Wholesale Customers Analysis

Analysis on undergraduate students that attend CMSU

Analysis of ABC asphalt shingles data

2021

SMDM PROJECT

Debadutta Mishra

PGP-DSBA Online

5/28/2021

Contents

Problem 1: **WHOLESALE CUSTOMERS ANALYSIS**

Executive Summary

A wholesale distributor operating in different regions of Portugal has information on annual spending of several items in their stores across different regions and channels. 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). We will analyze the data and try to make meaningful inferences.

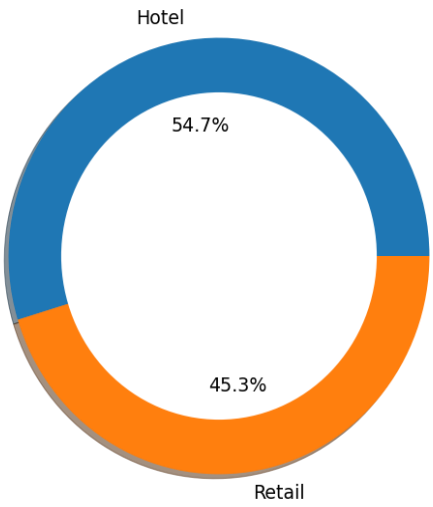
1.1) Use methods of descriptive statistics to summarize data.

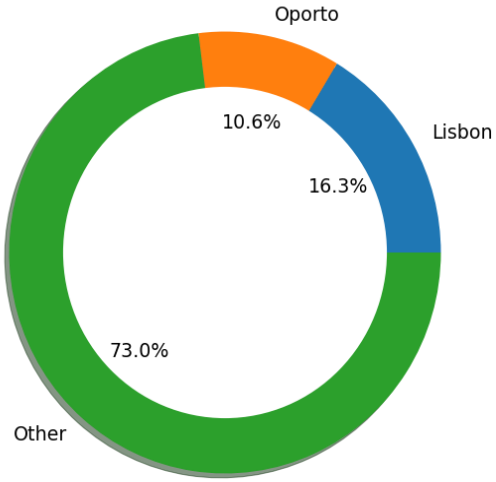
The below table describes the data present in the dataset:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Buyer/Spender	440.00	NaN	NaN	NaN	220.50	127.16	1.00	110.75	220.50	330.25	440.00
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	440.00	NaN	NaN	NaN	12000.30	12647.33	3.00	3127.75	8504.00	16933.75	112151.00
Milk	440.00	NaN	NaN	NaN	5796.27	7380.38	55.00	1533.00	3627.00	7190.25	73498.00
Grocery	440.00	NaN	NaN	NaN	7951.28	9503.16	3.00	2153.00	4755.50	10655.75	92780.00
Frozen	440.00	NaN	NaN	NaN	3071.93	4854.67	25.00	742.25	1526.00	3554.25	60869.00
Detergents_Paper	440.00	NaN	NaN	NaN	2881.49	4767.85	3.00	256.75	816.50	3922.00	40827.00
Delicatessen	440.00	NaN	NaN	NaN	1524.87	2820.11	3.00	408.25	965.50	1820.25	47943.00
Total	440.00	NaN	NaN	NaN	33226.14	26356.30	904.00	17448.75	27492.00	41307.50	199891.00

Above we can clearly see the descriptive statistics like mean, standard deviation, count etc. of the data wherever applicable. We observe that since the difference between the minimum and 25% values, maximum and 75% values of fields like Fresh, Milk, Grocery, Frozen, Detergents_Paper and Delicatessen is too big, we can infer that all these fields will have outliers.

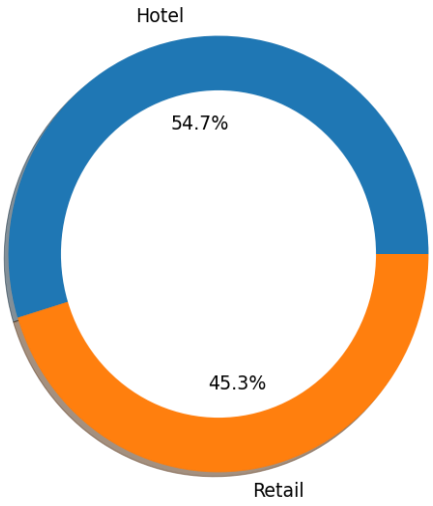
1.1.2) Which Region and which Channel spent the most?

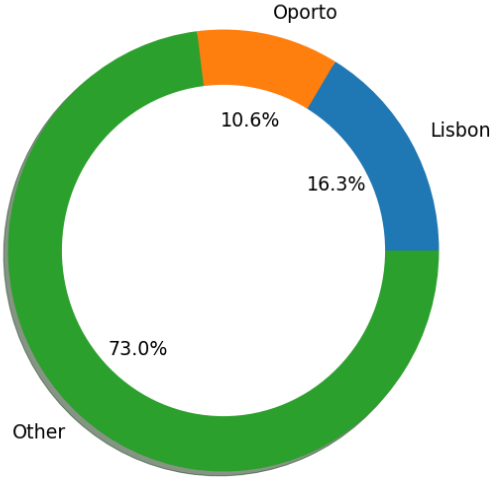




Above we can clearly see that Hotel channel and Other region spend the most

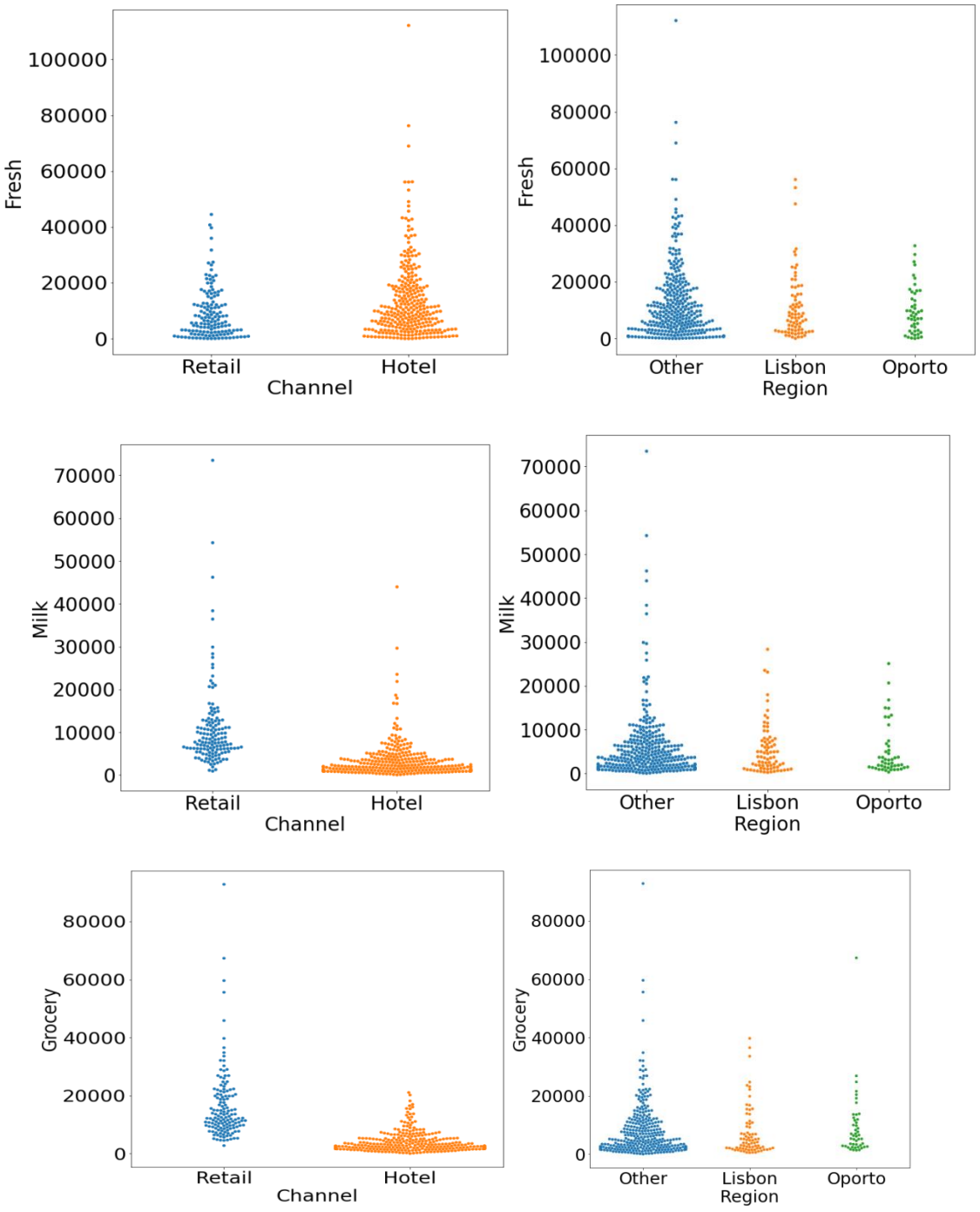
1.1.3) Which Region and which Channel spent the least?

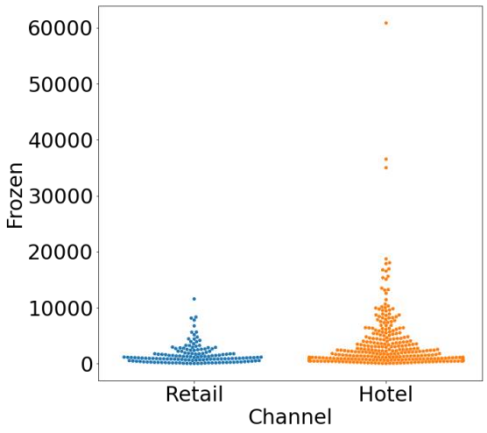


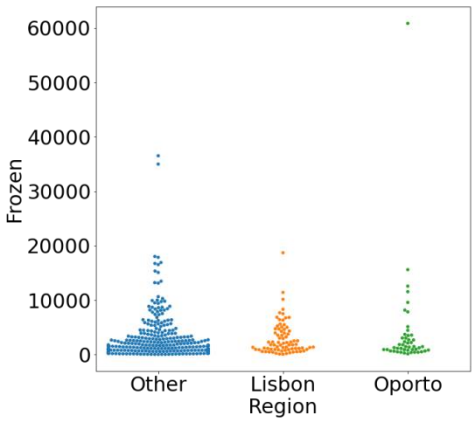


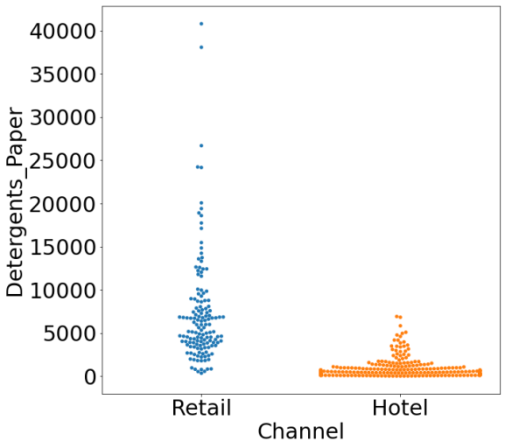
Above we can clearly see that Retail channel and Oporto region spend the least

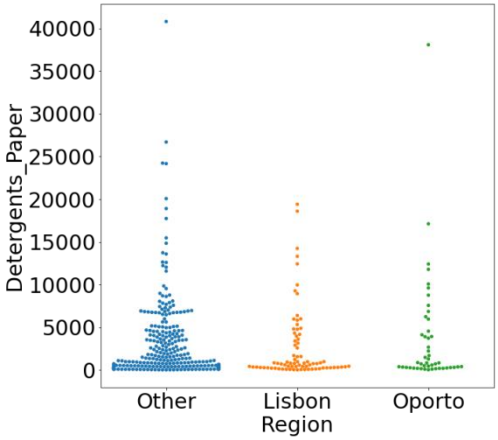
1.2) There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.

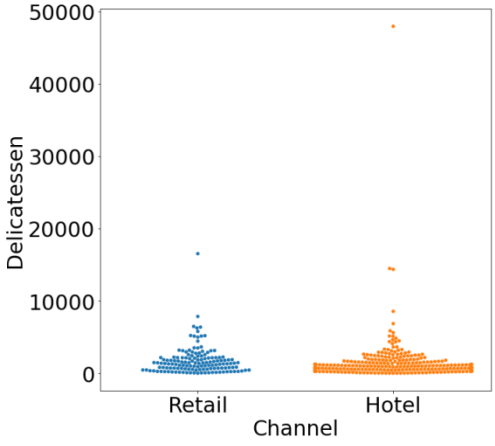


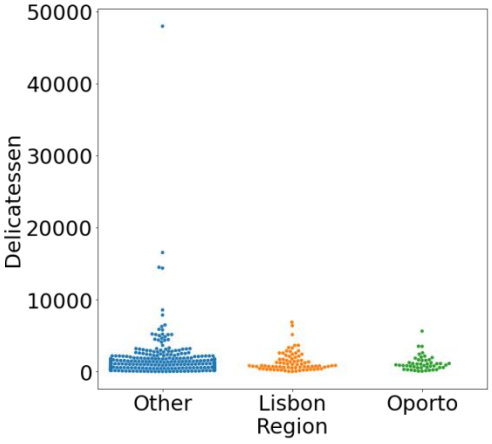

Frozen
60000
50000
40000
30000
20000
10000
0
Retail
Hotel
Channel


Frozen
60000
50000
40000
30000
20000
10000
0
Other
Lisbon
Oporto
Region


Detergents_Paper
40000
35000
30000
25000
20000
15000
10000
5000
0
Retail
Hotel
Channel


Detergents_Paper
40000
35000
30000
25000
20000
15000
10000
5000
0
Other
Lisbon
Oporto
Region


Delicatessen
50000
40000
30000
20000
10000
0
Retail
Hotel
Channel


Delicatessen
50000
40000
30000
20000
10000
0
Other
Lisbon
Oporto
Region

The above are swarm plots plotted for each item against channel and region. In the first item i.e Fresh we can see there is more spend in Hotel channel than Retail channel and most spend in 'Other' region.

In the second item i.e Milk we can see there is more spend in Hotel channel than Retail channel and most spend in 'Other' region. Similarly for all other items we can observe the spread is more for Hotel channel and 'Other' region than for any other. So we can infer the overall spend is more in Hotel and 'Other' region, confirming our inferences from 1.1.2 and 1.1.3.

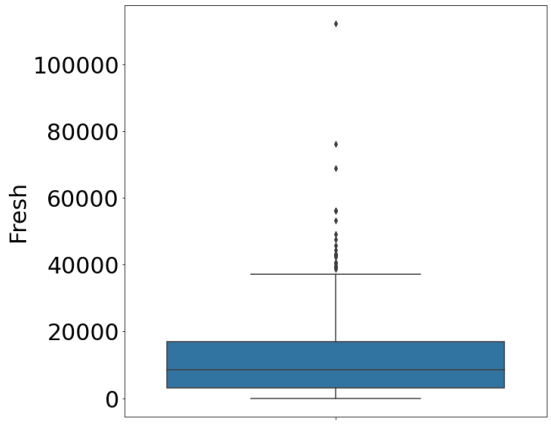
1.3) On the basis of the descriptive measure of variability, which item shows the most inconsistent behavior? Which items shows the least inconsistent behavior?

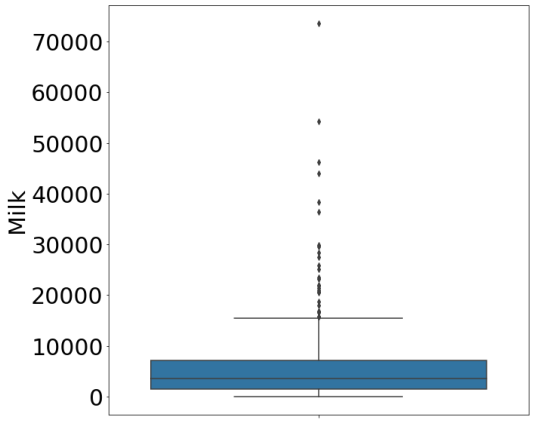
The below values show the coefficient of variance per item **(Level of consistency)**

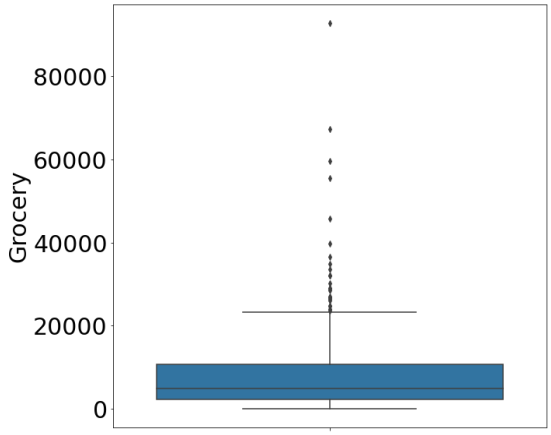
```
Fresh ---- 1.0539179237473149

Milk ---- 1.2732985840065414

Grocery ---- 1.1951743730016824

Frozen ---- 1.5803323836352914

Detergents_Paper ---- 1.6546471385005155

Delicatessen ---- 1.8494068981158382
```

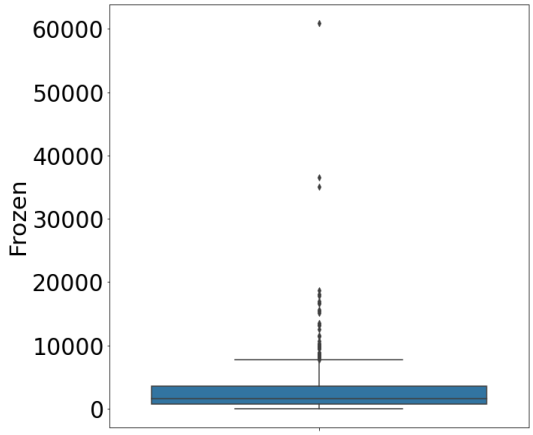
We can clearly see Fresh item is most consistent in terms of amount spent and Delicatessen is the most inconsistent

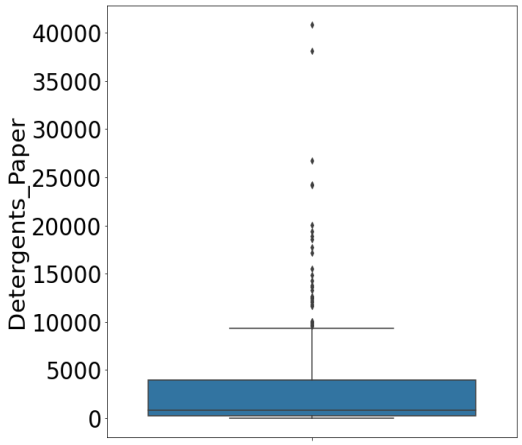
1.4) Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.

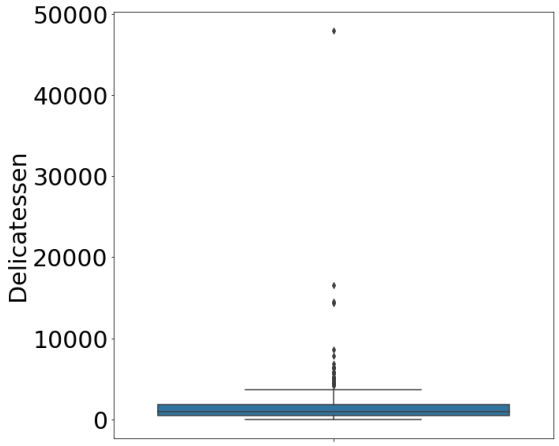












From the above box plots of all the items, we can clearly infer that all items have outliers i.e. there are spend values for each item which is greater that the Q3 by 1.5 times of inter-quartile range. This confirms our findings in 1.1).

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 above analysis of wholesale consumers we can clearly see that there is no big difference in the total spend of Hotel and Retail Channel, although hotel spends more than retail. Due to the presence of so many outliers, one can assume that the data might be faulty or the procedures used might be erroneous. It is difficult to put accurate statistical conclusions with so many outliers. It is important that we try to understand the nature of these outliers and decide whether to remove them or not.

Problem 2: **ANALYSIS ON UNDERGRADUATE STUDENTS THAT ATTEND CMSU**

Executive Summary

The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. CMSU creates and distributes a survey of 14 questions and receives responses from 62 undergraduates (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 Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
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 Gender	No	Undecided	Yes
Female	9	13	11
Male	3	9	17

2.1.3) Gender and Employment

Employment Gender	Full-Time	Part-Time	Unemployed
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 a 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 of a randomly selected student to be male is 46.77%

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

Probability of a randomly selected student to be female is 53.22%

2.3) Assume that the sample is a 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.

```
The Probability that a student is pursuing Management given that he is

male is  20.68965517241379 %


The Probability that a student is pursuing Retailing/Marketing given

that he is male is  17.241379310344826 %
```

The Probability that a student is pursuing Accounting given that he is male is 13.793103448275861 %

The Probability that a student is pursuing Economics/Finance given that he is male is 13.793103448275861 %

The Probability that a student is pursuing Other courses given that he is male is 13.793103448275861 %

The Probability that a student is still Undecided on the Maor given that he is male is 10.344827586206895 %

The Probability that a student is pursuing International Business given that he is male is 6.896551724137931 %

The Probability that a student is pursuing CIS given that he is male is 3.4482758620689653 %

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

The Probability that a student is pursuing Retailing/Marketing given that she is female is 27.272727272727277 %

The Probability that a student is pursuing Economics/Finance given that

she is female is 21.21212121212121 %

The Probability that a student is pursuing International Business given that she is female is 12.121212121212121 %

The Probability that a student is pursuing Management given that she is female is 12.121212121212121 %

The Probability that a student is pursuing Accounting given that she is female is 9.090909090909092 %

The Probability that a student is pursuing CIS given that she is female is 9.090909090909092 %

The Probability that a student is pursuing Other courses given that she is female is 9.090909090909092 %

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.

The probability that a randomly chosen student is a male and intends to graduate is 58.620689655172406 %

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

The probability that a randomly selected student is a female and does NOT have a laptop is 6.0606060606060606 %

2.5) Assume that the sample is a 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 a full-time employment

The probability that a randomly chosen student is a male or has full-time employment is 51.61290322580645 %

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 24.242424242424242 %

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 graduate intention and being female are independent events?

Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

The probability that the graduate intention is yes given it is a female student is 17.741935483870968 %,

Which is not equal to p(female)*p(graduate intention yes), i.e. 24.037460978147763.

But it is equal to p(graduate intention=yes)*p(Female/graduate intention= yes) i.e. 17.741935483870964 %.

Hence 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. 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?

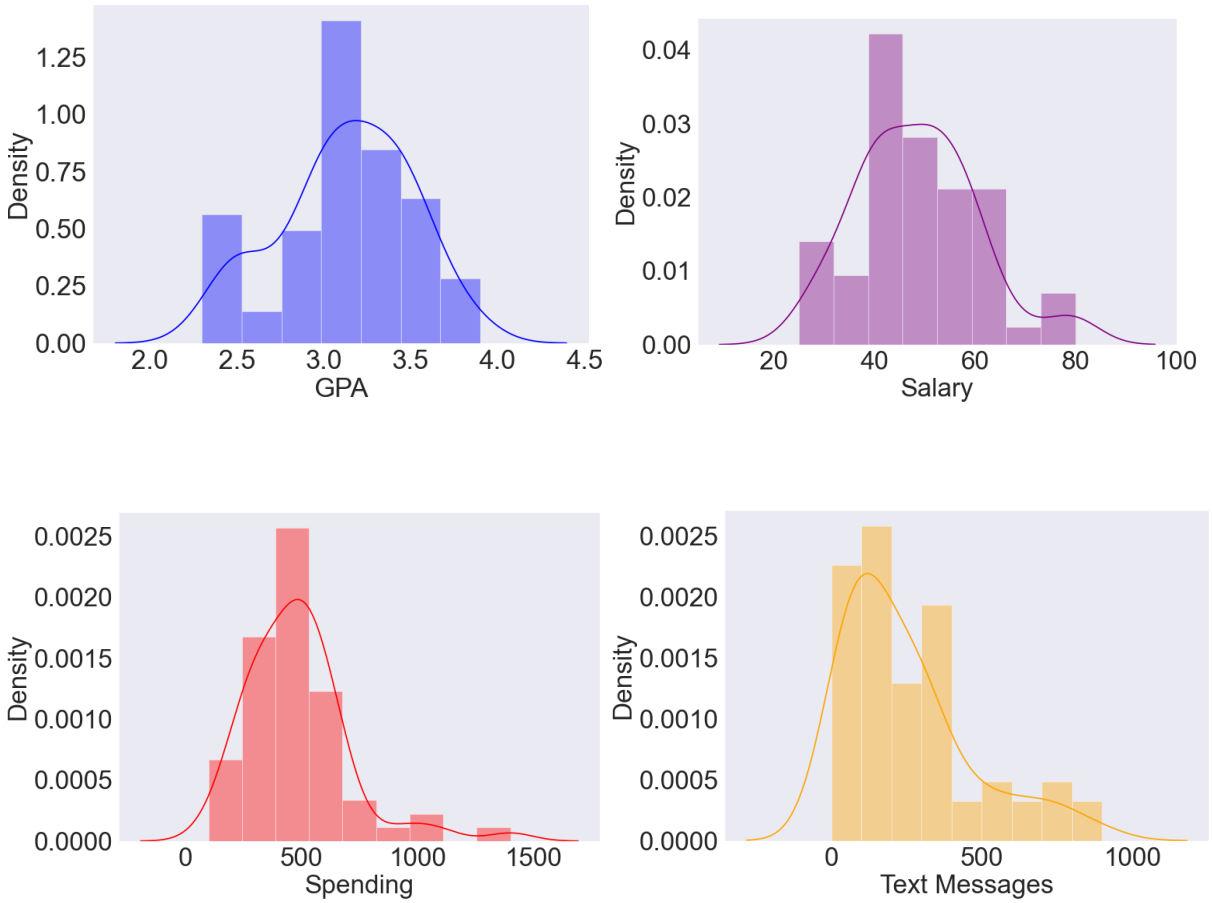
If a student is chosen randomly, the probability that his/her GPA is less than 3 is 27.419354838709676 %

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

The conditional probability that a randomly selected male earns 50 or more is 48.275862068965516 %.

The conditional probability that a randomly selected female earns 50 or more is 54.54545454545454 %.

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.



We can infer from the above distplots that GPA and Salary have almost a normal distribution. However Spending and Text Messages are slightly right skewed.

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

Upon doing a detailed analysis on **the** data about the undergraduate students that attend CMSU, we observed many interesting insights. E.g. the graduate intention and being female are not independent events. We can see most of the students use a laptop. We can also observe that female students have more affinity towards choosing 'Retailing/Marketing' as a major, whereas male students have more affinity towards choosing 'Management' as a major.

Problem 3: **ANALYSIS OF ABC ASPHALT SHINGLES DATA**

3.1) Do you think there is evidence that mean moisture contents in both types of shingles are within the permissible limits? State your conclusions clearly showing all steps.

We can state the null hypothesis and alternate hypothesis as below:

H0—Mean moisture content in the particular type of shingles is within the permissible limits

Ha— Mean moisture content in the particular type of shingles is not within the permissible limits

Conducting a single sample ttest individually on each type os shingles A and B, we draw the below inferences

We reject H0 for B type shingles as pvalue/2 < 0.05, i.e. there is no evidence that means moisture contents in B type shingles are within the permissible limits.

We accept H0 for A type shingles as pvalue/2 > 0.05, i.e. there is evidence that means moisture contents in A type shingles are within the permissible limits.

3.2) Do you think that the population means for shingles A and B are equal? Form the hypothesis and conduct the test of the hypothesis. What assumption do you need to check before the test for equality of means is performed?

We can state the null hypothesis and alternate hypothesis as below:

H0— Population mean of both A and B Shingles are same

Ha— Population mean of both A and B Shingles are not same

Conducting a double sample independent ttest on both types of shingles, we draw the below inferences

Since pvalue/2 > 0.05 we accept H0, i.e. population mean of both A and B Shingles are same.