/\*

a Data cleaning steps

1. Remove unnecessary columns by only selecting the ones we need

2. Extract the Youtube Channel names from the first columns

3. Rename the column names

\*/

/\* step 1 \*/

/\*

SELECT

NOMBRE,

total\_subscribers,

total\_videos,

total\_views

FROM

top\_uk\_youtubers\_data\_2024

\*/

-- this will display only the mentioned columns

/\* step 2 \*/

-- Charindex function

-- contains 2 arg, first the targetted char, 2nd the whole string

-- SELECT CHARINDEX('nice','i went to a nice beach') as pos\_of\_n\_in\_nice

-- this will show 13 as n is in the 13th position

-- we need to find the @ in all the names then we can extract whats on the left side to get the names

-- SELECT CHARINDEX('@', NOMBRE), NOMBRE FROM top\_uk\_youtubers\_data\_2024

-- this will show the position of the @ from the names in NOMBRE column

-- Substring

-- this takes 3 arg, first the column from where data will be taken, 2nd the starting point, 3rd the ending point

-- SELECT SUBSTRING(NOMBRE, 1, CHARINDEX('@', NOMBRE) -1) FROM top\_uk\_youtubers\_data\_2024

-- here the -1 will elimnate the @ symbol coming, bt this is not enuf we need to cast this into string type in sql server and for that

SELECT

CAST(SUBSTRING(NOMBRE, 1, CHARINDEX('@', NOMBRE) -1) as varchar(100)) as channel\_name,

total\_subscribers,

total\_videos,

total\_views

FROM top\_uk\_youtubers\_data\_2024

-- the hundred implies that the string limit is 100 in the channel names, we cant go over that and we named the column and selected the other columns as well

/\*

Data Quality Checks

1. The data needs to be 100 records of Youtube Channels (row count checks)

2. The data needs 4 fields (column count test)

3. The channel name column must be string format, and the other data need to be numerical data type (data type check)

4. Each record must be unique in the dataset (duplicate count check)

check 1: Row Count - 100 (passed)

check 2: Column count - 4 (passed)

DATA TYPES

check 3: Channel name = VARCHAR (passed)

total subscribers, views, videos = INTEGER

check 4: Duplicate Count = 0 (passed)

\*/

-- ROW COUNT CHECK

SELECT

COUNT(\*) as no\_of\_rows

FROM

view\_uk\_youtubers\_2024

-- this will give 100

-- COLUMN COUNT CHECK

SELECT

COUNT(\*) as column\_count

FROM

INFORMATION\_SCHEMA.COLUMNS

-- this basically knows all information about the database, columns rows etc etc, and the .column will check the column specific

WHERE

Table\_name = 'view\_uk\_youtubers\_2024'

-- this will give 4

-- DATA TYPE CHECK

SELECT

COLUMN\_NAME, -- this names we got when ww ran the prev function with only: \* in place of: COUNT(\*) as column\_count

DATA\_TYPE

FROM

INFORMATION\_SCHEMA.COLUMNS

WHERE

TABLE\_NAME = 'view\_uk\_youtubers\_2024'

-- this will view the data types

-- DUPLICATE CHECK

SELECT

channel\_name,

COUNT(\*) as duplicate\_count

FROM

view\_uk\_youtubers\_2024

GROUP BY

channel\_name

HAVING

COUNT(\*) > 1 -- having pairing more than 1

-- this returns nothing which means no duplicate