LESSON HANDOUT

Cleaning data

Cleaning, sometimes referred to as tidying, is a crucial element of the transform stage where we take raw, messy data and transform it into a format where it can be used for analysis. Cleaning data can be a technical part of the Data Analytics process, requiring programming skills (SQL) and advanced use of software tools.

Messy data

Messy data is data that has inconsistencies, missing values or errors which require fixing before it can be used. Messy data can occur for many reasons, but here are some key examples:

- **User input or error.** Often the data is the result of human error at the stage where data is inputted. This could be as simple as a spelling mistake, which is why mandatory and drop down fields are important for any user data collection system. It's far easier to fix these issues at the source than to clean these issues later on.
- **Multiple sources/systems.** Where we are combining data from multiple systems or reports, additional cleaning is required to standardise or normalise the data to ensure we are combining 'like for like'.
- **Machine data.** Cleaning is particularly taxing when the data that we are dealing with has been generated primarily for machine use. A large amount of data is generated by machines and for machines such as **metadata**: data which describes other data or **exhaust data**: the trail of data left by users. This data is normally used by machines to help user fetch or search data, or to enable machines to optimise processes. Increasingly this data is being harnessed for analysis, however, the obstacle is a high cleaning burden.

Data collection

As a Data Analyst, it is important that you influence data collection techniques which you find are sub-optimal: better collection leads to cleaner data which informs better analysis.

Common cleaning requirements

Common tasks performed at the cleaning stage are as follows:

- Dealing with missing or incorrect values
- Removing irrelevant data
- Transforming data to the correct format (text, number, date)
- Standardising categories - fixing spelling and abbreviations

- Removing duplicate rows
- Combining multiple rows into a single record

Clean data

What does **clean data** look like?

- A clean data table has one row per instance and one column per variable.
- Columns are clearly named and all data within columns will be the same format.
- Any missing data has been replaced, marked as missing or removed.

Row ID	Order ID	Date	State	Postal Code	Sales	Quantity	Discount	Profit
1	CA-2016-152156	1/12/2014	Kentucky	42420	261.96	2	0%	41.9136
2	CA-2016-152156	1/12/2014	Kentucky	42420	731.94	3	0%	219.582
3	CA-2016-138688	1/12/2014	California	90036	14.62	2	-	6.8714
4	US-2015-108966	1/12/2014	Florida	33311	957.5775	5	45%	-383.031
5	US-2015-108966	1/12/2014	Florida	33311	22.368	2	20%	2.5164
6	CA-2014-115812	1/12/2014	California	90032	48.86	7	-	14.1694
7	CA-2014-115812	1/12/2014	California	90032	7.28	4	0%	1.9656
8	CA-2014-115812	1/12/2014	California	90032	907.152	6	20%	90.7152
9	CA-2014-115812	2/12/2014	California	90032	18.504	3	20%	5.7825
10	CA-2014-115812	2/12/2014	California	90032	114.9	5	0%	34.47
11	CA-2014-115812	2/12/2014	Null	Null	1706.184	9	20%	85.3092

Figure 1 - Example of a clean data table

The Data Dictionary

The **data dictionary** is a description of all the data elements in a table and is used to describe the type and origin of the data to the outside user. We should always include a data dictionary with our analysis.

Data Dictionary Example

Table	Variable	Data Type	Description	Remarks
Sales	Row ID	Text	Table index	
Sales	Order ID	Text	Unique ID no	
Sales	Date	Date	Date order received	
Sales	State	Text	Delivery state	
Sales	Postal Code	Postal Code	Delivery postal code	Extracted and cleaned from Order table
Sales	Sales	Numerical	Total value of sale	
Sales	Quantity	Numerical	Quantity of items	
Sales	Discount	Numerical	% discount granted	Calculated from Discount field in Discount table
Sales	Profit	Numerical	Total value of profit	

Figure 2 0- Example of a data dictionary