Reshaping Data with Python

Pratheerth Padman
FREELANCE DATA SCIENTIST

Module Overview

Introduction to data reshaping

Reshape long data to wide data with the pivot function

Reshape wide data to long using the melt function

Reshaping data using the stack and unstack functions

Reshaping and aggregation with the pivot table function

Introduction to Data Reshaping

Data Shapes

Data shape refers to the way the data is arranged in rows and columns

ID	YEAR	HOUSE	COST
1	2015	House A	25000
2	2015	House B	41000
3	2016	House A	35000
4	2016	House B	37000
5	2017	House A	45000

ID	YEAR	HOUSE A	HOUSE B
1	2015	25000	41000
2	2016	35000	37000
3	2017	45000	NaN

Arrangement 1

Arrangement 2

Which Shape Is Better?

Analytical purpose of the data

Intended audience

Primary measures of interest

Data being compared / analyzed

Wide vs Long

ID	YEAR	CAR	COST
1	2015	House A	25000
2	2015	House B	41000
3	2016	House A	35000
4	2016	House B	37000
5	2017	House A	45000

ID	YEAR	HOUSE A	HOUSE B
1	2015	25000	41000
2	2016	35000	37000
3	2017	45000	NaN

Long (data analytics)

Wide (presentation)
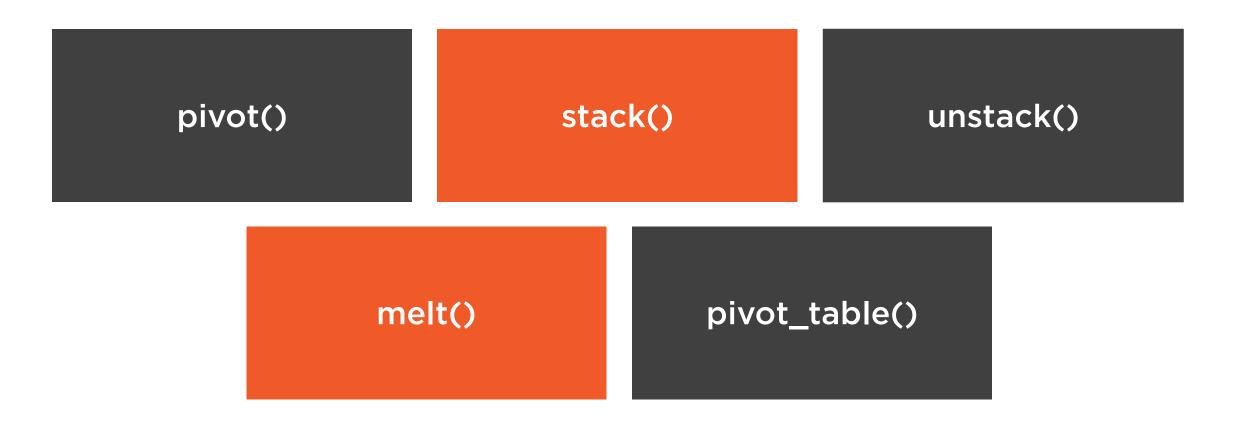
Data Reshaping

Technique of transforming data from one shape or representation to another based on your requirements

Reshape Long Data to Wide Data with the Pivot Function

Pivot Function

ID	YEAR	CAR	COST
1	2015	House A	25000
2	2015	House B	41000
3	2016	House A	35000
4	2016	House B	37000
5	2017	House A	45000

ID	YEAR	HOUSE A	HOUSE B
1	2015	25000	41000
2	2016	35000	37000
3	2017	45000	NaN

index – labels for the new frame's index

columns – labels for the new frame's columns

values – for populating the new frame's values

# Reshape Wide Data to Long Using the Melt Function
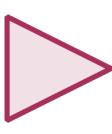
Melt Function
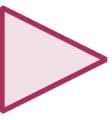
ID	YEAR	HOUSE A	HOUSE B
1	2015	25000	41000
2	2016	35000	37000
3	2017	45000	NaN
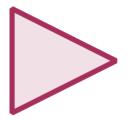
ID	YEAR	CAR	COST
1	2015	House A	25000
2	2015	House B	41000
3	2016	House A	35000
4	2016	House B	37000
5	2017	House A	45000

▶ frame – the dataframe to be reshaped

▶ Id_vars – columns to use as identifier variables

▶ value_vars – columns to unpivot

Reshaping Data Using the Stack and Unstack Functions

Stack and Unstack Function

Stack – to reshape from wide to long

**Unstack – to reshape from long to wide**

Used on a multi-indexed dataframe

Level parameter specifies the level at which stacking/unstacking occurs

Stack and Unstack

	Before Upskilling		After Upskilling	
	Age	**Salary**	**Age**	**Salary**
Marcus	25	45000	27	56000
Anthony	38	89000	41	102000
Mason	30	67000	31	70000
Daniel	45	130000	48	160000

Stack →

		After Upskilling	**Before Upskilling**
Marcus	Age	27	25
	Salary	56000	45000
Anthony	Age	41	38
	Salary	102000	89000
Mason	Age	31	30
	Salary	70000	67000
Daniel	Age	48	45
	Salary	160000	130000

Unstack →

	Before Upskilling		After Upskilling	
	Age	**Salary**	**Age**	**Salary**
Marcus	25	45000	27	56000
Anthony	38	89000	41	102000
Mason	30	67000	31	70000
Daniel	45	130000	48	160000

Reshaping and Aggregation with Pivot_Table

Pivot_Table

Pivot table function aggregates along with reshaping

Works only with numeric values, non-numeric values will be dropped

Mean is the default aggregation function, can be changed to others

Summary

5 techniques to reshape data – pivot, melt, stack, unstack and pivot_table

Long to wide:
- Pivot – single index
- Unstack – multi index

Wide to long:
- Melt – single index
- Stack – multi index

Pivot table – aggregation and reshaping