Survey Analysis

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 receive responses from 62 undergraduates (stored in the ***Survey*** data set).

Exploratory Data Analysis:

ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
1	Female	20	Junior	Other	Yes	2.9	Full-Time	50.0	1	3	350	Laptop	200
2	Male	23	Senior	Management	Yes	3.6	Part-Time	25.0	1	4	360	Laptop	50
3	Male	21	Junior	Other	Yes	2.5	Part-Time	45.0	2	4	600	Laptop	200
4	Male	21	Junior	CIS	Yes	2.5	Full-Time	40.0	4	6	600	Laptop	250
5	Male	23	Senior	Other	Undecided	2.8	Unemployed	40.0	2	4	500	Laptop	100

Dataset has 14 variables, which has the different values for the particular response. ID is the variable which has the unique row number for each response.

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

```
ID                       int64
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
dtype: object
```

Check for missing values in the dataset:

```
RangeIndex: 62 entries, 0 to 61
Data columns (total 14 columns):
ID                     62 non-null int64
Gender                 62 non-null object
Age                    62 non-null int64
Class                  62 non-null object
Major                  62 non-null object
Grad Intention         62 non-null object
GPA                    62 non-null float64
Employment             62 non-null object
Salary                 62 non-null float64
Social Networking      62 non-null int64
Satisfaction           62 non-null int64
Spending               62 non-null int64
Computer               62 non-null object
Text Messages          62 non-null int64
dtypes: float64(2), int64(6), object(6)
memory usage: 6.9+ KB
```

From the above description we see that there is no missing value present in the dataset.

Part I

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

Contingency Table: A cross-classification table showing the distribution of one row variable and a column variable. Contingency tables are useful to understand bivariate relationship between the constituent variables. Contingency tables may be constructed with more than 2 categorical variables.

a. Gender and Major

Major	**Accounting**	**CIS**	**Economics/Finance**	**International Business**	**Management**	**Other**	**Retailing/ Marketing**	**Undecided**
Gender								
Female	3	3	7	4	4	3	9	0
Male	4	1	4	2	6	4	5	3

b. Gender and Grad Intention

Grad Intention	**No**	**Undecided**	**Yes**
Gender			
Female	9	13	11
Male	3	9	17

c. Gender and Employment

Employment	Full-Time	Part-Time	Unemployed
Gender			
Female	3	24	6
Male	7	19	3

d. Gender and Computer

Computer	Desktop	Laptop	Tablet
Gender			
Female	2	29	2
Male	3	26	0

2) Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following questions:

a) What is the probability that a randomly selected CMSU student will be male?

Prob (Male)= (Total number of male students)/ (Total number of students at the university).
Prob (Male)= 29/62 = 0.468

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

Prob (Female)= (Total number of female students)/ (Total number of students at the university)
Prob (Female)= 33/62 = 0.532 = 1 – Prob (Male)

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

Count of Males = 29

P (Accounting| Male) = count of males selecting account/male count = 4/29 = 0.138
P (CIS| Male) = count of males selecting CIS /male count = 1/29 = 0.034
P (Economics| Male) = count of males selecting Economics /male count = 4/29 = 0.138
P (International| Male) = count of males selecting International /male count = 2/29 = 0.069
P (Mgmt.| Male) = count of males selecting Mgmt. /male count = 6/29 = 0.20
P (Other| Male) = count of males selecting other /male count = 4/29 = 0.138
P (Retail| Male) = count of males selecting Retail /male count = 5/29 = 0.172
P (Undecided| Male) = count of males are Undecided /male count = 3/29 = 0.103

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

Note that sum of the above conditional probabilities is 1

Count of Female = 33

P (Accounting| Female) = count of Female selecting account/ Female count = 3/33 = 0.091

P (CIS| Female) = count of Female selecting CIS / Female count = 3/33 = 0.091
P (Economics| Female) = count of Female selecting Economics / Female count = 7/33 = 0.21
P (International| Female) = count of Female selecting Intl / Female count = 4/33 = 0.12
P (Mgmt.| Female) = count of Female selecting Mgmt. / Female count = 4/33 = 0.121
P (Other| Female) = count of Female selecting other / Female count = 3/33 = 0.091
P (Retail| Female) = count of Female selecting Retail / Female count = 9/33 = 0.28
P (Undecided| Female) = count of Female are Undecided / Female count = 0/33 = 0

Note that sum of the above conditional probabilities is 1

c) Let the event that a randomly chosen students is Male be denoted by M
The event that a randomly chosen student Intends to graduate be denoted by G
Prob (Male AND Intends to graduate) = P(M ∩ G)

From the contingency table Gender and Grad Intention, there are 17 male students who intend to graduate

Hence
P(M ∩ G) = 17 / 62 = 0.274

d) Let the event that a randomly chosen students is Female be denoted by F
The event that a randomly chosen student has a laptop be denoted by L
Hence the event that a randomly chosen student does not have a laptop be denoted by L^c

Prob(Female AND Does not have a laptop) = P(F ∩ L^c)

From the contingency table gender and computer the number of female students not having a laptop is 2 + 2 = 4. (having desktops and tablets)

Hence
P(F ∩ L^c)= 4 / 62 = 0.06

2.2.4

a) Let the event that a randomly chosen students is Male be denoted by M
Let the event that a randomly chosen students has full-time employment be denoted by E
Prob(Male OR full-time employment) = P(M U E) = P(M) + P(E) - P(M ∩ E)
Where (M ∩ E) denotes the event that a randomly chosen student is a male AND has full-time employment.

P(M) = 29/62 = 0.468
P(E) = 10/62 = 0.16
P(M ∩ E) = 7/62 = 0.11

Hence P(M U E) = P(M) + P(E) - P(M ∩ E) = 0.468 + 0.16 – 0.11 = 0.518

b) **When dealing with conditional probability that the students chosen is a female, only the row where gender = Female in the table Gender and Major is of concern.**
Prob(International Business OR Management) = $(4 + 4) / 33 = 0.242$

	No	Yes	Total
Female	9	11	20
Male	3	17	20
Grand Total	12	28	40

2.2.5

Refer to the table above.

$P(F) = 20/40 = 0.5$
$P(Yes) = 28/40 = 0.7$

If being female and graduate intention are independent, the $P(F \cap Yes) = P(F)P(Yes)$

$P(F \cap Yes) = 11 / 40 = 0.275$
$P(F)P(Yes) = 0.5(0.7) = 0.35 \neq P(F \cap Yes)$

The two events are not independent

2.4

Prob($GPA < 3$) = $17 / 62 = 0.274$

Prob($Salary \geq 50$ | Male) = $14/29 = 0.48$
Prob($Salary \geq 50$ | Female) = $18/33 = 0.545$

PART 2: 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 we will test empirical rule: The empirical rule states that for a normal distribution, nearly all of the data will fall within three standard deviations of the mean. The empirical rule can be broken down into three parts:

- 68% of data falls within the first standard deviation from the mean.
- 95% fall within two standard deviations from the mean
- 99.7% fall within three standard deviations from the mean

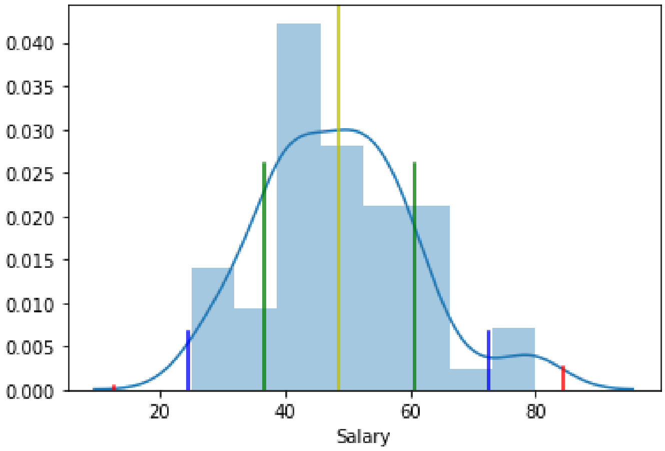
The rule is also called the 68-95-99.7 Rule or the Three Sigma Rule.

First we will calculate the mean and median and standard deviation for the variables.

Salary Variable:

Salary Mean: 48.55
Salary Median: 50.0
Salary Standard Deviation: 12.08

Histogram for Salary variable:

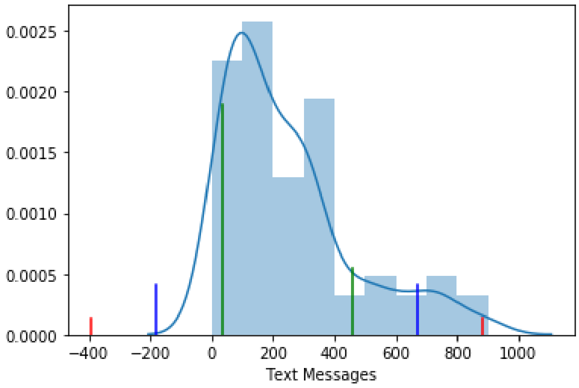


Text Messages Variable:

Text Messages Mean: 246.21
Text Messages Median: 200.0
Text Messages Standard Deviation: 214.47

Since mean and median of the Text Messages column has huge difference. It results that that data is highly skewed.

Histogram of Text messages

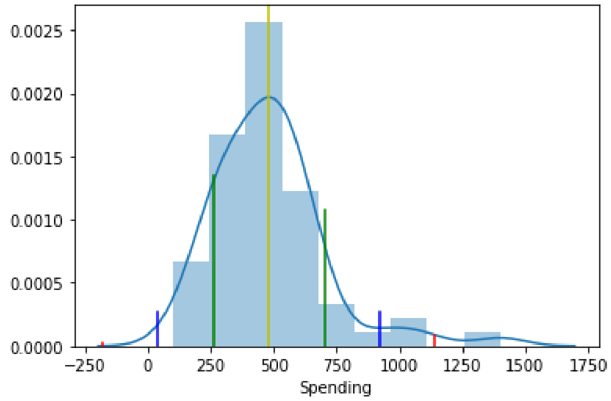


Spending Variable:

Spending Mean: 482.02
Spending Median: 500.0

Spending Standard Deviation: 221.95

Histogram of Spending



From the above analysis, we came to the result that variable (Salary, Text messages and Spending) are not normally distributed. Since the data is skewed (mean =! median) and the empirical rule also failed to propose that the data is normally distributed.