$Problem 1$

WHOLESALE COSTOMER DATA ANALYSIS

In [1]:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import scipy.stats as stats
import math
sns.set(color_codes=True)
from scipy.stats    import ttest_1samp, ttest_ind
```

In [2]:

```
whsale=pd.read_csv("Wholesale+Customers+Data.csv")
```

In [3]:

```
whsale.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 440 entries, 0 to 439
Data columns (total 9 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Buyer/Spender     440 non-null    int64
 1   Channel           440 non-null    object
 2   Region            440 non-null    object
 3   Fresh             440 non-null    int64
 4   Milk              440 non-null    int64
 5   Grocery           440 non-null    int64
 6   Frozen            440 non-null    int64
 7   Detergents_Paper  440 non-null    int64
 8   Delicatessen      440 non-null    int64
dtypes: int64(7), object(2)
memory usage: 31.1+ KB
```

In [4]:

```
whsale.size
```

Out[4]:

```
3960
```

In [5]:

```
whsale.head(10)
```

Out[5]:

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicate
0	1	Retail	Other	12669	9656	7561	214	2674	
1	2	Retail	Other	7057	9810	9568	1762	3293	
2	3	Retail	Other	6353	8808	7684	2405	3516	
3	4	Hotel	Other	13265	1196	4221	6404	507	
4	5	Retail	Other	22615	5410	7198	3915	1777	
5	6	Retail	Other	9413	8259	5126	666	1795	
6	7	Retail	Other	12126	3199	6975	480	3140	
7	8	Retail	Other	7579	4956	9426	1669	3321	
8	9	Hotel	Other	5963	3648	6192	425	1716	
9	10	Retail	Other	6006	11093	18881	1159	7425	

Dataset has 9 variables Buyer/ Spender, Channel, Region, Fresh, Milk, Grocery, Frozen, Detergents_Paper & Delicatessen. Channel and Region both are categorical columns while Buyer/ Spender, Fresh, Milk, Grocery, Frozen, Detergents_Paper & Delicatessen are integer.

In [6]:

```
whsale.Region.value_counts()
```

Out[6]:

```
Other      316
Lisbon      77
Oporto      47
Name: Region, dtype: int64
```

In [7]:

```
whsale.Channel.value_counts()
```

Out[7]:

```
Hotel      298
Retail     142
Name: Channel, dtype: int64
```

In [8]:

```
whsale.isnull().sum()
```

Out[8]:

```
Buyer/Spender       0
Channel             0
Region              0
Fresh               0
Milk                0
Grocery             0
Frozen              0
Detergents_Paper    0
Delicatessen        0
dtype: int64
```

In [9]:

```
corr=whsale.corr()
corr
```

Out[9]:

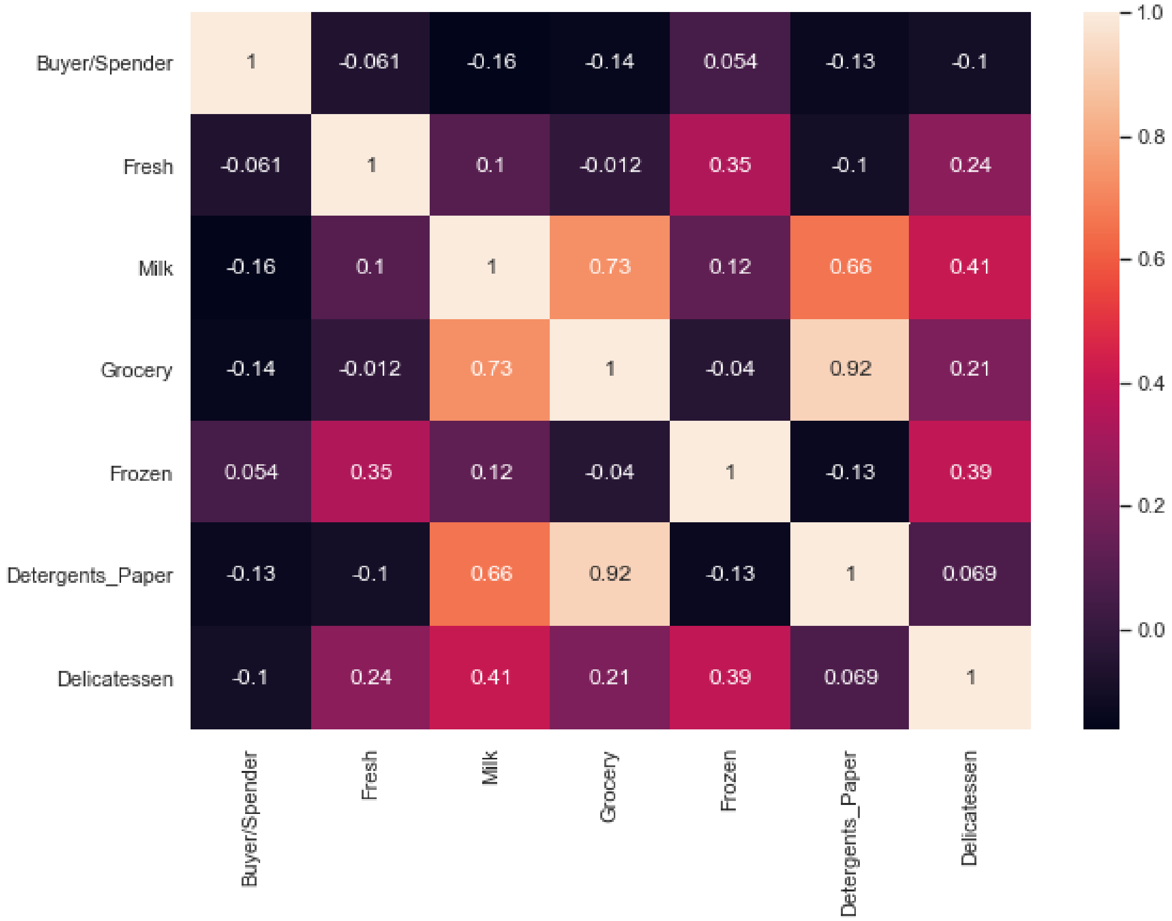
	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper
Buyer/Spender	1.000000	-0.061151	-0.162290	-0.140509	0.053802	-0.134365
Fresh	-0.061151	1.000000	0.100510	-0.011854	0.345881	-0.101953
Milk	-0.162290	0.100510	1.000000	0.728335	0.123994	0.661816
Grocery	-0.140509	-0.011854	0.728335	1.000000	-0.040193	0.924641
Frozen	0.053802	0.345881	0.123994	-0.040193	1.000000	-0.131525
Detergents_Paper	-0.134365	-0.101953	0.661816	0.924641	-0.131525	1.000000
Delicatessen	-0.101845	0.244690	0.406368	0.205497	0.390947	0.069291

In [10]:

```
plt.figure(figsize=(10,7))
sns.heatmap(corr,annot=True)
```

Out[10]:

```
<AxesSubplot:>
```

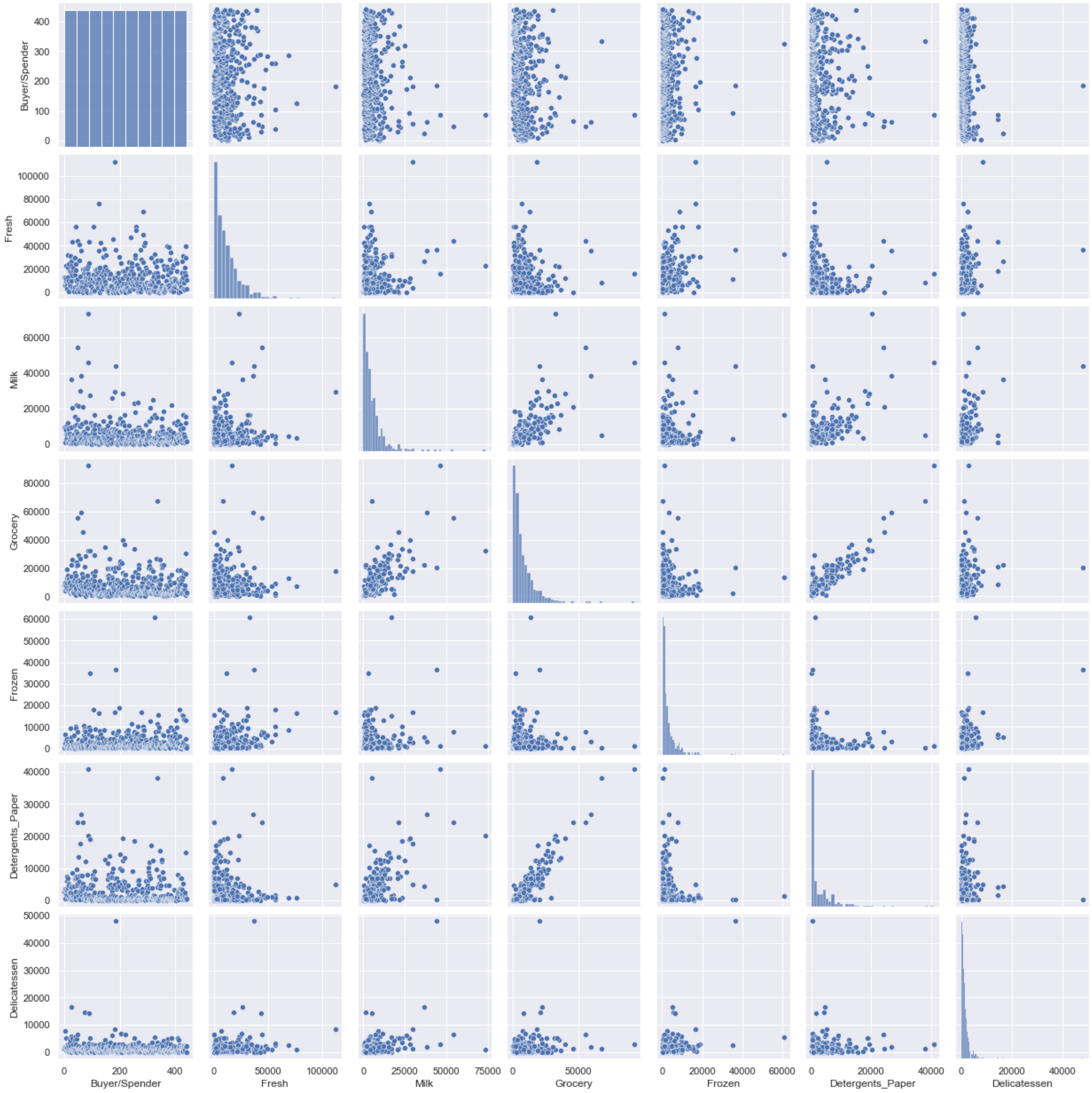


In [11]:

```
sns.pairplot(whsale)
```

Out[11]:

```
<seaborn.axisgrid.PairGrid at 0x1ff3f925730>
```



As per plot shown below, There is a linear relationship between Grocery and Detergents Paper.

Q1.1.1 Use methods of descriptive statistics to summarize data.

In [12]:

```
whsale.describe(include="all")
```

Out[12]:

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen
count	440.000000	440	440	440.000000	440.000000	440.000000	440.000
unique	NaN	2	3	NaN	NaN	NaN	I
top	NaN	Hotel	Other	NaN	NaN	NaN	I
freq	NaN	298	316	NaN	NaN	NaN	I
mean	220.500000	NaN	NaN	12000.297727	5796.265909	7951.277273	3071.931
std	127.161315	NaN	NaN	12647.328865	7380.377175	9503.162829	4854.673
min	1.000000	NaN	NaN	3.000000	55.000000	3.000000	25.000
25%	110.750000	NaN	NaN	3127.750000	1533.000000	2153.000000	742.250
50%	220.500000	NaN	NaN	8504.000000	3627.000000	4755.500000	1526.000
75%	330.250000	NaN	NaN	16933.750000	7190.250000	10655.750000	3554.250
max	440.000000	NaN	NaN	112151.000000	73498.000000	92780.000000	60869.000

There are two unique Channel and 3 region, where Hotel channel is the top most in the channel values column and other region has the top most values in the channel column

In [13]:

```
whsale["varieties_sum"]= whsale["Fresh"]+whsale["Milk"]+whsale["Grocery"]+whsale["Frozen"]+
whsale
```

Out[13]:

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delic
0	1	Retail	Other	12669	9656	7561	214	2674	
1	2	Retail	Other	7057	9810	9568	1762	3293	
2	3	Retail	Other	6353	8808	7684	2405	3516	
3	4	Hotel	Other	13265	1196	4221	6404	507	
4	5	Retail	Other	22615	5410	7198	3915	1777	
...	...	...	...	...	...	...	...	...	
435	436	Hotel	Other	29703	12051	16027	13135	182	
436	437	Hotel	Other	39228	1431	764	4510	93	
437	438	Retail	Other	14531	15488	30243	437	14841	
438	439	Hotel	Other	10290	1981	2232	1038	168	
439	440	Hotel	Other	2787	1698	2510	65	477	

440 rows × 10 columns

Q1.1.2 Which Region and which Channel spent the most?

Q1.1.3 Which Region and which Channel spent the least?

In [14]:

```
pd.DataFrame(whsale.groupby("Channel").varieties_sum.sum())
```

Out[14]:

	varieties_sum
Channel	
Hotel	7999569
Retail	6619931

In [15]:

```
pd.DataFrame(whsale.groupby("Region").varieties_sum.sum())
```

Out[15]:

	varieties_sum
Region	
Lisbon	2386813
Oporto	1555088
Other	10677599

Q1.2. 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.

In [16]:

```
pd.DataFrame(whsale.groupby("Channel").sum())
```

Out[16]:

	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Channel							
Hotel	71034	4015717	1028614	1180717	1116979	235587	421955
Retail	25986	1264414	1521743	2317845	234671	1032270	248988

In [17]:

```
pd.DataFrame(whsale.groupby("Region").sum())
```

Out[17]:

	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	va
Region								
Lisbon	18095	854833	422454	570037	231026	204136	104327	
Oporto	14899	464721	239144	433274	190132	173311	54506	
Other	64026	3960577	1888759	2495251	930492	890410	512110	

Q1.3 On the basis of the descriptive measure of variability, which item shows the most inconsistent behaviour? Which items shows the least inconsistent behaviour?

$Formula$ of $CV = Std/Mean$

In [18]:

```
whsale.describe().T
```

Out[18]:

	count	mean	std	min	25%	50%	75%	max
Buyer/Spender	440.0	220.500000	127.161315	1.0	110.75	220.5	330.25	44
Fresh	440.0	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	1121
Milk	440.0	5796.265909	7380.377175	55.0	1533.00	3627.0	7190.25	734
Grocery	440.0	7951.277273	9503.162829	3.0	2153.00	4755.5	10655.75	927
Frozen	440.0	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	608
Detergents_Paper	440.0	2881.493182	4767.854448	3.0	256.75	816.5	3922.00	408
Delicatessen	440.0	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	479
varieties_sum	440.0	33226.136364	26356.301730	904.0	17448.75	27492.0	41307.50	1998

In [19]:

```
CV_for_Fresh= (12647.328865/12000.297727)
CV_for_Fresh
```

Out[19]:

```
1.0539179237648593
```

In [20]:

```
CV_for_Milk= (7380.377175/5796.265909)
CV_for_Milk
```

Out[20]:

```
1.2732985841005522
```

In [21]:

```
CV_for_Grocery =(9503.162829/7951.277273)
CV_for_Grocery
```

Out[21]:

```
1.1951743729613995
```

In [22]:

```
CV_for_Frozen = (4854.673333/3071.931818)
CV_for_Frozen
```

Out[22]:

```
1.5803323838615222
```

In [23]:

```
cv_for_Detergents_Paper =(4767.854448/2881.493182)
cv_for_Detergents_Paper
```

Out[23]:

```
1.6546471384293562
```

In [24]:

```
CV_for_Delicatessen = (2820.105937/1524.870455)
CV_for_Delicatessen
```

Out[24]:

```
1.849406897322304
```

After calculate the data we show that CV (coefficient of variation) value as per formula $CV = Std/Mean$, As per descriptive $Fresh$ item behaviour are least inconsistent behaviour compare to rest of item.

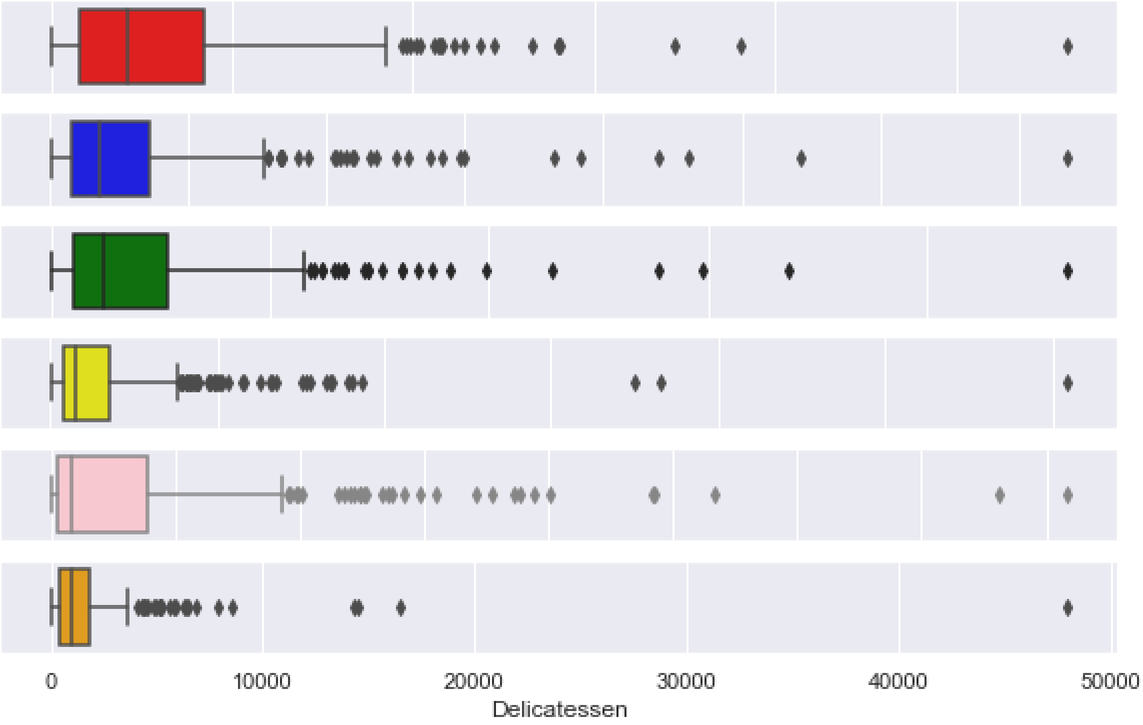
Q1.4 Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.

In [25]:

```
plt.figure(figsize=(10,6))
plt.subplot(6,1,1)
sns.boxplot(x="Fresh",data=whsale,color="red")
plt.subplot(6,1,2)
sns.boxplot(x="Milk",data=whsale,color="blue")
plt.subplot(6,1,3)
sns.boxplot(x="Grocery",data=whsale,color="green")
plt.subplot(6,1,4)
sns.boxplot(x="Frozen",data=whsale,color="yellow")
plt.subplot(6,1,5)
sns.boxplot(x="Detergents_Paper",data=whsale,color="pink")
plt.subplot(6,1,6)
sns.boxplot(x="Delicatessen",data=whsale,color="orange")
```

Out[25]:

```
<AxesSubplot:xlabel='Delicatessen'>
```



As per boxplot we see that all 6 $item$ have $outliers$...

$$Problem - 2$$

Clear Mountain State University Data Analysis

In [26]:

```
cmsu=pd.read_csv("Survey-1.csv")
```

In [27]:

```
cmsu.info()
```

```
<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
```

In [28]:

```
cmsu.size
```

Out[28]:

```
868
```

In [29]:

```
cmsu.isnull().sum()
```

Out[29]:

```
ID                   0
Gender               0
Age                  0
Class                0
Major                0
Grad Intention       0
GPA                  0
Employment           0
Salary               0
Social Networking    0
Satisfaction         0
Spending             0
Computer             0
Text Messages        0
dtype: int64
```

In [30]:

```
cmsu.head()
```

Out[30]:

	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	S
0	1	Female	20	Junior	Other	Yes	2.9	Full-Time	50.0	1	
1	2	Male	23	Senior	Management	Yes	3.6	Part-Time	25.0	1	
2	3	Male	21	Junior	Other	Yes	2.5	Part-Time	45.0	2	
3	4	Male	21	Junior	CIS	Yes	2.5	Full-Time	40.0	4	
4	5	Male	23	Senior	Other	Undecided	2.8	Unemployed	40.0	2	

In [31]:

```
cmsu.describe(include="all")
```

Out[31]:

	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employm
count	62.000000	62	62.000000	62	62	62	62.000000	
unique	NaN	2	NaN	3	8	3	NaN	
top	NaN	Female	NaN	Senior	Retailing/Marketing	Yes	NaN	Part-T
freq	NaN	33	NaN	31	14	28	NaN	
mean	31.500000	NaN	21.129032	NaN	NaN	NaN	3.129032	N
std	18.041619	NaN	1.431311	NaN	NaN	NaN	0.377388	N
min	1.000000	NaN	18.000000	NaN	NaN	NaN	2.300000	N
25%	16.250000	NaN	20.000000	NaN	NaN	NaN	2.900000	N
50%	31.500000	NaN	21.000000	NaN	NaN	NaN	3.150000	N
75%	46.750000	NaN	22.000000	NaN	NaN	NaN	3.400000	N
max	62.000000	NaN	26.000000	NaN	NaN	NaN	3.900000	N

In [32]:

```
corr=cmsu.corr()
corr
```

Out[32]:

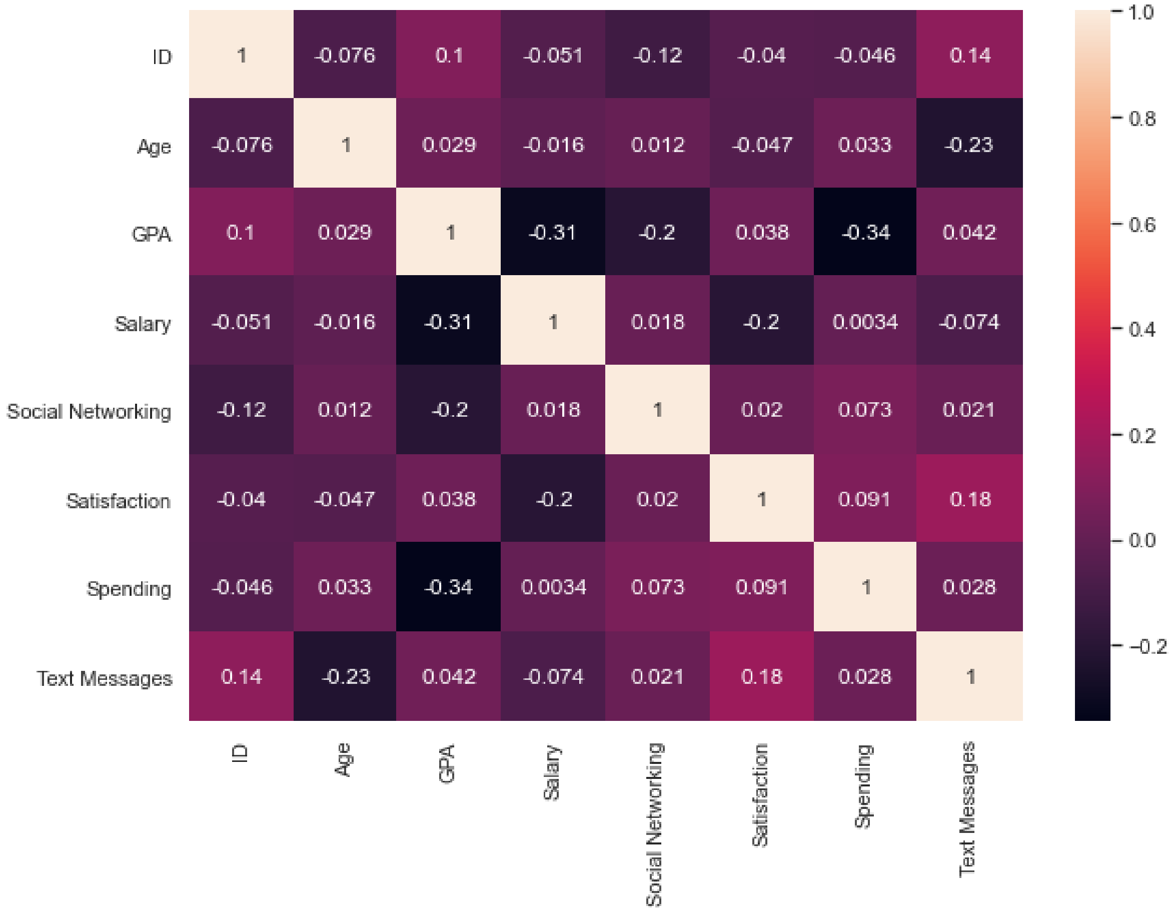
	ID	Age	GPA	Salary	Social Networking	Satisfaction	Spending	Te› Me
ID	1.000000	-0.075545	0.102328	-0.051484	-0.118383	-0.039676	-0.046230	0.
Age	-0.075545	1.000000	0.029370	-0.015536	0.011815	-0.046572	0.032968	-0.
GPA	0.102328	0.029370	1.000000	-0.308643	-0.197002	0.038097	-0.343403	0.
Salary	-0.051484	-0.015536	-0.308643	1.000000	0.017601	-0.197013	0.003402	-0.
Social Networking	-0.118383	0.011815	-0.197002	0.017601	1.000000	0.020125	0.073088	0.
Satisfaction	-0.039676	-0.046572	0.038097	-0.197013	0.020125	1.000000	0.090500	0.
Spending	-0.046230	0.032968	-0.343403	0.003402	0.073088	0.090500	1.000000	0.
Text Messages	0.138066	-0.227753	0.042195	-0.073640	0.020940	0.177548	0.028489	1.

In [33]:

```
plt.figure(figsize=(10,7))
sns.heatmap(corr,annot=True)
```

Out[33]:

```
<AxesSubplot:>
```

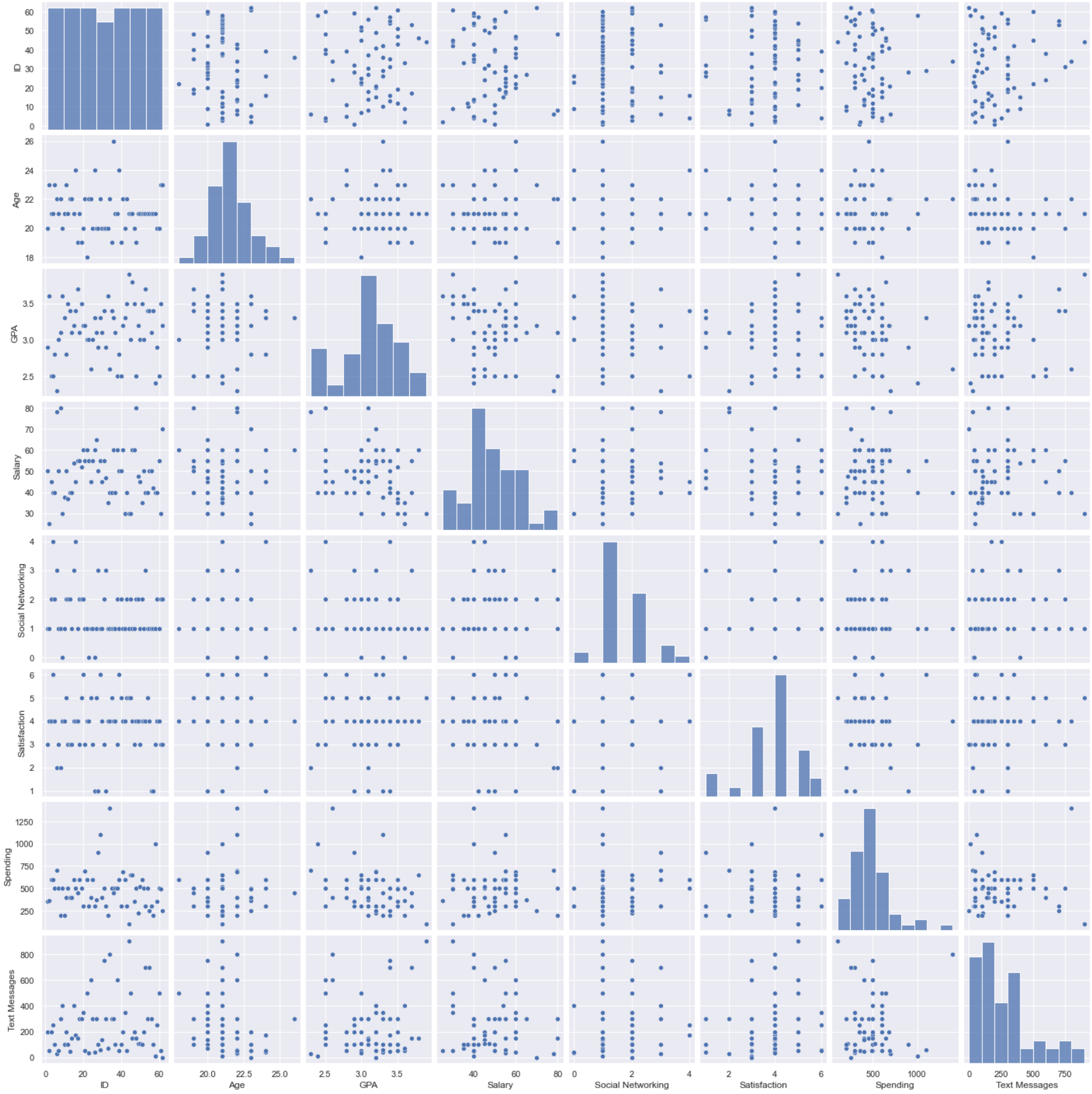


In [34]:

```
sns.pairplot(cmsu)
```

Out[34]:

```
<seaborn.axisgrid.PairGrid at 0x1ff4219bac0>
```



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

Q 2.1.1. Gender and Major

In [35]:

```
df_Major=pd.crosstab(cmsu["Gender"],cmsu["Major"],margins=True)
df_Major
```

Out[35]:

Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marke
Gender							
Female	3	3	7	4	4	3	
Male	4	1	4	2	6	4	
All	7	4	11	6	10	7	

Q 2.1.2. Gender and Grad Intention

In [36]:

```
pd.DataFrame(cmsu.groupby("Grad Intention").Gender.value_counts())
```

Out[36]:

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

Q 2.1.3. Gender and Employment

In [37]:

```
pd.DataFrame(cmsu.groupby("Gender").Employment.value_counts())
```

Out[37]:

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

Q 2.1.4. Gender and Computer

In [38]:

```
pd.DataFrame(cmsu.groupby("Gender").Computer.value_counts())
```

Out[38]:

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

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

In [39]:

```
pd.DataFrame(cmsu.Gender.value_counts())
```

Out[39]:

	Gender
Female	33
Male	29

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

$Solution$ -Probability of Male student randomly selected ,using formula= P(A)=m/n. I have two popution taking M is "Male" and B is Major. m=no of ways of accurrence of Male . n= no of total outcome .

In [40]:

```
m=29
n=62
print("probability that a randomly selected CMSU male student",m/n)
```

```
probability that a randomly selected CMSU male student 0.46774193548387094
```

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

$Solution$ -Probability of Male student randomly selected ,using formula=$P(A) = m/n$. I have two popution taking B is "Female" and C is Major. m=no of ways of accurrence of Female . n= no of total outcome .

In [41]:

```
m=33
n=62
print("probability that a randomly selected CMSU female student",m/n)
```

```
probability that a randomly selected CMSU female student 0.532258064516129
```

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

In [42]:

```
df_Major=pd.crosstab(cmsu["Gender"],cmsu["Major"],margins=True)
df_Major
```

Out[42]:

Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marke
Gender							
Female	3	3	7	4	4	3	
Male	4	1	4	2	6	4	
All	7	4	11	6	10	7	

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

$Solution$ - conditional Probability of Male different major ,using formula = $P(A|B) = P(A \text{ Int } B)/P(B)$. I have two popution taking A is "Major" and B is Male. P(A Int B)=intersection A and B . P(B)= m/n .

In [43]:

```
Int=6/29 #(gender -male, major of Management intersection(P(A Int B)=Int) )
P=29/62 # (m/n,male-26 & n-total outcome(P(B)=P))
print("conditional probability of male for Management ",Int/P)
```

```
conditional probability of male for Management  0.4423305588585018
```

In [44]:

```
Int=5/29 #(gender -male ,major of Retailing/Marketing intersection(P(A Int B)=Int) )
P=29/62 # (m/n,male-26 & n-total outcome(P(B)=P))
print("conditional probability of male forRetailing/Marketing ",Int/P)
```

```
conditional probability of male forRetailing/Marketing  0.36860879904875155
```

In [45]:

```
Int=4/29 #(gender -male ,major of Accounting intersection(P(A Int B)=Int) )
P=29/62 # (m/n,male-26 & n-total outcome(P(B)=P))
print("conditional probability of male for Accounting ",Int/P)
```

```
conditional probability of male for Accounting  0.2948870392390012
```

In [46]:

```
Int=4/29 #(gender -male, major of Economics/Finance intersection(P(A Int B)=Int) )
P=29/62 # (m/n,male-26 & n-total outcome(P(B)=P))
print("conditional probability of male for Economics/Finance ",Int/P)
```

```
conditional probability of male for Economics/Finance  0.2948870392390012
```

In [47]:

```
Int=4/29 #(gender -male, major of Other  intersection(P(A Int B)=Int) )
P=29/62 # (m/n,male-26 & n-total outcome(P(B)=P))
print("conditional probability of male for Other ",Int/P)
```

```
conditional probability of male for Other  0.2948870392390012
```

In [48]:

```
Int=3/29 #(gender -male, major of Undecided intersection(P(A Int B)=Int) )
P=29/62 # (m/n,male-26 & n-total outcome(P(B)=P))
print("conditional probability of male for Undecided ",Int/P)
```

```
conditional probability of male for Undecided  0.2211652794292509
```

In [49]:

```
Int=2/29 #(gender -male, major of International Business intersection(P(A Int B)=Int) )
P=29/62 # (m/n,male-26 & n-total outcome(P(B)=P))
print("conditional probability of male for International Business ",Int/P)
```

```
conditional probability of male for International Business  0.14744351961950
06
```

In [50]:

```
Int=1/29 #(gender -male, major of CPI Business intersection(P(A Int B)=Int) )
P=29/62 # (m/n,male-26 & n-total outcome(P(B)=P))
print("conditional probability of male for CPI ",Int/P)
```

```
conditional probability of male for CPI  0.0737217598097503
```

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

$Solution$ - conditional Probability of Female different major ,using formula = $P(A|B) = P(A \text{ Int } B)/P(B)$. I have two popution taking A is "Major" and B is Female. P(A Int B)=intersection A and B . P(B)= m/n .

In [51]:

```
Int=9/33 #(gender -female, major of Retailing/Marketing intersection(P(A Int B)=Int) )
P=29/62 # (m/n,male-26 & n-total outcome(P(B)=P))
print("conditional probability of female for Retailing/Marketing ",Int/P)
```

```
conditional probability of female for Retailing/Marketing  0.583072100313479
7
```

In [52]:

```
Int=7/33 #(gender -female, major of Economics/Finance intersection(P(A Int B)=Int) )
P=29/62 # (m/n,male-26 & n-total outcome(P(B)=P))
print("conditional probability of female for Economics/Finance ",Int/P)
```

```
conditional probability of female for Economics/Finance  0.4535005224660397
```

In [53]:

```
Int=4/33 #(gender -female, major of International Business intersection(P(A Int B)=Int) )
P=29/62 # (m/n,male-26 & n-total outcome(P(B)=P))
print("conditional probability of female for International Business ",Int/P)
```

```
conditional probability of female for International Business  0.259143155694
87983
```

In [54]:

```
Int=4/33 #(gender -female, major of Management intersection(P(A Int B)=Int) )
P=29/62 # (m/n,male-26 & n-total outcome(P(B)=P))
print("conditional probability of female for Management ",Int/P)
```

```
conditional probability of female for Management  0.25914315569487983
```

In [55]:

```
Int=3/33 #(gender -female, major of International Business intersection(P(A Int B)=Int) )
P=29/62 # (m/n,male-26 & n-total outcome(P(B)=P))
print("conditional probability of female for Accounting ",Int/P)
```

```
conditional probability of female for Accounting  0.1943573667711599
```

In [56]:

```
Int=3/33 #(gender -female, major of CIS intersection(P(A Int B)=Int) )
P=29/62 # (m/n,male-26 & n-total outcome(P(B)=P))
print("conditional probability of female for CIS ",Int/P)
```

```
conditional probability of female for CIS  0.1943573667711599
```

In [57]:

```
Int=3/33 #(gender -female, major of Other intersection(P(A Int B)=Int) )
P=29/62 # (m/n,male-26 & n-total outcome(P(B)=P))
print("conditional probability of female for Other ",Int/P)
```

```
conditional probability of female for Other  0.1943573667711599
```

Q2.4.1. Find the probability That a randomly chosen student is a male and intends to graduate.

In [58]:

```
pd.DataFrame(cmsu.groupby("Grad Intention").Gender.value_counts()).T
```

Out[58]:

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

Solution -Probability of Male student intends to graduate ,using formula= P(A)=m/n. I have two popution taking "M" is "Male" and "B" is Yes(Graduate Intention). m=no of ways of accurrence of Male . n= no of total outcome .

In [59]:

```
m=17#(male student intends graduate is Based on the data )
n=62#(total student)
print(" Probability intends to graduate student is a male", m/n)
```

```
 Probability intends to graduate student is a male 0.27419354838709675
```

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

In [60]:

```
pd.DataFrame(cmsu.groupby("Gender").Computer.value_counts()).T
```

Out[60]:

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

Solution -Probability of Female student does NOT have a laptop,using formula= P(A)=m/n. I have two popution taking "F" is "Female" and "B" is NOT have a laptop (have=Desktop+Tablet). m=no of ways of accurrence of Female(have=Desktop+Tablet) . n= no of total outcome .

In [61]:

```
m=4#((female the do't have loptop , have =Desktop+Tablet))
n=62#(total Computer type )
print(" probability female student does NOT have a laptop ", m/n)
```

```
 probability female student does NOT have a laptop  0.06451612903225806
```

Q 2.5.1. Find the probability that a randomly chosen student is either a male or has full-time employment?

In [62]:

```
pd.DataFrame(cmsu.groupby("Gender").Employment.value_counts())
```

Out[62]:

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

$Solution$ -finding probability of all condition fast either a male , secound codition is male & female and thard codition is only male & full time. using formula $P(A \text{ UNI}B) = P(A) + P(B) - P(A \text{ INT } B)$

In [63]:

```
#P(A UNI B)=?
A=29/62 #P(A)=(All empl are Male)
B=10/62 #P(B)(full time male & female)
C=7/62 #P(A Int B)(only Male full time)
print("probability either a male or has full-time employment ",A+B-C)
```

```
probability either a male or has full-time employment  0.5161290322580645
```

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

In [64]:

```
df_Major=pd.crosstab(cmsu["Gender"],cmsu["Major"],margins=True)
df_Major
```

Out[64]:

Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marke
Gender							
Female	3	3	7	4	4	3	
Male	4	1	4	2	6	4	
All	7	4	11	6	10	7	

$solution$ - Two Event multually exclusive or assuming marginal probability .

In [65]:

```
probability_ib_m=4/33+4/33
print("Probability of female for international business or management",probability_ib_m)
```

```
Probability of female for international business or management 0.24242424242
424243
```

Q2.6. 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?

Solution - *contingency* table

```
                     " yes       no          Total"
          "Male"      "17"          "3"         " 20"
          "Female"   "11"           "9"          "20"
          "Total     "28"           "12"         "40"
```

solution-Probability of Female studend ratio of unconditinal and conditional probability are having larger difference , and ucoditional probility is 28% larger than conditional probility.so we can say that are not independent and these 2 event are dependent.

In [66]:

```
P_female=20/40#(unconditional probability )
print("female total graduate intention ",P_female)
P_female_2=11/28#(conditional prabability)
print("female only yes graduate intention ",P_female_2)
```

```
female total graduate intention  0.5
female only yes graduate intention  0.39285714285714285
```

In [67]:

```
ratio=0.5/0.39
print("ratio of both event",ratio)
```

```
ratio of both event 1.282051282051282
```

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

In [68]:

```
pd.DataFrame(cmsu.groupby("GPA").Gender.value_counts()).head(11)
```

Out[68]:

		Gender
GPA	**Gender**	
2.3	**Female**	1
2.4	**Female**	1
2.5	**Male**	4
	Female	2
2.6	**Male**	2
2.8	**Male**	2
	Female	1
2.9	**Female**	3
	Male	1
3.0	**Female**	5
	Male	2

$Solution$ -Probability of GPA<3.0 student in graduae ,using formula=$P(A) = m/n$. I have two popution taking A is GPA<3.0 and B is total student. m=no of ways of accurrence of GPA<3.0 student . n= no of total outcome student.

In [69]:

```
m=17 #( GPA<3.0 student)
n=62 #(total no outcome)
print("probability of GPA<3.0 student less GPA",m/n)
```

```
probability of GPA<3.0 student less GPA 0.27419354838709675
```

Q 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.

In [70]:

```
pd.DataFrame(cmsu.groupby("Salary").Gender.value_counts()).tail(13)
```

Out[70]:

		Gender
Salary	**Gender**	
50.0	**Female**	5
	Male	4
52.0	**Male**	1
54.0	**Male**	1
55.0	**Female**	5
	Male	3
60.0	**Female**	5
	Male	3
65.0	**Male**	1
70.0	**Female**	1
78.0	**Female**	1
80.0	**Female**	1
	Male	1

In [71]:

```
df_salary=pd.crosstab(cmsu["Gender"],cmsu["Salary"],margins=True)
df_salary
```

Out[71]:

Salary	**25.0**	**30.0**	**35.0**	**37.0**	**37.5**	**40.0**	**42.0**	**45.0**	**47.0**	**47.5**	**50.0**	**52.0**	**54.0**	**55.0**	**60.0**
Gender															
Female	0	5	1	0	1	5	1	1	0	1	5	0	0	5	5
Male	1	0	1	1	0	7	0	4	1	0	4	1	1	3	3
All	1	5	2	1	1	12	1	5	1	1	9	1	1	8	8

$Solution$ -Probability of salary>=50 ,using formula=$P(A) = m/n$. I have two popution taking A is salary>=50 and B is only Female & Male.

In [72]:

```
p_50_more_female=18/33
p_50_more_female
```

Out[72]:

```
0.5454545454545454
```

In [73]:

```
P_50_more_male=14/29
P_50_more_male
```

Out[73]:

```
0.4827586206896552
```

There are 18 female are earn solary = >50 & more out of 33 female ,and prabability of 54.54% famale papution are take salary=>50
There are 14 males who earns 50 or more out of 29 males. So required probability that a randomly selected male earns 50 or more is 48.2%

Q 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.

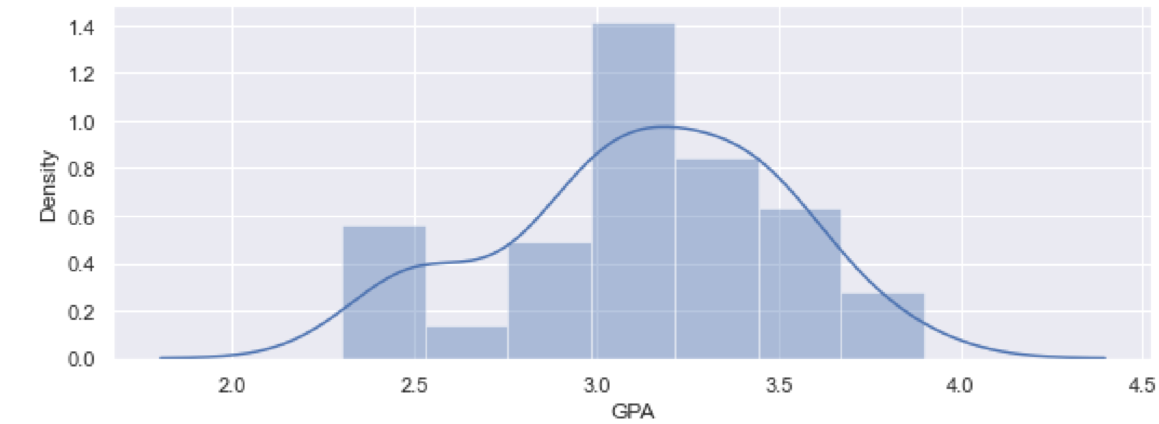
In [74]:

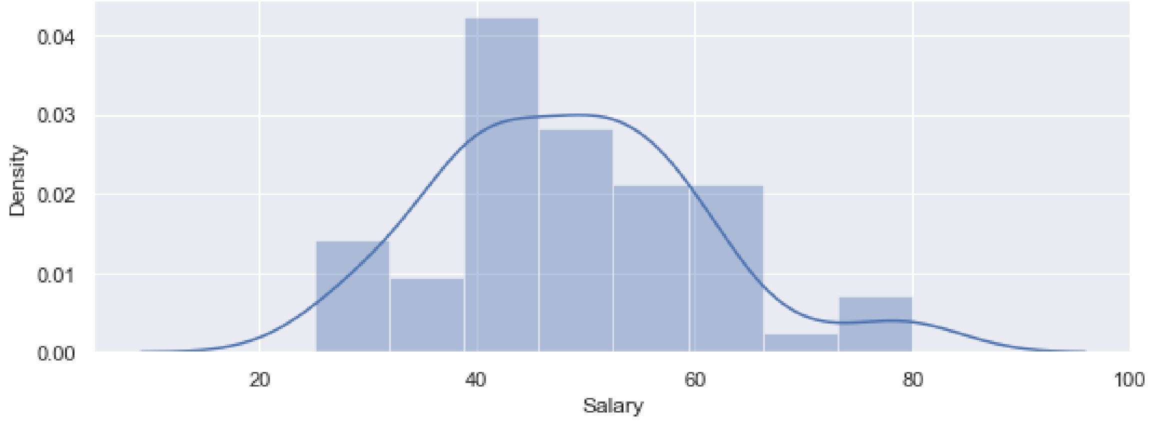
```
plt.figure(figsize=(10,16))
plt.subplot(4,1,1)
sns.distplot(cmsu.GPA)
plt.subplot(4,1,2)
sns.distplot(cmsu.Salary)
plt.subplot(4,1,3)
sns.distplot(cmsu.Spending)
plt.subplot(4,1,4)
sns.distplot(cmsu["Text Messages"])
```

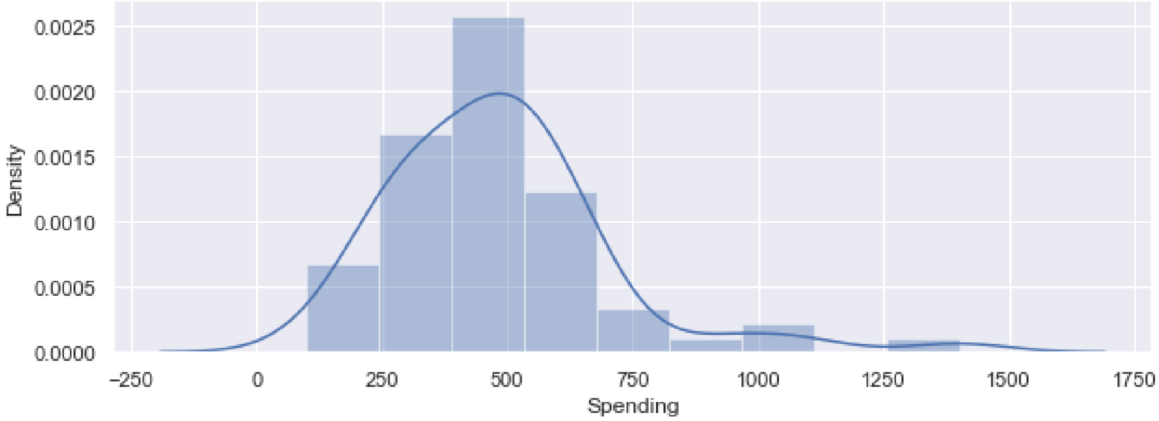
```
C:\Users\rahul\anaconda3\lib\site-packages\seaborn\distributions.py:2551: Fu
tureWarning: `distplot` is a deprecated function and will be removed in a fu
ture version. Please adapt your code to use either `displot` (a figure-level
function with similar flexibility) or `histplot` (an axes-level function for
histograms).
  warnings.warn(msg, FutureWarning)
C:\Users\rahul\anaconda3\lib\site-packages\seaborn\distributions.py:2551: Fu
tureWarning: `distplot` is a deprecated function and will be removed in a fu
ture version. Please adapt your code to use either `displot` (a figure-level
function with similar flexibility) or `histplot` (an axes-level function for
histograms).
  warnings.warn(msg, FutureWarning)
C:\Users\rahul\anaconda3\lib\site-packages\seaborn\distributions.py:2551: Fu
tureWarning: `distplot` is a deprecated function and will be removed in a fu
ture version. Please adapt your code to use either `displot` (a figure-level
function with similar flexibility) or `histplot` (an axes-level function for
histograms).
  warnings.warn(msg, FutureWarning)
C:\Users\rahul\anaconda3\lib\site-packages\seaborn\distributions.py:2551: Fu
tureWarning: `distplot` is a deprecated function and will be removed in a fu
ture version. Please adapt your code to use either `displot` (a figure-level
function with similar flexibility) or `histplot` (an axes-level function for
histograms).
  warnings.warn(msg, FutureWarning)
```

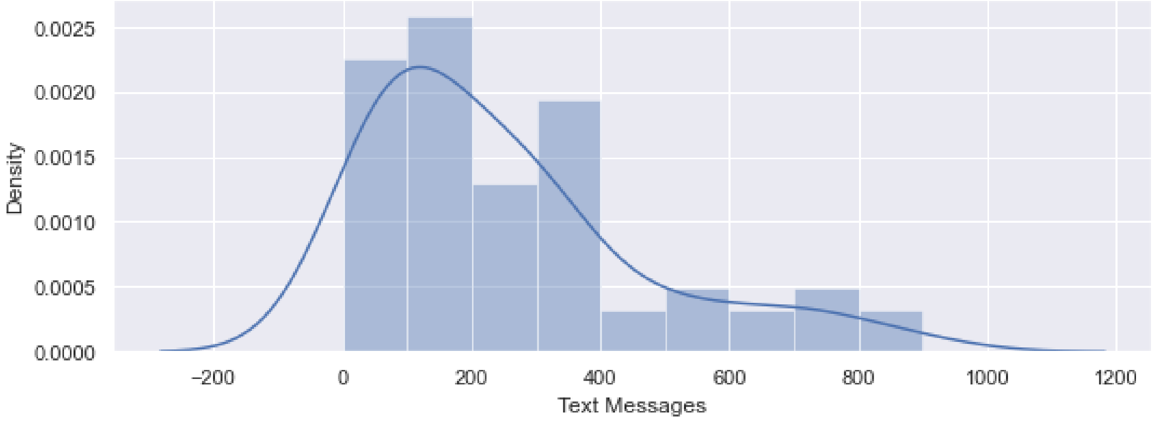
Out[74]:

```
<AxesSubplot:xlabel='Text Messages', ylabel='Density'>
```









$solution$ 1.$report$ GPA as per report of graph 3.0 to 3.5 GPA are more student gating result . 2.$report$ salary as per report of graph 40 to 60 salary more employer take . 3.$report$ Spending as per report of graph 250 to 600 hay value . 2.$report$ Text Messages$ as per report of graph 0 to 300 massage recived .

$$Problem - 3$$

manufacturers of ABC asphalt shingles Data Analysis

In [75]:

```
abc=pd.read_csv("A+&+B+shingles (1).csv")
```

In [76]:

```
abc.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36 entries, 0 to 35
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   A       36 non-null     float64
 1   B       31 non-null     float64
dtypes: float64(2)
memory usage: 704.0 bytes
```

In [77]:

```
abc.head()
```

Out[77]:

	A	B
0	0.44	0.14
1	0.61	0.15
2	0.47	0.31
3	0.30	0.16
4	0.15	0.37

In [78]:

```
abc.isnull().sum()
```

Out[78]:

```
A    0
B    5
dtype: int64
```

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.

In [79]:

```
abc.describe().T
```

Out[79]:

	count	mean	std	min	25%	50%	75%	max
A	36.0	0.316667	0.135731	0.13	0.2075	0.29	0.3925	0.72
B	31.0	0.273548	0.137296	0.10	0.1600	0.23	0.4000	0.58

As per calculated for other A shingle and B singles are 0.3166 and 0.27354 respectively. which are within permissible limit with is 0.35pound per 100 square feet.

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?

'''Forming hypothesis to perform the hypothesis test

```
HO: Poputation mean for shingles A and shingles B are not equal. (null  hypothesi
s) .
Hi: Population mean for shingle A and shingles B are equal .(alternate  hypothesi
s).
```

''' Alpha=0.05(leveal of significance)

Ttest

In [80]:

```
t_statistic ,p_value=ttest_ind(abc["A"],abc["B"],nan_policy="omit")
print("ttest value",t_statistic)
print("p_value",p_value)
```

```
ttest value 1.2896282719661123
p_value 0.2017496571835306
```

As per result fail to reject null hypothesis p_value>Alpha(0.5) .we show that A and B are not equail.

In [ ]:

```

```