SMDM PROJECT BUSSINESS REPORT

Kratik Mehta

PGP-DSBA Feb 2022

Table of Contents

List of Figures

List of Tables

Problem 1

Executive Summary

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). In this problem statement, we will perform some basic exploratory data analysis on the annual spending's of different items across different regions and channels.

Sample of the Wholesale Customers Dataset

	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

Table 1: Wholesale Customers Dataset Sample

Checking the types of variables in the dataset.

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

From the above output we can see that:

- There are 440 observations of different buyers in the data.
- There are 9 variables, out of which, 7 are of integer type and 2 are of object(categorical) type.
- Out of the 7 integer type variables, **Buyer/Spender** is *discrete numerical variable* while the others are *continuous numerical variables*.
- The dataset does not have any missing values.

Checking the distributions of the continuous variables.

From the below plot, it can be seen that the data for the continuous variables is not normally distributed.

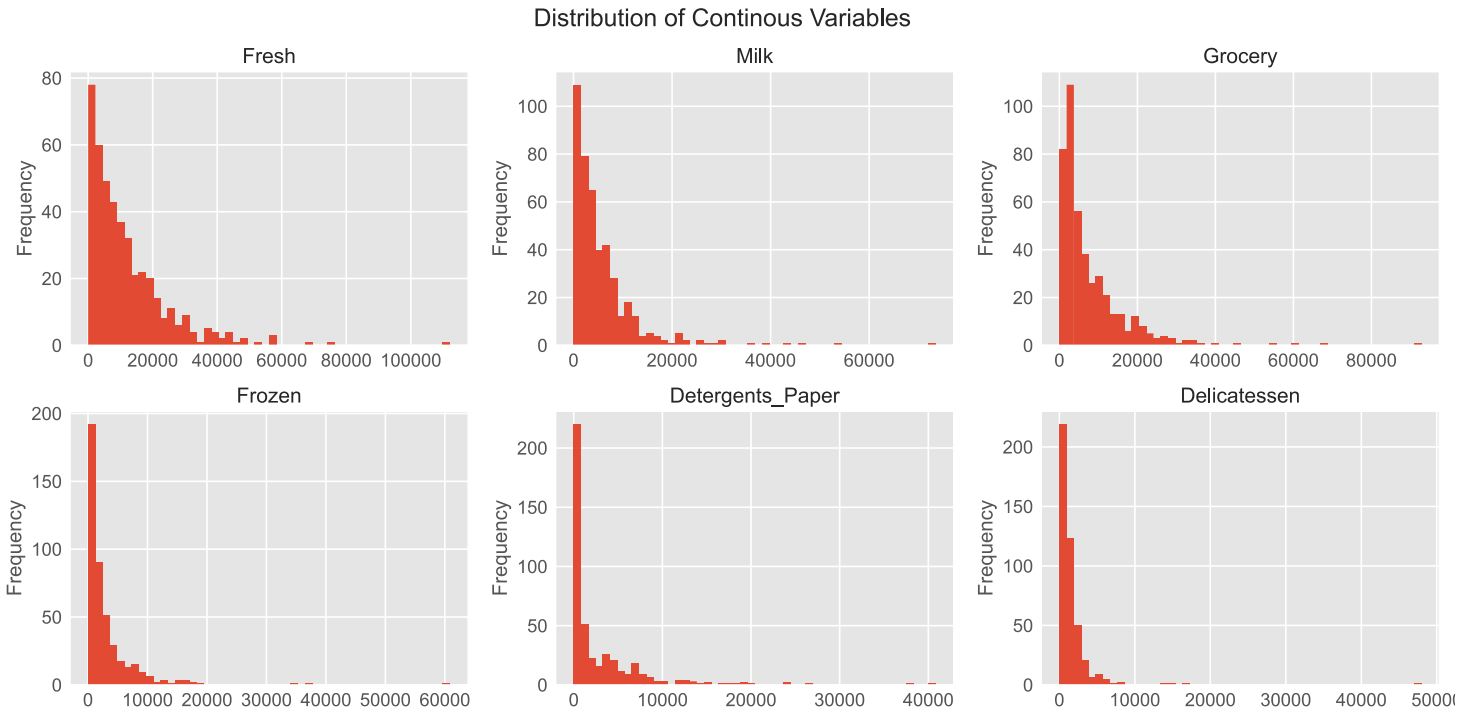

Figure 1: Distributions of the continuous variables in Wholesale Dataset

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?

Descriptive Statistics of the Wholesale Customers Dataset

Descriptive Statistics tells us a brief summary of the dataset. It can include the 5-point summary for the numerical variables and frequency distribution of categorical variables.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Buyer/Spender	440.0	NaN	NaN	NaN	220.5	127.161315	1.0	110.75	220.5	330.25	440.0
Channel	440	2	Hotel	298	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Region	440	3	Other	316	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Fresh	440.0	NaN	NaN	NaN	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0
Milk	440.0	NaN	NaN	NaN	5796.265909	7380.377175	55.0	1533.0	3627.0	7190.25	73498.0
Grocery	440.0	NaN	NaN	NaN	7951.277273	9503.162829	3.0	2153.0	4755.5	10655.75	92780.0
Frozen	440.0	NaN	NaN	NaN	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0
Detergents_Paper	440.0	NaN	NaN	NaN	2881.493182	4767.854448	3.0	256.75	816.5	3922.0	40827.0
Delicatessen	440.0	NaN	NaN	NaN	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0

Table 2: Descriptive Statistics of the Wholesale Customers Dataset

From the above output we can see that:

- The **Channel** variable has 2 unique values, with *Hotel* being the most frequent, occurring 298 times.
- The **Region** variable has 3 unique values, with *Other* being the most frequent, occurring 316 times.
- The **Fresh** items have the highest average annual spending in the dataset.
- The maximum annual spending values of all the items are quite high. These values might be outliers.
- The standard deviations of the annual spending are also large. This indicates that the data is spread over a big range.

NaN shows that the values cannot be calculated for that particular variable. Like we cannot calculate mean for a categorical variable. And in a same way unique value cannot be calculated for a numerical variable.

Region wise Total Annual Spending's for the items

Region	Total_Spendings
Oporto	1555088
Lisbon	2386813
Other	10677599

Table 3: Region wise Total Annual Spending's for the items

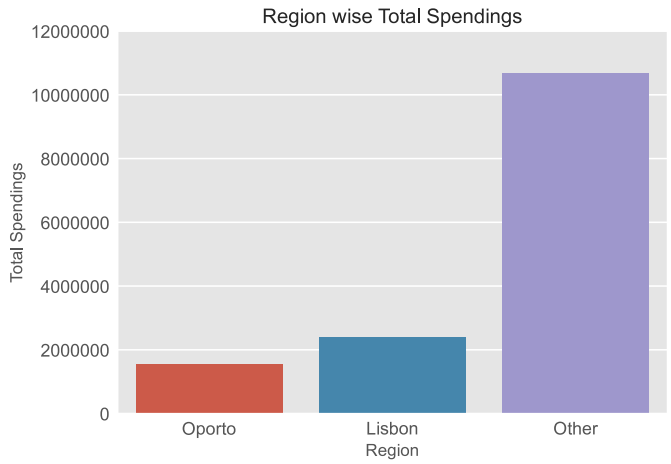

Figure 2: Region wise Total Annual Spending's

After calculating the Region wise Total Spending's, we found that:

- The buyers spent the *least* in the *Oporto* region.
- The buyers spent the *most* in the *Other* regions.

Channel wise Total Annual Spending's for the items

Channel	Total_Spendings
Retail	6619931
Hotel	7999569

Table 4: Channel wise Total Annual Spending's for the items

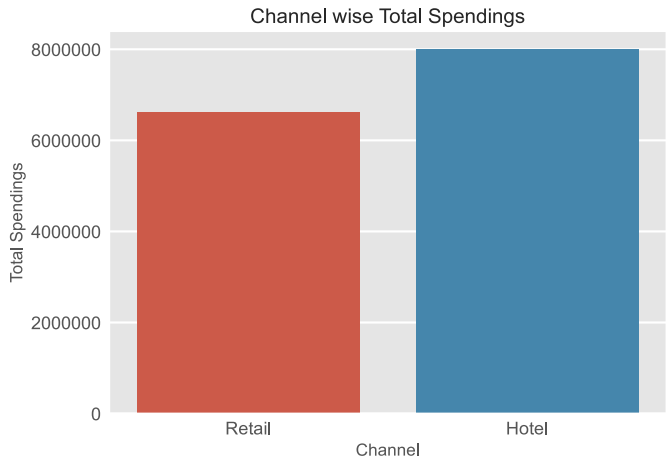

Figure 3: Channel wise Total Annual Spending's

After calculating the Channel wise Total Spending's, we found that:

- The buyers spent the *least* in the *Retail* channels.
- The buyers spent the *most* in the *Hotel* channels.

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.

Region wise Average spending for different variety of Items.

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

Table 5: Region wise Average spending for different variety of Items

Region wise Coefficient of Variation of spending for different variety of Items.

Coefficient of Variations is a measure of relative dispersion around the mean. It is used to compare variables with different ranges and units of measurements.

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Region						
Lisbon	1.041049	1.039815	1.147670	1.030599	1.587430	0.993008
Oporto	0.848318	1.145076	1.176182	2.262291	1.766718	0.906043
Other	1.068277	1.327648	1.207808	1.446761	1.630040	1.994680

Table 6: Region wise Coefficient of Variation of spending for different variety of Items

From the above tables we can conclude that:

- For all the Regions in Portugal: -
  - Highest Averages spending's are for *Fresh* items.
  - Lowest Averages spending's are for *Delicatessen* items. This means that people spend less on expensive food stores.
  - In general, spending's for all the items are largely spread out around the mean.
- In Lisbon: -
  - The spending's for **Detergents_Paper** is highly spread out about its mean in comparison to other items. This could mean that the actual spending's for **Detergents_Paper** are away from its mean and near the extreme values.
- In Oporto: -
  - The spending's for **Frozen** items is highly spread out about its mean in comparison to other items. This could mean that the actual spending's for **Frozen** are away from its mean and near the extreme values.
- In Other regions: -
  - The spending's for **Delicatessen** is highly spread out about its mean in comparison to other items. This could mean that the actual spending's for **Delicatessen** are away from its mean and near the extreme values.

Channel wise Average spending's for different variety of Items.

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

Table 7: Channel wise Average spending's for different variety of Items

Channel wise Coefficient of Variation of spending's for different variety of Items.

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Channel						
Hotel	1.026428	1.260867	0.894849	1.505745	1.396596	2.222828
Retail	1.009365	0.903246	0.751543	1.096932	0.865408	1.114267

Table 8: Channel wise Coefficient of Variation of spending's for different variety of Items

From the above tables we can conclude that:

- For all the Channels in Portugal: -
  - In general, spending's for all the items are largely spread out around the mean.
- For Hotels in Portugal: -
  - Highest Averages spending's are for *Fresh* items.
  - Lowest Averages spending's are for *Detergents_Paper*.
  - The spending's for **Delicatessen** is highly spread out about its mean in comparison to other items. This could mean that most of the values are less than the mean while only few are more than the mean.
- For Retail outlets in Portugal: -
  - Highest Averages spending's are for *Grocery* items.
  - Lowest Averages spending's are for *Frozen* items.
  - The spending's for **Delicatessen** items is highly spread out about its mean in comparison to other items. This could mean that most of the values are less than the mean while only few are more than the mean.

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?

Using the Coefficient of Variation (CV) as a measure of variability, as the means of all the items are highly different.

```
Fresh              1.053918
Milk               1.273299
Grocery            1.195174
Frozen             1.580332
Detergents_Paper   1.654647
Delicatessen       1.849407
dtype: float64
```

From the above table we can see that:

- *Delicatessen* items show the most inconsistent behaviour as its CV is very high.
- *Fresh* items show the least inconsistent behaviour as its CV is very low.

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

Box Plot is a very effective visualization to detect outliers. It also shows the five point summary of the variables, which helps us to identify the distribution of the variables.

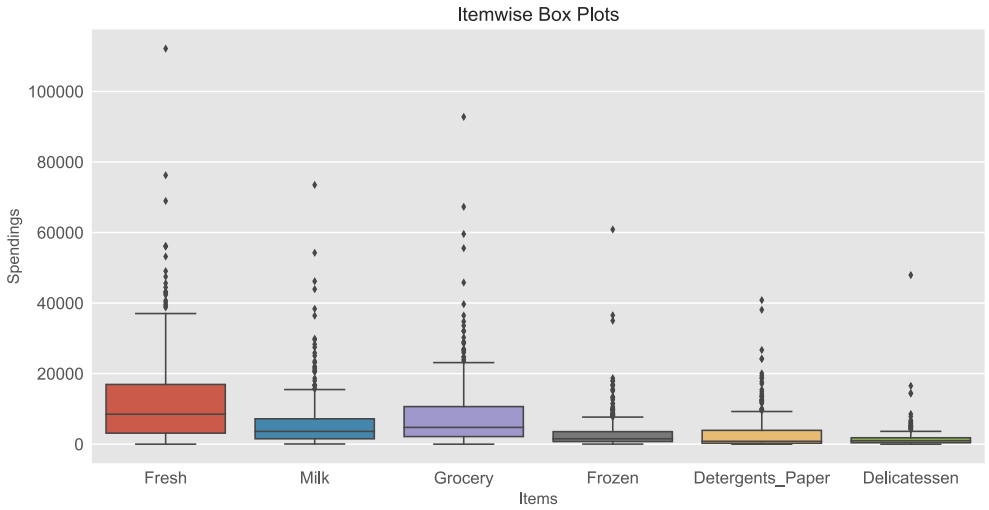

Figure 4: Item wise Box plots for Wholesale Dataset

From the above box plot we can see that the spending's for all the items in the dataset have outliers. The plot also confirms that the spending's are Positively skewed as the spread of data above the median is more than spread of data below the median. These outliers can be buyers who buy in large quantities at once or buyers who buy expensive items.

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

- From the above histograms we can see that, most of the buyers spend less for all items. So, the wholesaler can keep more stock of reasonable items as compared to expensive items to meet demand expectations.
- **Fresh** items are the most popular items bought by the customers. Keeping higher stock of the items might help increase sales and meet demands.
- **Delicatessen** items are the least bought items. Reducing the stock of these items will help decrease inventory costs.
- **Oporto** region has the lowest annual spending's. Increasing the stock of **Fresh** and **Grocery** items in this region might help increase sales.
- In **Lisbon** region, **Detergent_Paper** and **Delicatessen** items have the lowest sales. Reducing their stocks will help decrease inventory.
- The **Other** regions combined have the highest sales. Increasing the customer base and keeping the prices reasonable can help increase sales and profit in this regions.
- For **Hotel** channels, providing offers on **Detergent_Paper** and **Delicatessen** can help increase sales of these items. Also keeping high stocks of **Fresh** items will help meet demand expectations.
- For **Retail** channels, providing offers on **Frozen** and **Delicatessen** can help increase sales of these items. Also keeping high stocks of **Milk** and **Grocery** items will help meet demand expectations.

Problem 2

Executive Summary

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). In this problem statement, we will find various probabilities of different attributes of the students in the data set.

Sample of the Survey Dataset

	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

Table 9: Survey Dataset Sample

Checking the types of variables in the dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62 entries, 0 to 61
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   ID                 62 non-null     int64
 1   Gender             62 non-null     object
 2   Age                62 non-null     int64
 3   Class              62 non-null     object
 4   Major              62 non-null     object
 5   Grad Intention     62 non-null     object
 6   GPA                62 non-null     float64
 7   Employment         62 non-null     object
 8   Salary             62 non-null     float64
 9   Social Networking  62 non-null     int64
 10  Satisfaction       62 non-null     int64
 11  Spending           62 non-null     int64
 12  Computer           62 non-null     object
 13  Text Messages      62 non-null     int64
dtypes: float64(2), int64(6), object(6)
memory usage: 6.9+ KB
```

From the above output we can see that:

- There are 62 observations from different students in the data.
- There are 14 variables, out of which, 6 are of integer type, 2 are of float type and 6 are of object(categorical) type.
- The dataset does not have any missing values.

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	Management	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

Table 10: Contingency Table for Gender and Major

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

Table 11: Contingency Table for Gender and Grad Intention

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

Table 12: 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

Table 13: 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?

Out of the total 62 students, 29 students are male. Therefore,

$$P(Male) = \frac{n(Male)}{n(Total\ Students)} = \frac{29}{62} = 0.4677$$

Hence the probability that a randomly selected CMSU student will be male is 0.468.

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

Out of the total 62 students, 33 students are female. Therefore,

$$P(Female) = \frac{n(Female)}{n(Total\ Students)} = \frac{33}{62} = 0.5323$$

Hence the probability that a randomly selected CMSU student will be female is 0.532.

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.

$$P(\ Major \mid Male\ ) = \frac{P(Major \cap Male)}{P(Male)}$$

Conditional probability that a Male student has an Accounting major is 0.138.
Conditional probability that a Male student has a CIS major is 0.034.

```
Conditional probability that a Male student has an Economics/Finance major is 0.138.
Conditional probability that a Male student has an International Business major is 0.069.
Conditional probability that a Male student has a Management major is 0.207.
Conditional probability that a Male student has Other major is 0.138.
Conditional probability that a Male student has a Retailing/Marketing major is 0.172.
Conditional probability that a Male student has an Undecided major is 0.103.
```

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

$$P(\,Major \mid Female\,) = \frac{P(Major \cap Female)}{P(Female)}$$

```
Conditional probability that a Female student has an Accounting major is 0.091.
Conditional probability that a Female student has a CIS major is 0.091.
Conditional probability that a Female student has an Economics/Finance major is 0.212.
Conditional probability that a Female student has an International Business major is 0.121.
Conditional probability that a Female student has a Management major is 0.121.
Conditional probability that a Female student has Other major is 0.091.
Conditional probability that a Female student has a Retailing/Marketing major is 0.273.
Conditional probability that a Female student has an Undecided major is 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.

Out of the 62 students, 17 students are male who intends to graduate. Therefore,

$$P(Male\ and\ intends\ to\ graduate) = \frac{P(Male \cap Intends\ to\ Graduate)}{P(Total\ Students)} = \frac{17}{62} = 0.2742$$

```
Hence the probability that a randomly chosen student is a male and intends to graduate is
0.274.
```

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

Out of the 62 students, 4 students are female who do not have a laptop. Therefore,

$$P(Female\ and\ do\ not\ have\ a\ laptop) = \frac{P(Female \cap No\ Laptop)}{P(Total\ Students)} = \frac{2\ +\ 2}{62} = 0.0645$$

```
Hence the probability that a randomly chosen student is a female and does NOT have a laptop is
0.065.
```

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?

$$\begin{aligned} &P(Male \cup Full - Time\ Employment) \\ &\quad = P(Male) + P(Full - Time\ Employment) - P(Male \cap Full - Time\ Employment) \\ &\quad = \frac{29\ +\ 10\ -\ 7}{62} = 0.5161 \end{aligned}$$

Out of the 62 students, 29 are male, 10 have full time jobs, and 7 are both male and have full time jobs.
Hence the probability that a randomly chosen student is a male or has full-time employment is 0.516.

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

Out of the 33 female students, 4 are majoring in International Business and 4 are majoring in Management. Therefore for mutually exclusive events:

$$P(\mathit{International\ Business} \cup \mathit{Management} \mid \mathit{Female})$$
$$= P(\mathit{International\ Business} \mid \mathit{Female}) + P(\mathit{Management} \mid \mathit{Female}) = \frac{4+4}{33}$$
$$= 0.2424$$

Hence the conditional probability that given a female student is randomly chosen, she is majoring in international business or management is 0.242.

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?

Grad Intention	No	Yes	Total
Gender			
Female	9	11	20
Male	3	17	20
Total	12	28	40

Table 14: Contingency table for Gender and Intent to Graduate at 2 levels

Out of the 40 students, 20 are female, 28 intent to graduate, and 11 are both female and intent to graduate. Therefor for independent events:

$$P(Female \cap Intends\ to\ Graduate) = P(Female)P(Intends\ to\ Graduate)$$

$$P(Female \cap Intends\ to\ Graduate) = \frac{11}{40} = 0.275$$

$$P(Female)P(Intends\ to\ Graduate) = \frac{20}{40} * \frac{28}{40} = 0.35$$

Hence using the above-mentioned formula, we find that the graduate intention and being female are not independent events.

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

Out of the 62 students, 17 students have GPA less than 3.
Hence if a student is chosen randomly, the probability that his/her GPA is less than 3 is 0.274.

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.

```
Out of the 29 male students, 14 students have salaries more than or equal to 50.
Hence the conditional probability that a randomly selected male earns 50 or more is 0.483.

Out of the 33 female students, 18 students have salaries more than or equal to 50.
Hence the conditional probability that a randomly selected female earns 50 or more is 0.545.
```

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.

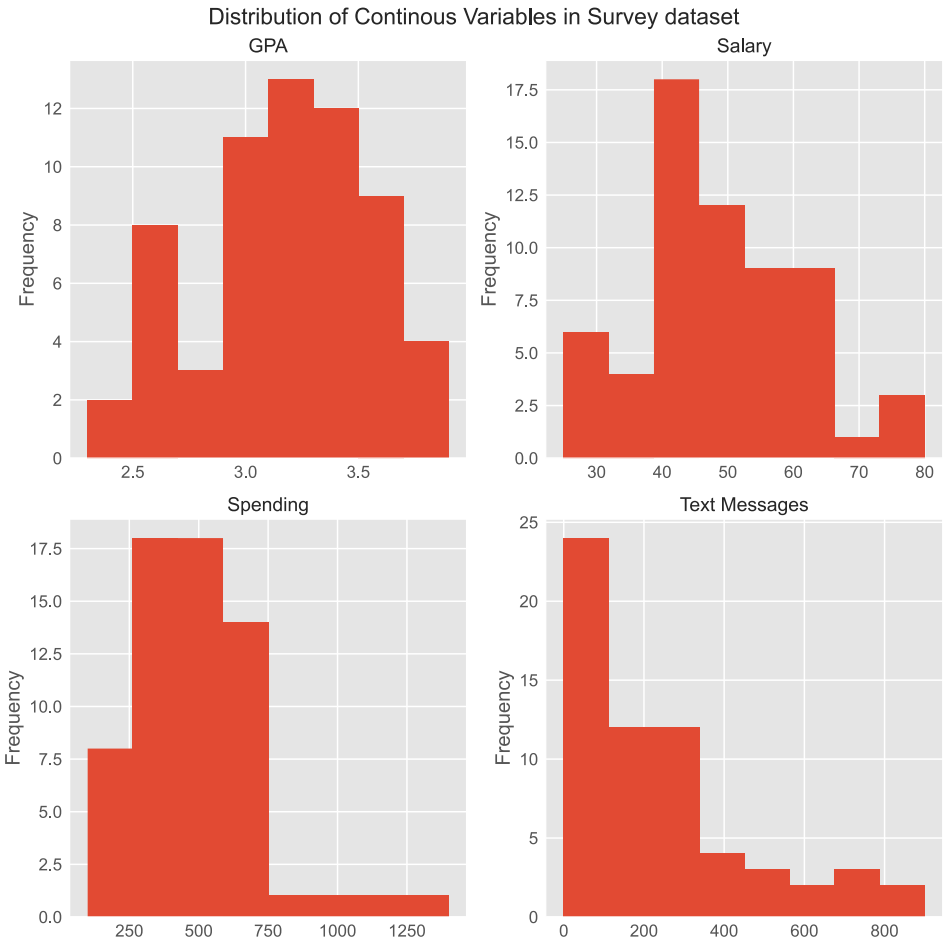

Figure 5: Distribution of Continuous Variables in Survey Dataset

Theoretically the mean, median and mode of a normal distribution are equal.

Distribution of GPA

```
For the GPA variable:
Mean = 3.1
Median = 3.2
Mode = [3.0, 3.1, 3.4]
```

Figure 6: QQ-Plot for GPA

From the above outputs we can conclude that the GPA variable in the Survey dataset approximately follows a normal distribution. This can also be verified from the histogram in Figure 5.

Distribution of Salary

```
For the Salary variable:
Mean = 48.5
Median = 50.0
Mode = 40.0
```

Figure 7: QQ-Plot for Salary

The mean, median and mode of Salary are very different from each other but its QQ-Plot is close to the straight line. Hence the Salary variable approximately follows a normal distribution. This can also be verified from the histogram in Figure 5.

Distribution of Spending

```
For the Spending variable:
Mean = 482.0
Median = 500.0
Mode = 500
```

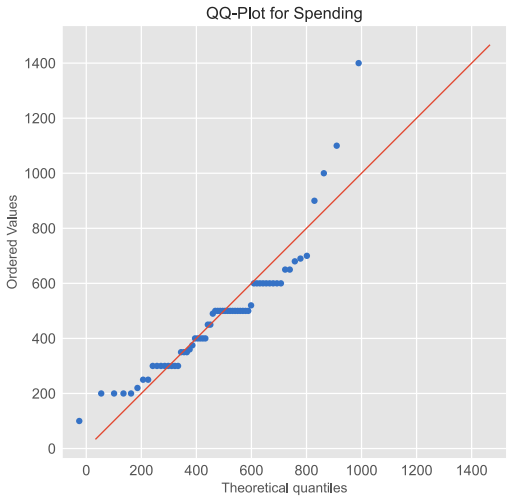

Figure 8: QQ-Plot for Spending

Even though the mean, median and mode of Spending are close to each other, its QQ-Plot varies a lot from the straight line. Hence the Spending variable does not follow a normal distribution. This can also be verified from the histogram in Figure 5.

Distribution of Text Messages

```
For the Text Messages variable:
Mean = 246.2
Median = 200.0
Mode = 300
```

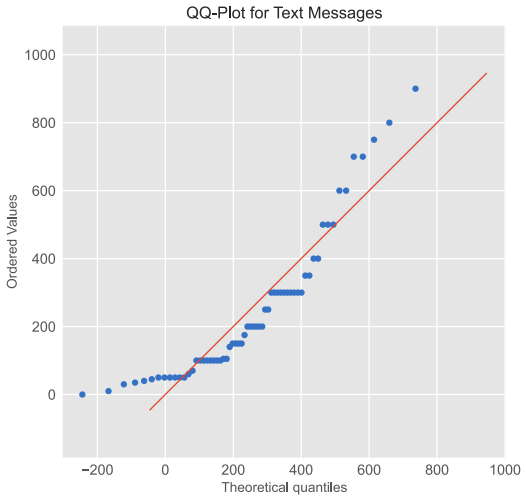

Figure 9: QQ-Plot for Text Messages

As the mean, median and mode of Text Messages variable are very different from each other and its QQ-Plot is not close to the straight line, the Text Messages variable does not follow a normal distribution. This can also be verified from the histogram in Figure 5.

Conclusion

1. The variables GPA and Salary approximately follow a normal distribution.
2. The variables Spending and Text Messages does not follow a normal distribution.

Problem 3

Executive Summary

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 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. In this problem statement, we will test some hypothesis regarding the moisture content of the A and B shingles using appropriate test statistic.

Sample of the A & B Shingles Dataset

	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

Table 15: A & B Shingles Dataset Sample

Checking the types of variables in the dataset.

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

- As mentioned in the problem statement, we can confirm that A shingles column has 36 measurements and B shingles column has 31 measurements.
- Both the columns are float type variables.
- Column B has 5 null values, which is expected from the problem statement.

Checking the distribution of both variables in the Shingles dataset.

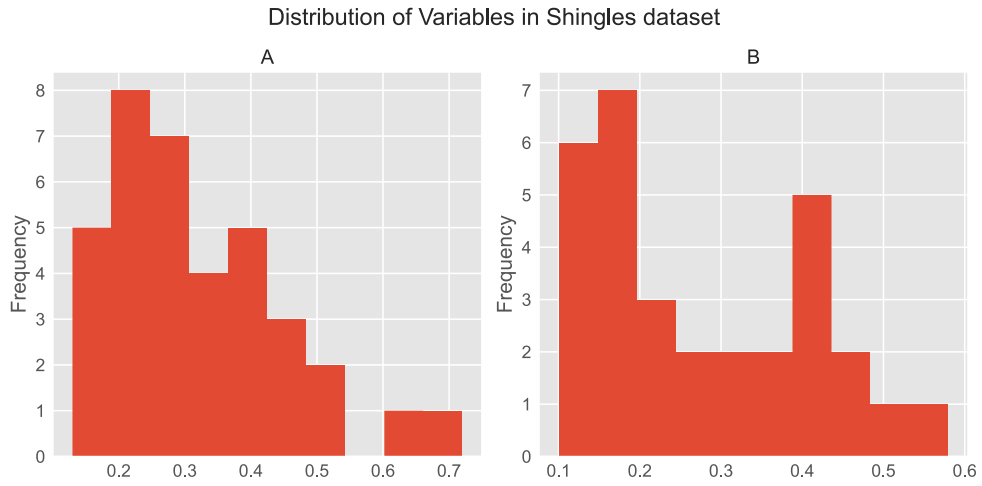

Figure 10: Distribution of Variables in Shingles Dataset

From the above histograms we can see that the measurements for both A and B shingles are not normally distributed.

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

Stating the Null and Alternative Hypothesis for both A and B shingles.

The company's current status is that the mean moisture content is less than 0.35 pounds per 100 square feet. If there is enough evidence against this claim, then the company will have to take preventive actions.
Hence the Null and Alternative Hypothesis for both A and B shingles are:

$$H_0: \mu_{moisture\ content} \leq 0.35$$

$$H_A: \mu_{moisture\ content} > 0.35$$

Deciding on the Type of Test to use.

Here we have to check whether the moisture content in both shingles is within permissible limits. Also, the population parameters are not provided. Assuming that the samples are randomly selected, independent and their population follows a normal distribution, we are going to use the One-sample t-test for both the shingles separately. From the Alternative Hypothesis we can conclude that this is a right tailed t-test.

Deciding on the Significance Level.

As the significance level is not mentioned, here we consider it to be 5%. Hence,

$$\alpha = 0.05$$

Hypothesis test for A shingles

```
t_statistic = -1.4735046253382782 and p_value = 0.9252236685509249
```

As the $p\,value > 0.05$, we fail to reject the Null Hypothesis. In other words, we do not have enough evidence to reject the claim that the moisture content is less than 0.35 pounds per 100 square feet for the A shingles.

Hypothesis test for B shingles

```
t_statistic = -3.1003313069986995 and p_value = 0.9979095225996808
```

As the $p\,value > 0.05$, we fail to reject the Null Hypothesis. In other words, we do not have enough evidence to reject the claim that the moisture content is less than 0.35 pounds per 100 square feet for the B shingles.

Conclusion

We have enough evidence to conclude that the moisture content is less than 0.35 pounds per 100 square feet for both A and B shingles.

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?

Stating the Null and Alternative Hypothesis

Here we have to check whether the mean moisture content is equal for both A and B shingles.
Hence the Null and Alternative Hypothesis are:

$$H_0: \mu_A - \mu_B = 0$$

$$H_A: \mu_A - \mu_B \neq 0$$

Deciding on the Type of Test to use.

The population parameters are not provided. Assuming that the samples are randomly selected, independent and their population follows a normal distribution, we are going to use the Two-sample independent t-test for both the shingles. This is a two-tailed t-test.

Assumptions to check for Two-sample t-test

To perform a 2-sample t-test, we assume that the sample variance of both the samples are equal.

```
The ratio of the variance of both samples is 0.9773231765154546.
As the ratio is close to 1.0, we can proceed with the 2-sample t-test.
```

Deciding on the Significance Level.

As the significance level is not mentioned, here we consider it to be 5%. Hence,

$$\alpha = 0.05$$

Hypothesis Test

```
t_statistic = 1.2896282719661123 and p_value = 0.2017496571835306
```

Conclusion

As the $p\,value > 0.05$, we fail to reject the Null Hypothesis. In other words, we have enough evidence to conclude that the mean moisture content of both A and B shingles are equal to each other.