**SMDM Business Report Project**

**Statistical Methods for Decision Making**

**Student Name – Amrita Jena**

**Course- PGP-DSBA**

**Batch - PGPDSBA Online May\_B 2021**

**INDEX**

**1-Wholesale Customers Analysis :**

* **Problem 1-**

**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?**

**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.**

**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?**

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

**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**

**2. The Student News Service at Clear Mountain State University (CMSU) (Survey) Analysis:**

* **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 (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**

**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?**

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

**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.**

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

**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.**

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

**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?**

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

**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?**

**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?**

**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.**

**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.**

**3. A and B Shingles Hypothesis Test Analysis:**

* **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 pounds per 100 square feet.**

**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.**

**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?**

**Problem 1-**

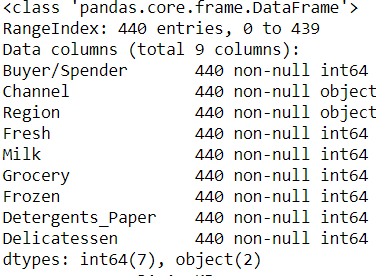
**Wholesale Customers Analysis**

**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).**

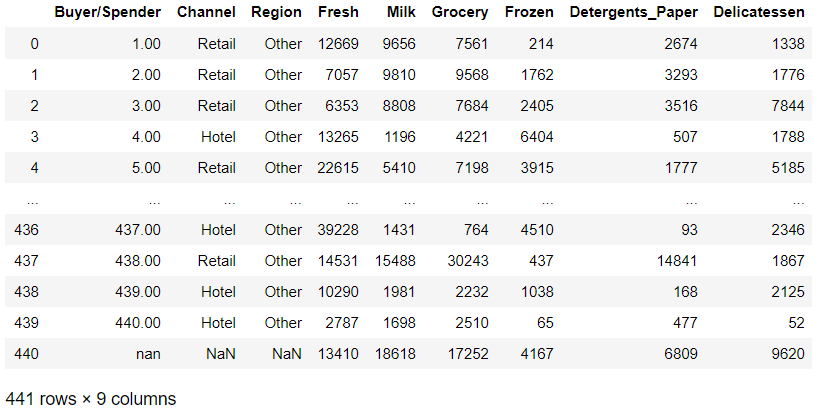
**1.1** **Use methods of descriptive statistics to summarize data. Which Region and which Channel spent the most? Which Region and which Channel spent the least?**

The given data has 440 rows and 9 columns with no null values.





By using describe function we could infer that:



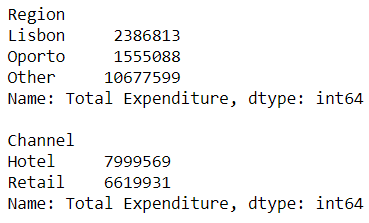
Out of 9 columns 6 are the category of products/items across 3 regions (Lisbon, Oporto, Other) and across 2 different sales channel (Hotel, Retail ).

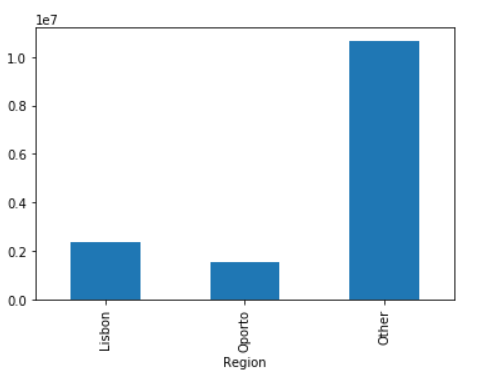
Where Fresh,Milk,Grocery,Frozen,Detergents\_Paper,Delicatessen are continuous in nature and Channel ,Region are nominal in nature. Among Channels, the most frequently occurring, 298 0ut of 440 is ‘Hotel’ and among Region, the most frequently occurring, 316 out of 440 is ‘Other’.

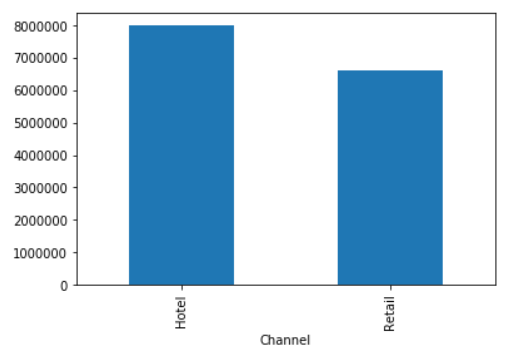
By creating a new column of ‘Total Expenditure’ and plotting bar graph with Region and Channel we can identify region as well as channel with maximum spend and minimum spend. Below is the bar graph representation

Hotel spend amount is highest with **7999569 PTE** (Portuguese Escudo i.e Portuguese currency) spend amount and Retail spend amount is the least with **6619931 PTE** spend amount based on Channel.

Other region spend amount is highest with the highest spend amount of **10677599 PTE** and Oporto is the least spending region with lowest spend amount of **1555088 PTE.**







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.

By using describe () we could infer that:

Fresh has a mean of 12000.3, standard deviation of 12633.13, with min value of 3 and max value of 112151.The quartile ranges are Q1(25%) is 3136.00, Q2(50%) is 8533 and Q3(75%) is 16933.00 with IQR = Q3-Q1 = 16933.8-3136.00 = 13,797.8

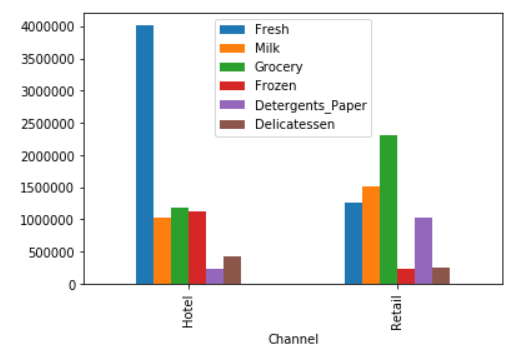
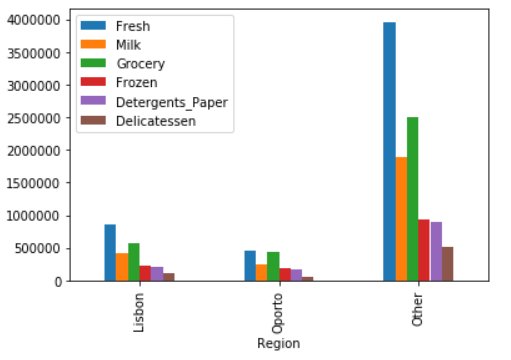
Milk has a mean of 5825.34, standard deviation of 7397.23, with min value of 55 and max value of 73498.The quartile ranges are Q1(25%) is 1534.00, Q2(50%) 3634.00 and Q3(75%) is 7209.00 with IQR = Q3-Q1 = 7209.00-1534.00 = 5,675.00

Grocery has a mean of 7972.37, standard deviation of 9502.68, with min value of 3 and max value of 92780.The quartile ranges are Q1(25%) is 2155.00, Q2(50%) 4757.00 and Q3(75%) is 10685.00 with IQR = Q3-Q1 = 10685.00-2155.00 = 8,530.00

Frozen has a mean of 3074.41, standard deviation of 4849.43, with min value of 25 and max value of 60869.The quartile ranges are Q1(25%) is 744.00, Q2(50%) 1535.00 and Q3(75%) is 3570.00 with IQR = Q3-Q1 = 3570.00-744.00 = 2,826.00

Detergents Paper has a mean of 2890.40, standard deviation of 4766.10, with min value of 3 and max value of 40827.The quartile ranges are Q1(25%) is 257.00, Q2(50%) is 820.00 and Q3(75%) is 3961.00 with IQR = Q3-Q1 = 3961.00-257.00 = 3,704.00

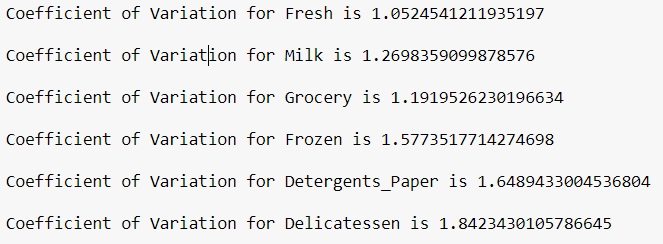
Delicatessen has a mean of 1543.2, standard deviation of 2843.15, with min value of 3 and max value of 47943.The quartile ranges are Q1(25%) is 409.00, Q2(50%) is 967.00 and Q3(75%) is 1824.00 with IQR = Q3-Q1 = 1824.00-409.00 = 1,415.00



From the above bar plots we could infer that category like Fresh, Frozen, Grocery have higher spend in Hotel and Grocery, Milk, Fresh, Detergents Paper have higher spend in the Retail channel across regions. On the other hand, Fresh and Grocery have higher consumption in the Hotel channel versus Retail, across all regions.

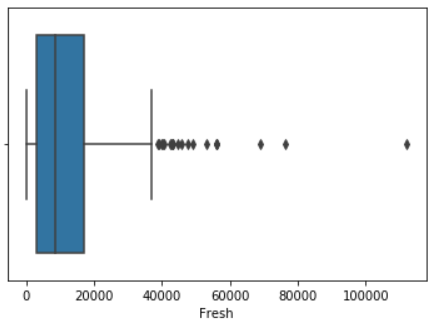
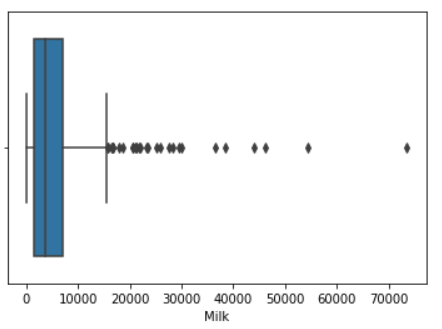
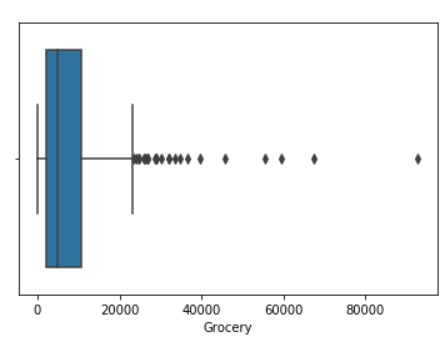
**1.3 On the basis of a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?**

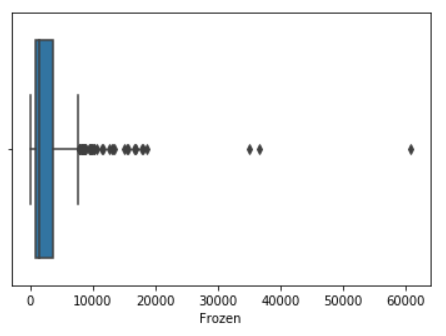
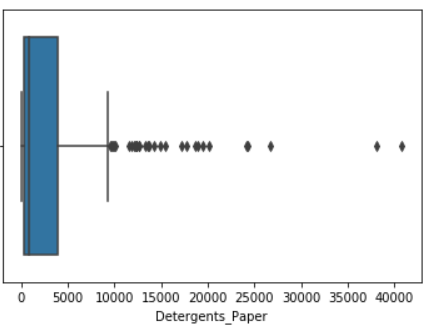
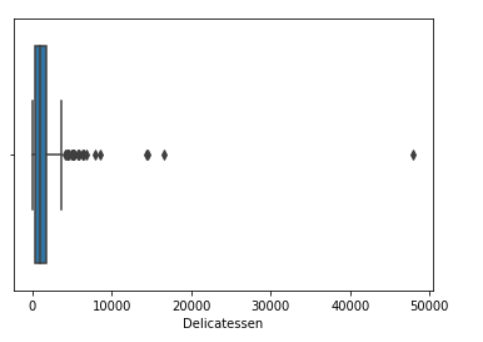
By applying Coefficient of Variation, we could infer which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour.



Using Coefficient of Variation, we observe that the least value is of “Fresh” (1.05) and highest value is of “Delicatessen” (1.84). So, by observing the values of Coefficient of Variation we can say that the most inconsistent behaviour is shown by Delicatessen and the least inconsistent behaviour is shown by Fresh.

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

By observing the boxplots, we could infer that outliers are present in all data.

**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**

As per the data analysis, I could state that there are inconsistencies in spending of different items (by calculating Coefficient of Variation), which should be minimized as it may affect the business flow. The spending scenario of Hotel and Retail channel are varying which should be uniform or more or less equal for unwrinkled flow of the retail business. Effort must be taken to spend more in Milk , Frozen , Detergent Paper and Delicatessen as they are the least spend items across the region.

|  |
| --- |

**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 (stored in the Survey data set)**

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

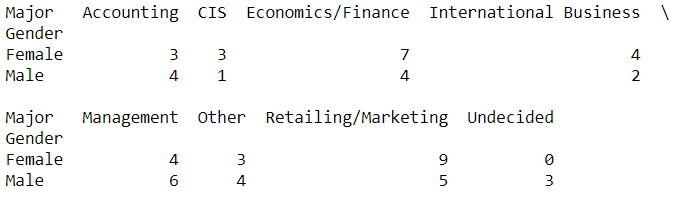
**2.1.1. Gender and Major**

**2.1.2. Gender and Grad Intention**

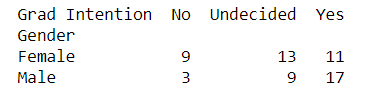
**2.1.3. Gender and Employment**

**2.1.4. Gender and Computer**

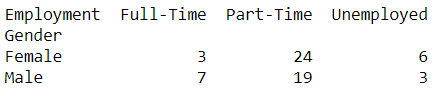
2.1.1- Answer-



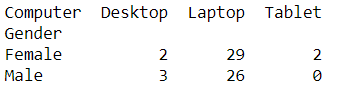
2.1.2-Answer-



2.1.3-Answer-



2.1.4-Answer-



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 selected randomly.

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

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 **53.2%** student will be female in CMSU if selected randomly.

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.

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

Below is the output from Python –

Total male opting Accounting major. is **13.8%**

Total male opting CIS major. is **3.5%**

Total male opting Economics/Finance major. is **13.8%**

Total male opting International Business. is **6.9%**

Total male opting Management.is **20.7%**

Total male opting Others.is **13.8%**

Total male opting Retailing/Marketing is **17.2%**

Total male still Undecided.is **10.3%**

From the above result we can infer that most of the males 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 of CMSU.

Using contingency tables of Gender and Majors we will get the total number of females and number of females opting for different majors .

Below is the output from Python –

Total male opting Accounting major. is **9**.1%

Total male opting CIS major. is **9.1%**

Total male opting Economics/Finance major. is **21.2%**

Total male opting International Business. is **12.1%**

Total male opting Management.is **12.1%**

Total male opting Others.is **9.1%**

Total male opting Retailing/Marketing is **27.3%**

Total male still Undecided.is **0.0%**

From the above result we can infer that most of the females students prefer Retailing/Marketing as Majors .

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.

Solution:

Using contingency tables of Gender and Grad Intention we will get the total numbers of males and number of

males intend to be graduate and after calculation we find out that - Probability of Males and intends to be Graduate. is **27.4%**

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 will get the total numbers of females and number of

females do not have a laptop and after calculation we find out that , probability of randomly selected student is a Female and does NOT have a laptop is **6.5%.**

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?

Solution:

Using contingency tables of Gender and Employment we will get the total number of males and number of males

who are full time employed and after calculation we find out that the Probability of randomly chosen student is either Male or has full time employment. is **51.6%**

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

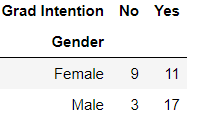
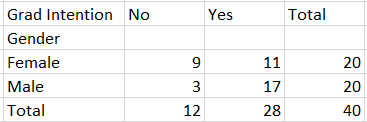
Solution:

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 after calculation we find out that the Probability that given a female student is randomly chosen, she is majoring in international business or management is **24.2%**

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?

Solution:

The Probability that a randomly selected student ‘being female’ and The Probability that a randomly selected student the graduate intention and being female are as follows:

P(Grad Intention Yes) = **28/40 = 0.7**

P(Grad Intention Yes | female) = **11 / 20 = 0.55**

These probabilities are not equal and one event doesn't change the probability of the occurrence of the other

Event so we can tell that the two events are 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?

Using contingency tables of Gender and GPA we will get the total number of students and number of students

GPA less than 3 and after calculation we get that the - Probability that student is chosen randomly and that his/

her GPA is less than 3 is **27.4%**

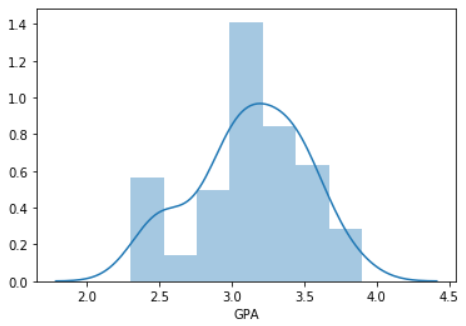
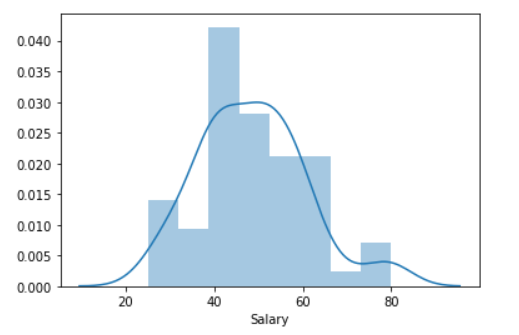
2.7.2. Find the conditional probability that a randomly selected male earns 50 or more. Find the conditional probability that a randomly selected female earns 50 or more.

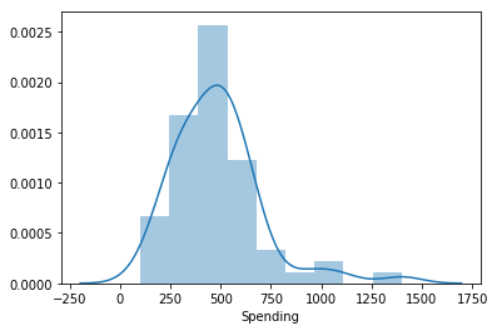
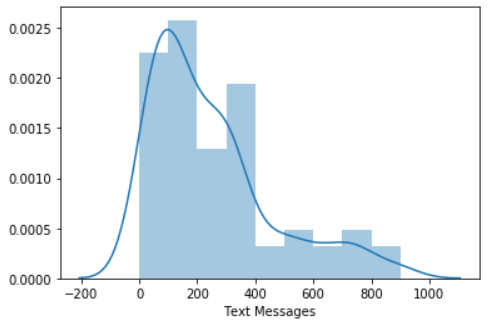
Using contingency tables of Gender and Salary we will get the total numbers of Male and Female and number of male and female earning 50 or more And and after calculation we get that the - Probability that randomly selected male earns 50 or more is **48.3%** and Probability that randomly selected female earns 50 or more is **54.5%.**

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.



From the above plot we can infer that GPA and Salary are normally distributed as both the data are forming bell shape where as in Spending, Text messages it could be observed that the variables are moderately skewed

towards right.

Problem 3 (Download Data)

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.

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 .

From the analysis we could infer that 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 as p-value = 0.0748 (rounding up the p value )

3.1.2-

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 - One sample t test t statistic: **-3.1003313069986995** p value: **0.0020904774003191826**

Since pvalue is < 0.05, we reject H0 .

From the analysis we could infer that 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 as p-value = 0.0021.

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?

Solution:

H0 : μ(A)= μ(B)

Ha : μ(A)!= μ(B)

α = 0.05

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)))

Result- t\_statistic=**1.29** and pvalue=**0.202**

As the pvalue > 0.05 we do not reject the H0 and we can infer that population mean for shingles A and B are equal. When running a two-sample t-test, the basic conditions are that the distributions of the two populations are normal, and that the variances of the two distributions are the same.