SMDM PROJECT

Name: Swetha Kunapuli

Batch & Course: PGP-DSBA

Online June Batch

Date: 07/08/2021

Table of Contents

Problem 1

Problem 2

Check for types of

variables in the data frame

Check for missing values in the dataset

2.1. For this data, construct the following contingency tables

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 a 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 a 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.

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

2.5.1 Find the probability that a randomly chosen student is a male or has a 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 graduate intention and being female are independent events?

2.7 Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending and Text Messages. Answer the following questions based on the data

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

2.7.2 Find conditional probability that a randomly selected male earns 50 or more. Find 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.

2.8.1 For each of them comment whether they follow a normal distribution.

2.8.2 Write a note summarizing your conclusions.

Problem 3

Executive Summary

Introduction

Data Description

Sample of the dataset

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.

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

PROBLEM 1

Executive Summary:

The wholesale customer data set refers to clients of a wholesale distributor. It includes the annual spending in monetary units (m.u.) on diverse product categories.

Introduction:

The purpose of this whole exercise is to explore the dataset. This data set is recommended for learning and practicing your skills in exploratory data analysis, data visualization.

The dataset gives data about sales of 6 category of products across 3 regions via 2 channels.

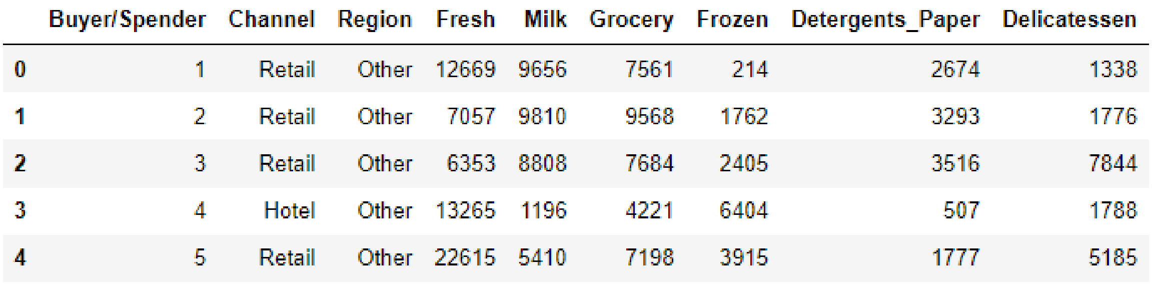
The dataset has total 440 rows and 9 columns.

Data Description:

Description of variables is as follows:

- Fresh: Annual spending (m.u.) on fresh products (Continuous).
- Milk: Annual spending (m.u.) on milk products (Continuous).
- Grocery: Annual spending (m.u.) on grocery products (Continuous).
- Frozen: Annual spending (m.u.) on frozen products (Continuous).
- Detergents_Paper: Annual spending (m.u.) on detergents and paper products (Continuous).
- Delicatessen: Annual spending (m.u.) on and delicatessen products (Continuous).
- Channel: Customers Channel - Hotel (Hotel/Restaurant/Cafe) or Retail channel (Nominal).
- Region: Customers Region Lisbon, Oporto or Other (Nominal).
- Buyer/Spender: It is showing running id number (assumption it is index) (Continuous).

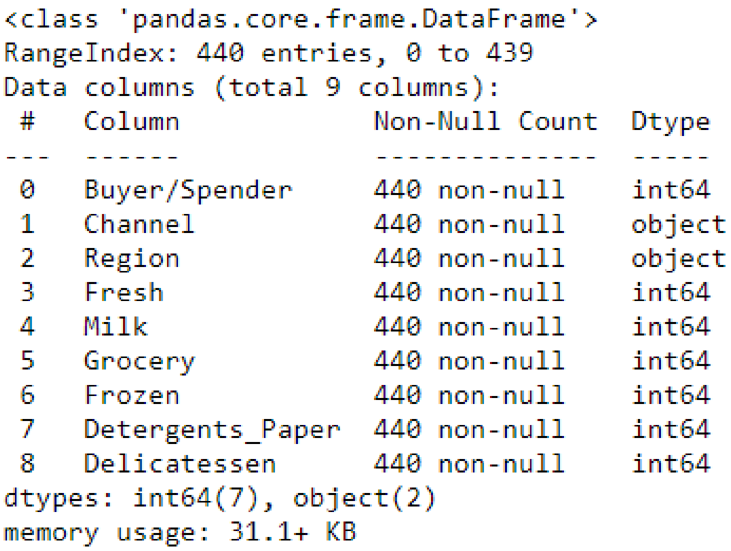
Sample of the dataset:



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

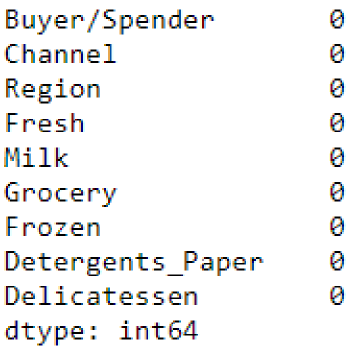
Exploratory Data Analysis:

Check for types of variables in the data frame:



```
<class 'pandas.core.frame.DataFrame'>
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
```

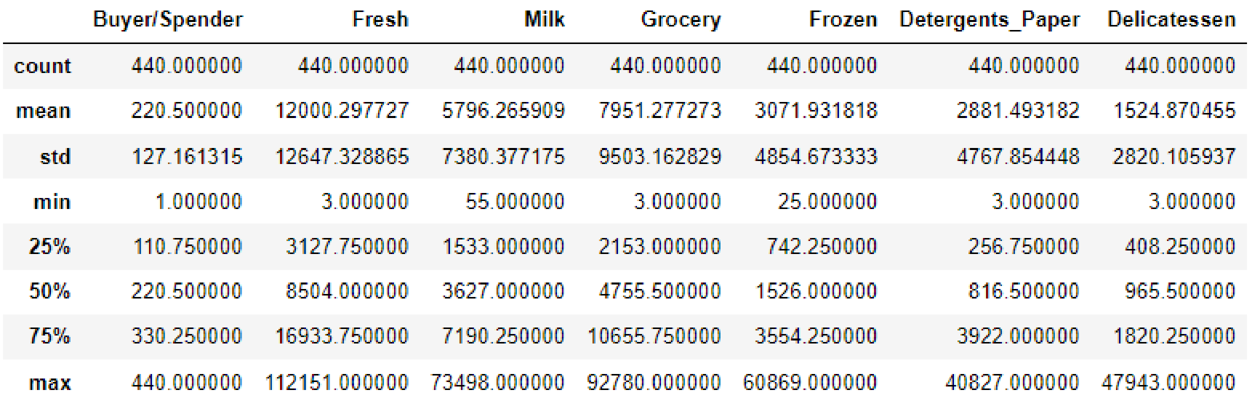
Check for missing values in the dataset:



```
Buyer/Spender       0
Channel             0
Region              0
Fresh               0
Milk                0
Grocery             0
Frozen              0
Detergents_Paper    0
Delicatessen        0
dtype: int64
```

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

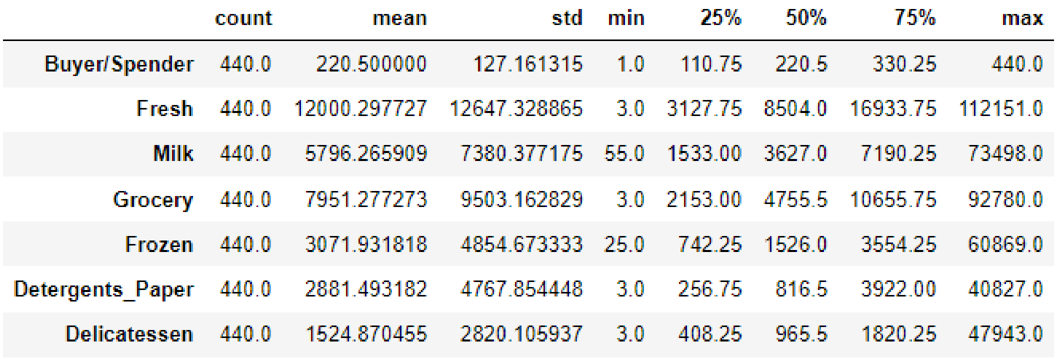
Descriptive statistics help describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data. The most recognized types of descriptive statistics are measures of centre: the mean, median, and mode, which are used at almost all levels of math and statistics.



	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
mean	220.500000	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	127.161315	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	1.000000	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	110.750000	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000
50%	220.500000	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000
75%	330.250000	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000
max	440.000000	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000

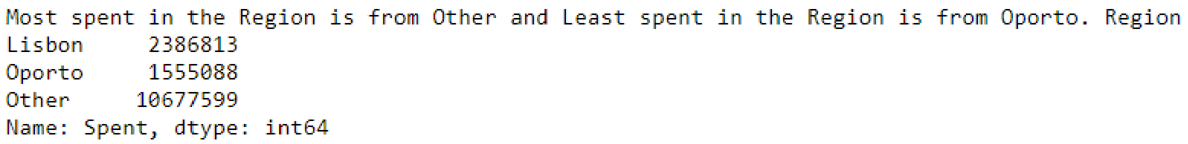
From the above figure, we can see there are 6 category of products and their measures of centre such as mean, median, mode, standard deviation, etc are calculated and represented.

1.1.1 Use methods of descriptive statistics to summarize data.



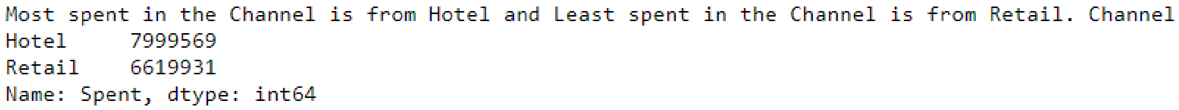
	count	mean	std	min	25%	50%	75%	max
Buyer/Spender	440.0	220.500000	127.161315	1.0	110.75	220.5	330.25	440.0
Fresh	440.0	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0
Milk	440.0	5796.265909	7380.377175	55.0	1533.00	3627.0	7190.25	73498.0
Grocery	440.0	7951.277273	9503.162829	3.0	2153.00	4755.5	10655.75	92780.0
Frozen	440.0	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0
Detergents_Paper	440.0	2881.493182	4767.854448	3.0	256.75	816.5	3922.00	40827.0
Delicatessen	440.0	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0

1.1.2 Which Region and which Channel spent the most?



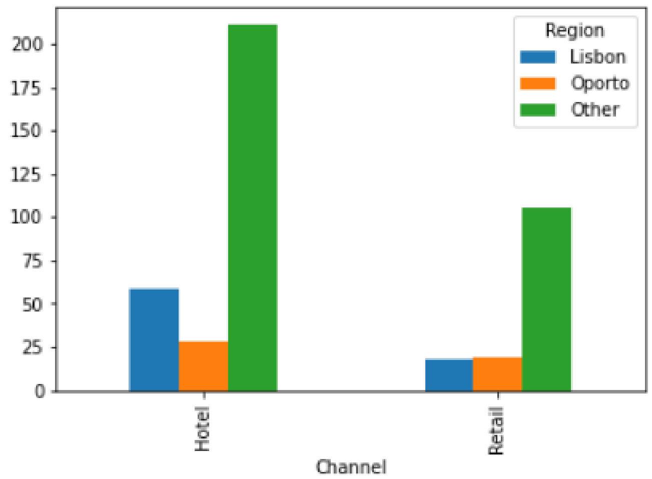
```
Most spent in the Region is from Other and Least spent in the Region is from Oporto. Region
Lisbon      2386813
Oporto      1555088
Other      10677599
Name: Spent, dtype: int64
```

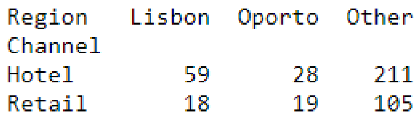
1.1.3 Which Region and which Channel spent the least?



```
Most spent in the Channel is from Hotel and Least spent in the Channel is from Retail. Channel
Hotel     7999569
Retail    6619931
Name: Spent, dtype: int64
```

Graphical Representation:





```
Region    Lisbon  Oporto  Other
Channel
Hotel         59      28    211
Retail        18      19    105
```

Here we are taking sum of all products by calculating region wise and channel wise that are most spent and least spent.

As per above graphical representation,

Most spent in the Region is from Other and Least spent in the Region is from Oporto.

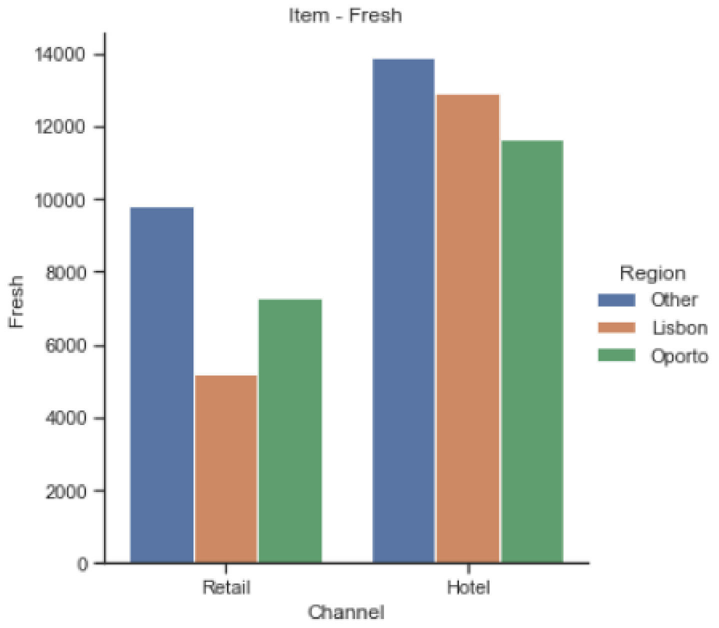
Most spent in the Channel is from Hotel and Least spent in the Channel is from Retail.

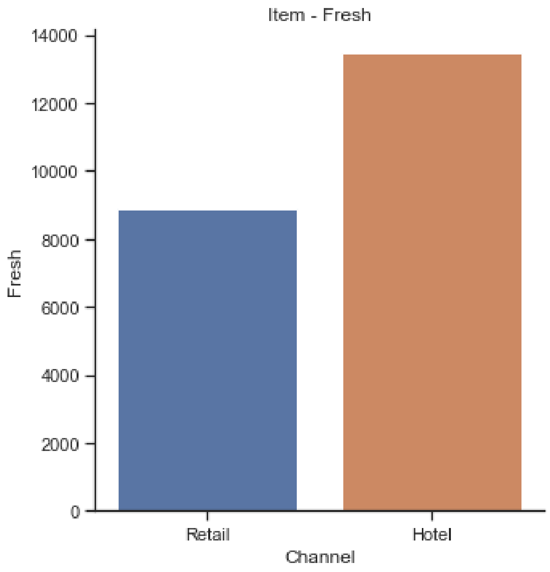
1.2. There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.

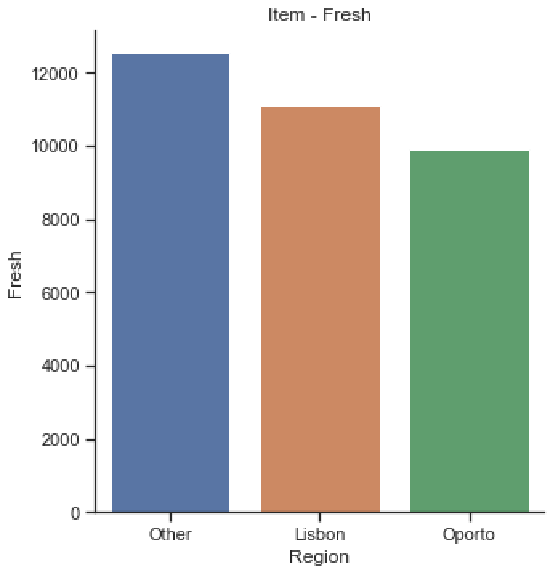
In the given wholesale customer dataset, there are 6 different varieties of items across region and channel are considered. We can analyse the data using bar plot representation.

Below are the plots represented analysing each item individually by categorizing region and channel wise.

Comparing Fresh item Region wise and Channel Wise:

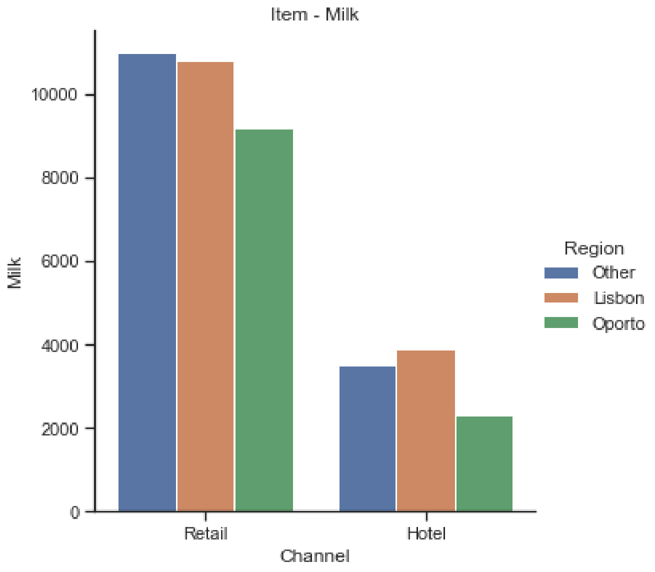


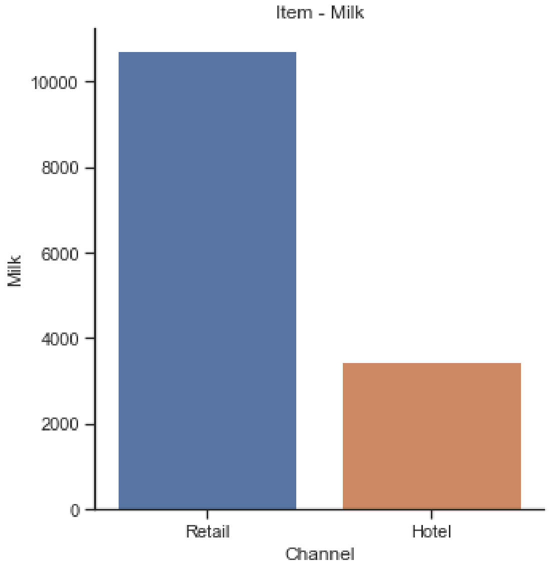


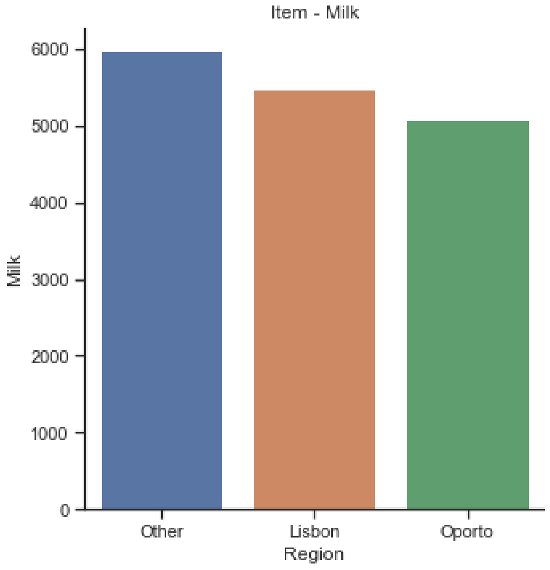


*As per above plots Fresh items are most sold in Hotel under Channel and under Region its Other.

Comparing Milk item Region wise and Channel Wise:

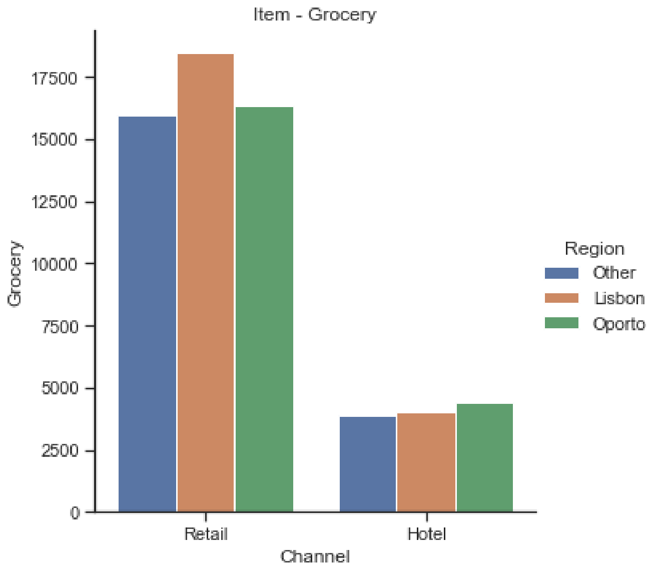


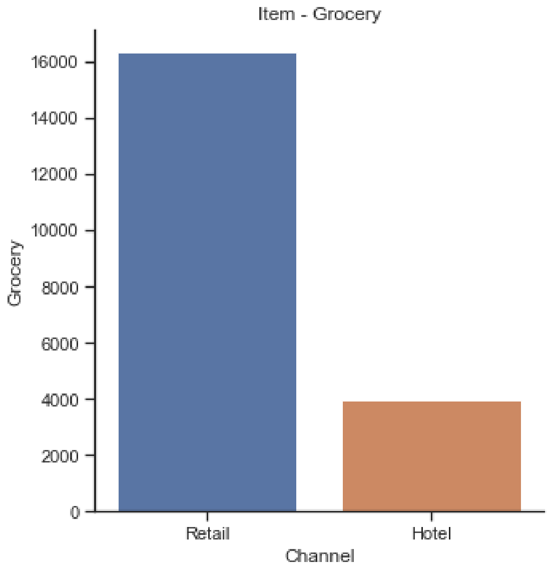


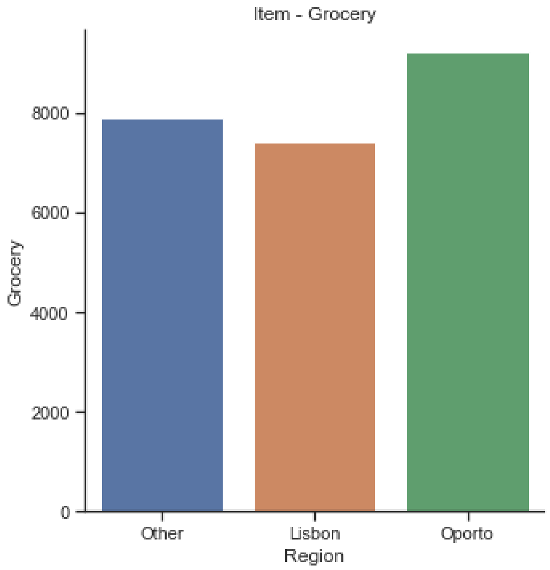


*As per above plots Milk item is most sold in Retail under Channel and under Region its Other.

Comparing Grocery item Region wise and Channel Wise:

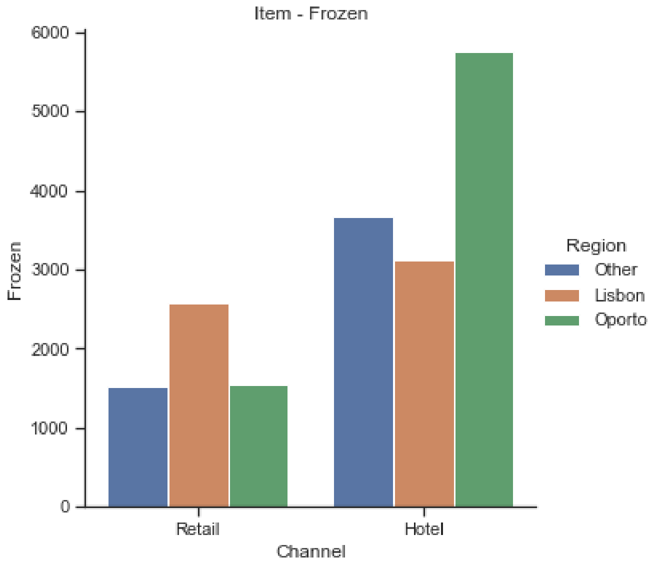


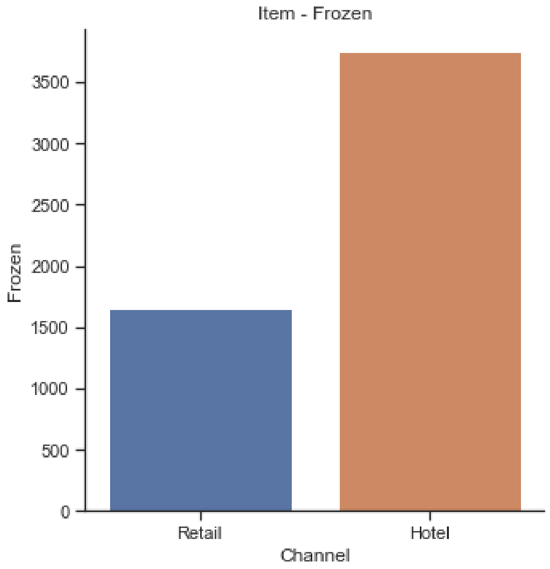


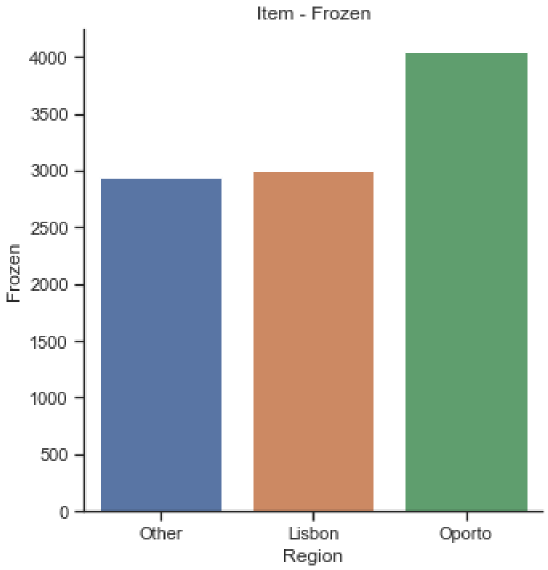


*As per above plots Grocery items are most sold in Retail under Channel and under Region its Oporto.

Comparing Frozen item Region wise and Channel Wise:

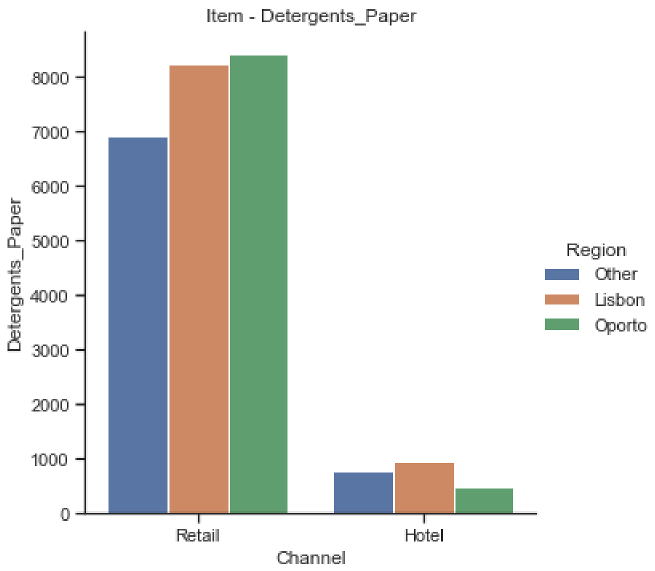


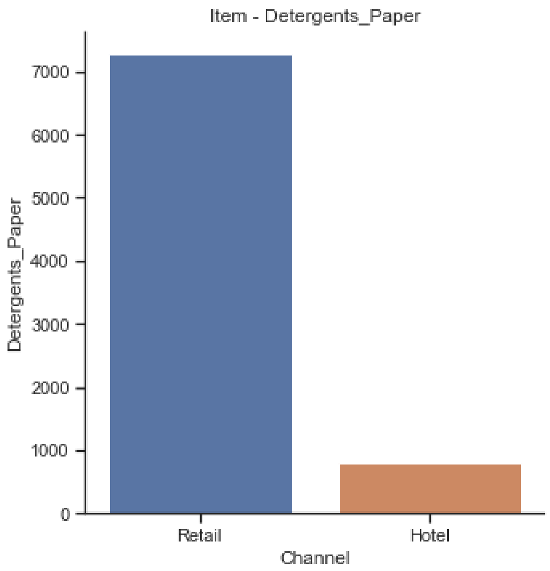


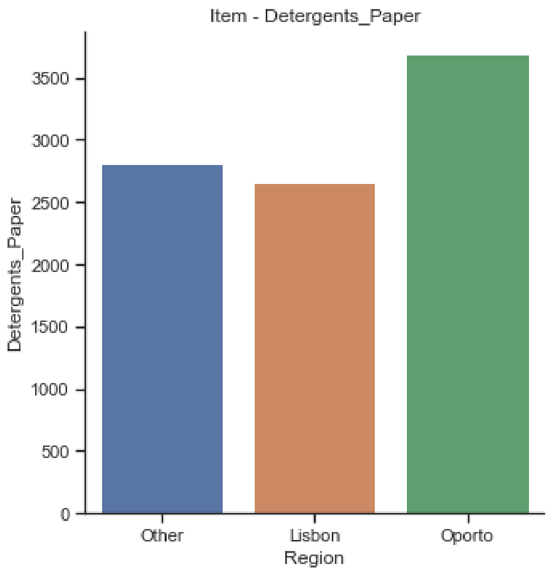


*As per above plots Frozen items are most sold in Hotels under Channel and under Region its Oporto.

Comparing Detergents-Paper item Region wise and Channel Wise:

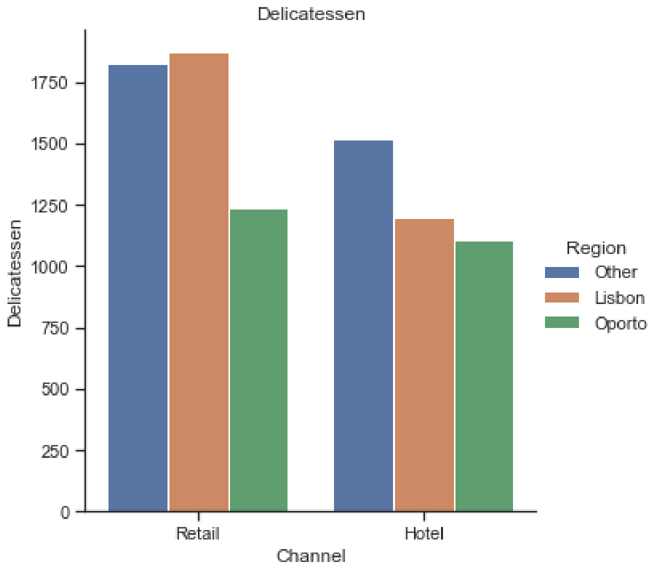


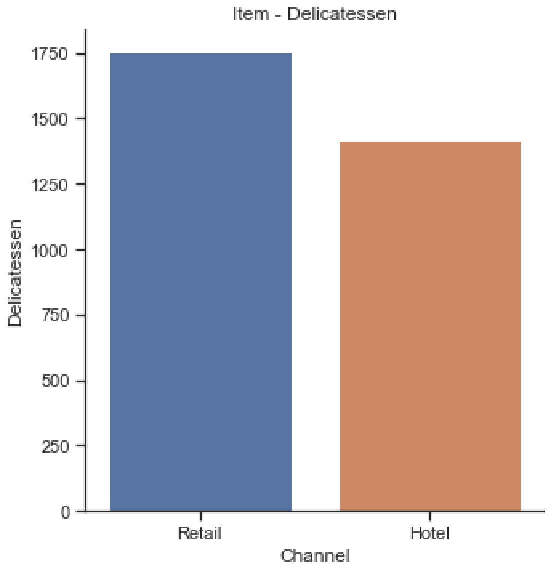


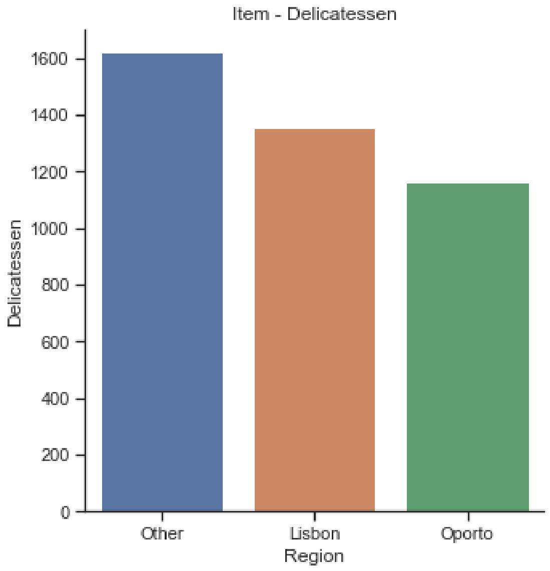


*As per above plots Detergents_Paper items are most sold in Retail under Channel and under Region its Oporto.

Comparing Delicatessen item Region wise and Channel Wise:







*As per above plots Delicatessen items are most sold in Retail under Channel and under Region its Other.

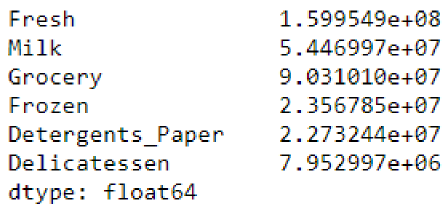
1.3 On the basis of the descriptive measure of variability, which item shows the most inconsistent behaviour? Which items shows the least inconsistent behaviour?

It talks about the descriptive measures of variability like IQR, Standard deviation, Coefficient of Variance, etc. Looking at the problem objective, Coefficient of Variance is the best measure to be computed to find the most inconsistent/least inconsistent behaviour.

Inconsistency here means variability, least inconsistent means less varying and most inconsistent means high variation. so, analysing each item is varying high/low in terms of price/quantity.

Coefficient of Variance of product = Standard Deviation of the Product/Mean of the Product.

Below is the data calculated and represented using Python codes.



```
Fresh               1.599549e+08
Milk                5.446997e+07
Grocery             9.031010e+07
Frozen              2.356785e+07
Detergents_Paper    2.273244e+07
Delicatessen        7.952997e+06
dtype: float64
```

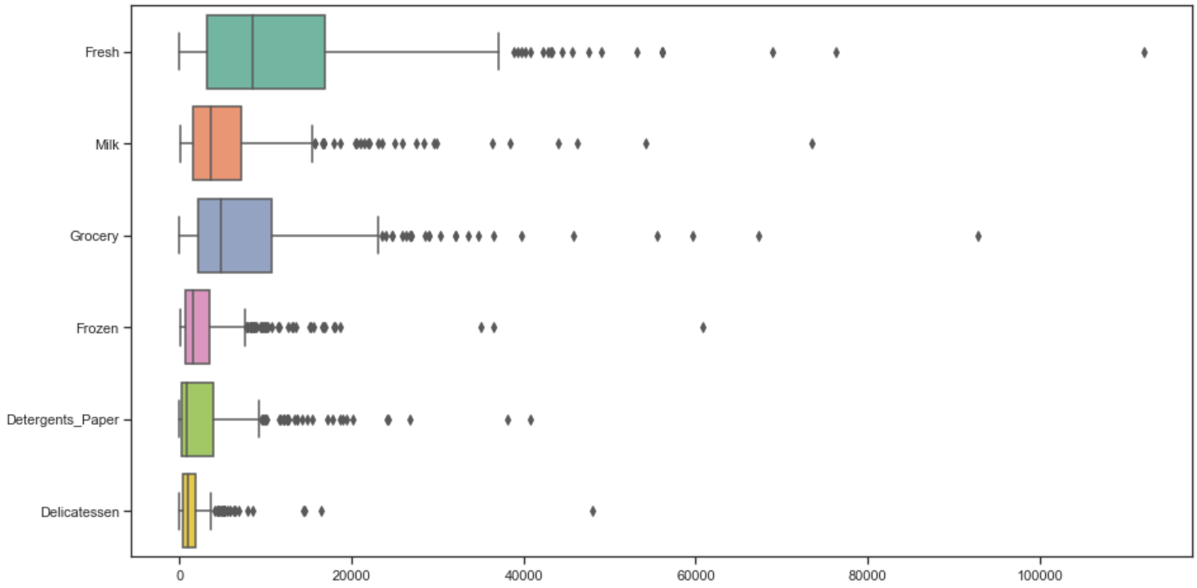
“Fresh” items have lowest coefficient of variation, so it is consistent.

“Delicatessen” items have highest coefficient of variation, so it is inconsistent.

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

We can analyse outliers by using box plots. As per below box plot, yes there are outliers in all the items across the product range (Fresh, Milk, Grocery, Frozen, Detergents_Paper & Delicatessen)

Outliers are detected but not necessarily removed, it depends on the situation. Here I will assume that the wholesale distributor provided us a dataset with correct data, so I will keep them as it is.



1.5 On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem? Answer from the business perspective.

As per the analysis, I find that there are inconsistencies in spending of different items by calculating Coefficient of Variance, which is to be minimized. The spending of Hotel and Retail channel are different which should be more, less or equal. And spent should equal for different regions. Need to focus on rest of the items other than “Fresh” and “Grocery”.

1.Channel: Unique values - 2 Frequent value - Hotel (298 out of 440 transactions.) 67.7 % of spending comes from the "Hotel" channel.

2.Region: Unique values - 3 Frequent value - Other (316 out of 440 transactions.) 71.8 % of spending comes from the "Other" region

3.Fresh: is the most consistent behaviour

4.Delicatessen: is the most inconsistent behaviour

5.Highest amount of money was spent in the region is Other and the total money spent was 10677599

6.Highest amount of money was spent in the channel is Hotel and the total money spent was 7999569

7.Lowest amount of money was spent in the region is Oporto and the total money spent was 1555088.

8.Lowest amount of money was spent in the channel is Retail and the total money spent was 6619931.

PROBLEM 2

Executive Summary:

The survey data set refers to student's survey analysis from CMSU which includes Grad Intention, Employment, GPA etc details of students.

Introduction:

The purpose of this whole exercise is to explore the dataset. This data set is recommended for learning and practicing skills in exploratory data analysis by using statistical methods and computing probabilities.

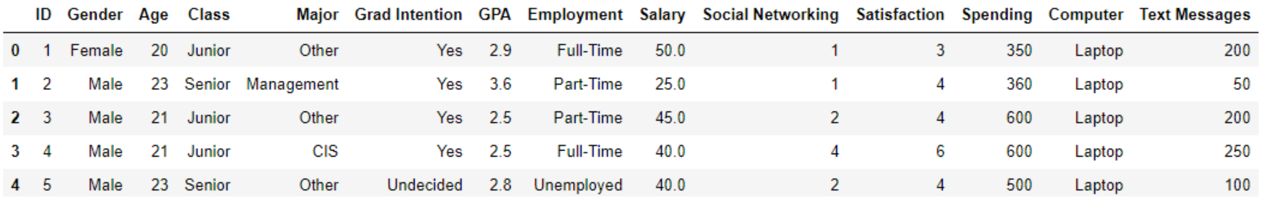
The survey dataset has 62 rows and 13 columns about student's details.

Data Description:

Description of variables is as follows:

- GPA: Annual score of the students in CMSU (Continuous).
- Salary: Annual spending of the students in CMSU (Continuous).
- Text Messages: Total count of messages (Continuous).
- Spending: Annual spending (Continuous).
- Gender: Gender details of the students in CMSU (Nominal).
- Age: Age of the students in CMSU (Nominal).
- Class: Junior/Senior students in CMSU (Ordinal).
- Major: course details of the students in CMSU (Ordinal).
- Grad Intention: (Ordinal).

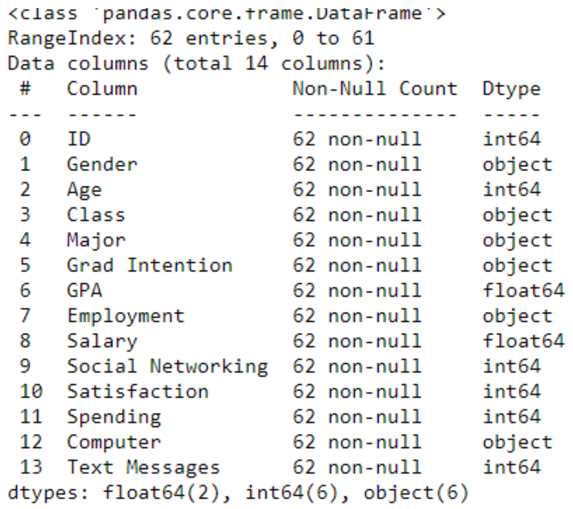
Sample of the dataset:



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

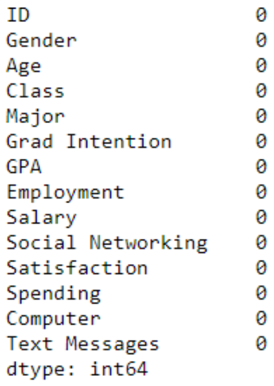
Exploratory Data Analysis

Check for types of variables in the data frame:



```
<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)
```

Check for missing values in the dataset:

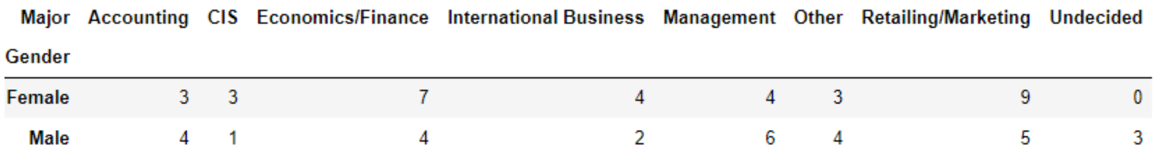


```
ID                   0
Gender               0
Age                  0
Class                0
Major                0
Grad Intention       0
GPA                  0
Employment           0
Salary               0
Social Networking    0
Satisfaction         0
Spending             0
Computer             0
Text Messages        0
dtype: int64
```

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

2.1.1. Gender and Major

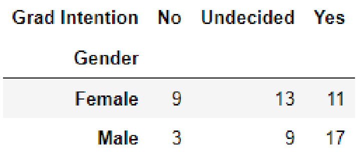
Contingency table based on Gender and Major details.



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

2.1.2. Gender and Grad Intention

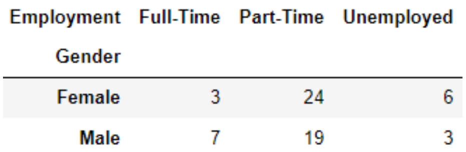
Contingency table based on Gender and Grad Intention details.



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

2.1.3. Gender and Employment

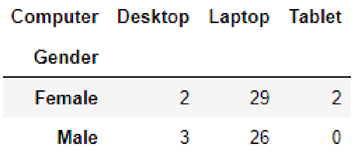
Contingency table based on Gender and Employment details.



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

2.1.4. Gender and Computer

Contingency table based on Gender and Computer details.



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

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

Based on given Gender details, number of male and female students need to be calculated using python codes. Total number of students in CMSU are 62 among which male students are 29 and female students are 33.

2.2.1 What is the probability that a randomly selected CMSU student will be male?

Here, the formula is (Probability of Male) / (Total Students)

The probability that a randomly selected CMSU student will be male is: 46.7% (or) 0.46774193548387094

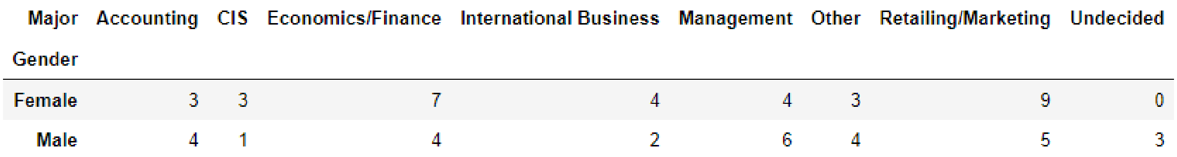
2.2.2 What is the probability that a randomly selected CMSU student will be female?

Here, the formula is (Probability of Female) / (Total Students)

The probability that a randomly selected CMSU student will be male is: 53.2% (or) 0.532258064516129

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

Based on the contingency table details of Gender and Major columns, probability is to be calculated. Total number of students in CMSU are 62 among which male students are 29 and female students are 33.



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

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

P (major | male) = P (major ∩ male)/ P(male)

Among Male Students in CMSU:
The probability of students in Accounting is: 13.7% (or) 0.13793103448275862
The probability of students in CIS is: 03.44% (or) 0.034482758620689655
The probability of students in Economics/Finance is: 13.7% (or) 0.13793103448275862
The probability of students in International Business is: 06.8% (or) 0.06896551724137931
The probability of students in Management is: 20.6% (or) 0.20689655172413793
The probability of students in Other is: 13.7% (or) 0.13793103448275862
The probability of students in Retailing/Marketing is: 17.2% (or) 0.1724137931034483
The probability of students in Undecided is: 10.3% (or) 0.10344827586206896

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

P (major | female) = P (major ∩ female)/ P(female)

Among Female Students in CMSU:
The probability of students in Accounting is: 09.09% (or) 0.09090909090909091
The probability of students in CIS is: 09.09% (or) 0.09090909090909091
The probability of students in Economics/Finance is: 21.2% (or) 0.21212121212121213
The probability of students in International Business is: 12.1% (or) 0.12121212121212122

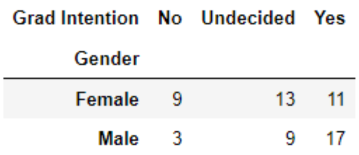
The probability of students in Management is: 12.1% (or) 0.12121212121212122
The probability of students in Other is: 09.09% (or) 0.09090909090909091
The probability of students in Retailing/Marketing is: 27.2% (or) 0.2727272727272727
The probability of students in Undecided is: 0% (or) 0.0

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.

Based on contingency table constructed for columns Gender and Grad Intention, probability is to be calculated.

P (Grad Intention ∩ Male) = P (Grad Intention| Male) x P (male)



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

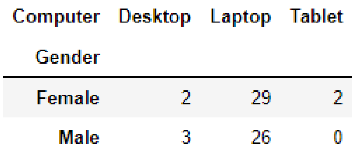
Here 17 are No. of male students who intend to graduate out of 29 total male students and 62 are total students in CMSU.

Hence the probability that a randomly chosen student is male and intends to graduate is 27.4% (or) 0.27419354838709675.

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

Based on contingency table constructed for columns Gender and Computer, probability is to be calculated.

P (No Laptop ∩ Female) = P (No Laptop| Female) x P (Female)



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

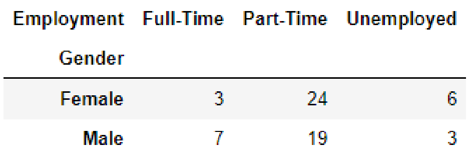
Here 4 are No. of female students who do not have laptop out of 33 total female students and 62 are total students in CMSU.

Hence the probability that a randomly selected student is a female and do not have a laptop is 06.4% (or) 0.06451612903225806

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

2.5.1 Find the probability that a randomly chosen student is a male or has a full-time employment

Based on contingency table constructed for columns Gender and Employment, probability is to be calculated.



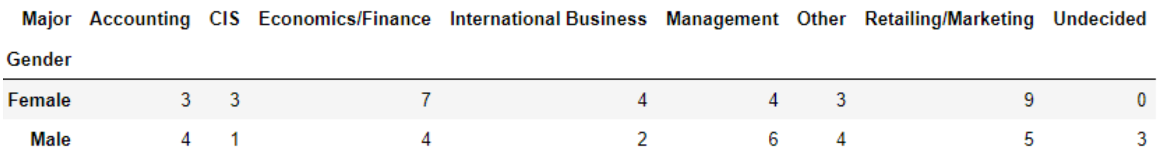
Employment	Full-Time	Part-Time	Unemployed
Gender			
Female	3	24	6
Male	7	19	3

P (A or B) = P (A) + P (B) – P (A n B)

The probability that a randomly chosen student is a male or has a full-time employment is 46.7% (or) 0.4678.

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

Based on contingency table constructed for columns Gender and Major, probability is to be calculated.



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

P (A / B) = P (A n B) / P (B)

The conditional probability that given a female student is randomly chosen; she is majoring in international business or management is 24.2% (or) 0.24242424242424243.

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 graduate intention and being females are independent events?

P (Grad Intention ∩ Female) = P (Grad Intention) * P (Female)

Here we need to calculate probability of female which is 53.2% (or) 0.532258064516129 and probability of Grad Intention which is 45.1 % (or) 0.45161290322580644.

Then multiply both probabilities i.e., P (Grad Intention) * P (Female) which is 24.03% (or) 0.24037460978147762.

Probability of combined event i.e., P (Grad Intention ∩ Female): is 17.7% (or) 0.1774193548387097.

These are not independent events as probability multiplication of both events is not equal to combined event. So, a graduate intention and being female candidate are not independent events.

2.7 Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending and Text Messages. Answer the following questions based on the data.

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

Here I am using normal distribution to calculate the probability. I have calculated me an value which is 3.129032258 and Standard deviation which is 0.377388393 using Excel.

My z value is -5.291278470771623.

By using Normal Distribution formula stats.norm.cdf(z), If a student is chosen rando mly, the probability that his/her GPA is less than 3: 6.073212770308564e-08.

2.7.2 Find conditional probability that a randomly selected male earns 50 or more.

In this case it’s a straightforward answer, probability of male students whose salary is 50 or more is 14 and divided by total no. of male students is 29.

The conditional probability that a randomly selected male earns 50 or more is 48.2% (or) 0.482758620689655.

Find conditional probability that a randomly selected female earns 50 or more.

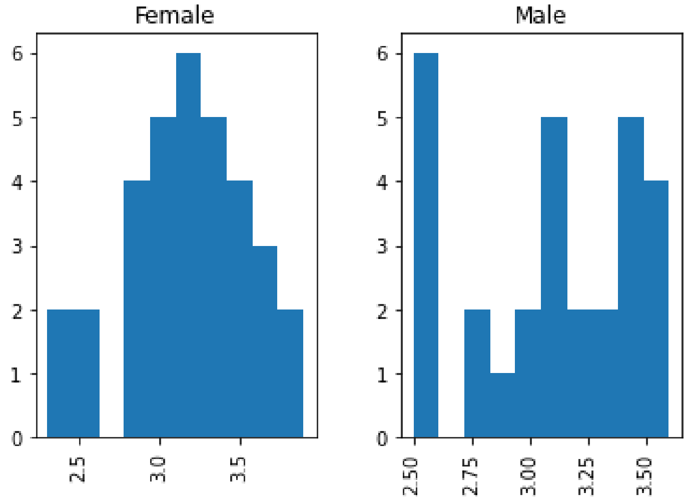
Same as above, probability of female students whose salary is 50 or more is 18 and divided by total no. of female students is 33.

The conditional probability that a randomly selected female earns 50 or more is 54.5% (or) 0.5454545454545454.

2.8.1 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.

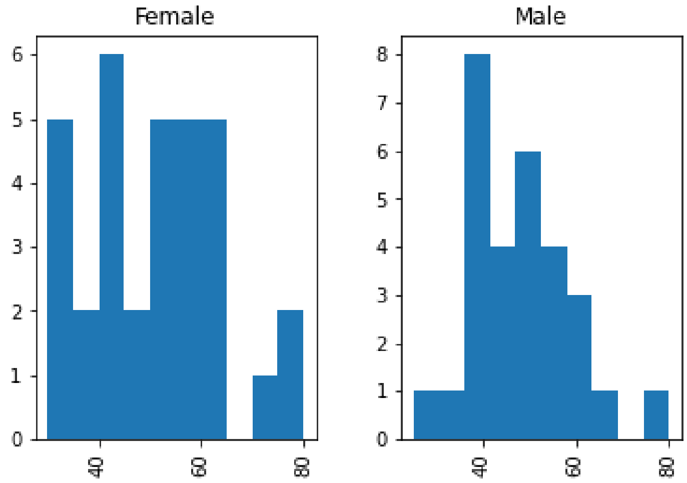
Below are the histograms plotted for continuous variables GPA, Salary, Spending and Text Messages segregating Gender wise.

Histogram for GPA



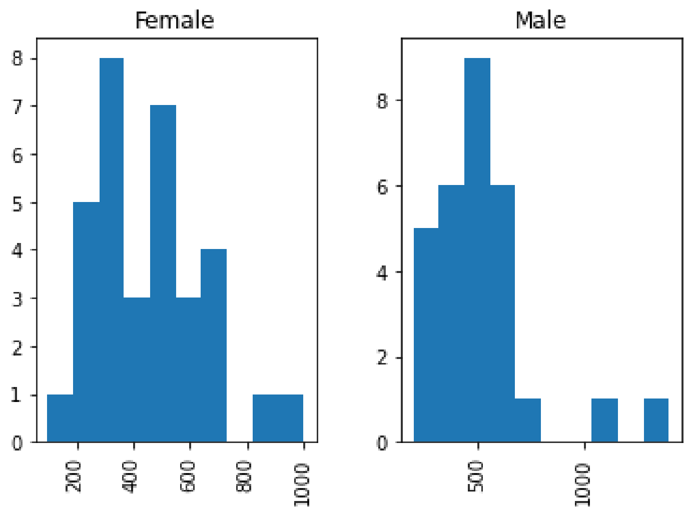
*Based on the above plot GPA for Female students is normally distributed whereas for male students it is not normally distributed.

Histogram for Salary



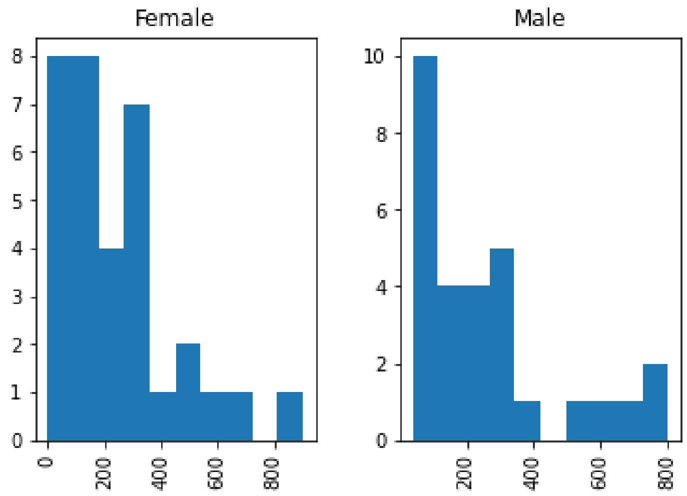
*Based on the above plot salary for male students is high compared to female students and normally distributed whereas for female students it is not normally distributed.

Histogram for Spending



*Based on the above plot spending is equal among male and female students, also spending among both are not normally distributed.

Histogram for Text Messages



*Based on the above plot Text messages among both male and female students are not normally distributed.

2.8.2 Write a note summarizing your conclusions

For the given problem with available data, I have constructed various contingency tables such as Gender-Major, Gender-Grad Intention, Gender-Employment, Gender-Computer.

In the survey dataset representative population of CMSU is given and based on it probability to identify randomly selected student is male or female is identified.

Also, with this methodology the conditional probability of different majors among male and female students, conditional probability that randomly selected male and randomly selected female who earn 50 or more in CMSU is identified.

Similarly male student who intends to graduate, female student not having laptop, student who is either male or has full-time employment, female student who is randomly chosen - that she is majoring in international business or management, variables in the dataset such as - GPA, Salary, Spending, Text Messages are identified.

PROBLEM 3

Executive Summary:

The company here already has a process in place where it keeps reducing the moisture content until the moisture content becomes less than 0.35. So it will try to check every time whether the moisture content is still greater than 0.35 pounds per 100 square feet. So, when we say that the company would like to show that the mean moisture content is less than 0.35 pounds per 100 square feet, it is looking at whether the moisture content is still greater than 0.35 pounds per 100 square feet. Hence, for every moisture test, the claim to check here becomes whether the moisture content is still greater than 0.35 pounds per 100 square feet.

Introduction:

A & B Shingles dataset is recommended for learning and practicing skills in exploratory data analysis by performing hypothesis testing. The expectation is to execute the hypothesis test and interpret the result based on the given hypothesis.

Data Description:

Alternative hypothesis (HA): mean moisture content > 0.35 and,

Null hypothesis (H0): mean moisture content <= 0.35.

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

H0: mean moisture content <=0.35.

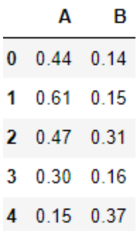
HA: mean moisture content > 0.35.

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

H0: mean moisture content <=0.35

HA: mean moisture content > 0.35

Sample of the dataset:



	A	B
0	0.44	0.14
1	0.61	0.15
2	0.47	0.31
3	0.30	0.16
4	0.15	0.37

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.

H0: mean moisture content <=0.35

HA: mean moisture content > 0.35

Alpha value is considered as 0.05

Performing one sample t-test individual for A & B

One sample t-test for A

Using python code t-statistic and p-value are calculated as below.

t-statistic: -1.4735046253382782, p-value: 0.07477633144907513

Conclusion

Since p-value > 0.05, do not reject Null Hypothesis (H0).

There is not enough evidence to conclude that the mean moisture content for Sample A shingles is less than 0.35 pounds per 100 square feet. p-value = 0.0748. If the population mean moisture content is in fact no less than 0.35 pounds per 100 square feet, the probability of observing a sample of 36 shingles that will result in a sample mean moisture content of 0.3167 pounds per 100 square feet or less is 0.0748.

One sample t-test for B

Using python code t-statistic and p-value are calculated as below.

t-statistic: -3.1003313069986995, p-value: 0.0020904774003191826

Conclusion

Since p-value < 0.05, reject Null Hypothesis (H0).

There is enough evidence to conclude that the mean moisture content for Sample B shingles is not less than 0.35 pounds per 100 square feet. p-value = 0.0021. If the population mean moisture content is in fact no less than 0.35pounds per 100 square feet, the probability of observing a sample of 31 shingles that will result in a sample mean moisture content of 0.2735 pounds per 100 square feet or less is .0021.

3.2 Do you think that the population means 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?

H0: μ(A)= μ(B)

HA: μ(A)! = μ(B)

Alpha is considered as 0.05

This is a Two sample t-test

Using python code t-statistic and p-value are calculated as below.

t-statistic=1.29 and p-value=0.202

Conclusion

As the p-value > 0.05, do not reject Null Hypothesis(H0).

We can say that population mean for shingles A and B are equal test assumptions when running a two-sample t-test, the basic assumptions are that the distribution of the two population are normal, and that the variances of the two distribution is the same. If those assumptions are not likely to be met, another testing procedure could be used.

-------XXX-------