Preliminary Analysis of Data

Sonia Xu

November 9, 2017

Checking the Data

Checking the data against ESPN's play-by-play shows that the SPORTSVU data matches the boxscores in their online datset. In the game explored–St. John's vs. Duke, Tyus has 22 points according to the play-by-play. When comparing Jahlil's blocked shots, the values also match. Thus, it appears as though the boxscores were correctly recorded.

Another problem was data completion, since not all the SportsVu keys were available publicly. Thus, via collaboration and discussion, below are the current keys that have been noted as the "truth." It is necessary to watch more film to confirm the validity of these event-id to event-description pairings.

In terms of combining the data, the best way to combine all the datasets is via the $TIME$ feature. There is no missing data in the final game data when merging the different datasets with this feature.

event-id	event-descrip
25	assist
24	blocked shot
6	defensive rebound
21	dribble
5	offensive rebound
22	pass

Initial exploratory analysis provides interesting insight onto the game. In terms of dribbling, defined by SportsVu as: • Player is closest to the ball. • Player is within 3.5 feet of the ball. • Player has been closest to ball for 5 or more frames. • Ball drops lower than 1.5 feet. • Ball is within the boundaries of the court. • Dribbles rely exclusively on optical data

We assume that the point guard will dribble the most. However, that is not the case for this specific game. As noted by the table, Tyus Jones is the 7th most frequent dribbler, whereas Rasheed Sulaimon, a post dribbles the most for Duke.

first.name	last.name	dribbles	drib_calc
Tyus	Jones	421	420
Rysheed	Jordan	251	250
Quinn	Cook	164	164
Phil	Greene IV	132	132
Sir'Dominic	Pointer	125	125
D'Angelo	Harrison	107	107

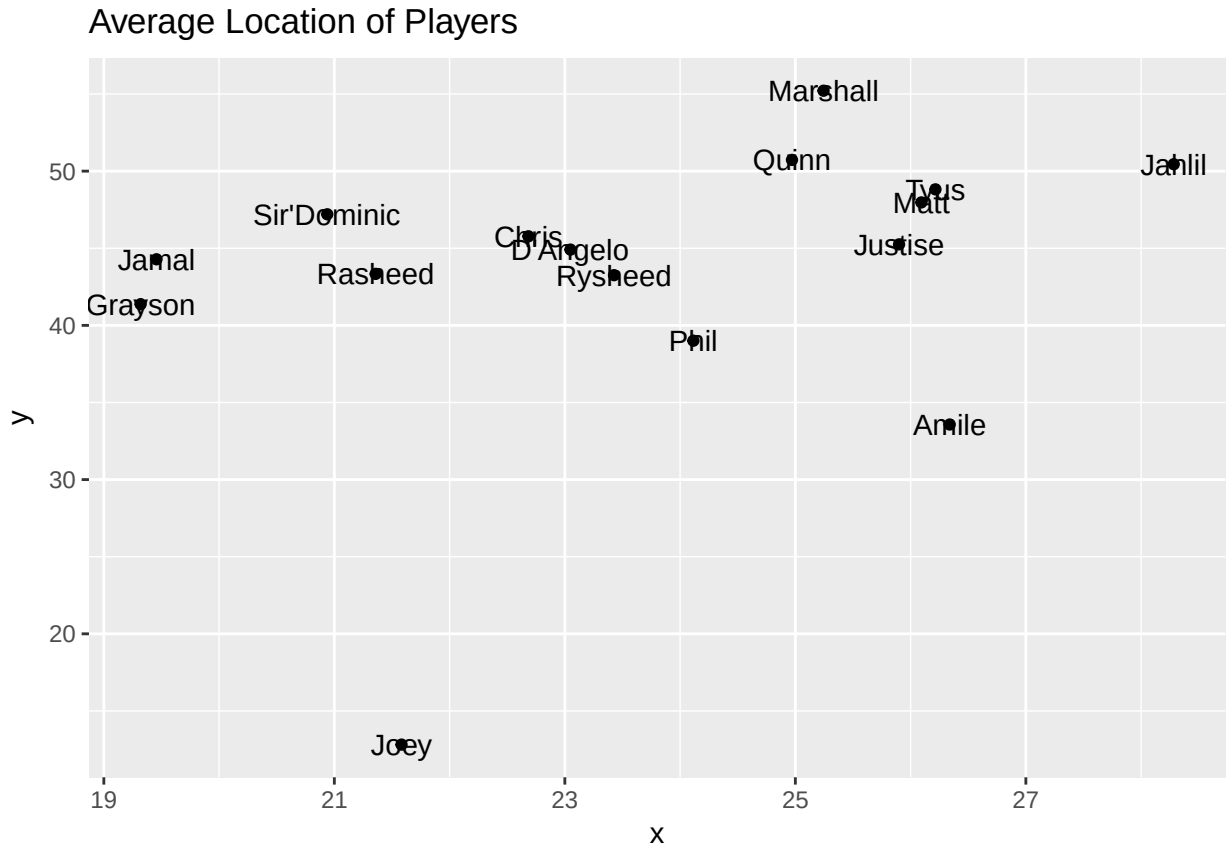
Counting the number of dribles via $event - id$ vs. the boxscore, the numbers also add up overall, for the difference between the two counts are off by a maximum of one dribble.

Plotting the Location of the Ball

Looking at the ball location in a 3-D graph, it is apparent that the ball spends less time in the air than close to the ground.

Plotting the Average Location of Players

Assuming that similar players (1-5) have similar average locations, we expect to see Grayson Allen in the same position as Matt Jones. However, since Grayson did not play a lot of minutes until after his freshman season, perhaps this skews the data. Jahlil, Amile, and Marshall appear to be the farthest away. Perhaps, Duke had greater spacing in their playstyle across the court compared to St. John's.



Other Interesting Things

Most missed shots occured with a lot of time left in the shot clock:

first.name	last.name	shot_time
Amile	Jefferson	21.21333
Chris	Obekpa	22.28500
D'Angelo	Harrison	21.79778
Grayson	Allen	31.56000
Jahlil	Okafor	29.70500
Jamal	Branch	22.41000

On the other hand, made shots occured around a similar time, but there is more variance with how long it takes for a player to take a good shot.

first.name	last.name	shot_time
Amile	Jefferson	23.64200
Chris	Obekpa	30.00000
D'Angelo	Harrison	26.49600
Jahlil	Okafor	28.71286
Marshall	Plumlee	18.00000
Matt	Jones	15.24500
‘‘‘		

Next steps, I will attempt to plot the x,y coordinates of each player as a heatmap to see where most players prefer to shoot the ball on a game-level and on a season-level to see how players change their shooting positions throughout the years. Do they move farther from the basket or stay consistent with their range?