Project 2 - SMDM Project Report

Tribid Maji

Index

1.Wholesale Customers Analysis …………………………………………………………….3

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?........3

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……………….3

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?..........................3

1.4 Are there any outliers in the data?........................................................................................3

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……………………………………………………………………………………..4

-------------------------------------------------------------

2.Survey

**2.1. For this data,construct the following contingency tables (Keep Gender as row variable)……………………………………………………………………………………….4**

**2.1.1. Gender and Major………………………………………………………………………5**

**2.1.2. Gender and Grad Intention……………………………………………………………..5**

**2.1.3. Gender and Employment……………………………………………………………….5**

**2.1.4. Gender and Computer…………………………………………………………………..5**

**2.2. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:………………………………………………………………..5**

**2.2.1. What is the probability that a randomly selected CMSU student will be male?............6**

**2.2.2. What is the probability that a randomly selected CMSU student will be female?.........6**

**2.3. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:……………………………………………………………….7**

**2.3.1. Find the conditional probability of different majors among the male students in CMSU…………………………………………………………………………………………7**

**2.3.2 Find the conditional probability of different majors among the female students of CMSU…………………………………………………………………………………………7**

**2.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question: ………………………………………………………….7**

**2.4.1. Find the probability That a randomly chosen student is a male and intends to graduate………………………………………………………………………………………..8**

**2.4.2 Find the probability that a randomly selected student is a female and does NOT have a laptop. …………………………………………………………………………………………8**

**2.5. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question: ………………………………………………………….……8**

**2.5.1. Find the probability that a randomly chosen student is either a male or has full-time employment?..............................................................................................................................8**

**2.5.2. Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management……………………………………………9**

**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?...................................................9**

**2.7. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages………..…………………………………………………….10**

**2.7.1. If a student is chosen randomly, what is the probability that his/her GPA is less than 3?.............................................................................................................................................10**

**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………………11**

**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**……………………………………11

--------------------------------------------------------------------

3.A & B Shingle……………………………………………………………………………..13

**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…………….14**

**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?.................................................................15**

**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. Which Region and which Channel seems to spend more? Which Region and which Channel seems to spend less?**

**ANS: Region with highest spending is: "OTHER".**

**ANS: Channel with highest spending is: "HOTEL".**

**ANS: Region with lowest spending is: "OPORTO".**

**ANS: Channel with lowest spending is: "RETAIL".**

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

**ANS: all varieties show similar behaviour across Region, "Grocery" spending highest via "Retail" channel.**

**ANS: all varieties show similar behaviour across Region, "Fresh" spending highest via "Hotel" channel.**

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

**ANS: which item shows the most inconsistent behaviour for the item is: Fresh.**

**ANS: Which items show the least inconsistent behaviour for the item is: Delicatessen.**

* 1. **Are there any outliers in the data?**

**ANS:** **all products (Fresh, Milk, Grocery, Frozen, Detergents\_Paper, Delicatessen) has outliers consist.**

* 1. **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**

**ANS: #Based on the provided data I have come up with couple of recommendation, which can help solve our business problem.**

**#the data is based on, in Portugal (Lisbon, Oporto, Other Region) product (Fresh, Milk, Grocery, Frozen, Detergents\_Paper, Delicatessen) consumption or spending by people.**

**#1) Product (consumption) spending is lowest in "Oporto" Region in Portugal.**

**#2) Product (consumption) spending is highest in "Other" Region in Portugal.**

**#3) Product (consumption) spending is lowest via "Retail" channel in Portugal.**

**#4) Product (consumption) spending is highest via "Hotel" channel in Portugal.**

**#5) all varieties show similar behaviour across Region, Grocery spending highest via Retail channel.**

**#6) all varieties show similar behaviour across Region, Fresh spending highest via Hotel channel.**

**#7) the most inconsistent behaviour for the item is: Fresh.**

**#8) the least inconsistent behaviour for the item is: Delicatessen.**

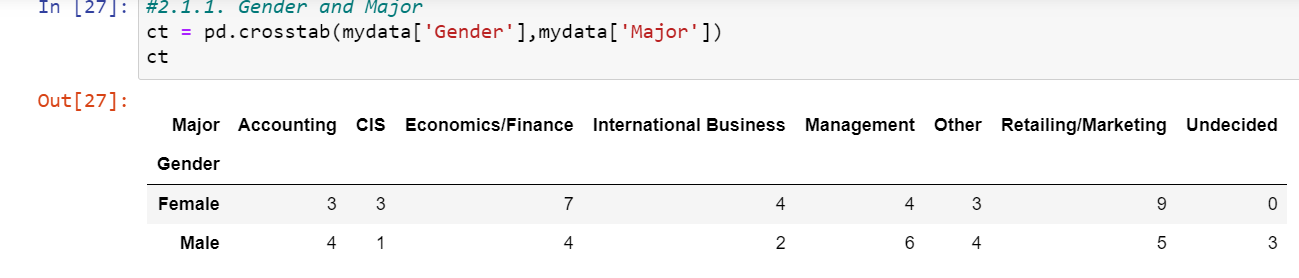
**================================================================**

Problem 2 -

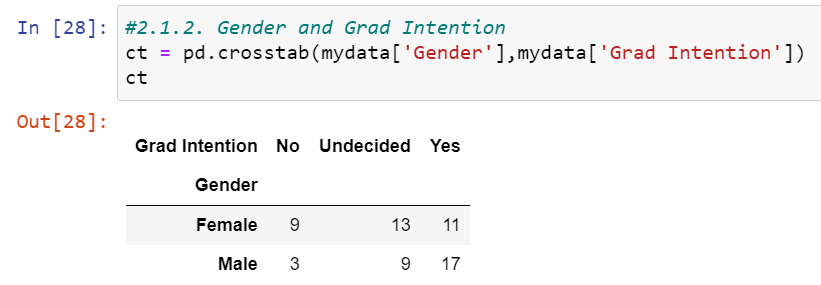
The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. CMSU creates and distributes a survey of 14 questions and receives responses from 62 undergraduates (stored in the **Survey** data set).

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

**2.1.1. Gender and Major**



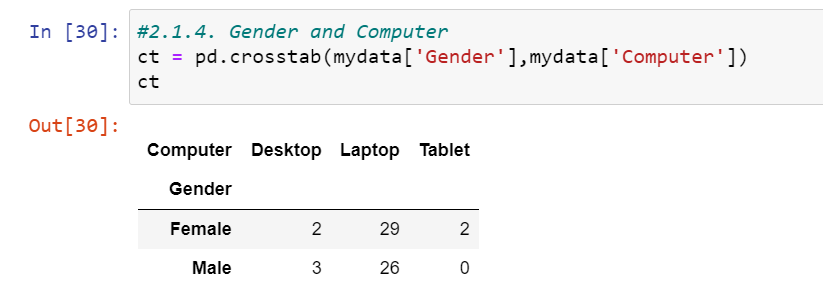
**2.1.2. Gender and Grad Intention**



**2.1.3. Gender and Employment**



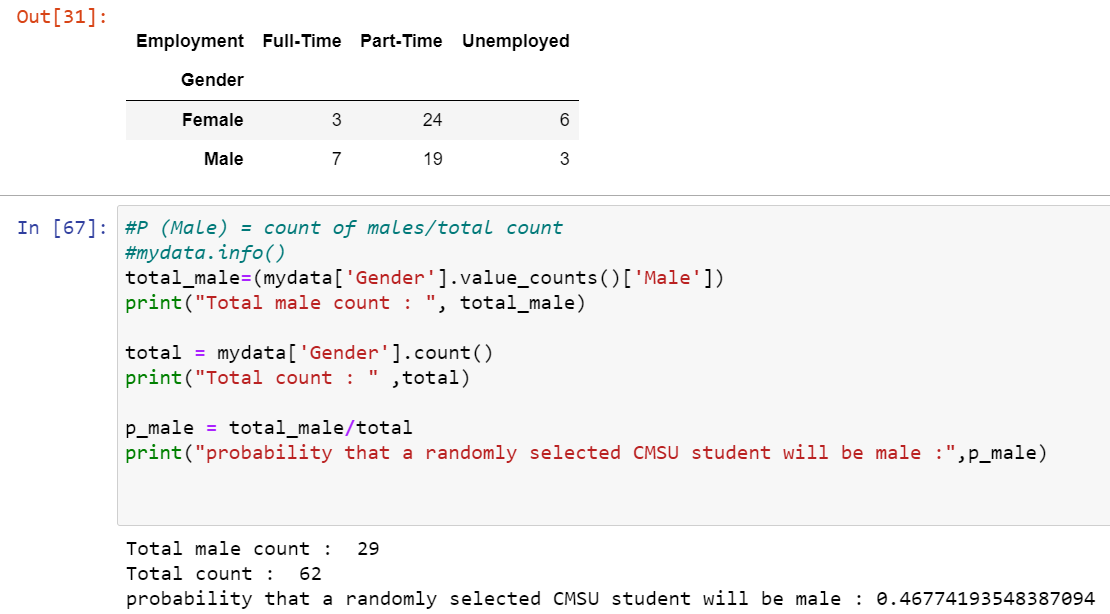
**2.1.4. Gender and Computer**



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

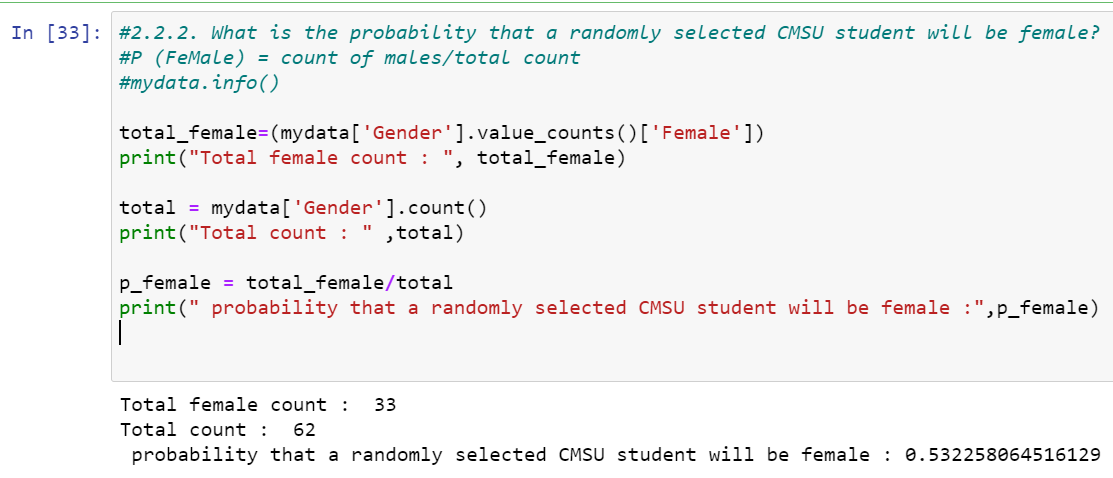
**2.2.1. What is the probability that a randomly selected CMSU student will be male?**

**ANS: Probability that a randomly selected CMSU student will be male: 0.46774193548387094**



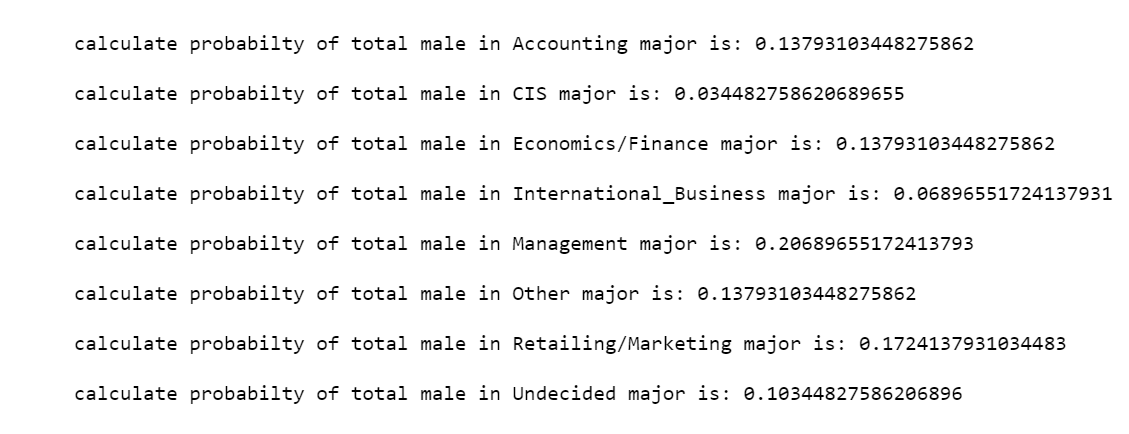
**2.2.2. What is the probability that a randomly selected CMSU student will be female?**

**ANS: Probability that a randomly selected CMSU student will be female: 0.532258064516129**

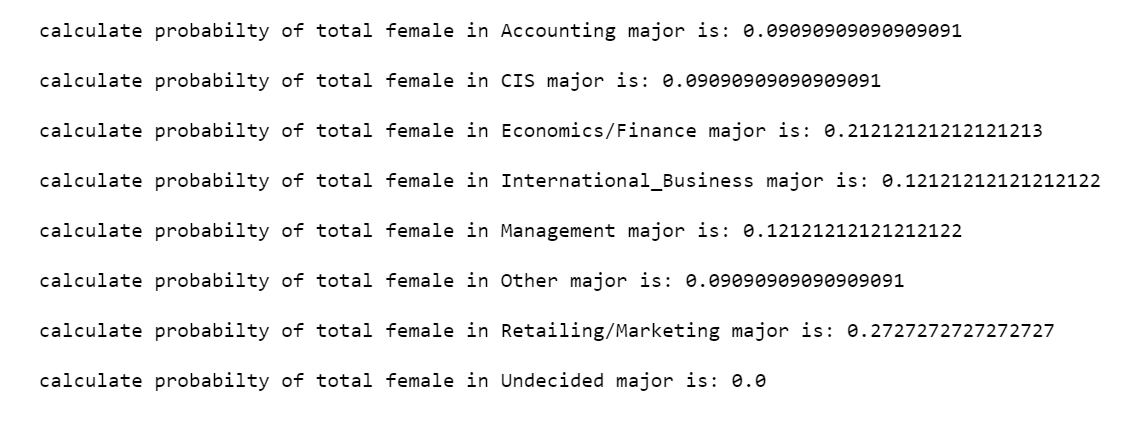


**2.3. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:**

**2.3.1. Find the conditional probability of different majors among the male students in CMSU.**

**ANS: Please find below screenshot.**

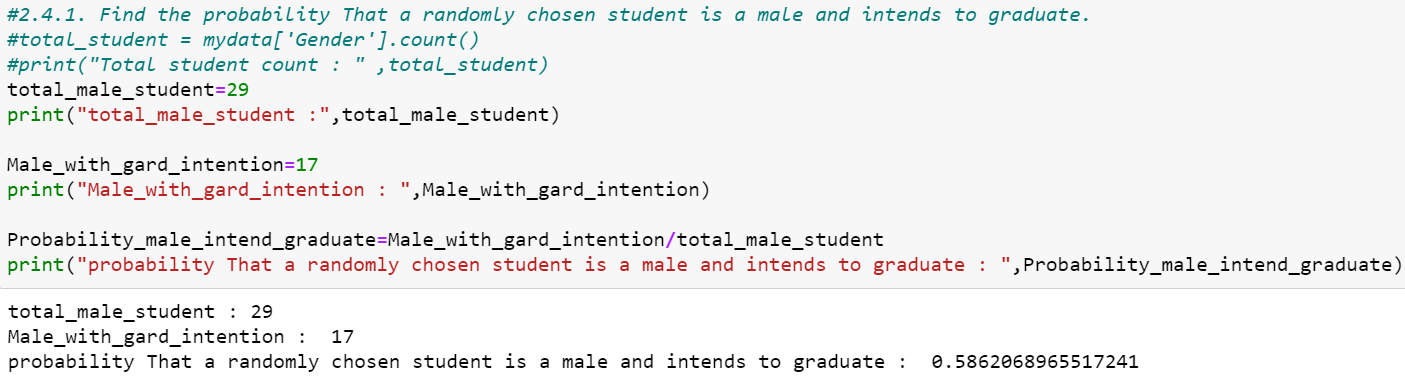
**2.3.2 Find the conditional probability of different majors among the female students of CMSU.**

**ANS: Please find the below screenshot.** 

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

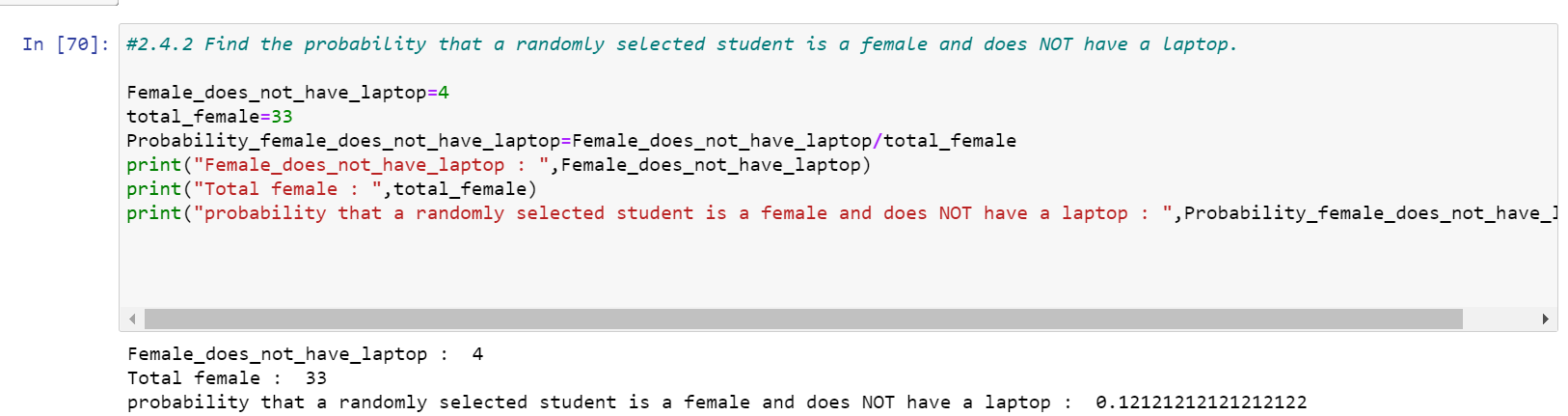
**2.4.1. Find the probability That a randomly chosen student is a male and intends to graduate.**

**ANS: Probability That a randomly chosen student is a male and intends to graduate: 0.5862068965517241**



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

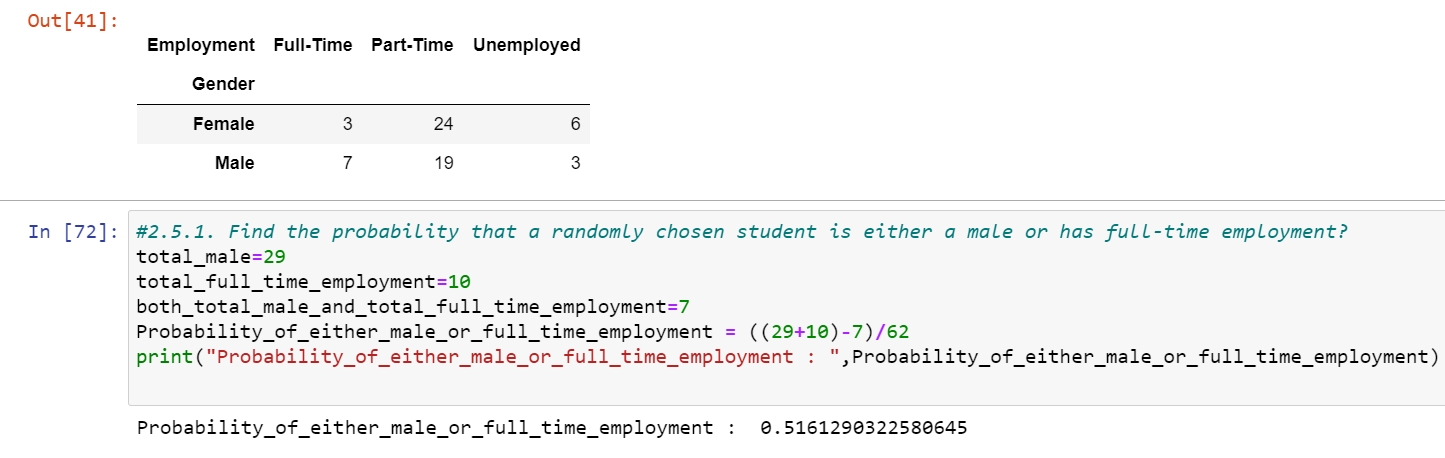
**ANS: Probability that a randomly selected student is a female and does NOT have a laptop: 0.12121212121212122**



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

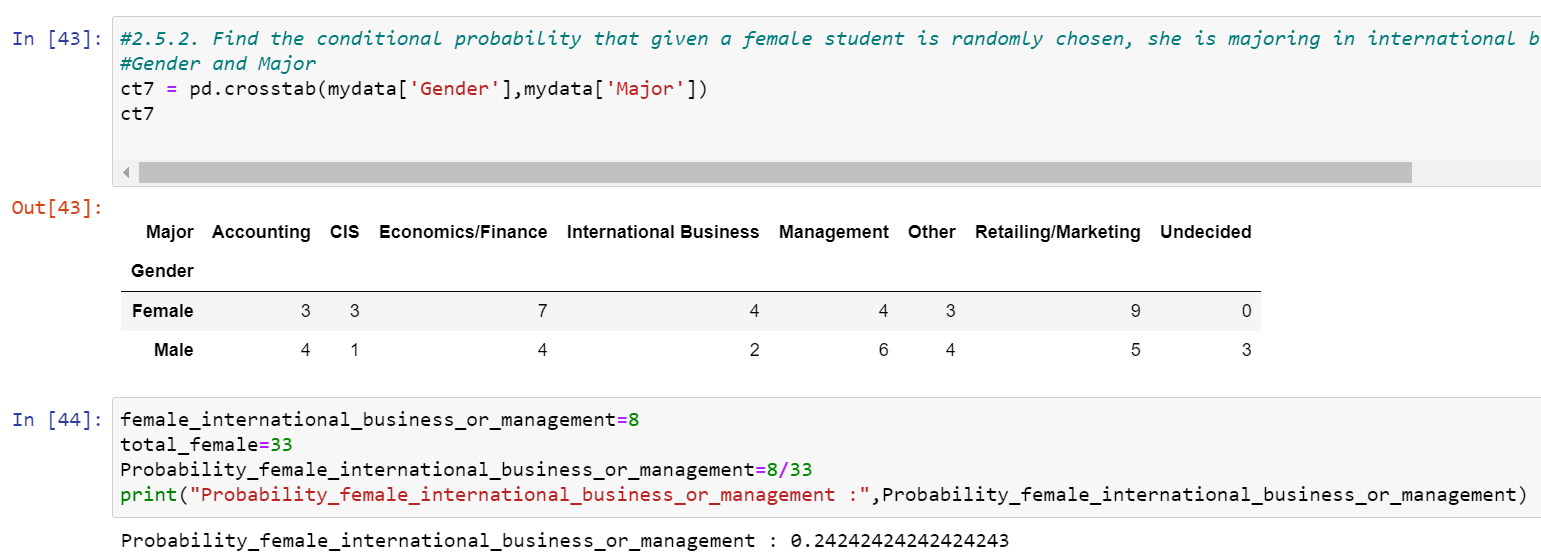
**2.5.1. Find the probability that a randomly chosen student is either a male or has full-time employment?**

**ANS: Probability of either male or full time employment: 0.5161290322580645**



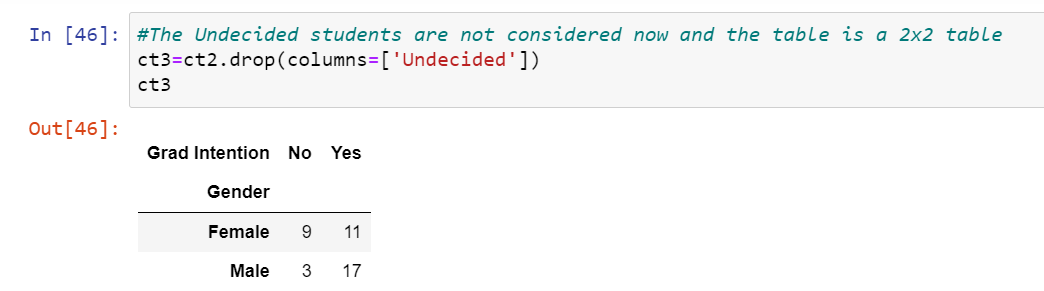
**2.5.2. Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management.**

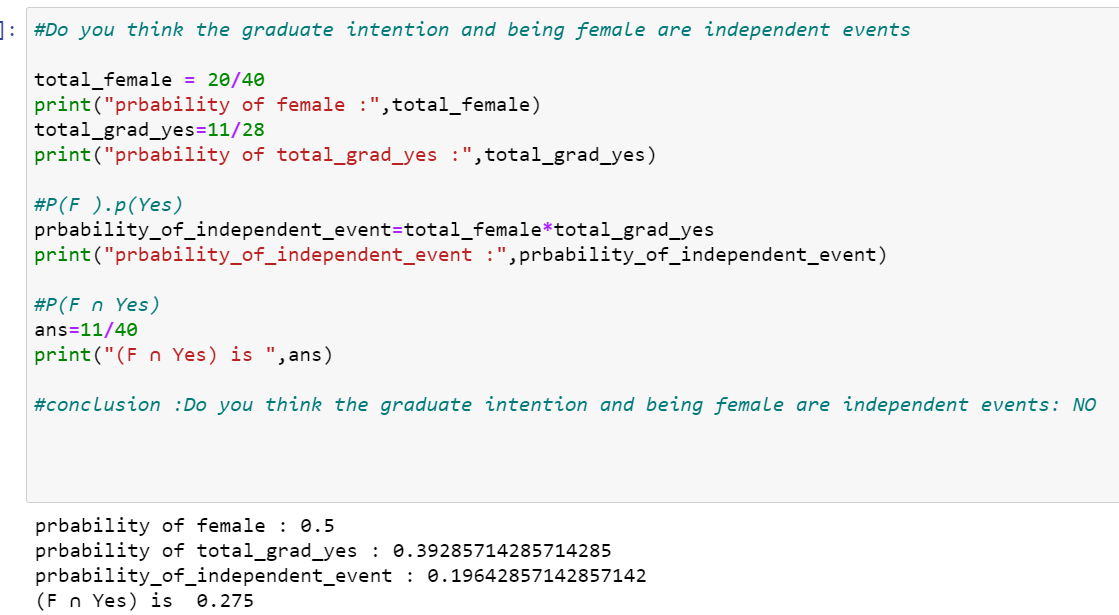
**ANS: Probability female international business or management: 0.24242424242424243**



**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?**

**ANS: Please find below 2X2 table.**





**#conclusion: Do you think the graduate intention and being female are independent events: NO.**

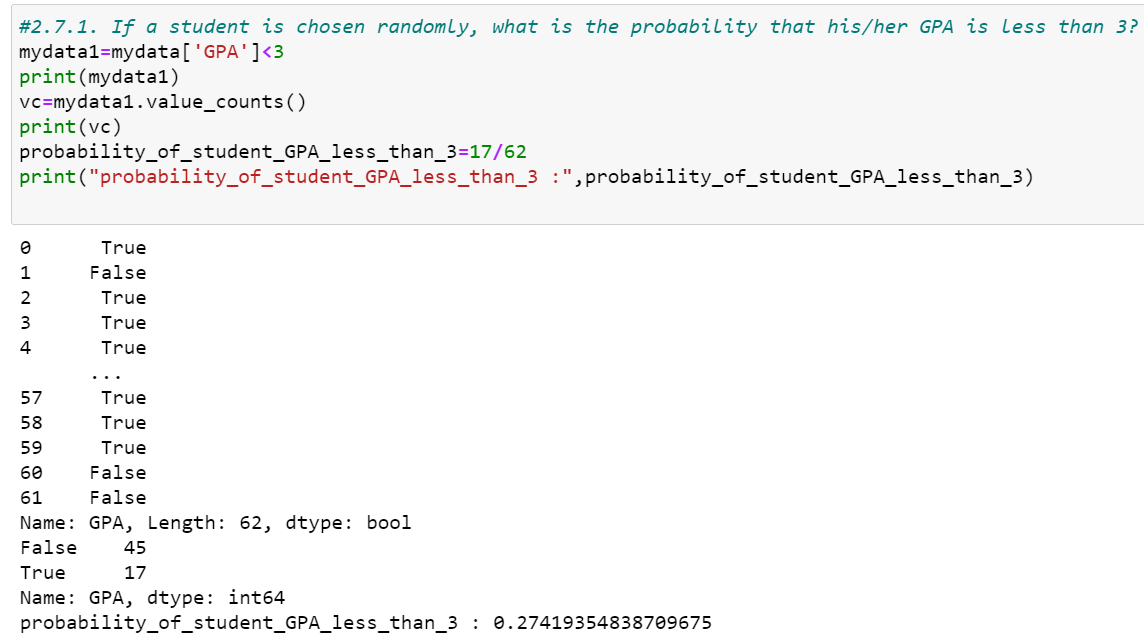
**Because, P(F).P(yes) not equal to P(F n Yes).**

**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?**

**ANS: Probability of student GPA less than 3 : 0.27419354838709675**



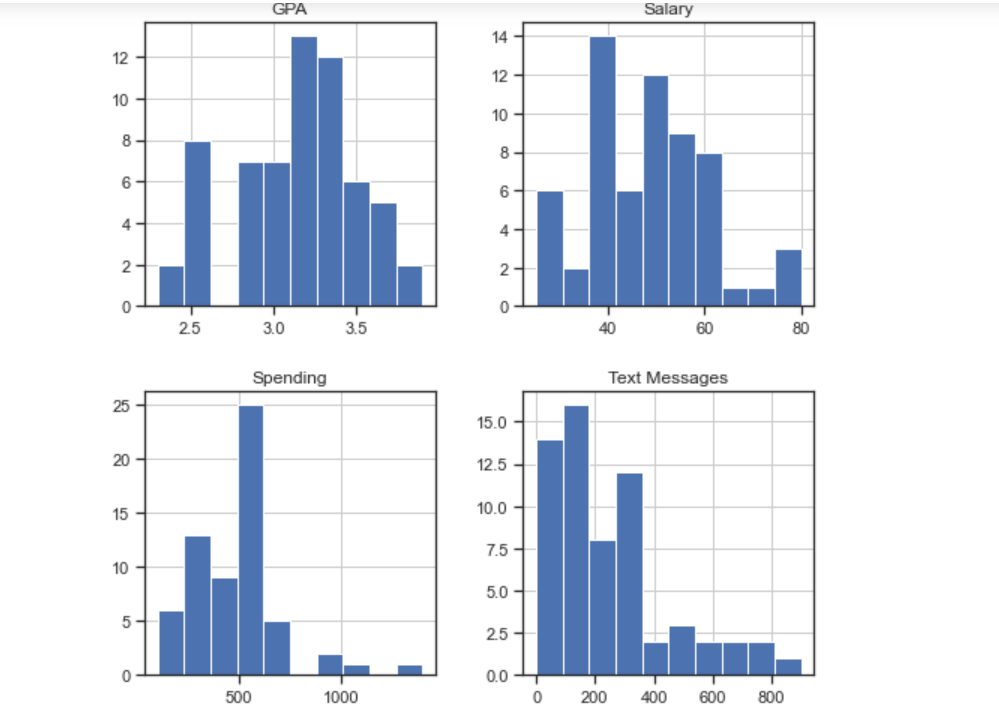
**2.7.2. Find the conditional probability that a randomly selected male earns 50 or more. Find the conditional probability that a randomly selected female earns 50 or more.**

**ANS: Probability randomly selected male earns 50 or more: 0.4827586206896552**

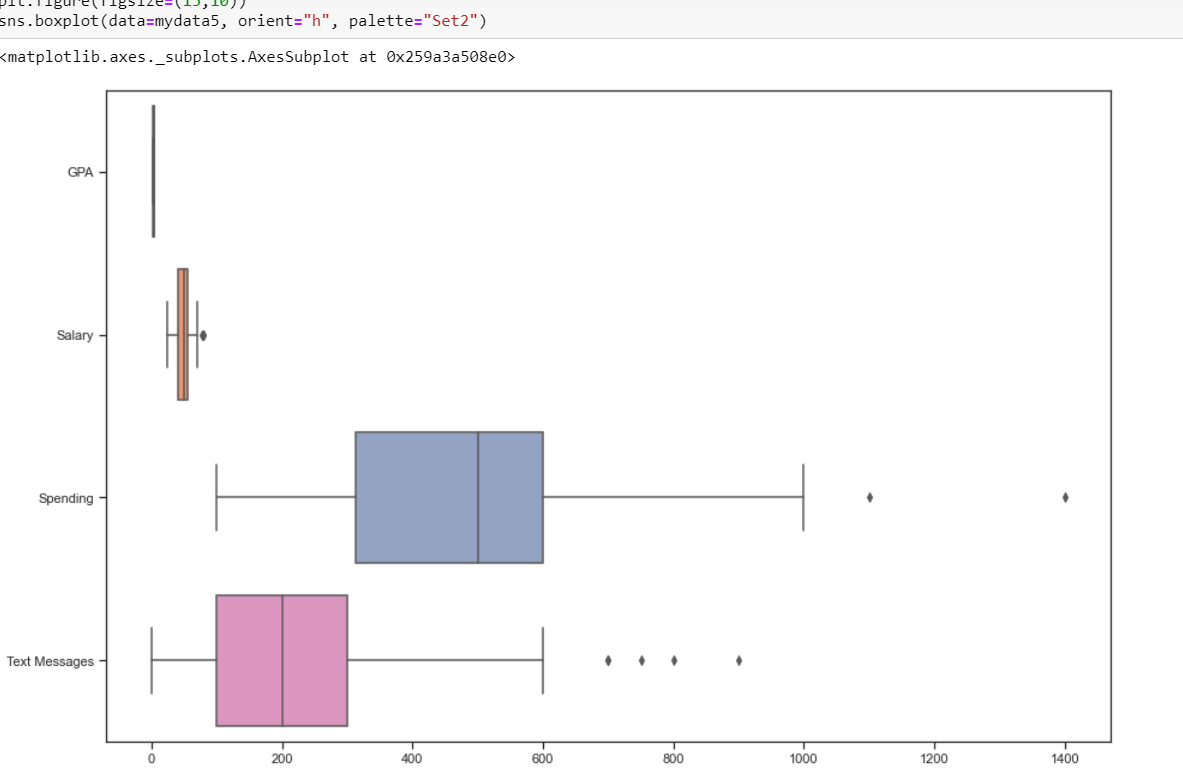
**ANS: Probability randomly selected female earns 50 or more: 0.5454545454545454**

# 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.

**ANS: Histogram Graph of (GPA, Salary, Spending, and Text Messages).**



**BoxPlot Graph (GPA, Salary, Spending, and Text Messages).**



**Write a note summarizing your conclusions.**

**#After analysing histogram and boxplot.**

**#I can conclude only GPA and Salary follow normal distribution.**

**#Spending and Text messages does not follow normal distribution because to many outliers present as per box plot graph.**

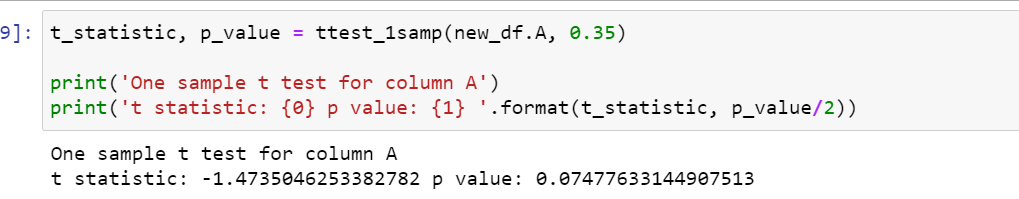
Problem 3

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 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 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](https://olympus.greatlearning.in/courses/32924/files/1735908/download?verifier=Obqplxdw312eNGcCIQQca22T4zboaI4xYff8ssjH&wrap=1)) 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.

# ANS: Please fins the P-value and T-statistics value of column A.

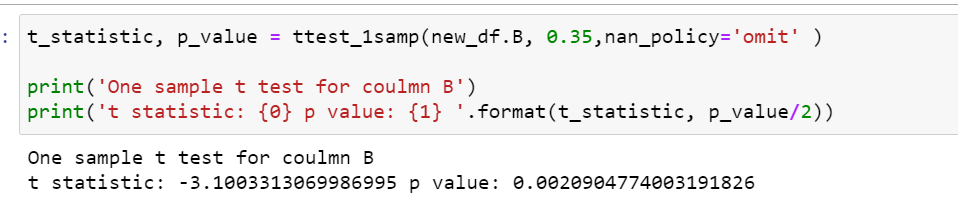


**#Since p-value > 0.05, do not reject H0 . There is not enough evidence to conclude that the mean moisture content for Sample A shingles is less than 0.35 pounds per 100 square feet.**

**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 0.0748.**

# ANS: Please fins the P-value and T-statistics value of column B.



**#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 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?

# ANS: Please find the conclusion below.

# 

# #As the p\_value > alpha, 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.