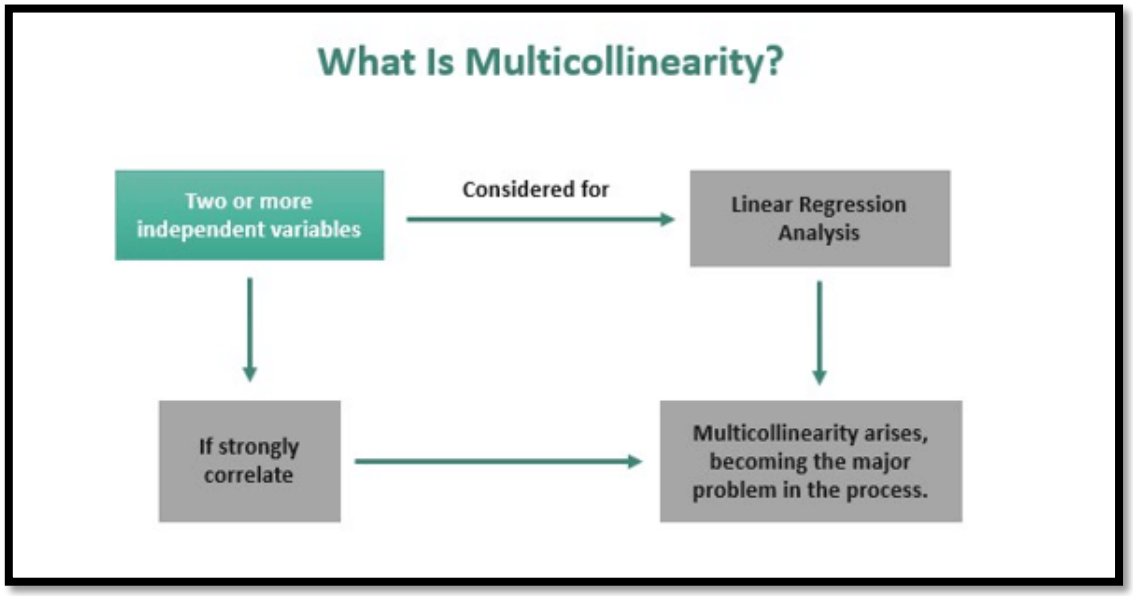
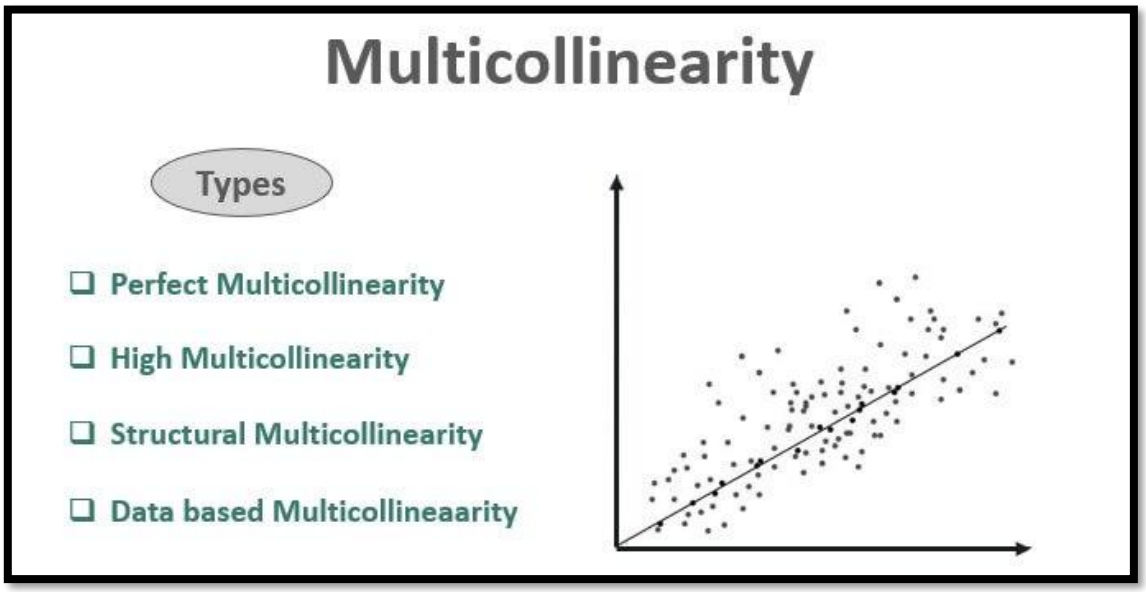
MULTICOLLINEARITY

Multicollinearity Definition

Multicollinearity refers to the statistical phenomenon where two or more independent variables are strongly correlated. It marks the almost perfect or exact relationship between the predictors. This strong correlation between the exploratory variables is one of the major problems in linear regression analysis.



Types of Multicollinearities



Multicollinearity exists in four types:

- **High Multicollinearity:** It signifies a high or strong correlation between two or more independent variables, but not a perfect one.
- **Perfect Multicollinearity:** This degree of collinearity indicates an exact linear relationship between two or more independent variables.
- **Data-based Multicollinearity:** The possibility of collinearity, in this case, arises out of the selected dataset.
- **Structural Multicollinearity:** This issue arises when researchers have a poorly designed framework for the regression analysis.

Why Is It a Problem?

- Inflates standard errors of coefficients.
- Makes coefficients unstable and sensitive to small changes in data.
- Reduces the interpretability of the model.
- Can lead to incorrect conclusions about variable importance.

How to Detect It

- **Variance Inflation Factor (VIF):** Values > 10 suggest serious multicollinearity.
- **Correlation Matrix:** Look for pairwise correlations > 0.8 or < -0.8.
- **Condition Number:** Values > 30 indicate multicollinearity.
- **Eigenvalues of Correlation Matrix:** Very small eigenvalues suggest linear dependence.

How to Fix It

- Remove one of the correlated variables.
- Combine variables (e.g., average scores).
- Use Principal Component Analysis (PCA) to transform variables.
- Apply Ridge Regression or Lasso Regression to regularize coefficients.
- Centering or standardizing variables may help in some cases.

Methods to detect and address multicollinearity, along with Python syntax examples to help you apply them:

1. **VIF(Variance Inflation Factor):**
   - VIF measures how much the variance of a regression coefficient is inflated due to multicollinearity among the independent variables.
   - It tells you how strongly a variable is linearly related to the other predictors in the model.
2. **Correlation Matrix:**
   - A simple but effective way to spot pairwise multicollinearity.
   - Look for correlation coefficients close to ±1.
   - You can also visualize it with a heatmap
3. **Condition Number:**
   - This checks for near-linear dependencies in the design matrix.
   - A condition number > 30 suggests multicollinearity.
4. **Eigenvalues of the Correlation Matrix**
   - Small eigenvalues indicate multicollinearity.
   - If any eigenvalue is close to zero, it means some variables are linearly dependent.
5. **Ridge Regression (to handle multicollinearity)**
   - Instead of removing variables, you can regularize them.
   - Ridge regression shrinks coefficients to reduce the impact of multicollinearity.
6. **Principal Component Analysis (PCA)**
   - Transforms correlated variables into uncorrelated components.
   - Use PCA components instead of raw features in your regression model.

GitHub links for this:

https://github.com/DhanasekarDomain/Data-Science-Basic/blob/e8f6acde3fc19625d5da7d310f8f07e23d192b88/Biveriate.ipynb