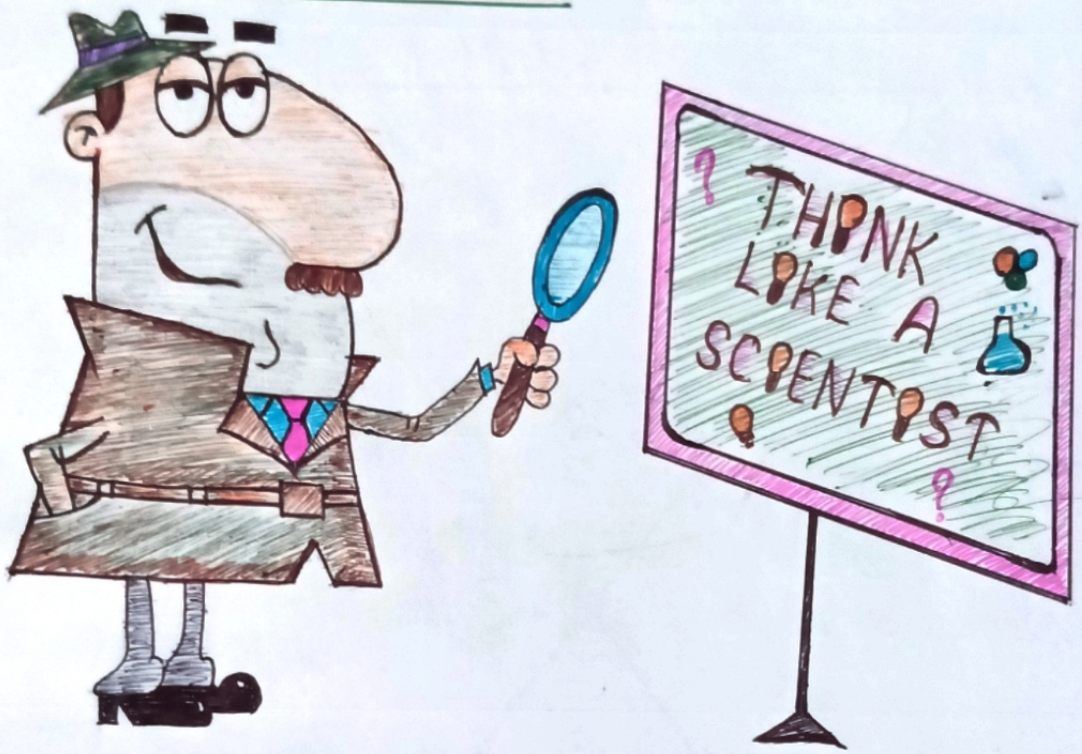
Removing observations with missing data in Pandas



Complete Case Analysis (CCA), also called list-wise deletion of cases, consists of discarding those observations where the values in any of the variables are missing. CCA can be applied to categorical and numerical variables. CCA is quick and easy to implement and has the advantage that it preserves the distribution of the variables, provided the data is missing at random and only a small proportion of the data is missing. However, if data is missing across many variables, CCA may lead to the removal of a big portion of the dataset.

How to do it....

Let's begin by loading pandas and the dataset:

1. First, we'll import the pandas library:

```
import pandas as pd
```

2. Let's load the Credit Approval Data Set:

```
data=pd.read_csv('creditApprovalUCI.csv')
```

3. Let's calculate the percentage of missing values for each variable and sort them in ascending order:

```
data.isnull().mean().sort_values(ascending=True)
```

The output of the preceding code is as follows:

```
A11    0.000000
A12    0.000000
A13    0.000000
A15    0.000000
A16    0.000000
A4     0.008696
A5     0.008696
A6     0.013043
A7     0.013043
A1     0.017391
A2     0.017391
```

```
A14     0.018841
A3      0.133333
A8      0.133333
A9      0.133333
A10     0.133333
dtype: float64
```

4. Now, we'll remove the observations with missing data in any of the variables:

```
data_cca = data.dropna()
```

To remove observations where data is missing in a subset of variables, we can execute `data.dropna(subset=['A3','A4'])`. To remove observations if data is missing in all the variables, we can execute `data.dropna(how='all')`.

5. Let's print and compare the size of the original and complete case datasets:

```
Print('Number of total observations: {}'.format(len(data)))
Print('Number of observations with complete case:
{}'.format(len(data_cca)))
```

Here, we removed more than 100 observations with missing data, as shown in the following output :

```
Number of total observations : 690
Number of observations with complete cases : 564
```

We can use the code from step3 to corroborate the absence of missing data in the complete case dataset.

How it works ....

In this method, we determined the percentage of missing data for each variable in the Credit Approval Data Set and removed all observations with missing information to create a complete case dataset.

First, we loaded the data from a CSV file into a dataframe with the pandas read_csv() method. Next, we used the pandas isnull() and mean() methods to determine the percentage of missing observations for each variable. We will check these methods in the Quantifying missing data method. With pandas sort_values(), we ordered the variables from the one with the fewest missing values to the one with the most.

To remove observations with missing values in any of the variables, we used the pandas dropna() method, thereby obtaining a complete case dataset. Finally, we calculated the number of observations we removed using the Python built-in method len, which returned the number of rows in the original and complete case datasets. Using format, we included the len output within the {} in the print statement, thereby displaying the number of missing observations next to the text.