RemoveDups

November 20, 2016

```
In [1]: import pandas as pd

In [7]: user_cols = ['user_id', 'age', 'gender', 'occupation', 'zip_code']
        users = pd.read_table('http://bit.ly/movieusers', sep='|', header=None, nam

In [3]:

Out[3]:    order_id  quantity                              item_name  \
        0         1         1           Chips and Fresh Tomato Salsa
        1         1         1                                   Izze
        2         1         1                       Nantucket Nectar
        3         1         1  Chips and Tomatillo-Green Chili Salsa
        4         2         2                           Chicken Bowl

                                         choice_description item_price
        0                                               NaN     $2.39
        1                                      [Clementine]     $3.39
        2                                           [Apple]     $3.39
        3                                               NaN     $2.39
        4  [Tomatillo-Red Chili Salsa (Hot), [Black Beans...    $16.98

In [8]: users.shape

Out[8]: (943, 4)

In [10]: users.zip_code.duplicated().head() # returns True if entry previous to it

Out[10]: user_id
         1    False
         2    False
         3    False
         4    False
         5    False
         Name: zip_code, dtype: bool

In [11]: users.zip_code.duplicated().sum() # find number of dupes

Out[11]: 148
```

```
In [12]: users.duplicated().head() # if entire row is identical to previous row, ou

Out[12]: user_id
         1    False
         2    False
         3    False
         4    False
         5    False
         dtype: bool

In [13]: users.duplicated().sum() # sum of all rows for this

Out[13]: 7

In [14]: users.loc[users.duplicated(), :] # identifies seven rows that are duplicat

Out[14]:          age gender occupation zip_code
         user_id
         496       21      F    student    55414
         572       51      M   educator    20003
         621       17      M    student    60402
         684       28      M    student    55414
         733       44      F      other    60630
         805       27      F      other    20009
         890       32      M    student    97301

In [15]: users.loc[users.duplicated(keep='first'), :]
         # mark dupes as true except for first occurrence

Out[15]:          age gender occupation zip_code
         user_id
         496       21      F    student    55414
         572       51      M   educator    20003
         621       17      M    student    60402
         684       28      M    student    55414
         733       44      F      other    60630
         805       27      F      other    20009
         890       32      M    student    97301

In [16]: users.loc[users.duplicated(keep = 'last'), :]
         # want to keep last => keeping the later dupes as opposed to first occurre

Out[16]:          age gender occupation zip_code
         user_id
         67        17      M    student    60402
         85        51      M   educator    20003
         198       21      F    student    55414
         350       32      M    student    97301
         428       28      M    student    55414
         437       27      F      other    20009
         460       44      F      other    60630
```

```
In [17]: # marks all dupes as True so you see all occurrences of duped data
         users.loc[users.duplicated(keep=False), :]
```

```
Out[17]:          age gender occupation zip_code
         user_id
         67        17      M    student    60402
         85        51      M   educator    20003
         198       21      F    student    55414
         350       32      M    student    97301
         428       28      M    student    55414
         437       27      F      other    20009
         460       44      F      other    60630
         496       21      F    student    55414
         572       51      M   educator    20003
         621       17      M    student    60402
         684       28      M    student    55414
         733       44      F      other    60630
         805       27      F      other    20009
         890       32      M    student    97301
```

```
In [18]: # default is to keep the first and drop the last
         # notice initial 943 so 7 dupes dropped
         # does not keep inplace by default
         users.drop_duplicates(keep='first').shape
```

```
Out[18]: (936, 4)
```

```
In [19]: # drops all duplicates => both first AND last => 14 total
         users.drop_duplicates(keep=False).shape
```

```
Out[19]: (929, 4)
```

```
In [20]: # in age and zip_code, there are 16 total rows with dupes,
         users.duplicated(subset=['age', 'zip_code']).sum()
```

```
Out[20]: 16
```

```
In [21]: # drops those 16 rows
         users.drop_duplicates(subset=['age','zip_code']).shape
```

```
Out[21]: (927, 4)
```

```
In [ ]:
```