Data set conversion between Spark Dataframe & Pandas Dataframe

DataFrame Spark: Dataset is a new interface added in Spark 1.6 that provides the benefits of RDDs (strong typing, ability to use powerful lambda functions) with the benefits of Spark SQL's optimized execution engine. It's immutable data set

```
Dataframe Pandas: two-dimensional size-mutable, potentially heterogeneous
tabular data structure with labeled axes (rows and columns). A Data frame is a
two-dimensional data structure, i.e., data is aligned in a tabular fashion in
rows and columns
```

In [3]:
```
import findspark
import pyspark
findspark.init()
```

In [4]:
```
from pyspark.sql import SparkSession
from pyspark import SparkContext, SparkConf

spark = SparkSession.builder.appName('abc').getOrCreate()
sc = spark.sparkContext
```

In [5]:
```
from sklearn.datasets import load_boston
import pandas as pd
import numpy as np

data = load_boston()

df = pd.DataFrame(data.data, columns=data.feature_names)
df['target'] = data['target']
df.head(4)
```

Out[5]:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94

From Pandas to Spark Dataframe

In [6]:
```
df = spark.createDataFrame(df)
```

In [7]:
```
df.show(5)
```

```
+-------+----+-----+----+-----+-----+----+------+---+-----+-------+------+-----
+------+
|   CRIM|  ZN|INDUS|CHAS|  NOX|   RM| AGE|   DIS|RAD|  TAX|PTRATIO|     B|LSTAT
|target|
+-------+----+-----+----+-----+-----+----+------+---+-----+-------+------+-----
+------+
|0.00632|18.0| 2.31| 0.0|0.538|6.575|65.2|  4.09|1.0|296.0|   15.3| 396.9| 4.98
|  24.0|
|0.02731| 0.0| 7.07| 0.0|0.469|6.421|78.9|4.9671|2.0|242.0|   17.8| 396.9| 9.14
|  21.6|
|0.02729| 0.0| 7.07| 0.0|0.469|7.185|61.1|4.9671|2.0|242.0|   17.8|392.83| 4.03
|  34.7|
|0.03237| 0.0| 2.18| 0.0|0.458|6.998|45.8|6.0622|3.0|222.0|   18.7|394.63| 2.94
|  33.4|
|0.06905| 0.0| 2.18| 0.0|0.458|7.147|54.2|6.0622|3.0|222.0|   18.7| 396.9| 5.33
|  36.2|
+-------+----+-----+----+-----+-----+----+------+---+-----+-------+------+-----
+------+
only showing top 5 rows
```

In [8]:
```
df.select('CRIM','NOX','AGE').show(5)
```

```
+-------+-----+----+
|   CRIM|  NOX| AGE|
+-------+-----+----+
|0.00632|0.538|65.2|
|0.02731|0.469|78.9|
|0.02729|0.469|61.1|
|0.03237|0.458|45.8|
|0.06905|0.458|54.2|
+-------+-----+----+
only showing top 5 rows
```

Convert Dataframe to Pandas Dataframe

In [9]:
```
dfPandas=df.toPandas()
dfPandas.head(5)
```

Out[9]:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33

In [ ]: