MULTICOLLINEARITY

DEFINITION:

- Multicollinearity is the occurrence of high intercorrelations among two or more independent variables in a multiple regression model.
- Multicollinearity can lead to skewed or misleading results when a researcher or analyst attempts to determine how well each independent variable can be used most effectively to predict or understand the dependent variable in a statistical model.

- Multicollinearity is a statistical concept where several independent variables in a model are correlated.
- Two variables are considered perfectly collinear if their correlation coefficient is +/- 1.0.
- Multicollinearity among independent variables will result in less reliable statistical inferences.
- When you're analysing an investment, it is better to use different types of indicators rather than multiple indicators of the same type to avoid multicollinearity.
- Multicollinearity can lead to less reliable results because the results you're comparing are generally the same.

VIF (Variance Inflation Factor) is a hallmark of the life of multicollinearity, and statsmodel presents a characteristic to calculate the VIF for each experimental variable and worth of greater than 10 is that the rule of thumb for the possible lifestyles of high multicollinearity. The excellent guiding principle for VIF price is as follows, VIF = 1 manner no correlation exists, VIF > 1, but < 5 then correlation exists.

$$VF_i = 1/1 - R_i^2$$

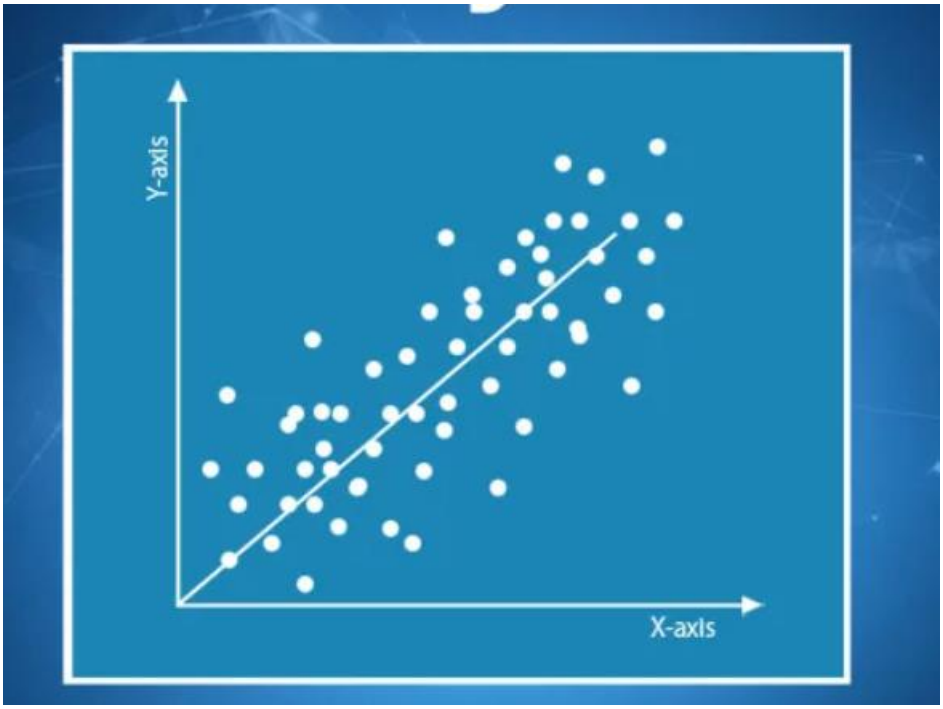
where R^2 is the coefficient of determination of variable.

TYPES OF MULTICOLLINEARITY:

Data-based multicollinearity: A poorly designed experiment or data collection process, such as using observational data, generally results in data-based multicollinearity, where data is correlated due to the nature of the way it was collected. Some or all of the variables are correlated.

Structural multicollinearity:

- Structural multicollinearity occurs when you use data to create new features. For instance, if you collected data and then used it to perform other calculations and ran a regression on the results, the outcomes will be correlated because they are derived from each other.
- This is the type of multicollinearity seen in investment analysis because the same data is used to create different indicators.



STEPS TO DETECT MULTICOLLINEARITY:

- Scatter plot and Correlation Matrix
- Correlation Heatmap
- VIF- Variation Inflation Matrix

HANDLING MULTICOLLINEARITY:

- Check for multicollinearity: Before taking any action, it's important to identify multicollinearity in your dataset. You can use methods such as correlation matrices, Variance Inflation Factor (VIF), or Eigenvalues to detect multicollinearity among predictors.
- Remove one of the correlated variables: If you have identified pairs of variables with high correlation, consider removing one of them from the

model. Choose the variable that is less theoretically relevant or less important to the research question.

- Feature selection: Use feature selection techniques to choose a subset of features that are most relevant to the target variable. Techniques such as forward selection, backward elimination, or stepwise regression can help in this regard.
- Principal Component Analysis (PCA): PCA is a dimensionality reduction technique that can be used to reduce the number of correlated predictors into a smaller set of uncorrelated components. These components can then be used in the regression model instead of the original predictors.
- Ridge Regression or Lasso Regression: Regularization techniques like Ridge Regression and Lasso Regression can help in reducing the impact of multicollinearity by adding a penalty term to the regression coefficients. Ridge Regression penalizes the sum of squared coefficients, while Lasso Regression penalizes the sum of absolute coefficients. These techniques can help in stabilizing coefficient estimates and reducing multicollinearity effects.
- Collect more data: Sometimes multicollinearity can arise due to a limited sample size. Collecting more data may help in reducing the correlation between predictors.
- Transform variables: Transforming variables using mathematical functions like log, square root, or inverse can sometimes reduce multicollinearity by changing the scale or distribution of the variables.

STEPS TO AVOID MULTICOLLINEARITY:

- Set VIF value and remove variables above the value.
- Use Regularization techniques like Ridge, Lasso and Elastic Net.
- Using heatmap/Correlation Matrix to detect the highly collinear variables and remove them manually
- Using Feature Engineering, Combine the correlated variables.