SMDM PROJECT REPORT

PGDSBA SEP\_D

DEEPAK SINGH

CONTENTS

Problem1……………………………………………………………………………………………………………………….

**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 behavior? Which items show the least inconsistent behavior?**

**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**

Problem 2………………………………………………………………………………………………………………………..

**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.**

Problem 3……………………………………………………………………………………………………………………

**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?**

**Please reflect on all that you have learnt while working on this project. This step is critical in cementing all your concepts and closing the loop. Please write down your thoughts**[**here**](https://forms.gle/mwnvE7zmZMxu4NrRA)**.**

Problem 1

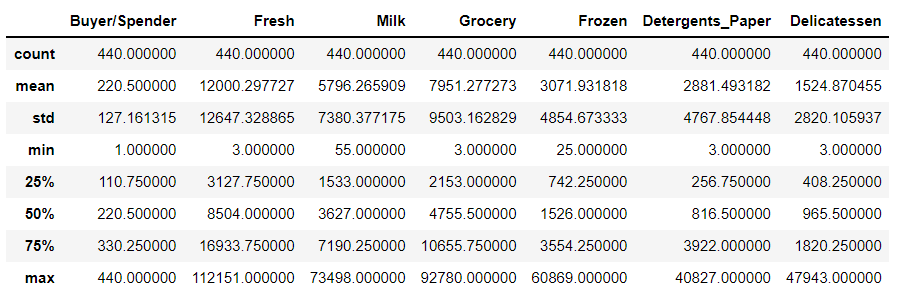
**Wholesale Customers Analysis**

**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).

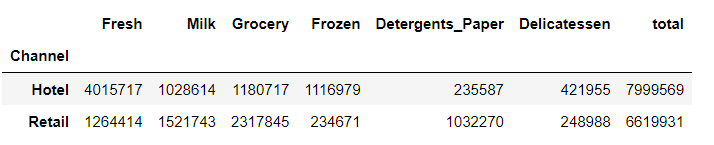
* 1. **Use methods of descriptive statistics to summarize data.**

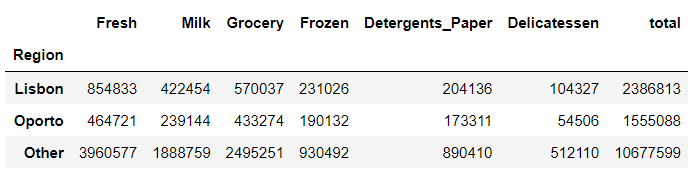
We imported the ‘Wholesale Customer data’ dataset in python to analyze the spend under each store items across regions and channel to find solutions to each problem. Below is the detailed approach and answer.



**1.2 Which Region and which Channel spent the most? Which Region and which Channel spent the least?**

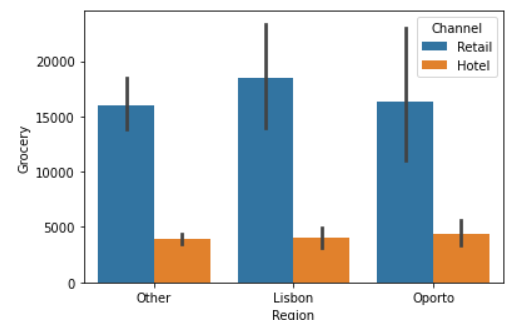
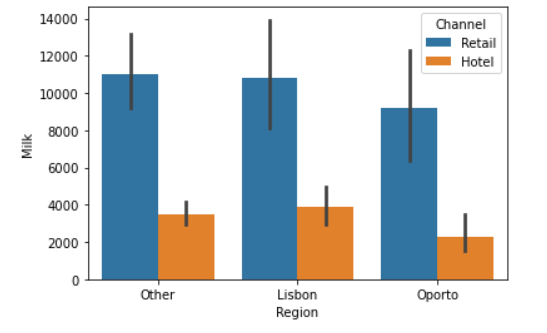
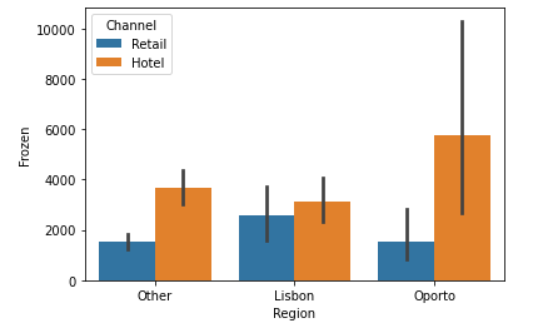
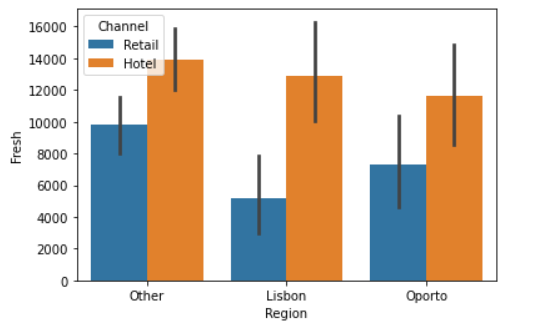
1. Hotel channel spend amount is **7999669** with the highest spend amount and
2. Retail spend amount **6619931** has least spend amount based on Channel.
3. **Other regions** spend amount is **10677599** with the highest spend amount and
4. **Oporto region** spend amount is **1555088** and has least spend amount by region

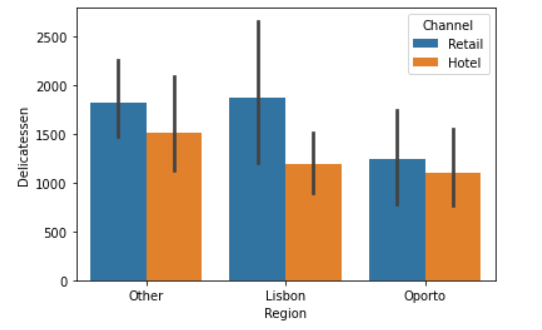
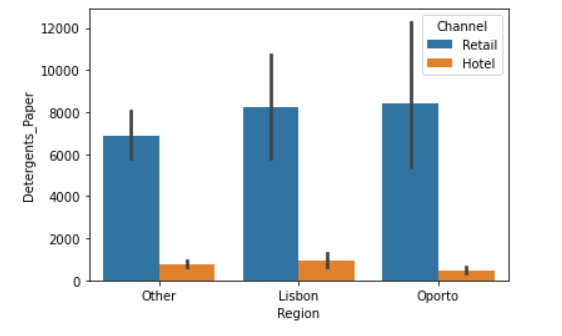




**1.2 There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer**

Looking at the below tables, we see that some categories like Milk, Grocery & Detergents 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.

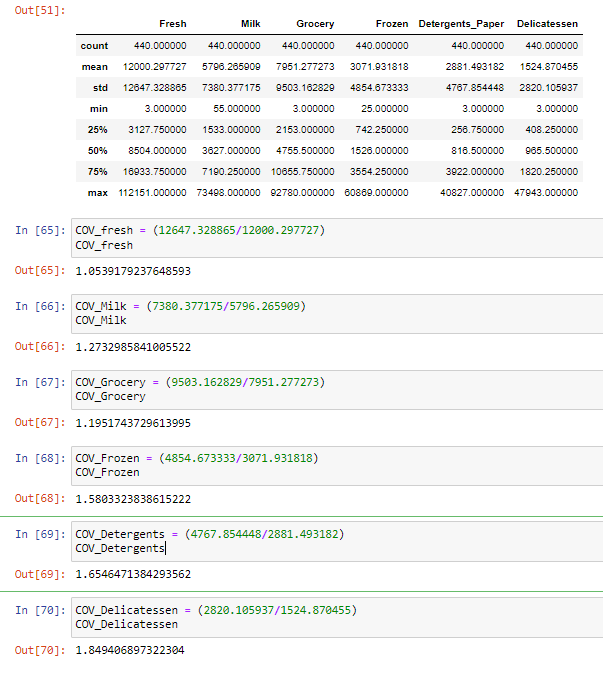




**1.3 On the basis of a descriptive measure of variability, which item shows the most inconsistent behavior? Which items show the least inconsistent behavior?**

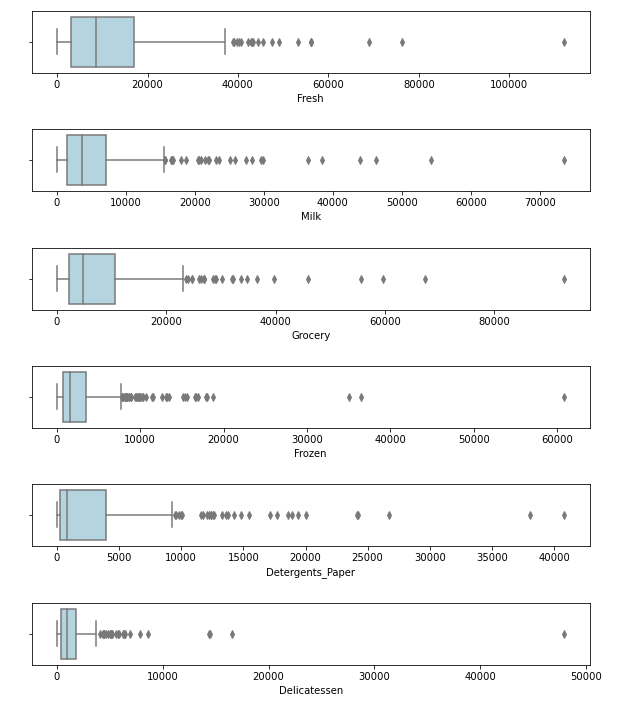
Using coefficient of Variation, we find out the least value is of Category FRESH = 1.05 and highest id of category DELICATESSEN = 1.84. So, from the given dataset it is clear that most inconsistent behavior shown by category is DILICATESSEN and least inconsistent behavior shown by category is FRESH.

Please find below the output code from Python -



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

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



**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 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”

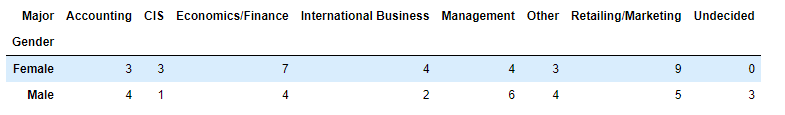
**Problem 2 -**

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

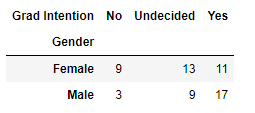
**2.1. For this data, construct the following contingency tables (Keep Gender as row variable)**

**2.1.1. Gender and Major**

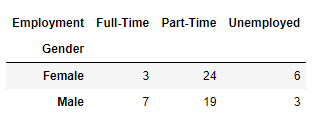
Below is the output code from python.



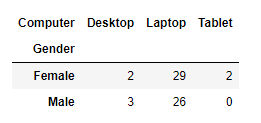
**2.1.2. Gender and Grad Intention**



**2.1.3. Gender and Employment**



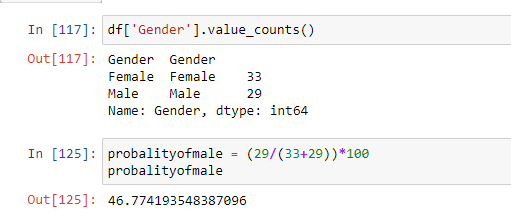
**2.1.4. Gender and Computer**



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

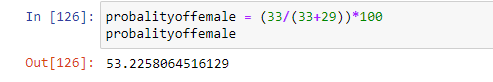
**2.2.1. What is the probability that a randomly selected CMSU student will be male?**

Probability of 46.77% student will be male in CMSU if randomly selected



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

Probability of 53.23% student will be female in CMSU if 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.**

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 International Business. 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 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 of CMSU.**

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 –

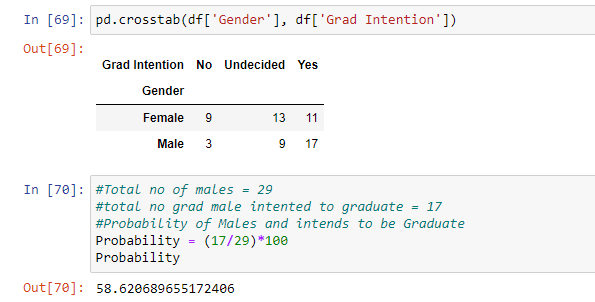
* Probability of Females opting for Accounting. is 9.09%
* Probability of Females opting for CIS. is 9.09%
* Probability of Females opting for Economics/Finance. is 21.21%
* Probability of Females opting for International Business. is 12.12%
* Probability of Females opting for Management. is 12.12%
* 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 female’s 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.**

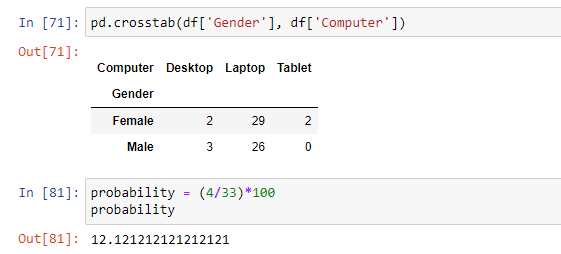
Using contingency tables of Gender and Grad Intention we got the total numbers of males and number of males intends to be graduate. And post calculation we find out that - Probability of Males and intends to be Graduate. is 58.62%



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

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

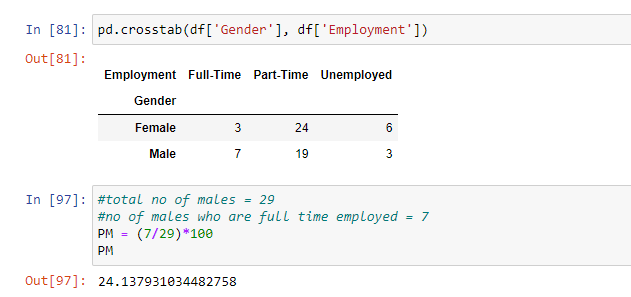
And post calculation we find out that - Probability of randomly selected student is a Female and does NOT have a laptop. is 12.12%



**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?**

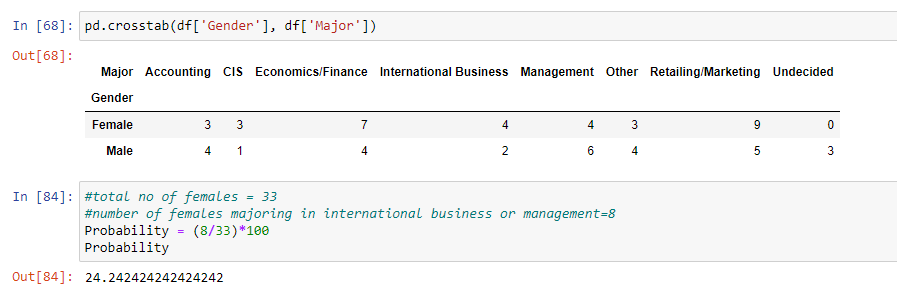
We find out that - Probability of randomly chosen student is either Male or has full time employment. is 24.13%

****

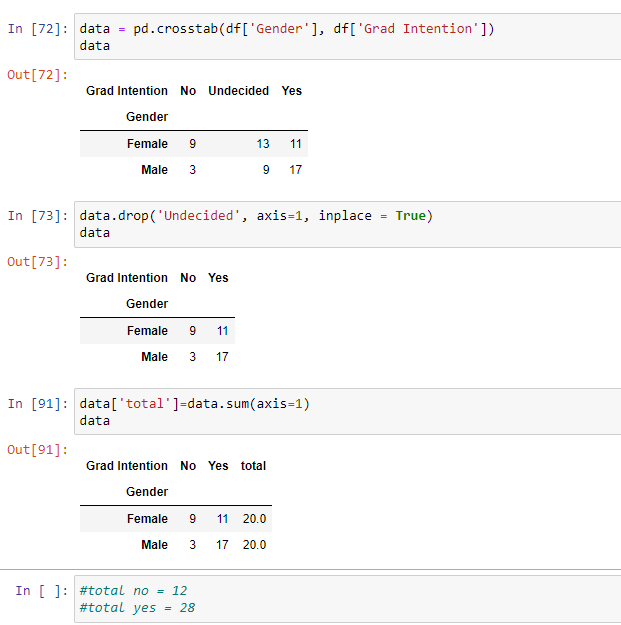
**2.5.2. Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management.**

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%.

Using contingency tables of Gender and Major we got the total numbers of females and number of females 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?**

****

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

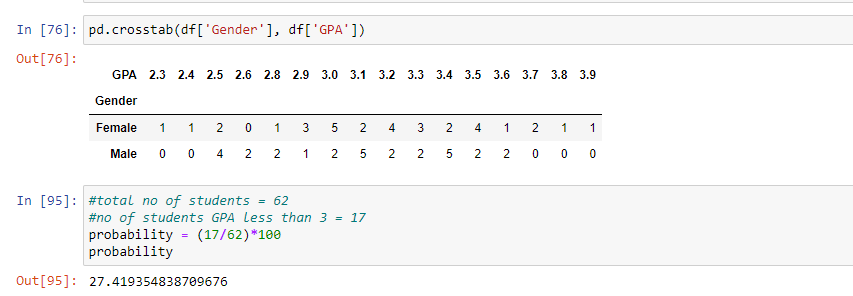
These probabilities are not equal. This suggests that the two events are independent

2.7. 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?

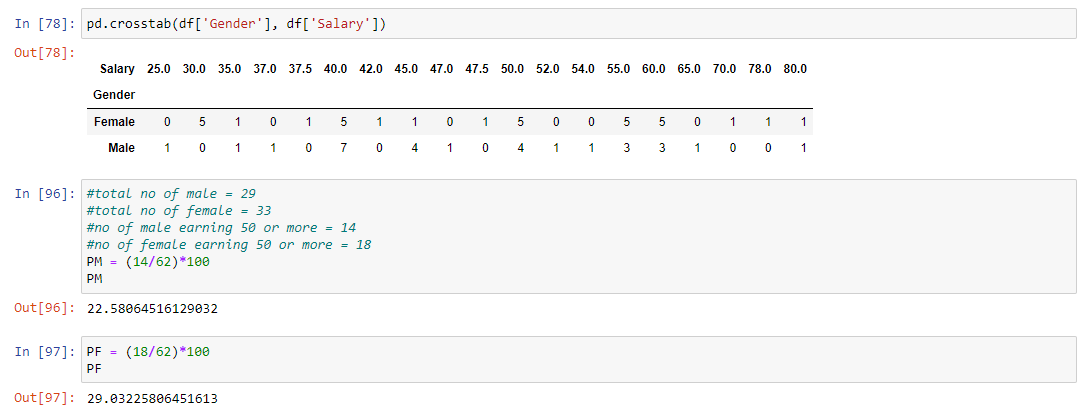
Using contingency tables of Gender and GPA we got the total numbers of students and number of students GPA less than 3. And post calculation we find out that - Probability that student is chosen randomly and that his/her GPA is less than 3 is 27.41%

****

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 got the total numbers of Male and Female and number of male and female earning 50 or more

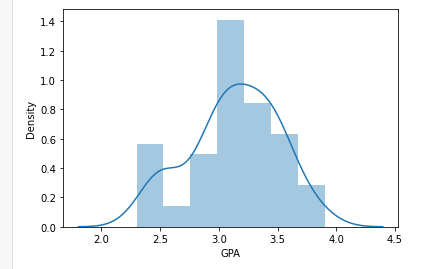
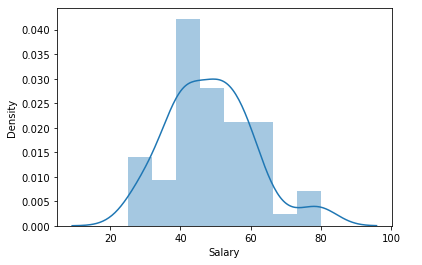
And post calculation we find out that - Probability that randomly selected male earns 50 or more is 22.58% And Probability that randomly selected female earns 50 or more is 29.03%

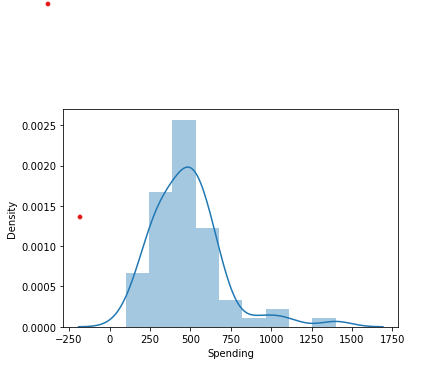
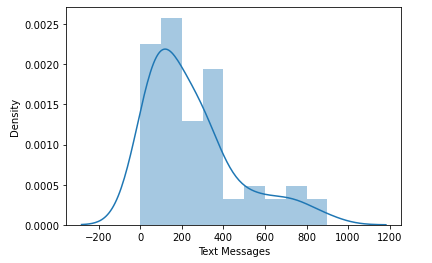
****

**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.**

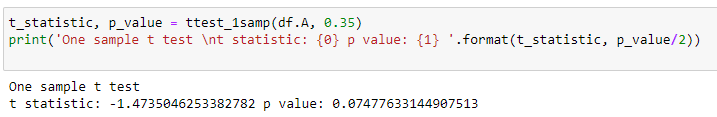
Used distplot to know the normal distribution of these four numerical (continuous) variables in the data set – GPA, Salary, Spending and Text Messages

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

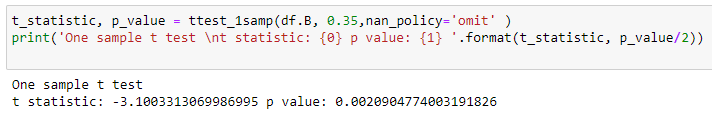
 

**** ****

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.

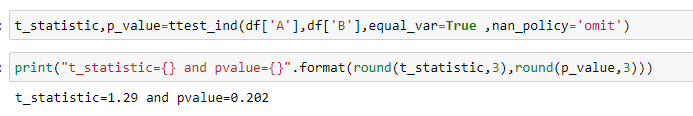
****

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.



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 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?

****

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.