SMDM PROJECT REPORT

Jyoti Ranjan Padhiary 1/30/22 PGDSBA

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 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. Which Region and which Channel spent the most? Which Region and which Channel spent the least?

	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

Figure 1 Sample of Data

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total_Spend
count	440.000000	440	440	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
unique	NaN	2	3	NaN	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	Hotel	Other	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	298	316	NaN	NaN	NaN	NaN	NaN	NaN	NaN
mean	220.500000	NaN	NaN	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455	33226.136364
std	127.161315	NaN	NaN	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937	26356.301730
min	1.000000	NaN	NaN	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000	904.000000
25%	110.750000	NaN	NaN	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000	17448.750000
50%	220.500000	NaN	NaN	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000	27492.000000
75%	330.250000	NaN	NaN	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000	41307.500000
max	440.000000	NaN	NaN	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000	199891.000000

Figure 2 Descriptive Statistics of the Data

Summary: We have the annual spending data of six different items in 440 retail stores across 3 regions and 2 channels.

The highest average spending among all verities is on "Fresh" items and lowest average spending is on "Delicatessen" items.

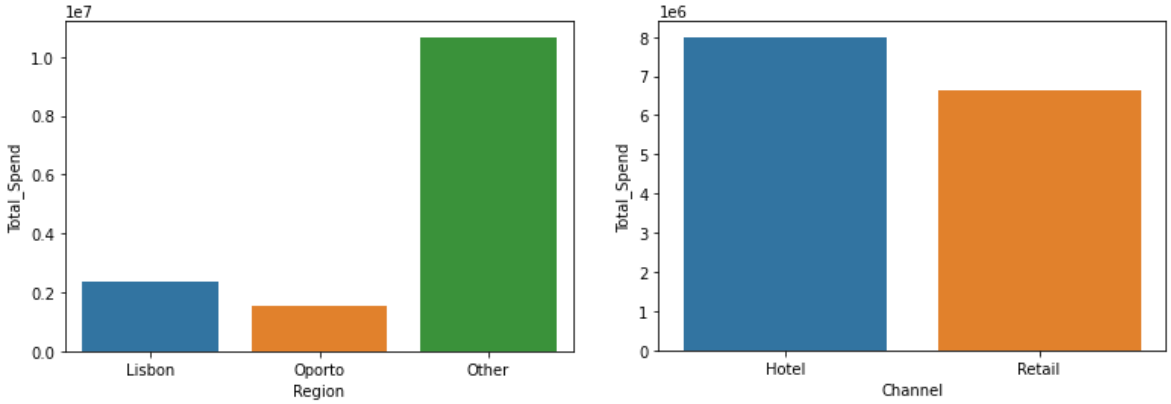

Figure 3 Total Spend across regions and channels

Answer: "Other" region and "Hotel" channel spent the most and "Oporto" region and "Retail" channel spent the least.

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.

Region	Channel	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Lisbon	Hotel	761233	228342	237542	184512	56081	70632
	Retail	93600	194112	332495	46514	148055	33695
Oporto	Hotel	326215	64519	123074	160861	13516	30965
	Retail	138506	174625	310200	29271	159795	23541
Other	Hotel	2928269	735753	820101	771606	165990	320358
	Retail	1032308	1153006	1675150	158886	724420	191752

Figure 4 Item wise total spending across region and channels

Answer: In all the regions maximum spending on Fresh items is in Hotels.

Hotels in Oporto region are spending least in Detergents paper.

Hotels in Lisbon are spending more on Milk items than the Retail stores in the region, as compared to other regions.

In all the items, hotels are spending more than retail stores in all three regions.

Spending in Milk items is more the hotels of Lisbon region but less as compared to retail stores in other and Oporto region.

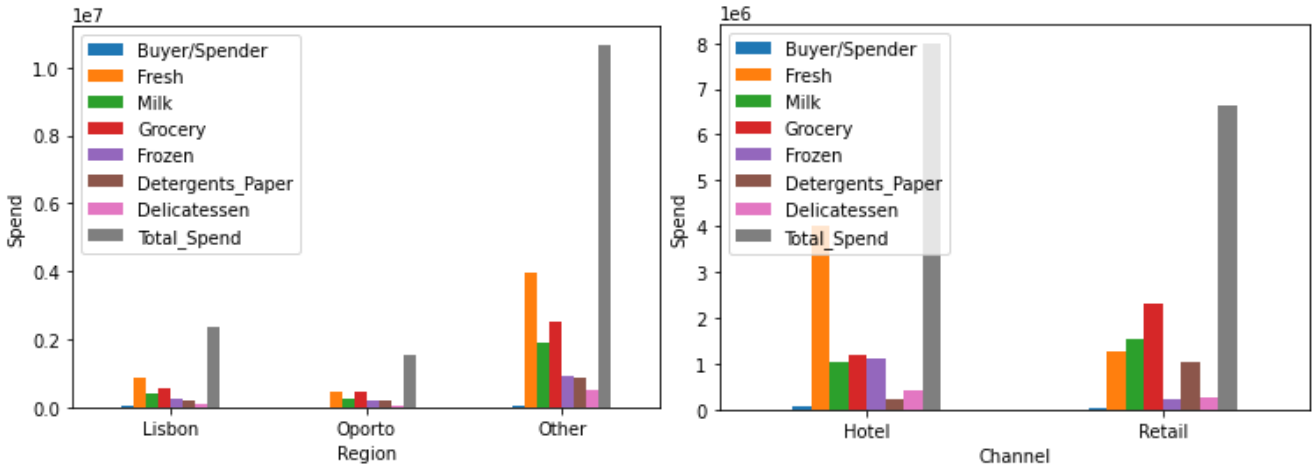

Figure 5:Item wise bar graph of spending in regions and channels

As per the above graph it can be commented that the maximum spending is on “Fresh” items in hotels.

Retail stores are spending more on “Grocery” items.

If we look at the total annual spending, then it is maximum in “Other” region and “Hotels”

“Frozen” items are seen very less spending in retail stores as compared to Hotels.

Retail stores are spending more on Milk, Grocery, Detergent powder items.

Other region has comparatively higher sending in all the items.

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?

Here we will use analysis of coefficient of variation to find out which item is most inconsistent and which is least inconsistent.

In fig 6 the coefficient of variation CV of all the items are plotted and we can see that Fresh items having lowest and Delicatessen having highest CV.

Answer: From the above analysis we can conclude that Fresh item is the least inconsistent as in the case of coefficient of variation its value is lowest, and Delicatessen is the most inconsistent with highest CV value.

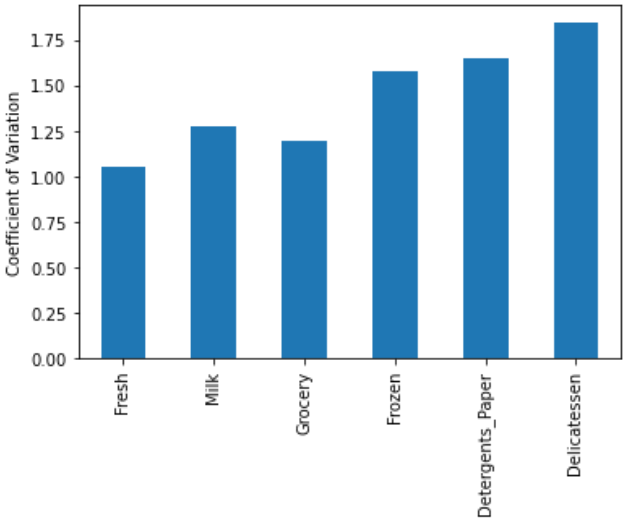

Figure 6: Coefficient of variation

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

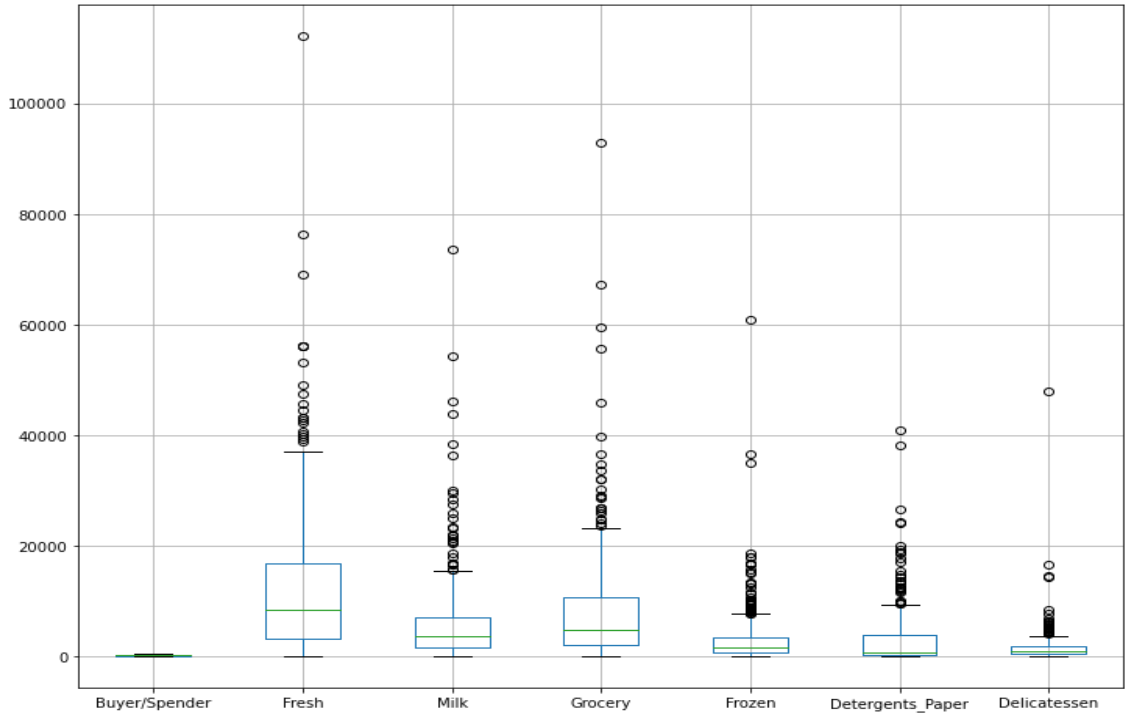

Figure 7: Box plot of all items

Answer: From the fig 7 it is clear that all the items are having outliers.

Fresh: by looking at the box plot we can see that the median is at the centre and both Q1 and Q2 are of equal distant so it represents a perfect symmetry and widely distributed data and is a normal distribution.

Milk: here the median is closer to Q1 and the minimum is also closer to Q1, hence the data is not symmetrically distributed but is right skewed. The maximum value is very far from the distribution.

Grocery: This plot shows that the median is very close to Q1 and minimum value, hence it's a right skewed distribution.

Frozen: Here the median is closer to Q3 so it represents a left skewed distribution.

Detergent Paper: As we can see that the median is almost equal to Q3, it represents a highly negative skewed distribution.

Delicatessen: The median is nearer to Q1 and the dataset is positively skewed.

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.

Observations:

From the above analysis it is observed that there is inconsistency in spending in all the items, which needs to be looked at.

If we look at the number of channels in each region, then the retail outlets are almost half of the hotel outlets and this explains the less spending in retail channels.

In the retail channels the spending on Detergent paper, milk and grocery has to be very high, but it is quite low.

In the hotels it is obvious that frozen items spending will be highest and its ok.

Region	Channel	Buyer/Spender
Lisbon	Hotel	59
	Retail	18
Oporto	Hotel	28
	Retail	19
Other	Hotel	211
	Retail	105

Figure 8: Number of channels across regions

Oporto region has the least number of channels which explains its least spending across all items. Fresh items are seen highest spending in hotels, but very wide gap as compared to retail channels.

Recommendations:

1. The number of channels should be increased in Lisbon and Oporto region. Especially in Lisbon region more retail outlets should be opened as there the gap is highest.

2. More fresh items should be supplied to retail stores as it is fast moving and can increase the spending.

3. The ratio of hotels to retail stores should be improved.

4. The stock of Fresh items can be maintained properly as the spending in this is least inconsistent and safe for business, but stock of delicatessen can be maintained as per order received or make to order.

Benefits:

If the number of channels increased in Lisbon and Oporto region, then the spending can increase in these two regions. Spending on fresh items is highest and least inconsistent so maintaining its stock will not be a loss for the company, but company has to be careful in maintaining stock of delicatessen items as spending on this is highly inconsistent, and company can avoid making wastage and increase in inventory. Similarly, the company has to take decisions on detergent paper and frozen items as these are seen high inconsistent spending too to avoid high inventory and loss.

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

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
All	10	43	9	62

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?

Total Number of students = 62

Total Number of male students = 29

P(Male) = 29/62 = 46.77%

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

Total Number of students = 62

Total Number of female students = 33

P(female) = 33/62 =53.22%

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
Male	0.137931	0.034483	0.137931	0.068966	0.206897	0.137931	0.172414	0.103448
All	0.137931	0.034483	0.137931	0.068966	0.206897	0.137931	0.172414	0.103448

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

Major Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing
Female	0.090909	0.090909	0.212121	0.121212	0.121212	0.090909	0.272727
All	0.090909	0.090909	0.212121	0.121212	0.121212	0.090909	0.272727

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

P (male and intend to graduate) = (17/29)× (29/62)= 0.274 or 27.4%

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

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

P (female and does not have laptop) = (4/33)× (33/62)= 0.064 or 6.45%

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	All
Gender				
Female	3	24	6	33
Male	7	19	3	29
All	10	43	9	62

P (male) = 29/62 = 0.46

P (full time employment) = 10/62 = 0.16

P (male and full time employment) = 7/62 = 0.11

P (male or full time employment) = (P (male) + P (full time employment))- P (male and full time employment) = (0.46+0.16)-0.11 = 0.51 or 51%

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

P (female student majoring in International Business) = 4/33 = 0.12

P (female student is majoring management) = 4/33 = 0.12

Major Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	All
Female	3	3	7	4	4	3	9	33
All	3	3	7	4	4	3	9	33

Both the events are mutually exclusive so,

P (female student majoring in international business or management) = P (female student majoring in international business) + P (female student majoring in management)

0.12 + 0.12 = 0.24 or 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 Gender	No	Yes
Female	9	11
Male	3	17

For two events A and B to be independent it should satisfy the below condition

$P(A \cap B) = P(A) \times P(B)$

P (Grad intention Yes) = 28/40 = 0.7

P(female) = 33/40 = 0.8

P (Grad intention Yes) × P(female) = 0.7 × 0.8 = 0.56

P (female and grad intention) = 11/40 =0.275

$0.56 \neq 0.275$

So the event of being female and grad intention are not dependent.

2.7 Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages, based on this answer below questions.

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

We have count the number of students who have less than 3 GPA using python which is =17

Total number of students = 62

P (random student will have <3 GPA) = 17/62 = 0.27 or 27%

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.

We have calculated the number of male / female earning more than 50 using python which is = 14 and 18 respectively

Total number of male = 29, female = 33

P (random male earns 50 or more) = 14/29 = 0.48 or 48%

P (random female earns 50 or more) = 18/33 = 0.54 or 54%

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. Write a note summarizing your conclusions.

We have done a shapiro test to check if the distribution is normal or not.

Here we have taken the hypothesis as follows:

H0: Distribution is normal

Ha: Distribution is not normal

If p value < 0.05 we reject the H0 and if p value > 0.05 then we fail to reject the H0

From fig 9 it can be observed that the distribution of GPA data points follows a bell curve. After performing a shapiro test we found p value = 0.11 which is > 0.05

Hence we fail to reject the H0 and it is a normal distribution curve.

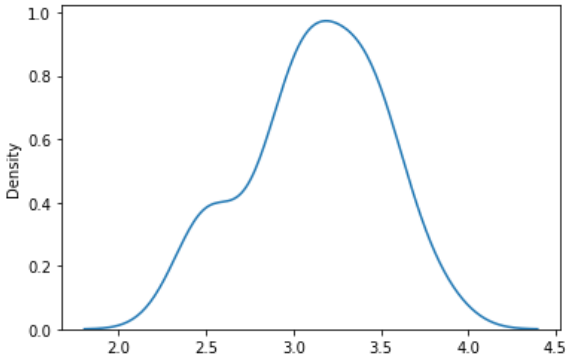

Figure 9: distribution of GPA

In fig 10 we have plot the distribution of salary, and performed a shapiro test where we found the

P value = 0.02 which is <0.05

Hence reject the H0 and it is not a normal distribution.

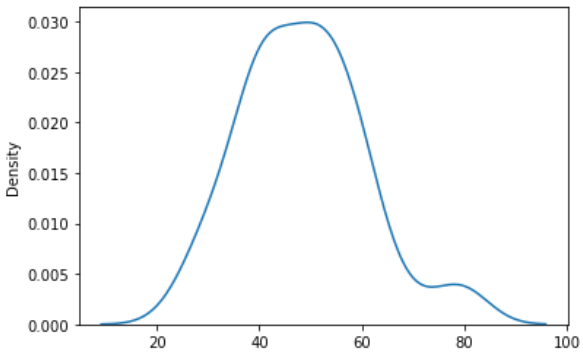

Figure 10: Distribution of salary

from fig 11 we can see that the distribution of spending is right skewed.

After performing a shapiro test we found the

P value = 1.6854e-05

Which is <0.05 hence we reject the H0

Hence it is not a normal distribution

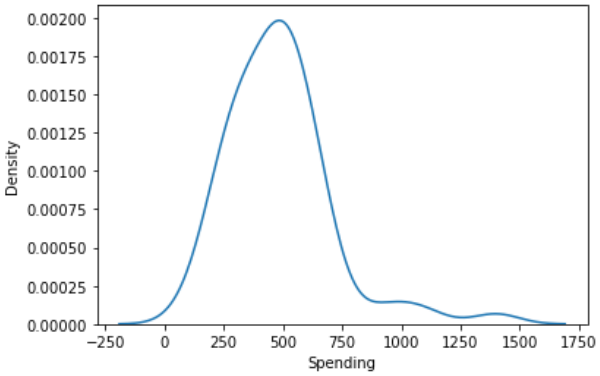

Figure 11: Distribution of spending

Here in fig 12 we have plot the distribution of text messages and it can be seen that it is slightly right skewed.

Here shapiro test gave p value = 4.320e-06

Which is < 0.05 hence we reject H0

Hence it is not a normal distribution.

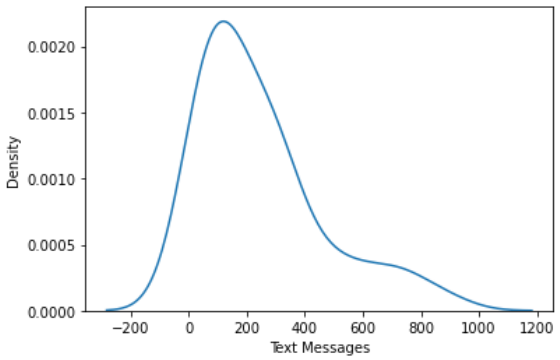

Figure 12: Distribution of text messages

2.8.2 Summary:

Distribution of GPA only is normal, which means most number of students score around the mean GPA. Distribution of Spending and text messages is not normal and it is right skewed. Distribution of salary is not normal. Then we have analysed the mean median and mode and below conclusions are made:

Most students earn a salary of 40 and spend 500

Most of the students have rated their satisfaction as 4

Most of the students are working as part time.

More than 30 students rated social networking as 1.

Most number of students have a laptop, very less students using a tablet.

Whereas max students intend to graduate, around 22 students have not decided yet.

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 are calculated. The company would like to show that the mean moisture content is less than 0.35 pounds per 100 square feet.

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

H0: Mean moisture of shingles A is =<0.35

Ha: Mean moisture of shingles A is > 0.35

We have done a one sample t test here using python.

The results are as below

T statistics: -1.47 and P value: 0.07

Here we have received p value > 0.05 hence we fail to reject the H0 and conclude that the mean moisture for shingles A is =< 0.35.

H0: Mean moisture of shingles B is =<0.35

Ha: Mean of moisture of shingles B is > 0.35

We have done a one sample t test using python.

The results are as below

T statistics: -3.1 and P value: 0.002

Here we have received a p value < 0.05 hence we reject the H0 and conclude that the mean moisture for shingles B is > 0.35

The moisture content in shingles A is within permissible limits but in shingles B it is more than permissible limit.

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?

H0: mean A = mean B

Ha : mean A ≠ mean B

Here we have done a two sample t test using python and the results are as below

T statistics= 1.29 and P value = 0.2

Here we have received a p value > 0.05 hence we fail to reject the H0 and conclude that the

mean A = mean B

Assumptions: Distribution of two populations are normal and the variance of the two population are same. If these assumptions are not likely to be met, then we have to perform another test.

The End