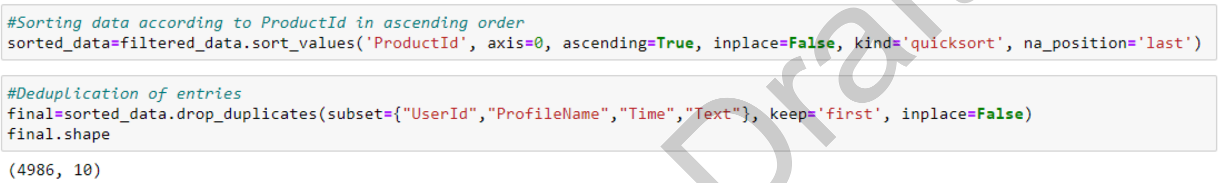
28.2 Data Cleaning: Deduplication

It is observed that the given dataset had many duplicate entries. Hence it is necessary to remove duplicates in order to get unbiased results for the analysis of the data.

In our data, we are terming two or more reviews as duplicates, if they have the same values for the columns ‘UserID’, ‘ProfileName’, ‘Time’, and ‘Text’. Before we delete the duplicate entries, we first have to sort all the reviews on the basis of the ‘ProductId’ column using pandas.sort_values(). After sorting the reviews, we drop the duplicates using panda.drop_duplicates().

Below is the line of code that was discussed starting from the timestamp 7:00 which sorts all the reviews on the basis of the ‘ProductId’ column. After sorting the reviews, we are dropping the duplicates and it was discussed at the timestamp



```
#Sorting data according to ProductId in ascending order
sorted_data=filtered_data.sort_values('ProductId', axis=0, ascending=True, inplace=False, kind='quicksort', na_position='last')
```

```
#Deduplication of entries
final=sorted_data.drop_duplicates(subset={"UserId","ProfileName","Time","Text"}, keep='first', inplace=False)
final.shape
```

```
(4986, 10)
```

Deduplication is a technique used to improve storage utilization and can also be applied to the network data transfers to reduce the number of bytes that must be read.