databricksExercise2-Assessment1

```
data = sc.textFile("/FileStore/tables/data.csv")
head = data.first()
data.take(5)
```

```
Out[37]: ['InvoiceNo,StockCode,Description,Quantity,InvoiceDate,UnitPrice,Cust
omerID,Country',
 '536365,85123A,WHITE HANGING HEART T-LIGHT HOLDER,6,12/1/2010 8:26,2.55,1785
0,United Kingdom',
 '536365,71053,WHITE METAL LANTERN,6,12/1/2010 8:26,3.39,17850,United Kingdo
m',
 '536365,84406B,CREAM CUPID HEARTS COAT HANGER,8,12/1/2010 8:26,2.75,17850,Uni
ted Kingdom',
 '536365,84029G,KNITTED UNION FLAG HOT WATER BOTTLE,6,12/1/2010 8:26,3.39,1785
0,United Kingdom']
```

```
import re


data_produced = data.filter(lambda line: line != head)\
               .map(lambda line: re.sub(r'(?!(([^"]*"){2})*[^"]*$),', '',
line))\
               .map(lambda line: line.split(','))\
               .map (lambda arr: list(map(str.strip, arr)))\
               .map(lambda arr: arr[:3] + [ int(arr[3]), arr[4].split(' ')
[0],float(arr[5]) ] + arr[6:] )\
               .filter(lambda row: ( (row[3] > 0) and row[0].isnumeric() ) )



data_produced.collect()
```

```
Out[40]: [['536365',
  '85123A',
  'WHITE HANGING HEART T-LIGHT HOLDER',
  6,
  '12/1/2010',
  2.55,
  '17850',
  'United Kingdom'],
 ['536365',
  '71053',
  'WHITE METAL LANTERN',
  6,
```

```
  '12/1/2010',
  3.39,
  '17850',
  'United Kingdom'],
 ['536365',
  '84406B',
  'CREAM CUPID HEARTS COAT HANGER',
  8,
```

```
data_rdd_4 = data_produced.filter(lambda arr: tuple(arr))
data_rdd_4.take(5)
```

```
Out[47]: [['536365',
  '85123A',
  'WHITE HANGING HEART T-LIGHT HOLDER',
  6,
  '12/1/2010',
  2.55,
  '17850',
  'United Kingdom'],
 ['536365',
  '71053',
  'WHITE METAL LANTERN',
  6,
  '12/1/2010',
  3.39,
  '17850',
  'United Kingdom'],
 ['536365',
  '84406B',
  'CREAM CUPID HEARTS COAT HANGER',
  8,
  '12/1/2010',
```

```
damaged_list = ['lost','damaged','Damaged','unsaleable','throw
away','lost??','check','LOST','MISSING','DAMAGED','missing','AWAY','UNSALEABLE'
]
```

```
Filter_damaged_rdd = data_rdd_4.filter(lambda row: (True if
(row[0].startswith("C")) else False)| (True if(any(ele in row[2] for ele in
damaged_list)) else False))
Damaged_products = Filter_damaged_rdd.map(lambda row:(row[1],
(row[2],row[3],row[4])))
#Damaged_products.take(5)
```

```
Out[82]: [('21830', ('damaged', 192, '9/16/2011')),
 ('85135B', ('check', 3, '10/28/2011')),
 ('22117', ('check', 184, '10/31/2011')),
 ('46000U', ('check', 10, '11/1/2011')),
 ('22812', ('check', 48, '11/1/2011')),
 ('84050', ('check', 14, '11/1/2011')),
 ('47503A', ('check', 65, '11/2/2011')),
 ('21644', ('check', 27, '11/2/2011')),
 ('35968', ('check', 9, '11/2/2011')),
 ('21539', ('check', 4, '11/2/2011')),
 ('21700', ('check', 26, '11/2/2011')),
 ('82600', ('check', 3, '11/2/2011')),
 ('16207A', ('check', 2, '11/2/2011')),
 ('22848', ('check', 2, '11/2/2011')),
 ('84659A', ('check', 19, '11/2/2011')),
 ('21349', ('check', 1, '11/2/2011')),
 ('22925', ('check', 1, '11/2/2011')),
 ('23091', ('check', 2, '11/2/2011')),
 ('22606', ('check', 1, '11/2/2011')),
 ('21804', ('check', 36, '11/2/2011')),
 ('10080', ('check', 22, '11/10/2011')),
```