Reading files into collections

Let us understand how to read data from files into collections.

- Python have simple and yet rich APIs to perform file I/O
- We can create a file object with open in different modes (by default read only mode)
- To read the contents from the file into memory, we have APIs on top of file object such as read()
- read() will create large string using contents of the files
- If the data have multiple records with new line character as delimiter, we can apply splitlines() on the output of read
- splitlines() will convert the string into list with new line character as delimiter

```
%%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
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']
```

```
len(orders) # same as number of records in the file
```

```
68883
```
