SMDM ASSIGMENT

Submitted By

Vinit Sharma

Table of Contents

Problem 1: Wholesale Distributor **4**

Problem 1.1 ..... **4**

Problem 1.2 ..... **5**

Problem 1.3 ..... **8**

Problem 1.4 ..... **9**

Problem 1.5 ..... **9**

Problem 2: CMSU Data **10**

Problem 2.1 ..... **10**

Problem 2.2 ..... **11**

Problem 2.3 ..... **11**

Problem 2.4 ..... **12**

Problem 2.5 ..... **12**

Problem 2.6 ..... **13**

Problem 2.7 ..... **14**

Problem 2.8 ..... **14**

Problem 3: A & B Shingles **16**

Problem 3.1 ..... **16**

Problem 3.2 ..... **17**

Figures of Contents

Figure 1: Wholesale Data information 4
Figure 2 Description summary of wholesale data 4
Figure 3: Total spent analysis based on Region 5
Figure 4: Total spent analysis based on Channel 5
Figure 5 Bar chart based on Region (Left) and Channel (Right) for six varieties 5
Figure 6 Stacked Bar chart based on Region (Left) and Channel (Right) for six varieties 6
Figure 7: Pie chart for region coverage 6
Figure 8: Pie chart for channel coverage 6
Figure 9 Varieties distribution in pie chart 7
Figure 10: Correlation Matrix 7
Figure 11: Box plot analysis for outlier 8
Figure 12: Contingency table for Gender and Major 10
Figure 13: Contingency table for Gender and Grad Intention 10
Figure 14: Contingency table for Gender and Employment 10
Figure 15: Contingency table for Gender and Computer 11
Figure 16: Distribution chart for GPA (Top Left), Salary (Top Right), Spending (Bottom Left) and Text Messages (Bottom Right) 15

Problem 1: Wholesale Customers Analysis (Download Data)

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).

1.1 Use methods of descriptive statistics to summarize data. Which Region and which Channel spent the most? Which Region and which Channel spent the least?

Use methods of descriptive statistics to summarize data.

The data consists of 440 large retailers' annual spending on 6 different varieties of products. This data also includes 3 different regions (Lisbon, Oporto, Other) and two different sales channel (Hotel, Retail). There are no null values in the data.

```
<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: 31.0+ KB
```

FIGURE 1: WHOLESALE DATA INFORMATION

The data describes 440 counts for each column. Data also includes mean, std, max, min and other values for each column. Analysis also consists of a five-point analysis for the given data.

	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440.00	440.00	440.00	440.00	440.00	440.00	440.00
mean	220.50	12000.30	5796.27	7951.28	3071.93	2881.49	1524.87
std	127.16	12647.33	7380.38	9503.16	4854.67	4767.85	2820.11
min	1.00	3.00	55.00	3.00	25.00	3.00	3.00
25%	110.75	3127.75	1533.00	2153.00	742.25	256.75	408.25
50%	220.50	8504.00	3627.00	4755.50	1526.00	816.50	965.50
75%	330.25	16933.75	7190.25	10655.75	3554.25	3922.00	1820.25
max	440.00	112151.00	73498.00	92780.00	60869.00	40827.00	47943.00

FIGURE 2 DESCRIPTION SUMMARY OF WHOLESALE DATA

Total spending based on Region and Channel have been calculated for the wholesale data. Information related to most and least expense are gathered from the data

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

FIGURE 3: TOTAL SPENT ANALYSIS BASED ON REGION

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

FIGURE 4: TOTAL SPENT ANALYSIS BASED ON CHANNEL

Based on data, "Other" region (10677599) and "Hotel" channel (7999569) spent the most. Based on data, "Oporto" region (1555088) and "Retail" channel (6619931) spent the least.

1.2 There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.

Fresh varieties work better for the other and Lisbon Region. Similarly, Grocery shows better result for Oporto. On the other side, Expenditure on Delicatessen are the least for Lisbon and Oporto Region and expenditure on Frozen is the least for the other region as shown in figure below.

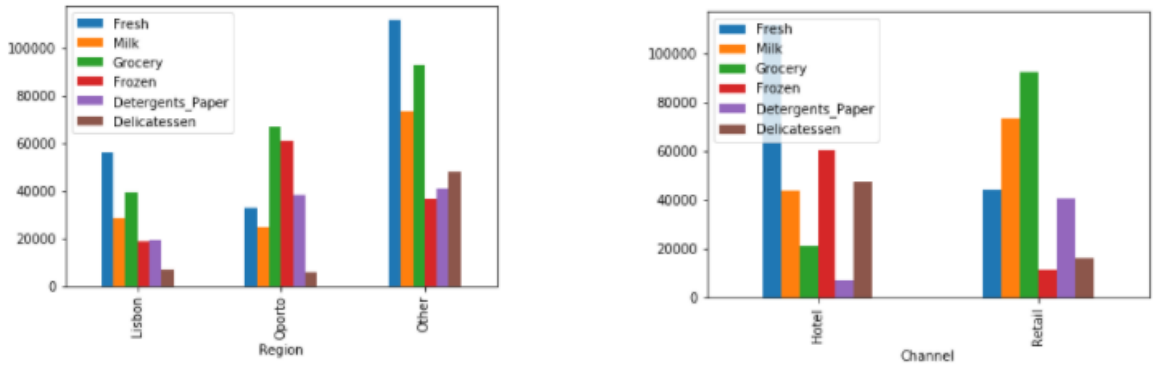

FIGURE 5 BAR CHART BASED ON REGION (LEFT) AND CHANNEL (RIGHT) FOR SIX VARIETIES

On the view for Channel, Fresh and Grocery varieties work better for the Hotel and Retail respectively. Similarly, Expense in Detergents Paper and Frozen are the least for Hotel and Retail respectively.

Stacked plot has also been drawn for region and channel based on the six varieties. For example, Other region consists of 400000 as a whole which distributes among six varieties in different proportion. Similarly, stack plot for Oporto and Lisbon have been plot in the figure. Total expense in Other region is far more than Lisbon region. Stacked plot has also been drawn for Hotel and Retail channel. In a view of channel, both are almost same.

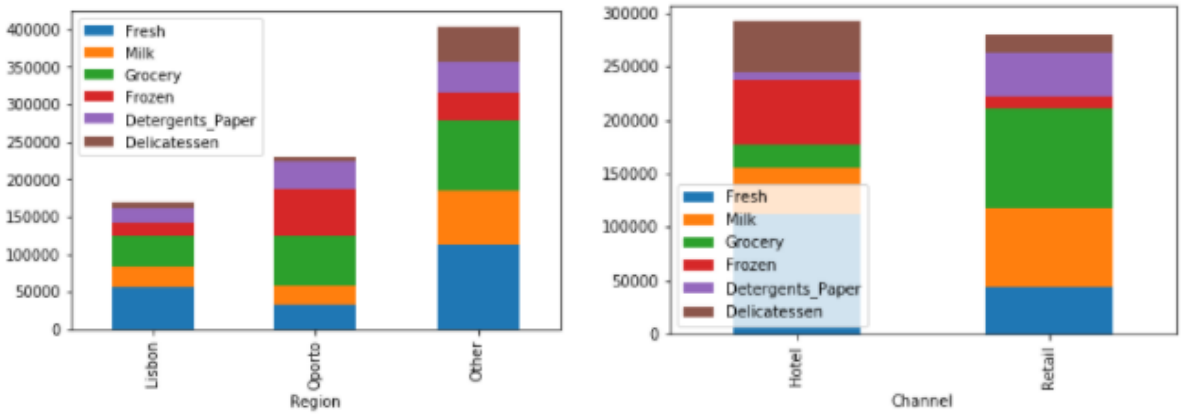

FIGURE 6 STACKED BAR CHART BASED ON REGION (LEFT) AND CHANNEL (RIGHT) FOR SIX VARIETIES

Pie chart for the region clearly shows that "Other" distribution is almost 75% and rest Oporto and Lisbon covers the area of rest 25% of the pie.

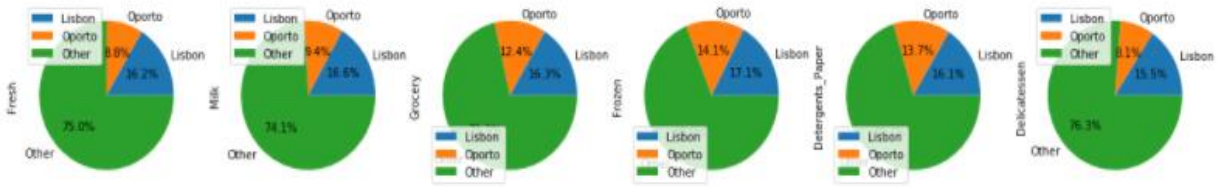

FIGURE 7: PIE CHART FOR REGION COVERAGE

Similarly, Pie chart for the Channel clearly shows that "Fresh, Frozen, Delicatessen" varieties on Hotel covers more than 50% of the pie chart and "Milk, Grocery and Detergents paper" varieties on Retail covers more than 50% of the pie chart.

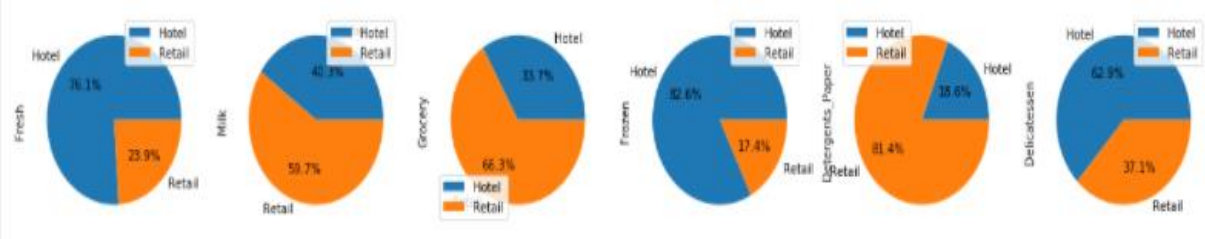

FIGURE 8: PIE CHART FOR CHANNEL COVERAGE

Varieties distribution pie chart has been created based on spent for each variety, Fresh covers the major portion which is 36.1%. Grocery and Milk are next in the list with 23.9 and 17.4 % respectively. Delicatessen covers the least portion in the chart with 4.0%. Rest Detergents and Frozen are distribute with 8.7% and 9.2% respectively.

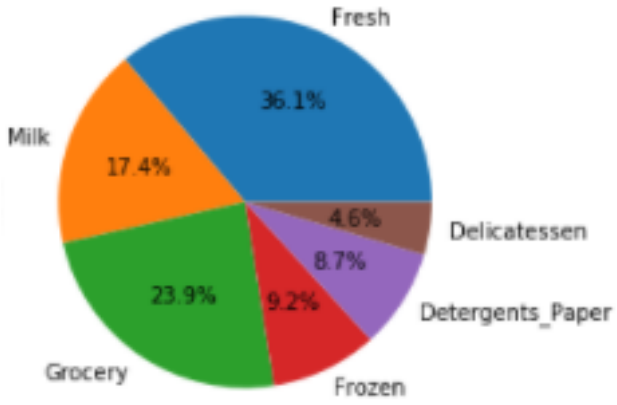

FIGURE 9 VARIETIES DISTRIBUTION IN PIE CHART

Below observation shows the correlation among the six varieties. Grocery and detergents paper are highly correlated with 0.91. In a next, Grocery and milk are also heavily correlated with 0.73 as value. Here, we have some negative correlated observation as well for example fresh and detergent paper.

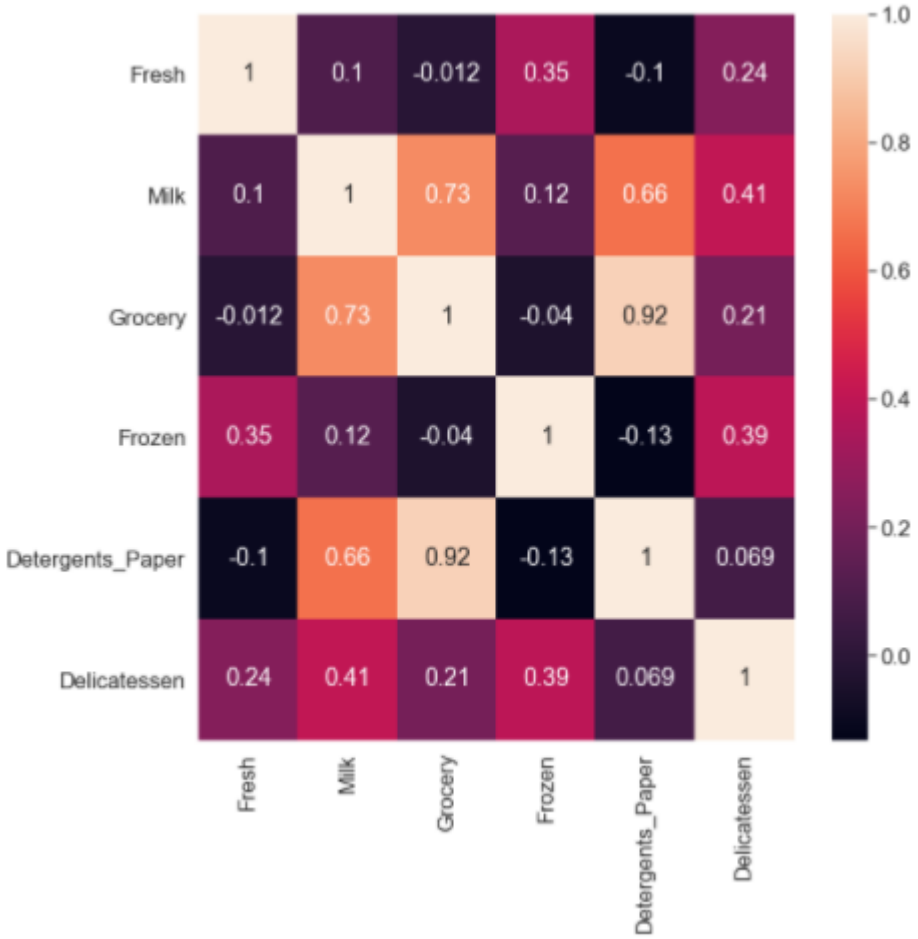

FIGURE 10: CORRELATION MATRIX

1.3 On the basis of a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?

We have calculated the measure of variability and its result has been shown below.

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

This result shows that coff. of variance for Fresh is 105.25% and for Delicatessen is 184.94%. Hence, Fresh product reflects the least inconsistent and Delicatessen shows the most inconsistent behaviour.

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

All the six variables ('Fresh, Milk, Grocery, Frozen, detergent paper, **Delicatessen**) have outliers. As per below box plot, we have identified in each variety many points lying outside the whisker of the boxplot.

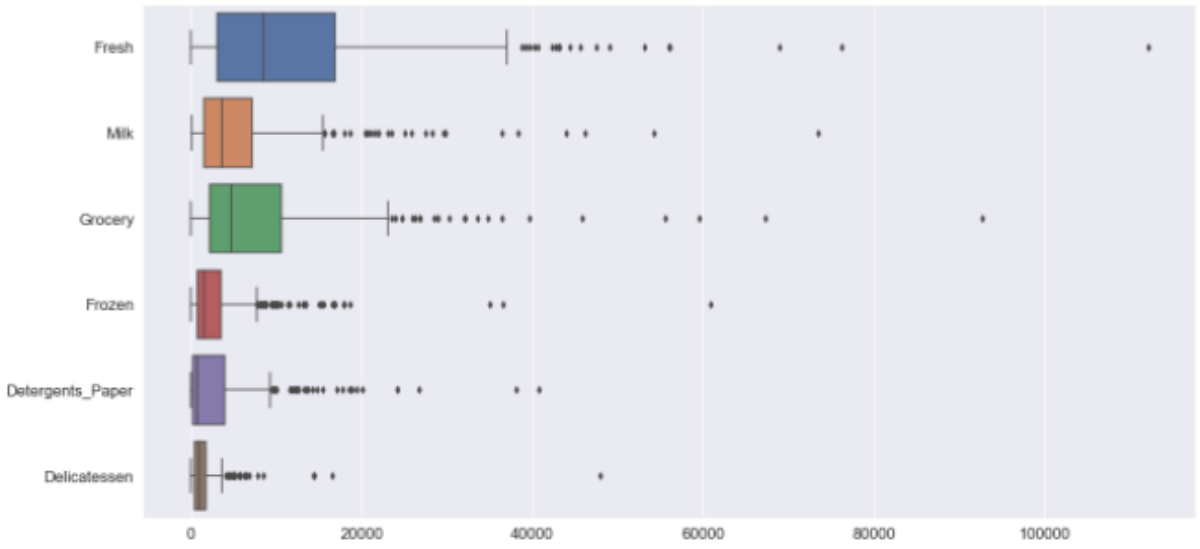

FIGURE 11: BOX PLOT ANALYSIS FOR OUTLIER

1.5 On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem? Answer from the business perspective

Based on wholesale distribution analysis, Hotel in channel and Other in region should be the main focus for the business. Fresh variety contributes the most in “other” region and “Hotel” channel. Business unit can invest more to provide "fresh" varieties in abundance.

Grocery is highly correlated with detergents and milk so business can utilize this correlation for their profit. Whenever grocery items are selected milk and detergents items also more favourable to get purchase. To provide attractive sale offer, business can use this correlation.

Based on observation, Fresh have the least inconsistency in the data which reflects investment in the fresh items would be a good option.

Problem 2 - Clear Mountain State University (CMSU) (Download Data)

The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. CMSU creates and distributes a survey of 14 questions and receives responses from 62 undergraduates (stored in the ***Survey*** data set).

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

Below are the contingency table between different variables.

2.1.1. Gender and Major

Major Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	Total
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

FIGURE 12: CONTINGENCY TABLE FOR GENDER AND MAJOR

2.1.2. Gender and Grad Intention

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

FIGURE 13: CONTINGENCY TABLE FOR GENDER AND GRAD INTENTION

2.1.3. Gender and Employment

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

FIGURE 14: CONTINGENCY TABLE FOR GENDER AND EMPLOYMENT

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

FIGURE 15: CONTINGENCY TABLE FOR GENDER AND COMPUTER

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

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

Probability of male student = Total male student/ total student

= 29/62

= 0.4677

Probability of a randomly selected CMSU student will be male is 46.77%.

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

Probability of female student = Total female student/ total student

= 33/62

= 0.5322

Probability of a randomly selected CMSU student will be female is 53.22%.

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

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

Following results shows the conditional probability of different majors among the male students in CMSU.

```
Probability for  Accounting ----> 13.793103448275861 %
Probability for  CIS ----> 3.4482758620689653 %
Probability for  Economics/Finance ----> 13.793103448275861 %
Probability for  International Business ----> 6.896551724137931 %
Probability for  Management ----> 20.689655172413794 %
Probability for  Other ----> 13.793103448275861 %
Probability for  Retailing/Marketing ----> 17.24137931034483 %
Probability for  Undecided ----> 10.344827586206897 %
```

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

Following results shows the conditional probability of different majors among the female students in CMSU.

```
Probability for  Accounting ----> 9.090909090909092 %
Probability for  CIS ----> 9.090909090909092 %
Probability for  Economics/Finance ----> 21.21212121212121 %
Probability for  International Business ----> 12.121212121212121 %
Probability for  Management ----> 12.121212121212121 %
Probability for  Other ----> 9.090909090909092 %
Probability for  Retailing/Marketing ----> 27.27272727272727 %
Probability for  Undecided ----> 0.0 %
```

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

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

Probability That a randomly chosen student is a male and intends to graduate = no. of male which intends to grad. /total no. of students

= 17/62

= 0.274193

Probability That a randomly chosen student is a male and intends to graduate is 27.41%.

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

Probability That a randomly chosen student is a female and does not have a laptop= no. of female which does not have a laptop /total no. of students

= (33-29)/62

= 0.0645161

Probability That a randomly chosen student is a female and does not have a laptop is 6.45%.

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

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

Probability of student is a male = 29/62

Probability of student has full time employment = 10/62

Probability of student is a male and employment = 7/62

Probability That a randomly chosen student is a male or has full time employment=

= (29/62)+(10/62)-(7/62)

= (32/62)

=0.516129

Probability That a randomly chosen student is a male or has full time employment is 51.61%.

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

Conditional probability that given a female student is randomly chosen, she is majoring in international business or management = P(IB|F)+P(M|F)

= 4/33+4/33

= 8/33

=0.2424242424

Conditional probability that given a female student is randomly chosen, she is majoring in international business or management is 24.24%.

2.6. Construct a contingency table of Gender and Intent to Graduate at 2 levels (Yes/No). The Undecided students are not considered now and the table is a 2x2 table. Do you think the graduate intention and being female are independent events?

Following table shows a contingency table of gender and intent to graduate at 2 level without undecided students.

Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

Condition for event A and B will be independent event when,

$$P(A \cap B) = P(A) \cdot P(B)$$

So, Probability of graduation intention= 28/40

Probability of female = 20/40

Probability of graduation intention and female = 11/40

Probability of graduation intention and female= Probability of female* Probability of graduation intention

- ➔ (11/40) != (28/40)*(20/40)
- ➔ 0.275 != 0.35 (!= means “NOT EQUAL TO”)

This means **the graduate intention and being female are independent events are dependent events.**

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

Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

Answer the following questions based on the data

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

Probability (GPA less than 3) = 17/62

= 0.27419354

Probability of GPA less than 3 is 27.41%.

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

Salary	50.0	52.0	54.0	55.0	60.0	65.0	70.0	78.0	80.0	Total
Gender										
Female	5	0	0	5	5	0	1	1	1	18
Male	4	1	1	3	3	1	0	0	1	14
Total	9	1	1	8	8	1	1	1	2	32

Conditional Probability for selected male earns 50 or more = total of selected male earns 50 or more/total male

=14/29

=0.482758620

Conditional Probability for selected male earns 50 or more is 48.27%.

Conditional Probability for selected female earns 50 or more = total of selected female earns 50 or more/total female

=18/33

=0.54545454

Conditional Probability for selected female earns 50 or more is 54.54%.

2.8. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarizing your conclusions.

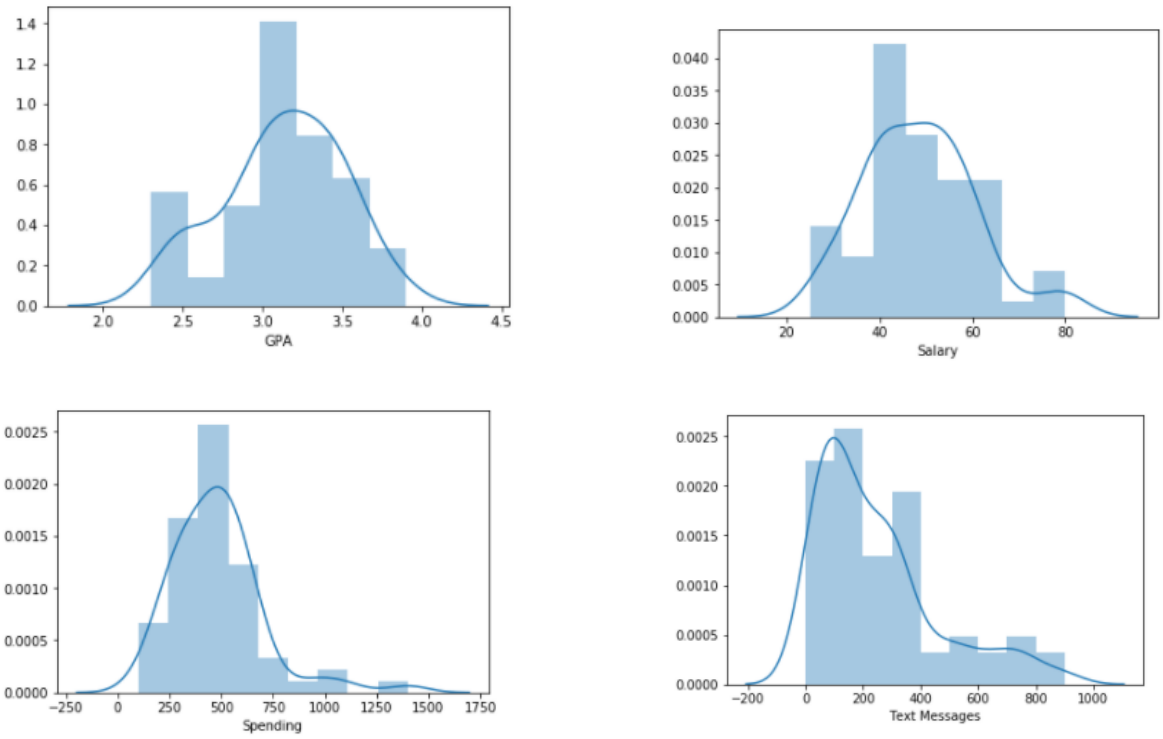

FIGURE 16: DISTRIBUTION CHART FOR GPA (TOP LEFT), SALARY (TOP RIGHT), SPENDING (BOTTOM LEFT) AND TEXT MESSAGES (BOTTOM RIGHT)

To test whether the data variables normally distribute or not, we use Shapiro Wilk test on different variables.

Shows the result of P value:

- GPA - `0.11204058676`
- Salary - `0.028000`
- Spending - `1.6854661225806922e-05`
- Text Messages - `4.324040673964191e-06`

`Based on result, GPA and Salary data are normal` distributive (P value more than 0.05).
Similarly, Spending and Text Messages are not normal distributive (P value less than 0.05).

Similarly, check for the skewness of data:

```
GPA             -0.31
Salary           0.53
Spending         1.59
Text Messages    1.30
dtype: float64
```

Based on skewness result,

GPA is more towards the left tail distributed. In a same way, Salary, Spending and text messages are right tail distributed.

Problem 3: A & B shingles (Download Data)

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 are calculated. The company would like to show that the mean moisture content is less than 0.35 pounds 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.

3.1 Do you think there is evidence that means moisture contents in both types of shingles are within the permissible limits? State your conclusions clearly showing all steps.

For Sample A:

Step 1: Define null and alternative hypotheses

H0 : mean of moisture content = 0.35

H1 : mean of moisture content < 0.35

Step 2: Decide the significance level

Here we select α = 0.05.

Step 3: Identify the test statistic

we use the t distribution and the tST test statistic

Step 4: Calculate the p - value and test statistic

scipy.stats.ttest_1samp calculates the t test for the mean of one sample given the sample observations and the expected value in the null hypothesis. This function returns t statistic and the two-tailed p value.

```
One sample t test
t statistic: -1.4735046253382782 p value: 0.14955266289815025
```

Step 5: Decide to reject or accept null hypothesis

```
Level of significance: 0.05
We have no evidence to reject the null hypothesis since p value > Level of significance
Our one-sample t-test p-value= 0.14955266289815025
```

For Sample B:

Step 1: Define null and alternative hypotheses

H0 : mean of moisture content = 0.35

H1 : mean of moisture content < 0.35

Step 2: Decide the significance level

Here we select α = 0.05.

Step 3: Identify the test statistic

we use the t distribution and the tST test statistic

Step 4: Calculate the p - value and test statistic

scipy.stats.ttest_1samp calculates the t test for the mean of one sample given the sample observations and the expected value in the null hypothesis. This function returns t statistic and the two-tailed p value.

```
One sample t test
t statistic: -3.0405827185417524 p value: 0.00445138286795499
```

Step 5: Decide to reject or accept null hypothesis

```
Level of significance: 0.05
We have evidence to reject the null hypothesis since p value < Level of significance
Our one-sample t-test p-value= 0.00445138286795499
```

3.2 Do you think that the population mean for shingles A and B are equal? Form the hypothesis and conduct the test of the hypothesis. What assumption do you need to check before the test for equality of means is performed?

Step 1: Define null and alternative hypotheses

- H_0: μA- μB = 0 i.e μA = μB
- H_A: μA - μB ≠ 0 i.e $\mu A \neq \mu B$

Step 2: Decide the significance level

Here we select α = 0.05.

Step 3: Identify the test statistic

we use the t distribution and the tST test statistic for two sample

Step 4: Calculate the p - value and test statistic

```
tstat  0.905
p-value for one-tail: 0.18586475076401154
```

Step 5: Decide to reject or accept null hypothesis

```
Paired two-sample t-test p-value= 0.18586475076401154
We do not have enough evidence to reject the null hypothesis in favour of alternative hypothesis
```