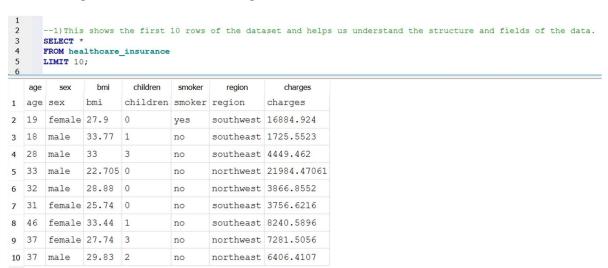
Author: Ria Verma

Dataset Source: https://www.kaggle.com/datasets/willianoliveiragibin/healthcare-insurance

**Goal:** The goal of this project is to analyze medical insurance data to uncover trends and patterns using a mix of basic and intermediate SQL queries. Through this exploratory data analysis project, I want to understand what demographic and lifestyle factors such as age, smoking, and region most impact healthcare insurance costs in America.

# 1) Understanding the Dataset

**1.1 Show First 10 Rows:** Before diving into more advanced queries, it is important to understand the structure of the dataset. This is a simple query to show the first 10 rows of the dataset. We can see that there are 7 fields: age, sex, BMI, children, smoker/non-smoker, region, and insurance charges.



**1.2 Unique Regions:** Now, I want to understand how many and which specific regions we are looking at as part of this dataset. This will help me later when I compare regions and see which region has the highest cost. We can see we are looking at 4 distinct regions.

```
--2) Now, we want to see how many specific regions we are looking at as part of this dataset (using the keyword distinct).

SELECT DISTINCT region

FROM healthcare insurance;

-- We can see are looking at 4 regions: southwest, southeast, northwest, and northeast

region

region

southwest

southeast

northwest

northwest

northeast
```

**1.3 Avg, Min, Max Charges**: Now I want to understand the general range of healthcare insurance costs, so I used the SQL built in functions for average, minimum, and maximum. This gives me a sense of the general insurance costs and some variations.

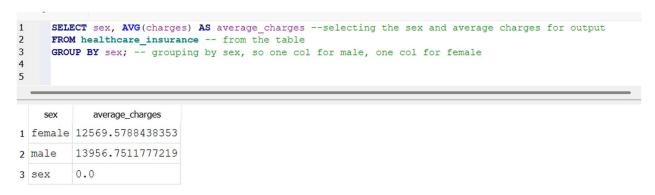
```
11
12
      --3) Summarizing the data to find the average, maximum, and minumum charges.
13
     SELECT
     AVG (charges) AS avg charges,
14
     MIN(charges) AS min_charges,
15
     MAX(charges) AS max_charges -- no comma after last column in select statement
17
      FROM healthcare insurance; -- semicolon ends a sql statement, use commas in between within a single query
18
      avg_charges
                   min_charges max_charges
1 13260.5115689014 10043.249 charges
```

**1.4. Count of Males and Females:** I want to see the approximate ratio of males to females in the dataset, so I used the COUNT function and GROUP BY to group by sex. We can see the ratio is about 1:1.

```
-/*
1
2
      SELECT COUNT (*)
3
      FROM healthcare insurance
4
     WHERE sex = "female";
5
6
     SELECT COUNT (*)
7
     FROM healthcare insurance
     WHERE sex = "male" ; */
8
9
10
      SELECT sex, COUNT (*)
11
     FROM healthcare insurance
12
      GROUP BY sex;
13
14
      --SELECT DISTINCT sex
15
      --FROM healthcare insurance;
```

	sex	COUNT(*)
1	female	662
2	male	676
3	sex	1

#### 1.5 Gender Distribution of Charges



This query helps explore if gender influences healthcare charges. From the output, we can see that males tend to have higher average charges.

# 2) Basic SQL Queries

#### 2.1 Count of Smokers and Non-Smokers:

```
SELECT smoker, COUNT(*) AS count -- selecting the value of smoker so we can see if its yes or no +count
FROM healthcare_insurance -- from our table
GROUP BY smoker; -- group by smoker means we want to group the data BY the column smoker and return vals

smoker count
no 1064
smoker 1
yes 274
```

This helps us see how many smokers versus nonsmokers are in this dataset and quantify the impact when we later on see how this impacts insurance costs.

## 2.2 Impact of Smoking on Charges

```
SQL 1* SELECT smoker, AVG(charges) -- select the smoker and the average charges
FROM healthcare_insurance -- from the table
GROUP BY smoker; -- group[ by whether they are smoker or not (so yes or no)

smoker AVG(charges)
1 no 8434.2682978562
2 smoker 0.0
3 yes 32050.2318315328
```

This query compares average charges between smokers and non smokers, which highlights the financial impact of smoking on healthcare costs.

### 2.2b Impact of Smoking on Charges

```
SQL 1* Select the result of subtracting two conditional averages

SELECT

AVG (CASE WHEN smoker = 'yes' THEN charges END) -- Calculate average charges for smokers only
- AVG (CASE WHEN smoker = 'no' THEN charges END) -- Subtract average charges for non-smokers
AS avg_charge_difference -- set the name of the result as avg_charge_difference

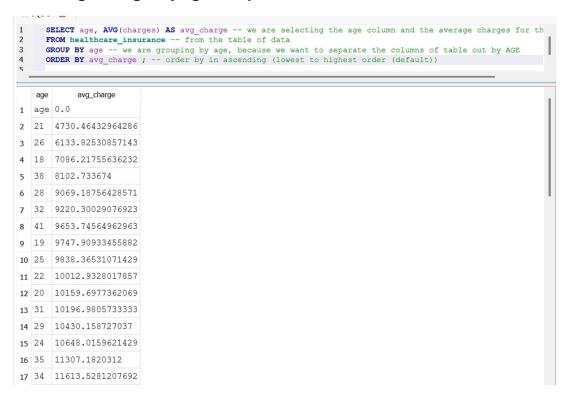
FROM healthcare_insurance; -- From the insurance table

avg_charge_difference

23615.9635336766
```

This query just takes the difference between the average charges of people who smoke versus people who don't smoke. This tells us that on average, people who smoke have to pay more than 23000\$ in insurance costs then people who do not.

### 2.3 Average Charges by Age Group

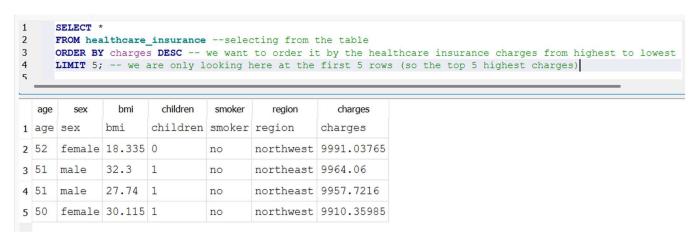


J.	age	avg_charge
32	50	15663.0033006897
33	51	15682.2558672414
34	44	15859.396587037
35	53	16020.930755
36	55	16164.5454884615
37	57	16447.18525
38	47	17653.9995931035
39	37	18019.9118772
40	52	18256.2697193103
41	54	18758.5464753571
42	59	18895.8695316
43	62	19163.8565734783
44	43	19267.2786533333
45	63	19884.9984608696
46	60	21979.4185073913
47	61	22024.4576086957
48	64	23275.5308372727

This query helps me analyze how age affects the cost of insurance. I am looking at the average charges for each age group and ordering them from the lowest to highest charge. The general trend is that as age increases, so do healthcare insurance costs, however there are some outliers to this trend such as maybe age 19 and age

3) Intermediate SQL Queries

#### 3.1. Top 5 people who have the highest healthcare charges.

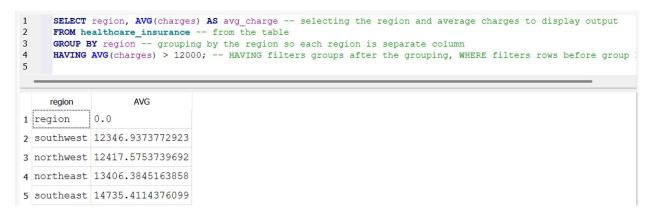


This query identifies outliers and the people who have paid extremely high charges. These may be influenced by smoking, age, or BMI. Some similarities I am seeing through these results are that the regions are all in the northern part and they are in the age range of 50s.

#### 3.2. Average Charges by Region

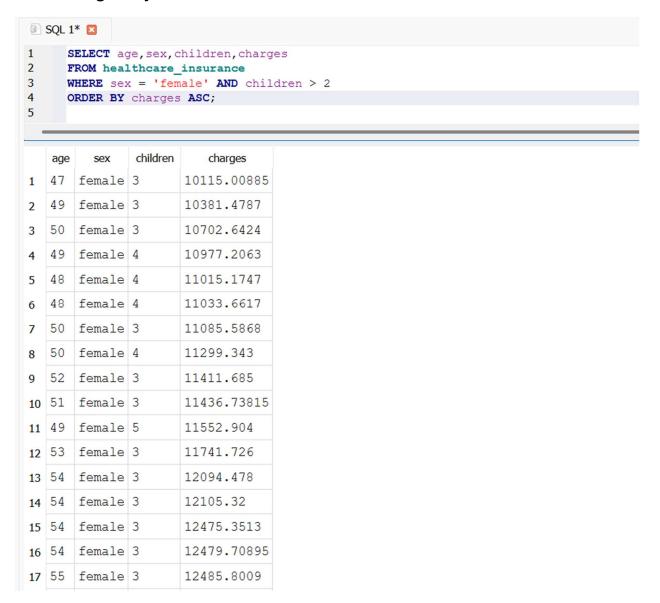
This query compares insurance charges by region and lets us see which regions have higher versus lower charges. The region with the highest charges is the southeast region and region with lowest is the southwest region.

#### 3.3 Filtering out only high average charges which exceed \$12,000



This query selects only the regions which have average charges over 12,000 to pinpoint which regions may be driving up insurance costs the most. We can see the southeast region seems to be the highest cost region.

#### 3.4. Filtering out by all females with more than 2 children



	age	sex	children	charges
19	52	female	5	12592.5345
20	55	female	3	13047.33235
21	56	female	3	13430.265
22	59	female	3	14001.1338
23	59	female	3	14001.2867
24	59	female	3	14007.222
25	59	female	3	14382.70905
26	57	female	4	14394.39815
27	59	female	3	14590.63205
28	62	female	3	15612.19335
29	40	female	4	15828.82173
30	64	female	3	16085.1275
31	27	female	3	16420.49455
32	64	female	3	16455.70785
33	18	female	3	18223.4512
34	30	female	3	18765.87545
35	27	female	3	18804.7524

This query filters out all females who have more than 2 children and arranges the results in order from lowest to highest. This allows us to explore if having dependents such as number of children increases charge and if coverage differs for larger families.

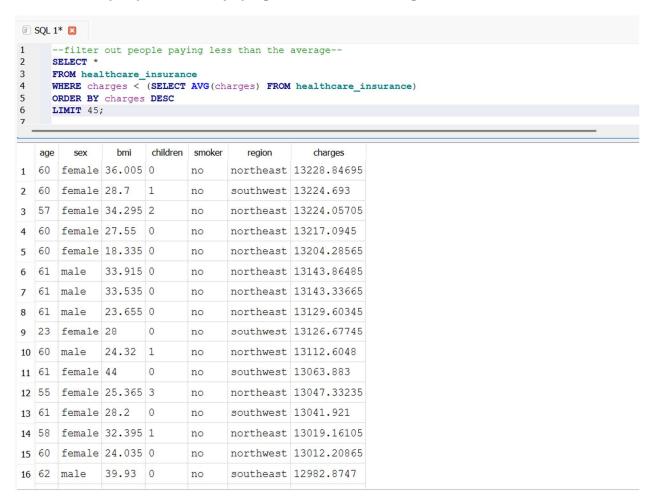
#### 3.5 Charges based on number of children

```
SELECT children, AVG(charges) AS average_charges -- selecting the children col. and average charges
    FROM healthcare_insurance --from the table
2
3
    GROUP BY children -- grouping by children for the results, 0 1 2 3 4 children
4
    ORDER BY children ASC; --ordering from highest to lowest charges
   children
              average_charges
         12365.9756016359
1 0
2 1
           12731.1718316358
3 2
           15073.5637339583
4 3
           15355.3183668153
5 4
          13850.6563112
6 5
         8786.03524722222
7 children 0.0
```

This query explores how the number of children affects healthcare insurance costs and can inform policy changes about family insurance coverage. We can see families with more children tend to have lower insurance charges and vice versa.

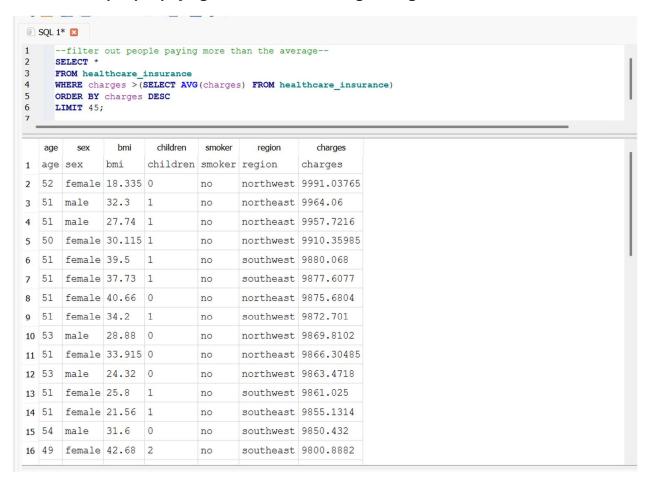
# 4) Using Subqueries

### 4.1 Filter out people who are paying less than the average



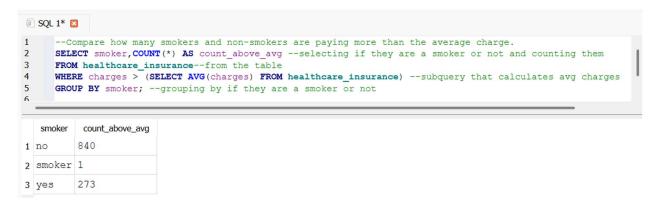
This query filters out the first 45 people who are paying less than the average charges and orders them from greatest to least. Some insights from this are that the top demographics who are paying less than the average charges are females in the age range of late 50s-60 who are non smokers. From 1.3, the average charge is \$13260.

#### 4.2 Filter out people paying more than the average charge



This filters out people who are paying more than the average charge from greatest to least. Most seem to be in the 50s age range and have 0-1 children.

### 4.3 Show smokers who pay more than the average charges.



This query compares and counts the number of smokers versus non smokers who are paying more than the average charges. We can see that people who do not smoke and paying above the average.