Transcript for Session 013

Listen to the podcast session, see resources & links:

http://chandoo.org/session13/

Transcript:

Hi there. Welcome back to chandoo.org podcast session 13. Chandoo.org podcast is dedicated to making you awesome in data analysis, charting, dashboards and VBA using Microsoft Excel.

First of all let me thank you for joining me in another episode of our podcast. I really appreciate your taking time to learn from me through this audio podcast method. I hope you are enjoying chandoo.org's podcasts. If I may ask you, would you please take a minute or two at the end of this podcast to go to our iTunes page and leave your feedback there - whether you like the show or you have some suggestions for improving the show. Please leave your comments there so that I can help you even more. Any feedback that you leave there will help us to reach out to more people and share this knowledge with them. Thank you.

Now let's talk about one thing that's on most of our minds at this time of the year. A quick reminder - this podcast is recorded on 1st July 2014 so some of the material that I will talk about here is time sensitive. So if you are listening to this way out in the future, you might think that this is not interesting anymore!

The thing that I am talking about is the FIFA football world cup. This is an event that happens once every four years and it is an event unlike anything else that humanity sees. For example, the last time that FIFA world cup took place back in 2010, estimates say that almost a billion people around the world tuned it to the final match between Spain and Netherlands - they followed it, watched it on TV, were present in the stadium or tuned in to the radio etc. A billion people! That's almost 20% of humanity following this one particular sport! So it makes football one of the most popular sporting events on earth. I know that my friends, colleagues, listeners and readers from the United States and from Canada (to a large extent) might think, "Hey, football - what is it? Isn't it the same as American football?" That's changing. I think that this year many of you from the States might also be enjoying football because the United States reached the last 16 stage. And, as of today, on 1st July 2014, they are playing against Belgium in the last 16 match.

The reason why I am talking about football is because we are going to do something really interesting in this podcast episode. We'll talk about football, but from an analysis point of view - how you would use an example from football, as an Analyst, to learn things and do better analysis.

Before I jump into our topic, let me also tell you the teams that I am rooting for - Netherlands, Germany, Argentina and Brazil. I'm following these teams actively. Last night has been really hectic because there was a match between Germany and Nigeria which was very interesting. Anyway, let's talk about the topic of the day.

I just want to quickly remind you that you can visit http://chandoo.org/session13/ to access all the show notes, resources and links that I talk about in this podcast.

I also have a special bonus tip at the end of this podcast on the theme of 13. If that sounds intriguing and interesting enough, stay tuned!

Let's talk about a familiar problem now. I see many business people rely on this. While watching the football match the other day (I think on the 23rd of June), the commentator made a statement towards the end of the match which went something like, "This has been a FIFA world cup of late goals." This is because there were no goals for the first seventy odd minutes in that match and then there was a goal. And, while people were still celebrating the joy of a goal (I think it was Netherlands that scored the goal), the commentator said that this has been a world cup of late goals. That got me thinking - "Hey, is it really a world cup of late goals? Or, is he just saying it because that's how he felt?" Now, you might be thinking, "Hey Chandoo, how does this relate to the world of analysis?" Let me relate some examples.

We hear these kinds of statements quite often in business settings - "product A is selling slower than the rest of the products". We hear these kinds of statements in sales presentations, status/review meetings or business planning sessions. People also say things like - "productivity of our Texas plant is way better than all of our other plants." We could consider this as a generalization or it might be a statement backed by solid data. We don't know. People make the statements and we just take it for granted.

Another example is - "we seem to pick up more sales on the weekends." Again, many business people make statements like this.

The statement made by the world cup commentator is no different. He said - "it has been a late goal world cup so far". As an Analyst, your antenna should go up as soon as you hear these kinds of statements. The very first thing that you should do is validate the statement. In many business settings it's very simple. You could just go to the data and check if that hypothesis or statement is true.

Let's do that the same thing with the statement that the 2014 world cup has been a late goal world cup so far. To understand whether this is a true statement or a statement made out of emotion (because the commentator was waiting for 70 minutes to say something exciting related to the match!) we need to gather the data and analyze it. That's what I did.

The very first thing that you should do whenever you hear a statement like this - a generalization that is probably not backed up by any data - is to set up a context. It was a match that I was watching on television, so I had no real idea what the context of the commentator was. I had to make a reasonable assumption. So, I assumed that when somebody says that it has been a late goal world cup so far, they mean it in the context of comparing it with other football world cups. So I assumed that the commentator was saying this in relation to the previous two world cups that we had, i.e. in 2006 and 2010. Technically, what he was saying was that in comparison to the previous two world cups, we are having a lot more late goals this time. This is one context.

When you want to prove/validate or invalidate a particular idea or hypothesis, you should set up a context around it. So, the context that I chose are the three world cups - 2006, 2010 and 2014. If you

look at the goal time data for all these three world cups, the thing that we want to validate or invalidate is that 2014 has had more late goals than the other two.

Likewise, you should also add some context to exclude anything that is too tricky or difficult to track. For example, in this scenario, I assumed that we can safely exclude the goals scored in the extra time. For those of you who have no idea about football, let me quickly explain how it works. I am not an expert; I think I have probably played a sum total of 3 football matches in my life and I didn't score in any of them! I think I was kicked out of the very first game I played at School because I was doing something silly and ridiculous that put my team in jeopardy!

A typical football match is of a duration of 90 minutes - teams play for 45 minutes, take a break for 10 minutes for drinks, regrouping and forming strategy for the next half and then they play for another 45 minutes. That's what happens in a typical match. Within these 90 minutes if one team is dominating the other by scoring more goals, then the match is considered done. So, at the end of 90 minutes if the score is 1-0 or 0-1, then you're good to go. Football can get really boring if you don't like the sport or if you're only watching it for action because sometimes you have to wait all of 89 minutes to see a goal. Many times, at the end of the 90 minutes, neither team has scored a goal. Or, both teams have scored equal goals. At such a stage, a match is considered to be drawn. A drawn match has an equal score at the end of 90 minutes. In certain competitions, a draw is allowed. If you're playing at a stage of the competition where there is no necessity of eliminating a team, a draw is okay. This is what happens in the world cup as well. In the initial stages, a match will not continue beyond 90 minutes if there is a draw. Both teams will get 1 point each in such a case. If a team wins, they get 3 points.

However, what happens if there is a chance of elimination? It's just like tennis. At the end of a set, if both players have an equal score, you have a tie-breaker in tennis. Similarly, there is a tie-breaker in football too. The way it works is that the teams are asked to play for another 30 minutes. So there are the first 90 minutes and then another 30 minutes of extra time. Within these 30 minutes, the teams play for 15 minutes, take a break for a minute or two, and then continue for another 15 minutes. If the match is still drawn after these 120 minutes then the teams have a penalty shoot-out. All this is too technical and irrelevant for our podcast. I'm only explaining these terms to clarify what extra time means.

From my basic analysis I concluded that there were only five goals scored in the extra time in the previous two world cups and the current world cup. This is as of 23rd June 2014 (the date on which I did the analysis). There were more than 400 goals scored altogether so far and only 5 were scored in the extra time.

So I thought that instead of analyzing all the data, I could exclude those five goals for the analysis. It will keep the analysis simple since I'm analyzing only the first 90 minutes of the data. Anything after 90 minutes is too minute; 5 goals out of 400 is only 1% of the data and we can ignore it. This is also another way of setting the context. The context is defined for our analysis as the current world cup and the previous two world cups, i.e. 2014, 2010 and 2006. And, the goals scored in the extra time (after the 90 minute mark) are excluded. This way we can put a boundary around the data that we are considering and focus on the analysis.

In many business cases you would do the same. For example, if somebody makes a statement like -

"Product A is selling a lot slower than other products" - we need to set a boundary in order to prove or dis-prove the statement. Is the statement in relation to the current month's sale, or the last twelve months sales, the previous week's sales or the latest quarter's sales? What is it? We have to set a boundary on the timeline, the products that are to be included in the analysis etc. Is product A to be considered along with the other products in the product group or with all the products that the company is selling? This is very important. We need to set the context and for our analysis the context has been set.

What I did next is that I made several attempts to validate whether the data is validating the statement made by the commentator. I will walk you through the various attempts that I made and I'll try to make them as clear as possible in an audio podcast format. If you have any difficulty visualizing what I am saying (there's not much to visualize, it would be easy to imagine since most of us follow one sport or another), head on to http://chandoo.org/session13/ where I will add some charts and link to an article that I've already written on this topic.

The very first attempt that I made was - "what if we could visualize the distribution of goals on a timeline of 0 to 90 minutes?" 0 to 90 minutes is the boundary of our analysis and so I thought that it would be good to add the distribution of goals on a timeline, i.e. whenever there is a goal on a timeline of 0 to 90 minutes, we'll add a dot to the timeline at that time. If there are more goals around a certain point of time, there will be more dots there. If there are fewer goals at a certain point of time, there will be fewer dots. The thing that I was trying to identify here is that according to the statement of the commentator we should see more dots towards the end of the timeline (probably in the 60 to 90 minutes mark) for 2014 whereas we should see more dots at the beginning of the timeline for the previous two world cups (2006 and 2010). That's one way to say that 2014 has indeed been a slow or late goal football world cup.

I plotted these distributions depicting three lines with lots of dots on them - there is a line each for 2006, 2010 and 2014. One thing to consider here is that the number of goals scored for each world cup is different. For 2006 and 2014, we have more than 145 goals in each of them - I think 145 in 2006 and 147 in the 2010 world cup. However, since the 2014 world cup is ongoing (as of 23rd June when the commentator made the statement and even as of today, 1st July 2014, when I am recording this podcast), so we only have partial data for 2014 whereas we have the entire data for 2006 and 2010. This is something that you need to keep in mind as there will obviously be fewer dots for 2014 but that's because we have less data. As we progress, we may see more goals in 2014 and the density of dots would probably look more or less the same as 2006 and 2010. Nevertheless, I went ahead and plotted the data. And finally I could conclude that there was no significant disturbance in the dots. The dots seemed all over the place for all three years. There was no discernible pattern to say that the earlier world cups had more dots in the beginning of the timeline as opposed to more dots towards the end of the timeline for 2014. There was something that could be spotted but it could be easily attributed to having fewer goals in 2014 since we have only partial data for 2014 as of now. This is something that I realized after the first attempt.

I made a few more attempts but I'm not going to go into the details in the podcast. However, I'll add a link at http://chandoo.org/session13/ where I'll link to the entire analysis and the article on the 2014 world cup football goals.

The second thing that I thought about was that maybe we should only consider the first 100 goals of each world cup. This was because we are kind of comparing apples to oranges in the earlier scenario. Instead of comparing all the goals of the previous world cups with partial goals of the current world cup, what if we made it consistent by taking the first 100 goals from each world cup and compared their distributions. This is similar to using the YTD, MTD or QTD sales. When we are analyzing year to date sales of a product, we should take the YTD of the previous year in order to prove whether it's more or less than the previous year. We should not compare the first six months sales of this year with sales for 12 months of the previous year. It's always going to be less and we might think that we're doing badly. But that's not true. If we compare six months of this year with the same six months of the previous year, we might be able to conclude a valid story and prove or dis-prove something.

I decided to do the same. I took just the first 100 goals of all the three world cups and plotted them on the lines to see the distribution. There is not much clarity in the charts for the kind of theory that we are trying to prove here, i.e. that there are more late goals in 2014. The charts look more or less the same for all three world cups. There are lots of dots all over the place and the density of the dots does not vary much between 2006, 2010 and 2014. This proved a little inconclusive. One way to look at it, from an analysis point of view, is to simply give up and say that whatever the commentator said was wrong. But that's not what smart Analysts should do. We should probably spend a little more time investigating the matter further. Maybe we are choosing the wrong type of data to analyze this.

So, I tried to think like the commentator. What is a person who makes a statement like "this has been a late goal world cup" really thinking in their mind? It seemed to me that we feel the slowness of the game if we have to wait a long time for the very first goal of the match. This made me think that maybe I should just consider the first goal of each match. If the first goal is scored early on and if there is a follow up goal later in the game, it doesn't matter because there is already excitement in the match as some team took the lead, dominated the game or gave away the goal by making an error. The slowness is not felt because there is some excitement already in the match.

I defined the context for the analysis by adding one more boundary condition, i.e. "include only the first goal in any match." We won't consider any goal after the first goal for the analysis.

When I did this and made the distribution, it gave me a better idea. I could clearly see that there were more first goals in the 2006 and 2010 world cups as compared with 2014. But this is obviously because in 2006 and 2010 there are 64 matches and so there was a maximum of 64 first goals. In 2014, as of the time when I analyzed the data, we'd had only 36 matches. Again, we only had partial data. But I could see that the density of the dots was towards the first half of the match between 0 to 45 minutes. Again, I found this to be a little inconclusive.

So, I thought that maybe I should calculate the average waiting time for the first goal. This came out to be 33 minutes for 2014, i.e. in 2014 we had to wait for an average of 33 minutes to see the very first goal. In comparison, this wait was only 30 minutes in 2006. 2006 was 10% faster as compared to 2014. The waiting time in the world cup that took place in South Africa in 2010 was 36 minutes. In relation we could say that maybe 2006 was the fastest world cup in terms of the goals and 2010 was the slowest (among these three). And, 2014 is neither the fastest nor the slowest, it's in between. Again, the statement made by the commentator was not proved but we discovered something else.

Keep in mind that calculating the average is not the best way to analyze the data. As described in an earlier podcast 'averages are mean'! So don't take averages on their face value. Instead you should investigate a little more. But for this analysis, since we already had the distributions, the average gave me an extra data point to chew on.

Finally I thought that maybe the whole approach of distributions was wrong and I decided to plot the cumulative percentage of goals scored by time. By this I mean that instead of seeing the individual goals as dots, we'd see the cumulative goals as percentages. In 2014, 100% of the goals would have been scored by the 90 minute mark because we're excluding the goals after the 90 minute mark. Hence, by the 90th minute all 100% of the goals are scored in all the editions. But what does the cumulative distribution look like?

This gives us three lines starting from 0% and going to 100% - three curves that go from 0% to 100% at their own pace. If a particular world cup is slow, then the line for that corresponding world cup would be at the bottom and it'll only climb to the top towards the end. And that is what happened for 2014. When I plotted the cumulative percentage goals, I could easily see that 2014 was the slowest world cup of all because it was lagging behind the other world cups in terms of cumulative percentage of goals. Again, we cannot compare absolute numbers because the numbers of goals are different and so we have to compare percentages in order to stay at the same base-point.

Here is an interesting nugget. As an example, in 2014, 50% of all the goals scored came within the first 60 minutes. Out of the 90 minute football match, 50% of the goals came in the first one hour of the match itself. However, in 2006, the same 50% of the goals came in the first 45 minutes. A lag of 15 minutes could be seen between 2006 and 2014. 2006 is not exactly 45 minutes, but it's almost there. This means that there is a waiting time of 15 odd minutes to get the same perception of goals in 2014. That kind of creates the impression that this has been a slow goal world cup. I want to remind you that all of this analysis is based on the data as of 23rd June 2014 and I am recording this podcast a week later. There are probably another 40 more goals added between then and now and if we include that data we might see different conclusions. Nevertheless, you would see some story or another.

This is how various attempts are made to figure out whether the statement made by the commentator is right or wrong or what kind of chart would prove or dis-prove the statement.

Let's talk a little bit about what the important lessons for Analysts are when we do any kind of analysis. The very first thing that strikes me is that, as an Analyst, we should never take any generalization or statement at its face value. When somebody makes a sweeping statement like - "young adults between 14 to 21 years are our largest customer base" - if you hear your Marketing Chief making a statement like this without any solid data, charts or something backing up the statement, you should always question it. That's what we should do as Analysts.

As Analysts it's our job to investigate and discover things. If we just take things at face value and proceed to an analysis we might make a false conclusion. It's almost like making a building on a poor foundation. For any analysis the basic facts that you are considering should always stem from the data. If I have to say it in the words of my Management School professor, the very first thing that he said to us when we joined B-school was "Assumptions and Presumptions are the biggest enemies of Managers." The same goes for any Analyst. If you're making a lot of assumptions and if you're taking things for

granted by presuming them, then you'll make a lot of mistakes and eventually your analysis will be poor. As an Analyst, don't assume much. Instead, you should always try to investigate and stay as close to the raw data as possible.

The second thing is that, as an Analyst, you should try various approaches. We should not stop after making one chart and conclude that whatever statement was made is wrong because this chart proves so. We should probably revisit the chart, revise our context or criteria, or define the boundaries much more clearly in order to prove or dis-prove a particular statement. This is very important.

The third thing that is very important is that, as an Analyst, you should never marry a particular idea or theme very tightly. You should be able to divorce the idea if the data says so. This is like holding an opinion. You should not hold an opinion about your data. Instead, you should go and discover the opinion that your data is conveying and then stick to it. If you start the analysis by making an assumption then you will never reach good conclusions. But if you ask the data, "Hey, what are you saying?", then you might be able to extract better stories. So that's why you should not marry a particular idea. You should be a little more flexible and discover what your data is trying to convey.

The fourth and final lesson for us Analysts is to enjoy a little football! This is not a real lesson but what it really means is that we can always get some inspiration, fresh ideas and maybe some scope for furthering our ability to question, inquire and discover in another field. If our business data is not giving as much as we want, we can always go and find something outside. That's what sports do for me. Whenever I watch a sport, I'm always keen to understand the performance numbers and analyze them to uncover some interesting patterns.

These are the four lessons. I want to quickly recap them.

- The very first lesson is - don't take any generalizations or statements for granted. You should always start with a blank state and make assumptions only if the data is convincing.
- You should always try various approaches before concluding that a particular statement is wrong or right. Only trying one approach and concluding is like giving up midway.
- The third lesson is - don't get married to a particular idea. Instead, use your data to discover the idea that is important and the idea that the data is conveying.
- The fourth one is to love football, but that's a little silly!

We have finished our analysis of the statement - "is this a FIFA world cup of late goals" and more or less stayed inconclusive on the statement. We can technically say that, after looking at the cumulative data, the 2014 world cup is indeed slow but that's only as of 23rd June. If I plot the same chart on 1st July, I might discover different behavior. If I plot the same chart again after 13th July, when the world cup concludes, we might see something different. So we have to wait until the end of the world cup to make any conclusive statement.

Now that we've finished the analysis, let me remind you to head on to http://chandoo.org/session13/ to access the link on the entire analysis of the FIFA world cup goals along with charts as well as other links and interesting material on this podcast.

Let's talk about the bonus tip now. This is the 13th podcast session and some people consider 13 to be an unlucky number. I don't particularly consider 13 or any other number to be lucky or unlucky; I believe

that all numbers are equal. But, the number 13 does have a lot of special connotations in our mind. For example, many airlines don't have a row numbered 13, you see row 14 after row 12! Likewise, quite a few buildings don't have a 13th floor. So the number 13 does have connotations pertaining to scariness, bad omens and evil thoughts. Especially in terms of dates, we hear about Friday the 13th. When a date happens to be Friday and also the 13th, many people believe that it is going to be an evil or scary day, and bad things will happen that day. If I am not wrong, I think there are a couple of movies with that theme as well.

A while ago when it was Friday the 13th, I wrote a blog post entitled "How to find the next Friday the 13th using Excel formulas". Today I just wanted to share a quick tip about this. I'm not going to talk about the entire formula about how to find the next Friday the 13th from the current date. Instead, I'll link to it in the podcast session notes. But, I just want to share a quick tip on how to find out if a certain date is Friday the 13th. Friday the 13th is an interesting way to see it.

A date will be Friday the 13th if two conditions are met - the day of the date is 13 and the weekday is Friday. Assuming date represents the date, the same can be represented in Excel as:
=AND(day of date=13, weekday of date=6).

6 here stands for Friday. The weekday functions, by default, will give you 1 through 7 - 1 for Sunday and 7 for Saturday. Thus 6 is Friday. So the formula is:
=AND(day of date=13, weekday of date=6).

This is how you would check a date to see if it happens to be Friday the 13th. I hope you find this little tip interesting. Many people dread working with DATE formulas especially when they start using Excel. Once you understand how dates work, you'll find that they're really simple to work with and there are many powerful functions in Excel that can help you.

If you want a little more challenge, head on to http://chandoo.org/session13/ where I will link to the Friday the 13th article that I wrote a while ago. In that you will find a lot of interesting formulas and a challenge to calculate the gap between the next two Friday the 13th's.

I'll leave you with that. I hope you enjoyed this episode of chandoo.org's podcast. Please visit http://chandoo.org/session13/ to access all the show notes, resources and links mentioned in this podcast episode.

Thank you so much for listening to this episode. I hope you have enjoyed the little rant on football and the interesting analysis of late goals in the 2014 FIFA world cup. Thank you so much and have a wonderful day. Bye.