



Transcript for Session 049

Listen to the podcast session, see resources & links:

<http://chandoo.org/session49/>

Transcript:

Hi and welcome to <http://chandoo.org> podcast. This is session number 49. We are getting close to session 50 of our podcast!

Today, I want to share with you a very interesting story/principle when it comes to making awesome charts. The basic principle is very short - "**Don't Do Data Dumps**" (DDDD). In fact you could shrink it and make it DDD (Don't Dump Data). But, that sounds a little serious and misleading and so I went with "Don't Do Data Dumps" (DDDD). Let's talk about it.

You might be thinking why we are talking of data dumps all of a sudden. Did I come across some serious data dumps? Well, that's true. Recently, I ran a contest on <http://chandoo.org> where I shared some data about KPIs - base value, current value, previous month value, targets, corresponding month value and stuff like that - and I asked our readers to come up with visualizations. We got more than 60 entries for that particular contest. This week I have been reviewing the entries to shortlist some of them and publish them on the blog as well as figure out who won the contest. There will be 2 winners and I wanted to announce them before the 1st week of December so that they have enough time to spend the gift money and buy something nice for Christmas.

While all of this was going on, I was looking at all these charts and I basically saw two categories of charts. The first comprise of a really good effort on charts and dashboards. Essentially, people put in a lot of effort; they thought about data, and they came up with novel, creative, informative and insightful ways to present it. This is the first category. The second category comprised of people who just selected all the data and inserted one of the random charts whether a bar chart, line chart or a pie chart, and just sent it to me. This is what I call **data dump syndrome**. A data dump is when you take all your data without caring what the data is (maybe you do care because you have been working with it) and go to the Insert Chart option in the Insert ribbon and insert a chart. It could be a beautiful looking column chart or it could be a clumsy 75 slice pie-chart. You insert a chart and send it to your audience. It could be your boss, clients, colleagues or to Chandoo in this case. I am running a contest and you send me the entry. This is what I call a data dump. **A data dump occurs when you take all the data and don't really try to isolate information from the data.** You just look at all the numbers and think about how to make a 3-D column chart out of all these 200 numbers, and you make it and send it. **Data dumps are really**



bad forms of visualization. It doesn't matter what best practices you adhere to. As long as you are dumping your data on the chart, no matter how beautiful your chart looks, it is still a data dump.

I am going to give some examples of data dumps on the show notes pages available at <http://chandoo.org/session49/>.

The basic idea is very simple - don't do data dumps. You might be thinking what else is in this podcast. We have already covered the basic topic and what are we going to do now. This is supposed to be a short podcast because I am in between a few things and I want to rush back home as well. But, this is how it is. Don't do data dumps.

I gave you one example about contest entries. Here is one more example. I see these kinds of examples all the time especially when people post their problems on Excel forums or when they present in various meetings or things like that. Let's say you have 50 products and you are trying to understand their performance - it could be sales, number of customers, costs, quantities or whatever.

A classic case of data dump is just taking all the data for the 50 products and making a chart. The **worst kind of data dump** in that case would be a **pie-chart** or a line chart. A pie-chart with 50 slices is a classic case of data dump. We don't care what our audience wants to understand; we just want to get the chart out and rush out of the office. If this is the attitude that the person has then you will end up with a 50 slice pie-chart. If you think I am kidding, I am not. I have seen several pie-charts like this when I was working. And, I see several pie-charts like this even today in business publications, media and blogs etc. Either people are criticizing them or people are just creating and posting them because they think it is fine. That's your worst type of data dump.

The **next level of data dump** is making a **line chart with 50 lines**. We know you can't draw a line from product 1 to 2, and 2 to 7, and 7 to 15. There is really no linear or line-like relationship as they are individuals. One of them could be cookies, and the other could be diapers. And, you cannot draw a line from diapers to cookies as that doesn't make any sense. But, still, people do that, and that is another kind of data dump.

A **slightly better kind of data dump** in this case would be a **column chart with 50 columns**. People can argue that a column or bar chart with 50 values is still readable. But, the problem in this case is that there are just too many columns and drawing any kind of meaningful insights out of that mass of 50 columns is really difficult.



There are various levels of data dumps and no matter what level you are doing it, **you should ask yourself if it is a data dump or if you are providing information.** That is the **basic question** that you should be asking when you are creating a chart.

So, these are examples of data dumps. Why do we dump data? I have done data dumps earlier in my life and maybe even today I do a data dump occasionally. Why do we do this? The **basic reason** seems to be **lack of time.** We don't have time to clearly understand the data, analyze it and come up with information and insights out of it. We just want to dump it in the chart and be done with it, or play with the chart. So, lack of time is the first reason.

The **second reason** is the **ease of creation of charts.** When you select a bunch of cells in Excel, especially in 2013 and 2016 versions, Excel automatically shows a little box alongside the data. This is supposed to be your Insights or Analyze box; I don't remember exactly what it is called where there is an option to create lots of different charts. Why would anybody want to create charts right out of the box? Usually, the process would be that you have data, you identify what you want to convey from this data and then find that information from the data, and then create charts. But, as soon as you have some raw data in Excel, because Excel cannot distinguish between data and information (everything is a number for it), it will show a suggestion box right next to the data and ask/provoke you to create a chart. Probably, this will prompt many people to go and experiment with a 50 slice pie-chart, or a 76 column bar-chart etc. Those will be disasters as you can imagine and understand. So, lack of time and convenience of creating charts are two reasons. There is no similar box that will provoke you to do a SUMPRODUCT or an INDEX-MATCH on the data. Of course, there are auto sum and auto count features but we know that those are rudimentary. If you want to do serious analysis, you are not going to do auto sums. You are going to do a lot more than that. Since the convenience of creating charts is available, people may do data dumps.

The other reason is that people **don't know better.** This is something that happened with me when I as learning as an Analyst. Very early on in my career, I have created many data dumps. Every time that I presented these to my bosses, colleagues or clients, there would be this dead silence in the room. People were too embarrassed and shocked that somebody would come up and present something like that. Maybe they were just too nice and didn't want to hurt a rookie's feelings and so nobody would tell me. But, at a later stage in my career, maybe after 3-4 months of doing this, I realized and got feedback that doing data dumps is really wrong. So, I went and tried to study better ways to visualize and analyze data and all that. So, **lack of skill** is another reason. That is excusable when you are in the very early stages of your career. When you are fresh out of college and don't know any better, you create a pie-chart with 50 slices. But, that is not excusable after 10-15 years of working. But, people still do this. I have had many colleagues and superiors who have been working with me for 5-10 years who would still create these kinds of data dumps. Again, I am not sharing all this to put them in bad light or anything. It is just that the lack of skill is such a pervasive thing and that's why we see data dumps.



If these are the reasons why data dumps happen - lack of time, lack of skill and convenience of creation - **how do we avoid data dumps** as smart and awesome Analysts or Managers or CEO's? How do you avoid data dumps? The process is very simple. You go from data to information and then **do an information dump**. A data dump is bad whereas an information dump can be awesome or overwhelming. Any time that you see a lot of information, the first thing that you get in your mind is that it is overwhelming. But, the second thing is that you get a lot of joy. You, as in either you or your audience, can see a lot of information. You can see and connect the dots and draw a full picture and make more sense out of the information. So, try to go from a data dump to an information dump.

But, what if you don't have the time? You don't have 7 hours in front of you to analyze all the million rows of data. How do we then avoid data dumps but still make good use of the 10-15 minutes of time we have? The easiest choice in such cases is to at least sort your data dump. Don't put everything just like that. Instead, try to arrange them in some sort of meaningful order. It could be alphabetical or by size or by volume or by growth rate or whatever. At least arrange it so that if you have a 50 column chart with sales for all the 50 products, you could at least arrange them in descending order of sales. Although, technically, it is a data dump, it is now presenting some sort of insight. It tells the people that these are our top products and these are our worst performing products. That kind of takes the data dump slowly towards the information dump. So, **sorting is one way to kind of overcome the data dump problem**.

The **second option is filtering**. You can take all the 50 products and create a column chart but maybe add a filter so that users can filter by the product category. So, they don't have to see all the 50 products all the time. They can choose to see only cookies or baby-care products or groceries or whatever and only those categories of products will be displayed and you can see them. To do this you could use features like slicers or form controls and these are pretty enjoyable and fun experiences. So, even in my contest that I recently ran, we had some data dumps that I have accepted as contest entries. I have discarded the other data dumps but accepted some because people have used features like slicers or form controls so that the chart does not feel like a data dump. It feels more like a controlled data dump. So, sorting is one option and filtering is another option.

If you can't avoid a data dump and you really don't have any time then, instead of making a chart, just give your users the **raw data as a table**. Just format everything as a table and give it to them and maybe slap a bit of **conditional formatting** on top of it like a heat map or icon sets or sparklines or data bars or whatever so that it doesn't feel overwhelming. When we are looking at a table of data, we naturally expect that the table will have more information. Of course we are not used to seeing a table with a million rows everyday but we are much more used to seeing a 50 row table than a 50 slice pie-chart. So, you can give your users a table and maybe arrange a sort order on the table. That will feel more like raw data rather than a data dump.



These are the ways to **avoid data dumps**. I'll quickly **summarize** them. The first and best way to avoid a data dump is to **go for an information dump**. If you cannot do the information dump and you don't have time then **sort your data dump**. Also, **consider adding a filter** on your data dump. So, sorting and filtering are powerful ways to reduce the amount of data in the data dump and kind of improve the chances of people not getting offended by that. Finally, **if you must give all the data, just give it as a table** rather than a chart or something like that because charts distort the data and many times take a lot more effort to understand as compared to a raw data table.

I have a couple of resources for you. In case you want to make better charts, I would recommend that since you are already listening to the podcast, listening to a few other podcasts is the best way to learn. There are three podcasts that I would recommend:

1. **Episode 29** (<http://chandoo.org/session29/>) which talks about a **6-step road map to create awesome charts** that impress your bosses and colleagues. This road map will help you create better charts every time.
2. **Episode 32** (<http://chandoo.org/session32/>) which talks about **5 rules for making awesome column charts**. Many times we use column or bar charts, and these rules will help you make better column or bar charts.
3. **Episode 38** (<http://chandoo.org/session38/>) which talks about **data-to-ink ratio**. This is a key principle and if you understand it and apply it well, your charts are going to look world class and awesome all the time.

So, the episodes are 29, 32 and 38 and you can access them at <http://chandoo.org/session> followed by the number 29 or 32 or 38.

That's all for now. This is a short format podcast and so I am not going to talk for all 30 minutes. I will stop now. I hope you enjoyed this particular episode on data dumps. If you come across any data dumps or any interesting stories about it from your past or your work environment, or have a friend or colleague who keeps doing them, please share the story in the comments section of this episode at <http://chandoo.org/session49/>. You can also access some resources about making better charts and how to avoid data dumps on that page. In case you are enjoying <http://chandoo.org> podcasts, I would really appreciate if you can take a minute and write a review for our podcast on either iTunes or your Android podcast page. Please visit <http://chandoo.org/session49/> to access a link to drop your review.

Thank you so much. Have an awesome day. Bye.