COURSE04: Process Data from Dirty to Clean!

Instructor: Sally — Measurement & Analytical Lead

Page No.
Date 20 01 23

*Introduction:→

Sample Size
Random Sampling
Testing data
Data cleaning Technique for spreadsheets & Databases
Verify & report your cleaning results.
Cleaning could be the highlight of your resume!

Course Contents:-
1. Ensuring data integrity
2. Understanding clean data
3. Cleaning data using SQL
4. Verifying & Reporting cleaning results
5. Adding data to your Resume (optional)
6. Challenge

Data Integrity:-
A strong analysis depends on integrity of data.

Defn: The accuracy, completeness, consistency, and trustworthiness of data throughout its lifecycle.

Low data integrity can cause anywhere from loss of a single pixel in a image to incorrect medical decision.

Good News * DW or DE team takes care of data Integrity

* Data integrity can be compromised in lots of different way:-
  ① Replicated
  ② Transfered
  ③ manipulated

* Data Replication:-
  Process of storing data in multiple locations
  ↓
  Theres a chance that data will be out of sync
  ↓
  means that different people won't be using the same data for their findings
  ↓
  leading to inconsisties.

* Data Transfer
  Process of copying data from storage device to memory or from one computer to another.
  (interruptions causes incomplete data)

* Data manipulation:-
  The process of changing data to make it more organized and easier to read.

* Other threats to data integrity:-
  - Human Error
  - Hacking
  - Viruses
  - System Failure
  - malware

VLOOKUP (value, array, index_col, T/F) ⟶ Corresponding value of index col

DATEDIF (start, end, unit) ⟶ Returns number of d/m/y

"m" "y" "d"

A slight coverage on VLOOKUP() & DATEDIF() & ~~DAYS360~~

21/01/23

Week 3/5

Dealing with insufficient data!

Challenges are bound to come out but once you know your business objective you'll be able to recognize whether you have enough data.

↓

If you don't you'll be able to deal with it before starting your analysis.

*Types of insufficient data:→

- > Data from only one source
- > Data that keep updating (not complete)
- > Outdated data
- > Geographically limited data

* Ways to address insufficient data:→

- > Identify trends with the available data
- > Wait for more data if time allows
- > Talk with stakeholders and adjust your objective
- > Look for a new data

* The need to take these steps will depend on your role in the company and needs of the wider industry!

* Importance of size :-

Population :- All possible data values in a certain dataset

100 % population use for analysis is ideal
↓
Its almost impossible to calculate information about entire population.
↓
Too time consuming / expensive.

Ex. See which toy cat owners in California prefer
→ Impossible
Now-Now ; SAMPLE SIZE

Sample Size :
A part of population that is representative of the population
↓
The goal is to get enough information from small group within a population to make predictions or conclusions about the whole population.

It Also represents confidence : degree to which you can be confident that your conclusions accurately represent the population.

* Creating sample size takes place before you even get to the data.

Page No.
Date

Downside :→ Small sample leads to uncertainty. You're not sure 100%. if your statistics are complete & accurate. representation of population.

* This is called Sample Bias

Random sampling:
A way of selecting a sample from a population so that every possible type of sample has an equal chance of being chosen.

Population :→ Entire group

Sample :→ Subset of population

Margin of Error :→ Difference between sample's results & Entire population's result.

Confidence Level :→ How confident are you in your survey

Confidence Interval :→ Range of values (popul^n's result) with certain Confidence level

Statistical Significance :→ Determines if your result is random chance or not (the less, the less due to chance)

* Don't use sample size less than 30
• CI is 95% most common, 90% also can be used.

why minimum size is 30 ↦

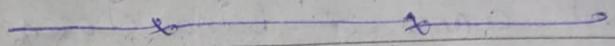
30 is the least value for which Central Limit Theorem holds

For Higher CL, use larger sample size

To decrease Margin of Error use ~~delarger~~ sample size

for Greater Statistical significance, use larger sample size.

+ Larger sample size do have higher cost



Week 1
4/5

22/01/23

Statistical Power:-
The probability of getting meaningful results from a test.

Hypothesis testing:
A way to see if a survey or experiment has meaningful results.

There are ways to accurately calculate Statistical Power

↓
Its calculated out of 1
0.6 ⟶ 60 %.

Statistical power means

↓

we say its 60 % statistically significant

Statistical Significance:-
If a test is statistically significant, it means the results of the test are real and not an error caused by random chance.

- Usually, you need a statistical power of atleast 0.8 or 80% to consider your results statistically significant.

Thinking like an analyst:-
We have a chain and want to test a same location the Birthday cake flavoured milkshake,

first :-> Think about what might prevent you from getting statistically significant results.

(1) Are there restraints running any promotions that might bring in new customers

(2) Do restraint have customers that buy newest ite regardless of what it is?

(4) Do some restraints have construction around/on them preventing customers going in?

To get higher statistical power, we'd have to consider all these before we include any locations in our study!

We wanna make sure any effect is due to milkshake & not other measurable effect: sales or customer at locations.

Page No.
Date

Statistical significance:→

If a test is statistically significant, it means the results of the test are real and not an error caused by random chance.

- Usually, you need a statistical power of atleast 0.8 or 80% to consider your results statistically significant.

*** Thinking like an analyst:→

We have a chain and want to test a some location the Birthday cake flavoured milkshake,

Test:→ ① Think about what might prevent you from getting statistically significant results.

② Are there restraints running any promotions that might bring in new customers

③ Do restraint have customers that buy nearest item regardless of what it is?

④ Do some restraints have constructions around/on them preventing customers going in?

To get higher statistical power, we'd have to consider all these before we include any locations in our study!

*** we wanna make sure any effect is due to milkshake & not other

*** measurable effect: sales or customer at location.

week 7
5/5

Page No.
Date 23 01 23

Margin of Error

As a DA we should always calculate variables like CL & MoE before running any kind of test or survey!
↓
Its the best way to make sure our results are objective and it gives us better chance of getting statistically significant results.

Defn: The maximum amount that the samples results are expected to differ from those of the actual population

> It tells us how reliable our data from Hypothesis testing is.

> The close to zero MoE, the closer our results to results from overall population.

Ex. survey nationwide 5 day vs 4 day (workweek)

60 % prefer 4 day workweek
with MoE = 10%.

i.e. 50% ↔ 70 % might like the idea

If we actually make a nationwide survey 50-70 % people would like this idea!
"Yes"

Since our MoE overlaps 50% mark, we can't say for sure that public like 4 day workweek INCONCLUSIVE

The more sample size the more the accuracy

Page No.
Date

When you want to lower MoE
You have to increase the Sample Size.

* To calculate MoE we need:
  - Popln size
  - Sample size
  - Confidence Level

Ex. Effectiveness of drug

Sample size = 500
Population = 80,000,000
Confidence level = 99%
] → 5.77

5.77 close to 6%

Calculators like these could be used to ensure data integrity.

Its good to ~~check~~ check for data integrity and aligning data with our objectives
↓
This puts you in good shape to complete your analysis

Week 2 | 1/3

Sparkling Clean Data

Data cleaning is a must

Page No.
Date

* Clean it Up!

Yearly cost of poor-quality data is $3.1T in US alone.

#1 Cause: HUMAN ERROR

* "Pasta & Cashew Nuts have similar tastes."

Dirty data can be:→
* Incorrect data (Spelling mistake)
* Inconsistent formatting (% in $ field)
* Blank fields
* Duplicates

Defn: Dirty Data is incomplete, incorrect or irrelevant to the problem you're trying to solve.

Clean data is complete, correct and relevant to the problem you're trying to solve.

Compares Dirty Cleaning Data with Brushing teeth

* Internal data is vaulted & cared for by your company's data engineers and data warehouse team.
↓
Its more likely to be clean.

* Data Engineers:
Transform data into a useful format for analysis and give it a reliable infrastructure

↓

This means, they develop, maintain and test databases, data processors & related systems.

* Data Warehousing specialists:
Develop processes and procedures to effectively store and organize data.

↓

They make sure data is available, secure & backed up to prevent losses.

If data passes through Data Engineers or Data Warehousing specialists, you know you're off to a good start.

Even if this happens you're still likely to clean your own data

NO DATASET is perfect, so its always a great way to examine your data before beginning analysis.

Data cleaning becomes more important when working with external data.

Page No.
Date

Especially which comes from Multiple Sources.

Null: Indication that a value doesn't exist in a dataset.

What to do with Nulls?

- Filter out the dataset → smaller sample size
- Keep them in & learn the fact that, customers didn't provide any response.

Reasons:→ (1) Questions weren't well written
(2) Confusing / Biased questions

Angres Program Manager of Engineering

"Cleaning Data is the heart and soul of Data Analysis"

1 Million $ worth burger, cause: getting multiplied.

Recognize & Remedy Dirty Data:

* Different types of currencies don't get them mixed up.
* we want to find errors & fix them.

* Field :
Single piece of information from a row or column of spreadsheet

* Field length:
A tool for determining how many characters can be keyed into a field.

↓

Assigning fields with field length is a great way to prevent errors.

Ex. Birthyear → 4 characters only.

* Data Validation:
A tool for checking the accuracy and quality of data before adding or importing it.

Spreadsheets offer lots of great functions for removing spaces or blanks automatically.

Ex. Spell check, auto correct, conditional formatting

You can always use your hands to clean data (manually)

Week 2/3

Page No.
Date 25 01 23

Data Cleaning Tools & Techniques.

* How to remove unwanted data:
Clean up text to remove extra spaces an blanks, fix typos, and make formatting consistent.
↓
Before Removing Unwanted Data
Always make a copy of the original Dataset.

* Removing Duplicates
* Irrelevant Data (takes a little time since you have to figure out the difference betn data you need and data you don't.
But little time invested here saves tons on time down the road.

* Removing Extra spaces & blanks.
↓
Cause errors while sorting filtering or searching through data.
* Fixing misspellings
* Inconsistent Capitalization
* Incorrect punctuation and other typos.
* Removing formatting (data from different sources have different formatting)

* There's also a "Clear Formats" option in Spreadsheet applications

* Cleaning multiple datasets.

Cleaning data that comes from two or more sources is very common for DA!

Mergers:
An agreement that unites two organisations into a single new one.

They do so to sustain in the market.

Data Merging:
The process of combining two or more datasets into a single dataset.

This proposes a challange, when two totally unique/different datasets are combined the information is almost gauranteed to be inconsistent & misaligned.

* In the example, Global log. has separate column for suite/apartment/unit no.

whereas International log combines it with the street address.

* Also, Global log. uses mail ID as member ID

International log. uses numbers as member ID

Global uses "Yoyg Professional" & International logistic uses "Stident Associate" for Interns

So we need to get rid of these kinds of inconsistency

*Compatibility:
How well two or more datasets are able to work together.

Questions to ask:
> Do I have all the data I need?
> Does the data I need exists within these datasets?
> Do the datasets need me to clean or are they ready to use?
> Are the datasets cleaned to the same standard
↓
what fields are regularly updated
How are missing values handled
How recently does the data updated

*Programming languages like R are very useful for cleaning data.

Page No.
Date

Activity:

(1) Create filter in Gsheets/Excel
Check all the rows for missing cells & delete rows accordingly.

make sure there are no missing values.

(2) Transpose the dataset
long form ⟶ Wide form
~~more rows~~ more cols than rows
~~than cols~~

Copy the data you want to transpose
go to desired location
Right click > Paste Special > Transposed!

Now delete the original data

(3) Get rid of spaces (extra) in string data.
Data > Data cleanup > Trim whitespaces.

Gsheets

Excel ⟶ TRIM(A1) & then drag

(4) Change Text Upper / Lower / Proper

Tools Download Change Case Extension & use it

(5) Delete All formatting

Clear any or all cell's formatting by Highlight them > Format > Clear Formatting

Excel : Home > Clear > Clear Formats.

26/01/23

Week 2
3/3

We have seen how we can clean data manually by searching for & fixing misspellings or removing empty spaces and duplicates

* Also that spreadsheet applications have tools & processes that simplify data cleaning processes.
  - > Conditional Formatting
  - > Removing Duplicates
  - > Formatting date
  - > Fixing text strings & substrings
  - > Splitting text to columns

(*) Conditional Formatting:

Defn A spreadsheet tool that changes how cells appear when values meet specific conditions.

performed Conditional formatting on International logistics dataset.

↓

Select required columns.

> format > Conditional Formatting

> Decide the format rule

> and format style

> Hit Done

↓

we can select more than one range by doing A:E, G:K, I:L etc

(*) Remove Duplicates:-

first of all create a copy of the dataset

↓

Right click on the shit & hit duplicate

this file will have Conditional formatting

↓

"Basically a proper copy"

This dataset has a member listed twice for this, we go to Data > Remove Duplicates (select "has a header row" if) & Hit the button

Defn A tool that automatically searches for and eliminates duplicates entries from a spreadsheet

Page No.
Date

(*) Consistent formatting:

Ex. Dates : Sometimes dates aren't in standard format, this could lead to confusion while calculating dates

↓

So we need to have a consistent date formate

↓

Select the column > Format > Number > Date

Boom!

All the dates in same numerical format DDMMYYYY

(*) Text String:

A group of characters within a cell, most often composed of letters, numbers or both

↓

Important characteristic : Length

Substring : Smaller subset of a text string

ⓐ Split:

A tool that divides text around a specified character and puts each fragment into a new, separate cell.

↓

Helpful when you have more than one piece of data in a cell & you want to separate them out.

Ex. | Fname Lname | → | Fname | Lname |

Specified text separator = delimiter

How to: Highlight > Data > Split text to column (, default)

(num)

* A column has a text value so its not able to get multiplied

So we do text to column & it resolves!

Split

Now it converts all to Number

(*) Concatenate:

A function that joins multiple text strings into a single string.

* Optimizing data cleaning processes:-

Function:-

A set of instructions that performs a specific calculation using the data in a spreadsheet

(1) COUNTIF : A function that returns the number of cells that match a specified value.

Syntax: A predetermined structure that includes all required information and its proper placements.

=COUNTIF(range, "value") | COUNTIF(I2:I72, <

Page No.
Date

(2) LEN : A function that tells you the length of a text string by counting the number of characters it contains.

↓

for fixing a length on certain column, etc.

Ex 6-digit Identification ID, so all member should have 6 digit only i.e. Len = 6.

Syntax : LEN(range) Ex Len(A2) = 6

(3) LEFT / RIGHT

A function that gives you a set number of characters from the left/right side of a text string.

Syntax: LEFT (range, no. of characters) LEFT(A2,5)
RIGHT (range, no. of characters) RIGHT(A2,4)

MID : A function that gives you a segment from the middle of the string.

Syntax :→ MID (range, start, length)

(4) Concatenate : syntax : CONCATENATE (item1, item2)

(5) TRIM : A function that removes leading, trailing & repeated spaces in data. (Removes Extra spaces)

Syntax : = TRIM (range)