***SMDM Project Business Report***

***DSBA***

***Sanjay Srinivasan***

***PGP-DSBA Online***

***JULY’ 21 Batch***

***Date: 09-10-2021***

# *INDEX*

|  |  |  |
| --- | --- | --- |
| ***S. No*** | ***Contents*** | ***Page* No** |
| **1.** | **Problem - 1** | 5 |
|  | Summary | 5 |
|  | Introduction | 5 |
|  | Data Description | 5 |
|  | Sample Dataset | 5 |
|  | Exploratory Data Analysis | 6 |
|  | Checking for missing values in the dataset | 6 |
|  | 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? | 6 |
|  | 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. | 8 |
|  | 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? | 9 |
|  | 1.4 Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments. | 9 |
|  | 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 | 10 |
| **2.** | **Problem - 2** | 11 |
|  | Summary | 11 |
|  | Introduction | 11 |
|  | Data Description | 11 |
|  | Sample Dataset | 11 |
|  | Exploratory Data Analysis | 12 |
|  | Checking for missing values in the dataset | 12 |
|  | 2.1. For this data, construct the following contingency tables (Keep Gender as row variable) | 12 |
|  | 2.1.1. Gender and Major | 12 |
|  | 2.1.2. Gender and Grad Intention | 13 |
|  | 2.1.3. Gender and Employment | 13 |
|  | 2.1.4. Gender and Computer | 13 |
|  | 2.2. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question | 13 |
|  | 2.2.1. What is the probability that a randomly selected CMSU student will be male? | 14 |
|  | 2.2.2. What is the probability that a randomly selected CMSU student will be female? | 14 |
|  | 2.3. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question | 14 |
|  | 2.3.1. Find the conditional probability of different majors among the male students in CMSU. | 14 |
|  | 2.3.2 Find the conditional probability of different majors among the female students of CMSU. | 15 |
|  | 2.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question | 15 |
|  | 2.4.1. Find the probability That a randomly chosen student is a male and intends to graduate. | 15 |
|  | 2.4.2 Find the probability that a randomly selected student is a female and does NOT have a laptop. | 16 |
|  | 2.5. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question | 16 |
|  | 2.5.1. Find the probability that a randomly chosen student is a male or has full-time employment? | 16 |
|  | 2.5.2. Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management. | 16 |
|  | 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? | 16 |
|  | 2.7. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. | 17 |
|  | 2.7.1. If a student is chosen randomly, what is the probability that his/her GPA is less than 3? | 17 |
|  | 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. | 17 |
|  | 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. | 17 |
| ***3.*** | **Problem - 3** | 19 |
|  | Summary | 19 |
|  | Introduction | 19 |
|  | Data Description | 19 |
|  | Sample Dataset | 19 |
|  | Exploratory Data Analysis | 19 |
|  | Checking for missing values in the dataset | 19 |
|  | 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. | 20 |
|  | 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 |

# *List Of Tables*

|  |  |  |
| --- | --- | --- |
| ***S.No*** | ***Content*** | ***Page No*** |
| 1 | Dataset sample | 5 |
| 1.1.1 | Summary of the data | 7 |
| 1.1.2 | Calculating total spent | 7 |
| 1.1.3 | Total vs. Region | 7 |
| 1.1.4 | Total vs. Channel | 8 |
| 1.2 | Items across regions and channels | 8 |
| 2 | Dataset sample | 11 |
| 2.1.1 | Gender vs. Major | 13 |
| 2.1.2 | Gender vs. Grad Intension | 13 |
| 2.1.3 | Gender vs. Employment | 13 |
| 2.1.4 | Gender vs. Computer | 13 |
| 2.2 | Gender count | 13 |
| 2.3 | Gender vs. Major | 14 |
| 2.4.1 | Gender vs. Grad Intension | 15 |
| 2.4.2 | Gender vs. Computer | 16 |
| 2.5.1 | Gender vs. Employment | 16 |
| 2.5.2 | Gender vs. Major | 16 |
| 2.6 | Gender vs. Grad Intension | 17 |
| 3 | Dataset sample | 19 |

***List Of Figures***

|  |  |  |
| --- | --- | --- |
| ***S.No*** | ***Content*** | ***Page No*** |
| 1.1.1 | Total vs. Region bar plot | 7 |
| 1.1.2 | Total vs. Channel bar plot | 8 |
| 1.2 | Items across regions and channels | 9 |
| 1.3 | Calculating variance for each items | 9 |
| 1.4 | Boxplots for each items | 10 |
| 2.8.1 | GPA Bell curve | 17 |
| 2.8.2 | Salary Bell curve | 18 |
| 2.8.3 | Spending Bell curve | 18 |
| 2.8.4 | Text Messages Bell curve | 18 |

# Problem - 1

*Summary*

A wholesale distributor operating in different regions of Portugal has information on the annual spending of several items in their stores across different regions and channels. The dataset consists of annual spending cost for the products from the large retailers on 6 different varieties of products in 3 different regions (Lisbon, Oporto, Other) and across different sales channels (Hotel, Retail).In this problem statement, we will explore the different attributes of cost spent on different products based on region and channel.

*Introduction*

The purpose of this exercise is to explore the dataset. The exploratory data analysis of this dataset is as wholesale distributors operating in different regions of Portugal. The annual spending of several items in their stores across different regions and channels. The dataset consists of annual spending costs for the products from the large retailers on 6 different varieties. Analyzing the different spending costs of the items can help in analyzing the amount spends across different regions and channels. This assignment helps in exploring the summary statistics.

*Data Description*

1. Buyer/Spender: No of buyer buying the items.

2. Channels: Distributors operating in the channels(for example: Hotel, Retail).

3. Regions: Distributors operating in the regions(for example: Lisbon, Oporto, Other) .

4.Fresh: Amount spending in the Fresh item.

5.Milk: Amount spending in the Milk item.

6.Grocery: Amount spending in the Grocery item.

7.Frozen: Amount spending in the Frozen item.

8.Detergents Paper: Amount spending in the Detergents Paper item.

9.Delicatessen: Amount spending in the Delicatessen item.

*Sample of the dataset:*

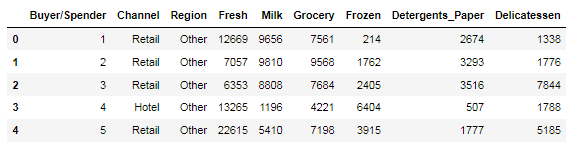


Table 1.Dataset Sample

Dataset has 9 variables with 6 different types of the items. Each item has different spending cost. Based on the characteristic amount spends on each item is defined.

*Exploratory Data Analysis*

*Let us check the types of variables in the data frame.*

|  |  |
| --- | --- |
| Buyer/Spender | int64 |
| Channel | object |
| Region | object |
| Fresh | int64 |
| Milk | int64 |
| Grocery | int64 |
| Frozen | int64 |
| Detergents\_Paper | int64 |
| Delicatessen | int64 |
| dtype: object |  |

There are total 440 rows and 9 columns in the dataset. Out of 9, 2 columns (Regions and Channels) are of object type and rest 7 are of either integer data type.

*Check for missing values in the dataset:*

RangeIndex: 440 entries, 0 to 439

Data columns (total 9 columns):

Buyer/Spender 440 non-null int64

Channel 440 non-null object

Region 440 non-null object

Fresh 440 non-null int64

Milk 440 non-null int64

Grocery 440 non-null int64

Frozen 440 non-null int64

Detergents\_Paper 440 non-null int64

Delicatessen 440 non-null int64

dtypes: int64(7), object(2)

memory usage: 31.0+ KB

### 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?

Descriptive statistics helps to describe and understand the features of a specific dataset by giving short summaries about the sample and measures of the data. The most recognized types of descriptive statistics are measures of centre: the mean, median, and mode, which are used at almost all levels of math and statistics.

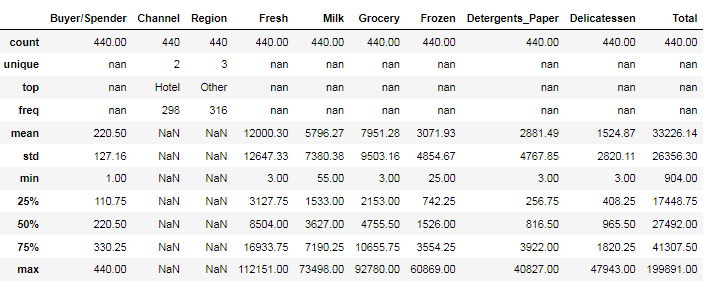


Table- 1.1.1. Summary of the data

From the descriptive statistics, we can see that there are 440 Buyers/Spender in the dataset. ‘Hotel’ is the most frequent Channel and ‘Other‘ is most frequent Region. The average price of the overall dataset is 33226.14. There are 2 unique Channel and 3 unique Region in the dataset.  **“Nan”** shows that the values cannot be calculated for that particular variables. Like we can calculate mean for a categorical/object type variable. And in a same way unique value for a numerical variable.

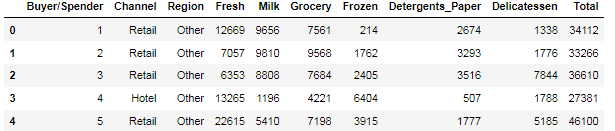


Table-1.1.2. Calculating total spent.

**Calculating the Highest and Least amount spent on each Region**

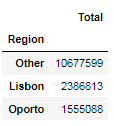


Table-1.1.3. Total vs Region

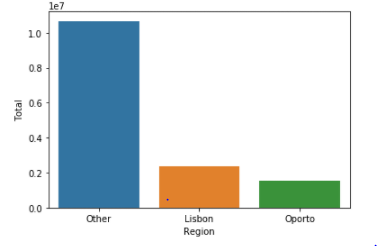


Fig - 1.1.1 Total vs. Region Bar Plot

Inferred from the above tables, we found the amount spent on the **other region** is the **highest** and the amount spent on the **Oporto region** is the **least**.

**Calculating the Highest and Least amount spent on each Channel**

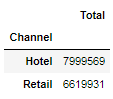


Table-1.1.5. Total vs Channel

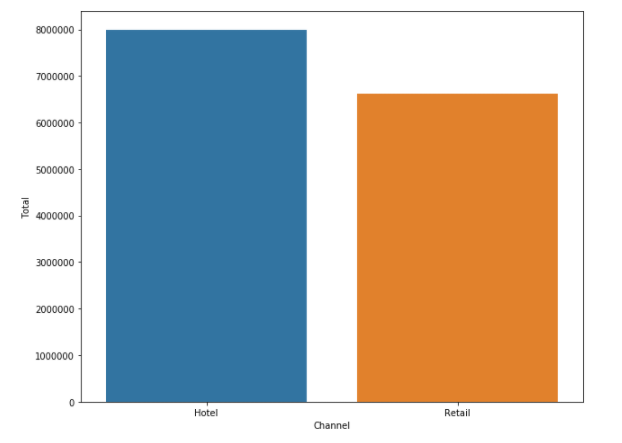


Fig – 1.1.2. Total vs Channel. Total Bar Plot

Inferred from the above tables, we found the amount spent on the **hotel channel** is the **highest** and the amount spent on the **Retail channel** is 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.

There are 6 different varieties of items are considered across regions and channel.

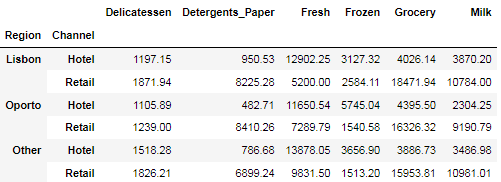


Table – 1.2 Items across Regions and Channels

In 'Lisbon' Region 'Delicatessen' is higher in the 'Retail' channel with that of the 'Hotel' channel. Highest amount spent on the item 'Fresh' than the other items in 'Hotel' channel, whereas in 'Retail' channel 'Grocery' has the highest amount spent.

In 'Oporto' Region, 'Detergents Paper' is the least amount spent in the 'Hotel' channel whereas 'Detergents Paper' is the third highest amount spent in the 'Retail' Channel. 'Fresh' and 'Grocery' is the item where highest amount spent in the channels 'Hotel' and 'Retail'.

In 'Other' Region, the amount spent on items 'Milk', 'Frozen' and 'Grocery' was almost the same in the 'Hotel' channel. In 'Retail' Channel, the amount spent on the 'Frozen' is least amount spent.

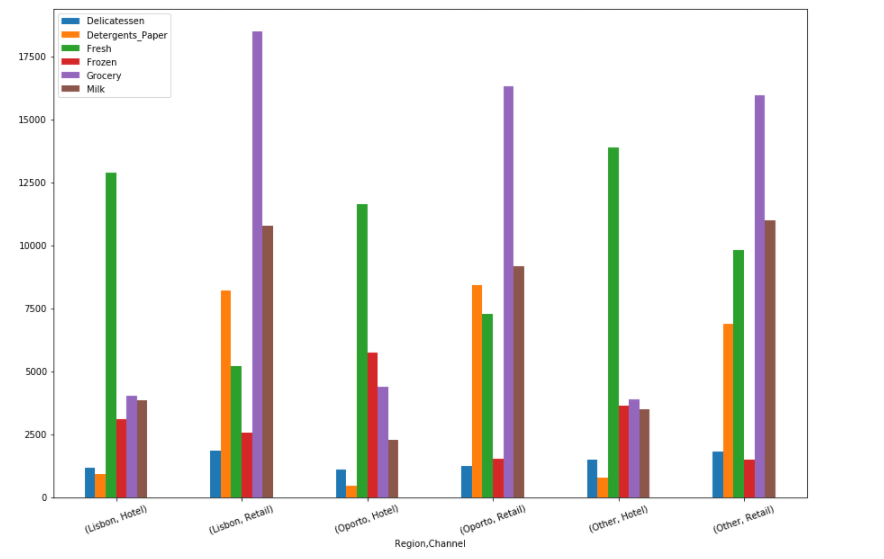


Fig – 1.2 Items across Regions and Channels

### 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?

## Calculating the variance

The variance is the ratio of standard deviation to that of the mean.

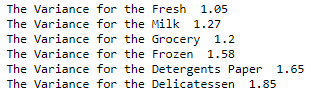


Fig – 3 Calculating variance for each items

From the descriptive statistics, we can calculate variance for each of the items. Which item has the least variance has the least inconsistent behaviour. From this inference, **Fresh** item has the **least inconsistent** behaviour with the value of **1.05** and the **Delicatessen** item has the **most inconsistent** behaviour with the value of **1.85**

### 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 descriptive statistics, we can find the outliers by plotting the values in the boxplot.

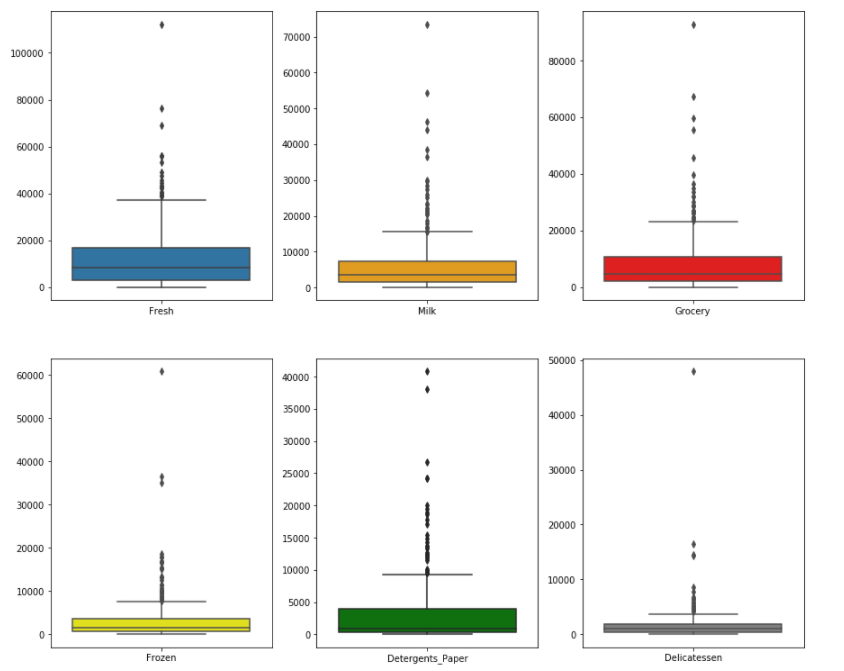


Fig – 4 Boxplots for each item

All the values of each item from the dataset have been plotted in the boxplot. Inferred from the boxplot, Items (Fresh, Milk, Grocery, Frozen, Detergents\_Paper, Delicatessen) has the 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 Descriptive analysis, I find out from the Coefficient of Variation that there are inconsistencies in spending of different items, which should be reduced. The spending for the Hotel and Retail channel should be more or less equal, but here they are different which needs to be addressed. And also spending should be equal for different regions. Need to focus on spending for other items than the “Fresh” and “Grocery”.

# Problem – 2

*Summary*

The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. The dataset consists of responses from 62 undergraduates. In this problem statement, we will explore the different questions that has been created by CMSU for the survey and analysis the response from the undergraduates.

*Introduction*

The purpose of this exercise is to explore the dataset. The exploratory data analysis of this dataset is the Student News Service at Clear Mountain State University (CMSU). The survey took from the undergraduates students at the Student News Service at Clear Mountain State University (CMSU). The dataset consists of survey responses of 62 undergraduates from the clear mountain state university. This assignment helps in exploring the responses made by the undergraduates provide inferential statistics.

*Data Description*

1. ID: No of undergraduates responded.

2. Gender: Describe the sex (for example: Male, Female).

3. Age: Describe the responded graduate age.

4. Class: Describe the grades.

5. Major: Describe the department of their studies.

6. Grad Intention: Describe about their further education.

7. GPA: Marks persuaded in the class.

8. Employment: Employment status of the undergraduates.

9. Salary: Undergraduate employment salary details.

10. Social Networking: No of social network.

11. Satisfaction: Rated how satisfied they are.

12. Spending: Describe the expenditure of each undergraduate.

13. Computer: Describe the computer of each undergraduate.

14. Text Messages: Describe the number of text messages.

*Sample of the dataset:*

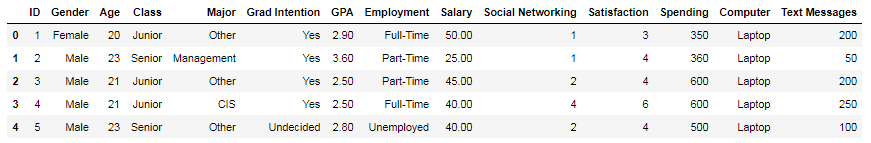
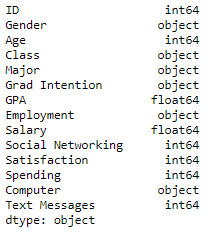


Table 2.Dataset Sample

Dataset has 14 variables with student details. Based on the characteristic Student details in the CMSU is defined.

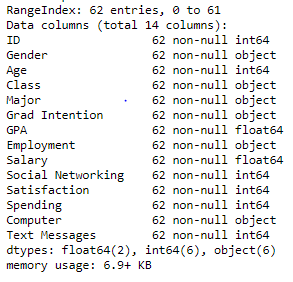
*Exploratory Data Analysis*

*Let us check the types of variables in the data frame.*



There are total 62 rows and 14 columns in the dataset. Out of 14, 6 columns are of object type, 2 columns are of float (Decimal value) type and rest 6 are of either integer data type.

*Check for missing values in the dataset:*



From this, it is clear that there are no null values present in the dataset.

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

Contingency Table: A table showing the distribution of one variable in rows and another in columns, used to study the correlation between the two variables.

#### 2.1.1. Gender and Major

Contingency table is plotted against gender vs. Major. From this we can infer that no of female and male persue across different majors.



Table 2.1.1.Gender vs. Major

**2.1.2. Gender and Grad Intension**

Contingency table is plotted against gender vs. Grad Intension. From this we can infer that no of female and male have graduation intensions.

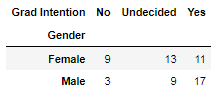


Table 2.1.2.Gender vs. Major

**2.1.3. Gender and Employment**

Contingency table is plotted against gender vs. Employment. From this we can infer that no of female and male have Employed and Unemployed.

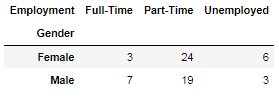


Table 2.1.3.Gender vs. Employment

**2.1.4. Gender and Computer**

Contingency table is plotted against gender vs. Computer. From this we can infer that no of female and male have which type of computer.

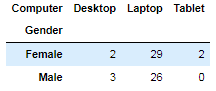


Table 2.1.4.Gender vs. Computer

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

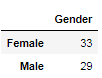


Table 2.2 Gender count

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

Probability of male = total male student / total student

***Probability\_male = 29 / 62 = 0.46774193548387094***

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

Probability of female = total female student / total student

***Probability\_female = 33 / 62 = 0.532258064516129***

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



Table 2.3 Gender vs. Major

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

Probability of Accounting male = Total Accounting male student / Total Male student

***Probability\_Accounting\_male = 4 / 29 = 0.13793103448275862***

Probability of CIS male = Total CIS male student / Total Male student

***Probability\_CIS\_male = 1 / 29 = 0.034482758620689655***

Probability of Economics/Finance male = Total (Economics/Finance) male student / Total Male student

***Probability\_Economics\_Finance\_male = 4 / 29 = 0.13793103448275862***

Probability of International Business male = Total International Business male student / Total Male student

***Probability\_International\_Business\_male = 2 / 29 = 0.06896551724137931***

Probability of Management male = Total Management male student / Total Male student

***Probability\_Management\_male = 6 / 29 = 0.20689655172413793***

Probability of Other male = Total Other male student / Total Male student

***Probability\_Other\_male = 4 / 29 = 0.13793103448275862***

Probability of Retailing male = Total Retailing male student / Total Male student

***Probability\_Retailing\_male = 5 / 29 = 0.1724137931034483***

Probability of Undecided male = Total Undecided male student / Total Male student

***Probability\_Undecided\_male = 3 / 29 = 0.10344827586206896***

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

Probability of Accounting female = Total Accounting female student / Total female student

***Probability\_Accounting\_female = 3 / 32 = 0.09090909090909091***

Probability of CIS female = Total CIS female student / Total female student

***Probability\_CIS\_female = 3 / 32 = 0.09090909090909091***

Probability of Economics/Finance female = Total (Economics/Finance) female student / Total female student

***Probability\_Economics\_Finance\_female = 7 / 32 = 0.21212121212121213***

Probability of International Business female = Total International Business female student / Total female student

***Probability\_International\_Business\_female = 4 / 32 = 0.12121212121212122***

Probability of Management female = Total Management female student / Total female student

***Probability\_Management\_female = 4 / 32 = 0.12121212121212122***

Probability of Other female = Total Other female student / Total female student

***Probability\_Other\_female = 3 / 32 = 0.09090909090909091***

Probability of Retailing female = Total Retailing female student / Total female student

***Probability\_Retailing\_female = 9 / 32 = 0.2727272727272727***

Probability of Undecided female = Total Undecided female student / Total female student

***Probability\_Undecided\_female = 0 / 32 = 0.0***

**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.**

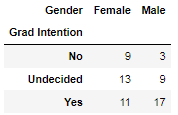


Table 2.4.1 Gender vs. Grad Intension

Probability of Male with grad Intensions = (Total male student / Total student)\*(male with grad intentions / total male student)

***Probability\_male\_with\_grad\_intension = (29 / 62)\*(17 / 29) = 0.27419354838709675***

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

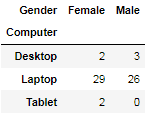


Table 2.4.2 Gender vs. Computer

Probability of Female without laptop= (Total female student / Total student)\*((female with desktop + female with tablet)/ total female student)

***Probability\_female\_without\_laptop = (33 / 62)\*((2 + 2) / 33) = 0.06451612903225806***

**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?**

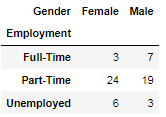


Table 2.5.1 Gender vs. Employment

Probability of male or full-time Employment = (Total Male student + Total student – Total male student having full-time employment / total student)

***Probability\_male\_or\_full-time\_employment = (29 + 10 – 7 / 62) = 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.**



Table 2.5.2 Gender vs. Major

Probability of male or full-time Employment = (Total Male student + Total student – Total male student having full-time employment / total student)

***Probability\_male\_or\_full-time\_employment = (29 + 10 – 7 / 62) = 0.5161290322580645***

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?

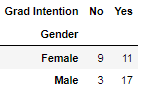


Table 2.6 Gender vs. Grad Intention

Inference from the above contingency table, ***being female and having Grad Intention are not two independent variables.***

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?**

Probability of student GPA less than 3 = (Total student with GPA less than 3 / total student)

***Probability\_student\_GPA\_LT\_3 = 17 / 62 = 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.**

Probability of Male earns greater than 50 = (Total male student and earns greater than 50 / total student)

***Probability\_Male\_and\_earns\_GT\_50 = 14 / 62 = 0.22580645161290322***

Probability of Female earns greater than 50 = (Total female student and earns greater than 50 / total student)

***Probability\_Female\_and\_earns\_GT\_50 = 18 / 62 = 0.2903225806451613***

**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.**

|  |  |  |  |
| --- | --- | --- | --- |
| Column  Name | Mean, Mode, Median, Shapiro test Result | Normal Distribution plot | Comment follows Normal Distribution(N.D) |
| *GPA* | ***mean*** 3.129032258064516  ***Mode*** - 0 3.00  1 3.10  2 3.40  dtype: float64  ***Median*** - 3.1500000000000004  ***Shapiro test***  ***P- value*** - 0.11204058676958084 | Fig 2.8.1. GPA bell curve | Mean, Mode and Median are approximately equal and from the bell curve and from the Shapiro test, ***GPA follows the normal distribution*** |
| Salary | ***mean*** - 48.54838709677419  ***Mode*** - 0 40.00  dtype: float64  ***Median*** - 50.0  ***Shapiro test***  ***P- value*** - 0.028000956401228905 | Fig 2.8.2. Salary bell curve | Mean, Mode and Median are approximately equal and from the Shapiro test, ***salary does not*** ***follows*** ***the*** ***normal distribution*** |
| Spending | ***mean*** - 482.01612903225805  ***Mode*** - 0 500  dtype: int64  ***Median*** - 500.0  ***Shapiro test***  ***P- value*** - 1.6854661225806922e-05 | Fig 2.8.3. Spending bell curve | Mean, Mode and Median are approximately equal and from the Shapiro test, ***Spending does not follows the normal distribution*** |
| Text Messages | ***mean*** - 246.20967741935485  ***mode*** - 0 300  dtype: int64  ***median*** - 200.0  ***Shapiro test***  ***P- value*** - 4.324040673964191e-06 | Fig 2.8.4. Text Messages bell curve | Mean, Mode and Median are approximately equal and from the Shapiro test, ***Text Messages does not follows the normal distribution*** |

# Problem – 3

*Summary*

The two companies has decided to gather data about the ABC asphalt Shingles. The dataset consists of amount of moisture contains in the shingles. In this problem statement, we will explore the different questions that has been created from the customer feedback about the moisture content present in the shingles is less than 0.35 pounds per 100 square feet and analysis the response from the hypothesis testing.

*Introduction*

The purpose of this exercise is to explore the dataset. The exploratory data analysis of this dataset is the shingles contain moisture content. The dataset consists of shingles with the moisture content. This assignment helps in exploring the Shingles manufacturing company whether shingles contain moisture content is less than 0.35 pounds 100 square feet with the hypothesis testing.

*Data Description*

1. A: Company A with moisture content in shingles.

2. B: Company B with moisture content in shingles.

*Sample of the dataset:*

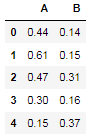


Table 3.Dataset Sample

Dataset has 2 variables, A company and B company moisture content in Shingles. Based on the characteristic moisture content in Shingles A & B company is defined.

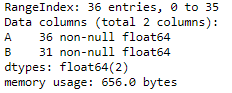
*Exploratory Data Analysis*

*Let us check the types of variables in the data frame.*



There are total 36 measurements (in pounds per 100 square feet) for A shingles and 31 measurements (in pounds per 100 square feet) for B shingles. 2 columns are of float (Decimal value).

*Check for missing values in the dataset:*



From this, it is clear that there are no null values present in the dataset. Column A has 36 measurements and 31 measurements

### 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.

For the A shingles, the null and alternative hypothesis to test whether the mean moisture content is less than 0.35 pound per 100 square feet is given:

H0 > 0.35

H1 < 0.35

**Null Hypothesis:** Mean moisture content greater than or equal to 0.35 moisture content.

**Alternate Hypothesis:** Mean moisture content lesser than 0.35 moisture content at 95% confidence level.

The level of significance – 0.05

One sample t test is used for finding the moisture content in company A and company B

**For Company A**



From this, we get T statistics and P value for Company A.



**For Company B**



From this, we get T statistics and P value for Company B.



From the T statistics, At 95% confidence level, where Company A has the moisture level is lesser than 0.35 pounds per 100 square feet. Whereas, Company B has a higher moisture content in Shingles.

### 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?

For the A shingles, the null and alternative hypothesis to test whether the mean moisture content is less than 0.35 pound per 100 square feet is given:

H0 : mu\_A == mu\_B

H1 : mu\_A != mu\_B

**Null Hypothesis:** Mean moisture content in company A Shingles is equal to moisture content in company B Shingles.

**Alternate Hypothesis:** Mean moisture content in company A Shingles is not equal to moisture content in company B Shingles.

The level of significance – 0.05

Two sample t test is used for finding the moisture content in company A and company B,

The output from two sample t test, we get



From this, we get T statistics and P value.



From this, we conclude that, Both Company A and Company B mean and moisture content in shingles is not equal.

**Assumptions,**

The Population to be normal and the variance of the two companies needs to be equal. If these assumptions were not met, then needs to proceed with alternate hypothesis testing.