Business Analytics with Excel

Data Analysis Using Statistics

Learning Objectives

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

- Create a moving average chart
- Perform ANOVA to compare means of different groups
- Identify relationships between variables using covariance and correlation
- Calculate regression for the given data
- Create normal distribution for the given data

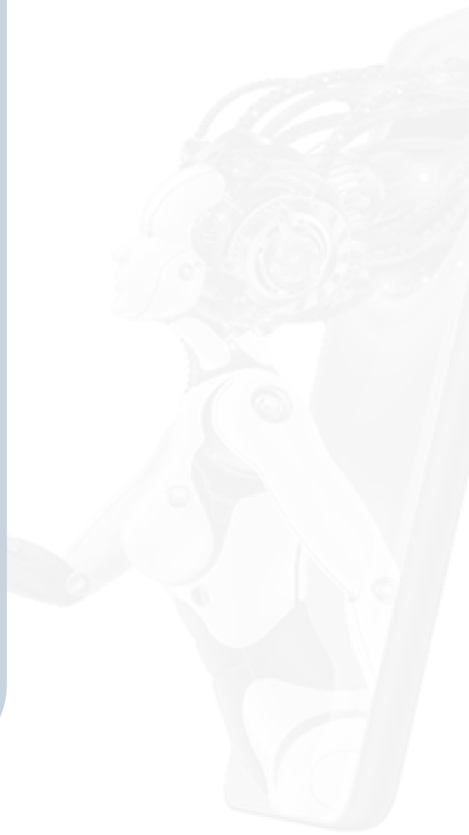
A Day in the Life of Business Analyst

As a business analyst of an organization:

You are required to do forecasting and planning for sales data

Along with the prediction models, you need to co-relate existing data and test any hypothesis.

This lesson will help you understand the usage of statistics for data analytics and predictions.



Introduction to Statistical Analysis

Statistical Analysis

It involves the collection, examination, summarization, manipulation, and interpretation of quantitative data to discover underlying causes, patterns, relationships, and trends.

Need for Statistical Analysis

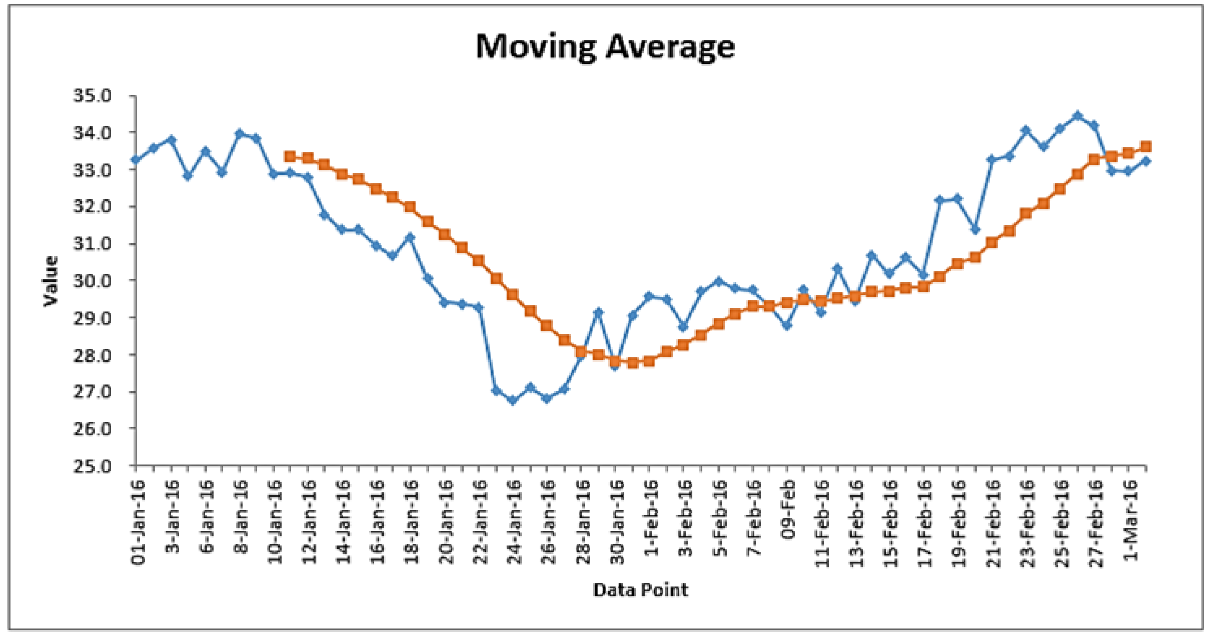
It reveals the overall pattern and behaviour of the data.

It is useful when you have a set of data and want to see a summary of that data set.

Statistical Analysis: Example

ABC LLC is a financial analytics and research organization that needs to determine how stock prices are fluctuating in various emerging economies.

Statistical Analysis: Example

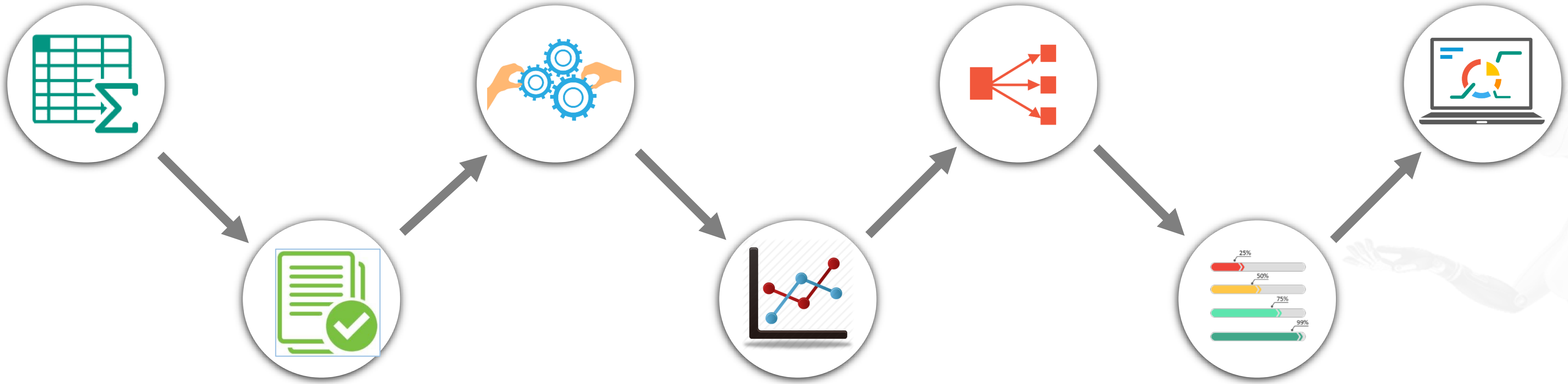


The firm can use the moving average tool based on the historical records and stock market data.

This tool forecasts the price trends for any number of days.

It predicts the trends for the upcoming month by creating a moving average chart.

Statistical Analysis: Tools



Moving Average

Hypothesis Testing

ANOVA

Covariance

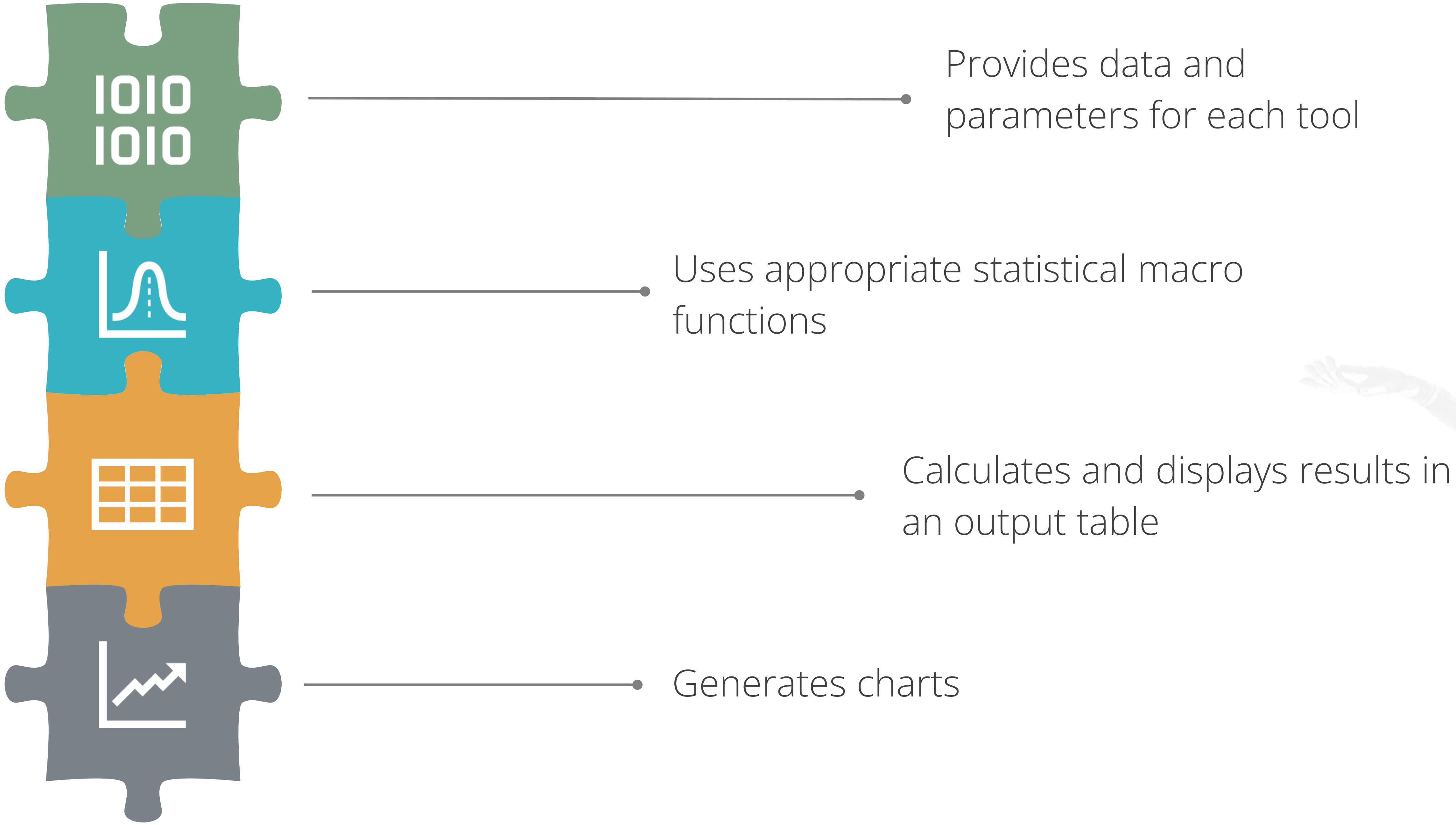
Correlation

Regression

Normal Distribution

Statistical Analysis in Excel

Excel is widely used to understand statistical concepts and perform calculations.



Provides data and parameters for each tool

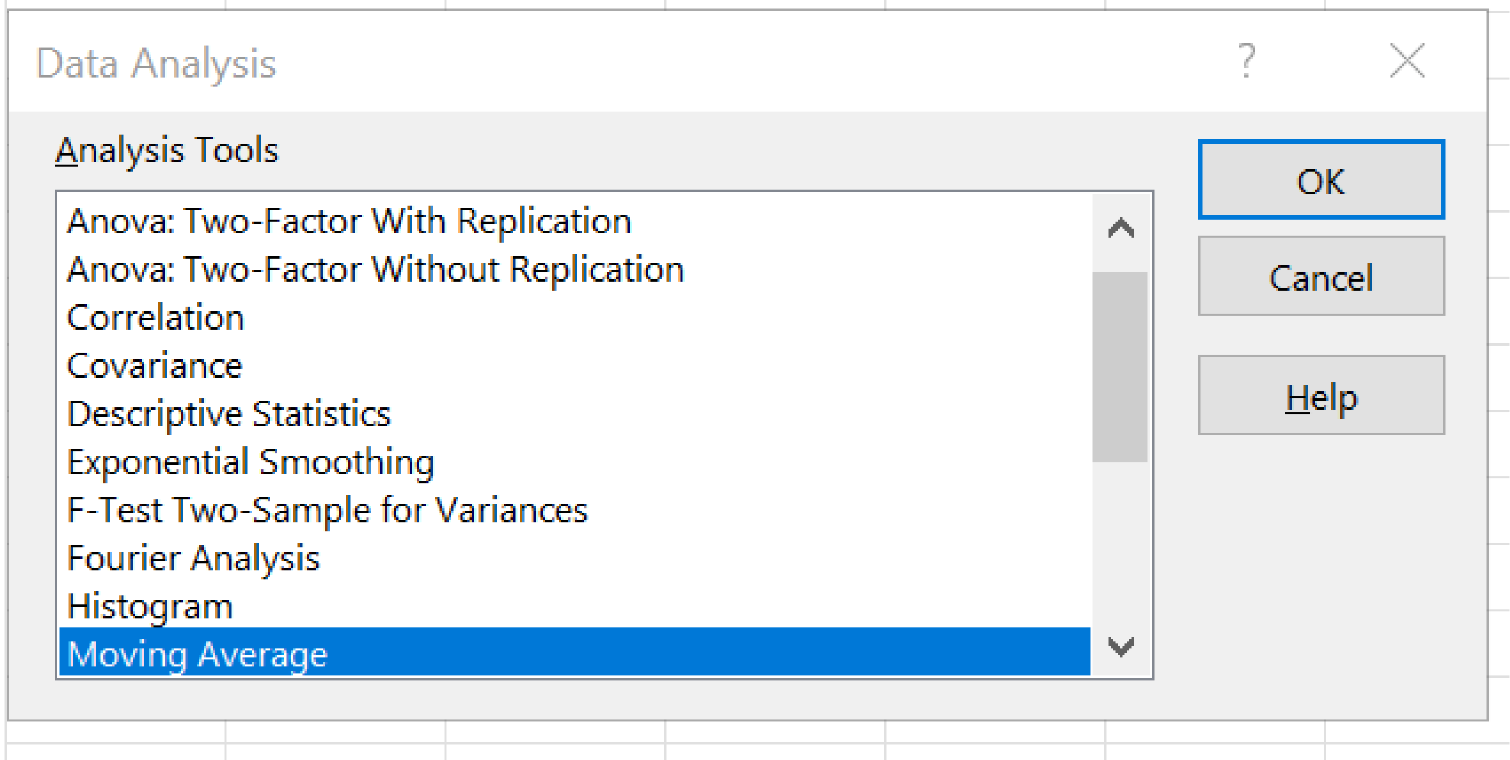
Uses appropriate statistical macro functions

Calculates and displays results in an output table

Generates charts

Data Analysis on Command

Data analysis tools are available under the Data Analysis command under Data tab.



Data Analysis

Analysis Tools

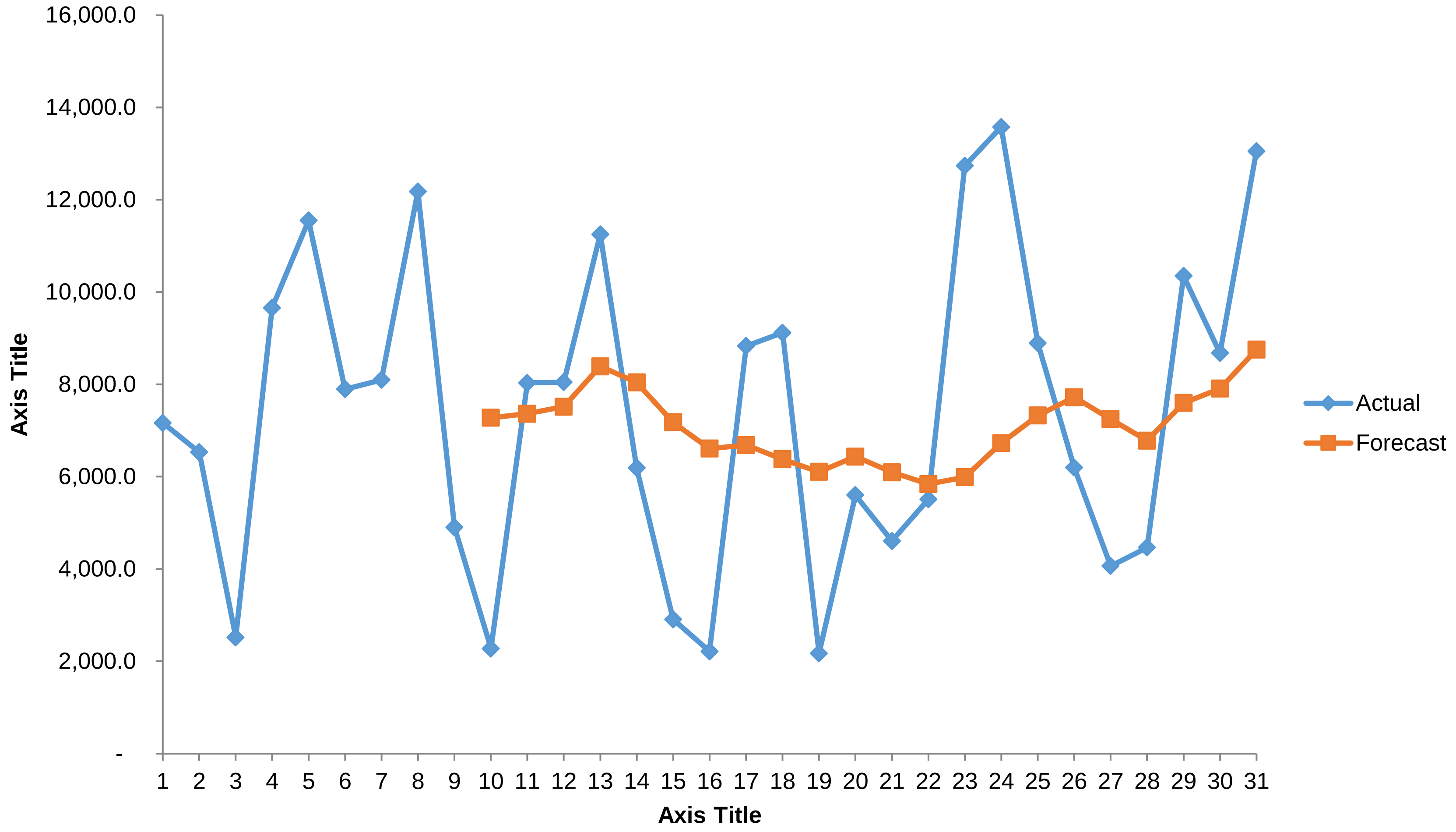
- Anova: Two-Factor With Replication
- Anova: Two-Factor Without Replication
- Correlation
- Covariance
- Descriptive Statistics
- Exponential Smoothing
- F-Test Two-Sample for Variances
- Fourier Analysis
- Histogram
- Moving Average

Analysis ToolPak add-in needs to be loaded if the Data Analysis command is not available.

Moving Average: Introduction

Moving Average

It evaluates data points by creating a series of averages of different subsets of the complete dataset.



A moving average is used to **smooth out irregularities and** easily **recognize trends**.

Moving Average

It is mainly used to forecast long-term trends in the data.

Moving Average can be calculated for any period of time.

Assisted Practice: Create Moving Average Chart

Problem statement:

Demonstrate how to create a Moving Average chart in Excel.

Assisted Practice Guidelines

Steps to follow:

Step 1: Open the Excel file

Step 2: Moving average

Hypothesis Testing: Introduction

Hypothesis Testing

It is used to determine whether there is enough evidence in a data sample to infer that a certain condition is true for the entire population.

Hypothesis Testing

To understand the characteristics of general population:

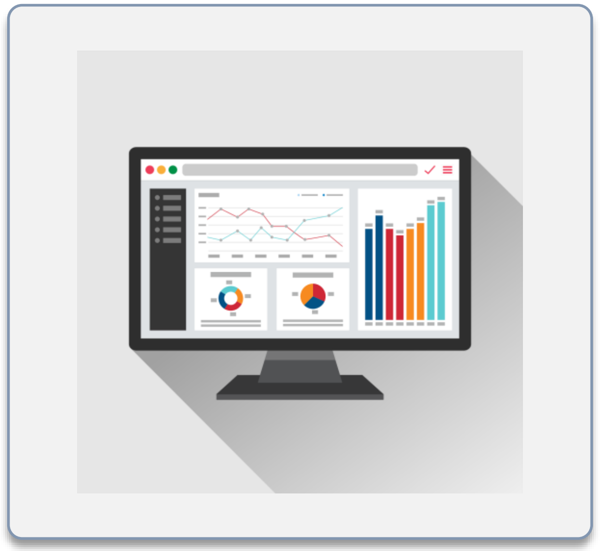
Take a random sample.

Analyze the properties of the sample.

Test whether the identified conclusions represent the population correctly or not.

Hypothesis Testing

A hypothesis about a population parameter is generated.



Sample statistics are used to assess the likelihood that the hypothesis is true.

Hypothesis Testing

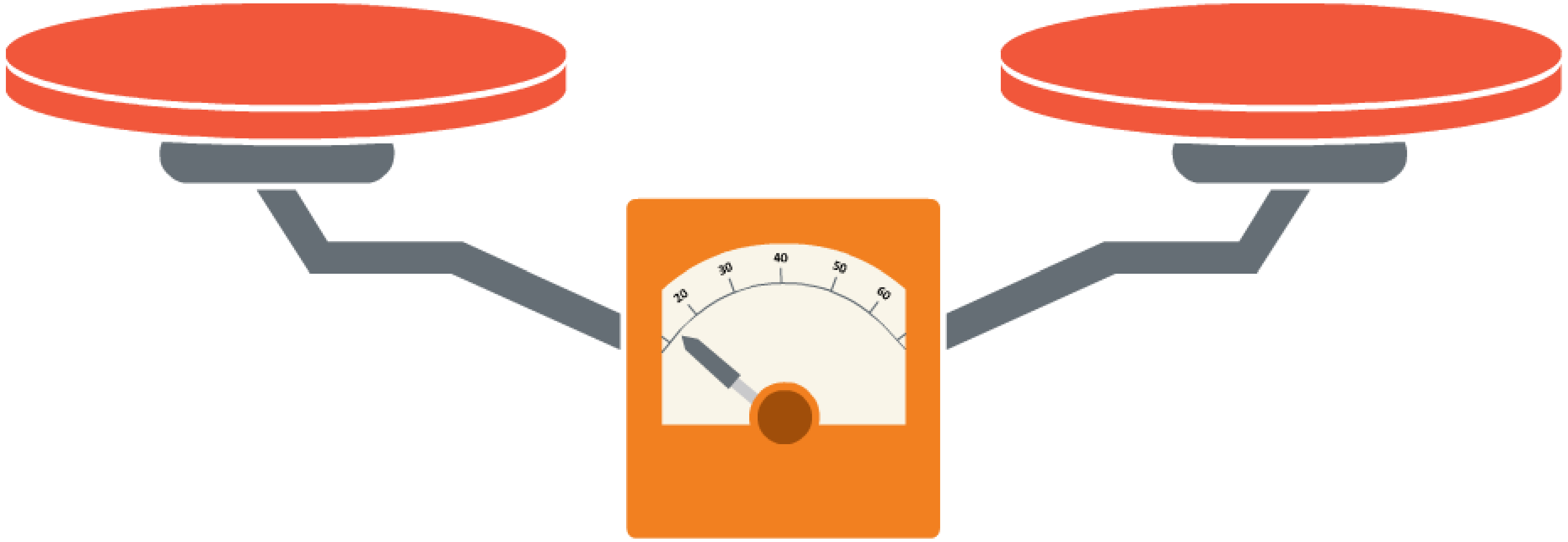
It is formulated in terms of two hypotheses:

Null Hypothesis, which is referred to as H_0, is assumed to be true unless there is strong evidence to the contrary.

Alternate Hypothesis, which is referred to as H_1, is assumed to be true when the null hypothesis is false.

Hypothesis Testing

The **Hypothesis Test (t–test)** is used to test the null hypothesis (H_0), which assumes that the mean or average of two populations is equal.



Assisted Practice: How to use Hypothesis Testing

Problem statement:

Demonstrate how to use Hypothesis Testing to determine Null Hypothesis for two variables.

Assisted Practice Guidelines

Steps to follow:

Step 1: Open the Excel file

Step 2: Hypothesis testing

ANOVA

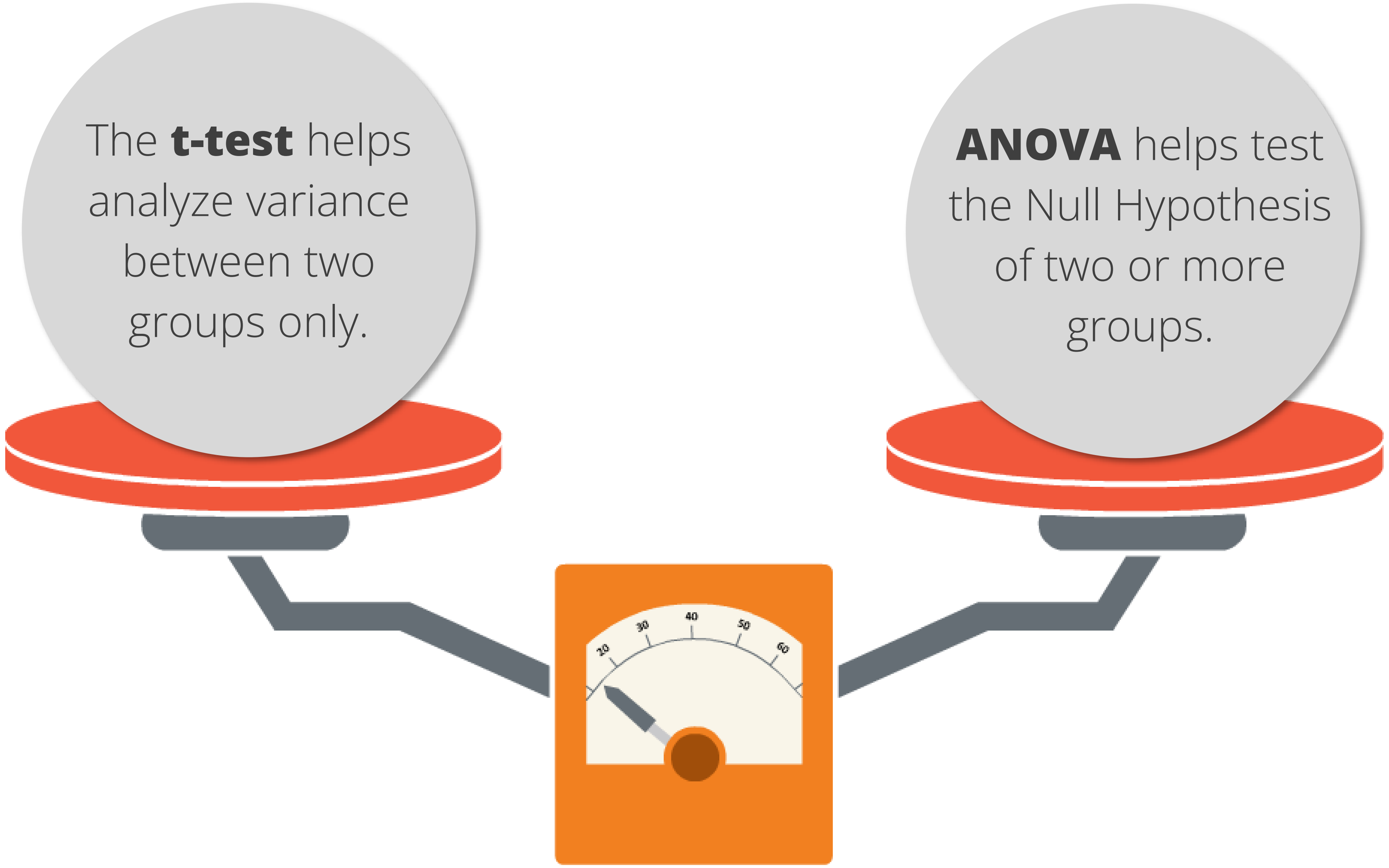
ANOVA

It is a statistical method that stands for analysis of variance.

The logic behind this analysis is to identify variance in the population.

ANOVA is a collection of statistical methods used to compare the means of different groups.

T-Test



The **t-test** helps analyze variance between two groups only.

ANOVA helps test the Null Hypothesis of two or more groups.

Assisted Practice: How to use ANOVA

Problem statement:

Demonstrate how to ANOVA to determine Null Hypothesis for two or more variables.

Assisted Practice Guidelines

Steps to follow:

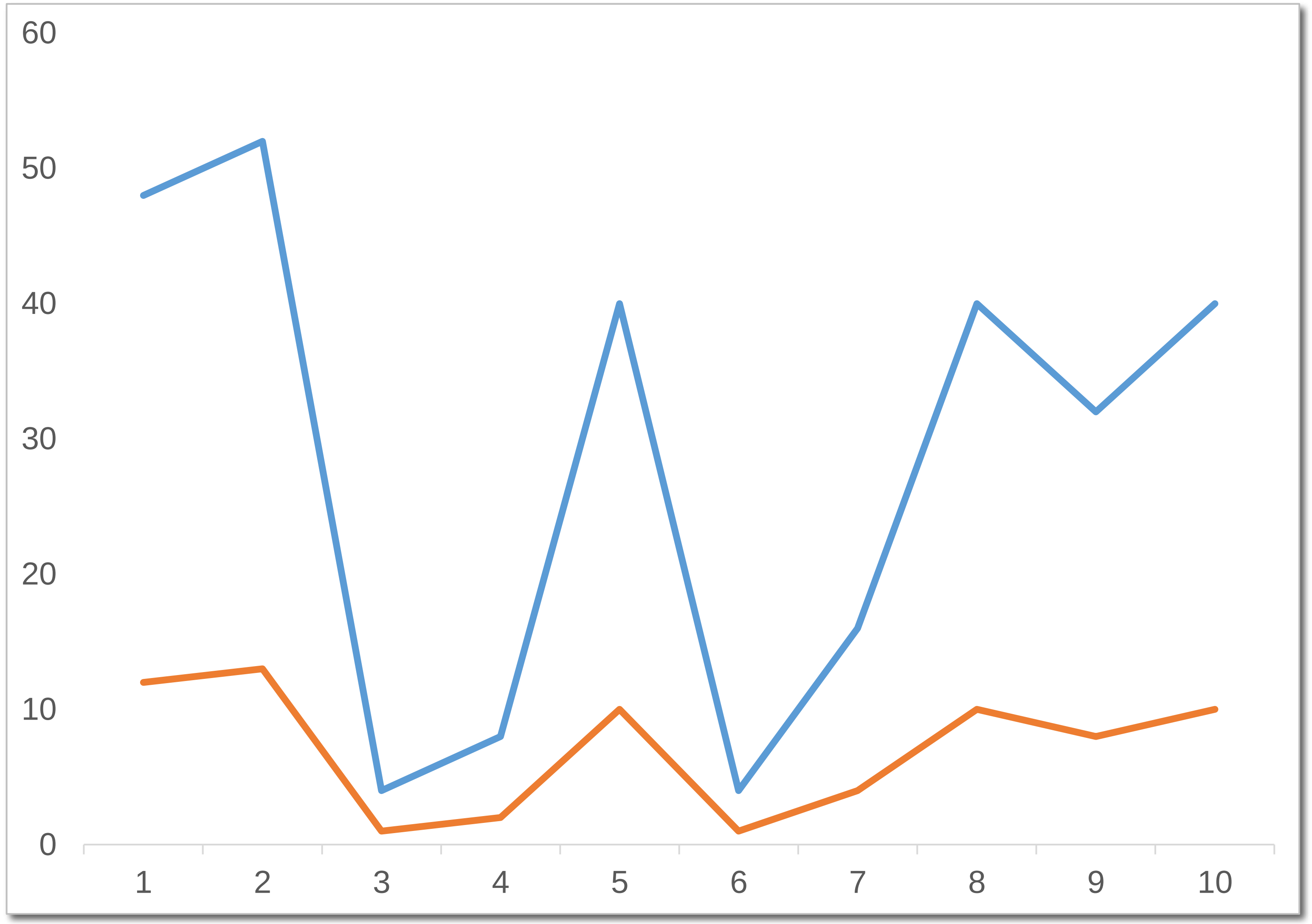
Step 1: Open the Excel file

Step 2: ANOVA testing

Covariance

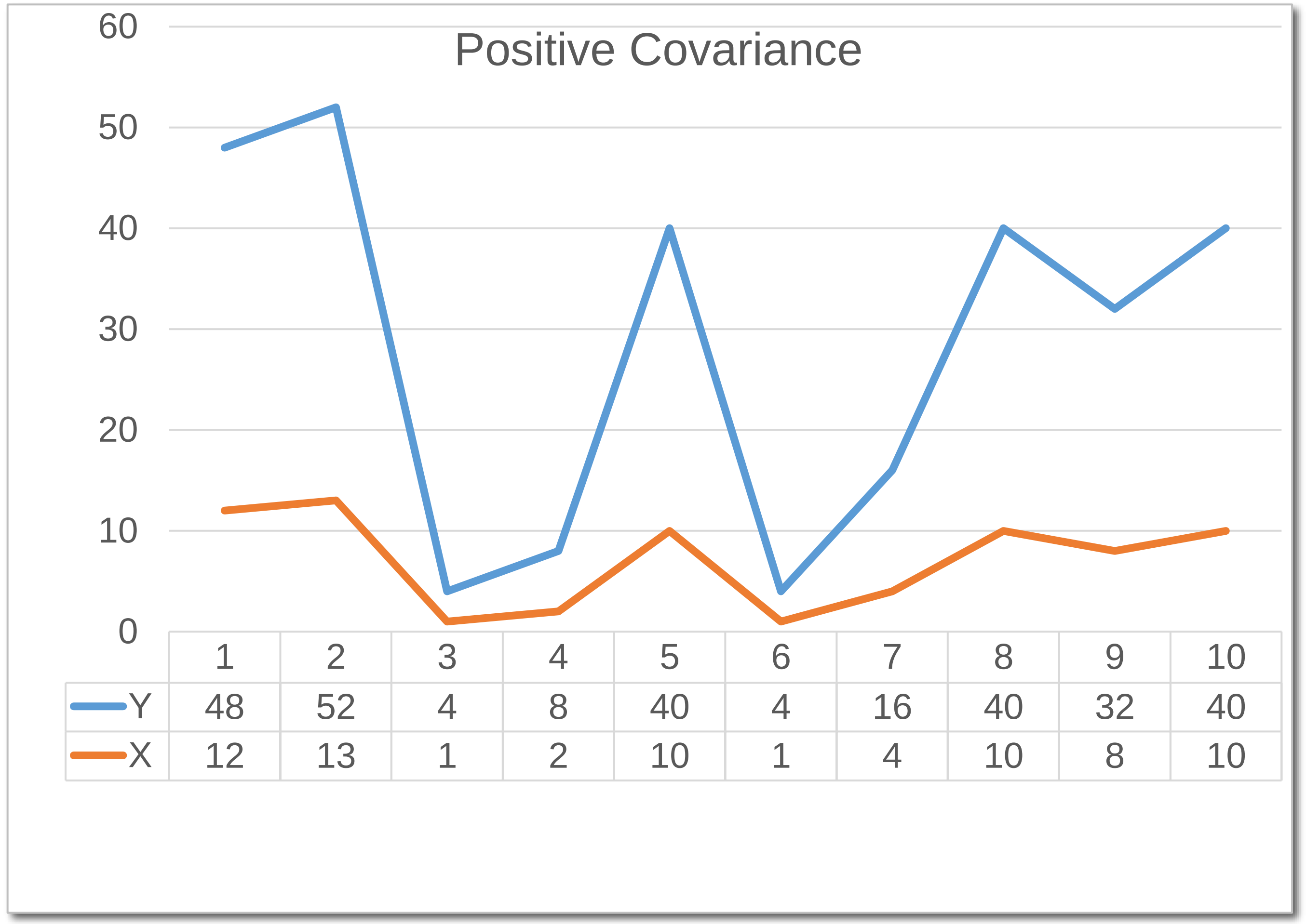
Covariance: Introduction

Covariance determines the relationship between two random variables and how they change together.



Covariance: Types

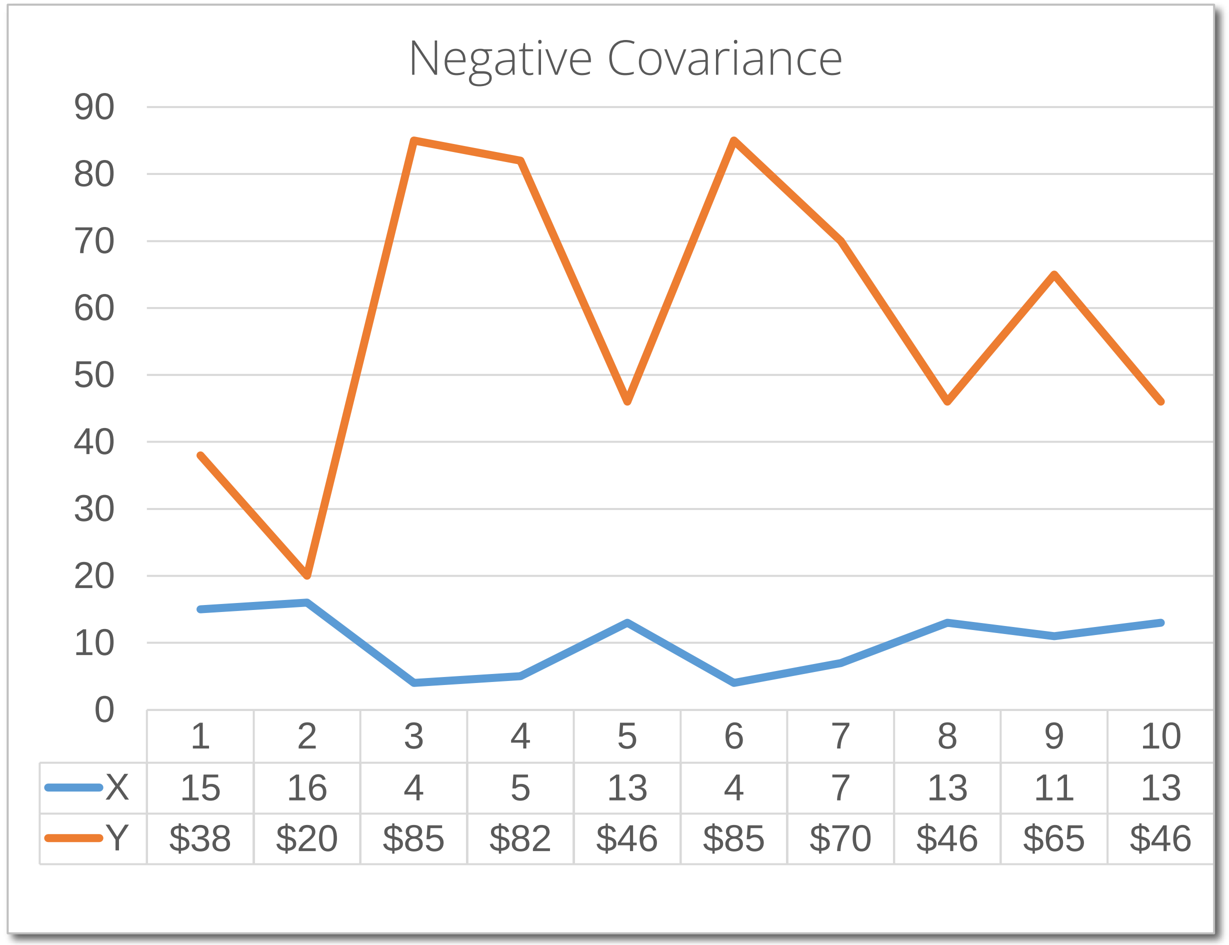
Let us suppose that X and Y are two random variables.



	1	2	3	4	5	6	7	8	9	10
Y	48	52	4	8	40	4	16	40	32	40
X	12	13	1	2	10	1	4	10	8	10

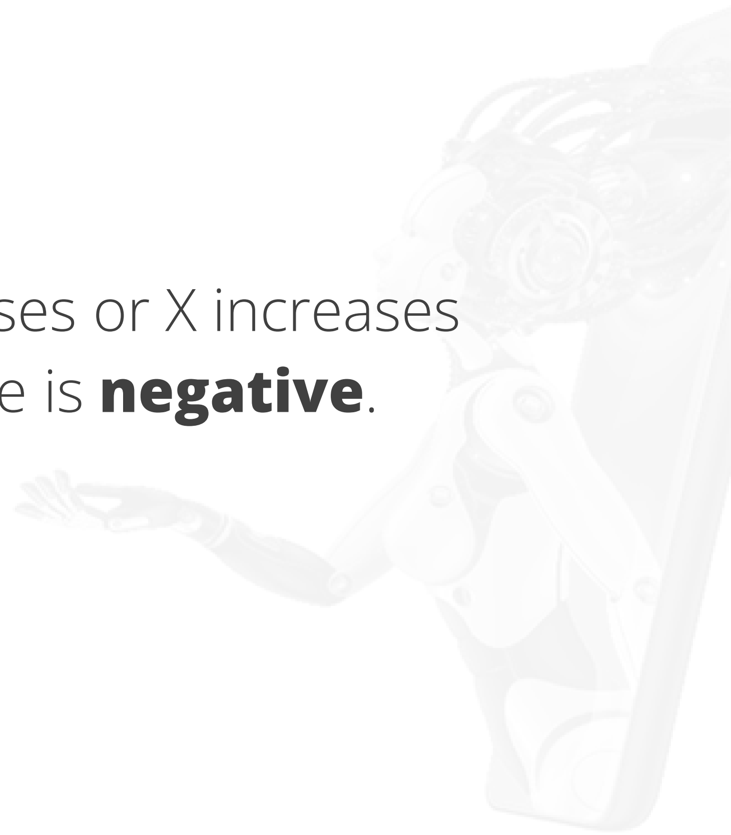
If variable X increases as Y increases or X decreases as Y decreases, then covariance is **positive**.

Covariance: Types



	1	2	3	4	5	6	7	8	9	10
X	15	16	4	5	13	4	7	13	11	13
Y	$38	$20	$85	$82	$46	$85	$70	$46	$65	$46

If variable X decreases as Y increases or X increases as Y decreases, then covariance is **negative**.



Assisted Practice: How to use Covariance

Problem statement:

Demonstrate how to use Covariance in Excel.

Assisted Practice Guidelines

Steps to follow:

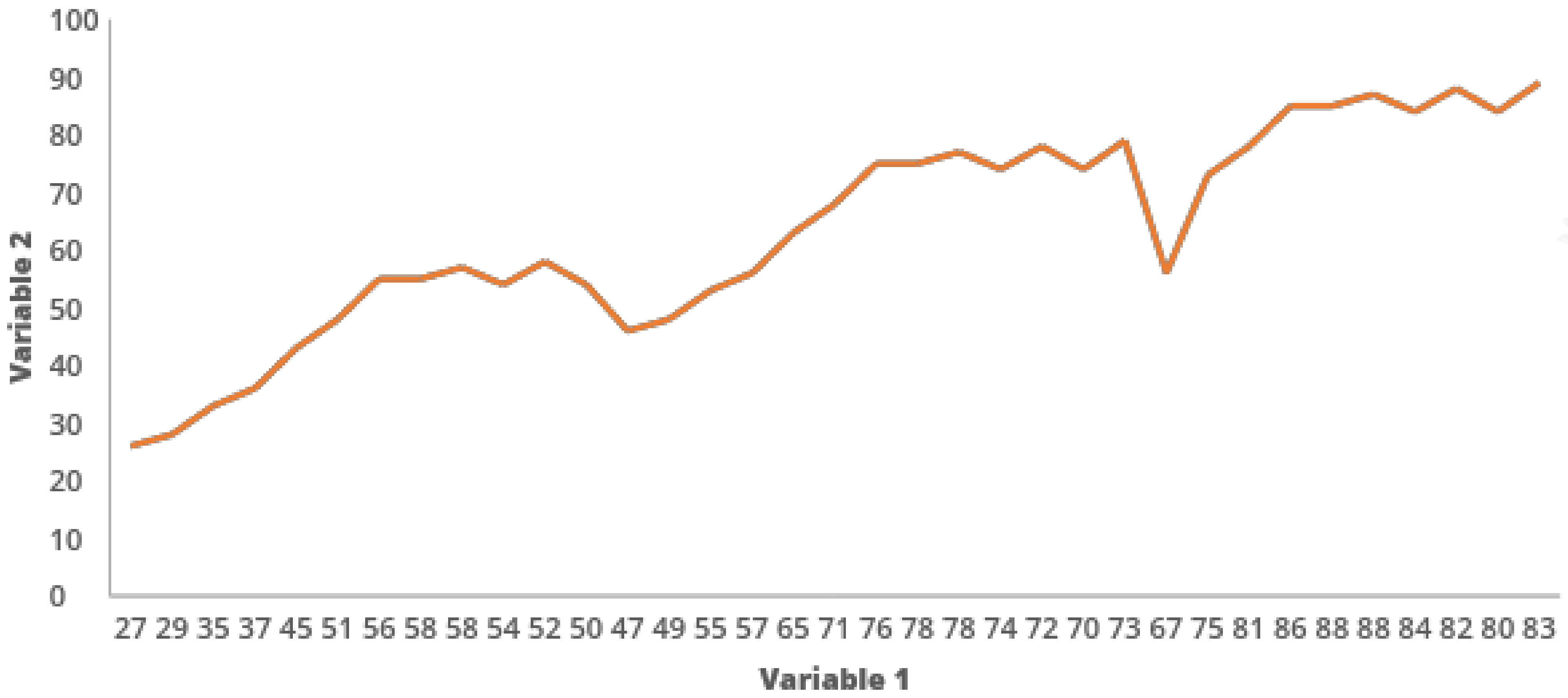
Step 1: Open the Excel file

Step 2: Use Covariance

Correlation

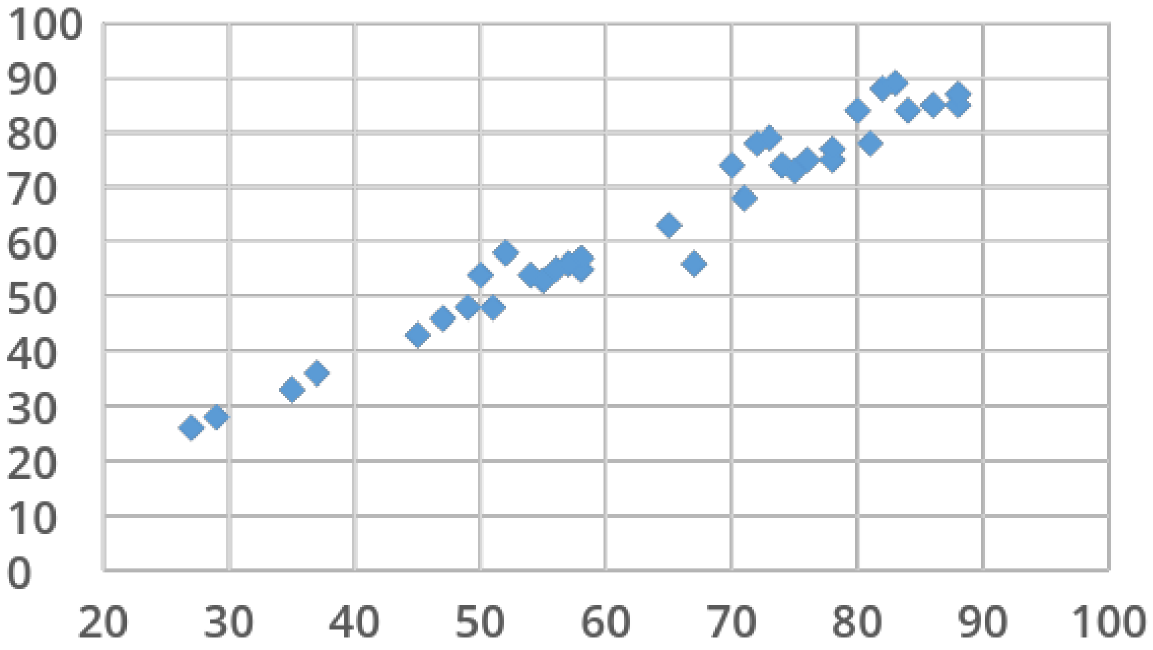
Correlation: Introduction

Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together.



Correlation Coefficient

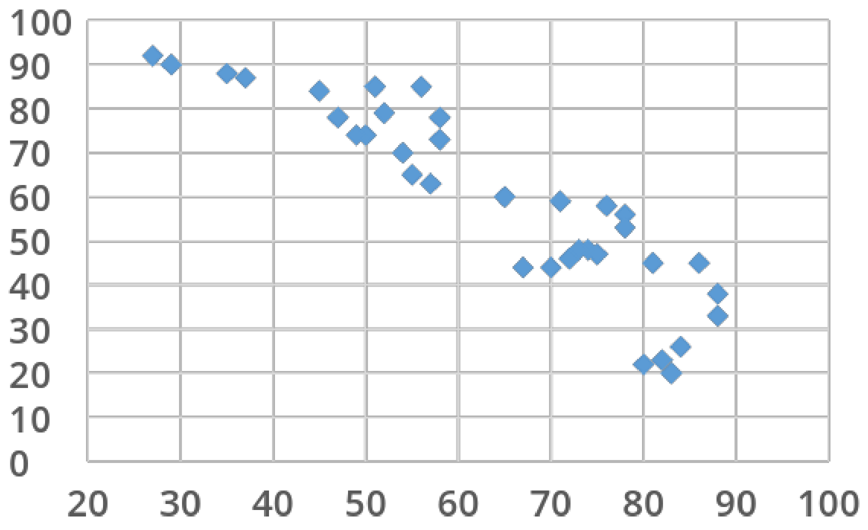
The correlation coefficient tells us how strongly two variables are related to each other and it has a value between -1 and +1.



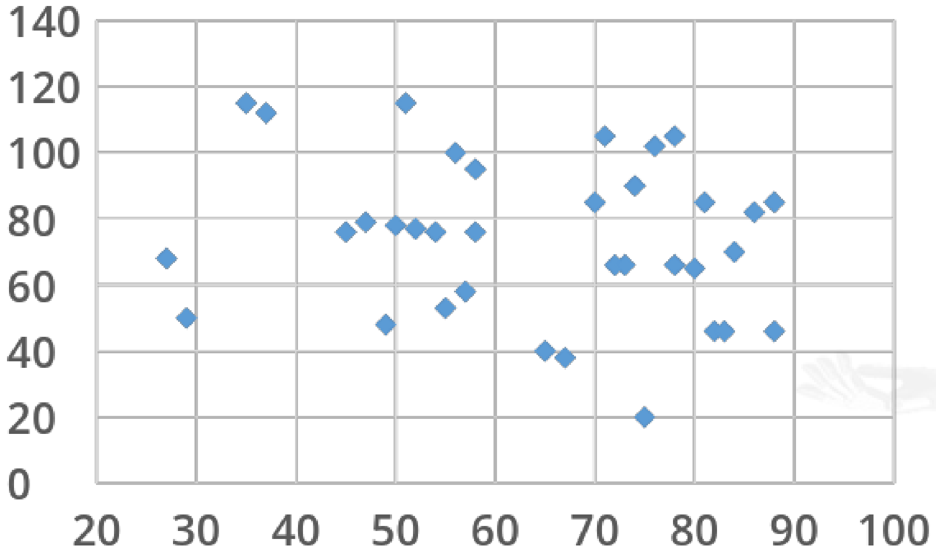
A correlation coefficient with value +1 indicates a perfect positive correlation.

Correlation Coefficient

In Excel, CORREL function is used to calculate correlation.



A correlation coefficient with value -1 indicates a perfect negative correlation.



A correlation coefficient with value 0 indicates no correlation.

Assisted Practice: How to use Correlation

Problem statement:

Demonstrate how to use Correlation in Excel.

Assisted Practice Guidelines

Steps to follow:

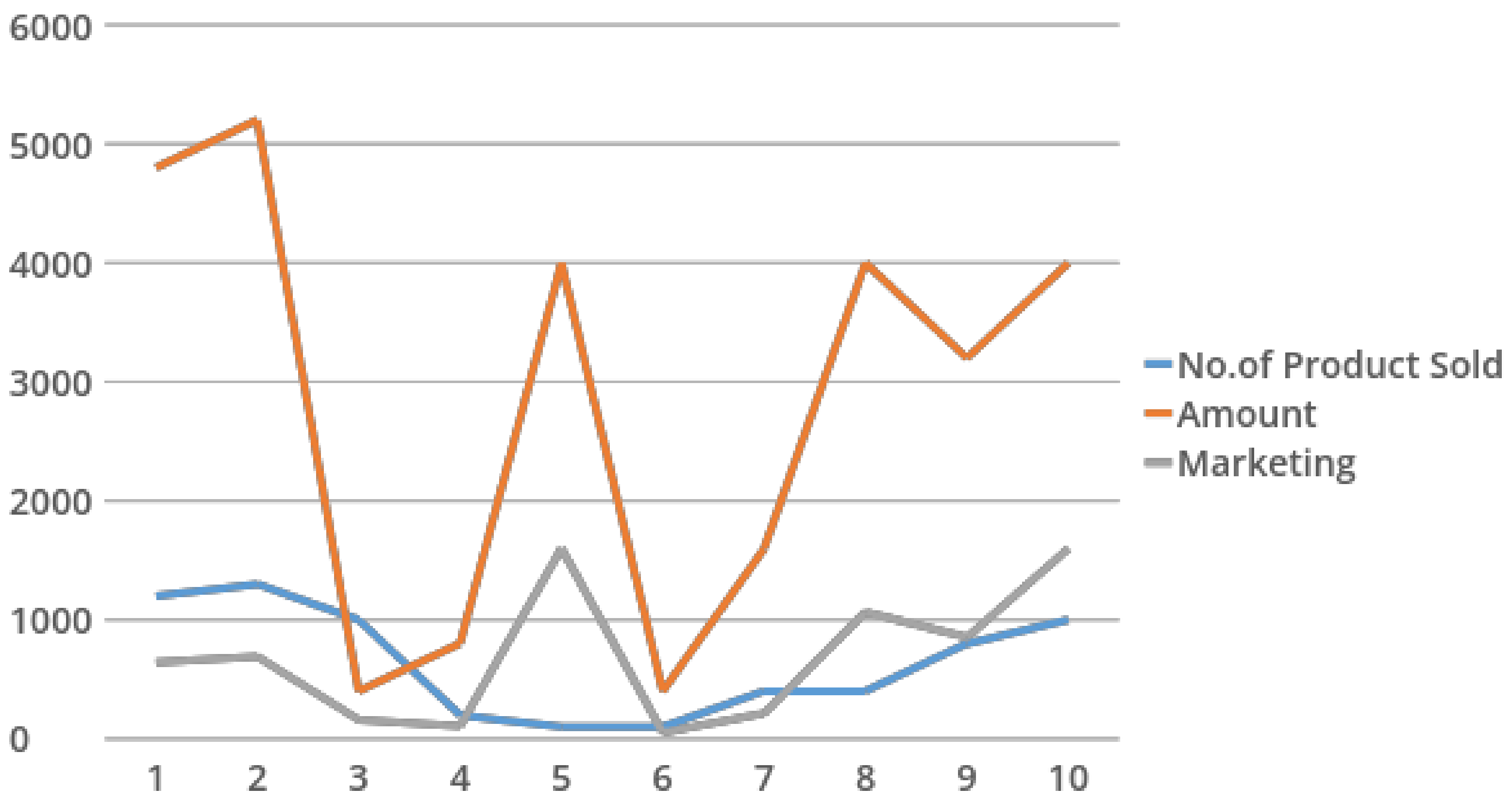
Step 1: Open the Excel file

Step 2: Use Covariance

Regression

Regression: Introduction

Regression is a statistical method for determining the strength of a relationship between one dependent variable and a set of independent variables that change over time.



Assisted Practice: How to use Regression

Problem statement:

Demonstrate how to use Regression to determine relationships between variables.

Assisted Practice Guidelines

Steps to follow:

Step 1: Open the Excel file

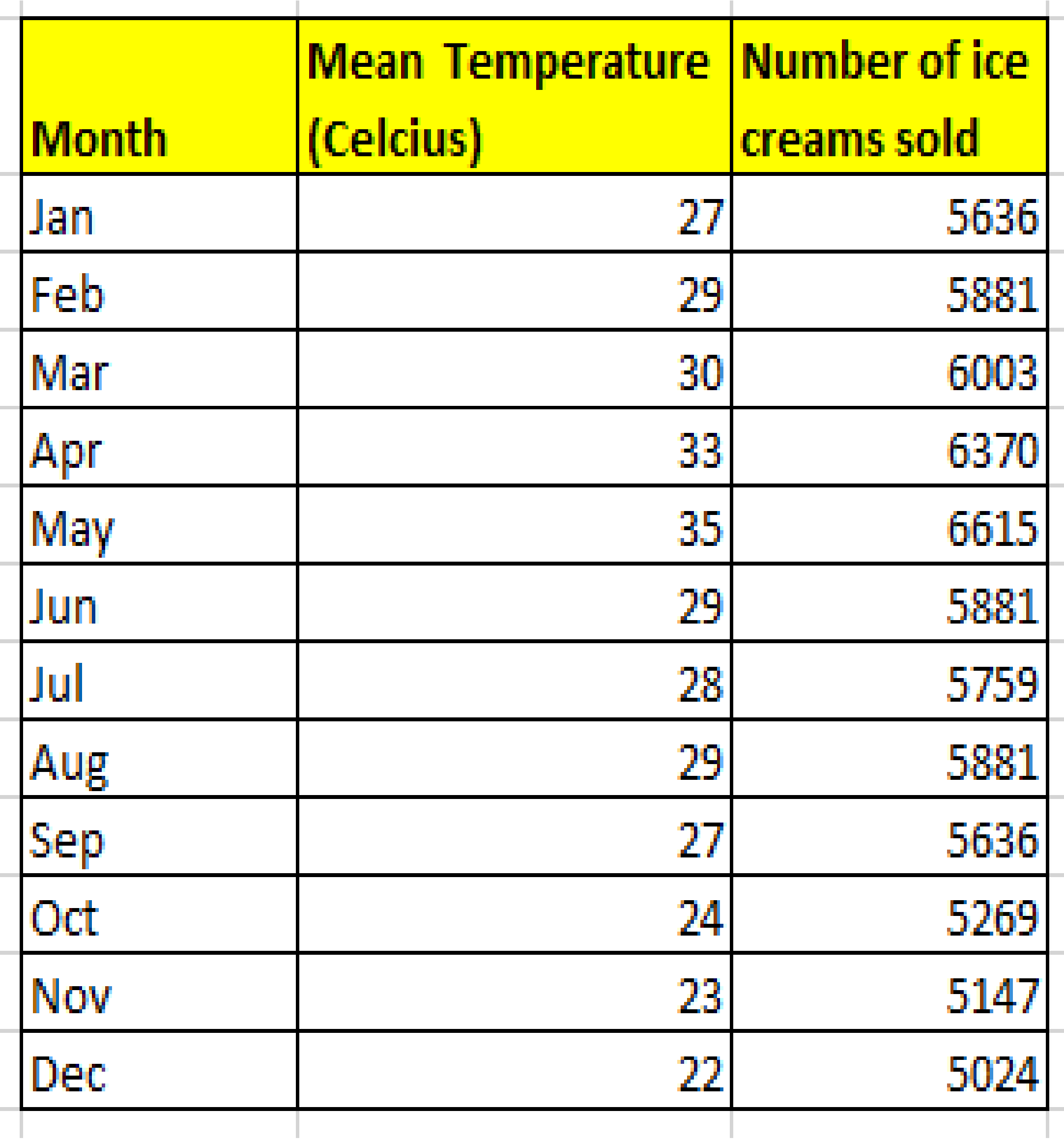
Step 2: Use Regression

Multiple Linear Regression

Simple Linear Regression

Simple Linear Regression (SLR) tries to find a linear representation between two variables x and y.

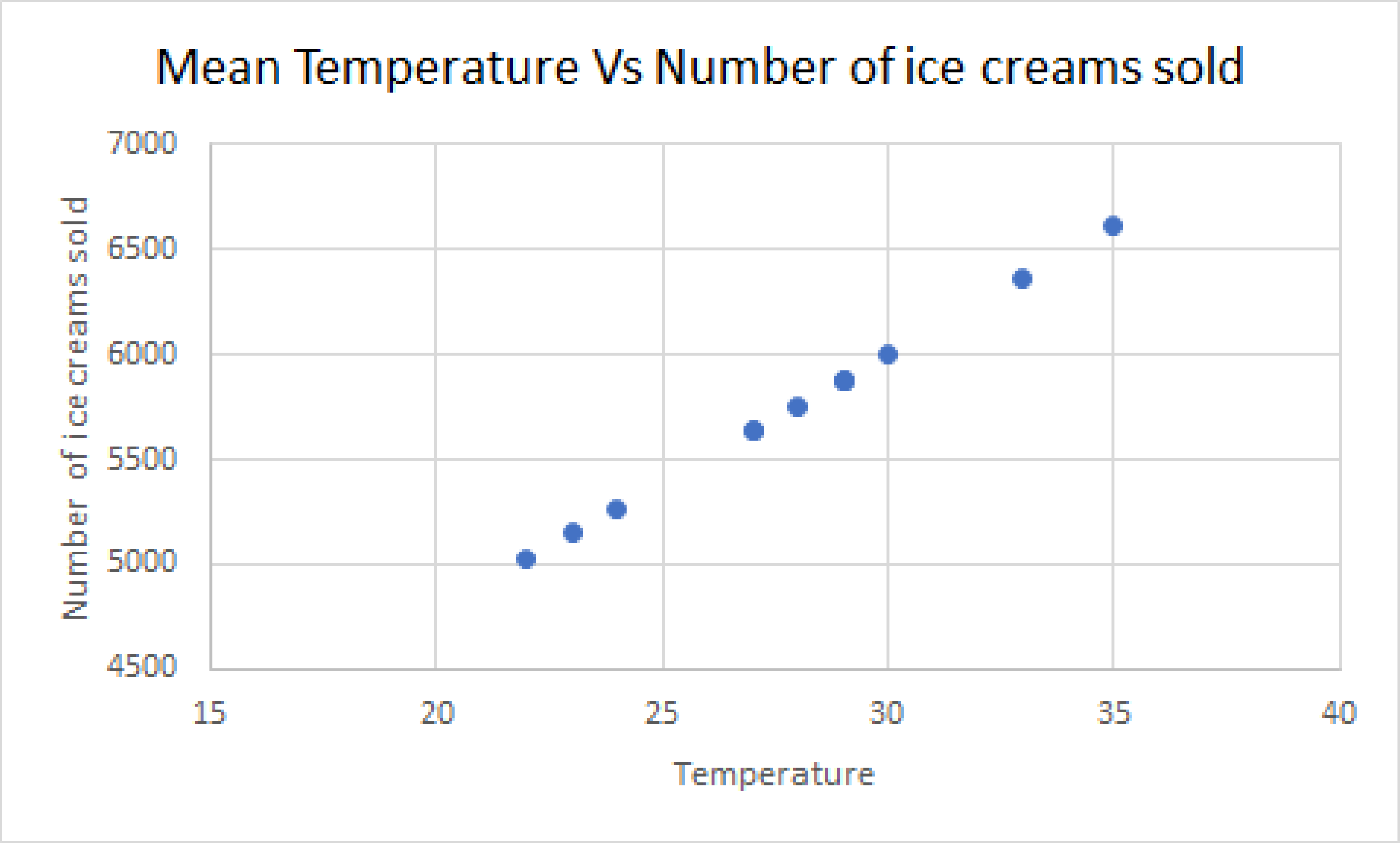
y = function(x)



Month	Mean Temperature (Celcius)	Number of ice creams sold
Jan	27	5636
Feb	29	5881
Mar	30	6003
Apr	33	6370
May	35	6615
Jun	29	5881
Jul	28	5759
Aug	29	5881
Sep	27	5636
Oct	24	5269
Nov	23	5147
Dec	22	5024

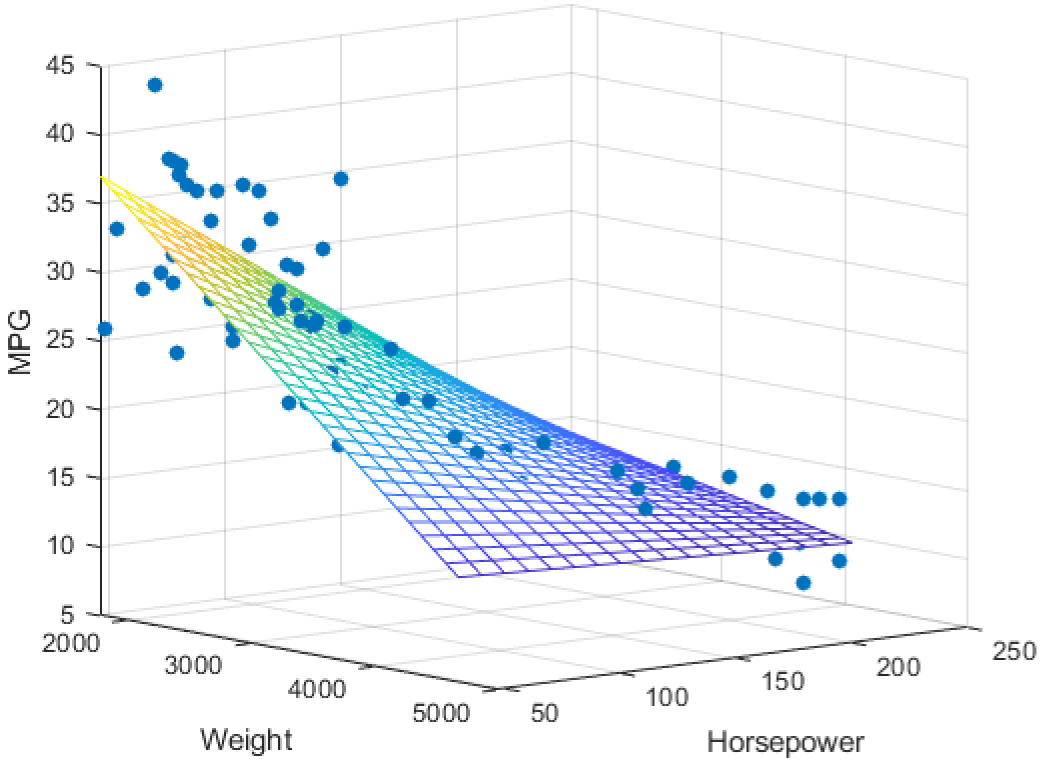
Simple Linear Regression

A linear relation of the temperature and number of ice creams sold can be observed using a scatter plot.



Multiple Linear Regression

Multiple Linear Regression (MLR) tries to find the relationship between multiple independent x's and a single independent y.



Source: *https://medium.com/analytics-vidhya/new-aspects-to-consider-while-moving-from-simple-linear-regression-to-multiple-linear-regression-dad06b3449ff*

Multiple Linear Regression

The approach is to build a fitting line in n-dimensional space to:

- Explain the effects of the independent variables on the y variable.
- Predict y value given in a new set of x variables.

Multiple Linear Regression

The data is fit into the following equation:

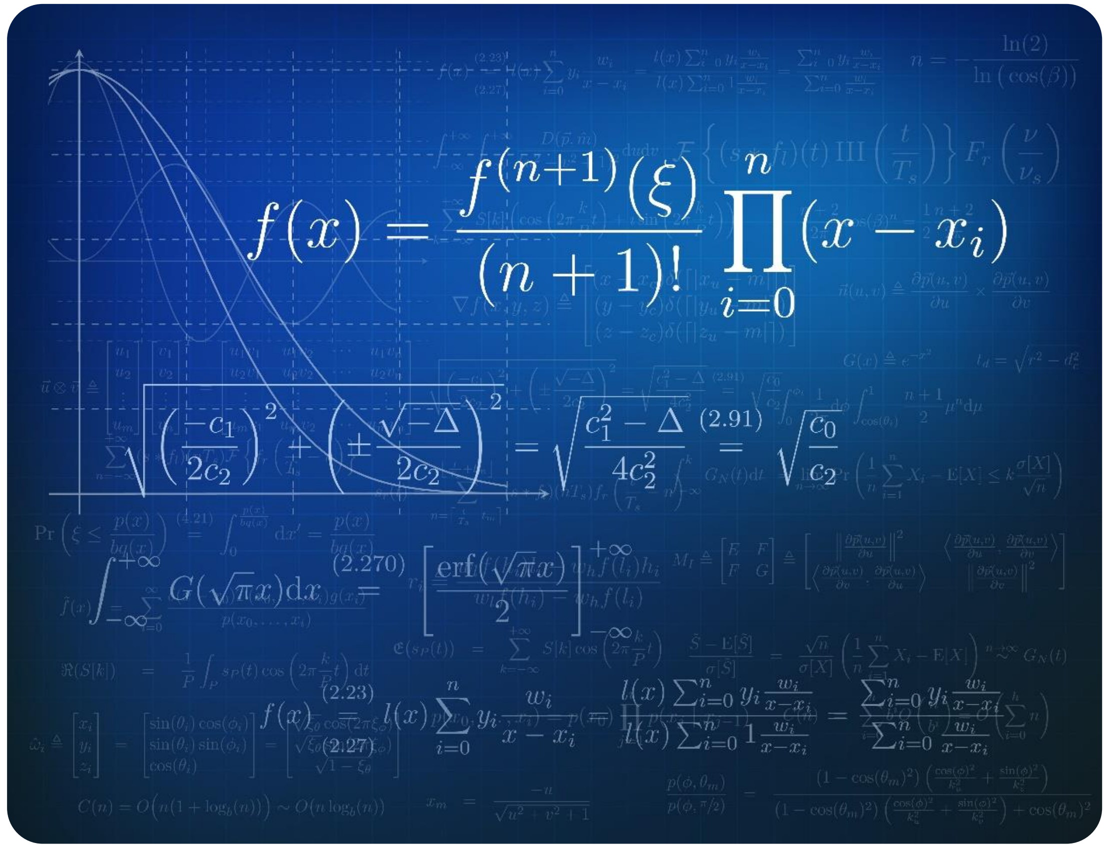
$$y=\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + e$$

Where:

- Y: dependent or resultant variable
- $x_1, x_2, x_3, \ldots, x_i$: independent variables
- β_0: constant term in the equation
- β_i: slope coefficients to each independent variable

Multiple Linear Regression

A multiple linear regression model can be built using Excel with at least 30 data points.


$f(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^{n} (x - x_i)$

The mathematical equation with the coefficients is derived instantly and used to predict new values.

Multiple Linear Regression

Consider the boston_housing.csv as the input data to build our model.

boston_housing.csv

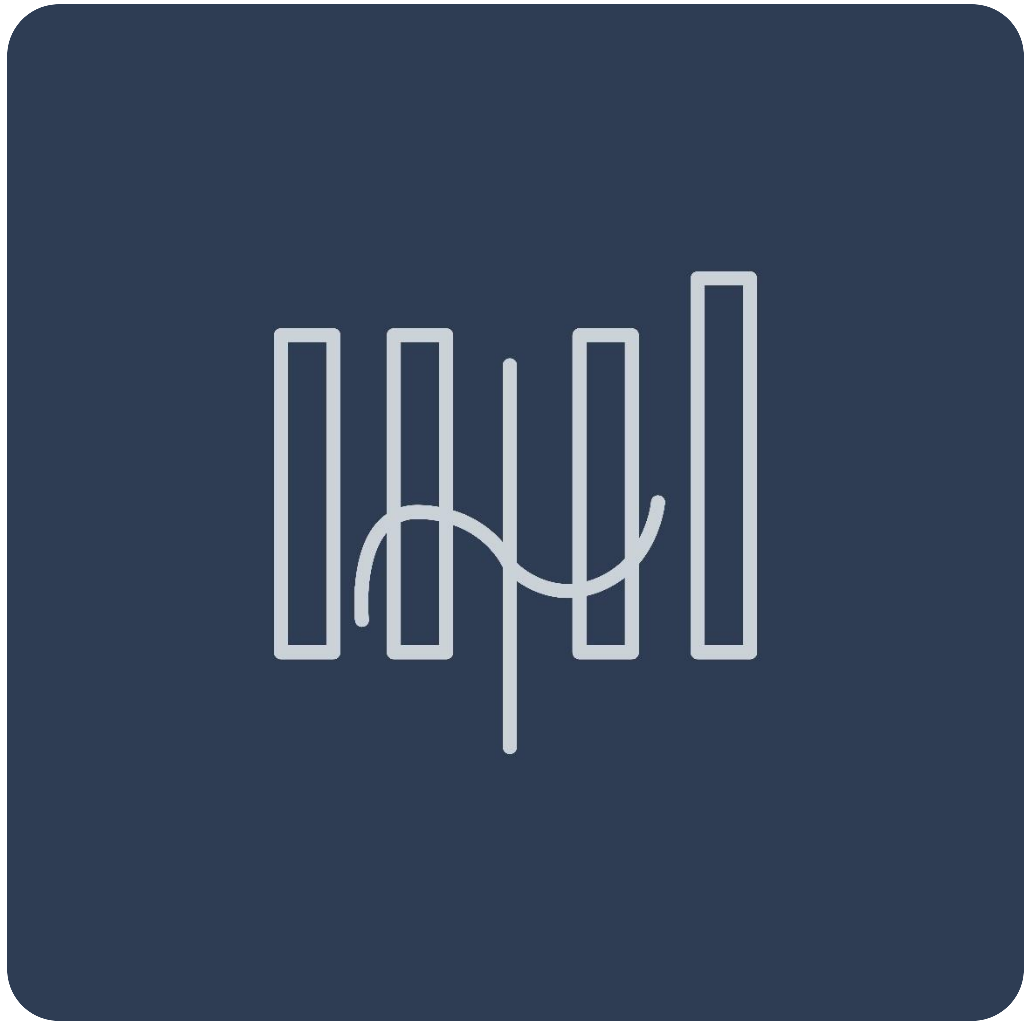
Multiple Linear Regression

The data set contains 13 independent variables which define the dependent variable MEDV.

MEDV is the median value of a house in Boston according to the data provided.

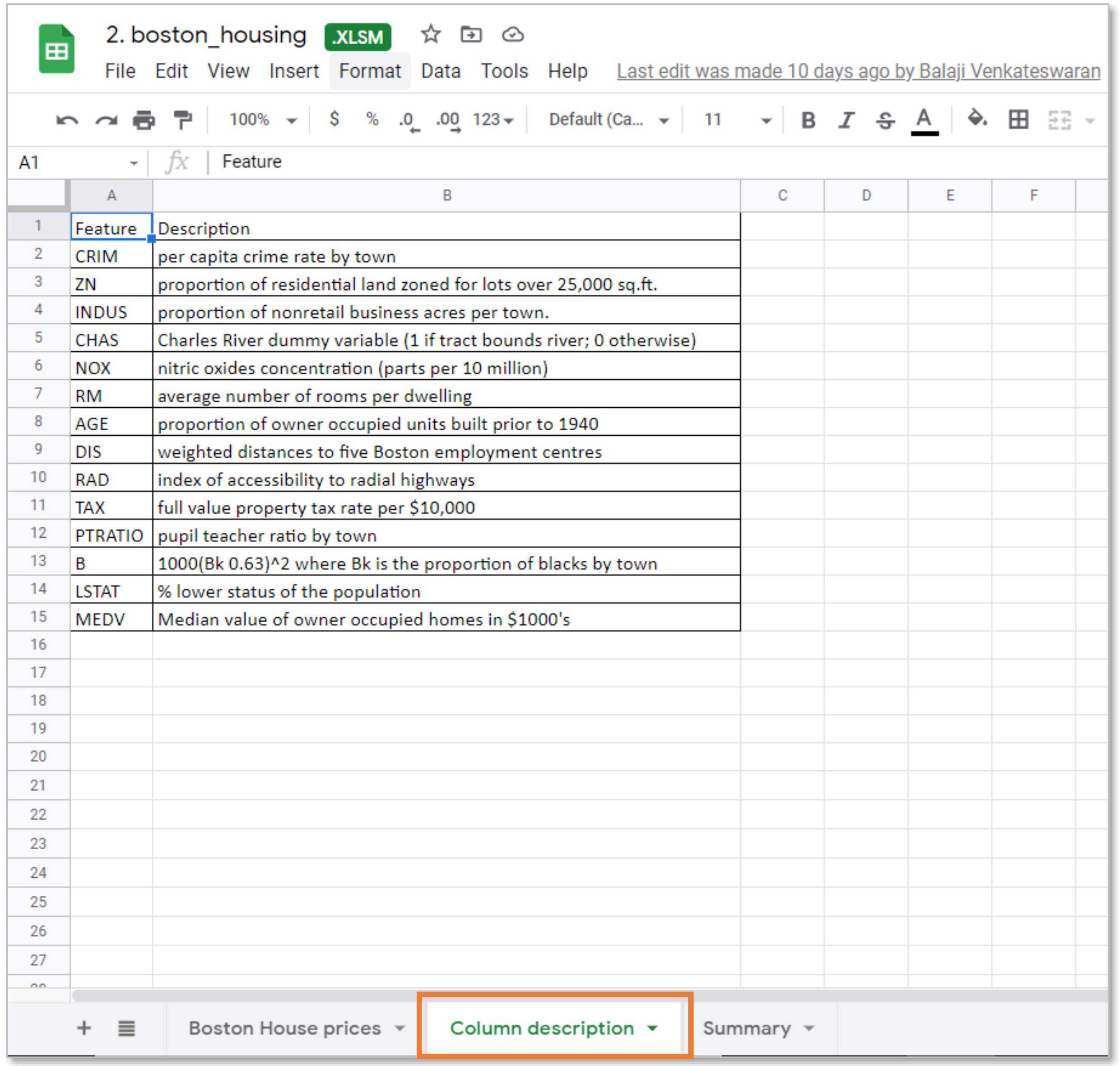
Multiple Linear Regression

A model built using this data can be used to predict the median value of a new house with the attributes of the house.



Multiple Linear Regression

The meaning of each attribute is given in the **Column description** tab.


2. boston_housing .XLSM
File Edit View Insert Format Data Tools Help
Last edit was made 10 days ago by Balaji Venkateswaran
100%
Default (Ca...
11
A1
Feature
Boston House prices
Column description
Summary

Feature	Description
CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	proportion of nonretail business acres per town.
CHAS	Charles River dummy variable (1 if tract bounds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
RAD	index of accessibility to radial highways
TAX	full value property tax rate per $10,000
PTRATIO	pupil teacher ratio by town
B	1000(Bk 0.63)^2 where Bk is the proportion of blacks by town
LSTAT	% lower status of the population
MEDV	Median value of owner occupied homes in $1000's

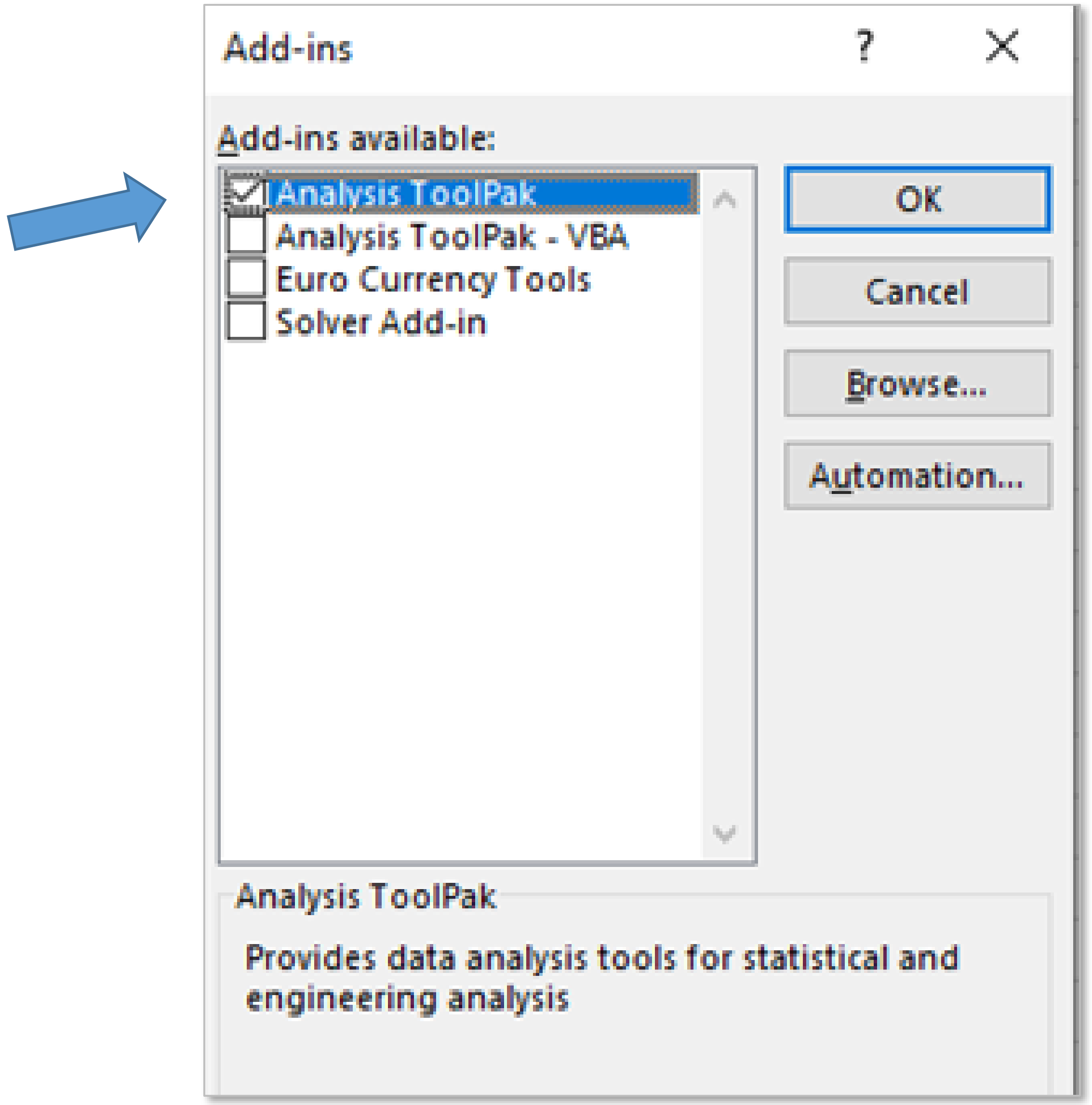
Create a Linear Regression Model

Choose the complete data after checking for any junk data

Click on Data Analysis in Data Tab.
If this does not appear, click on File -> Options -> Excel Add-ins and Go

Create a Linear Regression Model

Click on Analysis ToolPak to enable Data Analysis within Data



Add-ins

Add-ins available:

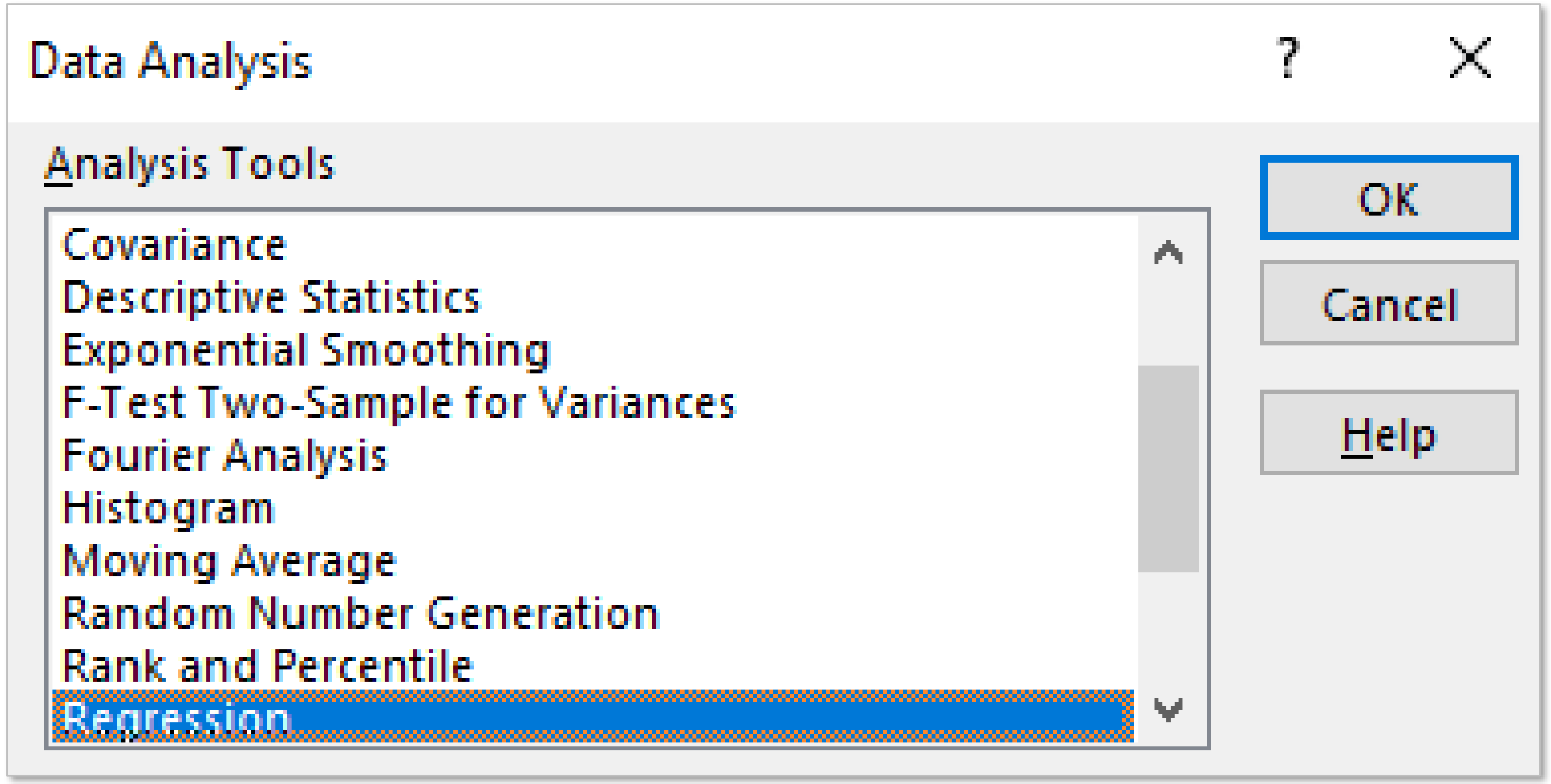
- ☑ Analysis ToolPak
- ☐ Analysis ToolPak - VBA
- ☐ Euro Currency Tools
- ☐ Solver Add-in

Analysis ToolPak

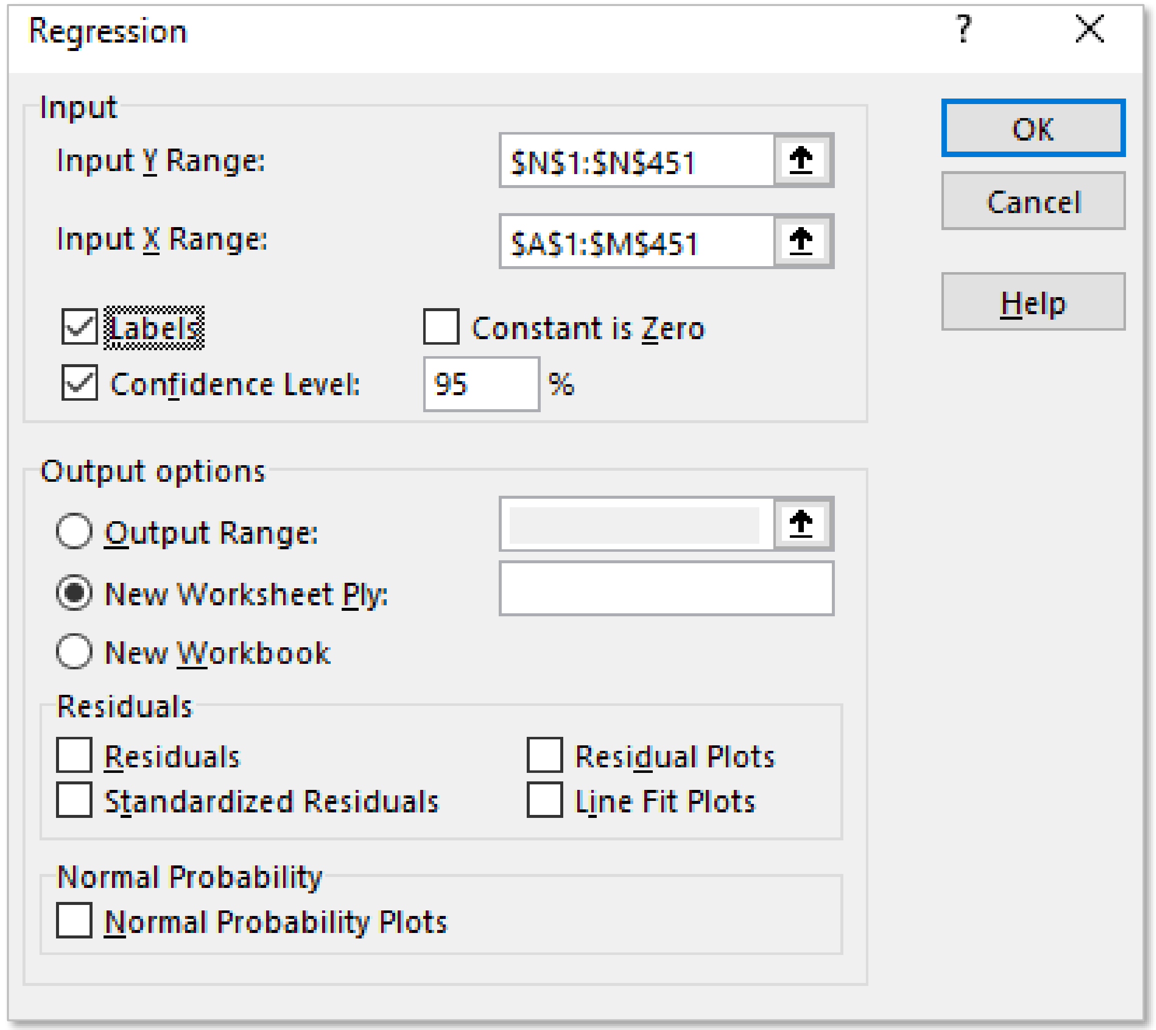
Provides data analysis tools for statistical and engineering analysis

Create a Linear Regression Model

Choose Regression from the Data Analysis dialog box



Create a Linear Regression Model



Regression

Input

Input Y Range: $N$1:$N$451

Input X Range: $A$1:$M$451

☑ Labels ☐ Constant is Zero

☑ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

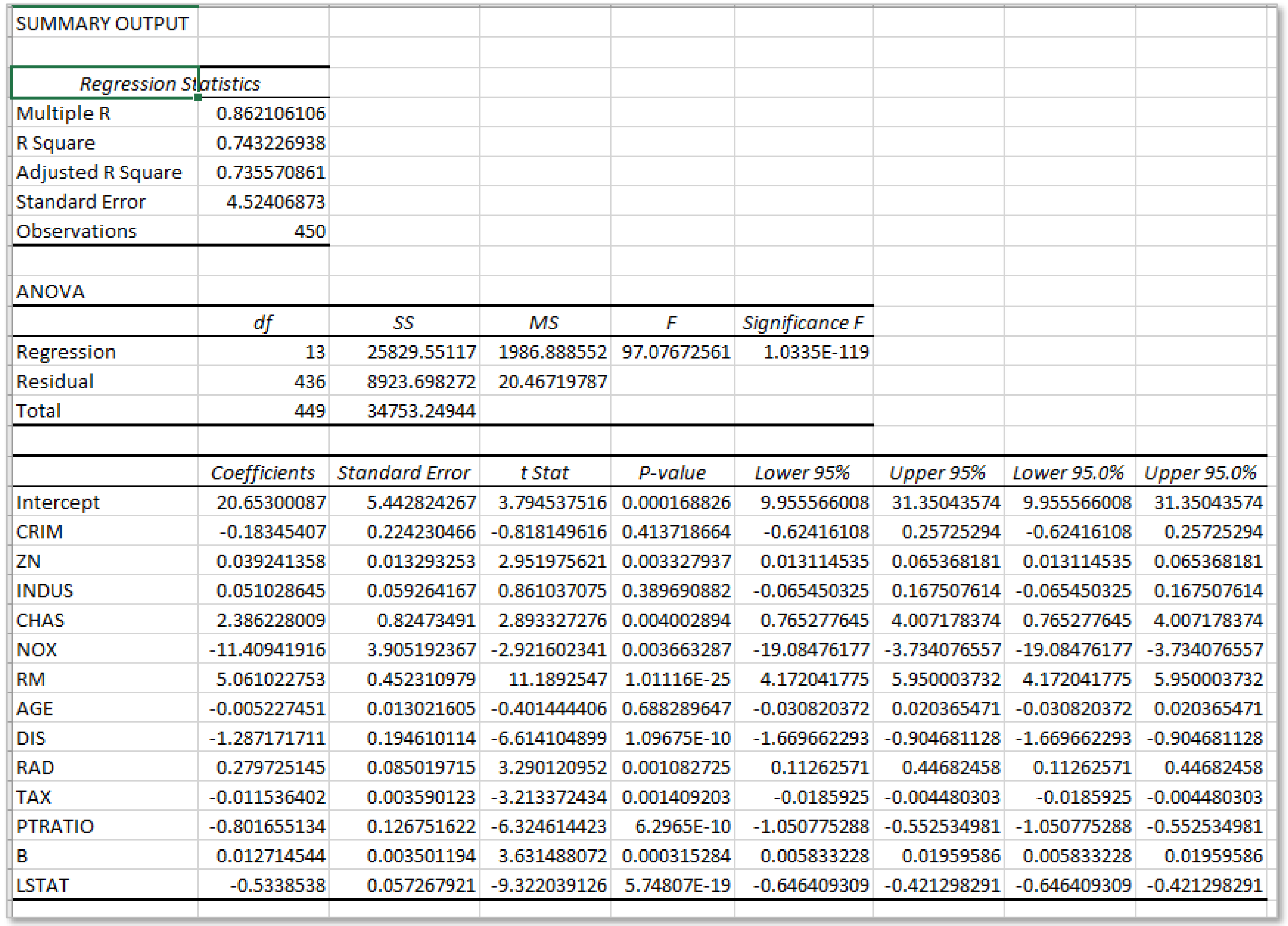
Normal Probability

☐ Normal Probability Plots

- Under Regression, choose rows and columns for the X range and column for the Y range
- Set Labels to present and the Confidence Level to 95%.

Create a Linear Regression Model

The results appear in a new worksheet, showing the regression data for the chosen data set.



SUMMARY OUTPUT

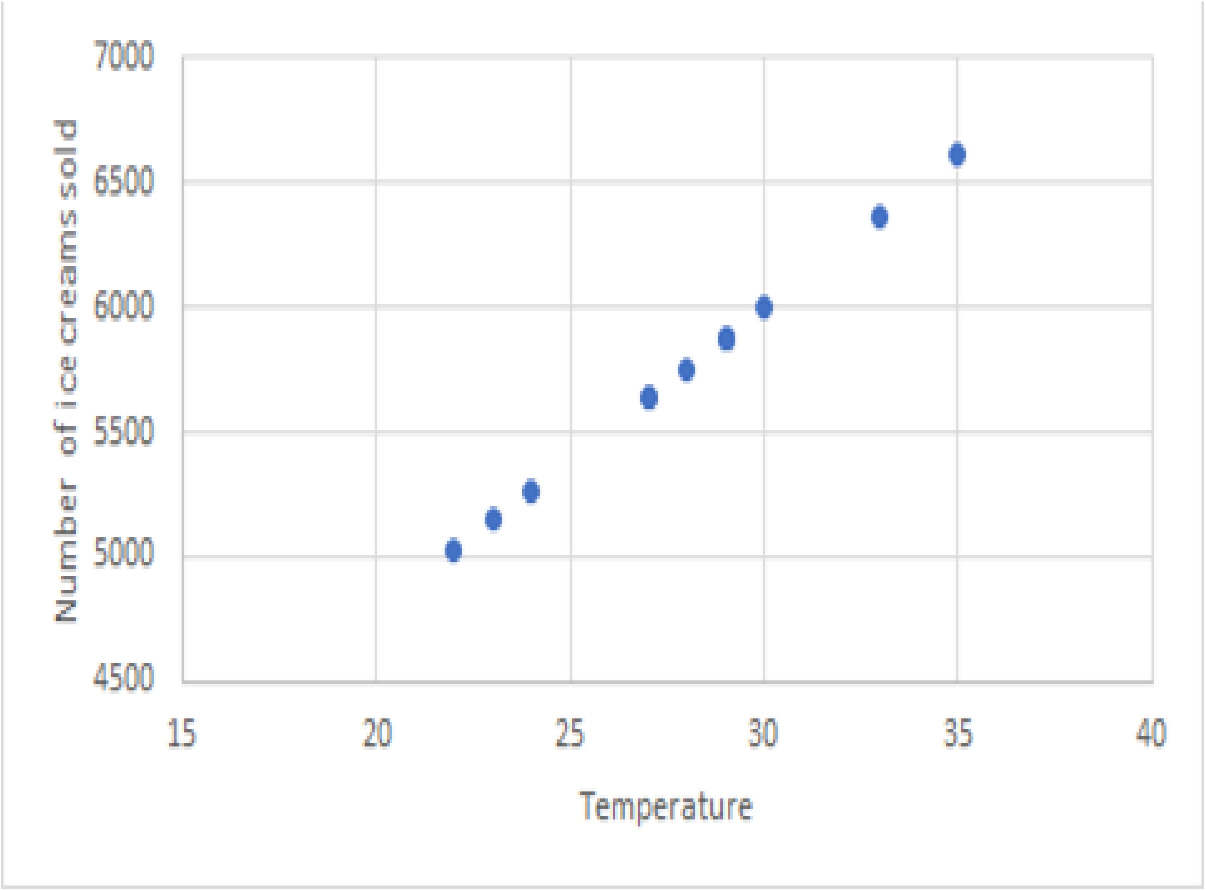
Regression Statistics	
Multiple R	0.862106106
R Square	0.743226938
Adjusted R Square	0.735570861
Standard Error	4.52406873
Observations	450

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	13	25829.55117	1986.888552	97.07672561	1.0335E-119
Residual	436	8923.698272	20.46719787		
Total	449	34753.24944			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	20.65300087	5.442824267	3.794537516	0.000168826	9.955566008	31.35043574	9.955566008	31.35043574
CRIM	-0.18345407	0.224230466	-0.818149616	0.413718664	-0.62416108	0.25725294	-0.62416108	0.25725294
ZN	0.039241358	0.013293253	2.951975621	0.003327937	0.013114535	0.065368181	0.013114535	0.065368181
INDUS	0.051028645	0.059264167	0.861037075	0.389690882	-0.065450325	0.167507614	-0.065450325	0.167507614
CHAS	2.386228009	0.82473491	2.893327276	0.004002894	0.765277645	4.007178374	0.765277645	4.007178374
NOX	-11.40941916	3.905192367	-2.921602341	0.003663287	-19.08476177	-3.734076557	-19.08476177	-3.734076557
RM	5.061022753	0.452310979	11.1892547	1.01116E-25	4.172041775	5.950003732	4.172041775	5.950003732
AGE	-0.005227451	0.013021605	-0.401444406	0.688289647	-0.030820372	0.020365471	-0.030820372	0.020365471
DIS	-1.287171711	0.194610114	-6.614104899	1.09675E-10	-1.669662293	-0.904681128	-1.669662293	-0.904681128
RAD	0.279725145	0.085019715	3.290120952	0.001082725	0.11262571	0.44682458	0.11262571	0.44682458
TAX	-0.011536402	0.003590123	-3.213372434	0.001409203	-0.0185925	-0.004480303	-0.0185925	-0.004480303
PTRATIO	-0.801655134	0.126751622	-6.324614423	6.2965E-10	-1.050775288	-0.552534981	-1.050775288	-0.552534981
B	0.012714544	0.003501194	3.631488072	0.000315284	0.005833228	0.01959586	0.005833228	0.01959586
LSTAT	-0.5338538	0.057267921	-9.322039126	5.74807E-19	-0.646409309	-0.421298291	-0.646409309	-0.421298291

Linear Regression Model

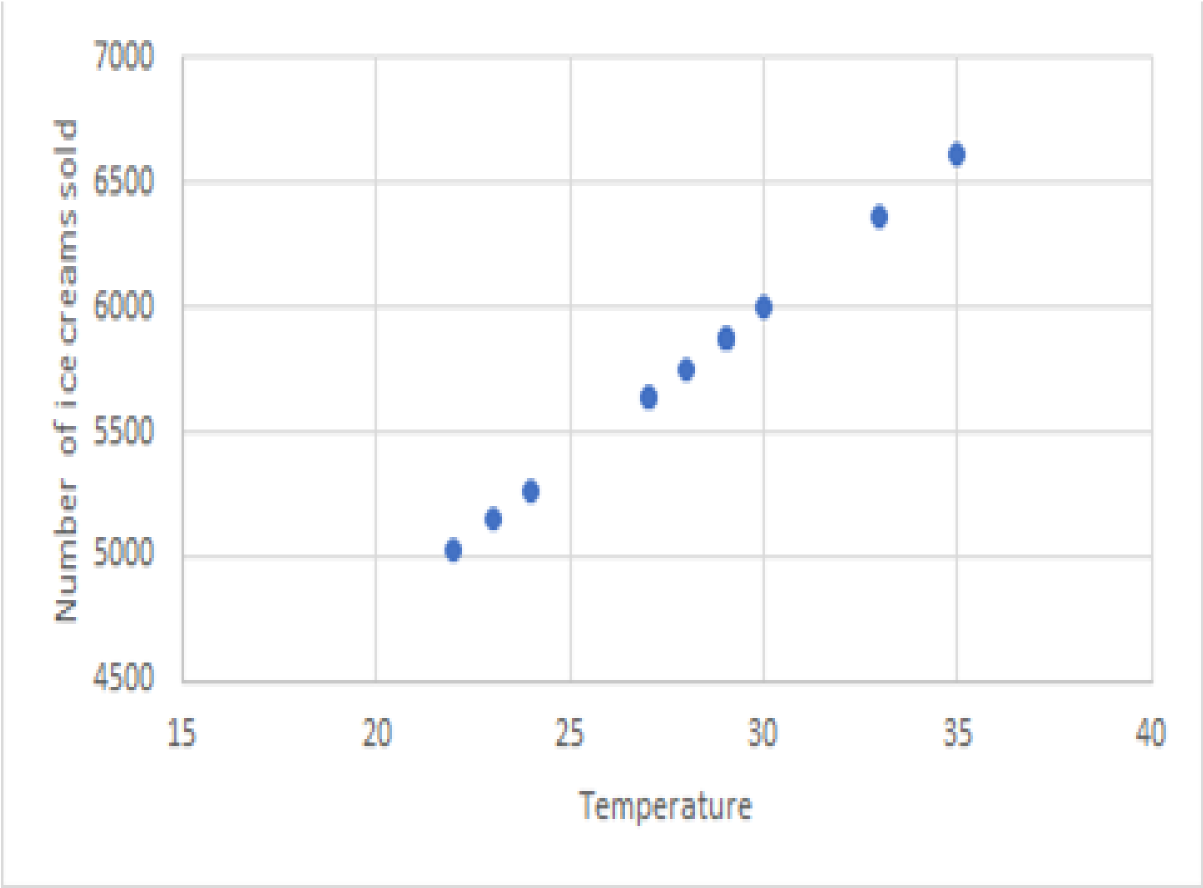


R-squared is a measure to indicate how much of the variance of y is explained by all x's. Closer to 1.0, better the model fit.

The **intercept coefficient** is β_0 in the multiple regression equation.

Other **coefficients** are β_i in the multiple regression equation.

Linear Regression Model



Standard error is a deviation from actual and the line of best fit line values.

P-value gives the significance of the feature on the dependent variable.

Linear Regression Model

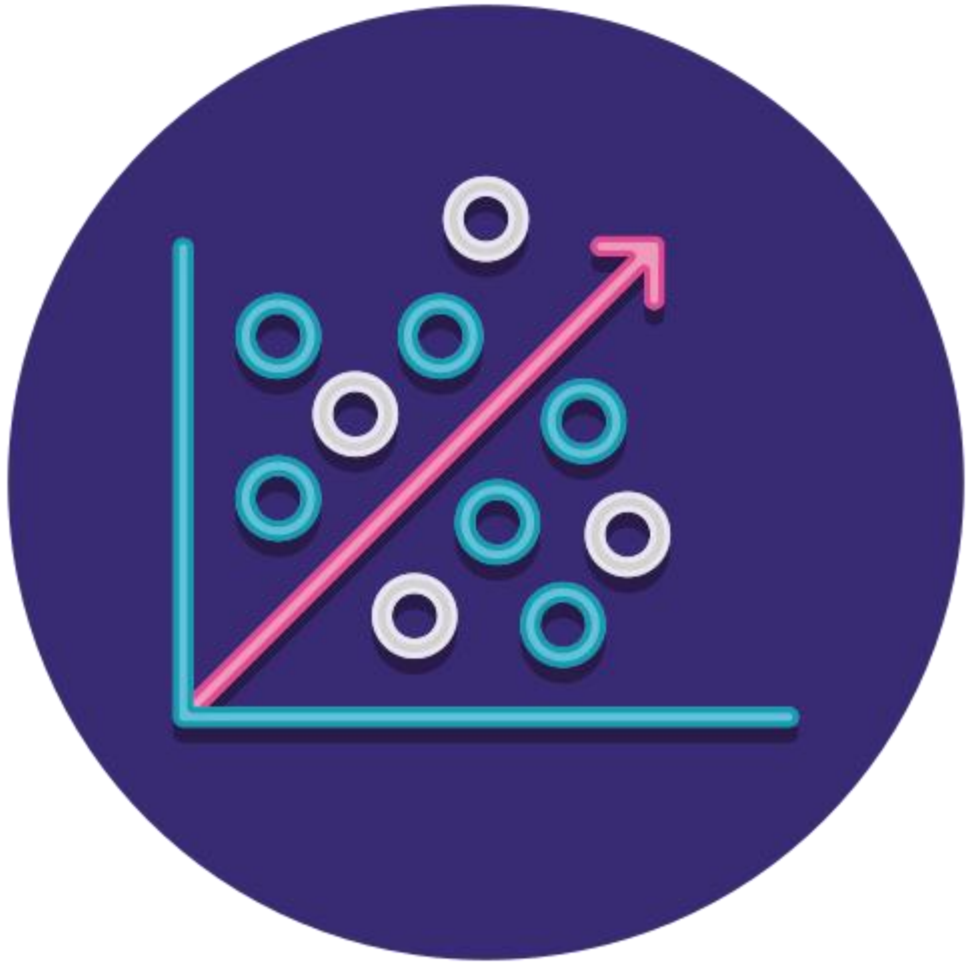
From the results it is understood that:

- The most and least important features determine the median price of the house.
- The value of y can be determined by using the equation with a new set of x values.

Logistic Regression

Logistic Regression

It is an algorithm for classification problems.



Though the name has the word regression, it is not a regression algorithm.

Logistic Regression

We have seen the following equation in linear regression:

$$y=\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + e$$

This equation cannot be used because:

- The value of y is not in **In odds** value
- The dependent variable y represents classes
- ***y*** is no more a continuous variable unlike regression
- log(ODDS) instead can help to arrive at a similar equation

Logistic Regression

Linear regression equation can be reused for logistic regression.

- By converting the y value in the classification problem to an ‘In odds’ value of the event
- $In(odds(E)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i + e$

Odds of Event

Odds of event (E) is defined as the probability of E happening divided by the probability of E not happening.

$$odds(E) = P(E)/1-P(E)$$

- The result of odds(E) is then converted to categorical values.
- Example: If y<= 0.5, then it is negative, or else it is positive.

Sigmoid Equation

If we solve for P(E) using the two odds equations, we get:

- $P(E) = 1/1+e^{-(\beta 0 + \beta 1x1+ \beta 2x2 + \dots + \beta ixi + e)}$
- The equation in this form is called the sigmoid equation.
- Example: If you take a numeric value of Y, it converts it into a probability value between 0 and 1.

Logistic Regression in Excel

To perform logistic regression in Excel, multiple regression equation is used which is created by using Data Analysis add-ins.

- It forms the equation of P(E), and
- Segregates the target values based on P(E)

Logistic Regression in Excel

When a new data is given to the model, the P(E) is calculated, and the target value is derived.

Steps to Derive Target Value

These are the steps to derive target values.

Step 1: Data items are encoded to numeric values

Steps to Derive Target Value

Step 2: The target values are encoded to numeric values

Steps to Derive Target Value

Step 3: Use add-ins of Data Analysis, to calculate the intercept and coefficients

Steps to Derive Target Value

Step 4: The linear regression equation arrives for each data row. This equation can be called **y**.

Steps to Derive Target Value

Step 5: P(E) is calculated as 1/(1+e^{-y})

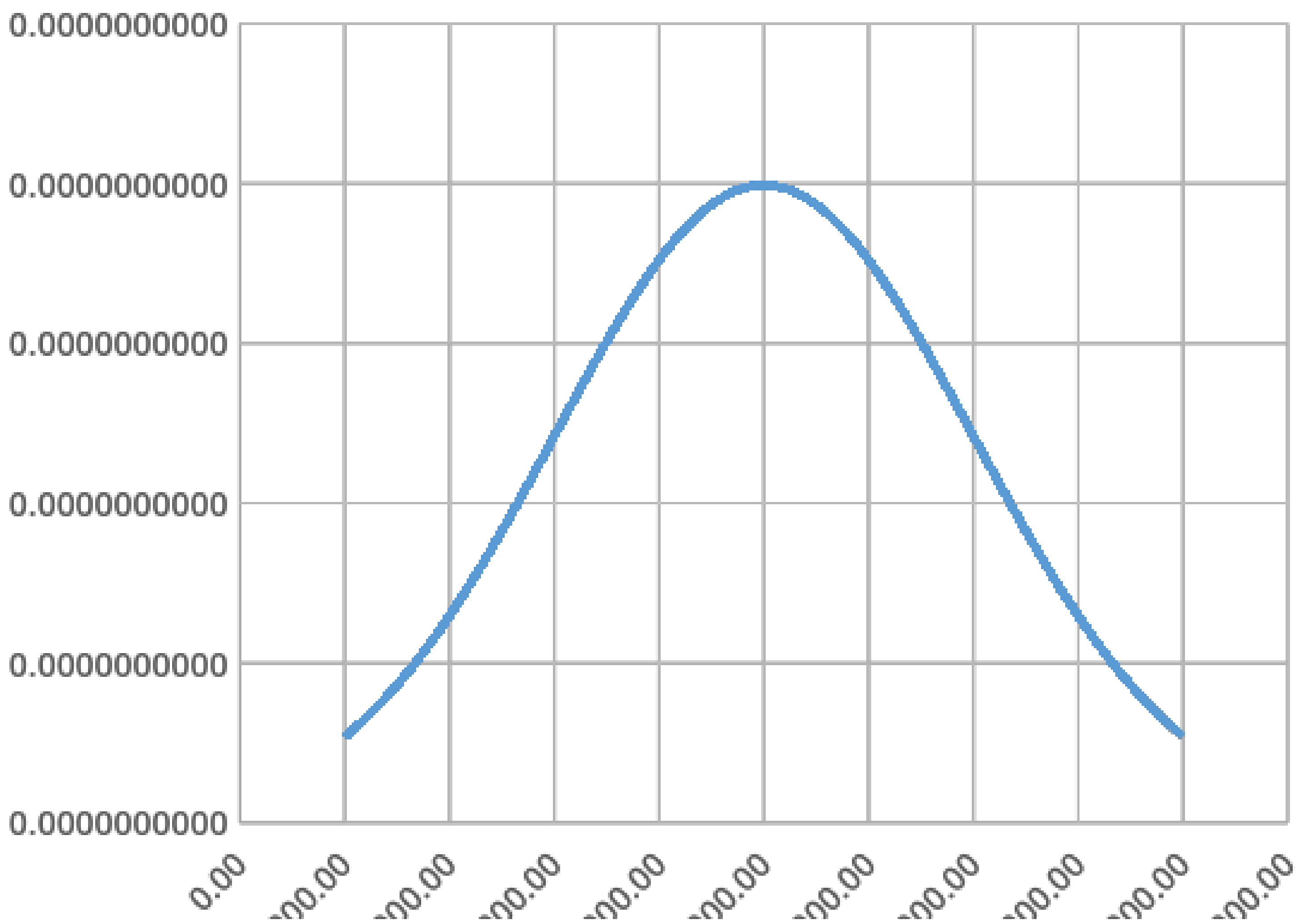
Steps to Derive Target Value

Step 6: A rule is applied on P(E) to get the target values

Normal Distribution

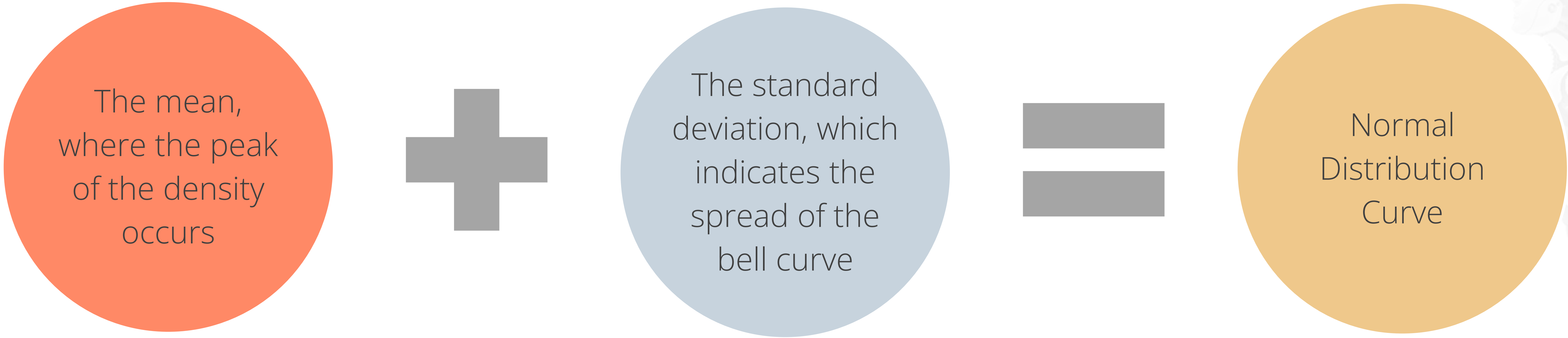
Normal Distribution: Introduction

All normal distributions are symmetric and have bell-shaped curves with a single peak.



Create Normal Distribution

Normal distribution helps find the probability distribution for various variables such as rainfall, height, weight, manufacturing error, weight error, and test scores.



The mean, where the peak of the density occurs

The standard deviation, which indicates the spread of the bell curve

Normal Distribution Curve

Normal Distribution: Empirical Rule

All normal density curves satisfy the Empirical Rule or (68-95-99.7% Rule) in Statistics.

68% of the observations fall within 1 standard deviation of the mean, i.e. between Mean – Standard Deviation and Mean + Standard Deviation.

95% of the observations fall within 2 standard deviations of the mean, i.e. between Mean – 2*Standard Deviation and Mean + 2*Standard Deviation.

99.7% of the observations fall within 3 standard deviations of the mean, i.e. between Mean – 3*Standard Deviation and Mean + 3*Standard Deviation.

Assisted Practice: Create Normal Distribution graph

Problem statement:

Demonstrate how to create a Normal Distribution graph in Excel.

Assisted Practice Guidelines

Steps to follow:

Step 1: Open the Excel file

Step 2: Create Normal Distribution

Key Takeaways

- A Moving Average evaluates data points by creating a series of averages of different subsets of the complete dataset.

- The Hypothesis Testing is used to test the null hypothesis.

- ANOVA is a collection of statistical methods used to compare the means of different groups.

- Covariance determines the relationship between two random variables— how they change together.

Key Takeaways

- Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together
- Regression is a statistical measure that determines the strength of the relationship between one dependent variable and a series of other changing variables.
- All Normal Distributions are symmetric and have bell-shaped curves with a single peak.

Knowledge Check

Knowledge Check

1

Which of the following statistical methods is used to analyze variance between more than two groups?

A. Hypothesis Testing

B. Histogram

C. ANOVA

D. Covariance

Knowledge Check

1

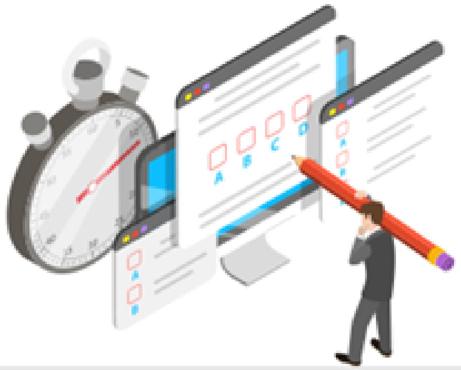
Which of the following statistical methods is used to analyze variance between more than two groups?

A. Hypothesis Testing

B. Histogram

C. ANOVA

D. Covariance



The correct answer is **C**

ANOVA is used to analyze variance between more than two groups.

Knowledge Check

2

What conclusion will you derive for the Null Hypothesis if “F > F crit” in ANOVA testing?

A. The Null Hypothesis is not rejected

B. The Null Hypothesis is rejected

C. There is no relationship with Hypothesis Testing

D. None of the above is correct

Knowledge Check

2

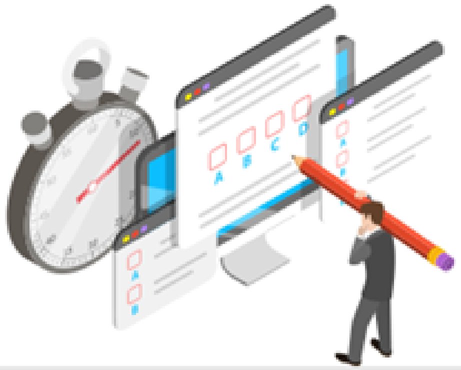
What conclusion will you derive for the Null Hypothesis if “F > F crit” in ANOVA testing?

A. The Null Hypothesis is not rejected

B. The Null Hypothesis is rejected

C. There is no relationship with Hypothesis Testing

D. None of the above is correct



The correct answer is **B**

In ANOVA testing if “F > F crit,” then the Null Hypothesis is rejected.

Knowledge Check

3

The Null Hypothesis means that the mean/average of two populations is equal.

A. True

B. False

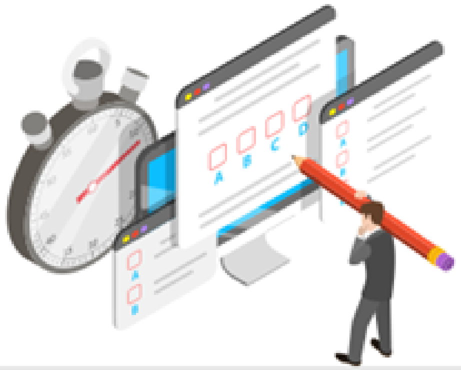
Knowledge Check

3

The Null Hypothesis means that the mean/average of two populations is equal.

A. True

B. False



The correct answer is **A**

The Null Hypothesis(H0) means that the mean/average of two populations is equal.

Knowledge Check

4

Which of the following is indicated if the Correlation Coefficient value is +1?

A. Perfect Positive Correlation

B. Zero Correlation

C. Perfect Negative Correlation

D. No Correlation

Knowledge Check

4

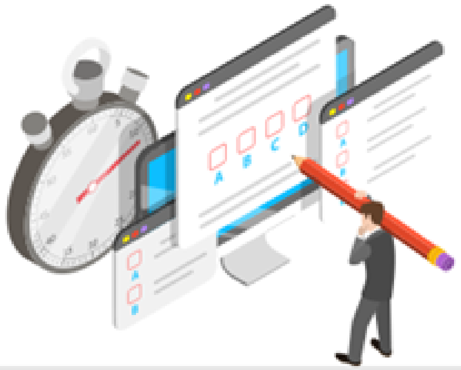
Which of the following is indicated if the Correlation Coefficient value is +1?

A. Perfect Positive Correlation

B. Zero Correlation

C. Perfect Negative Correlation

D. No Correlation



The correct answer is **A**

The Correlation Coefficient value of +1 indicates Perfect Positive Correlation.

Knowledge Check

5

Which statistical measure determines the strength between a dependent variable and an independent variable?

A. Histogram

B. Hypothesis Testing

C. Moving Average

D. Regression

Knowledge Check

5

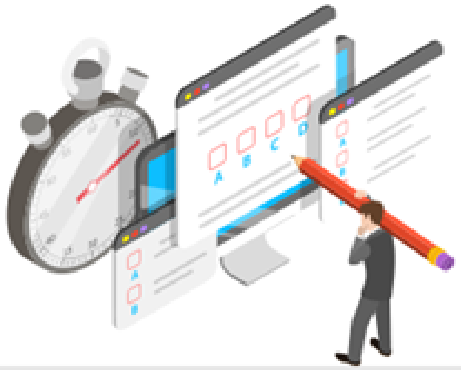
Which statistical measure determines the strength between a dependent variable and an independent variable?

A. Histogram

B. Hypothesis Testing

C. Moving Average

D. Regression



The correct answer is **D**

Regression determines the strength between a dependent variable and an independent variable.

Knowledge Check
6

What are the mandatory fields required while creating a Normal Distribution curve?

A. Mean and Standard Deviation

B. Mean and Maximum value

C. Maximum and Minimum value

D. Standard Deviation and Minimum Value

Knowledge Check

6

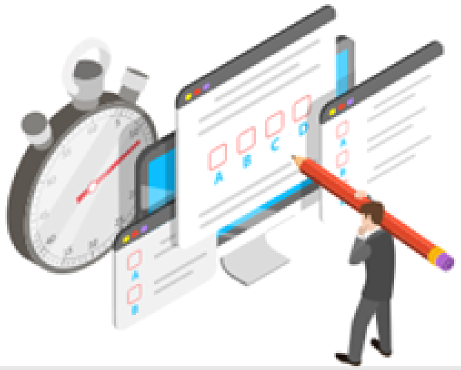
What are the mandatory fields required while creating a Normal Distribution curve?

A. Mean and Standard Deviation

B. Mean and Maximum value

C. Maximum and Minimum value

D. Standard Deviation and Minimum Value



The correct answer is **A**

To create Normal Distribution curve, we need to specify two quantities: the mean, where the peak of the density occurs, and the standard deviation, which indicates the spread of the bell curve.