

Data Aggregation: Takeaways

by Dataquest Labs, Inc. - All rights reserved © 2024

Syntax

GROUPBY OBJECTS

- Create a GroupBy object:

```
df.groupby('col_to_groupby')
```

- Select one column from a GroupBy object:

```
df.groupby('col_to_groupby')['col_selected']
```

- Select multiple columns from a GroupBy object:

```
df.groupby('col_to_groupby')[['col_selected1', 'col_selected2']]
```

COMMON AGGREGATION METHODS

- `mean()` : calculates the mean of groups
- `sum()` : calculates the sum of group values
- `size()` : calculates the size of groups
- `count()` : calculates the count of values in groups
- `min()` : calculates the minimum of group values
- `max()` : calculates the maximum of group values

GROUPBY.AGG() METHOD

- Apply one function to a GroupBy object:

```
df.groupby('col_to_groupby').agg(function_name)
```

- Apply multiple functions to a GroupBy object:

```
df.groupby('col_to_groupby').agg([function_name1, function_name2, function_name3])
```

- Apply a custom function to a GroupBy object:

```
df.groupby('col_to_groupby').agg(custom_function)
```

AGGREGATION WITH THE DATAFRAME.PIVOT_TABLE METHOD

- Apply only one function:

```
df.pivot_table(values='Col_to_aggregate', index='Col_to_group_by', aggfunc=function_name)
```

- Apply multiple functions:

```
df.pivot_table('Col_to_aggregate', 'Col_to_group_by', aggfunc=[function_name1, function_name2, function_name3])
```

- Aggregate multiple columns:

```
df.pivot_table(['Col_to_aggregate1', 'Col_to_aggregate2'], 'Col_to_group_by', aggfunc =  
function_name)
```

- Calculate the grand total for the aggregation column:

```
df.pivot_table('Col_to_aggregate', 'Col_to_group_by', aggfunc=function_name, margins=True)
```

Concepts

- **Aggregation** is applying a statistical operation to groups of data. It reduces dimensionality so that the DataFrame returned will contain just one value for each group. We can break down the aggregation process into three steps:
 - split the dataframe into groups
 - apply a function to each group
 - combine the results into one data structure
- The `groupby` operation optimizes the split-apply-combine process. We can break it down into two steps:
 - create a GroupBy object
 - call an aggregation function
- Creating the GroupBy object is an intermediate step that allows us to optimize our work. It contains information on how to group the DataFrame, but nothing actually gets computed until a function is called.
- We can also use the `DataFrame.pivot_table()` method to aggregate data; we can also use it to calculate the grand total for the aggregation column.