Statistical Methods for Decision Making

SMDM Project

ABSTRACT

This project, will show the analysis of three different case-study which includes,
(i) Wholesale distributer's food items.
(ii) CMSU's Graduated Student .
(iii) Manufacturing company ABC's shingles quality.

Selvathilagaraj

PGP-DSBA

August' 21

Table of Content

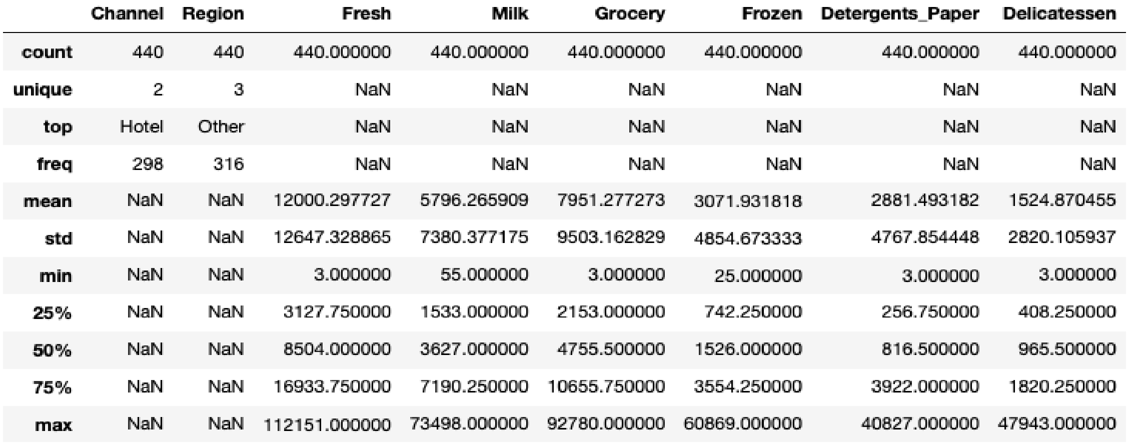
SMDM Project

Problem 1:

Wholesale Customers Analysis
Problem Statement:

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 Use methods of descriptive statistics to summarize data.



	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440	440	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
unique	2	3	NaN	NaN	NaN	NaN	NaN	NaN
top	Hotel	Other	NaN	NaN	NaN	NaN	NaN	NaN
freq	298	316	NaN	NaN	NaN	NaN	NaN	NaN
mean	NaN	NaN	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	NaN	NaN	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	NaN	NaN	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	NaN	NaN	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000
50%	NaN	NaN	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000
75%	NaN	NaN	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000
max	NaN	NaN	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000

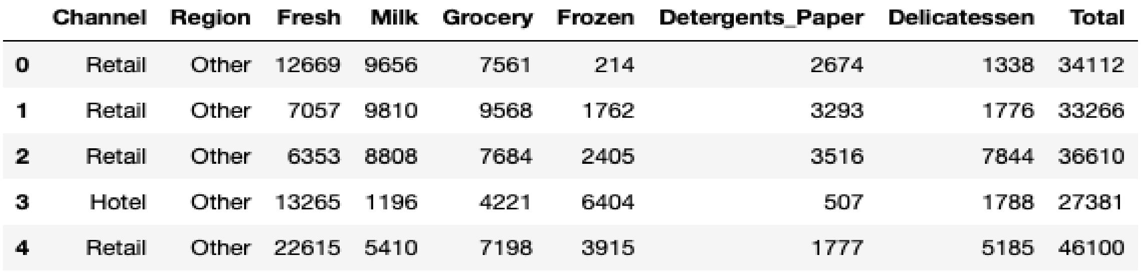
Summarization:

There are 6 numerical columns and 2 non-numerical columns, With no Null values.

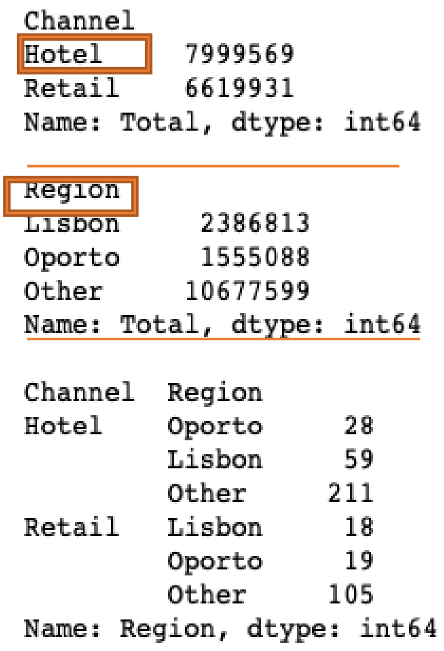
The mean is larger than the median, so it might be the Right skewed, hence the items spending at the OTHER Region is more when compared to Lisbon & Oporto

Max value is more than the Q3, so it is clearly infers that all the items have outliers

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



	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
0	Retail	Other	12669	9656	7561	214	2674	1338	34112
1	Retail	Other	7057	9810	9568	1762	3293	1776	33266
2	Retail	Other	6353	8808	7684	2405	3516	7844	36610
3	Hotel	Other	13265	1196	4221	6404	507	1788	27381
4	Retail	Other	22615	5410	7198	3915	1777	5185	46100



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

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

Channel  Region
Hotel    Oporto      28
         Lisbon      59
         Other      211
Retail   Lisbon      18
         Oporto      19
         Other      105
Name: Region, dtype: int64
```

Approach used:

Added new column TOTAL and cumulated all numeric variables.

Then grouped by channel and region individually

Infer Based on Channel wise

Hotel channel annual spend is nearly **8 Million USD**, which is the **highest spending**.

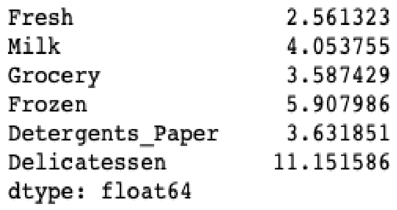
Retail channel annual spend is nearly **6.5 Million USD**, which is the **least spending**.

Infer Based on Region wise

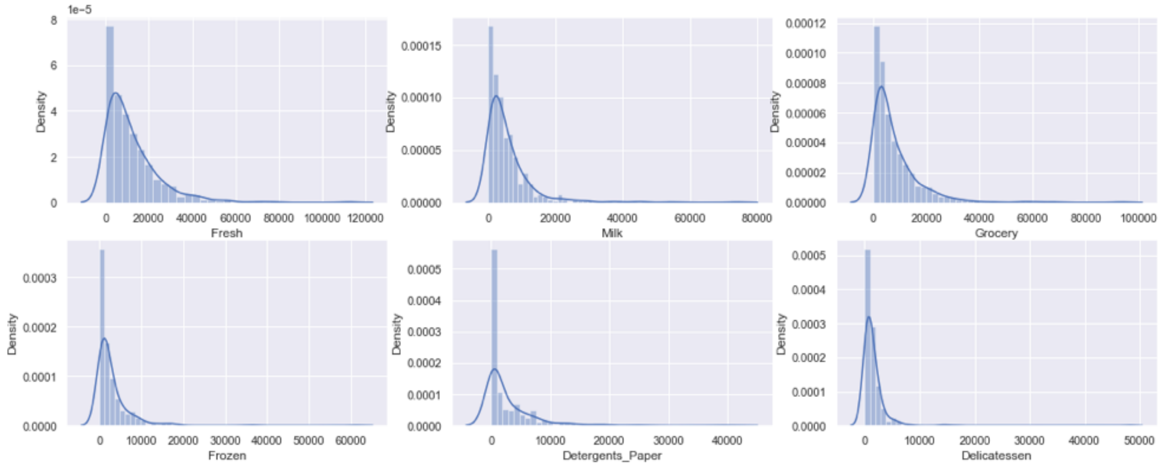
compared to Oporto and Lisbon; **OTHER** region annual spend is **highest spending**, which is 10.5 Million USD.

whereas, the **least spend** region is **OPORTO** which is 1.5 Million USD.

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                2.561323
Milk                 4.053755
Grocery              3.587429
Frozen               5.907986
Detergents_Paper     3.631851
Delicatessen        11.151586
dtype: float64
```



```
Region   Lisbon  Oporto    Other
Channel
Hotel    761233  326215  2928269
Retail    93600  138506  1032308
```

Total Spending for Fresh item is : 5,280,131 #1

```
Region   Lisbon  Oporto    Other
Channel
Hotel    228342   64519   735753
Retail   194112  174625  1153006
```

Total Spending for Milk item is : 2,550,357 #3

```
Region   Lisbon  Oporto    Other
Channel
Hotel    237542  123074   820101
Retail   332495  310200  1675150
```

Total Spending for Grocery item is : 3,498,562 #2

```
Region   Lisbon  Oporto   Other
Channel
Hotel    184512  160861  771606
Retail    46514   29271  158886
```

Total Spending for Frozen item is : 1,351,650 #4

```
Region   Lisbon  Oporto   Other
Channel
Hotel     56081   13516  165990
Retail   148055  159795  724420
```

Total Spending for Detergents-Paper item is : 1,267,857 #5

```
Region   Lisbon  Oporto   Other
Channel
Hotel     70632   30965  320358
Retail    33695   23541  191752
```

Total Spending for Delicatessen item is : 670,943 #6

Approach used:

Checked the skewness
Using pivot table, derived the output based on Each food items.
Insights:
Data are skewed right, magnitude of data is not symmetrical hence the items spending at the OTHER Region is more when compared to Lisbon & Oporto

Inference:

The wholesale distributor **spends most** for the **Fresh** Items then followed by **Grocery and Milk** in top three.

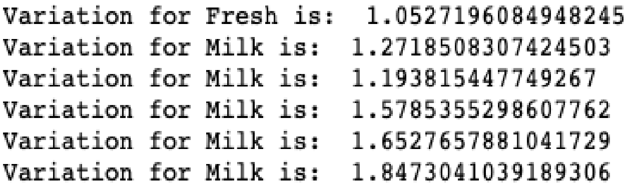
Whereas the **least spending** of the wholesale distributor is for **Delicatessen**.

Also the wholesale distributor almost **equally spend** for **Frozen and Detergents-Paper**.

On the whole we can clearly say the wholesale distributer spends ***mostly in Hotels as compared to Retail***.

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?

Aproached used:
Since the magnitude of data is not equal, so CV is the best option to describe the behaviour of the products.



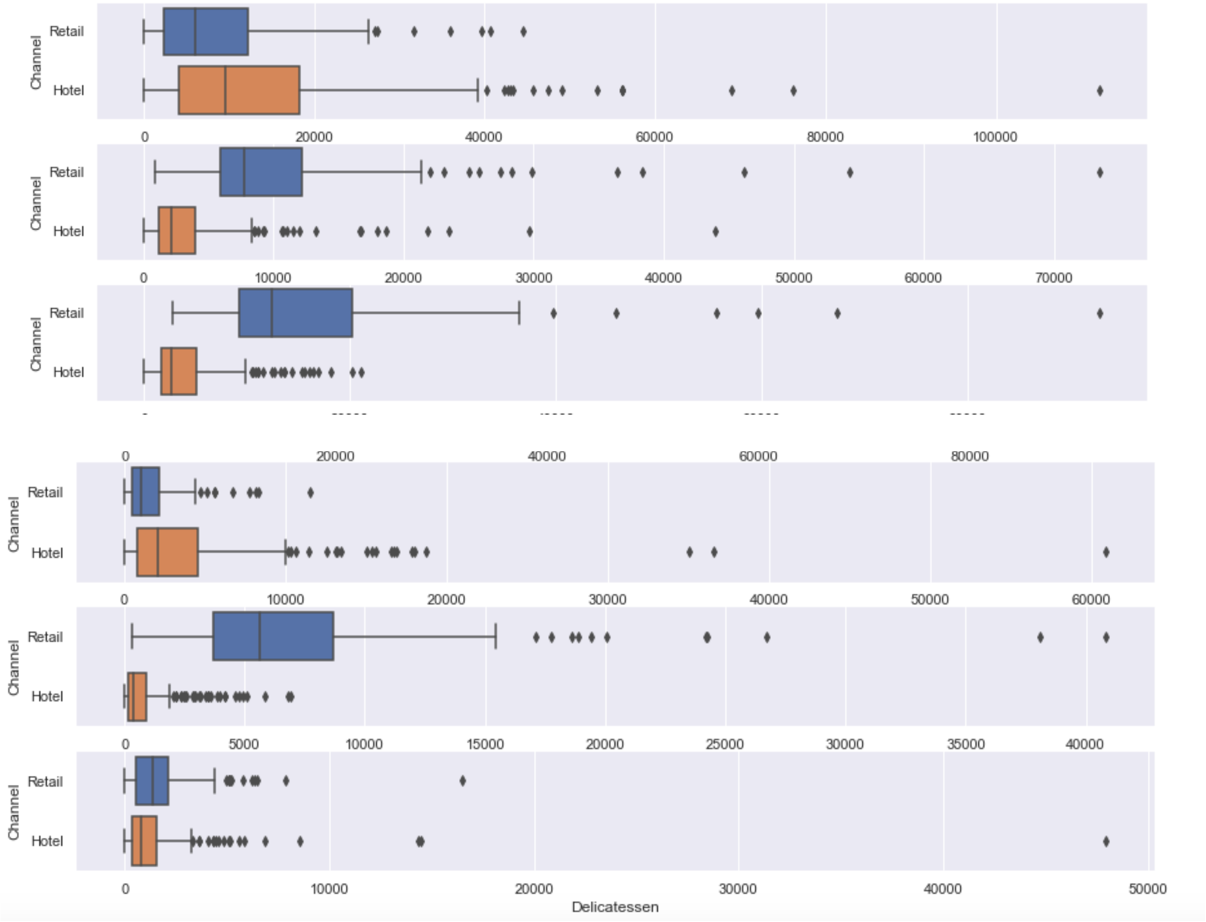
```
Variation for Fresh is:  1.0527196084948245
Variation for Milk is:  1.2718508307424503
Variation for Milk is:  1.193815447749267
Variation for Milk is:  1.5785355298607762
Variation for Milk is:  1.6527657881041729
Variation for Milk is:  1.8473041039189306
```

Insights:
Lower CV is lower risk.
Inference:
Using Coefficient of Variation we find out the least Category value is for “Fresh” items (1.05) and highest Category value is for “Delicatessen” (1.84)

So from the given data, it is clear that **most inconsistent behaviour** shown by item – **Delicatessen**

And **least inconsistent behaviour** shown by item – **Fresh.** *(More profit)*

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



Form the above box plot it is clear that , ***Outliers are present in all the food items*.**

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

As per the analysis,

I can say that, there are inconsistencies in spending of different items (by calculating Coefficient of Variation), which should be minimized. The spending of Hotel and Retail channel are irrational, but it should be more or less equal.

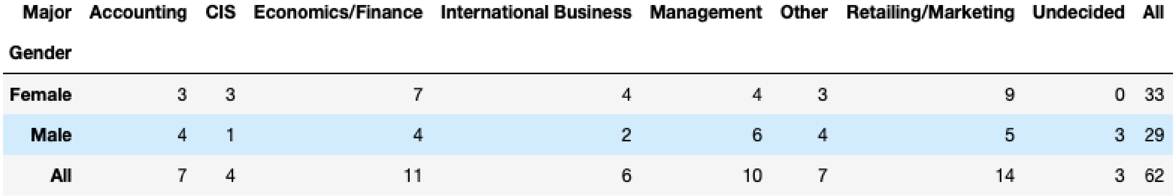
Also spending should be equal for different regions. The distributer needs to focus on other items also other than “Fresh” and “Grocery” items

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).

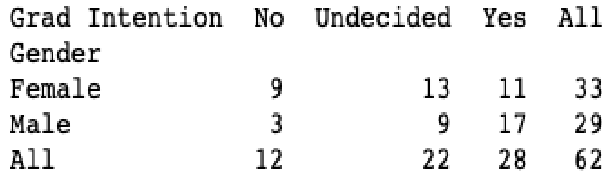
2.1. For this data, construct the following contingency tables (Keep Gender as row variable)

2.1.1. Gender and Major



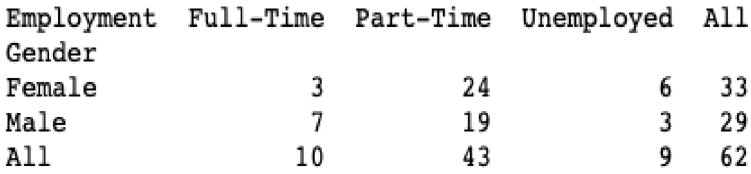
Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	All
Gender									
Female	3	3	7	4	4	3	9	0	33
Male	4	1	4	2	6	4	5	3	29
All	7	4	11	6	10	7	14	3	62

2.1.2. Gender and Grad Intention



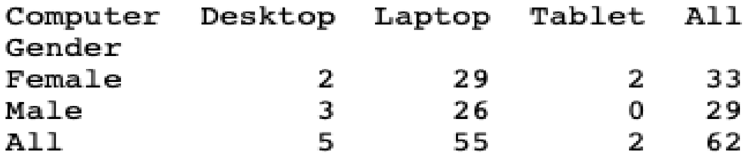
```
Grad Intention  No  Undecided  Yes  All
Gender
Female           9         13   11   33
Male             3          9   17   29
All             12         22   28   62
```

2.1.3. Gender and Employment



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

2.1.4. Gender and Computer



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

2.2. Assume that the sample is 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?

From the dataset
Total # of Female students= 33
Total # of Male students =29
Total # of students= 62
Probability that a randomly selected CMSU student will be male=
Total # of Male students / Total # of students

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?

From the dataset
Total # of Female students= 33
Total # of Male students =29
Total # of students= 62
Probability that a randomly selected CMSU student will be female=
Total # of Female students / Total # of students

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.

Total # of male=29
'Probability of students choose Accounting stream among Males is', 4/29* 100
'Probability of students choose CIS stream among Males is', 1/29* 100
'Probability of students choose Economics/Finance stream among Males is', 4/29* 100
'Probability of students choose International Business stream among Males is', 2/29* 100
'Probability of students choose Management stream among Males is', 6/29* 100
'Probability of students choose Other stream among Males is', 4/29* 100
'Probability of students choose Retailing/Marketing stream among Males is', 5/29* 100
'Probability of students choose Undecided stream among Males is', 3/29* 100

Hence:

Probability of students choose **Accounting** stream among Males is **13.79%**
Probability of students choose **CIS** stream among Males is **3.44%**
Probability of students choose **Economics/Finance** stream among Males is **13.79%**
Probability of students choose **International Business** stream among Males is **6.89%**
Probability of students choose **Management** stream among Males is **20.68%**
Probability of students choose **Other** stream among Males is **13.79%**
Probability of students choose **Retailing/Marketing** stream among Males is **17.24%**
Probability of students choose **Undecided** stream among Males is **10.34%**

Inference

From this output we can easily say that most of the **males** students prefer **Management Majors** as majority and **CIS** is the **least** preferred one

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

Total # of Female=33

'Probability of students choose Accounting stream among Males is', 3/33* 100
'Probability of students choose CIS stream among Males is', 3/33* 100
'Probability of students choose Economics/Finance stream among Males is', 7/33* 100
'Probability of students choose International Business stream among Males is', 4/33* 100
'Probability of students choose Management stream among Males is', 4/33* 100
'Probability of students choose Other stream among Males is', 3/33* 100
'Probability of students choose Retailing/Marketing stream among Males is', 9/33* 100
'Probability of students choose Undecided stream among Males is', 0/33* 100

Hence

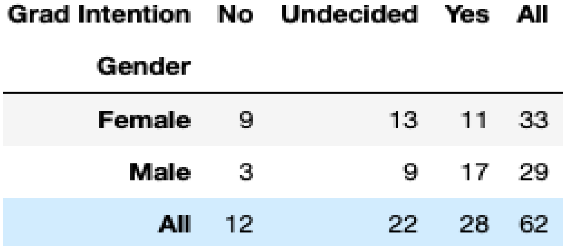
Probability of students choose **Accounting** stream among Males is **9.09%**
Probability of students choose **CIS** stream among Males is **9.09%**
Probability of students choose **Economics/Finance** stream among Males is **21.21%**
Probability of students choose **International Business** stream among Males is **12.12%**
Probability of students choose **Management** stream among Males is **12.12%**
Probability of students choose **Other** stream among Males is **9.09%**
Probability of students choose **Retailing/Marketing** stream among Males is **27.27%**
Probability of students choose **Undecided** stream among Males is **0.0**

Inference

From this output we can easily say that most of the **females** students prefer **Retailing/Marketing** as Majors.

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.

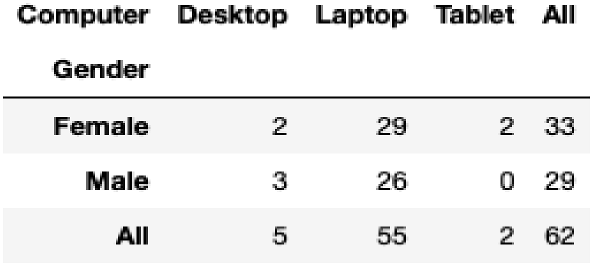


Grad Intention	No	Undecided	Yes	All
Gender				
Female	9	13	11	33
Male	3	9	17	29
All	12	22	28	62

Probability that a randomly chosen student is a male
a=29/62
Probability that a randomly chosen student is a male and intends to graduate
b=17/29
a * b

Probability That a randomly chosen student is a male and intend graduate is: **0.27419 or 27.419%**

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



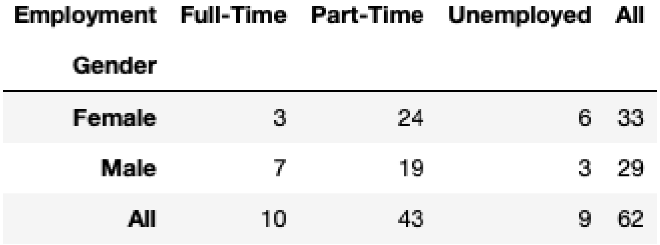
Computer	Desktop	Laptop	Tablet	All
Gender				
Female	2	29	2	33
Male	3	26	0	29
All	5	55	2	62

Probability that a randomly selected student is a female
a=33/62
Probability that a randomly selected student is a female and does NOT have a laptop
b=4/33
a * b

<u>Probability that a randomly selected student is a female and does NOT have a laptop is</u>**: 0 .06451 or 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?



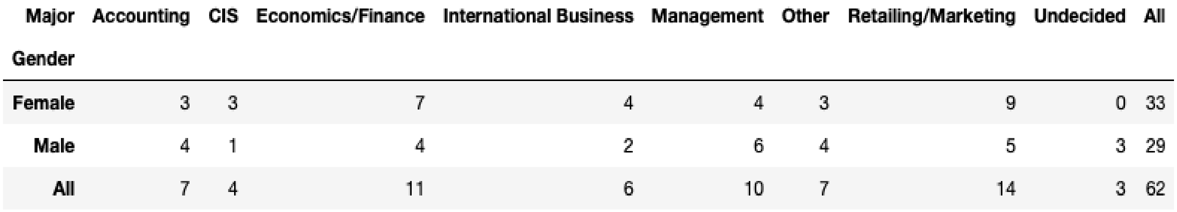
Employment	Full-Time	Part-Time	Unemployed	All
Gender				
Female	3	24	6	33
Male	7	19	3	29
All	10	43	9	62

Since it is not mutually exclusive
p(A u B)=p(A)+p(B)-p(A n B), will be the best option.
Probability that a randomly chosen student is a male
A=29/62
Probability that a randomly chosen student has full-time employment
B=10/62
Probability that a randomly chosen male has full-time employment
(A n B)=7/29

P(A u B)=A+B-(A n B)

<u>Probability that a randomly chosen student is a male or has full-time employment is :</u>
0.38765 or. 38.76%

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



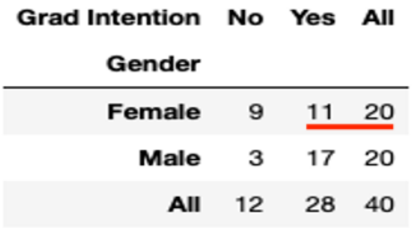
Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	All
Gender									
Female	3	3	7	4	4	3	9	0	33
Male	4	1	4	2	6	4	5	3	29
All	7	4	11	6	10	7	14	3	62

Probability that given a female student
C= 33/62
Female student is majoring in international business
A= 4/33
Female student is majoring in management
B=4/33

Since , we need to find the Majoring of female in either international business or management
P(A U B)= P(A) + P(B)
(4/33)+(4/33)
<u>Probability that given a female student is randomly chosen, she is majoring in international business or management is</u> : **0.242424 or 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?

Making a subset of dataset where *Grad Intention* column has been dropped by containing string ‘*Undecided’*



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

To Prove that, the graduate intention and being female are independent events…
We need to prove this formula: (A n B) = A * B for independent event)

Probability of being female
A=20/40
Probability of graduate intention
B=28/40
Graduate intention and being female

(A n B) = 11/20 = 0.55 (observed from the given contingency table)

(A n B)   =(20/40) * (28/40) = 0.35

(11/20)0.55 != 0.35

Here, p(A n B) is not equal to p(A) * p(B), which is clearly saying that the graduate intention and being female 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?

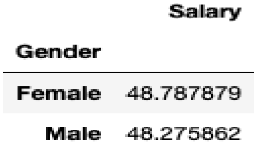
From the given data we know the sample. so it is **discreet distribution**. Also **Mean and stdev** can able to derive using GPA column.
So **Poisson Distribution** will be the best solution

We need to find the probability that his/her GPA is less than 3
(i.e) p(x) < 3. so cumulative of 2 will give the answer

Applying the derived inputs in Poisson distribution in Python the output is **0.39491 or 39.49%**

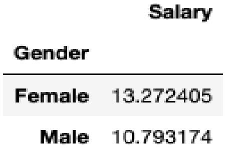
2.7.2. Find the conditional probability that a randomly selected male earns 50 or more. Find the conditional probability that a randomly selected female earns 50 or more.

Mean



	Salary
Gender	
Female	48.787879
Male	48.275862

Stdev



	Salary
Gender	
Female	13.272405
Male	10.793174

Probability that a randomly selected male earns 50 or more (i.e. p(x1)>=50)
Cumulative of selected male till **50** will help to find 50 or more.
Substituting Male sample mean and sigma in normal distribution will give the probability

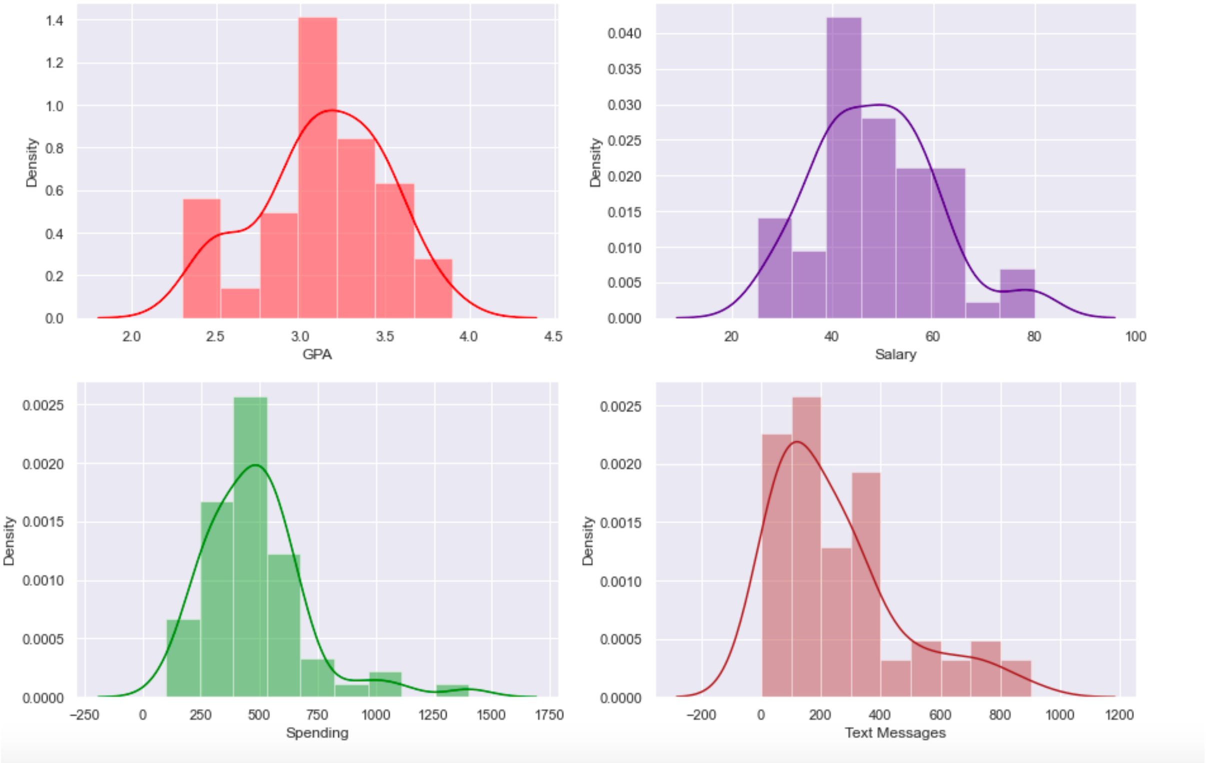
Applying the derived inputs in Normal distribution in Python.

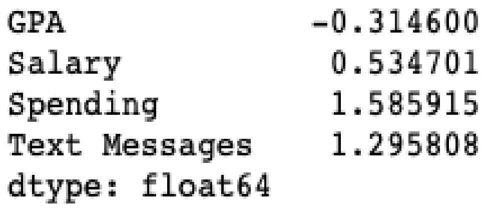
Probability that a randomly selected male earns 50 or more is **0.43654 or 43.65%**

Similarly, by substituting Female sample mean and sigma in normal distribution will give the pr obability

Probability that a randomly selected Female earns 50 or more is **0.46361 or 46.61%**

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.





```
GPA              -0.314600
Salary            0.534701
Spending          1.585915
Text Messages     1.295808
dtype: float64
```

GPA (Left Skewed):= Student's Average GPA is close to normal distribution, but ***few*** students getting less GPA. Majority of students getting good GPA.

Salary (Right Skewed):= Student's Average Salary is very close to normal distribution, Most of the students getting good Salary, whereas very few students getting less salary.

Spending (Right Skewed):=From the given data set, Students spending more money and very few students spending less.

Text Messages (Right Skewed):=similarly, Students sending more Text messages, only very few are sending less messages.

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

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.

The company would like to show that the mean moisture content is less than 0.35 pounds per 100 square feet. based on the data the company like to show (i.e. status quo)

So **Ho<=0.35**(based on historical data)

Customers feels that they have purchased a product lacking in quality, which means the moisture is more than 0.35

So **H(a)>0.35**

<u>Approach:</u>
Since population stdev is not given; T- Test is best option

Also it is a 1-tail test(u > 0.35)

Since significance level(alpha) is not there we can take 0.05 by default

<u>T- Test for A Shingles:</u>
Python output for Test stats and P-Value is
-1.4735046253382782 || 0.14955266289815025

Since **P-Value is > than significance level**, <u>failed to reject Ho</u> ,

<u>Inference:</u>
For Shingles 'A', the company have enough evidence to show that the mean moisture content is less than 0.35 pounds per 100 square feet.

<u>T- Test for B Shingles</u>
Python output for Test stats and P-Value is
-4.311710524179449 0.00012557068120902648

Since **P-Value is < than significance level**, <u>reject Ho</u> ,

<u>Inference:</u>
For shingles 'B', there is enough evidence to show Customers feeling towards the product quality is lacking (i.e. the moisture content is more than 0.35)

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?

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

so **Ho(mu) population mean for shingles A = population mean for shingles B**

Eventually, **H1(mu) population mean for shingles A != population mean for** shingles B

Approach:

Here we need to compare both the Shingles, so Two sample T-Test will be the best option.

Based on python output, below is the values.
Test stats=2.3257710269401746, P-Value=0.023007859248632315

since **P-Value is < than significance level**, reject Ho ,

Hence the population mean for shingles A is not equal to population mean for shingles B.

My assumption is:
Shingles 'A' having more number of measurement compared to Shingles B.

If Shingles 'B' also perform same number of measurement, there is more chances for equality of means, where the Company will have more ***strong evidence to show that Both shingles contains the moisture level within permissible limit*** (i.e. less than 0.35 pounds per 100 square feet)

********************************End-of-project******************************