SMDM PROJECT BUSINESS REPORT

DSBA

**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 ([Wholesale Customer.csv](https://olympus.greatlearning.in/courses/13133/files/983183/download?verifier=tC5pFQQUSfglzxuUmCNzPCHqB0nmBEfVtotlf53b&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, Retail).

* 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?

**Solution:**

**Channel**

* Total count = 440
* Number of unique values = 2
* Maximum count = Hotel (298)

**Region**

* Total count = 440
* Number of unique values = 3
* Maximum count = Other (316)

**Fresh**

* Total count = 440
* Mean= 12000.3
* Standard Deviation = 12647.3
* Min value = 3
* Max value = 112151
* First quartile (Q1) = 3127.75
* Median (Q2) = 8504
* Third quartile (Q3) = 16933.8
* Range = max value – min value = (112151 – 3) = 1121148
* IQR = Q3-Q1 = 16933.8 - 3127.75 = 13806.05
* Lower Limit = Q1 – 1.5\*IQR = 3127.75 – (1.5\*13806.05) = -17581.345
* Upper Limit = Q3 + 1.5\*IQR = 16933.8 + (1.5\*13806.05) =37642.875

**Milk**

* Total count = 440
* Mean= 5796.265909
* Standard Deviation = 7380.377175
* Min value = 55
* Max value = 73498.0
* First quartile (Q1) = 1533
* Median (Q2) = 3627.0
* Third quartile (Q3) = 7190.25
* Range = max value – min value = (73498– 55) = 5657.25
* IQR = Q3-Q1 = 7190.25 - 1533 = 13806.05
* Lower Limit = Q1 – 1.5\*IQR = 1533 – (1.5 \* 5657.25) = - 6952.875
* Upper Limit = Q3 + 1.5\*IQR = 7190.25 + (1.5 \* 5657.25) = 15676.125

**Grocery**

* Total count = 440
* Mean= 7951.277273
* Standard Deviation = 9503.162829
* Min value = 3
* Max value = 92780.0
* First quartile (Q1) = 2153.00
* Median (Q2) = 4755.5
* Third quartile (Q3) = 10655.75
* Range = max value – min value = (92780.0 – 3) = 92777
* IQR = Q3-Q1 = 10655.75 – 2153 = 8502.75
* Lower Limit = Q1 – 1.5\*IQR = 2153 – (1.5 \* 8502.75) = 12754.125
* Upper Limit = Q3 + 1.5\*IQR = 10655.75 + (1.5 \* 13806.05) = 20709.075

**Frozen**

* Total count = 440
* Mean= 3071.931818
* Standard Deviation = 4854.673333
* Min value = 25
* Max value = 60869.0
* First quartile (Q1) = 742.25
* Median (Q2) = 1526.0
* Third quartile (Q3) = 3554.25
* Range = max value – min value = (60869 - 25) = 60844
* IQR = Q3-Q1 = (3554.25 - 742.25) = 2812
* Lower Limit = Q1 – 1.5\*IQR = 742.25 – (1.5 \* 2812) = - 3475.75
* Upper Limit = Q3 + 1.5\*IQR = 3554.25 + (1.5 \* 2812) = 7772.25

**Detergents\_Paper**

* Total count = 440
* Mean= 2881.493182
* Standard Deviation = 4767.854448
* Min value = 3
* Max value = 40827.0
* First quartile (Q1) = 256.75
* Median (Q2) = 816.5
* Third quartile (Q3) = 3922.00
* Range = max value – min value = (40827 – 3) = 40824
* IQR = Q3-Q1 = (3922.00 - 256.75) = 3665.25
* Lower Limit = Q1 – 1.5\*IQR = 256.75 – (1.5 \* 3665.25) = - 5241.125
* Upper Limit = Q3 + 1.5\*IQR = 3922 + (1.5 \* 3665.25) = 9419.875

**Delicatessen**

* Total count = 440
* Mean= 1524.870455
* Standard Deviation = 2820.105937
* Min value = 3
* Max value = 47943.0
* First quartile (Q1) = 408.25
* Median (Q2) = 965.5
* Third quartile (Q3) = 1820.25
* Range = max value – min value = (47943 – 3) = 47940
* IQR = Q3-Q1 = (1820.25 - 408.25) = 1412
* Lower Limit = Q1 – 1.5\*IQR = 408.25 – (1.5 \* 1412) = - 1709.75
* Upper Limit = Q3 + 1.5\*IQR = 1820.25 + (1.5 \* 1412) = 3938.25
* Region which spends more is Other and Channel which spends more is Hotel
* Region which spends less is Oporto and Channel which spends less is Retail
  1. There are 6 different varieties of items are considered.  
     Do all varieties show similar behaviour across Region and Channel?

**Solution**

* Fresh is greater in Hotel channel and Other region
* Milk is greater in Retail channel and Other region
* Grocery is greater in Retail channel and Oporto region
* Frozen is greater in Hotel channel and Oporto region
* Detergents\_Paper is greater in Retail channel and Oporto region
* Delicatessen is greater in Retail channel and Other region
  1. On the basis of the descriptive measure of variability, which item shows the most inconsistent behaviour?

**Solution:**

**Based on Standard Deviation:**

Fresh has highest value of standard deviation so it shows the most inconsistent behaviour.

Delicatessen has lowest value of standard deviation so it shows the least inconsistent behaviour

**Based on Coefficient of Variation:**

Fresh has lowest value of Coefficient of variation so it shows the least inconsistent behaviour.

Delicatessen has highest value of Coefficient of variation so it shows the most inconsistent behaviour.

* 1. Are there any outliers in the data?

**Solution:**

Yes, there are outliers in all the items - Fresh, Milk, Grocery, Frozen, Detergents\_Paper &

Delicatessen

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

**Solution:**

* Region which spends more is Other and Channel which spends more is Hotel
* Region which spends less is Oporto and Channel which spends less is Retail
* Fresh is greater in Hotel channel and Other region
* Milk is greater in Retail channel and Other region
* Grocery is greater in Retail channel and Oporto region
* Frozen is greater in Hotel channel and Oporto region
* Detergents\_Paper is greater in Retail channel and Oporto region
* Delicatessen is greater in Retail channel and Other region
* Fresh has highest value of standard deviation so it shows the most inconsistent behaviour
* Delicatessen has lowest value of standard deviation so it shows the least inconsistent behaviour
* Fresh has lowest value of Coefficient of variation so it shows the least inconsistent behaviour
* Delicatessen has highest value of Coefficient of variation so it shows the most inconsistent behaviour

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/13133/files/909400/download?verifier=R4jt1Mm5hBGxj4Jkdliy6fzNmljgGRmJQqbs3AZF&wrap=1) file).

Part I

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

**Solution:**

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 a representative of the population of CMSU. Based on the data, answer the following questions:  
    
  2.2.1. What is the probability that a randomly selected CMSU student will be male?  
  What is the probability that a randomly selected CMSU student will be female?

**Solution**:

Male Probability = (Total number of male students)/ (Total number of students)

Prob\_Male = 29/62

Probability that a randomly selected CMSU student will be male 0.46774193548387094

Female Probability = (Total number of female students)/ (Total number of students)

Prob\_Female = 33/62

Probability that a randomly selected CMSU student will be female 0.532258064516129

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

**Solution:**

**Conditional probability of different majors among the male students:**

P(Accounting|Male) = P(Accounting and Male)/P(Male)

Probability that student is male and Accounting student is 0.13793103448275862

P(CIS |Male) = P(CIS and Male)/P(Male)

Probability that student is male and CIS student is 0.034482758620689655

P(Economics/Finance |Male) = P(Economics/Finance and Male)/P(Male)

Probability that student is male and Economics/Finance student is 0.13793103448275862

P(International Business |Male) = P(International Business and Male)/P(Male)

Probability that student is male and International Business student is 0.06896551724137931

P(Management |Male) = P(Management and Male)/P(Male)

Probability that student is male and Management student is 0.20689655172413793

P(Other |Male) = P(Other and Male)/P(Male)

Probability that student is male and Other student is 0.13793103448275862

P(Retailing/Marketing student |Male) = P(Retailing/Marketing student and Male)/P(Male)

Probability that student is male and Retailing/Marketing student is 0.1724137931034483

P(Undecided |Male) = P(Undecided and Male)/P(Male)

Probability that student is male and Undecided student is 0.10344827586206896

**Conditional probability of different majors among the female students:**

P(Accounting | Female) = P(Accounting and Female)/P(Female)

Probability that student is Female and Accounting student is 0.09090909090909091

P(CIS | Female) = P(CIS and Female)/P(Female)

Probability that student is Female and CIS student is 0.09090909090909091

P(Economics/Finance | Female) = P(Economics/Finance and Female)/P(Female)

Probability that student is Female and Economics/Finance student is 0.21212121212121213

P(International Business | Female) = P(International Business and Female)/P(Female)

Probability that student is Female and International Business student is 0.12121212121212122

P(Management | Female) = P(Management and Female)/P(Female)

Probability that student is Female and Management student is 0.12121212121212122

P(Other | Female) = P(Other and Female)/P(Female)

Probability that student is Female and Other student is 0.09090909090909091

P(Retailing/Marketing | Female) = P(Retailing/Marketing and Female)/P(Female)

Probability that student is Female and Retailing/Marketing student is 0.2727272727272727

P(Undecided | Female) = P(Undecided and Female)/P(Female)

Probability that student is Female and Undecided student is 0.0

2.2.3. Find the conditional probability of intent to graduate, given that the student is a male.  
Find the conditional probability of intent to graduate, given that the student is a female.

**Solution:**

P(Grad Intention | Male) = P(Grad Intention and Male)/P(Male)

Probability of intent to graduate, given that the student is a male is 0.5862068965517241

P(Grad Intention | Female) = P(Grad Intention and Female)/P(Female)

Probability of intent to graduate, given that the student is a female is 0.3333333333333333

2.2.4. Find the conditional probability of employment status for the male students as well as for the female students.

**Solution:**

**Conditional probability of employment status for the male students:**

P(Full\_Time | Male) = P(Full\_Time and Male)/P(Male)

Probability that employment status is Full\_Time for the male student is 0.2413793103448276

P(Part\_Time | Male) = P(Part\_Time and Male)/P(Male)

Probability that employment status is Part\_Time for the male student is 0.6551724137931034

P(Unemployed | Male) = P(Unemployed and Male)/P(Male)

Probability that employment status is Unemployed for the male student is 0.10344827586206896

**Conditional probability of employment status for the Female students:**

P(Full\_Time | Female) = P(Full\_Time and Female)/P(Female)

Probability that employment status is Full\_Time for the female student is 0.09090909090909091

P(Part\_Time | Female) = P(Part\_Time and Female)/P(Female)

Probability that employment status is Part\_Time for the female student is 0.7272727272727273

P(Unemployed | Female) = P(Unemployed and Female)/P(Female)

Probability that employment status is Unemployed for the female student is 0.18181818181818182

2.2.5. Find the conditional probability of laptop preference among the male students as well as among the female students.

**Solution:**

P(Laptop | Male) = P(Laptop and Male)/P(Male)

Probability of laptop preference among the male students is 0.896551724137931

P(Laptop | Female) = P(Laptop and Female)/P(Female)

Probability of laptop preference among the female students is 0.8787878787878788

* 2.3. Based on the above probabilities, do you think that the column variable in each case is independent of Gender?  
  Justify your comment in each case.

**Solution:**

According to the definition, two events are independent if and only if

P(A and B)=P(A)×P(B)

Prob\_Male = 0.46774193548387094

Prob\_Female = 0.532258064516129

**1)Gender and Major**

All the majors are dependent on gender as P(Major|Gender) is not equal to P(Major|Gender) \* P(Gnder)

**2) Gender and Graduation Intent**

As P(Grad Intention|Gender) != (P(Grad Intention|Gender) \* P(Gender), therefore Grad intention is dependent on gender

**3) Gender and Employment status**

As P(Employment|Gender) != P(Employment|Gender) \* P(Gender), therefore Employmentis dependent on gender

**4) Gender and Laptop**

As P(Laptop |Gender) != P(Laptop |Gender) \* P(Gender), therefore Laptopis dependent on gender

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 symmetric]

**Solution**:

| **Skewness** |
| --- |
| **Salary** | 0.521677 |
| **Spending** | 1.547285 |
| **Text Messages** | 1.264245 |

* Salary is very less skewed , almost follows a symmetric distribution.
* Spending is right skewed and does not follow a symmetric distribution.
* Text Messages is right skewed and does not follow a symmetric 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/13133/files/909401/download?verifier=IdywX6CyXpKMu3W4lixhDGShkWuVwqf4K0sNqD0q&wrap=1)) includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.  
  
For the A shingles, the null and alternative hypothesis to test whether the population mean moisture content is less than 0.35 pound per 100 square feet  is given:

LaTeX: H_0<=0.35

LaTeX: H_A>0.35  
  
For the B shingles, the null and alternative hypothesis to test whether the population mean moisture content is less than 0.35 pound per 100 square feet is given:

LaTeX: H_0<=0.35

LaTeX: H_A>0.35

3.1 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?

The population means for shingles A and B are equal

1. State the Null and Alternative Hypotheses

Ho: mu(A) = mu(B) ;( The moisture content of A-shingles is equal to moisture content of B-shingles)

Ha: mu(A) != mu(B) ;( The moisture content of A-shingles is not equal to moisture content of B-shingles)

MUa = 0.316667

MUb = 0.273548

alpha = 0.05

xbar\_A = 0.316666667

xbar\_B = 0.273548387

std\_dev\_A = 0.135731

std\_dev\_B = 0.137296

Sample size of A = 36

Sample size of B = 31

1. Calculate an appropriate test statistic (TS)

z = (x1-x2)/sqrt(((sd1)^2/n1) + ((sd2)^2/n2))

z = 1.2896282719661123

Now, Calculate p\_value

P\_value = 0.2017496571835306

Since p\_value > alpha , therefore failed reject Null Hypothesis.

Therefore moisture content of B-shingles is not equal to moisture content of A-shingles

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

**Assumption about the population/sample :**

MUa = 0.316667

MUb = 0.273548

alpha = 0.05

xbar\_A = 0.316666667

xbar\_B = 0.273548387

std\_dev\_A = 0.135731

std\_dev\_B = 0.137296

Sample size of A = 36

Sample size of B = 31