Week 3: Data Preliminaries for Analytics

Video 1: Data Preliminaries for Analytics: Overview

Hello, everybody. Welcome to data preliminaries for analytics. Namaste, Salaam and Sat Sri Akal. Let me take you to the course map first. Now, this is basically a depiction of the different components of the course in relation to each other.

As you can see, there are three main components, that's data pre-processing, followed by data analytics using various tools, including Machine Learning, and there is finally extraction insight from this process.

Now these three components are seemingly separate, but in different sessions you would have parts of them come through. Today's session is specifically focused on data pre-processing. This will also come up in bits and pieces elsewhere, but this is specific to data pre-processing and data preliminaries in general.

Video 2: Data Preliminaries: Motivating Example

All right. So, let me start with a motivating example to motivate data preliminaries, more generally, the value of data in a business setting. So, this takes me all the way back to March 2009. This was about the time that Uber technologies was founded. We all know what Uber is, right? By 2014, 5 years later, Uber's valuation was a mammoth $40 billion in five years. My question to you very generally would be why?

Why is Uber valued so highly? Of course, before I go there, the question would arise, what do we mean by valuation anyway? What does valuation mean? In technical terms, it is the money required to buy the company. But in business terms, effectively, what does valuation mean? Valuation is the Net Present Value.

The value in today's money. Of all the profits that that firm will earn over its lifetime, in principle and theory, that is what is being value effectively. That's what the market value, so to say. As of now, investors believe Uber is worth over $80 billion. That is the Net Present Value of the lifetime profits that Uber will earn. That's what investors believe. And time will prove them right or wrong, but there it is. With such wonderful valuation numbers, Uber sits in pretty interesting companies.

So, what accounts for these numbers? Why is Uber valued so much? Think about it. Right. Think about it. You think about it, well, Uber is a ridesharing business. There are other cab companies, so what is so special about Uber? Just as an example. So, what makes them so special? Well, is there some competitive advantage? Is there some strategic asset? Is there some enabling platform that could somehow explain at least some portion of this valuation number, effectively?

Well, let's have a look that. Now, Uber is asset-light. Okay. They don't even own the cabs. The Uber cabs that run are not owned by the company, so these are in some sense. However, they do have one asset. Okay.

One asset that Uber does own that lies at the heart of its revenue and hence, its profit projections. Uber owns data. All rights to every single bit of data from every passenger, every driver, every ride, every route on its network, Uber owns all rights to them. Period. It all goes

there. And that lies at the heart of how Uber is going to make money in the future. The possibilities for expansion, the possibilities for mining and extracting value.

Just how much data are we talking about? I mean, one could be incredulous, but still 80 billion is a big number, right? I mean, this is data about rides and passengers and really is it worth to that much? Have a look at this. Uber took 6 years to reach one billion rides. Okay. Around December 2015, starting from 2009.

They took about 6 odd years, right, to reach a billion rides. Guess how much time they took to reach the next billion rides. Six months. Six years to 6 months. We are talking exponential growth here. Yup. What does it mean that you had a billion rides in 6 months, a billion rides in 180 days, 5 and a half million rides per day? 5 and a half million rides per day. In fact, there was this prize that was announced for the 2 billionth ride. A 144 rides tied for that to the fraction of a second. Okay.

We are talking a data host, in some sense. All right. So, how does having data, even a lot of data, connect to analytics and thereafter to market value? That would be the question that I will try to dig into a little more as we proceed with this example. So, why care about data analytics? In some sense, data analytics is basically this, if analytics is the steam engine, data is the coal. Both are needed. There it is.

So, let's talk about the data piece and that is what this module is focused on. Why is data such a big deal? Think of data as a valuable asset, which brings up the question, what do we mean by asset anyway? So, very simply, an asset is any resource that yields returns over time. Assets are not consumed. Okay. So, we invest in assets and the yield returns over time. Data is a valuable asset and a special type of asset. It's not like your other fixed assets and so on. Data is a very special type of asset. Why?

Because it actually might help build sustainable competitive advantage. What is competitive advantage? Anything that gives you a leg up over competition, that's it. The problem with competitive advantage is that it is fleeting. It is transient. It is here today, gone tomorrow. Why? The other guy can see what you are doing and copy it. However, when the asset in concern is data, there is actually a chance to build sustainable competitive advantage because the other guy cannot copy your data. The sustainable part is key. What happens when competitive advantage becomes sustainable?

Let's have a look at that. When competitive advantage becomes sustainable, you get what economists call super normal profits. There is the normal profit curve and there is the super normal profit curve. Or what the finance types would call sustained abnormal returns. When you have that, folks, congratulations, we are in wonderful valuation territory. I mean, these go through the roof. The Net Present Value calculations for anything that is at the super normal profit curve is up there.

Well, let's basically look at another example. Now, you might say, hey, this is idiosyncratic. It's one example. Hold on. Let's have a look at a new versus old economy illustration for the same principle. Consider the stock performance and the stock performance effectively is market valuation of Amazon and Walmart. Okay. Amazon was born in the new economy, right? It was native to the web.

Walmart has been around since the early 60s. Okay. Huge corporation, very large revenues. All of that. Now, what you see is basically for a 5-year period, the differential stock performance of both these companies, by the way, both of them are incomparable markets. Both of them are B2C sellers, effectively, right now, so there it is. Okay. I've normalized it to hundred. So,

the stock is starting at the same point, so to say and then how it moves subsequently over that 5 year period.

There is no doubt about what happened. In February 2012, Walmart had a valuation of $200 billion and Amazon had a valuation of $82 billion. Okay. So, both of them are starting at 100, let's say, and then subsequently we trace what happened. Fast forward, 5 years, February 2017. Walmart's valuation is $210 billion. It is still where it was, 202 to 210 negligible up and down.

Amazon's valuation goes up from 82 billion to $400 billion. 5X. Now, even if you attribute half of that profit to AWS and the Cloud, but still 2.5%, a 150% jump effectively in valuation. And investors believe that lifetime profits of this company are going to go up that much. Which brings me to a favorite quote of mine about the age of data.

If land was the primary raw material of the agricultural age, what does it mean primary raw material of the agricultural age? In the agricultural age, millennia, centuries upon centuries, prior to the Industrial Revolution, the only way to get more agricultural product was to have more land. Period.

There was no other way. Iron and coal were the primary raw material of the Industrial Age. If you wanted more industrial product, you had to had more iron and coal. Period. Well, data is the primary raw material of the information age. You want more information product; you need more data.

Video 3: Data and Measurement Basics

All right. Let's dive into data and measurement basics in continuation of data preliminaries. Let's start with definitional preliminaries. What is Data? Basically, the first definition question that would come up. Now, there are many definitions and this one, for instance, is from Wikipedia. What I'm going to do is take a deeper dive at data. Let's go back to its origins, in some sense.

So, etymologically, the word data originates in philosophy, from the Latin word datum or datum, effectively meaning, given as true, that which is given, in some sense. It refers to that which is known, or assumed to be true with refers to facts. That's basically what it is. Key questions arise about data, data types, data structure, data properties, data size, data quality, processability, compatibility, evolution, etcetera. And each of these are multidimensional, so you can drill down into each of them.

Just as an example, if you take data quality, these are all the different aspects, in complete list by the way, of data quality. Accessibility, consistency, accuracy, timeliness, cost, and so on. Data alone is not enough to be useful, in some sense. Data has to be transformed into certain higher order entities. Data, in context, becomes information. Information with the meaning becomes knowledge. Knowledge from which insight is extracted, becomes wisdom.

So, all of these, the knowledge hierarchy, we call it wisdom, knowledge, information, all of them rest on this base of data. If there is no data, there is no, none of the other things as well. Basically, data underlies them all. And what underlies data? Measurement enables data, which leads to a host of other questions, and that is where I'm headed now. Why Measurement? Without measurement, there is no data. Without data, there is no analysis. Without analysis, there is no modeling, and this is a big one, analysis and modeling. Without modeling, any model is supposed to give you at least one of two things.

Okay. A model is supposed to give you either explanation, or prediction, ideally both. Without modeling, there is no prediction, there is no explanation. Without explanation, there is no insight, and without prediction, there is no optimization. Without optimization, and or insight, there is no management. Effectively, it all boils down, to data and measurement. Let's head there.

The data story in some sense, digging back to its origins, are coming to where we are today, where are we coming from to take that perspective. For millennia, I'm talking centuries, upon centuries, record keeping meant clay tablets, papyrus scrolls, parchments. So, this which you see, is a clear tablet, and this one is a parchment, both of them from ancient Egypt. I mean, how much information, or data, can you fit onto a clay tablet?

There's just so much you can do, but that's about it. Modern paper, hence, was an enormous advance, but what really, you could write a lot on paper, easy to store and carry, and all of that. However, what really set the revolution, the data revolution going, was the Gutenberg printing press, and this was centuries ago, 16th century. The printing press, changed things, revolutionized them, why? In 50 years, the printing press produced more books, than had been produced in all of prior history.

Certainly in the western world, I mean, out here we don't really know, but there it is. In subsequent centuries, you had the telegraph, and then the telephone, and then the radio, and then the T V, and finally, the big one, the computer. Just to tell you how much the computer, influence the data revolution. The year 1996, is a watershed. Why? For the first time in the U.S., in the year 1996, it became cheaper to store data on computers, rather than on paper. It was just in 1996 before that paper was actually cheaper to store data in. You can imagine the fall in the price of storage, digital storage since then.

This was 1996. Four years later, by the year 2000, 25% of all new data that was being produced, was being stored digitally. This is what happens when prices step effectively, when it becomes cheaper to do something digitally, more and more of it, certainly newer activities, would migrate to digital, and so there it was. 25%, only about a quarter four years later. Fast forward, seven years later.

By the year 2007, the proportion of new data that was being stored digitally was 94% in the West, 94%, a revolution happening right before our eyes. Now, forward another 10 years. So, let's look at this 2007 to 2017 timeframe. And this has implications, for where we are going and what we will be doing in Data Analytics. Every month, six billion photos are uploaded to Facebook. This is the amount of data being created. Every year, we are creating more data than had been created in the last century.

The blogosphere cannot include video blogs as well and podcasts, and so on, doubles in content volume every five months. Every five months, exponential growth. 72 hours of video uploaded onto YouTube, every single minute. 400 million tweets, every day on Twitter. The amount of data being generated and it has directly to do with the fall in storage costs, effectively.

Digital storage is now so cheap. In less than $100, you can buy a two terabyte hard drive for storage. Two things stand out, basically, where is all this coming from? One, there is ever more data being generated, year on year. 20 years ago, there were things that we wouldn't even consider data, which are data today. There are things we wouldn't know how to do 20 years ago, which we generate and store, as part of life. I mean, it's possible today, because storage and data have become that easy.

Two, ever more of that data, is native to digital. I mean, there is no paper anywhere in the picture. It directly goes to digital; it stays digital. For a lot of the people who were born in this century, who are digital natives, so to say, whenever go to people. It will remain forever in digital form. So, these two things I would like to stress upon. A lot of this data is what we will analyze for business insight in the coming sessions.

Video 4: Introduction to Data Dichotomies

Let me now introduce you to data dichotomies. There are 4 principal data dichotomies I will talk about. Three of them I will present right away, data types and dichotomies. So, let me start with a simple example. Consider the following data on public buses. So, there is a picture of a bus. You can see the registration number of the bus and the route number written up there.

The table that you see, let's say, is a small part of a large table. Give it a quick read. What do you see? you see? You see that this data now in a tabular form, is structured along rows and columns. And in data sciences, we would variously call the rows of a data table as observations or instances or cases. They all mean the same thing. They mean rows of a structured data table. That's basically what it means.

The variables, the columns are also variously called attributes or features and so on. Data columns, that's what they are. Look at the type of data presented on that table. You have date data. You have time data, which have an order into them. You have route number looks like a number, but really, I cannot add route number 83 and 84. It doesn't make sense. But there are numbers which are sensible as numbers. Ticket revenue, for instance, is sensible as a number.

I can add the revenue of two different routes. There are names, bus stations. You have Nellore and Vijaywada and so on. I mean, their bus stations about their names. You have name data in there. Registration number is this character string. I mean, you can see all of this playing out right there before you. This percentage data, occupancy and so on. So, we have all of that coming through. Now, having seen some basic data types, let's have a look at the 3 basic data dichotomies

So, what is a dichotomy? It's a split. It's like a branching off. It gives you 2 different classes. The first dichotomy is that between structured and unstructured data, so data comes in these two broad classes. structured data, unstructured data. Another basic dichotomy is perceptual versus objective data. So, data can be classified again into perceptual data, objective data. The third basic dichotomy concerns primary data versus secondary data. What are these data dichotomies?

Very briefly, structured versus unstructured structure data has to do with the intrinsic nature of the raw data. Does it require transformation? Does it require processing? Does it require additional effort from our side to make it structured effectively? That's the question we will ask. Two, perceptual versus objective. Whether the data collected is subjective in nature or objective in nature? This has implications for measurement and for analytics. We will see in one of the later modules in mapping of perceptual data.

Third, primary versus secondary, and this concerns the source of the data. Where are we collecting it from? How are we collecting it? This has implications for time, cost and so on, in data collection and in analysis.

Video 5: Data Dichotomies

Alright. So, let's explore these data dichotomies we talked about in a little more detail. When it comes to structured data, this data has pre-existing structure. It is in the form of well- defined variables and it can readily be recorded into data tables. An easy question to ask about whether some data is structured or not, is can you fit it into an excel sheet effectively into rows and columns, if you can, its structured data.

Databases, for instance, have structured data. We use SQL to query databases, the essence SQL is structured query language. This data, structured data, needs only minimal transformation and processing before it is ready to use. So, for instance, the public bus dataset, that small excerpt that i've shown you, this is basically structured data and its fitted into a table and it was pre-existing structure. Unstructured data, on the other hand is data that has no well-defined structure.

There are no ready to use variables. How do I work with unstructured data? The only way to work with unstructured data is to impose structure on data, to force structure out of that data. The choices we make when imposing such structure will deeply impact what analysis we can do and the quality of results we will get. We will come to this. For instance, think of a breakdown or an accident report. This is text text, text is technically unstructured we'll have to impose some structure on it.

There are decisions to be made in that regard, before we can go ahead and analyse that data. Alright, let me come to the second basic data dichotomy, perceptual versus objective data. What is perceptual data? Perceptual data refers to our perceptions. Data collected about human perceptions. It is subjective data. It is data about which two people can reasonably disagree.

But here is an example. Say I gave the captain of the Indian cricket team, Virat Kohli, an eight out of a 10. You give him a seven out of 10. Who is more correct? And the question doesn't make sense. It's perceptual, and technically we can reasonably disagree. Usually, perceptual data and business settings are collected about people's perceptions of quality, of service, of performance, et cetera. And this has huge implications for business outcomes to that extent we are quite interested in perceptual data.

Objective data on the other hand are facts that are independent of subjective perception, it doesn't matter what you think about Virat Kohli or what i think about Virat Kohli. Virat's strike rate is 83.3 that is objective, it's not perceptual data anymore. It comes from calculations. Usually about events measured and where does objective data come from? What is it usually about? So, this is usually about the events measured in physical attributes in terms of time, in terms of space, in terms of distance, mass, very importantly in terms of money.

All of it would be objective data to a large extent. The third and the last basic data dichotomy that we had referred to previously. There is one big one coming afterwards, but this is the third of it: Primary versus secondary data. Data collection for research and analytics, primary versus secondary. What is primary data? Primary data are data collected at source and hence primary in form.

Specifically for the research project at hand. In other words, this is data that would not have existed had we not gone and collected it for our work. This data is not data that gets collected automatically and that we can use, that happens to be lying around and that we can use. The data source could be individuals, groups, organizations, you name it. Surveys, interviews, focus groups, all of these are tools with which primary data can be collected.

Perceptual data typically has to be collected usually primary in form. Secondary data are all data that are not primary, data that were collected previously for some other purpose. They're just lying around we can use them. For instance, sales records. Sales records are there, the accounting department collects them for whatever reason, anyway. We can use that data.

The data exists regardless of whether or not we work on it or analyse it, that the data tends to exist. Such data would be secondary data: Sales records, industry reports, interview transcripts from past research. All of this is there anyway, it pre-exists. A very important source of secondary data is the API. I will come to this later today.

Video 6: Introduction to Data Types

All right. So, having seen the data dichotomies, let us do an exploration of basic data types, effectively data scales. Psychometric scaling is an important and interesting field. It is the intersection of psychology and statistics in some sense. The psychologist, Stanley Smith Stevens, introduced the theory of scales as different levels of measurement. There are 4 types of features ordered from low to high in terms of information content.

So, you have Nominal, you've Ordinal, you have Interval, and you have Ratio. The 4 primary scales. and data types corresponding to them. So, for example, for each type of feature, whether it is nominal or ordinal or interval, there are a specific set of permissible, analytic or statistical operations. Hence, it matters in which scale we have collected our data. Let's have a look at these data types and the permissible operations on them. All right. So, the four types of data and the four corresponding primary scales that are there, let's start with the first. Okay, we are going low to high.

The first is nominal. It comes from the Greek word nomen, which means name. So, nominal is often name. Nominal data are just labels. They are merely names. No further information can be gleaned beyond just that label, I mean, that's about it. For instance, A and B. Yeah, A and B. Ordinal implies order that comes basically from ordinality, which is the result of ordering that is implied. Ordered data would be ordinal. So, if the nominal data are arranged in a particular way, we might get ordinal.

So, this conveys up to preference information, okay. It conveys a direction, for instance, I prefer A to B. A is greater than B. A is more than B. A is better than B and so on. So, basically there is some sort of direction that is implied. There is some ordering that is applied. The third type of data is Interval data. Interval data and the name interval basically implies that between any 2 ratings, there is an uniform interval between them, equal length interval. It conveys a relative magnitude information in addition to preference information.

So, just going back a little. When you come to ordinal data and you say, "Hey, A is better than B." One ordinal data contains a nominal information, anyway, it contains the names A and B. In addition to that, it is saying, there is order, ordering implied. There is this, A ahead of B. When it comes to interval data, in addition to ordinal data and the nominal data, we have magnitude information. In ordinal, A is better than B, but how much better? By an inch or by a mile? Hard to say.

Here, we can actually make that judgement. Conveys relative magnitude information, interval data. Here is an example. I rate A as 7 and B a 4 on a scale of 10. One nominal information is contained, A and B are there. Ordinal information is contained, A is better than B. And see how much magnitude information is contained. Three points out of 10, basically. The last of the 4 primary scales is the ratio scale.

This is the gold standard in scales. This is the highest quality scale there is. It conveys information on an absolute level. Why? Because the absolute zero is objectively defined. It is independent of observer. I paid 11 rupees for A and 12 rupees for B. Ratio data. Why? Because zero rupees is objectively understood, independent of observer. In interval data, the point is not really independent of observer.

Only the intervals are considered uniform, which is the big difference there. Nominal, ordinal, interval, ratio. The 4 primary scales. Here is another quick example. Nominal, the numbers, imagine this athletic, 100 meter dash sprint. The numbers assigned to runners are nominal. The runner is wearing a number on the Jersey and somebody else is wearing the number 8. It doesn't really mean anything. These are just numbers, that's all. Ordinal, the rank order of the winners of the race. So, let's say number 3 came first and number 8 came second and number 7 came third.

There is a rank ordering, ordinal data. We know that first place better than second place better than third place. But we don't know what was the difference. The margin of victory. Interval data, performance rating on a zero to 10 scale. So, some sports, for instance, gymnastics. You have judges giving a rating out of 10. So, these are interval data effectively. And finally ratio data. Time to finish in seconds. So, A finished the race in 13.4 seconds. Jersey number 8 finished it in 14.1. Jersey number 7 in 15.2 seconds. The absolute 0, 0 seconds is understood independent of observer and hence, time to finish in second is considered ratio data.

Does it matter which scale you have? Well, it does. It matters for analysis. Let me give you an example. If you have nominal data, we have 3 types of customers. Type A, type B and type C. Nominal data, we just have A, B and C, for instance. Then all I can do is mode, A is a plurality. We can do some frequencies. A is 40% of the sample and we can do percentages. That's about it. We can't really do anything more with nominal. I mean, they are just labels. There's only so much I can do with them. I can count them, that's all. However, when you have ordinal data, then in addition to everything you can do with the nominal data, you can also do something more.

You can do median. Why? Because there is an ordering implied. The median is that, where half the sample is above the median, the other half is below the median. The above and below comes from the ordering that is implied on ordinate. You move from nominal to ordinal, you move a step. But when you move from ordinal to interval, you move a leap. You actually jumped quite a bit. Something fundamentally changes when you move from ordinal to interval and ratio.

Fundamentally, what is it that has changed? In addition to everything that nominal and ordinal does when you come to interval data, suddenly for the first time, the mean and the variance, arithmetic mean and variance become meaningful. At to this point, the entire battery of parametric statistical tests is now on the table. We can deploy this. At to this stage, things become really interesting indeed. The moment you have arithmetic mean and variance as meaningful, we enter the realm of what is called a metric space. Hence, interval and ratio data are considered metric scales. Right?

I mean, these scales are metric scales. The data you collect with a metric scale becomes metric data. Nominal and ordinal are non-metric data. Interval and ratio are metric data. Common wisdom as far as possible, collect your data using metric scales. Just to give you an example. Suppose you want to measure education level in a population. Now you could go the non-metric way and you could say, "okay, graduate, postgraduate, metric pass and so on. That's ordinal data at best.

Okay, nominal data otherwise. Or you could do what the economists do. You could count the education in number of years of schooling. So, a graduate would be 12 plus 3 or 12 plus 4, 14 or 15 years. 15 or 16 years. Postgrad would be that plus 2 or something. So, basically, you are collecting education level, reducing it down to a ratio measure. It matters for subsequent analysis. The quality of information contained in the data we have collected will matter.

Video 7: Introduction to Data Types

Okay. So, you've seen the activity that has been put up. What I shall do is basically debrief the activity. So, let me go through these questions and provide you my take on what the answers are, slash should be. Mr. Fernando measures favourability of the Airtel brand on a one to five scale. What kind of data are these among the four primary scales? If you take a look, this looks like it is going to be a rating scale interval data.

Now Jai gives Airtel a two, and Aditi gives it a four out of five, one to five scale. Now which of the following statements are true? A, Airtel is twice as much favoured by Aditi as Jai. True or False. Why? This is interval data and what this is asking of us is a ratio. And interval data are not capable of ratio responses,. A is false. B, the difference between Jai's and Aditi's ratings is two points. True or false? True.

Why? This is interval scale, and this is precisely what interval scales do very well. They give us differences across people, across sections. C, Jai is not favourably inclined towards Airtel, Aditi is. Well, what would you say? One to five scale, and it doesn't suggest, given a two, Aditi is given a four, it certainly looks like true. I would normally go with it however, there is a small problem here.

Let me mention it. Very unlikely, but I'll mention this. It is true if the scale is what we call a balance scale, which means that one is very unfavourable, and five is very favourable. In which case, this is true, right? So, Jai is on the unfavourable side. Three is neutral. However, if this is what we call an unbalanced scale, let's say if one is neutral and five is very favourable, then technically both are in favour.

The point of this particular answer option, it is necessary to know scale guidance in primary perceptual data. It will move the needle quite a bit. It is important to know this. D, on a one to nine scale Jai would have given a four, Aditi would have given a six. You can't prorate like this with the interval data. This just not happening. This is basically it requires some sort of ratio to be taken, and that is not possible with interval scales and hence the D is false. All right, let me go to the next debrief of the second question in the activity.

Mr. Fernando measures Airtel usage in minutes per day. What kind of data are these? These are effectively ratio data, minutes per day. Jai reports an average of 20 minutes, Aditi reports an average of 40 minutes. Now, which of the following statements are true? Let's see. A, Airtel is used twice as much as by Aditi as by Jai. True or false? 20 minutes Jai, 40 minutes Aditi, true, twice is true.

You can't take ratios with ratio data. B, the difference between Jai's and Aditi's average usage is 20 minutes. True or false? True. Why? Ratio data have all the properties of interval scales. So, you can't do what interval scales do certainly. C, Aditi uses Airtel more than Jai on any given day. True or False. False. Why? Because it says any given day. The important thing here and the reason I put this option in there, is to emphasize the fact that on the average is implied if it is not specifically mentioned, okay?

On any given day, anything could have happened, Aditi's battery could have run out, or the mobile battery could run out, and she didn't make any calls on that day let's say. So, C is false. We cannot make that inference based on the data available. D, Aditi's Airtel bill is higher than Jai's. I mean, we cannot say this. This depends on your plans and so on, right? And the point of putting that option in, I mean there are things we cannot say, and cannot infer even with ratio data. All right.

Last of the questions from the activity metric versus non metric and this is basically a table that contains. So, this is a salesforce table, sales territory period, sales actual, sales target, territory sales manager, all of those. Which of these variables are metric? And what kind of within metrics? So, is it interval or non? Or a ratio? Which of the variables are non metric? And what kind of it is non metric? Such as, Is it nominal, or is it ordinal, and so on. Let's have a quick guide round through this.

Territory one, two, three, these are sales territories. Okay? Now, one, two, and three look like numbers, but this is not numeric data. These are nominal data, okay? Easy way to look at this. If for that column you cannot take mean, arithmetic mean, for mean is not meaningful, then that is not metric data. Sales territory, nominal in this case. Period quarter1 2017, quarter2 2017, this is ordinal data, there is an ordering, to a time series.

Can I convert this to ratio? Yes. So typically, in UNIX one Jan 1970 is the cut off, and we take all time scales, as a comparison with one Jan 1970, in which case it becomes ratio. This one is ordinal. It is not metric. Sales actual, these are sales numbers. The absolute zero is meaningful. This is metric ratio data. Sales target, metric ratio data. Territory sales manager, these are names of people, nominal non metric data.

Sales for size, metric ratio data. Take a look at customer ratings and these are interval data, ratings are typically interval. Competitors sales, I mean these are ratio. Competitor sales, again ratio. I hope that clarifies, this ability to choose which variable is metric and non metric will be crucial going forward. Sales territory, it look like one, two, three. Bus numbers look like 83, 84, they are numbers, but they are actually names. It is important that we don't try arithmetic operations on nominal data, they are not meaningful.

Video 8A: Data Preprocessing for Analytics Using the Data-Preproc App: Part 1

Greetings. Let me walk you through the data pre-processing app, and we do need data pre-processing for analytics. If you're wondering why, let's go through that a little bit. So, why should we care about data pre-processing? To be useful, data must first be usable and to be usable, data must be clean and consistent.

Data cleaning is a huge job in data science, I mean, it takes up a lot of time and effort and deservedly so. It involves a variety of things. I'm going to provide you one platform that would take all of that into account. What does it mean the data clean and consistent? Well, part of it certainly means that there are no lost or missing values. Data imputation is one of the approaches we can take too dealing with this. It means there are no misidentified columns. Imagine a non-metric variable mistakenly used as metric.

Okay. Well let's say gender 01, you know, male-female, is being used directly as a metric variable, that would not be clean and consistent data. It should have sufficient variance in every variable. If there is no variance in a particular column, what we would end up with is a constant. Variants implies informativeness. And provided the data have adequate transformation.

So, think about re-scaling of variable, standardising variables, normalising them, creation of dummy variables and so on. All of these operations may be necessary to get to data ready for analysis. Okay. Usable form for analysis. The data PreProc app provides a one stop small sample way forward for us. So, let's have a look at this. Okay. I'm going to open the app and now that the app is open, let's get there. These are our output tabs. These are some input UI, User Interface elements.

There are some example datasets. So for instance, you can go to the diabetes data set and if you click here it gives you a description of the datasets. So there are these three example datasets. Let me walk you through some of this. So, what I'm going to do is basically introduce to you the input user interface elements. What are the output tabs? We will then go into each element and examine the workings.

All right. So, this is what it looks like. Now, this part is the app name and description, basic data preparation and analysis. This part is the file input field. This is where you can read the input file into. These are user input fields. The file may have different separators and so on. I will show you what these mean and different missing values. And finally, you have the output tabs. So each of these tabs you click on them, basically we can see what happens in the analysis output.

So, let me walk you through the data. I'm going to use the Indian diabetes dataset, this is up on your LMS as diabetes dot CSV. This is what the file looks like if you open it. So for instance insulin, all these NAs. NAs means "not available". These are all missing data fields for instance. So, clearly this data is not clean or consistent. We'll have to make it usable. How do we do that? And look at this outcome, Yes/ No. This is a non-metric variable, and so on. The variables are self explanatory. But like I showed you out there if you want to see, so go to diabetes, click on this question mark here, and you will get these other major variables in there.

So, let's see, which of these variables are metric? Which means either there numeric integer or ratio value. Pregnancies, number of times pregnant, metric. Glucose, well, glucose concentration, metric. BP, mm Hg, metrics, skin thickness measured in millimeters ratio metric, insulin metric.

This one is non-metric. Okay. It's basically a binary variable, yes or no, in some sense. So, let me get there and show you how this works. So, please go here. Each of these are different tabs. Please go here, click on "browse", and read the data in. You must have downloaded this from your LMS just to read it in. There it is. Once you read the data in, down here, you will see this has changed. And you can see all those NAs, they're all blanks now. These are missing values.

Now, if you try to run this through analysis directly, the machine will drop any rows that have missing values by default. We don't want that. I mean, we would be losing information. So, we would ideally want something else to do. Now, this is a CSV file comma separated. Which is why automatically, it will take the comma as a separator. Sometimes, the separator could be a semicolon, a tab, a space and so on. You can choose any of these. All right. So we are here, now what?

Let's go to the first output tab EDA. This is what it will look like. Okay, So we have read in the data, and we are now going to pre-process it, right? So, we figured out all of these things have happened. Some data are missing, some variables are non-metric. So now, let us go ahead and explore the EDA Output Tab. All right, now let us explore this second tab. The EDA output tab. And what is our main goal here? Pretty simple. We're going to screen the data for missing values and inconsistency.

So I've shown you what they look like, now let us dig a little deeper into this. Two particular questions of interest. What is the size of the dataset, set and which variables have been identified as factor or non-metric versus metric?

Which could be numerical integer. Let me go to the app and let's have a look at this. So let's go to the EDA tab, and you can see 4, 5 different sub tabs here, screen summary. I'm going to go through each of these. What is the data screening sub tab all about? Well, it screens data for missing values, verifies column names and data types. An important piece of information it gives us, is the size of the dataset, 768 rows times 9 columns. What are those 9 columns? These are those 9 variables as we can see, yeah, and for each of them it gives you what the data type is.

So, for instance pregnancies, glucose, blood pressure. These are all integer variables. BMI as a numeric variable integer and ratio. Numeric as ratio would be metric. In the entire 9 columns that we have, only 1 variable is non-metric. The factor variables, we also call them "outcome". And this has two levels, yes and no. Metric variables don't have any levels per say, district levels. Okay, then we have two other pieces of information: missing and passing date missing. Take blood pressure for instance. Okay. 35 out of 768 observations and blood pressure are missing, which corresponds to about 4.5% of observations missing. Overall, about 11 observations are missing. And there would be 432 rows that have at least one missing value.

So, if we did not impute missing values, we would lose over half our data. All right. Let's go to summary. Now, basically, it shows you all the variables that are there. Let's go with glucose for instance,

it gives you what is called a uni-variant analysis of glucose, descriptive statistics. Okay. It tells me there are 768 observations out of which 5 are missing. This is the mean or the average, median and mode. This is the variance. This is the standard deviation and the variance. This is the range maximum, minus minimum would be the range.

So, it just gives you information also about extreme values, about each of these variables. So, and also if you go to insulin for instance and submit, there it is, it will give you the data for insulin similarly. Okay. All right. If you come here, frequency qualitative, there is only one. Qualitative also means "non-metric", also means "factor". There is only one non-metric variable. Put submit in there, and this is what you get. Out of those two levels, no and yes, 500 are no, 268 are Yes, as in, the outcome is diabetic.

And you also get a small bar plot that comes along with it. The quantitative ones are all the metric variables. Okay. So let's go with, let's say blood pressure in this case, glucose, submit. It gives you a frequency table and a histogram. Ok. Well there are 768 observations. So let me take this all the way to 20 different bins, and you can see that there are now 20 of these bars, and there, this is the distribution of the glucose variable in the sample. I mean, that's what it would mean.

Finally, we have correlation. So, these are the seven variables of the features, and as you can see the diagonal is 1. The correlation of a variable with itself is always 1. So glucose correlated with glucose gives you 1. one. 1 means perfect correlation.

What does it mean? What does the correlation number mean? How do two variables move together? So, just as an example, age with insulin point 24. It means that if A as age goes up, insulin has a 24% chance of going up, and that's basically what it would. That's one way to look at correlation. As age goes up, there is a 13% chance 0.13, that BMI will go up as well. That's one way of looking at it. Now, let me come back here to the slides. So, what is the size of the dataset set? 768 times 9.

Which variables have been identified as factor? We did this one. We saw summary statistics, you could just choose that and click submit and you would see these things. We saw frequency qualitative and quantitative. We saw the bar plots that come along with qualitative, and with quantitative.

The histograms that come. And finally we saw correlation tables as well. What is the correlation between age and blood pressure? So this is age, this is blood pressure 0.34. As a goes up, there's a 34% chance your blood pressure will also go up. I mean that's what the data are saying for that sample. Next, we're going to get into non-metric detection and the conversion tab the next time.

Video 8B: Data Preprocessing for Analytics Using the Data-Preproc App: Part 2

All right. So, time to head to the non-metric detection and conversion tab. The big question we answer here is this one, are there any non-metric variables? OK. Factor variables, categorical variables that were erroneously identified as metric. Take the bus number example. We know that 20 is greater than 10. But what does it mean that bus route number 20 is greater than bus route number 10?

It doesn't make sense. Even though the column bus route number will read numeric or integer, it is actually a non-metric variable. So, these are the kind of things we have to be careful about because they can seriously impact downstream analysis. So, what I'm going to do is... Well, let's walk through this part. We were here at EDA. Let's get non metric detection and conversion.

You have, this is the uploaded data structure and this is the after conversion data structure. The two are identical now because we haven't converted anything yet. Now look at this, these are our variables. These are their classes. There is only one non metric so far, and this is the unique value count. What does it mean? Among the 780 odd, 736 observations that we have, pregnancies, the number of pregnancies basically take 17 unique values.

Glucose levels, take 136 unique values and so on. Now imagine for a minute, okay, that pregnancies, this particular variable is actually non metric or we want to treat it as non metric. That you know, a person who's had let's say, three pregnancies is different from a person who has had one in terms of propensity for diabetes. Okay, just to, let's consider that to be the case as an example. So, I'm going to convert pregnancies, which currently reads as a metric variable, to non-metric form.

Just click on convert and there it is. You will see that this variable has been converted to factor form, I mean, that's basically it. You can take other variables and convert them similarly. I'm using this one, it has the smallest unique value count, that is why. After you have converted these variables to factors, let's go to the next tab, which is missing value imputation. Now, we have seen a little bit of this in the EDA tab also.

These are the variables, 6 variables, in all which show missing values. Insulin has the highest number, almost 48% of the missing values. Now what do we do? If you don't do anything during analysis, the machine will automatically delete any row that has missing values. That would be a waste of information. In this particularly data, set view was based over half the rows that are available to us.

So let's go to the second sub tab, which is imputation. Automatically, all the metric continuous variables are taken up here directly, and we have a number of imputation methods. Imputation means we are going to guess the best value for that missing variable and put it there. I am

going to use the KNN or K-Nearest Neighbor method. It is a fairly solid, robust, popular method to do imputations with.

For categorical variables, we can use more. This is the only implemented method so far. So there it is. How do I impute these variables? Just click on "impute" and it is done. So, these are the variables number of values that have been imputed. If you go down here, you will see that there are no missing values anymore. Go back to the first tab and you see that there are missing values. There are all these missing values. But here, there are no missing values. These values have been imputed using the Knn If you go further to the left... So you click this, you know, you slide this lighter to decide. This also tells you which of the variables have imputed values.

So the first three, the top three in insulin were imputed. These three are imputed values. Because originally they were missing, and so they have been imputed in that steam. That's the way you look at it. We will head data transformation next. But before that, let me walk you through the slides that will summarise what I just did. We examine the unique value count and we found that, you know, this is the number of unique values for each of these variables. We picked pregnancies which has the smallest unique value count and we considered that maybe this is categorical.

Let's assume this is categorical. So, we converted this into a non-metric or a categorical factor variable. And when we did that, you click on "convert", it converts it to a factor firm, you saw this. I'm just summarising it one more time. Then we went to the missing values imputation tab, where we found these two sub tabs in there, these two, then we found these six variables having missing values. So, we then went to the second sub tab imputation, and there are some options for imputing the missing values.

These are the defaults that we used, and after we imputed them, we found that there are now no more missing values anymore. So basically, all of these have been imputed using K Nearest Neighbors, kNN method. Next, I'm going to go on to the next tab right after this. All right. So, let us now explore the remaining last two tabs data transformation.

We will start with that. Let me get here. So, after missing value imputation we go directly to data transformation. But these are all the numerical or the metric columns and we can transform them. They could use a variety of transformation methods, re-scaling methods. These are common in the literature.

Let me show you what normalisation does. For instance, look at this, glucose 148, 85, 183, 89 and so on. BP 72, 66, 64. All these columns have different scales, different minimum, different maximum, different variants, and so on. If I do a normalisation transformation, just click on this, and the variables are transformed. What it will do is, it will take the same variables and re-scale them to fall between 0 and 1. So think of normalisation as a percentile. The first one, here, glucose 148.

This is the 67th percentile in the sample. The second one was at the 26th percentile. 85 is at the 26th percentile, 90 percentile and so on. The same holds for all the other. Insulin, the 17th percentile, in the first case, the 17th percentile and so on. This is one way of transforming. Another transformation quite commonly used is where we do what is called "standardisation". Mean becomes 0, and variance becomes 1.

So, if I do like this, and you slide it towards this side, you will see that. Now the mean is 0. So, any positive number is above the mean. Any negative number is below the mean. So what does this mean, 0.87? It means that this value 148, is point 87, standard deviations above the mean.

It also gives us some idea about how far away it is from the mean. This one is two standard deviations above the mean, it's close to becoming an outlier, it would be above the 95th percentile. 2.47 standard deviations above the mean and so on. So, these are common things that we could do, and then you can download the transformed data for further analysis down the line. It is important that we standardise. In certain cases, it will become important. This is something that we could do directly.

Last but not least, we have dummy encoding. What does this mean? Now, this applies only to non-metric variables. So in this case, we have pregnancies and outcome. Just click on this and delete. I'm just going to delete it. Let's just stick with outcome. And what am I going to do? I'm going to... So, this is outcome. Outcome is yes, no. I am going to convert what are called "one hot encoding of dummy variables".

So once I convert it, what happens? Okay. So if you go here, this is what we see. Outcome no. If the outcome was... Remember outcome was either yes or no? If it is no, then this becomes 1, else it becomes 0. Okay. So, the second row outcome was no, the third row outcome was yes, the fourth row outcome was no and so on. And the opposite is true for outcome, yes. This is a powerful method, dummy encoding. A lot of the times, a lot of software will do it automatically, but the data PreProc app provides you a way to do it explicitly. And finally, you can download this and use it. Let me quickly go back and summarise what I have just mentioned.

So sometimes we transform the scale of metric variables. A lot of machine learning applications will require that the scale be the same. So for instance, if a variable is measured in Kgs, person's weight is let's say, 70 Kgs.

I transform that to grams, it becomes 70,000 grams. The scale has changed. Means the data still the same, but the scale has changed. We don't want scale to affect our analysis, so we may re-scale things. Standardisation and normalisation are different ways to do that. So this one, for instance, I chose standardisation. And when we do that, now, this is the original data. This is the original data.

And after transformation, if you slide it towards the right, you find the standardised data, for instance. So there it is. And then you can download the dataset for further analysis. The last tab, dummy encoding, was where we created these binary 0 1 variables, sometimes also called "one hot encoding". If you look at outcome, this was a non metric variable, non-metric variable. Yes, No. Two levels it had. And when I do conversion into dummy's, what do I get? I basically get these two columns.

Yeah. Outcome no, outcome yes. Corresponding to the two levels. When I need only one of them, no but, you know basically I'll get both columns out. So when this is yes, this becomes 1. When this is no, this guide becomes 1. And that's it. And you can download and carry on as usual. Time for a quick exercise based on what we've just seen.

Video 9: Basic Data Structures

Well, let me walk you through some basic data structures. Some of this may seem very basic, but some of it offers some perspective on what we mean by data sizes. Let's look at data structures and let's start simple. The very fundamental level, the very basic level, in terms of Data Algebra you have the Scalar, zero-dimensional array It is a single point of data that's about think of the number 42, think of the word India. These are scalar, zero dimensional arrays. A vector is a one-dimensional array.

The dimension is called length. An ordered collection of scalar. So imagine, I have a five lengths scalar 23,12,17,8 and 43. Five numbers in that order put together, that is a vector. These could be a scores in the last five games so to say. Vector of length 5. One-dimensional array, the array that dimension is called the length.

This one, for instance, is a 13-dimensional array, one-dimension 13 length vector. Now, we've matrices, two-dimensional array. Again remember, scalar, zero-dimensional vector, one-dimensional matrix, two-dimensional. What are the two dimensions of the matrix? Rows and columns.

Any structured data table with rows and columns is a matrix. Of course, all of them have to be numeric, typically we need to extend data science in numeric. So, imagine I have the readings of ten people on age, weight in kgs, height in centimeters than what I would get us a 10 by 3 matrix, 10 people 3 columns. This one is another example of a matrix, in some sense, the two-dimensional array.

The TSM is a name, but typically you know it would be..Now, we are generalising beyond two dimensions. We are going to the third dimension. The third dimension, 3-dimensional arrays are what we call tensors. You might have heard of Google's TensorFlow this is where it comes from. Tensors are three and higher dimensional generalizations of arrays. So, imagine image data. So, image data have height, breadth and color channels typically, RGB, the three primary colors.

So, imagine I have an image of 144 height, 256 length typical YouTube aspect ratios. And you have three primary colors. So, I would have a Tensor of size 144X 256X3, effectively that's what it would be, so there it is. Higher order tensors are available now, so if I take a tensor a collection of tensors, 3-dimensional tensors would become a 4-dimensional tensor a collection of 4-dimensional tensor becomes a 5-dimensional tensor and so on. Video data for instance, is a 4-dimensional tensor. Let me have a look at that. Let me come to storage and sizes, and we'll connect it also the data structures that we did just now. Fundamental unit of data storage is what we call a bit.

The idea basically comes from binary digit the B from binary and IT from the last words of digits. A Bit stores a binary value. Either a zero or a one. It is the most fundamental storage unit that is available zeros and ones that's how it is. A byte is typically a 8-bit storage unit that can encode up to 256, 2 power 8 values. You have bytes defined of different sizes but typically 8-bits as a byte is the most common. So, there it is. Up to 256 values can be stored in a byte, 8 bit byte.

This is a Wikipedia definition basically, the byte is the number of bits used to encode a single character of text in a computer it is the smallest addressable unit of memory. Basically historically this has been true. What is a Kilobyte? Kilobyte is a 1000 byte, with typically 1,000 we look at it as 2 power 10, 1024 bytes. A megabyte is a kilobyte square, so 1024 square a million bytes typically and so on.

You have a gigabyte, which is 1024 megabytes. Then you have a terabyte, which would be one 1024 gigabytes and so on. Let me give you an example of a 4-dimensional tensor and how much in terms of size it would be, how much storage space it will occupy. Imagine you have a still frame in a YouTube video, right? And it has a 144 x 256 pixels and it has three color channels. Imagine the length of the YouTube video is 60 seconds and the frame rate the number of frames per second is 4, 4 is very less, let's start with 4, 4fps. How many values do you need to store?

How many pixels... So, you have 3 color channels times 144x256, which is your image size, times 4 frames per second times 60 seconds, which comes to about 106 million change values. Now each of these values if you're using the 32 bit computer, 32 bit is now gone everyone uses a 64. Let's go with 32 bit, 32 bit float per value. So, if you're storing your floating point into a 32 bit storage unit then this would come to about 101.25 megabytes of storage, without compression, that is heavy.

That's just for a 60 seconds YouTube clip. Typically, it is smaller because we use compression algorithms lot of pixels are basically you know that I know it, you could in some sense compress them. But this would be how much space it would take in the normal course. So, quickly to summarise what we just saw about basic data structures. And we also saw about data storage in sizes. There are two different ways of looking at the data. One is on the software side, where we interact with data structures, we interact with vectors and matrices and tensors.

That's the way we program things. On the other side data storage and sizes basically are more related to the hardware aspect of things, kind of brought the two together it is necessary and important, even if we are on the software side to have an idea about the hardware side of things, at least you know so this particular data of this size, how much storage would it occupy, and vice versa as well to bring it all together and put it up there.

Video 10: Common Sources of Data

Let us explore some common data sources, which is an important part of data preliminaries. Some common secondary data sources, primary data, of course, has to be collected in context for particular projects. So, think of a business. What are the common data sources, secondary data sources that may exist? At one level, we can broadly classify them into internal sources, internal to the business and then there are external sources, which are external to the business. What kind of internal to the business data sources might there be?

Think of records. Sales records, Customer records, Customer complaint records. All sorts of records that typically kept in organizations. All of this is secondary data. All of this can be mined for insight. All of this is available. It exists. Think of systems that automatically collect data anyway. Think of an ERP system. Think of the amount of the data it collects. Right from material, all the way to inventory. I mean, you name it. ERP has collect a lot of information in manufacturing kind of settings.

Think of a CRM system or an SCM system, the amount of data and the kind of data they collect. I mean, they have data there. It's up to us to use analytics in some sense to analyse them, to define the problem, formulate the problem, to extract insight, all of it. Think of the business' website.

Huge trove of data. Clickstream data is available. Who is visiting your website? From where? What are pages did they visit after the landing page? How long did they stay there? Did they put up a query? Did they make an inquiry? Did they look up something of interest? A lot of mining can be done, a lot of insight to be had. Think of apps that businesses make, app-usage data, the very common data set, now a days available and so on.

I mean, these are internal to the business, typically, these are proprietary data sources that we have today. Among the external data sources, you can broadly divide it into public versus private/proprietary. Public data sources, well, you have the World Wide Web. You have Twitter data, for instance, is technically public though to collect it, and use it for commercial purposes,

one has to, basically ensure that we are clean on copyright. It is best to use their API. They define their terms of services there. You have government, a lot of government data, particularly in the West.

Now, also in India VSE. In the proprietary/private part of the external data sources that we have, typically, we have entities called syndicated data providers. These people, for them, data is their business. They collect the data, package it, sell it . Effectively or lease it out or provide it for a fee. Examples, well, we'll see a few. Let me start in some sense with government and web first on the public side and then we will come to the proprietary piece. This, for instance, is NYC open data. It's a website. You can google it up.

So, New York City is a city government has opened up a lot of data about NYC, on where can you find Wi-Fi in your neighbourhood and so on. These kind of things available. Where can you find the nearest doctor? The nearest hospital? The nearest school? I mean, all of these things. Once you give your geocoding, it basically gives you a lot of information of this kind. Kaggle, for instance, is a rich source of data.

Again, this is on the website, for instance. Businesses give data out here for contests, Kaggle datasets are all over now. Syndicated data providers on the proprietary side, a good example or good examples would be Capital IQ from S&P, for instance, or Bloomberg on the finance side. You have IRI and Nielsen on the marketing side. I mean, you have these syndicated data providers. You've got Gartner, Forrester, et cetera on the technology side.

All of them syndicate data, collect the data, lot of it primary, sometimes secondary. Package it and provide it for a fee. So, where do API's fit into this picture? I mean, they are somewhere between public and private. Let's have a look at them. What is an API? First off, I mean, you can look at the web definition. But basically, an API stands for what we call an application programming interface. It is an interface between two applications. That's what it is, two services.

So, what it does in some sense is allows for data transfer across the interface. You can send a query and that will send you data for a fee, of course. I mean, you have free APIs, but then not that many of them left anymore. Examples of APIs, I mean, this is the Yahoo suite of APIs. Yahoo weather API, Yahoo Finance API, and bunch of them out there. All right. There is an API for that. This is basically a take on an old Apple ad that says, there is an app for that. Today, there is an API for it. Why do firms put out APIs? Think of a Facebook or a Google. I mean, why do you put out APIs? Why do you want to provide a data even for a fee? Google maps API, for instance. Why?

Why would you do that? Well, data is currency. Data is money. You want to monetize your data. I said, this is one of the ways in which you could do it. Another big reason tech companies do this is they want to invite developers to come onto their platform and deploy their cool stuff there. These are not data APIs typically, but APIs all the same. And so, some examples of APIS and so on, out there.

Which domains might APIs most likely be found of in? So, you've different APIs and different domains. Let's have a quick look at this. And this is the growth since 2014 of APIs in different domains. The number one growth that you see out there, the highest growth by far is in data APIs, followed by financial APIs, followed by analytics APIs and so on. So, that's basically where the world is moving. That kind of data in some sense has a lot of value, where they're seeing the largest amount of growth.

Video 11: Data Preliminaries for Analytics: Summary

So, let me summarise all that we did and we did quite a bit in this module. We started off with the motivating example that sought to motivate data, its value and how it ties into business valuations. More generally, we saw the Uber technologies example. We then went into data and measurement basics.

We looked at definitional preliminaries, what is data, why measurement and so on. We then went into data types and dichotomies. We looked at the main data dichotomies, the three primary ones. We then went into psychometric scaling, the four primary scales and thereafter figured out the metric versus non-metric dichotomy as well, the fourth one. We then went into data pre-processing for analytics using the app. The app basically did a lot for us, If you recall all that happened.

We went into detection of metric and non-metric variables in a structure data table. We went into data imputation for missing values. We went into scaling of data, prior to analysis. We went into creation of dummy variables for non-metric data, I mean, dummy columns and so on, all of that. Then we came to basic data structures and we saw these structures along two lines on the software side and of the basically saw data algebra. You know, we saw scalars and vectors, matrices and tensors corresponding to 0 1 2 3 and higher dimensional arrays.

We also saw how these translate into data sizes on the hardware side, data storage requirements that would come up eventually. Data sizes are a big deal and we train machines. Well, machines need a lot of data for training and also for storage and so on. We will see. And finally we came to common sources of secondary data, where in addition to common sources, we also discussed the bare basics of APIs. With that, this module comes to a close.