SMDM PROJECT REPORT 2022

Sangeeth A
PGP-DSBA Online
July - 2022

OCTOBER 19

CONTENTS

LIST OF FIGURES

LIST OF TABLES

PROBLEM 1 - SUMMARY

A wholesale distributor operating in different regions of Portugal has information on the annual spending of several items in their stores across different regions and channels. The dataset consists of retailers' annual spending on 6 different products across 3 different regions and two different channels. In this problem, we will analyze the sales of the products across different channels and regions and discuss the measures to be taken into consideration to increase the sales of the products

INTRODUCTION

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using measures of central tendency, five point summary and other different plot methods and parameters. 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).

DATA DESCRIPTION

1. Buyer/Seller: Numbers(1ST Buyer, 2ND Buyer etc.)
2. Channel: Retail or Hotel
3. Region: Lisbon, Oporto and Other
4. Fresh: Product, total spend
5. Milk: Product, total spend
6. Grocery: Product, total spend
7. Frozen: Product, total spend
8. Detergents Paper: Product, total spend
9. Delicatessen: Product, total spend

SAMPLE OF THE DATASET

	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

Table1.1 – Sample dataset

The dataset has 9 variables, among which the buyer/spender column did not make any specific effect on the entire dataset, it only provides us with the information about how many customers are there so if we want we can remove the column buyer or spender.

EXPLORATORY DATA ANALYSIS

NO.	Column	Data type	Non null count
1	Buyer/Spender	Int64	440 non-null
2	Channel	object	440 non-null
3	Region	object	440 non-null
4	Fresh	Int64	440 non-null
5	Milk	Int64	440 non-null
6	Grocery	Int64	440 non-null
7	Frozen	Int64	440 non-null
8	Detergents_Paper	Int64	440 non-null
9	Delicatessen	Int64	440 non-null

Range Index: 440 entries, 0 to 439
Data columns (total 9 columns):

There are 440 entries and 9 columns are in the dataset and there is no null values are there in the dataset

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?

Column Name	Fresh
Minimum Value	3
25%	3127.75
50% or Median	8504.0
75%	16933.75
Maximum Value	112151
IQR	13806.0

Column Name	Milk
Minimum Value	55
25%	1533.0
50% or Median	3627.0
75%	7190.25
Maximum Value	73498
IQR	5657.25

Column Name	Grocery
Minimum Value	3
25%	2153.0
50% or Median	4755.5
75%	10655.75
Maximum Value	92780
IQR	8502.75

Column Name	Frozen
Minimum Value	25
25%	742.25
50% or Median	1526.0
75%	3454.25
Maximum Value	60869
IQR	2812.0

Column Name	Detergents_Paper
Minimum Value	3
25%	256.75
50% or Median	816.5
75%	3922.0
Maximum Value	40827
IQR	1412.0

Column Name	Delicatessen
Minimum Value	3
25%	408.25
50% or Median	965.5
75%	11820.25
Maximum Value	47943
IQR	1412.0

Table 1.1.1. Data Description along different products

We need find which channel and region spent the most and which channel and region s pent the least.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Buyer/Spender	440.00	NaN	NaN	NaN	220.50	127.16	1.00	110.75	220.50	330.25	440.00
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	440.00	NaN	NaN	NaN	12000.30	12647.33	3.00	3127.75	8504.00	16933.75	112151.00
Milk	440.00	NaN	NaN	NaN	5796.27	7380.38	55.00	1533.00	3627.00	7190.25	73498.00
Grocery	440.00	NaN	NaN	NaN	7951.28	9503.16	3.00	2153.00	4755.50	10655.75	92780.00
Frozen	440.00	NaN	NaN	NaN	3071.93	4854.67	25.00	742.25	1526.00	3554.25	60869.00
Detergents_Paper	440.00	NaN	NaN	NaN	2881.49	4767.85	3.00	256.75	816.50	3922.00	40827.00
Delicatessen	440.00	NaN	NaN	NaN	1524.87	2820.11	3.00	408.25	965.50	1820.25	47943.00

Table 1.1.2. Data Description

The above table 1.1.2 helps to describe the data. From table we can easily understands t he mean, standard deviation, minimum value and maximum value of different products , the same shown in table 1.1.1 too. From the table we can conclude that

- There are total 440 count values are present all the variables.
- There are two unique value present in channel (Hotel and Retail).
- Out of the two unique values in channel hotel has top count than the other.
- There are total 3 unique value present in region (Lisbon, Oporto, Other).
- The region ‘Other’ top among the 3 unique values in region.
- The minimum value among the products is 3 and for products Fresh, Grocery, Det ergents_paper, Delecatessen the minimum value is same that is 3.
- The maximum value among the products is 112151.0, it is for the item fresh.

To find out the maximum and minimum spending channel and region, we will add new c olumn total spending. The updated dataset sample is shown below.

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total_spending
0	1	Retail	Other	12669	9656	7561	214	2674	1338	34112
1	2	Retail	Other	7057	9810	9568	1762	3293	1776	33266
2	3	Retail	Other	6353	8808	7684	2405	3516	7844	36610
3	4	Hotel	Other	13265	1196	4221	6404	507	1788	27381
4	5	Retail	Other	22615	5410	7198	3915	1777	5185	46100

Table 1.1.3. Updated dataset sample

Let’s analyse the data with respect to region and total spending, and find out which regi on spent most and which region spent the least.

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

Table 1.1.4. Dataset with respect to region and spending

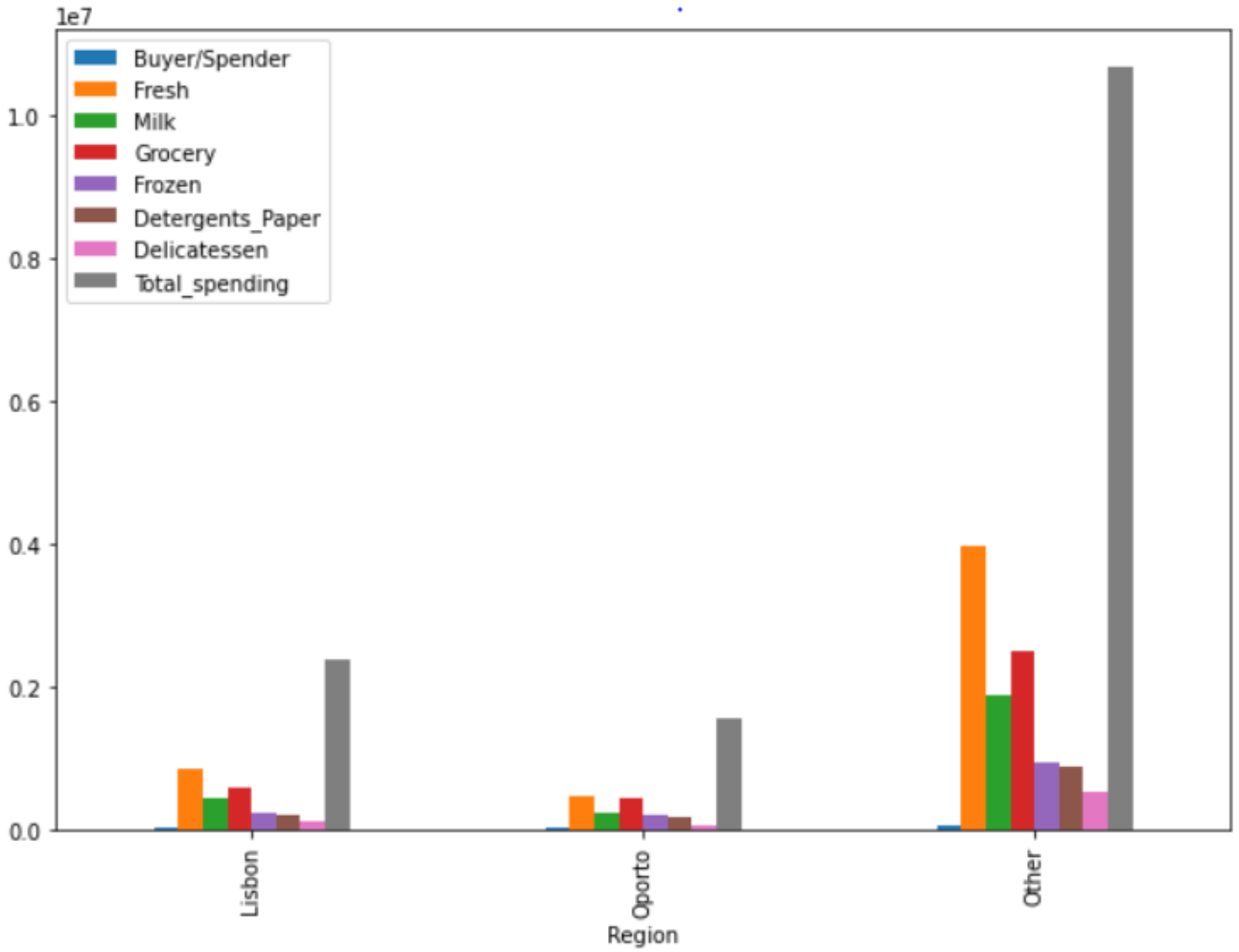

Fig. 1.1.1 – Region v/s spending bar plot

Let’s analyse the data with respect to channel and total spending, and find out which ch annel spent most and which region spent the least.

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

Table 1.1.5. Dataset with respect to channel and spending

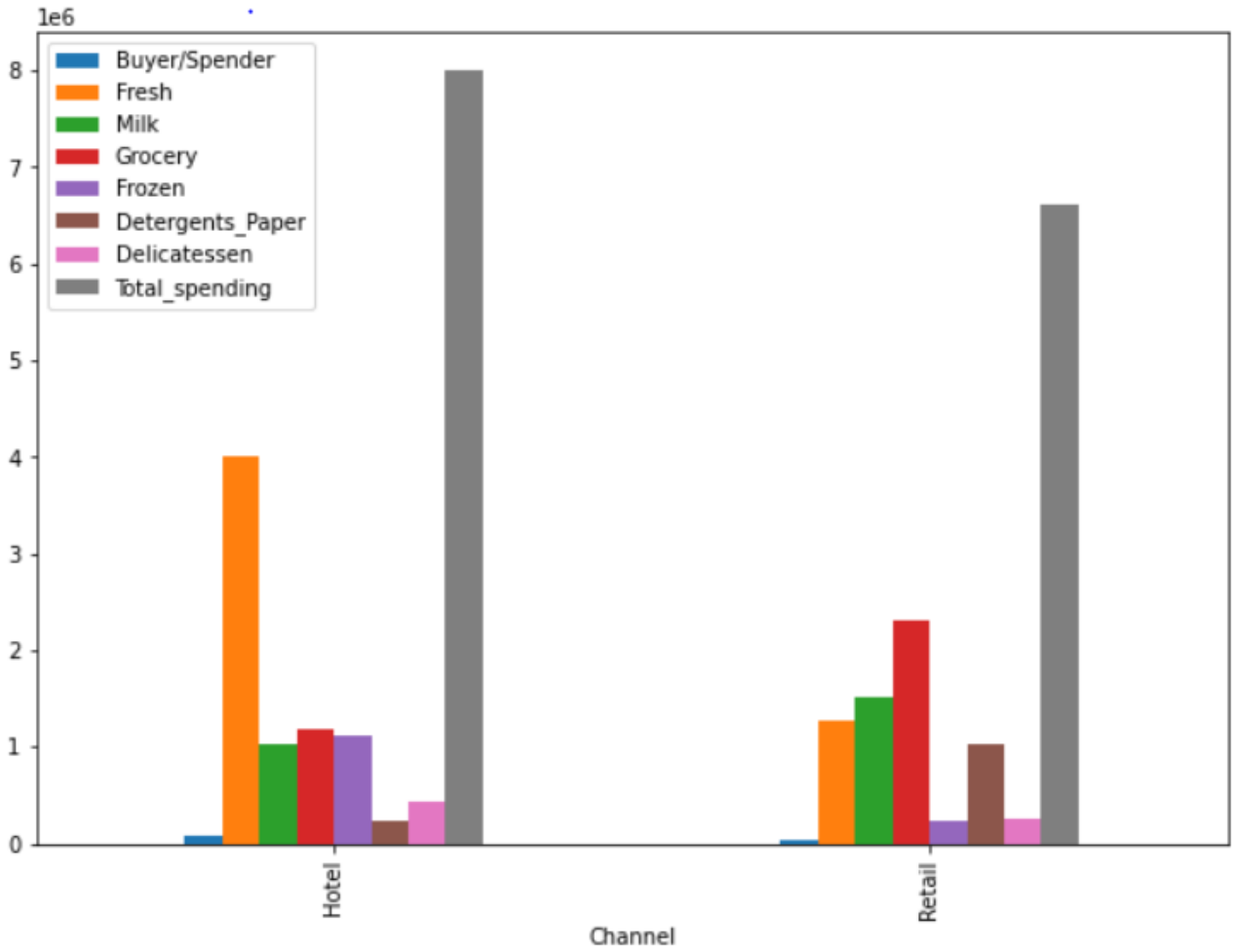

Fig. 1.1.2 – Channel v/s spending bar plot

By analyzing the above two bar plots we can conclude that,

- The region other spends the most
- The region Oporto spends the least
- The channel hotel spends the most
- The channel Retail spends the least

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.

We have 6 different varieties of items are there in the dataset, let us consider the same table 1.1.4 by removing the total spending column.

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

Table 1.2.1. Dataset with respect to region and spending

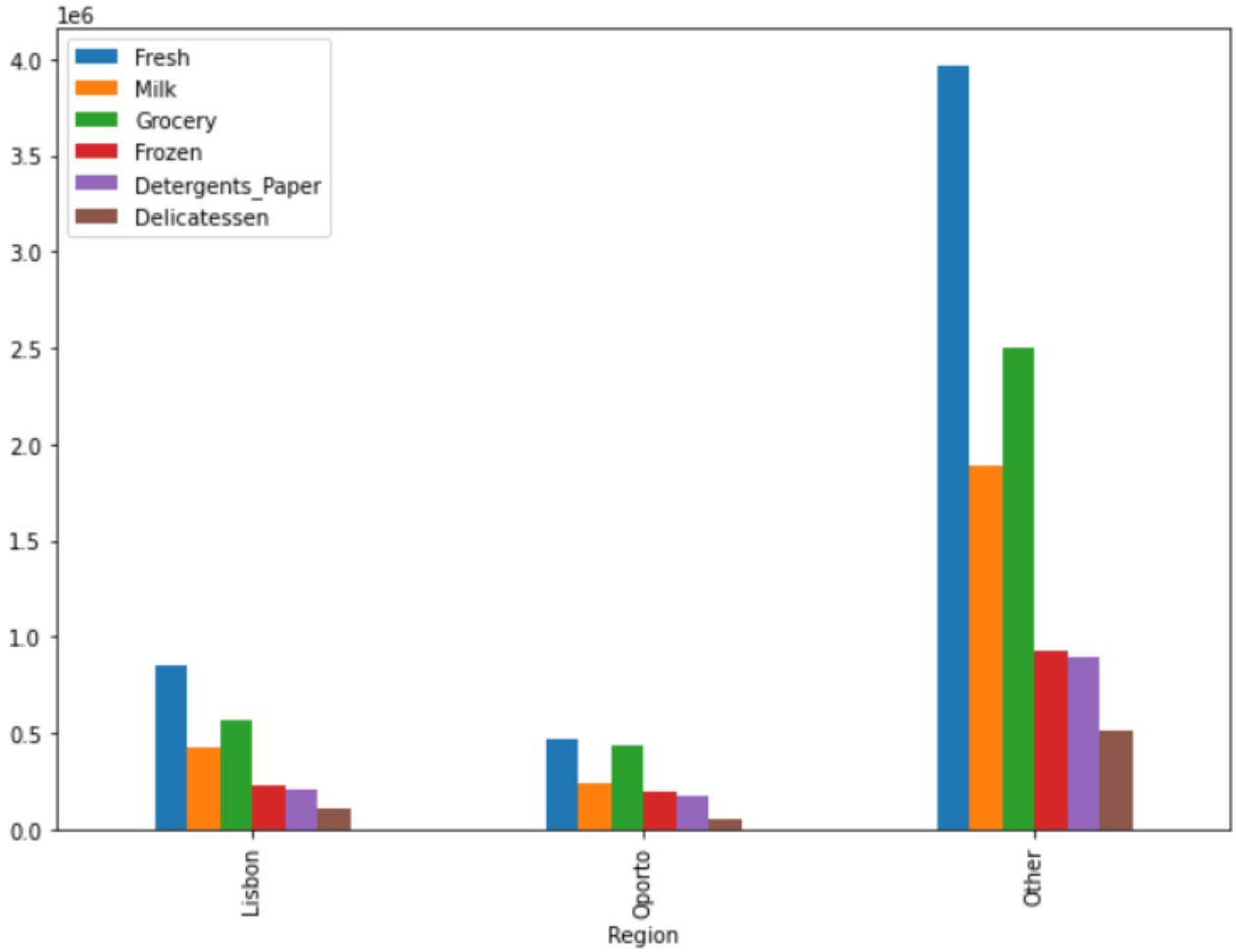

Fig. 1.2.1 – Region v/s spending bar plot

By analyzing the above region v/s spending bar plot we can conclude that

- In all the three regions people spend more on the fresh items
- The second favorite item in all the three region is grocery
- The third favorite item is milk
- The least favorite item in all the three region is frozen
- In the region other the spending is more on all the items considering the other two region

Now we can analyze the data with respect to channel and spending on different varieties

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

Table 1.2.2. Dataset with respect to Channel and spending

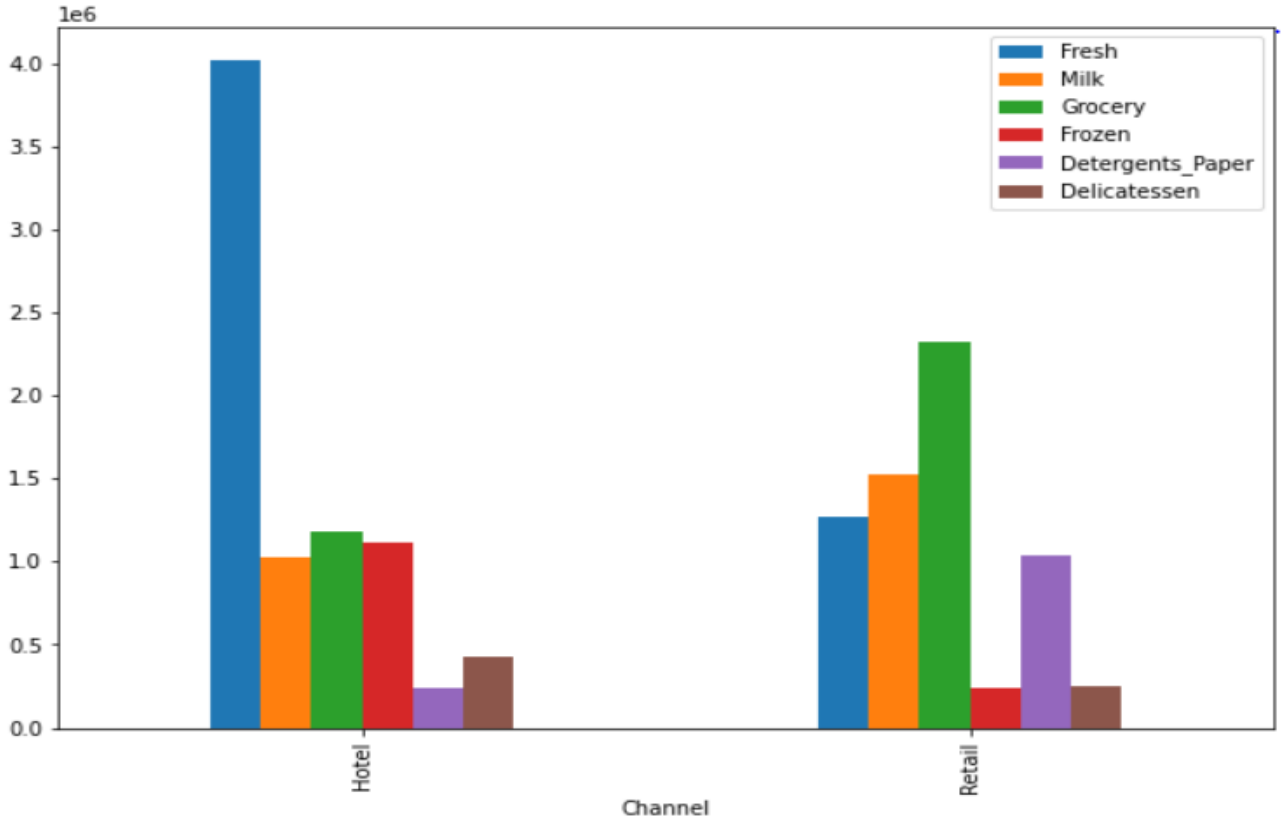

Fig. 1.2.2 – Channel v/s spending bar plot

By analyzing the above bar plot and table we can conclude that,

- Fresh item has more demand in channel hotel whereas retail channel item grocery has more demand.
- Least favorite item in channel hotel is detergent paper.
- In retail channel, the least favorite items are frozen and delicatessen.

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?

Let us check the standard deviation of the data

Item	Standard deviation
Fresh	12647.33
Milk	7380.38
Grocery	9503.16
Frozen	4854.67
Detergents_paper	4767.85
Delicatessen	2820.11

Table 1.3.1. Standard deviation

From the table the fresh item have more standard deviation.

To find out the item showing the most inconsistent and least inconsistent behavior let us calculate the coefficient of variance of the different items in the dataset,

Item	Co-eff of variance
Fresh	1.05
Milk	1.27
Grocery	1.20
Frozen	1.58
Detergents_paper	1.65
Delicatessen	1.85

Table 1.3.2. Coefficient of variance of items

We know that the co-efficient of variance and inconsistency is directly related, the co-efficient of variance of a item increases means the inconsistency of the item also increases.

Here we have to find the most inconsistent and least inconsistent item in the dataset while referring the table 1.3.1 we can see that the fresh item has less co-efficient of variance(1.05), so it is the least inconsistent item, whereas the item delicatessen has more co-efficient of variance(1.85), therefore it is the most inconsistent item in the dataset.

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

To find out the outliers in the data let us consider the boxplot method

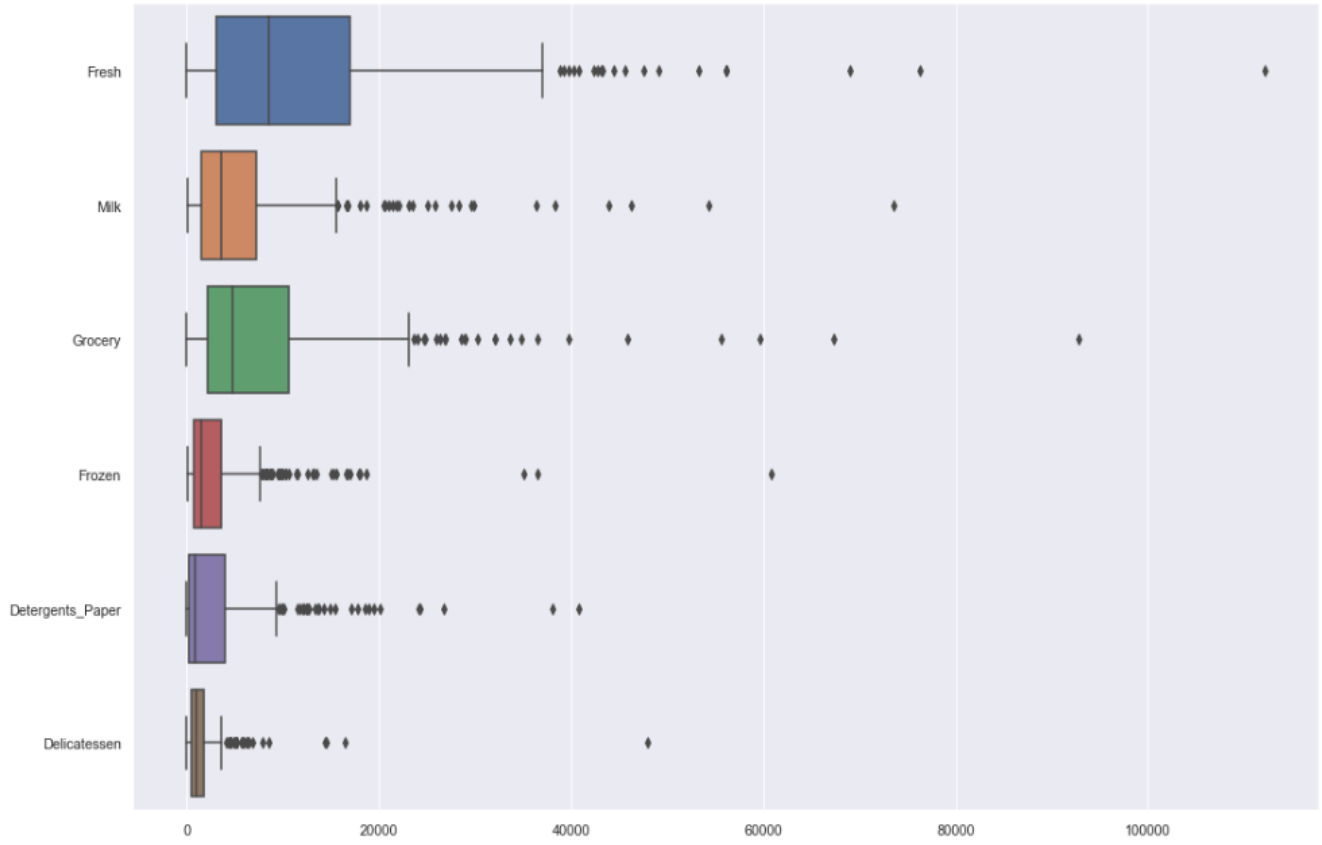

Fig. 1.4.1 – Item v/s spending box plot

From the above boxplot it is clearly visible that all the items have outliers.

Let us consider the IQR values of different items,

No.	Item	IQR
1	Fresh	13806.00
2	Milk	5657.25
3	Grocery	8502.75
4	Frozen	2812.00
5	Detergents_paper	3665.25
6	Delicatessen	1412.00

Table 1.4.1. IQR

From table 1.1.2 data description and table 1.4.1 IQR table let us check the values are coming within the range,

	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

Table 1.4.2. Outlier checking table

The true value from the table indicates the outlier. So from the box plot as well as the tables we can conclude that there are outliers are present in the data.

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.

The data provided contains 6 varieties of items which sales across 3 main regions Lisbon, Oporto and Other, the sale is mainly concentrated in Hotel and retail region. By analyzing the data it is clear that more consumers are present in Other region and less consumers are present in Oporto region, Fresh items are most favorite in all the region and

Delicatessen is the least favorite item among the customers. The sale of the items are limited to only two areas hotel and region and we have to find more sale areas to increase the sale.

Since Fresh items have more sale among other items consider increase the stock of the fresh items. Delicatessen, frozen and detergents paper are the least favorite items in the list so keep the stock of these items less considering other items and find out where these items are selling more and concentrate to increase the sale in those areas. For frozen items storage as well as electricity charge will affect the profit so this also need to taken into consideration before purchasing the stock.

PROBLEM 2 - SUMMARY

The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. They created a survey which contains 14 questions and receives responses from 62 undergraduates. In this problem, we will prepare different contingency tables according to the variables and find out the probabilities of different conditions.

INTRODUCTION

The purpose of this exercise is to explore the given dataset and prepare different contingency tables according to the answers to the 14 survey questions from 62 undergraduates. We also need to calculate the probability of different conditions.

DATA DESCRIPTION

1. ID: Numbers (1, 2, 3…)
2. Gender: Male or Female
3. Age: The age of the student
4. Class: Class of students (Junior, senior, and Sophomore)
5. Major: The area of study of the student
6. Grad intention: Whether the student is intended to graduate (yes, no, or undecided)
7. GPA: GPA of the student
8. Employment: Employment of the student (Full time, part-time or unemployed)
9. Salary: Current salary of the student
10. Social networking: How many social networking sites they are active
11. Satisfaction: Rating of how satisfied they are
12. Spending: How much money they are spending
13. Computer: Do they own a computer
14. Text messages: No. of text messages they are sending

SAMPLE OF THE DATASET:

	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

Table 2.1 – Sample dataset of survey

The dataset contains 14 different survey questions and receives responses from 62 undergraduates

EXPLORATORY DATA ANALYSIS

NO.	Column	Data type	Non null content
1	ID	Int64	62 non-null
2	Gender	object	62 non-null
3	Age	Int64	62 non-null
4	Class	object	62 non-null
5	Major	object	62 non-null
6	Grad Intention	object	62 non-null
7	GPA	float64	62 non-null
8	Employment	object	62 non-null
9	Salary	float64	62 non-null
10	Social Networking	Int64	62 non-null
11	Satisfaction	Int64	62 non-null
12	Spending	Int64	62 non-null
13	Computer	object	62 non-null
14	Text message	Int64	62 non-null

Table 2.2 – Survey data info.

Range Index: 62 entries, 0 to 61

Data columns (total 14 columns)

There are total 62 rows and 14 columns are there in the dataset, and also from the above table we can see that there is no null values present in the dataset.

DESCRIPTIVE DATA ANALYSIS

Let us check the descriptive analysis of the data,

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
ID	62.00	NaN	NaN	NaN	31.50	18.04	1.00	16.25	31.50	46.75	62.00
Gender	62	2	Female	33	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Age	62.00	NaN	NaN	NaN	21.13	1.43	18.00	20.00	21.00	22.00	26.00
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	62.00	NaN	NaN	NaN	3.13	0.38	2.30	2.90	3.15	3.40	3.90
Employment	62	3	Part-Time	43	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Salary	62.00	NaN	NaN	NaN	48.55	12.08	25.00	40.00	50.00	55.00	80.00
Social Networking	62.00	NaN	NaN	NaN	1.52	0.84	0.00	1.00	1.00	2.00	4.00
Satisfaction	62.00	NaN	NaN	NaN	3.74	1.21	1.00	3.00	4.00	4.00	6.00
Spending	62.00	NaN	NaN	NaN	482.02	221.95	100.00	312.50	500.00	600.00	1400.00
Computer	62	3	Laptop	55	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Text Messages	62.00	NaN	NaN	NaN	246.21	214.47	0.00	100.00	200.00	300.00	900.00

Table 2.3. Data Description

The above table helps to describe the data, from the table we can conclude the following points.

1. There are a total of 62 people participated in the survey.
2. Out of 62 female has the top frequency, which is 33.
3. The total number of males is 29.
4. There are a total of 8 different courses are there in the university.
5. The most popular course among people is Retailing/Marketing.
6. Out of the students who participated in the survey the minimum age is 18 and the maximum age is 26.
7. Out of the students who participated in the survey, their minimum GPA is 2.30 and their maximum GPA is 3.90. The group has an average GPA of 3.13
8. Most of the people work part-time.
9. Out of 62 students, 55 students use the laptop for study.

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	All
Female	3	3	7	4	4	3	9	0	33
Male	4	1	4	2	6	4	5	3	29
All	7	4	11	6	10	7	14	3	62

Table 2.1.1.Contingency table (Gender and Major)

2.1.2 Gender and Grad Intention

Grad Intention Gender	No	Undecided	Yes	All
Female	9	13	11	33
Male	3	9	17	29
All	12	22	28	62

Table 2.1.2.Contingency table (Gender and Grad intention)

2.1.3 Gender and Employment

Employment Gender	Full-Time	Part-Time	Unemployed	All
Female	3	24	6	33
Male	7	19	3	29
All	10	43	9	62

Table 2.1.3.Contingency table (Gender and Employment)

2.1.4 Gender and Computer

Computer Gender	Desktop	Laptop	Tablet	All
Female	2	29	2	33
Male	3	26	0	29
All	5	55	2	62

Table 2.1.4.Contingency table (Gender and Computer)

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 number of males = 29

Total number of people = 62

$$P(\text{Male}) = \frac{\text{Total number of male}}{\text{Total number of people}}$$

$$= \frac{29}{62} = 0.46774$$

Therefore the probability of randomly selected CMSU student will be male is 46.774%

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

Total number of females = 33

Total number of people = 62

$$P(\text{Female}) = \frac{\text{Total number of females}}{\text{Total number of people}}$$

$$= \frac{33}{62} = 0.53225$$

Therefore the probability of randomly selected CMSU student will be female is 53.225%

2.3 Assume that the sample is a 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

Consider the contingency table between gender and major,

Major / Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	All
Female	3	3	7	4	4	3	9	0	33
Male	4	1	4	2	6	4	5	3	29
All	7	4	11	6	10	7	14	3	62

Table 2.3.1.Contingency table (Gender and Major)

Also, consider the normalized contingency table between gender and major,

Major / Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Female	0.09	0.09	0.21	0.12	0.12	0.09	0.27	0.00
Male	0.14	0.03	0.14	0.07	0.21	0.14	0.17	0.10
All	0.11	0.06	0.18	0.10	0.16	0.11	0.23	0.05

Table 2.3.2.Normalized Contingency table (Gender and Major)

By analyzing the normalized contingency table between the gender and major,

- P (Male / Accounting major) = 0.14 = 14%
- P (Male / CIS major) = 0.03 = 3%
- P (Male / Economics or Finance major) = 0.14 = 14%
- P (Male / International business major) = 0.07 = 7%
- P (Male / Management major) = 0.21 = 21%
- P (Male / Other major) = 0.14 = 14%
- P (Male / Retailing or market major) = 0.17 = 17%
- P (Male / Undecided) = 0.10 = 10%

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

Major Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Female	0.09	0.09	0.21	0.12	0.12	0.09	0.27	0.00
Male	0.14	0.03	0.14	0.07	0.21	0.14	0.17	0.10
All	0.11	0.06	0.18	0.10	0.16	0.11	0.23	0.05

Table 2.3.3.Normalized Contingency table (Gender and Major)

Consider the normalized contingency table between the gender and major, by analyzing the table we can conclude the following conditional probabilities,

- P (Female / Accounting major) = 0.09 = 9%
- P (Female / CIS major) = 0.09 = 9%
- P (Female / Economics or Finance major) = 0.21 = 21%
- P (Female / International business major) = 0.12 = 12%
- P (Female / Management major) = 0.12 = 12%
- P (Female / Other major) = 0.09 = 9%
- P (Female / Retailing or market major) = 0.27 = 27%
- P (Female / Undecided) = 0.00 = 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.

Let us consider the contingency table gender and grad intention,

Grad Intention	No	Undecided	Yes	All
Gender				
Female	9	13	11	33
Male	3	9	17	29
All	12	22	28	62

Table 2.4.1. Contingency table (Gender and Grad intention)

Consider the below shown normalized table of gender and grad intention too,

Grad Intention	No	Undecided	Yes
Gender			
Female	0.27	0.39	0.33
Male	0.10	0.31	0.59
All	0.19	0.35	0.45

Table 2.4.2. Normalized Contingency table (Gender and Grad intention)

From the contingency table the probability of randomly chosen student is male and intent to graduate is,

$$P(\text{Male/intent to graduate}) = \frac{17}{29} = 0.59$$
$$= 59\%$$

Therefore the randomly chosen student is male and intent to graduate is 59%

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

Consider the contingency table between the gender and the computer,

Computer / Gender	Desktop	Laptop	Tablet	All
Female	2	29	2	33
Male	3	26	0	29
All	5	55	2	62

Table 2.4.3. Contingency table (Gender and Computer)

Consider the below normalized contingency table of gender and computer,

Computer / Gender	Desktop	Laptop	Tablet
Female	0.06	0.88	0.06
Male	0.10	0.90	0.00
All	0.08	0.89	0.03

Table 2.4.4. Normalized Contingency table (Gender and Computer)

By analyzing the both the table we can conclude that the probability of randomly selected student is female and does not have a laptop is,

$$P(\text{Female/Does not have laptop}) = 1 - \frac{29}{33}$$
$$= 1 - 0.8787 = 1 - 0.88$$
$$= 0.12$$

Therefore the randomly chosen student is female and does not have a laptop is 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.

Probability of randomly chosen student is male is,

$$P(M) = \frac{29}{62} = 0.4677$$

Refer the contingency table 2.1.3,

The probability of randomly chosen student having a job is,

$$P(J) = \frac{10}{62} = 0.1612$$

The probability of randomly chosen student is male and full time employment is,

$$P(M \cap J) = \frac{7}{62} = 0.1129$$

Here probability addition rule with mutually intersect event will apply, and the equation is,

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

$$P(M \cup J) = P(M) + P(J) - P(M \cap J)$$

$$= 0.4677 + 0.1612 - 0.1129$$

$$= 0.516$$

Therefore the probability of a randomly chosen student is male or has a full time employment is 51.6%

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

Major Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	All
Female	3	3	7	4	4	3	9	0	33
Male	4	1	4	2	6	4	5	3	29
All	7	4	11	6	10	7	14	3	62

Table 2.5.1. Contingency table (Gender and Major)

Consider the contingency table of gender and major, as per the contingency table, Probability of a female student is majoring in international business is,

$$P(I.B) = \frac{4}{33} = 0.1212$$

$$P(M) = \frac{4}{33} = 0.1212$$

Since the both events are independent to each other, as per probability addition rule equation is,

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

$$P(I.B \cup M) = P(I.B) + P(M)$$

$$= 0.1212 + 0.1212$$

$$= 0.2424$$

Therefore the conditional probability that given a female student is randomly chosen, 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?

Now we have to consider the gender and intent to graduate contingency table once again, this time as per condition we have to remove the rows of students who are still not decided. After removing those rows the sample updated dataset is shown below.

	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
8	9	Female	20	Junior	Management	Yes	3.60	Unemployed	30.00	0	4	500	Laptop	400

Table 2.6.1. Updated Dataset

Now let us construct the 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, and only 40 entries are present now.

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

Grad Intention Gender	No	Yes
Female	0.45	0.55
Male	0.15	0.85
All	0.30	0.70

Table 2.6.2. Contingency Table

Let us check the events are independent or not, the condition for two events are independent is if and only if

$$P(A|B) = P(A)$$

Probability of randomly selected student female is,

$$P(F) = \frac{20}{40} = 0.5$$

Probability of randomly selected student is female and intent to graduate is,

$$P(F|I.G) = \frac{11}{20} = 0.55$$

As per the output we received, we can conclude that,

$$P(F|I.G) \neq P(F)$$

Therefore the events 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?

Number of students whose GPA less than 3 is = 17

Total number of students = 62

$$P(GPA<3) = \frac{17}{62} = 0.2741$$

Therefore if a student is chosen randomly, the probability of his/her GPA is less than 3 is 27.41%

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.

i) Conditional probability that a randomly selected male earns 50 or more.

Total number of male earns less than 50 = 14

Total number of male = 29

$$P(\text{Male earns less than 50}) = \frac{14}{29} = 0.4827$$

Therefore the conditional probability that a randomly selected male earns 50 or more is 48.27%

ii) Conditional probability that a randomly selected female earns 50 or more.

Total number of female earns less than 50 = 18

Total number of male = 33

$$P(\text{Male earns less than 50}) = \frac{18}{33} = 0.5454$$

Therefore the conditional probability that a randomly selected female earns 50 or more is 54.54%

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 finding whether the data is normally distributed or not we can use distplot method and Shapiro test in python, if the value of p in Shapiro test is greater than 0.05 then we can conclude that the given data is normally distributed. Below shown figure shows the output from python using distplot,

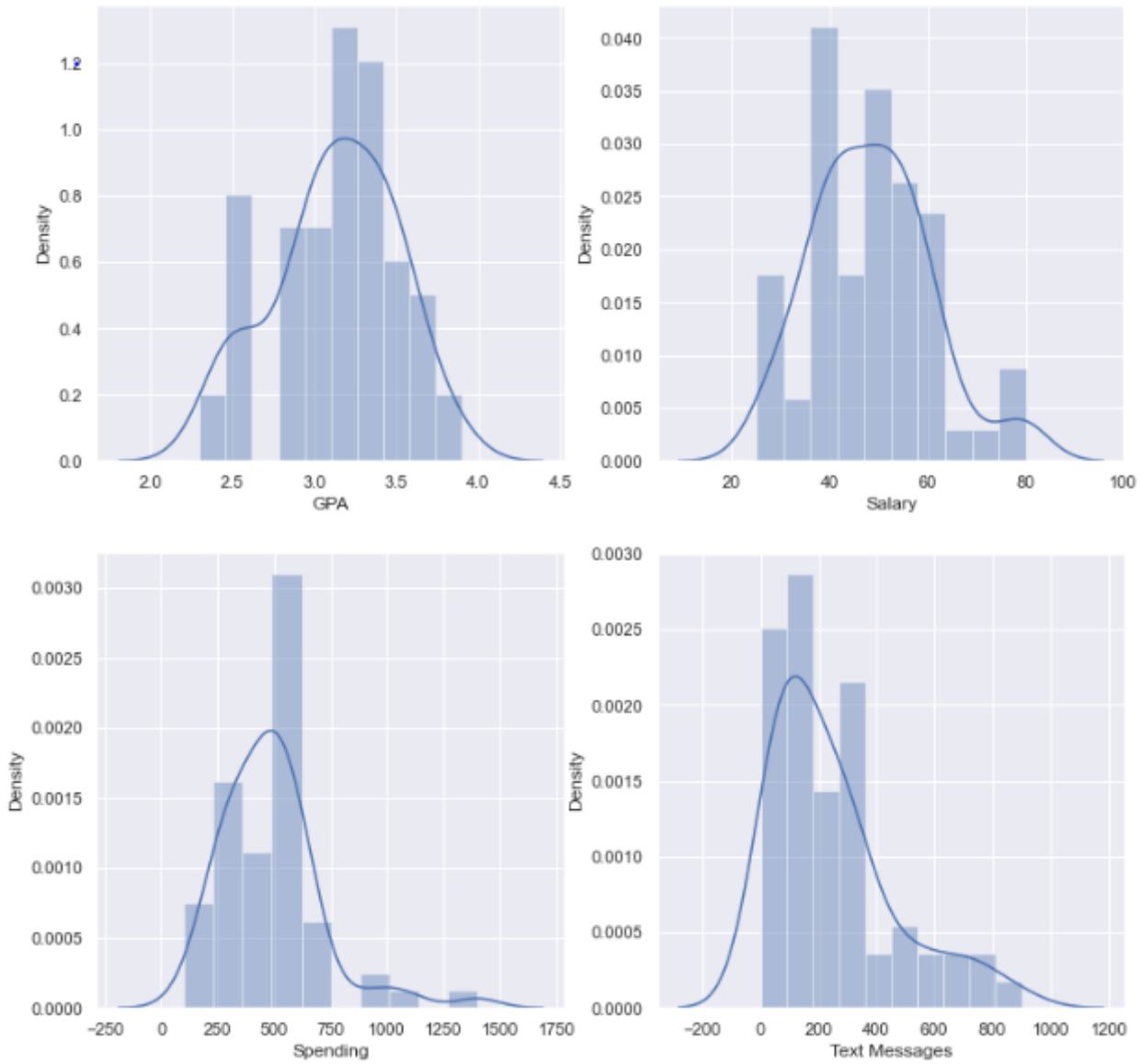

Fig. 2.8.1 – Distplot

Skew value of GPA, Salary, Spending, Text message is shown below,

- Skew value of GPA = -0.3146000894506981
- Skew value of salary = 0.5347008436225946
- Skew value of spending = 1.5859147414045331
- Skew value of text message = 1.2958079731054333

Shapiro test output from python is shown below,

- Shapiro result of GPA is Shapiro Result(statistic=0.9685361981391907, pvalue=0.11204058676958084)

- Shapiro result of Salary is Shapiro Result(statistic=0.9565856456756592, pvalue =0.028000956401228905)
- Shapiro result of Spending is Shapiro Result(statistic=0.8777452111244202, pvalue=1.6854661225806922e-05)
- Shapiro result of Text Messages is Shapiro Result(statistic=0.8594191074371338, pvalue=4.324040673964191e-06)

By comparing the distplot, skew value and Shapiro test results we can conclude that except GPA all others are right skewed and GPA is left skewed, and only the p value of GPA is greater than 0.05, GPA has a p value of 0.1120, therefore only the data GPA is normally distributed as per Shapiro test. All other data, salary, spending, text message did not follow a normal distribution.

CONCLUSION

The Student News Service at Clear Mountain State University (CMSU) has conducted a survey among 62 students and out of them 29 students are male and 33 students are female. Out of that most of the students are doing major in retailing and marketing and we can assume that it is the most popular course among the students. Most of the people is intent to graduate and most of them are working part time and most of them uses laptop for study. And the group has a mean salary near to 50.

PROBLEM 3 – SUMMARY

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. This problem deals with different hypothesis testing based on the different conditions.

INTRODUCTION

The purpose of this exercise is to explore the dataset and do exploratory data analysis, do the hypothesis testing based on the different assumptions we have to prove. The data includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.

DATA DESCRIPTION

The data includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.

SAMPLE OF THE DATASET

	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

Table 3.1. Shingles Sample Dataset

The data includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.

EXPLORATORY DATA ANALYSIS

NO.	Shingles	Data type	Null object
1	A	Float	36 non null
2	B	Float	31 non null

Table 3.2. Exploratory data analysis

There are a total of two columns and 36 rows, there are total of 5 null values present in column B

DESCRIPTIVE DATA ANALYSIS

Let us check the descriptive analysis of the data

	count	mean	std	min	25%	50%	75%	max
A	36.0	0.316667	0.135731	0.13	0.2075	0.29	0.3925	0.72
B	31.0	0.273548	0.137296	0.10	0.1600	0.23	0.4000	0.58

Table 3.3. Descriptive data analysis

- There are total 36 entries in singles A and 31 entries in shingles B
- The mean value of shingles A is greater than shingles B

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

3.1Do you think there is evidence that mean moisture contents in both types of shingles are within the permissible limits? State your conclusions clearly showing all steps.

We have to find whether the mean moisture content in both type of shingles are within the permissible limit, so we have to conduct hypothesis test in both sample A and B separately

Let us check hypothesis testing for Sample A

Step – 1: Define null and alternate hypothesis

i) Null hypothesis (H_0) = mean moisture content ≤ 0.35

ii) Alternate hypothesis (H_1) = mean moisture content > 0.35

Step – 2: Decide the significance level

Since the α value is not given we assume $\alpha = 0.05$

Step – 3: Identify the test statistic

Since we don't know the population standard deviation and the sample size (n) of sample A is less ie, 36 so we can proceed with 1 sample t-test.

Step – 4: Calculate the statistic and p-value

From python we will get the values of t-stat and p-value

t-stat = -1.4735046253382782

p-value = 0.07477633144907513

Step – 5: Decide to reject or accept the null hypothesis

Since the p-value is greater than the significance level (α) we will accept the null hypothesis.

Therefore we conclude that the mean moisture content is shingle A is less than or equal to 0.35

<u>Let us check hypothesis testing for Sample B</u>

Step – 1: Define null and alternate hypothesis

i) Null hypothesis (H_0) = mean moisture content ≤ 0.35

ii) Alternate hypothesis (H_1) = mean moisture content > 0.35

Step – 2: Decide the significance level

Since the α value is not given we assume α = 0.05

Step – 3: Identify the test statistic

Since we don't know the population standard deviation and the sample size (n) of sample A is less ie, 31 so we can proceed with 1 sample t-test.

Step – 4: Calculate the statistic and p-value

From python we will get the values of t-stat and p-value

t-stat = -3.1003313069986995

p-value = 0.0020904774003191826

Step – 5: Decide to reject or accept the null hypothesis

Since the p-value is less than the significance level (α) we will reject the null hypothesis.

Therefore we conclude that the mean moisture content is shingle B is greater than 0.35

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

Since both samples A and B are independent we have to conduct a independent t test here.

Step – 1: Define null and alternate hypothesis

Since we have to check whether the population means of singles A and B are equal, therefore

i) Null hypothesis (H_0): μ (Shingle A) = μ (Shingle B)

ii) Alternate hypothesis (H_1): μ (Shingle A) ≠ μ (Shingle B)

Step – 2: Decide the significance level

Since the α value is not given we assume α = 0.05

Step – 3: Identify the test statistic

We have two samples here and we do not know the standard deviation and sample size is very less so let us follow independent two sample t-test. Therefore it is a two tail test

Step – 4: Calculate the statistic and p-value

From python we will get the values of t-stat and p-value

t-stat = 1.289628271966112

p-value = 0.2017496571835328

Step – 5: Decide to reject or accept the null hypothesis

Since the p-value is greater than the significance level (α) we will accept the null hypothesis.

Therefore we conclude that the population means of singles A and B are equal.

The assumption we need to check we perform before the test of equality of mean is performed is we have to check the sample size first, the minimum sample size is 30. Then we have to assume that the data is normally distributed for confirming that we have to perform Shapiro test. If the p-value in Shapiro test is greater than 0.05 we can conclude that the data is normally distributed.

THE END

PLEASE CLICK HERE TO GO TO CONTENTS