

PRACTICE QUESTIONS

CertNexus

DSP-210

Certified Data Science Practitioner

SHARED BY [u/merrymarmot1979](#) on aiexamcerts

EDITION Free Set 1

QUESTIONS 10

SHARED October 15, 2026

EXAM CertNexus DSP-210, Certified Data Science Practitioner

Shared by [u/merrymarmot1979](#) on aiexamcerts

10 practice questions for the CertNexus DSP-210 exam, Certified Data Science Practitioner, each followed by its answer and a short explanation. More free sets and the full practice exam: www.aiexamcerts.com/certnexus/dsp-210. Spotted a mistake? Email support@aiexamcerts.com with the exam code DSP-210 and the question number.

1. A marketing dataset has 35 percent of values missing in a field that the model needs. Which handling choice is most defensible?

- A Remove all rows containing any missing value, whichever field is affected
- B Replace every missing value with the most common value without checks
- C Check why values are missing and how much signal the field carries
- D Keep missing values as text labels so that the model cannot read them

Answer: C

The cause of missingness determines whether imputation is safe, whether the field is worth keeping, or whether collection must be fixed. Replacing values with the most common value can bias results, and dropping every incomplete row can discard most of the dataset.

2. Converting a categorical column into binary indicator columns, one per category, is called one-hot _____.

Answer: encoding

One-hot encoding creates one column per category containing 0 or 1, which lets models use nominal categories without implying an order. Scikit-learn provides this through OneHotEncoder, and it is appropriate when categories have no natural ranking.

3. Two variables have a Pearson correlation of 0.9. What is the safest conclusion?

- A One variable causes the other because the correlation coefficient is this high
- B They move together strongly and linearly, but that alone does not prove causation
- C The relationship must be nonlinear because the coefficient is so large
- D The variables must be removed from every model because they are duplicates of each other

Answer: B

A high linear correlation shows the two variables move together, but a third factor or reverse causation may explain the pattern. Causation needs experimental or causal evidence. A large coefficient indicates linearity rather than nonlinearity, and high correlation does not make both variables duplicates.

4. Logistic regression is a regression method that predicts continuous numeric values despite its name.
- A True
 - B False

Answer: B (False)

Despite its name, logistic regression is a classification method that models the probability of class membership, typically through a logistic function. It does not predict continuous target values. The statement is false.

5. Which two practices make a data visualization easier to interpret accurately? (Choose two.)
- A Using a 3D pie chart with many slices to show small differences
 - B Labeling axes with units and stating what the chart measures
 - C Truncating a bar chart's y-axis to exaggerate small differences
 - D Adding colors without a legend that explains each one
 - E Ordering categories by value so that comparisons are direct

Answer: B, E

Labeled axes with units and a clear statement of the measure let readers interpret values correctly, and ordering categories by value makes comparisons direct. Truncated axes, undefined colors, and 3D pie charts all distort or obscure the comparison.

6. In k-means clustering, what does the parameter k represent?
- A The number of decision trees averaged together in the final forest model
 - B The number of nearest neighbors used to classify each new record in the dataset
 - C The number of clusters the algorithm will form
 - D The learning rate applied to each gradient update during optimization

Answer: C

In k-means, k is the number of clusters the algorithm partitions the data into, and it must be chosen before fitting. The number of neighbors belongs to k-nearest neighbors, a supervised method, which is the tempting confusion here.

7. What is the main purpose of a validation set when tuning hyperparameters?
- A To provide the unbiased final estimate after every choice is locked
 - B To compare hyperparameter settings without using the final test set
 - C To store the trained model artifact for production deployment
 - D To replace the training set once the model has been fitted one time already

Answer: B

The validation set lets you compare candidate hyperparameter settings while the test set stays untouched for the final estimate. Reporting the final estimate on validation data would be optimistic, since those data guided the choices. Storing artifacts and replacing training data are not its role.

8. A model is described as _____ when it performs much better on its training data than on unseen data.

Answer: overfitting (also accepted: overfit, overfitted)

A large gap between training and unseen-data performance is the signature of overfitting, where the model has learned noise or specifics of the training sample. Cross-validation, more data, simpler models, and regularization are the usual remedies.

9. A regression model has a root mean squared error of 4.2 on a target measured in dollars. What does this value mean?
- A The model explains only 4.2 percent of the variance in the target variable
 - B Exactly 4.2 percent of all predictions are wrong by any amount whatsoever
 - C Typical prediction errors are around 4.2 dollars, and large misses count more
 - D The training set contained roughly 4.2 thousand rows drawn from the source table

Answer: C

RMSE is expressed in the target's units and behaves like a typical error size, with large misses weighted more heavily because errors are squared before averaging. It is not a percentage of wrong predictions, a variance explained, or a row count.

10. Which chart is most suitable for showing the distribution of a single numeric variable?
- A A table listing only the category names
 - B A pie chart with one slice per row
 - C A scatter plot of two categorical columns
 - D A histogram

Answer: D

A histogram bins the values of one numeric variable and shows how frequently each range occurs, which reveals skew, modes, and gaps. A pie chart with one slice per row is unreadable, and a scatter plot needs two numeric axes.