Problem 2 – 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).

Solution:

Exploratory Data Analysis:

	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
0	1	Female	20	Junior	Other	Yes	2.9	Full-Time	50.0	1	3	350	Laptop	200
1	2	Male	23	Senior	Management	Yes	3.6	Part-Time	25.0	1	4	360	Laptop	50
2	3	Male	21	Junior	Other	Yes	2.5	Part-Time	45.0	2	4	600	Laptop	200
3	4	Male	21	Junior	CIS	Yes	2.5	Full-Time	40.0	4	6	600	Laptop	250
4	5	Male	23	Senior	Other	Undecided	2.8	Unemployed	40.0	2	4	500	Laptop	100

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

For the following results we used the crosstab function in python:

2.1.1: Gender and Major:

Major	Accounting	CIS	Economics/Finance	International Business	\
Gender					
Female	3	3	7	4	
Male	4	1	4	2	

Major	Management	Other	Retailing/Marketing	Undecided
Gender				
Female	4	3	9	0
Male	6	4	5	3

2.1.2: Gender and Grad Intention:

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

2.1.3: Gender and Employment:

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

2.1.4: Gender and Computer:

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

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

First, we create a crosstab of Gender, Major and Employment:

Employment		Full-Time	Part-Time	Unemployed
Gender	Major			
Female	Accounting	0	3	0
	CIS	0	3	0
	Economics/Finance	1	5	1
	International Business	0	4	0
	Management	0	1	3
	Other	2	1	0
	Retailing/Marketing	0	7	2
Male	Accounting	1	2	1
	CIS	1	0	0
	Economics/Finance	1	3	0
	International Business	0	2	0
	Management	1	5	0
	Other	0	3	1
	Retailing/Marketing	1	3	1
	Undecided	2	1	0

Then we get the count of the total number of Male and Female:

```
Female    33
Male      29
Name: Gender, dtype: int64
```

Then we get the total number of people in the different employment categories:

```
Employment
Full-Time     10
Part-Time     43
Unemployed     9
dtype: int64
```

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

Probability that a randomly selected CMSU student is a male is given by the formula,

Probability of a randomly selected CMSU student is a male = male students / Total number of students

Substituting the values, we get the answer as 0.46 or 46.77%

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

Probability that a randomly selected CMSU student is a female is given by the formula,

prob_female = (female_students/Total_number_of_students)*100

Substituting the values, we get the answer as 0.53 or 53.22%

Alternatively, we can also compute the same using the formula, 1 – (Probability of a randomly selected CMSU student being a male)

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

First, we place a sum column and a sum row in the crosstab of the different employment categories of students with different majors:

Employment	Full-Time	Part-Time	Unemployed	Total_Sum
(Female, Accounting)	0	3	0	3
(Female, CIS)	0	3	0	3
(Female, Economics/Finance)	1	5	1	7
(Female, International Business)	0	4	0	4
(Female, Management)	0	1	3	4
(Female, Other)	2	1	0	3
(Female, Retailing/Marketing)	0	7	2	9
(Male, Accounting)	1	2	1	4
(Male, CIS)	1	0	0	1
(Male, Economics/Finance)	1	3	0	4
(Male, International Business)	0	2	0	2
(Male, Management)	1	5	0	6
(Male, Other)	0	3	1	4
(Male, Retailing/Marketing)	1	3	1	5
(Male, Undecided)	2	1	0	3
Total	10	43	9	62

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

The conditional probability of the major: Accounting among the male students is given by the formula, P(Accounting|Male) = Number of males in accounting / Total Number of males

Substituting the values, we get the conditional probability of Accounting among the male students as 13.79%

The conditional probability of the major: CIS among the male students is given by the formula,

cond_prob_cis_male = (cis_male/male_students)*100

Substituting the values, we get the conditional probability of CIS among the male students as 3.44%

The conditional probability of the major: Economics Finance is given by the formula,

cond_prob_economicsfinance_male = (economics_finance_male/male_students)*100

Substituting the values, we get the conditional probability of Economics/Finance among the male students as 13.79%

The conditional probability of the major: International Business among the male students is given by the formula, cond_prob_internationalbusiness_male = (international_business_male/male_students)*100

Substituting the values, we get the conditional probability of International Business among the male students is 6.89%

The conditional probability of the major: Management among the male students is given by the formula, cond_prob_management_male = (management_male/male_students)*100

Substituting the values, we get the conditional probability of Management among the male students is 20.68%

The conditional probability of the major: Other among the male students is given by the formula,

cond_prob_other_male = (other_male/male_students)*100

Substituting the values, we get the conditional probability of Other among male students is 13.79%

The conditional probability of the major: Retailing Marketing among the male students is given by the formula, cond_prob_retailing_marketing_male = (retailing_marketing_male/male_students)*100

Substituting the values, we get the conditional probability of Retailing/Marketing among male students is 17.24%

The conditional probability of the major: Undecided among the male students is given by the formula,

cond_prob_undecided_male = (undecided_male/male_students)*100

Substituting the values, we get the conditional probability of undecided among male students is 10.34%

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

The conditional probability of the major: Accounting among the female students is given by the formula,

cond_prob_accounting_female = (accounting_female/female_students)*100

Substituting the values, we get the conditional probability of Accounting among the female students is 9.09%

The conditional probability of the major: CIS among the female students is given by the formula,

cond_prob_cis_female = (cis_female/female_students)*100

Substituting the values, we get the conditional probability of CIS among the female students is 3.03%

The conditional probability of the major: Economics Finance among the female students is given by the formula, cond_prob_economicsfinance_female = (economics_finance_female/female_students)*100

Substituting the values, we get the conditional probability of Economics/Finance among the female students is 12.12%

The conditional probability of the major: International Business among the female students is given by the formula, cond_prob_internationalbusiness_female = (international_business_female/female_students)*100

Substituting the values, we get the conditional probability of International Business among the female students is 6.06%

The conditional probability of the major: Management among the female students is given by the formula,

cond_prob_management_female = (management_female/female_students)*100

Substituting the values, we get the conditional probability of Management among the female students is 18.18%

The conditional probability of the major: Other among the female students is given by the formula,

cond_prob_other_female = (other_female/female_students)*100

Substituting the values, we get the conditional probability of Other among female students is 12.12%

The conditional probability of the major: Retailing Marketing among the female students is given by he formula, cond_prob_retailing_marketing_female = (retailing_marketing_female/female_students)*100

Substituting the values, we get the conditional probability of Retailing/Marketing among female students is 15.15%

The conditional probability of the major: Undecided among the female students is given by the formula, cond_prob_undecided_female = (undecided_female/female_students)*100

Substituting the values, we get the conditional probability of undecided among female students is 9.09%

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.

First, we create a crosstab of Gender and Grad Intention as follows:

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

Probability of all the males P(A) = 29/62 = 0.467

Probability of a student who intends to graduate P(B) = 28/62 = 0.4516

Probability of a male student given that he intends to graduate = P(A|B) = 17/28 = 0.607

Now as per the formula, P(A and B) = P(B) * P(A|B)

P(A and B) = 0.4516 * 0.607 = 0.27 or 27% probability

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.

We first create a crosstab of Gender and Computer as follows:

We have also included the Total column and row representing the sum of the row values and the sum of the column values respectively.

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

Probability of a female student P(A) = 33/62 = 0.53

Probability of having a laptop = 55/62 = 0.88

Probability of not having a laptop P(B) = (1 - 0.88) = 0.11

Now as per the formula, P(A and B) = P(B) * P(A|B)

P(A|B) = 0.57

P(A and B) = 0.57 * 0.11 = 0.06

Probability that a randomly selected student is a female and does not have a laptop is 0.06

2.5. Assume that the sample is a 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 either a male or has a full-time employment

Let, P(A) denote the probability of being a male and let P(B) denote the probability of having a full time employment.

Here, P(A or B) is asked where the events are non-mutually exclusive, so P(A or B) = P(A) + P(B) - P(A and B)

From df_crosstab5, we see that P(A) = 29/62 = 0.467

P(B) = 10/62 = 0.161

P(A and B) = 7/62 = 0.112

P(A or B) = 0.467 + 0.161 - 0.112 = 0.516

Hence the probability that a randomly selected student is either a male or has a full time employment is 0.51

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

We first create a new dataframe with the name extract1 with the Major column containing only International Business and Management categories using the isin function in python.

	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
1	2	Male	23	Senior	Management	Yes	3.6	Part-Time	25.0	1	4	360	Laptop	50
8	9	Female	20	Junior	Management	Yes	3.6	Unemployed	30.0	0	4	500	Laptop	400
12	13	Male	22	Senior	International Business	Undecided	3.4	Part-Time	40.0	2	3	400	Desktop	45
13	14	Male	22	Senior	International Business	Undecided	3.1	Part-Time	40.0	1	3	400	Laptop	150
14	15	Male	21	Senior	Management	Yes	3.2	Part-Time	54.0	3	4	600	Laptop	400
15	16	Male	24	Senior	Management	Undecided	3.4	Part-Time	45.0	4	4	500	Laptop	175
19	20	Female	20	Junior	Management	Undecided	3.2	Unemployed	60.0	2	6	300	Laptop	350
25	26	Male	24	Senior	Management	Yes	3.3	Full-Time	60.0	0	1	300	Laptop	40
27	28	Female	20	Junior	International Business	Yes	2.9	Part-Time	50.0	3	1	900	Laptop	100
36	37	Male	21	Senior	Management	Yes	3.1	Part-Time	40.0	1	4	500	Laptop	100
44	45	Female	21	Senior	International Business	No	3.0	Part-Time	30.0	2	5	650	Desktop	500
45	46	Female	21	Senior	Management	Undecided	3.8	Part-Time	60.0	1	4	650	Laptop	150
50	51	Female	21	Junior	Management	No	3.5	Unemployed	35.0	2	4	600	Tablet	100
51	52	Male	21	Senior	Management	No	3.0	Part-Time	50.0	1	4	500	Laptop	200
56	57	Female	21	Senior	International Business	Yes	3.4	Part-Time	42.0	1	1	200	Laptop	100
57	58	Female	21	Senior	International Business	No	2.4	Part-Time	40.0	1	3	1000	Laptop	10

We then create a crosstab using the Gender and Major columns from the above dataframe.

We also include the Total column and row to dispay the sum of the row and the columns respectively.

Major Gender	International Business	Management	Total
Female	4	4	8
Male	2	6	8
Total	6	10	16

Probability that a randomly selected female will be majoring in International Business OR Management:

P(International business or Management | Female) = 8 / 33 = 0.24

Probability that a randomly selected female will be 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 graduate intention and being female are independent events?

We first created a dataframe with the name: extract2 containing column Grad Intention and the subsequent Yes and No values.

Here is an extract of the actual dataframe:

	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
0	1	Female	20	Junior	Other	Yes	2.9	Full-Time	50.0	1	3	350	Laptop	200
1	2	Male	23	Senior	Management	Yes	3.6	Part-Time	25.0	1	4	360	Laptop	50
2	3	Male	21	Junior	Other	Yes	2.5	Part-Time	45.0	2	4	600	Laptop	200
3	4	Male	21	Junior	CIS	Yes	2.5	Full-Time	40.0	4	6	600	Laptop	250
8	9	Female	20	Junior	Management	Yes	3.6	Unemployed	30.0	0	4	500	Laptop	400
10	11	Female	23	Senior	Economics/Finance	Yes	2.8	Full-Time	50.0	2	5	400	Laptop	200
11	12	Male	21	Senior	Undecided	No	3.5	Full-Time	37.0	2	3	500	Laptop	100
14	15	Male	21	Senior	Management	Yes	3.2	Part-Time	54.0	3	4	600	Laptop	400
18	19	Male	19	Junior	Economics/Finance	Yes	3.5	Part-Time	52.0	2	5	500	Laptop	300
23	24	Male	22	Senior	Undecided	Yes	2.6	Full-Time	45.0	1	5	400	Laptop	600
24	25	Female	20	Junior	Economics/Finance	Yes	3.0	Part-Time	55.0	1	3	600	Laptop	300
25	26	Male	24	Senior	Management	Yes	3.3	Full-Time	60.0	0	1	300	Laptop	40
26	27	Male	20	Junior	Economics/Finance	Yes	3.1	Full-Time	65.0	1	5	375	Laptop	300

Then from the above dataframe we created a crosstab of Gender and Grad Intention.

Please note that the resulting crosstab will only count the Grad Intention sub categories: Yes and No as only these are mentioned in the dataframe above.

We also included the Total variables for the sum of column and row values.

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

Given that P(A) = Being a female = 20/40 = 0.5

Given that P(B) = Grad intention Yes = 28/40 = 0.7

If two events are independent, P(Female and Yes) = P(A) * P(B) = 0.50 * 0.7 = 0.37

To check if two events are independent if P(A|B) = P(A) and P(B|A) = P(B):

P(A|B) = 11/28 = 0.39

P(A) = 0.5 is not equal to P(A|B), i.e: not equal to 0.39

P(B|A) = 11/20 = 0.55

P(B) = 0.7

So, P(B|A) = 0.55 which is not equal to P(B) = 0.7

So the two events are not independent. So for dependent events, P(A and B) = P(A) * P(B|A)

P(B|A) = 11/20 = 0.55

P(A and B) = 0.55* 0.5 = 0.275

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?

We first created a dataframe with the name: df2 and included the columns of Gender, GPA and Salary.

	Gender	GPA	Salary
0	Female	2.9	50.0
1	Male	3.6	25.0
2	Male	2.5	45.0
3	Male	2.5	40.0
4	Male	2.8	40.0
...	...	...	...
57	Female	2.4	40.0
58	Female	2.9	40.0
59	Female	2.5	55.0
60	Female	3.5	30.0
61	Female	3.2	70.0

62 rows × 3 columns

We then created another dataframe from the above dataframe with the application of the filter condition: GPA < 3. The name of the resulting dataframe is df2_filtered

We also included the Total column and row to compute the sum of the row values and column values respectively.

		Gender	GPA	Salary	Total
0		Female	2.9	50.0	52.9
2		Male	2.5	45.0	47.5
3		Male	2.5	40.0	42.5
4		Male	2.8	40.0	42.8
5		Female	2.3	78.0	80.3
10		Female	2.8	50.0	52.8
23		Male	2.6	45.0	47.6
27		Female	2.9	50.0	52.9
31		Male	2.9	47.0	49.9
33		Male	2.6	40.0	42.6
37		Female	2.5	60.0	62.5
38		Male	2.8	50.0	52.8
39		Male	2.5	50.0	52.5
47		Male	2.5	80.0	82.5
57		Female	2.4	40.0	42.4
58		Female	2.9	40.0	42.9
59		Female	2.5	55.0	57.5
Total	FemaleMaleMaleMaleFemaleFemaleMaleFemaleMaleMa...		44.9	860.0	904.9

Then we used the count function to find the count of the same.

```
Gender    18
GPA       18
Salary    18
Total     18
dtype: int64
```

Probability of a student who is randomly chosen, that his GPA is less than 3 = Number of students with GPA less than 3/ Total number of students

Probability of a student who is randomly chosen, that his GPA is less than 3 = 18/62 = 0.2903

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.

We first created a new dataframe with the name df3 from the previous dataframe df2 with the application of the filter as: Salary >= 50

We also included the Total column and row to compute the sum of the row values and column values respectively.

An extract of the resulting dataframe df3 is shown as:

	Gender	GPA	Salary	Total
0	Female	2.9	50.0	52.9
5	Female	2.3	78.0	80.3
6	Female	3.0	50.0	53.0
7	Female	3.1	80.0	83.1
10	Female	2.8	50.0	52.8
14	Male	3.2	54.0	57.2
16	Female	3.7	55.0	58.7
17	Male	3.1	55.0	58.1
18	Male	3.5	52.0	55.5
19	Female	3.2	60.0	63.2
20	Female	3.2	55.0	58.2
21	Male	3.0	60.0	63.0
22	Female	3.0	55.0	58.0

2.7.2.a:

Probability that a randomly chosen male earns 50 or more = Number of male earning 50 or more/ Total number of male.

We then created a new dataframe called male50 from above dataframe df3 with the condition as Gender = Male. Then we used the count function to get a count of the values.

```
Gender    14
GPA       14
Salary    14
Total     14
dtype: int64
```

Number of Male earning 50 or more = 14

Total number of Male = 29

Probability that a randomly chosen male earns 50 or more = 14/29 = 0.48

2.7.2.b:

Probability that a randomly chosen female earns 50 or more = Number of female earning 50 or more/ Total number of female

We then created a new dataframe called female50 from above dataframe df3 with the condition as Gender = Female. Then we used the count function to get a count of the values.

```
Gender     18
GPA        18
Salary     18
Total      18
dtype: int64
```

Number of Female earning 50 or more = 18

Total number of Female = 33

Probability that a randomly chosen female earns 50 or more = 18/33 = 0.54

2.8.1 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.

We first created a new dataframe as df4 with the columns: GPA, Salary, Spending, Text Messages

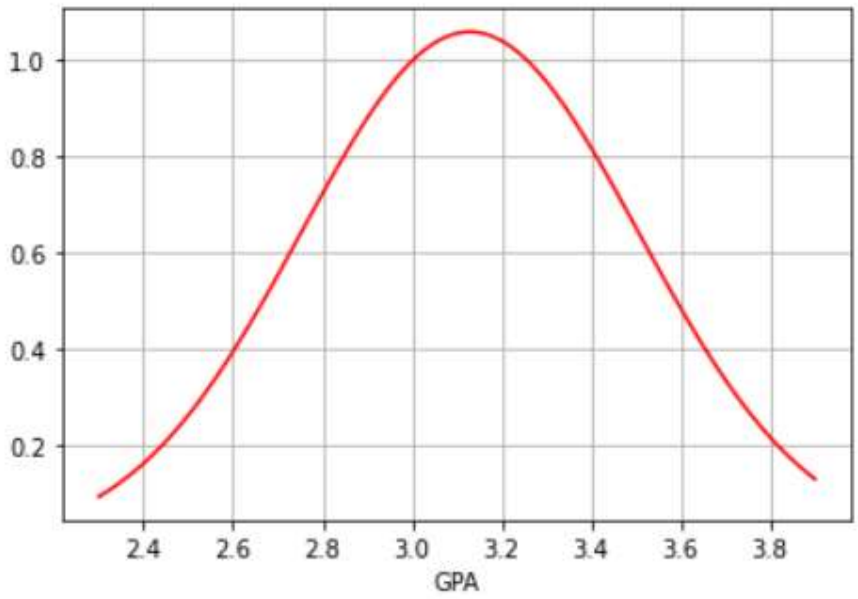
	GPA	Salary	Spending	Text Messages
0	2.9	50.0	350	200
1	3.6	25.0	360	50
2	2.5	45.0	600	200
3	2.5	40.0	600	250
4	2.8	40.0	500	100
...	...	...	...	...
57	2.4	40.0	1000	10
58	2.9	40.0	350	250
59	2.5	55.0	500	500
60	3.5	30.0	490	50
61	3.2	70.0	250	0

62 rows × 4 columns

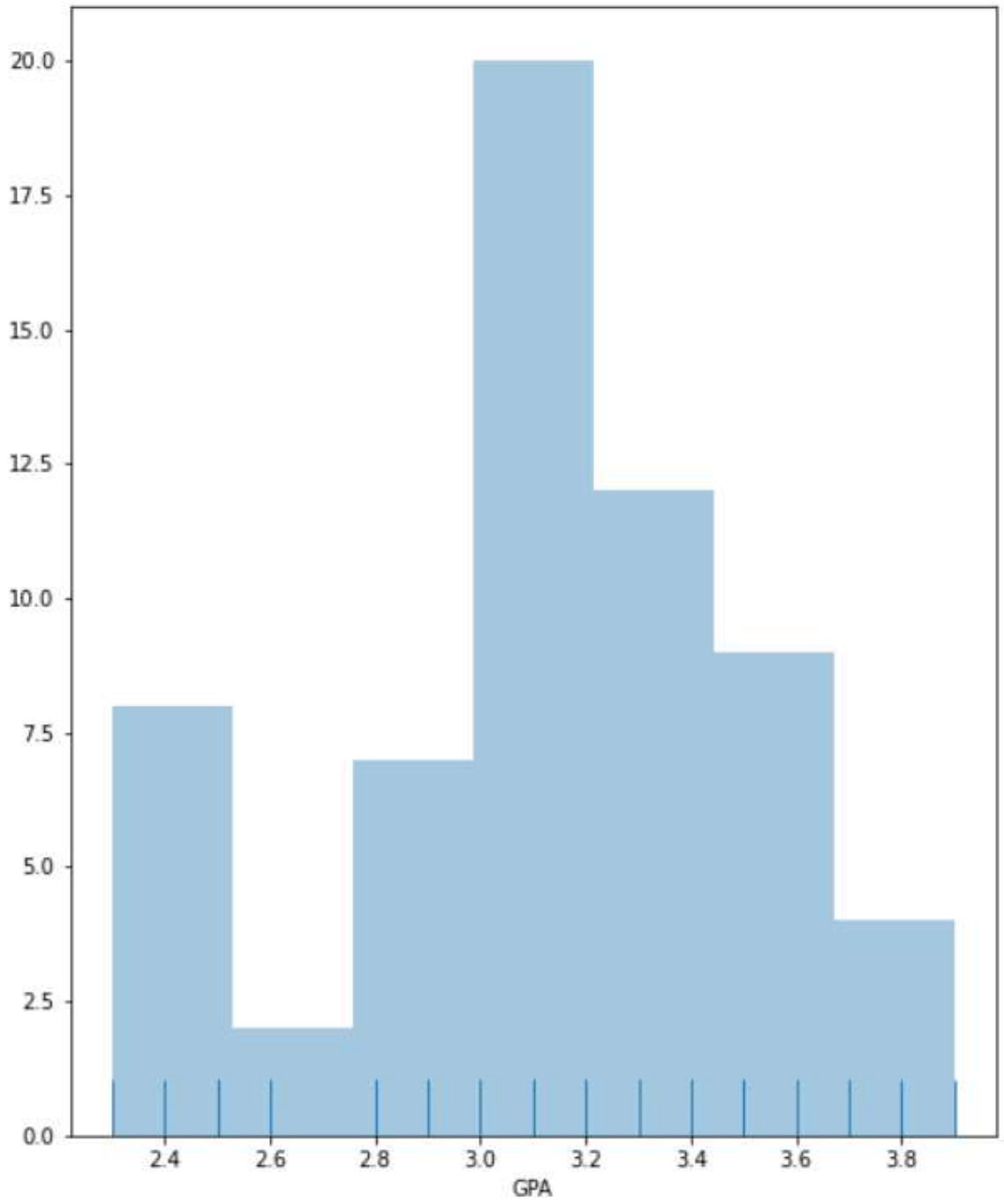
We then used the describe function on the above dataframe:

	GPA	Salary	Spending	Text Messages
count	62.000000	62.000000	62.000000	62.000000
mean	3.129032	48.548387	482.016129	246.209677
std	0.377388	12.080912	221.953805	214.465950
min	2.300000	25.000000	100.000000	0.000000
25%	2.900000	40.000000	312.500000	100.000000
50%	3.150000	50.000000	500.000000	200.000000
75%	3.400000	55.000000	600.000000	300.000000
max	3.900000	80.000000	1400.000000	900.000000

GPA Variable: To show that the GPA follows a Normal distribution, we plotted the probability distribution curve using matplotlib and histogram as follows:



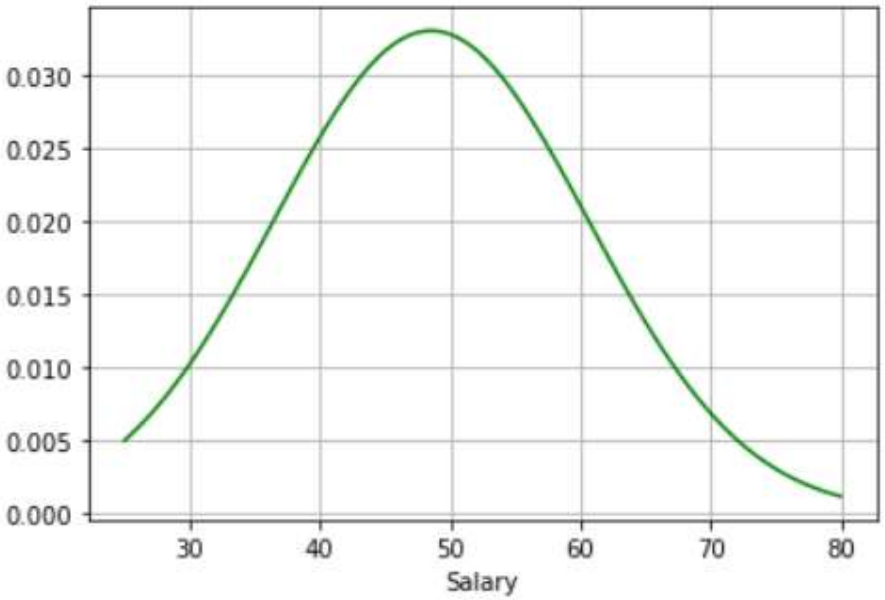
We also plot a Histogram to show the same.

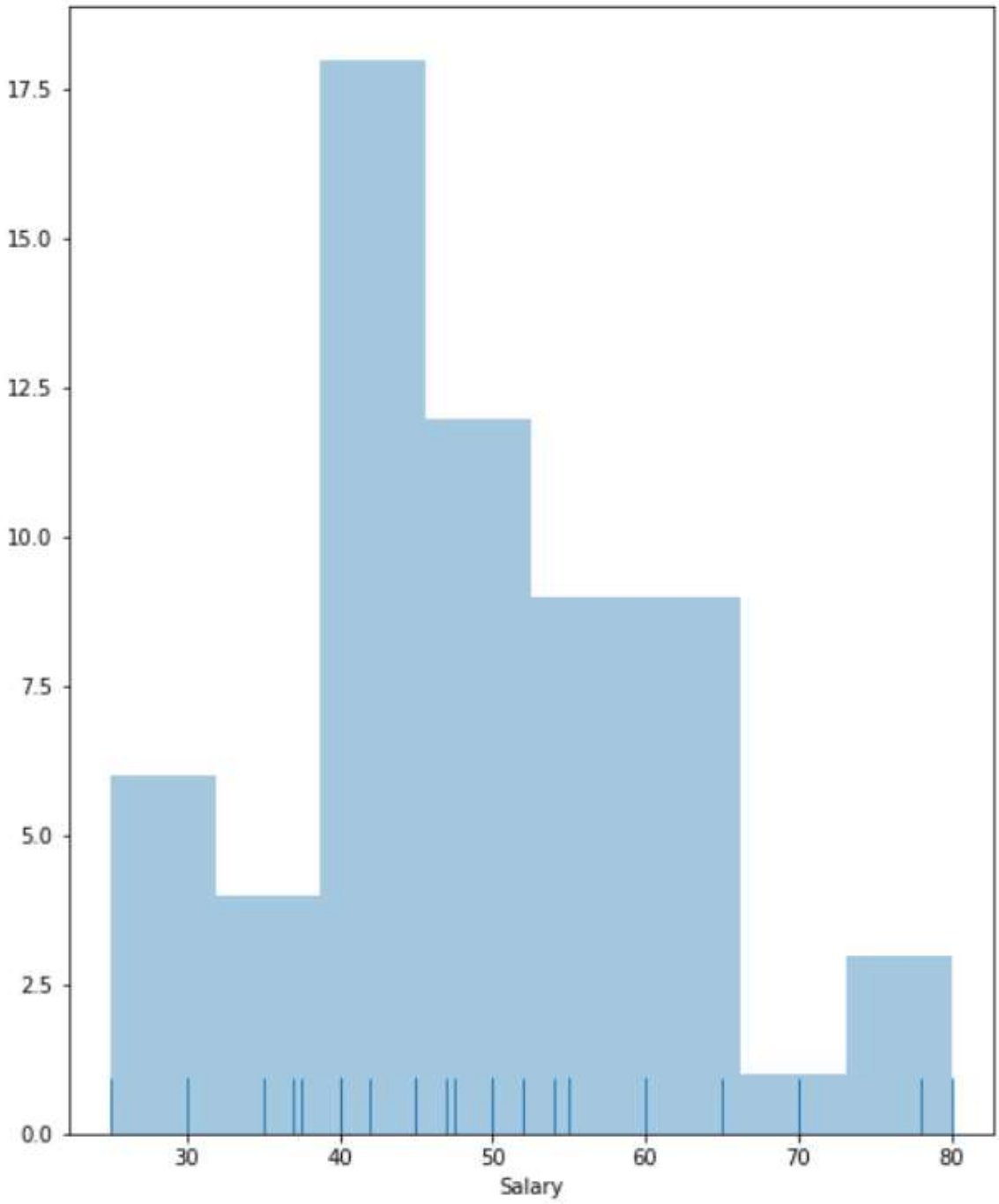


This shows that 68% of data fall between +/- 0.37 from the mean, 95% of data fall between +/- 0.74 from the mean and 99.7% of all the data fall between +/- 1.11 from the mean.

The data is perfectly symmetrical distribution for the GPA variable.

Salary variable: To show that the Salary follows a Normal distribution, we plotted the probability distribution curve using matplotlib and histogram as follows:


0.030
0.025
0.020
0.015
0.010
0.005
0.000
30
40
50
60
70
80
Salary

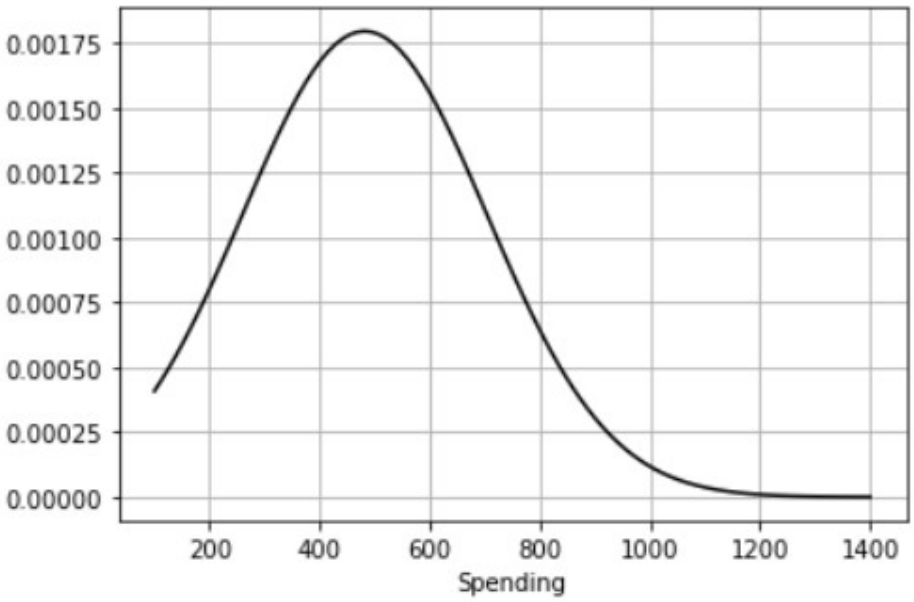


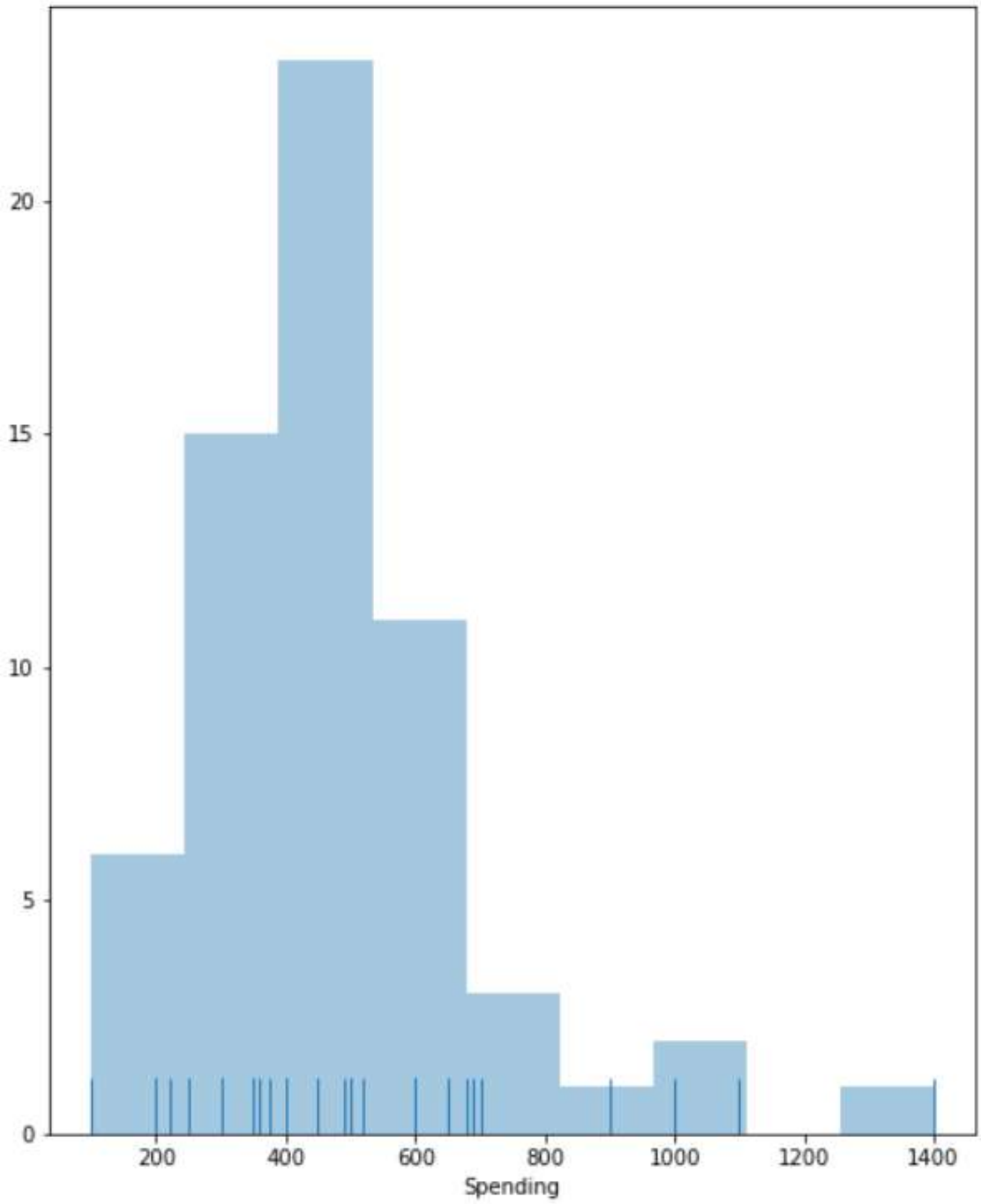
As we can see, the Salary is approximately showing a Normal distribution

This shows that 68% of data fall between +/- 12.08 from the mean, 95% of data fall between +/- 24.16 from the mean and 99.7% of all the data fall between +/- 36.24 from the mean.

The distribution for the Salary is a slightly right skewed or positive skewed distribution.

Spending variable: To show that the Spending follows a Normal distribution, we plotted the probability distribution curve using matplotlib and histogram as follows:


0.00175
0.00150
0.00125
0.00100
0.00075
0.00050
0.00025
0.00000
200
400
600
800
1000
1200
1400
Spending

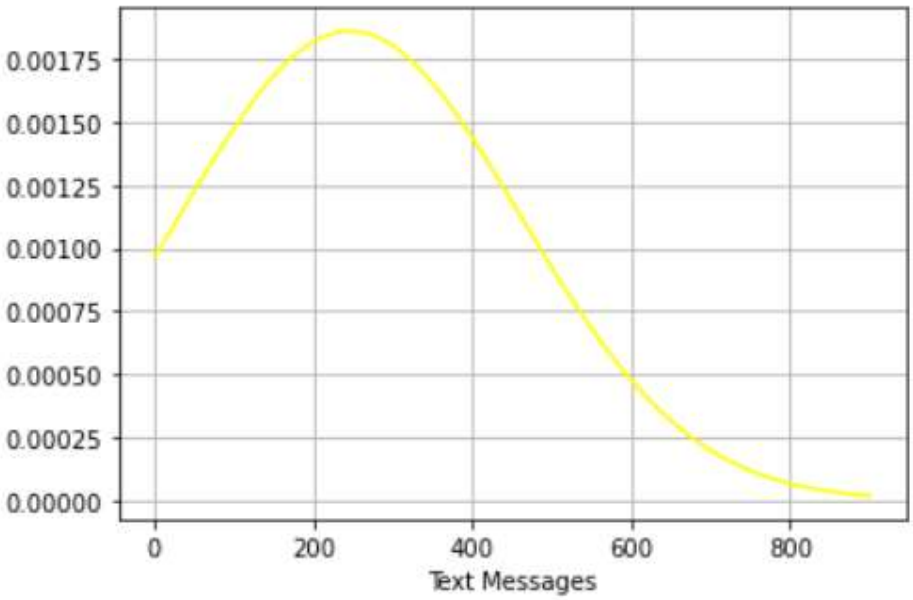


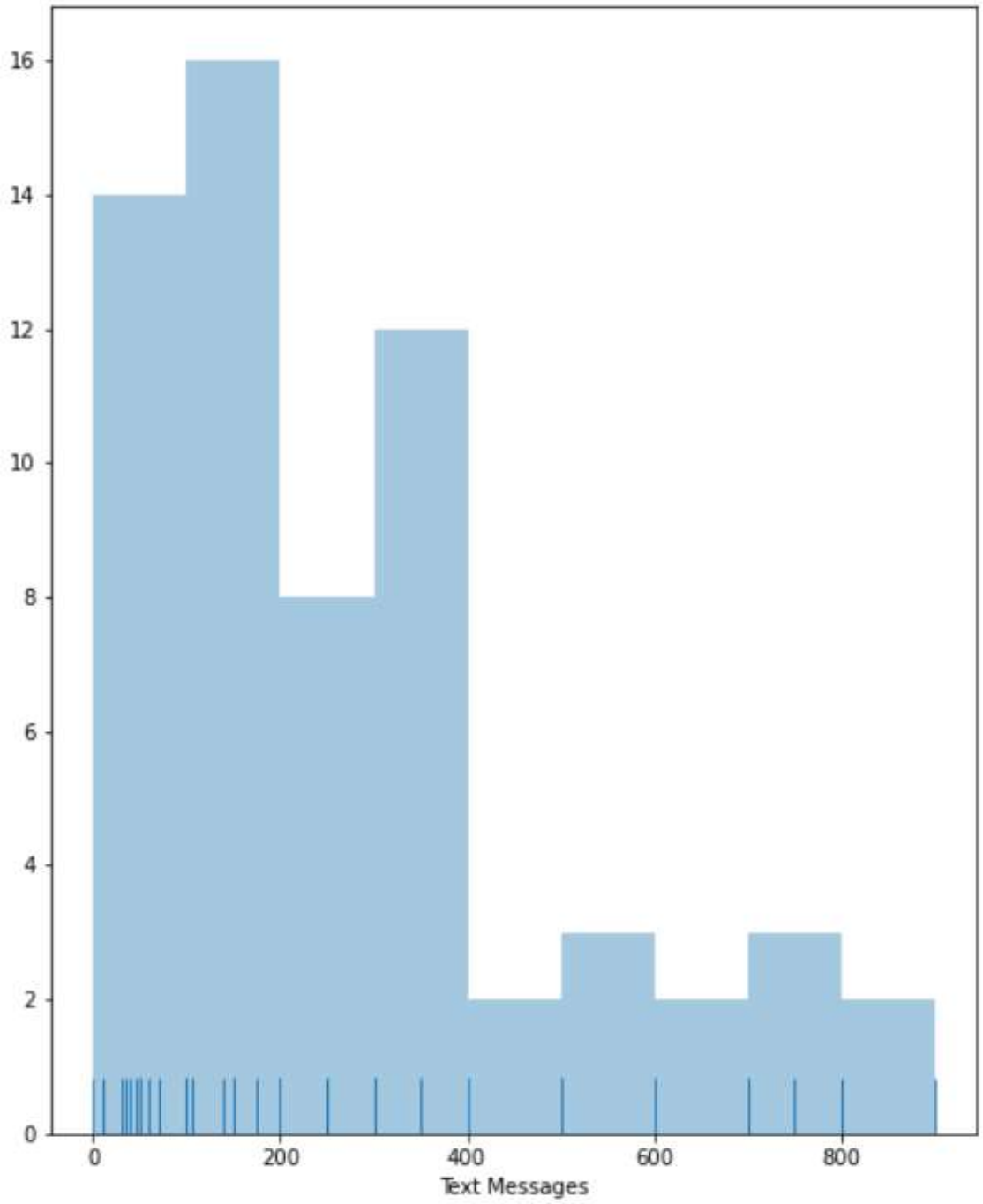
The Spending is showing a perfectly normal distribution until the value of 850.

This shows that 68% of data fall between +/- 221.953 from the mean, 95% of data fall between +/- 443.906 from the mean and 99.7% of all the data fall between +/- 665.859 from the mean.

It is clearly seen that the Spending distribution is a right tailed, or a positively skewed.

Text Messages variable: To show that the Text Messages follows a Normal distribution, we plotted the probability distribution curve using matplotlib and histogram as follows:


0.00175
0.00150
0.00125
0.00100
0.00075
0.00050
0.00025
0.00000
0
200
400
600
800
Text Messages



The Text Messages is showing a normal distribution until the first 500 counts.

This shows that 68% of data fall between +/- 214.465 from the mean, 95% of data fall between +/- 428.93 from the mean and 99.7% of all the data fall between +/- 643.395 from the mean.

The Text Messages distribution is a fully right skewed or a right tailed distribution.