12 - Pandas-Reshape

October 14, 2021

Table of Contents

[1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib as mpl
import seaborn as sns
from numpy.random import randn
```

[2]:
```
acs =    pd.read_csv('/Users/jimcody/Documents/2021Python/intropython/data/ga_a.
 ↪csv')
places = pd.read_csv('/Users/jimcody/Documents/2021Python/intropython/data/ga_p.
 ↪csv')

print(acs.shape)
print(places.shape)
```

```
(159, 13)
(159, 19)
```

[3]:
```
places.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 159 entries, 0 to 158
Data columns (total 19 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   Unnamed: 0       159 non-null    int64 
 1   StateAbbr        159 non-null    object
 2   StateDesc        159 non-null    object
 3   CountyName       159 non-null    object
 4   CountyFIPS       159 non-null    int64 
 5   TotalPopulation  159 non-null    int64 
```

```
 6   harthritis       159 non-null    float64
 7   hasthma          159 non-null    float64
 8   hbphigh          159 non-null    float64
 9   hcancer          159 non-null    float64
 10  hhighchol        159 non-null    float64
 11  hkidney          159 non-null    float64
 12  hcopd            159 non-null    float64
 13  hchd             159 non-null    float64
 14  hdiabetes        159 non-null    float64
 15  hmhlth           159 non-null    float64
 16  hphlth           159 non-null    float64
 17  hteethlost       159 non-null    float64
 18  hstroke          159 non-null    float64
dtypes: float64(13), int64(3), object(3)
memory usage: 23.7+ KB
```

[4]:
```
places = places.drop('Unnamed: 0',axis=1)
places.head()
```

[4]:

	StateAbbr	StateDesc	CountyName	CountyFIPS	TotalPopulation	harthritis	hasthma	hbphigh	hcancer	hhighchol	hkidney	hcopd	hchd	hdiabetes	hmhlth	hphlth	hteethlost	hstroke
0	GA	Georgia	Appling	13001	18507	32.2	10.5	39.0	7.4	37.8	4.1	11.8	10.0	16.3	16.7	18.2	24.6	5.1
1	GA	Georgia	Atkinson	13003	8297	29.0	10.7	38.3	6.5	37.3	4.1	11.8	9.8	16.8	17.7	18.8	28.1	5.1
2	GA	Georgia	Bacon	13005	11185	30.5	10.4	37.1	7.1	36.2	3.8	11.3	9.3	15.2	16.8	17.6	22.9	4.8
3	GA	Georgia	Baker	13007	3092	32.6	10.3	44.2	8.0	38.3	4.3	10.4	9.7	18.2	14.3	16.8	21.3	5.6
4	GA	Georgia	Baldwin	13009	44823	27.2	10.5	38.4	6.3	34.8	3.6	9.1	8.1	14.9	15.9	15.0	20.8	4.5

[5]:
```
df1 = {
    'fruit':['apples','pears','oranges','mangos'],
    'Jan':[100,87,45,56],
    'Feb':[78,43,78,89],
    'Mar':[34,67,54,98],
    'Apr':[102,98,105,154],
    'May':[1,2,3,4],
    'Jun':[10,20,40,70]
```

```
}
d1 = pd.DataFrame(df1)
d1
```

[5]:

	fruit	Jan	Feb	Mar	Apr	May	Jun
0	apples	100	78	34	102	1	10
1	pears	87	43	67	98	2	20
2	oranges	45	78	54	105	3	40
3	mangos	56	89	98	154	4	70

1 Melt - make a wide table narrow

[6]:
```
# id_vars -
# var_name -
# value_name -

d1.melt(id_vars=['fruit'])
```

[6]:

	fruit	variable	value
0	apples	Jan	100
1	pears	Jan	87
2	oranges	Jan	45
3	mangos	Jan	56
4	apples	Feb	78
5	pears	Feb	43
6	oranges	Feb	78
7	mangos	Feb	89
8	apples	Mar	34
9	pears	Mar	67
10	oranges	Mar	54
11	mangos	Mar	98
12	apples	Apr	102
13	pears	Apr	98
14	oranges	Apr	105
15	mangos	Apr	154
16	apples	May	1
17	pears	May	2
18	oranges	May	3
19	mangos	May	4
20	apples	Jun	10
21	pears	Jun	20
22	oranges	Jun	40
23	mangos	Jun	70

[7]:
```
#d1.melt(id_vars=['fruit'], var_name = 'Month', value_name = 'Picked')
d1.melt(id_vars=['fruit'], var_name = 'Month', value_name = 'Picked').
 ↪sort_values(by = 'fruit')
```

```
#d1.melt(id_vars=['fruit'], var_name = 'Month', value_name = 'Picked').
↪sort_values(by = ['fruit', 'Month'])
```

[7]:

	fruit	Month	Picked
0	apples	Jan	100
20	apples	Jun	10
16	apples	May	1
4	apples	Feb	78
12	apples	Apr	102
8	apples	Mar	34
19	mangos	May	4
15	mangos	Apr	154
11	mangos	Mar	98
3	mangos	Jan	56
7	mangos	Feb	89
23	mangos	Jun	70
10	oranges	Mar	54
22	oranges	Jun	40
6	oranges	Feb	78
14	oranges	Apr	105
18	oranges	May	3
2	oranges	Jan	45
9	pears	Mar	67
13	pears	Apr	98
5	pears	Feb	43
17	pears	May	2
1	pears	Jan	87
21	pears	Jun	20

2 Stack

[8]:
```
d1
```

[8]:

	fruit	Jan	Feb	Mar	Apr	May	Jun
0	apples	100	78	34	102	1	10
1	pears	87	43	67	98	2	20
2	oranges	45	78	54	105	3	40
3	mangos	56	89	98	154	4	70

[9]:
```
d1.stack()
```

[9]:
```
0  fruit     apples
   Jan          100
   Feb           78
   Mar           34
   Apr          102
   May            1
```

```
   Jun           10
1  fruit      pears
   Jan           87
   Feb           43
   Mar           67
   Apr           98
   May            2
   Jun           20
2  fruit    oranges
   Jan           45
   Feb           78
   Mar           54
   Apr          105
   May            3
   Jun           40
3  fruit     mangos
   Jan           56
   Feb           89
   Mar           98
   Apr          154
   May            4
   Jun           70
dtype: object
```

[10]:
```
fruit = ['apples','pears','oranges','mangos']
df2 = {
 #   'fruit':['apples','pears','oranges','mangos'],
    'Jan':[100,87,45,56],
    'Feb':[78,43,78,89],
    'Mar':[34,67,54,98],
    'Apr':[102,98,105,154],
    'May':[1,2,3,4],
    'Jun':[10,20,40,70]
}
d2 = pd.DataFrame(df2, index = fruit)
d2
```

[10]:

	Jan	Feb	Mar	Apr	May	Jun
apples	100	78	34	102	1	10
pears	87	43	67	98	2	20
oranges	45	78	54	105	3	40
mangos	56	89	98	154	4	70

[11]:
```
d2.stack()
```

[11]:
```
apples   Jan    100
         Feb     78
```

```
          Mar     34
          Apr    102
          May      1
          Jun     10
pears     Jan     87
          Feb     43
          Mar     67
          Apr     98
          May      2
          Jun     20
oranges   Jan     45
          Feb     78
          Mar     54
          Apr    105
          May      3
          Jun     40
mangos    Jan     56
          Feb     89
          Mar     98
          Apr    154
          May      4
          Jun     70
dtype: int64
```

[12]:
```
fruit = ['apples','pears','apple','pears']
state = ['MA','MA','VT','VT']
df2 = {
 #   'fruit':['apples','pears','oranges','mangos'],
    'Jan':[100,87,45,56],
    'Feb':[78,43,78,89],
    'Mar':[34,67,54,98],
    'Apr':[102,98,105,154],
    'May':[1,2,3,4],
    'Jun':[10,20,40,70]
}
d2 = pd.DataFrame(df2, index = [state,fruit])
d2
```

[12]:

		Jan	Feb	Mar	Apr	May	Jun
MA	apples	100	78	34	102	1	10
	pears	87	43	67	98	2	20
VT	apple	45	78	54	105	3	40
	pears	56	89	98	154	4	70

[13]:
```
d2_stacked = d2.stack()
d2_stacked
```

[13]:
```
MA  apples  Jan    100
            Feb     78
            Mar     34
            Apr    102
            May      1
            Jun     10
    pears   Jan     87
            Feb     43
            Mar     67
            Apr     98
            May      2
            Jun     20
VT  apple   Jan     45
            Feb     78
            Mar     54
            Apr    105
            May      3
            Jun     40
    pears   Jan     56
            Feb     89
            Mar     98
            Apr    154
            May      4
            Jun     70
dtype: int64
```

[14]:
```
d2_stacked.unstack()
```

[14]:

		Jan	Feb	Mar	Apr	May	Jun
MA	apples	100	78	34	102	1	10
	pears	87	43	67	98	2	20
VT	apple	45	78	54	105	3	40
	pears	56	89	98	154	4	70

3 Pivot

[15]:
```
df5 = {
    'state':['MA','MA','VT','VT'],
    'location':['bolton','berlin','boyleston','berlin'],
    'apples': [3, 2, 0, 1],
    'pears': [0, 3, 7, 2]
}
d5 = pd.DataFrame(df5)
d5
```

[15]:

	state	location	apples	pears
0	MA	bolton	3	0

```
1    MA     berlin       2      3
2    VT  boyleston       0      7
3    VT     berlin       1      2
```

[16]:
```
d5.pivot(index='state', columns = 'location')
```

[16]:
```
          apples                   pears
location berlin bolton boyleston berlin bolton boyleston
state
MA          2.0    3.0       NaN    3.0    0.0       NaN
VT          1.0    NaN       0.0    2.0    NaN       7.0
```

[17]:
```
d5.pivot(index='state', columns = 'location', values = 'apples')
```

[17]:
```
location  berlin  bolton  boyleston
state
MA           2.0     3.0        NaN
VT           1.0     NaN        0.0
```

4 Exercise - 10 minutes

[18]:
```
places.head()
```

[18]:
```
  StateAbbr StateDesc CountyName  CountyFIPS  TotalPopulation  harthritis  \
0        GA   Georgia    Appling       13001            18507        32.2
1        GA   Georgia   Atkinson       13003             8297        29.0
2        GA   Georgia      Bacon       13005            11185        30.5
3        GA   Georgia      Baker       13007             3092        32.6
4        GA   Georgia    Baldwin       13009            44823        27.2

   hasthma  hbphigh  hcancer  hhighchol  hkidney  hcopd  hchd  hdiabetes  \
0     10.5     39.0      7.4       37.8      4.1   11.8  10.0       16.3
1     10.7     38.3      6.5       37.3      4.1   11.8   9.8       16.8
2     10.4     37.1      7.1       36.2      3.8   11.3   9.3       15.2
3     10.3     44.2      8.0       38.3      4.3   10.4   9.7       18.2
4     10.5     38.4      6.3       34.8      3.6    9.1   8.1       14.9

   hmhlth  hphlth  hteethlost  hstroke
0    16.7    18.2        24.6      5.1
1    17.7    18.8        28.1      5.1
2    16.8    17.6        22.9      4.8
3    14.3    16.8        21.3      5.6
4    15.9    15.0        20.8      4.5
```

[19]:
```
# The places dataset has a separate column for each healthy outcome measure.
# Melt the table so that they are all in one column with their associated␣
 ↪values in a separate column.
```

```
#d1.melt(id_vars=['fruit'], var_name = 'Month', value_name = 'Picked').
 ↪sort_values(by = ['fruit', 'Month'])

places_long = places.
 ↪melt(id_vars=['StateAbbr','StateDesc','CountyName','CountyFIPS','TotalPopulation'],
                          var_name = 'Outcome', value_name='pct').
 ↪sort_values(by = 'CountyFIPS')
places_long.head(20)
```

[19]:

	StateAbbr	StateDesc	CountyName	CountyFIPS	TotalPopulation	Outcome	pct
0	GA	Georgia	Appling	13001	18507	harthritis	32.2
954	GA	Georgia	Appling	13001	18507	hcopd	11.8
1113	GA	Georgia	Appling	13001	18507	hchd	10.0
1590	GA	Georgia	Appling	13001	18507	hphlth	18.2
477	GA	Georgia	Appling	13001	18507	hcancer	7.4
1749	GA	Georgia	Appling	13001	18507	hteethlost	24.6
159	GA	Georgia	Appling	13001	18507	hasthma	10.5
1908	GA	Georgia	Appling	13001	18507	hstroke	5.1
795	GA	Georgia	Appling	13001	18507	hkidney	4.1
1272	GA	Georgia	Appling	13001	18507	hdiabetes	16.3
318	GA	Georgia	Appling	13001	18507	hbphigh	39.0
636	GA	Georgia	Appling	13001	18507	hhighchol	37.8
1431	GA	Georgia	Appling	13001	18507	hmhlth	16.7
1114	GA	Georgia	Atkinson	13003	8297	hchd	9.8
478	GA	Georgia	Atkinson	13003	8297	hcancer	6.5
1591	GA	Georgia	Atkinson	13003	8297	hphlth	
160	GA	Georgia	Atkinson	13003	8297	hasthma	
637	GA	Georgia	Atkinson	13003	8297	hhighchol	
1909	GA	Georgia	Atkinson	13003	8297	hstroke	
1	GA	Georgia	Atkinson	13003	8297	harthritis	

```
1591  18.8
160   10.7
637   37.3
1909   5.1
1     29.0
```

[20]:
```
places_long.shape
```

[20]:
```
(2067, 7)
```

[21]:
```
# This works but creates a potential problem.  What do you think it is?
```

[22]:
```
# Making sure TotalPopulation cannot be accidentally summarized
places_long = places_long.drop('TotalPopulation',axis=1)
```

[23]:
```
places_long.shape
```

[23]:
```
(2067, 6)
```