Multiple Linear Regression

**Interview Questions:**

1. **What is Normalization & Standardization and how is it helpful?**

**Normalization** - Normalization is the process of scaling the data to a fixed range, usually [0, 1]. And Standardization transforms the data to have a mean of 0 and standard deviation of

**Formula:**

Xnorm = X - Xmin / Xmax - Xmin

**Use :**  
Useful when you want to **preserve the shape** of the data distribution and the scale of different features varies greatly.

**Standardization -** Standardization transforms the data to have a mean of 0 and standard deviation of 1.

**Formula:**

Xstd= X−μ​ / σ

**Use**:

Commonly used in algorithms that assume normally distributed data, such as Linear Regression, SVM, Logistic Regression, etc.

1. **What techniques can be used to address multicollinearity in multiple linear regression?**

**Multicollinearity** occurs when two or more independent variables in a regression model are highly correlated. This can make it difficult to determine the effect of each predictor on the target variable, and it often inflates the variance of coefficient estimates, making the model unstable.

**Techniques to Address Multicollinearity:**

**1. Remove Highly Correlated Predictors**

* Drop one of the correlated variables.
* Example: If height and arm\_span are highly correlated, you may remove one.

**2. Combine Correlated Features**

* Create a new feature using transformation.
* Example: Combine height and arm\_span into a new average variable.

**3. Principal Component Analysis (PCA)**

* Reduces dimensionality and transforms features into uncorrelated components.
* Good for dealing with multicollinearity while preserving most information.

**4. Use Regularization Techniques**

* Apply **Ridge Regression (L2 Regularization)** – penalizes large coefficients and helps manage multicollinearity.
* **Lasso Regression (L1)** – can also help by shrinking some coefficients to zero, effectively doing feature selection.

**5. Drop One of the Dummy Variables (in case of categorical features)**

* Known as the **dummy variable trap**.
* Always exclude one category when using one-hot encoding.

**6. Increase Sample Size**

* Sometimes, adding more data can reduce the instability caused by multicollinearity.

**7. Domain Knowledge**

* Use your understanding of the domain to decide which variable to keep or drop.