Page No.
Date 27/01/23

Continue...

"If you change the way you look at things, the things you look at change."
-Wayne Dyer

No two projects in DA are exactly the same.
Different projects require us to focus on different information differently.

Today we'll look at such methods:
> Sorting
> Filtering
> Pivot tables
> VLOOKUP
> plotting

sorting: Arranging data into a meaningful order to make it easier to understand, analyze & visual
↓
Brings duplicate identities together for faster recognition

Filtering: showing only the data that meets a specific condition/criteria while hiding the rest.
↓
finding particular information

Pivot Table: Data summarization tool used in DAta processing
→ they sort, reorganize, group, count, total or average data stored in database
↓
Used to get quick clutter free view of data.

| $:Lock | $A2 ↑ Column Lock | A$2 Row Lock | $A$2 Cell Lock |
|---|---|---|---|

step 1 : Chose the desired range
step 2 : Insert > Insert Pivot Table
step 3 : New/Existing sheet → [Create]

---

* VLOOKUP: Vertical lookUP
A function that searches for a certain value in a column to return a corresponding piece of information.

Usually the data is spread across sheets or even database, that's when Vlookup comes in handy.

Syntax : =VLOOKUP(what, whe/Range, col to Ret, T/F)

what : what to look for
whe/Range: where to look, ! for range of other sheet
ex! K2:G47
Col to Ret: column to return if matched.
T/F : T: Exact match F: Approx match

Ex. VLOOKUP(A2, 'sheet2'!A1:B31, 2, False)

Searches for A2 in A1:B31 in sheet2 & returns value of B column (2) at exact match.

& we used find option in [Edit]

Data mapping gives us a clean roadmap to follow to make sure data arrives safely at its destination.

Page
Date

## (*) Plotting :→

When you plot data you put it in a graph, chart, table or other visual to help you quickly find what it looks like.

↓

Plotting is useful to identify skewed data or any outliers.

Select the colums & then
Insert > Chart > Choose the visual.

---

## (*) Data mapping :

process of matching fields from one data source to another.

↓

Data mapping is very imp. to the success of data migration, data integration and lots of other data mangement activites!

Depending upon the schema and number of primary and foreign keys in data source

↓

Data mapping can be simple or very complex.

Data mapping is so important because even one mistake when merging data can ripple throughout an organization.

Schema: A way of describing how something is organized.

* Data mapping tools can analyze field by field how to move data from one place to another then they can also automatically clean, match, inspect and validate the data.

↓

> Create consistent naming conventions
> When selecting, make sure it supports your files.

Manual way:→

> Check for the columns in both sheets, if they have the same format let them be
> Make any columns consistent for both sheets etc.

Now, merging two sheets can be done by:
> Querying
> Import wizards
> Simply Drag n Drop

Testing Phase of Data mapping:
You inspect sample piece to data to make sure its clean & properly formatted

Finally, once its confirmed to be clean & compatible we can use it for analysis.

Week 3
1/3

Page No.
Date 30 01 23

# Using SQL to clean data

This week :→
> Different data cleaning functions in sheets & SQL
> How SQL can be used to clean large datasets
> Apply basic SQL functions for transforming data and cleaning strings.

---

- SQL makes our lives easier when we are analyzing lots of data.
> SQL is a core skill for DA

Structured Query language.
Trillions of rows.

Ex. A list of names of all the peoples in the world (8 Billion)
↓
An average reader would require 101 years just to read those names.
↓
SQL can do this in seconds. lol!

Developing began in early 1970
In 1970 Edgar F. Codd developed RDBMS theory.
↓
At this time IBM was using RDBMS "System R"
↓
IBM computer scientists were having hard time manipulating & retrieving

So they switched to "Sequel"

In 1976, Sequel was released publically
In 1986, SQL became standard RDBMS language & it still is!

---

* SQL & Spreadsheets :-

> There are tools you can use in both spreadsheets & SQL.
> COUNT IF = COUNT + WHERE

> Both allows to filter data (WHERE & Filter)
> Both allows to sort data (ASC & DESC)
> Group data :
> Perform Calculation :

| SQL | Spreadsheet |
|---|---|
| > SUM (col_name) | > SUM (range) |
| > AVERAGE (col_name) | > AVERAGE (range) |
| > COUNT (--) | > COUNT (--) |
| > MIN (--) | > MIN (--) |
| > MAX (--) | > MAX (--) |

---

DISTINCT — Remove Duplicates

weeks 3/3

Just running query doesn't save it! You'll have to explicitly download it as sheet or save result in new table

Widely used SQL queries

31/01/22

> SELECT

↓

SELECT

```
SELECT name, city
FROM customer_data.customer_data;
```

INSERT

```
INSERT INTO cust_data.cust_address
    (Customer_ID, address, city ...)
VALUES
    (2645, '333 SQL ROAD', ...);
```

UPDATE

```
UPDATE customer_data.customer_address
SET address = '123 New Address'
WHERE customer_id = 2645;
```

DROP IF EXISTS

# When you're creating lots of tables in a database
You'll want to keep it clean so by using DROP TABLE IF EXISTS.

# Be careful, do not delete any important organisational data

---

* Evan, Portfolio Management

Easy to learn & even more fun to master

"Be super curious about whatever data set that you're given"

Totally Agree

Page No.
Date

## Cleaning String Variables :→

* SQL has DISTINCT for removing Duplicates

following Along :-

```
SELECT DISTINCT customer_id
FROM Customer_data.customer_address;
```

LEN/LENGTH :

```
SELECT LENGTH(country) AS letters_in_country
FROM customer_data.customer_address;

-- we need all country length = 2
-- Perform following query to check which country
   has length greater/lesser than 2

SELECT country
FROM customer_data.customer_address
WHERE LENGTH(country) <> 2;
```

<> → Not equal to

(*) SUBSTR(col, start, stop)

SUBSTR(country, 1, 3) → start 3 letters of country

```
SELECT DISTINCT customer_id
FROM customer_data.customer_address
WHERE SUBSTR(country, 1, 2)  -- ## for Consistency !
      = 'US';   -- USA, will also appear
```

* TRIM()

↓ Remove extra spaces

```
SELECT state
FROM customer_data, customer_address
WHERE LENGTH(state) >2;
```

↓

figure out what incorrectly entered states are,

↓

output _OH (Ohio)

↓

with extra space

Now we use the TRIM function in WHERE to list

* Query:

```
SELECT DISTINCT customer_id
FROM customer_data.customer_address
WHERE TRIM(state)= 'OH';
```

* SELECT customerid, InvoicesDate, Total
FROM invoices
WHERE Total >25;

# Advanced Data Cleaning Part 1

Page No.
Date 01/02/2023

## * CAST()

When you import data that doesn't already exists in SQL tables, the datatypes from new dataset might not have been imported correctly!

↓

CAST() can be use to convert anything from one data type to another.

```
-- sort prices descending order
SELECT purchase_price
FROM customer_data.customer_purchase
ORDER BY purchase_price DESC ;
```

↓

89.85
799.99 } wrong result due to wrong data type its string but it should be float

How:→ apple, orange Descending → Orange
Apple

799.99, 89.85 Descending:→ 89.85
799.99

Because it starts with the first letter 8, 7 8>7

Typecasting :→ Converting data from one type to another.

↓

```
CAST(purchase_price AS FLOAT64)
```

Page No.

BigQuery exercise

```
SELECT CAST(purchase_price AS FLOAT64)
FROM customer_data.customer_purchase
ORDER BY CAST(purchase_price AS FLOAT64) DESC;
```

---

CAST, CONCAT, COALESCE

```
--Purchases occured during December promotions

SELECT date, purchase_price
FROM custome_data.customer_purchase
WHERE date BETWEEN '2020-12-01'
                AND '2020-12-31';
```

↓

2020-12-12 T00:00:00
2020-12-28 T00:00:00

↓

This field looks odd because it's datetime we need to convert it to "date" only

We'll use the CAST function again

```
SELECT CAST(date AS date) AS date_only, p_price
FROM customer_data.customer_purchase
WHERE date
BETWEEN '2020-12-01' AND '2020-12-31';
```

* CONCATC) : Adds strings together to create new text strings that can be used as unique keys.

```
SELECT  CONCATC product_code , product_color) AS new_pro_code
FROM    customer_data.customer_purchase
WHERE   product = 'couch'
```

↓

output: SKU 31871grey
SKU 31871 white
SKU 31871 Blue
:

we can guage if people prefer one color or more & find which color is more popular.

---

* COALESCE (): Returns non-null values in a list

we want list of product names but if names aren't available (null) give us the product_code

```
SELECT  COALESCE (product , product_code)
FROM    customer_data.customer_purchase ;
```

---

COALESCE ( first , if_not_first)

↓

can save a lot of time on calculations to handle null values. /

Week 4

Y4

# Verifying & Reporting Results

A semicolon could shape/change the fate of an Org

* Verification:
An process to confirm that a data cleaning effort was well-executed and the resulting data is accurate and reliable.

Involves:
> Rechecking dataset
> Manual cleaning if needed
> Rethinking/focusing on objective

Most important aspect here is "Reporting on your Results"

Reporting:
> Opportunity to show stakeholders you're accountable
> Build trust with your team
> Being on same page about imp. details.

---

Coming Up

* Different strategies for reporting
* Creating data cleaning reports
* Documenting cleaning process
* Using Changelog

Changelog

A file containing a chronologically ordered list of modifications made to a project.

Usually organized by version & includes date followed by list of added improved and removed features.

Page No.
Date 02/02/23

## Verification

Step 01 :- Going back to original dataset and comparing with what you have now!
Review data & find common points

- (1) you may have had lots of nulls
  ↓
  check your cleaned dataset, that no nulls are present.
  ↓
  could search through data manually or use conditional formatting / filter
- (2) Maybe there was common spelling mistake (miss)
  ↓
  find the common miss-spelling & make sure it corrected

* Taking a big-picture view of the project to focus on the objective you're trying to reach.
In involves 3 things :→
  (1) Consider the business problem
  (2) Consider the goal
  (3) consider the data

> Sometimes, an analyst may lose sight of what he needs to do!
In those situations big-picture thinking is important
> Get feedbacks from colleagues
> Too familiar makes easy to miss something.

week 4
3/4

Page No.
Date

* final step in Data Cleaning :→

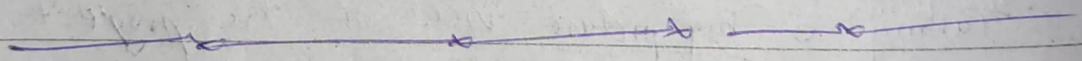
Working with Pivot Table.

what to DO:

when we encounter type
(1) find & Replace
(2) Pivot Table (Insert → Pivot Table)
↓

create Pivot Tables for focused / specific range for ease of working

COUNTA:
A functions that counts total number of values within a specified range.



In SQL

CASE statement:
The CASE statement goes through one or more conditions and returns a value as soon as a condition is met.

```
SELECT CASE
        WHEN condn THEN Soln
        WHEN ——————
        ELSE ——————
        END AS col
FROM
```

* Big Query has personal History & Pretty AMAZING
* Spreadsheets have version History

Ex.

```
SELECT customer_id
       CASE
            WHEN fname = 'Tonny' THEN 'Tony'
            ELSE fname
       END As Cleaned_Data
FROM customer_data.customer_name
```

---

* Documentation ↔
The process of tracking changes, additions, deletions and errors involved in your data-clearing effort

It should involve these 3 things:-
> Recalling data-cleaning Error
> Inform other users of changes
> Determine quality of data (help)

This is EPIC

* Sheet's version history : provides real-time tracker of all change & who made them from individual cells to entire worksheets

---

* Soft companies have their own softweres to keep track of change logs
> All you have to do is mention why you made the query you did. when you commit a query to the repo
> This let's company revert back to previous versions
> You could add comments too

Page No.
Date

## Why documentation is important?

- Documentation is the process of tracking changes, additions, deletions and errors involved in data cleaning.

- A changelog provides realtime account of every modification chronologically.

① Ways to create documentation :→

① List out the steps you took
> Remove duplicates,
> 33 → 32 observations
Sum(membership) decreased by 500$

in SQL we could add these as a comment

THIS IS BEING 100% TRANSPARENT about our data cleaning process.

---

Getting Feedback and using it :→

With consistent documentation and reporting we can uncover error patterns in data collection and entry procedures.

& Do take measures!

(*) Once Errors are identified & addressed, stakeholders can trust data for decision-making process.

Week 5
1/

# The DA Hiring process

Page No.
Date: 03 | 02 | 23

> Application Process
> Write or adjust your resume
> Examples
> Different types of DA jobs

---

Application Process :-

Step 1
> Job sites
> Company website

& search your interest
↓
Research on it

Have a MASTER Resume to tweak for different positions.

Most challenging part: Hearing "NO"

> People you reach out to might not be able to help you
> Companies you would love to work for might not have any openings.
> Jobs you apply for might be filled by someone else.

It's the part of Process, Don't be discouraged!
Be focused!

* Congrat's its a "YES"
> First point of contact is Recruiter
> Recruiter might reach out based on their Research
> Be professional & personable when you talk with the recruiter (zoom/call, etc).

① Recruiter
② Hiring Manager

> Using technical terms in interview help. terms like 'SQL', 'clean_data', etc

would show the recruiter you know what you're doing

* Recruiters don't go about ins & outs of job but they do want to know your capability.
↓
may give you prep material or other Ecoms

* Hiring Manager
> They see if you have the ability to work and if you are good fit for the team.
↓
Convince them "you're the Guy"

The more info you have on job the more chances of you getting it.
↓
If you're seen as fit you'll have atleast one more interview
↓
For giving future stakeholders & team mates a chance to decide if you're best candidate

If all goes well you'll get an official offer. first PHONE & then OFFICIAL letter
[Celebrate]

Find Balance bet^n what they give what you want & what is fair

Page No.
Date

Make sure its a competitive offer

> If they reach out to you, that means they want you as much as you want them

↓

If you're interviewing at other places, you can leverage this to figure out if negotiating for a more competitive offer is possible.

> You should research > Salaries
> Benefits
> vacation time and
> other imp factors (ace to you)

Company X gives y amount more for same job

↓ If you can show this kind of research

↓

There's usually room to negotiate your salary, vacation days or something else.

---

If everything set & done you're all set to start But wait, atleast 2 weeks before you officially start As its customary & polite to give at least a two-week notice at your old job before starting a new one

+

It's good to give yourself a break before starting new Adventure

Page No.
Date

# Resumes

- Brief
- Each Description in few Bullet points (2/4 enough)
  ↓ concise bullets
- One page will make you stick to who you are/ professionally what you wannabe
  ↓
  Only they can look 1 page (busy)

Contact Information : At the Top

- Name
- Address
- Phone Num (use most reachable & sound professional)
- Email Address (not multiple)
  ↓ Ex Janedoe17@email.com

You do have the right to keep anywhere

↓ Should match with online details

---

A format that focuses more on skills and qualifications is good for:

> People with gaps in work history
> Starting Out / Career Change

order ↦
> Most recent
> Earlier
> Earlier than that

Page No.
Date

additional

↓

* Summary :→ If you don't have relevant experience (career transition)

↓

Keep it to highlight your strengths & how it'll help the company you're applying to

↓

Includes positive words about yourself "dedicated" "proactive" support with numbers like number of years you've worked / tools you're experienced with SQL, spreadsheets

Ex :→
Hardworking customer service representative with over five years of experience

Ex :→
Entry-level data Analytics professional; recently completed the GDAPC

Change a little depending upon jobs.

Work Exp: > Jobs with companies
> Add volunteer / freelance positions if any

Key: Way in which you describe
Relate it for the position you're applying for

Clearly state qualifications, (preferred qualification)

Page No.
Date 04 02 23

(*) Resume should show :→

> You're a clear communicator (Be Direct & coherant)

First audience will be :→ Hiring manager & Recruiters

PAR :→ Problem
Action
Result

In Job desc. or Skills section

X Was responsible for writing two blogs a month

✓ Earned little-known website over 2,000 organic clicks through strategic bloging

Problem : little know websit
Action : strategic bloging
Result : 2000 new clicks

Talk about your technical exp:

Spreadsheets
SQL
Tableau
R

Skills/Qualifications > Prog. Langs
> SQL (Also can add top functions & packages youre comfortable with).
> R
> Python
> Excel (Pivot tables, etc)

* Translating past work Experience :→

> If you don't have work history
That's okay!
You can adjust your qualification and skills section

* Transferable skills :→
skills & qualities that can transfer from one job or industry to another.

① Communication skills
Something like :> Effectively implemented and communicated daily workflow procedures to fellow team members, resulting in an increase in productivity
15% increase ↑

* Increase Quantitative data if possible

② Problem-Solving

P
A
R

Problem: Previously absent workflow procedure
Action & Implement & comm. daily workflow procedure
Result: 15% increase in productivity.

Teamwork — Not only what your part of but the whole Org

Finding a Job is great,
Finding a Job you love is even Better!

Page
Date

* Soft skills :→
Non-technical traits & behaviours that relate to how you work.
  - → Detail oriented
  - → Perseverance

---

> Keep your interest in mind while searching for job

> You might apply for jobs you've background in!

GDAPC will be mostly applicable for Junior/Associate

Junior / Associate Data Analyst
- Healthcare Analyst (electronic health records)
- Marketing Analyst (quantitative & qualitative)
- BI Analyst (Increase profits)
- Financial Analyst (Rec. investment opportunity)

---

HYPED UP!