DataFrames Aggregations

DF Common Transformations

- ***select***

  df.select(cols) --> df

- ***groupBy***

  df.orderBy(cols) --> “grouped data df”

- ***agg***

  df.agg({op: col }) -- > df

- ***orderBy***

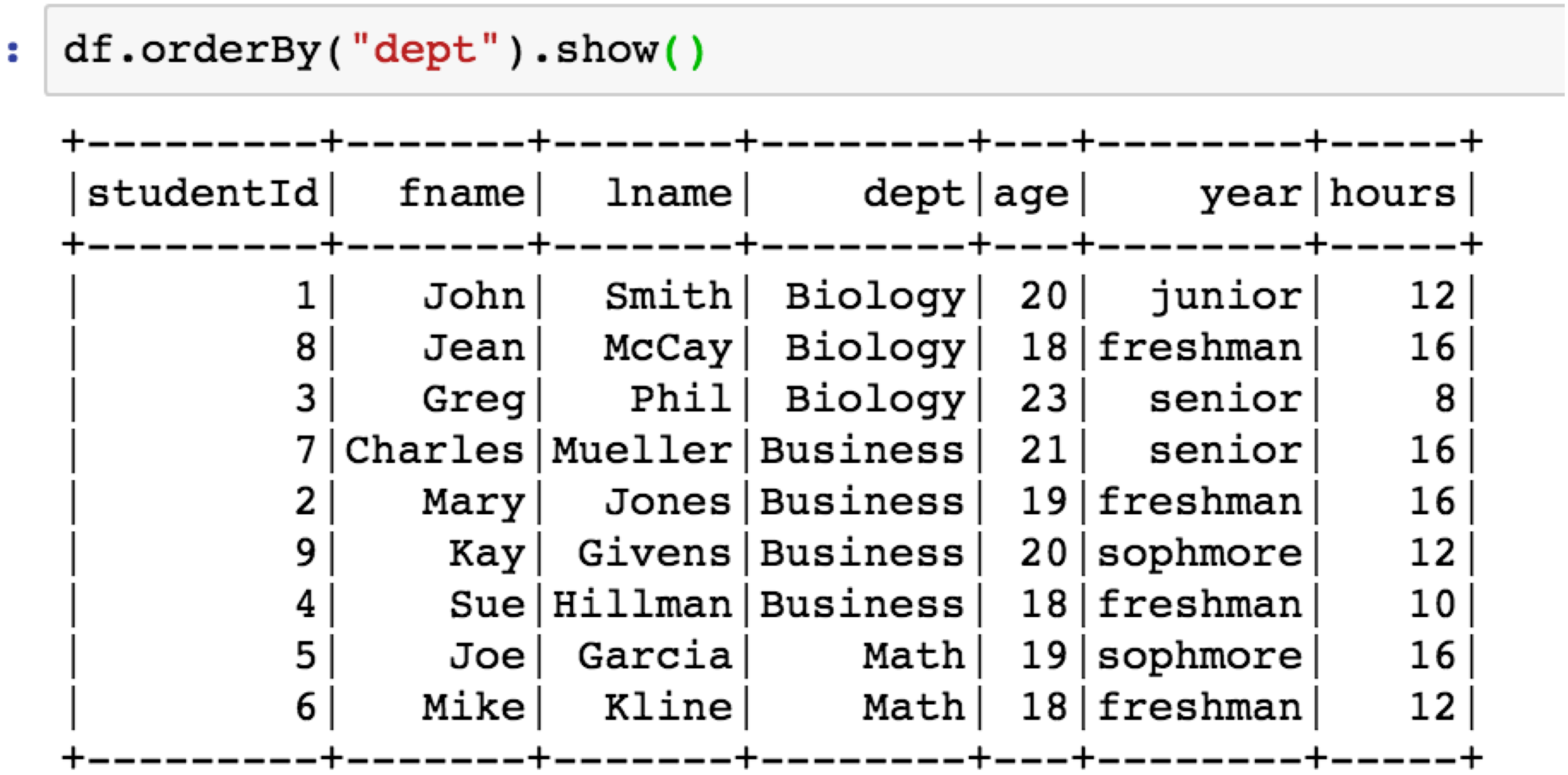
  df.sortBy(cols) → df

- ***join***

  df1.join(df2) → df

notebook : DFSortAggJoin

DF Transformation - OrderBy

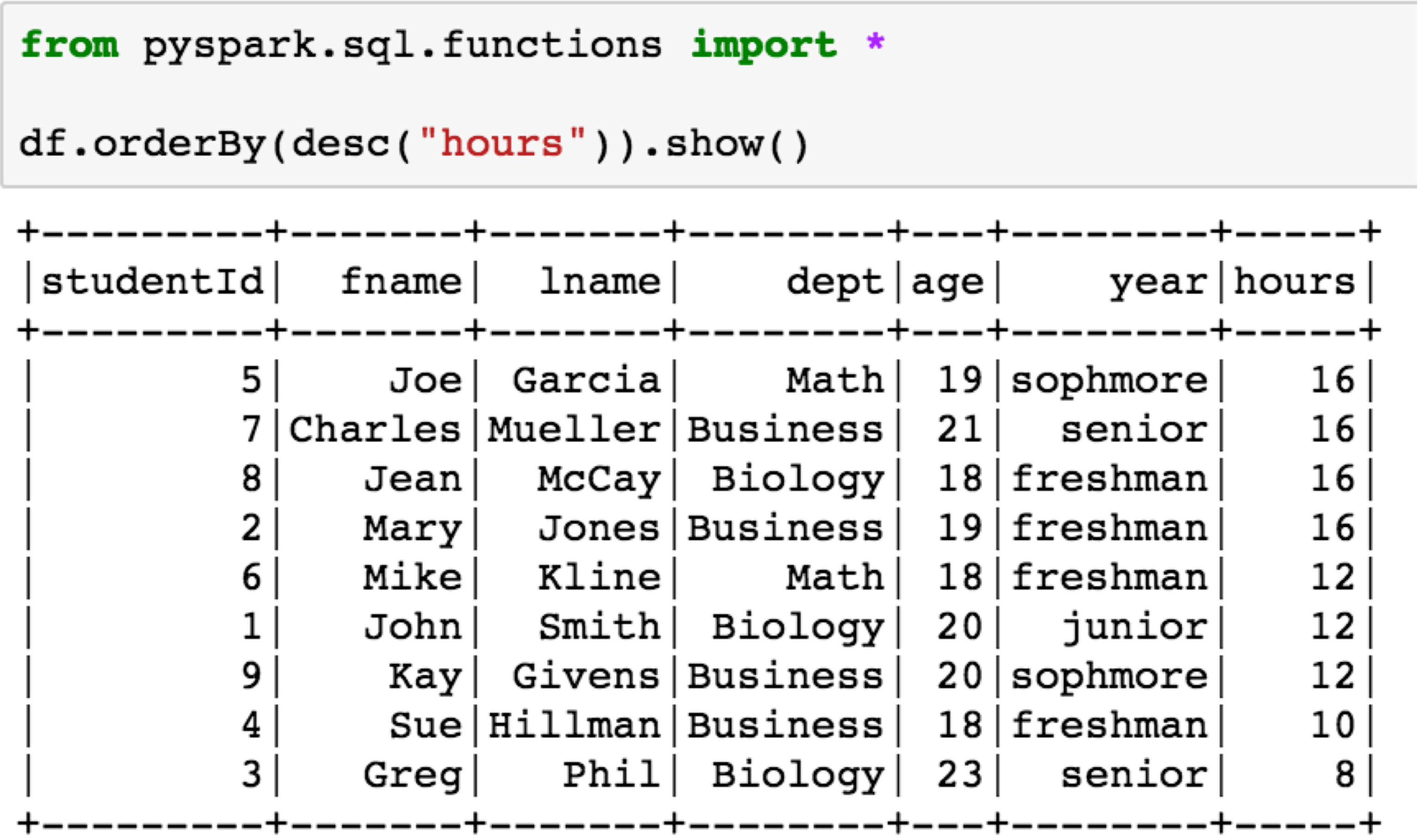


```
df.orderBy("dept").show()
```

```
+---------+-------+-------+--------+---+---------+-----+
|studentId|  fname|  lname|    dept|age|     year|hours|
+---------+-------+-------+--------+---+---------+-----+
|        1|   John|  Smith| Biology| 20|   junior|   12|
|        8|   Jean|  McCay| Biology| 18| freshman|   16|
|        3|   Greg|   Phil| Biology| 23|   senior|    8|
|        7|Charles|Mueller|Business| 21|   senior|   16|
|        2|   Mary|  Jones|Business| 19| freshman|   16|
|        9|    Kay| Givens|Business| 20|sophmore|   12|
|        4|    Sue|Hillman|Business| 18| freshman|   10|
|        5|    Joe| Garcia|    Math| 19|sophmore|   16|
|        6|   Mike|  Kline|    Math| 18| freshman|   12|
+---------+-------+-------+--------+---+---------+-----+
```

OrderBy – descending order

Note you must import pyspark.sql.functions



```
from pyspark.sql.functions import *

df.orderBy(desc("hours")).show()
```

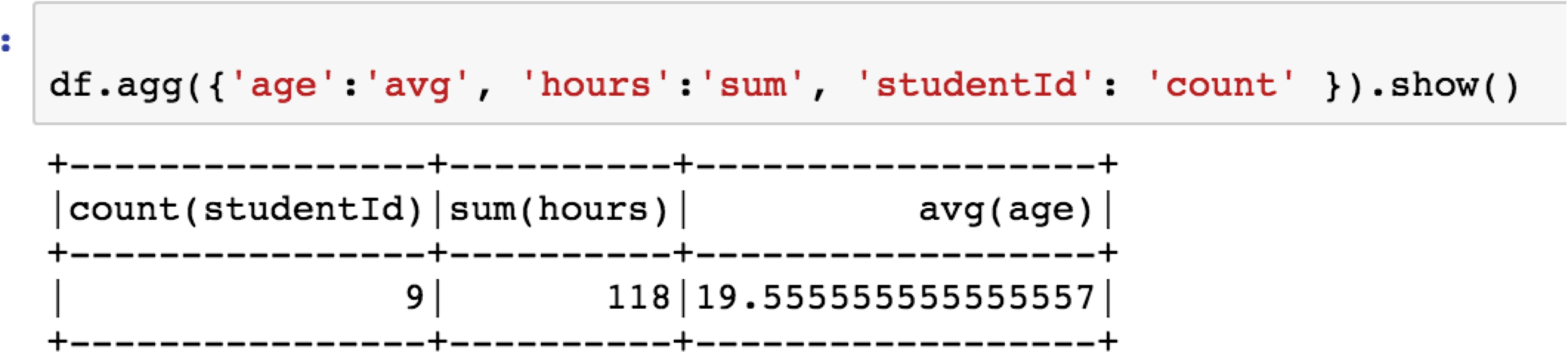
```
+---------+-------+-------+--------+---+--------+-----+
|studentId|  fname|  lname|    dept|age|    year|hours|
+---------+-------+-------+--------+---+--------+-----+
|        5|    Joe| Garcia|    Math| 19|sophmore|   16|
|        7|Charles|Mueller|Business| 21|  senior|   16|
|        8|   Jean|  McCay| Biology| 18|freshman|   16|
|        2|   Mary|  Jones|Business| 19|freshman|   16|
|        6|   Mike|  Kline|    Math| 18|freshman|   12|
|        1|   John|  Smith| Biology| 20|  junior|   12|
|        9|    Kay| Givens|Business| 20|sophmore|   12|
|        4|    Sue|Hillman|Business| 18|freshman|   10|
|        3|   Greg|   Phil| Biology| 23|  senior|    8|
+---------+-------+-------+--------+---+--------+-----+
```

Transformations - Aggregate

- Aggregate performs functions such as sum, count, avg … in a rather awkard way.

- df.agg({‘col1’:’op1', ‘col2’:’op2’, …. ‘coln' : ‘opn’}) returns a df

Transformations - Aggregate



```
df.agg({'age':'avg', 'hours':'sum', 'studentId': 'count' }).show()
```

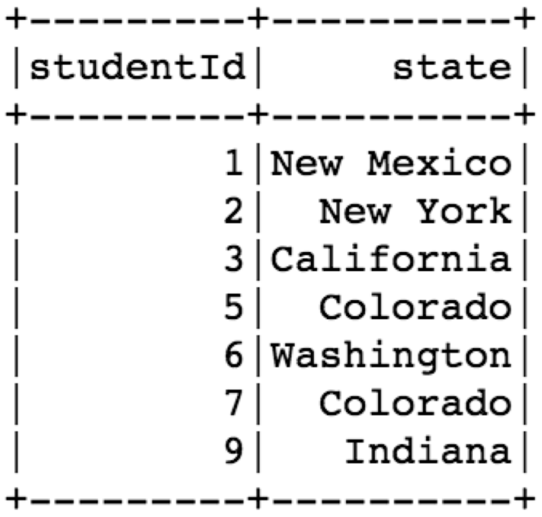
```
+----------------+----------+------------------+
|count(studentId)|sum(hours)|          avg(age)|
+----------------+----------+------------------+
|               9|       118|19.555555555555557|
+----------------+----------+------------------+
```

Transformations - Join

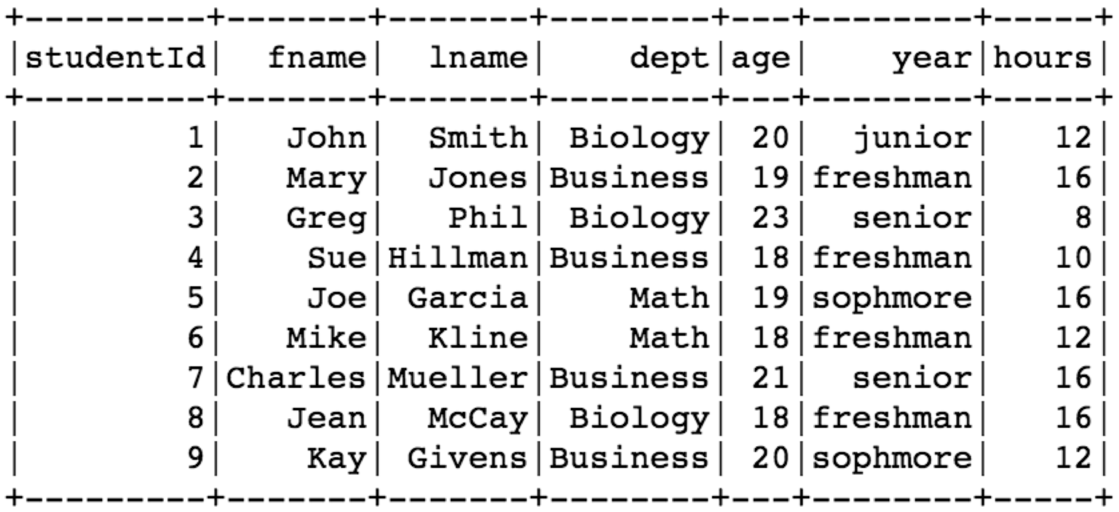
```
df.join( df2, df.key = df.key2)
```

inner, left, right joins are supported

Transformations - Join



```
+---------+----------+
|studentId|     state|
+---------+----------+
|        1|New Mexico|
|        2|  New York|
|        3|California|
|        5|  Colorado|
|        6|Washington|
|        7|  Colorado|
|        9|   Indiana|
+---------+----------+
```

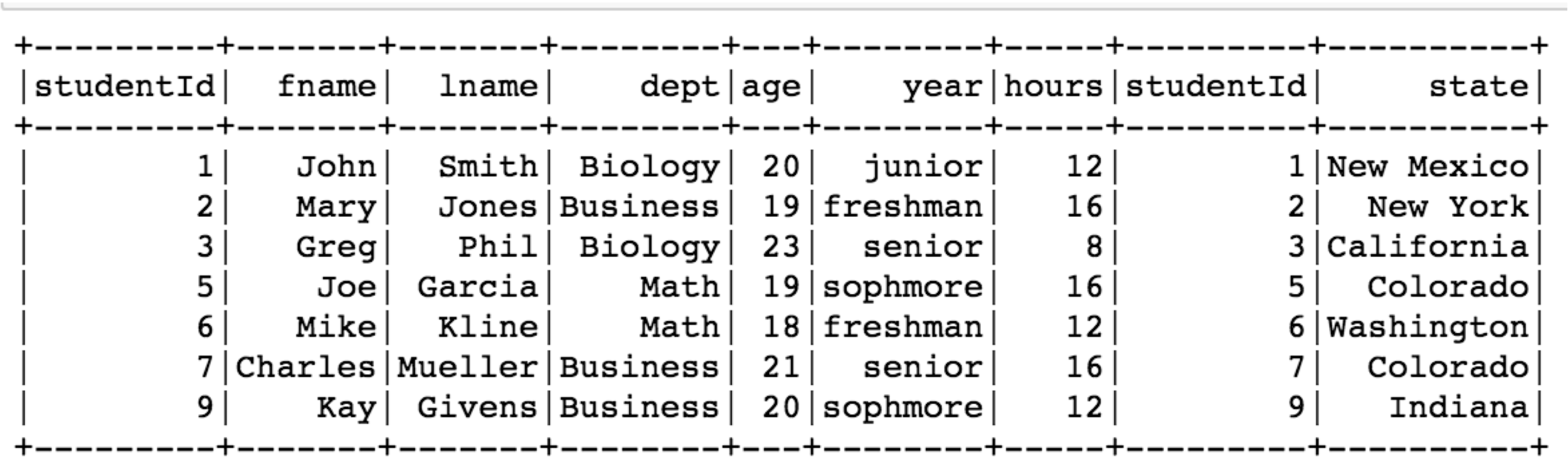


```
+---------+-------+-------+--------+---+--------+-----+
|studentId|  fname|  lname|    dept|age|    year|hours|
+---------+-------+-------+--------+---+--------+-----+
|        1|   John|  Smith| Biology| 20|  junior|   12|
|        2|   Mary|  Jones|Business| 19|freshman|   16|
|        3|   Greg|   Phil| Biology| 23|  senior|    8|
|        4|    Sue|Hillman|Business| 18|freshman|   10|
|        5|    Joe| Garcia|    Math| 19|sophmore|   16|
|        6|   Mike|  Kline|    Math| 18|freshman|   12|
|        7|Charles|Mueller|Business| 21|  senior|   16|
|        8|   Jean|  McCay| Biology| 18|freshman|   16|
|        9|    Kay| Givens|Business| 20|sophmore|   12|
+---------+-------+-------+--------+---+--------+-----+
```

```
df.join(df2, df.studentId == df2.studentId).show()
```

Transformations - Join

df.join(df2, df.studentId == df2.studentId).show()



```
+---------+-------+-------+--------+---+--------+-----+---------+----------+
|studentId|  fname|  lname|    dept|age|    year|hours|studentId|     state|
+---------+-------+-------+--------+---+--------+-----+---------+----------+
|        1|   John|  Smith| Biology| 20|  junior|   12|        1|New Mexico|
|        2|   Mary|  Jones|Business| 19|freshman|   16|        2|  New York|
|        3|   Greg|   Phil| Biology| 23|  senior|    8|        3|California|
|        5|    Joe| Garcia|    Math| 19|sophmore|   16|        5|  Colorado|
|        6|   Mike|  Kline|    Math| 18|freshman|   12|        6|Washington|
|        7|Charles|Mueller|Business| 21|  senior|   16|        7|  Colorado|
|        9|    Kay| Givens|Business| 20|sophmore|   12|        9|   Indiana|
+---------+-------+-------+--------+---+--------+-----+---------+----------+
```