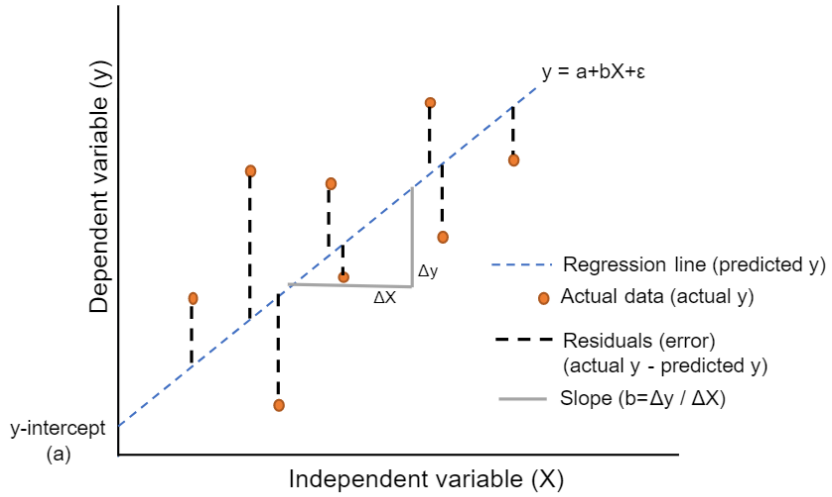
What is Linear Regression ?

1. Linear regression analysis is used to predict the value of a variable based on the value of another variable.

2. The variable you want to predict is called the dependent variable (Predictant).

3. The variable you are using to predict the other variable's value is called the independent variable (Predictor).

$$residuals = actual\ y(y_i) - predicted\ y\ (\hat{y}_i)$$

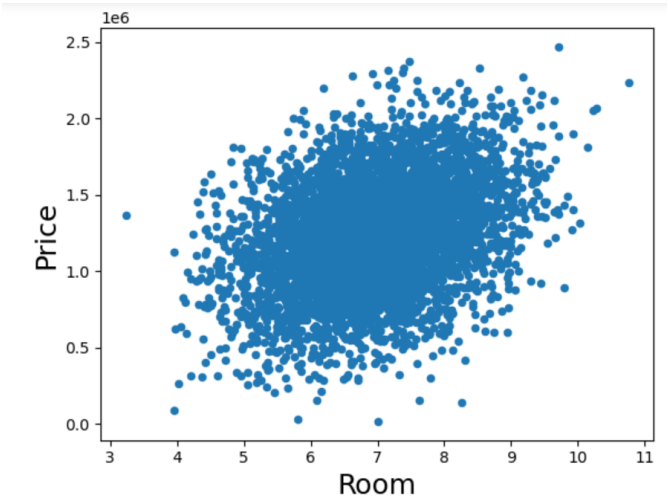


Assumptions for Linear Regression

1. **Linearity:** A linear relationship exists between the dependent variable, Y, and independent variable X.

   An easy way is to plot ***y*** against each explanatory variable ***x_j*** and visually inspect the scatter plot for signs of non-linearity.

```
: df.plot.scatter(x='Avg. Area Number of Rooms', y='Price')
plt.xlabel('Room', fontsize=18)
plt.ylabel('Price', fontsize=18)
plt.show()
df.corr()['Price']
```




```
Avg. Area Income                0.639734
Avg. Area House Age             0.452543
Avg. Area Number of Rooms       0.335664
Avg. Area Number of Bedrooms    0.171071
Area Population                 0.408556
Price                           1.000000
Name: Price, dtype: float64
```

2. No Multicollinearity

Multiple linear regression assumes that none of the predictor variables are highly correlated with each other.

When one or more predictor variables are highly correlated, the regression model suffers from multicollinearity, which causes the coefficient estimates in the model to become unreliable.

Multicollinearity may be checked multiple ways:

1) Correlation matrix – When computing a matrix of Pearson's bivariate correlations among all independent variables, the magnitude of the correlation coefficients should be less than .80.

```
In [10]: df.corr()
```

Out[10]:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price
Avg. Area Income	1.000000	-0.002007	-0.011032	0.019788	-0.016234	0.639734
Avg. Area House Age	-0.002007	1.000000	-0.009428	0.006149	-0.018743	0.452543
Avg. Area Number of Rooms	-0.011032	-0.009428	1.000000	0.462695	0.002040	0.335664
Avg. Area Number of Bedrooms	0.019788	0.006149	0.462695	1.000000	-0.022168	0.171071
Area Population	-0.016234	-0.018743	0.002040	-0.022168	1.000000	0.408556
Price	0.639734	0.452543	0.335664	0.171071	0.408556	1.000000

2) Variance Inflation Factor (VIF) – The VIFs of the linear regression indicate the degree that the variances in the regression estimates are increased due to multicollinearity. VIF values higher than 10 indicate that multicollinearity is a problem.

```
In [22]: vif_data = pd.DataFrame()
         vif_data["feature"] = df.columns
         vif_data["VIF"] = [variance_inflation_factor(df.values, i) for i in range(len(df.columns))]
         print(vif_data)
```

```
                       feature        VIF
0             Avg. Area Income  38.270629
1          Avg. Area House Age  29.097029
2    Avg. Area Number of Rooms  45.335953
3  Avg. Area Number of Bedrooms  14.542817
4             Area Population  14.397643
5                       Price  28.046950
```

```
: vif_data = pd.DataFrame()
  vif_data["feature"] = df.columns
  vif_data["VIF"] = [variance_inflation_factor(df.values, i) for i in range(len(df.columns))]
  print(vif_data)
```

```
                       feature        VIF
0             Avg. Area Income  33.549548
1          Avg. Area House Age  25.439225
2  Avg. Area Number of Bedrooms  10.329674
3             Area Population  13.650474
4                       Price  27.998285
```

3. **Residuals are Normally Distributed**

   If the error terms are non- normally distributed, confidence intervals may become too wide or narrow. Once confidence interval becomes unstable, it leads to difficulty in estimating coefficients based on minimization of least squares. Presence of non – normal distribution suggests that there are a few unusual data points which must be studied closely to make a better model.

4. **Residuals should be homoscedastic**

   Homoskedastic (also spelled "homoscedastic") refers to a condition in which the variance of the residual, or error term, in a regression model is constant. That is, the error term does not vary much as the value of the predictor variable changes. Another way of saying this is that the variance of the data points is roughly the same for all data points.

   Key Take Aways:

   1. Homoscedasticity occurs when the variance of the error term in a regression model is constant.
   2. If the variance of the error term is homoskedastic, the model was well-defined.
   3. If there is too much variance, the model may not be defined well.
   4. Adding additional predictor variables can help explain the performance of the dependent variable.
   5. Oppositely, heteroscedasticity occurs when the variance of the error term is not constant.