?



Yes, all the product variables have outliers. Fresh and grocery have the highest outliers

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.

Fresh, grocery and milk seem to be very inconsistent and getting a stable spending retailer can help the distributor (by comparing covariance).

Fresh and frozen seem to be spending a lot in hotel and not in retail, in all the other variables the retail has a lot of spending which should be matched.

The spend is different across different regions and can be seen in the above figures

The detergent papers are performing poorly in the hotel channel and the spending can be matched with the retail.

Every product is different in spending across different channels which should be made equal.

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.

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	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

2.1.2. Gender and Grad Intention

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

2.1.3. Gender and Employment

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

2.1.4. Gender and Computer

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

2.2. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

	Gender
Female	33
Male	29

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

P(Male) = no of males/total number of students

```
The probability of a student picked at random being male is 0.47
```

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

P(Female) = no of females/total number of students

```
The probability of a student picked at random being female is 0.53
```

2.3. Assume that the sample is representative of the population of CMSU. Base d on the data, answer the following question:

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

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

```
The probability a male studying Accounting is 13.79 %
The probability a male studying CIS is 3.45 %
The probability a male studying Economics/Finance is 13.79 %
The probability a male studying International Business is 6.9 %
The probability a male studying Management is 20.69 %
The probability a male studying Other is 13.79 %
The probability a male studying Retailing/Marketing is 17.24 %
The probability a male studying Undecided is 10.34 %
```

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

```
The probability a female studying Accounting is 9.09 %
The probability a female studying CIS is 9.09 %
The probability a female studying Economics/Finance is 21.21 %
The probability a female studying International Business is 12.12 %
The probability a female studying Management is 12.12 %
The probability a female studying Other is 9.09 %
The probability a female studying Retailing/Marketing is 27.27 %
The probability a female studying Undecided is 0.0 %
```

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.

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

P(male ∩ intends to graduate) = P(male)*P(Graduation Intention/male)

```
The probability a random student is a male and intents to graduate
is 0.27
```

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

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

P(Female ∩ no Laptop) = P(Female) * P(No Laptop/Female)

```
The probability a random student is a female and does not have a laptop
is 0.06
```

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 either a male or has full-time employment?

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

P(Male U Full-time) = P(male)+ P(Full-time) - P(Full-time/male)

```
The probability of a person chosen randomly is either a male or has
full-time employment is 0.52
```

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

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

P(Female U international business or Marketing) = P(International Business/Female) + P(Marketing/Female)

```
The probability that a given female student is chosen randomly is
majoring in international business or management 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?

Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

The graduate intention and being a female are not independent .

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

Answer the following questions based on the data:

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

Less than 3	False	True	All
Gender			
Female	25	8	33
Male	20	9	29
All	45	17	62

```
The probability that the GPA is less than 3 is 0.27419354838709675
```

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.

Salary 50 or more	False	True	All
Gender			
Female	15	18	33
Male	15	14	29
All	30	32	62

P(Male U Salary>50) = P(Salary>50/Male)

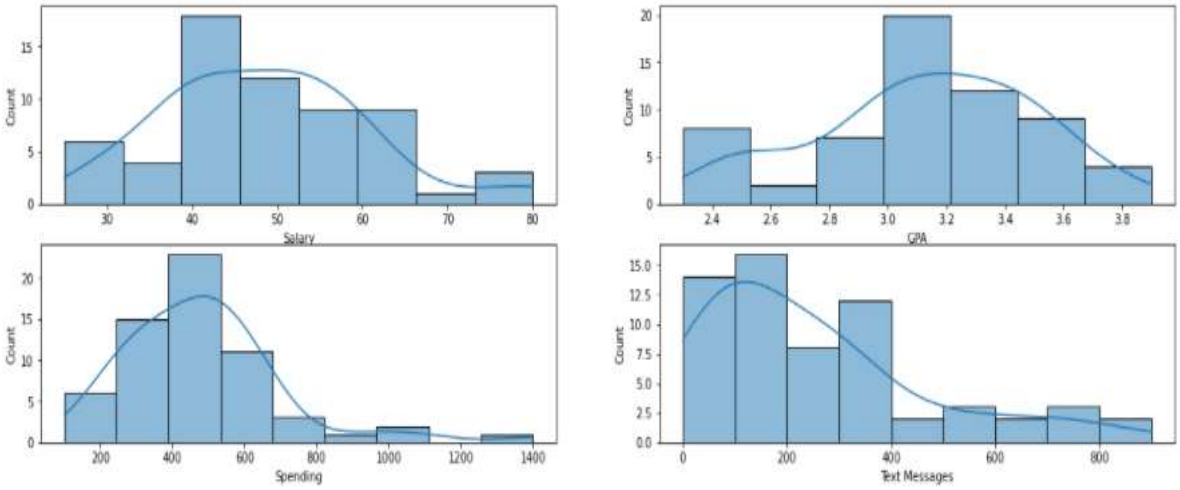
```
The probability that a male student salary is 50 or more is 0.48
```

P(Female U Salary>50) = P(Salary>50/Female)

```
The probability that a female student salary is 50 or more is 0.55
```

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.

To identify normality, we plot the distribution of the variables and apply shapiro-wilk test to find whether the variables are normally distributed or not




```
Salary distribution is not normal
GPA distribution is normal
Spending distribution is not normal
Text message distribution is not normal
```

From the above after performing the Shapiro-wilk test, we can tell that only GPA is normally distributed

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 colouring purposes to fall off the shingles resulting in appearance problems. To monitor the amount of moisture present, the company conducts moisture tests. A shingle is weighed and then dried. The shingle is then reweighed and based on the amount of moisture taken out of the product, the pounds of moisture per 100 square feet is calculated. The company would like to show that the mean moisture content is less than 0.35 pound per 100 square feet.

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.

Step 1: Formulate Hypothesis

H_0: $\mu \geq 0.35$

H_A: $\mu < 0.35$

Step 2: Alpha =0.05, let's consider the level of significance to be 0.05

Step 3: Performing 1 sample T-test on the sample

Step 4: Calculating the test-statistic and p-value

For Shingle A:

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

For Shingle B:

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

Step 5: Conclusion

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

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

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?

T-test is used since population mean is known and population standard deviation is unknown. Both the value has a before and after effect, so they are paired. And hence the 2 sampled pair T-test is conducted. The alpha is assumed to be 0.05 with a significance level of 95%.

Step 1: Formulate Hypothesis

H_0: $\mu_A = \mu_B$

H_A: $\mu_A \neq \mu_B$

Step 2: Alpha =0.05, let's consider the level of significance to be 0.05

Step 3: Performing 2 sampled paired T-test on the sample

Step 4: Calculating the test-statistic and p-value

```
Two sample paired t test
t statistic: 0.8445012483270872 p value: 0.4050738703654352
```

Step 5: Conclusions

```
We have no evidence to reject the null hypothesis since p value
0.40507 > 0.05 alpha
```

Since p-value > 0.05, do not reject H_0. So, we can say that population mean for shingles A and shingles B are equal.