Offense vs Defense: An Analysis in the NBA's Play Style

Team 22

2025-11-07

Abstract

Dunk or get dunked on — that's the name of the game when it comes to the NBA. You have twenty-four seconds to sink a shot or get one put down on your hoop, and currently the rules have seemed to strongly favor offenses and penalize physical defenses. Does that mean defense is less useful now, or does defense still play just as crucial a part and simply look different from what we expect? This paper seeks to investigate what efficient defenses and offenses look like by identifying key metrics that aggregate to describe efficient play, and ultimately analyze what wins games more, scoring points or allowing points. To conduct this analysis, we will develop models describing offensive and defensive efficiency and apply nested F-tests across the models and an aggregate to identify which contributes more to winning games. (Insert findings later)

Introduction

The NBA's rulebook since 2004 and ongoing regularly makes updates that drive the game to at least seem both extremely offense-driven and defense-second, with the latter seemingly being a mere afterthought if not outright not integrated into defensive philosophies. This perception of the game may have a strong sense of truth, but there is a very viable possibility that it only describes half the game and leaves sorely underdeveloped defensive outlook and strategy. To bring the best out of players' technical abilities, strengthen and develop the NBA-seeking talent pool with both the players and aspiring coaches, and help the NBA tangibly understand how to fight now-rampant criticisms of the NBA 'going soft', it is crucial to better understand the current defensive structure of the game and how to best develop it. By analyzing which key metrics contribute to defensive or offensive efficiency and developing models to understand how much each throughput affects winning outcomes, we can create this better understanding of the flipside to the offensive-minded game style. In order to best conduct this analysis, we will break the analysis into two phases.

For the first phase of analysis, we will first categorize the variables into defensive and offensive categories. Once we class the variables, we will test for intercategorical collinearity and develop a full model that selects one variable per collinear relationship to avoid redundancy. Once the full model is developed, we will employ techniques like backwards elimination to develop statistically significant offensive and defensive models that explain a team's win within a game accurately.

For the second phase of analysis, we will then create an aggregate model from the two initial models that is computed as a composite. Using all three models the, we will conduct nested F-tests to determine whether the offensive model or defensive model provides the stronger signal within the composite.

Data

In order to conduct this analysis on NBA gameplay statistics, we are using regular season data (original data) compiled over a time range from 2010-2024 (Korolyk, 2024). This dataset is compiled by Vasilii Korolyk and is publicly available at https://github.com/NocturneBear/NBA-Data-2010-2024 for academic use under its

MIT License. The original data contains over 33,000 observations and documents 57 variables over each entry.

Within the raw data given, one of the variables was not a statistic and was in fact a helpful utility called AVAILABLE_FLAGS, which indicated whether the data was healthy enough for use. As a result, when cleaning the data we initially dropped all entries that didn't have a value of 1, which indicated they were healthy. After dropping those entries, we then removed all variables that have no bearing on the intended research or methodologies along with fully-empty rows. After cleaning, we're left with around 28K observations and 27 variables. The variables for the fully cleaned data are as follows

Dimensions

- Season
- Team ID
- Team Abbreviation
- Team Name
- Game ID
- Game Date
- Matchup
- Win/Loss Status

Metrics

- Field Goals Made
- Field Goals Attempted
- Field Goal Accuracy
- Three Pointers Made
- Three Pointers Attempted
- Three Pointer Accuracy
- Free Throws Made
- Free Throws Attempted
- Free Throw Accuracy
- Offensive Rebounds
- Defensive Rebounds
- Total Rebounds
- Assists
- Turnovers
- Steals
- Blocks
- Blocks Against
- Personal Fouls
- Personal Fouls Drawn

Before we can use this data though, note something of key importance — three point shots are counted as field goals, since they are a type of field goal along with two-point shots. Mathematically,

$$\text{FG} = \text{3P} + \text{2P}$$

From here we can derive formulas to develop two-point metrics as follows

$$\text{2PA} = \text{FGA} - \text{3PA},\ \ \text{2PM} = \text{FGM} - \text{3PM},\ \ \text{2P\%} = \frac{\text{FGM} - \text{3PM}}{\text{FGA} - \text{3PA}}$$

After computing these derived metrics, general field goals is essentially rendered a composite of two available metrics, and hence obsolete. As a result, we drop the original field goal metricsm finally leaving us with a dataset of 28K observations and 27 variables named like so:

Dimensions

- Season
- Team ID
- Team Abbreviation
- Team Name
- Game ID
- Game Date
- Matchup
- Win/Loss Status

Metrics

- Three Pointers Made
- Three Pointers Attempted
- Three Pointer Accuracy
- Free Throws Made
- Free Throws Attempted
- Free Throw Accuracy
- Offensive Rebounds
- Defensive Rebounds
- Total Rebounds
- Assists
- Turnovers
- Steals
- Blocks
- Blocks Against
- Personal Fouls
- Personal Fouls Drawn
- Two Pointers Made
- Two Pointers Attempted
- Two Pointer Accuracy

As a final preliminary to EDA and data visualization, we will also categorize the metrics between offense and defense as follows along with rationale:

Offensive Metrics

- Three Pointers: Point-scoring event
- Free Throws: Point-scoring event
- Offensive Rebounds: Initiates offensive play via repossession on offensive drive
- Assists: Directly enables and precedes a point-scoring event (two/three pointers)
- Turnovers: Offensive loss of possession
- Two Pointers: Point-scoring event
- Personal Fouls Drawn: Enables a point-scoring event (free throw)
- Blocks Against: Disrupts a point-gaining opportunity for team

Defensive Metrics

- Defensive Rebounds: Initiates possession after defensive drive
- Steals: Interrupts opponent's offensive drive and initiates possession from defensive play
- Blocks: Disrupts a point-gaining opportunity for opponent
- Personal Fouls: Penalty for illegal defensive maneuver

EDA & Data Visualization

When computing models using the dataset, we must first analyze any potential colinearities between variables to identify potential redundancies in model construction prior to regression and pruning. To do so, we may produce a correlation matrix and run colinearity tests on variables of concern. When checking for colinearity, we will use the typical threshold of

$$|\rho| > 0.7$$

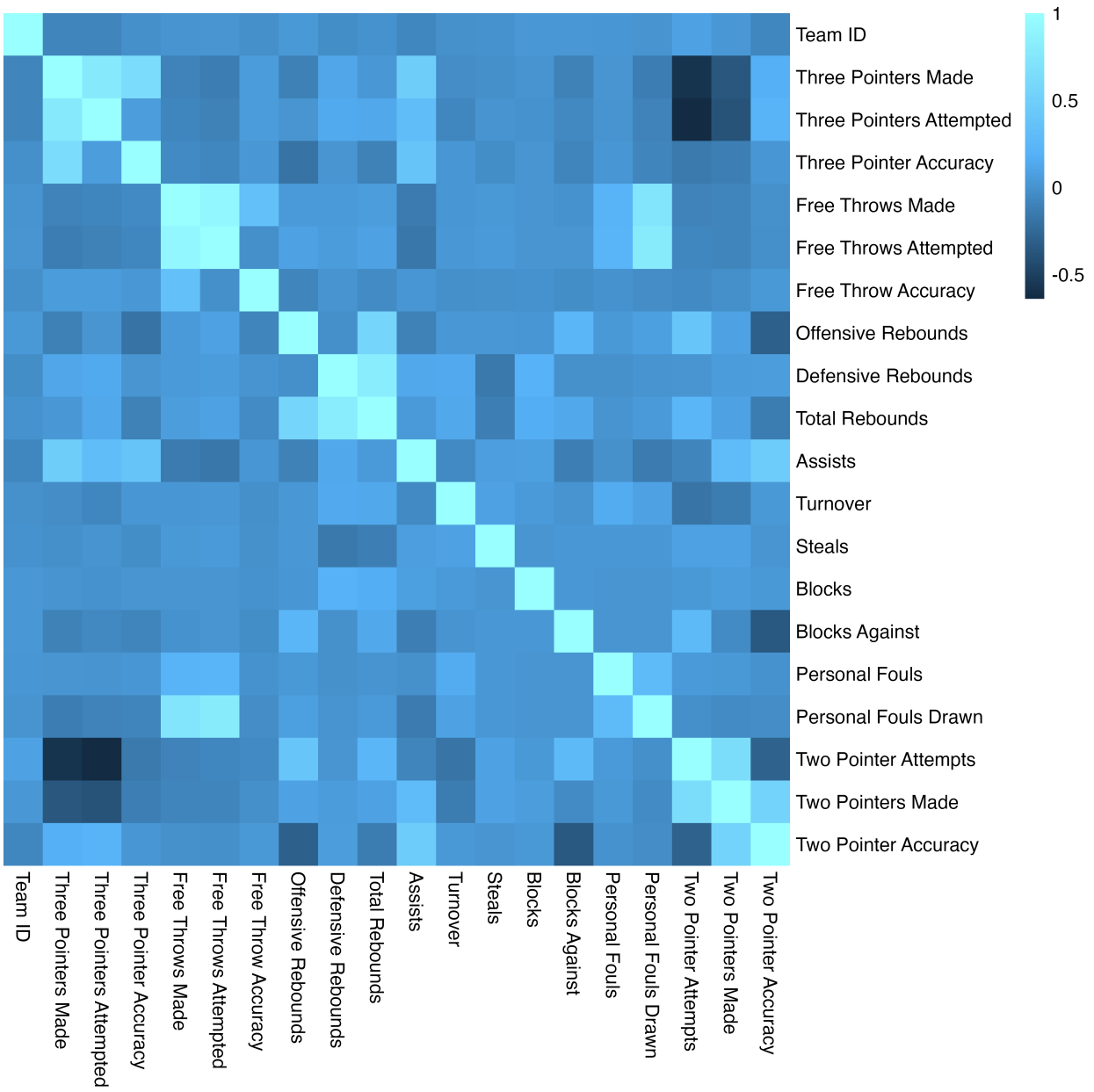
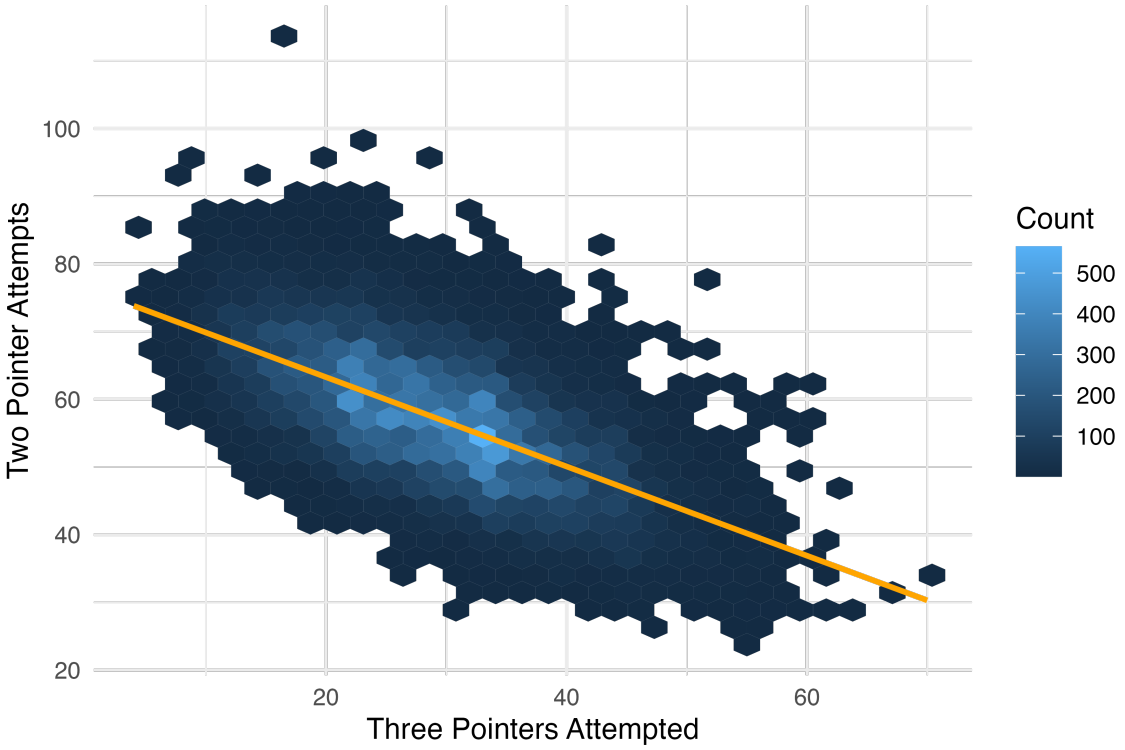


Figure 1.1: Correlation Matrix

When viewing the correlation heatmap (Figure 1.1) for the dataset, we must keep in mind the symmetry about the principal diagonal due to the mirroredness of the graph's structure. Furthermore, cells along the diagonal will trivially show and high colinearity due to this structure and how certain variables are calculated. With this in mind, only intensely-colored cells and clusters for intracategorical variables (i.e. three point and two point shots since both are offensive metrics) will be taken for concern. For simple verification, we may also want to test for colinearity for simple verification where there is a concern in an overlap of the space — practically and mathematically — certain variables are produced in (i.e. three point and two point shots since both are attempted during a drive as binary shot choices).

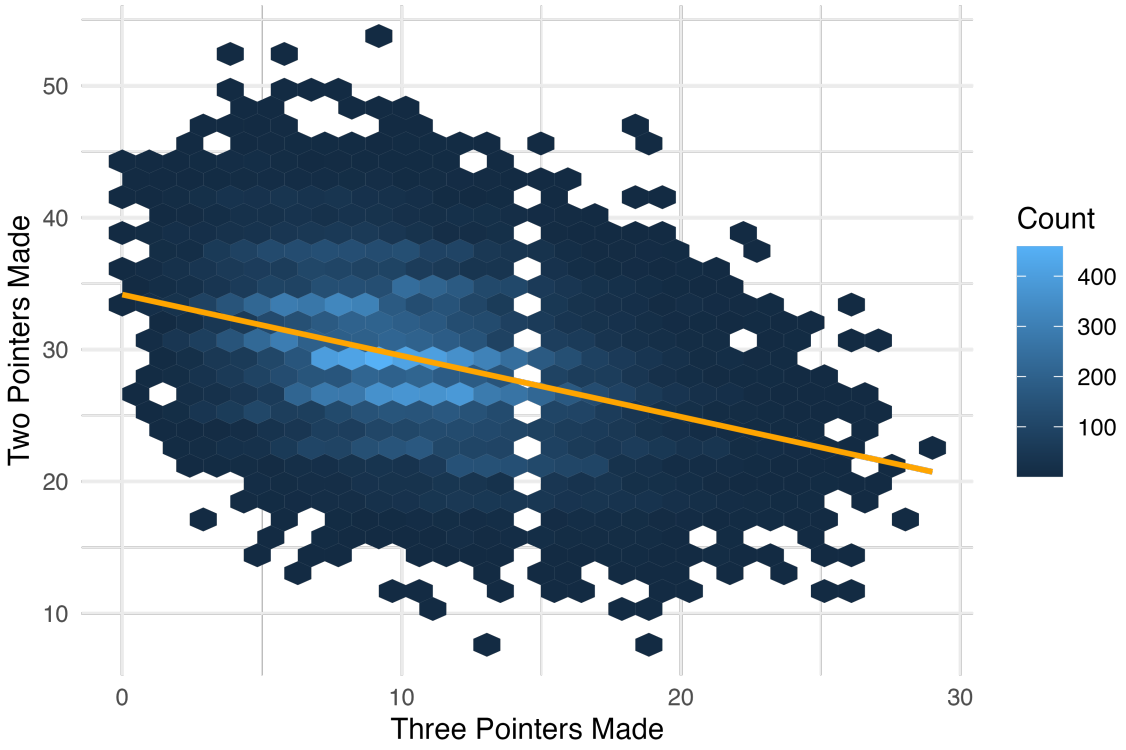
We first notice a dark cluster of cells between three-point and two-point attempts, so we check these two metrics for colinearity.



$$\rho = -0.64$$

$\rho = -0.64$ does not meet the required magnitude threshold to qualify as sufficiently colinear, both will be accounted for when constructing the offensive play model prior to F-testing.

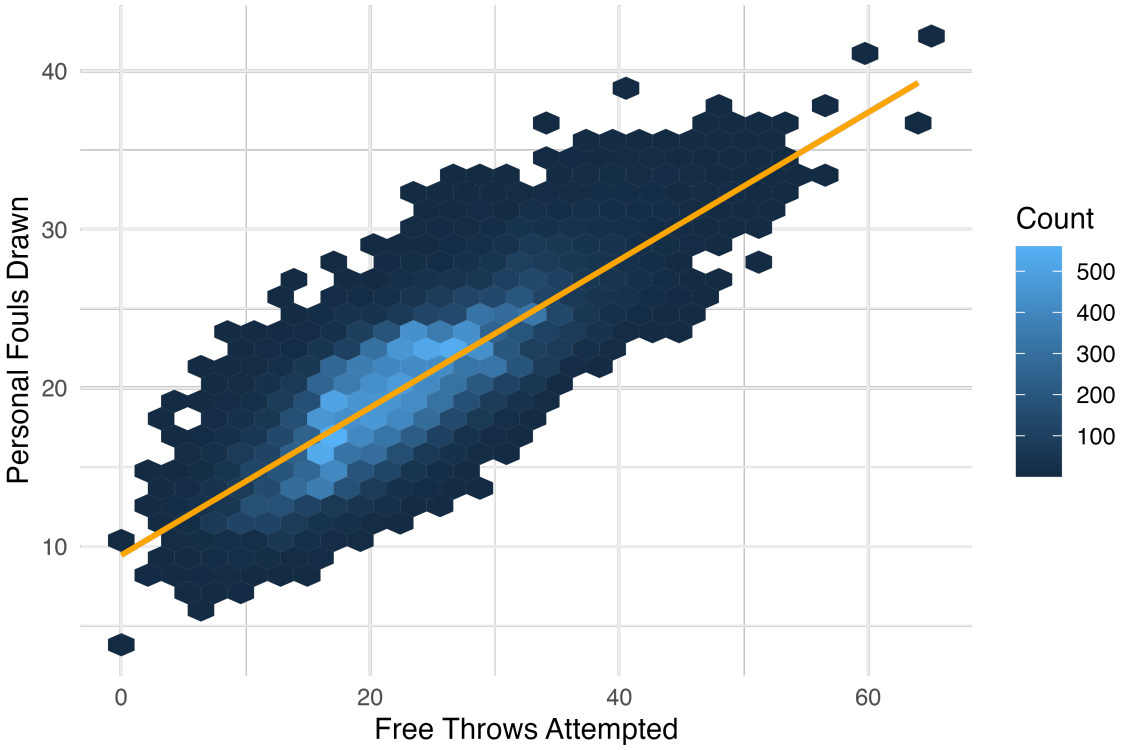
Similarly, ince three pointers and two-point shots are necessarily made during offensive drives, there is concern about colinearity between the successful attempts since they occupy the same production space. As a result, we will test for colinearity between the two.



$$\rho = -0.368$$

The plot indicates a very weak linear relation, and ρ fails the threshold test for colinearity. As a result, both variables can contribute to an initial model before backward elimination.

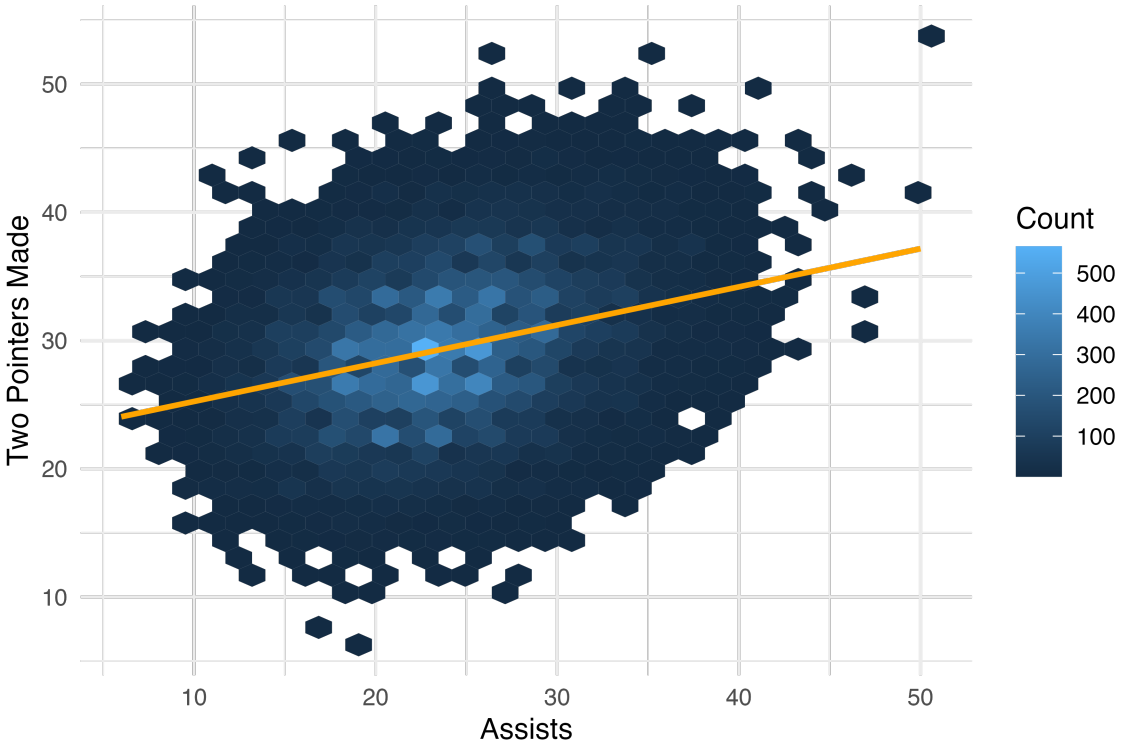
Due to the current NBA rules, personal fouls very frequently enable opponents to shoot free throws. Due to the confoundingness in the production of such metrics, colinearity must be tested.



$$\rho = 0.79$$

The plot indicates a very strong linear relation, with ρ meeting the threshold test for colinearity. Due to this, we must choose only one when constructing the full offensive model.

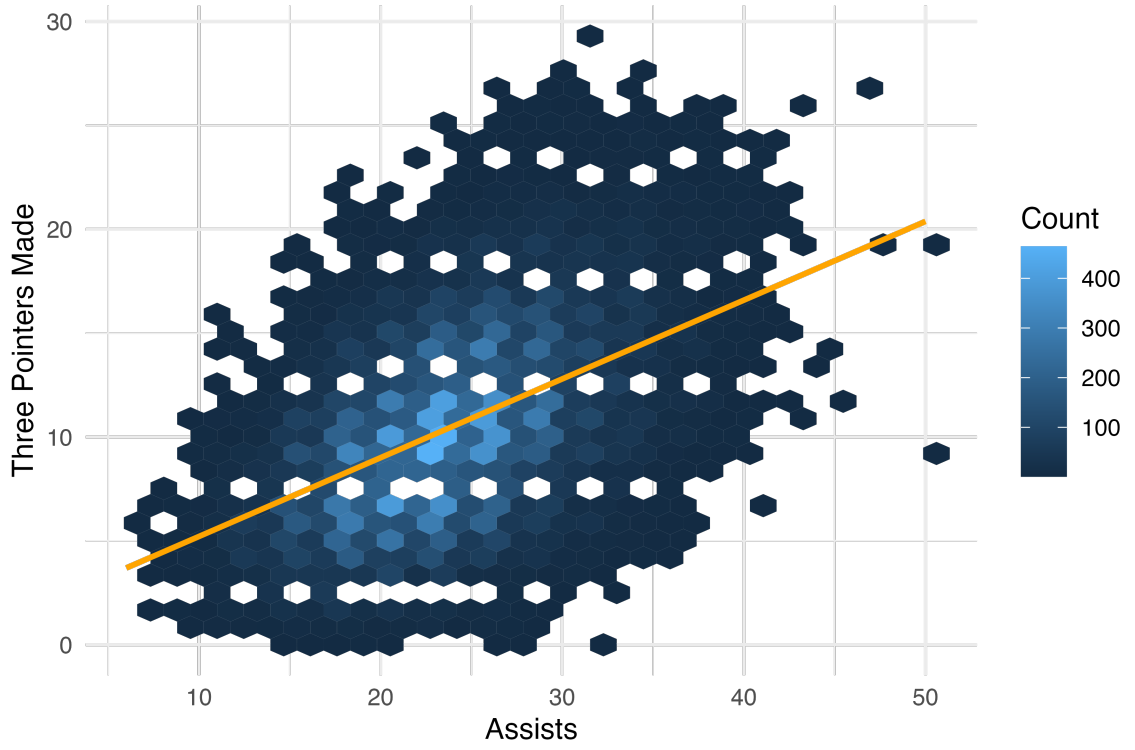
Similarly, assists necessarily precede a successful two-point attempt, which may be concerning when considering colinearity. Due to this concern, we will conduct a test for colinearity.



$$\rho = 0.298$$

The above plot displays a very weak linear relation, and furthermore fails the colinearity test. Consequently, both variables can contribute to the initial offense model prior to backward elimination.

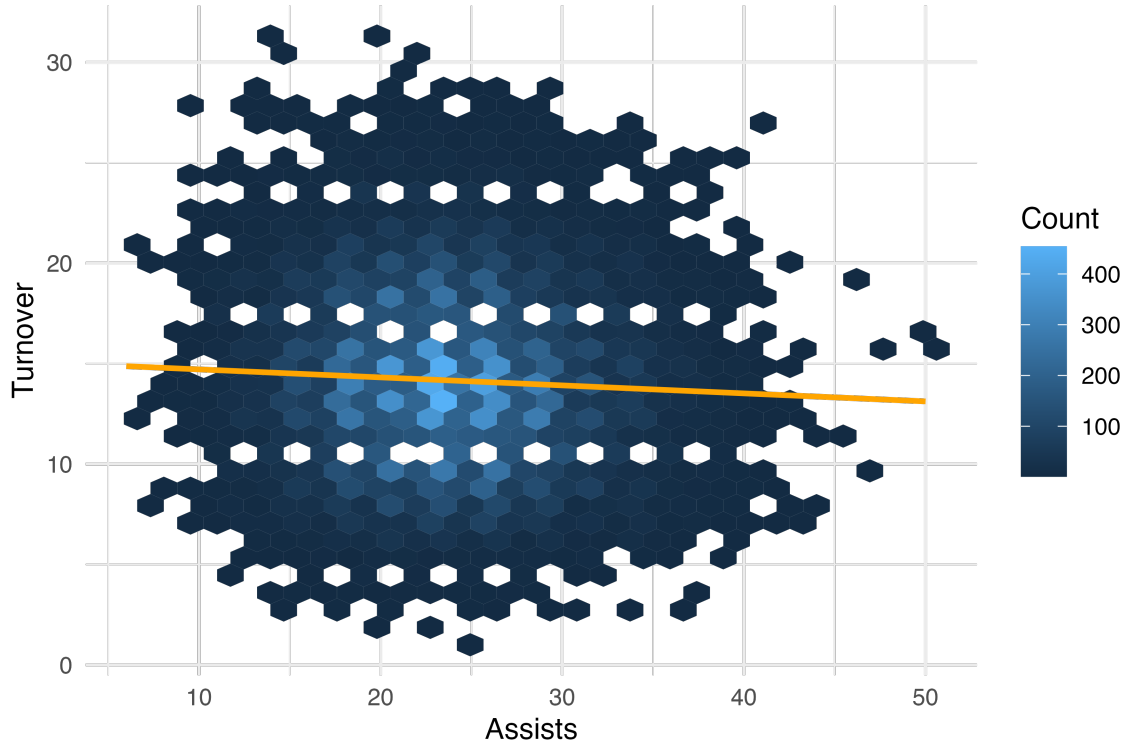
In much similar fashion to the former, assists also necessarily precede successful three-point shots, so we move ahead with colinear analysis.



$$\rho = 0.478$$

The plot indicates a very weak linear relation, and $\rho = 0.478$ fails the threshold test for colinearity. As a result, both metrics can contribute to an initial model before backward elimination.

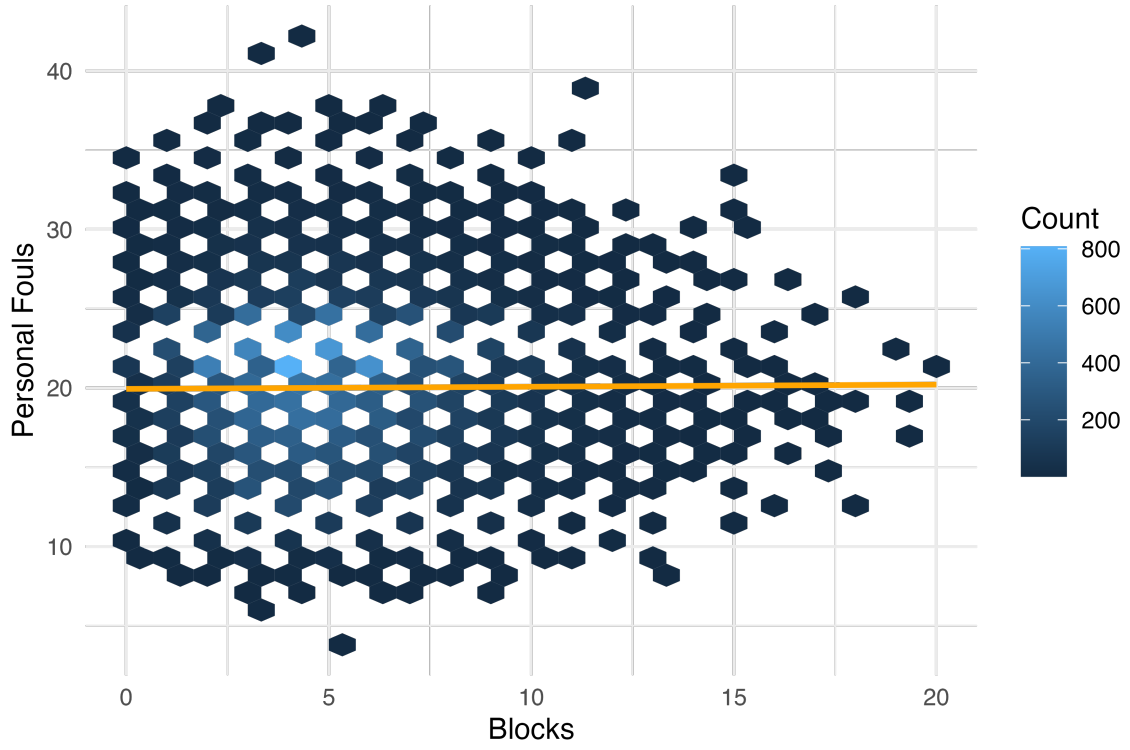
Turnovers and assists may possibly be confounding variables, especially due to the new 2024-2025 season rules that have strengthened the transition game and enabled very aggressive counter-offense. As a result, we will move forward with colinear testing.



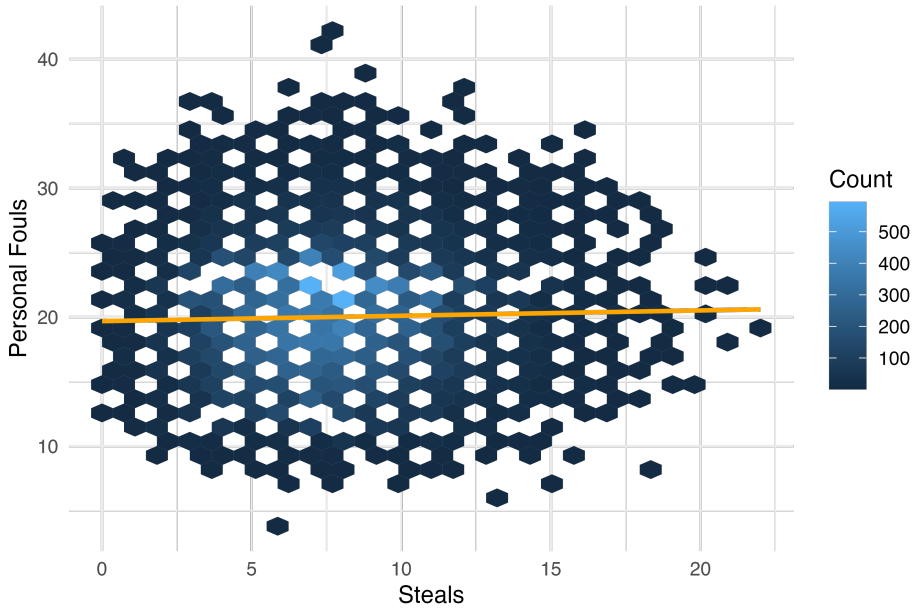
$$\rho = -0.053$$

The plot once again indicates a very weak linear relation, and $\rho = -0.053$ fails the threshold test for colinearity. As a result, both variables can be included in the offensive model before backward elimination.

In regards to defensive concerns, personal fouls are incurred during illegal physical contact with an offensive player. As a result, we must test both blocks and steals against personal fouls because both necessarily put the defensive player at risk of an illegal maneuver that constitutes the latter.



$\rho = 0.008$



$\rho = 0.028$

Both metrics have extremely low correlation magnitudes, and as a result, all three of these variables will be included in the full defensive model prior to pruning.

Analysis

Next Steps:

Full Offensive Model + Backwards Elimination

Full Defensive Model + Backwards Elimination

Develop Composite Model:

$$C = O + M$$

Nested F-test Analysis

Citations

Korolyk, Vitalii. "NBA Data 2010-2024 by NocturneBear." GitHub, NocturneBear, 2024, github.com/NocturneBear/NBA-Data-2010-2024.