

# Moving Between Long and Wide Data Formats

---



**Paweł Kordek**

SOFTWARE ENGINEER

@pawel\_kordek <https://kordek.github.io>



# Summary



Understanding the layout

Differences between formats

Pandas API



# Properties

---

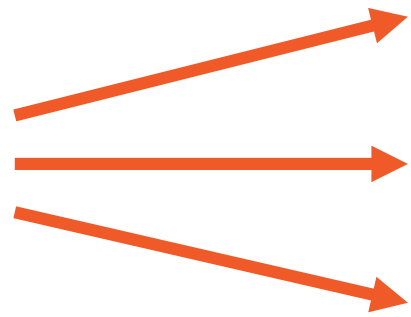


**playerID**      **goals**

nilsoma01	14
theodjo01	0
travepa01	2



Observations



playerID	goals
nilsoma01	14
theodjo01	0
travepa01	2



Observations

The diagram illustrates the relationship between observations and a measured variable. On the left, the word "Observations" is written in orange. Three orange arrows point from this text to the three rows of a table. Above the table, the column headers "playerID" and "goals" are written in orange. A blue arrow points from the text "Measured variable" (in blue) to the "goals" column header. The table itself has a light gray background and contains three rows of data.

playerID	goals
nilsoma01	14
theodjo01	0
travepa01	2



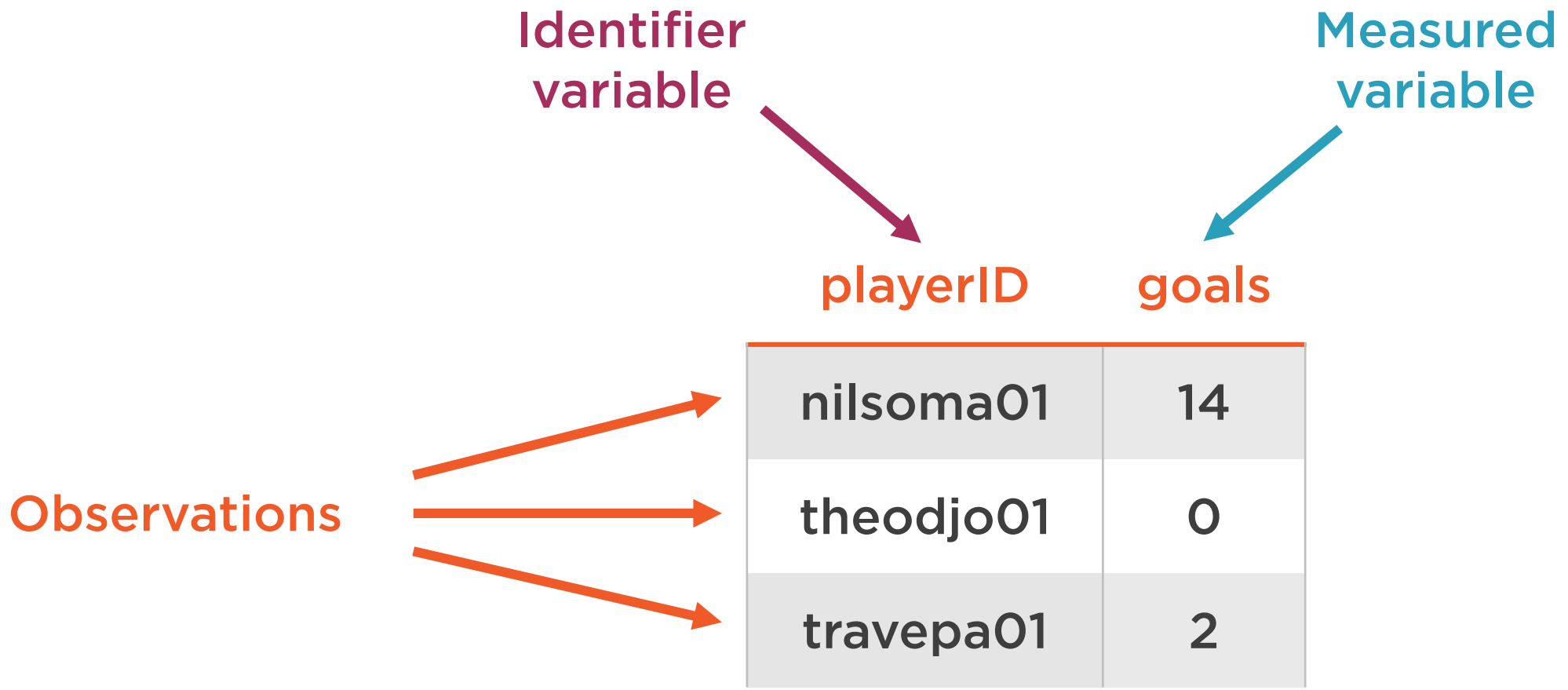
Identifier  
variable

Measured  
variable

playerID

goals

Observations



nilsoma01	14
theodjo01	0
travepa01	2



**playerID**      **year**      **goals**

nilsoma01	2001	14
nilsoma01	2002	15
theodjo01	2001	0
theodjo01	2002	0
travepa01	2001	2
travepa01	2002	0





playerID	year	goals
nilsoma01	2001	14
nilsoma01	2002	15
theodjo01	2001	0
theodjo01	2002	0
travepa01	2001	2
travepa01	2002	0

playerID	2001	2002
nilsoma01	14	15
theodjo01	0	0
travepa01	2	0



# Why Should You Care About This?



# Why Should You Care About This?



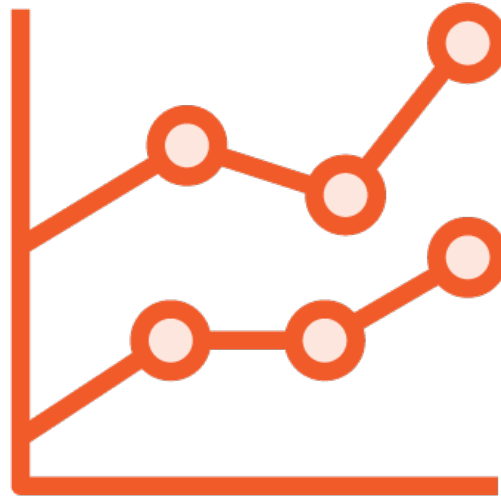
No standard in the wild



# Why Should You Care About This?



No standard in the wild



External tools



# Why Should You Care About This?



No standard in the wild



External tools



Easier analysis

# Simple Rules

---



playerID	year	goals
nilsoma01	2001	14
nilsoma01	2002	15
theodjo01	2001	0
theodjo01	2002	0
travepa01	2001	2
travepa01	2002	0

playerID	2001	2002
nilsoma01	14	15
theodjo01	0	0
travepa01	2	0



Column names should be  
variable names.





playerID	year	goals	assists
nilsoma01	2001	14	19
nilsoma01	2002	15	19



playerID	year	goals	assists
nilsoma01	2001	14	19
nilsoma01	2002	15	19

playerID	year	variable	value
nilsoma01	2001	goals	14.0
nilsoma01	2002	goals	15.0
nilsoma01	2001	assists	19.0



All your variables  
should correspond to  
column names.



**playerID**      **year**      **goals**      **assists**

nilsoma01	2001	14	19
nilsoma01	2002	15	19
theodjo01	2001	0	2
theodjo01	2002	0	2
travepa01	2001	2	3
travepa01	2002	0	13



In Pandas



```
df.melt(. . .)  
df.pivot(. . .)
```

In Pandas



# Summary



**Wide and long formats concept**

**Understanding the structure**

**Identifying variable types**