sivasree patnala

Sivasree234@gmail.com

SMDM project rteport

Sivasree patnala

**Contents:**

**Problem 1:……………………………………………………………………………………………….3**

* 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?
  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.
  3. On the basis of a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?
  4. Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.
  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

**Problem 2……………………………………………………………………………………………………6**

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?

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

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.

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

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.

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

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?

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

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?

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?

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.

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.

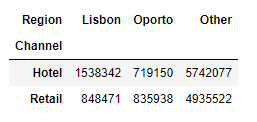
**Problem3………………………………………………………………………………………………….14**

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.

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?

**Problem 1:**

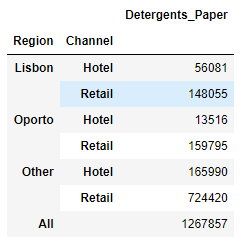
* 1. **Use methods of descriptive statistics to summarize data. Which Region and which Channel spent the most? Which Region and which Channel spent the least?**

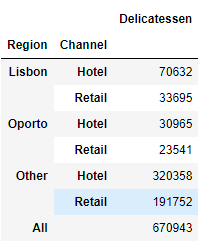


*If u consider the spending,*

* ***Hotels Channel in Other regions spent the most i.e., 5742077.***
* ***Hotels channel in Oporto region spent the least i.e., 719150.***
  1. **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.**

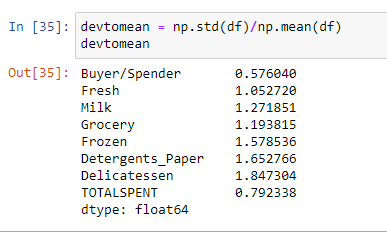
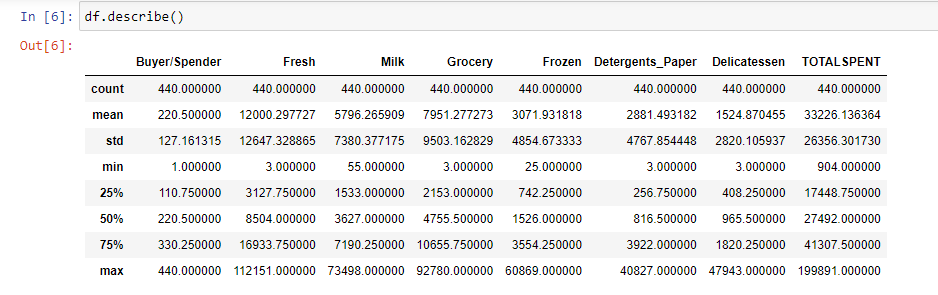






***Inferences:***

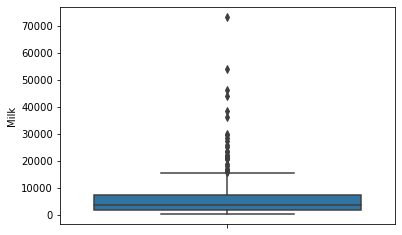
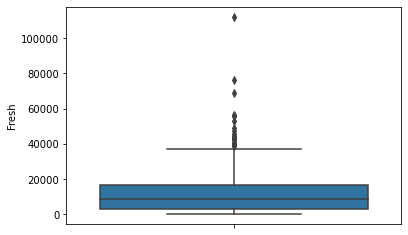
* ***Fresh varieties are most spent on with highest in hotel channel in other region least in Lisbon retail Channel.***
* ***Milk is more spent in Retail channel in Other region least in Oporto region hotel channel.***
* ***Grocery is second highest spend with major part in Retail Channel in Other region least in Oporto region hotel channel.***
* ***Frozen variety is most spent on by the hotels in Other region least in Lisbon retail***
* ***Detergent paper is more spent in Other regions and retail channel and least in Oporto Hotel channel.***
* ***Delicateessen is the least spent among all the varieties. Most in Others regions and hotel channel least in Oporto retail region.***
  1. **On the basis of a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?**

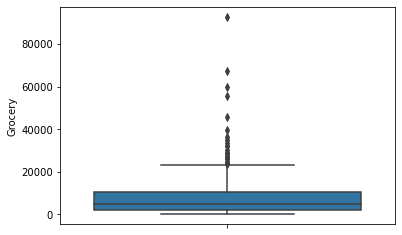
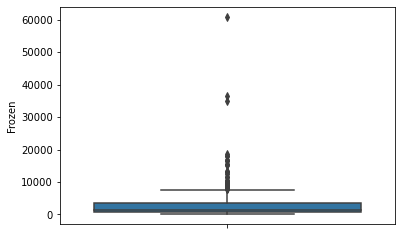
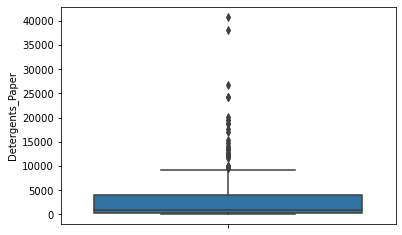
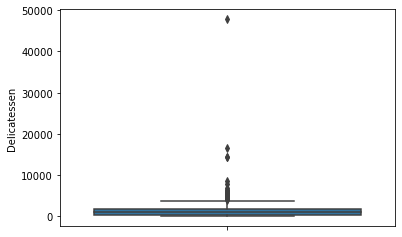


***From the dev to mean data frame we can see the std. dev to mean ratios lower the ratio higher the consistency. There fore least consistent is Delicatessen and most consistent is Fresh.***

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

*Yes, there are outliers by using boxplot as below I can say there are some outliers in the data.*

**

******

***As you can see all the varieties have the outliers in their data.***

* 1. **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**

*For effective business the recommendations I can suggest are:*

1. ***Hotels in Lisbon and Oporto have to spend more on the Fresh varieties rather than pushing Frozen Varieties.***
2. ***Retail channel for detergents\_papers has least spent when compared to the Hotel channel so the retail channel can have better offerings from the whole sale to boost the sales of the product in this Channel.***
3. ***Wholesaler can concentrate on the varieties Fresh, Milk, Grocery, Frozen in the lacking regions and channels as these varieties give maximum spent to reduce the variance and get better sales figures.***

***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*

**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**

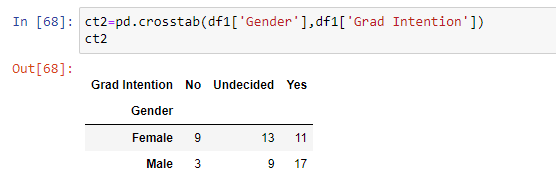
**2.1.3. Gender and Employment**

**2.1.4. Gender and Computer**

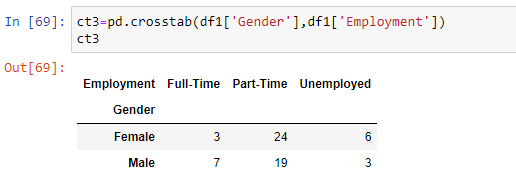
*2.1.1.*



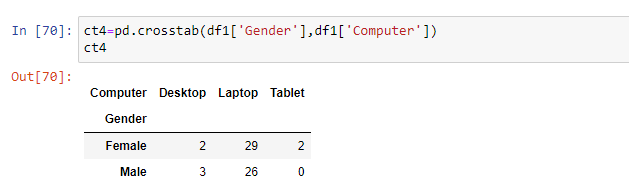
*2.1.2*



*2.1.3*



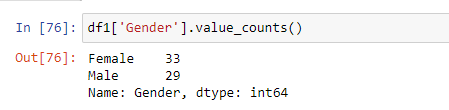
2.1.4



**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?**

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



***2.2.1: As from the Sample we have 33 females and 29 males out of the probability of getting a male when selected randomly is no. of males/total sample= 29//62= 0.460.***

***2.2.2: As from the Sample we have 33 females and 29 males out of the probability of getting a female when selected randomly is no. of females/total sample= 33//62= 0.540.***

**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.**

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

2.3.1:



*From the contingency table we can find the conditional probability for the male students for different majors mentioned as below:*

*Conditional probability for a mail with majors:*

1. *Accounting: males in accounting/total in accounting =* ***4/7***
2. *CIS: males in CIS/total in CIS =* ***¼***
3. *Economics/finance: males in Economics/total in economics = 4****/11***
4. *International Business: males in IB/total in IB =* ***2/6***
5. *Management: males in Management/total in management =* ***6/10***
6. *Other males in Others/total in others =* ***4/7***
7. *Retailing: males in retailing/total in retailing =* ***5/14***
8. *Undecided: males in undecided/total in undecided = 3/3 =* ***1***

*From the contingency table we can find the conditional probability for the female students for different majors mentioned as below:*

*Conditional probability for a mail with majors:*

1. *Accounting: females in accounting/total in accounting =* ***3/7***
2. *CIS: females in CIS/total in CIS =* ***3/4***
3. *Economics/finance: females in Economics/total in economics = 7****/11***
4. *International Business: females in IB/total in IB =* ***4/6***
5. *Management: females in Management/total in management =* ***4/10***
6. *Other: females in Others/total in others =* ***3/7***
7. *Retailing: females in retailing/total in retailing =* ***9/14***
8. *Undecided: females in undecided/total in undecided = 0/3 =* ***0***

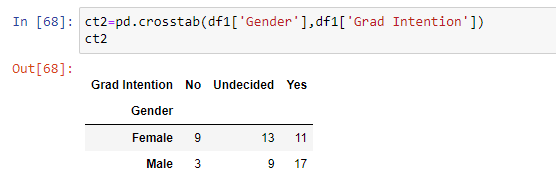
**2.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:**

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

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

*2.4.1:*

*From the contingency table between student gender and Grad intention given below:*



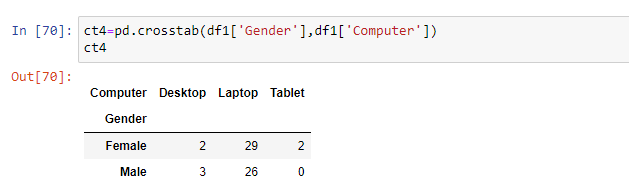
*probability That a randomly chosen student is a male and intends to graduate is given as:*

*P(M/IG) = total males intended to graduate/ total sample*

*= 17/62.*

*2.4.2:*

*From the contingency table between gender and laptop given below:*



*A female doesn’t have a laptop means females having desktop and tablet: i.e., 2+2=4*

*Probability that a randomly selected student is a female and does NOT have a laptop is given as total females with out laptop/total sample = 4/62.*

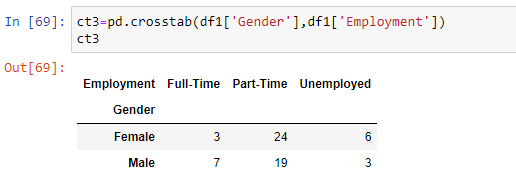
**2.5. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:**

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

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

2.5.1:

*From the contingency table between the gender and employment:*



*Find the probability that a randomly chosen student is a male or has full-time employment is given as:* ***32/62.***

*2.5.2:*

*From the contingency table between gender and majors:*



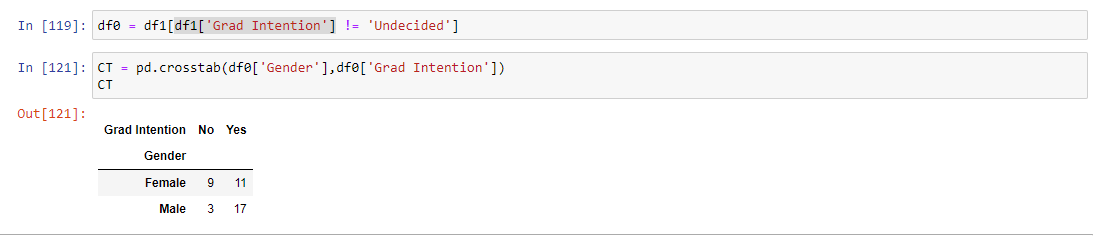
*Conditional probability that given a female student is randomly chosen, she is majoring in international business or management. Is given as:*

Equation for conditional probability.

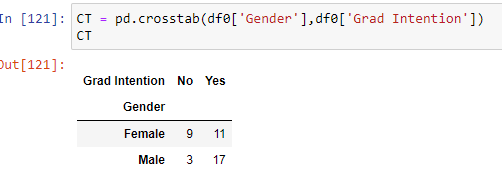
*i.e.,*

*P(IB or M|Female) = P(IB or M ∩ Female)/P(Female)*

*=* ***8/33***

**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?**

*Frome the Contingency table:*



*The graduate intention and being female are independent events are not independent events for as ‘An independent event is an event that has no connection to another event's chances of happening’ therefore the event of being female and grad Intention is not independent because there is not dependency on that.*

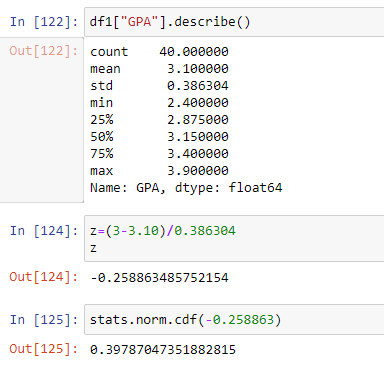
**2.7.1. If a student is chosen randomly, what is the probability that his/her GPA is less than 3?**

**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.**

**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.**

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

*: If we consider the distribution of the student GPA to be normally distributed the probability of the student getting GPA lest than 3 is given as: 0.397 i.e., 39.7%*

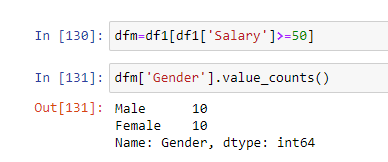


**2.7.2. Find the conditional probability that a randomly selected male earns 50 or more. Find the conditional probability that a randomly selected female earns 50 or more.**

*Conditional probability is given as:*

Equation for conditional probability.

1. conditional probability that a randomly selected male earns 50 or more is



*From the table we can see that there are 10 males and 10 females who earn more than or equal 50.*

*There for probability of earning more is 20/62.*

conditional probability that a randomly selected male earns 50 or more is from conditional probability table is

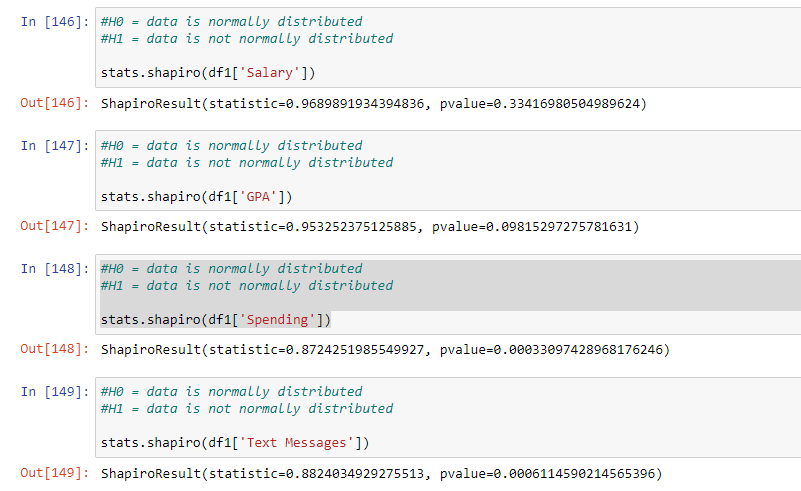
P(Male | earning >=50) = **(10/62)/20/62 = 10/20= 0.5**

1. *Similarly conditional probability that a randomly selected female earns 50 or more is:*

P(FeMale | earning >=50) = **(10/62)/20/62 = 10/20= 0.5**

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.**



* *Salary is normally distributed as the Shapiro test P-Value>0.05*
* *GPA is normally distributed as the Shapiro test P-Value>0.05*
* *Both spending and Test massages are not normally distributed as the Shapiro test P-Value<0.05*

***Problem 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.**

*Step1:*

*Hypothesis formulation: for shingle A*

*H0: Mu1=0.35*

*H1: Mu1 >0.35.*

*Step2:*

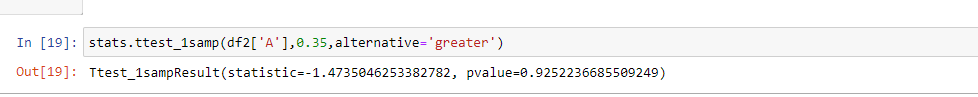
*Significance level(α) = 0.05 (since no value is given):*

*Step 3:*

*We will use the one sample t-test (right tailed) to know if the value is in the permissible limit or not:*

*Step 4:*

*The Test:*



*Since the P value of the t test is 0.92( twosided) for on sided we divide by two i.e. p-value is 0.46 which is greater than 0.05 therefore we fail to reject the null hypothesis.*

*Therefore, we can say that the A shingles are in permissible limit.*

*Similarly, for Shingle B:*

*Step1:*

*Hypothesis formulation: for shingle B*

*H0: Mu1=0.35*

*H1: Mu1 >0.35.*

*Step2:*

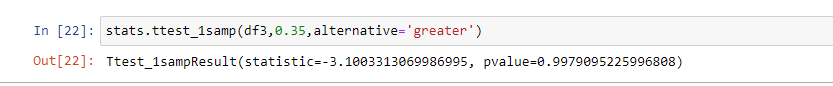
*Significance level(α) = 0.05 (since no value is given):*

*Step 3:*

*We will use the one sample t-test (Right tailed) to know if the value is in the permissible limit or not:*

*Step 4:*

*The Test:*



*Since the P value of the t test is 0.99(two sided) for on sided we divide by two i.e. p-value is 0.49 which is less than 0.05 therefore we fail to reject the null hypothesis.*

*Therefore, we can say that the B shingles are in permissible limit only.*

* 1. **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?**

*Step1:*

*Hypothesis formulation:*

*H0: Mu1=Mu2*

*H1: Mu1 != Mu2*

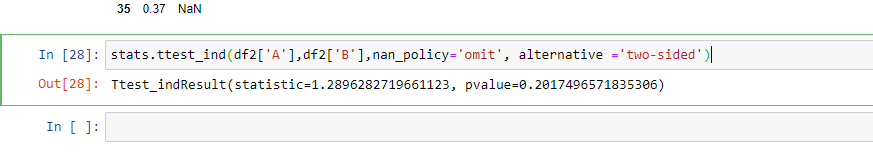
*Step2:*

*Significance level(α) = 0.05 (since no value is given):*

*Step 3:*

*We will use the two-sample t-test since the population mean is not know and there are two samples of unequal number of samples.*

*Step 4:*



*The P value is 0.201 for two sample T test which is greater than the significance value so we fail to reject the null hypothesis therefore both the population means are equal.*

***Assumptions:***

* *The samples are normally distributed.*
* *The significance level is 5% or the confidence interval is 95%*
* *The variances of two samples are equal.*