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

How well the independent variable explained the variation in the dependent variable

Rsquare = explained variation(SSR) / total variation(TSS)

**2. What are the assumptions of linear regression\***

1) x and y should linear in relationship

2) The observation (rows) should are independent of other

3) Homoscedasticity : The variance of the error are constant

4) Error should be normally distributed

**3. What is the difference between R-squared and Adjusted R-squared\***

Both of them are used to check the goodness of the model in different manner

In R-squared how well the independent(predictor) variable explained the variation of the dependent variable

Adj R- squared used to check the unnecessary feature are added or not in the model

**4. Why do we use Mean Squared Error (MSE)\***

Mean Squared Error is used as performance metrics

MSE is also used to get the best line is where the MSE is least

MSE is used as CF to get optimal M,C

**5. What does an Adjusted R-squared value of 0.85 indicate\***

The Adjusted R-squared indicates have a strong goodness in the model and the 85% variation on depended variable by accounting the number of predictors included

**6. How do we check for normality of residuals in linear regression\***

we can use the Visualize method for this

error = y\_actual – y\_predict

we can use an distribution plot for determine the Error normality

sns.distplot(error)

**7. What is multicollinearity, and how does it impact regression\***

Multicollinaertiy occur when two or more predictor variable in a regression model are highly colrelated

Making it difficulty to access their individual contribution to the outcome variables

**8 What is Mean Absolute Error (MAE)\***

Mean Absolute is a performance metrics and it measures the average absolute difference between y\_actual and y\_predict

**9 What are the benefits of using an ML**

1. **Scalability** - Handles large data and parallelizes tasks.
2. **Reproducibility** - Ensures consistent and trackable workflows.
3. **Modularity** - Enables easy updates and reusability of components.
4. **Automation** - Reduces manual effort and human errors.
5. **Maintainability** - Simplifies debugging and updating processes.
6. **Efficiency** - Optimizes resources and streamlines workflows.
7. **Collaboration** - Facilitates teamwork with standardized processes.
8. **Integration** - Simplifies deployment with seamless tool support.
9. **Experimentation** - Encourages fast, comparative testing.
10. **Monitoring** - Tracks performance and enables continuous improvement.

**10 Why is RMSE considered more interpretable than MSE\***

* **Same Units**: RMSE is in the same unit as the target variable, unlike MSE's squared units.
* **Intuitive**: RMSE reflects average error magnitude, making it easier to understand.
* **Communication**: RMSE is simpler to explain to non-technical audiences.
* **Comparability**: RMSE aligns better with data spread (e.g., standard deviation).

**11 What is pickling in Python, and how is it useful in ML\***

Pickling is the process of serializing (converting) Python objects into a byte stream using the pickle module. This byte stream can later be deserialized (unpickled) to reconstruct the original object.

**12 What does a high R-squared value mean\***

A high R-squared value indicates that a large proportion of the variance in the dependent variable is explained by the independent variables in the model.

Good fit one close to the observed data and may be there is chance of overfitting

**13 What happens if linear regression assumptions are violated\***

**Effects of Violating Linear Regression Assumptions:**

1. **Linearity**: Poor fit and biased predictions; use transformations or non-linear models.
2. **Independence**: Correlated errors lead to invalid tests; address with GLS or lag variables.
3. **Homoscedasticity**: Changing error variance affects reliability; use transformations or robust errors.
4. **Normality**: Non-normal errors affect confidence; use transformations or non-parametric methods.
5. **Multicollinearity**: Correlated predictors cause unstable coefficients; remove or regularize variables.
6. **Omitted Variables**: Missing predictors bias results; include all relevant variables.

**14. How can we address multicollinearity in regression\***

We can visualized by Clustermap can be we can not quantitative

So we can use the Variance\_inflation\_factor from statsmodels.stats.outlier\_influence import Variance\_inflation\_factor

First we need to empty dataframe for the variance\_inflation\_factor

The first columns as features of all the datas columns then measure the Vif

If the VIF> 10 then remove the featiure one by one

RFE >>Recursive\_Feature\_Elimination

Rfe is a feature selection method used in machine learning to identify most important predictor for a given model

**15 Why do we use pipelines in machine learning\***

**Reasons for Using Pipelines in Machine Learning:**

1. **Streamlined Workflows**: Automates repetitive tasks like preprocessing, feature selection, and model training.
2. **Reproducibility**: Ensures consistent and trackable processes for reliable results.
3. **Error Reduction**: Minimizes manual errors by chaining steps systematically.
4. **Modularity**: Allows easy replacement or adjustment of individual components.
5. **Efficiency**: Optimizes resource usage and reduces redundant computations.
6. **Deployment Readiness**: Simplifies end-to-end integration and model deployment.
7. **Experimentation**: Facilitates quick testing of different configurations and models.
8. **Collaboration**: Enhances teamwork by standardizing workflows and sharing components.

**16 How is Adjusted R-squared calculated\***

**Adjusted r2 = 1-(1-r2)(N-1)/N-P-1**

**N: no of datapoints**

**P:No: of independent variable**

**17 Why is MSE sensitive to outliers\***

If there is outlier then error is higher the squared value

**18 What is the role of homoscedasticity in linear regression\***

**Role of Homoscedasticity in Linear Regression**

1. **Accurate Inference**: Ensures reliable standard errors, p-values, and confidence intervals.
2. **Efficiency**: Enables optimal coefficient estimation with OLS.
3. **Model Fit**: Indicates consistent residual variance across predictors.
4. **Predictive Reliability**: Supports consistent error distribution for accurate predictions.

Violations lead to biased inferences; address with transformations or robust methods.

Also known as constane variance This assumption means the variance of the error are constant that means there is no pattern

**19 What is Root Mean Squared Error (RMSE)\***

RMSE is a metric used to evaluate the performance of regression models by measuring the average magnitude of errors between predicted and actual values.

Same unit it is also differentiable,less sensitive to outlier

**20 Why is pickling considered risky\***

**Risks of Pickling**

1. **Security**: Vulnerable to malicious code during deserialization.
2. **Compatibility**: Issues across Python versions or environments.
3. **Size**: Can result in large files.
4. **Debugging**: Binary format is not human-readable.
5. **Dependency**: Fails if object structure changes.

Use trusted sources or safer alternatives like JSON or joblib.

**21 What alternatives exist to pickling for saving ML models\***

**Alternatives to Pickling for ML Models**

1. **joblib**: Efficient for scikit-learn models.
2. **ONNX**: Cross-framework interoperability.
3. **HDF5**: Stores models and weights (e.g., TensorFlow/Keras).
4. **JSON**: Saves architecture; weights stored separately.
5. **PMML**: Standard for model sharing.
6. **Framework-Specific**: E.g., PyTorch (torch.save()), TensorFlow (SavedModel).

**22 What is heteroscedasticity, and why is it a problem\***

Heteroscedasticity occurs when the variance of the residuals (errors) in a regression model is not constant across all levels of the independent variables. In other words, the spread of errors increases or decreases as the value of the predictor variable changes.

**23 How does adding irrelevant predictors affect R-squared and Adjusted R-squared?**

R-squared:

Increases: Adding irrelevant predictors will always increase or keep R-squared the same, even if the predictors don’t improve the model. This is because R-squared measures the proportion of variance explained by the model, and more predictors typically capture more variance.

Adjusted R-squared:

Decreases or Remains Constant: Unlike R-squared, Adjusted R-squared accounts for the number of predictors in the model. Adding irrelevant predictors penalizes the model for unnecessary complexity, so Adjusted R-squared can decrease if the added predictors do not improve the model's explanatory power.