Healthcare Insurance in the United States Dataset (SQL Queries and Outputs)

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.

```sql

--1)This shows the first 10 rows of the dataset and helps us understand the structure and fields of the data.
SELECT *
FROM healthcare_insurance
LIMIT 10;

```

	age	sex	bmi	children	smoker	region	charges
1	age	sex	bmi	children	smoker	region	charges
2	19	female	27.9	0	yes	southwest	16884.924
3	18	male	33.77	1	no	southeast	1725.5523
4	28	male	33	3	no	southeast	4449.462
5	33	male	22.705	0	no	northwest	21984.47061
6	32	male	28.88	0	no	northwest	3866.8552
7	31	female	25.74	0	no	southeast	3756.6216
8	46	female	33.44	1	no	southeast	8240.5896
9	37	female	27.74	3	no	northwest	7281.5056
10	37	male	29.83	2	no	northeast	6406.4107

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.

```sql
--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
1	region
2	southwest
3	southeast
4	northwest
5	northeast

Healthcare Insurance in the United States Dataset (SQL Queries and Outputs)

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.

```sql
--3)Summarizing the data to find the average, maximum, and minumum charges.
SELECT
AVG(charges) AS avg_charges,
MIN(charges) AS min_charges,
MAX(charges) AS max_charges -- no comma after last column in select statement
FROM healthcare_insurance; -- semicolon ends a sql statement, use commas in between within a single query
```

	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.

```sql
/*
SELECT COUNT(*)
FROM healthcare_insurance
WHERE sex = "female" ;

SELECT COUNT(*)
FROM healthcare_insurance
WHERE sex = "male" ; */

SELECT sex,COUNT(*)
FROM healthcare_insurance
GROUP BY sex;

--SELECT DISTINCT sex
--FROM healthcare_insurance;
```

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

1.5 Gender Distribution of Charges

```sql
SELECT sex, AVG(charges) AS average_charges --selecting the sex and average charges for output
FROM healthcare_insurance -- from the table
GROUP BY sex; -- grouping by sex, so one col for male, one col for female
```

	sex	average_charges
1	female	12569.5788438353
2	male	13956.7511777219
3	sex	0.0

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:

```sql
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
1	no	1064
2	smoker	1
3	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*

```sql
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*

```sql
-- 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
1	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

```sql
SELECT age, AVG(charges) AS avg_charge -- we are selecting the age column and the average charges for th
FROM healthcare_insurance -- from the table of data
GROUP BY age -- we are grouping by age, because we want to separate the columns of table out by AGE
ORDER BY avg_charge ; -- order by in ascending (lowest to highest order (default))
```

	age	avg_charge
1	age	0.0
2	21	4730.46432964286
3	26	6133.82530857143
4	18	7086.21755636232
5	38	8102.733674
6	28	9069.18756428571
7	32	9220.30029076923
8	41	9653.74564962963
9	19	9747.90933455882
10	25	9838.36531071429
11	22	10012.9328017857
12	20	10159.6977362069
13	31	10196.9805733333
14	29	10430.158727037
15	24	10648.0159621429
16	35	11307.1820312
17	34	11613.5281207692

	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

Healthcare Insurance in the United States Dataset (SQL Queries and Outputs)

3) Intermediate SQL Queries

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

```sql
SELECT *
FROM healthcare_insurance --selecting from the table
ORDER BY charges DESC -- we want to order it by the healthcare insurance charges from highest to lowest
LIMIT 5; -- we are only looking here at the first 5 rows (so the top 5 highest charges)
```

	age	sex	bmi	children	smoker	region	charges
1	age	sex	bmi	children	smoker	region	charges
2	52	female	18.335	0	no	northwest	9991.03765
3	51	male	32.3	1	no	northeast	9964.06
4	51	male	27.74	1	no	northeast	9957.7216
5	50	female	30.115	1	no	northwest	9910.35985

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

```sql
SELECT region,AVG(charges) as AVG -- selecting the region and the avg charges to show up in output
FROM healthcare_insurance -- from the table
GROUP BY region -- group by region so each column of avg charges shows the avg for THAT region.
ORDER BY AVG -- ordering default from lowest to highest
```

	region	AVG
1	region	0.0
2	southwest	12346.9373772923
3	northwest	12417.5753739692
4	northeast	13406.3845163858
5	southeast	14735.4114376099

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

```sql
SELECT region, AVG(charges) AS avg_charge -- selecting the region and average charges to display output
FROM healthcare_insurance -- from the table
GROUP BY region -- grouping by the region so each region is separate column
HAVING AVG(charges) > 12000; -- HAVING filters groups after the grouping, WHERE filters rows before group
```

	region	AVG
1	region	0.0
2	southwest	12346.9373772923
3	northwest	12417.5753739692
4	northeast	13406.3845163858
5	southeast	14735.4114376099

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

SQL 1*

```sql
SELECT age,sex,children,charges
FROM healthcare_insurance
WHERE sex = 'female' AND children > 2
ORDER BY charges ASC;
```

	age	sex	children	charges
1	47	female	3	10115.00885
2	49	female	3	10381.4787
3	50	female	3	10702.6424
4	49	female	4	10977.2063
5	48	female	4	11015.1747
6	48	female	4	11033.6617
7	50	female	3	11085.5868
8	50	female	4	11299.343
9	52	female	3	11411.685
10	51	female	3	11436.73815
11	49	female	5	11552.904
12	53	female	3	11741.726
13	54	female	3	12094.478
14	54	female	3	12105.32
15	54	female	3	12475.3513
16	54	female	3	12479.70895
17	55	female	3	12485.8009

Healthcare Insurance in the United States Dataset (SQL Queries and Outputs)

	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

```sql
SELECT children, AVG(charges) AS average_charges -- selecting the children col. and average charges
FROM healthcare_insurance --from the table
GROUP BY children -- grouping by children for the results, 0 1 2 3 4 children
ORDER BY children ASC; --ordering from highest to lowest charges
```

	children	average_charges
1	0	12365.9756016359
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

SQL 1*

```sql
--filter out people paying less than the average--
SELECT *
FROM healthcare_insurance
WHERE charges < (SELECT AVG(charges) FROM healthcare_insurance)
ORDER BY charges DESC
LIMIT 45;
```

	age	sex	bmi	children	smoker	region	charges
1	60	female	36.005	0	no	northeast	13228.84695
2	60	female	28.7	1	no	southwest	13224.693
3	57	female	34.295	2	no	northeast	13224.05705
4	60	female	27.55	0	no	northeast	13217.0945
5	60	female	18.335	0	no	northeast	13204.28565
6	61	male	33.915	0	no	northeast	13143.86485
7	61	male	33.535	0	no	northeast	13143.33665
8	61	male	23.655	0	no	northeast	13129.60345
9	23	female	28	0	no	southwest	13126.67745
10	60	male	24.32	1	no	northwest	13112.6048
11	61	female	44	0	no	southwest	13063.883
12	55	female	25.365	3	no	northeast	13047.33235
13	61	female	28.2	0	no	southwest	13041.921
14	58	female	32.395	1	no	northeast	13019.16105
15	60	female	24.035	0	no	northwest	13012.20865
16	62	male	39.93	0	no	southeast	12982.8747

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

SQL 1*

```sql
--filter out people paying more than the average--
SELECT *
FROM healthcare_insurance
WHERE charges >(SELECT AVG(charges) FROM healthcare_insurance)
ORDER BY charges DESC
LIMIT 45;
```

	age	sex	bmi	children	smoker	region	charges
1	age	sex	bmi	children	smoker	region	charges
2	52	female	18.335	0	no	northwest	9991.03765
3	51	male	32.3	1	no	northeast	9964.06
4	51	male	27.74	1	no	northeast	9957.7216
5	50	female	30.115	1	no	northwest	9910.35985
6	51	female	39.5	1	no	southwest	9880.068
7	51	female	37.73	1	no	southeast	9877.6077
8	51	female	40.66	0	no	northeast	9875.6804
9	51	female	34.2	1	no	southwest	9872.701
10	53	male	28.88	0	no	northwest	9869.8102
11	51	female	33.915	0	no	northeast	9866.30485
12	53	male	24.32	0	no	northwest	9863.4718
13	51	female	25.8	1	no	southwest	9861.025
14	51	female	21.56	1	no	southeast	9855.1314
15	54	male	31.6	0	no	southwest	9850.432
16	49	female	42.68	2	no	southeast	9800.8882

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.

SQL 1*

```sql
--Compare how many smokers and non-smokers are paying more than the average charge.
SELECT smoker,COUNT(*) AS count_above_avg --selecting if they are a smoker or not and counting them
FROM healthcare_insurance--from the table
WHERE charges > (SELECT AVG(charges) FROM healthcare_insurance) --subquery that calculates avg charges
GROUP BY smoker; --grouping by if they are a smoker or not
```

	smoker	count_above_avg
1	no	840
2	smoker	1
3	yes	273

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.