DS3500
Aggregation and Windowing

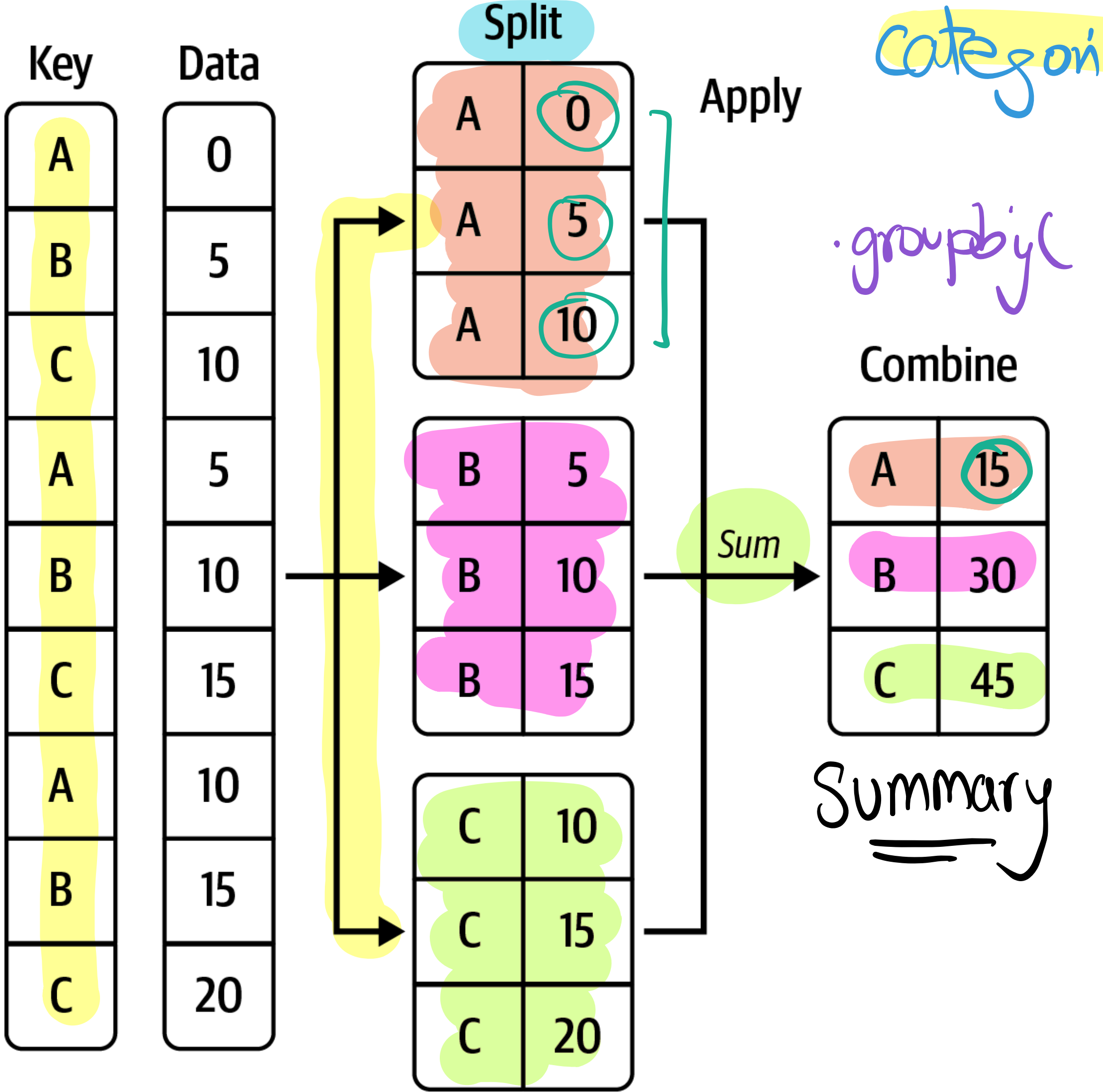
Rush Sanghrajka

Aggregation by Grouping

categorical variables

transform → original data set

.groupby( )



Key	Data
A	0
B	5
C	10
A	5
B	10
C	15
A	10
B	15
C	20

A	0
A	5
A	10

B	5
B	10
B	15

C	10
C	15
C	20

A	15
B	30
C	45

Summary

sum( )

mean( )

median( )

quantiles( ) Q1, Q3

min( )

max( )

first/last( )

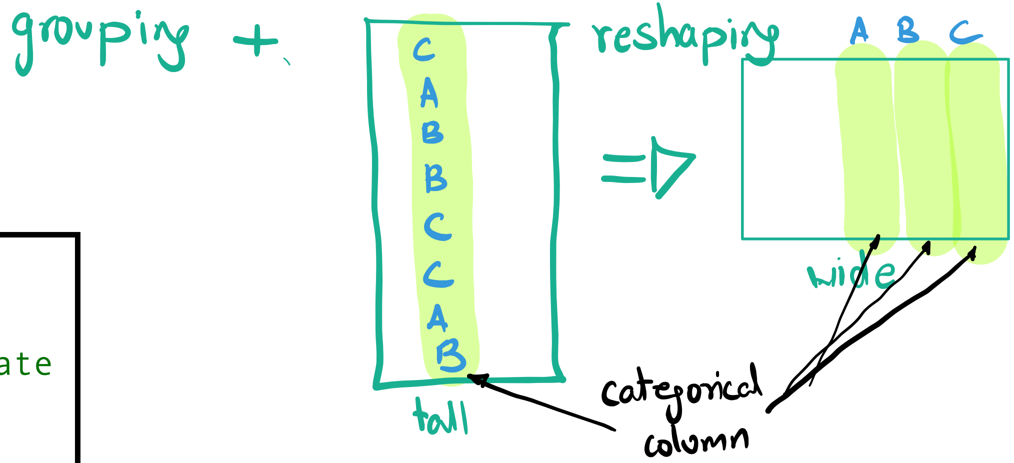
count( )

Pivot Tables*

(wrapper)

```
df.pivot_table(
    values='value',              # What to aggregate
    index='date',               # Rows
    columns='neighborhood',     # Columns
    aggfunc='mean'              # How to aggregate
(default: mean)
)


Special options:
  margins=True                  # Add row/column
totals
  fill_value=0                 # Replace NaN with
value
  aggfunc=['mean', 'max']      # Multiple functions
```



Uses grouping and reshaping to essentially convert a tall table to a wide one

Useful for 2+ categorical variables

Can reshape a categorical variable in one column into multiple columns

What drinks sell best?

Order_ID	Day	Time_Period	Drink_Type	Size	Price
1	Monday	Morning	Latte	Medium	4.50
2	Monday	Morning	Espresso	Small	3.00
3	Monday	Afternoon	Latte	Large	5.50
4	Tuesday	Morning	Latte	Medium	4.50
5	Tuesday	Morning	Cappuccino	Large	5.00
6	Tuesday	Afternoon	Espresso	Small	3.00
7	Wednesday	Morning	Latte	Medium	4.50
8	Wednesday	Afternoon	Cappuccino	Medium	4.50
9	Wednesday	Afternoon	Latte	Large	5.50

groupby(drink type) [price].sum() .count

Drink type	Total Sales	# sold
Latte	70	5
Espresso	60	2
Capp...	50	2
Coffee	65	0

$7	10
$4	15
$5	10

pd.pivot_table(df,)

Drink Type	Morning	Afternoon	Total
Latte	3	2	5
Espresso	1	1	2
Cappuccino	1	1	2
Total	5	4	

What drinks sell best during mornings vs afternoon?

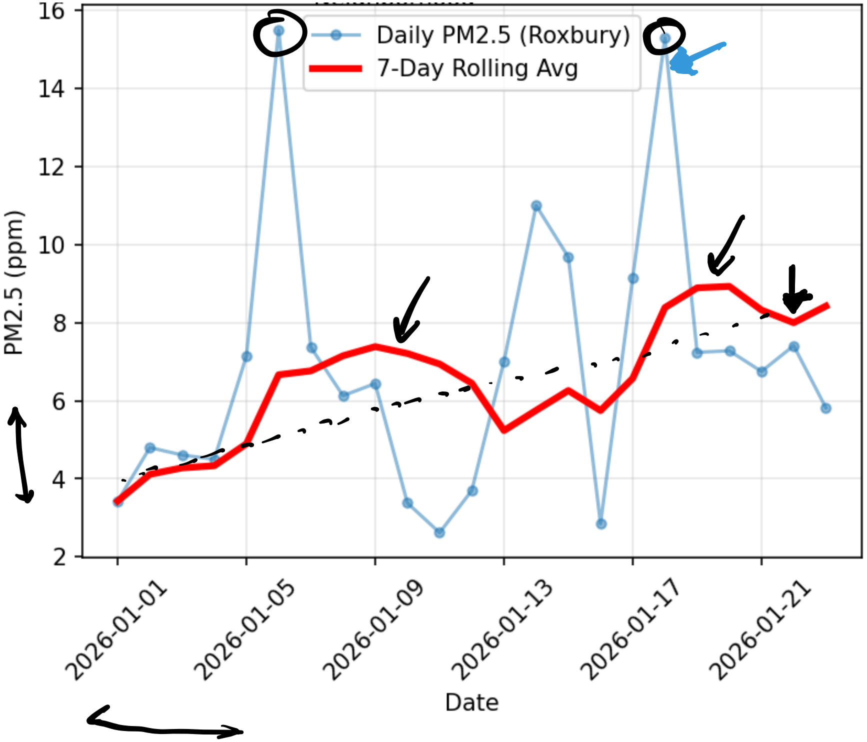
Order_ID	Day	Time_Period	Drink_Type	Size	Price
1	Monday	Morning 1	Latte	Medium	4.50
2	Monday	Morning	Espresso	Small	3.00
3	Monday	Afternoon 1	Latte	Large	5.50
4	Tuesday	Morning 2	Latte	Medium	4.50
5	Tuesday	Morning	Cappuccino	Large	5.00
6	Tuesday	Afternoon	Espresso	Small	3.00
7	Wednesday	Morning 3	Latte	Medium	4.50
8	Wednesday	Afternoon	Cappuccino	Medium	4.50
9	Wednesday	Afternoon 2	Latte	Large	5.50

What is each student's average score across all homeworks?

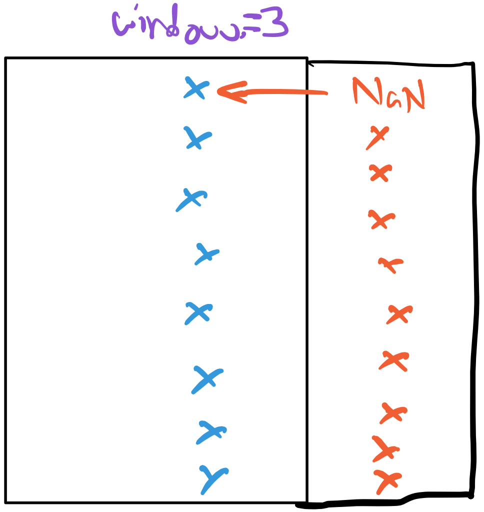
Group by student name, aggregate score with function mean

Student_Name	Assignment	Week	Status	Score	Hours_Spent
Alice	HW1	1	On-time	95	4
Alice	HW2	2	On-time	88	5
Alice	HW3	3	Late	92	6
Bob	HW1	1	On-time	78	3
Bob	HW2	2	Late	82	4
Bob	HW3	3	On-time	85	4
Carol	HW1	1	On-time	92	5
Carol	HW2	2	On-time	95	4
Carol	HW3	3	On-time	90	5

Windowing



suppress noise



```
df['col'].rolling(window=7).mean()
# 7-day average

df['col'].rolling(window=7,
min_periods=1).mean() # Include partial windows

df['col'].rolling(window=7).max()
# 7-day maximum

# use min_periods if you want to have means at
the edges
```