SMDM PROJECT BUSINESS REPORT

BY

RITUSRI MOHAN

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 ?

Descriptive Statistics of data including Channel & Retail:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Buyer/Spender	440	NaN	NaN	NaN	220.5	127.161	1	110.75	220.5	330.25	440
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	NaN	NaN	NaN	12000.3	12647.3	3	3127.75	8504	16933.8	112151
Milk	440	NaN	NaN	NaN	5796.27	7380.38	55	1533	3627	7190.25	73498
Grocery	440	NaN	NaN	NaN	7951.28	9503.16	3	2153	4755.5	10655.8	92780
Frozen	440	NaN	NaN	NaN	3071.93	4854.67	25	742.25	1526	3554.25	60869
Detergents_ Paper	440	NaN	NaN	NaN	2881.49	4767.85	3	256.75	816.5	3922	40827
Delicatessen	440	NaN	NaN	NaN	1524.87	2820.11	3	408.25	965.5	1820.25	47943

- Channel has two unique values, with Hotel as most frequent with 298 out of 440 transactions. i.e 67.7 percentage of spending comes from Hotel channel.
- Retail has three unique values, with Other as most frequent with 316 out of 440 transactions. i.e.71.8 percentage of spending comes from Other region.

The Highest & Lowest Spend

Region	Spending
Lisbon	2386813
Oporto	1555088
Other	10677599

The highest spend in Region is from Other which is $10677599 and lowest spend in the region is from Oporto which is $1555088.

Channel	Spending
Hotel	7999569
Retail	6619931

The highest spend in the Channel is from Hotel which is $7999569 and lowest spend in the Channel is from Retail which is $6619931

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.

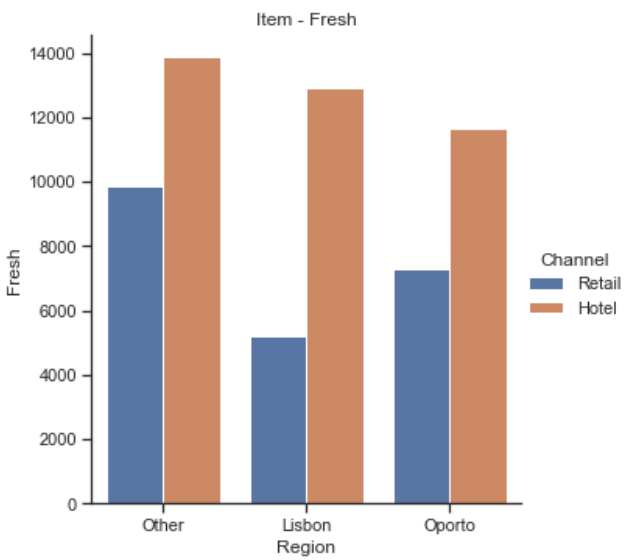


Fig.(a)

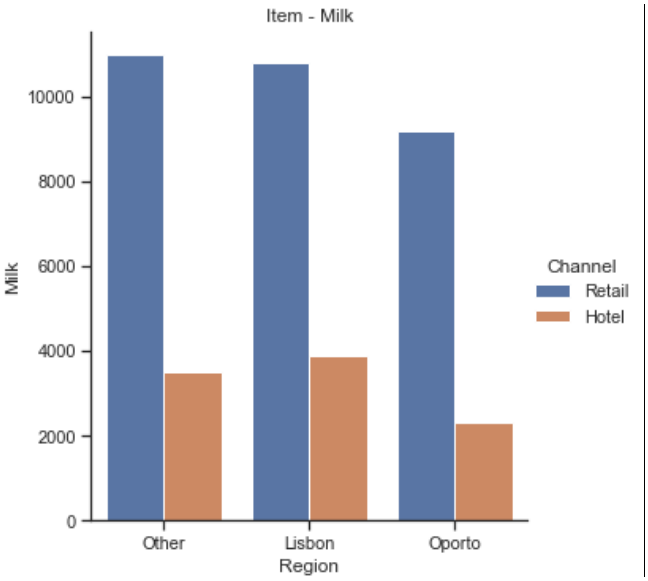


Fig.(b)

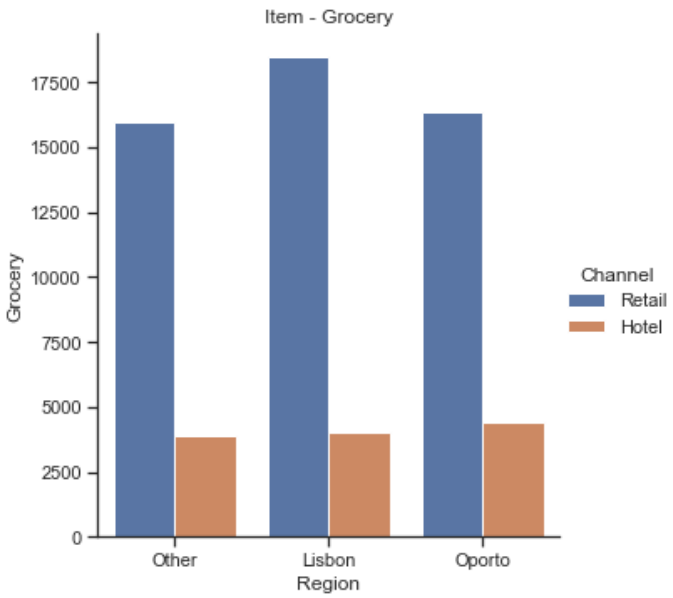


Fig.(c)

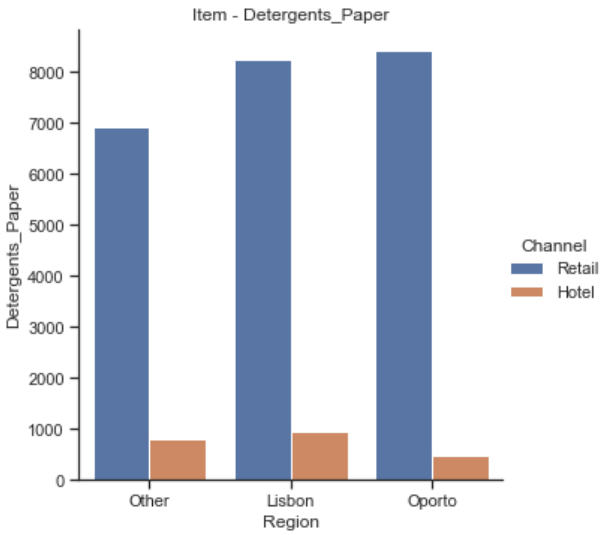


Fig.(d)

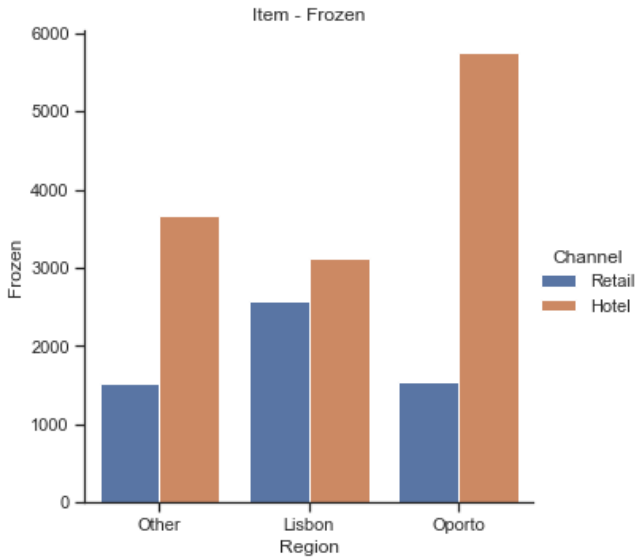


Fig.(e)

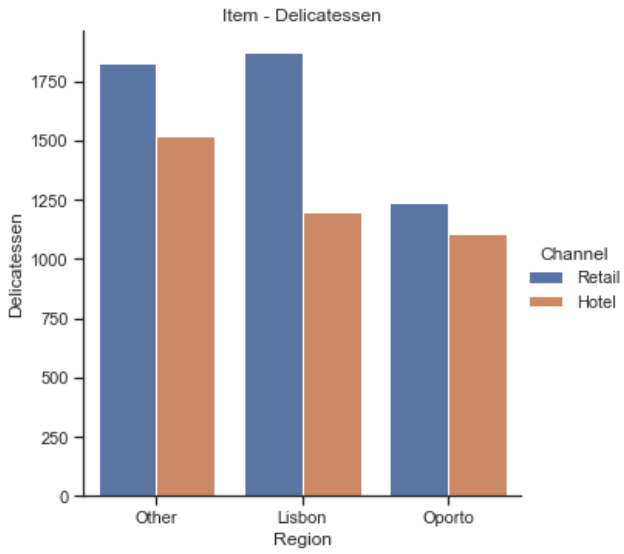


Fig.(f)

In the above plots-

- From Fig.(a) (i.e., item-fresh) and Fig.(e) (i.e., item-frozen), it is seen that, the spend for Channel Hotel is more than Retail across all the three Regions.
- From Fig.(b) (i.e., item-milk), Fig.(c) (i.e., item-grocery), Fig.(d) (i.e., item-detergents paper) and Fig.(f) (i.e., item-delicatessen), it is seen that, the spend for Channel Retail is more than Hotel across all the three Regions.

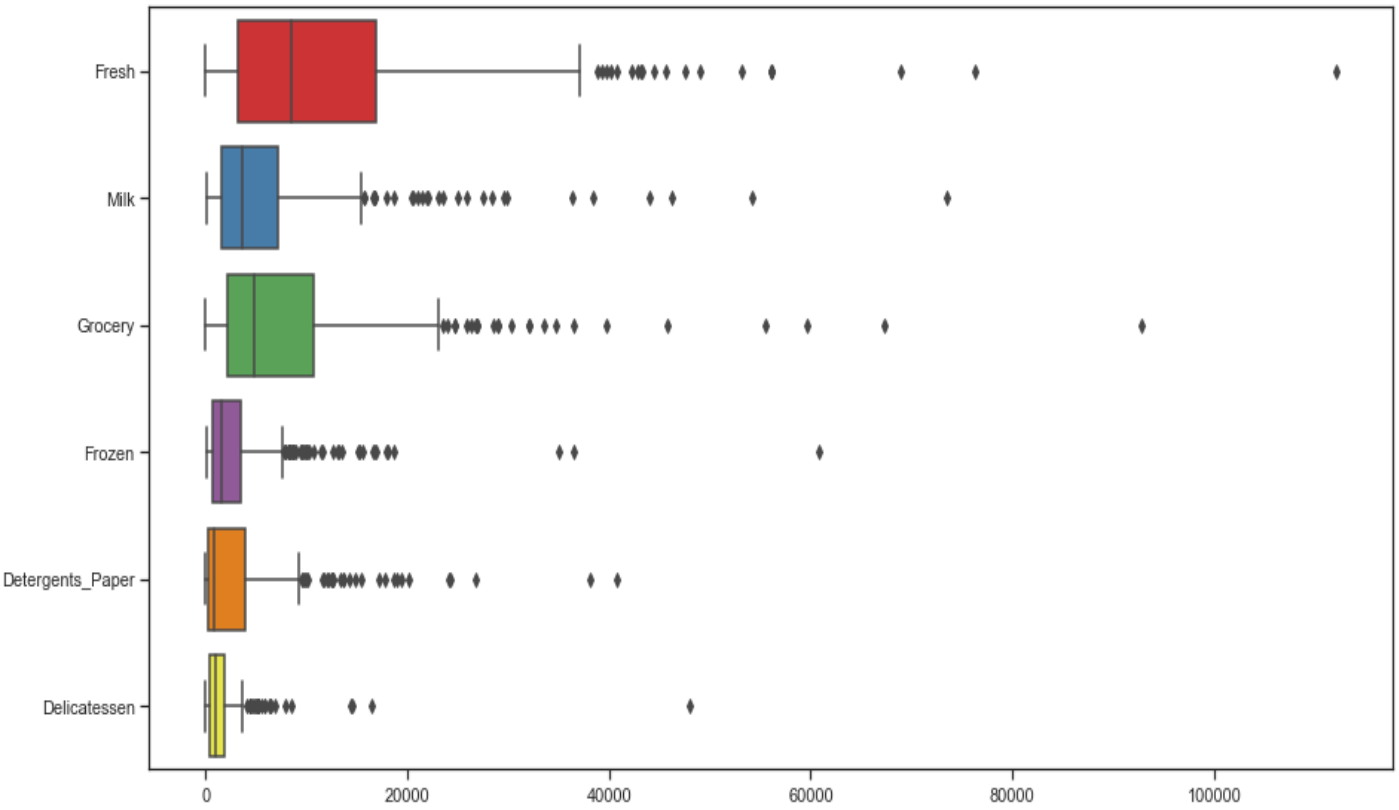
1.3 On the basis of a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?

Using Coefficient of Variation, we find out the item with least and most inconsistent behaviour. The Coefficient of Variation for the 6 items is as follows-

Fresh is 1.05
Milk is 1.27
Grocery is 1.19
Frozen is 1.57
Detergents Paper is 1.65
Delicatessen is 1.84

- From the results, it is clear that most inconsistent behavior shown by item – Delicatessen (highest value of coefficient of variation) and least inconsistent behavior shown by item – Fresh (least value of 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.



Yes, there are outliers in the given data.

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

- There are inconsistencies in spending of different items (by calculating Coefficient of Variation), which should be minimized.
- The spending of Hotel and Retail channel are different which should be more or less equal.
- Spend for different regions should be tried to made equal.

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	Accounting	CIS	Economics/ Finance	International Business	Management	Other	Retailing/ Marketing	Undecided
Gender								
Female	3	3	7	4	4	3	9	0
Male	4	1	4	2	6	4	5	3

2.1.2. Gender and Grad Intention

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
Gender			
Female	3	24	6
Male	7	19	3

2.1.4. Gender and Computer

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

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?

- Probability that a randomly selected candidate will be male: 46.774193548387096

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

- Probability that a randomly selected candidate will be female: 53.225806451612904

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.

Among MALE candidates:
Probability of being an Accounting student: 13.79%
Probability of being a CIS student: 3.44%
Probability of being an Economics/Finance student: 13.79%
Probability of being a student of International Business: 6.89%
Probability of being a Management student: 20.68%
Probability of being a student of Other major: 13.79%
Probability of being a Retailing/Marketing student: 17.24%
Probability of being a student of Undecided category: 10.34%

- From this output we can conclude that most of the male students prefer Management as Majors and CIS is least preferred.

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

Among FEMALE candidates:

Probability of being an Accounting student: 9.09%

Probability of being a CIS student: 9.09%

Probability of being an Economics/Finance student: 21.21%

Probability of being a student of International Business: 12.12%

Probability of being a Management student: 12.12%

Probability of being a student of Other major: 9.09%

Probability of being a Retailing/Marketing student: 27.27%

Probability of being a student of Undecided category: 0.0

- From this output we can easily say that most of the female students prefer Retailing/Marketing as Major.

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.

- Probability that a randomly chosen student is a male and intends to graduate: 58.620%

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

- Probability that a randomly chosen student is a female and does not have a laptop: 12.12%

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?

- Probability that a randomly chosen student is a male or has full-time employment: 51.61%

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

- Probability that given a female student is randomly chosen, she is majoring in international business or management: 24.24%

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
Gender		
Female	9	11
Male	3	17

- The Probability that a randomly selected student has an intention to graduate: 0.7
- The Probability that a randomly selected student has graduation intention and is female: 0.55.
- These probabilities are not equal. This suggests that the two events are independent.

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.

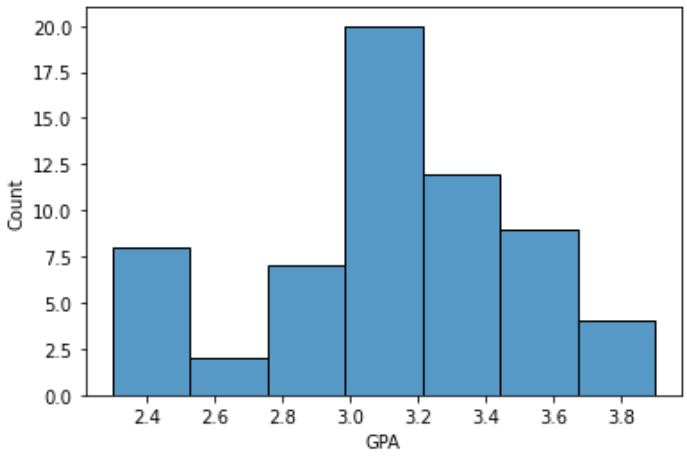
2.7.1. If a student is chosen randomly, what is the probability that his/her GPA is less than 3?

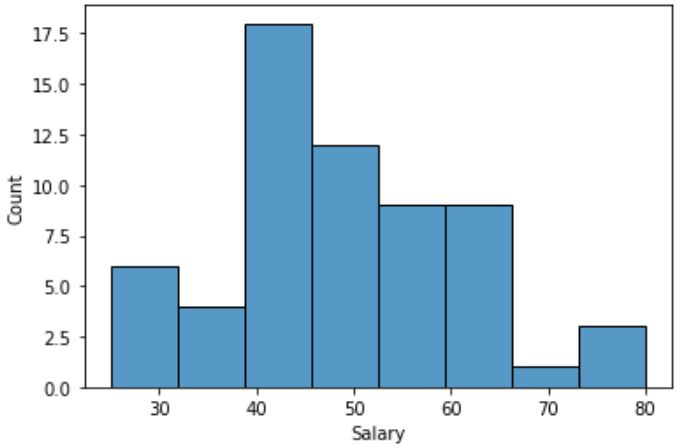
- Probability of choosing a student whose GPA is less than 3: 27.419%

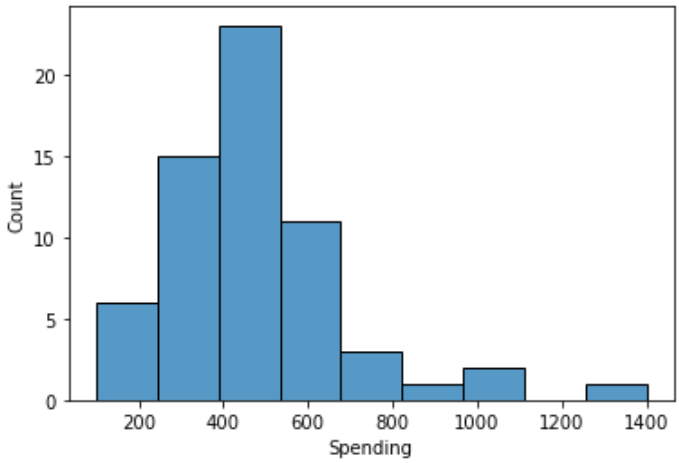
2.7.2. Find the conditional probability that a randomly selected male earns 50 or more. Find the conditional probability that a randomly selected female earns 50 or more.

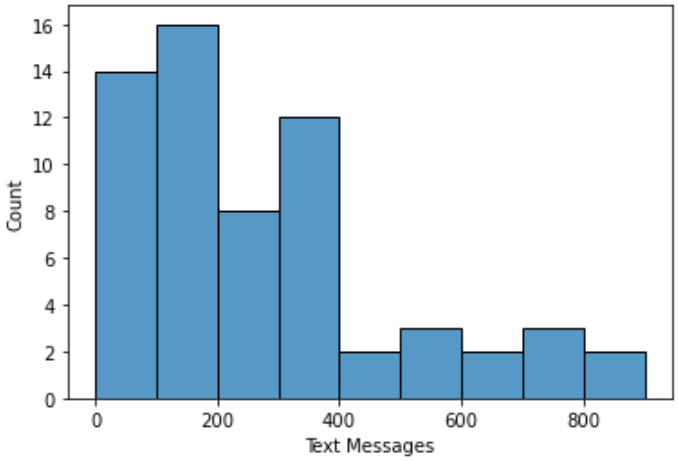
- Probability that a randomly selected male earns 50 or more: 34.48275862068966
- Probability that a randomly selected female earns 50 or more: 39.39393939393939

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.









- Shapiro test results for GPA-
  Statistics: 0.9685361981391907, p value: 0.11204058676958084

- Shapiro test results for Salary-
  Statistics: 0.9565856456756592, p value: 0.028000956401228905

- Shapiro test results for Spending-
  Statistics: 0.8777452111244202, p value: 1.6854661225806922e-05

- Shapiro test results for Text Messages-
  Statistics: 0.8594191074371338, p value: 4.324040673964191e-06

From the above results, we can conclude that ‘GPA’ and ‘Salary’ are following normal distribution whereas other two ‘Spending’ and ‘Text Messages’ are not following the normal distribution.

Problem 3

An important quality characteristic used by the manufacturers of ABC asphalt shingles is the amount of moisture the shingles contain when they are packaged. Customers may feel that they have purchased a product lacking in quality if they find moisture and wet shingles inside the packaging. In some cases, excessive moisture can cause the granules attached to the shingles for texture and coloring 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.

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.

A Shingles-

The null and alternative hypothesis to test whether the population mean moisture content is less than 0.35 pound per 100 square feet is given:

$$H_o: \mu = 0.35$$
$$H_a: \mu < 0.35$$

Since the test is conducted to check whether the population mean moisture content is less than 0.35, the alternate hypothesis is taken as: Ha < 0.35, and the null hypothesis is taken as: Ho = 0.35 because the hypothesis testing is done at a value.

The level of significance value, $\alpha = 0.05$.

The One-Sample t test was implemented. A python code was written in Jupyter Notebook to test the hypotheses.

Results-
t statistic: -1.4735046253382782
p value: 0.07477633144907513

- Since p value > 0.05, Ho cannot be rejected.
- There is not enough evidence to conclude that the mean moisture content for Sample A shingles is less than 0.35 pounds per 100 square feet. If the population mean moisture content is not less than 0.35 pounds per 100 square feet, the probability of observing a sample of 36 shingles that will result in a sample mean moisture content of 0.3167 pounds per 100 square feet or less, is 0.0748.

B Shingles

The null and alternative hypothesis to test whether the population mean moisture content is less than 0.35 pound per 100 square feet is given:

$$H_o: \mu = 0.35$$
$$H_a: \mu < 0.35$$

Since the test is conducted to check whether the population mean moisture content is less than 0.35,

the alternate hypothesis is taken as: Ha< 0.35, and the null hypothesis is taken as: Ho = 0.35 because the hypothesis testing is done at a value.

The level of significance value, α=0.05.

The One-Sample t test was implemented. A python code was written in Jupyter Notebook to test the hypotheses.

Results-
t statistic: -3.1003313069986995
p value: 0.0020904774003191826

- Since p value < 0.05, we reject Ho.
- There is enough evidence to conclude that the mean moisture content for Sample B shingles is not less than 0.35 pounds per 100 square feet. If the population mean moisture content is in fact no less than 0.35pounds per 100 square feet, the probability of observing a sample of 31 shingles that will result in a sample mean moisture content of 0.2735 pounds per 100 square feet or less is 0.0021.

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?

The null and alternative hypothesis to test whether the population mean moisture content of shingles A and B are equal, the following null hypothesis and alternative hypothesis were formed.

$$Ho: \mu(A) = \mu(B)$$
$$Ha: \mu(A) ! = \mu(B)$$

The level of significance value, α=0.05.

The Two-Sample t test was implemented. A python code was written in Jupyter Notebook to test the hypotheses.

Results-
t statistic: 1.2896
p value: 0.2017

- Since p value > 0.05, Ho cannot be rejected.
- It is possible that population mean for shingles A and B are equal. The test assumptions while doing the Two Sample t test are that – the distributions of both the populations (A & B shingles) are normal, and the variances are also same.