PGP - Data Science and Business Analytics

SMDM PROJECT REPORT

Akshaya Nallathambi

9th May, 2021

Table Of Contents

List of Figures

List of Tables

List of Graphs

Problem 1

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 channels (Hotel, Retail).

Data Description-

Buyer/Spender: Customer ID(continuous from 1 to 440)

Channel: Retail, Hotel

Region: Other, Lisbon, Oporto

Fresh: Amount spent in this product

Milk: Amount spent in this product

Grocery: Amount spent in this product

Frozen: Amount spent in this product

Detergents_Paper: Amount spent in this product

Delicatessen': Amount spent in this product

Sample of the dataset-

FIGURE 1

Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
1	Retail	Other	12669	9656	7561	214	2674	1338
2	Retail	Other	7057	9810	9568	1762	3293	1776
3	Retail	Other	6353	8808	7684	2405	3516	7844
4	Hotel	Other	13265	1196	4221	6404	507	1788
5	Retail	Other	22615	5410	7198	3915	1777	5185

There are 9 variables out of which 2 are categorical and 7 are continuous. The Buyer/Spender column is more like an ID to the individual customer and the respective row gives the amount they spent on each product. There are no null values.

Types of variables in the data frame-

TABLE 1

Buyer/Spender	int64	Continuous
Channel	object	Categorical
Region	object	Categorical
Fresh	int64	Continuous
Milk	int64	Continuous
Grocery	int64	Continuous
Frozen	int64	Continuous
Detergents_Paper	int64	Continuous
Delicatessen	int64	Continuous

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?

Descriptive statistics are brief descriptive coefficients that summarize a given data set, which can be either a representation of the entire or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables, kurtosis, and skewness. [1]

The below table gives the summary of all the variables of the given data -

FIGURE 2

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Buyer/Spender	440	NaN	NaN	NaN	220.5	127.161	1	110.75	220.5	330.25	440
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	NaN	NaN	NaN	12000.3	12647.3	3	3127.75	8504	16933.8	112151
Milk	440	NaN	NaN	NaN	5796.27	7380.38	55	1533	3627	7190.25	73498
Grocery	440	NaN	NaN	NaN	7951.28	9503.16	3	2153	4755.5	10655.8	92780
Frozen	440	NaN	NaN	NaN	3071.93	4854.67	25	742.25	1526	3554.25	60869
Detergents_Paper	440	NaN	NaN	NaN	2881.49	4767.85	3	256.75	816.5	3922	40827
Delicatessen	440	NaN	NaN	NaN	1524.87	2820.11	3	408.25	965.5	1820.25	47943

There are **2** unique *Channels* and **3** unique *Regions* in the given data set. The '***Hotel***' is the most frequently occuring *Channel* and the '***Other***' is the most frequently occuring *Region* in the data set. We can arrive at a conclusion that all the rows except the *Buyer/Spender* are highly skewed when the mean, standard deviation and the maximum value is taken into account.

The NaN values are displayed for those variables where the respective measure cannot be calculated. That is mean cannot be calculated for categorical/object type variable and frequency cannot be computed for the continuous variable.

The Region "***Oporto***" and the Channel "***Hotel***" spent the most annually on the products. The Region "***Lisbon***" and the Channel "***Retail***" spent the least annually on the products.

The below tables show the amount spent by each Region (*Table 2*) and Channel (*Table 3*) respectively-

TABLE 2

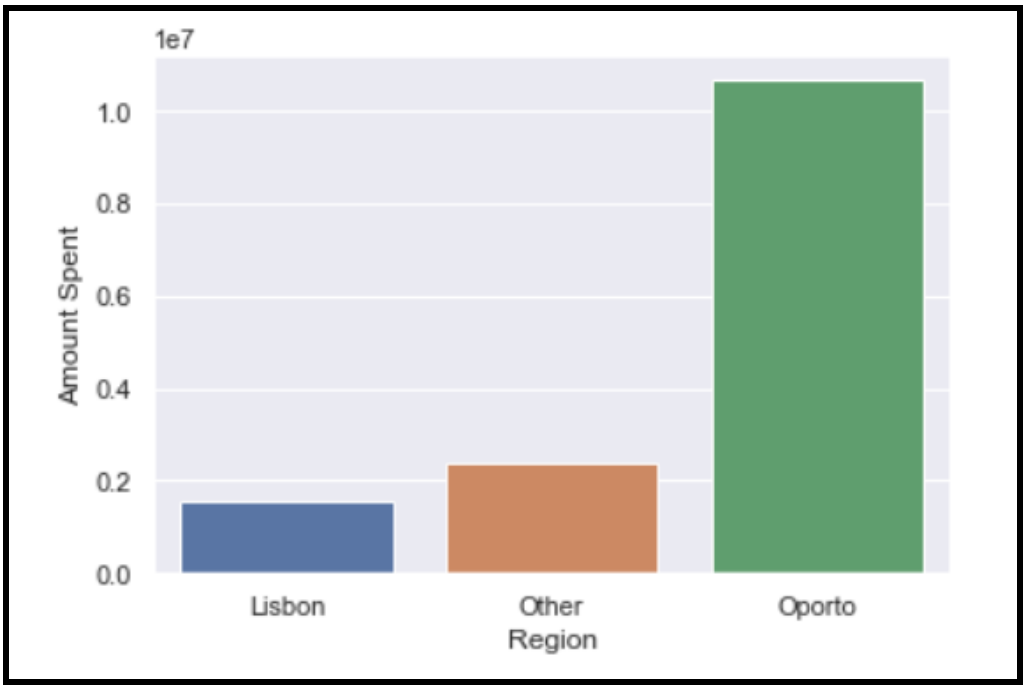
Region	Amount spent annually
Oporto	1,06,77,599
Other	23,86,813
Lisbon	15,55,088

TABLE 3

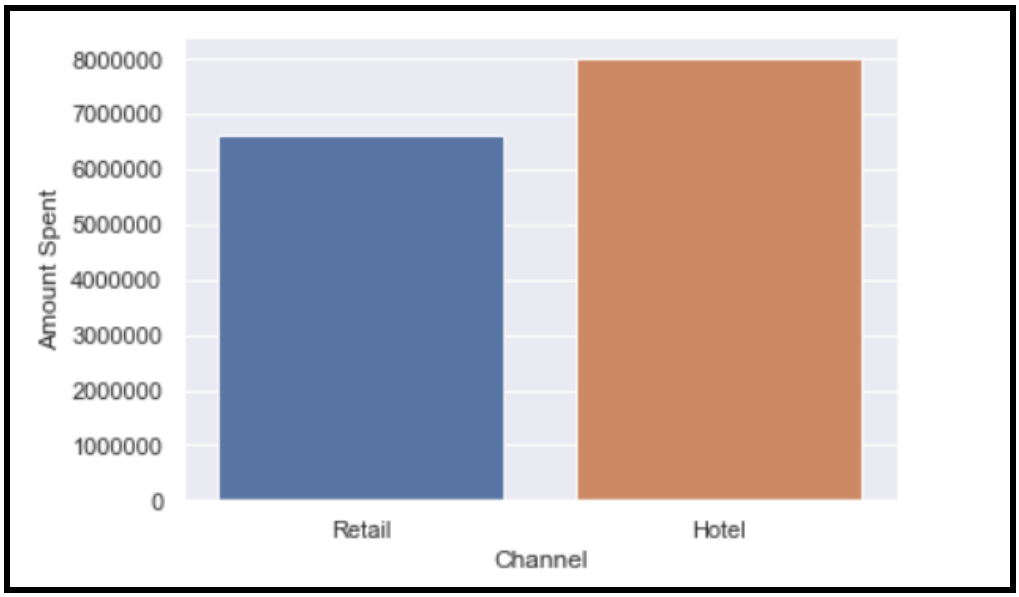
Channel	Amount spent annually
Hotel	79,99,569
Retail	66,19,931

This can be supported by the following graphs for Region (*Graph 1*) and Channel (*Graph 2*) respectively-

GRAPH 1



GRAPH 2



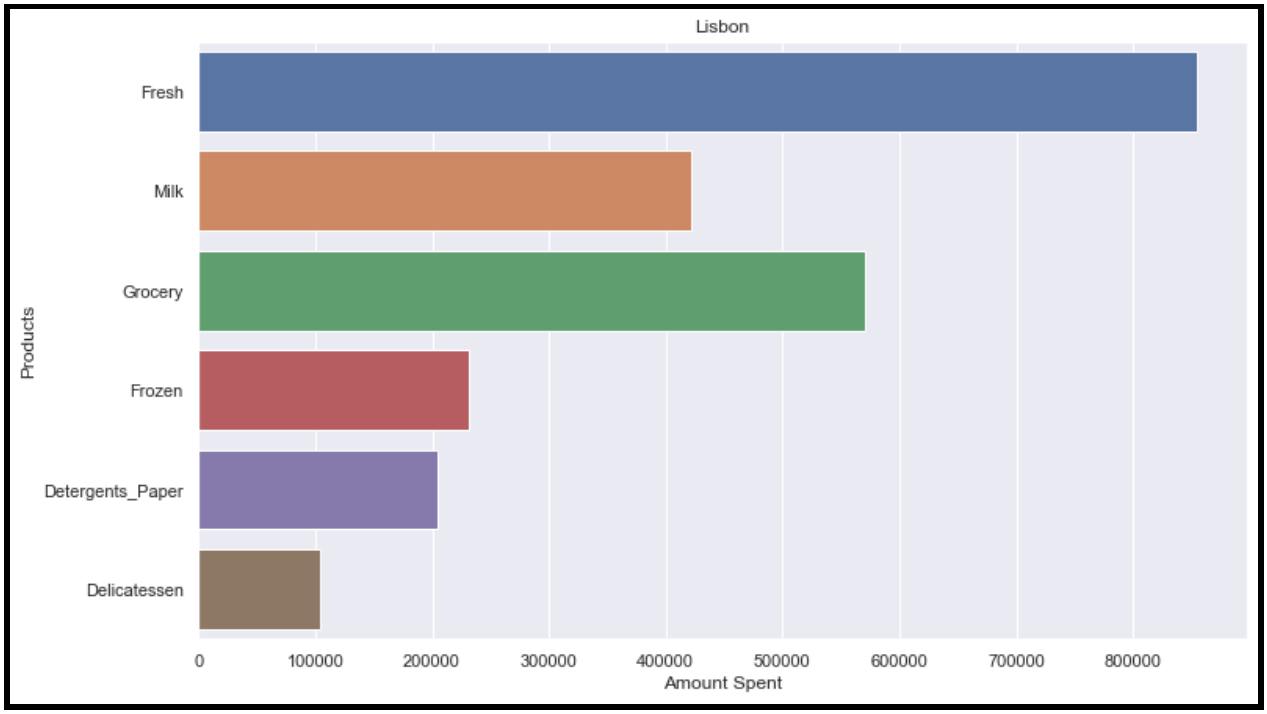
1.2 There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.

REGION

Lisbon
The people in Lisbon spent most in *Fresh* products and least in *Delicatessen* products. The following bar graph gives a pictorial representation to support my comment-

Fresh: 854833
Milk 422454
Grocery: 570037
Frozen: 231026
Detergents Paper: 204136
Delicatessen: 104327

GRAPH 3

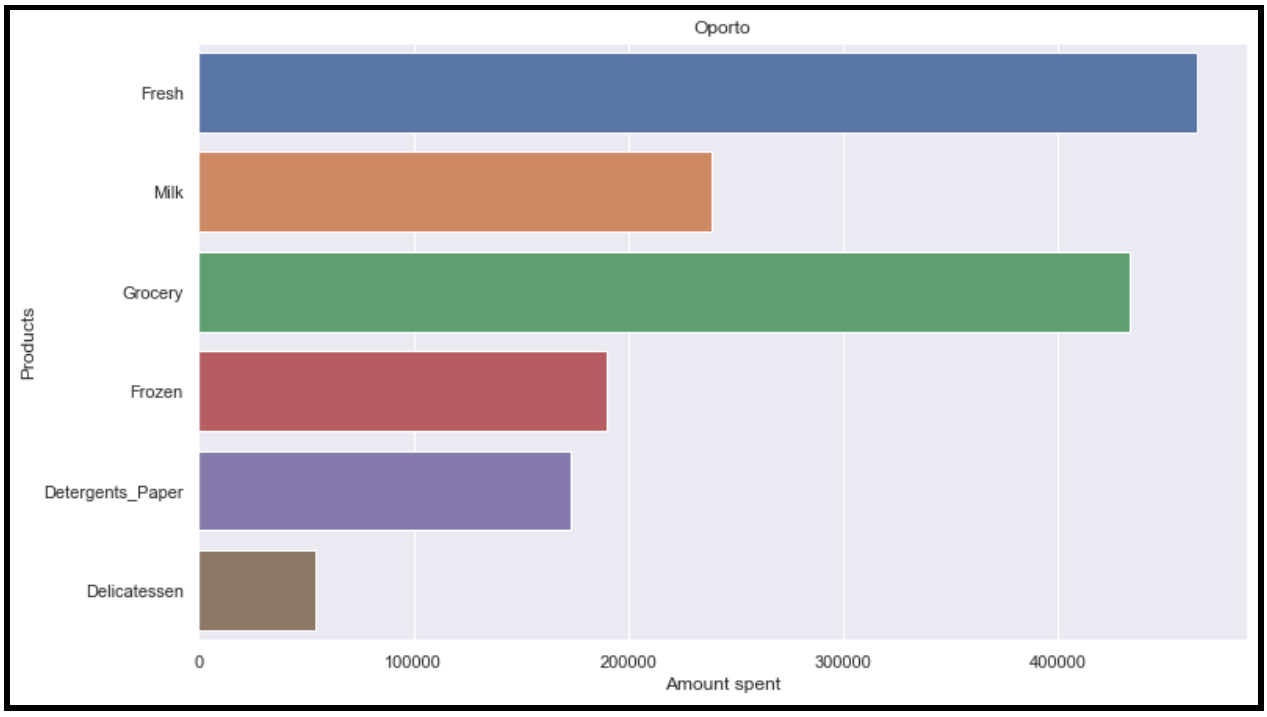


Oporto

The people in Oporto spent most in *Fresh* products and least in *Delicatessen* products. The below bar graph gives a pictorial representation to support my comment-

Fresh: 464721
Milk: 239144
Grocery: 433274
Frozen: 190132
Detergents Paper: 173311
Delicatessen: 54506

GRAPH 4

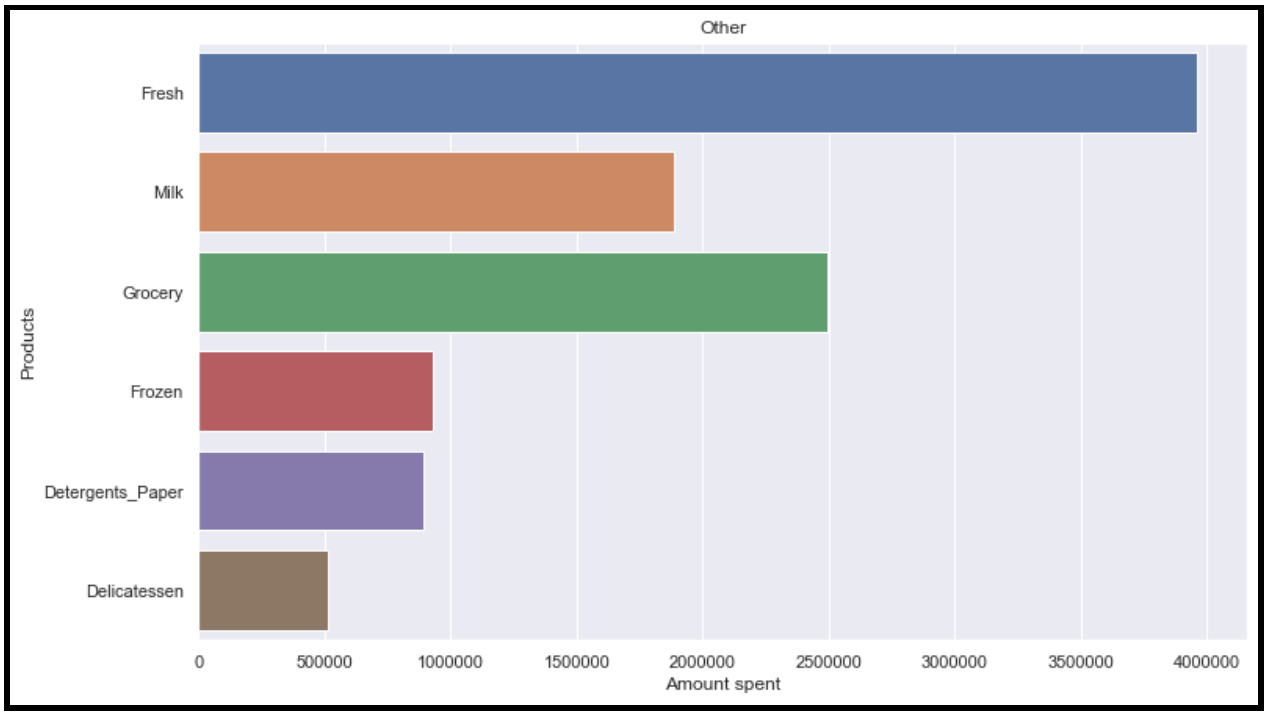


Other

The people in Other regions spent most in *Fresh* products and least in *Delicatessen* products. The below bar graph gives a pictorial representation to support my comment-

Fresh: 3960577
Milk: 1888759
Grocery: 2495251
Frozen: 930492
Detergents Paper: 890410
Delicatessen: 512110

GRAPH 5



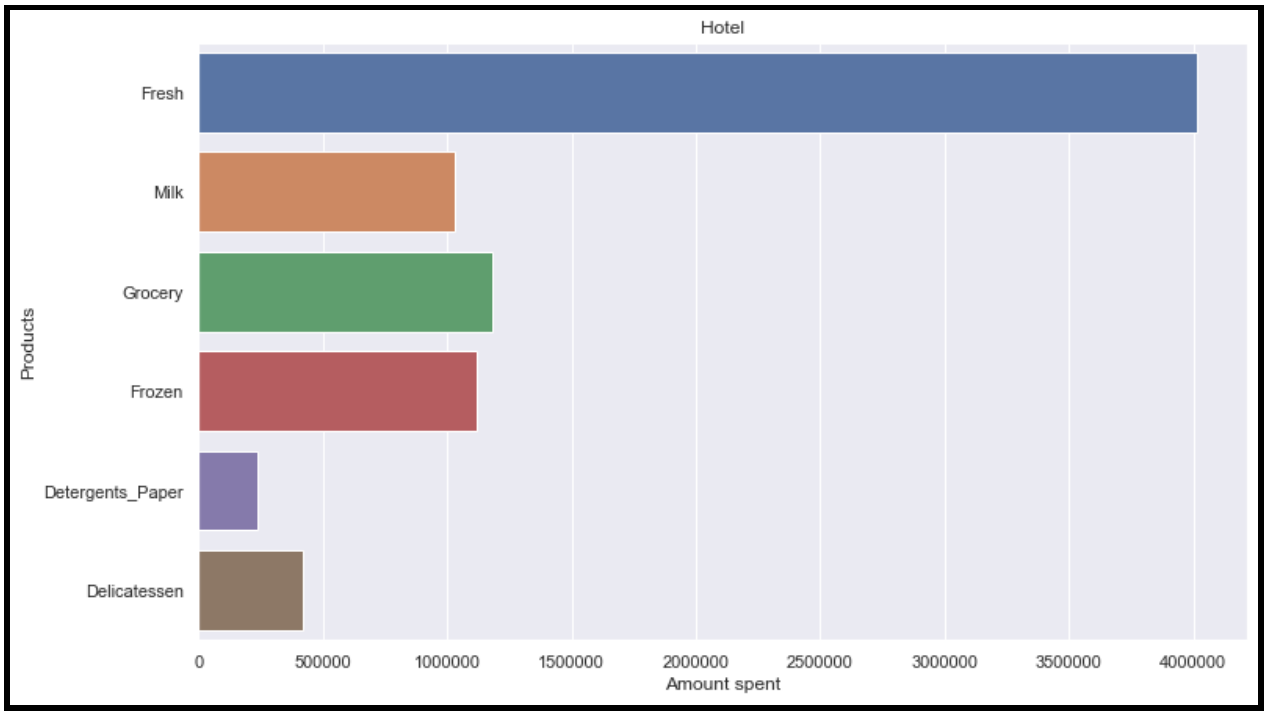
CHANNEL

Hotel

The Hotel channel spent most in *Fresh* products and least in *Detergents Paper* products. The below bar graph gives a pictorial representation to support my comment-

Fresh: 4015717
Milk: 1028614
Grocery: 1180717
Frozen: 1116979
Detergents Paper: 235587
Delicatessen: 421955

GRAPH 6

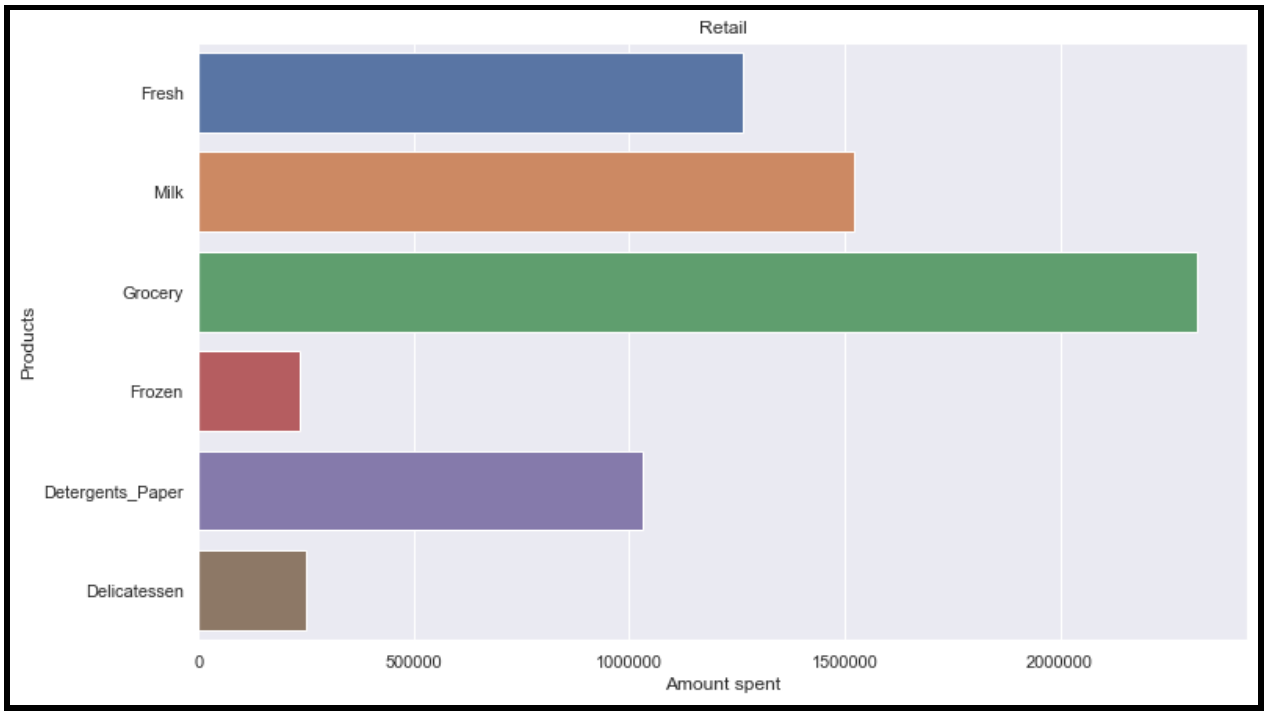


Retail

The Retail channel spent most in *Grocery* products and least in *Frozen* products. The below bar graph gives a pictorial representation to support my comment-

Fresh: 1264414
Milk: 1521743
Grocery: 2317845
Frozen: 234671
Detergents Paper: 1032270
Delicatessen: 248988

GRAPH 7



The table shows the five point summary of the continuous attributes grouped by region (*Table 4*) and channel (*Table 5*) respectively-

TABLE 4

	Region	Lisbon	Oporto	Other
Fresh	count	77	47	316
	mean	11101.72727	9887.680851	12533.47152
	std	11557.43858	8387.899211	13389.21312
	min	18	3	3
	25%	2806	2751.5	3350.75
	50%	7363	8090	8752.5
	75%	15218	14925.5	17406.5
	max	56083	32717	112151
Milk	count	77	47	316
	mean	5486.415584	5088.170213	5977.085443
	std	5704.856079	5826.343145	7935.463443
	min	258	333	55
	25%	1372	1430.5	1634
	50%	3748	2374	3684.5
	75%	7503	5772.5	7198.75
	max	28326	25071	73498
Grocery	count	77	47	316
	mean	7403.077922	9218.595745	7896.363924
	std	8496.287728	10842.74531	9537.287778
	min	489	1330	3
	25%	2046	2792.5	2141.5
	50%	3838	6114	4732
	75%	9490	11758.5	10559.75
	max	39694	67298	92780
Frozen	count	77	47	316
	mean	3000.337662	4045.361702	2944.594937
	std	3092.143894	9151.784954	4260.126243
	min	61	131	25
	25%	950	811.5	664.75
	50%	1801	1455	1498
	75%	4324	3272	3354.75
	max	18711	60869	36534

	count	77	47	316
	mean	2651.116883	3687.468085	2817.753165
	std	4208.462708	6514.717668	4593.051613
	min	5	15	3
	25%	284	282.5	251.25
	50%	737	811	856
	75%	3593	4324.5	3875.75
Detergents_Paper	**max**	19410	38102	40827
	count	77	47	316
	mean	1354.896104	1159.702128	1620.601266
	std	1345.42334	1050.739841	3232.58166
	min	7	51	3
	25%	548	540.5	402
	50%	806	898	994
	75%	1775	1538.5	1832.75
Delicatessen	**max**	6854	5609	47943

TABLE 5

	Channel	**Hotel**	**Retail**
Buyer/Spender	**count**	298	142
	mean	238.369128	183
	std	120.910343	132.136132
	min	4	1
	25%	137.25	61.25
	50%	241.5	166.5
	75%	344.5	303.75
	max	440	438
Fresh	**count**	298	142
	mean	13475.5604	8904.323944
	std	13831.6875	8987.71475
	min	3	18
	25%	4070.25	2347.75
	50%	9581.5	5993.5
	75%	18274.75	12229.75
	max	112151	44466

Milk	**count**	298	142
	mean	3451.724832	10716.5
	std	4352.165571	9679.631351
	min	55	928
	25%	1164.5	5938
	50%	2157	7812
	75%	4029.5	12162.75
	max	43950	73498
Grocery	**count**	298	142
	mean	3962.137584	16322.85211
	std	3545.513391	12267.31809
	min	3	2743
	25%	1703.75	9245.25
	50%	2684	12390
	75%	5076.75	20183.5
	max	21042	92780
Frozen	**count**	298	142
	mean	3748.251678	1652.612676
	std	5643.9125	1812.803662
	min	25	33
	25%	830	534.25
	50%	2057.5	1081
	75%	4558.75	2146.75
	max	60869	11559
Detergents_Paper	**count**	298	142
	mean	790.560403	7269.507042
	std	1104.093673	6291.089697
	min	3	332
	25%	183.25	3683.5
	50%	385.5	5614.5
	75%	899.5	8662.5
	max	6907	40827

Delicatessen	**count**	298	142
	mean	1415.956376	1753.43662
	std	3147.426922	1953.797047
	min	3	3
	25%	379	566.75
	50%	821	1350
	75%	1548	2156
	max	47943	16523

INFERENCE:

It is clear that, in all the regions *Fresh* products are given more importance and *Delicatessen (processed food)* products are given least importance. But the channels show a different behaviour. While the Hotels still give importance to *Fresh* products, the least important product here is *Detergents Paper* products whereas the Retail channel gives importance to *Grocery* products and *Frozen* products are purchased the least. The numbers are highly skewed in each product as there is a huge difference when the mean, standard deviation and maximum is taken into consideration.

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?

The *Fresh* product shows the least inconsistent behaviour and the *Delicatessen* product shows the most inconsistent behaviour.

This can be proved by calculating the Coefficient of variation for all the six products. The Coefficient of variation is given by the ratio of *Standard deviation* to *mean* (*Figure 3*).

FIGURE 3

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Standard deviation	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
Mean	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
Coefficient of variation	1.053918	1.273299	1.195174	1.580332	1.654647	1.849407

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

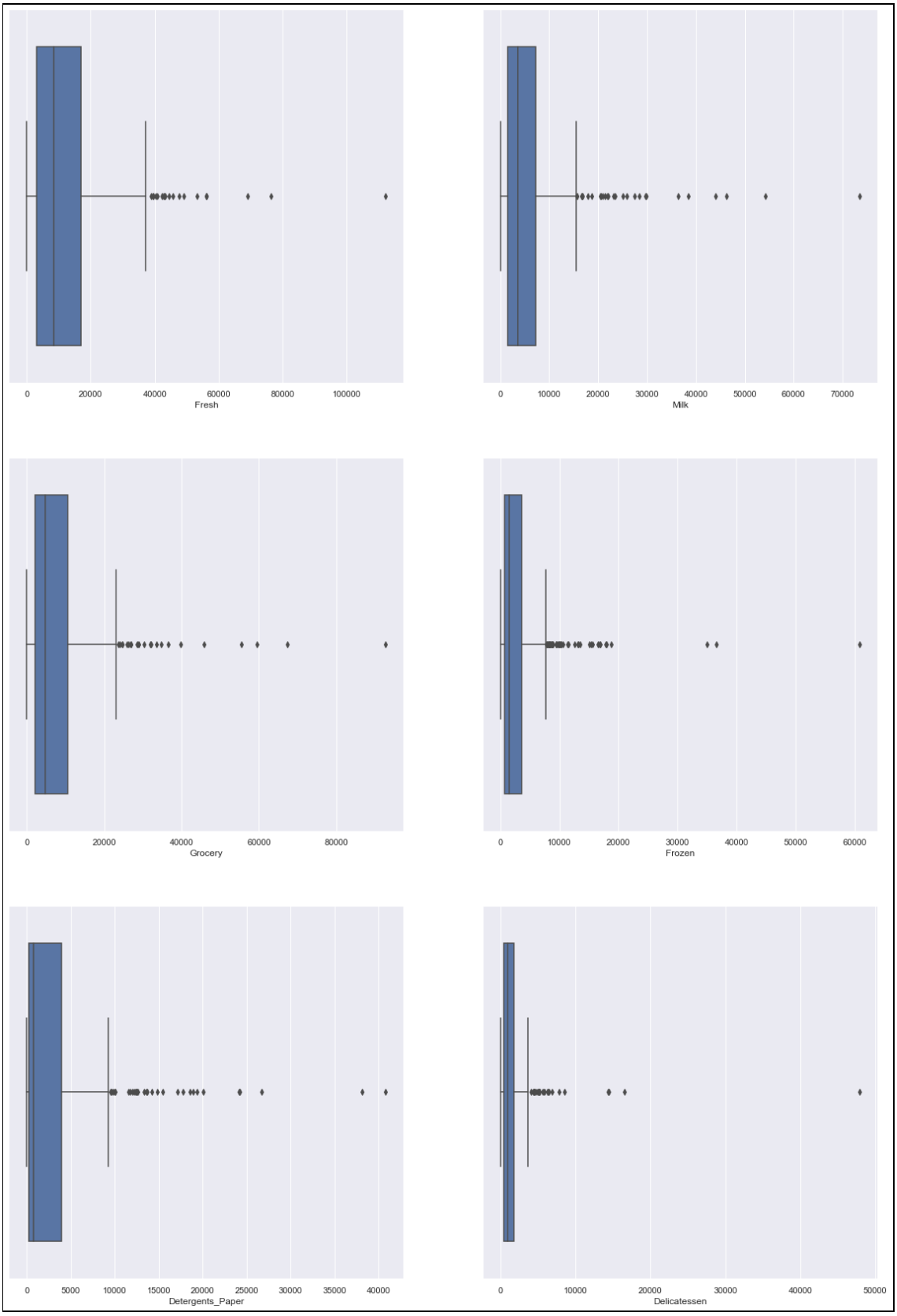
Yes, all the products (Fresh, Milk, Grocery, Frozen, Detergents Paper, Delicatessen) have outliers in it. This can be backed up by the following box plots (*Graph 8*). Also, it is clear from the graph that all are right skewed box plots. This means that the average (mean) amount spent is greater than the median.

To further support my point the below skewness table shows that all the products have a positive skewness value. This can tell us that the data definitely has outliers.

FIGURE 4

	Skewness
Fresh	2.552583
Milk	4.039922
Grocery	3.575187
Frozen	5.887826
Detergents_Paper	3.619458
Delicatessen	11.113534

GRAPH 8


0
20000
40000
60000
80000
100000
Fresh
0
10000
20000
30000
40000
50000
60000
70000
Milk
0
20000
40000
60000
80000
Grocery
0
10000
20000
30000
40000
50000
60000
Frozen
0
5000
10000
15000
20000
25000
30000
35000
40000
Detergents_Paper
0
10000
20000
30000
40000
50000
Delicatessen

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

In most of the analysis, I can see that *Fresh* products are the most purchased by the customers. The *Grocery* products are second best in all the regions and channels. It can be understood that most of the people give importance to their health. Considering this as a factor, the suppliers can focus on improving the sales of these products by coming up with a sales campaign promoting health. The least purchased products are *Delicatessen* and *Frozen.* Processed food is mostly bought by the people who do not have much time to cook. So an advertisement promoting the advantages of using processed foods will encourage working professionals to buy them that will save them a lot of time.

Problem 2

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.

Data Description-

ID: Student ID (continuous from 1 to 62)

Gender: Female, Male

Age: The age of the student (ranging from 18 to 26)

Class: Junior, Senior, Sophomore

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

Grad Intention: Yes, Undecided, No

GPA: The GPA of the student (ranging from 2.3 to 3.5)

Employment: Full-Time, Part-Time, Unemployed

Salary: The salary of the student (ranging from 25 to 80)

Social Networking: The social networking interest (ranging from 0 to 4)

Satisfaction: The satisfaction rating of the student (ranging from 1 to 6)

Spending: The expense of the student (ranging from 100 to 1400)

Computer: Laptop, Tablet, Desktop

Text Messages: The number of text messages of the student (ranging from 0 to 900)

Sample of the dataset-

FIGURE 5

ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
1	Female	20	Junior	Other	Yes	2.9	Full-Time	50.0	1	3	350	Laptop	200
2	Male	23	Senior	Management	Yes	3.6	Part-Time	25.0	1	4	360	Laptop	50
3	Male	21	Junior	Other	Yes	2.5	Part-Time	45.0	2	4	600	Laptop	200
4	Male	21	Junior	CIS	Yes	2.5	Full-Time	40.0	4	6	600	Laptop	250
5	Male	23	Senior	Other	Undecided	2.8	Unemployed	40.0	2	4	500	Laptop	100

There are 14 variables out of which 6 are categorical and 8 are continuous. There are no null values.

Types of variables in the data frame-

TABLE 6

ID	int64	Continuous
Gender	object	Categorical
Age	int64	Continuous
Class	object	Categorical
Major	object	Categorical
Grad Intention	object	Categorical
GPA	int64	Continuous
Employment	object	Categorical
Salary	int64	Continuous
Social Networking	int64	Continuous
Satisfaction	int64	Continuous
Spending	int64	Continuous
Computer	object	Categorical
Text Messages	int64	Continuous

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

2.1.1. Gender and Major

The contingency table for Gender and Major is below-

FIGURE 6

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

The contingency table for Gender and Grad Intention is below-

FIGURE 7

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

2.1.3. Gender and Employment

The contingency table for Gender and Employment is below-

FIGURE 8

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

2.1.4. Gender and Computer

The contingency table for Gender and Computer is below-

FIGURE 9

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

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

2.2.1. What is the probability that a randomly selected CMSU student will be male?

P (M): Probability that a randomly selected CMSU student will be male

Number_of_male: 29
Total number of students: 62

P(M) = 29 / 62
= 0.46774193548387094

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

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

P (F): Probability that a randomly selected CMSU student will be female

Number of female: 33
Total number of students: 62

P(F) = 33 / 62
= 0.532258064516129

The probability that a randomly selected CMSU student will be female is *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.

The below table gives the conditional probability of student taking different majors among the male students in CMSU-

TABLE 7

Major	Conditional probability
Accounting	0.13793103448275862
CIS	0.034482758620689655
Economics/Finance	0.13793103448275862
International Business	0.06896551724137931
Management	0.20689655172413793
Other	0.13793103448275862
Retailing/Marketing	0.1724137931034483
Undecided	0.10344827586206896

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

The below table gives the conditional probability of student taking different majors among the female students in CMSU-

TABLE 8

Major	Conditional probability
Accounting	0.09090909090909091
CIS	0.09090909090909091
Economics/Finance	0.21212121212121213

International Business	0.12121212121212122
Management	0.12121212121212122
Other	0.09090909090909091
Retailing/Marketing	0.2727272727272727
Undecided	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.

P (M): Probability that a randomly selected CMSU student will be male
P (G): Probability that a randomly selected CMSU student intends to graduate

Number of male: 29
Number of male students intends to graduate: 17
Total number of students: 62

$P(G \cap M) = P(G \mid M) \times P(M)$
$= (17 / 29) * (29 / 62)$
$= 0.27419354838709675$

The probability that a randomly chosen student is a male and intends to graduate is *0.27419354838709675.*

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

P (F): Probability that a randomly selected CMSU student will be female
P (L_c): Probability that a randomly selected CMSU student has no laptop

Number of female: 33
Number of students without a laptop: 7
Total number of students: 62

P(L_c ∩ F) = P (L_c | F) x P (F)
= (4 / 33) * (33 / 62)
= 0.06451612903225806

The probability that a randomly chosen student is a female and does not have a laptop is *0.06451612903225806.*

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

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

P (M): Probability that a randomly selected CMSU student will be male
P (FT): Probability that a randomly selected CMSU student has full-time employment

Number of male: 29
Number of students with full-time employment: 10
Number of male students with full-time employment: 7
Total number of students: 62

P(M U FT) = P (M) + P (FT) - P(M ∩ FT)
= (29 / 62) + (10 / 62) - (7 / 62)
= 0.5161290322580645

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

P(IB_M | F): Conditional probability that given a female student is randomly chosen, she is majoring in international business or management

Number of female: 33

Number of female International Business students: 4

Number of female Management students: 4
P(IB_M | F) = (4+4) / (33)
= 0.24242424242424243

The conditional probability that given a female student is randomly chosen, she is majoring in international business or management is *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?

FIGURE 10

Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

The multiplication law states that when two events are independent then,

$$P(F \cap Yes) = P(F)P(Yes)$$

P(F): Probability that a randomly selected CMSU student will be female
P(Yes): Probability that a randomly selected CMSU student intends to graduate
P(F ∩ Yes): Probability that a randomly chosen student is a female and intends to graduate

P(F) = 20 / 40
= 0.5
P(Yes) = 28 / 40
= 0.7

P(F)P(Yes) = (0.5) * (0.7)
= **0.35**

P(F ∩ Yes) = 11 / 40
= **0.275**

$P(F \cap Yes) \neq P(F)P(Yes)$
As the multiplication rule is not satisfied the events are not independent.

The graduate intention and being 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?

P(GPA < 3): Probability that a randomly selected CMSU student has a GPA less than 3

Total number of students: 62
Number of students with GPA less than 3: 17

P(GPA < 3) = 17 / 62
= 0.27419354838709675

The probability that a randomly selected CMSU student has a GPA less than 3 is *0.27419354838709675.*

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

P(Salary >= 50 | M): Conditional probability that a randomly selected male earns 50 or more

Total number of male students: 29
Number of male students who earns 50 or more: 14

P(Salary >= 50 | M) = 14 / 29
= 0.4827586206896552

P(Salary >= 50 | F): Conditional probability that a randomly selected female earns 50 or more

Total number of female students: 33
Number of female students who earns 50 or more: 18

P(Salary >= 50 | F) = 18 / 33
= 0.5454545454545454

The conditional probability that a randomly selected male earns 50 or more is *0.4827586206896552.*
The conditional probability that a randomly selected female earns 50 or more is *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 for this whole Problem 2.

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve. [2]

Properties of a normal distribution

- The mean, mode and median are all equal.
- The curve is symmetric at the center (i.e. around the mean, μ).
- Exactly half of the values are to the left of center and exactly half the values are to the right.
- The total area under the curve is 1. [3]

Skewness is a number that indicates to what extent a variable is asymmetrically distributed. The below table represents the level of skewness and the respective skewness value. [4]

TABLE 9

Skewness level	Value
Symmetrical or Not Skewed	0
Less Skewed Data	± 0.5 to 1
Highly Skewed Data	Greater than ±1

When the skewness value is positive it is considered as right skewed data and when the skewness value is negative it is considered as left skewed data.

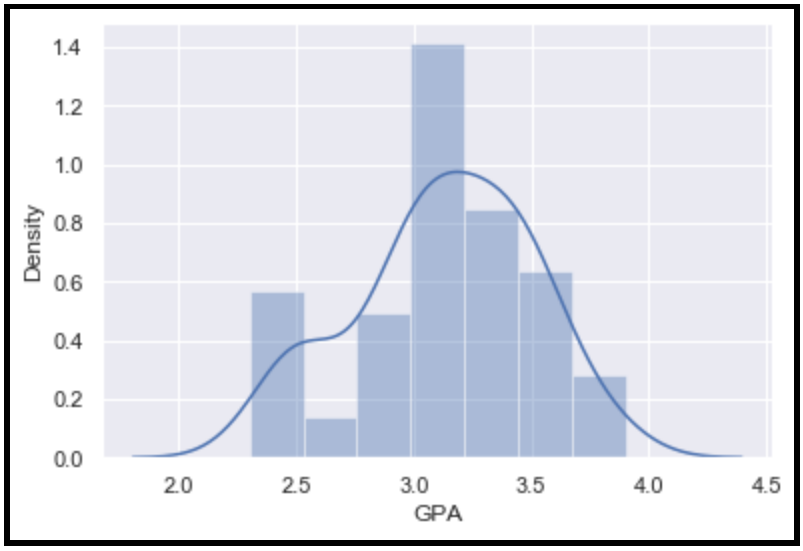
GPA:

TABLE 10

Mean	3.129032
Median	3.150000
Mode	3.0, 3.1, 3.4
Standard deviation	0.377388
Skewness	-0.306937

As the mean, median and mode are not equal, the variable GPA does not follow a normal distribution.
This can also be backed up by the skewness value. As it is a negative value it is skewed in the left.
The following graph can also be considered as a visual representation to understand that the data is not normally distributed.

GRAPH 9



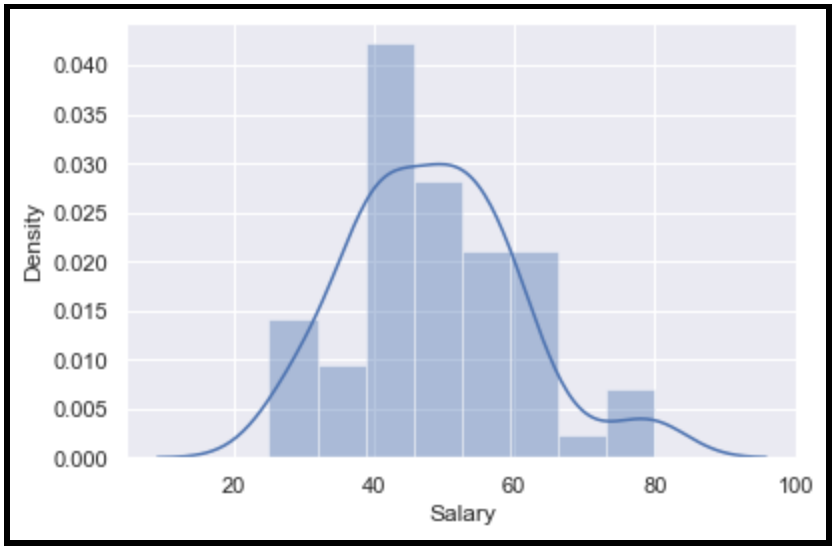
Salary:

TABLE 11

Mean	48.548387
Median	50
Mode	40
Standard deviation	12.080912
Skewness	0.521677

As the mean, median and mode are not equal, the variable Salary does not follow a normal distribution.
This can also be backed up by the skewness value. As it is a positive value it is skewed in the right.
The following graph can also be considered as a visual representation to understand that the data is not normally distributed.

GRAPH 10



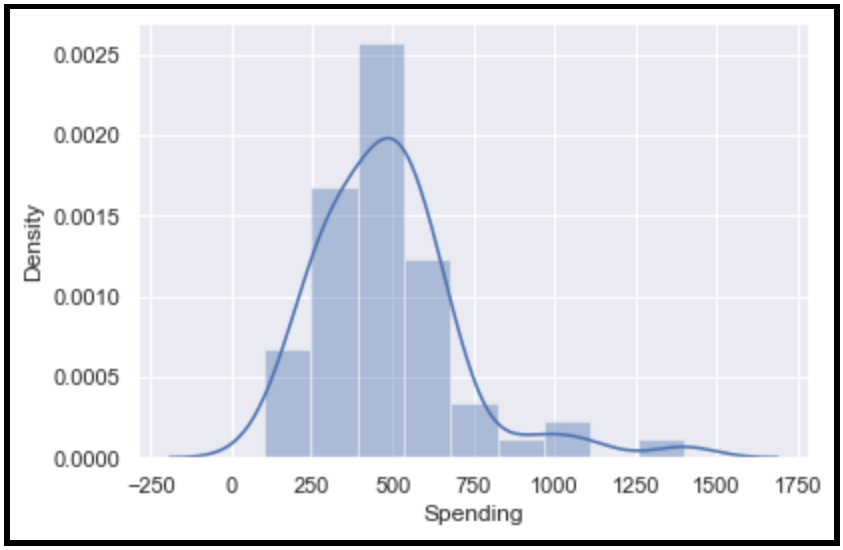
Spending:

TABLE 12

Mean	482.016129
Median	500
Mode	500
Standard deviation	221.953805
Skewness	1.547285

As the mean, median and mode are not equal, the variable Spending does not follow a normal distribution.
This can also be backed up by the skewness value. As it is a positive value it is skewed in the right.
The following graph can also be considered as a visual representation to understand that the data is not normally distributed.

GRAPH 11



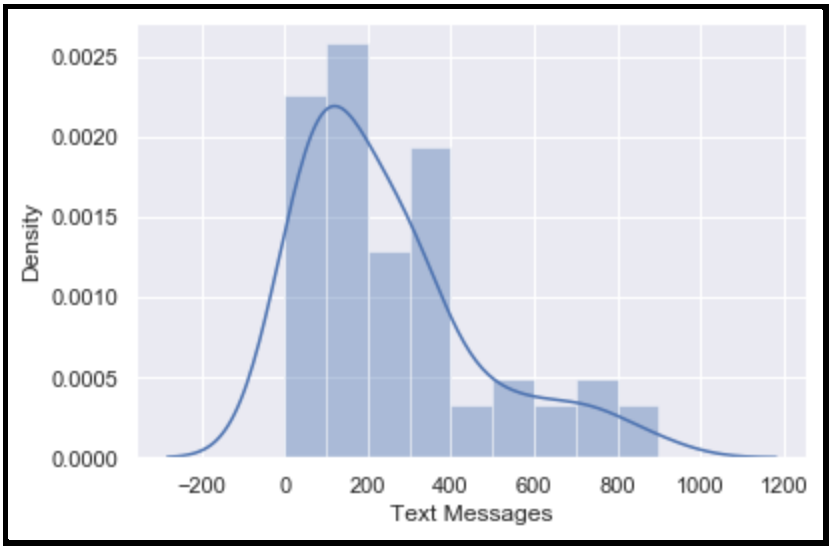
Text Messages:

TABLE 13

Mean	246.209677
Median	200
Mode	300
Standard deviation	214.465950
Skewness	1.264245

As the mean, median and mode are not equal, the variable Text Messages does not follow a normal distribution.
This can also be backed up by the skewness value. As it is a positive value it is skewed in the right.
The following graph can also be considered as a visual representation to understand that the data is not normally distributed.

GRAPH 12



Conclusion:

Even though the number of female students is more than the male students, the intention to graduate is less for females than males.

When it comes to employment as well, the number of employed female students is less than the number of employed male students in terms of both full time and part time employment.

Thus, the female students should be given more attention and should be encouraged more to complete their degree. Also, they should be given more employment opportunities.

Problem 3

Problem statement-

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 colouring 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 is calculated. The company would like to show that the mean moisture content is less than 0.35 pound per 100 square feet.

Data Description-

A: Moisture content for shingle A (ranging from 0.13 to 0.72)

B: Moisture content for shingle B (ranging from 0.1 to 0.58)

Sample of the dataset-

FIGURE 11

A	B
0.44	0.14
0.61	0.15
0.47	0.31
0.30	0.16
0.15	0.37

There are 2 continuous variables. There are 5 null values in column B.

Types of variables in the data frame-

TABLE 14

A	int64	Continuous
B	int64	Continuous

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.

TABLE 15

H_0	Null Hypothesis
H_a	Alternative hypothesis
μ	Hypothesized mean moisture content
α	Significance level

Shingle A

Step -1: Define null and alternative hypotheses

$H_0: \mu$ <=0.35

$H_a: \mu$ > 0.35

Step - 2: Decide the significance level

α = 0.05

Step - 3: Identify the test statistic

We do not know the population standard deviation and n = 36. So we use the t distribution and the t_{STAT} test statistic.

Step - 4: Calculate the p - value and test statistic

TABLE 16

t_{STAT}	-1.4735046253382782
p	0.07477633144907513

Step - 5: Decide to reject or accept null hypothesis

TABLE 17

p	0.07477633144907513
α	0.05

We have no evidence to reject the null hypothesis since p value > level of significance(α).

Therefore, at 95% confidence level, there is sufficient evidence to prove that mean moisture content is less than or equal to 0.35 in shingle A.

Shingle B

Step -1: Define null and alternative hypotheses

H_0: μ <=0.35

H_a: μ > 0.35

Step - 2: Decide the significance level

α = 0.05

Step - 3: Identify the test statistic

We do not know the population standard deviation and n = 31. So we use the t distribution and the t_{STAT} test statistic.

Step - 4: Calculate the p - value and test statistic

TABLE 18

t_{STAT}	-3.1003313069986995

p	0.0020904774003191826

Step - 5: Decide to reject or accept null hypothesis

TABLE 19

p	0.0020904774003191826
α	0.05

We have evidence to reject the null hypothesis since p value < level of significance (α).

Therefore, at 95% confidence level, there is sufficient evidence to prove that mean moisture content is greater than 0.35 in shingle B.

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?

TABLE 20

H_0	Null Hypothesis
H_a	Alternative hypothesis
μ_A	Hypothesized mean moisture content of shingle A
μ_B	Hypothesized mean moisture content of shingle B
α	Significance level

This is a two-sided test for the null hypothesis that 2 independent samples have identical average (expected) values. This test ***assumes*** that the populations have ***identical variances***.

I am going to assume that the ***variance*** is equal for this test and then compute the necessary statistical values.

Step -1: Define null and alternative hypotheses

H_0: $\mu_A = \mu_B$
H_a: $\mu_A \neq \mu_B$

Step - 2: Decide the significance level

α = 0.05

Step - 3: Identify the test statistic

We do not know the population standard deviation and n > 30. So we use the t distribution and the $tSTAT$ test statistic for two sample unpaired test.

Step - 4: Calculate the p - value and test statistic

TABLE 21

t_{STAT}	1.2896282719661123
p	0.2017496571835306

Step - 5: Decide to reject or accept null hypothesis

TABLE 22

p	0.2017496571835306
α	0.05

We have no evidence to reject the null hypothesis since p value > level of significance(α).

Therefore, at 95% confidence level, there is sufficient evidence to prove that mean mean for shingles A and B are equal.

References

Websites-

[1] https://www.investopedia.com/terms/d/descriptive_statistics.asp

[2] https://www.investopedia.com/terms/n/normaldistribution.asp

[3]https://www.statisticshowto.com/probability-and-statistics/normal-distributions/

[4] https://www.spss-tutorials.com/skewness/

Books-

[1] Statistics for business and economics, Eleventh Edition, *David R. Anderson*, *Dennis J. Sweeney*, *Thomas A. Williams.*

End of Project