SMDM Project Frequently Asked Questions

(Please read the document before attempting the project)

Question 1.1 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?

Are we expected to provide an answer with respect to the different combinations of Region and Channel?
Analysis and calculation should not be done on a combination of Region & Channel. It should be done individually on Region & Channel.

While comparing the regions and channels should I take the total spending or the average spend?
Use the Total Sum.

Which plot will best show max-min spending? Request your recommendation
Think of Barplot to show the same.

Please help in understanding the problem statement. When we say wholesale distributor has information on annual spending of several items in their stores across different regions and channels does that mean annual spending is amount of money distributor is spending to maintain stock of several items in their stores across different regions and channels or is it the amount of money it is generating when various retailers buy these items. Please clarify.
It's the amount of money the customers are spending

Question 1.2 There are 6 different varieties of items are considered. Do all varieties show similar behavior across Region and Channel?

What is the best check for 'Behaviour' and what does behavior indicate?

Behavior in general talks about the minimum, maximum, IQR, the spread of the variable, etc. Which can be concluded from describe() function only.

Do we need to describe the data based on Region and Channel?
Yes, it should on region and channel.
Typically, just a hint, while comparing among variables whose means show wide differences, CV is used instead of Standard Deviation.

For the above question can I use Skewness as a parameter to analyze
Behavior in general talks about the minimum, maximum, IQR, the spread of the variable, etc. Which can be concluded from describe() function only. Yes, skewness can be one of the parameters as well.

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

This can be calculated using STd or any plotting required? Are boxplots appropriate?
It talks about the descriptive measures of variability like IQR, Standard deviation, Coefficient of Variance, etc. Looking at the problem objective, kindly decide which is the best measure to be computed to find the most inconsistent/least inconsistent behavior.

Please explain what is the expectation

Inconsistency here means variability.
least inconsistent means less varying and most inconsistent means high variation.
so find which item is varying high/low in terms of price/quantity.

Question 1.4 Are there any outliers in the data?

Are there any outliers of the data? are we expected to draw the graph to check the outliers?
For outliers, boxplot is the ideal method.

Can we draw a box plot for all the variables in one set or box plot for all variables to be done separately?
Yes, you can draw box-plots for all the variables in one set/multiple sets.

Question 1.5 On the basis of this report, what are the recommendations?

What kind of recommendations to be provided?
You can put it in your own words. Insights can lead to recommendations. Frame in a way that can help business.

Should we plot heatmap and show how strongly elements are co-related?
The expectation is to state insights and recommendations on the basis of analysis/output/conclusions derived as part of other questions in this problem statement.

What does the report mean here?
Whatever EDA/analysis you have done on the wholesale customer analysis data you have to share your findings and business insights. That is the only recommendation it talks about which is based on all the steps you have done for that data.

Important Communication for Problem 2

Here, using excel spreadsheets for analysis of Problem 2 is also allowed and will be taken into consideration, but we would encourage using python as much as possible.

Question 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

Can I use the crosstab functions to create those
Yes, please go ahead using the same.

Question 2.2 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? What is the probability that a randomly selected CMSU student will be female?

How to approach this?
Firstly create the crosstab of the students by gender and major,employment etc.
Next just calculate the probability for each of the gender like following:
P (Male) = count of males/total count

Question 2.3 Assume that the sample is 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. Find the conditional probability of different majors among the female students of CMSU.

Should we check the unique majors or what does this different major denote here?

Ask is to compute separate probabilities across all majors provided the student is Male.

Should we calculate prob(Accounting.Male/Total.Male) similarly for all the majors?

Ask is to compute separate probabilities across all majors provided the student is Male.

Also, there are entries of undecided in majors, will that be also included?

Calculate for Undecided as well.

Can I use a contingency table and find the answers.

Yes, go ahead use the same and calculate the required ask of probabilities.

Are separate conditional probabilities need to be calculated here for each major?

Each major should be considered.

Question 2.4 Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question: a) Find the probability That a randomly chosen student is a male and intends to graduate. b) Find the probability that a randomly selected student is a female and does NOT have a laptop.

For part B, Is this the formula for the same p(a). p(B/A)?

It is Prob(Female AND Does not have a laptop) = P(F ∩ Lc). Look at the contingency table and derive your answer.

Here intends to graduate shall I consider the only case of yes or I should include undecided also.

Undecided should not be included in this calculation of intent to graduate.

Question 2.5 Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question: a) Find the probability that a randomly chosen student is either a male or has a full-time employment? b) Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management.

Is it P(Female|Internation Business) + P(Female|Management) ??

It should be calculated: Prob(International Business OR Management | Female)

Shall we solve this considering two events mutually exclusive OR assuming marginal probability?

It should be calculated: Prob(International Business OR Management | Female)

Question 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 graduate intention and being female are independent events?

How to remove the 'undecided' from the Intent to Graduate column?

You can create a subset of data and then use it in the cross-tab function as required.

Can you please clear this question and which method should be approached.
Condition to be checked: If being female and graduate intention are independent, the P(F ∩ Yes) = P(F)P(Yes)
Ask from the question is to compute the probability and check for the independence of the events.

Here graduate intention, YES only to be considered for independence, right? Also, as we discarded the x no. of undecided candidates, will my total sample size become (62 - x)?
That's correct. The total sample size would reduce considering this 2*2 table.

Question 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 a) If a student is chosen randomly, what is the probability that his/her GPA is less than 3? b) Find conditional probability that a randomly selected male earns 50 or more. Find conditional probability that a randomly selected female earns 50 or more.

Can you please help with the clarification of the same as which method should be approached?
It can also be solved taking the counts individually and computing the probabilities as per the required conditions.

Why does cdf function gives the wrong answer here? Is it because data is not normally distributed? Its distplot looks like a bell curve to me, so can we not assume it is normally distributed using CLT?
The concept of cdf and pdf is different when it comes to conditional probability. Here the ask is to calculate the absolute probabilities.

It says conditional probability, do we have to find the probability for a random student being selected or random male/female being selected.
Yes on random considering 50 or more specific to male and females

Do we have to find the same by computing z-statistic using normal distribution function or by using the addition/multiplication rule?
There is no need for z-stat, it can be calculated simply by Prob(GPA < 3)

Question 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.

Can you please explain this question? Should we calculate a normal distribution for each variable using the formula or just show a graphical representation?
Along with visual representation, you can calculate mean, median, and mode and test the empirical rule for the conclusion.

Do we have to identify the results using a Histogram and Boxplot or any other method?
A histogram would work. But conclude properly for each of the variables.

Important Communication for Problem 3

Here, we expect learners to consider themselves at a position from the company perspective and not any external entity. The company here already has a process in place where it keeps reducing the moisture content until the moisture content becomes less than 0.35. So it will try to check every time whether the moisture content is still greater than 0.35 pounds per 100 square feet.

So, when we say that the company would like to show that the mean moisture content is less than 0.35 pounds per 100 square feet, it is actually looking at whether the moisture content is still greater than 0.35 pounds per 100 square feet.

Hence, for every moisture test, the claim to check here becomes whether the moisture content is still greater than 0.35 pounds per 100 square feet.

Alternative hypothesis (HA) : mean moisture content > 0.35
And,
Null hypothesis (H0) : mean moisture content <= 0.35

For the A shingles, the null and alternative hypothesis to test whether the population mean moisture content is less than 0.35 pound per 100 square feet is given:
H0 : mean moisture content <=0.35
HA : mean moisture content > 0.35

For the B shingles, the null and alternative hypothesis to test whether the population mean moisture content is less than 0.35 pound per 100 square feet is given:
H0 : mean moisture content <=0.35
HA : mean moisture content > 0.35

Here are the main points:

1. The null hypothesis is the **current status or status quo.** The company's **current status** is that the mean moisture content is less than 0.35. Their current **status quo** needs to be refuted on the basis of very strong evidence. The company does this test on the basis of the assumption that their production process is under control. That is hinted as To monitor the amount of moisture present, the company conducts moisture tests
2. Hence H0: mean moisture content =< 0.35 (since it is current claim) Ha: mean moisture content > 0.35. The company is monitoring its quality control
3. This is a peculiar case of claim and status quo being the same i.e. Ho <= 0.35 while the test is carried out is to test for the opposite i.e. Ha > 0.35
4. The stated null and alternative hypothesis is **irrespective** of the sample values. **This is important to understand**. The highest tolerant value of moisture content is 0.35. The process is much better than the acceptable

value, hence the average content found from the sample is considerably less than the acceptable limit.

5. The rejection region in this case is **still on the right side**. Since the sample average is less than the null value, the p-value of the test will be greater than 0.5. The null hypothesis will not be rejected

The expectations of the learners is to execute the hypothesis test and interpret the result based on the given hypothesis. And based on the given hypothesis, you have to execute the test. You shall be evaluated based on the test and the interpretation of the given hypothesis only.

Question 3.1 Do you think there is evidence that mean moisture contents in both types of shingles are within the permissible limits? State your conclusions clearly showing all steps.

What is the value of alpha in this case?
Please consider alpha as 0.05, in case it is not specified.

Are we supposed to consider each Sample A and Sample B separately (separate data frame for each) and then evaluate scipy.stats.ttest_1samp ???
That is correct.

In question # 3 we need to use T-Test with 2 samples or one sample T-Test for A and B separately?
For 3.1: Run individual 1 sample T-Tests (ttest_1samp) for both "A" & "B". And 3.2 is a 2 sample T-Test (ttest_ind) problem.

What is the difference between 3.1 and 3.2?
Question 3.1 talks about the deriving conclusion for "A" & "B" separately. However, for Question 3.2, it talks about the comparison of means of "A" & "B".

The sample has two sets of data with different sample numbers.With the moisture content given it is one-tailed test. How do I find the tstat value? I've found out the p-value but not sure if I had to divide it by 2

You have to run individual 1 sample T-Tests and divide it by 2 if it has to be a one-sided test. Because by default in Python, ttest_1samp shows the result of 2-sided.

Is it Q of paired T-test, no of measurement in A is 36 and in B it is 31. do we need to drop Nan and conduct a paired T-test

We don't have to do a paired T-test, because A and B are independent samples and also it's not a before and after effect type of question.

Try to implement using nan_policy = 'omit' in the T-test function for the same.

If we just find the mean of both samples separately, that would work for this? or anything else is to be done? It is written "showing all steps", what do you mean?

This question should be solved by an appropriate test after framing the required hypothesis. Hint: T-Test can be used for the same.

Question 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?

How to approach this problem?

This question should be solved by an appropriate test after framing the required hypothesis. 2 sample T-Test (ttest_ind) can be used for the same. Write the relevant assumptions for the same.

How to go about the second part of the question?

The expectation is to write theoretical assumptions.