Problem 2

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).

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

contingency table is tabular representation of data and also Summation of row and columns.

2.1.1. Gender and Major

We will use crosstab function from Pandas library to plot Continency table.

```
# Gender vs Major
pd.crosstab(df['Gender'],df['Major'],margins=True)
```

In above code we are plotting Gender vs Major. Then Output can be seen. We can notice here All column and row, is summation of respective rows and columns.

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

2.1.2. Gender and Grad Intention

We will use crosstab function from Pandas library to plot Continency table.

```
pd.crosstab(df['Gender'],df['Grad Intention'],margins=True)
```

In above code we are plotting Gender vs Grad Intention. Then Output can be seen. We can notice here All column and row, is summation of respective rows and columns.

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

2.1.3. Gender and Employment

We will use crosstab function from Pandas library to plot Continency table.

```
pd.crosstab(df['Gender'],df['Employment'],margins=True)
```

In above code we are plotting Gender vs Grad Intention. Then Output can be seen. We can notice here All column and row, is summation of respective rows and columns.

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

2.1.4. Gender and Computer

We will use crosstab function from Pandas library to plot Continency table.

```
pd.crosstab(df['Gender'],df['Computer'],margins=True)
```

In above code we are plotting Gender vs Grad Intention. Then Output can be seen. We can notice here All column and row, is summation of respective rows and columns.

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

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?

To calculate probability we need total count, Male count, Female Count.

```
# to calculate probability we will first plot contigency table and its Row/column sum.
pd.crosstab(df['Gender'],df['Gender'],margins=True)
```

From below table we can get all the Male Female count easily.

Gender	Female	Male	All
Gender			
Female	33	0	33
Male	0	29	29
All	33	29	62

Now lets type below code, In which total male count is 29, Total Female count is 33, Total male + Female count will be 33.

Probability of Male will be = Total Male Count/ Total Student Count

```
# probability that a randomly selected CMSU student will be male = P_male
# P_male = No. of male/no of total students

# from above table we can get the values
total_male = 29
total_female = 33
total_mf=29+33

P_male= total_male/total_mf
P_male

print('probability that a randomly selected CMSU student will be male is',P_male)

# probability that a randomly selected CMSU student will be male=0.4677
```

Here we got probability of student will be male as 0.4677

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

To calculate probability we need total count, Male count, Female Count.

```
# probability that a randomly selected CMSU student will be male = P_female

P_female = total_female/total_mf
P_female
# probability that a randomly selected CMSU student will be female=0.5322
```

Total Female count is 33 and Total student count is 62

Probability that a randomly selected CMSU student will be female is 0.532258064516129.

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.

Here we have to find conditional probability, Student is male given

If we convert this problem into equation, it will be

P(Majors|Male) = P(Majors intersection Male)/P(Male)

```
#  We have to find P(Majors|Male)=P(Majors and Male)/ P(Male)

# Majors: Accounting,CIS,Economics/Finance,International Business,Management,Other,Retailing/Marketing,Undecided,All

# We will first present a dataframe which consist distribution of 29 male students according to major chosen-->Dist_Male
df_Male=pd.crosstab(df['Gender'],df['Major']).T
df_Male

# We will reset index to get Major and male as columns
Dist_Male=pd.DataFrame(data=df_Male, columns=['Male']).reset_index()
Dist_Male
```

We are trying to print a dataframe which will consist a tabular data, conditional probability among different majors, the below Dist_Male gives male count through different majors.

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

Now we create a null list con_prm.

After that by using for loop we will append all values in it.

I will be the specific male count for particular major.

29 will be the total count and to convert it into percent we will multiply it by 100.

```
# Creating a null list after that will append output of for loop to it
con_prm=[]


Dist_Male

for i in Dist_Male['Male'].values:
    con_prm.append(i/29*100)

con_prm
# here we have created con_prm list, by appending new values. it has conditional probabability of diff majors among
# male students

# Major_list is array of all Majors
Major_list=df['Major'].unique()
Major_list
```

Now we have Conditional Probabilities list and Major list, so we will create a new Dataframe by using dictionary

convert dictionary to dataframe by using pandas library

We can see all the conditional probabability of diff majors among male students

```
# Now we have Conditional Probabilities list and Major list, so we will create a new Dataframe by using dictionary

dict = {'Major': Major_list, 'Conditional Probability of Male': con_prm}

# Now convert dictionary to dataframe by using pandas library
df_m = pd.DataFrame(dict)
df_m

# We can see all the conditional probabability of diff majors among male students
```

The Major vise male conditional probability will be

	Major	Conditional Probability of Male
0	Other	13.793103
1	Management	3.448276
2	CIS	13.793103
3	Economics/Finance	6.896552
4	Undecided	20.689655
5	International Business	13.793103
6	Retailing/Marketing	17.241379
7	Accounting	10.344828

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

This problem is similar to above problem, Only difference is Female student is given

Here we will create a null list con_prf.

After that by using for loop we will append all values in it.

It will be the specific female count for particular major.

33 will be the total female count and to convert it into percent we will multiply it by 100.

```
# Creating a null list after that will append output of for loop to it
con_prf=[]

# First we will create sub table by selecting only Female column
Dist_Female=pd.DataFrame(data=df_Male,columns=['Female'])
Dist_Female

# For loop iterating through Female values, we divide this value by 33 as total female values are 33,
# multiplying it by 100 means converting it to Percent
for i in Dist_Female['Female'].values:
    con_prf.append(i/33*100)


# After appenidng values we are Selecting only values with index no 0 to 7
con_pr_female=con_prf[:8]
con_pr_female
Major_list=df['Major'].unique()
Major_list
```

Now we have Conditional Probabilities list and Major list, so we will create a new Dataframe by using dictionary

Now convert dictionary to dataframe by using pandas library,We can see all the conditional probabability of diff majors among female students

```
# Now we have Conditional Probabilities list and Major list, so we will create a new Dataframe by using dictionary

dict = {'Major': Major_list, 'Conditional Probability of Female': con_pr_female}

# Now convert dictionary to dataframe by using pandas library
df_f = pd.DataFrame(dict)
df_f

# We can see all the conditional probabability of diff majors among female students
```

The Output will be All the conditional Probability of different majors among female students

	Major	Conditional Probability of Female
0	Other	9.090909
1	Management	9.090909
2	CIS	21.212121
3	Economics/Finance	12.121212
4	Undecided	12.121212
5	International Business	9.090909
6	Retailing/Marketing	27.272727
7	Accounting	0.000000

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.

We have to find probability where random student is male AND he intends to Graduate

Probability of Male = 29/62

Probability of Male and he intends to Grad = 17/29 (Yes Count)

Total Probabilty of Male and he intends to graduate = Probability of Male * Probability of Male and he intends to Graduate

Probability where random student is male AND he intends to Graduate = 27.4193%

```
#We have to find probability where random student is male AND he intends to Graduate

Pr_Male = 29/62
Pr_Male

Pr_Male_intendsGrad = 17/29
Pr_Male_intendsGrad

Total_Prob_Male= Pr_Male * Pr_Male_intendsGrad
Total_Prob_Male

Percentage_Total_Prob_Male=Total_Prob_Male*100
Percentage_Total_Prob_Male

print('Probability where random student is male AND he intends to Graduate =',Percentage_Total_Prob_Male,'%')

# Probability where random student is male AND he intends to Graduate = 27.4193%
```

```
Probability where random student is male AND he intends to Graduate = 27.419354838709676 %
```

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

Here we first plot the count of Male, Female who have computer.

```
pd.crosstab(df['Gender'],df['Computer'],margins=True)
```

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

Probability of Female Student is getting selected is 33 out of 62 students.

Now Probability of female and not having a laptop will be is = Pr_female * Pr_Female_Nolaptop

```
Pr_female = 33/62
Pr_female

Pr_Female_Nolaptop=4/33
Pr_Female_Nolaptop

Pr_female_No_Laptop= Pr_female * Pr_Female_Nolaptop
Pr_female_No_Laptop

#  probability that a randomly selected student is a female and does NOT have a laptop is 0.06451612903225806

Percentage_Pr_female_No_Laptop = Pr_female_No_Laptop * 100
Percentage_Pr_female_No_Laptop

# probability that a randomly selected student is a female and does NOT have a laptop is 6.451612903225806%
```

Probability that a randomly selected student is a female and does NOT have a laptop is 0.0645%

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 either a male or has full-time employment?

Probability of Student is male = 29/62 = 0.4677

10 students have Full time employment out of 62

Probability of students have Full time employment = 10/62

Probability that a randomly chosen student is either a male OR has full-time employment

= Probability of Student is male + Probability of students have Full time employment

```
#Probability of Student is male = 29/62 = 0.4677
Pr_Male

#Student has fulll time employment = 10 students have Full time employment out of 62
pd.crosstab(df['Gender'],df['Employment'],margins=True)

Pr_Student_FullTimeEmp = 10/62
Pr_Student_FullTimeEmp

#probability that a randomly chosen student is either a male OR has full-time employment
Pr_Male_or_fulltimeEmp = Pr_Male + Pr_Student_FullTimeEmp
Pr_Male_or_fulltimeEmp

#probability that a randomly chosen student is either a male OR has full-time employment=0.6290322580645161
Percentage_Pr_Male_or_fulltimeEmp = Pr_Male_or_fulltimeEmp*100
Percentage_Pr_Male_or_fulltimeEmp

# Percent probability that a randomly chosen student is either a male OR has full-time employment = 62.903225806451616%
```

Percent probability that a randomly chosen student is either a male OR has full-time employment = 62.903225806451616%

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 interested in calculating Probability(majoring in international business or management|Female).

Pr_Female_Major_in_international_business_or_management

= Probability of Female Majors in _international_business or Majors in management

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

We will apply addition rule as OR is used

Conditional Probability(majoring in international business or management|Female)

= Pr_Female_Major_in_international_business_or_management/Pr_female

Probability(majoring in international business or management|Female)

= (8/33)/(33/62)

```
# Probability(majoring in international business or management|Female )?


Pr_female = 33/62
Pr_female=33/62


Pr_Female_Major_in_international_business_or_management = (4/33)+(4/33)
Pr_Female_Major_in_international_business_or_management=8/33

# Probability(majoring in international business or management|Female )
#           = Pr_Female_Major_in_international_business_or_management/Pr_female

# Lets say,conditional probability that given a female student is randomly chosen, she is majoring in international business
P_x = (8/33)/(33/62)

P_x

# Probability(majoring in international business or management|Female ) = 0.45546372819100095
# In Percentage,it will be 45.546%
```

Probability(majoring in international business or management|Female) = 45.5463728%

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?

Below is the code to plot contingency table of Gender and Intent to Graduate at 2 levels (Yes/No).We will first store gender vs grad intention in dataframe temp. Then we will drop column 'undecided' from it.

```
Temp=pd.crosstab(df['Gender'],df['Grad Intention'])
g=Temp.drop(['Undecided'],axis='columns')
g
```

Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

Above dataframe will give us idea of male and female are Intended to Graduation Yes/No.

We have discareded the total no of undecided Students, so sample space will be 62-22 = 40

Lets say

Event A : Being a Graduate Intention Yes

P(A)=28/40 ..... 28 Studnets said Yes out of 40.

P(A)=0.7

Event B : Being a Female

P(B) = 20/40

P(B)=0.5

Two events are Independednt when P(A intersection B) = P(A).P(B)

=0.7 * 0.5

=0.35

We have discareded the total no of undecided Students, so sample space will be 62-22 = 40

We have 20 Female Students

Student having Grad Intention is 28

11 Female Students has Grad Intention

If 17 students have intention Yes, 11 Students have Intention Yes and they are Female,9 are Female with no intention

Prob_A_intersection_B = 11/(17+11+9)

Prob_A_intersection_B

Prob_A_intersection_B is 0.2972972972972973.....1)

Prob_A * Prob_B = 0.35.....2)

Prob_A_intersection_B is not equal to Prob_A * Prob_B.

Two events are dependednt as P(A intersection B) not equal to P(A).P(B)

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

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

We can put variable G where all the values less than 3 are stored. Then we will perform sum.

```
G=df['GPA']<3
G.sum()
```

17 values are less than 3, Other values are greater than 3 that is 45 and total students are 62.

Probability that his/her GPA is less than 3 = 17/62

```
# 17 values are less than 3, Other values are greater than 3 that is 45 and total students are 62.
# probability that his/her GPA is less than 3 = 17/62

P_gpa_lessthan3=(17/62)*100
P_gpa_lessthan3

print('probability that his/her GPA is less than 3 is %1.2f' % P_gpa_lessthan3 +'%')
```

Probability that his/her GPA is less than 3 is 27.42%

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

Total no of Males = 29

Male those who are having salary equal to greater than 50 = 14

conditional probability that a randomly selected male earns 50 or more

```
# Total no of Males = 29
# Male those who are having salary equal to greater than 50 = 14

# Pr_M = conditional probability that a randomly selected male earns 50 or more

Pr_M = 14/29
Pr_M_percent = Pr_M*100

print('conditional probability that a randomly selected male earns 50 or more %1.2f' % Pr_M_percent +'%')
```

conditional probability that a randomly selected male earns 50 or more 48.28%

2.7.2.b.Find the conditional probability that a randomly selected female earns 50 or more.

Total no of females = 33

female those who are having salary equal to greater than 50 = 18

Pr_F = conditional probability that a randomly selected female earns 50 or more

Pr_F = 18/33

Pr_F_percent = Pr_F*100

```
# Total no of females = 33
# female those who are having salary equal to greater than 50 = 18

# Pr_F = conditional probability that a randomly selected female earns 50 or more

Pr_F = 18/33
Pr_F_percent = Pr_F*100

print('conditional probability that a randomly selected female earns 50 or more %1.2f' % Pr_F_percent +'%')
```

conditional probability that a randomly selected female earns 50 or more 54.55%

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.

This command will give us Mean, Std deviation, 25,50,75% distribution limits.

```
df.describe()
```

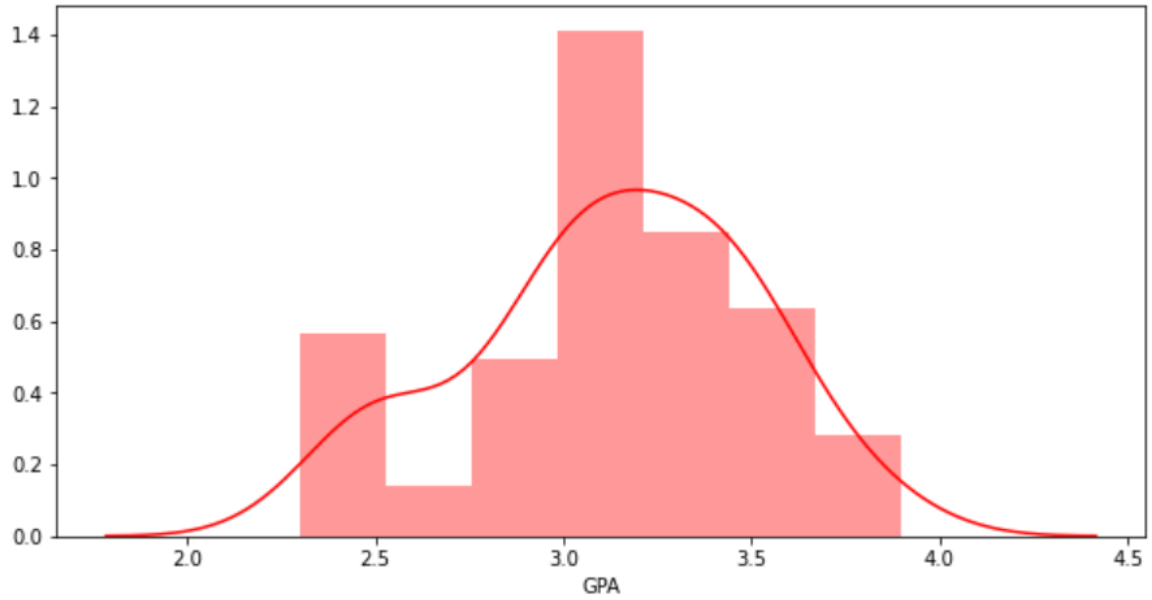
Output will be

	ID	Age	GPA	Salary	Social Networking	Satisfaction	Spending	Text Messages
count	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000
mean	31.500000	21.129032	3.129032	48.548387	1.516129	3.741935	482.016129	246.209677
std	18.041619	1.431311	0.377388	12.080912	0.844305	1.213793	221.953805	214.465950
min	1.000000	18.000000	2.300000	25.000000	0.000000	1.000000	100.000000	0.000000
25%	16.250000	20.000000	2.900000	40.000000	1.000000	3.000000	312.500000	100.000000
50%	31.500000	21.000000	3.150000	50.000000	1.000000	4.000000	500.000000	200.000000
75%	46.750000	22.000000	3.400000	55.000000	2.000000	4.000000	600.000000	300.000000
max	62.000000	26.000000	3.900000	80.000000	4.000000	6.000000	1400.000000	900.000000

Now we will plot boxplot and distplot to analyze it in detail whether it is Normally distributed or not.

For GPA,

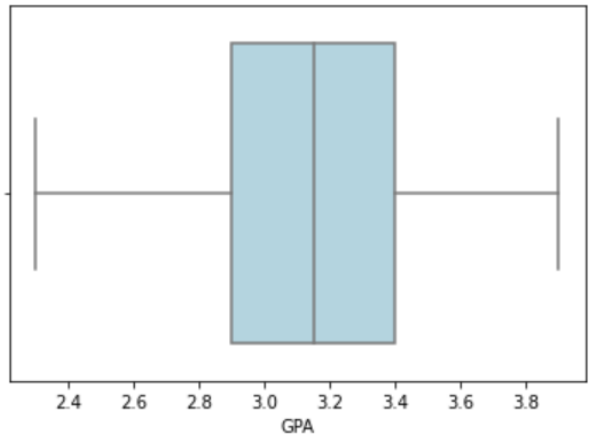
```
plt.figure(figsize= (10,5))
sns.distplot(df['GPA'], color ='Red');
```



If we see above graph visually we can see the points are distributed almost equally towards both side. Mean and Median Values of each sample are not much different.The GPA looks more symmetrically distributed.

Lets plot Boxplot to confirm,

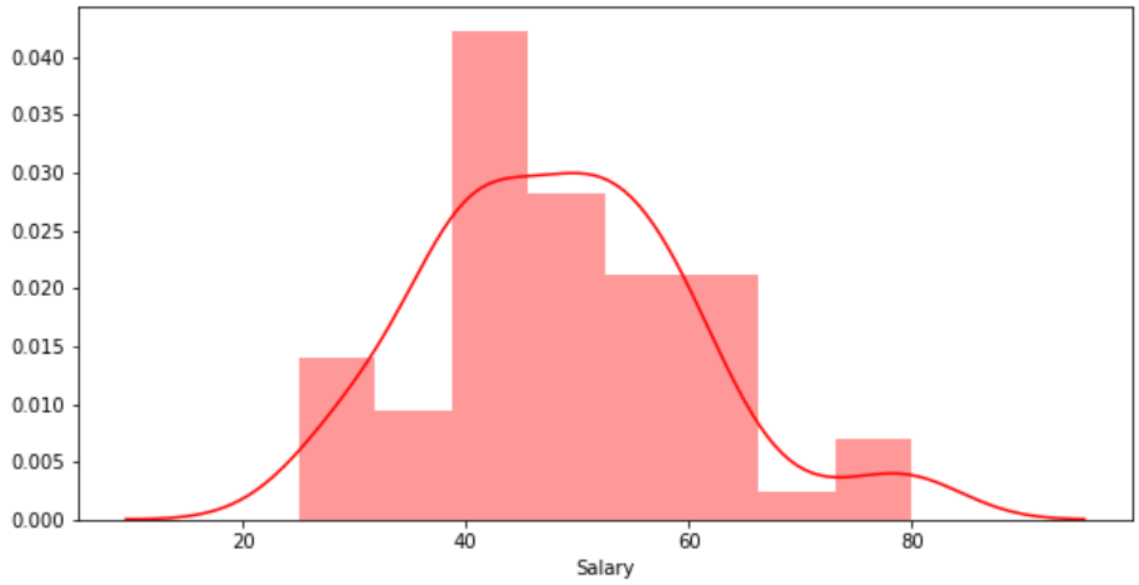
```
sns.boxplot(x= df['GPA'], color='lightblue')
plt.show()
```



Boxplot also shows similar behaviour, data is almost equally distributed we can confirm this by Describe() dataframe, and Mean and Median Values of each sample are not much different.

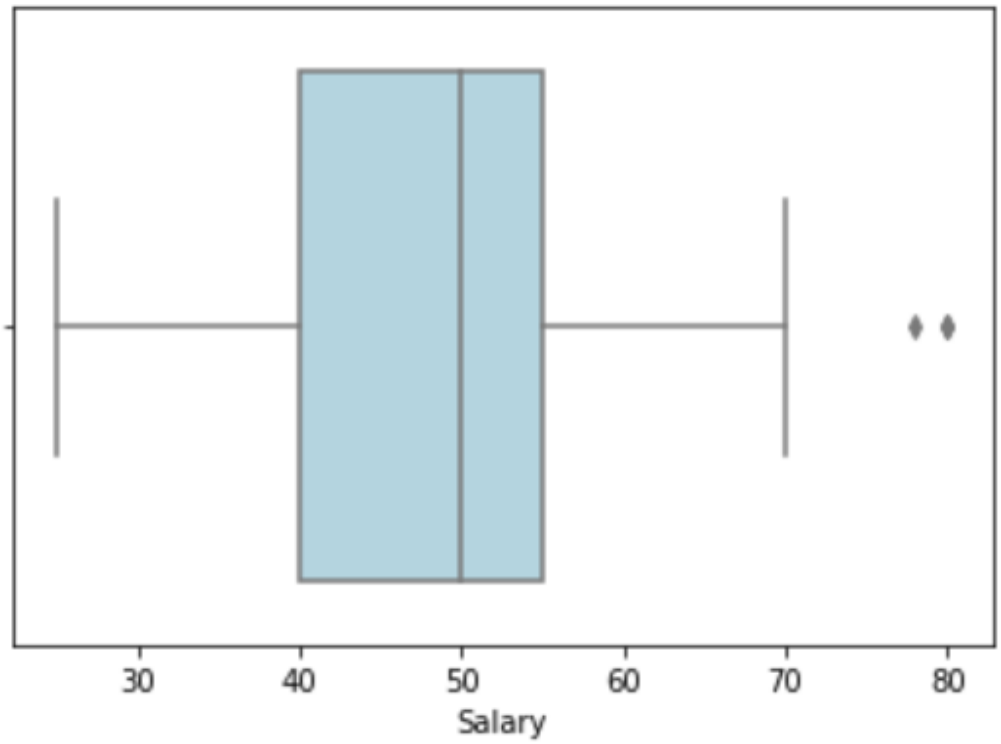
For Salary,

```
plt.figure(figsize= (10,5))
sns.distplot(df['Salary'], color ='Red');
```



If we see above graph visually we can see the points are distributed almost equally towards both side. but we are not sure, so we will analyse through boxplot

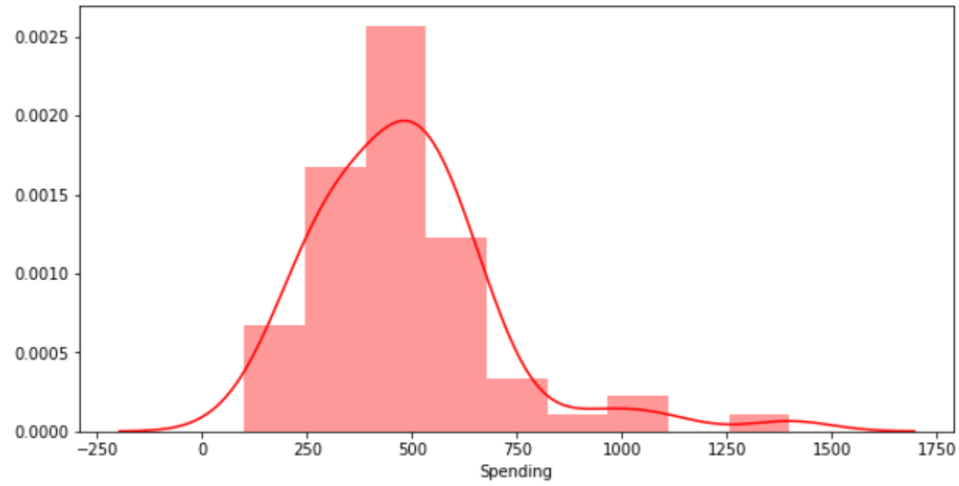
```
sns.boxplot(x= df['Salary'], color='lightblue')
plt.show()
```



Here we can easily see boxplot, the quartiles are not distributed equally so it is not normally distributed.

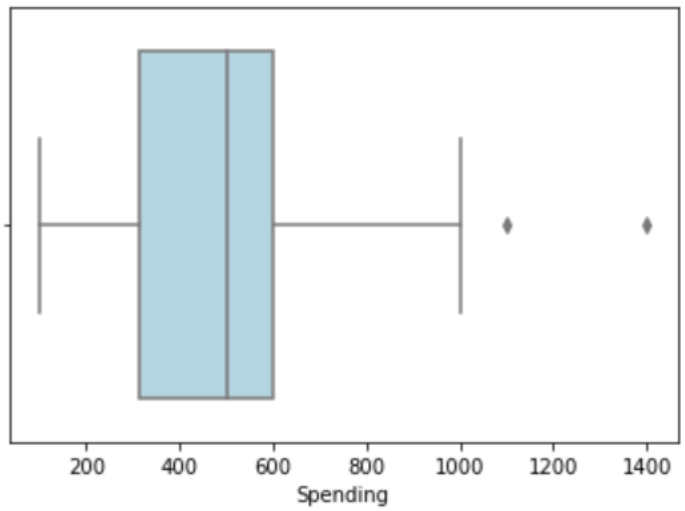
For Spending,

```
plt.figure(figsize= (10,5))
sns.distplot(df['Spending'], color ='Red');
```



In above distplot the data distribution is somewhat normal, Not a exact Normal distribution.

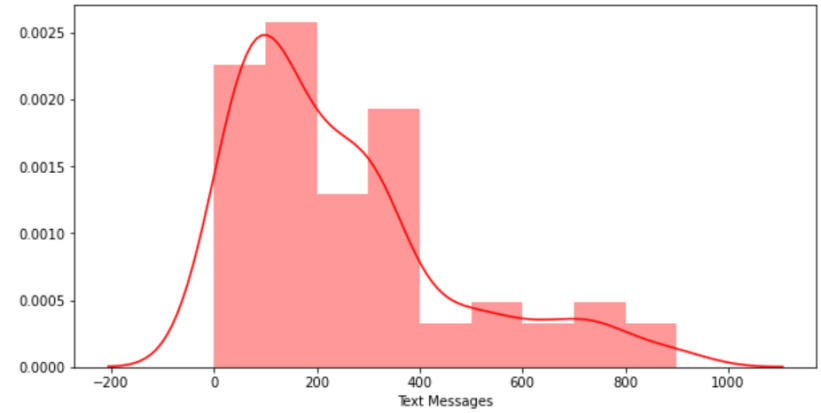
```
sns.boxplot(x= df['Spending'], color='lightblue')
plt.show()
```



By above box plot we can confirm that data is not distributed equally, It is not normally distributed

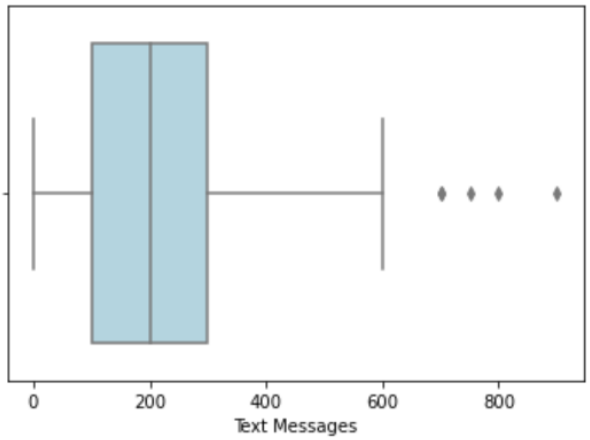
For Text Messages,

```
plt.figure(figsize= (10,5))
sns.distplot(df['Text Messages'], color ='Red');
```



It can be clearly seen in above plot Text messages count is not Normally distributed.

```
sns.boxplot(x= df['Text Messages'], color='lightblue')
plt.show()
```



In above box we can see data Text Messages is not normally distributed, Mean and median are not equal.

Overall only GPA is Normally distributed while Others are not.