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).

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
0	1	Retail	Other	12669	9656	7561	214	2674	1338
1	2	Retail	Other	7057	9810	9568	1762	3293	1776
2	3	Retail	Other	6353	8808	7684	2405	3516	7844
3	4	Hotel	Other	13265	1196	4221	6404	507	1788
4	5	Retail	Other	22615	5410	7198	3915	1777	5185

Channel – shows whether the channel is hotel or retail

Region – sales across different regions

Fresh, Milk, Grocery, Frozen, Detergents Paper, Delicatessen – These are the 6 items distributed in different regions.

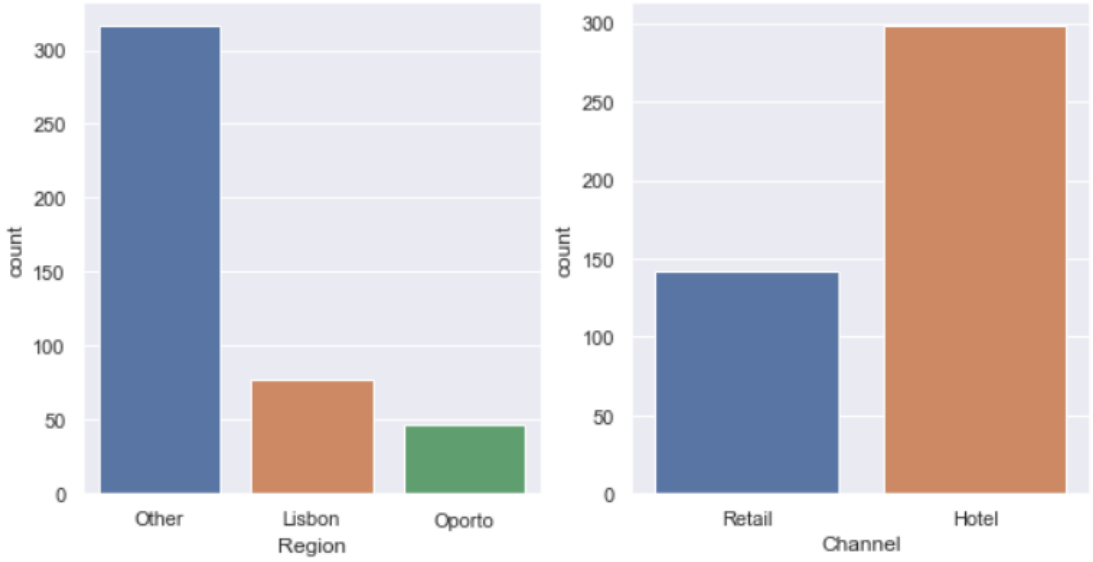
1.1 Use methods of descriptive statistics to summarize data. Which Region and which Channel spent the most? Which Region and which Channel spent the least?

	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

Observation :

- There are 440 counts in each and every column
- Channel has 2 unique values
- Region has 3 unique values
- Excluding the Buyer/Spender, we can say that Fresh has the highest mean and Delicatessen has the lowest mean.
- The minimum value is 3 for Fresh, Grocery, Detergents Paper and Delicatessen

- Fresh has the maximum value 112151
- 25%, 50%, 75% are the inter-quartile ranges which is nothing but IQR



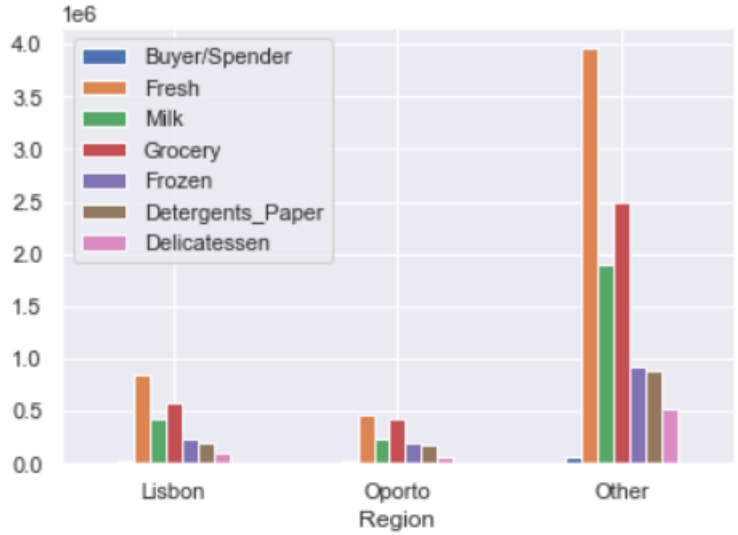
From the graph we can observe that:

- Other Region spends the more
- Hotel channel spends the more
- Oporto region spends the less
- Retail channel spends the less

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.

Varieties across Region:

Region	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Lisbon	18095	854833	422454	570037	231026	204136	104327
Oporto	14899	464721	239144	433274	190132	173311	54506
Other	64026	3960577	1888759	2495251	930492	890410	512110



From the graph, we can observe that sales based on regions:

- Fresh, Milk, Grocery, Frozen, Detergent Papers are the highest selling in all regions.
- Fresh, Milk, Grocery, Frozen, Detergents Paper are the highest selling products in other regions than the Lisbon and Oporto regions.
- Delicatessen is the lowest selling product in all regions but some what it is higher in the other region.
- Frozen and Detergents Paper are the least selling in all regions, it may be associated the person who is buying Frozen may also buy Detergents Papers also.

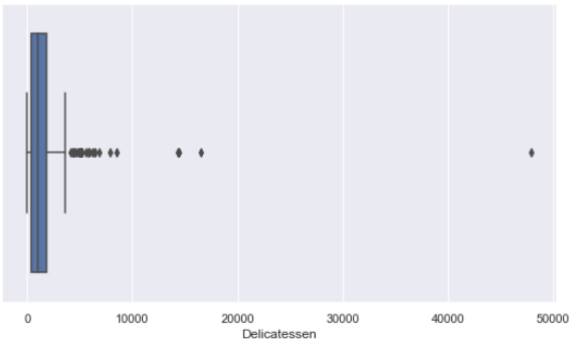
Varieties across Region:

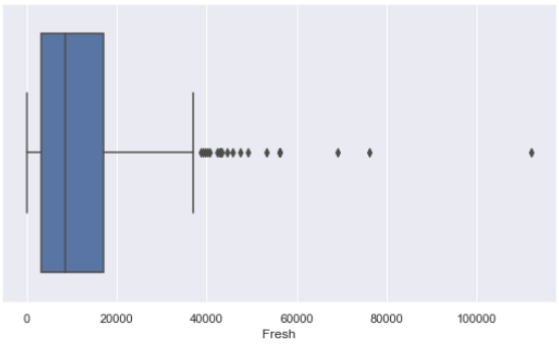
Channel	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Hotel	71034	4015717	1028614	1180717	1116979	235587	421955
Retail	25986	1264414	1521743	2317845	234671	1032270	248988

From the graph, we can observe that sales based on channel:

- It seems Fresh is the daily need for the hotel channels for their customers than the retail.
- Milk, Grocery and Frozen are not much deviating as it is a need for hotel channel but not more than Fresh.
- Grocery sales are higher than the Fresh in retail channel.
- Delicatessen item sale is high in hotel channel and low in retail channel.

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?

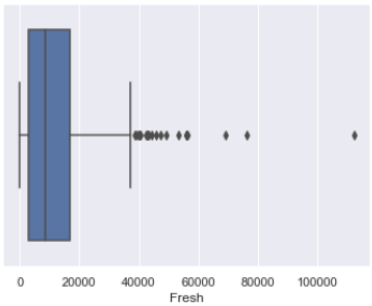


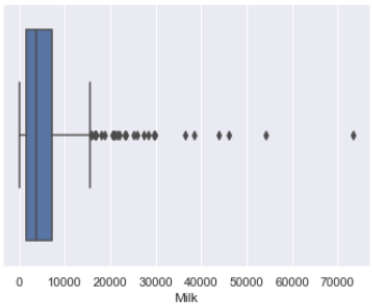


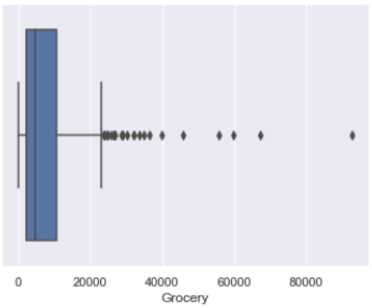
From the above box plot:

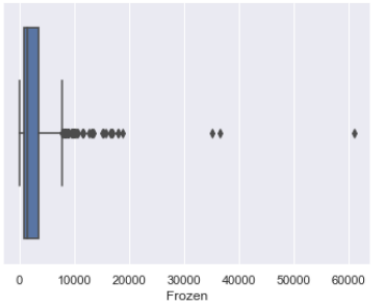
- we can observe that 'Delicatessen' item has the most inconsistent which is nothing but high variations and 'Fresh' item has the least inconsistent which is nothing but low variations

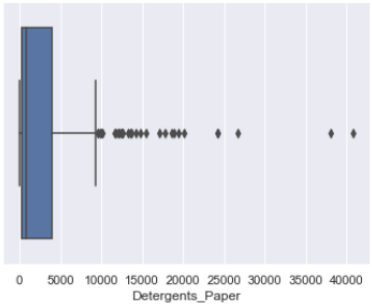
1.4 Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.

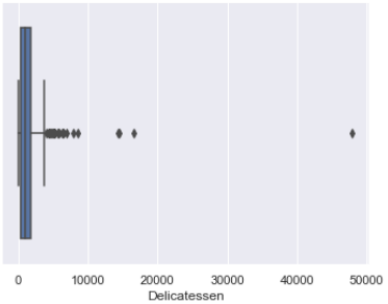












From the above plots:

- All the items from the plots have outliers extremely.
- These outliers may cause the business workflow

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

- Based on the analysis, we can observe that most of the buyers are from other regions than the Lisbon and Oporto regions. More preferably they buy the Fresh varieties and there are some daily needs like Grocery and Milk which all the regions buy.
- We found that varieties like Frozen, Detergents Paper and Delicatessen are not popular among the regions and retail channels. So, we can try to maximize these products where the demand is more.

My recommendations would be varieties like Fresh, Grocery and Milk to all regions and channels as it is daily needs. Sometimes the demand may be extremely higher or some what it can decrease but it won't decrease extremely low. Varieties like Frozen, Detergent Papers and Delicatessen can be sold in Hotel channels as there will be high usage and threshold stocks to retail channel as there will be minimal usage. So based on the population, usage and season we can plan the products need to be sold to various regions and channels.

SURVEY DATA ANALYSIS

Problem Statement:

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).

	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
0	1	Female	20	Junior	Other	Yes	2.9	Full-Time	50.0	1	3	350	Laptop	200
1	2	Male	23	Senior	Management	Yes	3.6	Part-Time	25.0	1	4	360	Laptop	50
2	3	Male	21	Junior	Other	Yes	2.5	Part-Time	45.0	2	4	600	Laptop	200
3	4	Male	21	Junior	CIS	Yes	2.5	Full-Time	40.0	4	6	600	Laptop	250
4	5	Male	23	Senior	Other	Undecided	2.8	Unemployed	40.0	2	4	500	Laptop	100

Id – unique number for each and every student
Age – age of each student
Class – whether the student is junior, senior or sophomore
Major – major subject in which student pursued
Grad Intention – student wants to graduate, not or undecided
GPA – score for each student
Employment – student current job situation
Salary – salary of each student
Social Networking – active socially platforms count
Satisfaction – feedback from students
Spending – amount spent by students
Computer – whether the student has laptop, desktop or laptop
Text Messages – count of words in feedback form

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
ID	62.0	NaN	NaN	NaN	31.5	18.041619	1.0	16.25	31.5	46.75	62.0
Gender	62	2	Female	33	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Age	62.0	NaN	NaN	NaN	21.129032	1.431311	18.0	20.0	21.0	22.0	26.0
Class	62	3	Senior	31	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Major	62	8	Retailing/Marketing	14	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Grad Intention	62	3	Yes	28	NaN	NaN	NaN	NaN	NaN	NaN	NaN
GPA	62.0	NaN	NaN	NaN	3.129032	0.377388	2.3	2.9	3.15	3.4	3.9
Employment	62	3	Part-Time	43	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Salary	62.0	NaN	NaN	NaN	48.548387	12.080912	25.0	40.0	50.0	55.0	80.0
Social Networking	62.0	NaN	NaN	NaN	1.516129	0.844305	0.0	1.0	1.0	2.0	4.0
Satisfaction	62.0	NaN	NaN	NaN	3.741935	1.213793	1.0	3.0	4.0	4.0	6.0
Spending	62.0	NaN	NaN	NaN	482.016129	221.953805	100.0	312.5	500.0	600.0	1400.0
Computer	62	3	Laptop	55	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Text Messages	62.0	NaN	NaN	NaN	246.209677	214.46595	0.0	100.0	200.0	300.0	900.0

- Male count is 29, Female count is 33

- 21 is the age mean and maximum age is 26
- Maximum GPA is 3.9 and Minimum GPA is 2.3
- 55 students use Laptop for education
- In employment part-time is more than others

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

2.1.3. Gender and Employment

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

2.1.4. Gender and Computer

Computer Gender	Desktop	Laptop	Tablet	All
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?

To find P(Male):
Count (male) = 29
Total count = 62

P(Male) = count(male)/Total count
P(Male) = 29/62 = 0.46774 or 46.77%

The 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?

To find P(Female):
Count (Female) = 33
Total count = 62

P(Female) = count(male)/Total count
P(Female) = 33/62 = 0.53225 or 53.22%

The 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.

Conditional probability of different majors,

Major Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Female	0.090909	0.090909	0.212121	0.121212	0.121212	0.090909	0.272727	0.000000
Male	0.137931	0.034483	0.137931	0.068966	0.206897	0.137931	0.172414	0.103448
All	0.112903	0.064516	0.177419	0.096774	0.161290	0.112903	0.225806	0.048387

Prob (Different Major)/count(male)
The above snippet shows the probability of male choosing different majors, the clear output is below

```
P(Male[Accounting]) =  0.13793103448275862
P(Male[CIS]) =  0.034482758620689655
P(Male[Economics/Finance]) =  0.13793103448275862
P(Male[International Business]) =  0.06896551724137931
P(Male[Management]) =  0.20689655172413793
P(Male[Other]) =  0.13793103448275862
P(Male[Retailing/Marketing]) =  0.1724137931034483
P(Male[Undecided]) =  0.10344827586206896
```

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

Conditional probability of different majors,

Major Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Female	0.090909	0.090909	0.212121	0.121212	0.121212	0.090909	0.272727	0.000000
Male	0.137931	0.034483	0.137931	0.068966	0.206897	0.137931	0.172414	0.103448
All	0.112903	0.064516	0.177419	0.096774	0.161290	0.112903	0.225806	0.048387

Prob (Different Major)/count(female)
The above snippet shows the probability of female choosing different majors, the clear output is below

```
P(Female[Accounting]) =  0.09090909090909091
P(Female[CIS]) =  0.09090909090909091
P(Female[Economics/Finance]) =  0.21212121212121213
P(Female[International Business]) =  0.12121212121212122
P(Female[Management]) =  0.12121212121212122
P(Female[Other]) =  0.09090909090909091
P(Female[Retailing/Marketing]) =  0.2727272727272727
P(Female[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.

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

Count(male) = 29
Male intends to graduate = 17

P (Male and intends to graduate) = male intends to graduate/ count(male)
= 17/29
= 0.5862 or 58.62%
The probability that a randomly chosen student is a male and intends to graduate is 58.62%.

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

Count(female) = 33
Female who doesn't have laptop = P (female with desktop) + P (female with tablet)
= 2+2
= 4
P (Female who doesn't have laptop) = Female who doesn't have laptop/count(female)
= 4/33
= 0.1212 or 12.12%
The probability that a randomly chosen student is a male and intends to graduate is 12.12%.

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

P (Male | Fulltime)
= P (Male count/ Total count) + P (Full-time/ Total full time) – P (Male. Full time/Total count)
= (29/62) + (10/62) – (7/62)
= 0.5161 or 51.61%

The probability that randomly chosen student is a male or has a full-time employment is 51.61%.

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

Count(female) = 33
International business or management = 4 + 4 = 8

P (international business or management | Female)
= international business or management / count(female)
= 8/33
=0.2424 or 24.24%

The conditional probability that given a female student is randomly chosen, she is majoring in international business or management is 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	Yes
Gender		
Female	0.45	0.55
Male	0.15	0.85
All	0.30	0.70

The graduate intention and being female are not independent events because only female has 50% but female who intends to graduate is 55%

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?

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	All
Gender																	
Female	1	1	2	0	1	3	5	2	4	3	2	4	1	2	1	1	33
Male	0	0	4	2	2	1	2	5	2	2	5	2	2	0	0	0	29
All	1	1	6	2	3	4	7	7	6	5	7	6	3	2	1	1	62

P (his/her GPA less than 3) = (1+1+6+2+3+4)/62
= 17/62
= 0.2741 or 27.41%

The probability that his/her GPA is less than 3 is 27.41%

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	All
Gender																				
Female	0	5	1	0	1	5	1	1	0	1	5	0	0	5	5	0	1	1	1	33
Male	1	0	1	1	0	7	0	4	1	0	4	1	1	3	3	1	0	0	1	29
All	1	5	2	1	1	12	1	5	1	1	9	1	1	8	8	1	1	1	2	62

Male earn 50 or more:

P (Male earn 50 or more) = P (male salary >= 50)/ count(male)

= (4+1+1+3+3+1+0+0+1)/29

= 14/29

= 0.4827 or 48.27%

The probability that a randomly selected male earns 50 or more is 48.27%.

Salary	False	True
Gender		
False	0.454545	0.545455
True	0.517241	0.482759
All	0.483871	0.516129

Female earn 50 or more:

P (Female earn 50 or more) = P (female salary >=50)/ count(female)

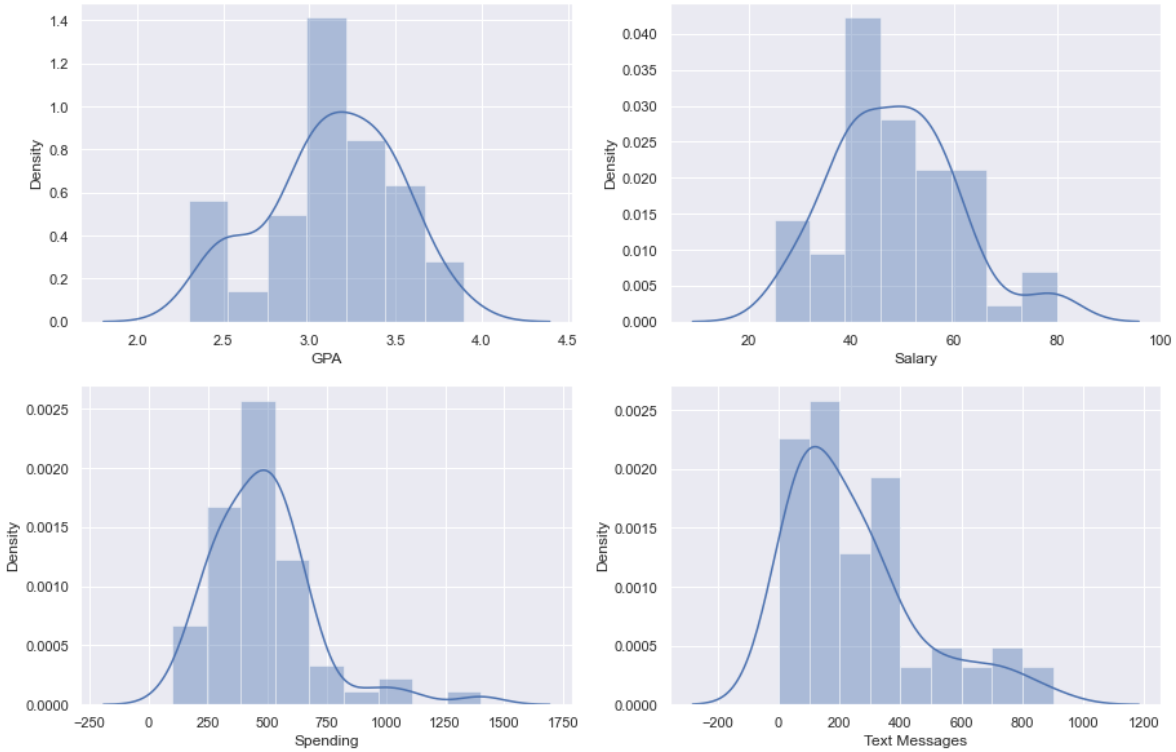
= (5+0+0+5+5+0+1+1+1)/33

= 18/33

= 0.5454 or 54.54%

The probability that a randomly selected female earns 50 or more is 54.54%.

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.



From the graph we can observe that:

- Salary, Spending and Text messages are normally distributed
- For the GPA, majority of data is distributed on the right side, hence it is negatively skewed
- If we are going with positive skewness then our algorithm going to make good predictions and very poor predictions on the higher side
- If we are going with negative skewness then our algorithm going to make good predictions on higher side and very poor predictions on the lower side
- By looking into positive skewness and negative skewness we find the likelihood situation

Conclusion:

Through out the analysis we can conclude that based on the survey of 62 students response in which both male and female are included. Many students have the intention of graduating in retailing/marketing. 75% are looking for a part time jobs. Approx 35% students have not decided about their career. Students who all looking for job they are expecting a mean salary around 50. 55 students have laptop for their education.

Shingles Analysis A & B

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.

	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

A – shingle is weighed and dried.
B – seems to be again reweighed readings.

```
A    0
B    5
dtype: int64
```

5 null values present in the data.

	count	mean	std	min	25%	50%	75%	max
A	36.0	0.316667	0.135731	0.13	0.2075	0.29	0.3925	0.72
B	31.0	0.273548	0.137296	0.10	0.1600	0.23	0.4000	0.58

- A has 36 counts, B has 31 counts
- The mean value of A is higher than B
- The max value for A is 0.72 and min value is 0.13
- The max value for B is 0.58 and min value is 0.10

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.

Given:
we have two independent samples of shingles A and B, where population standard deviation is unknown. So we will go with t-test. Since we have to find the mean moisture level is less than the permissible limit for the both samples we will use one sample t-test for sample A and sample B separately

For sample A:

step1: Define null and alternative hypothesis
H0: mu >= 0.35
HA: mu < 0.35

step2: Decide the significance level
alpha = 0.05

step3: identify the test statistic
from the given data and above table, we can observe that,
we have one sample A
sample size n>30
population standard deviation is not known
Hence we use tdist and tstat for one sample ttest. one tailed test is used here.

step4: calculate t_statistic and p_value
t_statistic : -1.4735046253382782
p_value : 0.14955266289815025

step5: Decide to reject or accept null hypothesis

(0.14955266289815025 > 0.05)
P_value is greater than alpha

- we fail to reject null hypothesis
- we conclude that moisture content is greater than permissible limit in sample A

For sample B:

step1: Define null and alternative hypothesis
H0: mu >= 0.35
HA: mu < 0.35

step2: Decide the significance level
alpha = 0.05

step3: identify the test statistic
from the given data and above table, we can observe that
we have one sample B
sample size n>30
population standard deviation is not known

omit the null values
Hence we use tdist and tstat for one sample ttest. one tailed test is used here.

step4: calculate t_statistic and p_value
t_statistic : -3.1003313069986995
p_value : 0.004180954800638365

step5: Decide to reject or accept null hypothesis

(0.004180954800638365 > 0.05)
P_value is lesser than alpha

- we reject null hypothesis
- we conclude that moisture content is less than permissible limit in sample B

From the above tests for sample A and B, we can observe that for sample B has enough evid ence to reject null hypothesis and hence we conclude that moisture content is less than the permissible limit in sample B

3.2 Do you think that the population mean for shingles A and B are equal? For m 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?

step1: Define null and alternative hypothesis
H0: muA != muB
HA: muA = muB

step2: Decide the significance level
alpha = 0.05

step3: identify the test statistic
from the given data and above table, we can observe that,
we have two samples A and B
sample size n>30
sample sizes for both samples are not same
population standard deviation is not known
omit the null values
Hence we use tdist and tstat for two sample ttest. Two-Tailed test is used here.

step4: calculate t_statistic and p_value
t_statistic : 1.2896282719661123
p_value : 0.2017496571835306

step5 : Decide to reject or accept null hypothesis

(0.2017496571835306 > 0.05)
P_value is greater than alpha

- we fail to reject null hypothesis
- we conclude that population means are not same