Google Discussions

Exam Professional Data Engineer All Questions

View all questions & answers for the Professional Data Engineer exam

Go to Exam

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 243 DISCUSSION

Actual exam question from Google's Professional Data Engineer

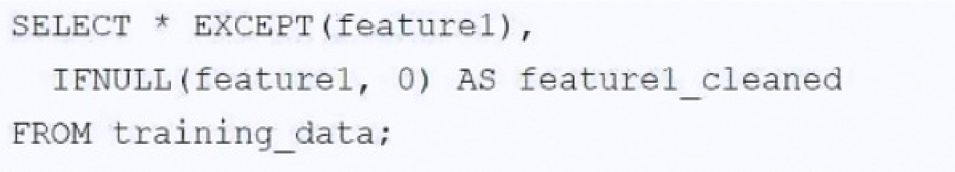
Question #: 243

Topic #: 1

[All Professional Data Engineer Questions]

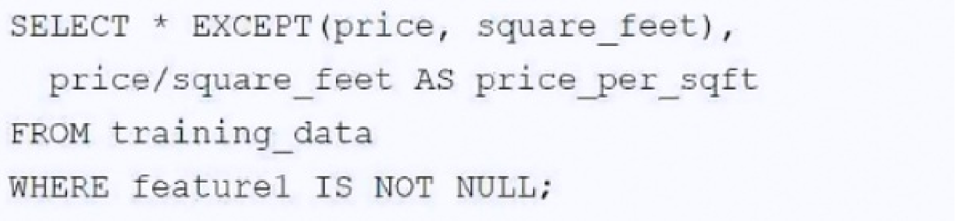
You are preparing data that your machine learning team will use to train a model using BigQueryML. They want to predict the price per square foot of real estate. The training data has a column for the price and a column for the number of square feet. Another feature column called 'feature1' contains null values due to missing data. You want to replace the nulls with zeros to keep more data points. Which query should you use?

A.



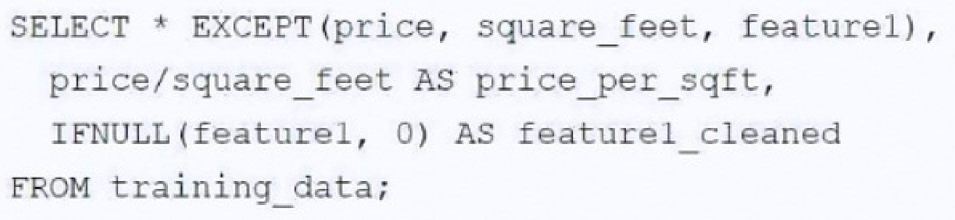
```
SELECT * EXCEPT(feature1),
  IFNULL(feature1, 0) AS feature1_cleaned
FROM training_data;
```

B.



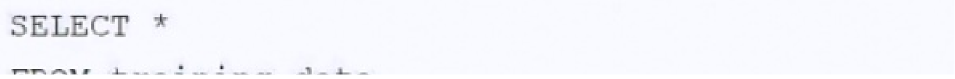
```
SELECT * EXCEPT(price, square_feet),
  price/square_feet AS price_per_sqft
FROM training_data
WHERE feature1 IS NOT NULL;
```

C.



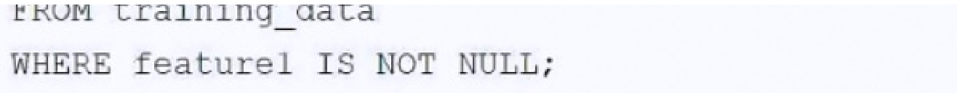
```
SELECT * EXCEPT(price, square_feet, feature1),
  price/square_feet AS price_per_sqft,
  IFNULL(feature1, 0) AS feature1_cleaned
FROM training_data;
```

D.



```
SELECT *
FROM training_data
```

D.



```
FROM training_data
WHERE feature1 IS NOT NULL;
```

Show Suggested Answer

by raaad at *Jan. 4, 2024, 10:45 p.m.*

Comments

Type your comment...

Submit

52ed0e5 Highly Voted 1 year, 1 month ago

Selected Answer: A

Option A is the correct choice because it retains all the original columns and specifically addresses the issue of null values in ‘feature1’ by replacing them with zeros, without altering any other columns or performing unnecessary calculations. This makes the data ready for use in BigQueryML without losing any important information.

Option C is not the best choice because it includes the EXCEPT clause for the price and square_feet columns, which would exclude these columns from the results. This is not desirable since you need these columns for the machine learning model to predict the price per square foot

upvoted 10 times

datapassionate Highly Voted 1 year, 3 months ago

Selected Answer: C

Correct answer is C.
It both replace NULL with 0 and pass price per square foot of real estate.

upvoted 7 times

George_Zhu 1 year, 2 months ago

Option C isn't a good practice. What if any 0 value is contained in the column of squre_feet, then price / 0 will throw an exception. IF(IFNULL(squre_feet, 0) = 0, 0, price/squre_feet).

upvoted 6 times

desertlotus1211 1 month, 1 week ago

The question is asking about Null values for the feature1 column, no other column with Null values.

upvoted 1 times

baimus 7 months ago

I think the assumption here is that no houses are zero feet in size. If they are, that should be caught in preprocessing, which is outside the short scope of this question. If the answer isn't C, then it's A, which would mean the question is suggesting you need an ML model to calculate price per square for data where you already have both price and square feet as features. In that instance you clearly need to only divide one by the other. Those columns must be intended to be the target, or the whole question is nonsense.

upvoted 4 times

MarcoPellegrino Most Recent 2 months, 1 week ago

Selected Answer: C

It's the only one that:
- computes the price per square foot of real estate. Note that "the training data has a column for the price and a column for the number of square feet" only.
- fills NAs with zeros

upvoted 2 times

LP_PDE 3 months, 1 week ago

Selected Answer: C

Not worded well but the best answer I would think would be C since it has price per square foot but I understand the argument for A.

upvoted 3 times

AWSandeep 4 months, 2 weeks ago

Selected Answer: C

Let's step away from GCP for a minute. If price & square feet are already features, then there is no need for BigQuery ML to do any prediction. You'd want to predict price per square foot based on other fields like location, weather, etc. The first sentence in the question indicates that a machine learning team is preparing the data for model training. Therefore, C's query a fantastic preparation step. If this query for any other use case, then A would've been the answer.

upvoted 2 times

Robbing_the_hood 4 months, 3 weeks ago

Selected Answer: C

To people saying you need price and square footage to predict price/sq. feet: you do not need an ML model then, you need a calculator. C is the correct answer because you want to predict price/sq. feet from the features EXCLUDING price and sq. footage.

upvoted 2 times

ToiToi 6 months ago

Selected Answer: C

Gemini told me C

Here's why it's the best of the limited choices:

Calculates price_per_sqft: It includes the calculation for the target variable your model needs.
Handles Nulls: It uses IFNULL(feature1, 0) to replace nulls in feature1 with 0, similar to COALESCE.
Most Comprehensive: While it excludes the original price, square_feet, and feature1 columns, it still retains any other columns that might be present in the training_data table.

upvoted 3 times

cloud_rider 5 months, 1 week ago

C is wrong as it excludes price and square feet values, what will model use to train model?

upvoted 1 times

Robbing_the_hood 4 months, 3 weeks ago

The features? Why would you train an ML model if you have price and sq. feet available?

upvoted 1 times

SamuelTsch 6 months, 1 week ago

Selected Answer: C

it should be C.

upvoted 2 times

baimus 7 months ago

Selected Answer: C

This must be C, though the wording isn't great. If price and square foot are included in the data, they are either intended to be the target, in which case you need to create that target as per C, or if they are genuinely features, you DO NOT need a machine learning model. If you already know price and square feet, price per square foot is just price/ft2. You don't need ML to predict that, it's just a division. The only context this makes sense in is if they mean "price and square foot are the target, and feature1 is the predictive feature", which means C is correct. The removing nulls from feature1 and the creation of price per square foot is C.

upvoted 2 times

47767f9 10 months ago

Selected Answer: C

Font Cloude 3.5 and GPT 4o, in theoy is better to keep the less amount of features, then price_per_sqft and feature1 cleaned is the best option

upvoted 2 times

srinidutt 1 year ago

EXCEPT means it won't select that column.

upvoted 2 times

demoro86 1 year, 2 months ago

Selected Answer: A

C is not a valid answer. You are introducing a redundant variable, that could be valid, but removing from the dataset 2 variables that exactly influence in the predictions you are trying to make.

upvoted 4 times

demoro86 1 year, 2 months ago

C is not a valid answer. You are introducing a redundant variable, that could be valid, but removing from the dataset 2 variables that exactly influence in the predictions you are trying to make.

upvoted 2 times

baimus 7 months ago

Just to clarify, they don't "influence" the prediction, they are in fact the target. The model needs to predict price per square foot. If you have price, and square foot, they are either 1) the prediction target price/squarefoot, or if not then you absolutely do not need a machine learning model, you just device price by square foot.

upvoted 2 times

PetrSz 1 year, 2 months ago

Selected Answer: C

Option C not only handles the null values in feature1 by replacing them with zeros (using IFNULL(feature1, 0) as feature1_cleaned), but it also creates a new feature price_per_sqft by dividing the price by the number of square feet (price/square_feet as price_per_sqft). This new feature directly corresponds to what your team wants to predict (the price per square foot of real estate), and could therefore be very useful for the machine learning model.

upvoted 2 times

cuadradobertolinisebastiancami 1 year, 2 months ago

Selected Answer: C

It should be C.

"They want to predict the price per square foot of real estate. The training data has a column for the price and a column for the number of square feet."

You need to create the column the model is going to predict.

upvoted 3 times

JyoGCP 1 year, 2 months ago

Selected Answer: A

Option A

upvoted 3 times

oleg25 1 year, 2 months ago

I didn't get why they mentioned in the task price and square feet columns. Just to irritate us? Do we need to do something with these columns or just with column feature1?

upvoted 5 times

d11379b 1 year, 1 month ago

I think they just want us to build a “label” (target) column ourselves since there’s no direct value in the training set

upvoted 1 times

d11379b 1 year, 1 month ago

But I still prefer to choose A since the square_feet column itself may have influence on price, which shouldn’t be removed

upvoted 1 times

Load full discussion...