SMDM PROJECT SAMPLE REPORT

DSBA

greatlearning

Learning for Life

Disclaimer: This document will act just as a reference for the learners on how the format of business report for submission is expected. The questions in this sample report might differ from the actual questions in the project.

Contents

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 (Wholesale Customer.csv) 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?

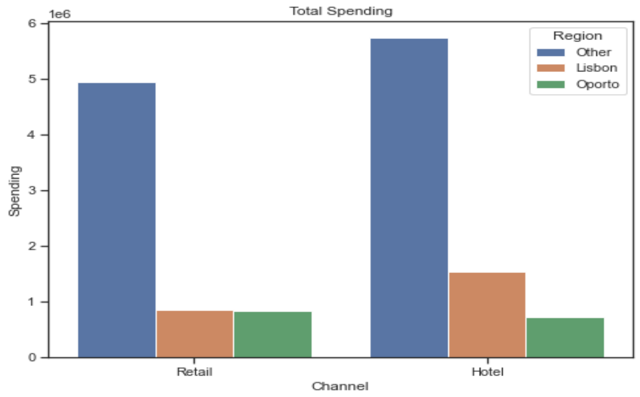
Which Region and which Channel spent the least?

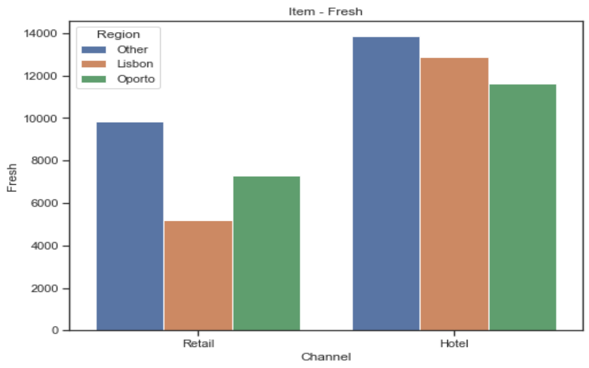
Using describe function in python we first looked at the basic descriptive statistics of the data set. Using bar graph with Region and Channel we were able to identify region with maximum spend and minimum spend. Below is the bar graph representation-Looking at the bar graph, Hotel Channel spends more and Retail spends least.

count	unique	top	freq	mean	std	min	25%	50%	75%	max	
Buyer/Spender	440.0	NaN	NaN	NaN	220.5	127.161315	1.0	110.75	220.5	330.25	440.0
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

- We can find the which region and channel spend most we need to find the first total spending to all the 6 items for all the retailers. For than we already plot the graph below





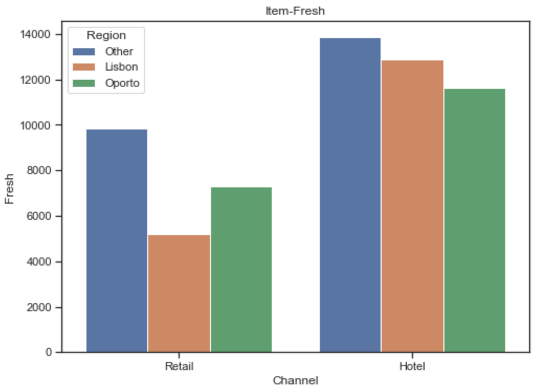
- The Region that spend the most is other and that the spent the least is Oporto.
- The channel that spent the most is Hotel and that spent the least is Retail.

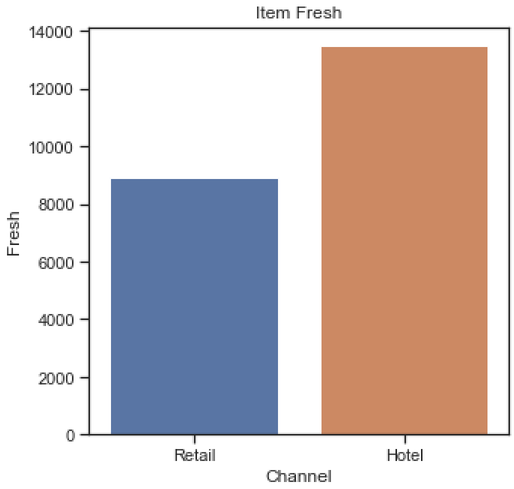
As per below details mentioned we can find the output

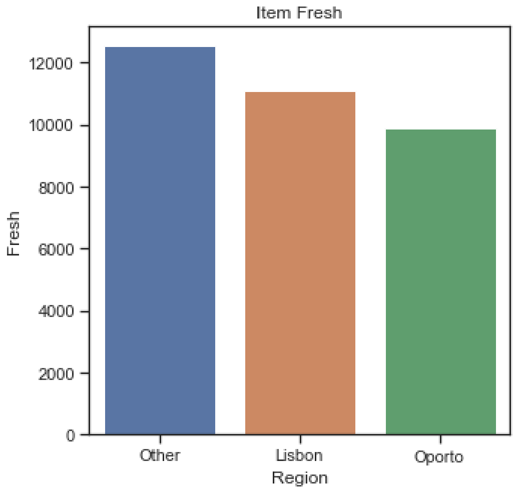
```
Region
Lisbon     2386813
Oporto     1555088
Other     10677599
Name: Spending.
Channel
Hotel     7999569
Retail    6619931
Name: Spending.
```

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.

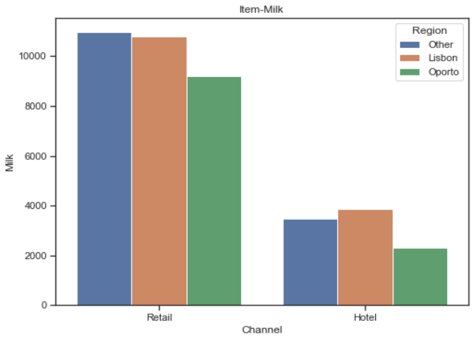
- Using bar plot we can see the behavior behaviour in all the channel and region. And using bar plot to find the out varieties across region and channel.

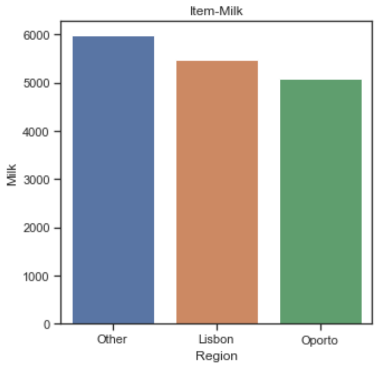


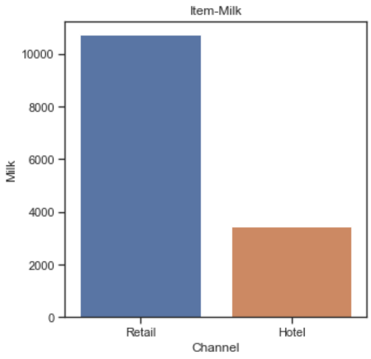




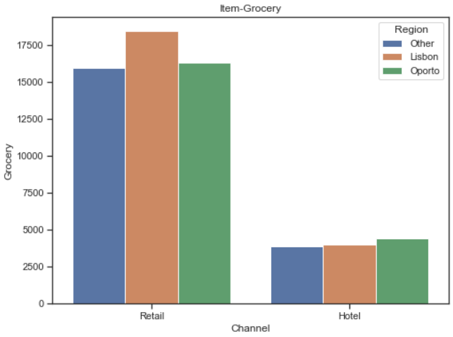
- Based on the plot, Fresh item is sold more in the Hotel channel.

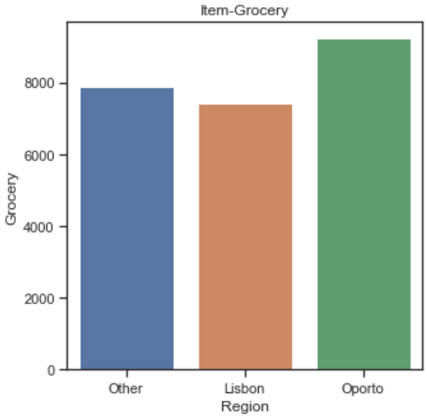


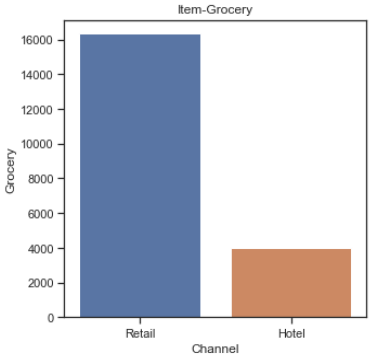




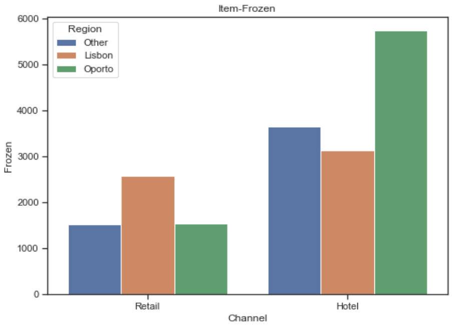
- Based on the plot, Milk item is sold more in the Retail channel

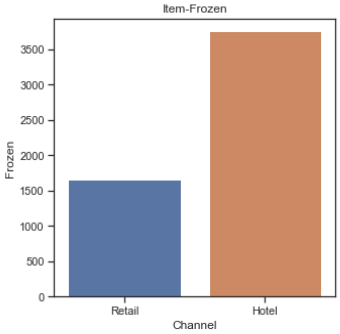


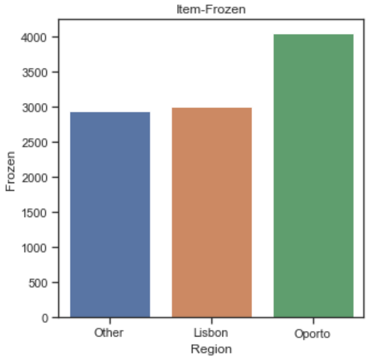




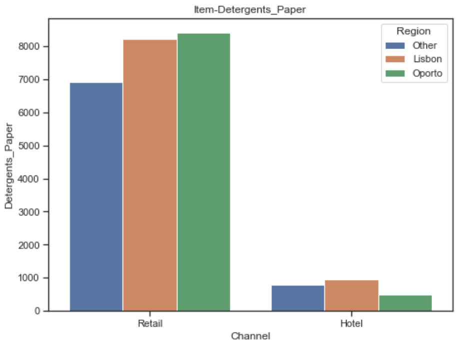
- Based on the plot, Grocery item is sold more in the Retail channel

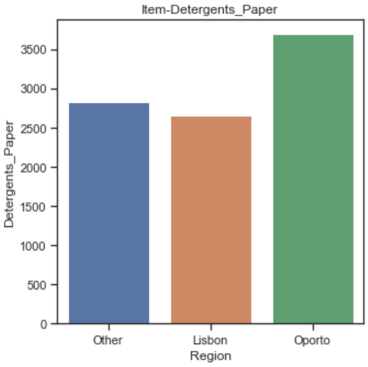


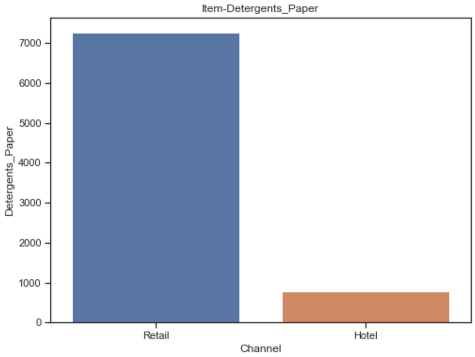




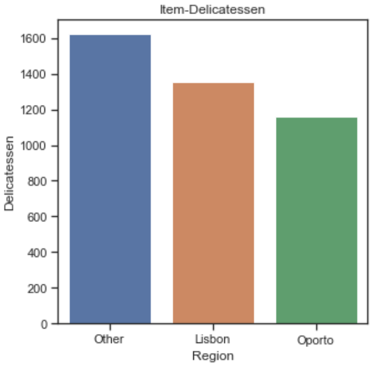
- Based on the plot, Frozen item is sold more in the Hotel channel

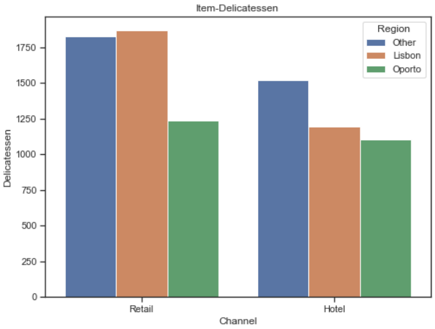


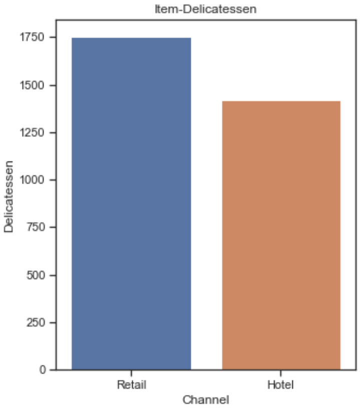




- Based on the plot, **Detergents Paper** item is sold more in the Retail channel







- Based on the plot, **Delicatessen** item is sold more in the Retail channel

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

```
Fresh            12647.0
Milk             7380.0
Grocery            9503.0
Frozen            4855.0
Detergents_Paper     4768.0
Delicatessen        2820.0
```

Below is the output from Python –

- Coefficient of Variation for Fresh is 1.0527196084948245

- Coefficient of Variation for Milk is 1.2718508307424503

- Coefficient of Variation for Frozen is 1.193815447749267

- Coefficient of Variation for Grocery is 1.5785355298607762

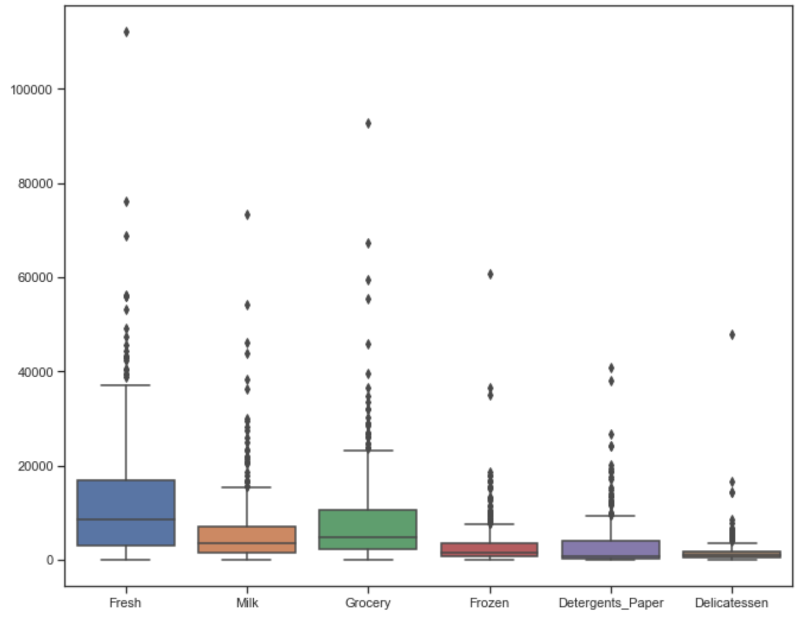
- Coefficient of Variation for Detergents Paper is 1.6527657881041729

- Coefficient of Variation for Delicatessen is 1.8473041039189306

- Fresh item has highest Standard deviation So that is Inconsistent.
- Delicatessen item have smallest Standard deviation, So that is consistent.

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

As we can see below box plot is showing that across all the product having outliers such as (Fresh, Milk, Grocery , Frozen, Detergents _Paper, Delicatessen)



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.

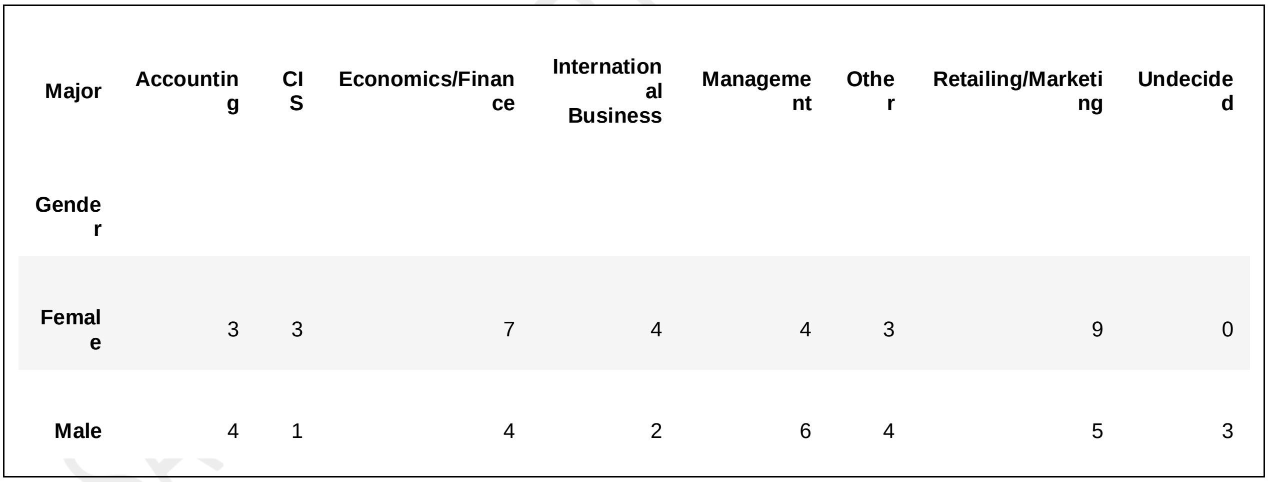
As per the analysis, I find out that there are inconsistencies in spending of different items (by calculating Coefficient of Variation), which should be minimized. The spending of Hotel and Retail channel are different which should be more or less equal. And also spent should equal for different regions. Need to focus on other items also than "Fresh" and "Grocery"

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.csv file).

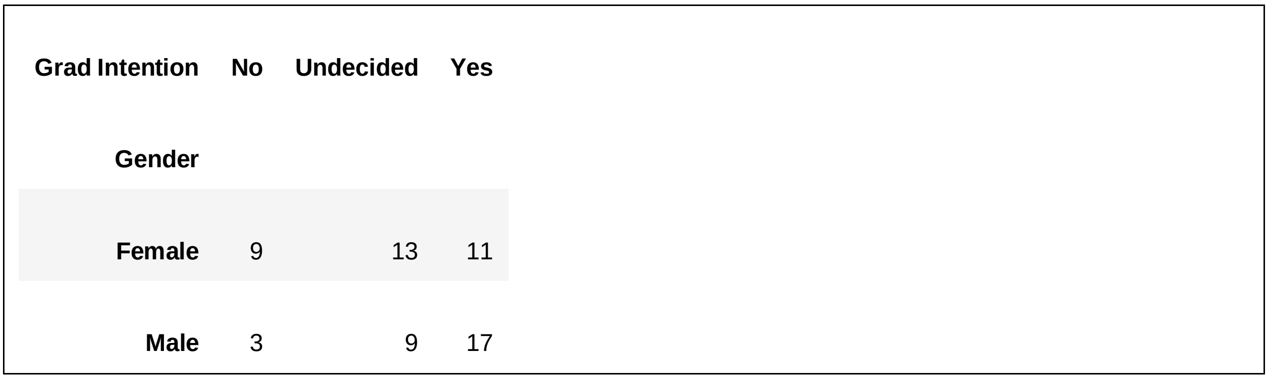
2.1. For this data, construct the following contingency tables (Keep Gender as rowvariable)

2.1.1. Gender and Major



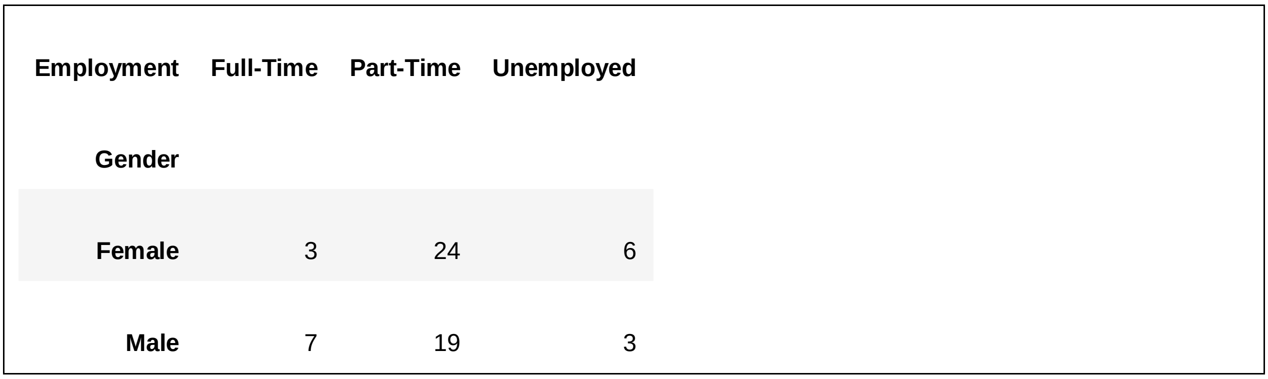
Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Gender								
Female	3	3	7	4	4	3	9	0
Male	4	1	4	2	6	4	5	3

2.1.2. Gender and Grad Intention



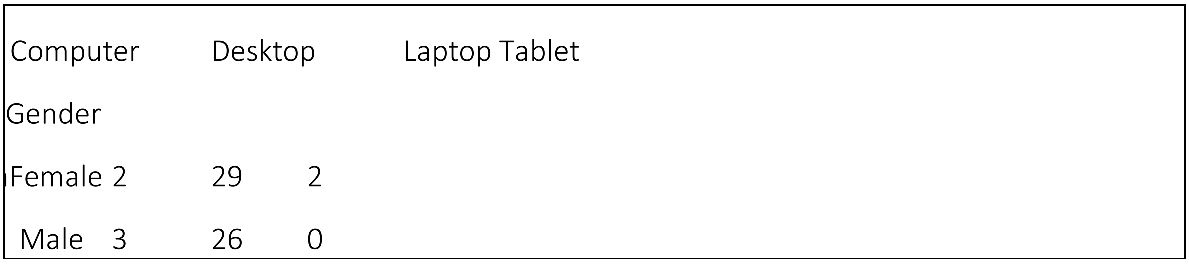
Grad Intention	No	Undecided	Yes
Gender			
Female	9	13	11
Male	3	9	17

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 a representative of the population of CMSU. Based onthe data, answer the following questions:

Calculate the probabilities as asked in python or excel and paste the results here and try to display in mathematical format for each.

Example: P(Head)=1/2=0.5

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

- probability that a randomly selected CMSU student will be male is 0.46774193548387094.
  - Percentage will be 46.77%
- probability that a randomly selected CMSU student will be female is 0.532258064516129
  - percentage will be 53.22%

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

- Among MALE students:
  - Probability of Accounting : 0.13793103448275862
  - Probability of CIS: 0.034482758620689655

- Probability of Economics/Finance : 0.13793103448275862
- Probability of International Business : 0.06896551724137931
- Probability of Management : 0.20689655172413793
- Probability of Other : 0.13793103448275862
- Probability of Retail/Marketing : 0.1724137931034483
- Probability of Undecided : 0.10344827586206896

- Among Female students:
  - Probability of Accounting: 0.09090909090909091
  - Probability of CIS : 0.09090909090909091
  - Probability of Economics/Finance: 0.21212121212121213
  - Probability of International Business: 0.12121212121212122
  - Probability of Management : 0.12121212121212122
  - Probability of Other : 0.09090909090909091
  - Probability of Retail/Marketing : 0.2727272727272727
  - Probability of Undecided : 0.0

2.2.3. Find the conditional probability of intent to graduate, given that the student is a male. Find the conditional probability of intent to graduate, given that the student is a female.

- Among MALE Students:
  - Probability of No as Grad Intention: 0.10344827586206896
  - Probability of Undecided as Grad Intention: 0.3103448275862069
  - Probability of a Grad Intention: 0.5862068965517241

- Among Female Students:
  - Probability of No as Grad Intention: 0.2727272727272727
  - Probability of Undecided as Grad Intention: 0.3939393939393939
  - Probability of a Grad Intention: 0.3333333333333333

2.2.4. Find the conditional probability of employment status for the male students as well as for the female students.

- Among MALE Students:
  - Probability of Full-Time as Employment status: 0.2413793103448276
  - Probability of Part-Time as Employment status: 0.6551724137931034
  - Probability of Unemployed as Employment status 0.10344827586206896

- Among Female Students:
  - Probability of Full-Time as Employment status: 0.09090909090909091
  - Probability of Part-Time as Employment status: 0.7272727272727273
  - Probability of Unemployed as Employment status: 0.18181818181818182

2.2.5. Find the conditional probability of laptop preference among the male students as well as among the female students.

- Among Female students:
  - Probability of Laptop as Computer: 0.8787878787878788

- Among Male students:
  - Probability of Laptop as Computer: 0.896551724137931

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

Using contingency tables of Gender and Employment we got the total numbers of males and number of males who are full time employed 24.10%

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

Using contingency tables of Gender and Major we got the total numbers of females and who are majoring in international business management 12.12 %

2.4 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	Total
Gender			
Female	9	11	20
Male	3	17	20

The Probability that a randomly selected student the graduate intention and female is
P(Grad Intention Yes) = 28/40 = 0.7
P(Grad Intention Yes | female) = 11 / 20 = 0.55 These probabilities are not equal.
This suggests that the two events are independent

2.5 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

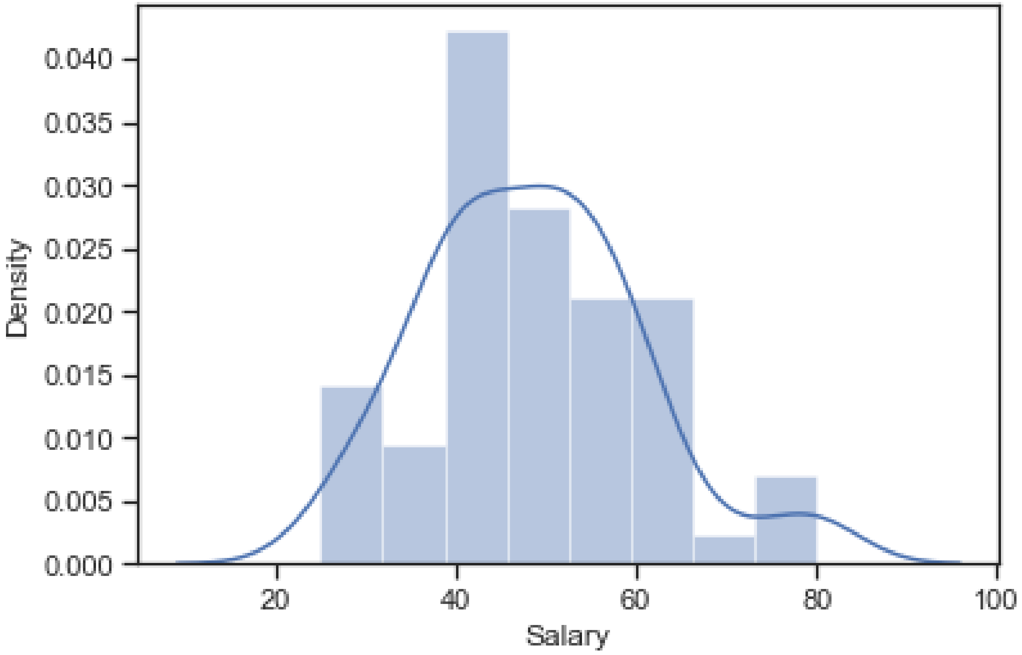
2.5.1 If a student is chosen randomly, what is the probability that his/her GPA is less than 3?

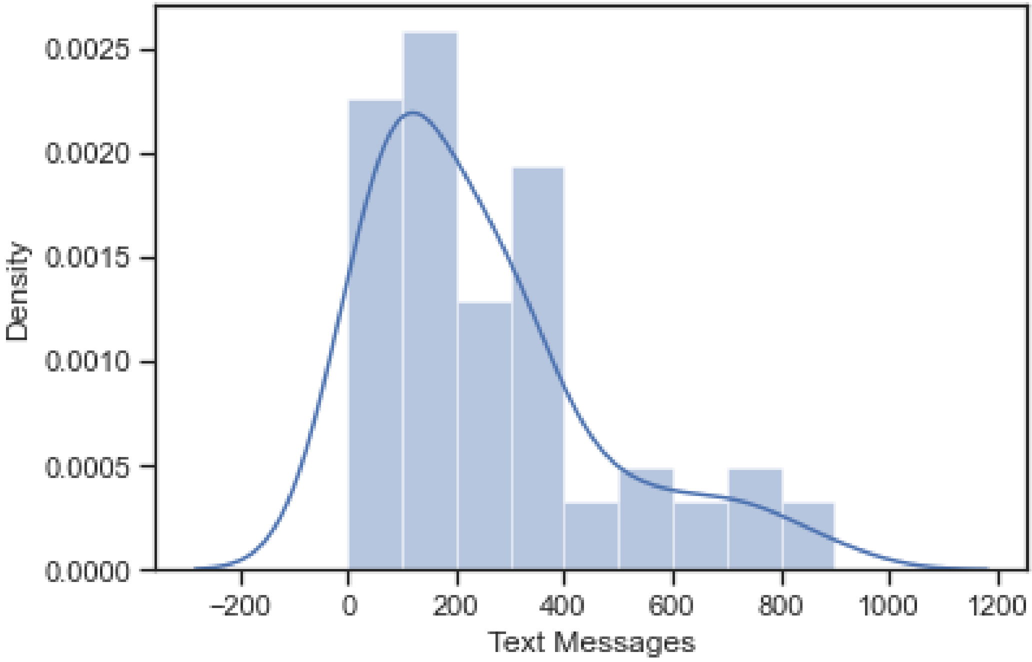
2.5.2 Find conditional probability that a randomly selected male earns 50 or more. Find conditional probability that a randomly selected female earns 50 or more.

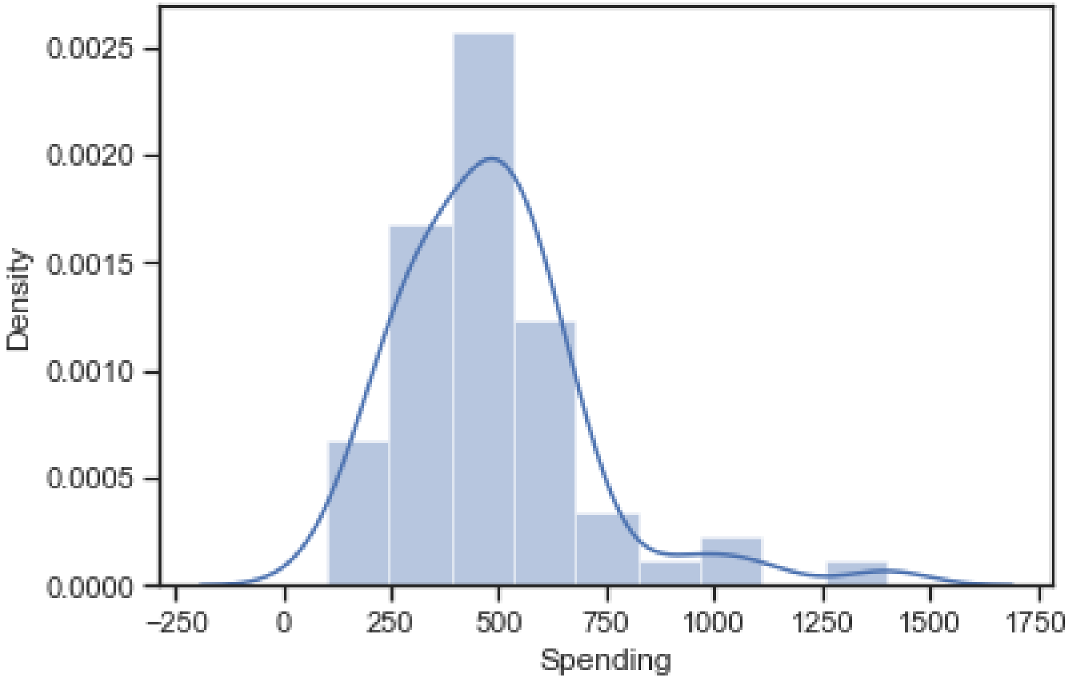
Randomly selected male earns 50 or more is 34.48% And
Probability that randomly selected female earns 50 or more is 30.3 %

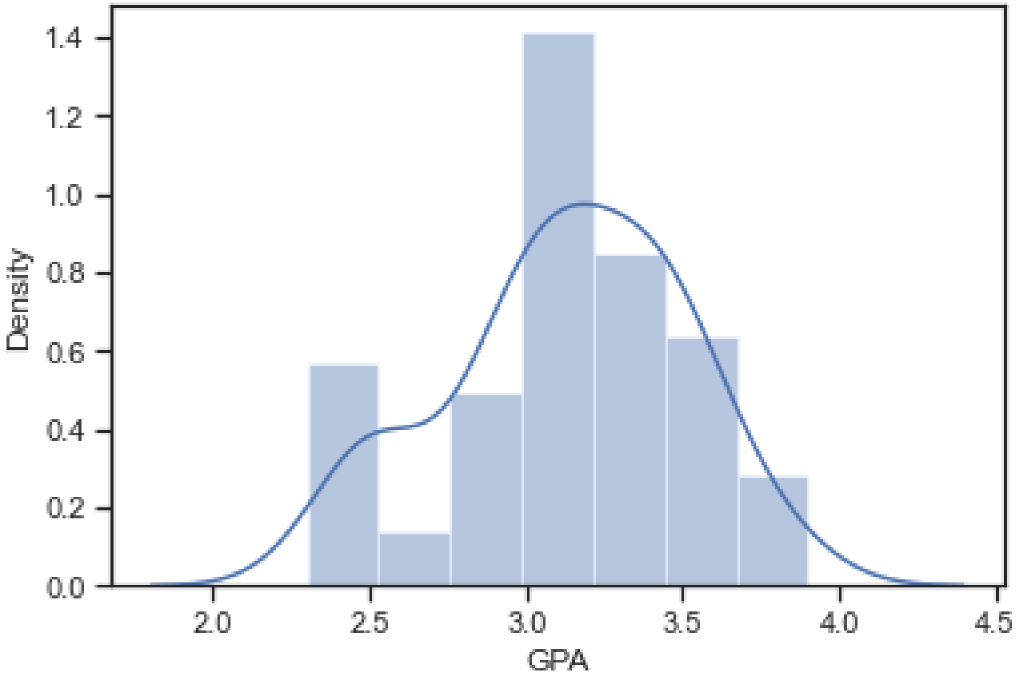
2.6Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending and Text Messages. For each of them comment whether they follow a normal distribution.

Used distplot to know the normal distribution of these four numerical (continuous) variables in the data set –GPA, Salary, Spending and Text Messages


0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
Density
20
40
60
80
100
Salary


0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
Density
−200
0
200
400
600
800
1000
1200
Text Messages





By the distplot which show above we observe that out of the given four data sets ‘GPA’ and ‘Salary’ are following normal distribution,
Other two ‘Spending’ and ‘Text Messages’ are not following the normal distribution.

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 is calculated. The company claims that the mean moisture content cannot be greater 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.

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:

$$H_0 \leq 0.35$$

$$H_1 > 0.35$$

For the B 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:

$$H_0 \leq 0.35$$

$$H_1 > 0.35$$

3.1 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?

Explanation :
Input will be mention below :-

```
t_statistic, p_value = ttest_1samp(data_df.A, 0.35)
```

```
print('1at  sample t test \nt statistic: {0} p value: {1} '.format(t_statistic, p_value/2))
```

Such as output will mentioned below

```
t test  statistic: -1.4735046253382782 p value: 0.07477633144907513
```

3.2 What assumption about the population distribution is needed in order to conduct the hypothesis tests above?

Explanation :
Input –

```
t_statistic,p_value=ttest_ind(data_df['A'],data_df['B'],equal_var=True ,nan_policy='omit')
print("t_statistic={} and pvalue={}".format(round(t_statistic,3),round(p_value,3)))
```

Output -

```
t_statistic=1.29 and pvalue=0.202
```

As the pvalue > α , do not reject H0;
and we can say that population mean for shingles A and B are equal Test Assumptions When running a two-sample t-test,
the basic assumptions are that the distributions of the two populations are normal,