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 spent the most? Which Region and which Channel spent the least?

Solution: We used describe() function to generate descriptive statistics on the contents of the data frame to show the central tendency, shape, distribution, and dispersion of variables. Examining descriptive statistics is the very first task in any quantitative data analysis.

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

```
In [8]: data["Channel"].value_counts()

Out[8]: Hotel     298
        Retail    142
        Name: Channel, dtype: int64
```

```
In [9]: data["Region"].value_counts()

Out[9]: Other     316
        Lisbon     77
        Oporto     47
        Name: Region, dtype: int64
```

Conclusions:

1. There are two unique values in "channel "and three unique values in "region".
2. 440 counts are there in every variable.

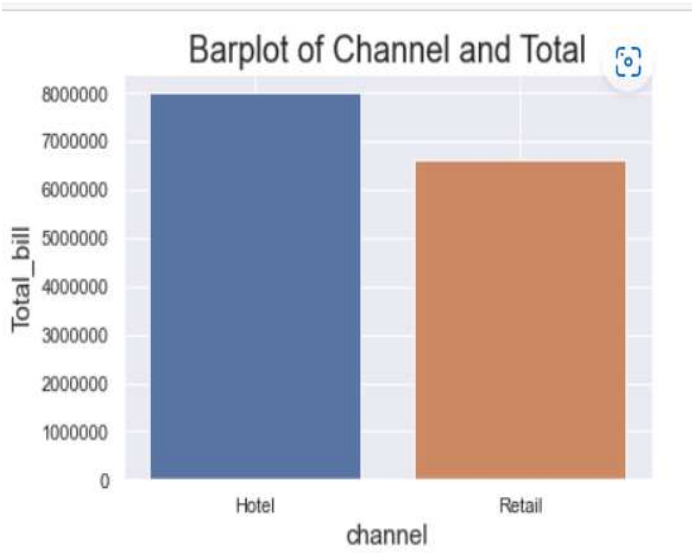
3. The mean of all the variables is different.
4. The minimum value is 1 of Buyer/Spender,3 of Fresh, Grocery, Detergents_Paper and Delicatessen, and 55 of milk.
5. Standard deviation is maximum of Fresh variable and minimum is of Buyer/spender.
6. There is total 440 counts in channel variable out of which 298 values are of hotel and 142 values are of retail.
7. There is total 440 counts in region variable out of which 316 values are of other variable,77 values are of Lisbon variable and 47 values are of Oporto variable.
8. The maximum value is 112151 of Fresh variable and 440 of Buyer/Spender.
9. Interquartile range (IQR) is the range of values that resides in the middle of the scores**.**

IQR = interquartile range=Q3-Q1

where, Q3 = 3rd quartile or 75th percentile and Q1 = 1st quartile or 25th percentile

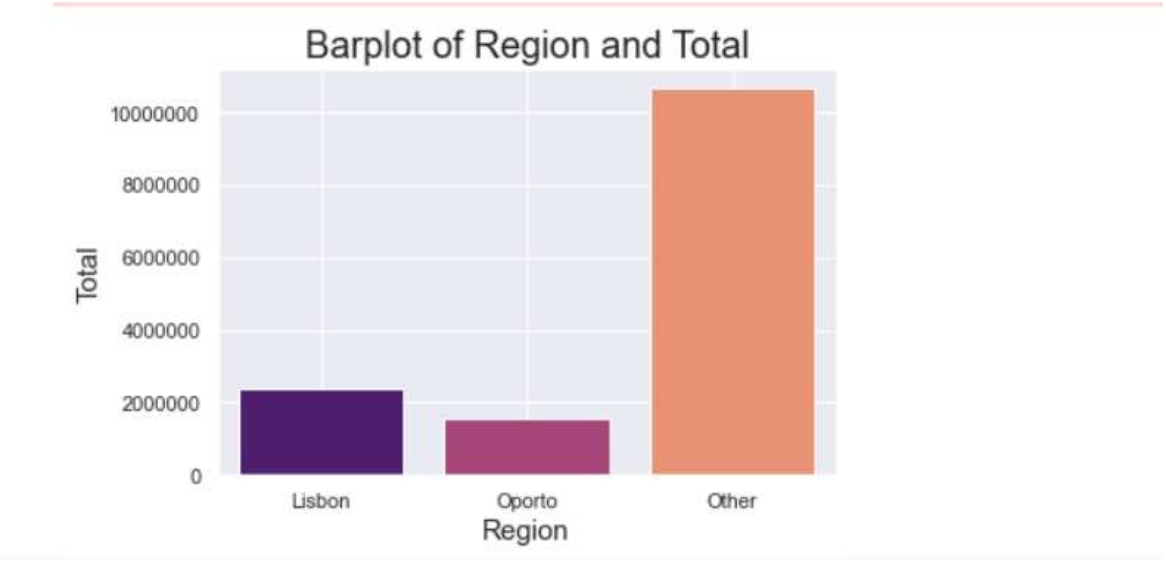
We've been asked which region and which channel spent the most and the least.

CHANNEL: HOTEL spends more, and RETAIL spends less.



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

REGION:OTHER spends more and OPORTO spends less



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

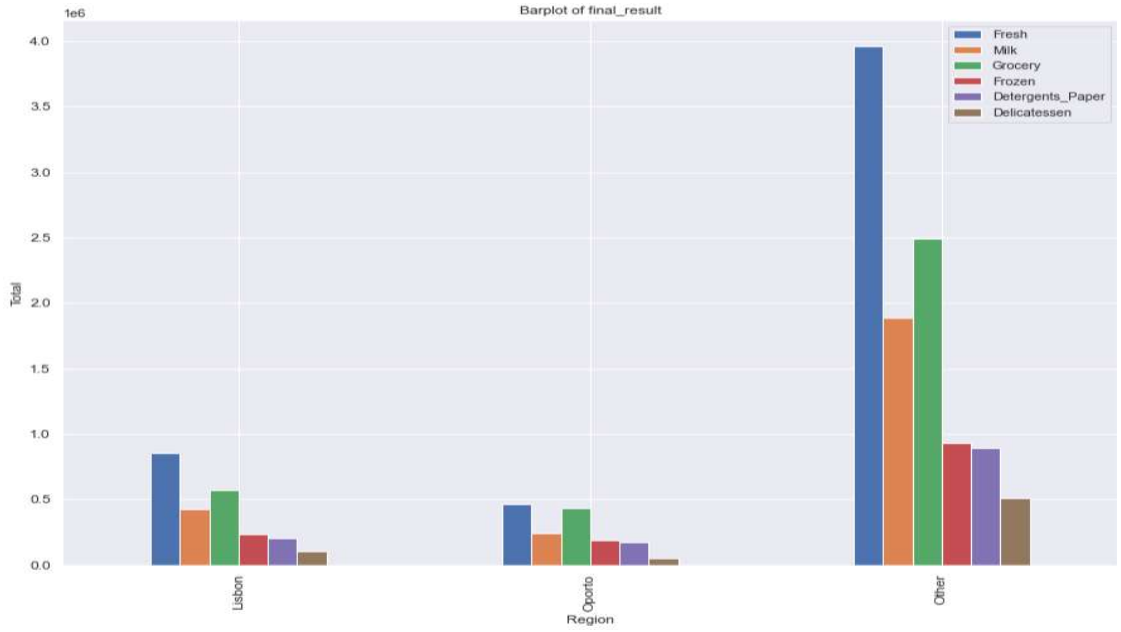
1.2. There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.

Solution:

Bar plot of different varieties across Region is shown.

In OTHER region,we are spending the maximum on all varieties
In OPORTO region,we are spending the minimum on all varieties.

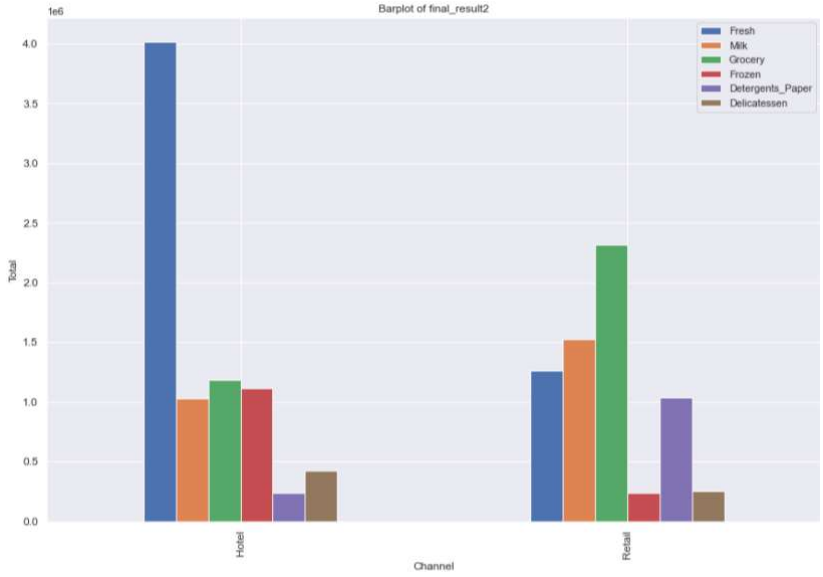
Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
Lisbon	854833	422454	570037	231026	204136	104327	2386813
Oporto	464721	239144	433274	190132	173311	54506	1555088
Other	3960577	1888759	2495251	930492	890410	512110	10677599



Bar plot of different varieties across channel is shown:

IF the spending on the varieties across channel is different from every product, we can see FRESH variety spends large in HOTEL channel while FROZEN variety spends the least in RETAIL channel.

Channel	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
Hotel	4015717	1028614	1180717	1116979	235587	421955	7999569
Retail	1264414	1521743	2317845	234671	1032270	248988	6619931



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?

Solution:

To find the most inconsistent behaviour and the least inconsistent behaviour of all, we can use Inter-Quartile range method or standard deviation.

The Inter Quartile range (length of the box) gives us an idea about how dispersed the data is. Here, dispersion is maximum in "Fresh" and minimum in "Delicatessen".

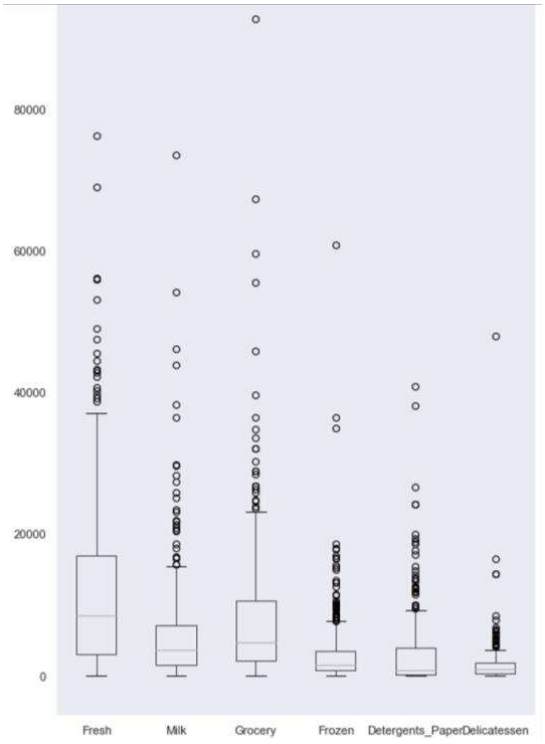
IQR = quantile (0.75) - quantile (0.25)

IQR for all the six varieties:

```
Fresh             13806.00
Milk               5657.25
Grocery            8502.75
Frozen             2812.00
Detergents_Paper   3665.25
Delicatessen       1412.00
dtype: float64
```

The most compact data is in "Delicatessen" and the most spread-out data is in "Fresh".

Boxplot of all the six varieties:



IF the box plot is relatively short, then the data is more compact. If the box plot is relatively tall, then the data is spread out. "Fresh “shows the most inconsistent behaviour and "Delicatessen “shows the least inconsistent behaviour.

Standard deviation of all the six varieties:

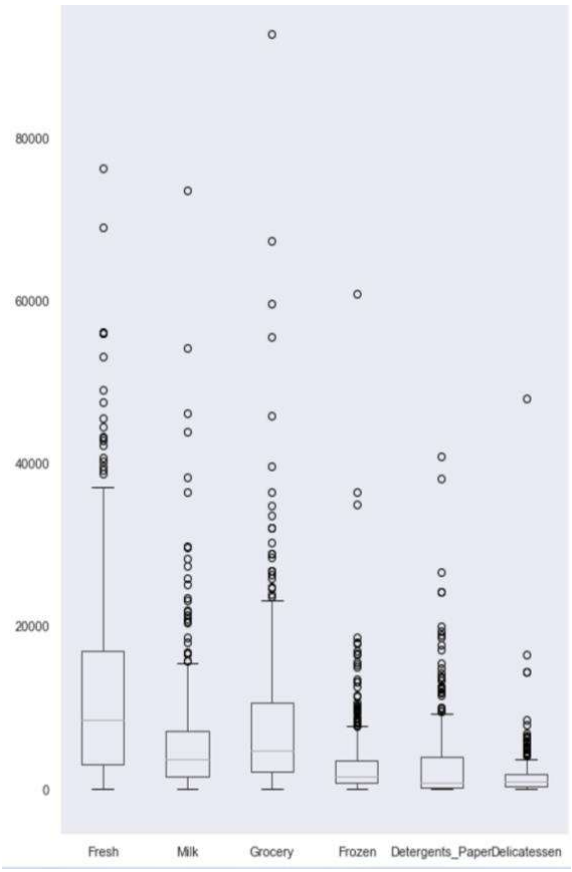
Standard deviation is a measure of how dispersed the data is in relation to the mean. The largest dispersion is in is in "Fresh" and the smallest dispersion is in "Delicatessen".

```
Fresh              12647.328865
Milk                7380.377175
Grocery             9503.162829
Frozen              4854.673333
Detergents_Paper    4767.854448
Delicatessen        2820.105937
dtype: float64
```

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

Solution: Yes, all the 6 varieties have many outliers in the data.

We can use IQR or boxplot to find the outliers in the data.



An outlier is an observation that is numerically distant from the rest of the data. When reviewing a box plot, an outlier is defined as a data point that is located outside the whiskers of the box plot.

By IQR method:

	Channel	Delicatessen	Detergents_Paper	Fresh	Frozen	Grocery	Milk	Region
0	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False
2	False	True	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False
4	False	True	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...
435	False	False	False	False	True	False	False	False
436	False	False	False	True	False	False	False	False
437	False	False	True	False	False	True	False	False
438	False	False	False	False	False	False	False	False
439	False	False	False	False	False	False	False	False

440 rows × 8 columns

Outliers are greater than (Q3+1.5*IQR) or less than (Q1-1.5*IQR).

Here, All the "True" indicate the presence of outliers.

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.

Solution:

Fresh" variety is widely used by the retailers across all regions and channels. But Frozen, Detergents Paper, Delicatessen are not used that much. We can try to improve the quality of these varieties and sell them where the demand is more. Also, there are many buyers/spenders in the "other" region, but they lack in other regions. So, we can try ways to optimise retail efficiency in other regions as well.

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).

Solution: The first step before doing any kind of data-related task, be it machine learning, data analytics, etc. is to see and analyse the data that is contained inside the data set that you have. We need to import NumPy, pandas, matplotlib, seaborn etc.

Importing necessary libraries:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

SHAPE= (62,14)

SIZE= 868

INFO- We've 62 rows and 14 columns and there is no null value.

FLOAT VALUES: GPA and SALARY

INTEGER VALUES: ID, age, social networking, spending, satisfaction, text messages

OBJECT VALUES: Gender, class, major, grad intention, employment and computer

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

NULL VALUES:0

```
ID                 0
Gender             0
Age                0
Class              0
Major              0
Grad Intention     0
GPA                0
Employment         0
Salary             0
Social Networking  0
Satisfaction       0
Spending           0
Computer           0
Text Messages      0
dtype: int64
```

DESCRIPTIVE STATISTICS OF THE DATASET:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
ID	62.0	NaN	NaN	NaN	31.5	18.041619	1.0	16.25	31.5	46.75	62.0
Gender	62	2	Female	33	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Age	62.0	NaN	NaN	NaN	21.129032	1.431311	18.0	20.0	21.0	22.0	26.0
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	62.0	NaN	NaN	NaN	3.129032	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	62.0	NaN	NaN	NaN	48.548387	12.080912	25.0	40.0	50.0	55.0	80.0
Social Networking	62.0	NaN	NaN	NaN	1.516129	0.844305	0.0	1.0	1.0	2.0	4.0
Satisfaction	62.0	NaN	NaN	NaN	3.741935	1.213793	1.0	3.0	4.0	4.0	6.0
Spending	62.0	NaN	NaN	NaN	482.016129	221.953805	100.0	312.5	500.0	600.0	1400.0
Computer	62	3	Laptop	55	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Text Messages	62.0	NaN	NaN	NaN	246.209677	214.46595	0.0	100.0	200.0	300.0	900.0

From descriptive analysis or five-point summary, we can say that

1) We've 2 unique values in GENDER,3 unique values in CLASS, GRAD INTENTION, EMPLOYMENT, COMPUTER and 8 unique values in MAJOR.
2) Number of females is 33 and males is 29.
3) 28 out of 62 people have grad intention while 34 don't.
4) The maximum age of students is 26, minimum age is 18 and median age is 21. The maximum salary is 80, minimum salary is 25 and median is 50. Students have scored 3.9 GPA Maximum while the minimum GPA seems to be 2.3.
5) 55 students have laptop for education.
6) 43/62 people are part time employees. Most of the people are part time employees.
7) Most people majored in Retailing/Marketing.

2.1. For this data, construct the following contingency tables (Keep Gender as row variable) 2.1.1. Gender and Major 2.1.2. Gender and Grad Intention 2.1.3. Gender and Employment 2.1.4. Gender and Computer. Add the contingency tables for each

Solution:

A contingency table is a type of table that summarizes the relationship between two categorical variables.

To create a contingency table in Python, we can use the crosstab () function.

2.1.1. Gender and Major

```
pd.crosstab(df["Gender"], df["Major"])          #A contingency table is a type of table that summarizes the relationship bet
                                                 #To create a contingency table in Python, we can use the pandas.crosstab()
```

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

2.1.2. Gender and Grad Intention

```
pd.crosstab(df["Gender"], df['Grad Intention'])
```

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

2.1.3. Gender and Employment

```
pd.crosstab(df["Gender"], df['Employment'])
```

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

2.1.4. Gender and Computer

```
pd.crosstab(df["Gender"], df['Computer'])
```

Computer	Desktop	Laptop	Tablet
Gender			
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?

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

Solution:

2.2.1. Total number of males=29

Total number of students=62

P (Male/Total students) = 29/62= 0.4677

2.2.2. Total number of females=33

Total number of students=62

P (female/total students) =33/62=0.532

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

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

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

Solution:

2.3.1. The conditional probability of different majors among the male students in CMSU=P (Different majors/male)= This table gives the probability of male choosing different majors :

Major Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Female	0.090909	0.090909	0.212121	0.121212	0.121212	0.090909	0.272727	0.000000
Male	0.137931	0.034483	0.137931	0.068966	0.206897	0.137931	0.172414	0.103448
All	0.112903	0.064516	0.177419	0.096774	0.161290	0.112903	0.225806	0.048387

2.3.2. The conditional probability of different majors among the female students of CMSU=P (Different majors/female) =This table gives the probability of females choosing different majors:

Major Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Female	0.090909	0.090909	0.212121	0.121212	0.121212	0.090909	0.272727	0.000000
Male	0.137931	0.034483	0.137931	0.068966	0.206897	0.137931	0.172414	0.103448
All	0.112903	0.064516	0.177419	0.096774	0.161290	0.112903	0.225806	0.048387

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.

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

Solution:

2.4.1 Probability that a randomly chosen student is a male and intends to graduate**:**

P (Intents to graduate/Male) =

Grad Intention	No	Undecided	Yes
Gender			
Female	0.272727	0.393939	0.333333
Male	0.103448	0.310345	0.586207
All	0.193548	0.354839	0.451613

Probability that a randomly chosen student is a male and intends to graduate is 0.586.

2.4.2 probability that a randomly selected student is a female and does NOT have a laptop:

P (Have a laptop/Female) =29/33=0.8787

Computer	Desktop	Laptop	Tablet
Gender			
Female	0.060606	0.878788	0.060606
Male	0.103448	0.896552	0.000000
All	0.080645	0.887097	0.032258

P(does not have a laptop/Female)=1-29/33=1-0.8787=0.1213

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?

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

Solution:

2.5.1

Probability that a randomly selected student is a male=0.467

Probability that a randomly selected student has full time employment=0.161

Probability that a randomly selected student is both a male and full employment=0.112

probability that a randomly chosen student is a male or has full-time employment=0.467+0.161-0.112=0.5161

2.5.2

Total females=33

Females majoring in international buisness=4

Females majoring in management=4

Total females majoring in international buisness and management=8

conditional probability that given a female student is randomly chosen, she is majoring in international business or management=8/33=0.2424

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?

Solution:

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

Grad Intention	No	Yes	All
Gender			
Female	0.225	0.275	0.5
Male	0.075	0.425	0.5
All	0.300	0.700	1.0

P(A)=The probability that a randomly selected Student is Female = 0.5

P(B)=The probability that a randomly selected student is graduate = 0.7

P(A intersection B)= The probability that a randomly selected student is female and intends to graduate=0.55

[A and B are known as independent events if $P(A \cap B) = P(B) .P(A)$]

here,p(A intersection B) !=P(A) *P(B)

Therefore,They 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?

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.

Solution: 2.7.1

Number of students having GPA less than 3=17

Total students=62

Therefore, if a student is chosen randomly, the probability that his/her GPA is less than 3 is 17/62=0.2741

2.7.2

Number of male students having salary greater than 50=14

Salary	False	True
Gender		
False	0.454545	0.545455
True	0.517241	0.482759

The conditional probability that a randomly selected male earns 50 or more is 0.482

Number of female students having salary greater than 50=18

Salary	False	True
Gender		
False	0.517241	0.482759
True	0.454545	0.545455

The conditional probability that a randomly selected female earns 50 or more is 0.545.

2.8. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarizing your conclusions.

Solution: 1st method: The Shapiro wilk test tests the null hypothesis that the data was drawn from normal distribution.H0: normal distribution, H1: Non-normal distribution

```
shapiro(df["Salary"])                          #pvalue<alpha   Reject H0
                                              #Data is not normally distributed
```

```
ShapiroResult(statistic=0.9565856456756592, pvalue=0.028000956401228905)
```

```
shapiro(df["Spending"])                        #pvalue<alpha #Data is not normally distributed
```

```
ShapiroResult(statistic=0.8777452111244202, pvalue=1.6854661225806922e-05)
```

```
shapiro(df["Text Messages"])                    #pvalue<alpha #Data is not normally distributed
```

```
ShapiroResult(statistic=0.8594191074371338, pvalue=4.324040673964191e-06)
```

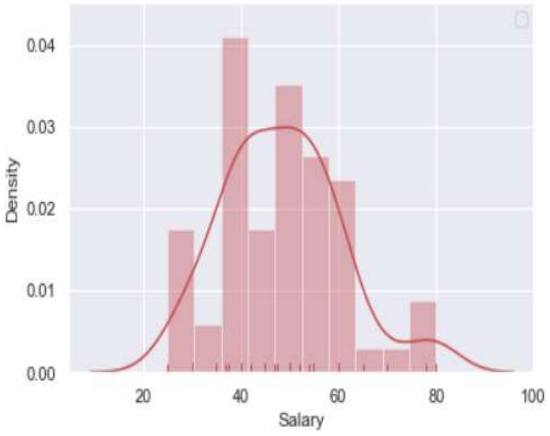
```
shapiro(df["GPA"])                              #pvalue>alpha  #Failed to reject H0
                                                # Data is normally distributed
```

```
ShapiroResult(statistic=0.9685361981391907, pvalue=0.11204058676958084)
```

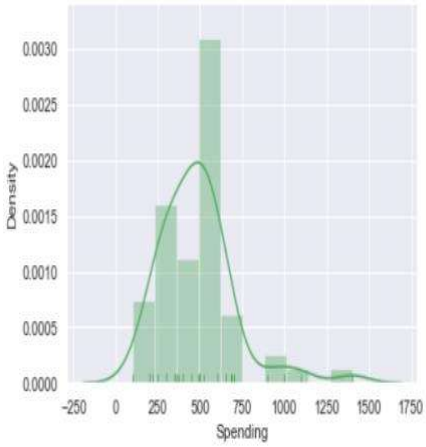
Distplot:

The distplot represents the univariate distribution of data i.e., data of a variable against the density distribution.

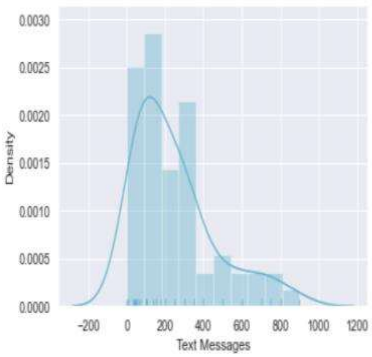
If the skew value is zero,the data is symmetric.If it is negative ,it means data is skewed left.And if it is right ,it means data is skewed right.

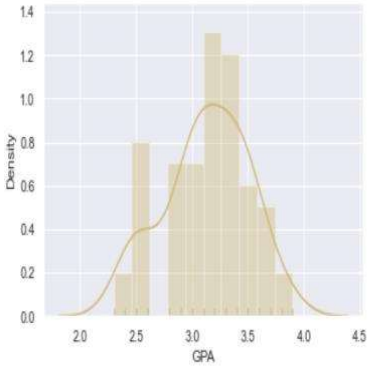


Skewness: 0.534701



Skewness: 1.585915

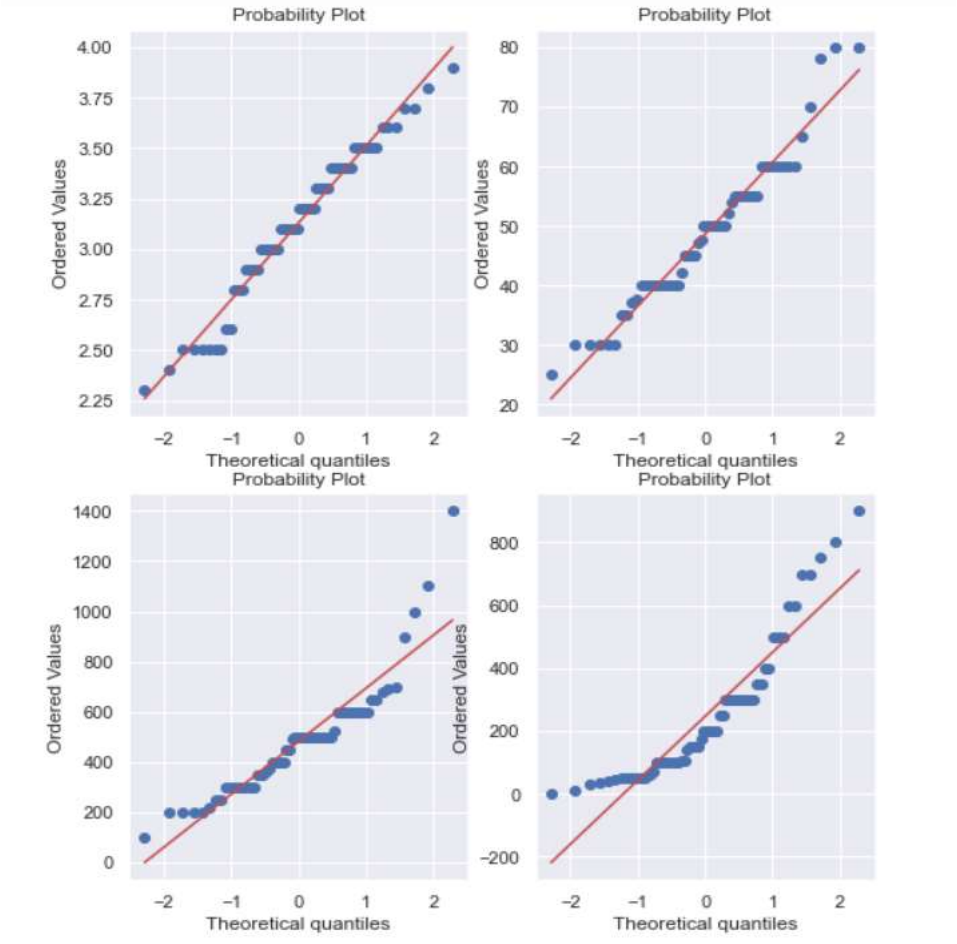




Skewness:1.295

Skewness:-0.314

Probability plot: The probabilty plot can be used to find if your data follows a normal distribution or not. Here all the points follow a straight line. Hence, they follow normal distribution.



Conclusion: The mean salary is around 50. Total numbers of students are 62 out of which females are 33. 43/62 people are part time employees. The maximum age of students is 26, minimum age is 18 and median age is 21. The maximum salary is 80, minimum salary is 25 and median is 50. Students have scored 3.9 GPA Maximum while the minimum GPA seems to be 2.3. Most of the people are part time employees. Most of the people majored in Retailing/Marketing. GPA is skewed left whereas text messages, spending and salary are skewed right.

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.

Solution:

Necessary Modules:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
from scipy.stats import ttest_1samp
from scipy.stats import ttest_ind
from scipy.stats import levene,wilcoxon,shapiro,normaltest
```

Five-point Summary:

	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

Null Values:

```
A    0
B    5
dtype: int64
```

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.

Solution:

Step 1: **Define null and alternate hypothesis.**

Alternative hypothesis (HA) :

mean moisture content> 0.35

And, Null hypothesis (H0) : mean moisture content <= 0.35

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

given:H0: mean moisture content <=0.35 HA: mean moisture content > 0.35

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

given:H0: mean moisture content <=0.35 HA: mean moisture content > 0.35.

Step2: **Write the significance level**

We'll assume alpha=0.05, since it's not given in the question

Step3: **Identify the test statistic:**

Population mean =0.35, n1=36, n2=31

We've been given two independent samples only.

Population standard deviation is not given. We've to check whether the moisture contents in bot h types of shingles are within the permissible limits.

Population standard deviation is not given so we can't perform z test. We'll have to do t test.

There will be 1_sampled t-test for both.

And since we are checking if the mean moisture content > 0.35. So.it will be 1-tailed test(one-sided).

Step4: **Calculate the p value and test statistic:**

```
ttest_1samp(a=df1["A"],popmean=0.35,alternative="greater")
```

```
Ttest_1sampResult(statistic=-1.4735046253382782, pvalue=0.9252236685509249)
```

```
ttest_1samp(a=df1["B"].dropna(),popmean=0.35,alternative="greater")
```

```
Ttest_1sampResult(statistic=-3.1003313069986995, pvalue=0.9979095225996808)
```

Step5: Conclusion:

Here pvalue > alpha. We fail to reject null hypothesis. We conclude that the mean moisture content in both the types of shingles are within permissible limits.

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?

Please reflect on all that you have learnt while working on this project. This step is critical in cementing all your concepts and closing the loop. Please write down your thoughts here.

Solution:

Step 1: Define null and alternate hypothesis: The null hypothesis states that the mean moisture content in both the types of shingles are equal

H0: mu_A=mu_B

The alternate hypothesis states that the mean moisture content in both the types of shingles are different.

H1: mu_1!= mu_B

Step 2: Write the significance level:

We'll assume alpha=0.05, since it's not given in the question.

Step 3: Identify the test statistic:

We have two independent samples, and we don't know the population standard deviation and variance. Sample sizes for both the samples are not same(n>30) Hence, we'll perform independent 2-sampled t-test.

Step 4:Calculate the pvalue and test statistic:

```
In [127]: ttest_ind(a=df1["A"],b=df1["B"],nan_policy="omit",equal_var=True,alternative="two-sided")
Out[127]: Ttest_indResult(statistic=1.2896282719661123, pvalue=0.2017496571835306)
```

Step 5: Conclusion:

Here pvalue > alpha. We fail to reject null hypothesis. It means (mu_A=mu_B) We conclude that the mean moisture content in both the types of shingles are equal.

Assumption do you need to check before the test for equality of means is performed:

1) The data is continuous

2) The two groups should be independent of each other.

3) The data should be normally distributed.

4) The groups should have equal variance.

5) Only two groups are compared.

levene test

It is performed to check if the variances are equal or not. The levene test tests the null hypothesis that all input samples are from population with equal variances.

H0: Variances are equal.

H1: Variances are unequal.

Shapiro Test

The Shapiro wilk test tests the null hypothesis that the data was drawn from normal distribution.

H0: normal distribution

H1: Non-normal distribution