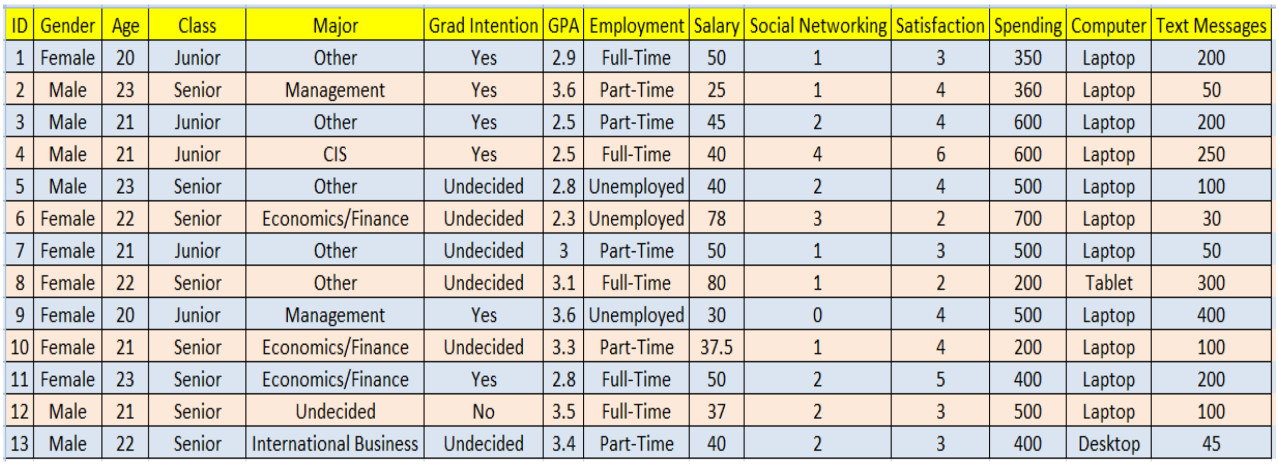
Problem 2 Statement

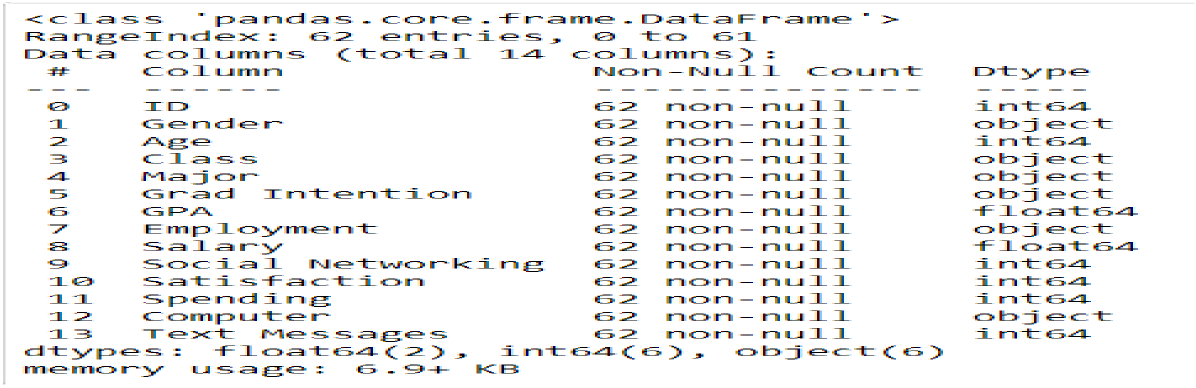
The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. CMSU creates and distributes a survey of 14 questions and receives responses from 62 undergraduates (stored in the Survey.csv file).

Exploratory Data Analysis :



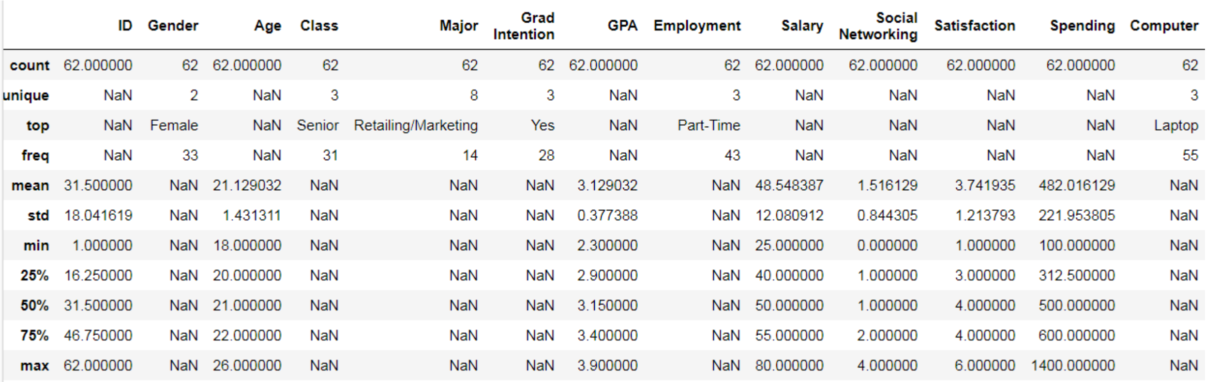
ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
1	Female	20	Junior	Other	Yes	2.9	Full-Time	50	1	3	350	Laptop	200
2	Male	23	Senior	Management	Yes	3.6	Part-Time	25	1	4	360	Laptop	50
3	Male	21	Junior	Other	Yes	2.5	Part-Time	45	2	4	600	Laptop	200
4	Male	21	Junior	CIS	Yes	2.5	Full-Time	40	4	6	600	Laptop	250
5	Male	23	Senior	Other	Undecided	2.8	Unemployed	40	2	4	500	Laptop	100
6	Female	22	Senior	Economics/Finance	Undecided	2.3	Unemployed	78	3	2	700	Laptop	30
7	Female	21	Junior	Other	Undecided	3	Part-Time	50	1	3	500	Laptop	50
8	Female	22	Senior	Other	Undecided	3.1	Full-Time	80	1	2	200	Tablet	300
9	Female	20	Junior	Management	Yes	3.6	Unemployed	30	0	4	500	Laptop	400
10	Female	21	Senior	Economics/Finance	Undecided	3.3	Part-Time	37.5	1	4	200	Laptop	100
11	Female	23	Senior	Economics/Finance	Yes	2.8	Full-Time	50	2	5	400	Laptop	200
12	Male	21	Senior	Undecided	No	3.5	Full-Time	37	2	3	500	Laptop	100
13	Male	22	Senior	International Business	Undecided	3.4	Part-Time	40	2	3	400	Desktop	45

We are provided with the above data set of 62 rows and 14 columns. Among the available columns 6 are categorical type, 6 are integer type and 2 are float type . The data has no Null values.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62 entries, 0 to 61
Data columns (total 14 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   ID                  62 non-null     int64
 1   Gender              62 non-null     object
 2   Age                 62 non-null     int64
 3   Class               62 non-null     object
 4   Major               62 non-null     object
 5   Grad Intention      62 non-null     object
 6   GPA                 62 non-null     float64
 7   Employment          62 non-null     object
 8   Salary              62 non-null     float64
 9   Social Networking   62 non-null     int64
 10  Satisfaction        62 non-null     int64
 11  Spending            62 non-null     int64
 12  Computer            62 non-null     object
 13  Text Messages       62 non-null     int64
dtypes: float64(2), int64(6), object(6)
memory usage: 6.9+ KB
```

Descriptive statistics for the dataset:



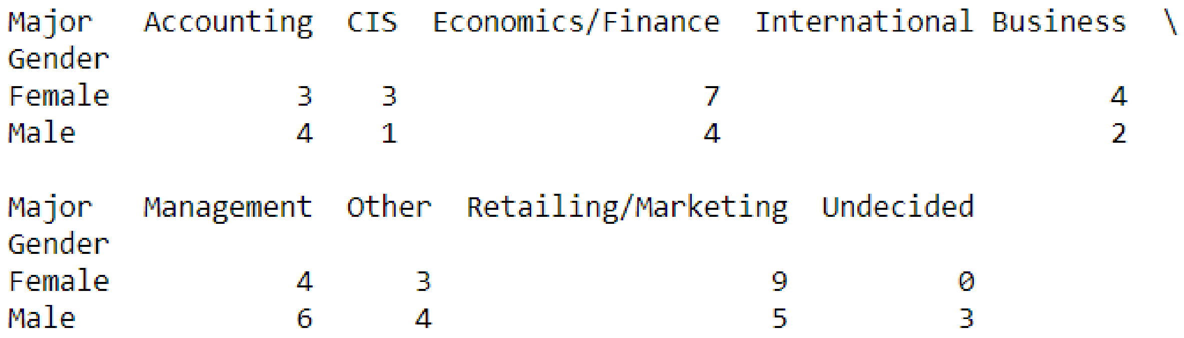
	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer
count	62.000000	62	62.000000	62	62	62	62.000000	62	62.000000	62.000000	62.000000	62.000000	62
unique	NaN	2	NaN	3	8	3	NaN	3	NaN	NaN	NaN	NaN	3
top	NaN	Female	NaN	Senior	Retailing/Marketing	Yes	NaN	Part-Time	NaN	NaN	NaN	NaN	Laptop
freq	NaN	33	NaN	31	14	28	NaN	43	NaN	NaN	NaN	NaN	55
mean	31.500000	NaN	21.129032	NaN	NaN	NaN	3.129032	NaN	48.548387	1.516129	3.741935	482.016129	NaN
std	18.041619	NaN	1.431311	NaN	NaN	NaN	0.377388	NaN	12.080912	0.844305	1.213793	221.953805	NaN
min	1.000000	NaN	18.000000	NaN	NaN	NaN	2.300000	NaN	25.000000	0.000000	1.000000	100.000000	NaN
25%	16.250000	NaN	20.000000	NaN	NaN	NaN	2.900000	NaN	40.000000	1.000000	3.000000	312.500000	NaN
50%	31.500000	NaN	21.000000	NaN	NaN	NaN	3.150000	NaN	50.000000	1.000000	4.000000	500.000000	NaN
75%	46.750000	NaN	22.000000	NaN	NaN	NaN	3.400000	NaN	55.000000	2.000000	4.000000	600.000000	NaN
max	62.000000	NaN	26.000000	NaN	NaN	NaN	3.900000	NaN	80.000000	4.000000	6.000000	1400.000000	NaN

There are 2 unique values in column **Gender**, out of which ‘**Female**’ appears mostly with frequency of 33 times, 3 unique values in column **Class**, out of which ‘**Senior**’ appears mostly with frequency of 31 times , 8 unique values in column **Major**, out of which ‘**Retailing/Marketing’** appears mostly with frequency of 14 times, 3 unique values in column **Grad Intention**, out of which ‘**Yes’** appears mostly with frequency of 28 times , 3 unique values in column **Employment**, out of which ‘**Part-Time’** appears mostly with frequency of 43 times and also 3 unique values in column **Computer** ,out of which ‘**Laptop**’ has highest repetition with frequency of 55 times.

Part I

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

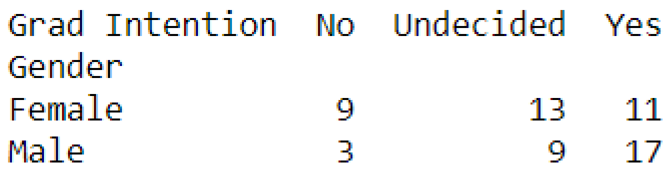
2.1.1. Gender and Major



```
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
```

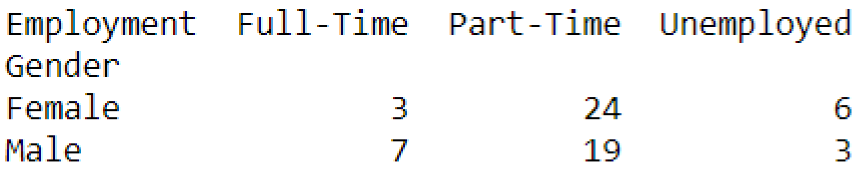
2.1.2. Gender and Grad Intention



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

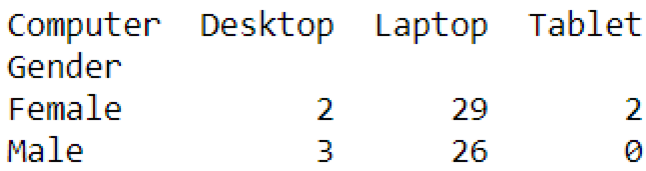
\

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 a representative of the population of CMSU. Based on the data, answer the following questions:

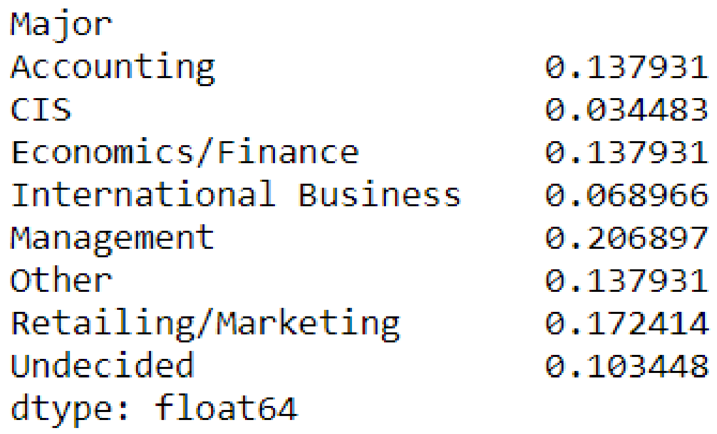
2.2.1. What is the probability that a randomly selected CMSU student will be male?

Probability that a randomly selected CMSU student will be male = 29/62 = 0.467742

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

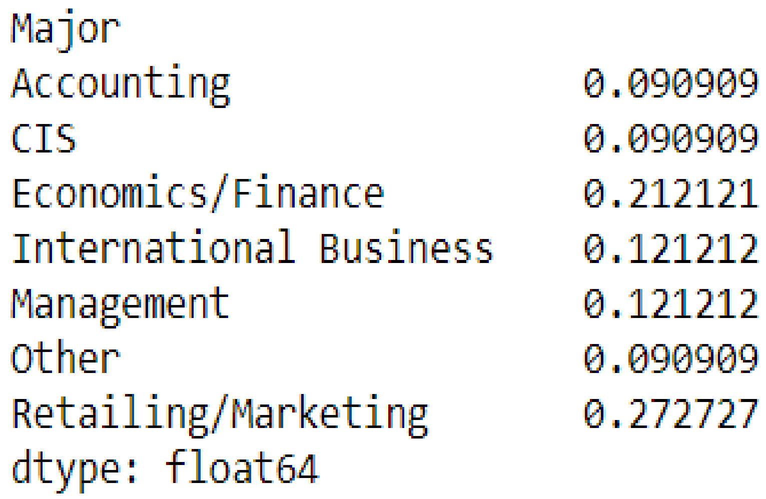
Probability that a randomly selected CMSU student will be female = 33/62 = 0.532258

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



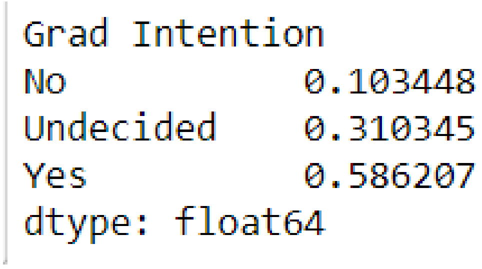
```
Major
Accounting                0.137931
CIS                       0.034483
Economics/Finance         0.137931
International Business    0.068966
Management                0.206897
Other                     0.137931
Retailing/Marketing       0.172414
Undecided                 0.103448
dtype: float64
```

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



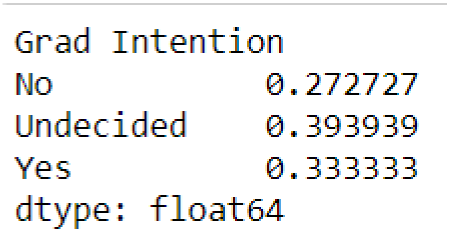
```
Major
Accounting                0.090909
CIS                       0.090909
Economics/Finance         0.212121
International Business    0.121212
Management                0.121212
Other                     0.090909
Retailing/Marketing       0.272727
dtype: float64
```

2.2.3. Find the conditional probability of intent to graduate, given that the student is a male.



```
Grad Intention
No               0.103448
Undecided        0.310345
Yes              0.586207
dtype: float64
```

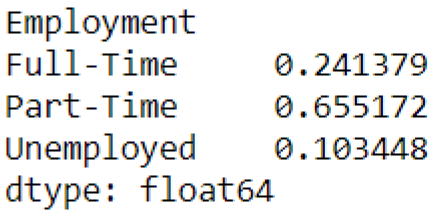
Find the conditional probability of intent to graduate, given that the student is a female.



```
Grad Intention
No               0.272727
Undecided        0.393939
Yes              0.333333
dtype: float64
```

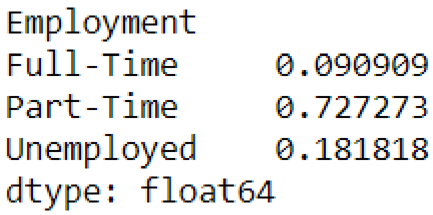
2.2.4. Find the conditional probability of employment status for the male students as well as for the female students.

Conditional probability of employment status for male students



```
Employment
Full-Time      0.241379
Part-Time      0.655172
Unemployed     0.103448
dtype: float64
```

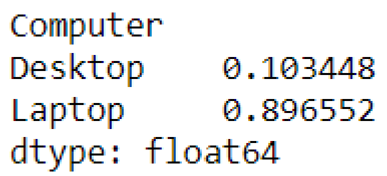
Conditional probability of employment status for female students



```
Employment
Full-Time      0.090909
Part-Time      0.727273
Unemployed     0.181818
dtype: float64
```

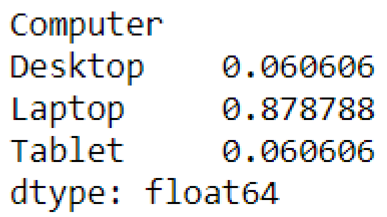
2.2.5. Find the conditional probability of laptop preference among the male students as well as among the female students.

conditional probability of laptop preference among the male students



```
Computer
Desktop    0.103448
Laptop     0.896552
dtype: float64
```

conditional probability of laptop preference among the female students

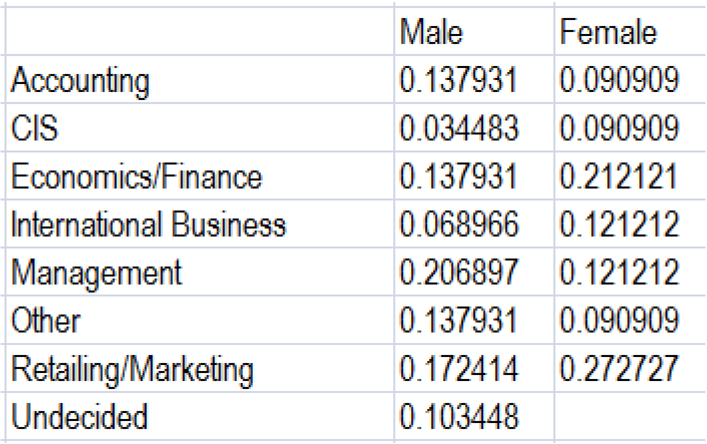


```
Computer
Desktop    0.060606
Laptop     0.878788
Tablet     0.060606
dtype: float64
```

2.3. Based on the above probabilities, do you think that the column variable in each case is independent of Gender?

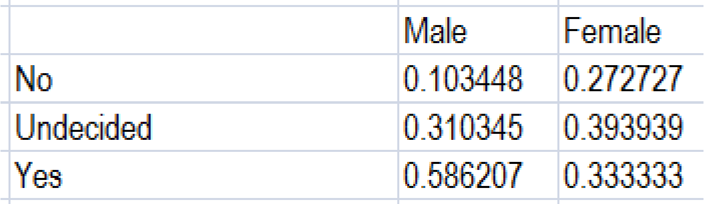
Justify your comment in each case.

- Considering column variable ‘Major’ we can see that none of the probability values are same. So we can say this variable is not independent of Gender



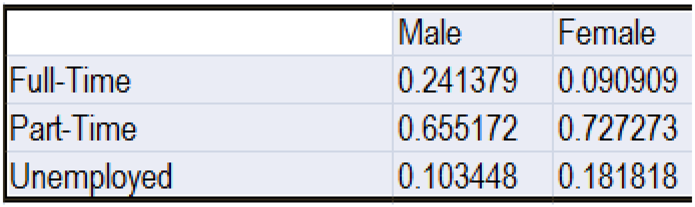
	Male	Female
Accounting	0.137931	0.090909
CIS	0.034483	0.090909
Economics/Finance	0.137931	0.212121
International Business	0.068966	0.121212
Management	0.206897	0.121212
Other	0.137931	0.090909
Retailing/Marketing	0.172414	0.272727
Undecided	0.103448	

- Considering column variable ‘Grad Intention’ we can see that none of the probability values are same. So we can say this variable is not independent of Gender



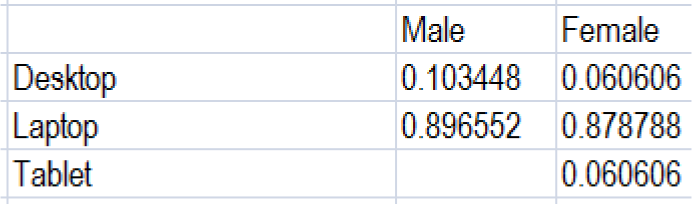
	Male	Female
No	0.103448	0.272727
Undecided	0.310345	0.393939
Yes	0.586207	0.333333

- Considering column variable ‘Employment’ we can see that none of the probability values are same. So we can say this variable is not independent of Gender



	Male	Female
Full-Time	0.241379	0.090909
Part-Time	0.655172	0.727273
Unemployed	0.103448	0.181818

- Considering column variable ‘Computer’ we can see that none of the probability values are same. So we can say this variable is not independent of Gender



	Male	Female
Desktop	0.103448	0.060606
Laptop	0.896552	0.878788
Tablet		0.060606

2.4. Note that there are three numerical (continuous) variables in the data set, Salary, Spending and Text Messages. For each of them comment whether they follow a normal distribution.
Write a note summarizing your conclusions.
[Recall that symmetric histogram does not necessarily mean that the underlying distribution is symmetric]

* Considering column variable **Salary** , we have

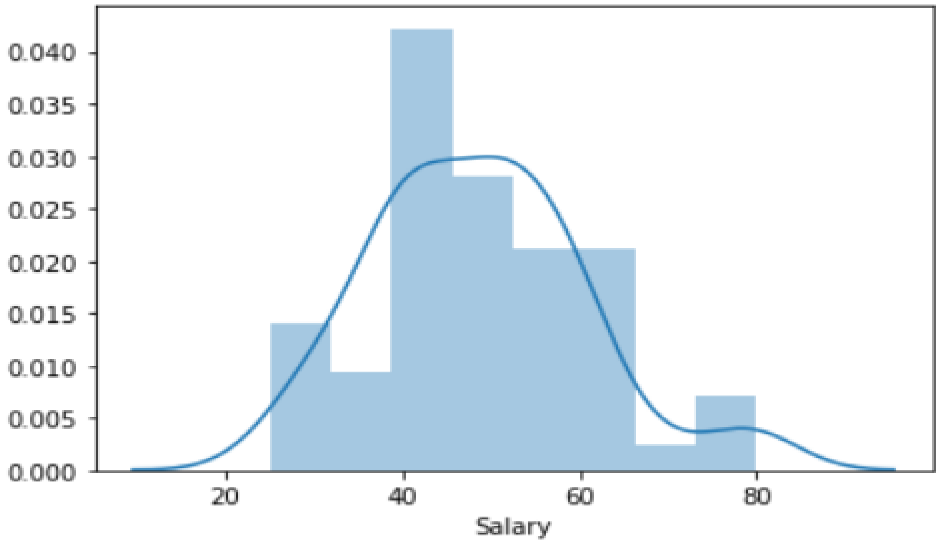
Count = 62
Mean = 48.548387
Std = 12.080912

As per empirical rule , 68-95-99 point rule ,

68% data should be within mean+/-one std i.e., between values 36.467 & 60.629
Total no of values between values 36.467 & 60.629 are 49
Total no of values in column = 62
Probabilty = 49/62 = 0.790323

95% should be within mean+/- two std i.e, between values 24.387 & 72.710
Total no of values between values 24.387 & 72.710 are 59
Total no of values in column = 62
Probabilty = 59/62 = 0.951613

99.7% should be within mean+/- three std i.e, between values 12.306 & 84.791
Total no of values between values 12.306 & 84.791 are 62
Probabilty = 62/62 = 1



This variable almost follows a normal distribution and it is little right skewed

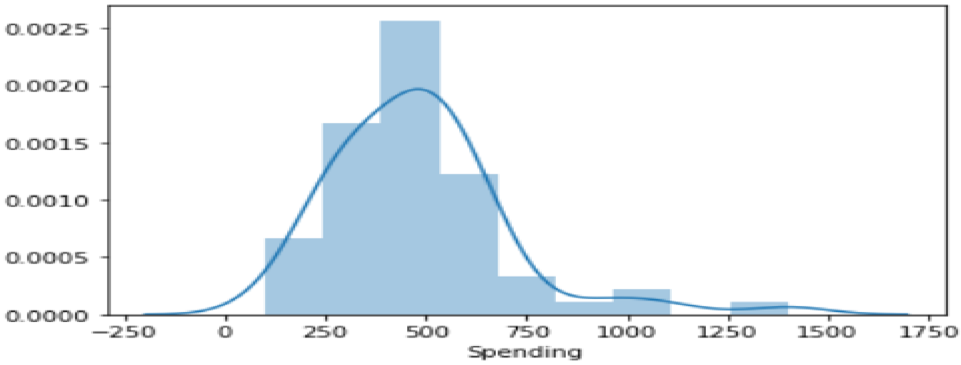
* Considering column variable **Spending** , we have

Count = 62
Mean = 482.016
Std = 221.954
As per empirical rule , 68-95-99 point rule ,

68% data should be within mean+/-one std i.e., between values 260.062 & 703.97
Total no of values between values 260.062 & 703.97 are 50
Total no of values in column = 62
Probabilty = 50/62 = 0.806452

95% should be within mean+/- two std i.e, between values 38.109 & 925.924
Total no of values between values 38.109 & 925.924 are 59
Total no of values in column = 59
Probabilty = 59/62 = 0.951613

99.7% should be within mean+/- three std i.e, between values -183.845 & 1147.878
Total no of values between values -183.845 & 1147.878 are 61
Probabilty = 61/62 = 0.983871



This variable doesn’t follows a normal distribution and it is highly right skewed

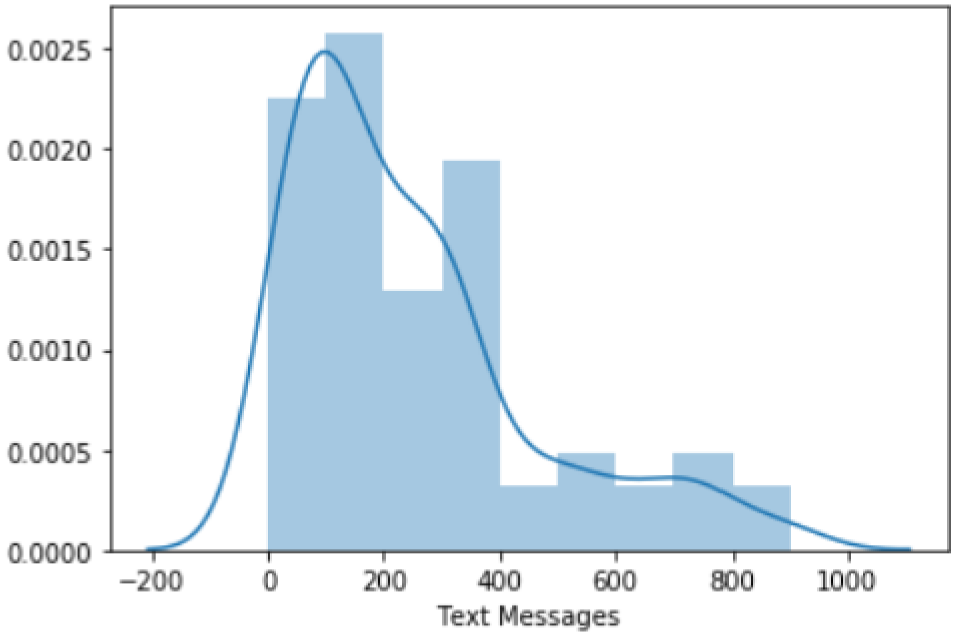
* Considering column variable **Text Messages** , we have

Count = 62
Mean = 246.21
Std = 214.466
As per empirical rule , 68-95-99 point rule ,

68% data should be within mean+/-one std i.e., between values 31.744 & 460.676
Total no of values between values 31.744 & 460.676 are 49
Total no of values in column = 62
Probabilty = 49/62 = 0.790322

95% should be within mean+/- two std i.e, between values -182.722 & 675.142
Total no of values between values -182.722 & 675.142 are 57
Total no of values in column = 57
Probabilty = 57/62 = 0.91935

99.7% should be within mean+/- three std i.e, between values -397.188 & 889.608
Total no of values between values -397.188 & 889.608 are 61
Probabilty = 61/62 = 0.983871



This variable doesn't follows a normal distribution and it is right skewed