Wholesale Customers Analysis

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).

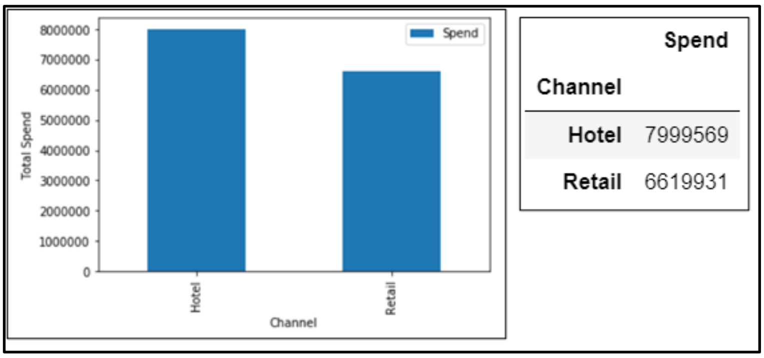
The dataset consists of 9 variables where Buyer/Spender is used as an index and has integer values, Channel and Region are categorical variables and have string values, and 6 different varieties of products i.e Fresh, Milk, Frozen, Grocery, Detergents Paper, Delicatessens have integer values.

There are 2 different Channels- Hotel and Retail, whereas there are 3 different categorisations of Regions- Lisbon, Oporto and Other where the wholesale distributor operates.

For analysing the overall spend across Channels and Regions, **we created a new variable in dataset as [Spend] = ∑ Money (spent by customers for all the 6 varieties).**

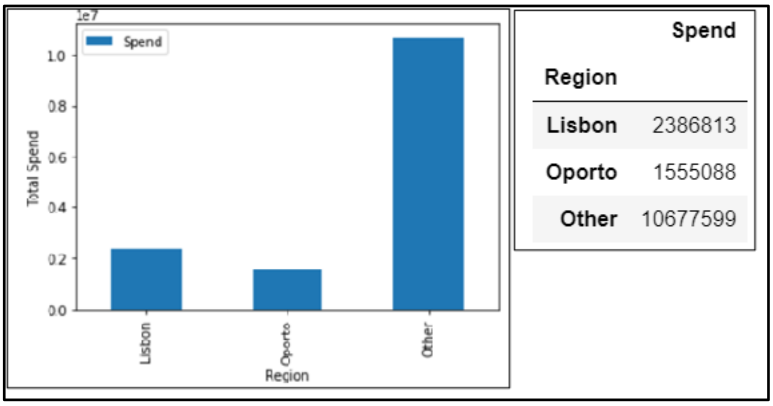
1.1 Use methods of descriptive statistics to summarize data. Which Region and which Channel seems to spend more? Which Region and which Channel seems to spend less?

So, to summarize which region and channel had maximum and minimum sales, we **grouped the data by Channel and Region respectively and used [Spend] value to decide the outcome** as it gives us a fair idea about the total expenditure by customers across all the 6 product categories. Following were our observations: -



	Spend
Channel	
Hotel	7999569
Retail	6619931

Hence, customers are spending across **Hotels(maximum)** than **Retail(minimum)** when looked **Channel-wise**.



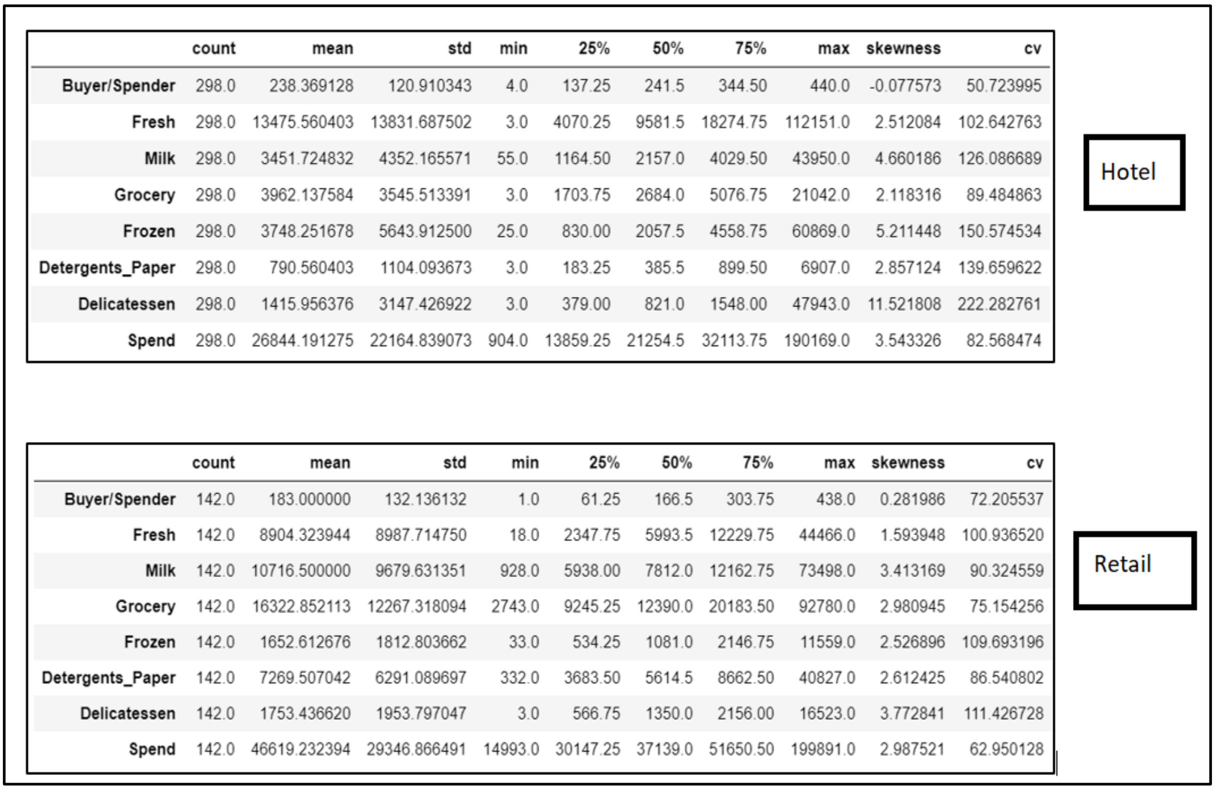
	Spend
Region	
Lisbon	2386813
Oporto	1555088
Other	10677599

Hence, customers in **Other regions are spending drastically more** compared to Lisbon and Oporto, the **least is being spent by customers in Oporto** at an overall level.

1.2 *There are 6 different varieties of items are considered. Do all varieties show similar behaviour across Region and Channel?*

To check the behaviour of items across different channels and regions, we **further sub-grouped** the data of Channel and Region based on their categories i.e **split Channel as Hotel and Retail a**nd **Region as Lisbon, Oporto and Others** and used **a describe() function** since it shows the max, min, mean, standard deviation, etc. We also **added functions like skewness and coefficient of variation** which would further help us in comparing their behaviour.

For Channel: -



	count	mean	std	min	25%	50%	75%	max	skewness	cv
Buyer/Spender	298.0	238.369128	120.910343	4.0	137.25	241.5	344.50	440.0	-0.077573	50.723995
Fresh	298.0	13475.560403	13831.687502	3.0	4070.25	9581.5	18274.75	112151.0	2.512084	102.642763
Milk	298.0	3451.724832	4352.165571	55.0	1164.50	2157.0	4029.50	43950.0	4.660186	126.086689
Grocery	298.0	3962.137584	3545.513391	3.0	1703.75	2684.0	5076.75	21042.0	2.118316	89.484863
Frozen	298.0	3748.251678	5643.912500	25.0	830.00	2057.5	4558.75	60869.0	5.211448	150.574534
Detergents_Paper	298.0	790.560403	1104.093673	3.0	183.25	385.5	899.50	6907.0	2.857124	139.659622
Delicatessen	298.0	1415.956376	3147.426922	3.0	379.00	821.0	1548.00	47943.0	11.521808	222.282761
Spend	298.0	26844.191275	22164.839073	904.0	13859.25	21254.5	32113.75	190169.0	3.543326	82.568474

Hotel

	count	mean	std	min	25%	50%	75%	max	skewness	cv
Buyer/Spender	142.0	183.000000	132.136132	1.0	61.25	166.5	303.75	438.0	0.281986	72.205537
Fresh	142.0	8904.323944	8987.714750	18.0	2347.75	5993.5	12229.75	44466.0	1.593948	100.936520
Milk	142.0	10716.500000	9679.631351	928.0	5938.00	7812.0	12162.75	73498.0	3.413169	90.324559
Grocery	142.0	16322.852113	12267.318094	2743.0	9245.25	12390.0	20183.50	92780.0	2.980945	75.154256
Frozen	142.0	1652.612676	1812.803662	33.0	534.25	1081.0	2146.75	11559.0	2.526896	109.693196
Detergents_Paper	142.0	7269.507042	6291.089697	332.0	3683.50	5614.5	8662.50	40827.0	2.612425	86.540802
Delicatessen	142.0	1753.436620	1953.797047	3.0	566.75	1350.0	2156.00	16523.0	3.772841	111.426728
Spend	142.0	46619.232394	29346.866491	14993.0	30147.25	37139.0	51650.50	199891.0	2.987521	62.950128

Retail

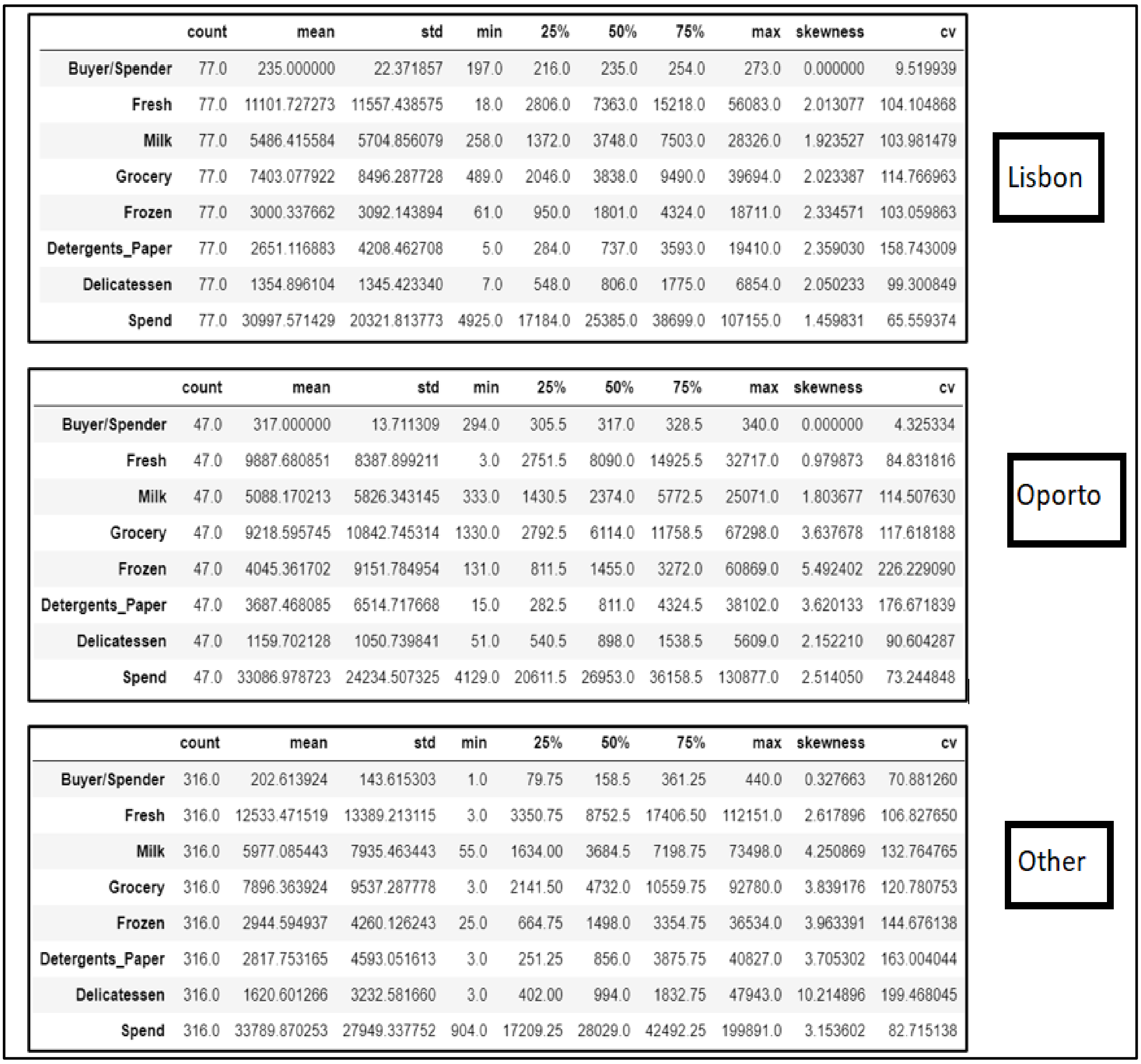
From the above data, we can say Fresh performs similar in both the channels despite the difference in the mean and standard deviation in Hotels and Retail because count was different in both the channels.

Milk, Grocery and Detergents Paper have lower mean and standard deviation in Hotels compared to Retail considering that most of the households would be purchasing (daily needs) from Retail stores, but the c.v for these varieties is greater in hotels.

Frozen products are used more in Hotels, and the data justifies that assumption as the mean, standard deviation and c.v are all higher in Hotels compared to Retail.

Delicatessens have a similar mean in Hotel and Retails but Hotels have a significantly higher standard deviation and hence higher c.v. This might be because the Hotels may be offering more varieties of delicatessens and meals.

For Region: -



	count	mean	std	min	25%	50%	75%	max	skewness	cv
Buyer/Spender	77.0	235.000000	22.371857	197.0	216.0	235.0	254.0	273.0	0.000000	9.519939
Fresh	77.0	11101.727273	11557.438575	18.0	2806.0	7363.0	15218.0	56083.0	2.013077	104.104868
Milk	77.0	5486.415584	5704.856079	258.0	1372.0	3748.0	7503.0	28326.0	1.923527	103.981479
Grocery	77.0	7403.077922	8496.287728	489.0	2046.0	3838.0	9490.0	39694.0	2.023387	114.766963
Frozen	77.0	3000.337662	3092.143894	61.0	950.0	1801.0	4324.0	18711.0	2.334571	103.059863
Detergents_Paper	77.0	2651.116883	4208.462708	5.0	284.0	737.0	3593.0	19410.0	2.359030	158.743009
Delicatessen	77.0	1354.896104	1345.423340	7.0	548.0	806.0	1775.0	6854.0	2.050233	99.300849
Spend	77.0	30997.571429	20321.813773	4925.0	17184.0	25385.0	38699.0	107155.0	1.459831	65.559374

Lisbon

	count	mean	std	min	25%	50%	75%	max	skewness	cv
Buyer/Spender	47.0	317.000000	13.711309	294.0	305.5	317.0	328.5	340.0	0.000000	4.325334
Fresh	47.0	9887.680851	8387.899211	3.0	2751.5	8090.0	14925.5	32717.0	0.979873	84.831816
Milk	47.0	5088.170213	5826.343145	333.0	1430.5	2374.0	5772.5	25071.0	1.803677	114.507630
Grocery	47.0	9218.595745	10842.745314	1330.0	2792.5	6114.0	11758.5	67298.0	3.637678	117.618188
Frozen	47.0	4045.361702	9151.784954	131.0	811.5	1455.0	3272.0	60869.0	5.492402	226.229090
Detergents_Paper	47.0	3687.468085	6514.717668	15.0	282.5	811.0	4324.5	38102.0	3.620133	176.671839
Delicatessen	47.0	1159.702128	1050.739841	51.0	540.5	898.0	1538.5	5609.0	2.152210	90.604287
Spend	47.0	33086.978723	24234.507325	4129.0	20611.5	26953.0	36158.5	130877.0	2.514050	73.244848

Oporto

	count	mean	std	min	25%	50%	75%	max	skewness	cv
Buyer/Spender	316.0	202.613924	143.615303	1.0	79.75	158.5	361.25	440.0	0.327663	70.881260
Fresh	316.0	12533.471519	13389.213115	3.0	3350.75	8752.5	17406.50	112151.0	2.617896	106.827650
Milk	316.0	5977.085443	7935.463443	55.0	1634.00	3684.5	7198.75	73498.0	4.250869	132.764765
Grocery	316.0	7896.363924	9537.287778	3.0	2141.50	4732.0	10559.75	92780.0	3.839176	120.780753
Frozen	316.0	2944.594937	4260.126243	25.0	664.75	1498.0	3354.75	36534.0	3.963391	144.676138
Detergents_Paper	316.0	2817.753165	4593.051613	3.0	251.25	856.0	3875.75	40827.0	3.705302	163.004044
Delicatessen	316.0	1620.601266	3232.581660	3.0	402.00	994.0	1832.75	47943.0	10.214896	199.468045
Spend	316.0	33789.870253	27949.337752	904.0	17209.25	28029.0	42492.25	199891.0	3.153602	82.715138

Other

From the above image we can say, the count is more in Other compared to Oporto and Lisbon as there maybe many cities considered in that grouping.

Fresh has higher mean and SD in Lisbon and Other regions compared to Oporto.

Milk has a similar mean across all 3 regions but the SD is higher in Other regions, hence the c.v is higher in Other regions.

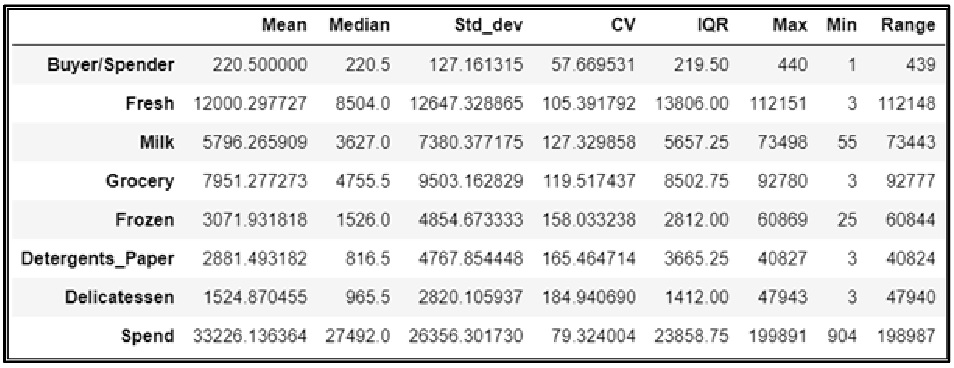
Grocery has a similar behaviour when we look at the c.v but the Oporto region has higher mean and SD compared to Other which has greater than Lisbon.

Frozen products and Detergents Paper have highest sales in Oporto when looked at mean and SD followed by Others (higher c.v) and Lisbon(lower c.v)

Delicatessens have the highest variations in Other regions.

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

To check the **consistency (less volatility)**, we used statistical functions like **mean, median, and measures of dispersion like standard deviation, coefficient of variation, IQR, range.**

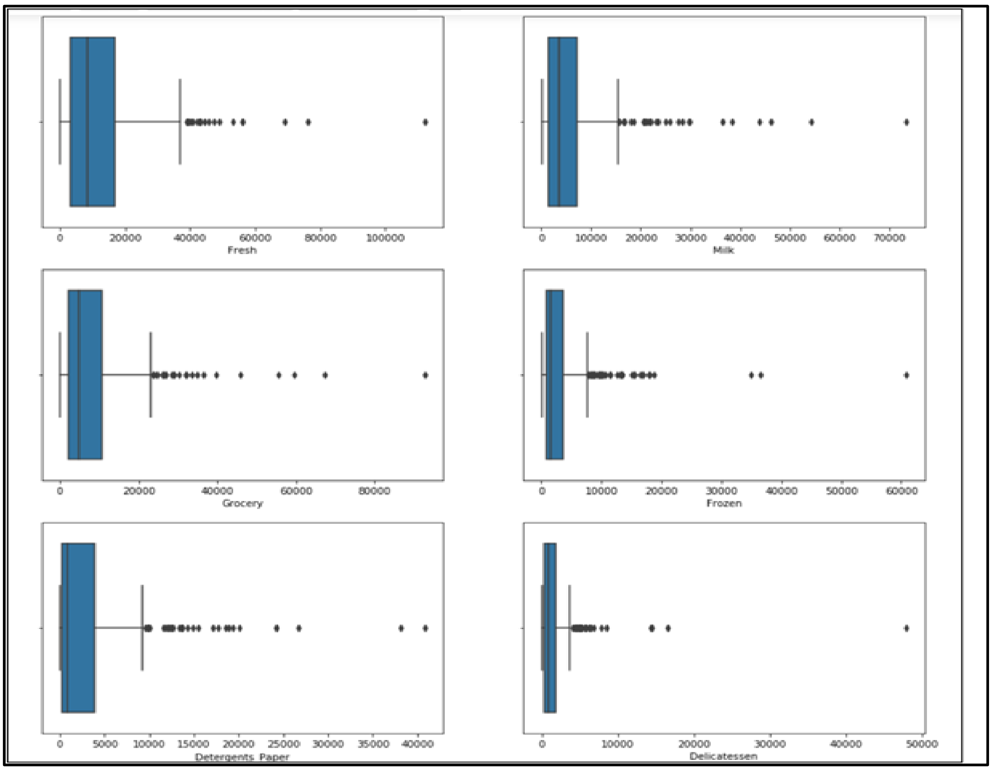


	Mean	Median	Std_dev	CV	IQR	Max	Min	Range
Buyer/Spender	220.500000	220.5	127.161315	57.669531	219.50	440	1	439
Fresh	12000.297727	8504.0	12647.328865	105.391792	13806.00	112151	3	112148
Milk	5796.265909	3627.0	7380.377175	127.329858	5657.25	73498	55	73443
Grocery	7951.277273	4755.5	9503.162829	119.517437	8502.75	92780	3	92777
Frozen	3071.931818	1526.0	4854.673333	158.033238	2812.00	60869	25	60844
Detergents_Paper	2881.493182	816.5	4767.854448	165.464714	3665.25	40827	3	40824
Delicatessen	1524.870455	965.5	2820.105937	184.940690	1412.00	47943	3	47940
Spend	33226.136364	27492.0	26356.301730	79.324004	23858.75	199891	904	198987

Looking at the output we can see that maximum range and minimum range difference were observed in Fresh and Detergents Paper respectively, but considering the volatility or difference in behaviour we considered **coefficient of variance (c.v) as the deciding parameter** as mean, standard deviations and range were significantly different for all variables and concluded that **Fresh was more consistent** compared to other items and **Delicatessen is less consistent** compared to the other varieties.
We could have considered Standard Deviation but since the data was skewed and not normally distributed, standard deviation would generate biased results

1.4 Are there any outliers in the data?

To check if there were any outliers in the data, we use a **simple box-plot** as it represents minimum value (Q1- 1.5*IQR), Q1, mean(Q2), Q3 and maximum value (Q3+ 1.5*IQR). **All the points above maximum value and lower than minimum values are considered as outliers.** So, the boxplots show if variables have the outliers present.



Looking at the above image, we can conclude that all the variables (all 6 varieties of products) i.e **Fresh, Milk, Groceries, Frozen, Detergents Paper and Delicatessen have outliers**.

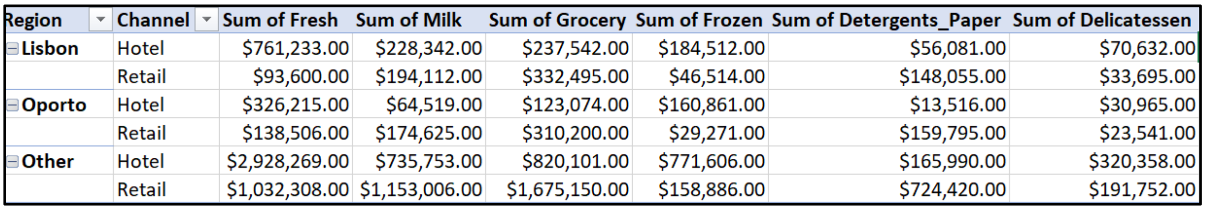
1.5 On the basis of this report, what are the recommendations?

Considering all the steps performed during the analysis, we can derive that: -
1. Based on the above report, we can see that the customers/retailers are spending more on the items in Hotel(7.99 mil) compared to the Retail(6.61 mil) and majority of retailers(298) spent on items in Hotel over 142 in Retail channel. But the average spend across Channels is greater in Retail (46.6K) compared to the Hotels (26.8K). Fresh performs similar in both the channels despite the difference in the mean and standard deviation in Hotels and Retail because count was different in both the channels. Milk, Grocery and Detergents Paper have lower mean and standard deviation in Hotels compared to Retail considering that most of the households would be purchasing (daily needs) from Retail stores, but the c.v for these varieties is greater in hotels. Frozen products are used more in Hotels, and the data justifies that assumption as the mean, standard deviation and c.v are all higher in Hotels compared to Retail. Delicatessens have a similar mean in Hotel and Retails but Hotels have a significantly higher standard deviation and hence higher c.v. This might be because the Hotels may be offering more varieties of delicatessens and meals.

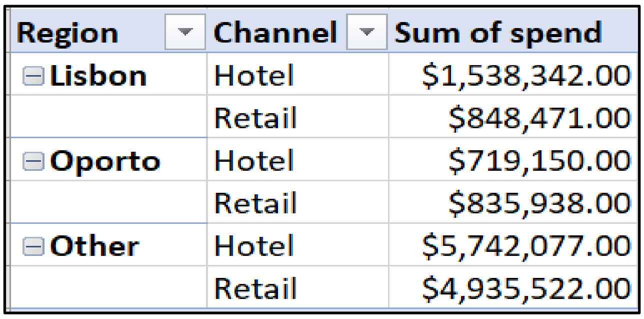
2.The spending (overall across all 6 categories of products) were in Other regions (10.6 mil) compared to Lisbon (2.38 mil) and Oporto (1.55 mil), hence the wholesaler should restock and supply more in other regions. But the average spent across regions was pretty similar i.e Lisbon (approx 31K), Oporto (approx 33K) and Others (approx 33.8K). The count is more in Other compared to Oporto and Lisbon as there maybe many cities considered in that grouping. Fresh has higher mean and SD in Lisbon and Other regions compared to Oporto. Milk has a similar mean across all 3 regions but the SD is higher in Other regions, hence the c.v is higher in Other regions. Grocery has a similar behaviour when we look at the c.v but the Oporto region has higher mean and SD compared to Other which has greater than Lisbon. Frozen products and Detergents Paper have highest sales in Oporto when looked at mean and SD followed by Others (higher c.v) and Lisbon(lower c.v). Delicatessens have the highest variations in Other regions.

3. Looking at the output we can see that maximum range and minimum range difference were observed in Fresh and Detergents Paper respectively, but considering the volatility or difference in behaviour we considered coefficient of variance (c.v) as the deciding parameter as mean, standard deviations and range were significantly different for all variables and concluded that Fresh was more consistent compared to other items and Delicatessen is less consistent compared to the other varieties. The boxplots tell us that all the varieties are positively-skewed /right-skewed as the median is closer to the bottom of the box, and the whisker is shorter on the lower end of the box.

So, we could say that the daily items should be produced and supplied more as it contributes to maximum sales mainly in the Retail channels, so we should collaborate with more retail channels and offer them the products of daily needs to increase our sales. When seen at a combined level (Region and Channel), we see that Lisbon has most of its revenue from Hotels, Oporto has better performance in Retail and Hotels perform better in Other regions.



Region	Channel	Sum of Fresh	Sum of Milk	Sum of Grocery	Sum of Frozen	Sum of Detergents_Paper	Sum of Delicatessen
⊟Lisbon	Hotel	$761,233.00	$228,342.00	$237,542.00	$184,512.00	$56,081.00	$70,632.00
	Retail	$93,600.00	$194,112.00	$332,495.00	$46,514.00	$148,055.00	$33,695.00
⊟Oporto	Hotel	$326,215.00	$64,519.00	$123,074.00	$160,861.00	$13,516.00	$30,965.00
	Retail	$138,506.00	$174,625.00	$310,200.00	$29,271.00	$159,795.00	$23,541.00
⊟Other	Hotel	$2,928,269.00	$735,753.00	$820,101.00	$771,606.00	$165,990.00	$320,358.00
	Retail	$1,032,308.00	$1,153,006.00	$1,675,150.00	$158,886.00	$724,420.00	$191,752.00



Region	Channel	Sum of spend
⊟Lisbon	Hotel	$1,538,342.00
	Retail	$848,471.00
⊟Oporto	Hotel	$719,150.00
	Retail	$835,938.00
⊟Other	Hotel	$5,742,077.00
	Retail	$4,935,522.00

Student Data Survey

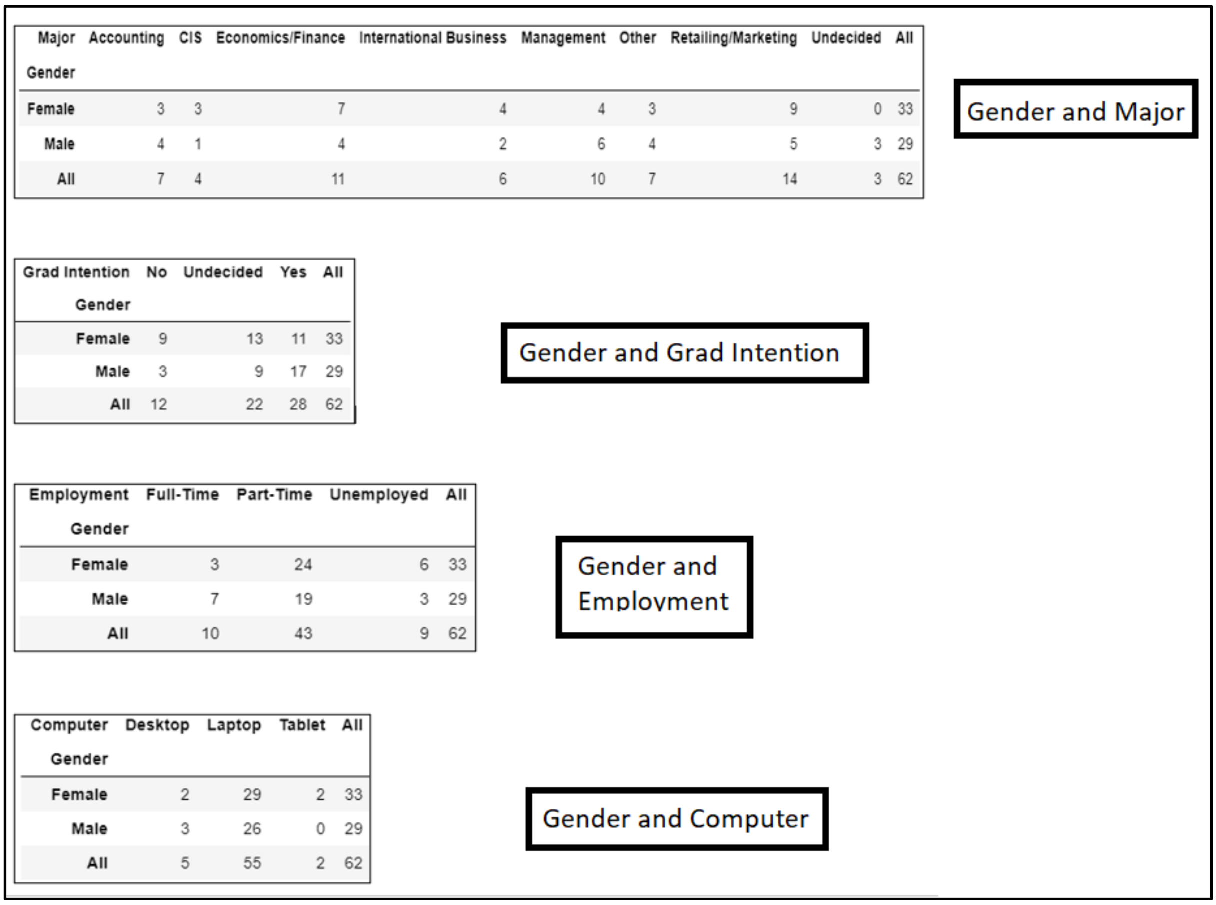
The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. CMSU creates and distributes a survey of 14 questions and receives responses from 62 undergraduates (stored in the *Survey* data set).

The dataset consists of 14 variables where Id is used as an index or primary key, Gender, Class, Major, Employment, Grad Intention, Computer are categorical variables , Age and GPA are quantities with continuous values , Social Networking and Satisfaction being ordinal attributes and other variables like Salary, Spending, Text Messages being numerical. The dataset is based on a survey of 14 questions where 62 students (29 males and 33 females) participated

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

For constructing a contingency table, **we use the crosstab function** that is provided by **pandas**.

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



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

Gender and Major

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

Gender and Grad Intention

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

Gender and Employment

Computer	Desktop	Laptop	Tablet	All
Gender				
Female	2	29	2	33
Male	3	26	0	29
All	5	55	2	62

Gender and Computer

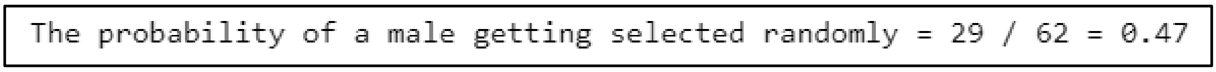
So, these tables will help us in doing analysis based on conditional probabilities.

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?

To answer the above questions, we do need to consider the number of males and females that participated in the survey. Using **count function on Gender**, we got the values, so to calculate the probability, we use the formula: -

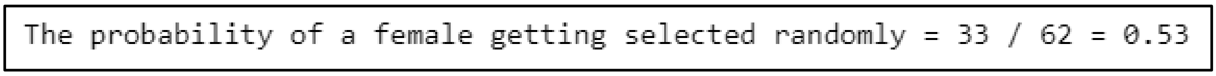
$$P(A) = \frac{\text{Number of favorable outcomes to A}}{\text{Total number of outcomes}}$$

Hence, for a **male getting selected**: -



```
The probability of a male getting selected randomly = 29 / 62 = 0.47
```

For a **female getting selected**: -



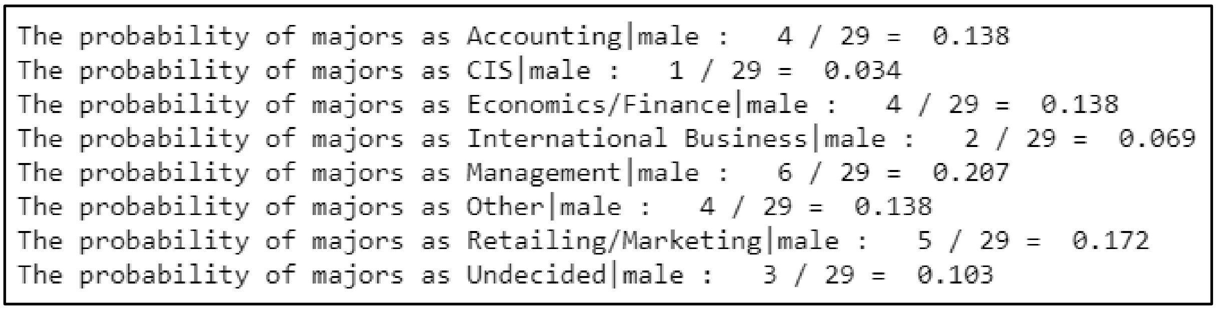
```
The probability of a female getting selected randomly = 33 / 62 = 0.53
```

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.

To calculate, conditional probability, we use the following formula: -

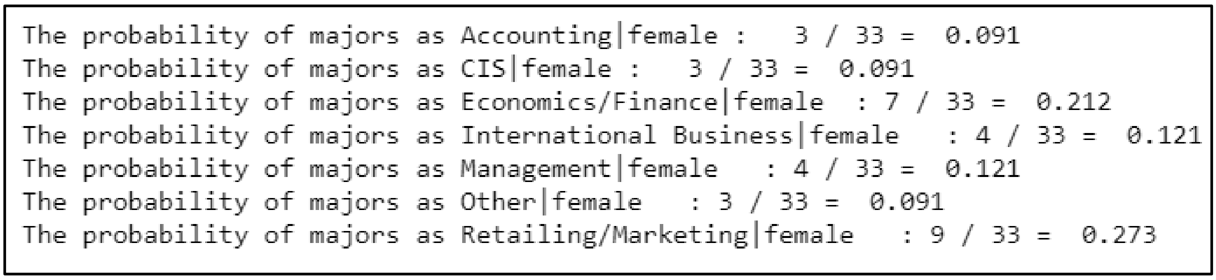
$$\underset{\text{Probability of event A given B has occured}}{P(A \mid B)} = \frac{\overset{\text{Probability of event A occured and event B occured}}{P(A \cap B)}}{\underset{\text{Probability of event B}}{P(B)}}$$

So, to answer the first question, **we select the number of males opting for that majors and divide by total number of males**, the distribution is given as: -



```
The probability of majors as Accounting|male :   4 / 29 =  0.138
The probability of majors as CIS|male :   1 / 29 =  0.034
The probability of majors as Economics/Finance|male :   4 / 29 =  0.138
The probability of majors as International Business|male :   2 / 29 =  0.069
The probability of majors as Management|male :   6 / 29 =  0.207
The probability of majors as Other|male :   4 / 29 =  0.138
The probability of majors as Retailing/Marketing|male :   5 / 29 =  0.172
The probability of majors as Undecided|male :   3 / 29 =  0.103
```

Similarly, for the second portion of the question, **we select the number of females opting for that majors and divide by total number of females**, the distribution is given as: -

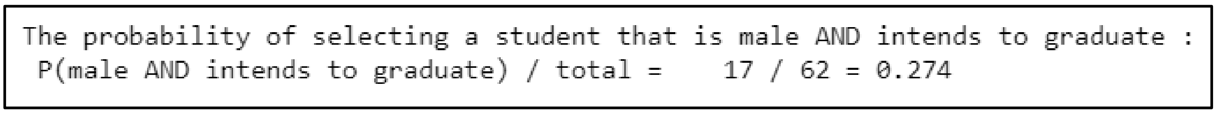


```
The probability of majors as Accounting|female :   3 / 33 =  0.091
The probability of majors as CIS|female :   3 / 33 =  0.091
The probability of majors as Economics/Finance|female  : 7 / 33 =  0.212
The probability of majors as International Business|female   : 4 / 33 =  0.121
The probability of majors as Management|female   : 4 / 33 =  0.121
The probability of majors as Other|female   : 3 / 33 =  0.091
The probability of majors as Retailing/Marketing|female   : 9 / 33 =  0.273
```

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.

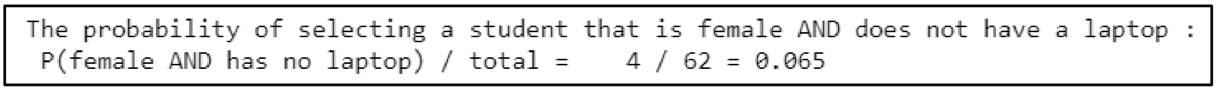
For this set of questions, **our denominator would be total students** as the probability is conditional and a random student will be selected from the entire set.

To answer the first part, we need to compute/count **the number of possible outcomes where a student is a male and intents to graduate(yes)**, so from the contingency table in Q2.1, we can see that there are **17** such possible outcomes.



```
The probability of selecting a student that is male AND intends to graduate :
 P(male AND intends to graduate) / total =    17 / 62 = 0.274
```

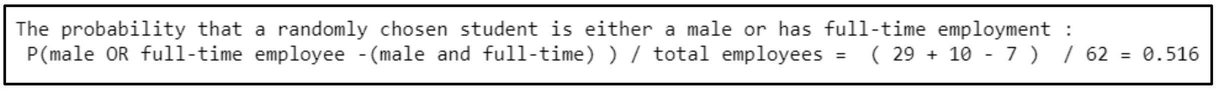
Similarly, in the second part, we need to consider **the students that are female and doesn’t have a laptop**, from the contingency table we can count number of females with laptop and subtract from total females i.e 33-29 =**4**



```
The probability of selecting a student that is female AND does not have a laptop :
 P(female AND has no laptop) / total =    4 / 62 = 0.065
```

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?

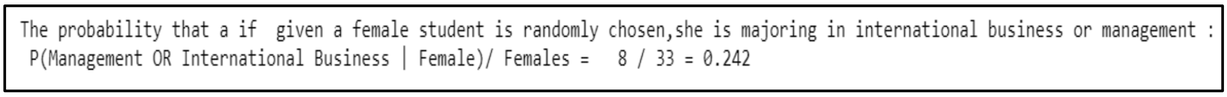
Here, we need to first count the number of possible outcomes i.e (Male OR Full-Time Employment), so we can do it applying the set formula i.e ∑ **(Male and Full-Time Employment) – (Males And Full-Time Employment)** = 29 + 10 – 7 = **32** possible outcomes .



```
The probability that a randomly chosen student is either a male or has full-time employment :
 P(male OR full-time employee -(male and full-time) ) / total employees =  ( 29 + 10 - 7 )  / 62 = 0.516
```

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

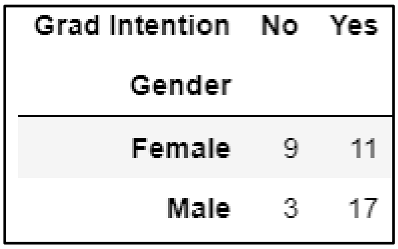
To count the total possible outcomes, we ∑ **students with majors in International Business and Management provided that she is a female** i.e (International Business OR Management | Female) : 4+4=**8** possible outcomes.



```
The probability that a if  given a female student is randomly chosen,she is majoring in international business or management :
 P(Management OR International Business | Female)/ Females =   8 / 33 = 0.242
```

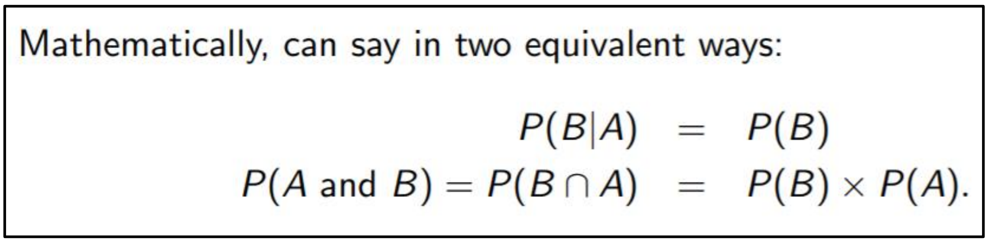
2.6. Construct a contingency table of Gender and Intent to Graduate at 2 levels (Yes/No). The Undecided students are not considered now, and the table is a 2x2 table. Do you think the graduate intention and being female are independent events?

For creating a new contingency table with males and females who have an intent to either graduate or not and eliminating the undecided candidates, we make a new dataframe where **we select students(gender) and grad intention and remove the values of grad intention as "undecided"** and then, create a contingency table using the **crosstab function** from the new dataframe.



Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

To answer the second portion of the question, i.e checking **if grad intention (Yes) and being females are independent events**, we check if they follow the rule: -

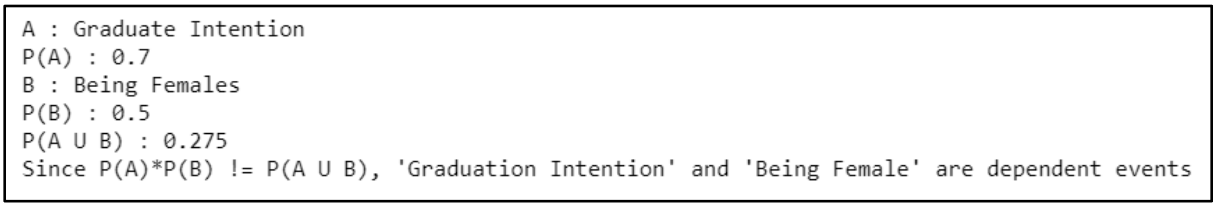


Mathematically, can say in two equivalent ways:

$$P(B|A) = P(B)$$

$$P(A \text{ and } B) = P(B \cap A) = P(B) \times P(A).$$

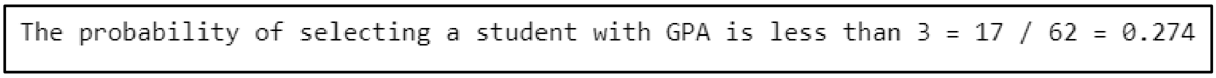
We have used the second method to verify if they are independent events,



```
A : Graduate Intention
P(A) : 0.7
B : Being Females
P(B) : 0.5
P(A U B) : 0.275
Since P(A)*P(B) != P(A U B), 'Graduation Intention' and 'Being Female' are dependent 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?

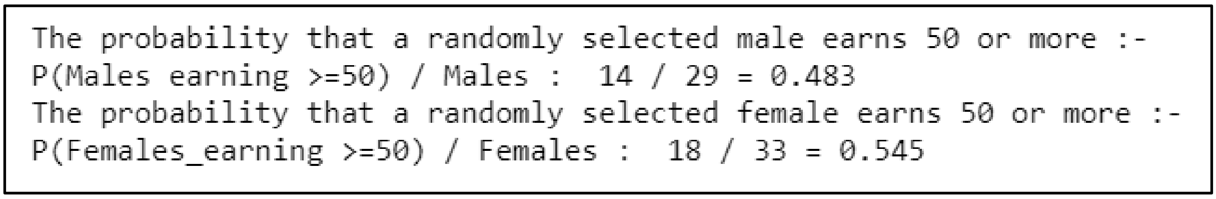
To get a count of number of students with GPA < 3, **we created a dataFrame from the existing dataFrame giving the condition as [GPA< 3.0]**. Now we can **take the number of rows in the newer dataFrame as our numerator using the shape[0]** and **the number of rows in the original dataFrame as the total count (denominator).**



```
The probability of selecting a student with GPA is less than 3 = 17 / 62 = 0.274
```

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.

Similar to the first part of the question 2.7.1, we created dataFrames specifying two conditions i.e **[Salary >= 50 & Gender =Male]** for the first portion of this question and **[Salary >= 50 & Gender =Female]** for the second portion. We used the **shape [0] to count the number of rows** i.e the numerator.



```
The probability that a randomly selected male earns 50 or more :-
P(Males earning >=50) / Males :  14 / 29 = 0.483
The probability that a randomly selected female earns 50 or more :-
P(Females_earning >=50) / Females :  18 / 33 = 0.545
```

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.

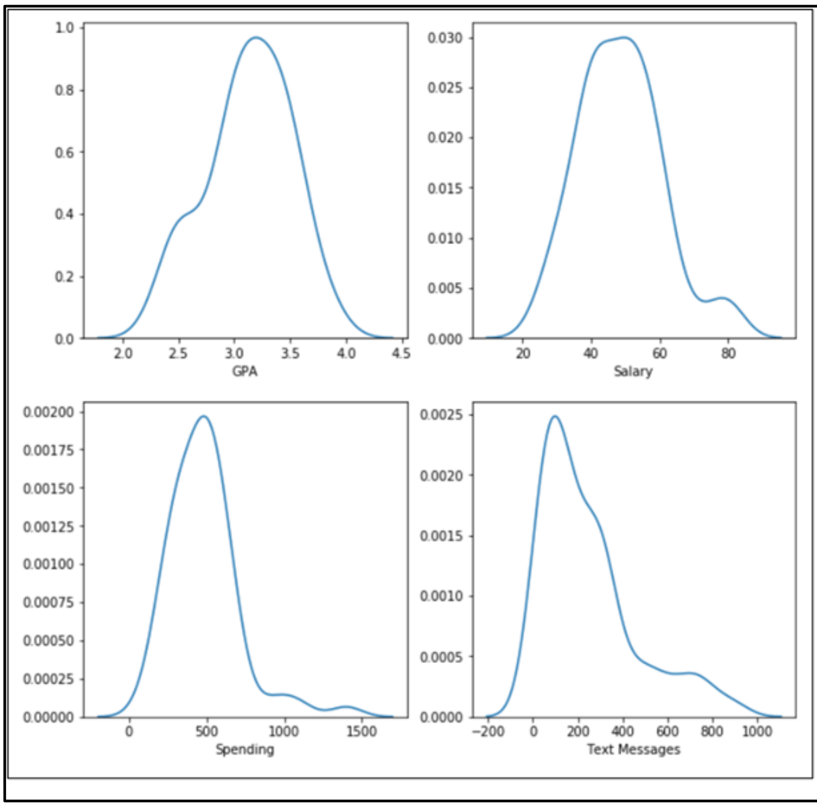
To check the distribution of variables, we used **the distplot() and set histogram=False to plot the the gaussian kernel density estimate, this shows us the curve of distribution**. It gives us an estimate about the **distribution being normal or skewed**.

Secondly, we plotted distplot() with the histogram and lines demarking area under **1 SD, 2 SD and 3 SD i.e 68%, 95% and 99.7%.** This would confirm our observations as the empirical rule for normal distribution says that if 68% of your observations will fall between one standard deviation of the mean, 95% will fall within two, and 99.7% will fall within three, it is a normal distribution.

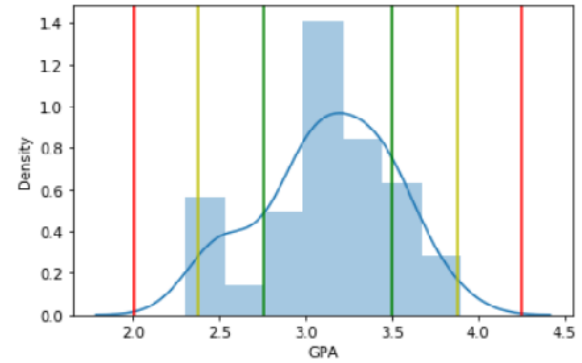
Also, if **mean=median=mode, the data is normally distributed; if mean< median, the data is left skewed and if mean> median, the data is right skewed**.

Following were our observations supported by graphs and numbers: -

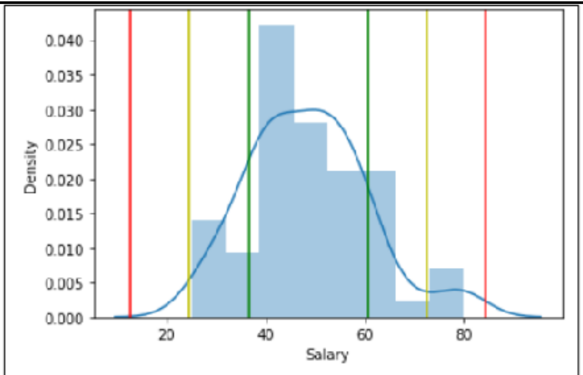
- **GPA** has a distribution that is similar to being normally distributed and maybe if we take more samples and apply central limit theorem, we can confirm the claim.
- **Salary** has a distribution that is similar to normal with most of data being in the centre but we observe some data in the right tail i.e 3rd SD which doesn't satisfy the empirical rule .
- **Spending** is left skewed since mean< median, but has more outliers at the right tail and **Text Messages** is right skewed since mean>median



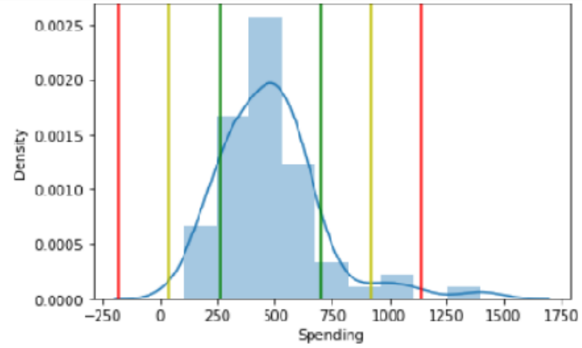
KDE plot



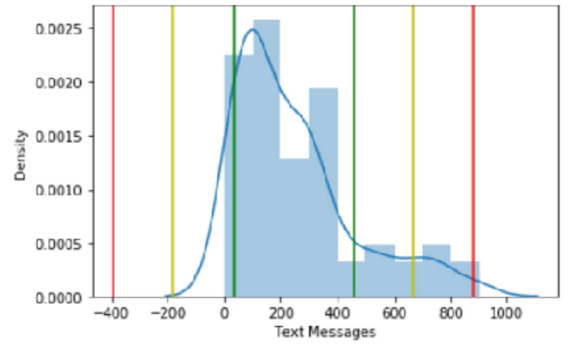
```
Mean of GPA : 3.129
Mode of GPA : [(3.0,), (3.1,), (3.4,)]
Median of GPA : 3.15
```



```
Mean of Salary : 48.548
Mode of Salary : [(40.0,)]
Median of Salary : 50.0
```



```
Mean of Spending : 482.016
Mode of Spending : [(500,)]
Median of Spending : 500.0
```



```
Mean of Text Messages : 246.21
Mode of Text Messages : [(300,)]
Median of Text Messages : 200.0
```

Asphalt Shilling Hypothesis and Test

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 are calculated. The company would like to show that the mean moisture content is less than 0.35 pound per 100 square feet.

The data 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 mean moisture contents in both types of shingles are within the permissible limits? State your conclusions clearly showing all steps.

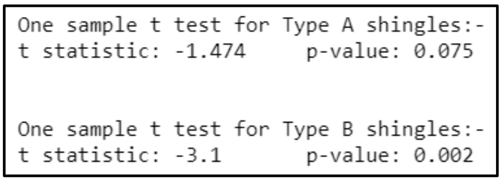
To check the claim, we will form hypothesis and do an appropriate test for the same. Also, we need to perform the test twice i.e Type A and Type B shingles differently. That is why we create two different dataFrames for data in Type A and data in Type B.
This is a 5-step process: -

Step 1: **Formulate the Hypothesis** (for both the cases- A and B**).**
HO: mean <= 0.35
HA: mean >0.35

Step 2: **Decide the α level** (confidence level)
α = 0.05 i.e confidence level is 5%

Step 3: **Decide the Test Statistic**
Since the standard deviation isn't specified, and we have to test both the samples individually, we will opt for the one-sample T-test (one-tailed)

Step 4: **Calculate the T-stat and p-value**
Python has built-in packages that enable us the test functionalities, we use **ttest_1samp** from scipy.stats, and then calculate the t-statistic value and p-value for both the types A and B. We have to use **omit='nan'** for the dataframe containing shingles of type B as there are 5 NaN values.
Also, python by default, python calculates p-value for 2-tailed using this function, so **we need to divide the p-value by 2 to get p-value for 1-tail test**
Following are the values obtained: -



```
One sample t test for Type A shingles:-
t statistic: -1.474     p-value: 0.075


One sample t test for Type B shingles:-
t statistic: -3.1       p-value: 0.002
```

Step 5: **Decide whether to accept or reject null hypothesis**
The criteria to decide the hypothesis is given as: **-** If the p-value is less than the confidence value (α value), we reject the null hypothesis, otherwise accept the null hypothesis.
In our case, the **p-value for Type A shingles is 0.075** which is **greater than α**=0.05, Hence we **accept the null hypothesis** i.e population mean moisture content is less than 0.35 pound per 100 square feet.
Whereas, the **p-value for Type B shingles is 0.002** which is **lesser than α=0.05**, Hence we **reject the null hypothesis and accept the alternative hypothesis** i.e population mean 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?

To check the claim, we will form hypothesis and do an appropriate test for the same. Here, unlike the test performed above, we will not need to perform the test twice as we are comparing the mean of shingles in Type A and Type B. We can directly use the 2 sample T-Test which is 2-tailed.

This is a 5-step process: -

Step 1: **Formulate the Hypothesis** (for both the cases- A and B)**.**

HO: mean (shingles A) = mean (shingles B)

HA: mean (shingles A) != mean (shingles B)

Step 2**: Decide the α level** (confidence level)

α = 0.05 i.e confidence level is 5%

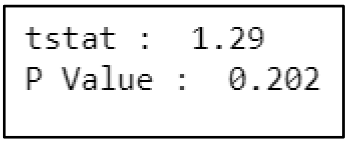
Step 3: **Decide the Test Statistic**

Since the standard deviation isn’t specified, and we are comparing both the samples, we will opt for the **2-sample T-test (two-tailed)**

Step 4: **Calculate the T-stat and p-value**

Python has built-in packages that enable us the test functionalities, we use **ttest_ind** from scipy.stats, and then calculate the t-statistic value and p-value for both the types A and B. We have to use **omit=’nan’** as shingles of type B have 5 NaN values.

Following are the values obtained: -



```
tstat :  1.29
P Value :  0.202
```

Step 5: **Decide whether to accept or reject null hypothesis**

The criteria to decide the hypothesis is given as: - If the p-value is less than the confidence value (α value), we reject the null hypothesis, otherwise accept the null hypothesis.

Since the **p-value we obtained (0.202) is greater than the α value (0.05)**, **we accept the H0** i.e the population mean for shingles A and B are equal