BUSINESS REPORT FOR SMDM PROJECT

Table of Contents:

Problem 1 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, Retail).

Exploratory Data Analysis:

	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

The Dataset has 9 variables. **Channel and Region are Categorical Variables**, the rest of them are of int64 datatype.

Descriptive Statistics for the dataset:

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440.000000	440	440	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
unique	NaN	2	3	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	Hotel	Other	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	298	316	NaN	NaN	NaN	NaN	NaN	NaN
mean	220.500000	NaN	NaN	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	127.161315	NaN	NaN	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	1.000000	NaN	NaN	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	110.750000	NaN	NaN	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000
50%	220.500000	NaN	NaN	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000
75%	330.250000	NaN	NaN	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000
max	440.000000	NaN	NaN	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000

We have two unique values in Channel column with "**Hotel**" being the top repetitive and three unique values from Region with the "Other" being the top repetitive.

A little surprising to see all the 6 columns (items – Fresh, Milk, Grocery, Frozen, Detergents_Paper and Delicatessen) have **higher** values in Standard Deviation than their Mean values which might indicate high variation between values, and not a perfect normal distribution for data.

Check for Null values

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

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

1.1. Use methods of descriptive statistics to summarize data.

Which Region and which Channel seems to spend more?

Region:

	Region	Milk	Fresh	Grocery	Frozen	Detergents_Paper	Delicatessen	Total_spent
0	Lisbon	422454	854833	570037	231026	204136	104327	2386813
1	Oporto	239144	464721	433274	190132	173311	54506	1555088
2	Other	1888759	3960577	2495251	930492	890410	512110	10677599

We can say from the above result that the Region called "**Other**" spends **more** Overall and the Region called "**Oporto**" spends **less** overall.

Channel:

	Channel	Milk	Fresh	Grocery	Frozen	Detergents_Paper	Delicatessen	Total_spent
0	Hotel	1028614	4015717	1180717	1116979	235587	421955	7999569
1	Retail	1521743	1264414	2317845	234671	1032270	248988	6619931

We can say from the above result that the Channel called "**Hotel**" spends **more** Overall and the Channel called "**Retail**" spends **less** overall.

When we consider from individual front (for all items) say Channel and Region Separately on average spending, we have(A little Extra Effort):

Spending More(By Average)

Region:

Region	Milk	Fresh	Grocery	Frozen	Detergents_Paper	Delicatessen
Lisbon	5486.415584	11101.727273	7403.077922	3000.337662	2651.116883	1354.896104
Oporto	5088.170213	9887.680851	9218.595745	4045.361702	3687.468085	1159.702128
Other	5977.085443	12533.471519	7896.363924	2944.594937	2817.753165	1620.601266

From the above result (Highlighted part) we can see that all the three Regions spends averagely more on "**Fresh**".

Channel:

Channel	Milk	Fresh	Grocery	Frozen	Detergents_Paper	Delicatessen
Hotel	3451.724832	13475.560403	3962.137584	3748.251678	790.560403	1415.956376
Retail	10716.500000	8904.323944	16322.852113	1652.612676	7269.507042	1753.436620

From the above result (Highlighted part) we can see that the Hotel spends averagely more in "**Fresh**" and Retail spends averagely more on "**Grocery**".

Spending Less(By Average)

Region:

Region	Milk	Fresh	Grocery	Frozen	Detergents_Paper	Delicatessen
Lisbon	5486.415584	11101.727273	7403.077922	3000.337662	2651.116883	1354.896104
Oporto	5088.170213	9887.680851	9218.595745	4045.361702	3687.468085	1159.702128
Other	5977.085443	12533.471519	7896.363924	2944.594937	2817.753165	1620.601266

From the above result (Highlighted part) we can see that all the three Regions spends averagely less on "**Delicatessen**".

Channel:

Channel	Milk	Fresh	Grocery	Frozen	Detergents_Paper	Delicatessen
Hotel	3451.724832	13475.560403	3962.137584	3748.251678	790.560403	1415.956376
Retail	10716.500000	8904.323944	16322.852113	1652.612676	7269.507042	1753.436620

From the above result (Highlighted part) we can see that the Hotel spends averagely less in "**Detergents_Paper**" and Retail spend less on "**Frozen**".

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

Do all varieties show similar behaviour across Region and Channel?

To Analyse the behaviour of all the items across Region and Channel, we are going to look upon skewness and how the information from describe function of pandas provides insights.

Skewness:

Channel Specific:

	Channel	Milk	Fresh	Grocery	Frozen	Detergents_Paper	Delicatessen
0	Hotel	4.660186	2.512084	2.118316	5.211448	2.857124	11.521808
1	Retail	3.413169	1.593948	2.980945	2.526896	2.612425	3.772841

From the above result we can see that all of items are right skewed(positive values) and the maximum skewness is seen in Delicatessen from the channel called "Hotel"(maximum deviation in normal distribution) and the minimum is seen in Fresh from a Channel called "Retail"(Still High Skewness based on rule of thumb).

Region Specific:

	Region	Milk	Fresh	Grocery	Frozen	Detergents_Paper	Delicatessen
0	Lisbon	1.923527	2.013077	2.023387	2.334571	2.359030	2.050233
1	Oporto	1.803677	0.979873	3.637678	5.492402	3.620133	2.152210
2	Other	4.250869	2.617896	3.839176	3.963391	3.705302	10.214896

From the above result we can see that all of items are right skewed(positive values) and the maximum skewness is seen in Delicatessen from the Region called "Other"(maximum deviation in normal distribution) and the minimum is seen in Fresh from a Region called "Oporto"(Falls under the acceptable region for normal distribution based on rule of thumb).

Interesting part is that most of the items have high skewness in the Other region than the other two.

It's time to look on individual Channels and Region for deeper analysis(Min, Max, Skewness analysis):

Channel : **Hotel**

	Milk	Fresh	Grocery	Frozen	Detergents_Paper	Delicatessen
count	298.000000	298.000000	298.000000	298.000000	298.000000	298.000000
mean	3451.724832	13475.560403	3962.137584	3748.251678	790.560403	1415.956376
std	4352.165571	13831.687502	3545.513391	5643.912500	1104.093673	3147.426922
min	55.000000	3.000000	3.000000	25.000000	3.000000	3.000000
25%	1164.500000	4070.250000	1703.750000	830.000000	183.250000	379.000000
50%	2157.000000	9581.500000	2684.000000	2057.500000	385.500000	821.000000
75%	4029.500000	18274.750000	5076.750000	4558.750000	899.500000	1548.000000
max	43950.000000	112151.000000	21042.000000	60869.000000	6907.000000	47943.000000
skew	4.660186	2.512084	2.118316	5.211448	2.857124	11.521808

Channel : **Retail**

	Milk	Fresh	Grocery	Frozen	Detergents_Paper	Delicatessen
count	142.000000	142.000000	142.000000	142.000000	142.000000	142.000000
mean	10716.500000	8904.323944	16322.852113	1652.612676	7269.507042	1753.436620
std	9679.631351	8987.714750	12267.318094	1812.803662	6291.089697	1953.797047
min	928.000000	18.000000	2743.000000	33.000000	332.000000	3.000000
25%	5938.000000	2347.750000	9245.250000	534.250000	3683.500000	566.750000
50%	7812.000000	5993.500000	12390.000000	1081.000000	5614.500000	1350.000000
75%	12162.750000	12229.750000	20183.500000	2146.750000	8662.500000	2156.000000
max	73498.000000	44466.000000	92780.000000	11559.000000	40827.000000	16523.000000
skew	3.413169	1.593948	2.980945	2.526896	2.612425	3.772841

Based on the analysis of Individual Channels we can see that the skewness still more in the Delicatessen, but the max spending is where we can start to see the difference. We can see that Hotel spends more in Fresh and Retail spends more in Grocery(which makes perfect sense when we think about it as Hotel tends to focus on Fresh items). Also, the spending seems to be more in the Grocery part in Retail and even the minimum spending is more in Grocery.

Region : **Lisbon**

	Milk	Fresh	Grocery	Frozen	Detergents_Paper	Delicatessen
count	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000
mean	5486.415584	11101.727273	7403.077922	3000.337662	2651.116883	1354.896104
std	5704.856079	11557.438575	8496.287728	3092.143894	4208.462708	1345.423340
min	258.000000	18.000000	489.000000	61.000000	5.000000	7.000000
25%	1372.000000	2806.000000	2046.000000	950.000000	284.000000	548.000000
50%	3748.000000	7363.000000	3838.000000	1801.000000	737.000000	806.000000
75%	7503.000000	15218.000000	9490.000000	4324.000000	3593.000000	1775.000000
max	28326.000000	56083.000000	39694.000000	18711.000000	19410.000000	6854.000000
skew	1.923527	2.013077	2.023387	2.334571	2.359030	2.050233

Region : **Oporto**

	Milk	Fresh	Grocery	Frozen	Detergents_Paper	Delicatessen
count	47.000000	47.000000	47.000000	47.000000	47.000000	47.000000
mean	5088.170213	9887.680851	9218.595745	4045.361702	3687.468085	1159.702128
std	5826.343145	8387.899211	10842.745314	9151.784954	6514.717668	1050.739841
min	333.000000	3.000000	1330.000000	131.000000	15.000000	51.000000
25%	1430.500000	2751.500000	2792.500000	811.500000	282.500000	540.500000
50%	2374.000000	8090.000000	6114.000000	1455.000000	811.000000	898.000000
75%	5772.500000	14925.500000	11758.500000	3272.000000	4324.500000	1538.500000
max	25071.000000	32717.000000	67298.000000	60869.000000	38102.000000	5609.000000
skew	1.803677	0.979873	3.637678	5.492402	3.620133	2.152210

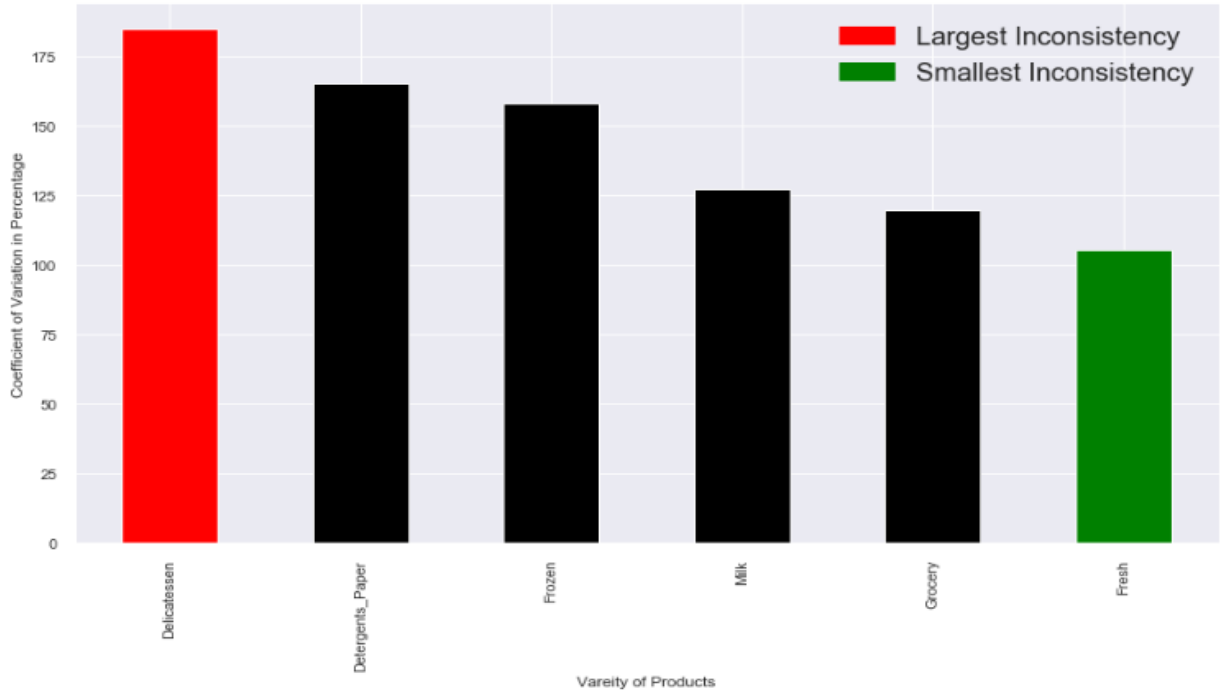
Region : **Other**

	Milk	Fresh	Grocery	Frozen	Detergents_Paper	Delicatessen
count	316.000000	316.000000	316.000000	316.000000	316.000000	316.000000
mean	5977.085443	12533.471519	7896.363924	2944.594937	2817.753165	1620.601266
std	7935.463443	13389.213115	9537.287778	4260.126243	4593.051613	3232.581660
min	55.000000	3.000000	3.000000	25.000000	3.000000	3.000000
25%	1634.000000	3350.750000	2141.500000	664.750000	251.250000	402.000000
50%	3684.500000	8752.500000	4732.000000	1498.000000	856.000000	994.000000
75%	7198.750000	17406.500000	10559.750000	3354.750000	3875.750000	1832.750000
max	73498.000000	112151.000000	92780.000000	36534.000000	40827.000000	47943.000000
skew	4.250869	2.617896	3.839176	3.963391	3.705302	10.214896

Based on the analysis of Individual Regions we can see that the skewness presents maximum in the Delicatessen in the Other region which cause the overall increase in trend when kept Region as a whole. Both Lisbon and Other shows relative max spending in the Fresh items, whereas Oporto looks to spend man in the Grocery part which makes me believe that it trends towards the Retail Mindset. The mean spending is definitely more in the Fresh items for all the Products. Oporto shows a significant change in behaviour when it comes to the Frozen items as comparing the other two regions, it spends more which is also reflected in the skewness of it.

1.3. On the basis of the descriptive measure of variability,

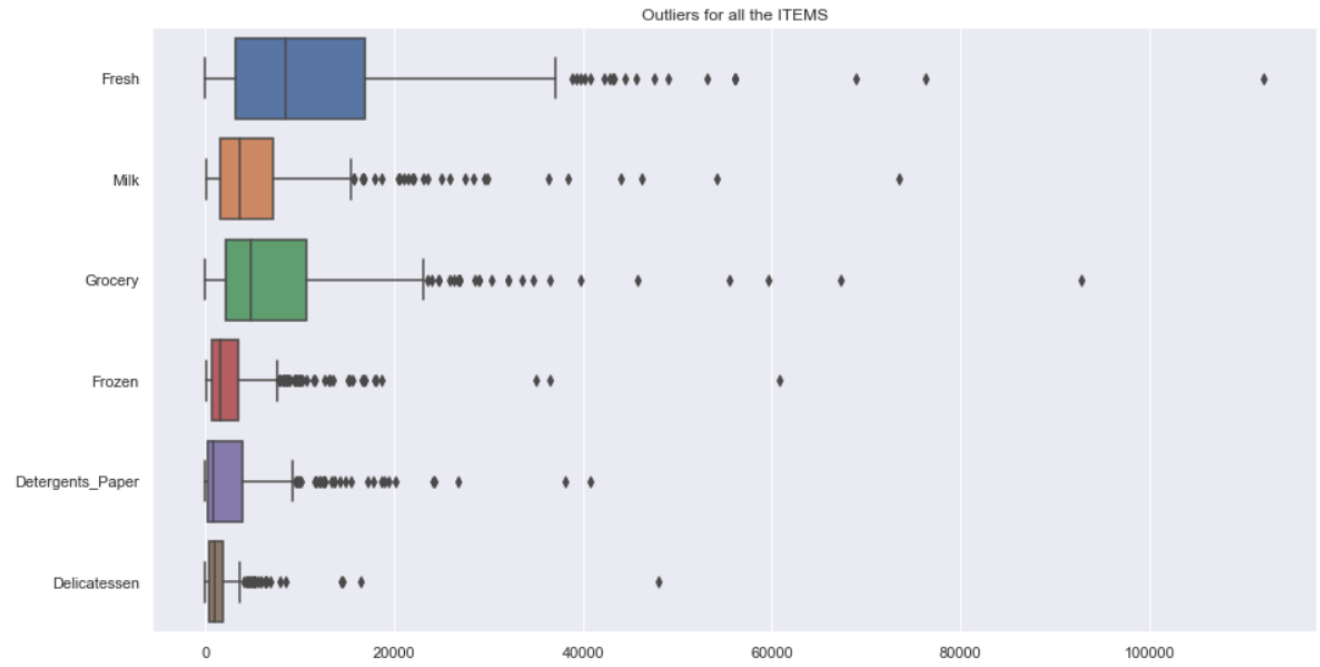
Which item shows the most inconsistent behaviour and Which items shows the least inconsistent behaviour?



Coefficient of Variation can provide more information on Inconsistency in the data as it normalizes the standard deviation to the mean. From the above result we can see that the item "**Delicatessen**" is the least consistent and "**Fresh**" is the most consistent(considering the data).

Also, on side note, since the percentages are more than 100, we can say that the data is not a generally seen normal distribution.

1.4. Are there any outliers in the data?



Only the 6 Items can be considered from the data can be considered here to be taken into account for calculating outliers.

From the above result we can see that all the 6 items have outliers, but all of these outliers come under the “excessive” category. Also looking at them we can see that only a handful of these outliers are extravagant i.e., only handful of these outliers are too much outside the acceptable zone and therefore outlier treatment for these points is highly recommended(Example : Replacing the values above 60000 for Fresh and replacing them with the mean/median value).

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

Analysing the results of the previous questions and doing a Pandas Profiling (Which can be seen in the notebook attached), here are the highlights:

- Detergents_Paper is highly correlated with Grocery (ρ = 0.98) and Fresh is Highly correlated with Delicatessen (ρ = 0.96) (Not surprising there).
- There are many outliers in the data (Please read the answer for the question 1.4 to resolve them).
- Spending seems to be more in "Other" Region and "Channel" Hotel and therefore needs to be look at closely if that directly relates to profit or not.
- Higher values in Standard Deviation than the items Mean Values indicate high variation between values, and abnormal distribution for data which should be rectified immediately.
- The Region called "Other" spends more in the 6 items in both "Hotel and Retail".
- Since there is a huge spending on Other region by retailers, we must try to look into the sales(what it does right in the region as more spent means more sales) that this region brings to the table and based on those results we must do actions in the rest of the regions.
-

Problem 1: Summary

1.1) **Spending More**:

Region: Other

Channel: Hotel

Spending Less:

Region: Oporto

Channel: Retail

1.2) **Behaviour of all the items across Region and Channel:**

Please refer to the detailed explanation provided for 1.2 as there are some noticeable behavioural changes when looked upon each Regions and Channels separately.

1.3) "**Fresh**" is the least consistent and "**Delicatessen**" is the most consistent.

1.4) All the 6 items(Fresh, Milk, Grocery, Frozen, Detergents_Paper, Delicatessen) **have outliers**.

1.5) Analysis and Recommendations:

- Detergents_Paper is highly correlated with Grocery (ρ = 0.98) and Fresh is Highly correlated with Delicatessen (ρ = 0.96) (Not surprising there).
- There are many outliers in the data (Please read the answer for the question 1.4 to resolve them).
- Higher values in Standard Deviation than the items Mean Values indicate high variation between values, and abnormal distribution for data which should be rectified immediately.
- The Region called "**Other**" spends more in the 6 items in both "**Hotel** and **Retail**".

Since there is a huge spending on Other region by retailers, we must try to look into the sales(what it does right in the region as more spent means more sales) that this region brings to the table and based on those results we must do actions in the rest of the regions.

Conclusion:

Looking into the data of "Wholesale Customer", we saw some insights on how the data is spread across the various regions and channels, which helps us identify various places that needs some damage control and various places that need some constant attention. Insights from those identified places can give tips for planning and procurement of items which can further help in building in forecasting of annual spending in Portugal.

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
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

The dataset has 8 Numeric Columns and 6 categorical Columns.

Numerical Columns : ID, Age, GPA, Salary, Social Networking, Satisfaction, Spending and Text Messages.

Categorical Columns: Gender, Class, Major, Grad Intention, Employment and Computer.

Descriptive Statistics for the dataset:

	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
count	62.000000	62	62.000000	62	62	62	62.000000	62	62.000000	62.000000	62.000000	62.000000	62	62.000000
unique	NaN	2	NaN	3	8	3	NaN	3	NaN	NaN	NaN	NaN	3	NaN
top	NaN	Female	NaN	Senior	Retailing/Marketing	Yes	NaN	Part-Time	NaN	NaN	NaN	NaN	Laptop	NaN
freq	NaN	33	NaN	31	14	28	NaN	43	NaN	NaN	NaN	NaN	55	NaN
mean	31.500000	NaN	21.129032	NaN	NaN	NaN	3.129032	NaN	48.548387	1.516129	3.741935	482.016129	NaN	246.209677
std	18.041619	NaN	1.431311	NaN	NaN	NaN	0.377388	NaN	12.080912	0.844305	1.213793	221.953805	NaN	214.465950
min	1.000000	NaN	18.000000	NaN	NaN	NaN	2.300000	NaN	25.000000	0.000000	1.000000	100.000000	NaN	0.000000
25%	16.250000	NaN	20.000000	NaN	NaN	NaN	2.900000	NaN	40.000000	1.000000	3.000000	312.500000	NaN	100.000000
50%	31.500000	NaN	21.000000	NaN	NaN	NaN	3.150000	NaN	50.000000	1.000000	4.000000	500.000000	NaN	200.000000
75%	46.750000	NaN	22.000000	NaN	NaN	NaN	3.400000	NaN	55.000000	2.000000	4.000000	600.000000	NaN	300.000000
max	62.000000	NaN	26.000000	NaN	NaN	NaN	3.900000	NaN	80.000000	4.000000	6.000000	1400.000000	NaN	900.000000

We have two unique values in Class column with "**Senior**" being the top repetitive ,8 unique values from Major with the "**Retailing/Marketing**" being the top repetitive, 3 unique values from Grad Intention with the "**Yes**" being top repetitive, 3 unique values from Employment with the "**Part-Time**" being top repetitive and 3 unique values from Computer with the "Laptop" being top repetitive .

The Standard Deviation of GPA is **0.377** from Mean value **3.129**(Well-read Students).

Check for Null values

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62 entries, 0 to 61
Data columns (total 14 columns):
ID                  62 non-null int64
Gender              62 non-null object
Age                 62 non-null int64
Class               62 non-null object
Major               62 non-null object
Grad Intention      62 non-null object
GPA                 62 non-null float64
Employment          62 non-null object
Salary              62 non-null float64
Social Networking   62 non-null int64
Satisfaction        62 non-null int64
Spending            62 non-null int64
Computer            62 non-null object
Text Messages       62 non-null int64
dtypes: float64(2), int64(6), object(6)
memory usage: 6.9+ KB
```

From the above result we can see that there are **no null values** in the data provided.

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

2.1.1. Gender and Major
2.1.2. Gender and Grad Intention
2.1.3. Gender and Employment
2.1.4. Gender and Computer

2.1.1:

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:

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

2.1.3:

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

2.1.4:

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?

Total Male = 29

Total Gender = 62

Therefore Probability = 29/62

Probability in Percentage = 47%

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

Total Female = 33

Total Gender = 633

Therefore Probability = 33/62

Probability in Percentage = 53%

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.

Answering both questions in a single Data frame, we have:

	Major	Female	Male	Total_Male	Total_Female	Total_Gender	P(Major|Male) in percentage	P(Major|Female) in percentage
0	Accounting	3	4	29	33	7	13.79	9.09
1	CIS	3	1	29	33	4	3.45	9.09
2	Economics/Finance	7	4	29	33	11	13.79	21.21
3	International Business	4	2	29	33	6	6.9	12.12
4	Management	4	6	29	33	10	20.69	12.12
5	Other	3	4	29	33	7	13.79	9.09
6	Retailing/Marketing	9	5	29	33	14	17.24	27.27
7	Undecided	0	3	29	33	3	10.34	0

From the Highlighted part of the above result we can see the Percentage values all the conditional probability involving Major|Gender.

P(Major | Male):

P(Accounting | Male) = 4/29 or 13.79%

P(CIS | Male) = 1/29 or 3.45%

P(Economics/Finance| Male) = 4/29 or 13.79%

P(International Business| Male) = 2/29 or 6.9%

P(Management | Male) = 6/29 or 20.69%

P(Other | Male) = 4/29 or 13.79%

P(Retailing/Marketing| Male) = 5/29 or 17.24%

P(Undecided| Male) = 3/29 or 10.34%

P(Major| Female):

P(Accounting| Female) = 3/33 or 9.09%

P(CIS | Female) = 3/33 or 9.09%

P(Economics/Finance | Female) = 7/33 or 21.21%

P(International Business | Female) = 4/33 or 12.12%

P(Management | Female) = 4/33 or 12.12%

P(Other | Female) = 3/33 or 9.09%

P(Retailing/Marketing | Female) = 9/33 or 27.27%

P(Undecided | Female) = 0/33 or 0%

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.

Answering both questions in a single Data frame, we have:

	Grad Intention	Female	Male	Total_Male	Total_Female	Total_Gender	P(Grad Intention|Male) in percentage	P(Grad Intention|Female) in percentage
0	No	9	3	29	33	12	10.34	27.27
1	Undecided	13	9	29	33	22	31.03	39.39
2	Yes	11	17	29	33	28	58.62	33.33

From the Highlighted part of the above result we can see the Percentage values all the conditional probability involving P(Grad Intention | Gender).

P(Grad Intention | Male):

P(No | Male) = 3/29 or 10.34%

P(Undecided| Male) = 9/29 or 31.03%

P(Yes | Male) = 17/29 or 58.62%

P(Grad Intention | Female):

P(No | Female) = 9/33 or 27.27%

P(Undecided| Female) = 13/33 or 39.39%

P(Yes | Female) = 11/33 or 33.33%

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

Answering both questions in a single Data frame, we have:

	Employment	Female	Male	Total_Male	Total_Female	Total_Gender	P(Employment|Male) in percentage	P(Employment|Female) in percentage
0	Full-Time	3	7	29	33	10	24.14	9.09
1	Part-Time	24	19	29	33	43	65.52	72.73
2	Unemployed	6	3	29	33	9	10.34	18.18

From the Highlighted part of the above result we can see the Percentage values all the conditional probability involving P(Employment | Gender).

P(Employment | Male):

P(Full-Time | Male) = 7/29 or 24.14%

P(Part-Time | Male) = 19/29 or 65.52%

P(Unemployed | Male) = 3/29 or 10.34%

P(Employment | Female):

P(Full-Time | Female) = 3/33 or 9.09%

P(Part-Time | Female) = 24/33 or 72.73%

P(Unemployed | Female) = 6/33 or 18.18%

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

Answering both questions in a single Data frame, we have:

	Computer	Female	Male	Total_Male	Total_Female	Total_Gender	P(Computer|Male) in percentage	P(Computer|Female) in percentage
0	Desktop	2	3	29	33	5	10.34	6.06
1	Laptop	29	26	29	33	55	89.66	87.88
2	Tablet	2	0	29	33	2	0	6.06

From the Highlighted part of the above result we can see the Percentage values all the conditional probability involving P(Computer | Gender).

P(Computer | Male):

P(Desktop | Male) = 3/29 or 10.34%

P(Laptop | Male) = 26/29 or 89.66%

P(Tablet | Male) = 0/29 or 0%

P(Computer | Female):

P(Desktop | Female) = 2/33 or 6.06%

P(Laptop | Female) = 29/33 or 87.88%

P(Tablet | Female) = 2/33 or 6.06%

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.

First before comparing the previous results we must look at the overall percentage between male and female students in this survey. They are almost half (47% Male and 53% Percent Female) and therefore we can use this data to further analyse the spread.

Looking at Major, the spread of male vs female in all the major looks dependent. For Example: P(Accounting | Male) = 4/29 or 13.79% and P(Accounting | Female) = 3/33 or 9.09% , we can see that from the results these two probabilities are not equal and if they were hypothetically equal then picking up Major wouldn't have mattered with genders and that's why we say it's Dependent. We can see the case in all the conditional probabilities related to Major with Gender.

Looking on the Grad Intention, we can say it Dependent to Gender. For Example: P(Yes | Male) = 17/29 or 58.62% and P(Yes | Female) = 11/33 or 33.33%, we can see that from the results these two probabilities are not equal and if they were hypothetically equal then Having Grad Intention wouldn't have mattered with genders and that's why we say it's Dependent. We can see the case in all the conditional probabilities related to Grad Intention with Gender.

Looking at the Employment category, we can again say that it Dependent to Gender. For Example: P(Full-Time | Male) = 7/29 or 24.14% and P(Accounting | Female) = 3/33 or 9.09%, we can see that from the results these two probabilities are not equal and if they were hypothetically equal then taking Full-Time job wouldn't have mattered with genders and that's why we say it's Dependent. We can see the case in all the conditional probabilities related to Employment Category with Gender.

Looking at the Type of Computer preference, we can say that it is Dependent of Gender. For Example: P(Desktop | Male) = 3/29 or 10.34% and P(Desktop | Female) = 2/33 or 6.06%, we can see that from the results these two probabilities are not equal and if they were hypothetically equal then preferring Desktop wouldn't have mattered with genders and that's why we say it's Dependent. We can see the case in all the conditional probabilities related to Computer preference with Gender.

Part II

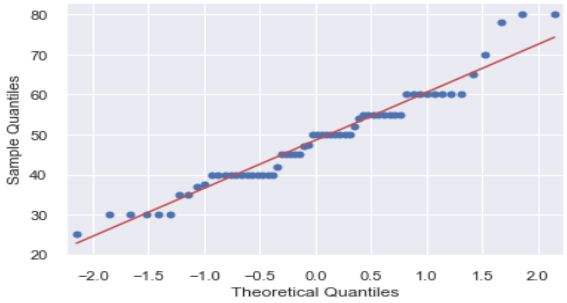
2.4. Note that there are three numerical (continuous) variables in the data set, Salary, Spending and Text Messages. For each of them comment whether they follow a normal distribution.

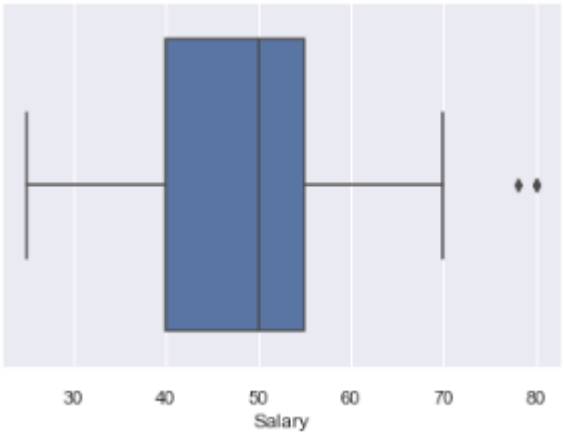
Write a note summarizing your conclusions.

[Recall that symmetric histogram does not necessarily mean that the underlying distribution is symmetric]

Salary

```
The Sknewness of Salary is 0.5347008436225946 - Right Skewed
```

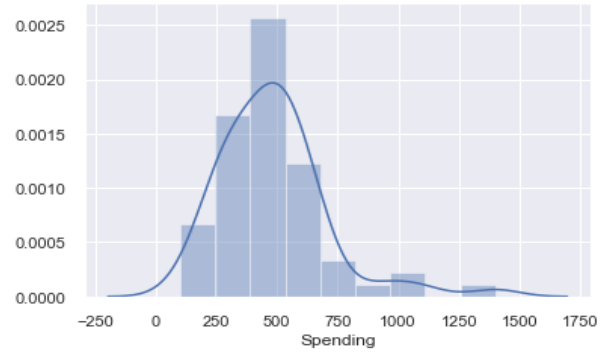


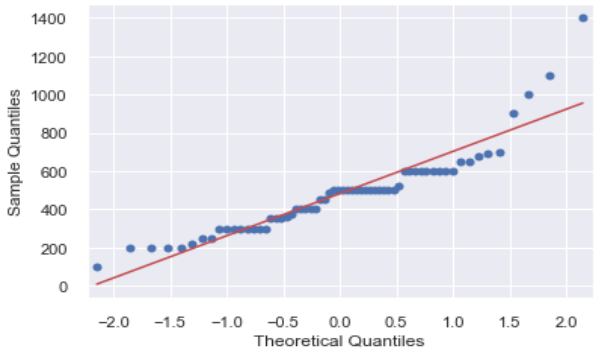


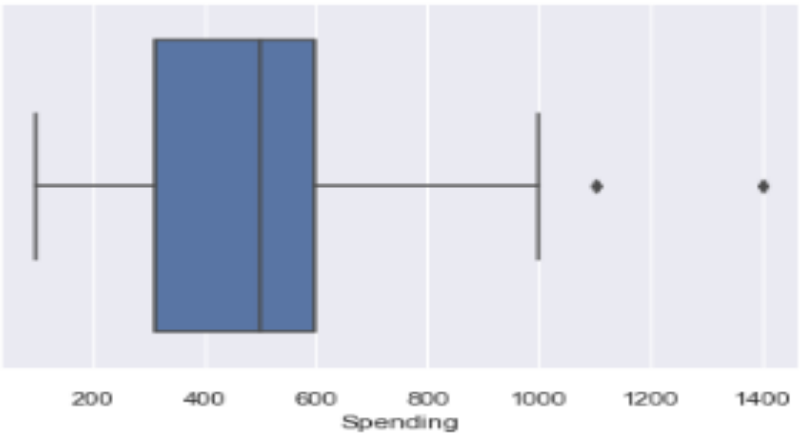
We can see from the Histogram , QQPLOT and Box Plot that **Salary** is Normally Distributed(not perfect though).

Spending

The Skewness of Spending is 1.5859147414045331 - Right Skewed



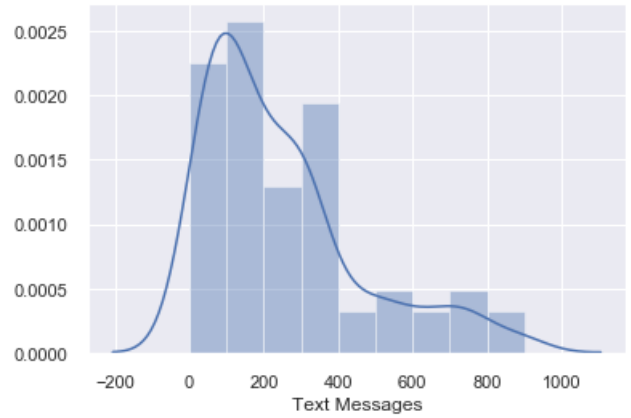


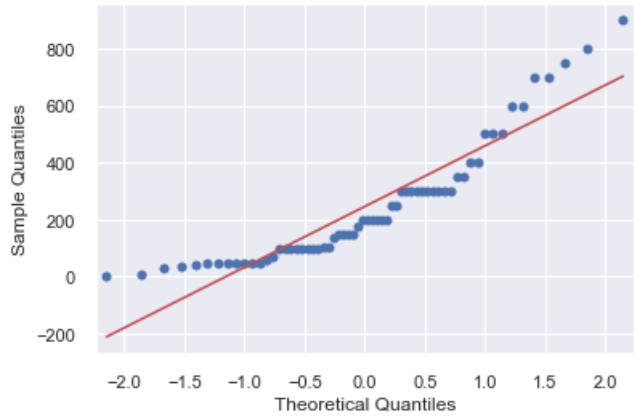


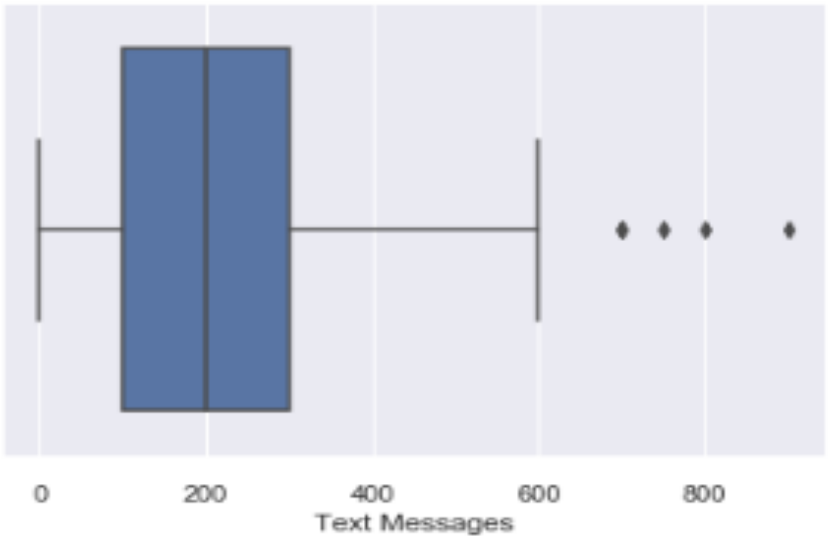
We can see from the Histogram , QQPLOT and Box Plot that **Spending** is Normally Distributed(not perfect though).

Text Messages

```
The Skewness of Text Messages is 1.2958079731054333 - Right Skewed
```







We can see from the Histogram , QQPLOT and Box Plot that **Text Messages** is Normally Distributed(not perfect though).

Although we have evidence that the three continuous columns are normally distributed, we can look at the describe function to see if 68% of values are in the IQR range to further prove it.

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

We can see that one position from either side of mean with Standard Deviation($\mu \pm \sigma$) falls in the IQR range(25% to 75%)

From the above result we can see that these continuous variables may not be proper normally distributed but shows signs of normally distribution.

Problem 2 Summary:

Part I

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

2.1.1. Gender and Major – A crosstab Dataframe was created with Gender as Rows and Major as Columns.

2.1.2. Gender and Grad Intention– A crosstab Dataframe was created with Gender as Rows and Grad Intention as Columns.

2.1.3. Gender and Employment– A crosstab Dataframe was created with Gender as Rows and Employment as Columns.

2.1.4. Gender and Computer– A crosstab Dataframe was created with Gender as Rows and Computer as Columns.

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:

Male:

Probability = 29/62 and Probability in Percentage = 47%

Female:

Probability = 33/62 and Probability in Percentage = 53%

2.2.2:

P(Major | Male):

P(Accounting | Male) = 4/29 or 13.79%

P(CIS | Male) = 1/29 or 3.45%

P(Economics/Finance| Male) = 4/29 or 13.79%

P(International Business| Male) = 2/29 or 6.9%

P(Management | Male) = 6/29 or 20.69%

P(Other | Male) = 4/29 or 13.79%

P(Retailing/Marketing| Male) = 5/29 or 17.24%

P(Undecided| Male) = 3/29 or 10.34%

P(Major| Female):

P(Accounting| Female) = 3/33 or 9.09%

P(CIS | Female) = 3/33 or 9.09%

P(Economics/Finance | Female) = 7/33 or 21.21%

P(International Business | Female) = 4/33 or 12.12%

P(Management | Female) = 4/33 or 12.12%

P(Other | Female) = 3/33 or 9.09%

P(Retailing/Marketing | Female) = 9/33 or 27.27%

P(Undecided | Female) = 0/33 or 0%

2.2.3:

P(Grad Intention | Male):

P(No | Male) = 3/29 or 10.34%

P(Undecided| Male) = 9/29 or 31.03%

P(Yes | Male) = 17/29 or 58.62%

P(Grad Intention | Female):

P(No | Female) = 9/33 or 27.27%

P(Undecided| Female) = 13/33 or 39.39%

P(Yes | Female) = 11/33 or 33.33%

2.2.4

P(Employment | Male):

P(Full-Time | Male) = 7/29 or 24.14%

P(Part-Time | Male) = 19/29 or 65.52%

P(Unemployed | Male) = 3/29 or 10.34%

P(Employment | Female):

P(Full-Time | Female) = 3/33 or 9.09%

P(Part-Time | Female) = 24/33 or 72.73%

P(Unemployed | Female) = 6/33 or 18.18%

2.2.5

P(Computer | Male):

P(Desktop | Male) = 3/29 or 10.34%

P(Laptop | Male) = 26/29 or 89.66%

P(Tablet | Male) = 0/29 or 0%

P(Computer | Female):

P(Desktop | Female) = 2/33 or 6.06%

P(Laptop | Female) = 29/33 or 87.88%

P(Tablet | Female) = 2/33 or 6.06%

2.3

All the column variables in each case is **Dependent** of Gender.

Part II

2.4 The continuous variables “Salary”, “Spending” and “Text Messages” are indeed normally distributed.

Conclusion:

Various Conditional Probabilities were applied on the Survey to find interesting insights on the spread of data. Part I provided us the information on Contingency tables and various probabilities derived from it. Also, it gave us a pretty good look on the how Gender influences in the Survey from various Conditional Probabilities. Part II gave us an additional takeaway like the distribution that the continuous variables in the Survey pose on(like since we have concrete evidence on Normal Distribution, we can use these columns to make some hypothesis testing).

Problem 3 Statement:

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 colouring 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 claims that the mean moisture content cannot be greater than 0.35 pound per 100 square feet.

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:

	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

We can see **two numerical columns** A and B with moisture in pound per 100 square feet(According to the Statement).

Descriptive Statistics for the dataset:

	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

The Standard Deviation of A and B is very nearly similar to each other unlike mean which shows noticeable difference(which we are going to check indefinitely with hypothesis testing.).

Check for NULL values:

Looking on the dataset we can see that we have 5 NULL values for column 'B'.

	Total Null for B	Total Null for A
0	5	0

Before Answering the questions posted, a **One sample TTest** was performed on Column 'A' and 'B' based on the null and alternate Hypothesis provided as below.

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

H0<=0.35

HA>0.35

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

H0<=0.35

HA>0.35

Basic Assumption is that the data is Normally Distributed and **Alpha Value** is 0.05

For A column:

P-Value:0.07477633144907513 T-Statistic:-1.4735046253382782

Since P-Value is greater than Alpha we **fail to reject** the Null Hypothesis to say that moisture is lesser than or equal to 0.35 pound per 100 square feet.

For B column:

P-Value:0.0020904774003191826 T-Statistic:-3.1003313069986995

Since P-Value is lesser than Alpha we **reject** the Null Hypothesis to say that moisture is greater than 0.35 pound per 100 square feet.

3.1 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?

First Forming the Null and the Alternate Hypothesis:

Null Hypothesis(H0) : $\mu A = \mu B$ or $\mu A - \mu B = 0$

Alternate Hypothesis(Ha or H1) : $\mu A \neq \mu B$ or $\mu A - \mu B \neq 0$

Finding the test to check the hypothesis:

We are going to perform a **Two Sample T-Test** here.

Alpha Value – Using Default 95% confidence interval or 0.05

Performed Two Sample T test and got :

T_Statistic value = 1.2896282719661123
P_Value = 0.2017496571835306

Since the P_Value we got is greater than Alpha **we fail to reject** the NULL hypothesis and therefore we can conclude that **Mean values of column 'A' and 'B' are equal**.

Some of the assumptions that we need to check before the test for equality of means is performed are:

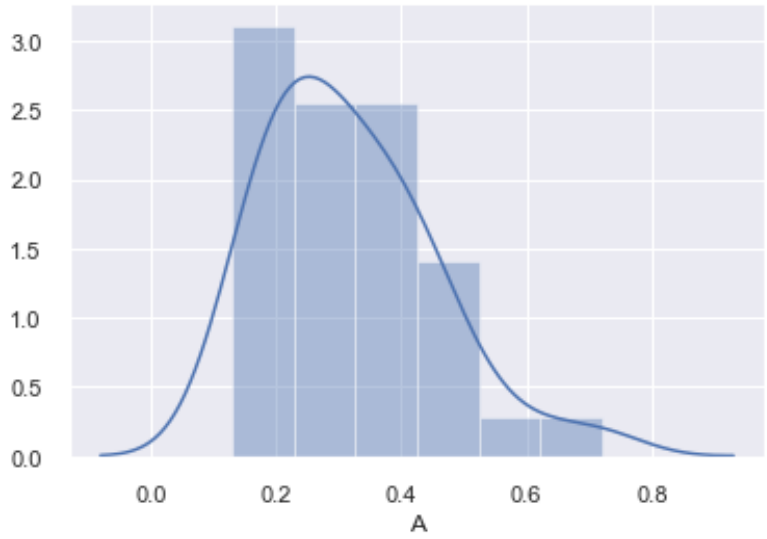
- Distribution is Normal(Check for calculations in 3.2)
- Alpha value is 0.05
- Columns are independent of each other(Mann-Whitney-U test – Refer Notebook attached).
- No significant outliers in the two groups(large set of outliers).

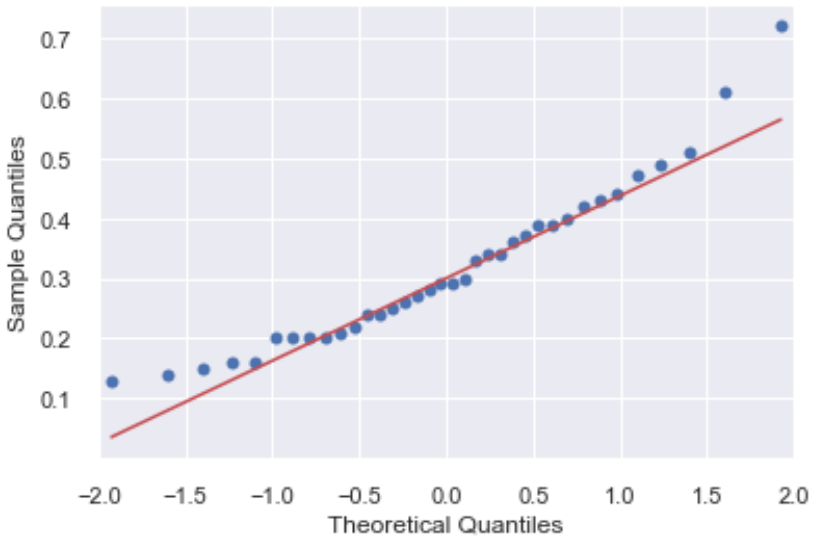
3.2. What assumption about the population distribution is needed in order to conduct the hypothesis tests above?

We assume a normal distribution of data generally to conduct Hypothesis. Therefore, looking at them individually, we have

For Column A:

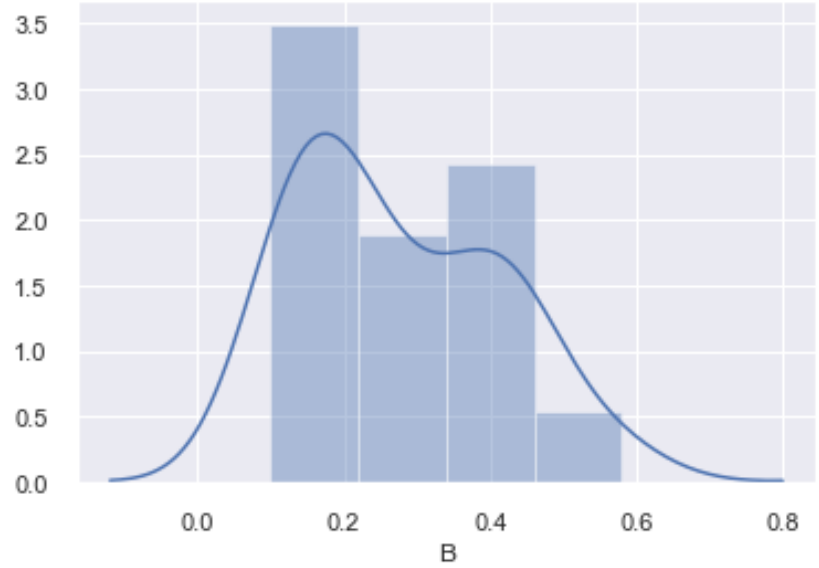
```
The Skewness of Text Messages is 0.9506185720492205 - Right Skewed
```

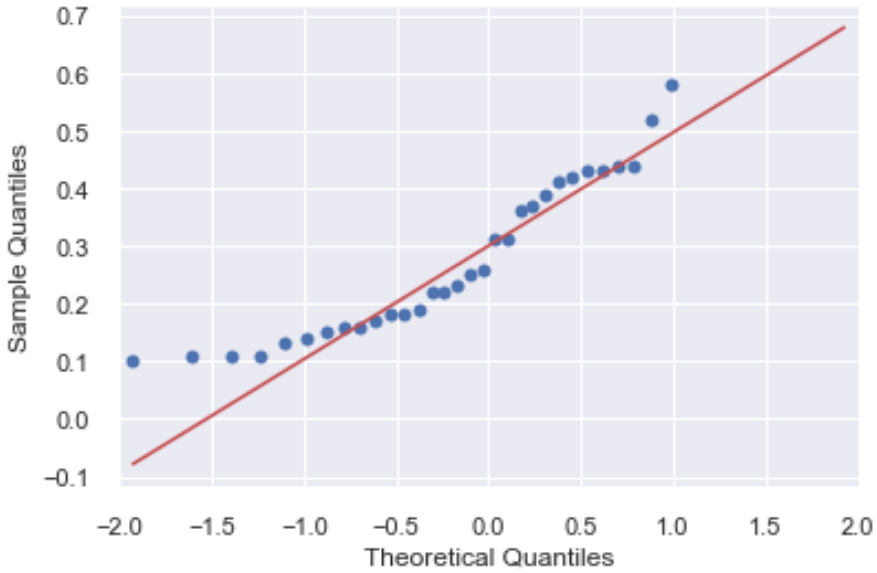




For Column B:

```
The Skewness of Text Messages is 0.5134239595793691 - Right Skewed
```





Although we assume that both A and B column are normally distributed, we can look at the above result and see that QQPlot and Distplot shows sign of Normal Distribution to further prove it (Empirical Rule - 68% of values are in the IQR range).

Problem 3 Summary:

3.1 : Mean values of column 'A' and 'B' are equal.

Assumptions:

- Distribution is Normal
- Alpha value is 0.05
- Columns are independent of each other(Mann-Whitney-U test – Refer Notebook attached).
- No significant outliers in the two groups(large set of outliers).

3.2 : Data should be Normally Distributed and based on the skewness, Distplot and QQplot it is indeed normally Distributed.

Conclusion:

Before going into the results of the questions asked, we saw that results of One Sample T Test which showed that Column 'B" showed mean values(moisture in pound per 100 square feet) greater than 0.35 with 95% confidence intervals which must be looked upon by the company. We also found that the population mean of A and B is not equal(No Surprise there) and the data is not a perfect Normal distribution(Which can be rectified by increasing the sample size – Proven results).