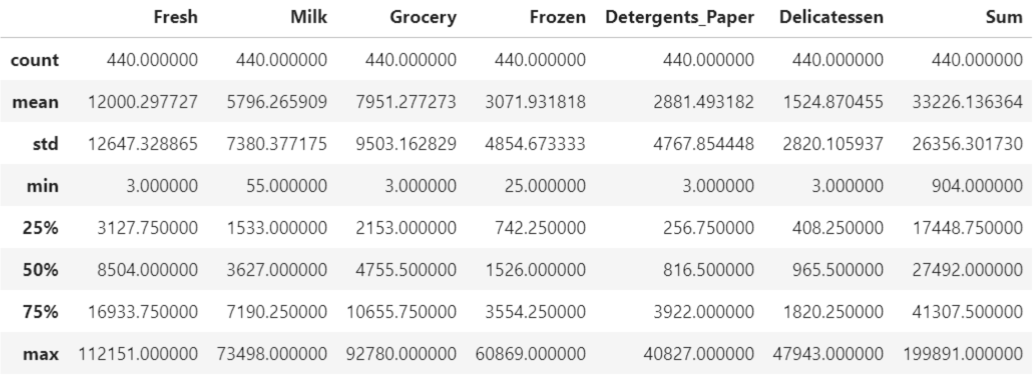
PG DSBA PROJECT 2

Contents

Problem 1

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).

1.1. Use methods of descriptive statistics to summarize data.

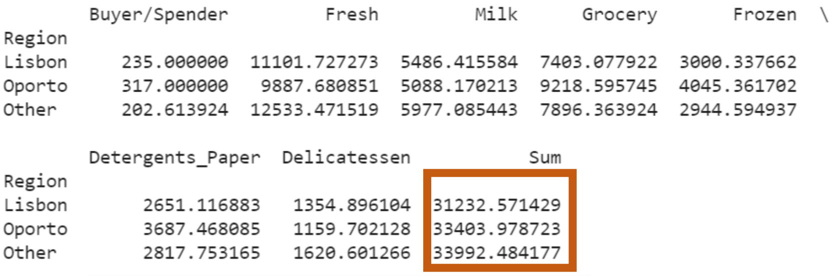


	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Sum
count	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
mean	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455	33226.136364
std	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937	26356.301730
min	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000	904.000000
25%	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000	17448.750000
50%	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000	27492.000000
75%	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000	41307.500000
max	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000	199891.000000

Figure 1: Descriptive statistics of data

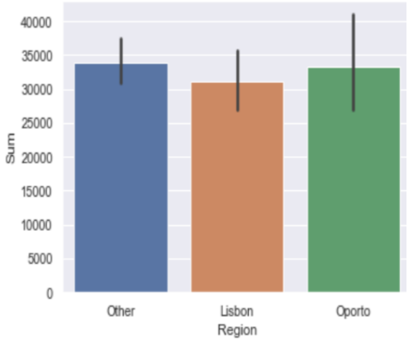
All the variables present in six products are continuous. The annual spending of the item Fresh is the highest among all the other products.

Which Region and which Channel seems to spend more?

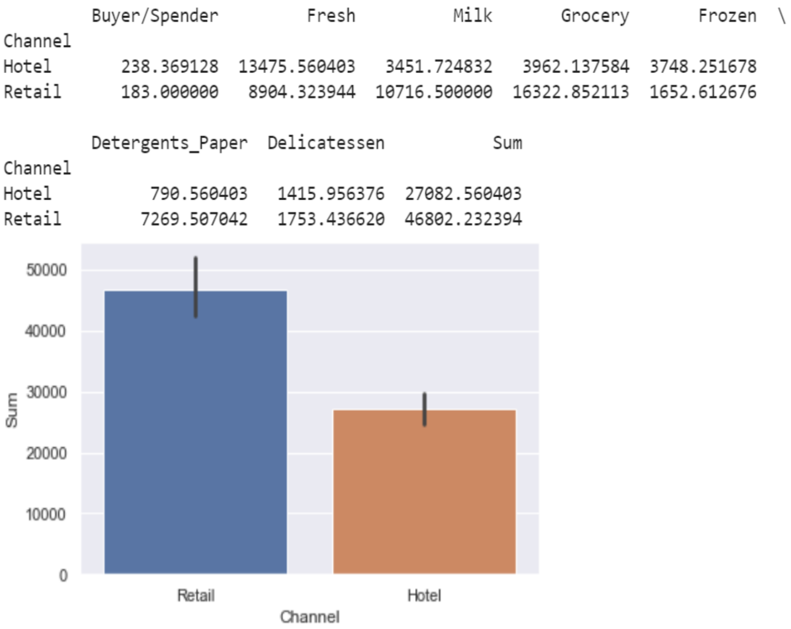


```
        Buyer/Spender         Fresh         Milk       Grocery        Frozen  \
Region
Lisbon     235.000000  11101.727273  5486.415584  7403.077922  3000.337662
Oporto     317.000000   9887.680851  5088.170213  9218.595745  4045.361702
Other      202.613924  12533.471519  5977.085443  7896.363924  2944.594937

        Detergents_Paper  Delicatessen           Sum
Region
Lisbon       2651.116883   1354.896104  31232.571429
Oporto       3687.468085   1159.702128  33403.978723
Other        2817.753165   1620.601266  33992.484177
```

As observed, the total sum of each channels are shown above and to get a visual representation of these values, bar plot is used.



From the above graph, **Region “Others”** is seen to spending **more** in comparison to the other two regions.



```
         Buyer/Spender         Fresh          Milk       Grocery       Frozen  \
Channel
Hotel       238.369128  13475.560403   3451.724832   3962.137584  3748.251678
Retail      183.000000   8904.323944  10716.500000  16322.852113  1652.612676

         Detergents_Paper  Delicatessen           Sum
Channel
Hotel          790.560403   1415.956376  27082.560403
Retail        7269.507042   1753.436620  46802.232394
```

Channel Retail is spending **more** than Hotel.

Which Region and which Channel seems to spend less?

After observations from the above graphs:

1.Among regions, **Lisbon Region** spends the **least.**

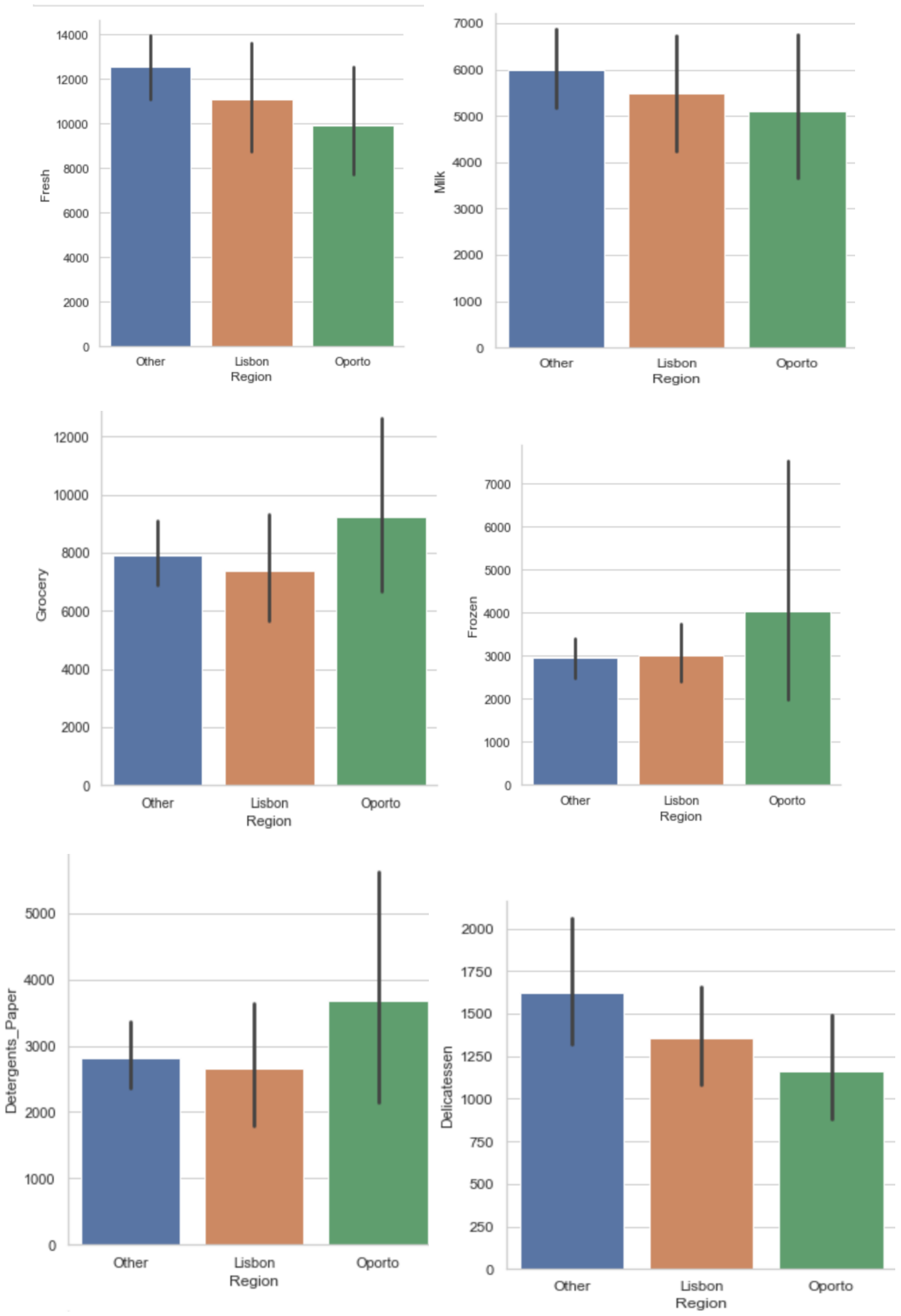
2.In channels, **Hotels** seems to spend the **least.**

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

I started with a simple approach for answering this question i.e. plotting a bar plot for all regions and channels. But using sns.barplot(), the visualized outputs were overlapping each other instead of getting a separate output graph for each product.

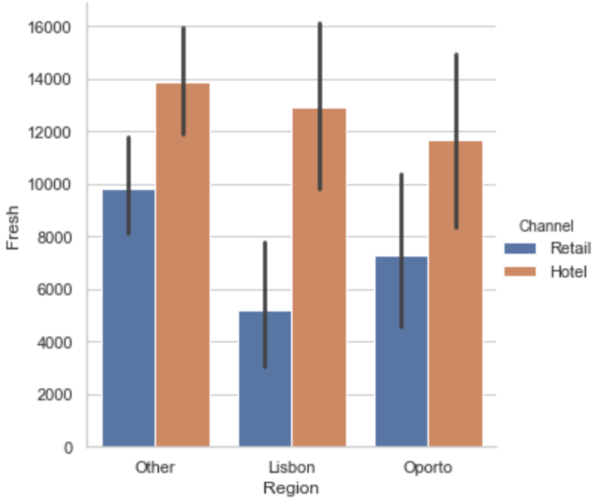
Therefore using sns.catplot(), the problem that occurred with the previous approach has been solved and I was able to get required plots to analyze whether there is a similar behavior among regions and channels.

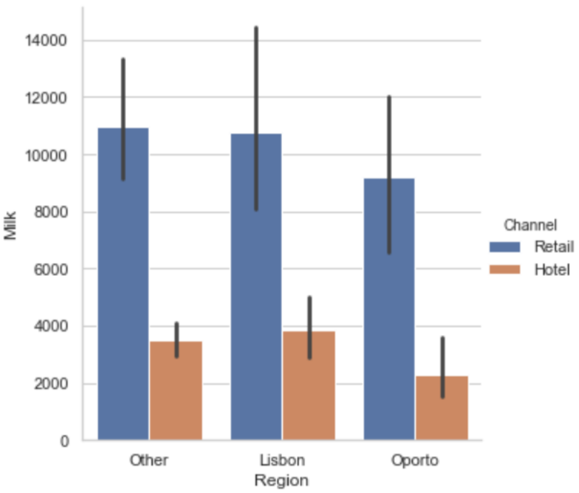
Region wise plots for all six products.

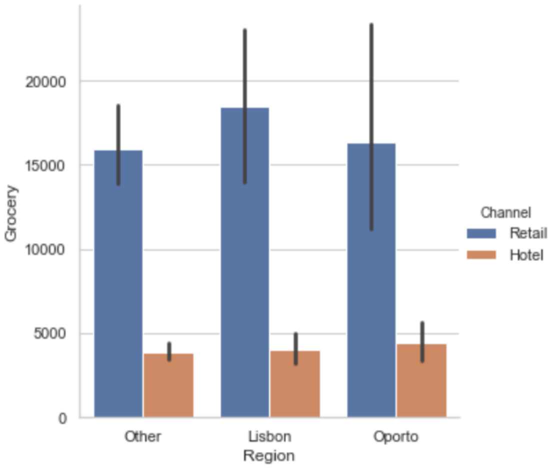


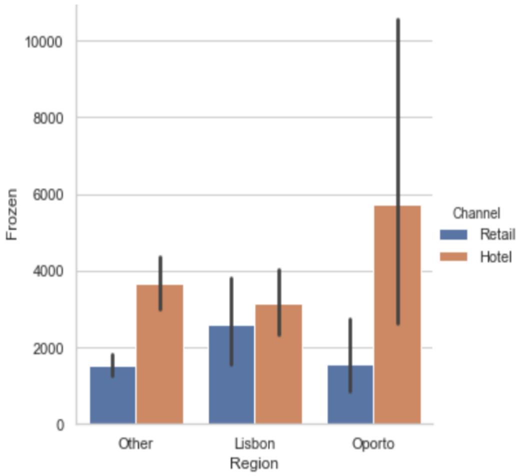
The above bar plots suggest that the behavior among regions is not entirely similar. In the above questions, We have notices that the Others region spends the most, whereas In the case of product Frozen, Lisbon region's spending in terms of Frozen is slightly higher than the 'Others' followed by Oporto. All the other category products except Frozen are similar in behavior in regards to regions

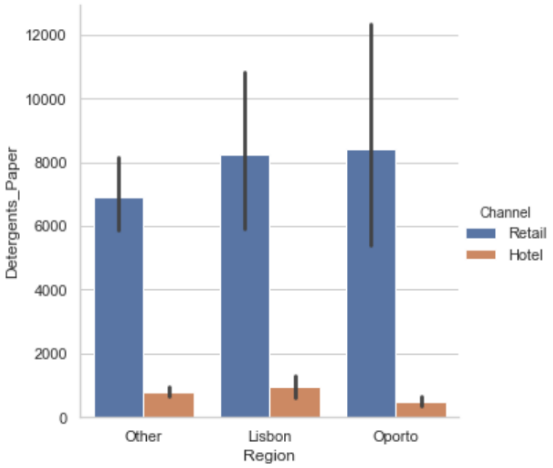
Channel wise plots

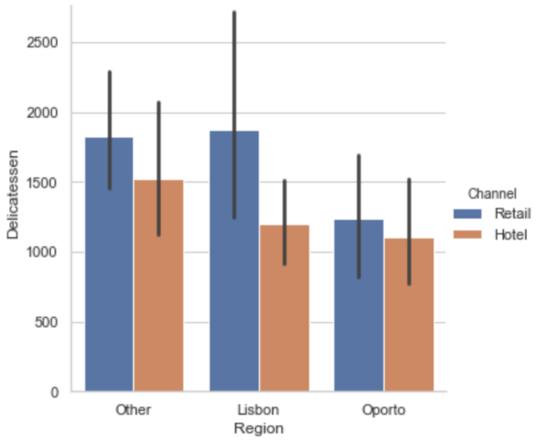












Behavior among channels for category products are:

- Detergent paper is sold significantly higher in Retail Channel than Hotels as there is a huge drop seen when compared.
- Fresh and Frozen products exhibit similar patterns, they are sold more in hotels channels than retails. This behavior varies from the observation made in 1.1Q that the region Retail spends the most.
- The rest, i.e. Milk, Grocery, Detergents Paper and Delicatessen products are sold more in retail channel than hotels.

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

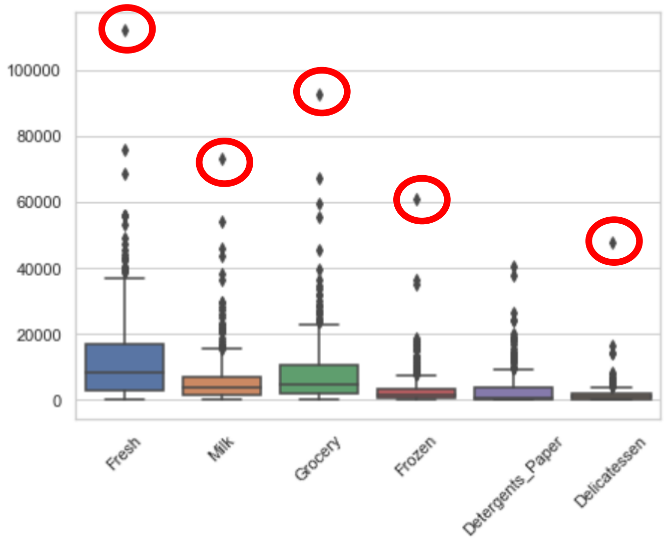
To solve this question, I have decided to use a basic python tool to index all products, so that it would be easier to calculate statistical measures of all products at once. After calculation, all the statistical data of the products are tabularized to draw insights.

Products	**Mean**	**Standard Deviation**	**Variance**	**Coefficient of Variation**
Fresh	12000.297727	12647.328865	1.595914e+08	1.053918
Milk	5796.265909	7380.377175	5.434617e+07	1.273299
Grocery	7951.277273	9503.162829	9.010485e+07	1.195174
Frozen	3071.931818	4854.673333	9.010485e+07	1.580332
Detergent Paper	2881.493182	4767.854448	2.268077e+07	1.654647
Delicatessen	1524.870455	2820.105937	7.934923e+06	***1.849407***

Delicatessen shows the most inconsistent behaviour since it's Coefficient of Variation is highest.

Fresh shows the least inconsistent behaviour since it's Coefficient of Variation is lowest.

1.4. Are there any outliers in the data?



The black diamond shapes are the outliers. Yes, the box plot clearly indicates that there are outliers present in all the items across the product range (Fresh, Milk, Grocery, Frozen, Detergents_Paper & Delicatessen)

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

Based on the exploratory data analysis done above, there is a strong correlation between detergents products and the grocery products. The customers tend to spend more money on these two types of products. The sales in hotels is half lower than the sales in retails. For better performance of channel Hotel, more analysis needs to be done. The sale strategies for Delicatessen, Frozen, detergent products need to be revisited and adapt to a new one if required for better growth of sales in these categories as the annual spendings by the customers are comparatively low in these products.

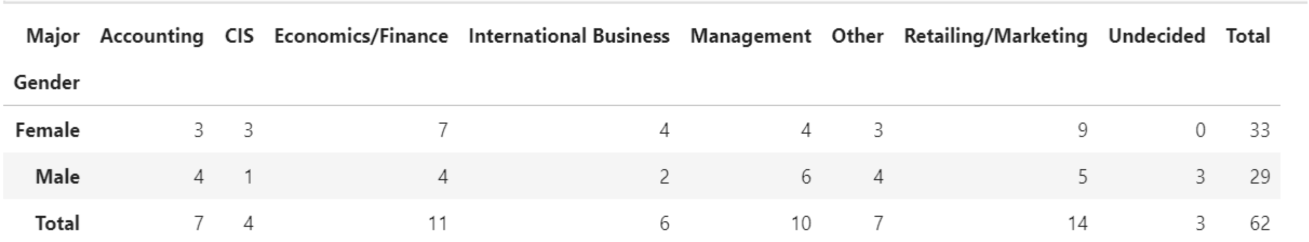
Problem 2

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).

Part I

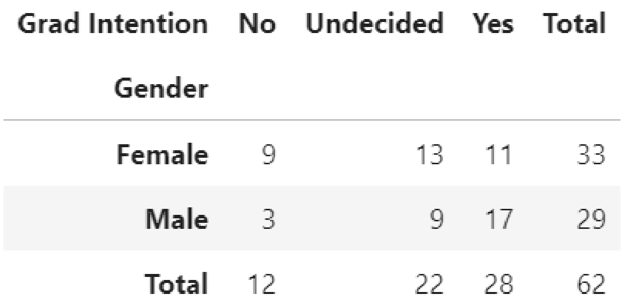
2.1. For this data, construct the following contingency tables (Keep Gender as row variable)

2.1.1. Gender and Major



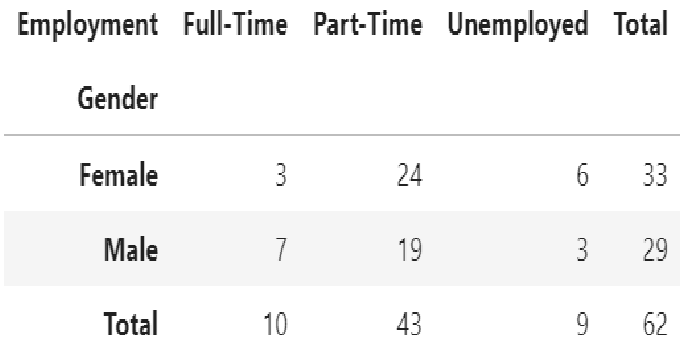
Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	Total
Gender									
Female	3	3	7	4	4	3	9	0	33
Male	4	1	4	2	6	4	5	3	29
Total	7	4	11	6	10	7	14	3	62

2.1.2. Gender and Grad Intention



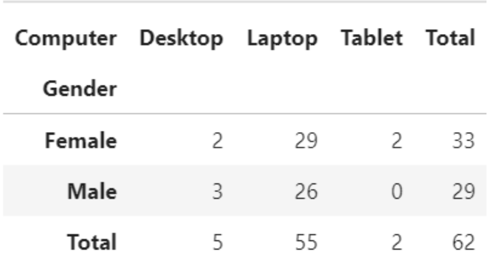
Grad Intention	No	Undecided	Yes	Total
Gender				
Female	9	13	11	33
Male	3	9	17	29
Total	12	22	28	62

2.1.3. Gender and employment



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

2.1.4. Gender and computer



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

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

Calculate the probabilities as asked in python or excel and paste the results here and try to display in mathematical format for each.

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?

Total number of females:33

Total number of males: 29

Total number of people/observations: 62

- Probability that a randomly selected CMSU student will be male is calculated as follows:

P(Male)= Total number of males /Total number of observations

= 29/62

= <u>0.47</u>

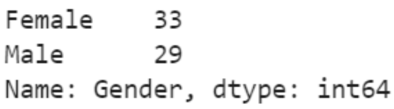
- Probability that a randomly selected CMSU student will be female is calculated as follows:

P(Male)= Total number of females /Total number of observations OR 1 – P(Male)

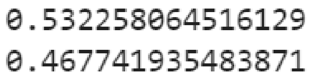
= 33/62 (1-0.47)

= <u>0.53</u>

The following images are the outputs of the calculated probabilities in Python:



```
Female    33
Male      29
Name: Gender, dtype: int64
```



```
0.532258064516129
0.467741935483871
```

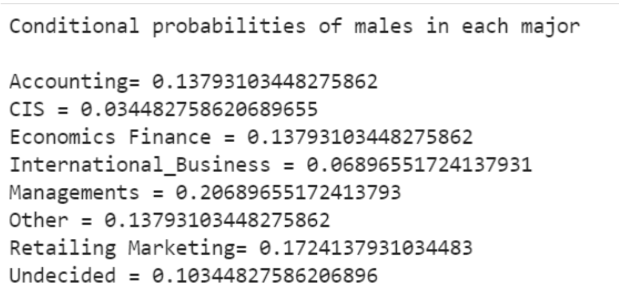
2.2.2. Find the conditional probability of different majors among the male students in CMSU. Find the conditional probability of different majors among the female students of CMSU.

Conditional probabilities of males in each major:

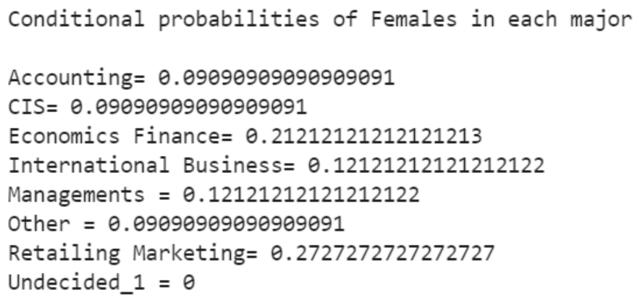
- P(Accounting) = 4/29 = *0.137931*
- P (CIS)= 1/29= *0.034483*
- P (Economics Finance) = 4/29= *0.137931*
- P (International Business) = 2/29= *0.068966*
- P ('Managements) =6/29= *0.206897*
- P (Other) = 4/29= *0.137931*
- P (Retailing Marketing) = 5/29= *0.172414*
- P (Undecided) =3/29= *0.103448*

Conditional probabilities of Females in each major:

- P(Accounting) = 3/33= *0.090909*
- P (CIS)= 3/33= *0.090909*
- P (Economics Finance) = 7/33= *0.212121*
- P (International Business) = 4/33= *0.121212*
- P ('Managements) = 4/33= *0.121212*
- P (Other) = 3/33= *0.090909*
- P (Retailing Marketing) = 9/33= *0.272727*
- P (Undecided) = *0*



```
Conditional probabilities of males in each major

Accounting= 0.13793103448275862
CIS = 0.034482758620689655
Economics Finance = 0.13793103448275862
International_Business = 0.06896551724137931
Managements = 0.20689655172413793
Other = 0.13793103448275862
Retailing Marketing= 0.1724137931034483
Undecided = 0.10344827586206896
```



```
Conditional probabilities of Females in each major

Accounting= 0.09090909090909091
CIS= 0.09090909090909091
Economics Finance= 0.21212121212121213
International Business= 0.12121212121212122
Managements = 0.12121212121212122
Other = 0.09090909090909091
Retailing Marketing= 0.2727272727272727
Undecided_1 = 0
```

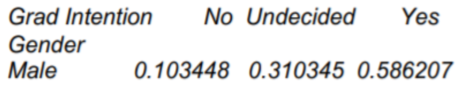
2.2.3. Find the conditional probability of intent to graduate, given that the student is a male. Find the conditional probability of intent to graduate, given that the student is a female.

Conditional probability of intent to graduate in CMSU given that a student is male:

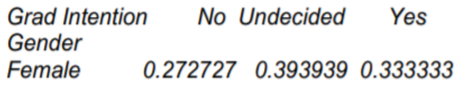
- P(Grad_male) = 17/29= 0.586

Conditional probability of intent to graduate in CMSU given that a student is female:

- P(Grad_Female) =11/33= 0.333



```
Grad Intention        No  Undecided       Yes
Gender
Male            0.103448   0.310345  0.586207
```



```
Grad Intention        No  Undecided       Yes
Gender
Female          0.272727   0.393939  0.333333
```

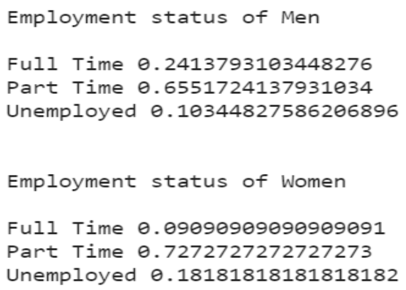
2.2.4. Find the conditional probability of employment status for the male students as well as for the female students.

Conditional probability of employment status in CMSU given that a student is male:

- P (full time) = 7/29 = *0.241379*
- P (Part Time) = 19/29= *0.655172*
- P (Unemployed) = 3/29=*0.103448*

Conditional probability of employment status in CMSU given that a student is male:

- P (full time) = 3/33= *0.090909*
- P (Part Time) = 24/33=*0.727273*
- P (Unemployed) = 6/33= *0.181818*

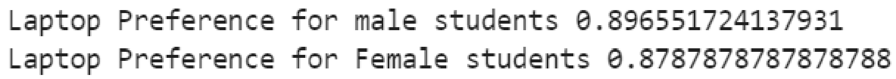


```
Employment status of Men

Full Time 0.2413793103448276
Part Time 0.6551724137931034
Unemployed 0.10344827586206896


Employment status of Women

Full Time 0.09090909090909091
Part Time 0.7272727272727273
Unemployed 0.18181818181818182
```

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

- Conditional Probability of Laptop Preference for male students = 26/29 = 0.896551724137931

- Conditional probability of Laptop Preference for Female students = 29/33 =0.878787878787878



```
Laptop Preference for male students 0.896551724137931
Laptop Preference for Female students 0.8787878787878788
```

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

Case 1

Streams		**Female**		**Male**	
P(Accounting)	**0.11**	P(Accounting|Female)	**0.09**	P(Accounting|Male)	**0.14**
P(CIS)	**0.06**	P(CIS|Female)	**0.09**	P(CIS|Male)	**0.03**
P(Economics/Finance)	**0.18**	P(Economics/Finance|Female)	**0.21**	P(Economics/Finance| Male)	**0.14**
P(International)	**0.10**	P(Economics/Finance|Female)	**0.12**	P(International|Male)	**0.07**
P(Business/Management)	**0.16**	P(Business/Management |Female)	**0.12**	P(Business/Management|Male)	**0.21**
P(Other)	**0.11**	P(Other|Female)	**0.09**	P(Other|Male)	**0.14**
P(Retailing/Marketing)	**0.23**	P(Retailing/Marketing|Female)	**0.27**	P(Retailing/Marketing| Male)	**0.17**

From the above table it is clearly evident that Column variable isn't independent of Gender as the probability of column variable if not equal to the probability of gender

Case 2

Grad Intention		**Female**		**Male**	
P(Yes)	0.45	P(Yes|Female)	0.33	P(Yes|Male)	0.59

Column variable is also independent of Gender in this case.

Case 3

Employment		**Female**		**Male**	
P(full-time)	0.16	P(Full-Time|Female)	0.09	P(Full-Time|Male)	0.24
P(Part – time)	0.69	P(Part-Time|Female)	0.73	P(Part-Time|Male)	0.66
P(Unemployed)	0.15	P(Unemployed|Female)	0.18	P(Unemployed|Male)	0.10

In case 3, Column variable is also independent of Gender in this case.

Case 4

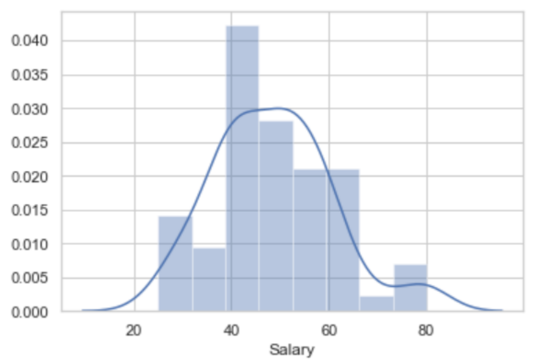
Column		Female		Male	
Desktop	**0.08**	P(Desktop|Female)	**0.06**	P(Desktop|Male)	**0.10**
Laptop	**0.89**	P(Laptop|Female)	**0.88**	P(Laptop|Male)	**0.90**
Tablet	**0.03**	P(Tablet|Female)	**0.06**	P(Tablet|Male)	**0.00**

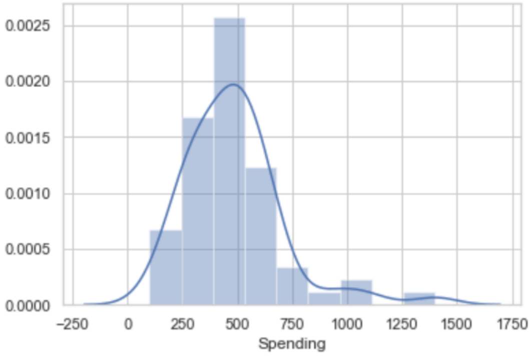
The probability of laptop column is almost equivalent to the probability of column/gender. Hence in thus case, the column variable laptop is independent of Gender.

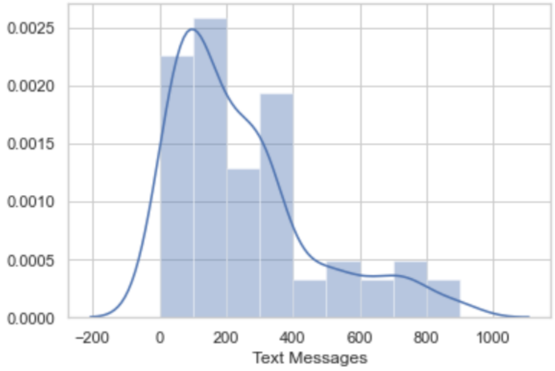
Part II

• 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]

The initial approach used for this question was by plotting distplots for each variable.But it is not reliable as a symmetric plot doesn't necessarily mean symmetric distribution







After researching about a few normality tests used, I have decided to go with Shapiro-Wilk Test. It evaluates a dataset and identifies if the data is drawn from a normal distribution or not.

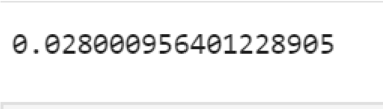
The shapiro() SciPy function will calculate the Shapiro-Wilk on a given dataset. The function returns both the W-statistic calculated by the test and the p-value. [1]

If the value of the Shapiro Wilk test i.e. p-value is greater than 0.05, then the data is normal.

I have performed this test on all the three continuous variables in the dataset, the results are all three of them do not follow a normal distribution.

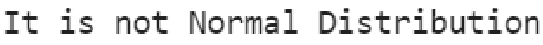
- **Salary**

P-VALUE:



```
0.028000956401228905
```

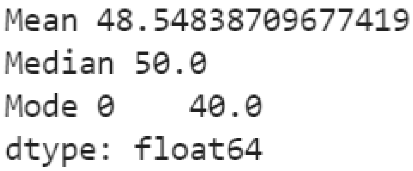
RESULT:



```
It is not Normal Distribution
```

Calculation of mean, median and mode has also been done. As observed below, the values of all three measures should be equal for a normal distribution.

Therefore, it further proves that Salary variable does not follow a Gaussian distribution.



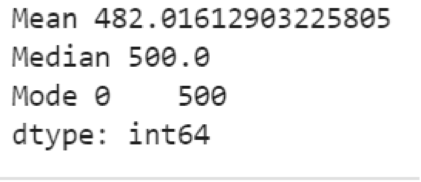
```
Mean 48.54838709677419
Median 50.0
Mode 0    40.0
dtype: float64
```

- **Spending**

P-VALUE: 1.68

RESULT:

It is not Normal Distribution

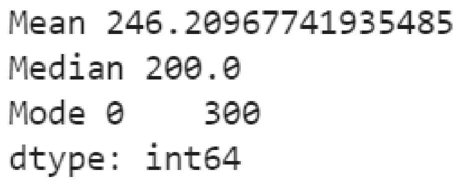


```
Mean 482.01612903225805
Median 500.0
Mode 0    500
dtype: int64
```

The values are not equal, hence not a normal distribution

- **Text Messages**

Result of Wilko Sharpio test is also as the above two i.e. not a normal distribution.



```
Mean 246.20967741935485
Median 200.0
Mode 0    300
dtype: int64
```

To conclude, none of the continuous variables in the dataset follow a Normal Distribution as proved by Wilko Sharpio Test. Also, mean, median and mode values of the variables are not equal for it to be called a normal distribution.

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. 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.

$H0 \le 0.35$ $H1 > 0.35$

Let the Random variable X represent the moisture content for shingles A in pounds per 100 Sq feet

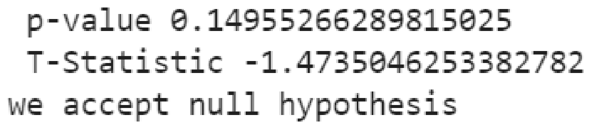
Null Hypothesis:

Ho: u > 0.35

Alternate Hypotheses:

Ha <= 0.35

A single tailed t test is performed.



```
 p-value 0.14955266289815025
 T-Statistic -1.4735046253382782
we accept null hypothesis
```

Hence as per the given sample, Mean moisture content is greater than 0.35 pounds per 100 feet

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:

$H0 \leq 0.35$ $H1 > 0.35$

Let the Random variable Y represent the moisture content for shingles B in pounds per 100 Sq feet

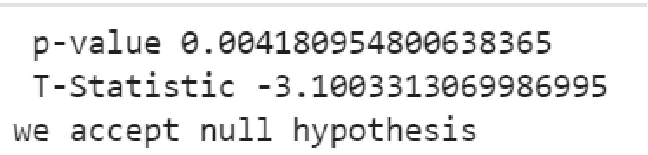
Null Hypothesis:

Ho: u > 0.35

Alternate Hypotheses:

Ha <= 0.35

A single tailed t test is performed. As there are missing values in the dataset, This test is performed with nan_policy = "omit" argument.



```
  p-value 0.004180954800638365
  T-Statistic -3.1003313069986995
we accept null hypothesis
```

Hence as per the given sample, Mean moisture content is greater than 0.35 pounds per 100 feet

3.1 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?

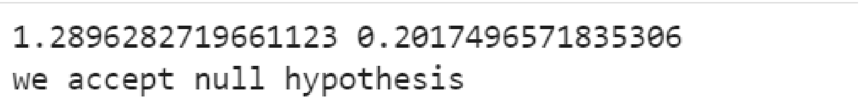
Null Hypothesis

Ho: Ua=Ub (both the data sets are equal)

Alternate hypotheseis

Ha: Ua<>Ub (both the data sets are not equal)

A two sample t test is performed with nan_policy = "omit" argument due to the missing values.



```
1.2896282719661123 0.2017496571835306
we accept null hypothesis
```

Null hypothesis is accepted. Hence we conclude that the population means of both the shingles are equal.

Assumptions before performing the test:

1. Data is continuous in nature

2. Large sample size is used.

3. The data follows Gaussian/Normal Distribution

4. Data is collected as representative of the original population

5. Variance is homogeneous.

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

As mentioned in the above assumptions, the data should be collected as a representative of the original population and as well as it needs to be normal i.e. follow a normal distribution.