Problem Statement

- **Domain** – Auto Insurance Claims – Risk Assessment
- **Business Context** – A famous Auto Insurance provider in France is trying to understand the underlying factors associated on making an auto insurance claim for a particular city. The data has been gathered by a third-party vendor from different Auto Insurance providers that has customer demographic information and if they have made claims or not. *You as an analyst are given the dataset to explore and extract important insights from the data which will help the marketing team to better make pricing decisions and assess the underlying risk elements. The data dictionary below provides the details of the dataset*.

- **Data Dictionary** – Table name - **auto_insurance_risk**
  - **IDpol** The policy ID (used to link with the claims dataset).
  - **ClaimNb** Number of claims during the exposure period.
  - **Exposure** The exposure period.
  - **Area** The area code.
  - **VehPower** The power of the car (ordered categorical).
  - **VehAge** The vehicle age, in years.
  - **DrivAge** The driver age, in years (in France, people can drive a car at 18).
  - **BonusMalus** Bonus/malus, between 50 and 350: 100 means malus in France. (https://en.wikipedia.org/wiki/Bonus-malus)
  - **VehBrand** The car brand (unknown categories).
  - **VehGas** The car gas, Diesel or regular.
  - **Density** The density of inhabitants (number of inhabitants per km2) in the city the driver of the car lives in.
  - **Region** The policy regions in France (based on a standard French classification)

- **Objective of the project** – Perform EDA and extract important insights from the data. SQL is an essential tool to efficiently query and get quick insights from the data. Therefore, it is a must have skill for a good data analyst or data scientist. The following questions shall help us in getting the bottom of the data and provide insights to our business stakeholders. Some of the following questions also touch base on the important conceptual aspects of SQL and RDBMS.

Questions –

Project Based

#1. Write a query to calculate what % of the customers have made a claim in the current exposure period[i.e. in the given dataset]? (2)
Hint: There are customers who have claimed more than once and they should be regarded only once in the % calculation.

#2. 2.1. Create a new column as 'claim_flag' in the table 'auto_insurance_risk' as integer datatype. (1.5)
2.2. Set the value to 1 when ClaimNb is greater than 0 and set the value to 0 otherwise. (1.5)

#3. 3.1. What is the average exposure period for those who have claimed? (1)
3.2. What do you infer from the result? (1)
Hint: Use claim_flag variable to group the data.

#4. 4.1. If we create an exposure bucket where buckets are like below, what is the % of total claims by these buckets? (2)
4.2. What do you infer from the summary? (1)
Hint: Buckets are => E1 = 0 to 0.25, E2 = 0.26 to 0.5, E3 = 0.51 to 0.75, E4 > 0.75, You need to consider ClaimNb field to get the total claim count.

#5. Which area has the higest number of average claims? Show the data in percentage w.r.t. the number of policies in corresponding Area. (2)
Hint: Use ClaimNb field for this question.

#6. If we use these exposure bucket along with Area i.e. group Area and Exposure Buckets together and look at the claim rate, an interesting pattern could be seen in the data. What is that? (3)
Note: 2 Marks for SQL and 1 for inference.

#7. 7.1. If we look at average Vehicle Age for those who claimed vs those who didn't claim, what do you see in the summary? (1.5+1 = 2.5)
7.2. Now if we calculate the average Vehicle Age for those who claimed and group them by Area, what do you see in the summary? Any particular pattern you see in the data? (1.5+1=2.5)

#8. If we calculate the average vehicle age by exposure bucket(as mentioned above), we see an interesting trend between those who claimed vs those who didn't. What is that? (3)

#9. 9.1. Create a Claim_Ct flag on the ClaimNb field as below, and take average of the BonusMalus by Claim_Ct. (2)
9.2. What is the inference from the summary? (1)

Note: Claim_Ct = '1 Claim' where ClaimNb = 1, Claim_Ct = 'MT 1 Claims' where ClaimNb > 1, Claim_Ct = 'No Claims' where ClaimNb = 0.

#10. Using the same Claim_Ct logic created above, if we aggregate the Density column (take average) by Claim_Ct, what inference can we make from the summary data?(4)
Note: 2.5 Marks for SQL and 1.5 for inference.

#11. Which Vehicle Brand & Vehicle Gas combination have the highest number of Average Claims (use ClaimNb field for aggregation)? (2)

#12. List the Top 5 Regions & Exposure[use the buckets created above] Combination from Claim Rate's perspective. Use claim_flag to calculate the claim rate. (3)

#13. 13.1. Are there any cases of illegal driving i.e. underaged folks driving and committing accidents? (1)
13.2. Create a bucket on DrivAge and then take average of BonusMalus by this Age Group Category. WHat do you infer from the summary? (2.5+1.5 = 4)
Note: DrivAge=18 then 1-Beginner, DrivAge<=30 then 2-Junior, DrivAge<=45 then 3-Middle Age, DrivAge<=60 then 4-Mid-Senior, DrivAge>60 then 5-Senior

Conceptual

#14. Mention one major difference between unique constraint and primary key? (2)

#15. If there are 5 records in table A and 10 records in table B and we cross-join these two tables, how many records will be there in the result set? (2)

#16. What is the difference between inner join and left outer join? (2)

#17. Consider a scenario where Table A has 5 records and Table B has 5 records. Now while inner joining Table A and Table B, there is one duplicate on the joining column in Table B (i.e. Table A has 5 unique records, but Table B has 4 unique values and one redundant value). What will be record count of the output? (2)

#18. What is the difference between WHERE clause and HAVING clause? (2)