1 - first column without name, I suppose it is an ID given that it is unique

2 - duplicate columns: remaining\_min\_1 smallint, power\_of\_shot\_1 tinyint, knockout\_match\_1 tinyint, remaining\_sec\_1 smallint, distance\_of\_shot\_1 tinyint

- Be careful about the duplicate columns! One strategy to manage missing values is to get the value from the first column (for example remaining\_min) and if it is a null take the value from the second column (for example remaining\_min\_1), or viceversa. But we decided to adopt another strategy for missing values given that it seems that the duplicated column doesn't contain the right value. For example:

Immagine che contiene tavolo

Descrizione generata automaticamente

Id = 4 is a missing value, and the value of remaining\_min\_1 is equal to 43. It is obviously not correct given should be a number between 6 and 9. Furthermore, there are decimal values for the duplicated columns#2 but this is not correct because, for example remaining\_min\_1 should be an integer and the seconds should go in remaining\_sec\_1, the same for power of shot should be a range between 1 and 7 rather than a decimal value:

Immagine che contiene tavolo

Descrizione generata automaticamente

So, we decided just to drop the columns that ends with \_1 because data seems not to be correct

3 - the column lat/lng contains commas (csv separator) and to resolve it is possible to use FIELD\_OPTIONALLY\_ENCLOSED\_BY = '"' during the creation of file format

4 - be careful to game\_season "2000-01" considered as a string but should be a date

5 – IN 1999 Ronaldo dindn’t played for Manchester United

6 – Manchester United always play home? Never Away?

7 - missing values

**FIRST ROUND**:

- ID, OK

- MATCH\_EVENT\_ID, we can't do nothing given that is an FK

- SHOT\_ID\_NUMBER, the same as MATCH\_EVENT\_ID

- LOCATION\_X, LOCATION\_Y filled with 0

- REMAINING\_MIN, we compute an average for a specific match between previous and next row

- POWER\_OF\_SHOT, we use the most frequent value in the table (MODE)

- GAME\_SEASON, we can retrieve the value given that we verified that the same value is unique for each match\_id.   
 For the NULL we just take the information from another row.

- REMAINING\_SEC, we do something of similar as REMANING\_MIN but considering not only the single match

but also the minute.

* AREA\_OF\_SHOT, in this case, we decide to leave null values and connect to the dimensions using the dummy row
* SHOT\_BASICS, also in this case, we manage null with dummies
* RANGE\_OF\_SHOT studying other data, we defined the following ranges:

Immagine che contiene testo

Descrizione generata automaticamente

For this we also create another column called RANGE\_OF\_SHOT\_IN\_FEETS

* TEAM\_NAME, according to transfert market we defined the following range:



* DATE\_OF\_GAME is always populated in other rows but one (match id = '29600031'). So we can get the same info from other rows
* LAT\_LNG:
  + - We created LATITUDE column
    - And LONGITUDE column
* TYPE\_OF\_SHOT and TYPE\_OF\_COMBINED\_SHOT is merged in one column called TYPE\_OF\_SHOT and we created a flag (column IS\_COMBINED). This because we verified that a Shot can’t be both combined and not combined
* MATCH\_ID and TEAM\_ID are always populated
* KNOCKOUT\_MATCH Same logic of Date of game, we can retrieve the information from another row
* DISTANCE\_OF\_SHOT, we use the most frequent value in the table (MODE)
* HOME\_AWAY Same logic of Date of game, we can retrieve the information from another row

**SECOND ROUND**

* Integration of team names in two different columns (home and away) from home\_away. Team names comes from excel file
* Game season
* EXTERNAL FILE FOR TEAMS AND FOR IS\_GOAL (REGRESSION)