Quizzes : Module 2 Part 2

Exploratory Data Analysis

In the LAMS Sequence, you have learned the theory behind this module. It is also expected that you have attempted the quizzes embedded within the LAMS Sequence, and have used the “unlimited attempts” opportunity to score 100%. Here are the quiz questions, consolidated with their answers and corresponding feedback. This is for your after-LAMS revision.

Question 1

Suppose the cost of having dinner (for two) at Clark Quay follows a Normal Distribution with Mean = SGD 50 and Standard Deviation (SD) = SGD 15. What is the probability that you will pay LESS than or equal to SGD 65 if you go for dinner at Clark Quay with a friend?

Answer Choice	Verdict	Explanation
Around 0.84 or around 84% chance	Correct	Correct answer. You are looking for probability “equal to or below” Mean + SD = SGD 65, which is around 0.84.
Around 0.5 or around 50% chance	Wrong	Nope. Check the distribution again. 50% should be the probability for less than the Mean = SGD 50.
Around 0.68 or around 68% chance	Wrong	Nope. Be careful. 68% is between Mean - SD to Mean + SD. The question asks something different.
Around 0.16 or around 16% chance	Wrong	Quite close. You are thinking in the right direction. 16% is for being "equal to or above" SGD 65 or Mean + SD.

Reference	Module 2 Extra Topic : Normal Distribution	No specific slide. The overall concept matters.

Question 2

As a follow up of the previous question (Mean = SGD 50 and SD = SGD 15), what is the probability that you will pay more than SGD 100 for a dinner for two at Clark Quay?

Answer Choice	Verdict	Explanation
Less than 0.00135 or less than 0.135% chance	Correct	Correct answer. SGD 100 is more than Mean + 3 * SD, and it is only one side of the tail. Hence half of 0.0027.
Less than 0.0027 or less than 0.27% chance	Almost there ...	Correct line of thought, but 0.27% is for both tails of the distribution. More than SGD 100 is only on one side. Think again.
Less than 0.05 or less than 5% chance	Close, but ...	This is one of the correct answers, but 5% is a huge margin. I am sure you can narrow it down. Check again.
Probability 0 or 0% chance	Wrong	For a Normal Distribution, we can't have strict maximum. There is always a probability for paying more than SGD 100.

Reference	Module 2 Extra Topic : Normal Distribution	No specific slide. The overall concept matters.

Question 3

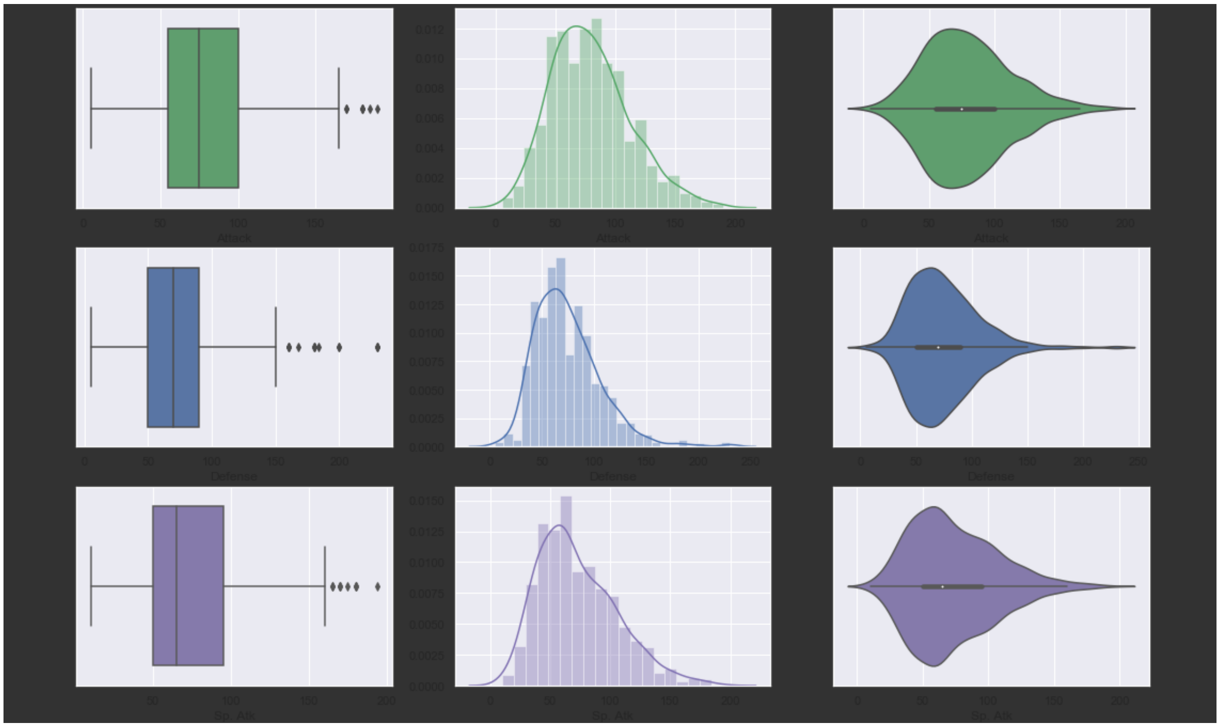
You know the Normal Distribution of cost of dinner (for two) at Clark Quay (Mean = SGD 50 and SD = SGD 15). What do you think 50% of the dinner options (or more), that is, more than or equal to half of the dinner choices at Clark Quay, cost for a dinner for two?

Answer Choice	Verdict	Explanation
Between SGD 35 and SGD 65	Correct	Correct answer. The range is Mean - SD to Mean + SD, and thus more than 50% of the options (around 68%) should fall here.
Less than or equal to SGD 50	Correct	Correct answer. Exactly 50% or half of the options are less than or equal to the Mean in a Normal Distribution.
Between SGD 50 and SGD 80	Wrong	Nope. SGD 50 is the Mean, and SGD 80 is Mean + 2 * SD. This range contains less than 50% of the places. Think again.
Less than or equal to SGD 40	Wrong	Nope. We can't say this as SGD 40 is considerably (SGD 10) less than the Mean, or SGD 50. Thus less than 50% options lie below SGD 40.

Reference	Module 2 Extra Topic : Normal Distribution	No specific slide. The overall concept matters.

Question 4

Study the following three uni-variate distributions (green, blue, magenta). Which one has the highest Median out of the three? Assume that the scale (markers) of the x-axis in each of the following plots is similar -- so, don't worry about the x-values.

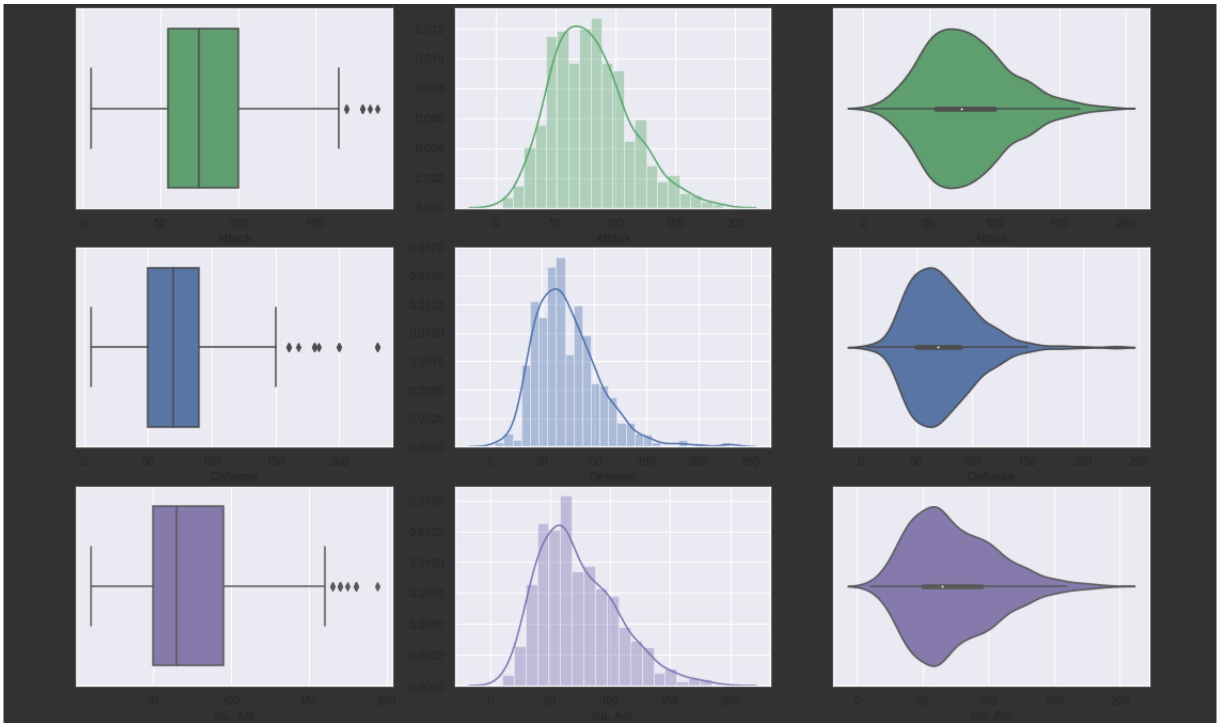


Answer Choice	Verdict	Explanation
The first distribution (green)	Correct	Correct answer. Medians are straight-forward to compare between distributions if the x-axis is the same.
The second distribution (blue)	Wrong	Nope. Check the distributions again. Median is the vertical line in the middle of the Box in a box-plot.
The third distribution (magenta)	Wrong	Nope. Check the distributions again. Median is the vertical line in the middle of the Box in a box-plot.
Impossible to determine as it depends on the outliers	Wrong	Nope. Median is generally not affected too much by outliers. Check the box-plots carefully, once again.

Reference Module 2 Topic 4 : Multi-Variate Exploration Slide 5 and Slide 6

Question 5

Study the following three distributions (green, blue, magenta) once again. Which one of these above three distributions is the most dissimilar to a Normal Distribution?

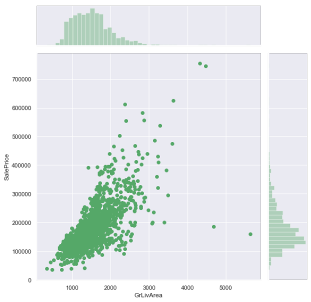


Answer Choice	Verdict	Explanation
The third distribution (magenta).	Correct	Correct answer. This one is the farthest from Normal, that is, the most skewed distribution. Look at the Median.
The second distribution (blue)	Wrong	Not so much. It is definitely not Normal, but not too far (skewed) either. If we drop the outliers, it may actually be Normal. There's another one more dissimilar to Normal.
The first distribution (green)	Wrong	Not so much. It is definitely not Normal, but not too far (skewed) either. There's another one more dissimilar to Normal.

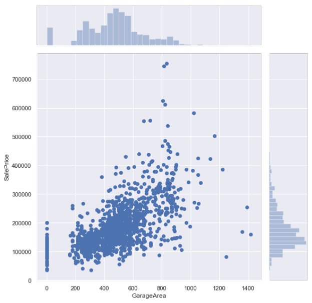
Reference Module 2 Topic 4 : Multi-Variate Exploration Slide 5 and Slide 6

Question 6

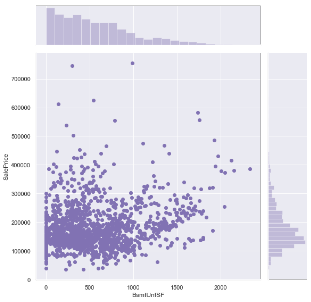
Study the joint-plots, and arrange them in order of Correlation -- highest to lowest.



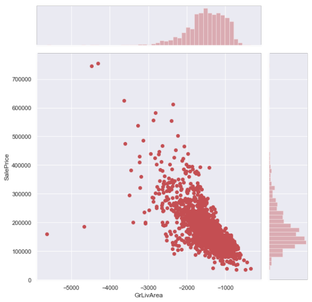
Possibly > 0.7



Possibly ~ 0.5



Possibly < 0.3



Possibly < -0.7

Note that correlation can be both positive and negative (+1 to -1). Absolute value depicts the dependence.

Reference Module 2 Topic 3 : Bi-Variate Exploration Slide 9 and Slide 10

Question 7

Suppose that the "Time students take to complete this LAMS Sequence" and the "Marks students score in this LAMS Quiz" have a correlation of 0.8. What can you infer from this?

Answer Choice	Verdict	Explanation
The data for LAMS Quiz Scores have a linear relationship with the data for Time taken to complete LAMS Sequence.	Correct	Correct answer. The linear relationship is the most you can infer from a high correlation. No "causality" can be claimed.
To score more marks in the LAMS Quiz, one must take a long time to complete the LAMS Sequence.	Wrong	Nope. This is a common misconception. High correlation DOES NOT mean that one variable "causes" the other.
If You scored a high mark in the LAMS Quiz, it is clear that You must have taken a long time to complete the Sequence.	Wrong	Nope. This is a common misconception. You, specifically, may be an outlier or an anomaly in the dataset, and may not follow the norm. High correlation only says that you are "likely" to have taken a long time, but it can't be claimed with certainty.

Reference Module 2 Topic 3 : Bi-Variate Exploration Slide 11 and Slide 12

Question 8

Suppose we look at Sale Prices (SalePrice) of houses in Singapore, and find that it has 0.7 correlation with the General Living Area (GrLivArea) of the houses. What can we infer?

Answer Choice	Verdict	Explanation
General Living Area (GrLivArea) may be an important variable in "predicting" Sale Prices (SalePrice) of houses in Singapore.	Correct	Correct answer. Indeed, we will try to use GrLivArea to predict the SalePrice. We will use the technique of Linear Regression.
General Living Area (GrLivArea) has no linear relationship with Sale Prices (SalePrice) of houses in Singapore.	Wrong	Nope. The correlation 0.7 is quite high, denoting a significant linear relationship between GrLivArea and SalePrice.
Increase in General Living Area (GrLivArea) causes the Sale Prices (SalePrice) of houses in Singapore to go higher.	Wrong	Nope. Once again, "causality" is not implied by correlation. However, we can try to use the high correlation to predict.

Reference	Module 2 Topic 3 : Bi-Variate Exploration	Slide 11 and Slide 12

Question 9

In a multi-variate dataset, we find 30 numeric variables -- one of them is the SalePrice, and the others other general living area, lot area, basement area, garage area, etc. 10 out of the 30 variables have strong positive correlation (above 0.6) with SalePrice, 5 out of the 30 have strong negative correlation (below -0.6) with SalePrice, and others have weak correlation (between 0.2 to -0.2) with SalePrice. What would your next step be?

Answer Choice	Verdict	Explanation
Consider the 10 variables with strong positive correlation (above 0.6) important for predicting SalePrice.	Correct	Correct answer. Strong positive correlation denotes strong linear relationship, hence important as predictors.
Consider the 5 variables with strong negative correlation (below -0.6) important for predicting SalePrice.	Correct	Correct answer. Strong negative correlation denotes strong linear relationship, hence important as predictors.
Can't decide which variables are important to predict SalePrice, and hence, will consider all variables equal.	Wrong	Nope. At least in case of a Linear Regression, strong positive or negative correlation helps in prediction.
Can't decide which variables are important to predict SalePrice, as strong correlation does not imply causality.	Wrong	Nope. You are right that strong correlation does not imply causality, but it sure helps in predicting SalePrice. :-)

Reference	Module 2 Topic 4 : Multi-Variate Exploration	Slide 8 and Slide 9

Question 10

True or False : Strong correlation of SalePrice with a categorical variable (like Type of the House) also denotes strong relationship.

Answer Choice	Verdict	Explanation
False	Correct	Correct answer. First, you have to decide what the definition of "Correlation" is in case of a categorical variable. Then, you have to see if the definition makes sense in terms of a strong relationship. You can't imply strong relationship based on a wrong definition of correlation.
True	Wrong	Nope. First, you have to decide what the definition of "Correlation" is in case of a categorical variable. Then, you have to see if the definition makes sense in terms of a strong relationship. You can't imply strong relationship based on a wrong definition of correlation.

Reference   Module 2 Topic 3 : Bi-Variate Exploration   Slide 11 and Slide 12