**Project Subject:** NBA information

**Project team:**

Cuitláhuac Maldonado

Guadalupe Vázquez

**Summary**

In this project we choose to explore and work with information related to the NBA. For this, we practice the information gathering using web scraping from the NBA stats web page, once we have the information for the seasons we wanted, he use pandas for formatting and clean the data already gathered and finally we load this information into a Database no relational as Mongo.

We decided to use this information because is a common interest topic and we have a good level of knowledge about to handle the content of the information.

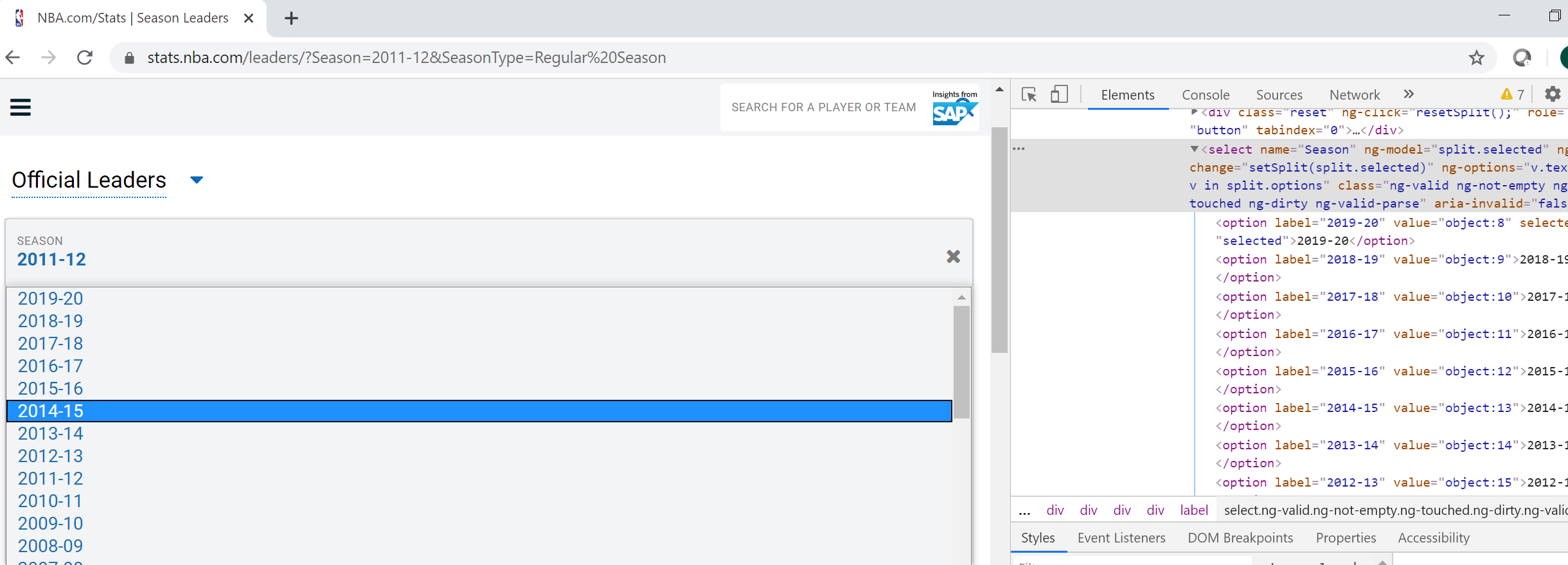
**Extract**

To extract the information we use the following link that contains a good amount of information about the performance of the players and teams form all seasons. For this project we decided to use the information of the season 2000 to the last documented 2019.

URL: <https://stats.nba.com/>

To obtain the information, we use the BeautifulSoup and splinter libraries to extract the tables of the players and the teams by sweeping through the different seasons, additionally, we use the option to inspect to locate the object corresponding to the particular object “season” and by using a for loop, the information was obtained from both: players and teams for the defined period of seasons.

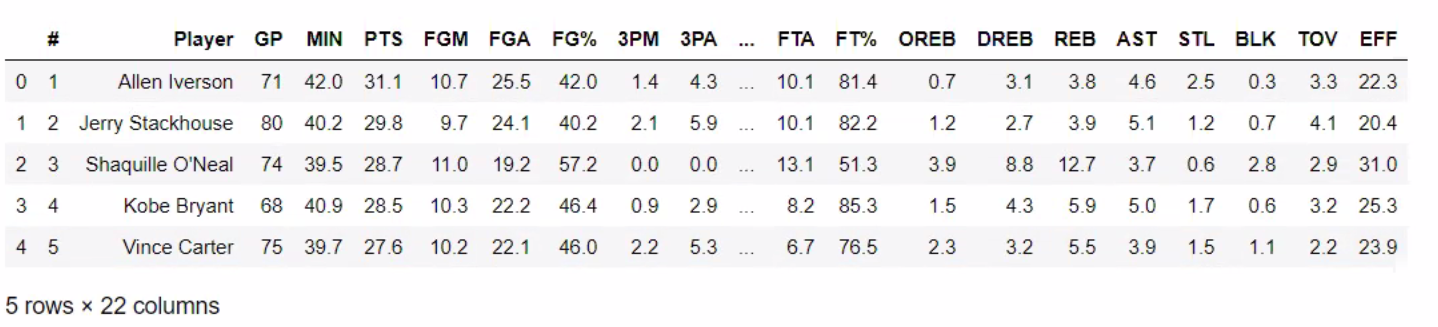
Bs4

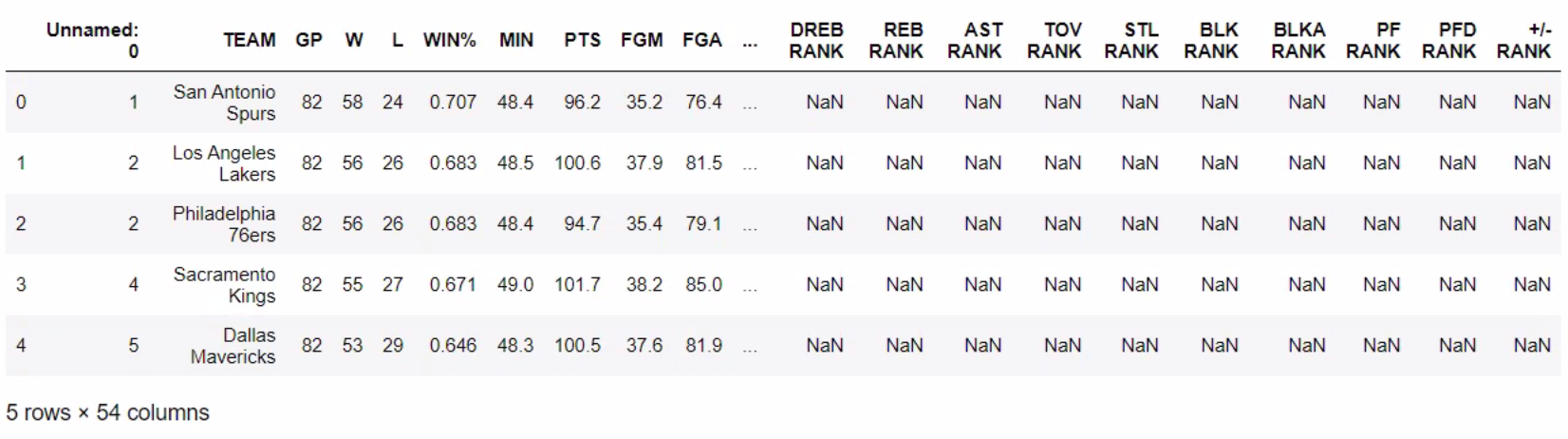


It is important to say that to be able to visualize the information of the tables correctly as a data frame it was necessary to apply the attribute of "prettify ()" to the html data since if this option did not return the information in the appropriate format.

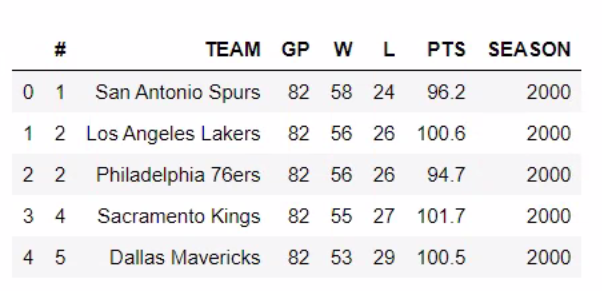
**Transform**

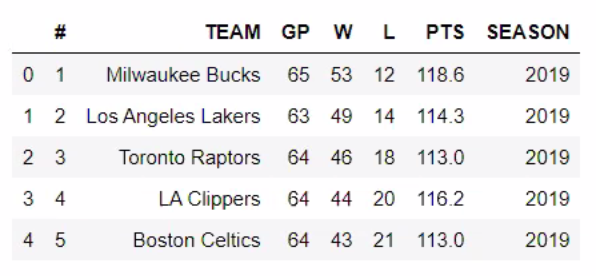
Once we have the information stored in the different tables by season, both players and teams, all the fields that did not bring information were removed, additionally, the fields that we consider relevant to be included in the database were selected and added into a data frame, other fields less important for us, like offensive reb, defensive reb, minutes played, and other that we didn´t know the meaning were removed.





Having all data frames for each of the seasons, it was necessary to add a column in each data frame with the specific season to which they belonged, so that when loaded into the database we could gather all the information of players in one table and that of teams in another and do not have different collections for each season.





**Load**

Finally, to load the information we used Mongo DB because it is a database non-relational that provide more flexible schema to data management, so as our information does not have relation between the collections we decided it is the better option for us.

We load the information using a for loop for each row in the data frame to insert each document in the proper collection.

This database has two collections: players\_seasons and teams\_seasons and the structure of the document in each collection is as follows:

players\_seasons: {'name', 'games', 'points','rebounds', 'assists', 'blocks', 'season'}

teams\_seasons: {'rank', 'name', 'games', 'wins', 'loses', 'points', 'season'}

Bellow the output of each collection in the terminal.

