list and set - Usage

Let us see some real world usage of list and set while building Python based applications.

- `list` is used more often than `set`.
  - Reading data from file into a `list`
  - Reading data from a table into a `list`
- We can convert a `list` to `set` to perform these operations.
  - Get unique elements from the `list`
  - Perform `set` operations between 2 lists such as union, intersection, difference etc.
- We can convert a `set` to `list` to perform these operations.
  - Reverse the collection
  - Append multiple collections to create new collections while retaining duplicates
- You will see some of these in action as we get into other related topics down the line

```
%%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
```

```
# Reading data from file into a list
path = '/data/retail_db/orders/part-00000'
# C:\\users\\itversity\\Research
orders_file = open(path)
```

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

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

```
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
```

```
# Get unique dates
dates = ['2013-07-25 00:00:00.0', '2013-07-25 00:00:00.0', '2013-07-26 00:00:00.0', '2014-01-25
00:00:00.0']
```

```
dates
```

```
['2013-07-25 00:00:00.0',
 '2013-07-25 00:00:00.0',
 '2013-07-26 00:00:00.0',
 '2014-01-25 00:00:00.0']
```

```
len(dates)
```

```
4
```

```
set(dates)
```

```
{'2013-07-25 00:00:00.0', '2013-07-26 00:00:00.0', '2014-01-25 00:00:00.0'}
```

```
len(dates)
```

```
4
```

```
# Creating new collection retaining duplicates using 2 sets
s1 = {'2013-07-25 00:00:00.0', '2013-07-26 00:00:00.0', '2014-01-25 00:00:00.0'}
```

```
s2 = {'2013-08-25 00:00:00.0', '2013-08-26 00:00:00.0', '2014-01-25 00:00:00.0'}
```

```
s1.union(s2)
```

```
{'2013-07-25 00:00:00.0',
 '2013-07-26 00:00:00.0',
 '2013-08-25 00:00:00.0',
 '2013-08-26 00:00:00.0',
 '2014-01-25 00:00:00.0'}
```

```
len(s1.union(s2))
```

```
5
```

```
s = list(s1) + list(s2)
```

```
s
```

```
['2013-07-26 00:00:00.0',
 '2013-07-25 00:00:00.0',
 '2014-01-25 00:00:00.0',
 '2014-01-25 00:00:00.0',
 '2013-08-26 00:00:00.0',
 '2013-08-25 00:00:00.0']
```

```
len(s)
```

```
6
```