Using NBA Stats to Analyze Interesting Facts on Basketball

Hang Xu

Date: April 19th 2015
Goal: primary analysis on the data of final project
Data collected from : NBAStats API
Data Range: Players'/NBA team's data in season 2013-2014 regular season (previous season)
Data statistics:
References:
http://www.danielforsyth.me/exploring_nba_data_in_python/
http://nbviewer.ipython.org/urls/dl.dropbox.com/s/dtfjeafyo7ut6f7/nbastats_docs.ipynb?create=1
http://www.cbssports.com/nba/eye-on-basketball/24840935/rockets-choose-to-live-by-the-three-as-they-shoot-historic-numbers

---

Introduction

As a basketball fan, I am really enjoy NBA games. At this time, it is NBA playoff season and every team wants to win the league. This makes me curious is there a same statistical pattern for every single winning team in NBA. In other words, does every winning team have a higher field goal percentage; does every winning team is the one that has more assists than those of their opponents, etc. Therefore, along with my curiosity, my final project uses NBAstats API, which is a python module written by its author, Bradley Fay that makes it easier and faster to retrieve data from stats.nba.com, to analyze some interesting facts on basketball, such as "Is there a correlation between height and free-throw percentage", "Does the team which has more assists in a game more likely actually win the game?", "Is three-pointers really crucial for winning the game?".

Hypothesis

We could make as many as hypothesis we want for this topic. Given the deadline of this project, we just choose several main hypotheses to analyze. Feel free to make more hypothesis and perform the analysis.

1. Is there a correlation between height and free-throw percentage?
2. Does the team which has more assists in a game more likely actually win the game?
3. Does three-pointers really help teams win the game?

Besides these three hypotheses, we also perform some other trivial analysis over the entire project. We do not want to expand them because we have limited time for this project. However, they are definitely very fun topics for other people to explore.

Data Collection

Importing the API to the ipython, we can use many commands to collect data. "ShortChart" is a method to see information on a single player short chart. It needs the player id as a parameter and we can easily to find that in NBA stats webpage. For this project, I choose to analyze ten different players: **Tim Duncan (PF), Dirk Nowitzki (PF), Lebron James (SF), DeAndre Jordan (C), Marc Gasol (C), Kevin Durant (SF), James Harden (SG), Klay Thompson (SG) Chris Paul(PG), and Stephen Curry (PG).** They are all top players in NBA and has different positions. In addition, they all play more than fifty-eight (seventy percent of total eighty-two season games) games and qualify the NBA stats minimum requirements for taking into account. Since this API call returns specific data in pandas dataframe, it is very easy for us to manipulate and adjust the data we would with a normal pandas dataframe that we studies in lectures and assignments.

This API retrieves data from 2013-2014 season, which is the previous season. It works better than we use the data from this season because, it is too soon to analyze the players stats for this year, since the regular season game is just finished and staffs in NBA might still need to edit and correct the stats in the follow months.

Furthermore, it is hard to find a distribution for the data we collected since the stats are not random variables and it is just counts of a players' performance in games. We should put our attention on what the actual meaning of these stats represents and how it helps with accepting or rejecting our hypothesis we made before performing the analysis.

How to justify our hypothesis using the data we collected?

For our first hypothesis: Is there a correlation between height and free-throw percentage?

We could collect fifty NBA active players and their free throw percentage, comparing them and using linear and logistic regression model to fit them and see whether we could find any correlation.

For the second hypothesis: Does the team which has more assists in a game more likely actually win the game?

We could find fifty games' data using their game id and find the counts of assists between the winning team and losing them, plotting their distributions and observe if there is a pattern justifying our hypothesis. We could also find several pairs of two teams and compare their counts of assists per game. If most team that has a higher average assist per game wins, it will help us justify our hypothesis as well. In addition, we could group the all winning games of a team together, and then the losing games, observing if there is any difference on the assist category.

For the third hypothesis: Does three-pointers really help teams win the game?

Since three-point is really more and more popular in NBA among all the teams, we could collect the three-point category and observe if there is a correlation between it and winning. One way to see that is to find top 5 winning teams in NBA and analyze the three-point portion among all the shots. Another way is to find a team and group their winning games together first, and then group the losing games together to analyze the relation between three-point and winning/losing the game.

Implementation details and Process

Please see the ipython notebook in the same folder as "Data Statistics"

Future

There is will be info for this section after discussing with our TF, Harry during his office hour.