Wide and long data formats

RESHAPING DATA WITH PANDAS



Maria Eugenia Inzaugarat

Data Scientist



You will learn

- Wide and long formats
- Long to wide transformation
- Wide to long transformation
- Stacking and unstacking columns
- Reshaping and handling complex data, such as string columns or JSON data

Why it is important

- Tidy datasets
- Data is not in the appropriate format for analysis:
 - Human readable vs. statistical analysis
- Nested data in DataFrames is complex to handle
- Get summary statistics for multi-level index DataFrames

Shape of data

The way in which a dataset is organized in rows and columns

```
fifa_players = pd.read_csv("fifa_players.csv")
fifa_players
```

```
nationality
                                                     club
                        age
                name
        Lionel Messi
                                               Barcelona
                                Argentina
                         32
0
  Cristiano Ronaldo
                                 Portugal
                                                Juventus
                         34
     Neymar da Silva
                                   Brazil Saint-Germain
                         27
2
```

```
fifa_players.shape
```

(3, 4)

fifa_players

```
nationality
                                                    club
                        age
                name
       Lionel Messi
                                               Barcelona
                                Argentina
0
                         32
  Cristiano Ronaldo
                                 Portugal
                                                Juventus
                         34
    Neymar da Silva
                                   Brazil Saint-Germain
                         27
```



```
fifa_players
```

```
VV
                                 nationality
                                                        club
                          age
                 name
        Lionel Messi
                         | 32
                                  Argentina
                                                  Barcelona
0
   Cristiano Ronaldo
                         34
                                   Portugal
                                                   Juventus
                                     Brazil Saint-Germain
     Neymar da Silva
                         27
2
                           \wedge \wedge
```

Each feature is in a separate column

```
fifa_players
```

```
nationality
                                                    club
                name
                        age
        Lionel Messi
                                Argentina
                                               Barcelona <--
                         32
0
  Cristiano Ronaldo
                                 Portugal
                         34
                                                Juventus <--
     Neymar da Silva
                         27
                                   Brazil Saint-Germain <--
```

- Each feature is in a separate column
- Each rows contains many features of the same player

```
fifa_players
```

- Each feature is in a separate column
- Each rows contains many features of the same player
- No repetition but large number of missing values
- Simple statistics and imputation

fifa_players_long.head()

```
variable
                                    value
              name
O Cristiano Ronaldo nationality
                                  Portugal
1 Cristiano Ronaldo
                                  Juventus
                           club
      Lionel Messi
                                        32
                            age
      Lionel Messi nationality
                                 Argentina
3
      Lionel Messi
                           club
                                 Barcelona
```



```
fifa_players_long.head()
```

```
name variable value

O Cristiano Ronaldo nationality Portugal <--

1 Cristiano Ronaldo club Juventus

2 Lionel Messi age 32

3 Lionel Messi nationality Argentina <--

4 Lionel Messi club Barcelona
```

Each row represents one feature

```
fifa_players_long.head()
```

```
variable
                                     value
              name
   Cristiano Ronaldo
                      nationality
                                     Portugal <--
   Cristiano Ronaldo
                                     Juventus <--
                             club
      Lionel Messi
                                         32
2
                             age
      Lionel Messi nationality
3
                                 Argentina
      Lionel Messi
                                  Barcelona
                            club
```

- Each row represents one feature
- Multiple rows for each player

```
fifa_players_long.head()
```

```
variable
                                                         value
                       name
     Cristiano Ronaldo | nationality
                                                    Portugal
     Cristiano Ronaldo |
                                           club
                                                     Juventus
2 |
            Lionel Messi |
                                                             32
                                            age
3 |
            Lionel Messi | nationality
                                                   Argentina
            Lionel Messi |
                                                   Barcelona
                                           club
           \wedge \wedge
```

- Each row represents one feature
- Multiple rows for each player
- A column (name) to identify same player

```
fifa_players_long.head()
```

n	name variable	value
0 Cristiano Rona	aldo nationality	Portugal
1 Cristiano Rona	aldo club	Juventus
2 Lionel Me	essi age	32
3 Lionel Me	essi nationality	Argentina
4 Lionel Me	essi club	Barcelona

- Each row represents one feature
- Multiple rows for each player
- A column (name) to identify same player
- Tidy data:
 - Better to summarize data
 - Key-value pairs
 - Preferred for analysis and graphing

- Transforming a DataFrame or Series structure to adjust it for analysis
 - Transposing a DataFrame

```
fifa_players.set_index('club')
```

```
nationality
                                  age
                            name
        club
   Barcelona
                    Lionel Messi
                                         Argentina
                                  32
               Cristiano Ronaldo
                                 NaN
                                          Portugal
    Juventus
Saint-Germain
                                            Brazil
                 Neymar da Silva
                                   27
```

- Transforming a DataFrame or Series structure to adjust it for analysis
 - Transposing a DataFrame

```
fifa_players.set_index('club')[['name', 'nationality']]
```

```
name nationality
club
Barcelona Lionel Messi Argentina
Juventus Cristiano Ronaldo Portugal
Saint-Germain Neymar da Silva Brazil
```

- Transforming a DataFrame or Series structure to adjust it for analysis
 - Transposing a DataFrame

```
fifa_players.set_index('club')[['name', 'nationality']].transpose()
```

club	Barcelona	Juventus	Saint-Germain
name	Lionel Messi	Cristiano Ronaldo	Neymar da Silva
nationality	Argentina	Portugal	Brazil

- Converting data from wide to long format and vice versa
- Unit of analysis:
 - Long format -> characteristic of a player
 - Wide format -> each player

Wide to long transformation

- Performed using pandas functions, such as:
 - o .melt()
 - o .wide_to_long()

Long to wide format

- Transform data using pandas methods, for example:
 - o .pivot()
 - o .pivot_table()

Let's practice!

RESHAPING DATA WITH PANDAS



Reshaping using pivot method

RESHAPING DATA WITH PANDAS



Maria Eugenia Inzaugarat

Data Scientist



From long to wide

- Demonstrate relationship between two columns
- Time series operations with the variables
- Operation that requires columns to be the unique variable

¹ https://pandas.pydata.org/docs/user_guide/reshaping.html



From long to wide

	Name	Year	Weight
0	John	2013	80
1	Mary	2013	65
2	Mary	2014	68
3	John	2014	83
4	Laura	2014	71

	Name	Year	Weight
0	John	2013	80
1	Mary	2013	65
2	Mary	2014	68
3	John	2014	83
4	Laura	2014	71

Name	John	Mary	Laura
Year			
2013	80	65	NaN
2014	83	68	71

df.pivot(, ,

	Name	Year	Weight	
0	John	2013	80	
1	Mary	2013	65	
2	Mary	2014	68	
3	John	2014	83	
4	Laura	2014	71	

Name	John	Mary	Laura
Year			
2013	80	65	NaN
2014	83	68	71

	Name	Year	Weight
0	John	2013	80
1	Mary	2013	65
2	Mary	2014	68
3	John	2014	83
4	Laura	2014	71

	Name	Year	Weight				
0	John	2013	80	Name	John	Mary	
1	Mary	2013	65	Year			Ī
2	Mary	2014	68	2013	80	65	İ
3	John	2014	83	2014	83	68	
4	Laura	2014	71				

	Name	Year	Weight
0	John	2013	80
1	Mary	2013	65
2	Mary	2014	68
3	John	2014	83
4	Laura	2014	71

df.pivot(index="Year", columns="Name", values="Weight")

	Name	Year	Weight
0	John	2013	80
1	Mary	2013	65
2	Mary	2014	68
3	John	2014	83
4	Laura	2014	71

df.pivot(index="Year", columns="Name", values="Weight")

```
fifa = pd.read_csv('fifa_players.csv')
fifa.head()
```

```
variable
                                   metric_system
                                                   imperial_system
                 name
   Cristiano Ronaldo
                           weight
                                               83
                                                            183.00
             J. Oblak
                           weight
                                                            191.00
                                              87
   Cristiano Ronaldo
                           height
                                                              6.13
                                             187
      J. Oblak
                                                              6.16
3
                           height
                                             188
```

```
fifa.pivot(index='name'
)
```



```
fifa.pivot(index='name', columns='variable'
)
```



```
fifa.pivot(index='name', columns='variable', values='metric_system')
```

```
variable height weight
name
Cristiano Ronaldo 187 83
J. Oblak 188 87
```



Pivoting multiple columns

```
fifa.pivot(index='name', columns='variable', values=['metric_system', 'imperial_system'])
```

```
metric_system
                                      imperial_system
        variable
                   height weight
                                      height
                                             weight
            name
Cristiano Ronaldo
                                        6.13
                                                183.0
                      187
                               83
        J. Oblak
                                                191.0
                      188
                               87
                                        6.16
```



Pivoting multiple columns

	Name	Year	Weight	Age
0			80	30
0	John	2013	00	30
1	Mary	2013	65	28
2	Mary	2014	68	29
		2014	02	21
3	John	2014	83	31
4	Laura	2014	71	34

df.pivot(index="Year", columns="Name")

Age

Mary

28

29

Laura

NaN

34

John

30

31

Laura

NaN

71

Pivoting multiple columns

```
fifa.pivot(index="name", columns="variable")
```

	metric_system		imperial_system	
variable	height	weight	height	weight
name				
Cristiano Ronaldo	187	83	6.13	183.0
J. Oblak	188	87	6.16	191.0



Duplicate entries error

```
another_fifa.head()
```

	name	variable	metric_system	imperial_system
0	Cristiano Ronaldo	weight	83	183.00
1	J. Oblak	weight	87	191.00
2	Cristiano Ronaldo	height	187	6.13
3	J. Oblak	height	188	6.16
4	Cristiano Ronaldo	height	187	6.14



Duplicate entries error

```
another_fifa.head()
```

```
variable
                                   metric_system
                                                  imperial_system
                 name
   Cristiano Ronaldo
                           weight
                                                            183.00
                                               83
                                                            191.00
             J. Oblak
                           weight
                                               87
      Cristiano Ronaldo
                             height
                                               187
                                                                6.13 <--
             J. Oblak
                           height
                                                              6.16
3
                                             188
      Cristiano Ronaldo
                             height
                                                                6.14 <--
                                               187
  4
```



Duplicate entries error

```
another_fifa.pivot(index="name", columns="variable")
```

ValueError: Index contains duplicate entries, cannot reshape

```
another_fifa = another_fifa.drop(4, axis=0)
another_fifa.pivot(index="name", columns="variable")
```

	metric	_system	imperial	_system
variable	height	weight	height	weight
name				
Cristiano Ronaldo	187	83	6.13	183.0
J. Oblak	188	87	6.16	191.0



Let's practice!

RESHAPING DATA WITH PANDAS



RESHAPING DATA WITH PANDAS



Maria Eugenia Inzaugarat
Data Scientist



Pivot method limitations

```
another_fifa.head()
```

```
variable metric_system
                                              imperial_system
             name
Cristiano Ronaldo
                       weight
                                          83
                                                       183.00
         J. Oblak
                       weight
                                          87
                                                       191.00
Cristiano Ronaldo
                       height
                                         187
                                                         6.13
 J. Oblak
                                                         6.16
                       height
                                         188
                       height
Cristiano Ronaldo
                                                         6.14
                                         187
```

```
another_fifa.pivot(index="name", columns="variable")
```

```
Traceback (most recent call last):
```

ValueError: Index contains duplicate entries, cannot reshape



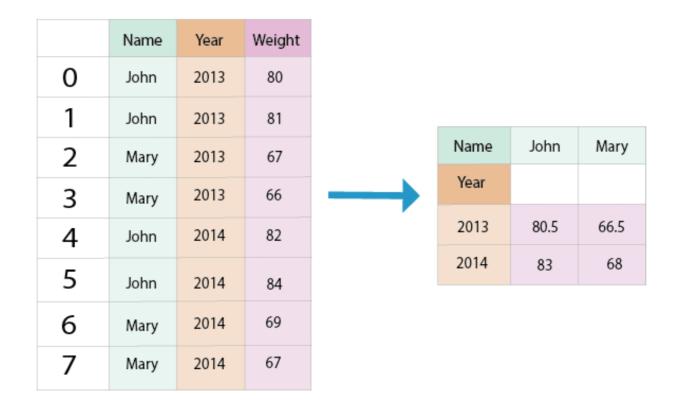
Pivot method limitations

- General purpose pivoting
- Index/column pair must be unique
- Cannot aggregate values

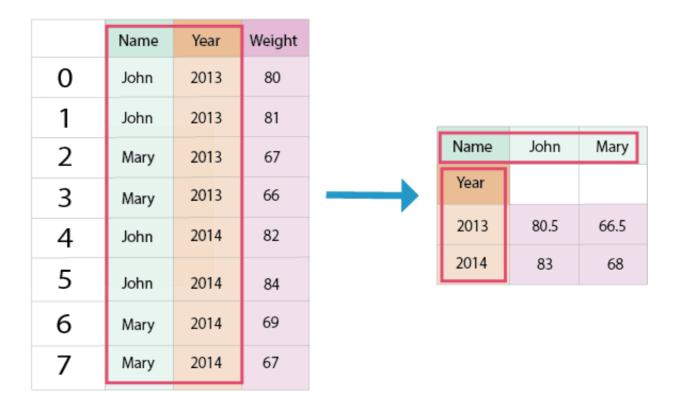
• A DataFrame containing statistics that summarizes the data of a larger DataFrame

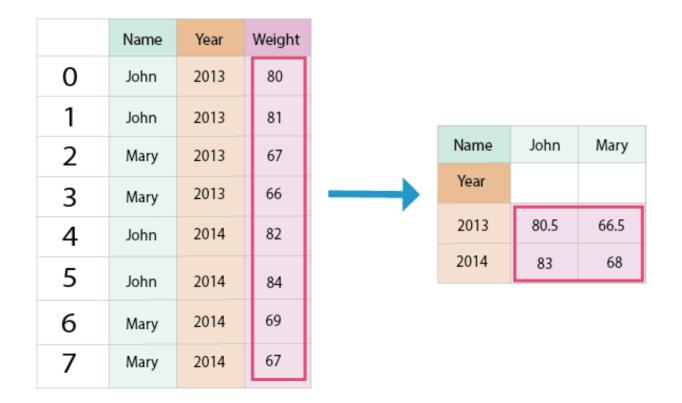
Name	John	Mary
Year		
2013	80.5	66.5
2014	83	68

	Name	Year	Weight
0	John	2013	80
1	John	2013	81
2	Mary	2013	67
3	Mary	2013	66
4	John	2014	82
5	John	2014	84
6	Mary	2014	69
7	Mary	2014	67



df.pivot_table(, , , , ,





df.pivot_table(index="Year", columns="Name", values="Weight", aggfunc="mean")

```
another_fifa.pivot_table(index="name", columns="variable", aggfunc="mean")
```

	metric	_system	imperial	_system
variable	height	•	height	weight
name				
Cristiano Ronaldo	187	83	6.135	183.0
J. Oblak	188	87	6.160	191.0



```
fifa_players.head(6)
```

first	last	movement	overall	attacking
0 Lionel	Messi	shooting	92	70
1 Cristiano	Ronaldo	shooting	93	89
2 Lionel	Messi	passing	92	92
3 Cristiano	Ronaldo	passing	82	83
4 Lionel	Messi	passing	96	88
5 Cristiano	Ronaldo	passing	89	84

```
fifa_players.head(6)
```

```
overall attacking
      first
                last
                       movement
                       shooting
     Lionel
               Messi
                                       92
                                                  70
1 Cristiano Ronaldo
                       shooting
                                       93
                                                  89
     Lionel
                        passing
               Messi
                                       92
                                                  92
3 Cristiano Ronaldo
                                       82
                                                  83
                        passing
                        passing
     Lionel
               Messi
                                       96
                                                  88
5 Cristiano Ronaldo
                        passing
                                       89
                                                  84
```

```
fifa_players.pivot_table(index= , columns="movement", values= , aggfunc= )
```

```
fifa_players.head(6)
```

```
overall attacking
      first
                last
                       movement
                       shooting
     Lionel
               Messi
                                       92
                                                  70
1 Cristiano Ronaldo
                       shooting
                                       93
                                                  89
                        passing
     Lionel
               Messi
                                       92
                                                  92
3 Cristiano Ronaldo
                                       82
                        passing
                                                  83
                        passing
    Lionel
               Messi
                                                  88
                                       96
5 Cristiano Ronaldo
                        passing
                                       89
                                                  84
```



```
fifa_players.head(6)
```

```
overall attacking
      first
               last
                       movement
     Lionel
              Messi
                       shooting
                                                 70
                                      93
1 Cristiano Ronaldo
                       shooting
                                                 89
     Lionel
              Messi
                       passing
                                      92
                                                 92
3 Cristiano Ronaldo
                       passing
                                                 83
     Lionel
              Messi
                       passing
                                      96
                                                 88
5 Cristiano Ronaldo
                                                 84
                       passing
                                       89
```

```
fifa_players.pivot_table(index=["first", "last"], columns="movement", values=["overall", "attacking"], aggfunc="max")
```

```
attacking overall
movement passing shooting passing shooting
first last
Cristiano Ronaldo 84 89 89 93
Lionel Messi 92 70 96 92
```



Margins

```
fifa_players.pivot_table(index=["first", "last"], columns="movement", aggfunc="count", )
```



Margins

```
fifa_players.pivot_table(index=["first", "last"], columns="movement", aggfunc="count", margins=True)
```

			attac	king		ove	rall
	movement	passing	shooting	All	passing	shooting	All
First	Last						
Cristiano	Ronaldo	2	1	3	2	1	3
Lionel	Messi	2	1	3	2	1	3
All		4	2	6	4	2	6

Pivot or pivot table?

Does the DataFrame have more than one value for each index/column pair?

Do you need to have a multi-index in your resulting pivoted DataFrame?

Do you need summary statistics of your large DataFrame?

Yes! Use .pivot_table()



Let's practice!

RESHAPING DATA WITH PANDAS

