

# Course 4

## Overview

1. Data integrity
2. Clean Data for more accurate insights
3. Data cleaning with SQL
4. Verify / report cleaning results

## Module 1

**Data Integrity**: Consistency, accuracy, completeness

- Data replication: Chance of missing sync
- Data transformation: if interrupted, lack of completeness
- Data manipulation: Error during manipulation can compromise the efficiency.

\* With incomplete data, it's hard to see the big **Picture**

\* Always **Align Business objectives** and **data**

Deal with insufficient data: One source do not show the 'trend'

- Data that keeps updating
  - Outdated data
- ↳ multiple data source may contain diff insights

## Case:

1. No data: Draw small scale to perform preliminary analysis.  
Then, complete the analysis when you have enough data.
2. Little data: Do the analysis using proxy data along with actual data.
3. Wrong data: Identify errors, try correct them, if not ignore them



The importance of Sample size : Cost effective

- Possibly get some results as using entire size

Sampling bias : Sample isn't representative of data

Random sampling : Selecting every possible type of the sample

- Sample must not be less than '30'

Larger sample for, higher confidence level  
to decrease the margin of error  
greater statistical significance.

Module 2:

Dirty Data : Duplicate, incorrect, incomplete, inconsistent

inaccurate data can be costly

Issues : Spelling, Duplicates, Blank  
↓ misspell      ↑ User Input      ← null

Data cleaning is essential for decision making.

excel:

Vlookup  
split  
left  
right  
mid  
countif

- pivot table : Summarize table for presentation.

Checklist for data cleaning:

- Determine size of the dataset
- Determine the # of categories or labels
- Identify missing data
- Identify unformatted data
- Explore the different data types



## Module 3

SQL

vs

Excel

- Large data
- complicated analysis
- Extract in depth data

- Small Data
- simple analysis
- hard to extract in depth data

- Language can be used to interact with data

- Only built in functions available

- Data can be stored across DB

- Data is stored locally

Select

From

Where

Length( ) , Substr(col, 1, 2) , Trim(stute) = 'OH'

ABC = AB

↳ removes empty spaces  
so 'OH\_' return also

CAST(price As float64)  
CAST(date As date)

date Between '2020-12-01' And '2020-12-31'

Concat(A, B)

Coalesce (A, B)

return A  
return B if A is null

Order by

CAST(price As float64) Desc

Update

Table

Set

Condition

(OP) Where

Condition