Business Report

Project –SMDM

Created by Amit Jain

Table of Contents

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).

Introduction: This report explains the business requirements and provide the detailed solution based on the data provided for each problem statement. given in the assignment.

Step of understanding the data:

- Import the data: Imported the data using Python notebooks and analyzed the Channel and Region by spending amount for each Product category.

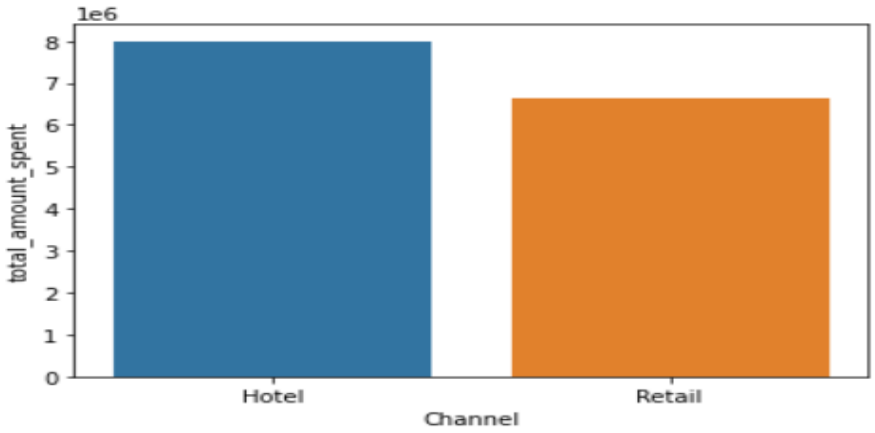
This is how the data look like:

	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

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: Based on analysis of the data, we Grouped the data for Channel wise and then Region wise and checked the total amount spent for all the products and generated the Bar Chart/Following are the Bar Chart Channel wise and region wise along with the data:

	Channel	total_amount_spent
0	Hotel	7999569
1	Retail	6619931



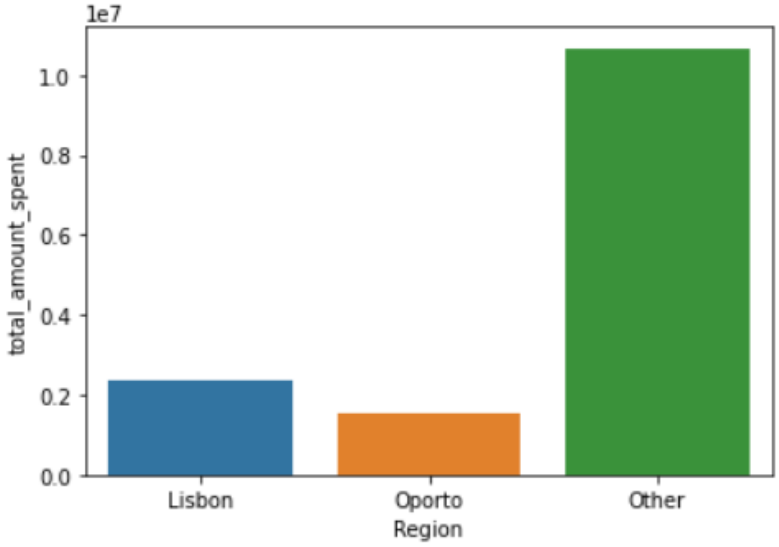
We can clearly see that Channel wise Hotel segment has Spent more than Retails Channel.

Hotel channel spend amount is $7999569 with the highest spend amount and,

Retail spend amount $6619931 has least spend amount based on Channel.

Similarly, Following are the bar plots for Region wise data set:

	Region	total_amount_spent
0	Lisbon	2386813
1	Oporto	1555088
2	Other	10677599



We can clearly see that Region wise Regions in Other segment has Spent most and Oporto spent least amount of money in total product category

Other regions spend amount is $10677599 with the highest spend amount and Oporto region spend amount is $ 1555088 and has least spend amount by Region.

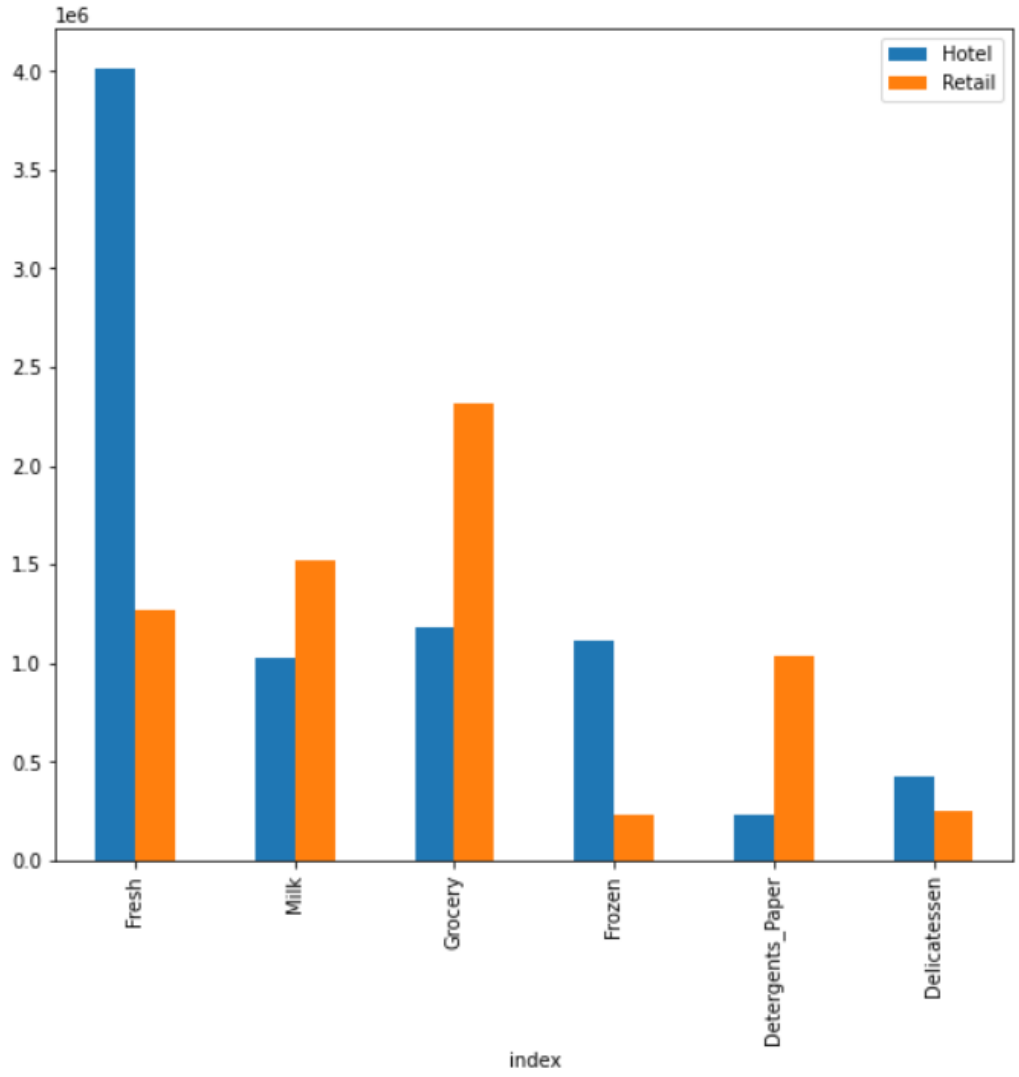
1.2 There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.

Warehouse data is the combination of 2 Channels (Hotel and Retail), gathered from 3 segment of Regions (Lisbon, Oporto, and Other regions) also it shows Amount spent by each customer in following 6 categories of products:

Fresh Milk Grocery Frozen Detergents Paper Delicatessen total_amount_spent

Based on the analysis of the all categories of the product, we Grouped data for Channel wise and taken the summation of total amount spent on each product line. Following numbers and plotted in a bar chart when we arranged data Channel wise:

	index	Hotel	Retail
0	Fresh	4015717	1264414
1	Milk	1028614	1521743
2	Grocery	1180717	2317845
3	Frozen	1116979	234671
4	Detergents_Paper	235587	1032270
5	Delicatessen	421955	248988

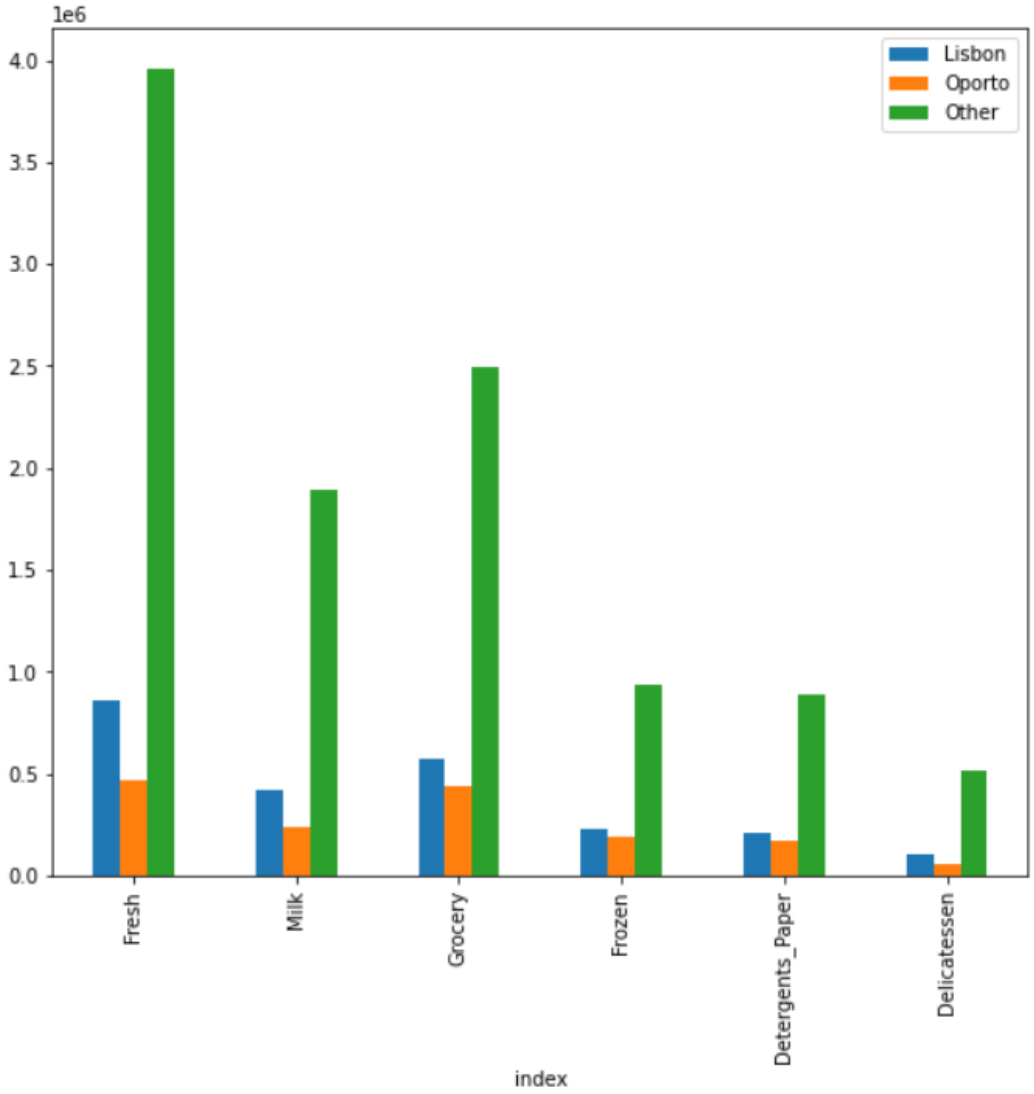


From Above graph we can see that "Hotel" has spent most of the amount in "Fresh" product and least in "Detergents Paper"

And "Retail" has spent most of the amount in "Grocery" product and least in "Frozen"

Based on the analysis of the all categories of the product, we Grouped data for Region wise and taken the summation of total amount spent on each product line. Following numbers and plotted in a bar chart when we arranged data Region wise:

	index	Lisbon	Oporto	Other
0	Fresh	854833	464721	3960577
1	Milk	422454	239144	1888759
2	Grocery	570037	433274	2495251
3	Frozen	231026	190132	930492
4	Detergents_Paper	204136	173311	890410
5	Delicatessen	104327	54506	512110



From Above graph we can see that "Lisbon" has spent most of the amount in "Fresh" product and least in "Delicatessen"

And "Oporto" has spent most of the amount in "Fresh" product and least in "Delicatessen"

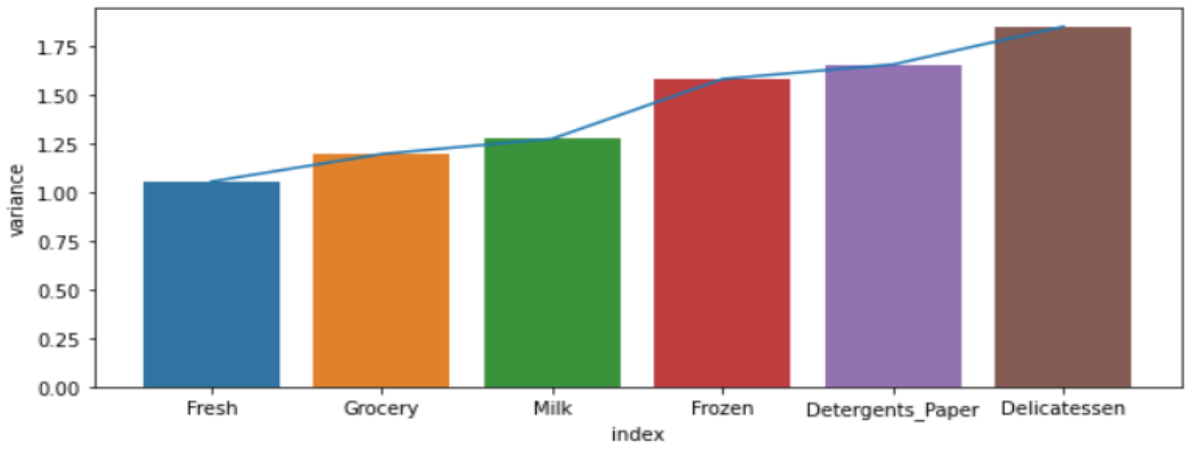
On the other hand, regions in "Other" Categories has spent most of the total amount in "Fresh" product and least in "Delicatessen"

1.3 Based on a descriptive measure of variability, which item shows the most inconsistent behavior? Which items show the least inconsistent behavior?

Solution: Using Coefficient of Variation we find out the least value is of Category “Fresh” (1.05) and highest value is of Category “Delicatessen” (1.85) So from the given data it is clear that most inconsistent behavior shown by item – Delicatessen And least inconsistent behavior shown by item – Fresh

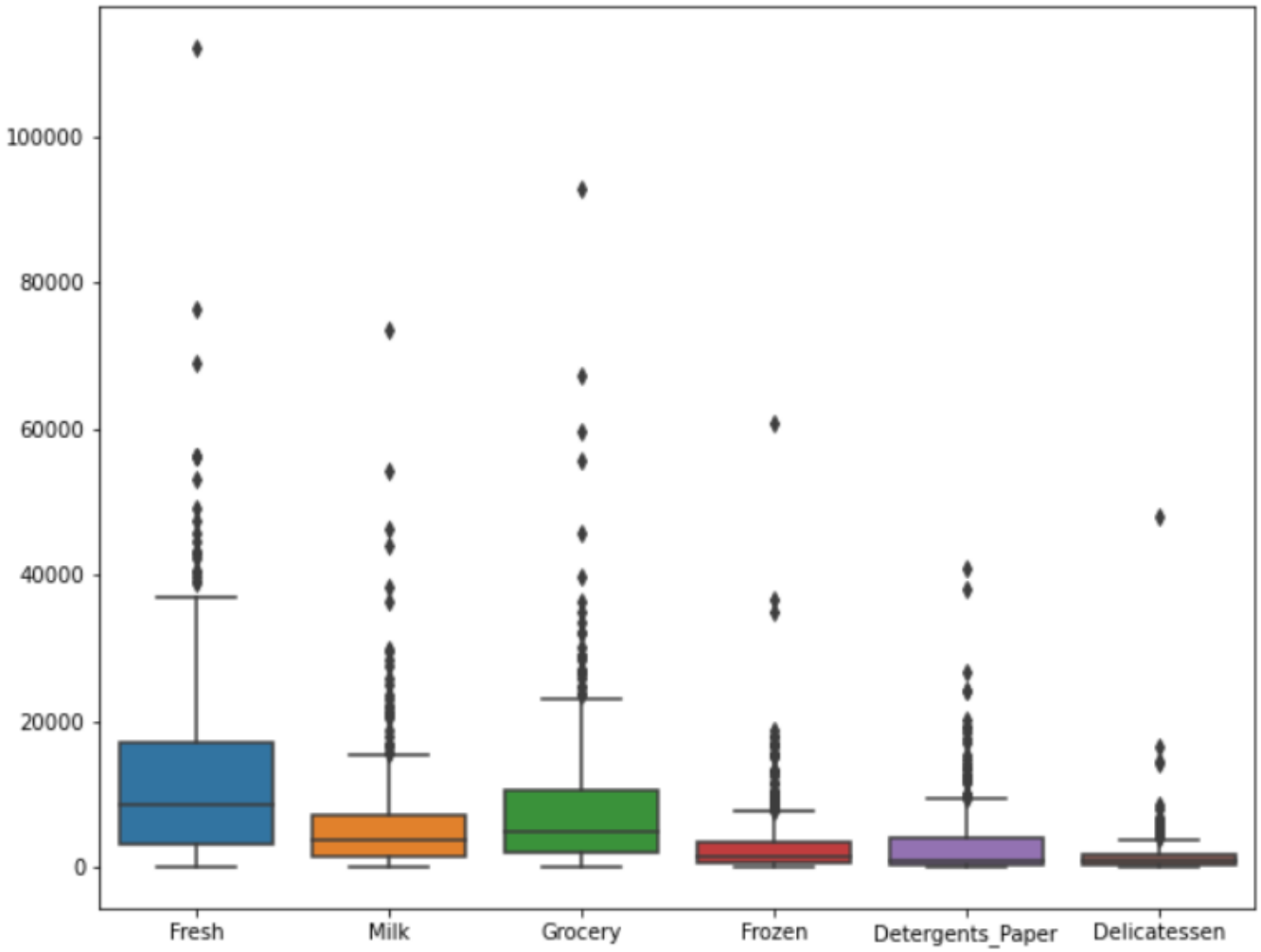
Below is the output from Python –

	index	std	mean	variance
0	Fresh	12647.33	12000.30	1.05
2	Grocery	9503.16	7951.28	1.20
1	Milk	7380.38	5796.27	1.27
3	Frozen	4854.67	3071.93	1.58
4	Detergents_Paper	4767.85	2881.49	1.65
5	Delicatessen	2820.11	1524.87	1.85



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

We plotted boxplot and the output gives the details that in all the data there are outliers:



1.5 Based on 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

Answer Since we do not have enough data available in data source, example: profit or loss information, we cannot judge if spending more money on which product is good for the business or if it is not a profitable product then minimize the Spending on that product.

But still If we ASSUME that profits are same in ratio for all the products , and based on variance in above chart we can suggest that Spending behavior has lot of variations in Product "Delicatessen" so amount spend should be equally distributed in all of the products, to foreseen the future market.

2. CMSU Student Analysis

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).

Step of understanding the data:

- Import the data: Imported the data using Python notebooks and analyzed the Student records for CMSU university.

This is how the data look like:

	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
0	1	Female	20	Junior	Other	Yes	2.90	Full-Time	50.00	1	3	350	Laptop	200
1	2	Male	23	Senior	Management	Yes	3.60	Part-Time	25.00	1	4	360	Laptop	50
2	3	Male	21	Junior	Other	Yes	2.50	Part-Time	45.00	2	4	600	Laptop	200
3	4	Male	21	Junior	CIS	Yes	2.50	Full-Time	40.00	4	6	600	Laptop	250
4	5	Male	23	Senior	Other	Undecided	2.80	Unemployed	40.00	2	4	500	Laptop	100

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

2.1.1 Gender and Major Solution:

Below is the contingency table for Gender and Majors

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

Below is the contingency table for Gender and Graduation Intention:

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

2.1.3 Gender and Employment

Below is the contingency table for Gender and Employment:

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

2.1.4 Gender and Computer

Computer	Gender	Desktop	Laptop	Tablet
0	Female	2	29	2
1	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?

For this we need to find out total male students out of whole student from the given data. The probability that a random selected CMSU student will be a male is: 47%

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

For this we need to find out total Female students out of whole student from the given data. The probability that a random selected CMSU student will be a female is: 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.

The probability that a random selected CMSU Male student taking Accounting as a major is 13.79%

Solution: Using contingency tables of Gender and Majors we got the total numbers of males and number of males opting for different majors Below is the output from Python –

The probability that a random selected CMSU Male student taking Accounting as a major is 13.79%

The probability that a random selected CMSU Male student taking CIS as a major is 3.4%
The probability that a random selected CMSU Male student taking International Business as a major 6.89%

The probability that a random selected CMSU Male student taking Management as a major is 20.68%

The probability that a random selected CMSU Male student taking Other as a major is 13.79%

The probability that a random selected CMSU Male student taking Retailing/Marketing as a major is 17.24%

The probability that a random selected CMSU Male student and not yet Decided for any major i s 10.3%

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

Solution: Using contingency tables of Gender and Majors we got the total numbers of females and number of females opting for different majors Below is the output from Python –

The probability that a random selected CMSU Female student taking Accounting as a major is 9.1%

The probability that a random selected CMSU Female student taking CIS as a major is 3.03%

The probability that a random selected CMSU Female student taking Economics/Finance as a major is 21.21%

The probability that a random selected CMSU Female student taking International Business as a major is 12.12%

The probability that a random selected CMSU Female student taking Management as a major is 12.12%

The probability that a random selected CMSU Female student taking Other as a major is 9.1%

The probability that a random selected CMSU Female student and not yet Decided for any major is 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.

Solution: Using contingency tables of Gender and Grad Intention we got the total numbers of males and number of males intends to be graduate And post calculation we find out that - The probability That a randomly chosen student is a male and intends to graduate is 58.62%

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

Solution: Using contingency tables of Gender and Computer we got the total numbers of females and number of females does not have a laptop And post calculation we find out that The probability That a randomly chosen student is a Female and Don't have Laptop is 12.12%

2.5 Assume that the sample is a 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 a full-time employment

Solution: Using contingency tables of Gender and Employment we got the total numbers of mal es and number of males who are full time employed And post calculation we find out that - The probability of randomly chosen student is a male or has a full-time employment is 51.61%

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

Solution: Using contingency tables of Gender and Major we got the total numbers of females and number of females majoring in international business or management. And post calculation we find out that - The probability of randomly chosen student is a Female and she is majoring in international business or management is 24.24%

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 graduate intention and being female are independent events?

Solution: We call Events as independent when their probabilities are not equal, hence from abo ve result we can say that Being Female and graduate intention is not dependent

Probability of graduate intention is 70%
Probability of graduate intention being female is 55%

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?

Solution: We first checked how many students have GPA less than 3 and then taken probability for those Students getting less than 3 GPA based on the data:

Probability that if a student is chosen randomly his/her GPA is less than 3 is 27.42%

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

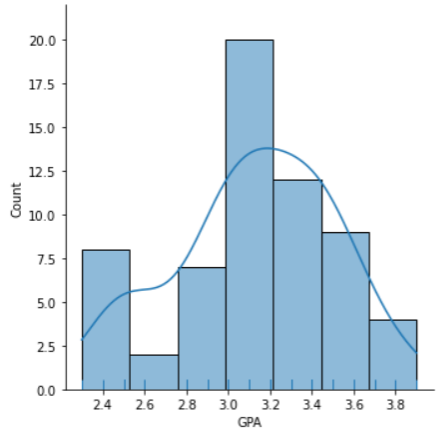
Solution: We first checked how many Female students are present and their Count for how many Female earns more than 50 Salary:

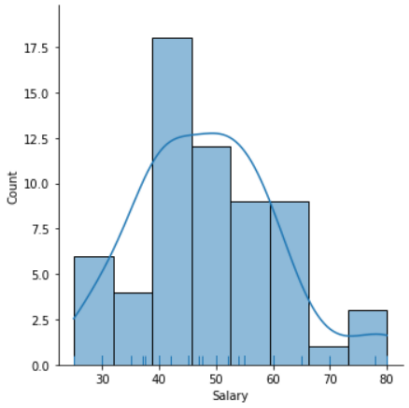
Probability that if randomly selected male earns 50 or more is 34.48%

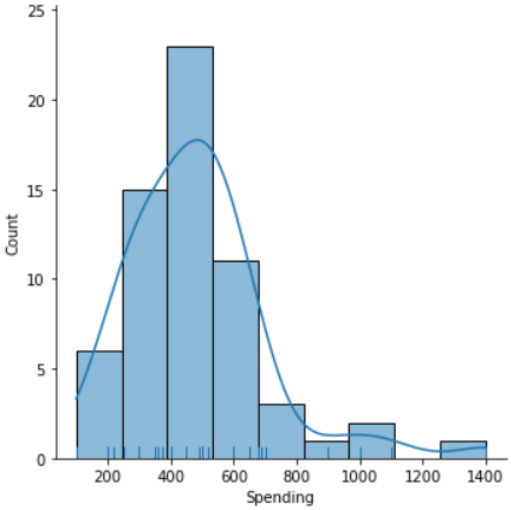
2.8 Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending and Text Messages. For each of them comment whether they follow a normal distribution.

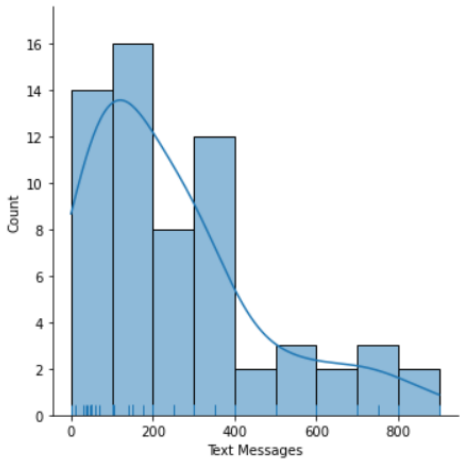
Based on distribution of the data for GPA, Salary, Spending and Text Messages, we plotted displot as well as QQ Plot, we can check that there is no relation between data for Spending and Text Messages , whereas GPA and Salary is following a pattern and data is distributed on a bell curve, which denotes that data for GPA and Spending is Normal distributed, whereas Text messages and salary data is not distributed Normally.

Graphs for the data are shows below:

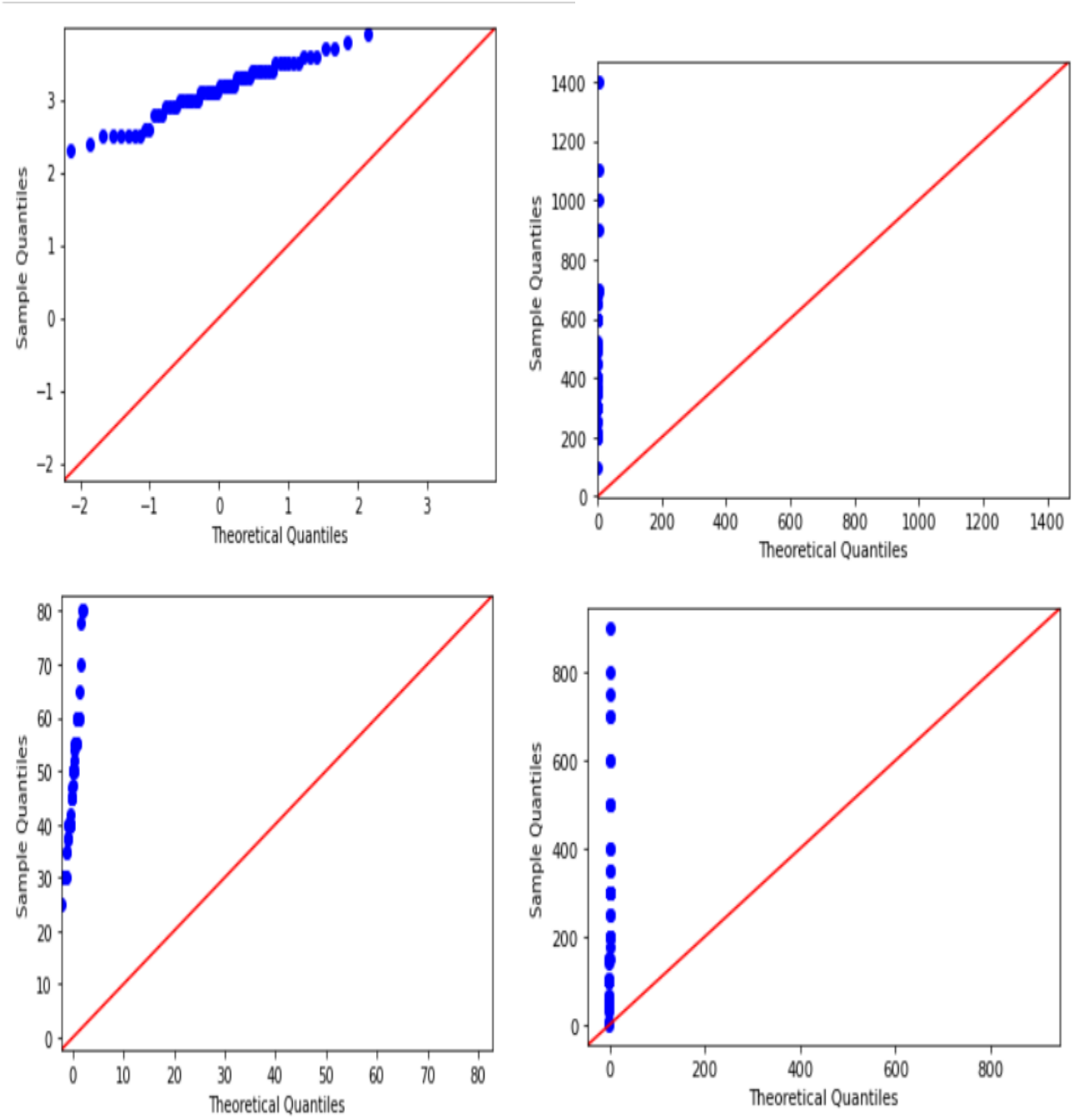








QQ Plots for the data are as follows:



Based on the Dist Plot and QQ Plot, we can see that GPA and salary are normally distributed but spending and Text messages are not normally distributed

3. Manufacturers of ABC Asphalt

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.

First, we Loaded the data using Python: This is how the data look like:

	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
5	0.24	0.18
6	0.16	0.42
7	0.20	0.58
8	0.20	0.25
9	0.20	0.41
10	0.26	0.17
11	0.14	0.13
12	0.33	0.23
13	0.13	0.11
14	0.72	0.10
15	0.51	0.19
16	0.28	0.22
17	0.39	0.44
18	0.39	0.11
19	0.25	0.11
20	0.16	0.31
21	0.20	0.43
22	0.22	0.26

20	0.16	0.31
21	0.20	0.43
22	0.22	0.26
23	0.42	0.18
24	0.24	0.44
25	0.21	0.43
26	0.49	0.16
27	0.34	0.52
28	0.36	0.36
29	0.29	0.22
30	0.27	0.39
31	0.40	NaN
32	0.29	NaN
33	0.43	NaN
34	0.34	NaN
35	0.37	NaN

We can see that total number of samples for Shingle A and B are different , also there is no relation between both data sets, and both can be treated indivisually.

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.

Declare Null and Alternate Hypothesis

Ho= moisture contents >= 0.35

Ha= moisture contents < 0.35

Check for Shingle A
We analyzed the sample records based on Shingle A samples and calculated the P_VALUE for Shingle A samples.

One sample t test
t statistic: -1.4735046253382782 p value: 0.07477633144907513
Assuming that Alpha value is 0.05

When actual_p_value < alpha_value:

Based on the rule:

When actual_p_value < Level of significance then We get evidence to reject the null hypothesis.

When actual_p_value > Level of significance then we consider that there is no evidence to reject the null Hypothesis.

And based on above data set, we can say that

We have no evidence to reject the null hypothesis since actual_p_value > Level of significance
Our one-sample t-test p-value= 0.07477633144907513
Null Hypothesis is True, it means Moisture content is Greater than 0.35

Check for Shingle B

We need to remove NULL values for Shingle B Sample data set, and then we can calculate the pv alue for Shingle B Samples. Upon calculations, we found that
We have evidence to reject the null hypothesis since p value < Level of significance
Our one-sample t-test p-value= 0.0020904774003191813
Alternate Hypothesis is True, it means Moisture content is Less than 0.35

3.2 Do you think that the population mean for shingles A and B are equal? Form the hypothesis and conduct the test of the hypothesis. What assumption do you need to check before the test for equality of means is performed?

Solution: Generate the NULL and Alternate hypothesis.

Ho= meanA = meanB

Ha= meanA != meanB

Since this time, we need to calculate comparative analysis of independent data then we will use 2 sample T test

Upon calculations, we found that tstat 1.2896282719661123
And P Value 0.2017496571835306

Based on the rule as mentioned above, we can say that We have no evidence to reject the null hypothesis since p value > Level of significance
Our one-sample t-test p-value= 0.0020904774003191813
,which means with 95% confidence we can say that mean for both shingles A and B will be same