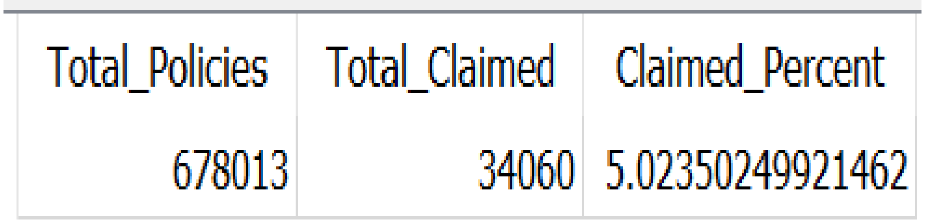
SQL GRADED PROJECT

Name: Panagam Mohitha

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]?**

```
SELECT
        COUNT(IDpol) AS Total_Policies,
        SUM(CASE WHEN ClaimNb >= 1 THEN 1 ELSE 0 END) AS Total_Claimed,
        SUM(CASE WHEN ClaimNb >= 1 THEN 1 ELSE 0 END)/6780.13  AS Claimed_Percent
FROM auto_insurance_risk;
```

Result:



Total_Policies	Total_Claimed	Claimed_Percent
678013	34060	5.02350249921462

Inference:

5.02% of the customers have made a claim in the current exposure period.

2. **2.1. Create a new column as 'claim_flag' in the table 'auto_insurance_risk' as integer datatype.**

```
ALTER TABLE Auto_insurance_risk ADD COLUMN claim_flag int;
```

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

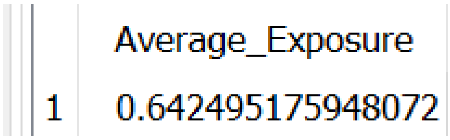
```
UPDATE Auto_insurance_risk
```

```
SET claim_flag = CASE WHEN ClaimNb > 0 THEN 1 ELSE 0 END;
```

3. 3.1. What is the average exposure period for those who have claimed?

```
SELECT

AVG(Exposure) AS Average_Exposure

FROM auto_insurance_risk

WHERE claim_flag IN (1);
```

Result:



	Average_Exposure
1	0.642495175948072

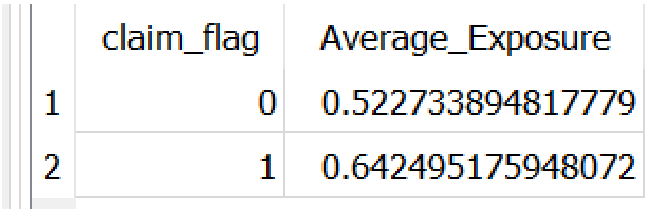
Inference:

0.643 is the average exposure period for those who have claimed.

3.2. What do you infer from the result?

```
SELECT

        claim_flag, AVG(Exposure) AS Average_Exposure

        FROM auto_insurance_risk

        GROUP BY claim_flag ;
```

Result:



	claim_flag	Average_Exposure
1	0	0.522733894817779
2	1	0.642495175948072

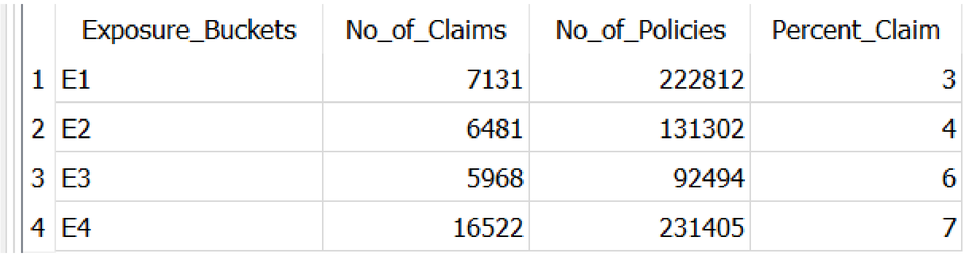
Inference:

0.643 is the average exposure period of those who have claimed. It is higher than those who have not claimed (0.523). Thus those with higher average Exposure tend to claim much more often as compared to the rest.

4. 4.1. If we create an exposure bucket where buckets are like below, what is the % of total claims by these buckets?

```
SELECT
        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"
                ELSE "E4"
                END AS Exposure_Buckets,
        SUM(ClaimNb) AS No_of_Claims,
        COUNT(IDpol) AS No_of_Policies,
        SUM(ClaimNb)*100/COUNT(IDpol) AS Percent_Claim
FROM auto_insurance_risk
        GROUP BY 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"
                ELSE "E4"
                END;
```

Result:



	Exposure_Buckets	No_of_Claims	No_of_Policies	Percent_Claim
1	E1	7131	222812	3
2	E2	6481	131302	4
3	E3	5968	92494	6
4	E4	16522	231405	7

4.2. What do you infer from the summary?

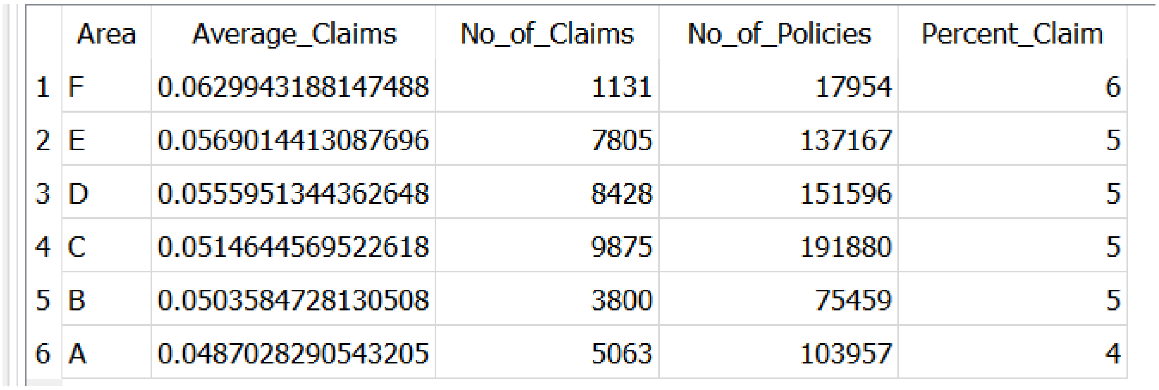
Inference:

E1 = 3%, E2 = 4%, E3 = 6%, E4 = 7%. As seen in previous question, 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% 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.**

```
SELECT

        Area,

        AVG(ClaimNb) as Average_Claims,

        SUM(ClaimNb) as No_of_Claims,

        COUNT(IDpol) as No_of_Policies,

        SUM(ClaimNb)*100/COUNT(IDpol) AS Percent_Claim

FROM auto_insurance_risk

        GROUP BY Area

        ORDER BY AVG(ClaimNb) DESC;
```

Result:



	Area	Average_Claims	No_of_Claims	No_of_Policies	Percent_Claim
1	F	0.0629943188147488	1131	17954	6
2	E	0.0569014413087696	7805	137167	5
3	D	0.0555951344362648	8428	151596	5
4	C	0.0514644569522618	9875	191880	5
5	B	0.0503584728130508	3800	75459	5
6	A	0.0487028290543205	5063	103957	4

Inference:

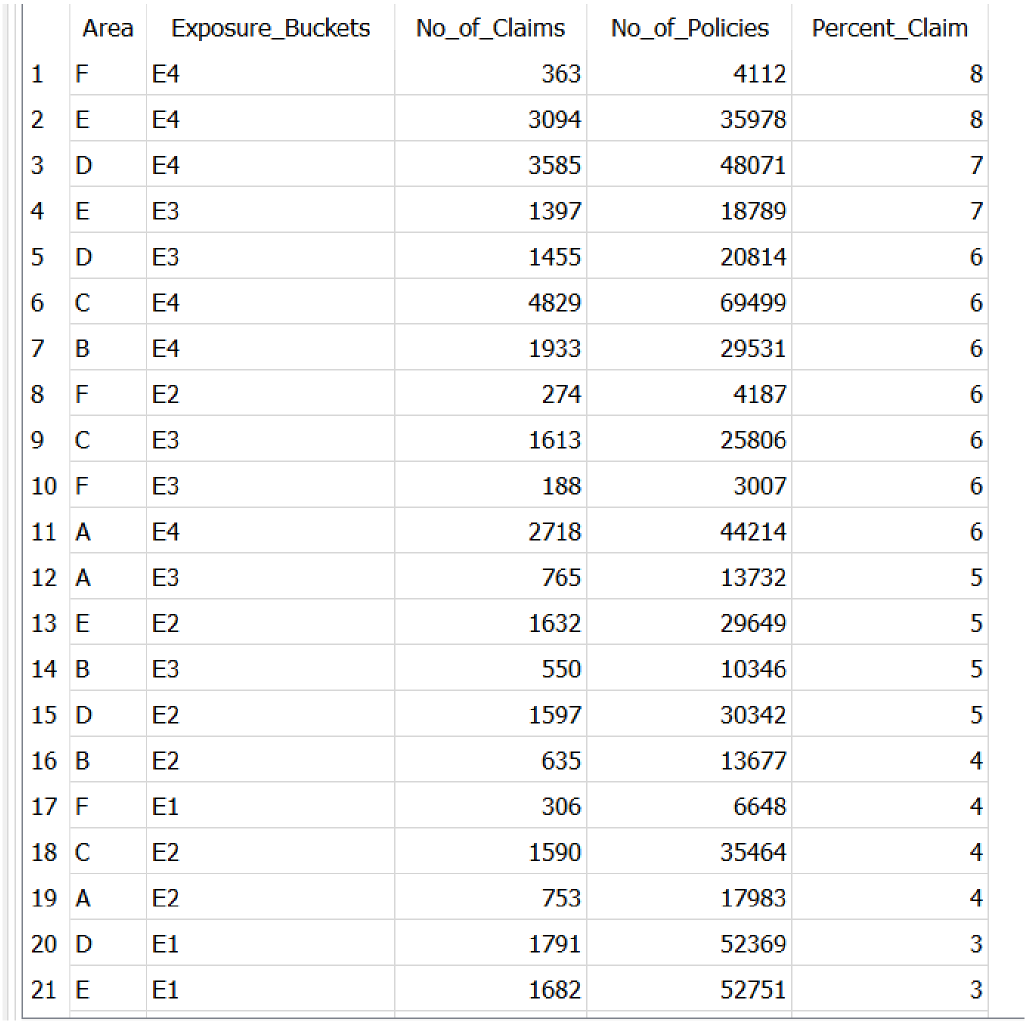
‘F’ area has the highest number of average claims (0.063) with 6% claim rate in that area.

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?**

```
SELECT
    Area,
    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"
        ELSE "E4"
        END AS Exposure_Buckets,
    SUM(ClaimNb) AS No_of_Claims,
    COUNT(IDpol) AS No_of_Policies,
    SUM(ClaimNb)*100/COUNT(IDpol) AS Percent_Claim
FROM auto_insurance_risk
    GROUP BY Area,
    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"
```

```
                    ELSE "E4"

                    END

ORDER BY SUM(ClaimNb)*100/COUNT(IDpol) DESC;
```

Result:



	Area	Exposure_Buckets	No_of_Claims	No_of_Policies	Percent_Claim
1	F	E4	363	4112	8
2	E	E4	3094	35978	8
3	D	E4	3585	48071	7
4	E	E3	1397	18789	7
5	D	E3	1455	20814	6
6	C	E4	4829	69499	6
7	B	E4	1933	29531	6
8	F	E2	274	4187	6
9	C	E3	1613	25806	6
10	F	E3	188	3007	6
11	A	E4	2718	44214	6
12	A	E3	765	13732	5
13	E	E2	1632	29649	5
14	B	E3	550	10346	5
15	D	E2	1597	30342	5
16	B	E2	635	13677	4
17	F	E1	306	6648	4
18	C	E2	1590	35464	4
19	A	E2	753	17983	4
20	D	E1	1791	52369	3
21	E	E1	1682	52751	3

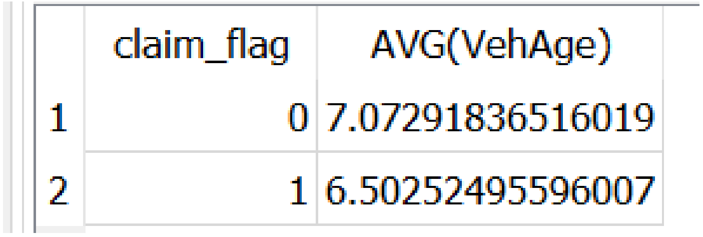
Inference:

For Area E & F, the exposure bucket E4 has the highest claim rate with 8% and 8% respectively. Also, Area F has relatively much higher claim rate for E4 bucket and Area E has higher claim rate in E3 bucket. Area C and Area A plays major role in Exposure buckets E2 and E1 respectively.

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?**

```
SELECT
        claim_flag,
        AVG(VehAge)
FROM auto_insurance_risk
GROUP BY claim_flag;
```

Result:



	claim_flag	AVG(VehAge)
1	0	7.07291836516019
2	1	6.50252495596007

Inference:

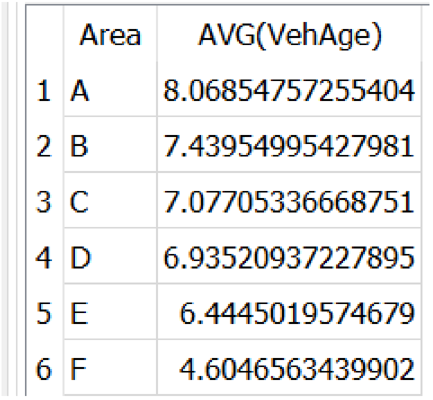
Average VehAge for those who claimed is 6.5 years while the same is 7.07 years for those who didn't claim. Those who did not claim have higher vehicle age as compared to those who claimed. Moreover, there is visually no difference between the same.

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?

```
SELECT
        Area,
        AVG(VehAge)
FROM auto_insurance_risk
```

```
GROUP BY Area

HAVING claim_flag = 1;
```

Result:



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

Inference:

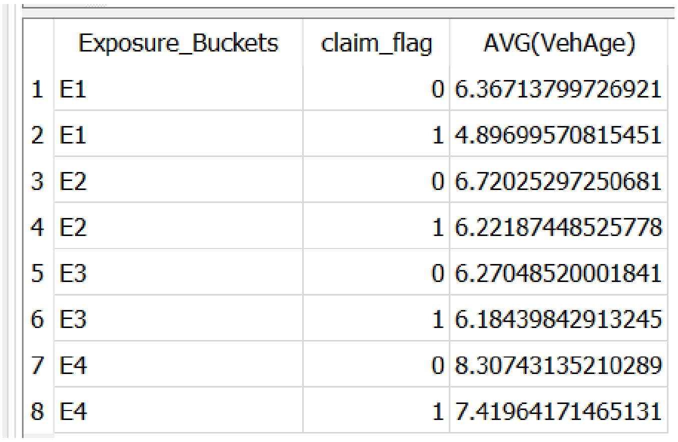
When we group the data by Area and filter on claim_flag = 1, we notice that the average vehicle age for those who claimed is highest ~8.07 in Area A while the same is least ~4.60 in Area F. It essentially means that the accident rate in Area A is much lower than Area F. It also indicates that the average age of vehicles in Area A is much higher than Area F.

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?**

```
SELECT

        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"

                ELSE "E4"
```

```
                    END AS Exposure_Buckets,
                    claim_flag,
                    AVG(VehAge)
FROM auto_insurance_risk
          GROUP BY 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"
                    ELSE "E4"
                    END,
                    claim_flag;
```

Result:



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

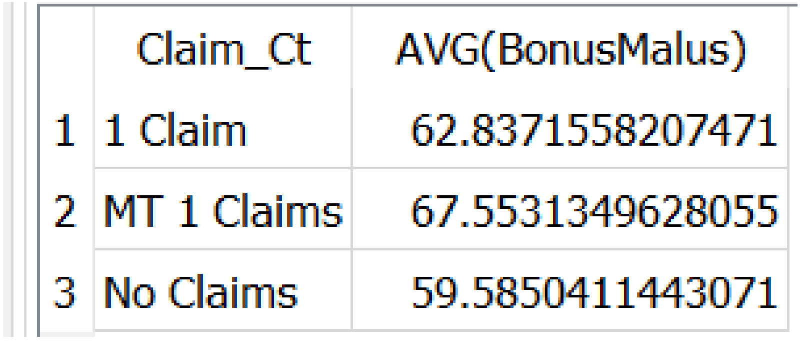
Inference:

Typically the average vehicle age is more for the higher exposure customers both those who claimed and those who didn't. However, 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 4.89 while the same for non claimers in 6.36 which makes the difference 1.47 years. It means relatively newer vehicles are at higher risk for lower exposure customers.

9. **9.1. Create a Claim_Ct flag on the ClaimNb field as below, and take average of the BonusMalus by Claim_Ct.**

```
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;
```

Result:



	Claim_Ct	AVG(BonusMalus)
1	1 Claim	62.8371558207471
2	MT 1 Claims	67.5531349628055
3	No Claims	59.5850411443071

9.2. What is the inference from the summary?

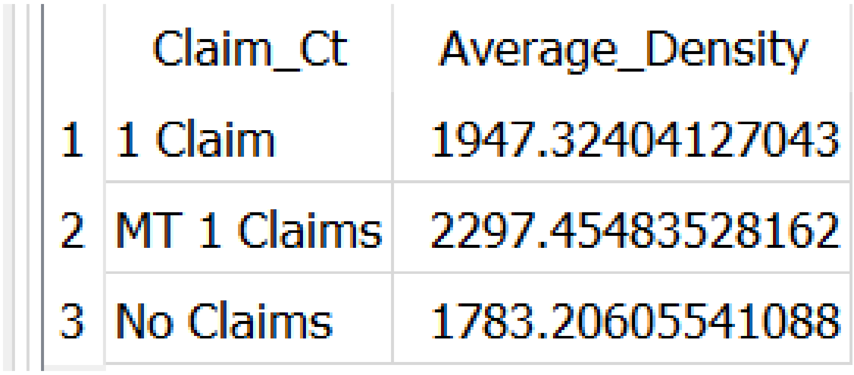
Inference:

We can see that the average BonuMalus is almost same for categories being a bit inclined towards those who have already claimed more than once. 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.

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?

```
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) AS Average_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;
```

Result:



	Claim_Ct	Average_Density
1	1 Claim	1947.32404127043
2	MT 1 Claims	2297.45483528162
3	No Claims	1783.20605541088

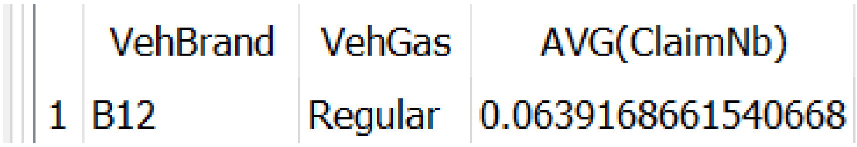
Inference:

Average Density is higher for those with more than one claims. It increases with the claims, thus being more for those who have claimed. 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.

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

```
SELECT
        VehBrand,
        VehGas,
        AVG(ClaimNb)
FROM auto_insurance_risk
GROUP BY VehBrand, VehGas
ORDER BY AVG(ClaimNb) DESC
Limit 1;
```

Result:



	VehBrand	VehGas	AVG(ClaimNb)
1	B12	Regular	0.0639168661540668

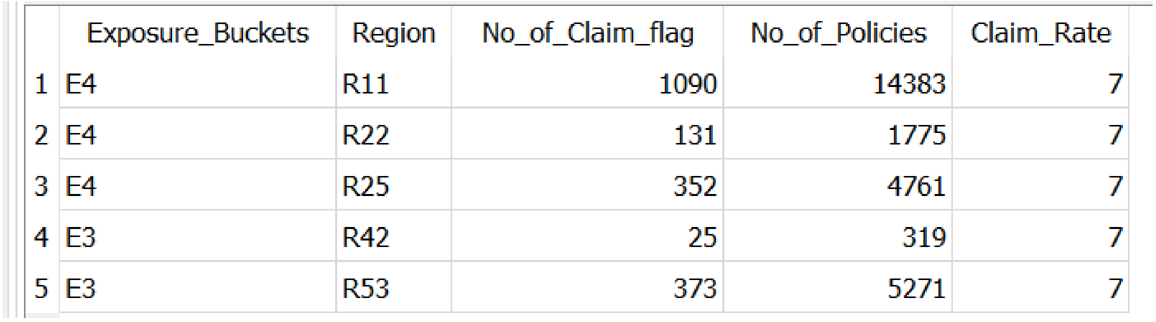
Inference:

Vehicle Brand B12 which is a Regular Vehicle Gas has the highest average claims among all the different Vehicle Brands and Gas types.

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.**

```
SELECT
        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"
                ELSE "E4"
                END AS Exposure_Buckets,
        Region,
        SUM(claim_flag) AS No_of_Claim_flag,
        COUNT(IDpol) AS No_of_Policies,
        SUM(claim_flag)*100/COUNT(IDpol) AS Claim_Rate
FROM auto_insurance_risk
        GROUP BY Region,
        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"

                ELSE "E4"

                END

ORDER BY SUM(claim_flag)*100/COUNT(IDpol) DESC

Limit 5;
```

Result:



	Exposure_Buckets	Region	No_of_Claim_flag	No_of_Policies	Claim_Rate
1	E4	R11	1090	14383	7
2	E4	R22	131	1775	7
3	E4	R25	352	4761	7
4	E3	R42	25	319	7
5	E3	R53	373	5271	7

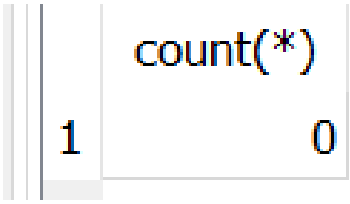
Inference:

The Top 5 Regions & Exposure Combination from Claim Rate's perspective are R11, R22, R25, R42, R53 with exposure Buckets of E4 and E3.

13. 13.1. Are there any cases of illegal driving i.e. underaged folks driving and committing accidents?

```
SELECT count(*)

FROM auto_insurance_risk

where DrivAge < 18;
```

Result:



```
  count(*)
1        0
```

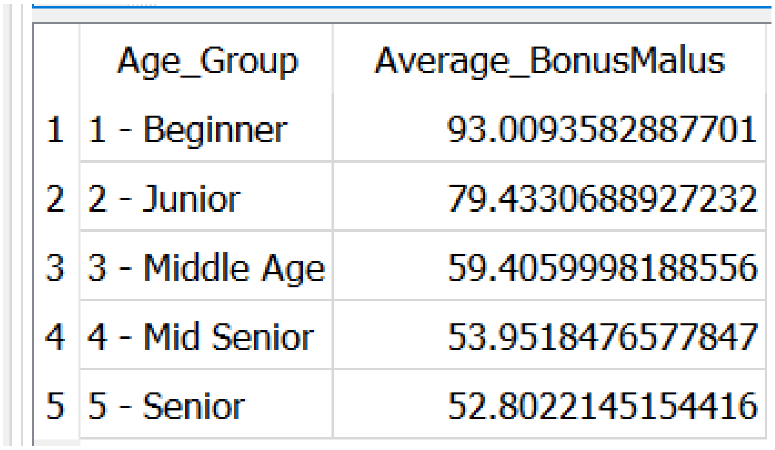
Inference:

No, there are no cases of illegal driving i.e. underaged folks driving and committing accidents.

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?

```
SELECT

CASE

        WHEN DrivAge = 18 THEN "1 - Beginner"

        WHEN DrivAge > 18 and DrivAge <=30 THEN "2 - Junior"

        WHEN DrivAge > 30 and DrivAge <=45 THEN "3 - Middle Age"

        WHEN DrivAge > 45 and DrivAge <=60 THEN "4 - Mid Senior"

        WHEN DrivAge > 60 THEN "5 - Senior"

        END as Age_Group,

        avg(BonusMalus) as Average_BonusMalus

FROM auto_insurance_risk

GROUP BY CASE

        WHEN DrivAge = 18 THEN "1 - Beginner"

        WHEN DrivAge > 18 and DrivAge <=30 THEN "2 - Junior"

        WHEN DrivAge > 30 and DrivAge <=45 THEN "3 - Middle Age"
        WHEN DrivAge > 45 and DrivAge <=60 THEN "4 - Mid Senior"

        WHEN DrivAge > 60 THEN "5 - Senior"

        END;
```

Result:



	Age_Group	Average_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:

We can see that BonusMalus i.e. which penalises them for making claims decreases with age. Therefore the discount given to these customers are much lower than other age groups. This can be due to the fact the that older people have much more experience in driving as compared to younger ones so they are expected to drive cautiously.

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

Primary Key: Primary key is an identifier in s database that references a column in which each value is unique. Only one primary key is allowed to use in a table, thus used to uniquely identify each record in the table. The primary key does not accept the any duplicate and NULL values

Unique Constraint: A column with a unique key constraint can only contain unique values. It is not a mandatory to have a unique key in a table. In this, values cannot have a duplicate.

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?**

    Cross Join: Returns the Cartesian product of the set of records from the two or more joined tables. Therefore, it returns 50 set of records i.e., 5*10 =50.

16. **What is the difference between inner join and left outer join?**

    Inner Join: Inner join returns rows when there is a match in both tables. It creates a new result table by combining column values of two values based upon the join predicate. The query compares each row of table 1 with each row of table 2 to find all pairs of rows which satisfy the join predicate.

    Left outer join: Left Outer join is an operation that returns a combined tuples from a specified table even the join condition will fail. It returns all records from the left table (Table 1) and matching records from the right table (Table 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?**

    Inner join returns rows when there is a match in both tables. Therefore, it returns 1 value.

18. **What is the difference between WHERE clause and HAVING clause?**

    WHERE Clause is used to filter the records from the table based on the specified condition whereas HAVING Clause is used to filter record from the groups based on the specified condition. WHERE Clause can be used without GROUP BY Clause, but HAVING Clause cannot be used without GROUP BY Clause.