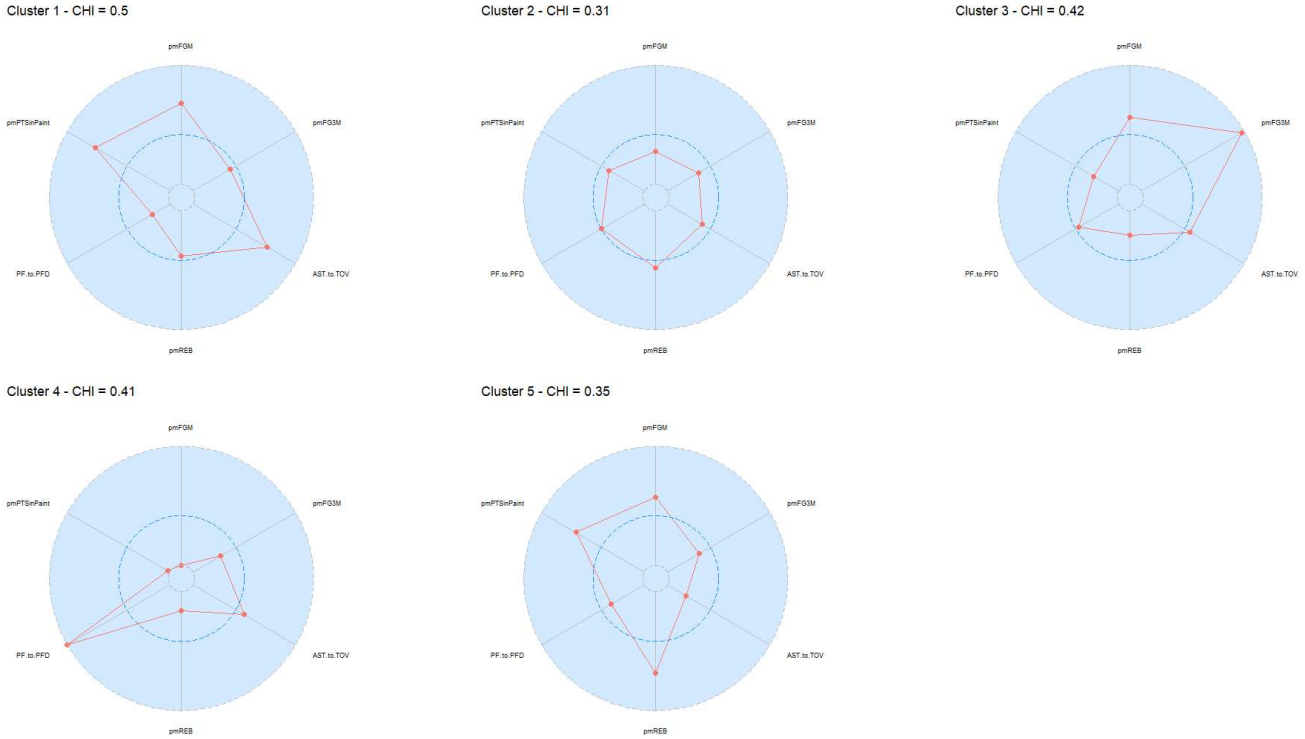
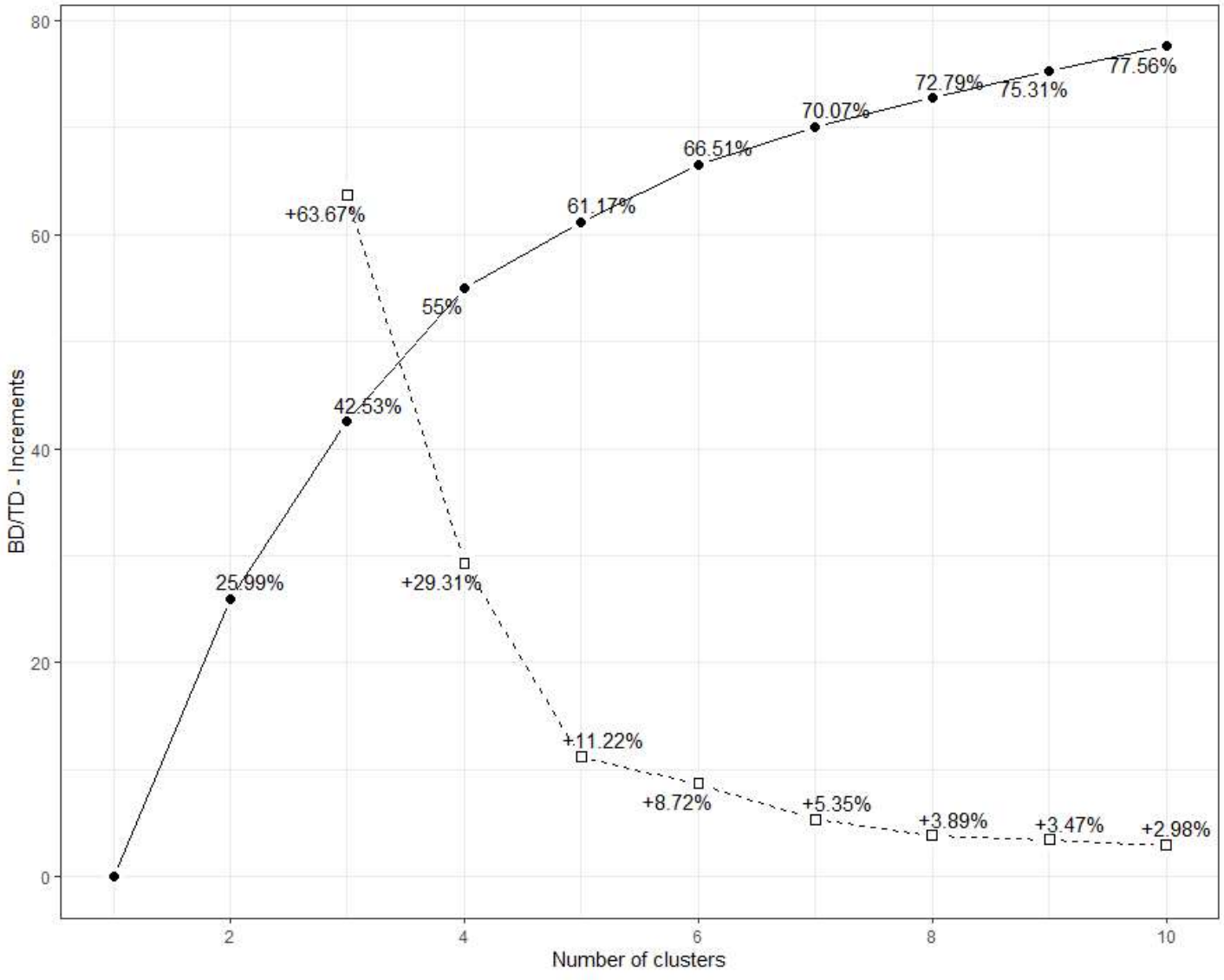
Alperen Sengun Percentile Rankings

Clustering

During ***Socrates Dergi***'s episode with Alperen Sengun, reporters mentioned how similar Domantas Sabonis, Nikola Jokic and Alperen Sengun are. So, I wondered if they'd fall under same cluster if I cluster centers, it turns out they do! They are members of "Cluster 5" in the plot below.

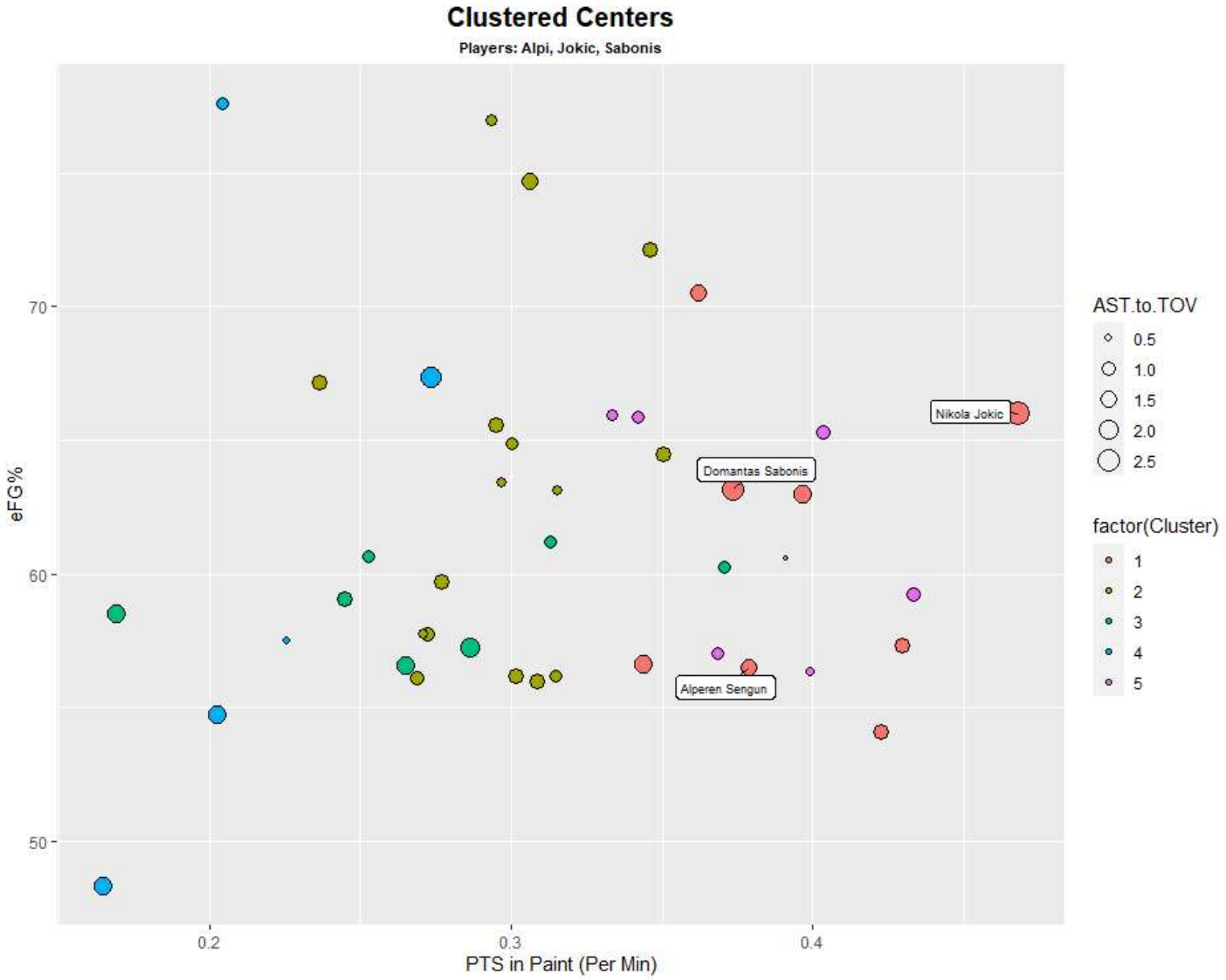

Cluster 1 - CHI = 0.5
Cluster 2 - CHI = 0.31
Cluster 3 - CHI = 0.42
Cluster 4 - CHI = 0.41
Cluster 5 - CHI = 0.35
pmFGM
pmFG3M
AST.to.TOV
pmREB
PF.to.PFD
pmPTSinPaint


BD/TD - Increments
Number of clusters
80
60
40
20
0
2
4
6
8
10
25.99%
42.53%
55%
61.17%
66.51%
70.07%
72.79%
75.31%
77.56%
+63.67%
+29.31%
+11.22%
+8.72%
+5.35%
+3.89%
+3.47%
+2.98%

I clustered centers to 5: Selected 5 since after 5, there isn't dramatic decrease in between deviance to total deviance ratio which means I don't get much information that would worth the complexity that comes with more clusters. CHI (**C**luster **H**eterogenity **I**ndex) values that are on top of the radial plot for each cluster is a metric for variability within the group: It measures the average variability within the cluster: $CHI_h = \frac{\sum_{j=1}^{p} \sigma^2{}_{jh}}{p}$

To cluster, I used variables per minute (field goal made, 3 pointers made, points in paint, rebound) and ratio variables (assists to turnover ratio, personal fouls to personal fouls drawn).

On the bubble plot below, vertical location gives the eFG% while horizontal location gives the points per minute. The size of the dots stands for assist to turnover ratio and the colors are assigned based on which cluster the player is under.

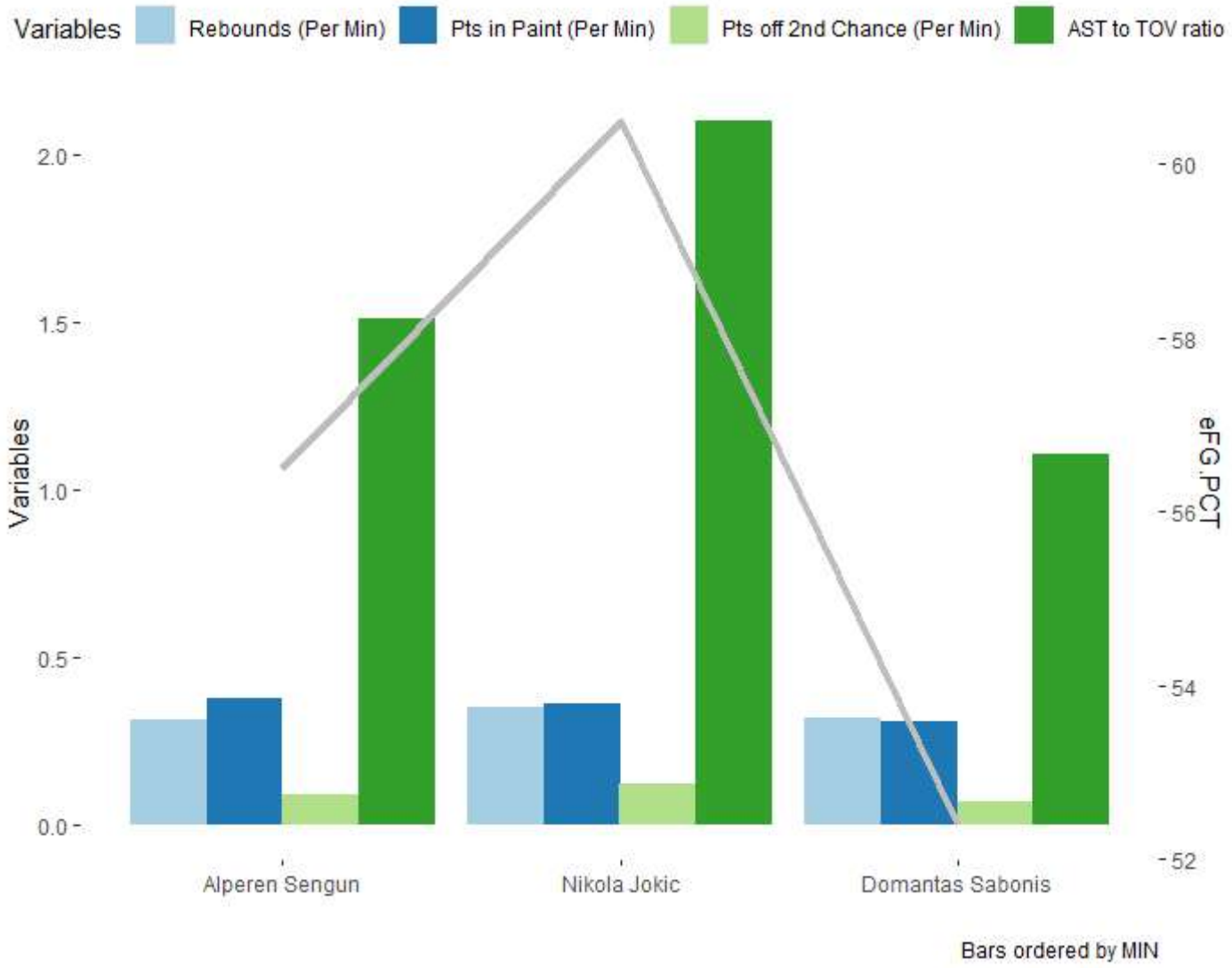


The bubble plot above was made using 2022-23 season for each player. However, Alperen is on his second year and others are veteran now. So, to be fair, I pulled Jokic's and Sabonis' second year data and compared then.

Descriptive Analysis

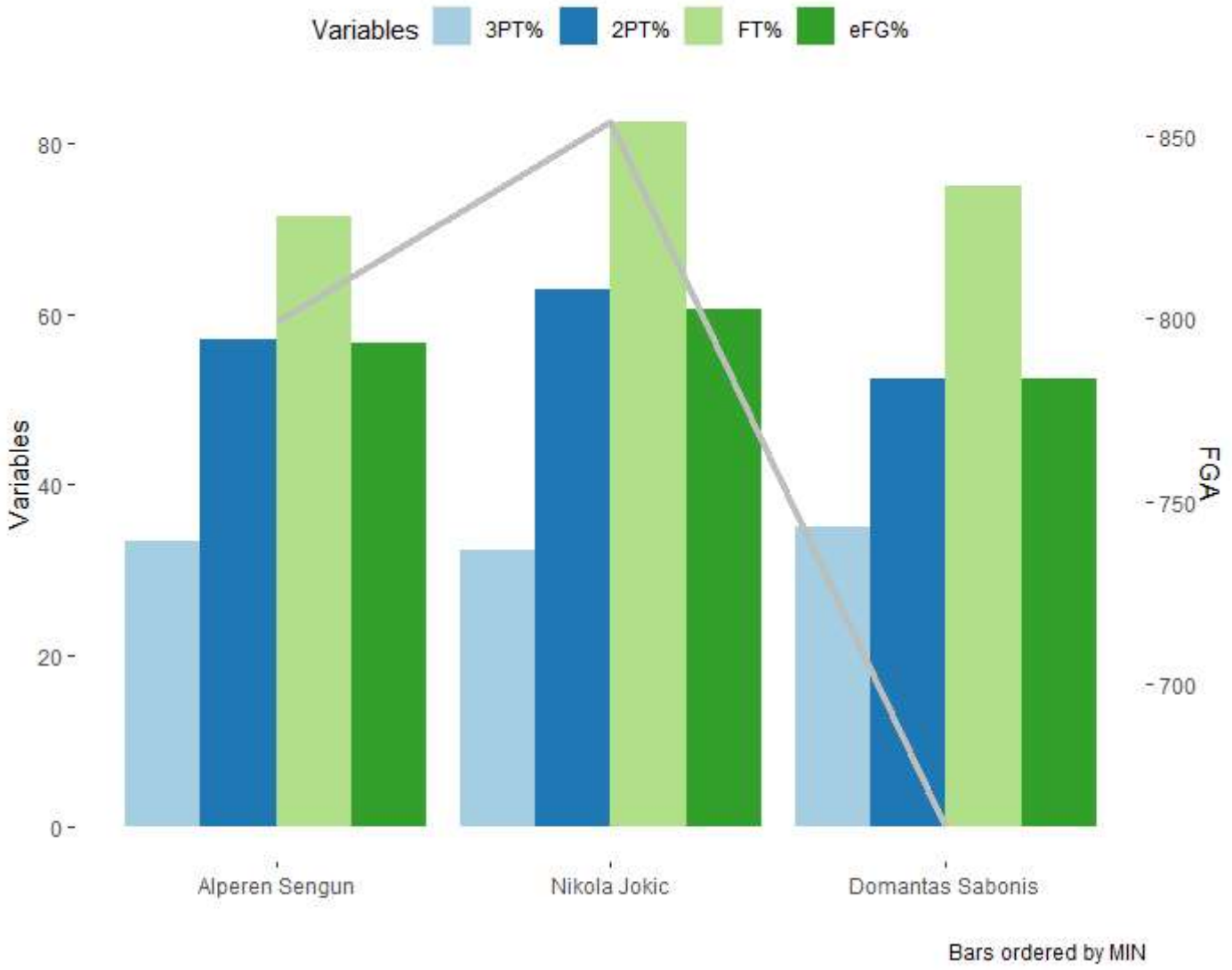
Second Year Performance Comparison

On the bar-line graphs below, the players are ordered based on how many minutes they played on their second season. On the first graph, vertical axis on the left corresponds to bar height while the one on the right corresponds to values for the line. One can see which variables represented by which bar, on the top. Alperen looks better than Sabonis and worse than Jokic considering these metrics.



Similar conclusion can be drawn when one looks at the shooting percentages. This time the line corresponds to field goal attempts (FGA): The Joker takes the top on this one with 854 FGA. While Sabonis shoots with 35.1% behind the line, compared to Alperen's 33.3%, he attempted less 3s than Alpi. Jokic is at the forefront on this one as well, with 139 shot attempts behind the line while playing with 32.4%.

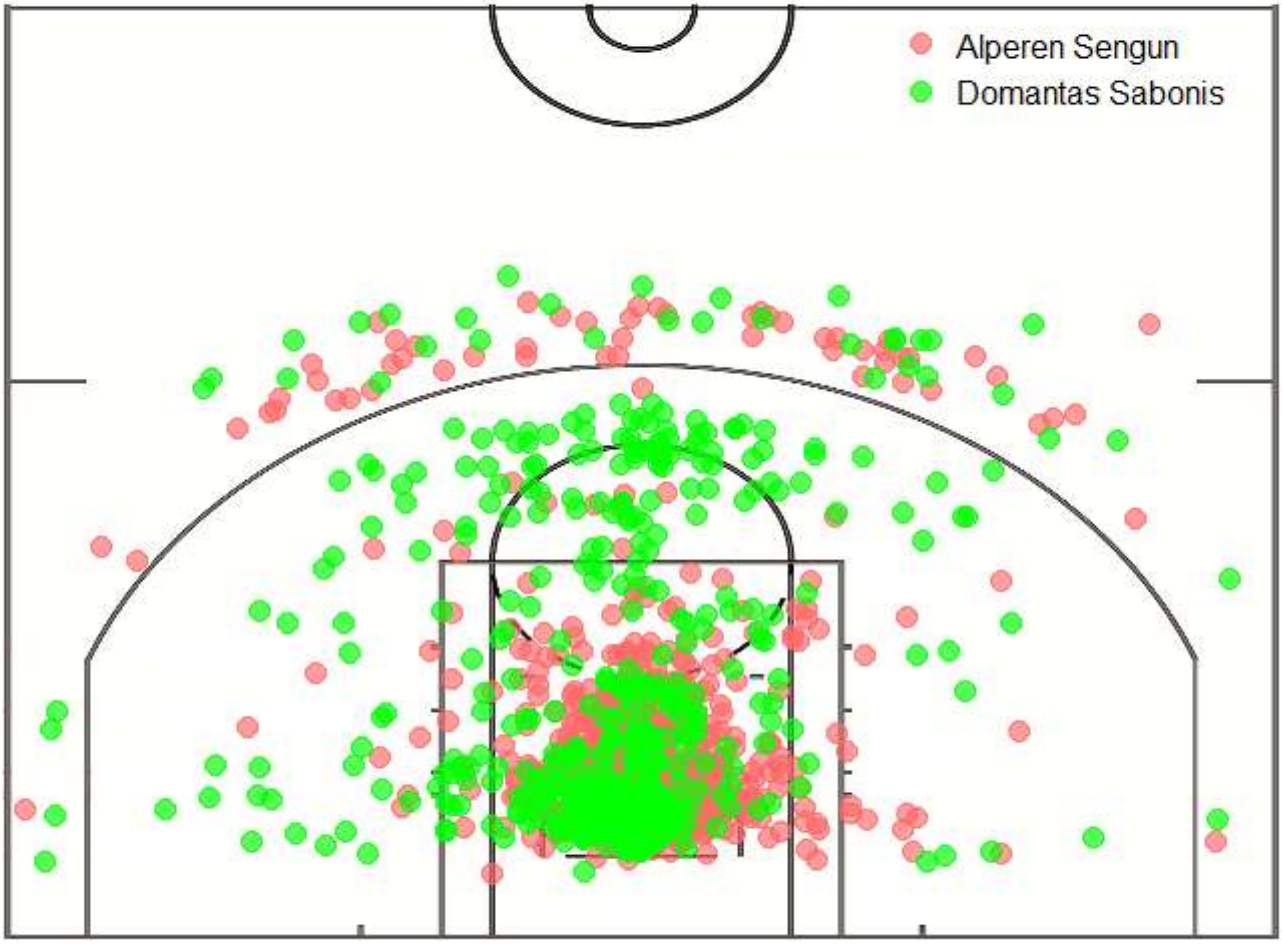
Alperen is the one who came to free throw line the most on his second year (330 times) but he is the worst in terms of free throw percentage (FT%): 71.5% compared to Sabonis' 75% and Jokic's 82.5%.

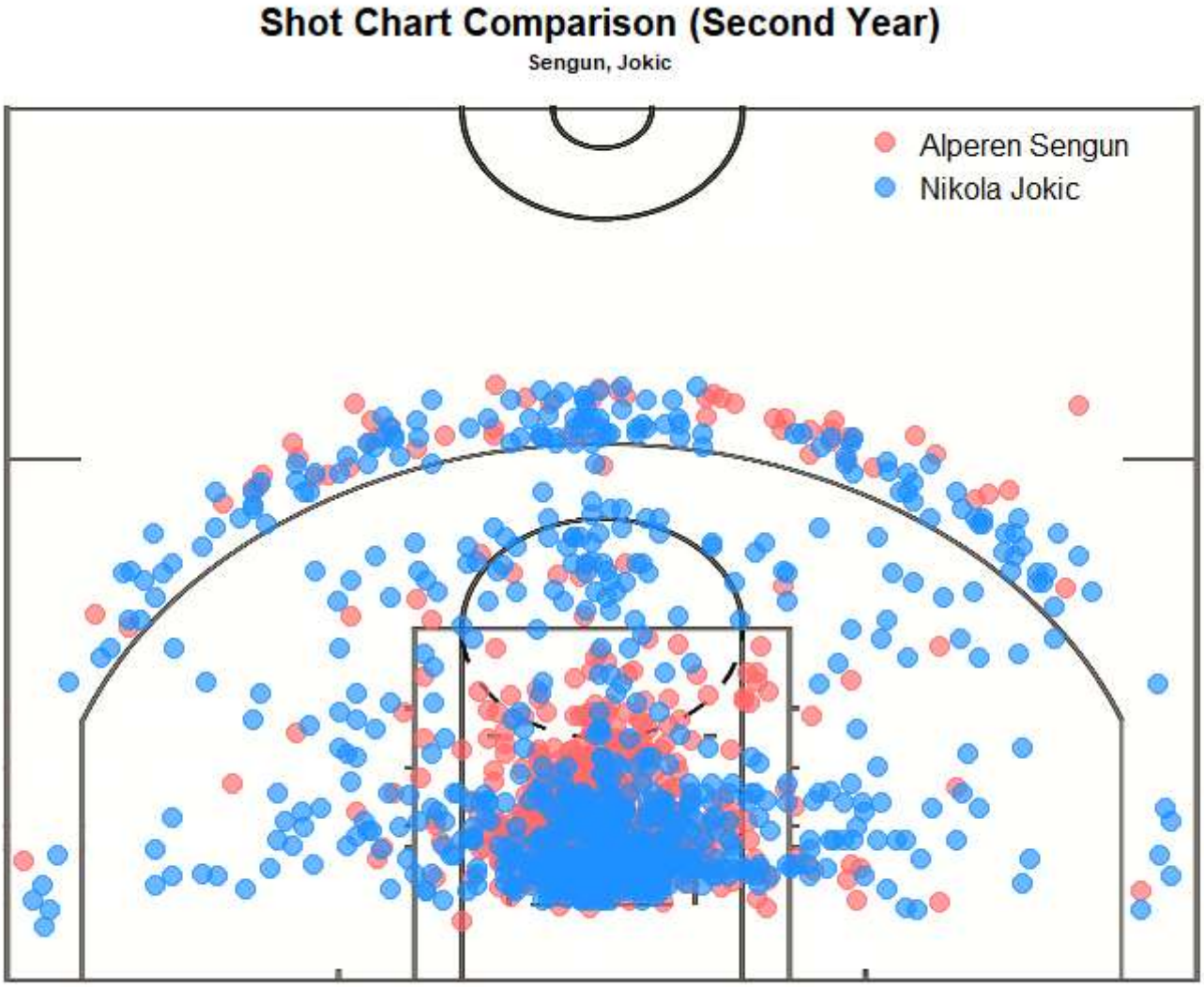


To understand where they take the shots most, here's the shot charts for their second years. For readability, I drew two charts instead of including all in one.

Shot Chart Comparison (Second Year)

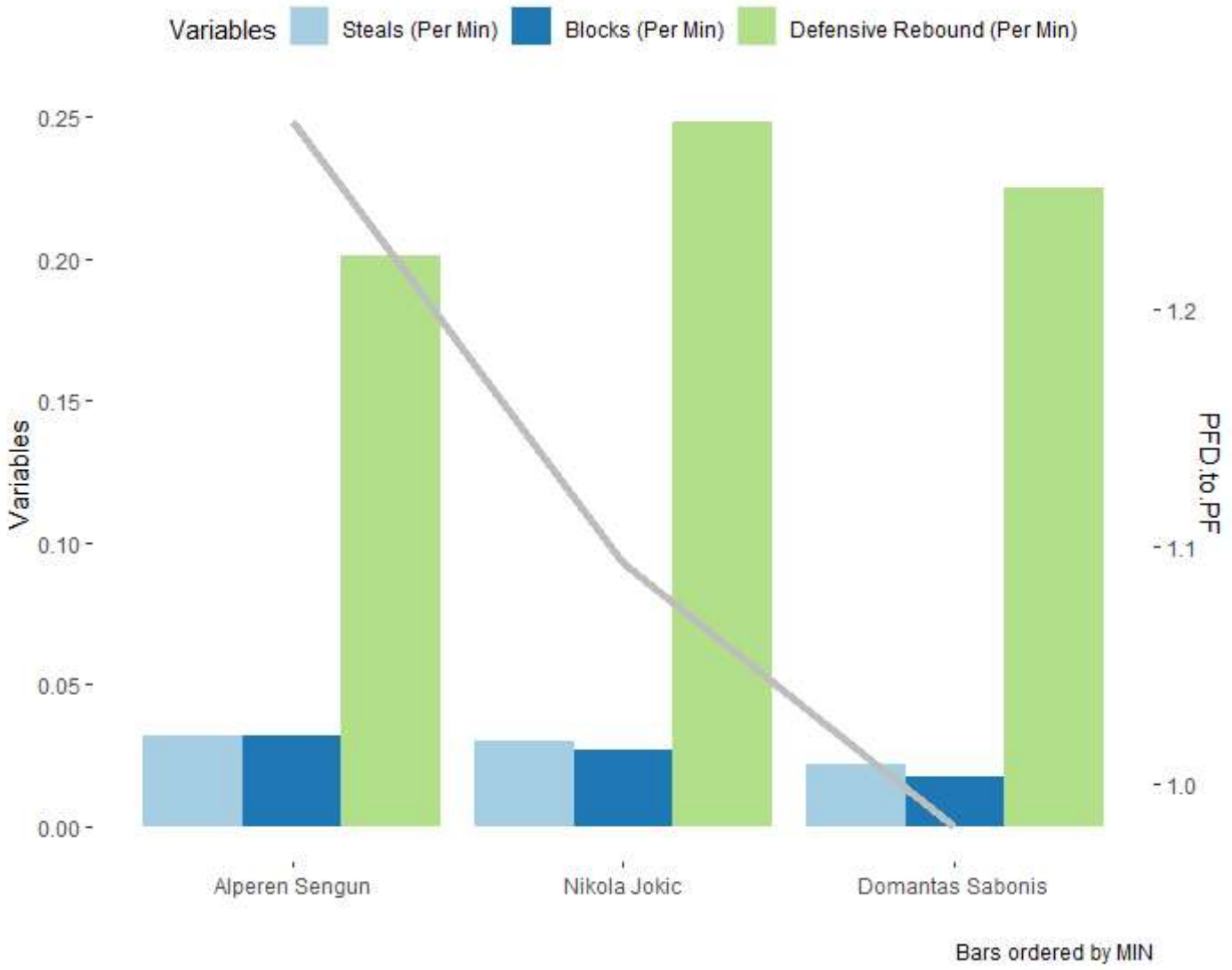
Sengun, Sabonis





Jokic's relatively more 3 point attempts are apparent in the chart and it looks like Sabonis' takes more "long 2s" compared to other two. Within the field, specifically behind the 3 pt line, Alperen attempts his shots from the center the most. On the other, Jokic's shots are more spread on both the horizontal and vertical axis, reflecting the variability of angles that he takes his shots from. While Sabonis' has less 3 pointer attempts than Alperen, he prefers corner 3s more.

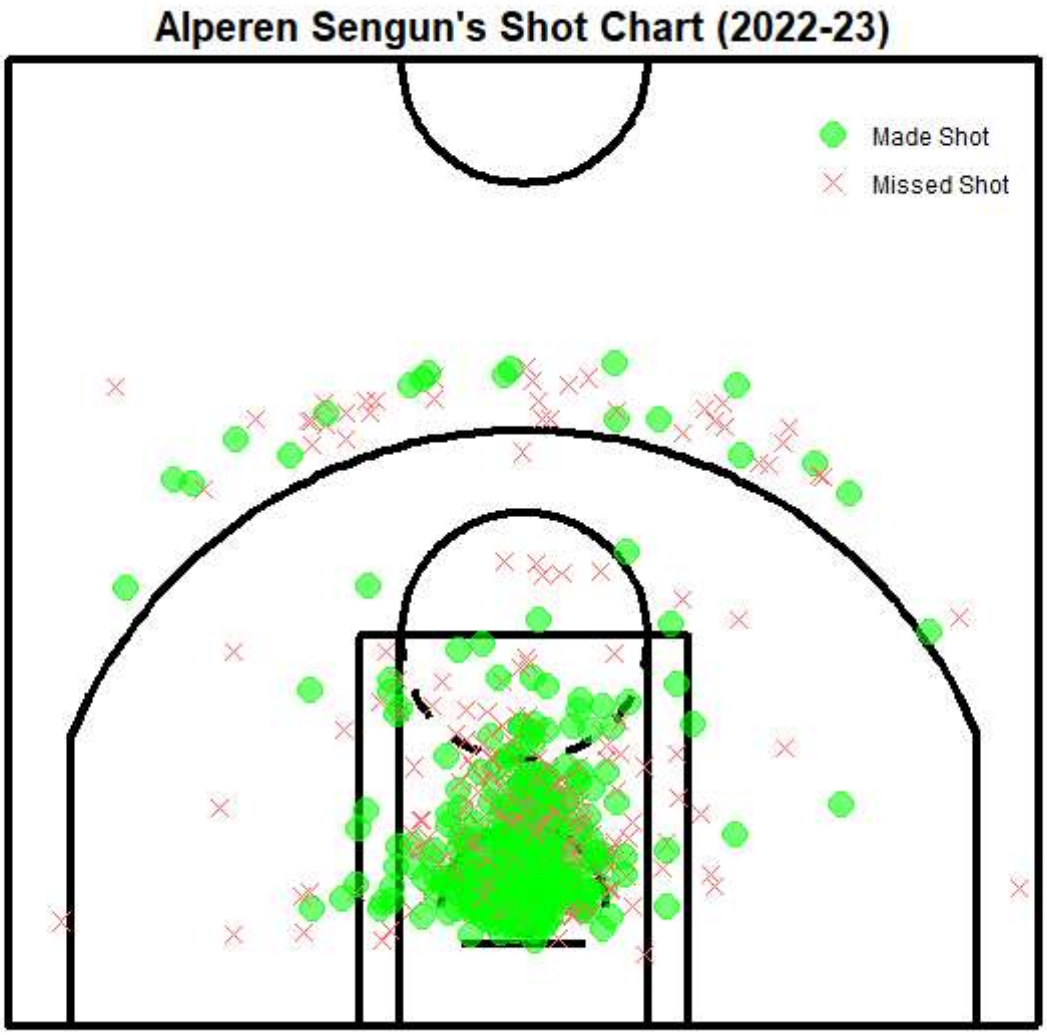
Let's take a look at the other end of the floor: Defense.



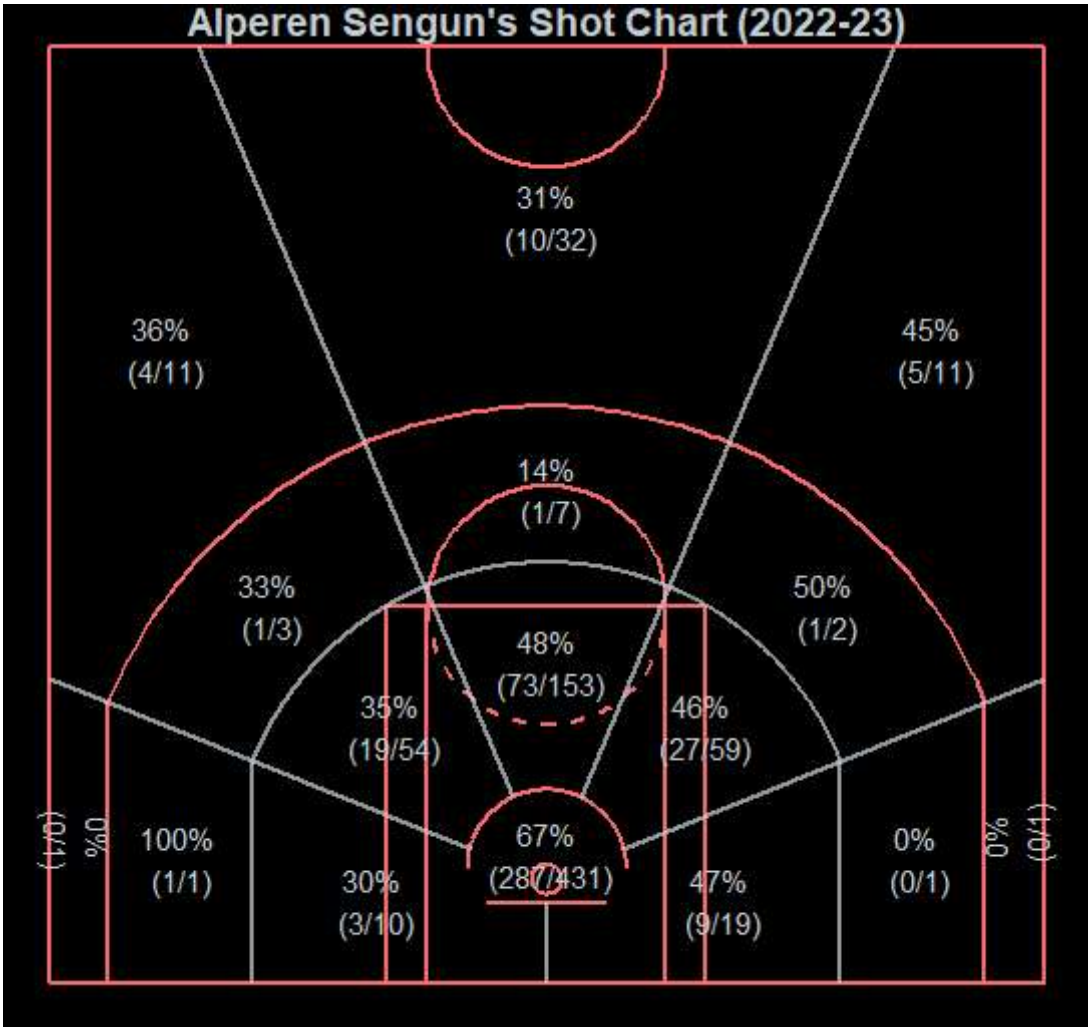
This time the line corresponds to personal fouls drawn to personal fouls ratio. Other than defensive rebounds, Alperen looks better than both of them.

Only Alperen

From now on, I'll solely focus on Alperen. To get a better sense of his shots, here's his shot chart with shot success.



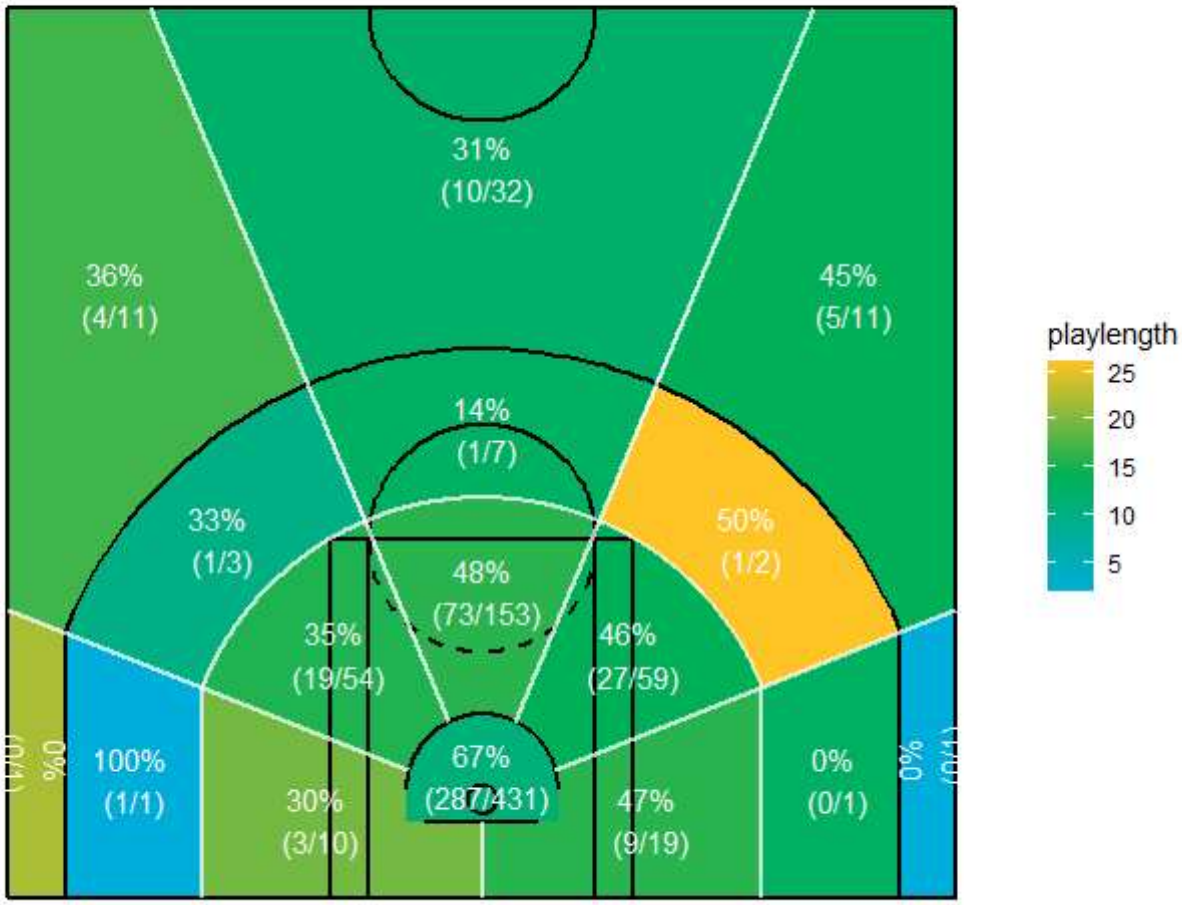
Segmenting the floor and showing the success percentage from each segment might provide more information.

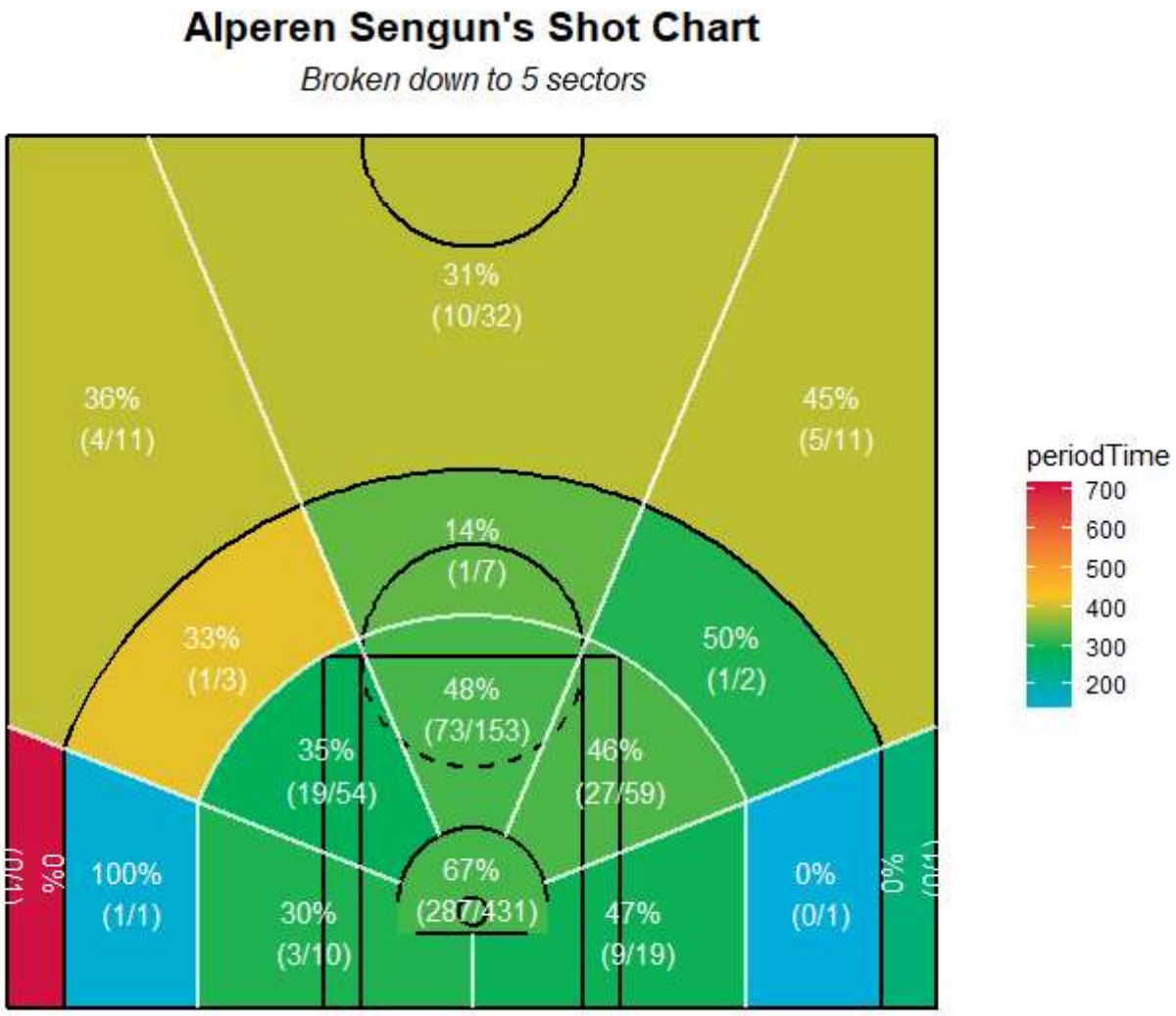


In the plot above, it looks like Alperen shoots better from the left side of the court. The most efficient areas for him are left center threes which he shoots with 45% and near the rim with 67%. A rough calculation for expected points for distances that corresponds to these areas are 1.35 (0.45 * 3) and 1.34 (0.67 * 2), respectively.

Alperen Sengun's Shot Chart

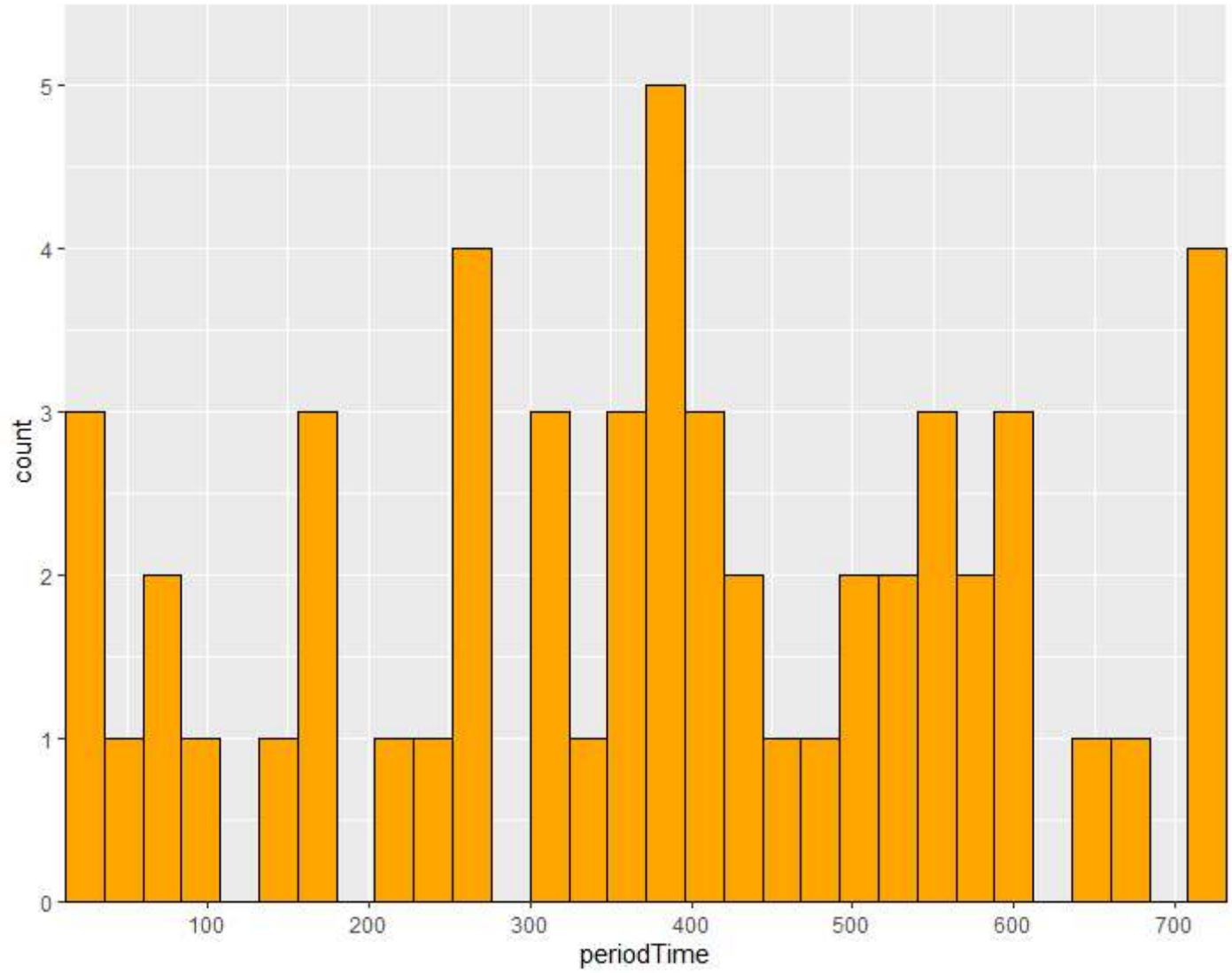
Broken down to 5 sectors



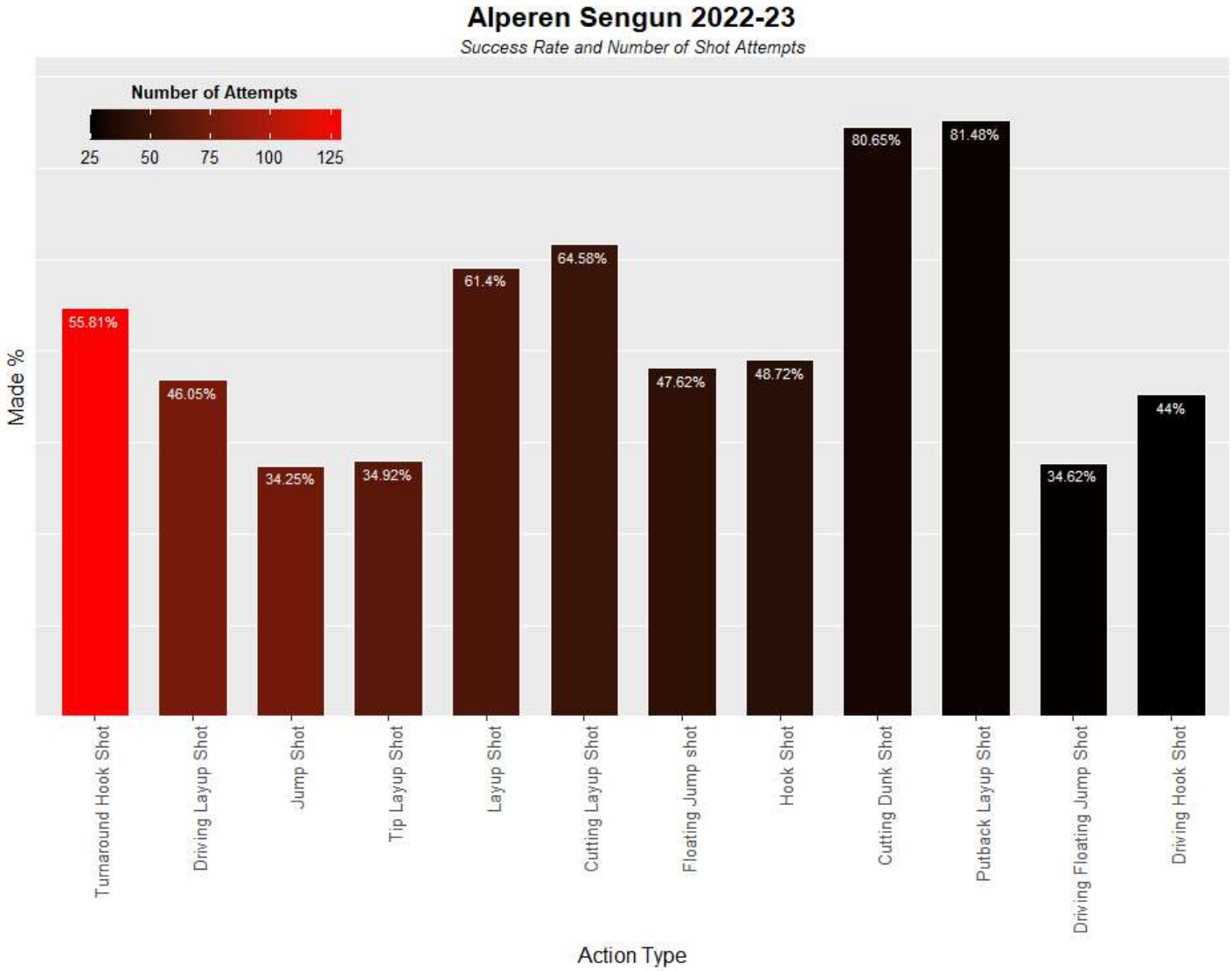


The two plots above carries additional information. On the first one, the sectors are colored based on “playlength” variable. Playlength describes the time since the preceding event (e.g., rebound, turnover, steal, shot etc.). In the light of this, Alperen’s near the rim shots occur around 15 seconds into the play, on average. Same comment can be made for threes (excluding corner ones).

On the second one, sectors are colored according to the “periodTime” variable, which gives the amount of time played in the quarter. It is easy to derive that the one corner three was taken towards the end of the period. In general, Alperen’s shots occur early to mid period, on average.



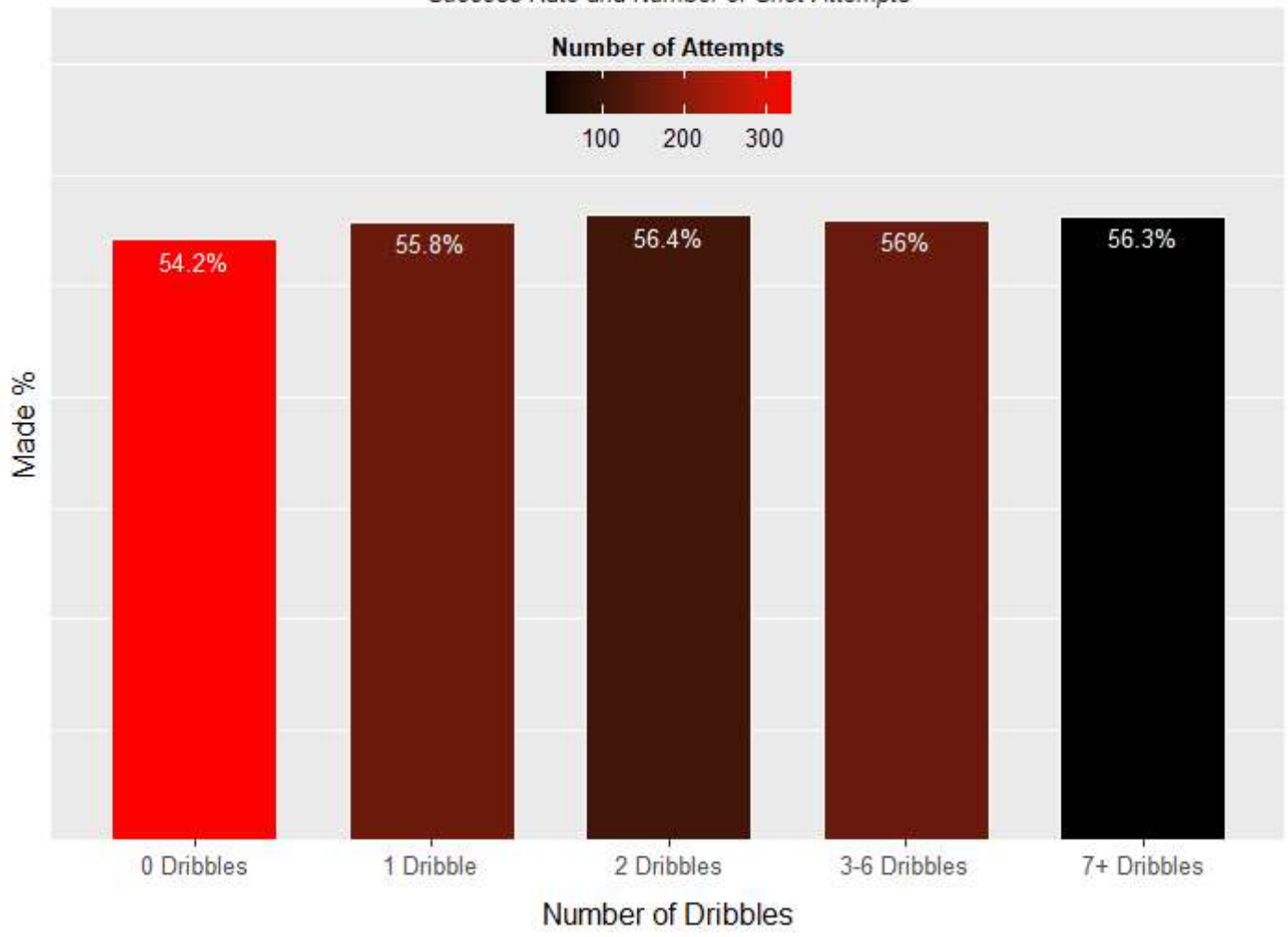
Alperen’s shot kinds can be found on the barchart below:

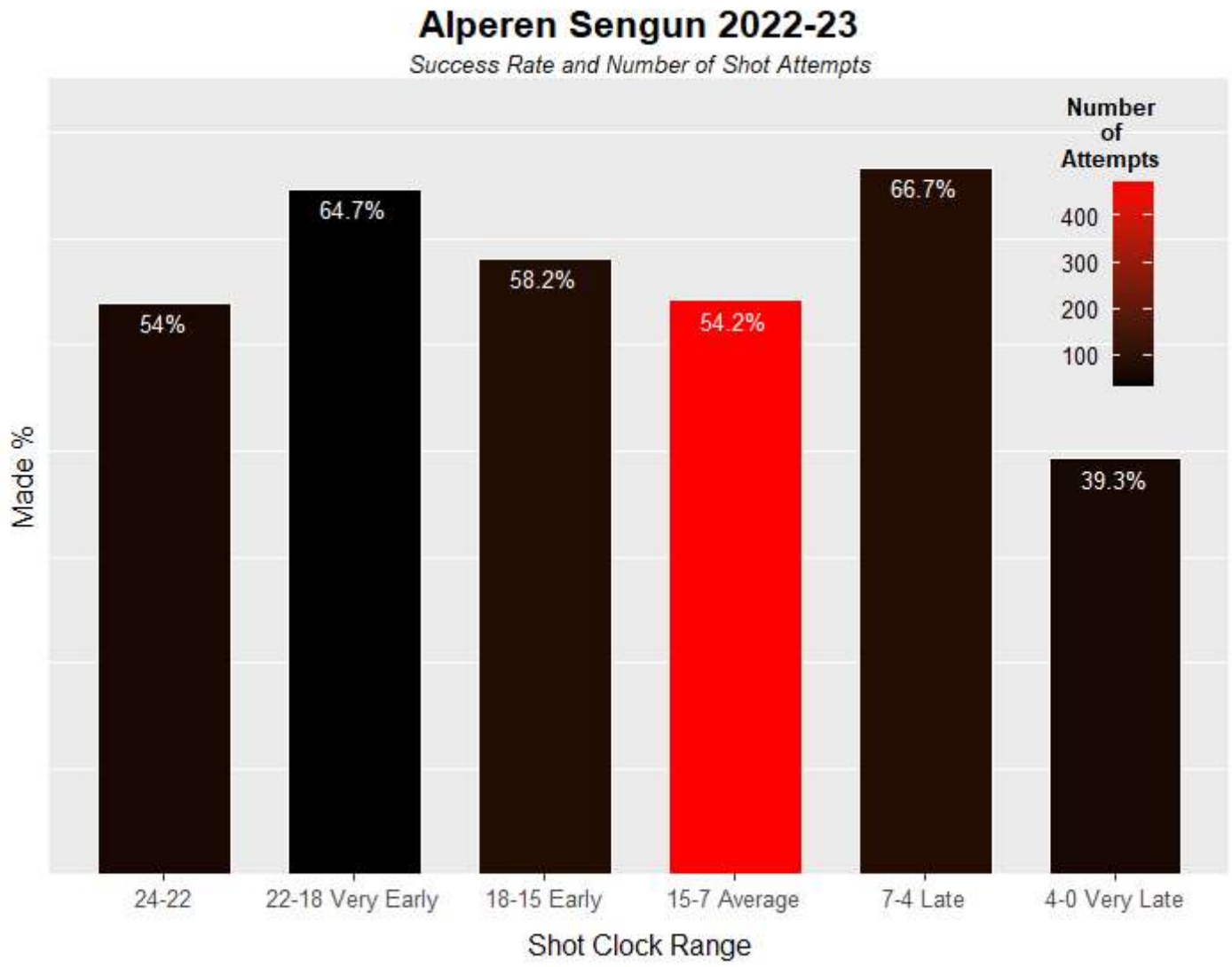


He attempts turnaround hook shots the most and it is expected due to his position and where he takes his shots generally. On the plot below, one can see the success percentage grouped by how many dribble Alperen took before taking the shot. It is hard to see significant difference in percentage based on number of dribbles Alperen took before shooting the ball.

Alperen Sengun 2022-23

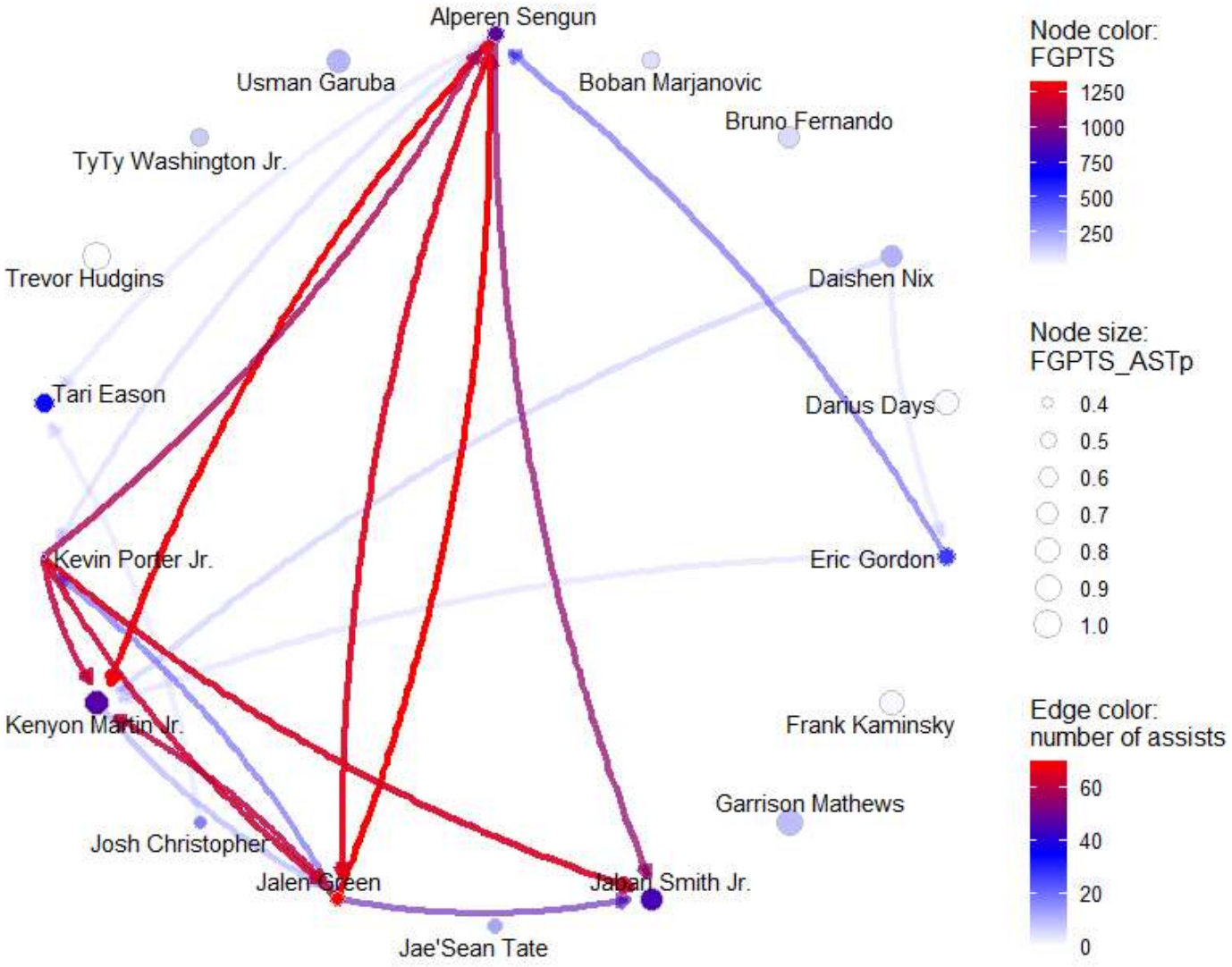
Success Rate and Number of Shot Attempts





On the other plot one can see the shooting percentages grouped by shot clock. Most attempts are in 15-7 seconds range while the most successful attempts came early or late range into the shot clock. Success in the early range might be a product of transition offense: When the defense is not ready yet. Success in the late range might be a product of good ball movement. With that being said, same cannot be said when it is “extremely late” since when there is less time left in the shot clock, defense increases its pressure on the ball handler since there is not much time for a passing option.

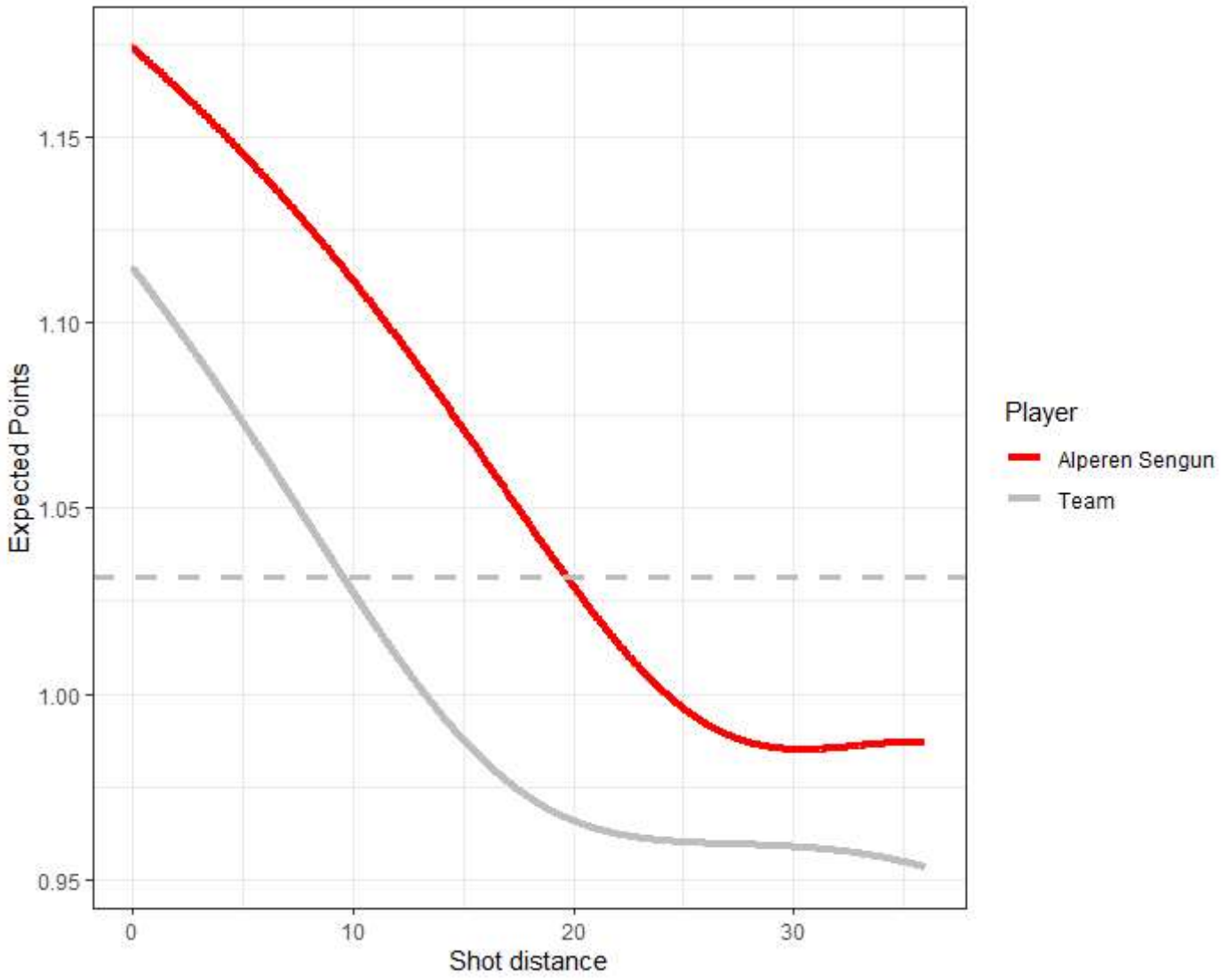
Besides his scoring abilities, Alperen has playmaking ability similar to how Denver runs their offense with Jokic. So, let’s take a look at Houston’s assist network:

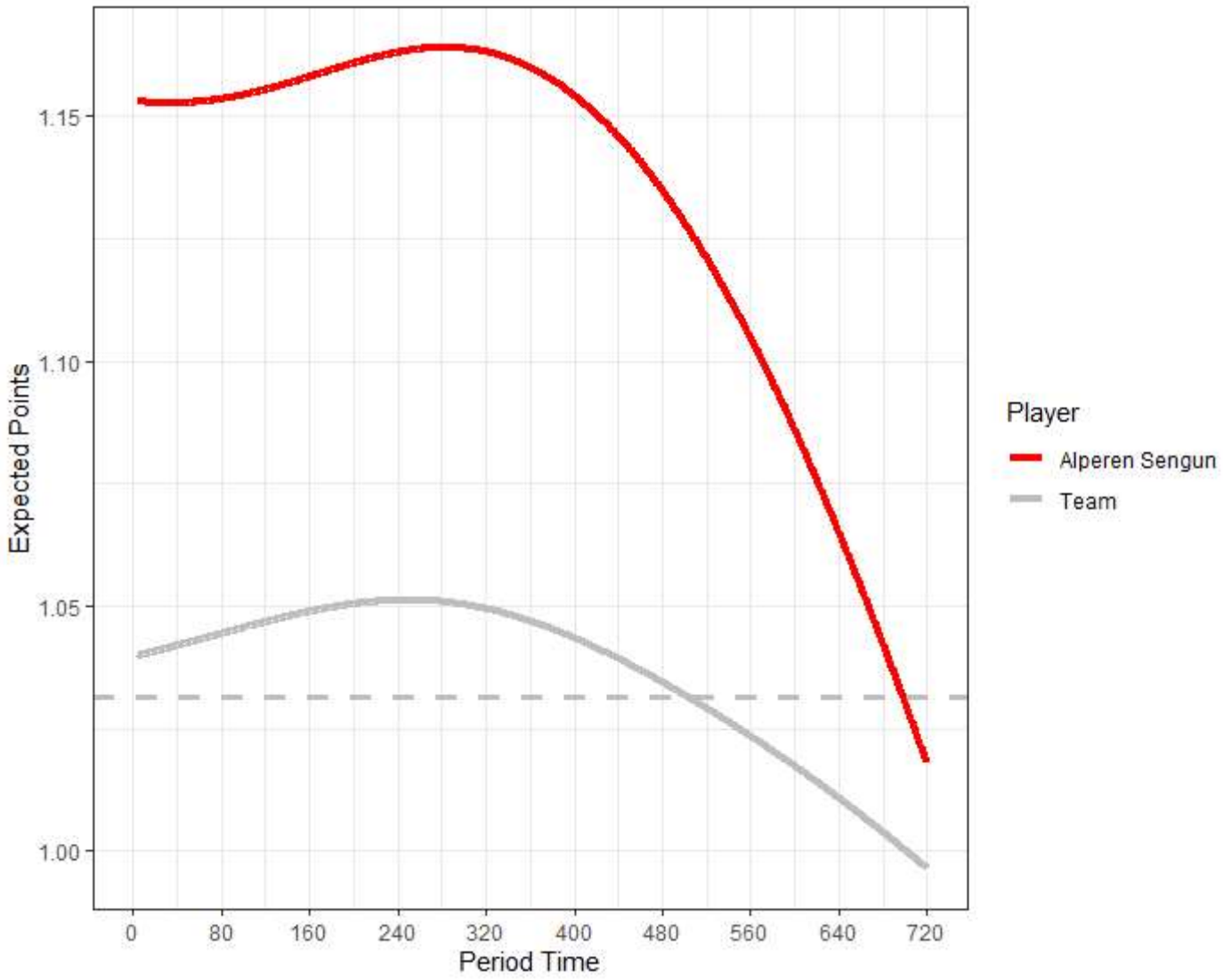


Node color represents how many points that the corresponding player scored during 2022-23 season, with field goals, while the size of the node represents the percentage of those points that comes from assists. Arrow color, on the other hand, represents the number of assists. Alperen makes most of his assists to Kenyon Martin Jr., Jalen Green, Jabari Smith Jr., respectively. There is an arrow pointing away from Jalen Green to Alpi, so for these two it is a two-way street. Other than him, Kevin Porter Jr. and Eric Gordon makes assists to Sengun.

Modelling

We did a rough calculation for expected points on shot chart section. However, we did not specify exact shot distance, rather we did it for certain distance range that corresponds to the area. This may be a bit flawed: Let's say one of the area when we divided the court into sectors correspond to 0 to 8 ft. And let's asssume all the misses that the player has correspond to 6 ft and longer. So, although the player plays with certain percentage within an area, it may not be distributed equally within it. Hence, to be more accurate, one needs to fit a function that is able to take every possible distance (or any other variable) as an input.


1.15
1.10
1.05
1.00
0.95
Expected Points
0
10
20
30
Shot distance
Player
Alperen Sengun
Team



Looking at the first graph, it is apparent that expected points decreases as the distance increases. However, around 24 ft. (around 3 point line) the decrease comes to an end. On the second graph, there is some increase as the period time gets larger until around 5 mins. Then, it starts decreasing with a huge negative slope and the slope decreases more for every value of period time (I'm trying to refer to derivate without mentioning it explicitly to make it more reader-friendly but I don't feel like I'm good at it).

While the approach above is univariate and gives us some sense of the relationship between the two variable, better approach would be to take variable interactions into account: For example, the effect of shot clock running down on shot probability might be different when shot clock is running down **and** there is relatively less time in the period **and** score difference is low. So, I used machine learning (ML) models to predict shot success.

I modeled 2 pointers and 3 pointers seperately. For both of the models, I used Zuccolotto et al. (2017) as a guide. Let me start with 3 pointer model:

3 PT Shots Model

For 3 pointer model, I removed two corner threes and one backcourt shot from the data which left me with 54 3 pointer attempt.

In addition to variables in the dataset (e.g., shot clock, shot distance, shot location, period time, period etc.), I created new ones such as the score difference before the shot attempt and the whether or not the last shot was made. Since the dataset is not large, I decided to go for a parametric method: logistic regression. Because of how

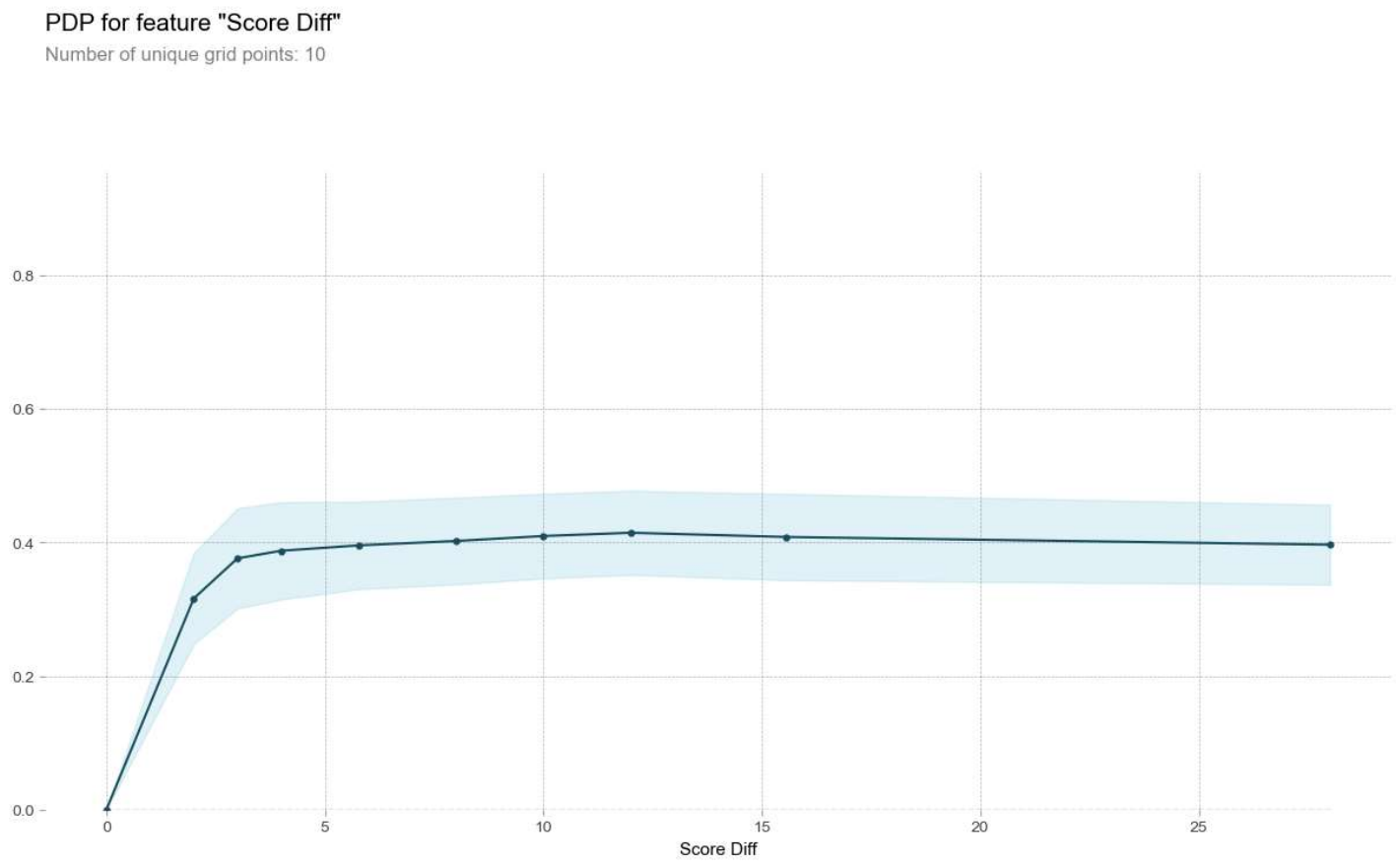
small the data is, I decided to uses “LeaveOneOut” cross validation which yielded average of 0.712 for accuracy. Whether or not the last shot was missed turned out to be the only statistically significant variable (with a coefficient of 2.506, which means that it increases the odds of making the shot, on logarithmic scale, by 2.5). However, with such a small dataset, it is hard to trust on standard error estimates so I did not remove any variable. Additionally, since I did not use a validation set that I could use for variable selection (and test the final model on test set) it wouldn’t be appropriate anyway. Which period the shot was taken had effect as well: Coefficient decreased for as the period increased — 4th period’s coefficient is -1.677.

2 PT Shots Model

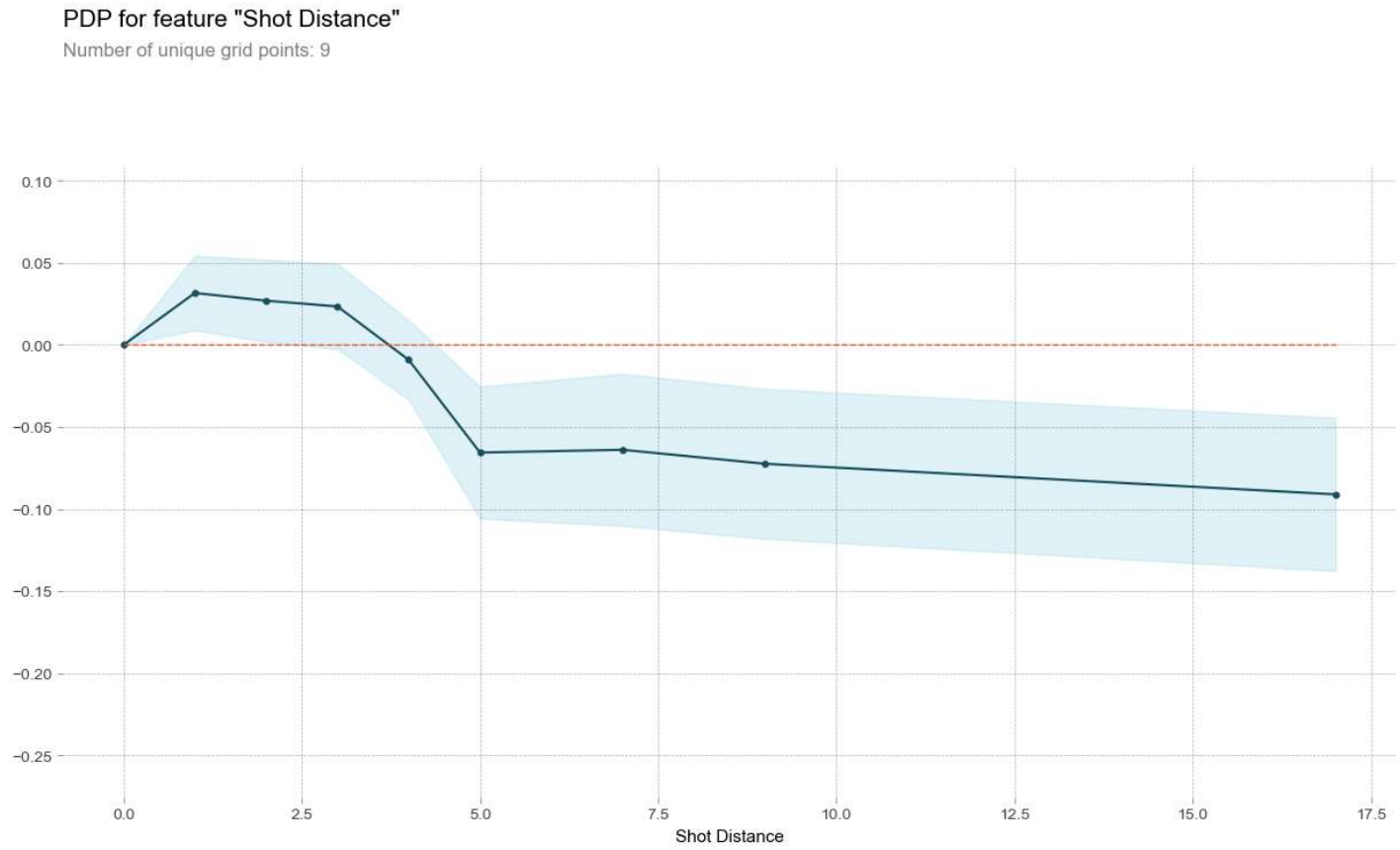
In addition to what I have included in the 3 pointer model, I added shot type as a variable. Even though it is not a very large dataset, I decided to try non-parametric models: Tuned random forest model was the best one based on how it performed on unseen data. Model fit was evaluated with area under the ROC curve, just like how Zuccolotto’s did and it yielded 0.705. Since the model contains many trees and runs a new observation on every tree and decides to assign its class then, it wouldn’t be wise to visualize a certain tree. So, I used model agnostic ML explainability methods to understand the model.

I started out with *permutation importance*: Select a variable of interest and shuffle its column randomly while not changing values in any other column, then make predictions with the shuffled dataset. If the predictions get affected a lot (i.e., how does your cost function changes) it means that the shuffled variable is important for the model. Score difference, shot distance, remaining time in period were the most important variables, respectively, according to permutation importance approach.

Although this approach tells us about what variables affect the model prediction the most, it does not tell us *how* they effect. *Partial dependence plots* (PDP) are helpful for answering such question.



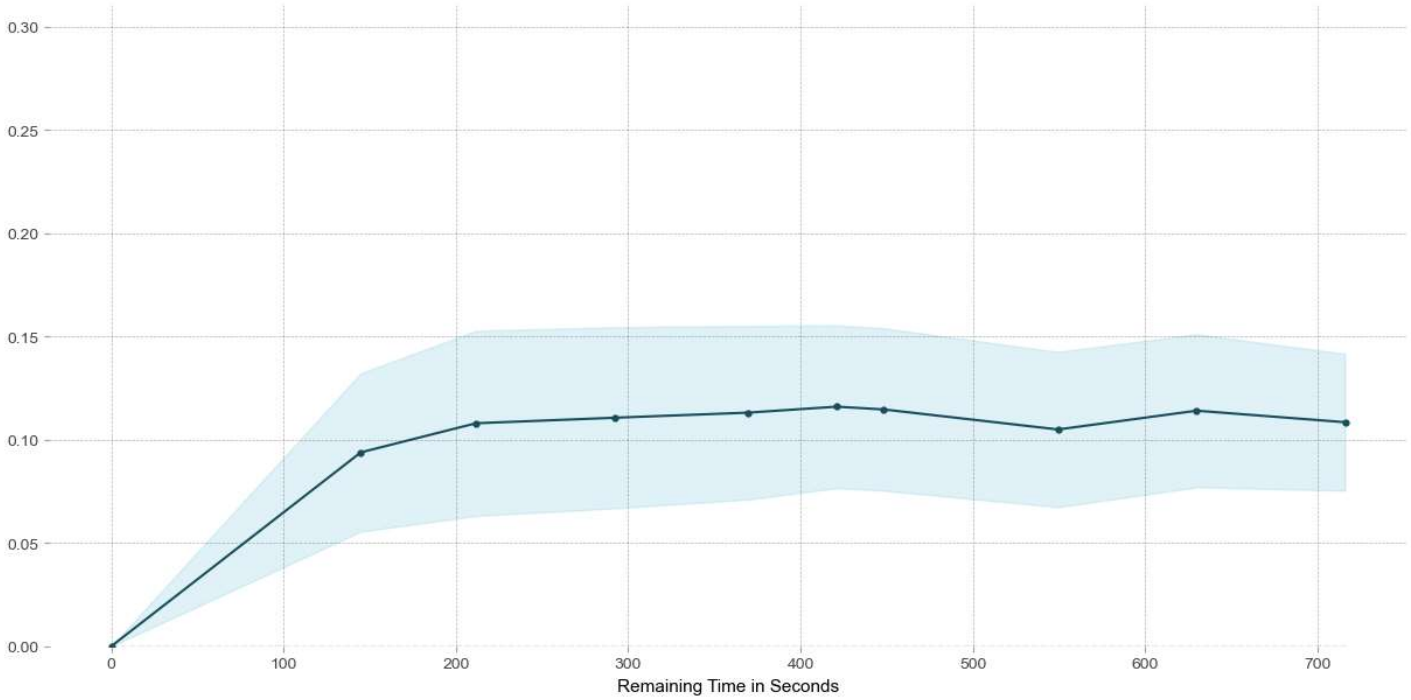
How does this work? Well, we select a single row from the dataset. On that row, we change the value of the variable we're interested in and see how it affects the prediction. We do that on more than one row (since interactions of variables for a single row result in a plot that is not generable) and note the average prediction for particular value $X = x$. For example, in the plot above it is apparent that as the score difference gets bigger the model thinks it's more likely for Alperen to make the 2 pointer (it levels off at around 10, there isn't much affect to predictions after that).



For shot distance, the model thinks Alperen is more likely to make shots around the basket and getting away from the basket after 3 to 4 ft. results in lower predictions. Partial dependence plots of other variables can be found below.

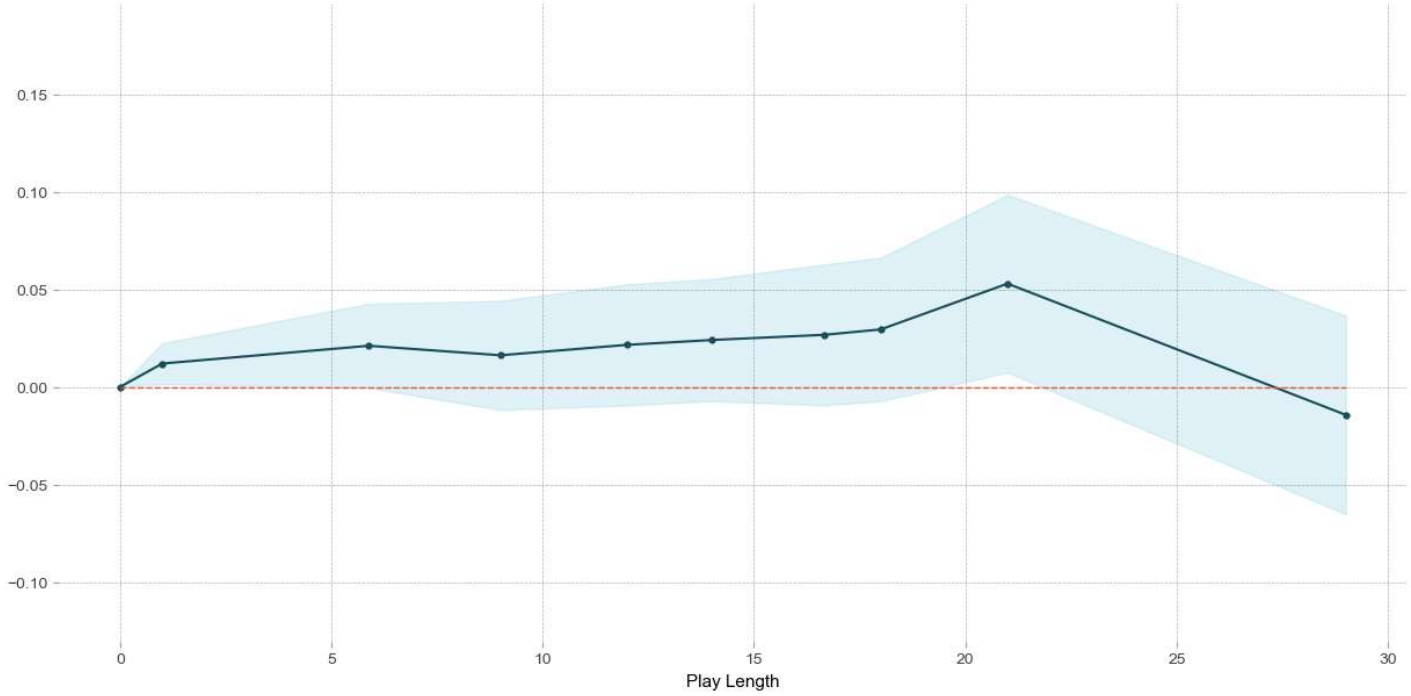
PDP for feature "Remaining Time in Seconds"

Number of unique grid points: 10

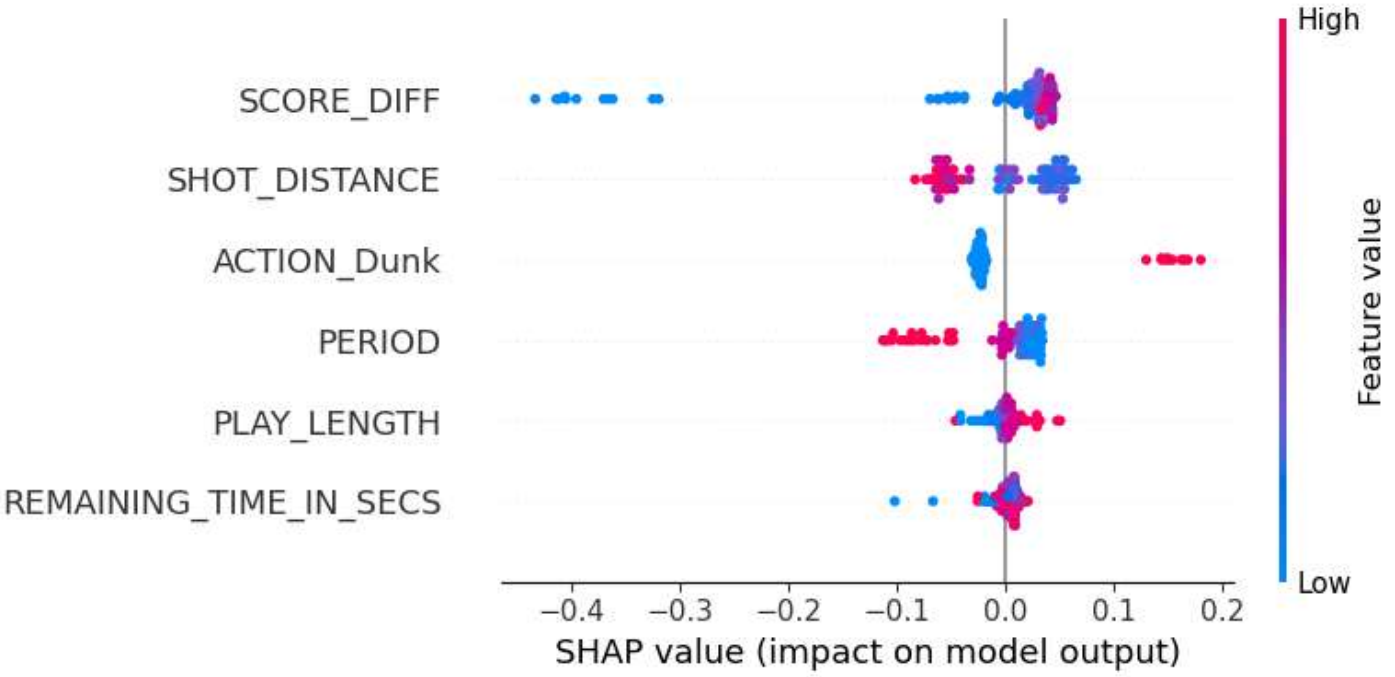


PDP for feature "Play Length"

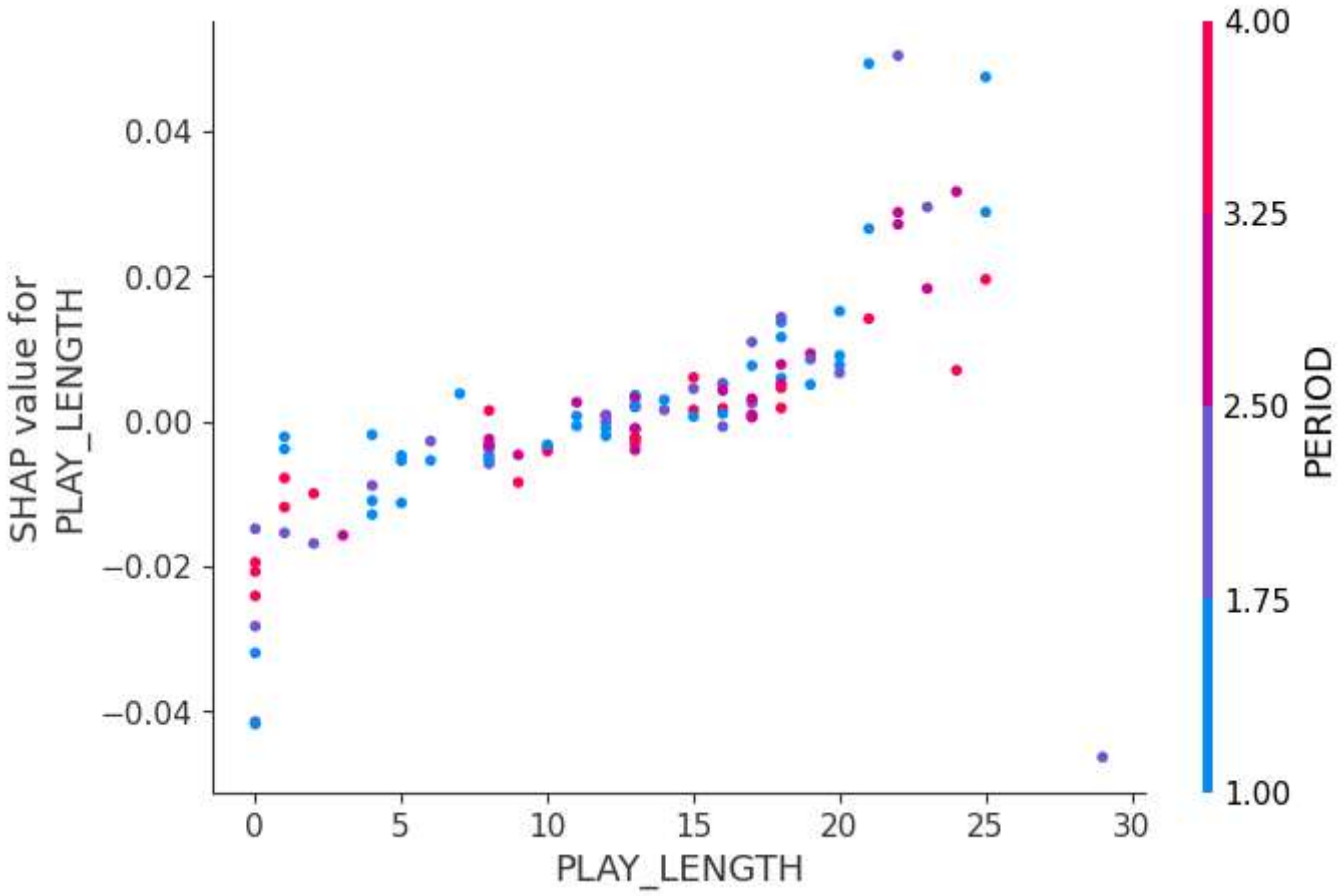
Number of unique grid points: 10



Furthermore, one can uses *shap* to summarise the effects of variables in one plot:



Vertical position in the graph corresponds to variables. Color of the dots gives us the value for that variable for that observation and horizontal location shows the effect of having that particullar value for that variable. Upside of this graph is one can understand how the variable importance forms: For instance, for score difference variable there is a group of observations on the very left side meaning that those observation might be pulling the variable importance up (additionally, when one does not consider them observations are still grouped in respect to that variable's values and their effect on the prediction). What I am trying to point out is when a variable has somehwat of a importance to a model it can be due to huge effect of few observations and nothing for the rest or it can be due to some moderate effect of many observations. What we have stated above in PDP section can be seen here in the row that corresponds to shot distance: low values resulted in higher prediction while high values resulted in lower prediction. Similar comments can be made for the period variable: Specifically 4th period seems to cause lower predictions.



This plot is similar to *partial dependence plot*, however, one can see interactions on this one: Effect of a value might depend on values of other variables (i.e., interaction).

First things first, since there is an upward trend, it can be said that as the playlength increases so does the prediction for shots to go in. However, there is spread vertically for same values of x-axis: For example, when x value (playlength value in the dataset) is around 21 (looking at 3 observations) the y value (effect of the x value to the prediction) is not the same. Taking a closer look, lower prediction corresponds to 4th period while the other two corresponds to first period.

Discussion

What else could have been done? I could have checked my assumptions under shot clock barchart by going to play by play data and grouping play length as it has been grouped for shot clock and see if my comments hold up (i.e., play length should be somewhat similar to shot clock). Additionally, I could have analyze free throws as well. Unfortunately, this whole analysis (specifically writing the comments) took more time than I was willing to spare so decided to stop at this very point. However, this is the first one as well which meant for me to write the codes from the ground up: On later ones, I'll just edit, adjust, tweak a bit which will make the process faster.

Anyway, I hope you enjoyed reading it. For more analysis you can follow me on instagram (@hooplytics).