## SMDM PROJECT REPORT

## ANSWER REPORT

## SULOCHANA

## 24-JULY-2022

## Report Overview:

This report gives proper descriptions of data by analyzing the dataset. Based on analysis, this document will explain what problems are encountered in this business as well as solutions for each problem.

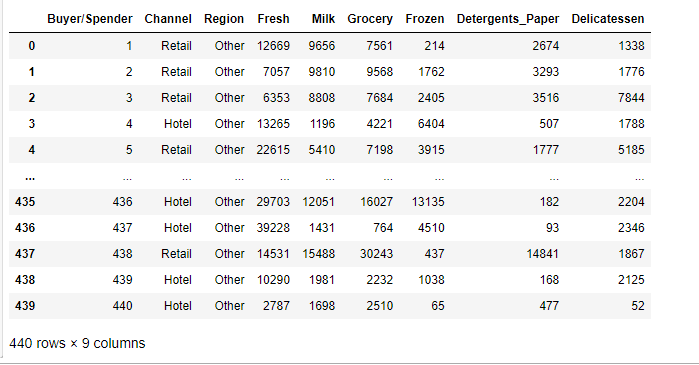
## Problems:

1. Wholesale Customer Analysis.
2. Clear Mountain State University (CMSU) Survey.
3. Hypothesis Testing for Quality of Shingles.

# Problem 1 - Wholesale Customers Analysis:

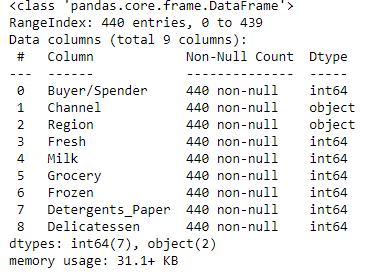
A wholesale distributor operating in different regions of Portugal has information on annual spending of several items in their stores across different regions and channels. The data consists of 440 large retailers’ annual spending on 6 different varieties of products in 3 different regions (Lisbon, Oporto, Other) and across different sales channel (Hotel, Retail).

The data is given in the File “Wholesale+Customers+Data.csv” as shown below.

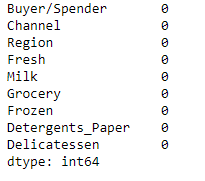


## EXPLORATORY ANALYSIS

* The data has 440 entries with 9 columns.
* It has two data types:
  + 7 integer type
  + 2 object type

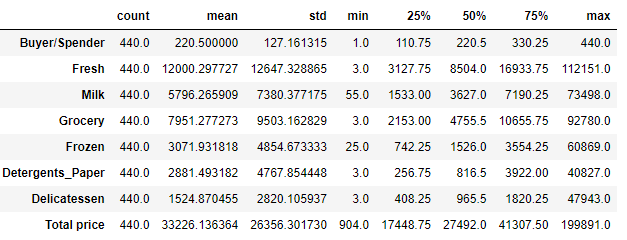


There are no null values in any of the column in the given data set.



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

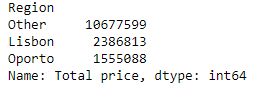
ANS:

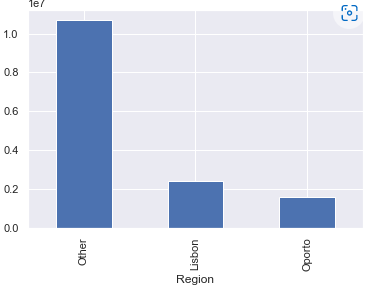


The above table summarizes the data by using descriptive statistics

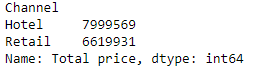
## Which Region and which Channel spent the most? Which Region and which Channel spent the least?

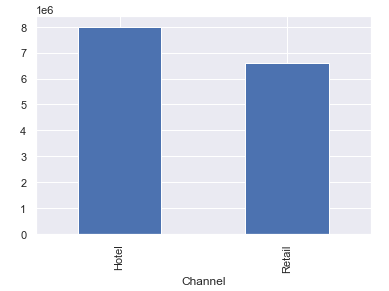
ANS:





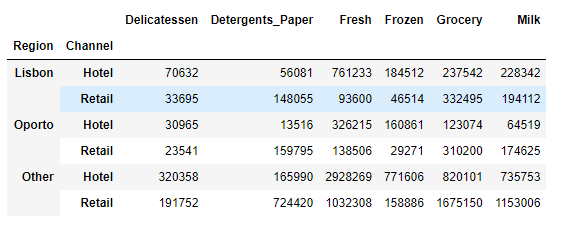
The above bar plot clearly explains the **other** region spent more compared to the reaming regions as well as the region **Oporto** spent the least.

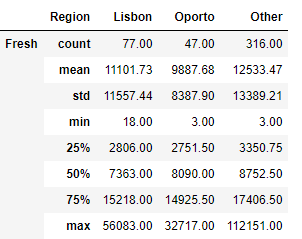


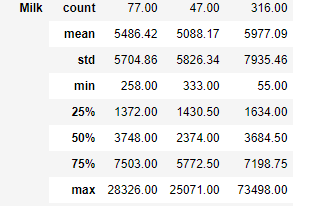


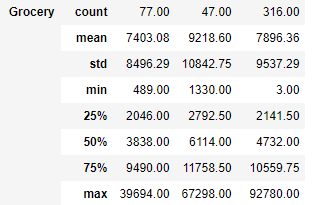
The above bar chart shows the channel **Hotel** spent the most and **Retail** spent less.

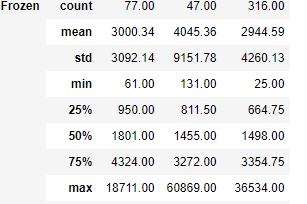
* 1. **There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.**

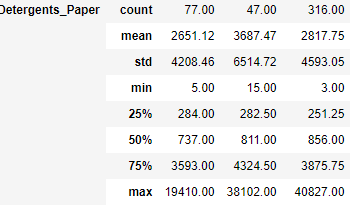


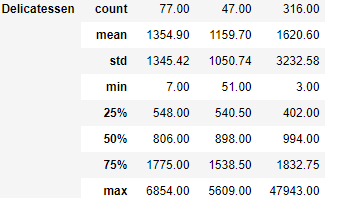




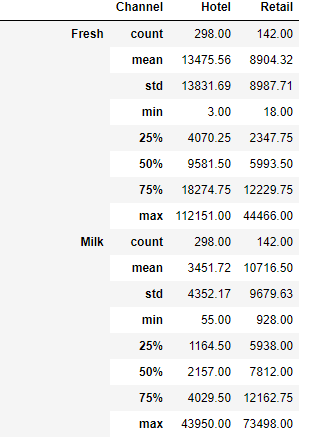


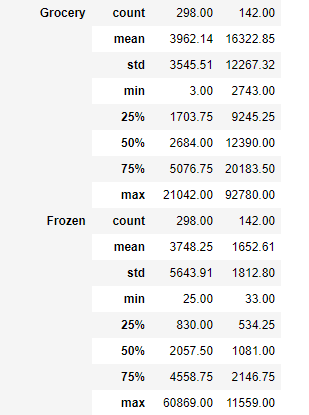


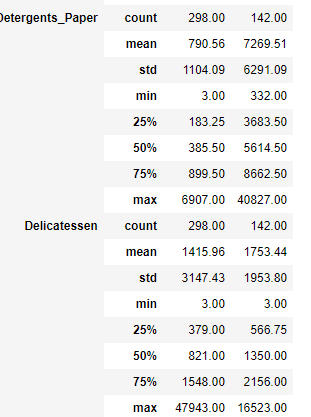




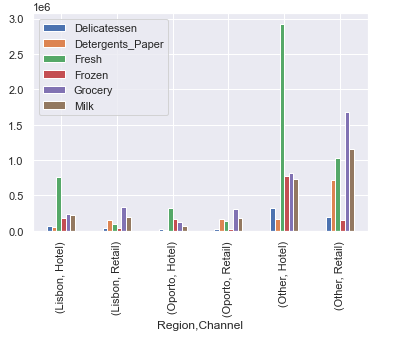
The above table describes about 6 varieties of products across region wise.





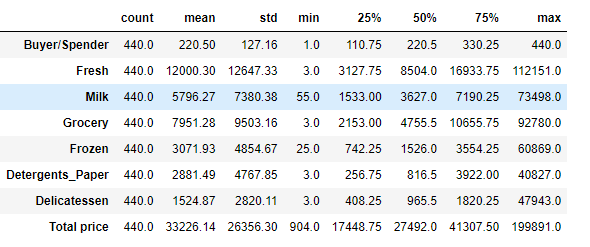


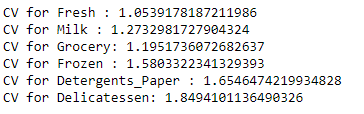
The above table describes about 6 varieties of products across channel wise.



The above tables and bar chart explains about 6 different varieties of products behavior across region wise as well as channel wise.

* 1. **On the basis of the descriptive measure of variability, which item shows the most inconsistent behavior? Which items shows the least inconsistent behavior?**





The above output represents the descriptive statistics of all the six Food Items **Fresh, Milk, Grocery, Frozen, Detergents Paper** and **Delicatessen.**

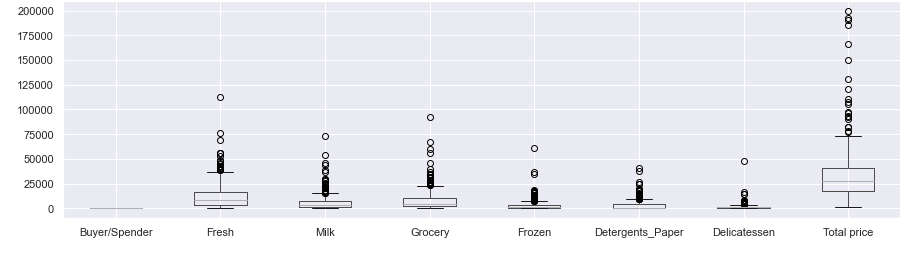
Here the consistency of all products calculated with the statistical measure **coefficient of variance.**

According to the above output, we can say that item delicatessen shows the highest inconsistent behavior in terms of price as the coefficient of variation is the highest for this item. On the other hand, fresh item shows the least coefficient of variation behavior in terms of price as the CV is the least.

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

ANS:

Here we can use box plot for identifying outliners in the data.



* According to the box plot outliners are present in the data.
* Outliners are present in the variables fresh, milk, frozen, detergent paper and delicatessen.
  1. **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.**
* Based on the analysis, the region **other** hashigher spending than the remaining regions. In the same way, channel **retail** has morespending when compared to other channels. So, the wholesale distributor can extend or open a new business in this region and channel. These changes will give profit to the business. In other regions like Lisbon and Oporto in the channel hotel, the fresh and grocery item shows high spending than other items. So, distributor can increase stock in these items in these regions.

# Problem 2 - Clear Mountain State University (CMSU) Survey:

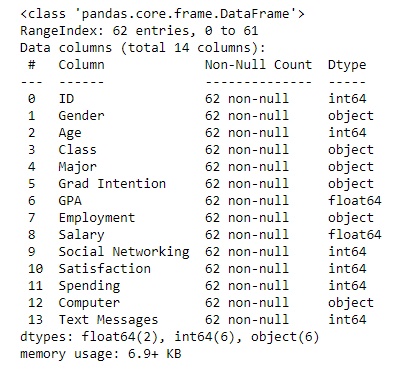
The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. CMSU creates and distributes a survey of 14 questions and receives responses from 62 undergraduates.

The Data is stored in the ***Survey*** data set as follows:

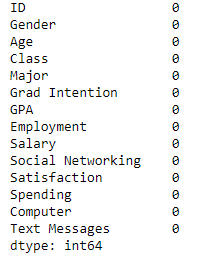


## Basic EDA

* The data has 62 entries with 14 columns.
* It has two data types:
  + 6 integer type
  + 6 object type
  + 2 Float

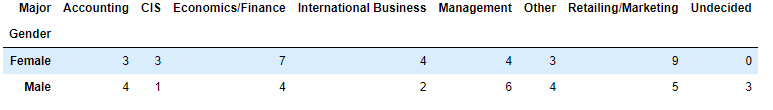


There are no null values in any of the column in the given data set.

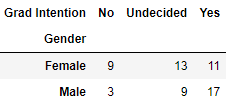


**2.1. for this data, construct the following contingency tables (Keep Gender as row variable)**

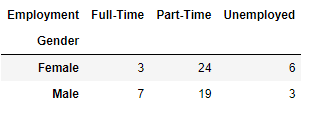
**2.1.1. Gender and Major**



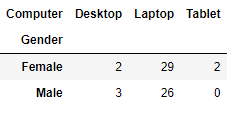
**2.1.2. Gender and Grad Intention**



**2.1.3. Gender and Employment**



**2.1.4. Gender and Computer**



**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?**



In this data set total 33 female and 29 male members are there.

The total numbers of students are 62.

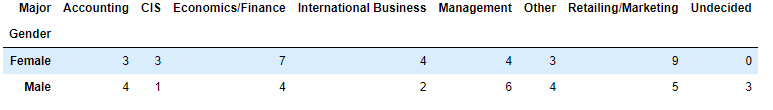
Probability that a randomly selected CMSU student will be male**: total number of male/total number of students =** 0.46774193548387094

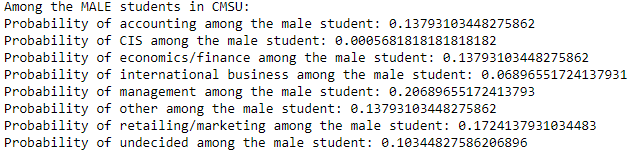
**2.2.2. What is the probability that a randomly selected CMSU student will be female?**

Probability that a randomly selected CMSU student will be female: **total number of female/total number of students =** 0.532258064516129

**2.3. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:**

**2.3.1. Find the conditional probability of different majors among the male students in CMSU.**





The probability of accounting among the male student: total no. of accounting students/total no. of male = 13.79%

The probability of CIS among the male student: total no. of accounting students/total no. of male = 3.45

The probability of economics/finance among the male student: total no. of accounting students/total no. of male = 13.79%

The probability of international business among the male student: total no. of accounting students/total no. of male = 6.9%

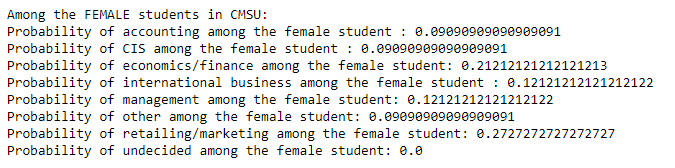
The probability of management among the male student: total no. of accounting students/total no. of male = 20.69%

The probability of other among the male student: total no. of accounting students/total no. of male = 13.79%

The probability of retailing/marketing among the male student: total no. of accounting students/total no. of male = 17.24%

The probability of undecided among the male student: total no. of accounting students/total no. of male = 10.34%

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



The probability of accounting among the female student: total no. of accounting students/total no. of male = 9.09%

The probability of CIS among the female student: total no. of accounting students/total no. of male = 9.09%

The probability of economics/finance among the female student: total no. of accounting students/total no. of male = 21.21%

The probability of international business among the female student: total no. of accounting students/total no. of male = 12.12%

The probability of management among the female student: total no. of accounting students/total no. of male = 12.12%

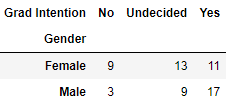
The probability of other among the female student: total no. of accounting students/total no. of male = 9.09%

The probability of retailing/marketing among the female student: total no. of accounting students/total no. of male = 27.27%

The probability of undecided among the female student: total no. of accounting students/total no. of male = 0%

**2.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:**

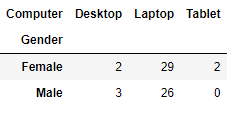
**2.4.1. Find the probability that a randomly chosen student is a male and intends to graduate.**

****

P (Grad Intention ∩ Male) = P (Grad Intention/ Male) x P (male) = 0.27419354838709675

**The probability that a randomly chosen student is a male and intends to graduate = 27.42%**

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

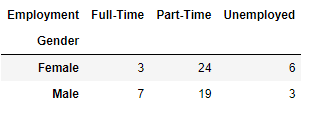


p(does not have a laptop ∩ Female) = P (does not have a laptop / Female) x P (Female) = 0.06451612903225806

**The probability that a randomly chosen student is a female and does not have a laptop = 6.45%**

**2.5. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:**

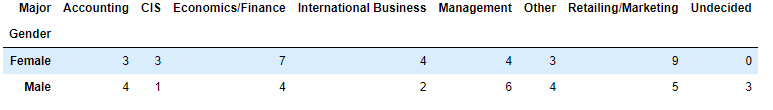
**2.5.1. Find the probability that a randomly chosen student is a male or has full-time employment?**



P (Male U full-time employment) = P (Male/Total Students) + P (Full time/Total Students) = 0.6290322580645161

**The probability that a randomly chosen student is a male or has full-time employment is = 62.90%**

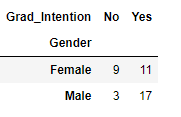
**2.5.2. Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management.**

****

p(international total female)+p(management total female) = 0.24242424242424243

**Conditional 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?**



P (F) P (Yes) = (20/40)\*(28/40))

P (F ∩ Yes) = (11/40))

P (Female) P (Yes) = 0.35

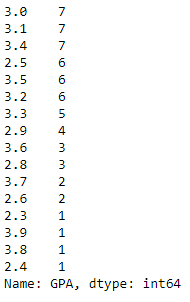
P (Female ∩ Yes) = 0.275

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.**

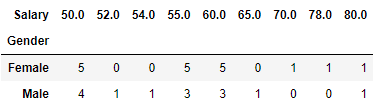
**Answer the following questions based on the data**

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



Randomly chosen student and the probability that his/her GPA is less than 3 = 0.27419354838709675

**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.**



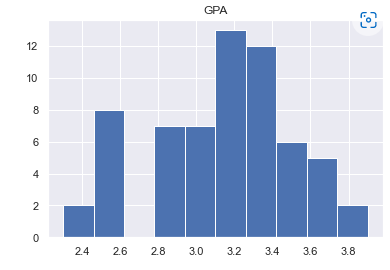
Conditional probability that a randomly selected male earns 50 or more = 22.58%

Conditional probability that a randomly selected female earns 50 or more = 29.03%

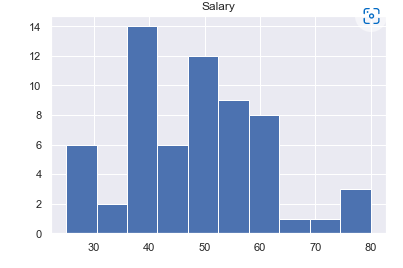
**2.8. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarizing your conclusions.**

Here histogram plot used to determine whether the GPA variable follow normal distribution or not.

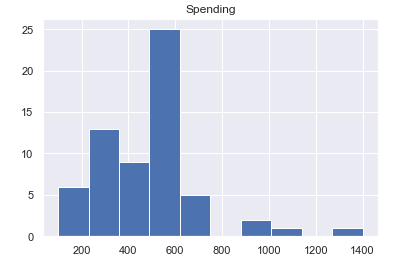
Histogram for GPA



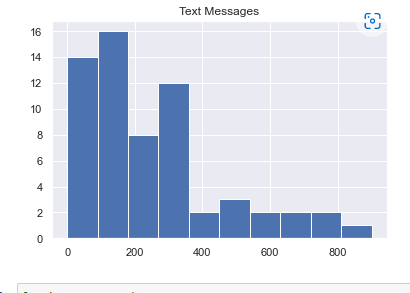
Histogram for salary



Histogram for spending



Histogram for text massages



* GPA look like normal distribution and it has very less Skewness and it is negative, so it is towards the left.
* Salary also looks like normal distribution but it has very less skewness. It is skewed towards the right.
* Spending is highly skewed.
* Text Message also highly skewed.

**Problem 3**

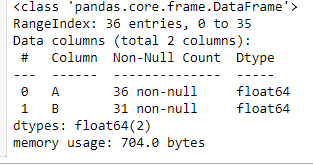
**Hypothesis Testing for Quality of Shingles**

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](https://olympus.mygreatlearning.com/courses/78177/files/5510637/download?verifier=WlrDhRuauwdWHHbuFxyKx9S5BFoEH1SVtxO3yMqb&wrap=1)) includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.



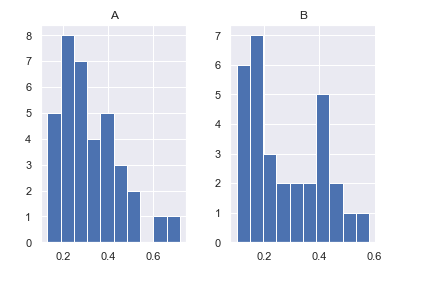
**Basic EDA**



There are total 36 rows and 2 columns in the dataset. These two columns are float data type.



The above table shows 5 null values in B variable.



The above histogram shows both A and B variable are highly right skewed.

**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.**

**For A shingles**

**A null hypothesis** a hypothesis that says there is no statistical significance between the two variables in the hypothesis. It is the hypothesis that the researcher is trying to disprove.

**An alternative hypothesis** simply is the inverse, or opposite, of the null hypothesis.

### Step 1: Define null and alternative hypotheses

* **Here for the A shingles, the null hypothesis to test whether the population mean moisture content is less than or equal to 0.35 pound per square feet is given.**
* **For A shingles an alternative hypothesis to test whether the population mean moisture content is more than or equal to 0.35 pound per square feet is given.**

**H0: mean moisture content <= 0.35**

**HA: mean moisture content > 0.35**

### Step 2: Decide the significance level

Level of significance: 0.05

### Step 3: Identify the test statistic

Here, we have a samples and mean value but we do not have the population standard deviation. And also the sample size is not a large. So, in this case we will use the t statistic test.

Since, we are testing for one sample shingles we will use one sample t test (**ttest\_1samp**) but in python by default it shows two side test results. Then we have to divide the p-value by 2 as our test is one sided.

### Step 4: Calculate the p - value and test statistic

The sample size of A shingles for this problem is 36.

One sample t test

**t statistic: -1.4735046253382782 p value: 0.14955266289815025**

### Step 5: Decide to reject or accept null hypothesis

Level of significance: 0.05

**We have no evidence to reject the null hypothesis since p value > Level of significance**

Our one-sample t-test p-value= 0.14955266289815025

**For B shingles**

### Step 1: Define null and alternative hypotheses

* **Here for the A shingles, the null hypothesis to test whether the population mean moisture content is less than or equal to 0.35 pound per square feet is given.**
* **For A shingles an alternative hypothesis to test whether the population mean moisture content is more than or equal to 0.35 pound per square feet is given.**

**H0: mean moisture content <= 0.35**

**HA: mean moisture content > 0.35**

### Step 2: Decide the significance level

Level of significance: 0.05

### Step 3: Identify the test statistic

Here, we have a samples and mean value but we do not have the population standard deviation. And also the sample size is not a large. So, in this case we will use the t statistic test.

Since, we are testing for one sample shingles we will use one sample t test (**ttest\_1samp**) but in python by default it shows two side test results. Then we have to divide the p-value by 2 as our test is one sided.

**Step 4: Calculate the p - value and test statistic**

The sample size of B shingles for this problem is 36.

One sample t test:

**t statistic: -3.1003313069986995**

**p value: 0.004180954800638365**

### Step 5: Decide to reject or accept null hypothesis

Level of significance: 0.05

**We have evidence to reject the null hypothesis since p value < Level of significance**

Our one-sample t-test p-value= 0.004180954800638365

**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?**

### Step 1: Define null and alternative hypotheses

## H0: mean moisture content of A = mean moisture content of B

## HA: mean moisture content of A ≠ mean moisture content of B

### Step 2: Decide the significance level

Level of significance: 0.05

### Step 3: Identify the test statistic

Here we have two samples A and B and we do not know the population standard deviation but we know the mean. So we will use the ttest for these samples.

Since we are testing two samples A and B we will use two sample t test(ttest\_1samp, ttest\_ind).

**Step 4: Calculate the p - value and test statistic**

**tstat 1.2896282719661123**

**P Value 0.2017496571835306**

### Step 5: Decide to reject or accept null hypothesis

Alpha value = 0.05

Two-sample t-test p-value= 0.2017496571835306

**We do not have enough evidence to reject the null hypothesis in favor of alternative hypothesis**

**We conclude that the population means for shingles A and B are equal.**