SMDM Project

Name: Gunjar Fuley

Phone: 9938126651

Email: gforgunjaar@gmail.com

Problem 1

Wholesale Customers Analysis (Download Data)

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).

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?

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440	440	440	440	440	440	440	440	440
unique	NaN	2	3	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	Hotel	Other	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	298	316	NaN	NaN	NaN	NaN	NaN	NaN
mean	220.5	NaN	NaN	12000.29773	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	127.161315	NaN	NaN	12647.32887	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	1	NaN	NaN	3	55	3	25	3	3
0.25	110.75	NaN	NaN	3127.75	1533	2153	742.25	256.75	408.25
0.5	220.5	NaN	NaN	8504	3627	4755.5	1526	816.5	965.5
0.75	330.25	NaN	NaN	16933.75	7190.25	10655.75	3554.25	3922	1820.25
max	440	NaN	NaN	112151	73498	92780	60869	40827	47943

The data doesn't have any null value. The data has 440 values under each item. The above table has 5 number summary of the data.

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

'Other' region has spent most and the 'Oporto' region has spent the most.

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

Channel 'Hotel' has spent more than channel 'Retail'.

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

Yes, most of the Regions & Channels shows similar behaviour.

Except for Channel- 'Hotel', where it shows different behaviour.

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?

Coefficient of Variation for different items are-

#Fresh - 1.0539179237648593

#Milk - 1.2732985841005522

#Grocery - 1.1951743729613995

#Frozen - 1.5803323838615222

#Detergents_Paper - 1.6546471384293562

#Delicatessen - 1.849406897322304

Most Consistent behaviour- 'Fresh'

Least Consistent behaviour- 'Delicatessen'

1.4 Are there any outliers in the data?

Yes there are outliers in the data. Outlines are clearly visible in the boxplot in the PYTHON file.

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

'Delicatessen' shows too much inconsistency therefore it needs to be taken care with high priority. However, other items have less inconsistent than 'Delicatessen' but they have huge scope of improvement.
The above suggestions are based on presence of too many outliers as well as the Coefficient of Variation.

Problem 2 -

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

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

2.1.1. Gender and Major

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

2.1.2. Gender and Grad Intention

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

2.1.3. Gender and Employment

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

2.1.4. Gender and Computer

Computer	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?

total = 62
male = 29
female = 33
prob_male = male/total
print(prob_male)

0.46774193548387094

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

prob_female = 1 - prob_male

print(prob_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.

Management

0.20689655172413793

Retailing/Marketing

0.1724137931034483

Accounting

0.13793103448275862

Other

0.13793103448275862

Economics/Finance

0.13793103448275862

Undecided

0.10344827586206896

International Business

0.06896551724137931

CIS

0.034482758620689655

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

Retailing/Marketing

0.2727272727272727

Economics/Finance

0.12121212121212122

Management

0.12121212121212122

International Business

0.12121212121212122

Accounting

0.09090909090909091

Other

0.09090909090909091

CIS

0.09090909090909091

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.

Graduation intention among males-

Yes 17

Undecided 9

No 3

Probability- 17/ 62

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

0.27419354838709675

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

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 either a male or has full-time employment?

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

= (29/62)+(10/62)-(7/62)
The probability that a randomly chosen student is either a male or has full-time employm ent
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.

female manegement = 4
female internationalbusiness = 4
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

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

2.6.1. If a student is chosen randomly, what is the probability that his/her GPA is less than 3?

Number of students less than 3 GPA=
17
Randomly chosen student having probability of GPA less than 3=
0.27419354838709675

2.6.2. 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.

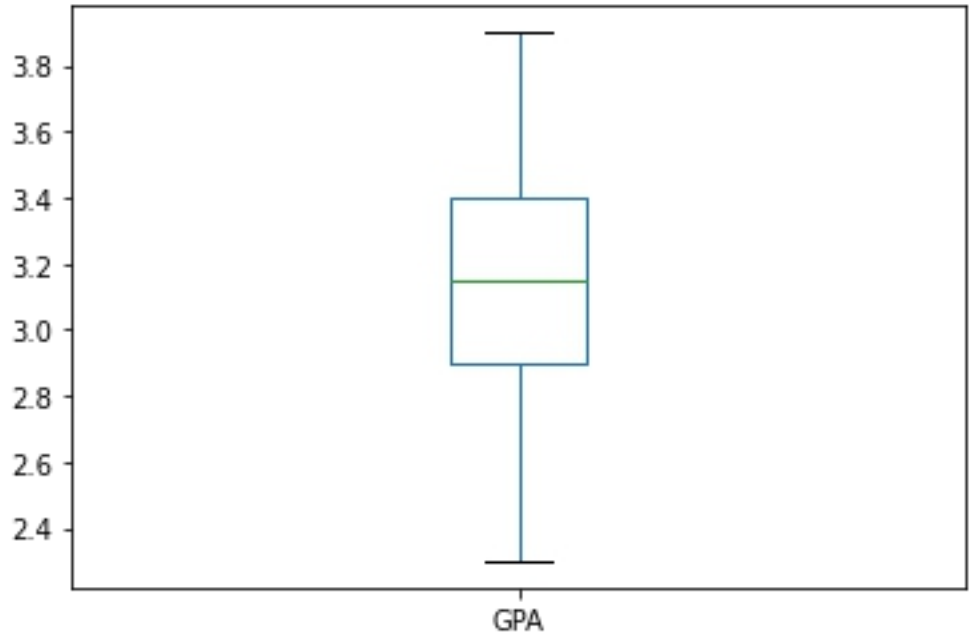
Males earning 50 and more= 14
The conditional probability that a randomly selected male earns 50 or more
0.4827586206896552

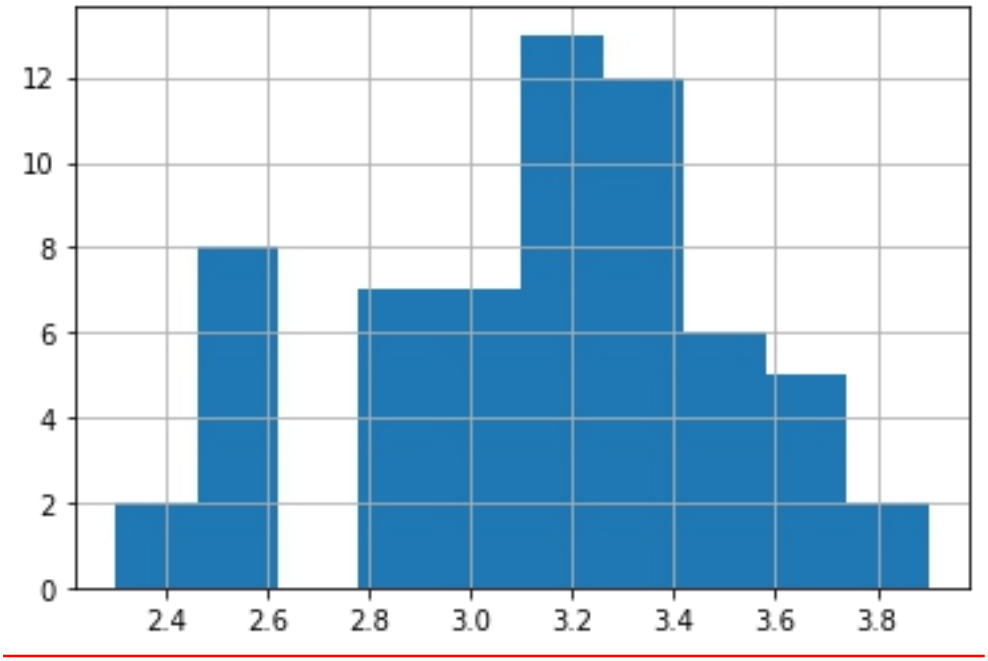
Females earning 50 and more= 18
The conditional probability that a randomly selected female earns 50 or more
0.5454545454545454

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.

GPA

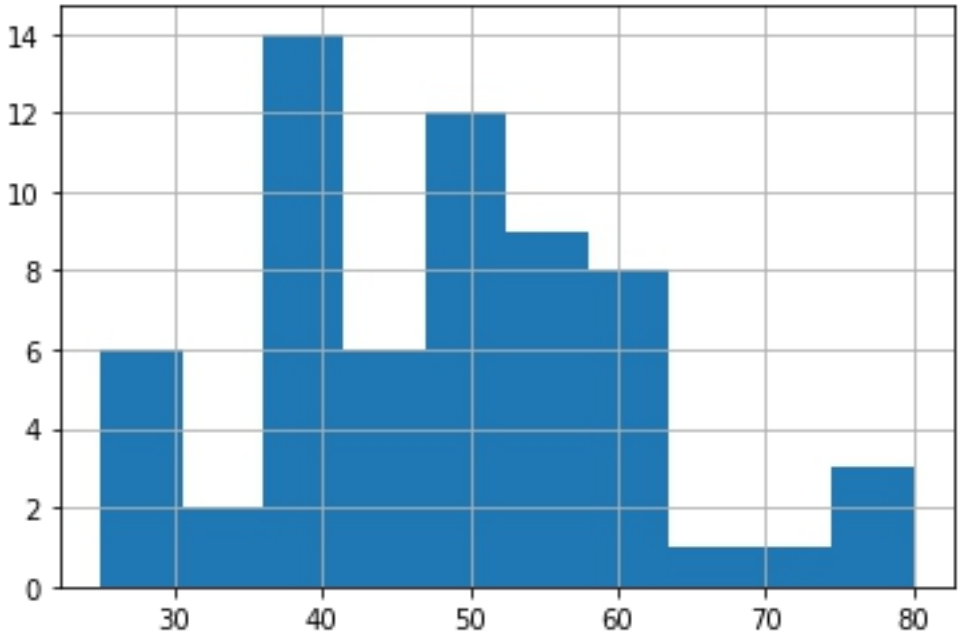
It follows normal distribution





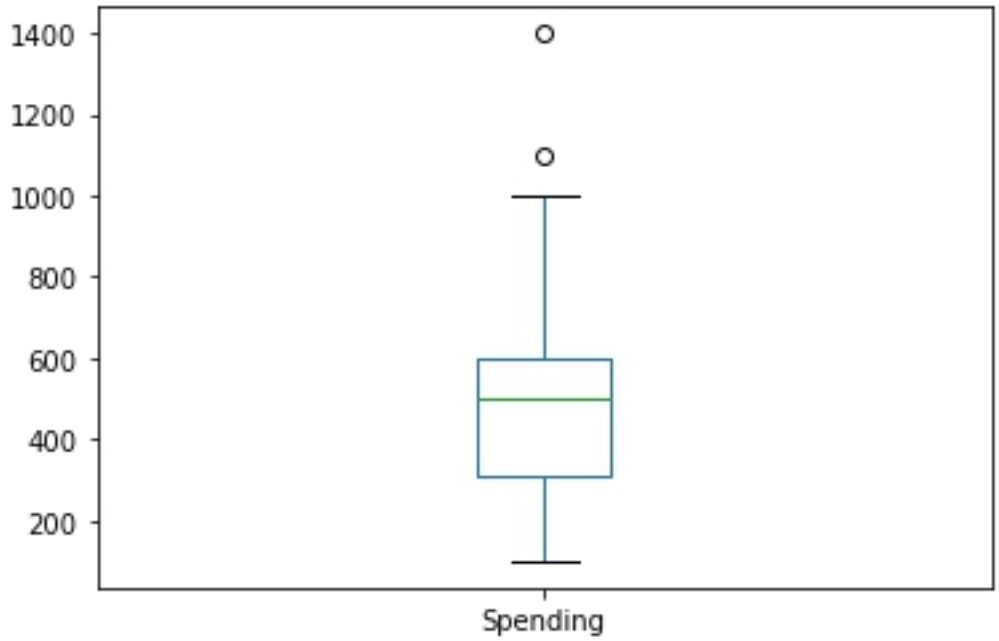
Salary

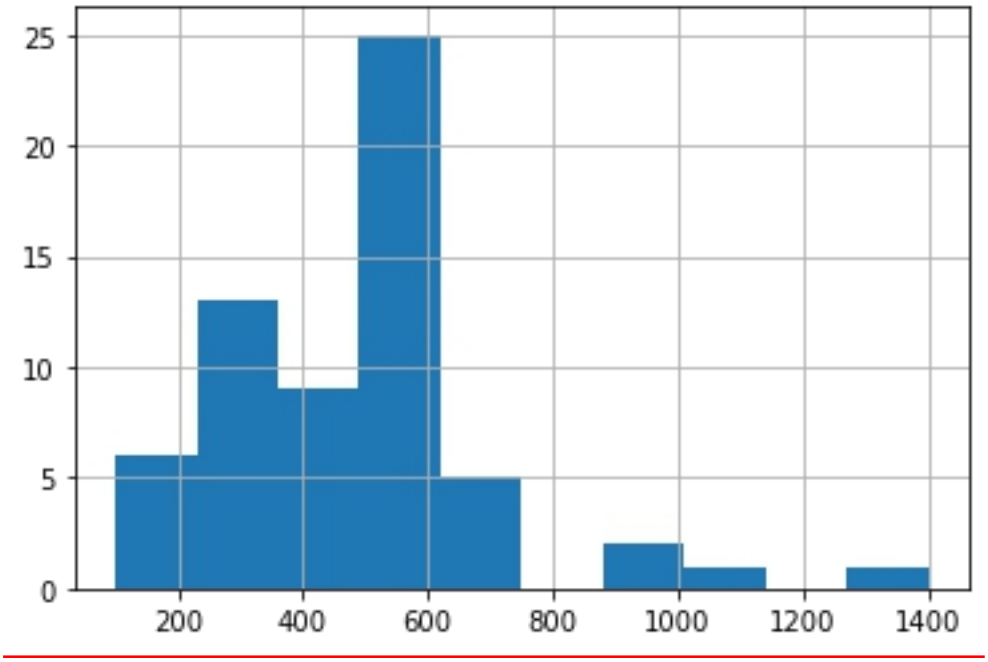
It doesn't follows normal distribution



Spending

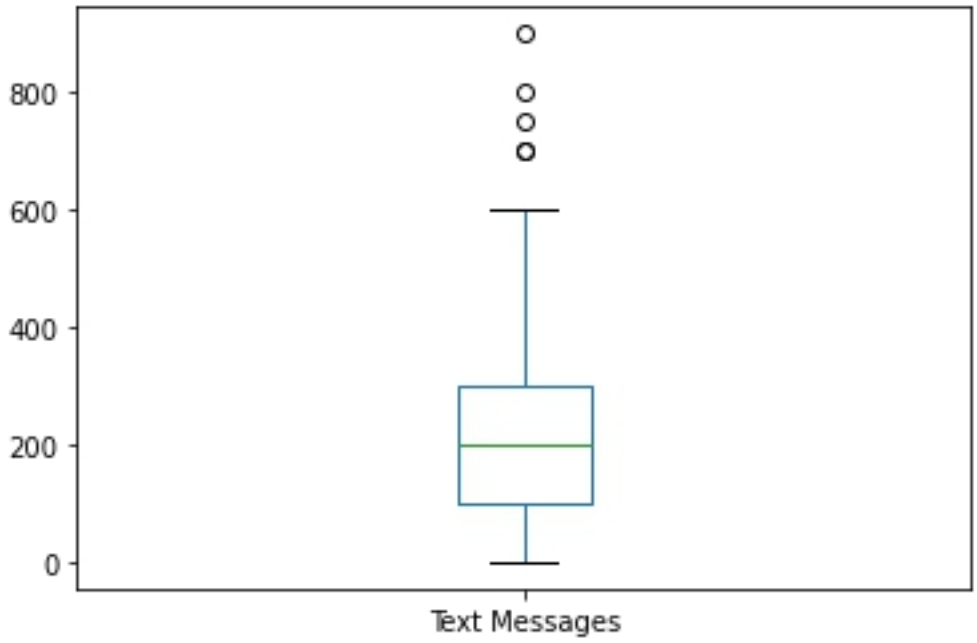
It doesn't follows normal distribution

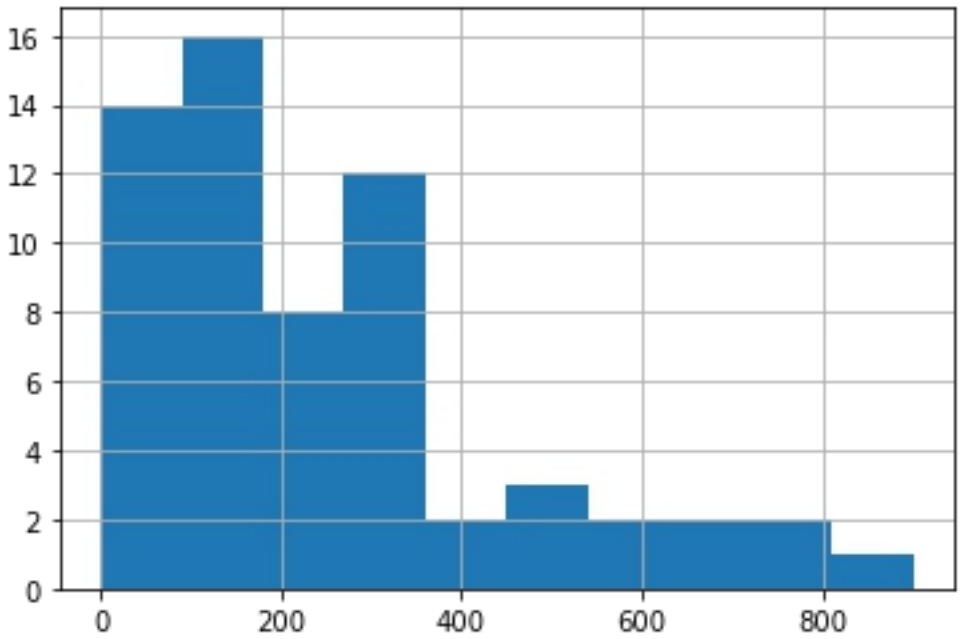




Text Messages

It doesn't follows normal distribution





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.

Mean of Shingles A = 0.3166666666666666

One Sample T Test T statistic: -1.4735046253382782 p value: 0.07477633144907513

pvalue is greater than 0.05, therefore we do not reject H0 .

No enough evidence to to conclude that the mean moisture content for Sample A shingles is less than 0.35 pounds per 100 square feet.

p value: 0.07477633144907513

The mean moisture content is not less than 0.35 pounds per 100 square feet, therefore the probability of observing a sample of 36 shingles that will result in a sample mean moisture content of 0.3166666666666666 pounds per 100 square feet or less is 0.07477633144907513.

Mean of Shingles B = 0.2735483870967742

One Sample T Test T Statistic: -3.1003313069986995 p value: 0.0020904774003191826

Now, in this case pvalue is less than 0.05, therefore we'll reject the Null hypothesis (H0) . Enough evidence is there to conclude that the mean moisture content for Sample B shingles is not less than 0.35 pounds per 100 square feet.

p-value = 0.0020904774003191826

The mean moisture content is not less than 0.35 pounds per 100 square feet, therefore the probability of observing a sample of 31 shingles that will result in a sample mean moisture content of 0.0020904774003191826 pounds per 100 square feet or less is 0.0020904774003191826.

3.2 Problem 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?

The mean for shingles A & B are different.
Mean of Shingles A = 0.3166666666666666
Mean of Shingles B = 0.2735483870967742

H0 : μ(A)= μ(B)
Ha : μ(A)!= μ(B)
α = 0.05

t_statistic=1.29 and pvalue=0.202

Here, we will not reject H0 because the pvalue > α.

The common assumptions made when doing a t-test include those regarding the scale of measurement, random sampling, normality of data distribution, adequacy of sample size and equality of variance in standard deviation.