

SMDM Project Report

By-Kunal Tiwari

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

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?

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

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

Problem 2 - (Download [Data](#))

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

2.1.2. Gender and Grad Intention

2.1.3. Gender and Employment

2.1.4. Gender and Computer

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?

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

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.

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

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.

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

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?

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

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?

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?

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.

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.

Problem 3 ([Download Data](#))

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.

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?

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

	Channel	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
0	Hotel	71034	4015717	1028614	1180717	1116979	235587	421955	7999569
1	Retail	25986	1264414	1521743	2317845	234671	1032270	248988	6619931

This clearly shows that the Hotels as a Channel are spending more i.e., 7,999,569 than Retailers which is spending 6,619,931.

This shows that hotels are spending 20.8% more as compared to Retail

```
440 rows x 9 columns

In [23]: regiondf = wholesale_customer_spending.groupby('Region')['Spending'].sum()
print(regiondf)
print()
channeldf = wholesale_customer_spending.groupby('Channel')['Spending'].sum()
print(channeldf)

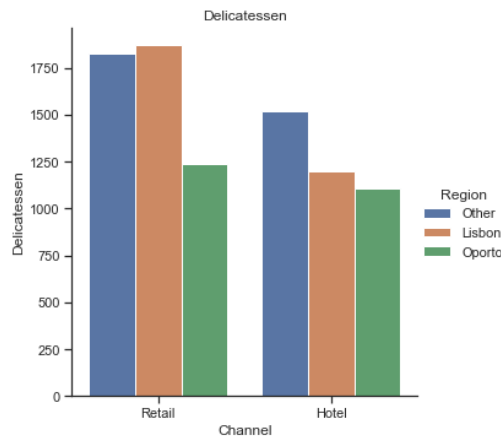
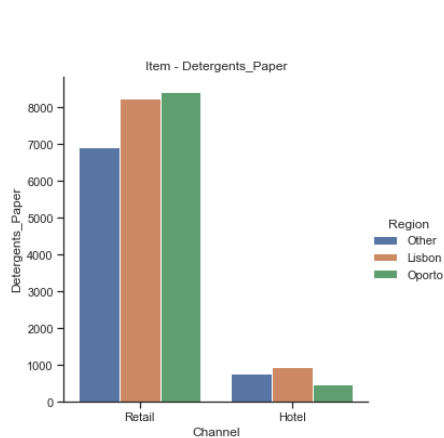
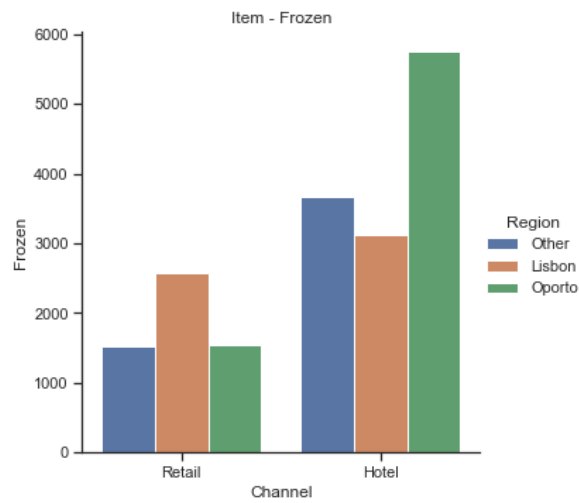
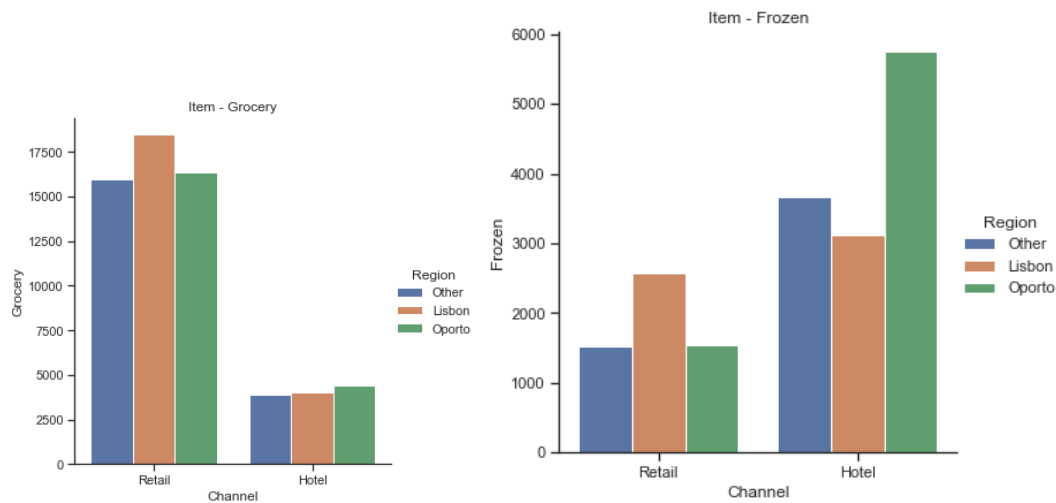
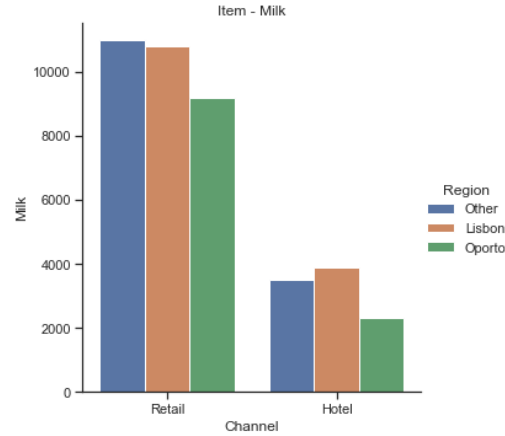
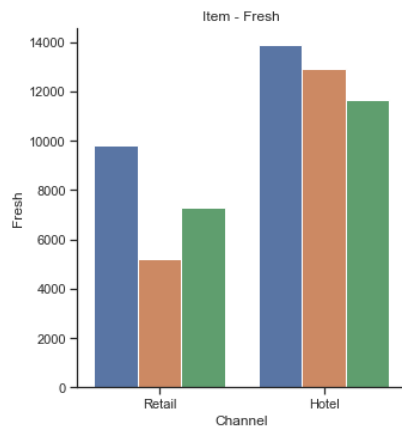
Region
Lisbon      2386813
Oporto      1555088
Other       10677599
Name: Spending, dtype: int64

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

Highest spend in the Region is from Others and lowest spend in the region is from Oporto.

Highest spend in the Channel is from Hotel and lowest spend in the Channel is from Retail.

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.



- The graph clearly shows that the amount spent on the Fresh items is more in Hotel channel as compared to the Retail channel. In Hotel channel spending on the Fresh items is maximum in every region as compared to Retail channel.
- Looking at the above graphs, we see that some categories like Milk, Grocery & Detergents_Paper have spent more in the Retail channel compared to Hotel across all regions. On the other hand, Fresh and Frozen have higher consumption in the Hotel channel compared to Retail across all regions.
- The average annual spending on the Fresh items is more in Lisbon region as compared to Oporto region in Hotel channel and viceversa.
- The average annual spending on Milk items is more in the Retail channel as compared to Hotel. Lisbon seems to be spending more as compared to Oporto region on Milk items in both channels

The average amount spent on the Grocery items is more in Retail

- channel as compared to the Hotel. Also, the Lisbon region is spending more in Grocery items via Retail channel as compared to the Hotel in which Oporto's annual spending is highest across the country.
- The average annual spending on Frozen items in Hotel is more as compared to Retail channel. Oporto region is the major contributor for Hotel whereas the average spending is highest in Retail channel.
- The average annual spending on Detergent Paper is very high across the Retail channel as compared to the Hotel. The Oporto region is spending the most via Retail whereas Hotels dominate the spending in Lisbon region.
- The average annual spending in Delicatessen items is less in hotel as compared to the Retail channel. The Lisbon region on average spends the highest via 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?

Standard Deviation of all items:

Fresh - 12647.33

Milk - 7380.38

Grocery - 9503.16

Frozen - 4854.67

Detergents_Paper - 4767.85

Delicatessen - 2820.11

#Coefficient of Variation

Fresh - 0.527196084948245

Milk - 1.271

Grocery - 1.193815

Frozen - 1.5785

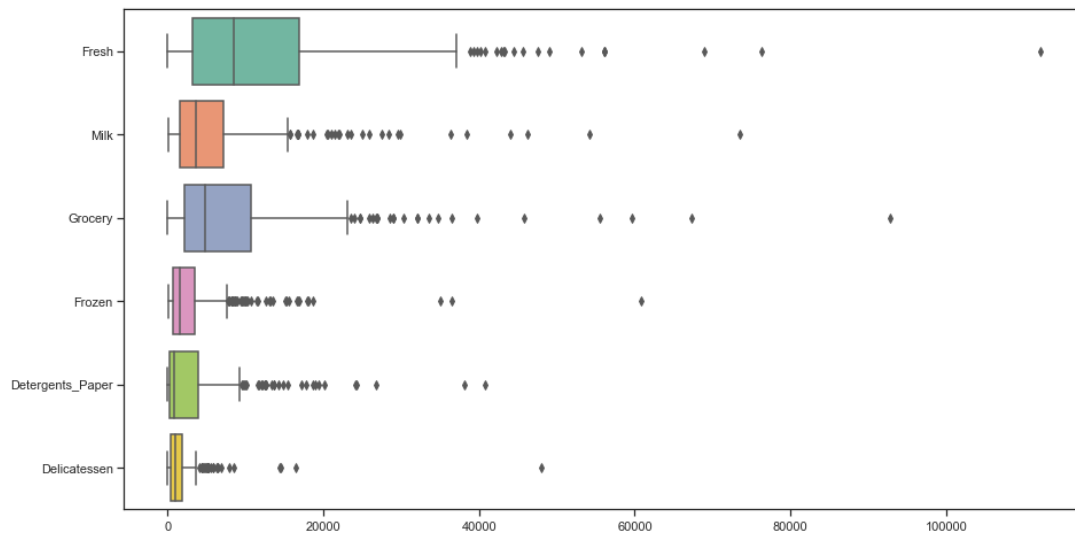
Detergents Paper - 1.6527

Delicatessen - 1.847

From this result we see that Delicatessen shows the most inconsistent behavior and Fresh shows the least inconsistent behavior.

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

The behavioral characteristics of the given data indicates that there are many outliers in the data. We can also confirm by doing a Boxplot for Region and Channel with respect to Total Amount spent for Items.



- Yes there are outliers in all the items across the product range(Fresh,Milk,Grocery,Detergents_Paper and Delicatessen). Outliers are detected but not necessarily removed, it depends of the situation. Here I assumed that the wholesome distributor provided us a dataset with corrected data, so I will keep them as is.

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

1. It can be noticed that the overall sales in Hotels is much more than the sales in Retail. The distributor may consider Retail channel as a target area for further expansion on growth.
2. Spend in Hotel needs to be increased in Milk, Grocery and Detergents_Paper.
3. Spend in Retail needs to be increased in Fresh, Frozen and Delicatessen.
4. The spending should be done carefully as Grocery items are also very inconsistent.
5. The annual spending in both channels by all the regions should be managed carefully especially in case of Fresh items because Fresh items have the highest standard deviation and are least inconsistent. So, the spending on this item should be done carefully.
6. The data is not normally distributed due to the presence of many outliers. This indicates that a large no of sales can be attributed to some specific buyers.
7. They should focus on increasing the Total in Lisbon, Oporto regions and Retail Channel to balance the reduce risk while increasing business.

Problem 2 - (Download [Data](#))

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)

Gender and Major:-

The below table denotes the variation between 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

The below table denotes the variation between Gender and Grad Intention.

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

2.1.3. Gender and Employment

The below table denotes the variation between Gender and Employment.

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

2.1.4. Gender and Computer

The below table denotes the variation between 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?

Total Students = 62

Total Male Students=29

Probability that a randomly selected student will be Male = $29/62$

$P(\text{Male}) = 0.4677$ or 46%

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

Total number of students = 62

Number of Female students = 33

Probability that a randomly selected student will be Female = $33/62$

$P(\text{Female}) = 0.532$ or 53.2%

Question 2.3: - Assume that the sample is representative of

the population of CMSU. Based on the data, answer the following question:

2.3 A) Find the conditional probability of different majors among the male students in CMSU.

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 students = 62

Number of males = 29

Below is the probability OF MALE CANDIDATES

Among Male candidates:

Probability of Males opting for Accounting: 0.13793103448275862

Probability of Males opting for CIS: 0.034482758620689655

Probability of Males opting for Economics/Finance 0.13793103448275862

Probability of Males opting for International Business: 0.06896551724137931

Probability of Males opting for Management: 0.20689655172413793

Probability of Males opting for Other : 0.13793103448275862

Probability of Males opting for Retailing/Marketing : 0.1724137931034483

Probability of Males opting for Undecided: 0.10344827586206896

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

Total number of students = 62

Number of females = 33

Below is the probability OF FE MALE CANDIDATES

Among Female candidates:

Probability of Female opting for Accounting: 0.09090909090909091

Probability of Female opting for CIS: 0.09090909090909091

Probability of Female opting for Economics/Finance : 0.21212121212121213

Probability of Female opting for International Business: 0.12121212121212122

Probability of Female opting for Management: 0.12121212121212122

Probability of Female opting for Other : 0.09090909090909091

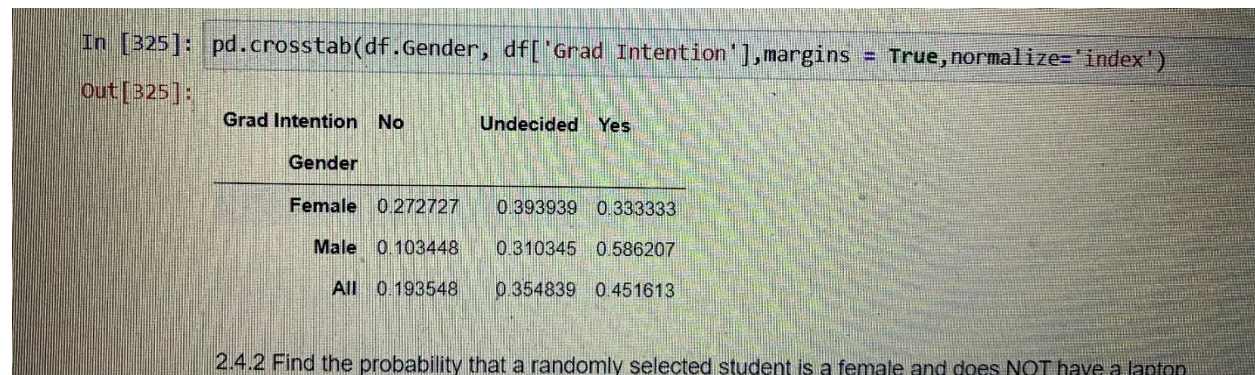
Probability of Female opting for Retailing/Marketing : 0.2727272727272727

Probability of Female opting for Undecided: 0.0

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.



The screenshot shows a Jupyter Notebook interface. The top part displays a pandas crosstab command and its output. The command is: `pd.crosstab(df.Gender, df['Grad Intention'], margins = True, normalize='index')`. The output is a table with 'Grad Intention' as columns (No, Undecided, Yes) and 'Gender' as rows (Female, Male, All). The bottom part of the screenshot shows a question: "2.4.2 Find the probability that a randomly selected student is a female and does NOT have a laptop."

```
In [325]: pd.crosstab(df.Gender, df['Grad Intention'], margins = True, normalize='index')
Out[325]:
```

Grad Intention	No	Undecided	Yes
Gender			
Female	0.272727	0.393939	0.333333
Male	0.103448	0.310345	0.586207
All	0.193548	0.354839	0.451613

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

$P(\text{Graduation Intent [Yes]} | \text{Male}) = 0.586$ or 58.6%

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

```
In [328]: pd.crosstab(df.Gender, df['computer'], margins = True, normalize='index')
```

Out[328]:

Gender	Desktop	Laptop	Tablet
Female	0.060606	0.878788	0.060606
Male	0.103448	0.896552	0.000000
All	0.080645	0.887097	0.032258

$$P(\text{No Laptop} \mid \text{Female}) = 0.87$$

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

Total number of students = 62

Number of males = 29

Number of full-time employees = 10

Number of males' full-time employees =

7 Probability Male = $29/62$

Probability full time employees = $10/62$

Probability males \cap full-time employees = $7/62$

$P(\text{Male} \cup \text{Full-Time Employment}) = \text{Probability Male} + \text{Probability fulltime employees} - \text{Probability males} \cap \text{full-time employees}$

Probability Male \cup Full Time Employment = $32/62$

$P(\text{Male} \cup \text{Full-Time Employment}) = 0.516$ or 51.6%

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

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 students = 62

Number of females = 33

Total number of female International Business or Management = 8

Probability International Business and Management \cap Female =

8/62 Probability Female = 33/62

Probability female International Business or Management =

(Probability International Business or Management \cap

Female)/ (Probability female)

Probability female International Business or Management = 8/33

P (International Business or Management| Female) = 0.242 or 24.2% .Probability t

hat given a

female student is randomly chosen, she is majoring in international business or ma
nagement 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

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

Computation of Probability- P (F \cap Yes) =

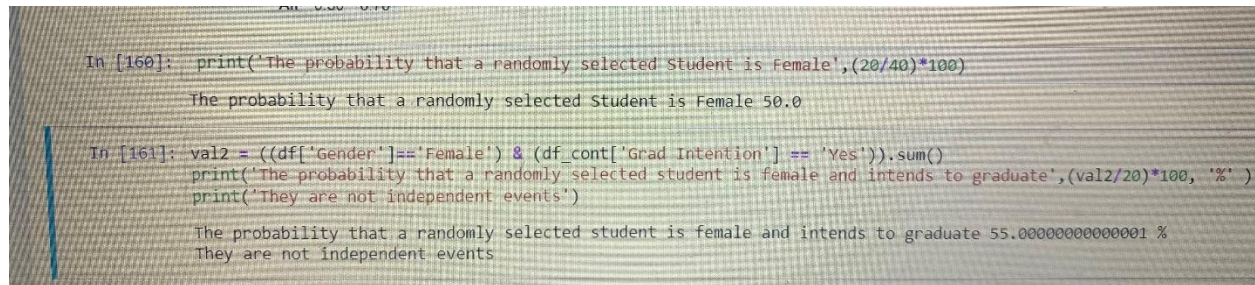
$P(F)P(\text{Yes})$ Total number of students = 62

Total Students Undecided Grad=22

Total Students new=Total Students old-Total Students Undecided

Grad=62-22=40

The probability that a randomly selected student is female and intends to graduate
55.00000000000001 %



```
In [160]: print('The probability that a randomly selected Student is Female',(20/40)*100)
The probability that a randomly selected Student is Female 50.0

In [161]: val2 = ((df['Gender']=='Female') & (df_cont['Grad Intention'] == 'Yes')).sum()
print('The probability that a randomly selected student is female and intends to graduate',(val2/20)*100, '%')
print('They are not independent events')
The probability that a randomly selected student is female and intends to graduate 55.00000000000001 %
They are not independent events
```

They 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?

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

Total number of students = 62

Number of students' GPA less than 3 = 17

Probability student's GPA less than 3 = 17/62

$P(\text{GPA} < 3.0) = 0.27$ or 27.4%

```
In [162]: No_of_stud_less = (df['GPA'] < 3).sum()
          print(No_of_stud_less)
          print(Total_value)

17
62

In [170]: pd.crosstab(df['Gender'],df['GPA'])
Out[170]:
```

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

```
In [164]: No_of_stud_less = (df['GPA'] < 3).sum()
          p_of_stud_less = No_of_stud_less/Total_value
          print('The probability that his/her GPA is less than 3 is', (p_of_stud_less)*100, '%')

The probability that his/her GPA is less than 3 is 27.419354838709676 %

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
```

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

Total number of students = 62

1) Total Male Students=29

Total Male Students $\text{sal_eq_grt50} = 14$

Probability of Random Male $\text{sal_eq_grt50} = \frac{\text{Total Male Students}}{\text{sal_eq_grt50}}$

Probability of Random Male $\text{sal_eq_grt50} = 14/29$

$P(\text{Salary} \geq 50 | \text{Male}) = 0.482$ or 48.2%

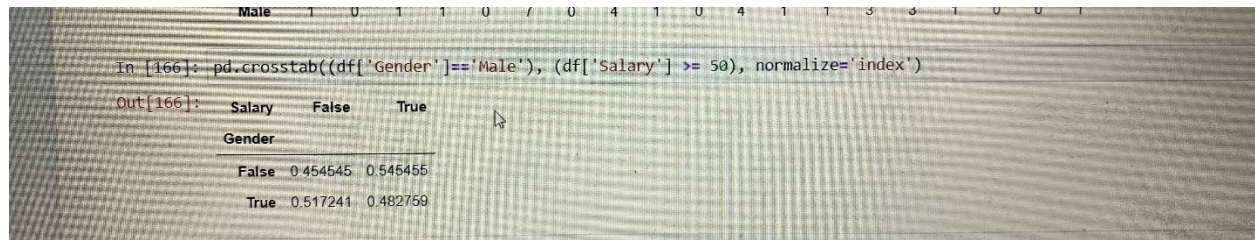
2) Number of females = 33

Number of females salary greater than 50 = 18

Probability of Random female $\text{sal_eq_grt50} = \frac{\text{Total female Students}}{\text{sal_eq_grt50}}$

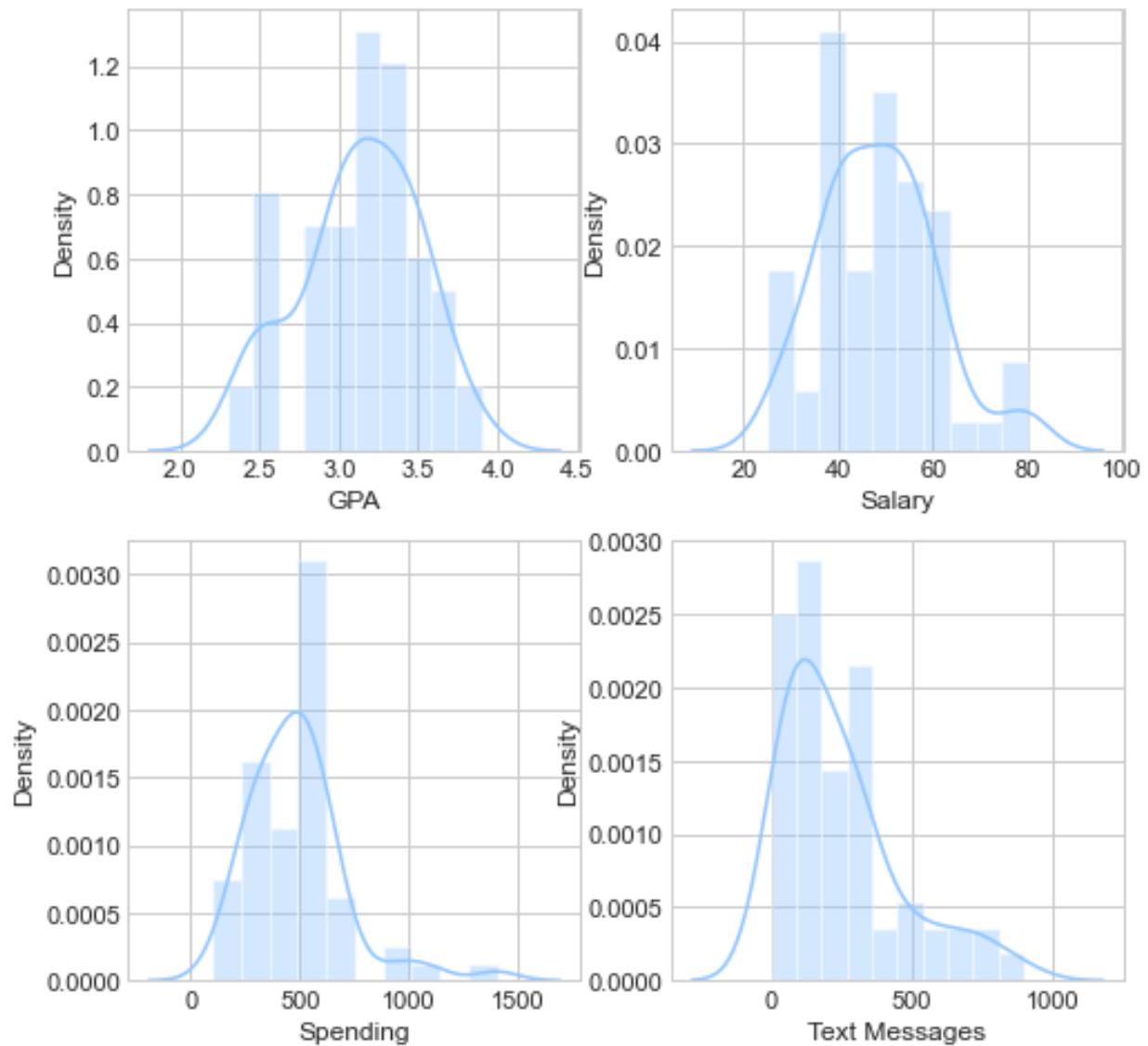
Probability of Random female $\text{sal_eq_grt50} = 18/33$

$P(\text{Salary} \geq 50 | \text{Females}) = 0.545$ or 54.5%



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.

Using the distplot, we plot the histogram plot for GPA,



From the above plots it appears that all four variables are following a Normal distribution. We can validate the same by using the empirical method for normal distribution.

The GPA box plot is normally distributed as the whiskers of the box plot are of the same length whereas the box plots of Salary, Spending, Text Messages have different whisker length and hence are not normally distributed.

.

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.

Alternative hypothesis (H_A) : mean moisture content > 0.35

Null hypothesis (H_0) : mean moisture content ≤ 0.35

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 : mean moisture content ≤ 0.35

H_A : mean moisture content > 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 : mean moisture content ≤ 0.35

H_A : mean moisture content > 0.35

	A	B
0	0.44	0.14
1	0.61	0.15
2	0.47	0.31
3	0.30	0.16
4	0.15	0.37
5	0.24	0.18
6	0.16	0.42
7	0.20	0.58
8	0.20	0.25
9	0.20	0.41

```
#H1=μ<0.35 pounds per 100 square feet

In [171]: # Executing one sample t-test

In [172]: from scipy.stats import ttest_1samp
           t_statistic,p_value=ttest_1samp(df.A,.35)
           t_statistic,p_value/2

Out[172]: (-1.4735046253382782, 0.07477633144907513)
```

One sample ttest Since $p\text{value} > 0.05$, do not reject H_0 . There is not enough evidence to conclude that the mean moisture content for Sample A shingles is no less than 0.35 pounds per 100 square feet. $p\text{ value} = 0.0748$. If the population mean moisture content is in fact no less than 0.35 pounds per 100 square feet, the probability of observing a sample of 36 shingles that will result in sample mean moisture content of 0.3167 pounds per 100 square feet or less is 0.0748.

```
73]: from scipy.stats import ttest_1samp
      t_statistic, p_value = ttest_1samp(df.B,.35,nan_policy='omit')
      t_statistic,p_value/2

173]: (-3.1003313069986995, 0.0020904774003191826)

#Inferences One sample ttest Since pvalue<0.05, reject H0. There is enough evidence to conclude that the mean moisture content for Sample B shingles is not less than 0.35 pounds per 100 square feet. p-value=0.0021. If the population mean moisture content is in fact no less than 0.35 pounds per 100 square feet, the
```

One sample ttest Since $p\text{value} < 0.05$, reject H_0 . There is enough evidence to conclude that the mean moisture content for Sample B shingles is not less than 0.35 pounds per 100 square feet. $p\text{-value} = 0.0021$. If the population mean moisture content is in fact no less than 0.35 pounds per 100 square feet, the probability of observing a

sample of 31 shingles that will result in a sample mean moisture content of 0.2735 pounds per 100 square feet or less is 0.0021

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?

Two sample t test As the pvalue $> \alpha$, do **not** reject H_0 ; **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, **and** that the variances of the two distributions are the same. If those assumptions are **not** likely to be met, another testing procedure could be use.

```
In [156]: from scipy.stats import ttest_ind
          t_statistic,p_value=ttest_ind(df['A'],df['B'],equal_var=True,nan_policy='omit')
          t_statistic,p_value
Out[156]: (1.2896282719661123, 0.2017496571835306)
```

Two sample ttest As the pvalue $> \alpha$, do not reject H_0 ; and we can say that population mean for shingles A and B are equal Test. Assumptions When running a two sample ttest the assumptions are that the distributions of the two populations are normal and that the variances of the distributions are the same.

Hence, at 95% confidence level, there is sufficient evidence to prove that population means in shingles A is equal to population mean in shingles B.