Video 30

In this section, we will cover an introduction to data visualisation.

In data analysis, many questions can be answered by using or querying databases and generating numbers. But some of those questions are better answered using visualisations which is nothing but a visual depiction of the data that you have generated or the numbers that you have generated. For example, supposing we were trying to answer this question, *what is growth of my sales quarter on quarter?* Option A could be to say the growth of my sales is 25% compounded annually. Option B could be to show the same information but in a visual depiction. For example, we have a sample depiction here of sales of I-pads, quarter on quarter. You can see that this is also talking about growth of sales but this is giving people a lot better information. For example, that is some quarters are growing faster than the others. Overall this is my growth rate, this is where I ended up in the last quarter.

Therefore, a very efficient and useful part of data analysis is data visualisation. Data visualisation techniques are used extensively in both data analysis and data science teams to both identify problems or issues or find interesting patterns and to convey information that will help with decision making. Data visualisation tools and techniques and interest in data visualisation is growing rapidly.

According to IDC, visual data discovery tools will be growing 2.5 times faster than rest of the BI market. And according to SAP, people report that on average, they spend 9 hours longer to find the same patterns or trends and correlations in the data if they are not using data visualisation tools. Many companies want to provide the power of data driven decision making to all their employees and data visualisation tools allow people who do not have specialised skills with advanced data analysis techniques to also generate insights from data relatively easily. Of course, to generate good insights implies that you have useful and appropriate data visualisation. It is very easy to draw the wrong conclusion when the data is not represented correctly. Therefore, it is a good idea to think about a framework for data visualisation.

Step 1: frame the problem statement correctly.
Step 2: extract or collect or combine your data to get the relevant data set for the analysis and the visualisation.
Step 3: process and analyse the data to get the answer to the question in mind in a numeric or tabular format.
4: choose the more appropriate visualisation for the question that you are trying to answer and
5: present that visualisation in a way where it is easy for people to generate the insight or see the insight.

Most times when people try to use data visualisation techniques, they start by building the visualisation directly. However, it is much better to instead generate the answers using whatever analysis technique that you want to use and get the answer in numeric or tabular format and then decide what that visualisation should be. Let's understand this framework and see how it is applied using a simple problem.

Let's say the business problem is, I want to understand how sales of MP3 players are doing across my stores. What would be a good way to answer this question and how would you use the visualisation to answer this question. The first thing is to start by defining the problem statement correctly. Now some of us may feel that the statement seems very straightforward. I want to understand how sales of MP3 players are doing across my stores? However, we want to add specifics to this problem statement, so we know exactly what it is that we are trying to analyse or answer. Rather than saying understand how sales of MP3 players are doing across stores, I will say, 'What are the monthly total unit sales figures for the category "MP3 players" across all stores for the last 12 months?' what is the difference in these two problem statements? The second problem statement adds a lot of specifics. It says, I want unit sales not revenue for category MP3 players across all stores in the last 12 months, so I am adding a lot of specifics, 12 months, unit sales, category MP3 etc. so that my data extraction and the analysis process is a lot more specific.

Once I have the problem statement, I then need to figure out where is the data that I am going to use for this analysis. Sometimes you may have to extract the data or sometimes you may already have the data, but it is in multiple sources and you have to combine all the data. In an example like this where I am looking at total unit sales, the data typically will be under transactional database. Now we have, let's say that we have queried the database and we have the following two pieces of information. We have a transactions file that contains transaction ID, payment method, timestamp, product code, items number and items amount and we have a separate product file which has the listing of the product category name against the product code and the price for that product. What we need to do is map the product code in the transaction files in order to get the product category name because we are interested in MP3 players. In other words, we have to combine these 2 files because I want it to be able to add up this unit sales for all the MP3 players. In this example, in the transaction file, items number is the number of units sold in each transaction. So I need to sum up the items number for product code which is equivalent to MP3 players. Let's use excel for this example, we will open up both the files in excel and then we will simply use the 'V look up' function to add the product category name to the transactions file using the product code.

This is a transactions file, you can see there is a transaction ID, a card ID, payment method, timestamp, product code, items number, items amount etc. This is a very big transactions file, we have lots and lots of data in fact, 350 thousand observations. You can see that there are a lot of product codes. We are interested in product code corresponding to MP3 players or we are interested in the category MP3 players, therefore, I need to add the product category names here that I will get from the product file.

This is the product category file, you can see there are product codes, category and the price for each of these product codes. You can see there are multiple categories here, game consoles, entertainment, hardware, Hi-fi and MP3 players, and then you can see there are many product codes for MP3 players. What we want to do is add this name of the product category to my transactions file and we can do that using 'Vlook up'. So you can see we have done that here, using 'V look up' . I am looking up this product code in the list in the product file and I am coming up with the name from the product code file. So if I do this, I get the actual product category name against each product code. Now I also want monthly unit sales, therefore, I need to derive the month from a date and the date is in the timestamp variable. So, here I use an excel function called month of this in order to generate the month of the transaction, so this is January, this is April, this is July and so on. Now that I have this data, remember what I want to do is figure out what are monthly unit sales of category MP3 players. I can do that using a pivot table. This pivot table is easy to generate. All we have to do is say insert pivot table, say ok and then, we simply choose the pivot… the data that we want to see in the pivot. So what we want to see is the sum of item number by month and we want to see this for in fact not product code, for product equal to MP3 players, so if we take a look at this pivot now, this instead of product all, we will choose product MP3 players and now, what I have is by month the sum of item numbers. Now I can generate a visualisation simply by going.... so this is my pivot table that shows the sum of item numbers by month. Now while this answers the question of what is my total category sale for MP3 players by month for the last 12 months. This information is better depicted in a data visualisation. While we have the data, it is important to choose the appropriate visualisation. Since we want to look at performance of sales by month, the simplest visualisation would be to create a line chart that will show the trend of sales by month. So we could simply take this data and say insert a chart and choose line chart and this gives me this visualisation which is simply sales by month, so a visualisation like this. However, while we have generated the visualisation, we still need to format and label the visualisation clearly because when people look at that visualisation, it should immediately tell them what they are looking at. So for example, we should update the title to reflect what is the data that we are capturing, we should add labels to the axis, we should change the scale of the axis if required, we should format numbers appropriately as required and choose appropriate colors. Remember, we do all of this to make sure that when people look at a visualisation, it is very very clear what it is that they are looking at.

How do we do that? We can simply update this chart by, for example changing the title so instead of saying count of items number, we can say, 'Total unit sales by Month category MP3 players, time period January to December 2001'. So this is a very self explanatory title. Here, in the label, instead of saying count of item numbers, we may want to change the label to say, month. So instead of count of item number, we call this Total unit sales. Now while this is total unit sales, we want to make sure that we also have a label for the horizontal axis, so we can go here to design and say we want to add axis titles. Primary horizontal is going to be month and primary vertical can be total units, total unit sold. We could also change this scale if we wanted, for example we don't really need 5000 and 10000 because in every month, at the minimum we have sold about 15000 and maximum 20000, so we could potentially if we want to see the difference is more, we could change the axis minimum to 10000, instead of 5000. So instead of using a minimum of 0, I am going to use a minimum of 10000. So you can see that there is a difference in the unit sales by month. We can also format these numbers to make it more readable, so instead of saying 20000 without the comma separator, I want to make this a number with 0 decimal places but I want to use a 1000 separator. Again, all of these essentially help us read the visualisation and the insight from the visualisation very very clearly.

So this is finally the answer to that question, ‘what are the sales of my MP3 players across all my stores’. So remember, data visualisation techniques are an effective way of data analysis and presentation. However, it is important to make sure that we choose the right set of visualisations given the problem statement and the available data. So how do we actually choose the right visualisations? We will look at, what are the appropriate visualisation types for different kinds of data and how we choose the right visualisation given the type of data that we are working with.