WAYS TO HANDLE MULTICOLLINEARITY

Multicollinearity is a condition in dataset which occurs when two or more independent variables of the dataset are correlated to each other. This affects on determining how individual independent variables impacts dependent variables.

To examine whether the dataset has multicollinearity or not, we can

- **Examine Correlation Matrices:** Correlation coefficient (r) is calculated for the variable. The closer r get to +1 or -1, the stronger is the linear correlation between variables.
- **Calculate VIF:** Variance Inflation Factor is a measure used in regression analysis to detect and quantify multicollinearity. **VIF=1/(1-** R_i^2**)**

  If
  VIF = 1, no correlation between independent variables
  VIF > 1, indicates presence of multicollinearity
  VIF > 5, high multicollinearity
- **Using Condition Index:** It is a tool to detect the presence of multicollinearity in regression model. Condition index is calculated from the correlation matrix of independent variables by **CI =** $\sqrt{\frac{\lambda_{max}}{\lambda_i}}$ where $\boldsymbol{\lambda_{max}}$ is largest eigen value and $\boldsymbol{\lambda_i}$ is individual eigen value

  If
  CI < 10, low multicollinearity
  CI = 10 to 30, moderate multicollinearity
  CI > 30, high multicollinearity

Ways to handle multicollinearity are:

1. **Drop Redundant variables:** removing highly correlated variable from the model can reduce multicollinearity.
2. **Transforming variables:**
   - **Combining Variables:** If two or more independent variables are highly correlated, they share same underlying concept. Combining these highly correlated variables reduces multicollinearity
   - **Standardization and Centering:** Transforming the variables to have mean = 0 and standard deviation = 1 using Z score to standardize independent variable. Or use centering method where mean of variable is calculated and mean value is subtracted from each datapoints to center the data around zero.
3. **Principal Component Analysis:** PCA transforms original variables into a set of uncorrelated variables. PCA is linear combinations of original variable, these PCA variables are chosen in such a way that they capture the maximum amount of variance in data.
   - The original variable is multiplied by a coefficient and sum up the results. These coefficients defines the direction of new axes

- First Principal Component captures the largest amount of variance in data
- Second Principal Component captures the most remaining variance and orthogonal to the first principal component.