Duplicates

This lesson will focus on how to deal with data that has duplicates.

WE'LL COVER THE FOLLOWING

Duplicates

Repeated data rows in the dataset are called **duplicates**. These can arise from a number of ways. The most common are:

- The same data is entered twice by accident, such as the same article is scraped twice or booking for an online product is made twice.
- If data is being collected in online forms or surveys and the user presses the submit button twice.
- If data is collected from multiple sources.

Detecting duplicates

In every dataset, there are some or one attribute that makes the records unique from each other. For instance, the order ID in the sales data, the student ID in the data of students, the longitude and latitude in the Census data, etc. We can search for duplicates using these key variables.

Pandas has a function, `duplicated` for finding duplicates. We specify the column names, and it gives us a list of Booleans (`True` and `False`) putting `True` against rows that have duplicates. It puts `False` for the first occurrence of a duplicate.

We will be using the Credit Card Default Dataset.

```
import pandas as pd
df = pd.read_csv('credit_card.csv')
print(df.columns)

# Find duplicates
duplicates = df.duplicated(subset = ['ID'])
print('\n\n',df[duplicates])
```

In **line 6** we use the function `duplicated`. We provide a list of column names as `subset` for which we want to find duplicates and it gives us a list. We index using that list in **line** 7 and get the duplicated rows.

Removing duplicates

Duplicates can be removed by using the function `drop_duplicates`. It has to be provided the list of column names in the same way as we provided to `duplicated` above.

```
import pandas as pd
df = pd.read_csv('credit_card.csv')
print('Shape of original dataframe : ',df.shape)

# Drop Rows that have Duplicates
new_df = df.drop_duplicates(subset = 'ID')
print('Shape of dataframe with duplicate rows dropped : ',new_df.shape)
```

We drop the duplicated columns in **line 6** and verify that by looking at the shape of the new dataframe. 8 rows have been dropped.

This was how to deal with duplicates in data. In the next lesson, we will focus on how to deal with inconsistent data.