Slide 1

In the last session we looked at the data exploration and the first three steps within the data exploration process for assessing data quality. First transform qualitative data into quantitative data, second generate derived variable, and third generate summary statistics.

When we talked about summary statistics for quantitative variables summary statistics are pretty straight-forward. Typically, the minimum, the maximum, the average, the standard deviation, the number of missing values.

For qualitative variables, summary statistics are usually generated via frequency distributions. What is the number of values or the number of observations for each value in a qualitative variable?

In this topic, we will take a look the next three steps within data exploration. Cross tables, graphical analysis, and anomaly detection.

Slide 2

What are cross tables? In data exploration, we are essentially assessing the quality of the data and making sure of the data. We

are making sure that we understand the patterns in the data and that we have good explanation for those patterns.

However, all the data exploration that we were doing so far has been at a single variable level; one variable at a time and sometimes that your data exploration will involve looking at patterns across groups of variables.

How do we do that? One way to do that is to generate cross tables. If you remember in the last session, we had looked at the distribution of age variable for the customers in the dataset and we had set that we expect the data to be skewed towards younger customers, which is not what the dataset was showing.

Now if your dataset is not confirming to our expectations, in this example, I was expecting higher proportion of people to be having younger age, it may or may not be a problem, in the sense that it could be that your data has not been extracted correctly or sampled correctly or it could be that it is really true of your data, but actually if you think about the age variable, we would be more interested in understanding is there a difference in attrition by age variable. Let us think that about for a minute.

Would you expect the difference in the percentage of customers leaving your service by age group? How do I look at that data? I would do a cross table, essentially, a frequency distribution of age along with attrition.

Slide 3

So if you look at this table you can see here, that we have taken the attrition variable, which has values of retained and lost and we are looking at proportion of attrition by age bucket.

Now this data is very interesting. It tells us that we have much higher levels of attrition for younger customers relative to older customers.

Now, is this confirming to expectations? Probably, because younger customers are more likely to move when they see an opportunity for improvement in customer service, or billing, or cost and so on.

Older customers are less likely to move. Now therefore, when I look at a cross table of age and attrition, I am seeing a pattern, which is easy to understand and which I was expecting to see in the data.

What is other cross tables that I could generate based on this dataset?
This is just one example, age versus attrition. There is many other cross tables that I might want to look at. For example, attrition and usage, I may want to look at is there a difference in attrition percentages based on high users, low users, medium users; attrition versus location that could be another useful cross table; attrition and promotion, if customers participate in promotions, are they less likely to leave or not.

All of these are examples of pattern that are useful to look at within the dataset that you can generate across variables using cross tables.

Slide 4

Here is another example of attrition by duration. It is a cross table of attrition by duration. Churn is essentially a word for attrition, and duration is how long the customer has stayed on as a customer.

You can see with this particular table that if I look at the percentage of attrition by duration for the first month, second month, third month, fourth month… those numbers are relatively stable. However,

when I get to the twelfth month the numbers are really shoots up 7.46%.

How would I actually use this information? This definitely tells me that at the time that one-year anniversary of a customer is coming up, there are lot more people that are switching service providers.

So this may be a useful piece of information for me to know in terms of preventing attrition. Maybe we need to generate strategies that reach out to customers at the one-year anniversary given them an incentive to stay on. So this is an example of cross tables where we are looking at relationships across variables, groups of variables rather than individual variables.

Slide 5

Another step in data exploration is graphical analysis.

Slide 6

In fact, a lot of people start with graphical analysis because data visualization is sometimes are much more powerful way of understanding data and patterns within the data.

When we talk about graphical analysis, essentially, we are talking about visualizations that can be used to determine distribution of the data to identify the spread of the data to look at whether there is biased or skewness in the data and also visualizations that help us identify outliers in the data.

We will talk in great detail about outliers when we talk about data preparation, but for now outliers are essentially unusual values in the data; data values that look like that they do not belong to that particular variable.

Slide 7

So remember graphical representation are essentially used to gain the visualizations of the data to understand in a much easier fashion than looking at summary statistics in tables.

There are many graphical representation including,

- Simple run charts
- Frequency distribution like histograms
- Range charts; box plots and stem and leaf plots
- Joint distribution charts and so on.

We are going to take a look at a few of these.

Slide 8

The easiest way to look at data for a single variable is to simply do run chart. For example, I have generated a chart, which has usage data for month 1 against the subscriber id. So each point in the visualization is one subscriber and their usage.

Now if I look at the run chart like this, what I can do with the run chart is essentially gain an understanding of the values of that variable. I can assess the distribution of the values, I can assess the spread of those values, and potentially I could identify anomalies or outliers.

If we look at this particular run chart, we can see that data values range from 0 all the way to little more than a 1,000. Most of the data are the average is somewhere here around the 200 range. There are some users that have high usage, and we can identify that around 600 to 700 range, and potentially there is this one user here around the 1,100 range who is maybe an anomaly because this particular user usage is very much different and very removed

from most of the other customers in the dataset.

So remember, this visualization again is used in the data exploration stage because we want to get a high level understanding of the data.

Slide 9

We could do the same visualization using a histogram, which is a frequency distribution. So I took the same data, but now instead of run chart I have generated a histogram.

Again, it used to identify spread, distribution of the data and potential outliers. We can see here that there is one point here that says very different from the all the other usage data in that variable.

Slide 10

Another visualizations could be a probability plot, this is an example of a normal probability plot.

Normal probability plot is used to check whether or not the data in your variable is normally distributed. The normal

distribution sometimes is a requirement for applying some analytical algorithms.

So you may want to check, is my data normally distributed? If your data is normally distributed then you should expect to see a straight line in the normal probability plot. In this particular example for minutes used 1, it is approximately straight. So this data is approximately normal.

However, you can still identify an outlier here. Remember that 1,020 point that we had seen in the histogram and in the run chart.

Slide 11

Another great way to look at data is to generate the box plot, especially if we want to identify spread and identify outliers. Box plots allow us to look at measures of central tendency. The lines in this box plot; this is the minimum, this is the maximum, this is 25th percentile, 75th percentile and this is the 50th percentile. The plus sign shows us where the average is.

If you see a long tail in a box plot then typically those are outliers. So we can see that again this is the same data, so you

can see that the same data point has been identified here as an outlier because it is very removed from the rest of the data.

Remember that all of these visualizations are essentially with the intent of understanding the data. Some of us may prefer to use box plot, and some of us may prefer to use histograms and there could be different visualizations for different kinds of data.

If I wanted to look at visualizations of two variables at a time, I would do like a scatter plot. A scatter plot or an XY plot will show me is there a relationship between two variables X and Y.

So there are many kinds of visualizations that we may employ in order to understand the quality of the data that is available to us. But remember ultimately the visualizations are used to assess the quality of the data to gain a big picture understanding of the data that is available to us.

Slide 12

So what we have looked at is cross tables and graphical analysis which are steps 4 and 5.

Slide 13

The final step is anomaly detection.

Now, what is an anomaly?
An anomaly is something that is an unusual value. So there were some anomalies that were evident from summary statistics. For example, negative minutes for the second minutes used variable, a 177,000 for the fourth minutes used variable and potentially truncated information for the zip code.

Now again, these are all anomalies because they do not seem right. There are many other anomalies that could potentially be present in the data. Some of them are evident from doing single variable analysis whether we do summary statistics or visualizations.

Other anomalies may be generated from looking at multiple variables at a time. For example, a potential anomaly in the data could be if service end date exists then does data exist for usage post the end date? If that does, then we know that we have a problem.

How do I identify anomalies?
Again one way to do that is summary statistics and visualizations. The other

way to do that is to look at variable values across multiple variables at a time.

But all this exceptions anomalies need to be investigated. Remember an anomaly is only a potential anomaly till we look at it and understand what is causing it. There may be a valid explanation for this extreme value or an anomaly.

We need to make sure that we do not take out variable values that we think our anomalies without understanding why that is being caused.

You do not want to create a dataset that has no anomalies at all because then your dataset is very generic and it may not capture the range or behaviour that you see in real life in your customers or in your transactions.

Again we will spend a lot of time on outlier detection and analysis when we do data preparation.

But at this point, it is important to understand that one outcome of exploratory data analysis is flagging potential outliers and anomalies.

Slide 14

So what we did in data exploration was? We answered the following three questions:

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

Remember is the data complete, is really answer that answer is going to come from the answers to the first two questions; what is the information contained in the data and what is the quality of this information. If we believe that the data is of good quality and the data is complete, then our next step will be to use the data for the actual analysis and the modelling process.

Slide 15

To summarize, remember, it is very critical to invest time in understanding the data because it is the starting point to build any solution and if your data is of bad quality, your output or your modelling result will also be of bad quality.

Model results will only be as good as the data that goes into it; and therefore model results will be reliable only if the right models are applied to complete data.

There are many techniques of Exploratory Data Analysis (EDA) that can be used to gain an understanding of the available data, the relationships between variables in your data, and flag potential issues with the available data.

These include summary statistics, graphical visualizations, correlations and cross tables; many techniques. It is however, very important to apply all of these to your data at the beginning of the analysis rather than investigate outcomes that seem unusual after the end of the analysis.

Slide 16

Let us quickly do a recap of what we have learned in the data exploration process.

Remember data exploration is an important component of the analytics methodology, and it is done in order to make sure that we are comfortable with the data that we are going to use for further analysis.

In the data exploration process, we answer the following questions about the data:

- What is the information contained in the data?

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

Slide 17

We looked at 6 steps within the data exploration process for assessing the quality of the data.

1. Transform qualitative data into quantitative data, where applicable.
2. Generate derived variables as required.
3. Create summary statistics or descriptive statistics to gain a big picture understanding of the data, and identify potential issues with the data.
4. Look at relationships across variables using cross tables.
5. Make use of visualization and graphical analysis to aid understanding of the data.
6. Use the data exploration process to detect noise or anomalies that maybe worth investigating.

Slide 18

What is the output of an Exploratory Data Analysis stage?

Sometimes or more often infact, the output is really internal. It is designed to convince the analyst that the data is good for analysis. So if the output is really going to an internal audience then essentially the output is aimed at making sure that,

- You have a good understanding of the data and all the variables
- You have clear identification and assessment of exceptions / outliers / wrong values / missing values
- Potentially some follow up questions for questions that you do not have answers.

Sometimes the data exploration process has an output that is presented to an external audience, in which case,

- You may have to have a more formal presentation and summarizing your assessment of the data
- It includes lot of visual showing interesting or useful patterns in the dataset
- Has a list of follow up questions that an external audience perhaps can help answer.

Remember investing time in Exploratory Data Analysis is very important to make

sure that you have an efficient and then interesting outcome of your data analysis.

A lot of times people skip the data exploration stage only to find at the end of the analysis that they do have to come back and figure out the issue with the data is.