

## NBA Team Data Project Description

### The Setup:

You have been hired as a sports analytics intern with an NBA team.

As the sport of basketball has evolved and taken on principles from software, baseball, finance, and other fields, the usage of sports analytics has taken on different identities over time. One of the age-old questions – and perhaps one of the most difficult ones – has been quantifying team “chemistry”, or discovering who can and cannot play well together. Given that basketball is played by humans with emotions and feelings, this can never be really quantified, but from a production standpoint, it can at least be discovered who can produce better together.

Therefore, one of the things the team would like to understand is how you can use analytics to understand what are the best lineups (which five players should be on the court at one time) in certain situations and against certain opponents. As more and more teams begin using analytics, using these types of analytics is not a vanity project, but increasingly a necessity to compete in the NBA. Luckily basketball is a game with many possessions (data points per game) as well as many games per season so there is plenty of data to work with.

What the analytics team really wants to know is can we find a “signal” in this data? Is there a way to separate the signal from the noise of basketball data (a highly dynamic and unscripted game)?

### The Team:

Any NBA team’s back office for analytics, as any professional sports team, is not a giant organization. Its job is to provide important analytics to the coach, the General Manager, and the President of Basketball Operations. If you can provide insights using data into how different players are contributing on the court, then the key stakeholders can use these insights in coaching their players and managing their roster.

### The Prompt:

You will be given play-by-play data on the entire set of games in the 2017-2018 NBA season. You will be asked to provide two sets of analysis and results:

**Insight #1:** You will be asked to predict the outcome (score) of 120 Games in the NBA season. You will be provided with play-by-play data for the remaining games in the season (1220 Games). The data will be at the individual play level and will contain information about each possession (turn a team has with the ball), as well as who was on the court and who was the opposition.

You will be judged based on: 1) How accurately you predict the box scores of the games. (50% Weight)  
2) How accurately you predict the +/- for each player, as well as the offensive efficiency. (25% + 25% weight). This will be weighted by the number of minutes each player gets in a game to adjust for the accuracy of the most important players.

The prediction accuracy of box scores will be calculated on the holdout sample as:

$$\text{Score} = \text{SUM}[(\text{Actual} - \text{Predicted})^2]$$

The prediction accuracy of player +/- will be calculated on the holdout sample as:

$$\text{Score} = \text{SUM}[(\text{Actual} - \text{Predicted})^2 \times \text{Share of Minutes}]$$

**Things to think about:** Consider the offensive (number of points scored) versus defensive (number of points allowed) efficiency of each lineup (5 players). Consider how much time they spend playing together on the court. Consider their opponents and the efficiency of those players.

**Insight #2:** Imagine you are in the final game of the 2018 season and you need to recommend a game plan to your team's coach before the game. Using the play-by-play data of the 81 games up until this point, you need to provide the coach with the following:

- The strategy or lineups that should be on the court (which players should he pick) to start and sub, and how much time should he allocate. (70% Weight)
- Who will be the most effective player on your team (in terms of points and +/-). (15% Weight)
- Who will be the most effective player from the opposition (in terms of points and +/-). (15% Weight)

You will be judged based on:

- 1) How the players in your lineup fared (plus minus, efficiency, etc.) weighted by the number of minutes they played. You will also be deducted a penalty if you allow any player to play more than one standard deviation above what they played for the first 81 games of the season. (This is to avoid having your best player play all the minutes). There is also a maximum of 48 minutes per game for any one player.

The prediction accuracy will be evaluated based on the following score.

$$\text{Score} = \text{SUM}[(|\text{Actual} - \text{Predicted}| - \text{Penalty})^2 \times \text{Share of Minutes}]$$

**Insight #3:** The predictions above will be your main task. However, as you go through that analysis you will undoubtedly look at different metrics, values and outcomes of games. In doing so, you will uncover patterns that may be interesting. Please document these findings in a graph or some other figure and present it as part of the memo. Focus on the one thing you have found most interesting.

What is most interesting is normative (i.e. subjective). This is true of any business setting. Your job will be to judge what could be an interesting insight or conclusion that your clients will respond to.

### **The Deliverable:**

You will be asked to provide a short summary (in the form of a 2-3 page memo or slide deck) detailing your insights, as well as a list of predictions in excel or csv format.

You may work in teams of 4 – 5 Students to analyze the data. Each student is expected to contribute to the project, and you will be asked to submit a breakdown of how much each student contributed at the end of the project.

You will be evaluated according to the following rubric.

<b>Grade Component</b>	<b>Description</b>	<b>Weight</b>
<i>Accuracy of Predictions</i>	Like in many data science type tasks, you will be judged based on how accurately you perform in predicting the outcomes (in Part 1 and 2 above).	30 + 30%
<i>Most Interesting Pattern or Insight</i>	Concisely explain what is the most interesting insight that you were able to generate by looking at this dataset (in Part 3 above).	30%
<i>Clarity and Presentation</i>	<p>You are preparing a very concise data brief. How you communicate and convey this is key. It has to be clear and impactful.</p> <p><i>As a general policy, I will not say anything about spacing, font, etc, on purpose. Do not think of this as an assignment where you have to follow my rules. Think about me as a client and showing me the best possible product!</i></p>	10%
<b>Total</b>		<b>100%</b>