Business Report

Problem1 Statement :

A wholesale distributor operating in different regions of Portugal has information on annual spending of several items in their stores across different regions and channels. The data (Wholesale Customer.csv) consists of 440 large retailers' annual spending on 6 different varieties of products in 3 different regions (Lisbon, Oporto, Other) and across different sales channel (Hotel/Restaurant/Café HoReCa, Retail).

Exploratory Data Analysis:

```
df.head()
```

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
0	1	Retail	Other	12669	9656	7561	214	2674	1338
1	2	Retail	Other	7057	9810	9568	1762	3293	1776
2	3	Retail	Other	6353	8808	7684	2405	3516	7844
3	4	Hotel	Other	13265	1196	4221	6404	507	1788
4	5	Retail	Other	22615	5410	7198	3915	1777	5185

Dataset has 9 variables Buyer/Spender , Channel , Region , Fresh , Milk , Grocery , Frozen , Detergents Paper and Delicatessen

Channel and Region both are categorical columns , while rest are integer columns.

Descriptive Statistics for the dataset:

Out[48]:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Buyer/Spender	440	NaN	NaN	NaN	220.5	127.161	1	110.75	220.5	330.25	440
Channel	440	2	Hotel	298	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Region	440	3	Other	316	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Fresh	440	NaN	NaN	NaN	12000.3	12647.3	3	3127.75	8504	16933.8	112151
Milk	440	NaN	NaN	NaN	5796.27	7380.38	55	1533	3627	7190.25	73498
Grocery	440	NaN	NaN	NaN	7951.28	9503.16	3	2153	4755.5	10655.8	92780
Frozen	440	NaN	NaN	NaN	3071.93	4854.67	25	742.25	1526	3554.25	60869
Detergents_Paper	440	NaN	NaN	NaN	2881.49	4767.85	3	256.75	816.5	3922	40827
Delicatessen	440	NaN	NaN	NaN	1524.87	2820.11	3	408.25	965.5	1820.25	47943

There are 3 unique value is in Region and 2 unique values in Channel.

Business Report

Check for Null values

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

From the above results, it is evident that there is no null values present in the dataset.

1.1. Use methods of descriptive statistics to summarize data.

A) Which Region and which Channel seems to spend more?

B) Which Region and which Channel seems to spend less?

Answer - By grouping by on the variables Channel and taking the sum of Buyer/Spender , we can identify channel which have spent more.

	Buyer/Spender
Channel	
Hotel	71034
Retail	25986

From the above result , it is evident Hotel in channel seems to spend more. Retail in channel seems to spend less.

Answer - By grouping by on the variables Region and taking the sum of Buyer/Spender , we can identify Region which have spent more and less

	Buyer/Spender
Region	
Lisbon	18095
Oporto	14899
Other	64026

From the above result , it is evident Other in 'Region' seems to spend more. Oporto in 'Region' seems to spend less.

Business Report

1.2. There are 6 different varieties of items are considered.

Do all varieties show similar behavior across Region and Channel?

Answer - First we will group by with Channel and Region and take count for all the other 6 variables.

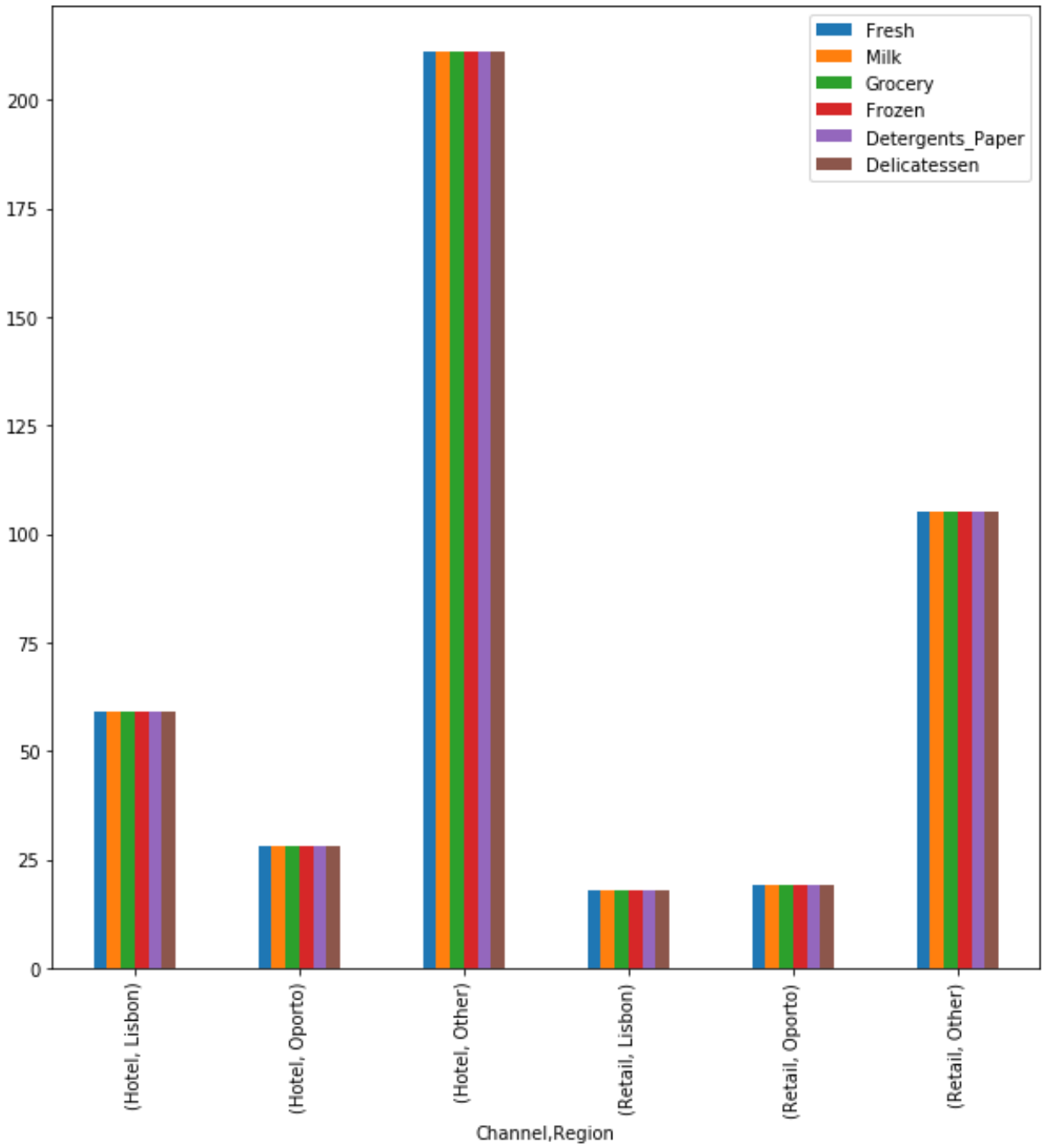
So , by doing so , we get the following data -

		Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Channel	**Region**						
Hotel	**Lisbon**	59	59	59	59	59	59
	Oporto	28	28	28	28	28	28
	Other	211	211	211	211	211	211
Retail	**Lisbon**	18	18	18	18	18	18
	Oporto	19	19	19	19	19	19
	Other	105	105	105	105	105	105

Now , plotting all the variables for Region and Channel on bar plot. From the bar plot it is evident that all varieties do not show similar behavior across Region and Channel.

Bar plot given below -

Business Report



Business Report

1.3. On the basis of the descriptive measure of variability, which item shows the most inconsistent behavior? Which items shows the least inconsistent behavior?

Answer- To measure the variability , we need to calculate the coefficient of variance by the formula -

Std/Mean *100

After calculating the coefficient of variance for all the 6 variables -

Fresh - 105.39179237473148
Milk - 127.32985840065413
Grocery - 119.51743730016824
Frozen - 158.03323836352914
Detergents_Paper - 165.46471385005154
Delicatessen - 184.94068981158384

From the above results , it is evident that Delicatessen shows the most inconsistent behavior. Fresh shows the least inconsistent behavior.

1.4. Are there any outliers in the data?

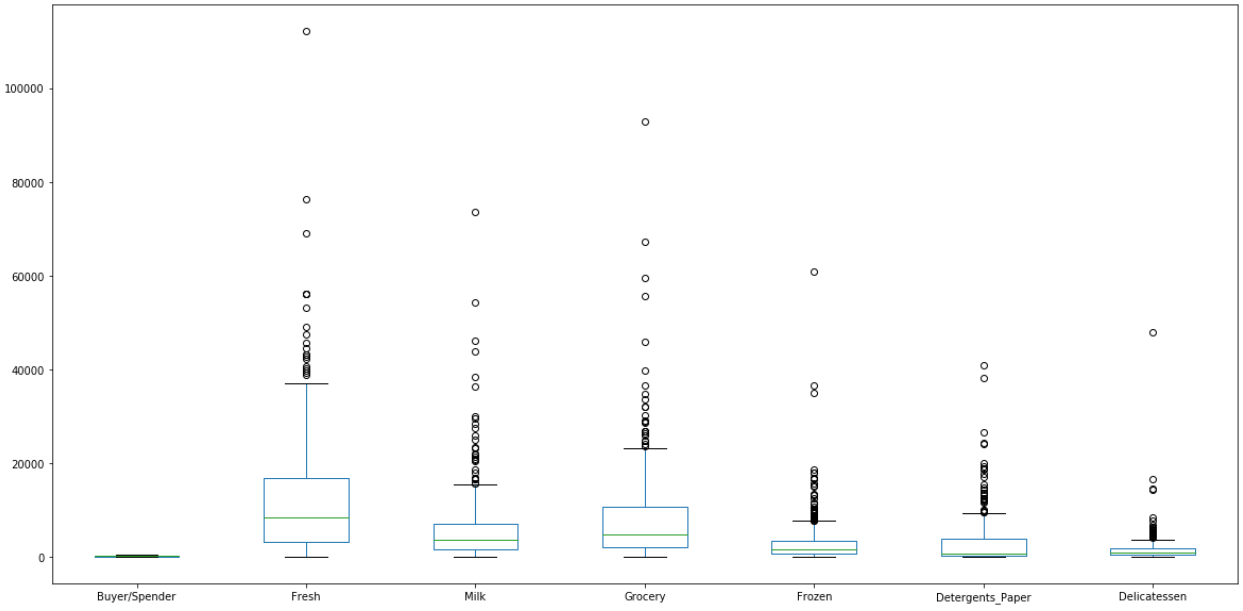
Answer - For this first we will calculate 1st and 2nd Quantile , and then Inter Quantile range(IQR). After calculating we have the following data for the variables -

```
Buyer/Spender        0
Channel              0
Delicatessen        27
Detergents_Paper    30
Fresh               20
Frozen              43
Grocery             24
Milk                28
Region               0.
```

Alternatively , we can see outliers by plotting the box plot.(Shown below)

By viewing the data and box plot we have come to conclusion that the data has outliers.

Business Report



Business Report

1.5. On the basis of this report, what are the recommendations?

From the above data analysis we can recommend-

1. All the 6 items show inconsistent behavior , so we can say where the items are less in Channel and region we can try to increase them.

2. We observed that some columns values have extreme values (outliers) , so we recommend to try to evenly distribute those items across other channel and region.

3. The item Fresh is having the least inconsistent behavior. So we can invest in this item more.

4. The Channel - Hotel and the Region - Other have more Buyers/Spenders, so we can invest more in this channel and region.

Business Report

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:

Dataset has 14 variables - ID, Gender , Age , Class , Major , Grad Intention , GPA , Employment , Salary , Employment , Social Networking , Satisfaction , Spending , Computer , Text Messages.

	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

Business Report

Descriptive Statistics for the dataset:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
ID	62	NaN	NaN	NaN	31.5	18.0416	1	16.25	31.5	46.75	62
Gender	62	2	Female	33	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Age	62	NaN	NaN	NaN	21.129	1.43131	18	20	21	22	26
Class	62	3	Senior	31	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Major	62	8	Retailing/Marketing	14	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Grad Intention	62	3	Yes	28	NaN	NaN	NaN	NaN	NaN	NaN	NaN
GPA	62	NaN	NaN	NaN	3.12903	0.377388	2.3	2.9	3.15	3.4	3.9
Employment	62	3	Part-Time	43	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Salary	62	NaN	NaN	NaN	48.5484	12.0809	25	40	50	55	80
Social Networking	62	NaN	NaN	NaN	1.51613	0.844305	0	1	1	2	4
Satisfaction	62	NaN	NaN	NaN	3.74194	1.21379	1	3	4	4	6
Spending	62	NaN	NaN	NaN	482.016	221.954	100	312.5	500	600	1400
Computer	62	3	Laptop	55	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Text Messages	62	NaN	NaN	NaN	246.21	214.466	0	100	200	300	900

Unique values present in variables Gender , class , Major , Grad Intention , Employment , Computer.

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

2.1.1. Gender and Major

Answer - Using Pandas crosstab function we can achieve this -

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

2.1.2. Gender and Grad Intention

Answer-

Grad Intention	No	Undecided	Yes
Gender			
Female	9	13	11

Business Report

Grad Intention	No	Undecided	Yes
Gender			
Male	3	9	17

2.1.3. Gender and Employment

Answer -

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

Answer -

Lets first group by Gender to find the count of Male and female -

	Gender	count
0	Female	33
1	Male	29

Total Count - 62

Business Report

Probability that a randomly selected CMSU student will be male is 29/62 = 0.4677 i.e 46.77%

Probability that a randomly selected CMSU student will be Female is 33/62 = 0.5322 i.e 53.22%

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

Answer- Lets first group by on gender and major and take count to find the actual number of counts of major among male and female.

Below the data shows the same -

	Major	Gender	count
0	Accounting	Female	3
1	Accounting	Male	4
2	CIS	Female	3
3	CIS	Male	1
4	Economics/Finance	Female	7
5	Economics/Finance	Male	4
6	International Business	Female	4
7	International Business	Male	2
8	Management	Female	4
9	Management	Male	6
10	Other	Female	3
11	Other	Male	4
12	Retailing/Marketing	Female	9
13	Retailing/Marketing	Male	5
14	Undecided	Male	3

- We know that there are total 29 Males . So to calculate conditional probability , we will take male count with respect to each Major and divide by total number of males.

For example -
Accounting = 4
Total_Male = 29
Prob_Accounting_given_Male = Accounting/Total_Male = 0.13793103448275862

1. Following is the Conditional probability of different major among Male

Business Report

Major	Gender	Probability
Accounting	Male	0.137 i.e 13.79%
CIS	Male	0.03448 i.e 3.44%
Economics/Finance	Male	0.137 i.e 13.79%
International Business	Male	0.068 i.e 6.89%
Management	Male	0.206 i.e 20.68 %
Other	Male	0.1379 i.e 13.79%
Retailing/Marketing	Male	0.172 i.e 17.24%
Undecided	Male	0.103 i.e 10.34 %

- Similarly , we know that the total number of counts for female is 33 . So to calculate conditional probability , we will take Female count with respect to each Major and divide by total number of Females.

For example -
Accounting = 4
Total_Female = 33
Prob_Accounting_given_FeMale = Accounting/Total_Female = 0.1212

2. Following is the Conditional probability of different major among Female

Major	Gender	Probability
Accounting	Female	0.0909 i.e 9.09%
CIS	Female	0.0909 i.e 9.09%
Economics/Finance	Female	0.2121 i.e 21.21%
International Business	Female	0.1212 i.e 12.12%
Management	Female	0.1212 i.e 12.12%
Other	Female	0.0909 i.e 9.09%
Retailing/Marketing	Female	0.2727 i.e 27.27%
other	Female	0.0 i.e 0%

Business Report

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

Answer- Lets first group by on gender and Grad Intention and take count to find the actual number of counts of Grad Intention among male and female.

Below the data shows the same -

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

Since we need to find the conditional probability for those who intent to graduate. So we will only take the count for yes.

We know the total number of Male are 29 and female are 33. 17 and 11 are the male and female which fall under the category 'Yes' respectively.

So conditional Probability for Male will be -

Total_Male = 29
Grad_intention_yes = 17
Prob_Grad_intention_given_male = Grad_intention_yes/Total_Male
Prob_Grad_intention_given_male = 0.5862 i.e 58.62%

So conditional Probability for Female will be -

Total_FeMale = 33
Grad_intention_Yes = 11
Prob_Grad_intention_given_FeMale = Grad_intention_Yes/Total_FeMale
Prob_Grad_intention_given_FeMale = 0.333333 = 33.33%.

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

Answer- Lets first group by on Gender and Employment and take count to find the actual number of counts of Employment status among male and female.

Below the data shows the same -

Business Report

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

We know the total number of Male are 29 and female are 33 . We will find the conditional probability by taking into account male and female counts in all the 3 categories of Employment.

1. Full time given Male

Total_Male = 29
Full_Time = 7
Prob_Full_Time_given_male = Full_Time/Total_Male
Prob_Full_Time_given_male = 0.2413 i.e 24.13%

2. Part time given Male

Total_Male = 29
Part_Time = 19
Prob_Part_Time_given_male = Part_Time/Total_Male
Prob_Part_Time_given_male = 0.65517 i.e 65.51%

3. Unemployed_given_male

Total_Male = 29
Unemployed = 3
Prob_Unemployed_given_male = Unemployed/Total_Male
Prob_Unemployed_given_male = 0.1034 i.e 10.34%

4. Full time given Female

Total_FeMale = 33
Full_Time = 3
Prob_Full_Time_given_FeMale = Full_Time/Total_FeMale
Prob_Full_Time_given_FeMale = 0.9009 i.e 9.09%

5. Part time given Female

Total_FeMale = 33
Part_Time = 24
Prob_Part_Time_given_FeMale = Part_Time/Total_FeMale
Prob_Part_Time_given_FeMale = 0.7272 i.e 72.72%

Business Report

6. Unemployed_given_Female

Total_FeMale = 33
Unemployed = 6
Prob_Unemployed_given_FeMale = Unemployed/Total_FeMale
Prob_Unemployed_given_FeMale = 0.1818 i.e 18.18%

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

Answer- Lets first group by on Gender and Laptop and take count to find the actual number of counts of Employment status among male and female.

```
Computer Gender
Desktop  Female      2
         Male        3
Laptop   Female     29
         Male       26
Tablet   Female      2>
```

We know the total number of Male are 29 and female are 33 . We will find the conditional probability by taking into account male and female counts in the Laptop category of Computer.

1. The conditional probability of laptop preference among the male students

Total_Male = 29
Laptop = 26
Prob_Lapotop_given_male = Laptop/Total_Male
Prob_Lapotop_given_male = 0.89655 i.e 89.65%

2. The conditional probability of laptop preference among the Female students

Total_FeMale = 33
Laptop = 29
Prob_Lapotop_given_FeMale = Laptop/Total_FeMale
Prob_Lapotop_given_FeMale = 0.8787 = 87.87%

Business Report

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.

Answer-

As we can see that column variable in each case is not independent of Gender as the events described above are not independent and have relationship with the column Gender.

Part II

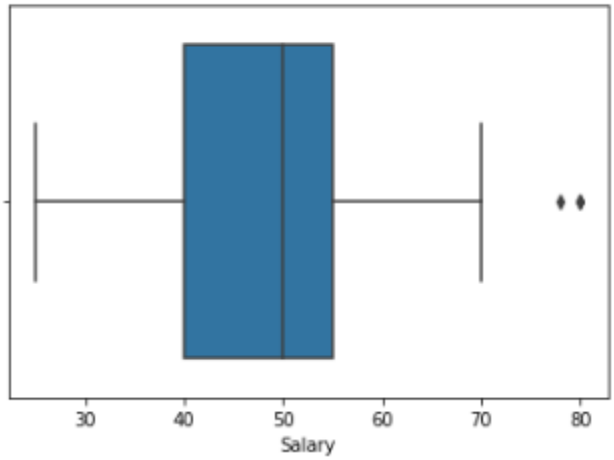
2.4. Note that there are three numerical (continuous) variables in the data set, Salary, Spending and Text Messages. For each of them comment whether they follow a normal distribution.
Write a note summarizing your conclusions.
[Recall that symmetric histogram does not necessarily mean that the underlying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

Answer - Lets check for the normal distribution for the column 'Salary' , 'Spending' , 'Text Messages'

1 . Salary -

First lets construct a box plot for Salary -

```
<matplotlib.axes._subplots.AxesSubplot at 0x1058a508>
```



From the above box plot it is evident that the column doesn't follow normal distribution as there is skewness in the box plot.

After doing Shapiro test also the p value comes to be 0.028 which is less than 0.05 , this confirms the column is not normally distributed.

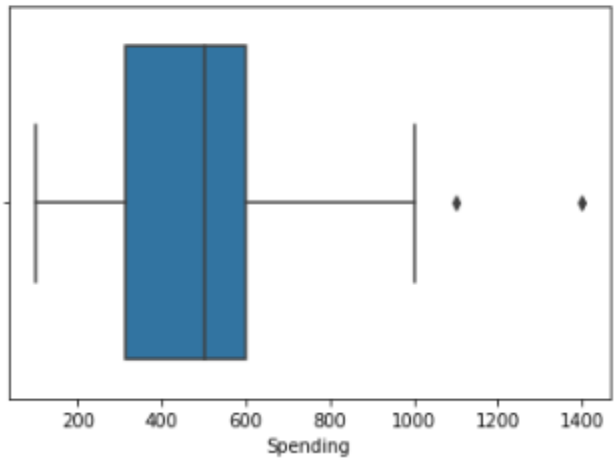
Business Report

```
(0.9565856456756592, 0.028000956401228905)
```

2. Spending-

First lets construct a box plot for Spending -

```
In [191]: sns.boxplot(x = Survey['Spending'])
Out[191]: <matplotlib.axes._subplots.AxesSubplot at 0x10cafd08>
```



From the above box plot it is evident that the column doesn't follow normal distribution as there is skewness in the box plot.

After doing Shapiro test also the p value comes to be very less than 0.05 , this confirms the column is not normally distributed.

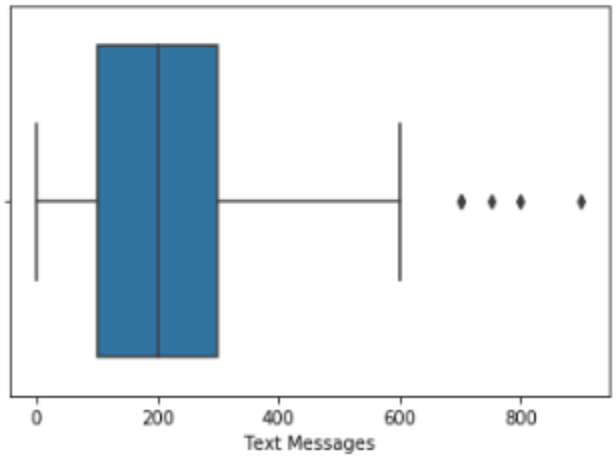
```
(0.8777452111244202, 1.6854661225806922e-05)
```

Business Report

3. Text Messages -

First lets construct a box plot for Text Messages -

```
In [125]: sns.boxplot(x = Survey['Text Messages'])
Out[125]: <matplotlib.axes._subplots.AxesSubplot at 0xbdc7a48>
```



From the above box plot it is shown that the plot doesn't have skewness. But the outliers are present. So need to do shapiro test to check for normal distribution.

After doing Shapiro test also the p value comes to be very less than 0.05 , this confirms the column is not normally distributed.

```
(0.8594191074371338, 4.324040673964191e-06)
```

Problem 3

An important quality characteristic used by the manufacturers of ABC asphalt shingles is the amount of moisture the shingles contain when they are packaged. Customers may feel that they have purchased a product lacking in quality if they find moisture and wet shingles inside the packaging. In some cases, excessive moisture can cause the granules attached to the shingles for texture and coloring purposes to fall off the shingles resulting in appearance problems. To monitor the amount of moisture present, the company conducts moisture tests. A shingle is weighed and then dried. The shingle is then reweighed, and based on the amount of moisture taken out of the product, the pounds of moisture per 100 square feet is calculated. The company would like to show that the mean moisture content is less than 0.35 pound per 100 square feet.

Business Report

The file (A & B shingles.csv) includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.

Exploratory Data Analysis:

-Checking the data head

```
In [36]: df6.head()
```

Out[36]:

	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

Checking the count,mean,std.

```
In [38]: df6.describe(include='all')
```

Out[38]:

	A	B
count	36.000000	31.000000
mean	0.316667	0.273548
std	0.135731	0.137296
min	0.130000	0.100000
25%	0.207500	0.160000
50%	0.290000	0.230000
75%	0.392500	0.400000
max	0.720000	0.580000

Checking Nulls. We see that there are 5 nulls in the column 'B'.

```
In [43]: df6.isnull().sum()
Out[43]: A    0
         B    5
         dtype: int64
```

Business Report

3.1. For the A shingles, form the null and alternative hypothesis to test whether the population mean moisture content is less than 0.35 pound per 100 square feet.

Answer -

#Ho = Mean moisture content is equal or greater than 0.35 pound per 100 square feet.
#Ha = Mean moisture content is less than 0.35 pound per 100 square feet.

Null and Alternate Hypothesis for Shingle A -

Null Hypothesis: Ho: u >= 0.
Alternate Hypothesis: Ha: u < 0.35

3.2. For the B shingles, form the null and alternative hypothesis to test whether the population mean moisture content is less than 0.35 pound per 100 square feet.

#Ho = Mean moisture content is equal or greater than 0.35 pound per 100 square feet.
#Ha = Mean moisture content is less than 0.35 pound per 100 square feet.

Null and Alternate Hypothesis for Shingle B -

Null Hypothesis: Ho : u >= 0.35
Alternate Hypothesis: Ha : u < 0.35

3.3. Do you think that the population means 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?

Answer -

Null and Alternate Hypothesis for the population means for shingles A and B are equal
Null Hypothesis: Ho: u1 = u2 (# population mean for Shingle A 'u1' is equal to population mean 'u2' for Shingle B)

Alternate Hypothesis: Ha: u1 != u2 (# population mean for Shingle A 'u1' is not equal to population mean 'u2' for Shingle B)

First we will assume that the two groups have same variance before the test for equality of means is performed.

Here we should note that we have dropped Null values as the null values were present in 'B'.

Business Report

After conducting the two t-test as the standard deviation is not known , we come to know that the p-value is greater than alpha (0.05). Therefore we failed to reject Null Hypothesis(Ho).That is population mean for Shingle A 'u1' is equal to population mean 'u2' for Shingle B

```
Two sample t-test result - 0.985249977839441 0.3284577916404776
```

3.4. What assumption about the population distribution is needed in order to conduct the hypothesis tests above?

Answer-

We need to Assume that the population distribution should be 'Normally Distributed' in order to conduct the hypothesis tests above.