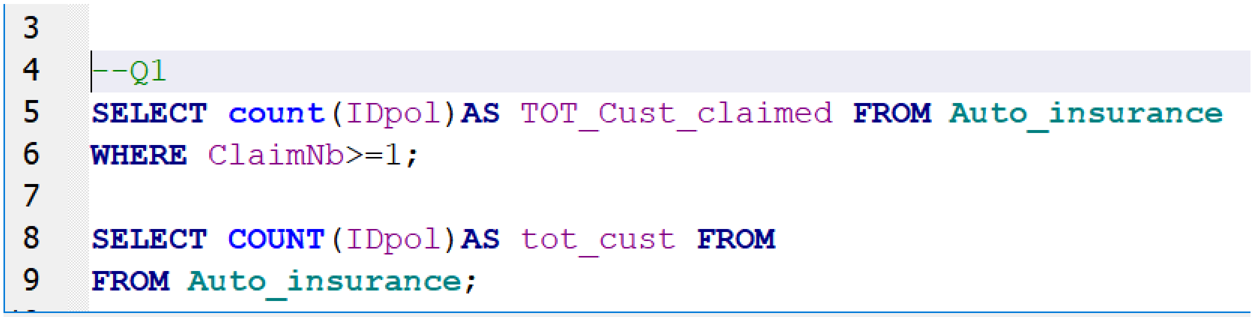
Ashwath J

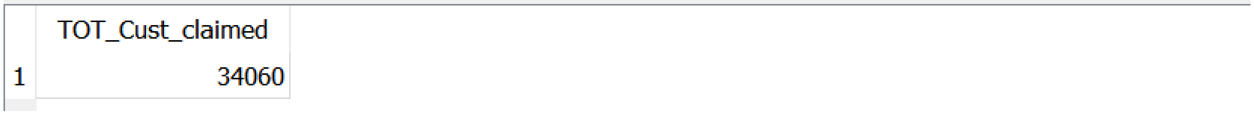
AUTO INSURANCE RISK - BUSINESS REPORT
GRADED PROJECT

Q1)Write a query to calculate what % of the customers have made a claim in the current exposure period[i.e. in the given dataset]?

Customer made claim = 34060

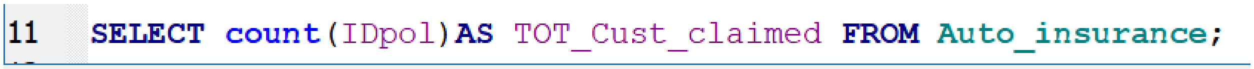


```sql

--Q1
SELECT count(IDpol)AS TOT_Cust_claimed FROM Auto_insurance
WHERE ClaimNb>=1;

SELECT COUNT(IDpol)AS tot_cust FROM
FROM Auto_insurance;
```

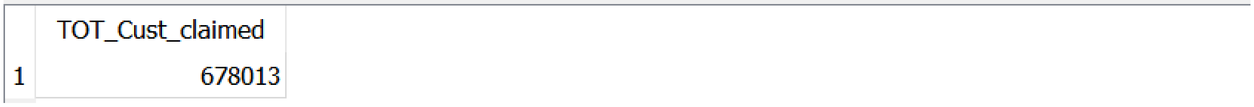


	TOT_Cust_claimed
1	34060

Total Customers = 678013



```sql
SELECT count(IDpol)AS TOT_Cust_claimed FROM Auto_insurance;
```

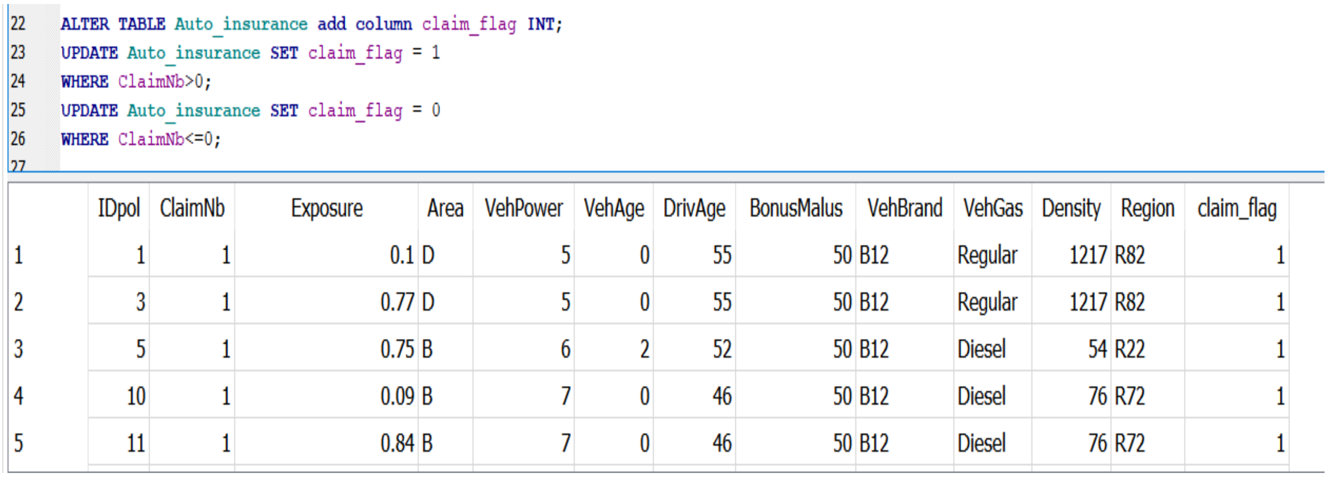


	TOT_Cust_claimed
1	678013

I.e, 5% of customers have claimed.

Ashwath J

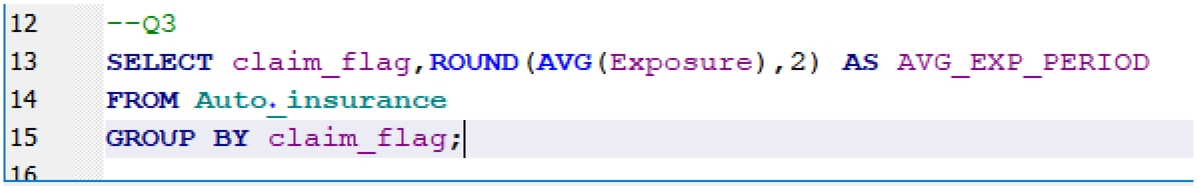
Q2)
2.1. Create a new column as 'claim_flag' in the table 'auto_insurance_risk' as integer datatype.
2.2. Set the value to 1 when ClaimNb is greater than 0 and set the value to 0 otherwise.



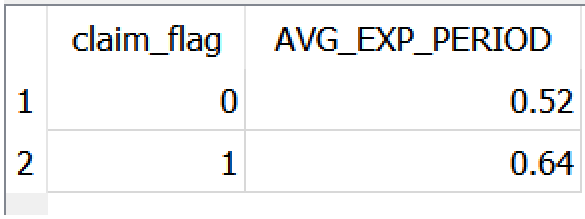
```
ALTER TABLE Auto_insurance add column claim_flag INT;
UPDATE Auto_insurance SET claim_flag = 1
WHERE ClaimNb>0;
UPDATE Auto_insurance SET claim_flag = 0
WHERE ClaimNb<=0;
```

	IDpol	ClaimNb	Exposure	Area	VehPower	VehAge	DrivAge	BonusMalus	VehBrand	VehGas	Density	Region	claim_flag
1	1	1	0.1	D	5	0	55	50	B12	Regular	1217	R82	1
2	3	1	0.77	D	5	0	55	50	B12	Regular	1217	R82	1
3	5	1	0.75	B	6	2	52	50	B12	Diesel	54	R22	1
4	10	1	0.09	B	7	0	46	50	B12	Diesel	76	R72	1
5	11	1	0.84	B	7	0	46	50	B12	Diesel	76	R72	1

Q3)
3.1. What is the average exposure period for those who have claimed?
3.2. What do you infer from the result? Hint: Use claim_flag variable to group the data.



```
--Q3
SELECT claim_flag,ROUND(AVG(Exposure),2) AS AVG_EXP_PERIOD
FROM Auto.insurance
GROUP BY claim_flag;
```

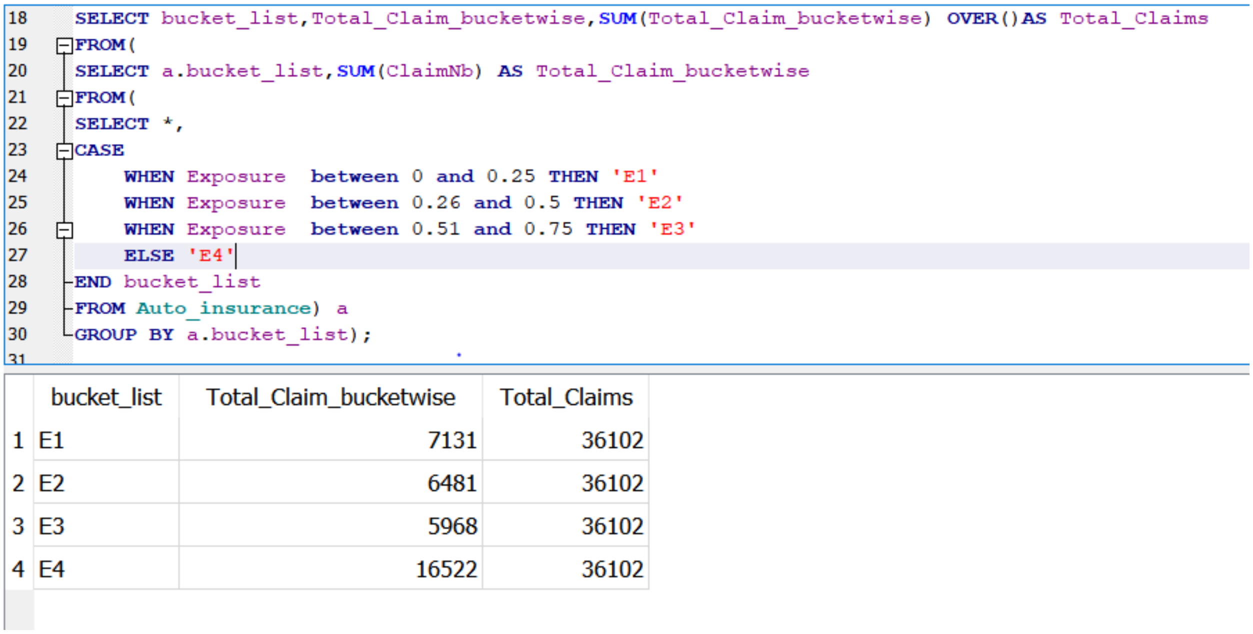


	claim_flag	AVG_EXP_PERIOD
1	0	0.52
2	1	0.64

Inference: Avg exposure period is high in case of customers who have claimed.

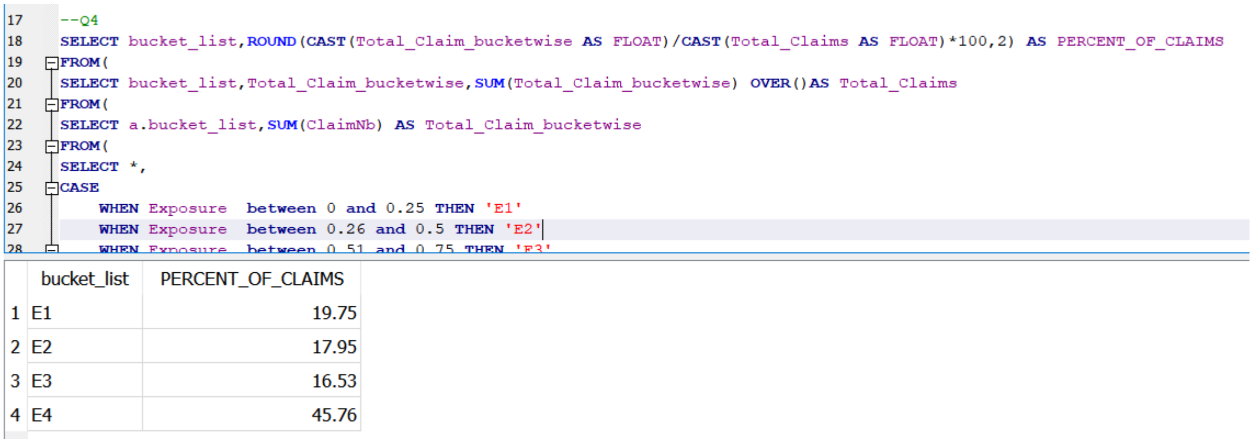
Ashwath J

Q4)
4.1. If we create an exposure bucket where buckets are like below, what is the % of total claims by these buckets?
4.2. What do you infer from the summary?
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 the ClaimNb field to get the total claim count.



```
SELECT bucket_list,Total_Claim_bucketwise,SUM(Total_Claim_bucketwise) OVER()AS Total_Claims
FROM(
SELECT a.bucket_list,SUM(ClaimNb) AS Total_Claim_bucketwise
FROM(
SELECT *,
CASE
    WHEN Exposure  between 0 and 0.25 THEN 'E1'
    WHEN Exposure  between 0.26 and 0.5 THEN 'E2'
    WHEN Exposure  between 0.51 and 0.75 THEN 'E3'
    ELSE 'E4'
END bucket_list
FROM Auto_insurance) a
GROUP BY a.bucket_list);
```

	bucket_list	Total_Claim_bucketwise	Total_Claims
1	E1	7131	36102
2	E2	6481	36102
3	E3	5968	36102
4	E4	16522	36102

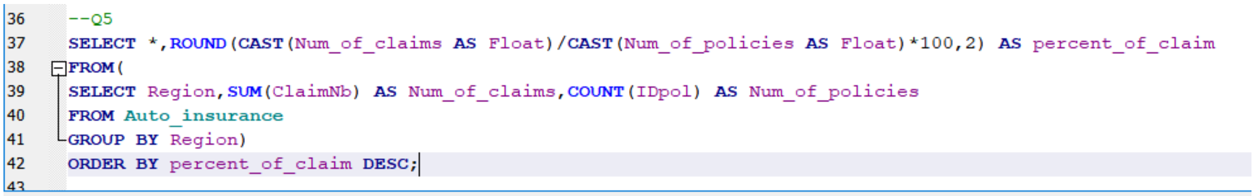


```
--Q4
SELECT bucket_list,ROUND(CAST(Total_Claim_bucketwise AS FLOAT)/CAST(Total_Claims AS FLOAT)*100,2) AS PERCENT_OF_CLAIMS
FROM(
SELECT bucket_list,Total_Claim_bucketwise,SUM(Total_Claim_bucketwise) OVER()AS Total_Claims
FROM(
SELECT a.bucket_list,SUM(ClaimNb) AS Total_Claim_bucketwise
FROM(
SELECT *,
CASE
    WHEN Exposure  between 0 and 0.25 THEN 'E1'
    WHEN Exposure  between 0.26 and 0.5 THEN 'E2'
    WHEN Exposure  between 0.51 and 0.75 THEN 'E3'
```

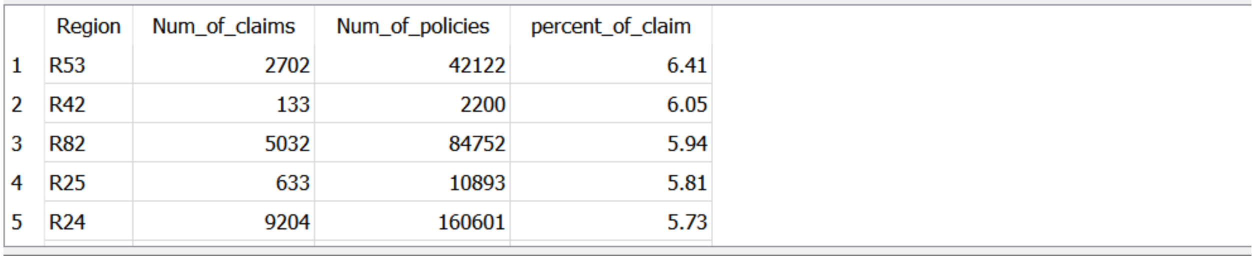
	bucket_list	PERCENT_OF_CLAIMS
1	E1	19.75
2	E2	17.95
3	E3	16.53
4	E4	45.76

- **Highest number of claims is done in exposure of more than 0.75**

Q5)Which area has the highest number of average claims? Show the data in percentage w.r.t. the number of policies in the corresponding Area.
 Hint: Use the ClaimNb field for this question.

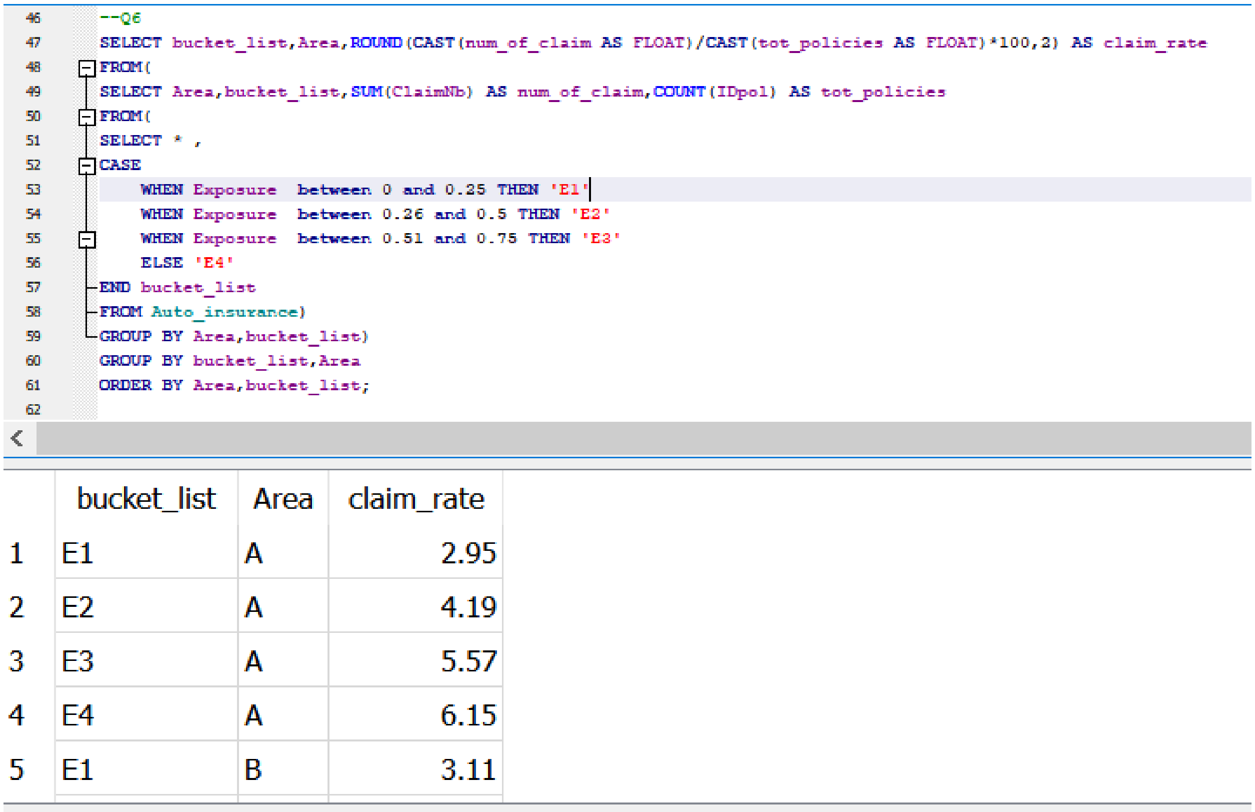


```
--Q5
SELECT *,ROUND(CAST(Num_of_claims AS Float)/CAST(Num_of_policies AS Float)*100,2) AS percent_of_claim
FROM(
SELECT Region,SUM(ClaimNb) AS Num_of_claims,COUNT(IDpol) AS Num_of_policies
FROM Auto_insurance
GROUP BY Region)
ORDER BY percent_of_claim DESC;
```



	Region	Num_of_claims	Num_of_policies	percent_of_claim
1	R53	2702	42122	6.41
2	R42	133	2200	6.05
3	R82	5032	84752	5.94
4	R25	633	10893	5.81
5	R24	9204	160601	5.73

Q6)
If we use these exposure buckets 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?



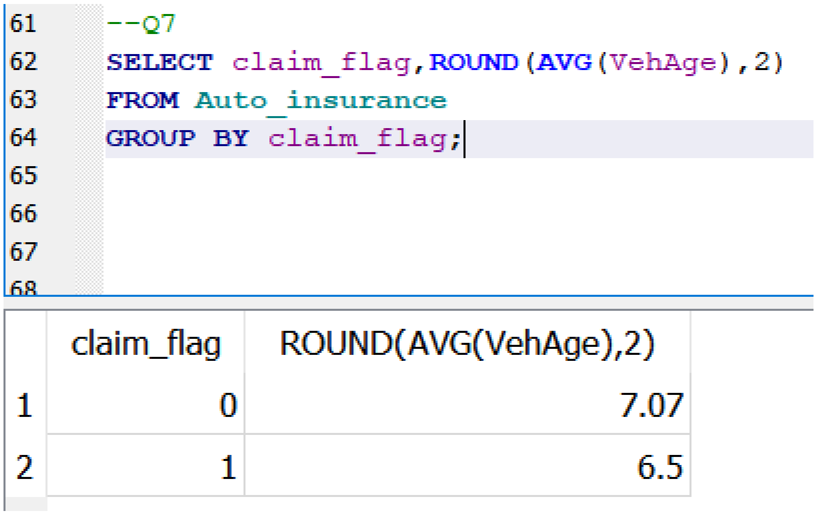
```
--Q6
SELECT bucket_list,Area,ROUND(CAST(num_of_claim AS FLOAT)/CAST(tot_policies AS FLOAT)*100,2) AS claim_rate
FROM(
SELECT Area,bucket_list,SUM(ClaimNb) AS num_of_claim,COUNT(IDpol) AS tot_policies
FROM(
SELECT * ,
CASE
    WHEN Exposure  between 0 and 0.25 THEN 'E1'
    WHEN Exposure  between 0.26 and 0.5 THEN 'E2'
    WHEN Exposure  between 0.51 and 0.75 THEN 'E3'
    ELSE 'E4'
END bucket_list
FROM Auto_insurance)
GROUP BY Area,bucket_list)
GROUP BY bucket_list,Area
ORDER BY Area,bucket_list;
```

	bucket_list	Area	claim_rate
1	E1	A	2.95
2	E2	A	4.19
3	E3	A	5.57
4	E4	A	6.15
5	E1	B	3.11

INFERENCE- For each area the claim_rate is increasing with the increase in the emposure bucket level.

Q7)

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)

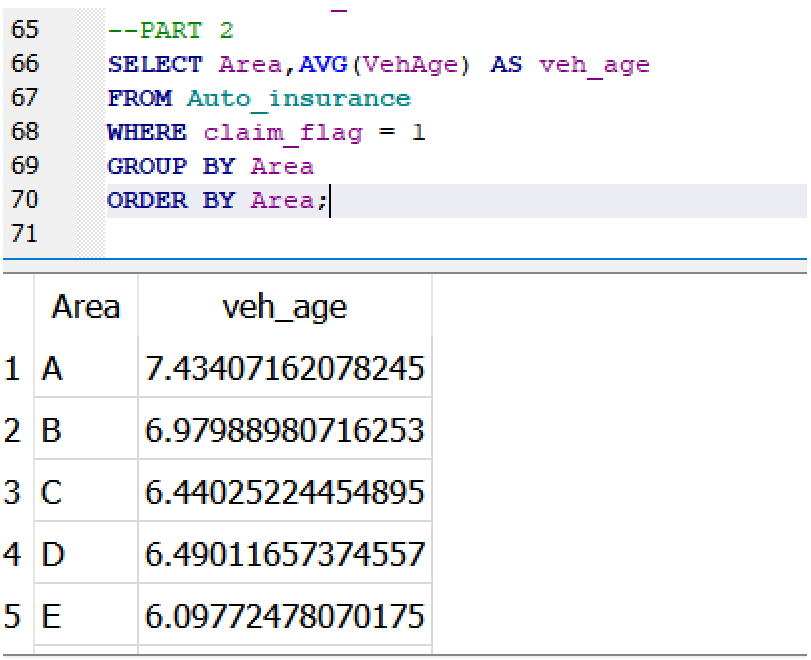


```
--Q7
SELECT claim_flag,ROUND(AVG(VehAge),2)
FROM Auto_insurance
GROUP BY claim_flag;
```

	claim_flag	ROUND(AVG(VehAge),2)
1	0	7.07
2	1	6.5

INFERENCE- The vehicle age of people who claim their insurance is less when compared with the vehicle age of people who don't claim their insurance,this can be taken in this way also people care a lot about newly bought bikes.

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)

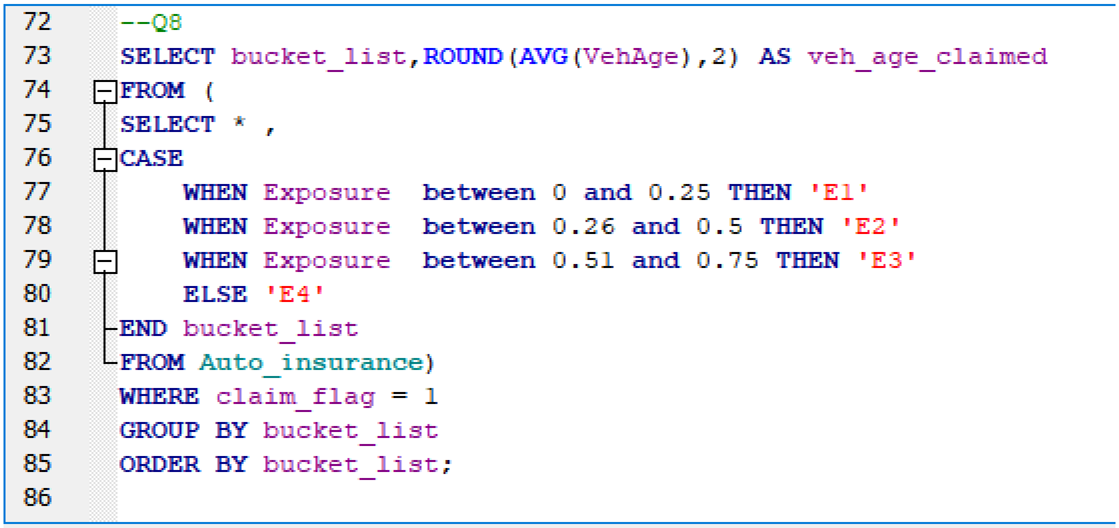


```
--PART 2
SELECT Area,AVG(VehAge) AS veh_age
FROM Auto_insurance
WHERE claim_flag = 1
GROUP BY Area
ORDER BY Area;
```

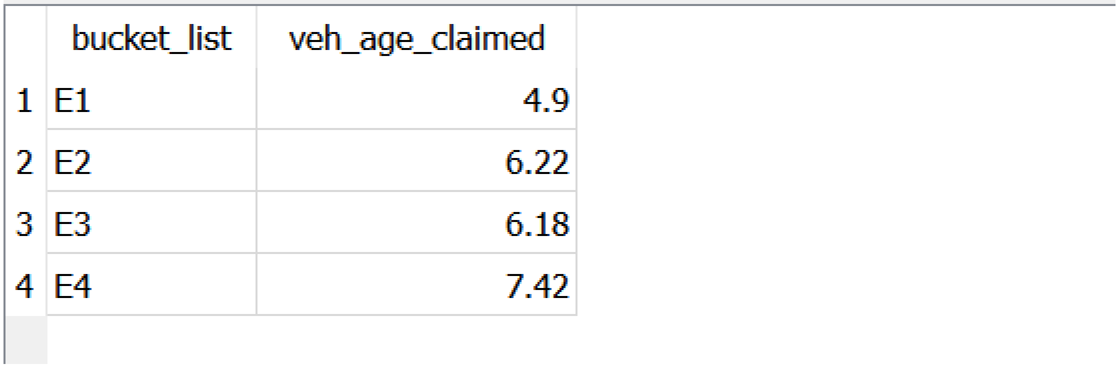
	Area	veh_age
1	A	7.43407162078245
2	B	6.97988980716253
3	C	6.44025224454895
4	D	6.49011657374557
5	E	6.09772478070175

INFERENCE- The vehicle age is continuously decreasing.

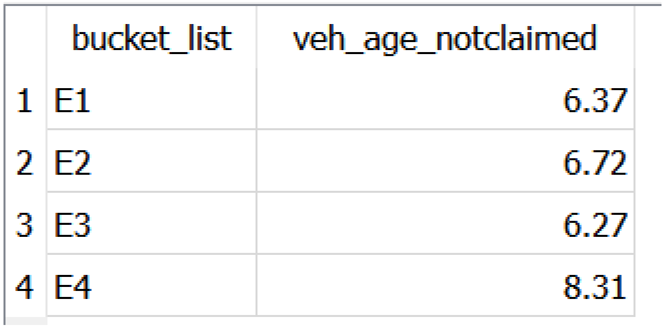
Q8). If we calculate the average vehicle age by exposure bucket(as mentioned above), we see an interesting trend between those who claimed vs those who didn't. What is that?



```
--Q8
SELECT bucket_list,ROUND(AVG(VehAge),2) AS veh_age_claimed
FROM (
SELECT * ,
CASE
    WHEN Exposure  between 0 and 0.25 THEN 'E1'
    WHEN Exposure  between 0.26 and 0.5 THEN 'E2'
    WHEN Exposure  between 0.51 and 0.75 THEN 'E3'
    ELSE 'E4'
END bucket_list
FROM Auto_insurance)
WHERE claim_flag = 1
GROUP BY bucket_list
ORDER BY bucket_list;
```



	bucket_list	veh_age_claimed
1	E1	4.9
2	E2	6.22
3	E3	6.18
4	E4	7.42

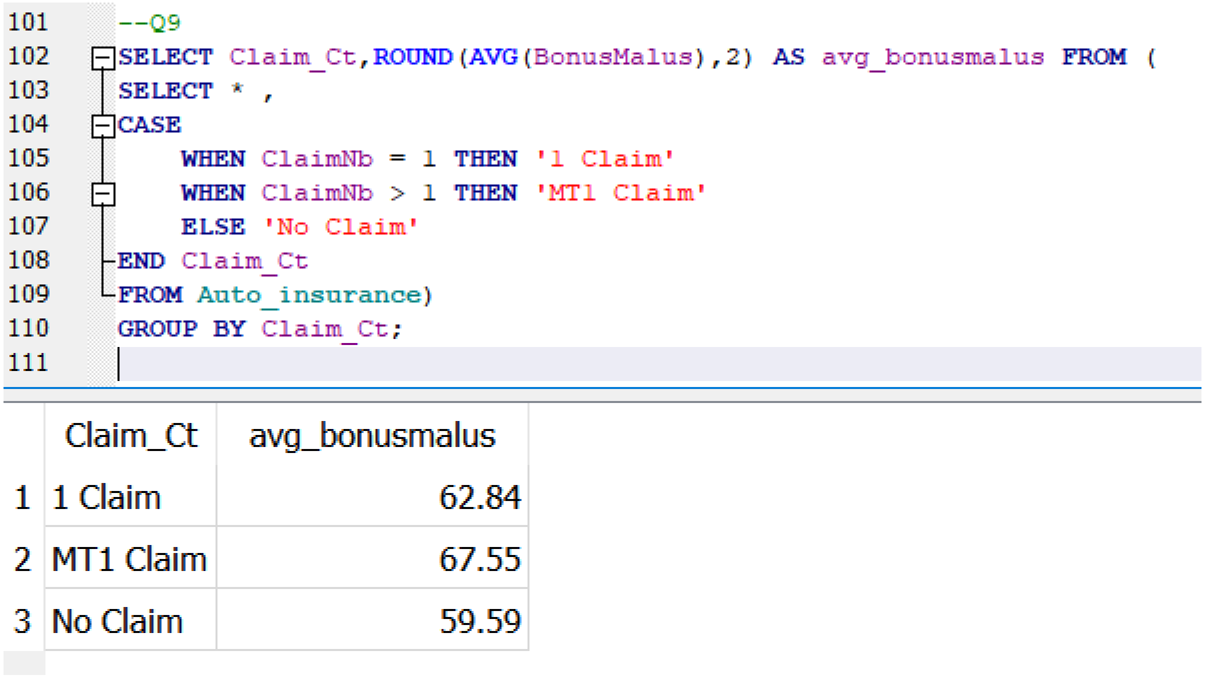


	bucket_list	veh_age_notclaimed
1	E1	6.37
2	E2	6.72
3	E3	6.27
4	E4	8.31

INFERENCE- There is no much increase in the veh_age between claimed and not claimed except in E1 bucket.

Q9)

9.1. Create a Claim_Ct flag on the ClaimNb field as below, and take average of the BonusMalus by Claim_Ct. (2)



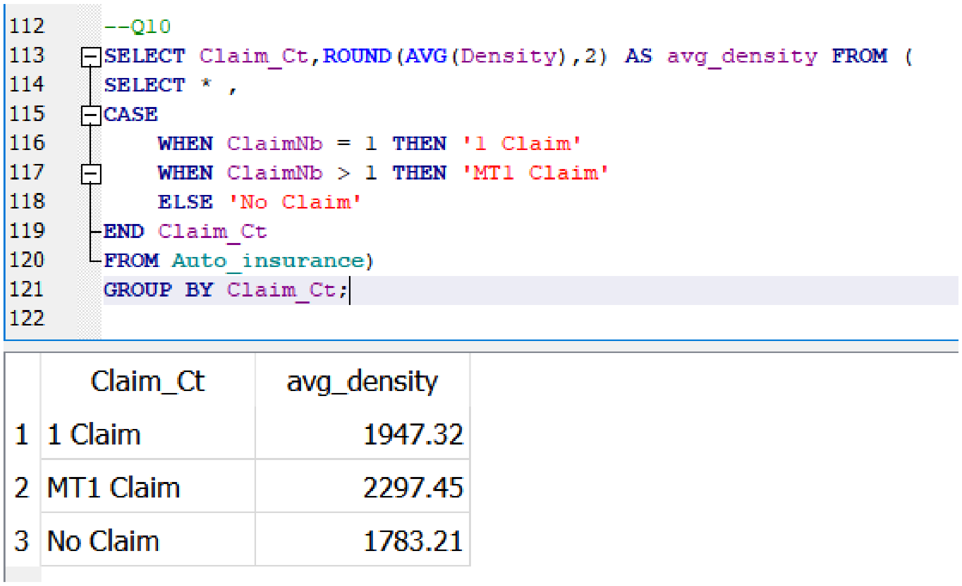
```
--Q9
SELECT Claim_Ct,ROUND(AVG(BonusMalus),2) AS avg_bonusmalus FROM (
SELECT * ,
CASE
    WHEN ClaimNb = 1 THEN '1 Claim'
    WHEN ClaimNb > 1 THEN 'MT1 Claim'
    ELSE 'No Claim'
END Claim_Ct
FROM Auto_insurance)
GROUP BY Claim_Ct;
```

	Claim_Ct	avg_bonusmalus
1	1 Claim	62.84
2	MT1 Claim	67.55
3	No Claim	59.59

9.2. What is the inference from the summary? (1)

INFERENCE - The average fine is more for people who have claimed more than once.

Q10) 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.

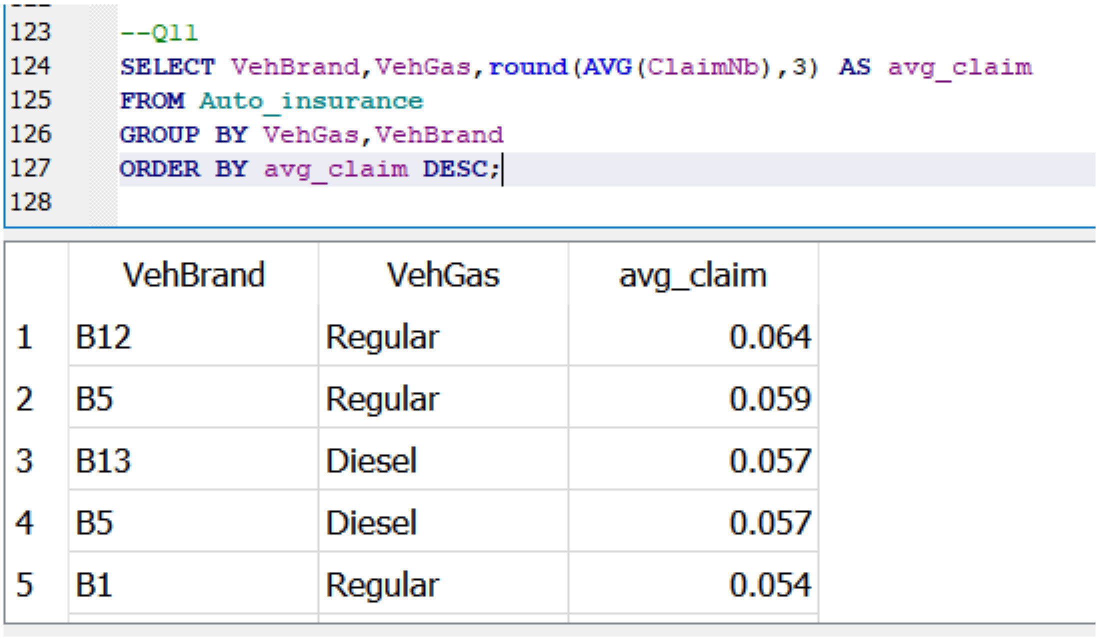


```
--Q10
SELECT Claim_Ct,ROUND(AVG(Density),2) AS avg_density FROM (
SELECT * ,
CASE
    WHEN ClaimNb = 1 THEN '1 Claim'
    WHEN ClaimNb > 1 THEN 'MT1 Claim'
    ELSE 'No Claim'
END Claim_Ct
FROM Auto_insurance)
GROUP BY Claim_Ct;
```

	Claim_Ct	avg_density
1	1 Claim	1947.32
2	MT1 Claim	2297.45
3	No Claim	1783.21

INFERENCE - More than one claim are mostly done in high dense cities/areas.A simple analogy is people in metro cities claim more than once than normal tier 2 and tier 3 places.

Q11) Which Vehicle Brand & Vehicle Gas combination have the highest number of Average Claims (use ClaimNb field for aggregation)? (2)

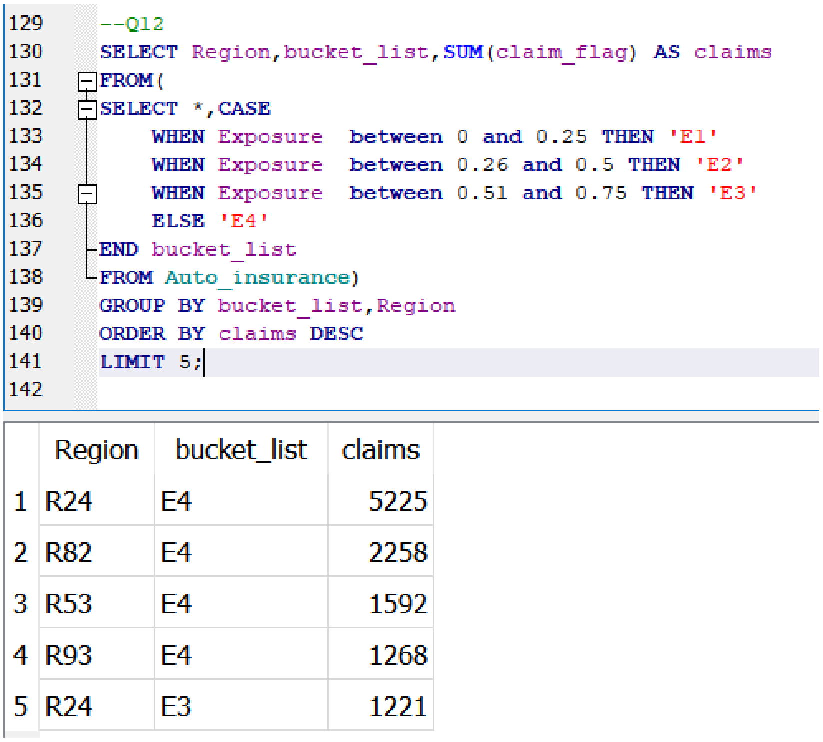


```
--Q11
SELECT VehBrand,VehGas,round(AVG(ClaimNb),3) AS avg_claim
FROM Auto_insurance
GROUP BY VehGas,VehBrand
ORDER BY avg_claim DESC;
```

	VehBrand	VehGas	avg_claim
1	B12	Regular	0.064
2	B5	Regular	0.059
3	B13	Diesel	0.057
4	B5	Diesel	0.057
5	B1	Regular	0.054

The B12 Regular model has the highest average claim among the other models.

Q12) 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)



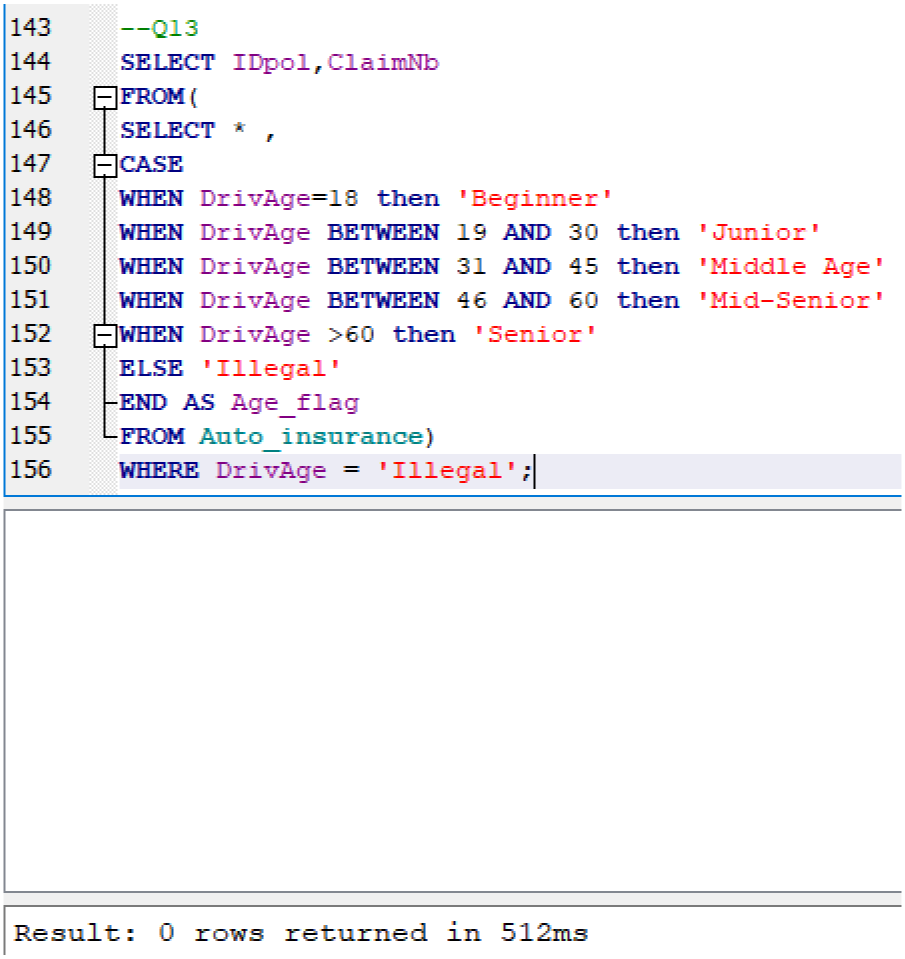
```sql
--Q12
SELECT Region,bucket_list,SUM(claim_flag) AS claims
FROM(
SELECT *,CASE
    WHEN Exposure  between 0 and 0.25 THEN 'E1'
    WHEN Exposure  between 0.26 and 0.5 THEN 'E2'
    WHEN Exposure  between 0.51 and 0.75 THEN 'E3'
    ELSE 'E4'
END bucket_list
FROM Auto_insurance)
GROUP BY bucket_list,Region
ORDER BY claims DESC
LIMIT 5;
```

	Region	bucket_list	claims
1	R24	E4	5225
2	R82	E4	2258
3	R53	E4	1592
4	R93	E4	1268
5	R24	E3	1221

Q13)

Ashwath J

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

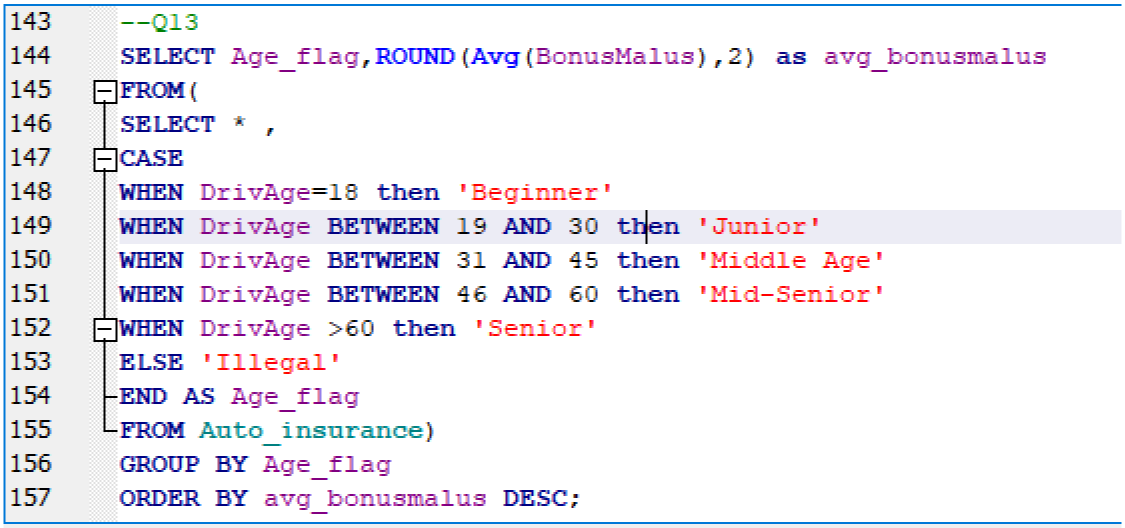


```
--Q13
SELECT IDpol,ClaimNb
FROM(
SELECT * ,
CASE
WHEN DrivAge=18 then 'Beginner'
WHEN DrivAge BETWEEN 19 AND 30 then 'Junior'
WHEN DrivAge BETWEEN 31 AND 45 then 'Middle Age'
WHEN DrivAge BETWEEN 46 AND 60 then 'Mid-Senior'
WHEN DrivAge >60 then 'Senior'
ELSE 'Illegal'
END AS Age_flag
FROM Auto_insurance)
WHERE DrivAge = 'Illegal';
```

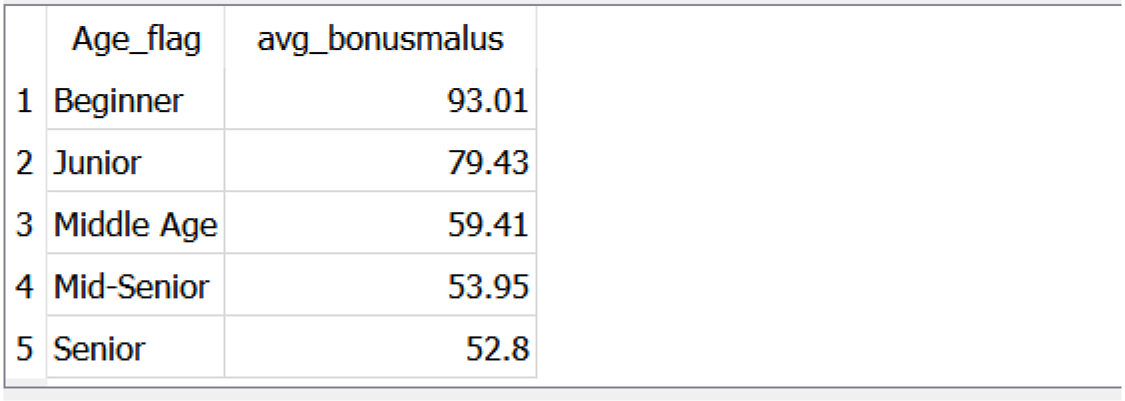
```
Result: 0 rows returned in 512ms
```

INFERENCE - No accident cases are committed by driver less than 18 years

13.2. Create a bucket on DrivAge and then take the 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.



```
--Q13
SELECT Age_flag,ROUND(Avg(BonusMalus),2) as avg_bonusmalus
FROM(
SELECT * ,
CASE
WHEN DrivAge=18 then 'Beginner'
WHEN DrivAge BETWEEN 19 AND 30 then 'Junior'
WHEN DrivAge BETWEEN 31 AND 45 then 'Middle Age'
WHEN DrivAge BETWEEN 46 AND 60 then 'Mid-Senior'
WHEN DrivAge >60 then 'Senior'
ELSE 'Illegal'
END AS Age_flag
FROM Auto_insurance)
GROUP BY Age_flag
ORDER BY avg_bonusmalus DESC;
```



	Age_flag	avg_bonusmalus
1	Beginner	93.01
2	Junior	79.43
3	Middle Age	59.41
4	Mid-Senior	53.95
5	Senior	52.8

INFERENCE - The inexperienced drivers i.e, drivers starting their driving career are the people paying most of the fines.

CONCEPTUAL QUESTIONS:

Q14. Mention one major difference between unique constraint and primary key? (2)
(A)Unique constraints can have null values also included but the primary key should be unique and not null.

Q15. 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)
(A)Cross join is a cartesian product of the number of rows in each table,So therefore the result set will be having 5*10 i.e,50 rows.

Q16. What is the difference between inner join and left outer join? (2)
(A)Inner join does not return null values but left outer join returns null values if there is no exact match on the right table.

Q17. 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 the record count of the output? (2)
(A)The output table will have 4 sets which are in common with both the tables.

Q18. What is the difference between WHERE clause and HAVING clause? (2)
(A) WHERE clause is used to filter the table values before GROUP BY but Having clause
Is used to filter the output after the GROUP BY clause.

THE END

Submitted by,
Ashwath J