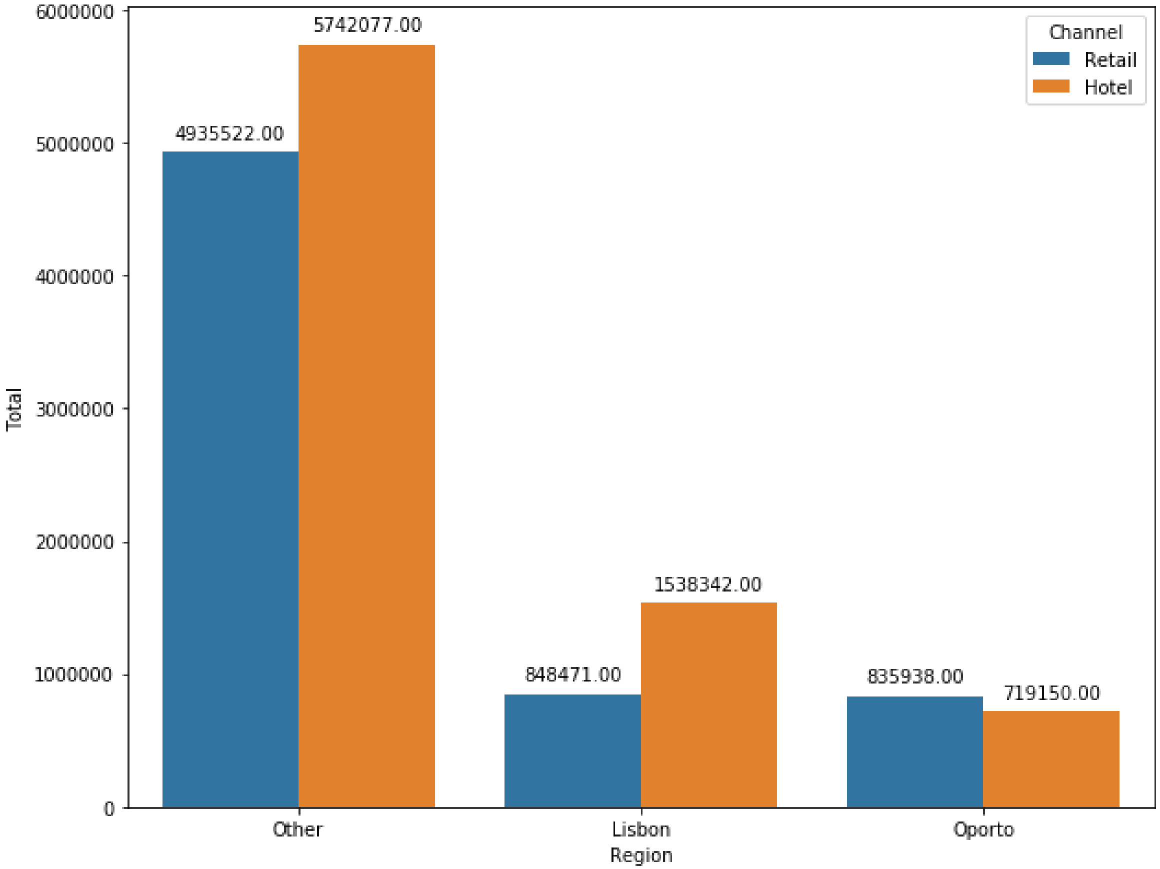
SMDM PROJECT REPORT

Monika Nanda

Problem 1

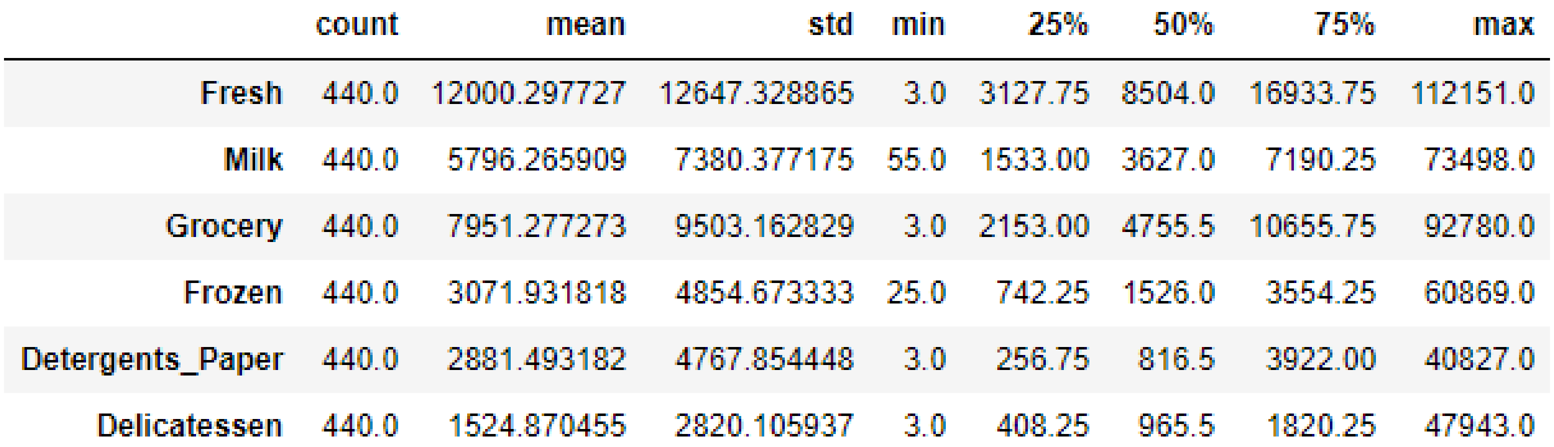
Wholesale Customer Data

- 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?



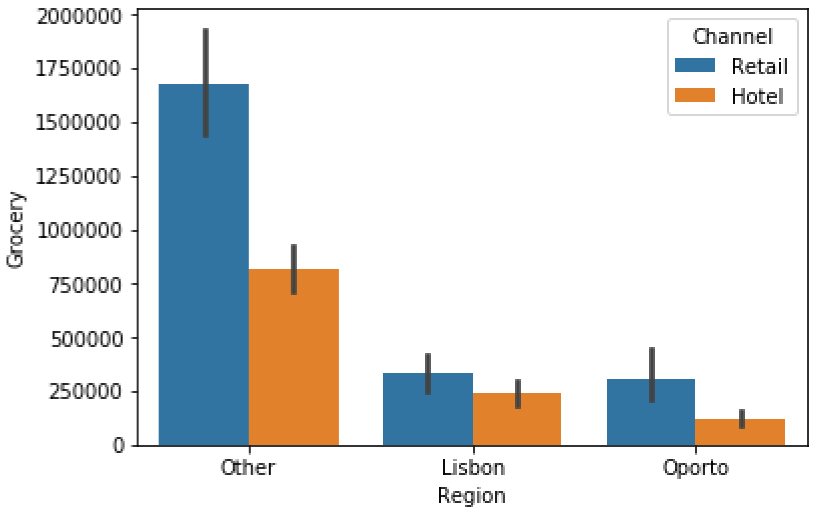
- From above matrix and graph we can conclude that Hotel of "Other" region spend the most and Hotel of "Oporto" region spend the least.

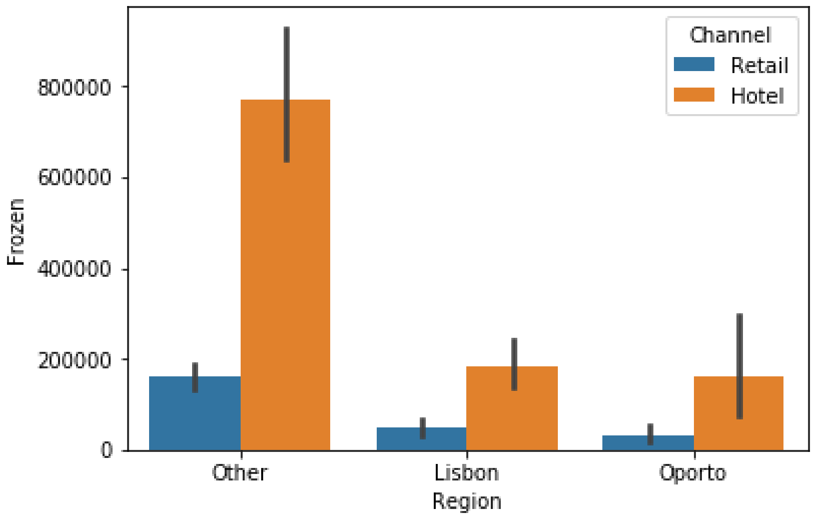
- 1.2. There are 6 different varieties of items.
  Do all varieties show similar behaviour across Region and Channel?

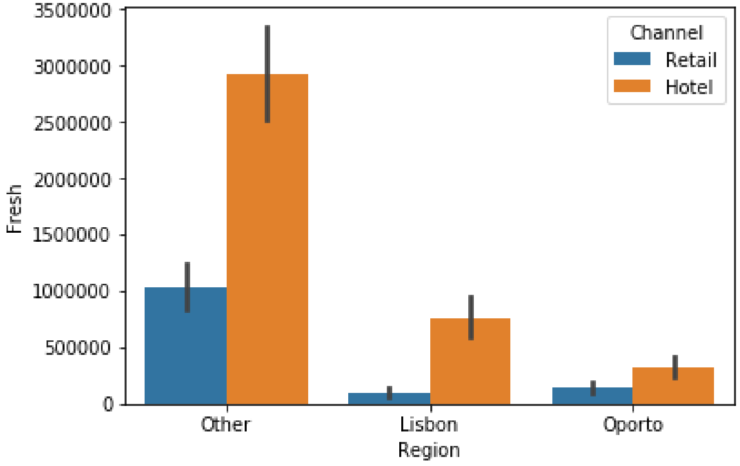


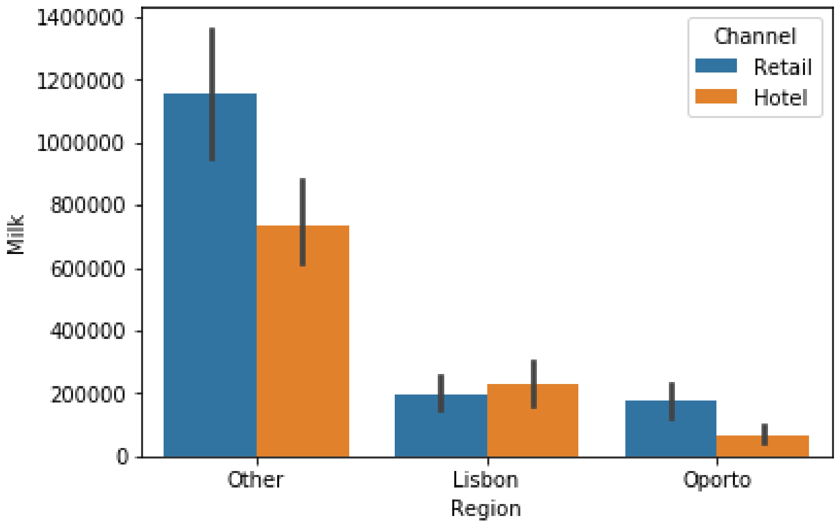
	count	mean	std	min	25%	50%	75%	max
Fresh	440.0	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0
Milk	440.0	5796.265909	7380.377175	55.0	1533.00	3627.0	7190.25	73498.0
Grocery	440.0	7951.277273	9503.162829	3.0	2153.00	4755.5	10655.75	92780.0
Frozen	440.0	3071.931818	4854.673333	25.0	742.25	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

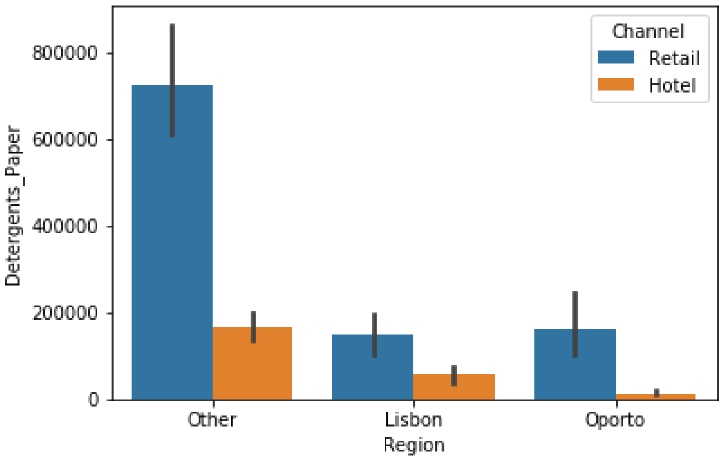
- Looking at the above summary statistics we can confirm that the six varieties do not show similar behaviour as the mean and standard deviation values are very different from each other.

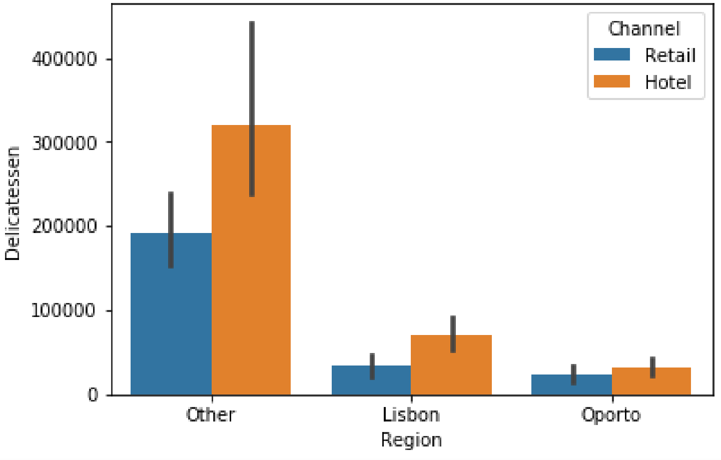

Channel
Retail
Hotel
Grocery
2000000
1750000
1500000
1250000
1000000
750000
500000
250000
0
Other
Lisbon
Oporto
Region


Channel
Retail
Hotel
Frozen
800000
600000
400000
200000
0
Other
Lisbon
Oporto
Region


Channel
Retail
Hotel
Fresh
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
Other
Lisbon
Oporto
Region


Channel
Retail
Hotel
Milk
1400000
1200000
1000000
800000
600000
400000
200000
0
Other
Lisbon
Oporto
Region

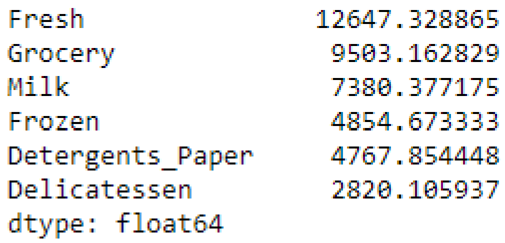




- Also, most varieties show similar spending behaviour across Region and Channel except “**Milk**” as the spend on Milk is more for “**Retail**” in **Other** and **Oporto** Region but it is more for “**Hotel**” for **Lisbon** Region.

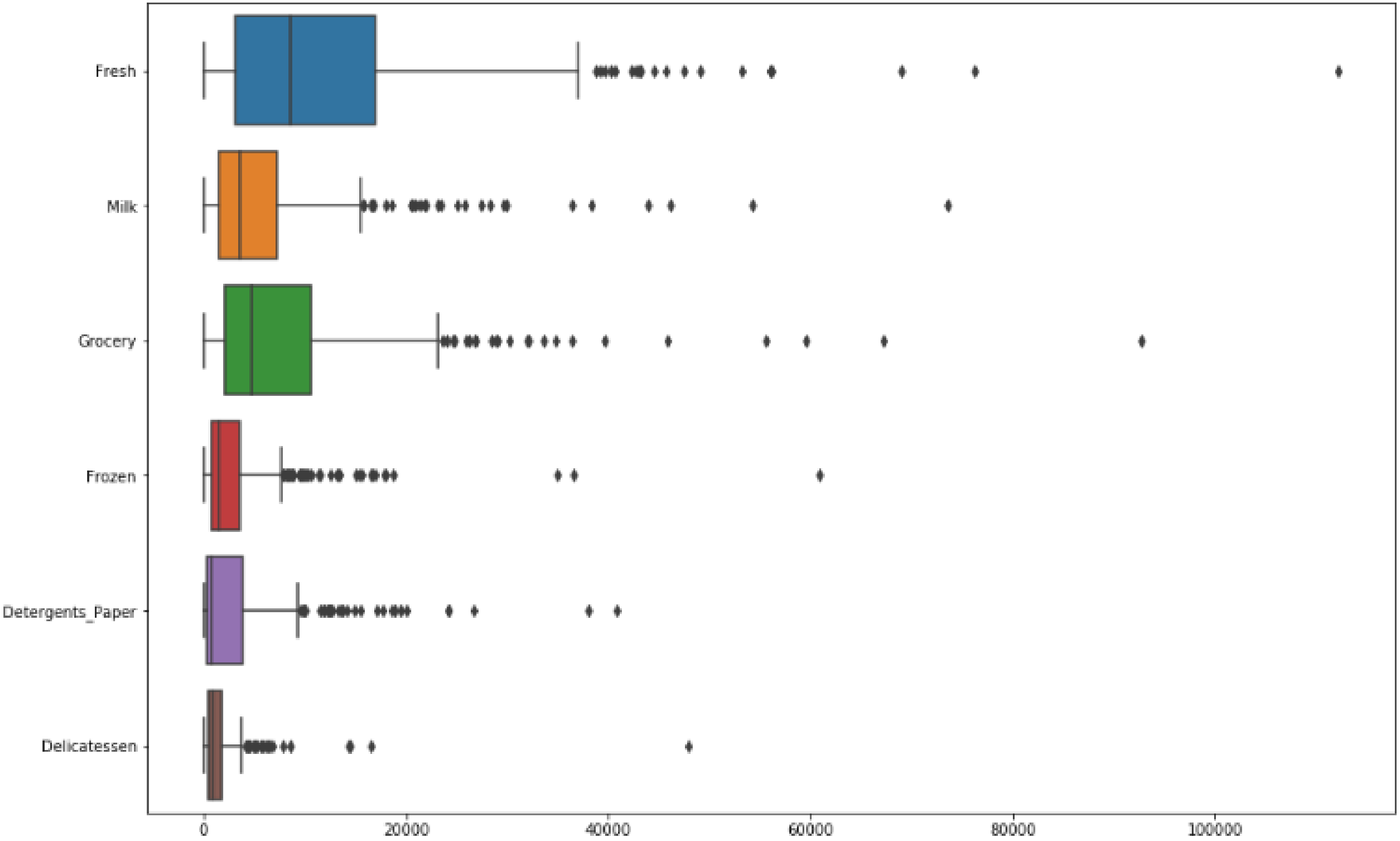
1.3. On the basis of the descriptive measure of variability, which item shows the most inconsistent behaviour? Which items shows the least inconsistent behaviour?

- On the basis of Standard Deviation, “Fresh” shows the most and “Delicatessen” shows the least inconsistent behaviour



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

1.4. Are there any outliers in the data?



- Yes, the above representation confirms that all items have outliers present.

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

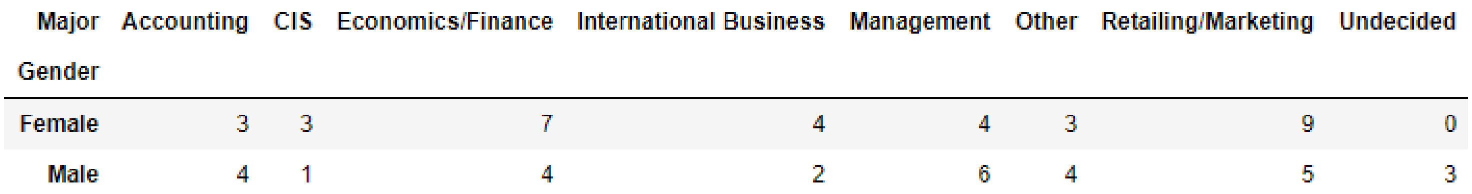
- As per the report, we can conclude that the data given is not consistent and no inference should be made for Business solutions looking at this data, as this report is just stating the spending region and channel wise.
- To conclude on a business solution, we would be needing more details about the problem statement for which the data as gathered
- This data set has outliers present which should be removed initially and then the standard deviation should be checked, if still inconsistency is present we should collect more data with other parameters like month to make better deductions of spending pattern.

Problem 2

Clear Mountain State University (CMSU)

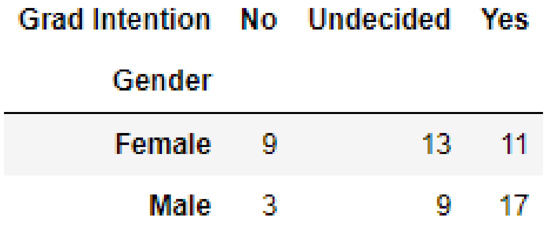
2.1. For this data, construct the following contingency tables (Keep Gender as row variable)

2.1.1. Gender and Major



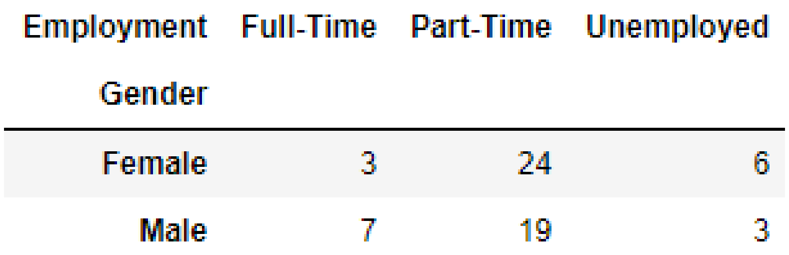
Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Gender								
Female	3	3	7	4	4	3	9	0
Male	4	1	4	2	6	4	5	3

2.1.2. Gender and Grad Intention



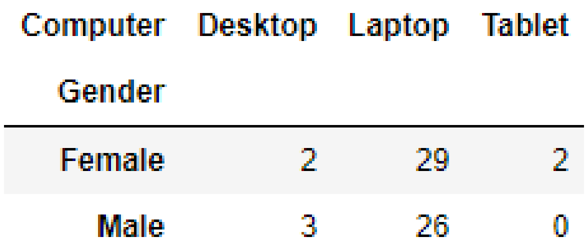
Grad Intention	No	Undecided	Yes
Gender			
Female	9	13	11
Male	3	9	17

2.1.3. Gender and Employment



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

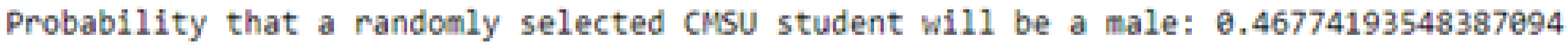
2.1.4. Gender and Computer



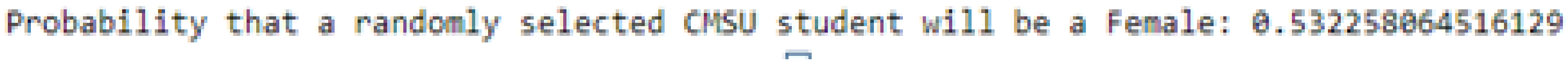
Computer	Desktop	Laptop	Tablet
Gender			
Female	2	29	2
Male	3	26	0

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

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



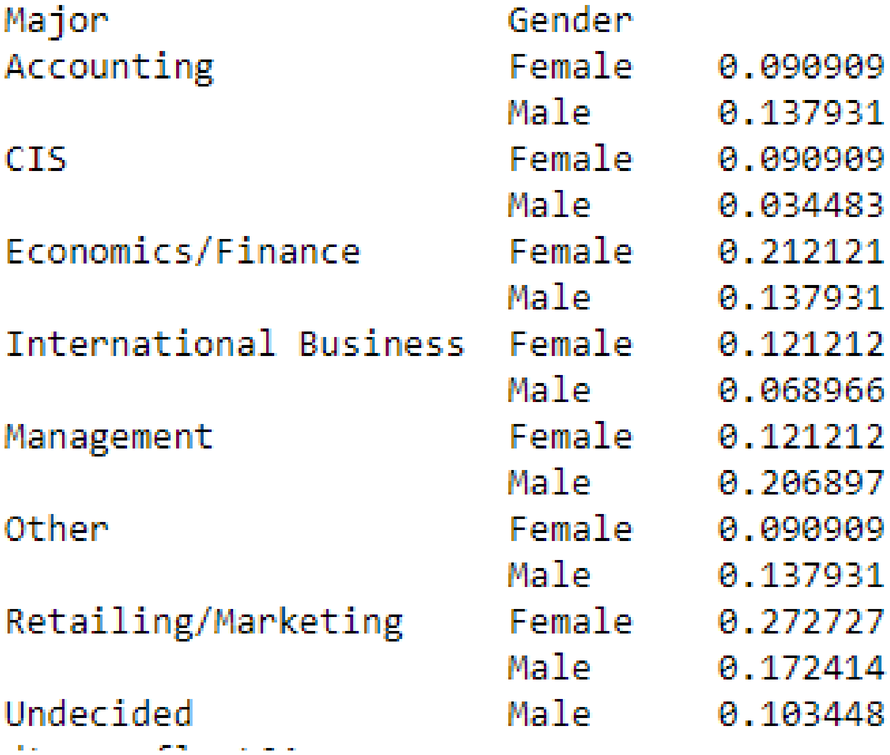
```
Probability that a randomly selected CMSU student will be a male: 0.46774193548387094
```



```
Probability that a randomly selected CMSU student will be a Female: 0.532258064516129
```

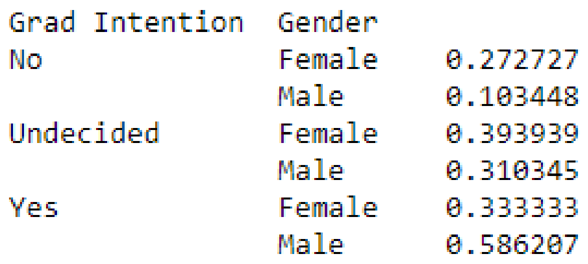
2.2.2. Find the conditional probability of different majors among the male students in CMSU.

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

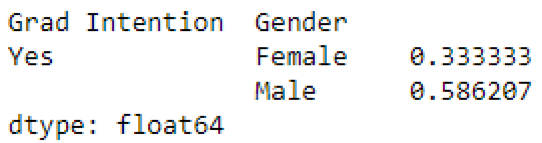


```
Major                   Gender
Accounting              Female    0.090909
                        Male      0.137931
CIS                     Female    0.090909
                        Male      0.034483
Economics/Finance       Female    0.212121
                        Male      0.137931
International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

2.2.3. Find the conditional probability of intent to graduate, given that the student is a male. Find the conditional probability of intent to graduate, given that the student is a female.

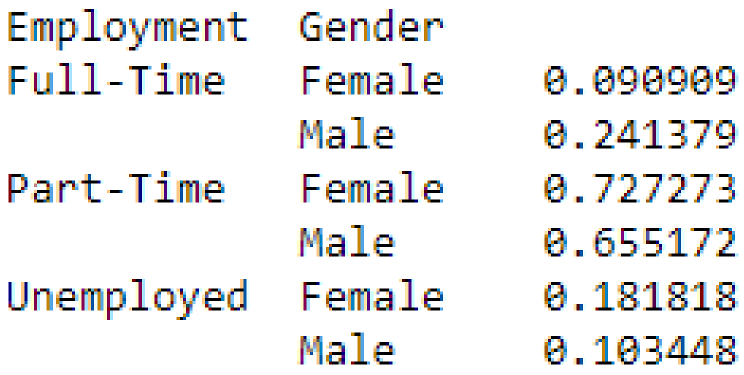


```
Grad Intention  Gender
No              Female    0.272727
                Male      0.103448
Undecided       Female    0.393939
                Male      0.310345
Yes             Female    0.333333
                Male      0.586207
```



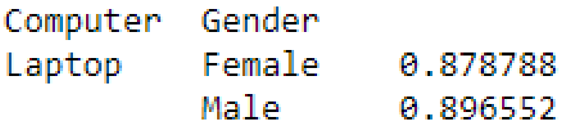
```
Grad Intention  Gender
Yes             Female    0.333333
                Male      0.586207
dtype: float64
```

2.2.4. Find the conditional probability of employment status for the male students as well as for the female students.



```
Employment  Gender
Full-Time   Female    0.090909
            Male      0.241379
Part-Time   Female    0.727273
            Male      0.655172
Unemployed  Female    0.181818
            Male      0.103448
```

2.2.5. Find the conditional probability of laptop preference among the male students as well as among the female students.



```
Computer  Gender
Laptop    Female    0.878788
          Male      0.896552
```

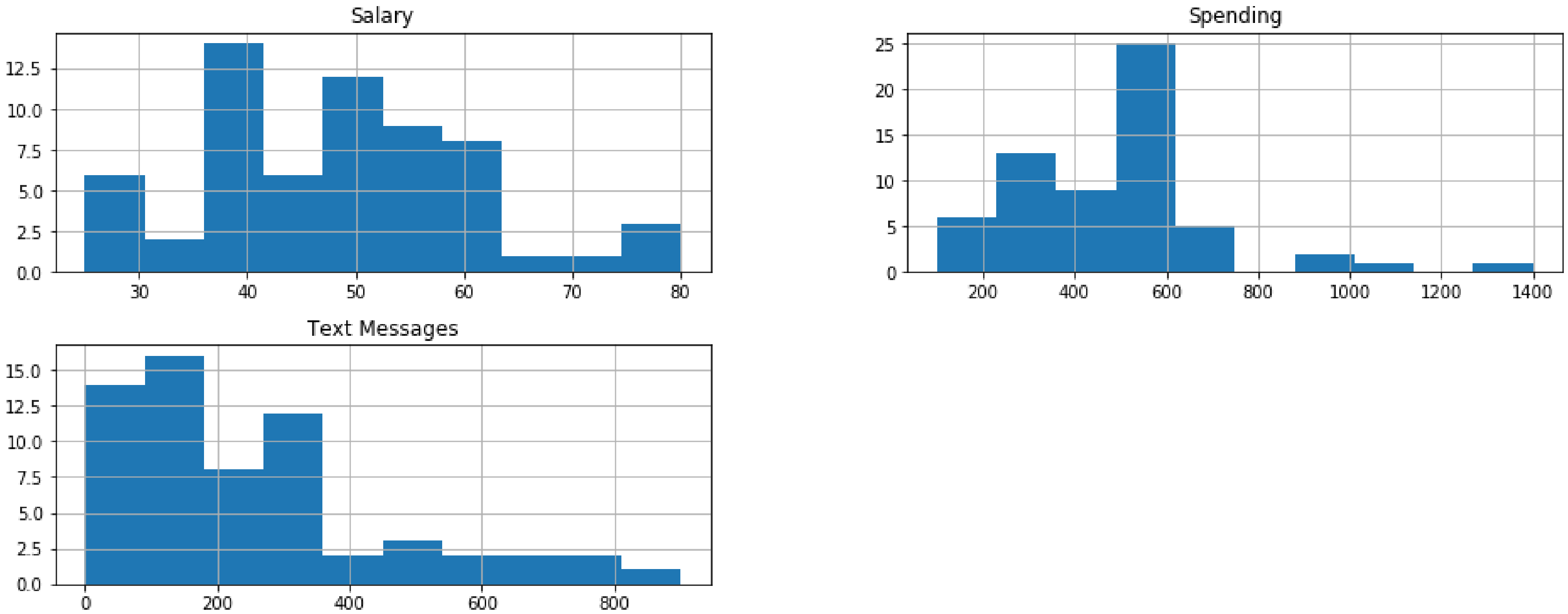
2.3. Based on the above probabilities, do you think that the column variable in each case is independent of Gender?
Justify your comment in each case.

- Major : Major is not independent of gender as the probability of choosing a course is different for Male and Female. For instance, it can be seen that Female are equally likely to choose Accounting and CIS whereas Male are more likely to choose Accounting over CIS.
- Grad Intention : From the conditional probabilities calculated, we can see that Male have higher intentions of graduation than female. Hence, we can say that Grad Intention is not independent of Gender.
- Employment :This column variable is also not independent of gender as we can see that the probabilities are different for male and female. For instance, Male have higher probability of Full time employment over female and female have better part-time employment rate than male.
- Laptop Preference : In this case P(Laptop) = 88.7%, P(Laptop | Male) = 89.6% and P(Laptop | Female) = 87.8% In both cases, the probability is very high but we still cannot say that the probability is independent of Gender as the two probabilities are not exactly same.

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]



Since the histogram do not give us sufficient information to conclude that they follow normal distribution, we will perform statistical tests to verify.

- **Statistical Test Chosen :** Shapiro
- **Level of Significance (alpha) :** 0.05
- **Null Hypothesis :** Population is normally distributed
- **Alternate Hypothesis :** Population is not normally distributed

- **Salary**
  - **P-value** ⟶ 0.028000956401228905

  Since p-value : 0.028 < 0.05, we reject the Null Hypothesis that Salary is Normally Distributed

- **Spending**
  - **P-value** → 1.6854661225806922e-05

  Since p-value : 1.6854661225806922e-05 < 0.05, we reject the Null Hypothesis that Spending is Normally Distributed

- **Text Messages**
  - **P-value** → 4.324040673964191e-06

  Since p-value : 4.324040673964191e-06 < 0.05, we reject the Null Hypothesis that Text Messages is Normally Distributed

Problem 3
ABC asphalt shingles

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

- Null Hypothesis : Mean >= 0.35 pound per 100 square feet
- Alternate Hypothesis : Mean < 0.35 pound per 100 square feet

3.2. For the A shingles, conduct the test of hypothesis and find the p-value.
Interpret the p-value.
Is there evidence at the 0.05 level of significance that the population mean moisture content is less than 0.35 pound per 100 square feet?

- Test : One Sample t-Test
- P-value : 0.14955266
- Since p-value : 0.14955266289815025 > 0.05, we fail to reject the Null Hypothesis. Therefore, for A shingles population mean moisture content is not less than 0.35 pound per 100 square feet

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

- Null Hypothesis : Mean >= 0.35 pound per 100 square feet
- Alternate Hypothesis : Mean < 0.35 pound per 100 square feet

3.4. For the B shingles, conduct the test of the hypothesis and find the p-value.
Interpret the p-value.
Is there evidence at the 0.05 level of significance that the population mean moisture content is less than 0.35 pound per 100 square feet?

- Test : One Sample t-Test
- P-value : 0.00418095
- Since p-value : 0.00418095 < 0.05, we reject the Null Hypothesis. Therefore, we conclude that there is evidence at the 0.05 level of significance that the population mean moisture content for B shingles is less than 0.35 pound per 100 square feet

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

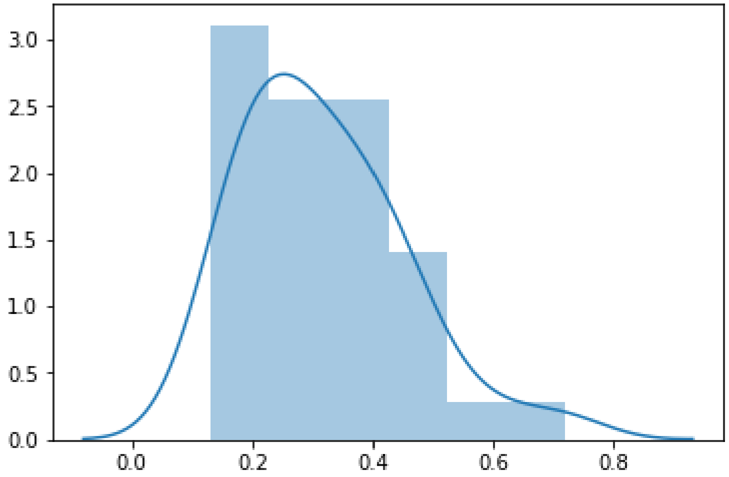
- ➢ Null Hypothesis : Shingles A and B have same population mean
- ➢ Alternate Hypothesis : Shingles A and B do not have same population mean
- ➢ Statistical Test : Two-sample t-test
- ➢ This test assumes the two groups have the same variance (can be checked with tests for equal variance - Levene)
- ➢ P-value : 0.20174966
- ➢ Since p-value : 0.20174966 > 0.05, we fail to reject the Null Hypothesis. Therefore, population means for shingles A and B are equal

3.6. What assumption about the population distribution is needed in order to conduct the hypothesis tests above?

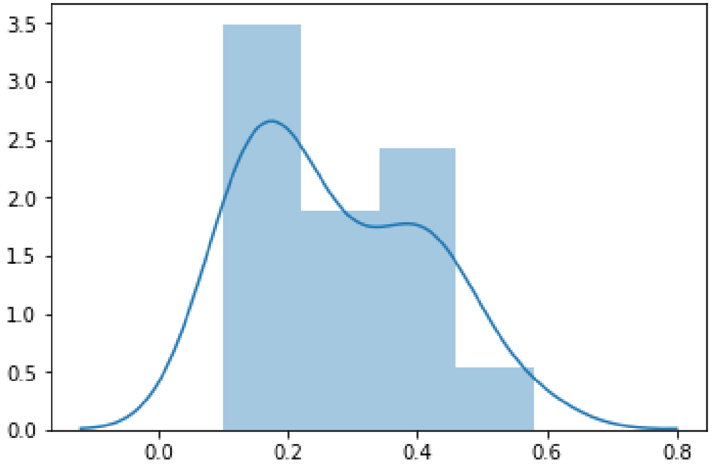
- ➢ The populations (where the samples come from) follow the Normal Distribution

3.7. Check the assumptions made with histograms, boxplots, normal probability plots or empirical rule.

A Singles



B Shingles



From the histogram, we can say that the A & B do not follow Standard Normal Distribution.

3.8. Do you think that the assumption needed in order to conduct the hypothesis tests above is valid? Explain.

- ➢ The populations (where the samples come from) follow the normal distribution.
  - To test this, we have to do a normality test : **Shapiro-Wilk test**

Sample	**P-Value**
A Shingles	0.04267051443457603**5**
B Shingles	0.020027840510010**72**

  - For both the cases p-value < 0.05, therefore we reject the hypothesis that the sample is Normally Distributed. But since our data set is large enough (>30) to apply parametric test although the normality criterion is violated, we can safely go ahead with our assumption and perform the test.

- The variances of the populations are also the same.
  - To test for variance, apply the **Levene test**
  - Null hypothesis : that all samples come from populations with equal variances. The **variance criterion holds true when p > a** (where a is the probability threshold usually set to 0.05)

P-Value	**0.62723121**

  - Since p-value 0.62723121 > 0.05, we fail to reject the Null Hypothesis. Therefore, we can conclude that two samples have equal variance. This assumption is valid.