Data Analyst Technical Interview Test

(If it's more convenient, you can use Sql extensions like Presto Sql, postgresQL, MS-SQL etc..)

Question 1-4 are based on the following 2 tables:

'Industries' table (Column Names are 'company_site_id', 'industry' and 'date')

company_site_id	industry	date
alexeastman.uk	Marketing	2022-01-31
bellacosta.net	Restaurant	2022-01-24
cupidhomesandsupportservices.co.uk	Marketing	2022-01-31
cybercolloids.net	Software	2022-01-24
glamcamps.ca	Software	2022-02-07
interpetcentral.com	Retail	2022-01-24
midwestxposure.com	Marketing	2022-01-31
selvanegrarestaurant.com	Restaurant	2022-01-31
shopmeaway.co.uk	Retail	2022-01-24
steelhorsellc.com	Construction	2022-01-31
taranash.com	Retail	2022-02-07
wedbridalboutique.com	Retail	2022-02-07

'Countries' table (Column Names are 'company_site_id', 'country_code' and 'date')

company_site_id	country_c	date
alexeastman.uk	UK	2022-02-07
canadianstroke.com	CA	2022-01-31
cupidhomesandsupportservices.co.uk	UK	2022-02-07
cybercolloids.net	IE	2022-01-24
fatissmartfuel.com	US	2022-01-24
geratek.qc.ca	CA	2022-02-07
glamcamps.ca	CA	2022-01-24
interpetcentral.com	UK	2022-01-31
jma.ie	IE	2022-02-07
midwestxposure.com	US	2022-02-07
shopmeaway.co.uk	US	2022-01-24
wedbridalboutique.com	UK	2022-01-31

Question 1

Write an SQL query, showing all industry names that start with an 'R'

—----------------------------

```
SELECT DISTINCT industry
FROM Industries
WHERE industry LIKE 'R%';
```

Question 2

Write an SQL query, showing all domains, their industry and their country that are Retail companies.

—----------------------------

```
SELECT co.company_site_id, i.industry, co.country_code
FROM Industries i
JOIN Countries co ON i.company_site_id = co.company_site_id
WHERE i.industry = 'Retail';
```

Question 3

Write an SQL query, showing the total numbers of 'company_site_id's for each industry in the industries table. (output should be 2 columns: industry name and number of site_id of this industry)

—----------------------------

```
SELECT industry, COUNT(DISTINCT company_site_id) AS number_of_sites
FROM Industries
GROUP BY industry;
```

Question 4

Using CTE / Temp Table (WITH temp_table AS (…)), create a temp table including all the domains and countries from the countries table for date '2022-01-31' and create another temp table with the same data but for date '2022-02-07'. Use the temp tables to return the domains and countries from both tables using a UNION statement.

—----------------------------

```
WITH TempTable1 AS (
    SELECT company_site_id, country_code
    FROM Countries
    WHERE date = '2022-01-31'
)
, TempTable2 AS (
    SELECT company_site_id, country_code
```

```
    FROM Countries
    WHERE date = '2022-02-07'
)
SELECT company_site_id, country_code
FROM TempTable1
UNION
SELECT company_site_id, country_code
FROM TempTable2;
```

Question 5

Listed below are 3 Regular Expression Rules:
"\b": Word boundary - marks the beginning/ending of the word = '\bstring\b'
Ex. \btext\b => 'text' only, NOT 'texts', NOT 'texting' etc.

"?": Optional addition to string
Ex1. 'apples?' => either 'apple' or 'apples'
Ex2. 'text(ing|s|ed)?' => 'text', 'texting', 'texts', 'texted'

"\W": Symbolizes text separators in text like space, underscore (ASCII letter, digit or underscore) in regex

Ex1. 'bed&breakfast' => 'bed\Wbreakfast'
Ex2. 'on\W?boarding' => 'on-boarding', 'on boarding', "on_boarding, 'onboarding'

Based on these rules, please write a pattern that will indicate both 'sub-title' and 'subtitles' words.

For example
- These texts should return no match (empty result) : “hello everyone how are you? ” , “this title is not good enough”
- These texts should return a match (empty result) : “you should to read the subtitles Mr. Jones ” , “can you add a sub-title to your article ”

—----------

```
\b(sub\W?titles?)\b
```

Question 6

This are the tables for this question

comapnies	
id	string Unique id of the company
state	string The state of which the company is operating
rev	integer anual revenue (M$)
no_of_emp	integer Number of employees
ind_code	integer The code of the industry of the company
clevel	
id	string Unique id of the company
c_name	string The name of the C-level
email	string the email of the C-level
d_cease	date The date of termination of the C-level
inds	
name	string The name of the industry of the company
code	integer The code of the industry of the company

You need to prepare an SQL query that will return the best scope for your new client - Lose-a-Fortune Bank.
They asked for their scope to include three industries that have between 10 to 14 in average revenue.
They prefer industries with a lower revenue in that range.

The query returns a table with three columns: industry_name, num_of_comp, average_rev
Please write the relevant SQL code:

—----------

```
SELECT
  inds.name AS industry_name,
  COUNT(comapnies.id) AS num_of_comp,
  AVG(comapnies.rev) AS average_rev
FROM comapnies
INNER JOIN inds
  ON comapnies.ind_code = inds.code
```

```
WHERE comapnies.rev BETWEEN 10 AND 14
GROUP BY inds.name
ORDER BY average_rev ASC
LIMIT 3;
```

Bonus Question

This is the table for this question.
There is also a sample data in the attached csv for your convenience

Table companies_details

tarci_id	revenue	no_of_employees	incorporation_date	main_address	alternative_address	industry
000021f3-bf9f-48						Real Estate
00004750-ca48-		11-50 employees			Higher Whitnell Farm	Restaurants
00005b65-d2a0-						Retail
00005fcd-43a5-4						Automotive
000073d6-142e-	0-500k	1-10 employees	02/04/2014	63 Hatton Garden Fifth Floor Ste 23, London, EC1N 8LE ENG, United Kingdom	63 Hatton Garden Fifth Floor S	Energy
000073d6-142e-	0-500k	1-10 employees	02/04/2014	63 Hatton Garden Fifth Floor Ste 23, London, EC1N 8LE ENG, United Kingdom	South Clapham	BI
000073d6-142e-	0-500k	1-10 employees	02/04/2014	63 Hatton Garden Fifth Floor Ste 23, London, EC1N 8LE ENG, United Kingdom		Real Estate
0000c57a-cd22-						Restaurants
0000d0b6-8b1b-						Retail
0000e377-5561-						Automotive
0000f309-22a6-4						Energy
000103ed-309e-						BI
00010705-856e-		1-10 employees			Office 83 Whins Road	Real Estate
0001115e-5d10-	5M-10M	1-10 employees	03/14/2013	The Old Forge East Street, Colchester, CO1 2TP ENG, United Kingdom	The Old Forge East Street	Restaurants
0001115e-5d10-	5M-10M	1-10 employees	03/14/2013	The Old Forge East Street, Colchester, CO1 2TP ENG, United Kingdom		Retail
00012d5b-3d3a-						Automotive
000154d9-e2d8-		1-10 employees			Unit 1 Mountbatten Industrial P	Energy
000154d9-e2d8-		1-10 employees				BI
000157bf-5544-4						Real Estate
00015a45-5fc0-4						Restaurants
00015c29-31e2-		1-10 employees				Retail
00015c29-31e2-		1-10 employees			Stockbridge Technology Centre	Automotive
0001731c-25f6-4						Real Estate
00018793-7dcc-						Restaurants
00018c73-6574-	500k-1M	11-50 employees	04/06/2010	Jigsaw Group 7 8 Delta Bank Road Metro Riverside Park, Gateshead Tyne Wear, NE	Tower Road	Retail
00018c73-6574-	500k-1M	11-50 employees	04/06/2010	Jigsaw Group 7 8 Delta Bank Road Metro Riverside Park, Gateshead Tyne Wear, NE	Jigsaw Group 7 8 Delta Bank R	Automotive
00018c73-6574-	500k-1M	11-50 employees	04/06/2010	Jigsaw Group 7 8 Delta Bank Road Metro Riverside Park, Gateshead Tyne Wear, NE11 9DJ ENG, United Kingdom		Energy

Question7:

You need to provide all the details of the top 10 companies from the companies_details table

The priority is:
Existence of main address
Existence of alternative_address
Existence of no_of_employees
For example:
if 000073d6-142e-45c5-9184-9178f8017e63 is included in the top tarci_ids/company then all of it's rows from companies details will be included in the output (rows 6,7,8 in the screenshot)

—----------

```
WITH TopCompanies AS (
    SELECT tarci_id
    FROM companies_details
    ORDER BY
        CASE WHEN main_address IS NOT NULL THEN 1 ELSE 2 END,
        CASE WHEN alternative_address IS NOT NULL THEN 1 ELSE 2 END,
        CASE WHEN no_of_employees IS NOT NULL THEN 1 ELSE 2 END
    LIMIT 10
)
SELECT *
FROM companies_details
WHERE tarci_id IN (SELECT tarci_id FROM TopCompanies);
```

This one is also hot:

```
SELECT * FROM (
  SELECT tarci_id, revenue, no_of_employees, incorporation_date, main_address, alternative_address, industry,
    (CASE WHEN main_address IS NOT NULL THEN 1 ELSE 0 END) +
    (CASE WHEN alternative_address IS NOT NULL THEN 1 ELSE 0 END) +
    (CASE WHEN no_of_employees IS NOT NULL THEN 1 ELSE 0 END) AS priority
  FROM companies_details) AS t
ORDER BY priority DESC, tarci_id ASC
LIMIT 10
```