

Week1: Organizing data in spreadsheet and SQL

GRIT!



Analysis

- the process used to make sense of the data collected.
- Goal: identify trends and relationships within data, so that you can accurately answer to questions you are asking.
- "* The 4 Phases of Analysis" → make it easier to search, understand, e.g. organize data into a table, hide / show column
 - ① Organize data → observing and organizing data in a way that's easy to reference.
 - ② Format & Adjust data → adjusting, make it more easy to digest like SORT & FILTER
 - finding outliers (data points that are very different from similarly collected data & might not be reliable values.)



- * Sort → arrange data into meaningful order (ASC, DESC), group similar data
- * Filter → seeing data that meets a specific criteria

- ③ Get input from others → checking others' analysis e.g. who experiences on this.
- ④ Transform data → identify patterns, calculations (e.g. correlation relationships between 2 variables),

Spreadsheet Workshop Commands →

- Sort
 - ✓ Sort sheet → all of the data in a spreadsheet is sorted by the ranking of a specific sorted column - data across rows is kept together.
 - ✓ Sort Range → Nothing else on the spreadsheet is rearranged besides the specified in a column.
- ✓ Sort Function → =SORT(A2:D6, sorted_by-column, asc/desc)
- ✓ Customized Sort Order → Sort data in a spreadsheet using multiple conditions:
on Data → Sort range > Advanced Range sorting ops.

Get to know
Sort
&
Order By

SQL Workshop Commands →

- ✓ WHERE clause → to filter only values that meet a specific criteria.
- ✓ SORT BY (ASC by default) → ASC or DESC

Week2: Convert and Format Data Combine multiple datasets

Incorrectly formatted data can:

- Lead to mistakes ✗
- Take time to fix 🕒 ...
- Affect stakeholder's decision-making 😬

Spreadsheet Workshop →

- ✓ Typecast numbers (Unit Conversion)
 - tab above (Format > Number)
 - =CONVERT(E2, "mph", "mls")
 - =CONVERT(E2, "C", "F")
(Convert cell E2 from Celsius to Fahrenheit)
- ✓ VALUE() convert text string to number.
- ✓ LEN, LEFT, RIGHT, FIND
 - Find position of text/string in a cell

v Data Validation (Module 4 Process)

In Spreadsheet:

- * Add dropdown lists with predefined options.
- Create custom checkboxes
- Protect structured data & formulas (Reject input) will protect from mistyping e.g. is valid
- * - Conditional Formatting highlight cells when matches the condition

SQL Workshop →

- * - CONCAT
 - CONCAT("google", ".com")
 - CONCAT()
 - CONCAT(=W1,W2, "google", ".com")
 - CONCAT_WSC()
 - Google + .com
 - CONCAT with +
- CASTC ~ AS int64)

Get support during analysis

- Encounter error when calculating time (start time is greater than end time)
change start-end to =IF(end>start, end-start, 1-start+end)
- ASK team
- ★ - Ask/find questions online
 - > Best practices for searching online
 - Thinking skills (mental model: your thought process & the way you approach a problem)
 - Data Analytics terms (use the right terms when searching) ⚡
 - Basic knowledge tools (can apply new with old knowledge) 📈

Week 3: Aggregate data for analysis
(VLOOKUP, JOINS, Subqueries)

- Aggregation → collecting or gathering many separate pieces into a whole.
→ organizing pieces of data by Average, min, Max, Sum

Helps:

- Identify trends ↗
- Make comparisons ⚡
- Gain insights 🌟

Tools for doing Aggregation:

* ① VLOOKUP (map one product/value to another table)

=VLOOKUP(A2, A2:A14, 2, FALSE)
 search key range index (return column) exact match

Spreadsheet

3

Troubleshooting Questions:

- How should I prioritize these issues?
- In a single sentence, what's the issue I'm facing?
- What sources can help me solve the problem?
- How can I stop this problem from happening in the future?

* You can protect sheet by clicking on Data > Protected sheets & ranges *
to protect the entire sheet / cell from editing *

- JOINS

- (INNER) JOIN (by default) ⚡
- LEFT JOIN
- RIGHT JOIN
- FULL OUTER JOIN



SQL

- COUNT [COUNT not count duplicates] COUNT DISTINCT return distinct values in that range.
- * Inner query executes first *
- Subqueries (or Nested Query / Inner Query / Inner Select)
 - HAVING (WHERE first) to aggregate oh statistics functions.
 - CASE → return record that matches like IF-THEN
 - Aliasing → easier & shorter to reference.



Nested Query

A	B	C	D
	Hours Worked	Rank	Employee #
1	10	1	FT12578
2	20	2	FT12579
3	20	3	FT12580
4	20	4	FT12581

Fix by using INDEX() or MATCH()

```

1 SELECT *
2 FROM (
3   SELECT warehouse.state,
4         COUNT(DISTINCT order_id) AS num_orders
5   FROM warehouse
6   GROUP BY warehouse.state
7   HAVING COUNT(DISTINCT order_id) <= 10
8 ) AS T1
9 WHERE T1.state NOT IN (
10   SELECT state
11   FROM (
12     SELECT warehouse.state,
13         COUNT(DISTINCT order_id) AS num_orders
14     FROM warehouse
15     GROUP BY warehouse.state
16   ) AS T2
17   WHERE T2.num_orders > 10
18 )
  
```

EXAMPLE:

Week 4: Performing Data Calculations
(Pivot, SQL calculations, Data-validation, Temp tables)