NBA Data Description

Ben Moolman, Craig Orman, Ethan Pross

Data Set Description:

- Obtained from the Kaggle Dataset NBA Boxscore - Season 2023 / 2024 by Alberto Filosa[1]
- There are 30 columns and 12300 rows. 7 columns are identifiers, 1 column is a row index, 2 are characteristic and comment columns, and then we have 20 numerical statistic columns.
- Each row significes a unique tuple of (game, player), that is to say there is a row for stats of each player in every game.
- There are 1230 unique games, 30 unique teams, and 392 unique players.
- There is no missing data because we are only analyzing rows where the player was a starter, and therefore always played and stats were recorded.
- Presumably, missing values in the starting position column refer to a non-starter player.

Variable	Description
Game_ID	Unique Identifier of the Game.
Team_ID	Unique Identifier of the Team. For Each GAME_ID there are only two TEAM_ID.
Player_ID	Unique Identifier of the Player.
FGA	Number of Field Goals (both 2 and 3 Points) Attempted by the player in the game.
FG_PCT	Percentage of Field Goals (both 2 and 3 Points) by the player in the game.
FG3A	Number of 3 Points Attempted by the player in the game.
FG3_PCT	Percentage of 3 Points by the player in the game.
REB	Number of Total Rebounds (Defensive and Offensive Rebounds) reached by the player in the single game.
AST	Number of Assists done by the player in the single game.
PTS	Number of Points done by the player in the single game.

Prior Analysis on this dataset

- There appears to be one analysis submitted for this Kaggle dataset, Done by Adam Briggs, they seem to have done mostly visuals and descriptive statistics, looking at best player on the team, and top performing player in various metrics.
- Basketball analytics is a very common field though, and there are entire websites dedictated to it such as one by Evan Miyakawa who launched their website as part of a PhD program with Baylor. Usually, analytics attempt to gain an edge in the industry by identifying undervalued players or by predicting the outcomes of games.

[1]NBA Boxscore - Season 2023 / 2024 by Alberto Filosa, https://www.kaggle.com/datasets/albi9702/nba-boxscore-season-2023-2024?resource=download

Important notes

- Missing data sampling can be done on games where the player did not play, using priors to consider the effect of previous games and the team.
- Data Analysis is on the raw data. We will be focusing on starting players.

Variable	Missing	Mean	SD	Median	Min	Max
FGM	0	5.8875610	3.3779162	5.000	0	25
FGA	0	12.2078862	5.9179516	12.000	0	47
FG_PCT	0	0.4805760	0.1799890	0.500	0	1
FG3M	0	1.6898374	1.6844233	1.000	0	12
FG3A	0	4.5466667	3.2885738	4.000	0	23
FG3_PCT	0	0.3090527	0.2646441	0.333	0	1
FTM	0	2.4943089	2.7565473	2.000	0	24
FTA	0	3.1325203	3.2594904	2.000	0	32
FT_PCT	0	0.5654188	0.4164934	0.667	0	1
OREB	0	1.3016260	1.5429185	1.000	0	15
DREB	0	4.3720325	2.9168990	4.000	0	20
REB	0	5.6736585	3.7084151	5.000	0	31
AST	0	3.7410569	2.9722169	3.000	0	23
STL	0	0.9500813	1.0429180	1.000	0	7
BLK	0	0.6769919	1.0118289	0.000	0	10
TO	0	1.7782927	1.5398969	1.000	0	11
PF	0	2.2308130	1.4528870	2.000	0	6
PTS	0	15.9592683	9.0763652	15.000	0	73
PLUS_MINUS	0	0.2706504	13.6022566	0.000	-58	46

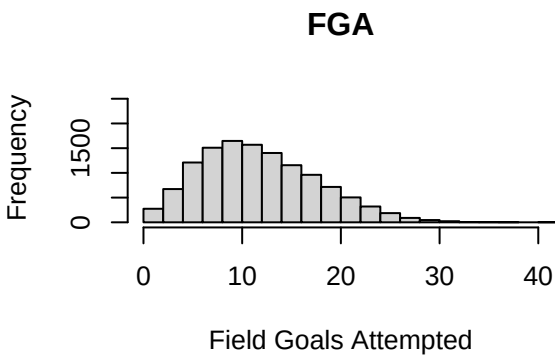
Exploratory Data Analysis

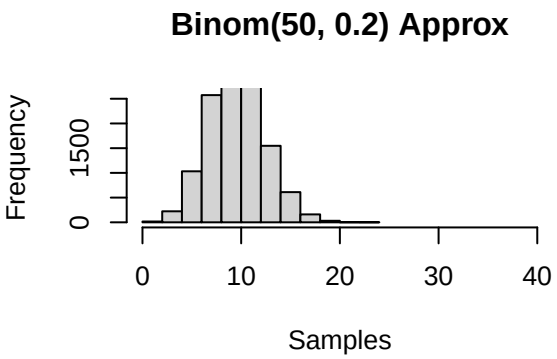
- Most of the numeric fields are right skewed
- FGA follows a binomial but with a heavier right tail

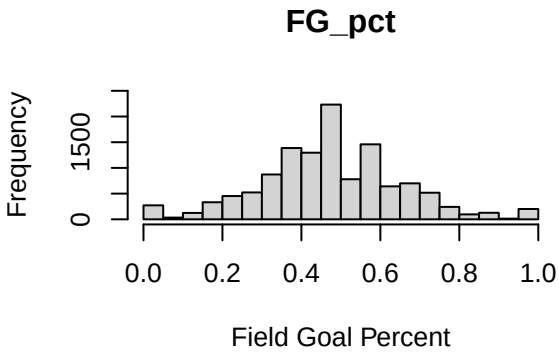
Sampling models

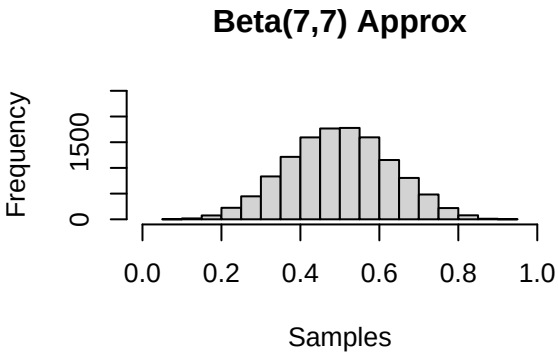
- Two variables of interest to predict are Field Goal Attempts and Field Goal Percent. Given below, we will attempt to model these with the beta and binomial respectively.

```
par(mfrow=c(2,2))
y1 = 2600
hist(original_tbl$FGA[original_tbl$START_POSITION != ""], breaks = 20, xlim=c(0,40), ylim = c(0, y1), ma
hist(rbinom(12300, 50, 0.2), xlim=c(0,40), ylim = c(0, y1), main="Binom(50, 0.2) Approx",  xlab= "Sample
y2 = 2500
hist(original_tbl$FG_PCT[original_tbl$START_POSITION != ""], breaks = 20, ylim=c(0, y2), main="FG_pct",
hist(rbeta(12300, 7, 7), xlim=c(0,1), ylim=c(0, y2), main="Beta(7,7) Approx",  xlab= "Samples")
```


FGA
Frequency
0
1500
0
10
20
30
40
Field Goals Attempted


Binom(50, 0.2) Approx
Frequency
0
1500
0
10
20
30
40
Samples


FG_pct
Frequency
0
1500
0.0
0.2
0.4
0.6
0.8
1.0
Field Goal Percent


Beta(7,7) Approx
Frequency
0
1500
0.0
0.2
0.4
0.6
0.8
1.0
Samples

Appendix

Variable	Description
GAME_ID	Unique Identifier of the Game.
TEAM_ID	Unique Identifier of the Team. For Each GAME_ID there are only two TEAM_ID.
TEAM_ABBREVIATION	Abbreviation of the Team (e.g. GWS - Golden State Warriors).
TEAM_CITY	NA
PLAYER_ID	Unique Identifier of the Player.
PLAYER_NAME	Complete Name (Name and Surname) of the Player who played the game.
NICKNAME	Nickname of the Player who played the game.
START_POSITION	Position in which the player started the game (If populated, the player started the game).
COMMENT	NA
MIN	Number of minutes in which the player played the game.
FGM	Number of Field Goals (both 2 and 3 Points) Made by the player in the game.
FGA	Number of Field Goals (both 2 and 3 Points) Attempted by the player in the game.
FG_PCT	Percentage of Field Goals (both 2 and 3 Points) by the player in the game.
FG3M	Number of 3 Points Made by the player in the game.
FG3A	Number of 3 Points Attempted by the player in the game.
FG3_PCT	Percentage of 3 Points by the player in the game.
FTM	Number of Free Throws Made by the player in the game.
FTA	Number of Free Throws Attempted Made by the player in the game.
FT_PCT	Percentage of Free Throws by the player in the game.
OREB	Number of Offensive Rebounds reached by the player in the single game.
DREB	Number of Defensive Rebounds reached by the player in the single game.
REB	Number of Total Rebounds (Defensive and Offensive Rebounds) reached by the player in the single game.
AST	Number of Assists done by the player in the single game.
STL	Number of Steals done by the player in the single game.
BLK	Number of Blocks done by the player in the single game.
TO	Number of Turnovers done by the player in the single game.

Variable	Description
PF	Number of Personal Fouls done by the player in the single game.
PTS	Number of Points done by the player in the single game.
PLUS_MINUS	Number of Plus Minus done by the player in the single game.

```
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
# Comes out as all NAs for some reason
no_na_numerics_tbl = na.omit(original_tbl[,12:30])
correlation = cor(no_na_numerics_tbl, method = "pearson")
par(mfrow=c(1,1))
corrplot(correlation) # colorful number
```