



7/11/2021

# SMDM Project Report PGP -DSBA

Saloni Juwatkar



**SALONI JUWATKAR**

PGP – DATA SCIENCE AND BUSINESS ANALYTICS

## Table of Contents

<b>1</b>	<b>Problem Statement: 1</b>	<b>5</b>
1.1	Use methods of descriptive statistics to summarize data. Which Region and which Channel spent the most? Which Region and which Channel spent the least? .....	7
1.2	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. ....	8
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? .....	9
1.4	Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments. ....	10
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.....	10
<b>2</b>	<b>Problem Statement: 2</b>	<b>11</b>
2.1	For this data, construct the following contingency tables (Keep Gender as row variable)	12
2.1.1	Gender and Major.....	12
2.1.2	Gender and Grad Intention.....	12
2.1.3	Gender and Employment.....	13
2.1.4	Gender and Computer .....	13
2.2	Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:.....	13
2.2.1	What is the probability that a randomly selected CMSU student will be male? .....	13
2.2.2	What is the probability that a randomly selected CMSU student will be female? .....	13
2.3	Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:.....	14
2.3.1	Find the conditional probability of different majors among the male students in CMSU.....	14
2.3.2	Find the conditional probability of different majors among the female students of CMSU. ....	14
2.4	Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:.....	15
2.4.1	Find the probability That a randomly chosen student is a male and intends to graduate. ....	15
2.4.2	Find the probability that a randomly selected student is a female and does NOT have a laptop. ....	15
2.5	Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:.....	16
2.5.1	Find the probability that a randomly chosen student is a male or has full-time employment? ..	16
2.5.2	Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management. ....	16
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? .....	17

2.7 Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. .... 17

Answer the following questions based on the data ..... 17

2.7.1 If a student is chosen randomly, what is the probability that his/her GPA is less than 3? ..... 17

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. .... 18

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. .... 19

### 3 Problem Statement 3 ..... 20

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. .... 21

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?..... 22

## List of Figures

Figure 1: Problem 1-Sample Dataset .....	5
Figure 2 : Problem Statement 1-Null Values.....	6
Figure 3: Problem 1: Heatmap.....	6
Figure 4:Problem 1- Descriptive Statistics .....	7
Figure 5: Channel and Region Annual Spending.....	7
Figure 6 : Plot of each product spending against each Region and Channel.....	8
Figure 7: Standard Deviation of 6 products.....	9
Figure 8: Coefficient of Variation.....	9
Figure 9: Box Plot of all Products .....	10
Figure 10: Problem 2 – Dataset .....	11
Figure 11: Problem 2 - Null data .....	12
Figure 12: Distplot.....	19
Figure 13: Shingles A and B Dataset .....	20
Figure 14: Problem 3 - Null Data.....	20

## List of Tables

Table 1: Contingency Table (Gender and Major).....	12
Table 2: Contingency Table (Gender and Grad Intention).....	12
Table 3: Contingency Table (Gender and Employment).....	13
Table 4: Contingency Table (Gender and Computer) .....	13
Table 5: Conditional Probabilities (Major and Male).....	14
Table 6: Conditional Probabilities (Major and Female) .....	15
Table 7: Contingency Table (Gender & Intent to Graduate (Yes or No)).....	17
Table 8: Contingency Table (Gender and GPA).....	17
Table 9: Contingency Table (Gender and Salary).....	18

Saloni Juwatkar

## 1 Problem Statement: 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).

### Data Description

1. **Buyer/Spender:** No of retailers. Index of the dataset.
2. **Channel:** Different channel across which items are sold.
3. **Region:** Different regions across which products are sold.
4. **Fresh:** Annual spending on fresh products.
5. **Milk:** Annual spending on milk products.
6. **Grocery:** Annual spending on grocery products.
7. **Frozen:** Annual spending on frozen products.
8. **Detergents\_Paper:** Annual spending on detergents paper.
9. **Delicatessen:** Annual spending on delicatessen.

### Sample of Dataset

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
0	1	Retail	Other	12669	9656	7561	214	2674	1338
1	2	Retail	Other	7057	9810	9568	1762	3293	1776
2	3	Retail	Other	6353	8808	7684	2405	3516	7844
3	4	Hotel	Other	13265	1196	4221	6404	507	1788
4	5	Retail	Other	22615	5410	7198	3915	1777	5185

Figure 1: Problem 1-Sample Dataset

The dataset has information about 440 retailer's annual spending across 6 products in 3 different regions (Lisbon, Oporto, Other) and across 2 different channels (Hotel, Retail.).

Assumption: the annual spending is done in INR i.e., Rs.

## Exploratory Data Analysis

Missing values in Dataset.

Channel	0
Region	0
Fresh	0
Milk	0
Grocery	0
Frozen	0
Detergents_Paper	0
Delicatessen	0

Figure 2 : Problem Statement 1-Null Values

From the figure, we can see that there are no missing values in dataset.

## Correlation

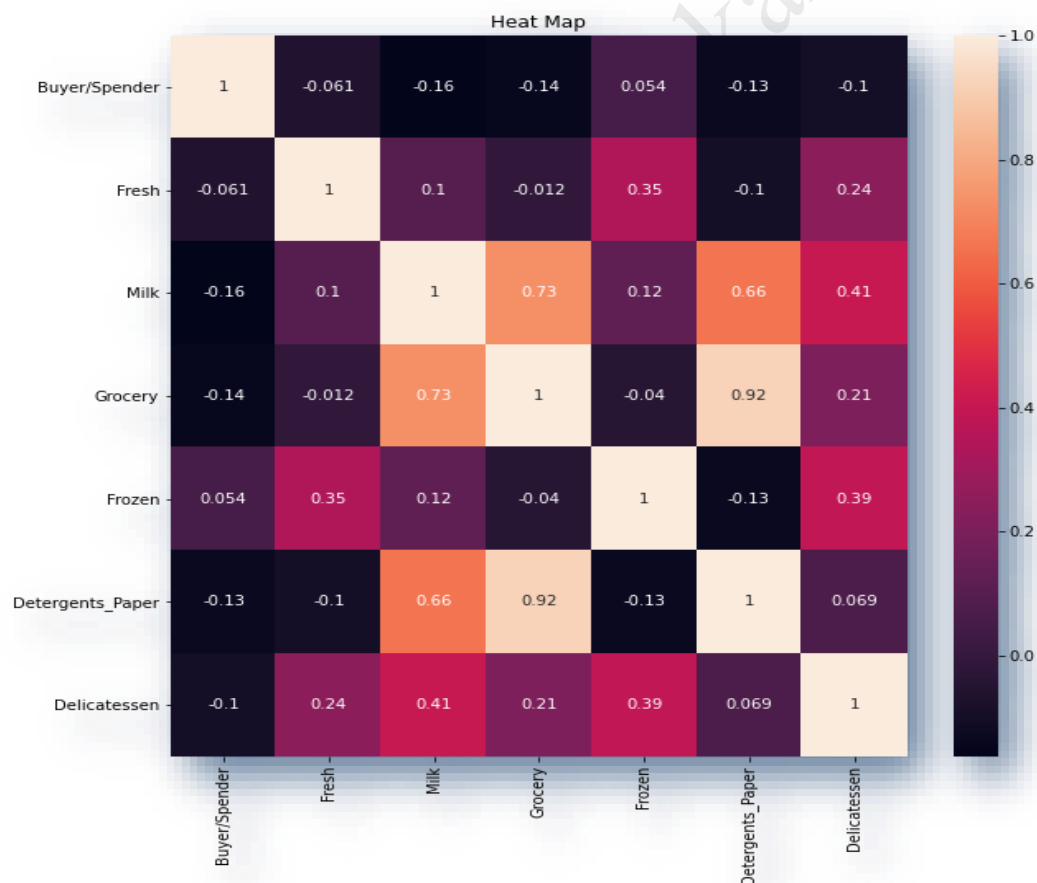


Figure 3: Problem 1: Heatmap

From the correlation plot, we can see that some variables (products) are highly correlated to one another. Correlation values near 1 or -1 are highly correlated whereas values 0 or near to 0 are not correlated.

### 1.1 Use methods of descriptive statistics to summarize data. Which Region 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
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	440.0	NaN	NaN	NaN	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0
Milk	440.0	NaN	NaN	NaN	5796.265909	7380.377175	55.0	1533.0	3627.0	7190.25	73498.0
Grocery	440.0	NaN	NaN	NaN	7951.277273	9503.162829	3.0	2153.0	4755.5	10655.75	92780.0
Frozen	440.0	NaN	NaN	NaN	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0
Detergents_Paper	440.0	NaN	NaN	NaN	2881.493182	4767.854448	3.0	256.75	816.5	3922.0	40827.0
Delicatessen	440.0	NaN	NaN	NaN	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0

Figure 4: Problem 1- Descriptive Statistics

- We use descriptive statistic to summarize the dataset.
- From the descriptive statistics we can see that there are 6 different types of products.
- The channels through which the retailers spend their annual spending on are of two categories: Hotels, Retail. Although it can be observed that the spending through hotels is higher (298 out of 440 retailers) among retailers.
- The regions in which products are sold are Lisbon, Oporto and other. It is observed that the products are sold the highest in other Regions (316 out of 440).
- Considering all the products it can be observed that the maximum spending was done on fresh products having a maximum spending of Rs.112151 and the least amount was spent on Detergent paper having maximum spending of Rs.40827.

```
Channel
Hotel    7999569
Retail    6619931
Name: Total, dtype: int64

Region
Lisbon    2386813
Oporto    1555088
Other    10677599
Name: Total, dtype: int64
```

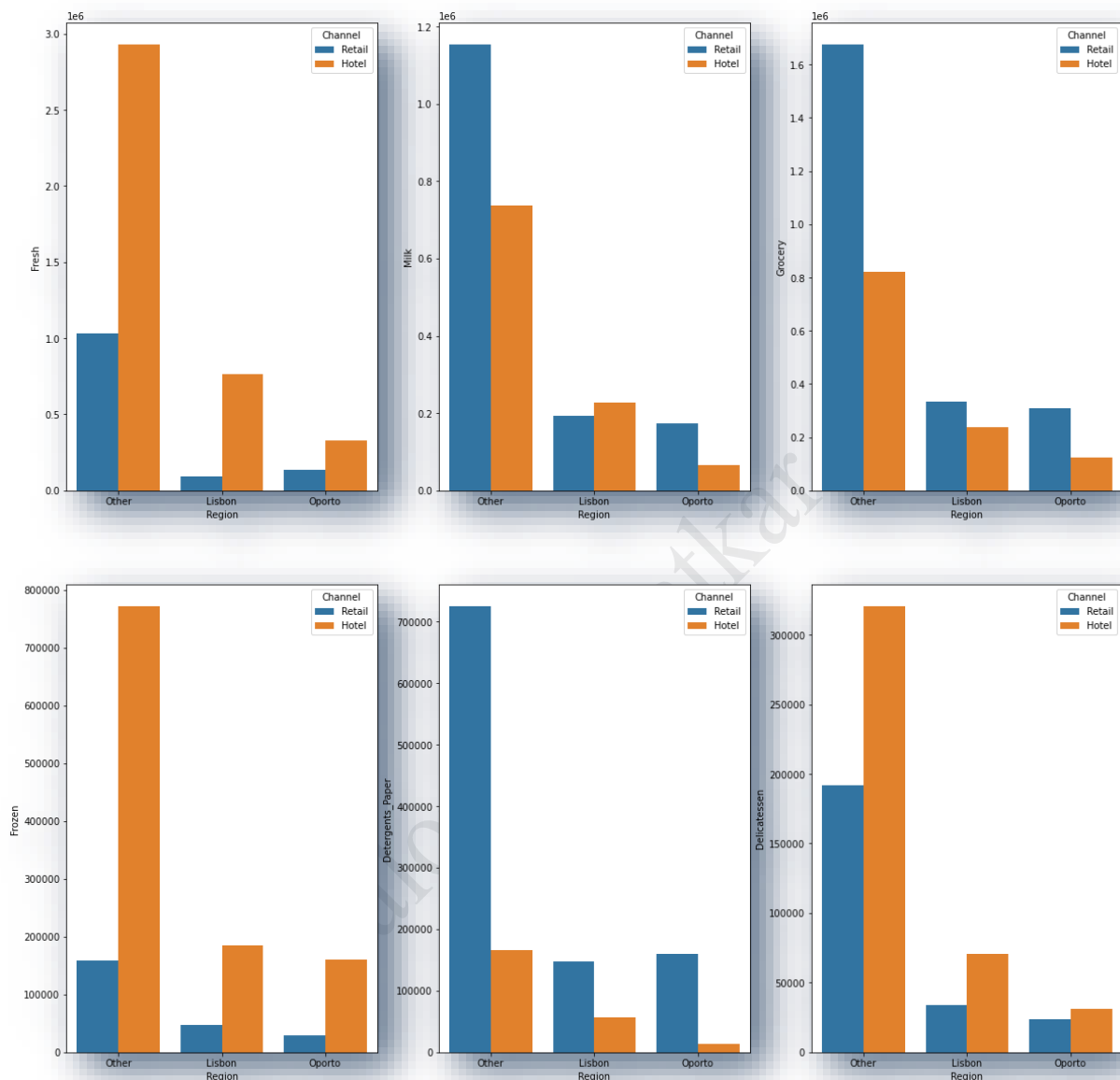
Figure 5: Channel and Region Annual Spending.

From the above analysis it can be determined that

‘Hotel’ **Channel (Rs. 7999569)** and ‘Other’ **Region (Rs. 10677599)** spends the most.  
‘Retail’s’ **Channel (Rs. 6619931)** and ‘Oporto’ **Region (Rs. 1555088)** spends the least on the products.



**1.2 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 6 : Plot of each product spending against each Region and Channel*

From the above plot, it can be observed that

- All 6 products have higher consumption in other regions.
- All 6 products have higher consumption in hotel channel.
- Fresh, Frozen and Delicatessen products have higher consumption in hotel channel.
- Milk, Grocery and Detergents paper have higher consumption in retail channel.

### 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?

If you consider the descriptive measure of variability, standard deviation and coefficient of variance would have to be considered.

#### Standard Deviation:

Standard Deviation of six products :	
Fresh	12647.328865
Milk	7380.377175
Grocery	9503.162829
Frozen	4854.673333
Detergents_Paper	4767.854448
Delicatessen	2820.105937
Total	26356.301730

Figure 7: Standard Deviation of 6 products.

**Fresh** products have the **highest** standard deviation and hence it shows **most inconsistent behaviour**.

**Delicatessen** has the **lowest** standard deviation and hence it shows the **least inconsistent** behaviour.

#### Coefficient of Variation:

Coefficient of Variation:	
Fresh	1.053918
Delicatessen	1.849407
Milk	1.273299
Grocery	1.195174
Frozen	1.580332
Detergents_Paper	1.654647

Figure 8: Coefficient of Variation

**Fresh** Product **least** variation hence it is least **inconsistent**, whereas delicatessen having the **highest variation** is the **most inconsistent**.

Therefore, from the derived values we can say that **delicatessen product** shows the **most inconsistent** behaviour and the **fresh product** shows the **least inconsistent** behaviour.

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

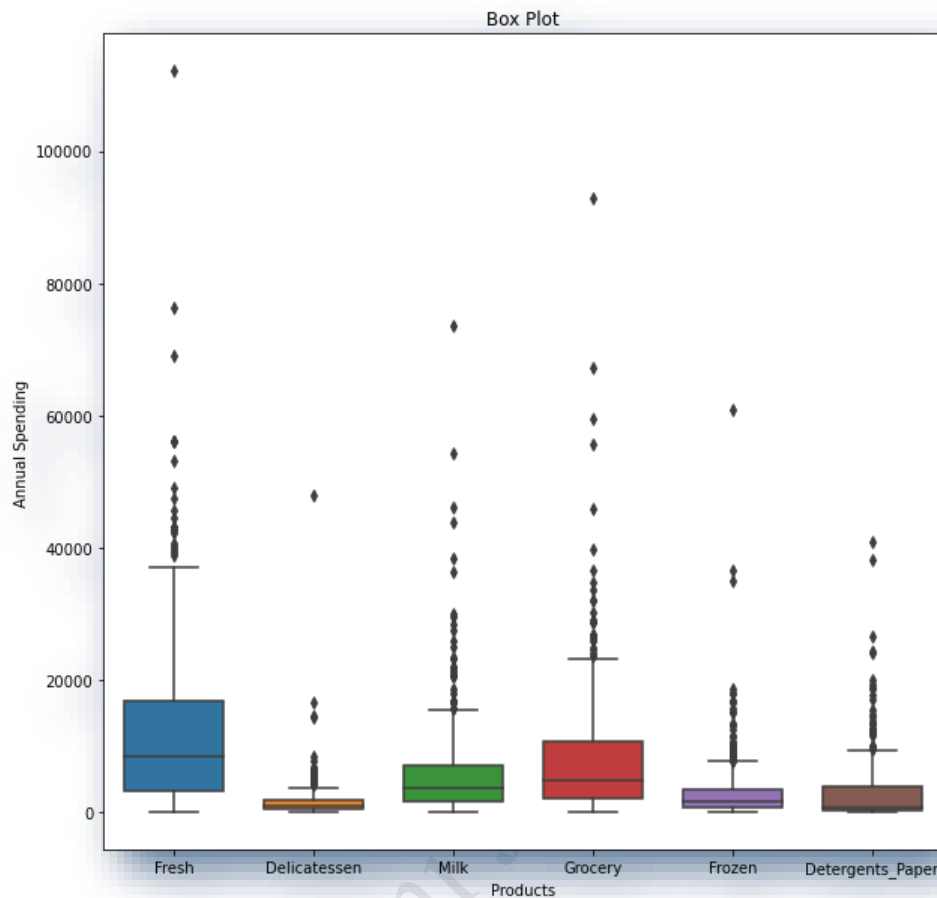


Figure 9: Box Plot of all Products

From the above plot it can be inferred that there are outliers present in the dataset.

**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**

From the analysis carried above we can say that there are inconsistencies when you calculate the measures of variability. There are some outliers present in the annual spending of the 6 products, so the spending can be minimized to remove inconsistencies and the outliers from the dataset. Also, annual spending should be done equally for all products in all regions and across all channels.

## 2 Problem Statement: 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

### Data Description

1. **ID:** no of Students.
2. **Gender:** Gender of all 62 students.
3. **Age:** The age of the CMSU student.
4. **Class:** Junior or Senior class student.
5. **Major:** Which major a student has opted for (Accounting, CIS, Economics/Finance, International Business, Management, Other, Retailing/Marketing and Undecided)
6. **Grad Intention:** If the student is planning to graduate.
7. **GPA:** GPA of the student.
8. **Employment:** If the student is Full-time, part-time or unemployed.
9. **Salary:** Salary of Student.
10. **Social Networking:**
11. **Satisfaction:** Rating from the student.
12. **Spending:** Spending of the student.
13. **Computer:** Whether the student has a computer or not.
14. **Text Messages:** No. of messages he sends.

### Sample of Dataset

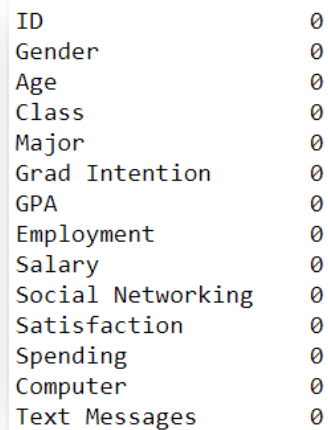
	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
0	1	Female	20	Junior	Other	Yes	2.9	Full-Time	50.0	1	3	350	Laptop	200
1	2	Male	23	Senior	Management	Yes	3.6	Part-Time	25.0	1	4	360	Laptop	50
2	3	Male	21	Junior	Other	Yes	2.5	Part-Time	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

Figure 10: Problem 2 – Dataset

The data contains survey of 62 undergraduates from CMSU. The survey has 14 question and all questions are answered (no null data).

## Exploratory Data Analysis.

### Missing Value Check



ID	0
Gender	0
Age	0
Class	0
Major	0
Grad Intention	0
GPA	0
Employment	0
Salary	0
Social Networking	0
Satisfaction	0
Spending	0
Computer	0
Text Messages	0

Figure 11: Problem 2 - Null data

There is no missing value in the dataset.

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

### 2.1.1 Gender and Major

Table 1: Contingency Table (Gender and Major)

Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Gender								
Female	3	3	7	4	4	3	9	0
Male	4	1	4	2	6	4	5	3

### 2.1.2 Gender and Grad Intention

Table 2: Contingency Table (Gender and Grad Intention)

Grad Intention	No	Undecided	Yes
Gender			
Female	9	13	11
Male	3	9	17

### 2.1.3 Gender and Employment

Table 3: Contingency Table (Gender and Employment)

Employment	Full-Time	Part-Time	Unemployed
Gender			
Female	3	24	6
Male	7	19	3

### 2.1.4 Gender and Computer

Table 4: Contingency Table (Gender and Computer)

Computer	Desktop	Laptop	Tablet
Gender			
Female	2	29	2
Male	3	26	0

## 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?

This is found out by finding the ratio of male students to the total no of CMSU students.

Based on the Dataset, the probability percentage that a randomly selected student is **male** is **46.77%**.

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

This is found out by finding the ratio of female students to the total no of CMSU students.

Based on the Dataset, the probability percentage that a randomly selected student is **female** is **53.23%**.

**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.**

Conditional Probability is calculated using the formula

$$p(A|B) = p(A \cap B) / p(B).$$

Therefore, using the contingency table (Table 1: Contingency Table (Gender and Major)) and the formula above, below conditional probabilities are obtained for the question.

*Table 5: Conditional Probabilities (Major and Male)*

Major	Conditional Probability (%)
Accounting	13.79
CIS	3.45
Economics/Finance	13.79
International Business	6.9
Management	20.69
Other	13.79
Retailing/Marketing	17.24
Undecided	10.34

It can be inferred that male student opt more for **Management** major.

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

Conditional Probability is calculated using the formula

$$p(A|B) = p(A \cap B) / p(B).$$

Therefore, using the contingency table( Table 1: Contingency Table (Gender and Major)) and the formula above, below conditional probabilities are obtained for the question.

Table 6: Conditional Probabilities (Major and Female)

Major	Conditional Probability (%)
Accounting	9.09
CIS	9.09
Economics/Finance	21.21
International Business	12.12
Management	12.12
Other	9.09
Retailing/Marketing	27.27
Undecided	0

It can be inferred that **Female** student opt more for **Retailing/Marketing** major.

## 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.

Here the contingency table (Table 2: Contingency Table (Gender and Grad Intention)) is used to find the probability of a random student being male and intending to graduate.

$$P(\text{Male and Intend to Graduate}) = P(\text{Intend to Graduate}|\text{Male}) * P(\text{Male})$$

Thus, Probability percentage of student chosen is **male and intends to graduate is 27.42 %**.

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

Here the contingency table( Table 4: Contingency Table (Gender and Computer)) is used to find the probability of a random student being female and not having a laptop.

$$P(\text{Female and No laptop}) = P(\text{No Laptop}|\text{Female}) * P(\text{Female})$$



Thus, Probability percentage that a randomly selected student is a **female and does NOT have a laptop 6.45 %**

**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?**

Here the contingency table (Table 3: Contingency Table (Gender and Employment)) to find the probability of student is a male or has a full-time employment.

$$P(\text{Male or Full Employed}) = P(\text{Male}) + P(\text{Full - Employed}) - P(\text{Male and Full Employed})$$

Probability percentage that chosen student is **male or has full time employment** is **46.77 %**

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

From conditional probability calculated in table (Table 6: Conditional Probabilities) we can say that the conditional probability percentage that a given female is majoring in international business or management is **24.24%**.

**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?**

The Contingency table for gender and intent to graduate according to the requirement is as:

*Table 7: Contingency Table (Gender & Intent to Graduate (Yes or No))*

Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

From the above table,

$$P(\text{Graduate Intent: Yes}) = 28/40 = 0.7$$

$$P(\text{Female and Graduate Intent: Yes}) = 11/20 = 0.55$$

Hence, it can be derived that the two events are **independent events** as both the probabilities are different.

**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?**

Here a contingency table of Gender vs GPA has to be constructed.

*Table 8: Contingency Table (Gender and GPA)*

GPA	2.3	2.4	2.5	2.6	2.8	2.9	3	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9
Gender																
Female	1	1	2	0	1	3	5	2	4	3	2	4	1	2	1	1
Male	0	0	4	2	2	1	2	5	2	2	5	2	2	0	0	0

From the above table it can be found that the probability percentage that his/her GPA is **less than 3 is 27.42 %**.

**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.**

Here a contingency table of Gender vs Salary has to be constructed.

*Table 9: Contingency Table (Gender and Salary)*

Salary	25	30	35	37	37.5	40	42	45	47	47.5	50	52	54	55	60	65	70	78	80
Gender																			
Female	0	5	1	0	1	5	1	1	0	1	5	0	0	5	5	0	1	1	1
Male	1	0	1	1	0	7	0	4	1	0	4	1	1	3	3	1	0	0	1

From the above table it can be inferred that probability percentage that a randomly selected **male** earns 50 or more than 50 is **48.28 %** and probability percentage that a randomly selected **female** earns 50 and more than 50 is **54.55%.**

**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 find whether the given continuous variables are normally distributed we need to construct a distplot for each of them.

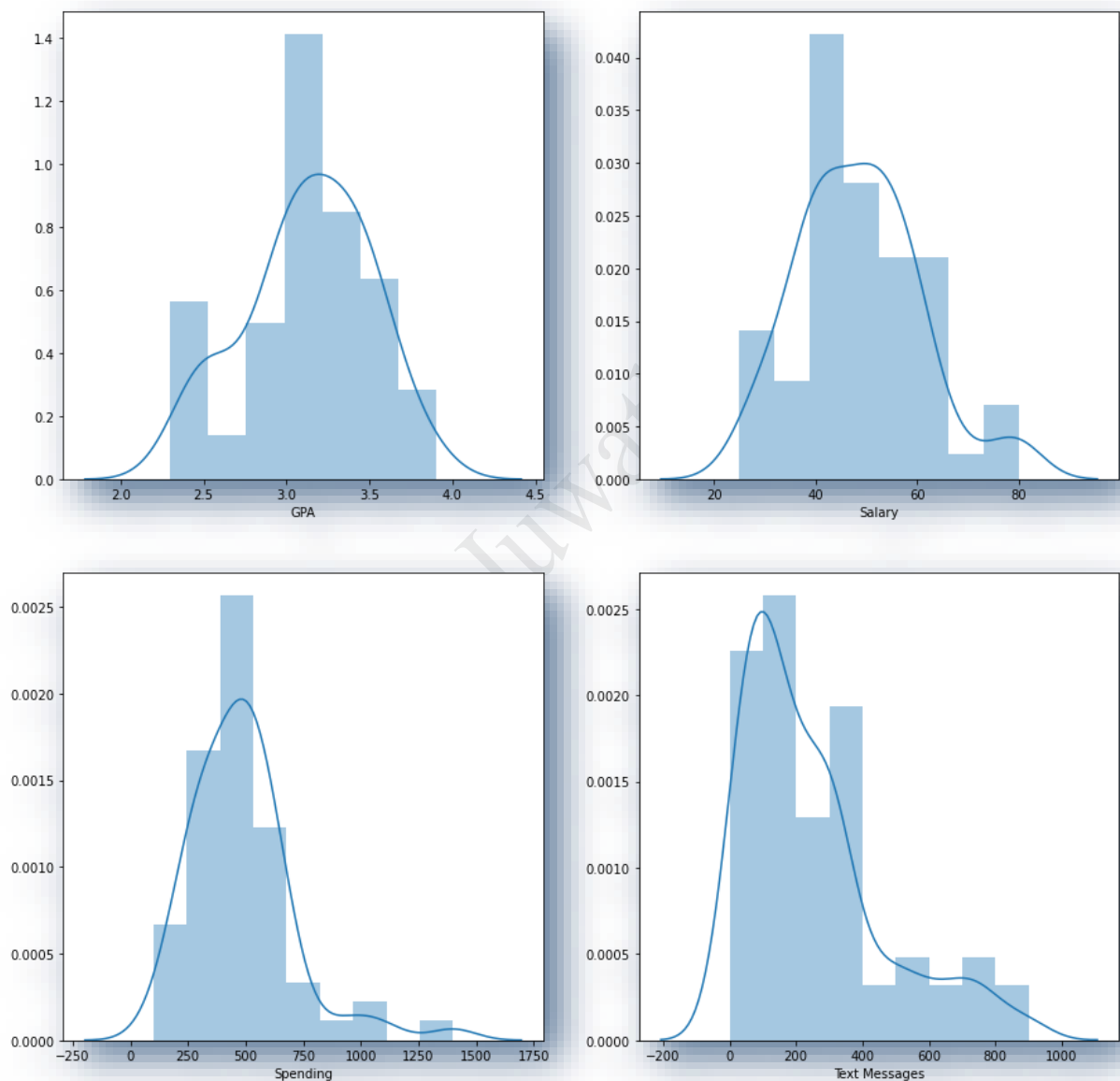


Figure 12: Distplot

Thus, it can be inferred that **GPA and Salary** are relatively **normally distributed** whereas **Spending and Text Messages** are **not normally distributed (right skewed)**.

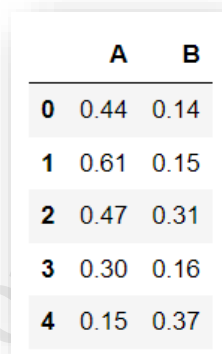
### 3 Problem Statement 3

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

#### Data Description

1. A: Shingle A weight in pounds.
2. B: Shingle B weight in pounds.

#### Sample of Dataset

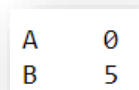


	A	B
0	0.44	0.14
1	0.61	0.15
2	0.47	0.31
3	0.30	0.16
4	0.15	0.37

Figure 13: Shingles A and B Dataset

#### Exploratory Data Analysis

##### Check missing value



A	0
B	5

Figure 14: Problem 3 - Null Data

From this we can see the shingle B data has some missing or null data, hence it has to be made sure to handle this while doing hypothesis testing.

### 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.

$\alpha$ : 0.05 (default)

**For shingles A**, 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:

**H0:**  $\mu \leq 0.35$

**H1:**  $\mu > 0.35$

From python calculation we can find the below values for one tailed t-test.

**t\_value:** -1.47

**p\_value:** 0.07 (one t test)

Here  $p\_value > \alpha$ , we **fail to reject** the null hypothesis (Ho).

Therefore, the mean moisture content is not within permissible limits (less than 0.35 pound per 100 square feet).

**For shingles B**, 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:

**H0:**  $\mu \leq 0.35$

**H1:**  $\mu > 0.35$

From python calculation we can find the below values for one tailed t-test.

**t\_value:** -3.1

**p\_value:** 0.002(one t test)

Here  $p\_value < \alpha$ , we to reject the null hypothesis (Ho).

Therefore, the mean moisture content is within permissible limits (less than 0.35 pound 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?**

**$\alpha$ :** 0.05 (default)

**H0:**  $\mu(A) = \mu(B)$

**H1:**  $\mu(A) \neq \mu(B)$

From python calculation we can find the below values.

**t\_value:** 1.29

**p\_value:** 0.202

Here,  $p\_value > \alpha$ , therefore **we fail to reject** null hypothesis (H0).

While testing for equality of means, we assume that that both distributions are normal and the variance of them are equal.

Thus, it can be inferred that the population mean of two shingles A and B are equal.

**The End!**