NBA Analytics

**Introduction:**

The power of modern-world technologies such as python have reshaped the way fans, reporters, players, and sports executives see and approach their respective business.

In the NBA, it wasn’t very long ago that:

* The 3-point shot was seen as inefficient and a last resort unless you were a specialty role-player.
* Executives firmly believed key to winning championships was an athletic center who could score and defend the rim.
* Players relied on an 18-foot Michael Jordanesque mid-range game that today is seen as the epitome of inefficient basketball.

**Interest:**

One of the main inspirations to pursuing the data visualization course is my admiration for the plethora of data that data analysts used to transform the NBA and other sports. Many players, commentators, and some executives deride analytics as short-sighted and do not believe they accurately measure their “intangible” qualities such as leadership and maturity.

That being said, there is zero doubt that the men and women who created meaning shot tracking heat maps and calculated the “efficiency threshold” – that is – the percentage at which a team must make 3s to make them worth more than a 2 point shot, have changed the way sports are played forever.

**Goals:**

While this visualization does not bring forward any “groundbreaking” statistical analysis or mind-blowing visualizations, it is my “beginners” approach to dashboard building while looking at surface-level stats to find meaning. This is a foundational approach to a pursuit of advanced sports analytics.

The first step I decided to take is to look at two polarizing players – James Harden and LeBron James. My aim is to scrape their game logs and piece them together in different fashions to observe any trends.

I scraped their game logs for the 2017, 2018, and 2019 seasons and aggregated their “basic” statistics into 3-year averages – points, rebounds, assists, minutes played, free throw, 2-point, and 3-point attempts as well as makes. There are more statistics in the dataframes I mined, but the primary focus was points, rebounds, and assists.

**Process, tools used:**

I used a combination of web-scraping via beautiful soup to scrape the game logs from basketball-reference.com, then put the data together into a pandas dataframe. The dataset is small – only 169 rows, but I did not want to be overwhelmed by performing an overarching amount of analysis. My primary aim here was to observe some surface-level trends while creating a somewhat smooth first-time dashboard.

Once I had my data and aggregated it appropriately, I bucketed it into two trends I wanted to explore: how the two players trended across time (via months) through the variance in the total of their points + rebounds + assists. The second trend I wanted to look at was how do they perform against specific teams? I grouped all of their games by opponent and once-again averaged out the total of their points, rebounds, and assists versus each opponent they faced.

The visualizations described above aren’t exactly “non-standard” but in my mind that is okay. Again, my primary aim here was to gain comfort with the visualization tools at hand so I could take the next step into deep, meaningful analysis.

The third visualization I did for each player was a heat-map of their plus-minus ranking versus each team across the 2017-2019 seasons. In NBA circles, the plus-minus rating (how many points you are either outscoring the opponent or being outscored by the opponent while you are on the floor) can be a debatable metric. Some say that it is more team-oriented and reflects poorly on superstar players. Others say it provides meaningful insight into a subset of players that are marked as poor defenders.

The data visualizations leveraged the following javascript libraries:

-D3

-Leaflet

-Anychart

**Challenges/Limitations:**

The primary…”all-skills-equal” (meaning if you are as slick as java deployment as you are at pandas analysis) challenge I faced was naturally with the data. Injuries are part of the game and some null values in the dataset really hampered a smooth processing of the data. In addition, basketball-reference game logs both refer to average points/rebounds/assists and total points/rebounds/assists using the same key! This was sort of new to me. I found a quick shortcut simply by writing the impacted dataframe to a CSV, then immediately reloading the CSV as a dataframe. PANDAs will update the dataframe with a “.1” on the right-most column.

That being said, I only spent about 1-2 days gathering the data I needed. The dashboard? That took me well over a week to get right. Including a 5-day marathon from 9AM Tuesday to 7am Saturday that saw me sleep only about 4-5 hours total during that timeframe. I’m not kidding when I say it was only yesterday, around 2:30AM Friday morning that I seriously contemplated just not showing up here today and for the rest of the semester and calling it quits. In that moment I knew I had made progress, but I still had no visualizations, a clunky interface, and little hope of finishing by the 10AM deadline.

**Conclusions and next steps**

Overall, I personally am very happy with the lightyears of progress I made with integrating flask, javascript, and server-side hosts. Now that I have what I feel is a dependable foundation, I can move into deeper analysis and use more advanced techniques.

It wasn’t too long ago that looking at all this stuff was like trying to decode the Matrix. Now that I can sort of understand it, I am excited both for the future of my own visualizations as well as eager to see what everyone else has to present today.