Group Assignment – SMDM

**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 ([Wholesale Customer.csv](https://olympus.greatlearning.in/courses/10646/files/1288578/download?verifier=VGYKkQ430EEom0EvN7pohPmv6tWcw2hT3mfQdh2G&wrap=1)) 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/Restaurant/Café HoReCa, Retail).

* 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?**

The given data is an information of annual spending of 6 different varieties of products across different Channels (i.e. Hotel and Retail) and Regions (i.e. Lisbon, Oporto, Other)

For the given data an additional column has been added to get information on total spending across different Channels and Regions of all 6 varieties of products.

Further, we have taken out the sum of total spending across different channels (i.e. Hotel and Retail) and the same is visually represented below.

Chart, pie chart

Description automatically generated

fig 1.1.1, spending across different channels.

Similarly, we have taken out the sum of total spending across different Regions (i.e. Lisbon, Oporto and Other) and the same is visually represented below.

Chart, pie chart

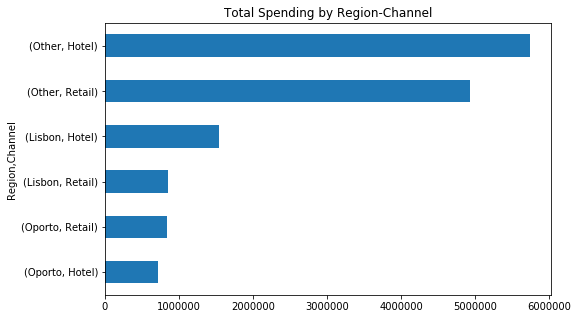
Description automatically generatedBackground pattern

Description automatically generated

fig 1.1.2, spending across different regions.

By grouping together region and channels, we can identify as a whole about which channel in which region is spending more or less.

Below graph represent the sum of spending Region-Channel wise:

****

**Conclusion :**

Based on above presentation it can be clearly said that,

In Channels – Hotels spending is more compared to Retail.

In Regions – Other spending is more compared to Lisbon and Oporto, Further, Oporto spending is least in the data set.

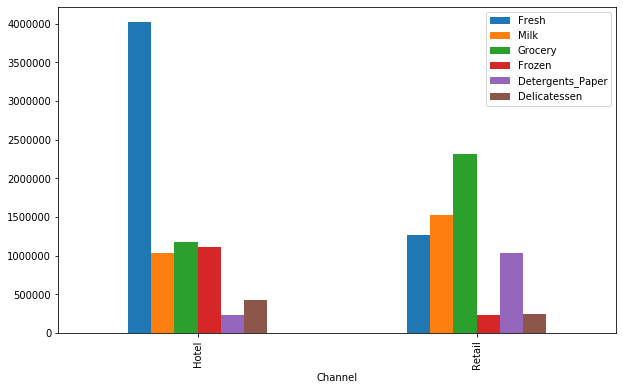
In Region-Channel: Other region and Hotel channel have highest spending, while Oporto region and hotel channel have lowest spending

**1.2. There are 6 different varieties of items are considered.**  
**Do all varieties show similar behaviour across Region and Channel?**

Data set contains spending 6 different varieties of products across different Channels and Regions.

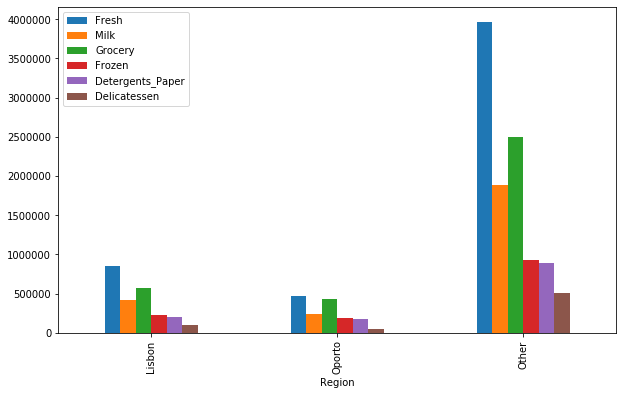
Spending of different varieties of products has been separately taken out from the whole data for different Channels and Regions.

Total spending of each product across different Channel has been considered and the same has been visually represented below.



fig, 1.2.1 spending of different products across channels

Similarly, total spending of each product across different Regions has been considered and the same has been visually represented below.



fig, 1.2.2 spending of different products across regions

**Conclusion:**

Based on above bar graphs,

In Channels – By comparing the spending pattern of each product individually in different channels (i.e. hotel and retail) is different. Hence, behaviour of different products across Channels is not similar to each other.

In Regions - By comparing the spending pattern of each product individually in different regions (i.e. Lisbon, Oporto and other) is similar. Hence, behaviour of different products across Regions is similar to each other.

**1.3. On the basis of the descriptive measure of variability,**

**Which item shows the most inconsistent behaviour?  
Which items shows the least inconsistent behaviour?**

**Solution:**

Mean value of each product vary from each other in the given data set.

To find the consistency we have considered co-efficient of variance as a measure of variability due to the different mean values of each product.

Co-efficient of variation is the ratio of the standard deviation to the mean

Co-efficient of variation (CV)= standard deviation/ mean.

Greater CV indicates more variability and smaller CV indicates less variability.

|  |  |
| --- | --- |
| Product | Co-efficient of variance |
| Delicatessen | 1.849407 |
| Detergents Paper | 1.654647 |
| Frozen | 1.580332 |
| Milk | 1.273299 |
| Grocery | 1.195174 |
| Fresh | 1.053918 |

Based on above table it can be said that ‘Delicatessen’ has the greater CV and ‘Fresh’ has the smaller CV.

**Conclusion:**

On the bases of above table it can be identified that ‘Delicatessen**’**  shows most inconsistent behaviour and ‘Fresh’ shows the most consistent behaviour.

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

**Solution:**

In order to check for the existence of outliers in the data we have visualised the data in the form of boxplots.

Channel wise representation:

Graphical user interface, application

Description automatically generated

Region wise representation:

A picture containing calendar

Description automatically generated

**Conclusion:**

Boxplots represent that, each variable has outliers excluding the Product Fresh in Oporto region.

**1.5 On the basis of this report, what are the recommendations?**

**Recommendations:**

Delicatessen - Marketing and sales strategy can be increased in this variety in Lisbon and Oporto region.

Detergent Paper - Marketing and sales strategy can be increased in this variety in all regions.

Fresh – This variety is consumed in bulk in hotel as there are many outliers, we can work on marketing to increase the sales in retail segment.

Grocery & Detergent Paper - Marketing and sales strategy can be increased in this variety in hotel segment.

**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.csv file).

**Part I**

**2.1. Construct the following contingency tables (Keep Gender as row variable)**

**2.1.1 Contingency Table for Gender and Major:**

| **Major** | **Accounting** | **CIS** | **Economics**  **/Finance** | **International Business** | **Management** | **Other** | **Retailing**  **/Marketing** | **Undecided** | **Total** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Gender** |  |  |  |  |  |  |  |  |  |
| Female | 3 | 3 | 7 | 4 | 4 | 3 | 9 | 0 | 33 |
| Male | 4 | 1 | 4 | 2 | 6 | 4 | 5 | 3 | 29 |
| Total | 7 | 4 | 11 | 6 | 10 | 7 | 14 | 3 | 62 |

**2.1.2. Contingency Table for Gender and Grad Intention:**

| **Grad Intention** | **No** | **Undecided** | **Yes** | **Total** |
| --- | --- | --- | --- | --- |
| **Gender** |  |  |  |  |
| Female | 9 | 13 | 11 | 33 |
| Male | 3 | 9 | 17 | 29 |
| Total | 12 | 22 | 28 | 62 |

**2.1.3. Contingency Table for Gender and Employment:**

| **Employment** | **Full-Time** | **Part-Time** | **Unemployed** | **Total** |
| --- | --- | --- | --- | --- |
| **Gender** |  |  |  |  |
| Female | 3 | 24 | 6 | 33 |
| Male | 7 | 19 | 3 | 29 |
| Total | 10 | 43 | 9 | 62 |

**2.1.4. Contingency Table Gender and Computer:**

| **Computer** | **Desktop** | **Laptop** | **Tablet** | **Total** |
| --- | --- | --- | --- | --- |
| **Gender** |  |  |  |  |
| Female | 2 | 29 | 2 | 33 |
| Male | 3 | 26 | 0 | 29 |
| Total | 5 | 55 | 2 | 62 |

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

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

**Solution:**

**Observations:**

Total number of male = 29

Total number of Female =33

Total number of students= 62

**Calculation Formula:**

Probability of Male/Female = (Number of Male/Female)/Total number of students

**Outcome:**

**Probability that student is Male = 0.47**

**Probability that student is Female = 0.53**

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

**Solution:**

**Observations:**

Values for each major based on gender are identified bycontingency Table for Gender and Major.

**Calculation Formula:**

Probability of Different majors = (number of male/female in a major)/(total number of male/female)

**Outcome:**

**Conditional probability of different majors among the male students:**

|  |  |
| --- | --- |
| **Majors** | **Probability** |
| Accounting | 0.1379 |
| CIS | 0.0345 |
| EF | 0.1379 |
| IB | 0.069 |
| Management | 0.2069 |
| Others | 0.1379 |
| RM | 0.1724 |
| Undecided | 0.1035 |

**Conditional probability of different majors among the female students:**

|  |  |
| --- | --- |
| **Majors** | **Probability** |
| Accounting | 0.0909 |
| CIS | 0.0909 |
| EF | 0.2121 |
| IB | 0.1212 |
| Management | 0.1212 |
| Others | 0.0909 |
| RM | 0.2727 |
| Undecided | 0 |

**2.2.3. Find the conditional probability of intent to graduate, given that the student is a male. Find the conditional probability of intent to graduate, given that the student is a female.**

**Solution:**

**Observations:**

By contingency Table for Gender and Grad Intention, we can identify the number of male/female who has value as yes in the Grad Intention.

**Calculation Formula:**

Probability male/female intent to graduate= (male/female having intent)/(total male/female)

**Outcome:**

|  |  |
| --- | --- |
| **Conditional probability of intent to graduate, given that the student is male** | **0.5862** |
| **Conditional probability of intent to graduate, given that the student is Female** | **0.3333** |

**2.2.4. Find the conditional probability of employment status for the male students as well as for the female students.**

**Solution:**

**Observation:**

By contingency Table for Gender and Employment, indentified employment status of male and female

**Calculation Formula:**

Probability of employment = (number of male/female based on employment)/(total of male/female)

**Outcome:**

**Conditional probability of employment status for Male Students:**

|  |  |
| --- | --- |
| **Fulltime employed and male:** | **0.2414** |
| **Part-time employed and male:** | **0.6552** |
| **Unemployed and male:** | **0.1035** |

**Conditional probability of employment status for Female Students:**

|  |  |
| --- | --- |
| **Fulltime employed and female:** | **0.0909** |
| **Part-time employed and female:** | **0.7273** |
| **Unemployed and female:** | **0.1818** |

**2.2.5. Find the conditional probability of laptop preference among the male students as well as among the female students**.

**Solution:**

**Observations:**

By contingency Table for Gender and Computer, we can identify the number of male/female who has Laptop.

**Calculation:**

Probability of laptop preference = (number of male/female having laptop)/(total number of male/female)

**Outcome**:

|  |  |
| --- | --- |
| **Probability of Laptop preference of Male:** | **0.8966** |
| **Probability of Laptop preference of Female:** | **0.8788** |

**2.3. Based on the above probabilities, do you think that the column variable in each case is independent of Gender? Justify your comment in each case. Part II**

**Solution:**

To check if the column variables are independent of Gender will do Test of Independence.

**Hypothesis for Test of Independence:**

Null Hypothesis (H0): Two variables are independent

Alternate Hypothesis (H1): Two variables are dependent

**Method Used: Chi –Squared test**

In Python that can be done using the chi2\_contingency()scipy Function and check for p-value returned by the function. If the p-value is less than 0.05(for 95%confidence), then will reject the null hypothesis and can say that the variables are dependent.

**Observations:**

**Based on the p-value returned by chi2\_contingency test on contingency table Gender-Major, Gender-Grad Intention, Gender-Employment and Gender-Computer:**

|  |  |
| --- | --- |
| **Column Variables** | **P-value** |
| Gender-Major | 0.42 |
| Gender-Grad intention | 0.09 |
| Gender-Employment | 0.23 |
| Gender-Computer | 0.91 |

**Conclusion: As the p-value is more than 0.05 in all cases, thus with 95% of confidence we can say that, we Failed to reject null hypothesis and the variable columns are independent of Gender.**

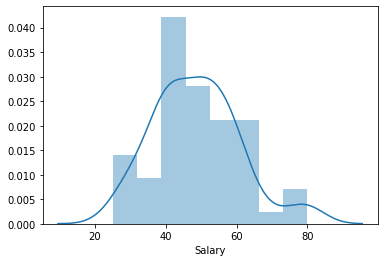
**2.4. Note that there are three numerical (continuous) variables in the data set, Salary, Spending and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarizing your conclusions. [Recall that symmetric histogram does not necessarily mean that the underlying distribution is symmetric]**

**Solution:**

**Method used:**

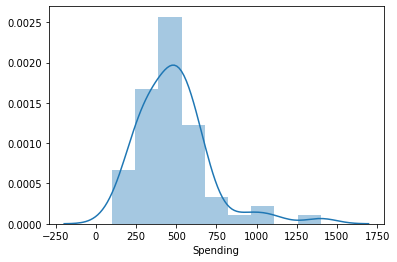
To check if the variable Salary, Spending and Text Message is normally distributed, will go for Shapiro test for normality and Distance Plot.

**Distance plot and p-value returned by Shapiro test for Salary variable:**



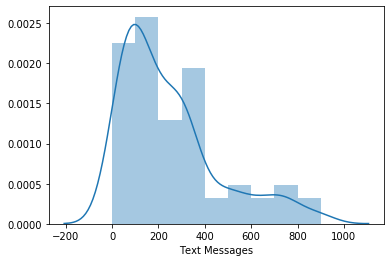
**P-value: 0.028**

**Distance plot and p-value returned by Shapiro test for Spending variable:**

****

**P-value: 1.6854661225806922e-05**

**Distance plot and p-value returned by Shapiro test for Text Messages variable:**



**P-value: 4.324040673964191e-06**

**Conclusion:**

**The p-value from Shapiro test for normality in all cases is less than 0.05 thus we can say that with 95% confidence that the distribution of variables Salary, Spending and Text messages is not normal.**

**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 colouring 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 is calculated. The company claims that the mean moisture content cannot be greater than 0.35 pound per 100 square feet.The file ([A & B shingles.csv](https://olympus.greatlearning.in/courses/10646/files/711871/download?verifier=Af8h0FkzTi9v6yu6PMJT0StHHlqwpiAcqrbqGYkf&wrap=1)) includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.

**3.1 Do you think that the population means 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?**

**Solution:**

**Assumptions:**

**Null Hypothesis (H0): mean of sample A = mean of sample B**

**Alternate Hypothesis (H1): means are not equal**

**Null hypothesis for shapiro test(H0): Sample A & B are normally distributed**

**Alternate hypothesis for shapiro test(H1): Sample A & B are not normally**

**distributed**

**Null hypothesis for levene test(H0): Sample A and Sample B variance are equal**

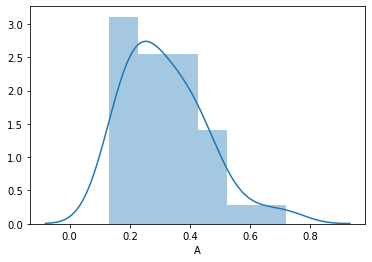
**Alternate hypothesis for levene test(H1): Variance of samples are not equal**

**Outcomes:**

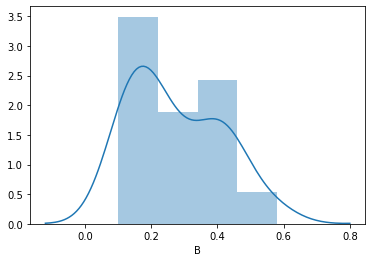
**To check if the samples are normally distributed will conduct Shapiro test and plot Histogram.**

|  |  |
| --- | --- |
| Sample | P value from Shapiro test |
| A | 0.04 |
| B | 0.02 |

Histogram for Sample A:



Histogram for sample B:



**Since the p-value is less than 0.05 in both cases we will reject null hypothesis. Thus we can say that the distribution of sample A and B is not normal.**

**To check if the variance of sample A and B are equal will perform levene test for equality of variance.**

|  |  |
| --- | --- |
| P-value returned by levene | 0.627 |

**Since the p-value is more than 0.05 we fail to reject null hypothesis. Thus we can say that the variance of sample A and B are equal.**

**By Shapiro test we conclude that the distribution is not normal and By levene test we can see that the variances are equal. So based on the variance test we will conduct parametric test for two unpaired samples i.e T-test of independence.**

|  |  |
| --- | --- |
| P-value returned by ttest\_ind | 0.2017 |

**Since the p-value is more than 0.05 we failed to reject null hypothesis and we can say that with 95% of confidence that the mean of sample A and mean of sample B are equal**

**3.2 What assumption about the population distribution is needed in order to conduct the hypothesis tests above?**

**Solution:**

**To conduct the above test we have following assumptions for the population distribution:**

1. **Samples are random and independent.**
2. **We assume that the mean of both the sample A and B are equal.**
3. **We assume both the sample A and B are normally distributed.**
4. **We also assume that the variance of both the sample A and B are equal.**