Analyze Data to Answer Questions

Saturday, January 28, 2023 22:22

Data Analytics basics

Sunday, October 23, 2022 19:16

Analysis: the process used to make sense of the data collected, identify trends and relationships within data so you can accurately answer the question you're asking

4 Phases of Analysis

1. organize data
2. format and adjust data - streamlines & saves time, filtering & sorting
3. get input from others
4. transform data - identify relationships & patterns in the data

Organize data

Most data will be organized in **tables**. They help organize similar kinds of data into categories and subject areas. They also help figure out what variables are needed and the data type those variables should have. Both filters and sorts are affected by the type of the data.

Format and Adjust

A filter can help find errors or outliers that can then be fixed or flagged. **Outliers** are data points that are very different from similarly collected data and might not be reliable values.

Sorting: arranging data into a meaningful order to make it easier to understand, analyze, and visualize. Ranks data based on a specific metric, e.g. ascending or descending. Helps grouping similar data together.

Filtering: used to only show data that meets a specific criteria and hide the rest. Especially useful when there is a lot of data.

↳ WHERE clause in SQL

Sorting data in Spreadsheets

- ascending/descending, numbers/letters, color

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 cells in a column.

The **SORT** function: use with **FILTER** function for powerful results

= SORT (range, column # to sort by, asc. / desc. TRUE/FALSE)

Customized sort order: when you sort data in a spreadsheet using multiple conditions

Sorting in SQL

ORDER BY:

```
SELECT *
FROM tablename <WHERE ...
ORDER BY column_name
```

> defaults to **ascending** order, should always be the **last line**

```
ORDER BY column_name DESC
```

Convert and format data

Saturday, October 29, 2022 14:39

Incorrectly formatted data can:

- Lead to mistakes
- Take time to fix
- Affect stakeholder's decision-making

Spreadsheets

- Format → Number → choose the desired type
- `= CONVERT(cell num, "current unit", "convert to unit")`
- **Data Validation** (within Spreadsheets) allows you to control what can and can't be entered in your work sheet
  - add dropdown lists with predetermined options
  - create custom checkboxes
  - protect structured data and formulas
    - > reject invalid inputs
- **Conditional formatting**: a spreadsheet tool that changes how cells appear when values meet specific conditions, helps understand spreadsheet at a glance

Transforming data in SQL

- **COERCION** - to work with big numbers, implicit conversion
- **UNIX_DATE** - returns the number of days that have passed since Jan 1 1970, used to work with dates across multiple time zones
- Common conversions using **CAST**: ex:

Numeric (number) → Integer
→ Numeric (number)
→ Big number
→ Floating Integer

```
SELECT
    CAST(mycount AS STRING)
FROM
    mytable
```

mytable

	→ Big number
	→ Floating Integer
	→ String

String	→ Boolean
	→ Integer
	→ Numeric (number)
	→ Big number
	→ Floating Integer
	→ String
	→ Bytes
	→ Date
	→ Date time
	→ Time
	→ Timestamp

Date	→ String
	→ Date
	→ Date time
	→ Timestamp

- SAFE_CAST : returns a value of Null instead of an error when a query fails, same syntax as CAST

Combine multiple datasets

Tuesday, November 1, 2022 08:08

SQL

```
CONCAT(start_station_name, " to ", end_station_name) AS route
GROUP BY
```

Spreadsheets

=FIND(" ",C3) returns location of substring

=LEFT(D2, 11)

=RIGHT(D2,8)

=LEN(C2)

Manipulating strings in SQL

CONCAT	adds strings together to create new text strings that can be used as unique keys	CONCAT('Google','.com')
CONCAT_WS	adds two or more strings together with a separator	CONCAT_WS('.', 'www', 'Google', 'com') ↑ separator first
CONCAT with +	adds two or more strings together using the + operator	'Google' + '.com'

Aggregate Data for analysis

Thursday, November 3, 2022 19:51

R: A programming language frequently used for statistical analysis, visualization, and other data analysis.

Data aggregation: the process of gathering data from multiple sources in order to combine it into a single summarized collection

- identify trends
- make comparisons
- gain insights

Data can be aggregated over a given time period to provide statistics such as: averages, minimums, maximums, sums.

Subquery: a query within another query

VALUE() a function that converts a textstring that represents a number to a numerical value

VLOOKUP(103, A2:B26, 2, FALSE**)**

- 103: value to search for
- A2:B26: range that will be searched
- 2: column number in the rang containing the return value
- FALSE: find exact match (is_sorted)

- only returns the first match it finds
- can only return a value from the data to the right (can't look left)
- TRUE -> approximate matches, FALSE -> exact matches

MATCH() a function used to locate the position of a specific lookup value

IFNA(value, value_if_na**)** -> IFNA(#N/A, "Does not exist"), replaces the #N/A error message with something more descriptive

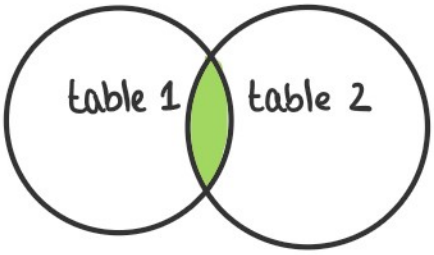
IFNA(value, value_if_na) → IFNA(#N/A, "Does not exist"),
replaces the #N/A error message with something more descriptive

- - -

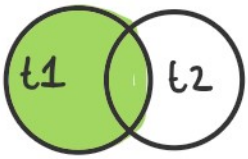
JOIN a SQL clause that is used to combine rows from two or more tables based on a related column

common JOINS · (the table mentioned first is left)

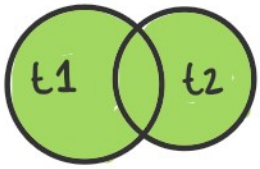
· inner



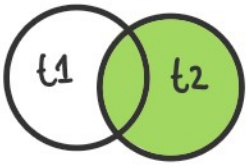
· left



· full outer



· right



INNER JOIN a function that returns records with matching values in both tables

LEFT JOIN returns all the records from the left table and only the matching records from the right table

RIGHT JOIN returns all the records from the right table and only the matching ecords from the left table

OUTER JOIN returns all records in both tables

example:

```
SELECT
    employees.name AS employee_name,
    departments.name AS department_name
FROM
  employees
INNER JOIN
    departments ON
    employees.department_id = departments.department_id
```

```
INNER JOIN
    departments ON
    employees.department_id = departments.department_id
```

AS creates an alias, or temporary name, for a column or table

```
SELECT column_name(s)
FROM table_name AS alias_name;
```

or

```
SELECT column_name AS alias_name
FROM table_name;
```

some SQL databases don't support AS, in that case, it can simply be left out.

```
FROM table_name alias_name
```

COUNT (in SQL) a query that returns the number of rows in a specified range

COUNT DISTINCT returns the number of distinct values in a specified range

```
SELECT
    COUNT(DISTINCT warehouse.state) AS num_states
...
```

Subqueries

Sunday, November 6, 2022 18:41

Subquery: a SQL query that is nested inside a larger query
- also 'inner query' or 'inner select'
- inner query executes first, so that results can be passed on to the outer query to use

HAVING: allows you to add a filter to your query instead of the underlying table that can only be used with aggregate functions

CASE: returns records with your conditions by allowing you to include if/then statements in your query

Data Calculations

Sunday, November 6, 2022 20:23

Common calculation formulas

= SUM()

= AVERAGE()

= MIN()

= MAX()

= COUNTIF(range,"value") COUNTIF(B2:B50,"=1")

= SUMIF(range,criteria/condition, [sum_range])

= SUMIFS(sum_range, criteria_range1, criterion1, [criteria-range2, criterion2, ...])

= COUNTIFS(criteria_range1, criterion1, [criteria_range2, criterion2, ...])

Composite functions

SUMPRODUCT multiplies arrays and returns the sum of those products

= SUMPRODUCT(array1, [array2]...)

ARRAY: a collection of values in cells

Pivot Tables

Tuesday, November 8, 2022 18:27

Calculated field : a new field within a pivot table that carries out certain calculations based on the values of other fields

Elements of a Pivot table

- **rows** : organize and group data horizontally
- **columns** : organize and display values vertically
- **values** are used to calculate and count data, the Values editor creates columns for the pivot table, e.g. using functions like SUM, AVERAGE
- **filters** in a pivot table work like in regular spreadsheets

SQL calculations

Wednesday, November 9, 2022 20:18

Operator: a symbol that names the type of operation or calculation to be performed in a formula

```
SELECT
    column A,
    column B,
    column A + columnB AS columnX
```

+, -, *, /, %, AVG

Group by: a command that groups rows that have the same values from a table into summary rows, SELECT - FROM - WHERE - GROUP BY - ORDER BY

Extract: pulls one part of a given date to use

```
EXTRACT( YEAR FROM STARTTIME)
```

Data validation

Wednesday, November 9, 2022 21:54

Data validation process: checking and rechecking the quality of your data so that it is complete, accurate, secure and consistent

Types of data validation

1) Data Type

- purpose: check that the data matches the data type defined for a field

2) Data Range

- purpose: check that the data falls within an acceptable range of values defined for the field

3) Data constraints

- purpose: check that the data meets certain conditions or criteria for a field, this includes type of data entered as well as other attributes of the field, such as number of characters

4) Data Consistency

- purpose: check that the data makes sense in the context of other related data

5) Data Structure

- purpose: check that data follows or conforms to a set structure

6) Code validation

- purpose: check that the application code systematically performs any of the previously mentioned validations during user input

Temporary Tables

Thursday, November 10, 2022 18:59

Temporary table: a database table that is created and exists temporarily on a database server, the **WITH** clause is a type of temporary table that you can query from multiple times, can also use **SELECT INTO** or **CREATE TABLE** clause

SELECT INTO: copies data from one table into a new table, but doesn't add the new table to the database

CREATE TABLE: good option when several people need to access the same temp table, adds the table into the database

```
CREATE TABLE AfricaSales AS
(
SELECT *
FROM GlobalSales
WHERE Region = "Africa"
)
```