On-Demand Training: R Integration Transcript

Hello, and welcome to analyzing with R in Tableau. In this video, we'll be going over a few analyses common to R, and show you how to do them with Tableau. Included in these examples will be analyzing outliers, building cluster analyses using K means, and working with regression models. We'll also be including an example on hypothesis testing, which will be included in the attached workbook.

One caveat before we get started, we're assuming that those watching the video have some experience and knowledge using R and some basic statistical knowledge as well.

So let's go ahead and start with our first example. Outliers. What is an outlier? Douglas Hawkins wrote a whole book on the topic and stated that an outlier is an observation which deviates so much from the other observations as to arouse suspicion that it was generated by a different mechanism. That is to say, outliers are anomalies. They sometimes stick out like sore thumbs, but sometimes they're not so easy to see with the naked eye. And that's where things like R can come in handy and Tableau helps, as well.

So as you can see above, we have a chart looking at profit over time, and each orange dot denotes where profit was exceptionally high and where it was exceptionally low. This gives you great insight, seeing where things are significantly different from the rest of what's going on.

Let's take a look at how to build this and get a feeling for how R interacts with Tableau. I'll go ahead and create a new worksheet. Now we'll bring Order Date out to Columns. Then we'll bring Profit out to Rows. You see Tableau gives us a line here, but we want to see each individual observation separate from the others. So go ahead and make this a circle.

And now let's take a look at that outlier calculation. As we start, we tell Tableau that we're going to run some R script with a real number. First we call the package in our library, the outlier package. Next we run any functions from that package. And we have this, called ".arg1", which is a placeholder for any measures that we want to use from Tableau. Below that, we have the sum of Profit, which will be passed through .arg1, into R. And you notice that all of this R script is nested within one of Tableau's if statements. So, if the R script returns zero, then it's an outlier. Else, it's normal.

So let's go ahead and bring that out onto Color. You notice that, at the year level, we have no outliers. If this were R, and we had written all of this script out, and we wanted to ask "Well, if years don't have outliers, then what about quarters, or what about months, or what about weeks?". How do we drill down? We have to write new script. That would take a lot of time and effort. With Tableau, we're able to just go ahead and drill down seamlessly, just like any other Tableau experience. We can continue to drilldown as we see fit. And you'll see that more and more outliers stick out. This is a great example of the way that Tableau and R work seamlessly and allow you to have that drag and drop feel and let you drill down into your data without having to create any long, new scripts. You'll also notice that all the formatting was done by Tableau for us. So none of those GG plots or anything like that are needed, either.

Let's go on to the next example, Clusters. Here you'll see three different clusters. This is from a rather popular example, which is Fisher's Iris Data. If you've ever taken a college course on statistics, you've probably seen it. What these scientists did in the 1930s was they took a lot of observations of different flowers from different species. They then measured different parts of the flowers and wanted to find a statistical method to predict which flower belonged

to which species. Not only is this powerful in Biology, but it can also be powerful in any other application where you want to predict who or what belongs to which group. It helps you to see similar items.

So let's go ahead and recreate this and get a better feel for how it works. Go ahead and create a new worksheet. First we're going to bring out the petal lengths and petal widths out to Column. Next we'll bring the sepal lengths and sepal widths out to Rows. As in the previous example, we want to disaggregate this data so as to see each individual observation separate from the others. You can see we're getting close. Now we just want to go ahead and look at that cluster calculation and see how it works. What we're doing here, is we're creating a variable called "Result" and passing through the R function, k-means, across all of our measures. Arg1 being petal length, arg2 petal width, arg3 sepal length, and arg4 sepal width. Then we're just telling it to create three clusters, and return that to us. Then click OK. Like in the previous example, bring clusters out to Color, and there you have it. Our cluster analysis.

We also have an actual, so you can see what the real data is for the actual flowers, inside of their species versus the predictions. You'll see there's not many differences. It's a pretty accurate measure to predict who belongs to which group. It's a very powerful tool and Tableau makes it a seamless drag and drop experience with this analysis.

Let's go ahead and go to our last example which will be regression analysis. Regression analysis is a really powerful tool that helps you to estimate the relationship between different measures. Above, we're using World Bank indicator data, which can be found with any copy of Tableau. In this case, we're looking at the relationship between life expectancy and the infant mortality rates across different countries, each country, being its own observation across each measure. So, let's go ahead and look at how this works. You'll notice that, just like any Tableau analysis, we have our tooltip. We get a number of statistics with that. We also notice that each country is colored, whether it's below or above the trend. How did we do this? We used R to do this.

So let's go ahead and recreate this and see how this all works. Create a new worksheet. Next, you'll notice that my health-related measures are all grouped together. This is a really useful way to organize all those items that are similar. So bring out Infant Mortality to Columns, and Life Expectancy to Rows. Again, we're going to want to disaggregate this data, and you'll notice that we have a lot of observations. This is because we only want to look at 2010, so we'll have to filter that by the years and pick 2010. And you'll also notice that we have nulls. This is really because of the data that we have. Not every country has information on infant mortality rate and life expectancy, so we can go ahead and filter out both of those.

Now we have nice, clean data to work with. Go ahead and close that folder and now we'll just look at whether or not these items are above or below the trend. So go ahead and bring this out to Color and you'll notice that it worked. But let's go ahead and look at that calculation. How is it working? This calculation, in and of itself, doesn't have R script in it. It's using R script that was written in other calculations to pass it through a Tableau calculation. Here we're looking at the average life expectancy across each observation and we're looking at whether that's greater than the intercept and the coefficients multiplied by the natural log of the average infant mortality rates. This is across countries. So these intercepts and coefficients are R script that we're actually calling in to. And we have them over here.

So let's go ahead and see how those work. Here you can see that we're using the lm function from R and we're

passing it across one argument, which is the life expectancy, and the log of the other argument, which is infant mortality rates. And we're bringing that back through this vector. Let's go ahead and look at how the incept works. It works very similarly. So, we're just passing through the arguments, through the lm function, and we pass both of these together back into whether or not each of these observations is above the trend or below the trend.

Let's go ahead and bring that trend line out. You'll notice that it gives us two of them. So, what we'll do here first is we'll edit. We'll make it logarithmic and we'll tell it just to create one trend line. So there you have it. This is a really powerful way to look at the overall relationship between different measures and then also see who is below and who is above. You can also add on more functionality as you add on R skills and also get better at Tableau.

This concludes our examples for this video, but there's no doubt in my mind that the community will definitely add a lot more depth and knowledge as to what can be done with Tableau and R. Thank you for joining me today and I hope you have fun doing analysis with R in Tableau.