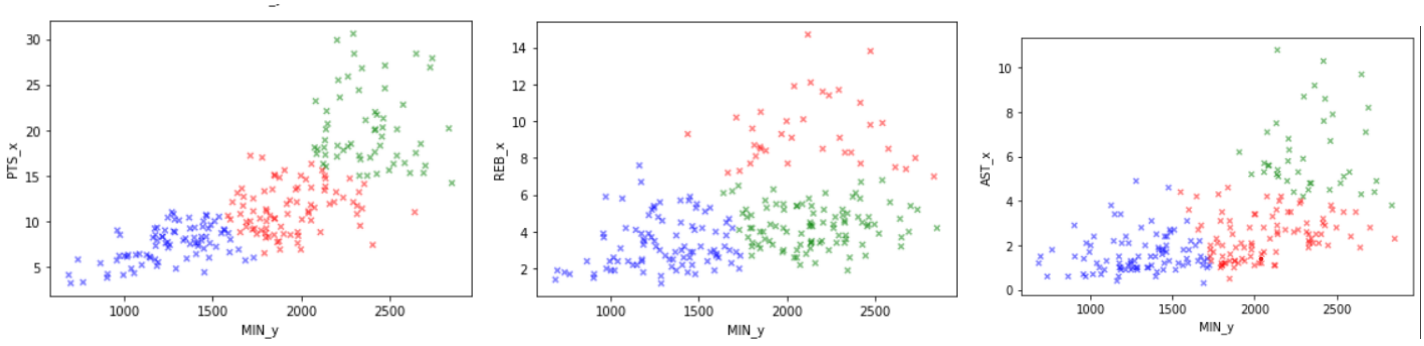
Elijah Malabuyo
DATA 301

As a fan of basketball and the NBA, I wanted to analyze NBA data to see if there were any interesting insights that I could uncover with player statistics. I began by first collecting the data from the NBA.com statistics page. I was able to utilize their free API to retrieve a json object that provided a list of players per game averages from the season of 2021-2022. I then utilized a second API from their page to collect a second json object that had a list of players totals from the season. The two datasets I created were one of player averages and one of player totals. During the cleaning process, I decided to merge the two datasets together to get a whole dataset; however, during this process the two dataframes were of different length. After more research on the player averages API, the NBA.com statistics page calculated averages based on effective minutes played, therefore players with minutes averaged below 10 were considered "not impactful" and removed. Some players had multiple entries, but I decided to keep their duplicates because several players played on different teams during the season. I also kept values of 0 for all quantitative variables. I created a function to create another categorical variable for each player labeled "ALLSTAR", 1 for yes the player is an allstar and 0 for a non-allstar. This categorical variable will be utilized later in a classification model. Before I clustered and classified my data, I also standardized my variables through scikit-learn's standard scaler.

I wanted to focus more on offensive statistics therefore I found which offensive stats had more impact on the game by looking at the correlation with total \minutes played. I found that the three categories with the highest correlation to total minutes played were points, assists, and rebounds.

After processing these offensive stats I wanted to utilize these three categories to assign new roles to players based on their play style. Traditionally there are 5 positions in basketball, however, a recent report from an ex-NBA Coach stated, "It (Basketball) may be as simple as three positions now, where you're either a ball-handler, a wing or a big". I wanted to determine if there was any truth behind this statement. Therefore I utilized a K=3 means clustering to cluster players based off of the three impactful in-game stats. Below shows a scatterplot of the attributes group from the k means cluster result.



The green cluster appeared to have higher statistics in the PTS and AST category. The red cluster had higher statistics in the REB category. The blue cluster had the lowest average statistics for all three categories. After sampling several players from the cluster, I was able to label a new position for these three clusters.

GREEN PLAYER = Star Player
Player that performs well on all three stats
Sample Green Player

	PLAYER_x	PTS_x	AST_x	REB_x
2	Luka Doncic	28.4	8.7	9.1

RED PLAYER = BigMan
Player that is more rebound dominant
Sample Red player

	PLAYER_x	PTS_x	AST_x	REB_x
56	Rudy Gobert	15.6	1.1	14.7

BLUE PLAYER = Role Player
Player that adapts to the other teammates play and is utilized for multiple roles
Sample Blue player

	PLAYER_x	PTS_x	AST_x	REB_x
224	Juan Toscano-Anderson	4.1	1.7	2.4

The result from KMeans clustering and its interpretations are quite interesting and provide some statistical backing to the NBA coach's statement.

Next, I wanted to formulate an alternate all-star team utilizing the 2021-2022 all star statistics. I utilized a K-nearest neighbor classification to find the closest neighbors to each of the all star players to find another player with similar stats. Utilizing the custom categorical variable ALLSTAR, I trained on the seven averages per game in the following categories: PTS_x", "AST_x", "REB_x", "STL_x", "EFF_x", "BLK_x", "FG_PCT_x" (note that the x at the end indicates season averages while a y at the end indicates totals for the season). These seven categories are evaluated along with public votes to determine all-star players throughout the season (NBA.com).

```
Allstar: Joel Embiid REPLACEMENT: Nikola Vucevic
Allstar: Giannis Antetokounmpo REPLACEMENT: Domantas Sabonis
Allstar: Luka Doncic REPLACEMENT: Russell Westbrook
Allstar: Trae Young REPLACEMENT: D'Angelo Russell
Allstar: DeMar DeRozan REPLACEMENT: CJ McCollum
Allstar: Nikola Jokic REPLACEMENT: Domantas Sabonis
Allstar: Jayson Tatum REPLACEMENT: Julius Randle
Allstar: Devin Booker REPLACEMENT: Khris Middleton
Allstar: Donovan Mitchell REPLACEMENT: Terry Rozier
Allstar: Stephen Curry REPLACEMENT: Terry Rozier
Allstar: Karl-Anthony Towns REPLACEMENT: Christian Wood
Allstar: Darius Garland REPLACEMENT: Kyle Lowry
Allstar: Dejounte Murray REPLACEMENT: Chris Paul
Allstar: Fred VanVleet REPLACEMENT: Jrue Holiday
```

I formulated a new dataset from these results to showcase the alternate all-star team. It is important to note that for the all-star players Devin Booker and Dejounte Murray their nearest neighbors were already considered on the all-star team as their stats are similar in the model categories. I evaluated this classification model to get an F1 score of 0.9788.

From this k-nearest neighbor classification model, it was interesting to see which players could have been All-Stars as well during the 2021-2022 season. Through the in-game stats considered by the NBA, we can see that public fan voting plays a significant part in this decision process and is likely biased. While the true All-Star players perform at a high level, there are also other players in the NBA with similar in-game averages that should be given the chance and recognition in the league.