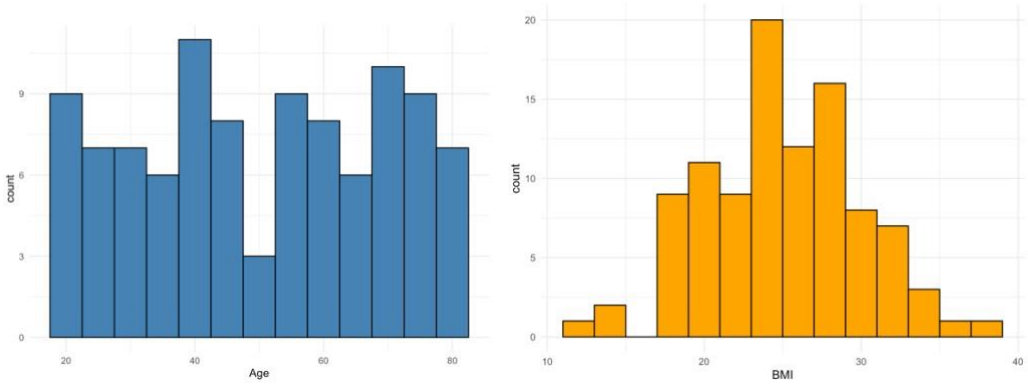
Multiple Linear Regression Report

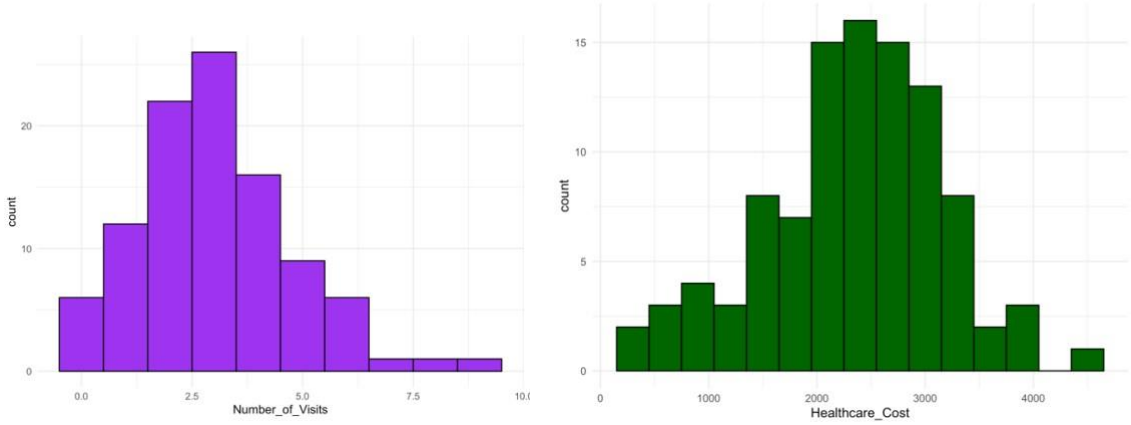
1. Data Exploration:

The dataset includes health and lifestyle information for 100 patients. The goal was to understand what factors affect healthcare costs the most.

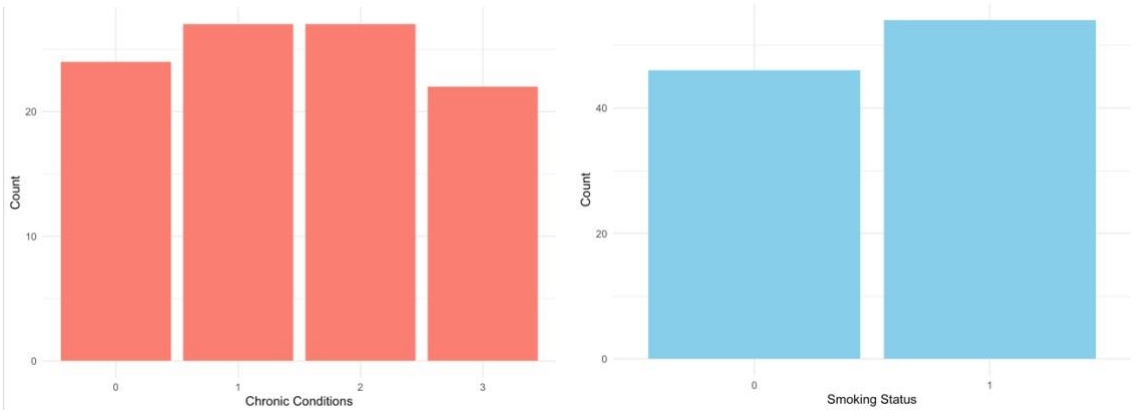
- Age: Patients are evenly spread across ages 19 to 79. So, there are both younger and older adults.



- BMI: Most patients fall within a healthy to slightly overweight range (BMI 20–30), and the shape of the graph looks close to a bell curve.



- Number of Hospital Visits: Most people visited the hospital 2–4 times last year. Very few had many visits.
- Healthcare Costs: Costs were mostly between 2000 and 3000 units. The distribution looked roughly normal, which is good for regression.



- Chronic Conditions: The number of chronic conditions per person is balanced, some had none, others had 1–3.
- Smoking Status: Slightly more smokers than non-smokers.

There was no missing data, all columns were complete and ready for analysis.

2. **Model Building**

I wanted to predict a person's total healthcare cost using the following: Age, BMI, Smoking Status, Number of hospital visits, Number of chronic conditions

The assumption was: The more health problems or risk factors someone has, the more they'll likely spend on healthcare.

3. Model Results

I first ran a full model with all five predictors. Then refined it by removing the ones that weren't statistically important.

Original model findings:

- Age and chronic conditions had a strong, significant impact on healthcare costs.
- BMI and smoking status didn't seem to matter much in this case.
- Number of visits had a small effect, but not strongly significant.

Refined model

I kept just age, number of visits, and chronic conditions.

- The model still explained about 49% of the variation in healthcare costs, which is very close to the full model. All predictors in this model were useful and easy to interpret.

Therefore, from refined model:

Older patients spend more, about 27 more units for every year of age. Each hospital visit adds around 63 units to the cost. Each chronic condition increases cost by about 136 units.

4. Model Check:

I used 4 diagnostic plots to check if the model followed important assumptions:

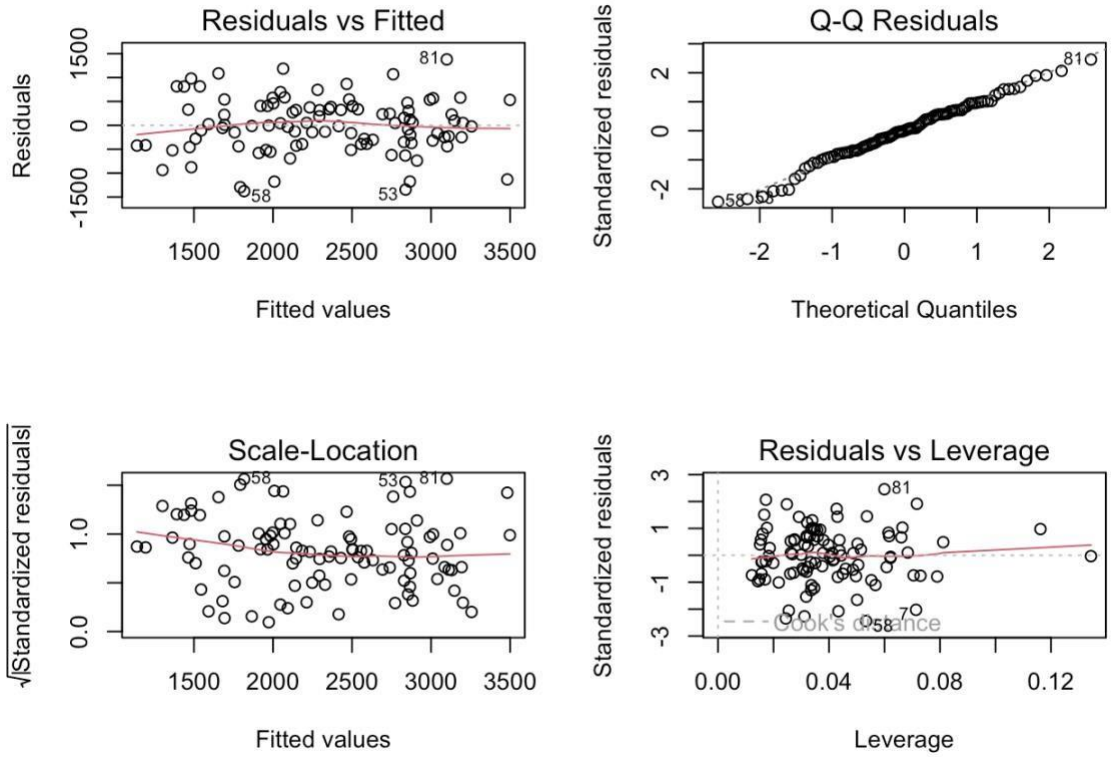
Residuals vs Fitted: Points were scattered randomly, the relationship between predictors and cost looks good.

Q-Q Plot: Residuals mostly followed a straight line; they are approximately normally distributed.

Scale-Location: Spread of points was even, the model's errors are consistent across the predictions.

Residuals vs Leverage: A couple of points had slightly more influence (like patient 58 and 81), but none were strong enough to hurt the model.

The model passed all the basic checks and is reliable.



5. Summary:

The refined model is the better choice, because it is easier to explain, it performs just as well as the complex one, it focuses only on the factors that really matter.
In this dataset, the main factors that drive healthcare costs are older age, a higher number of hospital visits, and having more chronic conditions. These variables showed a clear and significant relationship with increased healthcare expenses. On the other hand, BMI and smoking status did not show a strong connection to cost in this analysis. However, it's important to note that even though they weren’t significant in this model, they may still affect overall health outcomes in the long term.

Healthcare providers should consider focusing their intervention efforts on older adults and patients with multiple chronic conditions, as these groups are associated with higher healthcare costs. Additionally, frequent hospital visits can contribute significantly to overall expenses, so encouraging preventive care and offering more outpatient support could help reduce unnecessary hospital use. Although smoking did not emerge as a major cost driver in this analysis, it remains an important health risk and should continue to be a focus in long-term health promotion and disease prevention strategies.