* 1. **Use methods of descriptive statistics to summarize data. Which Region and which Channel seems to spend more? Which Region and which Channel seems to spend less?**

ANS:- I have used describe() and info() function’s to get the overall descriptive analysis of the data. In which we can see the Mean, median, null, not null, IQR and other details of the data.

* Using box plot found that the number of hotels is more at all locations. This may be because of standard of living of the population.
* Out of the data we can conclude that the Hotel channel at other region tend to spend the most and Hotel channel at Oporto Region observed to spend the least.
  1. **There are 6 different varieties of items are considered. Do all varieties show similar behaviour across Region and Channel?  Provide justification for your answer**

ANS:- I have used describe() function to get all the details about the 6 different varieties. I have observed via pair plot that all the varieties are right skewed. However looking at the std. dev., min and max value we can conclude that all these varieties are not showing similar behaviour across region and channel.

Calculating Coefficient Of Variation: -

Fresh == 0.58

Milk == 1.27

Grocery == 1.19

Frozen == 1.58

Detergents\_Paper == 1.65

Delicatessen == 1.84

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

ANS:- the consistency of a variable and be calculated by ratio of standard deviation by its mean and below is calculation for each variable:-

Fresh == 0.58 – **Least in-consistent**

Milk == 1.27

Grocery == 1.19

Frozen == 1.58

Detergents\_Paper == 1.65

Delicatessen == 1.84 **-- Most In-consistent**

* 1. **Are there any outliers in the data?**

ANS:- Used box plot function and got that all the 6 varieties have outliers in the data.

**1.5 On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem? Answer from the business perspective**

**ANS:- From the descriptive analysis of the data we can see that the business needs to work in introducing more number of retail chains at the locations and should be educated advised to provide better products and services to match the level of hotels at their location.**

**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**

**ANS:- Used pd.crosstab function with gender at row and other parameters at columns.**

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

**ANS:- p(male) = total number of male students/total students= 29/62=0.467**

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

**ANS:- p(male) = 1-p(male)= 1-0.467= 0.532 OR**

**p(female) = total number of female students/total students= 33/62=0.532**

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

**ANS:- We got the counts of each majors bifurcating among male and female students using contingency tables and below calculation is based on the same numbers:-**

**Total Male Students = 29**

**p(accounting|male)= 4/29= 0.138**

**p(CIS|male)= 1/29 = 0.034**

**p(Economics/Finance|male)= 4/29 = 0.138**

**p(International Business|male)= 2/29 = 0.069**

**p(Management|male)= 6/29 = 0.207**

**p(Other|male)= 4/29 = 0.138**

**p(Retailing/marketing|male)= 5/29 = 0.172**

**p(undicided|male)= 3/29 = 0.103**

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

**ANS:- We got the counts of each majors bifurcating among male and female students using contingency tables and below calculation is based on the same numbers:-**

**Total female Students = 33**

**p(accounting|female)= 3/33= 0.091**

**p(CIS|female)= 3/33 = 0.091**

**p(Economics/Finance|female)= 7/33 = 0.212**

**p(International Business|female)= 4/33 = 0.121**

**p(Management|female)= 4/33 = 0.121**

**p(Other|female)= 3/33 = 0.091**

**p(Retailing/femarketing|male)= 9/33 = 0.273**

**p(undicided|female)= 0/33 = 0.000**

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

**ANS:- we got the counts of male and females for their Grad intensions and below is the calculation for probability:-**

**P(male|intend to graduate) = male intend to graduate / total male = 17/29= 0.586**

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

**ANS:- we got the counts of male and females for their deasktop, laptop and tablets and below is the calculation for probability:-**

**P(female|no laptop) = total female – no of female with laptop/ total female = 4/33= 0.121**

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

**ANS:- total number of students = 62**

**Number of male students = 29**

**Number of full time employment students= 10**

**P = 29/62+10/62 = 39/62 = 0.629**

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

**ANS:- total female student count = 33**

**Total female students with international business = 4**

**Total female students with management = 4**

**P = 4+4/33 = 8/33 = 0.242**

**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?**

**ANS:- new total student count = 40**

**P(f) = 20/40 = 0.5**

**P(yes) = 28/40 = 0.7**

**Test for independent events = P(F ∩ Yes) = P(F)P(Yes)**

**P(F ∩ Yes)= 11/40 = 0.275**

**P(F)P(Yes) = 0.5\*0.7 = 0.35**

**P(F ∩ Yes) is not equal to P(F)P(Yes) so the events are dependent.**

**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**

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

**ANS:- Total number of students = 62**

**Total number of students with GPA < 3 = 17**

**P= 17/62 = 0.274**

**2.6.2. 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.**

**ANS:- Total male students = 29**

**Total male students earning > 50 = 14**

**P = 14/29 = 0.483**

**Total female student = 33**

**Total female students earning > 50 = 18**

**P = 18/33 = 0.545**

**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.**

**ANS:- for visual representation I have created histogram for all numerical continuous variable and found that the variables GPA and salary are following a normal distribution however, the variables spending and text messages are not following normal distribution and are right skewed.**

**THE MEAN, MEDIAN AND MODE OF AN NORMAL DISTRIBUTION ARE EQUAL.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | GPA | Salary | Spending | Text Messages |
| Mean | 3.129032 | 48.54839 | 482.0161 | 246.2096774 |
| Median | 3.15 | 50 | 500 | 200 |
| Mode | 3 | 40 | 500 | 300 |

**Write a note summarizing your conclusions**.

**ANS:- we can conclude that the GPA and salary are following normal distribution and independent of gender.**

**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.**

**ANS:-**

**#Ho = Moisture content = 0.35**

**#Ha = Moisture content>0.35**

**Using python I got the below response:**

**ta statistics: [-1.47350463] pa\_value: [0.14955266]**

**considering alpha as 0.5 (as its not mentioned)**

**I have divided p/2 as python automatically gives 2 direction result and we need only 1 directional result.**

**For sample A after performing the 1 sample T-test, found that P value is more than alpha we fail to reject the null hypothesis and conclude that the moisture content of samples are with the permissible limit.**

**However, on conducting the same for sample B, we got the following T and P value:-**

**tb statistics: [-3.10033131] pb\_value: [0.00418095]**

**I have divided p/2 as python automatically gives 2 direction result and we need only 1 directional result.**

**The P value is less than alpha and we reject the null hypothesis. For sample B we can conclude that the moisture content is more than the permissible limit of 0.35.**

**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?**

**ANS:-** **Ho = Xbar of A = Xbar of B**

**Ha = Xbar of A is not equal to XBar of B**

**the sample size of both A and B are not same.**

**I have used 2 sample T test for independent events and below are T and P value of the test:-**

**tstats 1.2896282719661123**

**p value 0.2017496571835306**

**used nan\_policy = 'omit' to handle null values in the data.**

**from the above test we can conclude that P value is more than alpha of 0.05 and we fail to reject Null hypothesis. so the mean of sample A is equal to mean of Sample B.**