Learning R through Sports, or Learning Sports through R: CodeMash

by CodeMash

Learning R through Sports, or Learning Sports through R

Learning R through Sports, or Learning Sports through R

[Autogenerated] My name's Mike Ross. Nick, I'm here to talk to you about our and about sports. Honestly, mostly baseball. Uh, just because baseball, if you know who here follow sports a lot. But baseball is the most stats driven sport, and more importantly for us trying to do stuff with the stats. It's the sport that has the most easily accessible amount of stats you can go get for free. Uh, you could do a lot of cool stuff that I'm showing you with other sports, but it's usually a big pain to go get the big databases of stats. So baseball it is. I've struck a question. Who here has used our before? All right, Who here has seen a baseball game before? Awesome. All right, Perfect. So what is our It's interpreted language. It's heavily used for advanced mass staffs. Analysis, predictive analysis, things like that. And when people ask me Well, what in the world is our when they don't know it, they know other languages. I compared a lot to Matt Lab. Has a lot of people use that in school, and that's the way people can kind of get their head around it of what it. ISS and Matt Labs like a very, very, very simple version of our That doesn't do any the cool stuff, our

house. So there we go. That's that's your slides. Now we're actually looking code because that's the cool stuff. I'm assuming if you guys have used our most of you use our studio, that really is the best idea. There are other ones out there to use, But I don't know why you wouldn't use our studio. So that's gonna be using here, animal step through. So some basics of R and then I'm gonna shows some little bit more advanced stuff with the baseball stats I was talking about and we'll go from there, feel free, Stop me with any questions anybody has, and I will do my best to answer them. All right, So to start off pretty easy here, you want to add numbers. Like I said, it's interpreted all you d'oh four plus three says, Hey, that's seven. Isn't that amazing? But look at that. You can now make four plus three, turn it into a variable called seven and then print out the variable. Very complicated stuff. I know, but if you haven't used our before. This is a good way to kind of show you some of the basics of it. Like here. You can make a raise of numbers. You haven't re of numbers. Then with that rain numbers, you can do stuff like get the mean, get the max. We'll go down here, we'll jump into smart. You can actually, you know, directly access numbers in your ray. Pretty somewhere to most, uh, languages. You've probably used nothing too crazy yet same thing because math language has things like square root built into it. I have to do anything crazy to get that. Now we're starting to cool stuff. As you'll see over here on the screen is this is my history of everything that I've done. You can see it. That's just what I'm running through. So you don't have to watch me horribly type numbers and things like that. I would do it. What you can do is you can download some really handy stuff for our, especially when it comes to doing baseball. Ah, there's our tools, which is a pretty standard package of tools. It's your standard set, usually. But then there's also this library called Lonmin Leben. I don't know how you pronounce it. It's someone's name. I'm sure I pronounce it wrong. What? That hasn't it? It's just a ton of baseball stats, so you could just It's free to download. Downloaded. I think it's ah updated to 2018. Currently, baseball is really the good sport to go with if you want to do this with sports because that stuff is available. Also, say if anybody ever taken notes, slides are available. This whole history, everything I'll have available on Get Hub. There's Lincoln the end. You can get it so you don't have to worry about writing everything down. So you go. You get this great library, you choose to use this library, and what we're gonna do is we're gonna make some cool stuff with it. Has dysfunction here, batting stats, taking that, making it into be stats. What is that? There you go. You can see here kind of some of these batting stats that it shows, but an even better view of it. See appear. It shows, actually, everything that's in this batting stats database. You've got ideas for players years team I D. S, and then a whole bunch of stats. If you're a baseball person, you love these and you probably pour over the stuff like crazy like I do just tons of stance for a bunch of different players. Now, what we're gonna do is gonna take, take those batting staffs, put another variable called batting and going here. Well, miss first line, There we go. Make that. And what

we've done now is we've merged in a couple different sets of data and I can show you hear what that looks like. Remember what looked like before it is too much of stat straight across, and I pull in these staffs up, basically updated it by merging these to databases together. Looks like basically the same stats for everything. But now I've added in salaries. Granite doesn't have salaries for every player because there's not really salaries from before, sometime in the late seventies. But it's out there so you can actually pull stats for people based on how much money they made. For a year now, I have here another database here, Master info. Pretty simple. It's just the playersnames. That's all it is. I mean, this is a pretty simple table structure. You can probably get your head around it. What? What's going on here? So, uh, what we're gonna do next? It's probably Ah, very unexpected. But we're merging in those two tables. You could see all you do there to merge them. Very simple. Still calling this thing batting using the merge command for batting and master info. Then we go in here and look at batting. It's that same table that we had originally with all your stats there with your salary. Now we have those that player info from master info Birth year first in ___ lane when what they batted left your writing stuff like that again. You know, it's pretty easy stuff, but gives you a good kind of general sense of some of these things you could do with our any questions so far. Okay. I did not emerge them. Well, here in the Emerge command merge these two and the all that version with that and it base it on the, uh what? What is the name of the right there? If you seon master info. We had player i d, and that's the key for that table. And then the key for the batting table before emerging Master info was also player I d. Yes. I just matches. It's smart enough to know when you when you merge, it takes those keys and put them together. Yes, that is a good question. And I'll tell you, honestly, I'm not totally sure that's where you start getting to some of the our stuff that I don't totally know. If someone wants to answer, feel free. Okay. Yes. No, they don't have been the first column, but they have to be matching. They have thio sebi matching, but they have to be the same. Same name, same column name for it. But they could be in any any com with it, but be designated as the key for that table. And the questions thes one in particular. I don't actually know. Um, in there. Yeah, I don't know. If they are, you can, uh, load them up with different types. Like you can have, like, a regular like, like sequel database, where you have, you know, in so our characters of our cars or something like that, These ones in particular, I'm not sure what they're loaded up as all right where we were here. So I got all those guys we got Thea all the batting info. Now we have another thing here. We're gonna do pretty simple, are pretty similar to that they're setting up here. It is getting a subset of this batting data we're taking the you can see on their with batting. That's the batting data the year and minus your I d minus birth year. And, uh, if if els you guys carry that I don't need to read that stuff for you, so it gives us Ah, it was just a subset of data there. Now we're going in here using another library here. This pl why our library you can see

down here in this area has all of our packages as well that we was installed. Um, I'll show a couple different ways and stop act is in a moment. But just know you can do everything through the council. You can also, if you have something pre installed, you have this option on this pain here to actually just go and check stuff so you can go and take mark off the packages that you want a couple different ways. You can do it, So install this Pl Why are huh package they can see here is tools for splitting, applying, combining data Just a day manipulation package. So we're gonna go here, We're going to again update that batting table and we're arranging it. And if we view it again, it's not gonna look that different because it just kind of re sorted stuff out. But we're just looking at the top. A little bit of it. Vacancy. Same stuff in there. Now we're gonna start doing a little bit more with it. We're doing to create something called eligible hitters. And what that is is you can see here. It's a subset of that batting data where the year is at least 1900. So it's everybody that's 1900 till I said, I believe this goes up 2018 and they've had at least 450 plate appearances. So what? That is, uh, if you know baseball, that's people that that's eligible for the batting title for under 50 plate appearances and just from 1900 eyes. Some that's general used when doing baseball stats, because before 1900 the stats are really bad. They're missing. They're usually wrong. People find mistakes in them all the time. So generally, 1900 is kind of that clean cut off that's used. And so you could see it's a table exact same format of what we had of data. But if we look over here at our environment variables here we have our batting table and it was, ah, 105,000 Rosen. It eligible batters goes down to 14,000 because a lot of those batters in that table were from before 1900 or didn't have the number of plate appearances that they need it or their pictures. So everybody kind of getting what we're doing here with the data. How are manipulating it? I said it. It's pretty simple, a lot of this stuff, and this is why our is used for a lot of these things is things are built in pretty generally and make it very simple, like it's all here to get that subset. It really is just subset we're taking the subset from, and what you wantto used to kind of collect that subset. It's very simple to make those smaller sets of data. All right, so we have eligible hitters. Now we're gonna do is we have our top hitters here, not top hitters. If you can kind of read that again pretty easy to read it. You can see here taking the eligible hitters and then we're going Thio. We have the year as the I D right there. We're taking a subset from those hitters where the batting average you guys can kind of file along. It's it's pretty simple. The batting average is the max batting average of that year from this year I D or Banning out is over 400. So what we're gonna get from this subset, then with these top hitters is we're going to get all the batting leaders for every year, along with anyone that's hit over 400 because there are years where you hit 400. But there were other people that hit 400 as well. So it's just it's a way to get kind of those top players top hitters. DDP L Y. That is a, uh what does it actually stand for? Um forget it. It actually stands for, but what that does is you can see and hear how we have that, like

same subset command that we used before, is it takes that subset command to make that subset table. But then we're able to pass in the year here that we wanted to use for that. So it's a way of kind of adding variable into making that subset forget it. Actually, the actual name of it is, but it's like a fancier merged tool that allows you to use variables with it. So here we go. We have the, uh, top hitters here. It's that same table looks pretty much the same data in there. But if we go over here to our environment, you can see here the top hitters isn't even smaller table. Only 127 rose in it. So much less status or whittling down. Whittling down our data. I think about it. We initially started with everybody that had a plate appearance in Major league baseball from 1900. We had everybody that plate appearance of major league baseball that's in this database ever up till 2018 will it? Down to everybody. That was 1900 till 2018 that had over 450 plate appearances and then whittled that down, even Maur to all the people that led their league in batting or over 400 for a year. Just keep whittling it down, whittling it down with Lincoln down. Now what we're gonna do with that? Well, one thing we can do is we can make here a, uh, another subset of that data and that is with your with Thea dollar sign here. It's just a variable that's used with that subset of data to call it out. So we're using top hitters taken that database and doing anyway with the batting average, at least 400 so that everybody was a 400 hitter that's in there so they could have led the league. Could not let Lee. Doesn't matter. Everybody hit 400 for a year. Same thing here. We're taking the big names. Make sure you get all of this taking these, uh, making a big name Variables. Another big names database. We're taking a subset of that top hitters database everybody. That was the 400 hitter. It's taking a subset then of the top hitters database. That was any year after 1950 and they're batting average was over 3 80 So that makes sense. What we're doing with that whittling that down even more and what that's gonna do, You see here what that's gonna give us. It's an even smaller amount of data because all we get here or to the very end with names, we're going to get anybody hit over 400 for a season. I want to get anybody, though. That hit over 3 80 since 1950 because there haven't been a lot of 400 hitters since 1950. So it's been a few. Now we go in there were taken, ah, updating that big names database even more, making another subset kind of itself. And I were just taking on Lee the these couple of calms That's always did, as we just selected out these couple of columns player I D. Your I d. First last name and their batting average. So it's a really concise piece of data there that we took out. Same thing here we have top hitters did very much exact same thing there that we just did. I won't show it again, but it's same thing we just cut it down the mountain and because we really don't care about if guys were left here righty or how many doubles they had in the year or what? Their O. P. S waas? Any other stats that were in there was a huge amount of staffs, but we're just caring right now about batting average. So we're just taken that pertinent info willing down that data and a lot more reasonable read a readable and reasonable amount of data Now we're gonna

do is we're gonna graph it. That's one of the cool things that Arkan do is It can graft stuff pretty well. So I'm going to, uh, released couple of commands here. This is going thio set up are X and Y, and I'm including here. There's this library called G plot. I happen to be G plot to is the newest one. I was going here and you could see this is a pretty big command that it runs, but I think it's pretty easy to follow. Um I mean, it's just running in this G plot, which is going to make us a graph. What's it gonna show on the graph already set exes and wise up here. But here we go. We're taking top hitters. Uh, we're setting our x and y doing year on the X batting average on the Why are putting in some points you'll see here. You kind of get how these points work. When the graft pops up here, you could see where you are labeling stuff and I'll just bring the graph up. It's not easy to explain with it, and here we go. So by running that this is what we get the built in graphing that's in that G plot package. I can see it's got on here, everybody. That was in those, uh, that small side of day that we pulled. That's every batting title winner from 1900 to 2018 along with anyone else. They hit over 400 for the year. You can see what they're trying to do. Here is there was the smooth command in here to try and make some semblance of smoothing this graph out so you could try and figure out the plot for it, as he doesn't work very well. That's how baseball is. Nothing really makes sense, and it's all a lot of guesses, but that's the best they could do is coming up with that. But remember, we had that other subset of data That was all the 400 hitters. And right here is the line of 400 everybody They hit 400. We planted that as well. Close this guy. You can kind of see that. So you see, there's the graph. That's what we got. So we haven't hear with it. Thio, get to that. We find it. Yep. Yeah, there we go is Ah, because big names was Thank you. That was that table we had that had the people that hit 400 or hit over 3 80 since 1950. And so we're calling that out as a subset of data here plotting that on the same graph, you know, same x and y. And because I mean year of the year so it can share the same thing. It puts that guy in there and then it adds a label to each one of those points of the last name, which we didn't have on these other ones. We don't have a last name getting labeled. It also adds in and changes the color thio red on there. So you get that. That's a pretty simple kind of walk through some of the real basic stuff you could do with data and our and, uh, gonna get no stray a little bit more cooler stuff that we have that you can also. D'oh! Any questions on that? So far, Yes. Um, yeah, yeah, I have not done that. So there there might be a way to do it out, ladies. Yep. Now question mark first. Oh, they must build on that. Yeah. There we go. Well, there you go. That's what you can get that. Thank you. And I didn't lie about the poor typing skills. Uh, so at one other demo is gonna show you guys with data and you could see the whole bunch stuff on here preloaded stuff, because, I don't know, the internet connection be. And sometimes it takes forever. Uh, but what I did, because I can show you the commands. It's really easy using a lot of the same stuff we used before. But we added in this pitch our ex library, which, if you follow,

follow baseball talk, that's a big thing that they talk about now. Attract every pitch every game. It's okay. And a speed And, uh what type of pitcher waas and whether the batter swung on it. Hit it. Uh, so this is a library that uses that data. It can go get it. That pitch FX data is freely available. Uh, it's available back to I believe 2014 is the first year it's available for every game. That means literally every pitch in every game in the major leagues since 2014 is available for you to download for free and run any kind of queries in analysis. On that, you want to again why baseball is really the good sport to try learning this stuff with. I've looked it up for, like, previously for, like, hockey, basketball and football really doesn't have hardly anything honestly. And hockey and basketball are just in the last year starting to add more advanced, uh, stats for things and advanced. That's for them is very, very minimal compared to a baseball has. So all right, so what is? I called in this pitch our ex library and I created this data set that just went and scraped all the data from that between these dates. So between 10 6 2016 and 11 2 of 2016 which, if anyone knows what that is that is a 20 in post season. So and this is what it did, pull down everything. You see, a lot of all the stuff pop up here it goes and grabs. All the data was pulling in every pitch on those days. It's a lot of pitches, even though it's just the postseason I think about in a baseball game, there's probably around 300 pitches thrown, at least so a lot of pitches. Now what I take that if you guys can't tell what I'm doing, I'm looking at I should tell you here as I'm taking a uh uh subset of that data and I'm just looking at Columns five through 13. There's a whole bunch extra data in the in here. We don't care about those. The columns that have kind of important stuff we want it can see here because I'm calling it Kluiber. Obviously, I'm looking at Cory clubbers 2016 postseason run. So I put it in this Kluiber ah database here, and so you can see what kind of data we get is we're getting every pitchy through twice 16 postseason, so pitch type. So you have like curveball, slider ah ff his fastball. Ah, start speed, X and Y position of where they believe as best, their guesses of where across the plate and the outcome of it so called strikes balls Ah, foul and then stand is if the batter that was at bad time for the left. Your writing. There's a whole bunch of other stuff you can get on these. You can get a spin on the ball. You can get how much the X and Y changed during the flight of the ball on there, so you can really get crazy in depth. If you want to see who's got the highest spin rate on their curveballs for a season or who's slider has, you know, the biggest extra, why change for a season or a game or one batter or one picture his best game at that or something like that? Okay, I'm sorry I pulled in the ER when I pulled in the data for it. It, um, waas. I pulled the data for it was Kluiber data in there, so I messed. I skipped out of line on here From that, there was another scrape on there that pulls in that pulls in just kluver sorry I missed it on there, so I pulled in. I pulled in. The dates, then pulled include more than pulled in those. Right. Um, now we're taking this. We're going to look for the, uh, filtering this down. This is same thing we were doing before filtering down data taken all these pictures right here. Anything

that has a description of swinging because if it's swinging, it's general gonna be a swinging strike. We're just taking those. Bring those in its own well database. You can see it's even smaller, all swinging strikes right there. Um, you can see here. This is why you look for swinging because you can have stuff like this where it's swinging. Strike blocked, still swinging Strike has blocked in there. So you have Thio. You can have to look through some of this data and get to be a little bit more familiar with it because there is. Sometimes it's a little too specific, especially on this _____ pitch FX data. It could be a little too granular. Then you think it's going to be s so we got that. So this is all of the swinging strikes from Cory Clubber in the 2016 postseason has taken. Those were going Thio plot those over here. P is going to be our graph that we're gonna we're going to make, replied ng. From that and we're going here using the X and Y is going to be the X and Y, and we're basing the color on the start speed. Um, there is You can go and you can change like the color. If you want to be different colors to go from, you know, say, like blue toe orange or red or something like that. That's what they allow time show during games, but we're just using the default colors for everything. The next thing we do here is you can see we're creating the X's and y's and are basing those Ah, Green, the X's and y's and what we're doing here. You can see we're plotting out the, uh, strike box so very simple or play up strike box. Putting a label on it, um, putting different labels on their catchers view philosophy, things like that. It was kind of making a little bit easier to read. We're just changing the size on these guys. A couple more of these this is just to make it look a little bit prettier. And finally, what you wind up with is that every swinging strike from Cory Clubber put the wrong title on there. Nothing. But it's from 2016 postseason. You can see on their it plotted all those in there. This is from the catcher view. So if the catcher, well, he's looking and seeing this is where it's at, um, regardless of it was a lefty or righty batter, we don't care. And if you want to see the speed, this is why we're changing the color. Based on that is this is but it does with it. So you can see here what you would expect swinging strikes that are slower the swinging strikes at her up more where the harder, harder pitches. So this is this is how I got started with our This is kind of the very, uh I think introductory cool stuff you can do with it. It just dissecting this data. I did a lot of these kind of pitching graphs, and you can also get this same data. But instead of scraping for a pitcher, you could scrape for batters. So you can see every all that you know, every time a batter made contact, you get something like this every time he swung at something and missed. Well, you know, get something like this or every fastball that somebody faced over whole season. If you want to, What do you do with all of them? You can do that. It's not. It's not difficult. As you can see with our it's very easy to take those data sets and whittle them down because, as you can imagine, a data set of every pitch for even just a game. Like I said, it's around. Probably 300 or so pitches in a game. That's a lot of stuff we can whittle it down on. It goes fairly fast, too, to get up. And then all this graphing stuff is it's all, uh, just easy to

download libraries that air with our it's not extra. Add on stuff that you have to get or third party things. It's just standard our materials. So it also want to show you guys is first off. This right here book is amazing. If this is stuff that interest you, they just put out a new addition uh, 2018. I believe analyzing baseball data with our stuff I showed you is the first probably three chapters or so of the book. Is them doing this? And the rest of it starts getting into some crazy awesome stuff up live updating graphs of batters and pictures and overall runs scored in a season. Things like that. Eso with that? Another thing that people do a lot of stuff is three guys. Noah. The shiny, shiny APS weaken those for a lot of people. Use those host APS online. It's a super easy thing to do with our There's our library that is for shiny to make it connect with it. And it's really easy to do in our studio. Yes, yes. And I have one here that Ah, they load up nicely, just kind of give you an idea of where you can come is this is ah, uh, one of the more simple ones that's out there, and this is all stuff that's built into our all of these elements. This isn't anything like crazy third party or a different language that you use for it or something like that. It was all there, So this is the runs per game in baseball over the years. You're just a slider, Thio. Select your years. You could make it a You know you that you can, uh, get your ailing NL separately. So it's a ton of data and you see how quickly that is to get that, too. So stuff works really fast on there. That's the nice thing with our is. It's built to work with large amounts of data, so that means it works with it. Well, which means it doesn't quickly. Yes, all right, over on spikes the nineties. Well, um, I let's see, uh, I mean, if it's steroids is the big one. Yeah, well, it's serious, but it's it's c. Could you ask me? So I have to think of a good answer because it's so it's it's steroid use, but it's also a lot of ah, a lot of pictures in the late eighties and early nineties, had a lot of success with four balls and a lot of dropping pitches was a big thing, thanks to Roger, Clemens was a big person with that. So a lot of people were trying to throw those in the mid nineties and then into the late nineties. But they couldn't because people aren't Roger Clemens. Uh, so you had a lot of these sinking pitches that weren't good. And we're going slow like it's a big difference if something's gonna drop out in front of you and it's going 95 96 versus drop out in front of you going, you know, 88 89. So that also added to it. That's why I think you had a lot of these guys that were, uh, not on steroids, Probably still hitting a lot of long balls. So So you had to get my credit for that. I had to get that. Um So here's like that. Here's an example. Then of something in d'oh. Uh, this is, ah, code posted on get hub that you can go get. You can mess around with it as much as you want and see some more stuff on here with just this. And this is just runs runs per game from baseball. So let me, uh, get back up here. Get Thio. Last thing here we have Ah, you want a little bit further. There is more links. Well beyond that stuff. That's really the cool thing to do. I think with it is all the graphical stuff when I was learning are it was really just as a math language and it was just tables like Oh, cool. I could make a really cool tables of data. But then, as I got more and more into it because I was

interested in baseball, I started learning all the graphical stuff that are does, like all the plotting, all the grafts. I didn't know that I could do websites right out of our I thought it'd take that data out and dump it into something else. But no, you could do websites out of there. Uh, other cool things you can do is predictive stuff. It can actually do a little bit of machine learning in there with it. Not too hard. So if you want to try and predict, you know, uh, what the next batting champ is gonna hit based on previous batting champs, you can do that kind of stuff as well. All that stuff is in our There you go. That's everything I got. Guys, if you have any further questions or questions now, even you can ask him and get a hold of me. I have this stuff on. Get hope. I have the Q R code there to take you to the get hub. It's gonna have this amazingly dense slide deck, but more importantly, we'll have all of the code that I ran on this. And I also have some more links for further reading on baseball stuff because there is a ton of baseball nerds out there, probably nerdier than I am about it. Definitely know are better than I d'oh! And they get way in depth with a lot of those more advanced features that I showed you. Thank you very much.

Course author

CodeMash

CodeMash is a unique event that educates developers on current practices, methodologies and technology trends in variety of platforms and development languages such as Java, .NET, Ruby and PHP....

Course info

Level	Beginner
Rating	★★★★★
My rating	★★★★★
Duration	0h 42m
Released	7 Feb 2020

Share course