**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.**

ANS: - Total 5% of the customers have claimed in mentioned exposure period.

SQL- SELECT

sum(case when CLaimNb > 0 then 1 else 0 end) as total\_claimed

,count(IDpol) as total\_customer

,sum(case when CLaimNb > 0 then 1 else 0 end)/count(IDpol) as per\_claimed

FROM auto\_insurance\_risk;

**2. 2.1. Create a new column as 'claim\_flag' in the table 'auto\_insurance\_risk' as integer datatype. (1.5)**

SQL-

ALTER TABLE auto\_insurance\_risk ADD claim\_flag integer;

**2.2. Set the value to 1 when ClaimNb is greater than 0 and set the value to 0 otherwise. (1.5)**

SQL-

UPDATE auto\_insurance\_risk

SET claim\_flag = case when CLaimNb > 0 then 1 else 0 end;

**3.1. What is the average exposure period for those who have claimed? (1)**

**Hint: Use claim\_flag variable to group the data.**

SQL-

select avg(Exposure) FROM Auto\_insurance\_risk WHERE ClaimNb > 0;

**3.2. What do you infer from the result? (1)**

The result states that average exposure for those who have not claimed is 643953 and for those who have claimed is 34060. Therefore, the average exposure is more for those who have not claimed.

**4.1. If we create an exposure bucket where buckets are like below, what is the % of total claims by these buckets? (2)**

SQL:-

SELET

case when exposure <= 0.25 then 'E1' when exposure <= 0.50 then 'E2' when exposure <= 0.75 then 'E3' else 'E4'

end as exposure\_bkt

,sum(ClaimNb) as num\_claims

,count(IDpol) as num\_pol

,sum(ClaimNb)/count(IDpol) as pct\_claim FROM auto\_insurance\_risk

GROUP BY case when exposure <= 0.25 then 'E1' when exposure <= 0.50 then 'E2' when exposure <= 0.75 then 'E3' else 'E4'

end;

**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.**

E1 = 3.2%, E2 = 4.9%, E3 = 6.5%, E4 = 7.1%. As seen in Q3, indeed higher exposure policies have higher claim rate. From the summary, we can see that customer with policies having exposure >0.75 [i.e. E4] has the highest claim rate ~7.1% which is more than double the claim rate of E1 bucket.

**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.

--A. Area - F has the highest number of average claims. It is 6.3% of the total number of policies in Area F.

--SQL- SELECT

Area

,avg(ClaimNb) as avgerage\_claims FROM auto\_insurance\_risk

GROUP BY Area;

--Alternate solution SELECT

Area

,sum(ClaimNb) as tot\_claims

,count(IDpol) as tot\_pol

,sum(ClaimNb)/count(IDpol) as per\_claim

FROM auto\_insurance\_risk

GROUP BY Area;

**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?**

**Note: 2 Marks for SQL and 1 for inference.**

For Area E & F, the exposure bucket E4 has the highest claim rate with 8.6% and 8.8% respectively. Also, Area F has relatively much higher claim rate for E2 bucket as well ~6.5%

SQL. SELECT

Area

,case when exposure <= 0.25 then 'E1' when exposure <= 0.50 then 'E2' when exposure <= 0.75 then 'E3' else 'E4'

end as exposure\_bkt

,sum(ClaimNb) as num\_claims

,count(IDpol) as num\_pol

,sum(ClaimNb)/count(IDpol) as pct\_claim FROM auto\_insurance\_risk

GROUP BY Area

,case when exposure <= 0.25 then 'E1' when exposure <= 0.50 then 'E2' when exposure <= 0.75 then 'E3' else 'E4'

end

ORDER BY sum(ClaimNb)/count(IDpol) DESC

**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)**

SQL:

SELECT avg(VehAge) FROM Auto\_insurance\_risk WHERE ClaimNb > 0 UNION SELECT avg(VehAge) FROM Auto\_insurance\_risk WHERE ClaimNb = 0 ;

The above query resulted in the 6.50252495596007 for those who claimed and 7.07291836516019 for those who didn’t claim i.e., the average vehicle age is lesser for those who claimed.

**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)**

SQL:

SELECT avg(VehAge), Area FROM Auto\_insurance\_risk WHERE ClaimNb > 0 GROUP BY Area;

**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)**

The average vehicle age is often greater for clients with more exposure, including those who claimed and those who didn't. If we look closely, however, we can see that the E1 bucket, which is the bucket with the least risk, has the greatest average difference vehicle age between claimants and non-claimants.

The difference of 1.5 years between the mean VehAge of E1 bucket for claimants and non-claimants is 4.9 against 6.4.

--It implies that consumers with less exposure are more at risk while driving relatively recent autos.

SQL. SELET

case when exposure <= 0.25 then 'E1' when exposure <= 0.50 then 'E2' when exposure <= 0.75 then 'E3' else 'E4'

end as exposure\_bkt

,claim\_flag

,avg(VehAge)

FROM auto\_insurance\_risk

GROUP BY case when exposure <= 0.25 then 'E1' when exposure <= 0.50 then 'E2' when exposure <= 0.75 then 'E3' else 'E4'

end

,claim\_flag;

**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.**

The greatest average bonusmalus—67.6—is for MT 1 Claims. It indicates that people who file claims more frequently generally receive the smallest insurance premium cut.

SQL:-

SELECT

case

when ClaimNb = 0 then 'No Claims' when ClaimNb = 1 then '1 Claim' when ClaimNb > 1 then 'MT 1 Claims' end as Claim\_Ct

,avg(BonusMalus) FROM auto\_insurance\_risk

GROUP BY case

when ClaimNb = 0 then 'No Claims' when ClaimNb = 1 then '1 Claim' when ClaimNb > 1 then 'MT 1 Claims'

end

**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 dat**

**Note: 2.5 Marks for SQL and 1.5 for inference**.

The places where a claim has been made have a substantially greater population density. The population density is much higher within the claim zones when there are many claims.

-- 1 Claim =1947.3

-- MT 1 Claims =2297.5

-- No Claims =1783.2

SQL :- SELECT

case

when ClaimNb = 0 then 'No Claims' when ClaimNb = 1 then '1 Claim' when ClaimNb > 1 then 'MT 1 Claims'

end as Claim\_Ct

,avg(Density)

FROM auto\_insurance\_risk GROUP BY case

when ClaimNb = 0 then 'No Claims' when ClaimNb = 1 then '1 Claim' when ClaimNb > 1 then 'MT 1 Claims'

end

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

VehGas = Regular & VehBrand = B12 has the highest Claim rate @6.4% among all the different Vehical Brands and Gas types.

SQL :- SELECT

VehGas

,VehBrand

,avg(ClaimNb)

FROM auto\_insurance\_risk GROUP BY VehGas

,VehBrand

ORDER BY avg(ClaimNb) DESC Limit 1;

**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)**

Here is a list of the Top 5 CLaim Rate Region & Exposure bucket combination exposure\_bkt Region claim\_cnt policy\_cnt claim\_rate

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| E3 | R42 | 25 |  | 319 |  |  | 0.078 |
| E4 | R82 | 2258 | 29617 |  | 0.076 |  |  |
| E4 | R11 | 1090 | 14383 |  | 0.076 |  |  |
| E4 | R53 | 1592 | 21318 |  | 0.075 |  |  |
| E4 | R25 | 352 |  | 4761 |  | 0.074 |  |

SQL: SELECT

case when exposure <= 0.25 then 'E1' when exposure <= 0.50 then 'E2'

when exposure <= 0.75 then 'E3' else 'E4'

end as exposure\_bkt

,Region

,sum(claim\_flag) as claim\_cnt

,count(IDPol) as policy\_cnt

,sum(claim\_flag)/count(IDPol) as claim\_rate FROM auto\_insurance\_risk

GROUP BY case when exposure <= 0.25 then 'E1' when exposure <= 0.50 then 'E2' when exposure <= 0.75 then 'E3' else 'E4'

end

,Region

ORDER BY sum(claim\_flag)/count(IDPol) DESC Limit 5;

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

No, there have never been any accidents involving drivers who were under the age of 18.

**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**

Since novices have a larger chance of getting into accidents and subsequently filing claims, the younger the person, the higher the BonusMalus.

Therefore the discount given to these customers are much lower than other age groups.

age\_group avg\_bonusmalus 1-Beginner 93.009

2-Junior 79.433

1. Middle Age 59.406
2. Mid-Senior 53.952

5-Senior 52.802

--SQL.

SELECT count(\*) --Count is Zero FROM auto\_insurance\_risk

where DrivAge < 18

SELECT

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' else '5-Senior'

end as age\_group

,avg(BonusMalus) as avg\_bonusmalus FROM auto\_insurance\_risk

GROUP BY 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' else '5-Senior'

end