SMDM Project

PRADEEP KUMAR MISHRA
PGP-DSBA Online
Jun_B_21
Date: 08:Aug:2021

Content View

Problem Statement - 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).

Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and other parameters. 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).

Sample of the 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

Dataset has 9 variables and 440 different types of Buyer/Spender .

Exploratory Data Analysis

Let us check the types of variables and missing values in the dataset

```
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)
```

- From the above results we can see that there is no missing value present in the dataset.
- There are a total 440 rows and 9 columns in the dataset. Out of 9, 2 columns are of object type and the rest 7 are int64 data types.

Total element in dataset

Total number of elements of dataset is 3960

Describe the dataset

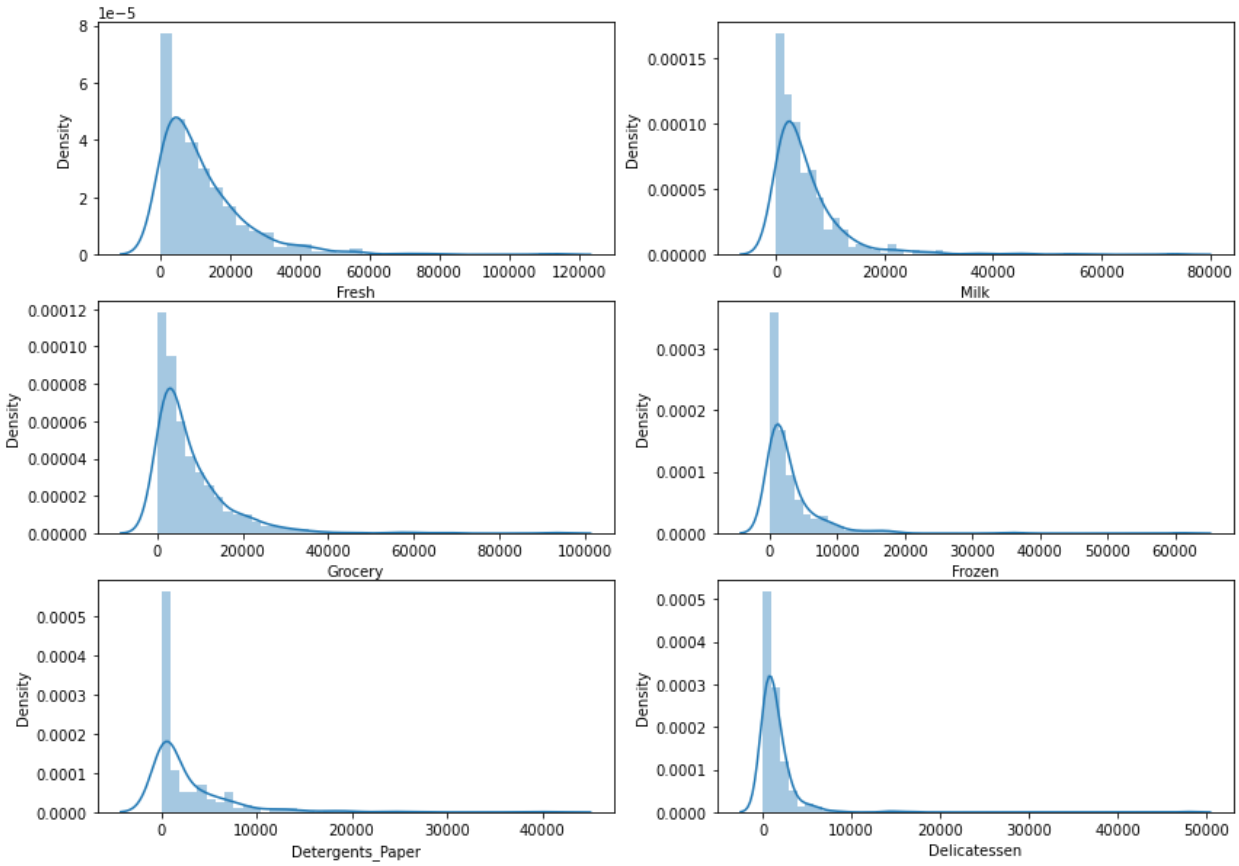
	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
mean	220.500000	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	127.161315	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	1.000000	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	110.750000	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000
50%	220.500000	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000
75%	330.250000	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000
max	440.000000	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000

	Channel	Region
count	440	440
unique	2	3
top	Hotel	Other
freq	298	316

from the above picture we can observe below points.

- We can observe the mean, medium, std, min and max of each variable.
- For the objective data types we can observe count, unique, top and freq.

Checking normality and skewness



Remarks: We can see above figure data is not normally distributed.

In each variable their right skewness.

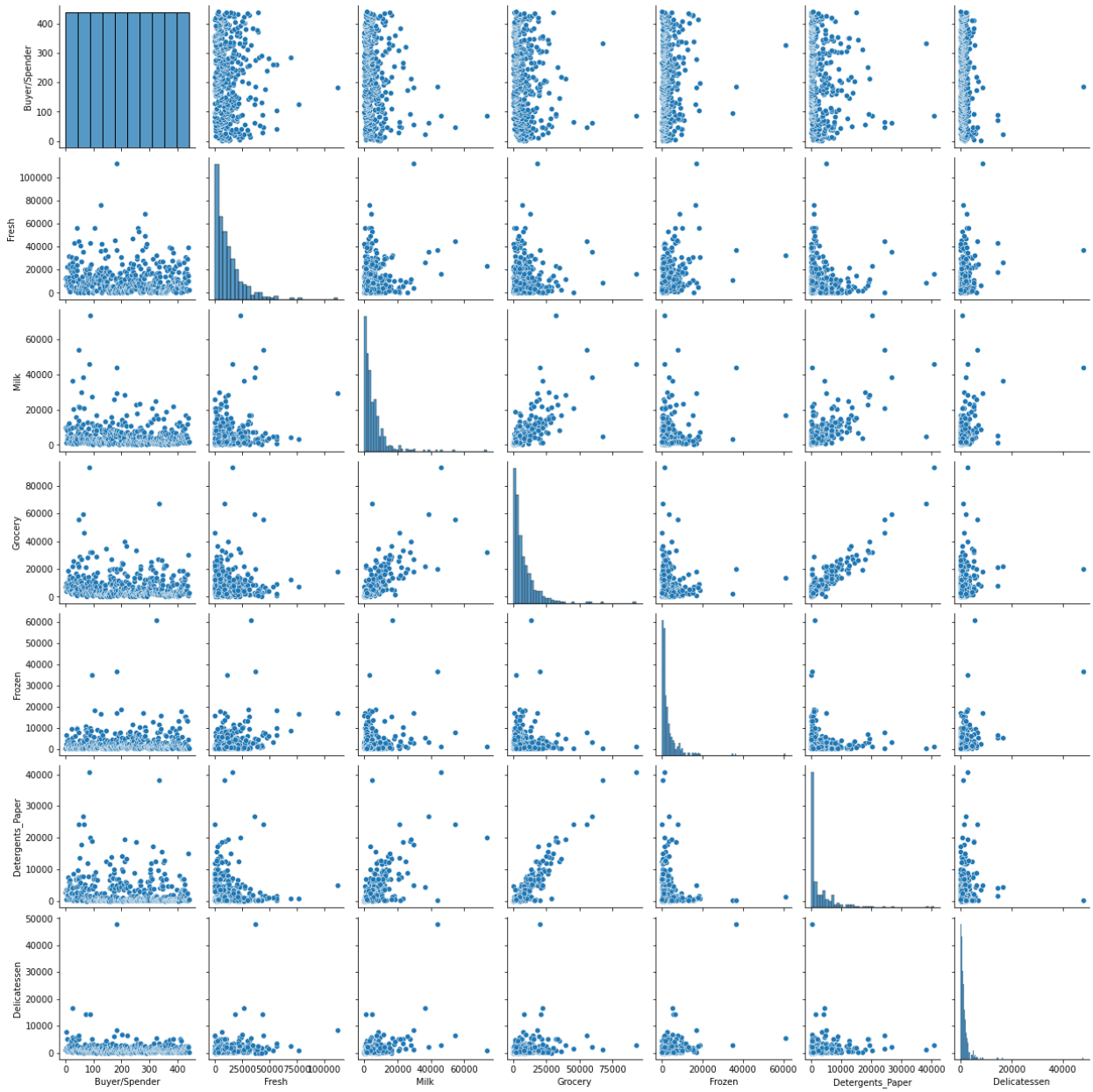
Correlation Plot

From the correlation plot, we can see that various attributes of the A wholesale distributor are not-highly correlated to each other. Correlation values near to 1 or -1 are highly positively correlated and highly negatively correlated respectively. Correlation values near to 0 are not correlated to each other.

Pairplot

Pairplot shows the relationship between the variables in the form of scatterplot and the distribution of the variable in the form of histogram.

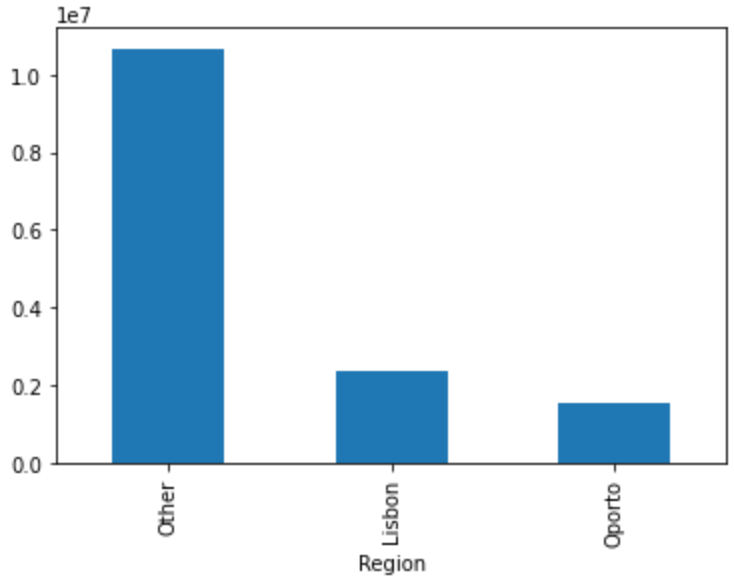
From the graph, we can see that there is a positive linear relationship between variables like Detergents_Paper and Grocery. From the histogram we can see that the price of the whole dataset is right skewed.



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?

Region

Other	10677599
Lisbon	2386813
Oporto	1555088

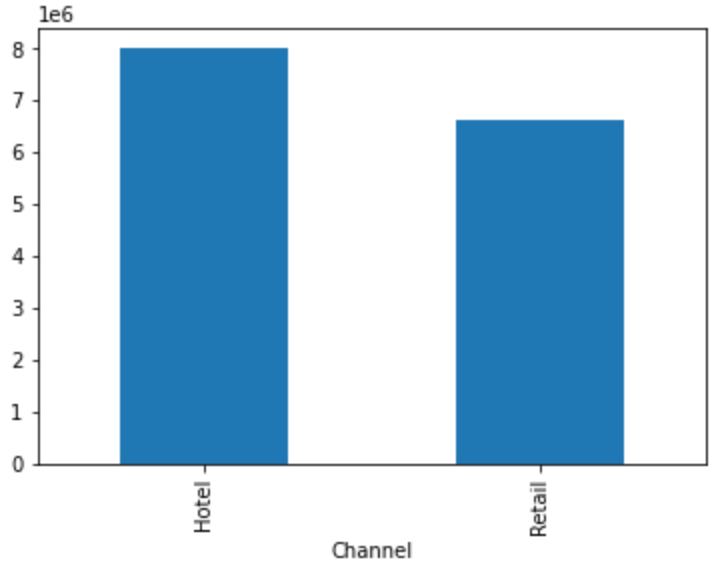


Conclusion of above table and figure

- Other is the highest region spent.
- Oporto is the least region spent.

Channel

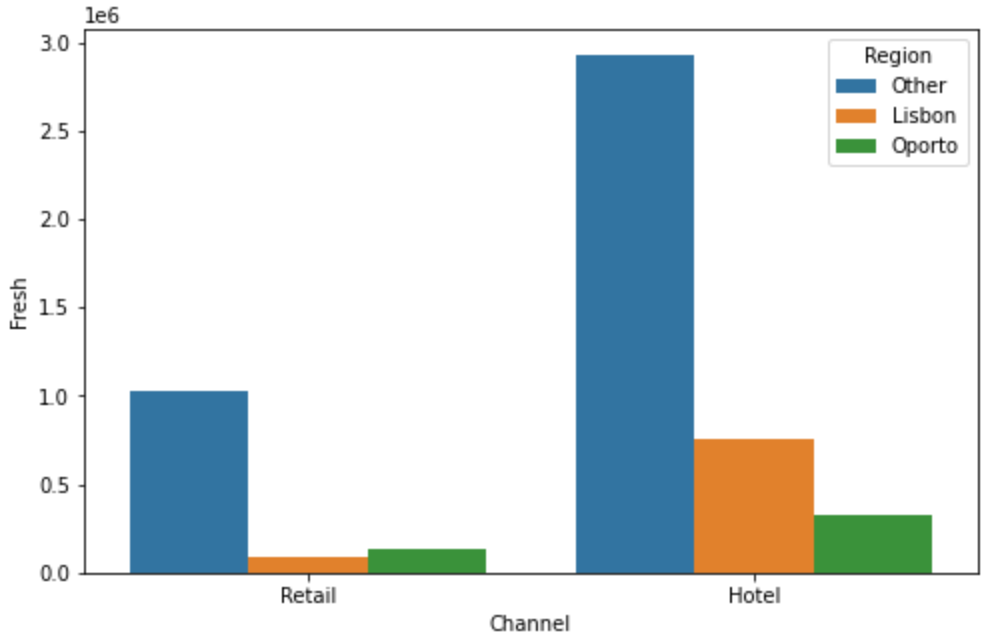
Hotel	7999569
Retail	6619931



Conclusion of above table and figure.

- Hotel is the highest Channel spent.
- Retail is the least Channel spent.

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.

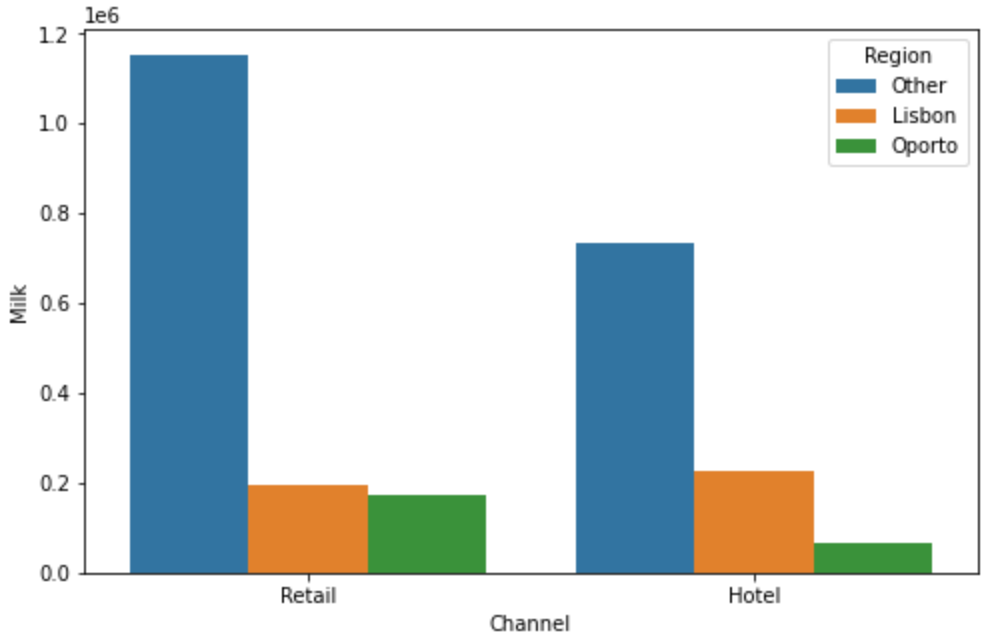


Channel	Hotel	Retail
count	298.000000	142.000000
mean	13475.560403	8904.323944
std	13831.687502	8987.714750
min	3.000000	18.000000
25%	4070.250000	2347.750000
50%	9581.500000	5993.500000
75%	18274.750000	12229.750000
max	112151.000000	44466.000000
CV	1.026428	1.009365

Conclusion of above table and figure.

- Mean and medium are not closed to each other so data is not following normal distribution.

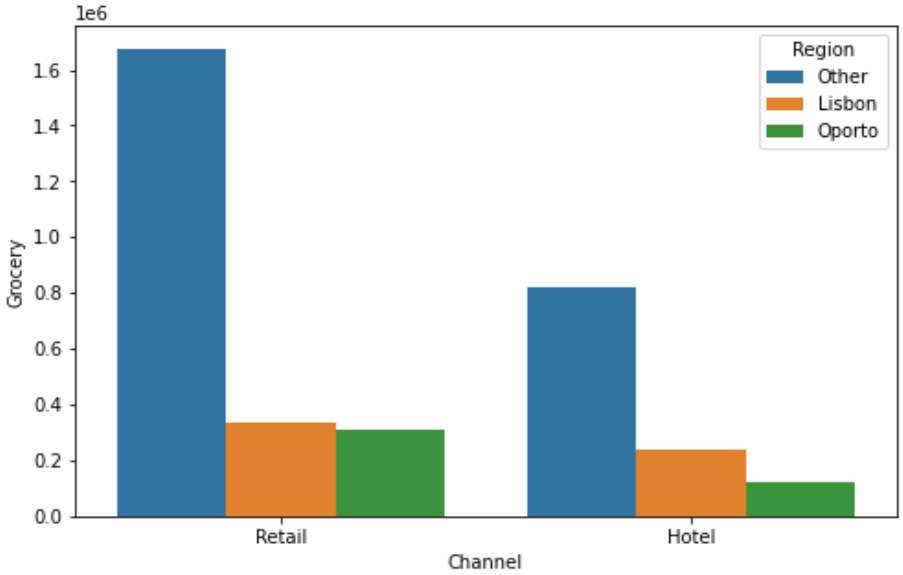
- Customers spending on Fresh is more inconsistent in hotels than retail.
- In the hotel channel other region total spent of customers is more and Oport is least.
- In the Retail channel other region total spent of customers is more and Lisbon is least.
- Total spending of customers in hotels is more than in retail.



Channel	Hotel	Retail
count	298.000000	142.000000
mean	3451.724832	10716.500000
std	4352.165571	9679.631351
min	55.000000	928.000000
25%	1164.500000	5938.000000
50%	2157.000000	7812.000000
75%	4029.500000	12162.750000
max	43950.000000	73498.000000
CV	1.260867	0.903246

Conclusion of above table and figure.

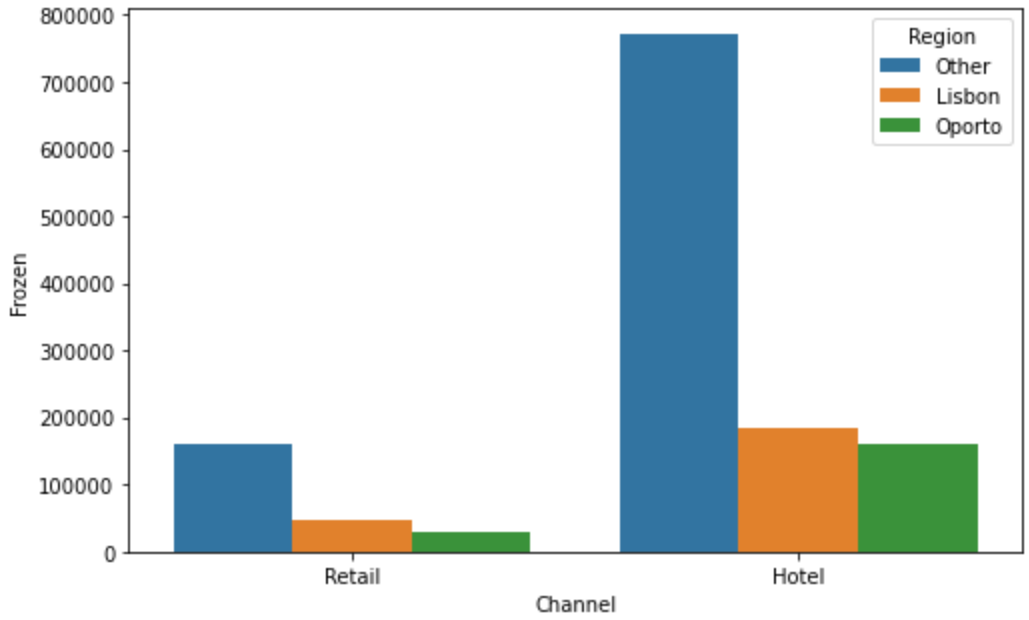
- Mean and medium are not closed to each other so data is not following normal distribution.
- Customers spending on Milk is more inconsistent in hotels than retail.
- In the hotel channel other region total spending of customers is more and Oporto is least.
- In the Retail channel other region total spent of customers is more and Oporto is least.
- Total spending of customers in hotels is less than in retail.



Channel	Hotel	Retail
count	298.000000	142.000000
mean	3962.137584	16322.852113
std	3545.513391	12267.318094
min	3.000000	2743.000000
25%	1703.750000	9245.250000
50%	2684.000000	12390.000000
75%	5076.750000	20183.500000
max	21042.000000	92780.000000
CV	0.894849	0.751543

Conclusion of above table and figure.

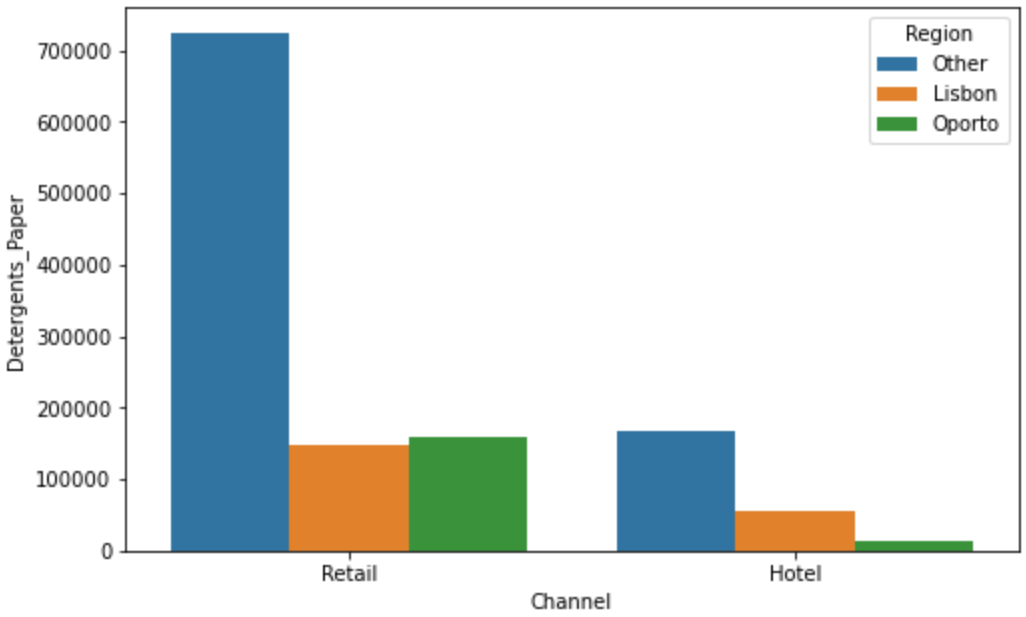
- Mean and medium are not closed to each other so data is not following normal distribution.
- Customers spending on Grocery is more inconsistent in hotels than retail.
- In the hotel channel other region total spending of customers is more and Oporto is least.
- In the Retail channel other region total spending of customers is more and Oporto is least.
- Total spending of customers in hotels is less than in retail.



Channel	Hotel	Retail
count	298.000000	142.000000
mean	3748.251678	1652.612676
std	5643.912500	1812.803662
min	25.000000	33.000000
25%	830.000000	534.250000
50%	2057.500000	1081.000000
75%	4558.750000	2146.750000
max	60869.000000	11559.000000
CV	1.505745	1.096932

Conclusion of above table and figure.

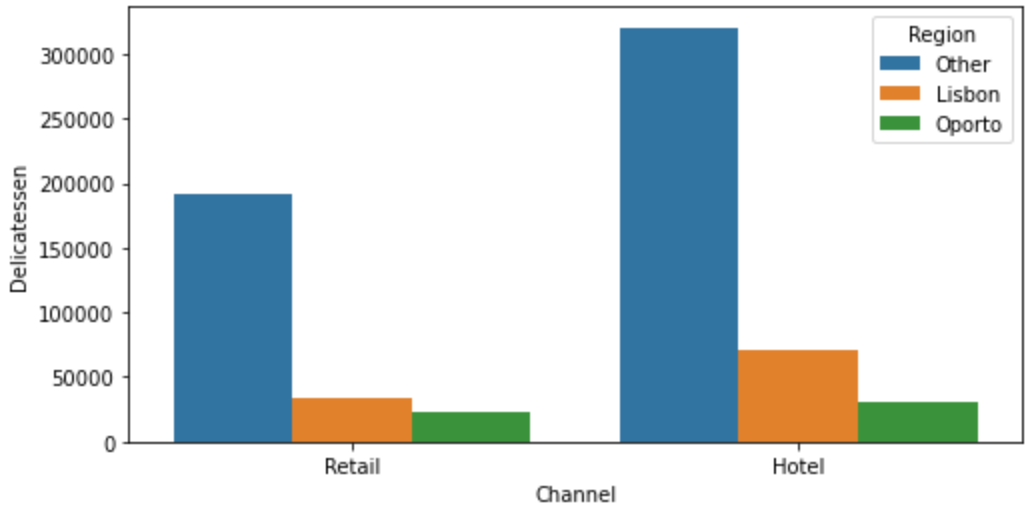
- Mean and medium are not closed to each other so data is not following normal distribution.
- Customers spending on Frozen is more inconsistent in hotels than retail.
- In the hotel channel other region total spending of customers is more and Oporto is least.
- In the Retail channel other region total spending of customers is more and Oporto is least.
- Total spending of customers in hotels is more than in retail.



Channel	Hotel	Retail
count	298.000000	142.000000
mean	790.560403	7269.507042
std	1104.093673	6291.089697
min	3.000000	332.000000
25%	183.250000	3683.500000
50%	385.500000	5614.500000
75%	899.500000	8662.500000
max	6907.000000	40827.000000
CV	1.396596	0.865408

Conclusion of above table and figure.

- Mean and medium are not closed to each other so data is not following normal distribution.
- Customers spending on Detergents_Paper is more inconsistent in hotels than retail.
- In the hotel channel other region total spending of customers is more and Oporto is least.
- In the Retail channel other region total spending of customers is more and Lisbon is least.
- Total spending of customers in hotels is less than in retail.



Channel	Hotel	Retail
count	298.000000	142.000000
mean	1415.956376	1753.436620
std	3147.426922	1953.797047
min	3.000000	3.000000
25%	379.000000	566.750000
50%	821.000000	1350.000000
75%	1548.000000	2156.000000
max	47943.000000	16523.000000
CV	2.222828	1.114267

Conclusion of above table and figure.

- Mean and medium are not closed to each other so data is not following normal distribution.
- Customers spending on Delicatessen is more inconsistent in hotels than retail.
- In the hotel channel other region total spending of customers is more and Oporto is least.
- In the Retail channel other region total spending of customers is more and Oporto is least.
- Total spending of customers in hotels is more than in retail.

Region Vs Fresh :-

Region	Lisbon	Oporto	Other
count	77.000000	47.000000	316.000000
mean	11101.727273	9887.680851	12533.471519
std	11557.438575	8387.899211	13389.213115
min	18.000000	3.000000	3.000000
25%	2806.000000	2751.500000	3350.750000
50%	7363.000000	8090.000000	8752.500000
75%	15218.000000	14925.500000	17406.500000
max	56083.000000	32717.000000	112151.000000
CV	1.041049	0.848318	1.068277

Conclusion of above table.

- Mean and medium are not closed to each other so data is not following normal distribution.
- Customers spend most inconsistent of other and Oporto less.
- Total spending of customers is more Other and Oporto is less.

Region Vs Milk

Region	Lisbon	Oporto	Other
count	77.000000	47.000000	316.000000
mean	5486.415584	5088.170213	5977.085443
std	5704.856079	5826.343145	7935.463443
min	258.000000	333.000000	55.000000
25%	1372.000000	1430.500000	1634.000000
50%	3748.000000	2374.000000	3684.500000
75%	7503.000000	5772.500000	7198.750000
max	28326.000000	25071.000000	73498.000000
CV	1.039815	1.145076	1.327648

Conclusion of above table.

- Mean and medium are not closed to each other so data is not following normal distribution.
- Customers spend mostinconsistent of other and Lisbon less.
- Total spending of customers is more Other and Oporto is less.

Region Vs Grocery

Region	Lisbon	Oporto	Other
count	77.000000	47.000000	316.000000
mean	7403.077922	9218.595745	7896.363924
std	8496.287728	10842.745314	9537.287778
min	489.000000	1330.000000	3.000000
25%	2046.000000	2792.500000	2141.500000
50%	3838.000000	6114.000000	4732.000000
75%	9490.000000	11758.500000	10559.750000
max	39694.000000	67298.000000	92780.000000
CV	1.147670	1.176182	1.207808

Conclusion of above table.

- Mean and medium are not closed to each other so data is not following normal distribution.
- Customers spend most inconsistent of other and Lisbon less.
- Total spending of customers is more Other and Oporto is less.

Region Vs Frozen

Region	Lisbon	Oporto	Other
count	77.000000	47.000000	316.000000
mean	3000.337662	4045.361702	2944.594937
std	3092.143894	9151.784954	4260.126243
min	61.000000	131.000000	25.000000
25%	950.000000	811.500000	664.750000
50%	1801.000000	1455.000000	1498.000000
75%	4324.000000	3272.000000	3354.750000
max	18711.000000	60869.000000	36534.000000
CV	1.030599	2.262291	1.446761

Conclusion of above table.

- Mean and medium are not closed to each other so data is not following normal distribution.
- Customers spend most inconsistent of Oporto and Lisbon less.
- Total spending of customers is more Other and Oporto is less.

Region Vs Detergents_Paper

Region	Lisbon	Oporto	Other
count	77.000000	47.000000	316.000000
mean	2651.116883	3687.468085	2817.753165
std	4208.462708	6514.717668	4593.051613
min	5.000000	15.000000	3.000000
25%	284.000000	282.500000	251.250000
50%	737.000000	811.000000	856.000000
75%	3593.000000	4324.500000	3875.750000
max	19410.000000	38102.000000	40827.000000
CV	1.587430	1.766718	1.630040

Conclusion of above table.

- Mean and medium are not closed to each other so data is not following normal distribution.
- Customers spend most inconsistent of Oporto and Lisbon less.
- Total spending of customers is more Other and Oporto is less.

Region Vs Delicatessen

Region	Lisbon	Oporto	Other
count	77.000000	47.000000	316.000000
mean	1354.896104	1159.702128	1620.601266
std	1345.423340	1050.739841	3232.581660
min	7.000000	51.000000	3.000000
25%	548.000000	540.500000	402.000000
50%	806.000000	898.000000	994.000000
75%	1775.000000	1538.500000	1832.750000
max	6854.000000	5609.000000	47943.000000
CV	0.993008	0.906043	1.994680

Conclusion of above table.

- Mean and medium are not closed to each other so data is not following normal distribution.
- Customers spend most inconsistent of Other and Oporto less.
- Total spending of customers is more Other and Oporto is less.

Dataset skewness

```
Fresh              2.561323
Milk               4.053755
Grocery            3.587429
Frozen             5.907986
Detergents_Paper   3.631851
Delicatessen      11.151586
```

We can see the above skewness of variables , there are right skewness.

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?

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
mean	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000
50%	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000
75%	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000
max	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000
CV	1.053918	1.273299	1.195174	1.580332	1.654647	1.849407

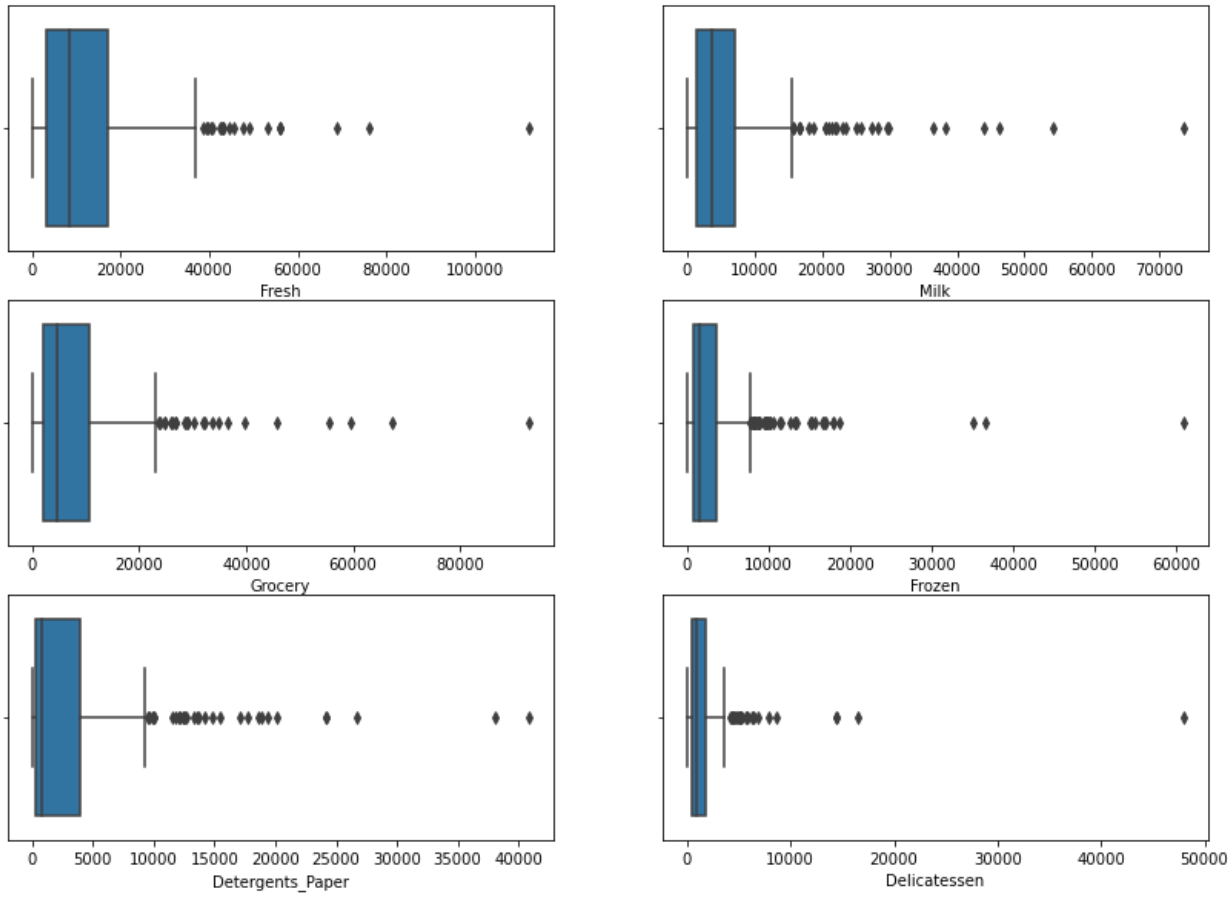
There are 6 variables, so we use CV for comparison of each one inconsistency. Because if we want to compare two or more variables with their relationship mean and std is not fair for it.

By the seeing above figure Fresh is least inconsistent CV(Fresh)= 1.053918 and Delicatessen is most inconsistent CV(Delicatessen)= 1.849407

Below is inconsistent level in descending order :-

Delicatessen>Detergents_Paper>Frozen>Milk>Grocery>Fresh

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



By using Boxplot we can observe that each variables have outliers.

There are lots of points in outside of box plots.

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

By solving this problem I found the points below.

- Hotel channel and other region spent more so keep maitenign this record and trying to more on that.
- Retail channel and oporto region spent less so they should try different methods and techniques to improve these records. They can give cash back offers to improve that.
- For improving sales they can run advertisements on social media based on behaviour of customers for a particular product so that customers engage easily.
- While analysing data I found in the retail channel Lisbon region very less amount spent they can improve these records .

Problem Statement - 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).

Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and find the probability of variables. The data consists of 62 undergraduate students and 14 questions to responses.

Sample of the 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

Dataset has 14 variables and 62 students.

Exploratory Data Analysis

Let us check the types of variables and missing values in the dataset

```
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)
```

- From the above results we can see that there is no missing value present in the dataset.
- There are a total 62 rows and 14 columns in the dataset. Out of 14, 2 columns are of float64 type, 6 columns are of int64 types and 6 columns are of object data types.

Total element in dataset

Total number of elements of the dataset is 868 .

Describe the dataset

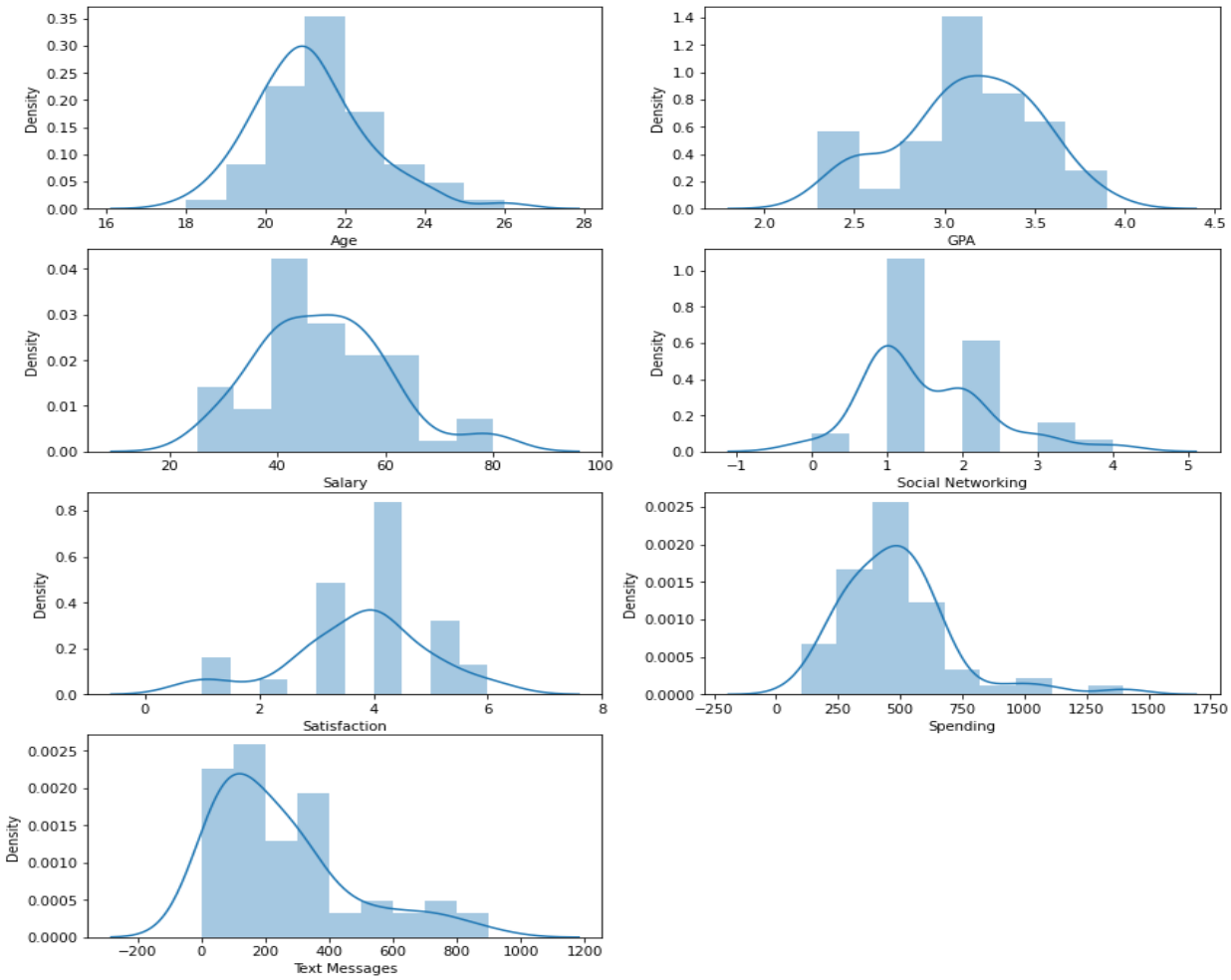
	ID	Age	GPA	Salary	Social Networking	Satisfaction	Spending	Text Messages
count	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000
mean	31.500000	21.129032	3.129032	48.548387	1.516129	3.741935	482.016129	246.209677
std	18.041619	1.431311	0.377388	12.080912	0.844305	1.213793	221.953805	214.465950
min	1.000000	18.000000	2.300000	25.000000	0.000000	1.000000	100.000000	0.000000
25%	16.250000	20.000000	2.900000	40.000000	1.000000	3.000000	312.500000	100.000000
50%	31.500000	21.000000	3.150000	50.000000	1.000000	4.000000	500.000000	200.000000
75%	46.750000	22.000000	3.400000	55.000000	2.000000	4.000000	600.000000	300.000000
max	62.000000	26.000000	3.900000	80.000000	4.000000	6.000000	1400.000000	900.000000

	Gender	Class	Major	Grad Intention	Employment	Computer
count	62	62	62	62	62	62
unique	2	3	8	3	3	3
top	Female	Senior	Retailing/Marketing	Yes	Part-Time	Laptop
freq	33	31	14	28	43	55

from the above picture we can observe below points.

- We can observe the mean, medium, std, min and max of each variable.
- For the objective data types we can observe count, unique, top and freq.

Checking normality and skewness

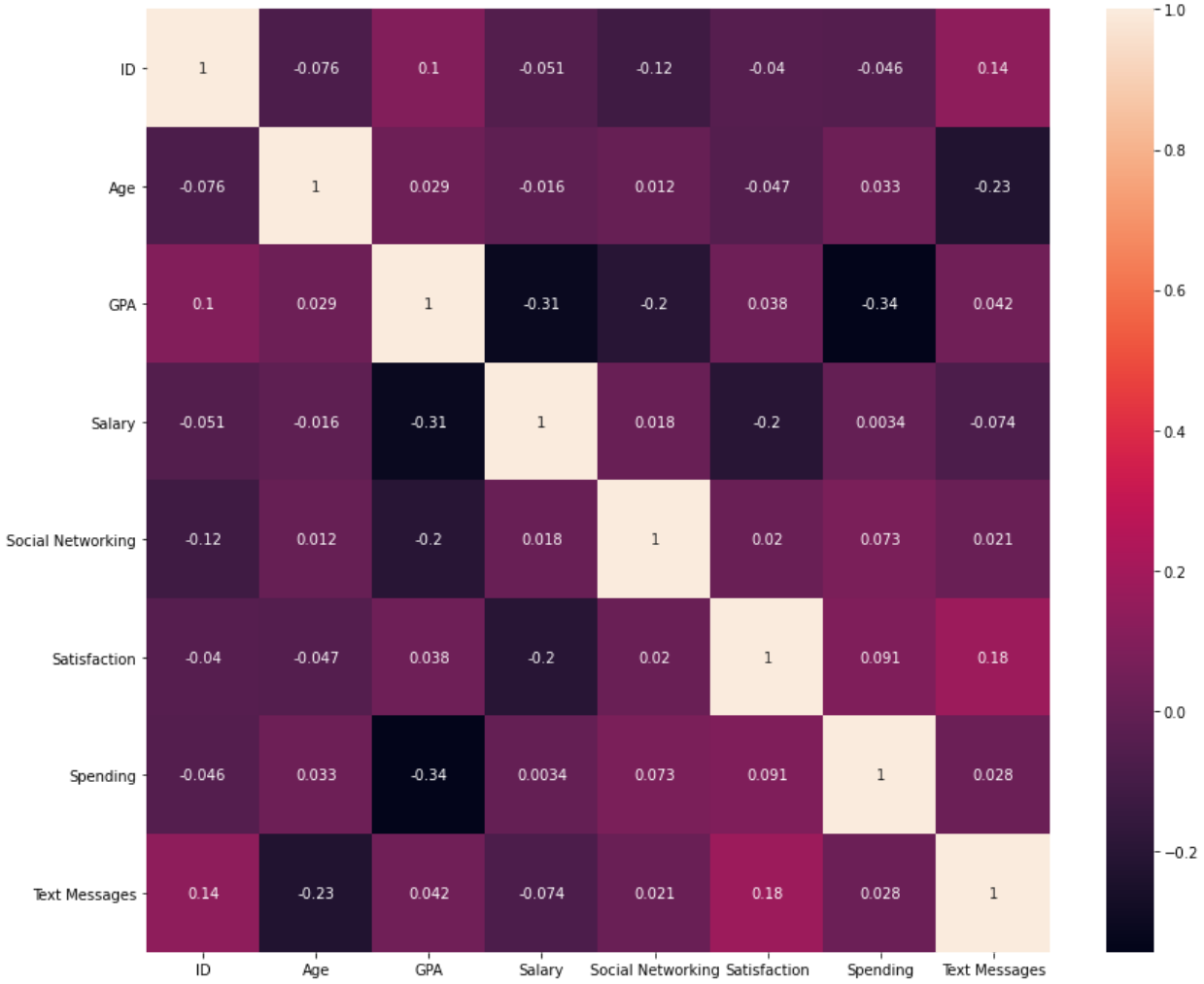


Remarks: We can see above figure data is not normally distributed.

- Age, Salary, Spending, Social Networking and Text Message are right skewness.
- GPA and satisfaction are left skewness

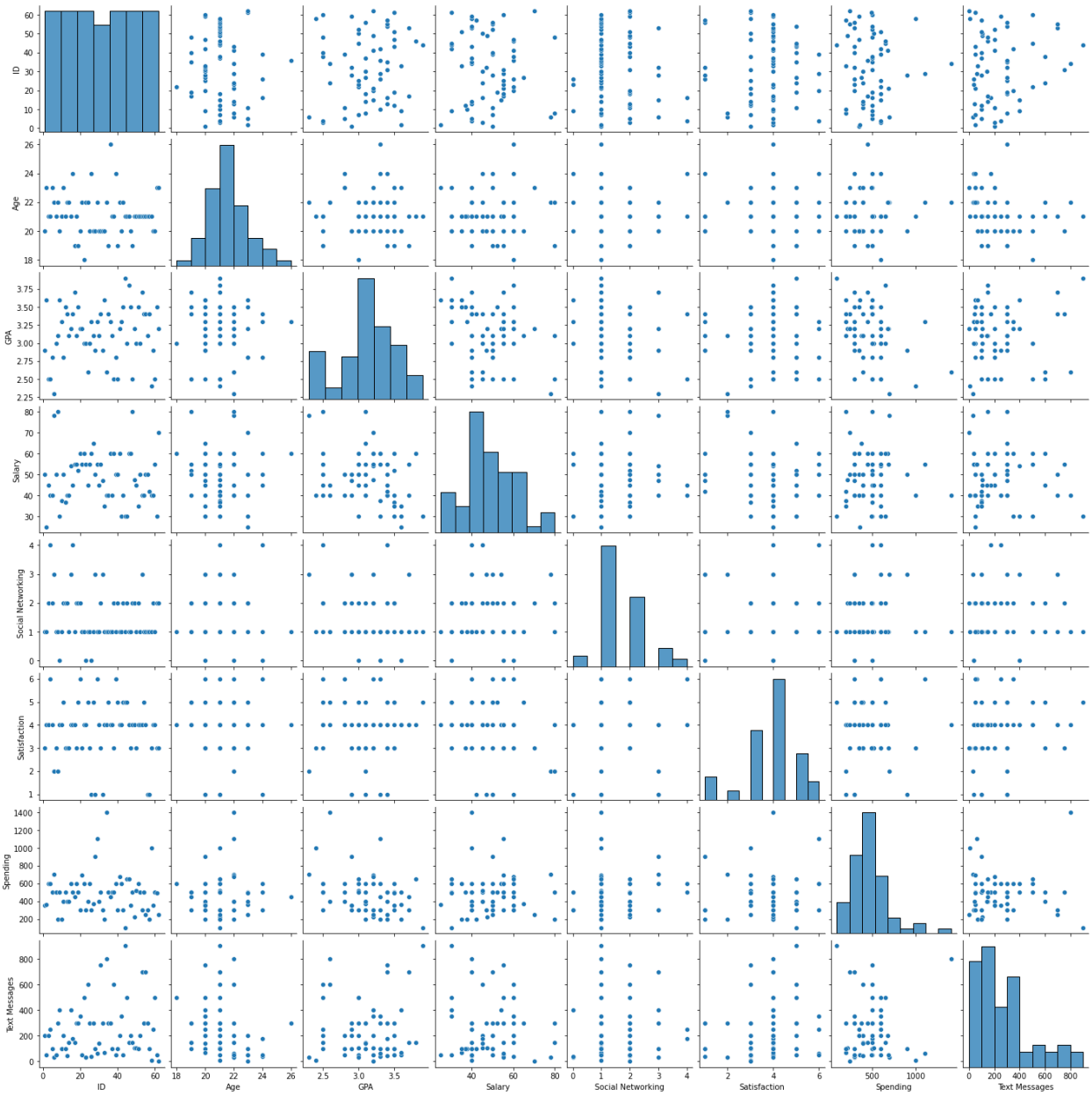
Correlation Plot

From the correlation plot, we can see that various attributes of the students are not-highly correlated to each other. Correlation values near to 1 or -1 are highly positively correlated and highly negatively correlated respectively. Correlation values near to 0 are not correlated to each other.



Pairplot

Pairplot shows the relationship between the variables in the form of scatterplot and the distribution of the variable in the form of histogram.



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 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?

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

The probability that a randomly selected CMSU student will be male is 0.46774193548387094

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

the probability that a randomly selected CMSU student will be female 0.532258064516129

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.

Gender Major	Female	Male	probability
Accounting	3	4	0.137931
CIS	3	1	0.034483
Economics/Finance	7	4	0.137931
International Business	4	2	0.068966
Management	4	6	0.206897
Other	3	4	0.137931
Retailing/Marketing	9	5	0.172414
Undecided	0	3	0.103448

Please follow the probability columns of each Major.

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

Gender Major	Female	Male	probability
Accounting	3	4	0.090909
CIS	3	1	0.090909
Economics/Finance	7	4	0.212121
International Business	4	2	0.121212
Management	4	6	0.121212
Other	3	4	0.090909
Retailing/Marketing	9	5	0.272727
Undecided	0	3	0.000000

Please follow the probability columns of each Major.

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.

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

probability That a randomly chosen student is a male and intends to graduate. 0.27419354838709675
0.27419354838709675

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

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

the probability that a randomly selected student is a female and does NOT have a laptop. 0.06451612903225806

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?

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

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

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

the conditional probability that given a female student is randomly chosen, she is majoring in international business or 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 the graduate intention and being female are independent events?

Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

Condition to be checked:If being female and graduate intention are independent, the P(F ∩ Yes) = P(F)P(Yes)

P(F)=20/40=0.5

P(YES)=28/40=0.7

If being female and graduate intention are independent, the P(F∩YES)=P(F)P(Yes)

P(F∩Yes)=11/40=0.275

P(F)P(Yes)=0.5*(0.7)=0.35 != P(F∩YES)

Hence P(F and Yes) is not equal to P(F)*P(Yes)

So we can say that graduate intention and being female are not independent.

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?

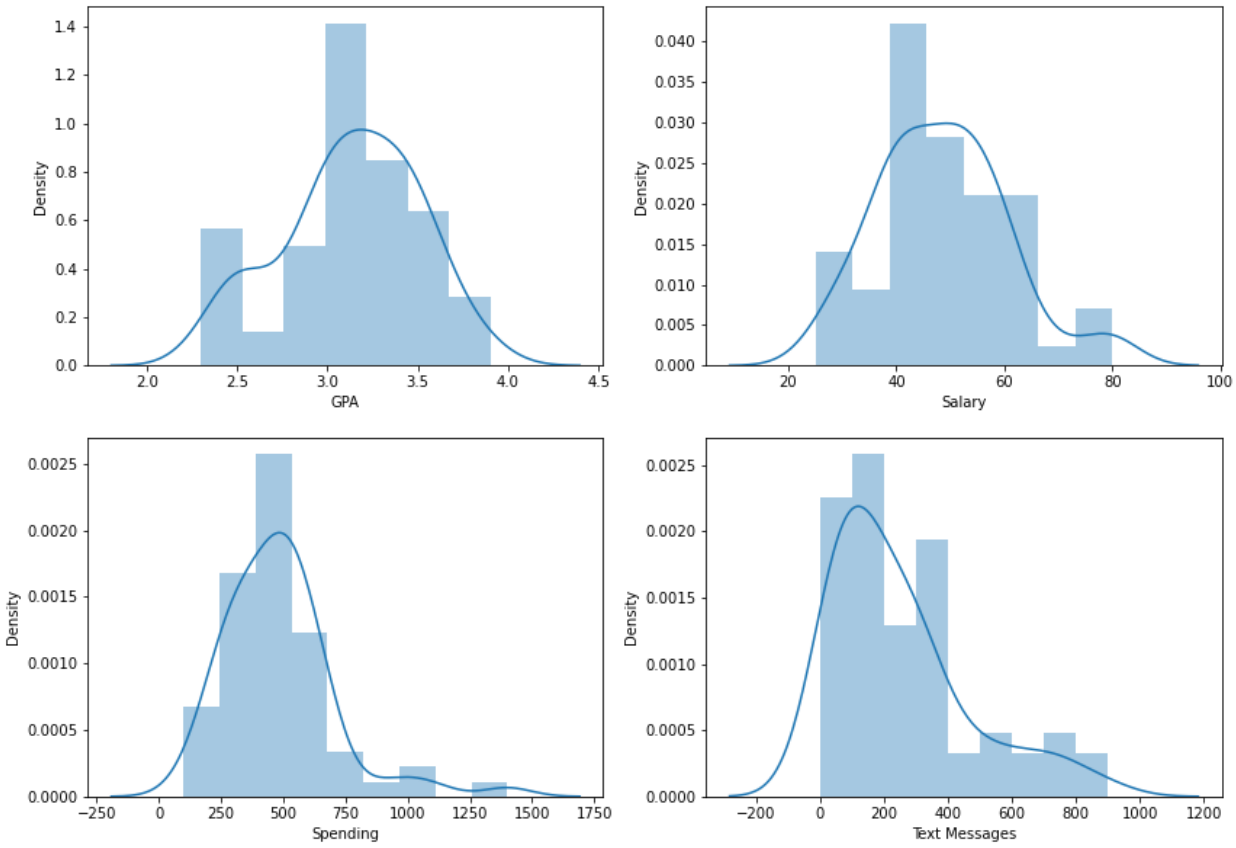
a student is chosen randomly, what is the probability that his/her GPA is less than 3 = 0.27419354838709675

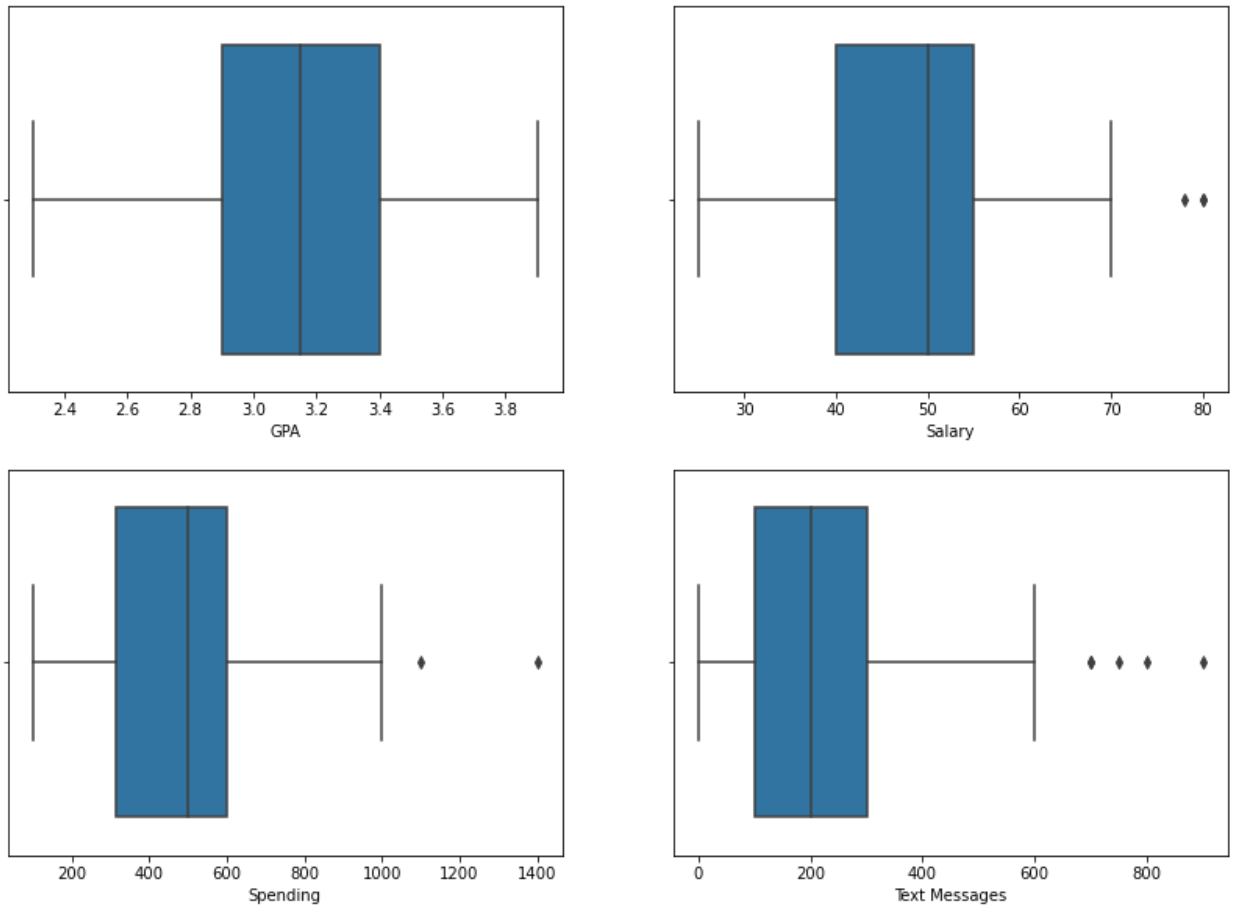
2.7.2. Find the conditional probability that a randomly selected male earns 50 or more. Find the conditional probability that a randomly selected female earns 50 or more.

Find the conditional probability that a randomly selected male earns 50 or more 0.4827586206896552

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





By observing figure and calculating shapiro wilk test we can accept only GPA data is normally distributed and other are not

Problem Statement - 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.

Introduction

The purpose of this whole exercise is to test Hypothesis. Do the exploratory data analysis.

Explore the dataset using central tendency and other parameters.

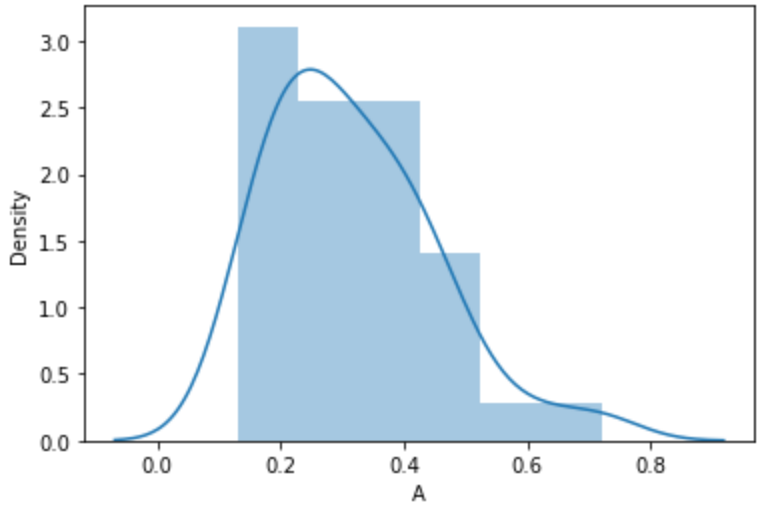
Describe the dataset

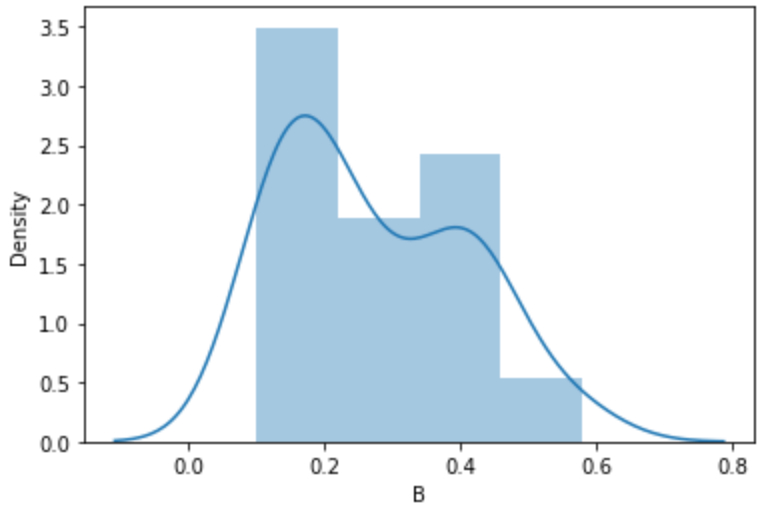
	A	B
count	36.000000	31.000000
mean	0.316667	0.273548
std	0.135731	0.137296
min	0.130000	0.100000
25%	0.207500	0.160000
50%	0.290000	0.230000
75%	0.392500	0.400000
max	0.720000	0.580000

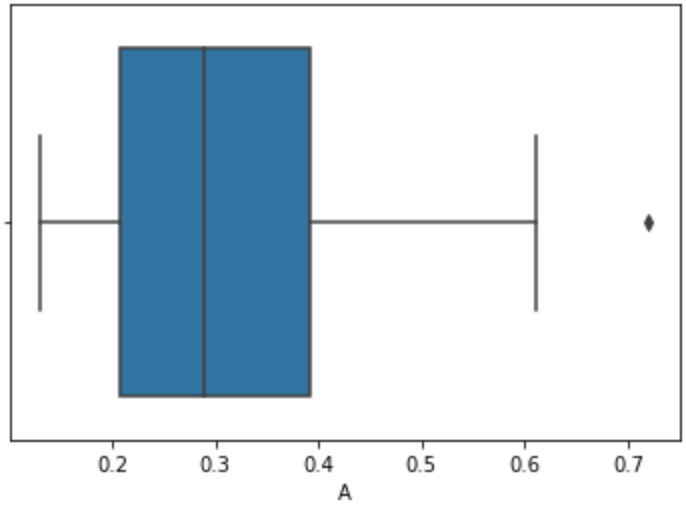
Checking for missing values

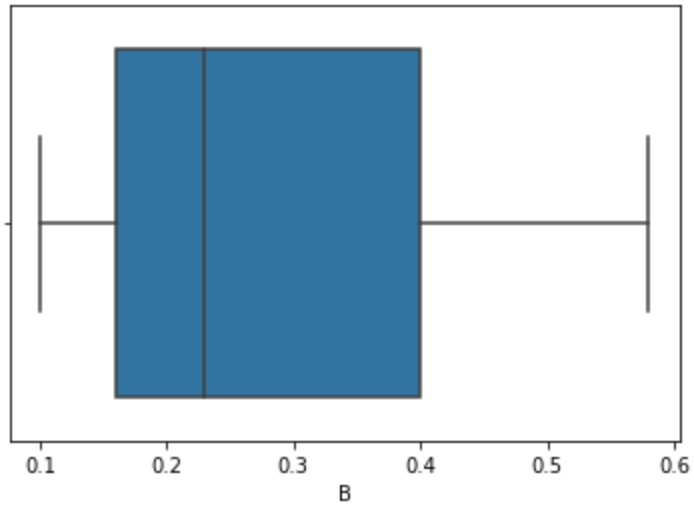
There are 5 missing values

Plot Histograms and boxplot









Remark: As per CLT Assumption we can assume that data is normally distributed

- The level of significance (Alpha) = 0.05.
- But since the population standard deviation (Sigma) is unknown, we have to use a T-test.
- We would prefer One-sided T-test

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.

Hypothesis Formulation:-

Use the T Test (one sample t-test): One-sided for Means:

H0 : mean moisture content <=0.35 (population mean moisture for shingles A less than)

HA : mean moisture content > 0.35 (population mean moisture for shingles A greater than)

In this scenario the p value is 0.07477633144907513 is greater than 0.05, Hence accept the null hypothesis that mean moisture content for shingles A less than

H0 : mean moisture content <=0.35 (population mean moisture for shingles B less than)

HA : mean moisture content > 0.35 (population mean moisture for shingles B greater than)

In this scenario the p value is 0.0020904774003191813 is less than 0.05, Hence reject the null hypothesis that mean moisture content for shingles B greater than

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?

Hypothesis Formulation:-

Use the T-test (Two sample t-test independent t-test): Two-sided

H0 : mean moisture for shingles A = mean moisture for shingles B

HA : mean moisture for shingles A != mean moisture for shingles B

In this scenario the p value is 0.2017496571835306 is greater than 0.05, Hence accept the null hypothesis.