SMDM PROJECT REPORT

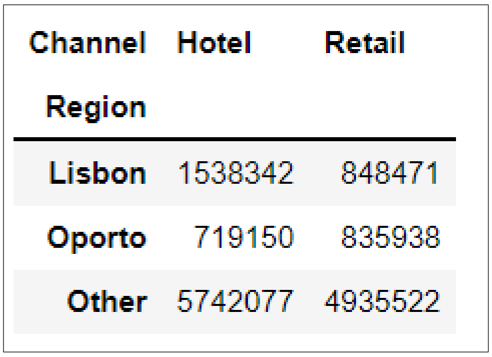
Wholesale Customers Analysis

Problem Statement:

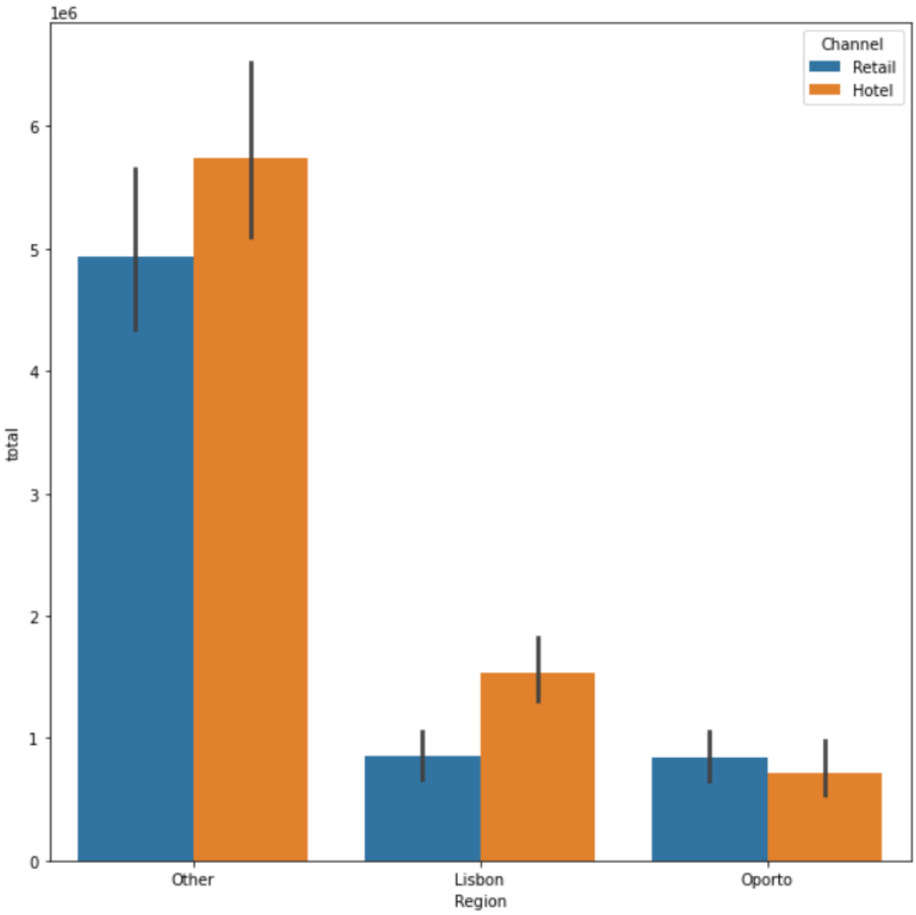
A wholesale distributor operating in different regions of Portugal has information on annual spending of several items in their stores across different regions and channels. The data consists of 440 large retailers' annual spending on 6 different varieties of products in 3 different regions (Lisbon,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

1.1 Use methods of descriptive statistics to summarize data. Which Region and which Channel spent the most? Which Region and which Channel spent the least?

The Hotel Channel in the Other Region spent the most and the Hotel channel in Oppoto Region spent the Least.



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

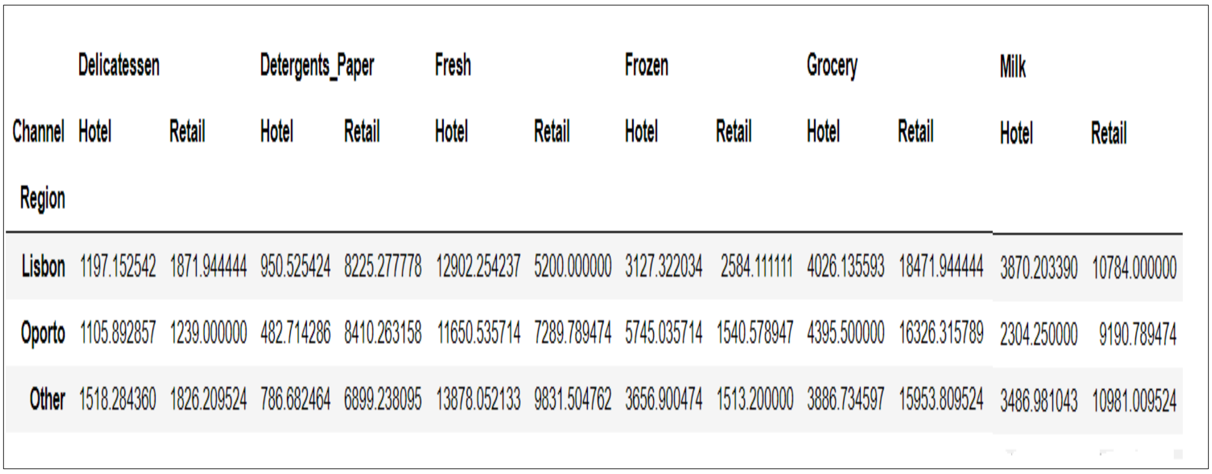


1.2 There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.

To analyse the spends of various items across Region and Channel Wise, I have used average spending instead of total spending. Because the no of distributors for each channel are not same.

And the below table shows the average spends of various items across Region and Channel Wise.

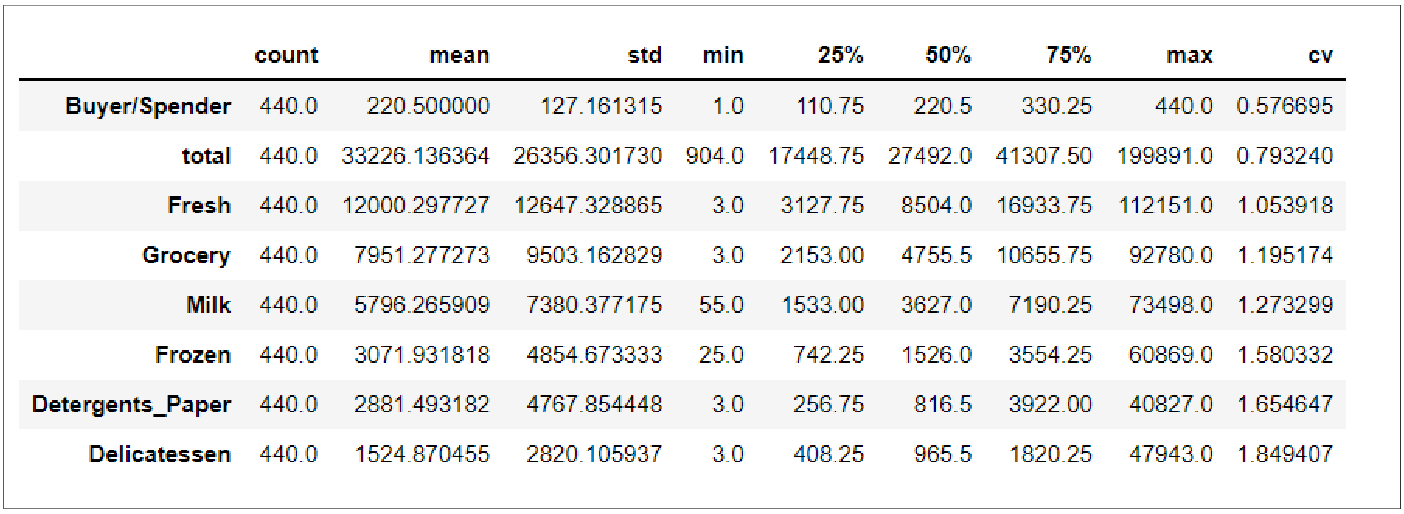
Item/Channel	Hotel			Retail		
	Lisbon	Oporto	Others	Lisbon	Oporto	Others
Delicatessen		Min	Max	Max	Min	
Detergents_Paper	Max	Min			Max	Min
Fresh		Min	Max	Min		Max
Frozen	Min	Max		Min	Max	
Grocery		Max	Min	Max		Min
Milk	Max	Min			Min	Max



	Delicatessen		Detergents_Paper		Fresh		Frozen		Grocery		Milk	
Channel	Hotel	Retail	Hotel	Retail	Hotel	Retail	Hotel	Retail	Hotel	Retail	Hotel	Retail
Region												
Lisbon	1197.152542	1871.944444	950.525424	8225.277778	12902.254237	5200.000000	3127.322034	2584.111111	4026.135593	18471.944444	3870.203390	10784.000000
Oporto	1105.892857	1239.000000	482.714286	8410.263158	11650.535714	7289.789474	5745.035714	1540.578947	4395.500000	16326.315789	2304.250000	9190.789474
Other	1518.284360	1826.209524	786.682464	6899.238095	13878.052133	9831.504762	3656.900474	1513.200000	3886.734597	15953.809524	3486.981043	10981.009524

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?

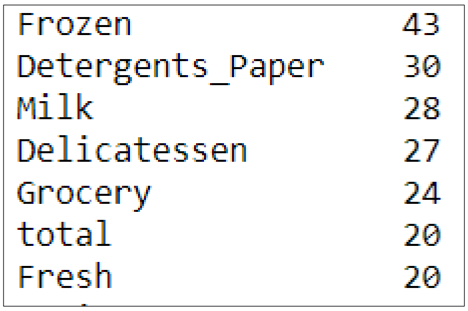
The Fresh item is least inconsistent and the Delicatessen item shows the most inconsistent behaviour.



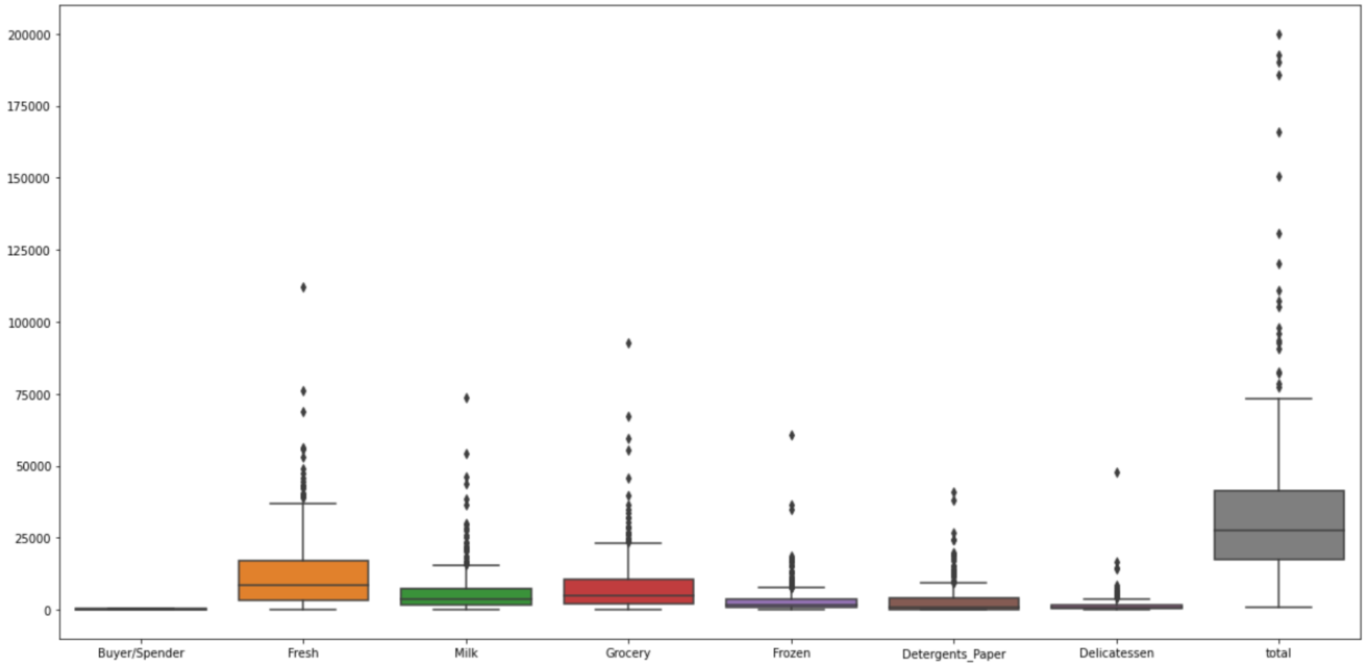
	count	mean	std	min	25%	50%	75%	max	cv
Buyer/Spender	440.0	220.500000	127.161315	1.0	110.75	220.5	330.25	440.0	0.576695
total	440.0	33226.136364	26356.301730	904.0	17448.75	27492.0	41307.50	199891.0	0.793240
Fresh	440.0	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0	1.053918
Grocery	440.0	7951.277273	9503.162829	3.0	2153.00	4755.5	10655.75	92780.0	1.195174
Milk	440.0	5796.265909	7380.377175	55.0	1533.00	3627.0	7190.25	73498.0	1.273299
Frozen	440.0	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0	1.580332
Detergents_Paper	440.0	2881.493182	4767.854448	3.0	256.75	816.5	3922.00	40827.0	1.654647
Delicatessen	440.0	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0	1.849407

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

Yes, there are outliers in the given dataset. And the number of outliers in each item are mentioned below.



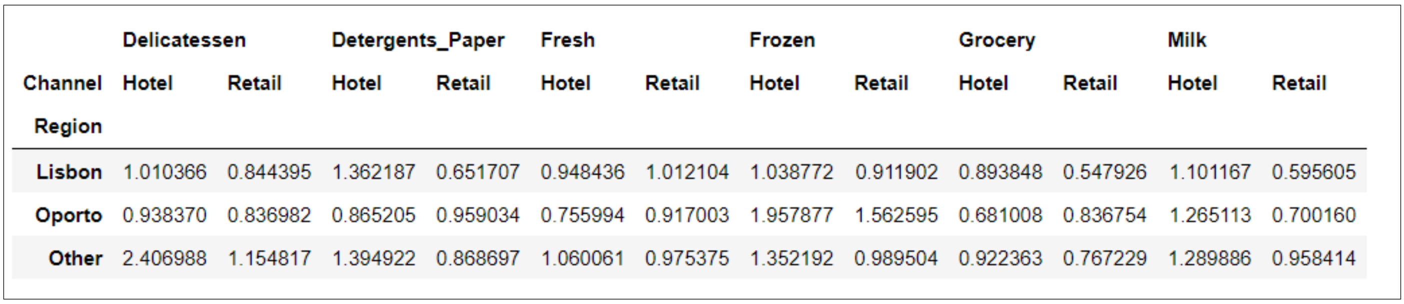
```
Frozen              43
Detergents_Paper    30
Milk                28
Delicatessen        27
Grocery             24
total               20
Fresh               20
```



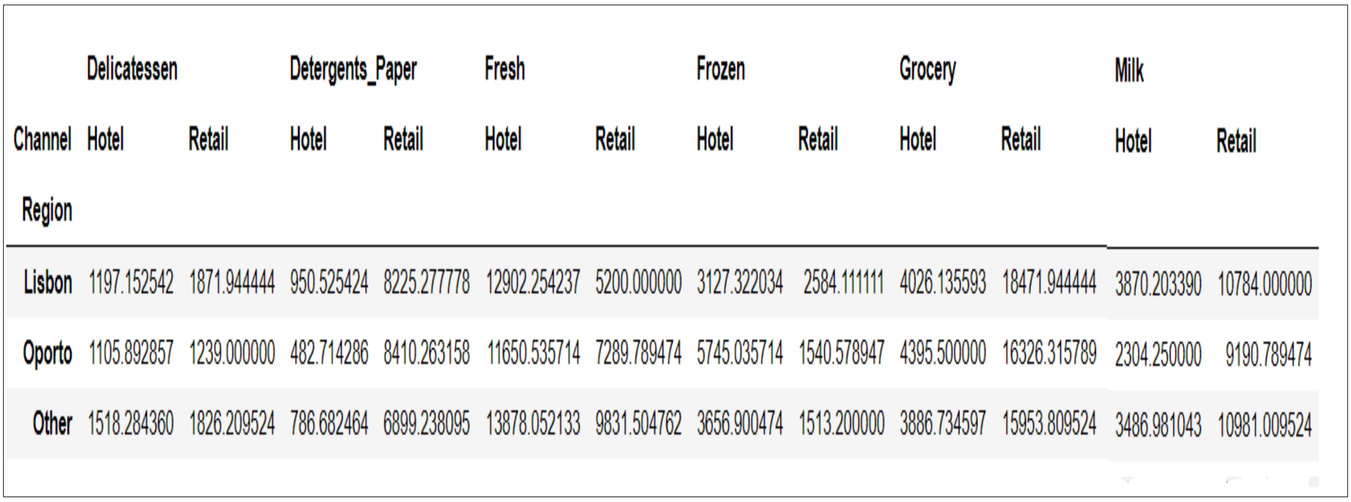
1.5 On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem? Answer from the business perspective

By Creating Pivot Table for CV of all items, Channel and Region wise. It will help us to analyse the spending and suggesting the proper investment based on CV.

From the below two outputs we can understand that, which item in which channel and region is more-inconsistent and by comparing this inconsistency with average spends we can decide how to spend wisely in future. This will improve the profits of the business as we will invest in the item which is being sold most.



	Delicatessen		Detergents_Paper		Fresh		Frozen		Grocery		Milk	
Channel	Hotel	Retail	Hotel	Retail	Hotel	Retail	Hotel	Retail	Hotel	Retail	Hotel	Retail
Region												
Lisbon	1.010366	0.844395	1.362187	0.651707	0.948436	1.012104	1.038772	0.911902	0.893848	0.547926	1.101167	0.595605
Oporto	0.938370	0.836982	0.865205	0.959034	0.755994	0.917003	1.957877	1.562595	0.681008	0.836754	1.265113	0.700160
Other	2.406988	1.154817	1.394922	0.868697	1.060061	0.975375	1.352192	0.989504	0.922363	0.767229	1.289886	0.958414



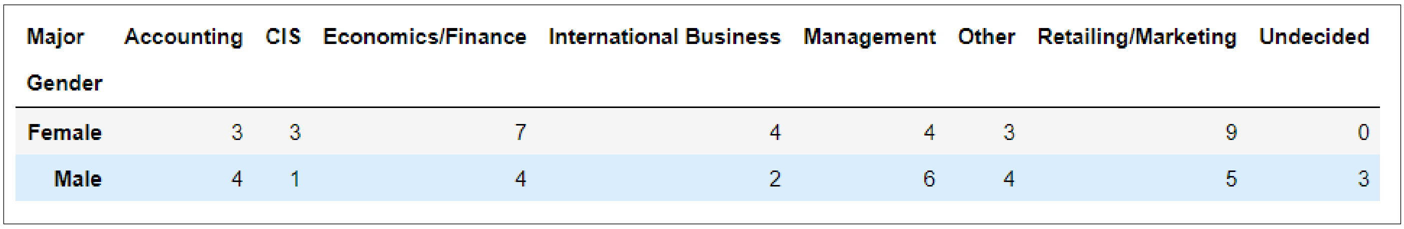
	Delicatessen		Detergents_Paper		Fresh		Frozen		Grocery		Milk	
Channel	Hotel	Retail	Hotel	Retail	Hotel	Retail	Hotel	Retail	Hotel	Retail	Hotel	Retail
Region												
Lisbon	1197.152542	1871.944444	950.525424	8225.277778	12902.254237	5200.000000	3127.322034	2584.111111	4026.135593	18471.944444	3870.203390	10784.000000
Oporto	1105.892857	1239.000000	482.714286	8410.263158	11650.535714	7289.789474	5745.035714	1540.578947	4395.500000	16326.315789	2304.250000	9190.789474
Other	1518.284360	1826.209524	786.682464	6899.238095	13878.052133	9831.504762	3656.900474	1513.200000	3886.734597	15953.809524	3486.981043	10981.009524

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).

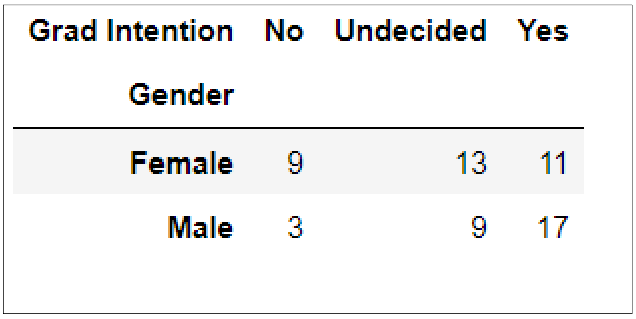
2.1. For this data, construct the following contingency tables (Keep Gender as row variable)

2.1.1. Gender and Major



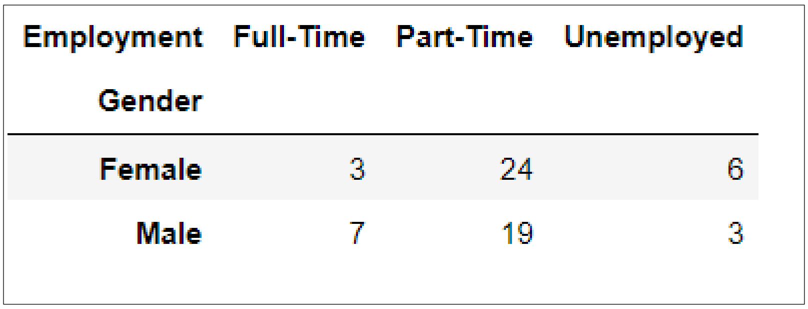
Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Gender								
Female	3	3	7	4	4	3	9	0
Male	4	1	4	2	6	4	5	3

2.1.2. Gender and Grad Intention



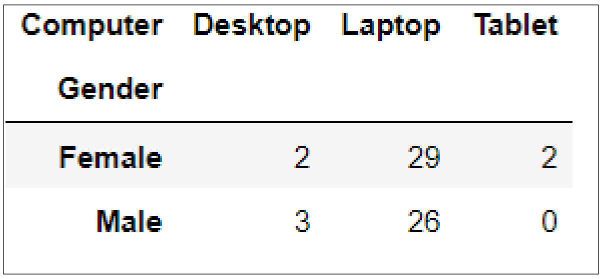
Grad Intention	No	Undecided	Yes
Gender			
Female	9	13	11
Male	3	9	17

2.1.3. Gender and Employment



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

2.1.4. 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?

Probability that a randomly selected CMSU student will be a male is **0.47**.

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

Probability that a randomly selected CMSU student will be a female is **0.53**.

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

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

Conditional probability of randomly selected CMSU student is Majors in accounting given that student is m ale is **0.14**.

Conditional probability of randomly selected CMSU student is Majors in CIS given that student is male is **0.03**.

Conditional probability of randomly selected CMSU student is Majors in Economics/Finance given that student is male is **0.14**.

Conditional probability of randomly selected CMSU student is Majors in International Business given that student is male is **0.07**.

Conditional probability of randomly selected CMSU student is Majors in Management given that student is male is **0.21**.

Conditional probability of randomly selected CMSU student is Majors in Others given that student is male is **0.14**.

Conditional probability of randomly selected CMSU student is Majors in Retailing/Marketing given that student is male is **0.17**.

Conditional probability of randomly selected CMSU student is Majors Undecided given that student is male is **0.1**.

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

Conditional probability of randomly selected CMSU student is Majors in accounting given that student is female is **0.09**

Conditional probability of randomly selected CMSU student is Majors in CIS given that student is female is **0.09**.

Conditional probability of randomly selected CMSU student is Majors in Economics/Finance given that student is female is **0.21**.

Conditional probability of randomly selected CMSU student is Majors in International Business given that student is female is **0.12**.

Conditional probability of randomly selected CMSU student is Majors in Management given that student is female is **0.12.**

Conditional probability of randomly selected CMSU student is Majors in Others given that student is female is **0.09**.

Conditional probability of randomly selected CMSU student is Majors in Retailing/Marketing given that student is female is **0.27**.

Conditional probability of randomly selected CMSU student is Majors Undecided given that student is female is **0**.

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.

Probability that a randomly chosen student is a male and intends to graduate is (0.27)

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

Probability that a randomly selected student is a female and does NOT have a laptop is 0.06)

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

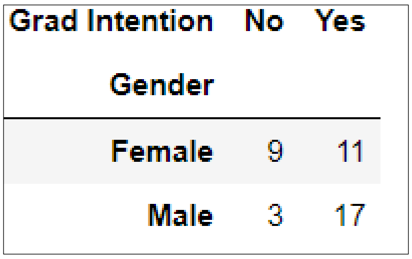
2.5.1. Find the probability that a randomly chosen student is a male or has full-time employment?

Probability that a randomly chosen student is a male or has full-time employment is **0.52)**

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

The conditional probability that given a female student is randomly chosen, she is majoring in international business or management is **0.24**.

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?



Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

The graduate intention and being female are not independent events based on chi2 hypothesis test.

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?

If a student is chosen randomly, the probability that his/her GPA is less than 3 is **0.27**.

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.

The conditional probability that a randomly selected male earns 50 or more is **0.48**.

The conditional probability that a randomly selected female earns 50 or more is **0.55**.

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.

GPA follows approximately Normal Distribution.

Salary, Spending, and text Messages do not follow Normal Distribution.

By observing the box plots of Salary, Spending, and text Messages, we can conclude.

Distribution of Salary is Left skewed.

Distribution of spending and distribution of Text Messages are right skewed.

And it is observed that outliers are present in Salary, Spending and Text 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 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.

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.

For Sample A;

H_0: mean moisture content in sample A = 0.35

H_1: mean moisture content in sample A < 0.35.

Level of significance (α) = 0.05

t test is adopted for verifying hypothesis.

t stat = -1.47

value 0f p = 0.07

Conclusion: As the value of p is greater than 0.05, we fail to reject null hypothesis. Hence, there is no strong evidence to conclude mean moisture content in type “A” shingles is within the permissible limit.

For Sample B;

H_0: mean moisture content in sample B = 0.35

H_1: mean moisture content in sample B < 0.35.

Level of significance (α) = 0.05

t test is adopted for verifying hypothesis.

t statistic = -3.1

value of p = 0.002

Conclusion: As the value of p is less than 0.05, we reject null hypothesis. Hence, mean moisture content in type “B” shingles is within the permissible limit.

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?

Test adopted is two sample t test. Before performing the test, we need to check variance of Sample A and Sample B.

It is found that variance of Sample A and variance of Sample B are equal and is **0.018**.

H_0: mean moisture content in sample A equal to mean moisture content in sample B

H_1: mean moisture content in sample A is not equal to mean moisture content in sample B

Level of significance (α) = 0.05

t statistic = 1.29

value 0f p = 0.2

Conclusion: As the value of p is greater than 0.05, we fail to reject null hypothesis. Hence, there is no strong evidence to state mean moisture content in sample A is not equal to mean moisture content in sample B.