Q1.

**1. What makes a model "linear"? "Linear" in what?** A model is called “linear” when it is linear in its **coefficients**, not necessarily in its variables. That means the model’s equation forms a straight-line relationship between the inputs (features) and the output using a linear combination of the coefficients.

**2. How do you interpret the coefficient for a dummy/one-hot-encoded variable?** The coefficient shows the difference in the outcome between that category and the reference category (the one left out). The trick is that this interpretation only works when the model has an **intercept**. If there’s no intercept, the meaning of the coefficient changes.

**3. Can linear regression be used for classification? Explain why, or why not.** It can be used, but it’s not ideal. For binary classification (like 0 and 1), it might work okay, but it doesn’t predict probabilities well and can give values outside of 0 and 1. Logistic regression is better for classification because it gives outputs between 0 and 1.

**4. What are signs that your linear model is overfitting?**

* High accuracy on training data but low accuracy on test data
* A very complex model with too many variables
* Large differences between training and validation errors
* The model is capturing noise, not patterns

**5. Clearly explain multicollinearity using the two-stage least squares technique.** Multicollinearity happens when your predictors are too closely related to each other, making it hard to tell which one is actually affecting the outcome. In two-stage least squares:

* First, we predict the problematic variable using other variables (instruments).
* Then, we use that predicted value in the main regression.  
   This helps reduce the collinearity issue.

**6. How can you incorporate nonlinear relationships into your analysis?** You can:

* Add polynomial terms Use log or square root transformations
* Include interaction terms
* Or switch to a model that handles nonlinearity, like decision trees or neural networks

**7. What is the interpretation of the intercept, a slope coefficient, and a dummy variable coefficient?**

* **Intercept**: The expected outcome when all predictors are zero
* **Slope coefficient**: How much the outcome changes with a one-unit change in that variable
* **Dummy variable coefficient**: The difference in the outcome between that category and the reference category (if an intercept is present)

Q9.

What worked: Scaling the features and using logistic regression gave solid results. The binary variables didn’t need extra encoding, which simplified things.

What didn’t: At first, I tried one-hot encoding everything, which was unnecessary and caused issues.

Was the code intentional? Yes, I followed a plan rather than randomly testing.

If I did this again: I’d check variable distributions and correlations early to avoid wasted steps and get to a better model faster.

Q8.

In this example, I performed a simple linear regression using the Boston Housing Dataset to predict the median home value based on the average number of rooms per dwelling. I evaluated the model's performance using MSE and R², and visualized the regression line.​ This foundational approach can be extended by incorporating more features, performing feature engineering, and experimenting with more complex models to improve predictive performance.