Pandas groupby

SHARP SIGHT

What you'll learn

- How to use the Pandas `groupby()` method

- How to "group" your data
  - how to analyze your data by a categorical variable (i.e., a grouping variable)
  - used in the "split, apply, combine" strategy of analysis

Pandas groupby overview

Groupby enables aggregation by a categorical variable

- The `groupby()` method enables you to "group" your data by a categorical variable
  - the "split, apply, combine" strategy

- The output of `groupby()` is a `DataFrameGroupBy` object

- Commonly used in conjunction with summary statistics
  - e.g., mean, median, sum, etc

- Enables you to calculate values by a categorical variable
  - e.g., mean by decade

Example: you have a dataframe

model	make	year	horsepower	weight
911 Turbo	Porsche	2013	513	1.67
Z4 3.0i	BMW	2007	215	1.36
Cayman	Porsche	2007	241	1.3
Veyron	Bugatti	2012	1184	1.84
550i	BMW	2013	444	1.9

Note: this is an abbreviated version of the `supercars` dataset that you'll work with in the code walkthroughs

You want to count the number of cars by "make"

model	make	year	horsepower	weight
911 Turbo	Porsche	2013	513	1.67
Z4 3.0i	BMW	2007	215	1.36
Cayman	Porsche	2007	241	1.3
Veyron	Bugatti	2012	1184	1.84
550i	BMW	2013	444	1.9

You can use groupby() with count()

```
supercars_group_make = supercars.groupby('make')
supercars_group_make.model.count()
```

model	make	year	horsepower	weight
911 Turbo	Porsche	2013	513	1.67
Z4 3.0i	BMW	2007	215	1.36
Cayman	Porsche	2007	241	1.3
Veyron	Bugatti	2012	1184	1.84
550i	BMW	2013	444	1.9

make	count
Porsche	2
Bugatti	1
BMW	2

You can use groupby() with count()

```
supercars_group_make = supercars.groupby('make')
supercars_group_make.model.count()
```

Here, we're using groupby to specify `make` as a grouping variable

And we're using `count()` to count the number of records by `make`

make	count
Porsche	2
Bugatti	1
BMW	2

Pandas groupby syntax

Syntax: pandas groupby

The name of the DataFrame you want to operate on

The name of the method (called with "dot" syntax)

```
myDataFrame.groupby(by-variable)
```

The name of the variable you want to group your data by

Parameters of Pandas groupby

The parameters of Pandas groupby

Parameter	What it does	Format
`by-variable`	Specify the variable on which to group the data	This can be a variable or list of variables. It can also be a function or mapping, but this is extremely uncommon.

Note: `groupby()` has several other parameters.
These are rarely used, so we will not discuss them.

The output of Pandas groupby

- When you call `groupby()` on a DataFrame, the output is a DataFrameGroupBy object
  - Typically, you will need to use this DataFrameGroupBy in conjunction with another function for it to be usable

- When you call `groupby()` on a Series, the output is a SeriesGroupBy object

Recap

Recap of what we learned

- You can use `groupby()` to group your data
  - i.e., group data, and then calculate statistics by a categorical variable

- Note: `groupby()` is very important for data analysis!
  - learn how to use it well

- Note: `groupby()` is a little abstract
  - best to learn it with concrete examples

- **Next Steps**: Review the code walkthrough videos for clear, step-by-step examples of `groupby()`