**Database**

* Database that we will use is PostgreSQL (PG Admin)

**Cleaning the Data Set**

* From Kaggle data set (Season Stats, Players, Player Data) the main one we will be using is Season Stats
* First, we drop unwanted columns like Games Started and Position among others from Season Stats
* The remove data from before the 2003 season as the data set is from 1950-2018 and we want 2003-2018
* Next, we would join on players from the Players csv to get the height and weight.
* Then, we would have to create a new column which will be the year the player is in, to get this we would take the year stated column from Plyer Data and subtract that number form the year column in Season Stats
* Lastly, we need to remove the duplicates from the team column, for when a player played on multiple teams in a year and keep the Total.
  + I.e., if a player plays on 2 teams, they would be shown 3 times, once with team 1, once with team 2 and once with Total, so we keep the total

**After Cleaning**

* We will feed the dataset into PCA