REPORT

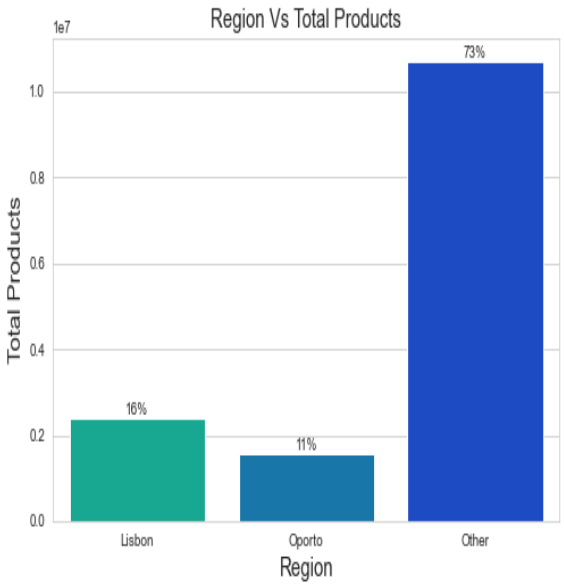
Author- Ritesh Singh

PROBLEM 1

1.1 Use methods of descriptive statistics to summarize data. Which Region and which Channel seems to spend more? Which Region and which Channel seems to spend less?

- REGIONWISE

Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total_Products
Lisbon	854833	422454	570037	231026	204136	104327	2386813
Oporto	464721	239144	433274	190132	173311	54506	1555088
Other	3960577	1888759	2495251	930492	890410	512110	10677599



A new column is added which has sum of all the spending's.
Total Products = **Fresh**+**Milk**+**Grocery**+**Frozen**+**Detergents_Paper**+**Delicatessen**
73% of Total products sold are from Other Regions, 16% from Lisbon and 11% from Oporto.

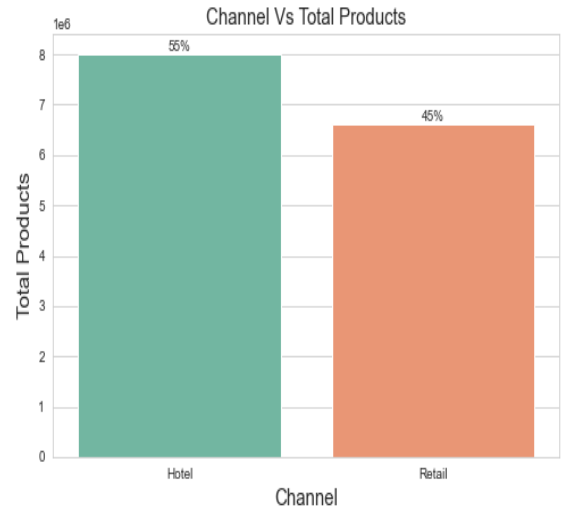
Then Data is grouped based on Region. Grouped data is used to plot a Bar plot showing Total Products and Regions.

Data was grouped by Regions (i.e. Lisbon, Oporto and Other). We can observe that spending on total products in **Other** region is maximum and minimum in **Oporto**. Following are Observations:

- The Population in Oporto is comparatively less than Lisbon and Other regions that is why less people buy products in these areas.
- Wrong Assortment in Lisbon and Oporto- People from different places has different taste of products. So, keeping a good mixture of products can do wonders. Assortment planning based on customer demand can be implemented in Lisbon and Oporto.
- Promotions Aren't enough - Includes offers, discounts etc.
- Competitors are highly active in Lisbon and Oporto.
- Need online presence, which includes locations, select and collect the product features.
- Price has to match with the competitions. If competitors are providing items at lower price.

- CHANNELWISE

Channel	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total_Products
Hotel	4015717	1028614	1180717	1116979	235587	421955	7999569
Retail	1264414	1521743	2317845	234671	1032270	248988	6619931



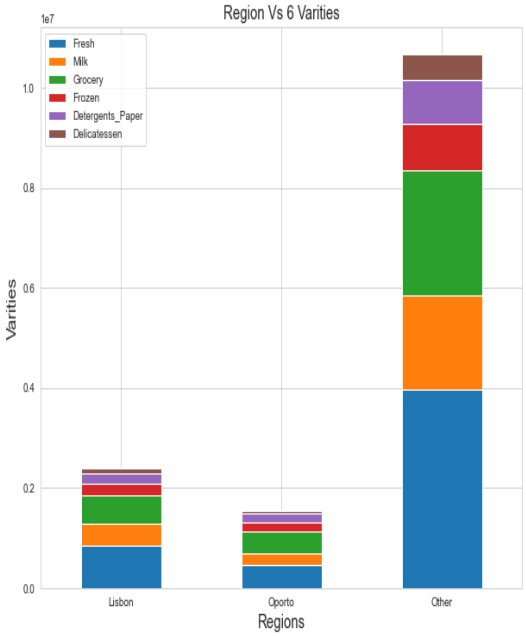
Data is grouped based on Region. Grouped data is used to plot a Bar plot showing Total Products and Regions.

56% of Total products are sold from hotels and 45% are sold from Retail

Data was grouped by Channel (i.e. Hotel and Retail). We can observe that spending on total products in **Hotels** are higher than **Retail**. Following are Observations

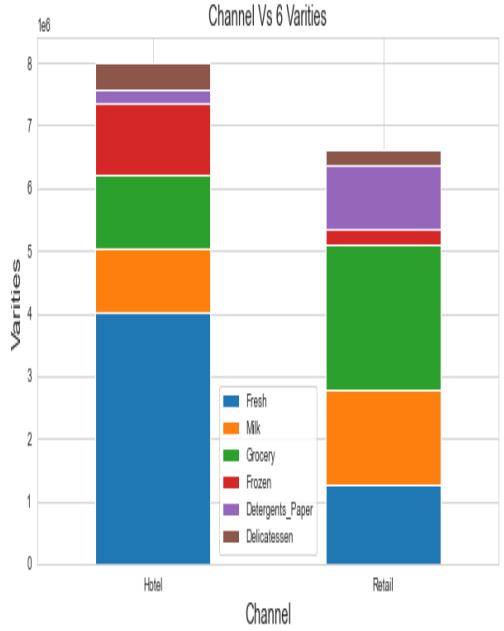
- The climate of Portugal is temperate and influenced by the Atlantic Ocean. In the north, the climate is cool and rainy so there is possibility that People like to stay indoors in hotels.
- Farm to Table Dining is the new trend so people are attracted to Hotels.
- Some Population is too busy so they do not have time to buy from retail and they get their services from hotels.
- Hotels sell at a higher rate, as ambiance and other services are also included.
- Number of Hotels are more compared to Retail store.
- We can increase the number of Retail stores, or Hotels can tie up to open a retail store attached to it.

1.2 There are 6 different varieties of items are considered. Do all varieties show similar behaviour across Region and Channel? Provide justification for your answer.



For Every Region Stacked Bar Plot is made which shows the data of 6 Varieties for every region. Py file contains separate boxplots and Describe Tables for this data. Region wise the behaviour is **similar** in all regions.
Fresh > Grocery > Milk > Frozen > Detergent paper > Delicatessen
Data was grouped by Regions (i.e. Lisbon, Oporto and Other). We can observe that spending on total products in **Other** region is maximum and minimum in **Oporto**. Following are Observations:

- More people are buying Fresh items. Grocery comes next in all three regions. Fresh and Grocery items have lot of large orders, which we can see in box plots. These orders may be made for some large gatherings like parties, Events etc in those areas.
- People are not attracted to Delicatessen products, indicates that foreign food items aren't liked much in these areas.
- Though we can see that the behaviour of all varieties is same in all the regions
- ***Fresh > Grocery > Milk > Frozen > Detergent paper > Delicatessen***
- Foreign food items can be launched slowly in the market that too in a lower price range. So that people will buy cheaper foreign food items more. Then in future slowly we can launch big products if market is fruitful.
- We can also observe that except Oporto all Other Regions has good amount of sales. The possible reason for this is a smaller number of hotels in Oporto.



For Every Channel Stacked Bar Plot is made which shows the data of 6 Varieties for every Channel. Py file contains separate boxplots and Describe Tables for this data.

Data was grouped by Channel (i.e. Hotel and Retail). We can observe that spending on total products in **Hotels** is more

than **Retail** stores. Following are Observations:

- When we see behaviour of all varieties in both the channels, we see a **different** pattern. In Hotels people tend to buy Fresh Items more and in Retail people are buying Grocery items more than anything else. It is because Hotels provide freshly prepared food items, number of hotels is more and demand of cooked fresh food is high.
- From Retail people are buying Grocery items, Detergent papers and Milk in large numbers. We can also observe large orders in all categories (py file-boxplot). Maybe customers buy once in a while and store the items.
- If customer is keeping items for a long time, we may focus on **SHELF LIFE** of the products.

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?

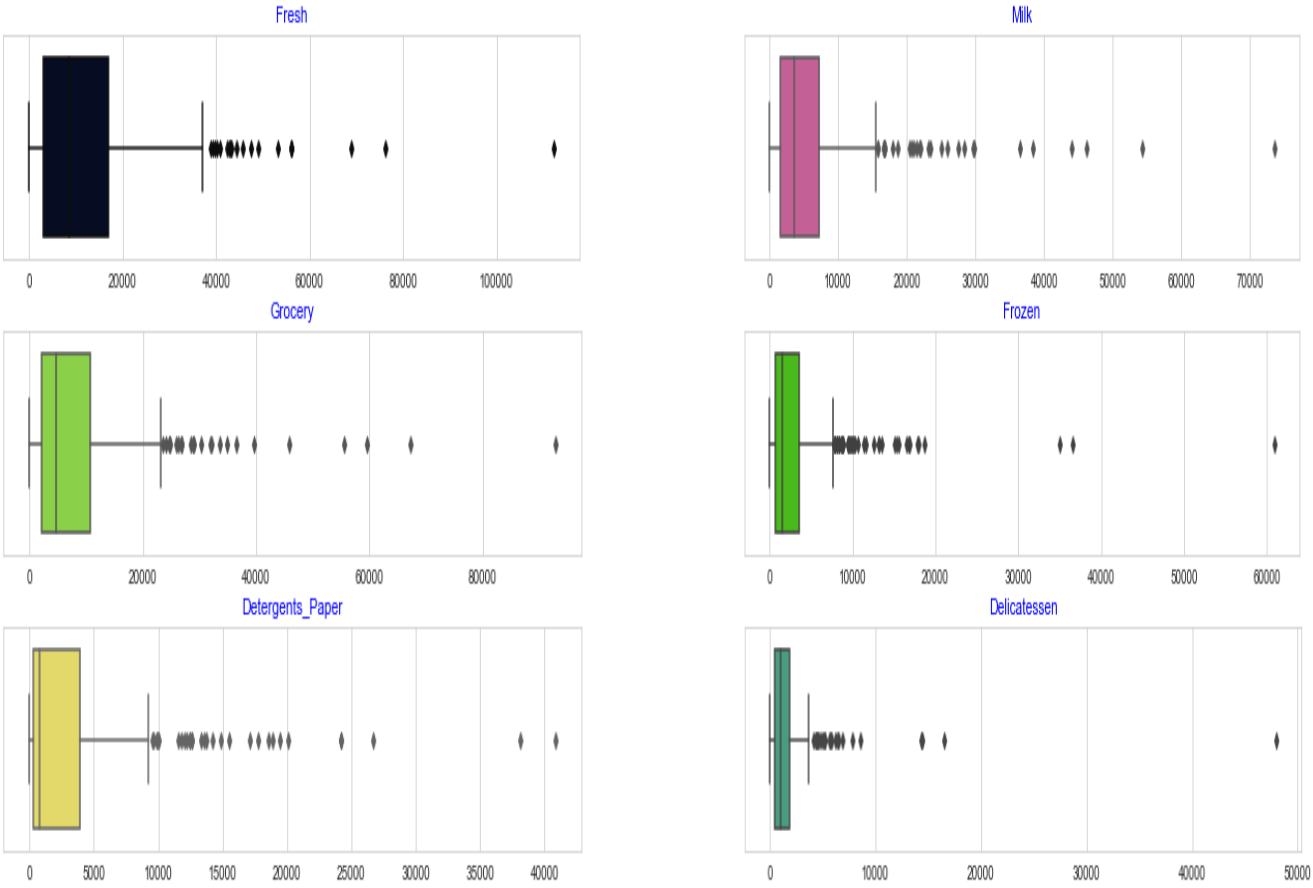
	mean	std	coefficient of variation (CV)
Fresh	12000.2977	12647.3289	1.053918
Milk	5796.26591	7380.37718	1.273299
Grocery	7951.27727	9503.16283	1.195174
Frozen	3071.93182	4854.67333	1.580332
Detergents_Paper	2881.49318	4767.85445	**1.654647**
Delicatessen	1524.87046	2820.10594	**1.849407**
Total_Products	33226.1364	26356.3017	0.79324

A new column for Coefficient of variation (CV) is added.

$$CV = \frac{Standard\ Deviation}{Mean}$$

we can see that CV of Fresh is less compared to other varieties. Only Fresh items shows least inconsistent behaviour, while Delicatessen and Detergents Papers shows more inconsistency.

1.4 Are there any outliers in the data?



Boxplots to show Outliers, uses values from Q1-1.5IQR and Q3+1.5IQR to plot outliers in data.

All Varieties have outliers towards right side of the mean. This means all are Right 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 in the data?

For Business to grow increasing the number of Retail outlets is one way. Future is in collaboration so wherever new Hotels are opened they can collaborate with the retail stores nearby. Marketers need to implement this strategy so that all can earn profits.

There are lot of customers who are buying in large amount, either these orders are for parties, some huge gatherings or events or customer is storing for a long time. Keeping both the things in mind giving discount and improving the shelf life of the products is one of the improvements.

Foreign products (Delicatessen) are not liked in Portugal. Before bringing Coffee as a beverage company launched a coffee toffee. Off course Toffees are cheaper, this is how customers become habitual to the products.
Same way before bringing a new costly product in the market launching a cheaper variant in the market can be a good idea for the Business.

Bundling and providing discounts can be implemented to improve the sales of Detergent papers. On the other hand, Market growth rate in Delicatessen is very low and possible relative market growth is also low. So, one of the options is to kill these products in low performing regions.

Hotels are covering 55% of Market and rest is covered by Retail. Fresh has a market of 36%, then 24% of Market is taken by Grocery. Proper Inventory management has to be implemented.

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

Following is the describe table for the data (only Numeric data).

	mean	std	min	0.25	0.5	0.75	max
GPA	3.129032	0.377388	2.3	2.9	3.2	3.4	3.9
Salary	48.548387	12.080912	25	40	50	55	80
Text Messages	246.20968	214.46595	0	100	200	300	900
Spending	482.01613	221.95381	100	313	500	600	1400

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

Contingency tables provide information on Joint, Marginal and Conditional Probabilities

2.1.1 Gender and Major

Gender/Major	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

Gender/Grad Intention	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

Gender/Employment	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

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

2.2.1 What is the probability that a randomly selected CMSU student will be male?

Total no. of students is: 62
Number of Male: 29
Probability that a randomly selected CMSU student will be male: 29 /62= 0.47

2.2.2 What is the probability that a randomly selected CMSU student will be female?
Total no. of students is: 62
Number of Female: 29
Probability that a randomly selected CMSU student will be female: 33 /62= 0.53

2.3.1 Find the conditional probability of different majors among the male students in CMSU.
P (Accounting / MALE) =4/29= 0.14
P (CIS / MALE) =1/29= 0.03
P (Economics/Finance / MALE) =4/29= 0.14
P (International Business / MALE) =2/29= 0.07
P (Management / MALE) =6/29= 0.21
P (Other / MALE) =4/29= 0.14
P (Retailing/Marketing / MALE) =5/29= 0.17
P (Undecided / MALE) =3/29= 0.1

Note: Formula used

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

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

P (Accounting / FEMALE) =3/33= 0.09
P (CIS / FEMALE) =3/33= 0.09
P (Economics/Finance / FEMALE) =7/33= 0.21
P (International Business / FEMALE) =4/33= 0.12
P (Management / FEMALE) =4/33= 0.12
P (Other / FEMALE) =3/33= 0.09
P (Retailing/Marketing / FEMALE) =9/33= 0.27
P (Undecided / FEMALE) =0/33= 0.0

2.4.1 Find the probability That a randomly chosen student is a male AND intends to graduate.

The probability That a randomly chosen student is a male and intends to graduate

P(Male and int to grad) = 17/62 = 0.274

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

The probability that a randomly selected student is a female and does NOT have a laptop is

P(Female and Desktop)+P(Female and Tablet)=4/62=0.0645

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

The probability that a randomly chosen student is either a male or has a full-time employment

P(Male U Full Time Education)=P(Male)+P(FullTime)-P(Male and FullTime)
= (29/62+10/62)-7/62
=0.5161

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

No. of Female students majoring in international business or management=8

Total Female students=33

The conditional probability that given a female student is randomly chosen, she is majoring in international business or management is

=8/33= 0.2424

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 graduate intention and being female are independent events?

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

P(F n Yes)=11/40
P(F)=20/40
P(Yes)=28/40
P(F n Yes) != P(F).P(Yes)
So graduate intention and being female are **NOT independent events**

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

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

$$P\left(\frac{\#\text{ of students scoring}<3}{Total\ number\ of\ students}\right)$$

= 17/62 = 0.27

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

Probability that a randomly selected Male earns 50 or more=

$$P\left(\frac{\#\text{ of Male students earning}\geq 50}{Total\ number\ of\ Male\ students}\right)$$

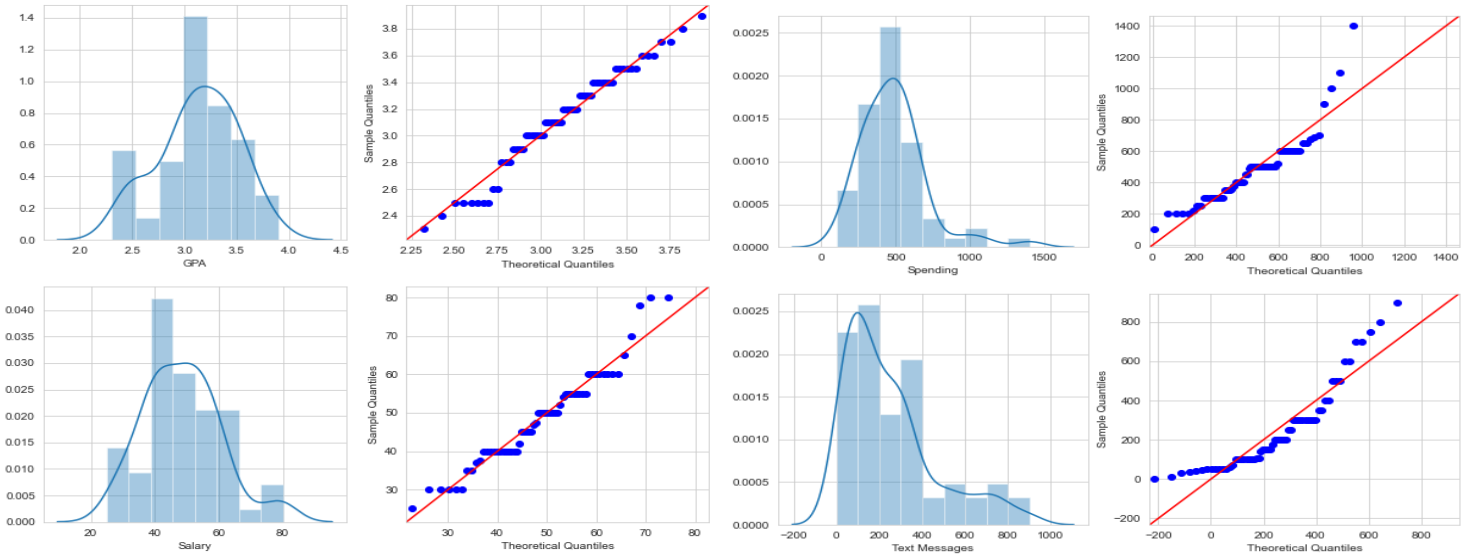
=14/29=0.483

Probability that a randomly selected Female earns 50 or more

$$= P\left(\frac{\#\text{ of Female students earning}\geq 50}{Total\ number\ of\ Female\ students}\right)$$

=18/33= 0.545

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.



	Skew
GPA	-0.3146
Salary	0.534701
Text Messages	1.295808
Spending	1.585915

Concept of QQ Plot and skewness is used to check how close the Distribution is to Normal Distribution.

- We can see that **GPA** has a negative skewness and hence has left skewness and follows a distribution which is very close to normal.
- We can see that **Salary** has a negative skewness and hence has Right skewness and follows a distribution which is close to normal.
- **Spending** is also a close call, if we have a greater number of samples, we may be able to comment better. Still it is close to Normal and is Right skewed.
- Text **Messages** follows a distribution which is close to normal and is Right Skewed.

Note: QQ Plot gives us better representation if we have a large sample size

2.8.2 Write a note summarizing your conclusions.

- Sample contains 47% Male students and 53% Female students. A lot of students like Retail/Marketing more than any other stream. Top three streams liked by students in order are:
  Retail/Marketing > Economics > Management
- More Female students are working in Part time jobs and more Male students are working in Full time jobs.
- 27% Male students intend to graduate.
- Most of the students owns Laptops.
- 4% females do not own a laptop.
- 51% of students are either male or have full time employment.
- 24% female students are majoring International Business or Management.
- Graduate intention and being female are not independent events.
- 48% Male students have salaries more than 50 and 54% of Female students have salaries more than 50.
- 27 % of Students have GPA less than 3. May be this is because they are working somewhere.
- Distributions in 4 Numerical variables are close to normal. We can have a better understanding of Distribution if Outliers are removed and sample size is increased.

PROBLEM 3

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 pound per 100 square feet.

- We want to take precautionary measures if there is more moisture content For the 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:

  H0= u<=0.35

  HA= u>0.35

- For the 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:

  H0: u<=0.35

  HA: u>0.35

- alpha=0.05

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

t-Test for one sample is used in this problem as population Standard Deviation is not known

$$\text{t-test statistic} = \frac{\boldsymbol{X} - \mu}{\frac{\boldsymbol{s}}{\sqrt{\boldsymbol{n}}}}$$

X=Sample mean
μ=Population mean/Theoretical mean
s=sample standard deviation
n=sample size

t-stat and p-value for **A Shingles** are -1.474 and **0.075**
Since in A Shingles **p value> alpha**
We have enough evidence to **accept NULL HYPOTHESIS**

t-stat and p-value for **B Shingles** are -3.1 and **0.002**
Since in B Shingles **p value< alpha**
We have enough evidence to **reject NULL HYPOTHESIS**

3.2 Do you think that the population means 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?

α = 0.05 and the population standard deviation is not known.
Assumptions:
-Both A and B are **INDEPENDENT variables.**
-Scale of measurement applied to the data collected follows a continuous scale.
-Variables follows a normal distribution

- We have two samples and we do not know the population standard deviation.
- Sample sizes for both samples are not same.

Null Hypothesis states that the population means for shingles A and B are equal

$$H0: \mu_A = \mu_B$$

Alternative Hypothesis states that that the population means for shingles A and B are not equal

$$HA: \mu_A! = \mu_B$$

We will perform two sample t test in this case

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

t-stat and p-value are 1.289 and 0.201 respectively.

Two-sample t-test p-value= 0.2017496571835306

Since p value > alpha

We have enough evidence to ACCEPT the null hypothesis.

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