The goal of baseball team is to make the playoffs.

The A's approach was to get to the playoffs by using analytics.

We'll first show how we can predict whether or not a team will make the playoffs by knowing how many games they won in the regular season.

We will then use linear regression to predict how many games a team will win using the difference between runs scored and runs allowed, or opponent runs.

We will then use linear regression again to predict the number of runs a team will score using batting statistics, and the number of runs a team will allow using fielding and pitching statistics.

We'll start by figuring out how many games a team needs to win to make the playoffs, and then how many more runs a team needs to score than their opponent to win that many games.

So our first question is how many games does a team need to win in the regular season to make it to the playoffs.

In Moneyball, Paul DePodesta reduced the regular season to a math problem.

He judged that it would take 95 wins for the A's to make it to the playoffs.

Let's see if we can verify this using data.

This graph uses data from all teams and seasons, from 1996 to 2001.

There is a point on the graph for every team and season pair.

They are sorted on the x-axis by number of regular season wins, and are ordered on the y-axis by team.

The gray points correspond to the teams that did not make the playoffs, and the red points correspond to the teams that did make the playoffs.

This graph shows us that if a team wins 95 or more games, or is on the right side of this line, they almost certainly will make it to the playoffs.

But if a team wins, say, 85 or more games then they're likely to make it to the playoffs, but it's not as certain.

And if a team wins, say, 100 or more games then they definitely will make it to the playoffs.

So this graph shows us that if a team wins 95 or more games then they have a strong chance of making it to the

playoffs, which confirms Paul DePodesta's claim.

So we know that a team wants to win 95 or more games.

But how does a team win games?

Well, they score more runs than their opponent does.

But how many more do they need to score?

The A's calculated that they needed to score 135 more runs than they allowed during the regular season to expect to win 95 games.

Let's see if we can verify this using linear regression in R.

In our R console, let's start by loading our data.

We'll call it baseball, and use the read.csv function to read in the data file, baseball.csv.

We can look at the structure of our data by using the str function.

This data set includes an observation for every team and year pair from 1962 to for all seasons with 162 games.

We have 15 variables in our data set, including runs scored, RS, runs allowed, RA, and Wins, W. We also have several other variables that we'll use when building models later on in the lecture.

Since we are confirming the claims made in Moneyball, we want to build models using data Paul DePodesta had in 2002.

So let's start by subsetting our data to only include the years before 2002.

We'll call our new data set, moneyball, and use the subset function to only take the rows of baseball for which year is less than 2002.

We can look at this structure of the data set, moneyball, by using the str function again.

Now, we have 902 observations of the same 15 variables.

So we want to build a linear regression equation to predict wins using the difference between runs scored and runs allowed.

To make this a little easier, let's start by creating a new variable called, RD, for run difference, and set that equal

to runs scored minus runs allowed.

If we look at the structure of our data again, we can see that we have a new variable called, RD.

So let's build a linear regression equation using the lm function to predict wins with RD as our independent variable, and using the data set, moneyball.

We can look at the summary of our linear regression equation, and we can see that both the intercept and our independent variable, RD, are highly significant.

And our R-squared is 0.88.

So we have a strong model to predict wins using the difference between runs scored and runs allowed.

Now, let's see if we can use this model to confirm the claim made in Moneyball that a team needs to score at least 135 more runs than they allow to win at least 95 games.

Our regression equation is wins equals 80.8814, our intercept term, plus the coefficient for run difference, 0.1058, times run difference, or RD.

We want wins to be greater than or equal to 95.

This will be true if and only if our regression equation is also greater than or equal to 95.

By manipulating this equation, we can see that this will be true if and only if run difference, or RD, is greater than or equal to 95 minus the intercept term, 88.8814, divided by 0.1058, which is equal to 133.4.

So this tells us that if the run difference of a team is greater than or equal to 133.4 then they will win at least 95 games, according to our regression equation.

This is very close to the claim made in Moneyball that a team needs to score at least 135 more runs than they allow to win at least 95 games.

So using linear regression and data, we were able to verify the claim made by Paul DePodesta in Moneyball.