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/Restaurant/Café HoReCa, Retail).

1.1. Use methods of descriptive statistics to summarize data.

Description of Categories

- *All Categories are in terms of annual spending hence are continuous in nature*

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
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

Which Region and which Channel seems to spend more?

Regions:

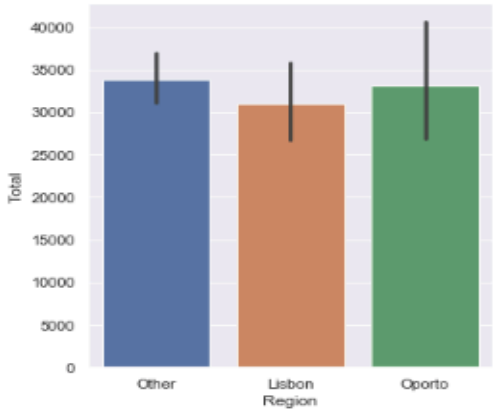
- *Maximum total spending is being done in Others followed by Opporto with Lisbon at the bottom. Category Wise spilt of the spending is given below*

```
Region wise spending is

                Fresh         Milk      Grocery       Frozen  Detergents_Paper  \
Region
Lisbon   11101.727273  5486.415584  7403.077922  3000.337662       2651.116883
Oporto    9887.680851  5088.170213  9218.595745  4045.361702       3687.468085
Other    12533.471519  5977.085443  7896.363924  2944.594937       2817.753165

         Delicatessen         Total
Region
Lisbon    1354.896104  30997.571429
Oporto    1159.702128  33086.978723
Other     1620.601266  33789.870253

<seaborn.axisgrid.FacetGrid at 0x1c21b382e8>
```



Channels:

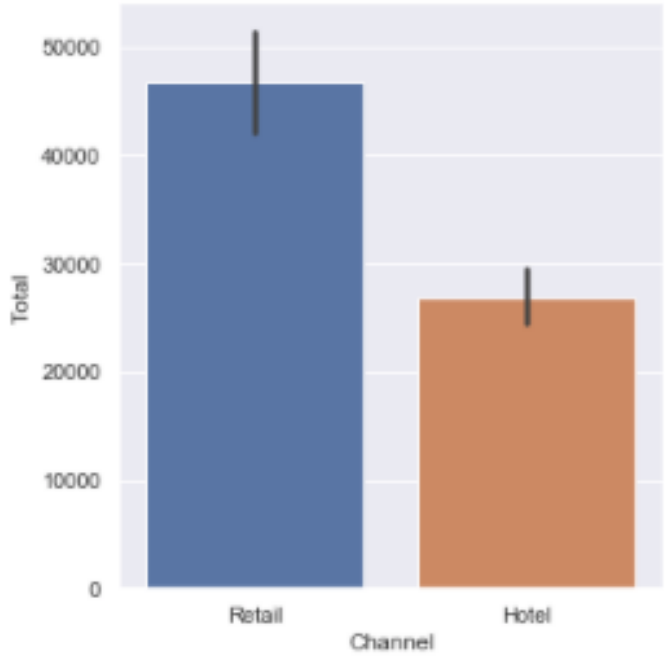
- *Maximum total spending is being done via Retail followed by Hotel. Category wise split is given below:*

```
Channel wise spending is

                Fresh          Milk       Grocery       Frozen  \
Channel
Hotel    13475.560403   3451.724832   3962.137584  3748.251678
Retail    8904.323944  10716.500000  16322.852113  1652.612676

          Detergents_Paper  Delicatessen         Total
Channel
Hotel           790.560403   1415.956376  26844.191275
Retail         7269.507042   1753.436620  46619.232394

<seaborn.axisgrid.FacetGrid at 0x1c21b78ac8>
```



Which Region and which Channel seems to spend less?

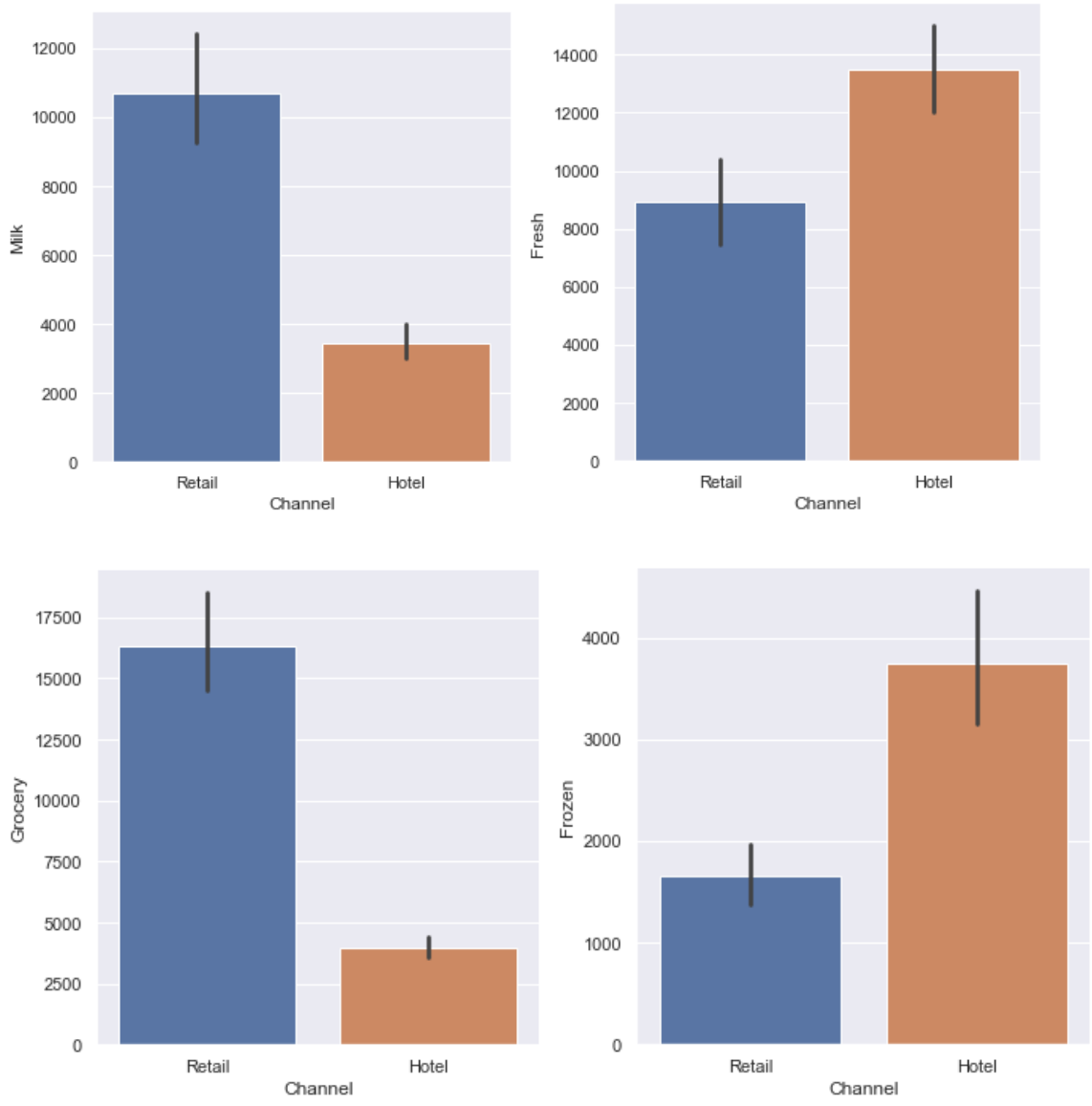
- *Lisbon Region spends the Least*
- *In channels Hotels seems to spend the lease*

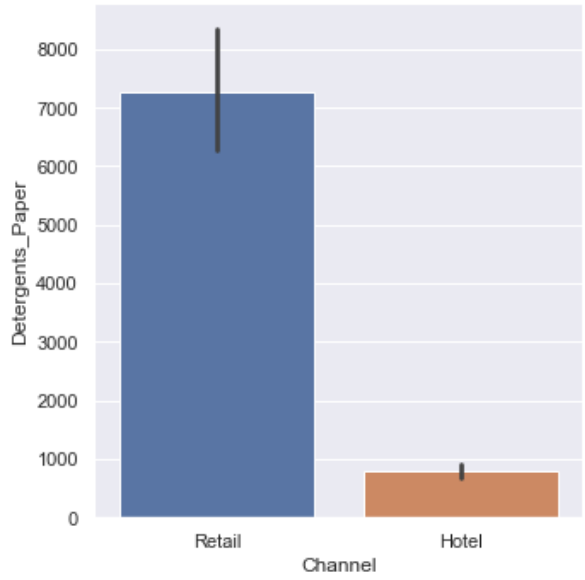
1.2. There are 6 different varieties of items are considered. Do all varieties show similar behaviour across Region and Channel?

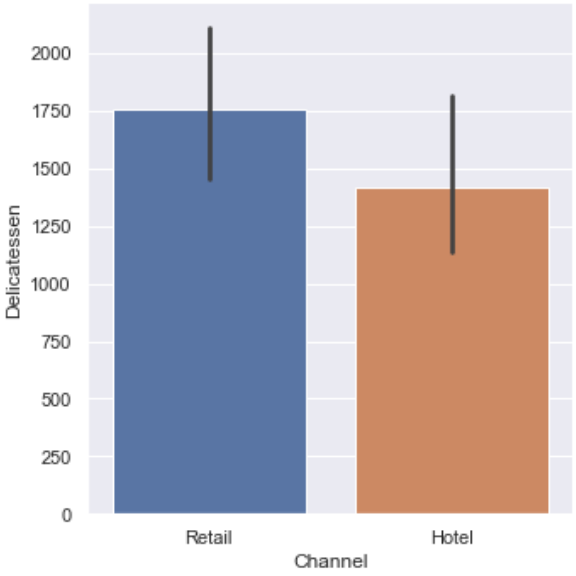
Behaviour of Categories across Channels:

- *The Behaviour Of categories across Channels is not the same.*

- *Milk, Grocery, Detergents Paper and Delicatessen exhibit similar patterns i.e. more in retail sales whereas Fresh and Frozen follow an opposite trend i.e. they are more sold via hotels*

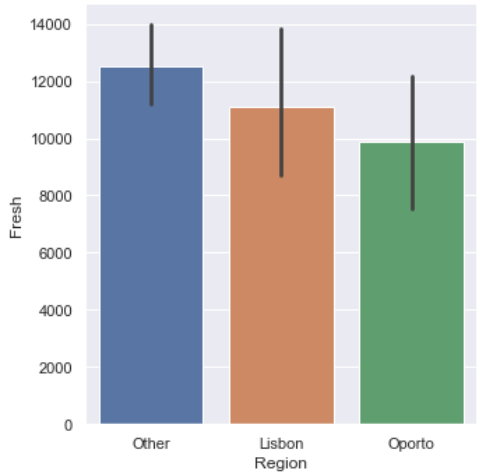


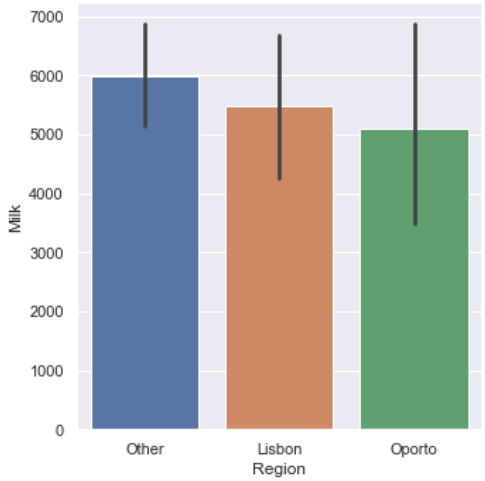


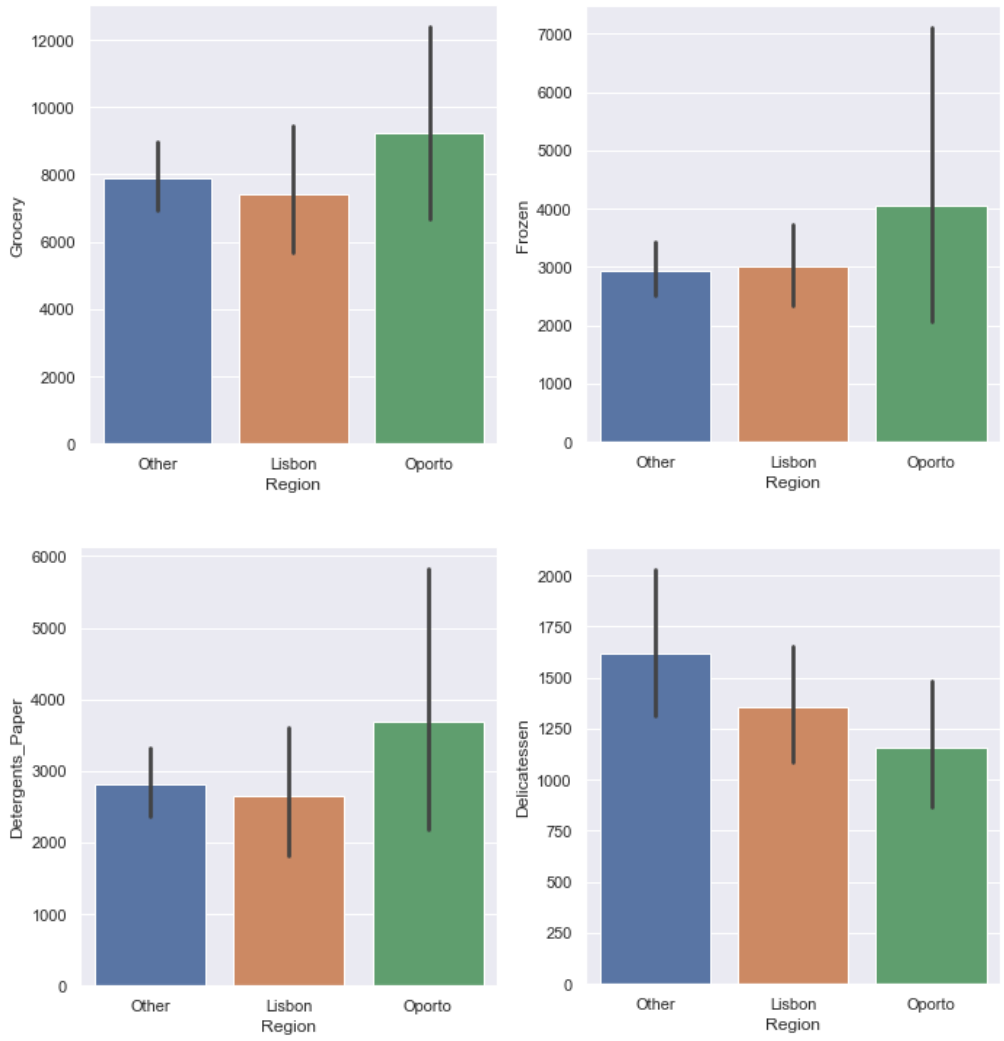


Behaviour of Categories Across Regions:

- *The Behaviour Of categories across Regions is not the same.*
- *Fresh, Milk, and Delicatessen exhibit similar patterns whereas Fresh and Frozen follow an opposite trend.*







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

- *After analysing the coefficient of Variation and the standard deviation, Delicatessen shows the most inconsistent behaviour since its standard deviation expressed as a percentage of mean is highest.*

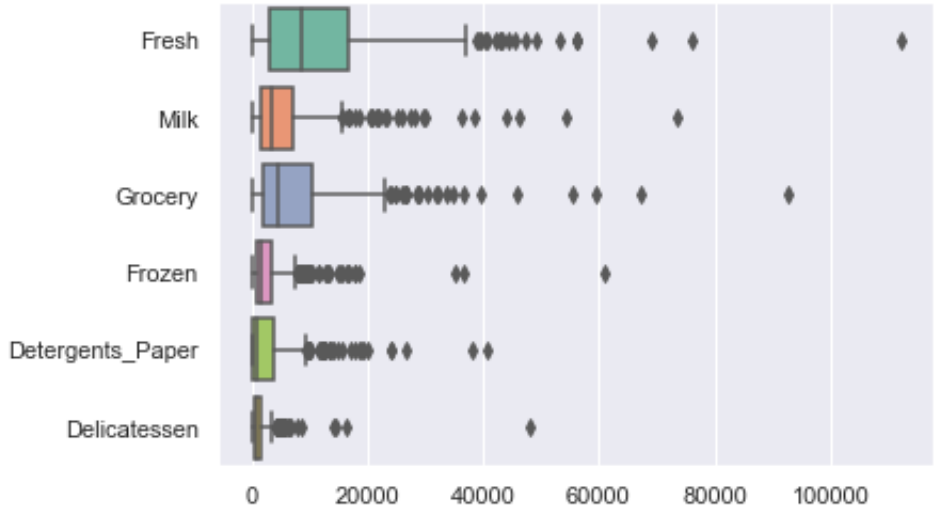
Category	*Mean*	*Standard Deviation*	*Coefficient of Variation*
Fresh	12000.2977	12647.3289	1.053918
Milk	5796.26591	7380.37718	1.273299
Grocery	7951.27727	9503.16283	1.195174
Frozen	3071.93182	4854.67333	1.580332
Detergents_Paper	2881.49318	4767.85445	1.654647
Delicatessen	1524.87046	2820.10594	1.849407

Which items shows the least inconsistent behaviour?

- *Fresh shows the least inconsistent behaviour since its standard deviation expressed as a percentage of mean is lowest.*

1.4. Are there any outliers in the data?

- *There are outliers in the data across the categories, these are shown in the boxplot below (represented by the black dots)*



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

- *Based on the above observations, we can see that the Categories: Frozen, Detergents Paper, and Delicatessen show high variability in sales.*
  - *We need further investigation as to determine the causes of high deviation and variation for these categories.*
- *The sales via hotels channel is almost half the sales via Retail. In categories such as milk, groceries, and Detergents paper, the share of Hotels in disproportionately small compared to the trend in other categories.*
  - *Hence there is ample scope for the improvement in sales in these categories via Hotels channel. Efforts should be made to increase the sales here.*
- *The shares of delicatessen, frozen, Detergents Paper are extremely low in the overall sales. Hence there is ample opportunity for these categories to grow.*
  - *The sales strategy for these categories should be revisited to increase the sales in these categories.*

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*

Tables Created in Python

```
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**	**TOTAL**
GENDER				
FEMALE	9	13	11	33
MALE	3	9	17	29
TOTAL	12	22	28	62

Tables Created in Python

```
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	TOTAL
GENDER				
FEMALE	3	24	6	**33**
MALE	7	19	3	**29**
TOTAL	**10**	**43**	**9**	**62**

Tables Created in Python

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	TOTAL
GENDER				
FEMALE	2	29	2	**33**
MALE	3	26	0	**29**
TOTAL	**5**	**55**	**2**	**62**

Tables Created in Python

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?

- *Total Females: 33*
- *Total Males: 29*
  - *probability that a randomly selected CMSU student will be male = 29/(29+33) = **0.47***
  - *probability that a randomly selected CMSU student will be female= 33/(29+33)= **0.53***

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 probability of different majors in CMSU, given that a student is mal e

Major Accounting CIS Economics/Finance International Business \ Gender
Male 0.137931 0.034483 0.137931 0.068966

Major Management Other Retailing/Marketing Undecided Gender
Male 0.206897 0.137931 0.172414 0.103448

conditional probability of different majors in CMSU, given that a student is fem ale

Major Accounting CIS Economics/Finance International Business \ Gender
Female 0.090909 0.090909 0.212121 0.121212

Major Management Other Retailing/Marketing Undecided Gender
Female 0.121212 0.090909 0.272727 0.0

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 m ale

Grad Intention No Undecided Yes
Gender
Male 0.103448 0.310345 0.586207

conditional probability of intent to graduate in CMSU given that a student is fe male

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

Employment	Full-Time	Part-Time	Unemployed
Gender			
Male	0.241379	0.655172	0.103448

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

Employment	Full-Time	Part-Time	Unemployed
Gender			
Female	0.090909	0.727273	0.181818

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 in CMSU, given that a student is male

Computer	Desktop	Laptop	Tablet
Gender			
Male	0.103448	0.896552	0.0

conditional probability of laptop preference in CMSU, given that a student is female

Computer	Desktop	Laptop	Tablet
Gender			
Female	0.060606	0.878788	0.060606

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.

Column variable will be independent of the gender if

P (Column Variable) = P (Column Variable | Gender)

Gender and Major:

P (Male) = 0.47

P (Female) = 0.53

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

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(International|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*

- *Here it is clear that P (Column Variable) is not equal to the P (Column Variable | Gender)*
- *Hence the two Column variable and the Gender are not independent*

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

P(No)	0.19	P(No|Female)	0.27	P(No|Male)	0.10
P(Undecided)	0.35	P(Undecided|Female)	0.39	P(Undecided|Male)	0.31
P(Yes)	0.45	P(Yes|Female)	0.33	P(Yes|Male)	0.59

- *Here it is clear that P (Column Variable) is not equal to the P (Column Variable | Gender)*
- *Hence the two Column variable and the Gender are not independent*

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**

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

- *Here it is clear that P (Column Variable) is not equal to the P (Column Variable | Gender)*
- *Hence the two Column variable and the Gender are not independent*

Gender and Computer

COMPUTER	DESKTOP	LAPTOP	TABLET	TOTAL
GENDER				
FEMALE	2	29	2	**33**
MALE	3	26	0	**29**
TOTAL	**5**	**55**	**2**	**62**

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

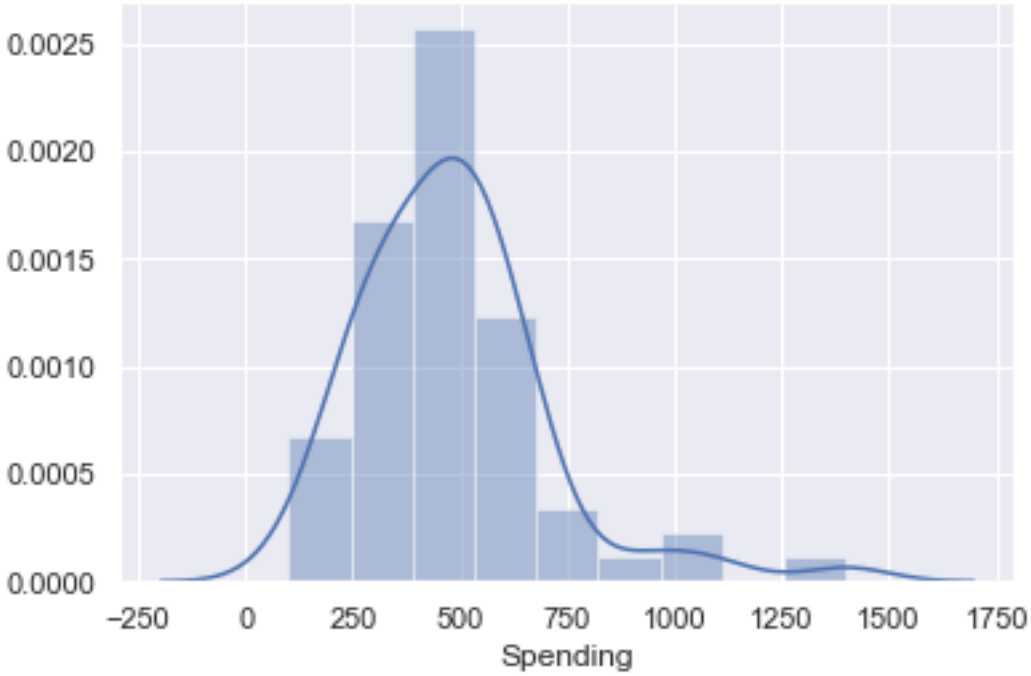
- *Here it is clear that probability of laptop variable is almost equal to the respective P (Column Variable | Gender)*
- *Hence Laptop is independent*
- *Hence the Desktop, tablet and Gender are not independent*

Part II

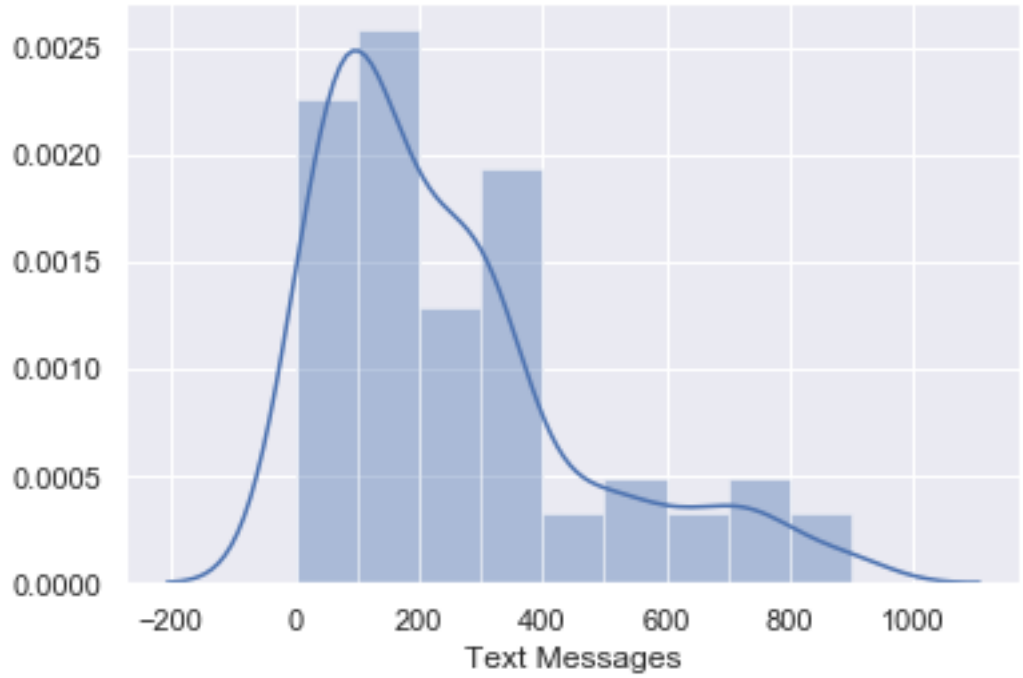
- **2.4. Note that there are three numerical (continuous) variables in the data set, Salary, Spending and Text Messages. For each of them comment whether they follow a normal distribution.**
- **Write a note summarizing your conclusions.**
  [Recall that symmetric histogram does not necessarily mean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

Salary:

Spending



Text Messages:



From the histograms, it is clear that Salary and Spending follow a normalised distribution where as text message may or may not follow a normalised 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.

Hypothesis formation:

For Shingles A:

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 with alpha = 0.05 is performed.

The result obtained is

Mean: 0.316667

t-statistic -1.4735046253382782

p-value 0.14955266289815025

we accept null hypothesis.

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

3.2. 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.

for Shingles B:

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 with alpha = 0.05 is performed.

The result obtained is

Mean of the sample: 0.273548

t-statistic -3.1003313069986995

p-value 0.004180954800638363

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

3.3. 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?

Hypothesis formed:

Null Hypothesis - Ho: Ua=Ub (means of both the data sets are equal)

Alternate hypotheseis - Ha: Ua<>Ub (means of both the data sets are not equal**)**

A two sample t test is performed with alpha=0.05 in python

Result

Mean of sample A: 0.316667

Mean of sample B 0.273548

t-statistic 1.289628271966112

p-value 0.2017496571835328

null hypothesis is accepted.

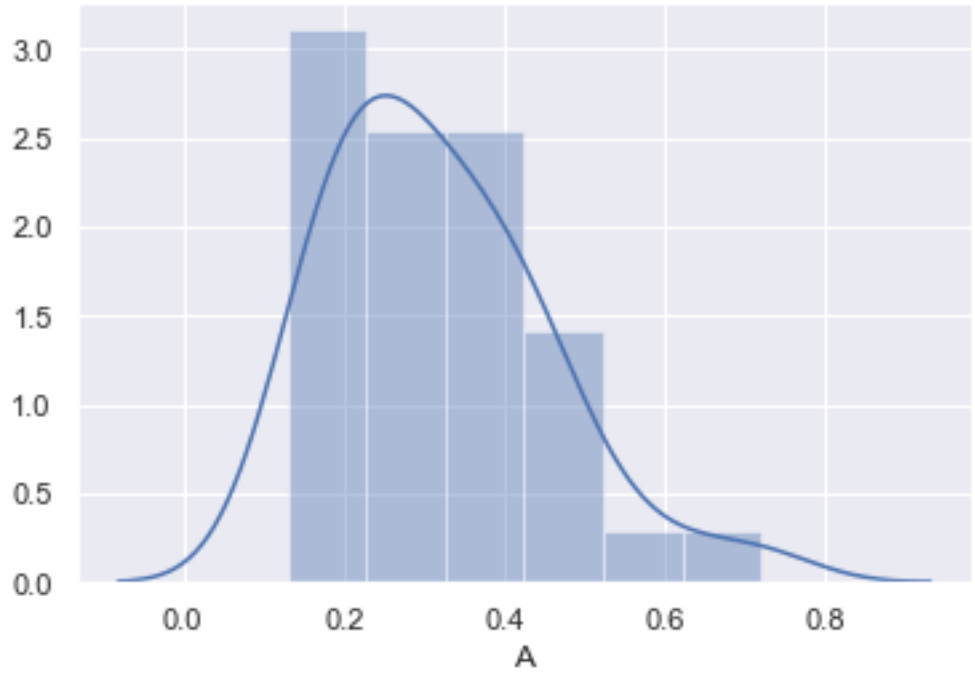
Hence we conclude that the Mean doe both the shingles are equal.

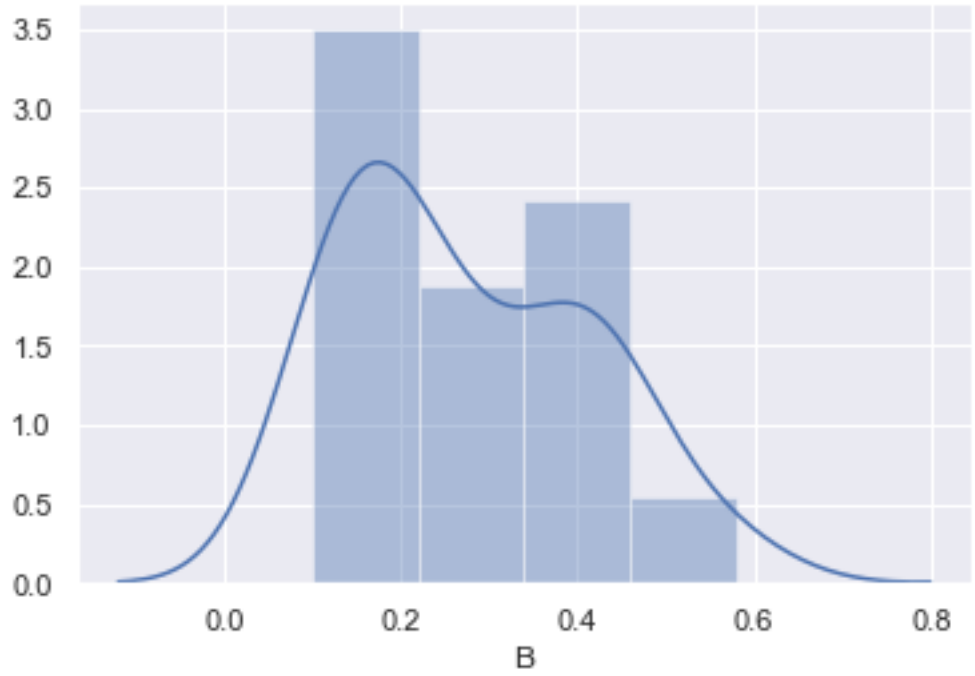
Assumptions:

1. Data is continuous
2. Data is collected is representative of the original population.
3. The data in the sample follow a normal distribution
4. Reasonably large sample size is used.
5. Variance is homogeneous

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

To conduct the hypothesis tests, the basic assumption is that the data follows a normal distribution.




3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
B