Nurbek: Contextualization and motivation, Clarity of research questions and hypotheses
Mihaly: Interpretation of results and appropriateness of drawn conclusions
Asset: Description and justification of data and methods (including assumption checks)
Identification of limitations and opportunities for future work

-

PREDICTING THE HEIGHT OF NBA PLAYERS BASED ON PERFORMANCE AND CLUSTERING THEM

Asset Kabdula, Mihaly Hanics, Nurbek Bektursyn
Central European University
Quellenstrasse 51, 1100 Vienna, Austria

ABSTRACT

The National Basketball Association hosts the most prestigious basketball league in the world: the NBA. Getting into the NBA is super hard even compared to other sports due to being almost totally limited to North America and the "incestuous" college draft system, and the players who get to join are already special, "chosen" players. Watching some stars shine in games, we can see that indeed players tend to be more singular and rely much more on their talent, than in soccer where you are expected to train and rely on a much broader skillset, having "more average" statistics. In this project, the aim was to find what is the correlation between a player's physical attributes and performance, enhancing making predictions (for example, with machine learning models, or regression) on one off the other. In particular, we used statistical methods to show the correlation between height, age of players and their performance statistics, and their "behavior" on the pitch. Does the habit of shooting from a far range typically mean that the player is small (as smaller players may tend to rely on their shooting skills, rather than powerfully scoring a point from close range)?
We work with the sub-official NBA API and fetched our data with it from the NBA website. Our methods include classical statistical tools like hypothesis testing, ANOVA, and longitudinal analysis. For further research, analysis of weight correlations should be made too, and we could extend the performance statistics used.

1 INTRODUCTION

Given both career and seasonal performance statistics of an NBA player, can we predict his height? It would be interesting to see how much we could tell about a player's physical attributes from performance, playstyle.
Some related work on this dataset has been done by Tal Boger, a Ph.D. student at John Hopkins University, founder of Dribble Analytics, he used machine learning and statistical algorithms (mostly LDA and Logistic Regression) on this dataset to predict the MVP each year, predict best defenders, comparing height's and wingspan's correlation with and so on.
Like Tal, we also worked with the sub-official NBA API. It gathers information from the nba.com website, fetching data through various player and team stats endpoints. This gave me enough information to work with, and build a model that 49.5% of the times can predict the height of a player with at most 1-inch error. This ensemble model, and other models can be later used to show that there is a playstyle and physical attribute correlation in the NBA.

2 DATA

The data is entirely fetched from the official NBA website [1]. The fetching process and code are available in the file *fetch_players.ipynb*, and the stored data is available in 4 files: *career.csv, career_filtered.csv, player_bios.csv, player_bios3.csv*. The tool used to fetch the data is the nba-api package, available on pip. Using this, we fetched two types of data: player career stats, using the *PlayerCareerStats* endpoint, and seasonal player data including biographical information through the LeagueDashPlayerBioStats endpoint. The fetching was done on all players who have played in the previous 20 seasons, fetching all seasonal data from 2003-04 till 2022-23. All information about gathering the data is available in the respective notebook.

2.1 Data description

The seasonal performance data from 2003-04 till 2022-23 is contained in career_filtered.csv, which has 9778 instances of player-season data, with 27 attributes including age, team, education, country, minutes and games played, and performance statistics, including points, shots scored and attempted, defensive stats (this set of data does not contain height and weight as an attribute yet). The most important

variables from this dataset are the number of offensive and defensive rebounds, which is not available in the other dataset.
The player bio dataset contains more biographical data of each player, but also seasonal. It contains height, weight, draft year, round and number, and more general statistics such as net rating, usage rate. For each season, the players playing in that season are included in the seasonal data, and we collected this data for all seasons between 2003-04 and 2022-23. The number of attributes is 24, but most attributes overlap with the other dataset. After combining the two sources and filtering for most relevant attributes, combining multiple one season stats for each player (e.g., player has played for two teams in one season) excluding players who played less than the length of 3 matches in a season, we had 8322 instances with 14 attributes. An instance:

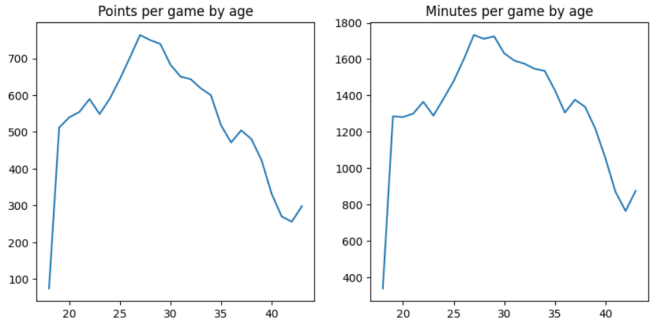
PLAYER_ID	SEASON_ID	HEIGHT_INCH	WEIGHT
1630639	*2022-23*	*78.0*	179.0
AGE	MIN	OREB_PCT	DREB_PCT
22.0	*217*	*0.046*	*0.152*
FGA_PM	FG_PCT	FG3A_PM	FG3_PCT
0.406	*0.663*	*0.230*	*0.267*
FTM_PM	BLK_PM	PF_PM	TS_PCT
0.018	*0.0*	*0.101*	*0.589*

As you can see, because of multiple seasons data for most players, the data is actually longitudinal. We analyzed some hypotheses with longitudinal statistical methods.
Important to mention that while height was a numerical variable, its values only included integers. This may mean that we would be better off using classification statistical methods, which we did not use, we assumed the statistic to be a continuous variable.

2.2 Data preprocessing
The steps we took in preparing the data:
Career:
1) Select the relevant attributes
2) Combine multiple one-season player data into one seasonal data (the case when a player plays for 2+ teams in one season)
3) Filter out instances with less than 3 matches of gametime (144 minutes)
4) Normalize adequate attributes by minutes
Biographical:
1) Select the relevant attributes
2) After-analysis-preprocessing
3) Combine data with career data

Post-analysis-preprocessing:
We found while looking at attribute-by-age plots that “too” young and “too” old players are outliers, for which we filtered them. We also found that very heavy players are sub-outliers too, but we did not filter them, as we did not use weight in any analysis.

Normalization:
A visual proof that we have to normalize some stats by minutes:



We normalized most statistics by minutes, except the ones where it would not make sense (percentages e.g. for accuracy).

2.3 Used data for analysis
A description and explanation of variables used for analysis:
- MIN: minutes, how many minutes a player played in a season. All players below 3 matches (144 minutes) worth of minutes are dropped
- FG, FG3, FT: Field goal (2 point goals), field goal 3 pointers, free throws ("penalty throws", 1 point)
  M: made (scored), A: attempted, PCT: percentage of how many are scored of attempts
- OREB/DREB: Offensive/defensive rebounds, the amount of regained possessions in the opponent’s / own half
- PF: Personal fouls (negative statistic)
- PTS: Points scored

Bio:
- TS_PCT: True shooting percentage. General measure for how well a player shoots.

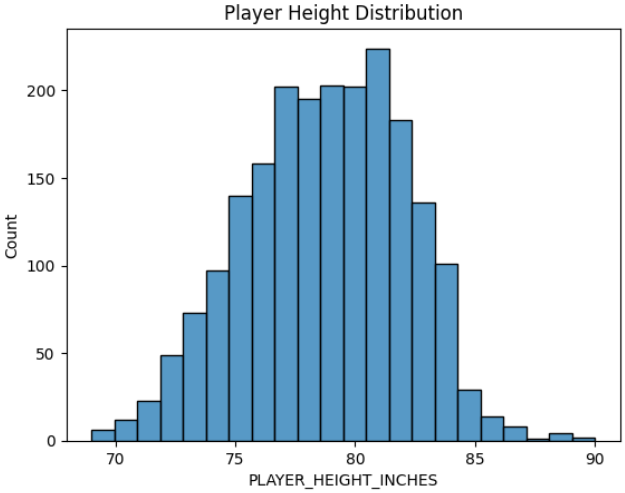
Player age is also a statistic, and it is an interesting one.

A question arises: if we have longitudinal data, data of players from different seasons, and our task isn’t typically something related to time or trends (like how a player’s career develops) but to analyze something constant throughout the time interval (career start and end), should we use approaches designed for longitudinal data, should we combine all season data for a player into one averaged career-average data, or keep it as it is? We choose the last

option. The argument against combining the data into one career-long average is that this way, we would have much less instances and more importantly much less diversity in stats. When combining, we'd have only around 1200 instances to analyze. This would mean that we have to oversample our instances. Oversampling is a dangerous tool as it creates bias, and with these limited instances of many attributes, we would advise against it. The little "variance" of player's stats over multiple seasons gives in fact a more general outlook than averaging and resampling the same data about as many times as many seasons the player played.
As a tradeoff, our models "oversample" players who have played multiple seasons, the rate of oversampling linear to how many seasons they played between 2003 and 2023. But from our point of view, even if this'd mean less accuracy on players who "come and go" from the NBA, it makes the analysis better suitable on players who have the ability to stay in the NBA for many seasons, who are "more interesting". This may even be more important than a more "fair" data analysis.

2.4 Bootstrapping

It is not uncommon that a statistical method (such as ANOVA) requires normally distributed data to use the method. For these cases, we created oversampled datasets with bootstrapping, e.g. to have a dataset where heights are more normally distributed.



The height distribution looks like it could be a candidate for a normal distribution, but after running a Kolmogorov-Smirnov test, we found that we cannot accept the hypothesis that this is normally distributed (p value less than 10^{-11}). The oversampled dataset passed this test (with p around 0.74).

3. HYPOTHESES AND METHODS USED

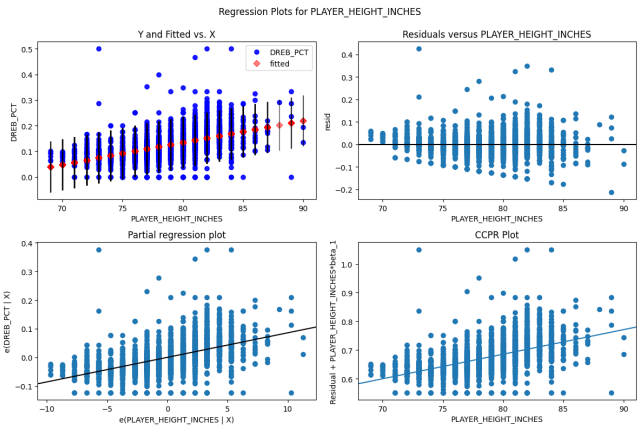
Our analyses can be split into categories in multiple ways: defensive statistics analyses, gametime correlations, and scoring analyses. Out of these, the analysis of personal fouls (per minute) and TS% over years of experience, and scores of players above age 30 are longitudinal analyses and we look at how these statistics change for a player over time. For other analyses, that is height and rebound correlation, 3-pointer accuracy and minutes played correlation, and age and minutes played correlation, we used cross-sectional data.

3.1 Defensive statistics relationships with height and age

Since players taking over more defensive roles seem to be taller and on average more experienced just by looking at big teams with our eyes, it'd be interesting to see if we see such phenomena generally too, using our data. Since height is only stored as integer values, we can classify players easily based on height.
Unlike for most other analyses, we use "per minute" normalized data here.

Hypothesis 1: Taller players collect more rebounds

To explore the truth of this hypothesis, firstly we used a one-way ANOVA to see if among all height values (remember: the set of height (in inches) values just includes integer numbers, those between 69 and 89, why we can do this), the mean defensive rebounds per minute differs from other values. The null hypothesis was that for all height values, the mean is the same. We use an oversampled (on personal fouls) dataset here, to artificially make the dependent variable normally distributed, which is a requirement for ANOVA. The test came out with a p-value less than 10^{-100}, so we can safely reject it, and start looking for what correlation these attributes have.



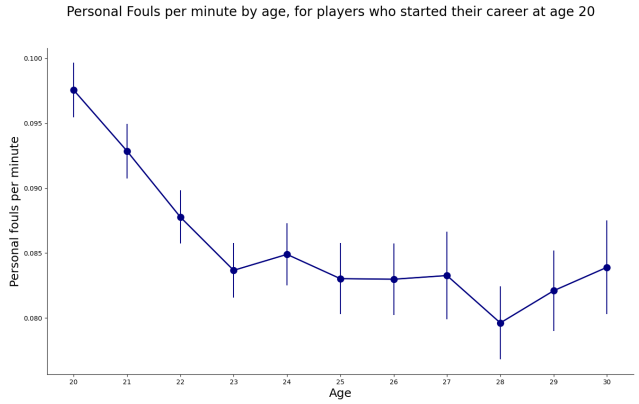
Since we ran ANOVA which is a regression in disguise, we might as well try OLS, but on the original dataset. We see on the graph above the nice gradual increase of defensive rebound percentage as we look at taller players. The results came out with R-squared value of 26%, which is not bad, and the Pearson correlation being 0.51, thus there is a correlation between the two. Since the value is positive (or just looking at the graph), indeed we can say that bigger height means bigger defensive rebound percentage. The exact step size is 0.86, meaning for every inch we "add", the defensive rebound percentage increases value by 0.86.

Hypothesis 2: Players commit less fouls as they age

To start our longitudinal analysis journey, firstly, we will look at whether players make fewer fouling mistakes the older and more experienced they get.

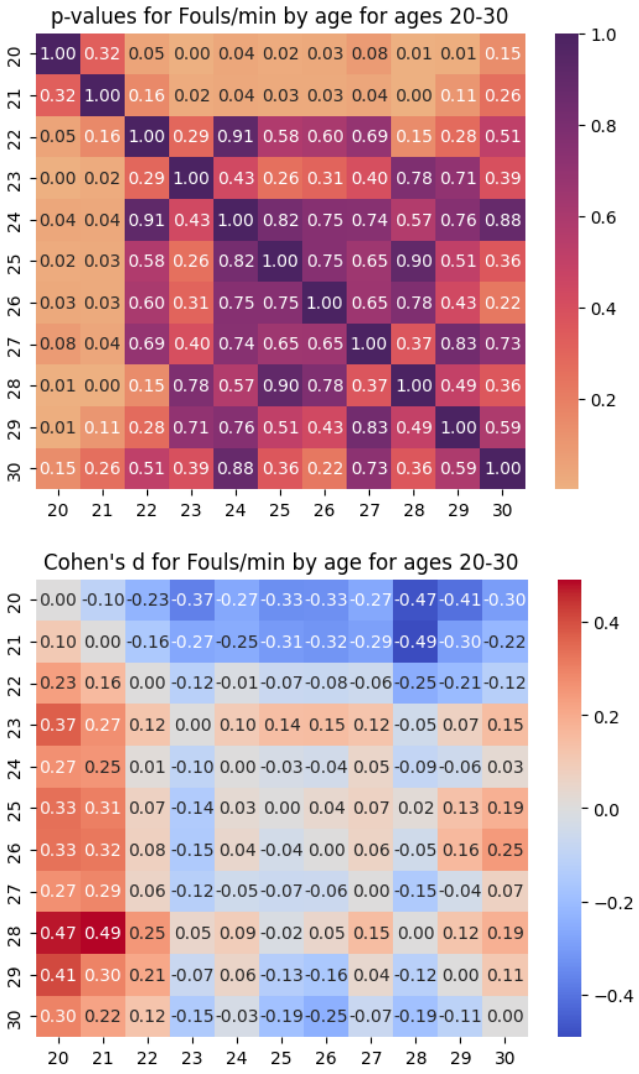
Before we get into the analysis, it is important to describe what we mean by "as they age". We looked at plainly age and performance correlations, and "experience" with performance correlation, that is performance by amount of seasons played.

What we found is that plotting average performance by age does not change that much. However, plotting some player's performance by age, we do see the trend that they improve in their 20s. How is that? From what we see, what actually matters is not how old a player is. It is about how many seasons of experience they have. Some players get drafted at 19, some get into the NBA later in their careers. Usually the "late joiners" bring down the averages. But there indeed seems to be a trend that players do better in certain statistics the more seasons they play.



To properly compare, we only looked at players who started off their careers at age 20, which is the most common age in the dataset. This way, amount of seasons played translates back to age, their 1st season is at age 20, their 5th is at 24, and so on.

We see a downward going trend, but differences are on a very small scale, so we have to run paired t-tests for all ages between 20 and 30 (years of experience between 1 and 11 seasons). We can plot the p-values and Cohen's distances on heatmaps, only looking at the upper-right triangle.

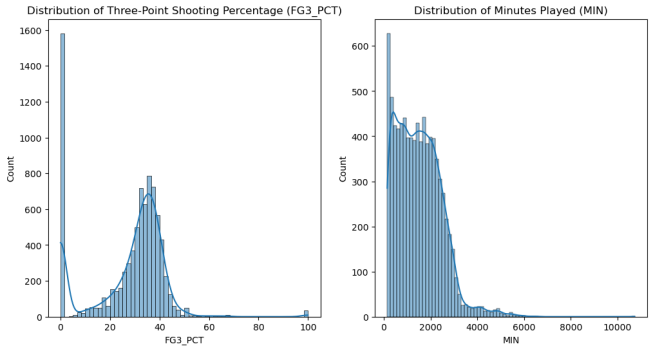


The paired t-tests show that for example, we can't reject the null hypothesis (of no difference in mean values) for the age-pairs (20,21) and (20,22), but there is a significant difference between the means for age 20 and age 23 and so on. Interestingly, we cannot statistically tell apart the means for any age combination from age 23 to 30, so seems like most of the improvement in making less personal fouls comes in the first 2-3 seasons, from that it does not really change. Looking at the Cohen's *d* values (just right upper corner), blue cells mean a negative value meaning at higher age, the mean is lower. We see that the majority of the cells are blue, meaning players tend to make fewer mistakes year by year (but mostly in the first 2-3 years).

3.2 Minutes played correlations

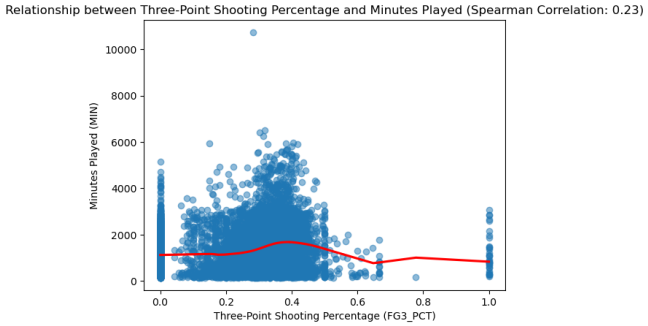
Before diving into exploring the relationship between 'Three-Point Shooting Percentage (FG3_PCT)' and 'Minutes Played (MIN)', let's first look at their distributions. The left distribution of FG3_PCT appears to be bimodal, implying two separate categories within the dataset: one clustering around 0 % and another around 30-40 %. The first category seems to be those who don't shoot three-pointers at all, which the second category seems to be more proficient in it.

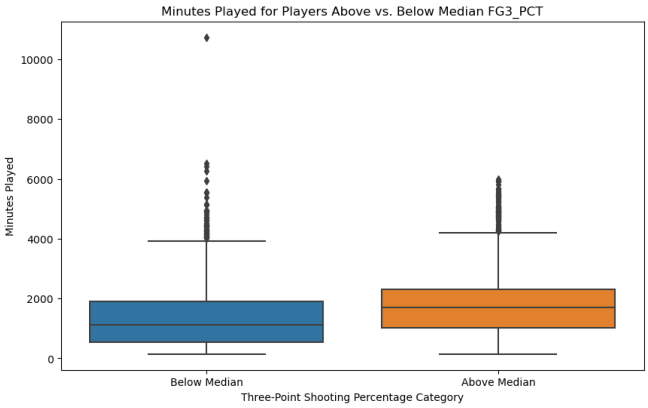
On the right, the distribution of The 'Minutes Played (MIN)' is right-skewed, indicating that a large number of basketball players have accumulated a relatively humble amount of minutes, with the frequency steadily decreasing as minutes increase.

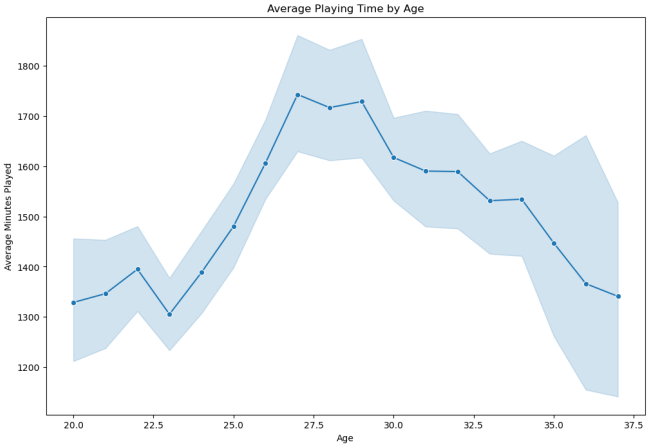


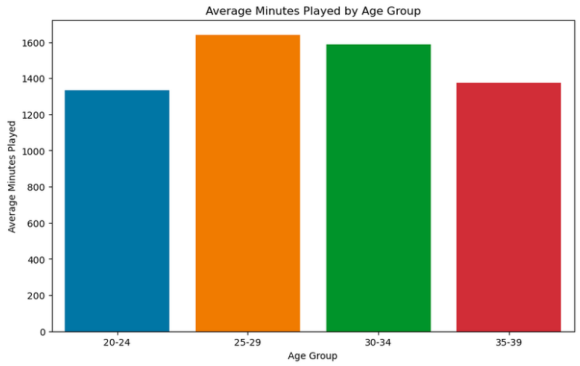
Hypothesis 3: Higher Three-Point Accuracy Correlates with More Playing Time

According to the third hypothesis, greater three-point accuracy corresponds with more playing time. The three-point shot has become a crucial component of modern basketball. Teams are increasingly relying on players who are capable of scoring outside the arc, providing additional advantage to the teams. Thus, there is a notion that coaches allow such players to play more time on the court. To test this notion, a Spearman's rank correlation analysis was performed. The scatter plot displaying the interaction between 'Three-Point Shooting Percentage (FG3_PCT)' and 'Minutes Played (MIN)' shows a moderately positive correlation, as proven by a Spearman correlation coefficient of 0.23 and a non-linear trend line that reflects the interaction's complexity. Noteworthy, the wide dispersion of data points indicates that although three-point competence is an indicator that determines playing time, it is not the only determinant factor. Outliers, especially at higher three-point accuracy levels, indicate uncommon occurrences when players may not get extra minutes regardless of their shooting competence. This visualization emphasizes the diverse decision-making process in moden professional basketball, which depend on multiple factors, including three-point shooting expertise.









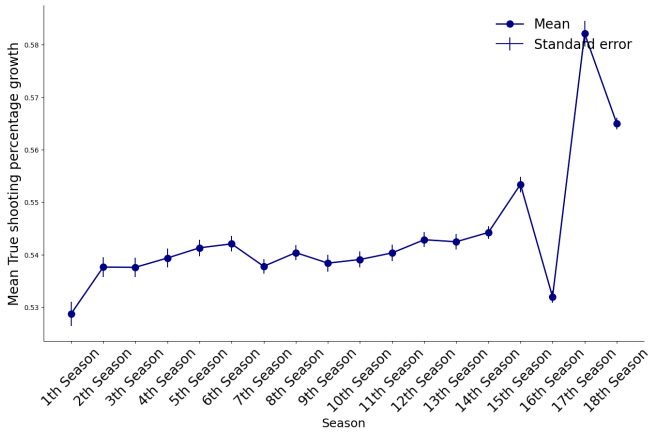
Hypothesis 4: Players' average minutes played peak within a certain age range and then decline as they age

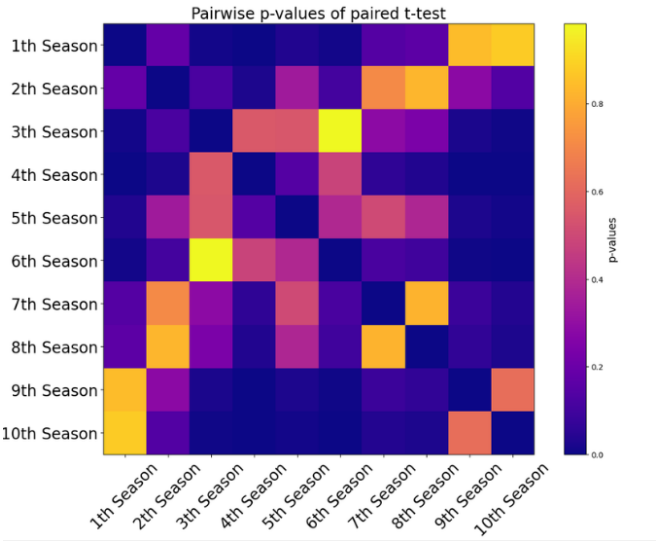
This analysis examines the relationship between basketball players' ages and their average playing time. The hypothesis suggests that there is a peak in playing time within a specific age range, after which playing time declines. An ANOVA test was employed to compare average playing times across different age groups. The test yielded an F-value of 59.77 with a highly significant p-value ($p<0.001$), indicating substantial differences between the groups. The line graph entitled 'Average Playing Time by Age' reveals a peak in playing time for players in their late 20s, followed by a decline into their mid-30s. The shaded area indicates variability within the age segments. The bar graph 'Average Minutes Played by Age Group' corroborates these findings, showing clear distinctions in playing time among the different age categories. The data confirms the initial hypothesis: basketball players' average playing time reaches its zenith when they are in an optimal age range and diminishes as they age. This trend reflects the typical arc of a basketball career.

3.3 Shooting and scoring over age

Hypothesis 5: Players tend to improve their true shooting percentage (TS_PCT) as they gain more experience in subsequent seasons

The analysis aims to explore changes in players' true shooting percentage (TS_PCT) over their careers, hypothesizing that experience correlates with increased scoring efficiency. We applied a paired t-test to seasons with a minimum of three years of data, specifically comparing the 1st and 3rd seasons. The test results were significant with a p-value of 6.316e-05, indicating a departure from the null hypothesis that there is no difference between seasons. The line graph displays the mean true shooting percentage growth across seasons, with an observable increase in efficiency after the 3rd season. The heatmap of p-values from paired t-tests across seasons further substantiates these differences. The paired t-test statistic of -4.0168 and Cohen's d of -0.130 suggest a small yet statistically significant improvement in scoring efficiency as players progress, aligning with the hypothesis that experience enhances performance. The average career span of 5 seasons indicates that this improvement is most critical in the early stages of a player's career.

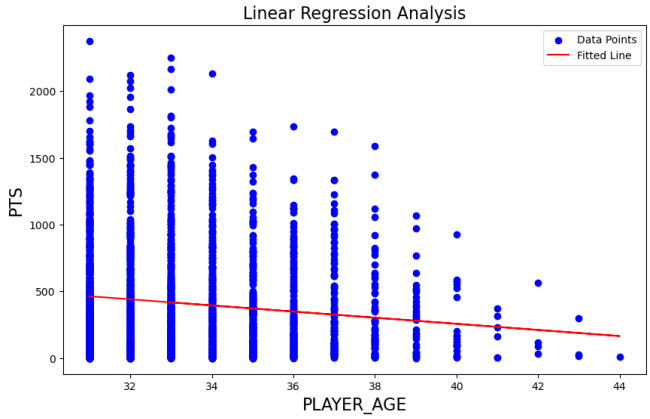




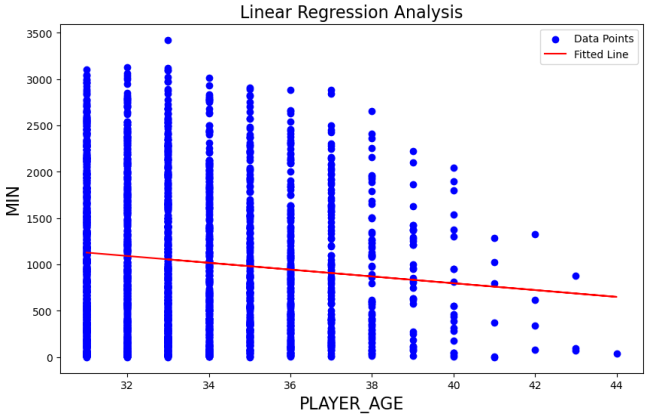
Hypothesis 6: Players tend to score less after age of 30 as their physical form becomes worse.

This analysis aims to evaluate the trend in scoring performance of basketball players post the age of 30, a

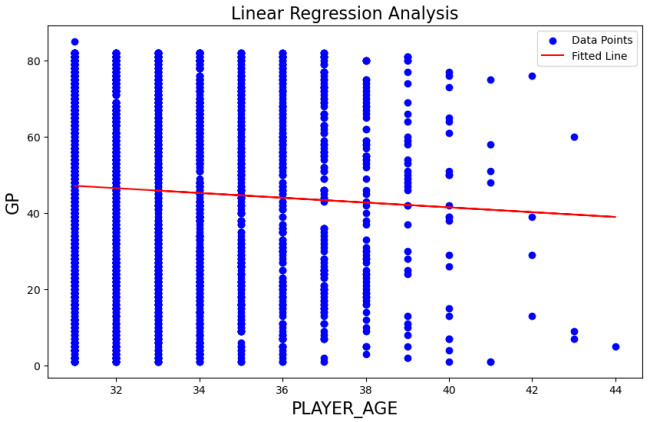
period generally associated with a decline in physical form. An Ordinary Least Squares (OLS) regression model was utilized to examine the relationship between players' ages and their performance metrics including points scored (PTS), minutes played (MIN), and games played (GP). The sample included players aged 30 and above, with 2179 observations. The regression analysis revealed a negative correlation between age and all three performance metrics, signifying a decrease in PTS, MIN, and GP as players aged. The OLS results were statistically significant, rejecting the null hypothesis of no difference in performance post age 30. Consistent with the alternative hypothesis, players tend to show a decline in scoring performance after the age of 30. However, we see that the reason for that may not be the decline in physical form of the basketball players. We see that the reason for scoring less after the age of 30 is that players tend to play less minutes and games. Thus, they score less. If we take a look at their true shooting percentage after age 30, we see that it does not correlate with the age. That's why we may conclude that players score less after the age of 30, however the reason for that is that they play less.



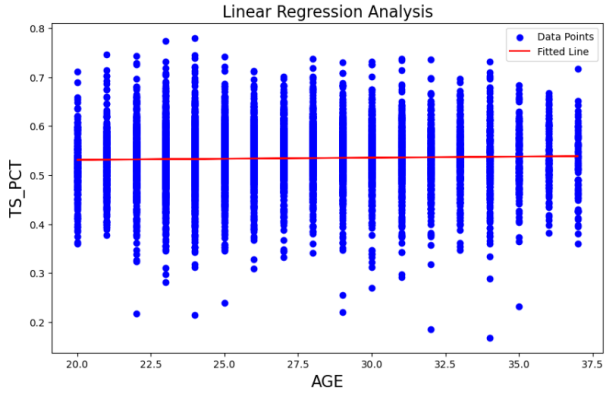
	coef	std err	t	P>|t|
const	1175.0585	119.256	9.853	0.000
PLAYER_AGE	-22.9519	3.569	-6.430	0.000



	coef	std err	t	P>|t|
const	2268.8477	237.826	9.540	0.000
PLAYER_AGE	-36.7730	7.118	-5.166	0.000



	coef	std err	t	P>|t|
const	66.5806	7.291	9.132	0.000
PLAYER_AGE	-0.6287	0.218	-2.881	0.004

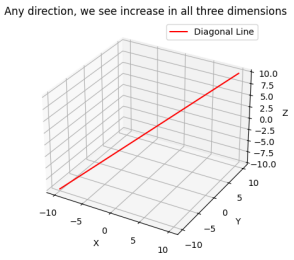


	coef	std err	t	P>|t|
const	0.5221	0.004	117.629	0.000
AGE	0.0004	0.000	2.723	0.006

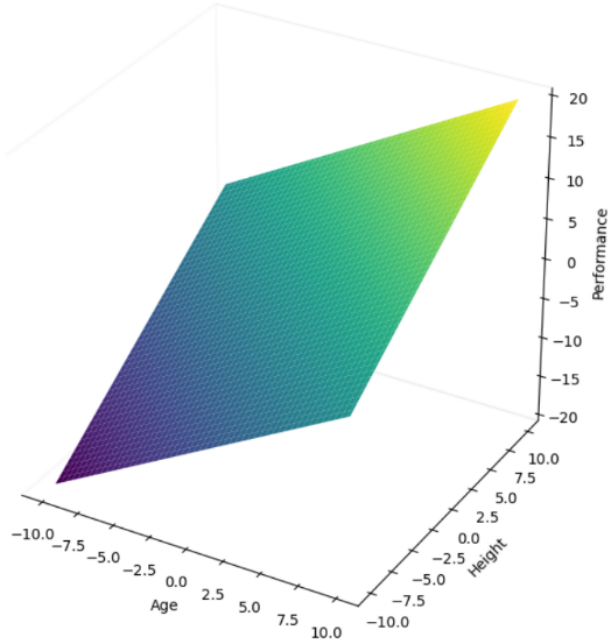
4. STATISTICAL TRAP: CORRELATION IS NOT TRANSITIVE

When looking at height and age correlations with performance statistics, we noticed an interesting phenomenon. We are trying to look at correlations of height and other statistics, and since for 99%+ of players that doesn't change with age. However, age can be a key point in how well one player does (although we said earlier that it matters less, and experience matters more, the stats show that players do best in most statistics at age 26-30, and they do "bad" at age 18-20), and we see positive correlation between height and performance, so if taller players play better (on average, do better in some statistic), and older/more experienced players also do better, one might get confused and think for a second: are taller players older? Of course not, they are statistically independent, even for NBA players (one may think only the best, tallest players manage to stay in the NBA after X seasons, but that is false). But this is not that clear sometimes, when looking at other research questions, having information on a variable A and its relationship with B and C, we can incorrectly make assumptions between B and C that are actually incorrect. Such as: if an increase in B means an increase in A, and C is increasing when A is, then an increase in B also means increase in C, which is generally not true, as proved with our performance-height-age example. Correlation is not transitive.

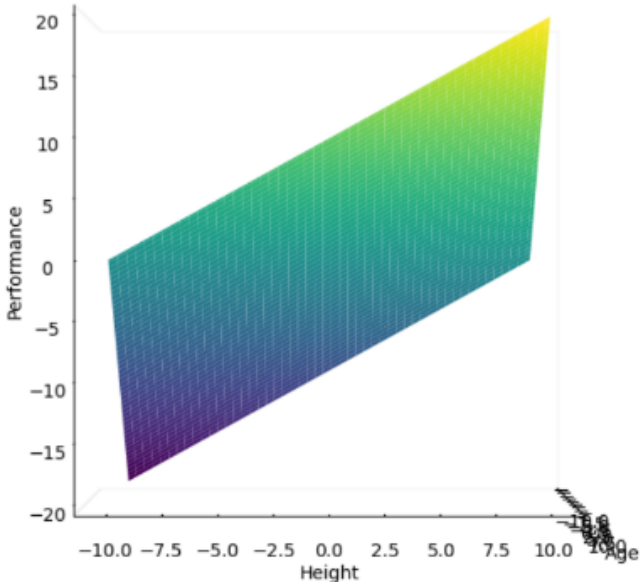
So what happens mathematically, why we are deceived? That is because we think in averages. If you look at average values of B over A and see a growing function (let's say, just a line with positive slope), and assume that all datapoints lay on that line, then in fact, it would be true that an increase in B also means an increase in C. Because in the 3D world, we'd have just a line of values, and that line has one direction with a positive slope over all variables (=all partial derivatives are positive, or all are negative).



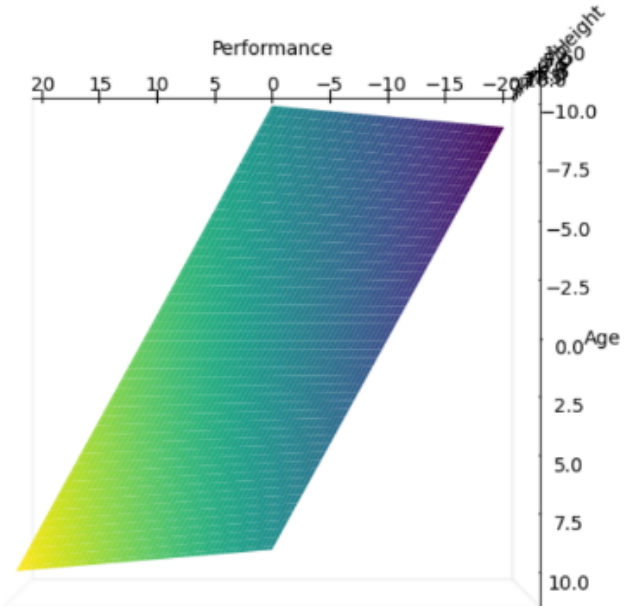
But when you look at not just the averages, you get back the dimensionality.
Let's assume the formula: Performance = Age + Height, linear combination of the two. Let's plot this function:



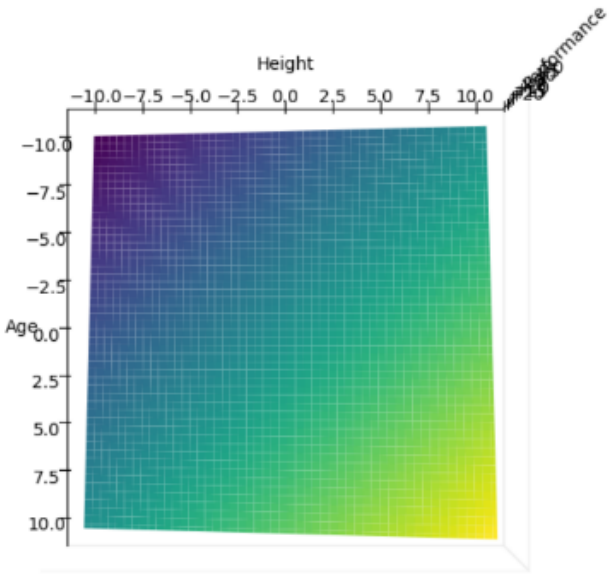
If we look from it from different sides, everytime one component "falls out" and we can see the correlation of just two variables:



See how much this plot resembles the bottom figure on Page 3?



Indeed we see the increases, if we'd draw a best fit line it would have positive slope. But for age vs. height:



We don't have any correlation, that is why the plot is the whole domain. These variables are independent.
On dealing with such effects, one may run multiple regression, which is not limited by pairwise relationships. The way this problem could come up in machine learning is this: When trying to predict height from performance, performance depends on age, so age matters indirectly. Should you include it as an attribute for building models? For deciding this, consider what the model builds upon. Trees, and thus, random forests only care about pairwise correlation, so they'd not make use of including age. A specially constructed neural network may (combinations of different outputs for prediction have been used in for example computer vision, such as taking the weather, time guessed from shadows etc. for prediction, and combining these "independent" pieces of information with what's seen in the picture).

5. CONCLUSION, FURTHER WORK

This project led us to understand more about how some biographical attributes and performance of basketball players playing in the NBA are correlated. We used statistical tools to find how players change in time, and how they play based on their physical attributes. For further work, we also created machine learning models (random forests, gradient descent), to try to predict height just from performance statistics. This lead to the "best we know" predictions of player height in the NBA. On the internet, we did not find any approaches that exhausted a large database and used some models to predict height, but any biographical statistic predicted from performance stats is rare, the vica versa is more common. These analyses, and a clustering of players we did to find interesting correlations between players (that lead to cluster the players into mostly positional/role categories) could help understand how the "playstyles" in the NBA are biased on physical attributes, leading to a big question: "are you, Mr. Basketball Player playing on the position you always wanted, or where your boss put you because of your size?"
Some things we learned in this project:

- Statistical tools are very powerful
- One has to be very careful when splitting their data manually… If it is biased in any way, that can hurt many test and analysis effectiveness. Always check the assumptions of a tool before using it.
- Statistical tools are complicated and designed centuries ago, modern data analysis is safer and easier to correctly execute, and combined with a good eye and domain knowledge is unbeatable, even by statistics and AI. We still prefer this.

References

[1] N. B. Association, „NBA Advanced Stats,” [Online]. Available: https://www.nba.com/stats. [2023].

[2] T. Boger, „Defining NBA players by role with k-means clustering,” [Online]. Available: https://dribbleanalytics.blog/2019/04/positional-clustering/.