How can I create categorical variables in Pandas, and what are some examples of doing so?”

Authored by

stats writer

June 26, 2024

Creating categorical variables in Pandas allows for easier manipulation and analysis of data that has distinct categories or groups. This can be achieved by using the "astype" function to convert a column of data into a categorical variable. Examples of creating categorical variables in Pandas include converting numerical data into categories such as age groups, or converting strings into categories such as different levels of education. Categorical variables can also be created by grouping and labeling data based on specific criteria, such as geographic regions or product categories. Using categorical variables in Pandas can improve data organization and facilitate more efficient data analysis.

Create Categorical Variables in Pandas (With Examples)

You can use one of the following methods to create a in pandas:

Method 1: Create Categorical Variable from Scratch

df =

Method 2: Create Categorical Variable from Existing Numerical Variable

df = pd.cut(df,
bins=,
labels=)

The following examples show how to use each method in practice.

Example 1: Create Categorical Variable from Scratch

The following code shows how to create a pandas DataFrame with one categorical variable called team and one numerical variable called points:

```
import pandas as pd

#create DataFrame with one categorical variable and one numeric variable
df = pd.DataFrame({'team': ,
'points': })

#view DataFrame
print(df)
```

```
team points
0 A 12
1 B 15
2 C 19
3 D 22
4 E 24
5 F 25
6 G 26
7 H 30
```

#view data type of each column in DataFrame
print(df.dtypes)

team object
points int64
dtype: object

By using df.dtypes, we can see the in the DataFrame.

We can see:

The variable team is an object.The variable points is an integer.

In Python, an object is equivalent to a character or "categorical" variable. Thus, the team variable is a categorical variable.

Example 2: Create Categorical Variable from Existing Numerical Variable

The following code shows how to create a categorical variable called status from the existing numerical variable called points in the DataFrame:

import pandas as pd

#create DataFrame with one categorical variable and one numeric variable
df = pd.DataFrame({'team': ,
'points': })

#create categorical variable 'status' based on existing numerical 'points' variable
df = pd.cut(df,
bins=,
labels=)

#view updated DataFrame
print(df)

team points status
0 A 12 Bad
1 B 15 Bad
2 C 19 OK
3 D 22 OK
4 E 24 OK
5 F 25 OK
6 G 26 Good
7 H 30 Good

Using the cut() function, we created a new categorical

variable called status that takes the following values:

'Bad' if the value in the points column is less than or equal to 15.Else, 'OK' if the value in the points column is less than or equal to 25.Else, 'Good'.

Note that when using the cut() function, the number of labels must be one less than the number of bins.

In our example, we used four values for bins to define the bin edges and three values for labels to specify the labels to use for the categorical variable.

The following tutorials explain how to perform other common tasks in pandas: