Data Analysis Report On

“Wholesale Customers”

“Student News Service at Clear Mountain State University (CMSU)”

&

“A & B shingles”

Name: - Ruchita Abrange

DSBA Online

April ’2021

Date: 13/06/2021

Table of Contents

Contents:-

"Wholesale Customers"

Questions

Table of Contents

Contents:-

"Student News Service at Clear Mountain State University (CMSU)"

Questions

Table of Contents

Table of Contents

Contents:-

"A & B shingles"

Questions

Data Analysis Report On

“Wholesale Customers”

About Us

Executive Summary:

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.

Introduction:

- The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and other parameters.
- 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).

Data Description:

1. Buyer/Spender: - Numbers of buyers or spender (Numerical numbers).
2. Channel: - Type of channel through which business is done (Hotel, Retail).
3. Region: - In which region the business is done (Other, Lisbon, Oporto).
4. Fresh: - Sale (Counts in numerical).
5. Milk: - Sale (Counts in numerical).
6. Grocery: - Sale (Counts in numerical).
7. Frozen: - Sale (Counts in numerical).
8. Detergent_Paper: - Sale (Counts in numerical).
9. Delicatessen: - Sale (Counts in numerical).

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

Table 1. Dataset Sample

Dataset has 9 variables with 2 different types of the Channel and 3 different types of Region. 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).

Exploratory Data Analysis:

Let us check the types of variables in the data frame.

```
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)
```

The number of rows are 440 and the columns are 9 and the total number of elements of the dataset are 396 0. Out of 8, 2 columns are of object type and rest are of integer data type.

Check for missing values in the dataset:

```
0   Buyer/Spender      440 non-null
1   Channel            440 non-null
2   Region             440 non-null
3   Fresh              440 non-null
4   Milk               440 non-null
5   Grocery            440 non-null
6   Frozen             440 non-null
7   Detergents_Paper   440 non-null
8   Delicatessen       440 non-null
```

Correlation Plot:

Figure 2. Correlation Heatmap

From the correlation plot, we can see that various attributes of the Buyer/spender and delicatessen are highly co rrelated to each other. An effect score closer to 0 translates to there being no relationship. A score closer to 1 or -1 is a positive or negative relationship. A perfect score of 1 is a direct correlation.

PairPlot:

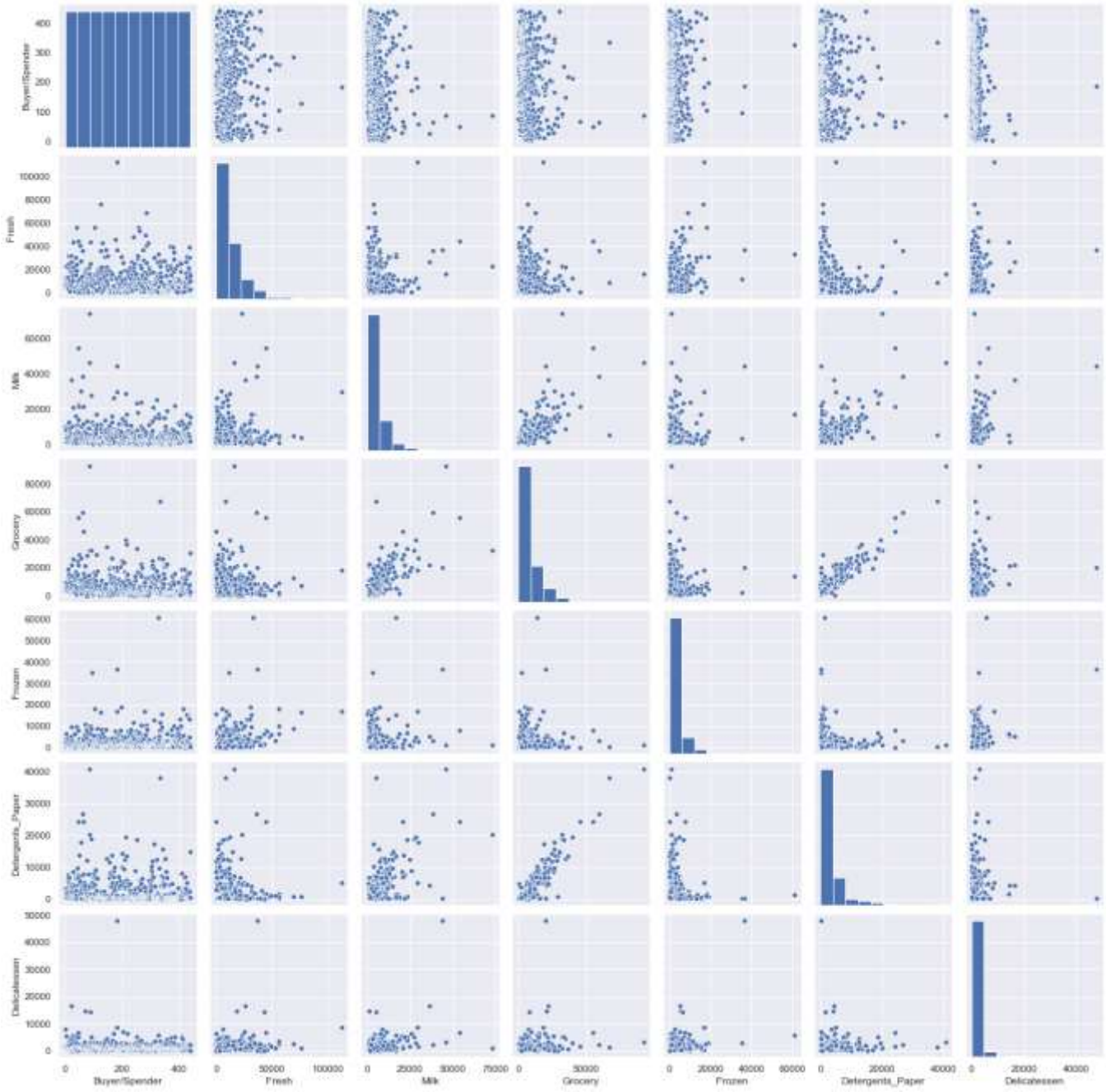


Figure 3. Pairplot

Pairplot shows the relationship between the variables in the form of scatterplot and the distribution of the variable in the form of histogram.
From the graph, we can see that there is positive linear relationship between variables like Grocery and Detergents_Paper. This says that customers often buys both.
From the histogram we can see that the price of the whole dataset is right skewed except for Buyer/Spender.

Questions:

Q1.1 Use methods of descriptive statistics to summarize data. Which Region and which Channel spent the most? Which Region and which Channel spent the least?

Ans1.1 Measure of Central Tendency are Mean, Median, mode, Measure of Dispersion - Range, IQR, Standard Deviation. The descriptive statistics of data is as follows:-

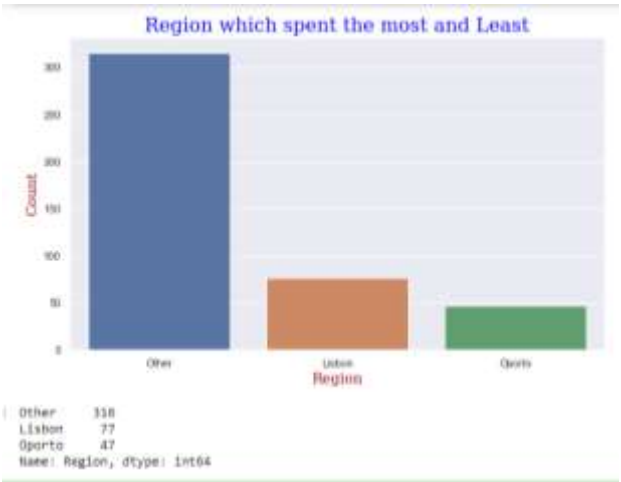
:

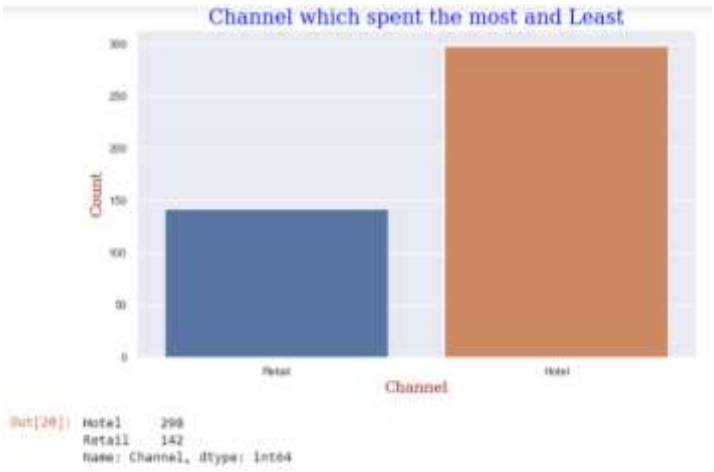
	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440.000000	440	440	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
unique	NaN	2	3	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	Hotel	Other	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	298	316	NaN	NaN	NaN	NaN	NaN	NaN
mean	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

We can infer that more frequency of buying is in other region through Hotel channel.

```
Region
Lisbon      2386813
Oporto      1555088
Other      10677599
Name: Spends, dtype: int64

Channel
Hotel      7999569
Retail     6619931
Name: Spends, dtype: int64
```

- The region which spends more is other and the Channel which spends more is Hotel. The region which spe nds least is Oporto and the Channel which spends least is Retail.

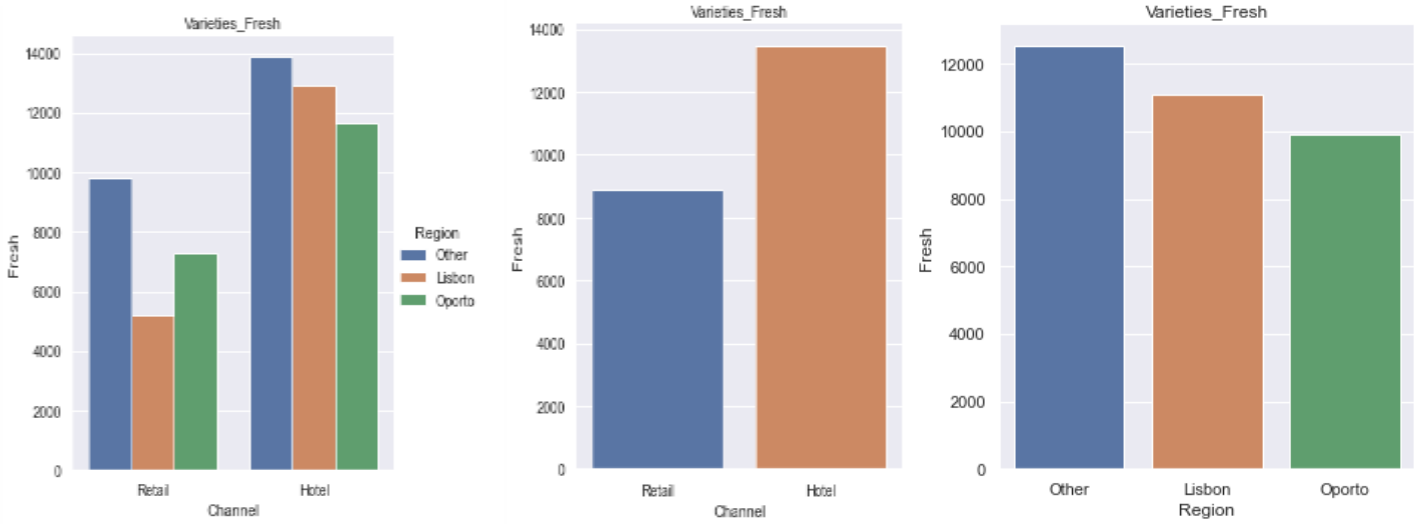




Questions:

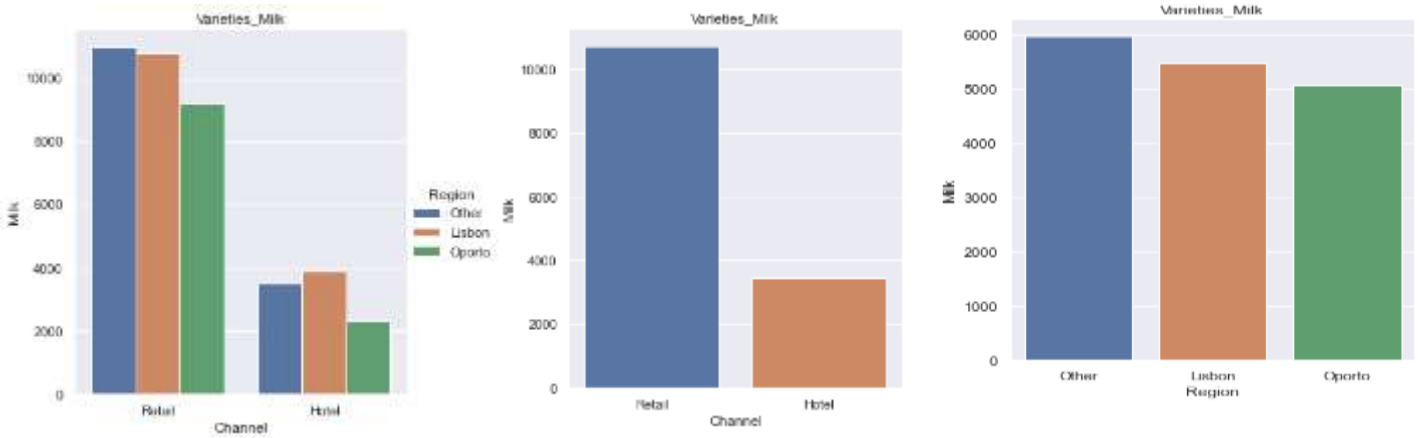
Q1.2 There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.

Ans1.2) For Varieties Fresh across Region and Channel



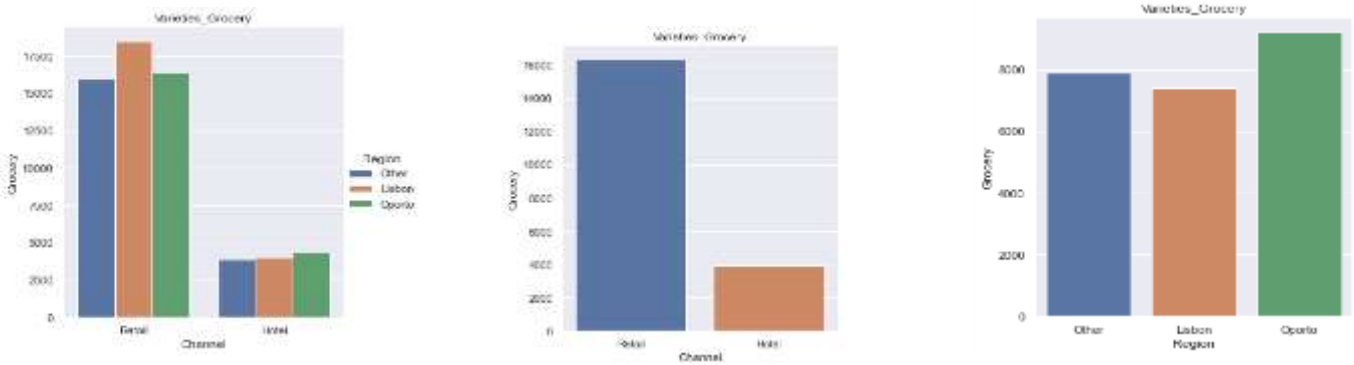
From above plot we can infer that Fresh is sold more in Hotel channel and in Other Region.

- **For Varieties Milk across Region and Channel**



From above plot we can infer that Milk is sold more in Retail channel and in Other Region.

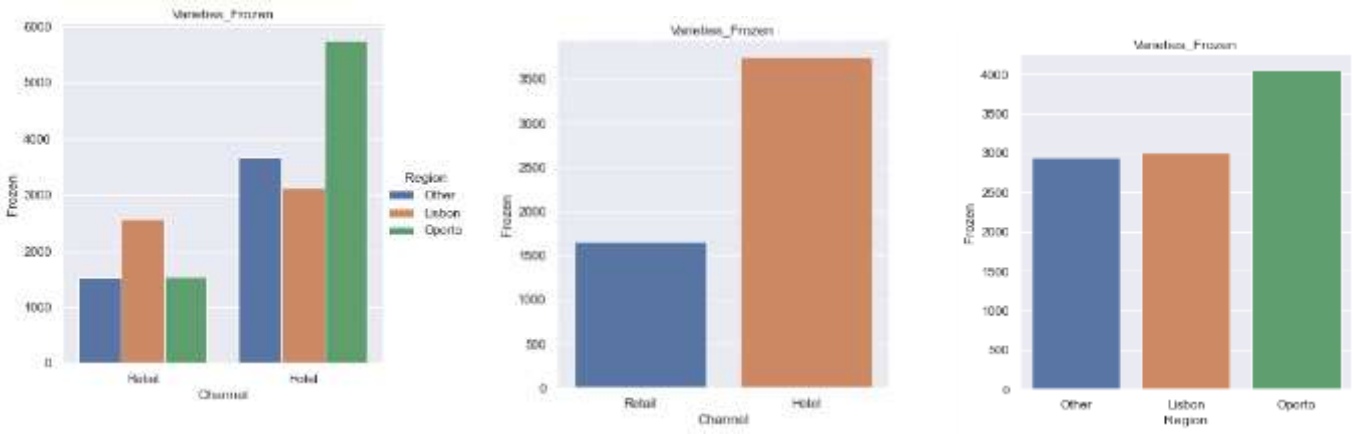
- **For Varieties Grocery across Region and Channel**



From above plot we can infer that Grocery is sold more in Retail channel and in Oporto Region.

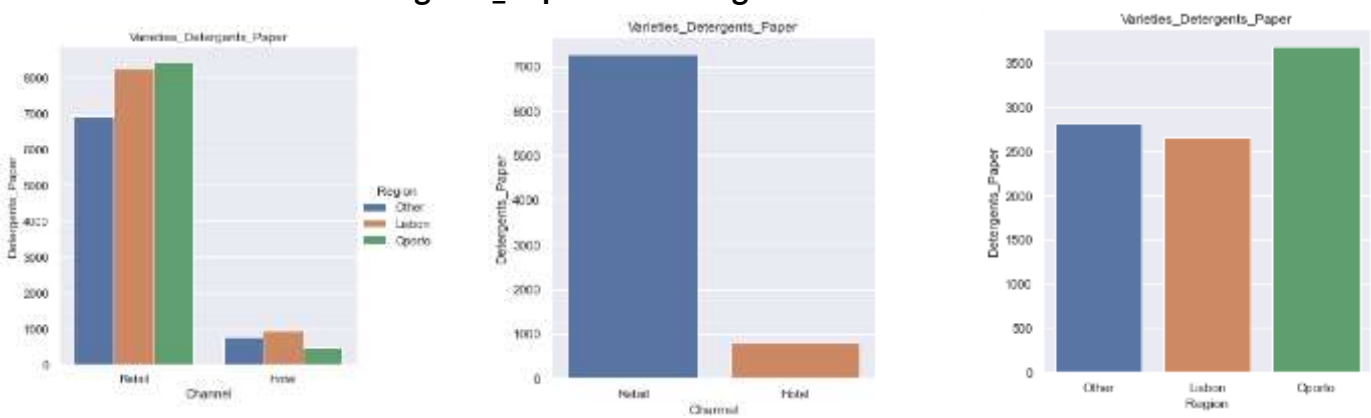
Questions:

- **For Varieties Frozen across Region and Channel**



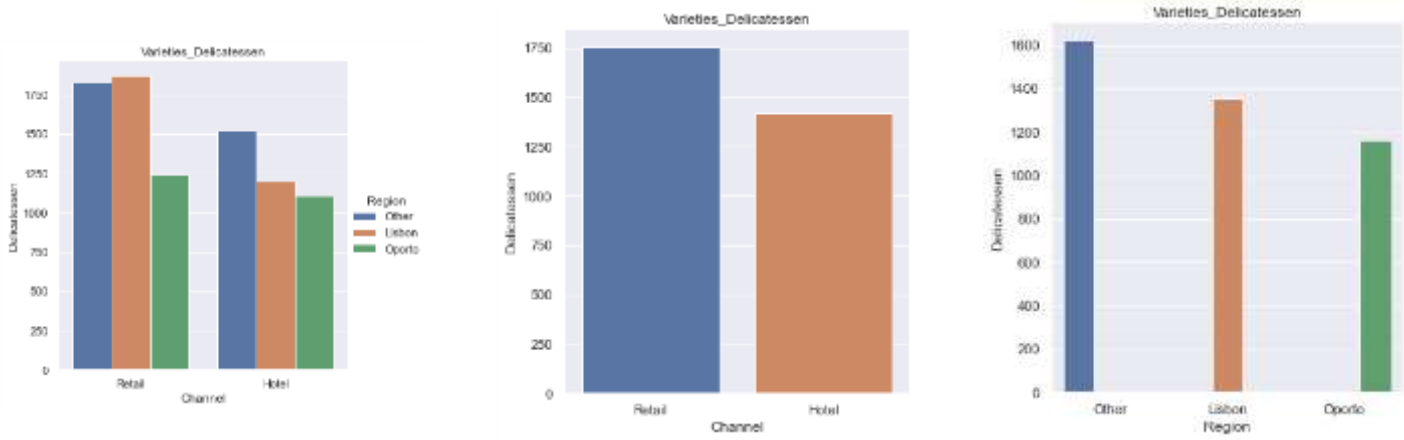
From above plot we can infer that Frozen is sold more in Hotel channel and in Oporto Region.

- **For Varieties Detergents_Paper across Region and Channel**



From above plot we can infer that Detergents_Paper is sold more in Retail channel and in Oporto Region.

- **For Varieties Delicatessen across Region and Channel**



From above plot we can infer that Delicatessen is sold more in Retail channel and in Other Region.

Questions:

Q1.3 On the basis of a descriptive measure of variability, which item shows the most inconsistent behaviour? W hich items show the least inconsistent behaviour?

Ans1.3) A measure of variability is a summary statistic that represents the amount of dispersion in a dataset. H ow spread out are the values?

- **Standard Deviation of varieties are as follows:-**

```
Fresh               12647.328865
Milk                 7380.377175
Grocery              9503.162829
Frozen               4854.673333
Detergents_Paper     4767.854448
Delicatessen         2820.105937

dtype: float64
```

Fresh is having highest number of standard deviation so it shows the most inconsistent behavior whereas Delicatessen sh ows is the most consistent behavior as its standard of deviation is low.

- Coefficient of variation are as follows :-

Fresh	1.0527196084948245
Milk	1.2718508307424503
Grocery	1.193815447749267
Frozen	1.5785355298607762
Detergents_Paper	1.6527657881041729
Delicatessen	1.8473041039189306

From the above table we can infer that Fresh has the least coefficient of variation so that is consistent whereas Delicatessen has the more number of variation so that is inconsistent.

- Variance are as follows :-

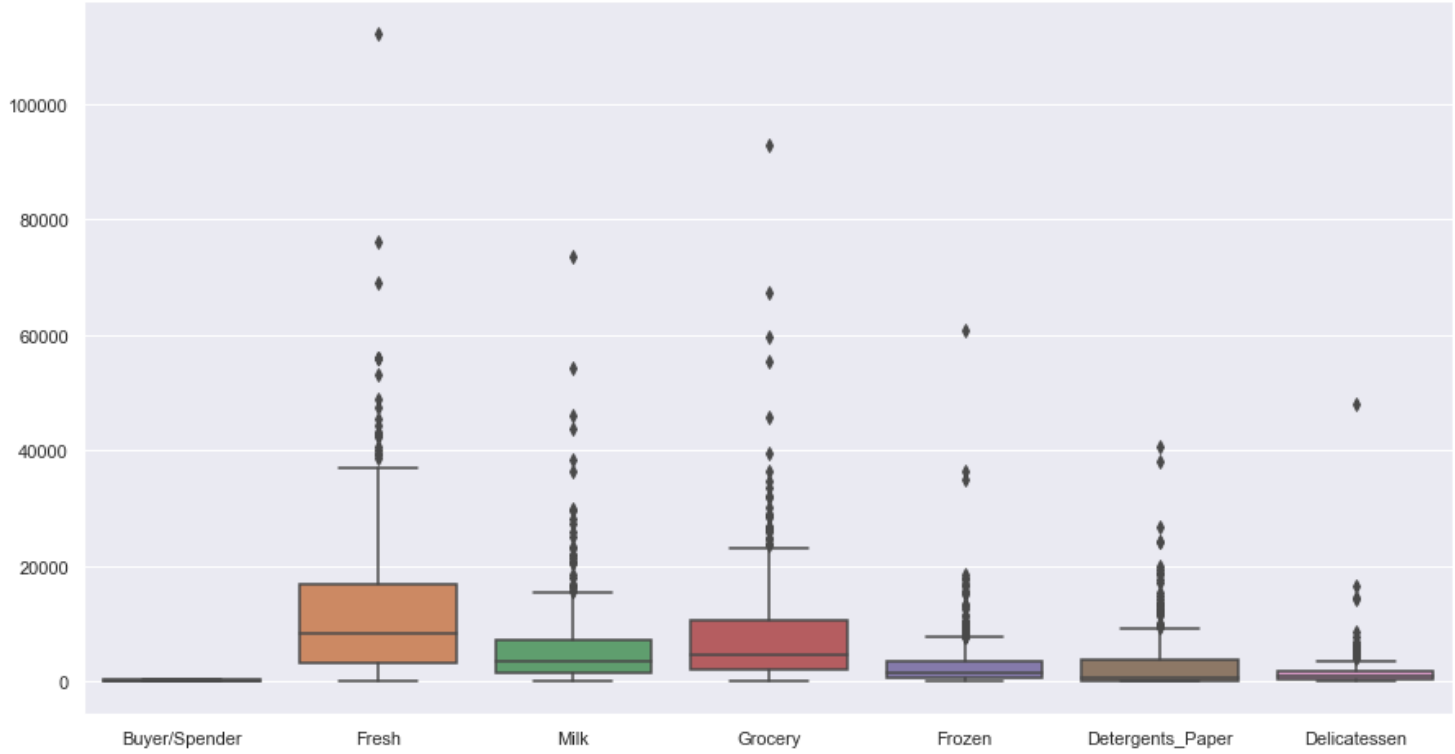
```
Fresh               1.599549e+08
Milk                5.446997e+07
Grocery             9.031010e+07
Frozen              2.356785e+07
Detergents_Paper    2.273244e+07
Delicatessen        7.952997e+06
dtype: float64
```

From the above table we can infer that Fresh has the least variance so that is consistent whereas Grocery has the more number of variance so that is inconsistent

Questions:

Q1.4 Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of d etailed comments.

Ans1.4) An outlier is an object(s) that deviates significantly from the rest of the object collection. It is an abnormal observation during the Data Analysis stage, that data point lies far away from other values. An outlier is an observa tion that diverges from well-structured data.



Yes, there are outliers in all the items. We can see that Grocery has more number of outliers.

Q1.5 On the basis of your analysis, what are your recommendations for the business? How can your analysis he lp the business to solve its problem? Answer from the business perspective?

Ans1.5) From the data exploration, we can suggest/recommend below suggestions to the Wholeseller:-

- It is recommended to sell Grocery and Detergents_Paper together to increase the sale.
- Frozen and delicatessen are rarely brought together. Between other, Lisbon and Oporto, its Oporto buys le ss.
- So frequency to sell items in that region can be made less.
- The wholeseller should concentrate more in Hotel Channel as its buys the most.
- Fresh and Frozen are sold most in Hotel so more number of this items should be pitched to hotel channel
- Frozen, Grocery and Detergents_Paper are sold more in Oporto region so the number of items should be in cresed in that region.
- The dataset is right skewed so the business is going in right direction.

THE END!

Data Analysis

“Student News Service at Clear Mountain State University (CMSU)”

About Us

Executive Summary:

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:

- The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using contingency table to find the Probablity and other parameters.
- The data consists of 62 responses from undergraduates for 14 columns of questions or conditions.

Data Description:

10. ID : - Numbers of CMSU Students who answered the questions (Numerical numbers).
11. Gender : - Type of gender (Male, Female).
12. Age : - Age of Student (18-26 integer).
13. Class: - In which class the students are (Junior, Senior, Sophomore).
14. Major: - In which subject students have done major (Accounting, CIS, Economics/ Finance, International Business, Management, Other, Retailing/Marketing, undecided).
15. Grade Intention: - Are the students intended for grad (Yes, No, Undecided).
16. GPA : - GPA is the average result of all your grades (Counts in float from 2.3 to 3.9).
17. Employment: - Student is fulltime, part-time or unemployed (object).
18. Salary : - Salary of students starting from 25-80 (Counts in float).
19. Social Networking :- Students have 0,1,2,3 or 4 accounts. (Count in integer)
20. Satisfied :- Satisfied in count from 1-6. (Count in integer).
21. Spending :- Spending in 100 – 1200 (Count in integer).
22. Computer :- Student having Laptop, Desktop or Tablet (Object)
23. Text Messages :- Text messages send in count from 0 – 900 (Count in integer).

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 1. Dataset Sample

Dataset has 14 variables with information of the students including the gender, age, class, Major, Grad intention, GPA, Employment, Salary, Social Networking, Satisfaction, Spending, computer, Text Messages. The data consists of 440 students whose survey has been taken on various variables and noted.

Exploratory Data Analysis:

Let us

```
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
```

The number of rows are 440 and the columns are 14 and the total number of elements of the dataset are 868
Out of 14, 6 columns are of integer type, 6 columns are of object type and rest are of float data type.

Check for missing values in the dataset:

```
#   Column             Non-Null Count
---  ------             --------------
 0   ID                 62 non-null
 1   Gender             62 non-null
 2   Age                62 non-null
 3   Class              62 non-null
```

```
4   Major              62 non-null
5   Grad Intention     62 non-null
6   GPA                62 non-null
7   Employment         62 non-null
8   Salary             62 non-null
9   Social Networking  62 non-null
10  Satisfaction       62 non-null
11  Spending           62 non-null
12  Computer           62 non-null
13  Text Messages      62 non-null
```

From the above results we can see that there is no missing value present in the dataset.

Correlation Plot:

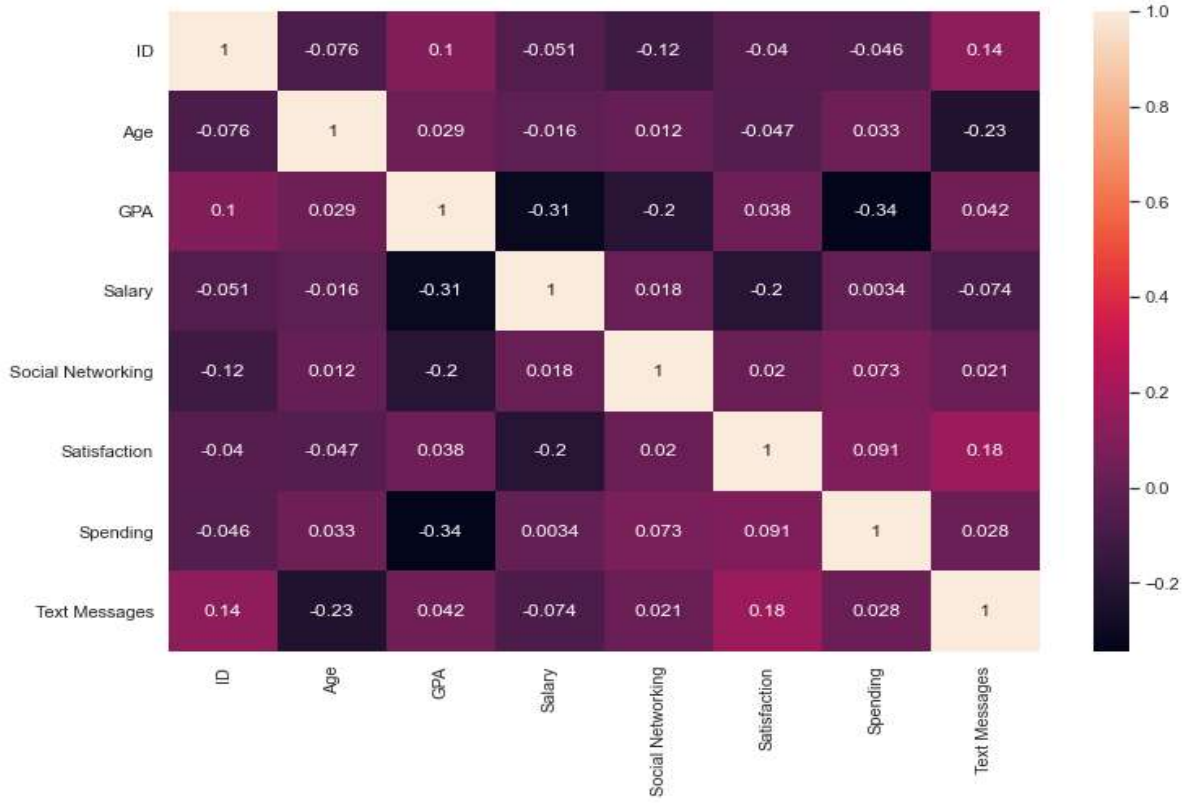


Figure 2. Correlation Heatmap

From the correlation plot, we can see that various attributes of the Text Messages and Id are highly correlated to each other. An effect score closer to 0 translates to there being no relationship. A score closer to 1 or -1 is a positive or negative relationship. A perfect score of 1 is a direct correlation.

PairPlot:

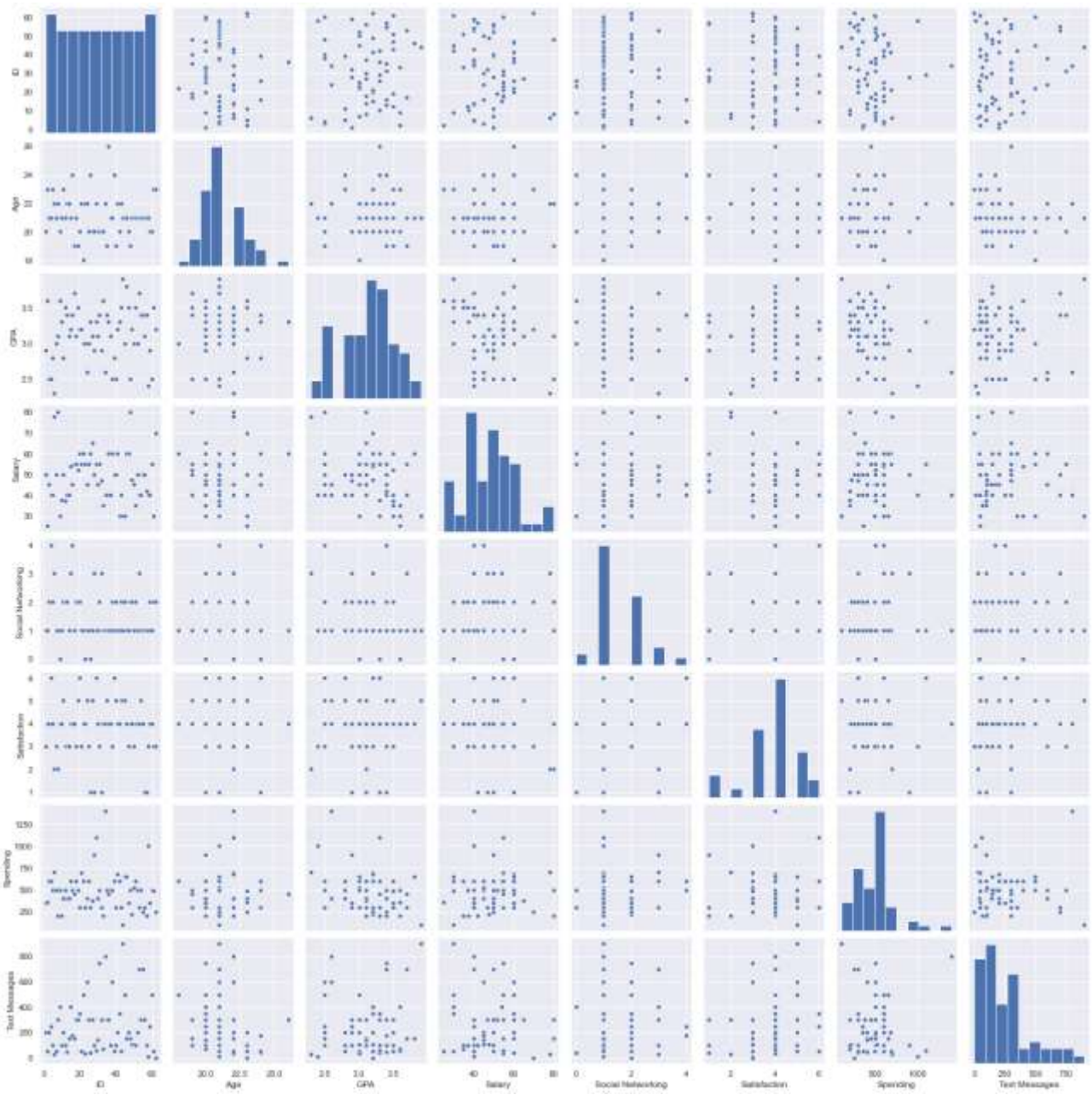


Figure 3. Pairplot

Pairplot shows the relationship between the variables in the form of scatterplot and the distribution of the variable in the form of histogram.
From the graph, we can see that there is relationship between ID and Text messages.

Questions:

Q2.1. For this data, construct the following contingency tables (Keep Gender as row variable)?

Ans 2.1

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

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

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

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

The contingency table for the Gender vs Major, Grad Intention, Employment and Computer has been showed.

Questions:

Q2.2. 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?

Ans2.2.1)

:

	Gender	Female	Male
Major	Employment		
Accounting	Full-Time	0	1
	Part-Time	3	2
	Unemployed	0	1
CIS	Full-Time	0	1
	Part-Time	3	0
Economics/Finance	Full-Time	1	1
	Part-Time	5	3
	Unemployed	1	0
International Business	Part-Time	4	2
Management	Full-Time	0	1
	Part-Time	1	5
	Unemployed	3	0
Other	Full-Time	2	0
	Part-Time	1	3
	Unemployed	0	1
Retailing/Marketing	Full-Time	0	1
	Part-Time	7	3
	Unemployed	2	1
Undecided	Full-Time	0	2
	Part-Time	0	1

Ans 2.2.1) Probability that a randomly selected CMSU student will be male: 0.467 ie. 46.7%

Q2.2.2. What is the probability that a randomly selected CMSU student will be female?

Ans 2.2.2) Probability that a randomly selected CMSU student will be female: 0.532 ie. 53.2%

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.

Ans 2.3.1) Among MALE candidates:
Probability of Accounting: 0.1379 ie. 13.79%
Probability of having a CIS: 0.0344 ie. 3.44%
Probability of being a Economics/Finance: 0.1379 ie. 13.79%

```
Probability of being International Business: 0.0689 ie. 6.89%
Probability of being a Management: 0.20689 ie. 20.68%
Probability of being Other: 0.1379 ie. 13.79%
Probability of being Retailing/Marketing: 0.1724 ie. 17.24%
Probability of being a Undecided: 0.1034 ie. 10.34%
```

From above probablity we can infer that most male students are from Management majors and CIS is the least preferred one.

Q2.3.2 Find the conditional probability of different majors among the female students of CMSU.

```
Among FEMALE candidates:
Probability of Accounting: 0.0909 ie. 9%
Probability of having a CIS: 0.0909 ie. 9%
Probability of being a Economics/Finance: 0.2121 ie. 21.21%
Probability of being International Business: 0.1212 ie. 12.12%
Probability of being a Management: 0.1212 ie. 12.12%
Probability of being Other: 0.0909 ie. 9%
Probability of being Retailing/Marketing: 0.2727 ie. 27.27%
Probability of being a Undecided: 0.0 ie. 0%
```

Ans2.3.2) From above probablity we can infer that most female students are from Retailing / Marketing.

Q2.4. 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.

```
P(Grad Intention ∩ Male) = P (Grad Intention| Male) x P (Male) =  0.27419354
838709675
```

Ans) From the above equation we can infer that a student who is randomly chosen is a male and intends to graduate is 27.4%

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

```
P(No Laptop ∩ Female) = P (No Laptop | Female) x P (Female) =  0.06451612903
225806
```

Ans) From the above equation we can infer that a student who is randomly chosen is a female and intends to graduate is 6.45%

Q 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 a male or has full-time employment?

Ans) The Probability % that a student is a male or has full-time employment is 46.7 %

Questions:

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

Ans) The Probability % that a student is a Female and has either majors in International Business or Managment is 24.2 %

Q2.6. Construct a contingency table of Gender and Intent to Graduate at 2 levels (Yes/No). The Undecided students are not considered now and the table is a 2x2 table. Do you think the graduate intention and being female are independent events?

Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

p(Female n Yes)= P(Female)P(Yes)

Ans) The Probability % that a student is a Female and has grad intention is 24.7%
So Graduate intention and being female are not independent events.

Q 2.7. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages.

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

Ans) From the excel sheet Here 17 students have less than 3 GPA, So P(GPA<3) = 17/62

Probability that a randomly selected CMSU student GPA is less than 3: 27.4%

2.7.2. Find the conditional probability that a randomly selected male earns 50 or more. Find the conditi onal probability that a randomly selected female earns 50 or more.

Ans) From the excel sheet we can make out that there are 14 Males whose Salary is 50 and more than 50. So P(Salary<=50) = 14/29
conditional probability that a randomly selected male earns 50 or more is : 48.27 %

From the excel sheet we can make out that there are 18 Females whose Salary is 50 and more than 50. So

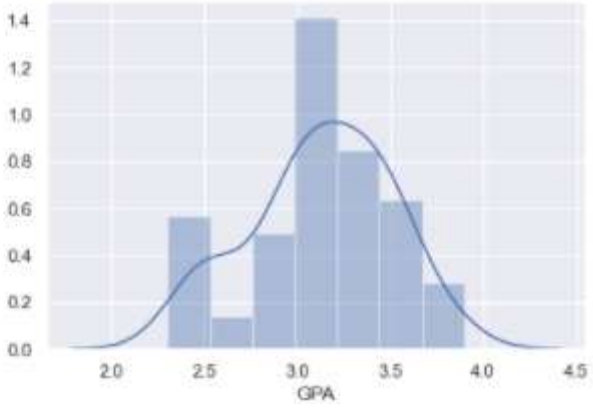
P(Salary<=50) = 18/33

conditional probability that a randomly selected Female earns 50 or more is : 54.54 %

Questions:

Q2.8. 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 not e summarizing your conclusions for this whole Problem 2.

```
In [100]: sns.distplot(dff['GPA']);
```

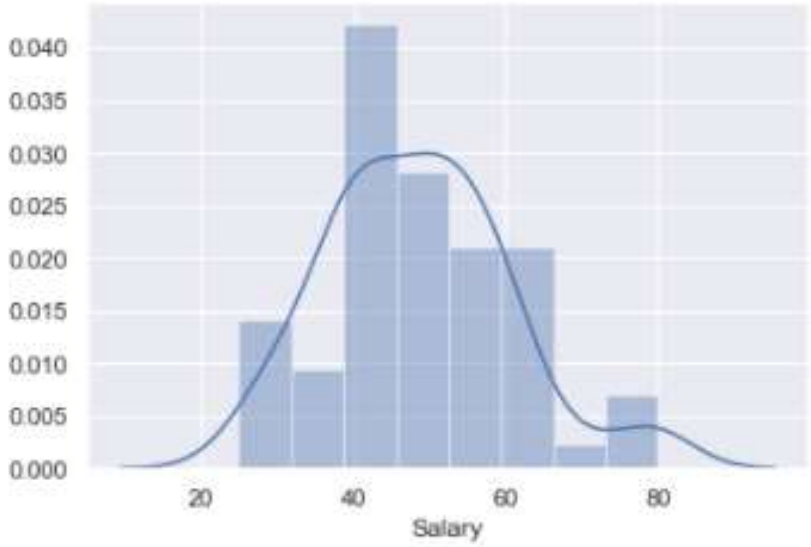



```
In [101]: dff['GPA'].mean()
Out[101]: 3.129032258064516

In [103]: dff['GPA'].median()
Out[103]: 3.1500000000000004

In [104]: dff['GPA'].mode()
Out[104]: 0    3.0
          1    3.1
          2    3.4
          dtype: float64
```

Hence GPA has Normal distribution.

Questions:

```
In [105]: sns.distplot(dff['Salary']);
```

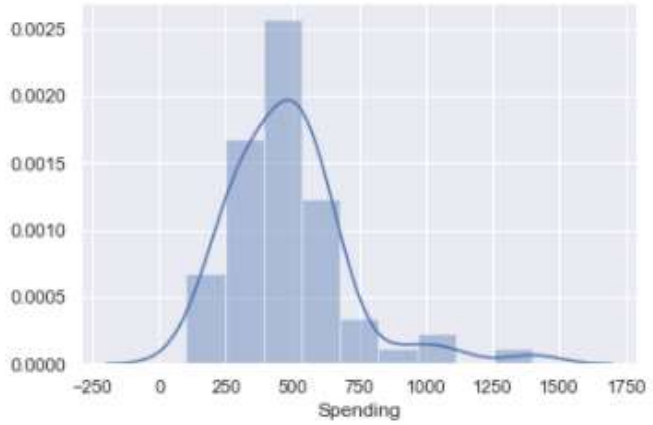



```
In [106]: dff['Salary'].mean()
Out[106]: 48.54838709677419

In [107]: dff['Salary'].median()
Out[107]: 50.0

In [108]: dff['Salary'].mode()
Out[108]: 0    40.0
          dtype: float64
```

Hence Salary has Normal Distribution

Questions:

```
In [109]: sns.distplot(dff['Spending']);
```

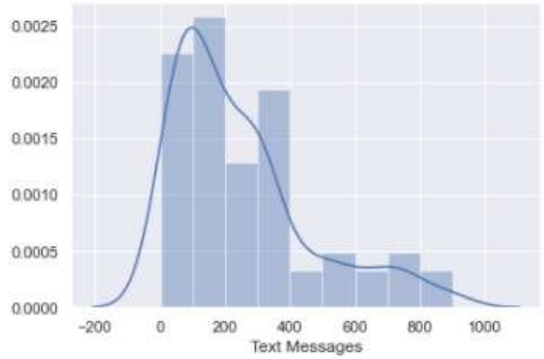



```
In [110]: dff['Spending'].mean()
Out[110]: 482.01612903225805

In [111]: dff['Spending'].median()
Out[111]: 500.0

In [112]: dff['Spending'].mode()
Out[112]: 0    500
          dtype: int64
```

Hence Spending is on little right side so it is right skewed and not normal distribution.

Questions:

```
In [113]: sns.distplot(dff['Text Messages']);
```




```
In [114]: dff['Text Messages'].mean()
Out[114]: 246.20967741935485

In [115]: dff['Text Messages'].median()
Out[115]: 200.0

In [116]: dff['Text Messages'].mode()
Out[116]: 0    300
          dtype: int64
```

Hence Text Messages has right skewed distribution and not normal distribution.

By these we can infer that GPA and Salary are Normally distributed but Spend ing and Text Messages are not following the normal distribution they are right skewed.

THE END!

Data Analysis

“A & B shingles”

About Us

Executive Summary:

An important quality characteristic used by the manufacturers of ABC asphalt shingles is the amount of moisture the shingles contain when they are packaged.

Introduction:

- 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.

Data Description:

1. A :- Includes 36 measurements (in pounds per 100 square feet) of Shingles.
2. B :- Includes 31 measurements (in pounds per 100 square feet) of Shingles.

Sample of 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

Table 1. Dataset Sample

Dataset has 2 variables with information of the Shingles A and B. The data consists of 72 elements. The number of rows are 36 and the columns are 2

Exploratory Data Analysis:

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

The number of rows are 36 and the columns are 2 and the total number of elements of the dataset are 72
Both the column are of float datatype.

Check for missing values in the dataset:

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

From the above results we can see that there is no missing value present in the dataset.

Correlation Plot:

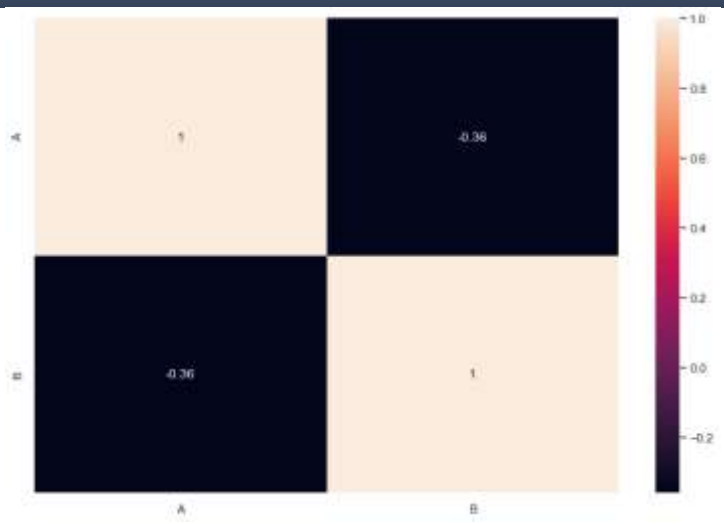


Figure 2. Correlation Heatmap

From the correlation plot, we can see that attributes of the A and B are highly correlated to each other. An effect score closer to 0 translates to there being no relationship. A score closer to 1 or -1 is a positive or negative relationship. A perfect score of 1 is a direct correlation.

Pair Plot:

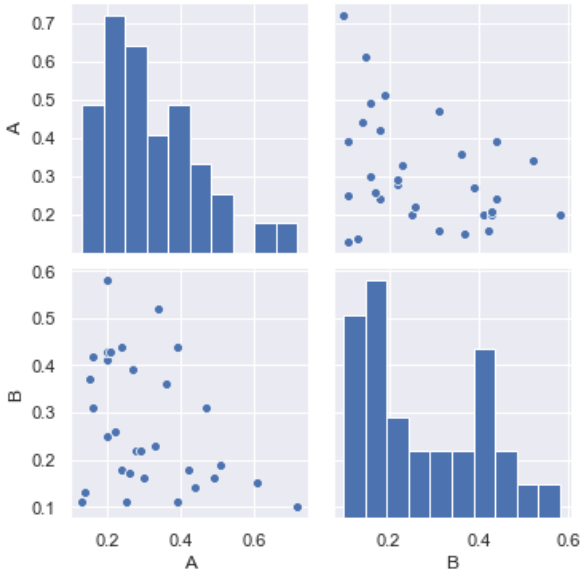


Figure 3. Pairplot

Pairplot shows the relationship between the variables in the form of scatterplot and the distribution of the variable in the form of histogram.
From the graph, we can see that there is relationship between A and B.

Questions:

Q 3.1 Do you think there is evidence that means moisture contents in both types of shingles are within the per missible limits? State your conclusions clearly showing all steps.

Ans 3.1 Solution: H0: The mean moisture content in Shingles is equal to 0.35

H1: company would like to show that the mean moisture content is less than 0.35 pound per 100 square feet.

For A Shingles:

```
In [24]: print('One sample t test \n t statistic:{0}, p value{1} '.format(tstat, pvalue/2))
         One sample t test
          t statistic:-1.4735046253382782, p value0.07477633144907513
```

After calculating the output comes to be
One sample t test
t statistic:-1.4735046253382782, p value0.07477633144907513

Since pvalue > 0.05 so we do not reject Ho. There is no evidence that means moisture contents in A shingles are less than 0.035 pounds per 100 sqft. Since the p value = 0.0748. the population mean moisture content is in fact no less than 0.35 pounds per 100 sqft the probability of observing a sample of 36 shingles that will result in a sample mean moisture content of 0.3167 pounds per 100 sqft or less is 0.0748

For B Shingles:

```
In [31]: print('One sample t test \n t statistic:{0}, p value:{1} '.format(tstat, pvalue/2))
         One sample t test
          t statistic:-3.1003313069986995, p value:0.0020904774003191826
```

After calculating the output comes to be
One sample t test
t statistic:-3.1003313069986995, p value:0.0020904774003191826

Since Pvalue < 0.05, reject Ho. There is enough evidence to conclude that the mean moisture cont ent for Sample B Shingles is not less than 0.35 pounds per 100 square feet.

So at 95% of significance level there is an enough evidence to prove that mean weight is equal to or less than 0.35

Questions:

Q3.2 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?

H0: population mean for shingles A and B are equal
H1: population mean for shingles A and B are not equal
a= 0.05

To perform Hypothesis Testing, the following assumptions must hold,
The variables must follow continuous distribution
The sample must be randomly collected from the population
The underlying distribution must be normal. Alternatively, if the data is continuous, but may not be assumed to follow a normal distributi on, a reasonably large sample size is required. CLT asserts that sample mean follows a normal distribution, even if the population distribut ion is not normal, when sample size is at least 30.
For 2 sample t-test, the population variances of 2 distributions must be equal.

Ans3.2)

```
In [27]: print("tstat={}, pvalue={}".format(round(tstat,3),round(pvalue,3)))
         tstat=1.29, pvalue=0.202
```

After calculating the output comes to be
Tstat=1.29, pvalue=0.202

As we can see from the output that the pvalue > a. So we fail to reject null hypothesis. So we can conclude that population mean for shingles A and B are equal

THE END!