Introducing Spark SQL

How to process Dataframe as SQL

```
In: geoip_df

Out: DataFrame[ip: string, code: string, country: string]
```

```
In: geoip_df
```

```
Out: DataFrame[ip: string, code: string, country: string]
```

show

```
In: geoip_df
```

```
Out: DataFrame[ip: string, code: string, country: string]
```

show

```
In: geoip_df.show(3)
```

```
+---------------+----+------------------+
|             ip|code|           country|
+---------------+----+------------------+
|194.120.126.123|  NL|       Netherlands|
| 94.126.119.173|  FR|            France|
|  193.46.74.166|  RU|Russian Federation|
+---------------+----+------------------+
only showing top 3 rows
```

select

select

```
In: geoip_df.select("country","ip")\
            .show(3)
```

```
+------------------+---------------+
|           country|             ip|
+------------------+---------------+
|       Netherlands|194.120.126.123|
|            France| 94.126.119.173|
|Russian Federation|  193.46.74.166|
+------------------+---------------+
only showing top 3 rows
```

where

where

```
In: geoip_df\
        .select("country","ip")\
        .where("country = 'Russian Federation'")\
        .show(3)
```

```
+------------------+--------------+
|           country|            ip|
+------------------+--------------+
|Russian Federation| 193.46.74.166|
|Russian Federation| 46.235.67.202|
|Russian Federation|193.161.193.64|
+------------------+--------------+
only showing top 3 rows
```

```
In: step1 = geoip_df.select("country","ip")
```

```
In: step2 = step1.where("country = 'Russian Federation'")
```

```
In: step3 = step2.show(3)
```

```
+------------------+--------------+
|           country|            ip|
+------------------+--------------+
|Russian Federation| 193.46.74.166|
|Russian Federation| 46.235.67.202|
|Russian Federation|193.161.193.64|
+------------------+--------------+
only showing top 3 rows
```

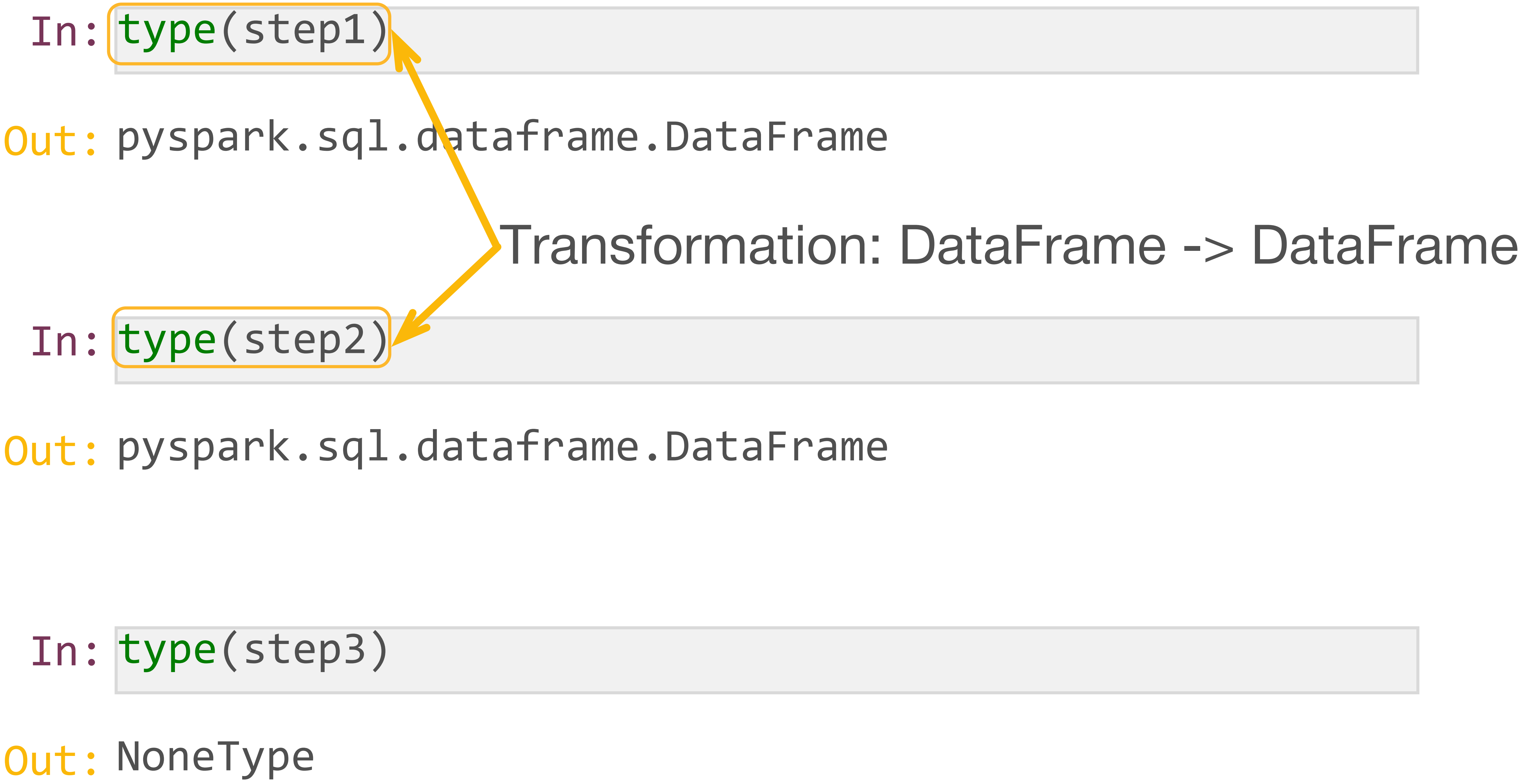
```
In: type(step1)

Out: pyspark.sql.dataframe.DataFrame
```

```
In: type(step2)

Out: pyspark.sql.dataframe.DataFrame
```

```
In: type(step1)

Out: pyspark.sql.dataframe.DataFrame
```

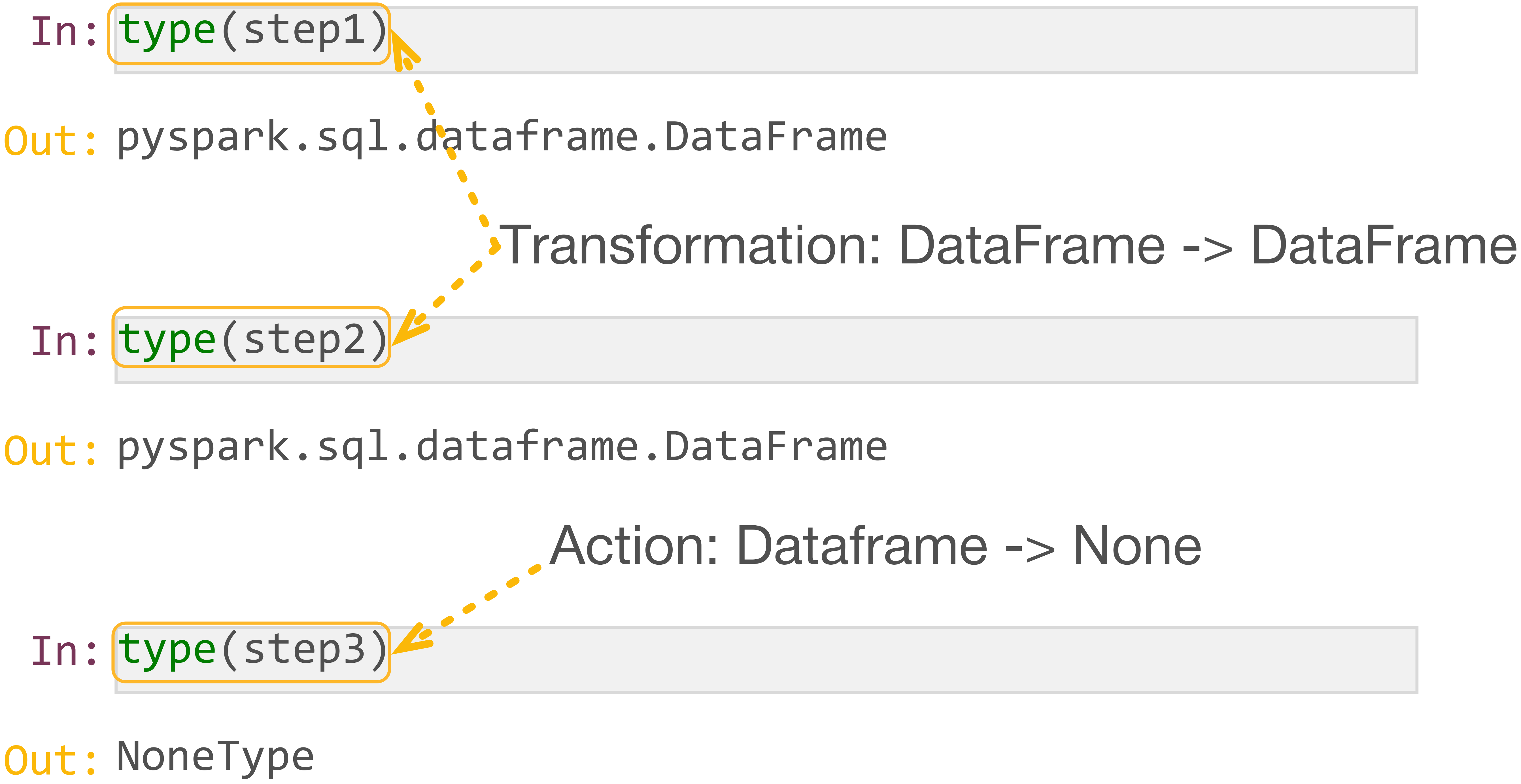
```
In: type(step2)

Out: pyspark.sql.dataframe.DataFrame
```

```
In: type(step3)

Out: NoneType
```



```
In:  type(step1)

Out: pyspark.sql.dataframe.DataFrame

In:  type(step2)

Out: pyspark.sql.dataframe.DataFrame

In:  type(step3)

Out: NoneType
```

Transformation: DataFrame -> DataFrame



```
In: type(step1)

Out: pyspark.sql.dataframe.DataFrame
```

Transformation: DataFrame -> DataFrame

```
In: type(step2)

Out: pyspark.sql.dataframe.DataFrame
```

Action: Dataframe -> None

```
In: type(step3)

Out: NoneType
```

In:

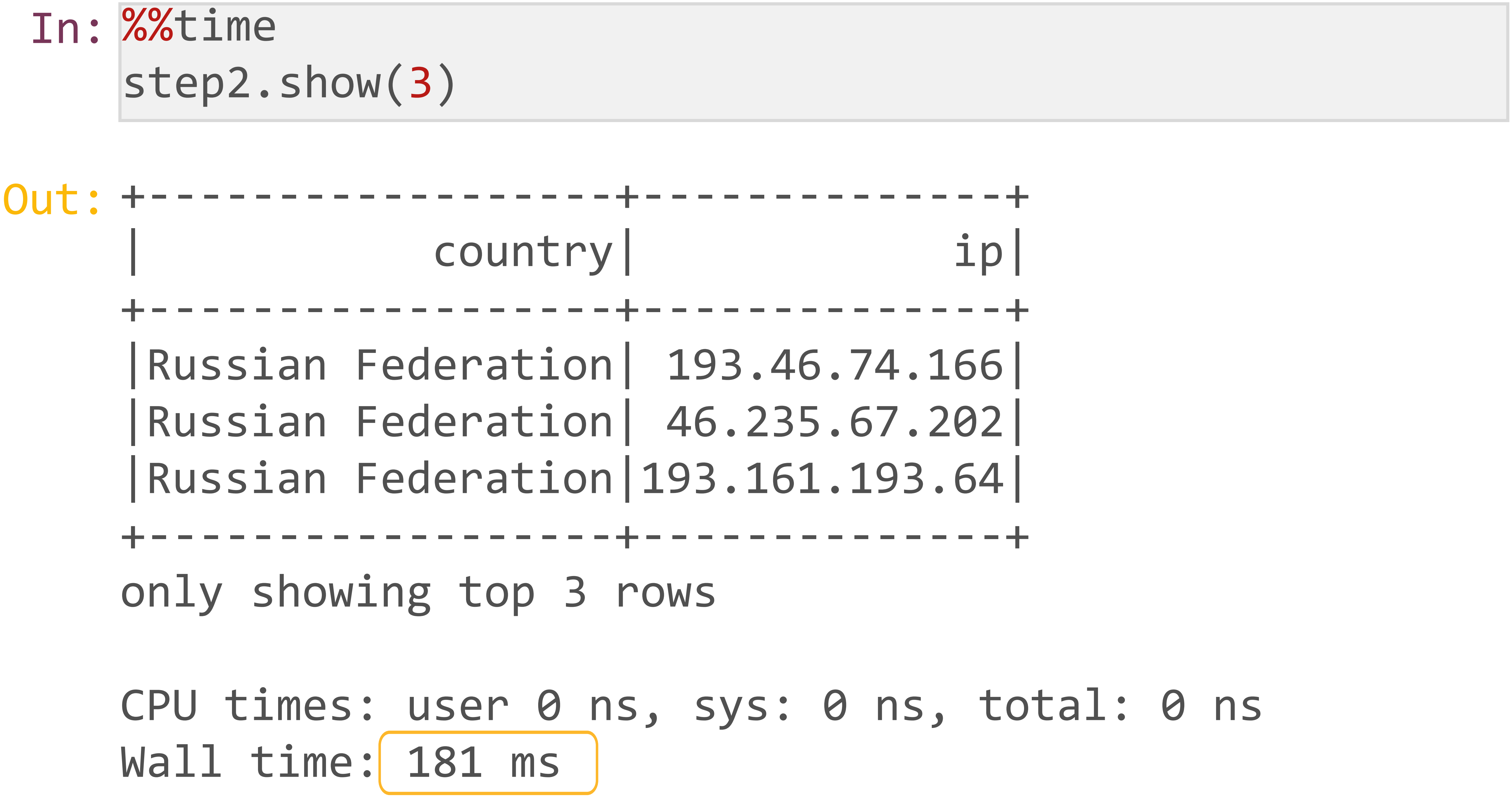
```
%%time
step1 = geoip_df.select("country","ip")
```

Out:

```
CPU times: user 4 ms, sys: 0 ns, total: 4 ms
Wall time: 28.8 ms
```

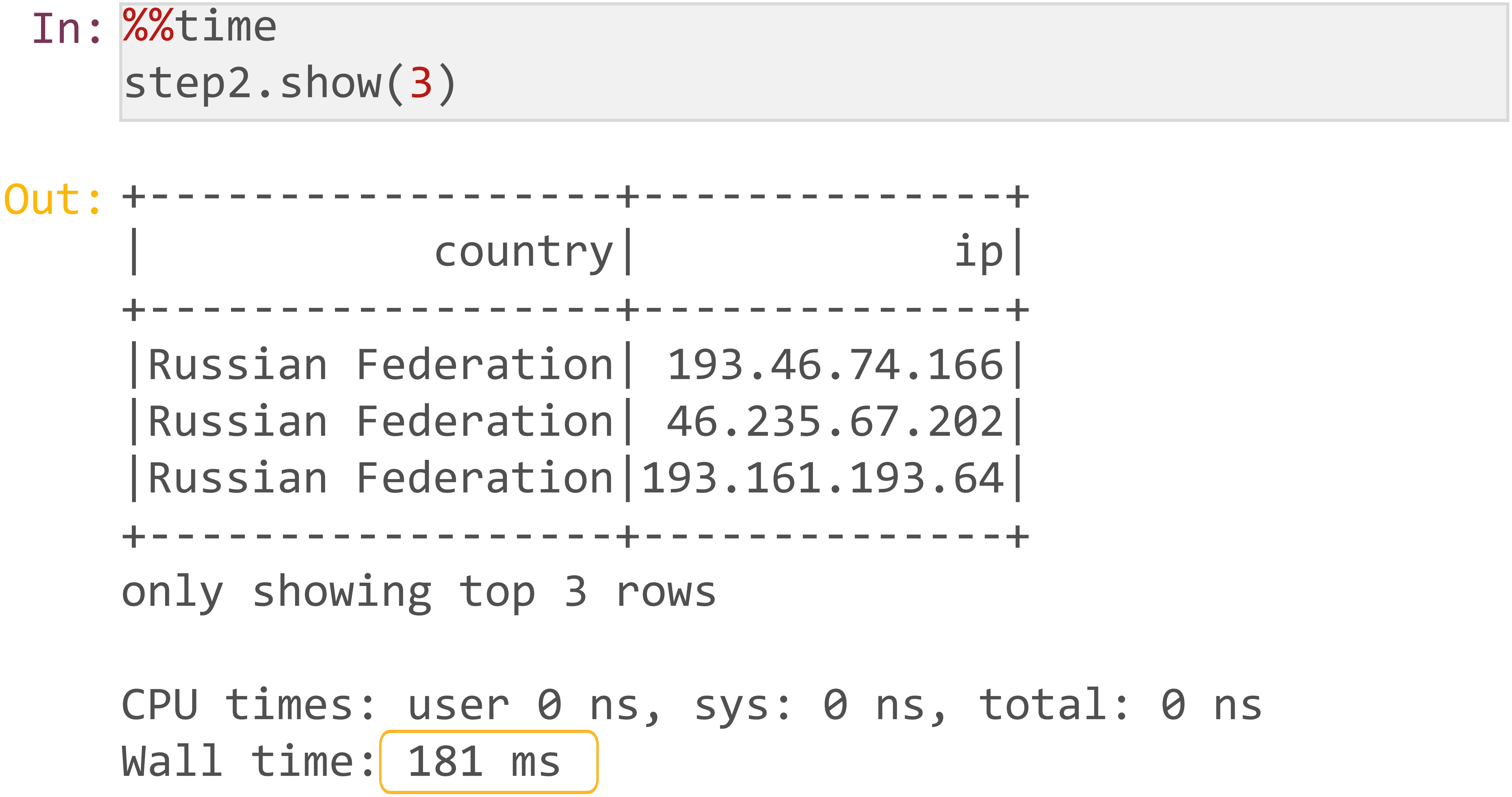
```
In: %%time
    step1 = geoip_df.select("country","ip")

Out: CPU times: user 4 ms, sys: 0 ns, total: 4 ms
     Wall time: 28.8 ms
```

```
In: %%time
    step2 = step1.where("country = 'Russian Federation'")

Out: CPU times: user 4 ms, sys: 0 ns, total: 4 ms
     Wall time: 13.8 ms
```



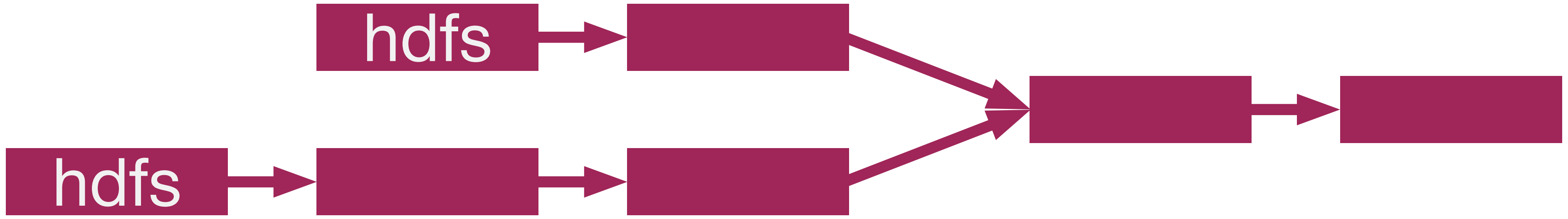
```
 In: %%time
     step2.show(3)
```

```
Out: +------------------+---------------+
     |           country|             ip|
     +------------------+---------------+
     |Russian Federation| 193.46.74.166|
     |Russian Federation| 46.235.67.202|
     |Russian Federation|193.161.193.64|
     +------------------+---------------+
     only showing top 3 rows

     CPU times: user 0 ns, sys: 0 ns, total: 0 ns
     Wall time: 181 ms
```

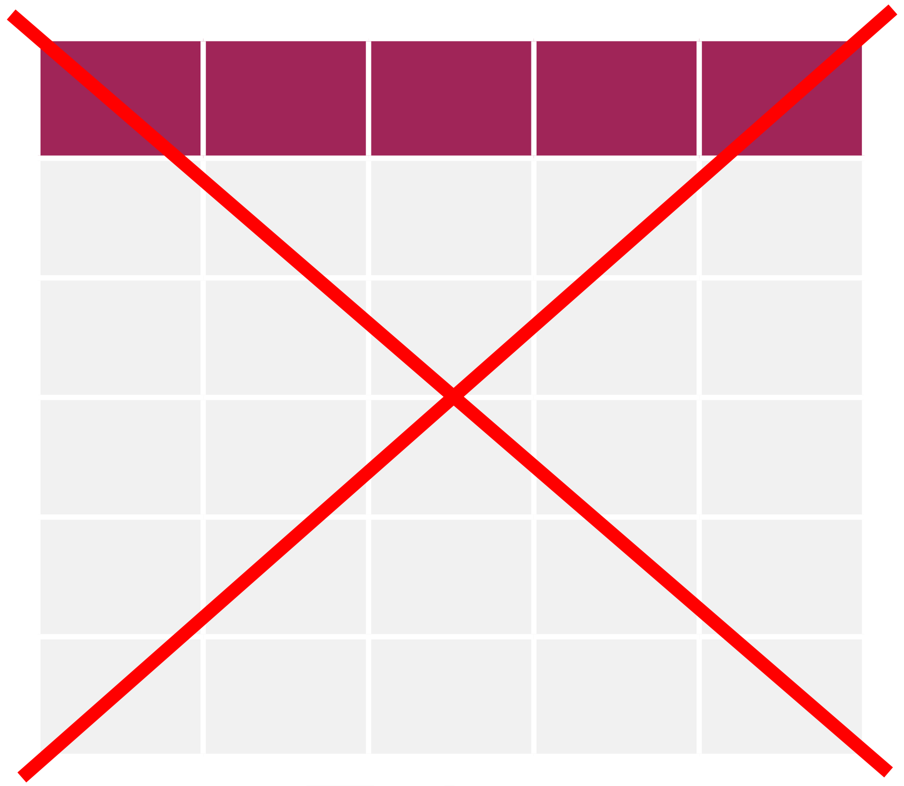


In:
```
%%time
step2.show(3)
```

Out:
```
+------------------+---------------+
|           country|             ip|
+------------------+---------------+
|Russian Federation| 193.46.74.166|
|Russian Federation| 46.235.67.202|
|Russian Federation|193.161.193.64|
+------------------+---------------+
only showing top 3 rows

CPU times: user 0 ns, sys: 0 ns, total: 0 ns
Wall time: 181 ms
```

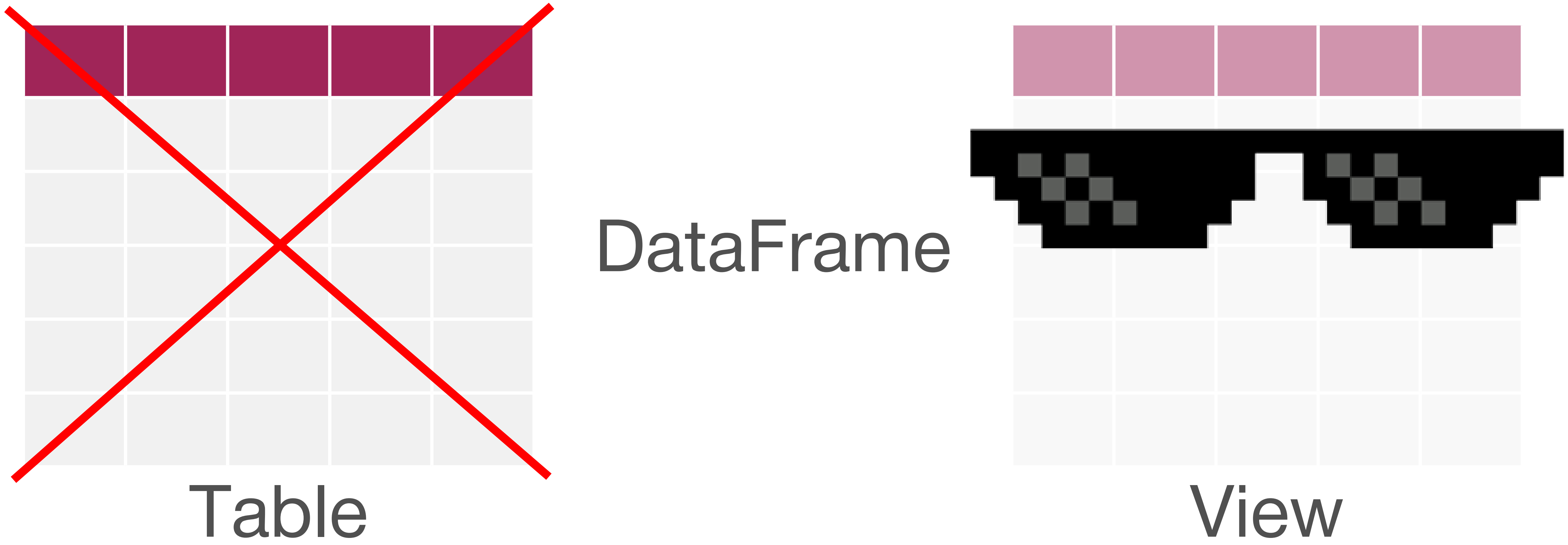


DataFrame



Table

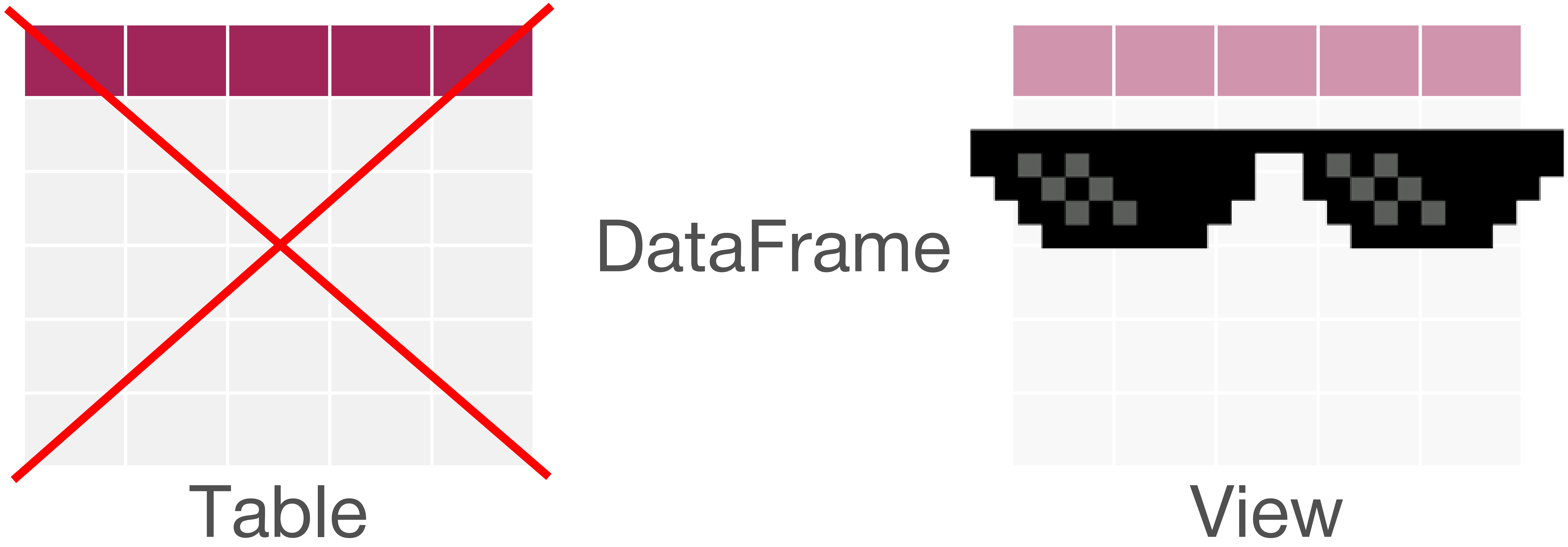
DataFrame



DataFrame

Table

View

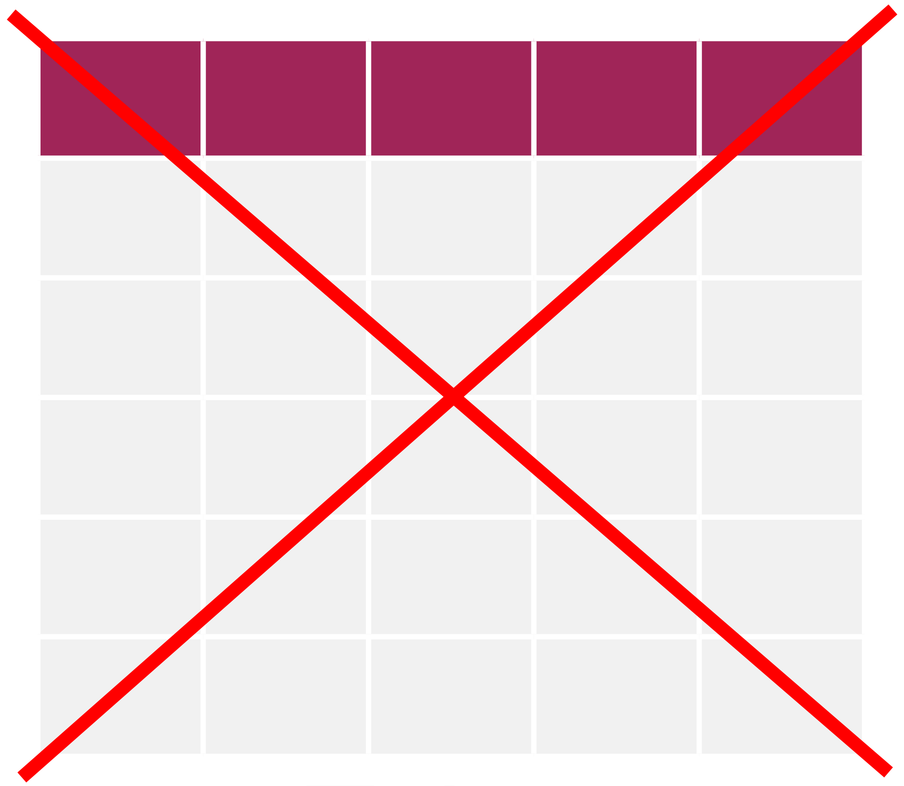


Table

DataFrame

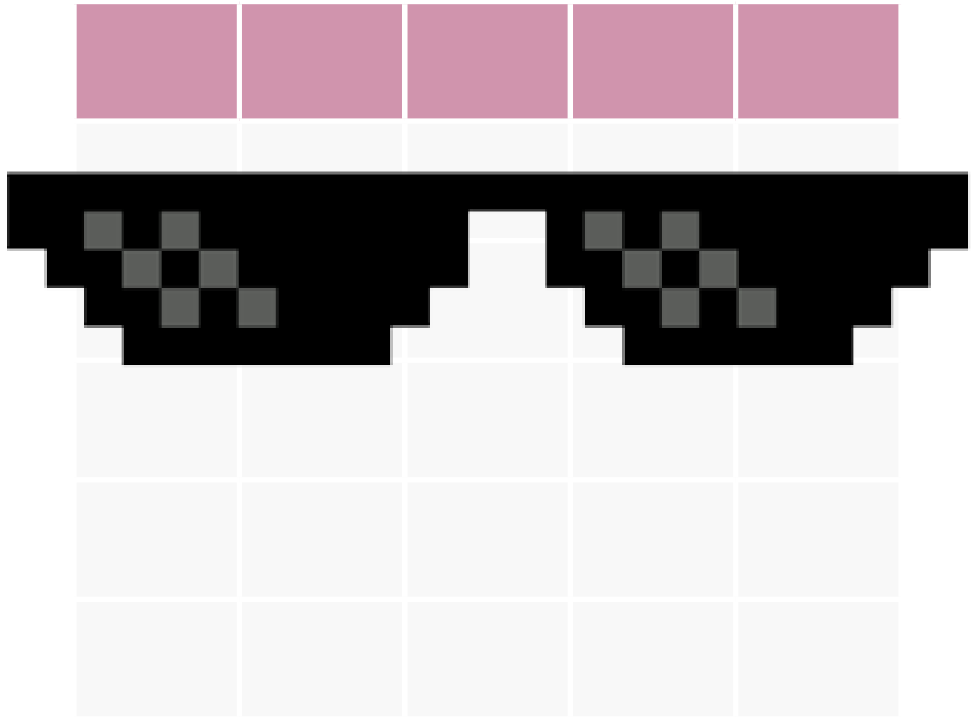
View

```
In: geoip_df.createTempView("geoip")
```



Table

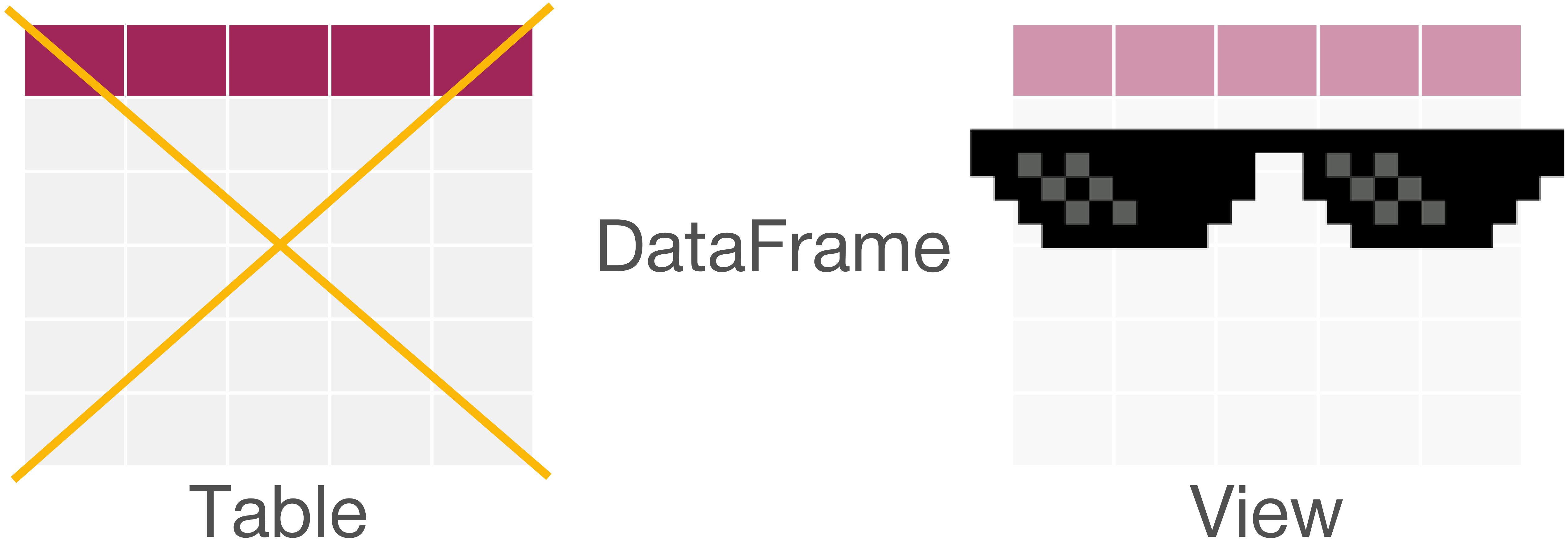
DataFrame



View

```
In: geoip_df.createTempView("geoip")
```

spark_session.sql



Table

DataFrame

View

```
In: geoip_df.createTempView("geoip")
```

spark_session.sql

```
In: spark_session.sql("""
        select country from geoip
    """)
```

```
Out: DataFrame[country: string]
```

```
In: counries_df = spark_session.sql("""
        select country from geoip
    """)
```

In:

```
counries_df = spark_session.sql("""
    select country from geoip
""")
```

In:

```
counries_df.show(3)
```

```
+------------------+
|           country|
+------------------+
|       Netherlands|
|            France|
|Russian Federation|
+------------------+
only showing top 3 rows
```

What have we learned:

- How to make simple operation with spark dataframes
- How to convert them into SQL views
- And how to execute sql queries on them