Data Analytics with R

simplilearn

DATA AND
ARTIFICIAL INTELLIGENCE

Introduction to Machine Learning

simplilearn

Business Scenario

- Anna works for a pharmaceutical company. The company is testing its new drug and wants to understand the impact of drug dosage on the blood pressure of patients.

  Anna is assigned the task of developing a relationship between the drug dosage and blood pressure and predicting the expected blood pressure based on the dosage.

Approach: To successfully develop the relationship for prediction, Anna must understand the concepts of regression.

Learning Objectives

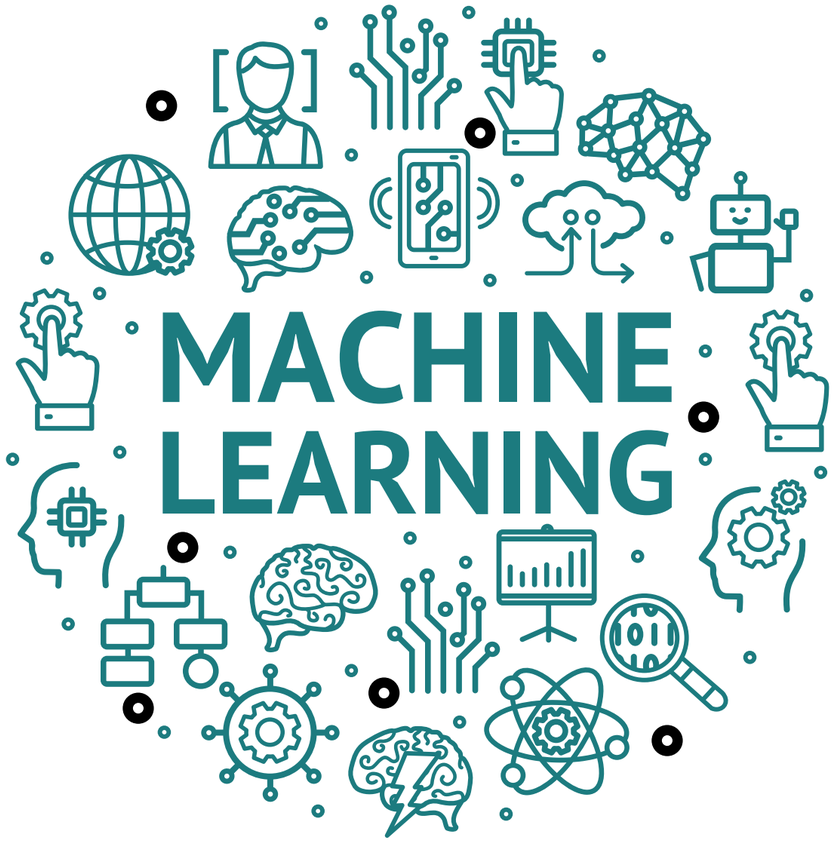
By the end of this lesson, you will be able to:

- Explain machine learning and its types with their applications
- Summarize the flow of supervised learning and unsupervised learning
- Perform regression analysis
- Build a regression model
- Define correlation
- Analyze the assumptions of regression

DATA AND
ARTIFICIAL INTELLIGENCE

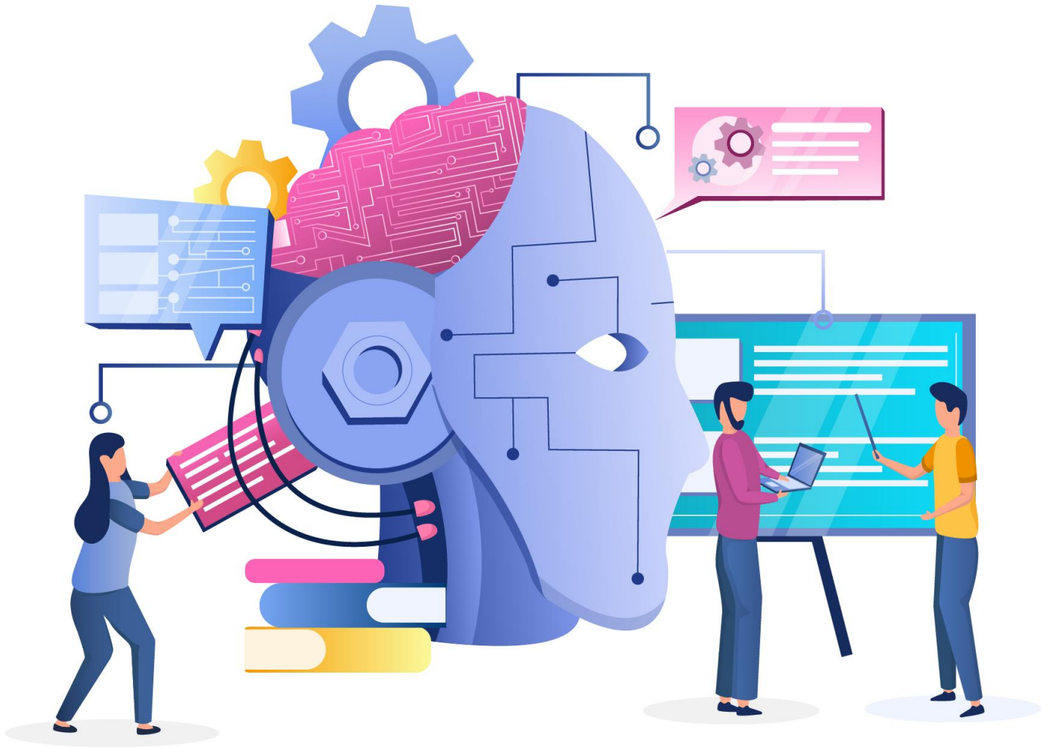
Machine Learning

What Is Machine Learning?



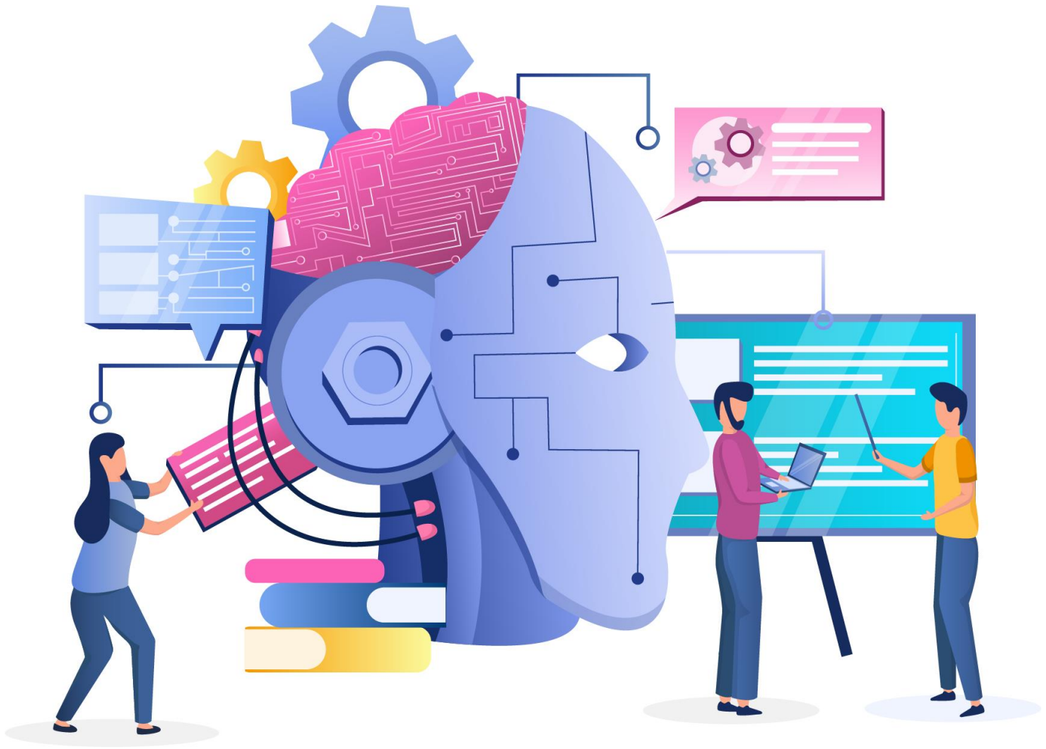
- Machine learning is a subset of artificial intelligence.
- ML allows software applications to become more accurate at predicting outcomes without being explicitly programmed.
- Machine learning algorithms use historical data as input to predict new output values.

Why Machine Learning?



- For many businesses, machine learning has become a crucial competitive differentiation.
- Machine learning is a fundamental aspect of the operations of leading companies, such as Facebook and Uber.
- Machine learning is important because it allows businesses to see trends in customer behavior.

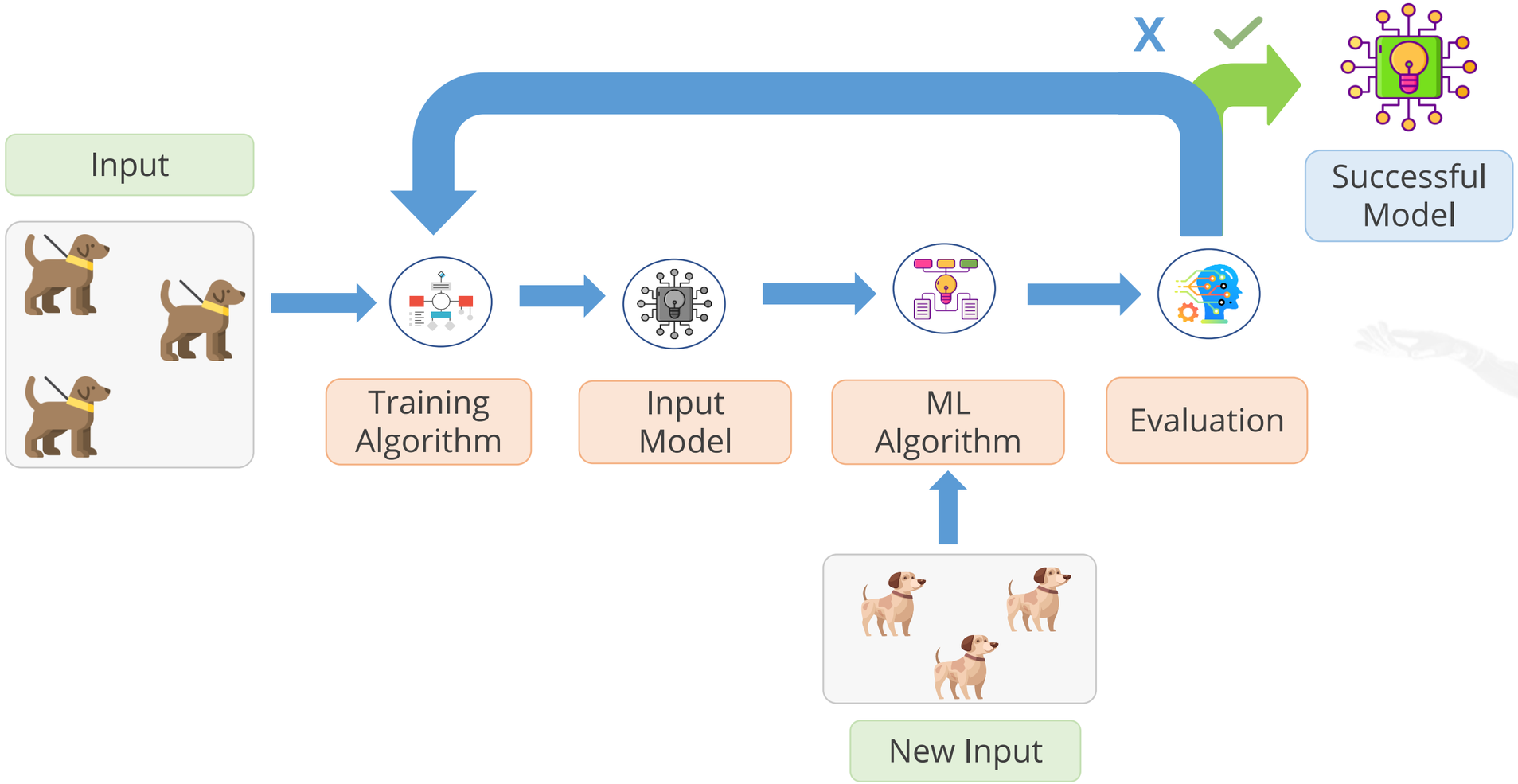
Why Machine Learning?



- It aids in the understanding of business operational patterns and the development of new products for businesses.
- Many businesses utilize machine learning in manufacturing to reduce costs, improve quality control, and streamline supply chains.

Machine Learning: Process Flow

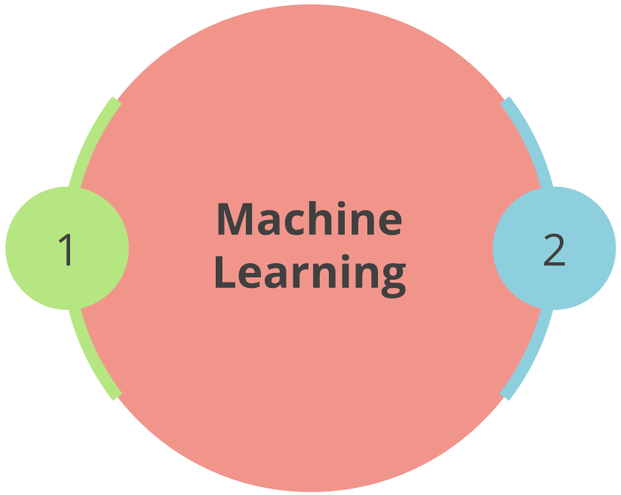
The machine learning process has several stages which are depicted below:



Machine Learning Types

Types of Machine Learning

There are two types of machine learning.



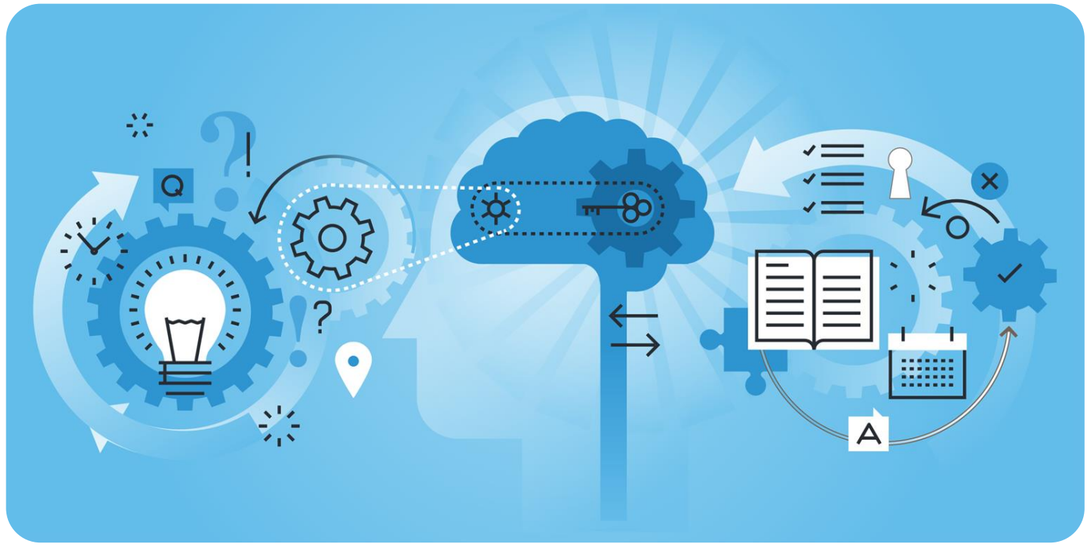
Supervised Learning

Unsupervised Learning

Supervised Learning

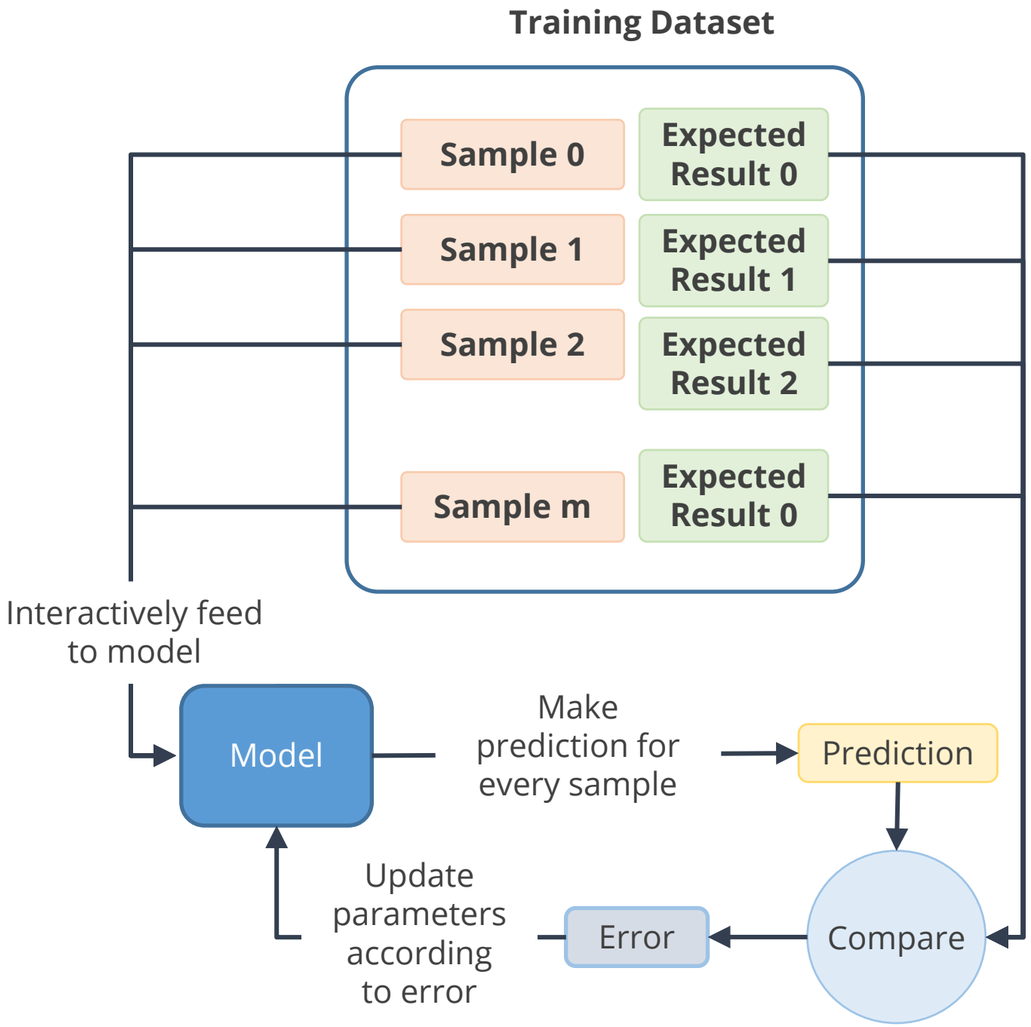
Supervised Learning

Supervised learning is used to train models using labeled training data. It provides the ability to predict the output of future or unseen data.



The goal of a supervised learning algorithm is to discover a mapping function that translates the input variable (x) to the output variable (y).

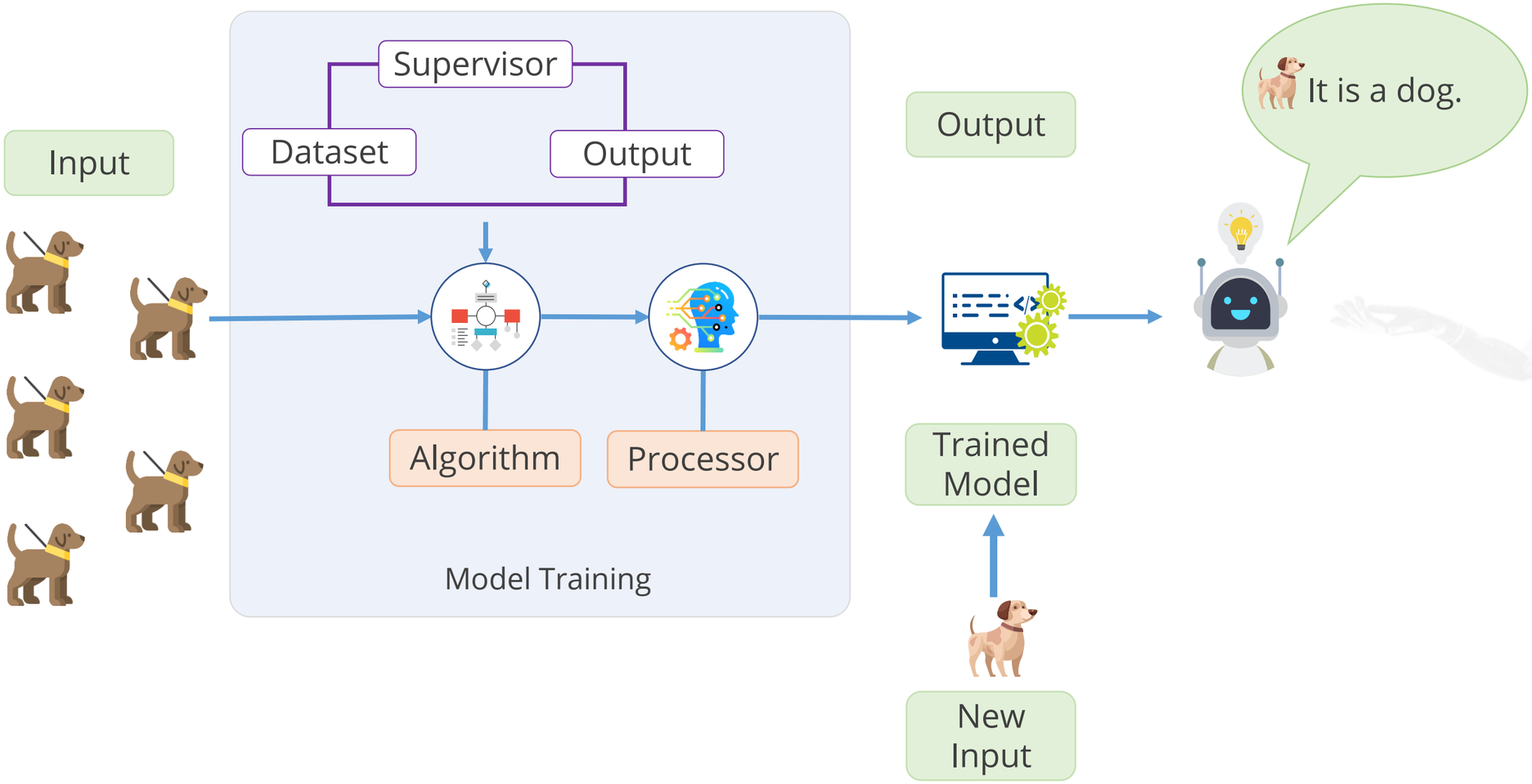
Supervised Learning: Process Flow



- The input-output pairs should make up the required dataset.
- Each pair consists of a data sample for prediction and a label for the expected outcome.
- The human supervisor is responsible for assigning labels to the data in the machine learning process.

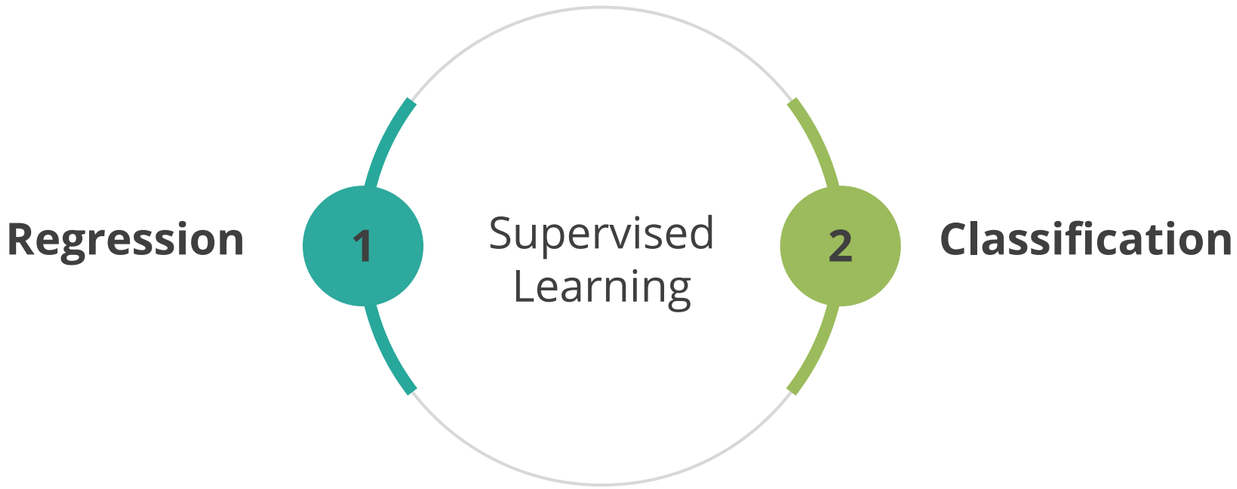
Supervised Learning: Example

The supervised learning process has several stages which are depicted below with an example:



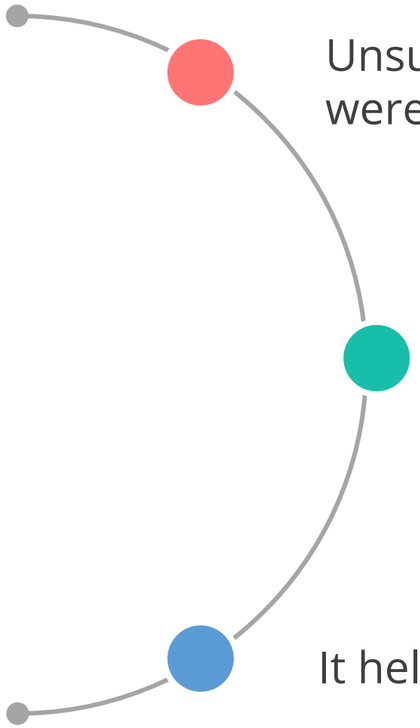
Types of Supervised Learning

An algorithm is selected based on the target variable from the two types, such as:



Unsupervised Learning

Unsupervised Learning



Unsupervised learning allows the model to discover patterns and information that were previously undetected.

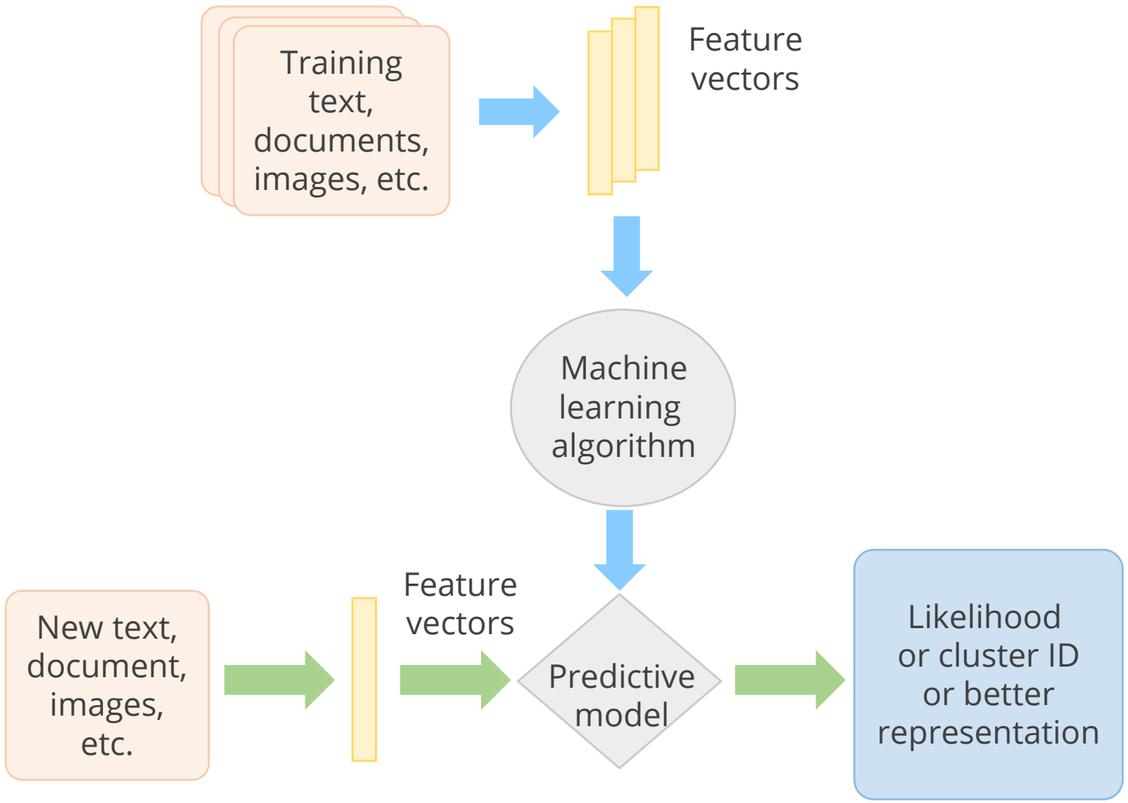
It helps to find all kinds of unknown patterns in data.

It helps to identify the features required for categorization.

Unsupervised Learning: Process Flow

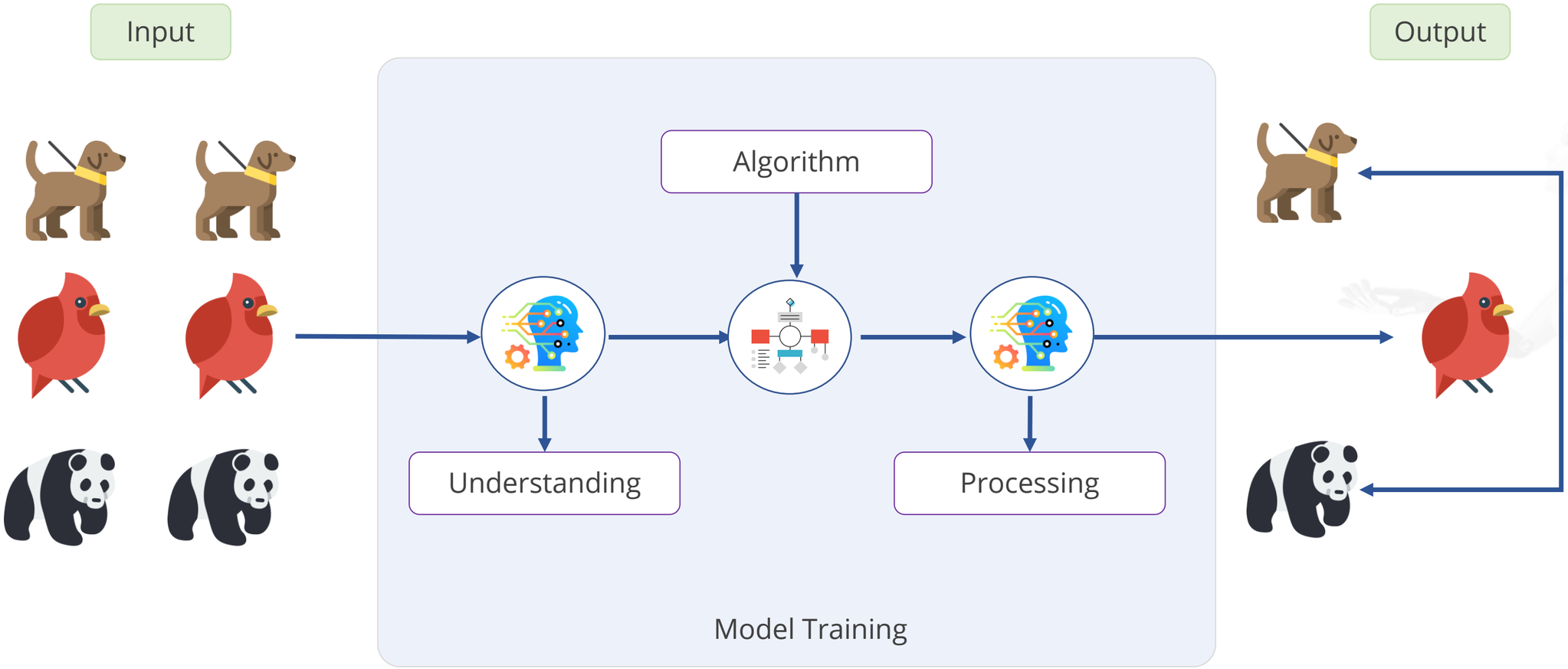
There are no labels on the data. The machine learning algorithm searches for patterns it can detect.

Unsupervised Learning Model



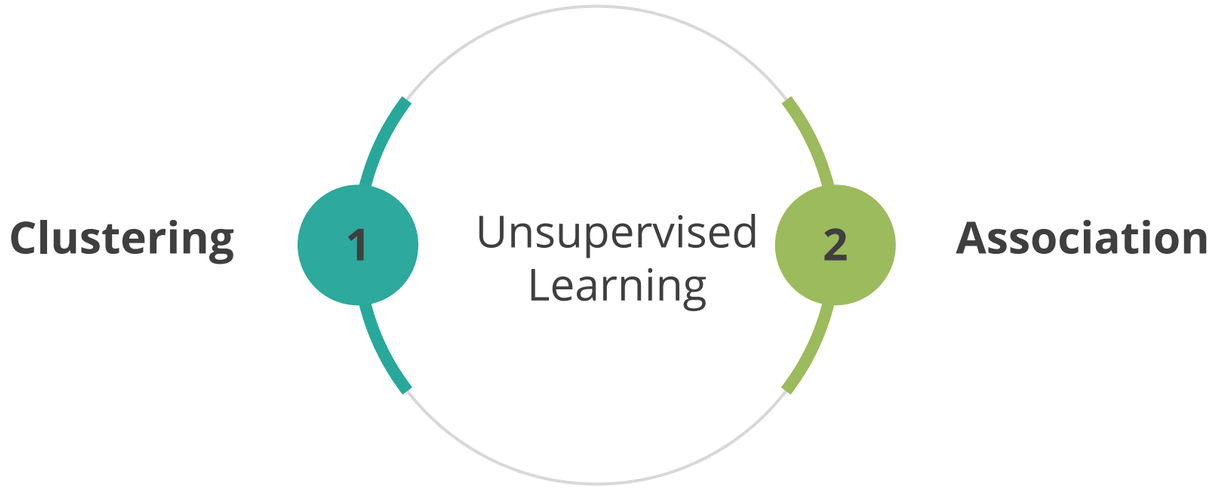
Unsupervised Learning: Example

The unsupervised learning process has several stages which are depicted below with an example:



Types of Unsupervised Learning

There are two types of unsupervised learning.

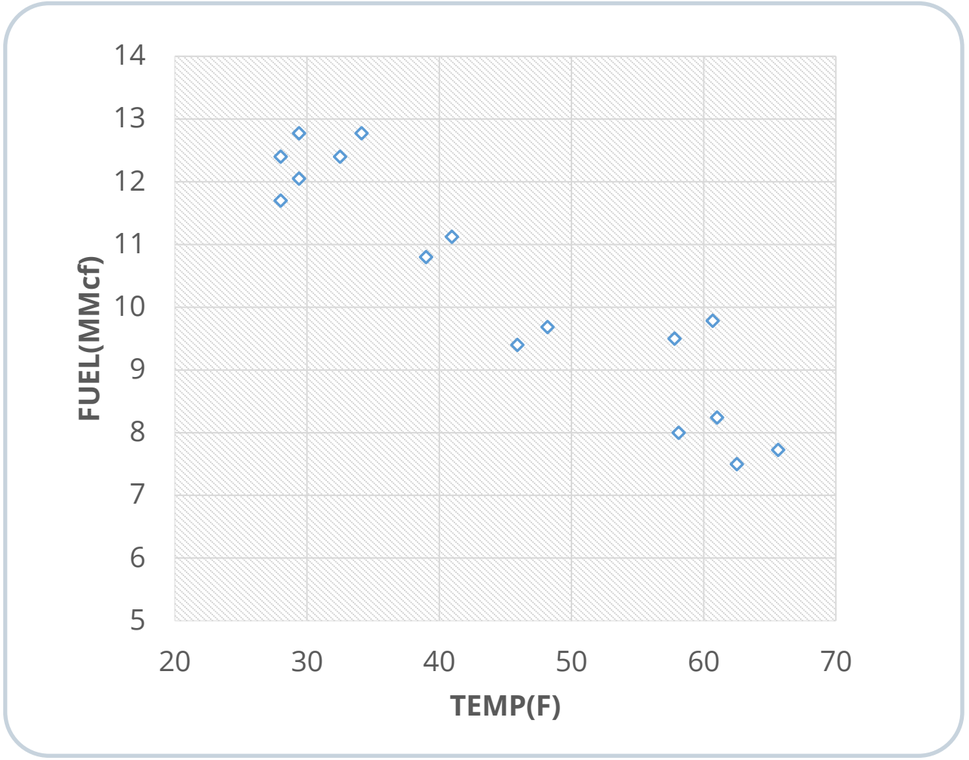


DATA AND
ARTIFICIAL INTELLIGENCE

Regression Analysis

Regression

To predict the weekly amount of natural gas required at homes and businesses in a city based on the week's average hourly temperature, companies leverage the regression analysis technique.

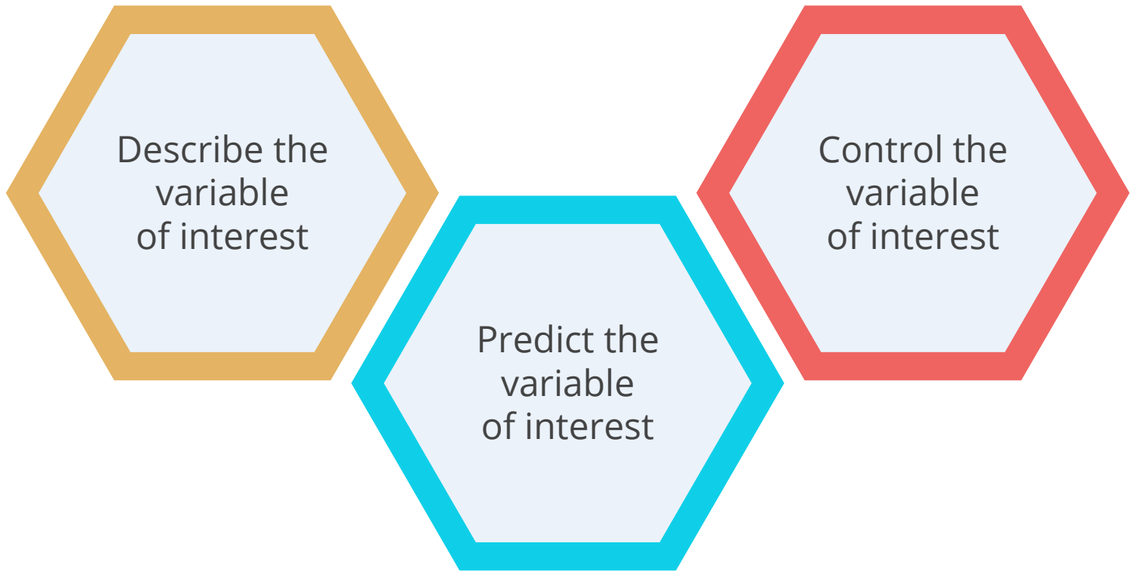


This analysis is used to work on the demand-supply gap.

What Is a Regression Analysis?

It is a statistical technique used to relate a variable of interest (dependent variable) to one or more independent or predictor variables.

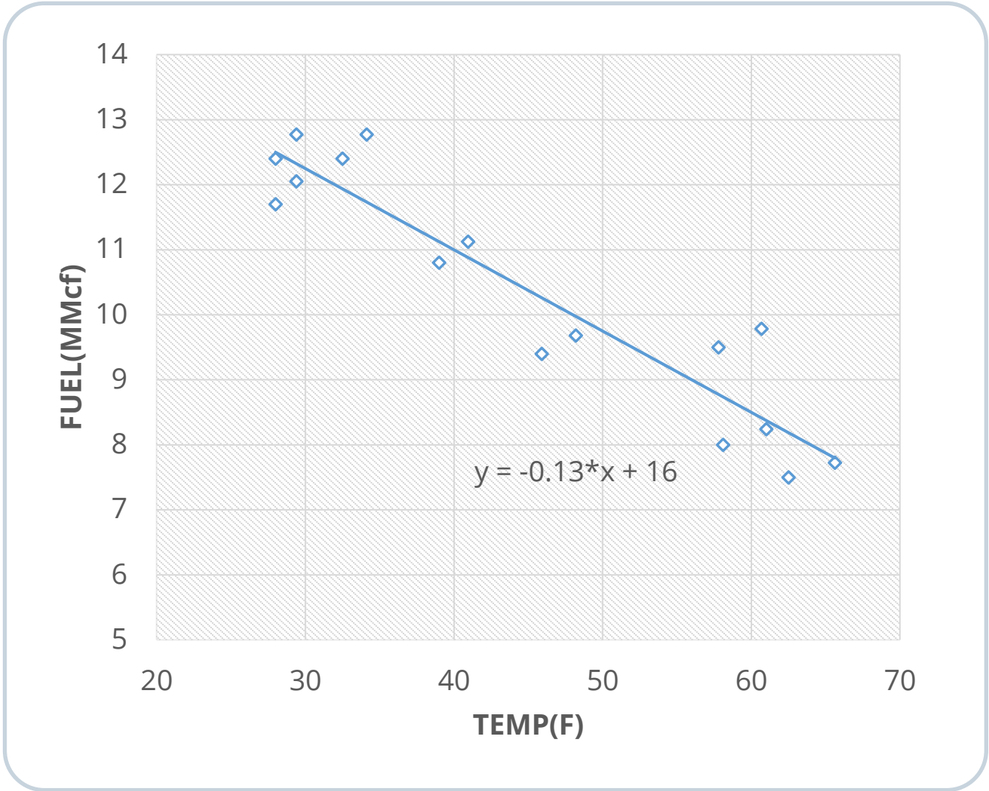
The objective is to build a statistical model to:



Simple Linear Regression

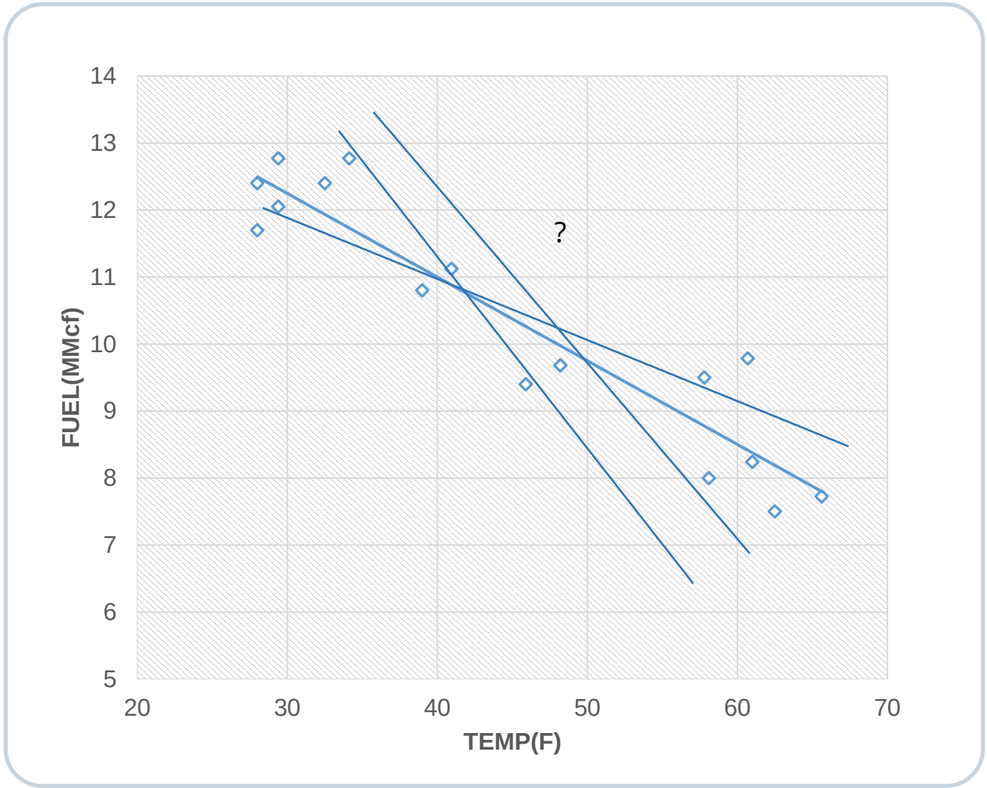
Simple linear regression is a linear regression model with a single predictor variable.

Here, a linear relationship between dependent and independent variables is established.

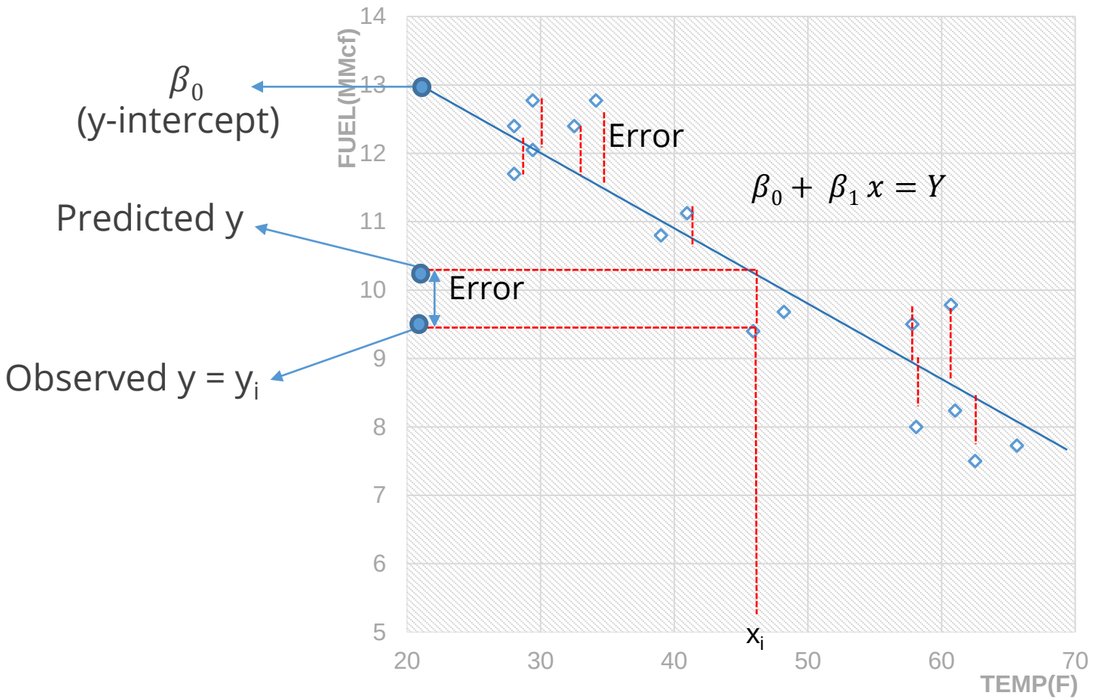


Choosing the Best Line

Out of so many lines passing through the points, let us find out the best line.



Ordinary Least Square Regression

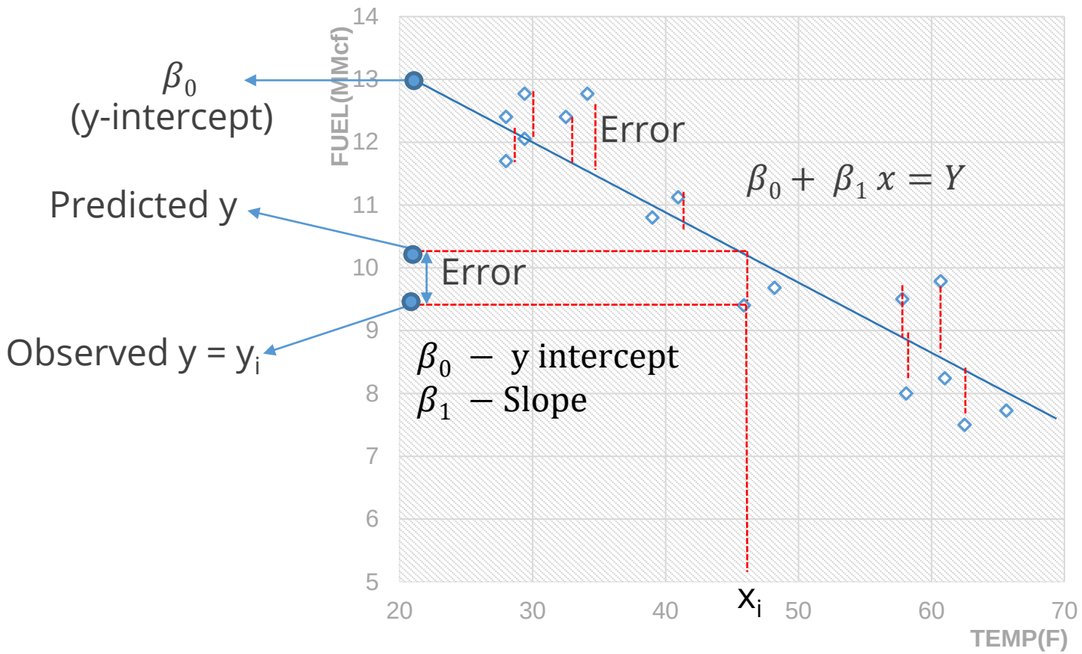


- Assume any line $\beta_0 + \beta_1 x = Y$ passing through the points.
- Here, β_0 is y intercept and β_1 is the slope of the line.

Find the appropriate values of β_0 and β_1 to get to the best−fit line.

Ordinary Least Square Regression

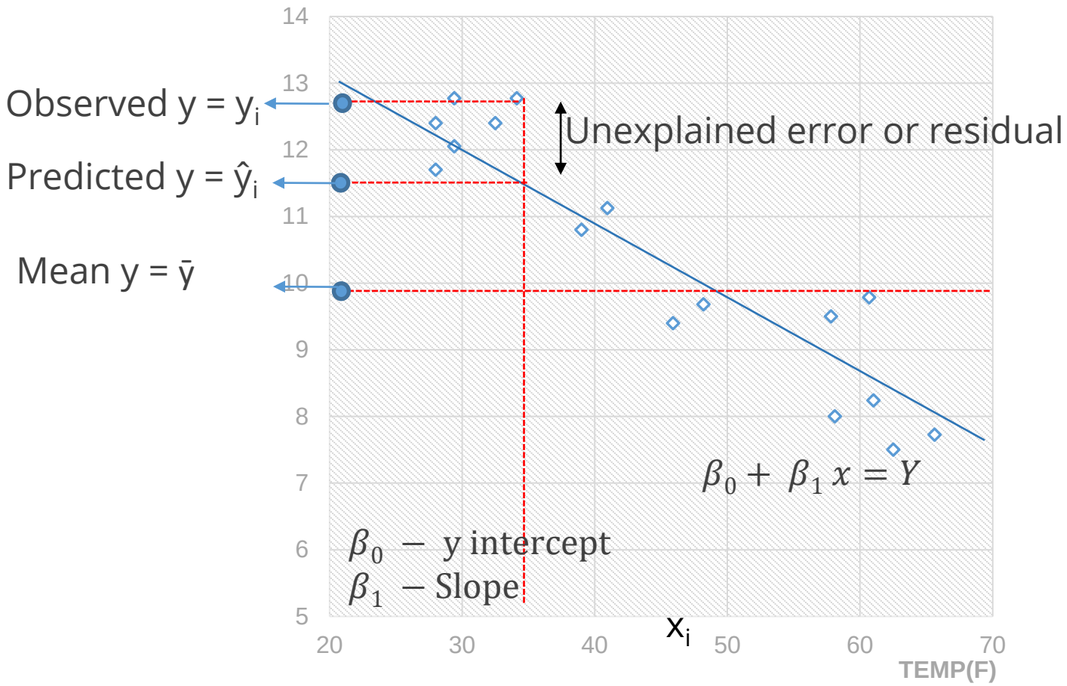
The idea is to find a line for which predicted y and observed y are close for all the points.



- Predicted y = $\beta_0 + \beta_1 xi$, find a line and β_0 and β_1 for which $\sum$(predicted y – observed y)^2 is minimum.
- Find β_0 and β_1 for which $\sum_{i=1}^{n}((\beta_0 + \beta_1 xi) - y_i)^2$ is minimum.

How Good is Regression?

Regression is assessed based on the residual deviations.



$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

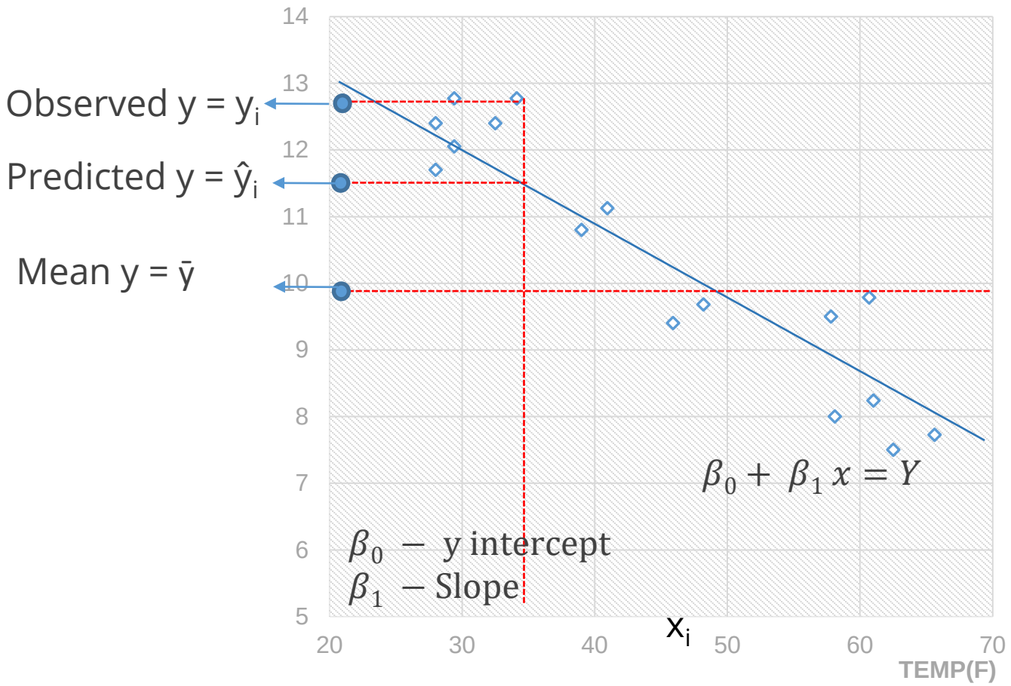
Total deviation = Unexplained deviation + Explained deviation

$$\Sigma(y_i - \bar{y})^2 = \Sigma(y_i - \hat{y}_i)^2 + \Sigma(\hat{y}_i - \bar{y})^2$$

SST (Total sum of squared)	SSE (Sum of squares of error)	SSR (Sum of squares of regression)

How Good Is Regression?

Once the linear relationship is determined, one can analyze the strength of the relationship.



Coefficient of determination = R squared

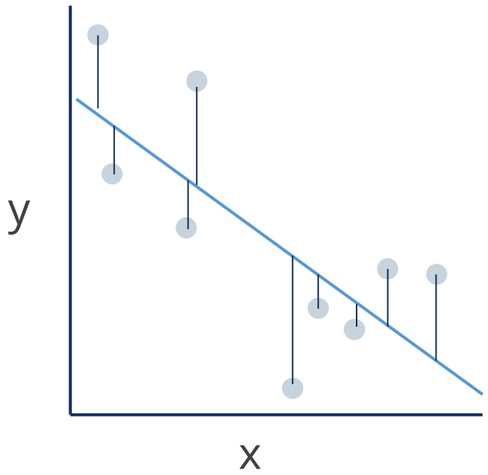
It is the proportion of the variation in y that is explained by the regression.

It is given by $r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

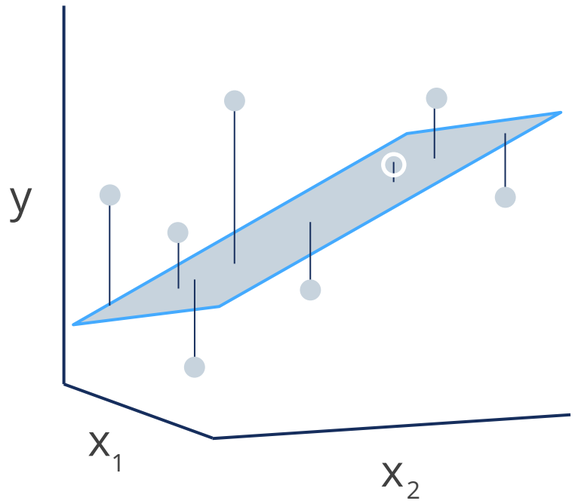
Multiple Linear Regression

Multiple linear regression is regression with multiple predictors.

Simple Linear Regression



Multiple Linear Regression
(Two independent variables (x_1, x_2))



$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Adjusted R^2

Adjusted (or corrected) R^2 is the coefficient of determination corrected for degree of freedom.

- Adjusted R^2 doesn't always increase when new variables are introduced in the regression model.
- Adjusted R^2 increases only when a new variable that is added to the model adds any additional value.
- It is given as: adjusted $R^2 = 1 - \frac{SSE/[n-(k+1)}{SST/(n-1)}$

  n = Sample size

  k = No. of predictors

Evaluation Metrics for Linear Regression

Evaluation metrics measure how good a model performs and how well it defines the relationships.

Other than R^2 and Adjusted R^2, evaluation metrics include:

Metric	Formula
MSE: Mean Squared Error	$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$
MAE: Mean Absolute Error	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert\hat{y}_i - y_i\rvert$
RMSE: Root Mean Squared Error	$MSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$

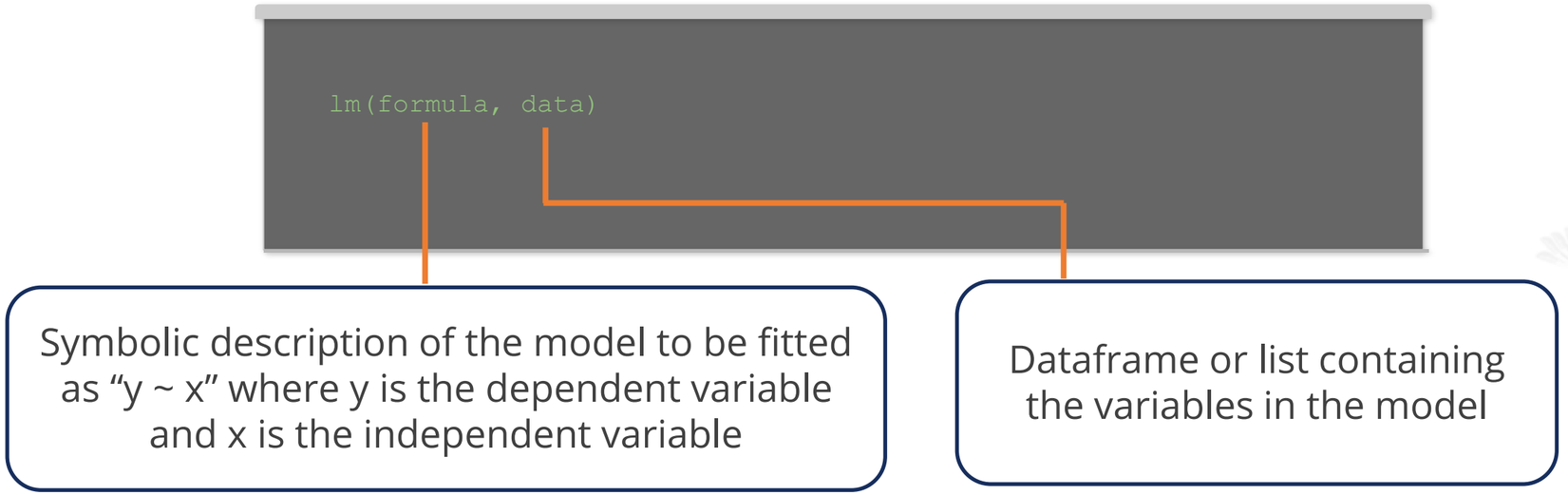
Note

A lower value of these metrics indicates a better model.

Linear Regression Using R

Linear regression can be implemented in R using the lm() function.

Syntax:



To include more predictors in the formula for multiple regression, variables names can be separated by +, for example, y ~ x1 +x2 +x3

House Prices Prediction

Duration: 10 minutes

Problem Scenario: The ***housing.csv*** file contains details of median price values of houses in Boston suburbs. The data describes the various features of the neighborhood.

Sam wants to buy a house in the Boston suburbs. He reaches out to his realtor to get a price estimate.

Build a regression model that will help the realtor give an estimate of the price for a house in a Boston suburb.

Evaluate the model using evaluation metrics and provide conclusions.

Note: Please download the data set and the solution document from the **Course Resources** section and follow the steps given in the document

ASSISTED PRACTICE

DATA AND
ARTIFICIAL INTELLIGENCE

Assumptions of Regression

Assumptions of Linear Regression

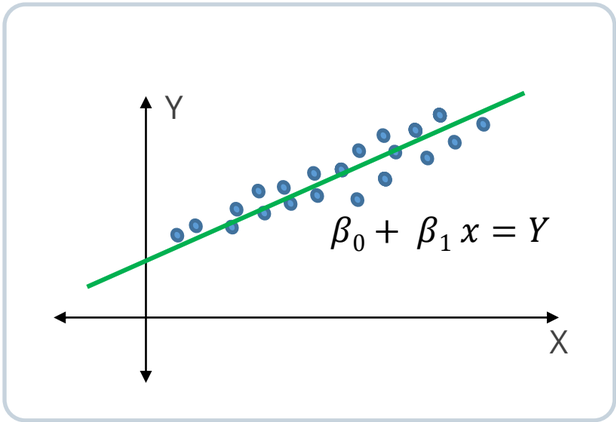
There are four assumptions associated with a linear regression model.

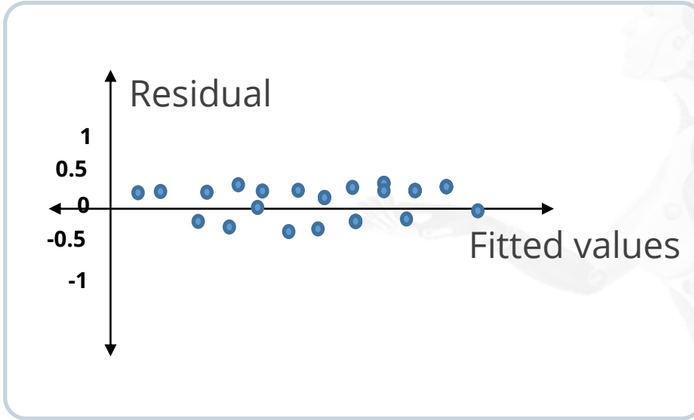
- Linear relationship
- Independence of error
- Normality of error terms
- Equality of variance

Assumption of Linear Regression

- **Linear relationship**
- Independence of error
- Normality of error terms
- Equality of variance

The relationship between the independent and dependent variables should be linear.





Assumption of Linear Regression

- Linear relationship
- **Independence of error**
- Normality of error terms
- Equality of variance

The residuals are independent.

There should be no correlation between consecutive residuals in a time-series data.

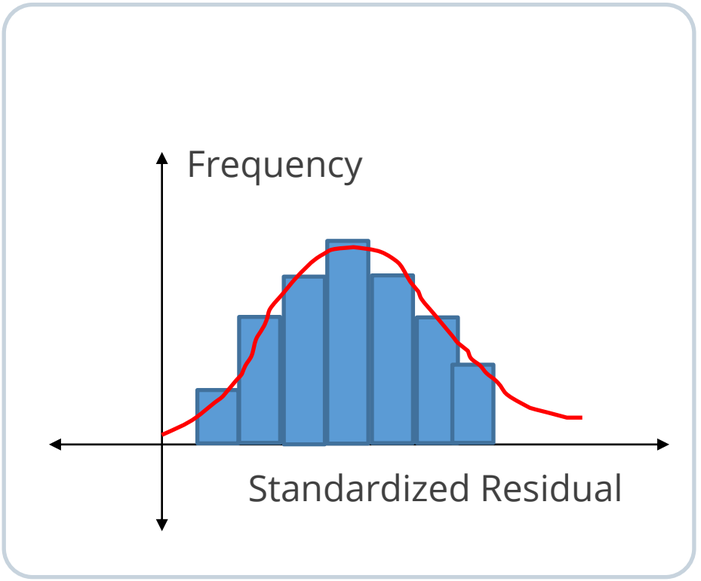
This assumption is important when there is longitudinal or time-series datasets, for instance stock price data.

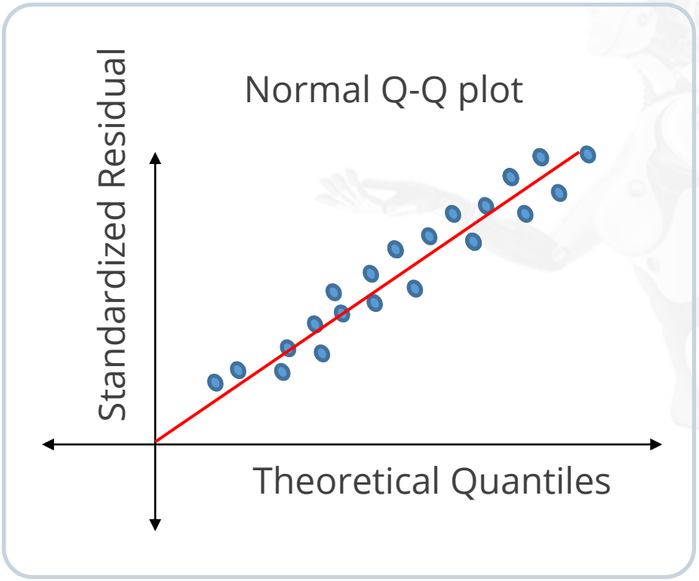
Assumption of Linear Regression

- Linear relationship
- Independence of error
- **Normality of error terms**
- Equality of variance

The error terms (residuals) are normally distributed.

Histogram and Quantile-Quantile plots are used to check this.





Assumption of Linear Regression

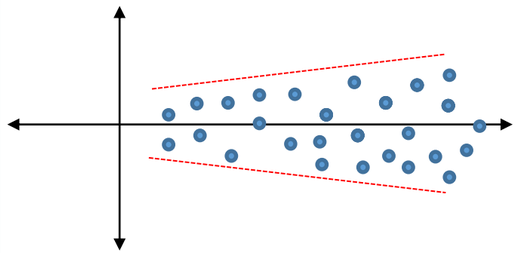
Linear relationship

Independence of error

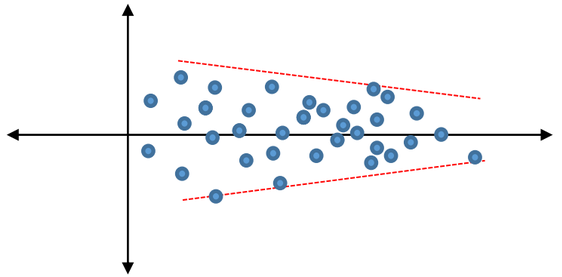
Normality of error terms

Equality of variance

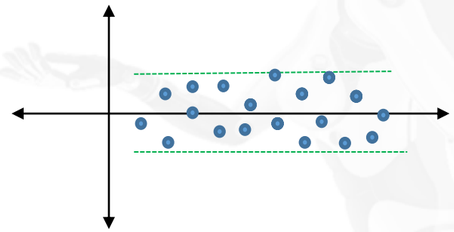
The error terms (residuals) have constant variance at every level of X. It is called homoscedasticity.



Heteroscedasticity
Increasing error variance



Heteroscedasticity
Decreasing error variance

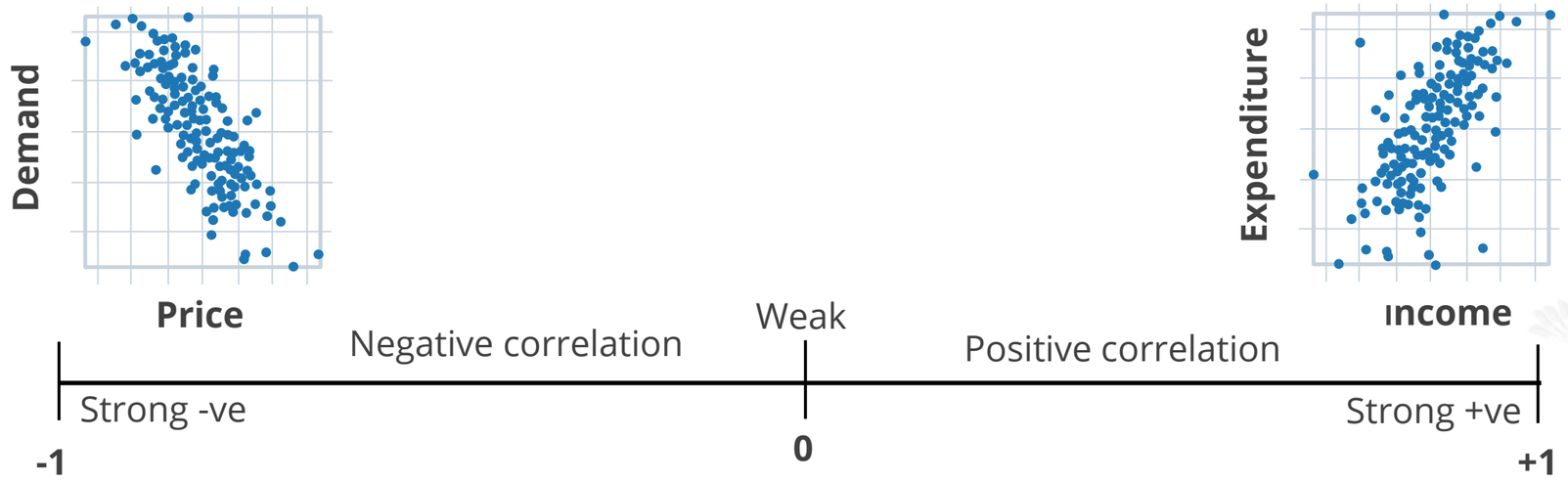


Homoscedasticity
Constant error variance

Correlation

Correlation

Strength and direction of the relationship between two variables is measured by the correlation coefficient.



Correlation coefficients range between -1 and +1.

Zero correlation indicates that the relationship is not linear but has some relation. It can be a strong curvilinear relationship.

Correlation Analysis: Points to Consider

Correlation analysis determines the degree of relationship between two or more variables.

Correlation should not be interpreted as a cause-and-effect relationship.

The existence of causation always implies correlation.

Multicollinearity

Multicollinearity in Regression

If the independent variables in regression model are correlated with one another, it is called multicollinearity.

Multicollinearity reduces the precision of estimated coefficients thereby reducing the statistical power of the model to identify statistically significant independent variables.

Multicollinearity in Regression

This problem is detected using:

01

Correlation coefficients:

initial inspection

02

Scatter diagram between independent variables:

visual inspection

03

Variance inflation factor (VIF):

diagnosis of the issue

Variance Inflation Factor

VIF measures the inflation of variance of an independent variable based on its interaction with other independent variables.

VIF for an independent variable is measured as:

$$VIF_i = \frac{1}{1 - Ri^2}$$

Multicollinearity in Regression

Some of the remedies to remove multicollinearity are:

Evaluate sample scheme and make changes if required

Drop collinear variables

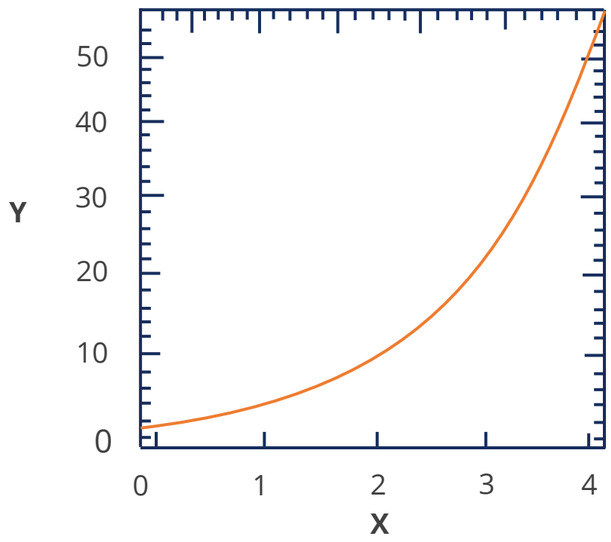
Create new variables using collinear variables and form new combination of X variables which are not correlated

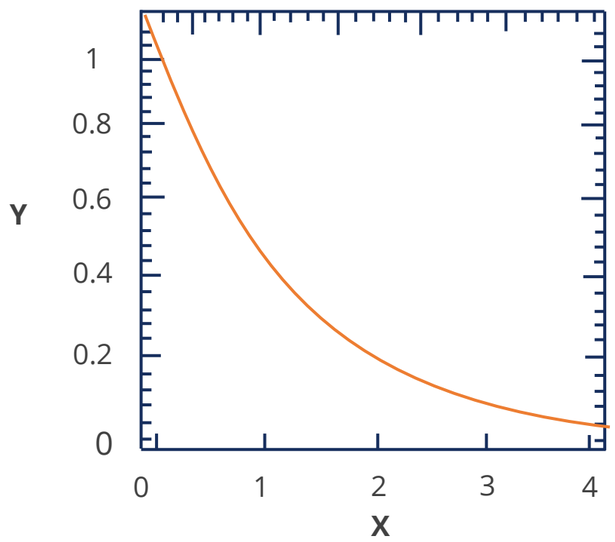
DATA AND
ARTIFICIAL INTELLIGENCE

Nonlinear Regression

Nonlinear Relationship and Transformation

Transformation is used to achieve linearity where there is a nonlinear relationship between the variables.





Nonlinear Relationship and Transformation

Transformation of a nonlinear relationship to a linear form can be seen below:

Nonlinear relationship Y = $\alpha\ e^{\beta X}$	Transformation steps	Linear form
After taking log on both the sides, log Y = log α + βX	Y′ = log Y	Y′ = log α + βX

DATA AND
ARTIFICIAL INTELLIGENCE

Working with Categorical Data

simplilearn

Working with Categorical Variables

Some potential predictors are categorical and qualitative.

To accommodate these variables in the regression model, they should be transformed into dummy variables.

Original Data

Price	LivingArea	Region
16858	1629	East
26049	1344	West
26130	822	East
31113	1540	East
40932	1320	West
44674	1214	North
44873	882	South
45004	960	North
49564	1363	West

Transforming Into Dummy Variables

Price	LivingArea	East	West	North
16858	1629	1	0	0
26049	1344	0	1	0
26130	822	1	0	0
31113	1540	1	0	0
40932	1320	0	1	0
44674	1214	0	0	1
44873	882	0	0	0
45004	960	0	0	1
49564	1363	0	1	0

Validation Framework

Creating a Validation Framework

To test the performance of a model in new scenarios, validation frameworks need to be created.

Popular validation frameworks include:

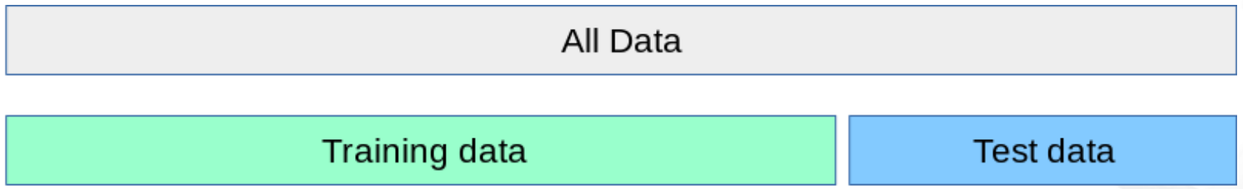
- Hold-out-based validation
- K-fold cross-validation

Hold-Out-Based Validation

Hold-out-based validation

K-fold cross Validation

The dataset is randomly split into training and test data.



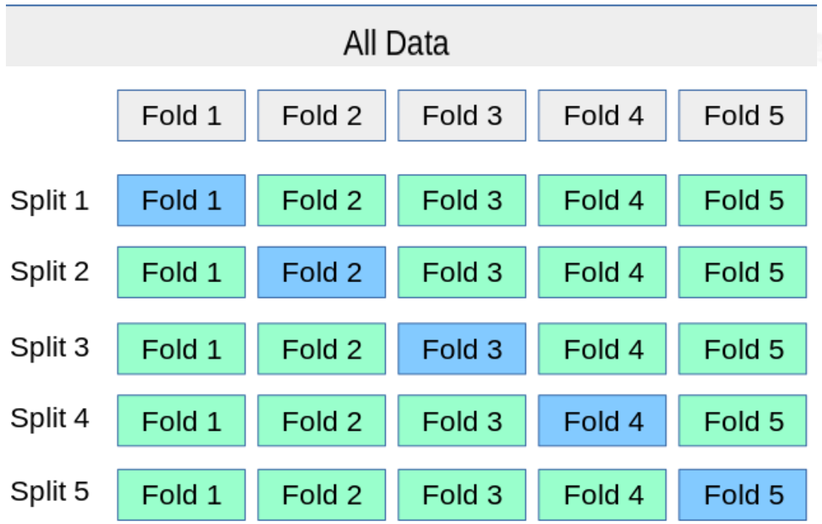
K-Fold Cross-Validation

Hold-out based Validation

K-fold cross-validation

The original dataset is equally partitioned into k subparts or folds.

Out of the k-folds, for each iteration, one group is selected as validation data; the remaining (k-1) groups are selected as training data.



Detecting and Eliminating Multicollinearity

Duration: 5 minutes

Problem Scenario: The ***housing.csv*** file contains details of median price values of houses in Boston suburbs. The data describes the various features of the neighborhoods.

Multicollinearity is a problem in regression as it generates a high variance of the estimated coefficients.

Check for multicollinearity in the regression model built on ***housing.csv*** data and devise a solution to reduce it in the model.

Note: Please download the data set and the solution document from the **Course Resources** section and follow the steps given in the document

ASSISTED PRACTICE

Key Takeaways

- Correlation describes the association between two numerical variables.
- Regression analysis estimates the relationship between dependent and independent variables.
- Simple linear regression uses only one independent variable to estimate relationships, while multiple linear regression uses more than one variable.
- Regression models can be evaluated using R squared, adjusted R squared, mean squared error, root mean squared error, or mean absolute error metrics.

Knowledge Check

Knowledge Check

1

Which of the statements are true?

A. R squared represents the variation in dependent variables explained by the independent variable.

B. Adjusted R squared value always increases as we increase the number of variables in the model.

C. High RMSE value indicates a good model.

D. All of the above statements are true.

Knowledge Check

1

Which of the statements are true?

A. R squared represents the variation in dependent variables explained by the independent variable.

B. Adjusted R squared value always increases as we increase the number of variables in the model.

C. High RMSE value indicates a good model.

D. All of the above statements are true.

The correct answer is **A**

Adjusted R squared does not need to increase after the inclusion of new variables. RMSE is a measure of error. Error term in any model should be minimized.

Knowledge Check

2

Which of the following can best detect multicollinearity?

A. Correlation coefficient

B. Variance inflation factor (VIF)

C. Scatter plot

D. None of the above

Knowledge Check

2

Which of the following can best detect multicollinearity?

A. Correlation coefficient

B. Variance inflation factor (VIF)

C. Scatter plot

D. None of the above

The correct answer is **B**

Multicollinearity occurs when there is a strong correlation between the independent variables. It is measured using VIF value. The higher the VIF, higher the multicollinearity.

Knowledge Check

3

What is true for dummy variables?

A. Dummy variables are categorical representation of numerical data.

B. Dummy variables are (n-1) binary variables created from a factor column with n levels.

C. Both A and B

D. None of the above

Knowledge Check

3

What is true for dummy variables?

A. Dummy variables are categorical representation of numerical data.

B. Dummy variables are (n-1) binary variables created from a factor column with n levels.

C. Both A and B

D. None of the above

The correct answer is **B**

Dummy variables are (n-1) binary variables created from a factor column with n levels.