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 seems to spend more? Which Region and which Channel seems to spend less?

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

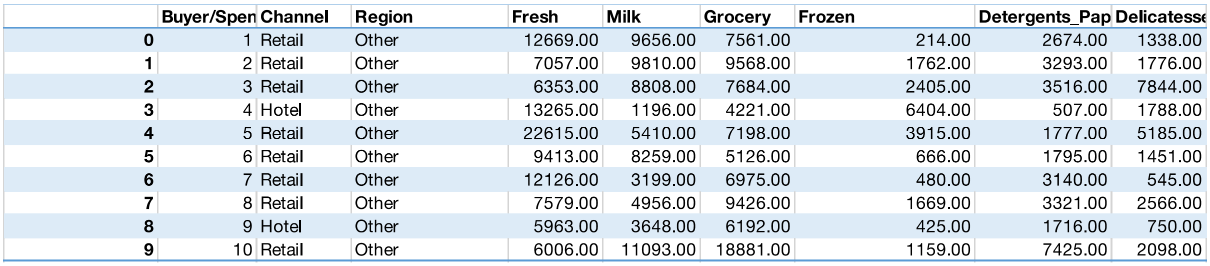
1.3 Based on a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?

1.4 Are there any outliers in the data?

1.5 based on this report, what are the recommendations?

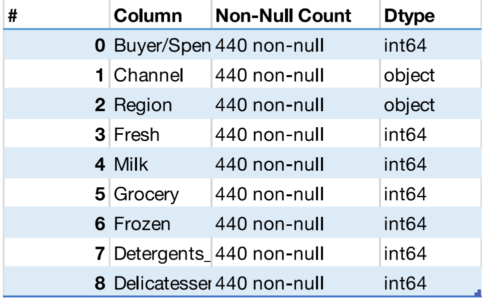
Exploratory Data Analysis:

The first 10 rows are shown below:



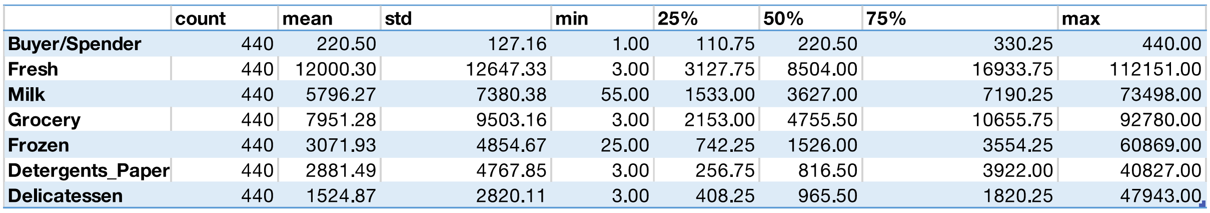
	Buyer/Spen	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Pap	Delicatesse
0	1	Retail	Other	12669.00	9656.00	7561.00	214.00	2674.00	1338.00
1	2	Retail	Other	7057.00	9810.00	9568.00	1762.00	3293.00	1776.00
2	3	Retail	Other	6353.00	8808.00	7684.00	2405.00	3516.00	7844.00
3	4	Hotel	Other	13265.00	1196.00	4221.00	6404.00	507.00	1788.00
4	5	Retail	Other	22615.00	5410.00	7198.00	3915.00	1777.00	5185.00
5	6	Retail	Other	9413.00	8259.00	5126.00	666.00	1795.00	1451.00
6	7	Retail	Other	12126.00	3199.00	6975.00	480.00	3140.00	545.00
7	8	Retail	Other	7579.00	4956.00	9426.00	1669.00	3321.00	2566.00
8	9	Hotel	Other	5963.00	3648.00	6192.00	425.00	1716.00	750.00
9	10	Retail	Other	6006.00	11093.00	18881.00	1159.00	7425.00	2098.00

The data has 440 rows and 9 columns of data. From the bellow we can say that there are no Null values in the data set. Columns, 'Channel' and 'Regional' are categorical while the remaining are all continuous variables and with int datatype.



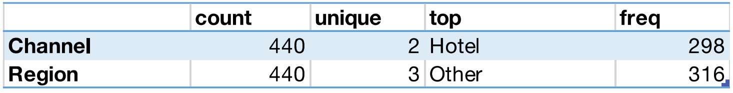
#	Column	Non-Null Count	Dtype
0	Buyer/Spen	440 non-null	int64
1	Channel	440 non-null	object
2	Region	440 non-null	object
3	Fresh	440 non-null	int64
4	Milk	440 non-null	int64
5	Grocery	440 non-null	int64
6	Frozen	440 non-null	int64
7	Detergents_	440 non-null	int64
8	Delicatesser	440 non-null	int64

The five number summary for all the continuous variables are shown below:



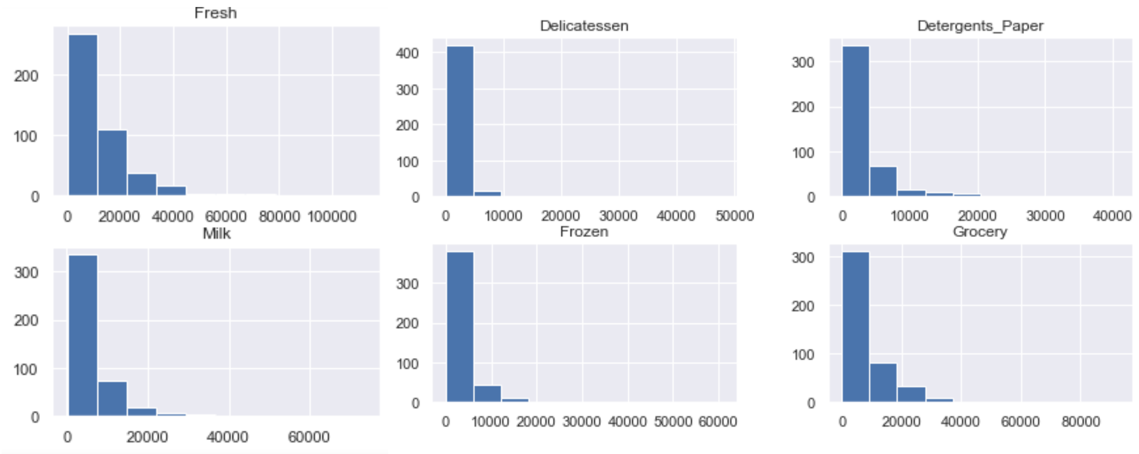
	count	mean	std	min	25%	50%	75%	max
Buyer/Spender	440	220.50	127.16	1.00	110.75	220.50	330.25	440.00
Fresh	440	12000.30	12647.33	3.00	3127.75	8504.00	16933.75	112151.00
Milk	440	5796.27	7380.38	55.00	1533.00	3627.00	7190.25	73498.00
Grocery	440	7951.28	9503.16	3.00	2153.00	4755.50	10655.75	92780.00
Frozen	440	3071.93	4854.67	25.00	742.25	1526.00	3554.25	60869.00
Detergents_Paper	440	2881.49	4767.85	3.00	256.75	816.50	3922.00	40827.00
Delicatessen	440	1524.87	2820.11	3.00	408.25	965.50	1820.25	47943.00

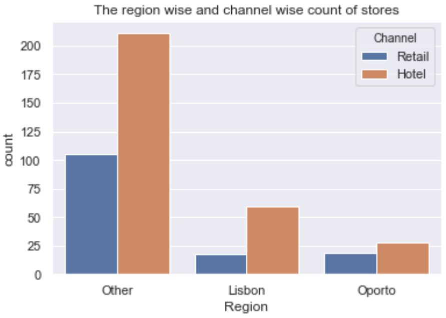
The summary for categorical variables are shown below:



	count	unique	top	freq
Channel	440	2	Hotel	298
Region	440	3	Other	316

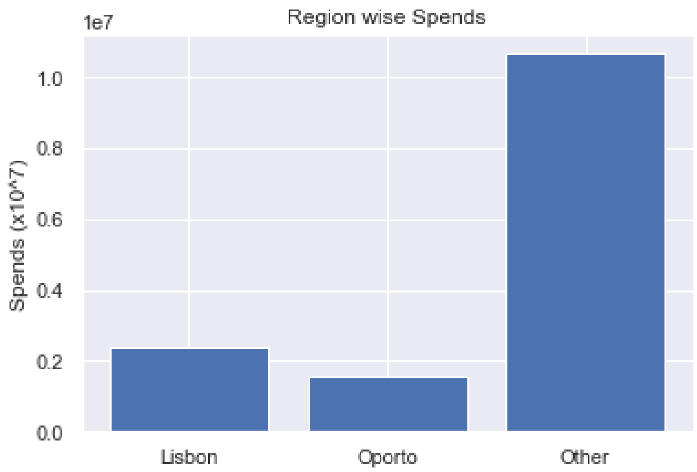
1. When compared to other items, Fresh, Milk and Grocery items have a comparatively high standard deviation.
2. 'Channel' has 2 unique items ('Hotel' and 'Retail'), amongst which 'Hotel' occurs maximum number of times with a frequency of 298.
3. 'Region' has 3 unique items('Other', 'Lisbon' and 'Oporto'), amongst which 'Others' occurs maximum times with a frequency of 316.
4. All the items seemed to be Right Skewed or Positive Skewed as mean is greater than the median. This can be confirmed by the histogram plotted below.
5. Items have high concentration on the far left/lower side of the distribution and the remaining right/higher sides are sparsely distributed.
6. There are higher number of hotels in each region than retails.

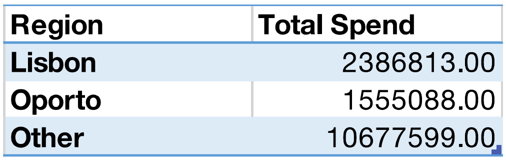




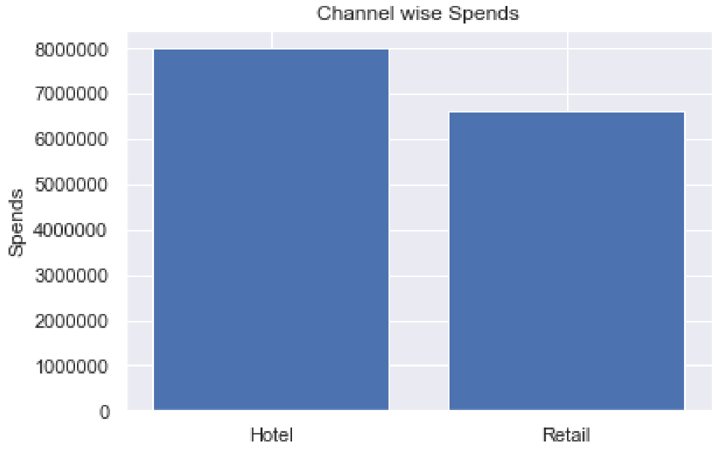
Question 1:

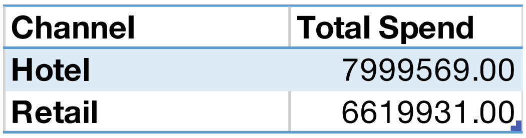
Which Region and which Channel seems to spend more? Which Region and which Channel seems to spend less?





Region	Total Spend
Lisbon	2386813.00
Oporto	1555088.00
Other	10677599.00



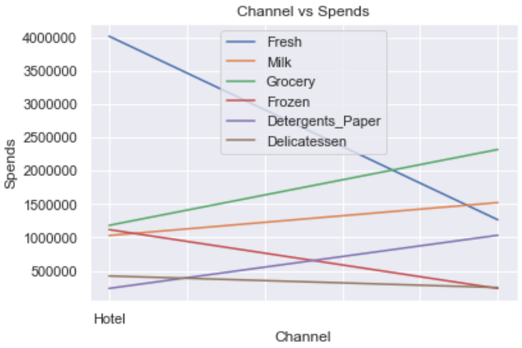


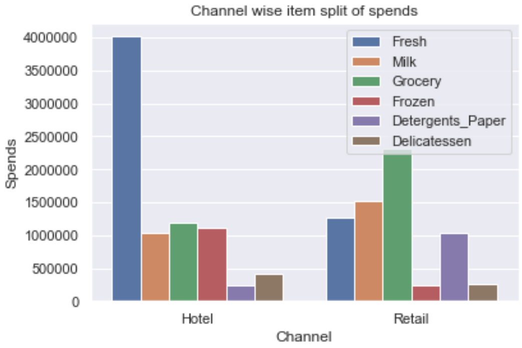
Channel	Total Spend
Hotel	7999569.00
Retail	6619931.00

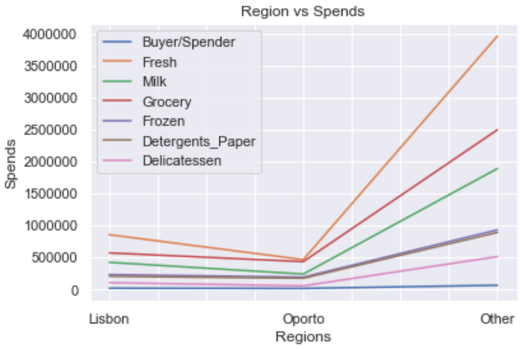
Answer:
‘Other’ region spends the maximum (i.e. 10677599).
‘Oporto’ region spends the minimum (i.e. 1555088).
‘Hotel’ channel spends the maximum (i.e. 7999569).
‘Retail’ channel spends the minimum (i.e. 6619931).

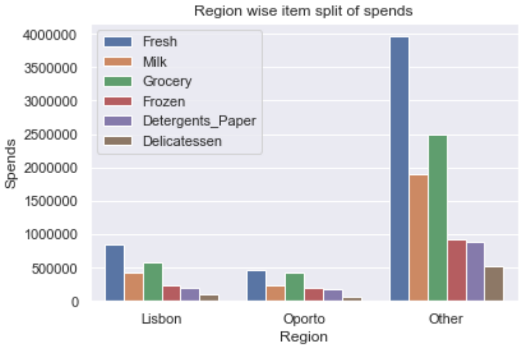
Question 2:

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

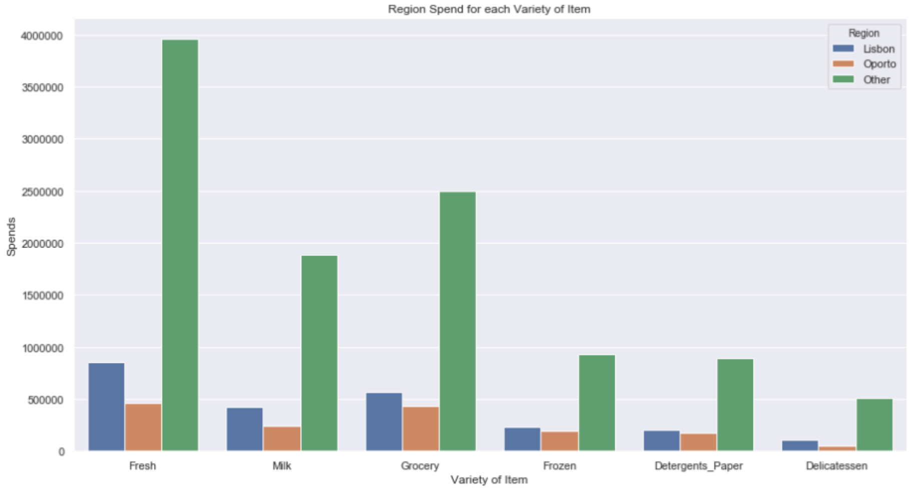


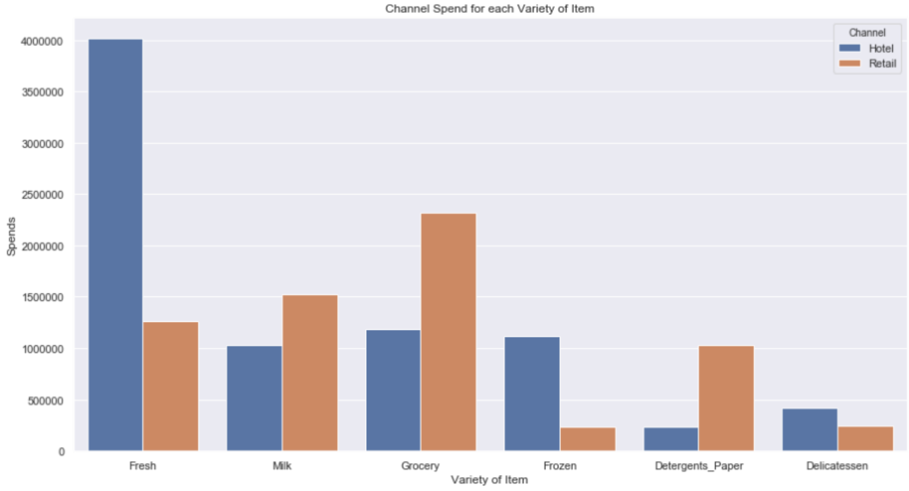




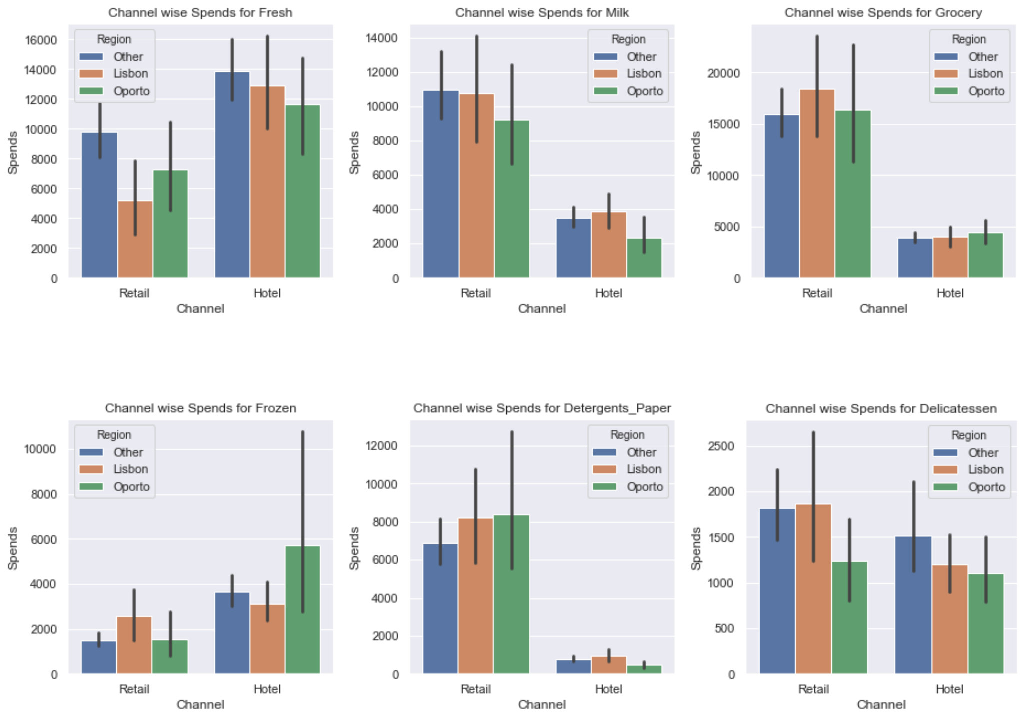


Channel wise and Region wise (Actual Spends):





Channel Wise (Average):


Channel wise Spends for Fresh

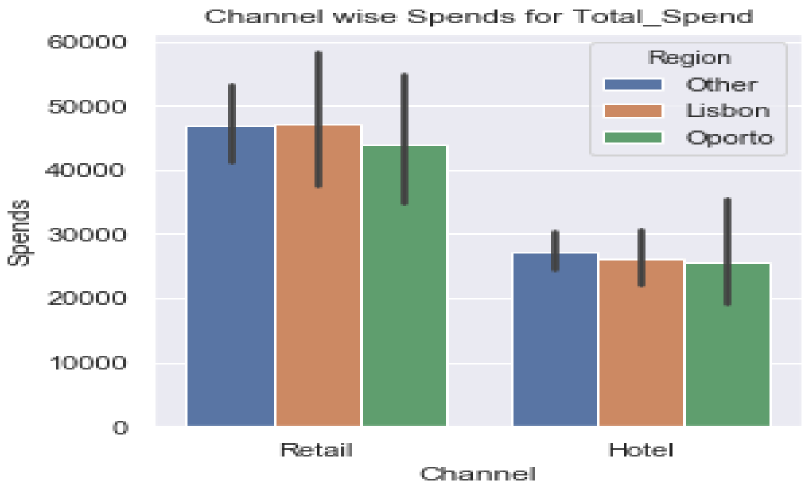
Channel wise Spends for Milk

Channel wise Spends for Grocery

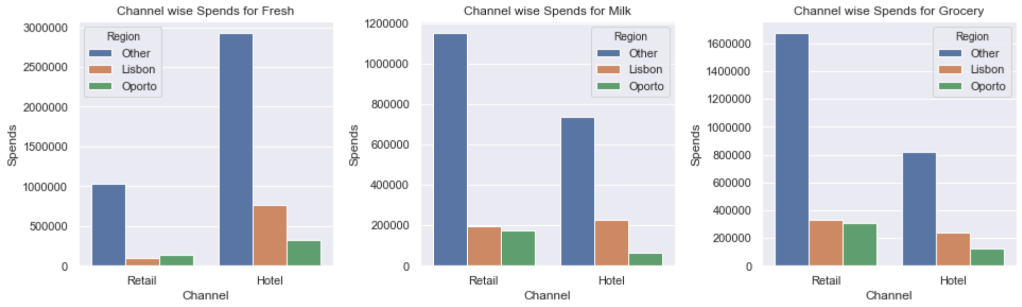
Channel wise Spends for Frozen

Channel wise Spends for Detergents_Paper

Channel wise Spends for Delicatessen

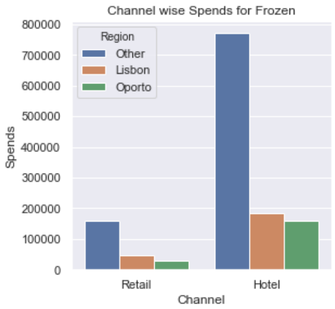

Channel wise Spends for Total_Spend

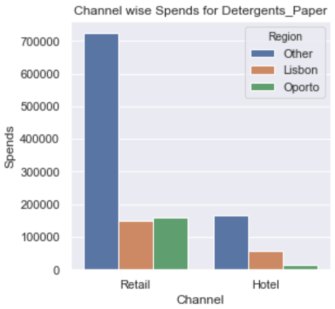
Channel Wise (Actual Spends):

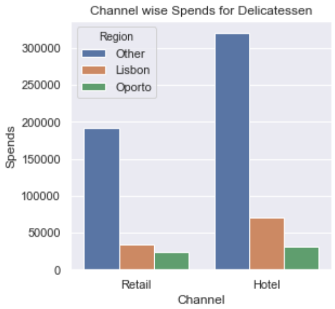

Channel wise Spends for Fresh

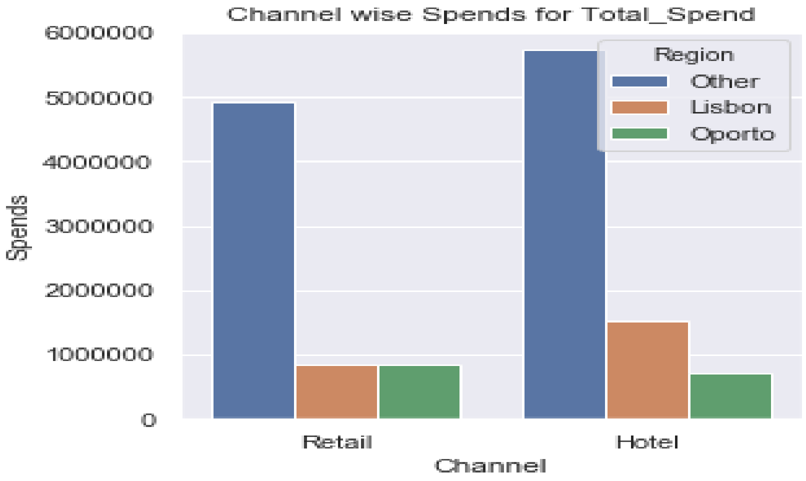
Channel wise Spends for Milk

Channel wise Spends for Grocery

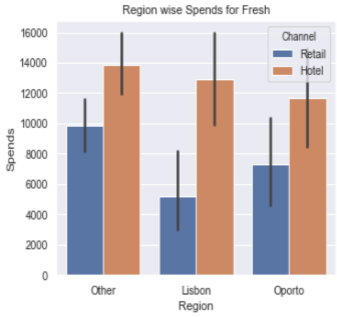

Channel wise Spends for Frozen

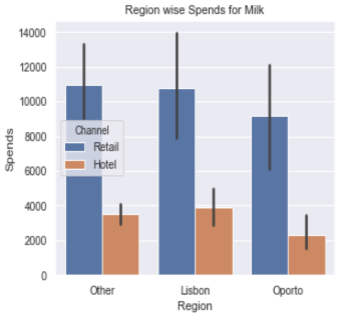

Channel wise Spends for Detergents_Paper

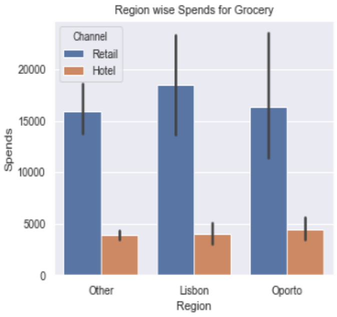

Channel wise Spends for Delicatessen

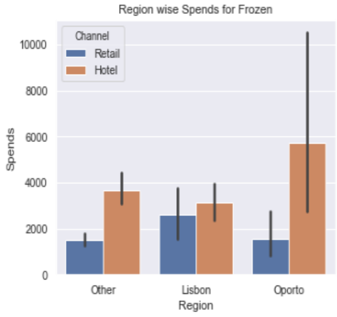

Channel wise Spends for Total_Spend

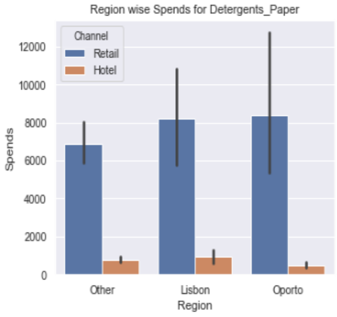
Region Wise (Average):

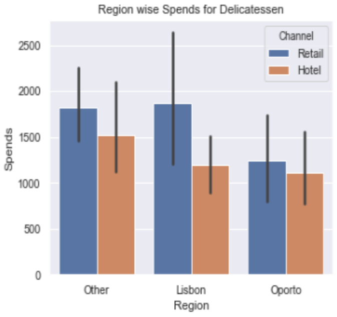

Region wise Spends for Fresh

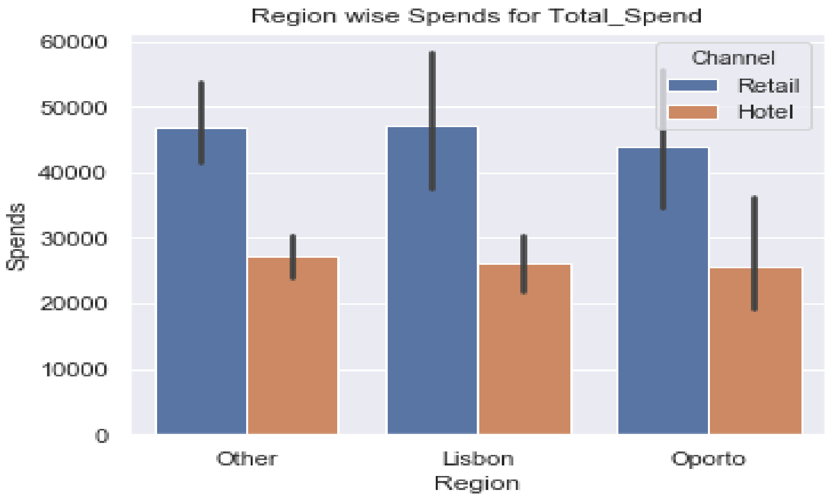

Region wise Spends for Milk


Region wise Spends for Grocery

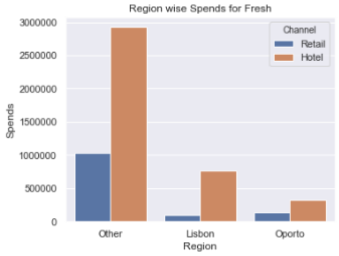

Region wise Spends for Frozen

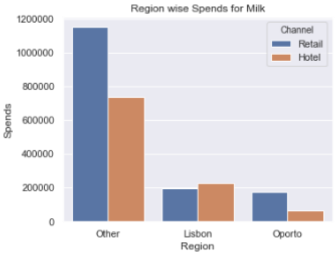

Region wise Spends for Detergents_Paper

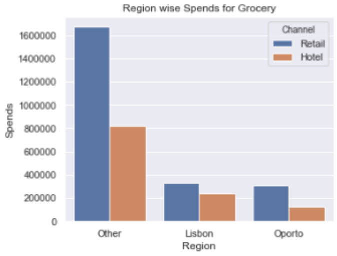

Region wise Spends for Delicatessen

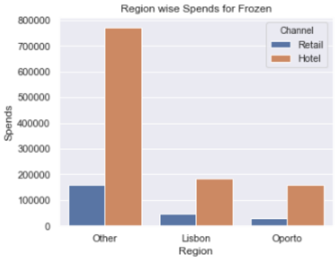


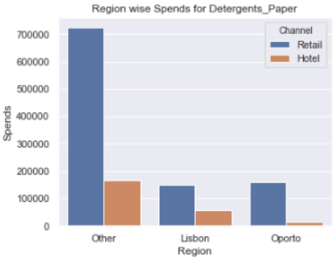
Region Wise (Actual Spends):

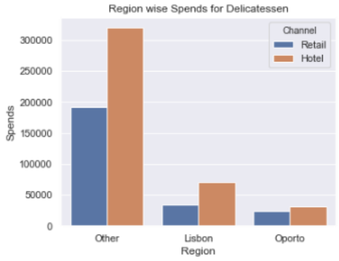


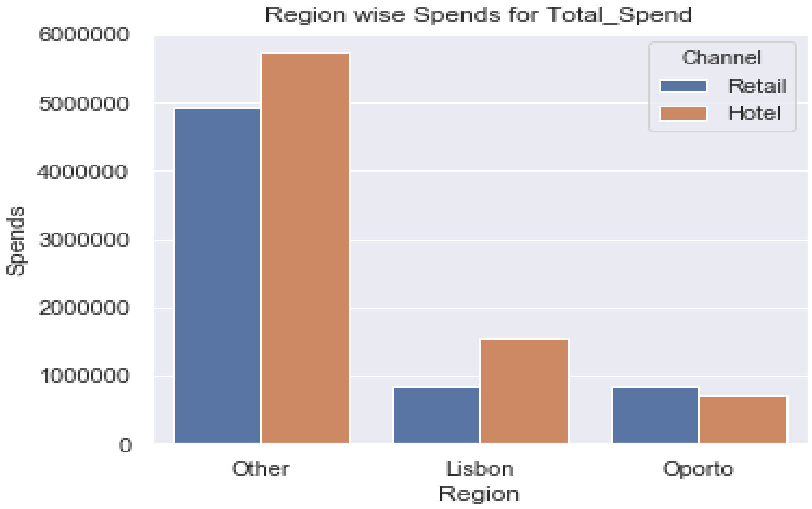












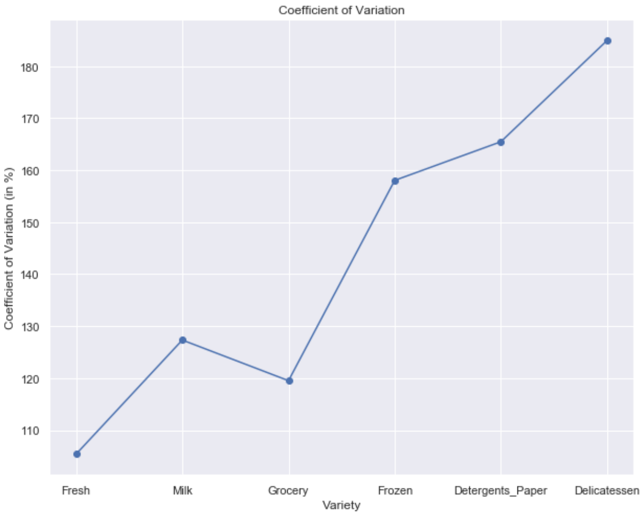
Answer:

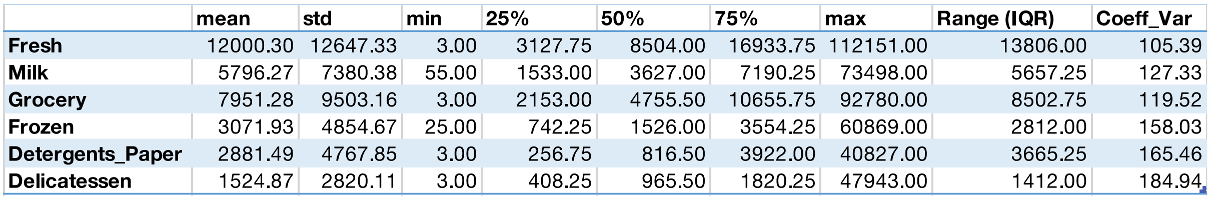
1. Form the first 2 graphs
   - It can be observed that all 6 varieties of items follow a similar spend pattern across the regions. The pattern observed is ascending order of spending Oporto, Lisbon and then Others.
   - No particular pattern of spend can be observed across channels.
2. From the Channel wise graphs:
   - The Other Region have the highest investment in all the item varieties. If average across channels is considered, the retail channel in Libson and the hotel channel in Other have the highest spends.
   - The average spend in Lisbon Retail is almost equal to spend in Other Retail. Similarly, Lisbon Hotel and Oporto Hotel are almost equal.
   - There is a an anomaly spotted between the actual and the average spends for the following combinations:
     - Milk item, Other region and Hotel channel
     - Grocery item, Other region and Retail channel
     - Fresh item, Other region and Retail channel
     - Fresh item, Oporto region and Retail channel
     - Detergent item, Other region and Retail & Hotel channel
     - Delicatessen item, Other region and Retail & Hotel channel

     This could be because there is higher spends and lower counts or vice versa.
3. From the Region wise Graphs:
   - On an average out of the 6 varieties, Fresh and Frozen have greater Hotel spends while the others have greater Retail spends.
   - On an overall average, the retail has significantly higher spends than the hotel channel in all regions.
   - Though the Delicatessen item has more actuals spends in Hotel channel but the average turns out that the Retail has more spends. This implies that there could be a more number of small spends in Hotel compared to retail or few number of high spends in Retails compared to hotels.
4. When compared to other items, Fresh, Milk and Grocery items have a comparatively high variation in data.
5. All the items seemed to be Right Skewed or Positive Skewed as mean is greater than the median. This can be confirmed by the histogram plotted below.
6. Items have high concentration on the far left/lower side of the distribution and the remaining right/higher sides are sparsely distributed.

Question 3:

Based on a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?





	mean	std	min	25%	50%	75%	max	Range (IQR)	Coeff_Var
Fresh	12000.30	12647.33	3.00	3127.75	8504.00	16933.75	112151.00	13806.00	105.39
Milk	5796.27	7380.38	55.00	1533.00	3627.00	7190.25	73498.00	5657.25	127.33
Grocery	7951.28	9503.16	3.00	2153.00	4755.50	10655.75	92780.00	8502.75	119.52
Frozen	3071.93	4854.67	25.00	742.25	1526.00	3554.25	60869.00	2812.00	158.03
Detergents_Paper	2881.49	4767.85	3.00	256.75	816.50	3922.00	40827.00	3665.25	165.46
Delicatessen	1524.87	2820.11	3.00	408.25	965.50	1820.25	47943.00	1412.00	184.94

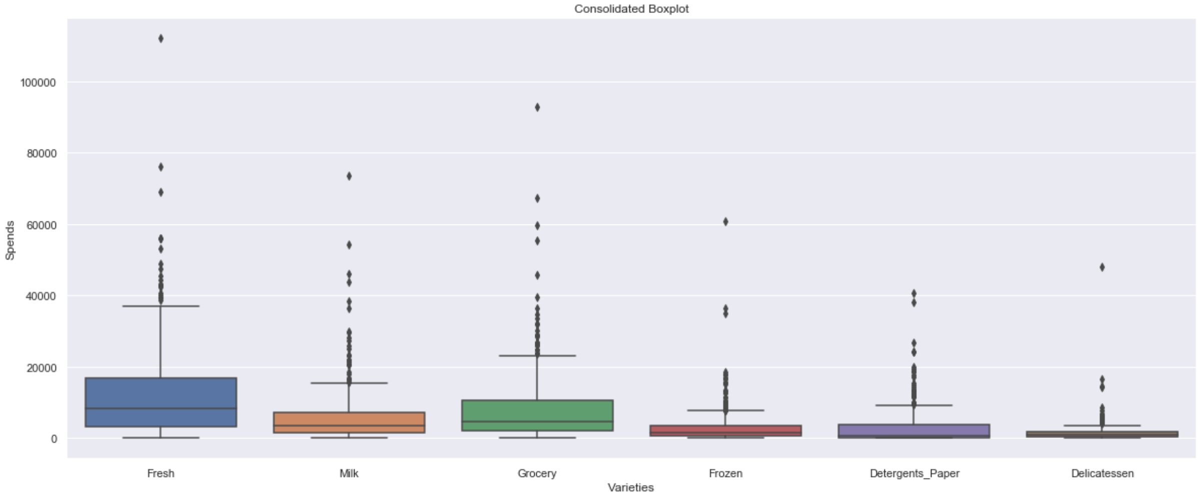
Answer:

From the five number summary it can be observed that the range of values are very high there could be probable outliers so we will calculate the Inter-Quartile Range. The Standard deviation is highest for Fresh variety and lowest for Delicatessen. But the standard deviation does not give the comparative variation so we calculate the Coefficient of Variation.

By plotting the coefficient of variation for the varieties and it can be observed that Delicatessen has the highest inconsistency in data and Fresh has the lowest inconsistency in data.

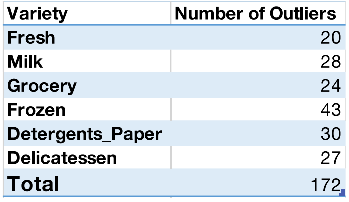
Question 4:

Are there any outliers in the data?



Answer:

All the varieties have outliers. The number of outliers are given in the table below:



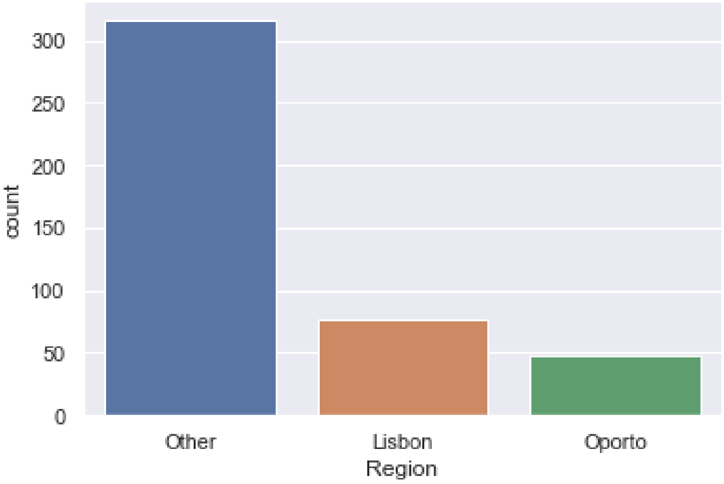
Variety	Number of Outliers
Fresh	20
Milk	28
Grocery	24
Frozen	43
Detergents_Paper	30
Delicatessen	27
Total	172

Question 5:

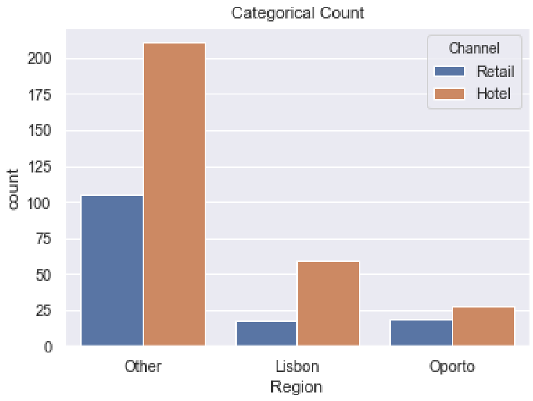
Based on this report, what are the recommendations?

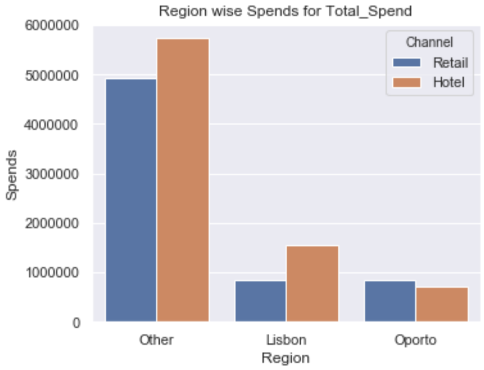
Answer:

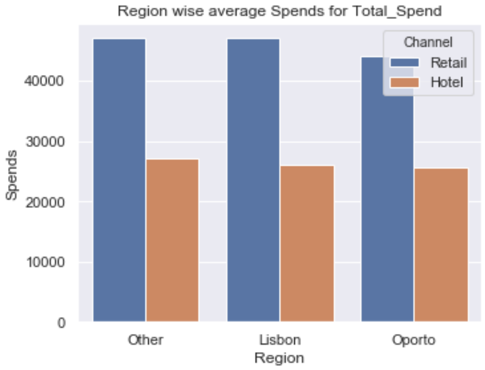
1. Form the below it can be observed that Other Region has a high market share. This will keep the company at high risk as there is no proportional distribution amongst the regions and any closure of the stores at Other will result in huge loss. Hence, it is recommended to venture further more into the Lisbon and Oporto.



2. Form the below it can be observed that the number of hotels are more that the number of retails. Proportionally the Total Spend is also proportional (except for the Oporto Region). However the average is reverse. The average when plotted shows that the retail spends more. The possible reason for this could be that they are small spends by many hotels or large spends by few retails. The recommendation is to revert back to the strategies followed in the other verities provided there is very little business impact.







3. From the average Region wise Spends Graphs it can be observed that across regions a constant average for the product is maintained. In that view,
   - Fresh variety in Libson and Oporto can increase the spends
   - Milk variety in Oporto can increase the spends
   - Frozen variety in Other and Lisbon can increase their spends in the Hotel channel
   - Delicatessen in Oporto needs to increase the Retail Spends considerably and the Hotel Spends slightly. The Lisbon, Retail can also be increased slightly.
   - The Total Spends is almost balanced except for the Retail Channel in the Oporto Region. The above fixes (specially Delicatessen) will set this imbalance right.
4. From Channel wise and Region wise (Actual Spends) graph, it can be said that Milk, Frozen and Detergent_Paper have moderate spends and this can be increased further more. Milk and Frozen might have over heads and short shelf life but Detergent_Paper may not incur a lot of over heads and there is not much concern of shelf life, making it a good focus for spending more.
5. In addition to the above Delicatessen has very little spends and can be increased further more. Since there is a huge variation in spends it can be observed that there are a good number of stores spending in huge quantity and the remaining spend very less. This could be because they are dedicated stores for Delicatessen (spending more) and small stores just trying the product (spending less). If this is a case of dedicated stores we will have to find out more stores like so or if the stores are still trying the product, we will have to advertise and market the products to get spends increased.

Survey

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).

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 either 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

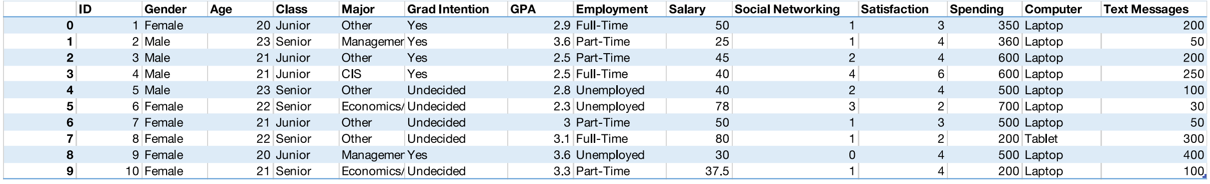
2.6.1. If a student is chosen randomly, what is the probability that his/her GPA is less than 3?

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.

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.

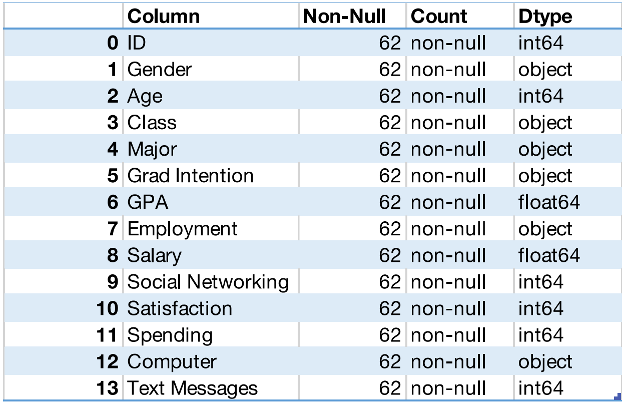
Exploratory Data Analysis:

The first 10 rows are shown below:



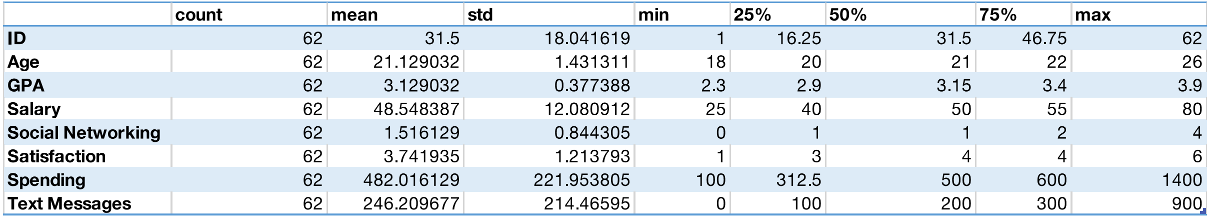
	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
0	1	Female	20	Junior	Other	Yes	2.9	Full-Time	50	1	3	350	Laptop	200
1	2	Male	23	Senior	Managemer	Yes	3.6	Part-Time	25	1	4	360	Laptop	50
2	3	Male	21	Junior	Other	Yes	2.5	Part-Time	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/	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
8	9	Female	20	Junior	Managemer	Yes	3.6	Unemployed	30	0	4	500	Laptop	400
9	10	Female	21	Senior	Economics/	Undecided	3.3	Part-Time	37.5	1	4	200	Laptop	100

- The data has 62 Rows (62 Undergraduates) and 14 Columns (14 Questions).
- From the below we can say that there are no Null values in the data set. ‘Gender’, ’Class’, ’Major’, ’Grad Intention’, ’Employment’, ’Computer’ are categorical variables.
- The columns ‘Social networking’ and ‘Satisfaction’ are numerical but are ordinate scale variables and will be treated as categorical variables.
- The remaining are continuous variables with integer and float type data types.



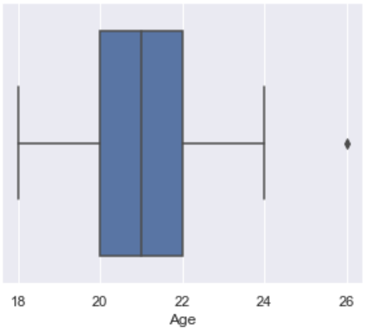
	Column	Non-Null	Count	Dtype
0	ID	62	non-null	int64
1	Gender	62	non-null	object
2	Age	62	non-null	int64
3	Class	62	non-null	object
4	Major	62	non-null	object
5	Grad Intention	62	non-null	object
6	GPA	62	non-null	float64
7	Employment	62	non-null	object
8	Salary	62	non-null	float64
9	Social Networking	62	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

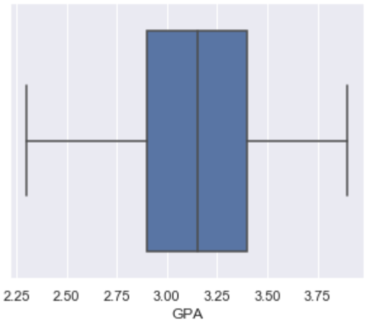
The five number summary for all the continuous variables are shown below:

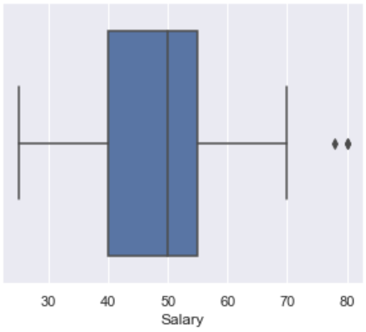


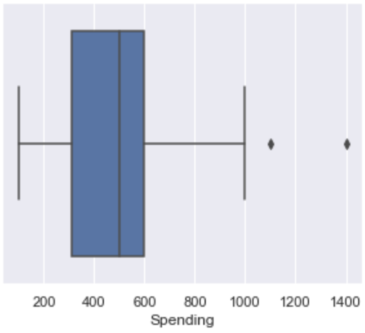
	count	mean	std	min	25%	50%	75%	max
ID	62	31.5	18.041619	1	16.25	31.5	46.75	62
Age	62	21.129032	1.431311	18	20	21	22	26
GPA	62	3.129032	0.377388	2.3	2.9	3.15	3.4	3.9
Salary	62	48.548387	12.080912	25	40	50	55	80
Social Networking	62	1.516129	0.844305	0	1	1	2	4
Satisfaction	62	3.741935	1.213793	1	3	4	4	6
Spending	62	482.016129	221.953805	100	312.5	500	600	1400
Text Messages	62	246.209677	214.46595	0	100	200	300	900

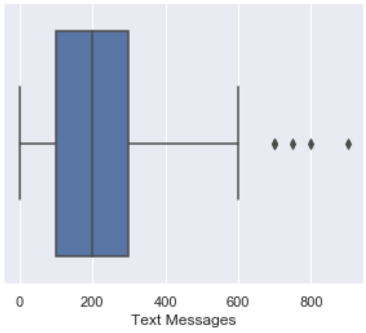
The skewness/distribution of data with outliers:



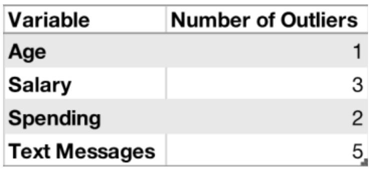






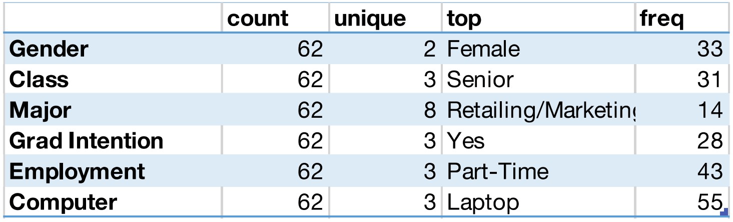


- Distribution of data:
  Age : Almost normal (Right Skewed)
  GPA : Almost normal (Left Skewed)
  Salary : Left Skewed
  Spending : Left Skewed
  Text Messages : Right Skewed
- All the variables have outliers except GPA. The number of outliers in each variable is shown below:



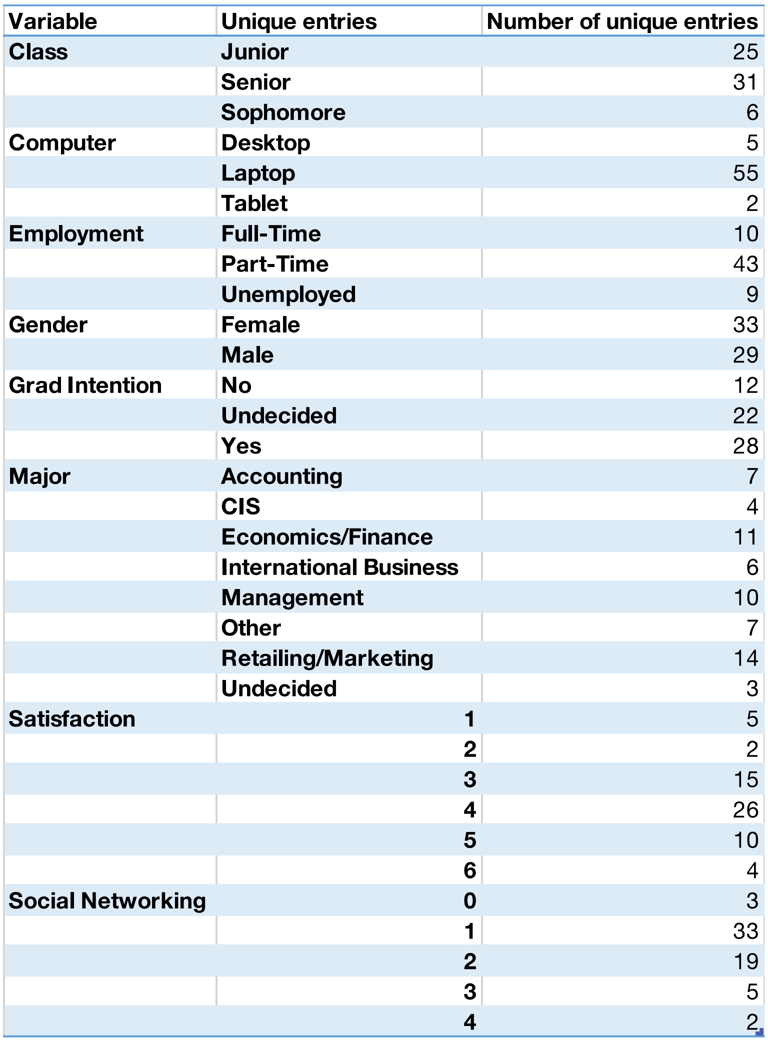
Variable	Number of Outliers
Age	1
Salary	3
Spending	2
Text Messages	5

The summary for categorical variables are shown below:

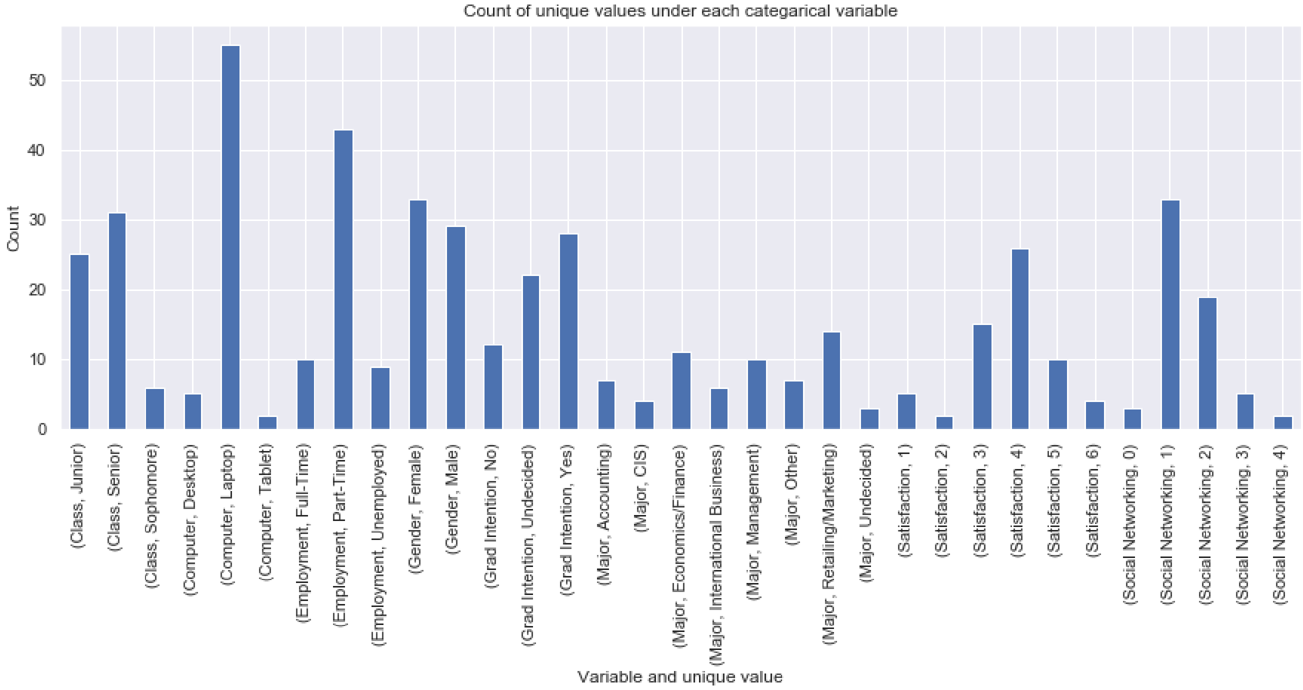


	count	unique	top	freq
Gender	62	2	Female	33
Class	62	3	Senior	31
Major	62	8	Retailing/Marketin	14
Grad Intention	62	3	Yes	28
Employment	62	3	Part-Time	43
Computer	62	3	Laptop	55

- The above picture gives rough idea about the variable and the entry within the variable that has the highest frequency and the corresponding frequency as well.
- The list of variables and the unique items within them is listed below:



Variable	Unique entries	Number of unique entries
Class	Junior	25
	Senior	31
	Sophomore	6
Computer	Desktop	5
	Laptop	55
	Tablet	2
Employment	Full-Time	10
	Part-Time	43
	Unemployed	9
Gender	Female	33
	Male	29
Grad Intention	No	12
	Undecided	22
	Yes	28
Major	Accounting	7
	CIS	4
	Economics/Finance	11
	International Business	6
	Management	10
	Other	7
	Retailing/Marketing	14
	Undecided	3
Satisfaction	1	5
	2	2
	3	15
	4	26
	5	10
	6	4
Social Networking	0	3
	1	33
	2	19
	3	5
	4	2



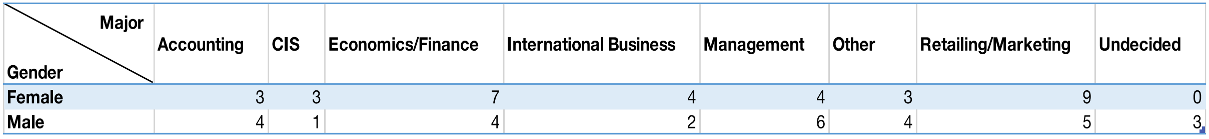
Question 1:

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

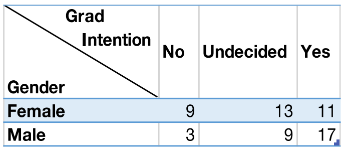
Answer:

2.1.1. Gender and Major



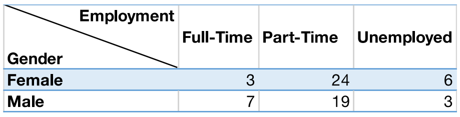
Major Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Female	3	3	7	4	4	3	9	0
Male	4	1	4	2	6	4	5	3

2.1.2. Gender and Grad Intention



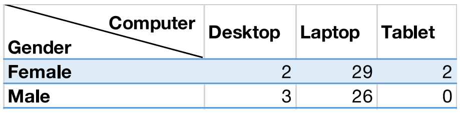
Grad Intention Gender	No	Undecided	Yes
Female	9	13	11
Male	3	9	17

2.1.3. Gender and Employment



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

2.1.4. Gender and Computer



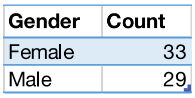
Computer Gender	Desktop	Laptop	Tablet
Female	2	29	2
Male	3	26	0

Question 2:

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?

Answer:



Gender	Count
Female	33
Male	29

Total Population: 62

2.2.1. P(Male) = $({}^{29}/_{62})$ * 100 = 46. 77%

The probability that a randomly selected CMSU student will be male is **46. 77%**

2.2.2 P(Female) = $({}^{33}/_{62})$ * 100 = 53.23%

The probability that a randomly selected CMSU student will be female is **53.23%**

Question 3:

2.3. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

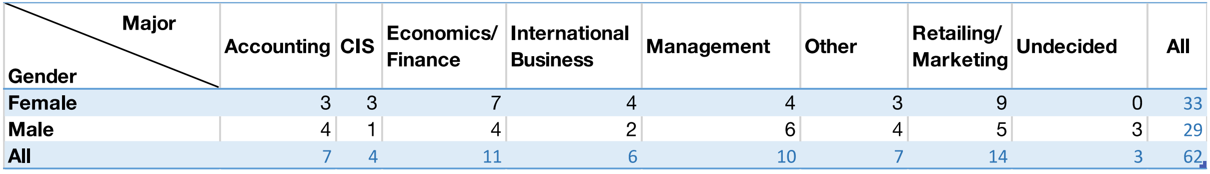
2.3.1. Find the conditional probability of different majors among the male students in CMSU.

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

Answer:

Since here the probability of majors for each gender is asked. i.e. the event of choosing the gender has already occurred hence we apply conditional probability.

Conditional Probability is P(B/A), Probability of B when A has already occurred



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

2.3.1. P($^{\text{Major}}/_{\text{Male}}$) = ($^{\text{Number of males for that major}}/_{\text{Number of Males}}$) * 100

- The probability of **Accounting** amongst the male students in CMSU is **13.79%**
- The probability of **CIS** amongst the male students in CMSU is **3.45%**
- The probability of **Economics/Finance** amongst the male students in CMSU is **13.79%**
- The probability of **International Business** amongst the male students in CMSU is **6.9%**
- The probability of **Management** amongst the male students in CMSU is **20.69%**
- The probability of **Other** amongst the male students in CMSU is **13.79%**
- The probability of **Retailing/Marketing** amongst the male students in CMSU is **17.24%**
- The probability of **Undecided** amongst the male students in CMSU is **10.34%**

2.3.2 P($^{\text{Major}}/_{\text{Female}}$) = ($^{\text{Number of females for that major}}/_{\text{Number of Females}}$) * 100

- The probability of **Accounting** amongst the female students in CMSU is **9.09%**
- The probability of **CIS** amongst the female students in CMSU is **9.09%**
- The probability of **Economics/Finance** amongst the female students in CMSU is **21.21%**
- The probability of **International Business** amongst the female students in CMSU is **12.12%**
- The probability of **Management** amongst the female students in CMSU is **12.12%**
- The probability of **Other** amongst the female students in CMSU is **9.09%**
- The probability of **Retailing/Marketing** amongst the female students in CMSU is **27.27%**
- The probability of **Undecided** amongst the female students in CMSU is **0%**

Question 4:

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.

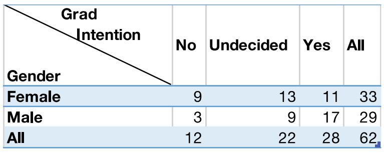
2.4.2 Find the probability that a randomly selected student is a female and does NOT have a laptop.

Answer:

2.4.1. Both (Male and Intend to grad) are dependant variables

P(Male and intend to graduate)

= P(Male)* P$(^{\text{Intend to grad}}/_{\text{Male}})$
= P(Intend to grad)*P$(^{\text{Male}}/_{\text{Intend to grad}})$
=Number of students male and intending to graduating/Total number of students



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

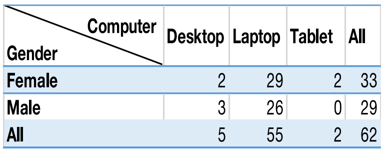
- P(Male) = $(^{29}/_{62})$
- P(Intend to grad/Male) = $(^{17}/_{29})$
- P(Male intending to graduate) = $(^{29}/_{62})$ x $(^{17}/_{29})$ x 100 = 27.41%

The probability that a randomly chosen student is a male and intends to graduate is **27.41%.**

2.4.2 Both (Female and not having laptop) are dependant variables

P(Female not having laptop)

= P(Female)* P$(^{\text{Not having laptop}}/_{\text{Female}})$
= P= P(Not having laptop)*P$(^{\text{Female}}/_{\text{Not having laptop}})$
= Number of female students not having laptop/Total number of students



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

- P(Female) = $(^{33}/_{62})$
- P$(^{\text{Not having laptop}}/_{\text{Female}})$ = $(^{(33-29)}/_{33})$ = $(^{4}/_{33})$
- P(Female not having laptop) = $(^{33}/_{62})$ x $(^{4}/_{33})$ x 100 = 6.45%

The probability that a randomly chosen student is a female not having a laptop is **6.45%.**

Question 5:

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?

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

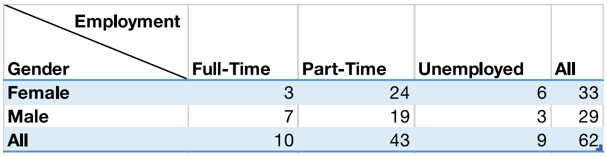
Answer:

2.5.1 Both (Male and full-time employment) are dependant variables

P(Male or full-time employment) (or) = P(Male) + P(Full-time employment) -P(Male and full-time employment)

P(Male) U P(Full-time employment)

= == Number of female students not having laptop/Total number of students



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

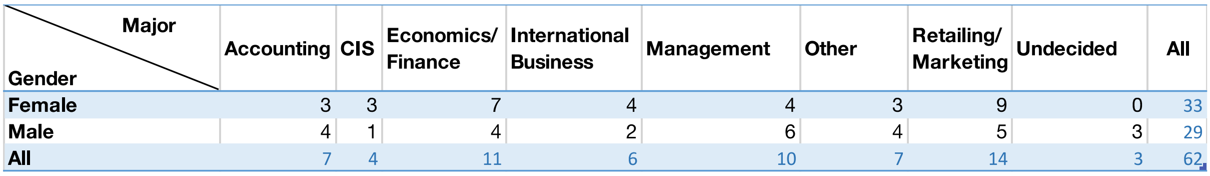
- P(Male) = $(^{29}/_{62})$
- P(Full-time Employment) = $(^{10}/_{62})$
- P(Male and Full-time Employment) = $(^{7}/_{62})$
- P(Male or Full-time Employment) = $(^{(29+10-7)}/_{62})$ * 100 =

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

2.5.2 Both (Female and international business or management) are dependant variables. Here the event of choosing a female student has already occurred and hence the probability is P$(^{\text{International business or management}}/_{\text{Female}})$

P(Female and international business or management)

= == Number of female students whose major is international business or management /Total number of female students



Gender \ Major	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

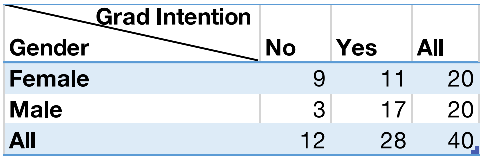
- P$(^{\text{International business or management}}/_{\text{Female}})$ = $(^{4+4}/_{33})$ = 24.24%

The probability that given a female student is randomly chosen, she is majoring in international business or management is **24.24%.**

Question 6:

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?

Answer:



Gender \ Grad Intention	No	Yes	All
Female	9	11	20
Male	3	17	20
All	12	28	40

Graduate intention and being female are independent events if

P(Graduate intention (Yes) and being female)

P(Graduate intention (Yes)) * P(Female)

= P(Graduate intention) * P(Female)
= $(^{28}/_{40})$ * $(^{20}/_{40})$ * 100

= 35%

However when calculated the basic way, i.e. favourable events/total events
Graduate intention and being female are independent events if

P(Graduate intention (Yes) and being female)
(or)

P(Graduate intention (Yes) ∩ Female)

= Number of Female students with graduate intent as yes/Total number of Female students
=$(^{11}/_{20})$
= 55%

Since P(Graduate intention (Yes)) * P(Female) ≠ P(Graduate intention (Yes) ∩ Female, it can be concluded that the two events Graduation intent and Gender are not independent events.

Question 7:

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.

Answer:

2.7.1. P(GPA < 3)

= Number of students scored less than 3 GPA/ Total Number of students
= 17/62
= 27.42%

The probability of the chosen student having GPA less than 3 is **27.42%.**

2.7.2 P$(^{\text{Salary of male} \geq 50}/_{\text{Male}})$

= (Number of males earning ≥ 50)/Total number of males
= $(^{14}/_{29})$
= 48.28%

P$(^{\text{Salary of female} \geq 50}/_{\text{Female}})$

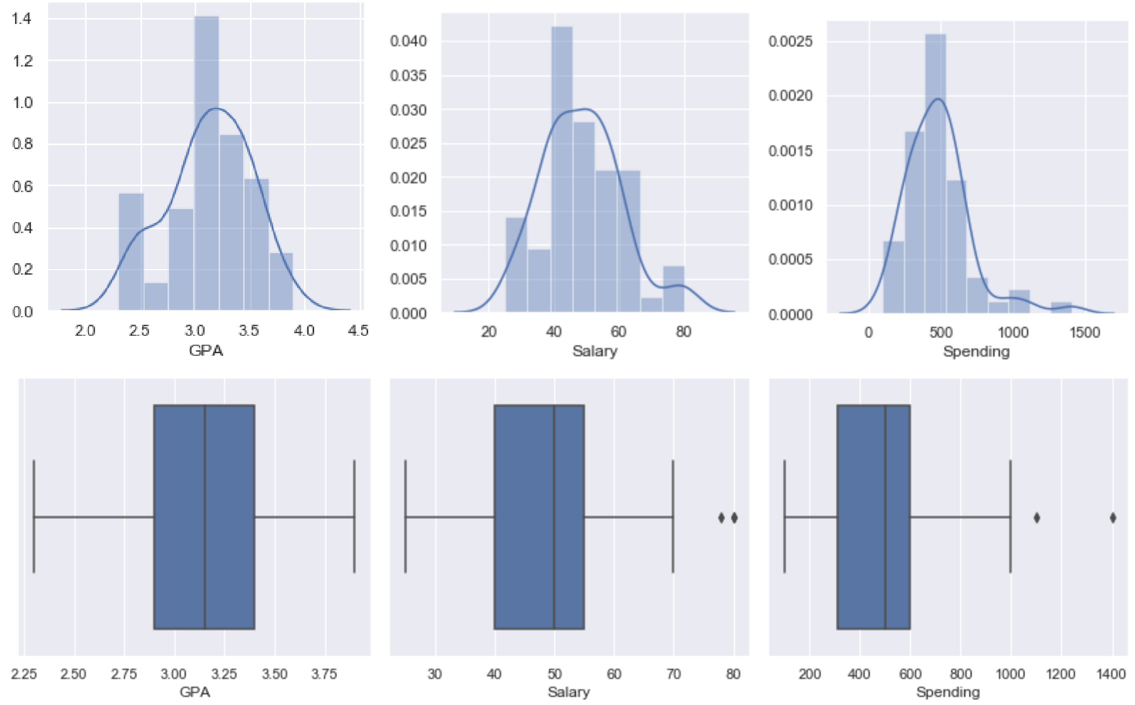
= (Number of females earning ≥ 50)/Total number of males
= $(^{18}/_{33})$
= 54.55%

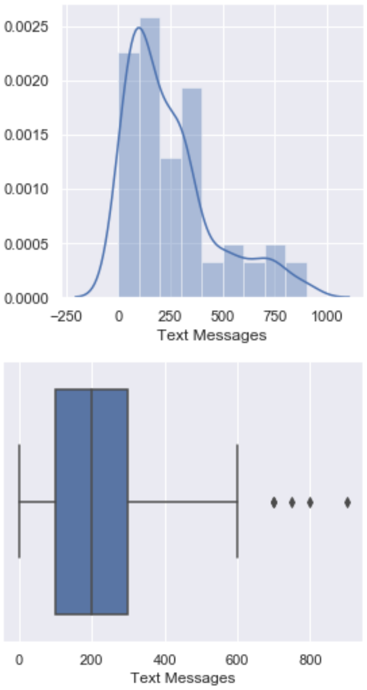
The probability of a selected male earning 50 or more is **48.28%.**
The probability of a selected female earning 50 or more is **54.55%.**

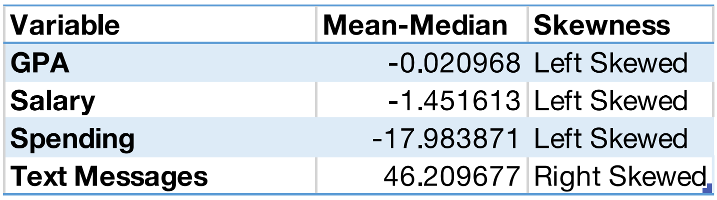
Question 8:

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.

Answer:



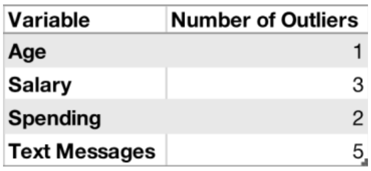




Variable	Mean-Median	Skewness
GPA	-0.020968	Left Skewed
Salary	-1.451613	Left Skewed
Spending	-17.983871	Left Skewed
Text Messages	46.209677	Right Skewed

From the histogram and box plots, GPA and Text Messages seemed to be like a normal distribution. The histogram and boxplots for Salary and Spending seem like left skewed. When the mean is less than median then it is left skewed and if median is greater than median then it is right skewed. From both the above observations the following can be concluded:

- GPA : Almost normal, slightly left skewed
- Salary : Left Skewed
- Spending : Left Skewed
- Text Messages : Almost normal, slightly right skewed
- Except GPA all other variables have outliers. The number of outliers in each are mentioned below:



Variable	Number of Outliers
Age	1
Salary	3
Spending	2
Text Messages	5

Shingles

Problem Statement:

An important quality characteristic used by the manufacturers of ABC asphalt shingles is the amount of moisture the shingles contain when they are packaged. Customers may feel that they have purchased a product lacking in quality if they find moisture and wet shingles inside the packaging.   In some cases, excessive moisture can cause the granules attached to the shingles for texture and colouring purposes to fall off the shingles resulting in appearance problems. To monitor the amount of moisture present, the company conducts moisture tests. A shingle is weighed and then dried. The shingle is then reweighed, and based on the amount of moisture taken out of the product, the pounds of moisture per 100 square feet is calculated. The company would like to show that the mean moisture content is less than 0.35 pound per 100 square feet.

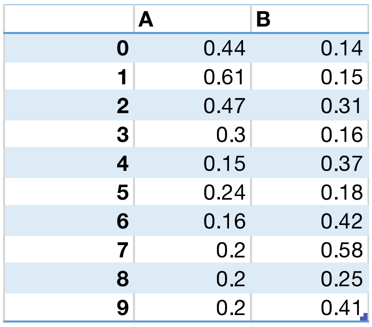
The file (A & B shingles.csv) includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.

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.

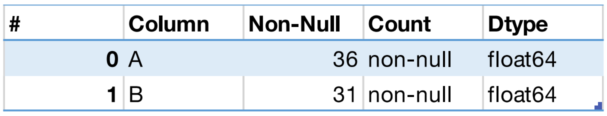
3.2 Do you think that the population mean for shingles A and B are equal? Form the hypothesis and conduct the test of the hypothesis. What assumption do you need to check before the test for equality of means is performed?

Exploratory Data Analysis:

The first 10 rows are shown below:

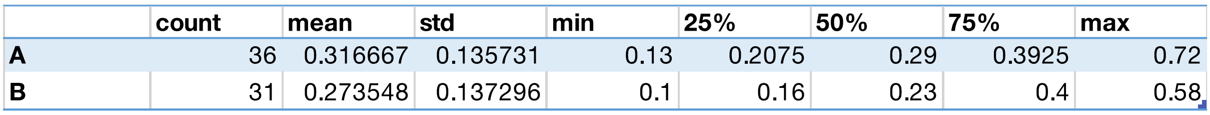


	A	B
0	0.44	0.14
1	0.61	0.15
2	0.47	0.31
3	0.3	0.16
4	0.15	0.37
5	0.24	0.18
6	0.16	0.42
7	0.2	0.58
8	0.2	0.25
9	0.2	0.41

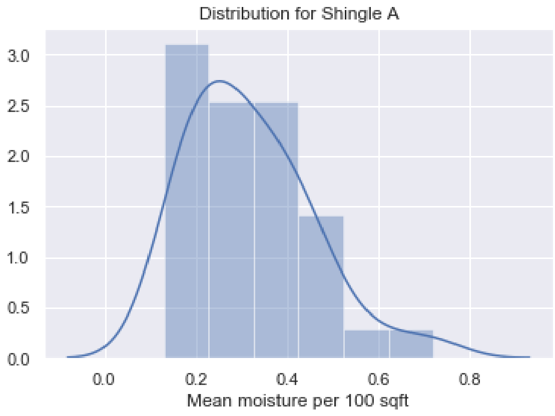


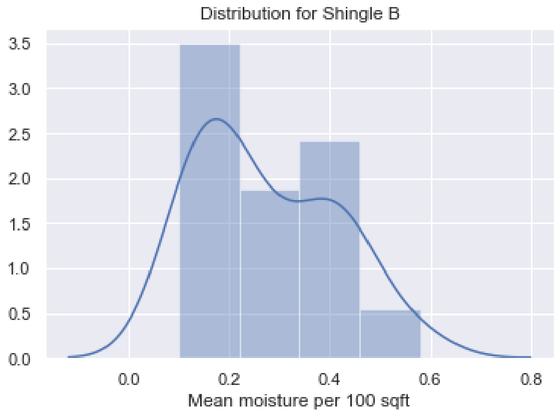
#	Column	Non-Null	Count	Dtype
0	A	36	non-null	float64
1	B	31	non-null	float64

The data has 36 observations for Shingle A and 31 observations for Shingle B.



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





- Shingle A and B is right skewed.

Question 1:

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.

Answer:

Shingle A:

1. Formulating the Null and Alternate Hypothesis
   - H_0: μ ≥ 0.35
   - H_A: μ < 0.35
2. For α = 0.05
3. Since the population deviation is not provided we will proceed with t test (1 sample t test).
4. This is a left tailed t test.
5. The p value calculated is 7.4% which is greater than 5% and hence we fail to reject the Null hypothesis.

Shingle A is **not** within the permissible limits (i.e. not less than 0.35 pounds per 100 SQFT).

Shingle B:

1. Formulating the Null and Alternate Hypothesis
   - H_0: μ ≥ 0.35
   - H_B: μ < 0.35
2. For α = 0.05
3. Since the population deviation is not provided we will proceed with t test (1 sample t test).
4. This is a left tailed t test.
5. The p value calculated is 0.2% which is lesser than 5% and hence we reject the Null hypothesis.

Shingle B is within the permissible limits (i.e. less than 0.35 pounds per 100 SQFT).

Question 2:

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?

Answer:

Assumption:

- Shingle A and B are 2 independent samples of 2 shingles.
- Since the deviation of the population is not given we will proceed with 2 sample t test
- Let us take the confidence interval of 95%. Hence $\alpha = 0.05$
- μ_A is the population mean of Shingle A and μ_B is the population mean of Shingle B.
- Assuming that I want to prove that the population means are not equal. I formulate the below hypothesis.

2 Sample T test:

1. Formulating the Null and Alternate Hypothesis
   - H_0: $\mu_A = \mu_B$
   - H_a: $\mu_A \neq \mu_B$
2. For $\alpha = 0.05$
3. Since the population deviation is not provided we will proceed with t test (2 sample t test).
4. This is a two tailed t test.
5. The p value calculated is 20% which is greater than 5% and hence we fail to reject the Null hypothesis.

The **population mean of both the shingles are equal** at 95% confidence interval.