Filling Missing values(#NA) using values from previous columns in Pandas

Gregory Murimi

1

Create a sample DataFrame with #NA values

```
In [2]: #create a dataframe

import pandas as pd


df = pd.DataFrame({
    'Name':['Angie','Louie','Jim'],
    'Arrival Date':['06/05/2023','08/05/2023','09/05/2023'],
    'Airport':['Yuma International','John F Kennedy','Denver'],
    'Cost Center':[pd.NA,pd.NA,pd.NA]
})

print(df)

    Name Arrival Date            Airport Cost Center
0  Angie   06/05/2023  Yuma International        <NA>
1  Louie   08/05/2023     John F Kennedy        <NA>
2    Jim   09/05/2023             Denver        <NA>
```

Because, I'm always extra 😊, I had to change the 'Arrival Date' datatype from object to datetime

```
In [3]: df.dtypes
```

```
Out[3]: Name           object
        Arrival Date   object
        Airport        object
        Cost Center    object
        dtype: object
```

```
In [4]: #Change the Arrival Date column from object(Text) to datetime

        df['Arrival Date'] = pd.to_datetime(df['Arrival Date'], format = '%d/%m/%Y')
```

```
In [5]: #Confirm column type chnaged

        df.dtypes
```

```
Out[5]: Name                   object
        Arrival Date   datetime64[ns]
        Airport                object
        Cost Center            object
        dtype: object
```

Fill in the missing values in Cost Center column using masks.

Masks are a Boolean series or arrays that you can use to select or filter values in a DataFrame based on some condition

```python
In [7]: #Fill in the missing values in Cost Center column using masks

        #create a mask of missing values from the 'Cost Center' column
        mask = df['Cost Center'].isna()

        #print mask, where the condition is met(missing values), it should return 'True'
        print(mask)

        0    True
        1    True
        2    True
        Name: Cost Center, dtype: bool
```

4

*Locate the mask in the 'Cost Center' column then use **'ffill'** to fill in values from the previous column (in this case, the cost centers are from the same Airport)*

.loc is a label-based indexing method used to select rows and columns from a DataFrame. It is used to filter data based on labels, instead of numerical positions.

*The **fillna** method is used to fill missing or NaN (not a number) values in a pandas DataFrame or Series with a specified value or method such as **'ffill'** (before), **'bfill'** (after the missing values) etc.*

*The **'ffill'** method is used to fill missing or NaN values with the previous non-missing value in a pandas DataFrame or Series.*

```
In [9]: #Locate the mask in the 'Cost Center' column then use ffill to fill in values from the previous column

        df.loc[mask,'Cost Center'] = df.loc[mask,'Airport'].fillna(method = 'ffill')

        #print the dataframe
        print(df)

            Name Arrival Date              Airport         Cost Center
        0  Angie   2023-05-06  Yuma International  Yuma International
        1  Louie   2023-05-08      John F Kennedy      John F Kennedy
        2    Jim   2023-05-09              Denver              Denver

In [ ]:
```

Done! 😊 *Missing Values have been successfully filled with values from the previous column!*

Pandas are beautiful! Aren't they?