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).

1.1.1 Use methods of descriptive statistics to summarize data.

The dimensions of the data set is (440, 9). It consists of 440 rows and 9 columns.

The size or the number of entries in the data set is 3960.

The below table gives the first 5 rows of the data set

Buyer / Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
1	Retail	Other	12669	9656	7561	214	2674	1338
2	Retail	Other	7057	9810	9568	1762	3293	1776
3	Retail	Other	6353	8808	7684	2405	3516	7844
4	Hotel	Other	13265	1196	4221	6404	507	1788
5	Retail	Other	22615	5410	7198	3915	1777	5185

The columns of the data set are 'Buyer/Spender', 'Channel', 'Region', 'Fresh', 'Milk', 'Grocery', 'Frozen', 'Detergents_Paper', 'Delicatessen' as seen by the below output.

```
Index(['Buyer/Spender', 'Channel', 'Region', 'Fresh', 'Milk', 'Grocery',
       'Frozen', 'Detergents_Paper', 'Delicatessen'],
      dtype='object')
```

The below gives the description of the data like the total count, mean, median or Q2 or 50th percentile, standard deviation, Q1 or 25th percentile, Q3 or 75th percentile and the minimum and maximum of each of the columns or attributes.

	count	mean	std	min	25%	50%	75%	max
Buyer/Spender	440	220.5	127.1613	1	110.75	220.5	330.25	440
Fresh	440	12000.3	12647.33	3	3127.75	8504	16933.75	112151
Milk	440	5796.266	7380.377	55	1533	3627	7190.25	73498
Grocery	440	7951.277	9503.163	3	2153	4755.5	10655.75	92780
Frozen	440	3071.932	4854.673	25	742.25	1526	3554.25	60869
Detergents_Paper	440	2881.493	4767.854	3	256.75	816.5	3922	40827
Delicatessen	440	1524.87	2820.106	3	408.25	965.5	1820.25	47943

The following output gives information like total number of entries which are non null and the data types of each of the attribute and memory usage. It indicates except for 2 attributes which are string, rest all are integer data types.

```
<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
```

The above output indicates that there are no missing values in the data which is also verified further. (Please see Python file for the same)

The below output indicates that there are 298 belonging to Hotel and 142 belonging to Retail in 'Channel' attribute and there are 316 records belonging to Other, 77 belonging to Lisbon and 47 in Oporto in 'Region' attribute

```
Hotel     298
Retail    142
Name: Channel, dtype: int64

Other     316
Lisbon     77
Oporto     47
Name: Region, dtype: int64
```

1.1.2 Which Region and which Channel spent the most?

The Channel, 'Hotel' has spent more and

the Region, 'Other' has spent more as seen in the below output.

```
Channel
Hotel     7999569
Retail    6619931
Name: Total Expenditure, dtype: int64

Region
Lisbon     2386813
Oporto     1555088
Other     10677599
Name: Total Expenditure, dtype: int64
```

1.1.3 Which Region and which Channel spent the least?

The Channel Retail has spent less and

the Region 'Oporto' has spent less as seen in the below output.

```
Channel
Hotel     7999569
Retail    6619931
Name: Total Expenditure, dtype: int64

Region
Lisbon     2386813
Oporto     1555088
Other     10677599
Name: Total Expenditure, dtype: int64
```

1.2. There are 6 different varieties of items that are considered. Describe and comment / explain all the varieties across Region and Channel? Provide a detailed justification for your answer. Do all varieties show similar behaviour across Region and Channel? Provide justification for your answer.

For Item Fresh

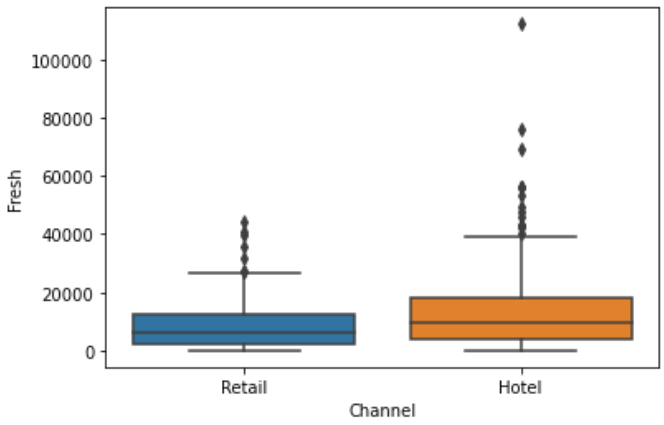
The mean and median expenditure of the item 'Fresh' across channel is given below.
It can be inferred that the mean and median expenditure for the item 'Fresh' is more for 'Hotel ' when compared to 'Retail'.
Also, it is seen that there is considerable variation in the values of mean and median.

Mean

```
Channel
Hotel     13475.56
Retail     8904.32
Name: Fresh,
dtype: float64
```

Median

```
Channel
Hotel     9581.5
Retail    5993.5
Name: Fresh,
dtype: float64
```



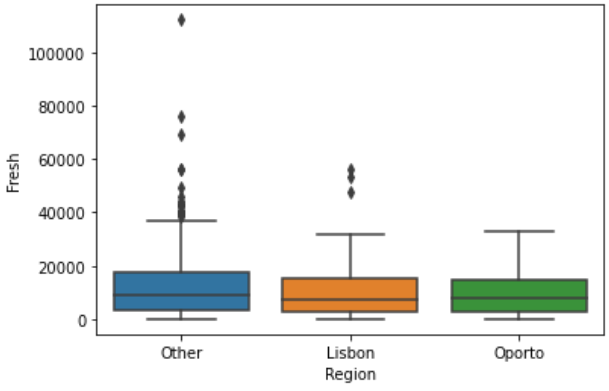
The mean and median expenditure of the item 'Fresh' across the 3 different regions is given below. It can be inferred that the mean and median expenditure for the item 'Fresh' in Region 'Other' is the higher than Lisbon and Oporto.

Mean

```
Region
Lisbon    11101.73
Oporto     9887.68
Other     12533.47
Name: Fresh, dtype: float64
```

Median

```
Region
Lisbon    7363.0
Oporto    8090.0
Other     8752.5
Name: Fresh, dtype: float64
```



For Item Milk

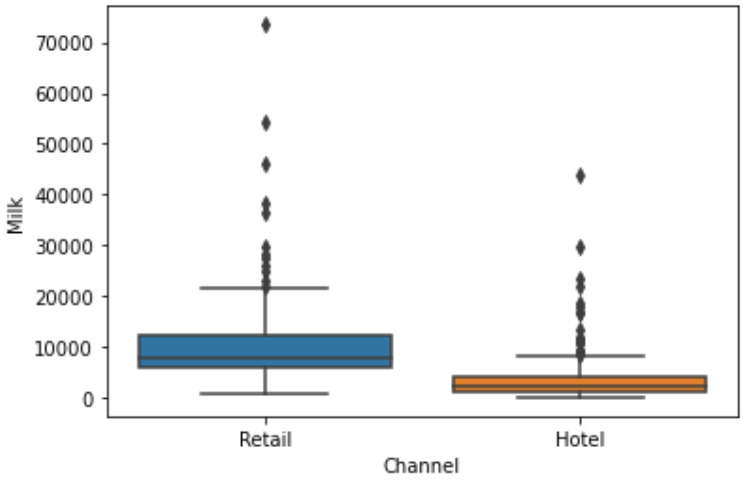
The mean and median expenditure of the item ‘Milk’ across channel is given below. It can be inferred that the mean and median expenditure for the item ‘Milk’ is more for Channel ‘Retail’ when compared to ‘Hotel’.

```
Mean

Channel
Hotel      3451.72
Retail    10716.50
Name: Milk,
dtype: float64

Median

Channel
Hotel     2157
Retail    7812
Name: Milk,
dtype: int64
```



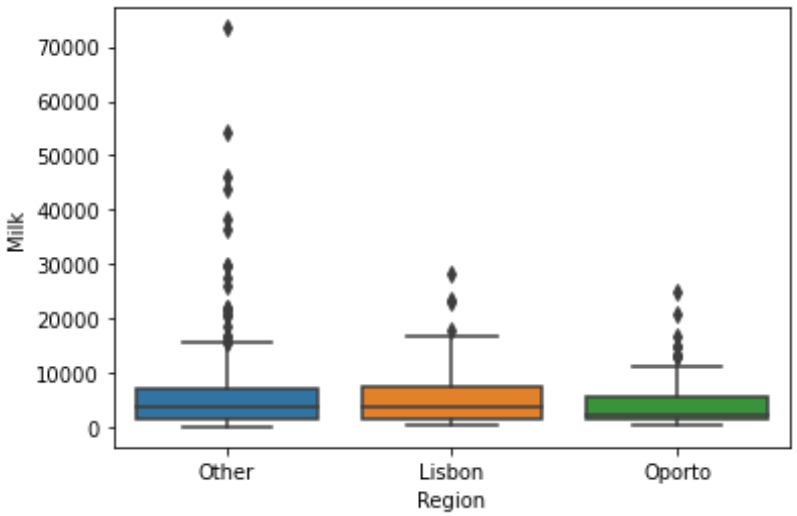
The mean and median expenditure of the item ‘Milk’ across the 3 different regions is given be low. It can be inferred that the mean expenditure for the item ‘Milk’ in Region ‘Other’ is the highest and the median expenditure for the item ‘Milk’ in Region ‘Lisbon’ is the highest. As there are more outliers beyond the 75th percentile, the Mean values are higher than the Median

Mean

```
Region
Lisbon    5486.42
Oporto    5088.17
Other     5977.09
Name: Milk,
dtype: float64
```

Median

```
Region
Lisbon    3748.0
Oporto    2374.0
Other     3684.5
Name: Milk,
dtype: float64
```



For Item Grocery

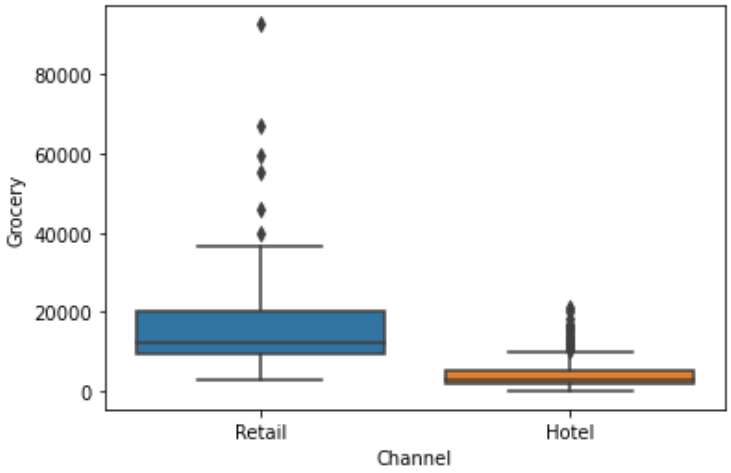
The mean and median expenditure of the item ‘Grocery’ across channel is given below. It can be inferred that the mean and median expenditure for the item ‘Grocery’ is more for Channel ‘Retail’ when compared to ‘Hotel’.

Mean

```
Channel
Hotel      3962.14
Retail    16322.85
Name: Grocery,
dtype: float64
```

Median

```
Channel
Hotel      2684
Retail    12390
Name: Grocery,
dtype: int64
```



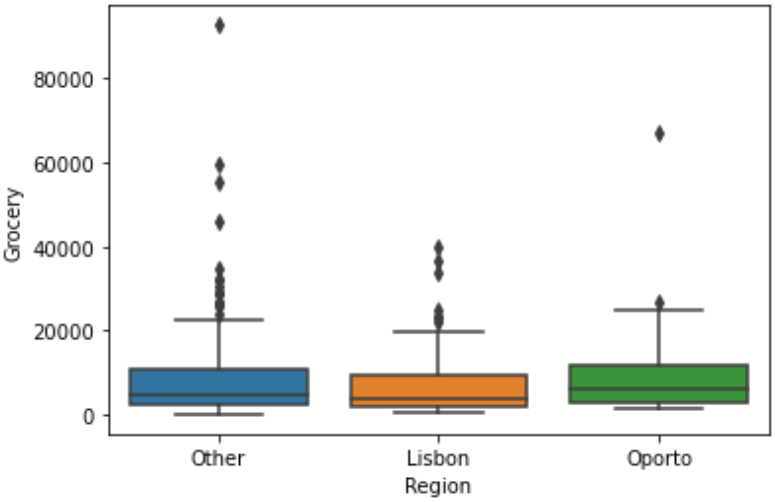
The mean and median expenditure of the item ‘Grocery’ across the 3 different regions is given below. It can be inferred that the mean and median expenditure for the item Grocery’ in Region ‘Oporto’ is the highest.

Mean

```
Region
Lisbon    7403.08
Oporto    9218.60
Other     7896.36
Name: Grocery,
dtype: float64
```

Median

```
Region
Lisbon    3838
Oporto    6114
Other     4732
Name: Grocery,
dtype: int64
```



For Item Frozen

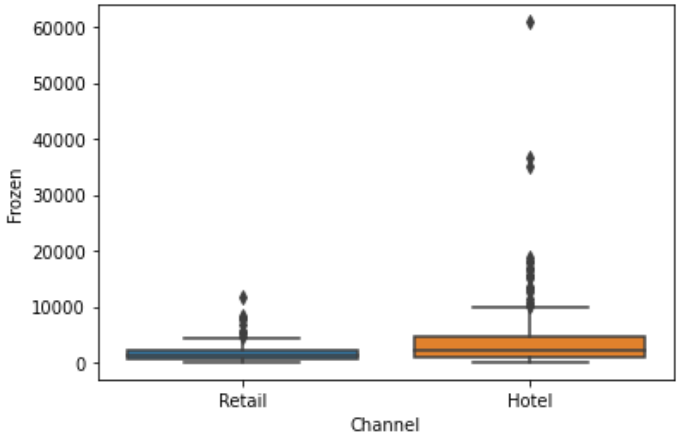
The mean and median expenditure of the item ‘Frozen’ across channel is given below. It can be inferred that the mean and median expenditure for the item ‘Frozen’ is more for Channel ‘Hotel’’ when compared to ‘Retail’.

Mean

```
Channel
Hotel     3748.25
Retail    1652.61
Name: Frozen, dtype: float64
```

Median

```
Channel
Hotel     2057.5
Retail    1081.0
Name: Frozen, dtype: float64
```



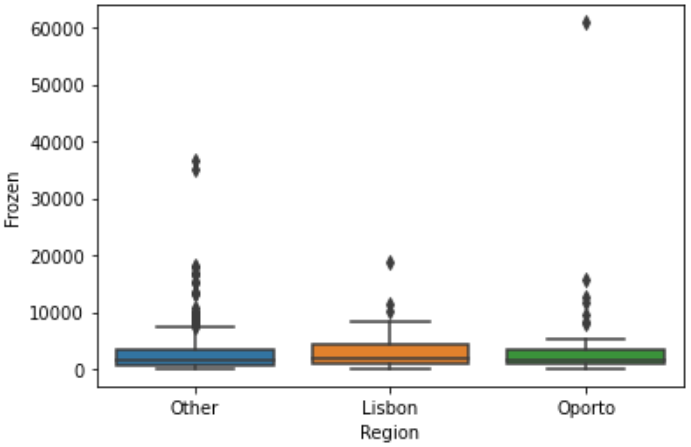
The mean and median expenditure of the item ‘Frozen’ across the 3 different regions is given below. It can be inferred that the mean expenditure for the item ‘Frozen’ in Region ‘Oporto’ is the highest and the median expenditure for the item ‘Frozen’ in Region ‘Lisbon’ is the highest. This difference is due to the presence of outliers beyond Q3, the 75th percentile as seen in the plot.

Mean

```
Region
Lisbon    3000.34
Oporto    4045.36
Other     2944.59
Name: Frozen, dtype: float64
```

Median

```
Region
Lisbon    1801
Oporto    1455
Other     1498
Name: Frozen, dtype: int64
```



For Item Detergents_Paper

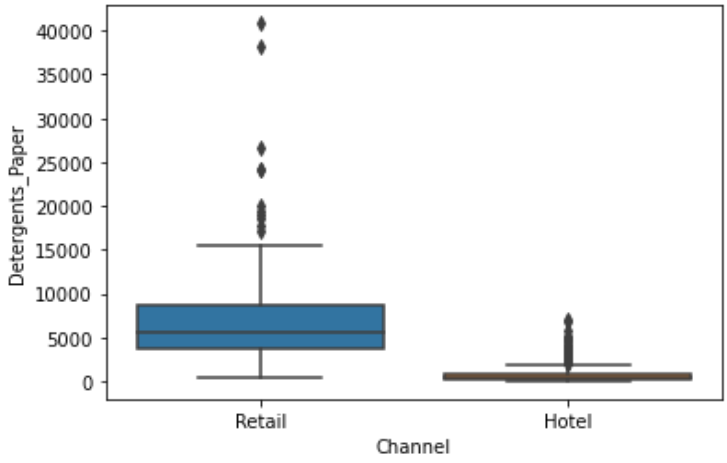
The mean and median expenditure of the item ‘Detergents_Paper’ across channel is given below. It can be inferred that the mean and median expenditure for the item ‘Detergents_ Paper’’ is more for Channel ‘Retail’ when compared to ‘Hotel’.

Mean

```
Channel
Hotel      790.56
Retail    7269.51
Name: Detergents_Paper,
dtype: float64
```

Median

```
Channel
Hotel      385.5
Retail    5614.5
Name: Detergents_Paper,
dtype: float64
```



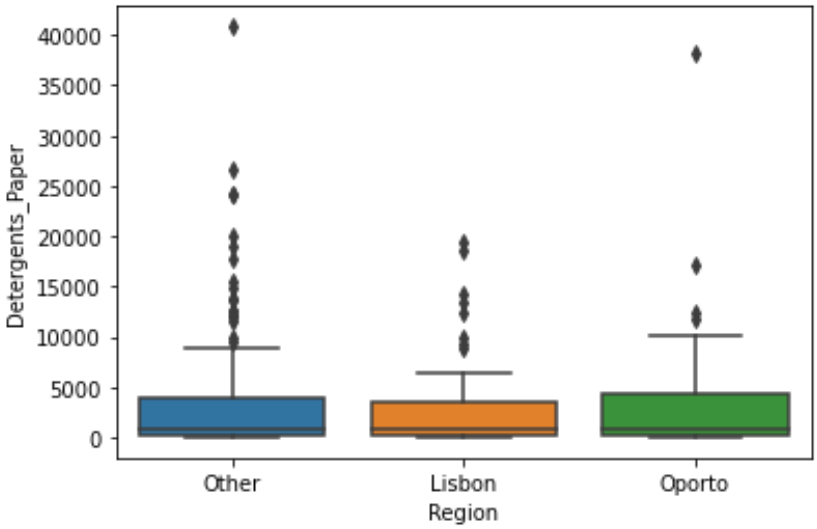
The mean and median expenditure of the item ‘Detergents_Paper’ across the 3 different regions is given below. It can be inferred that the mean expenditure for the item ‘Detergents_Paper’ in Region ‘Oporto’ is the highest and the median expenditure for the ‘Detergents_Paper’ in Region ‘Other’ is the highest.

Mean

```
Region
Lisbon    2651.12
Oporto    3687.47
Other     2817.75
Name: Detergents_Paper,
dtype: float64
```

Median

```
Region
Lisbon    737
Oporto    811
Other     856
Name: Detergents_Paper,
dtype: int64
```



For Item Delicatessen

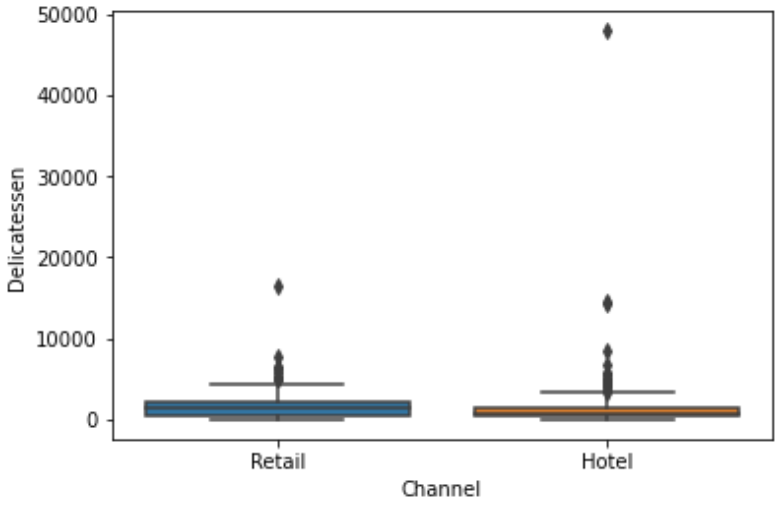
The mean and median expenditure of the item ‘Delicatessen across channel is given below. It can be inferred that the mean and median expenditure for the item ‘Delicatessen’ is more for Channel ‘Retail’ when compared to ‘Hotel’.

Mean

```
Channel
Hotel     1415.96
Retail    1753.44
Name: Delicatessen,
dtype: float64
```

Median

```
Channel
Hotel      821
Retail    1350
Name: Delicatessen,
dtype: int64
```



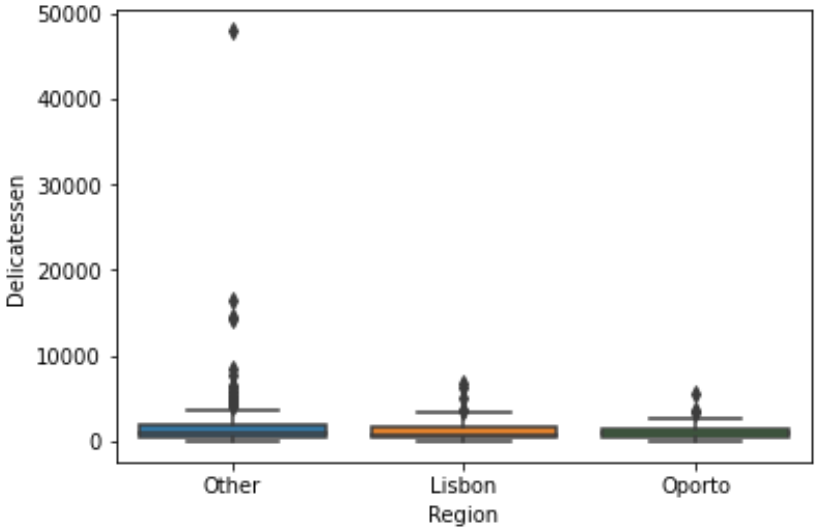
The mean and median expenditure of the item ‘Delicatessen’ across the 3 different regions is given below. It can be inferred that the mean and median expenditure for the item ‘Delicatessen’ in Region ‘Other’ is marginally higher.

Mean

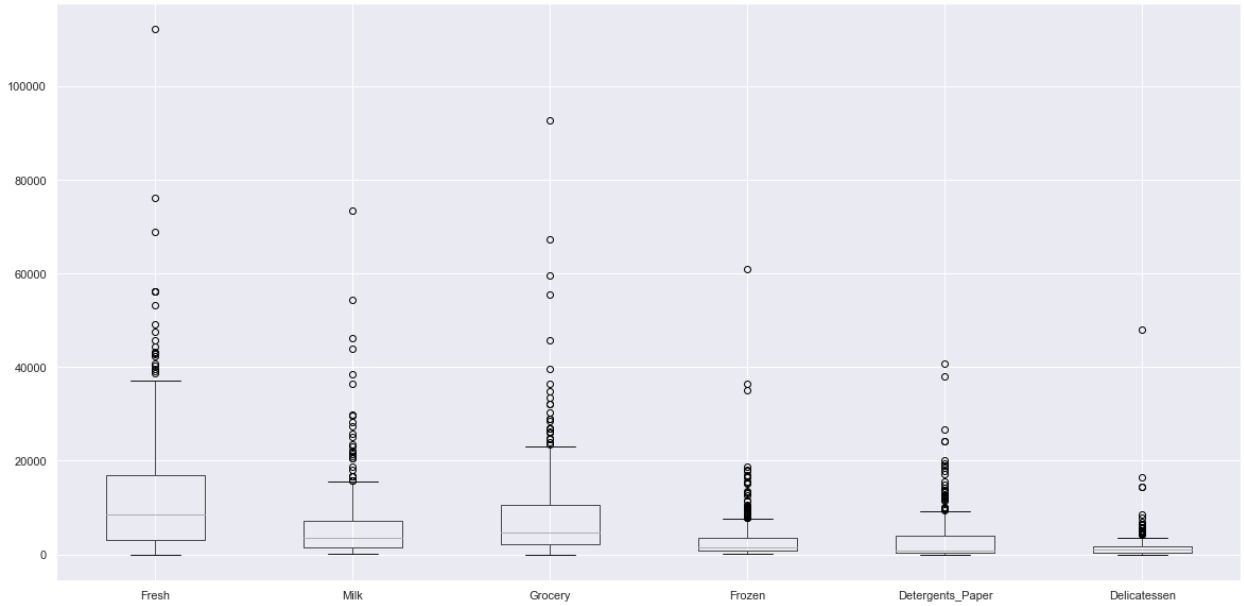
```
Region
Lisbon    1354.9
Oporto    1159.7
Other     1620.6
Name: Delicatessen,
dtype: float64
```

Median

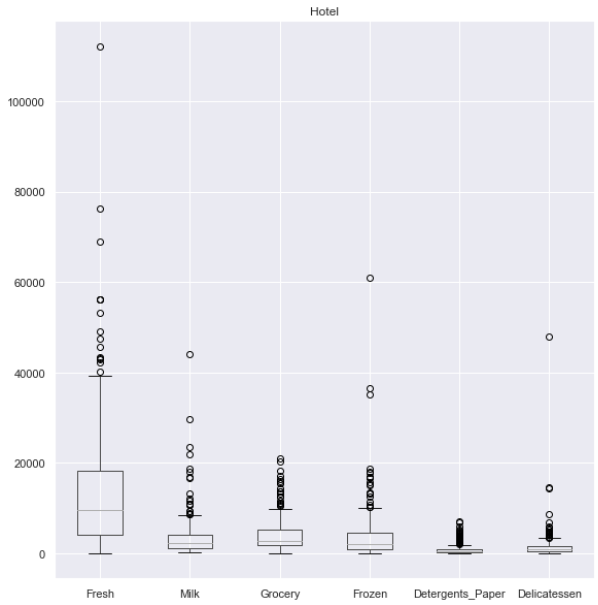
```
Region
Lisbon    806
Oporto    898
Other     994
Name: Delicatessen,
dtype: int64
```

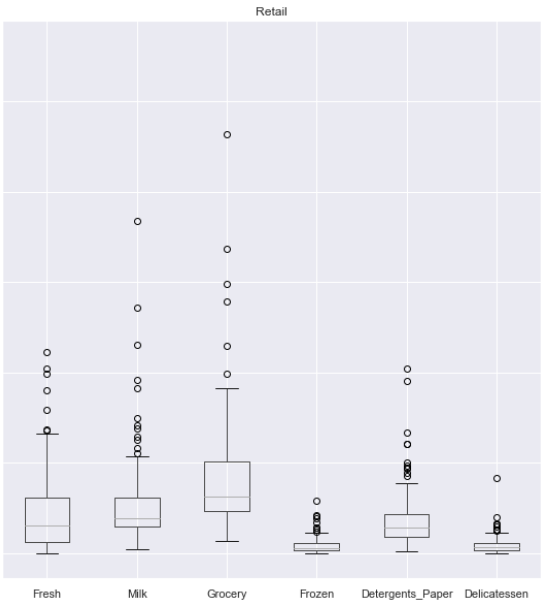


Of all the items, maximum spending was for item Fresh and least for Delicatessen. The expenditure for each item is presented below.

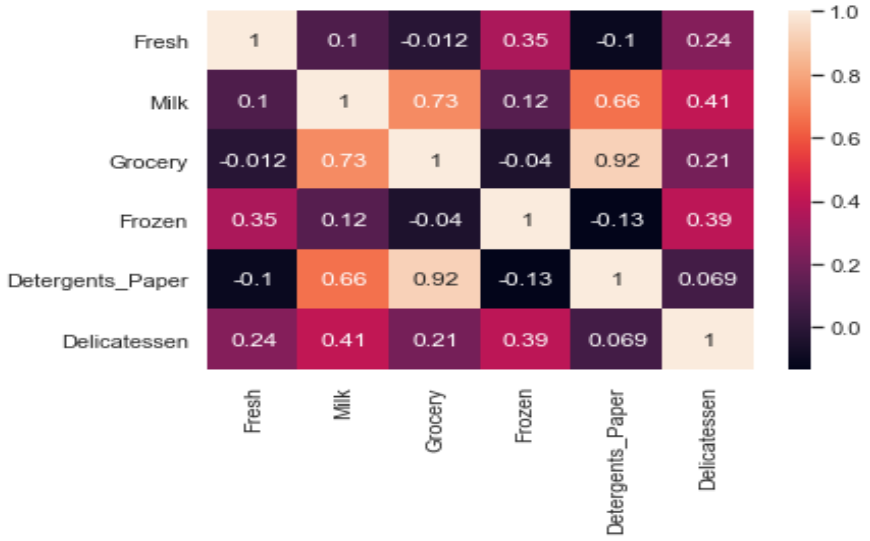


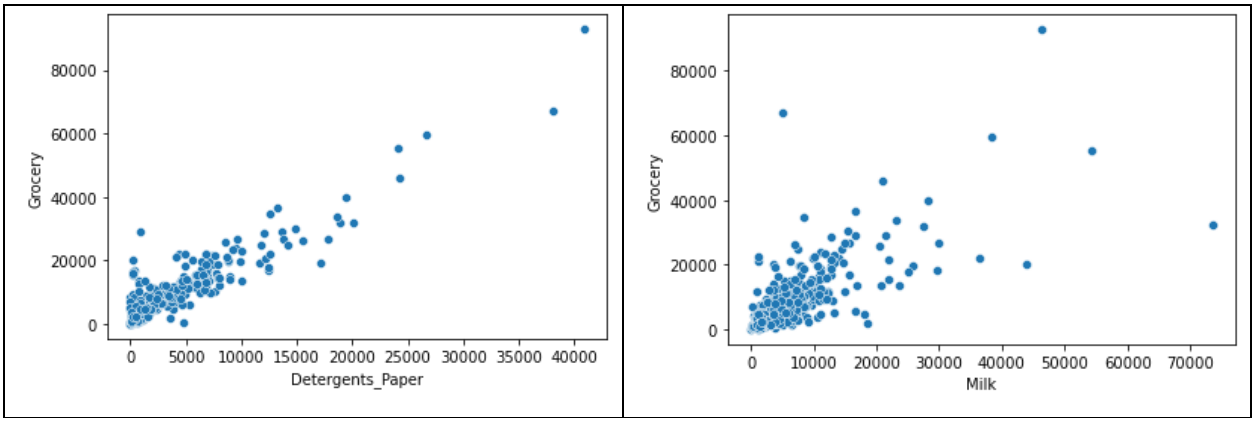
From the below box plots, Hotel has maximum spending for item Fresh and lower spending for items Detergents_Paper. In retail maximum spending was for item Grocery. Fresh and Frozen were found to have higher spending in Hotel and the other items were found to have higher spending in Retail





The correlation of the different items were studied and a positive high correlation was observed between Grocery and Detergent_Paper and Grocery and Milk. The correlation matrix and the correlation between Grocery and Detergent_Paper and Grocery and Milk are presented below.





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

The coefficient of variation gives the extent of variability of the data with respect to the mean. It is calculated as Standard Deviation / Mean

The item ‘Fresh’ shows the least inconsistent behaviour with a coefficient of variation equal to 105.39% and the item ‘Delicatessen’ shows the most inconsistent behaviour with a coefficient of variation equal to 184.94%.

But as far as this data is considered, all the coefficient of variation values for the items are extremely high (All > 100%). This shows that the data is very inconsistent.

1.3 Are there any outliers in the data?

Yes. There are several outliers in the data as can be seen from the boxplots. (boxplots given in solution to Q1.2)

1.5 On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem? Answer from the business perspective

Root cause for very high coefficient of variation for all items needs to be investigated.

The analysis shows that the expenditure for item ’Fresh’ was maximum. Hence, it is recommended that the business plans its inventory by stocking up sufficient quantity of item ‘Fresh’.

A positive high correlation was observed between Grocery and Detergent_Paper and Grocery and Milk. This leads to the conclusion that whenever grocery is sold, milk is also sold and vice versa. Similarly, whenever there is demand for grocery, there is demand for detergents paper

as well. Hence, it is advisable for the business to stock up grocery, milk and detergents paper proportionately.

These insights will help the distributor to procure and plan his stock and optimize on advertising.

The focus should be for stocking up the items Fresh and Frozen for the channel Hotel and the focus should be on Grocery, Milk, Detergent_Paper and Delicatessen for Retail.

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*** data set).

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

2.1.1. Gender and Major

Below is the contingency table output.

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

2.1.2. Gender and Grad Intention

Grad Intention → / Gender ↓	No	Undecided	Yes	All
Female	9	13	11	33
Male	3	9	17	29
All	12	22	28	62

2.1.3. Gender and Employment

Employment → / Gender ↓	Full-Time	Part-Time	Unemployed	All
Female	3	24	6	33
Male	7	19	3	29

2.1.4. Gender and Computer

Computer → / Gender ↓	Desktop	Laptop	Tablet	All
Female	2	29	2	33
Male	3	26	0	29
All	5	55	2	62

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?

The probability that a randomly selected CMSU student will be male is 29/62 =0.468

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

The probability that a randomly selected CMSU student will be female is 33/62= 0.532

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 probability that a randomly selected male student will be an accounting major is 4/29 = 0.138

The probability that a randomly selected male student will be a CIS major is 1/29 = 0.034

The probability that a randomly selected male student will be an Economics/Finance major is 4/29 = 0.138

The probability that a randomly selected male student will be a International Business major is 2/29 = 0.069

The probability that a randomly selected male student will be a Management major is 6/29 = 0.207

The probability that a randomly selected male student will be a Other major is 4/29 = 0.138

The probability that a randomly selected male student will be a Retailing/Marketing major is 5/29 = 0.172

The probability that a randomly selected male student will be Undecided is 3/29 = 0.103

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

The probability that a randomly selected female student will be an accounting major is 3/33 = 0.091

The probability that a randomly selected female student will be a CIS major is 3/33 = 0.091

The probability that a randomly selected female student will be an Economics/Finance major is 7/33=0.212

The probability that a randomly selected female student will be a International Business major is 4/33=0.121

The probability that a randomly selected female student will be a Management major is 4/33=0.121

The probability that a randomly selected female student will be a Other major is 3/33=0.091

The probability that a randomly selected female student will be a Retailing/Marketing major is 9/33=0.273

The probability that a randomly selected female student will be Undecided is 0/33 = 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.

The probability That a randomly chosen student is a male and intends to graduate is 17/62 =0.274

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 is 4/62 = 0.065

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?

P (Male student U Full Time Employment) = P(Male student) + P(Full Time Employment) - P(Male student with full time employment)

= 29/62 + 10/62 – 7/62 = 32/62 = 16/31 = 0.516

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

P(International business major or Management Major | Female Student) = P(International business major | Female Student) + P(Management Major | Female Student) = 4/33 + 4/33 = 8/33 = 0.242

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 graduate intention and being female are independent events?

Grad Intention → / Gender ↓	No	Yes
Female	9	11
Male	3	17

Yes. Graduate intention and being female are independent events as occurrence of one event does not depend on the other.

2.7 Note that there are four numerical (continuous) variables in the data set, GPA, Sala ry, Spending and Text Messages. Answer the following questions based on the data

2.7.1 If a student is chosen randomly, what is the probability that his/her GPA is less t han 3?

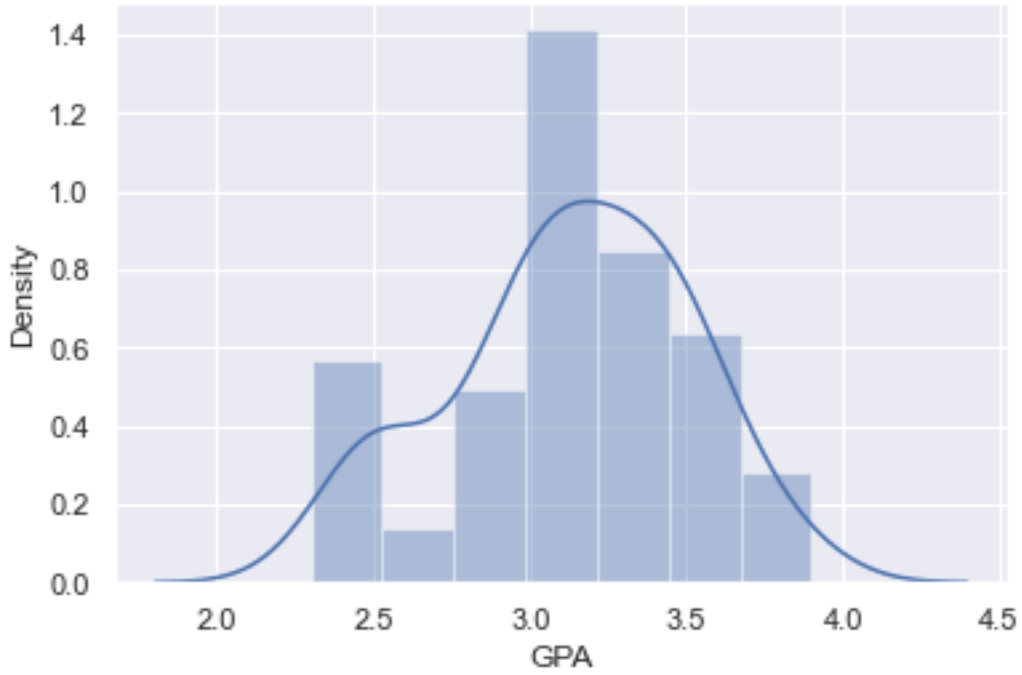
If a student is chosen randomly, the probability that his/her GPA is less than 3 is 17/62 = 0..274

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

The conditional probability that a randomly selected male earns 50 or more is 14/29=0.483
The conditional probability that a randomly selected female earns 50 or more is 18/33=0.545

2.8.1 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.

GPA



We use the Shapiro-Wilk Test to check normality of GPA

Null Hypothesis : H0-the population is normal

Alternate Hypothesis: H1- the population is not normal

#alpha = 0.05, if p value <alpha reject the null hypothesis,

if p value > alpha cannot reject(accept) the null hypothesis

For the attribute GPA, we find the W statistic and P -Value.

Shapiro Result (statistic = 0.9685361981391907, pvalue = 0.11204058676958084)

As pvalue > 0.05, we fail to reject Ho. This shows the 'GPA' distribution is normal.

From the histogram as well as the plot, it can be concluded that the GPA is normally distribute d.

Salary

For the attribute Salary, we find the W statistic and P -Value.

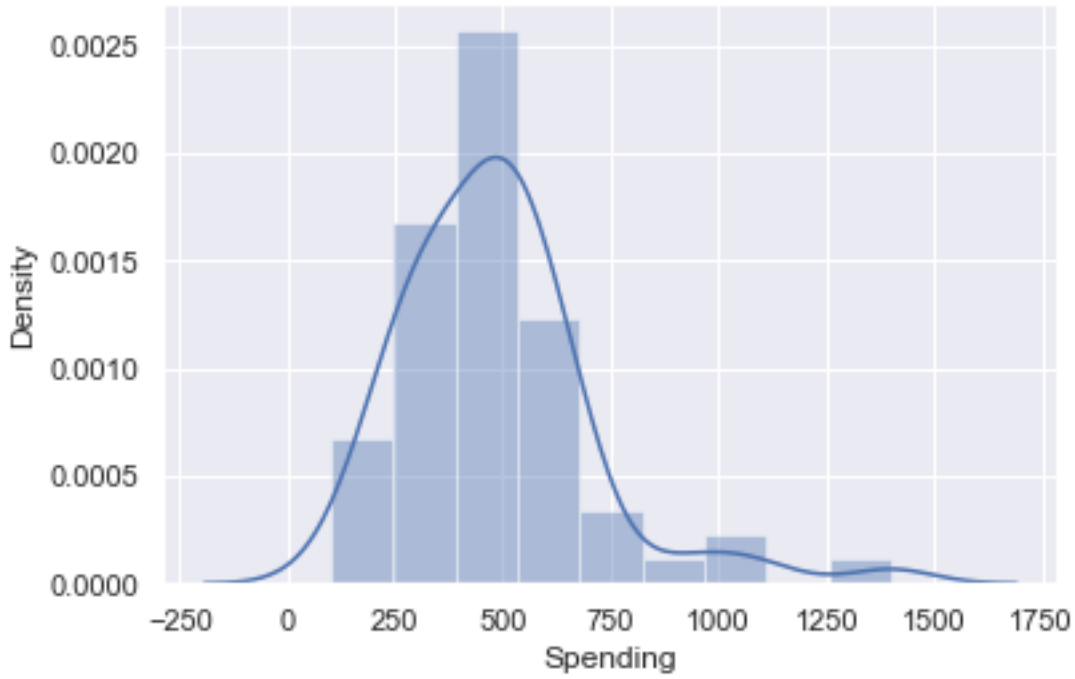
Shapiro Result (statistic = 0.9565856456756592, pvalue = 0.028000956401228905)

As pvalue < 0.05, we reject Ho.

This shows the 'Salary' distribution is not normal.

From the histogram and Shapiro Wilk test, it can be concluded that Salary is not normally distr ibuted.

Spending



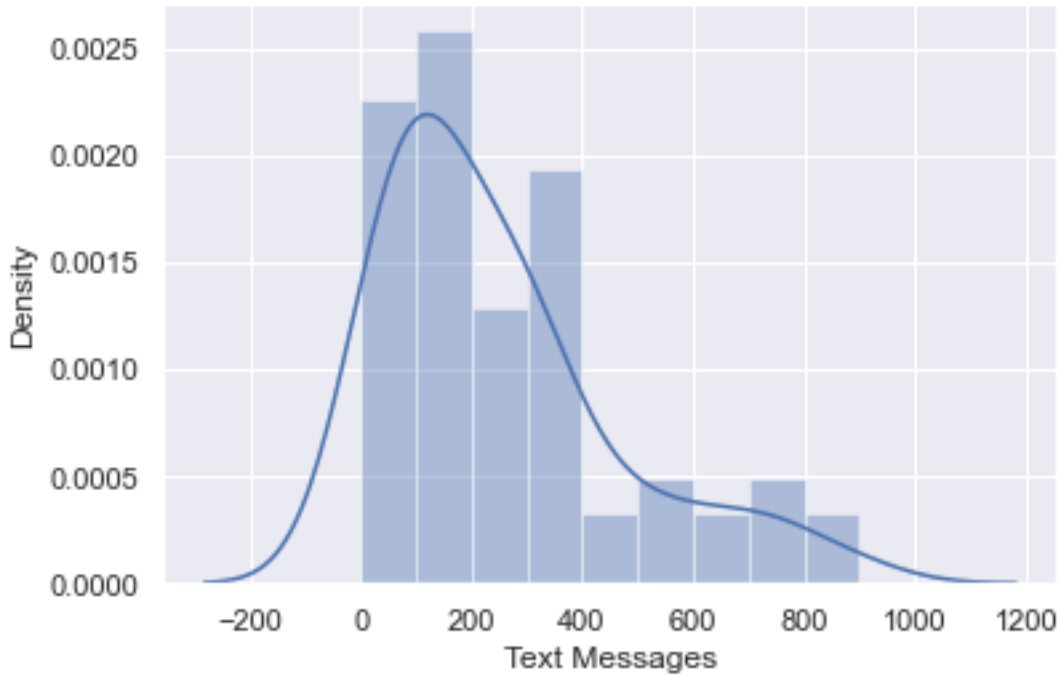
For the attribute Spending, we find the W statistic and P -Value.

Shapiro Result (statistic = 0.8777452111244202, pvalue = 1.6854661225806922e-05)

As pvalue < 0.05, we reject Ho. This shows the 'Spending' distribution is not normal.

From the histogram and Shapiro Wilk test, it can be concluded that Spending is not normally d istributed.

Text Messages



For the attribute Text Messages, we find the W statistic and P -Value.

ShapiroResult (statistic = 0.8594191074371338, pvalue = 4.324040673964191e-06)

As pvalue < 0.05, we reject Ho. This shows the 'Text Messages' distribution is not normal.

From the histogram and Shapiro Wilk test, it can be concluded that Text Messages is not norm ally distributed.

Further, mean and median values of GPA is 3.12 and 3.15 respectively which is very close. Thus, GPA distribution is normal. Whereas for the other 3 numeric variables mean and median are not equa l which shows the distributions are not normal.

2.8.2 Write a note summarizing your conclusions.

Analysis shows that most of the male students prefer management major and most female Students prefer Retail & Marketing.

Intention to graduate among male students is higher (58%) when compared to female students(33%).

The employment rates among male students and female students is almost comparable. (89.6 % vs 81.8%)

Laptop seems to be the preferred gadget among both male and female students when comp ared to desktops and tablets.

The average GPA of the students is 3.12 and GPA is normally distributed. None of the other numeric variables are normally distributed. Intending to graduate and gender of students are independent events (i.e) the occurrence or non occurrence of one does not affect the other.

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 mean moisture contents in both types of shingles are within the permissible limits? State your conclusions clearly showing all steps.

A Shingle

H_0 : the mean moisture content of A shingles is equal to 0.35 pound per 100 square feet.

Sample mean <= 0.35

H_1: the mean moisture content of A shingles is less than 0.35 pound per 100 square feet.

Sample Mean > 0.35

alpha = 0.05

As population standard deviation is not known, one sample t test is performed

T statistic and p-value are determined.

```
One sample t test
t statistic: -1.4735046253382782 p value: 0.07477633144907513

Level of significance: 0.05
We have no evidence to reject the null hypothesis since p value > Level of
significance
Our one-sample t-test p-value= 0.07477633144907513
```

Therefore, for A shingles, the evidence shows that the mean moisture content is less than or equal to 0.35 pounds per 100 square feet and within the permissible limits(i.e.) Sample mean <= 0.35

B Shingle

H_0 : the mean moisture content of B shingles is equal to 0.35 pound per 100 square feet.

Sample mean <= 0.35

H_1: the mean moisture content of B shingles is less than 0.35 pound per 100 square feet.

Sample Mean > 0.35

alpha = 0.05

As population standard deviation is not known, one sample t test is performed

T statistic and p-value are determined.

```
One sample t test
t statistic: -3.1003313069986995 p value: 0.0020904774003191813

Level of significance: 0.05
We have evidence to reject the null hypothesis since p value < Level of
significance
Our one-sample t-test p-value= 0.0020904774003191813
```

Therefore, for B shingles, the evidence shows that the mean moisture content is > 0.35 pounds per 100 square feet and not within the permissible limits(i.e.) Sample Mean > 0.35

3.2 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?

H_0: Population means for shingles A and B are equal ($\mu_A = \mu_B$) (Null Hypothesis)
H_1: Population means for shingles A and B are not equal ($\mu_A \neq \mu_B$) (Alternate Hypothesis)

alpha = 0.05

The 2 sample T test is used as there are 2 samples and the samples are independent and not dependent on each other

The p - value and test statistic are calculated

The scipy.stats.ttest_ind is used to calculate the t-test for the means of two independent samples of scores given the two sample observations. This function returns t statistic and two-tailed p value. This is a two-sided test for the null hypothesis that 2 independent samples have identical average (expected) values. This test assumes that the populations have identical variances.
For this exercise, we are going to first assume that the variance is equal and then compute the necessary statistical values.

tstat = 1.289628271966112
P Value = 0.2017496571835328

two-sample t-test p-value = 0.2017496571835328

As p-value >0.05, there is not enough evidence to reject the null hypothesis in favour of alternative hypothesis

It is concluded that the means of both A Shingles and B Shingles are the same.