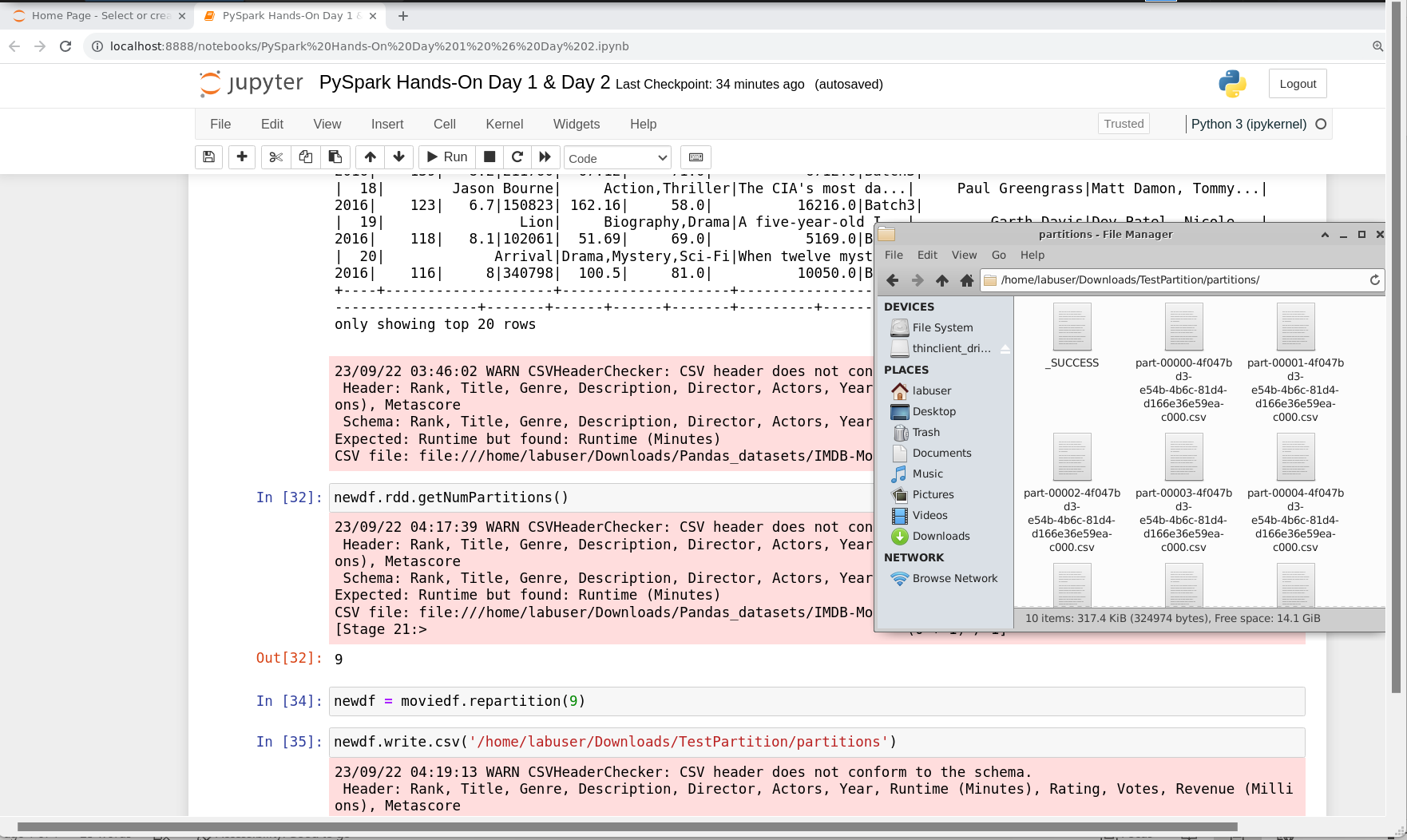
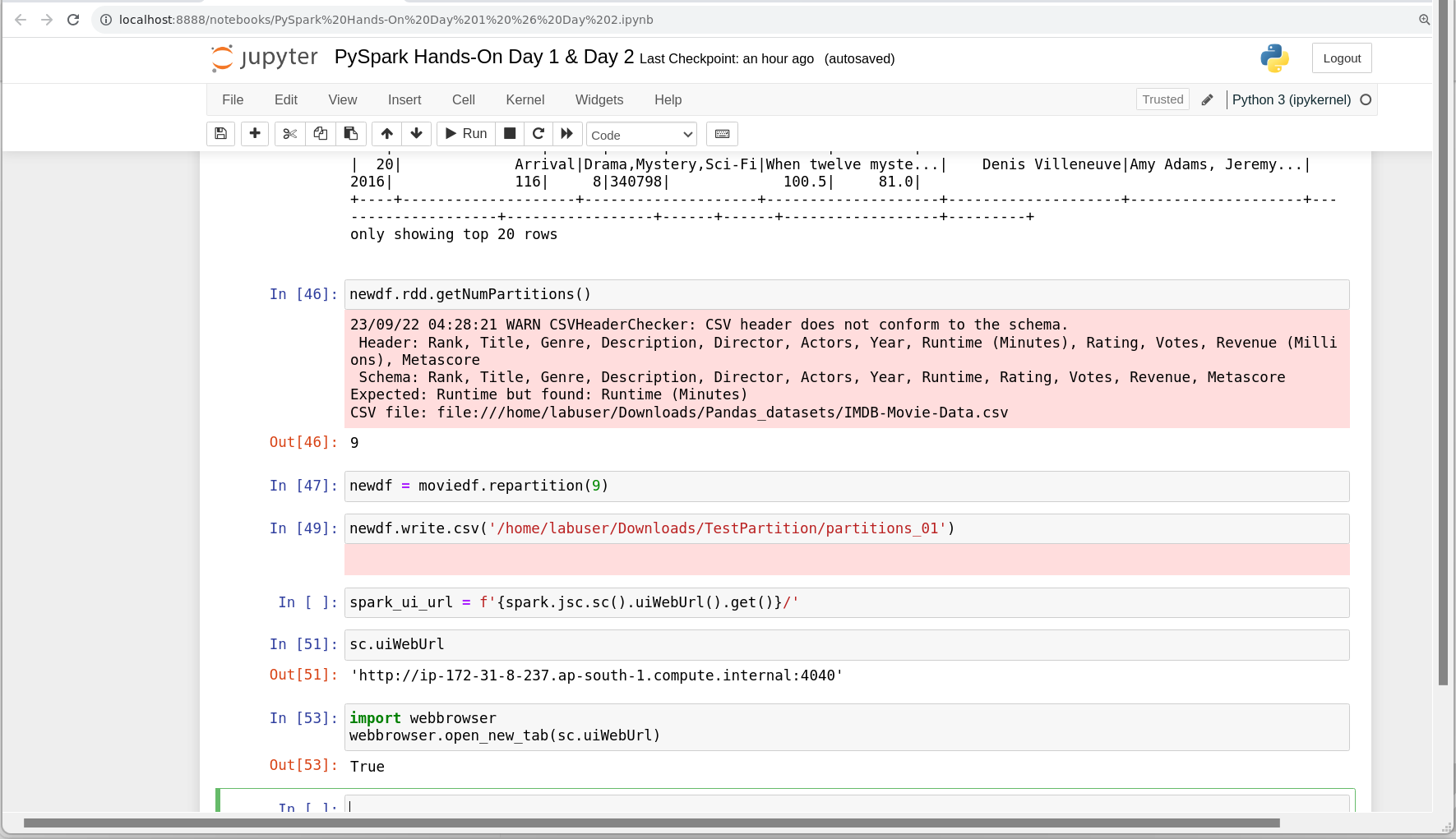
Custom Training

Day 17

**PySpark**

Partition size can be configured according to out need. This is can be re-configured also because it is runtime process.



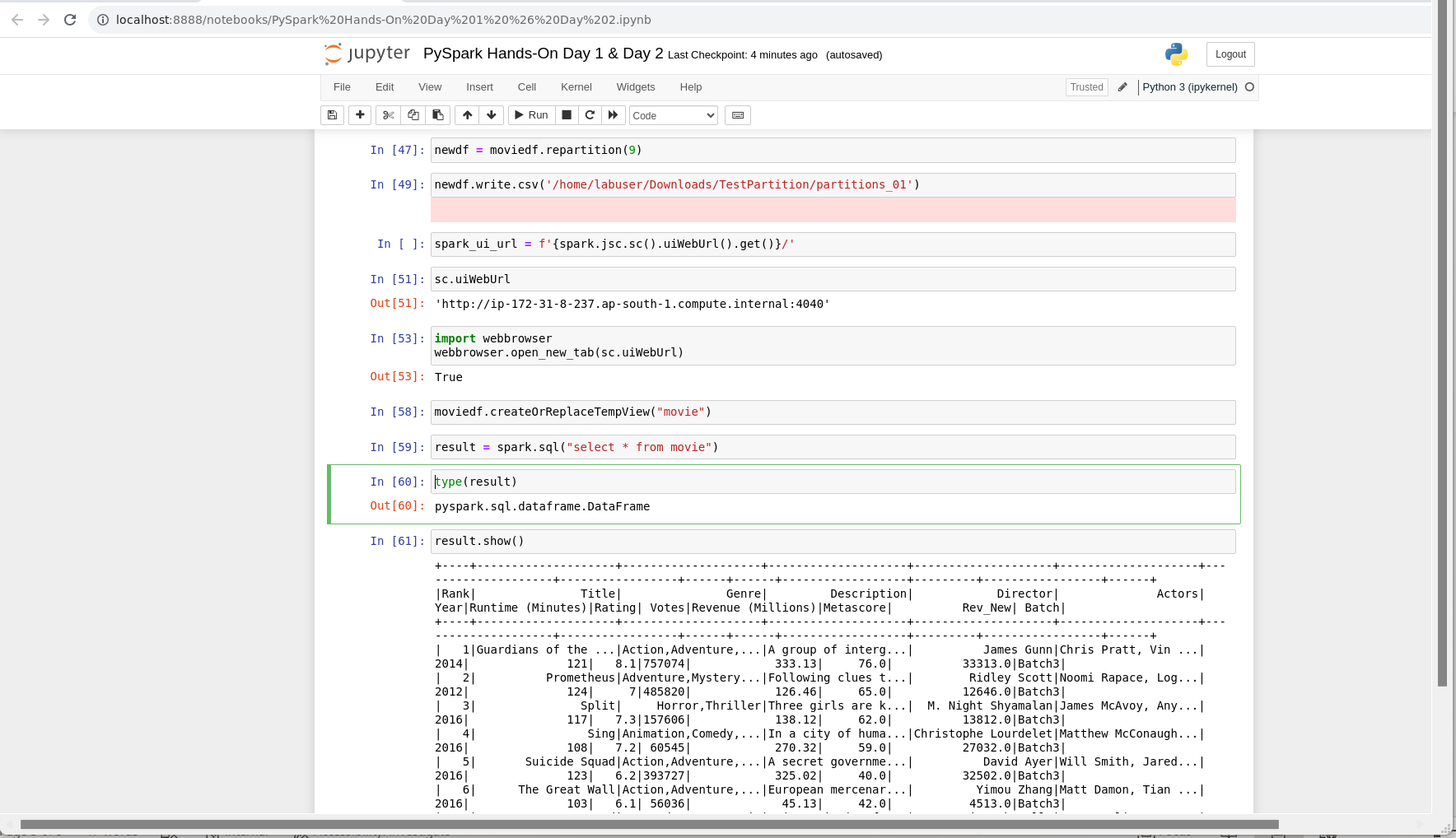


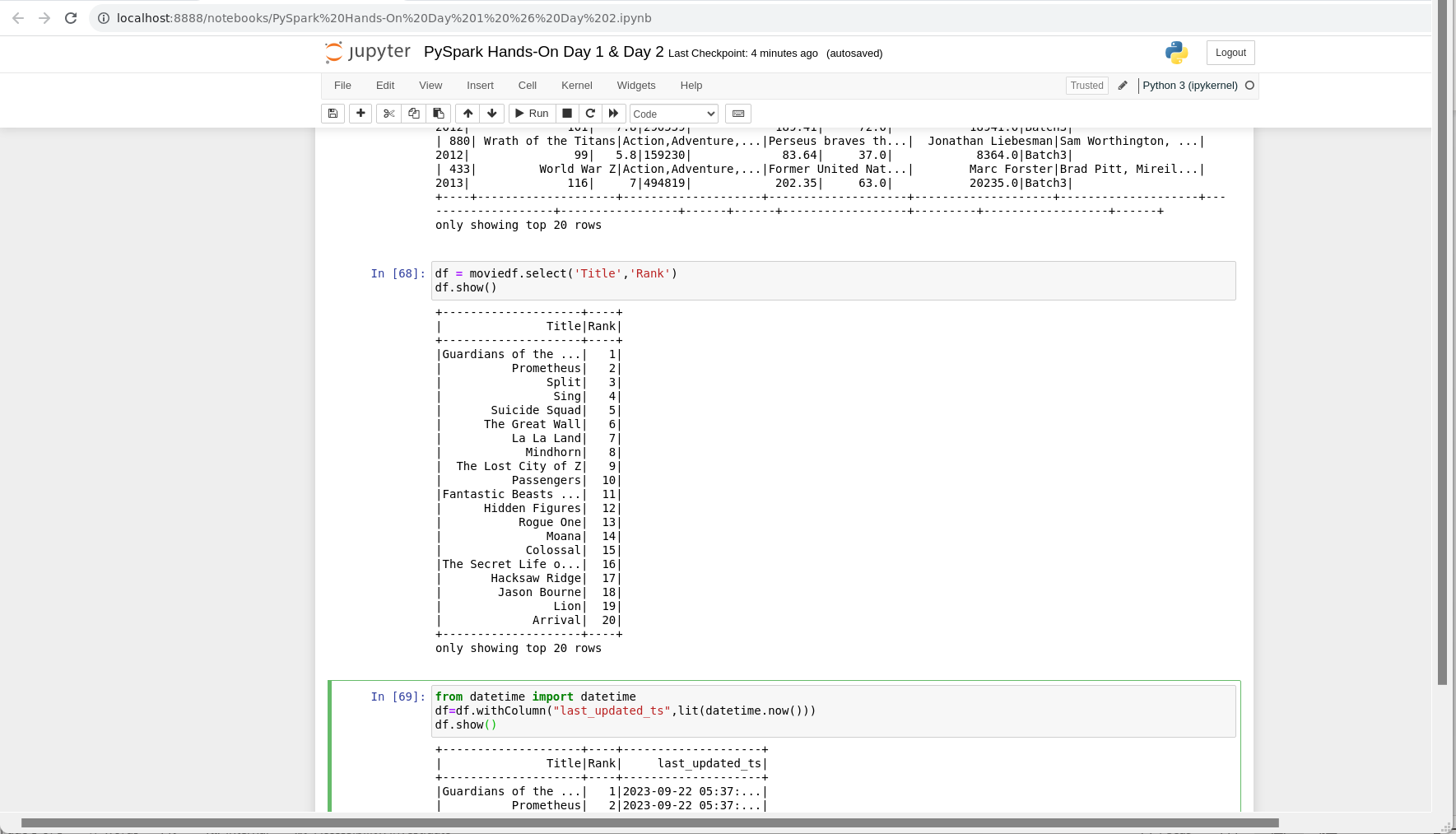
SparkContext is used for RDDs.

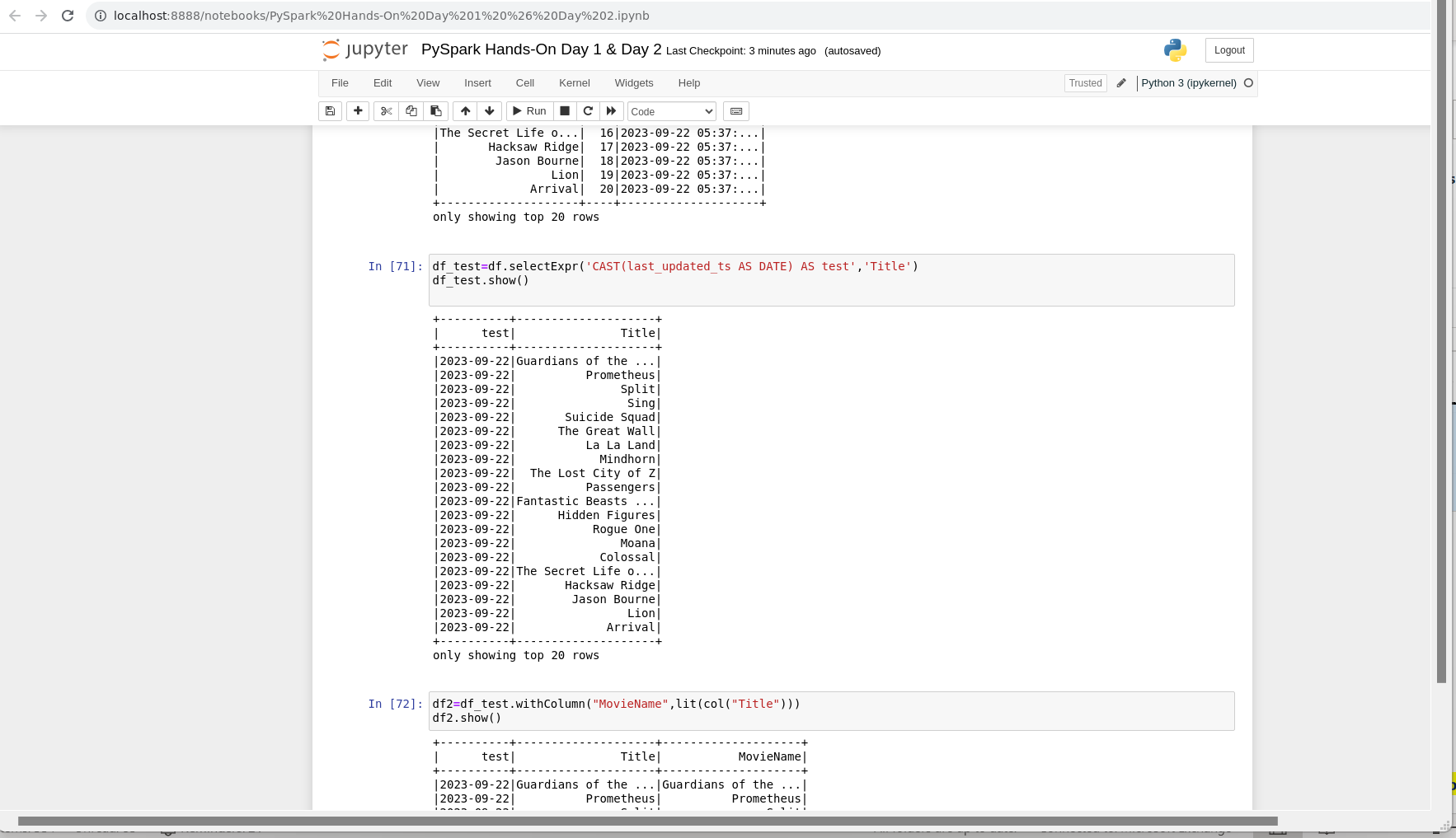
SparkSession is used for dataframes.

Both are entry points.

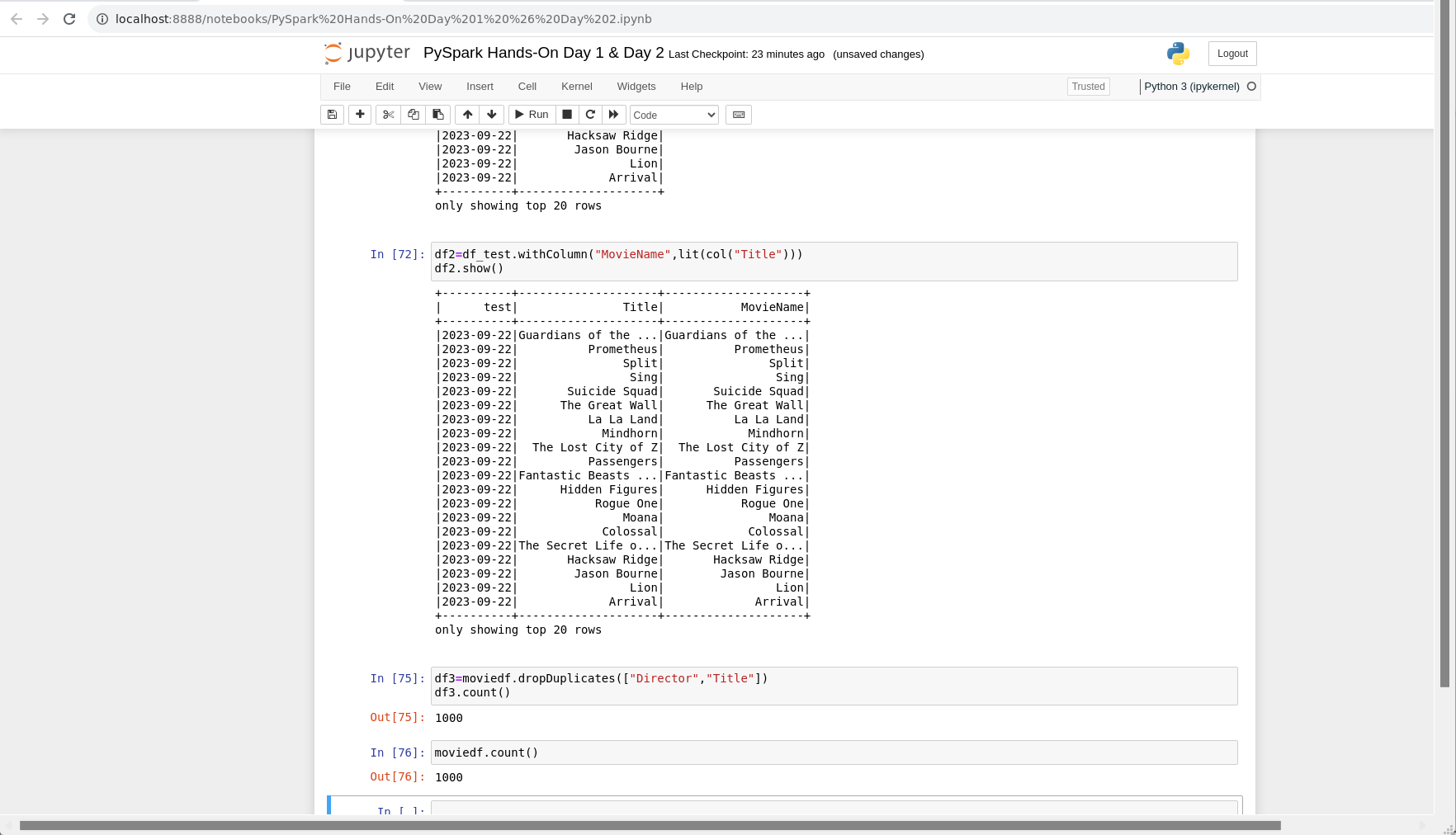
We can run sql commands using spark.sql .







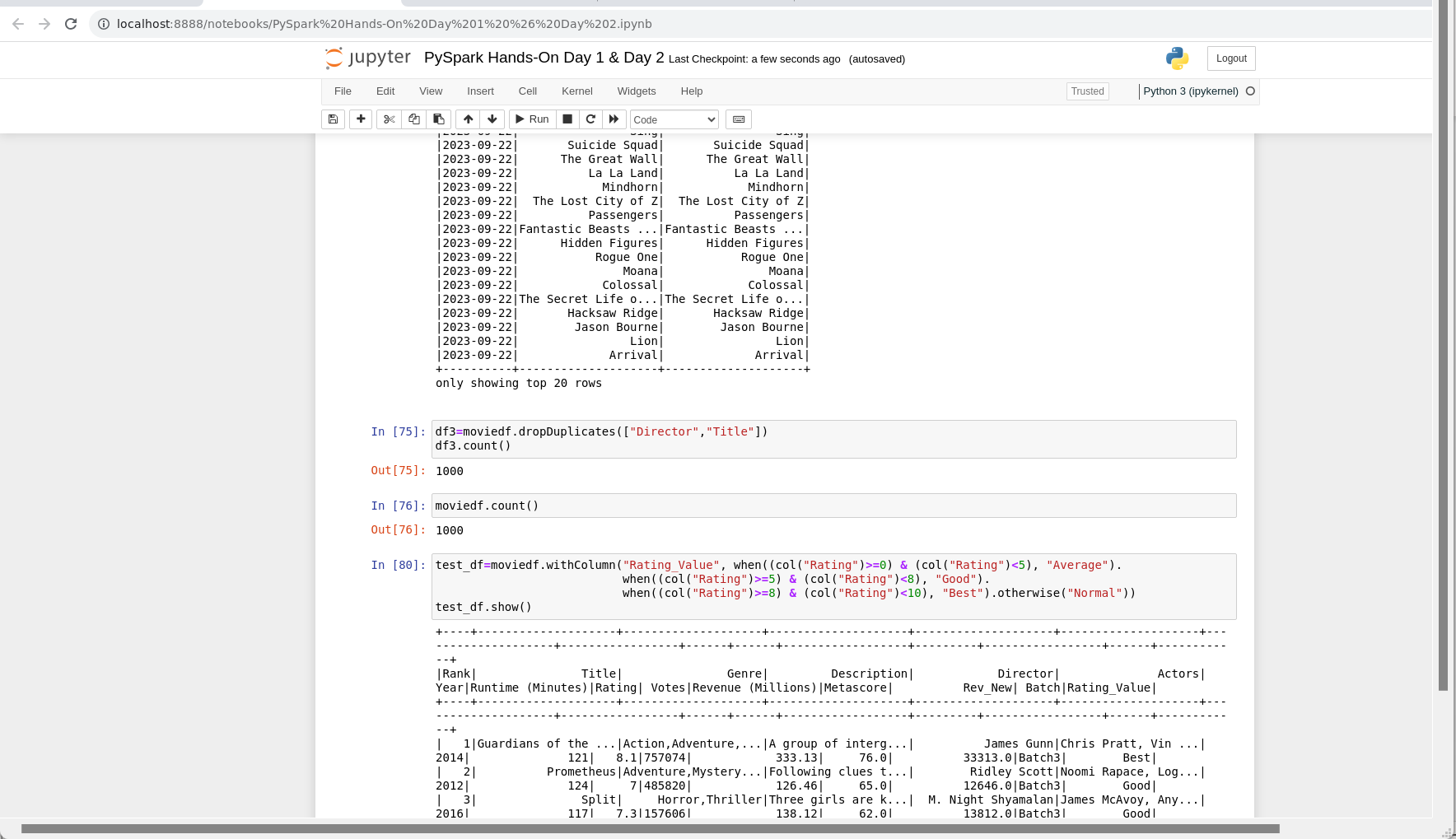
dropDuplicates: We can remove duplicates even by referring 1 or more columns at the same time.



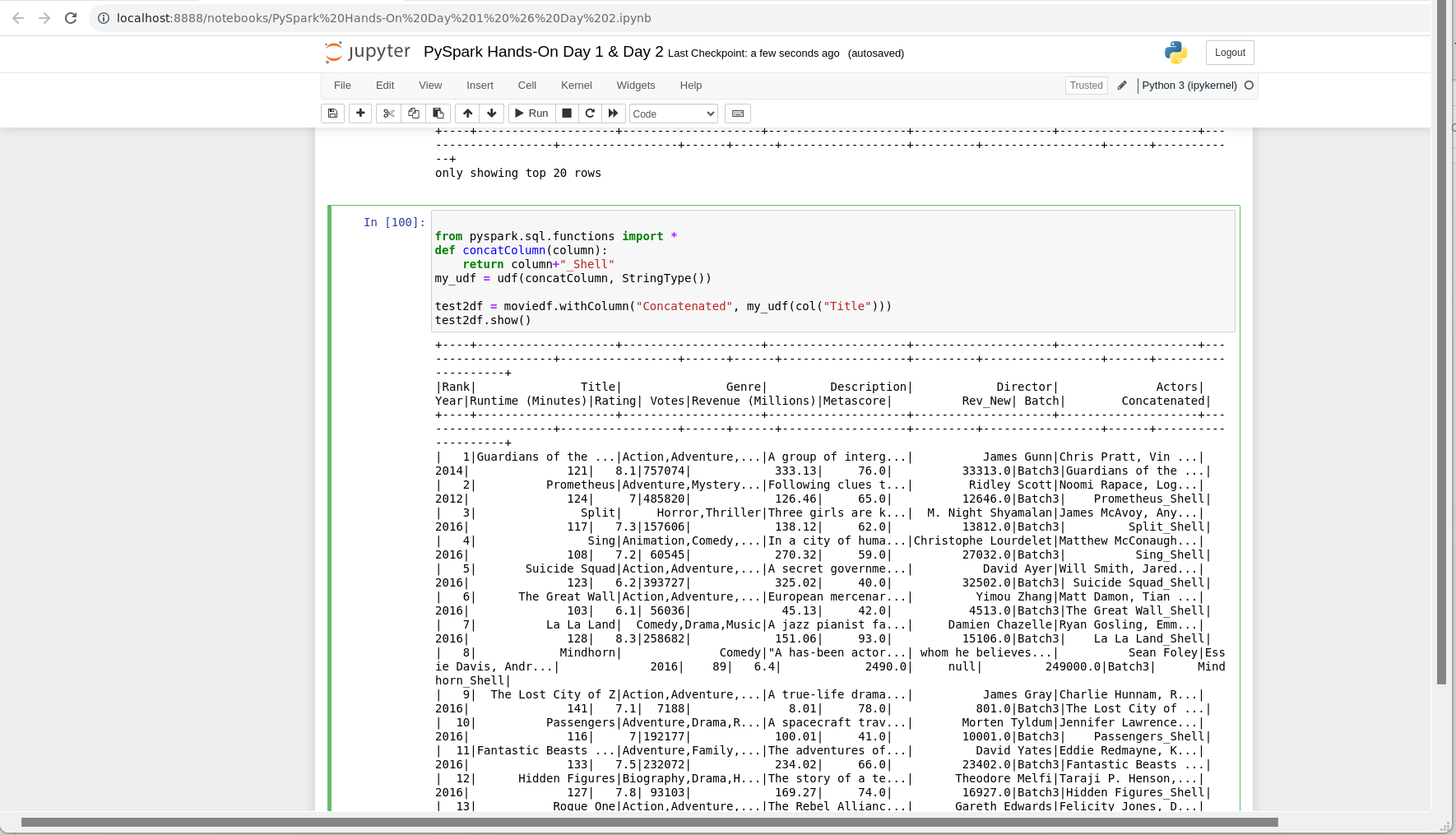
If we want to exclude columns from a huge number of columns:

Newdf=df.drop([excluded columns list])

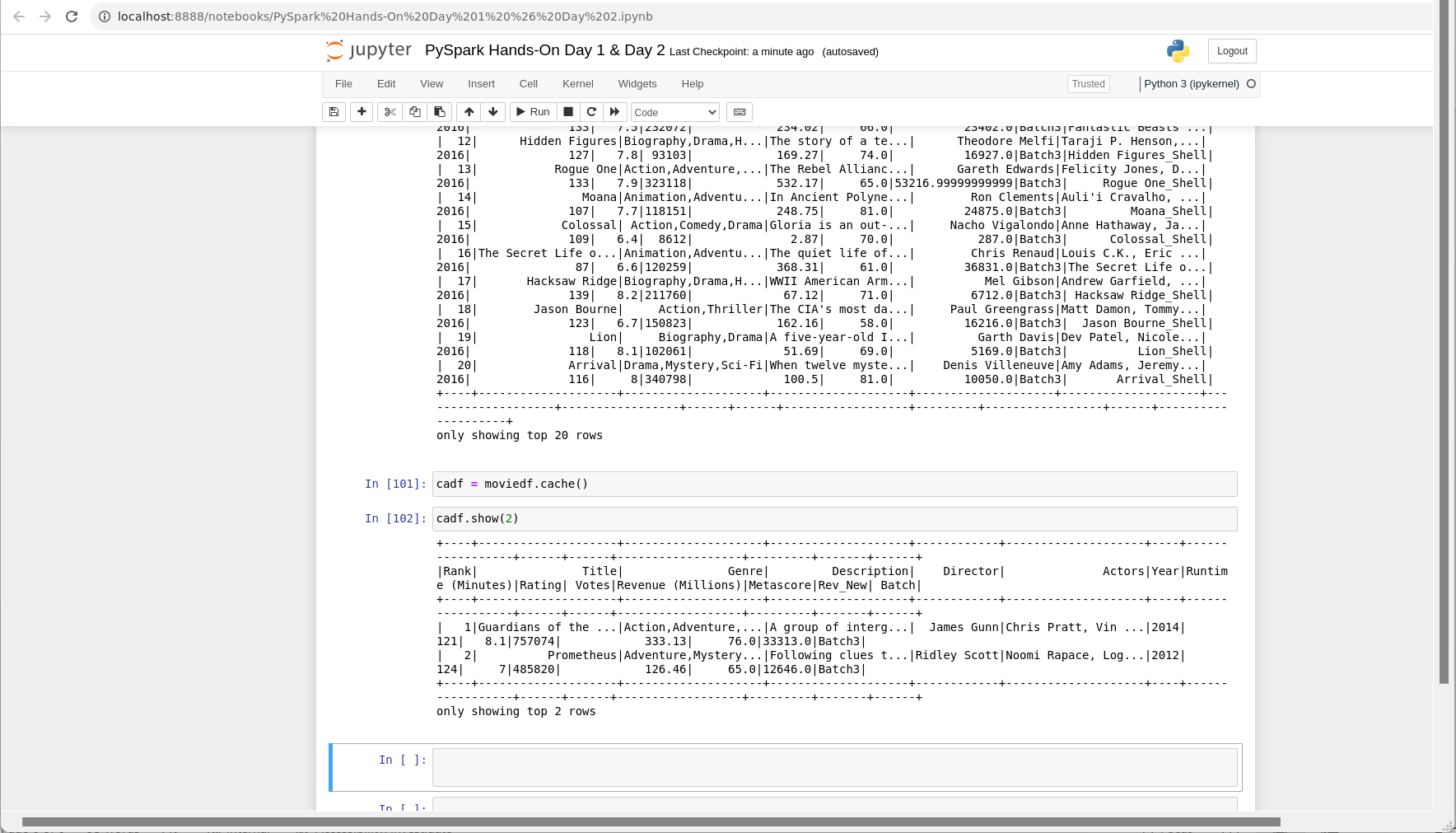
If-Else Condition:



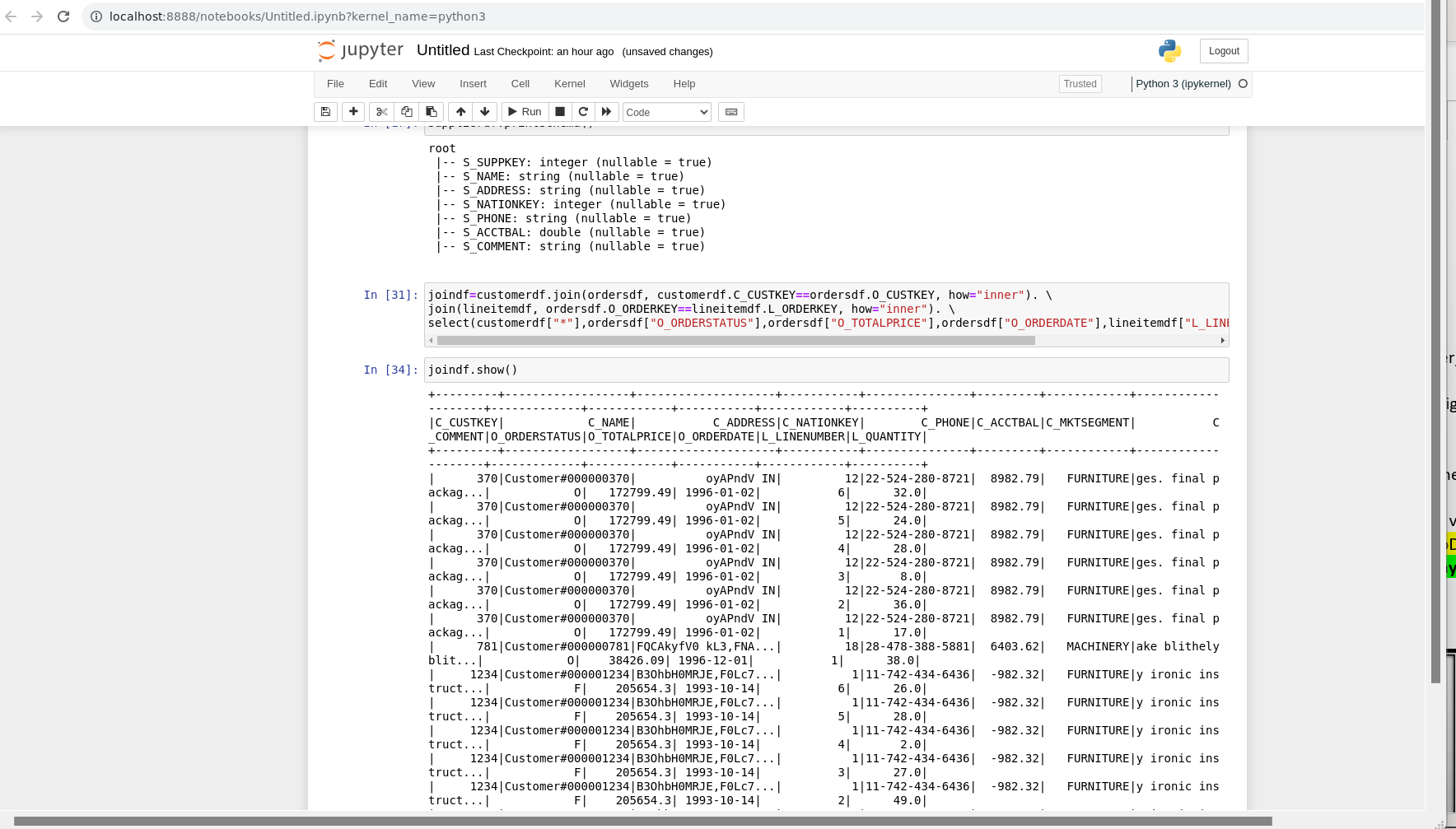
Using User-Defined Functions with PySpark:



Caching: To cache the data for faster fetching of data.



Joins:



Partitioning and Coalesce:

