Business Report
Project – SMDM (Oct / Nov 2020)

Student's Name – ----

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).

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.

1 – Wholesale Customer Data Analysis

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.

1.1Problem 1.1.1. Use methods of descriptive statistics to summarize data

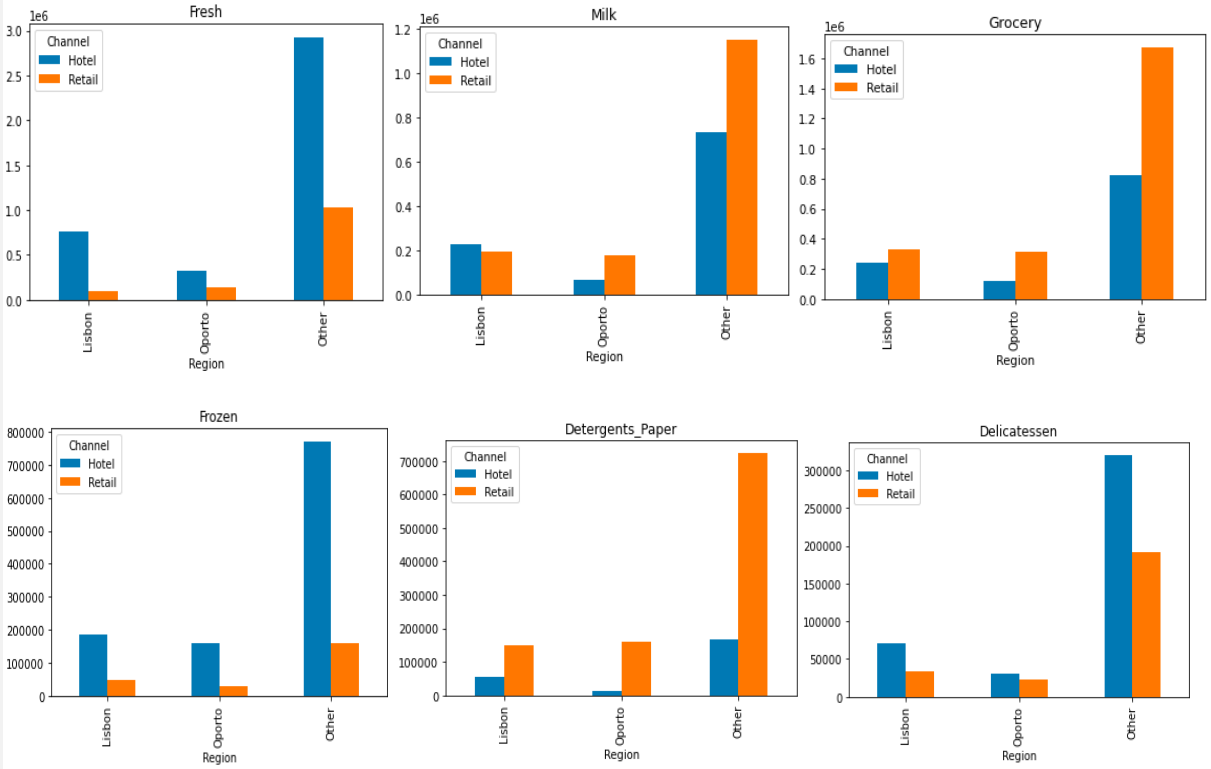
	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Buyer/Spender	440	NaN	NaN	NaN	220.5	127.161	1	110.75	220.5	330.25	440
Channel	440	2	Hotel	298	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Region	440	3	Other	316	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Fresh	440	NaN	NaN	NaN	12000.3	12647.3	3	3127.75	8504	16933.8	112151
Milk	440	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

1.1Problem 1.1.2. & 1.1.3 **Which region and channel *spend most* & *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..

- Hotel channel spend amount is **8070603$** with the highest spend amount and,
- Retail spend amount **6645917$** has least spend amount based on Channel.



Below is the output from Python
Channel
Hotel 8070603
Retail 6645917
dtype: int64

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

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.

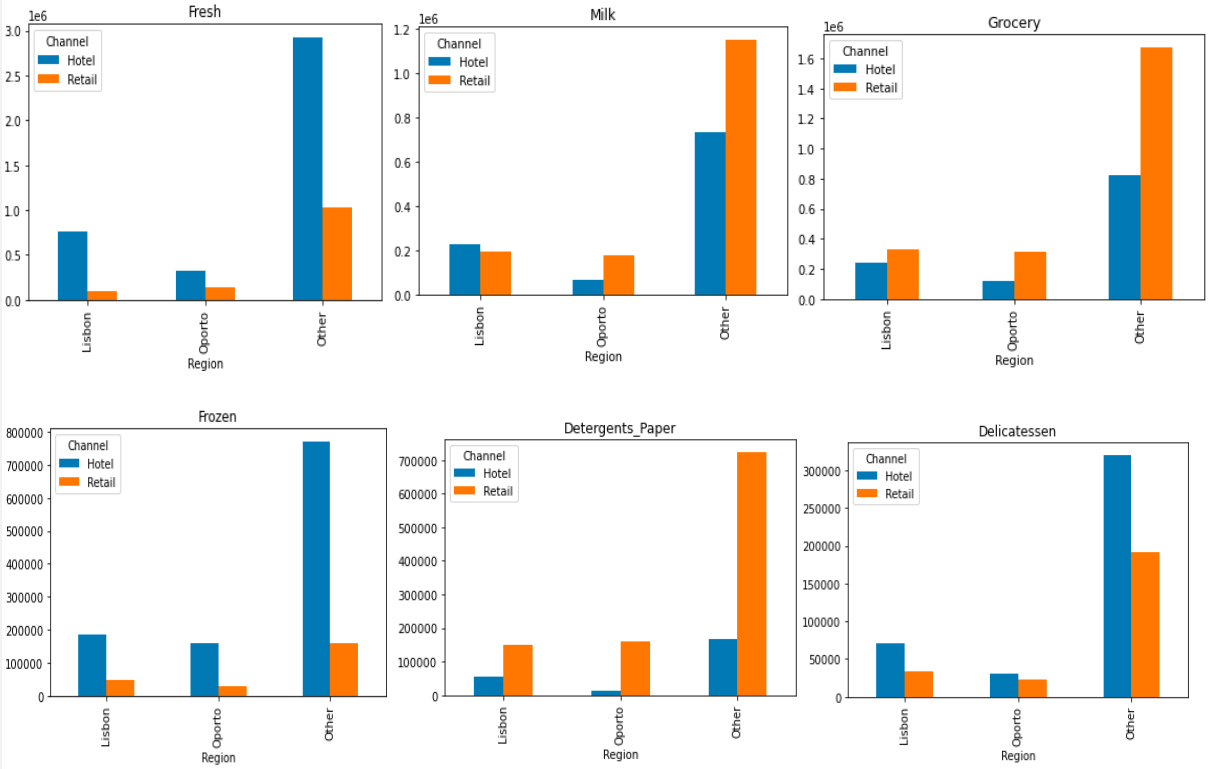
Below is the output from Python –

Region	
Lisbon	2404908
Oporto	1569987
Other	10741625

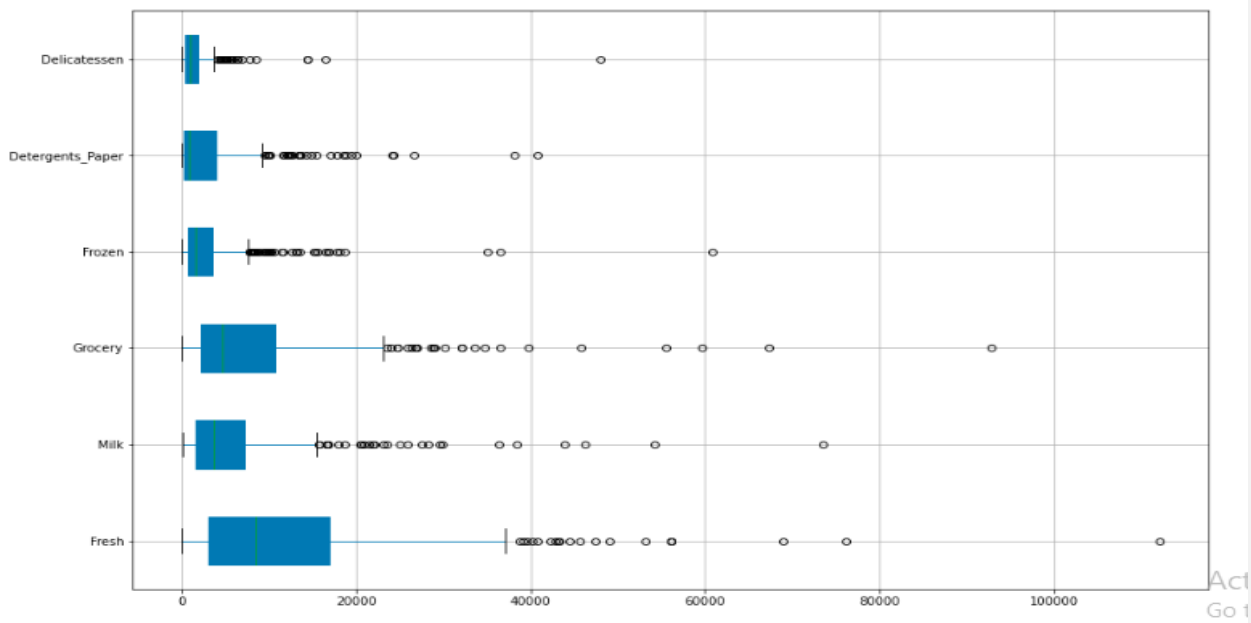
dtype: int64

1.2Problem 1.2 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 -



Looking at the above tables, we see that some categories like Milk, Grocery & Detergents_Paperhave higher spend 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 boxplot is below –



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” (1.05) and highest value is of Category “Delicatessen” (1.84)
So from the given data it is clear that most inconsistent behavior shown by item – Delicatessen
And least inconsistent behavior shown by item – Fresh

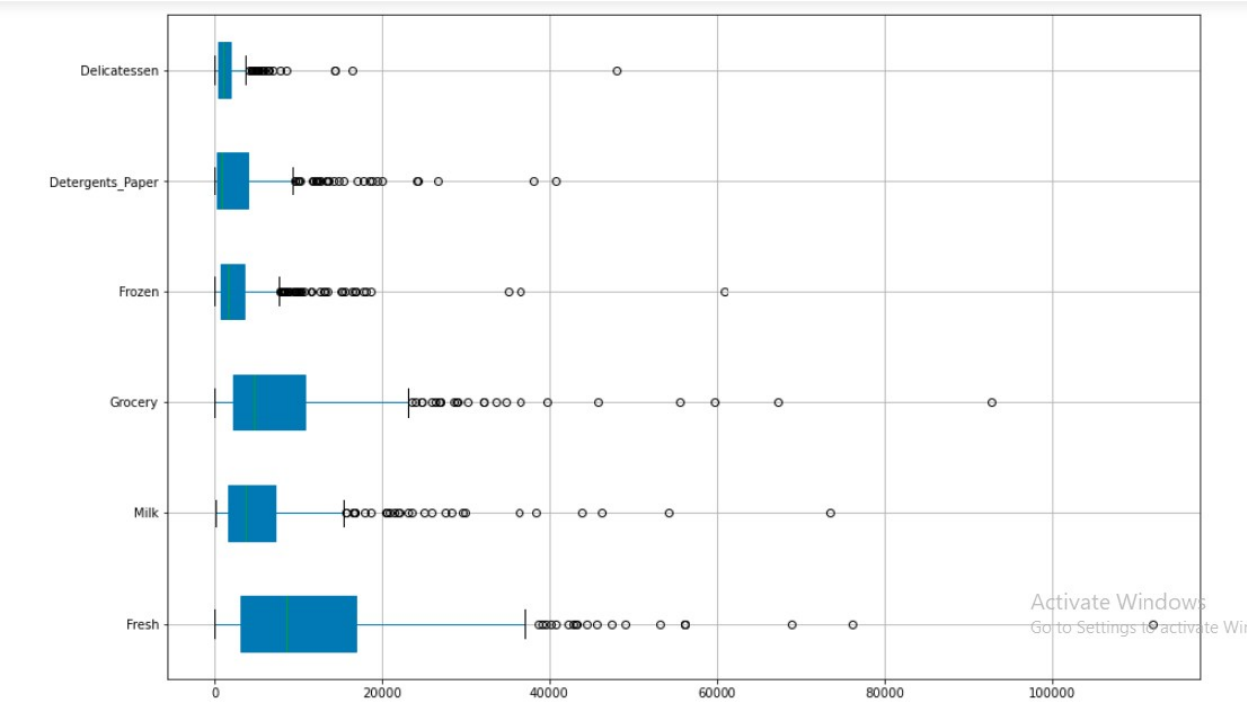
Below is the output from Python –

Coefficient of Variation for Fresh is 1.0539179237473144
Coefficient of Variation for Milk is 1.2732985840065412
Coefficient of Variation for Frozen is 1.5803323836352914
Coefficient of Variation for Grocery is 1.1951743730016822
Coefficient of Variation for Detergents_Paper is 1.654647138500516
Coefficient of Variation for Delicatessen is 1.849406898115838

1.4 Problem Are there any outliers in the data?

Solution:

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



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 the analysis, I find out that 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. And also spent should equal for different regions. Need to focus on other items also than “Fresh” and “Grocery”

*****

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.

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.

2 – CMSU Survey Data Analysis

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.

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

2.1Problem 2.1.1 Gender and Major

Solution:

Below is the output from Python

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

2.1Problem 2.1.2 Gender and Grad Intention

Solution:

Below is the output from Python

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

2.1Problem 2.1.3 Gender and Employment

Solution:

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

2.1Problem 2.1.4 Gender and Computer

Solution:

Below is the output from Python

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

2.2Problem. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

2.2Problem 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 46.77% student will be male in CMSU if randomly selected

2.2Problem 2.2.2 What is the probability that a randomly selected CMSU student will be female?

Solution:

For this we need to find out total female students out of whole student from the given data. After calculation we got the result that probability of 53.23% student will be female in CMSU if randomly selected

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

2.3 Problem 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 13.79%
Probability of Males opting for CIS. is 3.45%
Probability of Males opting for Economics/Finance. is 13.79%
Probability of Males opting for InternationalBusiness. is 6.90%
Probability of Males opting for Management. is 20.69%
Probability of Males opting for Other. is 13.79%
Probability of Males opting for Retailing/Marketing. is 17.24%
Probability of Males opting for Undecided. is 10.34%

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

2.3 Problem 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 females opting for different majors

Below is the output from Python –

Probability of Females opting for Accounting. is 9.09%
Probability of Females opting for CIS. is 9.09%
Probability of Females opting for Economics/Finance. is 21.21%
Probability of Females opting for InternationalBusiness. is 12.12%
Probability of Females opting for Management. is 12.12%
Probability of Females opting for Other. is 9.09%
Probability of Females opting for Retailing/Marketing. is 27.27%
Probability of Females opting for Undecided. is 0.00%

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

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

2.4Problem 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 and intends to be Graduate. is 58.62%

2.4Problem 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 randomly selected student is a Female and does NOT have a laptop. is 13.79%

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

2.5Problem 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 randomly chosen student is either Male or has full time employment. is 74.19%

2.5Problem 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 - Probability that given a female student is randomly chosen, she is majoring in international business or management is 24.24%

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:

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

2.7Problem 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 GPA less than 3

And post calculation we find out that - Probability that student is chosen randomly and that his/her GPA is less than 3 is 22.58%

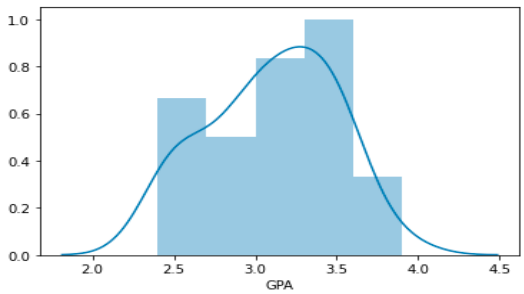
2.7Problem 2.7.2 Find the conditional probability that a randomly selected male earns 50 or more. Find the conditional probability that a randomly selected female earns 50 or more.

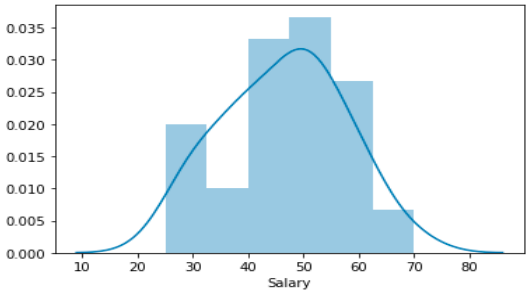
Solution:
Using contingency tables of Gender and Salary we got the total numbers of Male and Female and number of male and female earning 50 or more

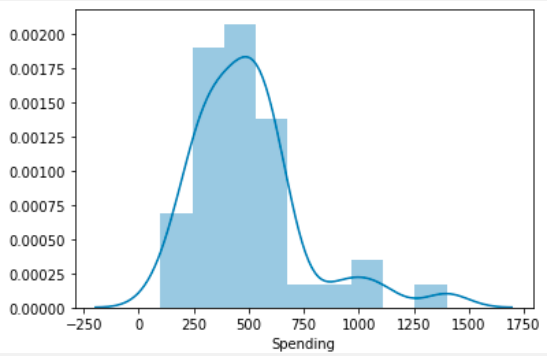
And post calculation we find out that - Probability that randomly selected male earns 50 or more is 34.48%
And Probability that randomly selected female earns 50 or more is 30.3%

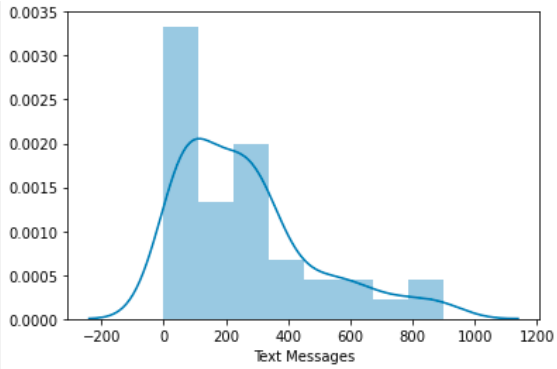
2.8 Problem Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarizing your conclusions.

Solution:
Used distplot 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 –

ShapiroResult(statistic=0.953252375125885, pvalue=0.09815297275781631)
ShapiroResult(statistic=0.9689891934394836, pvalue=0.33416980504989624)
ShapiroResult(statistic=0.8724251985549927, pvalue=0.00033097428968176246)
ShapiroResult(statistic=0.8824034929275513, pvalue=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 -

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 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 and provides relative information with regards to solving the problem.

3 – Asphalt Shingles Data Analysis

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

3.1Problem 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.

Solution:

Input - Python Jupyter

```
t_statistic, p_value = ttest_1samp(df.A, 0.35)
print('One sample t test \nt 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
```

Since pvalue > 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(df.B, 0.35,nan_policy='omit' )
print('One sample t test \nt 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: 0.0020904774003191826
```

Since pvalue < 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=ttest_ind(df['A'],df['B'],equal_var=True ,nan_policy='omit')
print("t_statistic={} and pvalue={}".format(round(t_statistic,3),round(p_value,3)))
```

Output from Python Jupyter

t_statistic=1.29 and pvalue=0.202

As the pvalue > α , do not reject H0; and we can say that population mean for shingles A and B are equal Test Assumptions When running a two-sample t-test, the basic assumptions are that 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.

*****