**Team Stats Explainer**

I think right now the idea is to end up with a dataframe with the columns shown below that is used to train the model. It will have 30 rows for the 30 teams in the NBA and will only include statistics from the 2017 season (for now). Every column other than team\_name, abv, and wins\_tot will have to be calculated from Seasons\_Stats.csv using group\_by functions, grouped by the Tm column. An important note is that when the team stats dataframe is created from all the groupby functions (before it’s joined to the columns from ranking.csv) it’s going to have 31 rows because there’s going to be a TOT row. The TOT row needs to be dropped from the dataframe before it can be joined to the rankings.csv columns otherwise the dimensions won’t match up.

Here are the columns that I think should be what we need for the model. I have explanations for each column as well and how they are calculated.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **team\_name** | **abv** | **pts\_per\_min** | **2fg\_pct** | **3fg\_pct** | **ts\_pct** | **dbpm\_tot** | **obpm\_tot** | **wins\_tot** |
|  |  |  |  |  |  |  |  |  |

**team\_name** – name of team from ranking.csv (TEAM column)

**abv** – team abbreviation from Seasons\_Stats.csv (Tm column)

**pts\_per\_min (points per minute) –** total points the team scored divided by minutes played (dividing by minutes played should help account for when some teams go into overtime more than others). Should be calculated:

**pts\_per\_min = SUM(PTS)/SUM(MP)**

**2fg\_pct (2-point field goal percentage) –** percentage of two-point shots the team made by how many they attempted. Should be calculated:

**2fg\_pct = SUM(2P)/SUM(2PA)**

**3fg\_pct (3-point field goal percentage) –** percentage of three-point shots the team made by how many they attempted. Should be calculated:

**3fg\_pct = SUM(3P)/SUM(3PA)**

**ts\_pct (true shooting percentage) –** this is pretty commonly used to compare shooting stats because it takes into account the fact that some players shoot more three-point shots than others. Here’s a link to the Wikipedia article: <https://en.wikipedia.org/wiki/True_shooting_percentage>

Should be calculated:

**ts\_pct = SUM(PTS)/(2\*(SUM(FGA) + 0.44\*SUM(FTA)))**

**dbpm\_tot (Defense Box Plus/Minus Total) –** DBPM is a number created to compare each player’s contributions to defense to each other no matter what team they play for. Here’s a more in-depth explainer of it: <https://www.sports-reference.com/blog/2014/10/introducing-box-plusminus-bpm-2/>

Basically the idea is to add up each team’s player’s DBPM to see which teams have the better defensive players. Since the number doesn’t take into account how many minutes each player plays, we have to weight it by each player’s minutes. Should be calculated like this:

**dbpm\_tot = SUM(DBPM\*MP)**

**obpm\_tot (Offensive Box Plus/Minus Total) –** Basically the same as DBPM but for offense. Calculated similarly:

**obpm\_tot = SUM(OBPM\*MP)**

**wins\_tot (Total Wins) –** Taken directly from the total wins (W column) in ranking.csv. This is the outcome variable we’re going to train the model to predict.