```
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--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.
--A. 5% of the customers have claimed in the given exposure period.
--SQL-
SELECT
     sum(case when CLaimNb > 0 then 1 else 0 end) as count_claimed
    ,count(IDpol) as count_cust
    ,sum(case when CLaimNb > 0 then 1 else 0 end)/count(IDpol) as pct_claimed
FROM auto_insurance_risk;
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--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)
--SQL-
ALTER TABLE auto_insurance_risk ADD claim_flag integer;

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)
--3.2. What do you infer from the result? (1)
--Hint: Use claim_flag variable to group the data.
--A. 0.64 is the average exposure period of those who have claimed. It is higher
than those who have not claimed(0.52). It suggests that typically customers
--with higher exposure to the insurance have higher propensity to claim.
--SQL-
SELECT
     claim_flag
    ,avg(Exposure) as avg_exposure
FROM auto_insurance_risk
GROUP BY claim_flag;
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--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.
--A. 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
--customers 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.
--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
    ,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;
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
--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 avg_claims
FROM auto_insurance_risk
GROUP BY Area;

--Alternate solution
SELECT
     Area
    ,sum(ClaimNb) as num_claims
      ,count(IDpol) as num_pol
      ,sum(ClaimNb)/count(IDpol) as pct_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.
--A. 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)
--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)
--A.1. Average VehAge for those who claimed is 6.5 years while the same is 7.1
for those who didn't claim. There is visually no difference between the same.
--A.2. When we group the data by Area and filter on claim_flag = 1, we notice
that the average vehicle age for those who claimed is highest ~7.4 in Area A
while the same is least ~4.04 in Area F.
--It essentially means that the accident rate in Area A is much lower than Area
```

```
F. It also indicates that the average age of vehicles in Area A is much higher
than Area F
--SQL.
SELECT
       claim_flag
      ,avg(VehAge)
FROM auto_insurance_risk
GROUP BY claim_flag

SELECT
       Area
      ,avg(VehAge)
FROM auto_insurance_risk
WHERE claim_flag = 1
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)
--A. Typically the average vehicale age is more for the higher exposure
customers both those who claimed and those who didn't.
--However, if we notice carefully, the difference of average vehicle age between
claimers and non-claimers is highest for E1 bucket which is the least exposure
bucket.
--The average VehAge of E1 bucket for claimers is 4.9 while the same for non-
claimers in 6.4 which makes the difference 1.5 years.
--It means relatively newer vehicles are at higher risk for lower exposure
customers.
--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
     ,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.
--A. The average bonusmalus is highest for MT 1 Claims which is 67.6. It means,
typically those who claim more frequently get least discount in insurance
premium.
--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
data?(4)
--Note: 2.5 Marks for SQL and 1.5 for inference.
--A. The population density is much higher for those areas where a claim has
been made. Within the regions of claim, where the claim counts are more than
one, the population density is even higher.
-- 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)
--A. 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)
--A. 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)
--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
--A.13.1. No, there is no such illegal cases of driving where folks with age
less than 18 have committed accident.
--A.13.2. The lower the age, the higher the BonusMalus, which means since they
are beginners, they have higher risk of committing accident and eventually make
claims.
--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
3-Middle Age                 59.406
4-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
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
```