

Project Title: Medical Insurance Dataset Analysis and Case Study

Description:

In this project, I utilized SQL to analyze the medical insurance dataset from Kaggle, uncovering key insights into insurance costs and demographic factors. Through detailed data exploration and visualization, the case study provides a comprehensive understanding of the factors affecting insurance premiums.

```
1 SELECT * FROM insurance.insurance_data;
2 use insurance
3 -- 1) Select all columns for all patients
4 select * from insurance_data;
5
6 -- 2) Display the average claim amount for patients in each region
7 select region , avg(claim) as avg_claim from insurance_data
8 group by region;
9
10 -- 3) Select the maximum and minimum BMI values in the table.
11 select max(bmi) as max_bmi, min(bmi) as min_bmi from insurance_data;
12
13 -- 4) Select the PatientID , age, and BMI for patients with a BMI b/w 40 and 50
14 select PatientID , age , bmi from insurance_data where bmi between 40 and 50;
15
16 -- 5) Select the num of smokers in each region.
17 select region , count(PatientID) as num_of_smokers from insurance_data where smoker = "Yes"
18 group by region;
19
20 -- 6) What is the average claim amount for patients who are both diabetic and smokers?
21 select avg(claim) as avg_claim_Amount from insurance_data where diabetic = "Yes" and smoker = "Yes"
22
23 -- 7) Retrieve all patients who have a BMI greater
24 -- than the avg BMI of patients who are smokers.
25 select * from insurance_data where smoker = "Yes" and bmi > (select avg(bmi) from insurance_data where smoker = "Yes");
26 -- select avg(bmi) from insurance_data where smoker = "Yes" ;-- 30.71
27
28 -- 8) Select the avg claim amount for patients in each age group.
29 select
30     case when age < 18 then "Under 18"
31     when age between 18 and 30 then "18-30"
32     when age between 31 and 50 then "31-50"
33     else "Over 50"
34 end as age_group,
35 round(avg(claim),2) as average_claim
36 from insurance_data
37 group by age_group;
38
39 -- 9) Retrieve the total claim amount for each patient,
40 -- along with the avg claim amount across all patients.
41 select PatientID,sum(claim) over(partition by PatientID) as total_claim from insurance_data, -- creating window
42 avg(claim) over() as avg_claim from insurance_data;
43
44 -- 10) Retrieve the top 3 patients with the highest claim amount, along with their
45 -- respective claim amounts and the total claim amount for all patients.
46 select PatientID, claim , sum(claim) over() as total_Claim from insurance_data
47 order by claim desc limit 3;
48
49 -- 11) select the details of patients who have a claim amount
50 -- greater than the avg claim amount for their region
51 select * from insurance_data t1
52 where claim > (select avg(claim) from insurance_data t2 where t2.region = t1.region); # correlation sub query
53
54
55 -- 12) Retrieve the rank of each patient based on their claim amount
56 select * from insurance_data
57 select *,rank() over(order by claim desc) from insurance_data;
58
59 -- 13) Select the details of patients along with their claim amount ,
60 -- and their rank based on claim amount within their region
61 select *,rank() over(order by claim desc) from insurance_data;
62 select * , rank() over(partition by region order by claim desc) from insurance_data
63
```

Code:

```
SELECT * FROM insurance.insurance_data;
```

use insurance

-- 1) Select all columns for all patients

```
select * from insurance_data;
```

-- 2) Display the average claim amount for patients in each region

```
select region , avg(claim) as avg_claim from insurance_data
```

```
group by region;
```

-- 3) Select the maximum and minimum BMI values in the table.

```
select max(bmi) as max_bmi, min(bmi) as min_bmi from insurance_data;
```

-- 4) Select the PatientID , age, and BMI for patients with a BMI b/w 40 and 50

```
select PatientID , age , bmi from insurance_data where bmi between 40 and 50;
```

-- 5) Select the num of smokers in each region.

```
select region , count(PatientID) as num_of_smokers from insurance_data where smoker = "Yes"
```

```
group by region;
```

-- 6) What is the average claim amount for patients who are both diabetic and smokers?

```
select avg(claim) as avg_claim_Amount from insurance_data where diabetic = "Yes" and smoker = "Yes"
```

-- 7) Retrieve all patients who have a BMI greater

-- than the avg BMI of patients who are smokers.

```
select * from insurance_data where smoker = "Yes" and bmi > (select avg(bmi) from insurance_data where smoker = "Yes");
```

```
-- select avg(bmi) from insurance_data where smoker = "Yes" ;-- 30.71
```

-- 8) Select the avg claim amount for patients in each age group.

select

case when age < 18 then "Under 18"

when age between 18 and 30 then "18-30"

when age between 31 and 50 then "31-50"

else "Over 50"

end as age_group,

round(avg(claim),2) as average_claim

from insurance_data

group by age_group;

-- 9) Retrieve the total claim amount for each patient,

-- along with the avg claim amount across all patients.

select PatientID,sum(claim) over(partition by PatientID) as total_claim from insurance_data, --
creating window

avg(claim) over() as avg_claim from insurance_data;

-- 10) Retrieve the top 3 patients with the highest claim amount, along with their

-- respective claim amounts and the total claim amount for all patients.

select PatientID, claim , sum(claim) over() as total_Claim from insurance_data

order by claim desc limit 3;

-- 11) select the details of patients who have a claim amount

-- greater than the avg claim amount for their region

select * from insurance_data t1

where claim > (select avg(claim) from insurance_data t2 where t2.region = t1.region); # correlation
sub query

-- 12) Retrieve the rank of each patient based on their claim amount

select * from insurance_data

select *,rank() over(order by claim desc) from insurance_data;

-- 13) Select the details of patients along with their claim amount ,

-- and their rank based on claim amount within their region

select *,rank() over(order by claim desc) from insurance_data;

select * , rank() over(partition by region order by claim desc) from insurance_data