**Business Report on**

**Statistical Methods for Decision Making**

***Submitted to***



**Great Learning Olympus**

***Prepared by***

**Mrinmoy Majumdar**

**June-D Batch**

**Post Graduate Program in Data Science & Business Analytics**

**From**

****

UT Austin

**August, 2021**

**CONTENT**

Case study 1- Wholesale Customers Analysis 5

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? 9

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 11

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? 14

1.4 Are there any outliers in the data? Backup your answer with a suitable

Plot /technique with the help of detailed comments 15

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 16

Case Study 2- Clear Mountain State University (CMSU) 17

2.1. For this data, construct the following contingency tables 19

(Keep Gender as row variable)

2.1.1. Gender and Major 19

2.1.2. Gender and Grad Intention 19

2.1.3. Gender and Employment 20

2.1.4. Gender and Computer 20

2.2. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question: 20

2.2.1. What is the probability that a randomly selected CMSU student will be male? 20

2.2.2. What is the probability that a randomly selected CMSU student will be female? 20

2.3. Assume that the sample is representative of the population of CMSU. Based

on the data, answer the following question: 21

2.3.1. Find the conditional probability of different majors among the male students in CMSU 21

2.3.2 Find the conditional probability of different majors among the female students of CMSU 22

2.4. Assume that the sample is a representative of the population of CMSU

Based on the data, answer the following question: 22

2.4.1. Find the probability that a randomly chosen student is a male and intends to

Graduate 22

2.4.2 Find the probability that a randomly selected student is a female and does not

have a laptop 22

2.5. Assume that the sample is representative of the population of CMSU. Based

on the data, answer the following question: 23

2.5.1. Find the probability that a randomly chosen student is either a male or has

full-time employment? 23

2.5.2. Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management 23

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? 24

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 25

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

than 3? 25

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 25

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 for this whole problem 26

Case Study 3- A&B Shingles 29

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 30

* 1. Do you think that the population mean for shingles A and B are equal? For 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? 31

Case Study 1 - Wholesale Customers Analysis

**Problem Statement:**

A wholesale distributor operating in different regions of Portugal has information on annual spending of several items in their stores across different regions and channels. The data consists of 440 large retailers’ annual spending on 6 different varieties of products in 3 different regions (Lisbon, Oporto, Other) and across different sales channel (Hotel, Retail).

**Data Description:**

The data at hand contains annual spending on items characterized by certain attributes.

# Domain:

Business

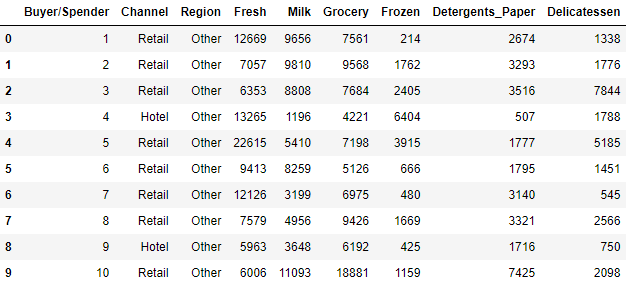
# Context:

Collection of annual spending information of customers is the driving force for most businesses. In the case of a wholesale distribution business, attributes of spending amounts according to different products and across various channels and regions like the ones mentioned below can be crucial in making business decisions. Hence, knowing to explore and generate value out of such data can be an invaluable skill to have.

# Attribute Information:

* Buyer/Spender : Customer ID of customers buying products from wholesale market
* Channel : Categories of channels from which items are bought by customers
* Region : Various regions in which wholesale business products are on sale
* Fresh : Annual spending on Fresh items
* Milk : Annual spending on Milk items
* Grocery : Annual spending on Grocery items
* Frozen : Annual spending on Frozen items
* Detergents\_Paper : Annual spending on Detergents and Paper items
* Delicatessen : Annual spending on Delicatessen products

**Displaying Wholesale Data:**

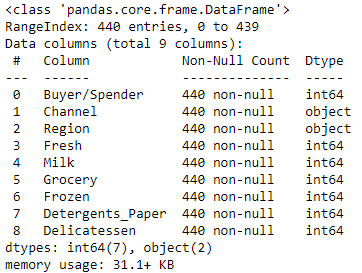


**Table 1: Top 10 rows of wholesale data Frame**

## **Basic EDA:**

* Find the shape of the data, data type of individual columns

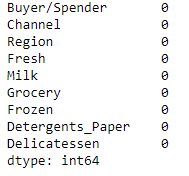
(440,9) – The data set contains 440 observations of data and 9 variables.



**Image 1: Information on wholesale dataset**

### The data has 440 instances with 9 attributes. 7 integer type and 2 object type (Strings in the column).

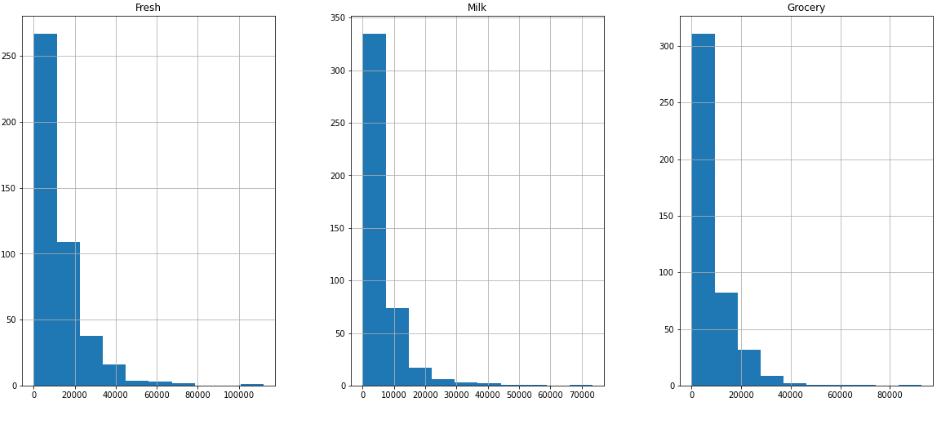
* Check the presence of missing values

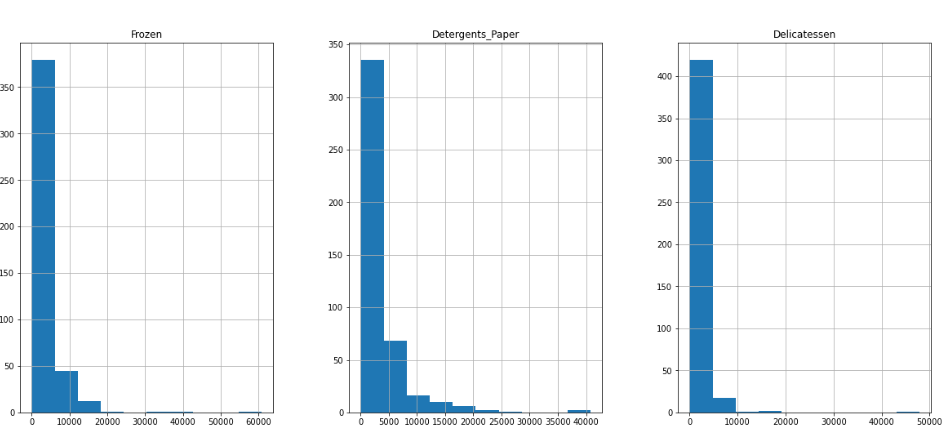


**Image 2: Checking null values in data**

### There are no null values in any of the columns.

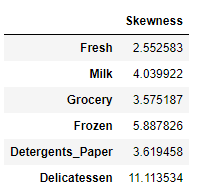
* Find the distribution of numerical columns and the associated skewness





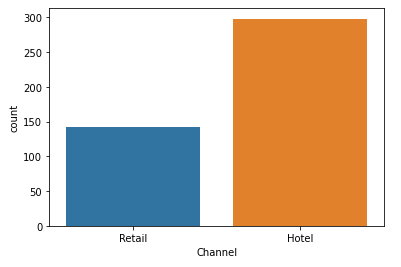
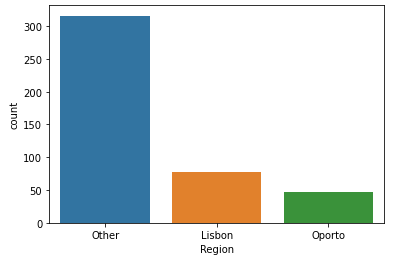
**Image 3: Histogram plot on all numerical columns of wholesale dataset**

Observation from histograms - All the numerical variables are right-skewed.



**Image 5: Table showing skewness of data**

* Skew of Fresh is the lowest amongst all numerical columns.
* Delicatessen have highest skewness.
* Distribution of categorical columns



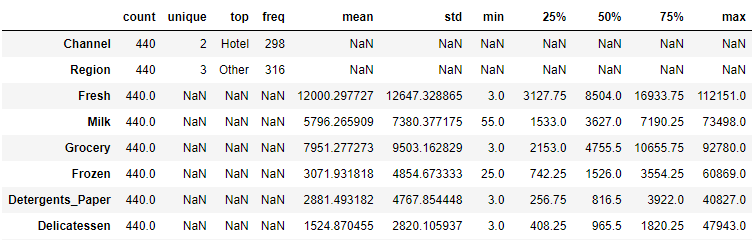
**Image 6: Count Plot of Region and Channel**

* There are a lot more Hotel channels than Retail channels in the data.
* Most instances of other regions compared to Lisbon and Oporto regions.

**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 is a type of statistic that is concerned with data summarization, graphs or charts and tables. It is one of the key pillars of Business Analytics. Summary measures include central tendency, mean, standard deviation, histogram, boxplots, scatter diagrams etc.**



**Image 7: Dataset Summary**

Data looks legit as all the statistics seem reasonable –There are 2 unique channels, 3 unique regions and 6 different items on sale. Unique, top and frequency cannot be calculated for numerical variables and similarly five number summary (min, 25%, 50%, 75% and max) standard deviation and mean cannot be measured for categorical variables which explains why dataset contains Nan values. Customers visiting the hotel channel and other region have spent the most on items as indicated by top. It is also clear from the dataset looking at max column that customers have maximum spending on item “Fresh” and minimum spending on “Detergents\_Paper”.

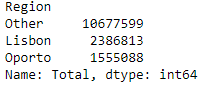


**Image 8: Highest earning channel**



**Image 9: Lowest earning channel**

As seen from the above outputs in jupyter notebook, channel with most spending is “Hotel” and least spending is “Retail”.

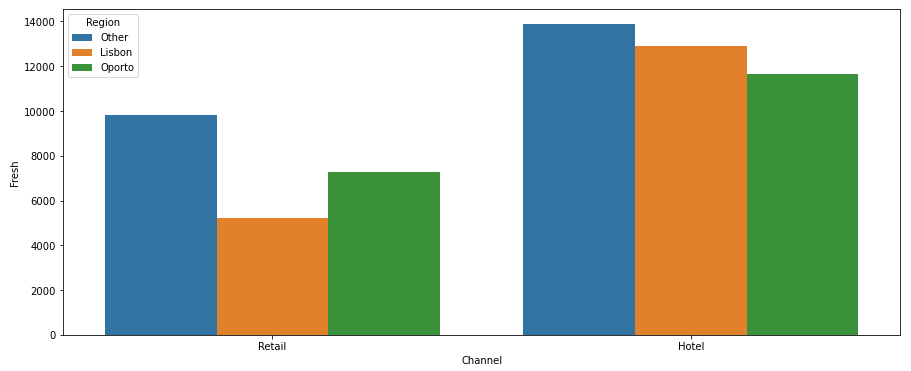


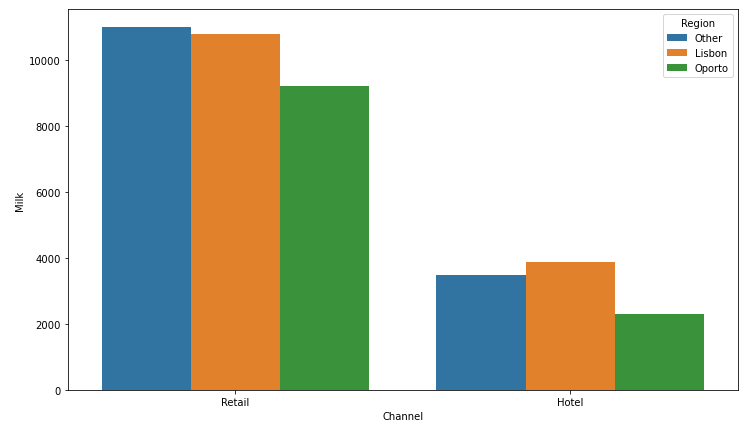
**Image 10: Maximum and minimum spending regions**

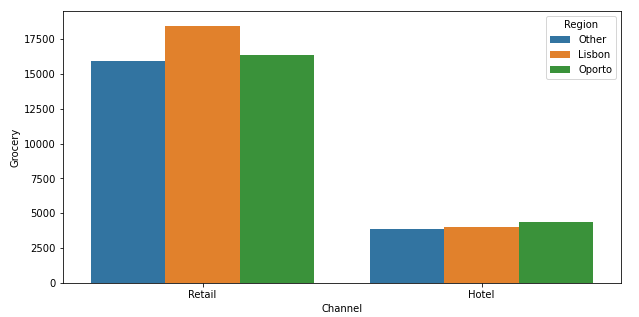
Inference from the above output is that region with most spending is “Other” and least spending is “Oporto”.

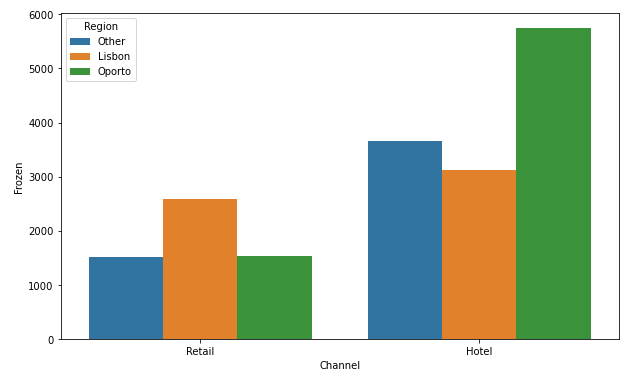
**1.2 There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.**

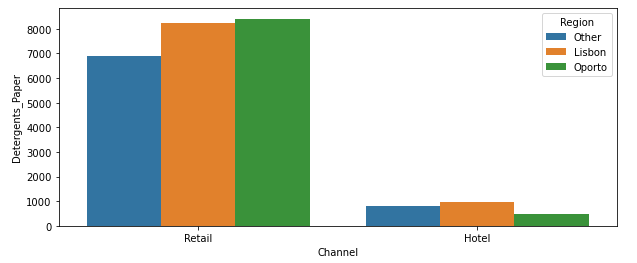
The behaviour of various items across different channels and regions can be explained with the help of bar plots as shown below –

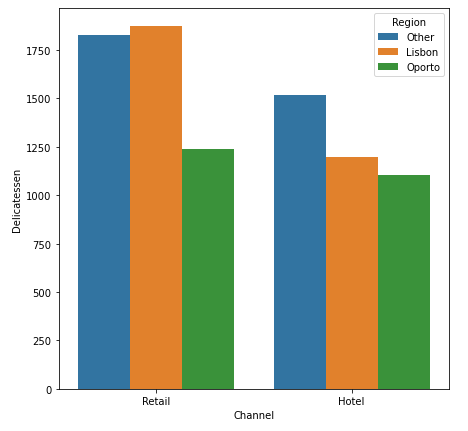












**Image 11: Bar Plot of different products across regions and channels**

Observations-

* The item “Fresh” has highest spending from “Hotel” channel and “Other” region and minimum spending from “Retail” channel and “Lisbon” region.
* The item “Milk” has highest spending from “Retail” channel and “Other” region and minimum spending from “Hotel” channel and “Oporto” region.
* The item “Grocery” has highest spending from “Retail” channel and “Lisbon” region and minimum spending from “Hotel” channel and “Other” region.
* The item “Frozen” has highest spending from “Hotel” channel and “Oporto” region and minimum spending from “Retail” channel and “Other” region.
* The item “Detergents\_Paper” has highest spending from “Retail” channel and “Oporto” region and minimum spending from “Hotel” channel and “Oporto” region
* The item “Delicatessen” has highest spending from “Retail” channel and “Lisbon” region and minimum spending from “Hotel” channel and “Oporto” region.

**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?**

**The descriptive measure of variability to find out the items with most and least inconsistent behaviour is Coefficient of Variation (CV).**

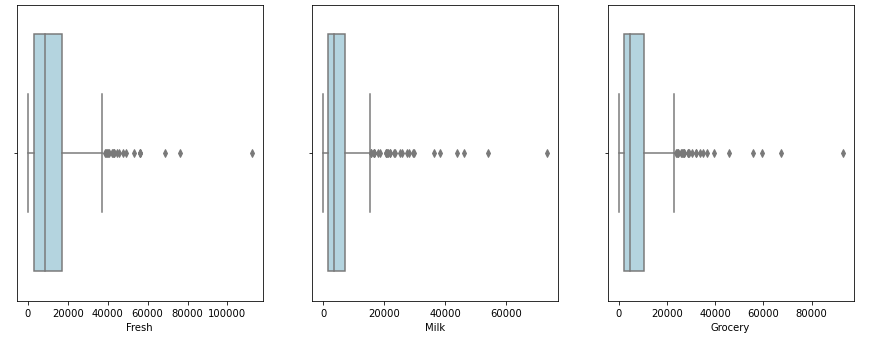
|  |  |
| --- | --- |
| **Items** | **Coefficient of Variation (CV)** |
| Fresh | 1.0527196084948245 |
| Milk | 1.2718508307424503 |
| Grocery | 1.193815447749267 |
| Frozen | 1.5785355298607762 |
| Detergents\_Paper | 1.6527657881041729 |
| Delicatessen | 1.8473041039189306 |

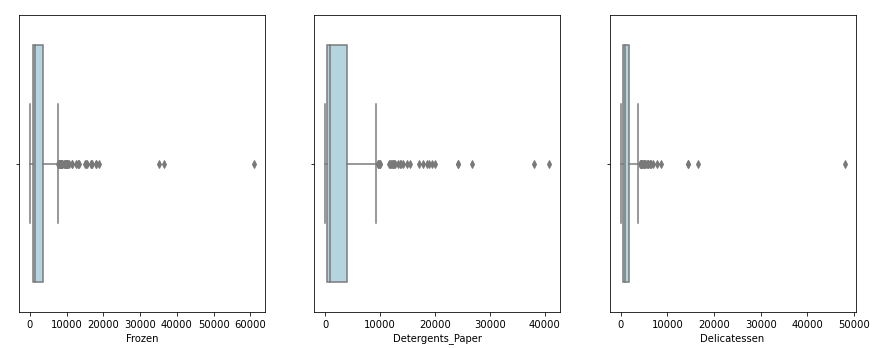
**Table 2: CV of different items**

“Delicatessen” item shows most inconsistent behaviour as CV is high and “Fresh” item shows least inconsistent behaviour as CV is low which is depicted by the above table.

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

Box plot will be plotted to check for outliers.





**Image 12: Box plot of different products**

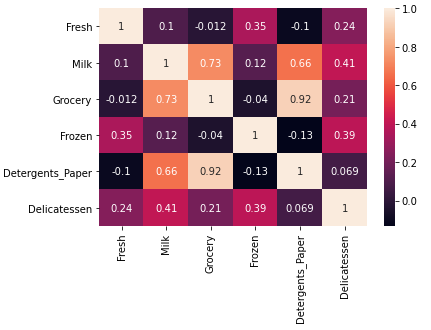
Observations –

* Box plot of 6 different items confirms the presence of many outliers in the data.
* All the items contains outliers as shown above.

**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**

**Recommendations for the business on the basis of data analysis –**

* Retail channels are lesser in number than hotel channels, hence hotel channels have more customers and higher sale of products than retail channel which should be equal. So, wholesale distributor should setup more retail channels.
* Customers spend the least on items found in “Oporto” region as its overall headcount is minimum. So, “Oporto” regions need more focus to increase sales.
* Customers are more likely to buy “Fresh” and “Grocery” irrespective of channels and regions and similarly least likely to buy “Delicatessen” as seen from average highest spending across all regions and channels. Therefore, wholesale distributor should provide discounts to attract customers to buy items other than “Fresh” and “Grocery”.
* There are inconsistencies in customer spending of different products which is clear from coefficient of variation and heat map, efforts should be made by distributors to eliminate such inconsistency.



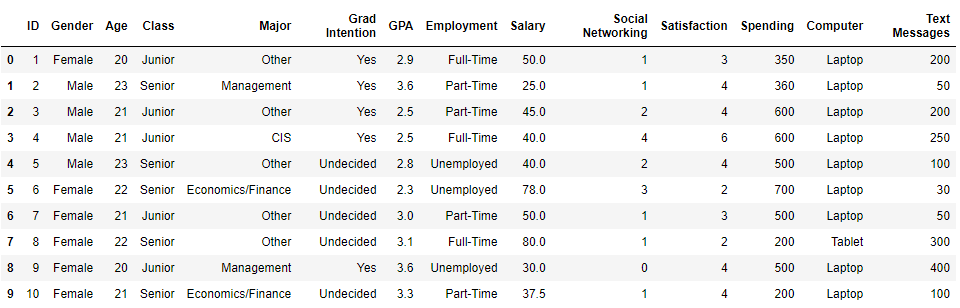
**Image 12: Heat map of various products**

**Case Study 2-Clear Mountain State University (CMSU) Survey**

**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).

Dataset Snapshot –



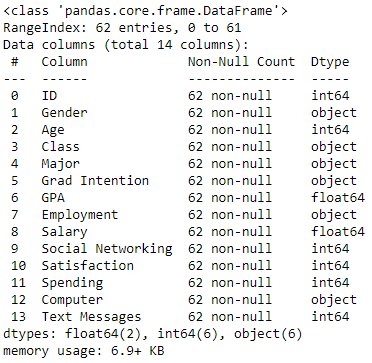
**Image 13: Head of CMSU dataset**

Basic EDA-

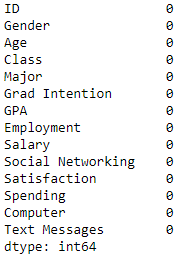
* (62, 14)-The dataset contains 62 observations of data and 14 variables.

### The data has 62 instances with 14 attributes. 6 integer type, 2 float types and 6 object type (Strings in the column).

* There are no null values in the given dataset.

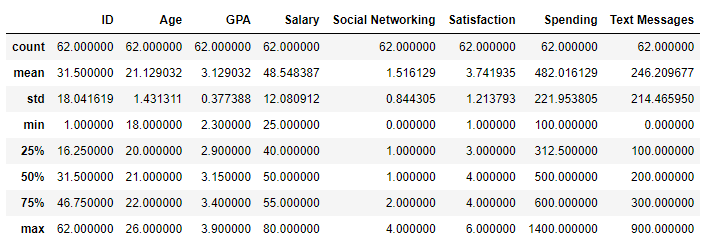


**Image 14: Information on CMSU datatypes**



**Image 15: Count of null values in each CMSU column**

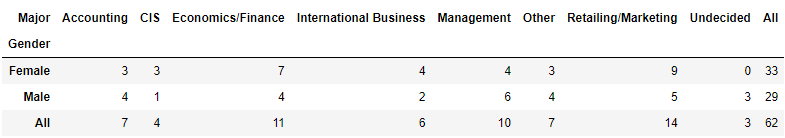
Summary of Dataset



**Image 16: Descriptive summary of CMSU data**

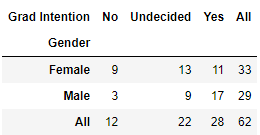
**2.1. For this data, construct the following contingency tables (Keep Gender as row variable)**

**2.1.1. Gender and Major**



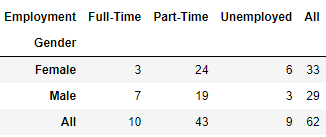
**Image 17: Contingency table of Gender and Major**

**2.1.2. Gender and Grad Intention**



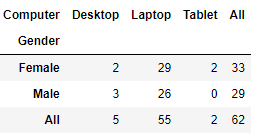
**Image 18: Contingency table on gender and grad intention**

**2.1.3. Gender and Employment**



**Image 19: Contingency table on gender and employment**

**2.1.4. Gender and Computer**



**Image 20: Gender and employment contingency table**

**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?**

**There are a total of 62 students in the university out of which 29 are male.**

**Probability that a randomly selected CMSU student is male is (29/62) which is approximately 0.468 rounded off to 3 decimal places.**

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

**There are 33 females out of 62 students in the university. Hence, probability that a randomly selected CMSU student is female = (33/62) which is 0.532 rounded off to 3 decimals.**

**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**

Among Male Students:

Probability of opting for Accounting-> 0.138

Probability of opting for CIS-> 0.034

Probability of opting for Economics/Finance-> 0.138

Probability of opting for International Business-> 0.069

Probability of opting for Management-> 0.207

Probability of opting for Other-> 0.138

Probability of opting for Retailing/Marketing-> 0.172

Probability undecided-> 0.103

The above code output is based on the contingency table between gender and major, can be mathematically explained with conditional probability formula as follows:

P (Major | Male) = P (Major ∩ Male)/ P (Male)

Total no of male students = 29

Total no of male Accounting students = 4

P (Accounting | Male) = 4/29 = 0.138

Total no of male CIS students = 1

P (CIS | Male) = 1/29 = 0.034

Total no of male Economics or Finance students = 4

P (Economics/Finance | Male) = 4/29 = 0.138

Male student count in International Business = 2

P (International Business | Male) = 2/29 = 0.069

Male student count in Management = 6

P (Management | Male) = 6/29 = 0.207

Count of male students in Others = 5

P (Others | Male) = 5/29 = 0.138

Male student count in Retail or Marketing = 4

P (Retail/Marketing | Male) = 4/29 = 0.172

Total count of undecided male students = 3

P (Undecided | Male) = 3/29 = 0.103

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

**Formula -** P (Major | Female) = P (Major ∩ Female)/ P (Female)

Among Female Students:

Probability of opting for Accounting- 0.091

Probability of opting for CIS- 0.091

Probability of opting for Economics/Finance- 0.212

Probability of opting for International Business- 0.121

Probability of opting for Management- 0.121

Probability of opting for Other- 0.091

Probability of opting for Retailing/Marketing- 0.273

Probability undecided- 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.**

Total number of male students = 29

Total number of male students who intends to graduate = 17

Probability of male student intending to graduate = (17/29) = 0.586

**2.4.2 Find the probability that a randomly selected student is a female and**

**Does NOT have a laptop.**

Total number of female students = 33

Total number of female students with no laptop = 4

Applying formula of conditional probability -

P (No Laptop ∩ Female)

= P (No Laptop | Female) x P (Female)

= ((4/33) \* (33/62))

= 0.065

**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 (A or B) = P (A) + P (B) – P (A and B) [Events are not mutually exclusive]

P (Male | Total Students or Fulltime Employment | Total Students)

= (P (Male | Total students) + P (Fulltime employment | Total Students) -

P (Male | Total Students ∩ Fulltime employment | Total Students)

= (29/62) + (10/62) – (7/62)

= 0.516

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

P (A or B) = P (A) + P (B) [Mutually Exclusive events]

P (Female | Total Students or International Business/Management | Total Students)

= P (Female | Total Students) + P (International Business/Management | Total Students)

= (4/33) + (4/33)

= 0.242

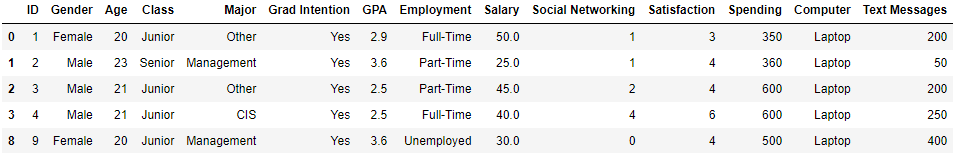
**2.6. Construct a contingency table of Gender and Intent to Graduate at**

**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?**

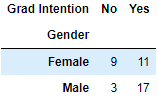
**After dropping undecided students from grad intention column, the sample dataset**

**is given below:**



**Image 21: Sample dataset after dropping unrequired columns**

Contingency table of gender and intent to graduate at 2 levels is shown below –



**Image 22: Contingency table of Gender and Grad Intention at 2 levels**

For 2 events to be independent, following condition needs to be satisfied:

P (A ∩ B) = P (A) \* P (B)

P (Grad Intention ∩ Female) = P (Grad Intention) \* P (Female)

P (Female) = 20/40 = 0.5

P (Grad Intention) = 28/40 = 0.7

P (Grad Intention) \* P (Female) = 0.7 \* 0.5 = 0.035

P (Grad Intention ∩ Female) = 11/40 = 0.275

This is not independent events as probability multiplication of both events is not equal to combined event, so being a female and graduate intention are not independent events.

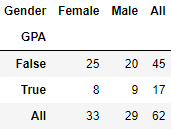
**2.7. Note that there are four numerical (continuous) variables in the data**

**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?**

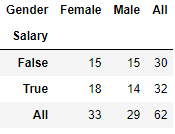


**Image 23: Contingency table of Gender and GPA less than 3**

The probability that a randomly chosen student has GPA less than 3

= 17/62 = 0.274

**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.**



**Image 24: Contingency table of Gender and Salary > = 50**

Conditional probability that a randomly selected male earns 50 or more

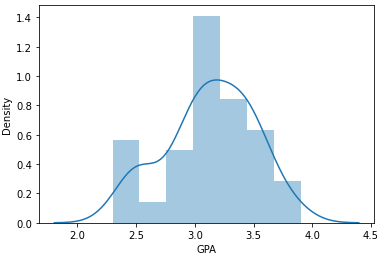
= 14 / 29 = 0.482

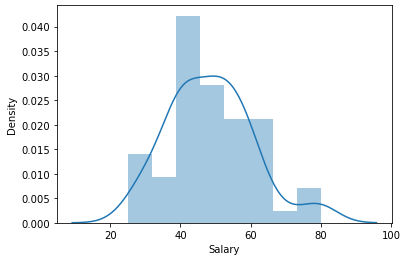
Conditional probability that a randomly selected female earns 50 or more

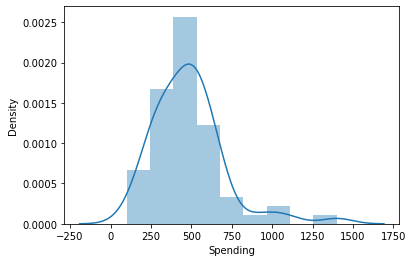
= 18 / 33 = 0.545

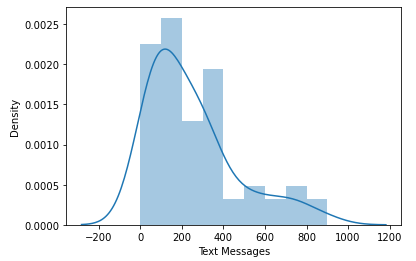
**2.8. Note that there are four numerical (continuous) variables in the data**

**set, GPA, Salary, Spending, and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarizing your conclusions.**









**Image 25: Distribution plot for continuous variables**

Conducting Shapiro test on various numerical variables in survey dataset,

* GPA follows Normal Distribution as p value is greater than level of significance.
* Spending, Salary and Text Messages do not follow normal distribution as their p values are very less.

**Shapiro Test results –**

(0.9685361981391907, 0.11204058676958084) 🡪 GPA column

(0.9565856456756592, 0.028000956401228905) 🡪 Salary Column

(0.8777452111244202, 1.6854661225806922e-05) 🡪 Spending Column

(0.8594191074371338, 4.324040673964191e-06) 🡪 Text Messages Column

First value of the tuple is statistic and second value is p value assuming level of significance as 0.05.

**Conclusion –**

1. **As per exploratory data analysis, the students studying in CMSU are aged between 21 and 26 earning salaries ranging from 48.5 to 80 and GPA of 3.12 to 3.90.**
2. **The count of female students is higher than male students in the university.**
3. **From contingency table analysis, students are more involved in part time employment possessing more laptops than desktop and tablets.**
4. **Female students are more likely to take up Retailing/Marketing whereas male students prefer Management more than other majors.**
5. **There are 3 male students who are undecided about taking up majors whereas female students are clear about their choice of majors.**
6. **The university should include new majors that the undecided male students may be interested in order to attract more male students to graduate.**

Case Study-3:A&B Shingles

Problem Statement

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.

Dataset Sample



**Image 26: Shingles dataset snapshot**

**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.**

### Step 1: Define null and alternative hypothesis

In testing whether the mean moisture content of both the shingles are within permissible limits, the null hypothesis states that the mean moisture content is less than or equal to 0.35.The alternative hypothesis states that the mean moisture content is greater than 0.35.

* H0:mu<=0.35
* HA: mu>0.35

### Step 2: Decide the significance level

Here we select α= 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.
* The samples are large samples with n1=36 and n2=31. So you use the t distribution and the tSTAT test statistic for two samples separately.

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

**Scipy.stats.ttest\_1samp calculates the t test for the mean of one sample given the sample observations and the expected value in the null hypothesis. This function returns t statistic and the two-tailed p value.**

**For Shingle A-**



**Image 27: T-test output for Shingle A**

For Shingle B –



**Image 28: T-test output for Shingle B**

### Step 5: Decide to reject or accept null hypothesis

For shingle A, p value is greater than 0.05 and hence we fail to reject the null hypothesis. So, we do not have enough evidence to prove that the mean moisture content is less than or equal to 0.35.

For shingle B, p value is less than 0.05 and hence we reject the null hypothesis. So, we have enough evidence to prove that mean moisture content is not less than or equal to 0.35.

**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?**

### Step 1: Define null and alternative hypothesis

In testing whether the population mean are same in both the shingles, the null hypothesis states that the population mean of the shingles are the same, μA equals μB. The alternative hypothesis states that the population mean of both shingles are different, μA is not equal to μB.

* H0: μA - μB = 0 i.e μA = μB
* HA: μA - μB ≠ 0 i.e.  μA ≠ μB

### Step 2: Decide the significance level

Here we select α = 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 shingles are different.
* So you use the t distribution and the tSTAT test statistic for two sample unpaired test.

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

**We use the scipy.stats.ttest\_ind to calculate the t-test for the means of TWO INDEPENDENT samples of shingles given the two sample observations. This function returns t statistic and two-tailed p value.**

**This is a two-sided test for the null hypothesis that 2 independent samples have identical average (expected) values. This test assumes that the populations have identical variances.**

For this exercise, we are going to first assume that the variance is equal and then compute the necessary statistical values.



**Image 29: Output for Independent Samples T-test**

### Step 5: Decide to reject or accept null hypothesis

We have enough evidence to accept the null hypothesis .We conclude that the

mean population of the two shingles are same.

Assumptions –

* The samples from population must be independent of one another.
* The population from which the samples are taken must be normally

distributed.

* Data from each group is obtained by random sampling from the population.