Google data analytics professional course

Week - 1

Data analysis basics

The analysis process

Analysis

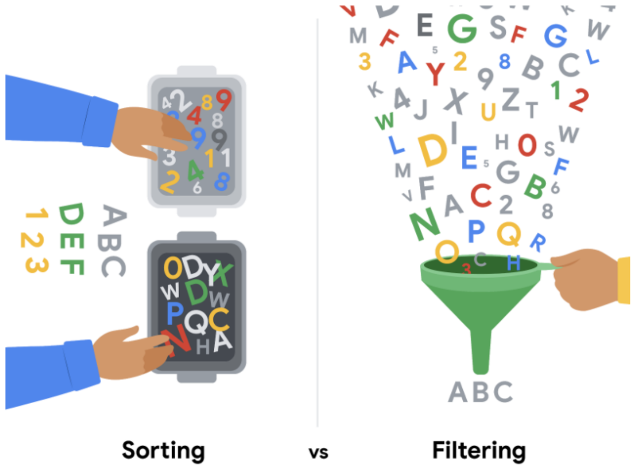
- Basically, analysis is the process used to make sense of the data collected.
- The goal of analysis is to identify trends and relationships within the data so that you can accurately answer the question you're asking.

The 4 phases of analysis:

- Organize data, -list the gift using sort and filter
- Format and adjust data, -convert one type to another
- Get input from others,and -what others bought
- Transform data

Transforming data means identifying relationships and patterns between the data, and making calculations based on the data you have.

Organize data for analysis



More on sorting and filtering

Sorting is when you arrange data into a meaningful order to make it easier to understand, analyze, and visualize.

In SQL

- *IS NOT NULL*
- *> or <*

Sorting in pivot table

Sort and filter data in spreadsheets

- *Sort sheet -all r sorted*
- *Sort range -only the specified range*
- *SORT function -https://support.google.com/docs/answer/3540681*
- *FILTER function -https://support.google.com/docs/answer/3093197?hl=en*

***A customized sort** order is when you sort data in a spreadsheet using multiple conditions.*

Sort and filter data using SQL

Refer saved Queries in BigQuery

- *AVG() as*
- *BETWEEN*
- *ORDER BY DESC*
- *WHERE*

Week - 2

Convert and format data

From one type to another

Currency conversion

Using $ in the spreadsheet or from data tab

=CONVERT

Fahrenheit to celsius

Eg: =CONVERT(B2,"F","C")

=CONVERT(D2, "mph", "m/s")

https://support.google.com/docs/answer/6055540?hl=en

Format data in SQL use CAST()

Converting data in spreadsheets

Help center

G-https://support.google.com/docs/?hl=en#topic=1382883

M-https://support.microsoft.com/

String to date

G-https://www.ablebits.com/office-addins-blog/2019/08/13/google-sheets-change-date-format/

M-https://www.ablebits.com/office-addins-blog/2015/03/26/excel-convert-text-date/#:~:text=Excel%20DATEVALUE%20function%20%2D%20change%20text,Excel%20recognizes%20as%20a%20date.&text=So%2C%20the%20formula%20to%20convert,stored%20as%20a%20text%20string.

String to numbers Combining columns

G-https://productivityspot.com/convert-text-to-numbers-google-sheets/

M-https://www.ablebits.com/office-addins-blog/2018/07/18/excel-convert-text-to-number/

Combining columns

G-https://www.techrepublic.com/article/how-to-split-or-combine-text-cells-with-google-sheets/

M-https://support.microsoft.com/en-us/office/combine-text-from-two-or-more-cells-into-one-cell-81ba0946-ce78-42ed-b3c3-21340eb164a6

Number to percentage

G-https://support.google.com/docs/answer/3094284?hl=en

M-https://support.microsoft.com/en-us/office/format-numbers-as-percentages-de49167b-d603-4450-bcaa-31fba6c7b6b4

Hands-On Activity: Combine multiple pieces of data

Spreadsheet Functions

CONCAT

Eg: =CONCAT(A2,B2)

Output: GeorgeWashington

CONCATENATE

Eg: =CONCATENATE(A3," ",B3)

Output: George Washington

Data validation

Data tab => Data validation

Data validation

- *Adding drop-down lists,*
- *Creating custom checkboxes, and*
- *Protecting structured data and formulas.*

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

Format=>Conditional Formatting

Conditional formatting

A spreadsheet tool that changes how cells appear when values meet specific conditions.

Transforming data in SQL

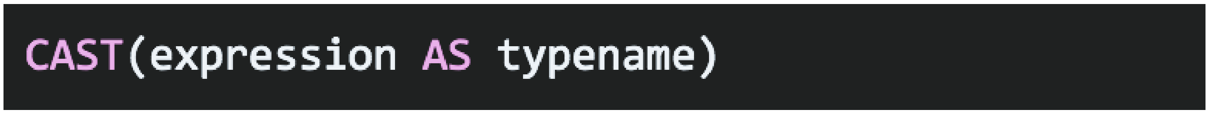
Some functions

- CAST
- COERCION
- UNIX_DATE

https://cloud.google.com/bigquery/docs/reference/standard-sql/conversion_rules

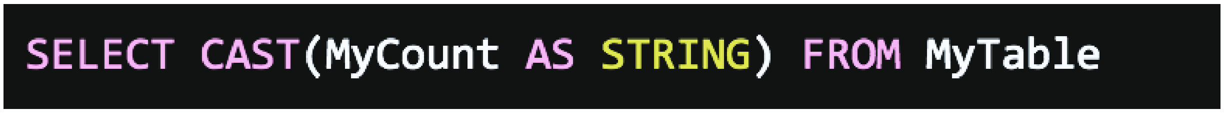
CAST

Syntax



```
CAST(expression AS typename)
```

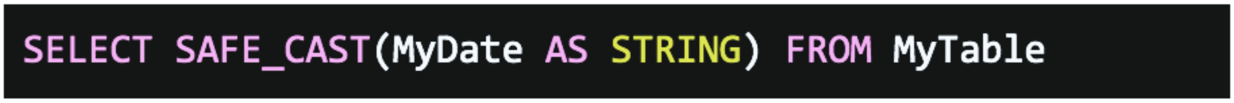
Eg:



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

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

Eg:



```
SELECT SAFE_CAST(MyDate AS STRING) FROM MyTable
```

Combine multiple datasets

Merging and multiple sources

CONCATENATE is a function that joins together two or more text strings.

```
SELECT
   usertype,
   CONCAT(start_station_name, " to ", end_station_name) as route,
   COUNT(*) as no_trip,
   ROUND(AVG(tripduration/60)) as duration
FROM
   `bigquery-public-data.new_york_citibike.citibike_trips`
GROUP BY
   route,
   usertype
ORDER BY
   no_trip DESC
LIMIT
   10
```

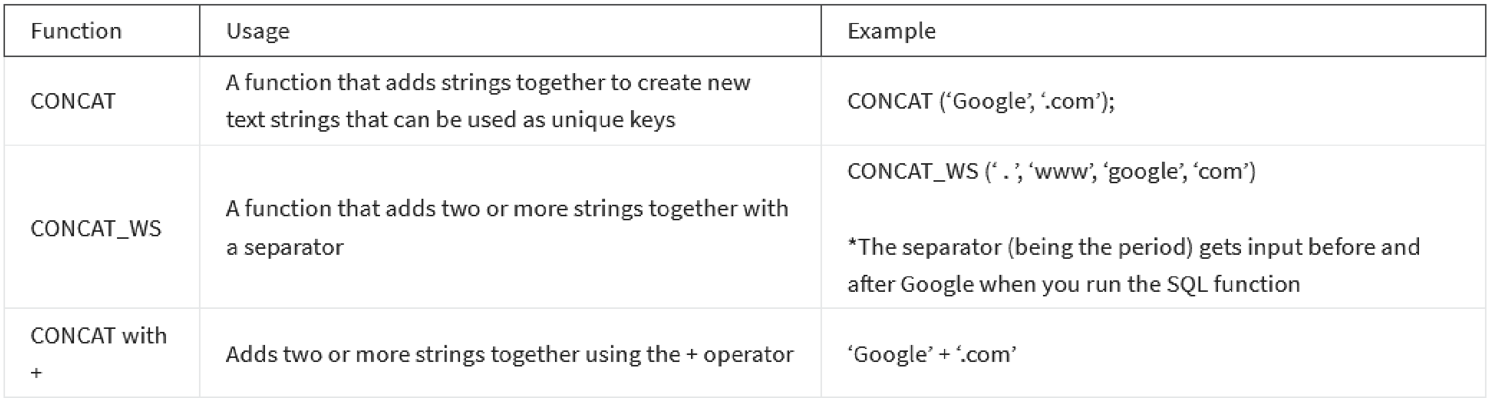
SELECT is based on GROUP BY and WHERE

Strings in spreadsheets

Already we saw this in Module 4 Week-2

- LEN
- LEFT
- RIGHT
- FIND

Manipulating strings in SQL



Function	Usage	Example
CONCAT	A function that adds strings together to create new text strings that can be used as unique keys	CONCAT (‘Google’, ‘.com’);
CONCAT_WS	A function that adds two or more strings together with a separator	CONCAT_WS (‘ . ’, ‘www’, ‘google’, ‘com’) *The separator (being the period) gets input before and after Google when you run the SQL function
CONCAT with +	Adds two or more strings together using the + operator	‘Google’ + ‘.com’

SQL FUNCTIONS

https://www.w3schools.com/sql/sql_ref_sqlserver.asp

SQL KEYWORDS

https://www.w3schools.com/sql/sql_ref_keywords.asp

CONCAT

https://www.w3schools.com/sql/func_sqlserver_concat.asp

CONCAT_WS

https://www.w3schools.com/sql/func_sqlserver_concat_ws.asp

CONCAT with +

https://www.w3schools.com/sql/func_sqlserver_concat_with_plus.asp

Quick Review

In Analysis

Step1 : Organize

Step2: Convert and Data format

Organize

IN SPREADSHEET	IN SQL
Filter	WHERE
Sort	ORDER BY

Convert and Data format

IN SPREADSHEET	IN SQL
Data tab $	CAST
CONVERT	SAFE_CAST
Some links	
CONCAT() and CONCATENATE()	CONCAT()
Data validation	
Conditional formatting	

STRINGS

IN SPREADSHEET	IN SQL
LEN	CONCAT
FIND	CONCAT_WS
LEFT	CONCAT with +
RIGHT	

Get support during analysis

- *Online support*
- *From teammate*

Advanced spreadsheet tips and tricks

Google Spreadsheet

Keyboard shortcuts

https://support.google.com/docs/answer/181110#zippy=%2Cpc-shortcuts

Function list

https://support.google.com/docs/table/25273?hl=en

20 Google Sheets Formulas You Must Know!

https://automate.io/blog/google-spreadsheet-formulas/

18 Google Sheets Formulas Tips & Techniques

https://www.benlcollins.com/spreadsheets/google-sheets-formulas-techniques/

Week - 3

VLOOKUP for data aggregation

Aggregate data for analysis

Aggregation means collecting or gathering many separate pieces into a whole.

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

Preparing for VLOOKUP

Spreadsheet Function

- *VALUE*
- *TRIM*
- *Remove duplicates*

VLOOKUP in action

- *VLOOKUP*
- *MATCH*

VLOOKUP core concepts

- *VLOOKUP()*
- *IFNA()*

https://infoinspired.com/sheets-vs-excel-formula/vlookup-formula-in-excel-and-google-sheets/

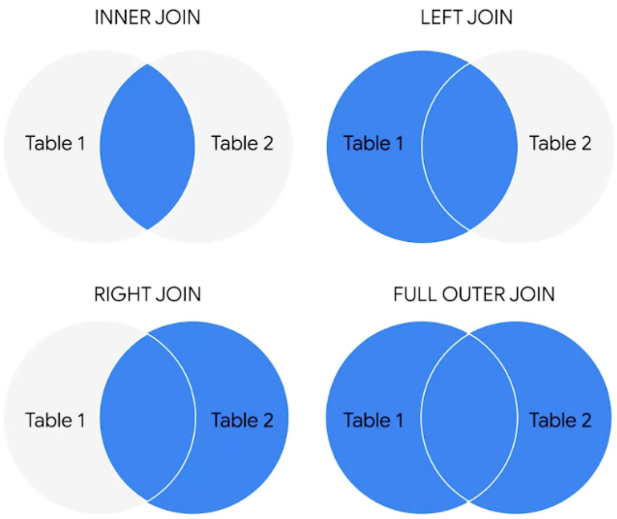
Use JOINS to aggregate data in SQL

Understanding JOINS

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

There are four common JOINs data analysts use,

- *inner,*
- *left,*
- *right, and*
- *outer*



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

A LEFT JOIN is a function that will return all the records from the left table and only the matching records from the right table.

The table mentioned first is left and the table mentioned second is right.

RIGHT JOIN does the opposite. It will return all records from the right table and only the matching records from the left.
OUTER join or FULL join combines RIGHT and LEFT JOIN to return all matching records in both tables.

Eg:

```
SELECT
    `dulcet-velocity-294320.Module_5_analysis.employee`.name as Work_title,
    `dulcet-velocity-294320.Module_5_analysis.employee`.department_id as ID,
    `dulcet-velocity-294320.Module_5_analysis.department`.name as Name
FROM
 `dulcet-velocity-294320.Module_5_analysis.employee`
full  join
`dulcet-velocity-294320.Module_5_analysis.department` on
    `dulcet-velocity-294320.Module_5_analysis.employee`.department_id =
`dulcet-velocity-294320.Module_5_analysis.department`.department_id
LIMIT 1000
```

Using aliases 'as'

```
SELECT
    e.name as Work_title,
    e.department_id as ID,
    d.name as Name
FROM
 `dulcet-velocity-294320.Module_5_analysis.employee` as e
full  join
`dulcet-velocity-294320.Module_5_analysis.department` as d on
    e.department_id = d.department_id
LIMIT 1000
```

COUNT and COUNT DISTINCT

- *COUNT*
- *COUNT DISTINCT*

COUNT is a query that returns the number of rows in a specified range. COUNT DISTINCT is a query that only returns the distinct values in that range. This means COUNT DISTINCT doesn't count repeating values.

Work with subqueries

Queries within queries

We can put Query in the Query

Locations: SELECT FROM WHERE

Using subqueries to aggregate data

- HAVING
- SELECT
- CASE
- CONCAT
- COUNT
- FROM
- LEFT JOIN
- GROUP BY
- WHERE
- Huge code refer video

SQL functions and subqueries

```
SELECT account_table.*
  FROM (
        SELECT *
          FROM transaction.sf_model_feature_2014_01
         WHERE day_of_week = 'Friday'
       ) account_table
 WHERE account_table.availability = 'YES'
```

There are a few rules that subqueries must follow:

- *Subqueries must be enclosed within parentheses*
- *A subquery can have only one column specified in the SELECT clause. But if you want a subquery to compare multiple columns, those columns must be selected in the main query.*
- *Subqueries that return more than one row can only be used with multiple value operators, such as the IN operator which allows you to specify multiple values in a WHERE clause.*
- *A subquery can't be nested in a SET command. The SET command is used with UPDATE to specify which columns (and values) are to be updated in a table.*

Week - 4

Get started with data calculations

Common calculation formulas

- *SUM*
- *% Growth finding*
- *AVERAGE*
- *Conditional formatting*
- *MIN*
- *MAX*

Functions and conditions

- *COUNTIF*
- *SUMIF*
- *AVERAGEIF*
- *COUNTIFS*
- *SUMIFS*

Composite functions

SUMPRODUCT *is a function that multiplies arrays and returns the sum of those products.*

The ***profit margin*** *is a percentage that indicates how many cents of profit have been generated for each dollar of sale.*

- *SUMPRODUCT*

Pivot...pivot...pivot…

Pivot table

- https://support.google.com/docs/answer/1272900?co=GENIE.Platform%3DDesktop&hl=en
- https://infoinspired.com/google-docs/spreadsheet/all-about-calculated-field-in-pivot-table-in-google-sheets/
- https://www.benlcollins.com/spreadsheets/pivot-tables-google-sheets/

Using pivot tables in analysis

- Perform calculations
- Sort your data
- Filter your data
- Format your data (group by)

Learn more SQL calculations

Queries and calculations

An operator is a symbol that names the type of operation or calculation to be performed in a formula.

Eg:

- +
- *
- -
- /

Common functions in spreadsheet and SQL

SUM	SUM
AVERAGE	AVG

Embedding simple calculations in SQL

```
/*SELECT
    Date,region,Small_Bags,Large_Bags,XLarge_Bags,Total_Bags,
    Small_Bags + Large_Bags + XLarge_Bags as TOTAL

FROM
    `dulcet-velocity-294320.Module_5_WEEK_4.avocado` LIMIT 1000*/

SELECT
    Small_Bags,
    Total_Bags,
    (Small_Bags/Total_Bags)*100 as small_bag_per
FROM
    `dulcet-velocity-294320.Module_5_WEEK_4.avocado`
WHERE
    Large_Bags != 0 -- or Large_Bags <> 0
```

Calculations with other statements

```
SELECT
      EXTRACT(YEAR FROM starttime) as Year,
      COUNT(*) as Total
FROM `bigquery-public-data.new_york_citibike.citibike_trips`
GROUP BY
      Year
ORDER BY
      Year
```

EXTRACT

The data-validation process

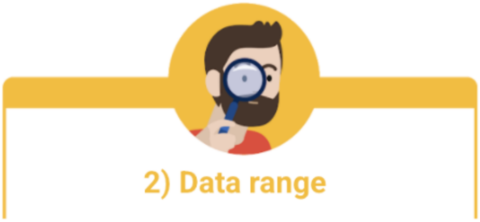
Check and recheck

Data validation process

This process involves checking and rechecking the quality of your data so that it is complete, accurate, secure, and consistent.

Types of data validation

- ***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 constraints

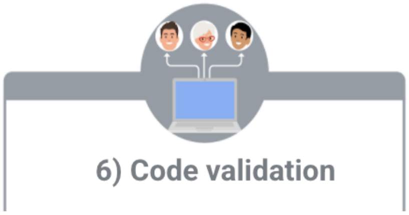
- ***Purpose:*** *Check that the data meets certain conditions or criteria for a field. This includes the type of data entered as well as other attributes of the field, such as number of characters.*
- ***Example:*** *Content constraint: Data values for school grades 1-12 must be whole numbers.*
- ***Limitations:*** *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.*



- *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.*

Using SQL with temporary tables

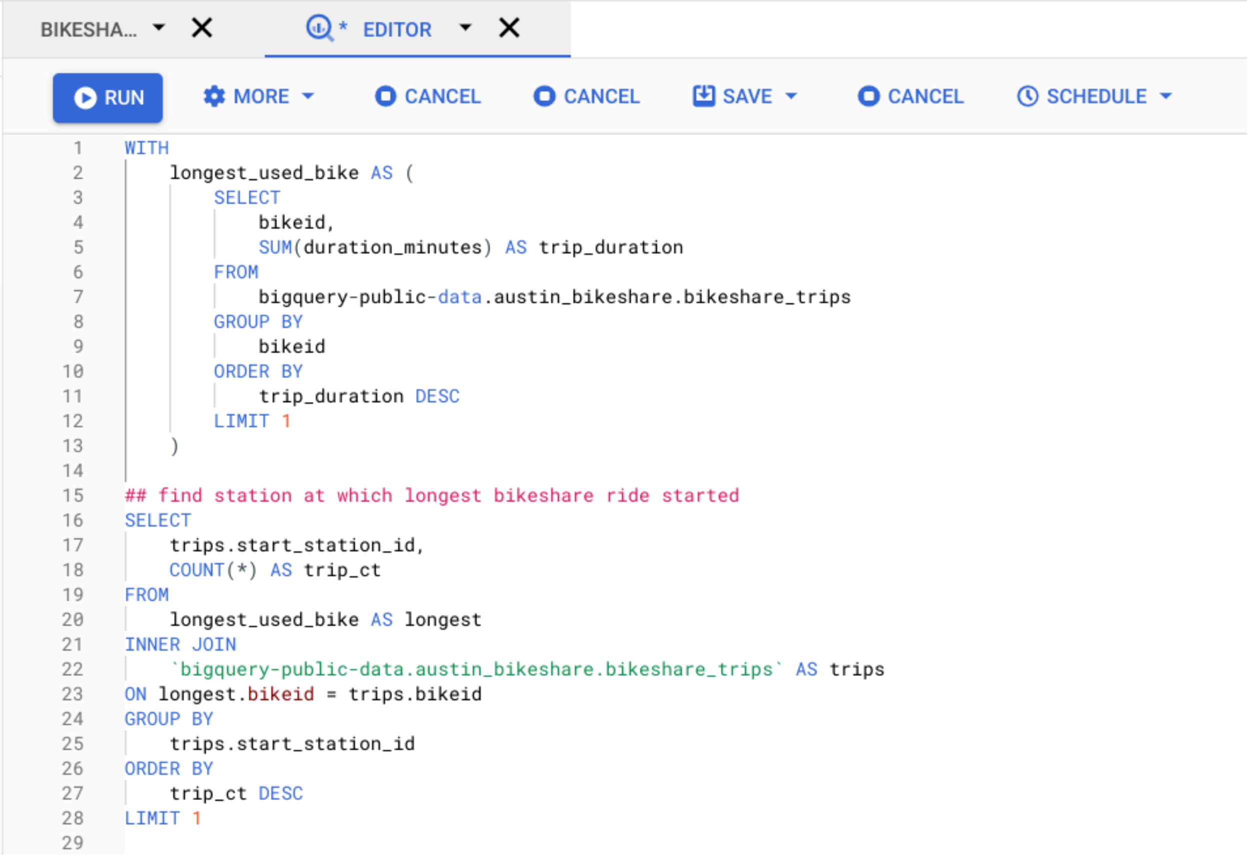
Temporary tables

*A **temporary table** is 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.*

```
WITH trip_1hr as (
    SELECT *
    FROM `bigquery-public-data.new_york_citibike.citibike_trips`
    WHERE tripduration >= 60

)

##count trip over  60 minits
--description line above
SELECT
    count(*)
FROM
    trip_1hr
```



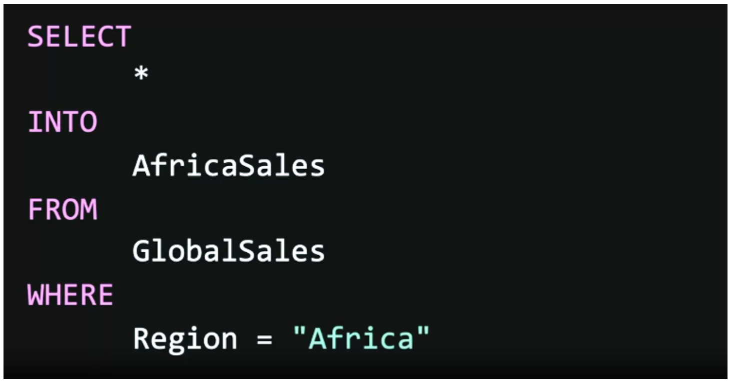
```
WITH
  longest_used_bike AS (
    SELECT
      bikeid,
      SUM(duration_minutes) AS trip_duration
    FROM
      bigquery-public-data.austin_bikeshare.bikeshare_trips
    GROUP BY
      bikeid
    ORDER BY
      trip_duration DESC
    LIMIT 1
  )

## find station at which longest bikeshare ride started
SELECT
  trips.start_station_id,
  COUNT(*) AS trip_ct
FROM
  longest_used_bike AS longest
INNER JOIN
  `bigquery-public-data.austin_bikeshare.bikeshare_trips` AS trips
ON longest.bikeid = trips.bikeid
GROUP BY
  trips.start_station_id
ORDER BY
  trip_ct DESC
LIMIT 1
```

Multiple table variations

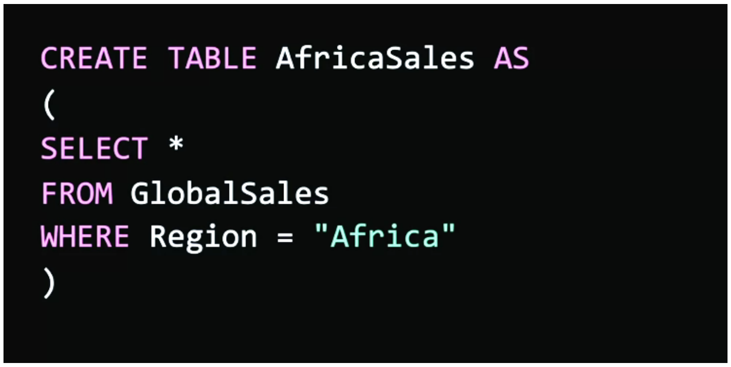
SELECT INTO

Temporary table creation in other databases (not supported in BigQuery)



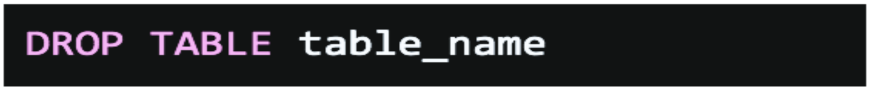
```
SELECT
        *
INTO
        AfricaSales
FROM
        GlobalSales
WHERE
        Region = "Africa"
```

CREATE TABLE



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

After you have completed working with your temporary table, you can remove the table from the database using the DROP TABLE clause.



```
DROP TABLE table_name
```

WITH clauses, CREATE TABLE statements, and CREATE TEMP TABLE statements all create temporary tables in queries.

Working with temporary tables

Best practices when working with temporary tables

- *Global vs. local temporary tables*
- *Dropping temporary tables after use*

For more information

- *BigQuery Documentation for Temporary Tables*
- *How to use temporary tables via WITH in Google BigQuery*
- *Introduction to Temporary Tables in SQL Server*
- *SQL Server Temporary Tables*
- *Choosing Between Table Variables and Temporary Tables*

Your intermediate guide to SQL

Refer pdf **" M5_W4_Your intermediate guide to SQL.pdf "**

Quick Review

Week-3

Data aggregation

Spreadsheet

Prepare for vlookup

VLOOKUP

SQL

JOIN

COUNT,DISTINCT

Subqueries

Week-4

Calculations

Spreadsheet

Calculations

Pivot table

SQL

Calculations

Data validation

Checklist provided

SQL

Temporary table

Dhamodharan

20/10/2021