Business Report-
SMDM Project
By
Kritika Kumari
Date – 24th April, 2022

Contents

Problem 1

Wholesale Customers Analysis

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).

Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and different measures of dispersion. The data consists of 440 retailers in different regions of Portugal. Analyse the spending of customers on the 6 different varieties of products in each of the regions and also in different channels that is Hotel and Retail.

Data Description

1. Buyer/Spender- It is a continuous variable which is used as a serial number for every row. We are not using this row for our analysis, so we are dropping it for doing our analysis.
2. Channel – It is a categorical variable. There are 2 types of channels in the dataset, Hotel and Retail.
3. Region- It is a categorical variable. There are 3 types of regions in the dataset, Lisbon, Oporto and Other.
4. Fresh- It is a continuous variable which has details of spending on the Fresh type of Product.
5. Milk- It is a continuous variable which has details of spending on Milk.
6. Grocery- It is a continuous variable which has details of spending on Grocery items.
7. Frozen- It is a continuous variable which has details of spending on the Frozen type of Product.
8. Detergents_Paper- It is a continuous variable which has details of spending on the Detergents_Paper .
9. Delicatessen- It is a continuous variable which has details of spending on the Delicatessen type of Product.

Dataset sample :-

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
0	1	Retail	Other	12669	9656	7561	214	2674	1338
1	2	Retail	Other	7057	9810	9568	1762	3293	1776
2	3	Retail	Other	6353	8808	7684	2405	3516	7844
3	4	Hotel	Other	13265	1196	4221	6404	507	1788
4	5	Retail	Other	22615	5410	7198	3915	1777	5185

Exploratory Data Analysis

Dataset Information

```
#   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
```

Wholesale Customer Data consists of 440 rows and 9 columns, out of which 2 columns are categorical of object type and the rest are continuous of integer type. There are no null values for any of the columns.

1.1 Use methods of descriptive statistics to summarize data. Which Region and which Channel spent the most? Which Region and which Channel spent the least?

Using the describe function, we get the information on the descriptive statistics for all the variables in the dataset. This gives us an idea on how the data is dispersed in the dataset for each variable.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Channel	440	2	Hotel	298	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Region	440	3	Other	316	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Fresh	440.0	NaN	NaN	NaN	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0
Milk	440.0	NaN	NaN	NaN	5796.265909	7380.377175	55.0	1533.0	3627.0	7190.25	73498.0
Grocery	440.0	NaN	NaN	NaN	7951.277273	9503.162829	3.0	2153.0	4755.5	10655.75	92780.0
Frozen	440.0	NaN	NaN	NaN	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0
Detergents_Paper	440.0	NaN	NaN	NaN	2881.493182	4767.854448	3.0	256.75	816.5	3922.0	40827.0
Delicatessen	440.0	NaN	NaN	NaN	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0

The wholesale customer data is a dataset which has information of annual spendings in the stores of Portugal across different regions and channels.
Channel- It is a categorical variable in the dataset whose domain value consists of Hotel and Retail. It has 440 rows out of which 298 rows have data for transactions in Hotel and the rest rows have data for transactions in Retail.

Region- It is a categorical variable in the dataset whose domain value consists of Lisbon, Oporto and Other. It has 440 rows out of which 316 rows have data for transactions in Hotel and the rest rows have data for transactions in Lisbon and Oporto.

There are 6 different varieties of products in these stores which are as follows:-
Fresh- This is a continuous variable in the dataset which consists of spending done on Fresh variety of products and has the following characteristics:-

- Mean = 12000.297727
- Standard deviation= 12647.328865
- Minimum = 3.0
- Maximum = 112151.0
- Range = Max - Min = 112148
- Q1= 3127.75
- Q2= 8504.0
- Q3= 16933.75
- Interquartile Range= Q3 - Q1 = 13806

Milk- This is a continuous variable in the dataset which consists of spending done on Milk and has the following characteristics:-

- Mean = 5796.265909
- Standard deviation= 7380.377175
- Minimum = 55.0
- Maximum = 73498.0
- Range = Max - Min = 73443
- Q1= 1533.0
- Q2= 3627.0
- Q3= 7190.25
- Interquartile Range= Q3 - Q1 = 5657.25

Grocery- This is a continuous variable in the dataset which consists of spending done on Grocery items and has the following characteristics:-

- Mean = 7951.277273
- Standard deviation= 9503.162829
- Minimum = 3.0
- Maximum = 92780.0
- Range = Max - Min = 92777.0
- Q1= 2153.0
- Q2= 4755.5
- Q3= 10655.75
- Interquartile Range= Q3 - Q1 = 8502.75

Frozen- This is a continuous variable in the dataset which consists of spending done on Frozen items and has the following characteristics:-

- Mean = 3071.931818
- Standard deviation= 4854.673333
- Minimum = 25.0

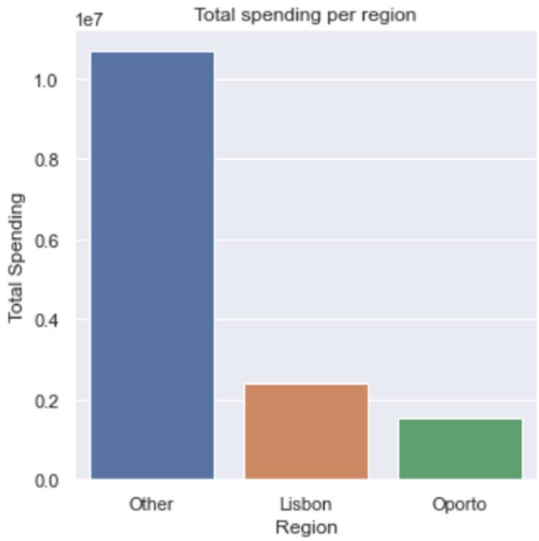
- Maximum = 60869.0
- Range = Max - Min = 60844.0
- Q1= 742.25
- Q2= 1526.0
- Q3= 3554.25
- Interquartile Range= Q3 - Q1 = 2812.0

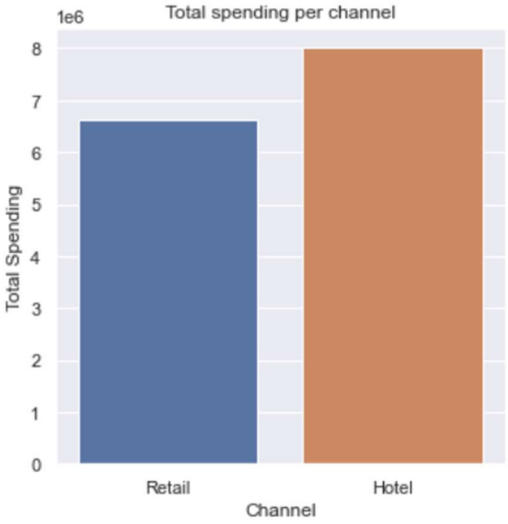
Detergents_Paper- This is a continuous variable in the dataset which consists of spending done on Detergents_Paper and has the following characteristics:-

- Mean = 2881.493182
- Standard deviation= 4767.854448
- Minimum = 3.0
- Maximum = 40827.0
- Range = Max - Min = 60844.0
- Q1= 256.75
- Q2= 816.5
- Q3= 3922.0
- Interquartile Range= Q3 - Q1 = 3665.25

Delicatessen- This is a continuous variable in the dataset which consists of spending done on Delicatessen and has the following characteristics:-

- Mean = 1524.870455
- Standard deviation= 2820.105937
- Minimum = 3.0
- Maximum = 47943.0
- Range = Max - Min = 47940.0
- Q1= 408.25
- Q2= 965.5
- Q3= 1820.25
- Interquartile Range= Q3 - Q1 = 1412.0



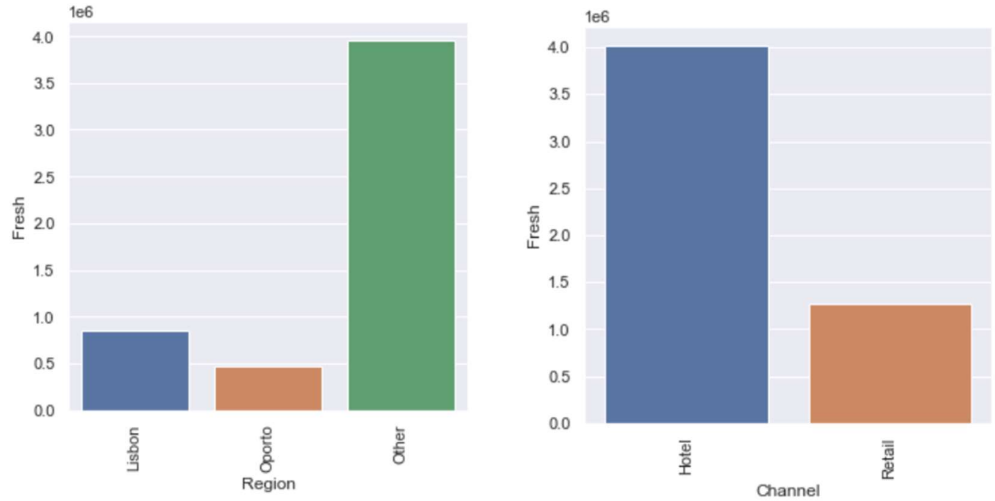


From the above plot charts, we can infer that amongst all the regions:-

1. Other region has the maximum spending and Oporto has the lowest spending.
2. And, amongst the two channels, Hotel has more spending than Retail.

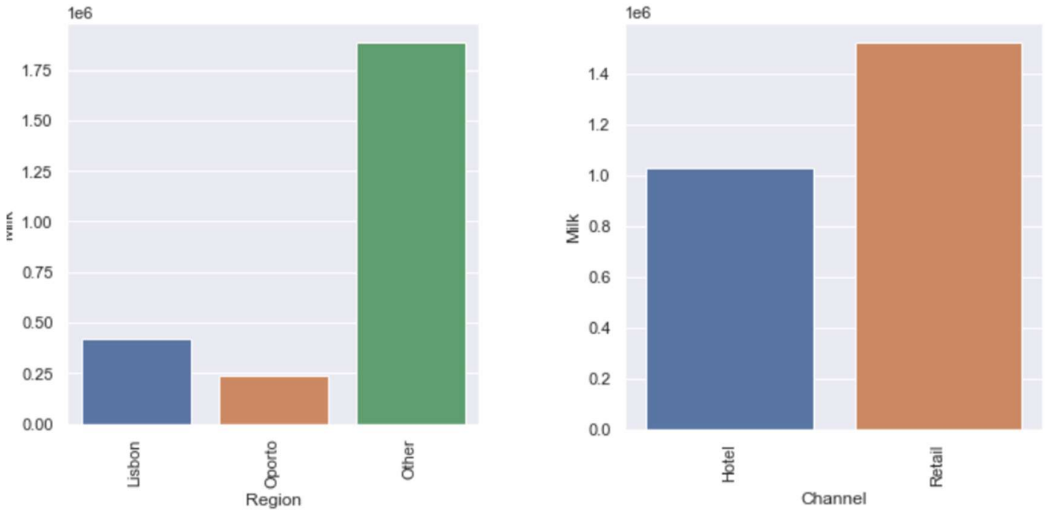
1.2 There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.

1. **Fresh product:-**



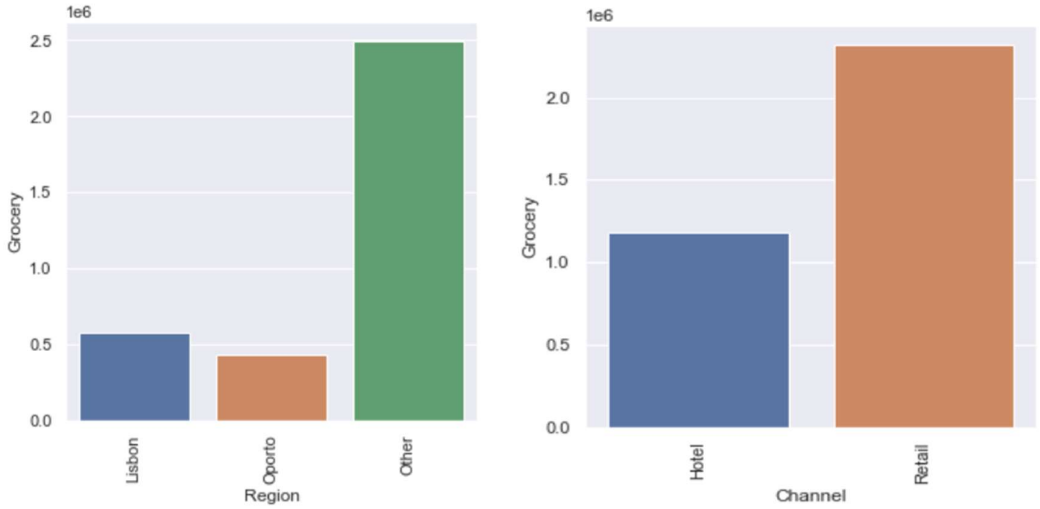
a. The spending on fresh product is the most in Other region and least in Oporto region.
b. The spending on fresh product is more in Hotel than in Retail.

2. Milk



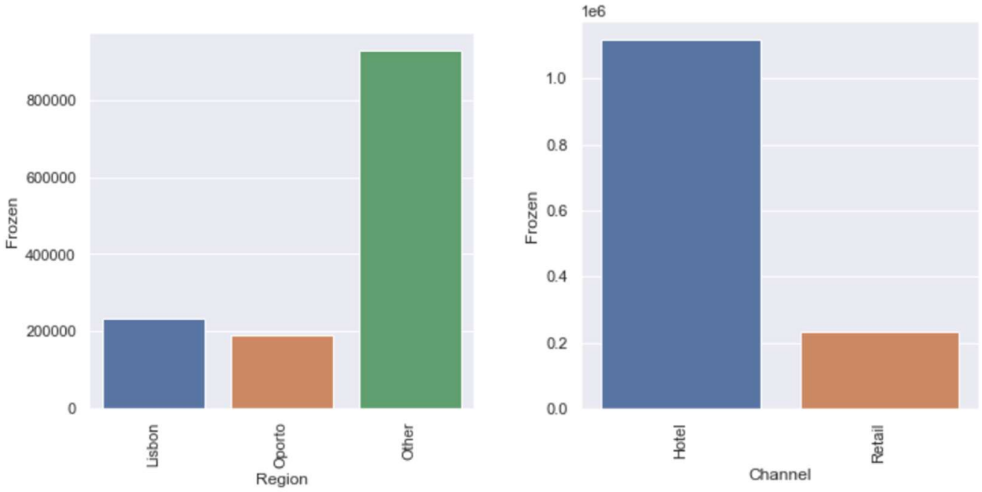
a. The spending on milk is the most in Other region and least in Oporto region.
b. The spending on milk product is more in Retail than in Hotel.

3. Grocery



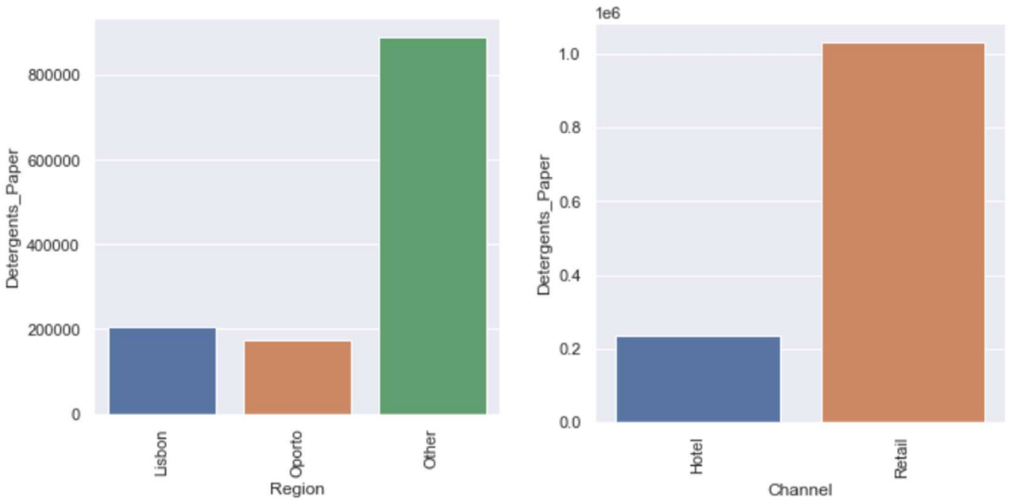
a. The spending on milk is the most in Other region and least in Oporto region.
b. The spending on milk product is more in Retail than in Hotel.

4. Frozen



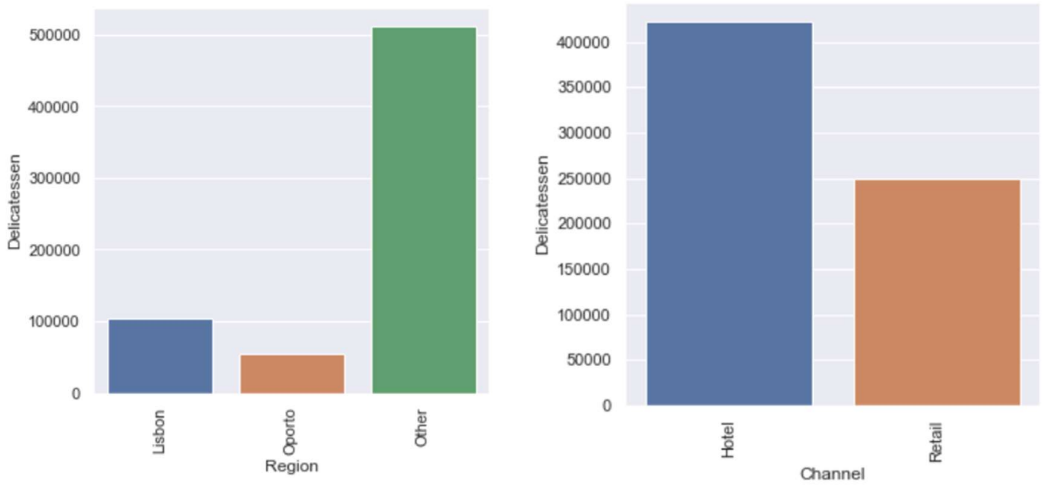
a. The spending on milk is the most in Other region and least in Oporto region.
b. The spending on milk product is more in Retail than in Hotel.

5. Detergents_Paper



a. The spending on Detergents_Paper items is the most in Other region and least in Oporto region.
b. The spending on Detergents_Paper items is more in Retail than in Hotel.

6. Delicatessen



a. The spending on delicatessen items is the most in Other region and least in Oporto region.
b. The spending on delicatessen items is more in Hotel than in Retail.

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total Spending
count	298.000000	298.000000	298.000000	298.000000	298.000000	298.000000	298.000000
mean	13475.560403	3451.724832	3962.137584	3748.251678	790.560403	1415.956376	26844.191275
std	13831.687502	4352.165571	3545.513391	5643.912500	1104.093673	3147.426922	22164.839073
min	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000	904.000000
25%	4070.250000	1164.500000	1703.750000	830.000000	183.250000	379.000000	13859.250000
50%	9581.500000	2157.000000	2684.000000	2057.500000	385.500000	821.000000	21254.500000
75%	18274.750000	4029.500000	5076.750000	4558.750000	899.500000	1548.000000	32113.750000
max	112151.000000	43950.000000	21042.000000	60869.000000	6907.000000	47943.000000	190169.000000

Spending on Products by Hotel Channel

From the above statistics, we can see that for Hotel channel the average spending is highest for Fresh product and the standard deviation is also the highest for Fresh. Detergents_Paper has the lowest spending.

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total Spending
count	142.000000	142.000000	142.000000	142.000000	142.000000	142.000000	142.000000
mean	8904.323944	10716.500000	16322.852113	1652.612676	7269.507042	1753.436620	46619.232394
std	8987.714750	9679.631351	12267.318094	1812.803662	6291.089697	1953.797047	29346.866491
min	18.000000	928.000000	2743.000000	33.000000	332.000000	3.000000	14993.000000
25%	2347.750000	5938.000000	9245.250000	534.250000	3683.500000	566.750000	30147.250000
50%	5993.500000	7812.000000	12390.000000	1081.000000	5614.500000	1350.000000	37139.000000
75%	12229.750000	12162.750000	20183.500000	2146.750000	8662.500000	2156.000000	51650.500000
max	44466.000000	73498.000000	92780.000000	11559.000000	40827.000000	16523.000000	199891.000000

Spending on Products by Retail Channel

From the above statistics, we can see that for Retail channel the average spending is highest for Grocery product and the standard deviation is also the highest for Grocery. Frozen item has the lowest spending by customers in Retail channel.

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total Spending
count	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000
mean	11101.727273	5486.415584	7403.077922	3000.337662	2651.116883	1354.896104	30997.571429
std	11557.438575	5704.856079	8496.287728	3092.143894	4208.462708	1345.423340	20321.813773
min	18.000000	258.000000	489.000000	61.000000	5.000000	7.000000	4925.000000
25%	2806.000000	1372.000000	2046.000000	950.000000	284.000000	548.000000	17184.000000
50%	7363.000000	3748.000000	3838.000000	1801.000000	737.000000	806.000000	25385.000000
75%	15218.000000	7503.000000	9490.000000	4324.000000	3593.000000	1775.000000	38699.000000
max	56083.000000	28326.000000	39694.000000	18711.000000	19410.000000	6854.000000	107155.000000

Spending on Products in Lisbon Region

From the above statistics, we can see that for Lisbon region the average spending is highest for Fresh product and the standard deviation is also the highest for Fresh. Delicatessen has the lowest spending.

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total Spending
count	47.000000	47.000000	47.000000	47.000000	47.000000	47.000000	47.000000
mean	9887.680851	5088.170213	9218.595745	4045.361702	3687.468085	1159.702128	33086.978723
std	8387.899211	5826.343145	10842.745314	9151.784954	6514.717668	1050.739841	24234.507325
min	3.000000	333.000000	1330.000000	131.000000	15.000000	51.000000	4129.000000
25%	2751.500000	1430.500000	2792.500000	811.500000	282.500000	540.500000	20611.500000
50%	8090.000000	2374.000000	6114.000000	1455.000000	811.000000	898.000000	26953.000000
75%	14925.500000	5772.500000	11758.500000	3272.000000	4324.500000	1538.500000	36158.500000
max	32717.000000	25071.000000	67298.000000	60869.000000	38102.000000	5609.000000	130877.000000

Spending on Products in Oporto Region

From the above statistics, we can see that for Oporto region the average spending is highest for Fresh product but maximum spending is done on Grocery product with the highest standard deviation. Delicatessen has the lowest spending.

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total Spending
count	316.000000	316.000000	316.000000	316.000000	316.000000	316.000000	316.000000
mean	12533.471519	5977.085443	7896.363924	2944.594937	2817.753165	1620.601266	33789.870253
std	13389.213115	7935.463443	9537.287778	4260.126243	4593.051613	3232.581660	27949.337752
min	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000	904.000000
25%	3350.750000	1634.000000	2141.500000	664.750000	251.250000	402.000000	17209.250000
50%	8752.500000	3684.500000	4732.000000	1498.000000	856.000000	994.000000	28029.000000
75%	17406.500000	7198.750000	10559.750000	3354.750000	3875.750000	1832.750000	42492.250000
max	112151.000000	73498.000000	92780.000000	36534.000000	40827.000000	47943.000000	199891.000000

Spending on Products in Other Region

From the above statistics, we can see that for Other region the average spending is highest for Fresh product and the standard deviation is also the highest for Fresh. Delicatessen has the lowest spending.

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?

```
Skewness of all the products

Fresh               2.552583
Milk                4.039922
Grocery             3.575187
Frozen              5.887826
Detergents_Paper    3.619458
Delicatessen       11.113534
dtype: float64
```

Skewness factor of all products are > 0. So the distribution of all items are positively Skewed. All the products are not normally distribute about the mean. So to check the inconsistent behaviour of the product , we will check the coefficient of variation as the average spending of all the products varies a lot and also there is a huge standard deviation in the spendings.

```
Fresh               1.054992
Milk                1.274186
Grocery             1.196121
Frozen              1.581045
Detergents_Paper    1.655327
Delicatessen        1.850010
Name: CV, dtype: float64
```

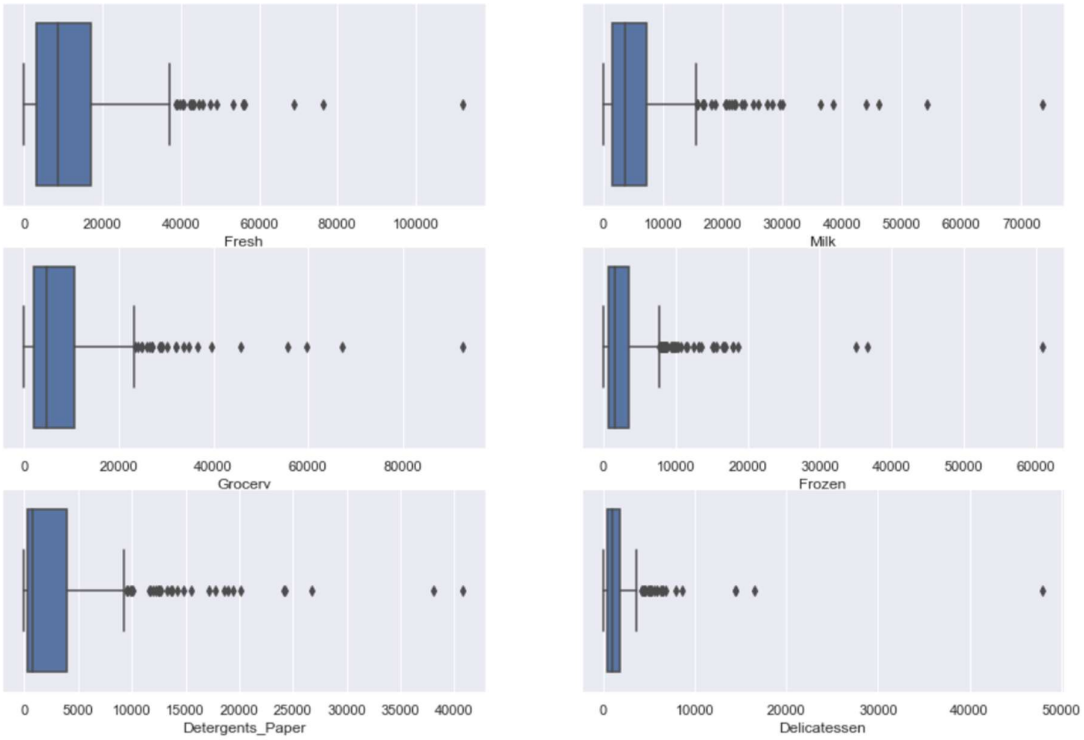
The **lowest Coefficient of variation is for Fresh**, so it is showing the least inconsistent behaviour. The **highest Coefficient of variation is for Delicatessen**, so it is showing the most inconsistent behaviour.

1.4 Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	441.000000	441.000000	441.000000	441.000000	441.000000	441.000000
mean	11973.088560	5783.125338	7933.249878	3064.969572	2874.962937	1521.416893
std	12645.864261	7377.148593	9499.903830	4851.357171	4764.407356	2817.832917
min	1.054992	1.274186	1.196121	1.581045	1.655327	1.850010
25%	3103.000000	1530.000000	2147.000000	737.000000	256.000000	406.000000
50%	8475.000000	3620.000000	4754.000000	1517.000000	813.000000	964.000000
75%	16933.000000	7184.000000	10646.000000	3549.000000	3909.000000	1819.000000
max	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000

Products description table

As we can see from the above table that there is a huge gap between the max spending on each product and Q3 . Same goes for the other side, i.e., there is a huge difference between min spending and Q1 of each product. So, we will plot a boxplot for each product to check if the particular product has outliers or not.



As per the above plots, outliers are present for all the products in the dataset due to which there is a huge standard deviation and the average of the spending is also not appropriate.

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.

1. Maximum spending is done in Other region and the least is done in Oporto . There is a need to increase the sales in Lisbon and Oporto region.
2. Hotel channel is spending more on the products than the Retail. So, the Customers should be encouraged to buy the products from these retail stores which can be influenced by introducing some discounts or offers.
3. Milk, Grocery and Detergents paper is showing more spending in Retail which means demand for these items are more in the retail stores.
4. Likewise, Fresh, frozen and delicatessen product has more spending in Hotel which shows that fancy items has more demand in Hotels than in Retails.
5. Lisbon region has a high demand of fresh products.
6. Oporto region has the highest demand of grocery items.
7. Other region has a high demand of fresh products.
8. Fresh and grocery products have very good sales, so some efforts should be done in order to increase the sales of the remaining products.
9. The dataset has outliers for each product which makes the data inconsistent. But for this analysis we have taken the outliers to be valid values.
10. The standard deviation for each product is very high which should be reduced in order to get a better picture.

Problem 2

Survey Data Analysis

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).

Survey Dataset Information

```
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   ID                62 non-null     int64
 1   Gender            62 non-null     object
 2   Age               62 non-null     int64
 3   Class             62 non-null     object
 4   Major             62 non-null     object
 5   Grad Intention    62 non-null     object
 6   GPA               62 non-null     float64
 7   Employment        62 non-null     object
 8   Salary            62 non-null     float64
 9   Social Networking 62 non-null     int64
 10  Satisfaction      62 non-null     int64
 11  Spending          62 non-null     int64
 12  Computer          62 non-null     object
 13  Text Messages     62 non-null     int64
dtypes: float64(2), int64(6), object(6)
```

There are a total of 14 columns in the dataset out of which 2 columns are of float type, 6 are of integer type and 6 are of object type. The dataset is clean as there are no null values in it.

Data description

1. ID- This is a continuous variable which has the Id number of every student.
2. Gender- This is a categorical variable with values as Male or Female .
3. Age- This is a continuous variable with age of very student.
4. Class- This is a categorical variable which denotes the class of every student.
5. Major- This is a categorical variable which has the name of the subject student is doing major in.
6. Grad Intention- This is a categorical variable which denotes if the student has decided to do the graduation or not.
7. GPA- GPA of the students.
8. Employment- A categorical variable to show what type of employment student has.
9. Salary- A continuous variable to show the salary of students.
10. Social Networking- A categorical variable
11. Satisfaction- A categorical variable
12. Spending- Continuous variable

13. Computer- Categorical variable for types of computer
14. Text Messages- Continuous variable

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

2.1.1. Gender and Major

Major Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	Total
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 Gender	No	Undecided	Yes	Total
Female	9	13	11	33
Male	3	9	17	29
Total	12	22	28	62

2.1.3. Gender and Employment

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

2.1.4. Gender and Computer

Computer Gender	Desktop	Laptop	Tablet	Total
Female	2	29	2	33
Male	3	26	0	29
Total	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?

Total students= 62
Total male students= 29
So, Probability that a randomly selected CMSU will be male= (Total male students)/(Total students)= 0.468

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

Total students= 62
Total female students= 33
So, Probability that a randomly selected CMSU will be female= (Total female students)/(Total students)= 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.

Major Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	Total
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

Total male students= 29
For Male:-
Total in accounting= 4
Total in CIS= 1
Total in Economics/Finance= 4
Total in International Business= 2
Total in Management= 6
Total in Other= 4
Total in Retailing/Marketing=5
Total in Undecided= 3
So, Probability of male students having major in Accounting = 4/29= 0.14
Probability of male students having major in CIS =1/29=0.03
Probability of male students having major in Economics/Finance= 4/29=0.14
Probability of male students having major in International Business=2/29=0.07
Probability of male students having major in Management= 6/29=0.21
Probability of male students having major in some other subject= 4/29=0.14
Probability of male students having major in Retailing/Marketing= 5/29=0.17
Probability of male students who have not decided about their major= 3/29= 0.1

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

Total female students= 33
For Female:-
Total in accounting= 3
Total in CIS= 3

Total in Economics/Finance= 7
Total in International Business= 4
Total in Management= 4
Total in Other= 3
Total in Retailing/Marketing=9
Total in Undecided= 0
So, the Probability of female students having major in Accounting=3/33= 0.09
Probability of female students having major in CIS= 3/33=0.09
Probability of female students having major in Economics/Finance= 7/33=0.21
Probability of female students having major in International Business =4/33=0.12
Probability of female students having major in Management =4/33=0.12
Probability of female students having major in some other subject =3/33= 0.09
Probability of female students having major in Retailing/Marketing= 9/33=0.27
Probability of female students who have not decided about their major=0/33=0.0

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.

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

Total male intending to graduate=17
Total students=62
Probability of male students who intends to Graduate is: 0.27

2.4.2 Find the probability that a randomly selected student is a female and does NOT have a laptop.

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

Total female with no laptop= 2+2=4
Total students=62
Probability that a randomly selected student is a female and does not have a laptop= 0.06

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 a male or has full-time employment?

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

Total male students=29
Total students=62
P(male)=29/62=0.467
Total full time employed=10
P(fullTimeEmp)= 10/62= 0.161
Total male full time employed= 7
P(male_fullTimeEmp)= 7/62= 0.112
p_maleOrFullTimeEmp=P(male)+P(fullTimeEmp)-P(male_FullTimeEmp)= 0.52
So, probability that a randomly chosen student is a male or has full-time employment= 0.52

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

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

Total female doing International Business major=4
Total female students=33
Total female doing Management major= 4
P(femaleIntBus)= 4/33= 0.121
P(femaleMgmt)= 4/33= 0.121
P(femaleIntBusOrMgmt)=P(femaleIntBus)+P(femaleMgmt) =0.24
So, the conditional probability that given a female student is randomly chosen, she is majoring in international business or management = 0.24

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 the graduate intention and being female are independent events?

Grad Intention	No	Yes	Total
Gender			
Female	9	11	20
Male	3	17	20
Total	12	28	40

For two events to be independent, p(A and B) should be equal to P(A)*P(B).
Total female students= 20

Total students=40
Total student with Grad intention=28
Total female with Grad intention=11
P (female and Grad Intention) = 11/40= 0.275
P(female)= 20/40=0.5
P (Grad intention) = 28/40=0.7
P(female) * P (Grad intention) =0.5 * 0.7 =0.35
So, we see that P (female and Grad Intention)! = P(female) * P (Grad intention)
Therefore, Graduate intention and being female are not independent events

2.7. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, 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 than 3?

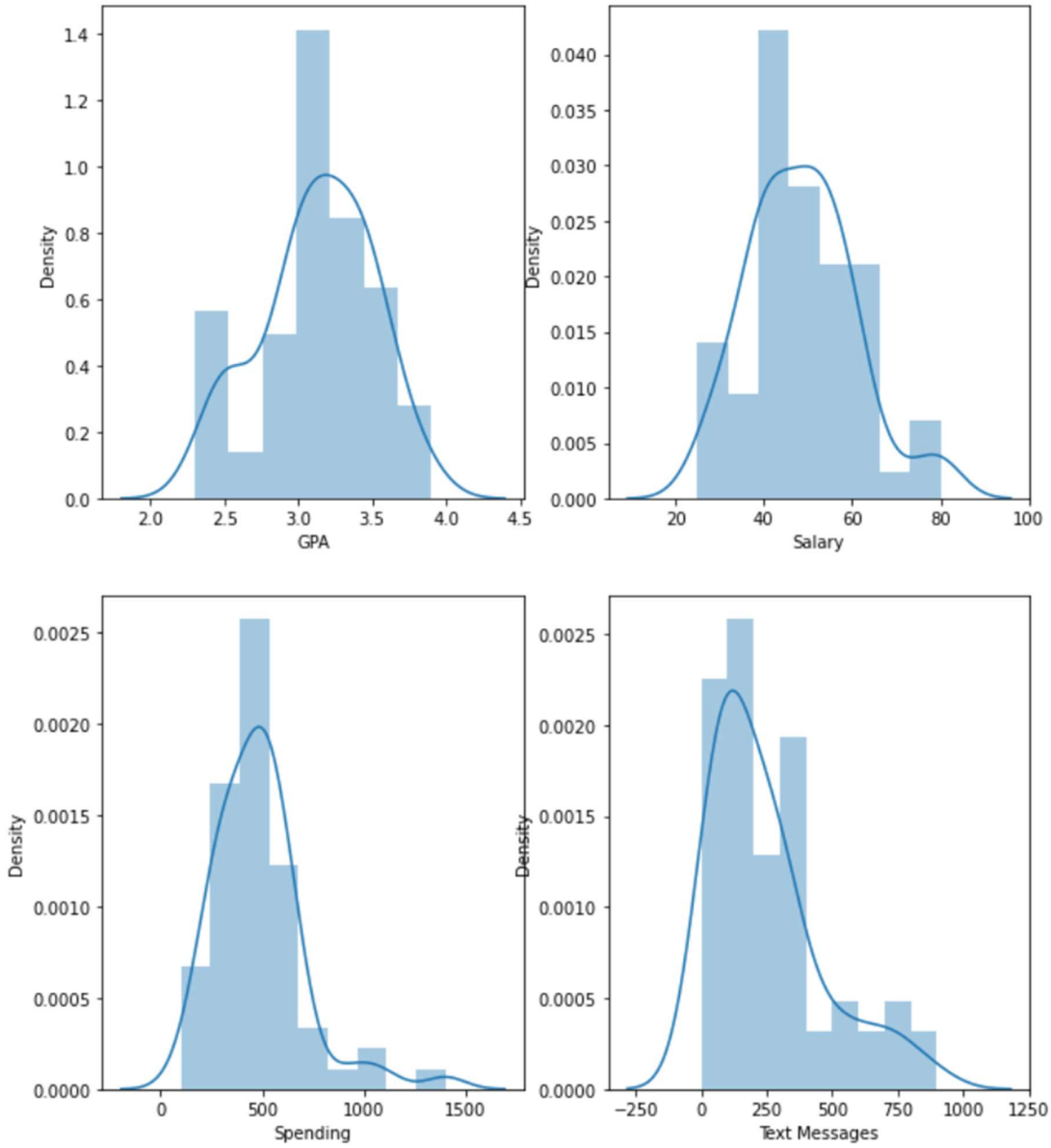
Total students=62
Total students with GPA less than 3= 17
P(GPA less than 3)=17/62= 0.274
So, the probability that GPA of a student is less than 3= 0.274

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

Total male students= 29
Total male students earning 50 or more =14
P(male earning more than 50)= 14/29= 0.482
So, the conditional probability that a randomly selected male earns 50 or more= 0.482
Total female students= 33
Total male students earning 50 or more =18
P(male earning more than 50)= 18/33= 0.545
So, the conditional probability that a randomly selected female earns 50 or more= 0.545

2.8. 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. Write a note summarizing your conclusions.

To check if the numerical continuous variables of the survey dataset follow a normal distribution or not, we will first see the plot , so that we can see the KDE of the particular variable



From the above plots, we can see that Salary and GPA are almost normally distributed but Spendings and Text messages seems to be somewhat rightwards skewed. To confirm we will now test if the mean, median and mode of these variables lie at the same place or not. If they lie at the same place, the variable should be normally distributed, else not.

GPA-

Mean of GPA: 3.129032258064516
Median of GPA: 3.1500000000000004
Mode of GPA: 0 3.0
1 3.1
2 3.4

Salary-
Mean of Salary: 48.54838709677419

Median of Salary: 50.0
Mode of Salary: 0 40.0

Spending-
Mean of Spending: 482.01612903225805
Median of Spending: 500.0
Mode of Spending: 0 500

Text Messages-
Mean of Text Messages: 246.20967741935485
Median of Text Messages: 200.0
Mode of Text Messages: 0 300

From the above measures of central tendency, we see here that mean, median and mode of GPA is very close, but for Salary, Spending and Text Messages we can't say the same.
To confirm the above inferences, we will now check the skewness factor and also do the Shapiro test which will give us a confirmation on the distribution type of these numeric variables.

Skewness factor-

Skewness factor of GPA is: -0.3146000894506981
Skewness factor of Salary is: 0.5347008436225946
Skewness factor of Spending is: 1.5859147414045331
Skewness factor of Text Messages is: 1.2958079731054333

Shapiro test-

Shapiro test p_value of GPA is 0.11204
Shapiro test p_value of Salary is 0.02800
Shapiro test p_value of Spending is 0.00002
Shapiro test p_value of Text Messages is 0.00000

From the above calculations for skewness factor and shapiro test, we see that skewness of only GPA lies in the range of -0.5 to 0.5 which proves it to be normally distributed. Also, the result of shapiro test shows us that only GPA has its p_value greater than the significance level, i.e, 0.05. So out of all the four numeric variables, only GPA is normally distributed and other variables are slightly skewed.

Problem 3

A & B shingles Data-

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 purchas ed a product lacking in quality if they find moisture and wet shingles inside the packaging. In some c ases, excessive moisture can cause the granules attached to the shingles for texture and colouring pu rposes to fall off the shingles resulting in appearance problems. To monitor the amount of moisture pr esent, 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 are calculated. The company would like to show that the mean moisture content is les s than 0.35 pounds per 100 square feet.

1.

Data Description-

2. A- 36 measurements (in pounds per 100 square feet) for A shingles.
3. B- 31 measurements (in pounds per 100 square feet) for A shingles

3.1 Do you think there is evidence that means moisture contents in both types of shingles are within the permissible limits? State your conclusions clearly showing all steps.

Step1: Defining the null and alternative hypothesis

H0: mean_moisture_content<= 0.35 pounds per 100 square feet (Null Hypothesis)

H1: mean_moisture_content> 0.35 pounds per 100 square feet (Alternative Hypothesis)

Step2: Defining the significance level(alpha)

alpha=0.05 (Here, we are taking the default value of alpha i.e., 5% significance level as it is not mentioned in the problem statement)

Step3: Selecting Test Statistics

We are testing hypothesis for two types of Shingles separately and as per the hypothesis statements and given data one sample T-Test will be used. It is a one tailed test (Left direction). Since, population standard deviation value is not known we are not considering normal distribution test.

Step4: Compute p-value and T-statistics

scipy.stats.ttest_1samp calculates the t test for the mean of one sample given the sample observations and the expected value in the null hypothesis. This function returns t statistic and the two-tailed p value.

Step5: Decide to reject or accept null hypothesis

```
One Sample T-Test Results For Type A Shingles
Test_Statistics_Type_A :-1.4735046253382782, P_Value_Type_A: 0.9252236685509249

One Sample T-Test Results For Type B Shingles
Test_Statistics_Type_B :-3.1003313069986995, P_Value_Type_B: 0.9979095225996808
```

Step6: Conclusion

Type A Shingles

Basis the hypothesis test performance for the given sample of 36 observations of Type A Shingles at 95% Confidence level we failed to reject the null hypothesis i.e. There are not enough evidence to prove that population mean moisture content of type A Shingle products is greater than 0.35 pounds per 100 square feet. So population mean moisture content of type A Shingle products is maintained at value less than or equals to 0.35 pounds per 100 square feet

Type B Shingles

Basis the hypothesis test performance for the given sample of 31 observations of Type B Shingles at 95% Confidence level we are in a position to reject the null hypothesis i.e. There are enough evidence to prove that population mean moisture content of Type B Shingle products is greater than 0.35 pounds per 100 square feet.

3.2 Do you think that the population mean 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?

We will be performing the two sample t-test.

Assumptions to perform the two sample T-test are:

1. The samples for Type A shingle and Type b shingle are independent of one another

2. Data for each type of shingle is obtained by random sampling from the population

3. The populations from which the data in each group is taken must is normally distributed

Step1: Defining the null and alternative hypothesis

H0: mu_A=mu_B i.e, population mean of type A shingle= population mean of Type b shingle (Null Hypothesis)

H1: mu_A!=mu_B i.e, population mean of type A shingle!= population mean of Type b shingle (Alternative Hypothesis)

Step2: Defining the significance level(alpha)

alpha=0.05 (Here, we are taking the default value of alpha i.e., 5% significance level as it is not mentioned in the problem statement)

Step 3 - Selecting Test Statistics

In This problem we have to compare the population means of Two Independent Shingles Type i.e., Type A and Type B. So we will make use of Two Independent Sample T-Test.

Step 4 Compute P value and T-Statistics

We use the scipy.stats.ttest_ind to calculate the t-test for the population means of TWO INDEPENDENT samples of types of Shingles 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 population means (expected) values. This test assumes that the populations have identical variances.

T_statistics: - 1.289628271966112

P_value: - 0.2017496571835328

Step5: Either accept or reject Null Hypothesis based on the value of p

alpha = 0.05

Level of Significance 0.05

We don't have enough evidences to reject null hypothesis since P Value is greater than 5%

Level of Significance

Step 6 Conclusion

Basis the hypothesis test performance for the given samples of Type A and Type B Shingles at 95% Confidence level we failed to reject the null hypothesis i.e. There are not enough evidence to prove that population mean moisture content of type A and type B are not equal. So population mean for sample data of type A and Type B Shingles are equal

-----END-----