1. **What does R-squared represent in a regression model?**

    - R-squared measures how well the independent variables explain the variance in the dependent variable. A higher R-squared (closer to 1) indicates a better fit, while a low R-squared suggests the model doesn't explain much of the variance.
2. **What are the assumptions of linear regression?**

    - **Linearity**: The relationship between independent and dependent variables is linear.
    - **Independence**: Residuals (errors) are independent of each other.
    - **Homoscedasticity**: Constant variance of residuals across all levels of the independent variable.
    - **Normality**: Residuals follow a normal distribution.
    - **No Multicollinearity**: Independent variables are not highly correlated with each other.
3. **Difference between R-squared and Adjusted R-squared:**

    - **R-squared** increases as more predictors are added, even if they don't improve the model.
    - **Adjusted R-squared** corrects for the number of predictors, penalizing models with irrelevant variables. It can decrease if unnecessary predictors are added.
4. **Why use Mean Squared Error (MSE)?**

    - MSE penalizes larger errors by squaring them, providing a robust measure of model accuracy. It helps prioritize models that make smaller errors consistently.
5. **What does an Adjusted R-squared value of 0.85 indicate?**

    - 85% of the variance in the dependent variable is explained by the model, accounting for the number of predictors. It shows a strong model fit.
6. **How to check for normality of residuals in linear regression?**

    - **Visual Methods**:
        - Q-Q plot (quantile-quantile plot)
        - Histogram of residuals
    - **Statistical Tests**:
        - Shapiro-Wilk test
        - Anderson-Darling test
        - Kolmogorov-Smirnov test
7. **What is multicollinearity, and how does it impact regression?**

    - Multicollinearity occurs when independent variables are highly correlated, leading to unreliable and unstable coefficient estimates. This inflates the variance and makes it difficult to identify the true effect of each predictor.

8. **What is Mean Absolute Error (MAE)?**

    - MAE measures the average absolute difference between predicted and actual values. It is easy to interpret and not sensitive to outliers, unlike MSE.
9. **Benefits of using an ML pipeline:**

    - **Automation**: Streamlines data preprocessing, model training, and evaluation.
    - **Reproducibility**: Ensures consistent results.
    - **Efficiency**: Reduces redundancy and data leakage.
    - **Modularity**: Allows changes in individual steps without affecting the entire process.
10. **Why is RMSE more interpretable than MSE?**

    - RMSE is in the same unit as the dependent variable, making it easier to understand and interpret compared to MSE, which squares the errors.
11. **What is pickling in Python, and how is it useful in ML?**

    - Pickling serializes (saves) Python objects, including ML models, to a file. It allows models to be saved and reloaded for later use, preserving their state.
12. **What does a high R-squared value mean?**

    - A high R-squared indicates that the model explains most of the variance in the dependent variable, suggesting a good fit. However, it doesn't guarantee predictive accuracy.
13. **What happens if linear regression assumptions are violated?**

    - Violations can lead to:
        - **Biased estimates** (non-linearity, correlated errors)
        - **Inefficient estimates** (heteroscedasticity)
        - **Unstable coefficients** (multicollinearity)
        - **Poor predictive performance**
14. **How to address multicollinearity in regression?**

    - **Remove correlated predictors**
    - **Use Principal Component Analysis (PCA)**
    - **Apply Ridge or Lasso regression** (regularization techniques)
15. **Why use pipelines in machine learning?**

    - Pipelines prevent data leakage, automate workflows, and ensure consistent transformations across training and testing datasets.
16. **How is Adjusted R-squared calculated?**
    Adjusted R2=1−((1−R2)(n−1)n−p−1)Adjusted\ R^2 = 1 - \left(\frac{(1 - R^2)(n-1)}{n - p - 1} \right)
    Where nn = sample size and pp = number of predictors.

17. **Why is MSE sensitive to outliers?**

    - MSE squares the errors, giving greater weight to larger deviations. This makes it sensitive to outliers, which can disproportionately affect the model.
18. **Role of homoscedasticity in linear regression:**
    - Homoscedasticity ensures that the model's residuals have constant variance. If violated (heteroscedasticity), the model may produce inefficient estimates and unreliable predictions.
19. **What is Root Mean Squared Error (RMSE)?**
    - RMSE is the square root of MSE. It measures the standard deviation of prediction errors and provides a more interpretable error metric.
20. **Why is pickling considered risky?**
    - Pickled files can execute arbitrary code, posing a security risk if the file is tampered with or untrusted.
21. **Alternatives to pickling for saving ML models:**
    - **Joblib** (optimized for large objects)
    - **HDF5** (saves in hierarchical format)
    - **ONNX** (open standard for ML models)
    - **JSON** (lightweight and interpretable)
22. **What is heteroscedasticity, and why is it a problem?**
    - Heteroscedasticity occurs when residual variance changes across levels of the independent variable. This violates regression assumptions, leading to inefficient estimates and unreliable hypothesis tests.
23. **How does adding irrelevant predictors affect R-squared and Adjusted R-squared?**
    - R-squared **increases** as more predictors are added, regardless of relevance.
    - Adjusted R-squared **decreases** if the added predictors don't improve the model, penalizing overfitting.