**Introduction:**

The dataset given had 10 thousand rows and 11 columns. The dataset values are continuous in nature and using the Linear Regression model seems best, to predict the Target variable Y.

I have evaluated the Linear regression model’s performance based on different metrics such as Variance Inflation Factor, Homoscedastic nature, actual vs predicted plot etc. Based on this evaluation I opted to implement the Polynomial regression model of degree 3 for the dataset.

**Data Preparation and Feature selection:**

I used the following libraries for my analysis. sci-kit-learn, pandas, NumPy, seaborn and matplotlib. There were no null values in the dataset. I checked the distribution of the target and all the feature variables and found them to be normally distributed with negligent skew and few outliers. I used the correlation matrix metrics to select the features most relevant to predict the outcome and deleted one of the highly correlated variables. I divided the dataset into 75% training and 25% testing data for analysis.

**Model Selection and Performance Evaluation**:

1. **Linear Regression Model:** - I implemented the linear regression model on the dataset and made the following conclusions:
   1. The Actual vs predicted plot is curve in nature and hence linear assumption fails.
   2. The plot is heteroscedastic in nature.
   3. The Variance Inflation Factor is less than 5, hence the data is moderately correlated.
   4. The RMSE value (7.54) is low and the R2 value (0.774). This indicates model’s goodness of fit is not high.
   5. Multivariate normality plot shows the distribution of all the variables put together is not normal in nature.
2. **Polynomial Regression Model:** I implemented the Polynomial Regression model because the earlier Linear Regression model predicted the dataset is not linear in nature and could not explain the dataset properly.
   1. I used a Polynomial regression model of degree 3 and found the actual vs predicted value plot was a straight line.
   2. The dataset displayed Homoscedastic properties.
   3. The Variance Inflation factor was greater than 5 and so the variables were highly correlated.
   4. The model has low RMSE value (4.48) and high R2 value (0.9199) which indicated model’s goodness of fit is very high.
   5. Multivariate normality plot shows the distribution of all the variables put together is normal in nature.

Note – If you open the notebook in Google Colab some cells will be hidden. Required to open.