**STATISTICAL METHODS FOR DECISION MAKING PROJECT**

**SUBMITTED BY,**

**SHARA GEORGE VAIDIAN**

|  |  |  |
| --- | --- | --- |
| **NO** | **TITLE** | **PAGE NO:** |
| 1 | **Problem 1 – Wholesale Customer Analysis** |  |
|  | **Question 1.1** | **4** |
|  | **Question 1.2** | **6** |
|  | **Question 1.3** | **8** |
|  | **Question 1.4** | **9** |
|  | **Question 1.5** | **10** |
|  |  |  |
| **2** | **Problem 2** | **11** |
|  | **Question 2.1** |  |
|  | **Question 2.1.1** | **12** |
|  | **Question 2.1.2** | **12** |
|  | **Question 2.1.3** | **12** |
|  | **Question 2.1.4** | **12** |
|  | **Question 2.2** |  |
|  | **Question 2.2.1** | **13** |
|  | **Question 2.2.2** | **13** |
|  | **Question 2.3** |  |
|  | **Question 2.3.1** | **14** |
|  | **Question 2.3.2** | **15** |
|  | **Question 2.4** |  |
|  | **Question 2.4.1** | **16** |
|  | **Question 2.4.2** | **16** |
|  | **Question 2.5** |  |
|  | **Question 2.5.1** | **18** |
|  | **Question 2.5.2** | **18** |
|  | **Question 2.6** | **20** |
|  | **Question 2.7** |  |
|  | **Question 2.7.1** | **21** |
|  | **Question 2.7.2** | **21** |
|  | **Question 2.8** | **22** |
|  |  |  |
| **3** | **Problem 3** |  |
|  | **Question 3.1** | **23** |
|  | **Question 3.2** | **25** |

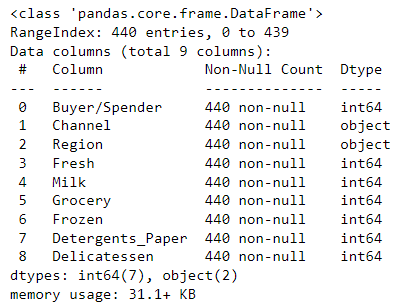
**PROBLEM:1**

**WHOLESALE CUSTOMER ANALYSIS**

Summary:

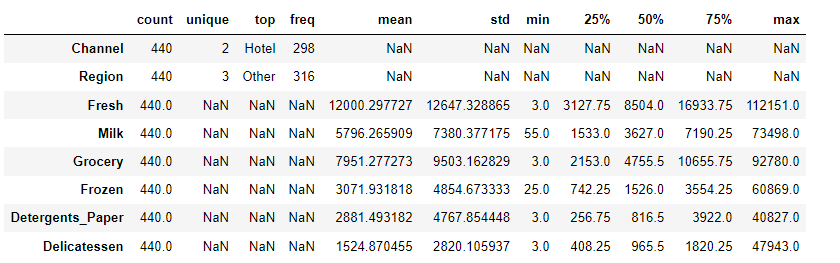
A wholesale distributor based in Portugal has the information on the annual spending of several items in their stores(Fresh, Milk, Grocery, Frozen, Detergents Paper, Delicatessen across different regions and channels. The data consists of the annual spending of 440 retailers on 6 different varieties of products in 3 different regions (Lisbon, Oporto, Other) and across different sales channel (Hotel, Retail).

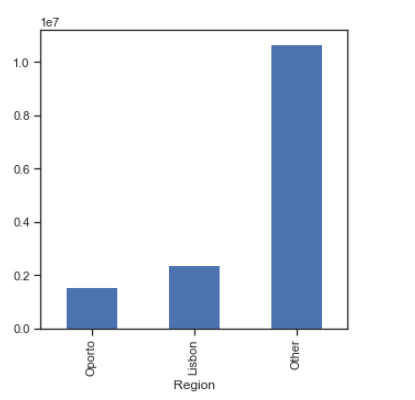
 We have imported the ‘Wholesale Customer data’ dataset in Jupyter Notebook to analyze the spending under each store items across regions and channel to find solutions to each problem.



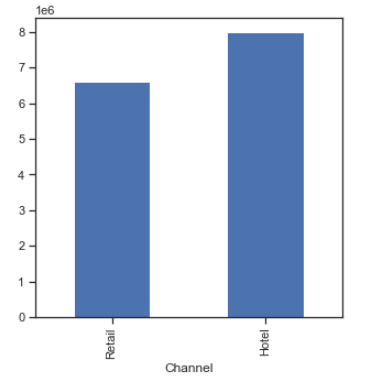
Here we can see that the dataset has 9 variables out of which 2 are categorical (Channel and Region) and 7 are integer type (Buyer/Spender, Fresh, Milk, Grocery, Frozen, Detergents\_Paper, Delicatessen). There are no null values in any of the columns.

**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?**



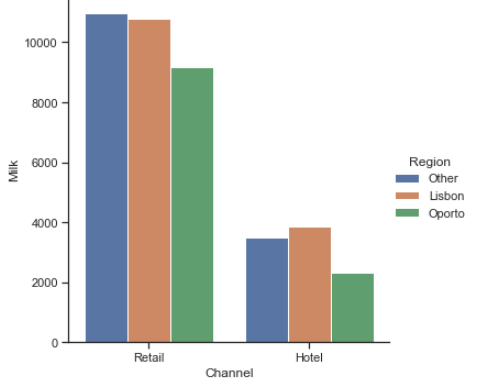
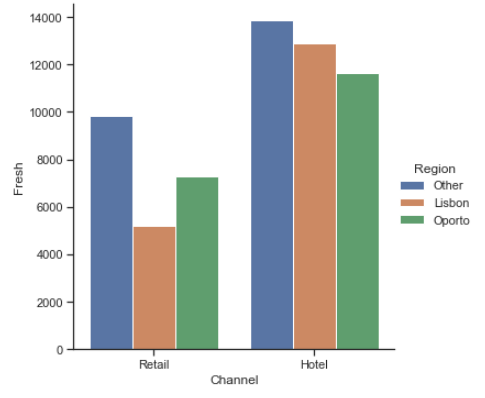
****

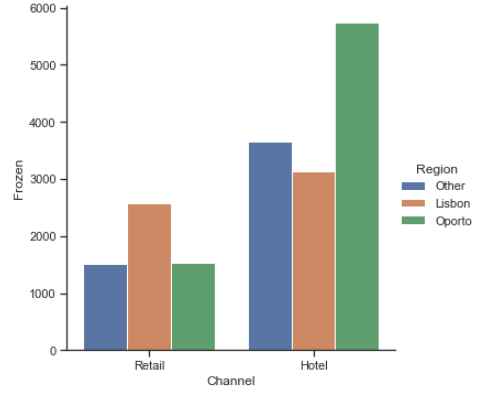
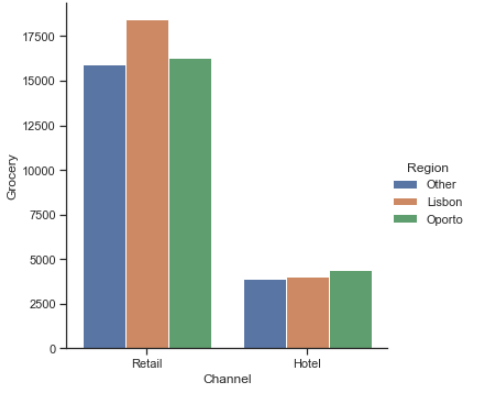
From the above plot it can be concluded that the region Other has the highest spending and the region Oporto has the lowest spending

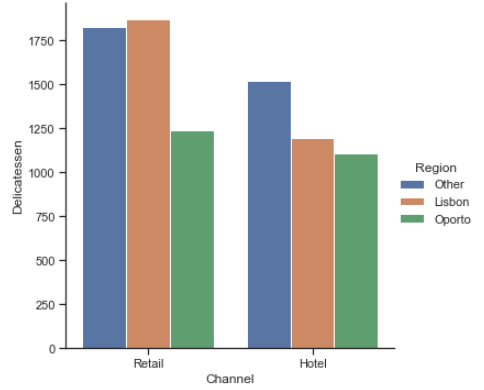
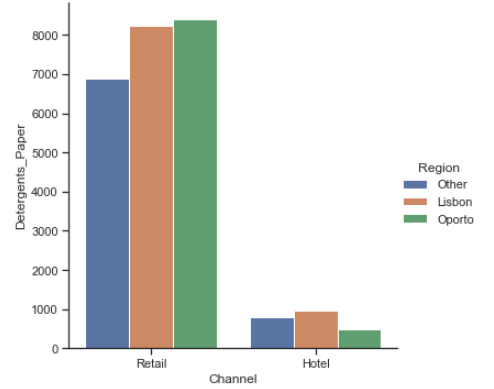
****

From the above plot it can be concluded that the channel Hotel has the highest spending when compared to the channel Retail

**Q1.2 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.**

****

****

****

Looking at the above tables, we can see that some categories like Milk, Grocery and Detergents\_Paper have a higher spend in the Retail channel than the Hotel Channel in all regions. On the other hand, Fresh and Frozen have higher consumption in the Hotel channel than the Retail Channel in all regions. Also, if we plot a box plot we can summarize that the spend for Fresh and groceries is the maximum across region and channel while for Delicatessen it is the minimum across region and channel.

Therefore we can conclude that the 6 varieties does not show similar behavior across the Region and Channel

**Q1.3 On the basis of a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?**

The Coefficient of Variations of all 6 varieties are given below:

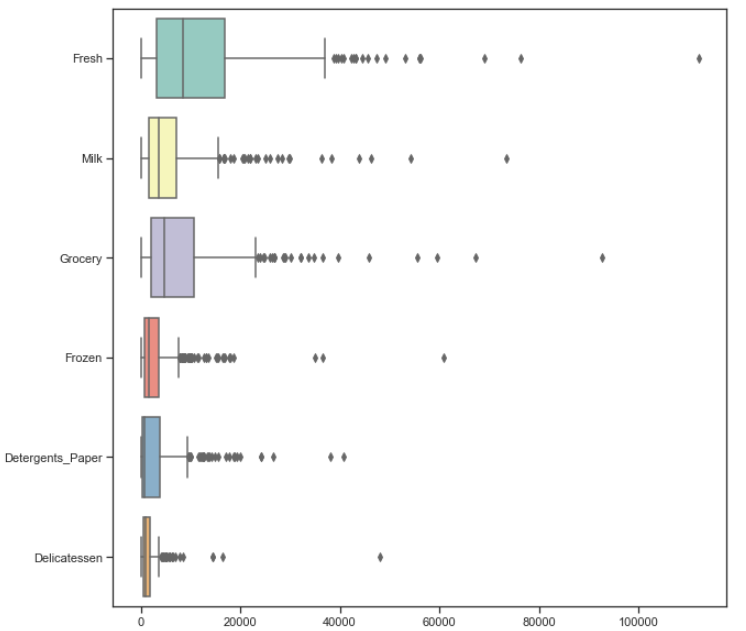
(calculated using the statistical function)

|  |  |
| --- | --- |
| Varieties | Coefficient of Variation |
| Fresh | 1.0527196084948245 |
| Milk | 1.2718508307424503 |
| Grocery | 1.193815447749267 |
| Frozen | 1.5785355298607762 |
| Detergents\_Paper | 1.65276588104729 |
| Delicatessen | 1.8473041039189306 |

From the table given above it is evident that the Coefficient of Variance is highest for variety Delicatessen and lowest for variety Fresh. Therefore it can be concluded that the variety that shows the most inconsistent behavior is Delicatessenand the variety that shows the least inconsistent behavior is Fresh.

**Q1.4 Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.**

To check whether any outliers are present or not we should create a Boxplot of all the variables. The corresponding Boxplot is given below:



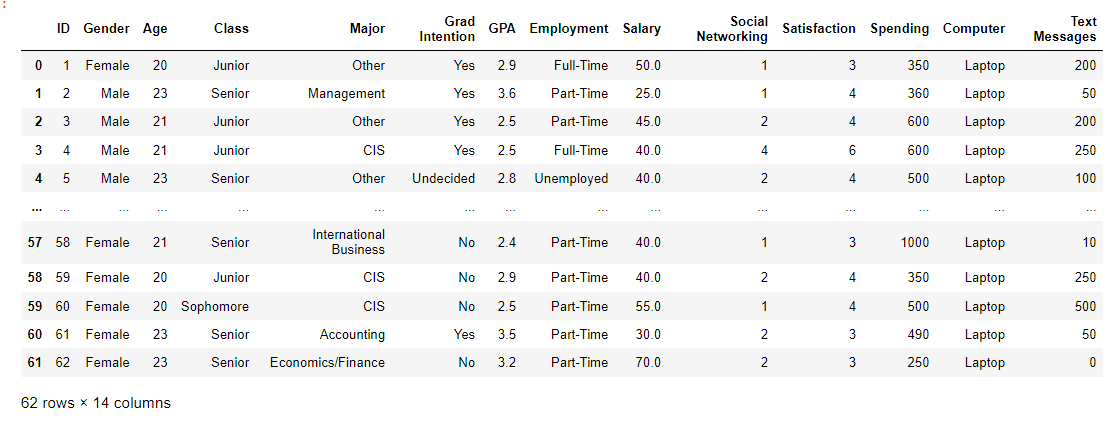
 From the Box plots of all the Variables as above it can be concluded that Yes, There are outliers in the Data.Outliers are present in the variables Fresh, Milk, Grocery, Frozen, Detergents\_Paper and Delicatessen

**Q1.5  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**

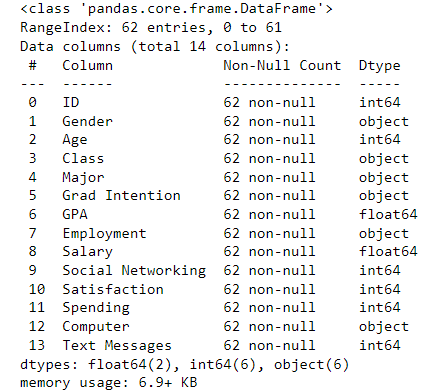
* While calculating the coefficient of variation we could find out that there are certain inconsistencies in the spending of the different varieties which should be minimized
* In all the regions the food item Fresh has the highest spending. It is followed by grocery and milk. Therefore the business needs to focus more on these items giving more priority to the Fresh food items
* Delicatessen items has less demand in all region. The distributor must focus on that.

**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 .The Data is stored in the dataset as follows:



The dataset has 62 rows and 14 columns.



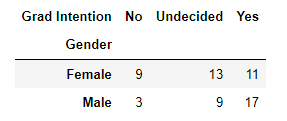
Here we can see that the dataset has 14 variables out of which 6 are categorical (Gender, Class, Major, Grad Intention, Employment, Computer) , 6 are integer type (ID, Age, Social Networking, Satisfaction, Spending, Text Messages ) and 2 float type (GPA, Salary).

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

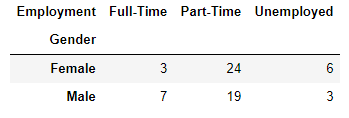
**2.1.1. Gender and Major**



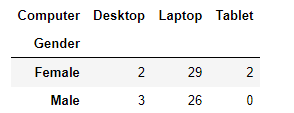
**2.1.2. Gender and Grad Intention**

****

**2.1.3. Gender and Employment**

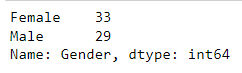
****

**2.1.4. 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?**

****

Using the value\_counts() function in python we get the total number of male and female students.

Total no: of students = 62

Total no: of male students = 29

Probability that a randomly selected CMSU student will be a male = (Total no: of male students/Total number of students)

=29/62

=0.46774194

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

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

Using the value\_counts() function in python we get the total number of male and female students.

Total no: of students = 62

Total no: of female students = 33

Probability that a randomly selected CMSU student will be a female = (Total no: of female students/Total number of students)

=33/62

=0.53225806

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

## 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.**

**Contingency table for Gender and Major**



From the contingency table it can be concluded that:

Among Male students:

Probability of Accounting - 4/29

Probability of CIS - 1/29

Probability of Economics/Finance - 4/29

Probability of International Business - 2/29

Probability of Management - 6/29

Probability of other - 4/29

Probability of Retailing/Marketing - 5/29

Probability of Undecided - 3/29

From the calculations we can conclude that:

Probability of Accounting - 4/29 = 0.13793103 = 13.79%

Probability of CIS - 1/29 = 0.03448276 = 3.45%

Probability of Economics/Finance - 4/29 = 0.13793103 = 13.79%

Probability of International Business - 2/29 = 0.06896552 = 6.9%

Probability of Management - 6/29 = 0.20689655 = 20.69%

Probability of other - 4/29 = 0.13793103 = 13.79%

Probability of Retailing/Marketing - 5/29 = 0.17241379 = 17.24%

Probability of Undecided - 3/29 = 0.10344828 = 10.34%

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

**Contingency table for Gender and Major**



From the contingency table it can be concluded that:

Among Female students:

Probability of Accounting – 3/33

Probability of CIS – 3/33

Probability of Economics/Finance - 7/33

Probability of International Business – 4/33

Probability of Management – 4/33

Probability of other – 3/33

Probability of Retailing/Marketing – 9/33

Probability of Undecided – 0/33

From the calculations we can conclude that:

Probability of Accounting – 3/33 = 0.09090909 = 9.09%

Probability of CIS – 3/33 = 0.09090909 = 9.09%

Probability of Economics/Finance - 7/33 = 0.21212121 = 21.21%

Probability of International Business – 4/33 = 0.12121212 = 12.12%

Probability of Management – 4/33 = 0.12121212 = 12.12%

Probability of other – 3/33 = 0.09090909 = 9.09%

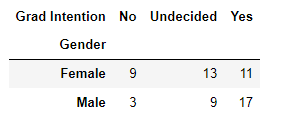
Probability of Retailing/Marketing – 9/33 = 0.27272727 = 27.27%

Probability of Undecided – 0/33 = 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.**

**Contingency table for Gender and Grad Intention:**

****

**From the contingency table we can find that:**

**Probability that a randomly chosen student is a male = 29/62**

**Probability of male that intends to graduate = 17/29**

**Therefore we can conclude that:**

**The probability that a randomly chosen student is a male and intends to graduate = (Probability that a randomly chosen student is a male) \* (Probability of male that intends to graduate)**

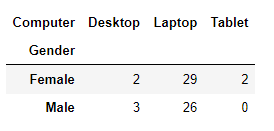
**= (29/62) \* (17/29)**

**= 0.27419355**

**=27.42%**

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

**Contingency Table for Gender and Computer :**



**From the contingency table we can find that:**

**Probability that a randomly chosen student is a male = 33/62**

**Probability of female that does not have a laptop = 4/33**

**Therefore we can conclude that:**

**The probability that a randomly chosen student is a male and intends to graduate = (Probability that a randomly chosen student is a female) \* (Probability of female that does not have a laptop)**

**= (33/62) \* (4/33)**

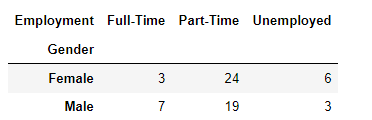
**= 0.06451613**

**=6.45%**

**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?**

**Contingency table for Gender and Employment**



From the contingency table we can find that :

Probability of a student being a male = 29/62 = 0.46774193548387094

Probability of a student having full time employment = 10/62 = 0.16129032258064516

Probability of male student having full time employment = 7/62 = 0.11290322580645161

Probability that a randomly chosen student is a male or has full time employment = Probability of a student being a male + Probability of a student having full time employment + Probability that a male student has full time employment

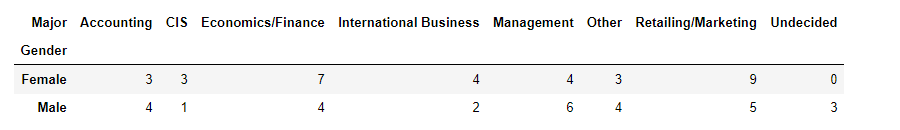
= 0.46774193548387094 + 0.16129032258064516 + 0.11290322580645161

= 0.74193547

= 74.19%

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

Contigency table for Gender and Major:



From the contingency table we can find that:

Probability of International Business given female = 4/62= 0.06451613

Probability of Management given Female = 4/62 = 0.06451613

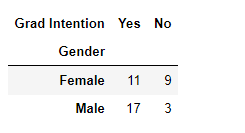
Probability of International Business or Management given female= Probability of International Business given female + Probability of Management given Female

=0.06451613 + 0.06451613

= 0.12903226

= 12.9%

**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?**



For 2 events to be independent, following condition is to be satified

P(A ∩ B) = P(A) \* P(B)

Total no: of students = 40

So, P (Grad Intention ∩ Female) = P(Grad Intention) \* P(Female)

P(Female) = 20/40 = 0.5

P(Grad Intention) = 28/40 = 0.7

P(Grad Intention) \* P(Female) = 0.5 × 0.7 = 0.35

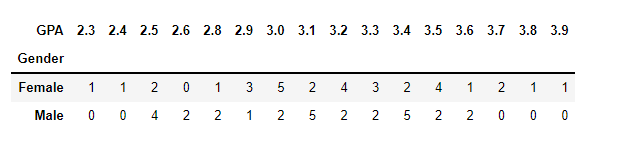
P (Grad Intention ∩ Female) = 11/40 = 0.275

This is not independent events as probability multiplication of both events is not equal to combined event, so being a Grad 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?**

Contigency table for Gender and GPA:



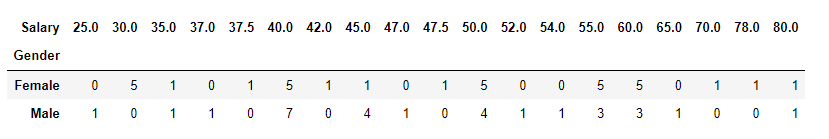
Total no: of students = 62

No: of students whose GPA is less than 3 = 17

Probability that a student’s GPA is less than 3 = 17/62 = 0.27

**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.**

**Contingency table for Gender and Salary**



From the contigency table we can find that the total number of male candidates who earn 50 or more is 14.

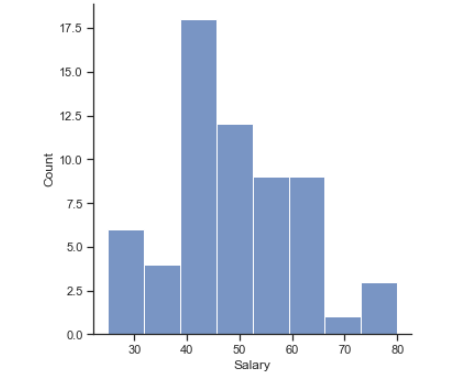
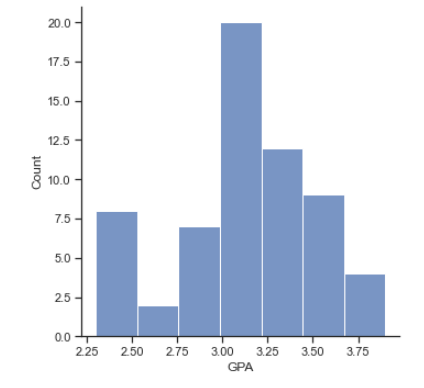
From the contigency table we can find that the total number of female candidates who earn 50 or more is 18.

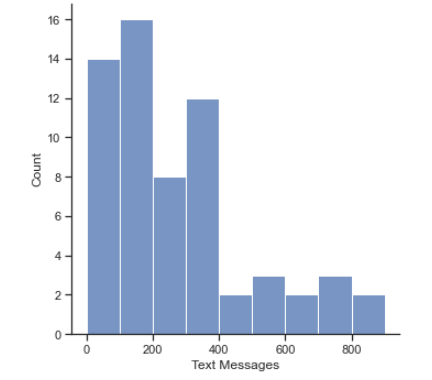
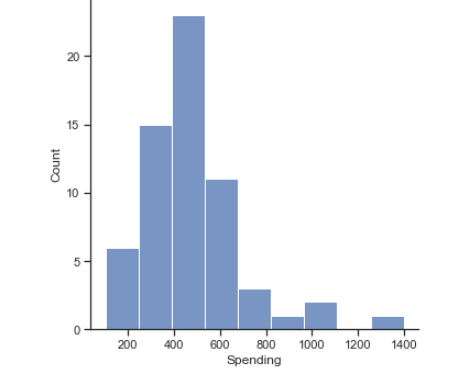
The probability that a randomly selected male earns 50 or more = 14/29 =0.48275862 = 48.27%

The probability that a randomly selected female earns 50 or more = 18/33=0.54545454= 54.5%

## 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

Using displot function in python , we can draw of the graphs for the four numerical variables (GPA, Salary, Spending, Text Messages)





From the graphs it can be concluded that:

* GPA is almost Normally Distributed with a slight skewness toward the left.
* Salary is also almost Normally Distributed with a slight skewness towards the right.
* Spending is not Normally distributed with a high skewness towards the right
* Text message is not Normally distributed with a high skewness towards the right

# 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.

**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.**

**For Sample A**

Step :1 Defining the null and alternate hypothesis

H0 : mean moisture content <=0.35

HA : mean moisture content > 0.35

Step :2 Deciding the level of significance

As the value of the level of significance is not given we can consider its value to be 0.05

Step :3 Identifying the test

Here the Sample size is 35 and the Population Standard Deviation is not known. So we use One Sample T Test here

Step :4 Compute P value

From the calculations in python we have computed the p value to be 0.07477633

Step :5 Conclusion

As the p-value is greater than the level of significance (0.07477633 > 0.05), we have no evidence to reject the null hypothesis. Therefore there is no evidence to reject the claim of moisture content of Shingles A is less than 0.35 pound per 100 square feet at 0.05 significance level.

**For Sample B**

Step :1 Defining the null and alternate hypothesis

H0 : mean moisture content <=0.35

HA : mean moisture content > 0.35

Step :2 Deciding the level of significance

As the value of the level of significance is not given we can consider its value to be 0.05

Step :3 Identifying the test

Here the Sample size is 35 and the Population Standard Deviation is not known. So we use One Sample T Test here

Step :4 Compute P value

From the calculations in python we have computed the p value to be 0.0020904774003191826

Step :5 Conclusion

As the p-value is less than the level of significance (0.0020904774003191826> 0.05), we have evidence to reject the null hypothesis. Therefore there is evidence to reject the claim of moisture content of Shingles B is less than 0.35 pound per 100 square feet at 0.05 significance level.

## 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?

Assumptions to be checked before the test for equality of means is performed:

* The standarad deviation and variance are similar (stated in the exploratory analysis)
* The samples must be normally distributed - The sample size of A is 35 and the sample size of B is 31, which proves that both A and B are normally distributed

Step :1 Defining the null and alternate hypothesis

HO : mean moisture of A is equal to the mean moisture of B

HA : mean moisture of A is not equal to the mean moisture of B

Step :2 Deciding the level of significance

As the value of the level of significance is not given we can consider its value to be 0.05

Step :3 Identifying the test

Here we have two samples and the population standard deviation is not known.

As we are testing the equality between two samples A and B, we use 2 sample T test

Step :4 Compute P value

From the calculations in python we have computed the p value to be 0.2017496571835306

Step :5 Conclusion

Here the pvalue is 0.2017496571835306 which is greater than the level of significance 0.05. Therefore we have no evidence to reject the null hypothesis as the p value is greater than the level of significance

Therefore we can conclude that the mean moisture content of A is equal to the mean moisture content of B