**ADVANCE STATISTICS BUSINESS REPORT**

**Problem 1: -**

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).

In basic EDA we could understand more about the data such as

Shape = (62, 14)

Head – We can use this to understand the first five rows and column is dataset.

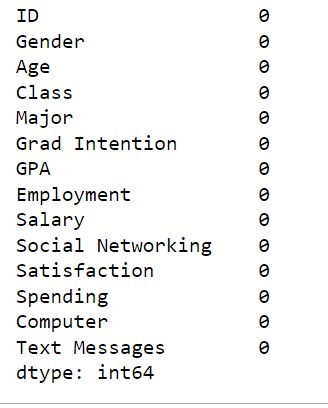
Info – we have 62 entries and 113 column and we don’t have any Null values

We have,

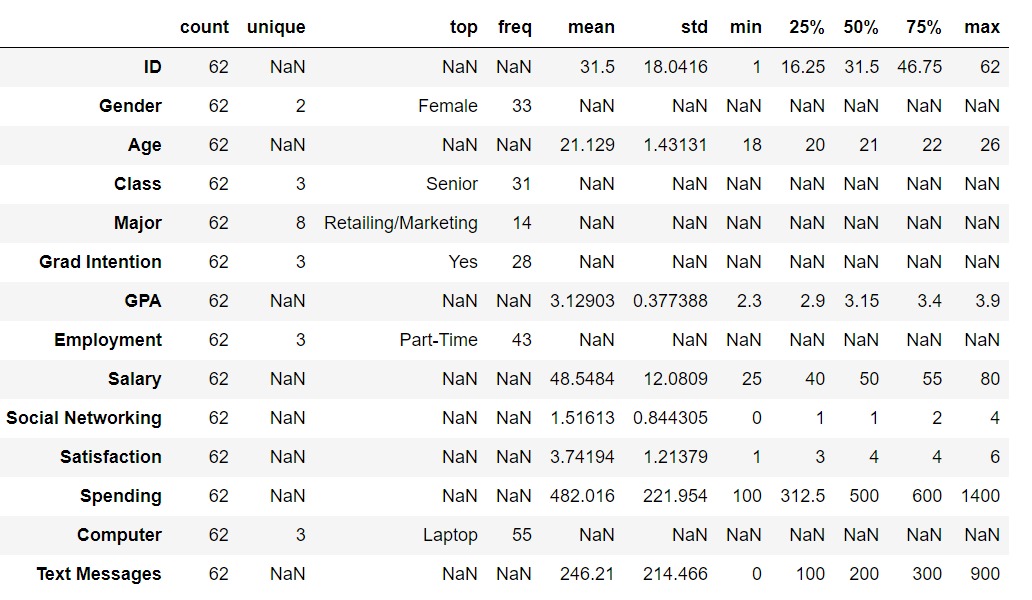
GPA and Salary as float values

Gender, Class, Major, Grade intention, Employment and Computer as object values

Null values = 0



Descriptive statistic of the data set



We have unique values in Gender, Class, Major, Grade Intention, Employment and Computer from descriptive statistics of five point summary.

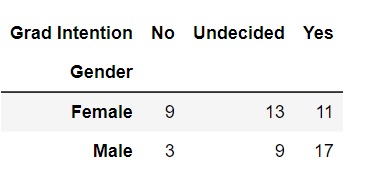
* Number of females is 33.
* Number of males is 29.
* The maximum age of students is 26 and median age 21
* Class we have 3 unique values senior, junior, sophomore.
* In majors we have 8 majors.
* Students have scored 3.9 GPA, but median remains around to be 3.15
* Retailing/marketing is the most preferred Major by students.
* 28/62 have grad intent.
* Part-time seems to be more when compared to fulltime job.
* 55 have laptop for the education.

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

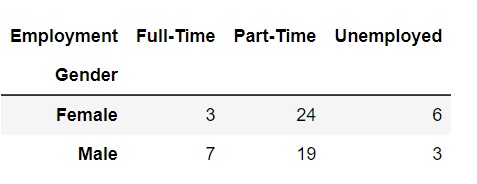
**1.1.1. Gender and Major**

****

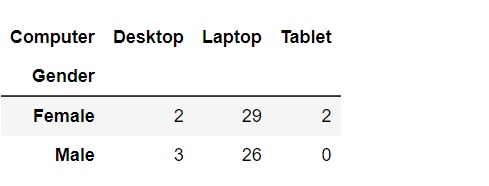
**1.1.2. Gender and Grad Intention**

****

**1.1.3. Gender and Employment**

****

**1.1.4. Gender and Computer**



**1.2. Assume that the sample is representative of the population of CMSU. Based on the data, answer the**

**following question:**

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

Number of male (A)

Total Number of students (B) = 62

P (A/B) = 29/62

The probability that a randomly selected CMSU student will be male is 46.774193548387096%.

In [17]:

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

Number of female (A) = 33

Total number of students (B) = 62

P (A/B) = 33/62

The probability that a randomly selected CMSU student will be female is 53.2258064516129%.

**1.3. Assume that the sample is representative of the population of CMSU. Based on the data, answer the**

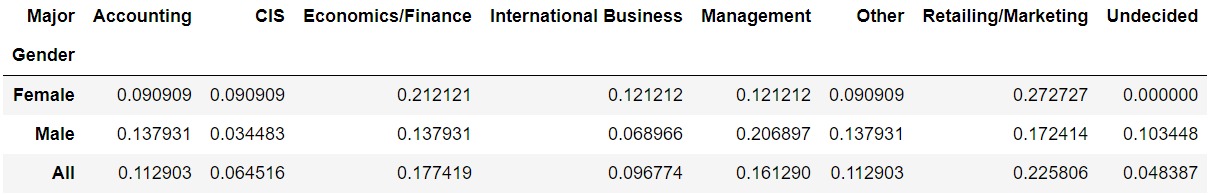
**following question:**

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

Conditional probability of different Majors

P (Different Majors/Male)

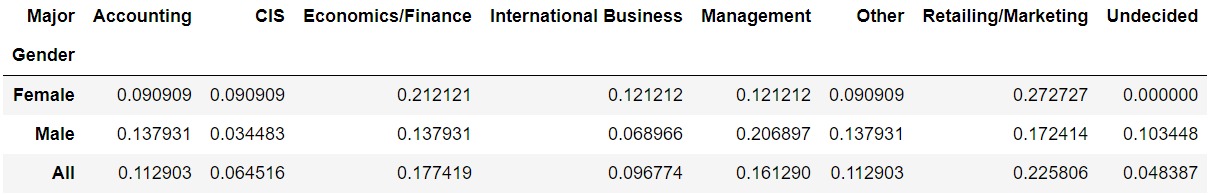
The snippet shows the probability of male choosing different majors.



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

P (Conditional Major/female)

The snippet shows the probability of female choosing different majors.

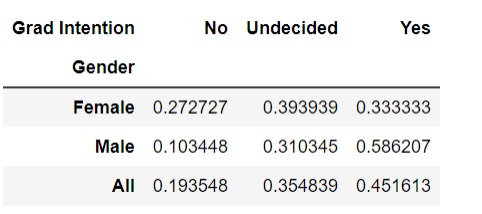
****

**1.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer**

**the following question:**

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

P (Grad intent yes/male) = 17/29

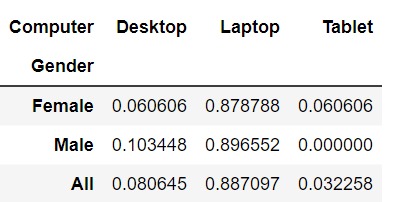


The probability that randomly chosen student is male and intends to graduate is 58.62%

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

P (Have a laptop/female) = 29/33

P (Does not have laptop/female) = 1-P (Have a laptop/Female) = 1-0.88=12%



The probability that randomly selected student is female and does not have a laptop is 12%

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

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

Probability that a randomly selected student is male P (A) = 46.77%

Probability of randomly selected student has a fulltime job P (B) = 16.13%

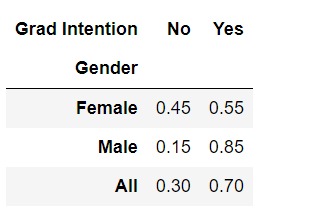
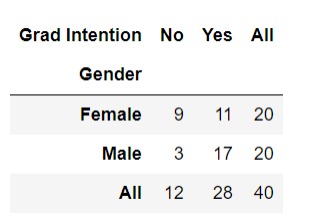
Probability of male having a fulltime job P (A and B) = 11.29%

The probability that a randomly chosen student is either a male or has full-time employment 51.61290322580645%

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

Probability that given a female student is randomly chosen, she is majoring in international business or management 24.44 %

**1.6.  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?**



**Conclusion: -**

The probability that a randomly selected Student is Female 50.0

The probability that a randomly selected student is female and intends to graduate 55.0%

They are not independent events.

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

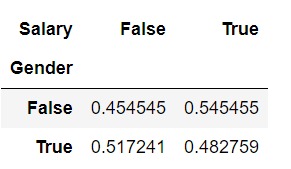
**Text Messages.**

**Answer the following questions based on the data.**

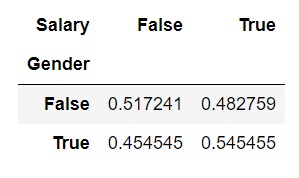
**1.7.1. If a student is chosen randomly, what is the probability that his/her GPA is less than 3?**

The probability that his/her GPA is less than 3 is 27.419354838709676 %

**1.7.2. 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.**

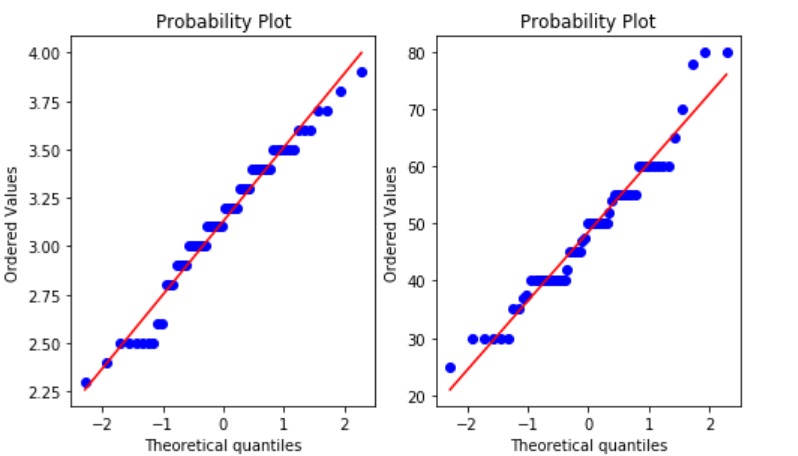
****

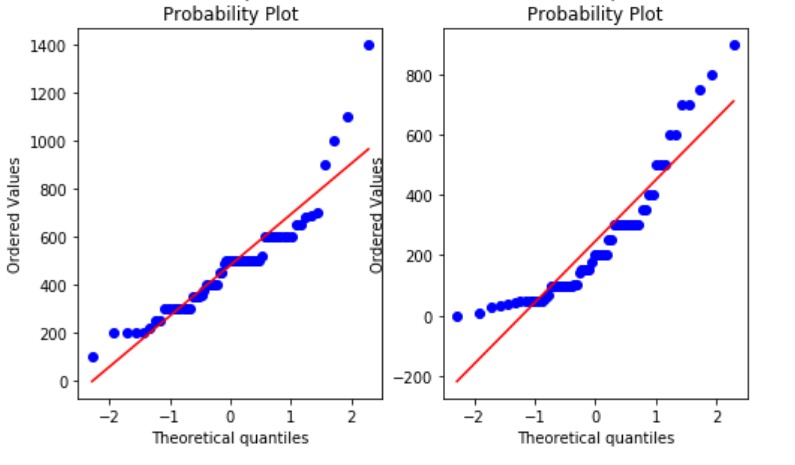
Probability that a randomly selected male is earns 50 or more is 48 %

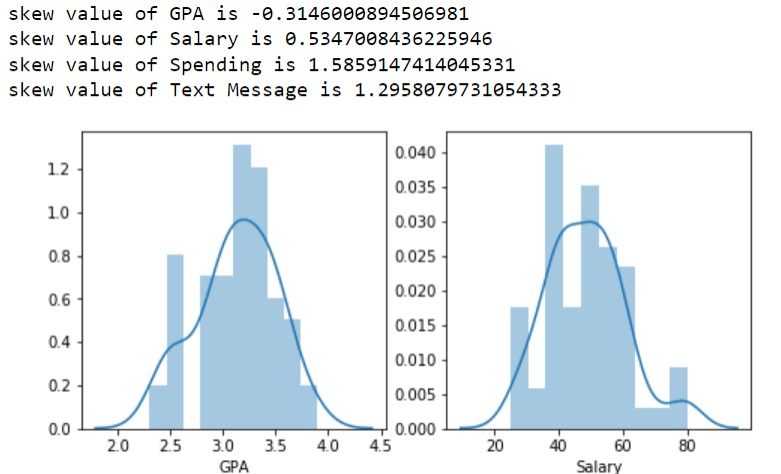


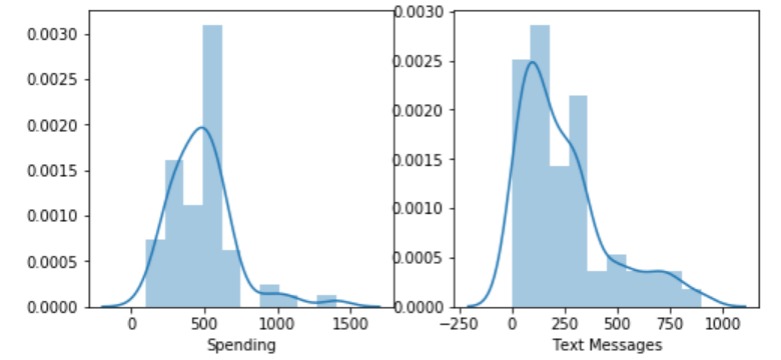
Probability that a randomly selected female is earns 50 or more is 54 %

**1.8. 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.**





****

****

The probability plot can be used to find the dataset follows a normal distribution or not, in our dataset we can find the points follow a straight line and we can say that all the GPA, salary, spending and text messages follow a normal distribution.

Looking at the skew value if the value is zero it is symmetric data, if we have a negative value for the skew that indicates that the data are skewed left and positive value of skew indicates the data are skewed toward right.

**Conclusion: -**

We have dataset of students answering to the survey and we have 62 responses from the students both male and female. We have almost equal number of male and female students. Many students have intention of graduating the retailing and marketing seem to have chosen by quite number of students. 2/3 of the students are looking for a part time job. The mean salary means to be around 50.

**Problem 2: -**

**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 pounds per 100 square feet.**

**The file (A &amp; B shingles.csv) includes 36 measurements (in pounds per 100 square feet) for A shingles and 31. for B shingles.**

**2.1 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.**

Solution:

In this problem we have provided with two independent samples A and B population standard deviation is unknown and hence we cant perform z test. So, we have to go with t-test.

Sample A

Step 1:

Define Null and alternate hypothesis.

The null hypothesis states that the moisture content of sample A is greater or than equal to the permissible limit, 𝜇 ≥ 0.35.

The alternative hypothesis states that the moisture content of sample A is less than permissible limit, 𝜇 < 0.35.

𝐻0: 𝜇 ≥ 0.35

𝐻𝐴: 𝜇 < 0.35

## Step 2: Decide the significance level.

Here we select 𝛼 = 0.05 as given in the question.

## Step 3: Identify the test statistic.

We have two samples (A and B) and we do not know the population standard deviation. Sample sizes for both samples are not the same. The sample size is , n > 30. So we use the t distribution and the 𝑡𝑆𝑇𝐴𝑇 test statistic for one sample test for A sample. One tail test for sample A

## Step 4: Calculate the p - value and test statistic.

Xbar = 0.316667

S = 0.135731

N = 36

Mu = 36

Tstat = -1.4735

(P value/2) = 0.0747

Step 5:

Decide to reject or accept null hypothesis.

Since tstat > p value, we fail to reject the null hypothesis

We conclude that the moisture content is greater than permissible limit in simple A.

### Define Null and alternate hypothesis for sample B

### step 1:

Testing whether the moisture content is less the permissible limit.

The null hypothesis states that the moisture content of sample B is greater or than equal to the permissible limit, 𝜇 ≥ 0.35

The alternative hypothesis states that the moisture content of sample B is less than permissible limit, 𝜇 < 0.35

𝐻0 : 𝜇 ≥ 0.35

𝐻𝐴 : 𝜇 < 0.35

**Step 2:**

Decide the significance level.

Here we select 𝛼 = 0.05 as given in the question.

Step 3: Identify the test statistic.

We have two samples (A and B) and we do not know the population standard deviation. Sample sizes for both samples are not the same. The sample size is , n > 30. So we use the t distribution and the 𝑡𝑆𝑇𝐴𝑇 test statistic for one sample test for B sample. one tail test for Sample B

Step 4: Calculate the p - value and test statistic

Xbar = 0.2735

S = 00.1372

N = 31

Mu = 0.35

Tstat = -3.10020

(P value/2) = 0.0020

### **Step 5:**

### Decide to reject or accept null hypothesis.

one-sample t-test p-value= 0.0020904774003191826

We have enough evidence to reject the null hypothesis in favour of alternative hypothesis

We conclude that the moisture content is less than permissible limit in sample B.

**2.2 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?**

In testing whether the mean for shingles A and Shingles B are the same, the null hypothesis states that the mean of shingle A to mean of shingle B are the same, µA equals µB. The alternative hypothesis states that the mean is different, µA is not equal to µB.

### Step 2: Decide the significance level.

Here we select  = 0.05 and the population standard deviation is not known.

### Step 3: Identify the test statistic.

We have two samples and we do not know the population standard deviation. Sample sizes for both samples are not the same. The sample size is, n > 30. So we use the t distribution and the 𝑡𝑆𝑇𝐴𝑇 test statistic for two sample test.

Step 4: Calculate the p - value and test statistic.

t = (M1 -M2)/(S2M1+S2M2)

calculation:

N1=36 N2=31

M1=0.32 M2=0.27

S21=0.02 S21= 0.02

DF1=35 DF2=30

**Tstat 1.2896282719661123**

**P Value 0.2017496571835306**