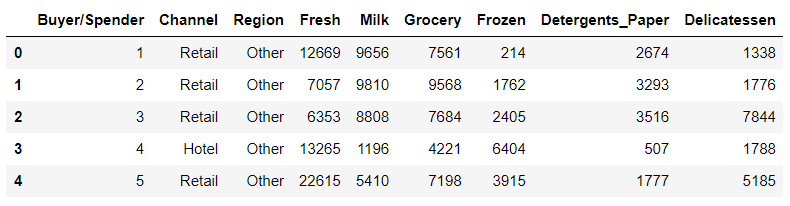
**Wholesale Customers Analysis**

**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).

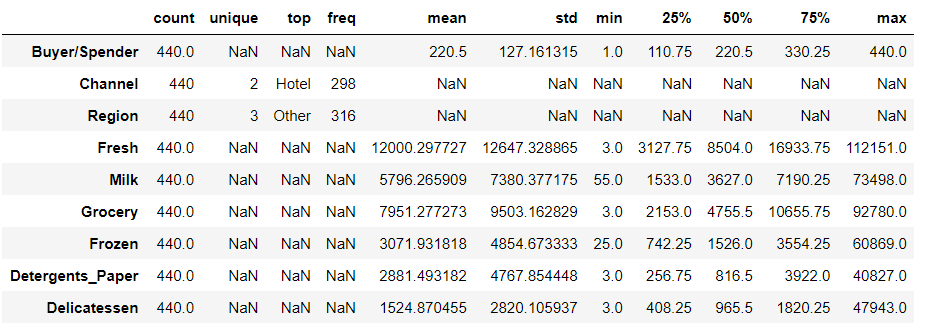


Channel – shows whether the channel is hotel or retail

Region – sales across different regions

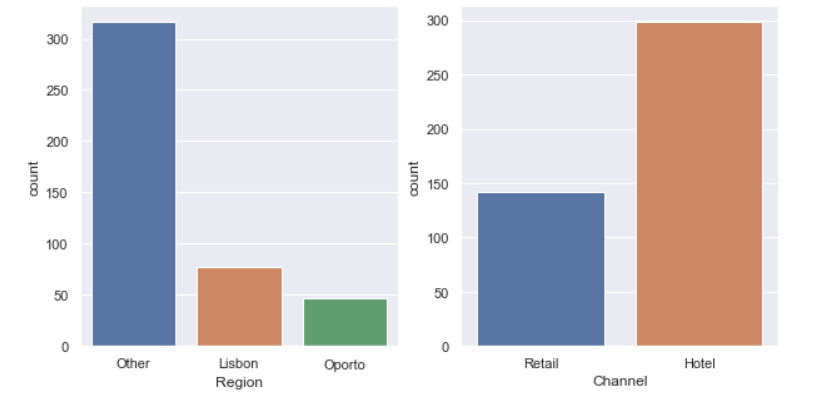
Fresh, Milk, Grocery, Frozen, Detergents Paper, Delicatessen – These are the 6 items distributed in different regions.

* 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?**



Observation :

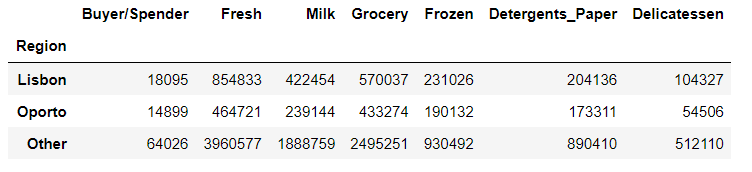
* There are 440 counts in each and every column
* Channel has 2 unique values
* Region has 3 unique values
* Excluding the Buyer/Spender, we can say that Fresh has the highest mean and Delicatessen has the lowest mean.
* The minimum value is 3 for Fresh, Grocery, Detergents Paper and Delicatessen
* Fresh has the maximum value 112151
* 25%, 50%, 75% are the inter-quartile ranges which is nothing but IQR

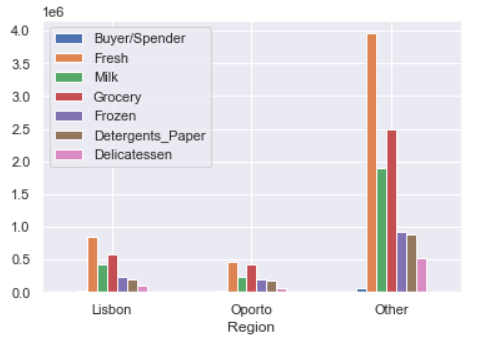
****

From the graph we can observe that:

* Other Region spends the more
* Hotel channel spends the more
* Oporto region spends the less
* Retail channel spends the less
  1. **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.**

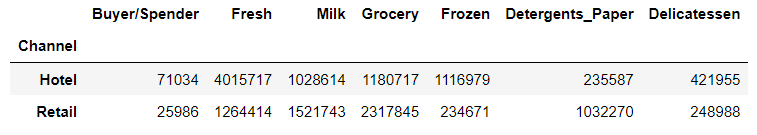
**Varieties across Region:**

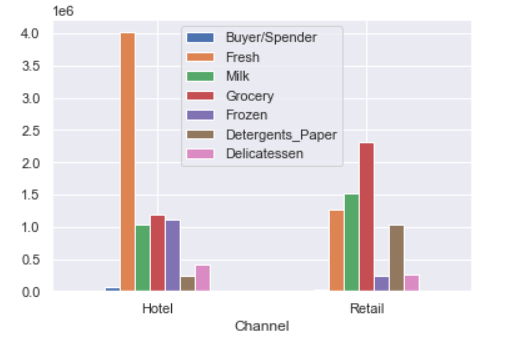
****

**  
From the graph, we can observe that sales based on regions:**

* **Fresh, Milk, Grocery, Frozen, Detergent Papers are the highest selling in all regions.**
* **Fresh, Milk, Grocery, Frozen, Detergents Paper are the highest selling products in other regions than the Lisbon and Oporto regions.**
* **Delicatessen is the lowest selling product in all regions but some what it is higher in the other region.**
* **Frozen and Detergents Paper are the least selling in all regions, it may be associated the person who is buying Frozen may also buy Detergents Papers also.**

**Varieties across Region:**

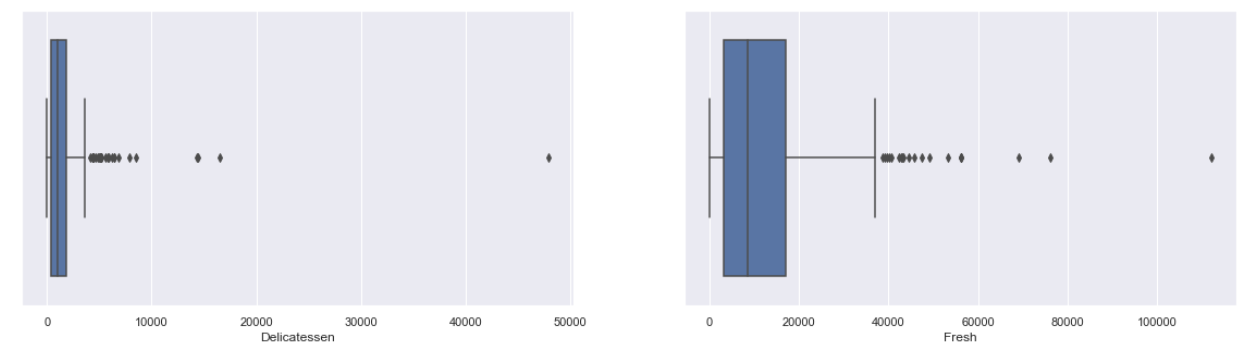




**From the graph, we can observe that sales based on channel:**

* **It seems Fresh is the daily need for the hotel channels for their customers than the retail.**
* **Milk, Grocery and Frozen are not much deviating as it is a need for hotel channel but not more than Fresh.**
* **Grocery sales are higher than the Fresh in retail channel.**
* **Delicatessen item sale is high in hotel channel and low in retail channel.**

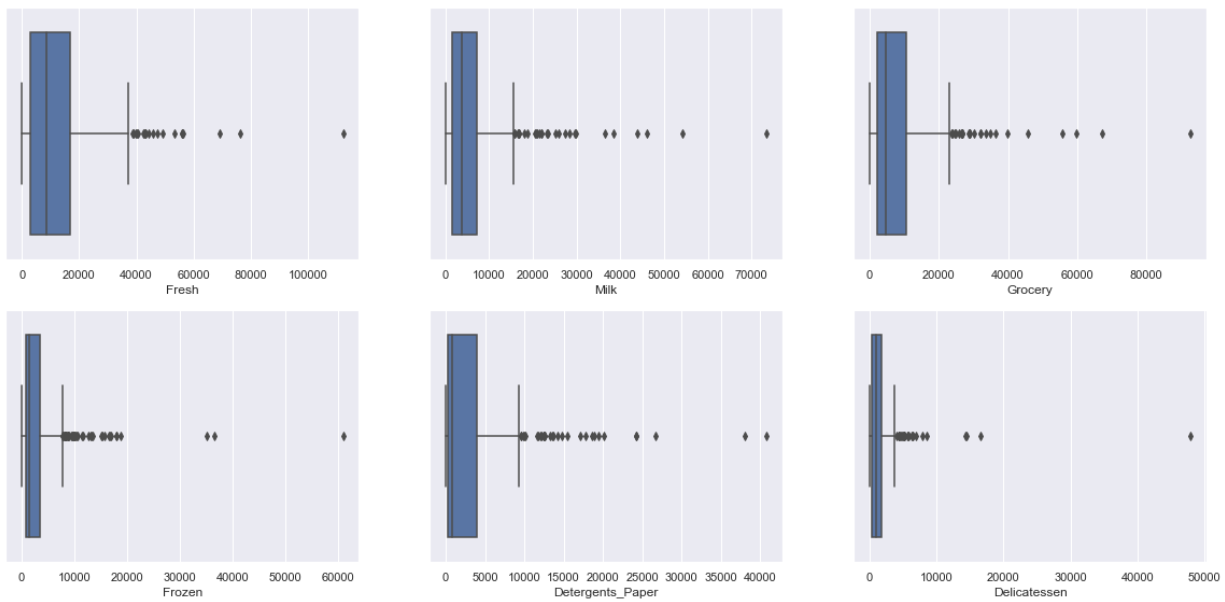
**1.3 On the basis of a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?**



**From the above box plot:**

* **we can observe that 'Delicatessen' item has the most inconsistent which is nothing but high variations and 'Fresh' item has the least inconsistent which is nothing but low variations**

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



**From the above plots:**

* **All the items from the plots have outliers extremely.**
* **These outliers may cause the business workflow**

**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**

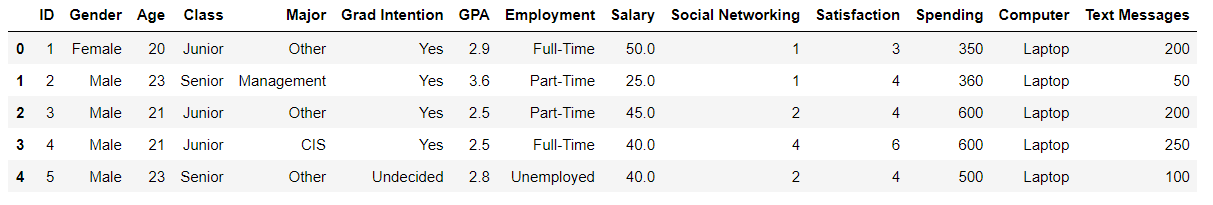
* **Based on the analysis, we can observe that most of the buyers are from other regions than the Lisbon and Oporto regions. More preferably they buy the Fresh varieties and there are some daily needs like Grocery and Milk which all the regions buy.**
* **We found that varieties like Frozen, Detergents Paper and Delicatessen are not popular among the regions and retail channels. So, we can try to maximize these products where the demand is more.**

**My recommendations would be varieties like Fresh, Grocery and Milk to all regions and channels as it is daily needs. Sometimes the demand may be extremely higher or some what it can decrease but it won’t decrease extremely low. Varieties like Frozen, Detergent Papers and Delicatessen can be sold in Hotel channels as there will be high usage and threshold stocks to retail channel as there will be minimal usage. So based on the population, usage and season we can plan the products need to be sold to various regions and channels.**

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



**Id – unique number for each and every student**

**Age – age of each student**

**Class – whether the student is junior, senior or sophomore**

**Major – major subject in which student pursued**

**Grad Intention – student wants to graduate, not or undecided**

**GPA – score for each student**

**Employment – student current job situation**

**Salary – salary of each student**

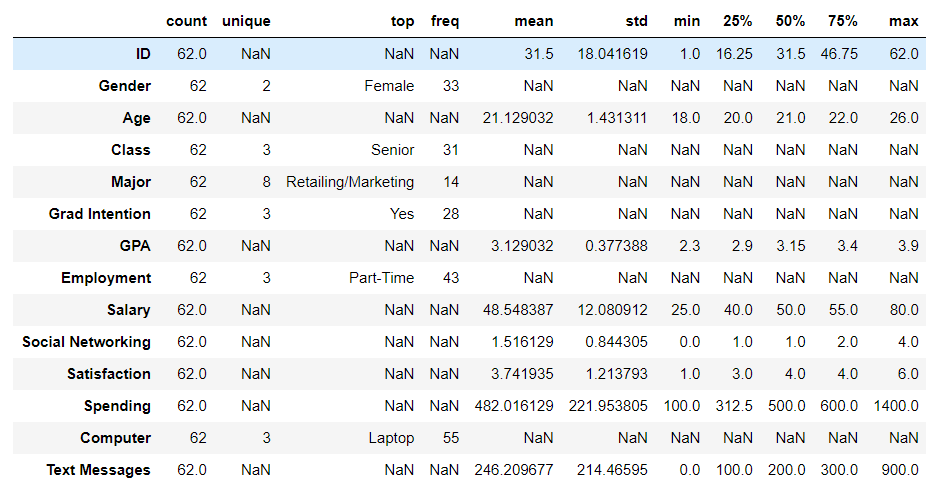
**Social Networking – active socially platforms count**

**Satisfaction – feedback from students**

**Spending – amount spent by students**

**Computer – whether the student has laptop, desktop or laptop**

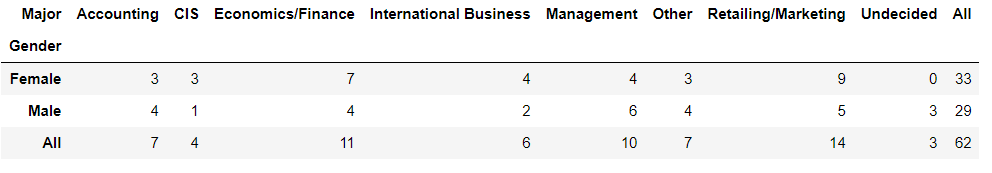
**Text Messages – count of words in feedback form**



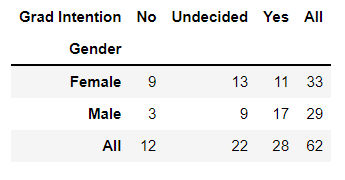
* **Male count is 29, Female count is 33**
* **21 is the age mean and maximum age is 26**
* **Maximum GPA is 3.9 and Minimum GPA is 2.3**
* **55 students use Laptop for education**
* **In employment part-time is more than others**

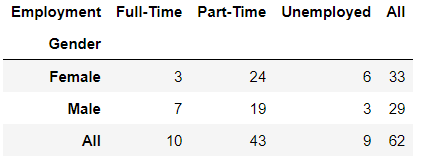
**2.1. For this data, construct the following contingency tables (Keep Gender as row variable)**

**2.1.1. Gender and Major**

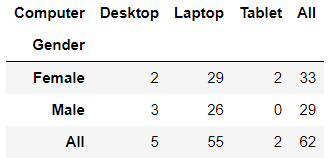
****

**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?**

**To find P(Male):**

**Count (male) = 29**

**Total count = 62**

**P(Male) = count(male)/Total count**

**P(Male) = 29/62 = 0.46774 or 46.77%**

**The probability that a randomly selected CMSU student will be male is 46.77%**

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

**To find P(Female):**

**Count (Female) = 33**

**Total count = 62**

**P(Female) = count(male)/Total count**

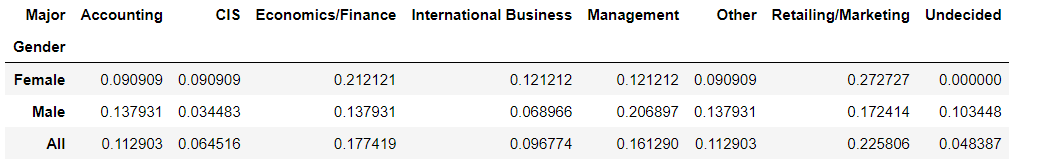
**P(Female) = 33/62 = 0.53225 or 53.22%**

**The probability that a randomly selected CMSU student will be female is 53.22%**

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

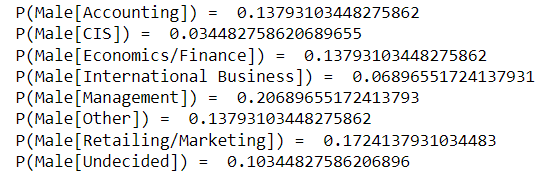
**2.3.1. Find the conditional probability of different majors among the male students in CMSU.**

**Conditional probability of different majors,**



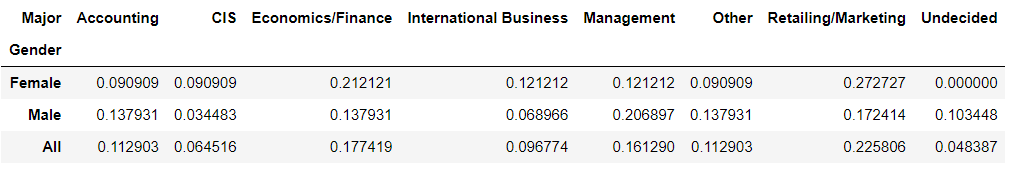
**Prob (Different Major)/count(male)**

**The above snippet shows the probability of male choosing different majors, the clear output is below**



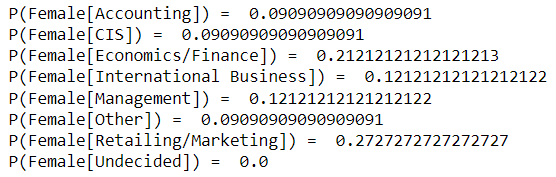
**2.3.2 Find the conditional probability of different majors among the female students of CMSU.**

**Conditional probability of different majors,**



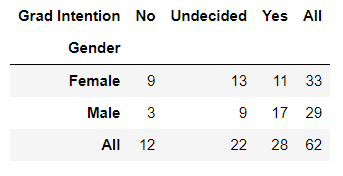
**Prob (Different Major)/count(female)**

**The above snippet shows the probability of female choosing different majors, the clear output is below**



**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.**



**Count(male) = 29**

**Male intends to graduate = 17**

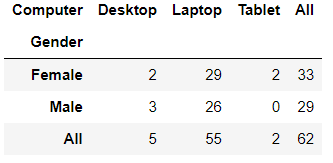
**P (Male and intends to graduate) = male intends to graduate/ count(male)**

**= 17/29**

**= 0.5862 or 58.62%**

**The probability that a randomly chosen student is a male and intends to graduate is 58.62%.**

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



**Count(female) = 33**

**Female who doesn’t have laptop = P (female with desktop) + P (female with tablet)**

**= 2+2**

**= 4**

**P (Female who doesn’t have laptop) = Female who doesn’t have laptop/count(female)**

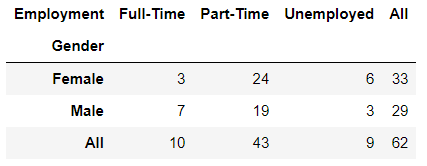
**= 4/33**

**= 0.1212 or 12.12%**

**The probability that a randomly chosen student is a male and intends to graduate 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?**



**P (Male | Fulltime)**

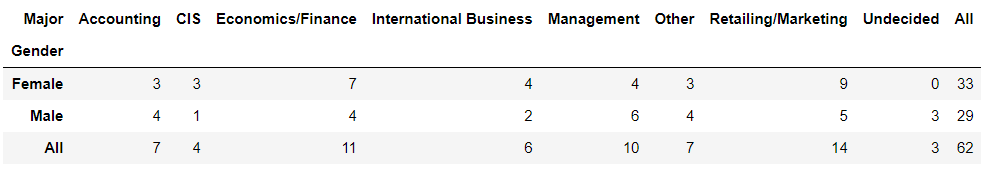
**= P (Male count/ Total count) + P (Full-time/ Total full time) – P (Male. Full time/Total count)**

**= (29/62) + (10/62) – (7/62)**

**= 0.5161 or 51.61%**

**The probability that randomly chosen student is a male or has a full-time employment is 51.61%.**

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



**Count(female) = 33**

**International business or management = 4 + 4 = 8**

**P (international business or management | Female)**

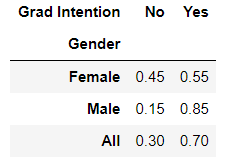
**= international business or management / count(female)**

**= 8/33**

**=0.2424 or 24.24%**

**The conditional probability that given a female student is randomly chosen, she is majoring in international business or management is 24.24%.**

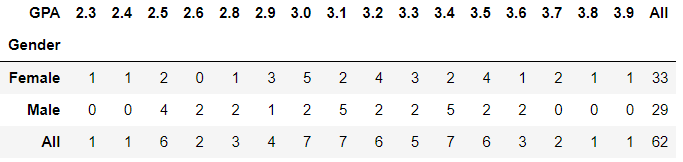
**2.6.  Construct a contingency table of Gender and Intent to Graduate at 2 levels (Yes/No). The Undecided students are not considered now and the table is a 2x2 table. Do you think the graduate intention and being female are independent events?**



**The graduate intention and being female are not independent events because only female has 50% but female who intends to graduate is 55%**

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

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



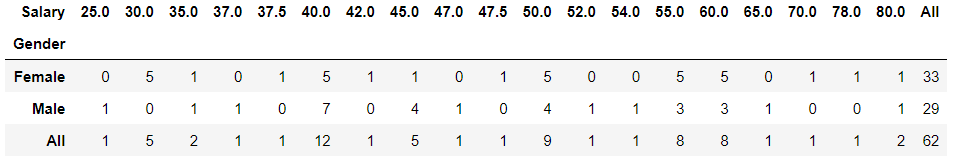
**P (his/her GPA less than 3) = (1+1+6+2+3+4)/62**

**= 17/62**

**= 0.2741 or 27.41%**

**The probability 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.**



**Male earn 50 or more:**

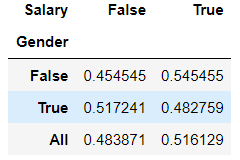
**P (Male earn 50 or more) = P (male salary >= 50)/ count(male)**

**= (4+1+1+3+3+1+0+0+1)/29**

**= 14/29**

**= 0.4827 or 48.27%**

**The probability that a randomly selected male earns 50 or more is 48.27%.**



**Female earn 50 or more:**

**P (Female earn 50 or more) = P (female salary >=50)/ count(female)**

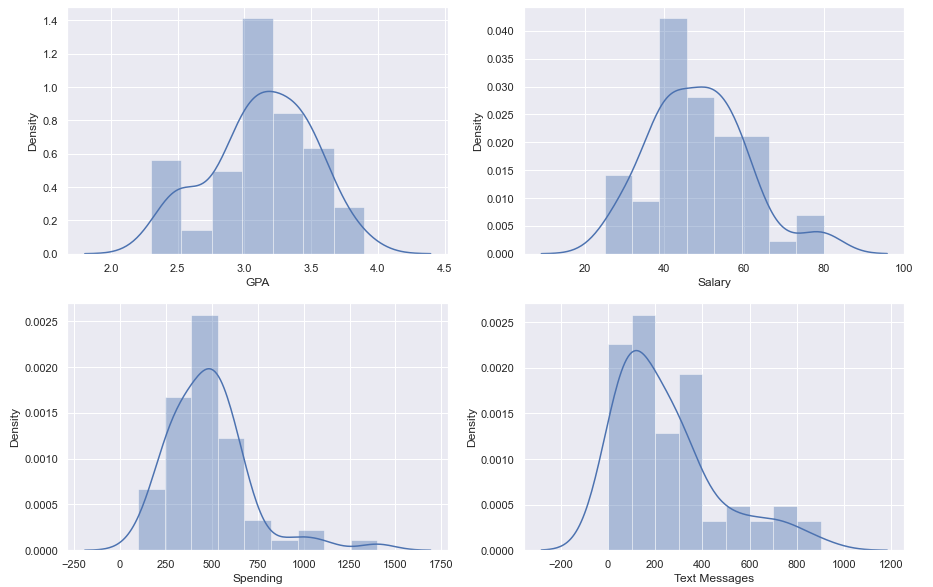
**= (5+0+0+5+5+0+1+1+1)/33**

**= 18/33**

**= 0.5454 or 54.54%**

**The probability that a randomly selected female earns 50 or more is 54.54%.**

**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 graph we can observe that:**

* **Salary, Spending and Text messages are normally distributed**
* **For the GPA, majority of data is distributed on the right side, hence it is negatively skewed**
* **If we are going with positive skewness then our algorithm going to make good predictions and very poor predictions on the higher side**
* **If we are going with negative skewness then our algorithm going to make good predictions on higher side and very poor predictions on the lower side**
* **By looking into positive skewness and negative skewness we find the likelihood situation**

**Conclusion:**

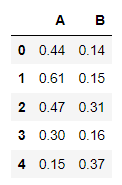
**Through out the analysis we can conclude that based on the survey of 62 students response in which both male and female are included. Many students have the intention of graduating in retailing/marketing. 75% are looking for a part time jobs. Approx 35% students have not decided about their career. Students who all looking for job they are expecting a mean salary around 50. 55 students have laptop for their education.**

Shingles Analysis A & B

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.

The file ([A & B shingles.csv](https://olympus.mygreatlearning.com/courses/63571/files/3937993/download?verifier=19GUsMb4t0pa3h8z5iRbD0FASfM0HF7jZuijAmMm&wrap=1)) includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.

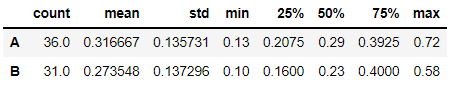


A – shingle is weighed and dried.

B – seems to be again reweighed readings.



5 null values present in the data.



* **A has 36 counts, B has 31 counts**
* **The mean value of A is higher than B**
* **The max value for A is 0.72 and min value is 0.13**
* **The max value for B is 0.58 and min value is 0.10**

**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.**

**Given:**

we have two independent samples of shingles A and B, where population standard deviation is unknown. So we will go with t-test. Since we have to find the mean moisture level is less than the permissible limit for the both samples we will use one sample t-test for sample A and sample B separately

**For sample A:**

**step1: Define null and alternative hypothesis**

**H0: mu >= 0.35**

**HA: mu < 0.35**

**step2: Decide the significance level**

**alpha = 0.05**

**step3: identify the test statistic**

**from the given data and above table, we can observe that,**

**we have one sample A**

**sample size n>30  
population standard deviation is not known**

**Hence we use tdist and tstat for one sample ttest. one tailed test is used here.**

**step4: calculate t\_statistic and p\_value**

t\_statistic : -1.4735046253382782

p\_value : 0.14955266289815025

**step5: Decide to reject or accept null hypothesis**

( 0.14955266289815025 **> 0.05 )**

**P\_value is greater than alpha**

* **we fail to reject null hypothesis**
* **we conclude that moisture content is greater than permissible limit in sample A**

**For sample B:**

**step1: Define null and alternative hypothesis**

**H0: mu >= 0.35**

**HA: mu < 0.35**

**step2: Decide the significance level**

**alpha = 0.05**

**step3: identify the test statistic**

**from the given data and above table, we can observe that**

**we have one sample B**

**sample size n>30**

**population standard deviation is not known**

**omit the null values**

**Hence we use tdist and tstat for one sample ttest. one tailed test is used here.**

**step4: calculate t\_statistic and p\_value**

t\_statistic : -3.1003313069986995

p\_value : 0.004180954800638365

**step5: Decide to reject or accept null hypothesis**

(0.004180954800638365 **> 0.05 )**

**P\_value is lesser than alpha**

* **we reject null hypothesis**
* **we conclude that moisture content is less than permissible limit in sample B**

**From the above tests for sample A and B, we can observe that for sample B has enough evidence to reject null hypothesis and hence we conclude that moisture content is less than the permissible limit in sample B**

**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?**

**step1: Define null and alternative hypothesis**

**H0: muA != muB**

**HA: muA = muB**

**step2: Decide the significance level**

**alpha = 0.05**

**step3: identify the test statistic**

**from the given data and above table, we can observe that,**

**we have two samples A and B**

**sample size n>30**

**sample sizes for both samples are not same**

**population standard deviation is not known**

**omit the null values**

**Hence we use tdist and tstat for two sample ttest. Two-Tailed test is used here.**

**step4: calculate t\_statistic and p\_value**

t\_statistic : 1.2896282719661123

p\_value : 0.2017496571835306

**step5 : Decide to reject or accept null hypothesis**

( 0.2017496571835306 **> 0.05 )**

**P\_value is greater than alpha**

* **we fail to reject null hypothesis**
* **we conclude that population means are not same**