Joining Multiple Data Tables

Paweł Kordek
SOFTWARE ENGINEER

@pawel_kordek https://kordek.github.io

Joins

Purpose

Types

Support in Pandas

RDBMS

Why We Need Joins

playerID	name
hopkide01	Dean Hopkins
mairad01	Adam Mair
chaseke01	Kelly Chase

playerID	goals
hopkide01	4
mairad01	1
chaseke01	5

playerID	name	goals
hopkide01	Dean Hopkins	4
mairad01	Adam Mair	1
chaseke01	Kelly Chase	5

Types

playerID	goals
hopkide01	4
mairad01	1
chaseke01	5
hoggaje01	2

playerID	name
hopkide01	Dean Hopkins
mairad01	Adam Mair
chaseke01	Kelly Chase
antonmi01	Mike Antonovich

playerID	goals
hopkide01	4
mairad01	1
chaseke01	5
hoggaje01	2

playerID	name
hopkide01	Dean Hopkins
mairad01	Adam Mair
chaseke01	Kelly Chase
antonmi01	Mike Antonovich

playerID	name	goals
hopkide01	Dean Hopkins	4
mairad01	Adam Mair	1
chaseke01	Kelly Chase	5

playerID	goals
hopkide01	4
mairad01	1
chaseke01	5
hoggaje01	2

playerID	name
hopkide01	Dean Hopkins
mairad01	Adam Mair
chaseke01	Kelly Chase
antonmi01	Mike Antonovich

playerID	name	goals
hopkide01	Dean Hopkins	4
mairad01	Adam Mair	1
chaseke01	Kelly Chase	5

Inner

playerID	goals
hopkide01	4
mairad01	1
chaseke01	5
hoggaje01	2

playerID	name
hopkide01	Dean Hopkins
mairad01	Adam Mair
chaseke01	Kelly Chase
antonmi01	Mike Antonovich

playerID	goals
hopkide01	4
mairad01	1
chaseke01	5
hoggaje01	2

playerID	name
hopkide01	Dean Hopkins
mairad01	Adam Mair
chaseke01	Kelly Chase
antonmi01	Mike Antonovich

playerID	name	goals
hopkide01	Dean Hopkins	4
mairad01	Adam Mair	1
chaseke01	Kelly Chase	5
hoggaje01	NaN	2

playerID	goals
hopkide01	4
mairad01	1
chaseke01	5
hoggaje01	2

playerID	name
hopkide01	Dean Hopkins
mairad01	Adam Mair
chaseke01	Kelly Chase
antonmi01	Mike Antonovich

playerID	name	goals
hopkide01	Dean Hopkins	4
mairad01	Adam Mair	1
chaseke01	Kelly Chase	5
hoggaje01	NaN	2

Left

playerID	goals
hopkide01	4
mairad01	1
chaseke01	5
hoggaje01	2

playerID	name
hopkide01	Dean Hopkins
mairad01	Adam Mair
chaseke01	Kelly Chase
antonmi01	Mike Antonovich

playerID	goals
hopkide01	4
mairad01	1
chaseke01	5
hoggaje01	2

playerID	name
hopkide01	Dean Hopkins
mairad01	Adam Mair
chaseke01	Kelly Chase
antonmi01	Mike Antonovich

playerID	name	goals
hopkide01	Dean Hopkins	4
mairad01	Adam Mair	1
chaseke01	Kelly Chase	5
antonmi01	Mike Antonovich	NaN

playerID	goals
hopkide01	4
mairad01	1
chaseke01	5
hoggaje01	2

playerID	name
hopkide01	Dean Hopkins
mairad01	Adam Mair
chaseke01	Kelly Chase
antonmi01	Mike Antonovich

playerID	name	goals
hopkide01	Dean Hopkins	4
mairad01	Adam Mair	1
chaseke01	Kelly Chase	5
antonmi01	Mike Antonovich	NaN

Right

playerID	goals
hopkide01	4
mairad01	1
chaseke01	5
hoggaje01	2

playerID	name
hopkide01	Dean Hopkins
mairad01	Adam Mair
chaseke01	Kelly Chase
antonmi01	Mike Antonovich

playerID	goals
hopkide01	4
mairad01	1
chaseke01	5
hoggaje01	2

playerID	name
hopkide01	Dean Hopkins
mairad01	Adam Mair
chaseke01	Kelly Chase
antonmi01	Mike Antonovich

playerID	name	goals
hopkide01	Dean Hopkins	4
mairad01	Adam Mair	1
chaseke01	Kelly Chase	5
hoggaje01	NaN	2
antonmi01	Mike Antonovich	NaN

playerID	goals
hopkide01	4
mairad01	1
chaseke01	5
hoggaje01	2

playerID	name
hopkide01	Dean Hopkins
mairad01	Adam Mair
chaseke01	Kelly Chase
antonmi01	Mike Antonovich

playerID	name	goals
hopkide01	Dean Hopkins	4
mairad01	Adam Mair	1
chaseke01	Kelly Chase	5
hoggaje01	NaN	2
antonmi01	Mike Antonovich	NaN

Outer

Source tables

id	val_1
A	1
B	2

id	val_2
A	11
C	3

Joins

Inner

id	v1	v2
A	1	11

Left

id	v1	v2
A	1	11
B	2	NaN

Outer

id	v1	v2
A	1	11
B	2	NaN
C	NaN	3

Right

id	v1	v2
A	1	11
C	NaN	3

Support in Pandas

`pd.merge`

Minimal Join in Pandas

```
pd.merge(df1, df2)
```

Minimal Join in Pandas

Two data frames are mandatory

```
pd.merge(df1, df2, how='left')
```

Minimal Join in Pandas

Type – default is 'inner'

playerID	goals	position
hopkide01	4	R
mairad01	1	C
chaseke01	5	R

playerID	name
hopkide01	Dean Hopkins
mairad01	Adam Mair
chaseke01	Kelly Chase

playerID	goals	position
hopkide01	4	R
mairad01	1	C
chaseke01	5	R

playerID	name
hopkide01	Dean Hopkins
mairad01	Adam Mair
chaseke01	Kelly Chase

playerID	goals	position	name
hopkide01	4	R	Dean Hopkins
mairad01	1	C	Adam Mair
chaseke01	5	R	Kelly Chase

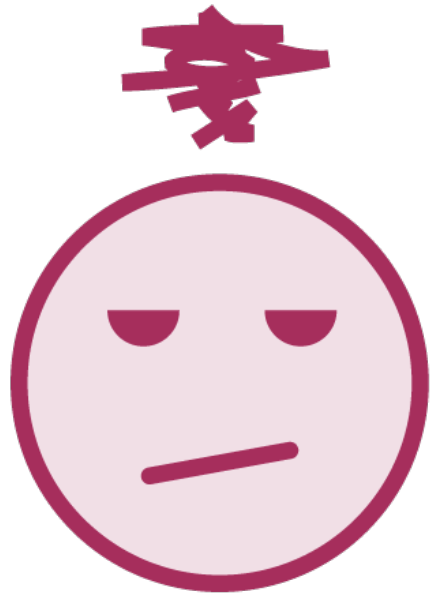
playerID	goals	position
hopkide01	4	R
mairad01	1	C
chaseke01	5	R

playerID	name	position
hopkide01	Dean Hopkins	Right Wing
mairad01	Adam Mair	Center
chaseke01	Kelly Chase	Right Wing

playerID	goals	position
hopkide01	4	R
mairad01	1	C
chaseke01	5	R

playerID	name	position
hopkide01	Dean Hopkins	Right Wing
mairad01	Adam Mair	Center
chaseke01	Kelly Chase	Right Wing

playerID	goals	position	name



```
pd.merge(df1, df2,
         on="playerID")
```

	playerID	goals	position
0	hopkide01	4	R
1	mairad01	1	C
2	chaseke01	5	R

	playerID	name
0	hopkide01	Dean Hopkins
1	mairad01	Adam Mair
2	chaseke01	Kelly Chase

```
pd.merge(df1, df2,
         left_on="playerID“,
         right_on="plrID")
```

	playerID	goals	position
0	hopkide01	4	R
1	mairad01	1	C
2	chaseke01	5	R

	plrID	name
0	hopkide01	Dean Hopkins
1	mairad01	Adam Mair
2	chaseke01	Kelly Chase

```
pd.merge(df1, df2,
         left_index=True,
         right_on="plrID")
```

	goals	position
hopkide01	4	R
mairad01	1	C
chaseke01	5	R

	plrID	name
0	hopkide01	Dean Hopkins
1	mairad01	Adam Mair
2	chaseke01	Kelly Chase