SMDM Project 2021

SEPTEMBER 9

Great Learning
Authored by: ANIMESH HALDER

Great Learning
POWER AHEAD

Content

List of Tables:

List of Figures:

Problem 1: Wholesale Customers Analysis

Problem statement: A wholesale distributor operating in different regions of Portugal has information on the 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 channels (Hotel, Retail).

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?

From the descriptive statistics, we can see that there are 2 unique types of the channel (namely Retail and Hotel), and 3 unique types of region (namely Lisbon, Oporto, and Other) available in the dataset. Among all, Hotel in Other region is the most frequent channel type. The average spending of Fresh is 12000.297 which is the maximum among all the item lists. The minimum quantity of the items varies from 3.0 to 55.0 while the maximum item quantity spent varies from 40827.0 to 112151.0. NaN shows that the values cannot be calculated for that particular variable. Like we can calculate mean for a categorical or object type variable, and in the same way unique value for a numerical variable.

Retail in the other region spent the most, and the total spent is around 0.2 million.

The hotel in Lisbon spent the least, and the total spent is around 900.

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.

A. Fresh: Hotel is consistently the best-preferred channel for the Fresh item in three regions, while the retail channel shows fluctuations. The hotel in Oporto is the stable buyer/spender of the Fresh item as there are no outliers. Except for hotel in Oporto, all other channels the tree regions show the presence of extreme values, and among them, hotel from other region is highly skewed, as there are quite a lot of extreme values.

B. Milk: No channels are consistent over three regions for the item Milk. Hotel over the three regions shows poor demand for Milk. Among them, retail in Oporto shows slightly skewed, but no outliers. While the other channels in the three regions show outliers and among them retail in other regions is highly skewed, there are quite a lot of extreme values.

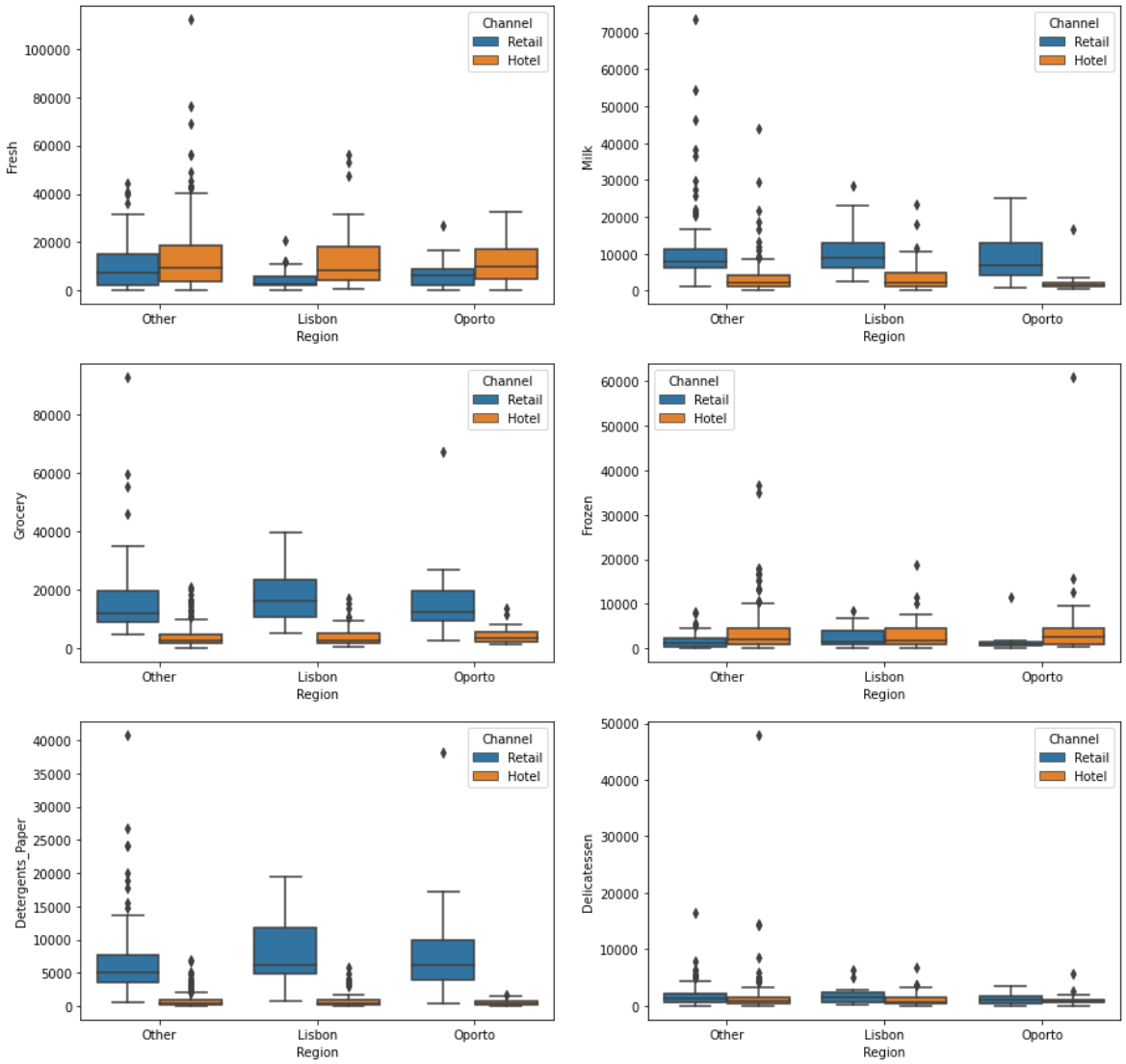

Figure 1: *Boxplot to display the distribution of 6 different varieties of items in three regions by two channels based on a five-number summary.*

C. Grocery: Both the channels show consistency over three regions. Though the count of this item in the retail channel is higher than in the hotel chain, their distribution is almost normal. While the hotel over the three regions is right-skewed. Except for retail in Lisbon, all the channels in the three regions shows the presence of outliers.

D. Frozen: its demand is consistent in the hotel rather than retail channels over three regions. The hotel in Oporto shows normal distribution while the retail is very low. Every channel in the three regions shows few outliers.

E. Detergents_Paper: The demand for this item in the retail channel over the three regions are higher than the hotel channel. The distribution of retails in Lisbon shows no outlier and Oporto shows one outlier respectively. The retail in other regions shows lots of extreme values and is found highly skewed while in Lisbon and Oporto the distributions are slightly skewed.

F. Delicatessen: The demand for this item is not at all promising in either channel over the three regions. All the plot shows the presence of outliers, except the retail in Oporto.

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

The coefficient of variations (CV) for each item is tabulated in Table 1. The coefficient of variation filter is the statistical parameter to measure the consistency of the samples across all experiments. A high CV value reveals inconsistency among the samples within the group.

***Table 1**: The coefficient of variations of different items spent in three regions.*

Items	Coefficient of Variation
Fresh	1.053918
Milk	1.273299
Grocery	1.195174
Frozen	1.580332
Detergents paper	1.654647
Delicatessen	1.849407

Thus, in the present context, ‘Fresh’ shows the least inconsistent behavior in the declared items as the coefficient of variation is the lowest for it. On the other hand, ‘Delicatessen’ shows the most inconsistent behavior in the declared items as the coefficient of variation is the highest.

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, each item has outliers as shown in Figure 2. The plot shows Fresh item is the most preferable one for though it has lots of extreme values, and found highly skewed. The same behaviour is observed in Grocery, which is the preferable item next to Fresh. No items show the normal distribution and all outliers are found on one side.

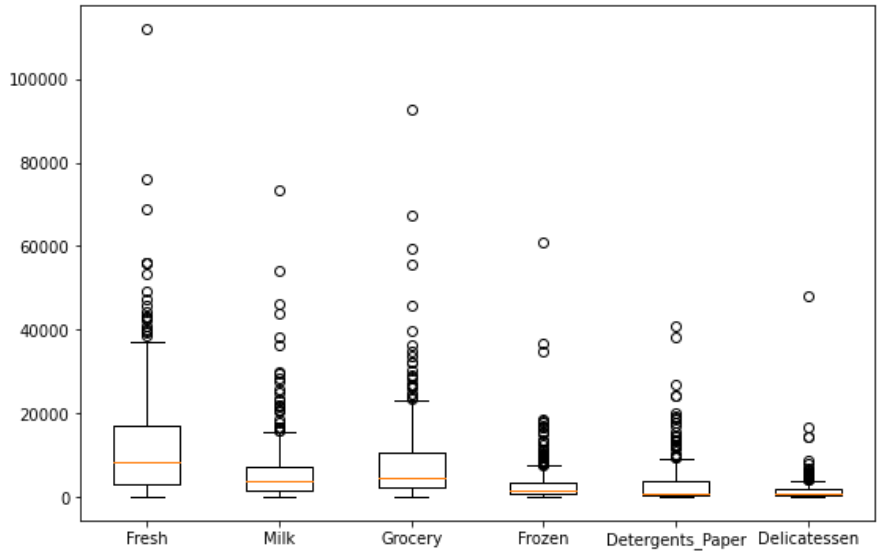

Figure 2: *Boxplot display to verify the presence of the outliers for the 6 different varieties of items.*

1.5 Based on 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.

A. There is much difference in the retail and hotel spent in the three regions. The minimum spent is not attractive and care must be taken to increase the minimum spent especially to the items like Fresh, Grocery, Detergents Paper, and Delicatessen.

B. The spent for all the items shows lots of extreme values, which increase the variability in the dataset. This intern reduces the statistical power.

Problem 2: A Survey

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

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

2.1.1 Gender and Major

Table 2: *Contingency tables constructed by the variables Gender and Major subjects taken by the undergraduate students of Clear Mountain State University.*

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

Table 3: *Contingency tables constructed by the variables Gender and Grad Intention of the undergraduate students of Clear Mountain State University.*

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

2.1.3 Gender and Employment

Table 4: *Contingency tables constructed by the variables Gender and Employment type that the students of Clear Mountain State University secure after graduation.*

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

2.1.4 Gender and Computer

Table 5: *Contingency tables constructed by the variables Gender and Major subjects taken by the undergraduate students of Clear Mountain State University.*

Computer	Desktop	Laptop	Tablet
Gender			
Female	3	3	7
Male	4	1	4

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?

From the given dataset, among 62 undergraduate students male are found 29 heads while female students are 33. Hence, the probability that a randomly selected CMSU student will be male is 0.468

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

From the given dataset, among 62 undergraduate students male are found 29 heads while the rest 33 students are female. Hence, the probability that a randomly selected CMSU student will be female is 0.532

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

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.

Table 6: *Probability of selection of major subjects by the students*

Probability of Major as	Male Students	Female Students
Accounting	0.138	0.091
CIS	0.034	0.091
Economics/Finance	0.138	0.212
International Business	0.069	0.121
Management	0.207	0.121
Other	0.138	0.091
Retailing/Marketing	0.172	0.273
Undecided	0.103	0.000

2.4. Assume that the sample is 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 probability that a randomly chosen student is a male and intends to graduate is 0.274.

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

The probability that a randomly selected student is a female and does NOT have a laptop is 0.758.

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 male or has full-time employment?

The probability that a randomly chosen student is male or has full-time employment is 0.709.

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

The conditional probability that given a female student is randomly chosen, she is majoring in international business or management is 0.242.

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

Table 7: *Contingency tables constructed by the variables Gender and Grad Intention of the undergraduate students of Clear Mountain State University at 2 levels.*

Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

The graduate intention and being female are not independent events, as P(Female)P(Graduate_intention) is not the same as P(Graduate_intention ∩ Female).

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?

From the descriptive Table 8, the probability of a randomly selected student, having a GPA less than 3 is obtained 0.366.

Table 8: *Summary of four continuous variables.*

	count	mean	std	min	25%	50%	75%	max
GPA	62.0	3.13	0.38	2.3	2.9	3.15	3.4	3.9
Salary	62.0	48.55	12.08	25.0	40.0	50.00	55.0	80.0
Spending	62.0	482.02	221.95	100.0	312.5	500.00	600.0	1400.0
Text Messages	62.0	246.21	214.47	0.0	100.0	200.00	300.0	900.0

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.

Table 9: *Contingency tables constructed by the variables Gender and Salary of the students.*

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

From the contingency Table 9, the conditional probability that a randomly selected male earns 50 or more is evaluated as 0.483, while the conditional probability that a randomly selected female earns 50 or more is calculated as 0.545.

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

The distributions for GPA is almost found normal (Figure 3). The distribution of the Salary is near normal, but the presence of the outliers make the distribution curve little rightly skewed. Presence of the many outliers in Spending and Text Message makes the distribution positively skewed.

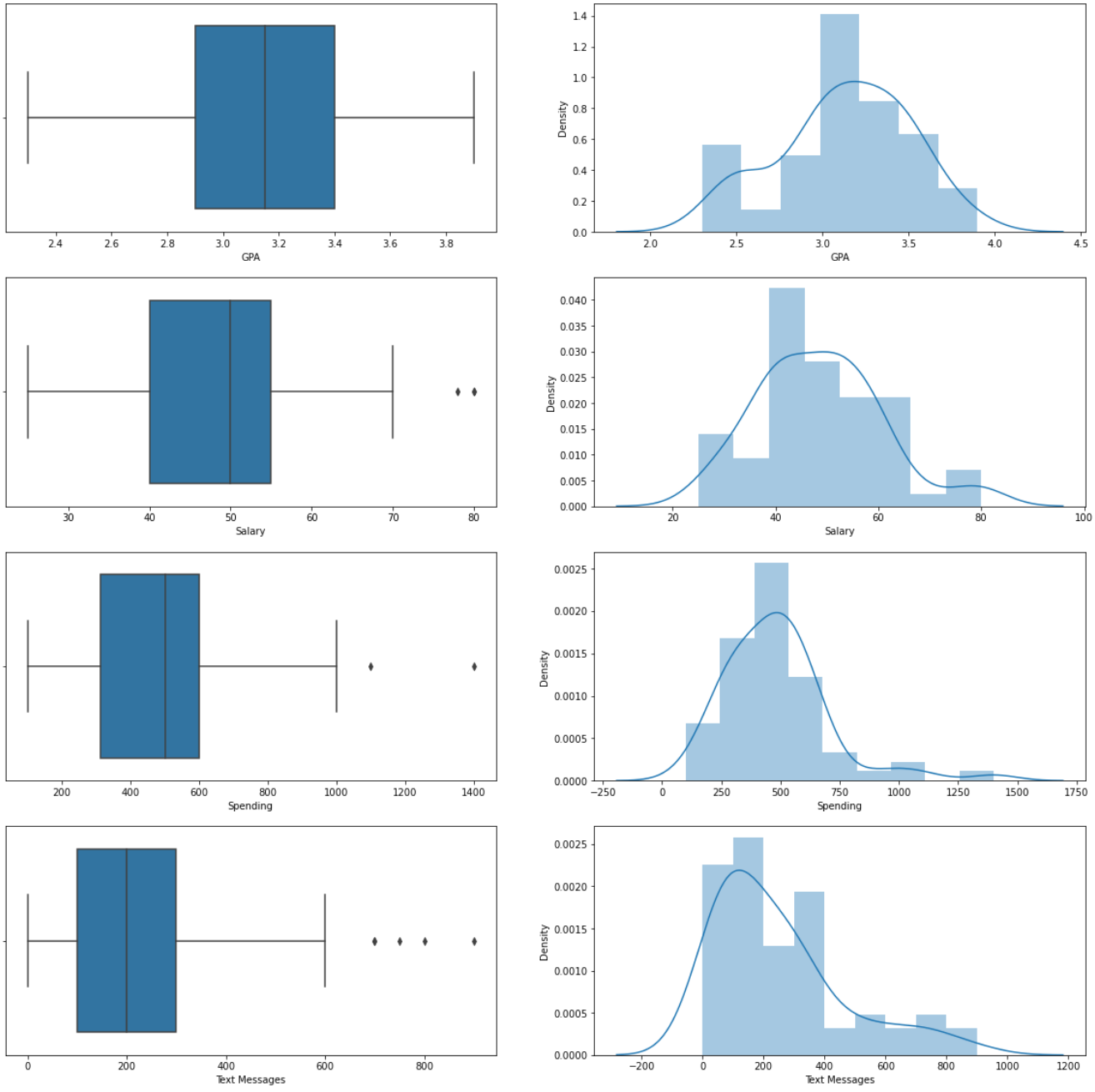

Figure 3: *Distribution of the four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages.*

Problem 3: Product Quality

Problem statement: An important quality characteristic used by the manufacturers of ABC asphalt shingles is the amount of moisture the shingles contain when they are packaged. Customers may feel that they have purchased a product lacking in quality if they find moisture and wet shingles inside the packaging. In some cases, excessive moisture can cause the granules attached to the shingles for texture and colouring purposes to fall off the shingles resulting in appearance problems. To monitor the amount of moisture present, the company conducts moisture tests. A shingle is weighed and then dried. The shingle is then reweighed, and based on the amount of moisture taken out of the product, the pounds of moisture per 100 square feet is calculated. The company would like to show that the mean moisture content is less than 0.35 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'.

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.

H_0 = The mean moisture content ≥ 0.35

H_A = The mean moisture content <0.35

n_A = 36 and n_B = 31

Since no population, the standard deviation is mentioned, so a T-test needed to be performed. The p-Values obtained from the T-test for the A and B types of shingles are 0.075 and 0.002 respectively.

Since p-Value for A-type shingles is > 0.05. So, no enough evidence to prove that the mean moisture content for 'A shingles' will be greater than and equal to 0.35 pounds per 100 square feet. Hence, we fail to reject the null hypothesis.

Similarly, the p-Value for B type shingles is < 0.05. So, the mean moisture content for 'B shingles' will be lesser than 0.35 pounds per 100 square feet. Hence, accept the null hypothesis.

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?

In this case, the hypothesis can be set as,

H_0: Population mean for shingles A = Population mean for shingles B

H_A: Population mean for shingles A ≠ Population mean for shingles B

To perform Hypothesis Testing, the following assumptions must hold,

1. The variables must follow a continuous distribution
2. The sample must be randomly collected from the population
3. The underlying distribution must be normal.
4. For a 2-sample t-test, the population variances of 2-distributions must be equal.

The p-Values obtained from the T-test is 0.202, which is greater than the level of significance value, i.e. 0.05.

So, we do not have enough evidence to reject the null hypothesis. Hence, we cannot refute the assumption that the population mean for shingles A and B are approximately the same.

A. Fresh: Hotel is consistently the best-preferred channel for the Fresh item in three regions, while the retail channel shows fluctuations. The hotel in Oporto is the stable buyer/spender of the Fresh item as there are no outliers. Except for the hotel in Oporto, all other channels the three regions show the presence of