BUSSINESS REPORT

By Abhinav Tyagi

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.

Project Based

Q.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.

QUERY :

```
Select (Count(ClaimNb)*100.0/(SELECT count(IDpol) from Auto_insurance_risk)) as Percentage_Claim
FROM Auto_insurance_risk
where ClaimNb >= 1;
```

OUTPUT:

	Percentage_Claim
1	5.02350249921462

Q.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)

QUERY:

```
ALTER TABLE Auto_insurance_risk ADD claim_flag INTEGER;

UPDATE Auto_insurance_risk
SET claim_flag = 1
WHERE ClaimNb > 1;

UPDATE Auto_insurance_risk
SET claim_flag = 0
WHERE ClaimNb < 1;
```

Q. 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.

QUERY:

```
SELECT claim_flag,avg(Exposure) from Auto_insurance_risk as Average_Exposure
group by claim_flag;
```

OUTPUT:

	claim_flag	avg(Exposure)
1	0	0.522733894817779
2	1	0.642495175948072

INFERENCE:

The average exposure for those who claimed is higher than who not claimed. .

Q 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.

QUERY:

```
'

ALTER TABLE Auto_insurance_risk
ADD exposure_bucket TEXT;

update Auto_insurance_risk
set exposure_bucket = case
when  Exposure >0 AND Exposure<= 0.25 then 'E1'
when  Exposure >=0.26 AND Exposure<=0.50 then 'E2'
when  Exposure >=0.51 AND Exposure<= 0.75 then 'E3'
when  Exposure >=0.75 then 'E4'
end;

select exposure_bucket,(count(ClaimNb)*100.0/(Select count(*)
from Auto_insurance_risk))as Percentage_Of_Claims  from Auto_insurance_risk
group by exposure_bucket;
```

OUTPUT:

	exposure_bucket	Percentage_Of_Claims
1	E1	32.8624967367882
2	E2	19.3657053773305
3	E3	13.6419213200927
4	E4	34.1298765657886

INFERENCE:

E4 exposure_bucket contains more Percentage of Claims.

Q5. 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

QUERY :

```
SELECT Area , avg(claimNB) from Auto_insurance_risk
group by area;
select area , count(claimNb)*100.0/sum(count(claimNb))over() as percentage
from Auto_insurance_risk
group by Area
having ClaimNb = 1;
```

OUTPUT:

	Area	percentage
1	A	15.3325968676117
2	B	11.1294326214984
3	C	28.3003423238197
4	D	22.3588633256294
5	E	20.230733039042
6	F	2.64803182239869

Q. 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.

QUERY:

```
SELECT Area,exposure_bucket,(Count(ClaimNb)* 100 / (Select Count(*) From Auto_insurance_risk))
as Claim_Rate
FROM Auto_insurance_risk
group by Area
order by Claim_Rate DESC;
```

OUTPUT:

	Area	exposure_bucket	Claim_Rate
1	C	E2	28
2	D	E1	22
3	E	E3	20
4	A	E1	15
5	B	E3	11
6	F	E2	2

INFERENCE:
Output shows the higher claim rate in Area C having exposure in between 0.25 to 0.50.

Q 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)

QUERY:

```
select claim_flag, avg(VehAge) from Auto_insurance_risk
group by claim_flag;
```

OUTPUT:

	claim_flag	avg(VehAge)
1	0	7.07291836516019
2	1	6.50252495596007

INFERENCE:

Those who claims have vehicle age around 6.5 years

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)

QUERY:

```
select claim_flag,Area,avg(VehAge) from Auto_insurance_risk
group by Area
having claim_flag=1;
```

OUTPUT:

	claim_flag	Area	avg(VehAge)
1	1	A	8.06854757255404
2	1	B	7.43954995427981
3	1	C	7.07705336668751
4	1	D	6.93520937227895
5	1	E	6.4445019574679
6	1	F	4.6046563439902

INFERENCE:

Area A contains higher Average of Vehicle Age.

Q8. 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)

QUERY:

```
select claim_flag,exposure_bucket, avg(VehAge) from Auto_insurance_risk
group by claim_flag,exposure_bucket;
```

OUTPUT:

	claim_flag	exposure_bucket	avg(VehAge)
1	0	E1	6.36713799726921
2	0	E2	6.72025297250681
3	0	E3	6.27048520001841
4	0	E4	8.30743135210289
5	1	E1	4.89699570815451
6	1	E2	6.22187448525778
7	1	E3	6.18439842913245
8	1	E4	7.41964171465131

Q 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.

QUERY:

```
ALTER TABLE Auto_insurance_risk
ADD Claim_Ct TEXT;

update Auto_insurance_risk
set Claim_Ct = case
when ClaimNb =1 then  '1 Claim'
when ClaimNb >1 then 'MT 1 Claims'
when ClaimNb =0 then 'No Claim'
end;

select Claim_Ct, avg(bonusmalus) from Auto_insurance_risk
group by Claim_Ct;
```

OUTPUT:

	Claim_Ct	avg(bonusmalus)
1	1 Claim	62.8371558207471
2	MT 1 Claims	67.5531349628055
3	No Claim	59.5850411443071

INFERENCE:

More than 1 claims have more average Bonus Malus involved.

Q. 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.

QUERY:

```
select Claim_Ct, avg(Density) from Auto_insurance_risk
group by Claim_Ct;
```

OUTPUT:

	Claim_Ct	avg(Density)
1	1 Claim	1947.32404127043
2	MT 1 Claims	2297.45483528162
3	No Claim	1783.20605541088

INFERENCE:
those who have more than 1 claim have highest average Density.

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

QUERY:

```
SELECT VehBrand, VehGas, max(maximum) as highest_number_of_claims
from(select VehBrand,VehGas, avg(claimNb)as maximum from Auto_insurance_risk)
group by VehBrand,VehGas;
```

OUTPUT:

	VehBrand	VehGas	highest_number_of_claims
1	B12	Regular	0.0532467666549166

Q. 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)

OUERY:

```
select Region,exposure_bucket,claimNb as maximum_Claim
from Auto_insurance_risk
where claim_flag = 1
order by claimNb DESC
limit 5;
```

OUTPUT:

	Region	exposure_bucket	maximum_Claim
1	R91	E2	16
2	R24	E4	11
3	R91	E1	11
4	R91	E1	11
5	R91	E1	9

Q. 13.1. Are there any cases of illegal driving i.e. underaged folks driving and committing accidents? (1)
QUERY:

```
select DrivAge , ClaimNb from Auto_insurance_risk
where DrivAge <18 and claim_flag =1;
-- no sign of underaged folks driving and commiting accidents
```

INFERENCE :
NO under aged driving commits accident.
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

QUERY:

```
ALTER TABLE Auto_insurance_risk
add DrivAge_Bucket TEXT;

UPDATE Auto_insurance_risk
set DrivAge_Bucket = case
when DrivAge =18 then '1-Beginner'
when DrivAge <=30 then '2-Junior'
when DrivAge <=45 then '3- Middle Age'
when DrivAge <=60 then '4-Mid-Senior'
when DrivAge >60 then '5-SEnior'
end;

SELECT DrivAge_Bucket ,avg(BonusMalus) FROM Auto_insurance_risk
GROUP by DrivAge_Bucket
ORDER by avg(BonusMalus) DESC;
```

OUTPUT:

	DrivAge_Bucket	avg(BonusMalus)
1	1-Beginner	93.0093582887701
2	2-Junior	79.4330688927232
3	3- Middle Age	59.4059998188556
4	4-Mid-Senior	53.9518476577847
5	5-SEnior	52.8022145154416

INFERENCE:

Beginner age have higher Bonus Malus involved in the claims

Conceptual

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

Unique constraints are also referenced by the foreign key of another table. It can be used when someone wants to enforce unique constraints on a column and a group of columns which is not a primary key.

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)

Cross join gives the cross product of two column in the table hence resultant table gives 50 records

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

Inner join	Outer Join
Returns the combined records between more than two tables	Returns the combined record from a specified table even join condition will fail
When attributes have no common values then it will return nothing	Does not depend upon the common values ,if the attribute is blank then it shows null
JOIN and INNER JOIN both clauses work the same	FULL OUTER JOIN and FULL JOIN both clauses work the same

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)

Inner Join generate the output as the common record so the value common in both the table gives the output

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

Where Clause- It is used to extract only those records that fulfill a specified criteria.

Having Clause - It is used to extract only those records that fulfill a specified criteria when condition should be applied on aggregated fields.

Having clause cannot be used without Group by clause where as WHERE Clause does not have any condition.