Row level transformations

Let us understand how to perform row level transformations using orders data set. Here are the details about orders.

- Data is in text file format
- Each line in the file contains one record.
- Each record contains 4 attributes which are separated by ","
    - order_id
    - order_date
    - order_customer_id
    - order_status

```
%%sh

ls -ltr /data/retail_db/orders/part-00000
```

```
-rw-r--r-- 1 root root 2999944 Nov 22 16:08 /data/retail_db/orders/part-00000
```

```
%%sh

tail /data/retail_db/orders/part-00000
```

```
68874,2014-07-03 00:00:00.0,1601,COMPLETE
68875,2014-07-04 00:00:00.0,10637,ON_HOLD
68876,2014-07-06 00:00:00.0,4124,COMPLETE
68877,2014-07-07 00:00:00.0,9692,ON_HOLD
68878,2014-07-08 00:00:00.0,6753,COMPLETE
68879,2014-07-09 00:00:00.0,778,COMPLETE
68880,2014-07-13 00:00:00.0,1117,COMPLETE
68881,2014-07-19 00:00:00.0,2518,PENDING_PAYMENT
68882,2014-07-22 00:00:00.0,10000,ON_HOLD
68883,2014-07-23 00:00:00.0,5533,COMPLETE
```

```
path = '/data/retail_db/orders/part-00000'
# C:\\users\\itversity\\Research\\data\\retail_db\\orders\\part-00000
orders_file = open(path)
```

```
type(orders_file)
```

```
_io.TextIOWrapper
```

```
orders_raw = orders_file.read()
```

```
type(orders_raw)
```

```
str
```

```
orders_raw.splitlines?
```

```
Docstring:
S.splitlines([keepends]) -> list of strings

Return a list of the lines in S, breaking at line boundaries.
Line breaks are not included in the resulting list unless keepends
is given and true.
Type:      builtin_function_or_method
```

```
orders = orders_raw.splitlines()
```

```
type(orders)
```

```
list
```

```
orders[:10]
```

```
['1,2013-07-25 00:00:00.0,11599,CLOSED',
 '2,2013-07-25 00:00:00.0,256,PENDING_PAYMENT',
 '3,2013-07-25 00:00:00.0,12111,COMPLETE',
 '4,2013-07-25 00:00:00.0,8827,CLOSED',
 '5,2013-07-25 00:00:00.0,11318,COMPLETE',
 '6,2013-07-25 00:00:00.0,7130,COMPLETE',
 '7,2013-07-25 00:00:00.0,4530,COMPLETE',
 '8,2013-07-25 00:00:00.0,2911,PROCESSING',
 '9,2013-07-25 00:00:00.0,5657,PENDING_PAYMENT',
 '10,2013-07-25 00:00:00.0,5648,PENDING_PAYMENT']
```

```
type(orders[0])
```

```
str
```

```
len(orders)
```

```
68883
```

```
%%sh

wc -l /data/retail_db/orders/part-00000
```

```
68883 /data/retail_db/orders/part-00000
```

Task 1

Get all order ids and associated statuses. Each record in the output should be comma separated string.

```
order = '1,2013-07-25 00:00:00.0,11599,CLOSED' # -> '1,CLOSED'
```

```
# We invokde join on delimiter
```

```
str.join?
```

```
Docstring:
S.join(iterable) -> str

Return a string which is the concatenation of the strings in the
iterable.  The separator between elements is S.
Type:      method_descriptor
```

```
':'.join(['1', '2', '3', '4'])
```

```
'1:2:3:4'
```

```
order.split(',')[0]
```

```
'1'
```

```
order.split(',')[3]
```

```
'CLOSED'
```

```
[order.split(',')[0], order.split(',')[3]]
```

```
['1', 'CLOSED']
```

```
','.join([order.split(',')[0], order.split(',')[3]])
```

```
'1,CLOSED'
```

```
l = [1]
```

```
l.append(2)
```

```
l
```

```
[1, 2]
```

```
order_statuses = []
for order in orders:
    order_statuses.append(','.join([order.split(',')[0], order.split(',')[3]]))
```

```
order_statuses[:10]
```

```
['1,CLOSED',
 '2,PENDING_PAYMENT',
 '3,COMPLETE',
 '4,CLOSED',
 '5,COMPLETE',
 '6,COMPLETE',
 '7,COMPLETE',
 '8,PROCESSING',
 '9,PENDING_PAYMENT',
 '10,PENDING_PAYMENT']
```

```
len(order_statuses)
```

```
68883
```

```
order_statuses = [','.join([order.split(',')[0], order.split(',')[3]]) for order in orders] # alternative solution
```

```
order_statuses[:10]
```

```
['1,CLOSED',
 '2,PENDING_PAYMENT',
 '3,COMPLETE',
 '4,CLOSED',
 '5,COMPLETE',
 '6,COMPLETE',
 '7,COMPLETE',
 '8,PROCESSING',
 '9,PENDING_PAYMENT',
 '10,PENDING_PAYMENT']
```

```
len(order_statuses)
```

```
68883
```

Task 2

Get all order ids, the dates on which order is placed and order status. Each record in the output should be dict with following column names as keys.

- order_id
- order_date
- order_status

```
{'order_id': 1, 'order_date': '2020-12-22', 'order_status': 'COMPLETE'}
```

```
{'order_id': 1, 'order_date': '2020-12-22', 'order_status': 'COMPLETE'}
```

```
def get_order_details(order):
    """Extract order details such as id, date as well as status and return as dict"""
    order_values = order.split(',')
    return ({
        'order_id': int(order_values[0]),
        'order_date': order_values[1],
        'order_status': order_values[3]
    })
```

```
get_order_details('1,2013-07-25 00:00:00.0,11599,CLOSED')
```

```
{'order_id': 1,
 'order_date': '2013-07-25 00:00:00.0',
 'order_status': 'CLOSED'}
```

```
order_details = []
for order in orders:
    order_details.append(get_order_details(order))
```

```
order_details[:10]
```

```
[{'order_id': 1,
  'order_date': '2013-07-25 00:00:00.0',
  'order_status': 'CLOSED'},
 {'order_id': 2,
  'order_date': '2013-07-25 00:00:00.0',
  'order_status': 'PENDING_PAYMENT'},
 {'order_id': 3,
  'order_date': '2013-07-25 00:00:00.0',
  'order_status': 'COMPLETE'},
 {'order_id': 4,
  'order_date': '2013-07-25 00:00:00.0',
  'order_status': 'CLOSED'},
 {'order_id': 5,
  'order_date': '2013-07-25 00:00:00.0',
  'order_status': 'COMPLETE'},
 {'order_id': 6,
  'order_date': '2013-07-25 00:00:00.0',
  'order_status': 'COMPLETE'},
 {'order_id': 7,
  'order_date': '2013-07-25 00:00:00.0',
  'order_status': 'COMPLETE'},
 {'order_id': 8,
  'order_date': '2013-07-25 00:00:00.0',
  'order_status': 'PROCESSING'},
 {'order_id': 9,
  'order_date': '2013-07-25 00:00:00.0',
  'order_status': 'PENDING_PAYMENT'},
 {'order_id': 10,
  'order_date': '2013-07-25 00:00:00.0',
  'order_status': 'PENDING_PAYMENT'}]
```

```
len(order_details)
```

```
68883
```

```
order_details = [get_order_details(order) for order in orders]
```

```
order_details[:10]
```

```
[{'order_id': 1,
  'order_date': '2013-07-25 00:00:00.0',
  'order_status': 'CLOSED'},
 {'order_id': 2,
  'order_date': '2013-07-25 00:00:00.0',
  'order_status': 'PENDING_PAYMENT'},
 {'order_id': 3,
  'order_date': '2013-07-25 00:00:00.0',
  'order_status': 'COMPLETE'},
 {'order_id': 4,
  'order_date': '2013-07-25 00:00:00.0',
  'order_status': 'CLOSED'},
 {'order_id': 5,
  'order_date': '2013-07-25 00:00:00.0',
  'order_status': 'COMPLETE'},
 {'order_id': 6,
  'order_date': '2013-07-25 00:00:00.0',
  'order_status': 'COMPLETE'},
 {'order_id': 7,
  'order_date': '2013-07-25 00:00:00.0',
  'order_status': 'COMPLETE'},
 {'order_id': 8,
  'order_date': '2013-07-25 00:00:00.0',
  'order_status': 'PROCESSING'},
 {'order_id': 9,
  'order_date': '2013-07-25 00:00:00.0',
  'order_status': 'PENDING_PAYMENT'},
 {'order_id': 10,
  'order_date': '2013-07-25 00:00:00.0',
  'order_status': 'PENDING_PAYMENT'}]
```

```
len(order_details)
```

```
68883
```

By Durga Gadiraju