Ways to Handle Multicollinearity

Multicollinearity:

- The situation in which more than 2 independent variables have a linear relationship (i.e. highly correlated with each other).

Disadvantage:

- Reduces the accuracy of the model prediction

Ways to Handle Multicollinearity:

1. **VIF (Variance Inflation Factor):**

   <u>Coding:</u>

```
from statsmodels.stats.outliers_influence import variance_inflation_factor

def vif(dataset):
    vif_data = pd.DataFrame()
    vif_data["feature"] = dataset.columns
    vif_data["VIF"] = [variance_inflation_factor(dataset.values, i)
                       for i in range(dataset.shape[1])]
    return vif_data
```

   Output:

   vif(dataset[['etest_p', 'hsc_p', 'ssc_p']])

S.No	Feature	VIF
0	etest_p	26.95899
1	hsc_p	45.815508
2	ssc_p	47.638794

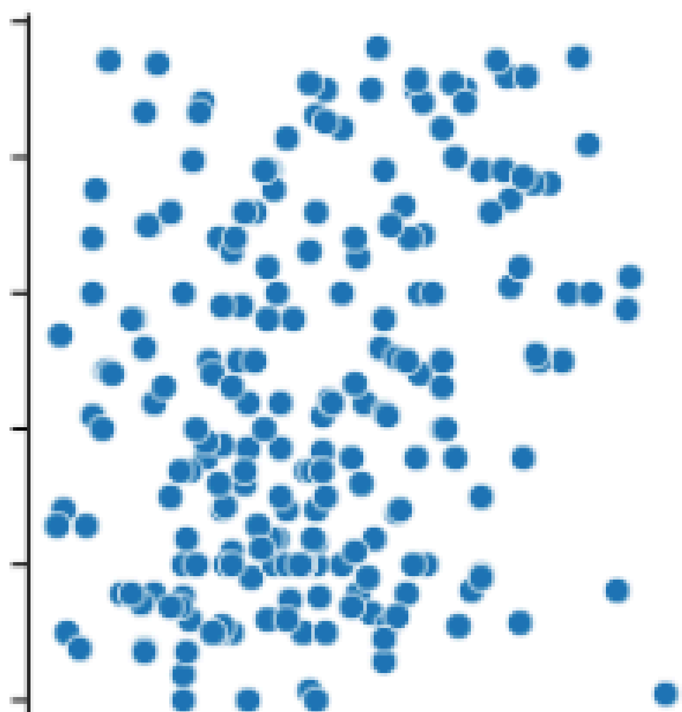
2. **PCA (Principal Component Analysis):** PCA is a dimensionality reduction technique that transforms the original features into a smaller set of uncorrelated components, which can help eliminate multicollinearity.

   Coding:

```
import pandas as pd
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt
import seaborn as sns

scaler = StandardScaler()
X_scaled = scaler.fit_transform(dataset[Quan])
pca = PCA(n_components=2)
reduced_data = pca.fit_transform(dataset[Quan])

pca_df = pd.DataFrame(reduced_data, columns=['mba_p', 'salary'])
plt.figure(figsize=(8, 6))
sns.scatterplot(x='mba_p', y='salary', data=dataset[Quan])
plt.title("PCA: First two principal components")
plt.show()
```



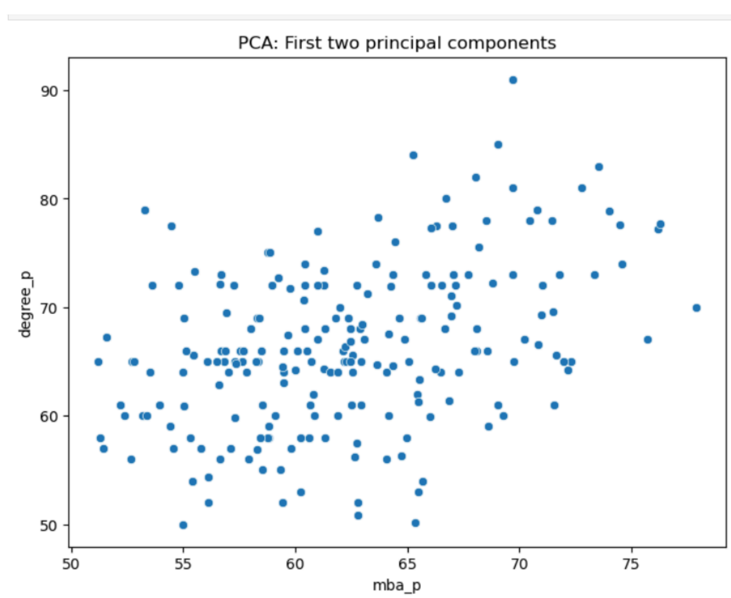


Image 1: Before
Image2: After

3. **Combining more related columns into single columns:** If two or more independent variables are highly correlated and measures similar aspects, we can combine them as a single variable
4. **Dropping one of the highly correlated variables**
5. **Ridge and Lasso Regression:** Adding penalty term to the regression model to shrink coefficient
    a. Ridge Regression (L2 regularization): Penalizes the sum of the squares of the coefficients.
    b. Lasso Regression (L1 regularization): Penalizes the sum of the absolute values of the coefficients and can shrink some coefficients to zero, effectively performing feature selection.

6. **Increase Sample Size:** Sometimes, small set of sample data may cause multicollinearity. Increasing the sample size might potentially reduce the error.