1.3

Problem 1: Wholesale Customers Analysis

Problem Statement:

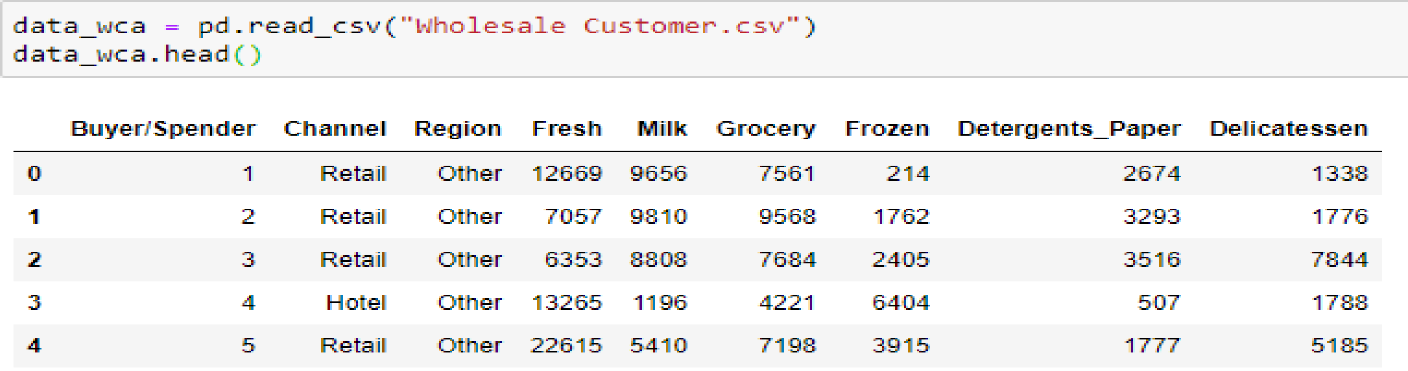
A wholesale distributor operating in different regions of Portugal has information on annual spending of several items in their stores across different regions and channels. The data consists of 440 large retailers' annual spending on 6 different varieties of products in 3 different regions (Lisbon, Oporto, Other) and across different sales channel (Hotel, Retail).

Solution:

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

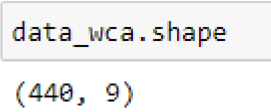
Dataset has below variables

1. 'Buyer/Spender'
2. 'Channel' (categorical variable)
3. 'Region' (categorical variable)
4. 'Fresh'
5. 'Milk'
6. 'Grocery'
7. 'Frozen'
8. 'Detergents_Paper'
9. 'Delicatessen'



```
data_wca = pd.read_csv("Wholesale Customer.csv")
data_wca.head()
```

	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

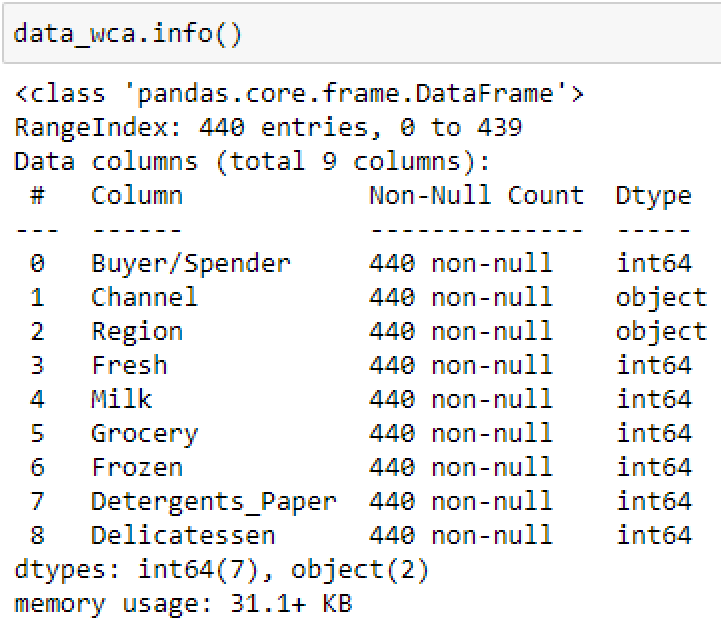


```
data_wca.shape
```

```
(440, 9)
```

Observations are 440 and variables are 9

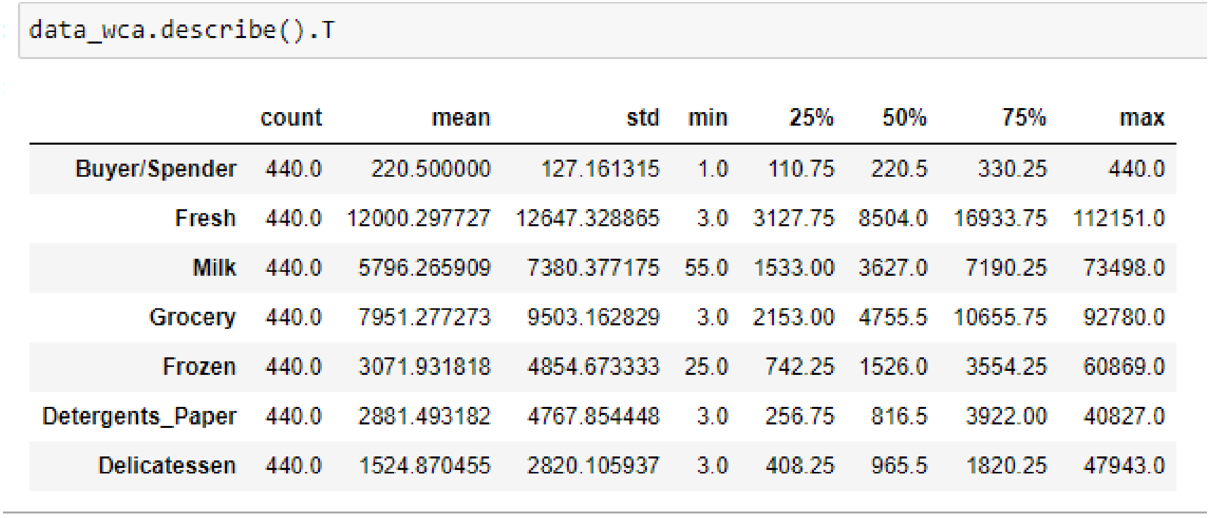
Let’s check datatypes of each variables



```
data_wca.info()
```

```
<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
```

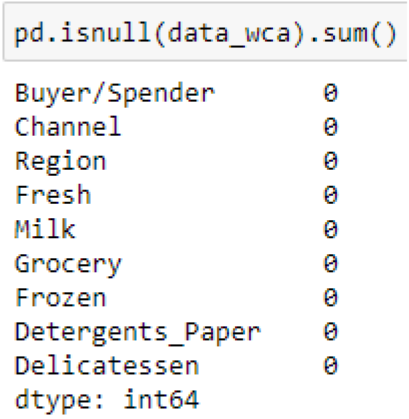
Let’s check descriptive statistics



```
data_wca.describe().T
```

	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

Let’s see if there are any null values present in dataset



```
pd.isnull(data_wca).sum()
```

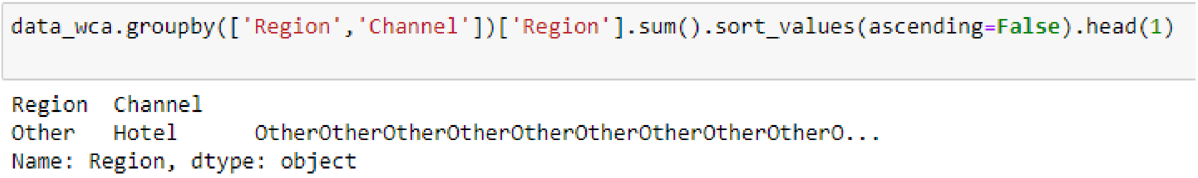
```
Buyer/Spender       0
Channel             0
Region              0
Fresh               0
Milk                0
Grocery             0
Frozen              0
Detergents_Paper    0
Delicatessen        0
dtype: int64
```

Let’s try to answer questions now,

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

I understand there are total four different questions to answer,

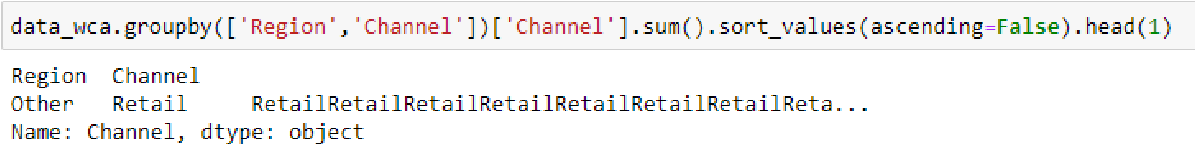
1. Which Region seems to spend more? – ‘Other’



```
data_wca.groupby(['Region','Channel'])['Region'].sum().sort_values(ascending=False).head(1)
```

```
Region  Channel
Other   Hotel      OtherOtherOtherOtherOtherOtherOtherOtherOtherO...
Name: Region, dtype: object
```

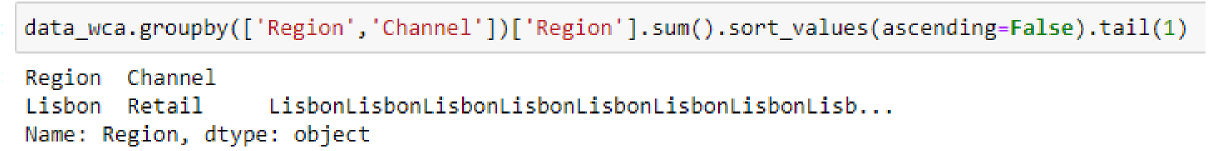
2. Which Channel seems to spend more? – ‘Retail’



```
data_wca.groupby(['Region','Channel'])['Channel'].sum().sort_values(ascending=False).head(1)
```

```
Region  Channel
Other   Retail     RetailRetailRetailRetailRetailRetailRetailReta...
Name: Channel, dtype: object
```

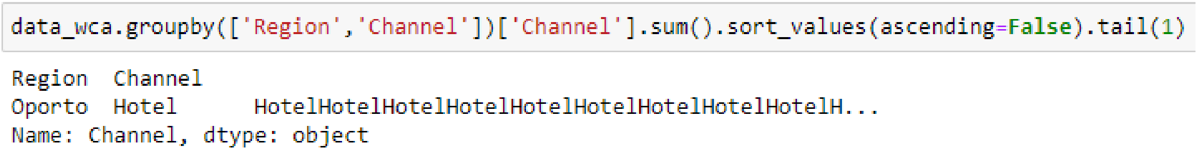
3. Which Region seems to spend less? – ‘Lisbon’



```
data_wca.groupby(['Region','Channel'])['Region'].sum().sort_values(ascending=False).tail(1)
```

```
Region  Channel
Lisbon  Retail     LisbonLisbonLisbonLisbonLisbonLisbonLisbonLisb...
Name: Region, dtype: object
```

4. Which Channel seems to spend less? - ‘Hotel’

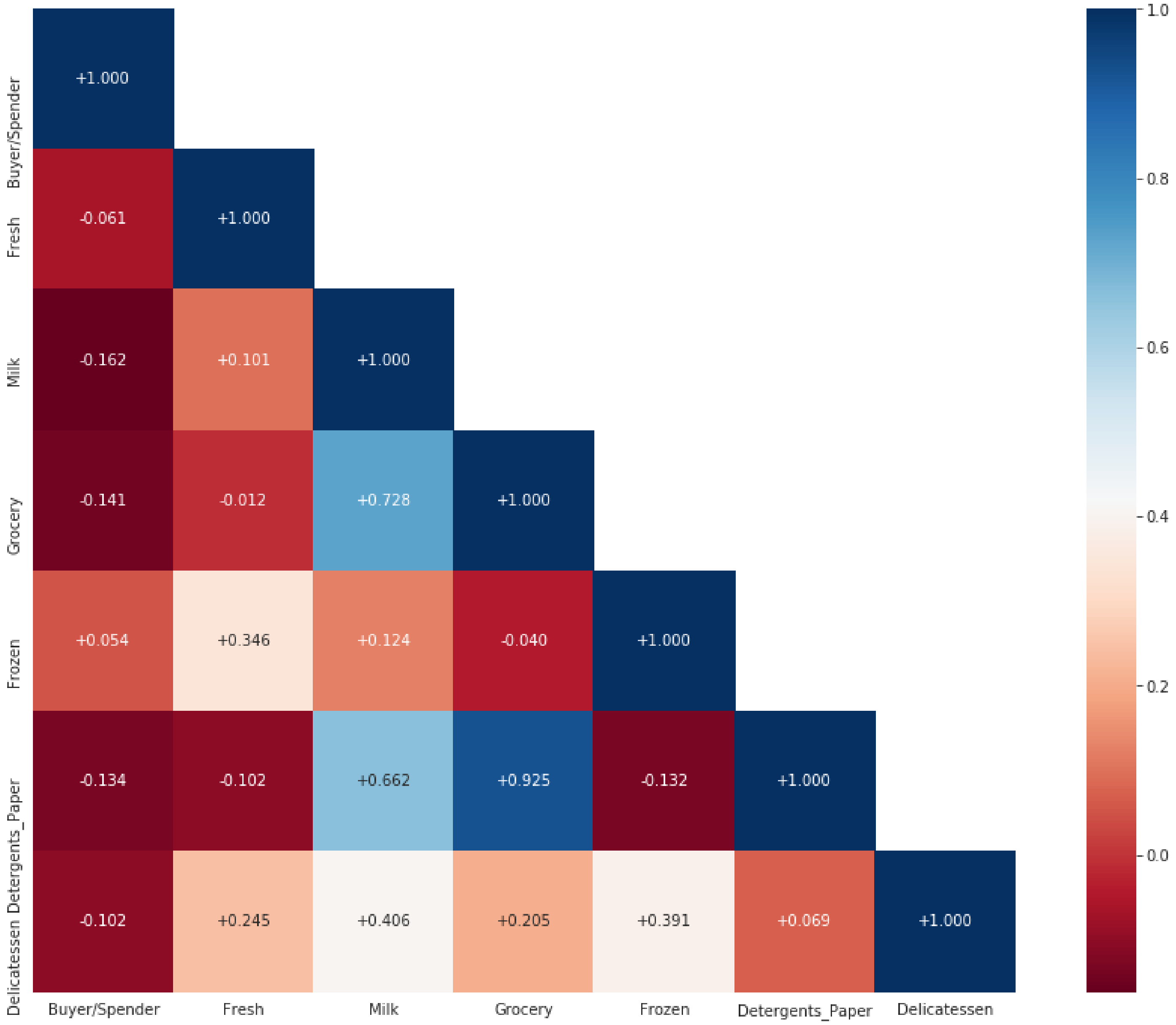


```
data_wca.groupby(['Region','Channel'])['Channel'].sum().sort_values(ascending=False).tail(1)
```

```
Region  Channel
Oporto  Hotel      HotelHotelHotelHotelHotelHotelHotelHotelHotelH...
Name: Channel, dtype: object
```

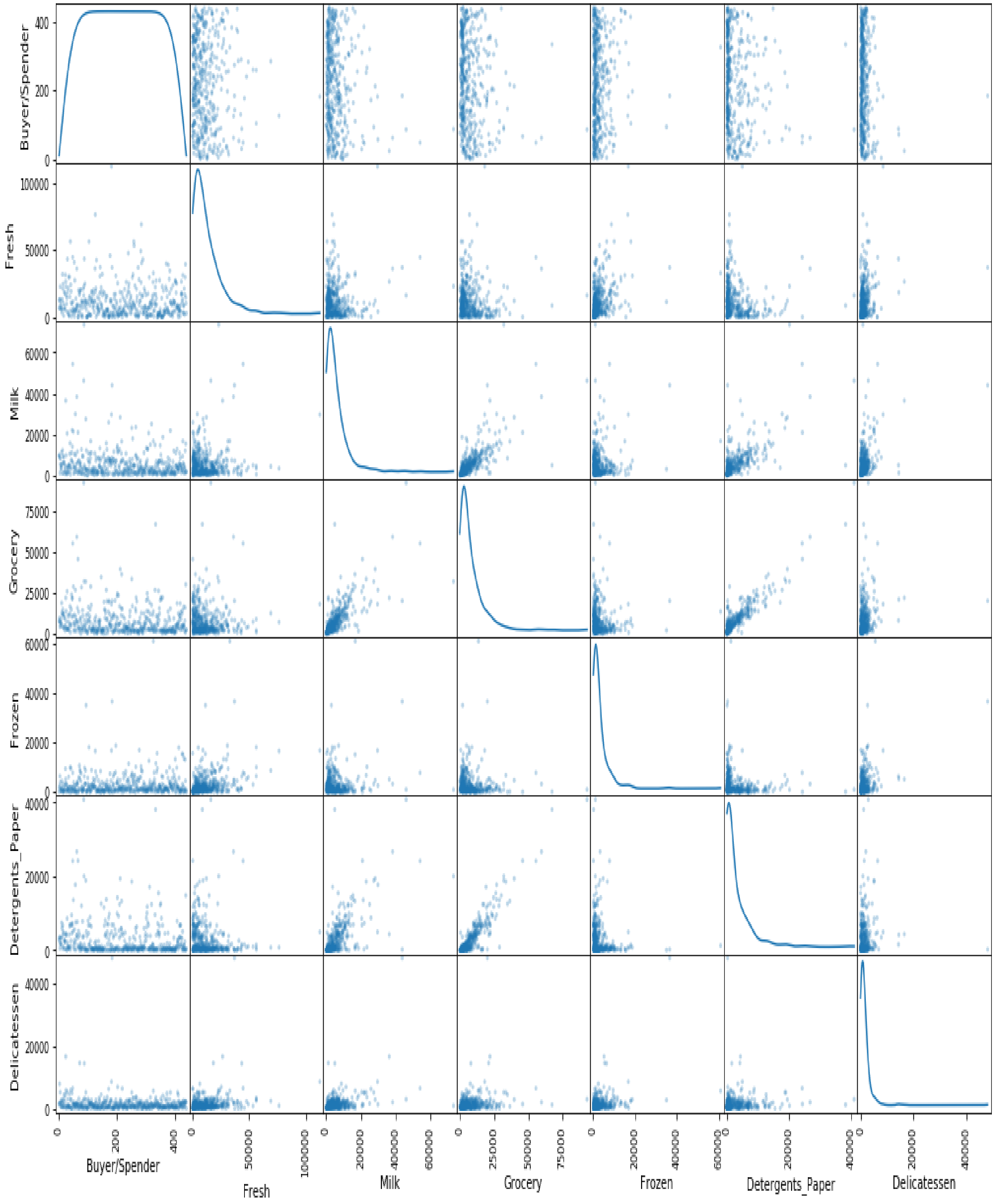
1.2 There are 6 different varieties of items are considered. Do all varieties show similar behaviour across Region and Channel?

Lets see how strong corelation present between variables, check this heatmap

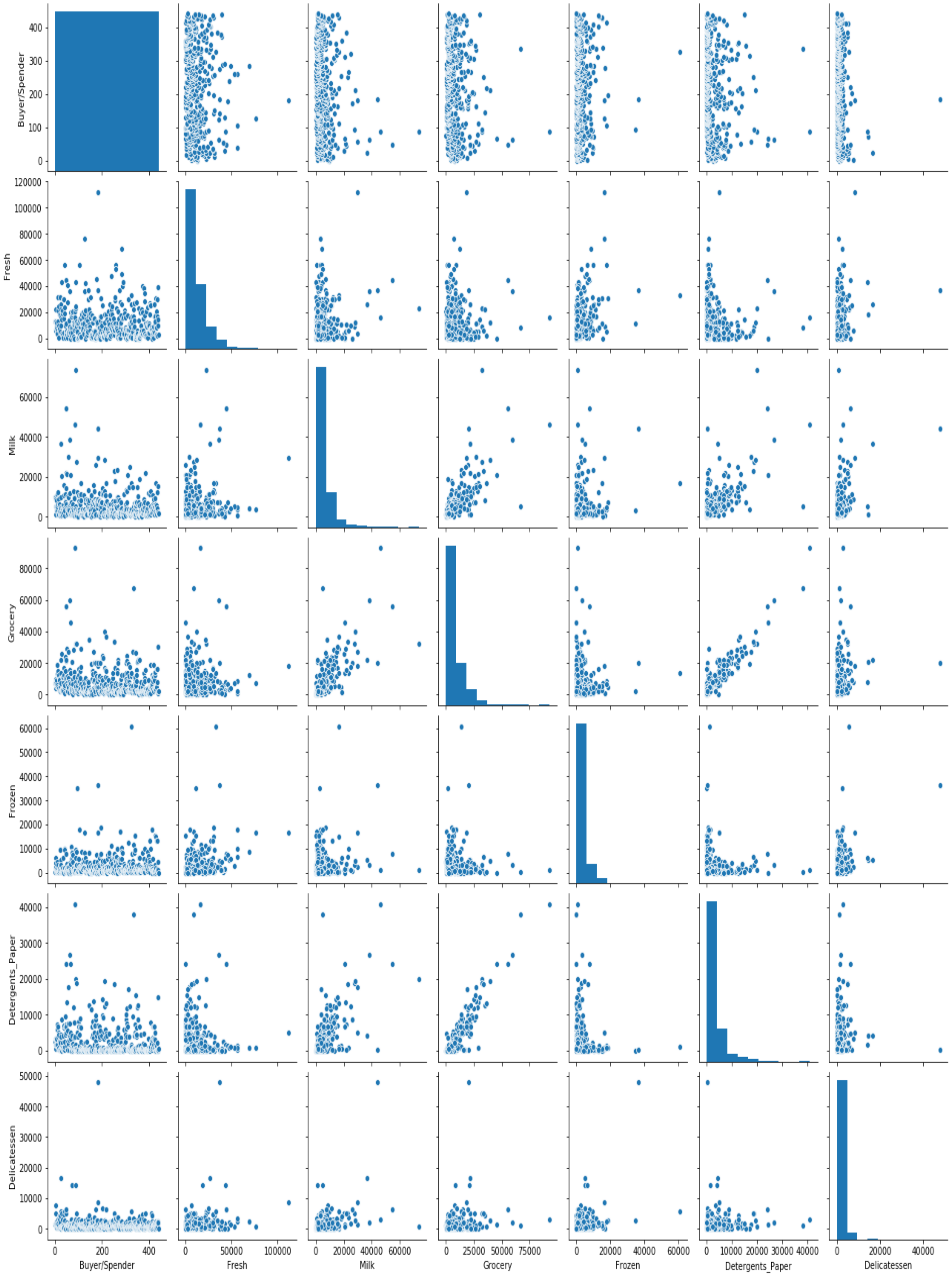


1.3 based on a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?

Hmm…. Let's see scatterplot also

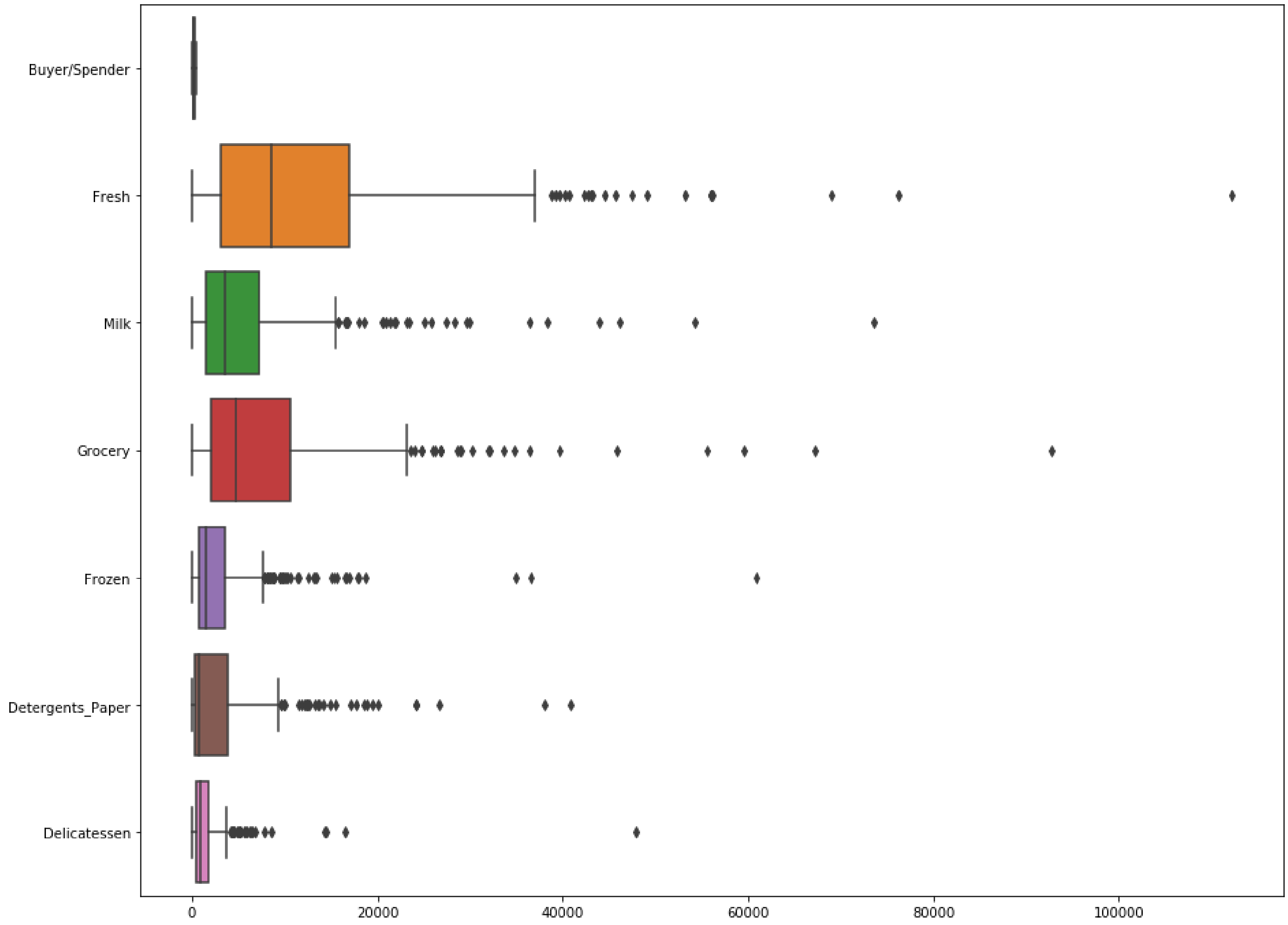


Pair plot,


Buyer/Spender
0
100
200
300
400
Fresh
0
20000
40000
60000
80000
100000
120000
Milk
0
20000
40000
60000
Grocery
0
20000
40000
60000
80000
Frozen
0
10000
20000
30000
40000
50000
60000
Detergents_Paper
0
10000
20000
30000
40000
Delicatessen
0
10000
20000
30000
40000
50000
0
200
400
Buyer/Spender
0
50000
100000
Fresh
0
20000
40000
60000
Milk
0
25000
50000
75000
Grocery
0
20000
40000
60000
Frozen
0
10000
20000
30000
40000
Detergents_Paper
0
20000
40000
Delicatessen

1.4 Are there any outliers in the data?

```
Below items hold outliers
1.Fresh
2.Milk
3.Grocery
4.Frozen
5.Detergents_Paper
6.Delicatessen
```



1.5 based on this report, what are the recommendations?

Recommendation

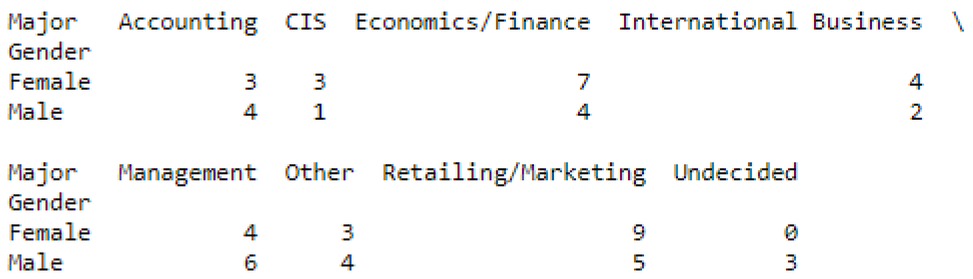
It appears that Grocery and Detergents_Paper have the strongest correlation of the pairs. It also looks like there is some correlation between Detergents_Paper and Milk, and Grocery and Milk. This confirms my suspicion above that Grocery was correlated with some other features that would allow for its value to be predicted with some degree of accuracy. All of the distributions appear to be skewed to the right, with more points hovering closer to the origin and some larger points extending it to the right. The shape of the distributions of Detergents_Paper, Grocery, and Milk are all quite similar.

Problem 2 - (Download **Survey** Data)

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

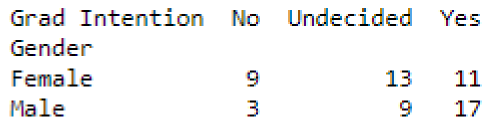
2.1. For this data, construct the following contingency tables (Keep Gender as row variable)

2.1.1. Gender and Major



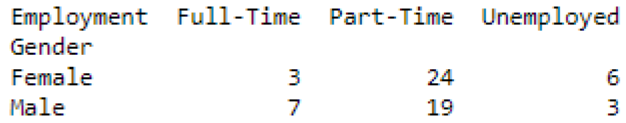
```
Major   Accounting  CIS  Economics/Finance  International Business  \
Gender
Female           3    3                  7                       4
Male             4    1                  4                       2

Major   Management  Other  Retailing/Marketing  Undecided
Gender
Female           4      3                    9          0
Male             6      4                    5          3
```

2.1.2. Gender and Grad Intention



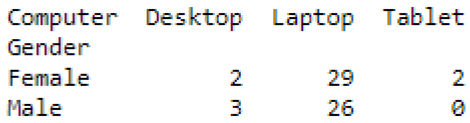
```
Grad Intention  No  Undecided  Yes
Gender
Female           9         13   11
Male             3          9   17
```

2.1.3. Gender and Employment



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

2.1.4. Gender and Computer



```
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:

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

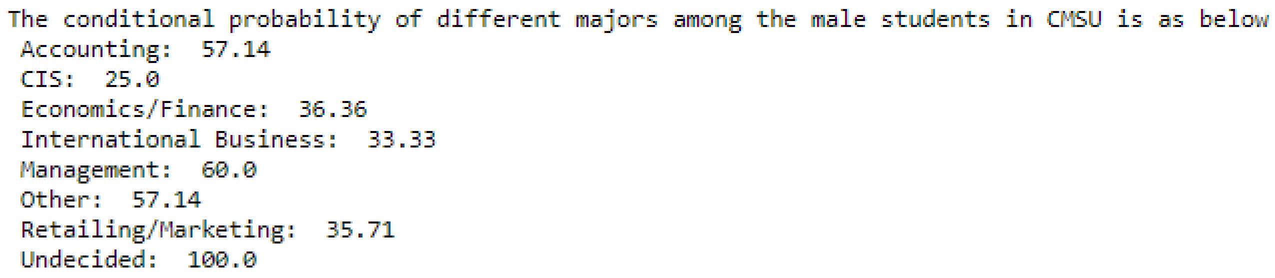
Probability that a randomly selected CMSU student will be male is 46.77 %

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

Probability that a randomly selected CMSU student will be Female is 53.23 %

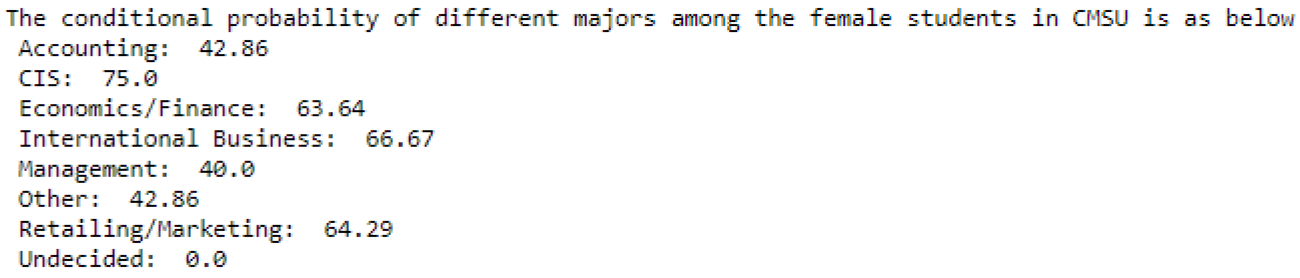
2.3. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

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



```
The conditional probability of different majors among the male students in CMSU is as below
 Accounting:  57.14
 CIS:  25.0
 Economics/Finance:  36.36
 International Business:  33.33
 Management:  60.0
 Other:  57.14
 Retailing/Marketing:  35.71
 Undecided:  100.0
```

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



```
The conditional probability of different majors among the female students in CMSU is as below
 Accounting:  42.86
 CIS:  75.0
 Economics/Finance:  63.64
 International Business:  66.67
 Management:  40.0
 Other:  42.86
 Retailing/Marketing:  64.29
 Undecided:  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.

Probability That a randomly chosen student is a male and intends to graduate is 58.62

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

The probability that a randomly selected student is a female and does NOT have a laptop 12.12

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

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

The probability that a randomly chosen student is either a male or has full-time employment is 14.32

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

The conditional probability that given a female student is randomly chosen, she is majoring in international business or management is 24.0

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?

No, they 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

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

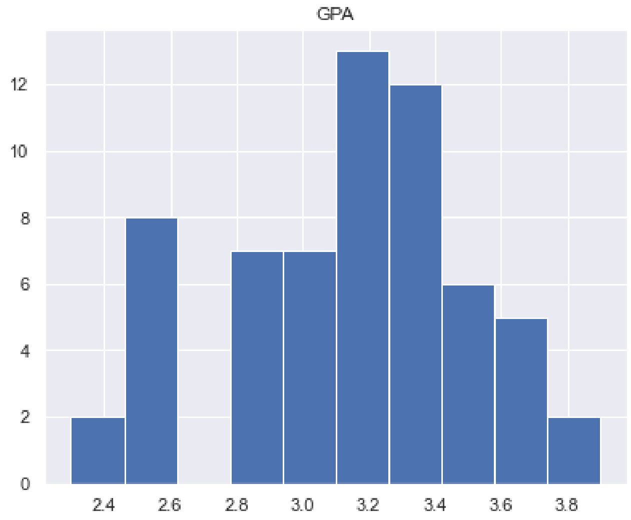
If a student is chosen randomly, the probability that his/her GPA is less than 3 is 27.0

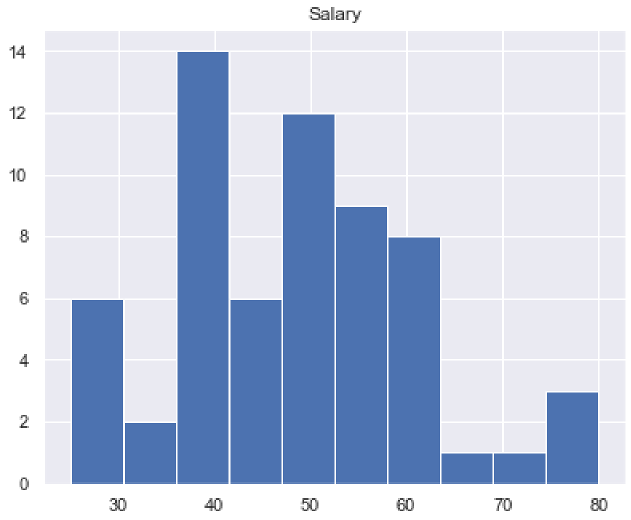
2.6.2. Find the conditional probability that a randomly selected male earns 50 or more. Find the conditional probability that a randomly selected female earns 50 or more.

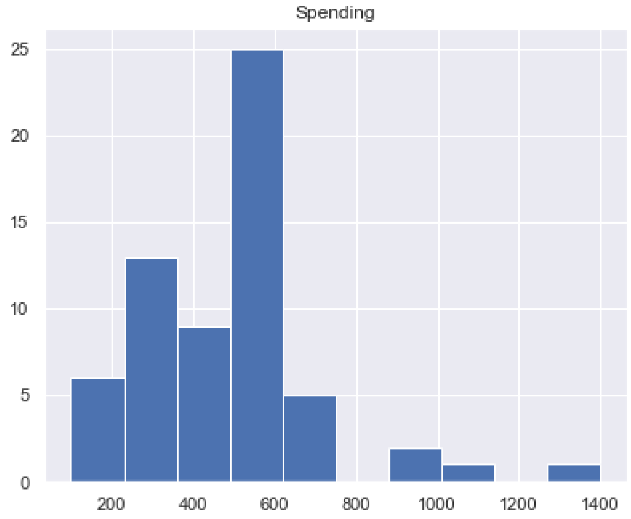
The conditional probability that a randomly selected male earns 50 or more is 48.28

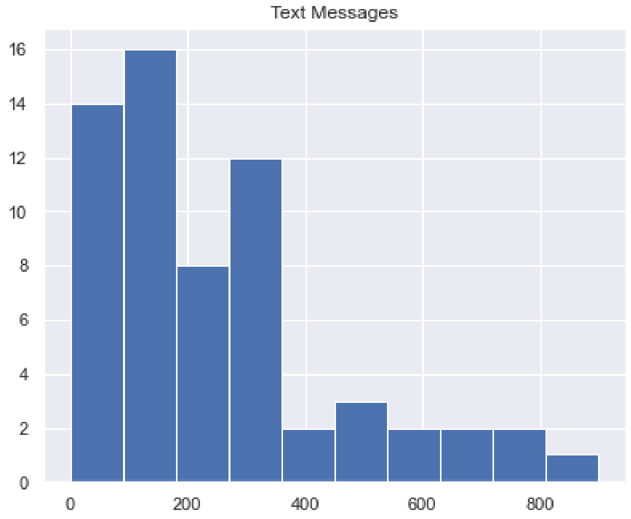
The conditional probability that a randomly selected female earns 50 or more is 54.55

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.









GPA and Salary seem follow normal distributions

Spending and Text-messages seem not to follow normal distribution. They seem right skew

Problem 3 -

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

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

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

1. Define null and alternative hypotheses

   H0 = mean moisture content is not equal to 0.35 pound per 100 square feet

   H1 = mean moisture content is less than 0.35 pound per 100 square feet

2. Decide the significance level

   Here we select α = 0.05 and the population standard deviation is not known

3. Identify the test statistic

   We have two samples and we do not know the population standard deviation.

   Sample sizes for both samples are not same. n1=36 n1=31

   We use two sample t-test.

4. Calculate the p - value and test statistic

   tstat 0.845

   p-value for one-tail: 0.2025369351827172

5. Decide to reject or accept null hypothesis

   Paired two-sample t-test p-value= 0.2025369351827172

   We do not have enough evidence to reject the null hypothesis in favour of alternative hypothesis

   We need to accept alternate hypothesis "mean moisture content is less than 0.35 pound per 100 square feet"

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

1. Define null and alternative hypotheses

   H0: μA - $\mu B \neq 0$

   HA: μA - μB = 0

2. Decide the significance level

   Here we select α = 0.05 and the population standard deviation is not known

3. Identify the test statistic

   We have two samples and we do not know the population standard deviation.

   Sample sizes for both samples are not same. n1=36 n1=31

   We use two sample t-test.

4. Calculate the p - value and test statistic

   tstat 0.985249977839441

   P Value 0.3284577916404776

5. Decide to reject or accept null hypothesis

   We do not have enough evidence to reject the null hypothesis in favour of alternative hypothesis