BUSINESS REPORT

SMDM PROJECT

SRINIDHI. D

BUSINESS REPORT

Problem -1: Wholesale Dataset

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 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).

About the Data: The dataset contains 8 variables: Channel, Region, Fresh, Milk, Grocery, Frozen, Detergents_Paper, Delicatessen.

1) Channel (Nominal): Retail or Hotel

2) Region (Nominal): Oporto or Lisbon or Other

3) Fresh (Continuous): Annual spending on fresh products

4) Milk (Continuous): Annual spending on milk products

5) Grocery (Continuous): Annual spending on grocery products

6) Frozen (Continuous): Annual spending on frozen products

7) Detergents_Paper (Continuous): Annual spending on detergents and paper products

8) Delicatessen (Continuous): Annual spending on delicatessen products

Basis this data, we need to find the Region and Channel that spent the most and least, the behaviour of the items across region and channel, and the item that shows the most consistent and inconsistent behaviour. Also, we need to suggest suitable recommendations.

	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	518

1.1 Descriptive Statistics:

Items	count	mean	std	min	25%	50%	75%	max
Fresh	440	12000.30	12647.33	3	3127.75	8504.00	16933.75	112151
Milk	440	5796.27	7380.38	55	1533.00	3627.00	7190.25	73498
Grocery	440	7951.28	9503.16	3	2153.00	4755.50	10655.75	92780
Frozen	440	3071.93	4854.67	25	742.25	1526.00	3554.25	60869
Detergents_Paper	440	2881.49	4767.85	3	256.75	816.50	3922.00	40827
Delicatessen	440	1524.87	2820.11	3	408.25	965.50	1820.25	47943

a. Count- Each variable contains 440 data points

b. Mean- This is highly affected by outliers; As observed, the average spending for 'Fresh' product is the highest and 'Delicatessen' product is the lowest.

c. Standard Deviation- It shows us how the data points are spread across the mean. Smaller the value, closer is the data point to mean and vice versa. 'Fresh' products seems to have the highest standard deviation and 'Delicatessen' products have the lowest standard deviation.

d. Median- Shows us the 50^{th} percentile of the dataset and is the least affected by the outlier.

Here, for all the six variables under consideration, mean is greater than the median implying that the data seems to be right skewed.

e. Minimum and Maximum- Shows the smallest and the largest values of the dataset. From the output, we can see that the minimum and maximum are quite far away from each other. In other words, it shows us the spread of the data.

f. Inter Quartile Range- This also helps assess the spread of the data. Higher the IQR, greater is the spread of the data.

Here, for all the variables, we see that IQR (the difference between Q1 and Q3), is greater implying the spread of the data is greater.

1.1 Which Region and which Channel seems to spend more? Which Region and which Channel seems to spend less?

Grouping the data by 'Channel' and 'Region' with respect to 'Items' and summing them across.

Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Hotel	Lisbon	761233	228342	237542	184512	56081	70632
	Oporto	326215	64519	123074	160861	13516	30965
	Other	2928269	735753	820101	771606	165990	320358
Retail	Lisbon	93600	194112	332495	46514	148055	33695
	Oporto	138506	174625	310200	29271	159795	23541
	Other	1032308	1153006	1675150	158886	724420	191752

Now summing all 6 items together and arranging them in descending order.

Channel	Region	Amount
Hotel	Other	5742077
Retail	Other	4935522
Hotel	Lisbon	1538342
Retail	Lisbon	848471
	Oporto	835938
Hotel	Oporto	719150

From the above table, we can derive the maximum and the minimum spent by the Channel and Region.

Channel	Region	Amount
Hotel	Other	5742077

Channel	Region	Amount
Hotel	Oporto	719150

'Hotel' under Channel and 'Other' under Region spent the maximum amount of money and 'Hotel' and 'Oporto' spent the minimum.

1.2. <u>There are 6 different varieties of items are considered. Do all varieties show similar behaviour across Region and Channel?</u>

<u>I. Region and Items:-</u>

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

	Region	Lisbon	Oporto	Other
Fresh	count	77	47	316
	mean	11101.73	9887.68	12533.47
	std	11557.44	8387.9	13389.21
	min	18	3	3
	25%	2806	2751.5	3350.75
	50%	7363	8090	8752.5
	75%	15218	14925.5	17406.5
	max	56083	32717	112151

	Region	Lisbon	Oporto	Other
Milk	count	77	47	316
	mean	5486.42	5088.17	5977.09
	std	5704.86	5826.34	7935.46
	min	258	333	55
	25%	1372	1430.5	1634
	50%	3748	2374	3684.5
	75%	7503	5772.5	7198.75
	max	28326	25071	73498

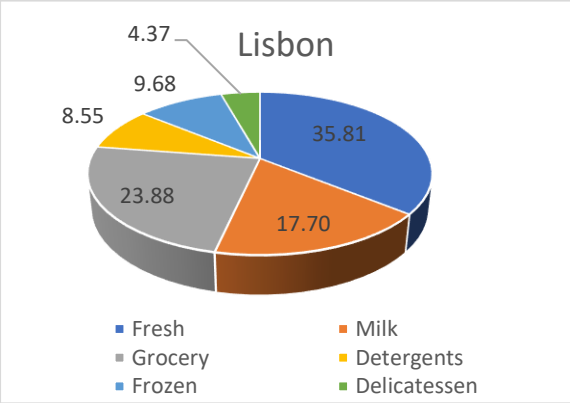
	Region	Lisbon	Oporto	Other
Grocery	count	77	47	316
	mean	7403.08	9218.6	7896.36
	std	8496.29	10842.75	9537.29
	min	489	1330	3
	25%	2046	2792.5	2141.5
	50%	3838	6114	4732
	75%	9490	11758.5	10559.8
	max	39694	67298	92780

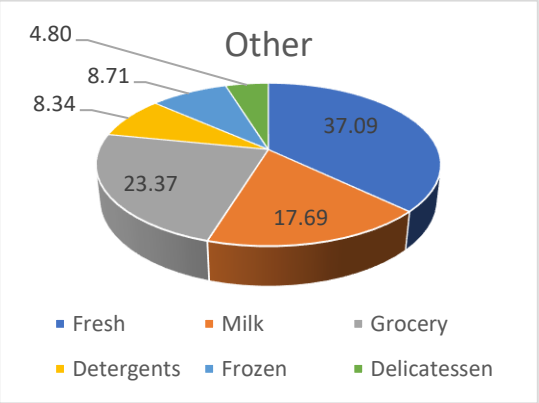
	Region	Lisbon	Oporto	Other
Detergents _Paper	count	77	47	316
	mean	2651.12	3687.47	2817.75
	std	4208.46	6514.72	4593.05
	min	5	15	3
	25%	284	282.5	251.25
	50%	737	811	856
	75%	3593	4324.5	3875.75
	max	19410	38102	40827

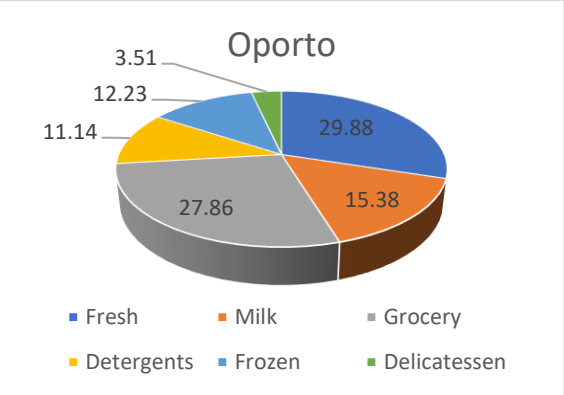
	Region	Lisbon	Oporto	Other
Frozen	count	77	47	316
	mean	3000.34	4045.36	2944.59
	std	3092.14	9151.78	4260.13
	min	61	131	25
	25%	950	811.5	664.75
	50%	1801	1455	1498
	75%	4324	3272	3354.75
	max	18711	60869	36534

	Region	Lisbon	Oporto	Other
Delicatessen	count	77	47	316
	mean	1354.9	1159.7	1620.6
	std	1345.42	1050.74	3232.58
	min	7	51	3
	25%	548	540.5	402
	50%	806	898	994
	75%	1775	1538.5	1832.75
	max	6854	5609	47943

Mean for Region and Items







- *As observed, the average spending for 'Fresh' product is the highest followed by 'Grocery', 'Milk', 'Frozen' and 'Detergents_Paper' under the 'Lisbon' region and 'Delicatessen' is lowest in the same region.
- This holds for both 'Oporto' and 'Other' regions as well.
- Mean across the six varieties under consideration for the 'Lisbon' and 'Other' regions does show a similar behaviour but that is not the case with 'Oporto' region.
- As we can see from the graphs, almost all the varieties under consideration for Lisbon and Other have a similar pattern in average spending but when Oporto is taken into account, we see that less than 30% is spent on average on Milk as compared to the other two. Also there is approximately 4% difference in (average) spending on Grocery in Oporto in comparison to Lisbon and Other regions. Oporto has a different pattern when considered 'Frozen' products.
- The order*(highest to lowest spent) is similar for all the three regions however the percentage (of amount spent) on each individual varieties of items are different. Again, we somewhat see a similar percentage spent in Lisbon and Other regions but that quite does not hold for Oporto region. This further requires a study of consumer behaviour in Lisbon, Other and Oporto region for deeper understanding of this trend.
- One pitfall is that if the data is highly skewed, mean cannot be accurate.

Median for Region and Items

- Considered one of the best measures of central tendency as it is least affected by the outliers present in the dataset. This shows the 50th percentile of the data.
- Under Lisbon region: 'Fresh' products has the highest median followed by 'Grocery', 'Milk', 'Frozen', 'Delicatessen' and 'Detergents_Paper' products has the lowest median.
- This holds for both 'Oporto' and 'Other' regions as well.
- One of the most important measures of central tendency and least affected by outliers in the data is the median. For the six variables under consideration, behaves in a similar fashion across all three regions.

Inter Quartile Range (IQR) for Region and Items

	Lisbon	Oporto	Other
Fresh	12412	12174	14056
Milk	6131	4342	5564.8
Grocery	7444	8966	8418.3
Detergents	3309	4042	3624.5
Frozen	3374	2460.5	2690
Delicatessen	1227	998	1430.8

- Helps assessing the spread of the dataset.
- It is the difference between the 3rd quartile and the 1st quartile.
- Under 'Lisbon' region: 'Fresh' has the highest IQR implying the spread is largest followed by 'Grocery', 'Milk', 'Frozen', 'Detergents_Paper' and 'Delicatessen' has the lowest IQR with lowest spread.
- Under 'Oporto' region: 'Fresh' has the highest IQR products followed by 'Grocery', 'Milk', 'Detergents_Paper', 'Frozen' and 'Delicatessen' products have the highest and lowest IQR respectively.
- The pattern of 'Oporto' region is followed by the 'Other' region as well.
- IQR, a measure of dispersion, shows us the spread of the data in terms of the six varieties under consideration follows a similar order in 'Oporto' and 'Other' regions but is different in 'Lisbon' region.
- Further dissecting the data into Region vs Varieties, we do not see any similar behaviour shown as far as IQR is considered.
- Between Regions, here, we observe that the spread of the data for each product is not much, but there is still some.
- If we consider the pattern from the above table, the spread of data highly varies highly varies within each varieties under consideration for the three regions.

Range for Region and Items

	Lisbon	Oporto	Other
Fresh	56065	32714	112148
Milk	28068	24738	73443
Grocery	39205	65968	92777
Detergents	19405	38087	40824
Frozen	18650	60738	36509
Delicatessen	6847	5558	47940

- One of the easiest way to see the spread of the data is to use the Minimum and Maximum values.
- The difference between max and min value gives us the Range of the dataset which is affected by the outliers in the dataset.
- In IQR also, we saw that the spread of the data under each varieties across the three regions highly varies. Within region for different category of items also, we see the range is spread out largely. The same observation can be made from the above Range table which also shows us there is presence of outliers.
- Range, the measure of dispersion, we see that the spread of the data in terms of the six varieties in 'Lisbon', 'Oporto' and 'Other' regions is different from each other.

So, in conclusion for Region and Items, taking all measures into consideration, the varieties of items does not necessarily show a similar behaviour across the three regions.

II. Channel and Items:-

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

	Channel	Hotel	Retail
Fresh	count	298	142
	mean	13475.56	8904.32
	std	13831.69	8987.71
	min	3	18
	25%	4070.25	2347.75
	50%	9581.5	5993.5
	75%	18274.75	12229.8
	max	112151	44466

	Channel	Hotel	Retail
Milk	count	298	142
	mean	3451.72	10716.5
	std	4352.17	9679.63
	min	55	928
	25%	1164.5	5938
	50%	2157	7812
	75%	4029.5	12162.75
	max	43950	73498

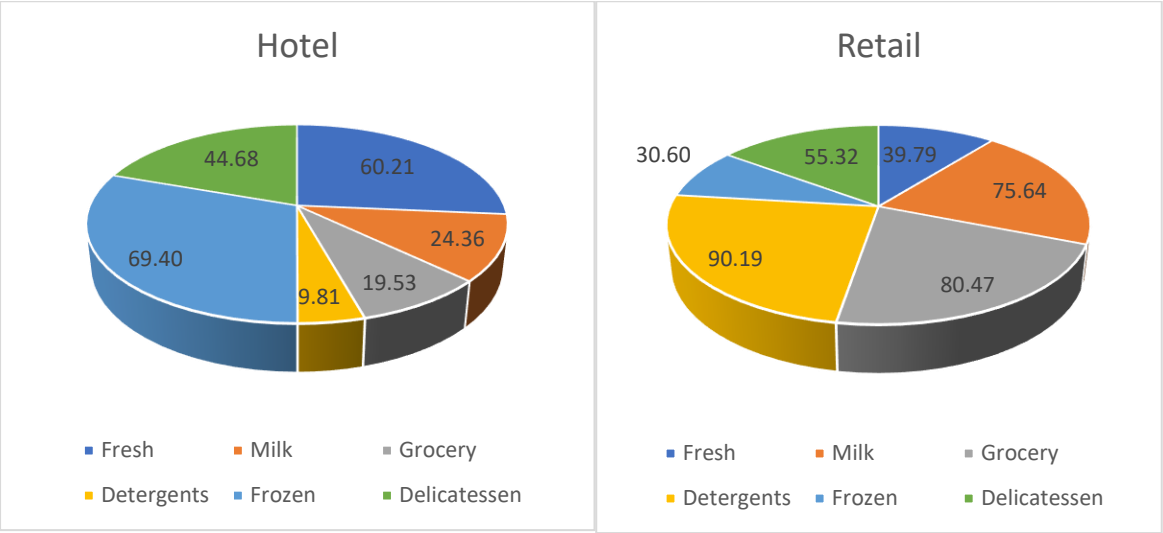
	Channel	Hotel	Retail
Grocery	count	298	142
	mean	3962.14	16322.85
	std	3545.51	12267.32
	min	3	2743
	25%	1703.75	9245.25
	50%	2684	12390
	75%	5076.75	20183.5
	max	21042	92780

	Channel	Hotel	Retail
Detergent_Paper	count	298	142
	mean	790.56	7269.5
	std	1104.09	6291.08
	min	3	332
	25%	183.25	3683.5
	50%	385.5	5614.5
	75%	899.5	8662.5
	max	6907	40827

	Channel	Hotel	Retail
Frozen	count	298	142
	mean	3748.25	1652.61
	std	5643.91	1812.8
	min	25	33
	25%	830	534.25
	50%	2057.5	1081
	75%	4558.75	2146.75
	max	60869	11559

	Channel	Hotel	Retail
Delicatessen	count	298	142
	mean	1415.95	1753.43
	std	3147.42	1953.79
	min	3	3
	25%	379	566.75
	50%	821	1350
	75%	1548	2156
	max	47943	16523

Mean for Channel and Items



- Under the 'Hotel' channel: As observed, the average spending for 'Fresh' product is the highest followed by 'Grocery, 'Frozen', 'Milk', 'Delicatessen' and 'Detergents_Paper' is lowest which can be interpreted in the following manner: Since Hotels mainly use Fresh products to cook for their customers daily, their average spending on the same is highest. On the other hand, their use of Detergents_Paper seems to be less explaining their less usage of these products within this particular channel.
- Under the 'Retail' channel: The average spending on 'Grocery' products is highest 'Milk', 'Fresh', 'Detergents_Paper', 'Delicatessen' and 'Frozen' with lowest, this implies that since retail store mainly serves the people nearby with their daily/monthly grocery items, the average spending on this is highest, however, it is interesting to see that the 'Frozen' items have the lowest mean amongst the six variables given, as Retail channels earn a lot of revenue through frozen products as well but their average spending on the same is the least.
- As far as Retail channel is considered, Detergents_Paper products most of its average spending (percentage) from the above pie chart whereas Fresh products constitute the maximum average spending (percentage) which is intuitive as discussed above.
- Almost the reverse situation is observed we see the lowest average spending on the items.
- Across the two channels, the six items under consideration does not seem to behave in a similar fashion as the average spending on these items are different in 'Retail' and 'Hotel' channels.

Median for Channel and Items

- Considered one of the best measures of central tendency as it is least affected by the outliers present in the dataset. This shows the 50^{th} percentile of the data.
- Under the 'Hotel' channel: The least median is observed in 'Detergents_Paper' products and highest in 'Fresh' products followed by 'Grocery', 'Milk', 'Frozen' and 'Delicatessen'.
- Under the 'Retail' channel: The median of 'Grocery' is the highest followed by 'Milk', 'Fresh', 'Detergents_Paper', 'Delicatessen' and 'Frozen' products has the least median.
- Considering the median value, there seems to be different behaviour of each of the six items with respect to the two channels.

Inter Quartile Range (IQR) for Channel and Items

	Hotel	Retail
Fresh	14204.5	9882
Milk	2865	3373
Grocery	3373	17440
Detergents	716.25	4979
Frozen	3728.75	1612.5
Delicatessen	1169	1589.25

- This is another way to assess the spread of the dataset.
- It is the difference between the 3^{rd} quartile and the 1^{st} quartile.
- Under 'Hotel' channel: 'Fresh' has the highest IQR implying the spread is largest in this case followed by 'Frozen', 'Grocery', 'Milk', 'Delicatessen' and 'Detergents_Paper' has the lowest IQR with lowest spread.
- Under 'Retail' channel: 'Grocery' has the highest IQR implying the spread is largest in this case followed by 'Fresh', 'Milk', 'Detergents_Paper', 'Frozen' and 'Delicatessen' has the lowest IQR with lowest spread.
- From the above table, we see that the IQR for the six items across the two channels highly varies.
- Dissecting the data into Retail and the six items, we see that the data is largely spread out and same is the case with Hotel and the six items.
- Between channels, the IQR for Milk and Delicatessen are the least fluctuating.
- Inter Quartile Range, which is also a measure of dispersion behaves in a different fashion for six items under consideration for both the channels.

Range for Channel and Items

	Hotel	Retail
Fresh	112148	44448
Milk	43895	72570
Grocery	21039	90037
Detergents	6904	40495
Frozen	60844	11526
Delicatessen	47940	16520

- One of the easiest way to see the spread of the data is to use the Minimum and Maximum values.
- The difference between max and min value gives us the Range of the dataset which is also affected by the outliers in the dataset.
- Under the 'Hotel' channel: The range for the 'Fresh' products is the highest followed by 'Frozen', 'Delicatessen', 'Milk', 'Grocery' and lowest for the 'Detergents_Paper' products.
- Under 'Retail' channel: The range for the 'Grocery' products is the highest followed by 'Milk', 'Fresh', 'Detergents_Paper', 'Delicatessen' and lowest for the 'Frozen' products.
- In IQR also, we saw that the spread of the data under each varieties across the two channels highly varies. The same observation can be made from the above Range table which also shows us there is presence of outliers.
- Within channel for different category of items also, we see the range is spread out largely.
- The Range, one of the measures of dispersion also behaves in a different fashion for six items under consideration for both the channels.

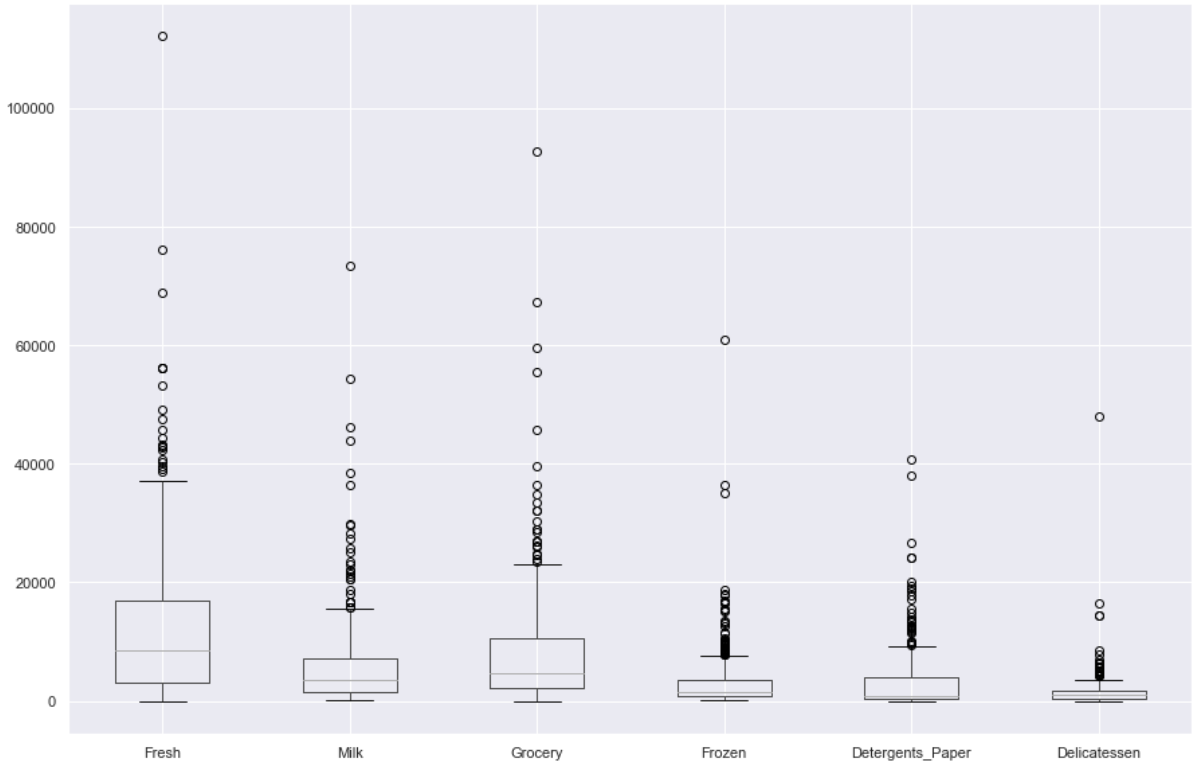
So, in conclusion for Channel and Items, taking all measures into consideration, the varieties of items does not necessarily show a similar behaviour across the two channels.

1.3. <u>On the basis of the descriptive measure of variability, which item shows the most inconsistent behaviour? Which items shows the least inconsistent behaviour?</u>

Items	Coefficient of Variation
Fresh	105.39
Milk	127.32
Grocery	119.51
Frozen	158.03
Detergents_Paper	165.46
Delicatessen	184.94

- Coefficient of variation is used as a measure of consistency of the dataset and values are on unitless scale.
- Higher the CV, higher is the inconsistency and the lower it is, lesser is the inconsistency.
- Delicatessen is the most inconsistent variable and Fresh is the least inconsistent variable.
- For this particular dataset, it is better to use Coefficient of Variation rather than standard deviation as the means for each variable is quite different from each other.

1.4. <u>Are there any outliers in the data?</u>



- Yes, there are outliers in the dataset.
- Outliers are defined as data points that are statistically inconsistent from the rest of the data.
- Outlier implies experimental error or occurs due to random variation. In other words, Outlier is an observation that appears far away and diverges from an overall pattern in a sample. Hence the treatment of the outliers become necessary before proceeding further.
- 'Fresh', 'Milk', 'Grocery', 'Frozen', 'Detergents_Paper' and 'Delicatessen' contains outliers.

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

Recommendations:

- ***Hotel Channel***:

Hotel is B2B in nature, we need to do a deeper research in all the regions to make an observation on the needs of the customers of this channel, the total number of hotels in those regions were the spending is not picking up to improve them further; It is also important to retain the existing customers and bring in new customers.

- ***Retail Channel***:

Retail is B2C in nature. Hence it is important to understand whether the retail who buys from this particular wholesaler buys from any other wholesaler or not. If yes, what are the products he is buying and why is he buying from that wholesaler instead of us (this wholesaler) and how can we convert this retailer to totally buy from us. In this case, 'loyalty programs' can encourage them to an extent. Doing a research on the competition (Wholesalers), will also provide a deeper insights on the prices they are offering and programs to hold the existing customers and to bring in new customers.

- ***Regions***:

We saw that spending in regions like Oporto (both Hotel and Retail channel) is not picking up whereas the Other regions in both channels were the highest spender. It is possible that customer understanding in that particular region might be lacking. Understanding on what could possibly be improved in the regions will help changing the situation in Oporto.

Problem -2: Survey Dataset

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.

About the Data:

1) ID- Student's Number

2) Gender- Gender of the students (Male/Female)

3) Age- Age (in number of years) of the student

4) Class- Class of the students (Junior/Senior/Sophomore)

5) Major- Major subject of the students (Accounting/CIS/Economics Finance/International Business/Management/Other/Retailing Marketing/Undecided)

6) Grad Intention- Whether students have the intention to graduate (Yes/No/Undecided)

7) GPA- GPA Score of the students

8) Employment- Employment status of the students (Full-Time/Part-Time/Unemployed)

9) Salary (Continuous)- Salary earned by the students

10) Spending (Continuous)- Spending by the students

11) Computer- Computer owned by the students (Desktop/Laptop/Tablet)

12) Text Message (Continuous)- Number of text messages by the students

Basis this data, we have to compute the conditional probabilities, interpret whether Major, Grad Intention, Employment and Computer are independent of Gender or not and also see if the Salary, Spending and Text Message follows normal distribution.

	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	1	3	350	Laptop	200
1	2	Male	23	Senior	Management	Yes	3.6	Part-Time	25	1	4	360	Laptop	50
2	3	Male	21	Junior	Other	Yes	2.5	Part-Time	45	2	4	600	Laptop	200
3	4	Male	21	Junior	CIS	Yes	2.5	Full-Time	40	4	6	600	Laptop	250
4	5	Male	23	Senior	Other	Undecided	2.8	Unemployed	40	2	4	500	Laptop	100

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

2.1.1. Gender and Major

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

- Keeping 'Gender' as the row variable and the 'Majors' as the column variable, the above contingency table shows us the majors chosen by each student, gender-wise.
- Of the total 62 students, 7 have chosen Accounting as majors, 4 chose CIS, 11 chose Economics/Finance, 6 in International Business, 10 in Management, 7 for Others, 14 in Retailing/Marketing and 3 students are under undecided category.
- As we see, of the total 62 survey responses, the total number of female responses are 33 and that of male are 29.

2.1.2. Gender and Grad Intention

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

- The above contingency table shows us the 'Grad Intention' amongst Male and Female.
- Of the total of 62 students, 12 have no intention of graduating, 22 are seen as undecided and 28 students have the intention to graduate.

2.1.3. Gender and Employment

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

- The above contingency table shows us the 'Employment' status amongst Male and Female.

- Of the total of 62 students, 10 have full-time job, 43 have part-time job and 9 of them are unemployed.

2.1.4. Gender and Computer

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

- The above contingency table shows us the 'Computer' owned by the students, gender-wise.
- Of the total of 62 students, 5 have desktop, 55 own a laptop and only 2 of them own a tablet.

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. <u>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?</u>

- The Probability % that a randomly selected CMSU student being male is 46.77 %
- The Probability % that a randomly selected CMSU student being female is 53.23 %
- The above probability is calculated as:

i. P(a randomly selected CMSU student being male) = $\frac{Number\ of\ Male\ students\ in\ the\ sample}{Total\ number\ of\ students}$

ii. P(a randomly selected CMSU student being female) = $\frac{Number\ of\ Female\ students\ in\ the\ sample}{Total\ number\ of\ students}$

- The above probability clearly shows us that number of female in the survey are greater than that of male students.

2.2.2. <u>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.</u>

	Conditional Prob		Conditional Prob
P(Accounting| Male)	0.14	P(Accounting| Female)	0.09
P(CIS| Male)	0.03	P(CIS| Female)	0.09
P(Eco-Fin| Male)	0.14	P(Eco-Fin| Female)	0.21
P(Intl Bus| Male)	0.07	P(Intl Bus| Female)	0.12
P(Mgmt.| Male)	0.21	P(Mgmt.| Female)	0.12
P(Others| Male)	0.14	P(Others| Female)	0.09
P(Ret-Mkt| Male)	0.17	P(Ret-Mkt| Female)	0.27
P(Undecided| Male)	0.10	P(Undecided| Female)	0.00

The above conditional probabilities are calculated as follows:

i. P(Accounting| Male) = $\frac{P(Accounting\ \&\ Male)}{P(Male)}$

ii. P(CIS | Male) = $\frac{P(CIS\ \&\ Male)}{P(Male)}$

iii. P(Economics/Finance | Male) = $\frac{P(Eco-Fin\ \&\ Male)}{P(Male)}$

iv. P(International Business| Male) = $\frac{P(International\ Business\ \&\ Male)}{P(Male)}$

v. P(Management| Male) = $\frac{P(Management\ \&\ Male)}{P(Male)}$

vi. P(Others| Male) = $\frac{P(Others\ \&\ Male)}{P(Male)}$

vii. P(Retailing & Marketing| Male) = $\frac{P\left(\frac{Retailing}{Marketing}\ \&\ Male\right)}{P(Male)}$

viii. P(Undecided| Male) = $\frac{P(Undecided\ \&\ Male)}{P(Male)}$

And similarly, calculated the conditional probabilities for Female.

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.

	Conditional Prob		Conditional Prob
P(No intent |Male)	0.10	P(No intent |Male)	0.27
P(Intent-Yes |Male)	0.31	P(Intent-Yes |Male)	0.39
P(Undecided |Male)	0.59	P(Undecided |Male)	0.33

i. P(No intent | Male) = $\frac{P(No\ Intent\ \&\ Male)}{P(Male)}$

ii. P(Yes Intent| Male) = $\frac{P(Yes\ Intent\ \&\ Male)}{P(Male)}$

iii. P(Undecided| Male) = $\frac{P(Undecided\ \&\ Male)}{P(Male)}$

And similarly, calculated the conditional probabilities for Female.

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

	Conditional Prob		Conditional Prob
P(Full time |Male)	0.24	P(Full time |Female)	0.09
P(Part time |Male)	0.66	P(Part time |Female)	0.73
P(Unemployed |Male)	0.10	P(Unemployed |Female)	0.18

i. P(Full time| Male) = $\frac{P(Full\ time\ \&\ Male)}{P(Male)}$

ii. P(Part time| Male) = $\frac{P(Part\ time\ \&\ Male)}{P(Male)}$

iii. P(Unemployed| Male) = $\frac{P(Unemployed\ \&\ Male)}{P(Male)}$

And similarly, calculated the conditional probabilities for Female.

2.2.5. <u>Find the conditional probability of laptop preference among the male students as well as among the female students.</u>

Conditional Prob			Conditional Prob
P(Laptop |Male)	0.89	P(Laptop |Female)	0.88

i. (Laptop| Male) = $\frac{P(Laptop\ \&\ Male)}{P(Male)}$

And similarly, calculated the conditional probabilities for Female.

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.

- Accounting and Gender

Male students choosing Accounting as their major is greater than that of female. It is a general perception that male students are good at maths compared to their female counterparts and also, they at future might want to pursue Chartered Accountancy which requires travelling from place to place. Hence it is intuitive that male prefer Accounting more than female.

- CIS and Gender

Computer sciences were pre-dominantly chosen by male students but this trend in recent times are changing because of more job availability in this field. Hence we see quite a difference in CIS major in female, which is higher than that of male.

- Economics/Finance and Gender

According to the IMF article, there is a closing gap in terms of career in Finance/Economics field, as women might be better in risk management than men and more women on boards will lead to better decisions. Seeing this as a career option can explain the above probability in this case.

- International Business and Gender

Business involves lot of risk-taking. 'Risk-Taking' has gender bias, according to the Harvard Business Review article, men are less risk-averse than women under stress. But the conditional probability of International Business given female is greater than that of male is an interesting phenomenon. This part can be understood as change in risk-averse minds of the people and that female will also have potential to take risk under stress.

- Management and Gender

Management also quite a gender bias. In many companies, top managers are mostly male rather than female due to the pressure from their families, only career-oriented women could achieve higher positions in management this might influence the students in choosing their majors and explains why the probability for male is much higher for female.

- Retailing/Marketing and Gender

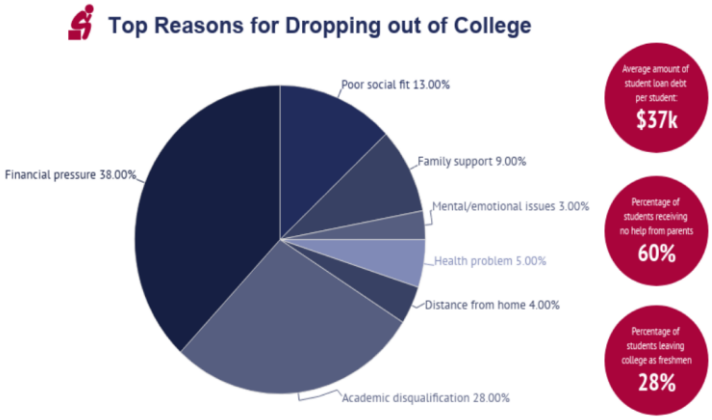
Women are blessed with attributes that make them crucial for retail. The ability to be pleasant, tolerance, compassion, stable, effectively communicate with customers makes women more suitable for specific retail roles. This explains why more female students prefer Retailing and Marketing as their major subject compared to their male counterparts.

- Undecided and Gender

As expected, the probability of 'undecided' is the least since most of the students are up-to data with the current situation in the market and they choose their majors basis that.

- Intent or No Intent to Graduate and Gender

'Intent' or 'No intent' to graduate is also closely related to 'Gender'. It is generally observed the drop-outs rate especially in developed countries like U.S. are very high and is even higher for the female population as parents or spouses become unemployed, financial aid status or eligibility change, family needs (caring for family members, pregnancy, or illness). One main consequences is that there would be a shortage of skilled workers in the working population.



- Employment and Gender

Employment and Gender is closely related to each other as well. According to the International Labour Organization report, finding a job is much tougher for women than it is for men. This explains the 'Unemployment' rate for male is lesser than for female in this case. When women are employed, they tend to work in low-quality jobs in vulnerable conditions. For women, working full-time tends to be difficult in general, which explains the lower probability in case of female than in male and hence they mostly prefer part-time jobs instead of full-time.

- Laptop Preference and Gender

Laptop preference is gender sensitive. Female students, in general, female is price sensitive and less sensitive to the features of laptop/desktop on the contrary to their male counterparts. That is why we see that laptop is preferred both by male and female students, however, probability of male over female is higher.

The above given variables (Majors, Intent to Graduate, Employment and Laptop Preferences) may directly or indirectly are linked to Gender and affect their decisions accordingly.

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]

Method-1: Shapiro-Wilk Test for Normality

The Shapiro-Wilks test for normality is one of three general normality tests designed to detect all departures from normality. It is comparable in power to the other two tests. It is more appropriate for sample less than size 50. However, they can handle sample sizes as large as 2000.

Null: The data is normally distributed

Alt: The data is not normally distributed

The null of normality is rejected if p-value is less than alpha (Here, it is 0.05).

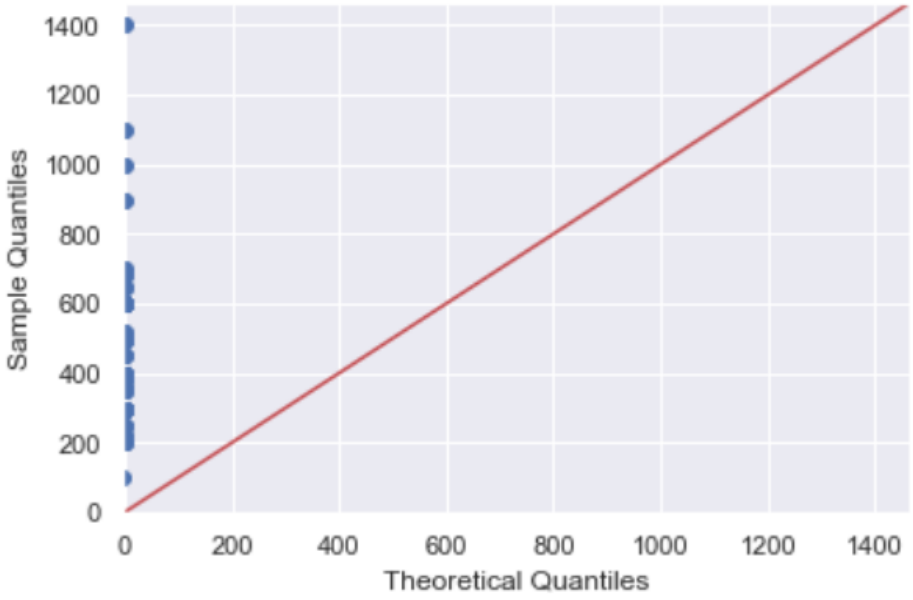
Variables	Shapiro Wilk Statistics	p-Value
Salary	0.95	0.028
Spending	0.88	0.000
Text Messages	0.86	0.000

From the above table, we see that the p-Value for Salary, Spending and Text Messages are all less than the alpha value which is 0.05. Therefore, we conclude that these three variables are not normally distributed.

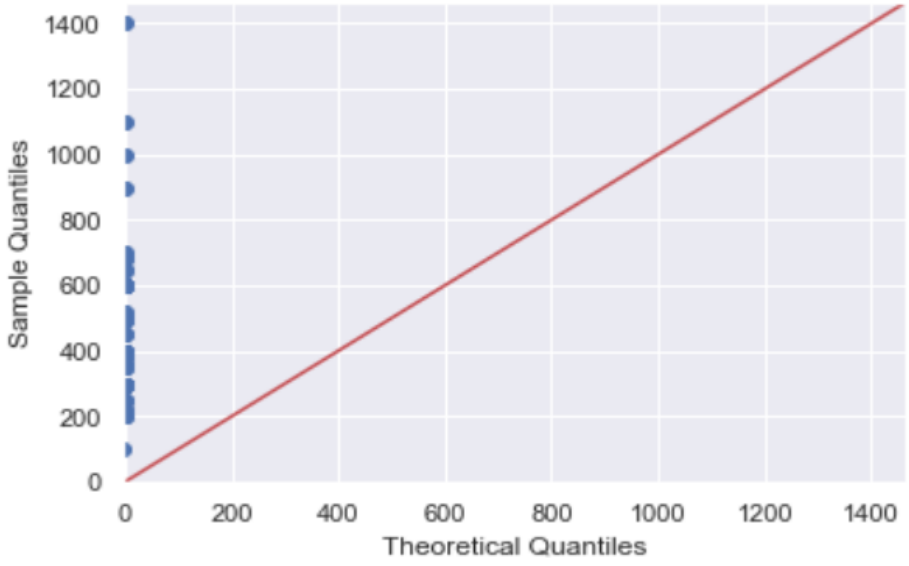
Method-2: Q-Q Plot

Q-Q Plot shows us whether the datasets come from the same distribution or not. Here, sample quantiles are plotted against theoretical quantiles. If the points are plotted around the 45° line, then we can say that the variable is normally distributed.

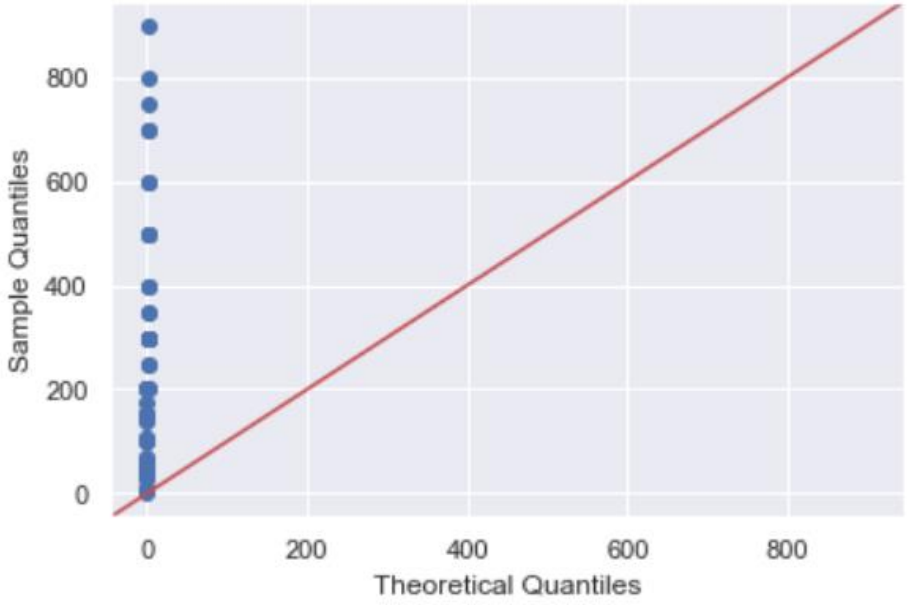
A. Q-Q PLOT FOR SALARY:



B. Q-Q PLOT FOR SPENDING:



C. Q-Q PLOT FOR TEXT MESSAGES:

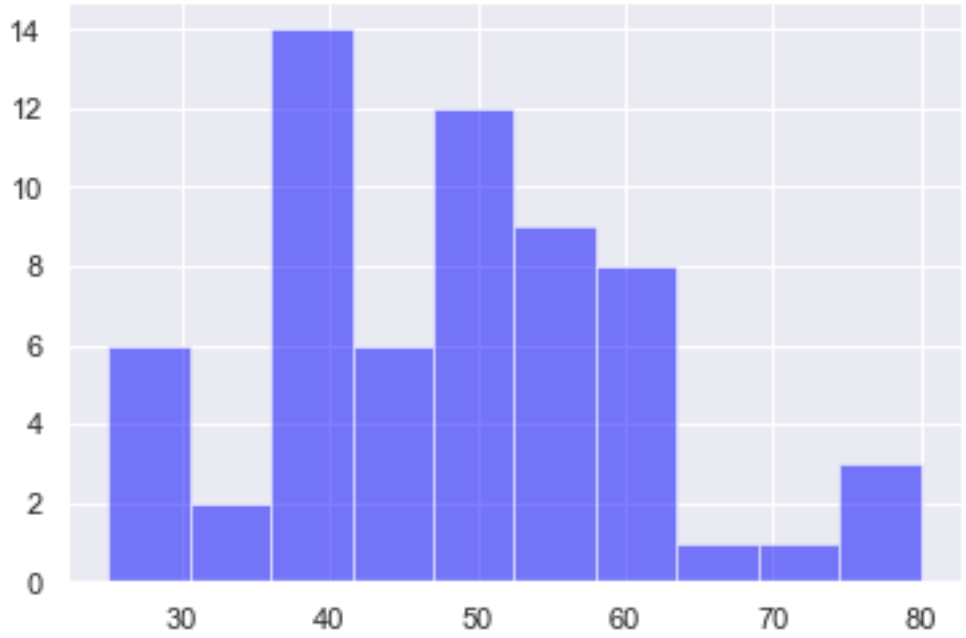


From the above plots, we can conclude that these three variables does not follow normal distribution.

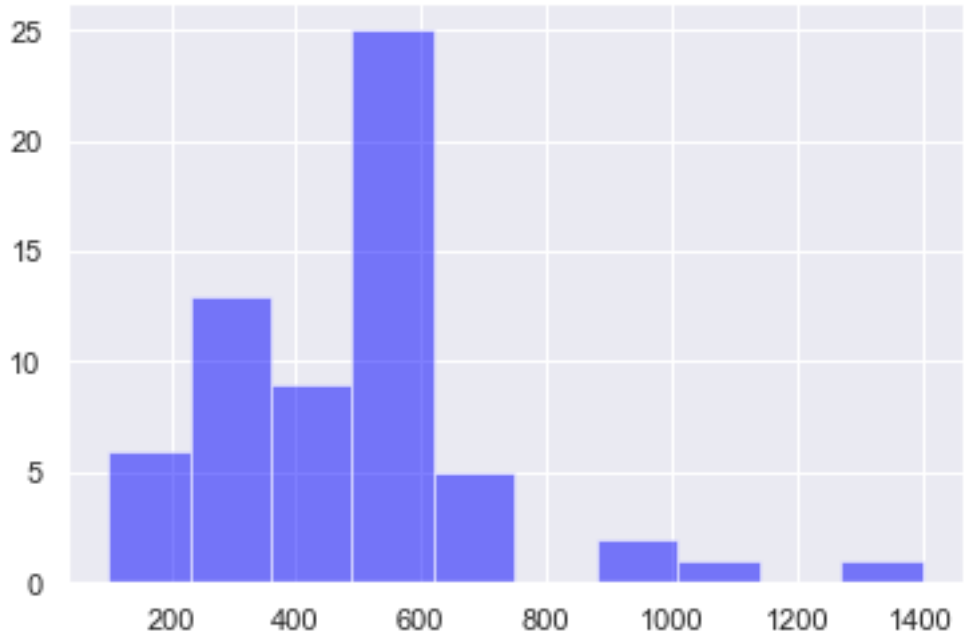
Method-3: Histogram

Histogram is commonly used to show the frequency distribution. Taller the bar for a particular range implies more data fall under that interval and vice versa.

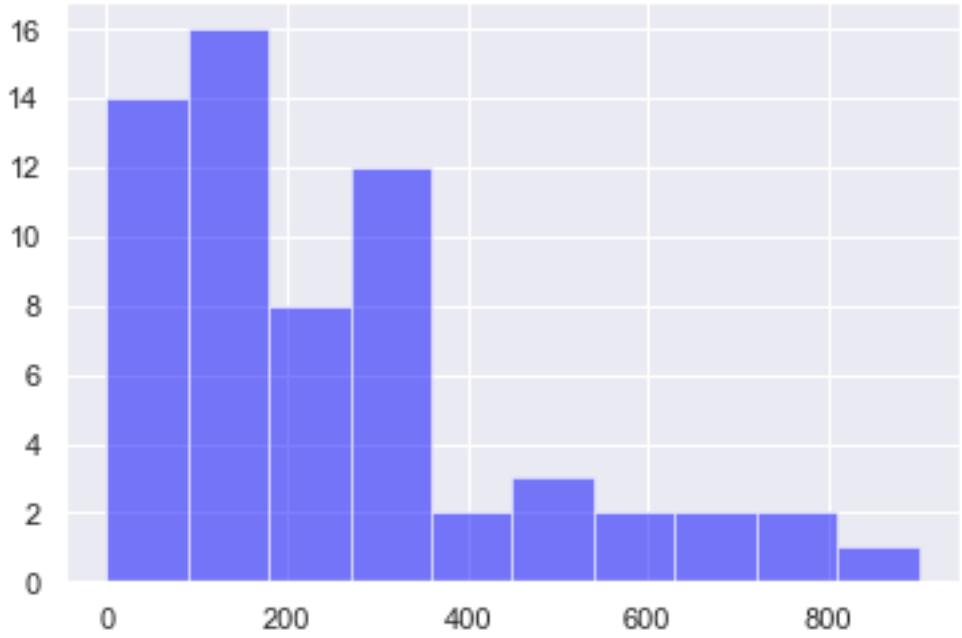
A. HISTOGRAM FOR SALARY:



B. HISTOGRAM FOR SPENDING:



C. HISTOGRAM FOR TEXT MESSAGES:



From the above three histograms for Salary, Spending and Text Messages, we can conclude that these variables do not follow Normal Distribution and they are right-skewed.
Going by 68-95-99.7, the empirical rule of Normal Distribution, the three variables under consideration does not seem to follow this rule.

From the above methods performed, we can conclude that Salary, Spending and Text Messages does not follow Normal Distribution.

Problem -3: A & B Shingles Dataset

An important quality characteristic used by the manufacturers of ABC asphalt shingles is the amount of moisture the shingles contain when they are packaged. 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. The company claims that the mean moisture content cannot be greater than 0.35 pound per 100 square feet. For the A and B shingles (each), the null and alternative hypothesis to test whether the population mean moisture content is less than 0.35 pound per 100 square feet.

	A	B
0	0.44	0.14
1	0.61	0.15
2	0.47	0.31
3	0.3	0.16
4	0.15	0.37

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.

Step-1: Hypothesis Formation

Null: μ (Shingles A) = μ (Shingles B) or μ (Shingles A) - μ (Shingles B) = 0

Alt: μ (Shingles A) ≠ μ (Shingles B) or μ (Shingles A) - μ (Shingles B) ≠ 0

Step-2: Deciding the Significance Level

By default, the alpha value is 0.05 with unknown population standard deviation

Step-3: Test Statistic

- We have two samples and we do not know the population standard deviation.
- Sample sizes for the two samples are not same.
- Here, we will use two-sample t-test (Independent Samples)

Step-4: t-Statistic, t-Critical and p-Value

The t-Statistics used is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

With degrees of freedom (as sample sizes are not the same):

$$df = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\left[\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}\right]}$$

The t-calculated value is: 1.2885

Degrees of Freedom: 63.3090

T critical value is: -1.9981

Step-5: Conclusion

- From the above, we can conclude that since t calculated > t critical, we fail to reject null and conclude that: Mu (Shingles A) = Mu (Shingles B) or Mu (Shingles A) - Mu (Shingles B) = 0
- In other words, population means of shingle A & shingle B are equal.

3.1. What assumption do you need to check before the test for equality of means is performed?

Assumption for the test for equality of means:

- Population is Normally Distributed
- Equal variance for both populations
- Sample values are independent of each other

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

To conduct the above test:

- Population is normally distributed
- Data are assumed to drawn independently from the population
- Here, n1 and n2 are greater than 30, if population distribution is not skewed then t-distribution is the preferred test in this case.

***References*:**

https://blogs.imf.org/2018/09/19/women-in-finance-an-economic-case-for-gender-equality/

https://link.springer.com/article/10.1007/BF00289552

https://www.fibre2fashion.com/industry-article/5642/women-in-retail

https://educationdata.org/college-dropout-rates/

https://www.ilo.org/infostories/en-GB/Stories/Employment/barriers-women#bridging-gap