**Business Report**

**Project – SMDM**

**MEGHA R**

**Contents**

1 - Wholesale Customer Data Analysis...................................................................................................2 1.1Problem 1.1.............................................................................................................................2 1.2Problem 1.2.............................................................................................................................5 1.3Problem 1.3.............................................................................................................................6 1.4Problem 1.4.............................................................................................................................7 1.5Problem 1.5.............................................................................................................................7

2 - Clear Mountain State University (CMSU) Survey...............................................................................8 2.1Problem 2.1.1..............................................................................................................................8 2.1Problem 2.1.2..............................................................................................................................8 2.1.Problem 2.1.3.............................................................................................................................8 2.1.Problem 2.1.4.............................................................................................................................8 2.2.Problem 2.2.1.............................................................................................................................9 2.2.Problem 2.2.2.............................................................................................................................9 2.3.Problem 2.3.1.............................................................................................................................9 2.3.Problem 2.3.2.............................................................................................................................9

2.4.Problem 2.4.1.............................................................................................................................10 2.4.Problem 2.4.2.............................................................................................................................10

2.5.Problem 2.5.1.............................................................................................................................10 2.5.Problem 2.5.2.............................................................................................................................10

2.6.Problem 2.6................................................................................................................................11 2.7.Problem 2.7.1.............................................................................................................................11

2.7.Problem 2.7.2.............................................................................................................................11

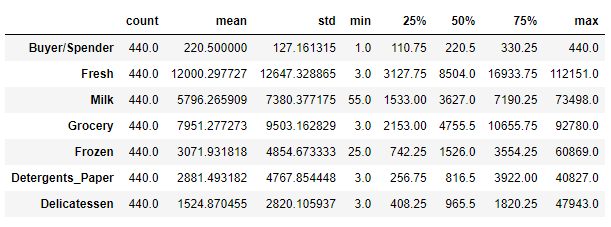
2.8.Problem 2.8...............................................................................................................................12

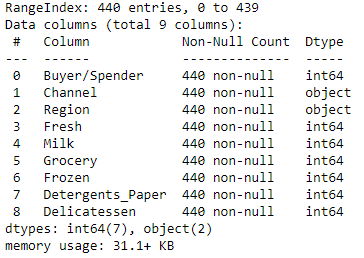
3 - Hypothesis Testing for Quality of Shingles.......................................................................................13 3.1.Problem 3.1.............................................................................................................................13 3.2.Problem 3.2.............................................................................................................................14

***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).

* 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?**



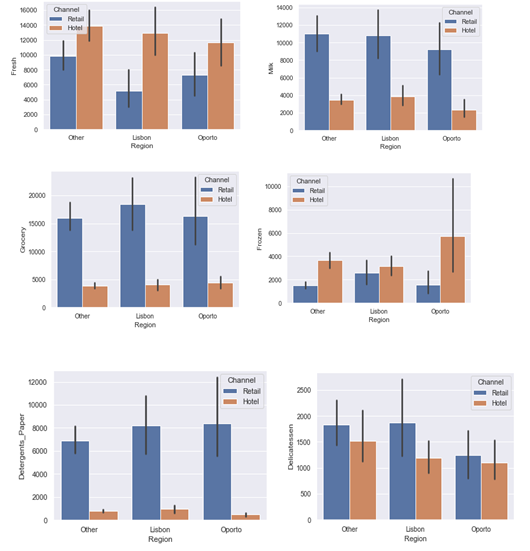


**From the summarized data we can say that,**

* **There are a total of 440 entries and the index range is from 0 to 439.**
* **A total of nine columns are present of which six different varieties of items.**
* **The Channel and Region are of Object datatypes and the rest are of Integer datatypes.**
* We can see that there is no null value and all the data inputs are having numeric datatype.
* **We can see that there are no missing values in the dataset. Thus, there is no need to use the dropna() function.**

Using bar graph with Region and Channel we were able to identify region with maximum spend and minimum spend. Below is the bar graph representation-

Looking at the bar graph, Hotel Channel spends more and Retail spends least.

****

* Hotel channel spend amount is 7974146$ (298 count) with the highest spend amount.
* Retail spend amount 6619931$ (142 count) has least spend amount based on Channel.

Below is the output from Python

Channel

Hotel 7974146

Retail 6619931

dtype: int64

Hotel 298

Retail 142

Name: Channel, dtype: int64

Similarly we grouped totals by region to get totals by region.

* Other regions spend amount is 10652176$ (316 count) with the highest spend amount
* Oporto region spend amount is 1555088$ (47 count) and has least spend amount by Region.

Below is the output from Python

Region

Lisbon 2386813

Oporto 1555088

Other 10652176

dtype: int64

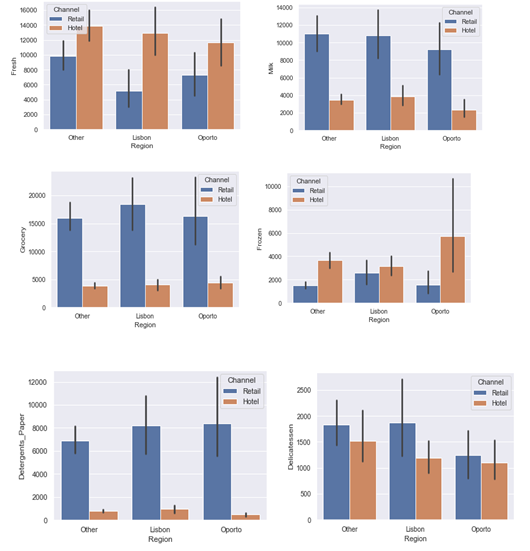
Other 316

Lisbon 77

Oporto 47

Name: Region, dtype: int64

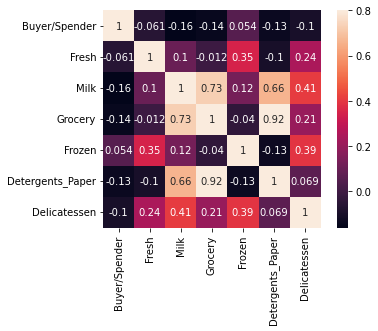
* 1. **There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.**



Looking at the above graph, we see that some categories like Milk, Grocery & Detergents\_Paper have higher spend in the Retail channel versus Hotel, across all regions.

On the other hand, Fresh and Frozen have higher consumption in the Hotel channel versus Retail, across all regions.

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



Looking at the plot above, there are a few pairs of features that exhibit some degree of correlation. They include:

Grocery and Detergents\_Paper are highly correlated - 0.92

Milk and Groceries - 0.73

Milk and Detergents\_Paper - 0.66

Buyer/Spender 1.617000e+04

Fresh 1.599549e+08

Milk 5.446997e+07

Grocery 9.031010e+07

Frozen 2.356785e+07

Detergents\_Paper 2.273244e+07

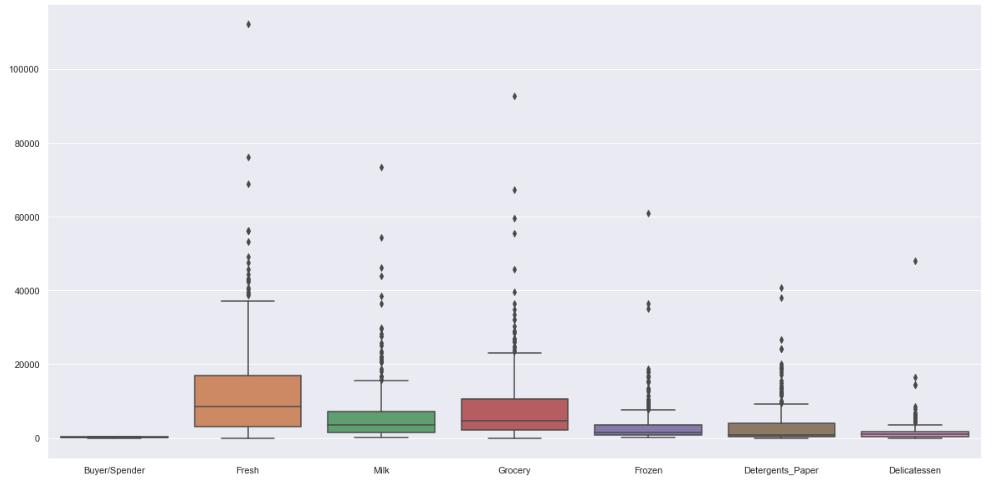
Delicatessen 7.952997e+06

dtype: float64

Using Coefficient of Variation we find out the least value is of Category “Fresh” (1.599) and highest value is of Category “Grocery” (9.031)

So from the given data it is clear that most inconsistent behavior shown by item – Grocery and least inconsistent behavior shown by item – 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 present in the data. To find out outliers we plotted boxplot and the output gives the details that in all the data there are outliers. The data appears to be very skewed and rich in outliers what can negatively affect our analysis.

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 least across region and channel.

**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**

As per the analysis, I find out that there are inconsistencies in spending of different items, which should be minimized. The spending of Hotel and Retail channel are different which should be more or less equal. And also spent should equal for different regions. Need to focus on other items also than “Fresh” and “Grocery”.

***Problem 2***

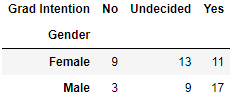
The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. CMSU creates and distributes a survey of 14 questions and receives responses from 62 undergraduates (stored in the **Survey** data set).

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

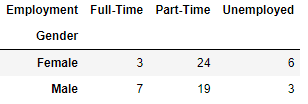
**2.1.1. Gender and Major**



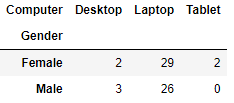
**2.1.2. Gender and Grad Intention**



**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?**

We find out total male students out of whole student from the given data. After calculation we got the result that probability of 46.77% student will be male in CMSU if randomly selected.

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

We find out total female students out of whole student from the given data. After calculation we got the result that probability of 53.23% student will be female in CMSU if randomly selected.

**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.**

Using contingency tables of Gender and Majors we got the total numbers of males and number of males opting for different majors.

Probability of Males opting for Accounting. is 13.79%

Probability of Males opting for CIS. is 3.45%

Probability of Males opting for Economics/Finance. is 13.79%

Probability of Males opting for InternationalBusiness. is 6.90%

Probability of Males opting for Management. is 20.69%

Probability of Males opting for Other. is 13.79%

Probability of Males opting for Retailing/Marketing. is 17.24%

Probability of Males opting for Undecided. is 10.34%

And from this output we can easily say that most of the males students prefer Management as Majors and CIS is the least preferred one.

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

Using contingency tables of Gender and Majors we got the total numbers of females and number of females opting for different majors.

Probability of Females opting for Accounting. is 9.09%

Probability of Females opting for CIS. is 9.09%

Probability of Females opting for Economics/Finance. is 21.21%

Probability of Females opting for InternationalBusiness. is 12.12%

Probability of Females opting for Management. is 12.12%

Probability of Females opting for Other. is 9.09%

Probability of Females opting for Retailing/Marketing. is 27.27%

Probability of Females opting for Undecided. is 0.00%

And from this output we can easily say that most of the females students prefer Retailing/Marketing as Majors.

**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.**

Using contingency tables of Gender and Grad Intention we got the total numbers of males and number of males intends to be graduate.

Probability of Males and intends to be Graduate is 58.62%

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

Using contingency tables of Gender and Computer we got the total numbers of females and number of females does not have a laptop.

Probability of randomly selected student is a Female and does NOT have a laptop is 13.79%

**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?**

Using contingency tables of Gender and Employment we got the total numbers of males and number of males who are full time employed.

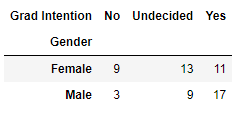
Probability of randomly chosen student is either Male or has full time employment is 74.19%

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

Using contingency tables of Gender and Major we got the total numbers of females and number of females majoring in international business or management.

Probability that given a female student is randomly chosen, she is majoring in international business or management is 24.24%

**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?**



Is the graduate intention and being female are independent events?

The Probability that a randomly selected student ‘being female’ The Probability that a randomly selected student the graduate intention and being female

P(Grad Intention Yes) = 28/40 = 0.7

P(Grad Intention Yes | female) = 11 / 20 = 0.55

These probabilities are not equal. This suggests that the two events are independent

**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?**

Using contingency tables of Gender and GPA we got the total numbers of students and number of students GPA less than 3.

Probability that student is chosen randomly and that his/her GPA is less than 3 is 22.58%

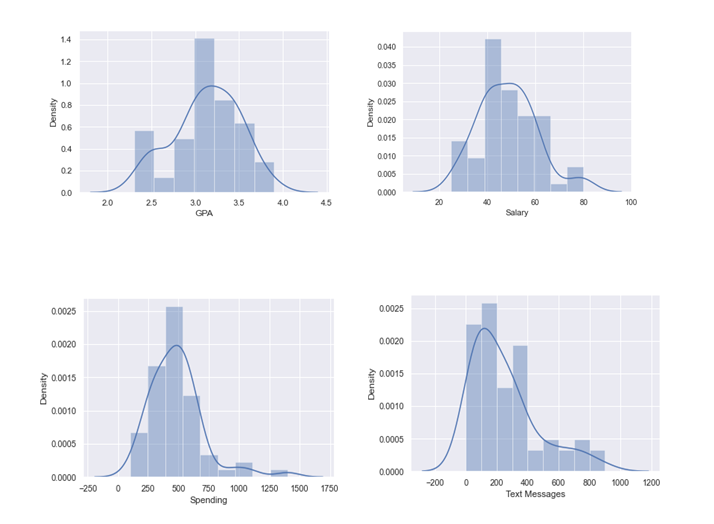
**2.7.2. Find the conditional probability that a randomly selected male earns 50 or more. Find the conditional probability that a randomly selected female earns 50 or more.**

Using contingency tables of Gender and Salary we got the total numbers of Male and Female and number of male and female earning 50 or more.

Probability that randomly selected male earns 50 or more is 34.48%

Probability that randomly selected female earns 50 or more is 30.3%

**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.**



And to confirm whether these four data sets are following normal distribution or not, we done the Shapiro– Wilk test and the output from Python we got – Shapiro results for ID

ShapiroResult(statistic=0.9551758170127869, pvalue=0.023849833756685257)

Shapiro results for Age

ShapiroResult(statistic=0.9281195998191833, pvalue=0.0013503588270395994)

Shapiro results for GPA

ShapiroResult(statistic=0.9685361981391907, pvalue=0.11204058676958084)

Shapiro results for Salary

ShapiroResult(statistic=0.9565856456756592, pvalue=0.028000956401228905)

Shapiro results for Social Networking

ShapiroResult(statistic=0.8165903687477112, pvalue=2.546349548993021e-07)

Shapiro results for Satisfaction

ShapiroResult(statistic=0.8981330990791321, pvalue=8.741713827475905e-05)

Shapiro results for Spending

ShapiroResult(statistic=0.8777452111244202, pvalue=1.6854661225806922e-05)

Shapiro results for Text Messages

ShapiroResult(statistic=0.8594191074371338, pvalue=4.324040673964191e-06)

By these details we confirm that out of the given four data sets ‘GPA’ and ‘Salary’ are following normal distribution whereas other two ‘Spending’ and ‘Text Messages’ are not following the normal distribution.

***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.

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 means moisture contents in both types of shingles are within the permissible limits? State your conclusions clearly showing all steps.**

One sample t test

t statistic: -1.4735046253382782 p value: 0.07477633144907513

Since pvalue > 0.05, do not reject H0 . There is not enough evidence to conclude that the mean moisture content for Sample A shingles is less than 0.35 pounds per 100 square feet. p-value = 0.0748. If the population mean moisture content is in fact no less than 0.35 pounds per 100 square feet, the probability of observing a sample of 36 shingles that will result in a sample mean moisture content of 0.3167 pounds per 100 square feet or less is .0748.

One sample t test

t statistic: -3.1003313069986995 p value: 0.0020904774003191826

Since pvalue < 0.05, reject H0. There is enough evidence to conclude that the mean moisture content for Sample B shingles is not less than 0.35 pounds per 100 square feet. p-value = 0.0021. If the population mean moisture content is in fact no less than 0.35 pounds per 100 square feet, the probability of observing a sample of 31 shingles that will result in a sample mean moisture content of 0.2735 pounds per 100 square feet or less is 0.0021.

**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?**

t\_statistic=1.29 and pvalue=0.202

As the pvalue > α , do not reject H0; and we can say that population mean for shingles A and B are equal Test Assumptions When running a two-sample t-test, the basic assumptions are that the distributions of the two populations are normal, and that the variances of the two distributions are the same.