Data Analysis Course 5: Analyze Data to Answer Questions

Module 1

Learn how to come up with clear and objective answers to questions
Organize and format data to do all kinds of calculations

Organize data to begin analysis
Format and adjust data
Aggregate data for analysis
Perform data calculations

Analysis: the process used to make sense of the data collected

- Goal of analysis is to identify trends and relationships within data so you can accurately answer the question you’re asking.

Four Phases of Analysis

1. Organize data
    1. Sorting and filtering
2. Format and adjust data
3. Get input from others
4. Transform data

Sorting is the process of arranging data into a meaningful order to make it easier to understand, analyze, and visualize.

Filtering: showing only the data that meets a specific criteria while hiding the rest.

- WHERE clause helps with this in SQL

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.

You can sort in the menu of a spreadsheet or by writing out a SORT function.

SORT Function
=SORT(A2:D6,2,TRUE)
=SORT(start range cell:end range cell, # of column to sort by, TRUE/FALSE)

- True = Ascending order
- False = Descending order

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

ORDER BY is always the last line in SQL query

```
SELECT
  *
FROM
  `bigquery-public-data.sdoh_cdc_wonder_natality.county_natality`
WHERE
  County_of_Residence = 'Erie County, NY'
  OR County_of_Residence = 'Niagara County, NY'
  OR County_of_Residence = 'Chautauqua County, NY'
ORDER BY
  County_of_Residence,
  Year;
```

```
\*The meteorologists you’re working with have asked you to obtain the temperature, wind speed, and precipitation for stations La Guardia and JFK, for every day in 2020. They’ve also requested the data be presented to them in descending order by date and ascending order by Station ID.*/

SELECT
  stn,
  date,
  IF(
      temp=9999.9,
      NULL,
      temp) as temperature,
  IF(
    wdsp="999.9",
    NULL,
    CAST(wdsp AS FLOAT64)) AS wind_speed,
  IF(
    prcp=99.99,
    0,
    prcp) AS precipitation
 FROM
  `bigquery-public-data.noaa_gsod.gsod2020`
WHERE
  stn = '725030' --La Guardia
  OR stn = '744860'-- JFK
ORDER BY
```

```
  date DESC,
  stn ASC;
```

IF function is used to replace values 9999, 999.9, and 99.99 with **NULL**. The dataset description explains that these are the default values when the measurement is missing.

Terms and definitions for Course 5, Module 1
ORDER BY: A SQL clause that sorts results returned in a query

1. Sort Sheet
2. WHERE
3. Organize data
4. Filtering, filtering, sorting enables
5. first option?
6. second option?
7. Oranize, format, get input
8. Second option?
9. Short sheet and SORT function, sort range
10. SECONd option

Module 2

Incorrectly formatted data can:

- lead to mistakes
- take time to fix
- affect stakeholder's decision-making

CONVERT function can change Fahrenheit to celsius, and so on.

Good idea to copy from formulas and select "Paste Special" > "Values Only"

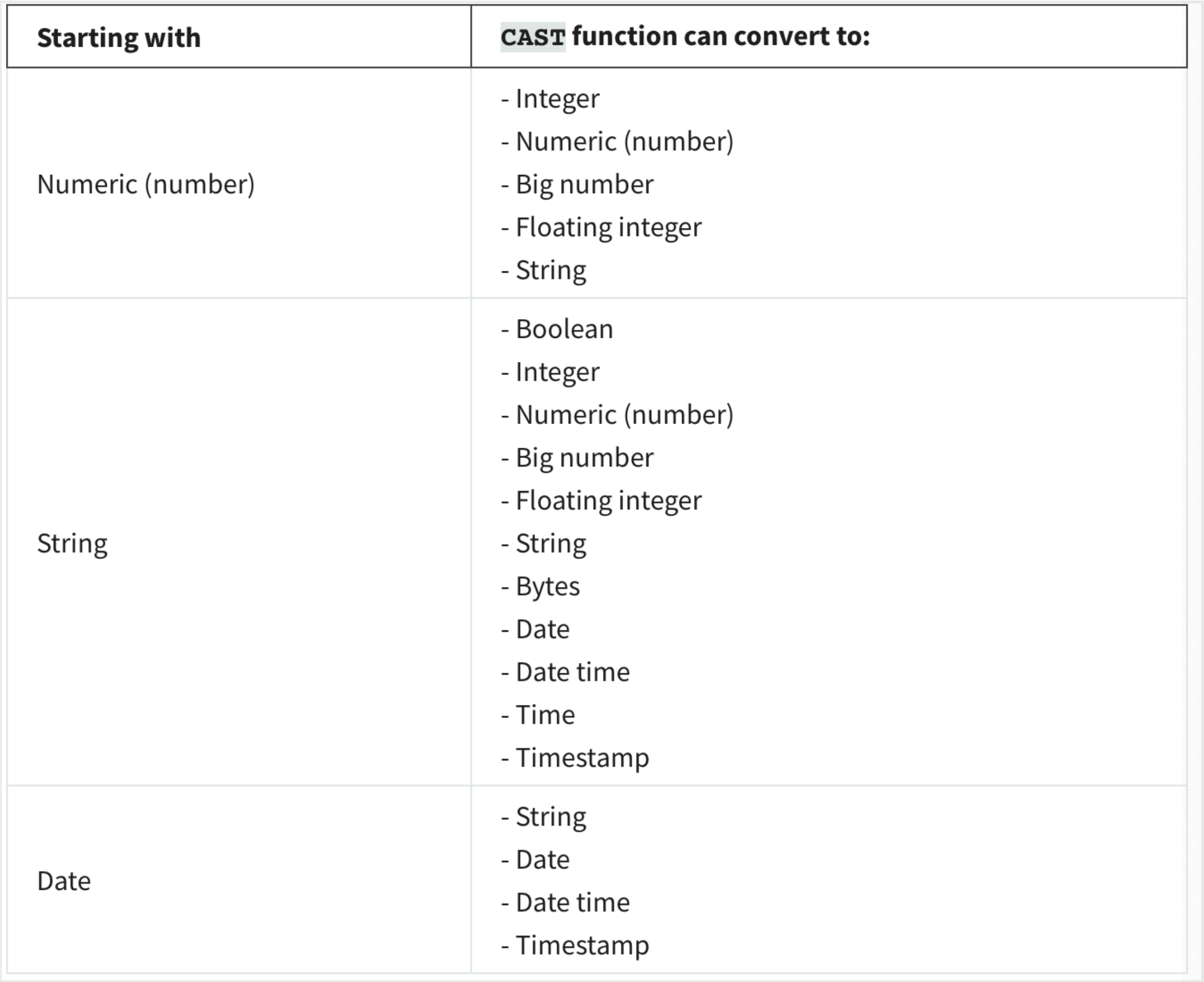
Data validation in spreadsheets allows you to control what can and can't be entered in your worksheet

- Add dropdown lists with predetermined options
- Create custom checkboxes
- Protect structured data and formulas
- Under the "Data" menu in Google Sheets

Data validation can help your team track progress, protect your tables from breaking when working in big teams, and help you customize tables to your needs.

Conditional formatting: a spreadsheet tool that changes how cells appear when values meet specific conditions.

Common uses of CAST function in SQL



Starting with	CAST function can convert to:
Numeric (number)	- Integer - Numeric (number) - 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

CAST syntax: CAST(expression AS typename)

- Where expression is the data to be converted and typename is the data type to be returned.

```
SELECT
      CAST(MyCount AS STRING)
FROM
      MyTable;
```

```
SELECT
      CAST(MyVarcharCol AS INT)
FROM
      MyTable
—INT = integer
```

Datetime values have the format of YYYY-MM-DD hh: mm: ss format, so date and time are retained together. The following **CAST** statement returns a datetime value from a date.

The SAFE_CAST function returns a value of Null instead of an error when a query fails.

Importing and combining data in spreadsheets and databases
Spreadsheets use IMPORTRANGE
=IMPORTRANGE(spreadsheet_url, range_string)

- range_string=cells you want to import (i.e. A2:D60)

SQL uses INSERT INTO command with a SELECT STATEMENT. Syntax below:

```
INSERT INTO [destination_table_name]
SELECT [column names, separated by commas, or * for all columns]
FROM [source_table_name]
WHERE [condition]
```

Concatenate
Spreadsheets: =CONCATENATE(item 1, item 2) AS alias
SQL: SELECT CONCAT(field1, " ", field2)
FROM [table_name]

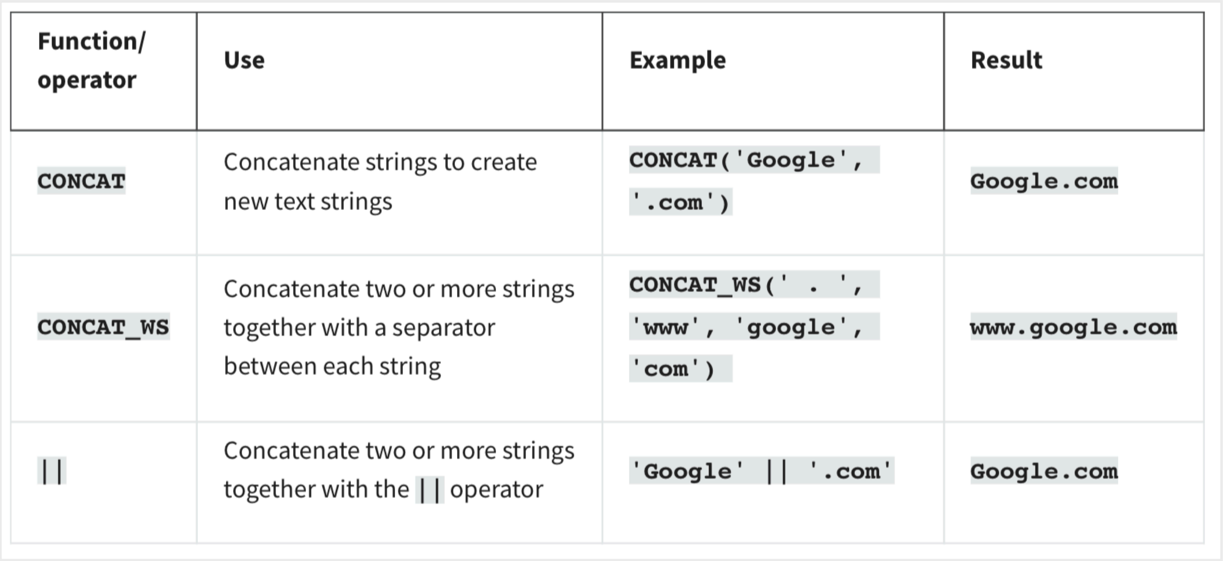
- Notice that this syntax includes " " so that there is a space between the combined fields. With this syntax, SQL combines field1 and field2 with a space between them.

```
SELECT
  usertype,
  CONCAT(start_station_name,' to ',end_station_name) AS route,
  COUNT(*) AS num_trips,
  ROUND(AVG(CAST(tripduration AS INT64)/60),2) AS duration
FROM
  `bigquery-public-data.new_york_citibike.citibike_trips`
WHERE
  LENGTH(usertype) > 0
GROUP BY
  start_station_name,
  end_station_name,
  usertype
ORDER BY
  num_trips DESC
LIMIT 10;
```

Separating values:

- use LEN, RIGHT, LEFT, FIND

Concatenate strings in SQL



Function/ operator	Use	Example	Result
`CONCAT`	Concatenate strings to create new text strings	`CONCAT('Google', '.com')`	`Google.com`
`CONCAT_WS`	Concatenate two or more strings together with a separator between each string	`CONCAT_WS(' . ', 'www', 'google', 'com')`	`www.google.com`
`||`	Concatenate two or more strings together with the `||` operator	`'Google' || '.com'`	`Google.com`

Best practices for searching online

- thinking skills
- data analytics terms
- basic knowledge of tools

mental model: your thought process and the way you approach a problem.

Be able to modify example code to fit your solution.

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

Asking questions on StackOverflow

- Keep it specific.
- Don’t use Stack Overflow to ask questions with opinion-based answers. For example:
  - “Which SQL function can I use to add two numbers together?” is an appropriate question.
  - “Which SQL function is your favorite?” is not.
- Before asking a question, search the Stack Overflow website first—someone may have already asked it. This reduces the number of redundant questions on the site and saves you time.
- Write clear and concise questions in complete sentences. People are more likely to understand what you are asking and can give you more specific or helpful answers.

ROUND: A SQL function that returns a number rounded to a certain number of decimal places

Module 3

Aggregation: collecting or gathering many separate pieces into a whole

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

- Helps analysts:
    - identify trends
    - make comparisons
    - gain insights

Data can also be aggregated over a give time period to provides statistics such as:

- Averages
- Minimums
- Maximums
- Sums

Functions help make data aggregation possible

Subquery: a query within another query

VLOOKUP (Vertical Lookup): a function that searches for a certain value in a column to return a corresponding piece of information.
=VLOOKUP("value/cell you want match of", range_start:range_end, column# you want to bring over, TRUE or FALSE)

- TRUE will return close match, where FALSE will return only an exact match

VALUE: a function that converts a text string that represents a number to a numerical value.

VLOOKUP searches for a search term called a search_key in one column of a spreadsheet. When the search_key is found the function returns the data from another column of the row from which it was located.

- VLOOKUP(search_key, range, index, is_sorted)
- only returns the first match that it finds.

search_key: This is the value the VLOOKUP function will search for. It can be a number, text string, or cell reference.

- The **search_key** must be to the left of the information you want the function to return. This may require you to move columns around before you use **VLOOKUP**.

index: This is the position of the column that contains the data to be returned. The first column in the range is column number 1, and each column is numbered sequentially to the right.

- For example, if the range is B2:D10 and you want to return a value from column D, the index number would be 3. If the index is not between 1 and the number of columns in range, the error message **#VALUE!** will be returned.

Troubleshooting questions

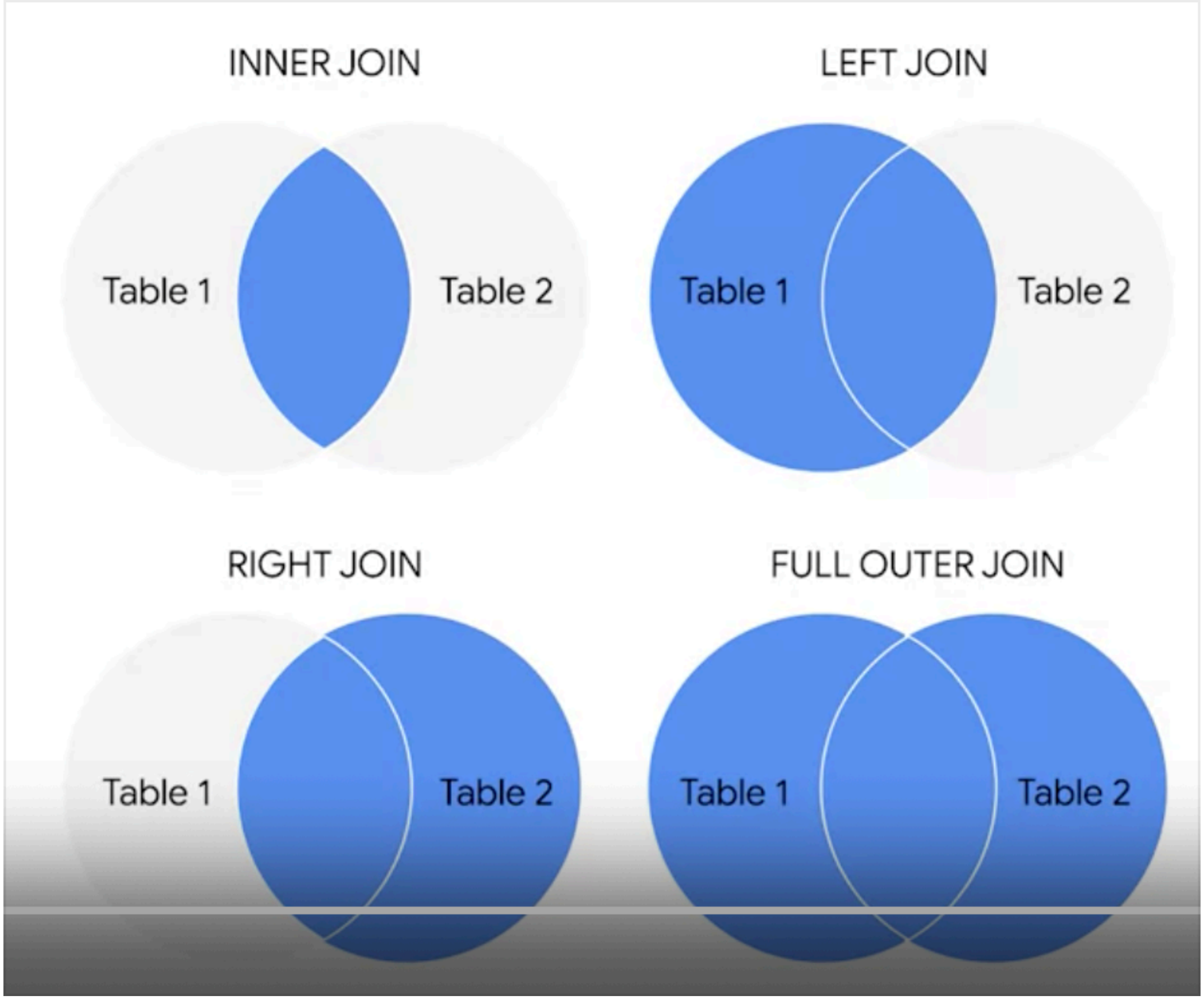
- How should I prioritize these issues?
- In a single sentence, what's the issue I'm facing?
- What resources can help me solve the problem?
- How can I stop this problem from happening in the future?

Absolute reference: a reference that is locked so that rows and columns won't change when copied

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

Common Joins:

- INNER JOIN: function that returns records with matching values in both tables
  - JOIN is shorthand for INNER JOIN
- LEFT JOIN: function that returns all records from left table and matching records from right table
- RIGHT JOIN: opposite of left join
- FULL OUTER JOIN: function that returns all records from both tables



```
SELECT
  employees.name AS employee_name,
  employees.role AS employee_role,
  departments.name AS department_name
FROM
  `silent-complex-430821-f3.employee_data.employees` AS employees
LEFT JOIN
  `silent-complex-430821-f3.employee_data.departments` AS departments
  ON employees.department_id = departments.department_id;
```

Aliases are used in SQL queries to create temporary names for a column or table.

- AS is used to create aliases

Subquery: a SQL query that is nested inside a larger query.

- inner and outer select
- innermost query executes firstc

Subqueries are also known as inner or nested queries

- they allow more complex questions to be answered in a single query

– also make your code more readable and maintainable.

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

- **HAVING**: The **HAVING** clause filters the results of a SQL query based on conditions applied after the grouping. Check out W3School’s HAVING overview
 for a tutorial on this clause
    - The **HAVING** clause **HAVING COUNT(order_items.product_id) > 100** indicates to only retrieve products that have been sold more than 100 times.
- **CASE**: **CASE** provides conditional logic in SQL queries, similar to an 'if-else' structure in programming languages. The W3School’s CASE overview explores the use of the **CASE** statement and how it works.
    - If/then
- **IF**: IF performs a simple conditional test and returns a value depending on the outcome. Review W3School’s IF overview
 for a tutorial of the **IF** function and examples that you can practice with.
- **COUNT**: **COUNT** performs a simple conditional test and returns a value depending on the outcome. Though it seems simple, the **COUNT** function is just as important as all the rest. The W3School’s COUNT overview provides a tutorial and examples.

Only one column in the SELECT clause of a subquery may be selected. To compare multiple columns, select them in the main query.

Clauses like HAVING and CASE paired with subqueries will help you build more and more complex queries, which lets you do more and more complex things in SQL.

Terms and definitions for Course 5, Module 3
Absolute reference: A reference within a function that is locked so that rows and columns won’t change if the function is copied
Aggregation: The process of collecting or gathering many separate pieces into a whole
Aliasing: Temporarily naming a table or column in a query to make it easier to read and write
COUNT DISTINCT: A SQL function that only returns the distinct values in a specified range
Data aggregation: The process of gathering data from multiple sources and combining it into a single, summarized collection
INNER JOIN : A SQL function that returns records with matching values in both tables

JOIN: A SQL function that is used to combine rows from two or more tables based on a related column
LEFT JOIN: A SQL function that will return all the records from the left table and only the matching records from the right table
LIMIT: A SQL clause that specifies the maximum number of records returned in a query
MATCH: A spreadsheet function used to locate the position of a specific lookup value
OUTER JOIN: A SQL function that combines RIGHT and LEFT JOIN to return all matching records in both tables
RIGHT JOIN: A SQL function that will return all records from the right table and only the matching records from the left.
Subquery: A SQL query that is nested inside a larger query
VALUE: A spreadsheet function that converts a text string that represents a number to a numeric value

Module 4

COUNTIF
=COUNTF(range,"criterion")

The range is the array (or collection) of cells that you are checking and the criteria is what you are checking for.

Summary table: a table used to summarize statistical information about data

SUMIF: a function that adds numeric data based on one condition
=SUMIF(range,"criterion",sum_range)

SUMIFS can include multiple conditions
=SUMIFS(sum_range, criteria_range1, "criterion1", criteria_range2, "criterion2", ...)

COUNTIFS can also include multiple conditions
=COUNTIFS(criteria_range1, "criterion1", criteria_range2, "criterion2", ...)

There are also functions for AVERAGEIF, MAXIFS, etc.

SUMPRODUCT: a function that multiples arrays and returns the sum of those products
=SUMPRODUCT(array1, array2,....)

Array: a collection of values in cells

Profit margin: a percentage that indicates how many cents of profit has been generated for each dollar of a sale.

Pivot tables: let you view data in multiple ways to find insights and trends

- They can help you quickly make sense of larger data sets by comparing metrics, performing calculations, and generating reports.
- Four basic parts
  - Rows: organize and group data you select horizontally
  - Columns: organize and display values from your data vertically.
  - Values: used to calculate and count data; where you input the variables you want to measure.
  - Filters: enables you to apply filters based on specific criteria.

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

Operator: a symbol that names the type of operation or calculation to be performed in a formula.
+ addition
- subtraction
* multiplication
/ division

The syntax of a query is its structure

```
SELECT
	columnA,
	columnB,
	columnA + columnB AS columnX
FROM
	table_name
```

Modulo: an operator (%) that returns the remainder when one number is divided by another

GROUP BY: a command that groups rows that have the same values from a table into summary rows

- in a basic query, GROUP BY comes at the end of the query

ORDER BY orders the results

- default is ascending order
- can also use DESC to use descending order

EXTRACT: lets us pull one part of a given date to use

```
SELECT
    EXTRACT(YEAR FROM start time) AS year
```

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

- to make sure your data makes sense
- combines business knowledge and technical expertise
- always make sure your calculations are functioning the right way

Types of data validation:

1. Data type
    - Purpose: Check that the data matches the data type defined for a field.
    - Example: Data values for school grades 1-12 must be a numeric data type.
    - Limitations: The data value 13 would pass the data type validation but would be an unacceptable value. For this case, data range validation is also needed.
2. Data range
    - Purpose: Check that the data falls within an acceptable range of values defined for the field.
    - Example: Data values for school grades should be values between 1 and 12.
    - Limitations: The data value 11.5 would be in the data range and would also pass as a numeric data type. But, it would be unacceptable because there aren't half grades. For this case, data constraint validation is also needed.
3. Data contraints
    - Purpose: Check that the data meets certain conditions or criteria for a field. This includes the type of data entered as well as other attributes of the field, such as number of characters.
    - Example: Content constraint: Data values for school grades 1-12 must be whole numbers.
    - Limitations: The data value 13 is a whole number and would pass the content constraint validation. But, it would be unacceptable since 13 isn’t a recognized school grade. For this case, data range validation is also needed.
4. Data consistency
    - Purpose: Check that the data makes sense in the context of other related data.

    - Example: Data values for product shipping dates can’t be earlier than product production dates.
    - Limitations: Data might be consistent but still incorrect or inaccurate. A shipping date could be later than a production date and still be wrong.
5. Data structure
    - Purpose: Check that the data follows or conforms to a set structure.
    - Example: Web pages must follow a prescribed structure to be displayed properly.
    - Limitations: A data structure might be correct with the data still incorrect or inaccurate. Content on a web page could be displayed properly and still contain the wrong information.
6. Code validation
    - Purpose: Check that the application code systematically performs any of the previously mentioned validations during user data input.
    - Example: Common problems discovered during code validation include: more than one data type allowed, data range checking not done, or ending of text strings not well defined.
    - Limitations: Code validation might not validate all possible variations with data input.

Temporary tables: 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.

- approximates a temporary table

```
WITH trips_over_1_hr AS (
SELECT *
FROM `bigquery-public-data.new_york_citibike_trips`
WHERE trip duration >= 3600
);

## Count how many trips are 60+ minutes long

SELECT
      COUNT(*) AS count
FROM
      trips_over_1_hr;
```

Almost always more than one way to get your analysis done.

BigQuery doesn’t currently recognize the SELECT INTO command

SELECT INTO

```
SELECT
      *
INTO
      AfricaSales
FROM
      GlobalSales
WHERE
      Region = “Africa”
```

- This is great if only you need to access the table.
- A data analyst uses SELECT INTO to copy data from one table into a temporary table without adding the new table to the database.

If others also need to access the same table CREATE TABLE

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

RDBMS: relational database management system

pre-processing data: using temp tables as a holding area for storing values if you are making a series of calculations.

staging: collecting results of multiple, separate queries in temp tables.

- useful if you need to perform a query on the collected data or merge the collected data.

Terms and definitions for Course 5, Module 4

Array: A collection of values in spreadsheet cells

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

Data security: Protecting data from unauthorized access or corruption by adopting safety measures

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

GROUP BY: A SQL clause that groups rows that have the same values from a table into summary rows

Modulo: An operator (%) that returns the remainder when one number is divided by another
Profit margin: A percentage that indicates how many cents of profit has been generated for each dollar of sale
Summary table: A table used to summarize statistical information about data
SUMPRODUCT: A function that multiplies arrays and returns the sum of those products
Temporary table: A database table that is created and exists temporarily on a database server
Underscores: Lines used to underline words and connect text characters