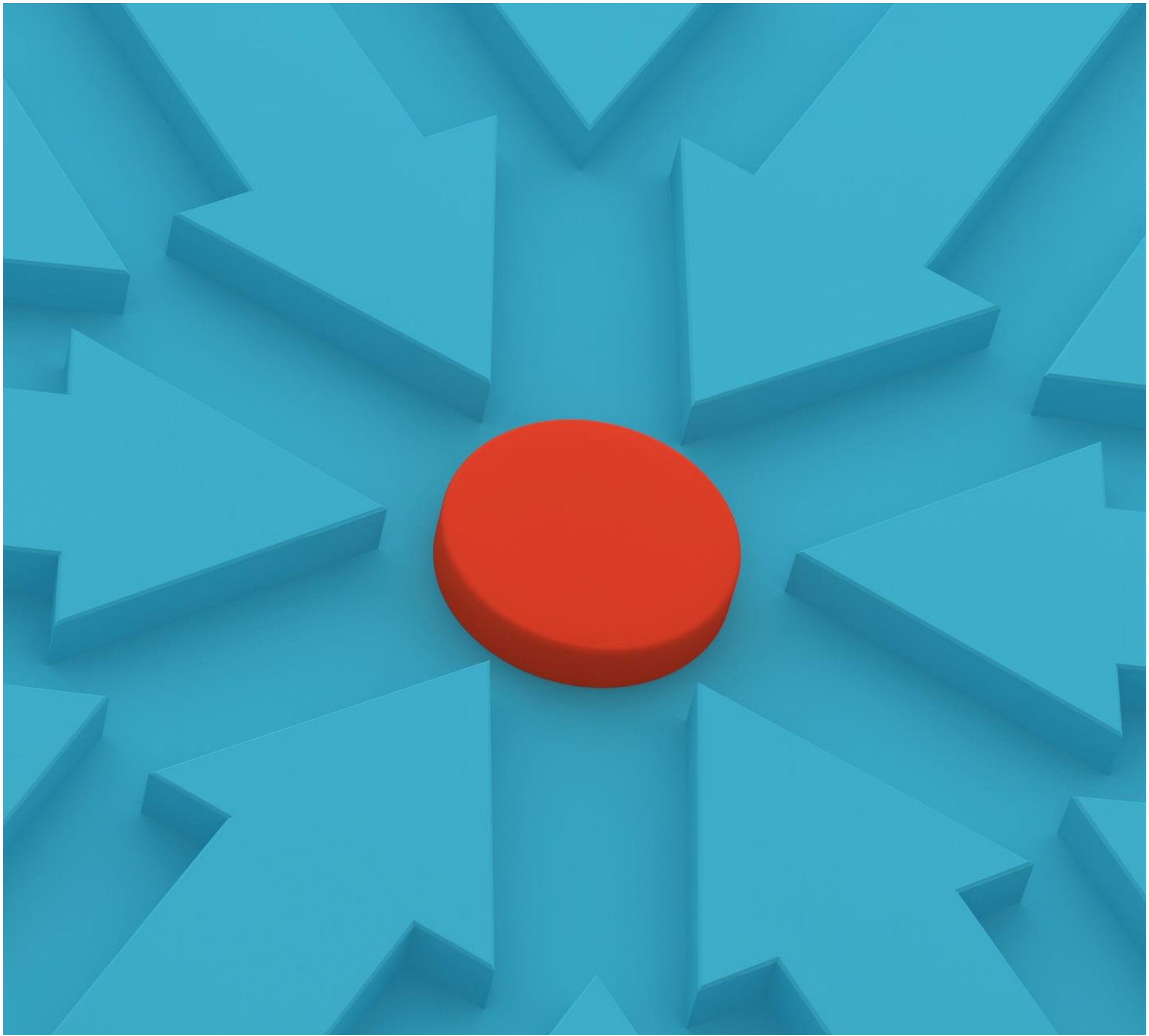


STATISTICAL METHODS FOR DECISION MAKING

PROJECT ANSWER REPORT

SUNDAR RAM S
PGPDSBA
ONLINE DEC_C 2021
27-FEB-2022

TABLE OF CONTENTS

Q 2.5.1

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

Q 2.5.2

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

Q 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 graduate intention and being female are independent events?

Q 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

Q 2.7.1

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

Q 2.7.2

Find conditional probability that a randomly selected male earns 50 or more. Find conditional probability that a randomly selected female earns 50 or more

Q 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.

Problem 3

Moisture Measurements of manufacturers of ABC Asphalt Shingles

Problem Statement

Q 3.1

Do you think there is evidence that mean moisture contents in both types of shingles are within the permissible limits? State your conclusions clearly showing all steps.

Q 3.2

Do you think that the population means 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?

Executive Summary

This report is based on the data of three different data sets – Annual spending of wholesale distributor, survey of Students News Service at Clear Mountain State University and A & B type Asphalt Shingles' moisture measurements of manufacturers of ABC Asphalt Shingles. Each of the dataset has a set of defined problems that are explored and solutions are recorded.

Introduction

The purpose of this report is to perform an exploratory data analysis and provide a detailed explanation on approach used, record inferences, insights and provide suitable business solutions. The techniques of Fundamentals of Business statistics, Inferential statistics and Hypothesis testing are leveraged in this exercise.

Problem 1 – Wholesale Customer Analysis

Problem Statement

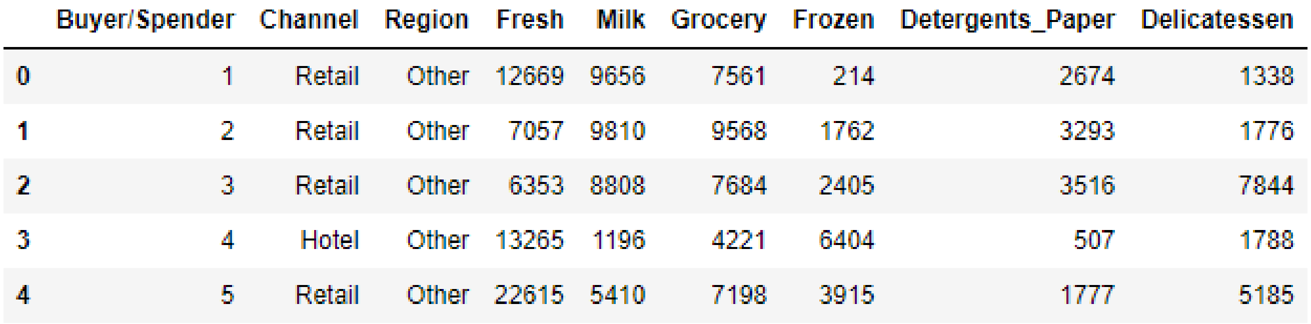
A wholesale distributor operating in different regions of Portugal has information on annual spending of several items in their stores across different regions and channels. The data consists of 440 large retailers' annual spending on 6 different varieties of products in 3 different regions (Lisbon,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

Dataset Description

1	Buyer/ Spender:	The ID of the buyer/spender
2	Channel:	The sales channel to which the store belongs (Hotel/ Retail)
3	Region:	The region in which the store is located (Lisbon/ Oporto/ Other
4	Fresh:	Annual Spending of each buyer on Fresh Products
5	Milk:	Annual Spending of each buyer on Milk
6	Grocery:	Annual Spending of each buyer on Grocery
7	Frozen:	Annual Spending of each buyer on Frozen Products
8	Detergents_Paper:	Annual Spending of each buyer on Detergents_Paper

9	Delicatessen:	Annual Spending of each buyer on Delicatessen

Sample of the Dataset



	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
0	1	Retail	Other	12669	9656	7561	214	2674	1338
1	2	Retail	Other	7057	9810	9568	1762	3293	1776
2	3	Retail	Other	6353	8808	7684	2405	3516	7844
3	4	Hotel	Other	13265	1196	4221	6404	507	1788
4	5	Retail	Other	22615	5410	7198	3915	1777	5185

Table 1.1 – Sample of the Dataset

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).

Exploratory Data Analysis – Problem 1

Variable Type

Buyer/ Spender:	int64
Channel:	object
Region:	object
Fresh:	int64
Milk:	int64
Grocery:	int64
Frozen:	int64
Detergents_Paper:	int64
Delicatessen:	int64

There are 440 rows and 9 columns in the dataset with 2 object type and 7 integer type data.

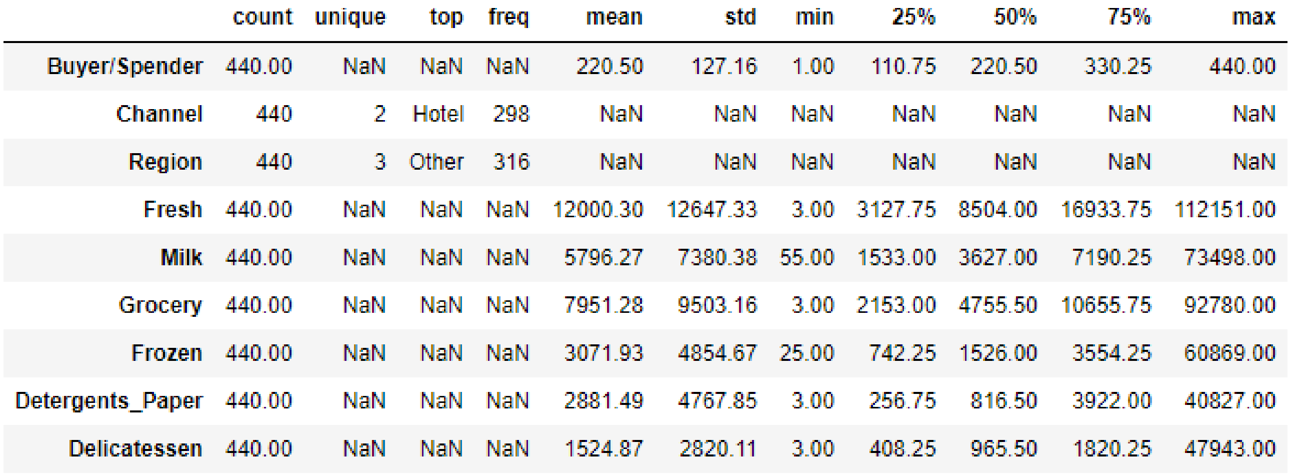
Check for missing values in the dataset

Buyer/ Spender:	440 non - null
Channel:	440 non - null
Region:	440 non - null
Fresh:	440 non - null
Milk:	440 non - null
Grocery:	440 non - null
Frozen:	440 non - null
Detergents_Paper:	440 non - null
Delicatessen:	440 non - null

It is evident that there are no missing values in the entire dataset

Q 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 us to summarize and understand the samples in the dataset. The categorical data is described in terms of count (number of entries), unique values, top occurrence and their frequency, while continuous arithmetic data are described in terms of their count (number of entries), mean, standard deviation, minimum value, maximum value, 25th percentile, median & 75th percentile.



	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Buyer/Spender	440.00	NaN	NaN	NaN	220.50	127.16	1.00	110.75	220.50	330.25	440.00
Channel	440	2	Hotel	298	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Region	440	3	Other	316	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Fresh	440.00	NaN	NaN	NaN	12000.30	12647.33	3.00	3127.75	8504.00	16933.75	112151.00
Milk	440.00	NaN	NaN	NaN	5796.27	7380.38	55.00	1533.00	3627.00	7190.25	73498.00
Grocery	440.00	NaN	NaN	NaN	7951.28	9503.16	3.00	2153.00	4755.50	10655.75	92780.00
Frozen	440.00	NaN	NaN	NaN	3071.93	4854.67	25.00	742.25	1526.00	3554.25	60869.00
Detergents_Paper	440.00	NaN	NaN	NaN	2881.49	4767.85	3.00	256.75	816.50	3922.00	40827.00
Delicatessen	440.00	NaN	NaN	NaN	1524.87	2820.11	3.00	408.25	965.50	1820.25	47943.00

Table 1.2 -Summary of the data

From the above summary table 1.2, we can infer that there are 440 entries. Out of the 440 entries, there are 298 entries from the channel Hotel and 316 entries with ‘other’ mentioned as region. We can also get an idea on the mean annual spending of the product varieties. Customers tend to spend more on the Fresh Food products with a mean spending of 12000.30, and they tend to spend the least for Delicatessen products with a mean spending value of 1524.87.

To find out the region and channel that spent the Most/Least, we have summed up the total spending of each customer in a new column and employed the groupby() functionality in python. On grouping the dataset based on region, it is evident that **“Other”** region spent the most with a value of 1,06,77,599 and **“Oporto”** spent the least with a value of 15,55,088. On grouping the dataset based on channel, it is evident that “**Hotel**” spent the most with a value of 79,99,569 and **“Retail”** spent the least with a value of 66,19,931.

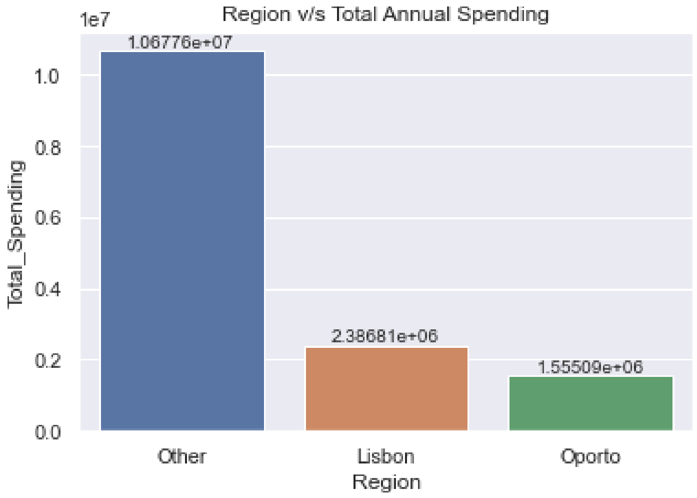


Figure 1.1 – Region v/s Total Annual Spending

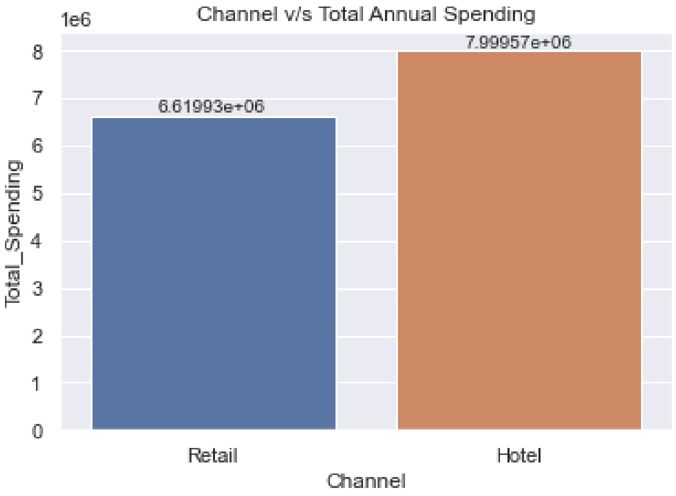
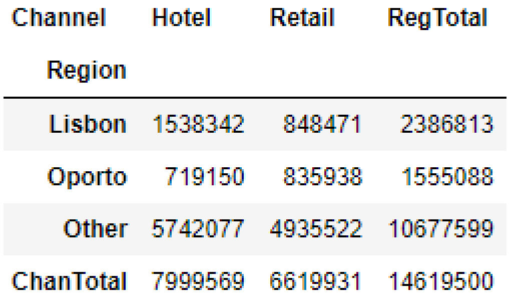


Figure 1.2 - Channel v/s Total Annual Spending

Alternatively, this can be analyzed and inferred using the crosstab() feature in Python through which a Pivot table can be developed.



Channel	Hotel	Retail	RegTotal
Region			
Lisbon	1538342	848471	2386813
Oporto	719150	835938	1555088
Other	5742077	4935522	10677599
ChanTotal	7999569	6619931	14619500

Figure 1.3 – Pivot table using crosstab()

Q 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 mean spending of Fresh and Frozen varieties is greater in “Hotel” channels across all the regions while the mean spending of Milk, Grocery, Detergents_paper & Delicatessen are grater in “Retail” channels across all the regions
- The maximum annual spending is done on Fresh products in “Hotel” channels of other regions, followed by hotels of Lisbon and Oporto. The maximum annual spending on Grocery products is done in the Retail channels of other regions, followed by retail channels of Oporto and Lisbon
- The product varieties of Fresh, Grocery, Detergents_paper & Delicatessen have the same minimum value in Annual spending across the “Retail” channel of “Other” regions, followed by Frozen and Milk products.

Q 1.3 On the basis of the descriptive measure of variability, which item shows the most inconsistent behavior? Which items shows the least inconsistent behavior?

Measure of variability describes the consistence of a variable. In Descriptive statistics, we measure the ratio of standard deviation to mean as Coefficient of Variation (CV) to calculate inconsistency. The higher the value of CV, more inconsistent is the variable. Below is the plot of CV in annual spending v/s Varieties in Food Products. From the plot, it is evident that, **“Delicatessen”** is the most inconsistent and **“Fresh”** is least inconsistent. However, the annual spending of all the varieties are inconsistent, as we are able to notice that the CV value is more than 1.

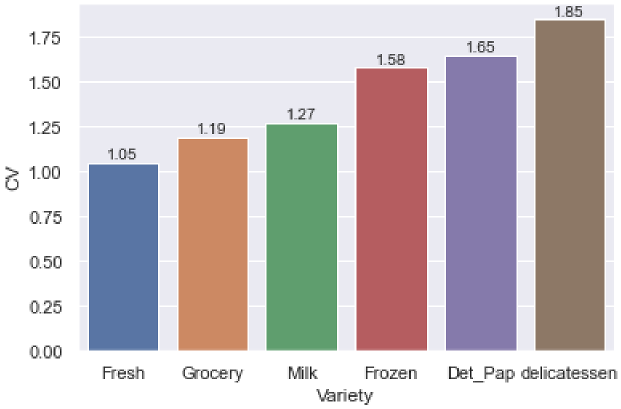


Figure 1.3 - CV in annual spending v/s Varieties in Food Products

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

Outliers are values that are abnormally away from the other values in the dataset. They tend to affect the arithmetic mean of the dataset, abnormally skewing the value to one side (upper or lower), depending on the presence of the outlier. The values below minimum **[Q1 - 1.5(Inter Quartile Range)]** are the outliers on the lower side of the dataset, while values above maximum **[Q3 + 1.5(Inter Quartile Range)]** are the outliers on the upper side of the dataset, where Q1 & Q3 are the 25th and 75th percentile respectively.

Boxplot is an excellent plot that gives us the 5 number summary (minimum, Q1, median, Q3, maximum). Q1, median & Q3 are represented by the box and the whiskers denote the values of maximum and minimum on either side of the box. The outliers are denoted by the points that fall either after the maximum whisker or below the minimum whisker.

From the below box plot, it is evident that all the varieties have outliers on the upper side.

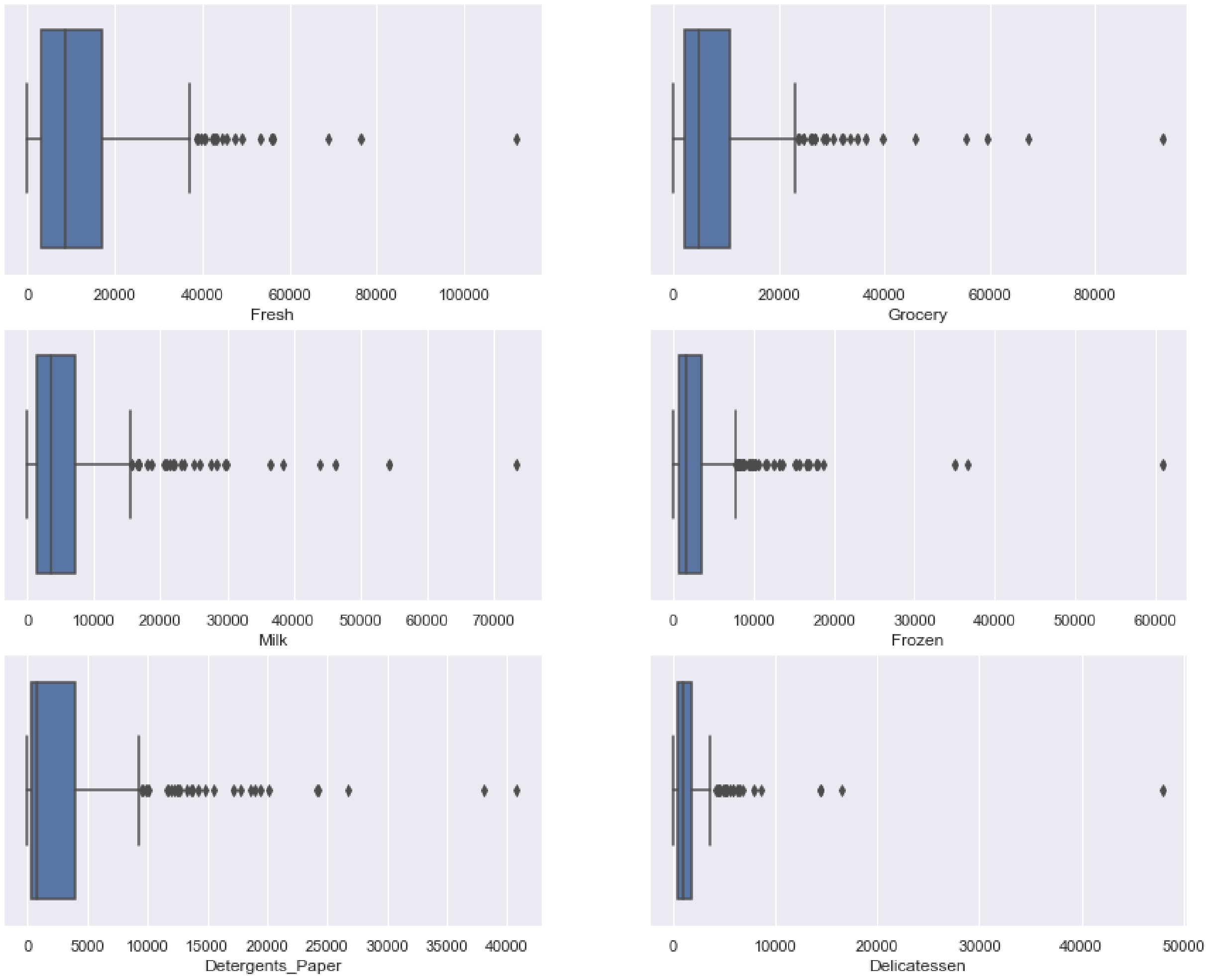


Figure 1.4 – Box plot of product varieties

Q 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

On the basis of the descriptive analysis performed, the business currently has a wide spread reach out across regions and channels. The fact that each channel across regions have the same varieties of food product that sells well. Hence, my recommendation to the distributor is to leverage on the current model, where the products of Fresh & Frozen varieties sell well in the Hotel channels across regions and the remaining varieties sell well on Retail channels across regions.

Problem 2 – Survey Data of CMSU

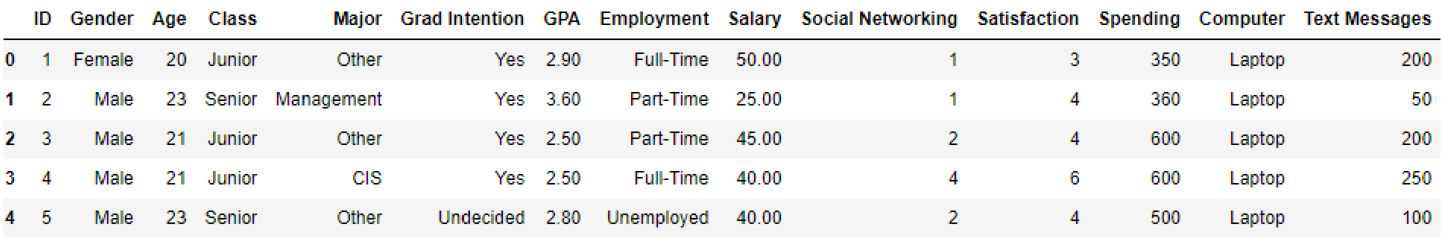
Problem Statement

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 Description

1	Gender:	Sex of the student (Male/ Female)
2	Age:	Age of the student in years
3	Class:	Class of the student (Junior/Senior/ Sophomore)
4	Major:	Field of study (Accounting/ CIS/ Economics/Finance/ International Business/ Management/ Other/Retailing/Marketing/Undecided)
5	Grad Intention:	Denotes the intention of the student in graduation (Yes/ No/ Undecided)
6	GPA:	Grade point average of the student
7	Employment:	Type of Employment (Full-time, Part-time, Unemployed)
8	Salary:	Earnings in Lacs per anuum
9	Social Networking:	Number of Social networks, the student is present
10	Satisfaction:	Satisfaction level of the student
11	Spending:	Average spending of the student
12	Computer:	The type of computer the student uses (Desktop/ Laptop/ Tablet)
13	Text Messages:	Average Text messages a student sends & receives

Sample of the Dataset



	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
0	1	Female	20	Junior	Other	Yes	2.90	Full-Time	50.00	1	3	350	Laptop	200
1	2	Male	23	Senior	Management	Yes	3.60	Part-Time	25.00	1	4	360	Laptop	50
2	3	Male	21	Junior	Other	Yes	2.50	Part-Time	45.00	2	4	600	Laptop	200
3	4	Male	21	Junior	CIS	Yes	2.50	Full-Time	40.00	4	6	600	Laptop	250
4	5	Male	23	Senior	Other	Undecided	2.80	Unemployed	40.00	2	4	500	Laptop	100

Table 2.1 – Sample of the dataset

The dataset consists of 62 student details across 13 variables.

Exploratory Data Analysis – Problem 2

Variable Type

Gender	object
Age	int64
Class	object
Major	object
Grad Intention	object
GPA	float64
Employment	object
Salary	float64
Social Networking	int64
Satisfaction	int64
Spending	int64
Computer	object
Text Messages	int64

The dataset consists of 2 float type and 6 each in integer and object data type variables.

Check for missing values in the dataset

Gender:	62 no non - null
Age:	62 no non - null
Class:	62 no non - null
Major:	62 no non - null
Grad Intention:	62 no non - null
GPA:	62 no non - null
Employment:	62 no non - null
Salary:	62 no non - null
Social Networking:	62 no non - null
Satisfaction:	62 no non - null
Spending:	62 no non - null
Computer:	62 no non - null
Text Messages:	62 no non - null

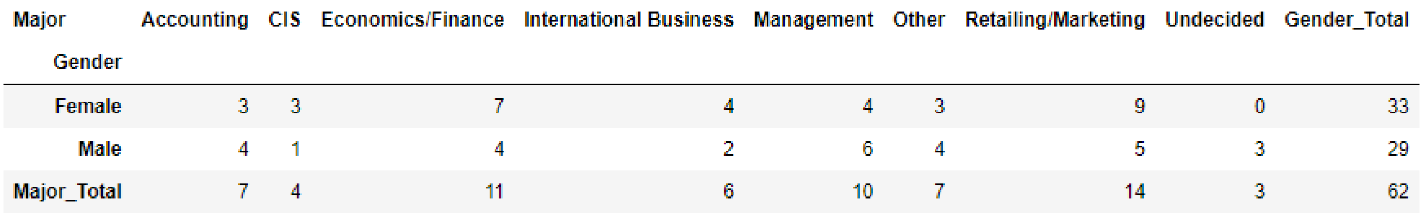
It is evident that there are no missing values in the entire dataset

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

Contingency table is a matrix in which the frequency distribution of the variable is indicated. This table plays a vital role in determining the conditional probability of events. In Python, crosstab() function is used to represent data in a Contingency table format.

Q 2.1.1. Gender and Major

The below Contingency table displays the spread of Frequency of Gender across their major field of study.

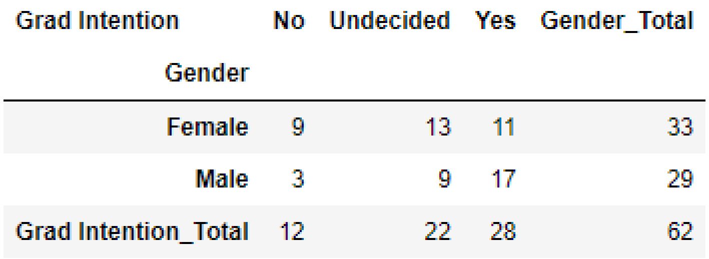


Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	Gender_Total
Gender									
Female	3	3	7	4	4	3	9	0	33
Male	4	1	4	2	6	4	5	3	29
Major_Total	7	4	11	6	10	7	14	3	62

Table 2.2 Contingency Table (Gender v/s Major)

Q 2.1.2. Gender and Grad Intention

The below Contingency table displays the spread of Frequency of Gender across their Grad intention

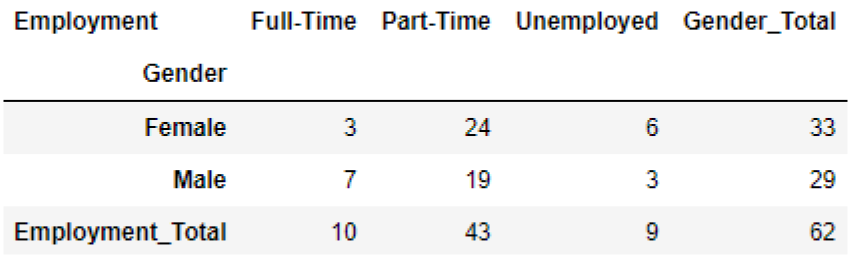


Grad Intention	No	Undecided	Yes	Gender_Total
Gender				
Female	9	13	11	33
Male	3	9	17	29
Grad Intention_Total	12	22	28	62

Table 2.3 Contingency Table (Gender v/s Grad Intention)

Q 2.1.3. Gender and Employment

The below Contingency table displays the spread of Frequency of Gender across their employment type.

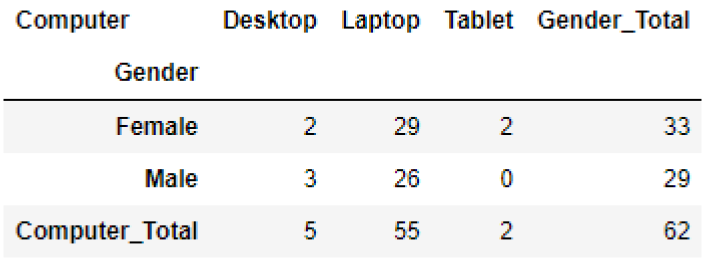


Employment	Full-Time	Part-Time	Unemployed	Gender_Total
Gender				
Female	3	24	6	33
Male	7	19	3	29
Employment_Total	10	43	9	62

Table 2.4 Contingency Table (Gender v/s Employment)

2.1.4. Gender and Computer

The below Contingency table displays the spread of Frequency of Gender across their computer type.



Computer	Desktop	Laptop	Tablet	Gender_Total
Gender				
Female	2	29	2	33
Male	3	26	0	29
Computer_Total	5	55	2	62

Table 2.5 Contingency Table (Gender v/s 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?

The probability is defined as the chance of an event to happen. Event is defined as the outcome of an experiment. Experiment is a process that is performed to understand the outcomes. Set of all outcomes of an experiment is called as the sample space.

Theoretically Probability of an event is defined as the ratio of number of ways that are likely to the occurrence of an event to the total number of outcomes of the experiment.

P(A) = m/n

Where, P(A) = Probability of an event A

m = number of ways that are likely to the occurrence of an event

n = size of the sample space

From the contingency tables we were able to infer that there are 29 males out of 62 students. Therefore, the probability that a randomly selected CMSU student will be male is 29 divided by 62, which is equal to 0.47.

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

From the contingency tables we were able to infer that there are 33 males out of 62 students. Therefore, the probability that a randomly selected CMSU student will be male is 33 divided by 62, which is equal to 0.53.

Hence, the Probability that a randomly selected CMSU student will be female is more than that of being a male.

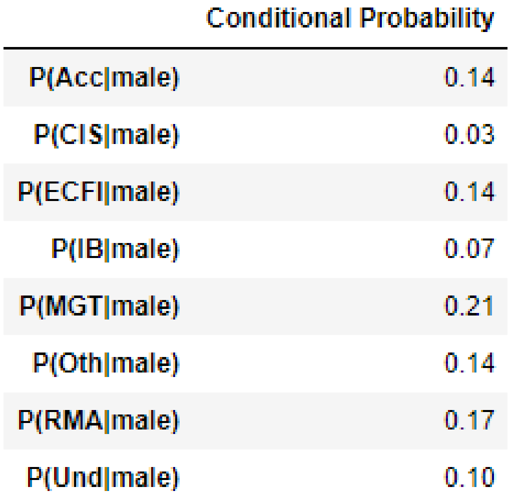
2.3. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

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

Conditional Probability is defined as the probability of an event given that another event has already occurred. The conditional probability is more accurate compared to the normal probability as the sample space size has been reduced. Conditional probability is defined theoretically as

P(A|B) = P(A intersection B)/P(B)

From the Contingency table (Table 2.2), we will be able to calculate the conditional probability of different majors among the male students. The calculated Conditional probabilities are as in the table below.



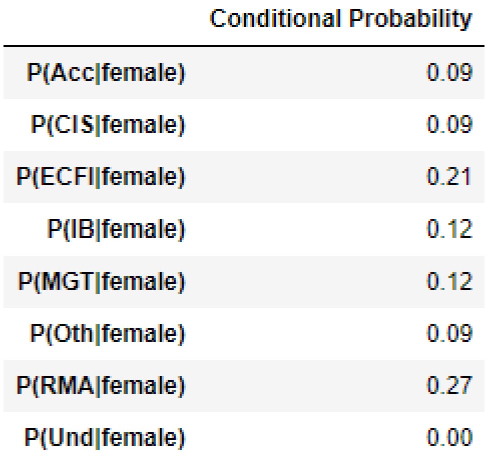
	Conditional Probability
P(Acc|male)	0.14
P(CIS|male)	0.03
P(ECFI|male)	0.14
P(IB|male)	0.07
P(MGT|male)	0.21
P(Oth|male)	0.14
P(RMA|male)	0.17
P(Und|male)	0.10

Table 2.6 – Conditional Probabilities of Major given male

From the above Table 2.6 it can be inferred that a randomly selected male student with major in Management has the highest probability.

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

From the Contingency table (Table 2.2), we will be able to calculate the conditional probability of different majors among the female students. The calculated Conditional probabilities are as in the table below



	Conditional Probability
P(Acc|female)	0.09
P(CIS|female)	0.09
P(ECFI|female)	0.21
P(IB|female)	0.12
P(MGT|female)	0.12
P(Oth|female)	0.09
P(RMA|female)	0.27
P(Und|female)	0.00

Table 2.7 – Conditional Probabilities of Major given female

From the above Table 2.7 it can be inferred that a randomly selected female student with major in Retail/ Marketing has the highest probability.

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.

From the Table 2.3, we can infer that the number of males intending to graduate is 17 out of 62 students. Hence, the probability that a randomly chosen student is a male and intends to graduate is 0.27

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

From the Table 2.5, we can infer that the number of females who doesn't have laptop is 4 out of 62 students. Hence, the probability that a randomly chosen student is female and does not have laptop is 0.06.

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?

Here, the event of choosing a male student or has a full-time employment are mutually exclusive events. Hence, as per the addition rule of mutually exclusive events

P(A or B) = P(A) + P(B) – P(A intersection B)

Assume,

A = event of choosing a male student

B = event of choosing a student with full-time employment

A intersection B is Male students who are having full-time employment.

Hence, From the Table 2.4, P(A) = 0.46, P(B) = 0.16 & P(A intersection B) = 0.11. Therefore, Probability that a randomly chosen student is a male or has full-time employment is 0.52

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

Here, We are asked to compute probability of two mutually exclusive events given a condition. i.e Give a female student is selected, we need to calculate the probability of her being a major in International Business or Management that are mutually exclusive.

Hence, from the contingency table 2.2, probability of selecting a female student is 0.53 and probability of a female majoring in International Business or Management is given as sum of the probabilities of a female majoring in International Business and Management. Hence, the value 0.12.

The required probability that given a female student is randomly chosen, she is majoring in international business or management is 0.24.

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

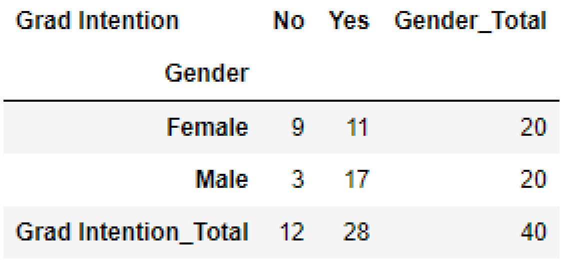
Here, We are asked to evaluate the test of independence.

Two events are said to be independent if the occurrence of one event is in no way influenced by the other event. This is given by the multiplication rule, as below

If two events A & B are independent, then

P(A intersection B) = P(A).P(B)

To create a contingency table of Gender and Intent to graduate at 2 levels, we first need to create a subset of the original data with only Intent to graduate as Yes/ No. With this subset, we shall create a contingency table using crosstab() function in Python.



Grad Intention	No	Yes	Gender_Total
Gender			
Female	9	11	20
Male	3	17	20
Grad Intention_Total	12	28	40

Table 2.8 – Contingency Table (Grad Intention – 2 levels)

Assume

A = event of choosing a female student

B = event of choosing a student intending to do graduation

From the Table 2.8, P(A) = 0.5, P(B) = 0.7, P(A intersection B) = 0.27.

RHS = P(A).P(B) = 0.35

Clearly, LHS is not equal to RHS. Hence graduate intention and being female are not independent events.

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

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?

From the dataset, we are having 17 students with GPA less than 3. Hence the probability of selecting a random student with GPA less than 3 is 0.27.

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.

Here, the condition is that, we have selected a student earning 50 or more salary

First part – selecting male given he earns 50 or more

From the given dataset we are having 14 males earning a salary of 50 or more and total students earning 50 or more is 32. Hence the conditional probability of selecting a male student given he earns 50 or more is 0.44

Second part – selecting female given he earns 50 or more

From the given dataset we are having 18 females earning a salary of 50 or more and total students earning 50 or more is 32. Hence the conditional probability of selecting a male student given he earns 50 or more is 0.56

We are able to infer that probability of a selecting a female earning 50 or more is greater than that of a male earning 50 or more.

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 a distribution to be normal, it must satisfy the below 2 points:

- The continuous distribution curve should look like a bell
- Mean, Median & Mode are all equal

Here, we shall infer the same using the distribution plot. The distribution plot of the four numerical continuous variables are as below

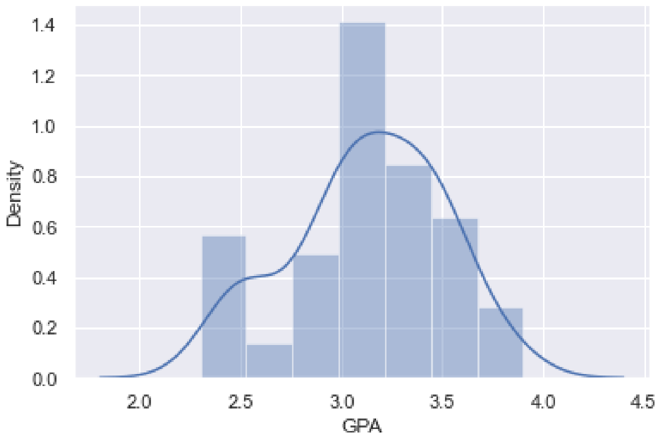


Figure 2.1 – Distribution Plot – GPA

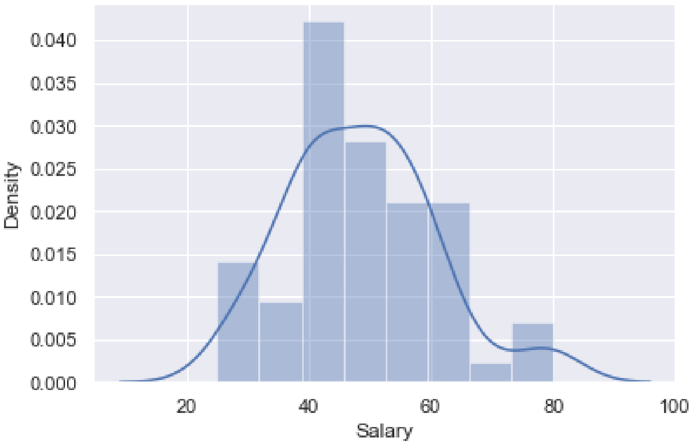


Figure 2.2 – Distribution Plot – Salary

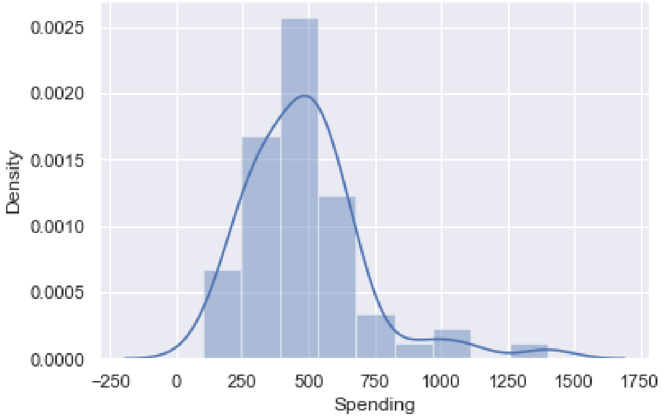


Figure 2.3 – Distribution Plot – Spending

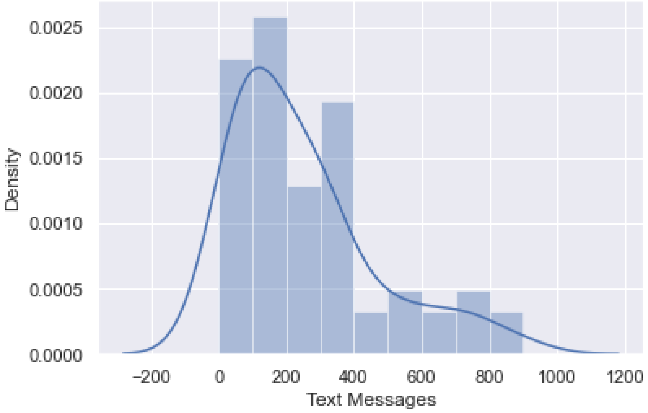


Figure 2.4 – Distribution Plot – Text Messages

None of the four variables follow a normal distribution as the distribution (density) curves(KDE) are not a bell curve. We can also conclude that the mean, median and mode would not be equal in any of these 4 variables.

Conclusion:

Considering the sample to be a representation of the population of students in CMSU, we can conclude that, there are more female graduating and placed in higher paid jobs than compared to the male students.

Problem 3 –Moisture Measurements of manufacturers of ABC Asphalt 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 coloring purposes to fall off the shingles resulting in appearance problems. To monitor the amount of moisture present, the company conducts moisture tests. A shingle is weighed and then dried. The shingle is then reweighed, and based on the amount of moisture taken out of the product, the pounds of moisture per 100 square feet are calculated. The company would like to show that the mean moisture content is less than 0.35 pounds per 100 square feet.

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

Q 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.

Hypothesis testing is done to verify if the right samples are selected and they truly represent a population. Hypothesis is a statement, an assumption and may or may not be true. A null hypothesis(H0) is a status quo and an alternate hypothesis (Ha) is the assumption usually which we try to prove.

Hypothesis testing follows a defined procedure as below:

Step 1 - Frame null and alternate hypothesis

H0 : population mean <= 0.35

Ha : population mean > 0.35

Step 2 – Decide the level of significance

Assume level of significance, alpha = 0.05

Step 3 – Decide the test type

1 Sample t test will be done as population std deviation is not known. Also, we assume that the samples are randomly selected and independent of each other.

Step 4 – calculate the test statistic and p value

Test Statistic and p value are calculated for both the types.

For the given data of A type and B type, we have the below results on using the scipy.stats.ttest_1samp in python. Here in the result, p value is divided by 2, as by default python calculates p value for 2 tailed test.

A type test Statistic value is -1.47
A type p value is 0.07
B type test Statistic value is -3.1
B type p value is 0.0021

Step 5: Statistical conclusions

Here, we can see that p value of type a is 0.07 and is greater than the level of significance (0.05).

Hence for type A, we fail to reject the null hypothesis and conclude that the mean moisture content is less than or equal to 0.35.

For Type B, the p value is 0.0021 and is less than the level of significance (0.05). Here, we reject the null hypothesis and conclude that the mean moisture content is greater than 0.35. Hence, the company needs to take immediate measures in the quality control process for Type B Asphalt Shingles.

Q 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 - Frame null and alternate hypothesis

H0 : a_mean - b_mean = 0

Ha : a_mean - b_mean != 0

Step 2 – Decide the level of significance

Assume level of significance, alpha = 0.05

Step 3 – Decide the test type

We have two independent samples with different sample sizes. The Population std deviation is not known. We need to compare means of the two samples. Hence, we will perform 2 Sample unpaired t test assuming the variances of both the samples are identical.

Step 4 – calculate the test statistic and p value

Test Statistic value is 1.29

p value is 0.2

Step 5: Statistical conclusions

Here, we can see that p value is 0.2 and is greater than the level of significance (0.05).

Hence, we fail to reject the null hypothesis and conclude that the population mean for both the types are equal.

The END