

Michael Ventoso

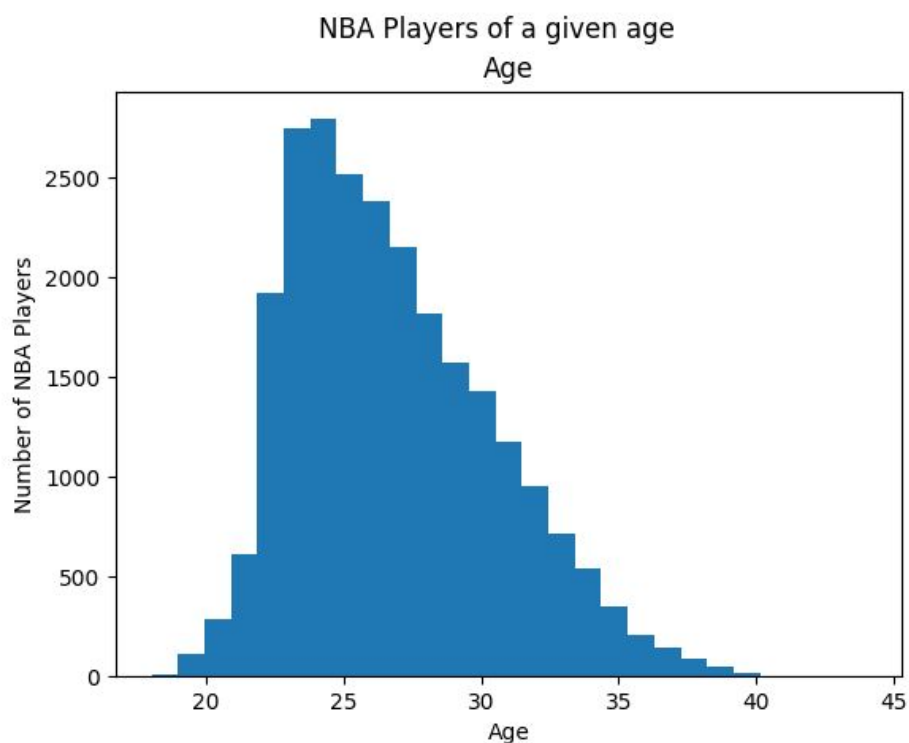
<MichaelVentoso@Gmail.com>

ProtAtOnce Interview Phase 3

Part A: NBA Dataset

For this part of the assignment, I wanted to focus on the players' ages as a central idea that would tie the project together. This also allowed me to use my data exploration as a chance to look at what sort of hypotheses I might draw from the data. I also decided to use Pandas Dataframes throughout these assignments, as they work well with large datasets, dealing with empty cells, and writing/reading to/from .csv files.

Data Exploration:



1. Age Histogram

My first idea for exploring this data with a focus on age was to make a histogram of how many players there were over time at each given age. After retrieving the highest and lowest recorded ages, I made the histogram with one bucket per each year (ie. a bucket

for 18 years old, 19 years old, etc.). The result is unsurprisingly a pretty normal distribution with a right-skew.

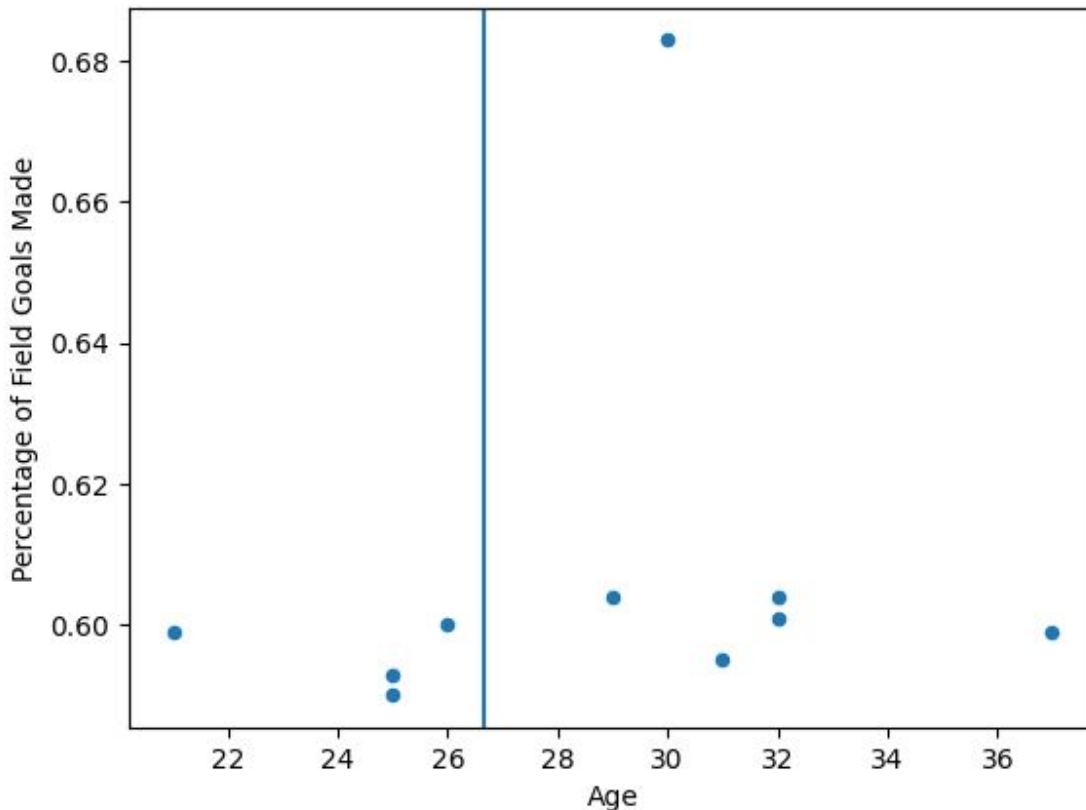
2. Team's Average Player Age

Next, I wondered if the age of players in the NBA was more evenly distributed, or if some teams had generally older or younger players. I calculated the average NBA player's age (26.66 years old) as well as the average player age for each team. Then I could look at how many teams have an average age above or equal to the NBA average (21) versus how many are lower (48). I have additionally uploaded a .txt file containing all NBA team's average player ages to the repository.

3. Ages of the Top Shooters

I then began to wonder if it is true when people comment that someone is 'too old' or 'past their prime', and aimed to analyze data on the ages of the NBA's best shots. For this analysis I focused on field goal percentage (both 2 and 3 point shots), but quickly discovered there are quite a few players who have a 100% field goal accuracy. After some quick research it appeared that anything over 50% is considered pretty good, so I figured a 100% accuracy likely means the player only attempted a couple shots in a short career. This led me to filter the data and consider a top shooter to be only someone who has spent enough time on the court to have attempted at least a thousand shots. Finally, in plotting the 10 best shooters (and showing the average age as a vertical line) it seems there is no trend. There is not a cluster of younger hot-shots as one might expect.

Top Shooters' FG% vs. Age



Statistical Hypothesis:

Looking at the graph of the top shooters vs their ages made me wonder if perhaps it only seems like there are no trends because the sample size is so small and because I am looking only at the very top of the barrel. I wondered if there could be a plausible correlation between age and field goal percentage across the entirety of the NBA.

My first step was to filter the dataset to remove outliers. This time, I figured having at least one hundred field goal attempts would be enough that nobody would have close to a 100% accuracy. I then needed to ensure the data was normally distributed. Since there were more than 5,000 samples, a Shapiro-Wilk normality test would not work, but an Anderson-Darling would be fine. A test statistic of 188.3 suggests this dataset is extremely normal, and therefore I could administer a Pearson correlation test comparing

the age of NBA players to how well they shoot. As it turns out, the correlation value is negligible at -0.004 and with a p-value of .613, suggesting there really is no correlation at all between a player's age and their ability to shoot basketballs.

