Multiple Linear Regression

**Interview Questions:**

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

* **Normalization** (also known as **min-max scaling**):  
  This technique rescales the data to a fixed range, typically [0, 1] or [-1, 1]. The formula is:

Xnorm=X−XminXmax−XminX\_{\text{norm}} = \frac{X - X\_{\text{min}}}{X\_{\text{max}} - X\_{\text{min}}}Xnorm​=Xmax​−Xmin​X−Xmin​​

**When is it helpful?**

* + Used when the data does not follow a Gaussian distribution.
  + Particularly effective for algorithms that rely on distance calculations, like K-Nearest Neighbors (KNN) and Gradient Descent-based methods.
* **Standardization** (also known as **Z-score normalization**):  
  This technique transforms the data to have a mean of 0 and a standard deviation of 1. The formula is:

Xstd=X−μσX\_{\text{std}} = \frac{X - \mu}{\sigma}Xstd​=σX−μ​

where μ\muμ is the mean and σ\sigmaσ is the standard deviation.  
**When is it helpful?**

* + Preferred for algorithms assuming normal distribution of data, such as Linear Regression, Logistic Regression, and Support Vector Machines (SVM).
  + Better when outliers are present, as it doesn't squash the data into a fixed range.

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

Multicollinearity occurs when independent variables are highly correlated, which can distort the estimates of regression coefficients.

**Techniques to Address Multicollinearity:**

* **Remove Highly Correlated Predictors:**  
  Identify and drop one of the variables that are highly correlated (using correlation matrix or VIF).
* **Principal Component Analysis (PCA):**  
  Transform the correlated variables into a set of uncorrelated components.
* **Regularization Techniques (Ridge & Lasso Regression):**
  + **Ridge Regression (L2 regularization):** Penalizes large coefficients but doesn’t eliminate them.
  + **Lasso Regression (L1 regularization):** Can shrink some coefficients to zero, effectively performing variable selection.
* **Variance Inflation Factor (VIF):**  
  Calculate VIF for each predictor. A VIF above 5-10 indicates high multicollinearity.
* **Combining Variables:**  
  Create composite variables by combining correlated predictors (e.g., through averaging or weighted sums).
* **Collecting More Data:**  
  Sometimes, increasing the sample size can help reduce the impact of multicollinearity.