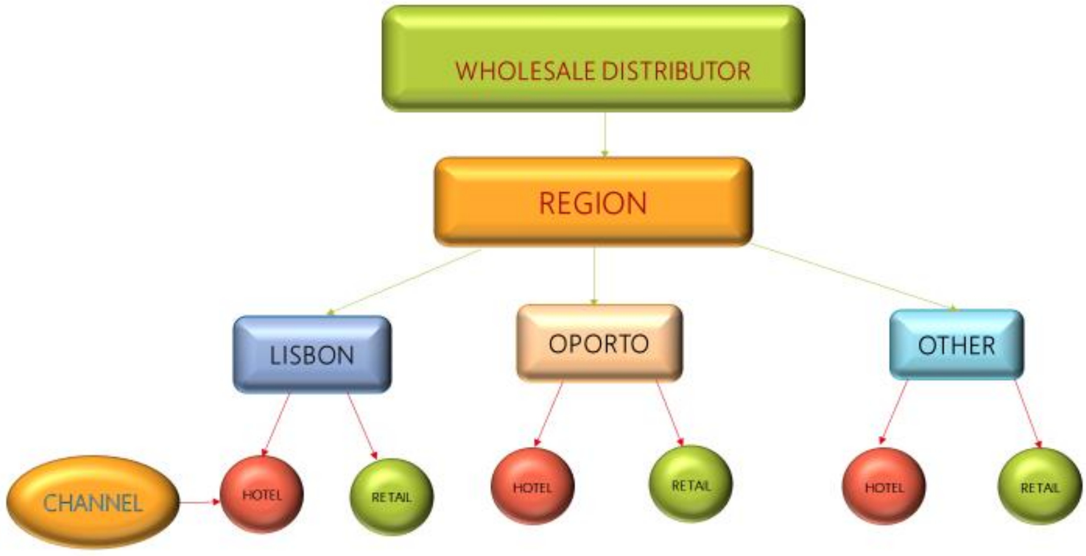
SMDM PROJECT

SUGANTHE RAMYA.M. K

Wholesale Customer Analysis

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 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).



Items sold by wholesale distributor

1) **FRESH:** annual spending on fresh products

2) **MILK**: annual spending on milk products

3) **GROCERY**: annual spending on grocery products

4) **FROZEN**: annual spending on frozen products

5) **DETERGENTS_PAPER**: annual spending on detergents and paper products

6) **DELICATESSEN**: annual spending on and delicatessen products

Wholesale dataset

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

Information on wholesale Customer Dataset

```
RangeIndex: 440 entries, 0 to 439

Data columns (total 9 columns):

#   Column          Non-Null Count  Dtype

---  ------          --------------  -----

0   Buyer/Spender    440 non-null    int64

1   Channel          440 non-null    object

2   Region           440 non-null    object

3   Fresh            440 non-null    int64

4   Milk             440 non-null    int64

5   Grocery          440 non-null    int64

6   Frozen           440 non-null    int64

7   Detergents_Paper  440 non-null    int64

8   Delicatessen      440 non-null    int64

dtypes: int64(7), object(2)

memory usage: 31.1+ KB
```

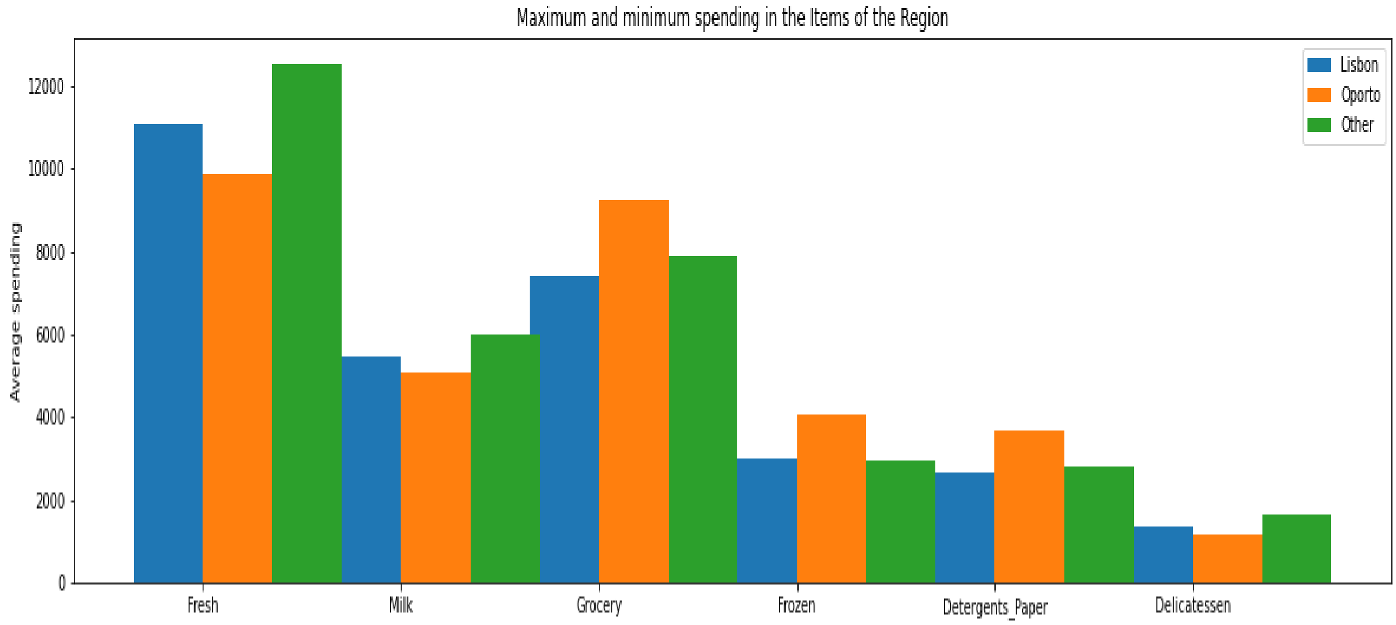
Inference

- This dataset consists of 7 continuous variables and 2 discrete variables
- Total number of entries = 440
- Total number of columns = 9
- There is no null values in this dataset

1.1 Use methods of descriptive statistics to summarize data. Which Region and which Channel seems to spend more? Which Region and which Channel seems to spend less?

REGION

	Fresh		Milk		Grocery		Frozen		Detergents_Paper		Delicatessen	
	mean	sum	mean	sum	mean	sum	mean	sum	mean	sum	mean	sum
Region												
Lisbon	11101.7	854833	5486.4	422454	7403.1	570037	3000.3	231026	2651.1	204136	1354.9	104327
Oporto	9887.7	464721	5088.2	239144	9218.6	433274	4045.4	190132	3687.5	173311	1159.7	54506
Other	12533.5	3960577	5977.1	1888759	7896.4	2495251	2944.6	930492	2817.8	890410	1620.6	512110



INFERENCE:

Based on region table and bar plot, We can conclude that :

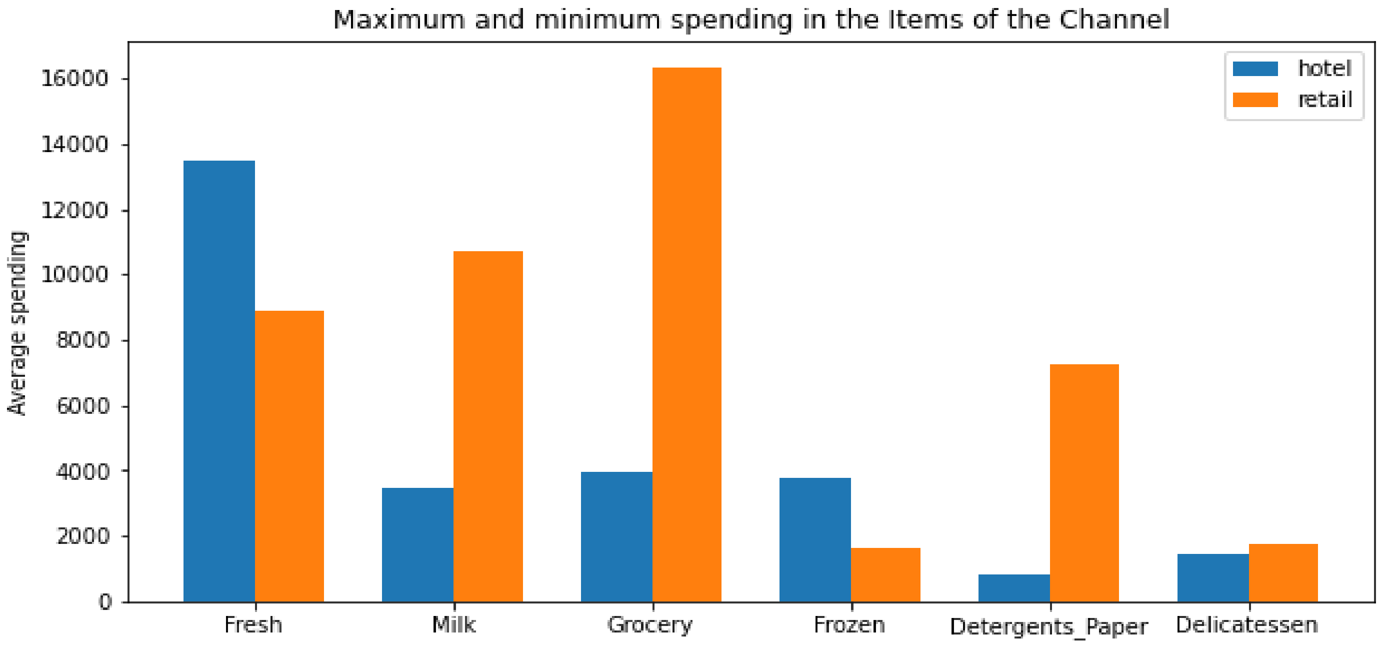
1.The Buyer/Spender seems spending more money on Other Region

2.The Buyer/Spender seems spending less money on Oporto Region

3. The Buyers seems buying more fresh varieties and less delicatessen

1.1Use methods of descriptive statistics to summarize data. Which Region and which Channel seems to spend more? Which Region and which Channel seems to spend less?

	Fresh		**Milk**		**Grocery**		**Frozen**		**Detergents_Paper**		**Delicatessen**	
	mean	sum	mean	sum	mean	sum	mean	sum	mean	sum	mean	sum
Channel												
Hotel	13475.6	4015717	3451.7	1028614	3962.1	1180717	3748.3	1116979	790.6	235587	1416	421955
Retail	8904.3	1264414	10716.5	1521743	16322.9	2317845	1652.6	234671	7269.5	1032270	1753.4	248988



INFERENCE:

Based on channel table and bar plot, We can conclude that :

1.The Buyer/Spender seems spending more money on Retail channel

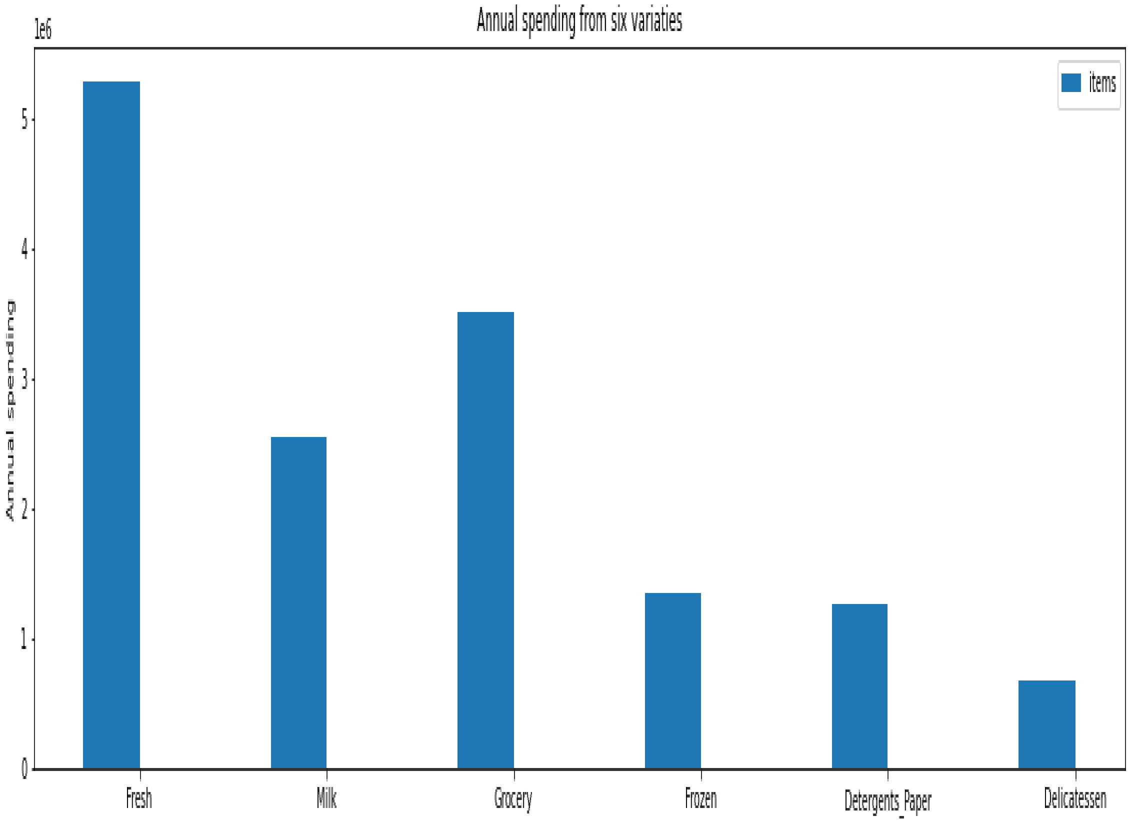
2.The Buyer/Spender seems spending less money on Hotel channel

3. Buyers are spending more on Grocery in retail channel

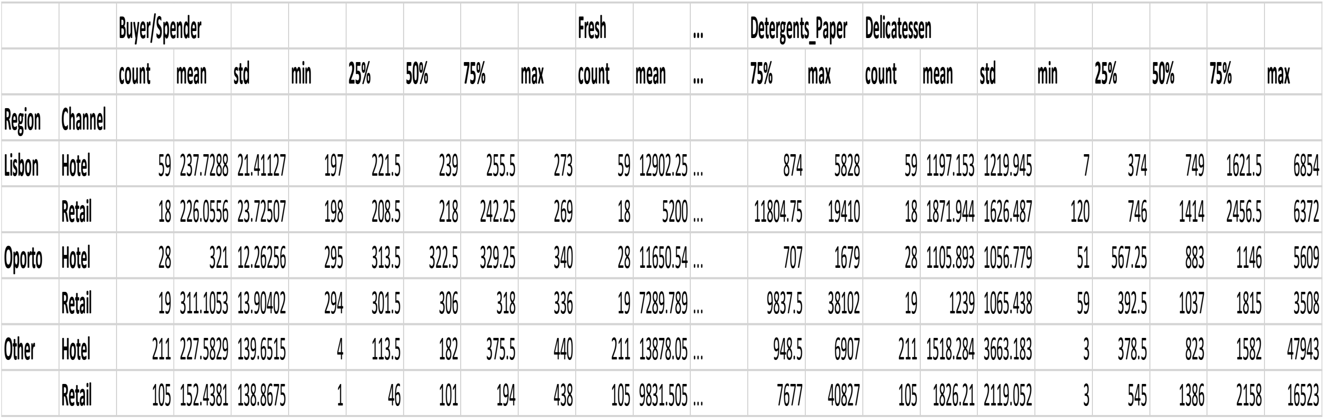
4. Buyers are spending less on Delicatessen in both retail and hotel channel

Annual spending

Varieties	Annual spending (Euro)
Fresh	5280131
Milk	2550357
Grocery	3498562
Frozen	1351650
Detergents_Paper	1267857
Delicatessen	670943
Total	14619500



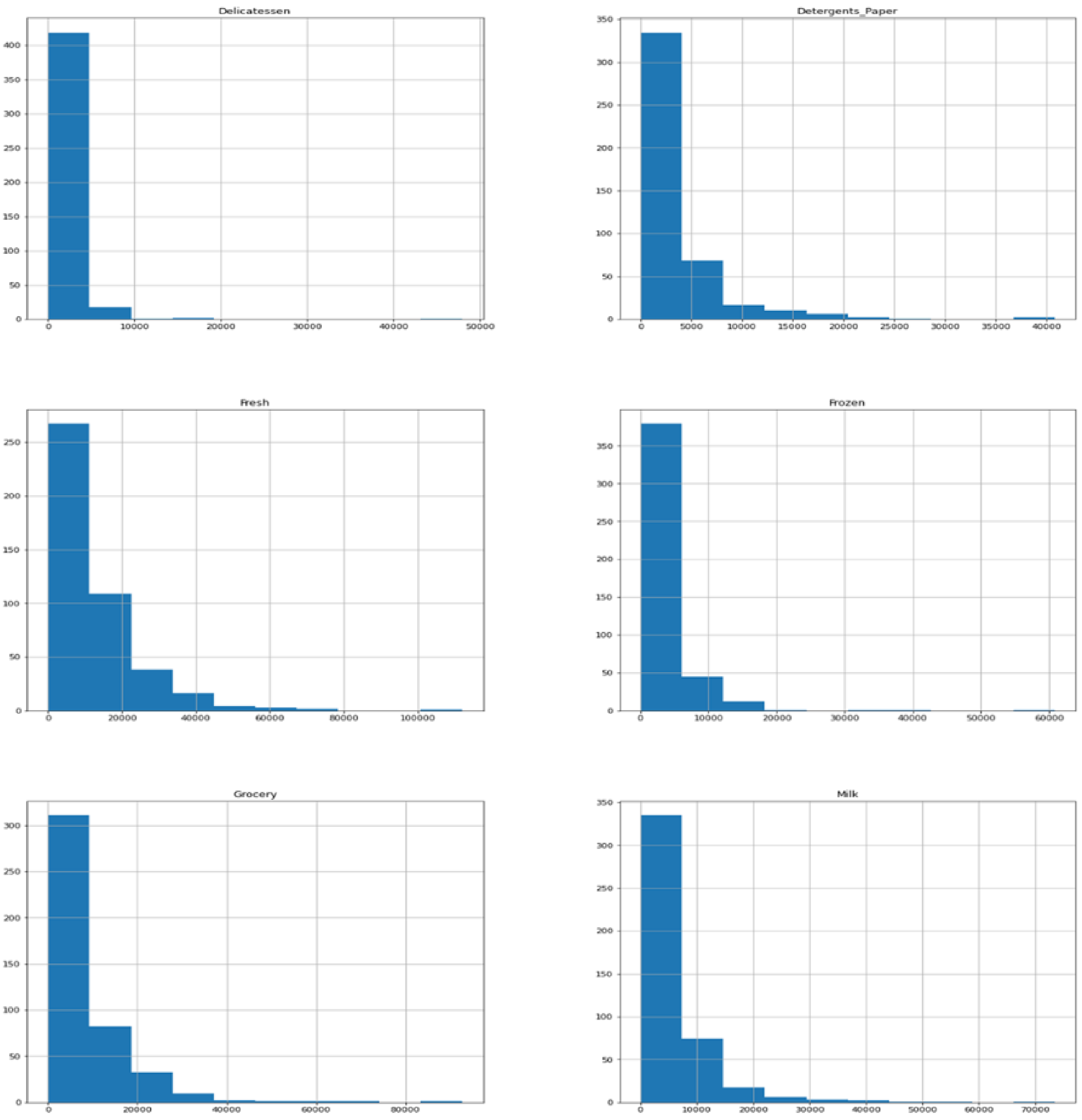
1.2 There are 6 different varieties of items are considered. Do all varieties show similar behavior across Region and Channel? Provide justification for your answer



		Buyer/Spender								Fresh		...	Detergents_Paper		Delicatessen							
		count	mean	std	min	25%	50%	75%	max	count	mean	...	75%	max	count	mean	std	min	25%	50%	75%	max
Region	Channel																					
Lisbon	Hotel	59	237.7288	21.41127	197	221.5	239	255.5	273	59	12902.25	...	874	5828	59	1197.153	1219.945	7	374	749	1621.5	6854
	Retail	18	226.0556	23.72507	198	208.5	218	242.25	269	18	5200	...	11804.75	19410	18	1871.944	1626.487	120	746	1414	2456.5	6372
Oporto	Hotel	28	321	12.26256	295	313.5	322.5	329.25	340	28	11650.54	...	707	1679	28	1105.893	1056.779	51	567.25	883	1146	5609
	Retail	19	311.1053	13.90402	294	301.5	306	318	336	19	7289.789	...	9837.5	38102	19	1239	1065.438	59	392.5	1037	1815	3508
Other	Hotel	211	227.5829	139.6515	4	113.5	182	375.5	440	211	13878.05	...	948.5	6907	211	1518.284	3663.183	3	378.5	823	1582	47943
	Retail	105	152.4381	138.8675	1	46	101	194	438	105	9831.505	...	7677	40827	105	1826.21	2119.052	3	545	1386	2158	16523

This describe() function clearly explains that

1. Buyer/Spender are spending more from other(hotel)
2. Buyer/Spender are spending very less in Delicatessen
3. Buyer/Spender with other products , buying more fresh items .



Inference

1.skewness > 0 : more weight in the left tail of the distribution.

2.Delicatessen is highly skewed to the left

3.All the items are skewed to the left

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?

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440	440	440	440	440	440
mean	12000.3	5796.27	7951.28	3071.93	2881.49	1524.87
std	12647.33	7380.38	9503.16	4854.67	4767.85	2820.11
min	3	55	3	25	3	3
25%	3127.75	1533	2153	742.25	256.75	408.25
50%	8504	3627	4755.5	1526	816.5	965.5
75%	16933.75	7190.25	10655.75	3554.25	3922	1820.25
max	112151	73498	92780	60869	40827	47943

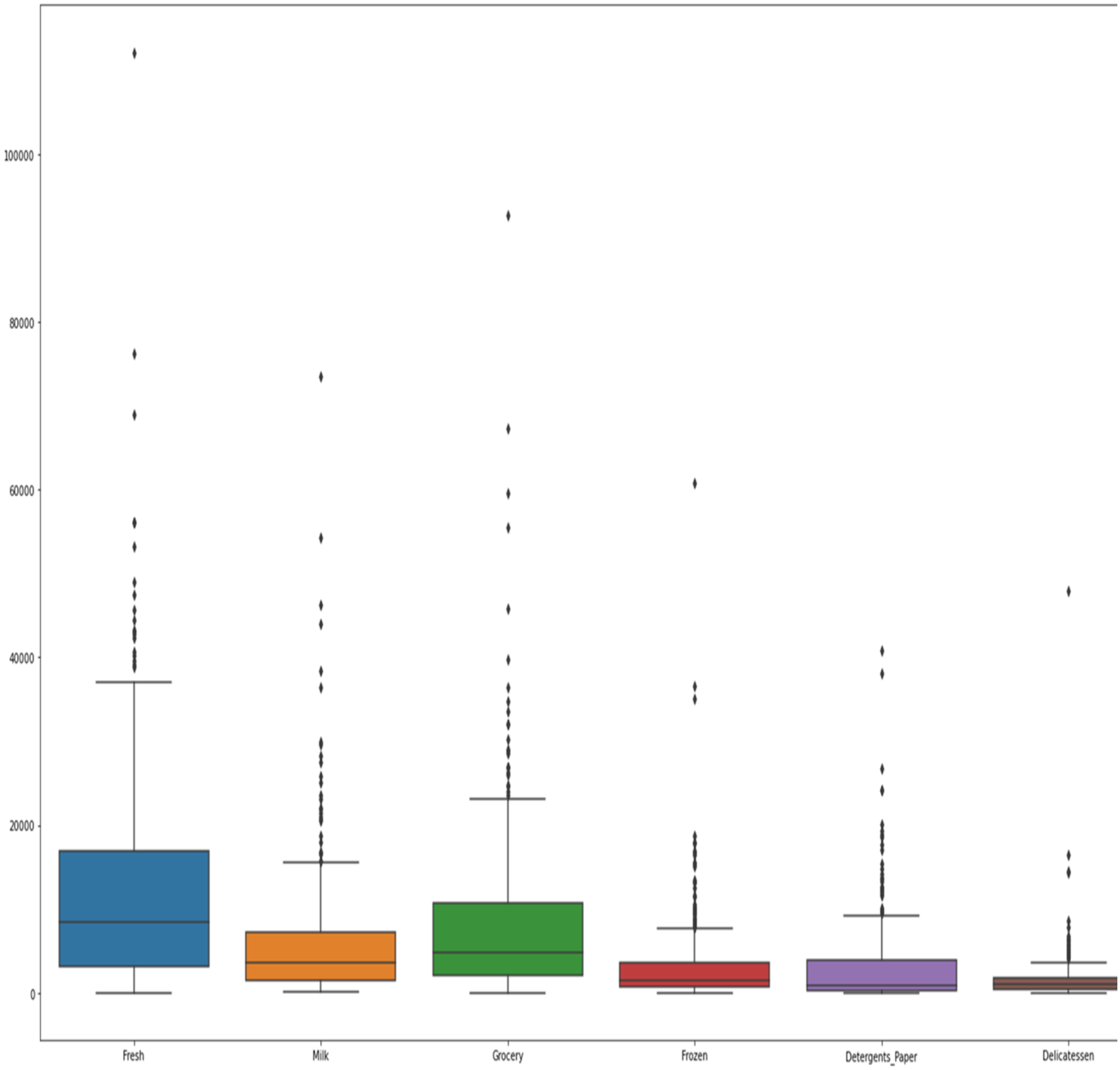
Variates	**CV**
Fresh	1.053918
Milk	1.273298
Grocery	1.195174
Detergents_paper	1.654647
Delicatessen	1.849410

Inference:

1.Delicatessen items has most inconsistent behavior

2.Fresh items has less inconsistent behavior

1.4 Are there any outliers in the data?



Based on above bar plot, all the varieties in the dataset have outliers

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 EDA analysis:

1. Wholesale distributor earned the total revenue of 14,619,500 euro from the six items in the three regions of the Portugal

Fresh :	5280131
Milk :	2550357
Grocery :	3498562
Frozen :	1351650
Detergents_Paper :	1267857
Delicatessen :	670943

2. From the six items , Buyers/spenders are spending more the fresh item and very less on the Delicatessen

3. There are more Retail buyers than the Hotel buyers¶

Conclusion for problem 1:

I recommend the wholescale distributor to increase the sale of fresh items in the Retail channel from other region of the Portugal.

Wholesale distributor also needs to find more Retail buyers ,so that there will be more sales

CLEAR MOUNTAIN STATE UNIVERSITY SURVEY

- **PROBLEM 2**
- **The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. CMSU creates and distributes a survey of 14 questions and receives responses from 62 undergraduates (stored in the Survey data set).**

CLEAR MOUNTAIN STATE UNIVERSITY SURVEY DATASET

	count	mean	std	min	25%	50%	75%	max
Age	62	21.12903	1.431311	18	20	21	22	26
GPA	62	3.129032	0.377388	2.3	2.9	3.15	3.4	3.9
Salary	62	48.54839	12.08091	25	40	50	55	80
Social Networking	62	1.516129	0.844305	0	1	1	2	4
Satisfaction	62	3.741935	1.213793	1	3	4	4	6
Spending	62	482.0161	221.9538	100	312.5	500	600	1400
Text Messages	62	246.2097	214.466	0	100	200	300	900

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62 entries, 0 to 61
Data columns (total 14 columns):
 #  Column       Non-Null Count     Dtype
---  ------         --------------       -----
 0  ID              62 non-null        int64
 1  Gender          62 non-null        object
 2  Age             62 non-null        int64
 3  Class           62 non-null        object
 4  Major           62 non-null        object
 5  Grad Intention 62 non-null         object
 6  GPA             62 non-null        float64
 7  Employment   62 non-null         object
 8  Salary          62 non-null        float64
 9  Social
    Networking     62 non-null        int64
 10  Satisfaction     62 non-null        int64
 11  Spending        62 non-null        int64
 12  Computer       62 non-null        object
 13  Text Messages 62 non-null         int64
dtypes: float64(2), int64(6), object(6)
memory usage: 6.9+ KB
```

INFERENCE

1. Variables in the data set, GPA, Salary, Spending, and Text Messages are numerical (continuous)

2. Total number of entries = 62

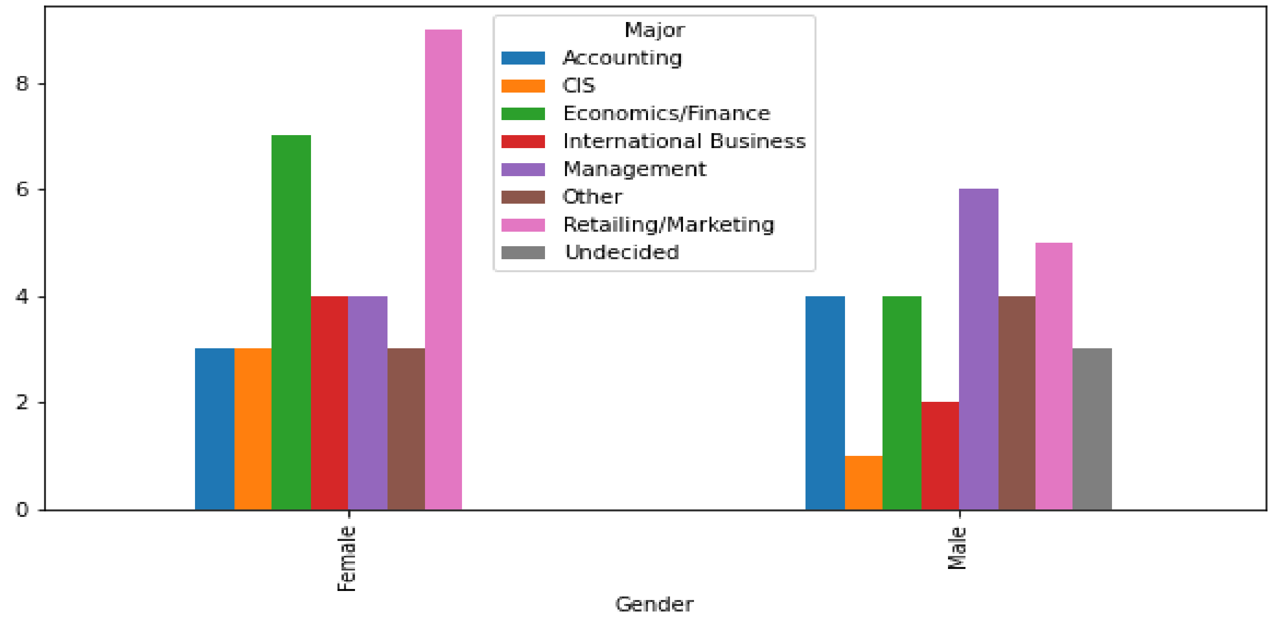
3. Total number of columns = 14

4. There is no null values in this dataset

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

2.1.1. Gender and Major

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



Based on plot, we can interpret

FEMALE:

1. Retailing/Marketing Major is highest chosen Major

2. Accounting, CIS and other majors are least chosen Major

3. All Female decided their majors

MALE:

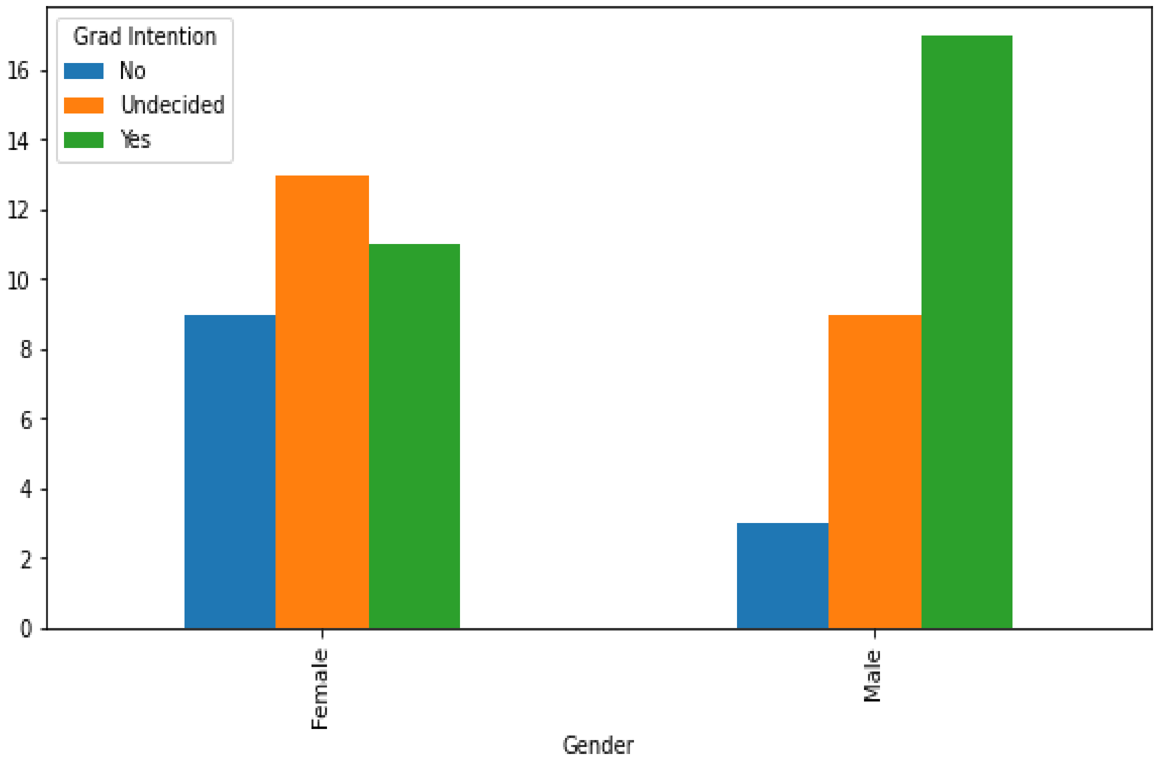
1. Management Major is highest chosen Major

2. CIS least chosen Major

3.Three male students undecided their major

2.1.2. Gender and Grad Intention

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



Based on plot, we can interpret

FEMALE:

1. **Most of the female students are undecided about**

grad intention

2. only few female students have no grad intention

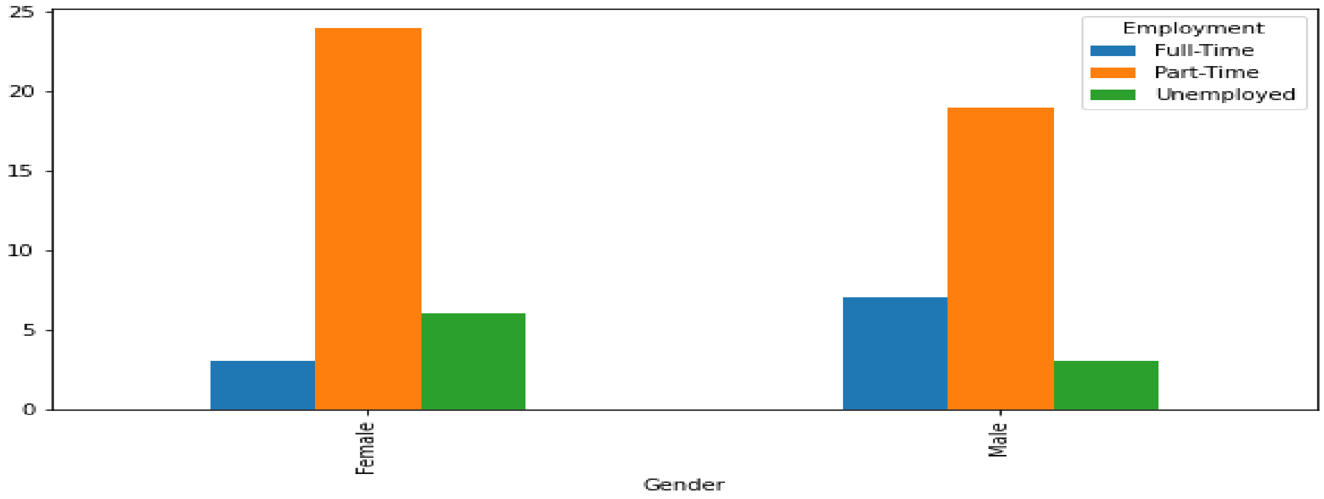
MALE:

1. Most of the male student indented to graduate

2. Very few male students have no grad intention

2.1.3. Gender and Employment

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



Based on plot, we can interpret

FEMALE:

1. Most of the female students have part-time employment

2. only few female students are unemployed

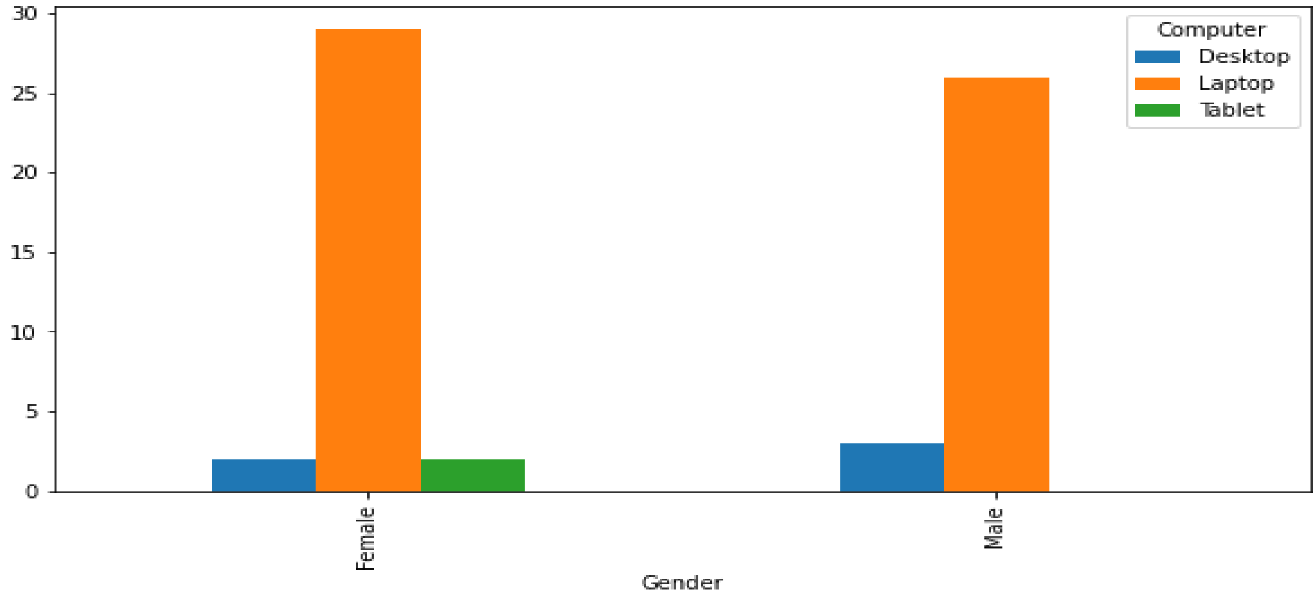
MALE:

1. Most of the male student have part-time employees

2. Very few male students are unemployed

2.1.4. Gender and Computer

Computer	Desktop	Laptop	Tablet
Gender			
Female	2	29	2
Male	3	26	0



1.Most of female and male students have laptops

2.Male students don’t have tablet

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?

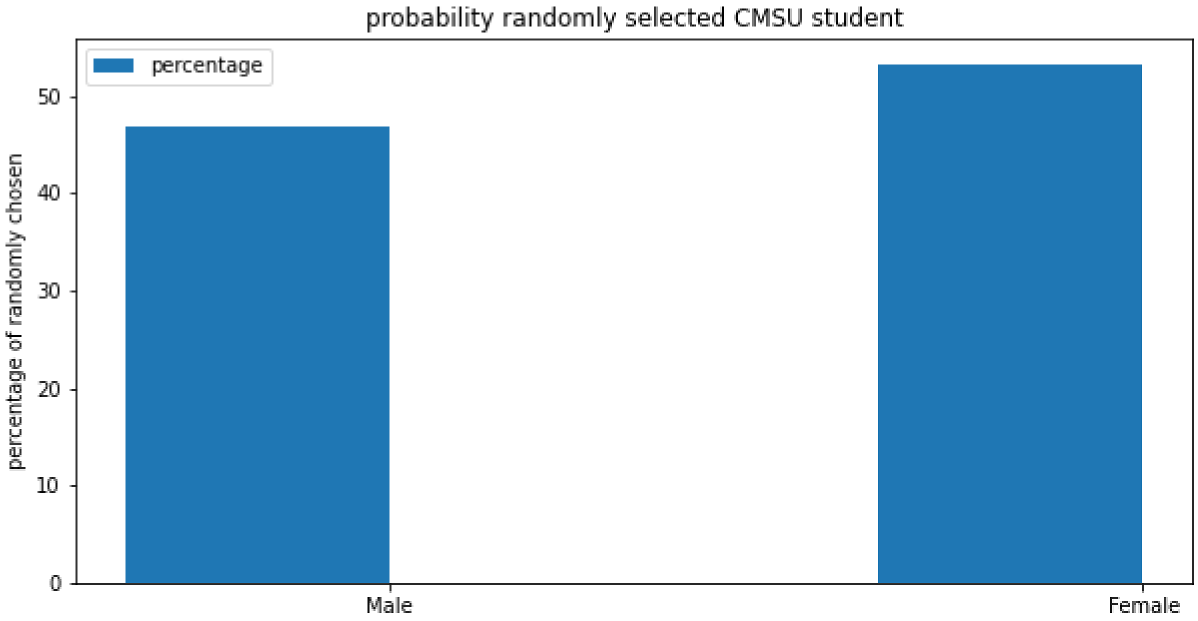
The probability of male randomly selected CMSU student : 46.77%

===

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

The probability of female randomly selected CMSU student : 53.23%

==



Inference

Female are more randomly chosen compare to male

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.

Major	Probability of male choosing different majors
Accounting	6.45 %
CIS	1.61 %
Economics/Finance	6.45%
International Business	3.23 %
Management	9.68 %
Other	6.45%
Retailing/Marketing	8.06%
Undecided	4.84 %

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

Major	Probability of Female choosing different majors
Accounting	4.84 %
CIS	4.84%
Economics/Finance	11.29%
International Business	6.45 %
Management	6.45 %
Other	4.84%
Retailing/Marketing	14.52%
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.

Randomly chosen male who intend to graduate: 27.0 %

===

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

Randomly selected female who does not have a laptop : 11.29 %

==

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 either a male or has full-time employment?

Randomly chosen Male has full time employment = 50.0 %

===

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

Female student randomly chosen international business or management: 12.90%

==

2.6. Construct a contingency table of Gender and Intent to Graduate at 2 levels (Yes/No). The Undecided students are not considered now, and the table is a 2x2 table. Do you think the graduate intention and being female are independent events?

Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

Probability of female graduate intention = 13.75 %

Yes, graduate intention and 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

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

	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	1	3	350	Laptop	200
2	3	Male	21	Junior	Other	Yes	2.5	Part-Time	45	2	4	600	Laptop	200
3	4	Male	21	Junior	CIS	Yes	2.5	Full-Time	40	4	6	600	Laptop	250
4	5	Male	23	Senior	Other	Undecided	2.8	Unemployed	40	2	4	500	Laptop	100
5	6	Female	22	Senior	Economics/Finance	Undecided	2.3	Unemployed	78	3	2	700	Laptop	30
10	11	Female	23	Senior	Economics/Finance	Yes	2.8	Full-Time	50	2	5	400	Laptop	200
23	24	Male	22	Senior	Undecided	Yes	2.6	Full-Time	45	1	5	400	Laptop	600
27	28	Female	20	Junior	International Business	Yes	2.9	Part-Time	50	3	1	900	Laptop	100
31	32	Male	20	Junior	Other	Yes	2.9	Part-Time	47	3	1	300	Laptop	300
33	34	Male	22	Senior	Retailing/Marketing	Yes	2.6	Full-Time	40	1	4	1400	Laptop	800
37	38	Female	21	Sophomore	Accounting	Yes	2.5	Part-Time	60	2	3	500	Laptop	600
38	39	Male	24	Junior	Economics/Finance	Yes	2.8	Part-Time	50	1	6	600	Laptop	50
39	40	Male	19	Sophomore	Retailing/Marketing	Yes	2.5	Unemployed	50	2	5	300	Laptop	100
47	48	Male	19	Sophomore	Undecided	Undecided	2.5	Part-Time	80	2	4	500	Laptop	150
57	58	Female	21	Senior	International Business	No	2.4	Part-Time	40	1	3	1000	Laptop	10
58	59	Female	20	Junior	CIS	No	2.9	Part-Time	40	2	4	350	Laptop	250
59	60	Female	20	Sophomore	CIS	No	2.5	Part-Time	55	1	4	500	Laptop	500

Probability of GPA his/her less than 3 = 27.42 %

2.6.2. 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.

	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
14	15	Male	21	Senior	Management	Yes	3.2	Part-Time	54	3	4	600	Laptop	400
17	18	Male	21	Junior	Economics/Finance	Undecided	3.1	Part-Time	55	2	3	600	Laptop	300
18	19	Male	19	Junior	Economics/Finance	Yes	3.5	Part-Time	52	2	5	500	Laptop	300
21	22	Male	18	Sophomore	Accounting	Undecided	3	Unemployed	60	1	4	600	Laptop	500
25	26	Male	24	Senior	Management	Yes	3.3	Full-Time	60	0	1	300	Laptop	40
26	27	Male	20	Junior	Economics/Finance	Yes	3.1	Full-Time	65	1	5	375	Laptop	300
28	29	Male	22	Senior	Retailing/Marketing	Yes	3.3	Part-Time	55	1	6	1100	Laptop	60
30	31	Male	20	Junior	Accounting	Undecided	3.4	Part-Time	55	2	3	500	Laptop	750
38	39	Male	24	Junior	Economics/Finance	Yes	2.8	Part-Time	50	1	6	600	Laptop	50
39	40	Male	19	Sophomore	Retailing/Marketing	Yes	2.5	Unemployed	50	2	5	300	Laptop	100
40	41	Male	22	Junior	Accounting	Yes	3.2	Full-Time	60	1	4	680	Desktop	200
47	48	Male	19	Sophomore	Undecided	Undecided	2.5	Part-Time	80	2	4	500	Laptop	150
51	52	Male	21	Senior	Management	No	3	Part-Time	50	1	4	500	Laptop	200
54	55	Male	21	Senior	Other	Yes	3.4	Part-Time	50	1	4	250	Desktop	700

Probability of male earn 50 or more = 22.58%

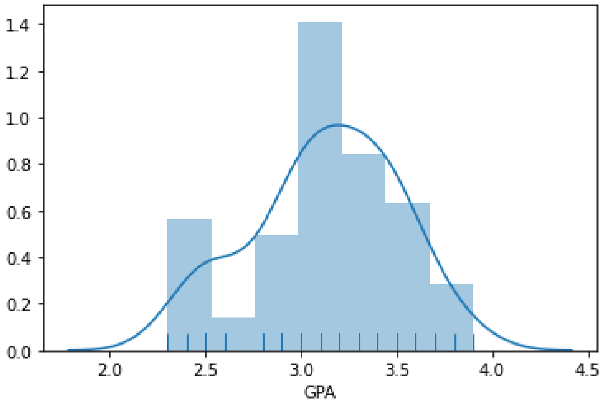
	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
0	1	Female	20	Junior	Other	Yes	2.9	Full-Time	50	1	3	350	Laptop	200
5	6	Female	22	Senior	Economics/Finance	Undecided	2.3	Unemployed	78	3	2	700	Laptop	30
6	7	Female	21	Junior	Other	Undecided	3	Part-Time	50	1	3	500	Laptop	50
7	8	Female	22	Senior	Other	Undecided	3.1	Full-Time	80	1	2	200	Tablet	300
10	11	Female	23	Senior	Economics/Finance	Yes	2.8	Full-Time	50	2	5	400	Laptop	200
16	17	Female	19	Junior	CIS	Undecided	3.7	Part-Time	55	1	4	450	Laptop	150
19	20	Female	20	Junior	Management	Undecided	3.2	Unemployed	60	2	6	300	Laptop	350
20	21	Female	22	Junior	Retailing/Marketing	Undecided	3.2	Part-Time	55	1	3	690	Laptop	50
22	23	Female	22	Senior	Retailing/Marketing	Undecided	3	Part-Time	55	0	4	300	Laptop	35
24	25	Female	20	Junior	Economics/Finance	Yes	3	Part-Time	55	1	3	600	Laptop	300
27	28	Female	20	Junior	International Business	Yes	2.9	Part-Time	50	3	1	900	Laptop	100
35	36	Female	26	Junior	Accounting	Yes	3.3	Part-Time	60	1	4	450	Desktop	300
37	38	Female	21	Sophomore	Accounting	Yes	2.5	Part-Time	60	2	3	500	Laptop	600
45	46	Female	21	Senior	Management	Undecided	3.8	Part-Time	60	1	4	650	Laptop	150
46	47	Female	20	Junior	Retailing/Marketing	Yes	3.5	Unemployed	60	1	3	350	Laptop	200
55	56	Female	21	Senior	Retailing/Marketing	No	3.1	Part-Time	50	1	1	300	Laptop	300
59	60	Female	20	Sophomore	CIS	No	2.5	Part-Time	55	1	4	500	Laptop	500
61	62	Female	23	Senior	Economics/Finance	No	3.2	Part-Time	70	2	3	250	Laptop	0

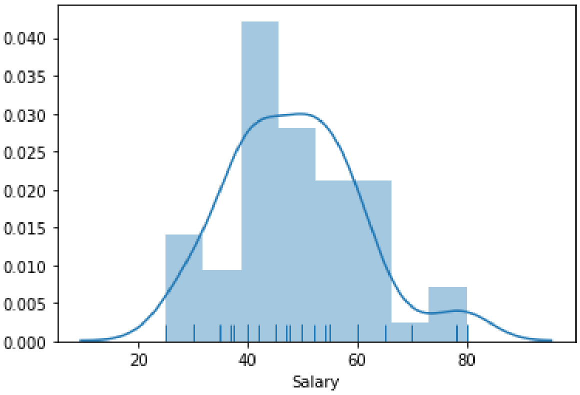
Probability of female earning 50 dollars or more = 29.03 %

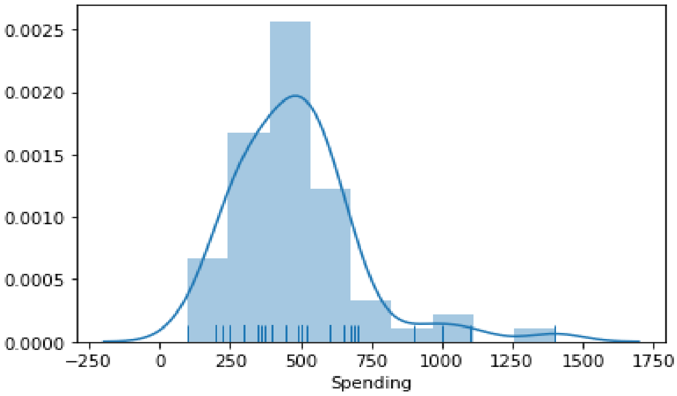
2.8. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarizing your conclusions.

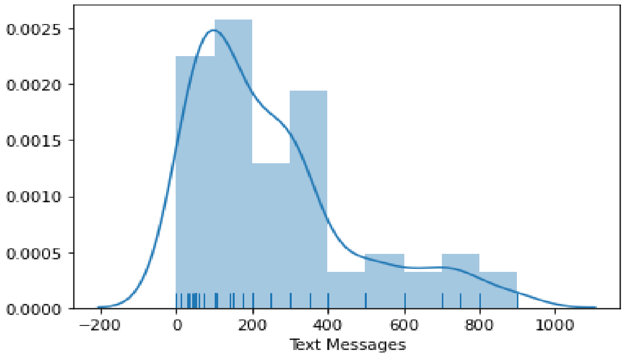
Summary of the dataset:

	count	mean	std	min	25%	50%	75%	max
GPA	62.0	3.129032	0.377388	2.3	2.9	3.15	3.4	3.9
Salary	62.0	48.548387	12.080912	25.0	40.0	50.00	55.0	80.0
Spending	62.0	482.016129	221.953805	100.0	312.5	500.00	600.0	1400.0
Text Messages	62.0	246.209677	214.465950	0.0	100.0	200.00	300.0	900.0









1. GPA ,Salary and spending seems to be normally distributed

2. Text messages is left skewed

CONCLUSION OF PROBLEM 2

- **Average salary of the students: 48.54 dollars**
- **Maximum salary of the students: 80 dollars**
- **Average GPA scored by students: 3.12**
- **Minimum GPA scored by students : 2.3**
- **Average spending by students: 482 dollars**
- **Minimum age of the students in the college : 18**
- **Maximum age of the students in the college : 26**
- **Average satisfaction level of the students : 3.74**

The manufacturers of ABC asphalt shingles

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 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.

A&B shingles dataset

	A	B
0	0.44	0.14
1	0.61	0.15
2	0.47	0.31
3	0.3	0.16
4	0.15	0.37

A&B shingles dataset summary

	A	B
count	36	31
mean	0.316667	0.273548
std	0.135731	0.137296
min	0.13	0.1
25%	0.2075	0.16
50%	0.29	0.23
75%	0.3925	0.4
max	0.72	0.58

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36 entries, 0 to 35
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   A       36 non-null     float64
 1   B       31 non-null     float64
dtypes: float64(2)
memory usage: 704.0 bytes
```

Inference

1. There are two continuous variables A and B

2. There are no null values

3. A has 36 values

4. B has 31 values

5. Both A and B variables has float datatypes

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.

A shingles

	A
0	0.44
1	0.61
2	0.47
3	0.3
4	0.15

Step 1:null and alternative hypotheses

For the A shingles, the null and alternative hypothesis to test whether the population mean moisture content is less than 0.35 per 100 square feet is given:

H0<=0.35

HA>0.35

Step 2:The significance level

Here we select α = 0.05.

The sample size for this problem is 36

Step 3: Identify the test statistic

We do not know the population standard deviation and n = 36. So we use the t distribution and the $tSTAT$ test statistic.

Step 4: Calculate the p - value and test statistic

One sample t test

t statistic: [-4406.51558207] p value: [4.02388859e-102]

Level of significance: 0.05

We have evidence to reject the null hypothesis since p value < Level of significance

Our one-sample t-test p-value= [4.02388859e-102]

Conclusion: A shingles moisture content is more than 0.35 pound per 100 square feet

B shingles

	A
0	0.44
1	0.61
2	0.47
3	0.3
4	0.15

Step 1:null and alternative hypothesis

For the B shingles, the null and alternative hypothesis to test whether the population mean moisture content is less than 0.35 per 100 square feet is given:

H0<=0.35

HA>0.35

Step 2:The significance level

Here we select α = 0.05.

The sample size for this problem is 36

Here there are 5 nan values, so during calculation t test we are using nan_policy='omit' to omit those values

Step 3: Identify the test statistic

We do not know the population standard deviation and n = 36. So we use the t distribution and the $tSTAT$ test statistic.

Step 4: Calculate the p - value and test statistic

One sample t test

t statistic: [-4044.1925072627105] p value: [1.29304495e-87]

Level of significance: 0.05

We have evidence to reject the null hypothesis since p value < Level of significance

Our one-sample t-test p-value= [1.29304495e-87]

Conclusion: B shingles moisture content is more than 0.35 pound per 100 square feet

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?

Step 1:null and alternative hypothesis

- $H0: \mu A = \mu B$ (The population mean for shingles A and B are equal)
  - $HA: \mu A \neq \mu B$ (The population mean for shingles A and B are not equal)

Step 2:The significance level

Here we select α = 0.05.

The sample size for this problem is 36

Here there are 5 nan values, so during calculation t test we are using nan_policy='omit' to omit those values

Step 3: Identify the test statistic

We do not know the population standard deviation and n = 36. So we use the t distribution and the $tSTAT$ test statistic.

Step 4: Calculate the p - value and test statistic

Two sample t test (ttest_ind)

tstat :1.2896282719661123

P Value :0.2017496571835306

RESULT:

two-sample t-test p-value= 0.2017496571835306

We do not have enough evidence to reject the null hypothesis in favor of alternative hypothesis

CONCLUSION:

We conclude that the population mean for shingles A and B are same.

CONCLUSION OF PROBLEM 3:

Based on hypothesis test , we can conclude that

One sample t test:

A Shingles

We have evidence to reject the null hypothesis since p value < Level of significance

A shingles moisture content is more than 0.35 pound per 100 square feet

B Shingles

We have evidence to reject the null hypothesis since p value < Level of significance

B shingles moisture content is more than 0.35 pound per 100 square feet

Two sample t test (ttest_ind)

We do not have enough evidence to reject the null hypothesis in favor of alternative hypothesis

The population mean for shingles A and B are same.