SMDM PROJECT REPORT

DSBA

Made by:
Anmol Tripathi
Data Science and Business Analytics (PGP-DSBA)
Online September_A 2021-22

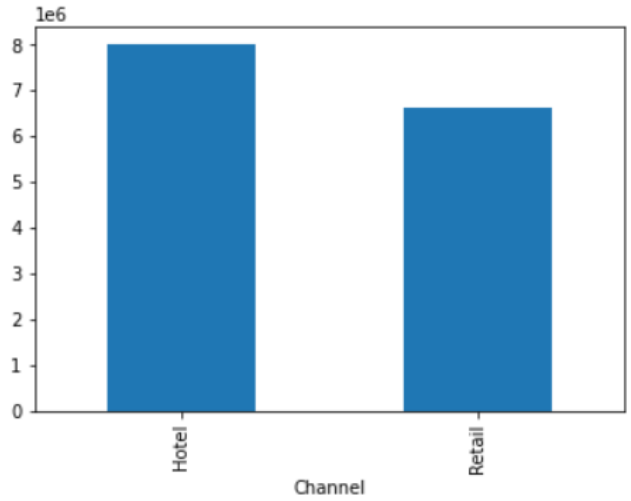
Contents:

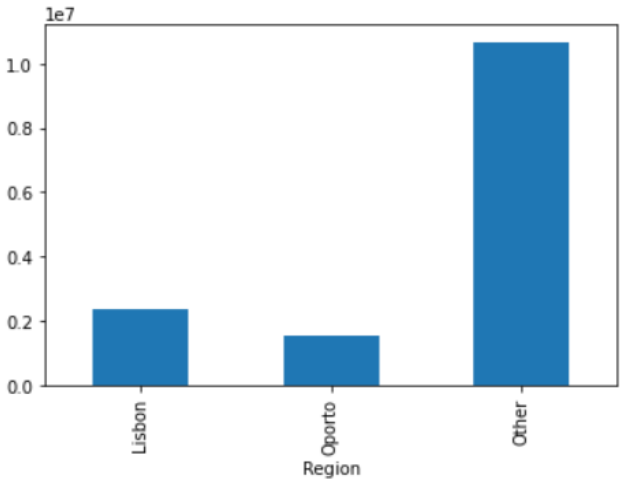
2.5.	Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:	12
2.5.1.	Find the probability that a randomly chosen student is a male or has full-time employment?	12
2.5.2.	Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management.	12
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?	12
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	13
2.7.1.	If a student is chosen randomly, what is the probability that his/her GPA is less than 3?	13
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.	13
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.	14-16
Problem 3:		17-18
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.	17-18
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?	18

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 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?



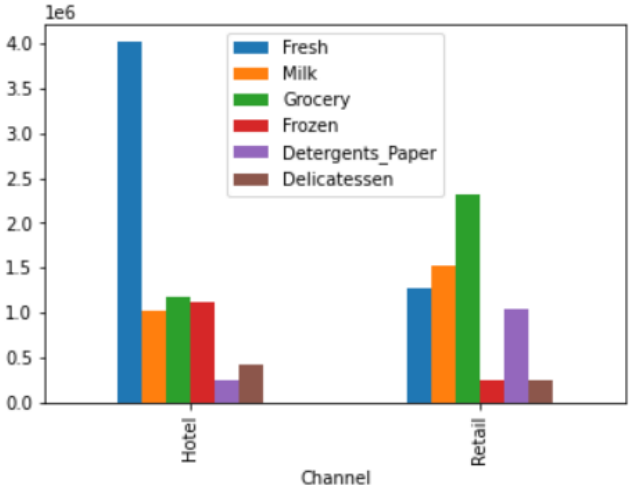


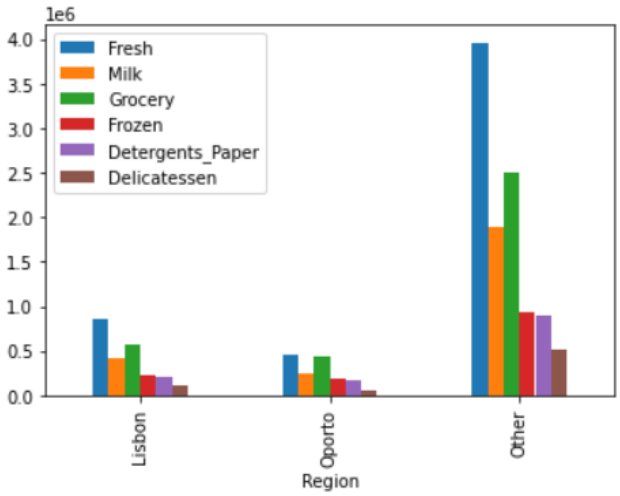
Here from the above graphs, we are easily able to get the following results.

Expenditure most: Channel: Hotel, Region: Others

Expenditure least: Channel: Retailers, Region: Oporto

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.






```
Buyer/Spender       0.00
Fresh               2.56
Milk                4.05
Grocery             3.59
Frozen              5.91
Detergents_Paper    3.63
Delicatessen       11.15
dtype: float64
```

Here from the above following results, we get to know that:

1) Region:

- best investment is Fresh products
- delicatessen remains last for all the three regions.

2) Channel:

- For Hotels Fresh product still remains the best investment
- Least attractive is of detergents
- Groceries are the best investment for the retailers

We can also conclude by stating that Fresh products, milk and groceries are really good to invest in.

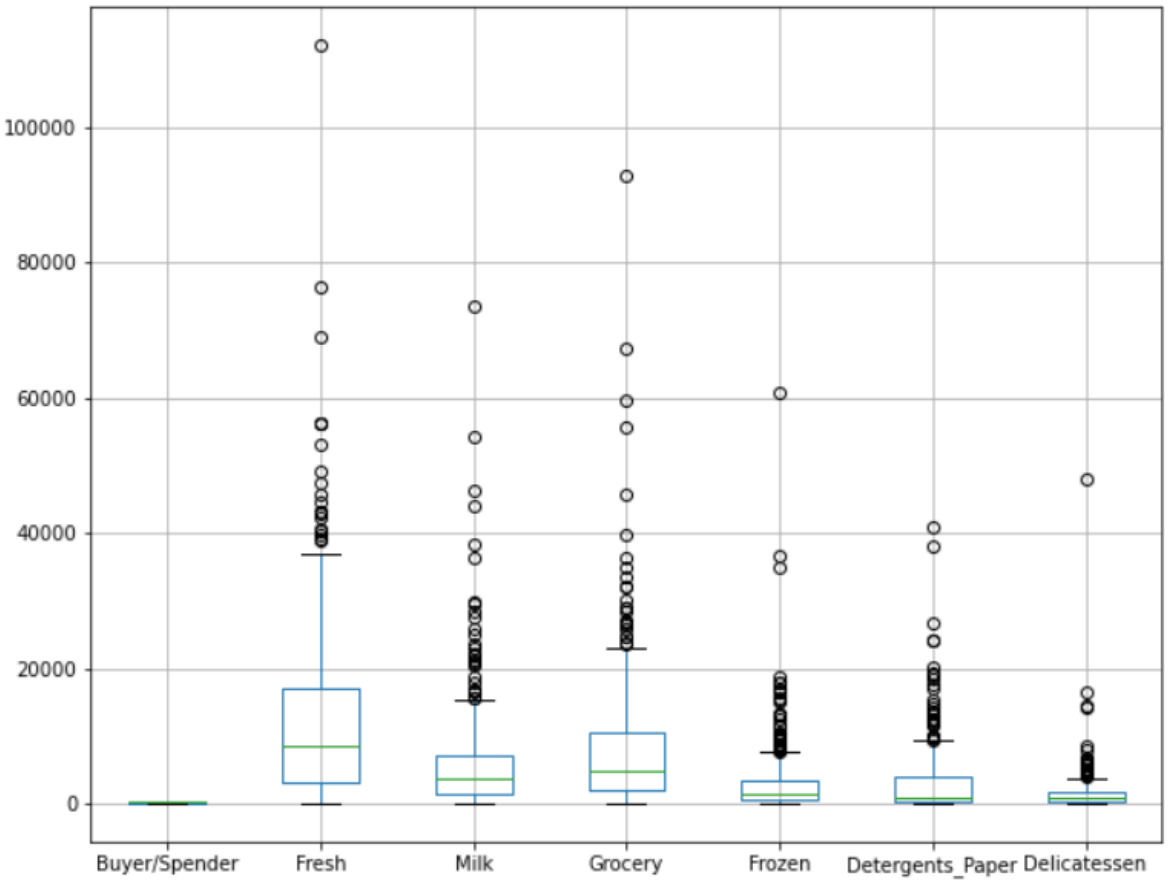
1.3 On the basis of a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?

	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
cv	0.58	1.05	1.27	1.20	1.58	1.65	1.85

Here from the above results, we can easily infer that:

The differences in the minimum/maximum values and mean of the products is high so we are using Coefficient of variance to make a conclusion. The CV (coefficient of variance) is lowest for Fresh products so it is least inconsistent while Delicatessen is most inconsistence. (Skewness is also considered)

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



From the above results, we can easily say that:

There are few outliers which are present in the data. The black coloured spots which are present above the box plot are the outliers itself which are present in the data. The highest individual value of the outlier is present in the Fresh segment.

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

	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
cv	0.58	1.05	1.27	1.20	1.58	1.65	1.85

We can conclude that:

The most consistent are fresh products and investing in them is the best. There is a lot of deviation in data which we can see from the above tabular result. Eg: The minimum and maximum values in the data have large difference and mean of different products also represents deviation. For the same thing, we have also made use of coefficient of variance for our analysis along with various other plots and skewness.

Below mentioned are further more analysis about the data:

1) Region:

- best investment is Fresh products
- delicatessen remains last for all the three regions.

2) Channel:

- For Hotels Fresh product still remains the best investment
- Least attractive is of detergents
- Groceries are the best investment for the retailers

We can also conclude by stating that Fresh products, milk and groceries are really good to invest in.

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*** data set).

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

2.1.3. Gender and Employment:

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

2.1.4. Gender and Computer:

Computer Gender	Desktop	Laptop	Tablet
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 question:

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

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

Probability that a randomly selected CMSU student will be male: 0.46774193548387094

= 46%

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

Probability that a randomly selected CMSU student will be female: 0.532258064516129

= 53%

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

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

Below is the Conditional Probablity formula:

P (Major | male) = P (Major ∩ male)/ P(male)

P (Major | female) = P (Major ∩ female)/ P(female)

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

```
 Probability of male students in different majors is as follows:
Probability of male students in Accounting: 0.13793103448275862
Probability of male students in CIS: 0.034482758620689655
Probability of male students in Economics/Finance: 0.13793103448275862
Probability of male students in International Business: 0.06896551724137931
Probability of male students in Management: 0.20689655172413793
Probability of male students in Others: 0.13793103448275862
Probability of male students in Retailing/Marketing: 0.1724137931034483
Probability of male students in Undecided: 0.10344827586206896
```

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

```
 Probability of Female students in different majors is as follows:
Probability of Female students in Accounting: 0.09090909090909091
Probability of Female students in CIS: 0.09090909090909091
Probability of Female students in Economics/Finance: 0.21212121212121213
Probability of Female students in International Business: 0.12121212121212122
Probability of Female students in Management: 0.12121212121212122
Probability of Female students in Others: 0.09090909090909091
Probability of Female students in Retailing/Marketing: 0.2727272727272727
Probability of Female students in 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.

P(Intent to graduate ∩ Male) = P (Intent to graduate| Male) x P (Male) = 0.27419354838709675

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

P(Females with no laptop ∩ Female) = P (Females with no laptop| Femal) x P (Female) = 0.06451612903225806

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

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

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

P(male U Fulltime)= P(male) + P(Fulltime) - P(male intersection Fulltime)

P(male U Fulltime)= P(male) + P(Fulltime) - P(male intersection Fulltime) = 0.5161290322580645

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

P(A OR B) = P(A) + P(B) as they are mutually exclusive

P(International Business OR Management| Female)= P(International Business) + P(Management) = 0.24242424242424243

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 graduate intention and being female are independent events?

Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

Condition for events being Independent: P(F ∩ Yes) = P(F)P(Yes)
P(F)= 0.825
P(Yes)= 0.7
P(F ∩ Yes)= 0.275
P(F)P(Yes)= 0.5774999999999999
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?

Total students with less than 3 GPA = 17
P(Probability of student with GPA less than 3)= No of students with GPA less than 3 / Total students: 0.27419354838709675

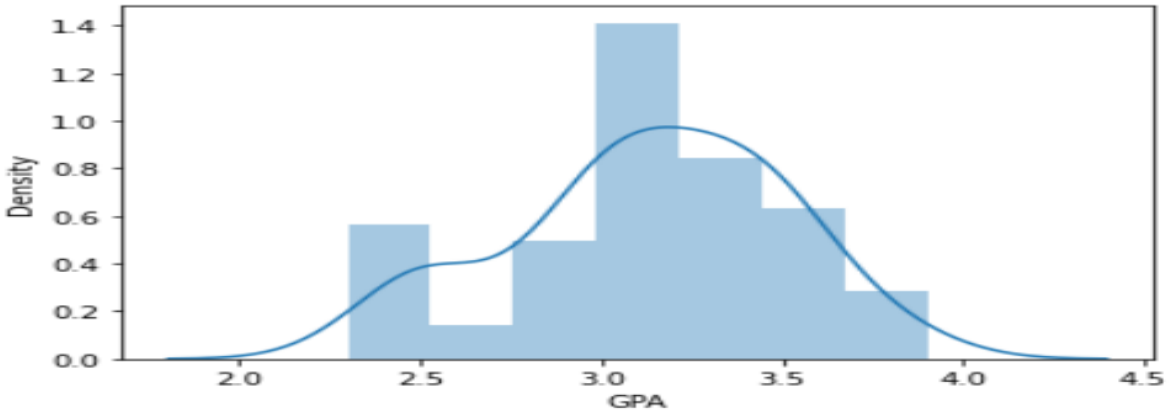
2.7.2 Find conditional probability that a randomly selected male earns 50 or more. Find conditional probability that a randomly selected female earns 50 or more.

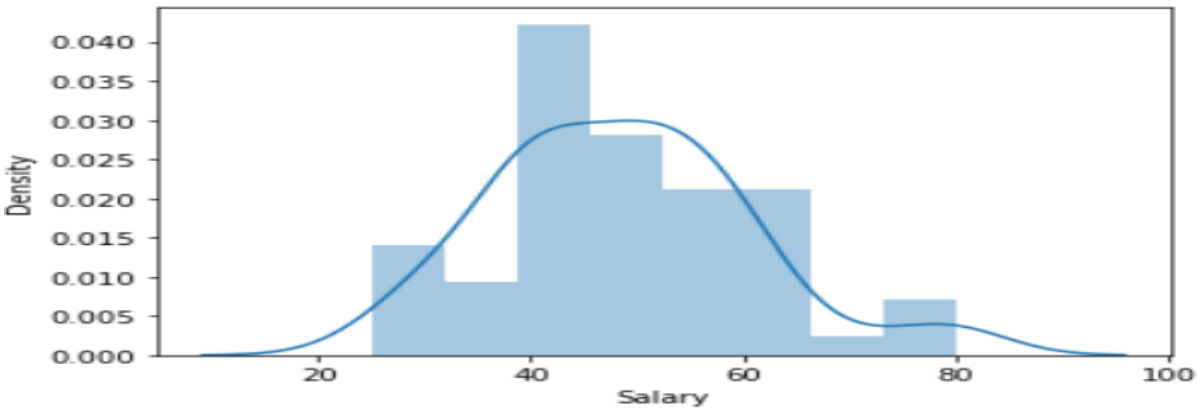
Salary	False	True
Gender		
Female	15	18
Male	15	14

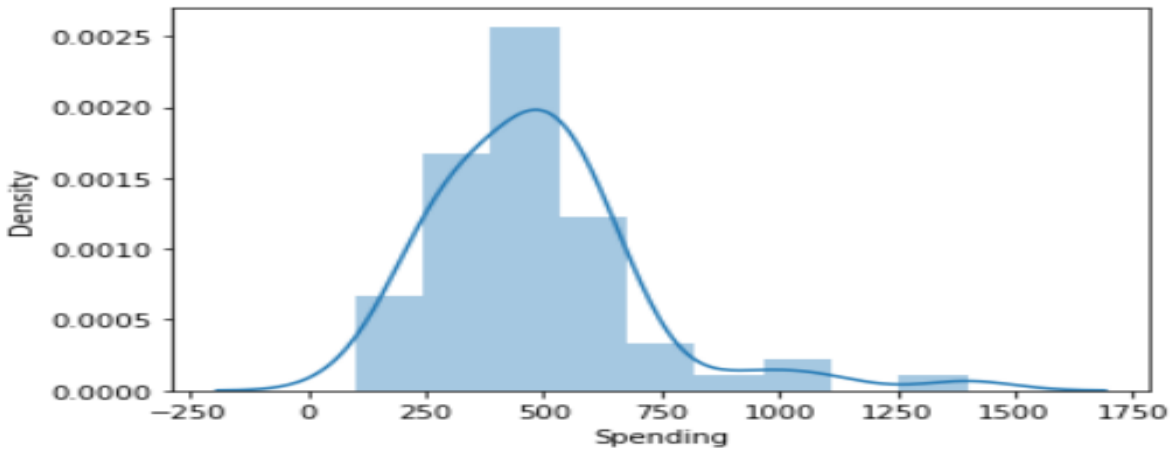
P(Probability of male earns 50 or more)= No of males with 50 and above salary / Total number of males: 0.4827586206896552

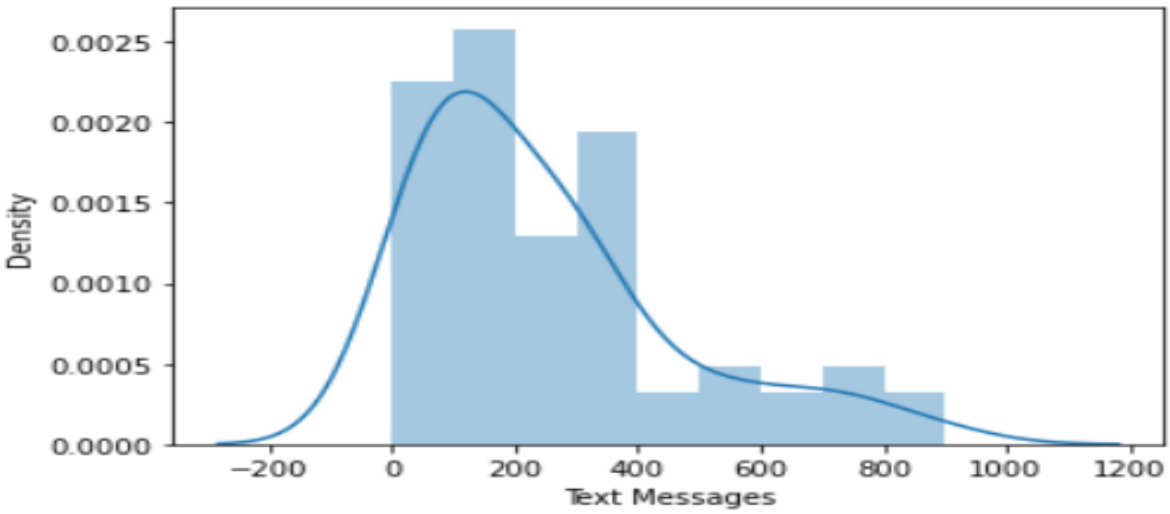
P(Probability of females earns 50 or more)= No of females with 50 and above salary / Total number of females: 0.5454545454545454

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.









	ID	Age	GPA	Salary	Social Networking	Satisfaction	Spending	Text Messages
count	62.00	62.00	62.00	62.00	62.00	62.00	62.00	62.00
mean	31.50	21.13	3.13	48.55	1.52	3.74	482.02	246.21
std	18.04	1.43	0.38	12.08	0.84	1.21	221.95	214.47
min	1.00	18.00	2.30	25.00	0.00	1.00	100.00	0.00
25%	16.25	20.00	2.90	40.00	1.00	3.00	312.50	100.00
50%	31.50	21.00	3.15	50.00	1.00	4.00	500.00	200.00
75%	46.75	22.00	3.40	55.00	2.00	4.00	600.00	300.00
max	62.00	26.00	3.90	80.00	4.00	6.00	1400.00	900.00

```
ID                  0.00
Age                 0.74
GPA                -0.31
Salary              0.53
Social Networking   0.96
Satisfaction       -0.51
Spending            1.59
Text Messages       1.30
dtype: float64
```

Conclusion:

Below are the outcomes which we get from the above results:

1) GPA: Normally distributed

2) Salary: Not Normally Distributed (due to the skewness and stretched tail in dis plot).

3) Spending: Not Normally Distributed (due to the skewness and stretched tail in dis plot).

4) Text Messages: Not Normally Distributed (due to the skewness and stretched tail in dis plot).

Problem 3

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.

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

$H_0 \leq 0.35$

$H_1 > 0.35$

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

$H_0 \leq 0.35$

$H_1 > 0.35$

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.

Considering for Shillings A:

H0: Mean Moisture content <=0.35 (Null Hypothesis)

HA: Mean Moisture content > 0.35 (Alternate Hypothesis)

Alpha = 0.05

T, p_value respectively are given below:

-1.4735046253382782 0.14955266289815025

After dividing p_value by 2

P value for A = 0.07477633144907513

p value for A > 0.05 , we will not reject H0.

Here we get that p2(B) < 0.05, thus we will reject the null Hypothesis. Thus we can say that we have enough evidence to conclude that mean moisture in sample B is not less than 0.35.

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?

H0: Mean of A = Mean of B (Null Hypothesis)

H1: Mean of A is not equal to Mean of B (Alternate Hypothesis)

Alpha= 0.05

T, p_value respectively are given below:

1.2896282719661123 0.2017496571835306

Here P value > 0.05 hence, we will not reject H0. Thus, we can say that the population means for samples A and B are equal.