Project Report: Hey now, you’re an All-Star!

Team 14: Joseph Mangahas and Sungbae Lee

11/16/2019

**Extract:**

For our data extraction, we obtained our data from websites such as, Kaggle and Data World. From Kaggle, we obtained a data set that contained player statistics from each NBA All-Star game from the year 1950 to 2009. This data contained information such as points made, blocks, rebounds, and other stats for each player. The data set we obtained from Data World stored general player info for each All-Star participant. General info such as player height, weight, and team they played for.

The Kaggle data set was stored as a .csv file and we used the following code to transform the data to a pandas data frame:



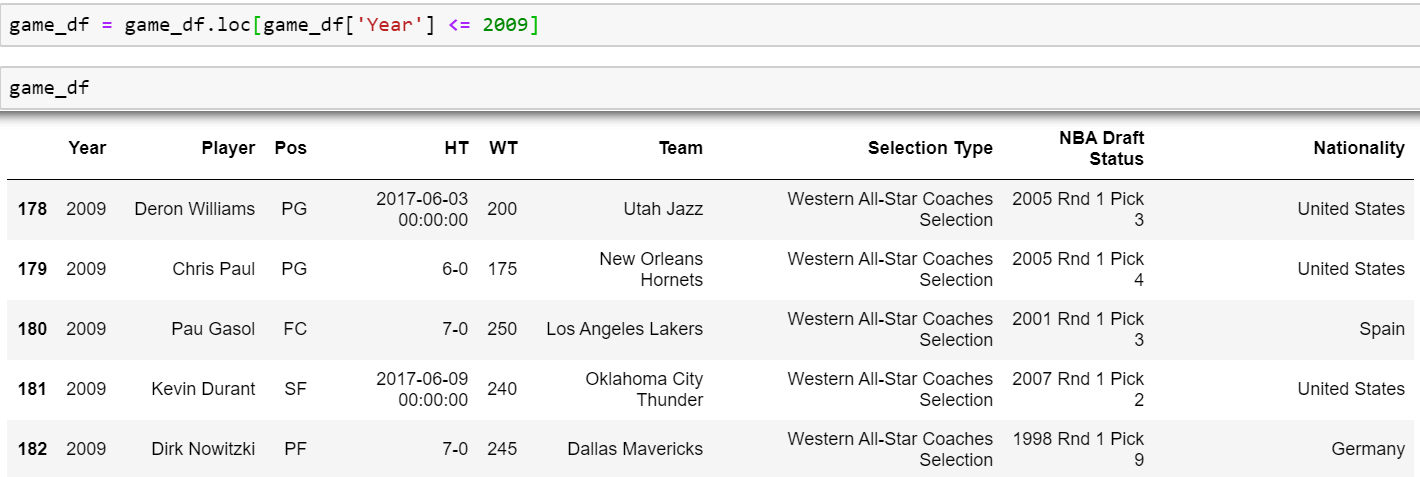
Our Data World data was stored as an excel file and we used, pd.read\_excel() in order to upload to a data frame. See coding below:

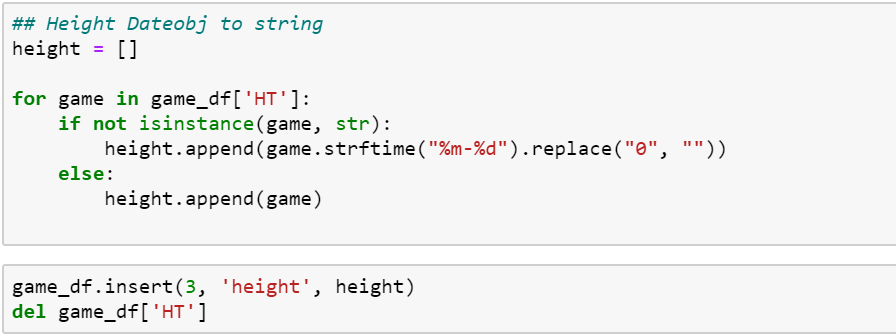


**Transformation:**

For our Kaggle data set, we had to drop all values with anything in relation to the season\_id (year) because, our Data World data set did not contain any player information from the year 1999 and earlier. We also had to drop all values after 2009 in our Data World set as our Kaggle data did not contain data after the year 2009.

Filtering by including only data from 2000 and after, and dropping all values from 1999 and before (Kaggle Set):

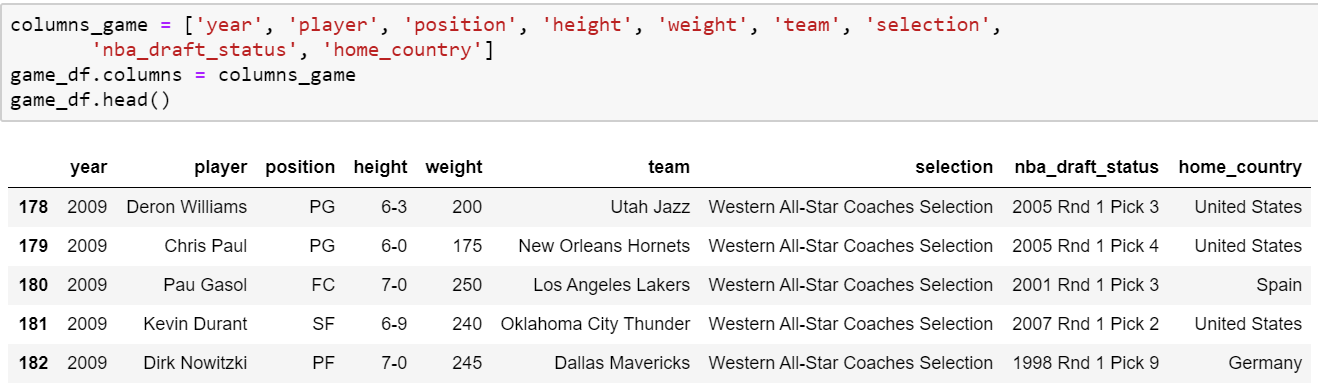
Filtering out data after the year of 2009 and only including data from before 2009 (Data World):

As also seen on our game\_df pictured above, the height (HT) outputted as a date. The code pictured to the right, is what we ran to fix this issue.

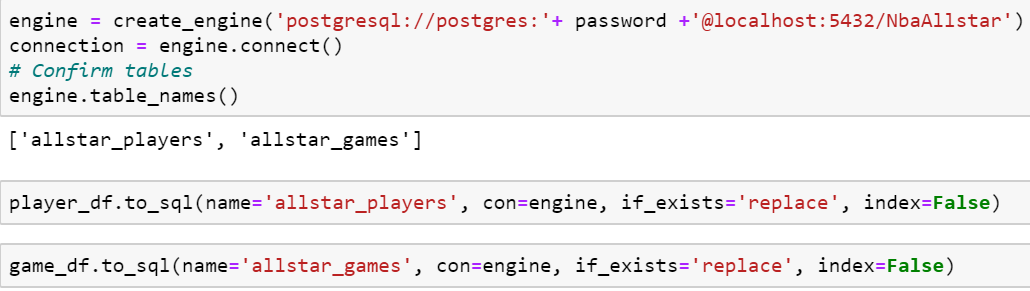
We also cleaned up our column names. Below is how we re-named our columns:

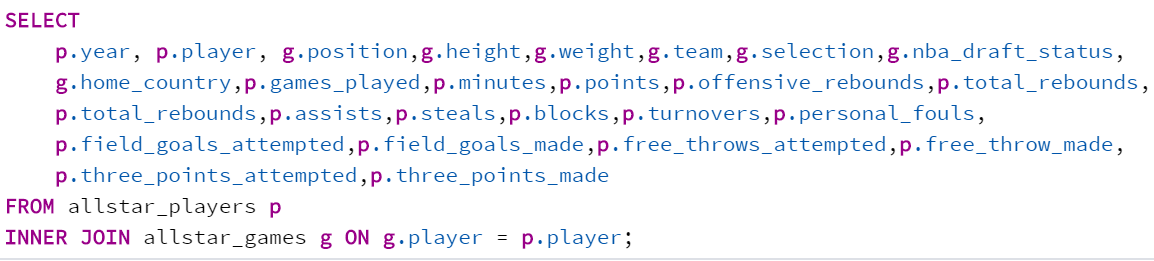
Before:

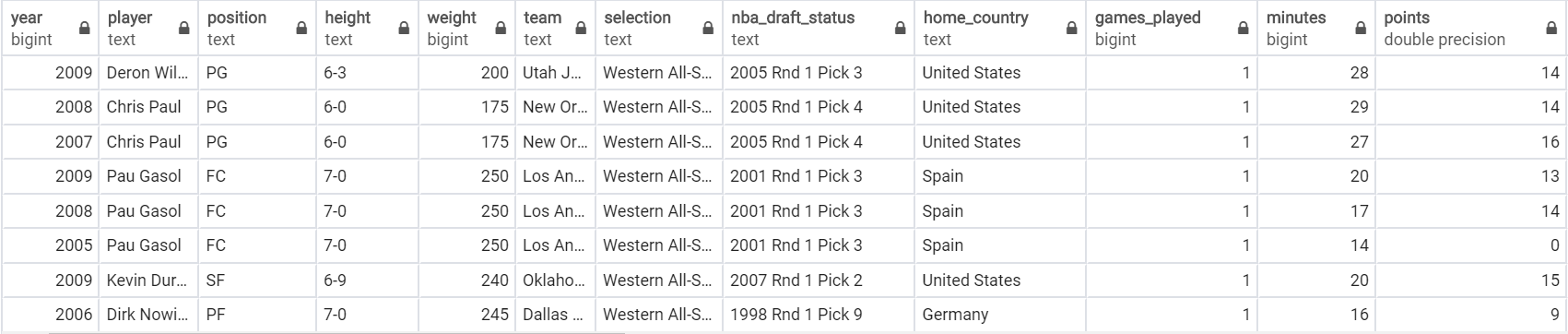
After:



**Load:**

Once we cleaned our data, we created to connection to load our data to postgres.

Below is how we joined our data:

Here is a sample of our joined tables:

We loaded our database as a relational database because both our data sets were stored in tables. And both tables shared the same column name of, player name. Postgres was the best tool to use to combine these tables, that shared the same column name. Our data also contained numerous columns and rows, and Postgres is the best way to interface with our final database.