

SMDM PROJECT REPORT

DSBA



February 6, 2022

Brijal Panchal

Contents

[Problem 1 3](#_Toc95002175)

[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? 3](#_Toc95002176)

[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. 4](#_Toc95002177)

[1.3 On the basis of the descriptive measure of variability, which item shows the most inconsistent behaviour? Which items shows the least inconsistent behaviour? 7](#_Toc95002178)

[1.4 Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments. 8](#_Toc95002179)

[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. 8](#_Toc95002180)

[Problem 2 9](#_Toc95002181)

[2.1. For this data, construct the following contingency tables (Keep Gender as row variable) 9](#_Toc95002182)

[2.1.1. Gender and Major 9](#_Toc95002183)

[2.1.2. Gender and Grad Intention 9](#_Toc95002184)

[2.1.3. Gender and Employment 9](#_Toc95002185)

[2.1.4. Gender and Computer 9](#_Toc95002186)

[2.2. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question: 9](#_Toc95002187)

[2.2.1. What is the probability that a randomly selected CMSU student will be male? 9](#_Toc95002188)

[2.2.2. What is the probability that a randomly selected CMSU student will be female? 9](#_Toc95002189)

[2.3. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question: 10](#_Toc95002190)

[2.3.1. Find the conditional probability of different majors among the male students in CMSU. 10](#_Toc95002191)

[2.3.2. Find the conditional probability of different majors among the female students of CMSU. 10](#_Toc95002192)

[2.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question: 10](#_Toc95002193)

[2.4.1. Find the probability That a randomly chosen student is a male and intends to graduate. 10](#_Toc95002194)

[2.4.2. Find the probability that a randomly selected student is a female and does NOT have a laptop. 11](#_Toc95002195)

[2.5. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question: 11](#_Toc95002196)

[2.5.1. Find the probability that a randomly chosen student is a male or has full-time employment? 11](#_Toc95002197)

[2.5.2. Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management. 11](#_Toc95002198)

[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? 11](#_Toc95002199)

[2.7. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. 12](#_Toc95002200)

[2.7.1. If a student is chosen randomly, what is the probability that his/her GPA is less than 3? 12](#_Toc95002201)

[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. 12](#_Toc95002202)

[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. 12](#_Toc95002203)

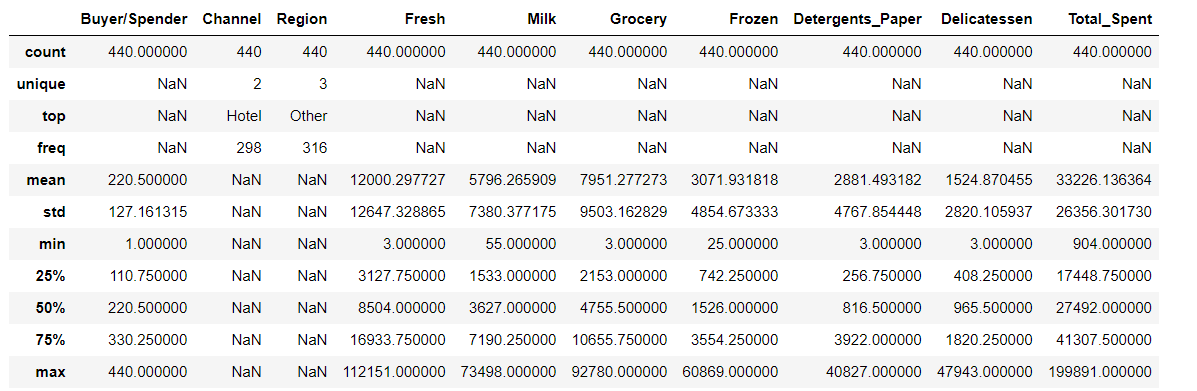
[Problem 3 13](#_Toc95002204)

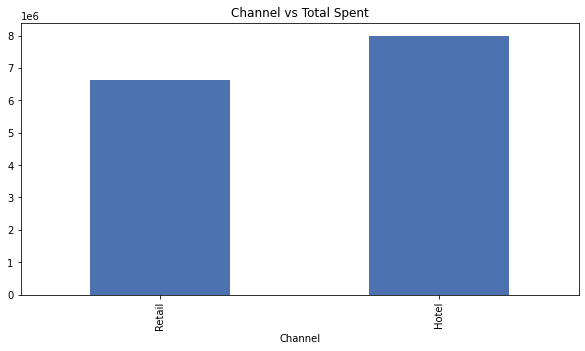
[3.1. Do you think there is evidence that means moisture contents in both types of shingles are within the permissible limits? State your conclusions clearly showing all steps. 13](#_Toc95002205)

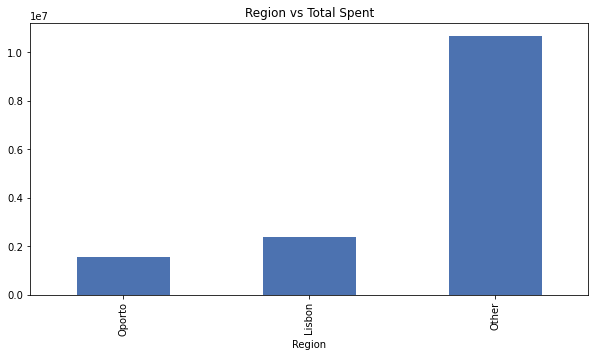
[3.2. Do you think that the population mean for shingles A and B are equal? Form the hypothesis and conduct the test of the hypothesis. What assumption do you need to check before the test for equality of means is performed? 14](#_Toc95002206)

# Problem 1

A wholesale distributor operating in different regions of Portugal has information on annual spending of several items in their stores across different regions and channels. The data consists of 440 large retailers’ annual spending on 6 different varieties of products in 3 different regions (Lisbon, Oporto, Other) and across different sales channel (Hotel, Retail).

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



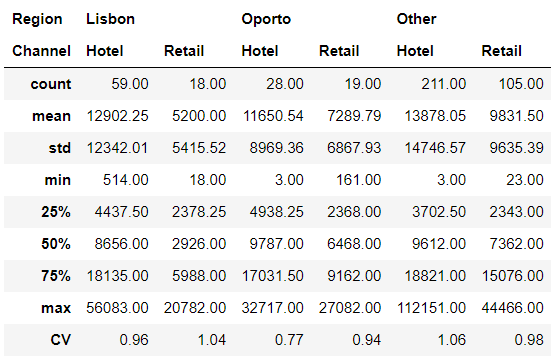


From the bar chart

1. Hotels and Other region spent the most
2. Retails and Oporto region spent less.

## 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.

1. **Fresh**



* 1. On an average highest spent for fresh items are from hotel of other region whereas on an average retailers of Lisbon region spent lowest.
  2. Hotel from Oporto region show the least inconsistent behaviour whereas hotel from other show the most inconsistent behaviour.

1. **Milk**



* 1. On an average highest spent for milk items are from retailer of other region whereas on an average hotel of Oporto region spent lowest.
  2. Retail from Lisbon region show the least inconsistent behaviour whereas hotel from other and Oporto show the most inconsistent behaviour.

1. **Grocery**



* 1. On an average highest spent for grocery items are from Retail of Lisbon region whereas on an average hotel of other region spent lowest.
  2. Retail from Lisbon region show the least inconsistent behaviour whereas hotel from other show the most inconsistent behaviour.

1. **Frozen**



* 1. On an average highest spent for frozen items are from hotel of Oporto region whereas on an average retailers of other region spent lowest.
  2. Retail from Lisbon region show the least inconsistent behaviour whereas hotel from Oporto show the most inconsistent behaviour.

1. **Detergents paper**



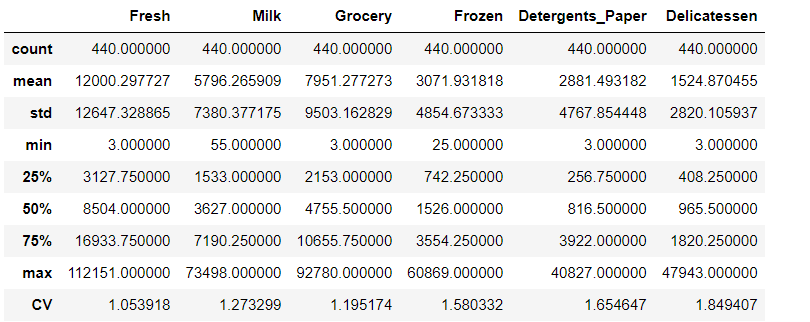
* 1. On an average highest spent for Detergents paper are from retail of Oporto region whereas on an average hotel of Oporto region spent lowest.
  2. Retail from Lisbon region show the least inconsistent behaviour whereas hotel from other show the most inconsistent behaviour.

1. **Delicatessen**



* 1. On an average highest spent for Delicatessen items are from retail of Oporto region whereas on an average hotel of Oporto region spent lowest.
  2. Retail from Lisbon region show the least inconsistent behaviour whereas hotel from other show the most inconsistent behaviour.

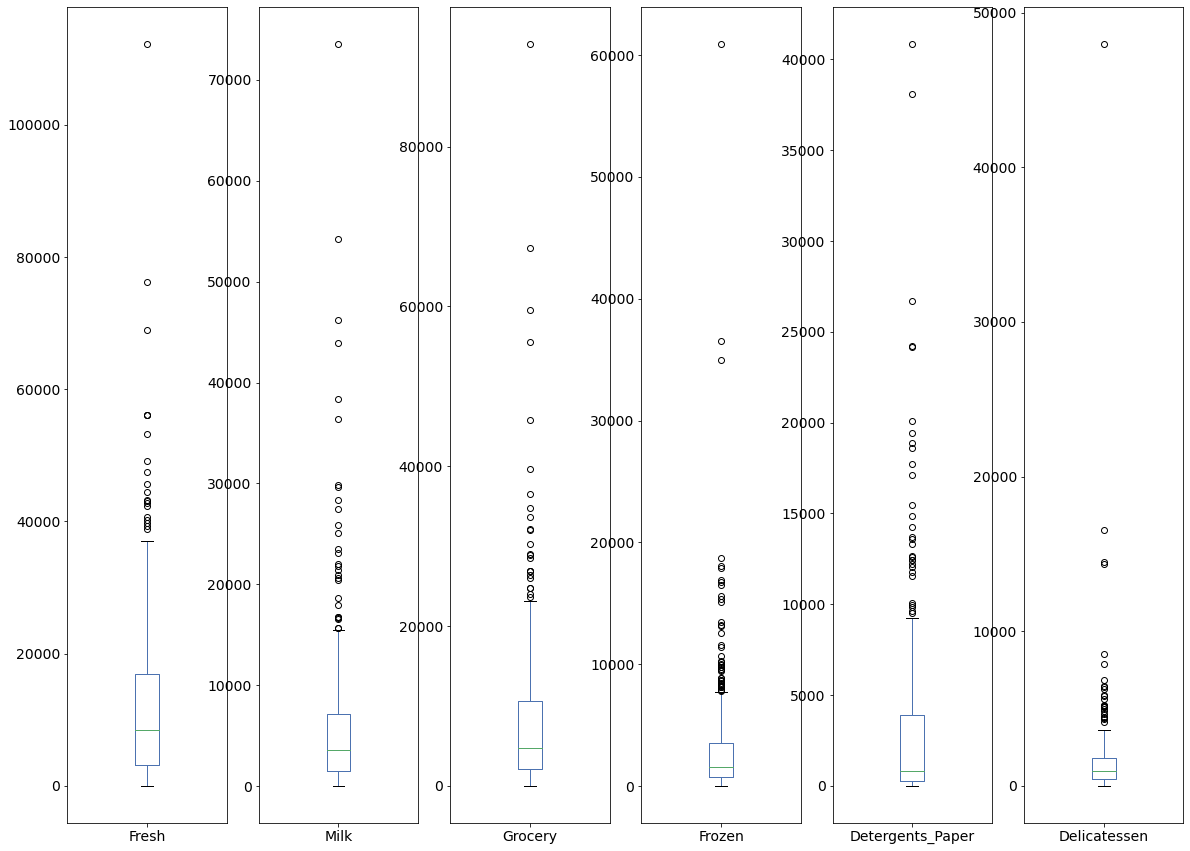
## On the basis of the descriptive measure of variability, which item shows the most inconsistent behaviour? Which items shows the least inconsistent behaviour?



From the table, we can say that,

1. Fresh item show the least inconsistent behaviour in the declared products as the coefficient of variation is the lowest for it.
2. Delicatessen item show the most inconsistent behaviour in the declared products as the coefficient of variation is the highest.

## Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.



As from Box plot, we can see that all items have outliers in data.

## 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.

From the data exploration, we can recommend below suggestions to the wholesale distributor:

1. Hotels and retailers spent less in frozen and delicatessen products across all region as we can see from the box plot and bar plot. Distributer need to explore more to increase the selling of these products.
2. Distributer need to reserve very high stock for Fresh product as we can see from descriptive measure of variability.

# Problem 2

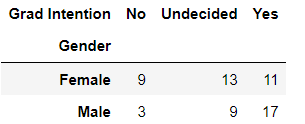
The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. CMSU creates and distributes a survey of 14 questions and receives responses from 62 undergraduates (stored in the Survey data set).

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

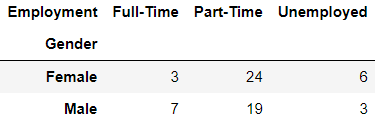
### Gender and Major



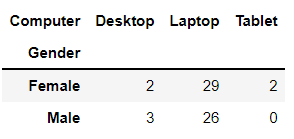
### Gender and Grad Intention



### Gender and Employment



### Gender and Computer



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

### What is the probability that a randomly selected CMSU student will be male?

Probability that a randomly selected CMSU student will be male: 0.46774193548387094

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

Probability that a randomly selected CMSU student will be male: 0.532258064516129

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

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

* Probability of Accounting among the male students in CMSU: 0.13793103448275862
* Probability of CIS among the male students in CMSU: 0.034482758620689655
* Probability of Economics/Finance among the male students in CMSU: 0.13793103448275862
* Probability of International Business among the male students in CMSU: 0.06896551724137931
* Probability of Management among the male students in CMSU: 0.20689655172413793
* Probability of Other among the male students in CMSU: 0.13793103448275862
* Probability of Retailing/Marketing among the male students in CMSU: 0.1724137931034483
* Probability of Undecided among the male students in CMSU: 0.10344827586206896

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

* Probability of Accounting among the Female students in CMSU: 0.09090909090909091
* Probability of CIS among the Female students in CMSU: 0.09090909090909091
* Probability of Economics/Finance among the Female students in CMSU: 0.21212121212121213
* Probability of International Business among the Female students in CMSU: 0.12121212121212122
* Probability of Management among the Female students in CMSU: 0.12121212121212122
* Probability of Other among the Female students in CMSU: 0.09090909090909091
* Probability of Retailing/Marketing among the Female students in CMSU: 0.2727272727272727
* Probability of Undecided among the Female students in CMSU: 0.0

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

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

Probability That a randomly chosen student is a male and intends to graduate = P (Grad Intention ∩ male) x P (male) = 0.27419354838709675

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

Probability that a randomly selected student is a female and does NOT have a laptop = P (Not have Laptop ∩ Female) x P (Female) = 0.06451612903225806

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

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

Probability that a randomly chosen student is a male or has full-time employment = P (Male) + P (Full time employment) - P (Male ∩ Full time employment)

= 0.3876529477196885

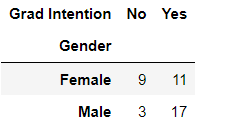
### Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management.

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

= P (International business ∩ Female) + P (Management ∩ Female)

= 0.24242424242424243

## 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?



For Two event to be independent bellow condition must be true.

P (A ∩ B) = P (A) \* P (B)

So,

P (Graduate intention ∩ Female) = P (Graduate intention) \* P (Female)

P (Graduate intention) = 0.7

P (Female) = 0.5

P (Graduate intention) \* P (Female) = 0.35

P (Graduate intention ∩ Female) = 33/62 = 0.275

This is not independent events as probability multiplication of both events is not equal to combined event, so graduate intention and being female are not independent.

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

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

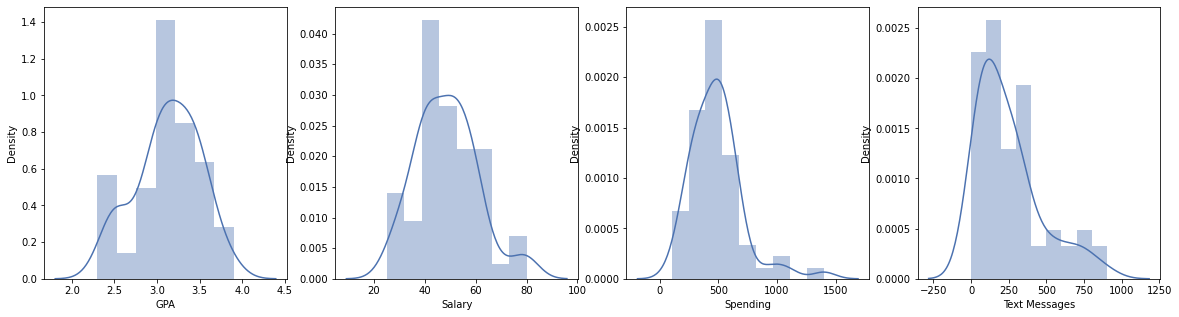
Probability that his/her (GPA is < 3): 0.27419354838709675

### 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.

1. Probability that randomly selected male earns >= 50: 0.24193548387096775

2. Probability that randomly selected female earns >= 50: 0.24193548387096775

## 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.



As per the distribution plot, we can conclude that GPA and Salary follows a normal distribution while Spending and Text Messages are not normally distributed data.

# 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) includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.

## 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.

**Assumptions:**

We assume that the samples are randomly selected, independent and come from a normally distributed population with unknown but equal variances. Level of significance is 5%.

* **Test 1: Mean moister contents in A shingles is greater than 0.35 pounds**

Solution:

**Step 1**: Define NULL and Alternate Hypothesis.

H0 : Mean moister contents in A shingles <= 0.35 pounds

H1 : Mean moister contents in A shingles > 0.35 pounds

**Step 2**: Write the significance level

Let’s assume signification level is 0.05 and it is given that sample size n = 36

**Step 3**: Identify the test to be performed

t-test one sample shall be used as n=36 and single sample is available.

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

p-value: 0.9252236685509249

t-state: -1.473504625338278

**Step 5:** Decide to reject or accept null hypothesis

At the 5% level of significance, we can see p-value of A > 0.05, we do not have enough evidence to prove that the Mean moister contents in A is less than 0.35 pounds. Hence, we fail to reject the null hypothesis.

* **Test 2: Mean moister contents in B shingles is greater than 0.35**

Solution:

**Step 1**: Define NULL and Alternate Hypothesis.

H0 : Mean moister contents in B shingles <= 0.35 pounds

H1 : Mean moister contents in B shingles > 0.35 pounds

**Step 2**: Write the significance level

Let’s assume signification level is 0.05 and it is given that sample size n = 36

**Step 3**: Identify the test to be performed

t-test one sample shall be used as n=36 and single sample is available.

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

p-value: 0.9979095225996808

t-state: -3.1003313069986995

**Step 5:** Decide to reject or accept null hypothesis

At the 5% level of significance, we can see p-value of B > 0.05, we do not have enough evidence to prove that the Mean moister contents in B is less than 0.35 pounds. Hence, we fail to reject the null hypothesis.

## 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**:

We assume that the samples are randomly selected, independent and come from a normally distributed population with unknown but equal variances. Level of significance is 5%.

**Solution**:

**Step 1**: Define NULL and Alternate Hypothesis.

H0 : Mean moister contents in A shingles = Mean moister contents in B shingles

H1 : Mean moister contents in A shingles != Mean moister contents in B shingles

**Step 2**: Write the significance level

Signification level is 0.05.

**Step 3**: Identify the test to be performed

Two-sample t test (including independent t test) shall be used as two independent samples are available.

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

p-value: 0.20174965718353277

t-state: 1.2896282719661123

**Step 5:** Decide to reject or accept null hypothesis

At the level of 5% significance, p-value = 0.2017.Since p-value > 0.05. We have no enough evidence to reject the null hypothesis. That means the population means for shingles A and B are equal at the 5% level of significance.