Problem 1: Descriptive Statistics

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 (Wholesale Customer.csv) 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).

Description of variables:

Buyer/Spender: Customer who makes/spends a purchase (Continuous data)

Fresh: annual spending on fresh products (Continuous data);

Milk: annual spending on milk products (Continuous data);

Grocery: annual spending on grocery products (Continuous data);

Frozen: annual spending on frozen products (Continuous data)

Detergents_Paper: annual spending on detergents and paper products (Continuous data)

Delicatessen: annual spending on and delicatessen products (Continuous data);

Channel: Hotel/Retail (Nominal data)

Region: Lisbon/Oporto/Other (Nominal data)

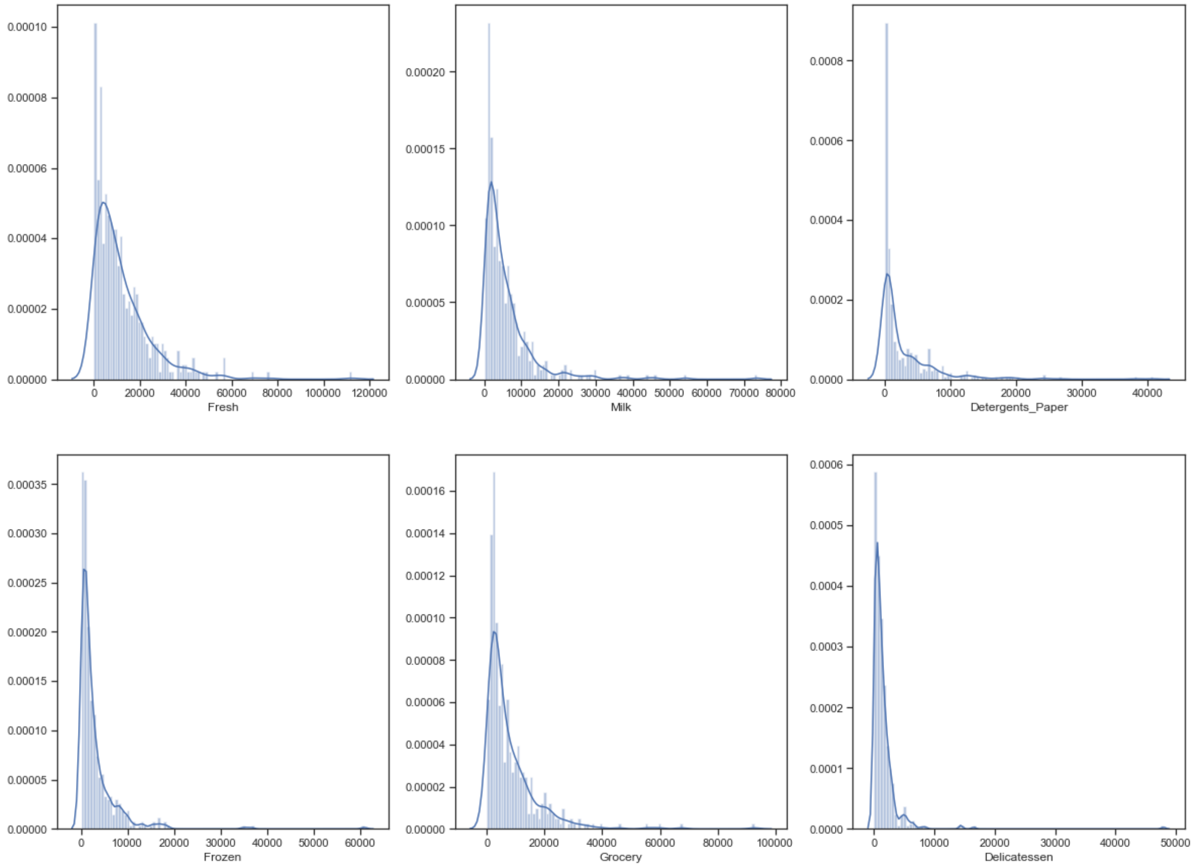
The above dataset gives data on sales of 6 category of products across 3 regions through 2 channel.

Dropping the non product column to see the product description.

Spread of Data

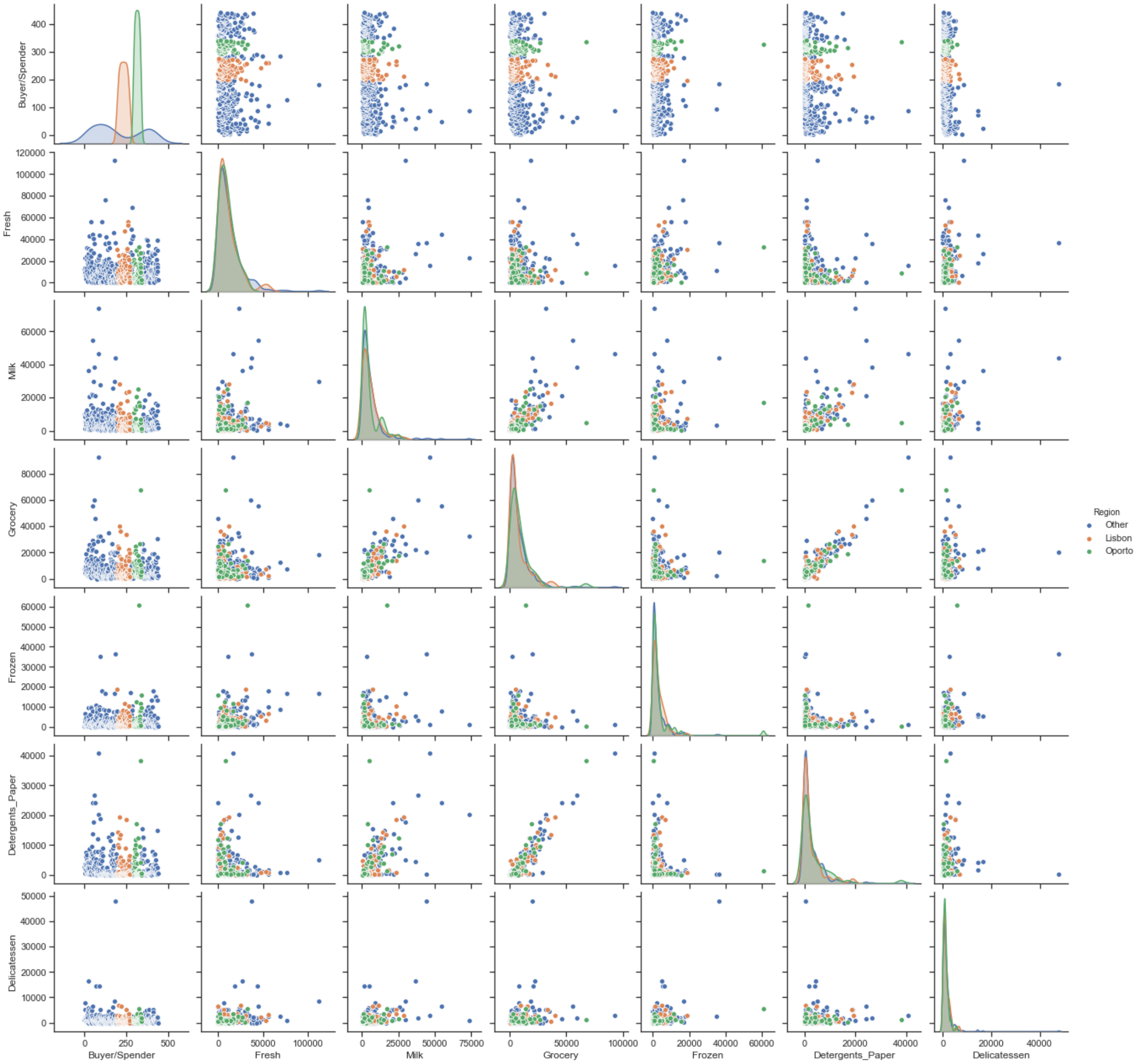
From the below plots,it seems the distribution is right skewed From the jupyter code, Low standard deviation 2820.105937 for item Delicatessen means data are clustered around the mean and high standard deviation 12647.328865 for item Fresh indicates data are more spread out

```
In [16]: All the variables are not normally distributed
```



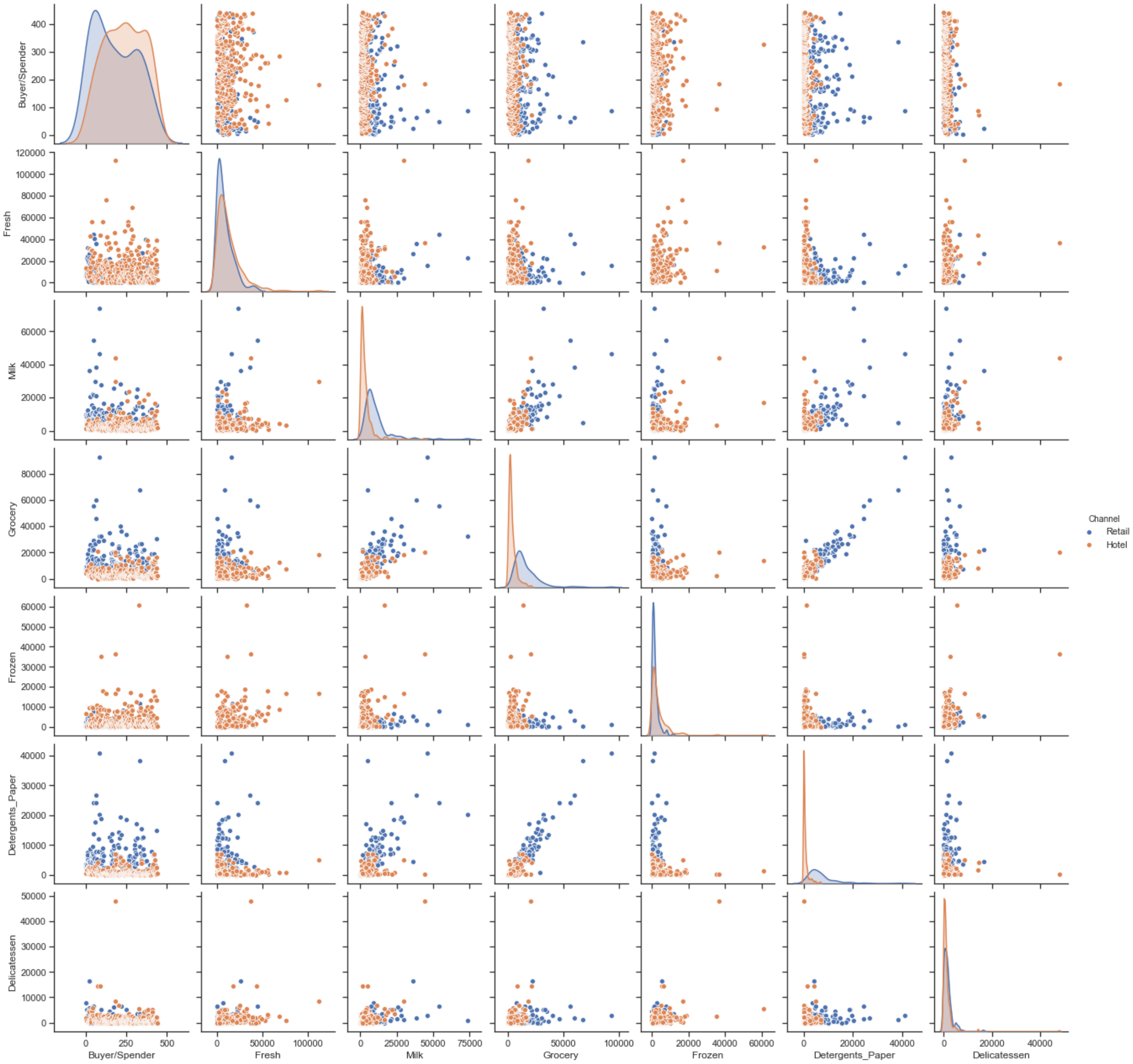
In [99]:

```
Observation :-

Below plot showing the relationship between the two continuous variables with
hue as Region wise
```



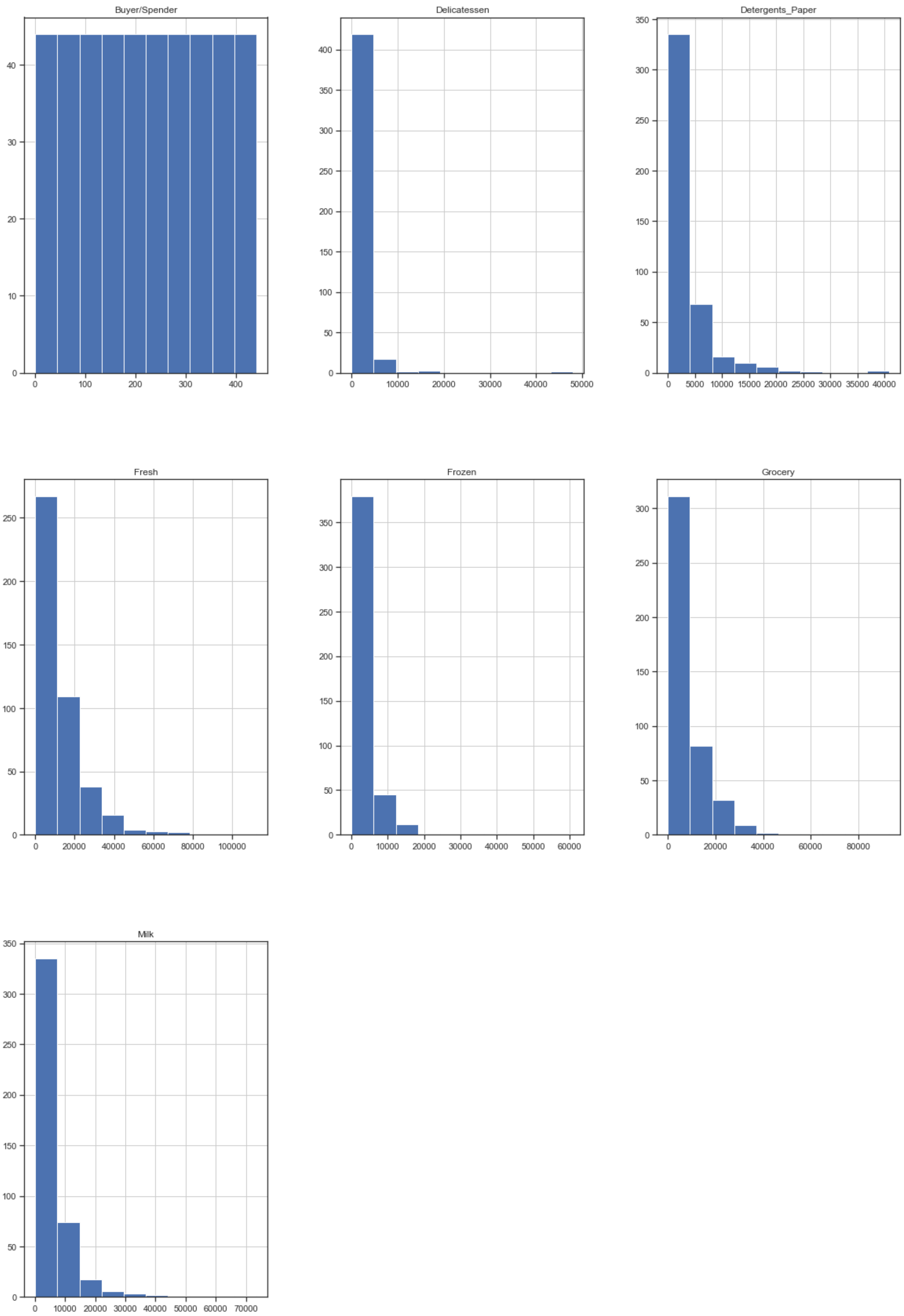
In [100]:

```
Observation :-

Below plot showing the relationship between the two continuous variables with
hue as Channel wise
```



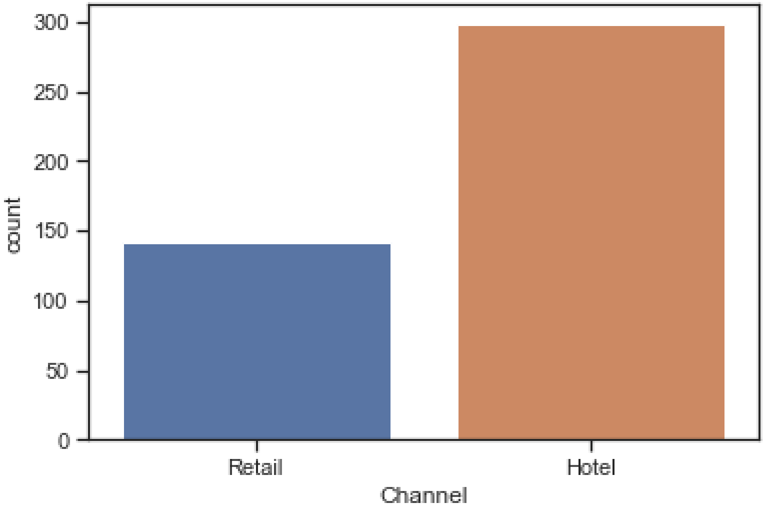
In [17]:

```
Below showing that the graph is right skewed
All variables are not normally distributed
```

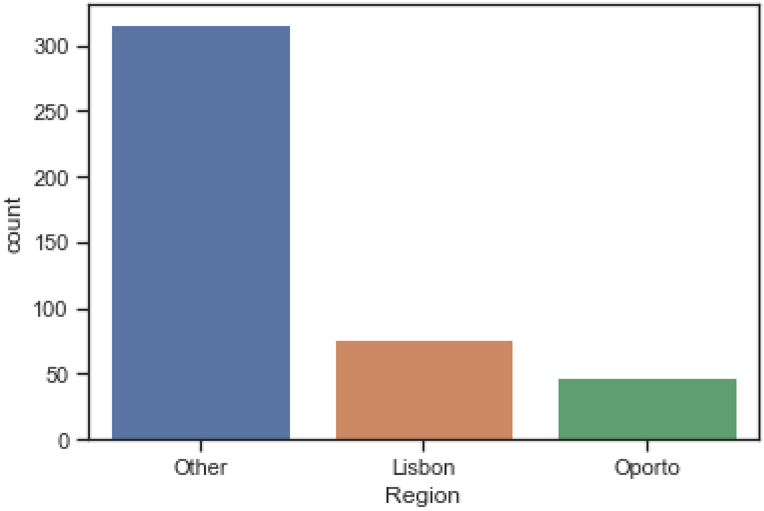


Use methods of descriptive statistics to summarize data. Which Region and which Channel seems to spend more? Which Region and which Channel seems to spend less?

In [44]:
```
We see that the most used channel is Hotel for selling the items
```

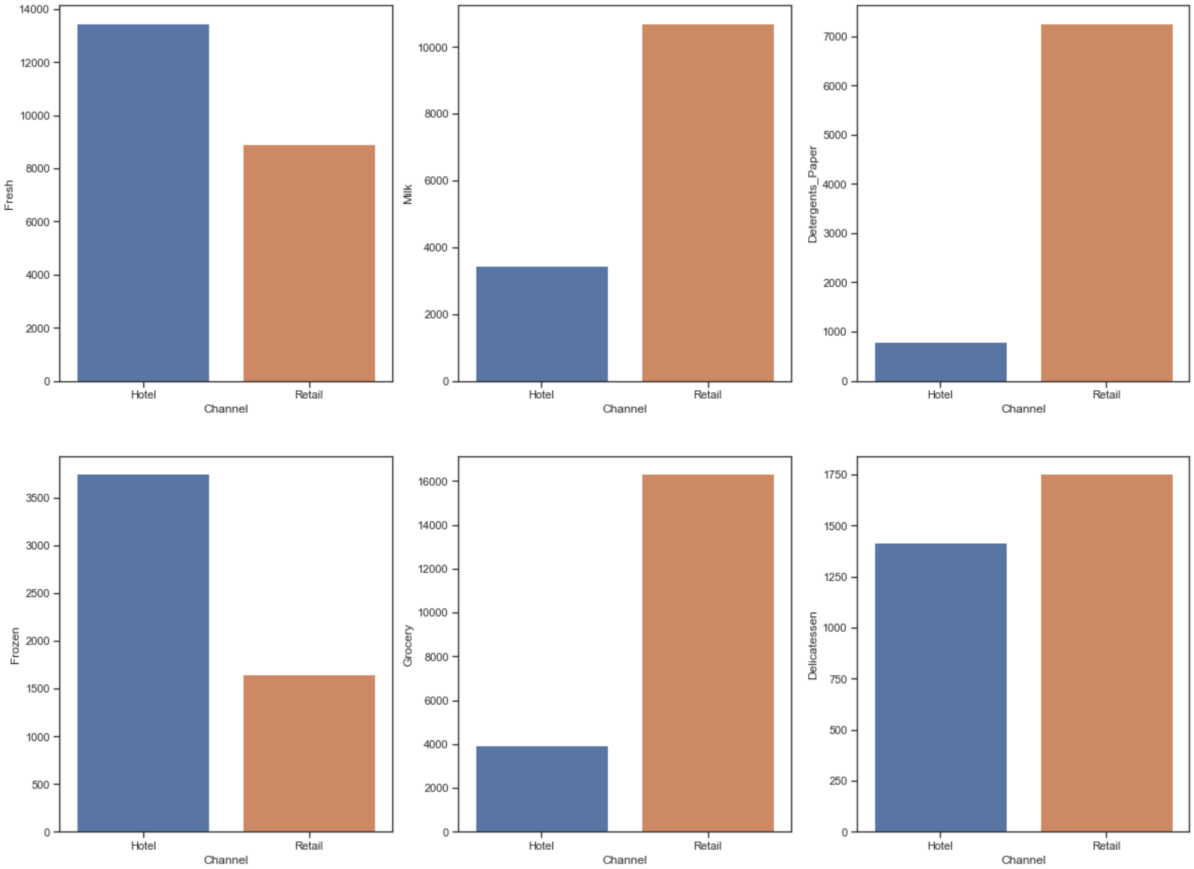


In [46]:
```
We see that the most used Region is Other for selling the items
```

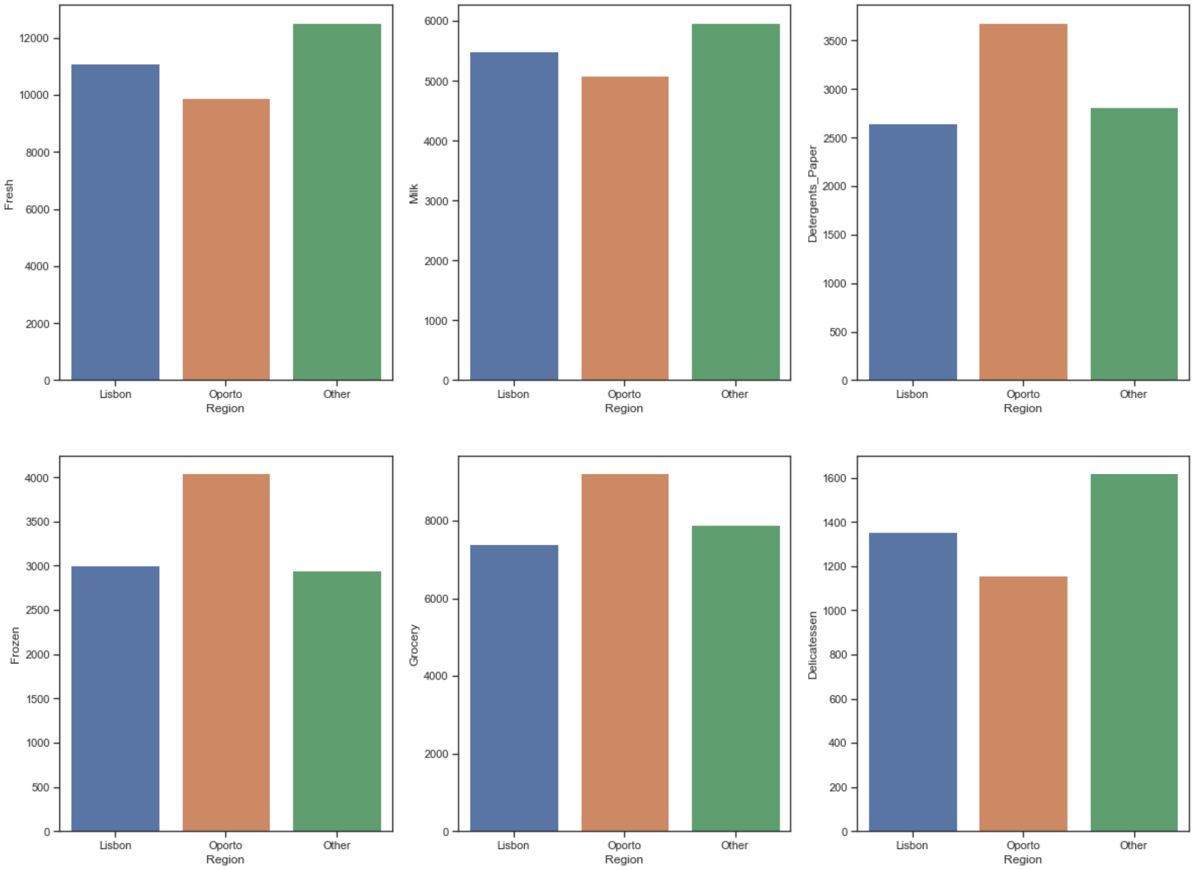


In [72]:

```
Observation :-

Dropping the Region data from the dataset and plotting the channel vs Items graph

We can see that in channel Hotel average highest spend is more in Fresh and average lowest spend is in Detergents_Paper

We can see that in channel Retail average highest spend is more in Grocery and average lowest spend is in Frozen items.
```



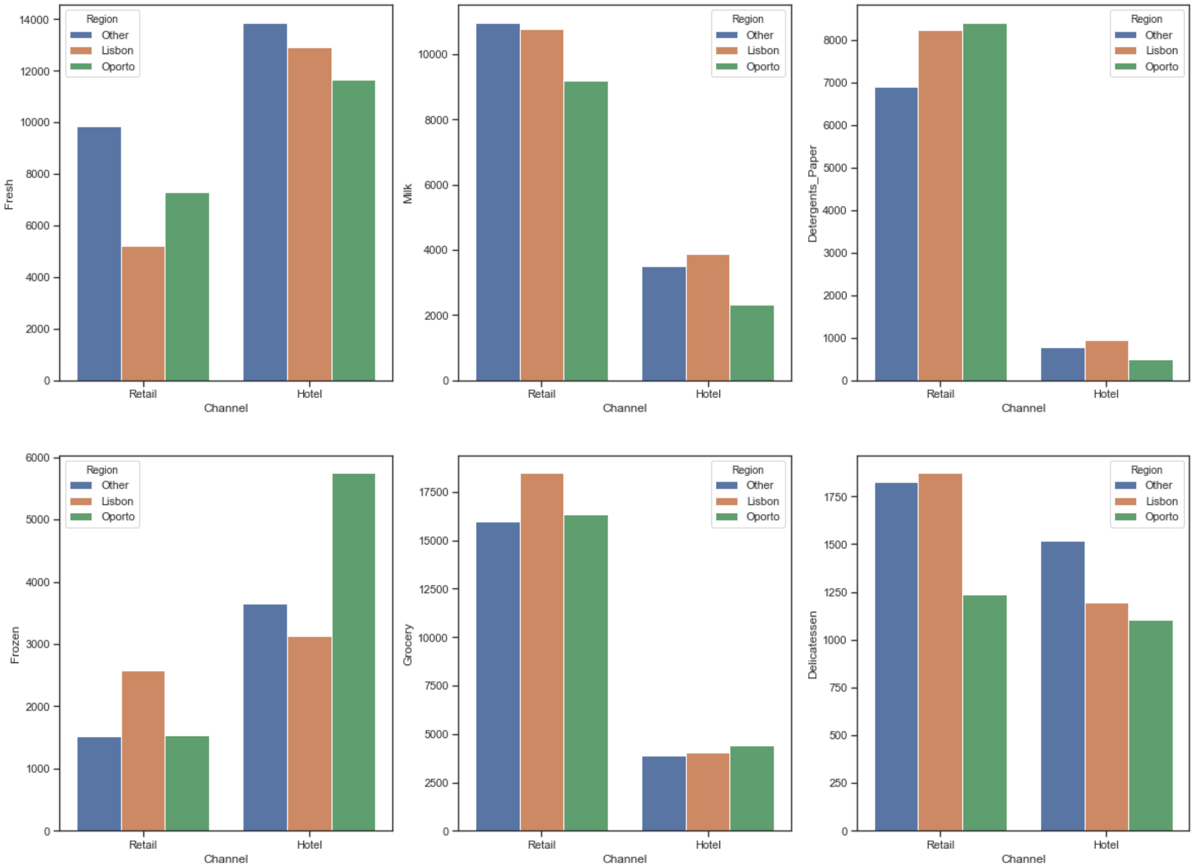
```
In [75]: Observation :-

Dropping the Channel data from the dataset and plotting the Region vs Items graph

In Region Lisbon Average Highest Spending is in Fresh and Lowest in Delicassen items.
In Region Oporto Average Highest Spending is in Fresh and Lowest in Delicassen items.
In Region Other Average Highest Spending in Fresh and Lowest in Delicassen items.
```



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

In [79]:

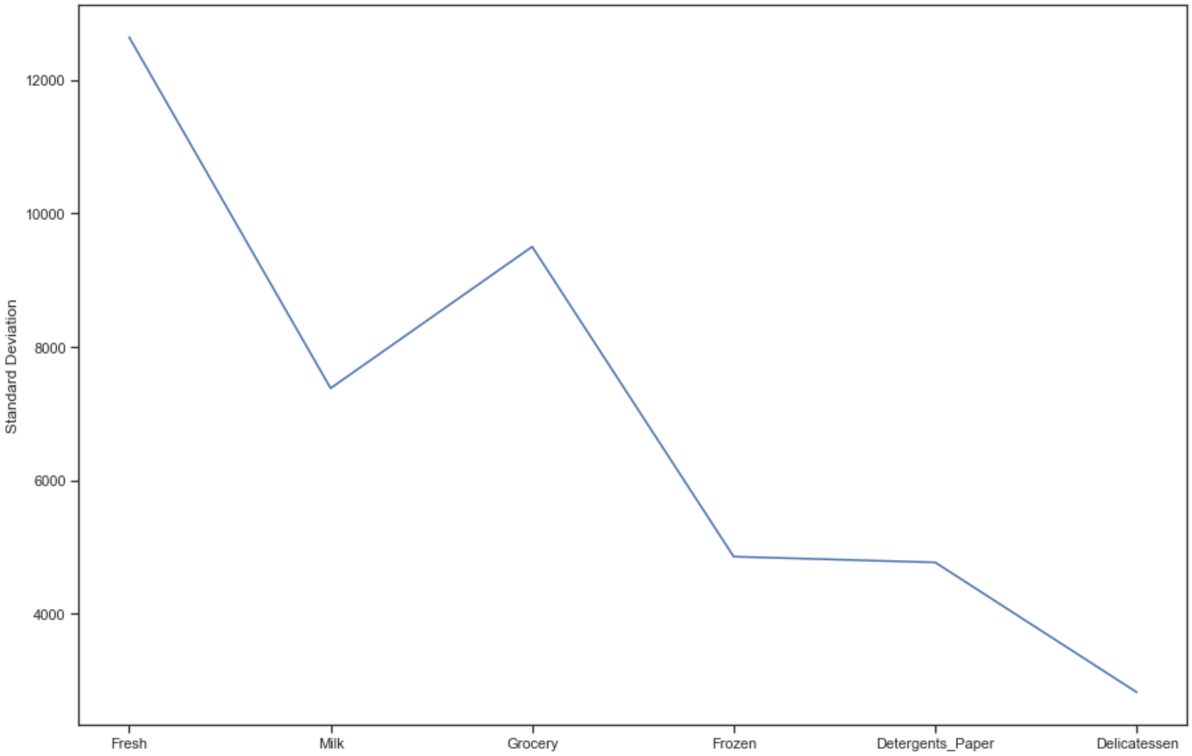
```
Observations :-

From the below graph we can infer that all varieties show different behaviour
across Region and Channel
```



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

In [92]:
```
Observations :-

From the below plot we can infer that the Fresh has the highest standard devia
tion it means it is the most inconsistent item
and Delicatessen has the lowest standard deviation it means it is the least in
consistent item
```

Out[92]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1f3a4f26508>
```



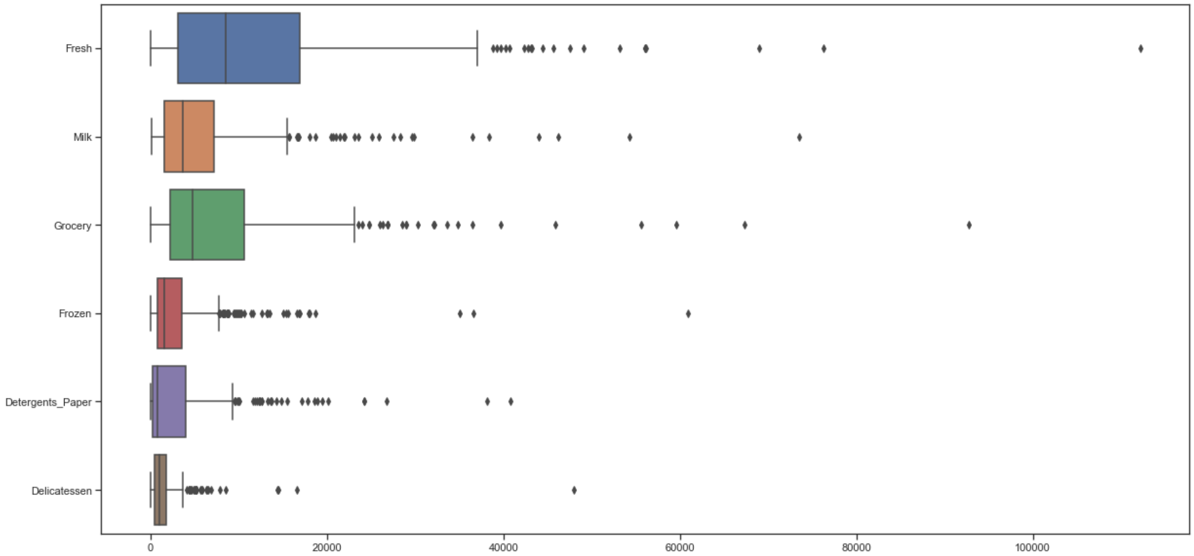
Are there any outliers in the data?

In [101]:

```
Observations :-

We can see from the below box plot that the distribution is right skewed i.e.
the mean is higher than the median.
Also notice that the tail of the distribution on the right hand (positive skew
ed)
```

Out[101]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1f3ab59d848>
```



On the basis of this report, what are the recommendations?

In [110]:

```
Lets choose the sample data from the wholesale data and see the recommendation
s:-
```

Out[110]:

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
1	7057	9810	9568	1762	3293	1776
2	6353	8808	7684	2405	3516	7844

In [111]:

```
Calculating the mean offset by subtracting the original data from the sample d
ata
```

	Buyer/Spender	Delicatessen	Detergents_Paper	Fresh	Frozen	Grocery	Milk
1	NaN	251.0	412.0	-4943.0	-1310.0	1617.0	4014.0
2	NaN	6319.0	635.0	-5647.0	-667.0	-267.0	3012.0

In [112]:
```
Calculating the median offset by subtracting the original data from the sample
data
```

	Buyer/Spender	Delicatessen	Detergents_Paper	Fresh	Frozen	Grocery	Milk
1	NaN	810.0	2477.0	-1447.0	236.0	4812.0	6183.0
2	NaN	6878.0	2700.0	-2151.0	879.0	2928.0	5181.0

In [ ]:
```
Observation/Recommendations :-

Customer a: It has high spending in Milk and Grocery, low spending in Fresh, F
rozen.
Medium spending in Detergents_Paper and Delicatessen . Hence it would be a Gro
cery shop.
They can think of investing more on the Fresh and Frozen Items

Customer b: It has high spending Delicatessen and Milk, low spending in Fresh,
Frozen and Grocery
and medium spending in Detergents_Paper
Its a ready to eat shop. They can expand and think investing more in low spend
ing items.
```

Problem 2 : Probability

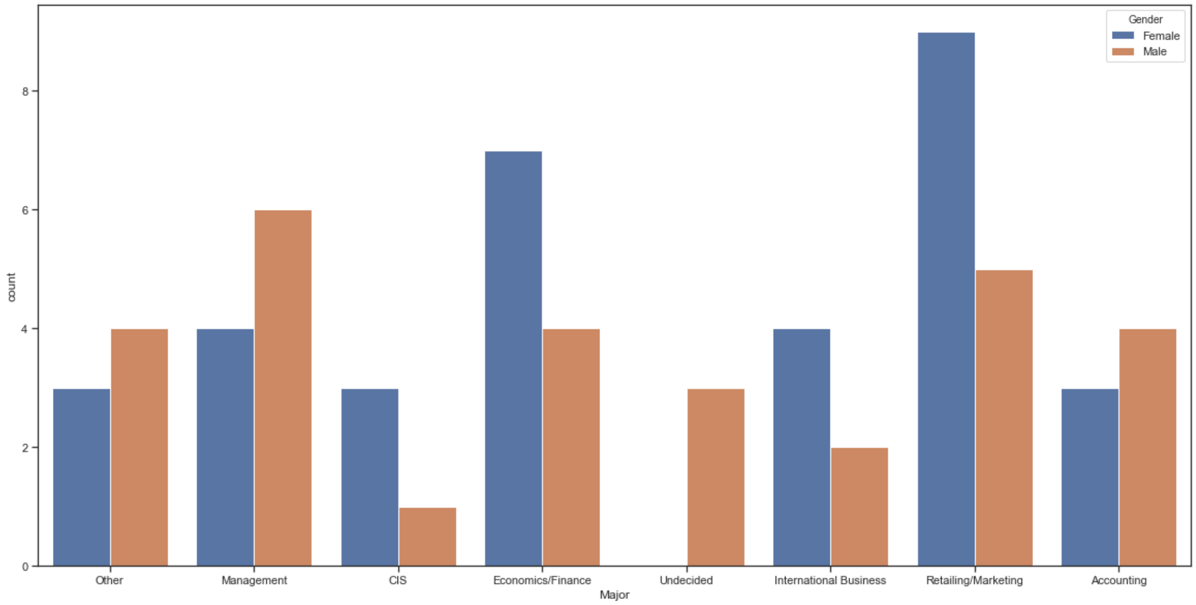
The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. CMSU creates and distributes a survey of 14 questions and receives responses from 62 undergraduates (stored in the Survey.csv file).

In [116]:

```
# 2.1.1. Gender and Major
From the below plot we observe that most of the Female opted for Retailing/Mar
keting specialization and
we see that Female are good in decision making as the number of Male we can se
e is not sure which major they should opt for.

We also see that most of the Male opted for Management specialization.

Only 1 male opted for CIS course
```

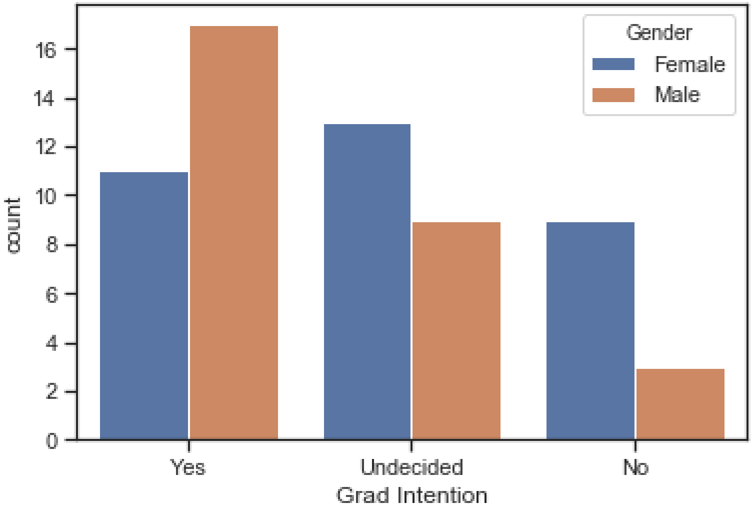
Out[116]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1f3abaaea88>
```



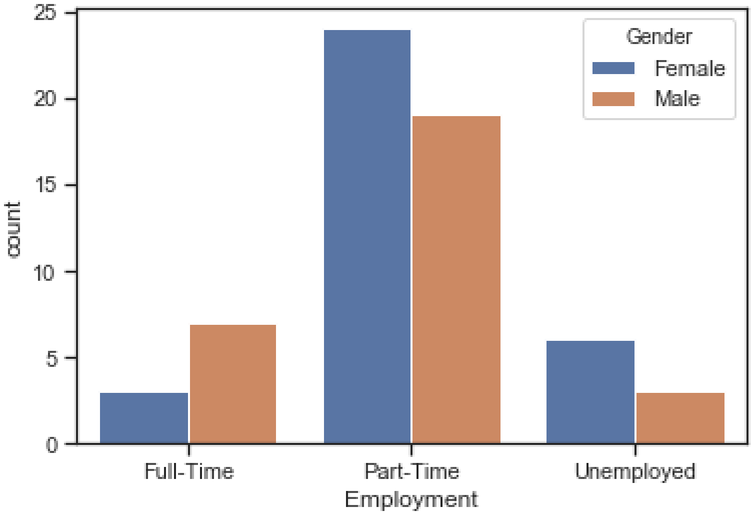
In [122]:

```
# 2.1.2. Gender and Grad Intention
From the below observation, we can say that number of males is more confident
in opting for the course
Number of more Females is not sure wether to go for the specific course or not
.
```



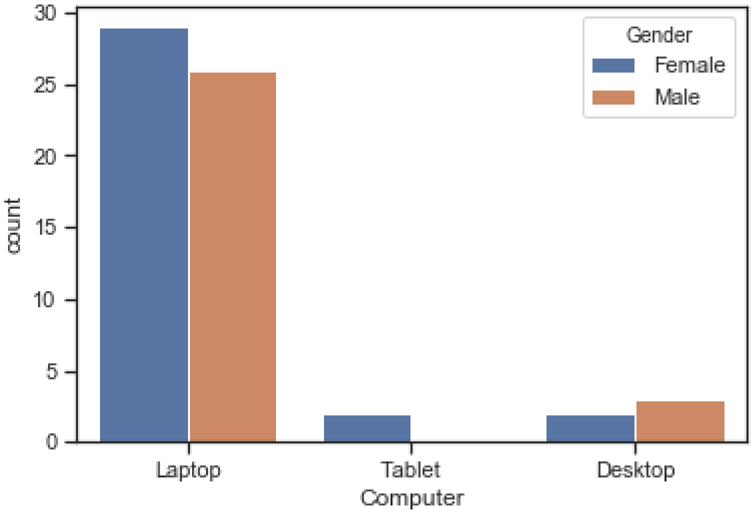
In [123]:

```
# 2.1.3. Gender and Employment
From the below graph we can say that both male and female prefer doing part-ti
me employment and more number of males going
for full time and more female numbers are unemployed
```



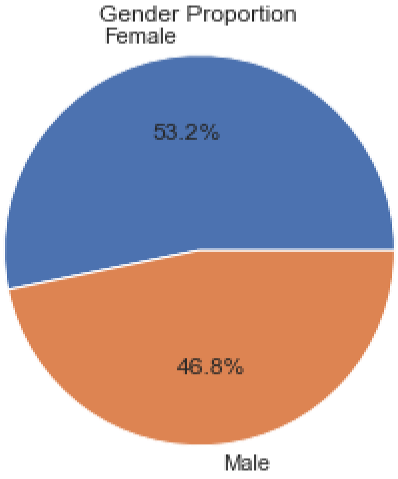
In [126]:

```
# 2.1.4. Gender and Computer

Males dont prefer tablet for their personal/business purpose and see that both
the gender prefers laptop as for their
personal/business work and a few male and female will buy desktop.
```



In [145]:

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

From the below observation we can say that the probability of selecting a rand
om male will be the number of males divided by
the total number of students which is 46.8%

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

From the below observation we can say that the probability of selecting a rand
om Female will be the number of Females divided by
the total number of students which is 53.2%
```

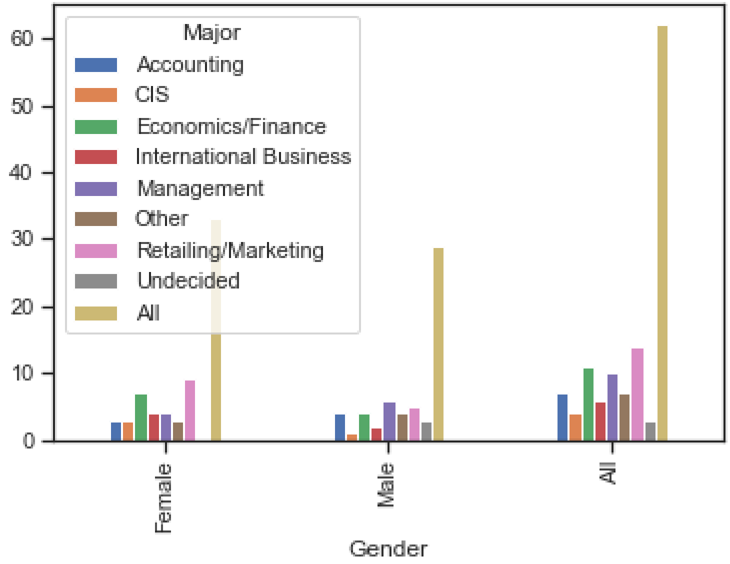


In [151]:

```
Find the conditional probability of different majors among the male students i
n CMSU.
Find the conditional probability of different majors among the female students
of CMSU.

From the below we observe that more Number of Males opt for Management course
From the below we observe that more Number of Female opts for Retailing/Market
ing
```

Out[151]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1f3b3ce4a48>
```

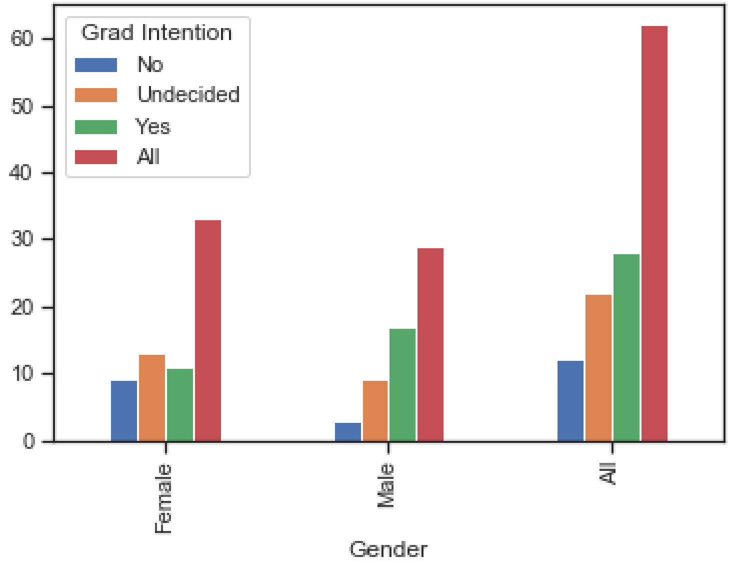


In [152]:

```
We can see that more number of males have decided yes to do the graduation cou
rse as compare to number of females
Similarly, less number of males have no intention of doing graduation as compa
re to Females.
```

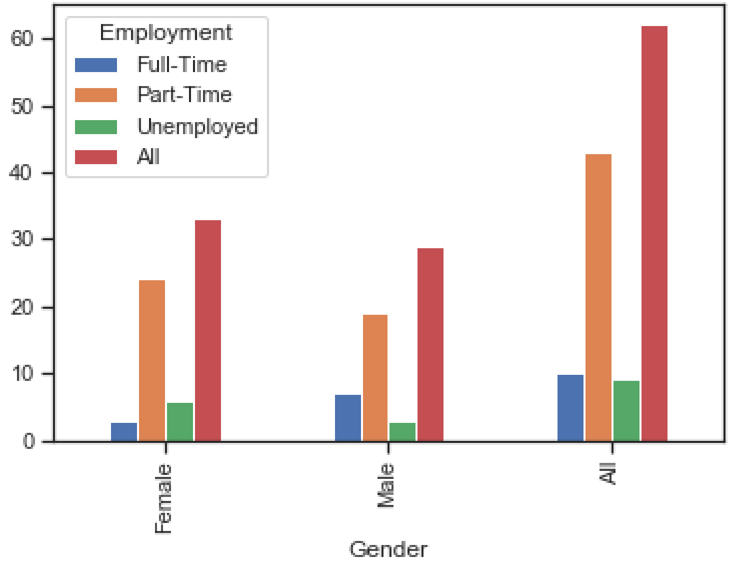
Out[152]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1f3b3d8f688>
```



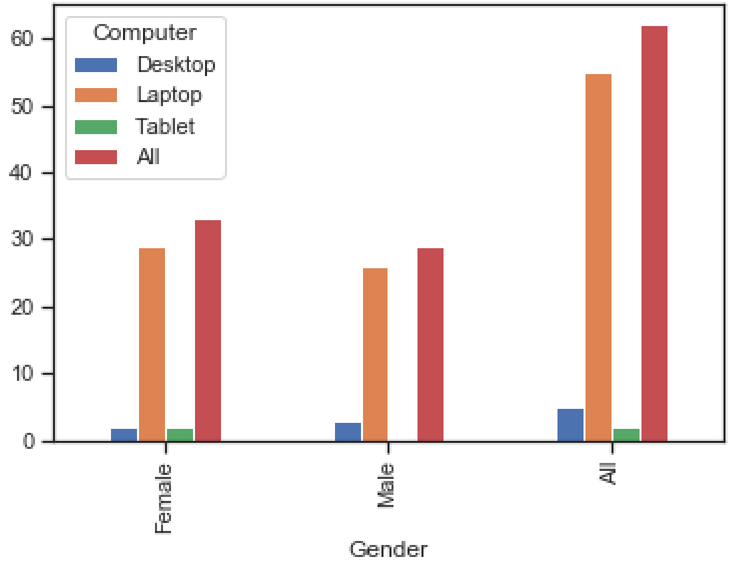
In [153]:
```
We observe that both females and males prefer doing part-time employment
Males prefer more full-time employment than the females
```

Out[153]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1f3b35e4cc8>
```



In [154]:
```
We observe that males dont prefer tablet at all and both gender prefers laptop
instead of desktop
(which we can see a few students took)
Very less Female prefers tablet
```

Out[154]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1f3b347f608>
```

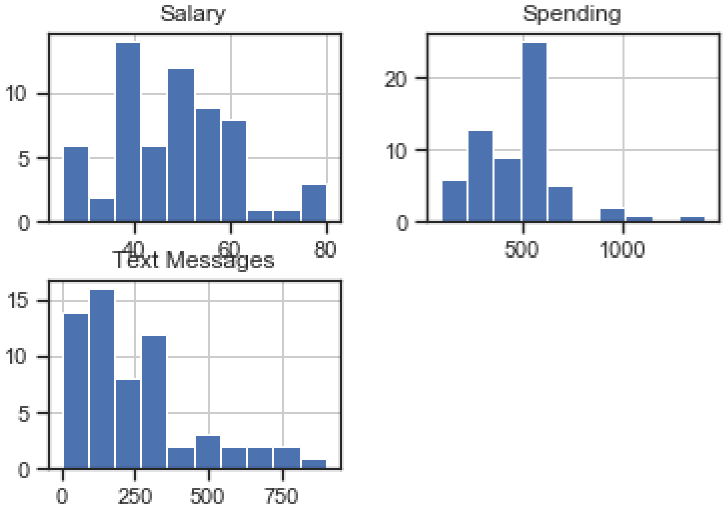


In [155]:

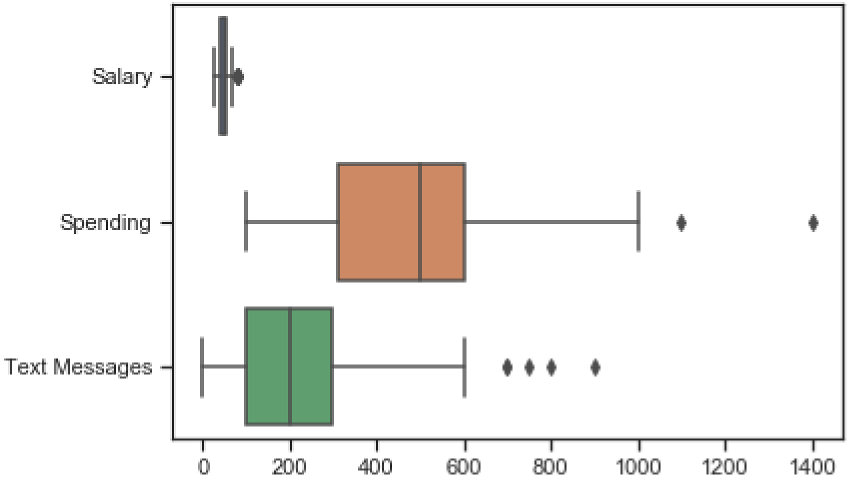
```
2.4. Note that there are three numerical (continuous) variables in the data set, Salary, Spending and Text Messages. For each of them comment whether they follow a normal distribution.
Write a note summarizing your conclusions.
[Recall that symmetric histogram does not necessarily mean that the underlying distribution is symmetric]

From below histogram, we see that it is not symmetric.
```

Out[155]:

```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x000001F3B338D308>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001F3B3328688>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x000001F3B3294C08>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001F3A6A3C248>]],
      dtype=object)
```



In [156]:



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 claims that the mean moisture content cannot be greater 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.

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

LaTeX: H_0<=0.35

LaTeX: H_A>0.35

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

LaTeX: H_0<=0.35

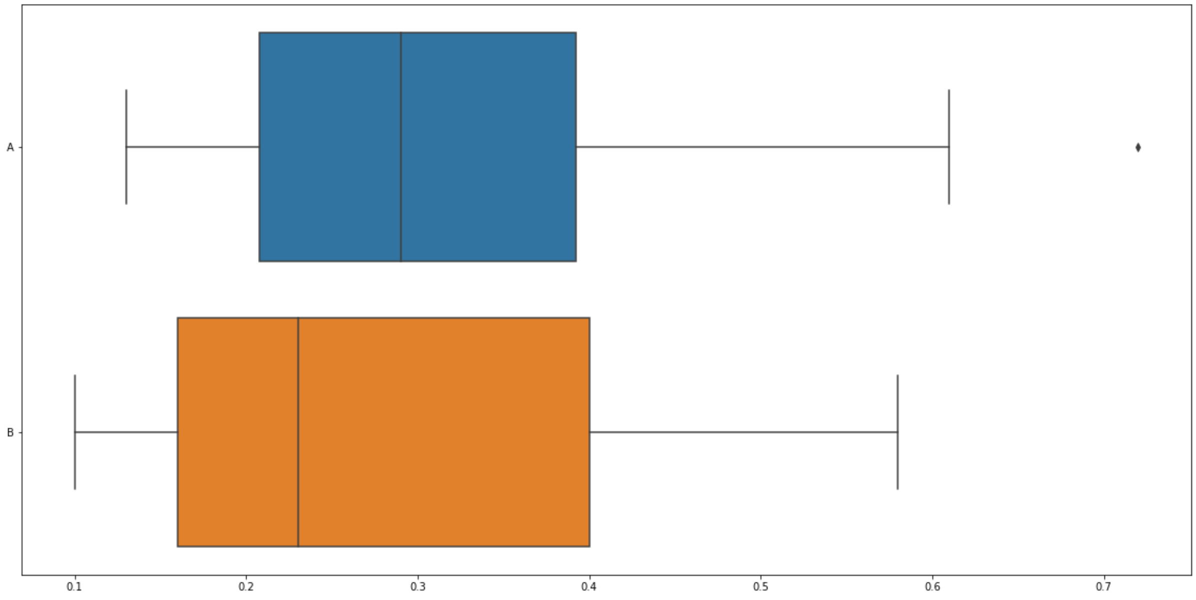
LaTeX: H_A>0.35

In [11]:

```
We can see the distribution of A shingles is normally distributed with approx
symmetry
We can see the distribution of B shingles is Right skewed
```

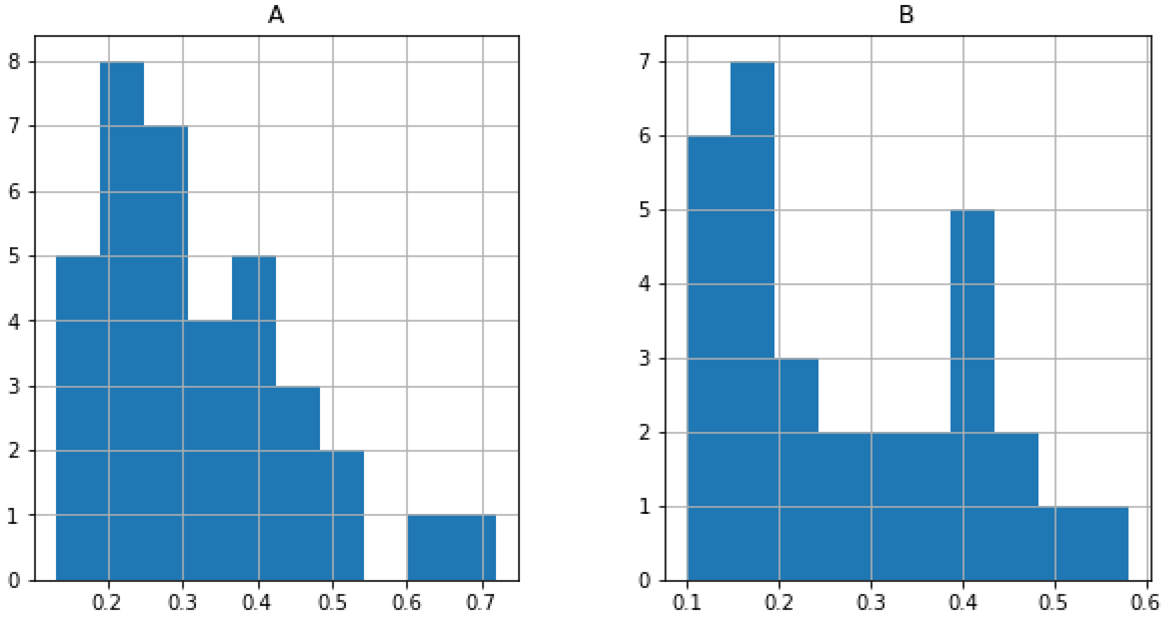
Out[11]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x27524cc6e48>
```



In [12]:

```
Showing histogram of both the column variables
```


A
8
7
6
5
4
3
2
1
0
0.2
0.3
0.4
0.5
0.6
0.7
B
7
6
5
4
3
2
1
0
0.1
0.2
0.3
0.4
0.5
0.6