# Pandas 4

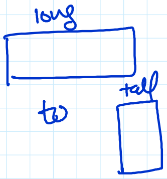
06 November 2024 20:56

Class starts at 9:03 pm

Colab Notes (Pandas-3) will also be shared

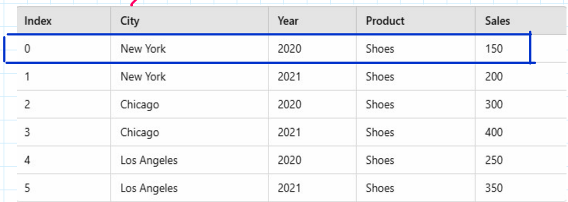
Agenda (Pandas coding) → ① (Multi indexing)

② Pd.melt()



③ Pivot Tables

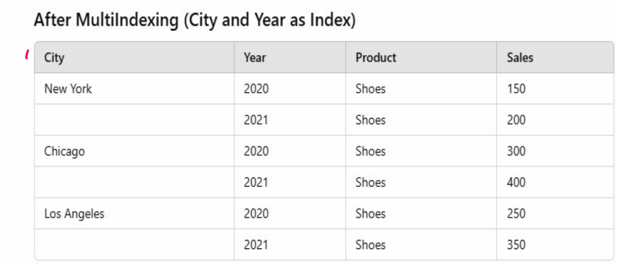
④ Binning



| Index | City | Year | Product | Sales |
|---|---|---|---|---|
| 0 | New York | 2020 | Shoes | 150 |
| 1 | New York | 2021 | Shoes | 200 |
| 2 | Chicago | 2020 | Shoes | 300 |
| 3 | Chicago | 2021 | Shoes | 400 |
| 4 | Los Angeles | 2020 | Shoes | 250 |
| 5 | Los Angeles | 2021 | Shoes | 350 |

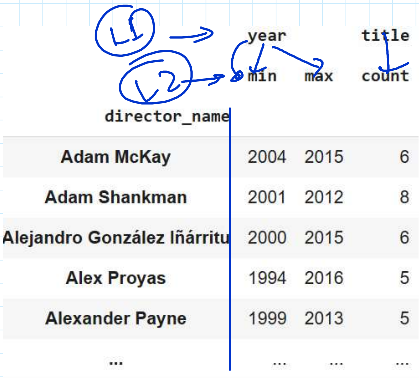
df.iloc[0]

indexing → Fetch the data fast (organise)



After MultiIndexing (City and Year as Index)

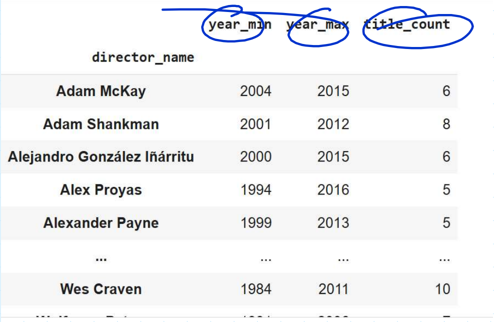
| City | Year | Product | Sales |
|---|---|---|---|
| New York | 2020 | Shoes | 150 |
|  | 2021 | Shoes | 200 |
| Chicago | 2020 | Shoes | 300 |
|  | 2021 | Shoes | 400 |
| Los Angeles | 2020 | Shoes | 250 |
|  | 2021 | Shoes | 350 |



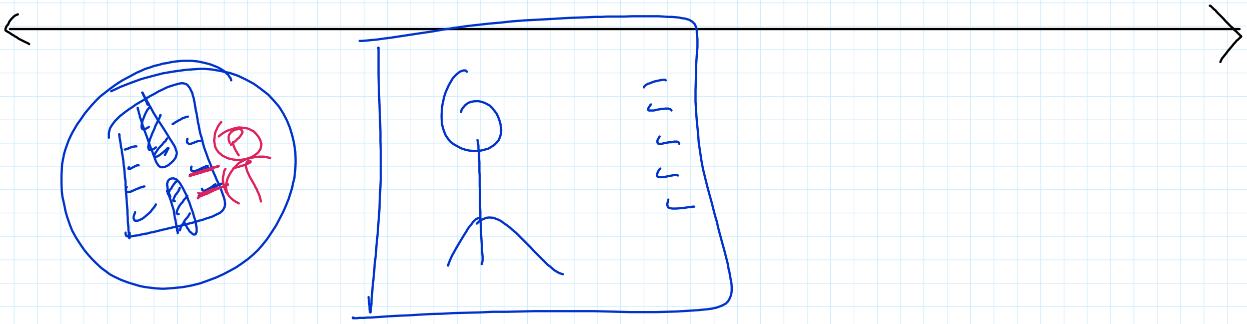
L1 →

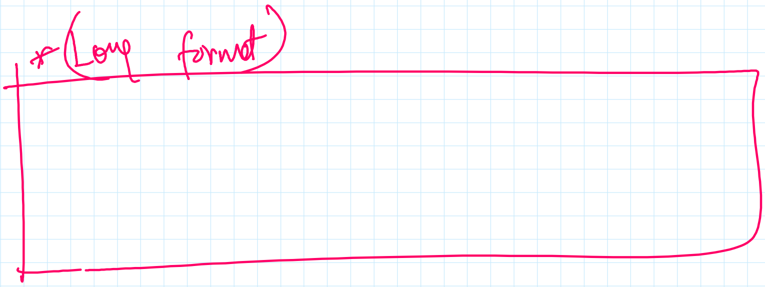
L2 →

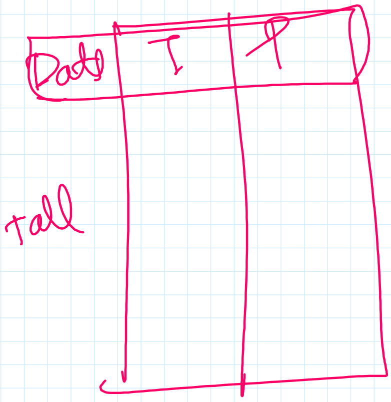
|  | year | | title |
|---|---|---|---|
|  | min | max | count |
| director_name | | | |
| Adam McKay | 2004 | 2015 | 6 |
| Adam Shankman | 2001 | 2012 | 8 |
| Alejandro González Iñárritu | 2000 | 2015 | 6 |
| Alex Proyas | 1994 | 2016 | 5 |
| Alexander Payne | 1999 | 2013 | 5 |
| ... | ... | ... | ... |

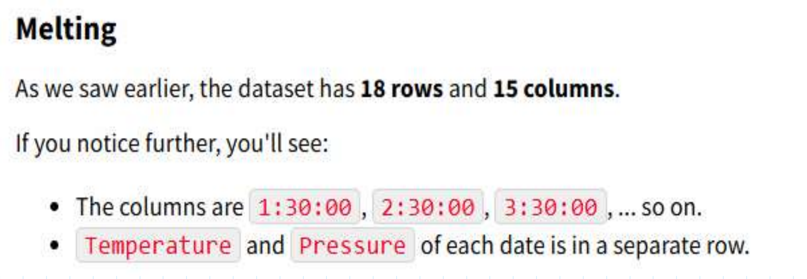


|  | year_min | year_max | title_count |
|---|---|---|---|
| director_name | | | |
| Adam McKay | 2004 | 2015 | 6 |
| Adam Shankman | 2001 | 2012 | 8 |
| Alejandro González Iñárritu | 2000 | 2015 | 6 |
| Alex Proyas | 1994 | 2016 | 5 |
| Alexander Payne | 1999 | 2013 | 5 |
| ... | ... | ... | ... |
| Wes Craven | 1984 | 2011 | 10 |









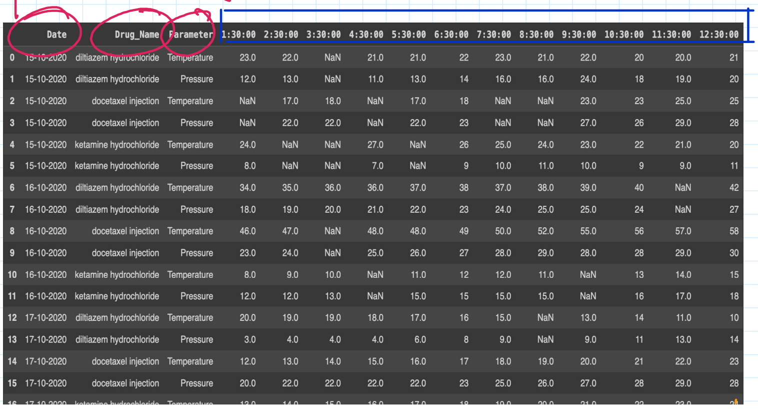
## Melting

As we saw earlier, the dataset has **18 rows** and **15 columns**.

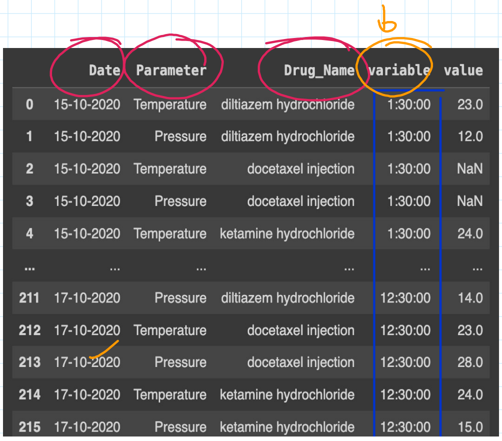
If you notice further, you'll see:

- The columns are `1:30:00`, `2:30:00`, `3:30:00`, ... so on.
- `Temperature` and `Pressure` of each date is in a separate row.

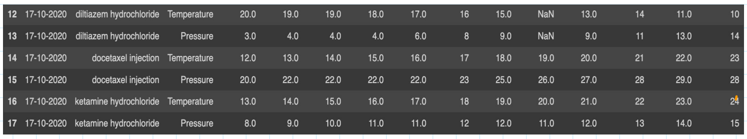
id_vars



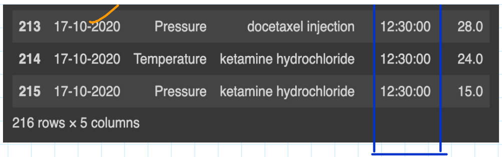
| | Date | Drug_Name | Parameter | 1:30:00 | 2:30:00 | 3:30:00 | 4:30:00 | 5:30:00 | 6:30:00 | 7:30:00 | 8:30:00 | 9:30:00 | 10:30:00 | 11:30:00 | 12:30:00 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 15-10-2020 | diltiazem hydrochloride | Temperature | 23.0 | 22.0 | NaN | 21.0 | 21.0 | 22 | 23.0 | 21.0 | 22.0 | 20 | 20.0 | 21 |
| 1 | 15-10-2020 | diltiazem hydrochloride | Pressure | 12.0 | 13.0 | NaN | 11.0 | 13.0 | 14 | 16.0 | 16.0 | 24.0 | 18 | 19.0 | 20 |
| 2 | 15-10-2020 | docetaxel injection | Temperature | NaN | 17.0 | 18.0 | NaN | 17.0 | 18 | NaN | NaN | 23.0 | 23 | 25.0 | 25 |
| 3 | 15-10-2020 | docetaxel injection | Pressure | NaN | 22.0 | 22.0 | NaN | 22.0 | 23 | NaN | NaN | 27.0 | 26 | 29.0 | 28 |
| 4 | 15-10-2020 | ketamine hydrochloride | Temperature | 24.0 | NaN | NaN | 27.0 | NaN | 26 | 25.0 | 24.0 | 23.0 | 22 | 21.0 | 20 |
| 5 | 15-10-2020 | ketamine hydrochloride | Pressure | 8.0 | NaN | NaN | 7.0 | NaN | 9 | 10.0 | 11.0 | 10.0 | 9 | 9.0 | 11 |
| 6 | 16-10-2020 | diltiazem hydrochloride | Temperature | 34.0 | 35.0 | 36.0 | 36.0 | 37.0 | 38 | 37.0 | 38.0 | 39.0 | 40 | NaN | 42 |
| 7 | 16-10-2020 | diltiazem hydrochloride | Pressure | 18.0 | 19.0 | 20.0 | 21.0 | 22.0 | 23 | 24.0 | 25.0 | 25.0 | 24 | NaN | 27 |
| 8 | 16-10-2020 | docetaxel injection | Temperature | 46.0 | 47.0 | NaN | 48.0 | 48.0 | 49 | 50.0 | 52.0 | 55.0 | 56 | 57.0 | 58 |
| 9 | 16-10-2020 | docetaxel injection | Pressure | 23.0 | 24.0 | NaN | 25.0 | 26.0 | 27 | 28.0 | 29.0 | 28.0 | 28 | 29.0 | 30 |
| 10 | 16-10-2020 | ketamine hydrochloride | Temperature | 8.0 | 9.0 | 10.0 | NaN | 11.0 | 12 | 12.0 | 11.0 | NaN | 13 | 14.0 | 15 |
| 11 | 16-10-2020 | ketamine hydrochloride | Pressure | 12.0 | 12.0 | 13.0 | NaN | 15.0 | 15 | 15.0 | 15.0 | NaN | 16 | 17.0 | 18 |
| 12 | 17-10-2020 | diltiazem hydrochloride | Temperature | 20.0 | 19.0 | 19.0 | 18.0 | 17.0 | 16 | 15.0 | NaN | 13.0 | 14 | 11.0 | 10 |
| 13 | 17-10-2020 | diltiazem hydrochloride | Pressure | 3.0 | 4.0 | 4.0 | 4.0 | 6.0 | 8 | 9.0 | NaN | 9.0 | 11 | 13.0 | 14 |
| 14 | 17-10-2020 | docetaxel injection | Temperature | 12.0 | 13.0 | 14.0 | 15.0 | 16.0 | 17 | 18.0 | 19.0 | 20.0 | 21 | 22.0 | 23 |
| 15 | 17-10-2020 | docetaxel injection | Pressure | 20.0 | 22.0 | 22.0 | 22.0 | 22.0 | 23 | 25.0 | 26.0 | 27.0 | 28 | 29.0 | 28 |
| 16 | 17-10-2020 | ketamine hydrochloride | Temperature | 13.0 | 14.0 | 15.0 | 16.0 | 17.0 | 18 | 19.0 | 20.0 | 21.0 | 22 | 23.0 | 24 |



| | Date | Parameter | Drug_Name | variable | value |
|---|---|---|---|---|---|
| 0 | 15-10-2020 | Temperature | diltiazem hydrochloride | 1:30:00 | 23.0 |
| 1 | 15-10-2020 | Pressure | diltiazem hydrochloride | 1:30:00 | 12.0 |
| 2 | 15-10-2020 | Temperature | docetaxel injection | 1:30:00 | NaN |
| 3 | 15-10-2020 | Pressure | docetaxel injection | 1:30:00 | NaN |
| 4 | 15-10-2020 | Temperature | ketamine hydrochloride | 1:30:00 | 24.0 |
| ... | ... | ... | ... | ... | ... |
| 211 | 17-10-2020 | Pressure | diltiazem hydrochloride | 12:30:00 | 14.0 |
| 212 | 17-10-2020 | Temperature | docetaxel injection | 12:30:00 | 23.0 |
| 213 | 17-10-2020 | Pressure | docetaxel injection | 12:30:00 | 28.0 |
| 214 | 17-10-2020 | Temperature | ketamine hydrochloride | 12:30:00 | 24.0 |
| 215 | 17-10-2020 | Pressure | ketamine hydrochloride | 12:30:00 | 15.0 |



| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 12 | 17-10-2020 | diltiazem hydrochloride | Temperature | 20.0 | 19.0 | 19.0 | 18.0 | 17.0 | 16 | 15.0 | NaN | 13.0 | 14 | 11.0 | 10 |
| 13 | 17-10-2020 | diltiazem hydrochloride | Pressure | 3.0 | 4.0 | 4.0 | 4.0 | 6.0 | 8 | 9.0 | NaN | 9.0 | 11 | 13.0 | 14 |
| 14 | 17-10-2020 | docetaxel injection | Temperature | 12.0 | 13.0 | 14.0 | 15.0 | 16.0 | 17 | 18.0 | 19.0 | 20.0 | 21 | 22.0 | 23 |
| 15 | 17-10-2020 | docetaxel injection | Pressure | 20.0 | 22.0 | 22.0 | 22.0 | 22.0 | 23 | 25.0 | 26.0 | 27.0 | 28 | 29.0 | 28 |
| 16 | 17-10-2020 | ketamine hydrochloride | Temperature | 13.0 | 14.0 | 15.0 | 16.0 | 17.0 | 18 | 19.0 | 20.0 | 21.0 | 22 | 23.0 | 24 |
| 17 | 17-10-2020 | ketamine hydrochloride | Pressure | 8.0 | 9.0 | 10.0 | 11.0 | 11.0 | 12 | 12.0 | 11.0 | 12.0 | 13 | 14.0 | 15 |

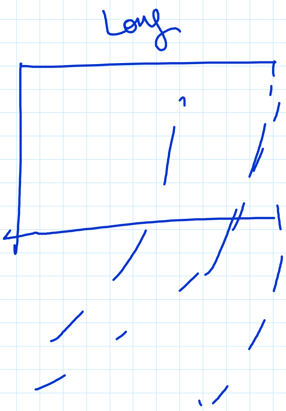


| | | | | | |
|---|---|---|---|---|---|
| 213 | 17-10-2020 | Pressure | docetaxel injection | 12:30:00 | 28.0 |
| 214 | 17-10-2020 | Temperature | ketamine hydrochloride | 12:30:00 | 24.0 |
| 215 | 17-10-2020 | Pressure | ketamine hydrochloride | 12:30:00 | 15.0 |

216 rows × 5 columns

## How to melt a dataframe ?

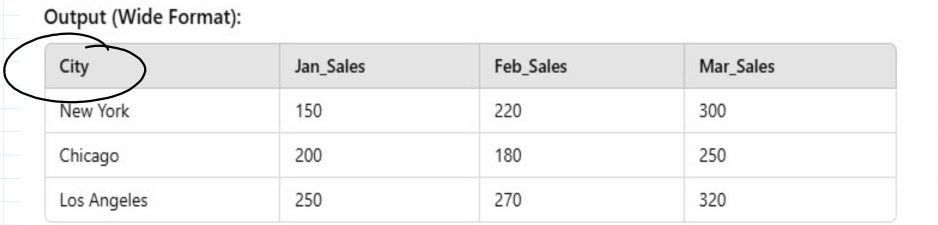
(No loss of data)
(Restructuring the existing data according to our needs)



Pd.melt( ) → id_vars = [ ] (list of variables / columns same in both df)

→ var_name = " "

→ value_name = " "

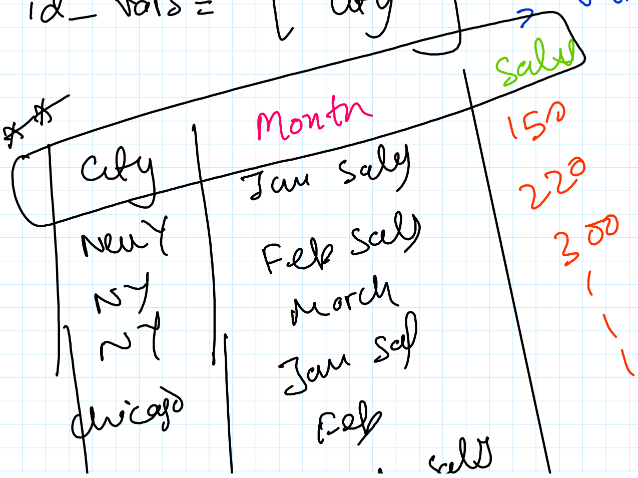


Output (Wide Format):

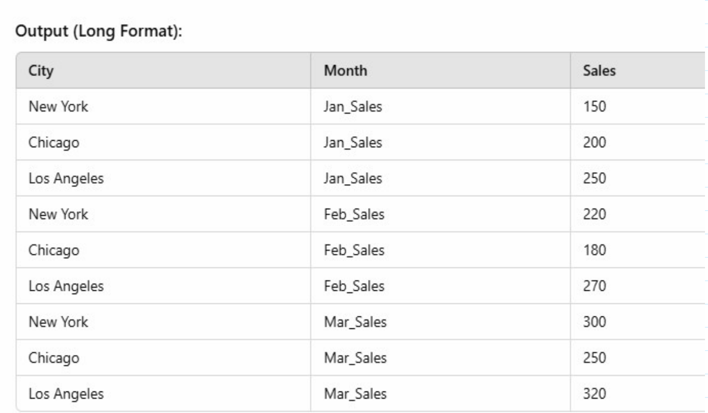
| City | Jan_Sales | Feb_Sales | Mar_Sales |
|---|---|---|---|
| New York | 150 | 220 | 300 |
| Chicago | 200 | 180 | 250 |
| Los Angeles | 250 | 270 | 320 |

(*)

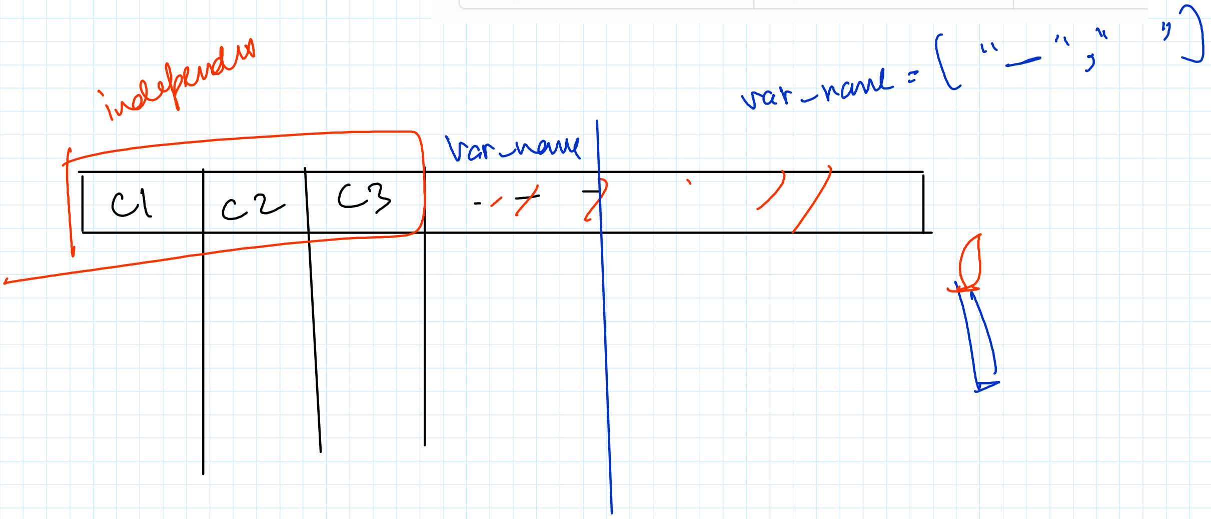
id_vars = ["city"] , var_name = "Month" , value_name = Sales



| City | Month | Sales |
|---|---|---|
| NewY | Jan Sales | 150 |
| NY | Feb Sales | 220 |
| NY | March | 300 |
| Chicago | Jan Sal | ) |
| | Feb | ) |
| | sales | |

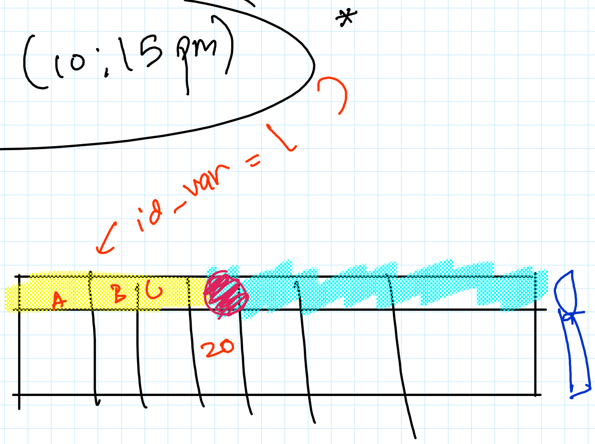

Output (Long Format):

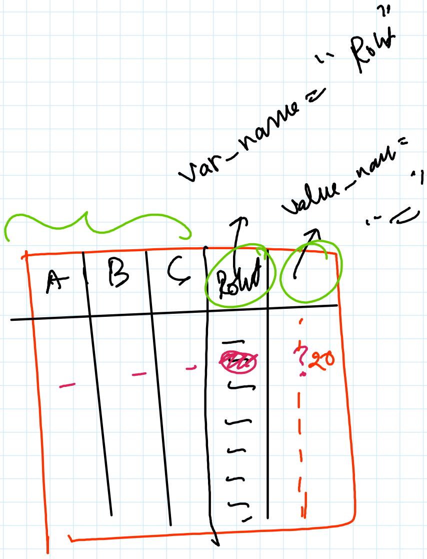
| City | Month | Sales |
|---|---|---|
| New York | Jan_Sales | 150 |
| Chicago | Jan_Sales | 200 |
| Los Angeles | Jan_Sales | 250 |
| New York | Feb_Sales | 220 |
| Chicago | Feb_Sales | 180 |
| Los Angeles | Feb_Sales | 270 |
| New York | Mar_Sales | 300 |
| Chicago | Mar_Sales | 250 |
| Los Angeles | Mar_Sales | 320 |



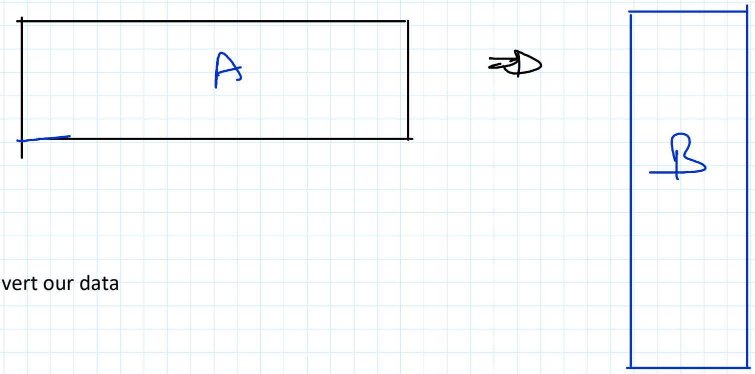
22:15 (10:15 pm)

pd.melt()

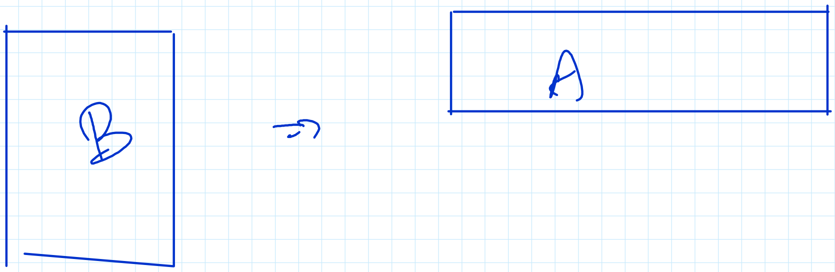




pd.melt( )



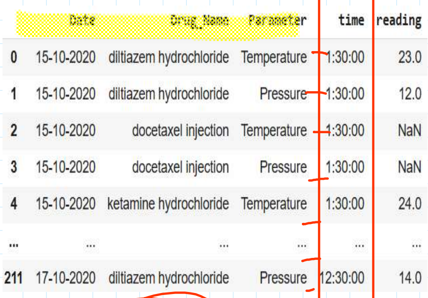
Now suppose we want to convert our data back to the **wide format**.



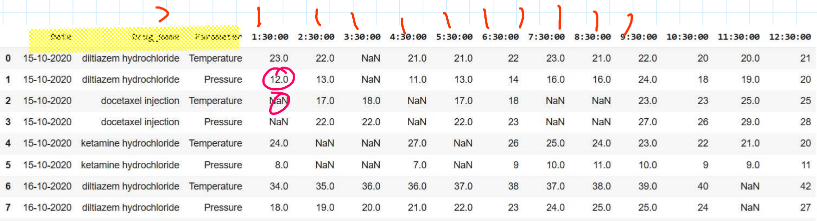
**How can we restructure our data back to the original wide format?**

df.pivot (index = ["date", "drug_name", "Parameter"], columns= "time", value = "reading")

(Reverse of melt)



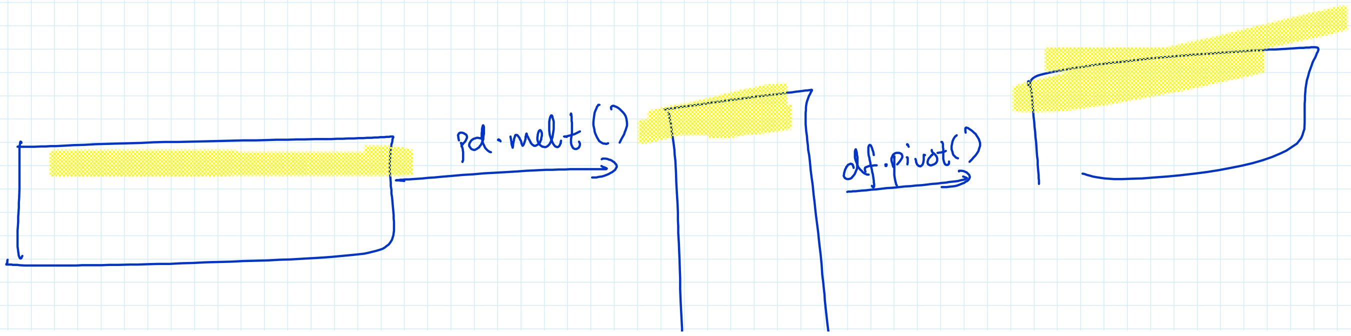
| | Date | Drug_Name | Parameter | time | reading |
|---|---|---|---|---|---|
| 0 | 15-10-2020 | diltiazem hydrochloride | Temperature | 1:30:00 | 23.0 |
| 1 | 15-10-2020 | diltiazem hydrochloride | Pressure | 1:30:00 | 12.0 |
| 2 | 15-10-2020 | docetaxel injection | Temperature | 1:30:00 | NaN |
| 3 | 15-10-2020 | docetaxel injection | Pressure | 1:30:00 | NaN |
| 4 | 15-10-2020 | ketamine hydrochloride | Temperature | 1:30:00 | 24.0 |
| ... | ... | ... | ... | ... | ... |
| 211 | 17-10-2020 | diltiazem hydrochloride | Pressure | 12:30:00 | 14.0 |



| | Date | Drug_Name | Parameter | 1:30:00 | 2:30:00 | 3:30:00 | 4:30:00 | 5:30:00 | 6:30:00 | 7:30:00 | 8:30:00 | 9:30:00 | 10:30:00 | 11:30:00 | 12:30:00 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 15-10-2020 | diltiazem hydrochloride | Temperature | 23.0 | 22.0 | NaN | 21.0 | 21.0 | 22 | 23.0 | 21.0 | 22.0 | 20 | 20.0 | 21 |
| 1 | 15-10-2020 | diltiazem hydrochloride | Pressure | 12.0 | 13.0 | NaN | 11.0 | 13.0 | 14 | 16.0 | 16.0 | 24.0 | 18 | 19.0 | 20 |
| 2 | 15-10-2020 | docetaxel injection | Temperature | NaN | 17.0 | 18.0 | NaN | 17.0 | 18 | NaN | NaN | 23.0 | 23 | 25.0 | 25 |
| 3 | 15-10-2020 | docetaxel injection | Pressure | NaN | 22.0 | 22.0 | NaN | 22.0 | 23 | NaN | NaN | 27.0 | 26 | 29.0 | 28 |
| 4 | 15-10-2020 | ketamine hydrochloride | Temperature | 24.0 | NaN | NaN | 27.0 | NaN | 26 | 25.0 | 24.0 | 23.0 | 22 | 21.0 | 20 |
| 5 | 15-10-2020 | ketamine hydrochloride | Pressure | 8.0 | NaN | NaN | 7.0 | NaN | 9 | 10.0 | 11.0 | 10.0 | 9 | 9.0 | 11 |
| 6 | 16-10-2020 | diltiazem hydrochloride | Temperature | 34.0 | 35.0 | 36.0 | 36.0 | 37.0 | 38 | 37.0 | 38.0 | 39.0 | 40 | NaN | 42 |
| 7 | 16-10-2020 | diltiazem hydrochloride | Pressure | 18.0 | 19.0 | 20.0 | 21.0 | 22.0 | 23 | 24.0 | 25.0 | 25.0 | 24 | NaN | 27 |

Tall $\xrightarrow{df.pivot()}$ wide

wide $\xrightarrow{pd.melt()}$ Tall



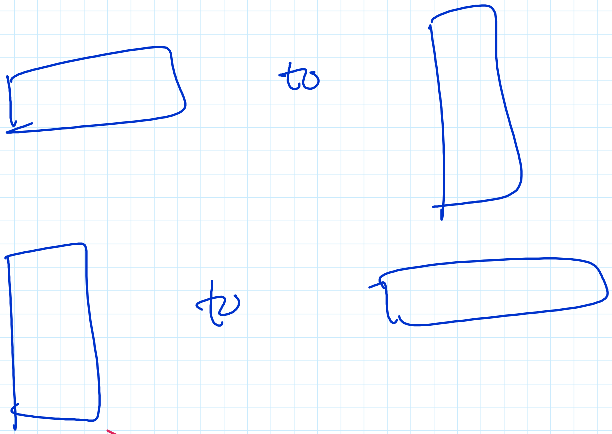
Pivot

Pivot_Table
(aggregation fn as well)

Summary

pd.melt()

df.pivot

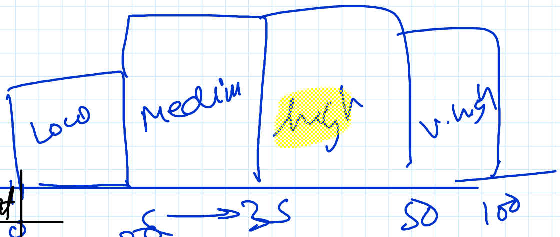


old times

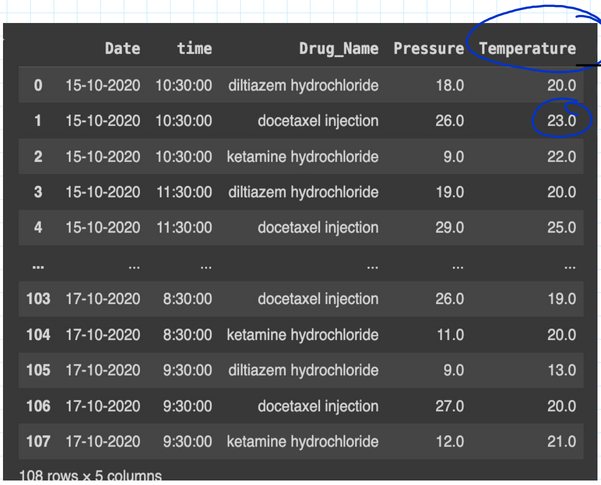
Pd.pivot_table( )

summary stat

Binning



*

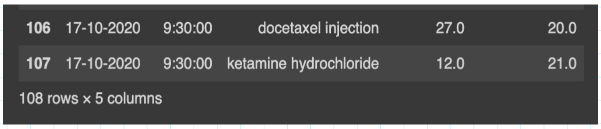


|  | Date | time | Drug_Name | Pressure | Temperature |
|---|---|---|---|---|---|
| 0 | 15-10-2020 | 10:30:00 | diltiazem hydrochloride | 18.0 | 20.0 |
| 1 | 15-10-2020 | 10:30:00 | docetaxel injection | 26.0 | 23.0 |
| 2 | 15-10-2020 | 10:30:00 | ketamine hydrochloride | 9.0 | 22.0 |
| 3 | 15-10-2020 | 11:30:00 | diltiazem hydrochloride | 19.0 | 20.0 |
| 4 | 15-10-2020 | 11:30:00 | docetaxel injection | 29.0 | 25.0 |
| ... | ... | ... | ... | ... | ... |
| 103 | 17-10-2020 | 8:30:00 | docetaxel injection | 26.0 | 19.0 |
| 104 | 17-10-2020 | 8:30:00 | ketamine hydrochloride | 11.0 | 20.0 |
| 105 | 17-10-2020 | 9:30:00 | diltiazem hydrochloride | 9.0 | 13.0 |
| 106 | 17-10-2020 | 9:30:00 | docetaxel injection | 27.0 | 20.0 |
| 107 | 17-10-2020 | 9:30:00 | ketamine hydrochloride | 12.0 | 21.0 |

108 rows × 5 columns

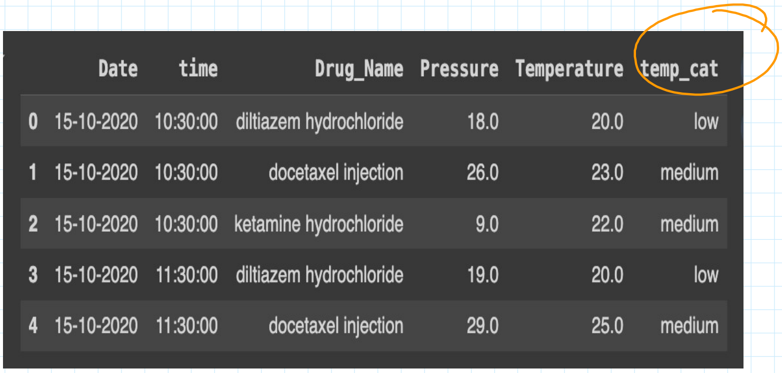
temp_cat
Low
Medium
Medium
Low
|
|
|
|

cut( )



```
106  17-10-2020  9:30:00       docetaxel injection  27.0  20.0
107  17-10-2020  9:30:00  ketamine hydrochloride    12.0  21.0
108 rows × 5 columns
```

pd.cut( )



|   | Date | time | Drug_Name | Pressure | Temperature | temp_cat |
|---|---|---|---|---|---|---|
| 0 | 15-10-2020 | 10:30:00 | diltiazem hydrochloride | 18.0 | 20.0 | low |
| 1 | 15-10-2020 | 10:30:00 | docetaxel injection | 26.0 | 23.0 | medium |
| 2 | 15-10-2020 | 10:30:00 | ketamine hydrochloride | 9.0 | 22.0 | medium |
| 3 | 15-10-2020 | 11:30:00 | diltiazem hydrochloride | 19.0 | 20.0 | low |
| 4 | 15-10-2020 | 11:30:00 | docetaxel injection | 29.0 | 25.0 | medium |

df.sort_index( )

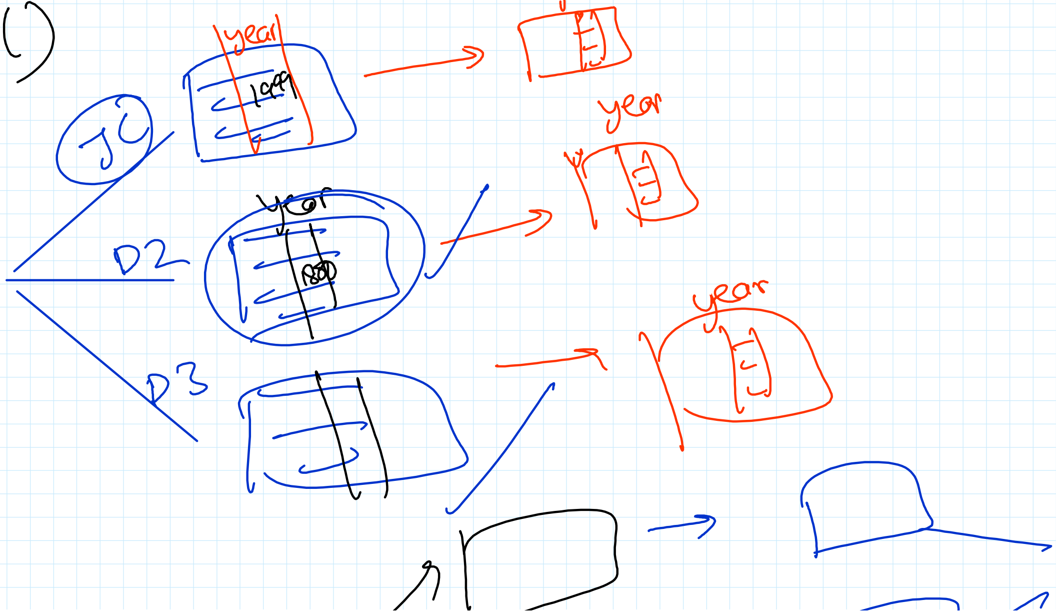
[# sort all the index]

Soln

| Store | Product |
|---|---|
| B | A1<br>A2 |
| A | B1<br>B2 |

year > 2003

groupby.filter( )



groupby.apply( )

# Apply custom function to each group