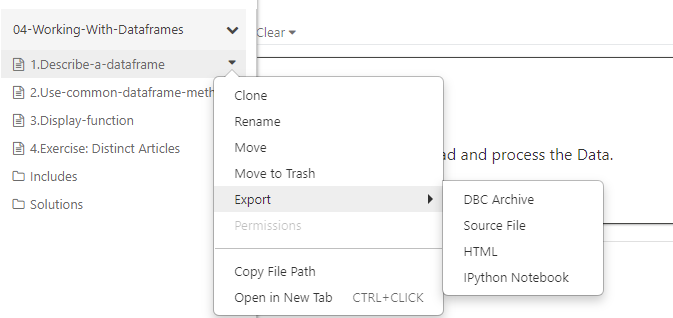
Lab3 Notes

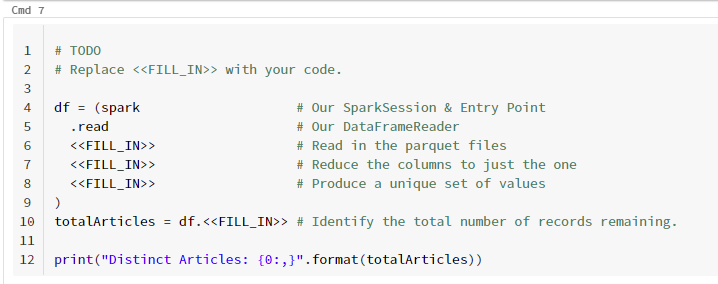
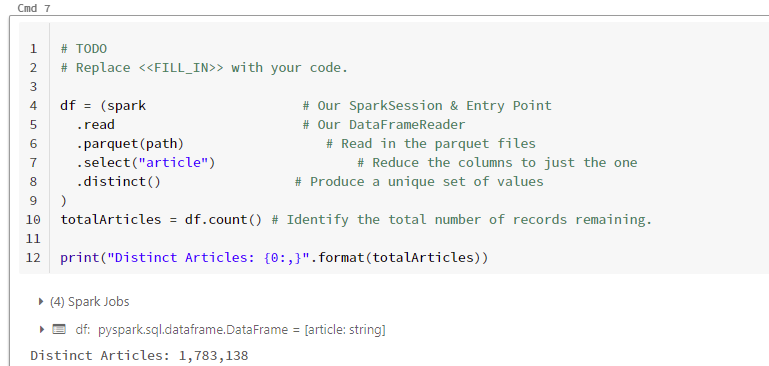
Exerc. 1

<https://github.com/MicrosoftLearning/DP-203-Data-Engineer/raw/master/Allfiles/microsoft-learning-paths-databricks-notebooks/data-engineering/DBC/04-Working-With-Dataframes.dbc>



At the end of the first notebook there is a task to view documentation. Click the Latest link, then scroll down to find and then click the Python API link. Click Get Started, and then read a few examples. Don’t spend more than 5 min on that task.

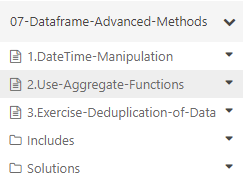
In the last notebook there is a cell you will complete the code:

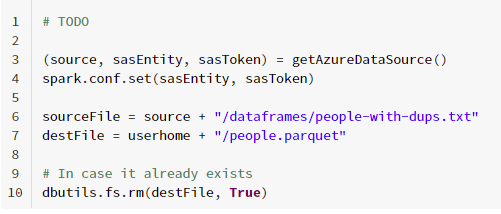
In the solution they don’t use the path variable that is already available.

Exerc. 2

<https://github.com/MicrosoftLearning/DP-203-Data-Engineer/raw/master/Allfiles/microsoft-learning-paths-databricks-notebooks/data-engineering/DBC/07-Dataframe-Advanced-Methods.dbc>



In the last notebook we don’t have a starter code to fill in the gaps. All we have is below:



A solution is provided. We should open it and study the content.

The basic idea is to introduce 4 columns that will make possible to find the duplicates, drop the duplicates and then drop these extra columns. The first 3 are just firstName, lastName and middleName all in lower case. The last one is the ssn without the dashes. We give other names to these 4 columns and pass the list of all columns with these 4 replacing the original ones to dropDuplicates([“col1”,”col2”,”col3”,”col4”,”gender”,”birthdate”,”salary”]) and only the new 4 columns to drop(“col1”, ,”col2”,”col3”,”col4”).

