Multicollinearity

> **Multicollinearity occurs when two or more independent variables are highly correlated with one another in a regression model.**

This means that an independent variable can be predicted from another independent variable in a regression model. For example, height and weight, household income and water consumption, mileage and price of a car, study time and leisure time, etc.

Multicollinearity can be a problem in a regression model because we would not be able to distinguish between the individual effects of the independent variables on the dependent variable. For example, let's assume that in the following linear equation:

$$Y = W0 + W1 * X1 + W2 * X2$$

Coefficient W1 is the increase in Y for a unit increase in X1 while keeping X2 constant. But since X1 and X2 are highly correlated, changes in X1 would also cause changes in X2 and we would not be able to see their individual effect on Y.

Detecting Multicollinearity using VIF

> ” VIF determines the strength of the correlation between the independent variables. It is predicted by taking a variable and regressing it against every other variable. “

or

> VIF score of an independent variable represents how well the variable is explained by other independent variables.

R^2 value is determined to find out how well an independent variable is described by the other independent variables. A high value of R^2 means that the variable is highly correlated with the other variables. This is captured by the VIF which is denoted below:

$$VIF = \frac{1}{1 - R^2}$$

So, the closer the R^2 value to 1, the higher the value of VIF and the higher the multicollinearity with the particular independent variable.

Lets look at the sample data now

In [299]:
```
1 df
```

Out[299]:

	Owner	HouseLocality	Length	Width	HouseArea	Price
0	Mukul	Saharanpur	129	129	16641	13700000
1	Rohan	Meerut	116	145	16820	49500000
2	Mayank	Agra	116	120	13920	43100000
3	Shubham	Saharanpur	135	101	13635	34500000
4	Aakash	Meerut	138	119	16422	40000000
5	Cierra	Saharanpur	103	126	12978	24800000
6	Vega	Delhi	120	102	12240	20200000
7	Alden	Agra	115	126	14490	26200000
8	Cantrell	Saharanpur	120	146	17520	21000000
9	Kierra	Saharanpur	119	131	15589	29400000
10	Gentry	Delhi	122	126	15372	11900000
11	Pierre	Agra	128	107	13696	17500000
12	Cox	Saharanpur	144	147	21168	40000000
13	Thomas	Meerut	124	148	18352	44700000
14	Crane	Delhi	127	124	15748	48000000
15	Miranda	Agra	105	133	13965	21600000
16	Shaffer	Saharanpur	147	104	15288	47800000
17	Bradyn	Delhi	122	133	16226	41000000
18	Kramer	Agra	137	144	19728	15500000
19	Alvaro	Meerut	136	118	16048	49400000
20	Mcgee	Delhi	132	137	18084	41200000

In [300]:
```
# before we do anything, encode the categorical columns
from sklearn.preprocessing import LabelEncoder
df["HouseLocality"] = LabelEncoder().fit_transform(df["HouseLocality"])
df["Owner"] = LabelEncoder().fit_transform(df["Owner"])

# dropping the target variable
df.drop(["Price"], axis=1, inplace = True)
```

In [301]:
```
# Import library for VIF

from statsmodels.stats.outliers_influence import variance_inflation_factor

def CalculateVIF(Data):

    # Calculating VIF
    vif = dict()
    vif["FeatureColumns"] = Data.columns
    vif["VIF"] = [variance_inflation_factor(Data.values, i) for i in range(Data.shape[1])]

    return(pd.DataFrame(vif))
```

In [302]:
```
CalculateVIF(df)
```

Out[302]:

	FeatureColumns	VIF
0	Owner	3.997694
1	HouseLocality	3.119227
2	Length	87.836604
3	Width	131.980725
4	HouseArea	49.524685

We know that, we can get the area of house by just multiplying Length and Width of the house, dropping both the columns and calculating the Variable Inflation Factor (VIF)

In [303]:
```
1 CalculateVIF(df.drop(["Length", "Width"], axis=1))
```

Out[303]:

	FeatureColumns	VIF
0	Owner	3.292545
1	HouseLocality	3.030030
2	HouseArea	1.049887

We were able to drop the variable Length and Height from the dataset because its information was being captured by the 'HouseArea' variable. This has reduced the redundancy in our dataset.

Dropping variables should be an iterative process starting with the variable having the largest VIF value because its trend is highly captured by other variables. If you do this, you will notice that VIF values for other variables would have reduced too, although to a varying extent.