Week 3: Cleeming with SQL.

Bizz: Healthcare, Advertising, E-commerce, Entertainment... everywhere data is used.

SQL: • Development begun in 1970.
- IBM first introduced - System R
- Then next version SQL.

- To find some specific information, in spreadsheet, you need to perform a lot of steps.

- With SQL it is generally done with single Statement.

Minor Differences...

Sheet	SQL
Small	• Large
Manual Entry	• Prepare data for further analysis
Create Visualization	
Built in spell check	• Fast
Solo project	• Collaborative work

"All data is stored in different places, maybe even in different formats... each location might have millions of rows and hundreds of tables".

Database: A computer program to store and process large amounts of data

Vendors of Database: MySQL, Postgresql, Oracle, SQL Server

SQL: A programming language used to talk to these databases...

Each vendor's product has its own SQL version called - SQL Dialect

- MongoDB, Cassandra → NoSQL.

- Duplicate Eliminated – DISTINCT

- Length – Consistency – String

- 1. Use trim
- 2. Remove duplicates
- 3. Locate Outliers
    - Length missmatch (Use sort - filter)
    - Not value in range (1 - 5 ...)
- 4. Fix Nulls.
- 5. Check spellings - Combine different spellings

- Use metadata to understand the range and eliminate outliers.

- Use unique values too much to get what fields are there

- WHERE col1 IS NULL;

↑ To check Nulls.

- AVG(COL)

↖ To get mean

- CAST()

CAST(COL1 AS FLOAT32)

- Use only Date when date require
  & Use only Time when time is ...

  Don't use DateTime together bro.

- Use BETWEEN ... AND ...

- CAST (COL1 AS date)

  ↑ Datatype

- CONCAT ( ) ↖ Used to join ...

COOL

- COALESCE ( ) ↖ like np.where

  Returns column's value if not null, otherwise will return other column's value.

  COALESCE (COL1, COL2)