SMDM PROJECT

REPORT

DSBA

Submitted By,

Ragavedhni K R

Contents

SMDM PROJECT REPORT 1

Problem 1

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

1.1. Use methods of descriptive statistics to summarize data. Which Region and which Channel seems to spend more? Which Region and which Channel seems to spend less?

- For calculating the total spends on all the items, we add a new column 'Spend' to the dataset. As there is vast difference in the Buyer/Spender, we calculate across the Region and Channel separately.
- By using the pivot table on the index as Region, we get that the total spend is high in the Other and the least spend is the Oporto.

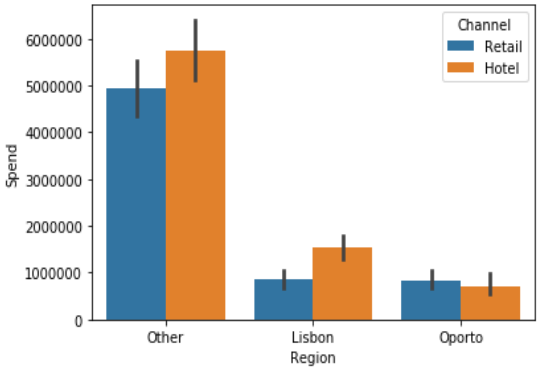
Region	Buyer/Spender	Delicatessen	Detergents_Paper	Fresh	Frozen	Grocery	Milk	Spend
Lisbon	18095	104327	204136	854833	231026	570037	422454	2386813
Oporto	14899	54506	173311	464721	190132	433274	239144	1555088
Other	64026	512110	890410	3960577	930492	2495251	1888759	10677599

- Considering the total spends across the Channel there is high spend in the Hotel and least spend in the Retail.

Channel	Buyer/Spender	Delicatessen	Detergents_Paper	Fresh	Frozen	Grocery	Milk	Spend
Hotel	71034	421955	235587	4015717	1116979	1180717	1028614	7999569
Retail	25986	248988	1032270	1264414	234671	2317845	1521743	6619931

- Grouping the Region and Channel, we see the most spend Region - Channel is Other- Hotel and the least spend is Oporto - Hotel as shown below.

```
Region  Channel
Oporto  Hotel       719150
        Retail      835938
Lisbon  Retail      848471
        Hotel      1538342
Other   Retail     4935522
        Hotel      5742077
Name: Spend, dtype: int64
```



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

- The items are considered Region and Channel wise separately to know their preference behaviour.
- Items sold in the order (High to Low) Region Wise:

```
Fresh              854833
Grocery            570037
Milk               422454
Frozen             231026
Detergents_Paper   204136
Delicatessen       104327
Name: Lisbon, dtype: int64
```

```
Fresh              464721
Grocery            433274
Milk               239144
Frozen             190132
Detergents_Paper   173311
Delicatessen        54506
Name: Oporto, dtype: int64
```

```
Fresh              3960577
Grocery            2495251
Milk               1888759
Frozen              930492
Detergents_Paper    890410
Delicatessen        512110
Name: Other, dtype: int64
```

- The order of preference of items is same across the regions, but the rate of preferring an item across the region is different.
- Items sold in the order (High to Low) Channel wise:

```
Fresh              4015717
Grocery            1180717
Frozen             1116979
Milk               1028614
Delicatessen        421955
Detergents_Paper    235587
Name: Hotel, dtype: int64
```

```
Grocery            2317845
Milk               1521743
Fresh              1264414
Detergents_Paper   1032270
Delicatessen        248988
Frozen              234671
Name: Retail, dtype: int64
```

- The order of preference of items across the channel is different.
- Grouping the Region and Channel for all the items, we get below details.

	Most Preferred		Least Preferred	
Items	**Region**	**Channel**	**Region**	**Channel**
Delicatessen	Lisbon	Retail	Oporto	Hotel
Detergents_Paper	Oporto	Retail	Oporto	Hotel
Fresh	Other	Hotel	Lisbon	Retail
Frozen	Oporto	Hotel	Other	Retail
Grocery	Lisbon	Retail	Other	Hotel
Milk	Other	Retail	Oporto	Hotel

- As the people preferences vary, we see noticeable behaviour differences of the items across the Region and the Channel.

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

- In statistics, there is a chance of occurrence of deviation or variability of data; which is bound to be dispersed around the mean.

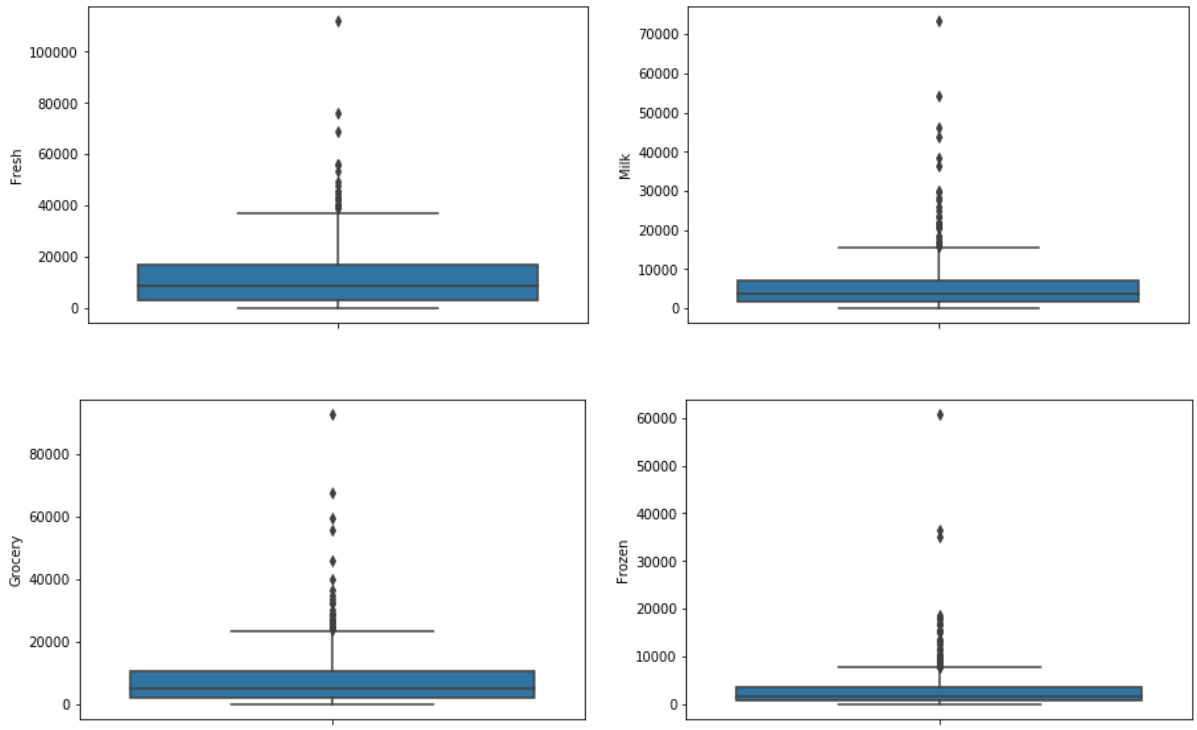
- The various descriptive measure of variability are Range, Inter-quartile range, Variance, Standard Deviation, Coefficient of variation, Empirical rules.
- The Coefficient of Variation (CV) helps to compare the degree of variability between different items from their mean, even if the mean values across the items vary drastically from one another.

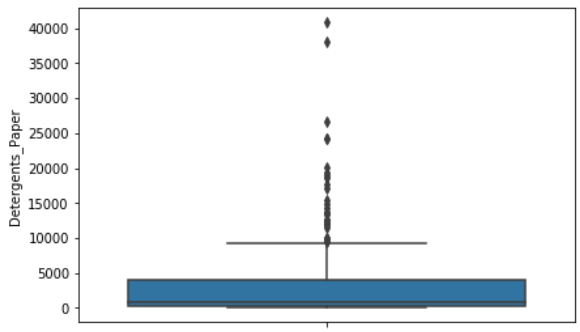
```
The Coefficient of Variation of Fresh: 1.0527196084948243
The Coefficient of Variation of Grocery: 1.1938154477492668
The Coefficient of Variation of Milk: 1.27185083074245
The Coefficient of Variation of Frozen: 1.578535529860776
The Coefficient of Variation of Detergents_Paper: 1.6527657881041735
The Coefficient of Variation of Delicatessen: 1.8473041039189302
```

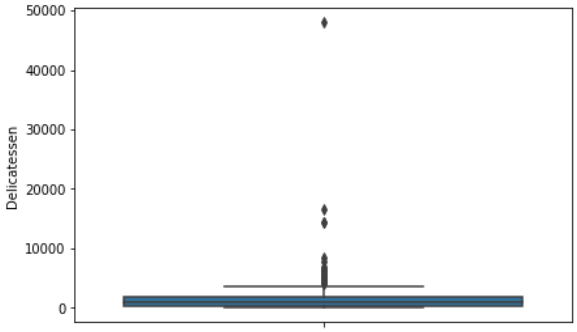
- The consideration for consistency is the smaller the CV, the higher the consistency (or least inconsistent).
- From the above result, we could see that the item 'Fresh' shows the least inconsistent behavior (more consistency).
- The most inconsistent behavior (less consistency) is shown by the item 'Delicatessen'.

1.4. Are there any outliers in the data?

Yes, the outliers are present in the data for all the items- Fresh, Milk, Grocery, Frozen, Delicatessen, and Detergents_Paper.
The box plot is plotted below to view the outliers for the items.







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

1. From the sample data, we correlate between two famous Holiday spots (Oporto and Lisbon) with the remaining region (Other) in Portugal. Oporto has the more floating population or few people stay based on the total buyers pattern.

2. Most of the population in the country stays in the Hotel (nearly thrice than Retail), it is obvious that people are mainly tourists or visitors.

Region	Buyer/Spender
Lisbon	18095
Oporto	14899
Other	64026

Channel	Buyer/Spender
Hotel	71034
Retail	25986

3. The people in the 3 regions prefer ‘Fresh’ items as they are healthier, whereas ‘Delicatessen’ is the least preferred as it an exotic food.

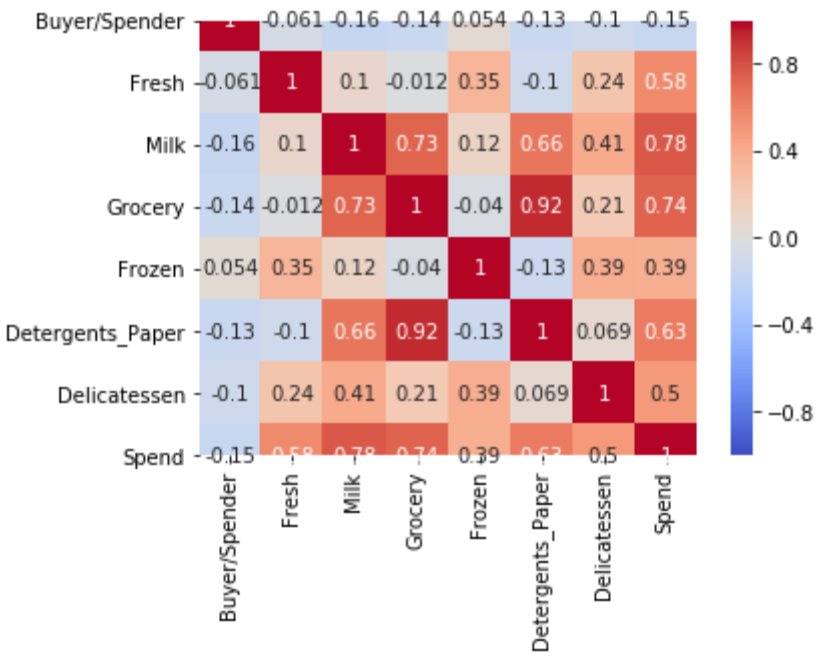
4. There is a pattern seen, where the items ‘Delicatessen’, ‘Fresh, ‘Frozen’ are sold high in the Channel Hotel. This can be based on the food habits and people’s affordability.

5. In the same way, a pattern is observed in the Channel Retail, where the items ‘Fresh’, ‘Grocery’, ‘Milk’ are sold high. The people staying in that place permanently would prefer to buy items and cook for themselves.

6. As the people are tourists or visitors, they would not prefer washing clothes during their short span of time, so ‘Detergents_Paper’ is least sold in Hotel.

7. In day-to-day pattern of living, the people prefer healthy food habits; avoiding the processed or frozen food items, so ‘Frozen’ items are less purchased in Retail.

8. There is a high correlation found between 'Grocery' and 'Detergents_Paper', and second most correlation between 'Grocery' and 'Milk', as they are basic necessities in a household.



9. The dataset has outliers so all the values are not considered for the items, we can try to remove the outliers using the median.

```
Other      316
Lisbon      77
Oporto      47
Name: Region, dtype: int64
```

```
Hotel      298
Retail     142
Name: Channel, dtype: int64
```

10. While considering the data count from the dataset, it is not uniform across the Regions and between the Channels. Hence we need higher sample size to get more insights on the items sold in Lisbon and Oporto.

Problem 2

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

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

2.1.1. Gender and Major

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

2.1.2. *Gender and Grad Intention*

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

2.1.3. *Gender and Employment*

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

2.1.4. *Gender and Computer*

Computer	Desktop	Laptop	Tablet
Gender			
Female	2	29	2
Male	3	26	0

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

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

```
Female    33
Male      29
Name: Gender, dtype: int64
```

P (Male) = 29/62 = 0.467742
P (Female) = 33/62 = 0.232258

```
Gender
Male    0.467742
dtype: float64
```

```
Gender
Female    0.532258
dtype: float64
```

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

Conditional Probability of Major and Gender

Major	Gender	
Accounting	Female	0.090909
	Male	0.137931
CIS	Female	0.090909
	Male	0.034483
Economics/Finance	Female	0.212121
	Male	0.137931
International Business	Female	0.121212
	Male	0.068966
Management	Female	0.121212
	Male	0.206897
Other	Female	0.090909
	Male	0.137931
Retailing/Marketing	Female	0.272727
	Male	0.172414
Undecided	Male	0.103448

dtype: float64

The conditional probability of different majors among the male students in CMSU:
P (Accounting | Male) = 0.137931
P (CIS | Male) = 0.034483
P (Economics/Finance | Male) = 0.137931
P (International Business | Male) = 0.068966
P (Management | Male) = 0.206897
P (Other | Male) = 0.137931
P (Retailing/Marketing | Male) = 0.172414
P (Undecided | Male) = 0.103448

The conditional probability of different majors among the female students in CMSU:
P (Accounting | Female) = 0.090909
P (CIS | Female) = 0.090909
P (Economics/Finance | Female) = 0.212121
P (International Business | Female) = 0.121212
P (Management | Female) = 0.121212
P (Other | Female) = 0.090909
P (Retailing/Marketing | Female) = 0.272727

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.

The conditional probability of intent of graduate, given that the student is a male:
P (Yes |Male) = 0.586207
P (No | Male) = 0.103448
P (Undecided | Male) = 0.310345

Conditional Probability of Intent to Grad and Gender

```
Grad Intention  Gender
No              Female    0.272727
                Male      0.103448
Undecided       Female    0.393939
                Male      0.310345
Yes             Female    0.333333
                Male      0.586207
dtype: float64
```

The conditional probability of intent of graduate, given that the student is a female:
P (Yes | Female) = 0.333333
P (No | Female) = 0.272727
P (Undecided | Female) = 0.393939

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

Conditional Probability of Employment and Gender

```
Employment  Gender
Full-Time   Female    0.090909
            Male      0.241379
Part-Time   Female    0.727273
            Male      0.655172
Unemployed  Female    0.181818
            Male      0.103448
dtype: float64
```

The conditional probability of employment status for the male students:
P (Full-Time | Male) = 0.241379
P (Part-Time | Male) = 0.655172
P (Unemployed | Male) = 0.103448

The conditional probability of employment status for the female students:
P (Full-Time | Female) = 0.090909
P (Part-Time | Female) = 0.727273
P (Unemployed | Female) = 0.181818

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

The conditional probability of Desktop, Laptop, Tablet preference among the male students:
P (Desktop | Male) = 0.103448
P (Laptop | Male) = 0.896552

Conditional Probability of Computer and Gender

```
Computer  Gender
Desktop   Female    0.060606
          Male      0.103448
Laptop    Female    0.878788
          Male      0.896552
Tablet    Female    0.060606
dtype: float64
```

The conditional probability of Desktop, Laptop, Tablet preference among the female students:
P (Desktop | Female) = 0.060606
P (Laptop | Female) = 0.878788
P (Tablet | Female) = 0.060606

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.

For two events A & B to be independent, P (A|B) = P (A) and P (B|A) = P (B).

CASE 1:
The conditional probability of different majors among the male students in CMSU:
P (Major = 'Accounting' | Gender= 'Male') = 0.137931.
The conditional probability of different majors among the female students of CMSU:
P (Major = 'Accounting' | Gender= 'Female') = 0.090909.

If the column 'Major' is independent of the 'Gender', then P (Major = 'Accounting' | Gender= 'Male') and P (Major = 'Accounting' | Gender= 'Female') would be same as P (Major = 'Accounting') = 0.112903, but here in this case it is dependent of the Gender= 'Male' and Gender = 'Female'.

Hence, the column variable 'Major' is not independent of 'Gender'.

CASE 2:
The conditional probability of intent to graduate, given that the student is a male:
P (Grad Intention= 'Yes'| Gender= 'Male') = 0.586207.
The conditional probability of intent to graduate, given that the student is a female:
P (Grad Intention= 'Yes'| Gender= 'Female') = 0.333333.

If the column 'Grad Intention' is independent of the 'Gender', then P (Grad Intention= 'Yes'| Gender= 'Male') and P (Grad Intention= 'Yes'| Gender= 'Female') would be same as P (Grad Intention= 'Yes') = 0.451613, but here in this case it is dependent of the Gender= 'Male' and Gender = 'Female'.

Hence, the column variable 'Grad Intention' is not independent of 'Gender'.

CASE 3:
The conditional probability of employment status among the male students:
P (Employment= 'Full-Time'| Gender= 'Male') = 0.241379.
The conditional probability of employment status among female students:
P (Employment= 'Full-Time'| Gender= 'Female') = 0.090909.

If the column 'Employment' is independent of the 'Gender', then P (Employment= 'Full-Time'| Gender= 'Male') and P (Employment= 'Full-Time'| Gender= 'Female') would be same as P (Employment= 'Full-Time')= 0.161290, but here in this case it is dependent of the Gender= 'Male' and Gender = 'Female'.

Hence, the column variable 'Employment' is not independent of 'Gender'.

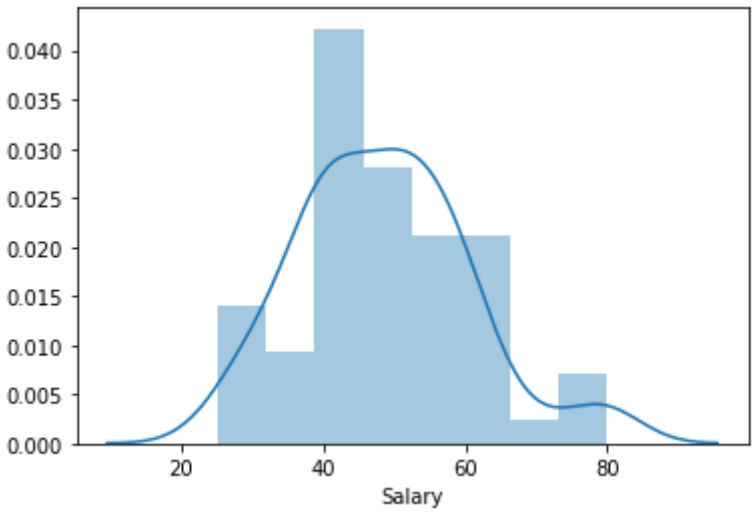
CASE 4:
The conditional probability of laptop preference among the male students:
P (Computer='Laptop'| Gender='Male) = 0.896552.
The conditional probability of laptop preference among the female students:
P (Computer='Laptop'| Gender='Female) = 0.878788.
If the column 'Computer' is independent of the 'Gender', then P (Computer='Laptop'| Gender= 'Male') and P (Computer= 'Laptop'| Gender= 'Female') would be same as P (Computer='Laptop') = 0.887097, but here in this case it is dependent of the Gender= 'Male' and Gender = 'Female'.

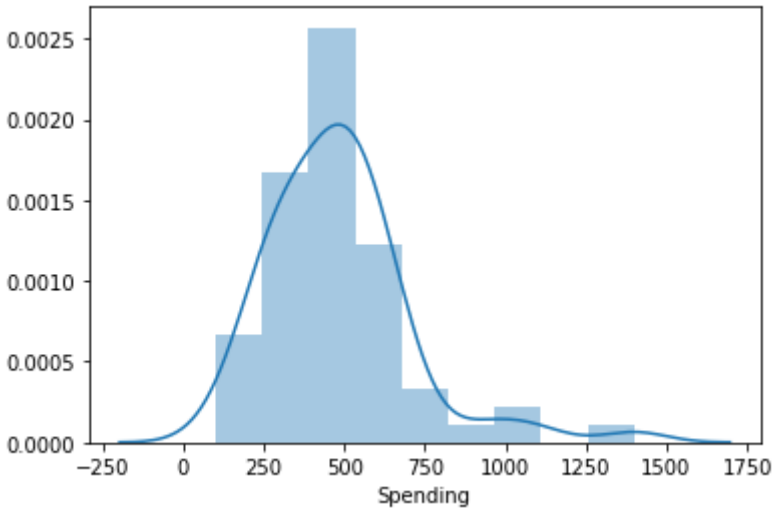
Hence, the column variable 'Computer' is not independent of 'Gender'.

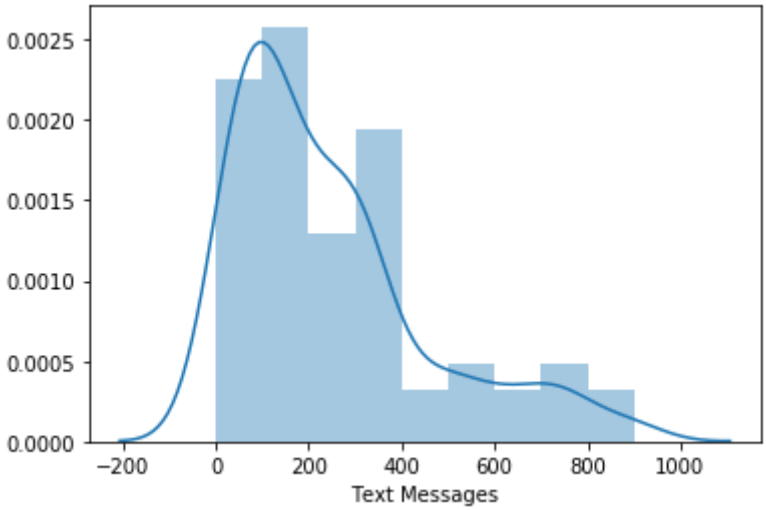
Therefore, all the columns Major, Employment, Graduate Intention, and Computer are dependent on Gender or they are biased to Gender.

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]

- Plotting with distribution graphs, it is evident that the continuous variables, Salary, Spending and Text Messages in the dataset follow the normal distribution.







- While checking the skewness of the columns, they are Right- Skewed distribution or Positive- Skew distribution.
- The column Salary is the least skewed towards the right and the column Spending is the most skewed towards the right.

Skewness of Spending, Salary, Text Messages

```
Spending         1.585915
Text Messages    1.295808
Salary           0.534701
dtype: float64
```

- In the below summary for the continuous variables, the mean of the data differs slightly from the median (50%) of the data. Eg: For Salary, mean is 48.548 and median is 50.000.
- The standard deviation (1σ) for the columns, Salary is 12.08, Spending is 221.953 and Text Messages is 214.465.

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

- Applying the Empirical rule in the region, this predicts that 68% of the data is distributed within one standard deviation (μ±σ), 95% of the data within two standard deviations (μ±2σ), and 99.7% of the data within three standard deviations (μ±3σ).

	Salary			Spending			Text Messages	
	+	-		+	-		+	-
μ±σ	36.46748	60.6293	μ±σ	260.0623	703.9699	μ±σ	31.74373	460.6756
μ±2σ	24.38656	72.71021	μ±2σ	38.10852	925.9237	μ±2σ	-182.722	675.1416
μ±3σ	12.30565	84.79112	μ±3σ	-183.845	1147.878	μ±3σ	-397.188	889.6075

- The columns Salary, Spending, Text Messages follow normal distribution.

Problem 3

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

To check whether the population means for shingles A and B are equal, we need to verify the means of shingle A and B respectively.

Calculating One-Sample t-test for A Shingles to verify the hypothesis:

Step 1: Formulate null hypothesis and alternate hypothesis

H0: μ <= 0.35 HA: μ > 0.35

Step 2: Decide the level of significance (α) as 0.05.(5%)

Step 3: Identify the test statistic

- We do not know the population standard deviation (σ).
- The sample size is n >30, so we use the t distribution and the t_{STAT} test statistic.
- To find the test statistic using 1 sample t-test for A Shingles.

Step 4: Calculate the test statistic and p_value

```
One sample t test for A Shingles
t_stat_A: -1.4735046253382782
p_value_A: 0.07477633144907513
```

Step 5: Decide to reject or accept null hypothesis

- We see that the p_value > α. Thus, it is confirmed we are failing to reject the null hypothesis with 5% level of significance.
- So, at 95% confidence level, there is sufficient evidence to prove that the population mean moisture content of A shingles is less than or equal 0.35 pound per 100 square feet.

Calculating One-Sample t-test for B Shingles to verify the hypothesis:

Step 1: Formulate null hypothesis and alternate hypothesis

H0: μ <= 0.35 HA: μ > 0.35

Step 2: Decide the level of significance (α) as 0.05. (5%)

Step 3: Identify the test statistic

- We do not know the population standard deviation (σ).
- The sample size is n >30, so we use the t distribution and the t_{STAT} test statistic.
- To find the test statistic using 1 sample t-test for B Shingles.

Step 4: Calculate the test statistic and p_value

```
One sample t test for B Shingles
t_stat_B: -3.1003313069986995
p_value_B: 0.0020904774003191826
```

Step 5: Decide to reject or accept null hypothesis

- We see that the p-value < α. Thus, it is confirmed we are rejecting the null hypothesis.
- With 95% confidence, there is no sufficient evidence to prove that the population mean moisture content of B shingles is less than 0.35 pound per 100 square feet.

The hypotheses to check the equality of population mean of A shingles and B shingles.

Step 1: Define null hypothesis and alternate hypothesis.

H0: μ (Population Mean A Shingles) = μ (Population Mean B Shingles)

H1: μ (Population Mean A Shingles) != μ (Population Mean B Shingles)

Step 2: Decide the level of significance (α) as 0.05.

Step 3: Identify the test statistic:

- We have two samples and we do not know the population standard deviation.
- Sample sizes for both samples are different, n1= 36 and n2= 31.
- We use the t distribution and the t_{STAT} test statistic for two sample unpaired test (Welch's t-test).

Step 4: Calculate the test statistic and p_value

```
Two sample t test for A Shingles and B Shingles
t_stat: 1.2885080295255027
p_value: 0.20225822050217818
```

Step 5: Decide to reject or accept null hypothesis

- The p_value is larger than alpha, therefore we are failing to reject the Null hypothesis with 5% significance level.
- We conclude that the evidence persuades that the two population means are same, even though we assume the two populations have (or may have) different standard deviations.

The assumptions we need to consider before the test for equality of means is performed:

- Welch t test assumes that both the group of data are sampled from Gaussian populations, but does not assume those two populations have the same standard deviation.
- For the unequal variance t test, the null hypothesis is that the two population means are the same but the two population variances may differ.
- For interpreting any P value, it is essential that the null hypothesis to be carefully defined.
- If the P value is large, you don't reject that null hypothesis, so conclude that the evidence does not persuade you that the two population means are different, even though you assume the two populations have (or may have) different standard deviations.
- The unequal variance t test reports a confidence interval for the difference between two means that is usable even if the standard deviations differ.

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

The assumptions to be considered before performing hypothesis test:

- The statistical tests are assumed as random sampling, means the data is collected from a representative (population).
- The tests of hypotheses about means also assume the interval-ratio level of measurement.
- The population under consideration be normally distributed, as one can specify a level of probability (level of significance, α= 5% or 0.05) as a criterion for acceptance.
- A larger sample size (n>=30) means the distribution of results should approach a normal bell-shaped curve, considering which hypothesis test to be considered.
- For Independent t-test, we consider the homogeneity of variance of both samples, when the standard deviations of samples are approximately equal.