SMDM BUSINESS REPORT

BY JASPER SHELDON M

I. Table of Contents

1. PROBLEM 1: 4

1.1.1. Use methods of descriptive statistics to summarize data. 4

1.1.2. Which Region and which Channel spent the most? 5

1.1.3. Which Region and which Channel spent the least? 5

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. 6

1.3. On the basis of a descriptive measure of variability, which item shows the most inconsistent behavior? Which items show the least inconsistent behavior? 7

1.4. Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments. 7

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. 8

2. PROBLEM 2: 9

2.1 For this data, construct the following contingency tables (Keep Gender as row variable 9

2.1.1 Gender and Major 10

2.1.2 Gender and Grad Intention 10

2.1.3 Gender and Employment 10

2.1.4 Gender and Computer 11

2.2 Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question 11

2.2.1 What is the probability that a randomly selected CMSU student will be male 11

2.2.2 What is the probability that a randomly selected CMSU student will be female? 11

2.3 : Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question 11

2.3.1 Find the conditional probability of different majors among the male students in CMSU. 11

2.3.2 Find the conditional probability of different majors among the female students of CMSU. 12

2.4 Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question 12

2.4.1 Find the probability that a randomly chosen student is a male and intends to graduate. 12

2.4.2 Find the probability that a randomly selected student is a female and does NOT have a laptop. 13

2.5 Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question 13

2.5.1 Find the probability that a randomly chosen student is a male or has full-time employment? 13

2.5.2 Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management. 14

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? 14

2.7 Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. 15

2.7.1 If a student is chosen randomly, what is the probability that his/her GPA is less than 3 .. 15

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. 15

2.8 16

2.8.1 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 for this whole Problem 2 16

2.8.2 Write a note summarizing your conclusions for this whole Problem 2. 17

3. PROBLEM 3 18

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. 18

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? 19

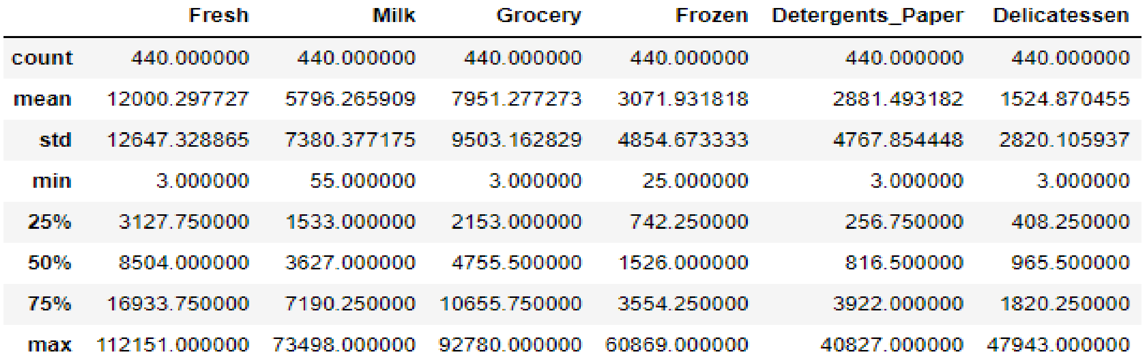
1. PROBLEM 1:

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

1.1.1. Use methods of descriptive statistics to summarize data.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 440 entries, 1 to 440
Data columns (total 8 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Channel           440 non-null    object
 1   Region            440 non-null    object
 2   Fresh             440 non-null    int64
 3   Milk              440 non-null    int64
 4   Grocery           440 non-null    int64
 5   Frozen            440 non-null    int64
 6   Detergents_Paper  440 non-null    int64
 7   Delicatessen      440 non-null    int64
dtypes: int64(6), object(2)
```

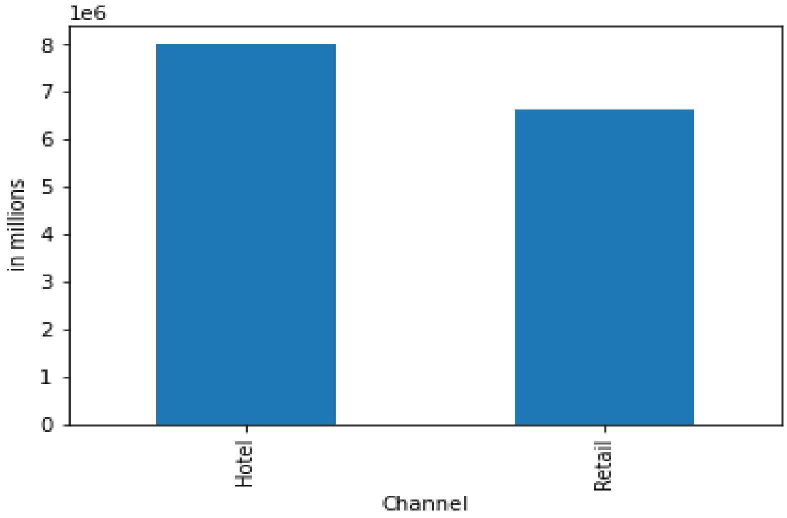
The above table shows the overall information about the data, we have totally 440 rows and 8 columns (Note: Buyer/Sender is considered as Index including it there are 9 columns present). 440 non null shown for each column this means there are no missing values in the data. 2 columns i.e. Channel and Region object type which means it is a textual data and rest of the columns are numerical values.



	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
mean	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000
50%	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000
75%	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000
max	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000

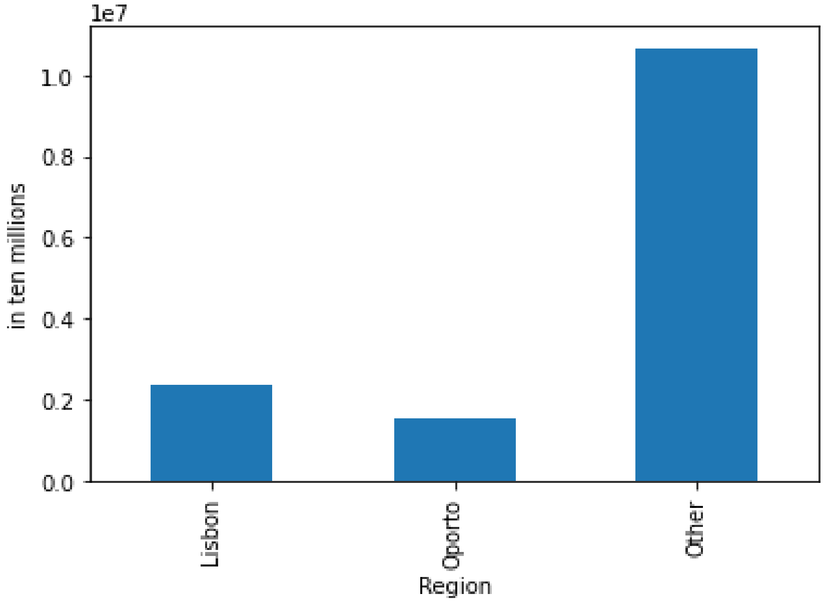
From the above table we can infer that Fresh, Grocery, Detergents Paper and Delicatessen recorded the minimum value. However fresh as recorded the maximum mean value of all, also the standard deviation for Fresh is also more which means it shows the maximum variation. Maximum value is also by 'Fresh'.

1.1.2. Which Region and which Channel spent the most?



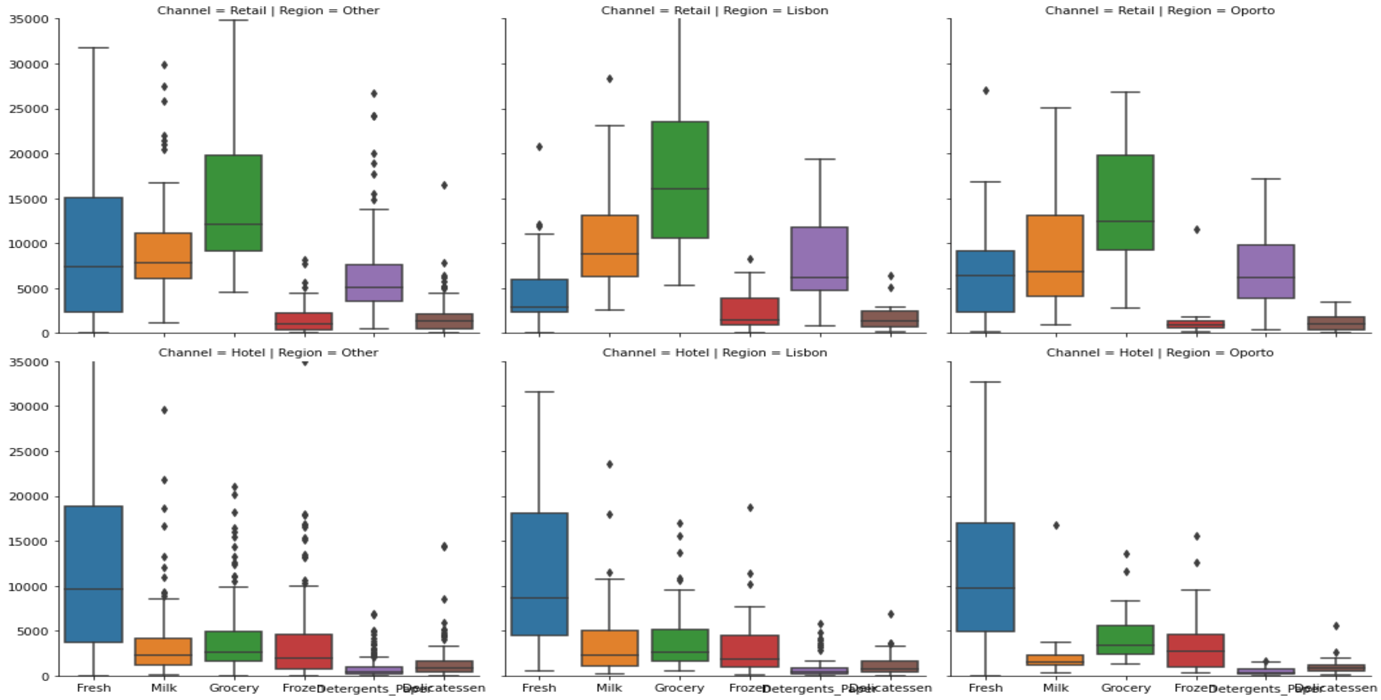
Within the channel we can see Hotel has spent the most and Retail has spent the least comparatively.

1.1.3. Which Region and which Channel spent the least?



When comparing the region, other places have spent the most and Lisbon is the second highest and Oporto is the least of all.

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.



Fresh : It has the maximum value and the maximum IQR in Other Region and the Hotel channel and the least median in the Lisbon region in Retail. Generally it has a wide IQR spread in the hotel channel compared to retail.

Milk: Milk has the maximum value in Retail in Oporto region, among the Retail it sells less in other regions. When comparing Hotels to retail it less very less in hotels as the IQR is less in hotels. It sells least in the Oporto Hotels.

Grocery: It has the maximum value in the Retail Lisbon and the least in Oporto Hotels. Generally in Retail the sales of grocery is more compared to Hotels as we can see IQR is more among the Retail.

Frozen Food: The sale of frozen food is more in hotels compared to retails. Among the hotels other regions sells maximum frozen foods and Lisbon sells the least. Overall the least sales is done by the retail in Oporto Region.

Detergent Papers: Across the entire region the sales of these papers are more in Retail sector than hotels. Within the retail Lisbon regions sells the most and other region sells the least. Within the hotels the sales are almost equally less but the minimum value is in Oporto Hotels.

Delicatessen: The overall sales of this item are relatively less compared to other items. However the sales are more in Retails than Hotels across all regions. Within the retail it has the maximum sale in Lisbon Retail and the least in Oporto Retail. Overall the minimum sales are shown in Oporto Hotels.

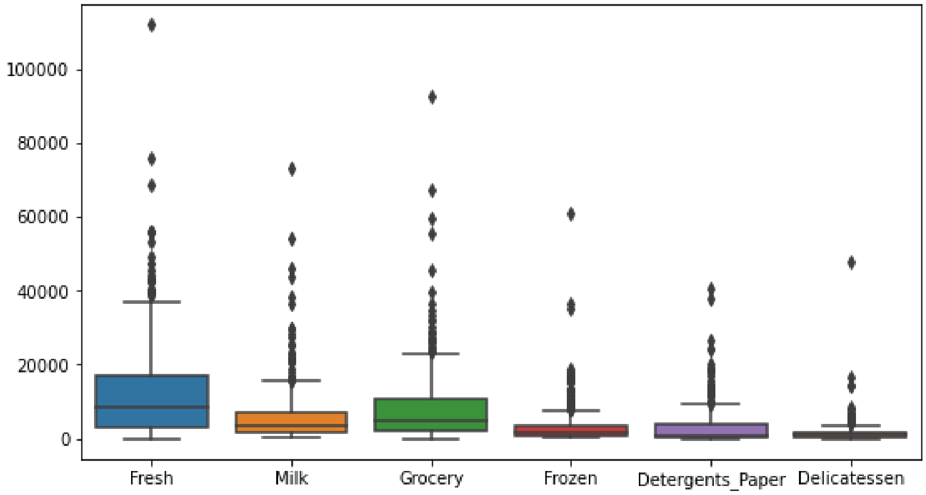
1.3. On the basis of a descriptive measure of variability, which item shows the most inconsistent behavior? Which items show the least inconsistent behavior?

```
Fresh               105.391792
Milk                127.329858
Grocery             119.517437
Frozen              158.033238
Detergents_Paper    165.464714
Delicatessen        184.940690
```

The above table shows the co efficient of variance for each variable in percentage from this the highest value is considered to be most inconsistent and the one with the least value is considered to be the least inconsistent.

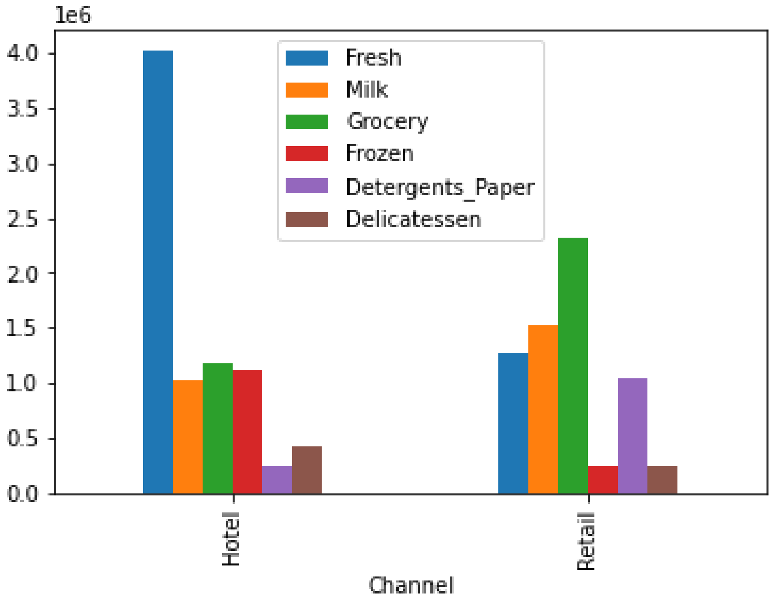
Delicatessen is the most inconsistent and fresh is the least inconsistent.

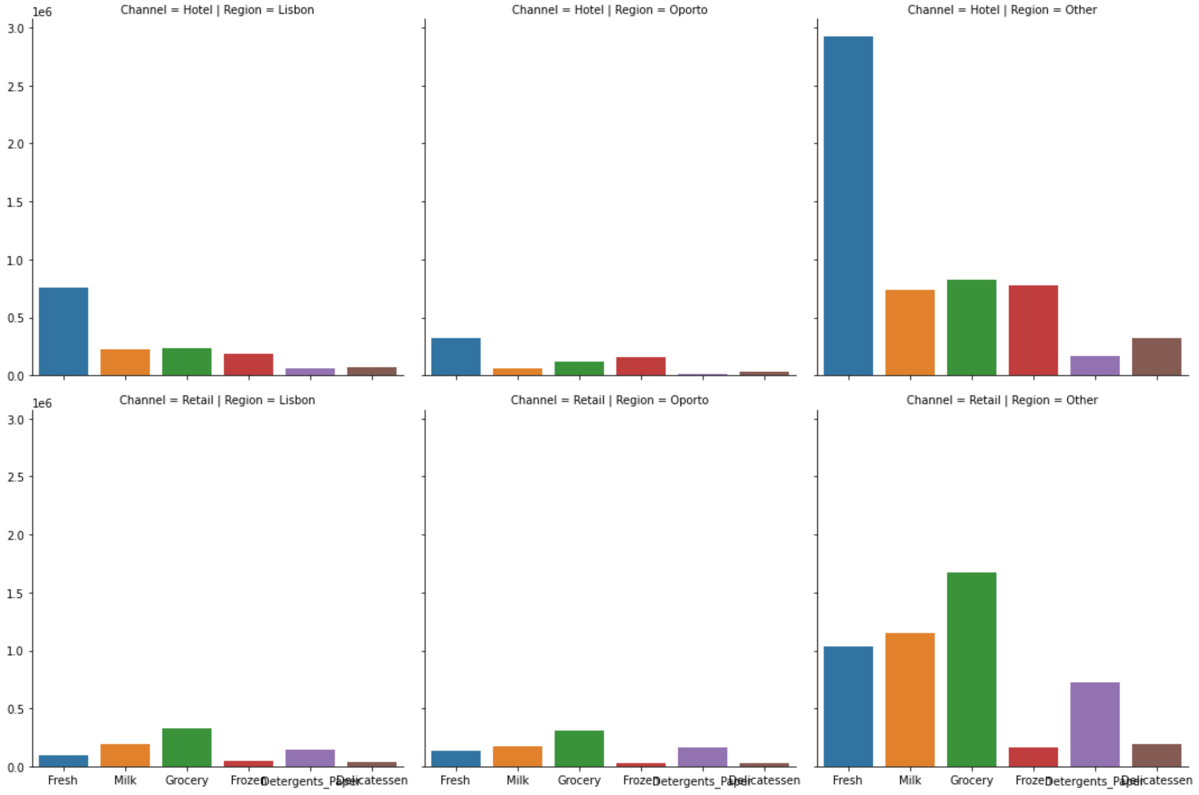
1.4. Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.



With the help of Boxplot we can clearly see there are outliers present in all six variables.

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 we can see that Retail shows fewer sales than Hotels so we can concentrate in bringing more retailers who works in retail channel. Of all the region, Oporto is the least of all that is because very less retailer are doing business in that region so we can canvas or advertise and reach out more retailers in that region.

- In Hotels of Oporto sale of milk and Grocery is less compared to Lisbon and other areas so we can try to get more retailers who sell more of milk and grocery

- In the retail side the sale of Frozen food is less we can check for the reason and improve the stocks management or marketing techniques.

- In retails of Oporto and Lisbon the sale of Fresh is less we can ask retailers to invest in fresh foods more.

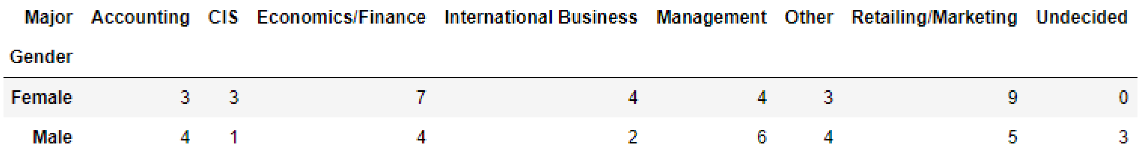
- The sale of Delicatessen is very unstable we can try to stabilize the sale by attracting more customers.

2. 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*** data set).

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

2.1.1 Gender and Major

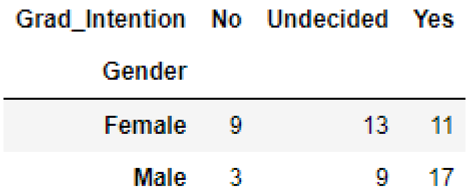


Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Gender								
Female	3	3	7	4	4	3	9	0
Male	4	1	4	2	6	4	5	3

It seems like female students prefer CIS, Economics/Finance, International Business and Retailing/Marketing and male students go for Accounting, Management.

There are no female students who haven't decided their major where as there are few male students who haven't decided yet

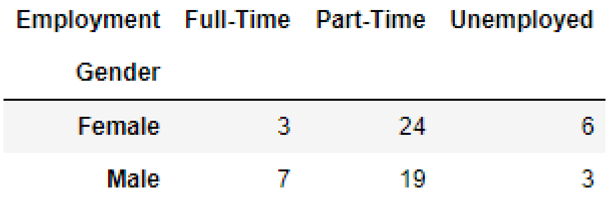
2.1.2 Gender and Grad Intention



Grad_Intention	No	Undecided	Yes
Gender			
Female	9	13	11
Male	3	9	17

Most of them have said yes, very less male students haven't decided or said no compared to female students.

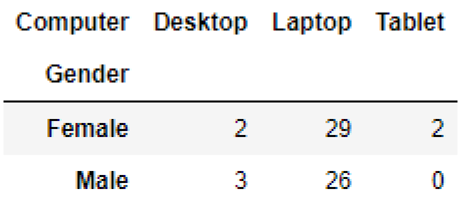
2.1.3 Gender and Employment



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

Most of the students are Part time working in general. Majority of men are full time working compared to women and majority of the women are unemployed compared to men.

2.1.4 Gender and Computer



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

Most of the students are using laptops very few students are using tablet and desktop

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

P (male) = No of male/ Total students =29/62

P (male) = 0.46774193548387094

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

P (Female) = No of female/ Total students =33/62

P (Female) = 0.532258064516129

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.

P (majors | male) =no of male students in each major / Total male

P(major|male)=

Accounting	0.137931

CIS	0.034483
Economics/Finance	0.137931
International Business	0.068966
Management	0.206897
Other	0.137931
Retailing/Marketing	0.172414
Undecided	0.103448

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

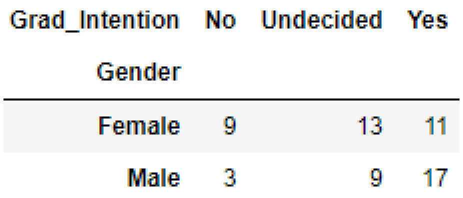
P (majors | Female) =no of female students in each major / Total female

P (major |female) =

Accounting	0.090909
CIS	0.090909
Economics/Finance	0.212121
International Business	0.121212
Management	0.121212
Other	0.090909
Retailing/Marketing	0.272727
Undecided	0.000000

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

2.4.1 Find the probability that a randomly chosen student is a male and intends to graduate.

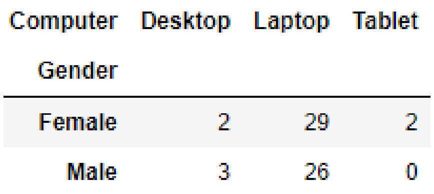


Grad_Intention	No	Undecided	Yes
Gender			
Female	9	13	11
Male	3	9	17

P (male and intend to graduate) =17/62

= 0.27419354838709675

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



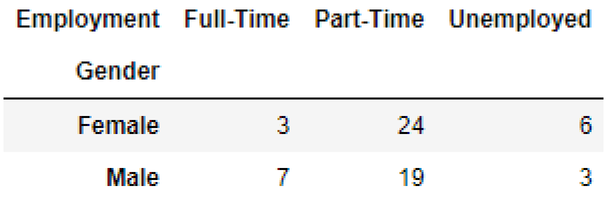
Computer	Desktop	Laptop	Tablet
Gender			
Female	2	29	2
Male	3	26	0

P (Female and not having laptop) = 4/62

=0.06451612903225806

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?



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

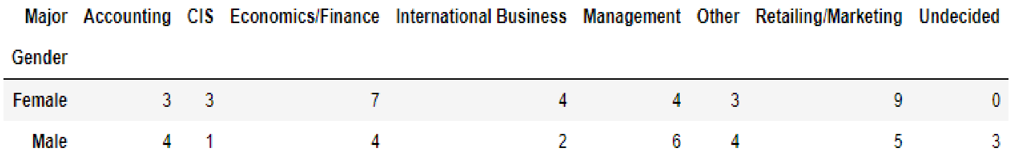
P (male or full time employment) =P (A) +P (B)-P (AnB)

= 29+10-7=32

=32/62

=0.5161290322580645

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	3	3	7	4	4	3	9	0
Male	4	1	4	2	6	4	5	3

Total number of female=33
P (international business or management | female) = P (international|Female) + P (management|Female)/Total Female
= 8/33

=0.24242424242424243

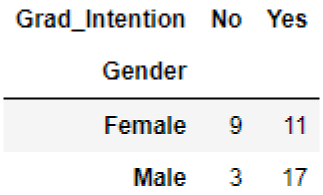
2.6 Construct a contingency table of Gender and Intent to Graduate at 2 levels (Yes/No). The Undecided students are not considered now and the table is a 2x2 table. Do you think the graduate intention and being female are independent events?

To check whether the events are dependent or not we can do a chi square test on both the variables.

Hypothesis Framing:

Ho: two variables are independent

Ha: two variables are dependent



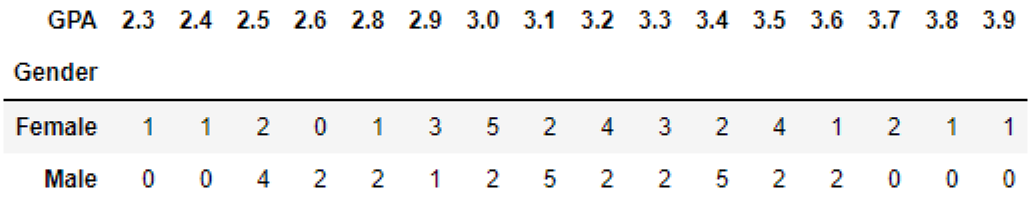
Grad_Intention	No	Yes
Gender		
Female	9	11
Male	3	17

On performing chi square test on the data the test static we obtained **=4.28571428571428** and p value **=0.03843393023678176**

Since we got the p value less than 0.05 (alpha) we have enough evidence to reject null hypothesis so we can declare that two variables are not independent events.

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

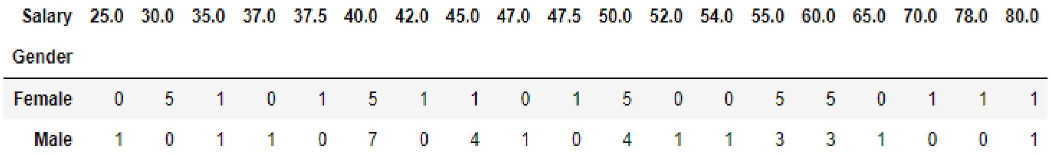
2.7.1 If a student is chosen randomly, what is the probability that his/her GPA is less than 3



GPA	2.3	2.4	2.5	2.6	2.8	2.9	3.0	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

P (Students < 3 GPA) =P (male<3 GPA) + P (Female< 3 GPA)
= (9+8)/62
=0.27419354838709675

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



Salary	25.0	30.0	35.0	37.0	37.5	40.0	42.0	45.0	47.0	47.5	50.0	52.0	54.0	55.0	60.0	65.0	70.0	78.0	80.0
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

P (Earns >= 50 |Male) = (4+1+1+3+3+1+1)/29)

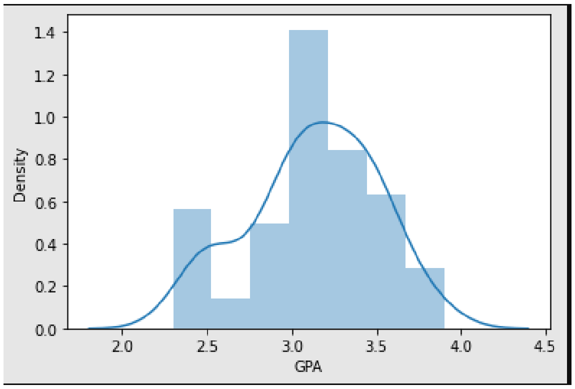
=0.4827586206896552

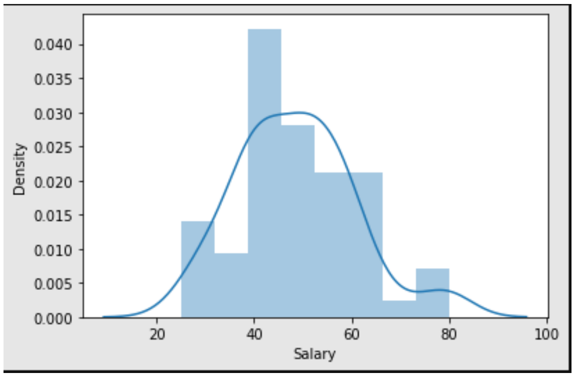
P (Earns >= 50 |Female) = (5+5+5+1+1+1)/33)

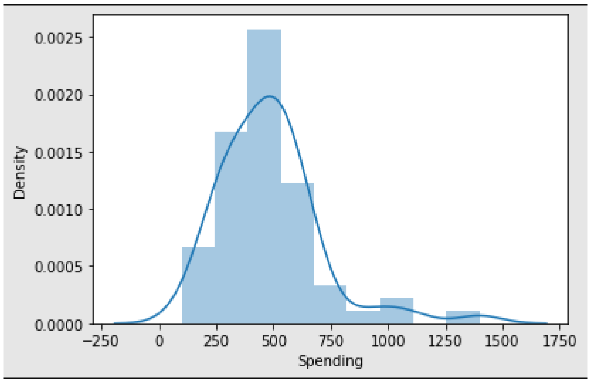
=0.5454545454545454

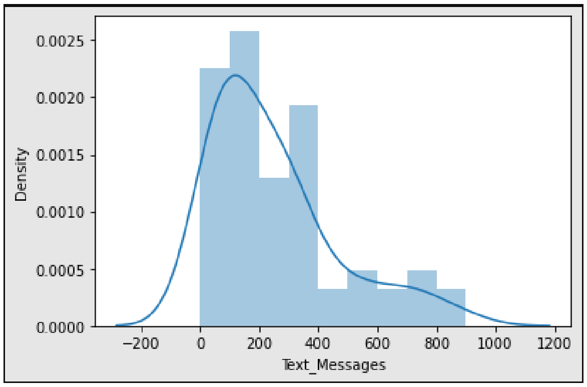
2.8 .

2.8.1 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 for this whole Problem 2



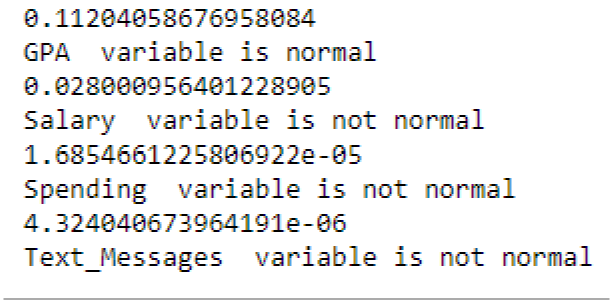






From the histograms we can see GPA tends to be almost normal, salary doesn't seem to be following normal distribution Spending and Text Messages are having right skewness to make a clear conclusion we can test for normality using Shapiro test.

Shapiro Test:
Hypothesis Framing
Ho: Distribution is normal
Ha: Distribution is not normal



```
0.11204058676958084
GPA  variable is normal
0.028000956401228905
Salary  variable is not normal
1.6854661225806922e-05
Spending  variable is not normal
4.324040673964191e-06
Text_Messages  variable is not normal
```

Based on the test, p value and based on the p value the result has been shown in the output. The following output confirms that GPA is normal and Salary, Spending, Text Messages are not normal.

2.8.2 Write a note summarizing your conclusions for this whole Problem 2.

In general female students prefer CIS, Economics/Finance and International Business and male students prefer accounting and management. Most of the female student hasn't decided about graduation. Most of the students studying are doing part time and majority of the full time workers are male. All most all of the students prefer laptops to study. There is a relationship between gender and intend to graduation that means being male or female might help us to decide whether they tend to graduate or not. The average GPA by the students is more than 3 and 50% of the students earn more than 50. Among all the variables only GPA tends to be normal.

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

The file (A & B shingles.csv) includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.

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.

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:

FOR SHINGLES A:

STEP 1: Framing the Ho and Ha

Ho: the mean of shingles <= 0.35

Ha: the mean of shingles >0.35

STEP 2: Deciding the alpha value 0.05

STEP 3: Deciding the test, in this case the test statistic is one sample T-test

STEP 4: Computing the test statistic and p value

Test statistic=-1.4735046253382782 and p value=0.07477633144907513

Conclusion: Since p value is not less than 0.05 (alpha) we fail to reject the null hypothesis, which means the mean of the shingles tested has the mean <=0.35.

At 95% confidence interval we have the evidence to say that the mean of shingles A is less than or equal to 0.35. So the company's claim stays true for this case

FOR SHINGLES B:

Ho: the mean of shingles <= 0.35

Ha: the mean of shingles >0.35

We are following the same steps above for sample B also .The statistic and p value are given below.

Test statistic=-3.1003313069986995 and p value=0.0020904774003191826

Conclusion: As we can see the p value is less than 0.05(alpha) we have enough evidence to reject the null hypothesis.

At 95% confidence interval we have the evidence to say that the mean of Shingles B is greater than 0.35 .The Company’s claim is false for this case.

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?

Assumptions to be made are:

> The samples must be continuous

> The sample is derived from a population in a random manner.

> Data plotted follows a normal distribution

> Equal variance between the two samples

STEP 1: Hypothesis framing

Ho: mean of A= mean of B

Ha: mean of A != mean of B

STEP 2: choosing level of significance 0.05

STEP 3: choosing the test statistic (paired sample t-test)

STEP 4: calculating test statistic and p value

Test statistic=1.2896282719661123 and p value=0.2017496571835306

P value is greater than 0.05 so we do not have enough evidence to reject null hypothesis.

At 95% confidence interval we have enough evidence to say that the population means of Shingles A and Shingles B are equal.