Analysing and Predicting the Peak Age of NBA players – a Data Science Project

Marcus Chua

June 1, 2020

Why I started:

The National Basketball Association is the top basketball league in the world, a stage where the most talented players around the world showcase their skills and fight for accolades. Being one of the many million fans of this league, I have watched sport analysts debate non-stop as they compare the teams and players. A major part of most debates is how to determine when a certain player has past his prime. While some players have found their best years to be before the age of 25, others have gone on to peak much later in their careers past the age of 30.

Using a variety of scraping, visualisation and modelling tools in Python, this study aims to determine when these athletes have reached their peaks and spot any trends in the peaks of retired/existing players. Thereafter, a model can be developed and used to predict the peak age of the players who have recently entered the league.

The methods and findings of this study would be of interest to any party who shares the love for this sport.

PART 1

Why I made my research different, and how:

Before embarking on my research, I looked through several existing publications of studies on NBA players' peaks. Most studies found the peak of players by studying their ages when they were selected for All-Star games or as the top players in that year. However, these accolades are given based on fan favorites and opinions of panelists. I decided to decipher the peaks of players using a composite indicator known as Player Efficiency Rating (PER). It is a per-minute rating developed by ESPN.com columnist John Hollinger, and widely regarded the go-to statistic to determine player performance.

In addition, I extended my research to cover every single player who played at least 5 years in the league and completed at least half of the games in these seasons. In order to increase the scope and rigour of my research, I looked for and included the data of all such players from 1950 to 2019.

Most importantly, I planned to perform analysis on the peak ages based on position and era of players. This were categories which I felt played a huge part in determining the best years of an athlete.

I used statistics which I could find from basketball-reference.com, the leading website for advanced and comprehensive basketball data. To do this, I made use of a client called basketball-reference-web-scraper

to scrap the data. The scraping enabled me to find the advanced statistics lines of each player within a certain season.

After thorough cleaning and filtering of the columns of data which I required for my study, the output of this (extremely long) loop process was 68 tables (one for each NBA season), with each table looking like this:

	playerID	name	position	age	PER	usage%	gp	yr
0	abdursh01	Shareef Abdur-Rahim	POWER FORWARD	27	21.2	24.8	53	2004
1	abdursh01	Shareef Abdur-Rahim	POWER FORWARD	27	16.5	23.3	32	2004
2	allenma01	Malik Allen	POWER FORWARD	25	10.5	18.3	45	2004
3	allenra02	Ray Allen	SHOOTING GUARD	28	21.7	27.8	56	2004
4	alstora01	Rafer Alston	POINT GUARD	27	13.7	17.3	82	2004

This sample data was taken from the year 2004. Immediately, a huge problem with the dataset was spotted – there was more than one row for a single player. This was because of mid-season transfers, where a player played for two or more teams in the same season. I aggregated the data for these cases and chose to average out the separate PER values.

I then began to collate the data I had by player, in order for me to identify their peaks. I wrote loops in order to search for and append these rows for each season into separate profiles for each player. I managed to obtain 3996 unique sets of NBA career data, each one detailing the information for a specific player. Here's an example of the data for

Kareem Abdul-Jabbar, who had arguably one of the longest and most decorated careers in the league's history:

	playerID	name	position	age	PER	usage%	gp	yr
0	abdulka01	Kareem Abdul-Jabbar	CENTER	22	22.5	0.0	82	1970
1	abdulka01	Kareem Abdul-Jabbar	CENTER	23	29.0	0.0	82	1971
2	abdulka01	Kareem Abdul-Jabbar	CENTER	24	29.9	0.0	81	1972
3	abdulka01	Kareem Abdul-Jabbar	CENTER	25	28.5	0.0	76	1973
4	abdulka01	Kareem Abdul-Jabbar	CENTER	26	24.4	0.0	81	1974
5	abdulka01	Kareem Abdul-Jabbar	CENTER	27	26.4	0.0	65	1975
6	abdulka01	Kareem Abdul-Jabbar	CENTER	28	27.2	0.0	82	1976
7	abdulka01	Kareem Abdul-Jabbar	CENTER	29	27.8	0.0	82	1977
8	abdulka01	Kareem Abdul-Jabbar	CENTER	30	29.2	27.0	62	1978
9	abdulka01	Kareem Abdul-Jabbar	CENTER	31	25.5	23.3	80	1979
10	abdulka01	Kareem Abdul-Jabbar	CENTER	32	25.3	24.1	82	1980
11	abdulka01	Kareem Abdul-Jabbar	CENTER	33	25.5	26.3	80	1981
12	abdulka01	Kareem Abdul-Jabbar	CENTER	34	23.4	25.6	76	1982
13	abdulka01	Kareem Abdul-Jabbar	CENTER	35	23.6	24.6	79	1983
14	abdulka01	Kareem Abdul-Jabbar	CENTER	36	21.3	25.1	80	1984
15	abdulka01	Kareem Abdul-Jabbar	CENTER	37	22.9	24.3	79	1985
16	abdulka01	Kareem Abdul-Jabbar	CENTER	38	22.7	26.6	79	1986
17	abdulka01	Kareem Abdul-Jabbar	CENTER	39	17.9	22.1	78	1987
18	abdulka01	Kareem Abdul-Jabbar	CENTER	40	15.8	21.4	80	1988
19	abdulka01	Kareem Abdul-Jabbar	CENTER	41	12.9	20.1	74	1989

Now, I could then proceed to identify the peak of each player's performance. I sorted the seasons in each player's career by PER (descending). And then I took out the top year from this arrangement.

At this point, I also took the opportunity to remove any players who had less than 5 years of playing experience and nulled the records of

players who had played less than half of any given season. This was to ensure that the peaks of each player was correctly identified, and players who barely featured in the league will not be part of the analysis.

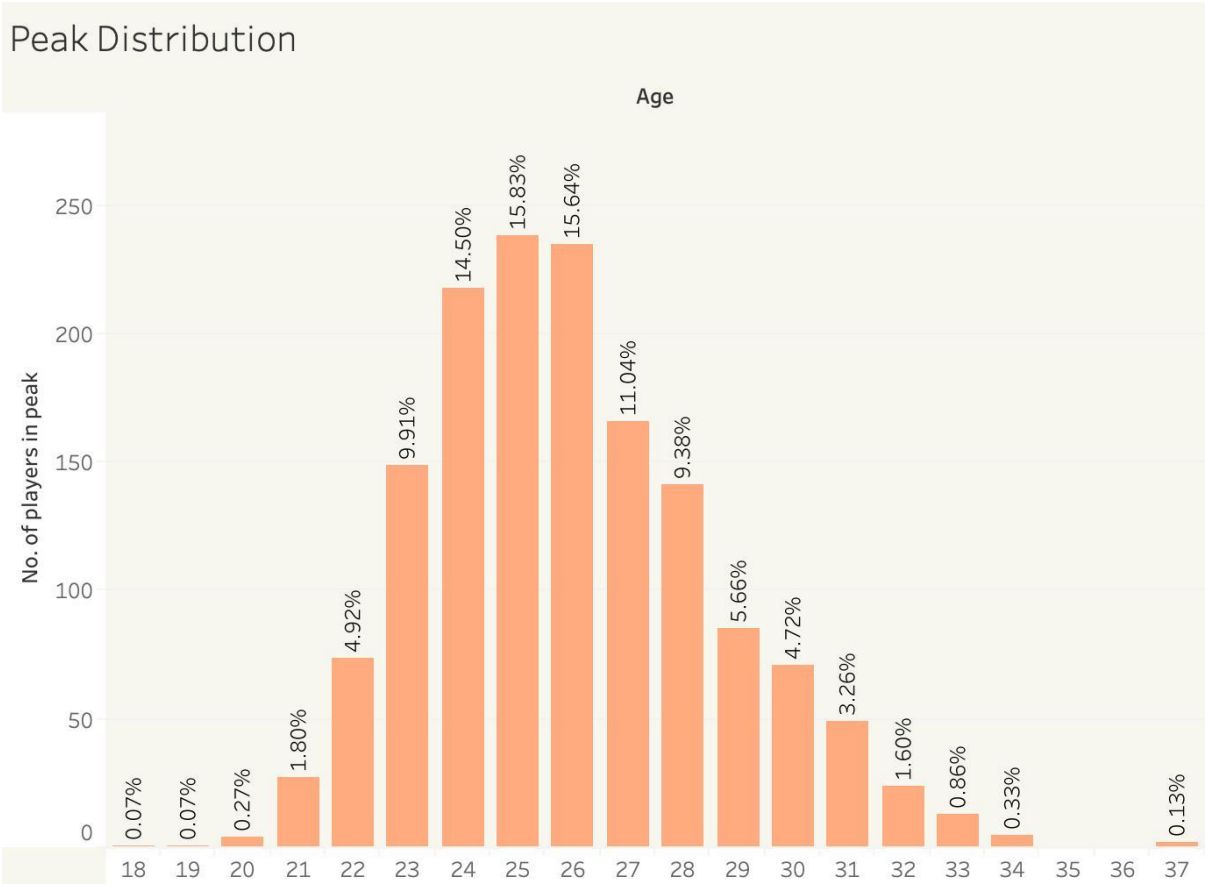
Finally, I rounded off the data collation here by finding each player's first year in the league. This will help us compare the players by decade in later parts of the study. This is an example of the peak year data for Kareem:

	playerID	name	position	age	PER	usage%	gp	yr	draftyr
2	abdulka01	Kareem Abdul-Jabbar	CENTER	24	29.9	0.0	81	1972	1970

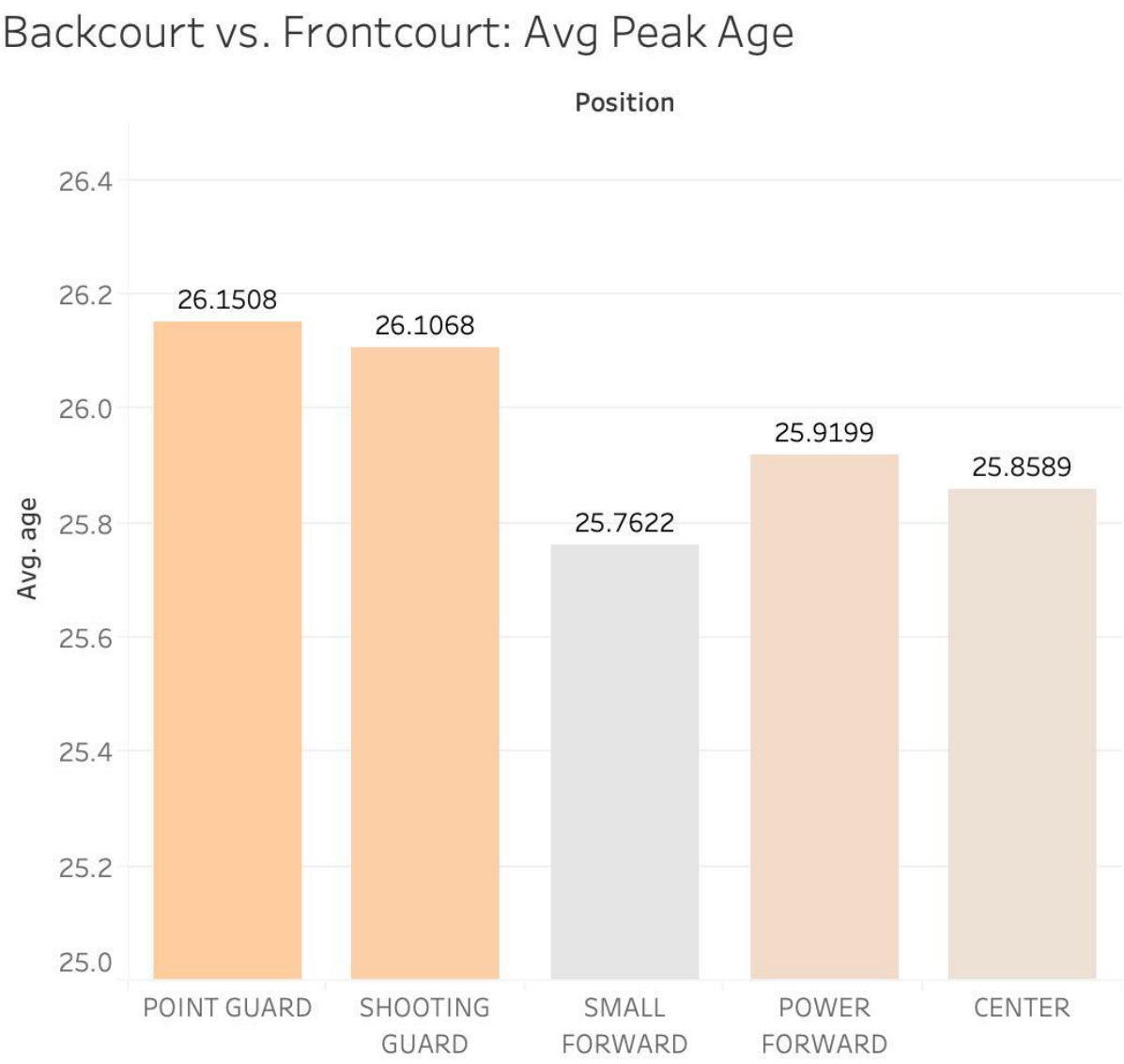
I took all the peak year data of different players and concatenated them into a large table. At this point, I had the data of 1503 players, with their peak age identified. Taking my research away from Python, I then used Tableau to visualise my findings.

What I found:

The average peak age of an NBA player in my study was 25.959 years. A player was able to be the most efficient on the court and show the most of his capabilities around this age. To analyse the trend, I looked at the distribution of peak ages:

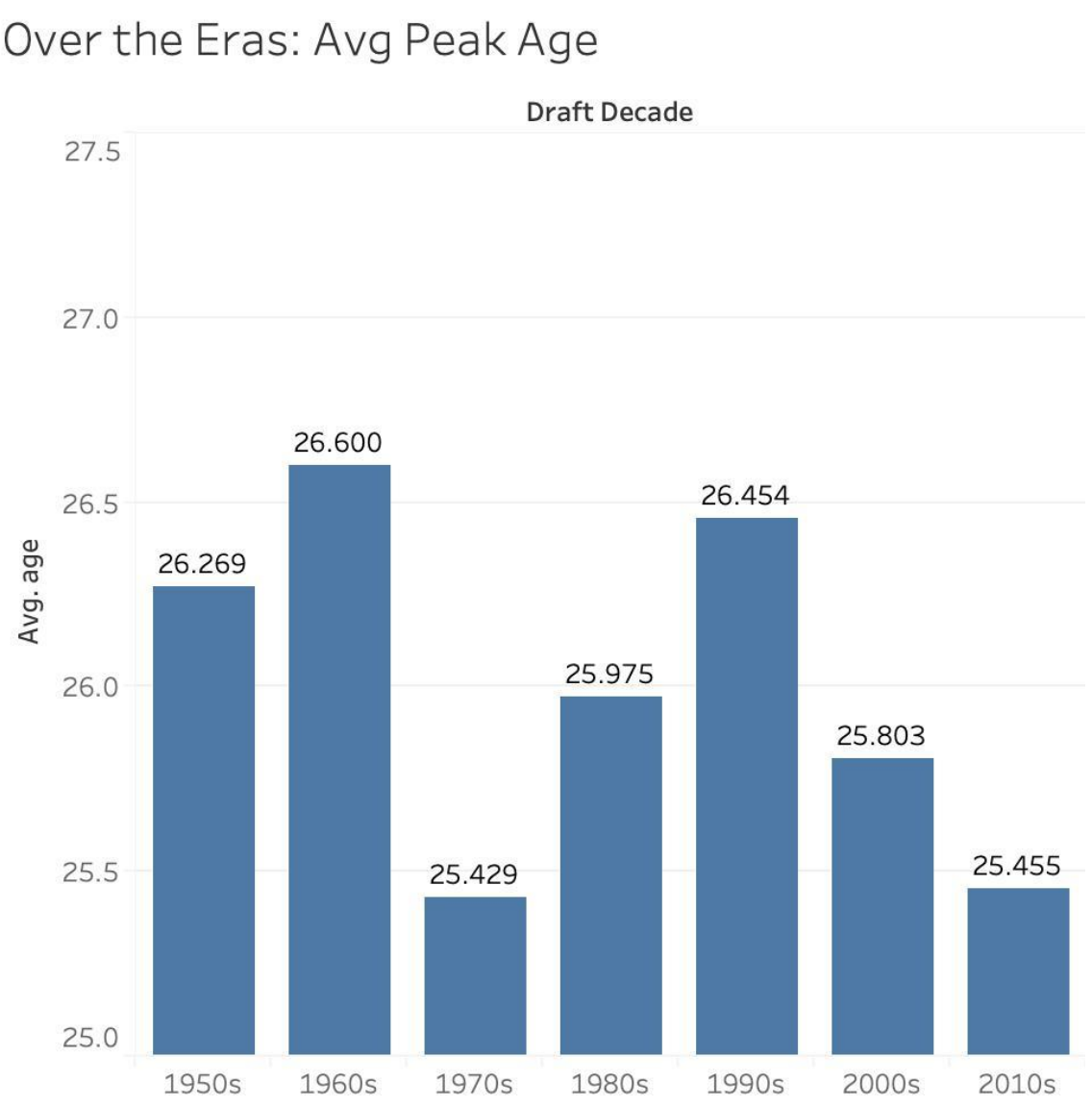


As expected, the distribution was bell-shaped with most players peaking at 25 or 26 years of age. Dissecting this information further, I classified the players by their positions. The results were rather insightful:



For the 5 positions in basketball, they were divided into 2 distinct groups by average peak age. While the guards tend to peak after 26, those in the front court found their best years to be on the other side of 26. This seems logical, as the guard positions are relatively more skill-based and usually result in less gruelling physical exertion than the forward and centre positions.

Moving on, I grouped the players by the decade in which they were drafted to see if there were any trends.



Players in the 50s, 60s and 90s peaked way later than those in other decades. While I am unable to find a definitive explanation for this trend, this could perhaps be due to the style of gameplay popularised in the different eras.

In the 1970s, the NBA and ABA merged. As the ABA players joined NBA teams, they brought their fast-paced and high-scoring style into the NBA. This required the players to be way fitter and nimble, which is why players would have done better in their early days.

In the 1980s to the 1990s, the pace of play in the league slowed down. The league was dominated by 'Showtime' in the 80s, and then by big, physical presences in the 90s – for example, Hakeem Olajuwon, Shaquille O'Neal and Patrick Ewing. These influences resulted in the game pace slowing down tremendously. The average NBA player was thus able to operate at a pace which did not take a huge toll on his body, and the peak age of a player rose tremendously.

As the NBA moved into the 2000s, isolation play gradually made way for selfless and high-intensity sequences of team play. This style of play involves every single player in the line-up fairly equally, and demands that each player is able to run the floor tirelessly. As such, the average peak age of players have started to fall again. A small number of playmakers on each team may peak later on in their careers as they pass the ball more efficiently with experience, but majority of the players who support them will need to be youthful and fit to keep up with the pace.

In all, I gained great insight on the effect of position and era on the average peak age of players. The table below provides a summary of all the data discussed.

Average Peak Age By Position and Era

	Draft Decade							
Position	1950s	1960s	1970s	1980s	1990s	2000s	2010s	Grand T..
POINT GUARD	27.267	27.000	25.565	25.534	26.825	26.208	25.581	**26.151**
SHOOTING GUARD	26.800	26.750	25.778	26.184	26.768	25.781	25.139	**26.107**
SMALL FORWARD	25.308	26.393	25.432	25.667	26.357	25.345	25.727	**25.762**
POWER FORWARD	26.462	27.519	24.794	26.233	26.478	25.429	24.938	**25.920**
CENTER	25.353	25.400	25.481	26.273	25.725	26.131	25.886	**25.859**
Grand Total	**26.269**	**26.600**	**25.429**	**25.975**	**26.454**	**25.803**	**25.455**	25.959

Average of age broken down by Draft Decade vs. Position.

Can we predict the peak age of players?

With the data that we had in our hands, I felt that it was only fitting that a model could be devised to determine the peaks of up-and-coming players who had just set foot in the NBA. Using the information which we had gathered about the 1500+ players who consistently featured in the league since 1952, we could train a model of choice to predict peak ages with accuracy! However, some research had to be done on the determinants of peak age. And so, Part 2 of my research begun.

PART 2

What factors determine the peak age of players?

Based on simple research and leafing through published reports, I came up with a list of factors to get my research going:

- Physical attributes (Height, Weight)
- Playing position

- Age entering league
- Draft year
- Average minutes per season (first 3 seasons)
- Average PER per season (first 3 seasons)

These factors were not meant to form an exhaustive list, but they were variables which were documented and easily obtainable. Each of these factors had clear links to the peak age of a player too. Another factor which I considered including was the college which players played for before joining the league. I would be determining the inclusion of this factor in my model in a later part.

Where could I get the data to start?

I repeated many steps which I had taken for Part 1 of the project. In terms of data, I reused the basketball-reference-web-scraper client to retrieve advanced statistical data for each player by season. However, I needed additional data this time. As there wasn't a complete dataset on the height and weight of each player and I could not find a suitable web client to scrape this data, I used a combination of two datasets which I found on Kaggle. These datasets spanned different time periods and combined to give me the height, weight and college data for players from 1952-2019. Finally, I obtained the peak age data from Part 1 of my research.

How did I begin?

First, I began with the advanced statistics from basketball-reference.com. The data was cleaned in similar fashion to Part 1, but this time including a column showing the minutes played (mp) per season. The mean PER and MP for each player was also calculated before each player's data was concatenated into a table. After which, I then proceeded to join the peak age information from my findings in Part 1.

After appending the peak age data, I then proceeded to join the height, weight and college columns from the two Kaggle datasets. These datasets had to be cleaned numerous times before they could be implemented. To complete the collating of information, I had to merge the height, weight and college columns from both datasets to ensure that each player had a maximum of one data value for each of these fields. The final dataframe of 4011 players looked like this:

	playerID	name	position	firstage	avgfirst3PER	avgfirst3MP	firstyr	peakage	height	weight	college
4006	welshth01	Thomas Welsh	CENTER	22	16.1	36.0	2019	NaN	213.36	115.665960	UCLA
4007	willijo04	Johnathan Williams	CENTER	23	15.1	372.0	2019	NaN	205.74	103.418976	Gonzaga
4008	willike04	Kenrich Williams	SMALL FORWARD	24	9.7	1079.0	2019	NaN	200.66	95.254320	None
4009	williro04	Robert Williams	CENTER	21	18.8	283.0	2019	NaN	NaN	NaN	NaN
4010	youngtr01	Trae Young	POINT GUARD	20	17.0	2503.0	2019	NaN	187.96	81.646560	Oklahoma

These players were at the tail end of the dataset, hence they did not have their peak ages identified yet. I did a quick check on other popular players, just to make sure that they already had their peak ages logged into the dataframe.

How did I prepare my data for modelling?

I split the players into 2 groups: my 'data set' of players who already had their peak ages identified in Part 1, and the 'prediction set' which had newer players whose peak ages I wanted to predict. The newer players were defined as those who were drafted into the league less than 3 seasons ago (before the start of the 2016/2017 season).

For players who had no height or weight data, I filled the NBA averages of 200 cm and 98 kg into their respective columns. I used one-hot encoding to assign binary values for the string data which were in the position and college columns. After removing the unnecessary columns like name, and player ID, I had 1500 anonymous rows of player data with peak ages identified, ready to be split into my training and testing data.

Using preprocessing tools from sklearn, I fit and transformed the data then split the rows into training and testing sets. The ratio was 4 to 1. Now, we could apply machine learning to the data.

What machine learning tools were used?

I tried to apply 5 different models to split the data, namely:

- Clustering (K Nearest Neighbours)
- Decision Tree Classifier
- Support Vector Machine

- Logistic Regression

- Multilinear Regression

My initial results were far off the mark, which led me to rethink my factors. I had an intuition that there were too many different types of colleges being assessed in the modelling which may have cluttered the machine learning, hence I removed the college as a factor and rewinded the steps. Sure enough, the accuracy of the predictions increased greatly. Using metrics from sklearn, I identified clustering as the most accurate method of the 5 for prediction of peak age.

How were the predictions like?

To round the study up, I cleaned the prediction set just as how I did for the train/test set, and applied the clustering technique to predict the peak ages. The raw data looked like this:

```
array([26, 26, 26, 24, 23, 24, 29, 23, 27, 31, 23, 26, 24, 23, 25, 24, 25,
       24, 23, 28, 25, 23, 25, 23, 24, 27, 25, 26, 28, 26, 26, 23, 24, 24,
       31, 26, 22, 25, 25, 23, 24, 22, 26, 24, 25, 23, 25, 24, 23, 23, 26,
       26, 24, 26, 24, 23, 28, 24, 24, 24, 24, 25, 25, 25, 27, 23, 23, 25,
       26, 25, 23, 26, 24, 23, 24, 24, 22, 24, 24, 26, 26, 25, 24, 24, 28,
       24, 23, 26, 26, 26, 25, 25, 28, 25, 26, 26, 25, 24, 24, 26, 28, 23,
       29, 24, 28, 24, 26, 25, 26, 23, 28, 21, 26, 26, 27, 25, 26, 24, 28,
       23, 28, 26, 24, 25, 22, 30, 28, 25, 28, 24, 26, 27, 28, 29, 23, 28,
       26, 26, 30, 24, 27, 24, 24, 27, 28, 24, 23, 28, 23, 24, 26, 26, 23,
       25, 26, 28, 25, 26, 25, 27, 27, 24, 24, 26, 23, 23, 26, 23, 27, 25,
       25, 27, 23, 25, 22, 26, 30, 24, 29, 26, 25, 23, 26, 26, 24, 24, 25,
       25, 25, 23, 23, 25, 24, 26, 22, 22, 23, 26, 27, 25, 29, 26, 26, 24,
       26, 23, 24, 26, 24, 28, 24, 28, 24, 26, 25, 26, 24, 26, 24, 25, 21,
       28, 25, 21, 30, 24, 23, 26, 25, 24, 23, 25, 24, 26, 25, 24, 26, 25,
       25, 23, 22, 28, 24, 25, 24, 22, 28, 25, 24, 24, 28, 26, 24, 26, 28,
       22, 24, 26, 24, 25, 23, 24, 25, 25, 24, 24, 24, 29, 25, 28, 24, 23,
       28, 26, 28, 24, 24, 24, 25, 24, 26, 24, 24, 27, 24, 28, 26, 27, 22,
       24, 30, 28, 25, 22, 21, 25, 24, 26, 25, 28, 23, 24, 24, 28, 26, 24,
       25, 26, 26, 26, 22])
```

Of course, I had to insert this data back into the prediction set to make the information relevant. A sample portion of my results looked like this:

playerID	name	position	firstage	avgfirst3PER	avgfirst3MP	firstyr	peakage	height	weight
abrinal01	Álex Abrines	SHOOTING GUARD	23	8.466667	925.666667	2017	26	NaN	NaN
bakerro01	Ron Baker	SHOOTING GUARD	23	6.000000	464.666667	2017	26	193.02	99.395120
baldwwa01	Wade Baldwin	POINT GUARD	20	8.133333	193.000000	2017	26	193.00	91.000000
beaslma01	Malik Beasley	SHOOTING GUARD	20	12.300000	875.666667	2017	24	195.79	88.452016
bembrde01	DeAndre' Bembry	SMALL FORWARD	22	9.066667	919.000000	2017	23	198.06	95.127160
...	...	...	...	...	...	...	...	...	...
welshth01	Thomas Welsh	CENTER	22	16.100000	36.000000	2019	25	213.36	115.665960
willijo04	Johnathan Williams	CENTER	23	15.100000	372.000000	2019	26	205.74	103.418976
willike04	Kenrich Williams	SMALL FORWARD	24	9.700000	1079.000000	2019	26	200.66	95.254320

The full dataset can be accessed as a CSV file on Github.

Closing Remarks

I was compelled to finish my investigation to find the correlation that career peaks shared with a player's era and position. While I planned my analysis very carefully, some limitations do exist.

A possible limitation in Part 1 is the accuracy of determining a player's peak using PER. Like every statistic, there is never one perfect measure and PER does have its own shortcomings too. Another limitation would be the possibility of career-changing injuries. Some irreversible damage to a player's physicality could result in him never being able to hit his true potential, which could affect our research. Given the physical nature of this sport, major injuries have been common and have affected countless players which I had included in my study. Finally, players who do reinvent themselves or undergo a positional change may be able to discover a 'second peak' in the

twilight years of their career. For example, LeBron James has become a guard and is increasingly efficient as he ages well past his prime years when he was a small forward.

Using machine learning, I ran the risk of omitting many key factors which could affect a player's peak. Again, the list of factors which I had compiled was by no means exhaustive. Further project extensions could involve me adding or removing these variables from the model to better predict the peak ages.

All in all, I thoroughly enjoyed myself throughout the course of this project.

References:

1. Cirtautas, J. (2020, March 8). NBA Players. Retrieved from https://www.kaggle.com/justinas/nba-players-data/data#
2. Goldstein, O. (2018, April 27). NBA Players stats since 1950. Retrieved from https://www.kaggle.com/drgilermo/nba-players-stats
3. Bradley, J. (n.d.). basketball-reference-web-scraper. Retrieved from https://pypi.org/project/basketball-reference-web-scraper/
4. Kareem Abdul-Jabbar Stats. (n.d.). Retrieved from https://www.basketball-reference.com/players/a/abdulka01.html
5. Calculating PER. (n.d.). Retrieved from https://www.basketball-reference.com/about/per.html