MITOCW | MIT15_071S17_Session_2.4.02_300k

In this recitation we will apply some of the ideas from Moneyball to data from the National Basketball Association-- that is, the NBA.

So the first thing we'll do is read in the data and learn about it.

The data we have is located in the file NBA_train and contains data from all teams in season since 1980, except for ones with less than 82 games.

So I'll read this in to the variable NBA, NBA = read.csv("NBA_train.csv").

OK.

So we've read it in.

And let's explore it a little bit using the str command, str(NBA).

All right.

So this is our data frame.

We have 835 observations of 20 variables.

Let's take a look at what some of these variables are.

SeasonEnd is the year the season ended.

Team is the name of the team.

And playoffs is a binary variable for whether or not a team made it to the playoffs that year.

If they made it to the playoffs it's a 1, if not it's a 0.

W stands for the number of regular season wins.

PTS stands for points scored during the regular season.

oppPTS stands for opponent points scored during the regular season.

And then we've got quite a few variables that have the variable name and then the same variable with an 'A' afterwards.

So we've got FG and FGA, X2P, X2PA, X3P, X3PA, FT, and FTA.

So what this notation is, is it means if there is an 'A' it means the number that were attempted.

And if not it means the number that we're successful.

So for example FG is the number of successful field goals, including two and three pointers.

Whereas FGA is the number of field goal attempts.

So this also contains the number of unsuccessful field goals.

So FGA will always be a bigger number than FG.

The next pair is for two pointers.

The number of successful two pointers and the number attempted.

The pair after that, right down here is for three pointers, the number successful and the number attempted.

And the next pair is for free throws, the number successful and the number attempted.

Now you'll notice, actually, that the two pointer and three pointer variables have an 'X' in front of them.

Well, this isn't because we had an 'X' in the original data.

In fact, if you were to open up the csv file of the original data, it would just say, 2P and 2PA, and, 3P and 3PA, without the 'X' in front.

The reason there's an 'X' in front of it is because when we load it into R, R doesn't like it when a variable begins with a number.

So if a variable begins with a number it will put an 'X' in front of it.

This is fine.

It's just something we need to be mindful of when we're dealing with variables in R.

So moving on to the rest of our variables.

We've got ORB and DRB.

These are offensive and defensive rebounds.

AST stands for assists.

STL for steals.

BLK stands for blocks.

And TOV stands for turnovers.

Don't worry if you're not a basketball expert and don't understand exactly the difference between each of these variables.

But we just wanted to familiarize you with some common basketball statistics that are recorded.

And explain the labeling notation that we use in our data.

We'll go on to use these variables in the next video.