**Business Report**

**Introduction:**

Problem1-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 ([Wholesale Customer.csv](https://olympus.greatlearning.in/courses/9064/files/629524/download?verifier=IYKYNzRyvIwkjMAJaeF7DOTsKBKgW4aRaufyl5Uv&wrap=1)) 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/Restaurant/Café HoReCa, Retail).

Data analysis:

Wholesale+Customers+Data Rows - 440 Columns - 9

Here list the evidence for that the dataset doesn’t contain any null values

Buyer/Spender 0

Channel 0

Region 0

Fresh 0

Milk 0

Grocery 0

Frozen 0

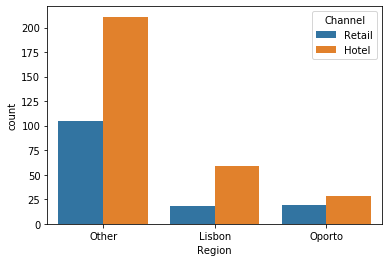
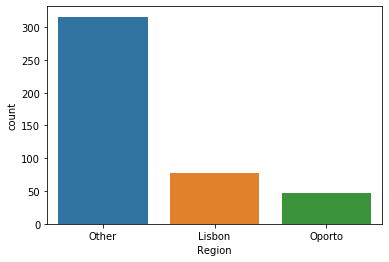
Detergents\_Paper 0

Delicatessen 0

1.Use methods of descriptive statistics to summarize data.

Which Region and which Channel seems to spend more?

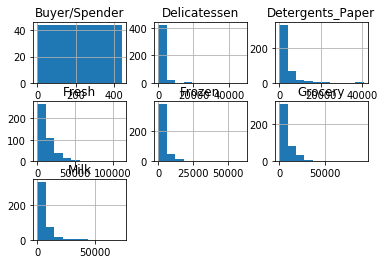
Which Region and which Channel seems to spend less?

 The visualized data clearly shows that the Region - “Other” and Channel - “Hotel” spends more and the Region - “Oporto” and Channel = “Retail” spend less.

2. There are 6 different varieties of items are considered.

Do all varieties show similar behaviour across Region and Channel?

Yes they have a similar behaviour of decreasing in order`.



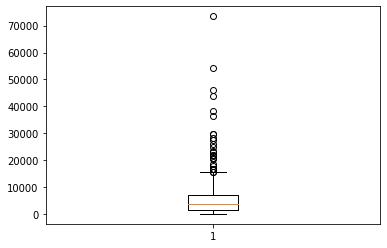
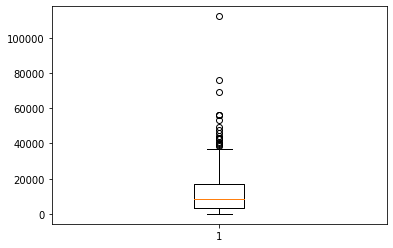
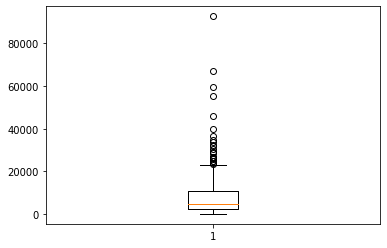
3. On the basis of the descriptive measure of variability, which item shows the most inconsistent behaviour?

Which items show the least inconsistent behaviour?

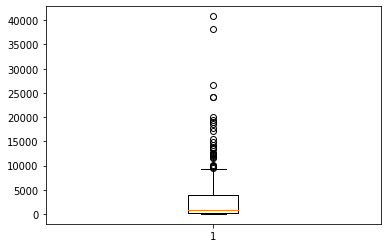
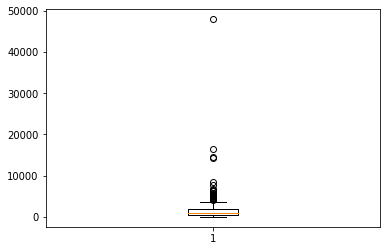
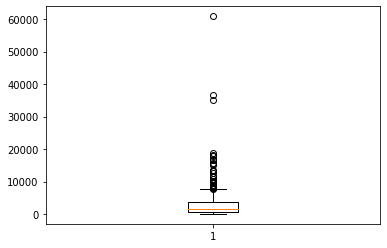
The Delicatessen is the most inconsistent behaviour and the least inconsistent behaviour is Milk.

4.Are there any outliers in the data? Yes the pictorial representation shows it clearly.

Fresh Milk Grocery

`

Frozen Detergents\_Paper Delicatessen



5. On the basis of this report, what are the recommendations?

As a result, the wholesale customer data shows that its spending on three main regions and two main channel. The dataset have six variety of items, which have the wide range of variations. The recordset have the similar behaviour of decreasing in its range as it clearly explains the sale having loss in the business. And the correlation between the items are nearby the negative correlation.

Problem 2 -The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. CMSU creates and distributes a survey of 14 questions and receives responses from 62 undergraduates (stored in the [Survey.csv](https://olympus.greatlearning.in/courses/9064/files/584888/download?verifier=B7Zbb8h1m0qZD6E1lh0AJ8YCM1ObPuZOILFZaIT2&wrap=1) file).

Data Analysis:

RangeIndex: 62 entries, 0 to 61

Data columns (total 14 columns):

ID 62 non-null int64

Gender 62 non-null object

Age 62 non-null int64

Class 62 non-null object

Major 62 non-null object

Grad Intention 62 non-null object

GPA 62 non-null float64

Employment 62 non-null object

Salary 62 non-null float64

Social Networking 62 non-null int64

Satisfaction 62 non-null int64

Spending 62 non-null int64

Computer 62 non-null object

Text Messages 62 non-null int64

1.. For this data, construct the following contingency tables (Keep Gender as row variable) 2.1.1. Gender and Major 2.1.2. Gender and Grad Intention 2.1.3. Gender and Employment 2.1.4. Gender and Computer

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

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

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

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

2.22.2.1. What is the probability that a randomly selected CMSU student will be male? 29/62=0.4677  
What is the probability that a randomly selected CMSU student will be female?33/62 =0.5322  
  
2.2.2. Find the conditional probability of different majors among the male students in CMSU.

Accounting = 7/29

CIS = 4/29

Economics/finance = 7/29

International business = 5/29

Management = 9/29

Other = 7/29

Retailing /Marketing = 7/29

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

Accounting = 3/33

CIS = 3/33

Economics/finance = 7/33

International business = 4/33

Management = 4/33

Other = 3/33

Retailing /Marketing = 9/33  
  
2.2.3. Find the conditional probability of intent to graduate, given that the student is a male. 17+9/29=0.8965  
Find the conditional probability of intent to graduate, given that the student is a female.=11+13/33=0.7272  
  
2.2.4. Find the conditional probability of employment status for the male students as well as for the female students.= 29/62+33/62 = 0.99  
  
2.2.5. Find the conditional probability of laptop preference among the male students as well as among the female students.= 55/62=0.8870

2.3. Based on the above probabilities, do you think that the column variable in each case is independent of Gender? No ,its all depends on the gender  
Justify your comment in each case.

Part II 2.4. Note that there are three numerical (continuous) variables in the data set, Salary, Spending and Text Messages. For each of them comment whether they follow a normal distribution.   
Write a note summarizing your conclusions.   
[Recall that symmetric histogram does not necessarily mean that the underlying distribution is symmetri

Salary : Mean = 48.5483 Standard Deviation = 12.0809

Comment : The data is not following a normal distribution, as it shows the left skewness among its distribution.

Spending: Mean = 482.016 Standard Deviation = 221.95

Comment : The data is slightly showing its skewness in left henceforth it’s clearly confirms that its non uniform distribution

Text Messages: Mean = 246.20 Standard Devation = 214.46

Comment : The three continuous variables does not forms a normal distribution.

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 claims that the mean moisture content cannot be greater than 0.35 pound per 100 square feet.

The file ([A & B shingles.csv](https://olympus.greatlearning.in/courses/9064/files/584889/download?verifier=3dBKeoA8vQYuRbtvU8261nYamSFTKl780zEGbXRo&wrap=1)) includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.

Data anlaysis:

Data columns (total 2 columns):

A 31 non-null float64

B 31 non-null float64

3.1. For the A shingles, form the null and alternative hypothesis to test whether the population mean moisture content is less than 0.35 pound per 100 square feet.

Step1 :It’s clear that it should not be greater than 0.35 in the moisture content.

Step 2 : Determining the statistics test - One Sample t-Test

Step 3: Sample Data - per 100 square feet

Step 4 : Finding the decision

Step 5 : Calculate p- value

T = -1.47 and the p-value is 0.14

Hence the p- value is greater than 0.05 , so that it confirms that fail to reject

Fail to reject 0.14>0.05

Step 6: Inference based on conclusion it’s fails to reject the null hypothesis.

3.2. For the B shingles, form the null and alternative hypothesis to test whether the population mean moisture content is less than 0.35 pound per 100 square feet.

Step1 :It’s clear that it should not be greater than 0.35 in the moisture content.

Step 2 : Determining the statistics test - One Sample t-Test

Step 3: Sample Data - per 100 square feet

Step 4 : Finding the decision

Step 5 : Calculate p- value

T = -3.100 and the p-value is 0.0041

Hence the p- value is less than 0.05 shows that 95 % level of significance and the p- value is less than 0.01 it confirm that tha level of significance to 99%

Fail to reject 0.041 < 0.05 and 0.041 < 0.05

Step 6: Inference based on conclusion it’s 99 % confirm that it rejects the null hypothesis.

3.3. 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?

To check for test for equality a best approach is two sample t-test

P value for shingles A and B is 0.3284 and t-statistic value is 0.9852

Hence 0.98 < 1.96 it shows that its alternate hypothesis with population mean of A is less than population mean of B.

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

Here we are using t-test in statistics to establish the values of two outcome are different from one another. Here we can assume the scale of measurement, Level of significance,normality of data distribution and equality of variance in standard deviation. In this i have find the difference between the means of two groups A and B based on a sample data.the t-test does not assume its a normally distributed data ony central limit theorem refines it. Here we concluded that the population mean for A is less than population mean for B .