**MLR - Interview Questions and Answers**

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

**Normalization** and **Standardization** are two techniques used in data preprocessing, especially in machine learning, to adjust the scale of data, making it more suitable for modelling.

**1. Normalization**

Normalization scales data to a specific range between 0 and 1.

**When to Use Normalization**

* When you **do not assume a normal distribution** of data.
* When the model, like **k-nearest neighbour’s (KNN)** or **neural networks**, is based on distance metrics (Euclidean, Manhattan, etc.).
* When you want your data to fall into a particular range (usually [0, 1]).

**Example**

For example, if you have the following data points for a feature: [10, 20, 30], and you want to normalize it, you will scale them to fall between 0 and 1.

**2. Standardization**

Standardization, also known as **z-score normalization**, scales the data to have a **mean of 0** and a **standard deviation of 1**.

**When to Use Standardization**

* When you **assume a normal distribution** of data.
* When algorithms like **support vector machines (SVM)** or **linear regression** expect features to have a normal distribution.
* When features have **different scales** but you want them to have the same importance in model training.

**These Techniques Helpful For:**

* **Improves performance of machine learning algorithms**: Models such as gradient descent-based algorithms, distance-based algorithms (like KNN), and neural networks perform better when features are on a similar scale.
* **Speeds up convergence in optimization problems**: In algorithms like gradient descent, when features are on a similar scale, the convergence rate increases.
* **Avoids bias**: Large-scale features will not dominate the learning process, allowing the model to treat each feature equally.

Both methods play a key role in preparing data for machine learning by ensuring that algorithms function optimally, even when data values differ significantly in scale.

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

Multicollinearity in multiple linear regression refers to a situation where two or more independent variables or predictors are highly correlated, meaning they provide overlapping information. So, this can lead to unstable estimates of regression coefficients, making it difficult to assess the importance of individual predictors.

**Techniques to Address Multicollinearity:**

1. Remove Highly Correlated Predictors

2. Principal Component Analysis (PCA)

3. Ridge Regression (L2 Regularization)

4. Lasso Regression (L1 Regularization)

5. Partial Least Squares (PLS) Regression

6. Centering the Variables (Mean Centering)

7. Increase Sample Size

**The choice of technique depends on the severity of multicollinearity and the goals of the analysis. If interpretation is important, removing predictors or using PCA may be the best option. If prediction is the focus, Ridge or Lasso are powerful regularization techniques.**