Week 1 1/5

Course 05 :→ Analyze Data to Answer Questions

Let's get organised

06 | 02 | 23

Instructor : Ayanna
Global Insight Manager

Best way to know if you're creating value is if you have evidence i.e. data

---

Created a learning log & took brief overview of what's to come

---

2/5 Analysis:
The process used to make sense of data collected

The goal of analysis is to identify trends and relationship within data so you can accurately answer the questions you're asking.

4 Phases of Analysis :-
① Organize Data : online Registry / etc
② Format and adjust data : sort & filter
③ Get input from others : Viewpoint you don't have
④ Transform data : Identifying Rel^n & pattern

"The most important thing that you need to have throughout this learning journey is grit".

3/5

Week 1

Page No.

Date 07 | 02 | 23

Always a need to Organize

* Sorting and filtering are affected by the type of data were working with.
  &
  Just have your data in the right format.

---

* Sorting & Filtering ↦

Ranks data based on specific metric

Sorting :
When you arrange 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.

```
SELECT * FROM movie-data.movies
WHERE Genre = 'Comedy';
```

---

* Advanced Sorting :

(*) 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 beside the specified cells in a column.

Doesn't keep information across rows together.

→ 2 methods for sorting :
① Menu ✓
② Writing Sort function

Sheets :→
Select specific column > Data > Sort by sheet / sort by Range

In the movie dataset we want to keep all of the information together, so first we'll sort sheet

⊛ SORT Function :
(index)
= SORT (Range, Sort_by, , TRUE/FALSE)
= SORT (A1:D6, 2, TRUE)
→ sorts the range by B col in Asc order

* Customized sort order
When you sort data in spreadsheets using multiple conditions. [order is important]

Data > sort > Advanced > Add multiple columns

Week 2
5/5

Page No.
Date 08 02 23

Sorting using SQL

* ORDER BY → Sorting
  ↓
  Last clause in a query!

WHERE → filtering

AND → Customized Sorting / multi-sort

Ex.① SELECT Name FROM Genre
ORDER BY Name;

↳ Alternative, Word

② SELECT * FROM Tracks
WHERE Composer = 'Chris Cornell'
ORDER BY GenreId DESC;
→ You Know My Name

In Video 1→ SELECT * FROM movies
WHERE Genre = "Comedy"
AND Revenue > 300,000,000
ORDER BY Release_Date DESC;

Week 2
1/3

Page No.
Date 09 02 23

You're as good of an analyst as your ability to ask right questions.

Incorrectly formatted data can:
- lead to mistakes
- Take time to fix
- Affect stakeholder's decision-making

* Sheets > Format > More format / chose

* Changing unit of measurement!

°F → °C Temperature

= CONVERT(C01, "F", "C")

Tip: After adding / performing any function. Copy & paste the ~~data~~ as values

paste special > values only

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

Page No.
Date

Data Validation (function)

Allows you to control what can and can't be entered in your worksheet

① > Add dropdown lists with predetermined options

•

•

we wanna add a list of options

Select desired column
> Data > Data Validation

Criteria: Not Yet started, In Progress, Ready

Done ✓

other thing we can do with data validation is

② > Custom checkboxes.

Criteria: checkbox

☑ Use custom cell values

Checked: Approved

Unchecked: Not Approved

③ Protect structured data & formulas.

Reject invalid inputs

Conditional formating

Select desired > Format > |

↓

Column to Set rule (1) & apply to (2)

Text is Exactly
- Not Yet Started → Color ⇒ Red
- In Progress → Color = Yellow
- Ready → Color = Green

Review date : Orange, Today

① Highlighting cells that contains the word (target)
② Color-coding cells that contain dates after today.

---

*IMP

CONCAT for only 2 variables and
CONCATENATE for more than two

CONCAT(A2, B2)
CONCATENATE(A2, " ", B2)

CONVERT(col, "F", "C")

week2

2/3

Page No.
Date 10/02/23

Merging & multiple sources

Query describing usertype their route and number of trips, duration.

```
SELECT usertype,
CONCAT(start_station, " to ", end_station) As route,
COUNT(*) as num_trips,
ROUND(AVG(CAST(tripduration as int64)/60; 2) as duration
From T_name
GROUP BY start_station, end_station, usertype
ORDER BY num_trips DESC
LIMIT 10;
```

LEN, LEFT, RIGHT, FIND in Spreadsheet

Same as that of in SQL

LEN(C2) → 14

FIND(" ", C2)

↑ substring

finds space in C2.

Right 8 chars: RIGHT(C2, 8) → 1 2 3 4 . . 8

Left 4 chars: LEFT(C2, 1) → 1 2 3 . . . 10 11 12 . .

* CONCAT ('Google', '.Com'); → Google.com

* CONCAT_WS ('.', 'www', 'google', 'com')
→ www.google.com

↓

Adds 2 or more strings with a ~~separator~~

* CONCAT with + is allowed.

Ex. 'Google' + '.com'

SELECT CASE WHEN A>10 THEN "A"+" is great";
END AS LOL
FROM LMAO;

* CASE Version:

SELECT CASE
WHEN condition THEN Soln
WHEN Condition → 1
WHEN Condition → 1
END As NewColumn
FROM Tname;

CAST SELECT CAST (myDate AS STRING) FROM Tname;
DATETIME

SAFE_CAST → returns null instead of Error

Week 2
3/3

Page No.
Date 13 02 23

* What to do when you get stuck?
  > Ask a senior or mentor
  > Consult a team member
  > Search online
  > Utilize problem solving techniques
  > Use online forums.

Reaching out is the way to go!

Layla : Analytical Lead (Google (Obv)
↓
Someone who helps advertisers understand the value of their money.

"The analyze stage is where you become Expert about your dataset"

You get to tell the story once the dataset is cleaned and loaded in Excel/V. tool

Creating a Pivot table to get the numbers, ~~facts~~ ~~fact~~ what is happening & what should they expect.

* Narrow down the error to ① Formula or ② Data
double check the formula & if not then data.

* Finding answers online to your problems can be empowering and can give you new knowledge for future.

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.

* Use the right terms.

* Upcoming we'll learn about R
> R is a programming language popularly used for statistical analysis, visualization and other data analysis.

* Being able to modify example code is a must know skill.

Week 3
1/3

Page No.
Date 14 | 02 | 23

VLookUp for data Aggregation

Data Aggregation:-
The process of gathering data from multiple sources in order to combine it into a single summarized collected

Puzzle pieces = data
Organisation = aggregation
Pile of pieces = summary
putting the pieces together = gaining insights

→ Identify trends
> make comparisons
> Gain insights.

Data can also be aggregated over a given time period to provide statistics such as:-
> Averages
> minimums
> maximums
> Sums

functions help make data aggregation possible.

* Subquery :→
A query within a query.

VLOOKUP (value, Range, Return index, TF) F => Exact

Page No.
Date

Data Aggregation tool

↓

VLOOKUP :

A function that searches for a certain value in a column to return a corresponding piece of information.

VALUE :

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

Before using the VLOOKUP ()

* Make sure of :-
  - > Data is formatted correctly : VALUE ()
  - > There are no extra spaces : TRIM ()
  - > Duplicates : Remove duplicates

Vlookup Action :-

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

* Most commonly Vlookup is used to populated data from one sheet to another sheet.

From the Example :-

=VLOOKUP (A2, 'Emp Sheet'! $A$2:$B$5, 2, FALSE)

Lock the sheet > Data
> Protect sheets and ~~ranges~~

VLOOKUP(key, range, index, T/F)

* Limitations of VLOOKUP & common problems &

Troubleshooting

① How should I prioritize these issues?
② In a single sentence, what's the issue I'm facing?
③ What resources can help me solve the problem?
④ How can I stop this problem from happening in the future.

(*) VLOOKUP only returns the first match it finds

(*) Can only the data from the right, can't look left.
[data we want] [lookup value]

* Absolute Reference
A ref. that is locked so that rows and cols ~~don't~~ change when copied.

* MATCH
A function used to locate position of specific lookup value

TRUE : Approximate match
FALSE : Exact match

Week 3
2/3.

Default Join ⇒ Inner Join

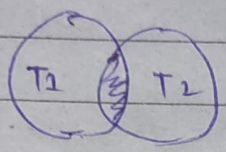
Page No:
Date 15 02 23

Start working with ~~First~~ Joins

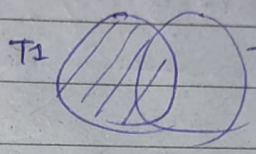
Joins :→

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

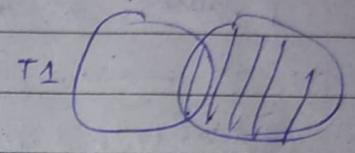
Common Joins :→



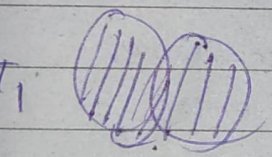
Inner Joins



Left Join



Right Join



full Outer Join

Inner Join :→ A function that returns records with matching values in both tables.

Left Join :→ A function that will return records from the left table and only the matching records from the right table

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

Right Join :→ A function that will return records from the right table and only the matching records from the left table.

Page No.
Date

Outer Joins

A function that combines RIGHT & LEFT join to return all matching records in both tables.

If there are no matches it'll create NULL values in those places.

Example:

```
SELECT
        employee.name       AS  employee_name,
        employee.role       AS  employee_role,
        departments.name    AS  department_name
FROM
        employee_data.employees
INNER JOIN
        employee_data.departments ON
        emp.dept_id = dept.dept_id
```

If we instead use LEFT JOIN
RIGHT JOIN
FULL OUTER JOIN
values will change accordingly!

We'll be using this to answer the question "How many?"

Page No
Date

* COUNT in Spreadsheet :→
Can be used to count the total number of numerical values within a specific range in spreadsheet.

* COUNT in SQL :→
A query that returns the number of rows in a specified range.

* COUNT DISTINCT :→
A query that only returns the distinct values in a specified range.

* Aliasing :→
When you temporarily name a table or column in your query to make it easier to read and write

Ex. ⇒

```
SELECT orders.*,
       ware.ware --alias
       ware.state
FROM ware.orders AS orders
JOIN
    ware-orders.warehouse ware ON orders.warehouse_id
```

```
SELECT COUNT(col)

SELECT COUNT(DISTINCT col)
```

Week 3
3/3

Page No.
Date 16/02/23

Queries inside query

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

Nesting Dolls → matryoshka (Russian Nesting dolls)

Keep in mind inner query executes first.

Usually subqueries are with FROM or WHERE Clause

Example →

```
SELECT AVG (salary)
FROM: (SELECT { NAME
       WHERE NAME LIKE "%e"
       FROM table });
```

IN, ANY, ALL

```
SELECT col_name FROM T_name
WHERE salary > ALL (SELECT ... );
```

Salary IN
Salary = ANY

Normal flow → WHERE & then GROUP BY

Page No.
Date

Wherever we want to have Group By first then WHERE clause, we need to use HAVING

* 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.

Clauses like HAVING & CASE, paired with subqueries will help you build more and more ~~subqueries~~ complex queries.

Justin, Data Analytics lead!

"Your career path is not always straightforward"

Keep changing little by little, figure out what's exciting about your role RN.

Find a job that let's you do more of what you like

Be curious! Ask why's