SMDM BUSINESS REPORT

This reports is about solving a business problem using different statistic method like Descriptive ,Inferential and Hypothesis testing and deriving some business insights to create better opportunities and results for the business.

By:-
SAHIL SAXENA

CONTENT

FIGURE CONTENT

TABLE CONTENT

INTRODUCTION

This report is about different problems, related to that problem the method we used to solve that problem. After solving the problem, we derive the business insight from that problem and figure out how to use this insight to create better opportunities and result.

PROBLEM – 1.

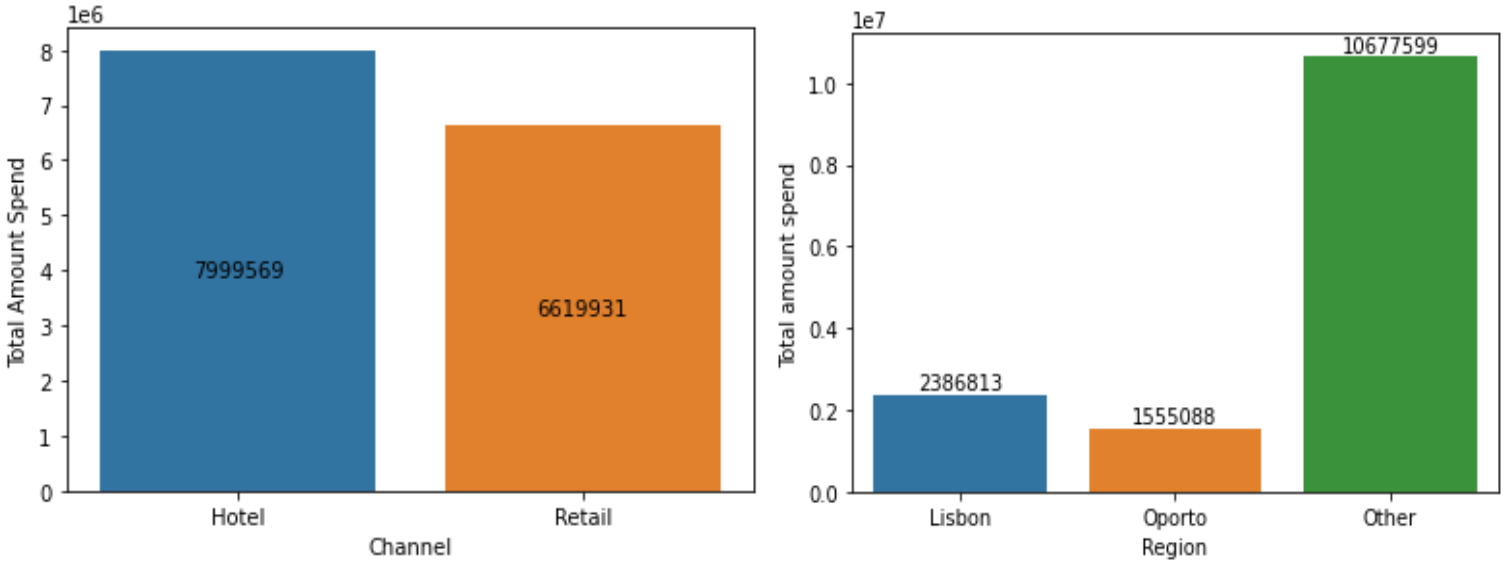
A wholesale distributor operating in different regions of Portugal has information on annual spending of several items in their stores across different regions and channels. The data consists of 440 large retailers' annual spending on 6 different varieties of products in 3 different regions (Lisbon, Oporto, Other) and across different sales channel (Hotel, Retail).

Q 1.1 -Use methods of descriptive statistics to summarize data.

1.1.1 Which Region and which Channel spent the most?

1.1.2 Which Region and which Channel spent the least?

Figure 1. 1 - Bar Plot – Total Amount Spend Vs Channel and Total Amount Spend Vs Region.



As we can from both plots which gives us the clear idea regarding the most and least spent channels and region which also gives the solution for our problems that

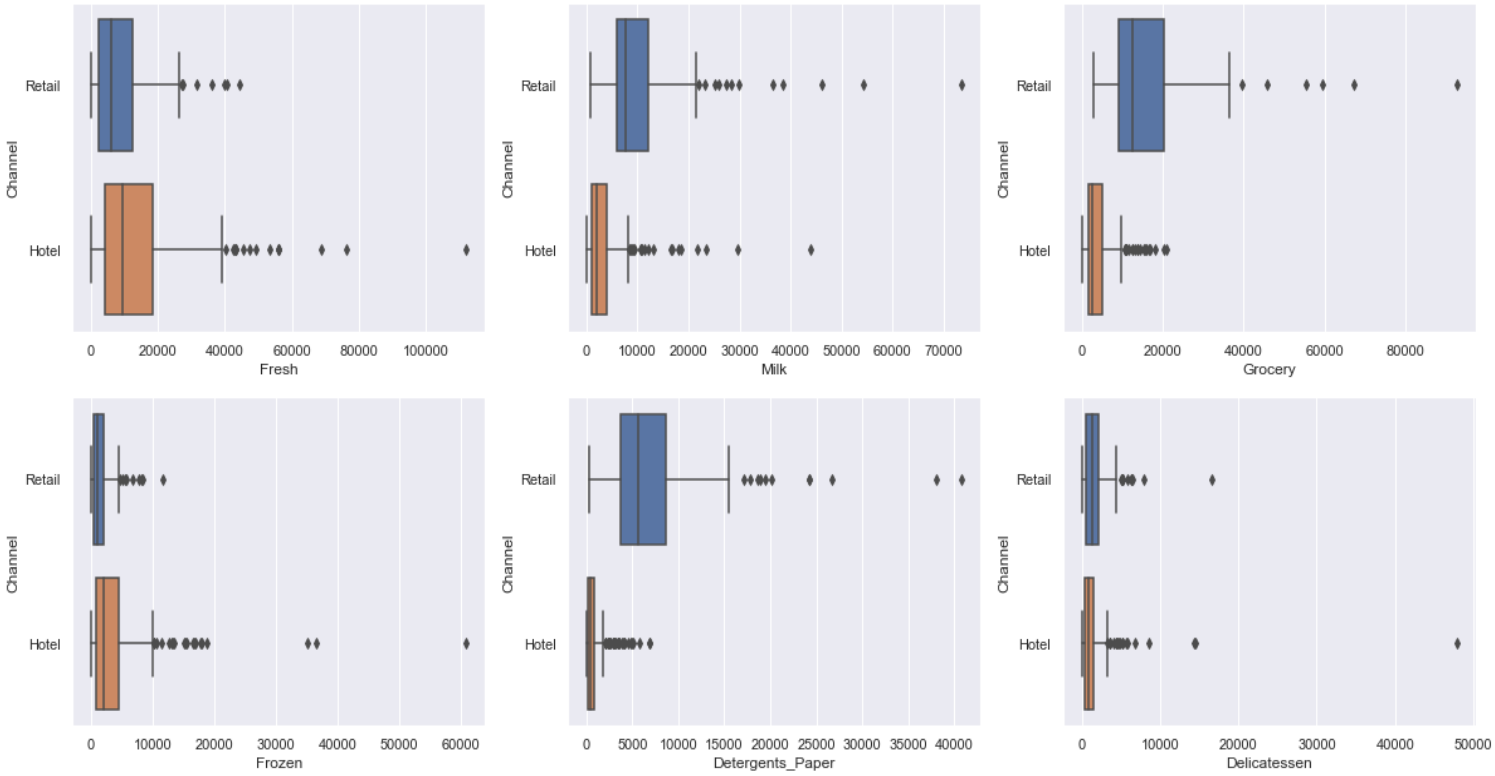
1.1 - Hotel in Channels and other in Regions, which spends the most that is around 79 lakhs and 1.06 crore respectively.

1.2 - Retails in Channels and Oporto region in Regions, which spends the least that is around 66 lakhs and 15.5 lakhs respectively.

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.

Items Behavior Across Channel

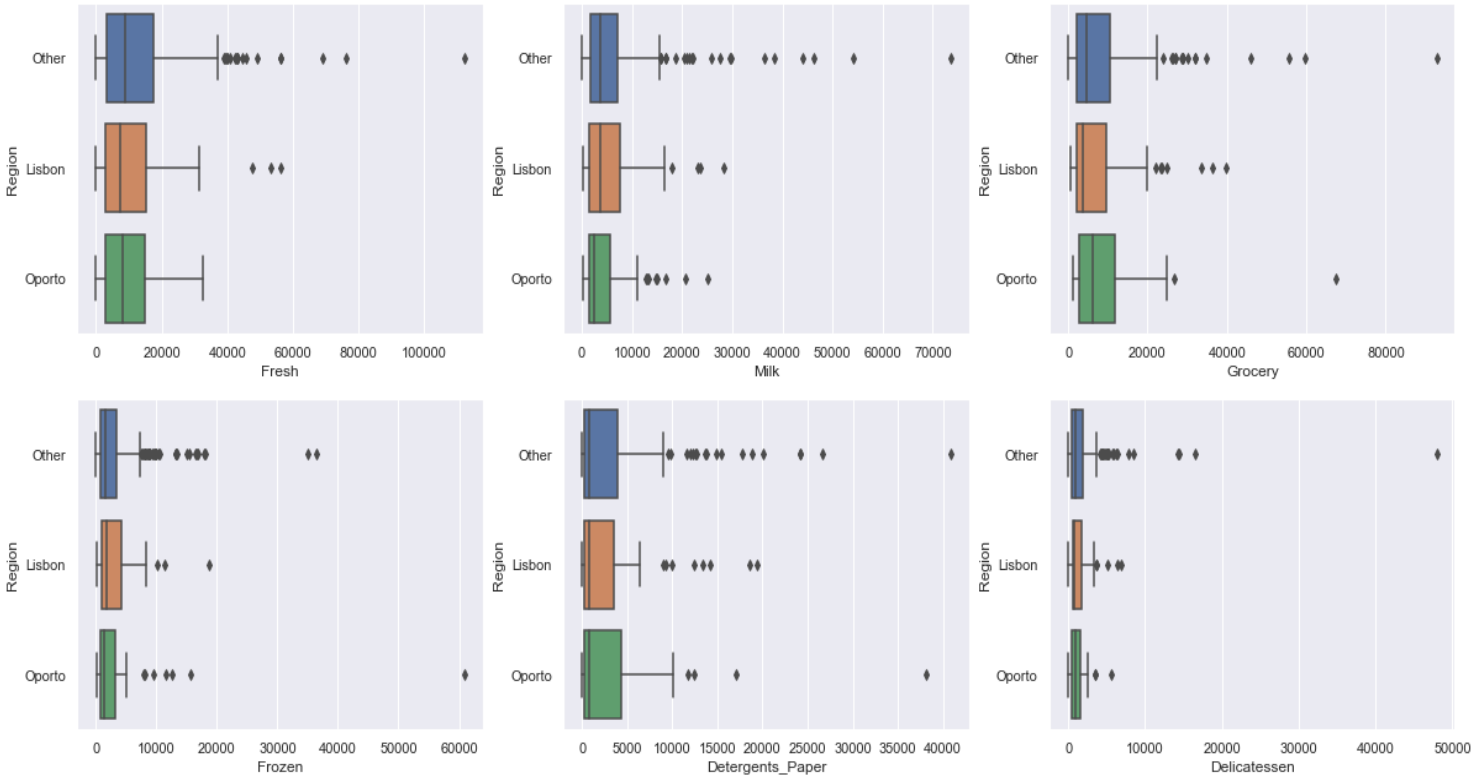
Figure 1. 2 - Box Plot – Items Across Channel



1. Fresh item distribution is good, very slightly right skewed in both hotel and Retail, CV value is low 0.97 and 0.99 respectively which means it's very consistent. There are too many outliers as an enormous difference in max and min with 3rd and 1st quartile, respectively. Hotel channel has high value of outliers as compare to retail.
2. Milk sale is more in retail channel in comparison to hotel channel. 1st quartile has significant difference in both channels. Distribution of milk in retail channel is right skewed ad has too many outliers, which means big amount of money has spent to buy milk through retail channels.
3. There is a substantial difference in spending of Grocery in both channels. Retail channel spends more as compare to hotel channel, distribution in both channels is positively skewed and both have outliers but retail channel has high value spending on grocery and outliers spreads much farther than the maximum value.
4. Frozen items are more preferred through hotel channels, the spread is closely spread and a bit right skewed. There are some outliers in both channels but the Hotel Channel has high value outliers.
5. Spending on Detergent papers are much more in retail channel as compare to hotel channel. The spread of the item in retail channel is wide spread and outliers has high values while Hotel channel is closely spread and outliers also spread closely beyond maximum value.
6. Delicatessen item spending is almost similar in both channels and it is also closely spread. tin both channels it follows almost normal distribution. The outliers value is much higher in hotel as compare to Retail Channel.

Items Behavior Across Region

Figure 1. 3 - Box Plot – Items Across Region

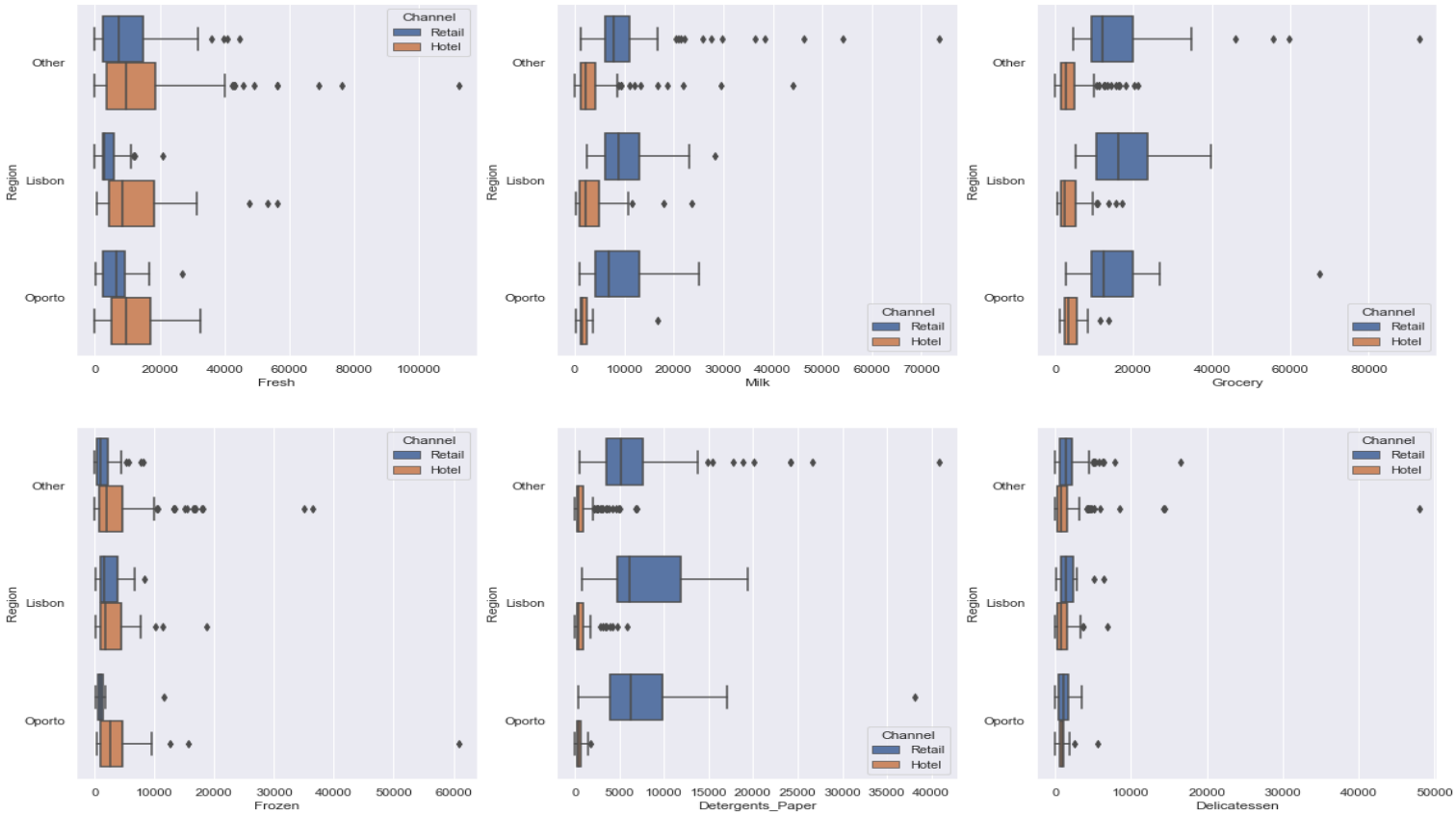


.

1. Fresh items are much more preferred in other region as compare to Lisbon and Oporto but there is not much difference in the spending. These distributions follow almost a normal distribution as they have little skewness. The Outliers are high valued in other region than Lisbon and Oporto.
2. The spending of money on milk is much higher in other region. The spread of variables is almost similar in Lisbon and other region but Oporto is bit closely spread and right skewed. There are outliers in all the region but highly and largely spread outliers are in other region.
3. Grocery item are preferred in all three regions and bit Oporto has higher spending in all three. All three region has wide spread of data and the distribution is also a bit right skewed. Oporto mean spending is higher but the spread of outliers is much wider in other area.
4. There is not much difference in spending of money to buy Frozen items in all three regions. The data is closely spread in all three regions and distribution is right skewed. Other region has much more outliers as compare to other region.
5. For detergents paper customer spends more in other region but there are too many outliers to affect the data. The distribution of the data is wide spread and highly right skewed. Other region has much more spread in outliers.
6. Delicatessen almost liked equally by all three region. The data is closely spread and IQR is small for all three regions and there are wide spread outliers present in other region

Items Behavior Across Region and Channel

Figure 1. 4 - Box Plot 3 – Items Across Channel and Region



1. Fresh items are mostly preferred by Hotel Channels in all three regions. Their IQR is wide spread in all three regions and the distribution is a bit right skewed. The outliers present in all three regions but most widely spread in other region with high values.
2. Milk is mostly preferred by retail channels in all three regions. The spread of data is much wider in Oporto region as compare to others and there are no outliers present in Oporto region which makes its more exact as other and Lisbon region has too many outliers which can affect the mean spending.
3. Grocery is also Preferred by Retail channels in all three regions but Lisbon the region where customer spends the most. The spread of data is also wider in Lisbon and almost follow a normal distribution.
4. Frozen items are preferred by hotel channels in all three regions as their average spending is almost similar, but Oporto region is higher in all three and follows almost normal distribution and closely spread data. there are outliers present in the data for all three regions which can affect the mean and standard deviation.
5. Detergent Paper has more spending by retail channels than hotel channels in all three regions, but the most spending is in Lisbon region. The data is wide spread in Lisbon region and its positively skewed. as there are outliers present in other and Oporto region that can affect is mean and standard deviation.
6. Average spending on Delicatessen is almost similar in both channels and in all three regions. Oporto is least spending area through hotel channel. All the data are closely spread. There are outliers present in all the regions and Channels.

Q 1.3 - On the basis of a descriptive measure of variability, which item shows the most inconsistent behavior? Which items show the least inconsistent behavior?

Table 1. 1 - Items Description

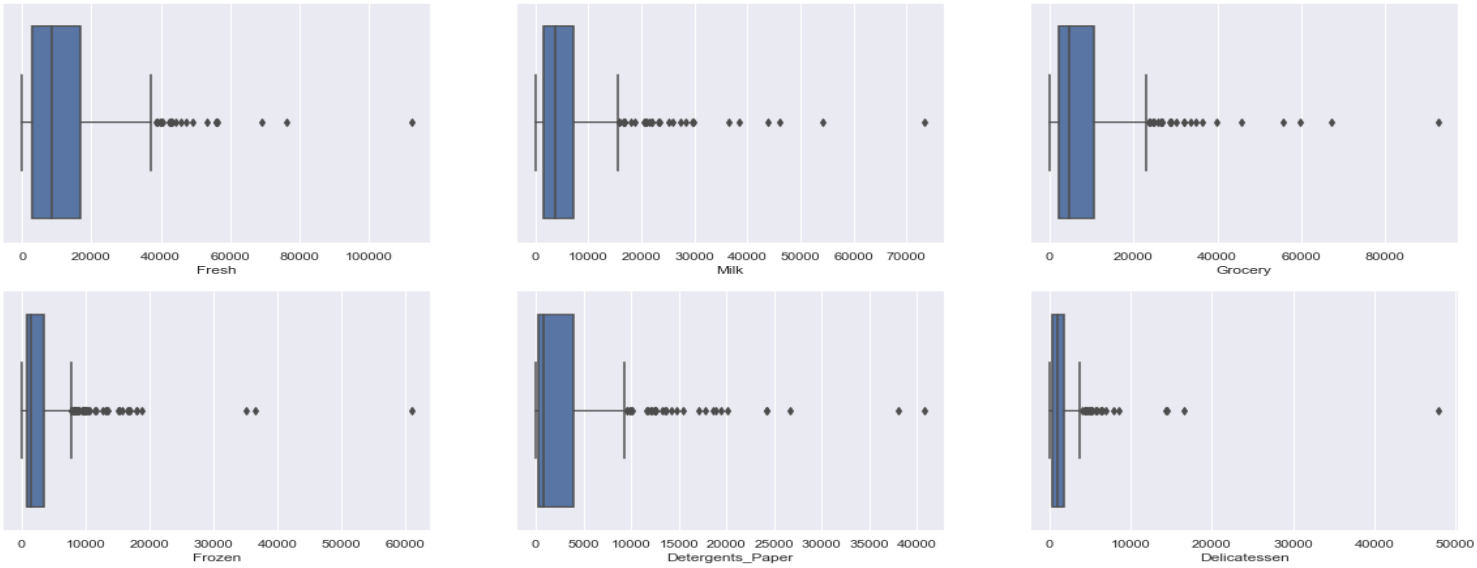
	Fresh	Milk	Grocery	Frozen	Detergents Paper	Delicatessen
count	440	440	440	440	440	440
mean	12000.29773	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	12647.32887	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	3	55	3	25	3	3
25%	3127.75	1533	2153	742.25	256.75	408.25
50%	8504	3627	4755.5	1526	816.5	965.5
75%	16933.75	7190.25	10655.75	3554.25	3922	1820.25
max	112151	73498	92780	60869	40827	47943
CV	1.053918	1.273299	1.195174	1.580332	1.654647	1.849407

Based on the table and values of Coefficient of Variance (CV)

1. The item that shows most inconsistence behavior is Delicatessen as it has highest CV value of 1.849407.
2. The item that shows least inconsistence behavior is Fresh Items as it has the lowest CV value of 1.053918.

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

Figure 1. 5 - Box Plot – Items Outliers.



From the plot we can see that every item has many outliers.

Q.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.

On the analysis of the above data there are few recommendations that we can make,

There is a substantial difference in the sales of given 6 items in hotel and retail channels, those which has good sales in hotel channels there is less sales in retail channels and vice versa. Like Fresh, frozen these items have very good sales in hotel channels, but we can change the strategies for the sales of these items in retail channels. We can find the root cause for such difference in sales for these items in both channels and change our strategies according to that and then try to increase sales. Similarly with Milk, Grocery and detergent paper have a good sale in retail channels as compare to hotel channels. we can around the root cause for these and can try different strategies and try to increase sales.

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).

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

2.1.1. Gender and Major

Table 2. 1 - Gender Vs Major

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

The above tables contain Males and Females distributed among their choice of major.

2.1.2. Gender and Grad Intention

Table 2. 2 – Gender Vs Grad Intention

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

This tables contains the intention of students intends to graduate.

2.1.3. Gender and Employment

Table 2. 3 – Gender Vs Employment

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

This tables have the details of students' who are working and not working.

2.1.4. Gender and Computer

Table 2. 4 - Gender Vs Computer

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

This tables gives the detail about what kind of computer does students have.

Q-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?

Solution:

Total number of males (A) = 29

Total number Students (B) = 62

P(A)= A/B = 29/62

The probability that a randomly selected CMSU student will be male is 46.77%.

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

Solution:

Total number of Females (A) = 33

Total number Students (B) = 62

P(A)= A/B = 33/62

The probability that a randomly selected CMSU student will be Female is 53.23%.

Q-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.

Table 2. 5 - Male Vs Major Probability

Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Male	4	1	4	2	6	4	5	3
Probability (%)	14	3	14	7	21	14	17	10

This Table comprises of conditional probability of different majors among the male candidates. Upper(1st) row consists the number of male students in their respective majors and below (2nd) row consists of probability of Male students in their respective major.

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

Table 2. 6 - Female Vs Major Probability

Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Female	3	3	7	4	4	3	9	0
Probability (%)	9	9	21	12	12	9	27	0

This Table comprises of conditional probability of different majors among the Female candidates. Upper(1st) row consists the number of Female students in their respective majors and below (2nd) row consists of probability of Female student in their respective major.

Q-2.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:

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

Solution:

Number of male students who intends to Graduate is = 17 .....(Refer Q 2.1.2, Table 2.1)

Total number of students = 62

Probability =17/62

The probability That a randomly chosen student is a male and intends to graduate is 27.42%.

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

Solution:

Number of Female who does have a laptop is = 29 .....(Refer Q 2.1.4, Table 2.4)

Total number of Female Students is = 33

Number of Female students who does not have a laptop is = 33-29 = 04

Total number of students = 62

Probability =4/62

The probability that a randomly selected student is a female and does NOT have a laptop is 6.45%

Q- 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?

Solution:

Total number of Male students (M) = 29

Total Number of Full-time employment students (FE) = 10 .....(Refer Q2.1.3, Table 2.3)

Total Number of Students who are male and Full-time employed $(M \cap FE) = 07$

P(M)=29/62 = 0.46774193548387094

P(FE)=10/62 = 0.16129032258064516

P(M ∩ FE) = 0.11290322580645161

P(M U FE) = P(M) + P(FE) - P(M ∩ FE)

The probability that a randomly chosen student is a male or has full-time employment is 0.516 i.e., 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:

Total Number of Female students (A) = 33

Total Number of Female students doing International Business (IB) = 04

Total Number of Female students doing Management (M) = 04

Probability of International or management given female is P((IB U M) | Female)

P((IB U M) | F) = P(IB|F) + P(M|F) - P((IB ∩ M) | F)

P(IB | F) = 4/33

P(M | F) = 4/33

P((IB ∩ M) | F) = 0 (as these are Mutually exclusive events)

P((IB U M) | F) = 4/33 + 4/33

The conditional probability that given a female student is randomly chosen, she is majoring in international business or management is. 24.24 %.

Q-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:

Graduate Intention considered to be Yes = GI

Female = F

For graduate intention and being female are independent events this following condition should be satisfied.

$P(GI \cap F) = P(GI) * P(F)$

Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

Table 2. 7 - Gender Vs Grad Intention (altered)

Let's Check,

P(GI)= 28/40 = 0.7

P(F) = 20/40 = 0.5

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

$P(GI \cap F)$ = 11/40 = 0.275

As we can check these are not independent events as probability multiplication of both events is not equal to combined event, so being a Graduate Intention and being female candidate are not independent events.

Q-2.7. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages.

Answer the following questions based on the data

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

Solution:

Total Number of students whose GPA is less than 3 are (A) = 17

Total Number of Students = 62

Probability of A i.e., P(A) = 17/62

If a student is chosen randomly, the probability that his/her GPA is less than 3 is 27.42 %

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?

Solution:

Part – 1.

The conditional probability that a randomly selected male earns 50 or more

Total Number of Males who earns 50 or more (A) = 14

Total Number of Males (B) = 29

Probability P (A | B) = 14/29

The conditional probability that a randomly selected male earns 50 or more is 34.48 %.

Part - 2.

The conditional probability that a randomly selected females earns 50 or more

Total number of Females who earns 50 or more (A) = 18

Total number of Females (B) = 33

Probability P (A | B) = 18/33

The conditional probability that a randomly selected female earns 50 or more is 39.39 %.

Q-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.

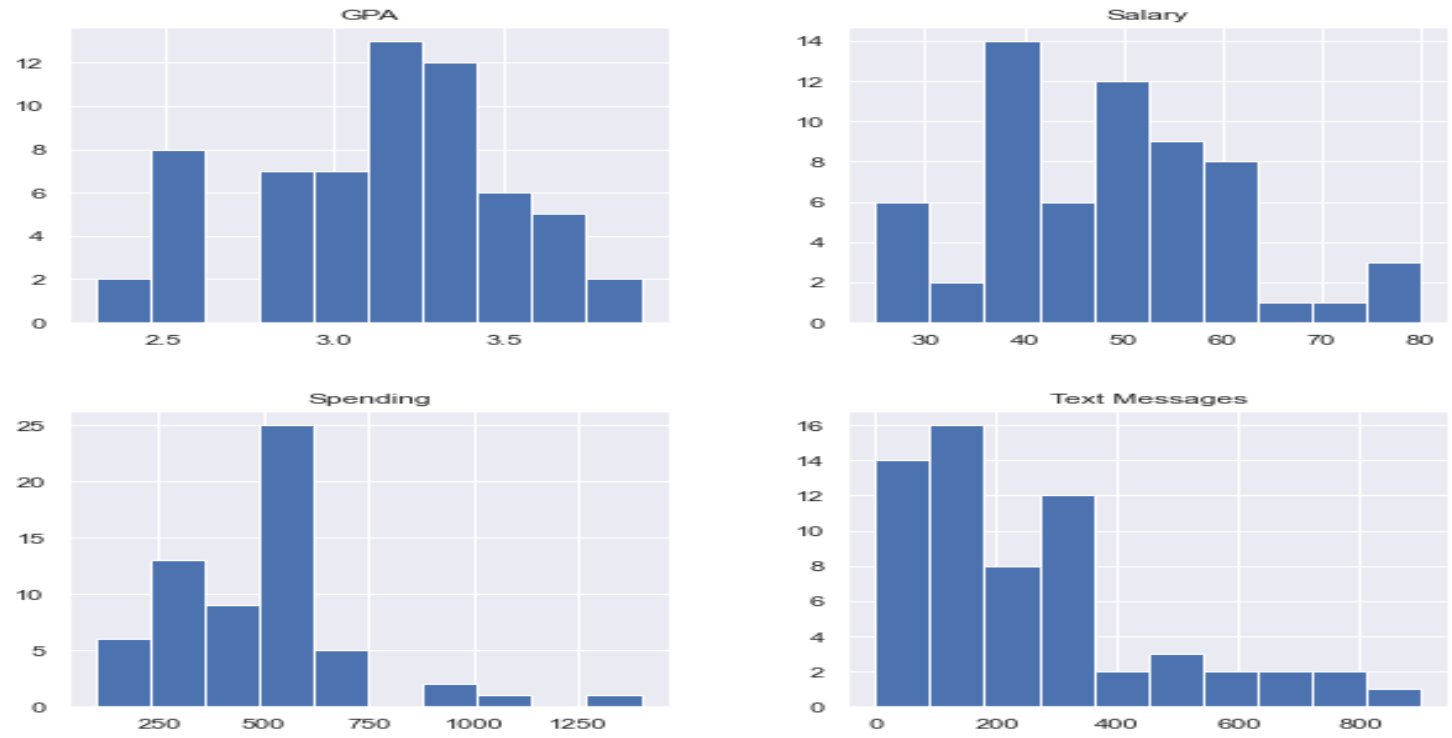


Figure 2.1 - Histogram Plot – Distribution Plot

Solution:

As we can see from the plots and stats that

For GPA,

Mean of GPA = 3.129032258064516
Median of GPA = 3.1500000000000004
Mode of GPA = 0 3.0
Skewness of GPA = -0.3069370607018797

Conclusion - GPA follows a normal distribution as its mean, median and mode are almost equal and skewness is also very slight

Skewness Limit

1. -0.5 to 0.5 – symmetrical distribution.
2. -1 to 1 – Moderately Skewed.
3. Above -1 or 1 – Highly Skewed

For Salary,

Mean of Salary = 48.54838709677419
Median of Salary = 50.0
Mode of Salary = 0 40.0
Skewness of Salary = 0.5216766008645851

Conclusion – As mean, median and mode are not equal and skewness is also a bit high, Salary does not follow a normal distribution.

For Spending,

Mean of Spending = 482.01612903225805
Median of Spending = 500.0
Mode of Spending = 500 .0
Skewness of Spending = 1.547250312929523

Conclusion – As we can check for spending mean, median and mode is not equal and skewness is also very high which means that spending does not follow a normal distribution.

For Text Messages,

Mean of Text Messages = 246.20967741935485
Median of Text Messages = 200.0
Mode of Text Messages = 0 300
Skewness of Text messages. = 1.2642446834439687

Conclusion – Text Message's mean , median and mode is not equal and its skewness is also high which means it does not follow normal distribution.

Q 2.8.2 Write a note summarizing your conclusions

Conclusion,

As we studied the survey data, we can say few things about the survey that, there are a greater number of males who intends to graduate as compare to female, as among females' lot of them is still undecided whether to graduate or not. students who intend to graduate most liked stream to major from among males is Management and Retail/Marketing and among females its Retail/marketing and Economics/finance, so we ca say among all students who intends to graduate most liked major is Retail/Marketing. Around 85% of students are working and most of them prefer part-time jobs and uses laptop instead of desktop or tablet. Among male students who earns good salary (more than or equal to 50) most of them have intention to graduate where as high paid working females (more than or equal to 50) have mixed feeling about Graduate intention.

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.¶

Note - The file (A & B shingles.csv) includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles

Q-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

Solution:

Step 1 – Hypothesis:

For A shingles

1. Null Hypothesis-H_0 μ <= mean weight 0.35 per 100 square feet
2. Alternative Hypothesis - H_A, μ > mean weight 0.35 per 100 square feet

For B Singles

1. Null Hypothesis- H_0 μ <= mean weight 0.35 per 100 square feet
2. Alternative Hypothesis - H_A, μ > mean weight 0.35 per 100 square feet

Table 3.1 – Shingles Description

	A	B
count	36	31
mean	0.316667	0.273548
std	0.135731	0.137296
min	0.13	0.1
25%	0.2075	0.16
50%	0.29	0.23
75%	0.3925	0.4
max	0.72	0.58

Step 2 - Significance Level α:

For Both samples of shingles (A & B)

alpha (α) = 0.05

Checking the hypothesis at 5 % significance level

Step - 3 Identifying t_{stat} for A and B samples

As n is larger than 30 in both cases but population standard deviation is not given

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

For A Shingles

X bar = 0.316667

μ = 0.35

S = 0.135731

N = 36

Degree of freedom = 35

t_{stat} = (0.316667 - 0.35) / (0.135731 /Sqrt (36)) or Using Python code ttest_1samp getting the same result.

t_{stat} = -1.4735046253382782

For B singles:

X bar = 0.273548

μ = 0.35

S = 0.137296

N = 31

Degree of freedom = 30

t_{stat} = (0.273548 - 0.35) / (0.137296 /Sqrt (31)) or Using Python code ttest_1samp getting the same result.

t_{stat} = -3.1003313069986995

Step 4 : Calculating p value

For A shingles

P value = 0.07477633144907513 > 0.05 (alpha)

For B shingles

P value = 0.0020904774003191826 < 0.05 (alpha)

Step 5: Conclusion

For Shingles A, p value > alpha

We fail to reject the null hypothesis and the process is still under control as mean weight shingles is less than the permissible limit 0.35 pounds per 100 square feet.

For B shingles, p value < alpha

We have enough evidence to Reject the null hypothesis and state that process is not under control as mean weight of shingles is more than the permissible limit 0.35 pounds per 100 square feet.

Q-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?

Step 1 - Hypothesis

1. Null Hypothesis- $\mu_A = \mu_B$ Population means are equal.
2. Alternative Hypothesis – $\mu_A \neq \mu_B$ Population means are not equal.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

Step 2 - Significance Level α

alpha (α) = 0.05

Checking the hypothesis at 5 % significance level

Step - 3 Identifying t$stat$ for A and B samples

X1 bar = 0.316667

X2 bar = 0.273548

S1 = 0.135731

S2 = 0.137296

N1 = 36

N2 = 31

t_{stat} = 1.2896282719661123 or Using Python Code ttest_ind for the same.

Assumptions need to check before the test,

1. Data values must be independent. Measurements for one observation do not affect measurements for any other observation.
2. Data in each group must be obtained via a random sample from the population.
3. Data in each group are normally distributed.
4. Data values are continuous.
5. The variances for the two independent groups are equal

Step 4: Calculating p value

P value = 0.2017496571835306 > 0.05 (alpha)

Step 5: Conclusion

We don't have evidence so; we fail to reject the null hypothesis and the mean weight of the samples is equal.