SMDM Project (DSBA)-June-12, 2022.

Submitted by,

Deepa. K

Table of Contents

List of Tables:

List of Figures:

Introduction:

The business report explains three different problems based on descriptive statistics, contingency table and probability. The exploration of data set using various attributes helps in analyzing the required information provided.

Problem Statement 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 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:

- Buyer/Spender: Number of person
- Channel: Hotel/Retail
- Region: Lisbon/Oporto/Other
- Fresh, Milk, Grocery, Frozen, Detergents, Delicatessen: Annual spending amount

Sample dataset:

Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatesen
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	Retail	Other	13265	1196	4221	6404	507	1788
5	Retail	Other	22615	5410	7198	3915	1777	5185

Table 2: Dataset Sample

Interpretation of Sample Dataset:

The sample dataset consists of 9 columns with 6 different items being spent by people of Portugal in several stores operating in Lisbon, Oporto and other region. The information displayed here is the annual expenditure of people across two sales channels Retail and Hotel.

Exploratory Data Analysis:

- Data Type Check:

Type of variables	Data Type
Buyer/Spender	int64
Channel	object
Region	object
Fresh	int64
Milk	int64
Grocery	int64
Frozen	int64
Detergents_ Paper	int64
Delicatessen	int64

Table 3: dtype Check

- Inference:

  From the above table, it is clear that there are 7 integer and 2 object data types available in the given dataset. The whole dataset contains 440 rows and 9 columns with no non null values (all values are present).

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?

1.1.1: Use methods of descriptive statistics to summarize data

The most recognized type of descriptive statistics to summarize data can be achieved by the Mean, Median and Mode. The below table explains the number of numerical values and categorical variables present across each column. Nan stands for not a number.

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_paper	Delicatessen
Count	440	440	440	440	440	440	440	440	440
unique	Nan	2	3	Nan	Nan	Nan	Nan	Nan	Nan
top	Nan	Hotel	Other	Nan	Nan	Nan	Nan	Nan	Nan
freq	Nan	298	316	Nan	Nan	Nan	Nan	Nan	Nan
mean	220.5	Nan	Nan	12000.29	5796.26	7951.27	3071.93	2881.49	1524.87
std	127.16	Nan	Nan	12647.32	7380.37	9503.16	4854.67	4797.85	2820.1
min	1.00	Nan	Nan	3.000	55.00	3.00	25.0	3.0	3.0
25%	110.75	Nan	Nan	3127.75	1533.00	2153.00	742.0	256.75	408.25
50%	220.50	Nan	Nan	8504.00	3627.00	4755.5	1526.00	816.5	965.5
75%	330.250	Nan	Nan	16933.75	7190.250	10655.75	3554.25	3922.0	1820.25
max	440.000	Nan	Nan	112151.00	73948.00	92780.00	6086.00	40827.0	47943.0

Table 4: Dataset Summary

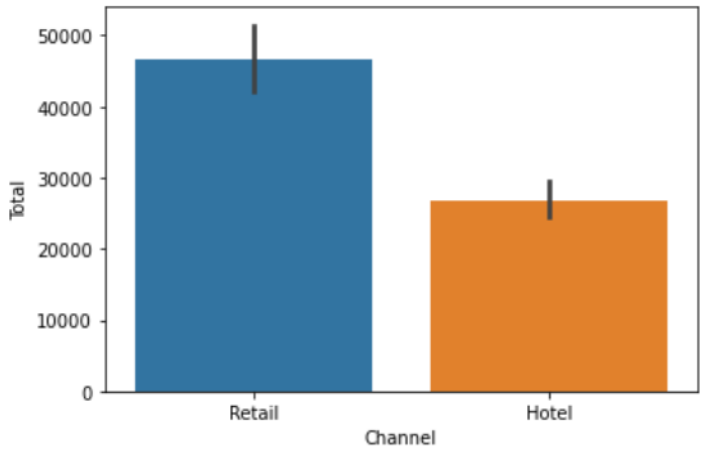

Figure 4: Channel Vs Total

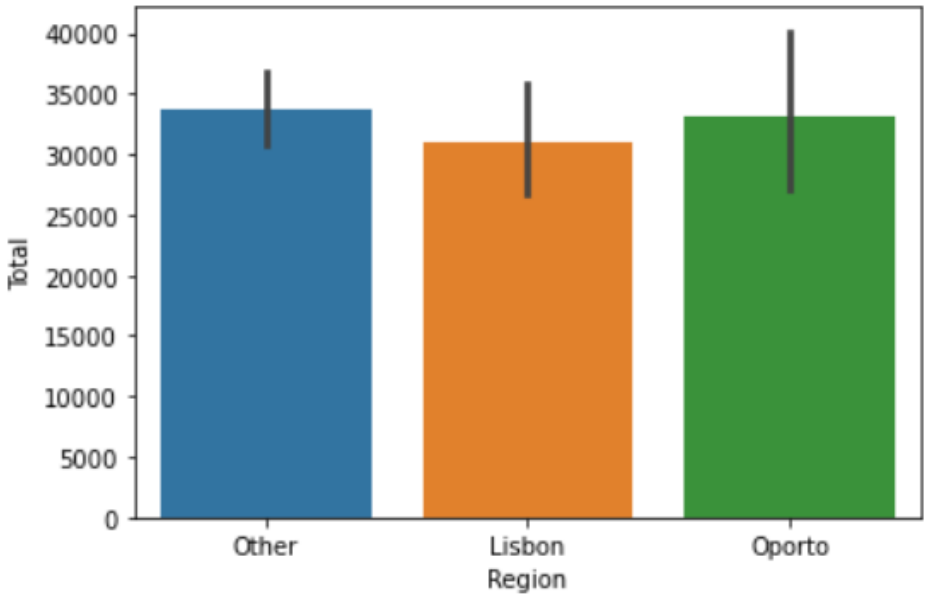

Figure 5: Region Vs Total

1.1.2 Which Region and which Channel spent the most?

The most annual spent was done in Other Region under Hotel Channel.

1.1.3 Which Region and which Channel spent the least?

The least annual spent was done in Oporto Region under Retail Channel.

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

Solution:

The annual spend on six different varieties of items are analyzed using bar plot among the regions and channels.

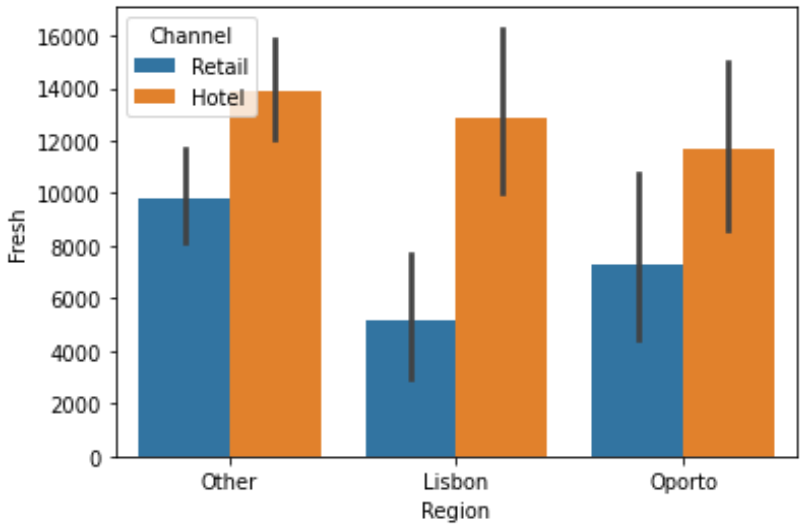

Figure 6: Region Vs Fresh bar plot

- The annual spend on Fresh is higher in other region under Hotel channel.

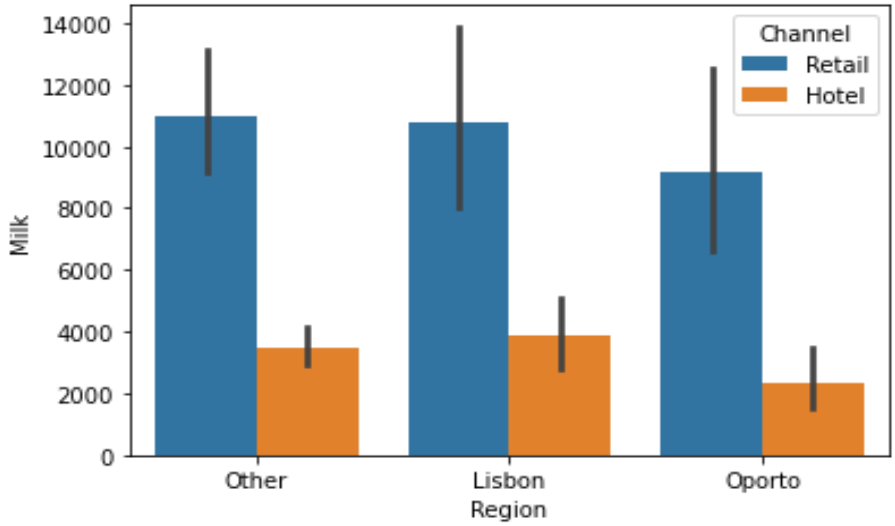

Figure 7: Region Vs Milk bar plot

- The annual spend on Milk is higher in Lisbon region under Retail channel and lesser in Oporto region under Hotel channel.

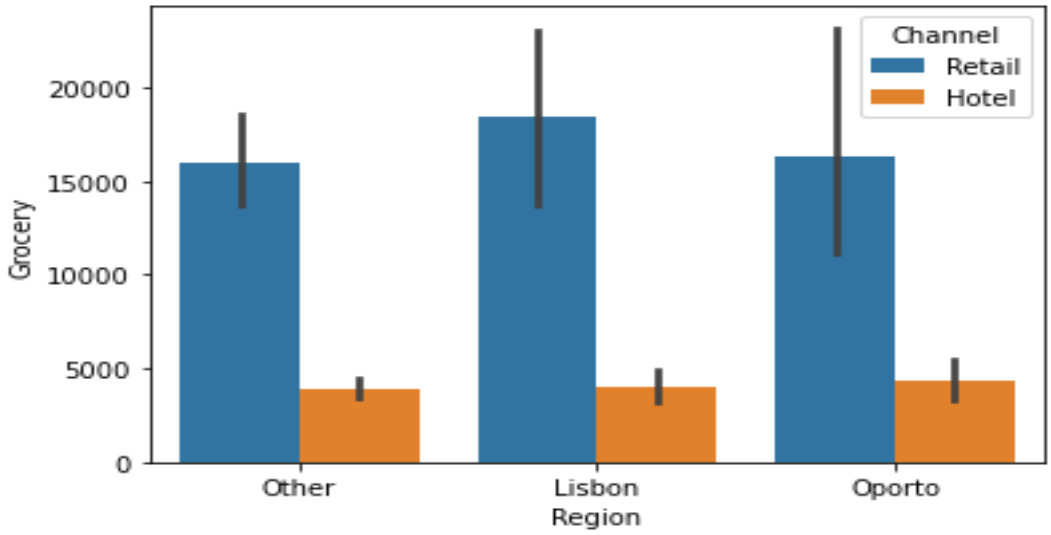

Figure 8: Region Vs Grocery bar plot

- The annual spend on Grocery is higher in Lisbon region under Retail channel and lesser in other region under Hotel channel.

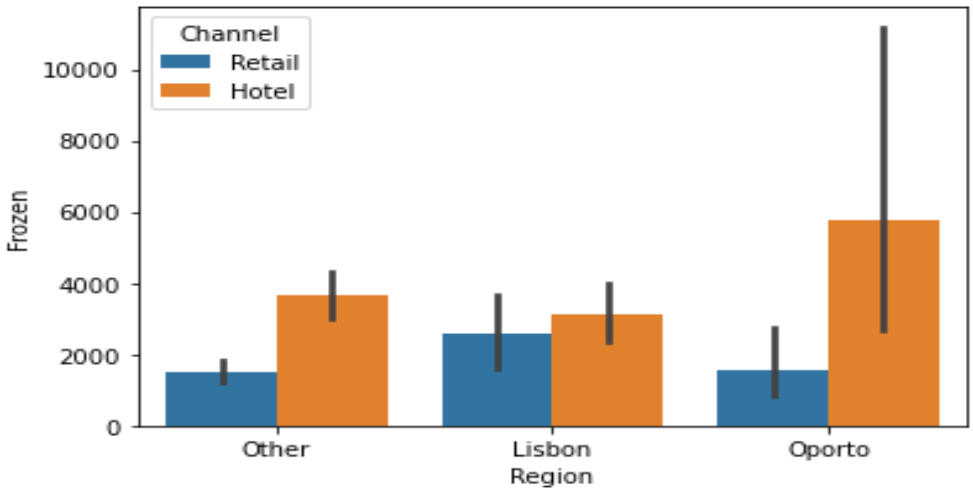

Figure 9: Region Vs Frozen bar plot

- The annual spend on Frozen is higher in Oporto region under Hotel channel and lesser in other region under Retail channel.

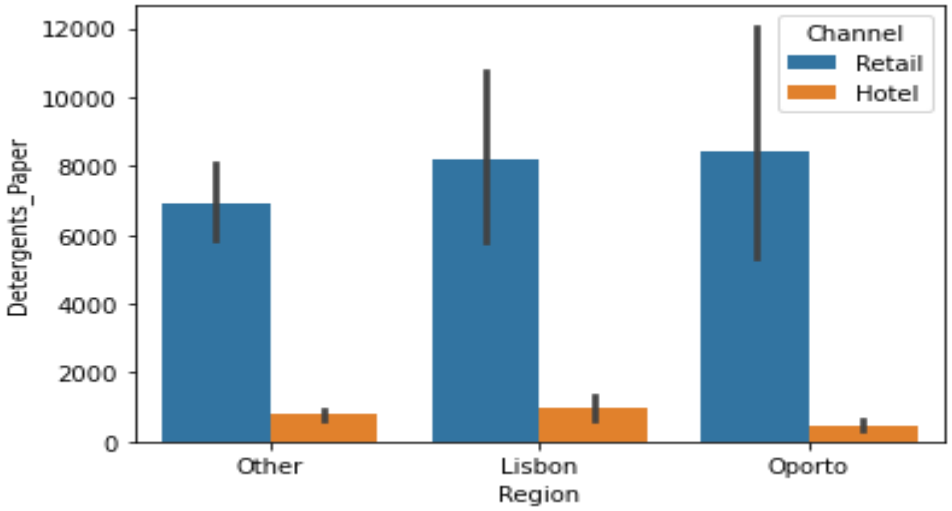

Figure 10: Region Vs Detergents_Paper bar plot

- The annual spend on Detergents_Paper is higher in Oporto region under Retail channel and lesser under Retail channel.

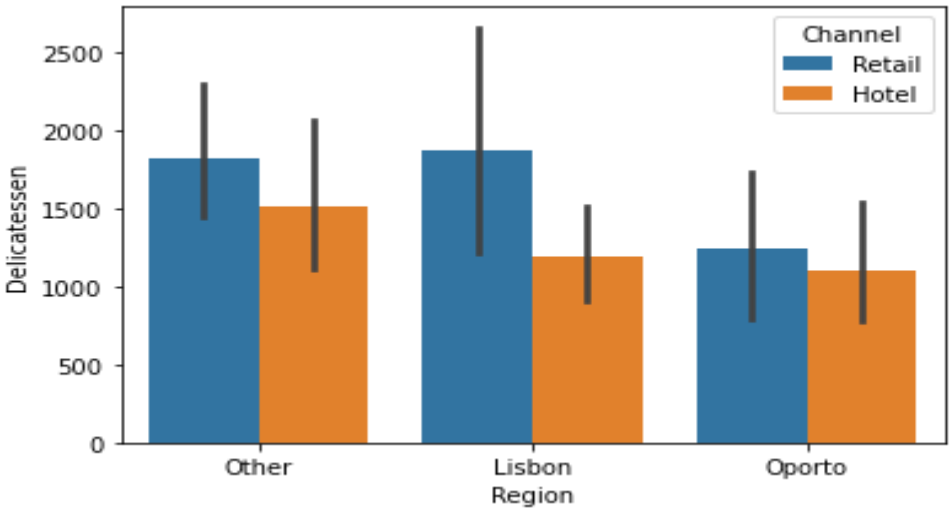

Figure 11: Region Vs Delicatessen bar plot

- The annual spend on Delicatessen is higher in Lisbon region under Retail channel and lesser in Oporto region under Hotel channel.

Justification:

The various items present in the dataset show the range of annual expenditure across each region and channel. The Lisbon and Oporto are the most and least spent regions from the given dataset.

1.3: On the basis of the descriptive measure of variability, which item shows the most inconsistent behaviour? Which items shows the least inconsistent behaviour?

Solution:

The descriptive measure of variability can be analyzed through Coefficient of variation.

Coefficient of variation is the ratio of standard deviation (σ) to mean (μ).

CV= σ/ μ for population data and S/$\bar{x}$ for sample data.

Fresh	12647.328865
Milk	7380.377175
Grocery	9503.162829
Frozen	4854.673333
Detergents_Paper	4767.854448
Delicatessen	2820.105937

Table 5: Standard Deviation (σ) values

Fresh	12000.297727
Milk	5796.265909
Grocery	7951.277273
Frozen	3071.931818
Detergents_Paper	2881.493182
Delicatessen	1524.870455

Table 6: Mean (μ) values

Fresh	1.053918
Milk	1.273299
Grocery	1.195174
Frozen	1.580332
Detergents_Paper	1.654647
Delicatessen	1.849407

Table 7: CV Values

From the above calculation,

- Delicatessen shows most inconsistent behavior among the other items.
- Fresh shows least inconsistent behavior among the other items.

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

Solution:

Box plots can be used to identify the outliers present in the dataset. The following are the box plots for 6 different items present in dataset.

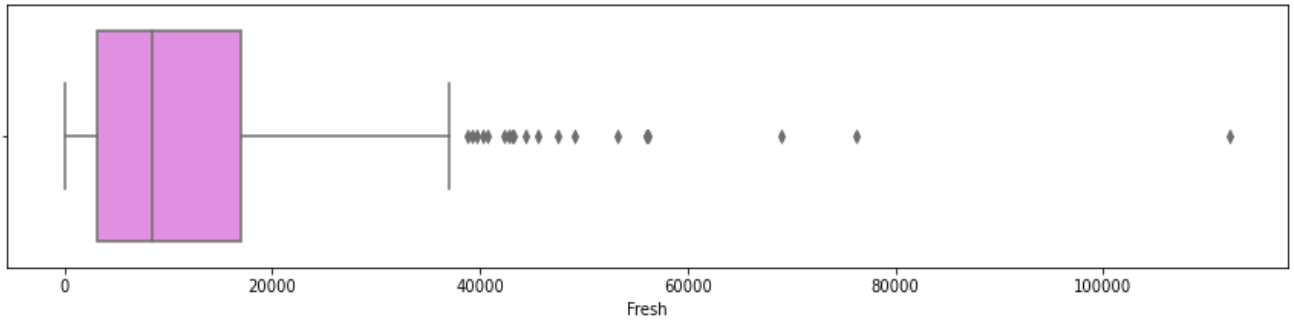

Figure 12 : Box plot for Fresh

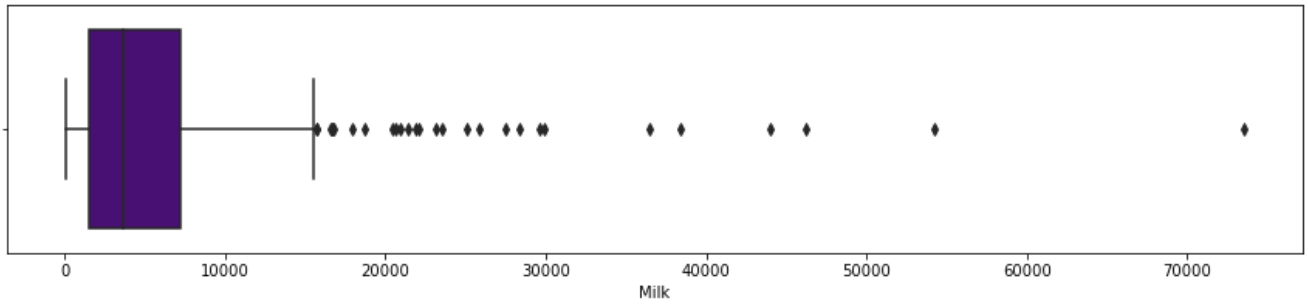

Figure 13: Box plot for Milk

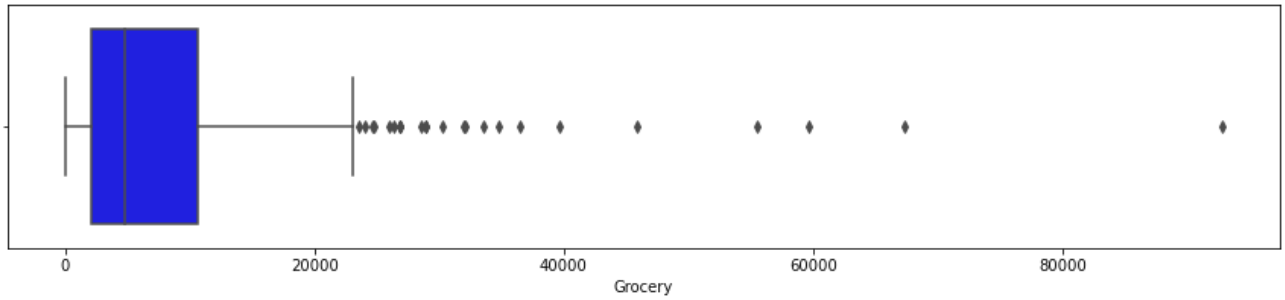

Figure 14: Box plot for Grocery

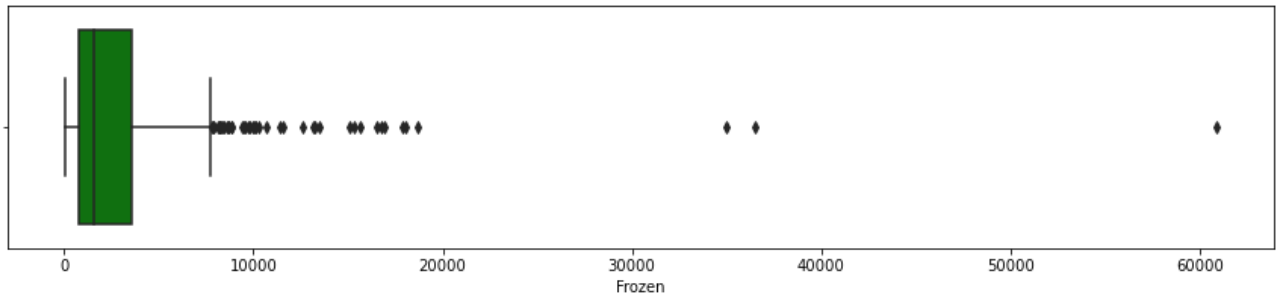

Figure 15 : Box plot for Frozen

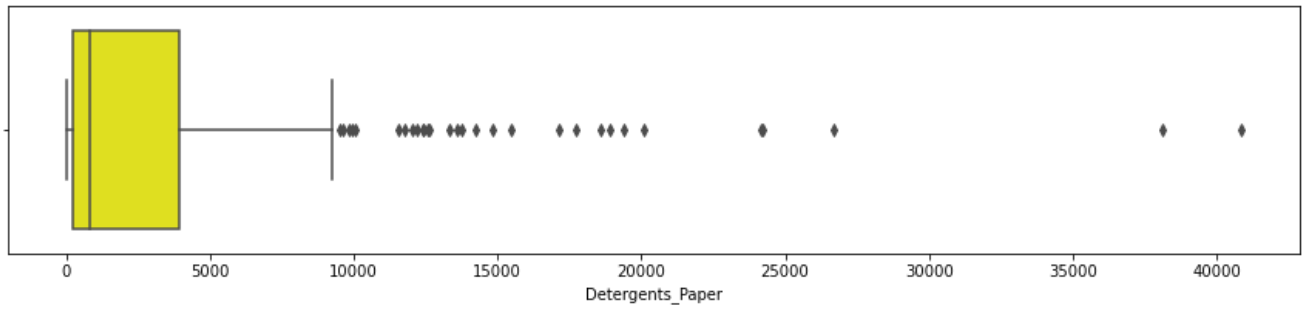

Figure 16: Box plot for Detergents_Paper

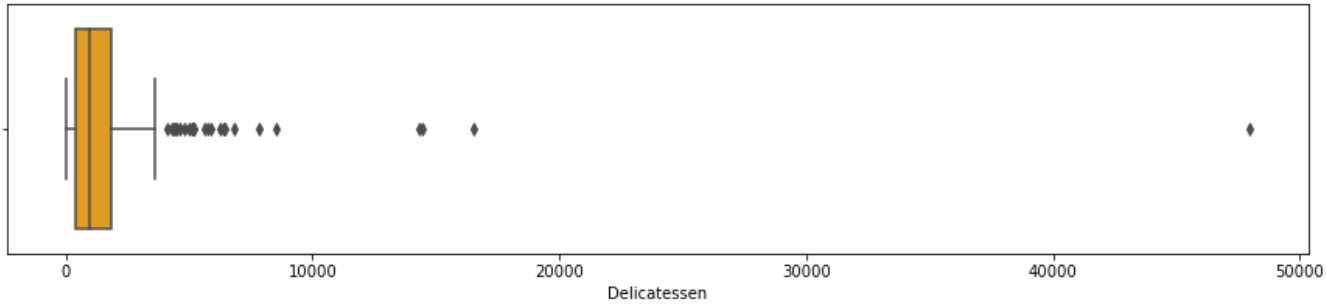

Figure 17: Box plot for Delicatessen

Comments:

From the above bar plots,

- Outliers/extreme values are present in all the items of the data set.

1. 5: On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem? Answer from the business perspective.

Solution:

- The region that spent most is other region under Hotel channel.
- The region that spent least is Oporto region under retail channel.
- There are outliers present in the 6 items mentioned in data set.

Problem Statement 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.

Data Description:

- Gender : Male/Female
- Age : in years
- Class :Junior/Senior
- Major : Accounting/CIS/Management/Other etc
- Grad Intention: Decided/undecided
- GPA: Score out of 10
- Employment :Full time/part time
- Salary :earned by students
- Social networking: Availability in media
- Satisfaction: rating out of 5
- Computer: Laptop/desktop

Exploratory Data Analysis:

- Data type Check:

Type of variables	Data Type
ID	int64
Gender	object
Age	int64
Class	object
Major	object
Grad Intention	object
GPA	float64
Employment	object
Salary	float64
Social Networking	int64

Satisfaction	int64
Spending	int64
Computer	object
Text messages	int64

Table 8: dtype Check

- Inference:

  There are totally 62 non null values present out of which 2 float (decimal value), 6 integer,6 object data types.

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

Figure 18: Gender & 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

Figure 19: Gender & 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

Figure 20: Gender & 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

Figure 21: Gender & Computer

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

Solution:

Probability = Number of Male students/Total population at CMSU

P (Male) = 29/62 = 0.4677

47% randomly selected student at CMSU will be male.

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

Solution:

Probability = Number of Female students/Total population at CMSU

P (Female) = 33/62=0.5322

53% randomly selected student at CMSU will be female.

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

Solution:

The below table shows the number of male candidates in each major.

Major	Male count
Management	6
Retailing/Marketing	5
Other	4
Economics/Finance	4
Accounting	4
Undecided	3
International Business	2
CIS	1

Table 9: Major count-M

The conditional probability of a male student choosing a Major is:
P (Male Major) = Number of Male count in each major /Total Male count

Major	Male count
Management	6/29 = 0.206897=21%
Retailing/Marketing	5/29 = 0.172414=17%
Other	4/29 = 0.137931=14%
Economics/Finance	4/29 = 0.137931=14%

Accounting	4/29 = 0.137931=14%
Undecided	3/29 = 0.103448=10%
International Business	2/29 = 0.068966=7%
CIS	1/29 = 0.034483=3%

Table 10: P (Male Major)

2.3.2: Find the conditional probability of different majors among the Female students in CMSU.

Solution:

The below table shows the number of female candidates in each major.

Major	Female count
Retailing/Marketing	9
Economics/Finance	7
Management	4
International Business	4
Other	3
CIS	3
Accounting	3

Table 11: Major count-F

The conditional probability of a male student choosing a Major is:
P (Female Major) = Number of female count in each major /Total female count

Major	Female count
Retailing/Marketing	9/33 = 0.272727=27%
Economics/Finance	7/33 = 0.212121=21%
Management	4/33 = 0.121212=12%
International Business	4/33 = 0.121212=12%
Other	3/33 = 0.090909=9%
CIS	3/33 = 0.090909=9%
Accounting	3/33 = 0.090909=9%

Table 12: P (female Major)

2.4.1: Find the probability that a randomly chosen student is a male and intends to graduate.

Solution:

The probability of getting a male student and intends to graduate can be defined by the number of male students who have intended "Yes" for graduation (17) divided by total students (62).

- P(M|G)=17/62=0.2741=27%

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

Solution:

The probability of getting a female student who doesn't have laptop can be defined by the number of female students who doesn't have laptop 4(2 tablet, 2 desktop) divided by total students (62).

- P(F|T) = 2/62 = 0.032258 (Female with Tablet)
- P(F|D) = 2/62 = 0.032258 (Female with Desktop)

Total female students who don't have laptop is 6%

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

Solution:

The probability of getting a male student or has full time employment can be,

- P(Male)= 29/62
- P(Full-time employment)=10/62
- P(Male ∩ Full-time employment)=7/62

P (Male ∪ Full-time employment) = P(Male)+ P(Full-time employment)- P(Male ∩ Full-time employment)

P (Male ∪ Full-time employment) = 29/62+10/62-7/62 = 32/62 = 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.

Solution:

The Conditional probability of getting a female student who is majoring in International business or management can be proved by,

- **P(IB or Management | Female) =** 4/33 = 0.1212

The given probability is ***mutually exclusive*** because if a female student chooses International business she cannot choose Management studies and vice versa**.**

- *P (IB U M)=P(IB)+P(M)=0.1212+0.1212=0.2424=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 the graduate intention and being female are independent events?**

Solution:

Grad Intention Gender	No	Yes
Female	9	11
Male	3	17

Figure 22: 2x2 contingency

Sample size = 62(Total students)-22 (students under undecided category)

Sample size = 40.

- P (Grad .Intention) = 28/40 = 0.7 (Students who opted yes)
- P (Grad ∩ Female) = 11/20 = 0.55 (Female students who opted yes)

The Graduate intention and being female are not independent events as they are not equal to each other.

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

Solution:

GPA	2.3	2.4	2.5	2.6	2.8	2.9	3.0	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	All
Gender																	
Female	1	1	2	0	1	3	5	2	4	3	2	4	1	2	1	1	33
Male	0	0	4	2	2	1	2	5	2	2	5	2	2	0	0	0	29
All	1	1	6	2	3	4	7	7	6	5	7	6	3	2	1	1	62

Figure 23: GPA<3

Total students = 62

- The probability of male students < 3 GPA: 9/62=0.1451=14.5%
- The probability of female students < 3 GPA: 8/62=0.1290=12.9%

The probability that his/her GPA<3 is 27.4%

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:

There are 14 male students who earns 50 or more and 18 female students who earns 50 or more.

Salary	25.0	30.0	35.0	37.0	37.5	40.0	42.0	45.0	47.0	47.5	50.0	52.0	54.0	55.0	60.0	65.0	70.0	78.0	80.0	All
Gender																				
Female	0	5	1	0	1	5	1	1	0	1	5	0	0	5	5	0	1	1	1	33
Male	1	0	1	1	0	7	0	4	1	0	4	1	1	3	3	1	0	0	1	29
All	1	5	2	1	1	12	1	5	1	1	9	1	1	8	8	1	1	1	2	62

Figure 24: Salary>=50

- The probability of male earns >=50: 14/29 = 48.27%
- The probability of female earns >=50: 18/33 = 54.54%

2.8.1: 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.

Solution:

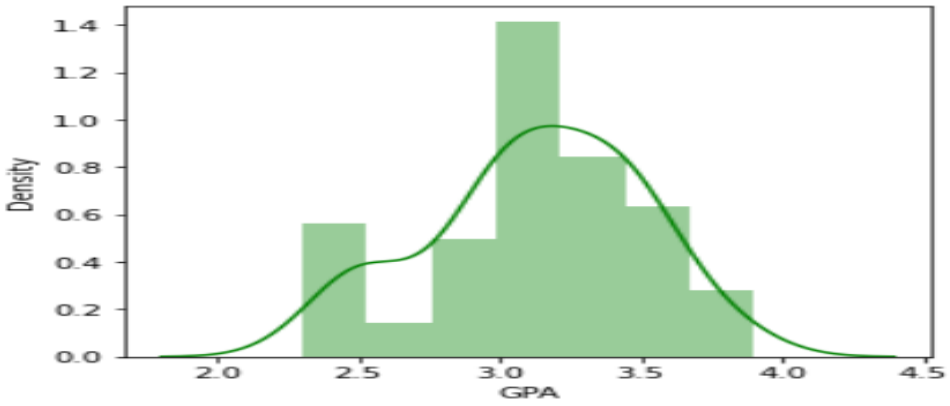

Figure 25: GPA Curve

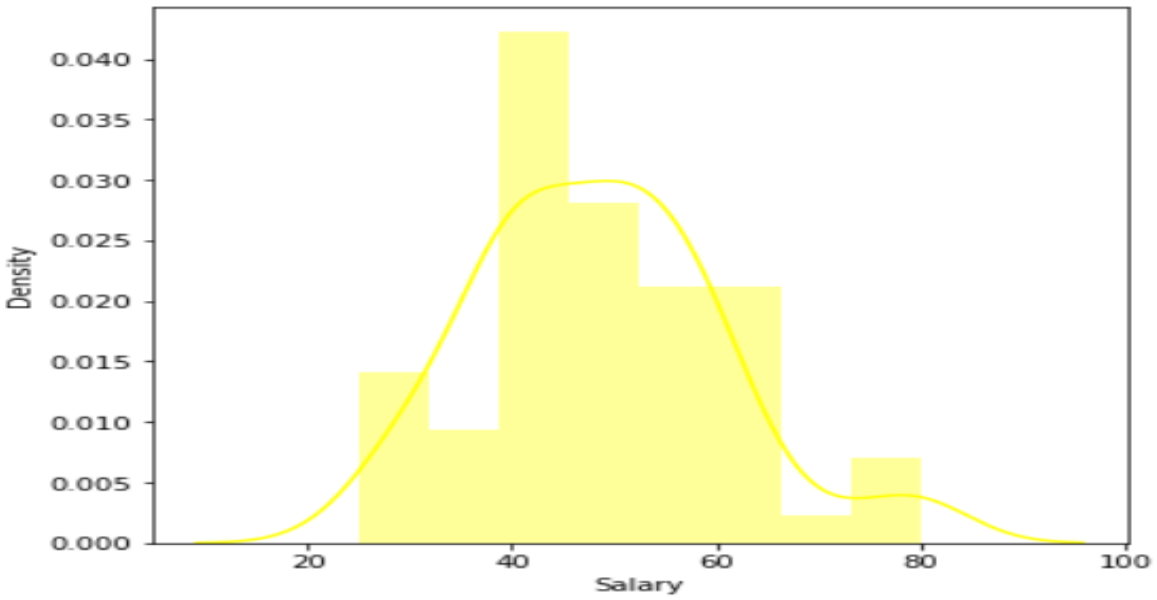

Figure 26: Salary curve

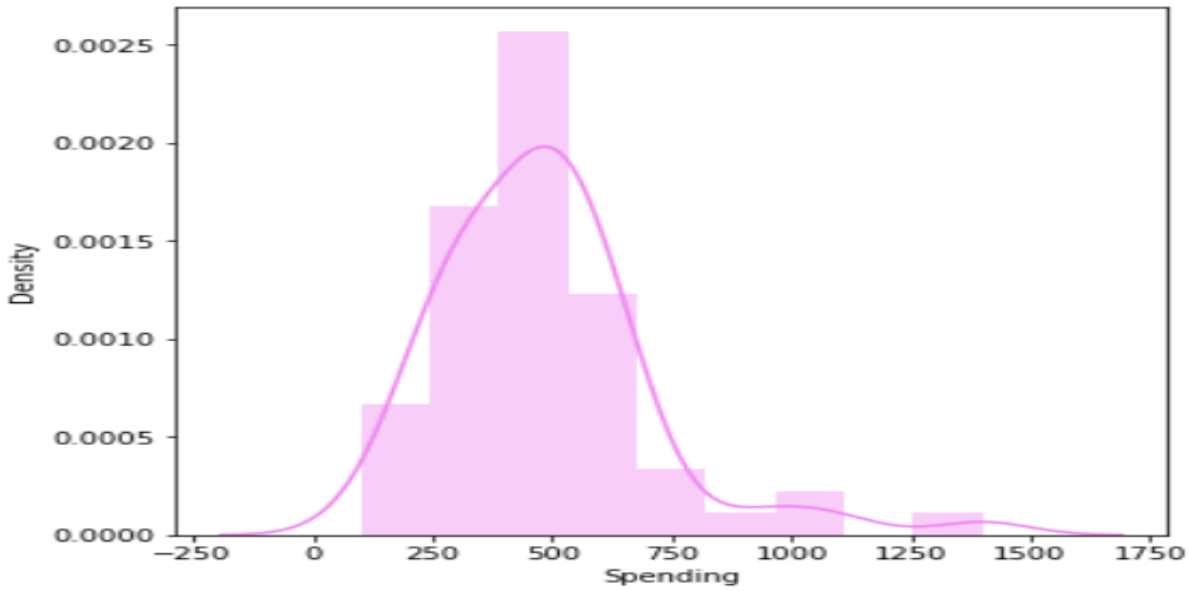

Figure 27: Spending curve

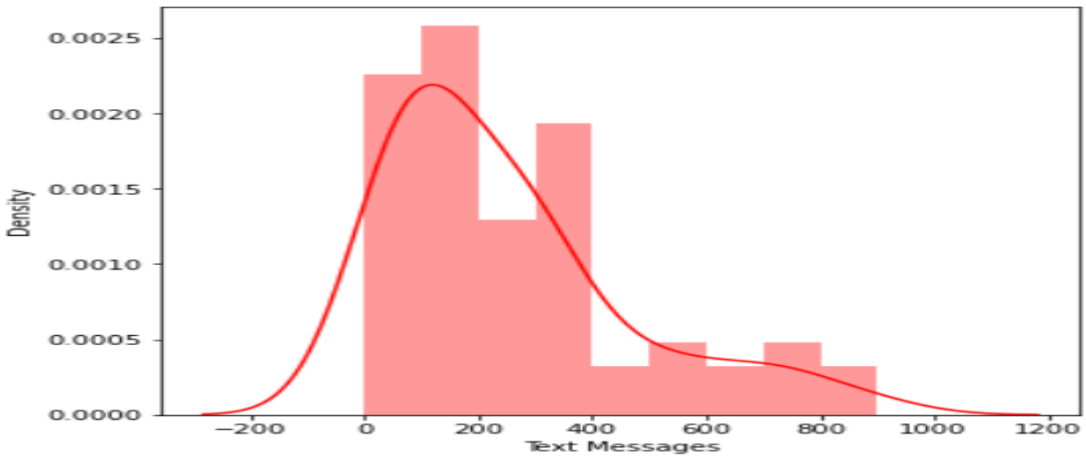

Figure 28: Text message curve

Comments:

All the numerical continuous variables are normally distributed (bell-shaped) as shown above.

Empirical rule for GPA:

- μ-σ = 3.12-0.37=2.75
- μ+σ=3.12+0.37=3.49

68% of students fall between 2.75 and 3.49 GPA.

- μ-2σ = 3.12-2*0.37=2.38
- μ+2σ = 3.12+2*0.37=3.86

95% of students fall between 2.38 and 3.86 GPA.

- μ-3σ = 3.12-3*0.37=2.01
- μ+3σ = 3.12+3*0.37=4.23

99.7% of students fall between 2.01 and 4.23 GPA.

Empirical rule for Salary:

- μ-σ=48.54-12.08=36.46
- μ+σ=48.54+12.08=60.62

68% of students earn between 36.46 and 60.62 salary.

- μ-2σ=48.54-2*12.08=24.38
- μ+2σ=48.54+2*12.08=72.7

95% of students earn between 24.38 and 72.7 salary.

- μ-3σ=48.54-3*12.08=12.29
- μ+3σ48.54+3*12.08=84.78

99.7% of students earn between 12.29 and 84.78 salary.

Empirical rule for Spending:

- μ-σ=482.01-221.95=260.06
- μ+σ=482.01+221.95=703.96

68% of students spend between 260.06 and 703.96.

- μ-2σ=482.01-2*221.95=38.11
- μ+2σ=482.01+2*221.95=925.91

95% of students spend between 38.11 and 925.91.

- μ-3σ=482.01-3*221.95= -183.83
- μ+3σ=482.01+3*221.95=1147.86

99.7% of students spend between -183.83 and1147.86.

Empirical rule for Text Messages:

- μ-σ=246.20-214.46=31.73
- μ+σ=246.20+214.46=460.65

68% of students receive between 31.73 and 460.65 text messages.

- μ-2σ=246.20-2*214.46= -182.72
- μ+2σ=246.20+2*214.46=675.12

95% of students receive between -182.72 and 675.12 text messages.

- μ-3σ=246.20-3*214.46= -397.18
- μ+3σ=246.20+3*214.46=889.57

99.7% of students receive between -397.18 and 889.57 text messages.

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

3.1: Do 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.

Solution:

A Null hypothesis (for shingles A& B) is a hypothesis that says there is no statistical significance between the two variables in the hypothesis. It is the hypothesis that the researcher is trying to disprove.

An Alternative hypothesis simply is the inverse, or opposite, of the null hypothesis.

- H_o: Mean moisture content <= 0.35
- H_a: Mean moisture content > 0.35

Alpha = 0.05(Level of significance of test)

Test:

One sample T-Test (Shingles A&B)

Shingle A:

- P value = 0.074776
- Tstat = -1.4735046

Inference:

- We failed to reject Null hypothesis as P value > 0.05.

Shingle B:

- P value = 0.002090
- Tstat = -3.10033130

Inference:

- We reject Null hypothesis as P value < 0.05.

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?

Solution:

- H_o: Population mean of shingle A = Population mean of shingle B
- H_a: Population mean of shingle A != Population mean of shingle B
- Alpha = 0.05 (Level of significance of test)

Assumptions:

- Define null and alternate hypothesis
- Assume alpha value if not mentioned.
- If σ known, proceed with one sample Z test.
- If σ unknown, proceed with one sample T test.
- If α is lesser than P value,reject null hypothesis.
- If α is greaterer than P value, fail to reject null hypothesis.

Test:

Two sample T-Test (Shingles A&B)

- P value = 0.20174
- Tstat = 1.28962

Inference:

- We failed to reject Null hypothesis as P value > 0.05

**********Thank You**********