SMDM Project Report

PGP-DSBA Online

May ’22

Date: 24/07/2022

Contents

1 Problem Statement……………………………………………………………………………………………………………3

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?...5

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…..6 justification for your answer.

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?...........9

1.4 Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments………………………………………………..9

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……………………………………………………………………………………………….12

2 Problem Statement………………………………………………………………………………………………………….13

2.1. For this data, construct the following contingency tables (Keep Gender as row variable)………………………………………………………………………………………………………………….13

2.1.1. Gender and Major…………………………………………………………………………………………13

2.1.2. Gender and Grad Intention……………………………………………………………………………13

2.1.3. Gender and Employment………………………………………………………………………………14

2.1.4. Gender and Computer…………………………………………………………………………………..14

2.2. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:………………………………………………………………..14

2.2.1. What is probability that a randomly selected CMSU student will be male?.....14

2.2.2. What is probability that a randomly selected CMSU student will be female?..14

2.3. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:………………………………………………………………..15

2.3.1. Find the conditional probability of different majors among the male students in CMSU. ………………………………………………………………………………………………………………..15

2.3.2 Find the conditional probability of different majors among the female students of CMSU. ………………………………………………………………………………………………………………..15

2.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question: ……………………………………………………………….15

2.4.1. Find the probability That a randomly chosen student is a male and intends to graduate. ……………………………………………………………………………………………………………….15

2.4.2 Find the probability that a randomly selected student is a female and does NOT have a laptop. ………………………………………………………………………………………………………..16

2.5. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question: ……………………………………………………………….16

2.5.1. Find the probability that a randomly chosen student is a male or has full-time employment? ………………………………………………………………………………………………………..16

2.5.2. Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management…………………………17

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? ……………………………………………………………………………………………………………….…17

2.7. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. …………………………………………………………………...17

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? ………………………………………………………………………………………………………………….17

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. …………………………………………………………………………………………………………………….18

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……………..18

3 Problem Statement…………………………………………………………………………………………………………22

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. …………………………………………………………………………………………………………………….22

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? ……………………………………23

**List of Figures**

Figure 1.1 Descriptive Statistics

Figure 1.2 Region and Channel that spent most

Figure 1.3 Region and Channel that spent least

Figure 1.4 Describing Fresh variety

Figure 1.5 Describing Milk variety

Figure 1.6 Describing Groceryvariety

Figure 1.7 Describing Frozen variety

Figure 1.8 Describing Detergents\_Paper variety

Figure 1.9 Describing Delicatessen variety

Figure 1.10 BoxPlot Fresh

Figure 1.11 BoxPlot Milk

Figure 1.12 BoxPlot Grocery

Figure 1.13 BoxPlot Frozen

Figure 1.14 BoxPlot Detergents\_Paper

Figure 1.15 BoxPlot Delicatessen

Figure 2.1 Contingency Table between Gender and Major

Figure 2.2 Contingency Table between Gender and Grad Intention

Figure 2.3 Contingency Table between Gender and Employment

Figure 2.4 Contingency Table between Gender and Computer

Figure 2.5 Contingency Table (2x2) between Gender and Grad Intention

Figure 2.6 Contingency Table between Gender and GPA < 3

Figure 2.7 Contingency Table between Gender and Salary > 50

Figure 2.8 DistPlot and ProbPlot for GPA

Figure 2.9 DistPlot and ProbPlot for Salary

Figure 2.10 DistPlot and ProbPlot for Spending

Figure 2.11 DistPlot and ProbPlot for Text Messages

**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).

Graphical user interface, application

Description automatically generated

*Figure 1.1 Descriptive Statistics*

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

The above Figure 1 summarizes the dataset “Wholesale+Customers+Data.csv” based on its characteristics.

We can infer that Channel and Region are categorical values, whereas the 6 varieties of products are numerical values.

There are 3 regions analysed here, Oporto, Lisbon and Other.

There are 2 channels analysed here, Hotel and Retail.

In the region category, Other regions occur more frequently than Lisbon or Oporto whereas in the Channel category, Hotel occurs more frequently than Retail.

The average of Fresh is highest and average of Delicatessen is lowest, which may tell us that people prefer to buy fresher items than delicacies.

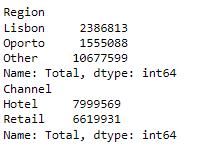
75% of Fresh items spent is less than 17000

Our inference from the above two figures could be that Other Regions spent the most on Fresh items for Hotel and least on Delicatessen.

This has been elaborated based on the different items below.

Summing up a total of the 6 products and taking their max value, will give us the below information

**Which Region and which Channel spent the most?**





*Figure 1.2 Region and Channel that spent most*

Therefore, we can conclude that region-Others and channel-Hotel have spent the most

**Which Region and which Channel spent the least?**

Text

Description automatically generated



*Figure 1.3 Region and Channel that spent least*

Therefore, we can conclude that region-Oporto and channel-Retail have spent the least

**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.**

Fresh:

Chart, bar chart

Description automatically generated

*Figure 1.4 Describing Fresh variety*

Fresh items are sold better to Hotels than Retail businesses.

Other regions lead the business for Hotels in this category.

Milk:

Chart, bar chart

Description automatically generated

*Figure 1.5 Describing Milk variety*

Milk items are sold better to Retail than Hotel businesses.

Other regions lead the business for Retail in this category.

Grocery:

Chart, bar chart

Description automatically generated

*Figure 1.6 Describing Grocery variety*

Grocery items are sold better to Retail than Hotel businesses.

Lisbon leads the business for Retail in this category.

Frozen:

Chart, bar chart

Description automatically generated

*Figure 1.7 Describing Frozen variety*

Frozen items are sold better to Hotels than Retail businesses.

Oporto leads the business for Hotels in this category.

Detergents\_Paper:

Chart, bar chart

Description automatically generated

*Figure 1.8 Describing Detergents\_Paper variety*

Detergents and Paper items are sold better to Retail than Hotel businesses.

Oporto leads the business for Retail in this category.

Delicatessen:

Chart, bar chart

Description automatically generated

*Figure 1.9 Describing Delicatessen variety*

Delicatessen items are sold better to Retail than Hotel businesses.

Lisbon leads the business for Retail in this category.

**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?**

Finding the coefficient of Variation (CV) for the different items will help to decide the consistency of each item. For instance, from our report we acquired the following values for the coefficient of Variation for the 6 items.

Cv\_fresh = 1.053

Cv\_milk = 1.273

Cv\_grocery = 1.195

Cv\_frozen = 1.580

cv\_Detergents\_Paper = 1.654

cv\_Delicatessen = 1.849

From the above result we can infer that Delicatessen items displayed the most inconsistent behaviour as the CV is highest and Fresh items displayed the least inconsistent behaviour as the CV is the lowest.

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

By plotting boxplots, we can identify whether we have outliers in the data. Below we can find the box plots for all the items sold by the distributors.

Fresh:

Chart, box and whisker chart

Description automatically generated

*Figure 1.10 BoxPlot Fresh*

Milk:

Chart, box and whisker chart

Description automatically generated

*Figure 1.11 BoxPlot Milk*

Grocery:

Chart, box and whisker chart

Description automatically generated

*Figure 1.12 BoxPlot Grocery*

Frozen:

Chart

Description automatically generated

*Figure 1.13 BoxPlot Frozen*

Detergents\_Paper:

Chart, box and whisker chart

Description automatically generated

*Figure 1.14 BoxPlot Detergents\_Paper*

Delicatessen:

Chart, scatter chart

Description automatically generated

*Figure 1.15 BoxPlot Delicatessen*

From the plots above, we can infer that there are outliers in the data. All of them are positively skewed and does not follow a Normal Distribution.

For Fresh items we have an outlier point which is above 100,000. Even though Fresh items have the largest outliers, the box plot shows that it is positively skewed, and the values are not very erratic. Hence these outliers are important in a business point of view.

We can also verify this with the descriptive statistics (*Figure 1*) discussed earlier. The 5 point summary showed us that mean of the Fresh items sold by all the retailers put together is greater than the Median itself, which means the plot will 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.**

1. From the analysis, we can suggest the wholesale distributor to focus on selling Delicatessen in Other regions to improve their business from a sales perspective.

2. They can also improve the promotion of fresh items in Oporto and Lisbon regions which will lead to a healthier lifestyle.

3. The data has been divided in 3 regions, viz. Lisbon, Oporto and Others. The Others comprises of the entire population except Lisbon and Oporto, which implies that they have very little distribution in smaller cities of Portugal. Targeting the local population usually brings more sales, and hence I would recommend the business to improve their sales in other cities of Portugal and consolidate a data out of it to delve deeper and analyse their sales and profits.

**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).

Descriptive analysis of the data

Graphical user interface, table

Description automatically generated with medium confidence

1. There are overall 62 students out of which 33 students belong to the female population and 29 students belong to the male population.
2. 3 different classes of students are present : Senior, Junior, Sophomore
3. There are 8 types of Majors
4. 43 students have taken part-time employment
5. 55 students use laptop for their studies

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

**2.1.1. Gender and Major**

Graphical user interface

Description automatically generated

*Figure 2.1 Contingency Table between Gender and Major*

**2.1.2. Gender and Grad Intention**

Table

Description automatically generated

*Figure 2.2 Contingency Table between Gender and Grad Intention*

**2.1.3. Gender and Employment**

Graphical user interface, application

Description automatically generated

*Figure 2.3 Contingency Table between Gender and Employment*

**2.1.4. Gender and Computer**

Table

Description automatically generated

*Figure 2.4 Contingency Table between Gender and Computer*

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

Total number of Students = 62

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

The probability that a randomly selected CMSU student will be male will be the total number of males among the total number of undergraduates.

That will be 29/62 = 0.46774193548387094

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

The probability that a randomly selected CMSU student will be female will be the total number of females among the total number of undergraduates.

That will be 33/62 = 0.532258064516129

**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.**

This can be derived by creating a contingency table between Gender and Major and finding the relevant probabilities.

P (Major | male) = P (major ∩ male)/ P(male)

Probability of Accounting among male students: 0.13793103448275862

Probability of CIS among male students: 0.034482758620689655

Probability of Economics/Finance among male students: 0.13793103448275862

Probability of International Business among male students: 0.06896551724137931

Probability of Management among male students: 0.20689655172413793

Probability of Other Major among male students: 0.13793103448275862

Probability of Retailing/Marketing among male students: 0.1724137931034483

Probability of Undecided among male students: 0.10344827586206896

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

P (Major | female) = P (major ∩ female)/ P(female)

Probability of Accounting among female students: 0.09090909090909091

Probability of CIS among female students: 0.09090909090909091

Probability of Economics/Finance among female students: 0.21212121212121213

Probability of International Business among female students: 0.12121212121212122

Probability of Management among female students: 0.12121212121212122

Probability of Other Major among female students: 0.09090909090909091

Probability of Retailing/Marketing among female students: 0.2727272727272727

Probability of Undecided among female students: 0.0

**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.**

For this we first need to create a contingency table between Gender and Grad Intention

Graphical user interface, text, application, table

Description automatically generated

This can be found out using the below formula

P(Graduate ∩ Male) = (17/29)

Therefore, the probability that a randomly chosen student is a male and intends to graduate : Graduate ∩ Male) = 0.5862068965517241

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

For this we first need to create a contingency table between Gender and Computer

Table

Description automatically generated

This can be found out using the below formula

P(No Laptop ∩ Female) = 1 – P(Laptop ∩ Female) = 1 – (29/33)

Therefore, the probability that a randomly chosen student is a female and does not have a laptop = 0.12121212121212122

**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?**

For this we first need to create a contingency table between Gender and Employment

Graphical user interface, application

Description automatically generated

The probability that a randomly chosen student is a male : 0.46774193548387094

The probability that a randomly chosen student has full-time employment : 0.16129032258064516

The probability that a randomly chosen student is a male or has full-time employment = P(AUB) = P(A) + P(B) - P(A ∩ B): 0.5161290322580645

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

For this we first need to create a contingency table between Gender and Major

Graphical user interface, text, application

Description automatically generated

The conditional probability that given a female student is randomly chosen: 0.532258064516129

The conditional probability that she is majoring in international business: 0.12121212121212122

The conditional probability that she is majoring in management: 0.12121212121212122

The conditional probability that the female student is majoring in international business or management:P(A|B) = P(A∩B) / P(B): 0.45546372819100095

**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?**

Graphical user interface, table

Description automatically generated with medium confidence

*Figure 2.5 Contingency Table (2x2) between Gender and Grad Intention*

Independent events check: P(Female ∩ GradYes) = P(Female)\*P(GradYes)

P(Female): 20/40 = 0.5

P(GradYes): 28/40 = 0.7

P(Female)\*P(GradYes): 0.5 \* 0.7 = 0.35

P(Female ∩ GradYes): 11/40 = 0.275

Therefore, we can conclude that graduate intention and being female are not 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?**

Contingency table for GPA < 3

Table

Description automatically generated

*Figure 2.6 Contingency Table between Gender and GPA < 3*

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

Contingency table for Salary >= 50 :

Graphical user interface

Description automatically generated with medium confidence

*Figure 2.7 Contingency Table between Gender and Salary > 50*

The conditional probability that a randomly selected male earns 50 or more:

P(A|B) = P(A∩B) / P(B) = 0.4827586206896552

The conditional probability that a randomly selected female earns 50 or more:

P(A|B) = P(A∩B) / P(B) = 0.5454545454545455

**2.8.1 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.**

We are going to look at this by plotting 2 plots at the same time for each of the 4 numerical variables. They are DistPlot and ProbPlot.

DistPlot is plotted to look at the type of curve and to see if the data is skewed right or left.

ProbPlot on the other hand is plotted to understand if the data points follow a normal distribution or not (the data points need to approximately fall into a straight line).

Below is the graphical representation for the summary above.

GPA:

Chart, histogram

Description automatically generatedChart, scatter chart

Description automatically generated

*Figure 2.8 DistPlot and ProbPlot for GPA*

From the DistPlot and ProbPlot above, we can conclude that GPA followed a normal distribution because the curve in distplot is a normal curve and the points on the line in the probplot have not deviated much from the line, rather follow the straight line.

Salary:

Chart, histogram

Description automatically generatedChart, scatter chart

Description automatically generated

*Figure 2.9 DistPlot and ProbPlot for Salary*

From the DistPlot and ProbPlot above, we can conclude that Salary followed a normal distribution because the curve in distplot is a normal curve and the points on the line in the probplot have not deviated much from the line, rather follow the straight line.

Spending:

Histogram

Description automatically generated with medium confidenceChart, scatter chart

Description automatically generated

*Figure 2.10 DistPlot and ProbPlot for Spending*

From the DistPlot and ProbPlot above, we can conclude that Spending followed a normal distribution because the curve in distplot is a normal curve and the points on the line in the probplot have not deviated much from the line, rather follow the straight line.

Text Messages:

Chart, histogram

Description automatically generatedChart, scatter chart

Description automatically generated

*Figure 2.11 DistPlot and ProbPlot for Text Messages*

From the DistPlot and ProbPlot above, we can conclude that Text Messages followed a normal distribution because the curve in distplot is a normal curve and the points on the line in the probplot have not deviated much from the line, rather follow the straight line.

###### **2.8.2 Conclusion**

From the survey of 62 students, we can conclude the following:

1. Most of the students wish to take up a part-time job rather than a full-time opportunity
2. The average Salary obtained is close to 50
3. Most of the students prefer laptops for their studies. Only 5 students use a desktop and only 2 students use tablets
4. 50% of the students surveyed belong to the Senior year
5. Even though the dataset contains 8 different majors, Retailing/Marketing is the major that is most preferred by the students

**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/78177/files/5510637/download?verifier=WlrDhRuauwdWHHbuFxyKx9S5BFoEH1SVtxO3yMqb&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.**

Graphical user interface, table

Description automatically generated with medium confidence

Given – Mean moisture content = 0.35

Since the sample sizes are not the same, we need to do a 1 sample t-test for both the samples independently.

For Sample A:

1. Define the Null and Alternate Hypothesis
   1. H0 : Mean moisture >= 0.35
   2. H1 : Mean moisture < 0.35
2. Perform t-test
   1. tStatistic = -1.4735046253382782
   2. pvalue = 0.07477633144907513
3. Significance level will be assumed as alpha = 0.05
4. Pvalue > alpha , this means we cannot reject the null hypothesis. Therefore, we can conclude that the moisture in Sample A is greater than the permissible limit.

For Sample B:

1. Define the Null and Alternate Hypothesis
   1. H0 : Mean moisture >= 0.35
   2. H1 : Mean moisture < 0.35
2. Perform t-test
   1. tStatistic = -3.1003313069986995
   2. pvalue = 0.0020904774003191826
3. Significance level will be assumed as alpha = 0.05
4. Pvalue < alpha , this means we can reject the null hypothesis. Therefore, we can conclude that the moisture in Sample B is lesser than the permissible limit.

**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?**

There are 2 samples and the population size > 30, hence we need to perform a 2 sample t-test

1. Defining the Hypothesis:
   1. H0 : Mean(A) <> Mean(B)
   2. H1 : Mean(A) = Mean(B)
2. Perform 2 sample t-test
   1. t\_statistic = 1.2896282719661123
   2. P Value = 0.2017496571835306
3. Significance level assumed to be alpha = 0.05
4. P\_Value > Alpha, hence we can conclude that we cannot reject the null hypothesis. Therefore, we infer that the population mean for Shingles(A) and Shingles(B) are not equal.