PARAKANDLA MANISHA

PGP – DSBA ONLINE

DATE: 12-09-2021

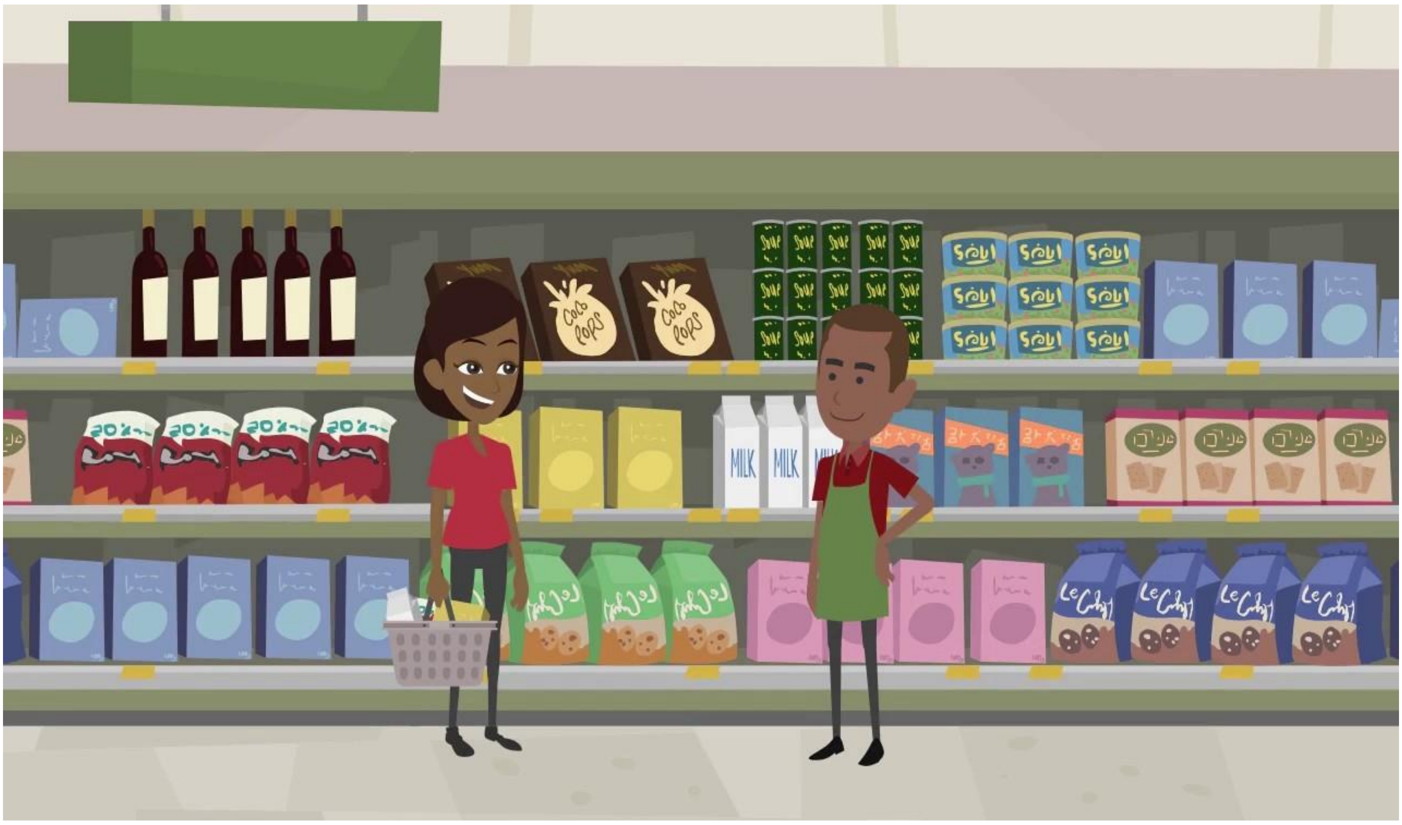
STATISTICAL METHODS OF DECISION MAKING

SMDM BUSINESS REPORT

Context: Page:

List of Figures:

List of Tables:



WHOLESALE CUSTOMER 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).

→INTRODUCTION

The data contains 440 observations (sample) of annual spending on 6 different product categories and number of customer spending on the annual wholesale. The objectives of the study is to determine the correlation between dependent variable which is the annual spending on Fresh, Milk, Grocery, Frozen, Detergent and Paper, Delicatessen products to determine the significant differences in mean of each annual spending on these products towards the region and to factorize the factors that are highly correlates to each other. We need to perform the exploratory data analysis on the Wholesale Customer dataset to draw the desired conclusions.

→DATA DESCRIPTION

The data contains 440 observations with 9 columns and 6 varieties of different products.

Channel: (Hotel, Retail) - categorical

Region: (Lisbon, Oporto, Other) - Categorical

Fresh: continuous

Milk: continuous

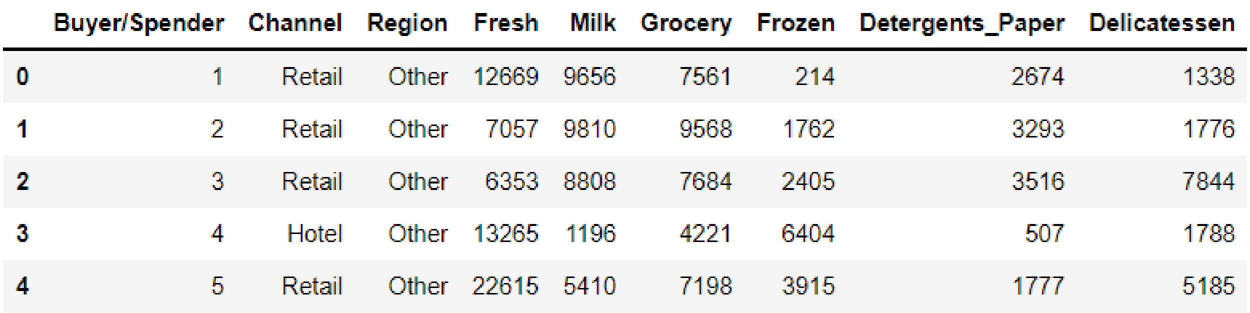
Grocery: continuous

Frozen: continuous

Detergents_Paper: continuous

Delicatessen: continuous

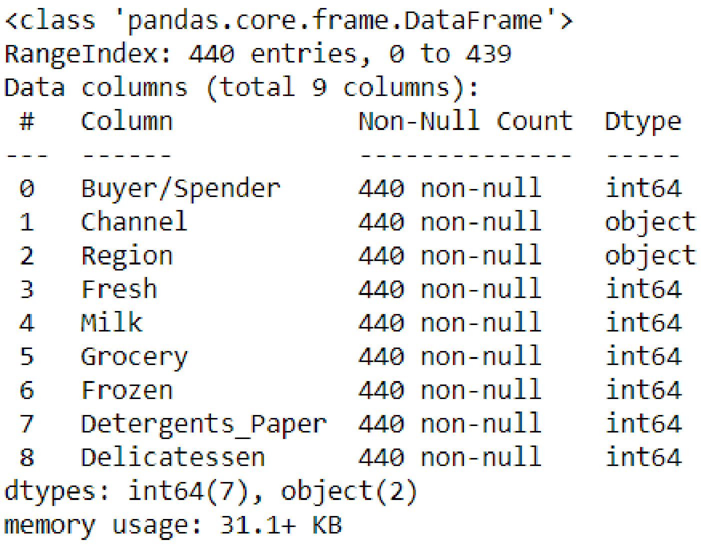
→SAMPLE OF A DATASET



	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

Table 1. Wholesale Customer dataset sample

Exploratory Data Analysis

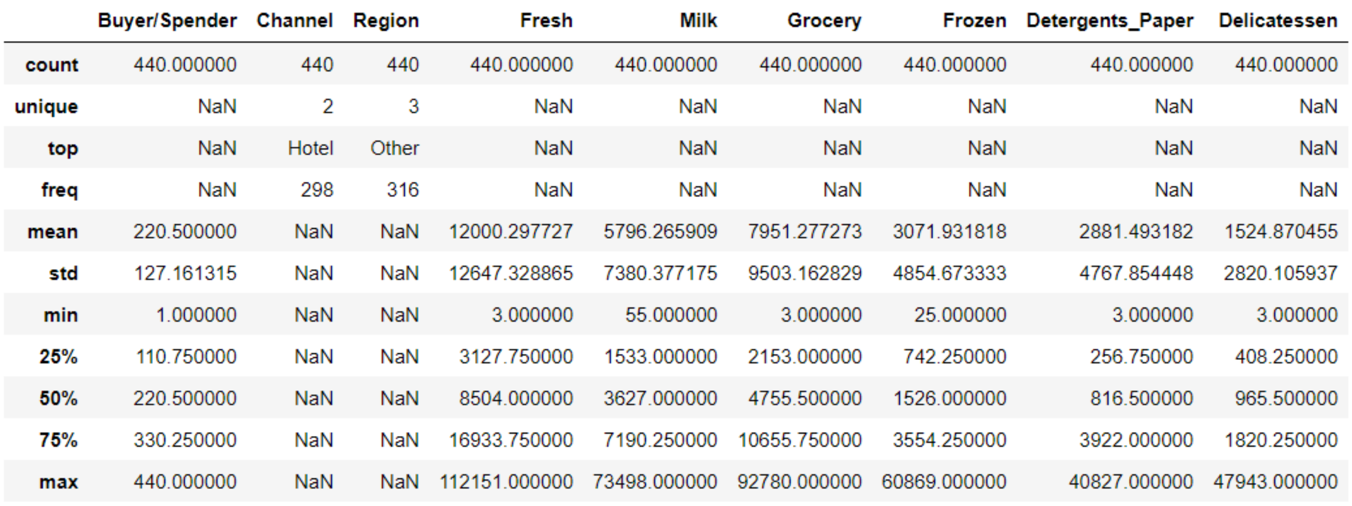


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 440 entries, 0 to 439
Data columns (total 9 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   Buyer/Spender     440 non-null    int64 
 1   Channel           440 non-null    object
 2   Region            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
```

We can now observe that there are no null values in the dataset and the datatypes of variables in the columns.

1.1Use 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 summarizes the characteristics of a dataset. It consists of two basic categories i.e., Measures of Central Tendency and Variability. It is used to describe the basic features of data like mean, median, Mode, count, std etc. Here we used the describe() function to summarize data.



	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440.000000	440	440	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
unique	NaN	2	3	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	Hotel	Other	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	298	316	NaN	NaN	NaN	NaN	NaN	NaN
mean	220.500000	NaN	NaN	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	127.161315	NaN	NaN	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	1.000000	NaN	NaN	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	110.750000	NaN	NaN	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000
50%	220.500000	NaN	NaN	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000
75%	330.250000	NaN	NaN	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000
max	440.000000	NaN	NaN	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000

Table 2. Summary of the dataset

Which Region and Which Channel spent the most? Which Region and Which channel Spent the least?

Using Group by function for both the region and channel columns we can find the highest spent and the least spent. Here we are using bar graph to represent the Highest and lowest spent on both region and channel. Through this we can conclude

The customers spent in “Other” region is more: 10741625.00 $

The customers spent in “Oporto” region is least: 1569987.00 $

The Hotel Channel spent more than Retail: 8070603.00 $

The Retail Channel spent the least: 6645917.00 $

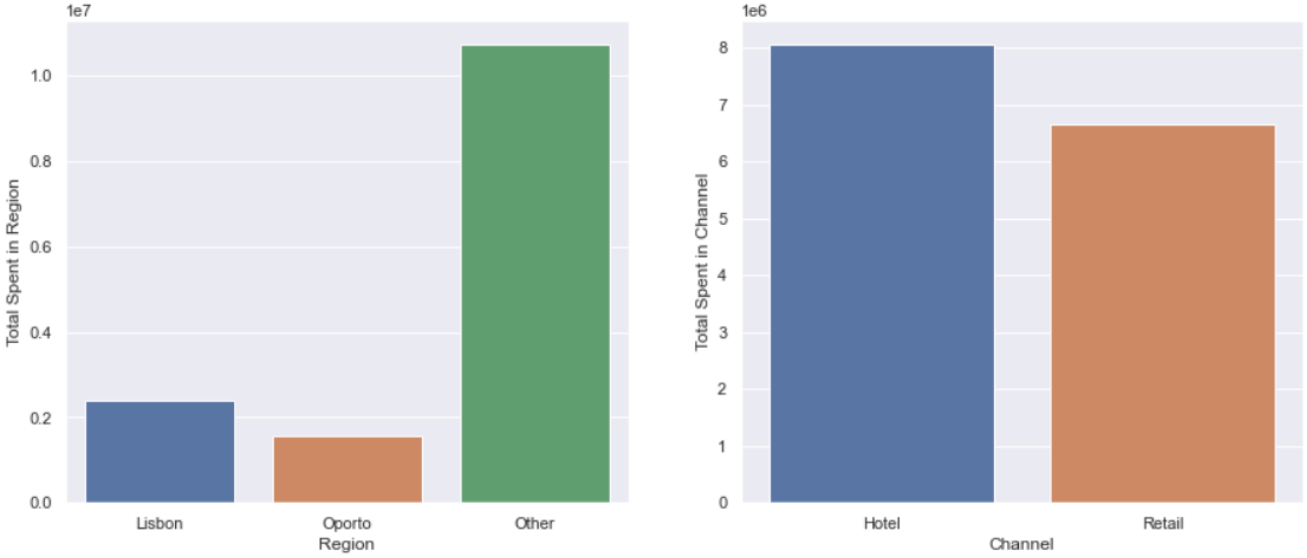
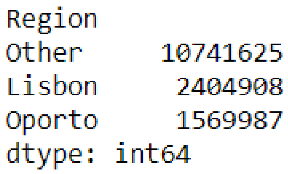


Fig1. Bar graph for Region & Channel

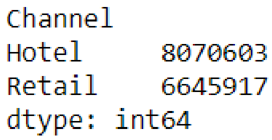
The Outputs from Python are:

For Region:



```
Region
Other     10741625
Lisbon     2404908
Oporto     1569987
dtype: int64
```

For Channel:



```
Channel
Hotel     8070603
Retail    6645917
dtype: int64
```

From the Analysis, we can conclude that the “Other” Region is the highest spent and “Oporto” is the least spent Region. The “Hotel” Channel is the highest spent and “Retail” is the least spent in Channel.

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

There are 6 continuous variables in the data according to Region & Channel.

We can describe the 6 different variables using the Pivot tables for each category and also check the spend across Region and Channel. This can be represented using plot as shown below:

Here, Milk, Grocery, Detergents_Powder are the highest spent in Retail channel compared to Hotel Channel across all regions. The Varieties Fresh & Frozen have highest spent in Hotel Channel compared to Retail Channel.

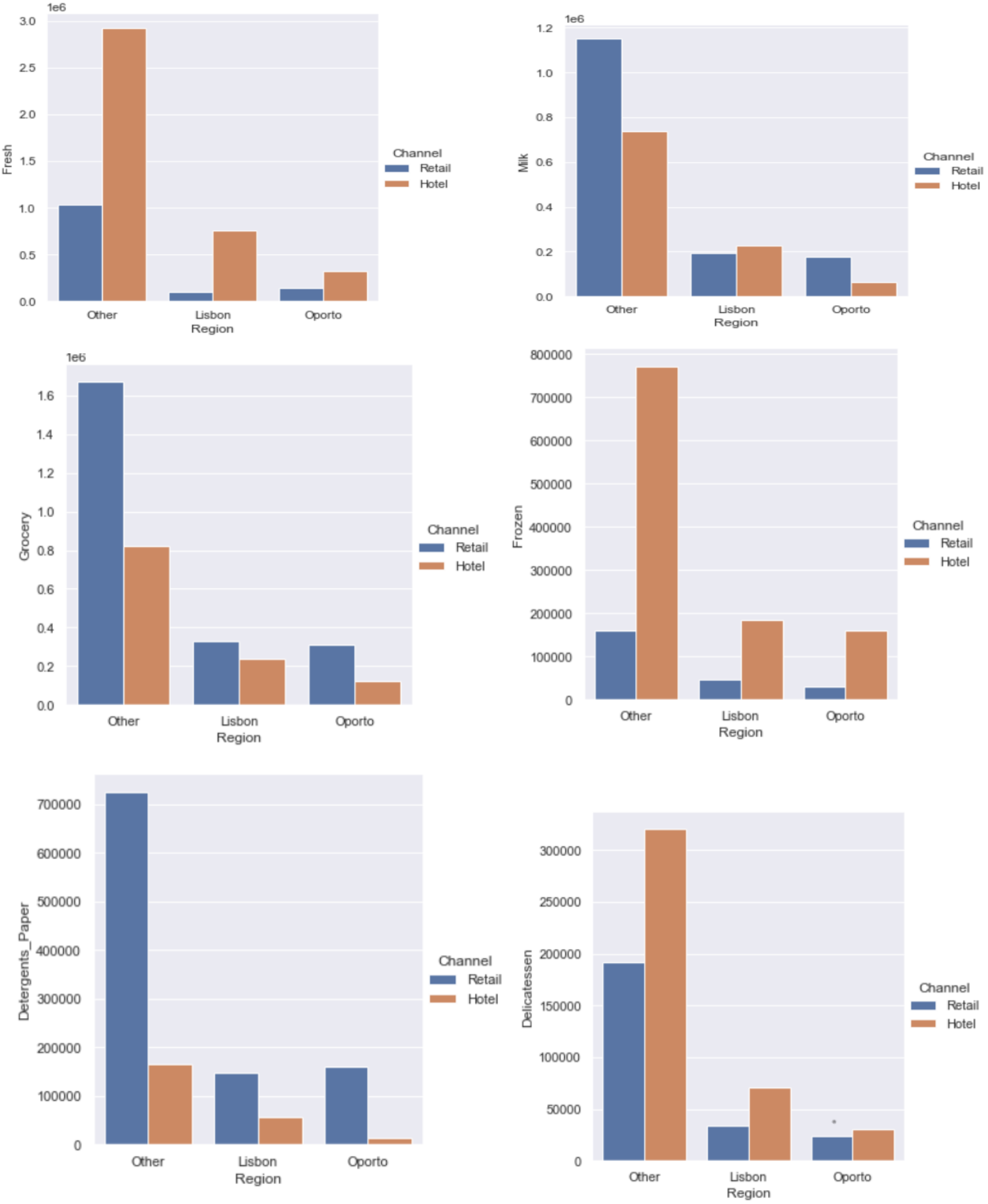


Fig2. Plots to describe the variable behaviour

We can also summarize the spending for Fresh and Groceries is Maximum in Region and Channel and the Delicatessen is the least in both the Region and Channel.

This can be represented using the boxplot. Using box plot we can also find the outliers present in each category.

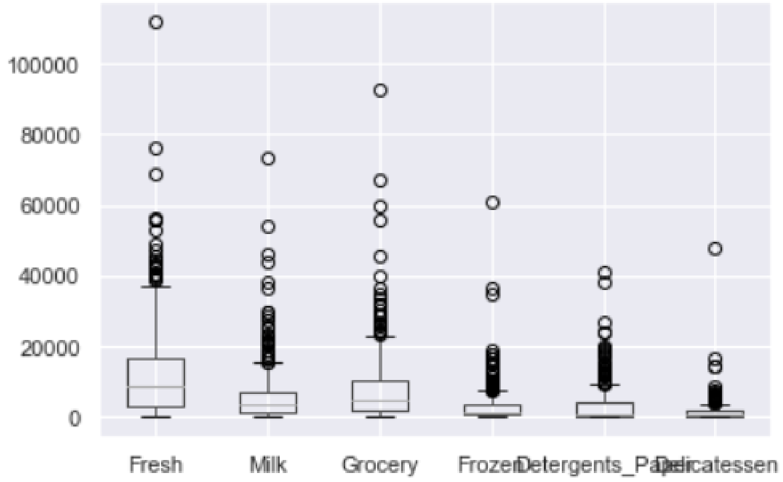
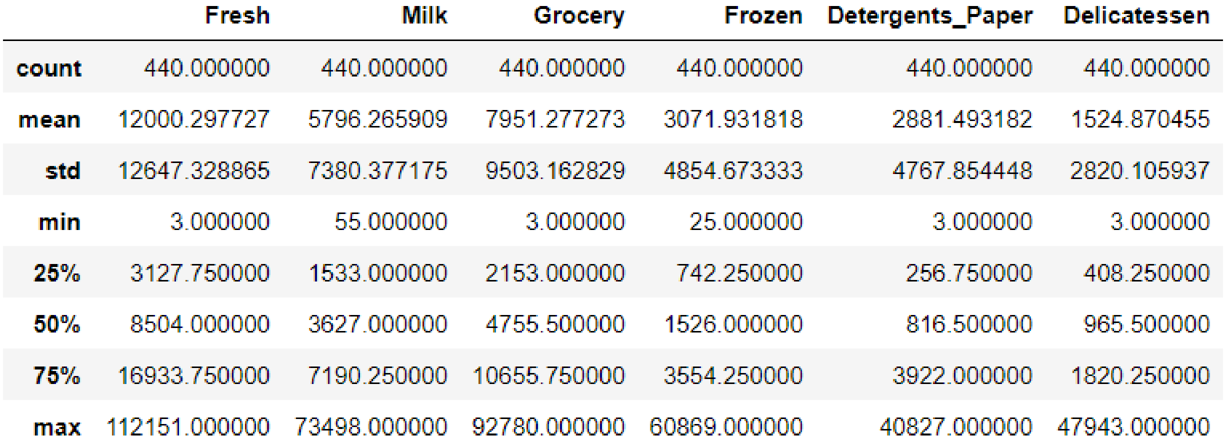


Fig3. Boxplot for six different variables

DESCRIPTIVE SUMMARY OF VARIABLES:



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

Table3. Summary table

Fresh: The total sale of fresh items in the three regions is 5280131 with mean 12000, SD 12647 and median 8504, and the data is right skewed, means range is huge which is approximately more than 1,12,151, which makes it the most popular item across the region.

Grocery: The total sale of grocery among the entire region is the second most selling category in wholesale market with total sale of 9503 mean 7951 and standard deviation 9503 which shows the distribution is not symmetrical and doesn't closely represent the normal curve.

Milk: The total sale of milk and its subsidiaries are the third highest selling product across the region with total sale of 7380, mean 5796, SD 7380 and median 3627. Higher SD tells us that the data is not following the normal distribution and it also has a wide range.

Frozen: Frozen items are not much sold across the region; it is the fourth highest selling category with mean 3071 and standard deviation 4854 while the median is 1526. As the mean is to the right of median and SD is high it means the data is not symmetric and it is slightly right skewed. The low sale of Frozen items also tells us the geographical condition of the region that the place might be colder region that is why the sales are low.

Detergents_Paper: It is the fifth highest selling category among the three regions with a mean 2881 and standard deviation 4767. And the mean is affected by high standard deviation otherwise the mean would have been less and close to the median which is 816. The stats suggest that the data is right skewed.

Delicatessen: It is the least selling category in the data with mean 1524 and SD 2820 and median 965 the stats suggest that the data is right skewed and the mean is affected by the outliers.

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

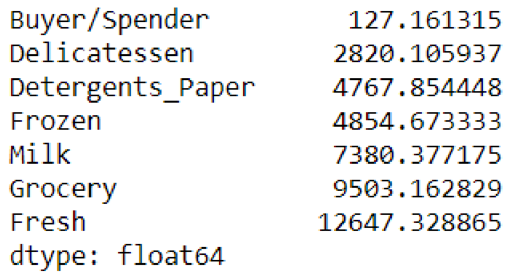
→Inconsistency here means variability.

→Least inconsistent means less variation and most inconsistent means high variation.

→Hence, we have to find which item is varying high/low in terms of price/quantity.

→std, IQR, Range, Skewness, Coefficient of Variation, Plot.

The standard deviation of the items:



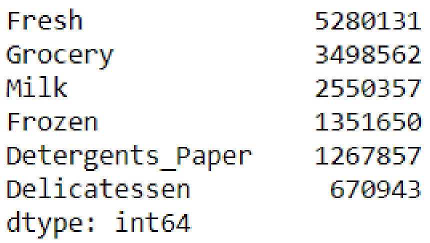
```
Buyer/Spender          127.161315
Delicatessen          2820.105937
Detergents_Paper      4767.854448
Frozen                4854.673333
Milk                  7380.377175
Grocery               9503.162829
Fresh                12647.328865
dtype: float64
```

Fresh item has the highest Standard Deviation. Hence, it is inconsistent.

Delicatessen item has least Standard Deviation. Hence, it is least inconsistent.

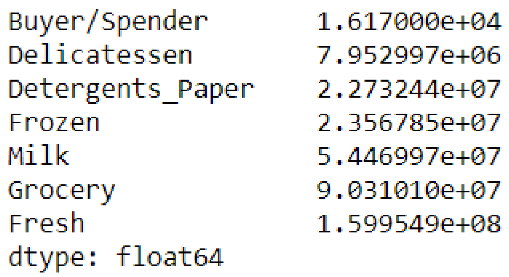
Outputs from Python:

The total sale of Items:



```
Fresh               5280131
Grocery             3498562
Milk                2550357
Frozen              1351650
Detergents_Paper    1267857
Delicatessen         670943
dtype: int64
```

Variance of items:

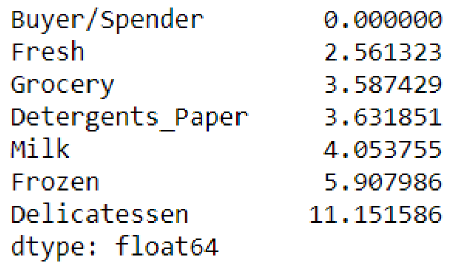


```
Buyer/Spender         1.617000e+04
Delicatessen          7.952997e+06
Detergents_Paper      2.273244e+07
Frozen                2.356785e+07
Milk                  5.446997e+07
Grocery               9.031010e+07
Fresh                 1.599549e+08
dtype: float64
```

Less Variance → Least Inconsistent

High Variance → Most Inconsistent

Skewness:



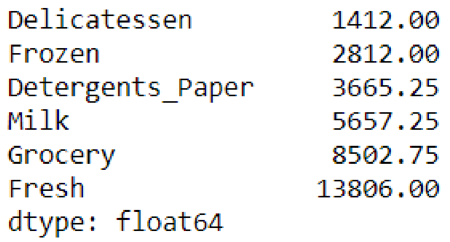
```
Buyer/Spender         0.000000
Fresh                 2.561323
Grocery               3.587429
Detergents_Paper      3.631851
Milk                  4.053755
Frozen                5.907986
Delicatessen         11.151586
dtype: float64
```

The least skewed is least consistent

Coefficient of variation:

Coefficient of variation of Fresh is: 1.0527196084948245
Coefficient of variation of Grocery is: 1.2718508307424503
Coefficient of variation of Milk is: 1.193815447749267
Coefficient of variation of Frozen is: 1.5785355298607762
Coefficient of variation of Detergents_Paper: 1.6527657881041729
Coefficient of variation of Delicatessen: 1.8473041039189306

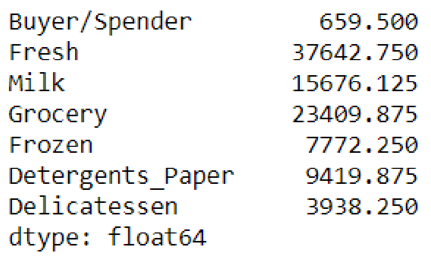
Inter Quartile Range(IQR):



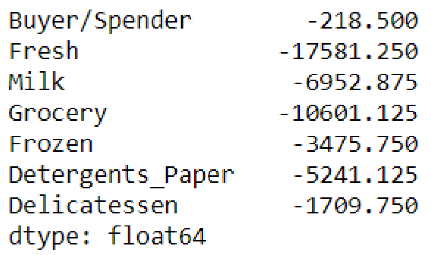
```
Delicatessen          1412.00
Frozen                2812.00
Detergents_Paper      3665.25
Milk                  5657.25
Grocery               8502.75
Fresh                13806.00
dtype: float64
```

Outliers(upper outliers, Lower Outliers):

Upper outliers



```
Buyer/Spender          659.500
Fresh                37642.750
Milk                 15676.125
Grocery              23409.875
Frozen                7772.250
Detergents_Paper      9419.875
Delicatessen          3938.250
dtype: float64
```



```
Buyer/Spender         -218.500
Fresh               -17581.250
Milk                 -6952.875
Grocery             -10601.125
Frozen               -3475.750
Detergents_Paper     -5241.125
Delicatessen         -1709.750
dtype: float64
```

Lower Outliers

Through Coefficient of variation, we can find that least value of item “Fresh” (1.05) and the highest value of item “Delicatessen” (1.84). Hence, from the analysis we can conclude that,

The item which shows the most inconsistent behaviour is “Delicatessen”. The item which shows the least inconsistent behaviour is “Fresh”.

Plot showing Distribution and Variability:

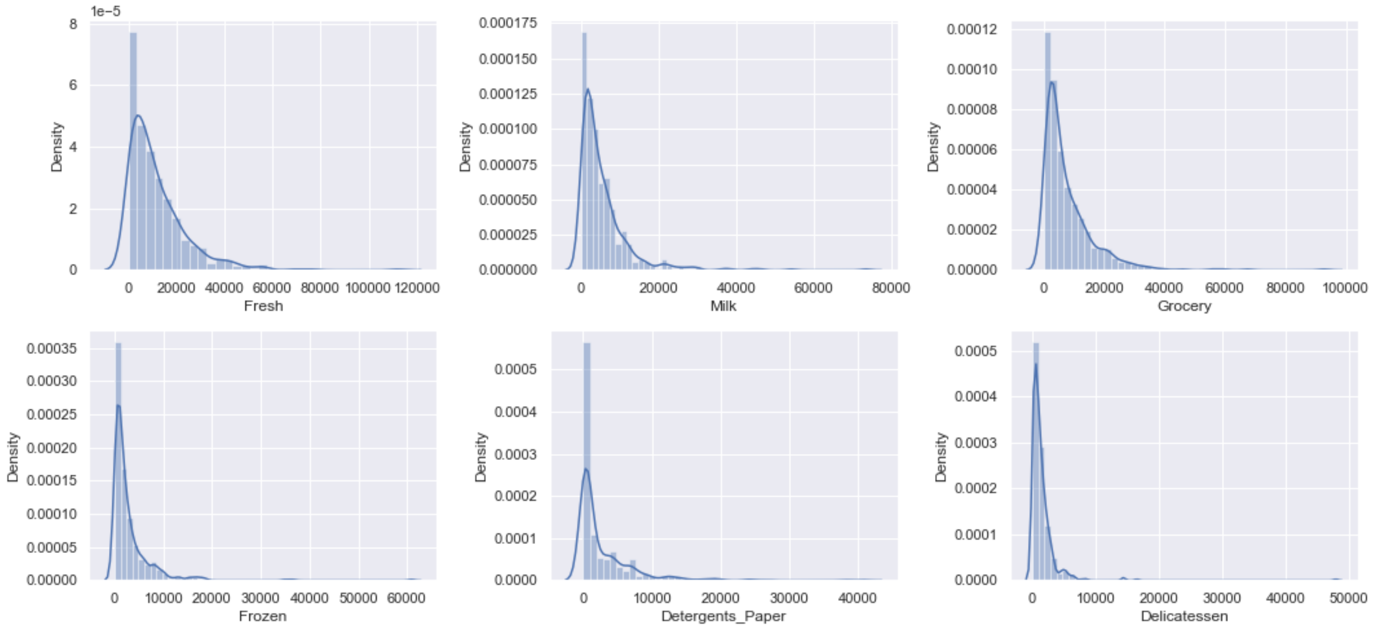


Fig 4. Distplot showing distribution of different items

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 use Box plot to represent/find the Outliers in the data. **Boxplot** is a method for graphically depicting groups of numerical data through their quartiles. Box plots may also have lines extending from the boxes (*Whiskers*) indicating variability outside the upper and lower quartiles, hence the terms **box-and-whisker plot** and **box-and-whisker diagram**. Outliers are plotted as individual points. The spacings between the different parts of the box indicate the degree of dispersion (spread) and skewness in the data, and show outliers.

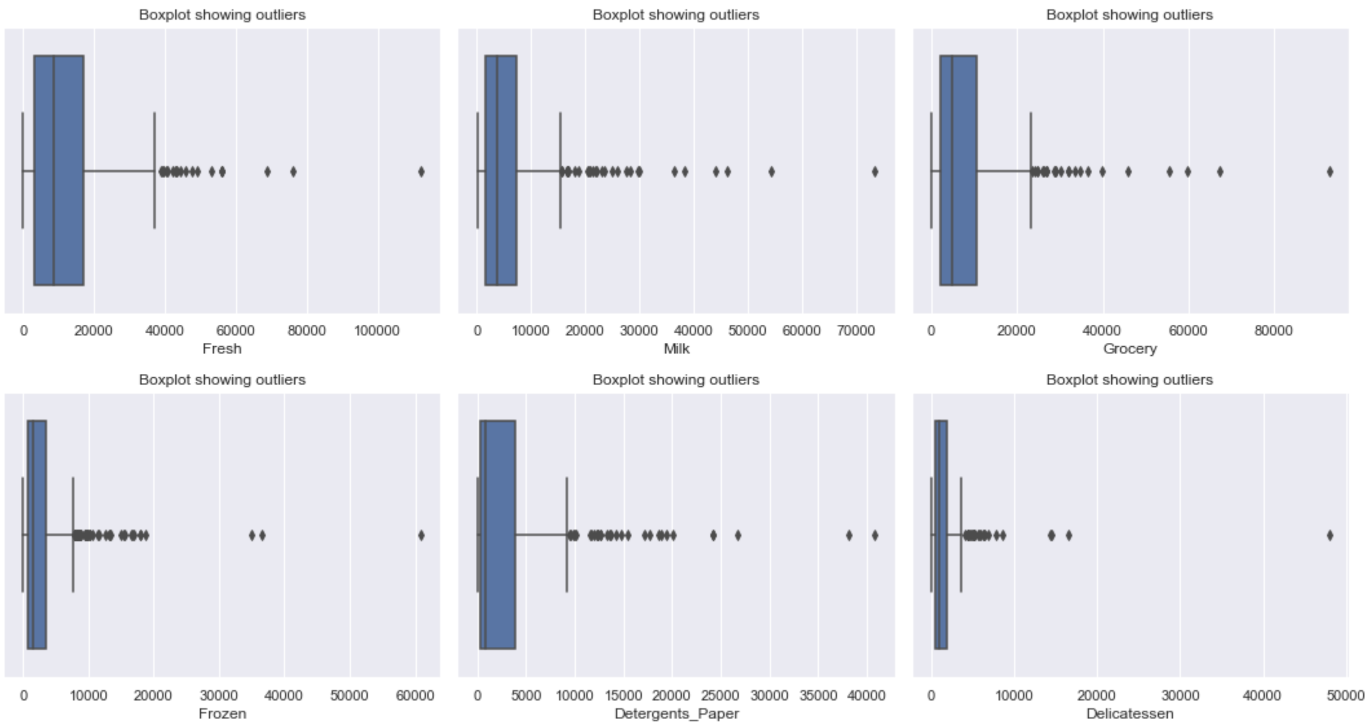
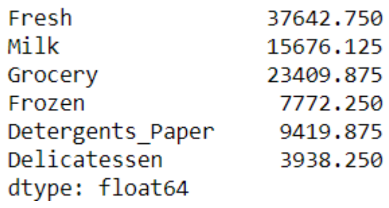


Fig 5. Boxplot showing Outliers

The Boxplot here shows the outliers in all the six varieties of items. All the six variables are having many outliers. We can also find the upper & Lower outliers using Python.

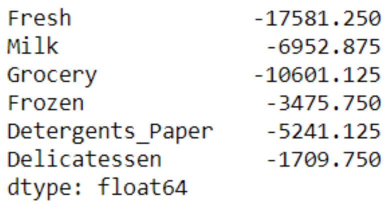
The outputs in Python for outliers are given below:

Upper Outliers:



```
Fresh               37642.750
Milk                15676.125
Grocery             23409.875
Frozen               7772.250
Detergents_Paper     9419.875
Delicatessen         3938.250
dtype: float64
```

Lower Outliers:



```
Fresh              -17581.250
Milk                -6952.875
Grocery            -10601.125
Frozen              -3475.750
Detergents_Paper    -5241.125
Delicatessen        -1709.750
dtype: float64
```

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

We can use the correlation plot here, as the Correlation plots can be used to quickly find insights. It is **used to investigate the dependence between multiple variables at the same time** and to highlight the most correlated variables in a data table. In this visual, correlation coefficients are coloured according to the value.

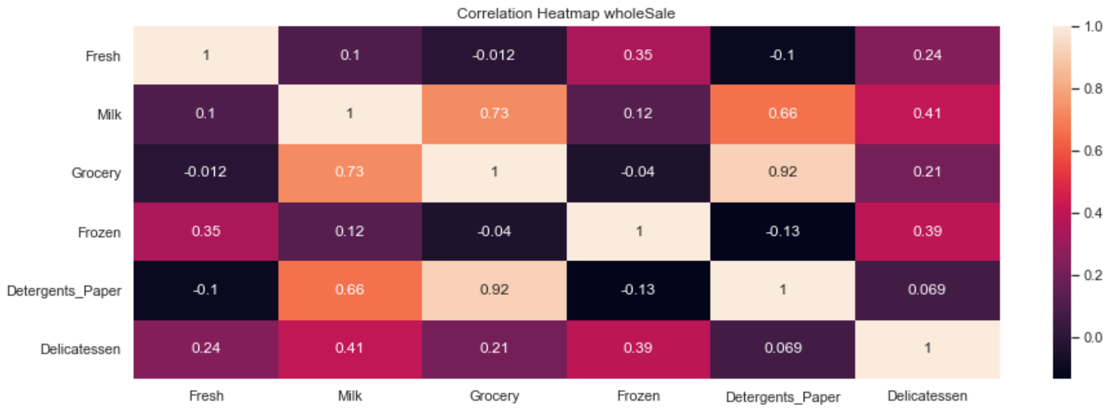


Fig 6. Correlation Plot

Conclusion & Recommendation:

From the Exploratory Data Analysis, we can suggest/recommend below suggestions to the Wholesalers:

1.There were Inconsistencies in the spending of the six different varieties of items which I have found while calculating coefficient of variation. It should be minimized.

2. The sale & spending over the Hotel and Retail in Channel are having lot of difference, this can be more or less equal and should be improved.

3.The Annual spending could be equal for different regions. As the Fresh and Grocery were far better, need to focus more on the other categories of items.

Clear Mountain State University (CMSU)

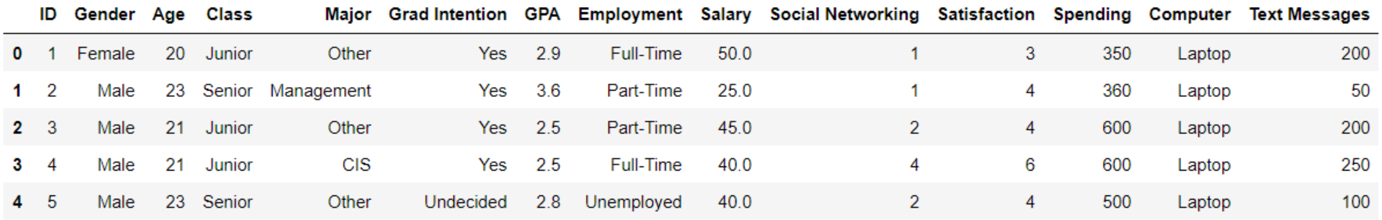
Problem Statement (Clear Mountain State University (CMSU))

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).

INTRODUCTION:

To Perform the exploratory data analysis we need to import the “Survey -1” Dataset. The dataset consists of 62 observations with 62 rows and 14 columns. With this data we can analyse the data about the Undergraduate students who attended CMSU.

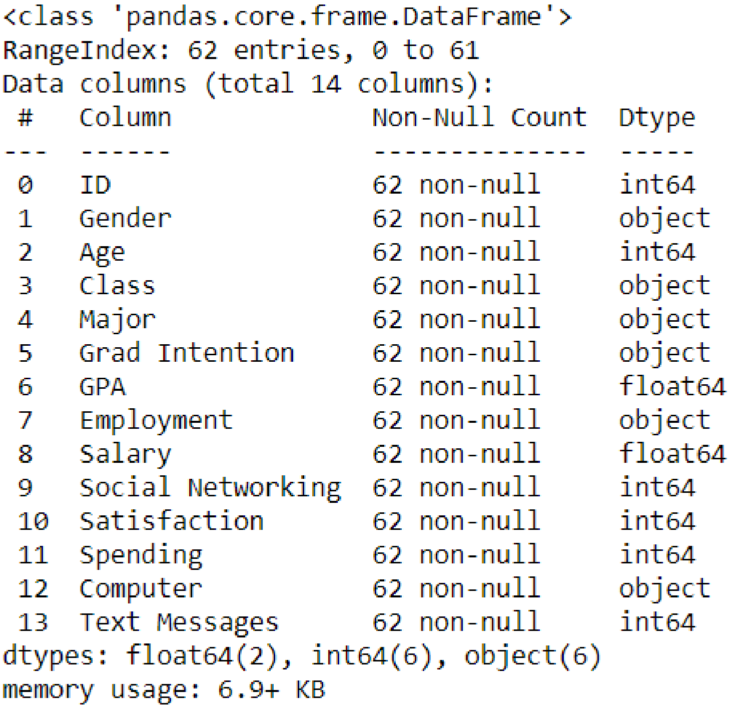
SAMPLE OF THE DATASET:



	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

Table 4. CMSU dataset sample

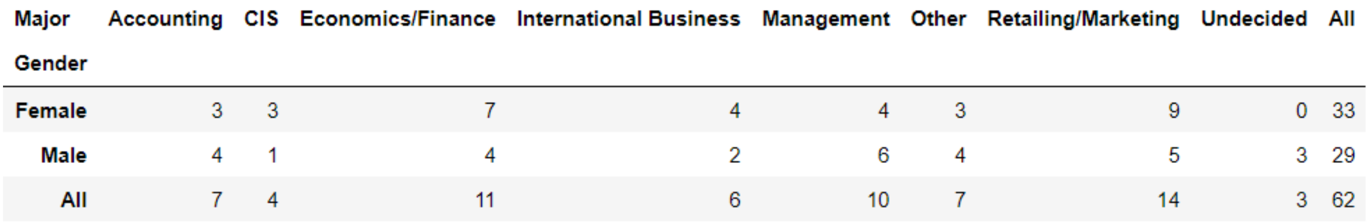
EXPLORATORY DATA ANALYSIS:



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

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

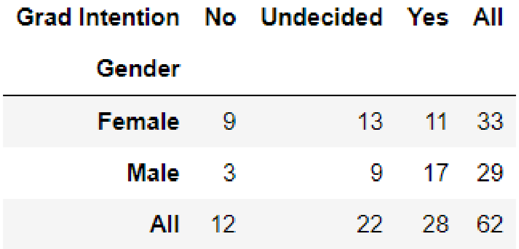
2.1.1. Gender and Major



Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	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

Table5. Gender_Major Contingency table

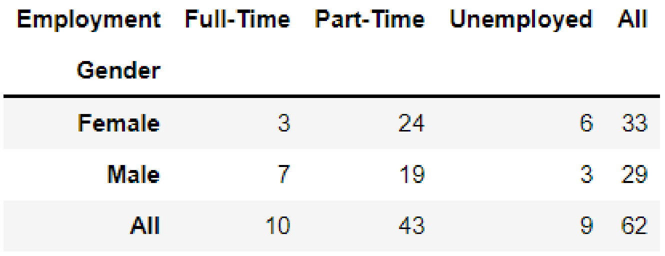
2.1.2. Gender and Grad Intention



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

Table6. Gender & Grad Intention Contingency table

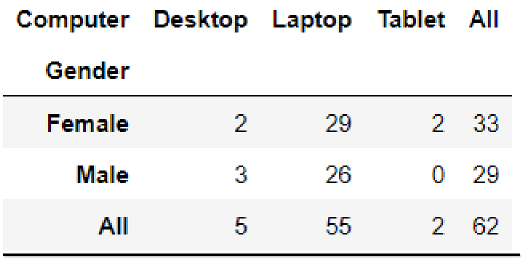
2.1.3. Gender and Employment



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

Table7. Gender & Employment Contingency table

2.1.4. Gender and Computer



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

Table8. Gender & Computer Contingency table

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?

Probability that a randomly selected CMSU student is a Male: 0.4677

For this we need to find the Total male students out of all students from the given data. After Calculation, The probability of a randomly selected CMSU student is a Male is **46.77%**

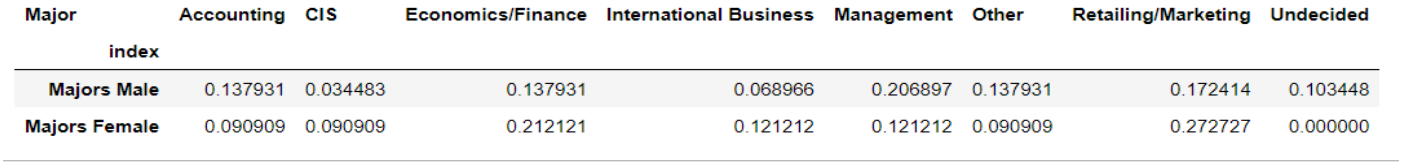
2.2.2. What is the probability that a randomly selected CMSU student will be female?

Probability that a randomly selected CMSU student is a Female: 0.5323

For this we need to find the Total female students out of all students from the given data. After Calculation,
The probability of a randomly selected CMSU student is a Female is **53.23%**

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

By using the contingency tables of Gender and Major, we can get the Total number of Males & Females opting for different Majors.



Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
index								
Majors Male	0.137931	0.034483	0.137931	0.068966	0.206897	0.137931	0.172414	0.103448
Majors Female	0.090909	0.090909	0.212121	0.121212	0.121212	0.090909	0.272727	0.000000

Table9.Majors Male & Female Contingency table

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

By Using the contigency tables of Gender and Major, we can get the Total number of Males opting for different Majors and the conditional Probability of different majors among male students in CMSU.

The Conditional Probability of male students opting for Accounting: **13.79%**
The Conditional Probability of male students opting for CIS: **3.45%**
The Conditional Probability of male students opting for Economics/Finance: **13.79%**
The Conditional Probability of male students opting for International Business: **6.90%**
The Conditional Probability of male students opting for Management: **20.69%**
The Conditional Probability of male students opting for Other: **13.79%**
The Conditional Probability of male students opting for Retailing/Marketing: **17.24%**
The Conditional Probability of male students opting for Undecided: **10.34%**

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

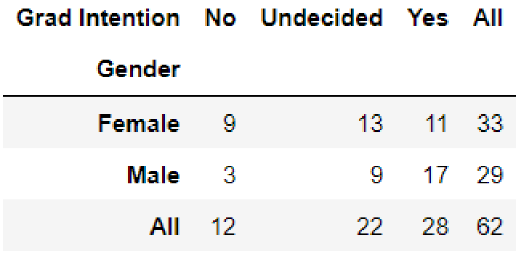
By Using the contingency tables of Gender and Major, we can get the Total number of Females opting for different Majors and the conditional Probability of different majors among Female students in CMSU

The Conditional Probability of male students opting for Accounting: **9.09%**
The Conditional Probability of male students opting for CIS: **9.09%**
The Conditional Probability of male students opting for Economics/Finance: **21.21%**
The Conditional Probability of male students opting for International Business: **12.12%**
The Conditional Probability of male students opting for Management: **12.12%**
The Conditional Probability of male students opting for Other: **9.09%**
The Conditional Probability of male students opting for Retailing/Marketing: **27.27%**
The Conditional Probability of male students opting for Undecided: **0.00%**

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.

By using the Gender and Grad intension contingency table, we can get the total number of males intends to graduate and the Probability of a randomly chosen student is a male and intends to graduate.



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

Table10.Gender & Grad Intention Contingency table

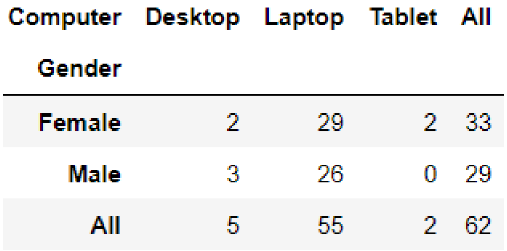
Probability (male and graduate intends) = probability(male) * Probability (graduate intends|male)
=17/62
= 0.2742

Probability that a randomly chosen student is male and intends to graduate is: 0.2742

Probability of a randomly chosen student is a male and intends to graduate: **27.42%**

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

By using the Gender and Computer contingency table, we can get the total number of female intends to graduate and the Probability of a randomly selected student is a female and does not have a Laptop.



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

Table11. Gender and computer Contingency table

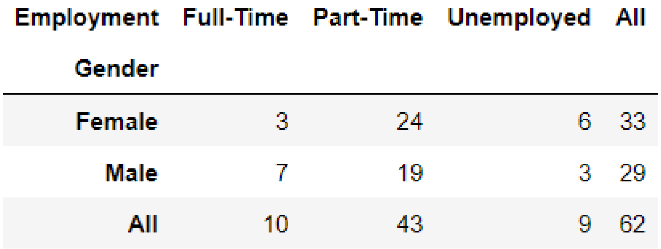
Probability (female and does not have a laptop) = Probability (no laptop|female) * Probability(female)

=33/62* 4/33 = 0.0645

Probability that a randomly selected student is a female and does not have a laptop is: 0.0645

Probability of a randomly selected student is a female and does not have a laptop is: **6.45%**

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



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

Table12. Gender & Employment Contingency table

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

P (male or full-time employment) = P(male) + P(full time employment) - P(male and full time employment)

Probability that a randomly selected student is either a male or has a full-time employment is: 0.5161

Probability that a randomly selected student is either a male or has a full-time employment is: **51.61%**

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

P (international business or management in female) = P (international business| female or management|female)

= P (international businees|female) + P(management|female) – P international business| female and management|female)

= 4/33+4/33 = 0.2424

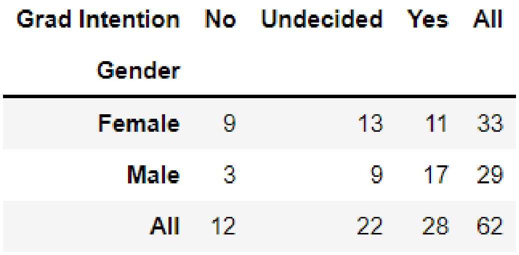
The P (international|girl and management|female)=0 G

Because international business and management are mutually exclusive events.

The Conditional Probability that a given female student is randomly chosen, she is majoring in International Business or Management = **24.24%**

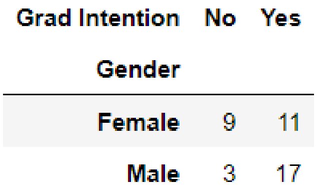
2.6. Construct a contingency table of Gender and Intent to Graduate at 2 levels (Yes/No). The Undecided students are not considered now and the table is a 2x2 table. Do you think the graduate intention and being female are independent events?

From Gender Grad Intention Contingency table:



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

Constructing a Contingency table of Gender and Intent to Graduate at 2 Levels (Yes/No):



Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

Table13. Contingency table of Gender & Intent to Graduate at 2 Levels

Do you think graduate intention and being female are independent events?

For Being Independent,

P(Female and graduate intent)= P(Female intersection graduate intent) = P(Female)*P(graduate intent|Female)

For Being independent,

P(Female and graduate intent) = P(Female)*P(graduate intent)

to test whether it is dependent or independent we have,

If P(graduate) = P(graduate intent|Female) # Independent

If P(graduate) != P(graduate intent|Female) # Not independent

prob_graduate_intent = 28/40

prob_Female = 20/40

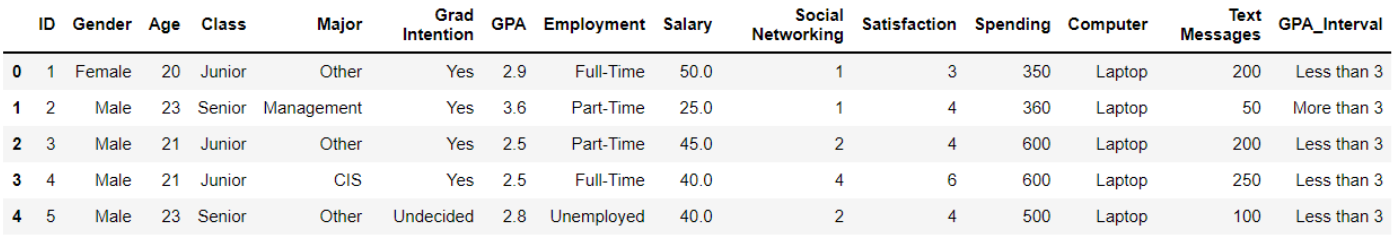
prob_Female_graduate_intent = 11/28

From the above information, we can conclude that Graduate intention and being Female 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?

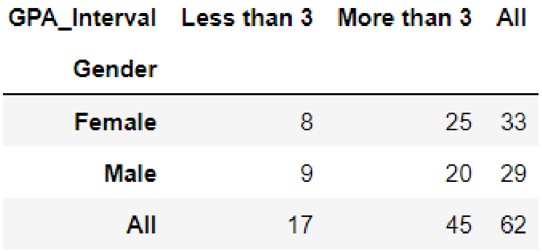
Here, we can cut the GPA into bins (0,2.99,5) and label them with Less than 3 and More than 3 and make a new column GPA_Interval.



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

Table14.sample data with GPA_Interval

Contingency table for Gender & GPA_Interval:



GPA_Interval	Less than 3	More than 3	All
Gender			
Female	8	25	33
Male	9	20	29
All	17	45	62

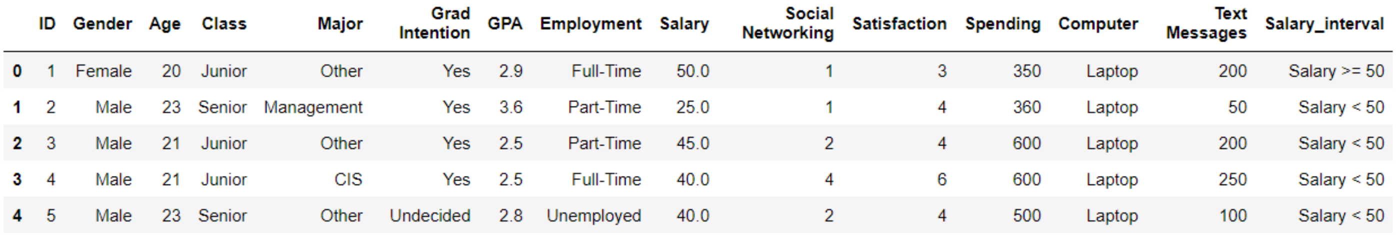
Table15. Contingency table for Gender & GPA_Interval

P(GPA Less than 3) = Total Less than 3 / All total = 17/62 = 0.2742

The Probability that a randomly chosen student has GPA Less than 3: **27.42%**

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.

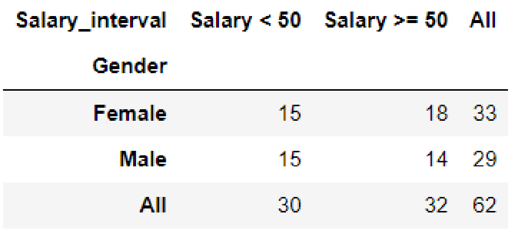
Cut the Salary into bins (0,49.99,100) and label them as 'Salary <50' & 'Salary >=50' and make a new column as Salary_Interval



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

Table16. Sample data with salary_Interval

Contingency table for Gender & Salary_Interval:



Salary_interval Gender	Salary < 50	Salary >= 50	All
Female	15	18	33
Male	15	14	29
All	30	32	62

Table 17. Contingency table for Gender & Salary Interval

P(earns 50 or more given male) = 14/29 = 0.4828

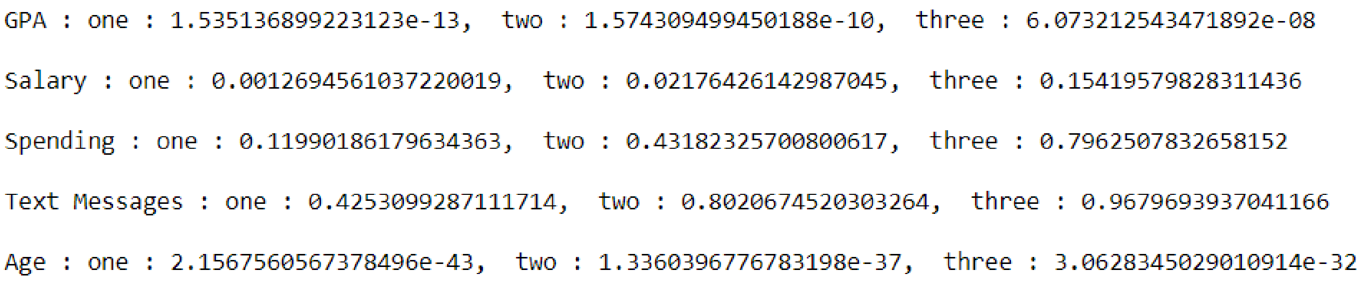
P(earns 50 or more given female) = 18/33 = 0.5455

Conditional Probability that a randomly selected male earns 50 or more: 48.28%

Conditional Probability that a randomly selected female earns 50 or more: 54.55%

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

The Normal Distribution follows the empirical rule. If the variables follow the empirical rule then we can say that it follows normal distribution otherwise not follows normal distribution.
We can check whether the given variables follow empirical rule or not using the information below:



```
GPA : one : 1.535136899223123e-13,  two : 1.574309499450188e-10,  three : 6.073212543471892e-08

Salary : one : 0.0012694561037220019,  two : 0.02176426142987045,  three : 0.1541957982831l436

Spending : one : 0.11990186179634363,  two : 0.43182325700800617,  three : 0.7962507832658152

Text Messages : one : 0.4253099287111714,  two : 0.8020674520303264,  three : 0.9679693937041166

Age : one : 2.1567560567378496e-43,  two : 1.3360396776783198e-37,  three : 3.0628345029010914e-32
```

From the above information, we can see that empirical rule is not followed by any of the variables. But the variable Text message is a bit close to normal distribution where the spending and other variables are too far to follow normal distribution.

In GPA, the distribution is left skewed and does not follow the normal distribution.

In Salary, the distribution is right skewed and does not follow the normal distribution.

In Spending, the distribution is slightly right skewed and not following normal distribution.

In, Text Messages, the distribution is slightly right skewed and a bit close to follow the normal distribution.

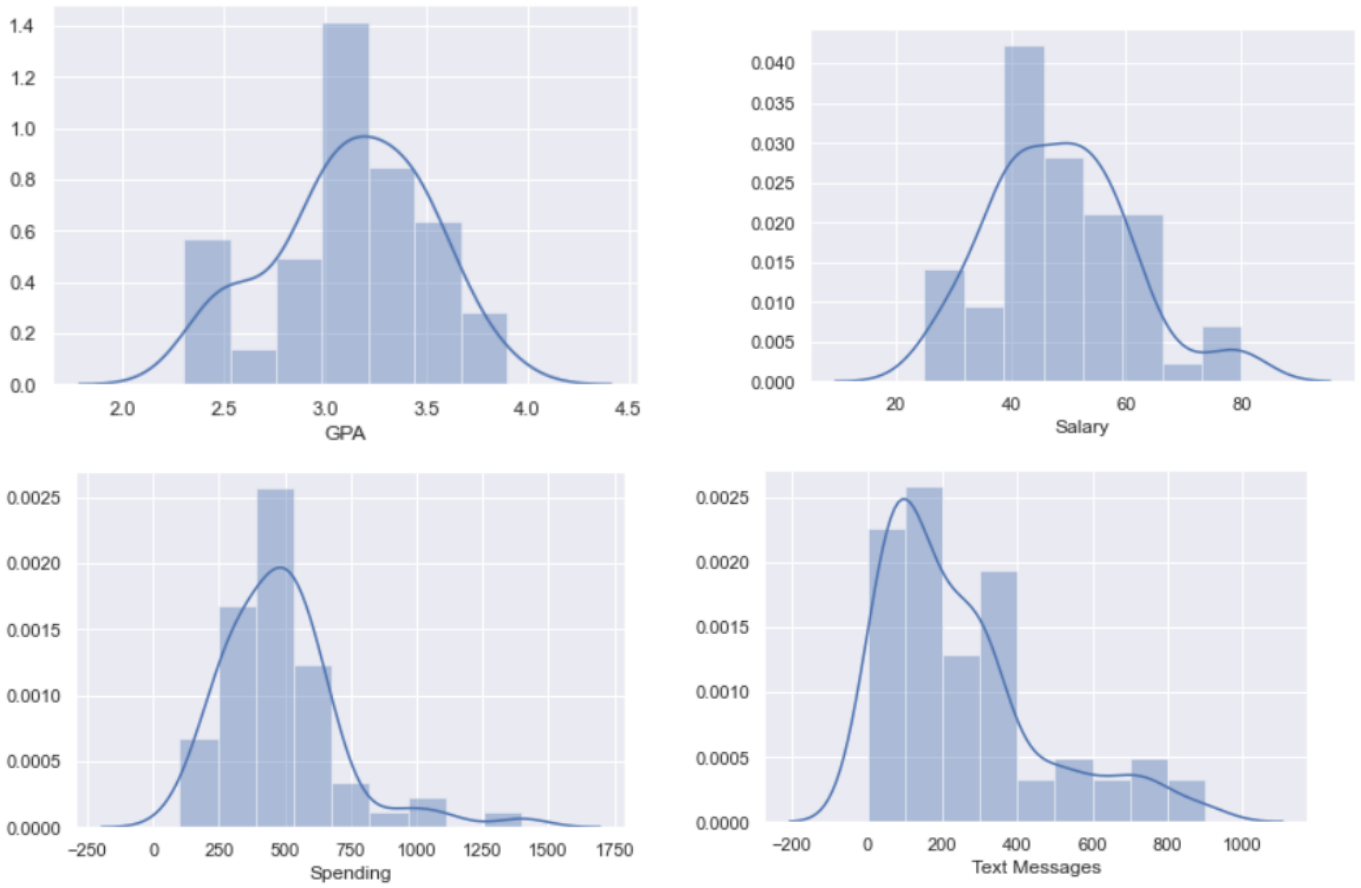


Fig7.Plot to find the spread & Normal distribution

ABC asphalt Shingles A & B

PROBLEM STATEMENT (ABC asphalt Shingles A & B):

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 pounds per 100 square feet.

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

For the two samples,

Ho: Mu <=0.35

Ha: Mu>0.35

The Population Standard Deviation is not mentioned in the data. Hence, we use **1 sample T-test** to test the null hypothesis Individually.

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.

Ho: Mu <=0.35

Ha: Mu>0.35

The company would like to show that the mean moisture content is less than 0.35 pounds per 100 square feet, it is actually 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.

Alternative hypothesis (HA): mean moisture content > 0.35

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

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

H0: mean moisture content <=0.35 HA : mean moisture content > 0.35

For 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

The null hypothesis is the current status or status quo. The company's current status is that the mean moisture content is less than 0.35. Their current status quo needs to be refuted on the basis of very strong evidence. The company does this test on the basis of the assumption that their production process is under control. That is hinted as to monitor the amount of moisture present, the company conducts moisture tests

Hence H0: mean moisture content =< 0.35 (since it is current claim) Ha: mean moisture content > 0.35. The company is monitoring its quality control. This is a peculiar case of claim and status quo being the same i.e. Ho <= 0.35 while the test is carried out is to test for the opposite i.e. Ha > 0.35

T-Test:

Alpha = 0.05

From one-sample T-Test we get,

t-statistics for sample A: -1.474

P- value for sample A: 0.07477

Hence, P value for sample A >0.05, we did not find enough evidence to reject the null hypothesis.

So, we fail to reject null hypothesis

We can say that, with 95% confidence level the mean moisture 100 sq.feet is less than 0.35 pounds for sample A.

t-statistics for sample B = -0.310033

P- value for sample B = 0.002090

Hence, P value for sample A <0.05, we find enough evidence to reject the null hypothesis. We reject the null hypothesis, we can say that with 95% confidence level that mean moisture per 100 sq.feet is more than 0.35 pounds for sample B.

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?

The Population mean For Shingles A & B are equal, we can conclude this using the 2-sample T-test

For the given samples,

Ho: MuA equal to MuB

Ha: MuA not equal to MuB

Now, we need to compare the two independent sample.

So, we are using **2-sample T-Test.**

T-Test:

Alpha = 0.05

t-statistic: 1.28962

p-value: 0.20174

Since, P-value is >0.05, we fail to reject the null hypothesis and we can say that with 95% confidence level that mean moisture per 100 sq.feet for sample A and B are equal.

Test Assumptions when running a 2- Sample T-test, the basic assumptions are the distributions of two populations are normal, and the variances of the two distributions are same. If these assumptions are not likely met, we should use another testing method.

THE END!!