Data Visualization for Storytelling and Discovery: Module 2 Introduction w/ Alberto Cairo

In week 1 you learn some elementary principles of data visualization. Now on week 2 you are going to see how those principles can be applied to real-world visualizations.

Data Visualization for Storytelling and Discovery: Module 2 video, part 1

As I mention at the beginning of the course data visualization can certainly be used for effective communication, but it can also be used for data discovery. To explore data and see what you can find in the data. What potential stories or insights you can find in the data. Originally actually, if you take a look at the literature about data visualization, the older books books that are considered foundational in the history of data visualization focus on this second purpose of visualization, the exploration side. Books such as for instance exploratory data analysis by statistician John Tukey or the Elements of Graphing Data by William Cleveland. Both of them are extraordinary books about data visualization focusing more on this statistical analysis and discovery.

And more recently books such as the Grammar of Graphics or more modernly even Visualization Analysis and Design by Tomas Munzner. If you are interested in learning more about how data visualization can be used to extract meaning from data, these are books that I highly recommend that you get uh in the future.

But anyway, I'm going to introduce you to some of the main principles that we can use to explore data using visualization. To do that I'm going to use a freely available and open source tool call iNZight. I'm going to bring it up into my browser. I'm let me explain very quickly what iNZight is and how it works. So iNZight is uh, freely available visualization tool that was designed and created by the Department of Statistics at the University of Auckland in New Zealand. Therefore, they played a little bit around with the title put in NZ in the world iNZight. It's a tool that you can either download and install in your computer. You just need to click over here and you can unload it and install it in your computer or you can use the web version the browser version which is the one that I usually use for most of my quick and dirty explorations of data. Before I proceed by the way, any data set that you see me using in any of these tutorials will be available in the page of the course. So you just need to look for it if you want to reproduce and recreate what I do in these videos.

Anyway, let me talk about the dataset that I'm about to use in this project. It's a data set that I used in the first few pages of my second book The Truthful Art. My book that focuses on how to use data visualization per say. In the intro pages of this book, I talked about a personal experience in dealing with data. I said and I wrote that when I moved to Miami for the first time and then later on when I bought a house in Miami and my family moved in, um, I wanted to learn a little bit about the quality of the public schools in the system, right the public schools that my kids were going to attend. So I was really interested in learning for example about how good or how bad the elementary, middle school, high schools in my area were.

Fortunately I could download data about these Miami-Dade County and the school system here has a data page where you can download the data from. I don't loaded data set that contains the quality metrics of each school in the system. There are more than 400 schools in Miami-Dade County. And the data set that I downloaded which I'm showing you right now on screen contains the following variables. First of all on the First Column we have this this school name on the second column. We have the Board District that each school belongs to because Miami is divided into nine board districts. These are Geographic areas Miami's divided into.

And then we have other columns with different metrics that measure quality. The first one is the school grade and this is the first one that I focused on right. The school grade is a letter grade that the county assigns or the school system assigns to each school in the system. I could be an A for and excellent school. A B for a good school. A C for an OK school and so on and so forth.
So I took a look at the data and I saw that the schools in my area the three of them elementary, middle, and high school. The three of them got A's and say well wonderful, right there are incredible schools the three of them got As. But then, you know knowing and having been educated a little bit about data thinking. I said well, wait a minute a number or a variable in this core value like in this case this letter grade only makes sense when you put it in context. When you compare it to other numbers. And the comparison that I made in my mind was okay so these three schools got As. How many other schools in the system also got A's? So I create a sort of a mental model of these data and I imagine that if I create a bar graph with the percentage of schools that got A's B's C's D's and so on and so forth. It will look more less like these right. Few schools that got A's few excellent schools. Plenty of schools are got B's. Plenty of schools that got C's. Then fewer schools that get D's F's and so on and so forth. Perhaps with another bar displaying the schools that didn't receive any grade for some reason, right?

So we have sort of a bell curve in this distribution right. Fewer A's more B's and so on and so forth. Now is this true or not? Let's visualize the data to put this model to the test. Let's do that. Practically. Let's go to iNZight. And here's how iNZight works. I'm going to go here. I'm going to go to where it says iNZight light and this will load the a web browser version of iNZight. Once it loads will have a menu that tells us upload the data here. It's the menu that says file. So I'm going to go here file and I'm going to tell the software import dataset.

Now that I have these window I'm going to move my cursor where it says browse. I'm going to browse my computer to locate the data that I want to visualize. I have a folder I create folders to store all my data. Data tutorials in this case. I'm going to find the data set that I want to use which is called schools SchoolsMiamiDade.csv. If you have never played with data CSV basically means comma separated values and it's a very common format in data visualization. Many data sets that you can find online are in CSV format, right?

By the way, you can open up this file in Excel or Google Sheets if you want to take a look at the take a look at the data set. All right. So once I have located it, I'm going to click on open and iNZight is going to import the data set and it's going to display the data set as it looks on a spreadsheet. We have the school name column, the Board District column, the school grade

column and so on and so forth, right. So and there are other variables that I will explain in just one minute.

So the next thing that I did was to visualize the number of schools that got A's B's C's and D's to compare that visualization to my mental model. See if it looks the same. I'm going to go over here to the upper menu where it says visualize and I'm going to select the variables that I want to visualize. And that is done on this column over here. I usually joke that iNZight is a great tool. But the way I still something that I forgot to mention that is important. That iNZight is a program that was created to teach statistics to high schoolers in at the high school level in New Zealand. So the high schools in New Zealand are using this software tool to teach students how to think about data right? So it's super simple to use and usually joke that is idiot proof almost because it says variable selection, please select the first variable that you want to visualize. What I want to visualize is the number of schools according to school grade right. These drop down menu over here. These items that you see over here correspond to the column headers that we had on the data set School named Board District. Let's select the school grade. Now iNZight is a program that will make smart decisions for you. If it detects that you are going to use this column to visualize. It will assume that what we want to do is to get the number of A's the number of B's the number of D's or the percentage. The percentage of schools that got A's B's C's D's and so on and so forth.

So I'm going to click on school grade and now automatically iNZight will create this bar graph that you have over here. And as you can see, it doesn't look like my mental model. The first bar that you have over here doesn't have a label. These are missing values most data sets have missing values that you need to solve somehow. You can perhaps find the right values by verifying the data with the source or you can just basically discard them if they are not there.
But here comes the interesting part. Take a look at the percentage of the percent of the schools that got A's. It's an enormous percent of schools more than 40 percent of schools in Miami-Dade got a great of A from the system. Then much fewer got B's much fewer got C's and much fewer got F's. All right, so.

Let me tell you why this is important. I have a mental model that look very different. Suddenly, I see the data visualized and I see that there is a huge contrast because between what I had in my mind what I'm seeing in the visualization, this is a clue. It's a clue to potentially finding a futurist story in the data. Whenever I see something like this something that looks interesting. What I do is to print out the graphic or take a screenshot and write a note like a reminder to myself look into these see what happens over here. Know that I usually wrote on this particular chart as you can see on the screen is these looks fishy. It looks very fishy to me that so many schools in Miami-Dade got a letter grade of A.

This could have happened for perhaps several reasons. I can think about two right. It could be that either schools in Miami are fantastic. Public schools are great so that's a reason why they got so many A's. But I know that that is not true. There are good schools and bad schools. And I don't think that there are so many great schools. Right? So the reason why it looks fishy is that it may be that the reason why we see so many A's is that there is something wrong in which in the way that

the system calculates or estimate and assigns later these letter grades. We don't know what the answer to that question is all that we know is that we have found something compelling and interesting in the data. Thanks to the visualization.

One key teaching in data visualization for exploration. Is that a visualization never gives you the answer to anything? What it does is to help you post better questions. We have found something interesting in the data. Now that we what we could do as a reporter, for example, imagine that I'm reporter working on the educational bit. Will be to pick up the phone call the county and said hey, here's what I have discovered. What is going on over here? Why so many schools are getting A's right? How do you calculate this variable? Right. So the visualization is a key element to find in that clue and do some further exploration.

Anyway in the data set in the data set. There are other variables that are much more interesting that this letter grade. For instance, there is a variable over here called reading. Actually there are two readings. Reading in 2012 reading in 2013 and the values underneath each one of these variables correspond to the percent of a student's on each one of those schools who were reading at grade level. On that particular school. So you see a number such as 1990. It means that 90% nine out of ten students in that school, we're reading at grade level in 2013. And we have a similar variable for math. We have the percent of a student's school by school who could do math at grade level. That's a more granular more interesting metric to measure the quality of a school. Right? How well students are performing on each one of these areas. We could have more right. We have science and so on and so forth. But this particular data set only has these two percentages.

I thought that they were compelling and I decided well, let's play around with the data. Let's see what I can find if I start visualizing the data. So let's go back to iNZight. I'm going to go back over here and instead of telling the software, please visualize the number of schools according to grade. Tell me, show me now a graphic that visualizes percentages of students who can read correctly or who could read at grade level.

By the way, before I forget it's very important to remember as long as I was talking about variables and values that if you want to use software tools such as iNZight or later on Flourish, which is another tool that we will use. It is very important that you pay attention at the format of the data. That's the reason why I recommended Data Wrangler on week number one, which lets you reach shape the data. It is very important for example that you use iNZight. Your data is shaped this way. All variable names need to be on the headers. Each variable needs to be one header and then all the values corresponding to that variable need to be placed right underneath header, right? So you need to structure your data this way if you want to visualize anything in iNZight.

Data Visualization for Storytelling and Discovery: Module 2 video, part 2

Anyway, so let's go back to iNZight and now I'm going to tell the software iNZight. Well don't visualize his school grade. Now. Let's visualize a quantitative variable right. School grade is a

qualitative variable is a categorical variable its names ABC and D. Those are also values, but they are not numbers.

Reading 2013 is a quantitative variable it's a continuous variable. We goes from zero percent up to 100% with all the steps in between right. So iNZight is going to visualize it differently as you may notice in just one second. I'm going to select it and the graphic that I get is a Dot Plot adopt lot that we could also call a histogram.

Let me explain how to read this chart. It's actually quite compelling. Each one of these thoughts corresponds to one school. The position of each one of these dots on the horizontal axis corresponds to the percent of the students on that particular school who could read well. So for instance on this schools here that you have on the right end of the spectrum a majority a huge majority of the students are reading correctly nearly 100% of the students on this school were reading correctly in 2013. If you go to the other end of the chart these four schools over here on those four schools. No students were reading at grade level just zero percent of the students were reading at grade level. And then most schools are somewhere in between. Most schools are cluster here in the middle in between the 40 percent threshold and the 65% threshold so to speak, right. Those are the percentages of each one of these schools.

So each dot is a school. Right? So this is showing us the shape of the data the distribution of the data and also the range of the data the range of the data is the minimum value and the maximum value of this distribution. In this case a minimum value is zero percent.

There are schools in which note students read well and 97 percent or something like that. That's the maximum number. That's the school the best school in terms of reading in our distribution. The other chart that you're seeing is this one down here. This is called a box plot. A box plot is a great way to summarize the distribution of a variable. The box plot basically tells you the same that you're seeing on the histogram, but in a different way. The box is red this way. This end of the box plot over here this represents the minimum value, right 0%. And then these end of the box over here of these line, this is the maximum value of the distribution. And then these lines that define the box itself. These are called the quartiles. The quartiles are the values in any distribution that divide that distribution into four portions of the same size.

That's quite a mouthful. So let me explain a different way. What the box plot is telling you is that 1/4 of a schools in Miami-Dade are here right? They have a percentage of the students who can correctly between 0% and 30 something percent. So 25 percent of the schools are here. Another quarter of the schools in Miami-Dade are between this line and this other line. This is the first quartile and this is the second quartile. Also called the medium which is a kind of average. Right? So another quarter of a schools are over here. Another quarter of a schools are over here. And then another quarter of the schools are over here. So the boxplot basically what it does is to summarize the range of the data, but also shows you how concentrated or how despairs the schools are in comparison to that average the median there in the middle. If you want to see the

numbers by the way behind any of these of these charts that I'm creating iNZight let's you do that. If you click on summary up here iNZight will show you the numbers behind that chart.

So we'll show you the minimum value is zero. The maximum value is somewhere over here 97. The first quartile is 37, the second quartile, which is the median is 52 point something, the third quartile, which is a 75 percentile is 68, and then the mean is that under standard deviation is that we don't need to worry about that if you know nothing about stats, but if you work with the stats this also displays the standard deviation of the data. And then the sample size which is basically the total amount of the schools that we have in our data set, which is 400 and something those are it's called the number of observations, right?

Anyway, we have discovered that schools in Miami-Dade are very unequal. Right? We have very schools in which students perform really well in reading and schools in which students don't perform that well in reading right. We could do something similar with math, right so I could go over here and what it says select first variable instead of visualizing reading I can visualize math and this will show me the distribution of math percents and this is actually quite more compelling.

Because I can see two different facts that look quite promising. First of all as potentially stories to tell from the data. For instance, I want to know what those schools are. So we have one, two, three, four five. I believe that eight or something like that in which no students can do math at grade level. Why? I don't know. The visualization never gives you the answer.

It's just giving you a clue to something that you can further explore later on. You can begin pick up the phone call the school system and say hey what is going on in this school? Right. That know students do math well. And then on the other hand you have one school over here that departs from the rest of the school's. Most schools are done here.

But suddenly you have an outlier. You have one school in which all is students. 100% of the students score perfectly at math. They are all do math at grade level. What is that school? Again, I don't know but it's it's worth exploring. What's going on?

One thing by the way, one feature great feature of iNZight is that I can find out what that school is, right? This this chart that you have over here is completely static, but I can make it interactive. I can go over here. I can go to interactive plot and then produce plot and the chart will eventually appear over here usually takes a while. This is the software again. It's online. It will depend on your internet connection.

Here we have the interactive version and then you can plot more variables right? You can say for example variables to display. You can display them all and then you can hover over and see the actual the actual scores over here. You can also see the table.

And then they plot that you producing in iNZight. This is an aside you can download for example, the interactive plot that I'm seeing right now can be downloaded as an HTML file. And the aesthetic

plot that I had before can be downloaded either as a JPEG, PNG, a PDF, or an SVG. These two formats PDF and SVG will help you later on to bring this chart into other software tools such as Inkscape or Adobe Illustrator or Photoshop or InDesign or any other design tool to style the chart a little bit a little bit more so you can download any chart that you create in iNZight.

Now the same way that I did before if I see these charts with so many potential compelling stories. What I do is to take a screenshot of them or print them out and they actually write notes like reminders to myself. Uh, salute things to look for right? So these are some notes that are wrote about this chart.

What is this school? What is that school over there? Why no students are reading correctly over here. Why are not not so many students doing math correctly over here. Those are potentially stories, right? Thing that we have discovered thanks to the visualization. We don't have this story but we have the clue to find impossible stories in the future.

Data Visualization for Storytelling and Discovery: Module 2 video, part 3

Now the next thing that I did with INZight was to ask myself. Well, is there any relationship between these two variables. Percent of students who can do math well and percentage of his percent of the students who can do who can read correctly. So INZight lets you do that because you can plot more than one variable. So I'm going to go over here. Select first variable. I'm going to do reading I'm going to go back to reading and then I'm going to select a second variable. Second variable is going to do math. And again INZight is a very smart piece of software. And if it sees that you're trying to visualize two quantitative variables. It will assume that you want to see the correlation that relationship between those two variables.

So I'm going to just like math and what you will get is what is calling visualization as scatter plot. So that's the scatter plot. This scatter plot shows you the relationship between these two variables. So on the x-axis you have math. On the vertical axis you have reading .The further to the right each dot is the larger the percent of the students in that school who could do math correctly. And the further up one of those dots is the larger the percent of students who could read at grade level. So as you can see that in general and that's the pattern that we have discovered thanks to the visualization. The further to the right a dot is that is the further up. It tends to be BB a positive correlation. The more you have from this the more you have from the other from the other verbal.

The correlation could be inverted by the way. Sometimes you have more than one variable and less than the other. They could happen that the more students do math correctly, the fewer students can read well. That's not true obviously, but it could have happened. That's something that we could also have discovered. Thanks to the visualization.

So this shot is not that compelling right because it's it's predictable that in general if students perform well in some tests. They will also perform relatively well in other tests. There are several potential schools here that may be worth looking into. For instance I saw this is school over here in

which fewer than 20% of the students could do math well, but more than 55 percent of the students could read correctly. So there is an imbalance between reading and math what's going on over here. It could be perhaps that that school puts a lot of emphasis in teaching reading but not so much emphasis in teaching math. We don't know but it's something that we have discovered thanks to the visualization. Now the most interesting part comes when I did something else which is to split the data up right.

Before I get there by the way if you're interested in statistics. This is just an aside for those of you who use his that INZight also lets you insert trendlines, right? So if you go to and I describe these in the practical tutorials that come after these videos. So if you go to the add to plot menu up here, there is one menu over here that says trend lines and curves and it will this will let you insert for example a linear model right regression line there in the middle that shows you the overall direction of the data. And you can insert quadratic lines or cubic lines different kinds of trend lines in case that you're the relationship between the two variables is not perfectly linear.
Anyway, so I was saying that the interesting the most interesting thing that I discovered in this data set appeared when I did something else which is to split the data up according to another variable. Remember that when we took a look at the data set I mentioned that there is one variable in the data set called board district, right. Miami is divided to nine board districts. These are geographic areas the city of Miami is divided into. I ask myself. Well right now on my scatter plot I am seeing all schools together right. All the schools in Miami-Dade together into the same chart. What happens if I create one scatter plot per board district. Let's see what I can find in the data. So I went to INZight and before I can do that, by the way, we're going to use the subset option right subset by board district, but we cannot use it right away. Because in order to subset the data correctly INZight needs to learn that the variable board district is not a number. It is a name. It looks like a number because if you take a look at the variable on the data set its board district 1,2, 3, 4, and 5.

So, you know these software tools can be really really smart. But in some cases they are really really dumb. And when they see a number they interpret that that is a number that is a quantitative variable 1 is less than 2 and 2 is less than 3 but that is not the case. This is a categorical variable. It's a name. So we need to tell INZight please don't interpret this variable as a number interpreted as name. Board District 1, Board District 2, Board District 3.

In order to do that I need to go to manipulate variables over here on the upper menu, and I'm going to go to numeric variables and I'm going to transform that variable from a number into a name. It will look the same. It will still be a name but inside is going to interpret that number as a name not as a number, right. So I'm going to go over here. And transformed variables I'm going to select the column that I want to transform which is going to be board district and I'm going to tell INZight this is not a number. This is a name. And that is done this way you select Board District and then learn here is select transformation. And you say change to factor. That just a fancy name for telling INZight please don't interpret this as a quantity as a quantity. This is a name. I'm going to change the factor. And then I'm going to click on transform and once I do that what INZight will do will be to

create an additional variable in the data set called factors board district. And that's the one that we are going to use to subset the data.

So if you click on transform that variable has already been created. I can come back to visualize. And I'm going to go over here until the software subset by factors board district. And this is going to create nine scatter plots, actually 10 is scatter plots because there is one school that doesn't belong to any of the board districts. It belongs to Board District 0 that doesn't exist in the data set. But all the other ones belong to Board District 1, Board District 2, Board District 3, Board District 4, 5 and so on and so forth. Now when I saw this chart. I started visualizing many many potential stories that we could pursue based on this very simple data set. So I took a screenshot. I downloaded this this graphic I click on PDF here and I downloaded the plot to my computer. I open it up in a software tool and I started adding annotations. Things that I would like interested. I would be interested in learning more about from these data sets.

So here here you have the same Scatter Plots over here with little circles and arrows pointing out. Things that I found compelling in the data for instance take a look at Board District Number 1. In Board District Number 1 most schools on that district are on the bottom left quadrant. That means that fewer than 50% of the students read well and fewer than 50% of the students can do math correctly, right? So that's a pattern most schools are on that bottom left.

But there's one exception when we explore data with visualization. We need to focus both on the trends and patterns and the exceptions to those trends and patterns. So that's the reason why I put I put a little arrow over there. Say well, we have a school in which students perform really well in math and reading in an otherwise relatively low performing board district. So what's going on in that school? That could be a potential story. We have a great school a school. That looks great in another wise not-so-great board district. What is that school doing well? Or could it be that the students that come to this school come to the school already really well prepared and that's the reason why they get better scores? We don't know. We don't have a causal explanation for this data to look like like it does right now, but we have discovered an insight from the data pattern an exception.

Now Board District Number 2 is very similar to this number one with the exception that is three schools that depart from the rest of the schools. This is school's over here. Now Board District Number 3 that's quite interesting as well. Board District Number 3 is a board district in which most schools are on the upper right quadrant of the spectrum. What's going on? Most schools are up there meaning a majority of the students read well and a majority of the students also do math well. There are no schools on the bottom left meaning there are no schools with a minority of the students who can really well and do math well, right. And then Board District Number 7, which is the Board District where I lived in uh is the is like the mirror image of Board District Number 1. If in Board District number one most schools were on the bottom left with one great exception on the top right. Board District Number 7 is the opposite. Most schools all schools are on the top a top right of the of the spectrum and there's one school down here and you can notice that I will put an air I put an arrow a pointing out that score. So we have a low performing school in another wise

great board district in terms of school. In terms of student performance in school. What's going on? We don't know but we have discovered an insight.

And then finally you may notice that I circle board district number six. I found that one really compelling and promising. The reason why that happened is that I saw that in most other board districts schools tend to be concentrated either on the bottom left or on the upper, right? You know low performing high performing but board number six these schools over here are really spread out and that's a reason why I mentioned before that when we visualize data for discovery, we need to pay attention on how much data concentrates or how much data disperses. In most school districts data is concentrated. Most schools are either here or here or here. But in Board District Number 6 we have screws on the upper right quadrant and schools on the bottom-left what's going on? We have low performing, high performing schools in the same district. Board districts are relatively small geographic areas. So why do we see so much school inequality?

The next step in this process was to design a map. I remember before moving to the United States that are read a lot of stories about a school quality and income inequality and the relationship between income inequality and school performance. This is our relationship that exists in all countries that I am aware of and all the countries where I lived in. It happens in Spain. It happens in Brazil and it happens also in the United States. It's only that in Spain the relationship between income high income and school or low-income and school performance is not as strong as it is in the United States right as far as I have read right. Now let's put that to the test. Let's see if there is actually a relationship between school performance and and income levels.

So I went to the Census Bureau. I downloaded the data set of um a income household income area by area in Miami. I drew that map. And then I overlaid a map of the boundaries of the board districts and I'm going to show you the uh the charts that I created side by side with the map and notice the patterns that we can discover over here. Board District Number One and Board District Number Two, which are the low performing districts are very low income areas in Miami. These are areas such as for example Over Town and Little Haiti, which tend to be low income in terms of household income. Board District Number Three, the third chart on this data set that we have over here. That's Miami Beach. Miami Beach obviously tends to be very high income.

My own area has both high income and relatively low income areas and that may explain that distortion in the data, which we have most schools on one quadrant and then we have a low performance school. Perhaps that low-performing school is in one of the low-income areas in my in my in my region. I don't know that's something that we need to look into right. And then the most compelling one that I mentioned before is Board District number six the one in which all these schools are very dispersed. We have high performance and low performing school. Now that inequality in terms of school performance is reflected in income inequality. Board District Number 6 is an area that contains at least two regions that are very very high income. One of them call Key Biscayne, which is an island, which is very very high income. And then our on the mainland you have a very narrow strip of land very close to the coast called Coconut Grove. And certain parts of Coconut Grove are very high income. Right?

But then on the other side of something if you pay a take a look at the map of Board District Number 6, you will see that they're very low income areas. It's like there is something that cuts the district in half. High income at the bottom low income on top of that. That thing that cuts the district in half is US 1 right? US 1, a road cuts the district in two halves high income and low income. Well, if I could explore the data further. The next step that I would do would be to find the coordinates the longitude and latitude of each one of these schools. Or I would use for example Google Maps to find them on a map. And then I would plot those schools onto the map to see whether the low-performing schools are on the low income side of Board District Number 6. And the high-performing schools are on the high income side on Board District Number 6. That will really really show the evidence behind that claim that in developing countries and developed countries like the United States, that there's a very strong relationship between income levels and school performance.

Now what we are exploring these things over here and I'm going to emphasize again. A visualization never gives you the answer to anything. It just lets you discover features of the data. What we are visualizing so far are relationships, we cannot establish a causal connection between these two relationships. We cannot say for example that low-income families leads to lower school performance. The relationship between those two variables could be much more complicated. We don't know why this relationship exists. The fact is that it does exist and this may lead us to further explore the data and uncover even more things. The process of using visualization for discovery can be endless. We could continue just digging deeper and deeper and deeper into the data finding other data set data sets that may supplement our exploration. We could also contact experts and that will be the next step that I would follow if I were interested in doing a data-driven story based on these data. I would pick up the phone go to the education department of the University. Talk to people who are experts on educational data in order to confirm whether any of these potential stories that I have this uncovered in the data thanks to the visualization have any merit. Or they are just a product of noise, which is something that sometimes happen when you explore data.

Data Visualization for Storytelling and Discovery: Module 2 video, part 4

If you want to learn more about iNZight and how to use iNZight for data exploration, you can refer to the Practical Videos this week in which I explore other datasets and I show several features that iNZight has such as for example how to color things differently how to arrange graphing graphics differently and so on and so forth. But there is a particular feature that I didn't cover in those videos because it was not in the program when I recorded them. It was recently included in the software which is the ability to design maps right in iNZight itself.

So I'm going to show you how that works very very quickly over here. So I'm going to upload a different data set. I'm going to go to the file menu import data set. I'm going to import a data set from gapminder, which is one of the sources that I mentioned on week number one. I'm going to import the data set called Gapminder2012, which is data set from country-by-country about

different kinds of features of those countries, such as obesity rates, child mortality, and so on and so forth. I'm going to open that file and once I have that I have done that. I'm going to go to iNZight to the upper menu where it says Advanced because all the new modules that are incorporated to iNZight little by little as the developers have time to create them. They are usually put the usually placed on the advanced on the advanced menu item.

So down here, you see that it says maps and this is how it works. Once you have that once you have the data set uploaded and you go to the maps. First of all you need to select the variables and there are two options over here. You can either use coordinates. So for instance if you had city level data, your data set should include the longitude and latitude of the center points of these cities if you want to visualize it on the map. In this case in the data set that I'm using I don't have longitudes and latitudes. What I have are is regions. I have regions of the world right? Therefore. I'm going to go here to regions and next a iNZight will ask you what is the map location? Map location in my case is the world this is world wide data, so I'm going to change that. But I need to tell it what the location variable is. In my case the location variable is country. There is a column in the data set called country and that includes all the country names. So I'm going to select country and then I need to tell it which variable I want to plot. So I'm going to open up this menu and over here as it happened before why do we have is all the column headers all the column names? So we have number of cell phones per person CO2 emissions, GDP. per capita ,, infant mortality, life expectancy and so on and so forth.

Let's display for example children per woman, which is the fertility rate. The average number of children per woman on each one of these countries from zero children per woman. Although there's no country with a fertility rate of zero, probably zero point five and the maximum value is like seven or eight children per woman in the poorest countries in the world. Let's display this data set on a map. If I click on children per woman automatically iNZight will generate a map of the world that would be color-coded according to the number of children per woman. The redder the color the larger the number of children per woman. As you can see the countries that have a larger number of children per woman a larger fertility rate are mostly in Africa with exception of Afghanistan. Afghanistan also has a very high um, uh fertility rate.

Now, there are many features that you can change on a map and I cover how to change the style of charts in the other in the practical videos that you can also watch. But if we're going to change the style of a map, you can just click on more options. And over here you have an option for example to change the color palette if you don't want to use red because red is a very aggressive color. Perhaps you can use blues for example, which is a calmer a color for these kinds of these kinds of data sets.

And then as it happens with any other kind of uh chart or map design with iNZight, you can download this map over here. You can just select the format that you want. PDF, for example. You can save the map. And then once you have you have save it, you can bring that map up into other software tools such as Photoshop or Illustrator or Inkscape to further to do some further styling.

Thank you for completing week two I hope that you have learned a lot about how to explore data using visualization Now we move on to week 3 in which I will teach you how to use these tools and these techniques to communicate with audiences.