Business Report Project

Sub: - SMDM

INDEX

Problem 1: - Wholesale Customers Analysis

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).

Answer: -

We imported the 'Wholesale Customer data' dataset in python to analyze the spend under each store items across regions and channel to find solutions to each problem. Below is the detailed approach and answer.

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 Statastics Data:
```

	count	mean	std	min	25%	50%	75%	max
Fresh	440.0	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0
Milk	440.0	5796.265909	7380.377175	55.0	1533.00	3627.0	7190.25	73498.0
Grocery	440.0	7951.277273	9503.162829	3.0	2153.00	4755.5	10655.75	92780.0
Frozen	440.0	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0
Detergents_Paper	440.0	2881.493182	4767.854448	3.0	256.75	816.5	3922.00	40827.0
Delicatessen	440.0	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0

(Fig1: - - Descriptive Statistics Data)

```
Descriptive Statastics Data with Channel & Retail:
```

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
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

(Fig2: - Descriptive Statistics of data with channel & Retail)

1.1.1 Which Region and which Channel spent the most?
1.1.2 Which Region and which Channel spent the least?

Solution: -

Using describe function in python we first looked at the basic descriptive statistics of the data set. Using bar graph with Region and Channel we were able to identify region with maximum spend and minimum spend. Below is the bar graph representation-Looking at the bar graph, Hotel Channel spends more and Retail spends least.

- Highest spend in the Channel is from Hotel = 5742077$
- Highest spend in the Region is from other = 10677599$

Below is the output from Python based on Channel

Channel

```
Channel
Hotel     7999569
Retail    6619931
Name: Spending, dtype: int64
```

Similarly, we grouped totals by region to get totals by region.

Below is the output from Python based on Channel and Region

```
Region  Channel
Lisbon  Hotel      1538342
        Retail      848471
Oporto  Hotel       719150
        Retail      835938
Other   Hotel      5742077
        Retail     4935522
Name: Spending, dtype: int64
```

- Other regions spend amount is 10741625$ with the highest spend amount and
- Oporto region spend amount is 1569987$ and has least spend amount by Region.

Below is the output from Python Based on Region

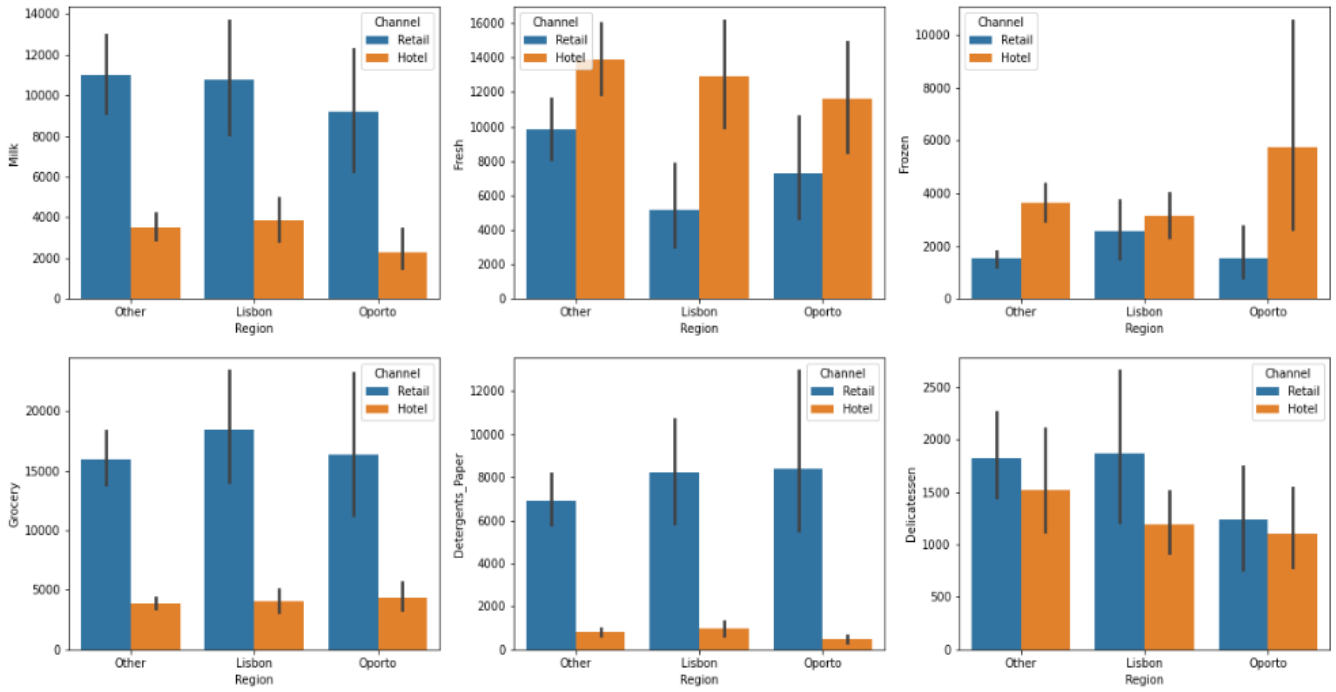
```
Region
Lisbon: -   2404908
Oporto: -   1569987
Other: -    10741625
Data type: int64
```

- lowest spend in the Region/Channel is from Oporto/Hotel = 719150$

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

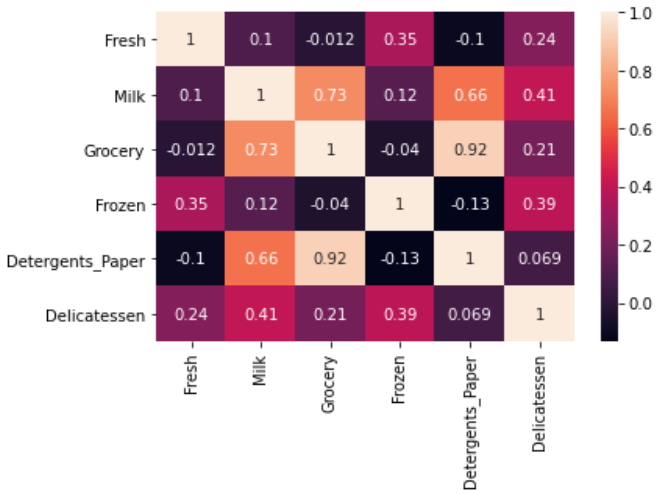
Solution:

Using pivot tables for each category and checking spend across Region and Channel we get the following outputs.



(Fig-3 bar plot graph)

Looking at the above charts, we see that some categories like Milk, Grocery & Detergent Paper have higher spent in the Retail channel versus Hotel, across all regions. On the other hand, Fresh and Frozen have higher consumption in the Hotel channel versus Retail, across all regions. Also, if we plot a box plot, we can summarize that the spend for Fresh and groceries is the maximum across region and channel while for Delicatessen it is the least across region and channel. The output heatmap is below.



(Fig4: - Correlation)

1.3 Problem On the basis of the descriptive measure of variability, which item shows the most inconsistent behavior? Which items shows the least inconsistent behavior?

Solution:

Using Coefficient of Variation, we find out the least value is of Category "Fresh" and highest value is of Category "Delicatessen."

- from the given data most inconsistent behavior shown by item is Delicatessen (1.847304)
- And least inconsistent behavior shown by item is Fresh. (1.052720)

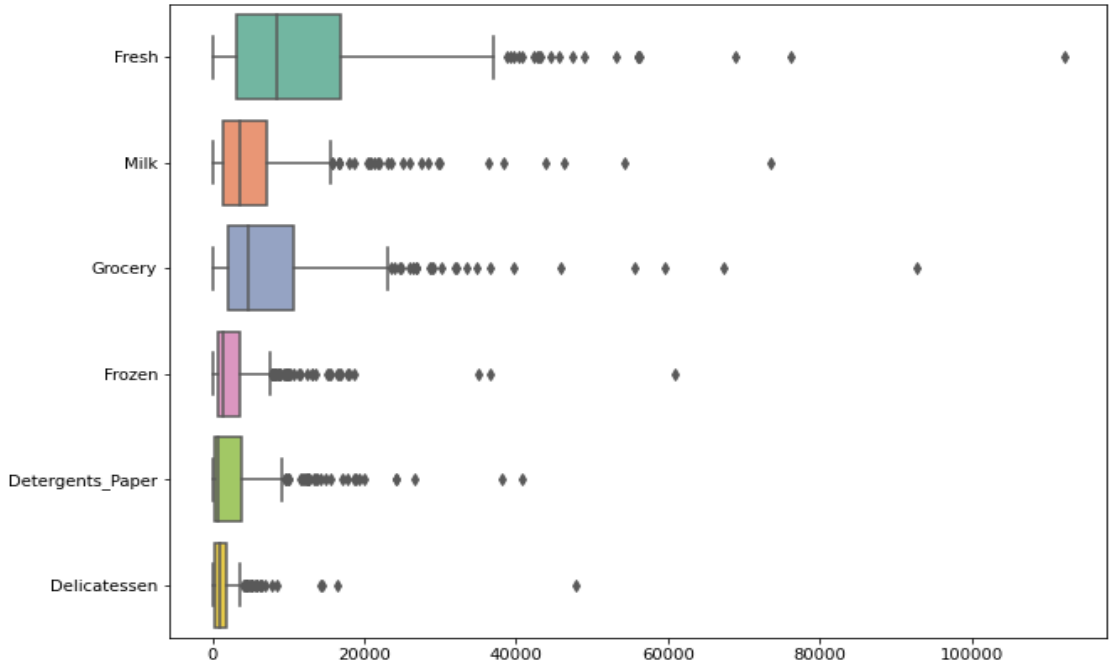
Below is the output from Python.

```
Covariance for Fresh is 1.052720
Covariance for milk is 1.271851
Covariance for Grocery is 1.193815
Covariance for Frozen is 1.578536
Covariance for Detergents_Paper is 1.652766
Covariance for Delicatessen is 1.847304
```

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

Solution:

To find out outliers we plotted boxplot and the output gives the details that in all the data there are outliers.



(Fig5:- Outliers)

List of outliers
- Fresh - 112151
- Milk - 73498
- Grocery - 92780
- Frozen - 60869
- Detergents Paper - 40827
- Delicatessen – 47943

1.5 Problem 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

Solution:

As per My analysis, I find out that there are inconsistencies in spending of different items by calculating Coefficient of Variation & Standard deviation, which should be helping to minimized. The spending of Hotel and Retail channel are different which should be more or less than or equal. And also spent should equal for different regions. Need to focus on other items also than “Fresh” and “Grocery”.

Problem 2: - CMSU Survey Data Analysis

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.

Summary

This business report provides detailed explanation of approach to each problem given in the assignment and provides relative information with regards to solving the problem.

We imported the 'CMSU Survey-1' dataset in python to analyze the data about the undergraduate students who attend CMSU. Below is the detailed approach and answer.

2.1. For this data, construct the following contingency tables.

(Keep Gender as row variable)

2.1.1 Gender **and Major**

Solution:

Below is the output.

Major Gender	Accounting	CIS	Economics/Finance	International Business
Female	3	3	7	4
Male	4	1	4	2

Major Gender	Management	Other	Retailing / Marketing	Undecided
Female	4	3	9	0
Male	6	4	5	3

2.1.2 Gender and Grad Intention

Solution:
Below is the output from Python.

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

2.1.3 Gender and Employment

Solution:

Below is the output from Python.

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

2.1.4 Gender and Computer

Solution:

Below is the output from Python.

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?

Solution:

For this we need to find out total male students out of whole student from the given data. After calculation we got the result that probability of 46.8% student will be male in CMSU if randomly selected

- probability = total number of males/ total number of entries
- total number of males = 29
- total number of entries = 62
- probability of males = 29/62
- 46.8% chances of males are probably randomly selected.

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

Solution:

For this we need to find out the total female students out of whole student from the given data. After calculation we got the result that probability of 53.2% student will be female in CMSU if randomly selected.

- probability = total number of females/ total number of Students
- total number of females = 33
- total number of Students = 62
- probability of males = 33/62
- 53.2% chances of females are probably randomly selected.

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.

Solution:

Using contingency tables of Gender and Majors we got the total numbers of males and number of males opting for different majors.

Below is the output from Python –

```
Probability of Males opting for Accounting is 13.8%
Probability of Males opting for CIS is 3.4%
Probability of Males opting for Economics/Finance is 13.8%
Probability of Males opting for International Business/Finance is 6.9%
Probability of Males opting for Management is 20.7%
Probability of Males opting for Other is 13.8%
Probability of Males opting for Retailing Marketing is 17.2%
Probability of Males opting for Undecided is 10.3%
```

And from this output we can easily say that most of the male's students prefer Management as Majors and CIS is the least preferred one

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

Solution:

Using contingency tables of Gender and Majors we got the total numbers of females and number of female spotting for different majors.

Below is the output from Python –

```
Probability of females opting for Accounting is 9.1%
Probability of females opting for CIS is 9.1%
Probability of females opting for Economics/Finance is 21.2%
Probability of females opting for International Business/Finance is 12.1%
Probability of females opting for Management is 12.1%
Probability of females opting for Other is 9.1%
Probability of females opting for Retailing Marketing is 27.3%
Probability of females opting for Undecided is 0.0%
```

And from this output we can easily say that most of the female students prefer Retailing/Marketing as Majors.

2.4 Problem. 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.

Solution:

Using contingency tables of Gender and Grad Intention we got the total numbers of males and number of males intends to be graduate.

And post calculation we find out that: -

```
Probability of males are graduate. is 58.6%
```

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

Solution:

Using contingency tables of Gender and Computer we got the total numbers of females and number of females does not have a laptop.

And post calculation we find out that: -

```
Probability of female students does not have laptop is 12.12%
```

2.5 Problem. 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 either a male or has full-time employment?

Solution:

Using contingency tables of Gender and Employment we got the total numbers of males and number of males who are full time employed.

And post calculation we find out that: -

```
Probability of total number of students has a male is 46.8%
Probability of total number of students has a full-time job is 11.3%
```

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

Solution:

Using contingency tables of Gender and Major we got the total numbers of females and number of females majoring in international business or management.

And post calculation we find out that: -

```
female student is randomly chosen, and she is majoring in international bu
siness or management is 24.2%
```

2.6 Problem. 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?

Solution:
Python output: -

Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

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

Is the graduate intention and being female are independent events?

- The Probability that a randomly selected student 'being female.'
- The Probability that a randomly selected student the graduate intention and being female.
- P (Grad Intention Yes) = 28/40 = 0.7
- P (Grad Intention Yes | female) = 11 / 20 = 0.55
- These probabilities are not equal. This suggests that the two events are independent.

2.7 Problem. 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?

Solution:

Using contingency tables of Gender and GPA we got the total numbers of students and number of students less than 3.

And post calculation we find out that: -

```
Probability of less than 3 GPA his/her students is 27.4%
```

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.

Solution:

Using contingency tables of Gender and Salary we got the total numbers of Male and Female and number of male and female earning 50 or more.

And post calculation we find out that: -

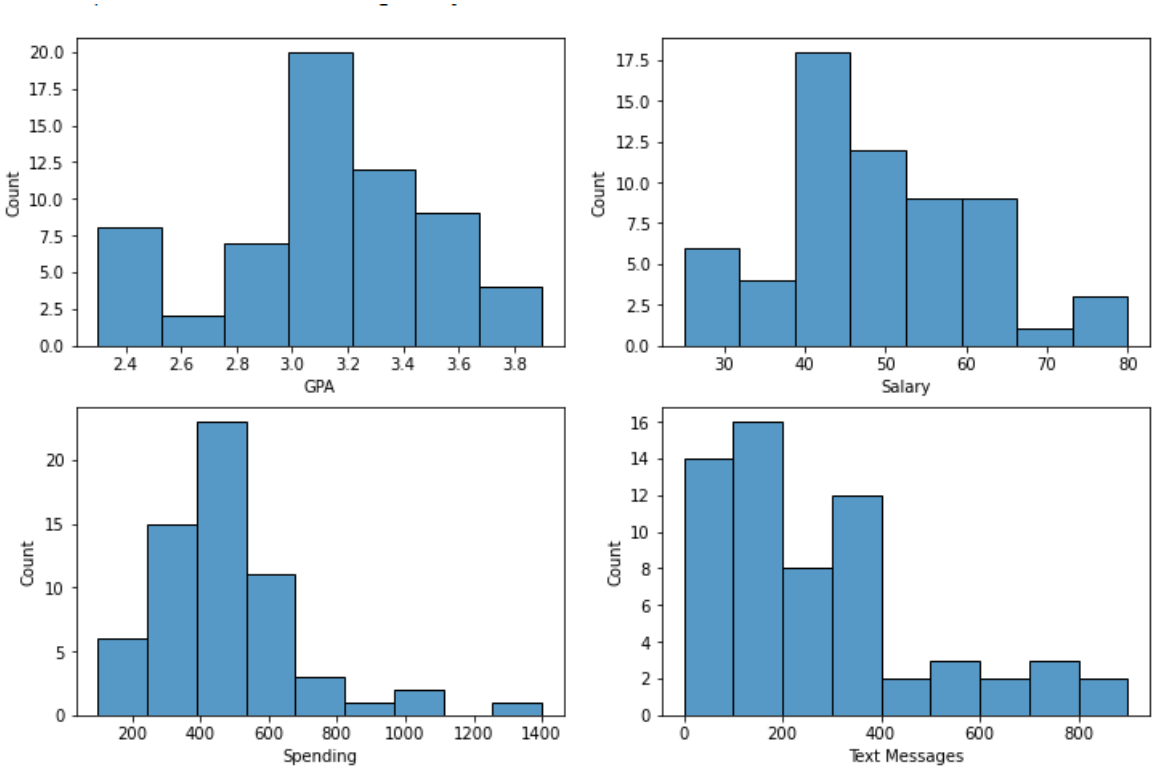
```
Probability of who earn more than or equal to 50 male students is 32.2%
Probability of who earn more than or equal to 50 female students is 35.7%
```

2.8 Problem 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.

Solution:

Used plt. histplot to know the normal distribution of these four numerical (continuous) variables in the data set –GPA, Salary, Spending and Text Messages

And to confirm whether these four data sets are following normal distribution or not, we done the Shapiro Wilk test and the output from Python we got.



(Fig6:- plot)

- Shapiro Result (statistic=0.953252375125885, p-value=0.09815297275781631)
- Shapiro Result (statistic=0.9689891934394836, p-value=0.33416980504989624)
- Shapiro Result (statistic=0.8724251985549927, p-value=0.00033097428968176246)
- Shapiro Result (statistic=0.8824034929275513, p-value=0.0006114590214565396)

By these details we confirm that out of the given four data sets ‘GPA’ and ‘Salary’ are following normal distribution whereas other two ‘Spending’ and ‘Text Messages’ are not following the normal distribution.

Problem 3: - Asphalt Shingles Data Analysis

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 is calculated. The company would like to show that the mean moisture content is less than 0.35 pound 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.

Summary-

This business report provides detailed explanation of approach to each problem given in the Assignment an provide relative information with regards to solving the problem.

We imported the 'A & B shingles' dataset in python to analyze the data about the Asphalt Shingles. Below is the details approach and answer.

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.

Solution:

Input -Python Jupyter

```
t_statistic, p_value =  ttest_1samp(Project_3_A,0.35)
print('t_statistic: {0} & p_value: {1} '.format(t_statistic, p_value/2))
```

Output from Python Jupyter

One sample t-test

```
t_statistic: -1.4735046253382782 & p_value: 0.07477633144907513
```

3.1.1 What assumption do i need to check before the test for equality of means is performed?

Since p-value > 0.05, do not reject H0 . 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. p-value = 0.0748. If the population mean moisture content is in fact no 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 .0748.

Input - Python Jupyter

```
t_statistic, p_value =  ttest_1samp(Project_3_B,0.35)
print('t_statistic: {0} & p_value: {1} '.format(t_statistic, p_value/2)
```

Output from Python Jupyter

One sample t test: -

```
t_statistic: -3.1003313069986995 & p_value: 0.0020904774003191813
```

3.1.2 What assumption do you need to check before the test for equality of means is performed?

Since p-value < 0.05, reject H0. 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. p-value = 0.0021. 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 .0021.

3.2 Problem 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?

Solution:

H0: μ(A)= μ(B)
Ha: μ(A)! = μ(B)α = 0.05

Input - Python Jupyter

```
t_statistic,p_value=stats.ttest_ind(Project_3_A,Project_3_B,equal_var=True ,nan_policy='omit')
print("t_statistic={} & pvalue={}".format(round(t_statistic,3),round(p_value,3)))
```

Output from Python Jupyter

```
t_statistic=1.29 & pvalue=0.202
```

3.2.1What assumption do you need to check before the test for equality of means is performed?

As the pvalue > a, do not reject HO; and we can say that population mean for shingles A and B are equal Test Assumption when running a two-sample t-test, the basic assumptions are the distributions of the two populations are normal, and that the variances of the two distributions are the same. If those assumptions are not likely to be met, another testing procedure could be use.