PROJECT REPORT ON STATISTICS FOR DECISION MAKING

Akshaya Parthasarathy

Batch: PGPDSBA_online_July E 2020

Problem 1

Wholesale Customers Analysis

Problem Statement:

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).

Exploratory Data Analysis:

	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

Index is reset to Buyer/Spender column.

Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
436	Hotel	Other	29703	12051	16027	13135	182	2204
437	Hotel	Other	39228	1431	764	4510	93	2346
438	Retail	Other	14531	15488	30243	437	14841	1867
439	Hotel	Other	10290	1981	2232	1038	168	2125
440	Hotel	Other	2787	1698	2510	65	477	52

Dataset has two categorical variables and six continuous variables of integer type.

Check for null values:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 440 entries, 1 to 440
Data columns (total 8 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Channel           440 non-null    object
 1   Region            440 non-null    object
 2   Fresh             440 non-null    int64
 3   Milk              440 non-null    int64
 4   Grocery           440 non-null    int64
 5   Frozen            440 non-null    int64
 6   Detergents_Paper  440 non-null    int64
 7   Delicatessen      440 non-null    int64
dtypes: int64(6), object(2)
memory usage: 30.9+ KB
```

From the above results it is evident that there are no null values in the dataset.

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?

Descriptive Statistics for the dataset:

	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440	440	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
unique	2	3	NaN	NaN	NaN	NaN	NaN	NaN
top	Hotel	Other	NaN	NaN	NaN	NaN	NaN	NaN
freq	298	316	NaN	NaN	NaN	NaN	NaN	NaN
mean	NaN	NaN	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	NaN	NaN	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	NaN	NaN	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	NaN	NaN	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000
50%	NaN	NaN	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000
75%	NaN	NaN	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000
max	NaN	NaN	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000

Out of 440 large retailers, there are

Two channels: Hotel and Retail

Wholesale distributors top contribution between two channels is Hotel with a count of 298.

Three Region: Lisbon, Oporto and Other

Here the top contribution is from Other regions with a count of 316.

Creating a column for 'Total Spend' for each entry which consists of the sum of spending in Fresh, Milk, Grocery, Frozen, Detergents_Paper and Delicatessen.

Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total Spend
1	Retail	Other	12669	9656	7561	214	2674	1338	34112
2	Retail	Other	7057	9810	9568	1762	3293	1776	33266
3	Retail	Other	6353	8808	7684	2405	3516	7844	36610
4	Hotel	Other	13265	1196	4221	6404	507	1788	27381
5	Retail	Other	22615	5410	7198	3915	1777	5185	46100
...	...	...	...	...	...	...	...	...	...
436	Hotel	Other	29703	12051	16027	13135	182	2204	73302
437	Hotel	Other	39228	1431	764	4510	93	2346	48372
438	Retail	Other	14531	15488	30243	437	14841	1867	77407
439	Hotel	Other	10290	1981	2232	1038	168	2125	17834
440	Hotel	Other	2787	1698	2510	65	477	52	7589

440 rows × 9 columns

Region and Total Spend: Groupby Region and calculate the sum of total spend

```
         Total Spend
Region
Oporto       1555088
Lisbon       2386813
Other       10677599
```

Other Region seems to spend more, and Oporto seems to spend less.

Channel and Total Spend: Groupby Channel and calculate the sum of total spend

```
          Total Spend
Channel
Retail        6619931
Hotel         7999569
```

Here, we can say that Hotel spends more, and retail shops spend less.

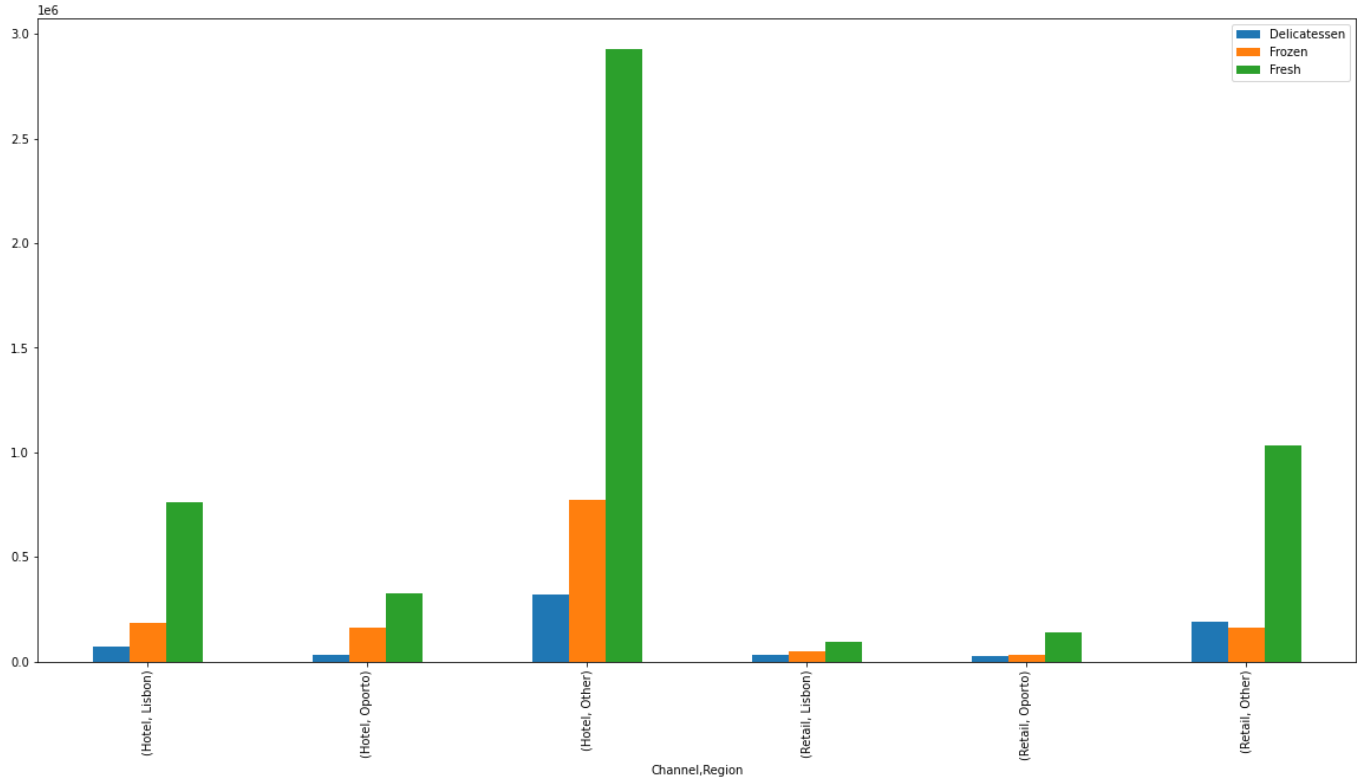
Region and Channel with Total Spend:

Region	Channel	Total Spend
Oporto	Hotel	719150
	Retail	835938
Lisbon	Retail	848471
	Hotel	1538342
Other	Retail	4935522
	Hotel	5742077

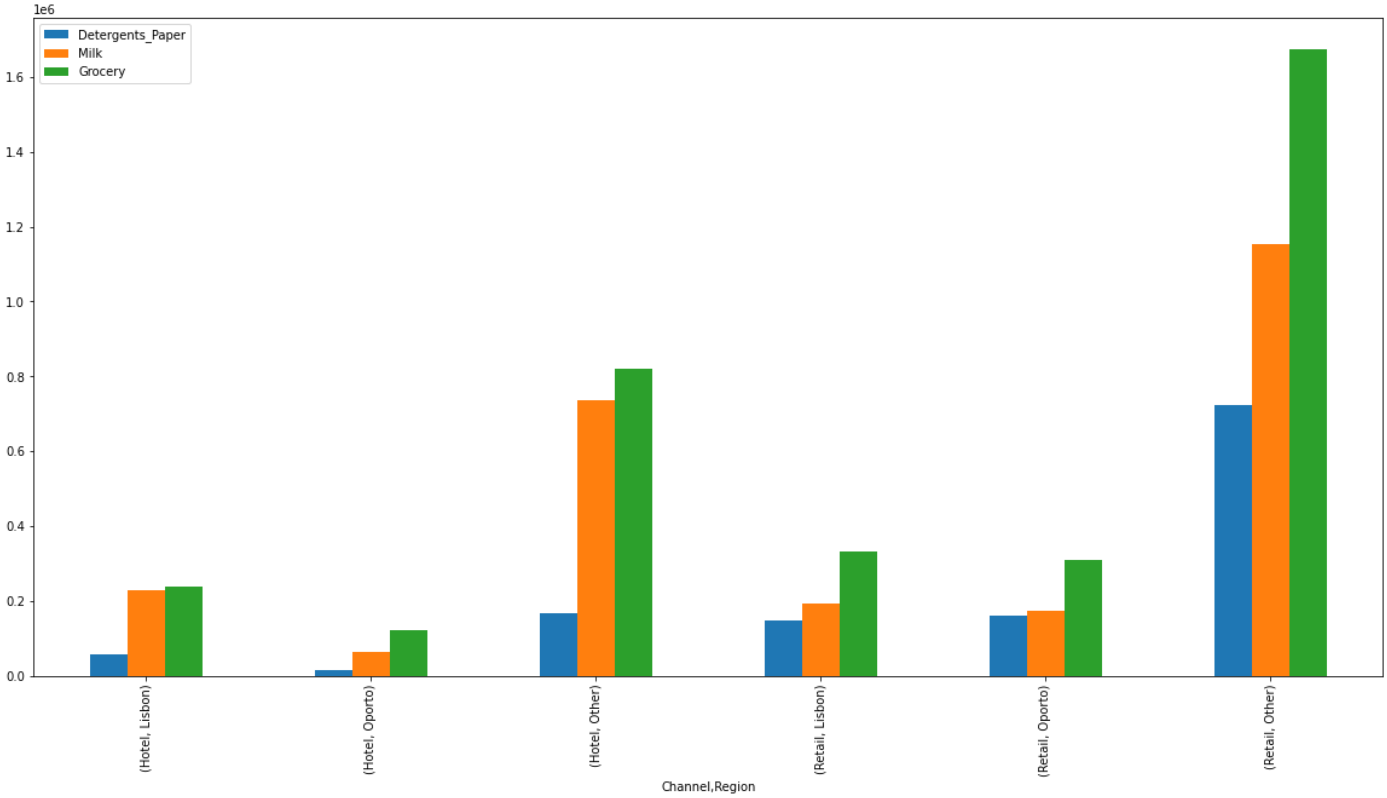
In Other region, Hotel channel spends more and in Oporto Region, Hotel channel spends less.

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

In order to show the behaviour of different varieties of items across Region and Channel, I have taken the total spending of each items to show the pattern.



From the above, we can say that the annual spending on items like Delicatessen, Frozen and Fresh is more in Hotel channel of all regions compared to the Retail channel. Among the items, Hotel seems to spend more on Fresh annually.



From here, we can say that the annual spending on items like Grocery, Milk and Detergents_Paper is more in Retail channel of all regions compared to the Hotel channel. Among the above items, in Retail outlets, Grocery has the highest annual spending.

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

	mean	std	CoV
Fresh	12000.297727	12647.328865	1.053918
Milk	5796.265909	7380.377175	1.273299
Grocery	7951.277273	9503.162829	1.195174
Frozen	3071.931818	4854.673333	1.580332
Detergents_Paper	2881.493182	4767.854448	1.654647
Delicatessen	1524.870455	2820.105937	1.849407
Total Spend	33226.136364	26356.301730	0.793240

Coefficient of Variation is used to measure the variability which is the relative standard deviation given by the formula

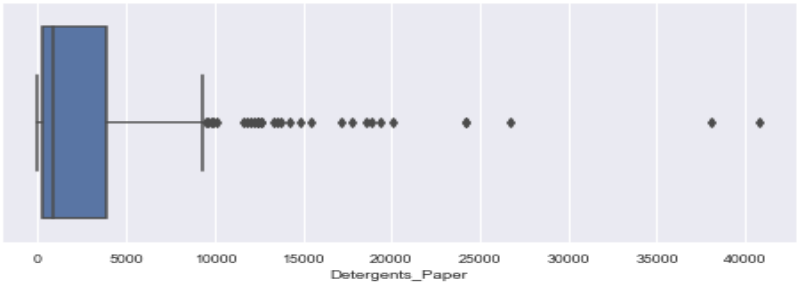
CoV = Ratio of Standard Deviation to mean

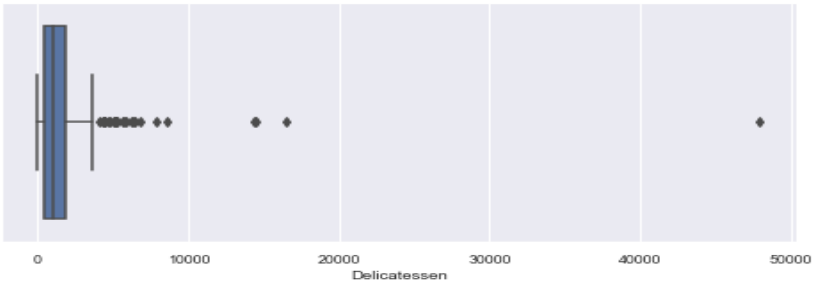
The item with least CoV is more consistent.

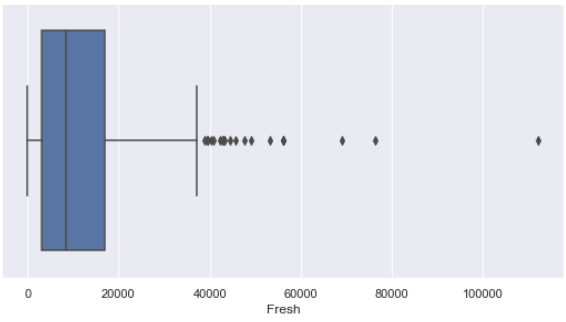
This table clearly shows that all the items are inconsistent. Among this, Fresh shows the least inconsistent behaviour and Delicatessen shows the most inconsistent behaviour.

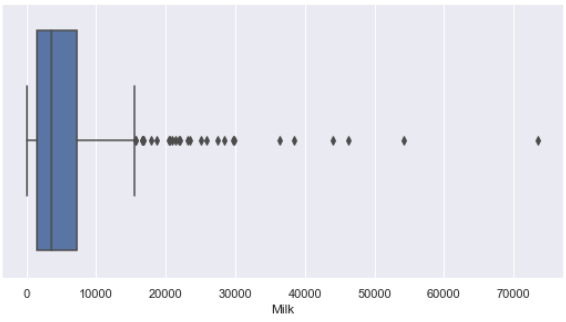
1.4 Are there any outliers in the data?

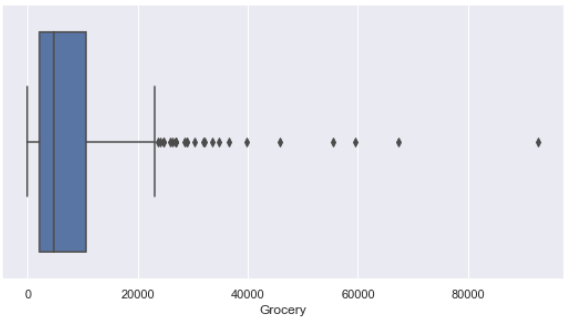
Boxplot is used to identify the outliers.

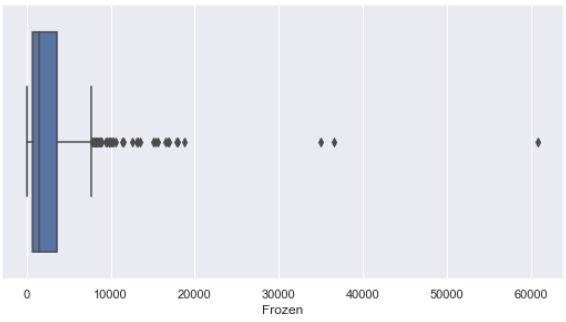












All the variables have outliers.

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

From the above analysis, we can infer that,

1. Wholesale distributors spending on different items is not consistent among region and channel. The annual spending for the items in 'Other' Region is comparatively more than Lisbon and Oporto.
2. Hotel channel spends more in the items like Frozen, Fresh and Delicatessen. Retail channel spends more in the items like Grocery, Milk and Detergents_Paper.
3. Delicatessen is the least spent item across channel and region. Fresh category is the most spent item across region and channel. It can be said that people prefer buying/spending on fresh items rather than ready-to-eat items (delicatessen).
4. In order to compensate for the inconsistency or to reduce it, the distributor can expand more in the retail channel of regions Lisbon and Oporto.
5. For treating the outliers in all the items, maybe the distributor can set a cap on annual spending of items across the region and channel. But the spending may also be due to the large population in that specific region or due to a greater number of Hotels rather than retail channel. So, it is inclusive from the given data on the possibility of the situation.

Problem 2

Survey of students

Problem Statement:

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).

Exploratory Data Analysis:

	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

Index is reset to ID column.

ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
58	Female	21	Senior	International Business	No	2.4	Part-Time	40.0	1	3	1000	Laptop	10
59	Female	20	Junior	CIS	No	2.9	Part-Time	40.0	2	4	350	Laptop	250
60	Female	20	Sophomore	CIS	No	2.5	Part-Time	55.0	1	4	500	Laptop	500
61	Female	23	Senior	Accounting	Yes	3.5	Part-Time	30.0	2	3	490	Laptop	50
62	Female	23	Senior	Economics/Finance	No	3.2	Part-Time	70.0	2	3	250	Laptop	0

Check for null values:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 62 entries, 1 to 62
Data columns (total 13 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Gender              62 non-null     object
 1   Age                 62 non-null     int64
 2   Class               62 non-null     object
 3   Major               62 non-null     object
 4   Grad Intention      62 non-null     object
 5   GPA                 62 non-null     float64
 6   Employment          62 non-null     object
 7   Salary              62 non-null     float64
 8   Social Networking   62 non-null     int64
 9   Satisfaction        62 non-null     int64
 10  Spending            62 non-null     int64
 11  Computer            62 non-null     object
 12  Text Messages       62 non-null     int64
dtypes: float64(2), int64(5), object(6)
memory usage: 6.8+ KB
```

From the above results it is evident that there are no null values in the dataset.

Descriptive Statistics for the dataset:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Gender	62	2	Female	33	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Age	62	NaN	NaN	NaN	21.129	1.43131	18	20	21	22	26
Class	62	3	Senior	31	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Major	62	8	Retailing/Marketing	14	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Grad Intention	62	3	Yes	28	NaN	NaN	NaN	NaN	NaN	NaN	NaN
GPA	62	NaN	NaN	NaN	3.12903	0.377388	2.3	2.9	3.15	3.4	3.9
Employment	62	3	Part-Time	43	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Salary	62	NaN	NaN	NaN	48.5484	12.0809	25	40	50	55	80
Social Networking	62	NaN	NaN	NaN	1.51613	0.844305	0	1	1	2	4
Satisfaction	62	NaN	NaN	NaN	3.74194	1.21379	1	3	4	4	6
Spending	62	NaN	NaN	NaN	482.016	221.954	100	312.5	500	600	1400
Computer	62	3	Laptop	55	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Text Messages	62	NaN	NaN	NaN	246.21	214.466	0	100	200	300	900

Class category: University consists of three class – Sophomore, Junior and Senior

Major: Accounting, CIS, Economics/Finance, International Business, Management, Other, Retailing/Marketing and Undecided

Grad Intention: Yes, No and undecided

Employment: Part-time, Full-time and unemployed

Computer: Laptop, Desktop and Tablet

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

NOTE: All the answers to be below questions are rounded off to two decimal points.

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?

```
Female    33
Male      29
Name: Gender, dtype: int64
Probability of randomly selected CMSU student will be male is: 29/62 =
0.47
```

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

```
Probability of randomly selected CMSU student will be female is:
1 – Probability(Male) = 0.53
```

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.

```
Out of 29 male students, the conditional probability of different major
s in CMSU is as follows:

P (Accounting | Male):              0.14
P (CIS | Male):                     0.03
P (Economics/Finance | Male):       0.14
P (International Business | Male):  0.07
P (Management | Male):              0.21
P (Other | Male):                   0.14
P (Marketing/Retailing | Male):     0.17
P (Undecided | Male):               0.1
```

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

```
Out of 33 female students, the conditional probability of different
majors in CMSU is as follows:

P (Accounting | Female):            0.09
P (CIS | Female):                   0.09
P (Economics/Finance | Female):     0.21
P (International Business | Female):0.12
P (Management | Female):            0.12
P (Other | Female):                 0.09
P (Marketing/Retailing | Female):   0.27
P (Undecided | Female):             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.

```
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:  17/62 = 0.27
```

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
Probability that a randomly selected student is a female and does not
have a laptop:  4/62 = 0.06
```

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 either a male or has full-time employment?

```
Employment  Full-Time  Part-Time  Unemployed
Gender
Female              3         24           6
Male                7         19           3
Probability that randomly chosen student is a male:  0.47
Probability that randomly chosen student has full-time employment: 0.16
Probability that randomly chosen student is a male and has a full-time
employment: 0.11

Probability that a randomly chosen student is either a male or has a fu
ll-time employment:  0.47 + 0.16 – 0.11 = 0.52 (Addition Rule)
```

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

```
Female students selecting international business or management:  8
Total female students:  33

Conditional probability that given a female student is randomly chosen,
she is majoring in international business or management: 8/33 = 0.24
```

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

```
Grad Intention  No  Yes
Gender
Female           9   11
Male             3   17

Event A: Graduate intention:  0.7
Event B: Being a female:  0.5
Grad intention and female:  0.275
Probability of event A * Probability of event B:  0.35
```

Since P(A) * P(B) is not equal to P(A and B), the event graduate intention and being a 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?

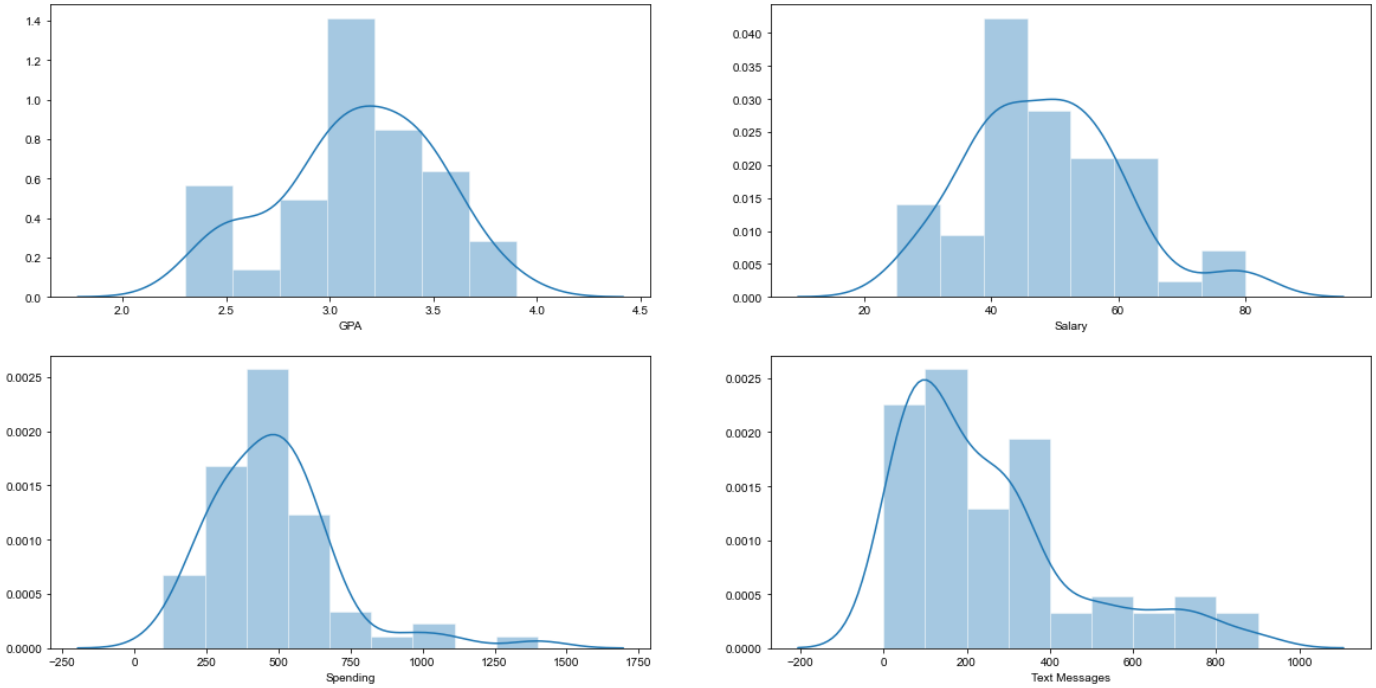
```
Probability that randomly chosen student GPA is less than 3: 17/62 =
0.27
```

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.

```
P (Earns 50 or more | Male):          14/29 =  0.48
P (Earns 50 or more | Female):        18/33 =  0.55
```

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.

Distplot is used to show the distribution of the data.



From the above plot, we can see that GPA is approximately symmetrical following a normal distribution while the other variables are right skewed distributions. Confirming the above using shapiro test for normal distribution and determine the skewness of the variables.

Results of the shapiro test and their corresponding p-values.

P-value > alpha, Normal distribution

P-value =< alpha, Not normally distributed

```
p ( 0.112 ) > alpha (0.05), hence the continuous variable GPA is normal
ly distributed
p ( 0.028 ) < alpha (0.05), hence the continuous variable Salary is not
normally distributed
p ( 0.000 ) < alpha (0.05), hence the continuous variable Spending is
not normally distributed
p ( 0.000 ) < alpha (0.05), hence the continuous variable Text messages
is not normally distributed
```

Skewness of the data can also be used to determine the distribution.

- Highly skewed: >= +1 and =< -1
- Moderately skewed: (-1 to -0.5) and (0.5 to 1)
- Approximately symmetrical: -0.5 to 0.5

```
Skew of GPA: -0.31
Skew of Salary: 0.52
Skew of Spending: 1.55
Skew of Text Messages: 1.26
```

Out of the four continuous variables, Only GPA follows normal distribution.

```
Approximately 68.27 % of the data falls within one standard deviation
from the mean
Approximately 95.45 % of the data falls within two standard deviation
from the mean
Approximately 99.73 % of the data falls within three standard deviation
from the mean
```

Problem 3

Shingles A & B

Problem statement:

An important quality characteristic used by the manufacturers of ABC asphalt shingles i s the amount of moisture the shingles contain when they are packaged. Customers ma y feel that they have purchased a product lacking in quality if they find moisture and w et shingles inside the packaging. In some cases, excessive moisture can cause the gran ules attached to the shingles for texture and colouring purposes to fall off the shingles r esulting in appearance problems. To monitor the amount of moisture present, the com pany conducts moisture tests. A shingle is weighed and then dried. The shingle is then r eweighed 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 t he mean moisture content is less than 0.35 pound per 100 square feet. The file (A & B s hingles.csv) includes 36 measurements (in pounds per 100 square feet) for A shingles a nd 31 for B shingles

Exploratory data Analysis:

	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

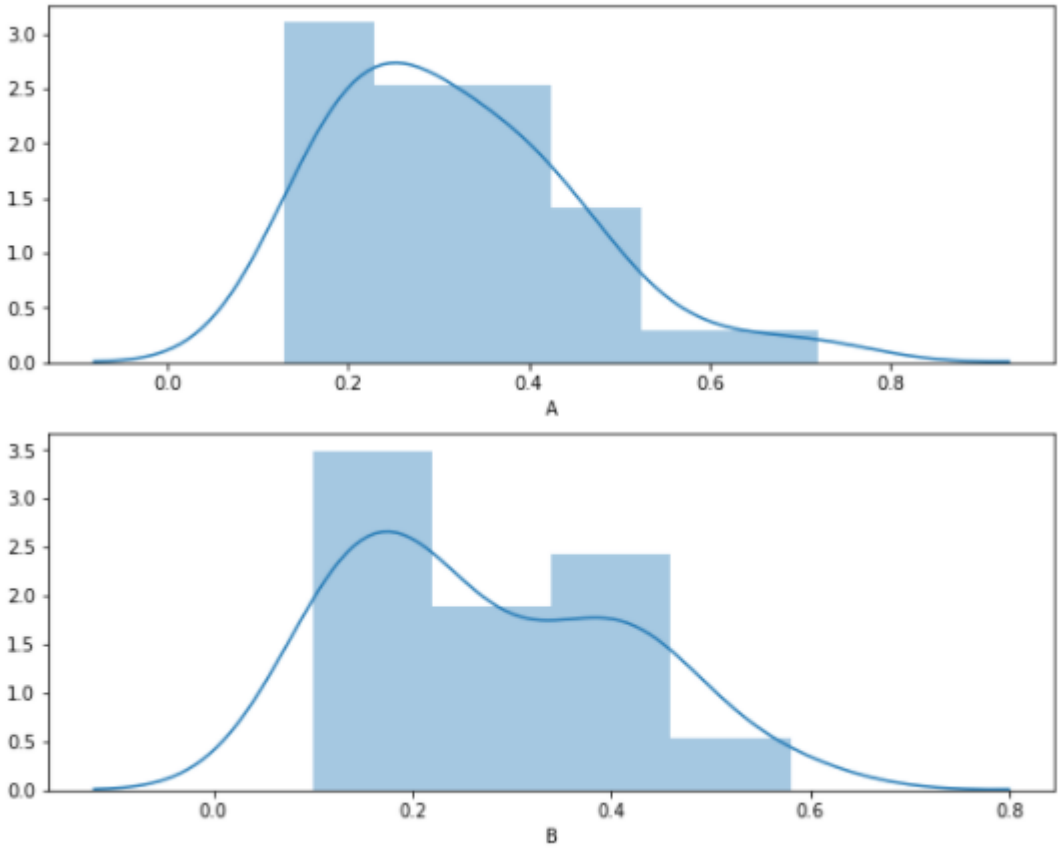
Descriptive statistics for the data set:

	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

From looking at the descriptive statistics of the data, we can say that both the shingles are within permissible limits, since the mean of both (A - 0.31 and B - 0.27) is less than

0.35 pound per 100 square feet. But we cannot be sure whether it is statistically significant.

Distribution plot for the dataset:



Both the distributions look normal.

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.

To show whether both types of shingles are within the permissible limits (less than 0.35 pound per 100 square feet). We perform one sample T-test for both the shingle types.

One sample T-test Assumptions:

- The data is continuous and follows normal distribution.

- Population standard deviation is not known.

- n (sample size) can be small or large.

- The samples are randomly selected, independent and come from a normally distributed population with unknown but equal variances.

STEP 1: Define Null and Alternate Hypothesis

In testing the mean moisture contents in both types of shingles are within the permissible limits.

- Null hypothesis states that mean moisture contents in both types of shingles are not within the permissible limits, μ is equal to 0.35.
- Alternative hypothesis states that the mean moisture content is less than 0.35 pound per 100 square feet., μ is less than 0.35.

$H0$: μ = 0.35
HA: μ < 0.35

STEP 2: Decide the significance level

Here we select α = 0.05.

STEP 3: Identify the test statistic:

We do not know the population standard deviation and n = 36 for shingles A and n = 31 for shingles B. So, we use the one sample t test for each shingle separately and the *t*STAT test statistic.

STEP 4: Calculate P-value and test statistic:

```
One sample t test for Shingle A
t statistic: -1.4735046253382809 p value: 0.07477633144907479
```

```
One sample t test for Shingle B
t statistic: -3.1003313069986995 p value: 0.0020904774003191826
```

STEP 5: Decide to reject or accept Null Hypothesis

For Shingles A,

Level of significance: 0.05
We have no evidence to reject the null hypothesis since p value > Level of significance
Our one-sample t-test p-value for Shingle A = 0.07477633144907479

Here, p value is 0.07477633 and it is greater than 5% level of significance

So, the statistical decision is failing to reject the null hypothesis at 5% level of significance.

Hence, at 95% confidence level, there is no sufficient evidence to prove that mean moisture contents in shingle A is within the permissible limits.

For Shingles B,

Level of significance: 0.05
We have evidence to reject the null hypothesis since p value < Level of significance
Our one-sample t-test p-value for Shingle B = 0.0020904774003191826

Here, p value is 0.002090477 and it is lesser than 5% level of significance

So, the statistical decision is to reject the null hypothesis.

Hence, we reject Null Hypothesis and conclude that there is sufficient evidence to prove that mean moisture contents in shingle B is within the permissible limits with 0.2% of being wrong.

3.2 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 show that the population means for shingles A and B are equal, we perform Two sample T-test with independent samples.

Two sample (Independent) T-Test Assumptions:

- The data should be continuous and follow the normal probability distribution.

- Both samples have identical variance.

- The two samples are independent. There is no relationship between the individuals in on e sample as compared to the other (as there is in the paired t-test).

- Both samples are simple random samples from their respective populations. Every sampl e in the population has an equal probability of being selected in the sample.

- Population Standard deviation not known.

STEP 1: Define null and alternate hypothesis

In testing the population means for shingles A and B, the null hypothesis states that the population means for shingles A and B are equal, μAequals μB. The alternative hypothesis states that the population means for shingles A and B are not equal, μA is not equal to μB.

- $H0$: μA - μB = 0; μA = μB
- HA: μA - μB ≠ 0; μA ≠ μB

It is a two tailed test.

STEP 2: Decide the significance level

Here we select α = 0.05.

STEP 3: Identify the test statistics:

- We have two samples and we do not know the population standard deviation.
- Sample sizes for both samples are different. (n=36 for Shingles A and n=31 for Shingles B)
- So, using the t distribution and the $tSTAT$ test statistic for two sample independent test.

STEP 4: Calculate the p-value and test statistic

```
Mean of A shingles 0.3166666666666666
Mean of B shingles 0.2735483870967742
STD of A shingles 0.13573082605973166
STD of B shingles 0.13729647694185443

Two sample independent t test for Shingle A&B
t statistic: 1.2896282719661123 p value: 0.2017496571835306
```

STEP 5: Decide to accept or reject Null:

```
Two-sample t-test p-value= 0.2017496571835306
We do not have enough evidence to reject the null hypothesis in favour
of alternative hypothesis
We conclude that population means of both the shingles are same.
```

Reservations about the test performed:

- If we compare the means of the two-sample distribution, we see that there is slight difference in the means, but statistically it does not hold good

- The smaller mean of shingle B can be due to the comparatively smaller number of measurements.

- The variance of the both the shingles are approximately equal.

- We are unsure of the sampling error present in the data.

- Even though the population means of both are equal, we can say from the above one sample t-test performed that only in Shingle B measurements, mean moisture content is within permissible limits.