SQL PROJECT

HEALTHCARE DATA ANALYSIS

Tools used: MySQL Workbench.

Dataset was gotten from Kaggle in a CSV format.

Dataset contained 7 columns which contained medical insurance cost information. It included demographic and health-related variables such as age, sex, BMI, number of children, smoking status, and residential region in the US.

The analysis was done with the following objectives in mind:

- Explore the demographic profile of insured individuals.
- Assess the role of BMI and smoking on healthcare charges.
- Identify high-risk groups higher medical costs.
- Analyze the impact of family size and region on charges.

Data analysis process

The tools used was MySQL Workbench. First, I created a new schema in Workbench. Then I imported the CSV file using the Table Data Import Wizard. Created a new table for the dataset and ensure the import configurations were right.

After importation, I created a duplicate of the dataset for backup purposes. Then checked all columns info to note that columns were correctly named and the data was of the right type. Noted those that weren't up to standard. I also added an ID column.

I used the ALTER TABLE and RENAME COLUMN commands to rename a few of the columns that weren't up to naming conventions. After that, I toggled with the SQL Set Updates settings to allow for modifications. Adjusted datatypes using the MODIFY COLUMN function.

Next came checking for duplicates. I used the ROW_NUMBER function in a CTE and removed them. I also checked for nulls and blanks, none were present. Next, I standardized the data. Adjusted spelling errors and capitalization. Did one last check to ensure everything was alright.

Also created new columns for grouping ages and BMIs into different groups for easier analysis.

1. Demographics

With an average age is 39.22 years (range: 18–64), our dataset contains a total of 1337 unique individuals, of which 662 are females and 675 are male, **showing an almost even split**. Analysis shows that the average charge for males is $13,975 and $12,569.58 for females.

Ages were split into categories for better understanding and analysis. Youth (18-29), Young Adults (30-39), Middle Age (40-49), and Seniors (50-64). Youths make up a greater portion of the population with 416 individuals, followed by Seniors (385), Middle age (279) and lastly, Young Adults (257). Despite the youth making up the greater part of the population, on average, they have the lowest charges. Seniors take the lead with an average of $17.9k, followed by Middle Age ($14.4k), then Young Adults ($11.7k) and lastly Youth ($9.2k). **This shows a clear upward trend i.e. older = higher charges.**

2. BMI & Health Risk

The dataset has an **a**verage BMI of 30.66 (range: 15.96 - 53.13). For better understanding the BMI was split into categories: Underweight (<18.5), Normal (18.5-24.9), Overweight (25-29.9), Obese (30+).

Obese takes the lead with 706 individuals, followed by Overweight (386), then Normal (225), and lastly Underweight (20). On average, the Obese category has the highest charge ($15.5k), followed by Overweight ($10.9k), then Normal ($10.4k), and lastly Underweight ($8.8k).

Insights: Obese individuals (53% of the dataset) have the highest average charges ($15.5k), nearly 50% more than Normal weight individuals ($10.4k). Underweight cases are rare (20 people) and show lower charges. The clear upward trend from Normal → Overweight → Obese suggests a strong correlation between BMI and insurance costs.

3. Lifestyle (Smoking)

Of a total of 1337 individuals, 1063 are non-smokers, and 274 are smokers. Overall average charge is $13.2k, but the average charge for smokers is $32k and $8.4k for non-smokers. **Smokers pay charges 3.8x higher than non-smokers.**

Age group	Smoker Avg charge	Non-smoker Avg charge	Insight
Youth	$27.5k	$4.4k	Noticeable gaps
Young Adult	$30.2k	$6.3k	Sharp rise in smoker costs
Middle Age	$32.6k	$9.1k	Impact of smoking widens
Senior	$38.7k	$13.4k	Burden grows too high

effects of smoking by age group.

Insight: Smoking increases the costs as age increases, showing cumulative health effects.

BMI group	Smoker Avg charge	Non-smoker Avg charge	Insight
Underweight	$18.8	$5.5k	Noticeable gaps
Normal	$19.9k	$7.6k	Smoking triples costs
Overweight	$22.5k	$8.2k	Consistent spike
Obese	$41.5k	$8.8k	Worst case scenario

effects of smoking by BMI group.

Insight: The combination of Smoking and Obesity is a risky one, leading to the highest insurance charges overall.

Insights & Analysis

Smoking has a direct and dramatic effect on insurance charges, independent of age or BMI. Also, older smokers face the steepest costs due to cumulative health risks. Obese smokers are the highest-cost group, with average charges nearly 5x non-smokers overall.

These findings highlight smoking as a major driver of insurance costs — a key factor policy makes must take into account.

4. **Family & Dependents**

Majority of families have 0–2 children, while families with 4 or 5 children are rare.

Children	Count	Avg charge
0	573	$12.3k
1	324	$12.7k
2	240	$15k
3	157	$15.3k
4	25	$13.8k
5	18	$8.7k

Insights: Families with 2–3 children have the highest average charges (~$15k+). Families with no children or 1 child are slightly lower (~$12k+). Families with 5 children have the lowest charges.

The relationship between children and costs is not strictly linear. Costs increase up to 2–3 children and then dips for 4–5 children. This suggests that other factors influence charges more than just number of children.

Children	Non-Smoker Avg charge	Smoker Avg charge	Insight
0	$7.6k	$31.3k	Largest cost gap
1	$8.3k	$31.8k	Smokers ~4x higher
2	$9.5k	$33.8k	Peak smoker cost
3	$9.6	$32.7k	Consistently high
4	$12.1k	$26.5k	Gap narrows slightly
5	$8.1k	$19k	Both groups lower

smoking + children

Insights: Smoking dominates charges — families with smokers pay 3–4x more regardless of the number of children. Families with 2–3 children who smoke face the highest costs (~$33k). Families with many children (4–5) still show lower average charges. Summarily, children alone do not strongly predict charges; smoking and BMI are stronger cost drivers.

Overall, charges peak for families with 2–3 children, especially when smoking is involved. Smoking remains the strongest factor in driving family insurance costs, overshadowing number of dependents.

5. Regional Insights

The data is spread over 4 regions: Southeast with 364 individuals and an average charge of $14.7k, the highest on both counts. It is followed by Southwest with 325 individuals and a $12.3k average. Northwest and Northeast have the same number of individuals: 324 but differing average. Northwest has an average of $12.4k and Northeast has $13.4k. Southeast has the highest average charges ($14.7k) and Southwest has the lowest average charges ($12.3k).

Analysis of each region by smokers' prevalence and average charges revealed some insights. Smokers are most common in the Southeast (91) and carry the highest average charge ($34.8k). In all regions, the average charges for non-smokers stay between $8k–9k. Regional differences are mostly explained by the number of smokers in the region.

Analysis of charges by region + BMI group shows that Obese individuals drive costs across all regions, with the highest costs in Northeast & Southeast ($16k+). Normal BMI charges vary sharply in some regions, Normal individuals in Southeast pay around $13.1k, which is the highest, while Normal individuals in Southwest pay about $7.2k (the lowest). Across all regions, the Underweight group remains the lowest cost.

Analysis of charges by region + age group shows that cross all regions, Seniors consistently drive the highest charges, followed by Middle Age groups. Particularly, Seniors in the Southeast have the highest average costs ($20.4k). Youths are the least costly, though Southeast youths still average $10k+, which is higher than other regions.

Overall, Southeast is the most expensive region, largely due to high smoker prevalence and high senior costs. The costliest combination still remains Smokers + Obesity + Senior age group, across all regions. Lastly, non-smokers and youths show the most stable and lowest costs regardless of location.

6. Financial Distribution & Outliers

Minimum charge: $1,121.87

Maximum charge: $63,770.40

Average charge: $13,279.12

Variation in charges is huge, with the highest case being 55x more than the lowest.

For further and deeper analysis, the charges were split into groups.

Charge Group	Count	Smoker Split	Insight
Low (<5k)	358	All non-smokers	These are mainly healthier, non-smoking, lower-risk individuals.
Medium (5k–20k)	706	644 non-smokers, 62 smokers	The "majority group" (~53%), still mostly non-smokers.
High (>20k)	273	61 non-smokers, 212 smokers	Heavily dominated by smokers.

Insights: As can be seen from the above, smokers dominate the high-charge group as about 78% of high-cost patients are smokers. The low-charge group has zero smokers, highlighting the direct cost burden of smoking.

Further analysis was carried out on the top 10 highest charge patients; the following were what they all had in common: all 10 are smokers and all fall into Obese category. Majority have 0–2 children, and they are mostly from the Southeast and the Northwest. Lastly, they are spread across all age groups, but the older groups are more.

Summary: The most extreme costs are tightly clustered among obese smokers, particularly in the Southeast region.

Obese Smokers have average charges > $41k, the single highest-risk lifestyle factor. Southeast Smokers have average charges of about $34.8k, the highest regional burden. Lastly, Senior Smokers consistently have averages of $38k+, making age and smoking the most expensive combo.

FINAL INSIGHTS

1. Smoking is the strongest cost driver. Smokers pay 3–6x more than non-smokers.
2. Obesity + smoking amplifies charges dramatically (obese smokers average $41k vs obese non-smokers $8.8k).
3. Age magnifies risk as senior smokers average ~$38k, compared to ~$13k for senior non-smokers.
4. Southeast region is costliest, especially for smokers and seniors.
5. Families with 0–3 kids see higher charges, but families with 4–5 kids show lower averages (possible area for further analysis with more data).
6. A small group of high-cost outliers (mostly obese smokers in the Southeast) drive a disproportionate share of spending.

CONCLUSIONS

This analysis highlights how lifestyle choices (especially smoking and obesity), combined with age and region, significantly impact medical insurance costs. These insights derived can be used to:

- Design targeted interventions for high-risk groups.
- Adjust premiums or wellness programs based on smoking/BMI profiles.
- Investigate regional disparities, particularly in the Southeast.