Slide -1

Hello, in this topic we will talk about data preparation.

We had looked at data pre-processing and reviewed data exploration method in our last session, data exploration and data preparation are usually done together.

We had looked at data exploration methodology including data assessment in the last couple of sessions, what we will look at now is how to actually prepare the data following data exploration.

So, why data does need to be prepared?

How data is prepared and why is it important to invest time in data preparation?

So let's answer the first question why does data need to be prepared, data we get as raw input may not always be ready for data analysis so there may be some cleaning, some preparation or some transformation that needs to be done in order to generate data that is usable for models, the data needs to be checked and treated for consistency and completeness and we may add additional variables that have to be created that are required for the actual data analysis, so all of these is part of data preparation.

Slide-2

Now lets look at how data is actually prepared. Essentially data preparation deals with couple of specific tasks.

1. Identifying and dealing with outliers.
2. Second identifying and dealing with missing values.
3. How to prepare a qualitative data.
4. Creating additional variables or derived variables.
5. Transforming a data.
6. Potentially some data reduction.

We will take a look on each of these in this session.

Slide-3

Remember that data preparation really is making sure that the data is ready to use.

Lots of times you may need to clean the data or process the data before the data is ready. Now most common issues with data cleaning tend to be around outliers and missing values.

Slide-4

Let's start with outliers first:

What exactly is an outlier?

An outlier is a value of a variable that appears to differ significantly from the rest of the values of that variable; notice here that there are two key terms, differ and significantly.

If you look at the visualization on the right which of this points are an outlier. Actually lot of us will point to this one, that this looks like an outlier but it may or may not be an outlier.

In fact if you look at its distance from the line, these points also are probably as distant or more distant we typically may not identify them as outliers. So how do we actually identify an outlier? In fact are extreme values outliers?

Slide-5

In fact let us talk about outliers a minute more.

Why are outliers a problem, are they really a problem?

If you think about business data, essentially outliers represent a variation in a sample. How can that be bad? Remember if we look at a standard business data set for example customer transactions.

Now we do understand all customers are not the same. We expect that there will be some customers that spend a lot and there will be some customers those don't spend at all, the majority of customers will be somewhere in between.

So if you see a customer that is doing a high spend, is that necessarily a problem? Isn't that a representative of what we would expect to see in real life? However what happens if those extreme values or those outlying values are surprising, very very extreme that might be unusual and may not be representative of behavior.

Remember when talk about outliers we should be very clear about one thing,

Are outlier impossible values?

No, impossible values are mistakes. For example a negative spend, unless we know that a negative spend is a credit, in general when we say spend, it has to be a number zero greater than zero.

So when we talk out outliers we are really talking about improbable values rather than impossible values but again when we talk about improbable values, remember that there is some probability that these values are possible.

Therefore before we decide a value is an outlier we have to be convinced about how likely or unlikely is this value. In fact we need to investigate and understand, are there any special circumstances or conditions that produce the outlying observations that may not apply to the problem at hand.

Let's take an example; let's say we were standing outside a big grocery store and we wanted to understand how much time customers take on average to shop at the store.

So we stand outside the store and we ask some respondents on their way out how much time did you spend at the store.

So you can see that are some times listed in the table here 20, 10, 25, 34 and so on all the way to 49. Now if we take a look at this data, what is an outlier? Do you think, there are some outlying values here? A lot of us may look at 49 and suggest that this might be an outlier because it seems to be a much larger number than other numbers in the data.

However supposing that we added little more information, like the age of the respondent. Now if you

look at the data does 49 look as surprising? Not necessarily because both 49 and 35 are from respondents that are older.

So again remember identifying the outlier does require some thought and some analysis. Just because value seems high doesn't mean that it is impossible. Or remember outliers can be very low values, may be 10, that could also be sometime identified as outlier. But for business analysis, we don't want to cut off extreme values simply because they are extreme. It is not necessary that all extreme values are outliers there may be a good reason for existing of those high values or low values and we need to understand them before we decide that a point is an outlier.

Slide-6

How are outliers really identified?

Remember at the data exploration stage, one of our key reasons for doing data exploration is to assess the quality of the data, and one of the things that we will do as part of data exploration is to identify outliers.

The easiest way to identify an outlier is by graphical visualization. Though off course you can also potentially identify outlier by doing summary statistics.

Let's take a look at graphical visualization. In the data explorations session we had looked at run charts, histograms, box plots, probability distribution charts, all of these are a great way to identify outliers. Sometimes outlier identification requires additional knowledge of the business situation or the domain.

Slide-7

Let's take a look at some visualization.

We had viewed these visualizations in the data exploration sessions. This is from the telecom data set where we are simply looking at total minutes used in the first month after joining the service.

So in this run chart you can clearly see this may be a potential outlier. Notice that I am saying that it is potential outlier, it is because we don't know yet if this is extremely unusual behavior. It is certainly true in the context of the first month that it is very high usage

but what about the month after? Perhaps this customer just unusually uses the phone a lot month on month in which case it may not be an outlier.

Slide-8

I could look at the same data in different ways, different visualizations like I could do a histogram.

Slide-9

We could do a box plot.

Slide-10

We could do a probability plot. Again the visualization itself depends on your preference but most of these visualizations will easily help you identify potential outliers.

Slide-11

Another thing to remember is that again looking at variables one at a time may not always lead to good identification of outliers, sometimes you may need to look at pairs of observations.

Now if we have a data set with many many variables, should we look at all possible pairs? Perhaps not, again we may need some knowledge of important variables to understand which pairs of observation that we would need to look at.

Here is an example of outlier detection using a multi-variant approach

Slide-12

Supposing that I show you this cater plot which is basically a X-Y chart. I have got usage of month one on X axis and usage in month 2 on Y axis, could you identify any outliers here? This point potentially is an outlier. Why is that? Because you can see somebody who has used a lot in the first month but has a low usage in the second month.

Now if you think about cell phone usage in general for customers, customers that tend to be high users in single month will tend to continue being high users in the second, third or the fourth month because it is behavior. Customers that tend to use the phone lot less in any month will also probably use the phone a lot less in all subsequent months.

So this is what we mean by detecting outliers on the basis of multi-variant analysis rather than a single variable. So this is potentially an outlier and these also could potentially be outliers. These points here again someone who is using a phone a lot in the second month but very little in the first month that seems like an unusual behavior on part of the customer.

Slide-13

Here is another example of multi-variant outlier detection.

Imagine we had years of college education on the X-axis and we had income associated for every customer on the Y-axis. So this is that list of customers, you have years of college education on the X-axis and their incomes on Y-axis. Could you identify any potential outliers in this data? Most of us will pick this point because it seems a very far away point.

However think about what the chart says. The scatter plot, this scatter plot says in general the higher the years of college education higher the income that we expect to see and using that logic this is certainly not an outlier. This is simply saying someone with 9 years of college education, I will expect to see a high incomes, so it is not unusual behavior in any sense.

However this point is probably an outlier, why is that, because contrary to the general trend in the data we have a point here that has someone with 2 years of college education but very high income, this is potentially an outlier rather than this point.

So remember it is very important to understand that just because you have an extreme value doesn't mean it is an outlier. Some outliers are extreme values but not all extreme values are outliers.

Slide-14

Off course, once we have identified outliers, whether it is on the basis of single variable descriptive statistics visualization or a multi-variant analysis, the more important step is to determine what to do with the outliers.

So what can you do with outliers? You could potentially delete the outlying values but what happens if you delete the outlying values.

If you have lots of data with many variables and many observations then deleting few outlier values may not have a large impact on your analysis.

However if you have relatively small data sets, then you may not like to delete the outlying value. Because let's say you have two variables and one of those variables has an outlying value, if you delete that then that entire observation is lost, meaning even if the variable two has a proper value for that observation because variable one has been deleted for that value, you can't use variable two value for that observation.

So there may be cost impact to deleting outlying values when you have small data sets.

So if you didn't want to delete the values, you could off course replace these values with other suitable values. So instead of losing data, instead of taking an observation out because it has an outlying value you may simply substitute a better value for that outlier.

How do we identify substitutes?

There are many ways of arriving at suitable values for replacement.

Slide-15

For example let's take this data set again from telecom data. Now if you look at this point here, this is potentially an outlier because if I look at that variable minused4, that number stands out as very very large. And let's say that I have looked at the data and I am convinced that this is an outlying value.

What do I do with this outlying value?

One I could simply delete that value so the minused4 will be replaced with nothing but with a missing value.

But remember if you do that, then for this entire record this entire record becomes not usable. In my analysis, if I wanted to use all these variables including minutes use 4 because this has a missing value.

Now I can't use any of this data. Now again if I had a large data set, this may not be a big problem.

Let's say that I have a small data set, in that case I don't want to risk losing the entire observation because of one outlying value in one variable. In that case I might want to substitute this 1133 with a better number.

What could be a reasonable substitute?

Intuitively one substitute simply could be the average of minused4, so we take the average of minused4 and we substitute that in place of 1133 instead of deleting the value. Now that is a substitute but it is not necessarily the best substitute.

A better substitute would to do what is called a similar case substitution. In that case what we would do is, we try and substitute this data set and find customers who are on nights and weekend plans, find customers whose average usage in first month was between say 400 and 500 in the second month between 800 and 850 in the third month between 800 and 900, essentially try to find customers that are as similar as possible to this particular customer and on that smaller subset take an average. So instead of taking a global average substitution, we are doing a similar case mean substitution.

Now remember outlier treatment, one of the things that impacts outliers treatment is the effectiveness and the impact, there is a cost versus benefit analysis here.

If you have lots of data, it may not be worth your time to spend lot of time coming out with substitutes for outlying values because it may simply be easier to delete the outlying values without large cost to your analysis. Because you have lots of data.

However if you have very little data, then obviously you will want to spend lot more time on outlier treatment and doing substitutions rather than losing the data.

Slide-16

There are other outlier treatment options for example data transformation. Sometimes taking a log will reduce the spread in the data, if you have variables with positive values then taking the log will reduce the spread and sometimes that may be an appropriate outlier treatment methodology.

Sometimes it may simply be easier to ignore outliers.

What are the implications of ignoring the outliers?

You know you can run your analysis with or without outliers, see if there is any material difference in the analysis outcome. Sometimes you may also want to generate statistics that are robust to the presence of outliers. For example, if you are running a regression model there is something called a **trim leap squared value** which is essentially a statistic that is robust to the presence of outliers.

Remember for outlier treatment and identification, you have to understand the cause of the outliers in order to arrive at the best method of dealing with them.

Slide-17

Just to do a quick recap, in data you will see the variation and you will see some extreme values. When we run an analysis, we want to run an analysis on a representative set of data. When we say representative, we wanted to behave like our customers behave in real life, like transactions behave in real life. Which means that there will be variations, some customers with high spend as an example and some customers with very low spend as an example.

What we don't want to do is reduce the data set to a generic data set in which case our analysis will only be good for a very very narrow set of customers.

We want our analysis to be applicable to as wide a set of customers as possible. However we don't want to

include very very unusual values in our data because they have an outsize impact on the analysis. So if a value is truly an outlier, we need to treat the outlier, sometimes by deleting the value, sometimes by substituting that value and sometimes by doing transformations.