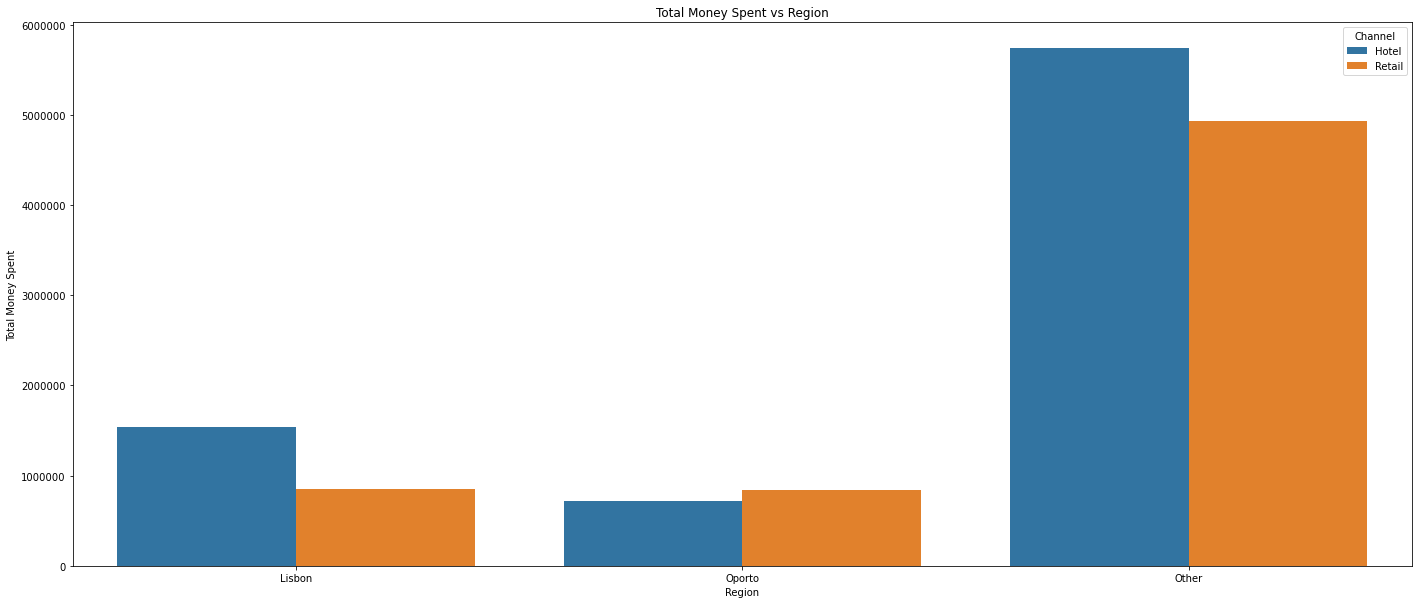
**Wholesale Customers Analysis (**[**Download Data**](https://olympus.greatlearning.in/courses/33031/files/2161312/download?verifier=BID0JlztIHpYpuAakKf4V1V0CcbtQla9bPYVx58V&wrap=1)**)**

**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).

* 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?**

Region wise summation of the data gives us an insight about the high and low selling Region and Channel. Below is the bar graph of the same.



**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**

**Across regions and channels the results were different, as the Standard Deviation was high, a more standard approach of relative dispersion, the CV (Coefficient of Variance) is calculated for Lisbon, Oporto and Other Regions and their channels is calculated.**

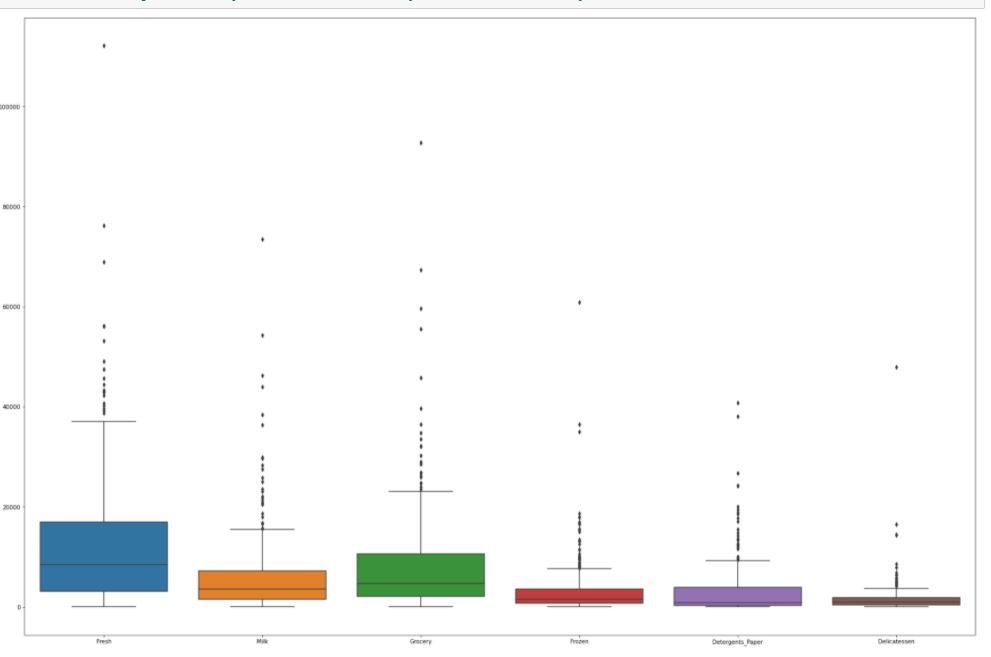
**It is understood that, Lisbon’s Hotel channel for all items the sales were inconsistent, Oporto’s Hotel and Retail Channels have inconsistencies, Other regions Retail channel is consistent compared to the Other regions Hotel channel.**

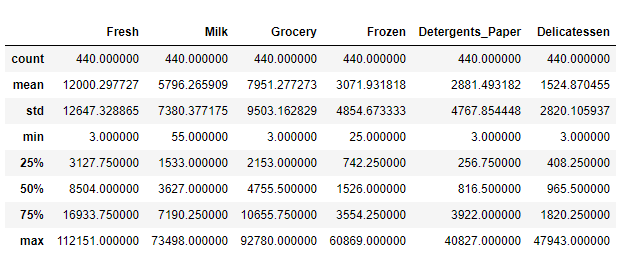
* 1. **On the basis of a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?**

**Of all the items Delicatessen item has Worst CV at 1.84 and hence making it least consistent and Of all the items Fresh item has less CV at 1.05 hence making it most consistent.**

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

Yes, there are huge outliers for Fresh, Milk, Grocery & Outliers do exist for Frozen and Delicatessen as well. Outliers will be the max values mentioned by the describe function.





* 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**

**Lisbon's Hotel Channel sales, Opportos Hotel and Retail Channel Sales, Other Region's Hotel Channel Sales need improvement,**

**they were observed to have more variance compared to their corresponding Channels. Also, Delicatessen items are not selling**

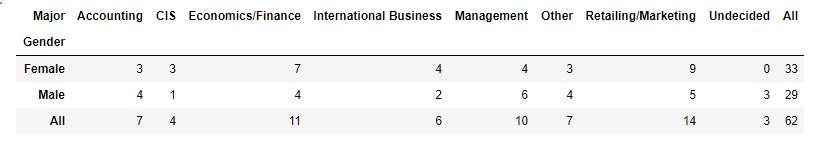
**consistently and they have worst variance across all items. Out of all Fresh items seem to main a consistent sales from customer.**

Problem 2 - (Download [Data](https://olympus.greatlearning.in/courses/33031/files/2161311/download?verifier=RQ6UFHXqGVNPx99o5t9fPQRqY9KNK8X9AKjgii5G&wrap=1))

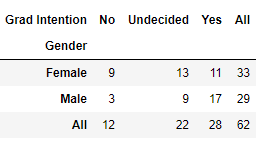
The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. CMSU creates and distributes a survey of 14 questions and receives responses from 62 undergraduates (stored in the **Survey** data set).

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

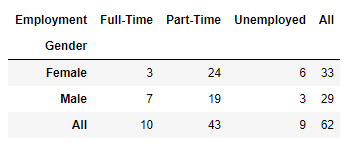
**2.1.1. Gender and Major**



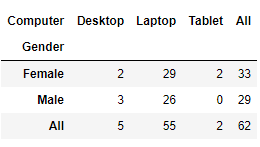
**2.1.2. Gender and Grad Intention**



**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?**

**Total Population is 62 and probability of selecting a male is 46.77%**

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

**Total Population is 62 and probability of selecting a female is 53.23%**

**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.**

**Percentage of male who chose Accounting as major is 14%**

**Percentage of male who chose CIS as major is 3%**

**Percentage of male who chose Economics/Finance as major is 14%**

**Percentage of male who chose International Business as major is 7%**

**Percentage of male who chose Management as major is 21%**

**Percentage of male who chose Other as major is 14%**

**Percentage of male who chose Undecided as major is 10%**

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

**Percentage of female who chose Accounting as major is 9%**

**Percentage of female who chose CIS as major is 9%**

**Percentage of female who chose Economics/Finance as major is 21%**

**Percentage of female who chose International Business as major is 12%**

**Percentage of female who chose Management as major is 12%**

**Percentage of female who chose Other as major is 9%**

**Percentage of female who chose Undecided as major is 0%**

**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.**

**The randomly chosen student is a male and intends to graduate is 27.42%**

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

**The randomly chosen student is a male and female and does NOT have a laptop 6.452%**

**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?**

**P(A) + P(B) – P(A & B)**

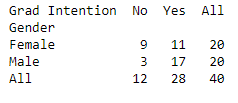
**probability that a randomly chosen student is either a male or has full-time employment is 51.61%**

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

**P(A) + P(B)**

Probability for female student is randomly chosen, she is majoring in international business or management is 12.903%

**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**

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

**gpaLessThan3/Total**

Probability of a student GPA to be less than 3 is 27.42%

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

**Men with Salary greater than 50/Total**

Probability that a randomly selected male earns 50 or more is 22.58%

**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**.

GPA:

Mean, Median and Mode are placed near, hence suggesting a normal distribution. GPA is symmetrically placed from mean with 1, 2 and 3 std dev from mean are in the range of Normal Distribution Values

Salary:

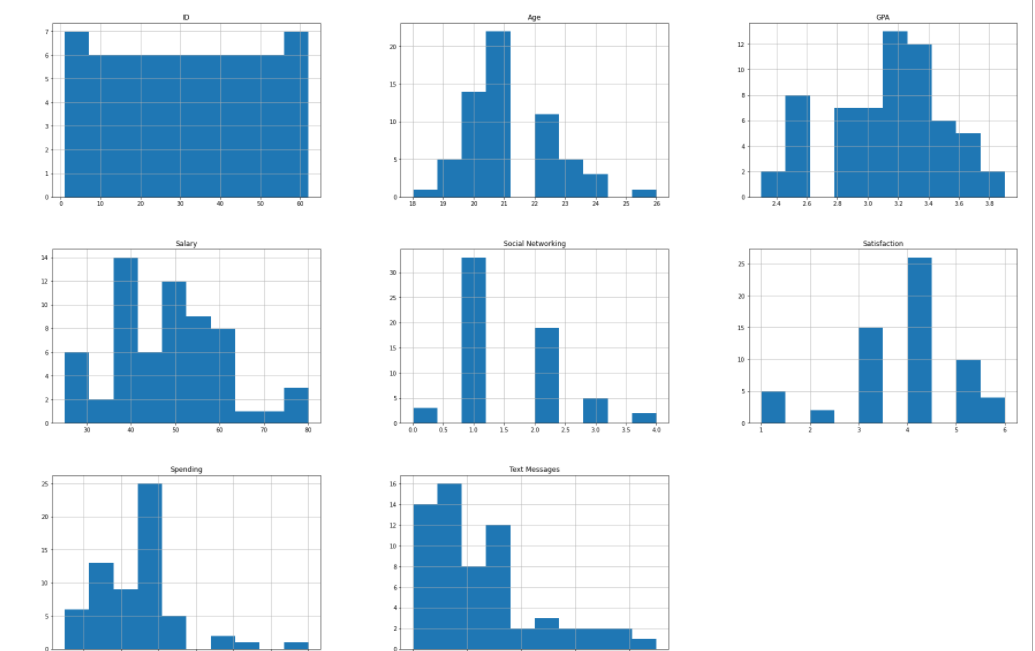
Mean, Median and Mode are distant and hence it suggests it's not a normal distribution. With 80% data in 1 std dev, the bell curve will be spread to left and right, instead of it's 68% range.

Spending:

Mean, Median and mode are not identical and the distribution is not a proper bell curve.

Text Messages:

Mean, Median and mode are not identical and the distribution is not a proper bell curve.



Problem 3 ([Download Data](https://olympus.greatlearning.in/courses/33031/files/2161310/download?verifier=M0IstI77NgDlE1Nlgg6KGVXcURSMjfbWo2WOPeiw&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 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.

The file ([A & B shingles.csv](https://olympus.greatlearning.in/courses/33031/files/1739538/download?verifier=qc06FoSSglTpSQtcE6uvfhBo776FSpWgstakWMbJ&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.**

We fail to reject null hypothesis for A as p\_value is 0.07477633144907513 and Mean moisture for A will be <= 0.35

Null hypothesis for B is rejected & Mean moisture for B will be > 0.35

**3.2 Do you think that the population mean for shingles A and B are equal? Form the hypothesis and conduct the test of the hypothesis.**

We fail to reject null hypothesis as p\_value is 0.1008748285917653 At 95% confidence interval, we can say that the mean moisture content will be less than or equal to 0.35

**Formula for 1 sample T Test:**

**xBar – mU/(s/sqrt(n)) ~ test\_statistic**

**Use TDIST of excel to get the PValue**

**Formula for 2 sample T Test:**

**X1Bar- x2Bar/sqrt(s1^2/n1 + s2^2/n2) ~ t(n1+n2 -2)**

**Use TDIST to get the P Value.**

**What assumption do you need to check before the test for equality of means is performed?**

Assumptions to perform a valid 2-sample T Test:

1. Sample data of the series should be independent. i.e. one observation should not affect other.

2. Sample data values of each series are continuous.

3. Sample data in each series should be a random sample from the population.

4. Variances (sigma^2) must be equal for the 2 series i.e. sigma1^2 - sigma2^2 = 0.

5. Sample data in each series is normally distributed.