USE insurance

-- First, I want to determine the percentage of smokers and non-smokers in this dataset. I will view these both as counts as well as percentages.

SELECT COUNT(\*) AS total,

COUNT(CASE WHEN smoker = 'yes' THEN 1 ELSE NULL END) AS smokers,

COUNT(CASE WHEN smoker = 'no' THEN 1 ELSE NULL END) AS non\_smokers,

ROUND(COUNT(CASE WHEN smoker = 'yes' THEN 1 ELSE NULL END) / COUNT(\*) \* 100, 2) AS percent\_smokers,

ROUND(COUNT(CASE WHEN smoker = 'no' THEN 1 ELSE NULL END) / COUNT(\*) \* 100, 2) AS percent\_non\_smokers

FROM insurance;

-- Although this provides me with the information I need, I would like to display it in a more visually intuitive way using the following query.

SELECT smoker,

COUNT(\*) AS count,

ROUND(COUNT(\*) / SUM(COUNT(\*)) OVER () \* 100, 2) AS percent

FROM insurance

GROUP BY smoker;

-- I can now see that smokers represent only 20.48% of the individuals in this dataset.

-- Next I want to break down the percentage of smokers by gender.

SELECT ROUND(COUNT(CASE WHEN sex = 'female' AND smoker = 'yes' THEN 1 END) / COUNT(CASE WHEN sex = 'female' THEN 1 END) \* 100, 2) AS percent\_female\_smokers,

ROUND(COUNT(CASE WHEN sex = 'male' AND smoker = 'yes' THEN 1 END) / COUNT(CASE WHEN sex = 'male' THEN 1 END) \* 100, 2) AS percent\_male\_smokers

FROM insurance;

-- 17.37% of females are smokers while 23.52% of males in this dataset are smokers.

-- Now I want to drill down and see how smoking and gender correlate with average charges. I begin by looking strictly at smoking across both genders.

SELECT smoker, AVG(charges) as average\_expenses

FROM insurance

GROUP BY smoker

-- There certainly seems to be a correlation, with smokers incurring charges that are, on average, 3.8 times greater than non-smokers.

-- Now I will analyze how these charges break down by both smoking status and gender.

SELECT smoker, sex, ROUND(AVG(charges), 2) as average\_expenses

FROM insurance

GROUP BY smoker, sex

ORDER BY sex

-- I would like a query that is easier to understand at a glance. A CASE statement should achieve this.

SELECT sex,

AVG(CASE WHEN smoker = 'yes' THEN charges ELSE NULL END) AS smoker\_charges,

AVG(CASE WHEN smoker = 'no' THEN charges ELSE NULL END) AS non\_smoker\_charges

FROM insurance

GROUP BY sex

-- I will also round two decimal places to make the output easier to understand.

SELECT sex,

COUNT(\*) AS count,

ROUND(AVG(CASE WHEN smoker = 'yes' THEN charges ELSE NULL END), 2) AS smoker\_charges,

ROUND(AVG(CASE WHEN smoker = 'no' THEN charges ELSE NULL END), 2) AS non\_smoker\_charges

FROM insurance

GROUP BY sex;

-- I can now see that smokers incur 7.41% higher average charges than female smokers.

-- I now want to see if smokers in certain regions incur higher charges than others.

SELECT region,

ROUND(AVG(CASE WHEN smoker = 'yes' THEN charges END), 0)AS avg\_smoker\_charges,

ROUND(AVG(CASE WHEN smoker = 'no' THEN charges END), 0)AS avg\_non\_smoker\_charges

FROM insurance

GROUP BY region;

-- The northeast has the highest average charges for non-smokers, but it also has the lowest average charges for smokers.

-- The next query will evaluate whether average charges correlate with number of children one has. Do average charges increase with the number of children one has?

SELECT

children,

COUNT(\*) AS count,

ROUND(AVG(charges),2) AS avg\_charges

FROM insurance

GROUP BY children

ORDER BY avg\_charges DESC;

-- Individuals with 4-5 children are outliers in this dataset. Only 18 people out of 1338 (or 1.3%) have 5 children, and just 25 people(or 1.8%) have 4 children.

-- I will modify the code to exlude the records where children equal 4 or 5.

SELECT

children,

COUNT(\*) AS count,

ROUND(AVG(charges),2) AS avg\_charges

FROM insurance

WHERE children NOT IN (4, 5)

GROUP BY children

ORDER BY avg\_charges DESC;

-- When those outliers are removed, I can see clearly that there is a positive correlation between number of children and average charges.

-- Average charges decreases as the number of children decreases

-- I will now analyze the data to see what correlations may exist between BMI and average charges.

-- In analyzing BMI, I based my CASE classifications on those used by the CDC : https://www.cdc.gov/obesity/basics/adult-defining.html

SELECT

CASE

WHEN bmi < 18.5 THEN 'Underweight'

WHEN bmi >= 18.5 AND bmi < 25 THEN 'Normal'

WHEN bmi >= 25 AND bmi < 30 THEN 'Overweight'

WHEN bmi >=30 THEN 'Obese'

END AS bmi\_category,

ROUND(AVG(charges),2) AS avg\_charges,

COUNT(\*)

FROM insurance

GROUP BY bmi\_category

ORDER BY avg\_charges DESC

-- Surpringly, the majority of people in this dataset fell into the "Obese" range. As can be seen below, the lower one's weight, the lower the average charges. As weight

-- decreases, so to do average charges.

-- It should be noted that there are only 20 individuals classified as "underweight".

-- Next I want to dig deeper into the data and see how obesity breaks down by both BMI and smoker status.

SELECT

CASE

WHEN smoker = 'yes' THEN 'smoker'

ELSE 'non-smoker'

END AS smoking\_status,

CASE

WHEN bmi < 18.5 THEN 'underweight'

WHEN bmi >= 18.5 AND bmi < 25 THEN 'normal'

WHEN bmi >= 25 AND bmi < 30 THEN 'overweight'

WHEN bmi >= 30 THEN 'obese'

END AS bmi\_category,

ROUND(AVG(charges), 2) AS avg\_charges,

COUNT(\*) AS count

FROM

insurance

GROUP BY smoking\_status , bmi\_category

ORDER BY count DESC

-- Smokers of normal weight still spend 2.6 times that of non-smokers of normal weight. It would seem that even if one is at a healthy weight, smoking alone still

-- drastically increases charges incurred. When we look at smokers who are classified as obese, however, we see an even more pronounced increase in charges. An obese smoker

-- will incur 41558 in charges compared to an obese non-smoker who will incur an average of 8842. So the obese smoker will spend, on average, 4.7 times that of an obese non-smoker. It is

-- very clear that smokers will incur higher average charges than non-smokers across all BMI classifications.

-- Finally, I want to look at age and average charges. Because I expect charges to increase with age, I am going to use a descending order on the age variable.

SELECT age, AVG(charges) AS average\_charges,

COUNT(\*) AS count

FROM insurance

GROUP BY age

ORDER BY age DESC;

-- Even though average charges vary across different age groups, it can still be seen that average charges increase with age.

-- It is worth noting that 18 or 19 year olds make up the largest age groups in this dataset, so for some analysis, there could be problems in terms of the representativeness of the data.