10 Questions About the Project

1. Why did you pick the Gradient Boosting model instead of something simpler?
Gradient Boosting gave me the best results. It handled complex patterns in the data better than simpler models like Linear Regression, especially when multiple features like age and BMI interact with each other.

2. Did you check if the model works equally well for different groups of people, like older adults or people from different regions?
Yes, I looked at model performance across variables like age and region. Smoking and BMI had the strongest influence, but results stayed consistent across regions.

3. What did you do about data points that looked really different or extreme, like very high charges?
I visualized and checked for outliers, especially very high insurance charges. While I didn't remove them, I used models like Gradient Boosting that are more robust to those extreme values.

4. Which mattered more for predicting cost—being a smoker or having a high BMI?
Smoking was the biggest cost driver by far. BMI also mattered, especially in combination with smoking, but smoking alone caused a much bigger jump in charges.

5. Why couldn't your model figure out who was a smoker? Isn't that kind of important?
Good question. The model couldn't predict smoking status because the other features didn't provide enough clear clues. It's likely that smoking behavior is independent of things like age or BMI in this dataset.

6. How do you make sure the model isn't unfair to certain groups of people?
I avoided using sensitive features like race. I also checked model outputs across groups and recommended fairness checks if this were deployed in the real world.

7. Would adding things like medical history make the model better, and is it okay to use that kind of private info?
Yes, medical history would definitely improve accuracy. But using that data must be done carefully—with consent, privacy protections, and clear ethical guidelines.

8. If an insurance company wanted to use this model, how would they actually do that?
They could integrate it into their pricing system by feeding in customer info and using the model to suggest risk-based pricing. They'd also need tools to explain the predictions to both staff and customers.

9. How often would they need to update or retrain it to keep it accurate?
It's best to retrain the model regularly—maybe every 6–12 months—to account for changes in health trends, policy, or population behavior.

10. What would change if you had more data, like step counts from a fitness tracker?
That would likely improve predictions a lot. Behavior data like physical activity could help personalize risk scores more precisely and encourage healthier choices too.