|  |
| --- |
| Photo displaying partial image of two pie charts on a canvas-textured page |
| SMDM Project Report  PGP-DSBA |
| |  |  |  | | --- | --- | --- | | Varun Kumar | 11/7/21 | Academic Project Report | |

Contents

[**Problem 1** 4](#_Toc87026595)

[**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?** 4](#_Toc87026596)

[**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.** 6](#_Toc87026597)

[**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?** 7](#_Toc87026598)

[**1.4** **Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.** 8](#_Toc87026599)

[**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** 9](#_Toc87026600)

[Problem 2 9](#_Toc87026601)

[**2.1. For this data, construct the following contingency tables (Keep Gender as row variable)** 9](#_Toc87026602)

[**2.1.2. Gender and Grad Intention** 10](#_Toc87026603)

[**2.1.3. Gender and Employment** 10](#_Toc87026604)

[**2.1.4. Gender and Computer** 10](#_Toc87026605)

[**2.2. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:** 11](#_Toc87026606)

[**2.2.1. What is the probability that a randomly selected CMSU student will be male?** 11](#_Toc87026607)

[**2.2.2. What is the probability that a randomly selected CMSU student will be female?** 11](#_Toc87026608)

[**2.3. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:** 11](#_Toc87026609)

[**2.3.1. Find the conditional probability of different majors among the male students in CMSU.** 12](#_Toc87026610)

[**2.3.2 Find the conditional probability of different majors among the female students of CMSU.** 12](#_Toc87026611)

[**2.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:** 13](#_Toc87026612)

[**2.4.1. Find the probability That a randomly chosen student is a male and intends to graduate.** 13](#_Toc87026613)

[**2.4.2 Find the probability that a randomly selected student is a female and does NOT have a laptop.** 13](#_Toc87026614)

[**2.5. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:** 13](#_Toc87026615)

[**2.5.1. Find the probability that a randomly chosen student is a male or has full-time employment?** 14](#_Toc87026616)

[**2.5.2. Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management.** 14](#_Toc87026617)

[**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?** 14](#_Toc87026618)

[**2.7. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages.** 15](#_Toc87026619)

[**2.7.1. If a student is chosen randomly, what is the probability that his/her GPA is less than 3?** 15](#_Toc87026620)

[**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.** 16](#_Toc87026621)

[**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.** 16](#_Toc87026622)

[Problem 3 18](#_Toc87026623)

[**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.** 19](#_Toc87026624)

[**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?** 20](#_Toc87026625)

[Figure 1 Bar Graph Between Region vs Spend 4](#_Toc87026381)

[Figure 2 Bar Graph Channels vs Spend 5](#_Toc87026382)

[Figure 3 Box Plot of all Items showing Outliers 7](#_Toc87026383)

[Table 1 Descriptive Summary of All the Six Items 3](#_Toc87026398)

[Table 2 Spend Region wise 4](#_Toc87026399)

[Table 3 Spend Channel wise 5](#_Toc87026400)

[Table 4 Spend Region and Channel wise across all items 6](#_Toc87026401)

[Table 5 Contingency Table, Gender & Major 8](#_Toc87026402)

[Table 6 Contingency Table, Gender & Grad Intention 9](#_Toc87026403)

[Table 7 Contingency Table, Gender & Employment 9](#_Toc87026404)

[Table 8 Contingency Table, Gender & Computer 9](#_Toc87026405)

[Table 9 Contingency Table, Gender & Grad Intention 14](#_Toc87026406)

[Table 10 Contingency table Gender & GPA<3 14](#_Toc87026407)

[Table 11 Contingency Table, Gender & Salary above or equal to 50 15](#_Toc87026408)

[Equation 1 11](#_Toc87026424)

[Equation 2 11](#_Toc87026425)

[Equation 3 13](#_Toc87026426)

[Equation 4 13](#_Toc87026427)

[Equation 5 14](#_Toc87026428)

[Equation 6 14](#_Toc87026429)

[Equation 7 15](#_Toc87026430)

[Equation 8 16](#_Toc87026431)

[Equation 9 16](#_Toc87026432)

# **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).

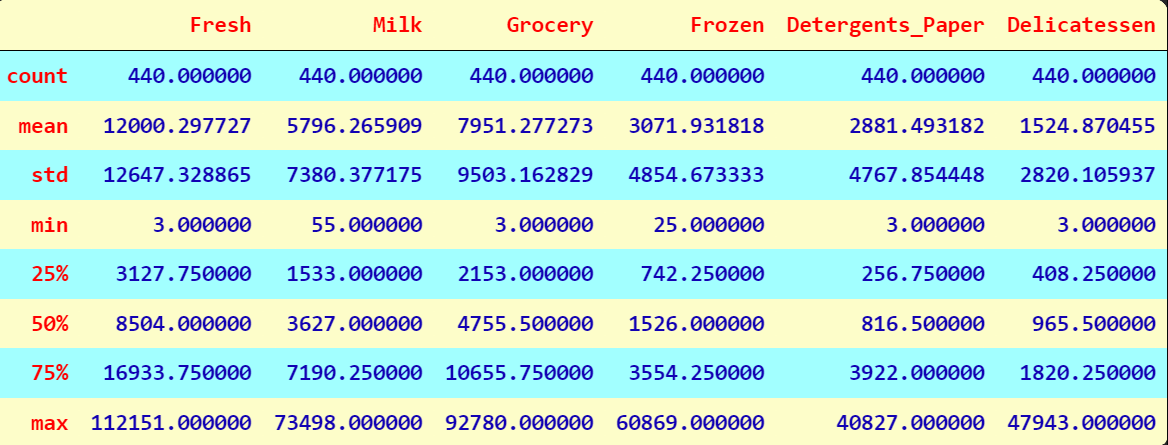
We imported the ‘Wholesale Customer data’ dataset in python to analyse the spend under each store items across regions and channel to find solutions to each problem. Below is the detailed approach and answer.

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

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

Using describe function in python we first looked at the basic descriptive statistics of the data set.

Table 1 Descriptive Summary of All the Six Items



Using bar graph with Region and Channel we were able to identify region with maximum spend and minimum spend. Below is the bar graph representation with the table.

Figure Bar Graph Between Region vs Spend

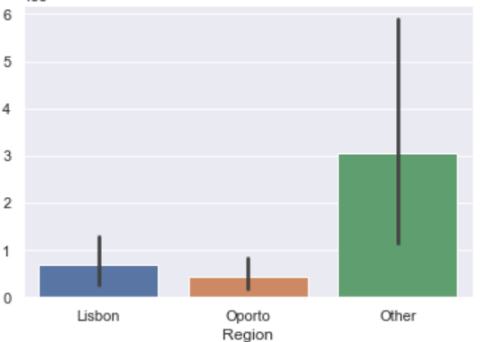
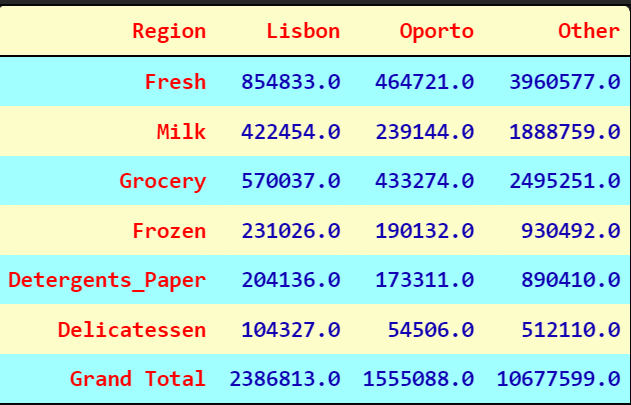


Table 2 Spend Region wise



Other regions spend amount is 10677599 with the highest spend amount and

Oporto region spend amount is 1555088 and has least spend amount by region.

Similarly, we grouped totals by channel to get totals by channel.

Figure Bar Graph Channels vs Spend

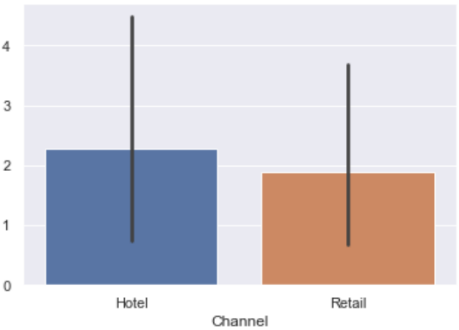
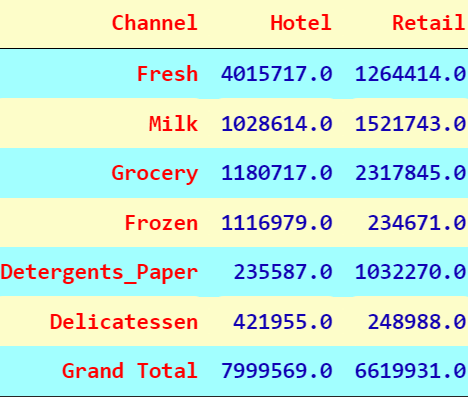


Table 3 Spend Channel wise

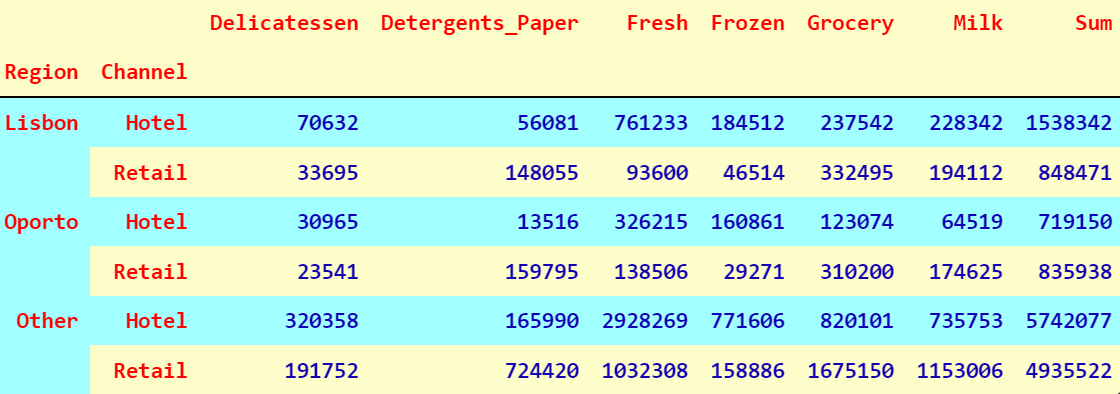


Hotel channel spend amount is 7999569 with the highest spend amount and Retail spend amount 6619931 has least spend amount based on Channel.

## **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.**

We can further breakdown the regions into channels. Using pivot tables for each category and checking spend across Region and Channel we get the following outputs -

Table 4 Spend Region and Channel wise across all items



As we already seen that Other’s spend most, but on further analysis we find that under the other’s and Lisbon region, Channel Hotel, spends more than Channel Retail. But in Oporto region, Channel Retail spends more. Looking at the above tables, we see that some categories like Milk, Grocery & Detergents Paper higher spend in the Retail channel versus Hotel, across all regions, except Lisbon (only in the case of “Milk”). On the other hand, Delicatessen, Fresh and Frozen have higher consumption in the Hotel channel versus Retail, across all regions. In my opinion overall sale should be equal across both the channels because both the channels have a great potential to expand.

## **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?**

Using Coefficient of Variation, we find out the least value is of Category “Fresh” (1.05) and highest value is of Category “Delicatessen” (1.84).

Coefficient of Variation for Fresh is 1.053918

Coefficient of Variation for Grocery is 1.195174

Coefficient of Variation for Milk is 1.273299

Coefficient of Variation for Frozen is 1.580332

Coefficient of Variation for Detergents Paper is 1.654647

Coefficient of Variation for Delicatessen is 1.849407

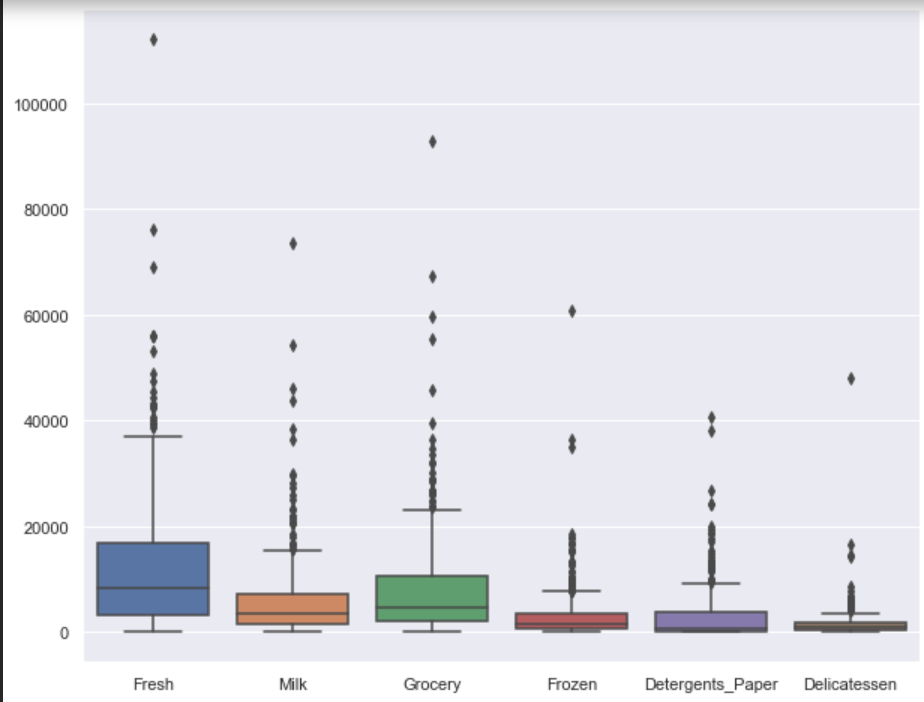
So, from the given data it is clear that most inconsistent behaviour shown by item – Delicatessen

And least inconsistent behaviour shown by item – Fresh

## **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.

Figure Box Plot of all Items showing Outliers



## **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 as well, not only “Fresh” and “Grocery”. Both the channels, Hotel & Retail, have great potential across all the regions. After this analysis, Distributor must need to find out the reasons that why items other than “Fresh” and “Grocery”, is not producing the desired sale? For that Distributor should conduct the survey among the markets and find out the reasons of low sales among other categories and need to understand the exact requirements of these two channels, consumers as well.

# Problem 2

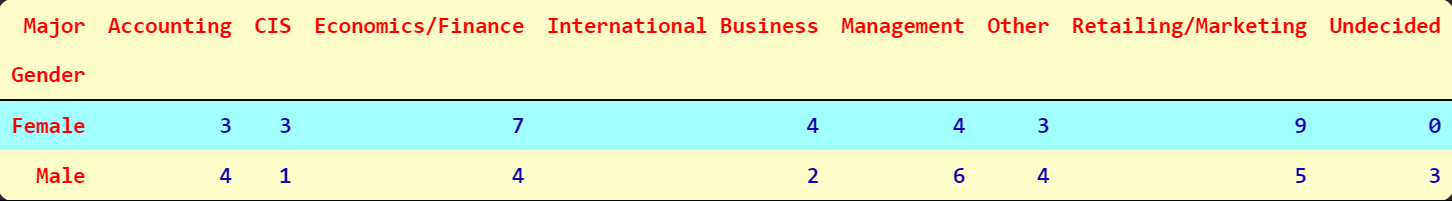
The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. CMSU creates and distributes a survey of 14 questions and receives responses from 62 undergraduates.

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

We imported the ‘Survey-1’ dataset in python, to analyse the students behaviour towards graduation and other aspects & to find solutions to each problem. Below is the detailed approach and answer.

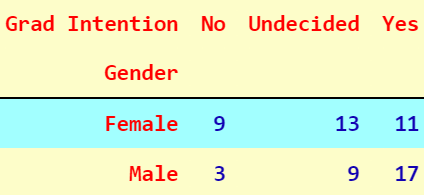
**2.1.1. Gender and Major**

Table 5 Contingency Table, Gender & Major



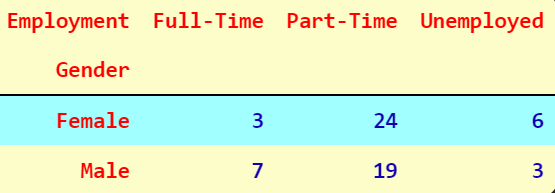
### **2.1.2. Gender and Grad Intention**

Table 6 Contingency Table, Gender & Grad Intention



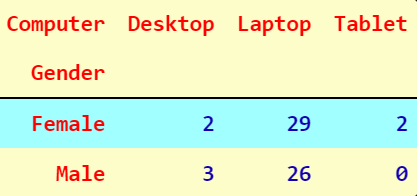
### **2.1.3. Gender and Employment**

Table 7 Contingency Table, Gender & Employment



### **2.1.4. Gender and Computer**

Table 8 Contingency Table, Gender & 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?**

For this we need to find out total male students out of whole student from the given data.

Total Students = 62

Number of male students = 29

Probability of a randomly selected CMSU student is male,

Equation 1

After calculation we got the result that 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?**

For this we need to find out total female students out of whole student from the given data.

Total Students = 62

Number of female students = 33

Equation 2

After calculation we got the result that 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

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 who are Undecided is 10.34%

And from this output we can easily say that most of the male 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

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 who are Undecided is 0.00%

And from this output we can easily say that most of the female 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 who intends to graduate.

Total Students = 62

Number of male students = 29

Number of male students who intends to graduate = 17

Equation 3

Probability of Males 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.**

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

Total Students = 62

Number of female students = 33

Number of female students who doesn’t have a laptop = 4

Equation 4

Probability of Females who doesn’t 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?**

Using contingency tables of Gender and Employment we got the total numbers of males and number of females.

Probability of Males is 29/62

Probability of Students with Full time employment is 10/62

Probability of Male Students with Full time employment is 7/62

Probability of randomly chosen student is Male or has Full time employment is

Equation 5

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

Using contingency tables of Gender and Major we got the total number of females.

Probability of Females majoring in, International Business is 4/33

Probability of Females majoring in, Management is 4/33

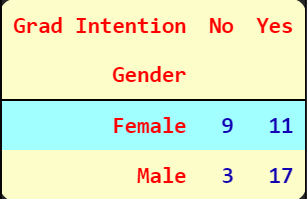
Probability of Female Student majoring in, International Business or Management is

Equation 6

## **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?**

Required contingency table is following

Table 9 Contingency Table, Gender & Grad Intention



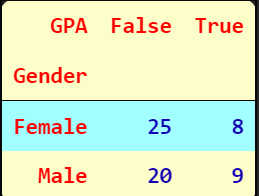
As we compare this contingency table & table 6, we can easily observe that students who has not decided or don’t have any Graduate Intention are from both the genders. Thus, we conclude that, the graduate intention & being female are independent events.

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

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

We need to create a contingency table between Gender and GPA<3

Table 10 Contingency table Gender & GPA<3



As we can see this table shows us Boolean Results. ‘True’ GPA<3, and ‘False’ for GPA≥3.

Moreover, this table also provides us the total number of Males & Females So we can easily calculate the required Probability.

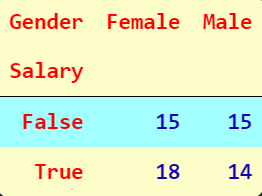
Probability of a randomly chosen student’s GPA less than three is,

Equation 7

### **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.**

We need to create a contingency table between Salary>=50 and Gender

Table 11 Contingency Table, Gender & Salary above or equal to 50



As we can see this table shows us Boolean Results. ‘True’ for the Salary is greater than or equal to 50, and ‘False’ for the Salary is less than 50.

Moreover, this table also provides us the total number of Males & Females So we can easily calculate the required Probability.

Probability that randomly selected Male earns 50 or more is

Equation 8

Probability that randomly selected Female earns 50 or more is 54.55%

Equation 9

## **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.**

GPA

We are going to perform Shapiro's Test to check the normal distribution of "GPA" in the given dataset.

For Shapiro's Test,

H0 is "GPA" is normally distributed

H1 is "GPA" is not normally distributed

Alpha is 0.05

Shapiro's Test indicates that p value is equal to 0.11204058676958084. Which is greater than 0.05, hence we fail to reject null hypothesis. With 95% confidence level we can say that "GPA" is normally distributed.

SALARY

We are going to perform Shapiro's Test to check the normal distribution of "Salary" in the given dataset.

For Shapiro's Test,

H0 is "Salary" is normally distributed

H1 is "Salary" is not normally distributed

Alpha is 0.05

Shapiro's Test indicates that p value is equal to 0.028000956401228905. Which is smaller than 0.05, hence we reject null hypothesis. We have enough evidence to reject the null hypothesis in favour of alternative hypothesis. We can say that "Salary" is not normally distributed.

SPENDING

We are going to perform Shapiro's Test to check the normal distribution of "Spending" in the given dataset.

For Shapiro's Test,

H0 is "Spending" is normally distributed

H1 is "Spending" is not normally distributed

Alpha is 0.05

Shapiro's Test indicates that p value is equal to 1.6854661225806922e-05. Which is smaller than 0.05, hence we reject null hypothesis. We have enough evidence to reject the null hypothesis in favour of alternative hypothesis. We can say that "Spending" is not normally distributed.

TEXT MESSAGES

We are going to perform Shapiro's Test to check the normal distribution of "Text Messages" in the given dataset.

For Shapiro's Test,

H0 is "Text Messages" is normally distributed

H1 is "Text Messages" is not normally distributed

Alpha is 0.05

Shapiro's Test indicates that p value is equal to 4.324040673964191e-06. Which is smaller than 0.05, hence we reject null hypothesis. We have enough evidence to reject the null hypothesis in favour of alternative hypothesis. We can say that "Text Messages" is not normally distributed.

As we can observe only GPA is normally distributed and rest others are not. Students earn different salaries, which varies between 25 to 80 so their spending’s also vary accordingly, with a big range. As far as text messages are concerned, we can say that few students are more actively use texts and others doesn’t active on texts, maybe they prefer calls or something else, that we can’t infer from the given data.

# 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 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.greatlearning.in/courses/59234/files/3628276/download?verifier=TzzyUGqlsBm6IdJNfKlSAsJ5si820ZN2WWwifG1i&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.**

We are going to conduct a hypothesis test.

H0 ≥ 0.35

H1 < 0.35

Alpha=0.05

SHINGLES A

After applying the hypothesis test in Jupyter Notebook, we get the following results:

t statistic: -1.4735046253382782, p value: 0.07477633144907513

Since p value > 0.05, we fail to 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.

SHINGLES B

After applying the hypothesis test in Jupyter Notebook, we get the following results:

t statistic: -3.1003313069986995, p value: 0.0020904774003191826

Since p value < 0.05, reject H0. There is enough evidence to conclude that the mean moisture content for Sample B shingles is less than 0.35 pounds per 100 square feet.

## **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?**

HYPOTHESIS

H0: μ(A)= μ(B)

H1: μ(A)≠ μ(B)

Alpha = 0.05

After performing the two sample ttest in Jupyter Notebook we get,

t statistic = 1.29 and p value = 0.202

As the p value > 0.05, we fail to 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
* the variances of the two distributions are the same.