Slide 1

In this section, we will talk about Exploratory Data Analysis.

Slide 2

What is Exploratory Data Analysis?

Essentially, it is the assessment of the quality of the data available to tackle business problems. When we do analytics on data, one of the first things, we want to ensure is that the data that we are using for our analysis is of good quality and has the attribute that we require for the data analysis.

At the end of this session, you will be able to answer the following questions about the data:

- What is data?
- If I have a dataset, can I immediately start creating models?
- What does it mean to talk about quality of the data?
- What is good data versus bad data?
- How do we look at characteristics of data?
- How do I summarize information in data?
- What happens if my data does not confirm to expectations? Is that a problem?

Slide 3

Where does data exploration fit in the analytics methodology?

Remember for any analytics project, the first step is to make sure that we define the right problems statement. Once we have a problem statement defined, the next step is to design the solution. Once the solution design has been accepted, the third-step, which is the longest step, is the actual implementation of the solution design.

Data exploration is typically is the first exploration within solution implementation. So what we are going to do is, we are going to look at how to explore data, once a solution design has been firmed up.

Slide 4

Typically, data exploration and data preparation are done together.

Within data exploration, we will take a look at data extraction, data integration, and the assessment of data quality. But data extraction and data integration are pretty straight-forward. We will spend most of our time in this topic on how to assess the data that is available to you for the analytics project.

Within data preparation, there is data cleaning, data transformation and data reduction. But for all of this, the input comes from the data exploration stage, which is why, we have mentioned that many times that data exploration and data preparation are done together.

Slide 5

We will understand data exploration also with the help of case study and some data.

The case study involves a mobile service provider that has noticed a lot of attrition in customers subscribing to their services. Attrition is essentially customers leaving or customers quitting, and obviously the mobile service provider would like to reduce the number of people that are leaving its service. In order to do so, it wants to understand, what exactly is driving attrition, and therefore, identify potential options to retain customers.

What is the data that are available to us? There are two kinds of data that are available i.e., usage-related information; how much our customers using the services, what type of plan our customer using, are they participating in promotions and so on. So usage-related data and there is also subscriber-related data. So the age of the subscriber, the location of the subscriber and so on…

Slide 6

Here some data from that particular case study. You can see that there are lot of variables, but this itself is a sample. There are more variables in the original dataset. There is also a data dictionary available to us, which has some definition of variables. So very quickly,

There is subscriber id, which is unique subscriber id, which is in column A, one row is per subscriber.

There are minutes used 1 to minutes used 21; 21 different variables that show usage of that particular subscriber in minutes of the mobile service in every month post acquisition. In our sample dataset, we have listed only four months of data, but remember that in the original dataset there are 21 months of data.

There is a third variable called plan type, which is essentially, what service plan that the subscriber start with at the time they started the service. So you can plan types has some different description of the service plans.

There are some promotion variables; promotion 1 to 21, which essentially are indicator variables. They have 0 and/or 1 values. This playing is whether or not the subscriber participated in some sort of promotion every month post acquisition. So you can see promotion 1 has a value of 0 for this particular subscriber, this means that this particular customer did not participate in any promotions in that month whereas for this particular customer '40992265' participated in a promotion in the second month, but not in the first month, the third month or the fourth month. Similar to usage data, we have promotion data for 21 months in our source dataset.

There is also Service Start Date and Service End Dates. These are date variables and essentially they list when a customer started the service and if the customer has quit, when the person quit the service.

There is some demographic information about the customer; there is birth date and zip code. Zip code is a US location code.

Finally, there is another variable called New Cell Indicator which essentially indicates there are not new cell phones was bundled with the service plan at the time of service plan start.

So this is the data that we have, and remember, what we want to do with the data eventually is understand factors that drive attrition in their customers.

Slide 7

But before we can do the entire analysis of the factors that drive attrition in customers, our first step will be,

- What do we think of the data?
- Is the data good enough for me to do an analysis?
- Is the data good enough for us to do an analysis?
- What is contained in the data?
- Is that data sufficient?
- Is it complete?
- Is there missing information?
- What is the quality of this data?

And that is what, we are going to do in data exploration.

Slide 8

Remember data exploration sometimes may start with data extraction and data integration.

Slide 9

Data extraction may simply be getting the data that you require from databases.

Sometimes, all your data may not be from a single source. Sometimes your data may come from single database table or it is already in a flat file format.

But at other times, you may need some specialized queries to extract the right data from multiple databases.

Slide 10

If your data is coming from multiple sources then after you extract the data, another step could be data integration, which is essentially putting all the data together to make a master dataset.

Again, you may need to write some specialized code to integrate the data in the correct manner because the data may or may not all be at the same level of aggregation.

Slide 11

But let us assume that our data extraction and your data integration process are done and that you have master dataset. So what we are going to do now is that we are going to take a look the master dataset and assess the data for completeness and for reliability.

Slide 12

In data assessment, typically we want to answer the following question:

- What is the information that is contained in the dataset?
- What is the quality of this information?
- Is my data complete?

Let us start with the first step, which is, 'what is the information contained in the data?"

Slide 13

Remember if we want to use an analytical approach to business problem then clearly we are focussing on a data driven analysis, which means, that our data has to be of primary importance.

Let us first think very quickly about what exactly is data. Data is really information in visible form. In other words, this is information of any kind that is essentially accessible to us for use. So it is collected, it is compiled, it is accessible for us.

Remember also that when we speak of data all of us are thinking about data being neutral and factual. In other words, we want to use data because we believe data is objective.

Slide 14

Now another thing that you may want to think about before we start looking at data assessment is how was the data collected that we are going to be using for analysis, and why is it important infact to understand mode of data collection because remember the mode of data collection will give us some insight into the quality of the data.

Data can be collected in many ways.

- They could be primary data versus secondary data.
- Data could be collected manually versus from an automated system.
- Data could be census data versus small subset which is a sample.

Again knowing how the data was collected will give some insight into assessment of the quality of the data.

Why is that?

A simple example could be, if we know that data was manually collected versus from an automated, let us say billing system, then when you see unusual values, it may be easier to conclude when you have manual data collection processes that there may be mistakes whereas if you see the same unusual values in a dataset that is coming from an automated system, it may be harder to conclude that there is a mistake.

Similarly, knowing whether we are dealing with a sample or a census gives us an insight into the assessment of the data. So remember when given a data set, one of the things you may want to think about is,

- How did all these data come together?
- How was it collected?
- Do you understand what the data collection process was?

Slide 15

Once you know what the data collection process was, another preliminary step could be just very quick classification of data.

How is data classified?
Typically, the easiest classification is to talk about qualitative data versus quantitative data.

Qualitative data is essentially text-based data. So something like a name, or an address, is qualitative.

Quantitative data is numeric data. Data that has numbers in it, within quantitative data; however, they could be again other sub-classifications. For example, discrete data or continuous data; discrete data is a data that can only be in whole numbers 1, 2, 3 etc., Continuous data is a data that can take decimal values like 3-1/2, 4.67 etc., they could also be date and time data.

Again within discrete data, there could be nominal data and ordinal data. Nominal data is when you have numbers, but the numbers do not have any ordering. For example, supposing that we give you a zip code or pin code that has number 560008 and there is another pin code that has number 560009.

It does not make sense to talk about 560009 being greater than 560008 because these are simply codes for a location. There is no ordering whereas many data have orders in it, which is that, it would make it ordinal data. For example, let us say we classify low, medium and high as 1, 2, and 3. Maybe it is low usage, medium usage, and high usage. Now 1, 2, 3 have an ordering, 2 is greater than 1, 3 greater than 2, so that would be ordinal data.

So a good way to take a look at your data at very high level would be to take all the variables and classify them in terms of either qualitative data, quantitative data and there could be other classifications schemes as well.

Slide 16

For example, you may want to look at whether the data is primary versus secondary data.

Primary data is data that you have collected that is coming directly to you. Secondary data is data that is collected by other people not necessarily for the purpose that you have in mind.

You could look at actual data versus derived data. Derived data being information that has been derived from some source data not directly captured.

You could have classification scheme based on usefulness for example, where you say these are critical variables and some of these are not so critical variables.

Slide 17

So let us go back to that data that we were looking at, that the mobile service provider. If I wanted to look at data classification for the variables that were contained in the data set, at a very high level, this is how I would list.

Some of these are quantitative variables such as subscriber id, minutes used 1, 2, 3 so on. Some of these are qualitative variables; the plan types are qualitative variable. Some of these are date variable.

Within these quantitative variables, subscriber id, for example, is a nominal variable. There is no ordering. So what this data classification helps us do is just get a high level sense of all the variables in my dataset and what kind of information they contain.

Slide 18

Remember in data exploration, we are going to answering three questions:

- What is the information contained in the data?
- What is the quality of this information?
- Is the data complete?

Slide 19

When we looked at the information contained in the data, we answered some of these questions.

- How was this data collected?
- Are the fields accurately labeled?
- Is there missing information?

At this stage, when we are looking at fields being accurately labeled or are there any missing information, we are essentially comparing this data to the data dictionary and making sure that we have the right data. We will also come back to missing information within the data itself in the next stage.

In the next topic, we will now spend time on how do we assess the quality of the information.

Slide 20

In this particular topic, we looked at,

- What is data exploration?
- Why we need data exploration?

And the first step within data exploration, which is, how do I look at what is the information that is contained in the data.

In the next topic, we will spend a lot more time in actually assessing the quality of the data that is contained in your dataset.