


# SMDM PROJECT BUSINESS REPORT

VAISHNAV U  
PGP-DSBA ONLINE  
19/10/2022



## Table of Contents

### Content

#### Problem-1

	<b>Summary</b>	<b>5</b>
	<b>Introduction</b>	<b>5</b>
	<b>Data Description &amp; EDA</b>	<b>5</b>
1.1	<b>Use methods of descriptive statistics to summarize data. Which Region and which Channel spent the most? Which Region and which Channel spent the least?</b>	<b>7</b>
1.2	<b>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.</b>	<b>11</b>
1.3	<b>On the basis of a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?</b>	<b>14</b>
1.4	<b>Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.</b>	<b>15</b>
1.5	<b>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.</b>	<b>17</b>

#### Problem-2

	<b>Summary</b>	<b>18</b>
	<b>Introduction</b>	<b>18</b>
	<b>Data Description &amp; EDA</b>	<b>18</b>
2.1	<b>For this data, construct the following contingency tables (Keep Gender as row variable)</b>	
2.1.1	<b>Gender and Major</b>	<b>21</b>
2.1.2	<b>Gender and Grad Intention</b>	<b>21</b>
2.1.3	<b>Gender and Employment</b>	<b>22</b>
2.1.4	<b>Gender and Computer</b>	<b>22</b>
2.2	<b>Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:</b>	

2.2.1	What is the probability that a randomly selected CMSU student will be male?	22
2.2.2	What is the probability that a randomly selected CMSU student will be female?	23
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.	23
2.3.2	Find the conditional probability of different majors among the female students of CMSU.	24
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.	25
2.4.2	Find the probability that a randomly selected student is a female and does NOT have a laptop.	26
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?	27
2.5.2	Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management.	28
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?	28
2.7	Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages.	29
	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?	29
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.	30
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.	30

### Problem-3

	<b>Summary</b>	<b>33</b>
	<b>Introduction</b>	<b>33</b>
	<b>Data Description &amp; EDA</b>	<b>33</b>
<b>3.1</b>	<b>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.</b>	<b>35</b>
<b>3.2</b>	<b>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?</b>	<b>37</b>

# Wholesale Customers Analysis

## Summary

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. Provided dataset will be analysed based on region and channel and proper business recommendations will be provided.

## Introduction

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 channels (Hotel, Retail). Dataset will be analysed based on descriptive statistics, various plots will be provided to shed insights into the distribution of data. The expectation is to state insights and recommendations based on analysis/output/conclusions derived as part of other questions in this problem statement.

## Data description

1	Buyer/Spender	Index number(eg. 1,2,3)
2	Channel	Hotel & Retail(Means by which items are dealt)
3	Region	Lisbon, Oporto, Other(Different regions of Portugal)
4	Fresh	Item (Spendings based on region and channel)
5	Milk	Item (Spendings based on region and channel)
6	Grocery	Item (Spendings based on region and channel)
7	Frozen	Item (Spendings based on region and channel)
8	Detergents_Paper	Item (Spendings based on region and channel)
9	Delicatessen	Item (Spendings based on region and channel)

## Exploratory Data Analysis

	Column	Non-Null Count	Data type
1	Buyer/Spender	440 non-null	int64
2	Channel	440 non-null	object
3	Region	440 non-null	object
4	Fresh	440 non-null	int64
5	Milk	440 non-null	int64
6	Grocery	440 non-null	int64
7	Frozen	440 non-null	int64
8	Detergents_Paper	440 non-null	int64
9	Delicatessen	440 non-null	int64

## Sample of dataset

Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
1	Retail	Other	12669	9656	7561	214	2674	1338
2	Retail	Other	7057	9810	9568	1762	3293	1776
3	Retail	Other	6353	8808	7684	2405	3516	7844
4	Hotel	Other	13265	1196	4221	6404	507	1788
5	Retail	Other	22615	5410	7198	3915	1777	5185

There are 440 rows and 9 columns in dataset. Channel and Region contains categorical variables and all other columns contains quantitative variables.

## Problem-1

Use methods of descriptive statistics to summarize data. Which Region

1.1 and which Channel spent the most? Which Region and which Channel spent the least?

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

### Descriptive statistics of the data shows :

- 1.Total of 440 counts are provided in data.
- 2.There are 2 unique values in channel and 3 unique values in region.
- 3.There are no null or missing values in the provided dataset.
- 4.Most occurring sales channel is Hotel(298) followed by Retail(142).
- 5.Among 3 region in Portugal the most occurring region is 'Other'(316) followed by 'Lisbon'(77) & 'Oporto'(47)
- 6.Mean,median and standard deviation of all items differs.

## Descriptive statistics of the 6 different items:

### Fresh Item

Count : 440  
 Mean : 12000.29  
 Standard deviation : 12647.32  
 Minimum: 3.0  
 25%(Q1): 3127.75  
 50% or Median: 8504.0  
 75%(Q3): 16933.75  
 Maximum: 112151.0  
 IQR(Q3-Q1) : 13806.0

### Grocery Item

Count : 440  
 Mean : 7951.27  
 Standard deviation : 9503.16  
 Minimum: 3.0  
 25%(Q1): 2153.0  
 50% or Median: 4755.5  
 75%(Q3): 10655.75  
 Maximum: 92780.0  
 IQR(Q3-Q1): 8502.75

### Detergents Paper Item

Count : 440  
 Mean : 2881.49  
 Standard deviation : 4767.85  
 Minimum: 3.0  
 25%(Q1): 256.75  
 50% or Median: 816.5  
 75%(Q3): 3922.0  
 Maximum: 40827.0  
 IQR(Q3-Q1): 3665.25

### Milk Item

Count : 440  
 Mean : 12000.29  
 Standard deviation : 12647.32  
 Minimum: 55.0  
 25%(Q1): 1553.0  
 50% or Median: 3627.0  
 75%(Q3): 7190.25  
 Maximum: 73498.0  
 IQR(Q3-Q1) : 5657.25

### Frozen Item

Count : 440  
 Mean : 3071.93  
 Standard deviation : 4854.67  
 Minimum: 25.0  
 25%(Q1): 742.25  
 50% or Median: 1526.0  
 75%(Q3): 3554.25  
 Maximum: 60869.0  
 IQR(Q3-Q1): 2812.0

### Delicatessen Item

Count : 440  
 Mean : 1524.87  
 Standard deviation : 2820.10  
 Minimum: 3.0  
 25%(Q1): 408.25  
 50% or Median: 965.5  
 75%(Q3): 1820.25  
 Maximum: 47943.0  
 IQR(Q3-Q1): 1412.0

## Region and Channel which spent the most & least :

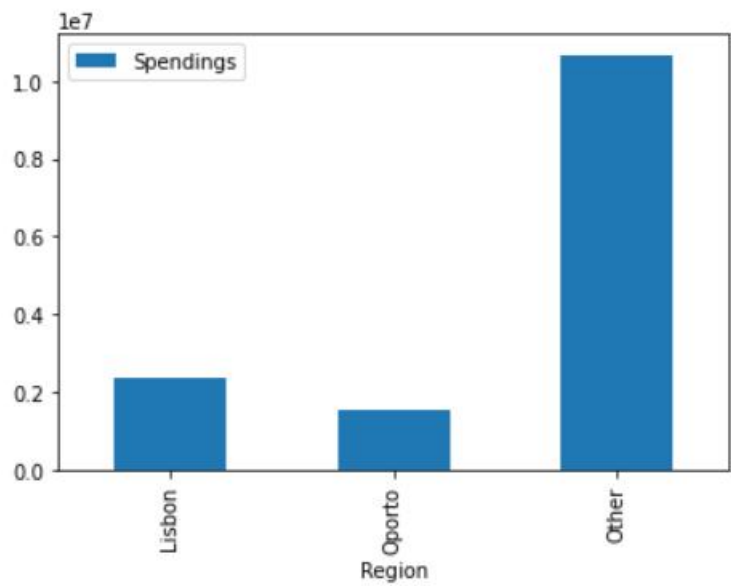
### Spending

Region	Channel	
Lisbon	Hotel	1538342
	Retail	848471
Oporto	Hotel	719150
	Retail	835938
Other	Hotel	5742077
	Retail	4935522



Region which spends the most:

Spendings	
Region	
Lisbon	2386813
Oporto	1555088
Other	10677599

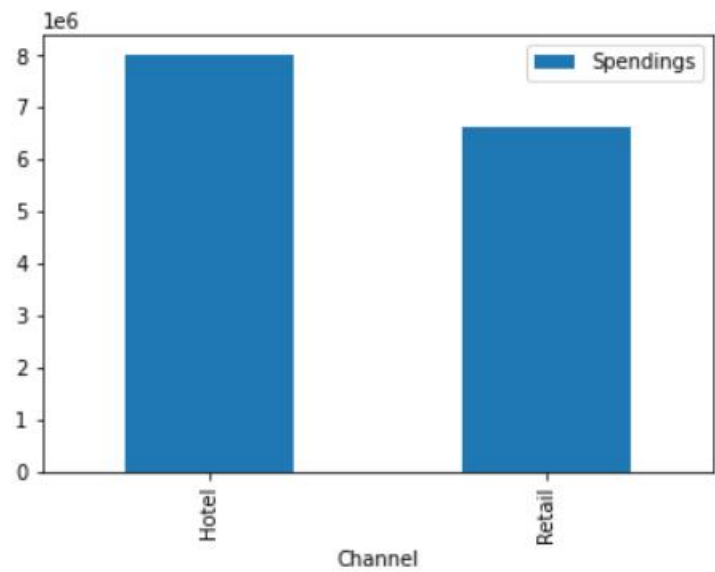


Other region spends most among the 3 region

Channel which spends the most:

Spendings	
Channel	
Hotel	7999569
Retail	6619931

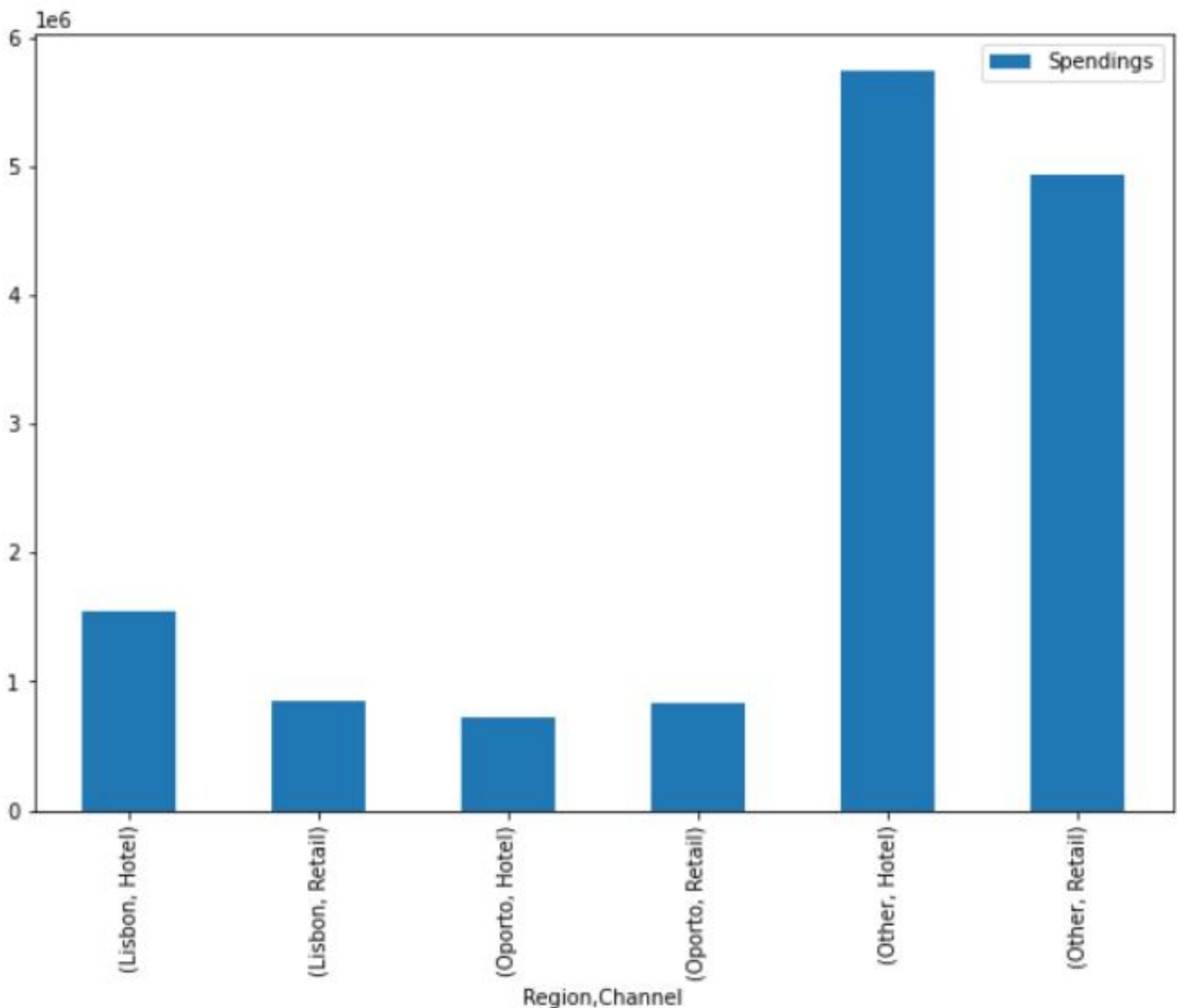
Bar plot to show the total amount spend by each region:



Bar plot to show the total amount spend by each channel:

Hotel spends most among channels

Graphical representation to show Region and Channel which spent the most and least:



From the graphical representations we can conclude that:

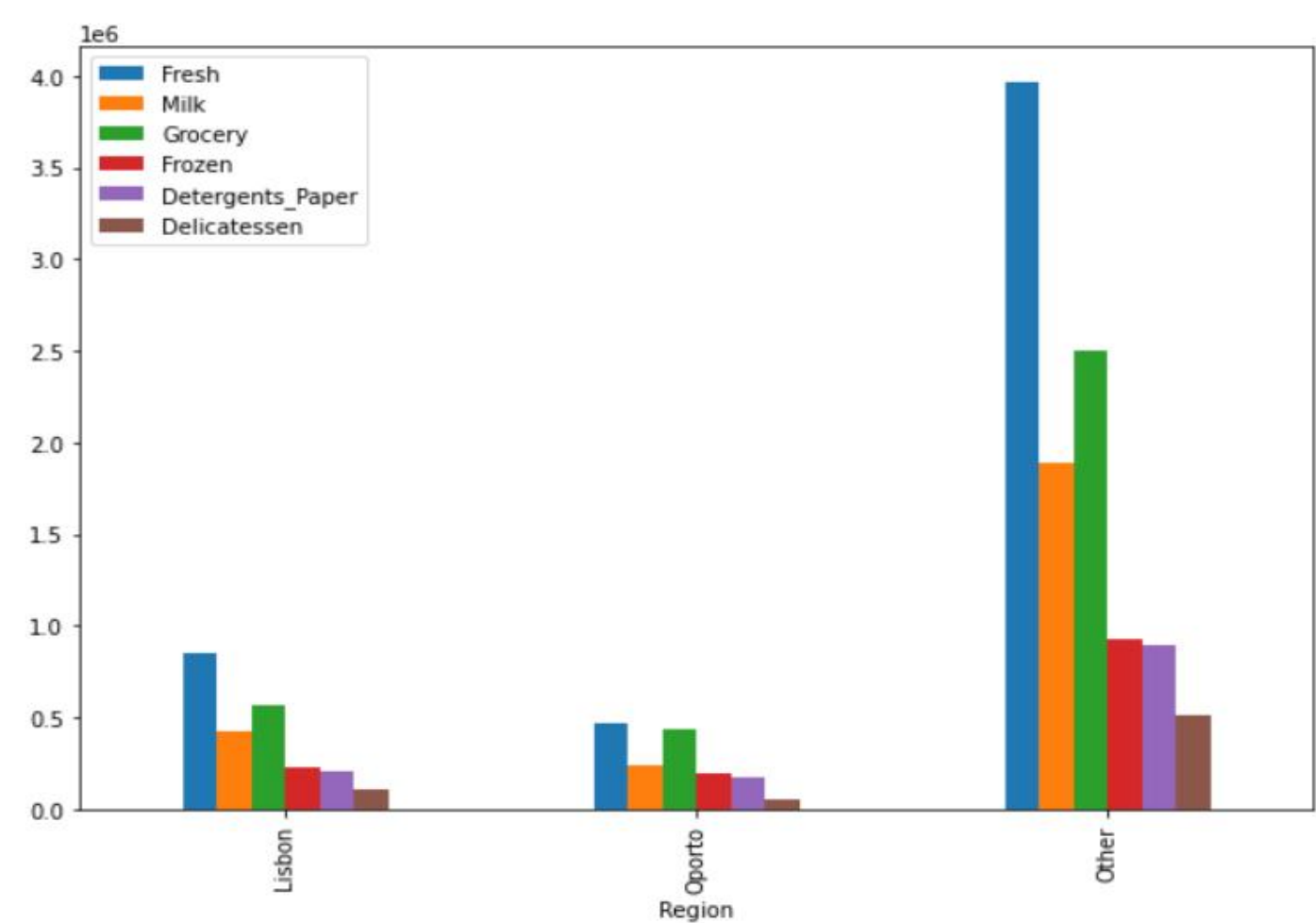
- Oporto region hotel spends the least followed by Lisbon region retail and Oporto region retail.
- Other region hotel spends the most followed by Other region retail.
- We can conclude Other region spends more on hotels and retail than Lisbon and Oporto also Oporto spends the least on both channels.
- Among hotel channel, Other region spends more followed by Lisbon and Oporto.
- Among retail channel, Other region spends more followed by Oporto and Lisbon.

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.

The data is grouped based on Region and money spend on each item:

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Region						
Lisbon	854833	422454	570037	231026	204136	104327
Oporto	464721	239144	433274	190132	173311	54506
Other	3960577	1888759	2495251	930492	890410	512110

Graphical representation based on above table :

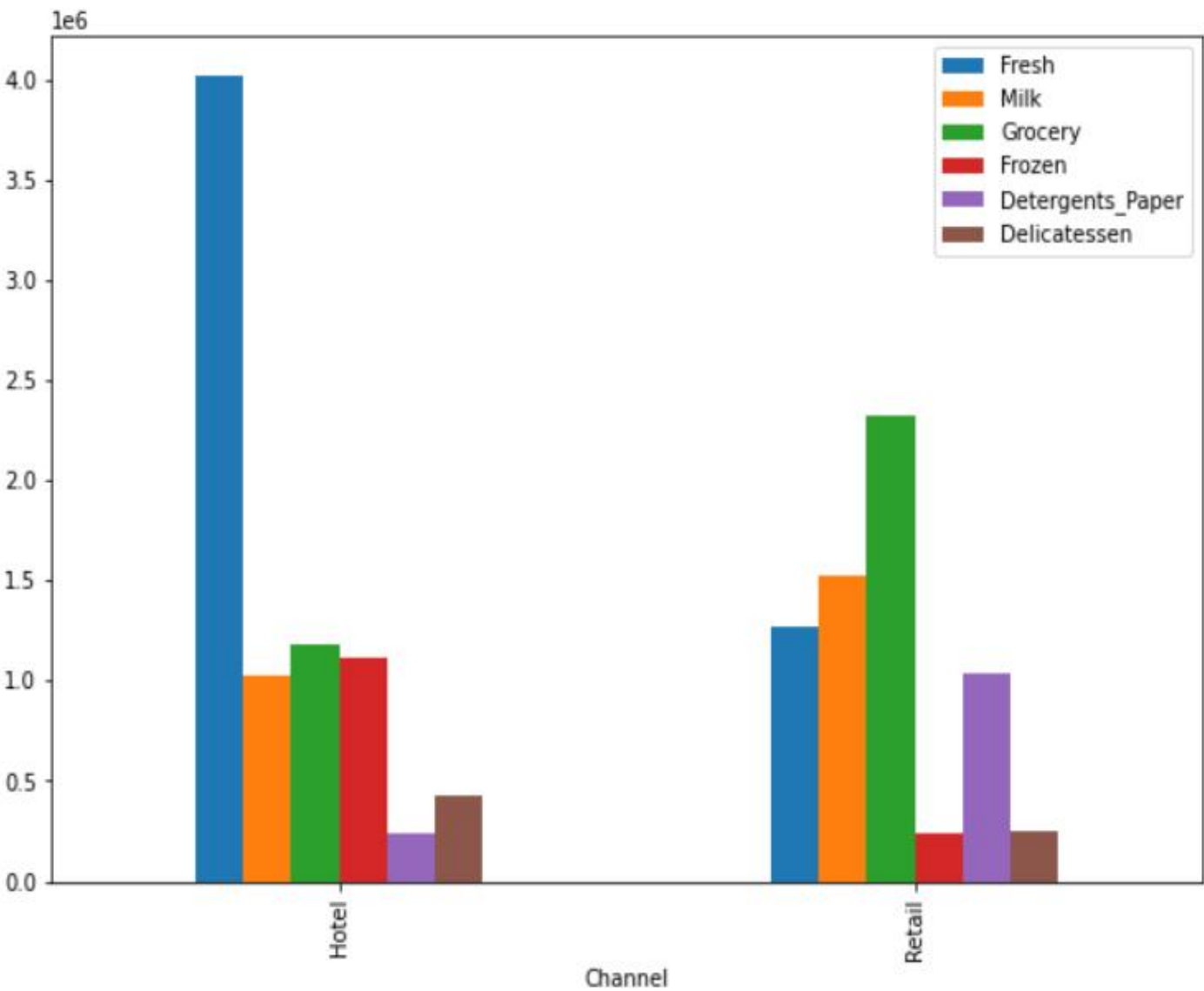


It is evident from the graph that Other region spends the most in all the 6 different items and Oporto region spends the least in all items.

The data is grouped based on channel(Hotel,Retail) and items:

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Channel						
Hotel	4015717	1028614	1180717	1116979	235587	421955
Retail	1264414	1521743	2317845	234671	1032270	248988

Graphical representation based on above table:



From the graph it is observed that Milk,grocery and detergents\_Paper items spends more across retail channel in all regions.

Fresh,frozen and delicatessen items spends more across hotel in all regions.

The below table shows the number of retail and hotel in 3 different regions

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

The number of buyer/spender in each region for all 6 items

Region						
	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Lisbon	77	77	77	77	77	77
Oporto	47	47	47	47	47	47
Other	316	316	316	316	316	316

The number of buyer/spender in each channel for all 6 items

Channel						
	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Hotel	298	298	298	298	298	298
Retail	142	142	142	142	142	142

From the above tables we can conclude that Other region has the most number of buyer(316) followed by Lisbon(77) and Oporto region(47).

There are more number of buyers in Hotel(298) than in Retail(142) across all regions.

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

	count	mean	std	min	25%	50%	75%	max
<b>Fresh</b>	440.0	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0
<b>Milk</b>	440.0	5796.265909	7380.377175	55.0	1533.00	3627.0	7190.25	73498.0
<b>Grocery</b>	440.0	7951.277273	9503.162829	3.0	2153.00	4755.5	10655.75	92780.0
<b>Frozen</b>	440.0	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0
<b>Detergents_Paper</b>	440.0	2881.493182	4767.854448	3.0	256.75	816.5	3922.00	40827.0
<b>Delicatessen</b>	440.0	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0

The standard deviation for Fresh item is 12632.95

The standard deviation for Milk item is 7371.99

The standard deviation for Grocery item is 9492.36

The standard deviation for Frozen item is 4849.15

The standard deviation for Detergents\_Paper item is 4762.43

The standard deviation for Delicatessen item is 2816.9

**Standard deviation of fresh items is the highest and delicatessen is the lowest.**

The coefficient of Variation for Fresh item is 1.05

The coefficient of Variation for Milk item is 1.27

The coefficient of Variation for Grocery item is 1.19

The coefficient of Variation for Frozen item is 1.58

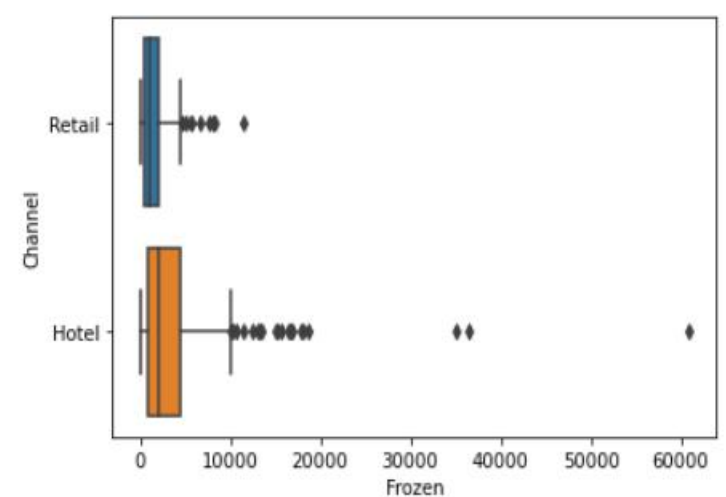
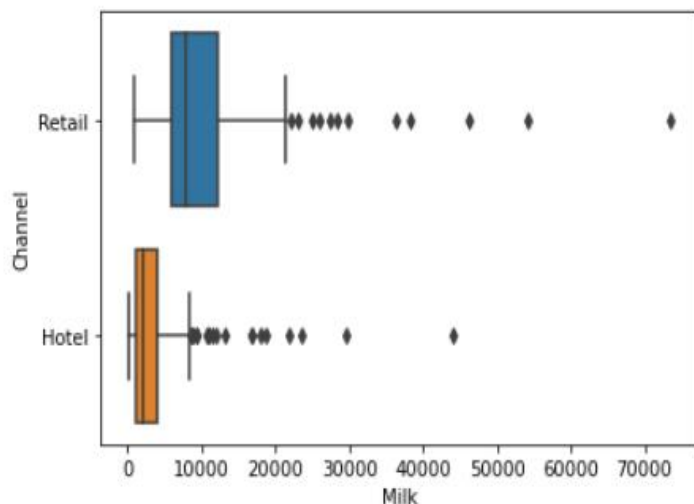
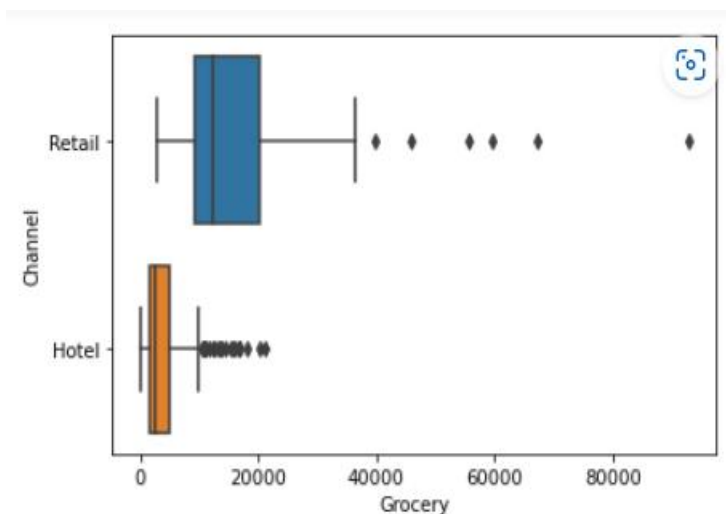
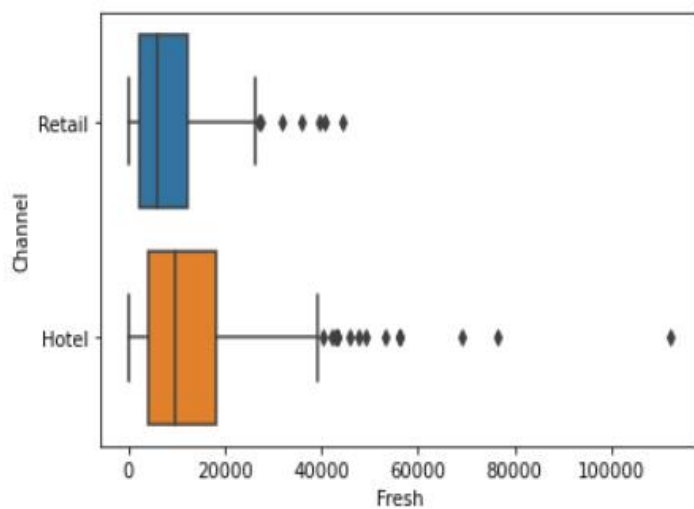
The coefficient of Variation for Detergents\_Paper item is 1.65

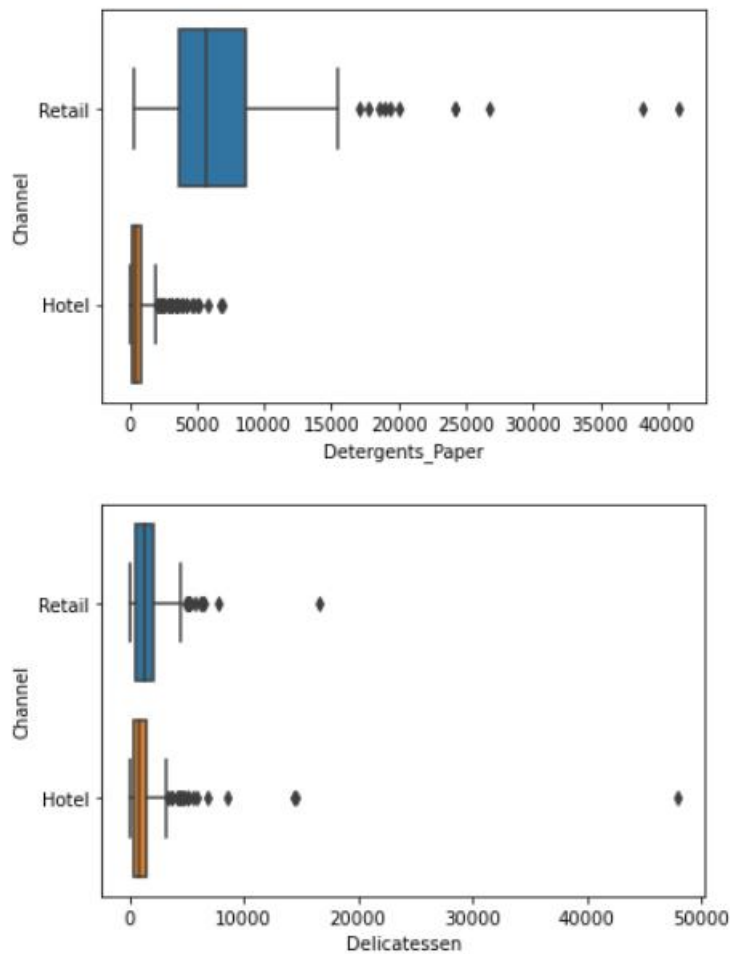
The coefficient of Variation for Delicatessen item is 1.85

**We will obtain following inferences from the coefficient of Variation(CV) obtained:**

- Fresh item have lowest coefficient of Variation
- Delicatessen item have highest coefficient of Variation.
- Even though the mean of these items may differ ,coefficient of Variation is useful measure for comparing the variability between two different datasets.
- Thus we may conclude fresh items are least inconsistent and Delicatessen are the most inconsistent ones.

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





The box plot shows that there are outliers in all the 6 items across all regions and channels.

Calculated maximum and minimum can be found out using the descriptive statistics:

Calculated minimum for Fresh is -17581.25

Calculated maximum for Fresh is 37642.75

Calculated minimum for Milk is -6952.875

Calculated maximum for Milk is 15676.12

Calculated minimum for Grocery is -10601.125

Calculated maximum for Grocery is 23409.88

Calculated minimum for Frozen is -3475.75

Calculated maximum for Frozen is 7772.25

Calculated minimum for Detergents\_Paper is -5241.125

Calculated maximum for Detergents\_Paper is 9419.88



Calculated minimum for Delicatessen is -1709.75

Calculated maximum for Delicatessen is 3938.25

Above the calculated maximum and below the calculated minimum all the values are considered as an outlier. Here all the outliers are above the calculated maximum. Also from the box plot it is evident that there are outliers in all the items.

- 1.5 On the basis of this report, what are the recommendations for the business? How can your analysis help the business to solve its problem?**
- Answer from the business perspective?**

Other region spends most in both hotel and retail channels. Oporto region can improve the sales by increasing the number of retail and hotels. The amount of money customer spends most is through hotel rather than retail channel in Lisbon and Other region. Oporto region customer spends most through retail channel. So it is advised to increase the number of retail stores in Oporto. Frozen, and Delicatessen are least spend through retail. Detergents paper is least spend through the hotel channel. Fresh item is the most consistent among all 6 products & Delicatessen is the inconsistent item.

# Survey

## Summary

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. We will be analysing data based on the questions provided.

## Introduction

The Student News Service at Clear Mountain State University (CMSU) survey dataset consists of 62 rows and 14 columns. Here we will be analysing the various factors such as age, gender, GPA, major and finding the probability of these aspects in the data.

## Data description

ID	Student ID(Count)
Gender	Student gender (Male/Female)
Age	Student age(18-26)
Class	Senior/Junior/Sophomore
Major	Accounting,CIS,Economics/Finance etc.
Grad Intention	Yes/No/Undecided
GPA	Average grade point(2.3-3.9)
Employment	Full-Time,Part-Time,Unemployed

Salary	Salary range between 25-50
Social Networking	Social Networking accounts(0-4)
Satisfaction	Satisfaction range between 1-6
Spending	Amount spend(100-1400)
Computer	Laptop,Desktop,Tablets
Text Messages	Text Messages range between 0-900

## Exploratory Data Analysis

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

## Descriptive statistics of data

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

There are 62 rows and 14 columns in dataset. Null values are not detected. Columns ID, Age, Social Networking, Satisfaction & Text Messages are integer values. Gender, Employment & Computer columns are objects. GPA and salary has float values. The frequency of females is more in dataset than males. There are 33 females and 29 males in selected 62 students. Minimum age of students is 18 maximum is 26 and average age is 21. Among the students 31 are seniors 25 are juniors and the rest are Sophomore. Most students majors in Retailing/Marketing(14) and the least majors in CIS also 3 students are undecided. 28 students have graduation intention. Minimum GPA is 2.3 ,maximum is 3.9 and an average of 3.12. Out of 62 CMSU students 43 have Part-Time employment while 10 have Full-Time employment and 9 are unemployed. 55 students have laptops with them while 5 uses desktop and 2 uses tablets.

## Sample of dataset

	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

## Problem-2

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

### **2.1.1. Gender and Major**

Contingency table for 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**

Contingency table for 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

Contingency table for 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

Contingency table for 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:

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

Total number of males = 29

Total students = 62

Probability of male students  $P(\text{Male}) = 29/62 = 0.4677$

The probability that a randomly selected CMSU student will be male is 46.77%

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

Total number of females = 33

Total students = 62

Probability of male students  $P(\text{Male}) = 33/62 = 0.5322$

The probability that a randomly selected CMSU student will be female is 53.22%

## 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.

Contingency table for 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

Normalized Contingency table for Gender and Major

Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Gender								
Female	0.090909	0.090909	0.212121	0.121212	0.121212	0.090909	0.272727	0.000000
Male	0.137931	0.034483	0.137931	0.068966	0.206897	0.137931	0.172414	0.103448
All	0.112903	0.064516	0.177419	0.096774	0.161290	0.112903	0.225806	0.048387

From the normalized table we can arrive at following conclusion:

- There is a probability 13.79% of Accounting major among the male students in CMSU.
- There is a probability 3.44% of CIS major among the male students in CMSU.



- There is a probability 13.79% of Economics/Finance major among the male students in CMSU.
- There is a probability 6.89% of International Business major among the male students in CMSU.
- There is a probability 20.68% of Management major among the male students in CMSU.
- There is a probability 13.79% of Other major among the male students in CMSU.
- There is a probability 17.2% of Retailing/Marketing major among the male students in CMSU.
- There is a probability 10.3% of Undecided major among the male students in CMSU.

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

Normalized Contingency table for Gender and Major

Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Gender								
Female	0.090909	0.090909	0.212121	0.121212	0.121212	0.090909	0.272727	0.000000
Male	0.137931	0.034483	0.137931	0.068966	0.206897	0.137931	0.172414	0.103448
All	0.112903	0.064516	0.177419	0.096774	0.161290	0.112903	0.225806	0.048387

From the normalized table we can arrive at following conclusion:

- There is a probability 9.09% of Accounting major among the female students in CMSU.
- There is a probability 9.09% of CIS major among the female students in CMSU.
- There is a probability 21.21% of Economics/Finance major among the female students in CMSU.
- There is a probability 12.12% of International Business major among the female students in CMSU.



- There is a probability 12.12% of Management major among the female students in CMSU.
- There is a probability 9.09% of Other major among the female students in CMSU.
- There is a probability 27.27% of Retailing/Marketing major among the female students in CMSU.
- There is a probability 0% of Undecided major among the female students in CMSU.

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.**

Contingency table for 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

Normalized Contingency table for Gender and Grad Intention

Grad Intention	No	Undecided	Yes
Gender			
Female	0.272727	0.393939	0.333333
Male	0.103448	0.310345	0.586207
All	0.193548	0.354839	0.451613

There is a probability 58.6% of student is a male and intends to graduate among the students in CMSU.

### 2.4.2 Find the probability that a randomly selected student is a female and does not have a laptop

Contingency table for Gender and Computer

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

Normalized Contingency table for Gender and Computer

Computer	Desktop	Laptop	Tablet
Gender			
Female	0.060606	0.878788	0.060606
Male	0.103448	0.896552	0.000000
All	0.080645	0.887097	0.032258

Probability of female students having laptop is 87.87%.

So probability of female students not having laptop is  $(100 - 87.87)\%$ . ie 12.12%.

Thus we can conclude that among students of CMSU 12.12% of females do not have a laptop.

Assume that the sample is representative of the population of CMSU.

## 2.5

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**

Contingency table for 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

Normalized contingency table for Gender and Employment

Employment	Full-Time	Part-Time	Unemployed
Gender			
Female	0.090909	0.727273	0.181818
Male	0.241379	0.655172	0.103448
All	0.161290	0.693548	0.145161

Probability of full time employed students :  $P(F) = 16.12\%$

Probability of choosing random male  $P(M) = 46.77\%$  (From question 2.2.1)

Probability of male having full time employment  $P(M \cap E) = 7/62 = 11.29\%$

Probability of randomly chosen student is a male or has a full-time employment:  $P(M \cup F)$

$$P(M \cup F) = P(F) + P(M) - P(M \cap E)$$

$$= 16.12 + 46.77 - 11.29 = 51.6\%$$

Probability that a randomly chosen student is a male or has a full-time employment is 51.6%

**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
Gender								
Female	0.090909	0.090909	0.212121	0.121212	0.121212	0.090909	0.272727	0.000000
Male	0.137931	0.034483	0.137931	0.068966	0.206897	0.137931	0.172414	0.103448
All	0.112903	0.064516	0.177419	0.096774	0.161290	0.112903	0.225806	0.048387

Probability of a random female student majoring in international business = 12.12 %

Probability of a random female student majoring in management = 12.12 %

Probability of a female student is randomly chosen, she is majoring in international business or management = 24.24%

**Construct a contingency table of Gender and Intent to Graduate at 2 levels**

**2.6** (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?

Contingency table of Gender and Intent to Graduate at 2 levels (Yes/No).

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

Normalized contingency table of Gender and Intent to Graduate at 2 levels (Yes/No).

Grad Intention	No	Yes
Gender		
Female	0.45	0.55
Male	0.15	0.85
All	0.30	0.70

s

Probability of female students intent to graduate:  $P(\text{Female} \mid \text{Grad}) = 55\%$

Probability of female students :  $P(F) = 50\%$

If  $P(\text{Female} \mid \text{Grad}) = P(F)$  they are independent events.

Probability of female students intent to graduate does not equals probability of female students. So they are not independent events.

**Note that there are four numerical (continuous) variables in the data set,**

**2.7 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?**

Total number of students having GPA less than 3 = 17.

Total number of students = 62

27.4% of students have GPA is less than 3

## 2.7.2 Find conditional probability that a randomly selected male earns 50 or more. Find conditional probability that a randomly selected female earns 50 or more.

Number of male students earns 50 or more = 14

Number of female students earns 50 or more = 18

Total number of male students = 29

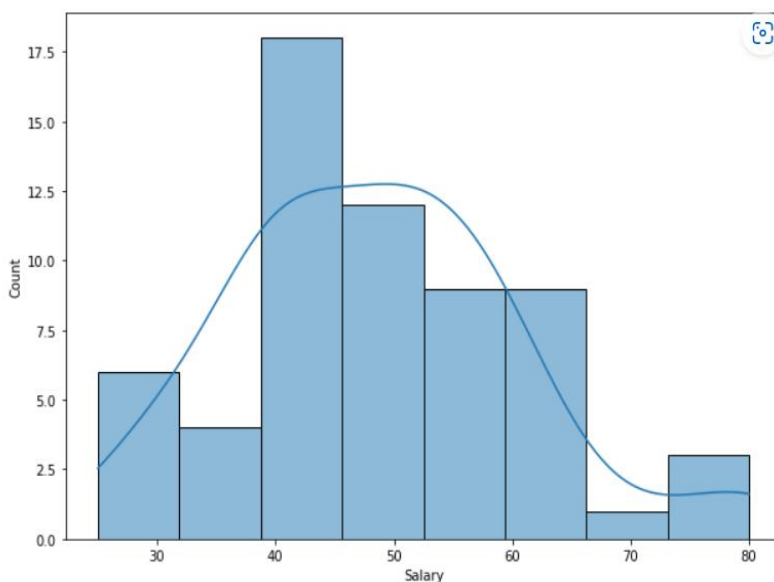
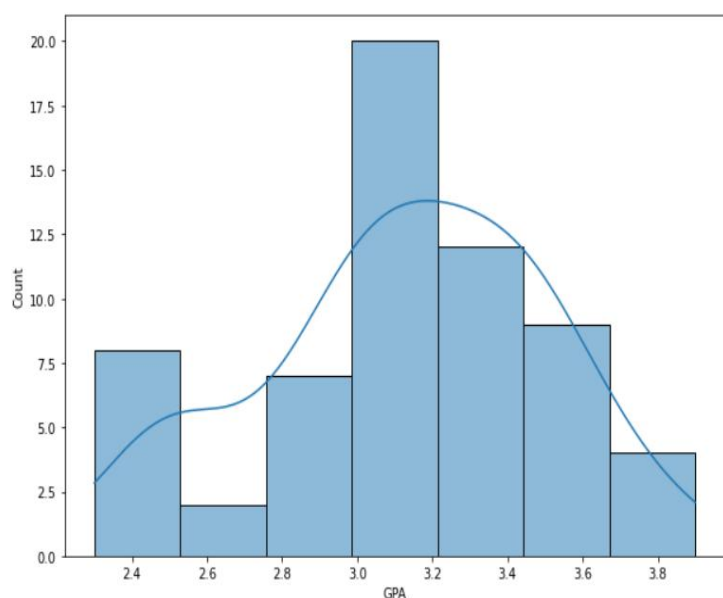
Total number of female students = 33

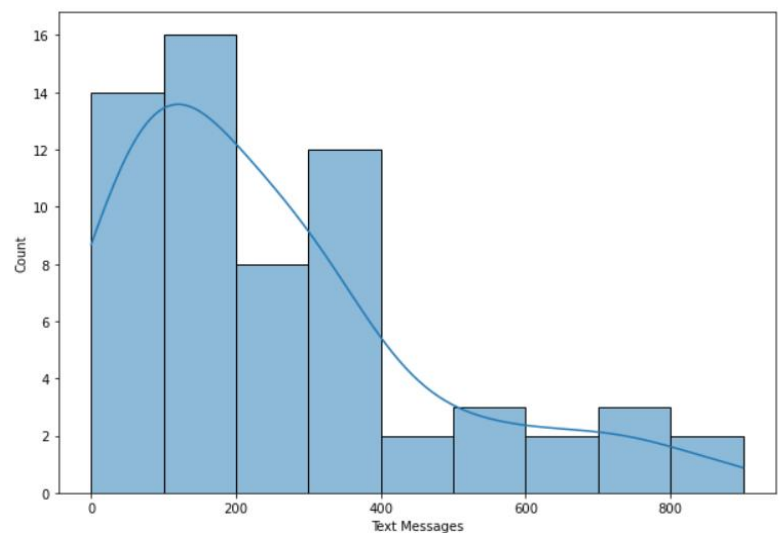
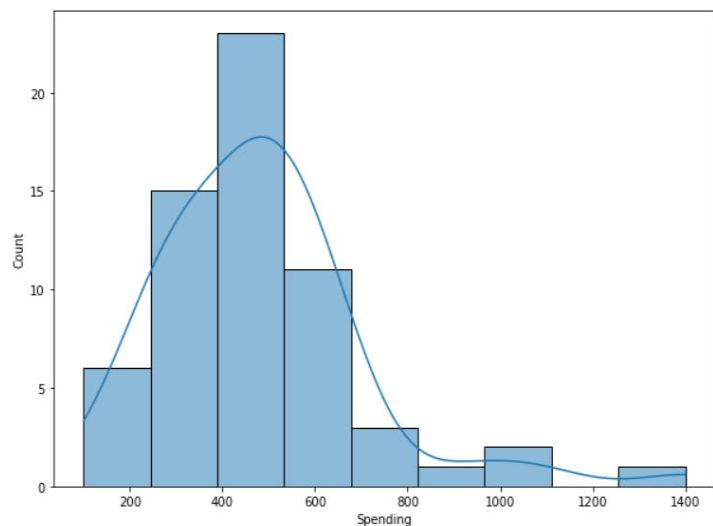
**Probability that a randomly selected male earns 50 or more is 48.27%**

**Probability that a randomly selected female earns 50 or more is 54.54%**

**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.**

To show GPA, Salary, Spending, and Text Messages columns in dataset follows a normal distribution histogram can be used.





From the histogram we can conclude that Spending, and Text Messages are right skewed data. GPA and salary follows a normal distribution. To confirm the conclusion we can use skewness and Shapiro test .

Skewness of these 4 columns:

GPA	-0.314600
Salary	0.534701
Spending	1.585915
Text Messages	1.295808

Shapiro test (p\_values) of these 4 columns

GPA	p_value = 0.11204058676958084
Salary	p_value = 0.028000956401228905
Spending	p_value = 1.6854661225806922e-05
Text Messages	p_value = 4.324040673964191e-06

From the Skewness and Shapiro test we can conclude that.

- Skewness value between -0.5 to 0.5 indicates the distribution is fairly symmetrical. Only GPA column follows this .
- If p\_value is greater than 0.05 for Shapiro test the variable follows a normal distribution. The only column which follows this condition is GPA. So it is evident that GPA follows normal distribution and Salary, Spending, and Text Messages does not follow normal distribution.

### **2.8.2 Write a note summarizing your conclusions.**

The data sample of 62 students in CMSU consists of 29 males and 33 females. Most of the students majors in Retailing/Marketing and the least in CIS. 20.68% of male students majors in Management and 21.21 % female students majors in Economics/Finance. 28 students intends to graduate in which 11 are females and rest are males. There is a probability 58.6% of student is a male and intends to graduate among the students in CMSU. With respect to employment 10 students are employed full-time while 43 are doing part time jobs. It is evident that students prefer part time jobs instead of full time . Most of the students own laptops. 27.4% of students have GPA less than 3. Probability that a randomly selected male and female earns 50 or more is 48.27% and 54.54% respectively.



## A&B Shingles

### Summary

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 are calculated. The company would like to show that the mean moisture content is less than 0.35 pounds per 100 square feet.

### Introduction

The moisture content in shingles when they are packed cause granules attached to the shingles for texture and colouring purposes to fall off the shingles resulting in appearance problems. Two samples named shingle A and shingle B are provided. There are 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles. The company want to the confirm that the mean moisture content in shingles are in permissible limits and the population means for shingles A and B are equal

### Data description

A Shingle A(pounds of moisture per 100 square feet are calculated)

B Shingle B(pounds of moisture per 100 square feet are calculated)

Exploratory Data Analysis

A	36-non_null	float64
B	31-non_null	float64

Descriptive statistics of data

	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

Descriptive statistics of data shows that there are 36 counts in shingle A and 31 counts in shingles B. There are 5 missing values in shingles B. The average moisture content in shingles A is 0.31 and in shingles B is 0.27 with a standard deviation of 0.13. The minimum value is 0.13 in shingle A and 0.10 in shingle B. The maximum is 0.72 and 0.58 for shingle A and B respectively.

Sample of dataset

	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

### Problem-3

- 3.1 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.

#### Shingles-A

##### Step 1: Define null and alternative hypotheses

$H_0$  : mean moisture content  $\leq 0.35$

$H_A$  : mean moisture content  $> 0.35$

##### Step 2: Decide the significance level

Level of significance = 0.05 (5%)

##### Step 3: Identify the test statistic

We have two independent samples, so we need to test our hypothesis for the 2 samples separately. As we have selected number of samples greater than 30 we can use 2 tailed 1 sample t test.

##### Step 4: Calculate the p - value and test statistic

Test statistics = -1.4735046

$p\_value/2 = 0.0747763$

As this is a 2 tailed test we have use  $p\_value/2$

##### Step 5 Decide to reject or accept null hypothesis

$p\_value$  is greater than level of significance for sample A. So we fail to reject null hypothesis

With 95% confidence we can state that average moisture content in shingle sample A is less than or equal to 0.35 pounds per 100 square feet.

### **Shingles-B**

#### Step 1: Define null and alternative hypotheses

$H_0$  : mean moisture content  $\leq 0.35$

$H_A$  : mean moisture content  $> 0.35$

#### Step 2: Decide the significance level

Level of significance = 0.05 (5%)

#### Step 3: Identify the test statistic

We have two independent samples, so we need to test our hypothesis for the 2 samples separately. As we have selected number of samples greater than 30 we can use 2 tailed 1 sample t test.

#### Step 4: Calculate the p - value and test statistic

Test statistics = -3.100331

$p\_value/2 = 0.0020904$

As this is a 2 tailed test we have use  $p\_value/2$

#### Step 5 Decide to reject or accept null hypothesis

$p\_value$  is less than level of significance for sample B. So we reject null hypothesis

With 95% confidence we can state that average moisture content in shingle sample B is greater than 0.35 pounds per 100 square feet.

**Do you think that the population mean for shingles A and B are equal?**

**3.2 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 hypotheses

$H_0$  : mean of A equals mean of B (mean A = B)

$H_A$  : mean of A does not equals mean of B (mean A  $\neq$  B)

Step 2: Decide the significance level

Level of significance = 0.05 (5%)

Step 3: Identify the test statistic

Shapiro test p\_value for shingle A = 0.0426705

Shapiro test p\_value for shingle B = 0.0200278

Levene's test p\_value for shingle A & B = 0.6272312

p\_value is greater than 0.05. So variances are equal as suggested by Levene's test. When there are more than about 25 observations per group and no extreme outliers, the t-test works well even for moderately skewed distributions of the outcome variable. Here we have to compare the population mean of sample A and B. So we can use 2 sample t test\_ind.

Step 4: Calculate the p - value and test statistic

Test statistics = 1.28962

p\_value = 0.201749

### Step 5 Decide to reject or accept null hypothesis

p\_value is greater than level of significance. So we fail to reject null hypothesis

With 95% confidence we can state that population mean in shingle A equals to population mean in shingle B

### **Assumptions to check before the test for equality of means is performed.**

- Data values are continuous.
- Sample data should be a representation of the population.
- Data values must be independent.
- Data should be a normal distribution. We can conduct a Shapiro test to verify normality.
- Sample size should be large enough .
- Homogeneity of variance of the samples should be approximately equal. We can conduct a Levene's test to verify the equality of variance.