DATA SCIENCE EAST AFRICA

INTRODUCTION TO DATA CLEANING.

Day 5/20

Introduction to Data Cleaning. (Part 1)

Data cleaning can also be referred to as data cleansing, data washing or data scrubbing. It can be defined as a process or technique of removing corrupt data. A data can be a record from a table, record set or database.

- **Data**

In statistics, data can be defined as facts or figures from which conclusions can be drawn. The singular of data is datum but this is rarely used because data is uncountable.

- **Table**

A table is the organization of data in rows and columns. It can also be referred to as a spreadsheet or datasheet. It is made up of records and fields.

If you are fond of confusing a “row” with a ”column”, the popular children song can help you: Row, row, row your boat. Gently down the stream. Merrily, merrily, merrily, merrily. Life is but a dream. You can row your boat from left to right not top to bottom.

Example:

USER ID	FIRST NAME	LAST NAME	AGE
1	DELE	GIWA	23
2	JAMES	COOK	40
3	SUSAN	PAIGE	34
4	OSENI	KUNLE	43

- **Record**

A record can be referred to as the row (x-axis). It contains all the fields present in the table. Considering the table above, the first record present is: “1, DELE, GIWA, 23” and the last record present is: “4, Oseni, Kunle,43”

- **Field**

A field can be referred to as the column (y-axis). It contains a single data entry in a record. Considering the table above, the AGE (field) for the first record is 23 (data) and the LAST NAME (field) for the last record is Kunle (data).

– **Record Set**

A set of record is a collection of records and it is referred to as a table or spreadsheet.

– **Database**

A database is a data structure that contains a collection of data. This data are stored as multiple tables.

Benefits of Data Cleaning.

Data cleaning is the most important as well as the most time consuming part of data science because of the numerous data generated on a daily basis. It has then become very essential to examine the generated data with a tool and fix the errors present in the data because it is no longer possible to skim through the large data pool.

Errors that can be present in the data includes:

- Irrelevant observations
- Unwanted outliers
- Missing data

– **Irrelevant observations**

Irrelevant observations may be a duplicate observation. This types of data are generated a lot. For example: If I consider my browsing history, there are some sites that I've checked more than once and I just need the list of websites that I have visited.

It’s imperative I clean the duplicate websites in my browsing history before performing other operations on my data. Irrelevant observations also can be observations that are not needed in the analysis.

– **Unwanted outliers**

An outlier is something different from others. An example of an unwanted outlier when calculating the mean wealth status in Kenya is chris Kirubi. Chris Kirubi’s worth is over $8 billion. So if we are to add his worth to the mean wealth status in Kenya, we will get a wrong result because less than 0.03% of the population in Kenya has such kind of worth. It’s better to clean such outliers in our data before performing any kind of operation.

– **Missing data**

In our tables, some important fields might be empty for a record. It is imperative that we either remove such records from our dataset before performing other operations on the data so we won’t get a skewed result.

Exercise:

1). Common data problems

- Research and read more on how you can overcome some of the most common dirty data problems.

- Research and read more on how you can convert data types, apply range constraints to remove future data points, and remove duplicated data points to avoid double-counting.

2). Text and categorical data problems

- Research and read more on how to fix whitespace and capitalization inconsistencies in category labels, collapse multiple categories into one, and reformat strings for consistency.

3). Record linkage

- What is record linkage and when do we use record linkage ?

- Research and read more on how to link records by calculating the similarity between strings, and use the skills to join two datasets into one clean master dataset.

All the best.

Regards Data Science East Africa.