Transcript for Session 052

Listen to the podcast session, see resources & links:

http://chandoo.org/session52/

Transcript:

Hi and welcome to http://chandoo.org podcast. This is session number 52. http://chandoo.org podcast is dedicated to making you awesome in data analysis, charting, dashboards and VBA using Microsoft Excel.

Thank you so much for tuning in to yet another episode of our podcast. I am so glad to see you here.

Before we make any progress, I must first explain why there has been such a big gap between episodes 51 and 52. If you remember, earlier on in the first or second week of January, we had our 51st episode of http://chandoo.org podcast, and we talked about frequently asked questions on the VLOOKUP formula. And, now, almost a month later, I am talking to you again. This is quite unusual as I try to have at least 2 or 3 episodes per month but quite a few things happened in between. We took our annual Christmas and New Year vacation after the New Year. I know it sounds a bit strange but, in India, the school holidays are usually given between the first and third weeks of January.

This year, we wanted to take the kids to some places in the Northern side of the country, and we took a pretty long trip, and were traveling on trains and back-packing for 2 weeks. It was quite fun but, at the same time, I couldn't set aside any time to work on Excel or recording podcasts or videos. So, I apologize in case you were waiting for the podcasts and I disappointed you. You might be thinking that that was the third week of January, and why I didn't put anything out in the next two weeks. That's because we had so much of back-log work once we got back, and we were pretty tired from the vacation, and were just lazing for the next two weeks. So, I couldn't really achieve much in the next two weeks, and I am really happy to present this podcast to you today. In fact when I started recording this podcast, it felt as if it was my very first podcast, and I was a bit nervous. I am not usually nervous when I am recording podcasts but, for this one, I felt a bit of nervousness because I wasn't sure how I used to do it. I hope you understand that. Anyway, that was the reason why I was away but I am hoping that, going forward in 2016, I will get to share so much more awesome stuff, and tips and ideas and personal experiences in the world of Excel with you through the podcasts. So, stay tuned, and let's jump in to this episode.

In this episode, we are going to talk about one of the newest books that I am reading now. It is called **M is for Data Monkey**. This is a book on **Power Query** if you haven't already guessed. It is written by **Ken Puls and Miguel Escobar**. Both of these guys are really extraordinary and awesome, and they are my friends. You might be thinking how come all the books that I review are from my friends. Well, it's not like I am partial to my friends or anything; it just so happens that the people who write Excel books are all a very close knit community, and we keep up with each other, and we interact through emails, collaborate on blog posts and chat on skype often. It so happens that I've known Ken and Miguel for quite some time. I met Ken, and chatted with Miguel and interviewed him for podcast episode 40 in which we talked about an introduction to Power Query. Both of these guys are really good. Miguel was talking about the book in the earlier podcast if you remember. Although the book has been out for quite some time, I haven't had a chance to purchase it in India. As soon as I got a chance, I ordered it through Amazon, and they shipped it from US to India, and I got the book about a week back, and I have been reading that.

In this podcast, I want to share some insights and ideas about the book, and probably encourage you to get a copy of it. Of course, you don't have to get the book. You can probably find a good bit of information about Power Query online, and tinker with it. I have been playing with Power Query for close to 2.5 years now. I started using it when it was called Data Explorer, and then renamed to Power Query. Nowadays, in Excel 2016, it is called 'Get & Transform Data'. I don't know where to begin my rant about that. Microsoft had a very good name in Power Query. Power Query really tells you what this is all about but then in 2016 they went ahead and started calling it **Get & Transform Data** which, according to their logic, is more approachable. But, to me, that kind of dilutes what Power Query is all about. Anyhow, no matter what it is called, a rose by any other name is still a rose, and Power Query by any other name is still a pretty awesome technology. So, Power Query is something that you can use to deal with data problems.

Let me now give you a very brief introduction of Power Query. In case you are looking for a more elaborate introduction and idea of what Power Query can achieve, I recommend that you tune in to the 40th episode. You can go to http://chandoo.org/session40/ and get that episode. That's where I interview Miguel Escobar, the author of this book as well and we talk about Power Query. So, tune in to that in case you want to learn more about Power Query.

Let me give you a very brief elevator pitch about Power Query before we jump in to the book. Essentially, as I said, Power Query is for dealing with data problems. If you dilute it a bit more, Power Query helps you when you've got dirty data. Usually, when you want to analyze data in business scenarios, the challenge you face is manifold. Your data in not in one place; it is usually scattered across multiple sources. Within the company itself, your data might be in CRM systems, SAP, Enterprise Resource Planning Systems, spreadsheets, daily dumps, text files, SQL server, MySQL, Oracle databases, and in some warehouses. The very first challenge that an Analyst must face is that the data is scattered and not in one place. Even if you could collect all this data into one place - let's say that you were able to

build a massive spreadsheet by bringing together all this data - the other day I got an email from one gentleman who said that he had the data spread across multiple places and somehow he ventured out and collected the data - but since it was more than a million rows, he couldn't collect it in an Excel sheet since spreadsheets max out at a million rows. So, he set up multiple tabs, i.e. one for each month, and collected the million rows for each month into one tab. Now, he has this workbook with 12 tabs with 12 months of data, and to do any kind of analysis, he must first combine all this 12 months of data, and do it, and that's dragging him down, and making everything slow. So, this is the first problem - your data is spread across and even if you collect all the data, you wouldn't be able to hold it in Excel because Excel has a maximum row limit of 1 million. These are the tip of the iceberg problems.

Then come the next layer of problems - you have duplicate data (the same data is maintained in two different systems). For example, the customer data is part of the CRM but, because certain data is required to maintain invoices, some of the customer data is also maintained in a finance system. How do you reconcile this and make sure that you are only dealing with one copy of the truth and one version of the truth, and not multiple versions of the truth? That's another problem. Likewise, you might have data that is not in the shape that works best for you. It's basically what we call unfit data. The data is there but it is in a weird shape. For example, you would expect customer names in one column down the rows but, for some weird reason, somebody was maintaining a spreadsheet, and the customer names are in columns and so it is basically in a transposed version of data, and how do you deal with it. Or, it is in a pivot table report kind of format and how do you deal with it. All these problems kind of bog us down and slow us down. When you want to analyze data, you have this grandiose vision - you want to do this kind of analysis and that kind of analysis, you want to make this type of a dashboard or a that kind of a visualization but, in order to even come up with a simple bar chart or a line chart, you must first battle with this data, and fight with it. How do you do all of this? The typical fighting pattern until now before Power Query is to set up massive spreadsheets, collect all the data, and use formulas to restructure it, and use VBA code or manual techniques. All of this takes up quite a bit of time. You start your project of data analysis and your first 35-40% of the time is spent just collecting and cleaning up the data. Only then do you get to spend the time on analysis. In some projects and real life situations, it is not unheard of where people spend close to 2-3 years just getting a sense of the data. I have been through these kinds of projects. For example, in a data warehouse implementation, the entire first 6 months is dedicated to just understanding data and maintaining it, or making sense of it. This is where Power Query helps you as an Analyst. It is a tool on your side wherein using Power Query you can easily handle all these data problems, and tackle them in a very simple click and drag kind of manner. You are not even writing any formulas; you are simply clicking and setting up a process for Excel to deal with data. This is not a one-time thing. When you set up a process, the process becomes repeatable. Any time that your original data changes, the same process gets applied, and you will get fresh, clean data into your data model or Excel spreadsheet so that you could do the analysis. This is what Power Query is supposed to do. It is a feature and an extra layer on top of Excel which talks to the original data, cleans it up, combines it, consolidates it, transforms it, purges it and then provides you the best version of the data that you want so that you can do your analysis. The process is a one-time process. You set it up, and any time that your source data changes, and any of the things change, Power Query kicks in again and cleans up the data, and so you don't have to worry about changes in the original data any more. That's

what Power Query is. I highly encourage that you tune in to the 40th episode of our podcast to get more insights about Power Query.

Now, let's talk about the book for a few minutes here. As I said, I have been using Power Query on and off for about 2.5 years now, and Power Query is a really slick and powerful piece of technology. So, naturally as an author and user of Excel, and as a trainer and enthusiast, I want to know more about it. So, I download it, play with it and pretty much self-learn the technology on and off. But, I find that Power Query is somewhat difficult to navigate. On the other hand, I found Power Pivot to be more approachable and understandable because it's essentially an extension of Excel formulas and pivot tables. So, there are some mental models that anyone can apply to self-learn at least the basics of Power Pivot. But, when it comes to Power Query, some of the things make sense because they seem like Excel formulas or pivot table actions but some of them are pretty far stretched out and confusing. So, it is a mixture of the ease of use of Excel and the power of SQL (Structured Query Language) that you use to deal with data. I am not sure how many of you are familiar with SQL so in case you are not an SQL user or have never heard about it, you can pretty much ignore the next one minute of the podcast. In SQL, we deal with data and we type a query and get results. So, we are not even doing any clicks or anything. We are just typing and getting the results and setting up queries and combining tables and all that. It's more of a programming kind of thing. However, in Power Query, you don't really write any code. You can write code but it is not something that you would do right out of the box. You would do it pretty much after you kind of establish yourself as a competent user of Power Query. Until that point, you're just clicking and dragging, and setting options. That's all that you're doing. So, it is giving you a powerful front-end to the kind of querying concept, but, at the same time, because of the front end, some of the things are jumbled and not everything is very clear or apparent. So, what I am trying to say is that while Power Query can be self-learned, it is a very difficult jungle to navigate. So, after trying it for a couple of years, I started reaching out to Ken when I heard that they launched a Power Query training program, and that's why I interviewed Miguel so that I could talk to him and get more of a sense of this technology. If you remember, in the 40th episode, I even recommend that you consider their live classes if you want to learn Power Query very quickly. But, again, I was waiting for their book, and, as soon the book was out, I wanted to get my hands on it. So, I am really glad that the book is here and I am reading it. This is where the book really helps. Power Query is a tricky piece of technology and unless you have a good background in technology and especially in databases, SQL and Excel, which is a weird combination, then you will find it difficult to understand what Power Query is doing. But, if you don't have that kind of a background, or even if you have that kind of background but you are a bit rusty in some of these areas, learning Power Query on your own is hard.

Since Power Query is a new technology, Microsoft is constantly working on it and improving it. These days, every week, I get an email from Microsoft or sometimes I see on my Twitter and Facebook streams that they are adding new features to Power BI or Power Query every week because it is on shaky ground where the product is always changing, and the documentation is not up to the mark. There is a function reference and other things on their website but I find that learning through that is hard. Likewise, let's say that you want to learn something about Power Query, and you go in to Google and type some

problem, chances are that you will find some blogs. You will find Ken's blog, you will find Chris BI's blog, and a few other blogs, but getting the solution to a particular problem or even understanding how you got yourself into that situation is hard because the Power Query community eco-system hasn't fully evolved yet. This is why I find the book to be an excellent resource for those of you who are starting to learn Power Query. There are a few other books as well. If you want you can order multiple books, i.e. this one and the other books as well. For me, I consider myself to be an intermediate user of Power Query, and I really want to get in to advanced stuff with Power Query. That's where I am finding this book to be quite handy and useful.

Let me briefly tell you about the topics covered. Although the book says M is for Data Monkey, this book is not a book about M language alone. I imagine that Ken and Miguel pick up that title because it is a bit cheeky and raises your curiosity but the book is more comprehensive than just being about M language. In fact I was expecting that when I start reading the book, they will start about M language in the very first chapter. But, that's not what the book is about. The book kind of gradually builds up to M language and then gets in to it. About two-thirds through the book is what I consider a very deep and thorough introduction of Power Query. Don't be confused by the word introduction here. It is quite an advanced, intensive course in Power Query. That's what you get in the first two-thirds of the book. The book starts off with very basic stuff like Introduction to Power Query and terminology. Then, the book talks about transformations which is where you take the data and transform it, i.e. change the shape of the data. This could be removing some columns, adding multiple tables, merging columns, combining tables, changing the shape of the data from vertical to horizontal and vice versa, pivoting and unpivoting, and those kinds of techniques. Then, we talk about the transformations where lots of things can be done. Around the same time, we also talk a little bit about the Power Query formulas. Some of the transformations involve changing individual characteristics of a column or data. For example, changing the text case or the precision of decimal values and things like that. Then, it talks a little bit more about using the Power Query formulas. This is where the M-language bit comes into the picture. M-language has two portions; one is the formulas where you are writing formulas, and the other is the scripting part where you can actually write the entire code instead of using the click and drag user interface of Power Query.

This is where I must really stop talking about the book because I haven't even read the chapters on M language. I am still reading the book. But, as soon as I read the first few chapters, I knew that this is something that I want to share with the podcast audience, and help you understand what this book is about, and maybe make a purchase decision in case you want to go for it. I haven't read the M language portion as yet but this is something that I am really eager to do. I am pacing myself, I am learning a chapter or two and practising simultaneously so that I don't miss out on the techniques and the concepts are solidified in my mind.

In a nutshell, the book basically has two-thirds of the portion dedicated to transformations, formulas, and deep introduction to Power Query, and the other one-third dedicated to the Power Query

programming language which is called M language. That's what this book is about as the name suggests. Although the name is kind of misleading here and suggests that it is a book about M language, you don't get M language stuff right off the bat. You will get it two-thirds of the way through in to the book.

Let me now share one awesome example just so that you know the kind of things that you can achieve with Power Query. We talked a little bit about what you can do with Power Query as a beginner user in the 40th episode of our podcast. Let me remind you that you can go to http://chandoo.org/session40/ and listen to that episode. In this one I am going to share something that I would consider pretty much impossible to do with Power Query if I hadn't read that chapter. Let's say that you are looking at a multi-level pivot table, something like region, product and month level classification of sales, customers and quantity of product. Can you imagine that kind of a pivot? It would probably hold regions in rows, and have two columns at the top - one with the product and the other with month. So, it is basically a multi-level pivot table and inside the pivot table, you not only have the sales value but you also have the number of customers and total quantity sold to them. So, you have such massive data. And, imagine that instead of getting raw data, somebody gave you this, and you are supposed to do some further analysis on it. Being an Excel user, you naturally think that this pivot format is not conducive for doing any kind of further analysis and that you should first unpivot the data and then start working with it. The challenge is that you cannot unpivot this straight away because of the multi-level structuring of the data. It looks like a pivot table but it is not really a two-level pivot table. It is a multi-level pivot table and the unpivot feature of Power Query doesn't work straight away. So, if someone gave such a table to me and asked me how I would unpivot that data, I would have simply told them that I would use VBA to do this because I would imagine that doing such a thing with Power Query is very tricky and not doable at all unless I knew a lot more about Power Query. But, I was reading that book, and chapter 15 talks about unpivoting complex data tables. I got curious and I flipped to that chapter and started reading it, and this was one of the examples there. I felt that it sounded really tricky and wondered how they do it. I immediately went to their website and downloaded the example file and started practicing step-by-step as I was reading the book. Can you believe that you can unpivot this fairly easily with Power Query? I was imagining that kind of thing was not possible. This is the kind of confidence and skill set that you will acquire if you read the book. You will learn how to deal with fairly complex data situations and not just simple situations, and fairly tricky type of data and how to use Power Query in such situations and get the best out of it. I highly recommend getting a copy of this book especially if you are an enthusiast into the Power BI world, or you are somebody who is using Power Pivot and Power Query off and on, and you want to polish your skills. Get this book and it will definitely help you.

Here is a small disclaimer. In case you are purchasing this book through the links mentioned on the http://chandoo.org podcast page, I do receive a small bit of commission through the Amazon Royalty Scheme but that's not why I am recommending this book. It's because I am learning a lot from this book and I feel you deserve to know about this book and benefit from it as well. That's a little bit about this book. I hope you enjoyed this review. I will get in touch with you again in the next podcast where we will talk about something interesting and useful for you as well.

If you want to know more about this book and get all the details and resources mentioned in this podcast, please visit http://chandoo.org/session52/ where you can find all the details. You will also find a link to the book and the http://powerquery.training website where you can find more information. All the best and talk to you again in the next episode. Thank you. Bye.