

Dealing with Missing Data - Dropping Columns Below Threshold

Question and Screenshot:

The screenshot shows a learning interface with a question on the left and a code editor on the right. The question is titled "Dealing with missing data" and describes a task involving a pandas DataFrame named 'planes'. The code editor shows a Python script that counts missing values, calculates a threshold, and drops missing values from columns below the threshold. The output of the code is displayed in the IPython Shell.

Exercise

Dealing with missing data

It is important to deal with missing data before starting your analysis.

One approach is to drop missing values if they account for a small proportion, typically five percent, of your data.

Working with a dataset on plane ticket prices, stored as a pandas DataFrame called `planes`, you'll need to count the number of missing values across all columns, calculate five percent of all values, use this threshold to remove observations, and check how many missing values remain in the dataset.

Instructions 3/3

- Create `cols_to_drop` by applying boolean indexing to columns of the DataFrame with missing values less than or equal to the threshold.
- Use this filter to remove missing values and save the updated DataFrame.

Take Hint (-9 XP)

```
script.py
1 # Count the number of missing values in each column
2 print(planes.isna().sum())
3
4 # Find the five percent threshold
5 threshold = len(planes) * 0.05
6
7 # Create a filter
8 cols_to_drop = planes.____[____ <= ____]
9
10 # Drop missing values for columns below the threshold
11 planes.____(____=____, inplace=____)
12
13 print(planes.isna().sum())
```

Run Code Submit Answer

IPython Shell

```
Destination    347
Route          256
Dep_Time       260
Arrival_Time   194
Duration       214
Total_Stops    212
Additional_Info 589
Price          616
dtype: int64
```

In [1]:

Question Explanation:

This task involves identifying columns in the 'planes' DataFrame where the count of missing values is less than or equal to the calculated five percent threshold. The identified columns are then used to drop missing values specifically from these columns.

Code Solution:

```
import pandas as pd
```

```
# Count the number of missing values in each column
print(planes.isna().sum())
```

```
# Find the five percent threshold
threshold = len(planes) * 0.05
```

```
# Create a filter for columns with missing values less than or equal to the
threshold
cols_to_drop = planes.columns[planes.isna().sum() <= threshold]
```

```
# Drop missing values for columns below the threshold
planes.dropna(subset=cols_to_drop, inplace=True)
```

```
# Print the remaining missing values in the DataFrame  
print(planes.isna().sum())
```

Solution Explanation:

1. The ``isna().sum()`` function identifies and counts missing values in each column.
2. The threshold is calculated as five percent of the total number of rows.
3. Boolean indexing is used to create ``cols_to_drop``, which includes column names with missing values less than or equal to the threshold.
4. The ``dropna()`` method removes rows with missing values only in the filtered columns, and the changes are applied in place.
5. The missing values in the updated DataFrame are printed for verification.