

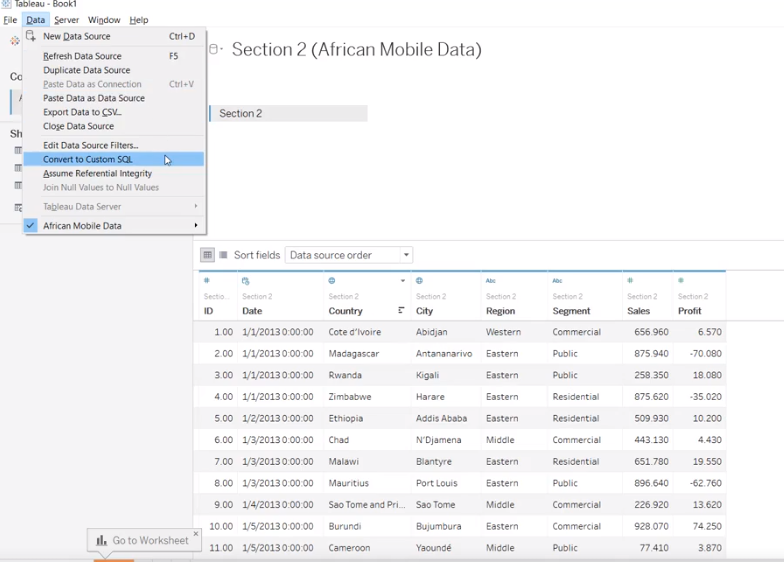
ALT+AC for short cut to calculation

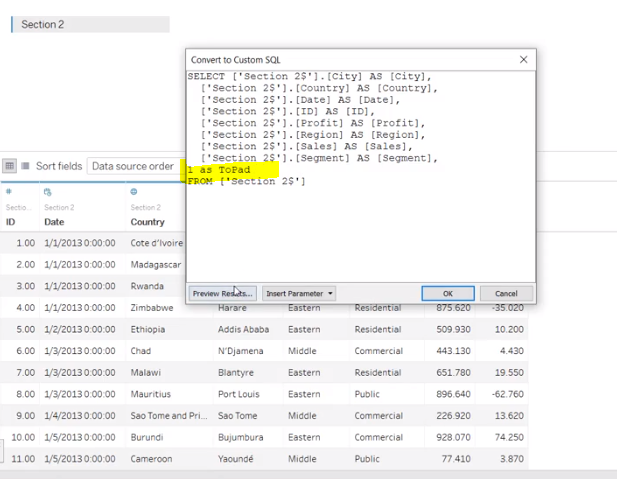
You can padd your data from 1 and 10 and then Bin my 2 in the Dimensions.

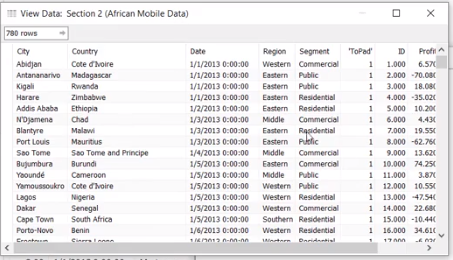
Go for the ALT+AC and put Index()

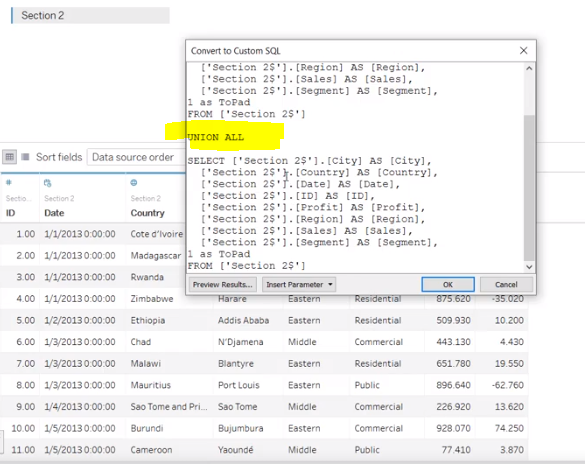
It will create index by ranking.

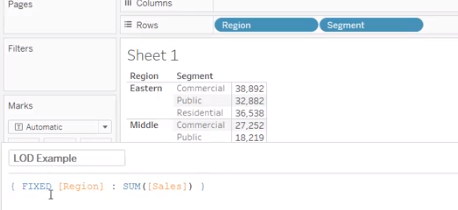
<https://www.microsoft.com/en-au/download/details.aspx?id=13255>











P-Value

There is no established association/relationship between p-value and R-square. This all depends on the data (i.e.; contextual).

R-square value tells you how much variation is explained by your model. So 0.1 R-square means that your model explains 10% of variation within the data. The greater R-square the better the model. Whereas p-value tells you about the F statistic hypothesis testing of the "fit of the intercept-only model and your model are equal". So if the p-value is less than the significance level (usually 0.05) then your model fits the data well.

Thus you have four scenarios:

1) low R-square and low p-value (p-value <= 0.05)

2) low R-square and high p-value (p-value > 0.05)

3) high R-square and low p-value

4) high R-square and high p-value

Interpretation:

1) means that your model doesn't explain much of variation of the data but it is significant (better than not having a model)

2) means that your model doesn't explain much of variation of the data and it is not significant (worst scenario)

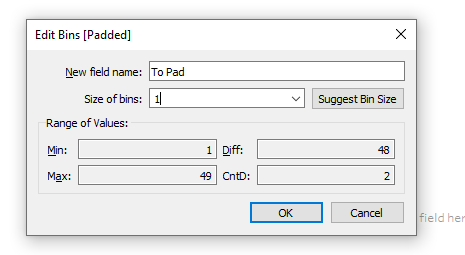
3) means your model explains a lot of variation within the data and is significant (best scenario)

4) means that your model explains a lot of variation within the data but is not significant (model is worthless)

<https://public.tableau.com/profile/mikevizneros#!/>

<https://twitter.com/mikevizneros?lang=en>

advance visualization, Data Visualization, hexbin chart, likert scale, sankey diagram, viola chart, Tableau, Udemy



To make 49 bins

Viola Chart Preparation:

Hello and welcome to the next video on setting up your viola chart inside of Tablo.

We're getting very close and it's exciting to be here.

We finished everything inside of Excel.

We finished everything in terms of data importing into Tablo and we created a data densification and

actually wanted to come in on that process a little bit.

I know it can seem a little bit tedious.

I know there was a lot we had to go through and you're doing great hanging in here.

The thing is I was trying with this challenge to replicate a real world situation.

A lot of times the classroom environments like this you will be given a data set.

It's nice and clean everything's done for you and you can go in and build your views in five minutes.

The thing is when working with real data it's just frankly not clean and you have to learn the process

of cleaning it up setting it up and making it so that you can use that data to build your views.

That being said all that's done now that's behind us.

So let's get in and actually build our views.

Right here we've got our data densification going on.

We have all these marks across the view.

We put it in a quick INDEX function.

We can then see we have values 1 through 49.

Now our first calculation is actually going to be a variant on that index.

We're going to call it T for reasons that you'll see a little bit.

So this is going to be open parentheses index minus 25 divided by 4.

What this does is it creates a scaling factor for t basically.

So if you remember index before we had the straight line.

This right here is also a straight line but it also allows us to include these negative values.

It also wraps all 49 marks between negative 6 and 6 helping us draw a nice smooth chart as we work through.

Now the reason that we named T is because t is actually a value in the formula for the curve that we

want to calculate.

It's going to be our next calculated field.

It's called sigmoid.

The formula for it is some straight math.

It's 1 over 1 plus E.

Now is a mathematical number that Tablo doesn't give you access to but we do have the EXPE function

the XP returns e raised to the power of the given number.

Example Expwy 5 is equal to either the fifth.

So each one is E to the first.

So we have 1 over 1 plus E raised to the negative.

Very important t

what this does is it creates a situation where that scaled index value is used as an input to the calculation.

So this curve is drawn across our padded values take that up you'll see we have this nice curve that

now exists across our dataset.

Now that's just the sigmoid though we need to create more calculations so that this sigmoid can be scaled

against the actual values we have.

So I'm going to need to create another curve calculation.

But first we need to bring in these two values.

So if you look at our stark percent or and present they'll show the same thing.

You realize we actually are in trouble in the sense that we only have two values and that makes sense.

We only have the value of one in 49 in our dataset.

That's because this is a non-Windows calculation.

So what we're going to do is create two calculations when doing that table calculation sorry similar

syntax and similar descriptions that can get mixed up sometimes but I'll call this a table start percent.

And we're going to wrap it in a zero.

No because we start with no right.

And so we want to make sure that starts at zero running average of the sum of our start percent because

Princie to wrap off that zero goal zero says if there's no value make it zero rather than just returning

blanks running average says take the average across the table of the sum of starting percent.

As you move across it.

So rather than window average which looks across the entire table as you go running average calculators

as you go along in this case you're going to have your first value of start percent averaged which is

always going to be exactly the same.

And then your last value which will be the exact same so you're going to have this constant start percent

across your entire table

right click on that and duplicate rather than table start present.

We want our table and present comp get rid of that copy as well and 0 no running average some of our

and present.

So now we have a starting point or ending point and we can create a calculation for our curve with cost

curve.

We'll take our STARTING POINT tables start percent and we'll add in the difference between our two endpoints

are two sides of the view.

So we have a table and present minus our table start percent.

Then you multiply that by the sigmoid.

So you can imagine table start present just imagine it down here is one table and percent imagine appears

10.

You have to get from one all the way up to 10 if you move across your table and subtract where you're

starting from from where you're ending up.

You'll eventually get there multiplying it by the sigmoid makes it so that it actually implements that

curve as you go across your table with that curve calculation.

We're actually all set now to build our view.

We're going to take a curve and T along with some dimensions here bring paddled over and we're building

out our view but we're going to save that for the next video you guys have done great hanging in here

Hexbin Chart:

You have underlying data points that if mapped that could be lost.

So you have too much detail on your view you're not really able to see anything by mapping things inside

of hex bins.

You can then show the relative frequency of them at an aggregate level while still maintaining the detail

of the view.

For instance this is not overlaid on a map but you can still tell that it's the United States.

Here's another example where we're looking at Steph Curry shot charts and you can see the percentages

based off of how many hits from the relative locations on the floor.

If you were to try and map every single dot from every single shot it could get busy very quickly.

A Mexican chart works past this by providing a level of aggregation in the form of a hex bin and then

plotting those as hexagons on your view at which you can then color by the frequency.

Then after that you'll see that we actually have.

It's almost like two passes through the chart.

The first we have a base Hexton calculations then we're going to go through to densification using our

level of detail calculation for these hexagons then we'll have our secondary calculations and we'll

build out the core chart then as we'll see there's actually a problem.

I'm not going to give away what it is right now.

So those of you who are paying attention are like a challenge.

And then if you drag up the census population to color all of a sudden we have a whole bunch of detail

here that viewed at an aggregate really is not that helpful.

Yes many of these individual zip codes here around Nevada you can actually see the individual zip codes

but you try and look anywhere with a dense population you quickly lose your data.

So Hexton makes it possible to bin these and actually see at a glance on a high level where the relative

concentrations are.

And so that's the technique we're going to show you today.

Oh a lot of credit to Alan Eldridge.

He did a lot of great work on this and we're using some of his techniques today.

And there's a couple other people that we'll talk about later.

But that's the Hixson Chert.

Challenge in the United States by law every ten years there is what's called a census.

A census is a way of counting the population and capturing certain data about the different demographics

across the United States.

This census data can be accessed online for free and is aggregated at many different levels.

But one of the most useful measures for population or for demographic studies is the zip code.

ZIP code is a level of aggregation right below city and 2 below state so each state will have multiple

cities.

Cities often will have multiple zip codes except for the very small ones which will only have one in

this data set.

We have the zip codes we have what state the zip code is in.

We have cities associated with the zip code.

You can see right here two zip codes in one city.

And then we have the population from the 2010 census that was shown in that zip code.

We also have a latitude and longitude value for every individual zip code which is very useful to have.

And the challenge you are trying to overcome is actually a text file.

So if you connect to it is that when you try to visualize this data to look for population density like

we showed in the introduction it becomes very difficult to pick out individual values very quickly.

So if we leave this chart on those automatic settings we can see some detail here but it quickly becomes

overwhelming if you zoom in.

You can then start to see the relative comparisons but you can also see how things are clustering around

city centers and you're losing out on a lot of detail when you have when you view it from a higher level

rather than from the zoomed in level.

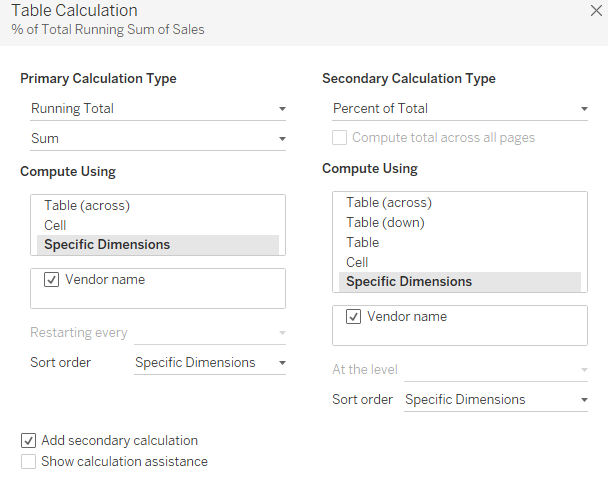
So we're looking for a chart that makes it possible to view this type of density and try to understand

what's going on underneath that as well as one that will respond to zooming in so we can still see this

individual detail as well.

Parito Principles Tableau

<https://www.thedataschool.co.uk/elnisa-marques/pareto-chart-dont-underestimate-peas-gave-us-pareto-principle-mendels-genetic-theory/>



Date

