Mapping Data and Finding Duplicates

In this lesson, data mapping will be disussed along with how to find and remove duplicates in data.

We'll cover the following

- Mapping data
- Removing duplicated data

Mapping data

As the name suggests, this technique is used to map the values of a `Series` or a `DataFrame`. The current values in a `Series` or a `DataFrame` are made equivalent to some other values. Then, the pandas `map` function is used to either replace the mapped values or join them together. The `map()` function can also be used to fill in the values of new columns.

The following example might make it clear.

```
import pandas as pd

df = pd.DataFrame({'City':['Lahore', 'Mumbai', 'Karachi', 'London'],
                   'AQI':[147, 166, 152, 81]})

print("The Original DataFrame")
print(df,'\n')

dict_map = {'Lahore':'Pakistan', 'Karachi':'Pakistan', 'Mumbai':'India', 'London':'UK'}

df['Country'] = df['City'].map(dict_map)

print("The Mapped DataFrame")
print(df)
```

Again a `DataFrame` is defined using the *dictionary* method like in the previous lesson. Different cities are mentioned along with their AQI (air quality index) values in the *dictionary*.

On **line 9**, a dictionary is defined with city names as the key attribute and a country name as their value.
On **line 11**, a new column `Country` is created and assigned values using the `map` function. This function takes the `dict_map` dictionary created on **line 9** as a parameter and automatically assigns the corresponding country values to the cities.

Removing duplicated data

Now, you'll see how to get rid of duplicates in a dataset. The `drop_duplcates()` function of pandas is used for this purpose.

```
import pandas as pd

df = pd.DataFrame({'Col1':['A', 'B', 'A', 'C', 'B', 'C'],
                   'Col2': [1, 2, 1, 3, 4, 3]})

print("The Original DataFrame")
print(df,'\n')

print("The DataFrame without duplicates")
print(df.drop_duplicates(),'\n')

print("The DataFrame without Column1 duplicates")
print(df.drop_duplicates(['Col1']))
```

The `drop_duplicates` method is used on **line 10** without any parameter. It removes all rows that occur more than once. It only removes the rows where each cell exactly matches any other row of the `DataFrame`.

The `drop_duplicates` method is used on **line 13**. It has a column name passed as a parameter. This method checks for duplicate values in the specified column only, not in the whole row. All the rows with the same column values are removed until only one remains.

df.drop_duplicates()

1

Duplicates

2

Duplicates Removed

df.drop_duplicates(['Col1'])

1

2

3

Col1 Dupicates Removed

Similarly, duplicates can also be dropped based on multiple columns.

In the next lesson, some more important functions of pandas are explored.