Transcript for Session 010

Listen to the podcast session, see resources & links:

http://chandoo.org/session10/

Transcript:

Welcome to chandoo.org podcast session 10. Chandoo.org podcast is designed to make you awesome in data analysis, charting, dashboards and VBA using Microsoft Excel.

Thank you so much for joining us in yet another episode of our podcast. In this episode we are going to talk about why 'averages are mean' and what kind of Excel techniques you should be familiar with. Please keep in mind that this is a continuation of the previous session of our podcast, and so this is actually part of our 'averages' or 'mean' podcast.

Since this is a continuation, it makes sense to do a quick recap of the techniques or ideas that we already discussed in the previous episode. In that we understood why averages are not such a good idea in many business scenarios and what to do about it. We started the session with a very simple example of sales for five different people in a typical sales department scenario, and understood why calculating the average of those five people might not reveal much.

We talked about some different examples around that and then we discussed five statistical concepts that are vital to understand if you want to analyze data better. The very first concept is standard deviation, which is really a number that can explain how spread across the values in the data are, as compared to the average. As an Analyst, you can make better sense of the data if you have both the average and the standard deviation. If the standard deviation is very high as compared to the average, then you could say that maybe the data is too spread across and all over the place and validates considering some other analytical choices rather than printing the average in your report.

The second concept that we talked about was the median which is the middle point of the data when arranged in ascending or descending order.

The third concept that we talked about is quartiles. There are two types of quartiles - 25th percentile and 75th percentile. These are similar to the median in terms of the definition but what they signify is different. So, the 25th percentile value tells us the value at which 25% of the items are less and 75% of the items are more.

Then we talked a little bit about outliers. This is where we introduced the example of Bill Gates house and we talked about not wanting to include extreme criteria like Bill Gates' house or a foreclosed house, while analyzing house prices in a county. Those things do not make much sense from a pure data analysis point of view, so it is important to understand what these outliers are and to treat them separately.

The fifth concept that we talked about was the distribution of values, i.e. how values are distributed. At this stage we also discussed the concept of box plots although we didn't delve into it too much. However, you understood what a box plot does.

This is where we concluded our session in the previous podcast. We talked a lot about conceptual things although we used plain English or business terminology as much as possible. Still all of these things are kind of nebulous; we didn't really understand how the concepts would play if you're making a dashboard, report or an analytical model.

So today I want to address those concerns and showcase some idea and techniques that are very useful when you are trying to implement these concepts using Excel. How do we analyze data applying those concepts? - That's the focus for us.

There are eight different things that I want you to be aware of; these are not exclusive of each other and you can combine these in any fashion so that your analytical or information needs are met. That's very important. There's no such as the golden measure or the golden statistic. In real life there is no such thing. As a smart Analyst, you should ask questions like - "there are 25 different things that I can do with this data so what 'n' numbers of things should I do?" For example, "should I do 3 things or should I do 4 things or should I do all 25 things?" That's what a smart Analyst would ask and he/she would eventually narrow it down to a meaningful set of statistical items and then go and calculate them and present them in the reports and dashboards.

This is where the 8 items that I have in mind come into the picture. From a podcast format, 8 seems to be a good number; even if you retain just half of these you would still walk away with a good amount of information.

The very first thing that I have for you is really a generic suggestion – 'start with the average'. When you get a bunch of data to analyze and you don't know where to begin, the average is always a good start. We know that it isn't the best way to represent the data or present information about the data so that decision makers can understand it, but it is an excellent start. Always start with the average. This is the easiest one. You would use the Excel formula '=average()' and just calculate the average of the numbers. So, start with average. Don't stop there; but use it as a starting point. In the earlier podcast, we hammered the concept that averages are really a bad way to represent data, but all said and done, probably 90% or more of business people easily understand averages. The moment you put up a slide or report that says that the average sales volume is 300 units, everybody would understand that instantly. You don't need to spell it out or explain it. On the other hand, if you present the following in your report - "the first quartile of sales volume is 74000" - I am sure that quite a few of your colleagues or Managers would come back and ask you what quartiles are.

Not many of us are familiar with quartiles. We learn the concepts of quartiles, median, standard deviation, variance and other things in school itself but we don't really connect the dots between that and what it means to our business immediately. I'm not trying to say that all Managers or clients are like that. I have come across quite a few people who know these things and demand them as opposed to just a plain average. But, it's always a good idea to start with the average.

As I said, I will talk about 8 different techniques and for many of my reports or analytical needs, I usually

mix and match them. So don't stop at picking one of these eight choices, instead try to pick a couple of them.

When it comes to the average, there are two other concepts that you should be familiar with. These are number 2 and 3 of our 8 items. The first one is - 'start with average'. The second one is 'moving averages' and the third one is 'weighted averages'. Let's understand what these two things are.

Talking about moving average, let's say that you are analyzing the sales of chocolates and you are working in a very big Supermarket like Target or WalMart, you're the head of the Chocolate department and you are selling millions of chocolates in any given year. This is quite an enviable position since you can pick up any candy or chocolate and eat it without anyone questioning you! This is the kind of power you have as the head of the chocolate department. Let's say that for some reason you want to understand the average sales of the chocolates you've been selling. WalMart is a very big store and imagine yourself as the head of its chocolate department, if there is such a thing. It's a massive company with hundreds of stores across many countries. They probably push about a billion dollars in chocolates alone every year (a wild guess!). When you're looking at so many items being sold, trying to print out all the individual transactions pertaining to chocolates and look at all those numbers to make sense isn't going to help, as you'd just drown in data.

So you might think of filtering the items where one of the items is chocolate and then averaging the total sales amount to get a picture of how much you are selling in an average sale. The biggest challenge for you here would be the fact that it's a really old company and they have been selling various things including chocolates for decades. To humor me, just imagine for a moment that all this data is residing in Excel instead of a database. So you're in Excel and you've filtered out and removed all the unnecessary items. You have the transaction date and time stamp in one column, in another column you have the type of chocolate that they have purchased, in another column you have the quantity and in the last column you have the total amount that they have paid.

A typical row of data would be that on 1st January 2012 somebody purchased 10 chocolate bars and paid $12. That would be a typical transaction detail in the data and you have millions of rows of this kind of information. You're looking at all this data and you want to calculate the average, so you average the amount column. Excel average is a very fast formula so it spits out a number like $7.25. This shows that $7.25 is the average price you are receiving for the chocolate sales. It's a good number and gives you an indication of what an average sale looks like. But the challenge is that it's not showing you a proper picture because you're looking at the data for the last several decades. You've calculated the average for all the data whereas to do any proper, meaningful analysis or make any decisions - for example if you have a hunch that people are eating more chocolates due to medical research or ongoing fashion trends, people have realized that eating a bit of dark chocolate every day is good for their heart or something like that (wild guesses here!) - So people have been purchasing chocolate and you want to see the trend - but the $7.25 per sale does not seem like a lot. This is because you are the head of the chocolate department and so you have a lot of market intelligence. And, you read in a research report that, on an average, Americans are purchasing $75 worth of chocolate every year.

But, your data is telling you something else. It's telling you that the average sales volume of chocolate is only $7.25. So, this and the $75 figures are contradictory. You're not really sure whether you should believe the data or the research report or come to a conclusion that, as a company, maybe WalMart has

a lot more to do before they can move up the average amount from $7.25 to $75. So you're not really sure.

Then you realize the mistake which is that we are averaging all the data way back from 1920. This is taking up all the tiny dollar amounts that people were spending in early years and adding them up with the high dollar amounts that people are presumably spending in the later years. That's why our average figure of $7.25 is inaccurate. It's not giving us the complete picture of what is what. This is where the concept of moving average comes into the picture. I realize that this is a really long introduction to moving average, but when I talk about chocolates, I'm like a kid and I obviously get excited!

Coming back, a moving average, in a very loose business sense, gives you the average for the last 'n' values like the last 100 values or the last 15 values etc. So, instead of averaging the values from 1920 to 2014, it would be a lot better if we could just average the data for the last 12 months. That's taking a subset of the data and calculating the average for it. If we do that, what we are calculating is technically the moving average. We're only calculating the average for the latest 12 months. When you do that, you might even come to the conclusion that you are selling about $72 of chocolate on average which is good and a lot better than the $7.25 figure that we got earlier. Again, there's nothing wrong with that figure as it's just giving you the average for all the numbers. This is what moving average is.

I want to clarify that the explanation that I gave for moving average is correct, but the definition is in very loose terminology. In the real world, moving average means that you would calculate such averages for every 'n' values. Imagine you are analyzing the sales for your department - instead of analyzing the sales per person, you are analyzing the sales by month - and you're looking at 24 months of data. The average for the 24 months would not be meaningful, but if you could take the average for 12 months, there would be 13 different averages as each of the 12 month combinations would be taken into account in succession. So January to December in the very first year would be the first, then February to January of the next year would be the second average etc. In this way you would end up with these average figures and this is what moving average signifies.

Again, from a pure calculation or Excel implementation point of view, the way to calculate moving average is very simple and straightforward. I could give you the formula in this podcast, but I realise that you're probably driving, commuting, or on a morning walk or evening jog, so now is not the right time for you to memorize the actual syntax of the formula. Instead, I am going to leave a link to a moving average example and a detailed tutorial on the show notes page. You can go to http://www.chandoo.org/session10/ to access the example and tutorial. That's about moving average.

The third item in our list of 8 items is weighted average. This is similar to moving average because it's also an average but what it signifies is something else altogether. Let's go back to the supermarket example, but this time let's not make things too complicated. Instead, let's keep it really simple. Imagine that you're running a supermarket that sells only two items, eggs and milk. On each and every aisle of the supermarket, all you can find are cartons of milk and eggs. You have one aisle dedicated to eggs and the other to milk. The store name is 'Eggs and Milk!'

You're looking at the sales figures for eggs and milk and your sales supervisor tells you that in the prior week you sold 500 boxes of eggs and 500 cartons of milk. Being a smart person, he goes on to add this nugget of wisdom - 'our average is 500 units' - since it's 500 units of eggs and 500 units of milk - adding

them up and dividing them by 2 categories gives us an average of 500 units per category. But, you know a lot better because you're listening to this podcast! So obviously you go and say that the statement doesn't make any sense. You can't add up milk and eggs. You want to calculate a better average. So what would you do in this case? A better number would be if I could somehow know the average revenue per category. We have only two categories so we might do a lot better by exploring it individually, category by category. But again, humor me, and think about what would happen if we want to know the average sale per category.

If I use the volume, we get 500 units per category as the sale, but that's not accurate. It's not really giving us the real picture as eggs and milk are two different things. So you'd go a little deeper and ask a question like - 'what is the price of a box of eggs?' In response, your sales supervisor tells you that it's $2. A box of eggs is sold for $2. Then, you also ask the question - 'what is the price of a carton of milk?' And, your sales supervisor tells you it's $4 per carton.

So you have these 500 units of eggs sold at $2 per box and 500 units of milk sold at $4 per carton. Now that you have these figures, the better way to calculate the average would be to weight it.

The weighted average concept is to simply take the 500 units and multiply it with the price of eggs and add to it the other 500 units after multiplying it with the price of milk. 500 units multiplied by $2 amounts to $1000 which is the revenue generated by eggs. Similarly, the revenue generated by milk is $2000 (500 units *$4). Our revenue is $1000 from eggs and $2000 from milk. If I calculate the average, the total revenue of $3000 would be divided by 1000 units. So, we would say that, on average, we're making revenue of $3 per item. This is what a weighted average is. Again, here we're no longer talking in terms of quantity. We've moved to the dollar domain and are talking in that direction. A weighted average concept is very popular and a better way to calculate average especially when you have disparate things like eggs and milk and you're trying to add them up.

There are many places where weighted average would be useful. On the show notes page, I will link to an article that clearly explains the Excel formulas that we need to use to calculate weighted average and showcases some examples of weighted average. Please go to http://www.chandoo.org/sesssion10/ to access the weighted average example. That's the third way to analyze the data.

The very first is 'starting with averages'. The second is using moving average or weighted average or both, depending on the type of data or situation that you have. These are three main techniques that surround averages.

The fourth is to use median and quartiles. Again, in the earlier podcast, we hammered the concept of median, quartiles and what they do. Now comes the time for you to calculate these things. There are two formulas in Excel that will help you do this.

The first one is the median formula to which you can pass off a range of values and get the median in return. It's a very straightforward and simple formula.

The second formula is called percentile and it can give you any percentile of your data. What is the first quartile? It's essentially the 25th percentile. So you would pass on the range and 25% to it and it'll give you the value of the 25th percentile. Likewise, you pass on the same range and give 75% as the value

and it'll give you the 75th percentile of the data. You'd use the percentile formula to get these. There are many variations of these formulas because in the actual science of statistics, the way you would calculate a median, standard deviation or percentile differs based on the kind of data that you have and whether you have the entire data or just a sample of it.

Excel has a lot of functions that'll help you. There's something called inclusion and exclusion as well. Again, these are too complicated from a business user point of view because I have never used the variations of the percentile formula personally. There are newer versions of the percentile formula that fix the floating point and some other calculation mistakes that usually happen in the earlier versions of Excel.

To cut the long story short, the important formulas you need to memorize are the median formula and the percentile formula. That's the fourth technique.

Again, the four techniques that we talked about are 'starting with averages', moving average, weighted average, and median and percentile.

The fifth technique is to try to use some outside benchmarks. This doesn't have anything to do with Excel. When you are calculating averages and are trying to present that information to your management, it makes a lot of sense if we could also figure out the benchmarks from outside. For example, let's put ourselves in the shoes of a car park manufacturer. We make very specific parts for cars. Let's take the steering wheel since it's a very simple part that we all can visualize. So you make these steering wheels and you sell them to various car makers in the world like BMW, Mercedes, Audi and Toyota. You calculate the defect rate for every thousand steering wheels. You manufacture millions of steering wheels since there are millions of cars all over the world and you are one of the prominent suppliers. For every thousand steering wheels that are manufactured in your plant, you maintain a defect log where you check the quality of the steering wheel in terms of whether it is round enough and has everything neatly mounted etc. If there are any defects, you add them to the defect log. So you maintain these defects by the lot, where each lot is 1000 units of steering wheels. So, the log might look something like this:
Lot 1 - 3 defects
Lot 2 - 1 defect
Lot 3 - 4 defects, etc.

After all the data is collected, you want to understand how many defects you are making on an average per lot. So you take the entire thing, average it out and reach the conclusion that, on an average, you are making 2 mistakes per lot. From a business point of view, you are quite satisfied because you feel that 2 mistakes are not bad since it means that you are 99.8% accurate and that equates to good quality for many business situations. If someone can boast of having 99.8% quality, it's a good number to rely on. But, that's where you might be making a mistake. By just looking at the average alone, you won't get the entire picture. If you could somehow benchmark your average with the industry average or with a competitor's average, then you get better insight. For example, your average mistake rate is 2 per lot, but you know that there is one other significant competitor as well. - And you have insider information, because they publish the rate in the stock market report or filing, and from that you know that their mistake rate is only 0.5 units per lot. For every 1000 steering wheels that they manufacture, only 0.5 steering is broken and everything else is perfect. This makes their defect rate to be 99.95%. So you feel

that you're in bad shape. Even though earlier 2 defects per thousand seemed like a good number; however when compared to 0.5 defects per lot from an outside average, you don't look good anymore. It could also happen the other way around - they might be making 10 mistakes per lot, so you're in better shape. Whatever may be the case, in order to get a better picture about your data and averages, sometimes you need to juxtapose it or add extra data from outside, which might be an industry average or benchmark or a competitor number or a number from your KPI target etc. So this is another technique that you can use. All of this can be done in Excel - you calculate the average and compare it with the benchmark so that you can see how good or bad the average is as compared to the benchmark.

The sixth technique is conditional averages. This basically refers to calculating the average for a subset of the data that meets certain conditions. For example, if you have a lot of invoices and you want to calculate the average duration that the customers are taking to pay the invoice. You're sending these invoices to 5 different companies that you deal with - Microsoft, Google, Apple, Samsung and Nokia. And, you know from experience that Microsoft is a very good client. They take their business very seriously; they immediately pay back as and when an invoice is sent. You have this invoice data and Microsoft invoices are almost always paid on the same day on which they are sent. Whereas all the other four companies usually pay within the due date, which is 30 days from the date that you send the invoice. If you send the invoice today, you're likely to receive the payment from those four companies within the next 30 days. Whereas, in the case of Microsoft, the money comes through immediately. You send the invoice today and the money is paid by tomorrow. You're looking at this data and you're analyzing the average time taken for payment. If you calculate the average for all of those dates, you won't get a good picture because you know that there is an outlier in the form of Microsoft. It would be better to exclude Microsoft and calculate the average.

This is where the Excel function called 'averageifs' comes in handy. You'd say:
=averageifs(all these numbers,company name,"<>Microsoft").

For this, Excel would calculate the average for the durations where the client name is not Microsoft, and it would tell you that on an average, if you remove Microsoft, the time taken is taken 20 days whereas if Microsoft is added, the time taken would be only 12 days. This is where the concept of conditional averages is useful.

There are two more really generic techniques that we are going to talk about. Sometimes, it is a lot better to visualize your data even before you calculate the average or any other metric. One simple way would be to take the data and sort it, after filtering out anything that you don't need. Once it's sorted, just select only the numbers that matter to you and create a line chart or column chart from it. Once you've sorted the data, it kind of exhibits the pattern that the data is following. This way you can see where the values are. If you're sorting the data of sales people and their sales volumes, and you're sorting it by sales volume. - When you sort it by descending order, the highest values would be at the top and lowest values would be at the bottom, and when you make a chart out of this, you get a spread of the values in a line or column chart. You can immediately see how tall the highest value and how short the lowest value is. If all of these appear to be within a band of a narrow range, it means that all the values are within a meaningful range and so you could calculate an average of them and explain them.

But, if the values are explained by a steep slanted line that goes from top to bottom, it means that the

values are all over the place and you'd be better off explaining them in a different way like maybe by excluding the outliers (the top performers and bottom performers) and then calculating the average. Or, maybe do something else like present the data in a way so that the top 10 or bottom 10 people can be shown. Likewise, if the data is for monthly, weekly, or daily sales, then you would sort it on the total volume or sales column, visualize it and see how these two are different and if there is any pattern that is exhibited by it. Then go and explain it with the right type of average, median or percentile data.

Visualizing is a very powerful way - this is not the final visualization that would go into your dashboard - but, it's something that you are using as an Analyst to understand the data better. When you have thousands of rows, it doesn't make sense to scan the entire thing and try to gauge it with your eyes. Instead, you could make a simple chart very quickly, after filtering out the things that you don't need, understand the spread of the data and then define your analytical motives.

The last one is - 'don't stop at averages.' We talked about this a little bit in the earlier episode as well. There is no point making a presentation or a dashboard that has a statement that the average sales are $24 or the standard deviation is $6. Most people would understand what average sales of $24 mean, but as a good Analyst, you should go one step further. Go ahead and explain what it means for a business. What does it mean when we say that the average sales are $24? What does it mean when we say that the standard deviation is 900? What is it that we want our Managers or clients to understand from statements like these? That's what a good Analyst should explain. Put it in plain words or depict it with a chart. So don't stop at averages, but explain what they mean.

This could be done in many ways. I'll share some of my favorite ways. I usually mention the average and then right beneath it, I also include some information about outliers. For example, I would say:
Average home price = $300,000
(highest priced home = $7 million dollars, lowest priced home = $60,000)

This will give a better picture immediately, rather than just the average price alone. Likewise, I also do comparisons with previous periods. This is very useful in business scenarios where we are reporting data for the latest month, quarter or year. If I am saying that the sales for April 2014 are $700,000, I would also add that the sales in April 2013 were $620,000 - which is the sales in the same month of the previous year. This will give a better picture of whether the figure of $700,000 is good or bad in relation to what happened in the past. Similarly, you could also add the target values, KPIs, benchmarks or industry averages etc.

This is what I mean by the statement - 'don't stop at averages'. If you stop at averages, it will be a very mean thing! It's just like putting a skirt on your dashboard. It's not going to reveal the full story. Instead, you should go a step further. Anytime that you are calculating the average and representing it in your report, ask yourself what the average is doing. What kind of insight, meaning or information is it trying to convey? If it's not conveying anything, remove it and put something else there like the median, quartile or distribution. That's how you should averages in an analytical situation.

To summarize the eight techniques that you should be using are:

- Start with averages
- Consider using a moving average or weighted average, depending on the data. In some cases, you need to use both,

- Consider using the median or quartiles (by using the median or percentile formulas in Excel).
- Compare with outside benchmarks like an industry average or competitor measure and co-relate those with your figures. Or, at least show them next to your numbers, so that you can assess if your numbers are good or bad.
- Use conditional averages, like the 'averageifs' formula in Excel, so that you could exclude or include the data what matters to you most and calculate the average only for that.
- Visualize the data before you calculate averages or any other type of metric. This helps you to understand the spread or distribution of the data better and then you can make a better statistical representation of the data, depending on that.
- Don't stop at averages. Instead, go an extra step and explain it.

That's about it for averages. I hope you have enjoyed this 2-part podcast on 'why averages are mean.' I hope you will apply these concepts to your day-to-day work and make better averages or reports. I use most of these techniques in my dashboards or reports all the time. I'm still learning a lot about these things, but I realize that I have progressed quite a bit from my early days when I would calculate the average for anything or everything. So, I'm in much better shape because of all these techniques and ideas that I have learnt. And, I want you to be in the same place.

Thank you so much for listening to this podcast and if you have enjoyed this episode, please go to http://www.chandoo.org/session10/ and leave a comment and tell me how you are applying these techniques. Also, if you are using averages in your business reporting, please tell us how you use it - the scenarios where you find averages to be quite useful and meaningful, how you avoid the mistakes or pitfalls of averages in your reports, and the kind of additional metrics you calculate so that the average values are explained better in your reports.

So, please share your experiences, thoughts, tips, techniques and tricks on our website. Please visit http://www.chandoo.org/session10/. On this page, you'll also find all the show notes, resources and links that I have been talking about, especially the articles on moving average, weighted average and a few other examples.

If you like the podcast, please take a minute and leave an honest review on our iTunes page. You can go to iTunes and search for chandoo.org. Or, you can go to http://www.chandoo.org/itunes/, which will re-direct you to iTunes to leave your feedback about our podcast so that more people can discover it and become awesome.

Thank you so much again. I know you're pretty awesome. Stay awesome and keep learning. Bye.