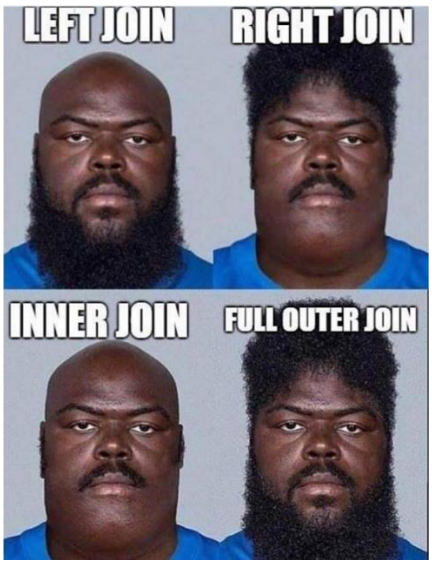
JOINS CHEATSHEET

Data:

```
val playersTeam = Seq (
(1,"Iniesta", "F.C Barcelona"),
(2,"Messi", "F.C Barcelona"),
(3,"Pique", "F.C Barcelona"),
(4,"Xavi", "F.C Barcelona"),
(5,"Puyol", "F.C Barcelona"),
(6,"Iker Casillas", "Real Madrid"),
(7,"Sergio Ramos", "Real Madrid"),
(8,"Cristiano Ronaldo", "Real Madrid"),
(9,"Raul", "Real Madrid"),
(10,"Benzema", "Real Madrid"),
(11,"Guti", "Real Madrid"),
(12,"Victor Valdes", "F.C Barcelona")
).toDF("Player_id","Player_name", "Team")
```

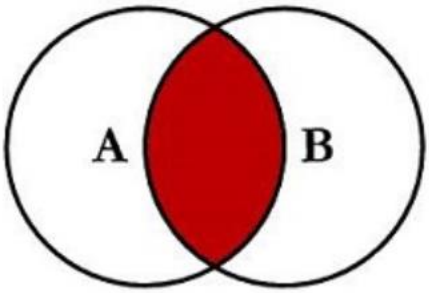
```
val statistics = Seq (
  (1,"Iniesta",24,2,10),
  (2,"Messi",65,4,20),
  (3,"Pique",78,20,21),
  (4,"Xavi",54,10,8 ),
  (5,"Puyol",110,15,25),
  (6,"Iker Casillas",100,10,20),
  (7,"Sergio Ramos",54,60,25),
  (8,"Cristiano Ronaldo",100,34,35),
  (9,"Raul",100,0,59),
  (13,"Neymar",25,25,8),
  (14,"Ronaldo",76,6,12)
).toDF("Player_id","Player_name", "matches","Cards","Cups")
```

The most common Joins I use in Spark:



INNER JOIN

The request would be… we are going to take into consideration the players who has information about the team he plays and his statistics. We do not want players with null information (in this case).

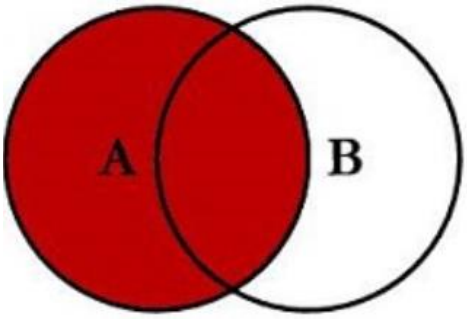



```
playersTeam.join (statistics, Seq ("Player_id","Player_name"),
"inner").show()
```

Player_id	Player_name	Team	matches	Cards	Cups
1	Iniesta	F.C Barcelona	24	2	10
2	Messi	F.C Barcelona	65	4	20
3	Pique	F.C Barcelona	78	20	21
4	Xavi	F.C Barcelona	54	10	8
5	Puyol	F.C Barcelona	110	15	25
6	Iker Casillas	Real Madrid	100	10	20
7	Sergio Ramos	Real Madrid	54	60	25
8	Cristiano Ronaldo	Real Madrid	100	34	35
9	Raul	Real Madrid	100	0	59

LEFT OUTER

The usual request to use this kind of join is: take all the players who has information about the team he plays and then fill the null values in the statistics with some value…

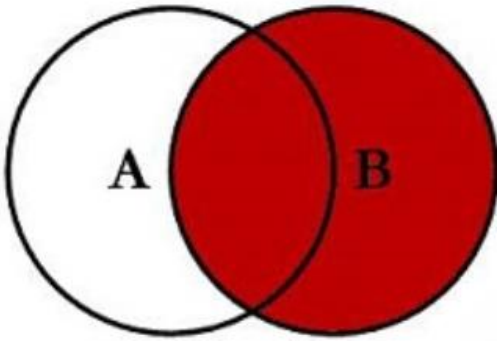



```
playersTeam.join (statistics, Seq ("Player_id","Player_name"),
"left_outer").show()
```

```
+---------+-----------------+-------------+-------+-----+----+
|Player_id|      Player_name|         Team|matches|Cards|Cups|
+---------+-----------------+-------------+-------+-----+----+
|        1|          Iniesta|F.C Barcelona|     24|    2|  10|
|        2|            Messi|F.C Barcelona|     65|    4|  20|
|        3|            Pique|F.C Barcelona|     78|   20|  21|
|        4|             Xavi|F.C Barcelona|     54|   10|   8|
|        5|            Puyol|F.C Barcelona|    110|   15|  25|
|        6|    Iker Casillas|  Real Madrid|    100|   10|  20|
|        7|     Sergio Ramos|  Real Madrid|     54|   60|  25|
|        8|Cristiano Ronaldo|  Real Madrid|    100|   34|  35|
|        9|             Raul|  Real Madrid|    100|    0|  59|
|       10|          Benzema|  Real Madrid|   null| null|null|
|       11|             Guti|  Real Madrid|   null| null|null|
|       12|    Victor Valdes|F.C Barcelona|   null| null|null|
+---------+-----------------+-------------+-------+-----+----+
```

RIGHT OUTER

The opposite of left outer: take the player who have information about his statistics and leave the team value with null.

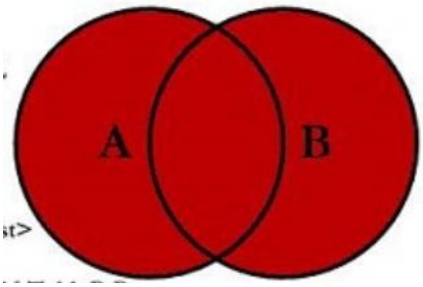



```
playersTeam.join (statistics, Seq ("Player_id","Player_name"),
"right_outer").show()
```

Player_id	Player_name	Team	matches	Cards	Cups
1	Iniesta	F.C Barcelona	24	2	10
2	Messi	F.C Barcelona	65	4	20
3	Pique	F.C Barcelona	78	20	21
4	Xavi	F.C Barcelona	54	10	8
5	Puyol	F.C Barcelona	110	15	25
6	Iker Casillas	Real Madrid	100	10	20
7	Sergio Ramos	Real Madrid	54	60	25
8	Cristiano Ronaldo	Real Madrid	100	34	35
9	Raul	Real Madrid	100	0	59
13	Neymar	null	25	25	8
14	Ronaldo	null	76	6	12

FULL OUTER

Load all the data, although columns team and statistics will have null values.

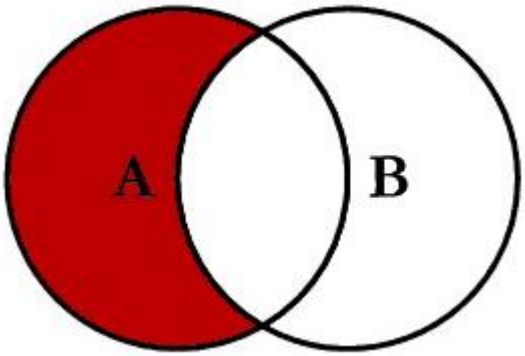



```
playersTeam.join (statistics, Seq ("Player_id","Player_name"),
"outer").show()
```

Player_id	Player_name	Team	matches	Cards	Cups
13	Neymar	null	25	25	8
14	Ronaldo	null	76	6	12
6	Iker Casillas	Real Madrid	100	10	20
1	Iniesta	F.C Barcelona	24	2	10
2	Messi	F.C Barcelona	65	4	20
4	Xavi	F.C Barcelona	54	10	8
5	Puyol	F.C Barcelona	110	15	25
9	Raul	Real Madrid	100	0	59
10	Benzema	Real Madrid	null	null	null
11	Guti	Real Madrid	null	null	null
12	Victor Valdes	F.C Barcelona	null	null	null
7	Sergio Ramos	Real Madrid	54	60	25
8	Cristiano Ronaldo	Real Madrid	100	34	35
3	Pique	F.C Barcelona	78	20	21

LEFTANTI

It takes only the columns from the left table and the data that are not in the right table. In this case the question from the analytics part is: HEY! **Tell me the players who he does not have statistics values.**




```
playersTeam.join (statistics, Seq ("Player_id","Player_name"),
"leftanti").show()
```

```
+---------+-------------+-------------+
|Player_id|  Player_name|         Team|
+---------+-------------+-------------+
|       10|      Benzema|  Real Madrid|
|       11|         Guti|  Real Madrid|
|       12|Victor Valdes|F.C Barcelona|
+---------+-------------+-------------+
```

LEFTSEMI

Take the left table data for these players who are in the statistics (right table):

```
playersTeam.join (statistics, Seq ("Player_id","Player_name"),
"leftsemi").show()
+---------+-----------------+-------------+
|Player_id|      Player_name|         Team|
+---------+-----------------+-------------+
|        1|          Iniesta|F.C Barcelona|
|        2|            Messi|F.C Barcelona|
|        3|            Pique|F.C Barcelona|
|        4|             Xavi|F.C Barcelona|
|        5|            Puyol|F.C Barcelona|
|        6|    Iker Casillas|  Real Madrid|
|        7|     Sergio Ramos|  Real Madrid|
|        8|Cristiano Ronaldo|  Real Madrid|
|        9|             Raul|  Real Madrid|
+---------+-----------------+-------------+
```