DATA SCIENCE AND BUSINESS ANALYSIS

Project 2

Piyush Kumar Singh
PGP – DSBA Online
May-21 Batch

Date: 05/07/2021

Table of Contents

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

Let's describe the data to find out column names and there various types, numbers of rows and columns, unique values of each column and missing values in the dataframe.

Data Description:

There are total 440 rows and 9 columns. Out of 9 columns only 2(Channel and Region) are categorical type and rest are continuous type. There are no missing values in any of the columns. Units of measurement for continuous data type is not mentioned.

#	Columns	Non Null Value Count	Data D-Type	Unique Values in Columns
0	Buyer/Spender	440	int64	Continuous Var
1	Channel	440	object	Hotel,Retail
2	Region	440	object	Other,Lisbon,Oporto
3	Fresh	440	int64	Continuous Var
4	Milk	440	int64	Continuous Var
5	Grocery	440	int64	Continuous Var
6	Frozen	440	int64	Continuous Var
7	Detergents_Paper	440	int64	Continuous Var
8	Delicatessen	440	int64	Continuous Var

Now, we create a dataframe using groupby of "Region" column and find out the sum of annual spending of all 6 varieties of products at that region.

- We can clearly see that Others spend the max amount of 10677599 and while Oporto spend the least amount of 1555088.

Again we can create a dataframe using groupby of "Channel" column and find out the sum of annual spending of all 6 varieties of products at that channel.

- We can clearly see that Hotels spend the max amount of 7999569 and while Retail spend th e least amount of 6619931.

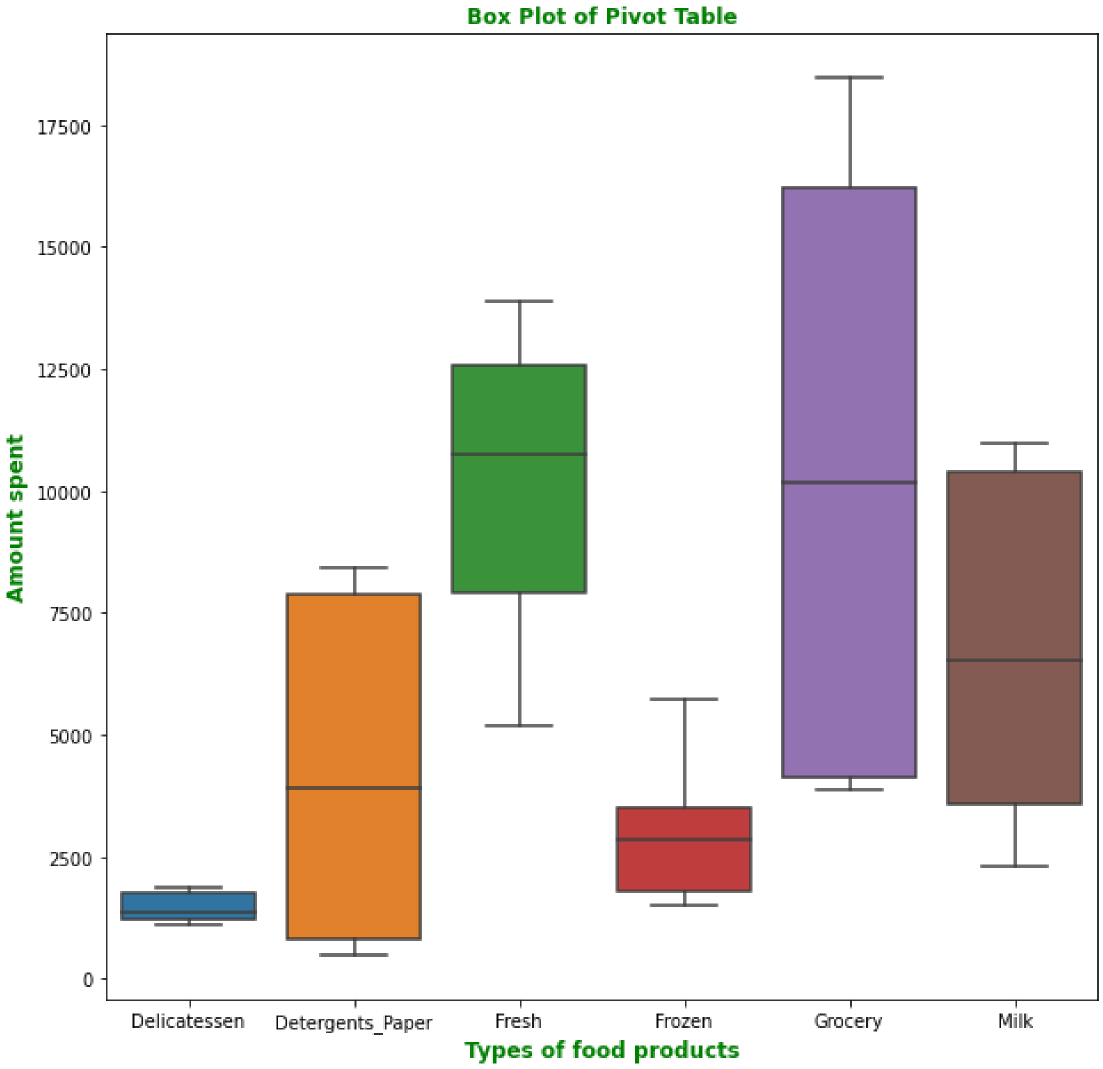
Please Note: that in both the calculation I have not include "Buyer/Spender" Column as they represent different 440 distributes index. Inclusion of this column will not give correct result of amount spend as its values will then also be then included in the calculations. Values in this columns is a representation of 440 number of distributors and not amount spend. So I have dropped the column before I performed the calculation for the question.

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.

We construct a pivot table using index as (Region and Channel) and compare it with the 6 types of products in dataset. By analysing the table we can see that amount spend on **Detergents_Paper, Milk, Grocery** is much higher in Retail channel than in Hotel across all region.

Whereas amount spend on **Fresh and Frozen** is higher in Hotel channel. **Delicatessen** has the least amount spend across all Region as well as Channel.

We can also visualise the amount spend in both the channels across all the region in total using a boxplot.



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?

I have used lambda function to find the Coefficient of Variation of each food category. Higher value of Coefficient of Variation means higher inconsistency and lower Coefficient of Variation means less inconsistency. We can clearly see that Fresh has the least value of 1.053918 and Delicatessen has the maximum value of 1.849407.

So, Most consistent behaviour is of Fresh.

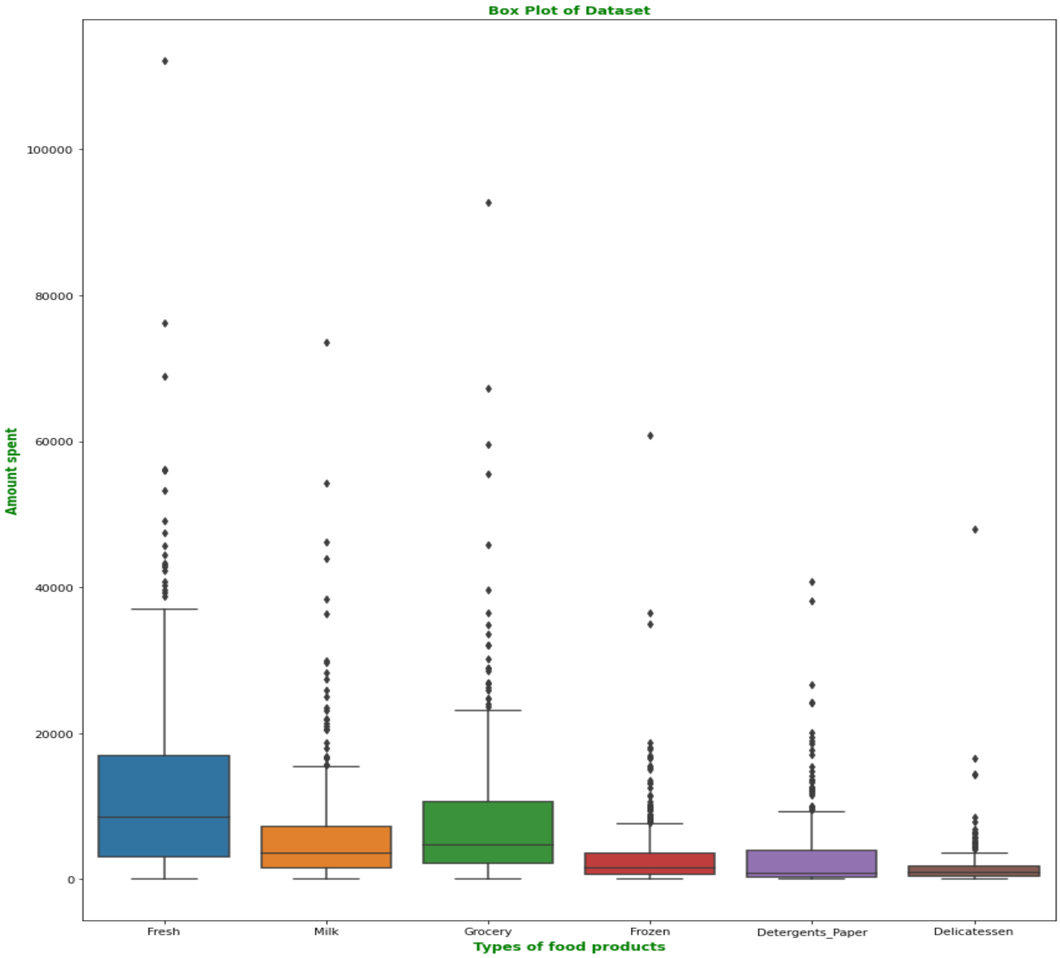
Least consistent behaviour is of Delicatessen.

Coefficient of Variation of each category:

Fresh	1.053918
Milk	1.273299
Grocery	1.195174
Frozen	1.580332
Detergents_Paper	1.654647
Delicatessen	1.849407

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

To find out outliers we use box plot. Here also we have dropped the Buyer/Spender column from the dataset we have visualised as they represents distributers. We can clearly see that all the category of food products have outliers present in them.



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

- On the basis of data some products are more sold on Retail and some are more sold on Hotel. There is significant difference between amounts spend between Retail and Hotel. Suggestion is reduce this difference between both channels.
- If we compare region wise then also others has max amount spend whereas there is significant difference between amounts spend between Lisbon and Oporto.
- On basis of coefficient of variance of all the food category there inconsistency in all the products. We need to implement changes so that the consistency in amount spent in all the various food products can be minimised as much as possible.
- Amount spend on Delicatessen and Frozen is very low across all the Region and Channels. Need changes to improve this.

Problem Statement:

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

To make a contingency table I have used crosstab between Gender and Major columns in dataset and used gender first to set as row variable.

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

Similarly, we can do the same for rest crosstab table's comparisons.

2.1.2. Gender and Grad Intention

Gender	No	Undecided	Yes
Female	9	13	11
Male	3	9	17

2.1.3. Gender and Employment

Gender	Full-Time	Part-Time	Unemployed
Female	3	24	6
Male	7	19	3

2.1.4. Gender and Computer

Gender	Desktop	Laptop	Tablet
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?

As per dataset Total Numbers of Female = 33

Total Numbers of Male = 29

Total Students = (33+29) = 62

Therefore, probability of random selected CMSU is a male is = 29/62 = 0.46 or 46.77 %

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

As per dataset Total Numbers of Female = 33

Total Numbers of Male = 29

Total Students = (33+29) = 62

Therefore, probability of random selected CMSU is a female is = 33/62 = 0.53 or 53.23 %

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.

There are 29 male students out of 62 total students in CMSU. We construct a pivot table between Major's and Gender to find out number of male and female in each major.

Here we find conditional probability of different majors given male.

P(Retailing/Marketing|Male) = 5/29 = 0.1724
There are 5 male students in Retailing/Marketing

P(Economics/Finance|Male) = 4/29 = 0.1379
There are 4 male students in Economics/Finance

P(Management|Male) = 6/29 = 0.2069
There are 6 male students in Management

P(Accounting|Male) = 4/29 = 0.1379
There are 4 male students in Accounting

P(Other|Male) = 4/29 = 0.1379
There are 4 male students in other

P(International Business|Male) = 2/29 = 0.069
There are 2 male students in International Business

P(CIS|Male) = 1/29 = 0.0345
There are 1 male students in CIS

P(Undecided|Male) = 3/29 =0.1034
There are 3 male students who are Undecided

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

There are 33 female students out of 62 total students in CMSU. We construct a pivot table between Major's and Gender to find out number of male and female in each major

Here we find conditional probability of different majors given female.

P(Retailing/Marketing|Female) = 9/33 = 0.2727
There are 9 female students in Retailing/Marketing

P(Economics/Finance|Female) = 7/33 = 0.2121
There are 7 female students in Economics/Finance

P(Management| Female) = 4/33 = 0.1212
There are 4 female students in Management

P(Accounting| Female) = 3/33 = 0.0909
There are 3 female students in Accounting

P(Other| Female) = 3/33 = 0.0909
There are 3 female students in other

P(International Business| Female) = 4/33 = 0.1212
There are 4 female students in International Business

P(CIS| Female) = 3/33 = 0.0909
There are 3 female students in CIS

P(Undecided| Female) = 0/33 = 0.00
There are 0 female students who are Undecided

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.

Details of students with intention to graduate:

Yes	-	28
Undecided	-	22
No	-	12

Probability of student with intend to graduate is 28/62 = 0.4516

Probability of male student is 29/62 = 0.4677

So, Probability of random selected student is a male and intends to graduate will be

= 0.4516*0.4677 = 0.2115 or 21.15%

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

Details of students with different computer type:

Laptop	-	55
Desktop	-	5
Tablet	-	2

Probability of student not having laptop = 5+2/62 = 0.1129

Probability of female student is 33/62 = 0.5322

So, probability that a randomly selected student is a female and does NOT have a laptop will be

= 0.1129*0.5322 = 0.0583 or 5.83%

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?

Details of students with employment:

Part-Time	-	43
Fulltime	-	10
Unemployed	-	9

Probability of student having full-time employment is 10/62 = 0.1613

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

= P(Male) + P(Full-Time Employment) – P(Male & Full-Time employment)

= 0.4677 + 0.1613 – (0.4677*0.1613) = 0.5548 or 55.48%

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

There are 8 total female doing Management and International Business, so the result is as follows:

Result = Probability of Female doing both subjects / Total Number of Females

Result = 8/33 = .2424 or 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?

Here we construct a crosstab table between Gender and Intent to Graduate and drop the Undecided column to make it a 2x2 table.

Gender	No	Yes
Female	9	11
Male	3	17

Yes being female and graduate intention are both independent events as intention to graduate has no relation with gender of a student in making a decision.

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?

First we find out how many students scored less than 3 GPA using pandas conditions out of total 62 students in college.

We get 17 students in total scored less than 3 GPA.

So, Probability that a random chosen student has less than 3 GPA = 17/62 = 0.2742

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.

First we found out how many students earn 50 or more using pandas conditions out of all the 62 students in college. There are 33 female and 29 male in total in college.

We get 32 students in total scored 50 or more. Out of which there are:

Males = 14

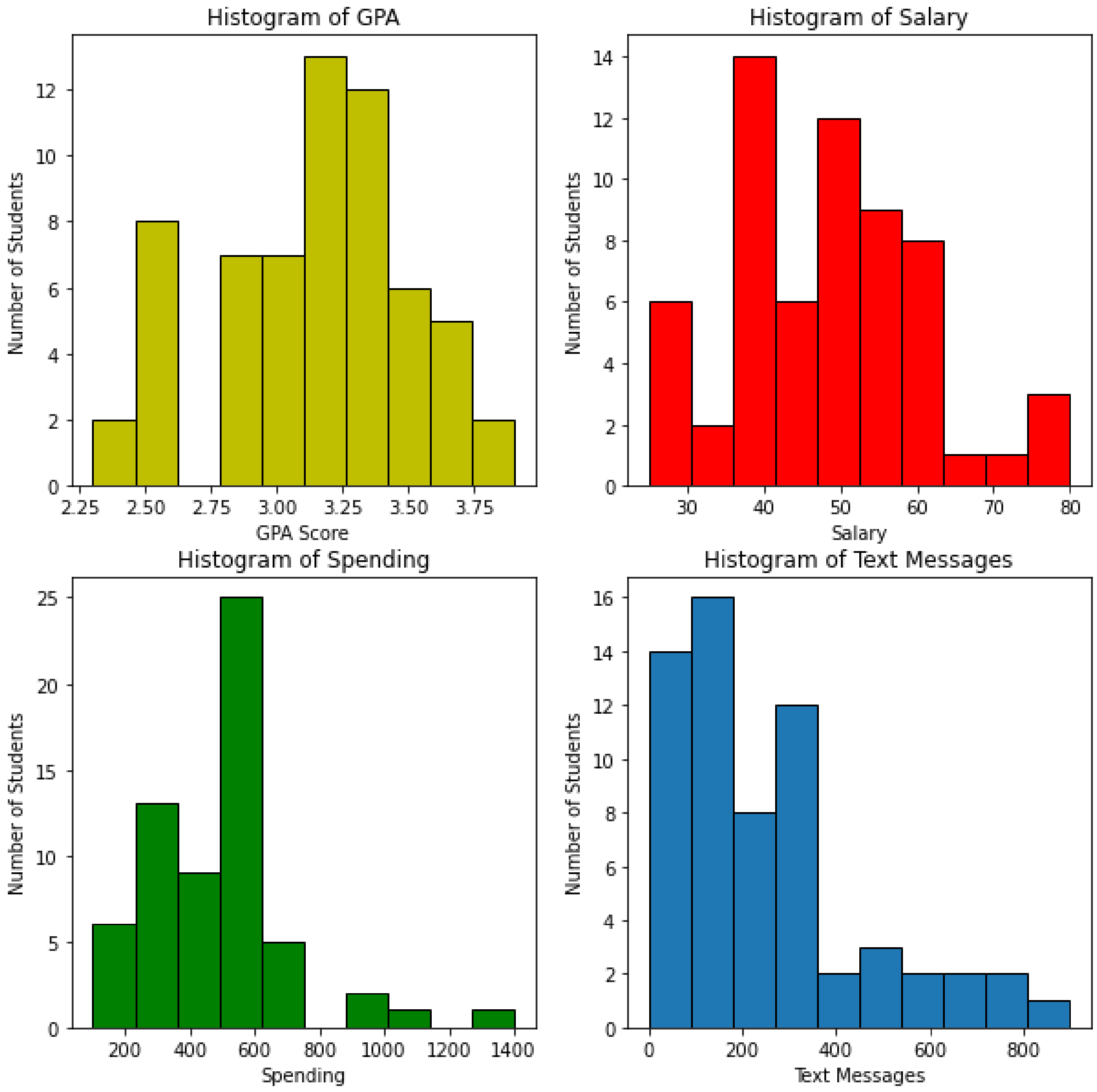
Females = 18

So, Conditional probability that a randomly selected male earns 50 or more is = 14/29 = 0.4828 or 48.28%

Conditional probability that a randomly selected female earns 50 or more is = 18/33 = 0.5455 or 54.55%

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.

Here we make histogram to check for the distribution of all the continuous variable mentioned in question. We have used histogram plot for all the variable separately and plotted all the 4 graphs next to each other using subplot method. Judging from the graphs we can tell the graph of **GPA Score and Salary** follow a normal distribution whereas graph of **Spending and Text Messages** is both right skewed distribution.



Problem Statement:

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.

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

Based on the question we can form our H0 and H1 hypothesis.

H0: Mean moisture content = 0.35 pounds per 100 square feet

H1: Mean moisture content < 0.35 pounds per 100 square feet

Now, this is a one tailed test as per our H1 formulation on negative side. Since nothing is given we take alpha to be 5% or 0.05.

We can perform 1 Sample T-Test for this problem for both A and B Shingles:

For Shingle A –

We get the p-value as 0.0748 and t statistic as -1.4735

Since p-value more than alpha we fail the reject H0 or accept H0 hypothesis.
Therefore no evidence to conclude that mean moisture content for Shingle A is less than 0.35 pounds per 100 square feet.

For Shingle B –

We get the p-value as 0.0021 and t statistic as -3.1003

Since p-value less than alpha we the reject H0 hypothesis.
Therefore there is enough evidence that we can to conclude that mean moisture content for Shingle B is less 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?

Based on the question we can form our H0 and H1 hypothesis.

H0: Mean Shingle A = Mean Shingle B

H1: Mean Shingle A != Mean Shingle B

Now, this is a 2-Tailed T test as per H1 formulation. Since nothing is given we take alpha to be 5% or 0.05.

We can perform 2 Sample T-Test for this problem for both A and B Shingles:

We get p-Value as 0.2017 and t statistic as 1.2896.

Since p-value more than alpha we fail the reject H0 or accept H0 hypothesis.

Therefore there is enough evidence that we can conclude that population mean of both shingles A and B is equal. Here we have assumed that the distribution of both shingle data is normal, and that the variance of the two distribution are the same. If these assumptions are not met, another test need to be used.

THE END