SMDM PROJECT SAMPLE REPORT

**Contents:**

**Problem 1:**

* 1. **Use methods of descriptive statistics to summarize data. Which Region and which Channel spent the most? Which Region and which Channel spent the least?**
  2. **There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.**
  3. **On the basis of a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?**
  4. **Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.**
  5. **On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem? Answer from the business perspective**

**Problem 2:**

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?**

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

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

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

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

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

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

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

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

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

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

**2.6.  Construct a contingency table of Gender and Intent to Graduate at 2 levels (Yes/No). The Undecided students are not considered now and the table is a 2x2 table. Do you think the graduate intention and being female are independent events?**

**2.7. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages.**

**Answer the following questions based on the data**

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

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

**2.8. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarizing your conclusions.**

**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 ([A & B shingles.csv](https://olympus.mygreatlearning.com/courses/67575/files/4317942/download?verifier=HMuRUxghivVnnk5VPUKphnGjW1wkeomMyRiU1bVF&wrap=1)) 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.**

**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?**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Channel | Region | Fresh | Milk | Grocery | Frozen | Detergents\_Paper | Delicatessen |  |
| 0 | Retail | Other | 12669 | 9656 | 7561 | 214 | 2674 | 1338 |
| 1 | Retail | Other | 7057 | 9810 | 9568 | 1762 | 3293 | 1776 |
| 2 | Retail | Other | 6353 | 8808 | 7684 | 2405 | 3516 | 7844 |
| 3 | Hotel | Other | 13265 | 1196 | 4221 | 6404 | 507 | 1788 |
| 4 | Retail | Other | 22615 | 5410 | 7198 | 3915 | 1777 | 5185 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 435 | Hotel | Other | 29703 | 12051 | 16027 | 13135 | 182 | 2204 |
| 436 | Hotel | Other | 39228 | 1431 | 764 | 4510 | 93 | 2346 |
| 437 | Retail | Other | 14531 | 15488 | 30243 | 437 | 14841 | 1867 |
| 438 | Hotel | Other | 10290 | 1981 | 2232 | 1038 | 168 | 2125 |

**Above is the data available**

**Q. 1.1 Use methods of descriptive statistics to summarize data. Which Region and which Channel spent the most? Which Region and which Channel spent the least?**

We can see from the above data that there are 7 kinds of products having numerical values of their sales and 2 categorical variables Region (showing sales in region) and Channel (showing sales in channel).

There are 439 entries

There are a total 3 Regions, namely Lisbon, Oporto & Other

There are a total 2 Channels, namely Hotel & Retail

**Summary of data**

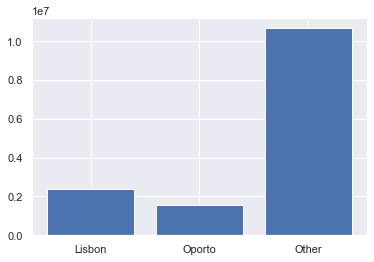
|  | **count** | **unique** | **top** | **freq** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Channel** | 440 | 2 | Hotel | 298 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **Region** | 440 | 3 | Other | 316 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **Fresh** | 440.0 | NaN | NaN | NaN | 12000.297727 | 12647.328865 | 3.0 | 3127.75 | 8504.0 | 16933.75 | 112151.0 |
| **Milk** | 440.0 | NaN | NaN | NaN | 5796.265909 | 7380.377175 | 55.0 | 1533.0 | 3627.0 | 7190.25 | 73498.0 |
| **Grocery** | 440.0 | NaN | NaN | NaN | 7951.277273 | 9503.162829 | 3.0 | 2153.0 | 4755.5 | 10655.75 | 92780.0 |
| **Frozen** | 440.0 | NaN | NaN | NaN | 3071.931818 | 4854.673333 | 25.0 | 742.25 | 1526.0 | 3554.25 | 60869.0 |
| **Detergents\_Paper** | 440.0 | NaN | NaN | NaN | 2881.493182 | 4767.854448 | 3.0 | 256.75 | 816.5 | 3922.0 | 40827.0 |
| **Delicatessen** | 440.0 | NaN | NaN | NaN | 1524.870455 | 2820.105937 | 3.0 | 408.25 | 965.5 | 1820.25 | 47943.0 |

**From the above tabular data we can find important meaningful data like most transactions are with Hotel.**

**Most Transactions also occur in Other region.**

**And mean, median, interquartile, minimum and maximum values can be seen for each numerical variable.**

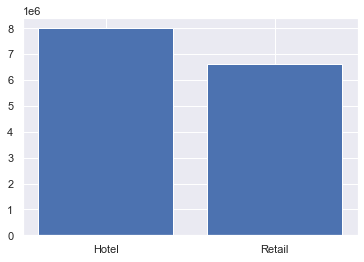
**From numeric data one can say that Fresh products have the highest average sale of** 12000 **but also have a high standard deviation of** 12647.33 **which could be because of outliers. Also mode for Fresh products is 8040. So as median is** 8504.0 **also the highest and close to mode we can safely say that Fresh product are very much in demand.**



|  | **Total Spending** |
| --- | --- |
| **Region** |  |
| **Lisbon** | 2386813 |
| **Oporto** | 1555088 |
| **Other** | 10677599 |

**From the graph we can understand that Other Region has spent the most and Oporto has spent the least**

|  | **Total Spending** |
| --- | --- |
| **Channel** |  |
| **Hotel** | 7999569 |
| **Retail** | 6619931 |

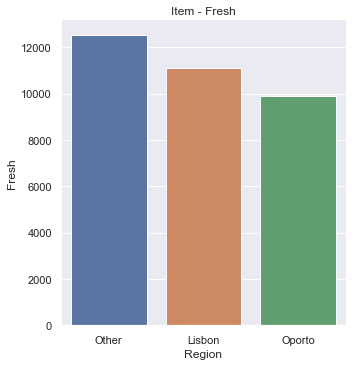


**From the graph above we can see that the Channel Hotel has spent the more than Channel Retail has spent. As there are only two Channels, Hotel has spent the most and Retail the least.**

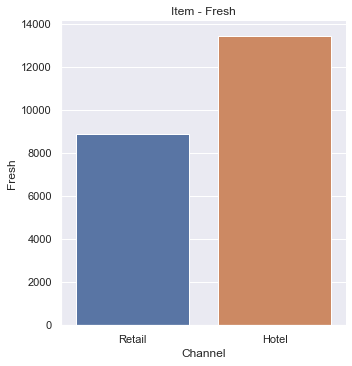
**Q. 1.2. There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.**

**There are 6 different varieties of items namely:**

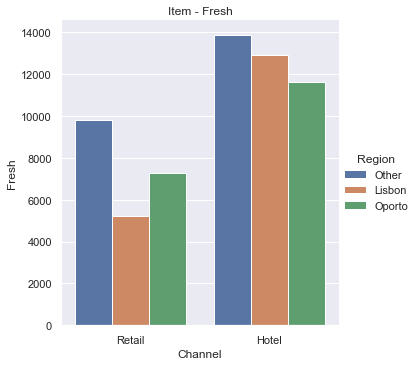
1. **Fresh**
2. **Milk**
3. **Grocery**
4. **Frozen**
5. **Detergents & Paper**
6. **Delicatessen**



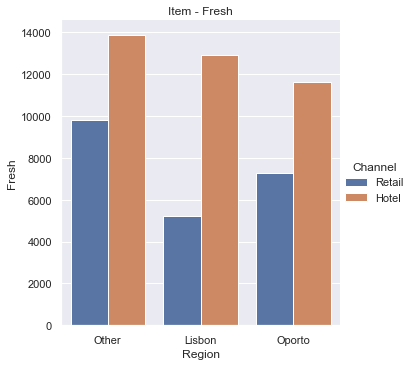
**Above graph shows us that the highest sale of Fresh product is in Other region.**



**Above graph shows that the highest sale of Fresh product is in Hotel channel.**

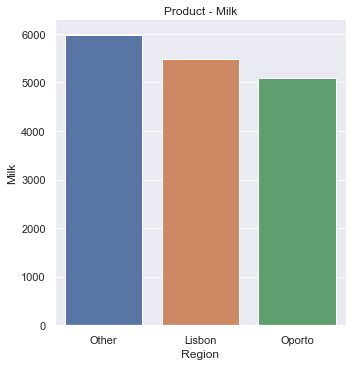


**This bar graph is showing the comparison of sale of Fresh product in different region in comparison with Retail and Hotel. We see that the sale of Fresh product is highest in Other region for both the channels.**

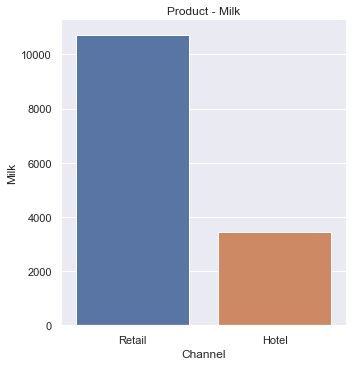


**This bar graph is showing the comparison of sale of Fresh product in different channel in comparison with region. We see that the sale of Fresh product is higher in Hotel channel in all the regions.**

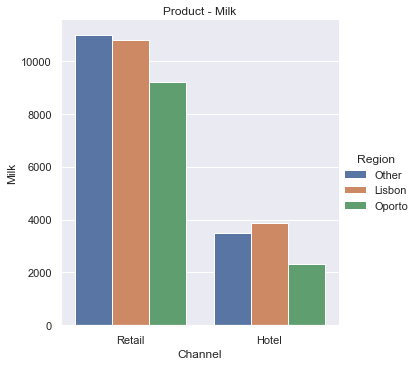
**2.Milk**



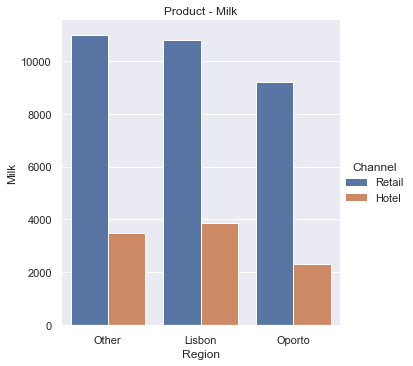
**We can find through this bar chart that the sale of Milk is highest in Other region.**



**We can find through this bar chart that the sale of Milk is highest in Retail channel.**

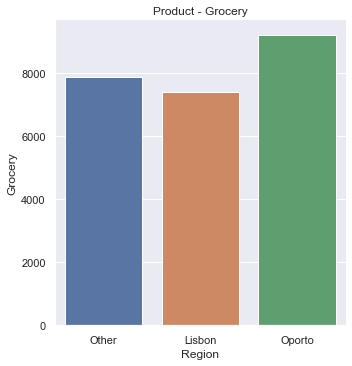


**This bar graph is showing the comparison of sale of Milk product in different region in comparison with Retail and Hotel. We can see that the sale of Milk product is highest in Other region in Retail channel but for Hotel channel it is highest in Lisbon. So population in Other channel consumes more milk in Other region but in Lisbon, Hotel channel has more demand for Milk.**

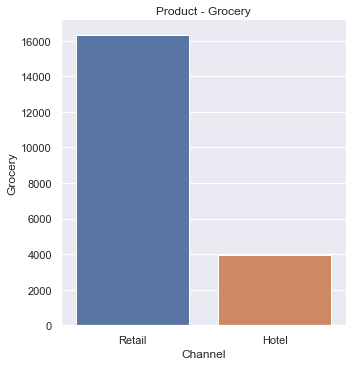


**This bar graph is showing the comparison of sale of Milk product in different channel in comparison with region. We see that the sale of Milk product is higher in Retail channel in all the regions.**

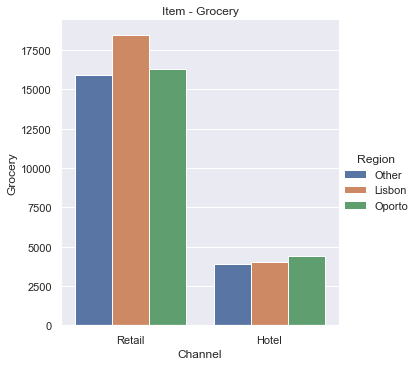
**3. Grocery**



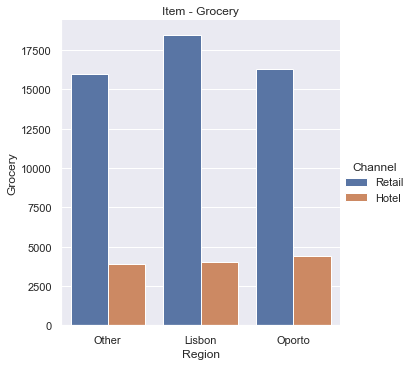
**We can find through this bar chart that the sale of Grocery is highest in Oporto region.**



**We can find through this bar chart that the sale of Grocery is highest in Retail channel.**

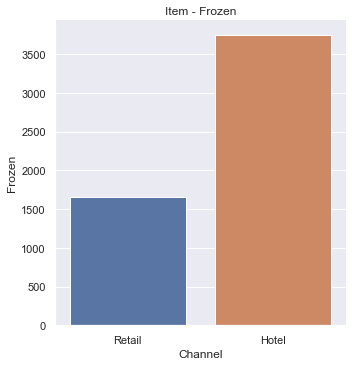
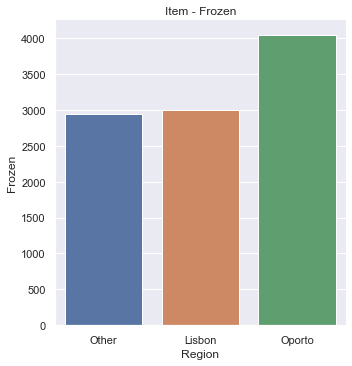


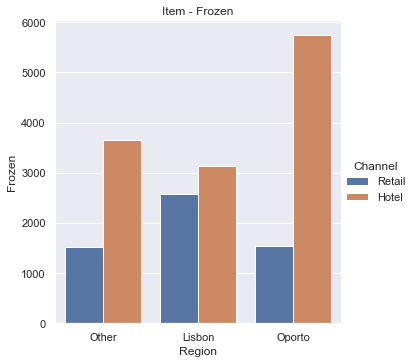
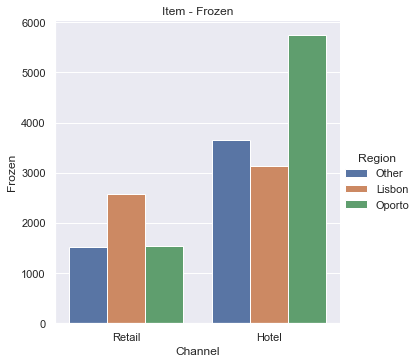
**This bar graph is showing the comparison of sale of Grocery product in different region in comparison with Retail and Hotel. We can see that the sale of Milk product is highest in Lisbon region in Retail channel but for Hotel channel it is highest in Oporto.**



**This bar graph is showing the comparison of sale of Grocery product in different channel in comparison with region. We see that the sale of Grocery product is higher in Retail channel in all the regions.**

**4. Frozen**

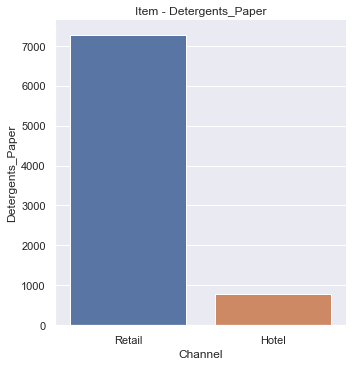
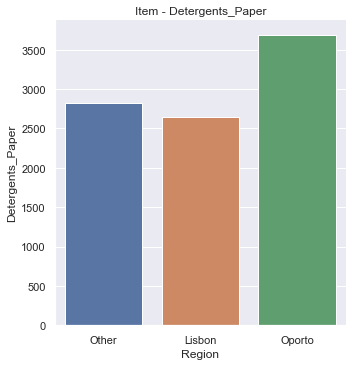
****

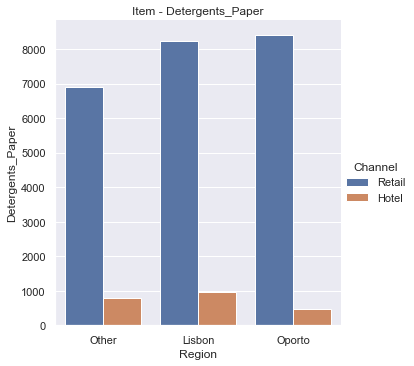
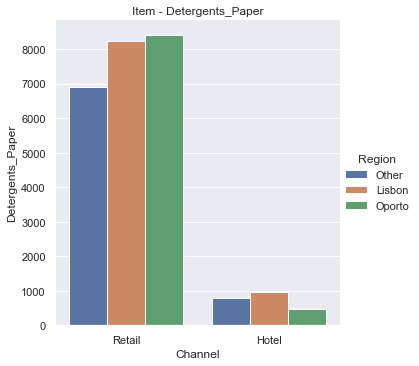
****

**From the above bar graphs one can understand the behaviour of the sales in different region to different channels. Here we see that Frozen product has the highest sale in Oporto and in Oporto majorly it is sold to the Hotel channel. Also it is sold more to the hotel channel in every region as compared to Retail channel.**

**We also see that it draws maximum sale from Lisbon in Retail channel.**

**5. Detergents and Paper**

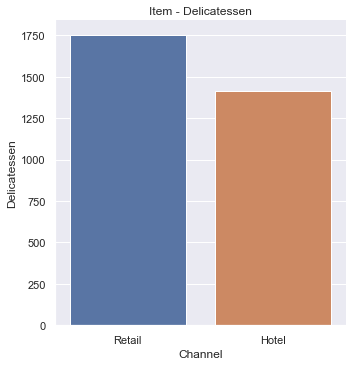
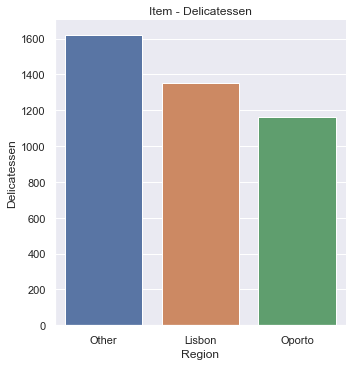
****

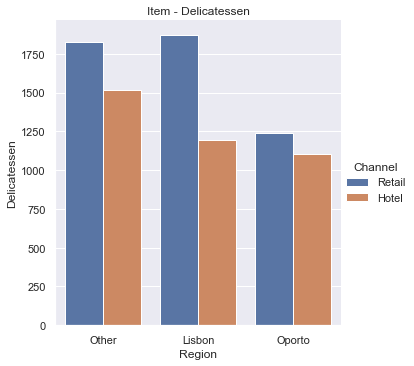
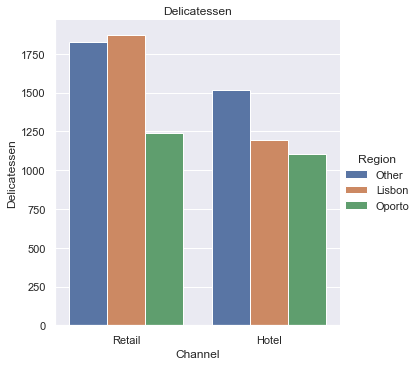
****

**From the above bar graphs one can understand the behaviour of the sales in different region to different channels. Here we see that Detergent and Paper product has the highest sale in Oporto and in Oporto majorly it is sold to the Retail channel. Also it is sold more to the Retail channel in every region as compared to Hotel channel.**

**We also see that it draws maximum sale from Lisbon in Hotel channel.**

1. **Delicatessen**

****

****

**From the above bar graphs one can understand the behaviour of the sales in different region to different channels. Here we see that Delicatessen product has the highest sale in Other and in Other region it is majorly sold in Retail. Also it is sold more to the Retail channel in every region as compared to Hotel channel.**

**We also interestingly see that it draws maximum sale from Lisbon in Hotel channel but overall sale in Other region is higher.**

**Q. 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?**

|  | **Fresh** | **Milk** | **Grocery** | **Frozen** | **Detergents\_Paper** | **Delicatessen** | **Total Spending** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **count** | 440.000000 | 440.000000 | 440.000000 | 440.000000 | 440.000000 | 440.000000 | 440.000000 |
| **mean** | 12000.297727 | 5796.265909 | 7951.277273 | 3071.931818 | 2881.493182 | 1524.870455 | 33226.136364 |
| **std** | 12647.328865 | 7380.377175 | 9503.162829 | 4854.673333 | 4767.854448 | 2820.105937 | 26356.301730 |
| **min** | 3.000000 | 55.000000 | 3.000000 | 25.000000 | 3.000000 | 3.000000 | 904.000000 |
| **25%** | 3127.750000 | 1533.000000 | 2153.000000 | 742.250000 | 256.750000 | 408.250000 | 17448.750000 |
| **50%** | 8504.000000 | 3627.000000 | 4755.500000 | 1526.000000 | 816.500000 | 965.500000 | 27492.000000 |
| **75%** | 16933.750000 | 7190.250000 | 10655.750000 | 3554.250000 | 3922.000000 | 1820.250000 | 41307.500000 |
| **max** | 112151.000000 | 73498.000000 | 92780.000000 | 60869.000000 | 40827.000000 | 47943.000000 | 199891.000000 |

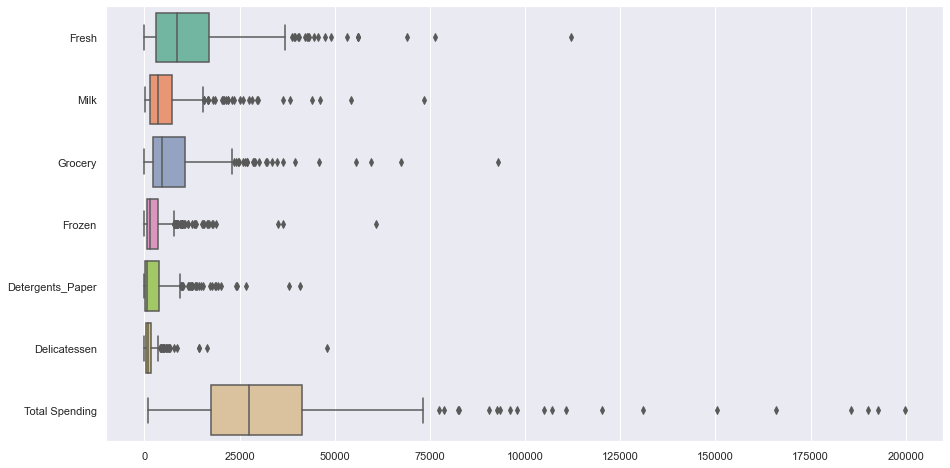
* **As coefficient of variance of Fresh product is the least i.e.** 1.0527196084948245, it is most

consistent.

* **As coefficient of variance of** Delicatessen **product is the highest i.e.** 1.8473041039189306,

it is least consistent.

**Q. 1.4. 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 that we can see in this box chart.**

**We see all the products have outliers in this data but take it as the data is totally right and collected correctly so these outliers are not a mistake and should be included in studying this data.**

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

**As we can see from the above data the maximum sales are in Other region and has a big market share. Still for some products there is less demand in Other region compared to Oporto and Lisbon like: Frozen, Grocery and Detergent & Paper.**

**It could be good to do business in Other region as it has high demand in almost all the products but also the competition could be high in this region and there could be untapped potential market in Lisbon and Oporto.**



**We can also see from this heatmap that Grocery and Detergent & Paper has a very high correlation.**

**That means if one purchases Grocery one is most probably to buy Detergent & Paper. Hence seller may plan to keep these products together for easy accessibility and promotion to increase sale.**

**Problem 2 - (Download**[**Data**](https://olympus.mygreatlearning.com/courses/67575/files/4631209/download?verifier=kzVSN9kUrVSLntf0rHmSKgJOdF1DjJWY0PIbQ0Ji&wrap=1)**)**

**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).**

|  | **Gender** | **Age** | **Class** | **Major** | **Grad Intention** | **GPA** | **Employment** | **Salary** | **Social Networking** | **Satisfaction** | **Spending** | **Computer** | **Text Messages** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | Female | 20 | Junior | Other | Yes | 2.9 | Full-Time | 50.0 | 1 | 3 | 350 | Laptop | 200 |
| **1** | Male | 23 | Senior | Management | Yes | 3.6 | Part-Time | 25.0 | 1 | 4 | 360 | Laptop | 50 |
| **2** | Male | 21 | Junior | Other | Yes | 2.5 | Part-Time | 45.0 | 2 | 4 | 600 | Laptop | 200 |
| **3** | Male | 21 | Junior | CIS | Yes | 2.5 | Full-Time | 40.0 | 4 | 6 | 600 | Laptop | 250 |
| **4** | Male | 23 | Senior | Other | Undecided | 2.8 | Unemployed | 40.0 | 2 | 4 | 500 | Laptop | 100 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **57** | Female | 21 | Senior | International Business | No | 2.4 | Part-Time | 40.0 | 1 | 3 | 1000 | Laptop | 10 |
| **58** | Female | 20 | Junior | CIS | No | 2.9 | Part-Time | 40.0 | 2 | 4 | 350 | Laptop | 250 |
| **59** | Female | 20 | Sophomore | CIS | No | 2.5 | Part-Time | 55.0 | 1 | 4 | 500 | Laptop | 500 |
| **60** | Female | 23 | Senior | Accounting | Yes | 3.5 | Part-Time | 30.0 | 2 | 3 | 490 | Laptop | 50 |
| **61** | Female | 23 | Senior | Economics/Finance | No | 3.2 | Part-Time | 70.0 | 2 | 3 | 250 | Laptop | 0 |

**Q. 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** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Gender** |  |  |  |  |  |  |  |  |
| **Female** | 3 | 3 | 7 | 4 | 4 | 3 | 9 | 0 |
| **Male** | 4 | 1 | 4 | 2 | 6 | 4 | 5 | 3 |

**2.1.2. Gender and Grad Intention:**

| **Grad Intention** | **No** | **Undecided** | **Yes** |
| --- | --- | --- | --- |
| **Gender** |  |  |  |
| **Female** | 9 | 13 | 11 |
| **Male** | 3 | 9 | 17 |

**2.1.3. Gender and Employment:**

| **Employment** | **Full-Time** | **Part-Time** | **Unemployed** |
| --- | --- | --- | --- |
| **Gender** |  |  |  |
| **Female** | 3 | 24 | 6 |
| **Male** | 7 | 19 | 3 |
|  |  |  |  |

**2.1.4. Gender and Computer:**

| **Computer** | **Desktop** | **Laptop** | **Tablet** |
| --- | --- | --- | --- |
| **Gender** |  |  |  |
| **Female** | 2 | 29 | 2 |
| **Male** | 3 | 26 | 0 |

**Q. 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?**

**Total Males are 29**

**Total sample population is 62**

**Hence probability is 29/62 = 0.46774193548387094**

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

**Total Males are 33**

**Total sample population is 62**

**Hence probability is 33/62 = 0.532258064516129**

**Q. 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.**

**Conditional probability for Retailing/Marketing in Male = 4/7 = 0.5714285714285714**

**Conditional probability for Economics/Finance in Male = 4/11 = 0.36363636363636365**

**Conditional probability for Management in Male = 6/10 = 0.6**

**Conditional probability for Other in Male = 4/7 = 0.5714285714285714**

**Conditional probability for  Accounting in Male = 4/7 = 0.5714285714285714**

**Conditional probability for  International Business in Male = 2/6 = 0.3333333333333333**

**Conditional probability for  CIS in Male = 1/4 = 0.25**

**Conditional probability for** Undecided **in Male = 3/3 = 1.0**

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

**CMSU.**

**Conditional probability for Retailing/Marketing in Female = 9/14 = 0.6428571428571429**

**Conditional probability for  Economics/Finance  in Female = 7/11 = 0.6363636363636364**

**Conditional probability for  Management** **in Female = 4/10 = 0.4**

**Conditional probability for  Other** **in Female = 3/7 = 0.42857142857142855**

**Conditional probability for  Accounting** **in Female = 3/7 = 0.42857142857142855**

**Conditional probability for  International Business in Female = 4/6 = 0.6666666666666666**

**Conditional probability for  CIS in Female = 3/4 = 0.75**

**Conditional probability for  Undecided in Female = 0/3 = 0.0**

**Q. 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.**

Total number of Male intend to do graduation = 17

Total sample population = 62

**Therefore probability = 17/62 = 0.27419354838709675**

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

Total number of Female not having a laptop = 4

Total sample of population = 62

Therefore probability = 4/62 = **0.06451612903225806**

**Q. 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?**

**Total number of Male = 29**

**Number of Male having Full time employment = 7**

**Therefore probability = 29/62+7/62-7/29 = 0.3392658509454949**

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

Total number of Female = 62

Number of Female intending to do graduation in international business or management = 4

Therefore probability = 8/62 = **0.12903225806451613**

**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?**

| **Grad Intention** | **No** | **Yes** |
| --- | --- | --- |
| **Gender** |  |  |
| **Female** | 9 | 11 |
| **Male** | 3 | 17 |

**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?**

| **GPA** | **2.3** | **2.4** | **2.5** | **2.6** | **2.8** | **2.9** | **3.0** | **3.1** | **3.2** | **3.3** | **3.4** | **3.5** | **3.6** | **3.7** | **3.8** | **3.9** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Gender** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Female** | 1 | 1 | 2 | 0 | 1 | 3 | 5 | 2 | 4 | 3 | 2 | 4 | 1 | 2 | 1 | 1 |
| **Male** | 0 | 0 | 4 | 2 | 2 | 1 | 2 | 5 | 2 | 2 | 5 | 2 | 2 | 0 | 0 | 0 |

**From the above data presentation we can easily see that total number of students having GPA less than 3 = 17.**

**Hence probability one chosen = 17/62 =** **0.27419354838709675**

**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** | **25.0** | **30.0** | **35.0** | **37.0** | **37.5** | **40.0** | **42.0** | **45.0** | **47.0** | **47.5** | **50.0** | **52.0** | **54.0** | **55.0** | **60.0** | **65.0** | **70.0** | **78.0** | **80.0** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Gender** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Female** | 0 | 5 | 1 | 0 | 1 | 5 | 1 | 1 | 0 | 1 | 5 | 0 | 0 | 5 | 5 | 0 | 1 | 1 | 1 |
| **Male** | 1 | 0 | 1 | 1 | 0 | 7 | 0 | 4 | 1 | 0 | 4 | 1 | 1 | 3 | 3 | 1 | 0 | 0 | 1 |

**Number of Male earning more than 50 = 14**

**Total number of Male = 29**

**Probability of Male earning 50 or more = 14/29 = 0.4827586206896552**

**Number of Female earning more than 50 = 18**

**Total number of Female = 33**

**Probability of Male earning 50 or more = 18/33 = 0.5454545454545454**

**Q. 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.**

|  | **Age** | **GPA** | **Salary** | **Social Networking** | **Satisfaction** | **Spending** | **Text Messages** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **count** | 62.000000 | 62.000000 | 62.000000 | 62.000000 | 62.000000 | 62.000000 | 62.000000 |
| **mean** | 21.129032 | 3.129032 | 48.548387 | 1.516129 | 3.741935 | 482.016129 | 246.209677 |
| **std** | 1.431311 | 0.377388 | 12.080912 | 0.844305 | 1.213793 | 221.953805 | 214.465950 |
| **min** | 18.000000 | 2.300000 | 25.000000 | 0.000000 | 1.000000 | 100.000000 | 0.000000 |
| **25%** | 20.000000 | 2.900000 | 40.000000 | 1.000000 | 3.000000 | 312.500000 | 100.000000 |
| **50%** | 21.000000 | 3.150000 | 50.000000 | 1.000000 | 4.000000 | 500.000000 | 200.000000 |
| **75%** | 22.000000 | 3.400000 | 55.000000 | 2.000000 | 4.000000 | 600.000000 | 300.000000 |
| **max** | 26.000000 | 3.900000 | 80.000000 | 4.000000 | 6.000000 | 1400.000000 | 900.000000 |

**Mode for GPA = 3.0, 3.1, 3.4**

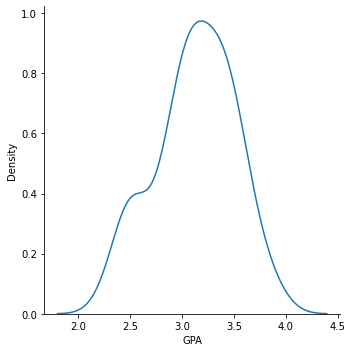
**Mode for Salary= 40**

**Mode for Spending = 500**

**Mode for Text Messages = 300**

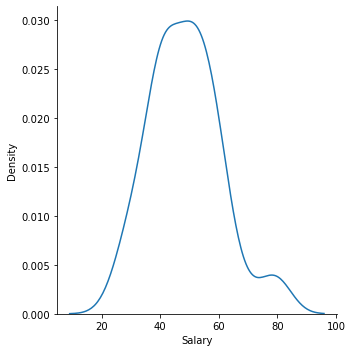
**GPA**

**We can see that mean and mode for GPA is close to each other but median is a little further but it is in bell curve hence it is close to normal distribution.**



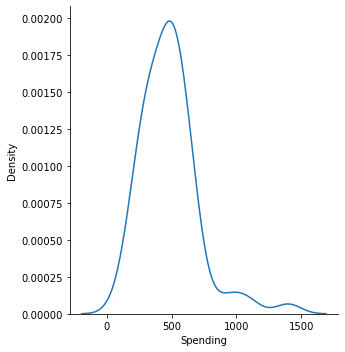
**Salary**

**We can see that mean, median for Salary are close to each other but mode is a little further but forms a bell curve hence normal distribution.**



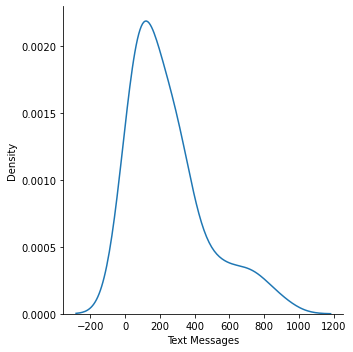
**Spending**

**We can see that mean, median and mode for Spending is close to each other but median is a little further hence it is normal distribution but it has many outliers.**

****

**Text Messages**

**We can see that mean, median and mode for Salary are far to each other hence it is skewed distribution.**

****

**Female student are more likely to opt for graduation and have more earning than their Male counterpart.**

**Trends mostly follow a normal distribution.**

**Females mostly choose Retail/Marketing and Economics/Finance.**

**Males choose Management and Retail/Marketing**

**Problem 3 (**[**Download Data**](https://olympus.mygreatlearning.com/courses/67575/files/4631208/download?verifier=yTey2g1Wj4JMdEOAnw0bOq0hkRUbIzIaS2fX2tkb&wrap=1)**)**

**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 (**[**A & B shingles.csv**](https://olympus.mygreatlearning.com/courses/67575/files/4317942/download?verifier=HMuRUxghivVnnk5VPUKphnGjW1wkeomMyRiU1bVF&wrap=1)**) includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.**

|  | **A** | **B** |
| --- | --- | --- |
| **0** | 0.44 | 0.14 |
| **1** | 0.61 | 0.15 |
| **2** | 0.47 | 0.31 |
| **3** | 0.30 | 0.16 |
| **4** | 0.15 | 0.37 |
| **5** | 0.24 | 0.18 |
| **6** | 0.16 | 0.42 |
| **7** | 0.20 | 0.58 |
| **8** | 0.20 | 0.25 |
| **9** | 0.20 | 0.41 |
| **10** | 0.26 | 0.17 |
| **11** | 0.14 | 0.13 |
| **12** | 0.33 | 0.23 |
| **13** | 0.13 | 0.11 |
| **14** | 0.72 | 0.10 |
| **15** | 0.51 | 0.19 |
| **16** | 0.28 | 0.22 |
| **17** | 0.39 | 0.44 |
| **18** | 0.39 | 0.11 |
| **19** | 0.25 | 0.11 |
| **20** | 0.16 | 0.31 |
| **21** | 0.20 | 0.43 |
| **22** | 0.22 | 0.26 |
| **23** | 0.42 | 0.18 |
| **24** | 0.24 | 0.44 |
| **25** | 0.21 | 0.43 |
| **26** | 0.49 | 0.16 |
| **27** | 0.34 | 0.52 |
| **28** | 0.36 | 0.36 |
| **29** | 0.29 | 0.22 |
| **30** | 0.27 | 0.39 |
| **31** | 0.40 | NaN |
| **32** | 0.29 | NaN |
| **33** | 0.43 | NaN |
| **34** | 0.34 | NaN |
| **35** | 0.37 | NaN |

**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.**

**P value for A =** 0.07477633144907513

As P value is 0.0748 > 0.05 we cannot conclude that the mean value of sample A is less than 0.35 pounds per 100 square feet.

P value for B = 0.0020904774003191826

As P value is 0.0021 < 0.05 we can conclude that the mean value of sample B is not less than 0.35 pounds per 100 square feet.

**Q. 3.2 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?**