Slide 1

In the previous topic, we had looked at what is data exploration and started with the first question, which is how do I understand the information contained in the data.

Slide 2

In this topic, we will take a look at how to assess the quality of this information. In the last session, we had looked at, what is data exploration, why we do data exploration, and what sort of questions do we answer with data exploration. We spent time on understanding, what is the information contained in the dataset in the whole.

What we will do in this session is spend lot more time in assessing the quality of the data that is contained in your data set.

Slide 3

A very important thing that we should understand at this stage is that data integrity and useful cannot be assumed. What that means is that given a dataset, we should not jump to the conclusion that the data is perfect, that the data is okay, that the data was extracted correctly. You must spend time in checking the data for data integrity and for reliability and usefulness.

Do you have all the data that you need to run your analysis and can you be sure that the data you have is of good quality. In other words, at a high level, what we are doing in data exploration is some basic sanity checks - do I see what I expect to see in the data? Of course, the next question is,

- Should I always see what I expect to see?
- Do I already know what is in the data?
- What happens if I see something unusual?
- If I see something unusual in the data, is that noise? Is that anomaly?
- How do I assess whether something is noise or is a problem?

Again, domain knowledge plays a very important part in assessing the quality of the data.

Slide 4

Remember when we are talking about the quality of the data, the questions that we specifically want to answer to gain an understanding of the dataset:

- How was the dataset collected?
- Is it the universe?
- Is it the census data?
- Was the data collected actively or passively?

- What are values of each variable?
- Do we know what values in each of those variables?
- Do we understand those values?
- Do we see a lot of missing information or missing data?
- Do we see a lot of unexpected values or anomalies?
- Is the data enough for our analysis? Do we need more variables?

Slide 5

Remember in order to run a good project, in order to perform a good analysis; you need to have good data. If we do not spend enough time assessing the quality of the data, we may generate an analysis, which may not generate any meaningful results.

So making sure that you do a thorough review of the data, it saves time towards the analysis in the end because we do not have to then come back and check whether there is a problem with our data when we see unexpected results in our analysis.

Slide 6

Let us actually look at a framework for assessing the quality of the data. There are 6

steps that we could execute in the data exploration.

1. Transform qualitative data into quantitative data.
2. Generate derived variables.
3. Generate some summary statistics.
4. Create some cross tables.
5. Maybe some graphical analysis and visualization.
6. Anomaly detection.

We will take a look at each of these steps in detail.

Slide 7

Let us start with transforming qualitative data to quantitative data, and we are going to use the telecom dataset to illustrate examples of each of these steps.

Slide 8

If we remember the qualitative data in the telecom case study, there was one variable called Plan Type. Plan Type was qualitative data because it had values like 200 for 10, Nights and Weekends, and Coast to Coast. These are essentially names of the service plans that customers have chosen.

Now if I wanted to use the Plan Type data in my analysis, I may not be able to directly use

text variables in a mathematical model. So what I will need to do is to essentially transform this qualitative data to quantitative data.

An easy way to do that is to simply recode, so where the data says 200 for 10; we use 1, where the data says Nights and Weekends; we use 2, and where the data says Coast to Coast; we use 3, so simply recoding qualitative data into quantitative data.

Slide 9

Again, you may not want to do this for every kind of qualitative data. For example, supposing we had name. Now we may not use name in the analysis in any case. Supposing I had address, we may or may not use address directly in my analysis. I certainly wouldn't convert that to a quantitative variable.

It is possible that we may take some information from the address data for an analysis. For example, maybe from address; we take city, maybe from address; we take zip code, if that is not separately available. However, the address itself is not something that is going to make it directly into the analysis.

So my final dataset may still have some qualitative data that is descriptive in nature

that I am not going to convert to quantitative values.

The next step is to generate derived values.

Slide 10

What is a derived variable?
Simple example in the telecom case study, remember the objective of the analysis is to understand attrition. Why people are leaving the service. Now if you think about the variables that we had listed, was there a variable that captured attrition directly? Actually, not.

But since it is my target variable, I will need a variable for attrition in the analysis. How do I get that? The information around attrition is actually already contained in the dataset via the service end date. Remember, if the service end data has date in it, it means the customer has left. If the service end date is blank, it means the customer is still an active customer.

Slide 11

If you take a look at this data, you can see that the first five customers are the customers that are quit because there is a service end date. What about customer 6, there is service start date 19th October, but there is no service

end date, which means that this is a customer that has not quit.

Therefore, I can derive an attrition variable from this variable. All I need to do is simply say where service end date is not missing, it will be… so let us say this is attrition and we will say this is 1, 1, 1, 1, 1, and 0. In this case, 1 is a customer that has left or attrited and 0 is a customer that is still a current customer not attrited. So attrition is a derived variable.

Slide 12

There are other examples of derived variables in datasets. A common example is something like a birth date. Again we may get birth date in a dataset for a customer, but birth date is not something that I could use directly. What I will do is to transform that or derived that into maybe an age variable. How I get age from birth date, I simply say minus the birthday that will give me the age variable. These are very straight-forward derived variables.

Sometimes the derived variables may need little additional work. For example, let us say that we have usage data and we want to use the usage data and bucket them into high user customers, medium user customers, and low user customers. How do we arrive at the cut-offs for high, medium and low.

Again the usage bucketing is a derived variable because we are deriving that from the usage data itself. However, in order to bucket customers into High Medium Low, I may need little more information. I may need some domain knowledge.

The company itself may have some idea of how it classifies high users versus medium users or I may take a look at the example usage and use some judgement to come up with user buckets. Again remember these user buckets are derived variables, but slightly more complex than something like an age.

Slide 13

The third step in that frame work was to generate summary statistics. Once we can transform qualitative data to quantitative where applicable, once we derive the variables that we need for information that is contained in the dataset, the third step is to generate summary statistics, and this is very much important and critical part of data exploration process.

What are summary statistics?
Essentially they are summarizing information about data contained in our dataset, and generating summary statistics gives us a very high level understanding of the data that is

included in your dataset, but it also helps us to identify if we have missing value issues, if we have potential anomalies or outliers in our data, and what are the range of values in my dataset.

Slide 14

Remember summary statistics can be used to assess how complete is your data, is there a lot of missing values, and do you have outliers in your data.

Why do we use summary statistics? Remember again for most analysis, you are dealing with very large data sets, which means that looking at the data may not be very easy to do when you have large volumes of data. So what we are doing is essentially an efficiency outcome - summarizing information about the large volumes of data using descriptive statistics.

Which descriptive statistics or summary statistics can we use? Most often, we will generate the minimum, the maximum, the mean, sometimes standard deviations, the number of missing observations, the median, the mode, all of these are useful summary statistics.

Slide 15

Remember summary statistics do two things. They generate an understanding of the big picture in the dataset, what is it that we are dealing with in the dataset. However, they are also useful for basic sanity checks.

In the telecom dataset, there are some basic sanity checks that we should apply to the data. For example,

- What is the earlier start date?
- What is the latest end date?
- If the customer has quit, is there usage data post the customer quitting?
- What is the maximum monthly usage in minutes?
- What is the minimum monthly usage in minutes?
- What percent of customers in the dataset have quit?

All of these are basic sanity checks that we should do against the data.

Slide 16

So what we are going to do now is to take a look at very simple descriptive statistics table for the quantitative variables in our dataset, which is the telecom dataset, and we will just take a look at the descriptive statistics to gain a big picture understanding about the data.

So you can see here that summary statistics have been generated for quantitative variables in our dataset. Plan type was qualitative; we have not changed it to quantitative yet, so we have not included in the table.

Now if you look at this table there are a couple of things that standout. For example, the first thing that most of us notice is that in the minutes used 2 variable; there is a negative number in the minimum -55. Now most of us understand that usage data can only be positive, it can be 0; you did not use your cell phone or you used it, in which case it will be a positive number. So it does not make sense to see a negative number in the usage data.

Now at this point, remember all we know is that there is atleast one negative number in minutes used 2. This is the minimum. We do not know, how many negative numbers there are. If there are lots of negative numbers, we may have a big problem, if there is only one negative number, we may conclude that there is a problem, but it is an isolated problem, also we may want to confirm what it is mean to see a negative number in a usage variable. It is possible that this is a credit and that your databases when people make entries are allowed to show credits as negative numbers.

Again, what else can we look at when we look at the descriptive a statistics table like this? Remember, you want to look at each of these statistics for each variable and convince yourself that you understand that what is going on with that variable.

So minutes used 3 for example has a minimum of 0 and maximum of 1,500, it seems pretty okay. But look at minutes used 4, it has minimum of 0, but the maximum of 177,700. Again something seems seriously wrong here. We have atleast one value that is very much high relative to the other values in that variable. Remember this is just the maximum. We do not know yet how many high values we have.

So we will want to take a look at what are the values in minutes used 4 near the maximum. Similarly, if you take a look at the zip code variable, now the zip code variable is a nominal variable. Remember that zip code does not really have any ordering. So it may not make sense to talk about a minimum, a maximum, a mean, a standard deviation for the zip code as well as the subscriber id variable.

However, when I look at the zip code summary statistics one of the thing that I notice is that the length of the zip code variable is different. I have atleast 1 three digit zip

code. Now again, in the US all zip code have to have five digits, so this may alert me to a potential problem that I have some zip that may have partial information.

What about the promotion variables. These are 1, 0 variables. Now again, does it make sense to talk about a minimum, a maximum, a standard deviation of 1, 0 variables, from the point of assessing data quality, yes. You should expect the promotion variables to have minimum 0 and maximum of 1, which seems to be the case, so we know that there is no problem with the values in the promotion variable.

What about the mean. The mean here actually contains information in participation rate. If I look at promotion variable promotion 2 and its mean is 0.36, it tells me that 36% of customers had a value of 1 and the remaining customers had a value of 0, and I can tell from the averages of the promotion variable that participation in promotion is reducing month on month. Month 2 had 36%, month 3 had 26%, month 4 had 24%, and month 5 had 12%.

So again, all we are doing is we are taking a look at the very much high level summary statistics of the quantitative variables in our dataset and we are using this to gain very basic understanding of what data is available to us in our dataset.

So some of the questions or initial findings from just looking at this table include:

1. Why do we have negative values for the minutes used for the first month?
2. Why do we have very very high monthly minutes used for the fourth month?
3. Should all my zip codes have a standard length?
4. If yes, why do we have atleast 1 three digit zip code? Is that loss of information?

Slide 17

So these are some initial findings from our data exploration. Of course, we may want to do a lot more investigation. For example, negative value of 55 in minutes used 1. Now what I might do is take a look at all the low values in that particular variable.

Remember, the minimum is only one number, but if I look at the bottom 5 values of that variable, I can see that there is only one negative number in the dataset. We can see that the lowest value is -55 but the next four lowest values are all 0. So it may be easy for me to conclude that this is a potential data entry error or problem value and for my analysis, I may want to exclude this value from my analysis.

Again same thing for the very very high value in minutes used 4. I have a 177,700. If I look at the top 5 highest values, we have 1,389, 1,500, 1,500, 1,500, and 177,000. So again this is potentially an error and an instance that I may want to exclude from my analysis because this is just an anomaly.

What about the zip code. If I look at the number of observations that I have three digit zip code, four digit zip code, and five digit zip code, I can see that almost 8% of my data has less than five digits in my zip code, and now this is not something that I could ignore because 8% of data is lot of data.

What should I do in this case? I may have to go back to the data extraction process and make sure that the zip code information is being pulled correctly or extracted correctly. If it is being extracted correctly, it is possible that for 8% of the variables, we just do not have good zip code data.

Slide 18

What do we do when we have qualitative variables? Remember we could convert qualitative variables to quantitative. The other way to do data exploration or summary statistics for qualitative variables is to do a frequency distribution. For example, plan

type. Plan type has three kinds. 200 for 10, Coast to Coast, and Nights and Weekends.

If I look at frequency distribution of this particular qualitative variable, I can see that the most popular plan type is 200 for 10. The least popular is Coast to Coast.

Similarly, New Cell Indicator, a qualitative variable, it have values Yes, No, and U. Now, if new cell indicator is essentially indicating whether or not a new cell phone was bundled with the service plan, we should expect to see yes and no. If you look at the data, however, and the frequency distribution of that variable, you will see that almost 80% of data actually has a value of U, which is unknown.

So this may be variable where even though there are lot of values in that variable most of them are unknown, which means that I may not be able to use this variable in any meaningful fashion for my analysis.

Slide 19

What about derived variables. Of course, you also want to take a look at summary statistics for derived variables. For example, the attrition variable, which is whether or not customer has left. Remember in data exploration, part of the reason for doing data exploration is to look at the values of the

variable and make sure that we do not have any values that we do not understand, but it also to get a big picture of understanding of the data.

If I look at a frequency of the distribution of attrition variable, which is retained versus lost, I can see that we have 40% attrition in the dataset. Now we need to make sure that we are comfortable with this 40%. It seems like a large number.

However, remember the objective of the analysis was to understand attrition. So clearly there is a problem with attrition in general and you will expect to see high values of attrition in your sample dataset.

Of course, how high is high. Supposing we had seen the attrition of 90% then definitely there may have been problem with the data extraction. If we had seen attrition to be only 5%, again maybe there is a problem with the data extraction. So these are things that you have to evaluate when you are looking at summary statistics or frequency distribution.

Another example of a derived variable is age. Supposing we were looking at age and we bucketed age in the following manner less than <25 years old, 25-35, 35-45, 45-55, and >55. If you look at the distribution of your age group, we have almost equal representation

from 25 all the way to 55 and above, again no skew by age. Now as an analyst, you want to make sure that you are comfortable with this conclusion that your data does not show skew by age.

Is this what you are expecting? Is this contrary to our expectations? If we think about subscribers to a cell phone service, we would expect a higher percentage of users to be from the younger age groups relative to older age groups. Now the data does not confirm to expectations in this case.

Therefore, this may be a question that is worth following up on. Can we be sure that the data was extracted correctly because the data does not show a distribution, the skewness by age group when we were expecting skewness by age group.

Slide 20

So what we were looking at in this data exploration process was the first three steps.

Transform qualitative data into quantitative data, generate derived variables, and generate summary statistics.

In the next topic, we will take a look at the next three steps, which is cross tables, graphical analysis, and anomaly detection.