Q.1 : Assume you're given the tables below about Facebook Page and Page likes (as in "Like a Facebook Page").

Write a query to return the IDs of the Facebook pages which do not possess any likes. The output should be sorted in ascending order.

`pages` Table:

Column Name	Type
page_id	integer
page_name	varchar

`pages` Example Input:

page_id	page_name
20001	SQL Solutions
20045	Brain Exercises
20701	Tips for Data Analysts

`page_likes` Table:

Column Name	Type
user_id	integer
page_id	integer
liked_date	datetime

`page_likes` Example Input:

user_id	page_id	liked_date
111	20001	04/08/2022 00:00:00
121	20045	03/12/2022 00:00:00
156	20001	07/25/2022 00:00:00

Q.2: Tesla is investigating production bottlenecks and they need your help to extract the relevant data. Write a query that determines which parts with the assembly steps have initiated the assembly process but remain unfinished.
Assumptions:

- `parts_assembly` table contains all parts currently in production, each at varying stages of the assembly process.
- An unfinished part is one that lacks a `finish_date`.

This question is straightforward, so let's approach it with simplicity in both thinking and solution.

Effective April 11th 2023, the problem statement and assumptions were updated to enhance clarity.

`parts_assembly` Table

Column Name	Type
part	string
finish_date	datetime
assembly_step	integer

`parts_assembly` Example Input

part	finish_date	assembly_step
battery	01/22/2022 00:00:00	1
battery	02/22/2022 00:00:00	2
battery	03/22/2022 00:00:00	3
bumper	01/22/2022 00:00:00	1
bumper	02/22/2022 00:00:00	2
bumper		3
bumper		4

Example Output

part	assembly_step
bumper	3

bumper	4

Q.3: Assume you're given a table Twitter tweet data, write a query to obtain a histogram of tweets posted per user in 2022. Output the tweet count per user as the bucket and the number of Twitter users who fall into that bucket.
In other words, group the users by the number of tweets they posted in 2022 and count the number of users in each group.

`tweets` Table:

Column Name	Type
tweet_id	integer
user_id	integer
msg	string
tweet_date	timestamp

`tweets` Example Input:

tweet_id	user_id	msg	tweet_date
214252	111	Am considering taking Tesla private at $420. Funding secured.	12/30/2021 00:00:00
739252	111	Despite the constant negative press covfefe	01/01/2022 00:00:00
846402	111	Following @NickSinghTech on Twitter changed my life!	02/14/2022 00:00:00
241425	254	If the salary is so competitive why won’t you tell me what it is?	03/01/2022 00:00:00
231574	148	I no longer have a manager. I can't be managed	03/23/2022 00:00:00

Example Output:

tweet_bucket	users_num
1	2
2	1

Q.4: Assume you're given the table on user viewership categorised by device type where the three types are laptop, tablet, and phone.
Write a query that calculates the total viewership for laptops and mobile devices where mobile is defined as the sum of tablet and phone viewership. Output the total viewership for laptops as `laptop_reviews` and the total viewership for mobile devices as `mobile_views`.

Effective 15 April 2023, the solution has been updated with a more concise and easy-to-understand approach.

`viewership` Table

Column Name	Type
user_id	integer
device_type	string ('laptop', 'tablet', 'phone')
view_time	timestamp

`viewership` Example Input

user_id	device_type	view_time
123	tablet	01/02/2022 00:00:00
125	laptop	01/07/2022 00:00:00
128	laptop	02/09/2022 00:00:00
129	phone	02/09/2022 00:00:00
145	tablet	02/24/2022 00:00:00

Example Output

laptop_views	mobile_views
2	3

Q.5: Given a table of candidates and their skills, you're tasked with finding the candidates best suited for an open Data Science job. You want to find candidates who are proficient in Python, Tableau, and PostgreSQL.
Write a query to list the candidates who possess all of the required skills for the job. Sort the output by candidate ID in ascending order.

Assumption:

- There are no duplicates in the `candidates` table.

`candidates` Table:

Column Name	Type
candidate_id	integer
skill	varchar

`candidates` Example Input:

candidate_id	skill
123	Python
123	Tableau
123	PostgreSQL
234	R
234	PowerBI
234	SQL Server
345	Python
345	Tableau

Example Output:

candidate_id
123

Q.6: Given a table of Facebook posts, for each user who posted at least twice in 2021, write a query to find the number of days between each user's first post of the year and last post of the year in the year 2021. Output the user and number of the days between each user's first and last post.

`posts` Table:

Column Name	Type
user_id	integer
post_id	integer
post_date	timestamp
post_content	text

`posts` Example Input:

user_id	post_id	post_date	post_content
151652	599415	07/10/2021 12:00:00	Need a hug
661093	624356	07/29/2021 13:00:00	Bed. Class 8-12. Work 12-3. Gym 3-5 or 6. Then class 6-10. Another day that's gonna fly by. I miss my girlfriend
004239	784254	07/04/2021 11:00:00	Happy 4th of July!
661093	442560	07/08/2021 14:00:00	Just going to cry myself to sleep after watching Marley and Me.
151652	111766	07/12/2021 19:00:00	I'm so done with covid - need travelling ASAP!

Example Output:

user_id	days_between
151652	2

661093	21

Q.7: Write a query to identify the top 2 Power Users who sent the highest number of messages on Microsoft Teams in August 2022. Display the IDs of these 2 users along with the total number of messages they sent. Output the results in descending order based on the count of the messages.
Assumption:

- No two users have sent the same number of messages in August 2022.

`messages` Table:

Column Name	Type
message_id	integer
sender_id	integer
receiver_id	integer
content	varchar
sent_date	datetime

`messages` Example Input:

message_id	sender_id	receiver_id	content	sent_date
901	3601	4500	You up?	08/03/2022 00:00:00
902	4500	3601	Only if you're buying	08/03/2022 00:00:00
743	3601	8752	Let's take this offline	06/14/2022 00:00:00
922	3601	4500	Get on the call	08/10/2022 00:00:00

Example Output:

sender_id	message_count

3601	2
4500	1

Q.8: Assume you are given the table below that shows job postings for all companies on the LinkedIn platform. Write a query to get the number of companies that have posted duplicate job listings.
Clarification:

- Duplicate job listings refer to two jobs at the same company with the same title and description.

`job_listings` Table:

Column Name	Type
job_id	integer
company_id	integer
title	string
description	string

`job_listings` Example Input:

job_id	company_id	title	description
248	827	Business Analyst	Business analyst evaluates past and current business data with the primary goal of improving decision-making processes within organizations.
149	845	Business Analyst	Business analyst evaluates past and current business data with the primary goal of improving decision-making processes within organizations.
945	345	Data Analyst	Data analyst reviews data to identify key insights into a business's customers and ways the data can be used to solve problems.

164	345	Data Analyst	Data analyst reviews data to identify key insights into a business's customers and ways the data can be used to solve problems.
172	244	Data Engineer	Data engineer works in a variety of settings to build systems that collect, manage, and convert raw data into usable information for data scientists and business analysts to interpret.

Example Output:

co_w_duplicate_jobs
1

Q.9:Assume you're given the tables containing completed trade orders and user details in a Robinhood trading system.
Write a query to retrieve the top three cities that have the highest number of completed trade orders listed in descending order. Output the city name and the corresponding number of completed trade orders.

`trades` Table:

Column Name	Type
order_id	integer
user_id	integer
price	decimal
quantity	integer
status	string('Completed' ,'Cancelled')
timestamp	datetime

`trades` Example Input:

order_id	user_id	price	quantity	status	timestamp
100101	111	9.80	10	Cancelled	08/17/2022 12:00:00

100102	111	10.00	10	Completed	08/17/2022 12:00:00
100259	148	5.10	35	Completed	08/25/2022 12:00:00
100264	148	4.80	40	Completed	08/26/2022 12:00:00
100305	300	10.00	15	Completed	09/05/2022 12:00:00
100400	178	9.90	15	Completed	09/09/2022 12:00:00
100565	265	25.60	5	Completed	12/19/2022 12:00:00

`users` Table:

Column Name	Type
user_id	integer
city	string
email	string
signup_date	datetime

`users` Example Input:

user_id	city	email	signup_date
111	San Francisco	rrok10@gmail.com	08/03/2021 12:00:00
148	Boston	sailor9820@gmail.com	08/20/2021 12:00:00
178	San Francisco	harrypotterfan182@gmail.com	01/05/2022 12:00:00
265	Denver	shadower_@hotmail.com	02/26/2022 12:00:00
300	San Francisco	houstoncowboy1122@hotmail.com	06/30/2022 12:00:00

Example Output:

city	total_orders
San Francisco	3
Boston	2
Denver	1

Q.10: Given the reviews table, write a query to retrieve the average star rating for each product, grouped by month. The output should display the month as a numerical value, product ID, and average star rating rounded to two decimal places. Sort the output first by month and then by product ID.

`reviews` Table:

Column Name	Type
review_id	integer
user_id	integer
submit_date	datetime
product_id	integer
stars	integer (1-5)

`reviews` Example Input:

review_id	user_id	submit_date	product_id	stars
6171	123	06/08/2022 00:00:00	50001	4
7802	265	06/10/2022 00:00:00	69852	4
5293	362	06/18/2022 00:00:00	50001	3
6352	192	07/26/2022 00:00:00	69852	3
4517	981	07/05/2022 00:00:00	69852	2

Example Output:

mth	product	avg_stars
6	50001	3.50
6	69852	4.00
7	69852	2.50

Q.11: Assume you have an events table on Facebook app analytics. Write a query to calculate the click-through rate (CTR) for the app in 2022 and round the results to 2 decimal places.
Definition and note:

- Percentage of click-through rate (CTR) = 100.0 * Number of clicks / Number of impressions
- To avoid integer division, multiply the CTR by 100.0, not 100.

`events` Table:

Column Name	Type
app_id	integer
event_type	string
timestamp	datetime

`events` Example Input:

app_id	event_type	timestamp
123	impression	07/18/2022 11:36:12
123	impression	07/18/2022 11:37:12
123	click	07/18/2022 11:37:42
234	impression	07/18/2022 14:15:12
234	click	07/18/2022 14:16:12

Example Output:

app_id	ctr
123	50.00
234	100.00

Q.12: Assume you're given tables with information about TikTok user sign-ups and confirmations through email and text. New users on TikTok sign up using their email addresses, and upon sign-up, each user receives a text message confirmation to activate their account.
Write a query to display the user IDs of those who did not confirm their sign-up on the first day, but confirmed on the second day.

Definition:

- `action_date` refers to the date when users activated their accounts and confirmed their sign-up through text messages.

`emails` Table:

Column Name	Type
email_id	integer
user_id	integer
signup_date	datetime

`emails` Example Input:

email_id	user_id	signup_date
125	7771	06/14/2022 00:00:00
433	1052	07/09/2022 00:00:00

`texts` Table:

Column Name	Type
text_id	integer
email_id	integer
signup_action	string ('Confirmed', 'Not confirmed')
action_date	datetime

`texts` Example Input:

text_id	email_id	signup_action	action_date
6878	125	Confirmed	06/14/2022 00:00:00
6997	433	Not Confirmed	07/09/2022 00:00:00
7000	433	Confirmed	07/10/2022 00:00:00

Example Output:

user_id
1052

Q.13: Your team at JPMorgan Chase is soon launching a new credit card, and to gain some context, you are analyzing how many credit cards were issued each month. Write a query that outputs the name of each credit card and the difference in issued amount between the month with the most cards issued, and the least cards issued. Order the results according to the biggest difference.

`monthly_cards_issued` Table:

Column Name	Type
issue_month	integer
issue_year	integer
card_name	string
issued_amount	integer

`monthly_cards_issued` Example Input:

card_name	issued_amount	issue_month	issue_year
Chase Freedom Flex	55000	1	2021
Chase Freedom Flex	60000	2	2021
Chase Freedom Flex	65000	3	2021

Chase Freedom Flex	70000	4	2021
Chase Sapphire Reserve	170000	1	2021
Chase Sapphire Reserve	175000	2	2021
Chase Sapphire Reserve	180000	3	2021

Example Output:

card_name	difference
Chase Freedom Flex	15000
Chase Sapphire Reserve	10000

Q.14: You're trying to find the mean number of items per order on Alibaba, rounded to 1 decimal place using tables which includes information on the count of items in each order (`item_count` table) and the corresponding number of orders for each item count (`order_occurrences` table).

`items_per_order` Table:

Column Name	Type
item_count	integer
order_occurrences	integer

`items_per_order` Example Input:

item_count	order_occurrences
1	500
2	1000
3	800
4	1000

There are a total of 500 orders with one item per order, 1000 orders with two items per order, and 800 orders with three items per order."

Example Output:

mean
2.7

Q.15: CVS Health is trying to better understand its pharmacy sales, and how well different products are selling. Each drug can only be produced by one manufacturer.
Write a query to find the top 3 most profitable drugs sold, and how much profit they made. Assume that there are no ties in the profits. Display the result from the highest to the lowest total profit.

Definition:

- `cogs` stands for Cost of Goods Sold which is the direct cost associated with producing the drug.
- Total Profit = Total Sales - Cost of Goods Sold

`pharmacy_sales` Table:

Column Name	Type
product_id	integer
units_sold	integer
total_sales	decimal
cogs	decimal
manufacturer	varchar
drug	varchar

`pharmacy_sales` Example Input:

product_id	units_sold	total_sales	cogs	manufacturer	drug
9	37410	293452.54	208876.01	Eli Lilly	Zyprexa
34	94698	600997.19	521182.16	AstraZeneca	Surmontil
61	77023	500101.61	419174.97	Biogen	Varicose Relief
136	144814	1084258	1006447.73	Biogen	Burkhart

Example Output:

drug	total_profit
Zyprexa	84576.53
Varicose Relief	80926.64
Surmontil	79815.03

Q.16: CVS Health is analyzing its pharmacy sales data, and how well different products are selling in the market. Each drug is exclusively manufactured by a single manufacturer. Write a query to identify the manufacturers associated with the drugs that resulted in losses for CVS Health and calculate the total amount of losses incurred.

Output the manufacturer's name, the number of drugs associated with losses, and the total losses in absolute value. Display the results sorted in descending order with the highest losses displayed at the top.

`pharmacy_sales` Table:

Column Name	Type
product_id	integer
units_sold	integer
total_sales	decimal
cogs	decimal
manufacturer	varchar
drug	varchar

`pharmacy_sales` Example Input:

product_id	units_sold	total_sales	cogs	manufacturer	drug
156	89514	3130097.00	3427421.73	Biogen	Acyclovir
25	222331	2753546.00	2974975.36	AbbVie	Lamivudine and Zidovudine

50	90484	2521023.73	2742445.90	Eli Lilly	Dermasorb TA Complete Kit
98	110746	813188.82	140422.87	Biogen	Medi-Chord

Example Output:

manufacturer	drug_count	total_loss
Biogen	1	297324.73
AbbVie	1	221429.36
Eli Lilly	1	221422.17

Q.17: CVS Health is trying to better understand its pharmacy sales, and how well different products are selling.
Write a query to find the total drug sales for each manufacturer. Round your answer to the closest million, and report your results in descending order of total sales.

Because this data is being directly fed into a dashboard which is being seen by business stakeholders, format your result like this: "$36 million".

`pharmacy_sales` Table:

Column Name	Type
product_id	integer
units_sold	integer
total_sales	decimal
cogs	decimal

manufacturer	varchar
drug	varchar

`pharmacy_sales` Example Input:

product_id	units_sold	total_sales	cogs	manufacturer	drug
94	132362	2041758.41	1373721.70	Biogen	UP and UP
9	37410	293452.54	208876.01	Eli Lilly	Zyprexa
50	90484	2521023.73	2742445.9	Eli Lilly	Dermasorb
61	77023	500101.61	419174.97	Biogen	Varicose Relief
136	144814	1084258.00	1006447.73	Biogen	Burkhart

Example Output:

manufacturer	sale
Biogen	$4 million
Eli Lilly	$3 million

Q.18: UnitedHealth has a program called Advocate4Me, which allows members to call an advocate and receive support for their health care needs – whether that's behavioural, clinical, well-being, health care financing, benefits, claims or pharmacy help.
Write a query to find how many UHG members made 3 or more calls. `case_id` column uniquely identifies each call made.

`callers` Table:

Column Name	Type
policy_holder_id	integer
case_id	varchar
call_category	varchar
call_received	timestamp
call_duration_secs	integer
original_order	integer

`callers` Example Input:

policy_holder_id	case_id	call_category	call_received	call_duration_secs	original_order
50837000	dc63-acae-4f39-bb04	claims	03/09/2022 02:51:00	205	130
50837000	41be-bebe-4bd0-a1ba	IT_support	03/12/2022 05:37:00	254	129
50936674	12c8-b35c-48a3-b38d	claims	05/31/2022 7:27:00	240	31
50886837	d0b4-8ea7-4b8c-aa8b	IT_support	03/11/2022 3:38:00	276	16
50886837	a741-c279-41c0-90ba		03/19/2022 10:52:00	131	325

50837000	bab1-3ec5-4867-90ae	benefits	05/13/2022 18:19:00	228	339

Example Output:

member_count
1

Q.19: UnitedHealth Group has a program called Advocate4Me, which allows members to call an advocate and receive support for their health care needs – whether that's behavioural, clinical, well-being, health care financing, benefits, claims or pharmacy help. Calls to the Advocate4Me call centre are categorised, but sometimes they can't fit neatly into a category. These uncategorised calls are labelled "n/a", or are just empty (when a support agent enters nothing into the category field).

Write a query to find the percentage of calls that cannot be categorised. Round your answer to 1 decimal place.

`callers` Table:

Column Name	Type
policy_holder_id	integer
case_id	varchar
call_category	varchar
call_received	timestamp
call_duration_secs	integer
original_order	integer

`callers` Example Input:

policy_holder_id	case_id	call_category	call_received	call_duration_secs	original_order
52481621	a94c-2213-4ba5-812d		01/17/2022 19:37:00	286	161
51435044	f0b5-0eb0-4c49-b21e	n/a	01/18/2022 2:46:00	208	225
52082925	289b-d7e8-4527-bdf5	benefits	01/18/2022 3:01:00	291	352
54624612	62c2-d9a3-44d2-9065	IT_support	01/19/2022 0:27:00	273	358
54624612	9f57-164b-4a36-934e	claims	01/19/2022 6:33:00	157	362

Example Output:

call_percentage
40.0