**SMDM PROJECT REPORT**

**SUBMITTED BY: RAGHAV GUPTA**

**DATE: 03-02-2021**

Table of Contents

[Problem 1 : Wholesale Customers Analysis 6](#_Toc60598225)

[Problem Statement: 6](#_Toc60598226)

[Data Description: 6](#_Toc60598227)

[Domain: 6](#_Toc60598228)

[Context: 6](#_Toc60598229)

[Attribute Information: 6](#_Toc60598230)

[1.1 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? 7](#_Toc60598231)

[Data Summary:- 7](#_Toc60598232)

[Frozen 7](#_Toc60598233)

[REGION WISE SPENDING: 7](#_Toc60598234)

[1.2 There are 6 different varieties of items are considered. Do all varieties show similar behaviour across Region and Channel?  Provide justification for your answer 10](#_Toc60598235)

[1.3 On the basis of a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour? 24](#_Toc60598236)

[1.4 Are there any outliers in the data? 24](#_Toc60598237)

[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 26](#_Toc60598238)

[Problem 2 31](#_Toc60598239)

[Problem Statement: 31](#_Toc60598240)

[Data Description: 31](#_Toc60598241)

[Domain: 31](#_Toc60598242)

[Context: 31](#_Toc60598243)

[Attribute Information: 31](#_Toc60598244)

[2.1. For this data, construct the following contingency tables (Keep Gender as row variable) 32](#_Toc60598245)

[2.1.1. Gender and Major 32](#_Toc60598246)

[2.1.2. Gender and Grad Intention 32](#_Toc60598247)

[2.1.3. Gender and Employment 32](#_Toc60598248)

[2.1.4. Gender and Computer 32](#_Toc60598249)

[2.2. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question: 34](#_Toc60598250)

[2.2.1. What is the probability that a randomly selected CMSU student will be male? 34](#_Toc60598251)

[2.2.2. What is the probability that a randomly selected CMSU student will be female? 34](#_Toc60598252)

[2.3. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question: 34](#_Toc60598253)

[2.3.1. Find the conditional probability of different majors among the male students in CMSU. 34](#_Toc60598254)

[2.3.2 Find the conditional probability of different majors among the female students of CMSU. 35](#_Toc60598255)

[2.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question: 35](#_Toc60598256)

[2.4.1. Find the probability that a randomly chosen student is a male and intends to graduate. 35](#_Toc60598257)

[2.4.2 Find the probability that a randomly selected student is a female and does NOT have a laptop. 35](#_Toc60598258)

[2.5. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question: 36](#_Toc60598259)

[2.5.1. Find the probability that a randomly chosen student is either a male or has full-time employment? 36](#_Toc60598260)

[2.5.2. Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management. 36](#_Toc60598261)

[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? 37](#_Toc60598262)

[2.7. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. 37](#_Toc60598263)

[2.7.1. If a student is chosen randomly, what is the probability that his/her GPA is less than 3? 38](#_Toc60598264)

[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. 38](#_Toc60598265)

[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. 38](#_Toc60598266)

[GPA 39](#_Toc60598267)

[Salary 40](#_Toc60598268)

[Spending 41](#_Toc60598269)

[Text Messages 42](#_Toc60598270)

[Problem 3 44](#_Toc60598271)

[Problem Statement: 44](#_Toc60598272)

[Data Description: 44](#_Toc60598273)

[Domain: 44](#_Toc60598274)

[Context: 44](#_Toc60598275)

[Attribute Information: 44](#_Toc60598276)

[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. 45](#_Toc60598277)

[Hypothesis Testing for Shingle A 45](#_Toc60598278)

[Hypothesis Testing for Shingle B 46](#_Toc60598279)

[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? 47](#_Toc60598280)

[Assumptions for T-Test that we need to check before the test 47](#_Toc60598281)

[Hypothesis Testing 47](#_Toc60598282)

# Problem 1 : ****Wholesale Customers Analysis****

## ****Problem Statement:****

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).

## Data Description:

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).

## Domain:

Retail

## Context:

Leveraging sales information is paramount for customers. In the case of a retail chains, analysis of products of retail chaing sold across different region under different sales channel are very crucial in making business decisions. Hence, knowing to explore and generate value out of such data can be an invaluable skill to have.

## Attribute Information:

Buyer/Spender

Channel

Region

Fresh

Milk

Grocery

Frozen

Detergents\_Paper

Delicatessen

## ****1.1 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?****

### ****Data Summary:-****

**Summary statistics of data is as shown below :-**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **Std** | **min** | **25%** | **50%** | **75%** | **max** |
| **Buyer/Spender** | 440.0 | 220.50 | 127.16 | 1.0 | 110.75 | 220.5 | 330.25 | 440.0 |
| **Fresh** | 440.0 | 12000.30 | 12647.33 | 3.0 | 3127.75 | 8504.0 | 16933.75 | 112151.0 |
| **Milk** | 440.0 | 5796.27 | 7380.38 | 55.0 | 1533.00 | 3627.0 | 7190.25 | 73498.0 |
| **Grocery** | 440.0 | 7951.28 | 9503.16 | 3.0 | 2153.00 | 4755.5 | 10655.75 | 92780.0 |
| **Frozen** | 440.0 | 3071.93 | 4854.67 | 25.0 | 742.25 | 1526.0 | 3554.25 | 60869.0 |
| **Detergents\_Paper** | 440.0 | 2881.49 | 4767.85 | 3.0 | 256.75 | 816.5 | 3922.00 | 40827.0 |
| **Delicatessen** | 440.0 | 1524.87 | 2820.11 | 3.0 | 408.25 | 965.5 | 1820.25 | 47943.0 |

**Findings from the summary statistics are:-**

* **First column which is just a reference of Buyer/Spender no which can be ignored for analysis**
* **Maximum values of all the column attributes are high as compared to the median value. Hence there seems to be many outliers in this data.**
* **On checking the median values (50%), it appears that retailers spend more on Fresh**

**products and grocery products as compared to others.**

* **75% of 440 retailers spend only 1820 or less annually on Delicatessen. So annual spend of Delicatessen appears to be least among all.**

### ****REGION WISE SPENDING:****

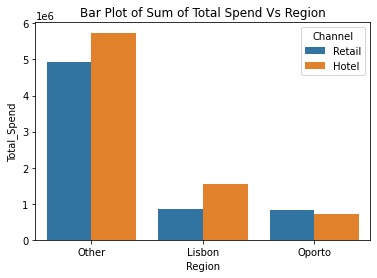
#### ****Pivot Table showing Region wise spending :-****

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Region** | **Buyer/Spender** | **Delicatessen** | **Detergents\_Paper** | **Fresh** | **Frozen** | **Grocery** | **Milk** | **Total\_Spend** |
| **Other** | 64026 | 512110 | 890410 | 3960577 | 930492 | 2495251 | 1888759 | 10677599 |
| **Lisbon** | 18095 | 104327 | 204136 | 854833 | 231026 | 570037 | 422454 | 2386813 |
| **Oporto** | 14899 | 54506 | 173311 | 464721 | 190132 | 433274 | 239144 | 1555088 |

**We can see from the above Pivot Table that 'Other' region is the highest spender. It seems like the obvious one because as seen in our EDA, 75% data is coming from Other category. Further we can see that lowest spenders are in 'Opporto' region.**

#### Bar Plot of Total Spend Vs Region

Bar Plot showing Region wise spending pattern for Total Spend is as shown below :



### ****CHANNEL WISE SPENDING:****

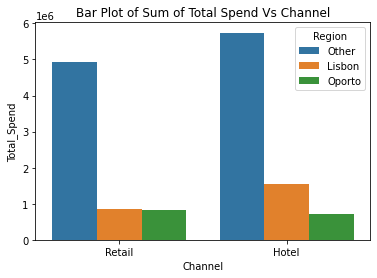
#### ****Pivot Table showing Channel wise spending :-****

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Channel** | **Buyer/Spender** | **Delicatessen** | **Detergents\_Paper** | **Fresh** | **Frozen** | **Grocery** | **Milk** | **Total\_Spend** |
| **Hotel** | 71034 | 421955 | 235587 | 4015717 | 1116979 | 1180717 | 1028614 | 7999569 |
| **Retail** | 25986 | 248988 | 1032270 | 1264414 | 234671 | 2317845 | 1521743 | 6619931 |

**We can see that spending is highest under 'Hotel' channel as compared to 'Retail' channel. Visually it can be seen under the Bar plot and Donut Charts. These charts are shown below.**

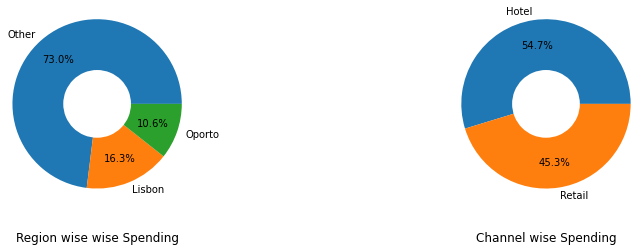
#### Bar Plot of Total Spend Vs Channel

Bar plot showing Channel wise spending pattern for Total Spend is as shown below :-

****

### ****Donut Chart showing region wise and channel wise spending****

**Donut showing spending pattern region wise and channel wise is shown below:-**

****

## ****1.2 There are 6 different varieties of items are considered. Do all varieties show similar behaviour across Region and Channel?  Provide justification for your answer****

Now for checking the behaviour of varieties across region/channel, I have compared the 5 point summary along with Coefficient of Variation and Skewness of each variety across 3 regions and 2 channels. Visually I have created a boxplot and swarmplot charts to see the distribution pattern across region and channels. Coding details are present in the notebook file attached along with this report.

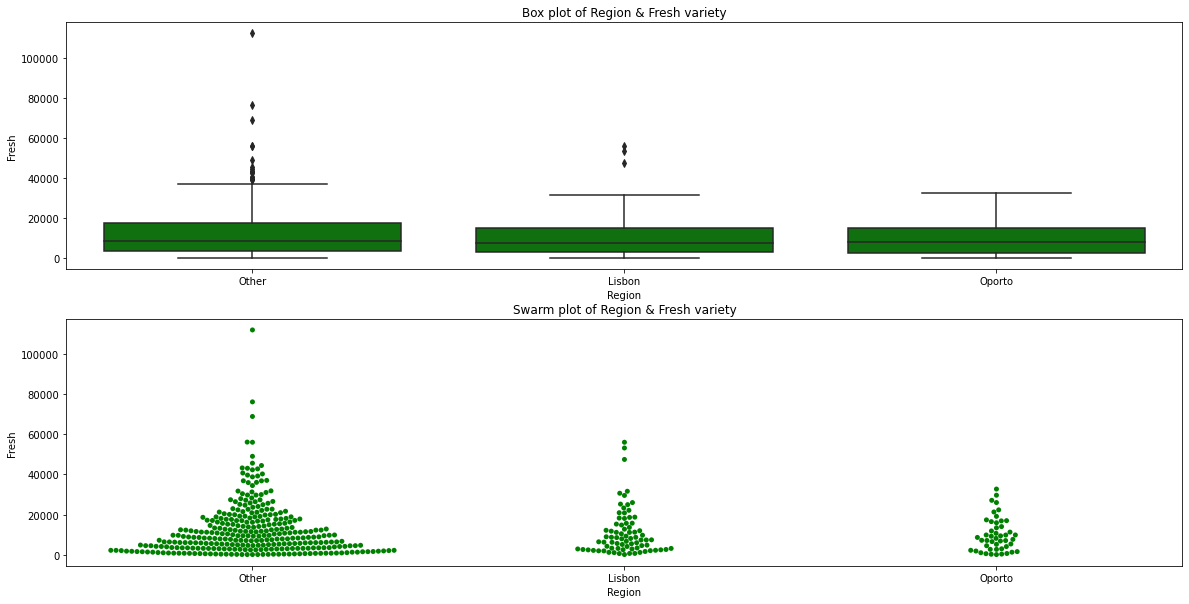
### Varieties across Regions

#### Fresh Variety behaviour across all three regions

##### 5 points summary , Coefficient of Variation and Skewness.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Fresh\_Other** | **Fresh\_Oporto** | **Fresh\_Lisbon** |
| **Count** | 316 | 47 | 77 |
| **Mean** | 12533.47 | 9887.68 | 11101.73 |
| **Std** | 13389.21 | 8387.9 | 11557.44 |
| **Min** | 3 | 3 | 18 |
| **25%** | 3350.75 | 2751.5 | 2806 |
| **50%** | 8752.5 | 8090 | 7363 |
| **75%** | 17406.5 | 14925.5 | 15218 |
| **Max** | 112151 | 32717 | 56083 |
| **CV** | 1.07 | 0.84 | 1.03 |
| **Skew** | 2.62 | 0.98 | 2.01 |
|  |  |  |  |

##### Box Plot and Swarm Plot

****

**Key points** as seen in the summary statistics and swarm plot/box plot,

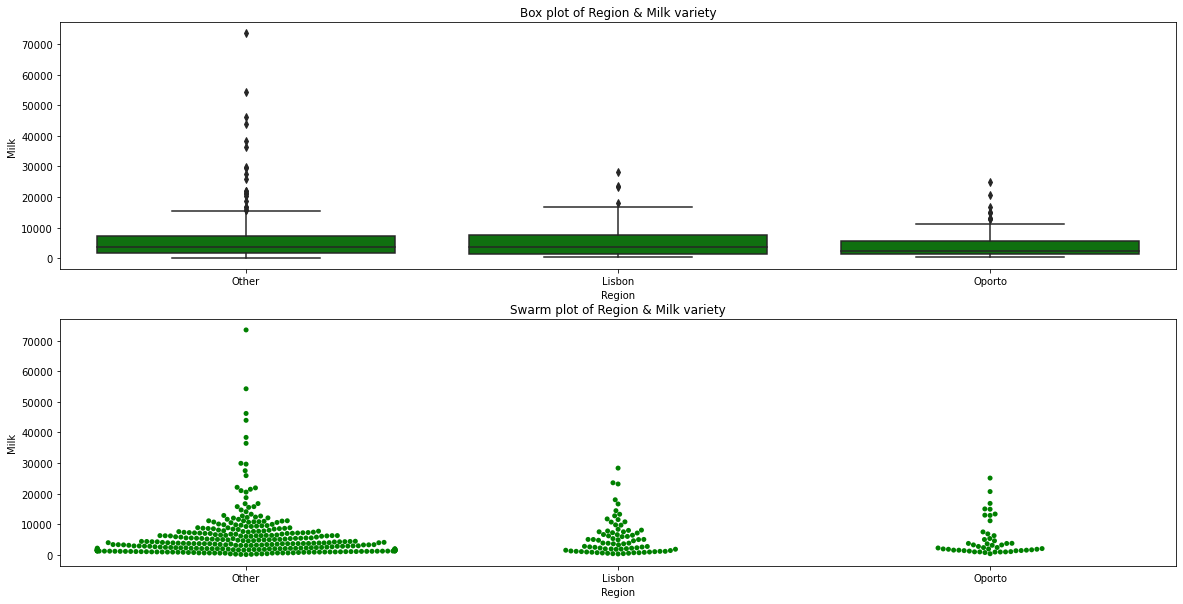
* Data contains more retailers in Other region as compared to Lisbon and Oporto.
* Except Oporto region, 'Other' and 'Lisbon' region Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
* Maximum annual spending in 'Other' region is very high as compared to other regions.
* Annual median spend of Other region is slightly higher(8752) than of Lisbon(7363) and Oporto region(8090)
* Although mean and median value of 'Other' region is highest but its volatility is also high i.e. it is the most inconsistent region for Fresh Variety
* Spread of data looks similar across all regions with distribution being right/positive skewed and 75% of retailers spending less than 17.5K annually across all three regions.
* Footfall is more for buyers under 'Other' Region and majority of buyers(75%) are spending less than around 8K across all three regions.

#### Milk Variety behaviour across all three regions

##### 5 points summary , Coefficient of Variation and Skewness

|  |  |  |  |
| --- | --- | --- | --- |
|  | Milk\_Other | Milk\_Oporto | Milk\_Lisbon |
| **Count** | 316 | 47 | 77 |
| **Mean** | 5977.09 | 5088.17 | 5486.42 |
| **Std** | 7935.46 | 5826.34 | 5704.86 |
| **Min** | 55 | 333 | 258 |
| **25%** | 1634 | 1430.5 | 1372 |
| **50%** | 3684.5 | 2374 | 3748 |
| **75%** | 7198.75 | 5772.5 | 7503 |
| **Max** | 73498 | 25071 | 28326 |
| **CV** | 1.33 | 1.13 | 1.03 |
| **Skew** | 4.25 | 1.8 | 1.92 |

##### Box Plot and Swarm Plot

****

**Key Points** as seen in the summary statistics and swarm plot/box plot,

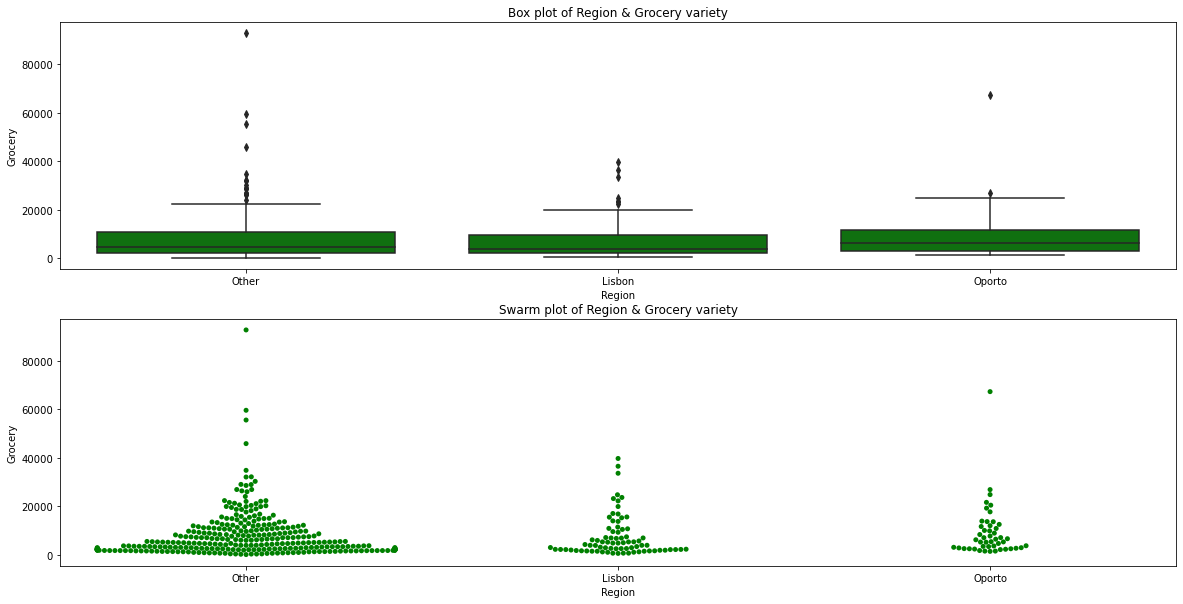
* Data contains more retailers in Other region as compared to Lisbon and Oporto.
* All three regions Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
* Maximum annual spending in 'Other' region is very high as compared to Lisbon/Oporto regions.
* Minimum annual spending in 'Other' region is low as compared to other Lisbon/Oporto regions.
* Annual median spend of Oporto region is slightly lower(2374) than of Lisbon(3748) and Other region(3684)
* Volatility of other region is highest among all i.e. it is the most inconsistent region for Milk Variety
* Spread of data looks similar across all regions with distribution being right/positive skewed and 75% of retailers spending less than 7.5K annually across all three regions.
* Footfall is more for buyers under 'Other' Region and majority of buyers(75%) are spending less than around 3K across all three regions.

#### Grocery Variety behaviour across all three regions

##### 5 points summary , Coefficient of Variation and Skewness

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Grocery\_Other** | **Grocery\_Oporto** | **Grocery\_Lisbon** |
| **Count** | 316 | 47 | 77 |
| **Mean** | 7896.36 | 9218.6 | 7403.08 |
| **Std** | 9537.29 | 10842.75 | 8496.29 |
| **Min** | 3 | 1330 | 489 |
| **25%** | 2141.5 | 2792.5 | 2046 |
| **50%** | 4732 | 6114 | 3838 |
| **75%** | 10559.75 | 11758.5 | 9490 |
| **Max** | 92780 | 67298 | 39694 |
| **CV** | 1.21 | 1.16 | 1.14 |
| **Skew** | 3.84 | 3.64 | 2.02 |

##### Box Plot and Swarm Plot

****

Key Points as seen in the summary statistics and swarm plot/box plot,

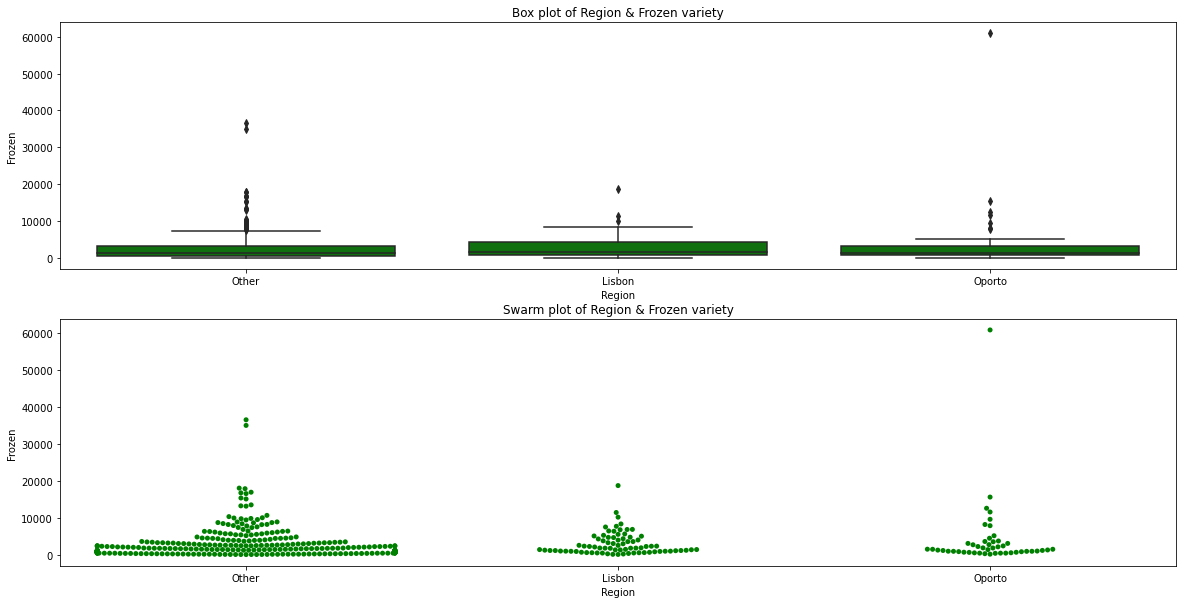
* + Data contains more retailers in Other region as compared to Lisbon and Oporto.
  + All three regions Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
  + Maximum annual spending in 'Other' region is very high as compared to Lisbon/Oporto regions.
  + Minimum annual spending in 'Other' region is low as compared to other Lisbon/Oporto regions.
  + Annual median spend of Oporto region is highest(6114) as compared to Lisbon(3838) and Other region(4732)
  + Volatility of other region is highest among all i.e. it is the most inconsistent region.
  + Spread of data looks similar across all regions with distribution being right/positive skewed and 75% of retailers spending less than 11.7K annually across all three regions.
  + Footfall is more for buyers under 'Other' Region and majority of buyers(75%) are spending less than around 10-11K across all three regions.

#### Frozen Variety behaviour across all three regions

##### 5 points summary , Coefficient of Variation and Skewness

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Frozen\_Other** | **Frozen\_Oporto** | **Frozen\_Lisbon** |
| **Count** | 316 | 47 | 77 |
| **Mean** | 2944.59 | 4045.36 | 3000.34 |
| **Std** | 4260.13 | 9151.78 | 3092.14 |
| **Min** | 25 | 131 | 61 |
| **25%** | 664.75 | 811.5 | 950 |
| **50%** | 1498 | 1455 | 1801 |
| **75%** | 3354.75 | 3272 | 4324 |
| **Max** | 36534 | 60869 | 18711 |
| **CV** | 1.44 | 2.24 | 1.02 |
| **Skew** | 3.96 | 5.49 | 2.33 |

##### Box Plot and Swarm Plot



**Key Points** as seen in the summary statistics and swarm plot/box plot,

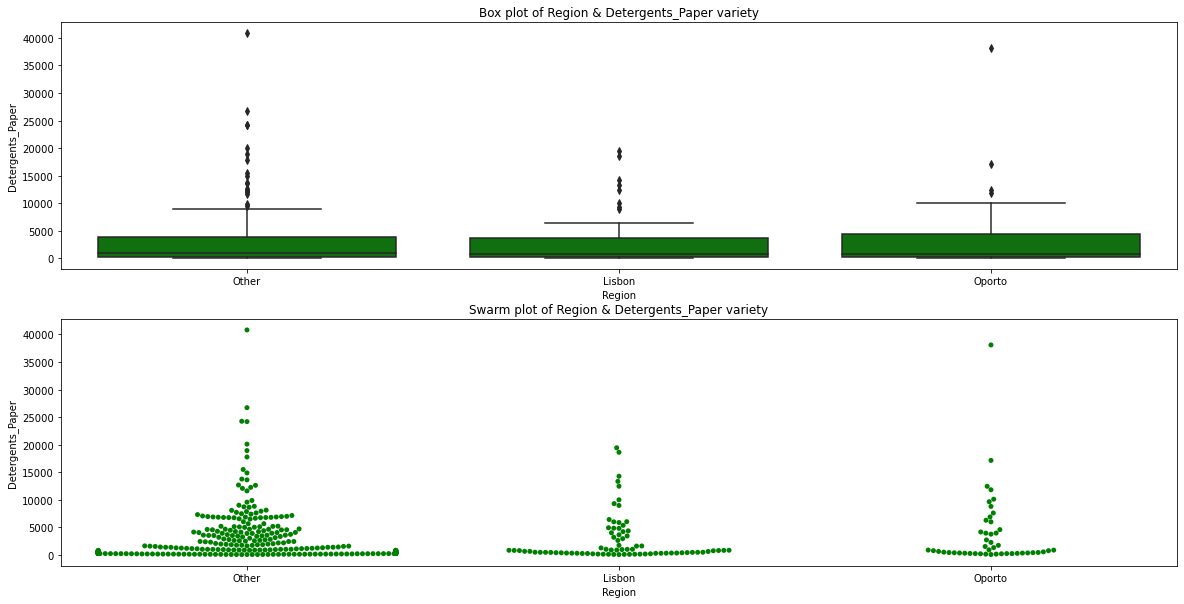
* Data contains more retailers in Other region as compared to Lisbon and Oporto.
* All three regions Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
* Maximum annual spending in 'Oporto' region is very high as compared to Lisbon/Other regions.
* Minimum annual spending in 'Other' region is low as compared to other Lisbon/Oporto regions.
* Annual median spend of Lisbon region is highest(1801) as compared to Oporto(1455) and Other region(1498)
* Volatility of 'Oporto' region is highest among all i.e. it is the most inconsistent region.
* Spread of data looks similar across all regions with distribution being right/positive skewed and 75% of retailers spending less than 4.3K annually across all three regions.
* Footfall is more for buyers under 'Other' Region and majority of buyers(75%) are spending less than around 3-4K across all three regions.

#### Detergents\_Paper Variety behaviour across all three regions

##### 5 points summary , Coefficient of Variation and Skewness

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Detergents\_Paper\_Other** | **Detergents\_Paper\_Oporto** | **Detergents\_Paper\_Lisbon** |
| **Count** | 316 | 47 | 77 |
| **Mean** | 2817.75 | 3687.47 | 2651.12 |
| **Std** | 4593.05 | 6514.72 | 4208.46 |
| **Min** | 3 | 15 | 5 |
| **25%** | 251.25 | 282.5 | 284 |
| **50%** | 856 | 811 | 737 |
| **75%** | 3875.75 | 4324.5 | 3593 |
| **Max** | 40827 | 38102 | 19410 |
| **CV** | 1.63 | 1.75 | 1.58 |
| **Skew** | 3.71 | 3.62 | 2.36 |

##### Box Plot and Swarm Plot



**Key Points** as seen in the summary statistics and swarm plot/box plot,

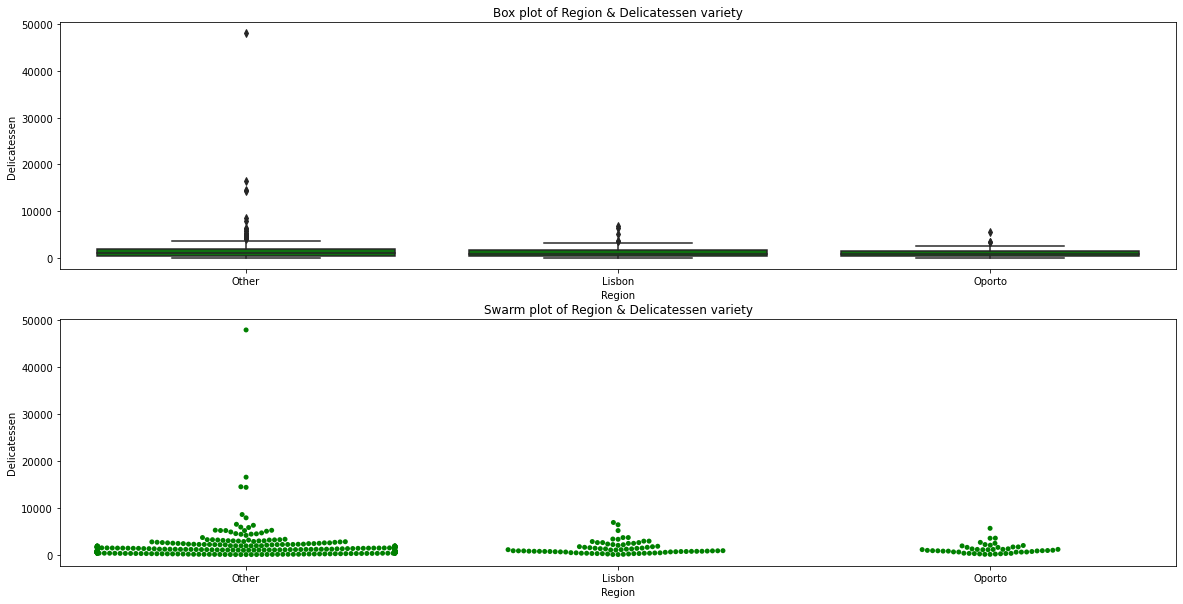
* Data contains more retailers in Other region as compared to Lisbon and Oporto.
* All three regions Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
* Maximum annual spending in 'Other' region is very high as compared to Lisbon/Oporto regions.
* Minimum annual spending in 'Other' region is low as compared to other Lisbon/Oporto regions.
* Annual median spend of 'Other' region is highest(856) as compared to Oporto(811) and Lisbon region(737)
* Volatility of 'Oporto' region is highest among all i.e. it is the most inconsistent region.
* Spread of data looks similar across all regions with distribution being right/positive skewed and 75% of retailers spending less than 4.3K annually across all three regions.
* Footfall is more for buyers under 'Other' Region and majority of buyers(75%) are spending less than 3-4K annually across all three regions.

#### Delicatessen Variety behaviour across all three regions

##### 5 points summary , Coefficient of Variation and Skewness

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Delicatessen\_Other** | **Delicatessen\_Oporto** | **Delicatessen\_Lisbon** |
| **count** | 316 | 47 | 77 |
| **mean** | 1620.6 | 1159.7 | 1354.9 |
| **Std** | 3232.58 | 1050.74 | 1345.42 |
| **Min** | 3 | 51 | 7 |
| **25%** | 402 | 540.5 | 548 |
| **50%** | 994 | 898 | 806 |
| **75%** | 1832.75 | 1538.5 | 1775 |
| **Max** | 47943 | 5609 | 6854 |
| **CV** | 1.99 | 0.9 | 0.99 |
| **Skew** | 10.21 | 2.15 | 2.05 |

##### Box Plot and Swarm Plot

****

**Key Points** as seen in the summary statistics and swarm plot/box plot,

* **Data contains more retailers in Other region as compared to Lisbon and Oporto.**
* **All three regions Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.**
* **Maximum annual spending in 'Other' region is very high as compared to Lisbon/Oporto regions.**
* **Minimum annual spending in 'Other' region is low as compared to other Lisbon/Oporto regions.**
* **Annual median spend of 'Other' region is highest(994) as compared to Oporto(898) and Lisbon region(806)**
* **Volatility of 'Other' region is highest among all i.e. it is the most inconsistent region.**
* **Spread of data looks similar across all regions with distribution being right/positive skewed and 75% of retailers spending less than 1.83K annually across all three regions.**
* **Footfall is more for buyers under 'Other' Region and majority of buyers(75%) are spending less than around 1.8K across all three regions.**

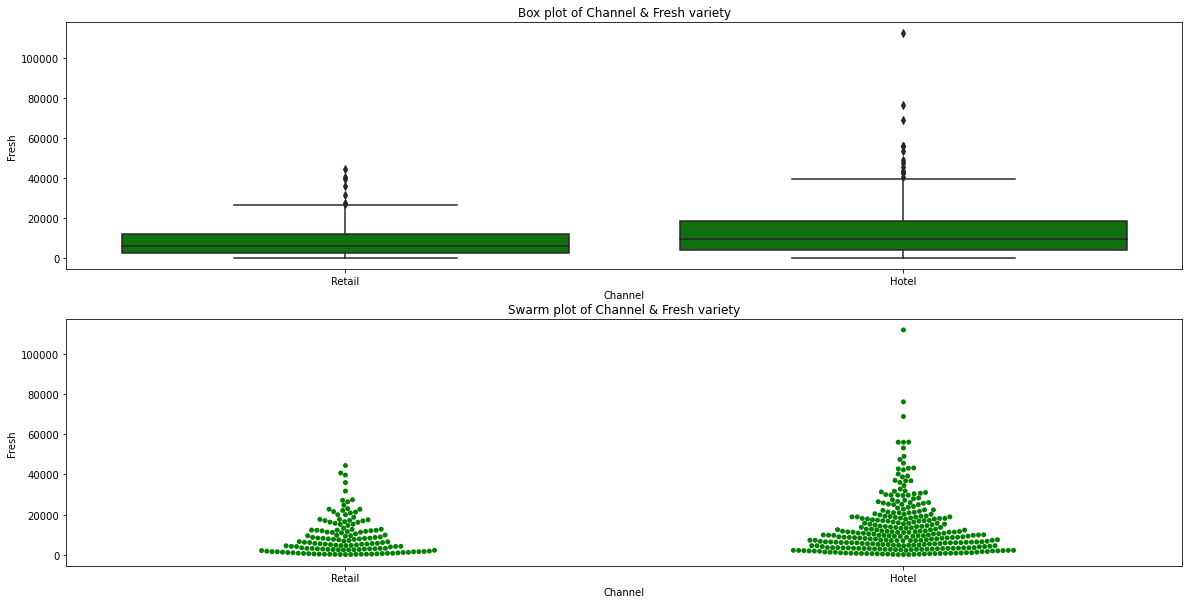
### Varieties across Channels

#### Fresh Variety behaviour across all 2 channels

##### 5 points summary , Coefficient of Variation and Skewness.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | **Fresh\_Hotel** | | **Fresh\_Retail** | |
| **Count** | | 298 | | 142 | |
| **Mean** | | 13475.56 | | 8904.32 | |
| **Std** | | 13831.69 | | 8987.71 | |
| **Min** | | 3 | | 18 | |
| **25%** | | 4070.25 | | 2347.75 | |
| **50%** | | 9581.5 | | 5993.5 | |
| **75%** | | 18274.75 | | 12229.75 | |
| **Max** | | 112151 | | 44466 | |
| **CV** | | 1.02 | | 1.01 | |
| **Skew** | | 2.51 | | 1.59 | |
|  | |  | |  | |  | |

##### Box Plot and Swarm Plot

****

**Key points** as seen in the summary statistics and swarm plot/box plot,

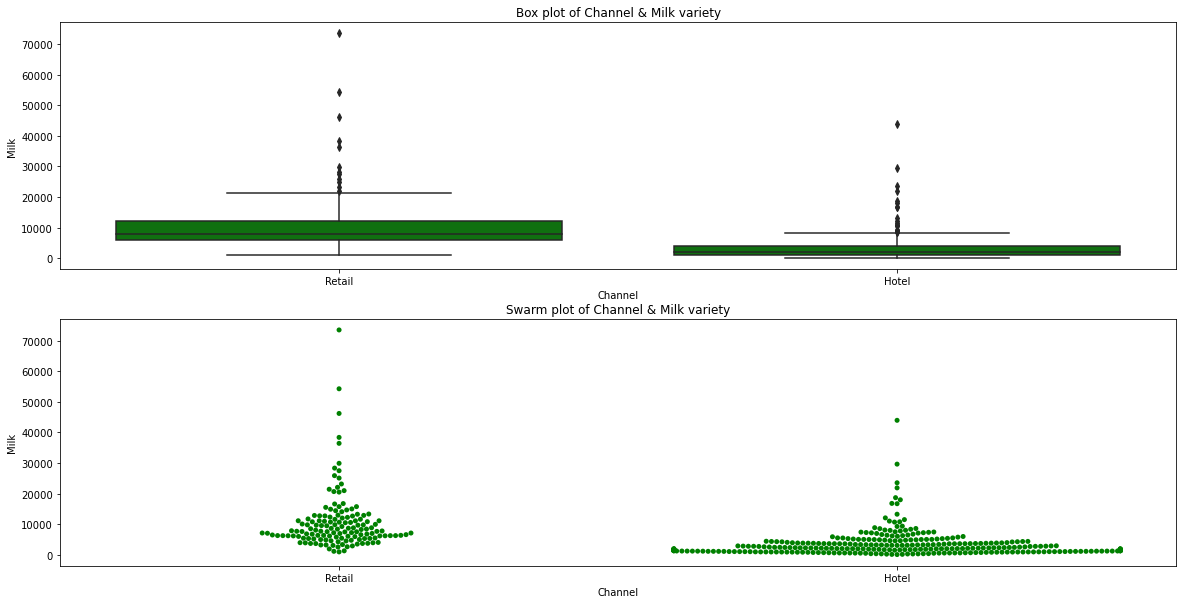
* Data contains more retailers in Hotel channel as compared to Retail channel
* Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
* Maximum annual spending in 'Hotel' Channel is high as compared to Retail Channel
* Min annual spending in 'Hotel' Channel is low as compared to Retail Channel
* Annual median spend of Hotel channel is higher(9581) than of Retail channel(5993)
* Volatility of both the channels is similar.
* Spread of data looks similar across both channels with distribution being right/positive skewed and 75% of retailers spending less than 18K annually on Hotel channel and 12.2K on Retail channel respectively.
* Footfall is more for buyers of Fresh variety under Hotel channel and buyers are spending much more money on Milk variety under Hotel channel.

#### Milk Variety behaviour across all channels

##### 5 points summary , Coefficient of Variation and Skewness

|  |  |  |
| --- | --- | --- |
|  | **Milk\_Hotel** | **Milk\_Retail** |
| **Count** | 298 | 142 |
| **Mean** | 3451.72 | 10716.5 |
| **Std** | 4352.17 | 9679.63 |
| **Min** | 55 | 928 |
| **25%** | 1164.5 | 5938 |
| **50%** | 2157 | 7812 |
| **75%** | 4029.5 | 12162.75 |
| **Max** | 43950 | 73498 |
| **CV** | 1.26 | 0.9 |
| **Skew** | 4.66 | 3.41 |

##### Box Plot and Swarm Plot

****

**Key Points** as seen in the summary statistics and swarm plot/box plot,

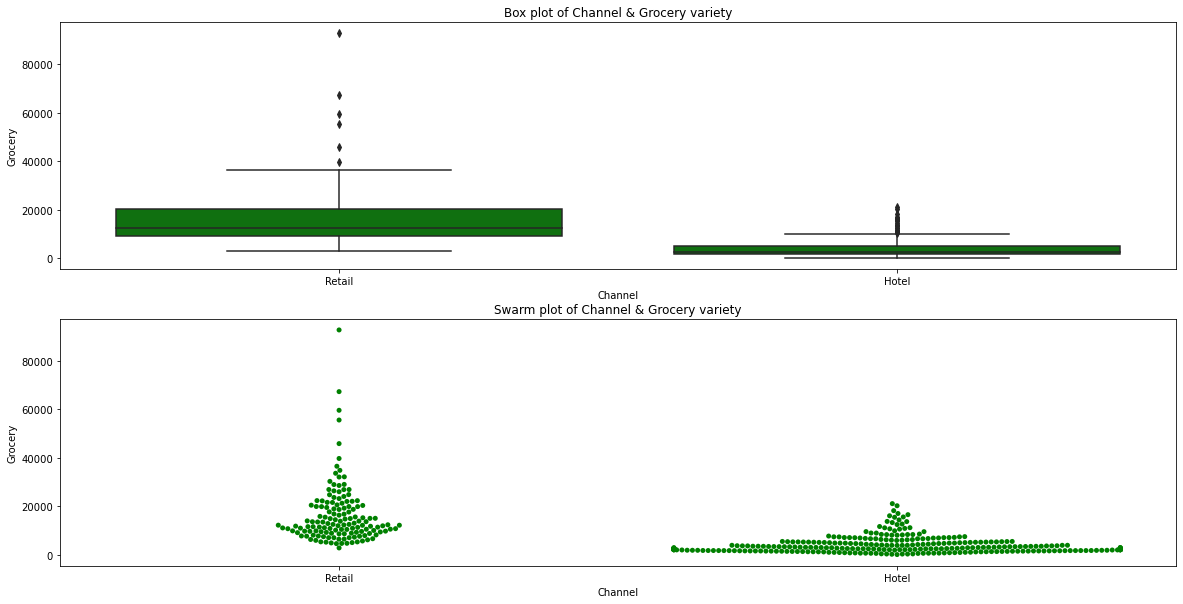
* + Data contains more retailers in Hotel channel as compared to Retail channel
  + Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
  + Maximum annual spending in 'Retail' Channel is high as compared to Hotel Channel
  + Min annual spending in 'Hotel' Channel is low as compared to Retail Channel
  + Annual median spend of Hotel channel(2157) is very low as compared to Retail channel(7812)
  + Volatility of Hotel channel is higher than of Retail channel.
  + Spread of data looks similar across both channels with distribution being right/positive skewed and 75% of retailers spending less than 4K annually on Hotel channel and 12K on Retail channel respectively.
  + Although footfall is more for buyers of Milk variety under Hotel channel but buyers are spending more money on Milk variety under Retail channel.

#### Grocery Variety behaviour across all channels

##### 5 points summary , Coefficient of Variation and Skewness

|  |  |  |
| --- | --- | --- |
|  | **Grocery\_Hotel** | **Grocery\_Retail** |
| **Count** | 298 | 142 |
| **Mean** | 3962.14 | 16322.85 |
| **Std** | 3545.51 | 12267.32 |
| **Min** | 3 | 2743 |
| **25%** | 1703.75 | 9245.25 |
| **50%** | 2684 | 12390 |
| **75%** | 5076.75 | 20183.5 |
| **Max** | 21042 | 92780 |
| **CV** | 0.89 | 0.75 |
| **Skew** | 2.12 | 2.98 |

##### Box Plot and Swarm Plot



Key Points as seen in the summary statistics and swarm plot/box plot,

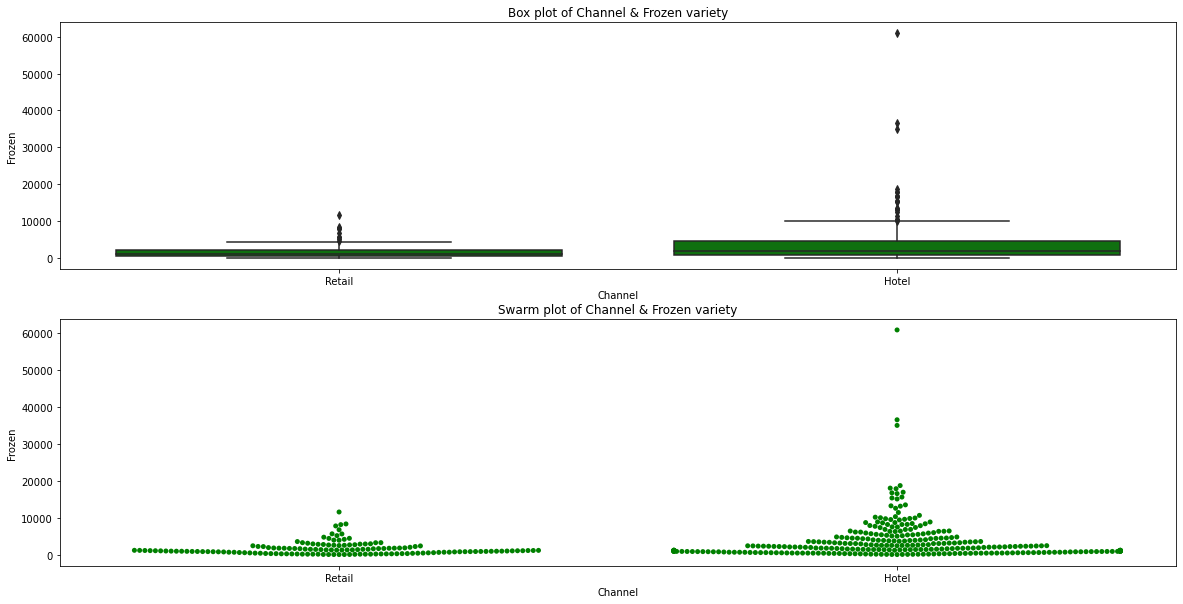
* Data contains more retailers in Hotel channel as compared to Retail channel
* Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
* Maximum annual spending in 'Retail' Channel is high as compared to Hotel Channel
* Min annual spending in 'Hotel' Channel is low as compared to Retail Channel
* Annual median spend of Hotel channel(2684) is very low as compared to Retail channel(12390)
* Volatility of Hotel channel is higher than of Retail channel.
* Spread of data looks similar across both channels with distribution being right/positive skewed and 75% of retailers spending less than 5K annually on Hotel channel and 20K on Retail channel respectively.
* Although Footfall is more for buyers of Grocery variety under Hotel channel but the buyers are spending more money on Grocery variety under Retail channel.

#### Frozen Variety behaviour across all channels

##### 5 points summary , Coefficient of Variation and Skewness

|  |  |  |
| --- | --- | --- |
|  | **Frozen\_Hotel** | **Frozen\_Retail** |
| **Count** | 298 | 142 |
| **Mean** | 3748.25 | 1652.61 |
| **Std** | 5643.91 | 1812.8 |
| **Min** | 25 | 33 |
| **25%** | 830 | 534.25 |
| **50%** | 2057.5 | 1081 |
| **75%** | 4558.75 | 2146.75 |
| **Max** | 60869 | 11559 |
| **CV** | 1.5 | 1.09 |
| **Skew** | 5.21 | 2.53 |

##### Box Plot and Swarm Plot



**Key Points** as seen in the summary statistics and swarm plot/box plot,

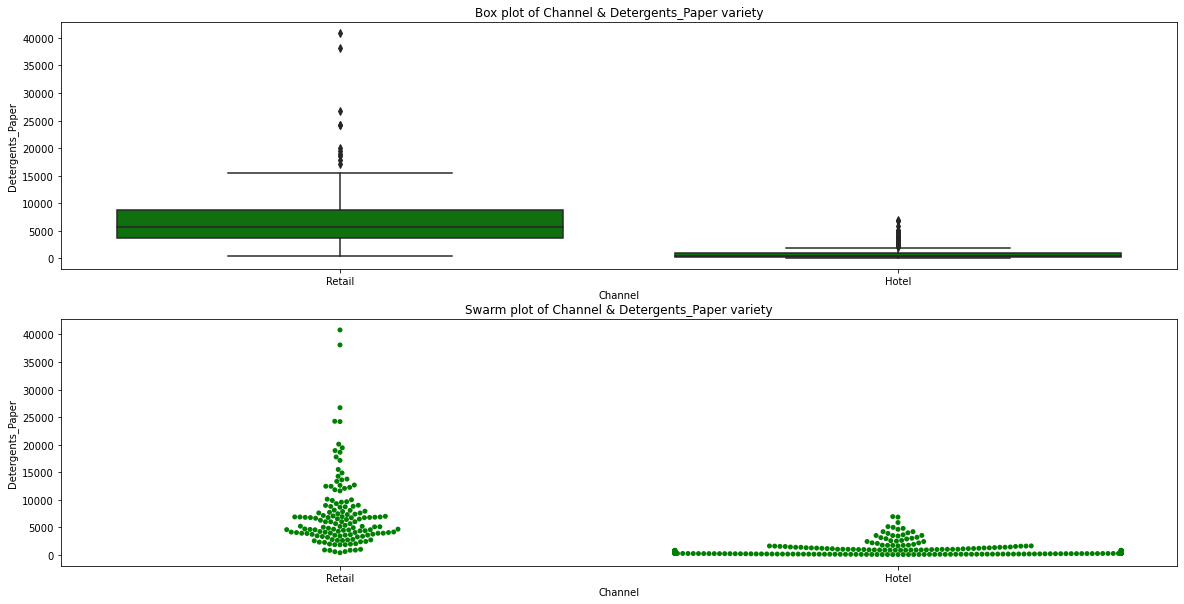
* Data contains more retailers in Hotel channel as compared to Retail channel
* Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
* Maximum annual spending in 'Hotel' Channel is high as compared to Retail Channel
* Min annual spending in 'Hotel' Channel is similar compared to Retail Channel
* Annual median spend of Hotel channel(2057) is high as compared to Retail channel(1081)
* Volatility of Hotel channel is higher than of Retail channel.
* Spread of data looks similar across both channels with distribution being right/positive skewed and 75% of retailers spending less than 4.5K annually on Hotel channel and 2.1K on Retail channel respectively.
* The Footfall is more for buyers under Hotel channel and the buyers are spending more money on Frozen variety under Hotel channel as compared to Retail channel

#### Detergents\_Paper Variety behaviour across all channels

##### 5 points summary , Coefficient of Variation and Skewness

|  |  |  |
| --- | --- | --- |
|  | **Detergents\_Paper\_Hotel** | **Detergents\_Paper\_Retail** |
| **Count** | 298 | 142 |
| **Mean** | 790.56 | 7269.51 |
| **Std** | 1104.09 | 6291.09 |
| **Min** | 3 | 332 |
| **25%** | 183.25 | 3683.5 |
| **50%** | 385.5 | 5614.5 |
| **75%** | 899.5 | 8662.5 |
| **Max** | 6907 | 40827 |
| **CV** | 1.39 | 0.86 |
| **Skew** | 2.86 | 2.61 |

##### Box Plot and Swarm Plot



**Key Points** as seen in the summary statistics and swarm plot/box plot,

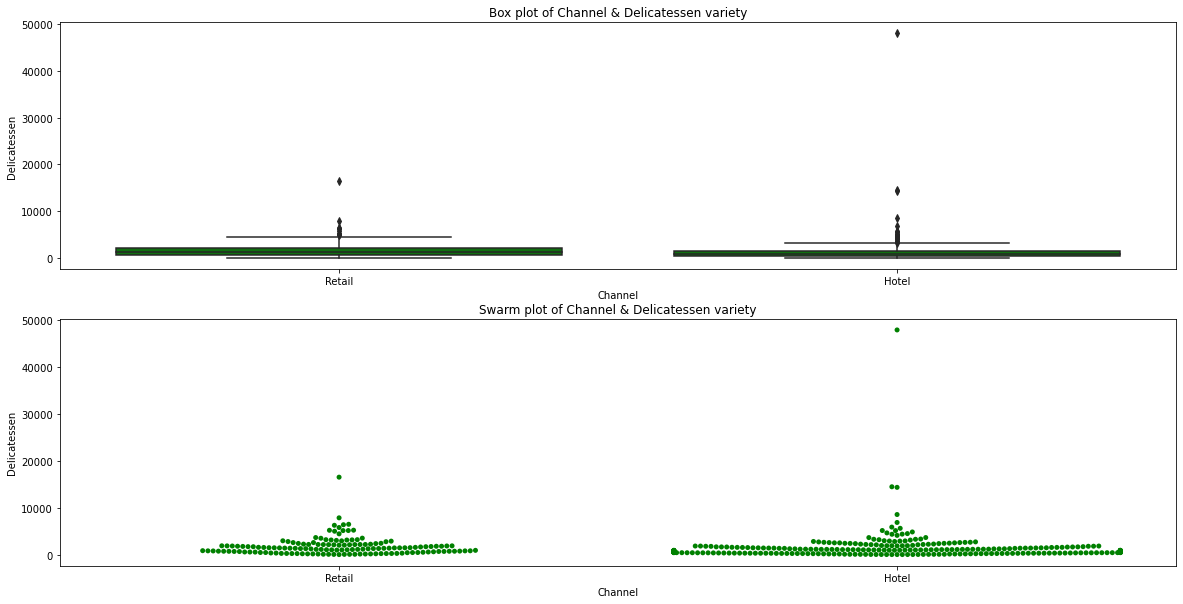
* Data contains more retailers in Hotel channel as compared to Retail channel
* Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
* Maximum annual spending in 'Retail' Channel is high as compared to Hotel Channel
* Min annual spending in 'Hotel' Channel is low as compared to Retail Channel
* Annual median spend of Hotel channel(385) is very low as compared to Retail channel(5614)
* Volatility of Hotel channel is higher than of Retail channel.
* Spread of data looks similar across both channels with distribution being right/positive skewed and 75% of retailers spending less than 0.9K annually on Hotel channel and 8.6K on Retail channel respectively.
* Although Footfall is more for buyers under Hotel channel but the buyers are spending more money on Detergent\_Paper variety under Retail channel.

#### Delicatessen Variety behaviour across all channels

##### 5 points summary , Coefficient of Variation and Skewness

|  |  |  |
| --- | --- | --- |
|  | **Delicatessen\_Hotel** | **Delicatessen\_Retail** |
| **Count** | 298 | 142 |
| **Mean** | 1415.96 | 1753.44 |
| **Std** | 3147.43 | 1953.8 |
| **Min** | 3 | 3 |
| **25%** | 379 | 566.75 |
| **50%** | 821 | 1350 |
| **75%** | 1548 | 2156 |
| **Max** | 47943 | 16523 |
| **CV** | 2.22 | 1.11 |
| **Skew** | 11.52 | 3.77 |

##### Box Plot and Swarm Plot



**Key Points** as seen in the summary statistics and swarm plot/box plot,

* Data contains more retailers in Hotel channel as compared to Retail channel
* Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
* Maximum annual spending in 'Hotel' Channel is high as compared to Retail Channel
* Min annual spending in 'Hotel' Channel is similar to Retail Channel
* Annual median spend of Hotel channel(821) is very low as compared to Retail channel(1350)
* Volatility of Hotel channel is higher than of Retail channel.
* Spread of data looks similar across both channels with distribution being right/positive skewed and 75% of retailers spending less than 1.5K annually on Hotel channel and 2.1K on Retail channel respectively.
* Although Footfall is more for buyers under Hotel channel but the buyers are spending more money on Delicateessen variety under Retail channel.

### ****Conclusions:-****

Spending pattern of all 6 items across region appears to be similar

Spending pattern of 6 items across channels gives us following points

* Fresh and Frozen variety is purchased more via Hotel Channel (annual median around 1.5 to 2 times) as compared to Retail channel
* Milk, Grocery is purchased more via Retail Channel(annual median around 4 to 5 times) as compared to Hotel Channel
* Detergents\_Paper is purchased very very less via Hotel Channel as compared to Retail Channel (annual median around 14 times)

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

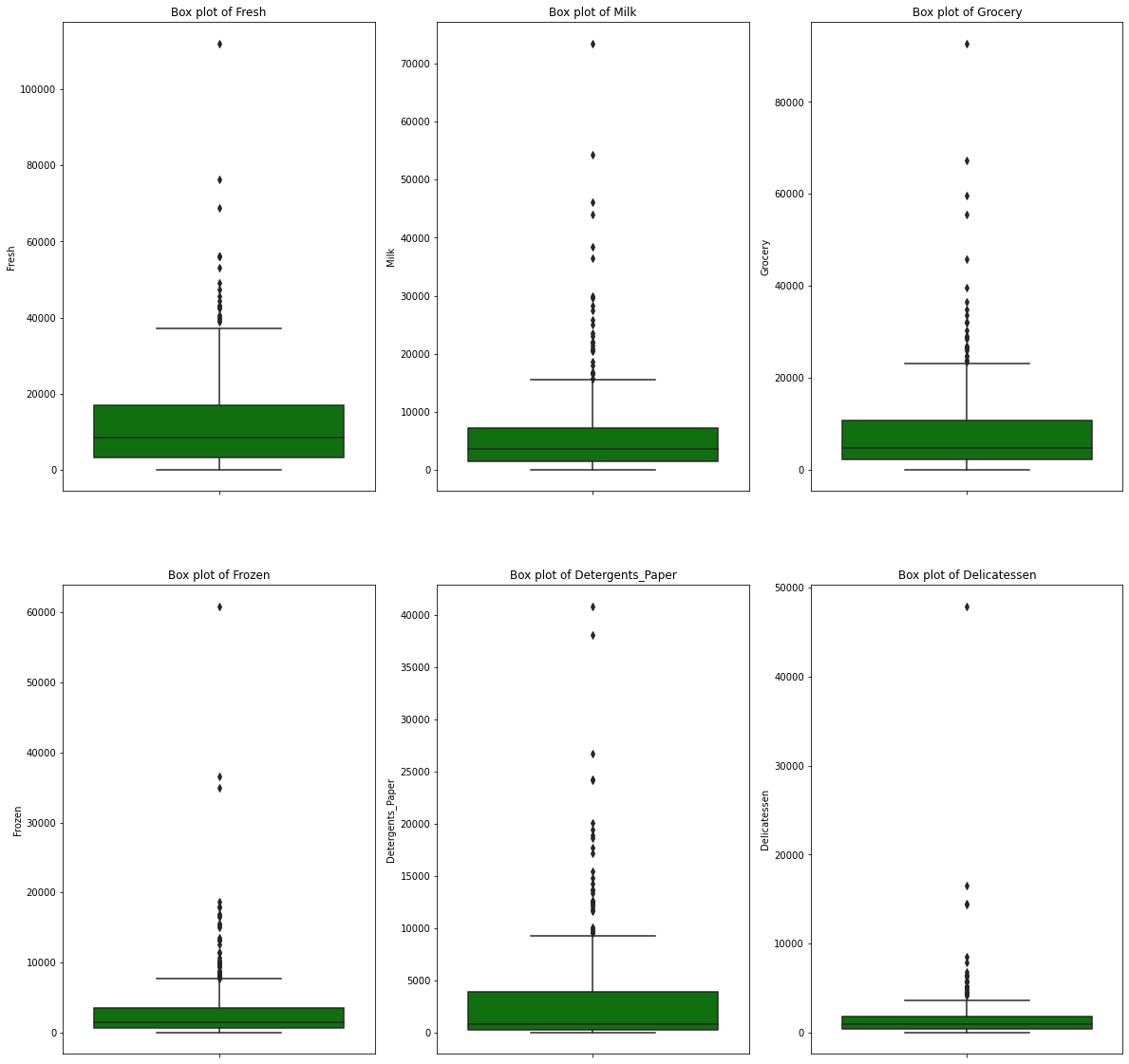
**For checking the most inconsistent behaviour, I have used CV (Coefficient of Variation) parameter for comparison. Code which is calculating the CV of all the 6 items can be seen in the notebook file attached. I have attached here is the summary report showing CV, mean & SD**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Fresh** | **Milk** | **Grocery** | **Frozen** | **Detergents\_Paper** | **Delicatessen** |
| **count** | 440 | 440 | 440 | 440 | 440 | 440 |
| **mean** | 12000.3 | 5796.27 | 7951.28 | 3071.93 | 2881.49 | 1524.87 |
| **std** | 12647.33 | 7380.38 | 9503.16 | 4854.67 | 4767.85 | 2820.11 |
| **min** | 3 | 55 | 3 | 25 | 3 | 3 |
| **25%** | 3127.75 | 1533 | 2153 | 742.25 | 256.75 | 408.25 |
| **50%** | 8504 | 3627 | 4755.5 | 1526 | 816.5 | 965.5 |
| **75%** | 16933.75 | 7190.25 | 10655.75 | 3554.25 | 3922 | 1820.25 |
| **max** | 112151 | 73498 | 92780 | 60869 | 40827 | 47943 |
| **CV** | 1.05 | 1.27 | 1.19 | 1.58 | 1.65 | 1.85 |

**As seen above from the CV values, item which is showing most inconsistent behaviour is 'Delicatessen' and item which is showing lowest inconsistent behaviour is 'Fresh'**

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

**I have plotted box plot for checking outliers. Copy of the boxplot image is attached below :-**



**As evident from above box plots, all attributes have outliers. All outliers are on maximum side i.e. there are few retailers which are spending much more than the majority of the retailers. Data appears to be right skewed**

## ****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****

**I have created a Pivot table on region and channel, and following table is the output**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Buyer/Spender | Delicatessen | Detergents\_Paper | Fresh | Frozen | Grocery | Milk |
| **Region** | Channel |  |  |  |  |  |  |  |
| **All** |  | 97020 | 670943 | 1267857 | 5280131 | 1351650 | 3498562 | 2550357 |
| **Other** | Hotel | 48020 | 320358 | 165990 | 2928269 | 771606 | 820101 | 735753 |
| **Retail** | 16006 | 191752 | 724420 | 1032308 | 158886 | 1675150 | 1153006 | 4935522 |
| **Lisbon** | Hotel | 14026 | 70632 | 56081 | 761233 | 184512 | 237542 | 228342 |
| **Retail** | 4069 | 33695 | 148055 | 93600 | 46514 | 332495 | 194112 | 848471 |
| **Oporto** | Retail | 5911 | 23541 | 159795 | 138506 | 29271 | 310200 | 174625 |
| **Hotel** | 8988 | 30965 | 13516 | 326215 | 160861 | 123074 | 64519 | 719150 |

In this table, we could see that Delicatessen total spent is very less as compared to other products. We could easily spot the highest selling varieties i.e. Fresh, Grocery and Milk.

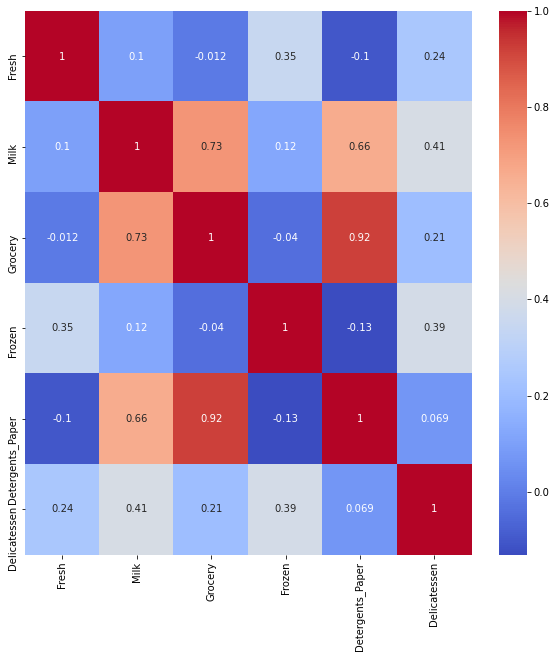
Then I created another Pivot table based on Channel and aggregated on median (not used mean as data contains outlier).

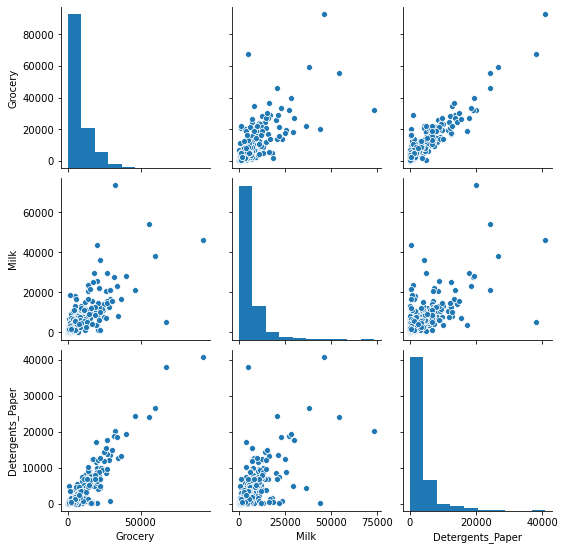
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Buyer/Spender** | **Delicatessen** | **Detergents\_Paper** | **Fresh** | **Frozen** | **Grocery** | **Milk** | **Total\_Spend** |
| **Channel** |  |  |  |  |  |  |  |  |
| **Hotel** | 241.5 | 821 | 385.5 | 9581.5 | 2057.5 | 2684 | 2157 | 21254.5 |
| **Retail** | 166.5 | 1350 | 5614.5 | 5993.5 | 1081 | 12390 | 7812 | 37139 |

In this table, we could see that some varieties are purchased more via Hotel channel (Fresh & Frozen) and some are purchased more via Retail channel (Milk & Grocery). I was not able to see any noticeable pattern in region wise pivot table.

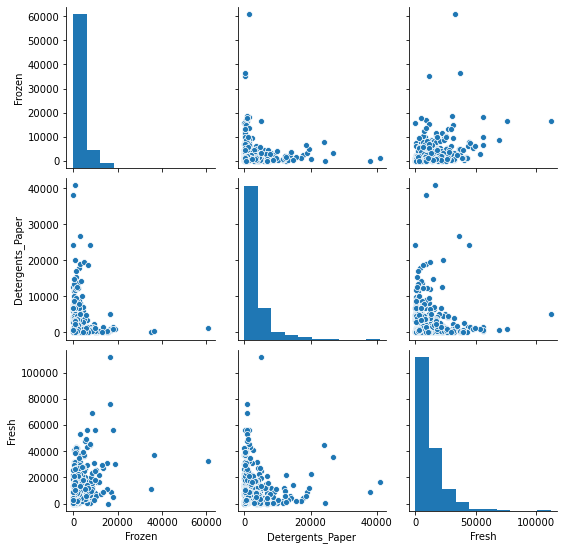
Then on checking the correlation we could see that there is very strong correlation between Grocery and Detergent\_Paper , Grocery and Milk. Also there exist a minor negative correlation between  Frozen and Detergent\_Paper , Detergent\_paper & Fresh. Heatmap and pairplot graphs are shown below :-

### Heatmap





### **Pairplot** showing strong positive correlation



### Pairplot showing negative correlation

***Thus after analysis , my recommendation for business is :-***

* As seen in summary and Pivot tables, Buyers are spending much more on Fresh Products, Grocery and Milk varieties. It means Fresh Products, Grocery and Milk varieties are high selling items and business should ensure proper supply of

these.varieties.

* As seen in summary, 75% of buyers are spending less than 4k annually on Frozen and Detergents Paper variety. Delicateessen variety is the least consumed

among all the 6 varieties with around 2K annual spend by 75% of buyers.

* + - * Delicateessen variety share is only 4.5% of total. May be price of these items could be very high causing low sales. So Business either need to drop this variety or perhaps needs put in more sales promotion/marketing efforts to boost its sale.
      * Fresh and Frozen variety is purchased more via Hotel Channel (annual median around 1.5 to 2 times) as compared to Retail channel
      * Milk, Grocery is purchased more via Retail Channel (annual median around 4 to 5 times) as compared to Hotel Channel
      * Detergents\_Paper is purchased very very less via Hotel Channel as compared

to Retail Channel (annual median around 14 times)

* + - * Thus Business should focus on sale of more Fresh and Frozen varieties under

Hotel Channel and more Milk, Grocery and Detergent\_Paper varities under

Retail Channel.

* + - * Since there is a strong correlation between Grocery & Milk, Grocery &

Detergent\_Paper business could give combo offers for these varieties like

BOGO scheme etc for sale promotion

* + - * Also clubbing of negative correlated varieties like Detergent\_Paper & Frozen, Detergent\_Paper & Fresh for any combo pack scheme should be avoided by

Business.

# Problem 2

## Problem Statement:

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).

## Data Description:

CMSU creates and distributes a survey of 14 questions and receives responses from 62 undergraduates (stored in the Survey data set)

## Domain:

Education

## Context:

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).

## Attribute Information:

ID

Gender

Age

Class

Major

Grad Intention

GPA

Employment

Salary

Social Networking

Satisfaction

Spending

Computer

Text Messages

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

### ****2.1.1. Gender and Major****

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

### ****2.1.2. Gender and Grad Intention****

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

### ****2.1.3. Gender and Employment****

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

### ****2.1.4. Gender and Computer****

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

## ****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?****

**Male\_student= 29**

**Total\_student=62**

**Probability that CMSU Student will be Male is 29/62 i.e. 0.468**

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

**Students can be either male or female i.e. both are mutually exclusive event**

**Prob Female student = 1 – Prob Male student = 1 – 0.468 = 0.532**

**Thus Probability that CMSU Student will be Female is 0.532**

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

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

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

Cond. Pr. that CMSU Student has Accounting major among male student is 4/29 = 0.138

Cond. Pr. that CMSU Student has CIS major among male student is 1/29 = 0.034

Cond. Pr. that CMSU Student has Economics/Finance major among male student is 4/29 = 0.138

Cond. Pr. that CMSU Student has International Business major among male student is 2/29 = 0.069

Cond. Pr. that CMSU Student has Management major among male student is 6/29 = 0.207

Cond. Pr. that CMSU Student has Other major among male student is 4/29 = 0.138

Cond. Pr. that CMSU Student has Retailing/Marketing major among male student is 5/29 = 0.172

Cond. Pr. that CMSU Student has Undecided major among male student is 3/29 = 0.103

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

Cond. Pr. that CMSU Student has Accounting major among female student is 3/33 = 0.091

Cond. Pr. that CMSU Student has CIS major among female student is 3/33 = 0.091

Cond. Pr. that CMSU Student has Economics/Finance major among female student is 7/33 = 0.212

Cond. Pr. that CMSU Student has International Business major among female student is 4/33 = 0.121

Cond. Pr. that CMSU Student has Management major among female student is 4/33 = 0.121

Cond. Pr. that CMSU Student has Other major among female student is 3/33 = 0.091

Cond. Pr. that CMSU Student has Retailing/Marketing major among female student is 9/33 = 0.273

Cond. Pr. that CMSU Student has Undecided major among female student is 0/33 = 0.000

## ****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.****

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

**P(Male and Intends to Graduate) = P(Male) \* P(Intends to Graduate| Male)**

**= 0.468 \* 17 / 29**

**= 0.274**

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

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

**P(Female and Doesnt have laptop) = P(Female) \* P(Doesnt have laptop| Female)**

**= 0.532 \* (2+2) / 33**

**= 0.064**

## ****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 either a male or has full-time employment?****

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

**P(Male or Full\_Time)** = P(Male) + P(Full\_Time) - P(Male and Full\_Time)

= 29/62 + 10/62 – 7/62

= 0.516

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

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

**P(International Business OR Management | Female)**

**= (Total\_international\_business\_female + Total\_management\_female) / Total\_female**

= (4 +4 )/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?****

**Contingency table after removing undecided students from the original data is as shown below :-**

|  |  |  |  |
| --- | --- | --- | --- |
| **Grad Intention** | **No** | **Yes** | **All** |
| **Gender** |  |  |  |
| **Female** | 9 | 11 | 20 |
| **Male** | 3 | 17 | 20 |
| **All** | 12 | 28 | 40 |

**Now for checking whether two events graduate intention and being female are independent or not, I am checking Multiplication rule of Probability.**

**P(A ∩ B) = P(A) \* P(B) if A & B are independent events.**

**Here A = Female and B = Yes (Graduate Intention)**

**So if P(Female ∩ Yes) equals to P(Female) \* P(Yes), then these are independent otherwise these are not.**

**P(Female ∩ Yes) = 11/40 = 0.275**

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

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

**P(Female) \* P(Yes) = 0.35**

**Now as P(Female) \* P(Yes) is not equal to P(Female ∩ Yes), we can confirm that both these events are not 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?****

**Total\_student = 62**

**Students with GPA less than 3 = 17**

**Probability that GPA less than 3 = 17/62 = 0.274**

### ****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.****

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Salary** | 25 | 30 | 35 | 37 | 37.5 | 40 | 42 | 45 | 47 | 47.5 | 50 | 52 | 54 | 55 | 60 | 65 | 70 | 78 | 80 | All |
| **Gender** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Female** | 0 | 5 | 1 | 0 | 1 | 5 | 1 | 1 | 0 | 1 | 5 | 0 | 0 | 5 | 5 | 0 | 1 | 1 | 1 | 33 |
| **Male** | 1 | 0 | 1 | 1 | 0 | 7 | 0 | 4 | 1 | 0 | 4 | 1 | 1 | 3 | 3 | 1 | 0 | 0 | 1 | 29 |
| **All** | 1 | 5 | 2 | 1 | 1 | 12 | 1 | 5 | 1 | 1 | 9 | 1 | 1 | 8 | 8 | 1 | 1 | 1 | 2 | 62 |

**P(student salary >=50 | male) = (4+1+1+3+3+1+0+0+1 ) / 29**

**= 0.483**

**P(student salary >=50 | female) = (5+0+0+5+5+0+1+1+1 ) / 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. Write a note summarizing your conclusions****.

I have used **Shapiro algorithm** for checking normal distribution. The Null Hypothesis of this test is that the population is normally distributed. Thus, if the p value is less than the chosen alpha level, then the null hypothesis is rejected and there is evidence that the data tested are not normally distributed. On the other hand, if the *p* value is greater than the chosen alpha level, then the null hypothesis (that the data came from a normally distributed population) cannot be rejected

Further I have also checked the **Empirical rule** of normality. It has the three conditions:-

Approx 68% of data should lie within 1 Standard deviation of the mean i.e mu +- 1 sigma

Approx 95% of data should lie within 2 Standard deviation of the mean i.e mu +- 2 sigma

Approx 99.7% of data should lie within 3 Standard deviation of the mean i.e mu +- 3 sigma

Then in the end, I have also compared the **mean, mode and median** values. These values should either be equal or should be very close to each other for normal distributions.

Visually I have plotted histogram for each continuous variable with Kernel Density on to check the shape of normal distribution (bell curve).

All the coding details are present in the notebook file attached. Results of all these test conditions for each of the 4 continuous variable is as shown below :-

### GPA

**Empirical result**

68% of the data should lie between 2.75 and 4

72.58% data lies LL and UL for 68%

95% of the data should lie between 2.37 and 4

96.77% data lies LL and UL for 95%

99% of the data should lie between 1.99 and 4

100.0% data lies LL and UL for 99%

**Shapiro**

Statistics = 0.969, p-value = 0.112

GPA follows a Normal Distribution

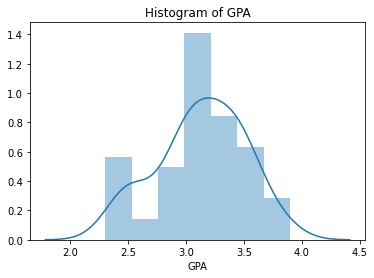
**Mean, Median & Mode**

mean of GPA is 3.13

mode of GPA is 3.00

median of GPA is 3.15

**Histogram**

****

### Salary

**Empirical result**

68% of the data should lie between 36.47 and 61

79.03% data lies LL and UL for 68%

95% of the data should lie between 24.39 and 73

95.16% data lies LL and UL for 95%

99% of the data should lie between 12.31 and 85

100.0% data lies LL and UL for 99%

**Shapiro**

Statistics = 0.957, p-value = 0.028

Salary doesnt follows a Normal Distribution

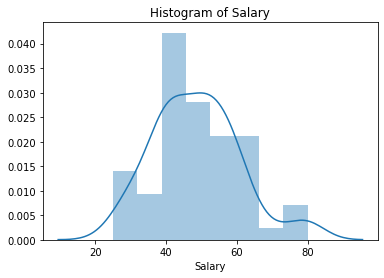
**Mean, Median & Mode**

mean of Salary is 48.55

mode of Salary is 40.00

median of Salary is 50.00

**Histogram**



### Spending

**Empirical result**

68% of the data should lie between 260.07 and 704

80.65% data lies LL and UL for 68%

95% of the data should lie between 38.12 and 926

95.16% data lies LL and UL for 95%

99% of the data should lie between -183.83 and 1148

98.39% data lies LL and UL for 99%

**Shapiro**

Statistics = 0.878, p-value = 0.000

Spending doesnt follows a Normal Distribution

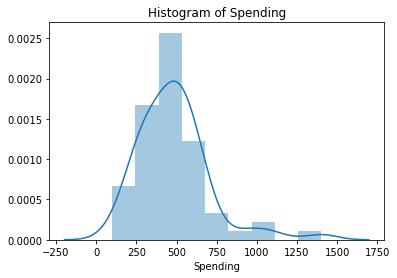
**Mean, Median & Mode**

mean of Spending is 482.02

mode of Spending is 500.00

median of Spending is 500.00

**Histogram**



### Text Messages

**Empirical result**

68% of the data should lie between 31.74 and 461

79.03% data lies LL and UL for 68%

95% of the data should lie between -182.73 and 675

91.94% data lies LL and UL for 95%

99% of the data should lie between -397.2 and 890

98.39% data lies LL and UL for 99%

**Shapiro**

Statistics = 0.859, p-value = 0.000

Text Messages doesnt follows a Normal Distribution

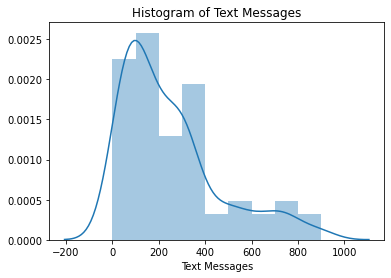
**Mean, Median & Mode**

mean of Text Messages is 246.21

mode of Text Messages is 300.00

median of Text Messages is 200.00

**Histogram**



**Summary :-**

I have used the Shapiro test for all the 4 continuous variables and found that at 95% confidence level, there is sufficient evidence to prove that only ‘GPA’ out of 4 given continuous variables is following the normal distribution. ‘Salary’, ‘Text Messages’ & ‘Spending’ are not following the normal distribution.

# Problem 3

## Problem Statement:

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 would like to show that the mean moisture content is less than 0.35 pound per 100 square feet.

## Data Description:

The file (A & B shingles.csv) includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.

## Domain:

Manufacturing

## Context:

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 would like to show that the mean moisture content is less than 0.35 pound per 100 square feet.

## Attribute Information:

A : 36 measurements for A shingles

B : 31 measurements 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.****

### ****Hypothesis Testing for Shingle A****

#### Step 1: Define null and alternative hypotheses for shingle A

#### Null hypothesis states that means moisture contents, 𝜇 is less than equal to 0.35

#### Alternative hypothesis states that the means moisture contents, 𝜇 is greater than 0.35.

* **𝐻0 H0: μ <= 0.35**
* **𝐻𝐴 HA: μ > 0.35**

#### Step 2: Decide the significance level

Here we select **α = 0.05.**

#### Step 3: Identify the test statistic

We do not know the population standard deviation and **n = 36.** So we use the t distribution and the tSTAT test statistic.

#### Step 4: Calculate the p - value and test statistic

**scipy.stats.ttest\_1samp calculates the t test for the mean of one sample given the sample observations and the expected value in the null hypothesis. This function returns t statistic and the two-tailed p value. Its values from the output of function are :-**

One sample t test for Shingle A

**t statistic: -1.4735046253382782 p value: 0.07477633144907513**

#### Step 5 Decide to reject or accept null hypothesis

In this example, p value is 0.07 and it is greater than 5% level of significance

So the statistical decision is **failing to reject the null hypothesis** at 5% level of significance.

##### **So at 95% confidence level, there is sufficient evidence to prove that means moisture contents of shingle A is within permissible limit of 0.35 pound per 100 square feet**[**¶**](http://localhost:8888/notebooks/Scripts/statistical%20method/Project/Raghav_Gupta_SMDM%20Project_31-12-2020.ipynb#So-at-95%-confidence-level,-there-is--sufficient-evidence--to-prove-that-means-moisture-contents-of-shingle-A-is-within-permissible-limit-of-0.35-pound-per-100-square-feet)

### ****Hypothesis Testing for Shingle B****

#### Step 1: Define null and alternative hypotheses for shingle B

#### Null hypothesis states that means moisture contents, 𝜇 is less than equal to 0.35

#### Alternative hypothesis states that the means moisture contents, 𝜇 is greater than 0.35.

* **𝐻0 H0: μ <= 0.35**
* **𝐻𝐴 HA: μ > 0.35**

#### Step 2: Decide the significance level

Here we select **α = 0.05.**

#### Step 3: Identify the test statistic

We do not know the population standard deviation and **n = 31**. So we use the t distribution and the tSTAT test statistic.

#### Step 4: Calculate the p - value and test statistic

**scipy.stats.ttest\_1samp calculates the t test for the mean of one sample given the sample observations and the expected value in the null hypothesis. This function returns t statistic and the two-tailed p value. Its values from the output of function are :-**

**.**

One sample t test for Shingle B

**t statistic: -3.1003313069986995 p value: 0.0020904774003191826**

#### Step 5 Decide to reject or accept null hypothesis

Level of significance: 0.05

Our one-sample t-test p-value=0.00

So the statistical decision is to reject the null hypothesis at 5% level of significance.

##### **Hence at 95% confidence level, there is sufficient evidence to prove that means moisture contents of shingle B is not in the permissible limit of 0.35 pound per 100 square feet**[**¶**](http://localhost:8888/notebooks/Scripts/statistical%20method/Project/Raghav_Gupta_SMDM%20Project_31-12-2020.ipynb#So-at-95%-confidence-level,-there-is--sufficient-evidence--to-prove-that-means-moisture-contents-of-shingle-B-is-not-in-the-permissible-limit-of-0.35-pound-per-100-square-feet)

## ****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 for T-Test that we need to check before the test

1. The scale of measurement applied to the data follows a continuous or ordinal scale
2. The data is collected from a randomly selected portion of the total population
3. The data collected follows a normal distribution
4. Sample size should be large

### Hypothesis Testing

#### Step 1: Define null and alternative hypotheses for shingle A and B equal Mean

#### Null hypothesis states that means moisture contents, μA is equal to μB

#### Alternative hypothesis states that the means moisture contents, μA is not equal to μB

* H0:  μA = μB
* HA:  μA ≠ μB

#### Step 2: Decide the significance level

Here we select α = 0.05.

#### Step 3: Identify the test statistic

We do not know the population standard deviation and **n1 = 36, n2=31.** So we use the t distribution 2 sample test statistic.

#### Step 4: Calculate the p - value and test statistic

**scipy.stats.ttest\_ind calculates the t test for the mean of two samples given the sample observations and the expected value in the null hypothesis. This function returns t statistic and the two-tailed p value. Its values from the output of function after omitting the null values are:-**

Two sample t test for Shingle A & B

**t statistic: 1.2896282719661123 p value: 0.2017496571835306**

#### Step 5 Decide to reject or accept null hypothesis

Level of significance: 0.05

Our one-sample t-test p-value=0.20

We have no evidence to reject the null hypothesis since p value > Level of significance

In this example, p value is 0.20 and it is greater than 5% level of significance

So the statistical decision is failing to reject the null hypothesis at 5% level of significance.

##### **So at 95% confidence level, there is sufficient evidence to prove that means moisture contents of shingle A and shingle B are equal**