Data Cleaning

Dr.P.Thiyagarajan
Assistant Professor
Department of Computer Science
Central University of Tamil Nadu
Thiruvarur – 610 005.

What is Data Cleaning?

- Data cleansing or data cleaning is the process of identifying and removing (or correcting) inaccurate records from a dataset, table, or database and refers to recognising unfinished, unreliable, inaccurate or non-relevant parts of the data and then restoring, remodelling, or removing the dirty or crude data.
- Data cleaning techniques may be performed as batch processing through scripting or interactively with data cleansing tools.
- After cleaning, a dataset should be uniform with other related datasets in the operation. The discrepancies identified or eliminated may have been basically caused by user entry mistakes, by corruption in storage or transmission, or by various data dictionary descriptions of similar items in various stores.

Why we need Data cleaning?

- Data Cleaning techniques are not only an essential part of the data science process – it's also the most time-consuming part. As the New York Times reported in a 2014 article called "For Big-Data Scientists, 'Jaitor Work' Is Key Hurdle to Insights"

"Data scientists … spend from 50 per cent to 80 per cent of their time mired in this more mundane labour of collecting and preparing unruly digital data, before it can be explored for useful nuggets"

Why we need Data cleaning?

- Unfortunately, Data Cleaning techniques are generally not spoken about in the media nor is it taught in most intro Data Scientist Course because it is not as important as training a model but to perform those things data cleaning plays a very important role.
- Without the data cleaning techniques, the model will not be as efficient as we want them to be.

Why we need Data Cleaning?

As you continue working with data, though, you'll find that much of the data you encounter will be messy, and lots of your time will be spent preparing it for analysis. These preparations, known as **data cleaning**, include:

- Removing data you don't need for your analysis.
- Removing duplicate data.
- Dealing with missing data and outliers.
- Creating new variables where necessary.
- Combining separate datasets.

Data Set

New York City Schools Data

- The datasets you'll be working with for this topic come from the New York City Department of Education. They contain data on NYC schools, including student demographics, test scores, graduation and dropout rates, and school locations.
- sat_results
- ap_2010
- class_size
- demographics
- graduation
- hs_directory

Which data do we need for our analysis?

- Often, you won't need all the data contained in datasets to answer your research questions, and it may make sense to create a new data frame containing only the necessary data. This will help you work more efficiently when you're working with large datasets.

- For example, we are interested in data for NYC high schools. However, the class_size and demographics data frames also contain information about elementary schools. Some of the data frames also contain observations from multiple years, and it may make sense to work with only the most recent years' data.

Do we need to create any new variables?

- Sometimes, your analysis will require a variable that is not currently defined in your data. For example, the sat_results data frame contains variables for students' scores on sections of the SAT: Math, Reading, and Writing.
- However, for our analysis, we may want a total SAT score variable. We would need to create it by calculating the sum of the section scores.

Are the data of the correct type?

- Are any of the data that you need to use to calculate new variables, for example, formatted as character instead of numeric?

Do we need to combine data frames?

- To analyze the NYC high schools data, we ultimately want to be able to perform summary calculations and visualize relationships between variables to understand how demographic factors affect test performance. We will need to combine the six data frames into a single, clean data frame.

- We can combine multiple data frames if they share a variable in common, and if that variable uniquely identifies observations. We call this variable a key.

- In the case of these data frames, we see that nearly all of them share a variable in common: the DBN, or district borough number, that uniquely identifies each school.

Do the data need to be reshaped?

- When you learned to visualize data, you saw that different arrangements of variables in rows and columns were needed for different tasks. You'll learn more about this concept and tools for reshaping data later in this course.

Are there missing data?

- In some data frames like demographics, there are many missing values (represented by NA). We will have to decide how to handle these missing values as we clean the data. Later in this course, you'll learn techniques for working with missing data and their implications for your analysis.

- In this mission, we'll guide you through the cleaning operations needed for each data frame. In the next mission, you'll learn about relational data and combining data frames. In later missions, as you begin analyzing the data, you'll learn techniques for reshaping data and handling missing values.

- Take a moment to preview the data frames and think about what data cleaning operations may be necessary.

THANK YOU !

Acknowledgement: Various Internet resources have been used to make this presentation and I acknowledge the same.