

DATA AND ARTIFICIAL INTELLIGENCE



Business Analytics with Excel



Using Macros for Analytics

Learning Objectives

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

- 🕒 Create Macros and functions
- 🕒 Examine mean of data using Macros
- 🕒 Describe the five point summary using Macros
- 🕒 Identify how to remove duplicates using Macros



A Day in the Life of Business Analyst

As a business analyst of an organization:

You are required to do few tasks in Microsoft Excel which are to be done repeatedly.

Also, you need to create and then run a macro that quickly applies these formatting changes to the cells that needs to be selected.

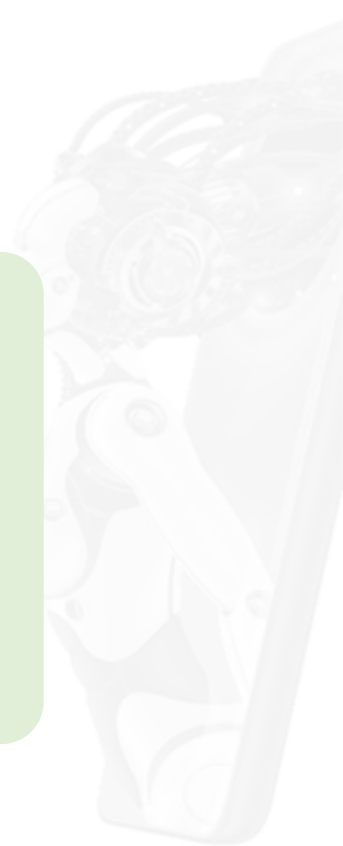
To achieve these tasks, you will be learning a few concepts, such as macros for analytics, means of data using macros, correlation coefficient and removing duplicates using macros



Using Macros for Analytics

Using Macros for Analytics

We use functions within Excel to perform data analysis, charting, and predictive analytics.



Using Macros for Analytics

Macros is an important feature in Excel which permits to do VBA programming within Excel workbook.



Source: https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.k2e.com%2Fseminars%2Fexcel-macros%2F&psig=AOvVaw2h6kc_fd2sSQnjL8L12diF&ust=1635577130758000&source=images&cd=vfe&ved=0CAsQjRxqFwoTCMj4qeqF7_MCFQAAAAAdAAAAABAD

VBA

Visual Basic for Applications (VBA) allows a programmable interface to Excel.



Source: https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.reddit.com%2Fr%2Fvba%2F&psig=AOvVaw0cti-XNoCvPvHQYrG3jy&ust=1635577238890000&source=images&cd=vfe&ved=0CAsQjRxqFwoTCJj-w5qG7_MCFQAAAAAdAAAAABAK

Create Macros and Functions

Macros and Functions are created to:

Perform operations
on the data



Validate data



Fix missing
values



Perform predictive
analysis



Types of Macros

There are 3 types of Macros:

- Event based
- Subroutine/Sub Procedure
- Functions



Event Based

This is based on a macro event. For instance, whenever the worksheet is activated, a message box is printed by the below macro. Example:

```
Private Sub Worksheet_Activate()  
    MsgBox ("Worksheet activated!")  
End Sub
```



Subroutine

This is a set of commands that does some processing in the worksheet and does not return any value.

Example:

```
Sub remove_duplicates()  
    Range("T1:T451").RemoveDuplicates Columns:=1, Header:=xlYes  
End Sub
```


Functions

They are similar to sub procedures, but they return some value to the calling sub procedure.
Example:

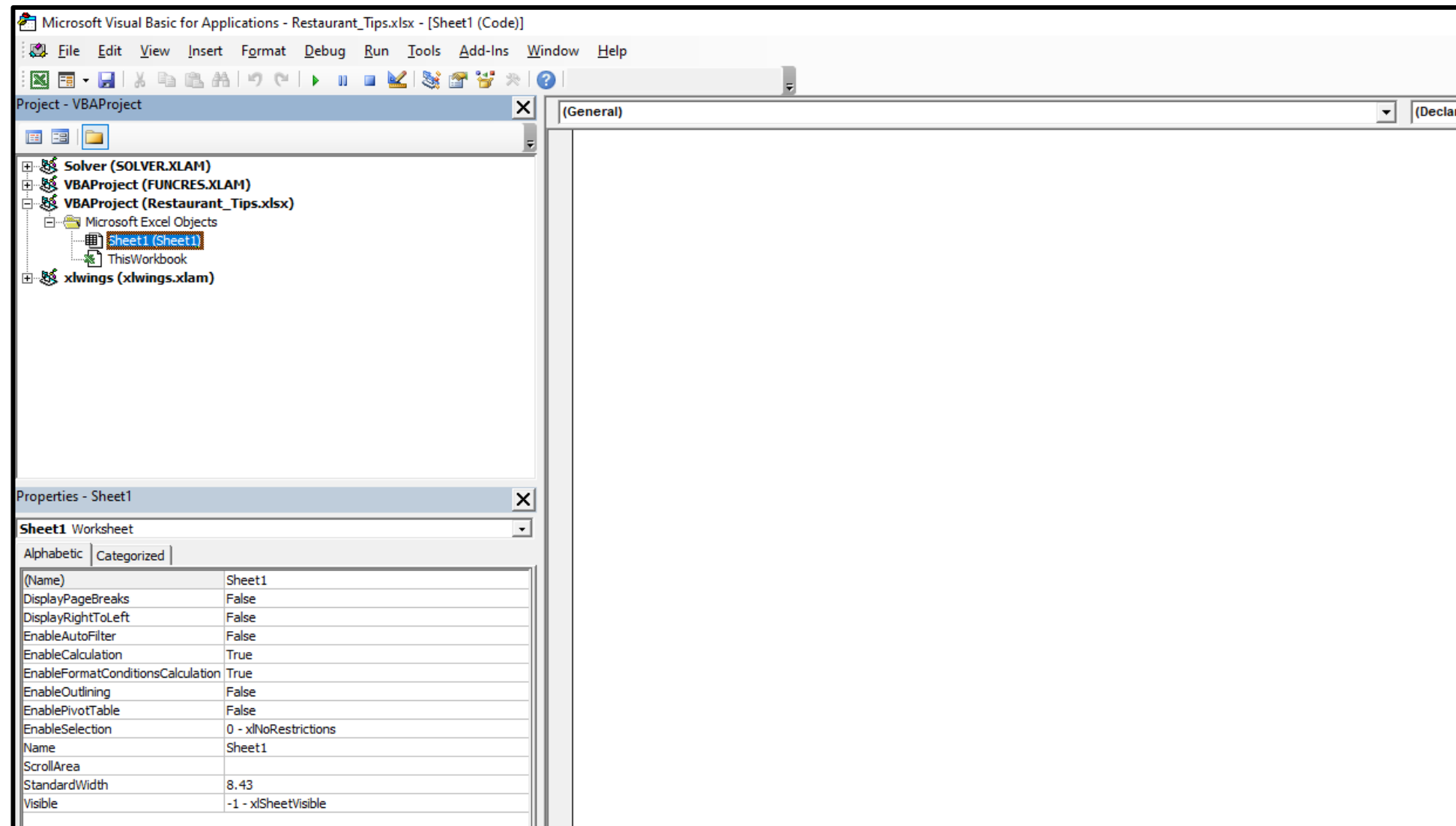
```
Public Function sqr(x)  
    sqr = x * x  
End Function
```

- To call this function we use: msgbox(sqr(9))
- This prints 81 in a message box.



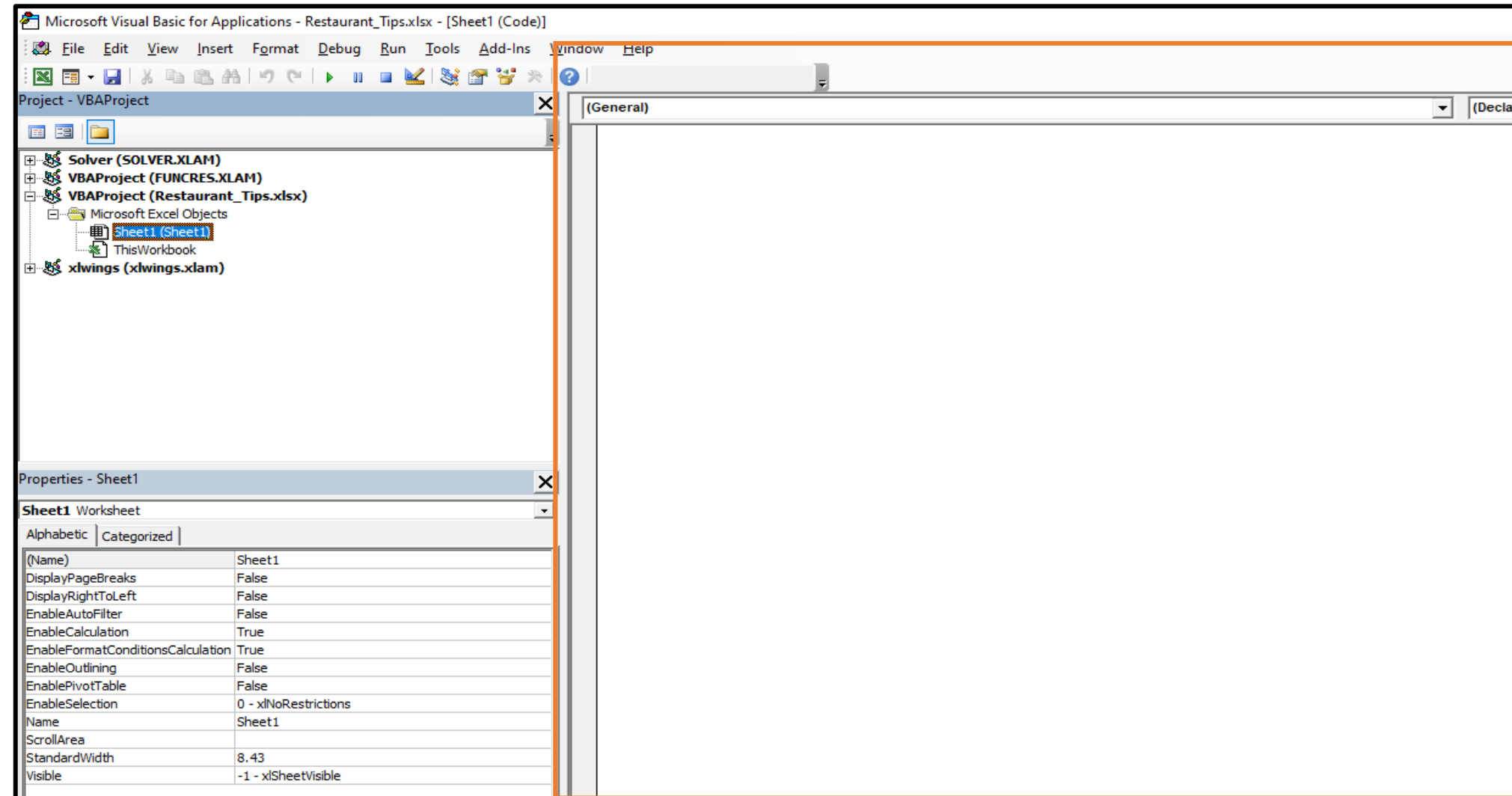
Create Macros and Functions

To create a macro, press **control+F11** on the sheet with the data



Create Macros and Functions

The white area on the right side can be used to create all the macro functions on the data.



Create Macros and Functions

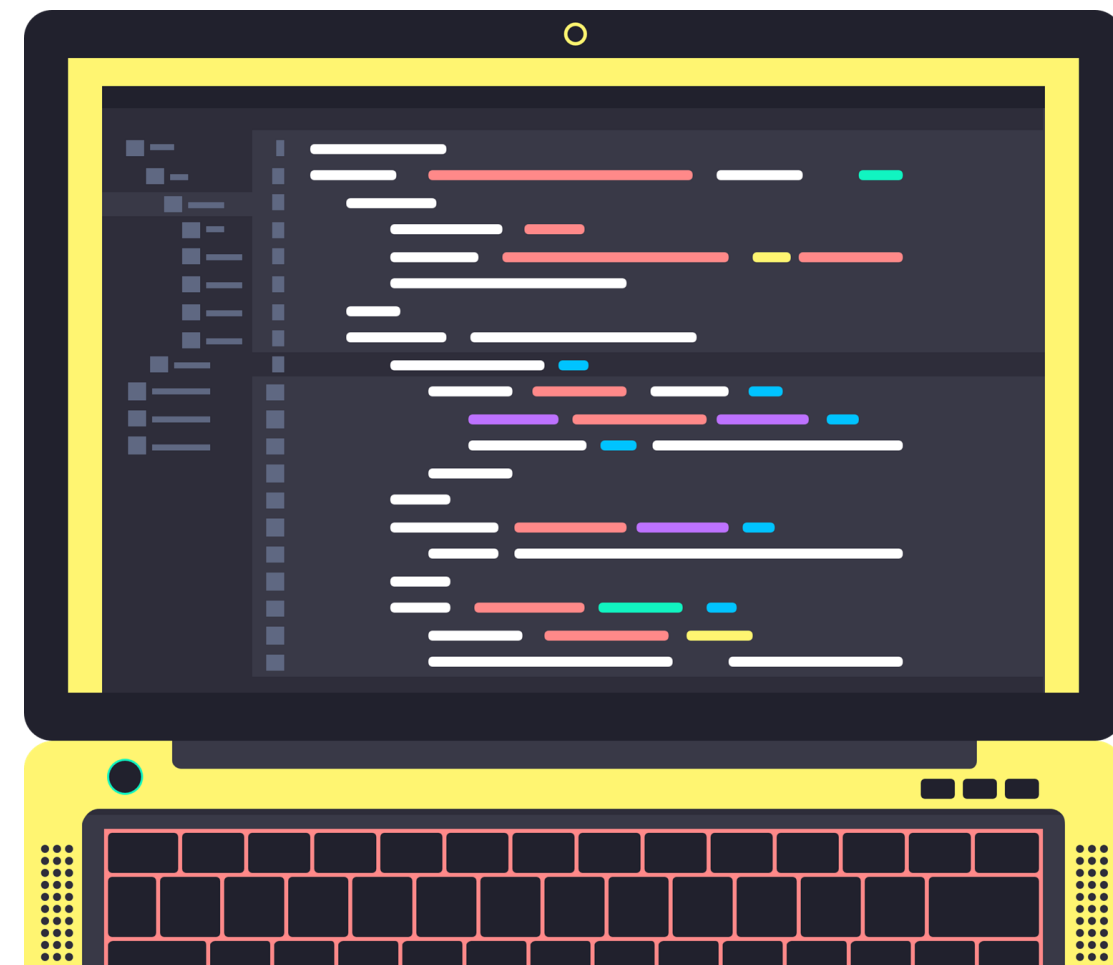
A function is usually written as:

```
Sub function_name()
```

```
...
```

```
...
```

```
End Sub
```



Create Macros and Functions

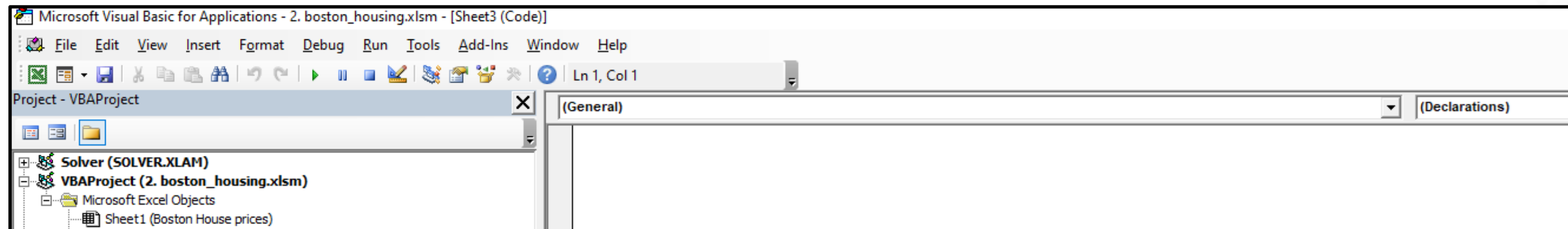
Regular VB programming can be done within the function.



A function can be based on any event done on the worksheet.

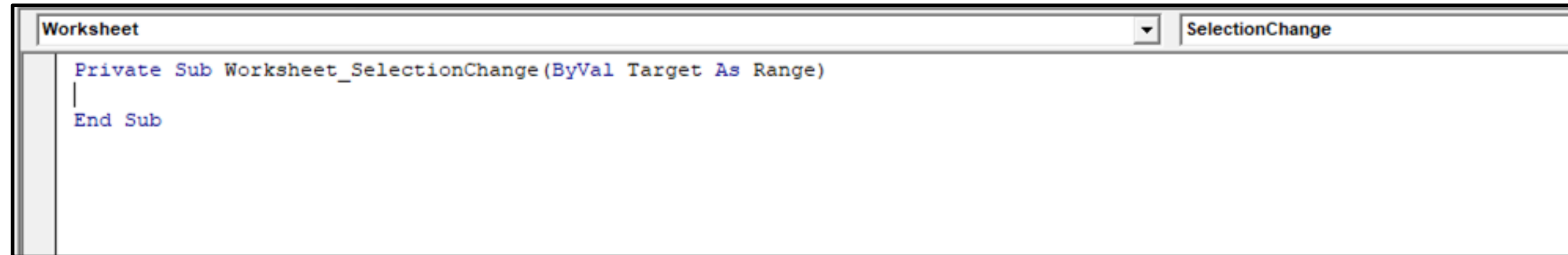
Create Macros and Functions

To choose an event to work on, choose Worksheet on the first drop down and then activate on the second drop down



Create Macros and Functions

This function will create a set of commands whenever the worksheet is activated.



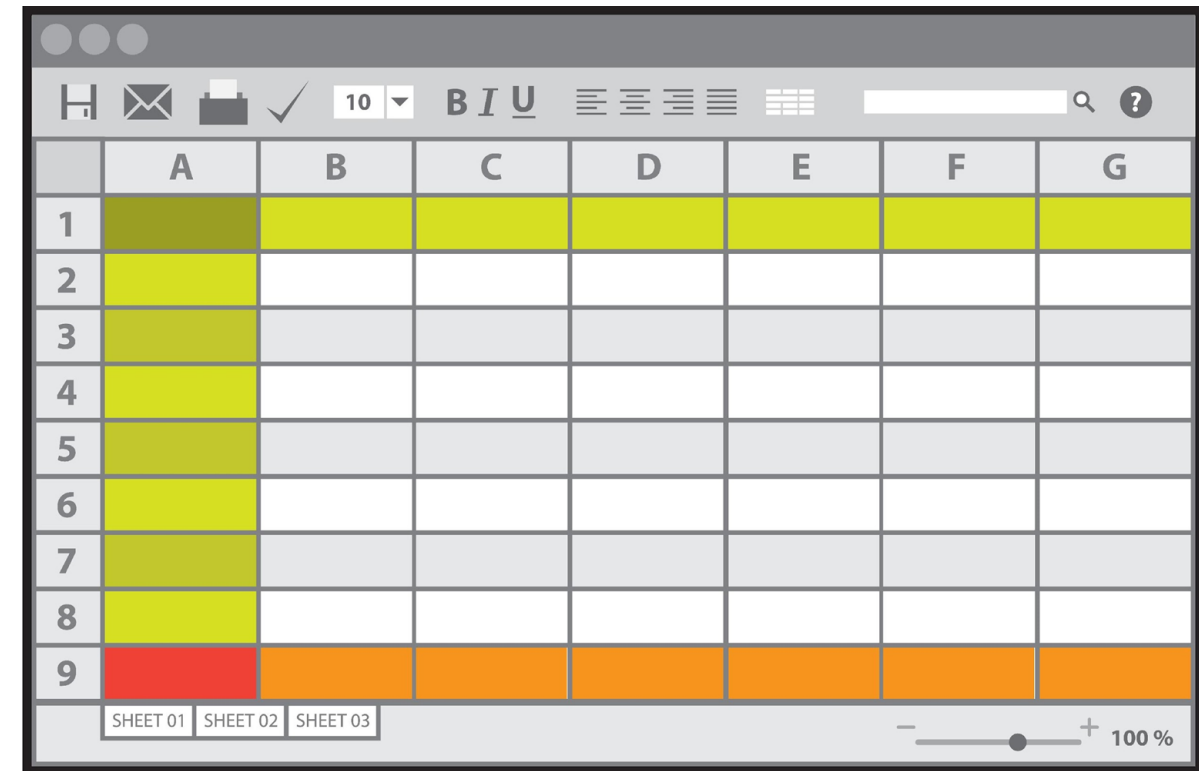
The screenshot shows the VBA editor window. The 'Worksheet' dropdown is selected, and the 'SelectionChange' event is chosen from the list. The code editor contains the following VBA code:

```
Private Sub Worksheet_SelectionChange(ByVal Target As Range)
|
End Sub
```

Create Macros and Functions

A cell value within the sheet can be accessed using the cells.

- Row starts with 1
- Column starts with 1



Mean of Data Using Macros

Mean

Mean is defined as the sum of values in a data set divided by the number of values in the data set.

Σ



Mean

We can find the mean of a column of values using macros for Excel.



Steps to Find Mean

Step 1: Open the boston_housing.xlsx file

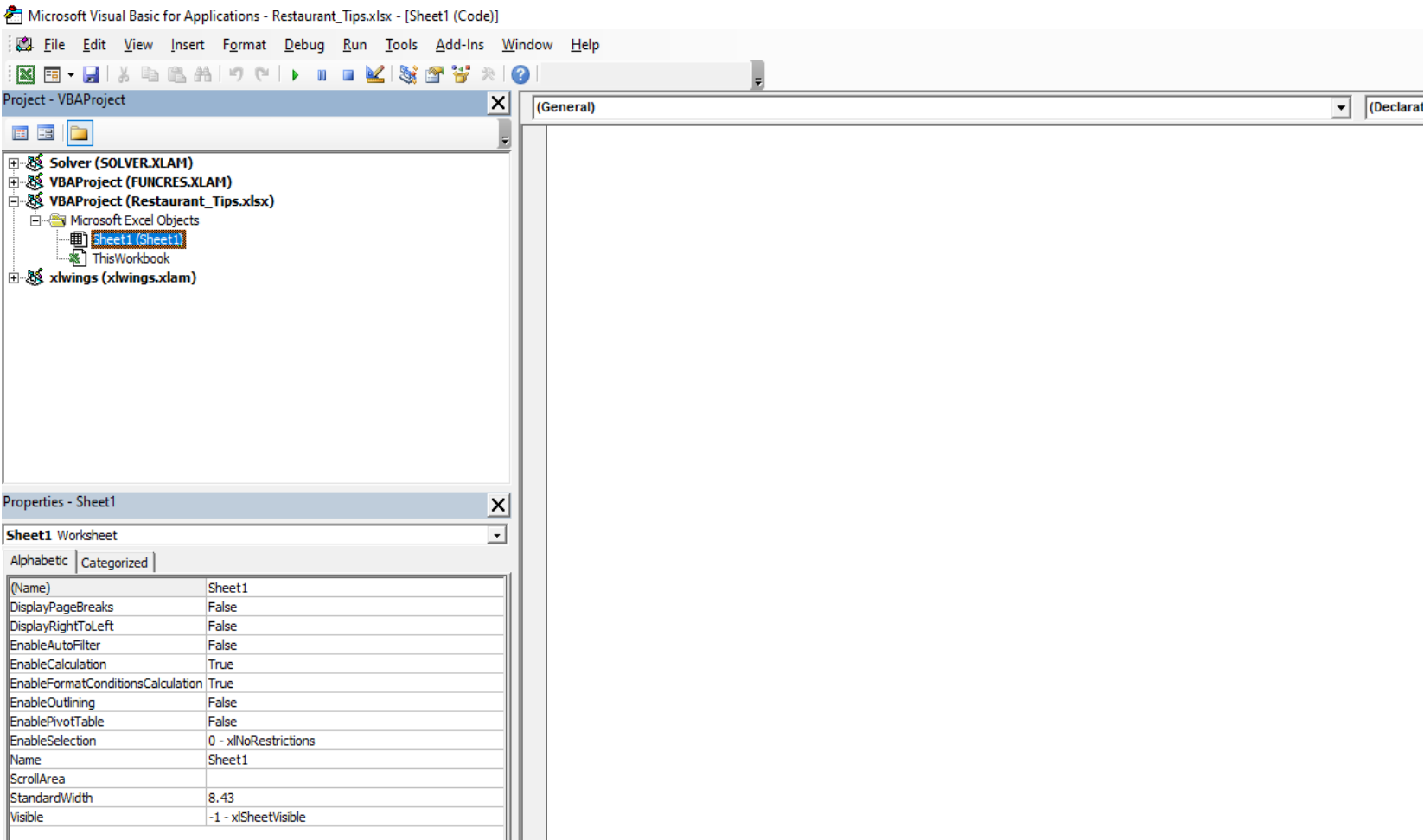


boston_housing



Steps to Find Mean

Step 2: Press **Alt+F11** to open the macro editor



Steps to Find Mean

Step 3: Create a macro with the below code:

```
(General)

Sub calculate_mean()

    n = 451

    Dim x(451)

    For c = 1 To 14
        For i = 1 To n
            x(i) = Cells(i + 1, c)
        Next i

        sum_x = 0

        For i = 1 To n
            sum_x = sum_x + x(i)
        Next i

        mean_x = sum_x / 451

        Cells(452, c) = mean_x
    Next c
End Sub
```



Steps to Find Mean

Step 4: Run the macro using the run button



Steps to Find Mean

The means are populated in row 452 for all 14 columns.

451	0.10959	0	11.93	0	0.573	6.794	89.3	2.3889	1	273	21	393.45	6.48	22	
452	1.404807	12.74945	10.28761	0.077605	0.540478	6.343871	65.498	4.048565	7.78714	376.8027	18.24279	369.7665	11.43361	23.77251	Macro calculated mean
453	1.407824	12.77778	10.28396	0.077778	0.540406	6.344569	65.464	4.051995	7.802222	377.0333	18.23667	369.7062	11.44151	23.79889	Mean function

Step 5: We can check the values with the average() function.

Five Point Summary Using Macros

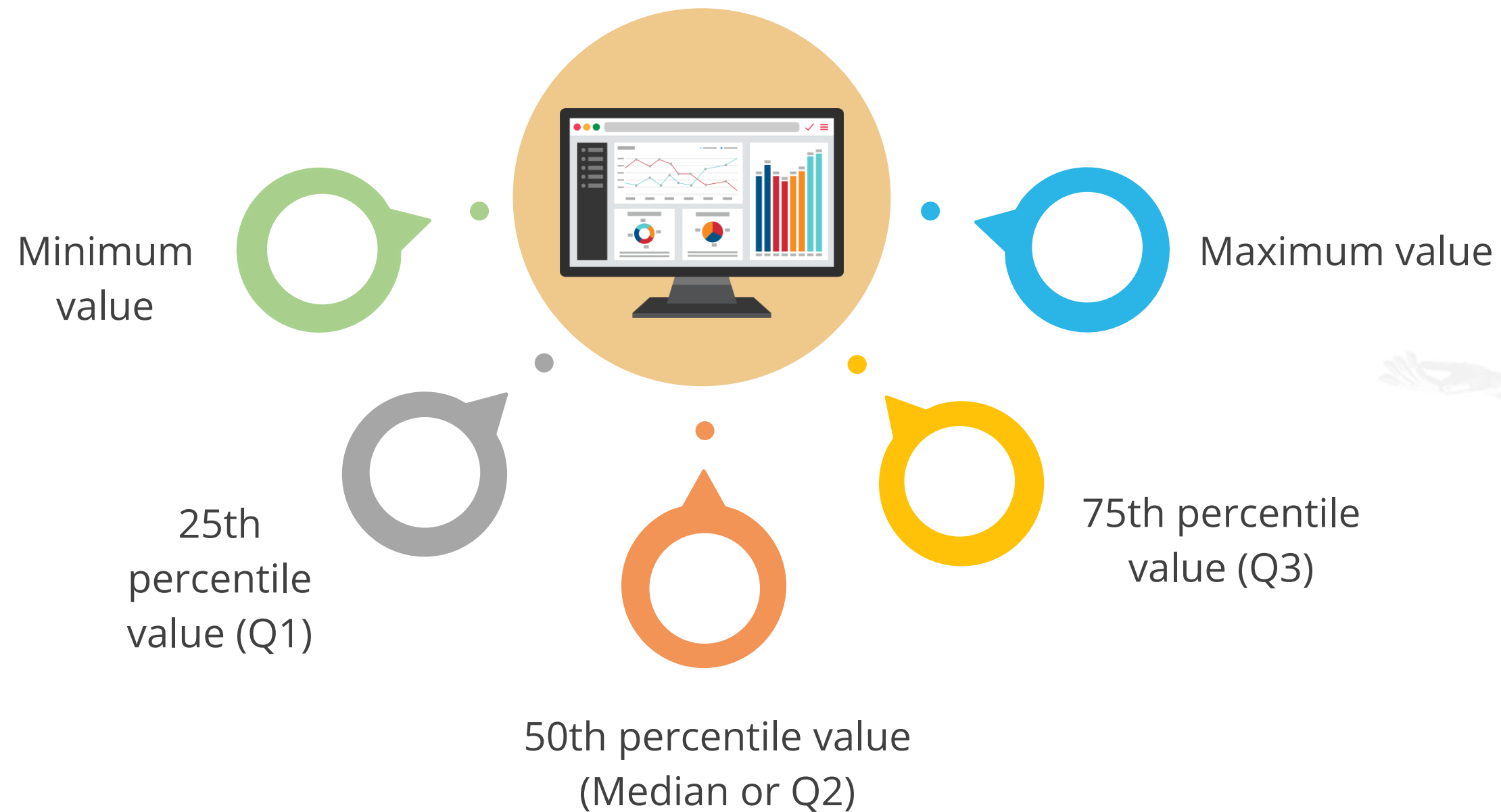
Five Point Summary

The five point summary in statistics specifies five values to describe a set of numeric values.



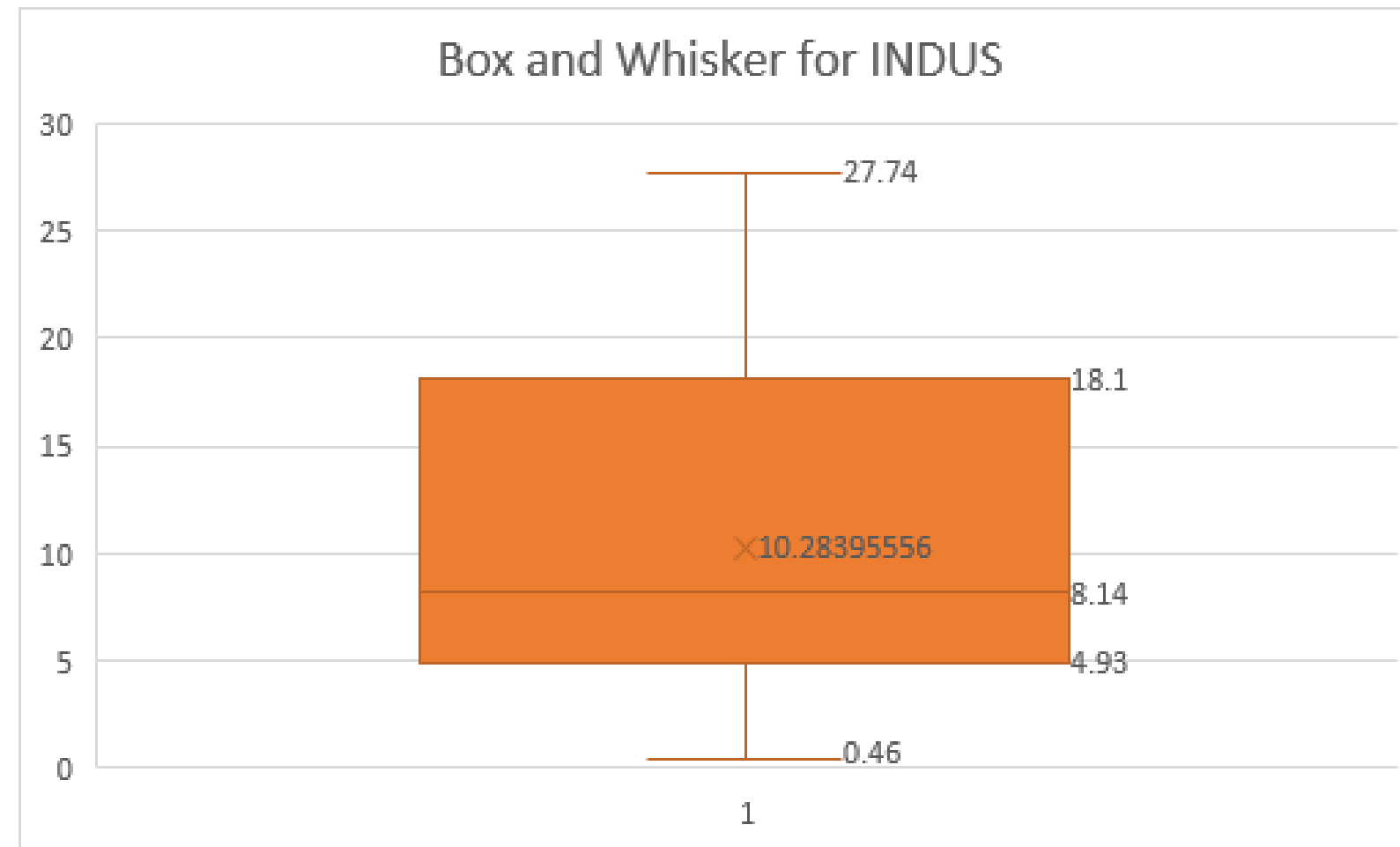
Values of Five Point Summary

The values are:



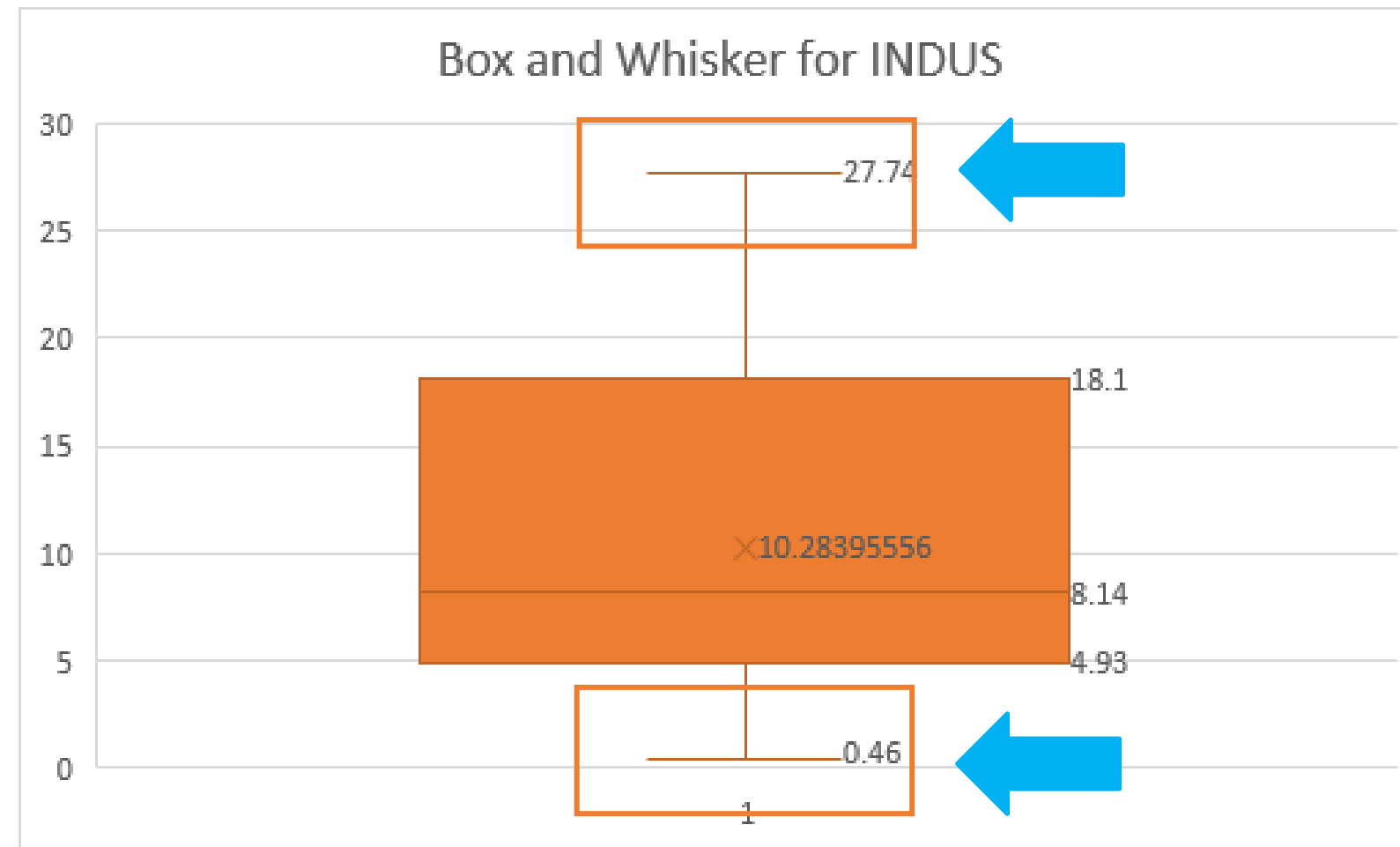
Values of Five Point Summary

The five point summary can be visualized using a box and whisker chart.



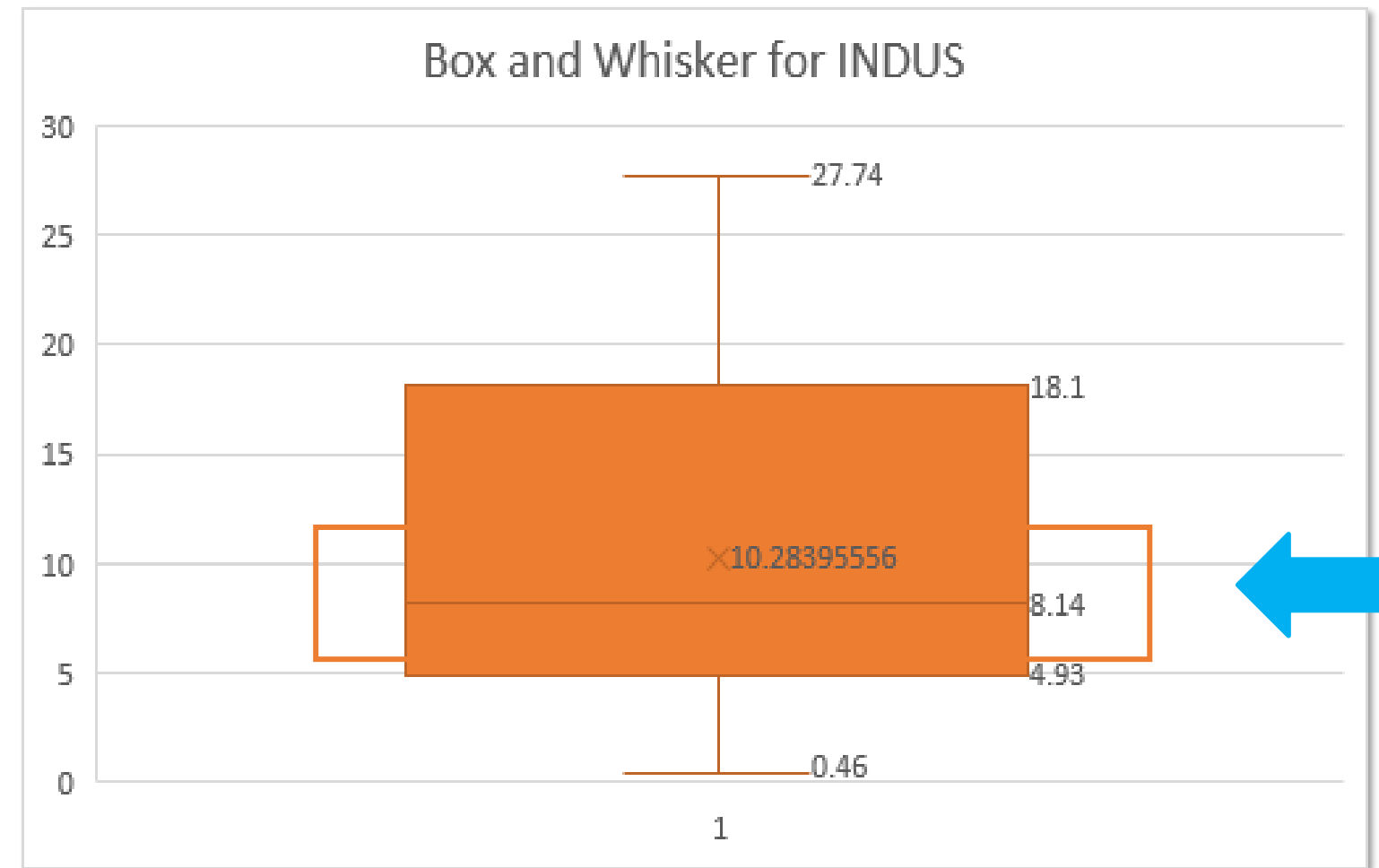
Values of Five Point Summary

The lowest point is minimum, and the topmost value is the maximum value.



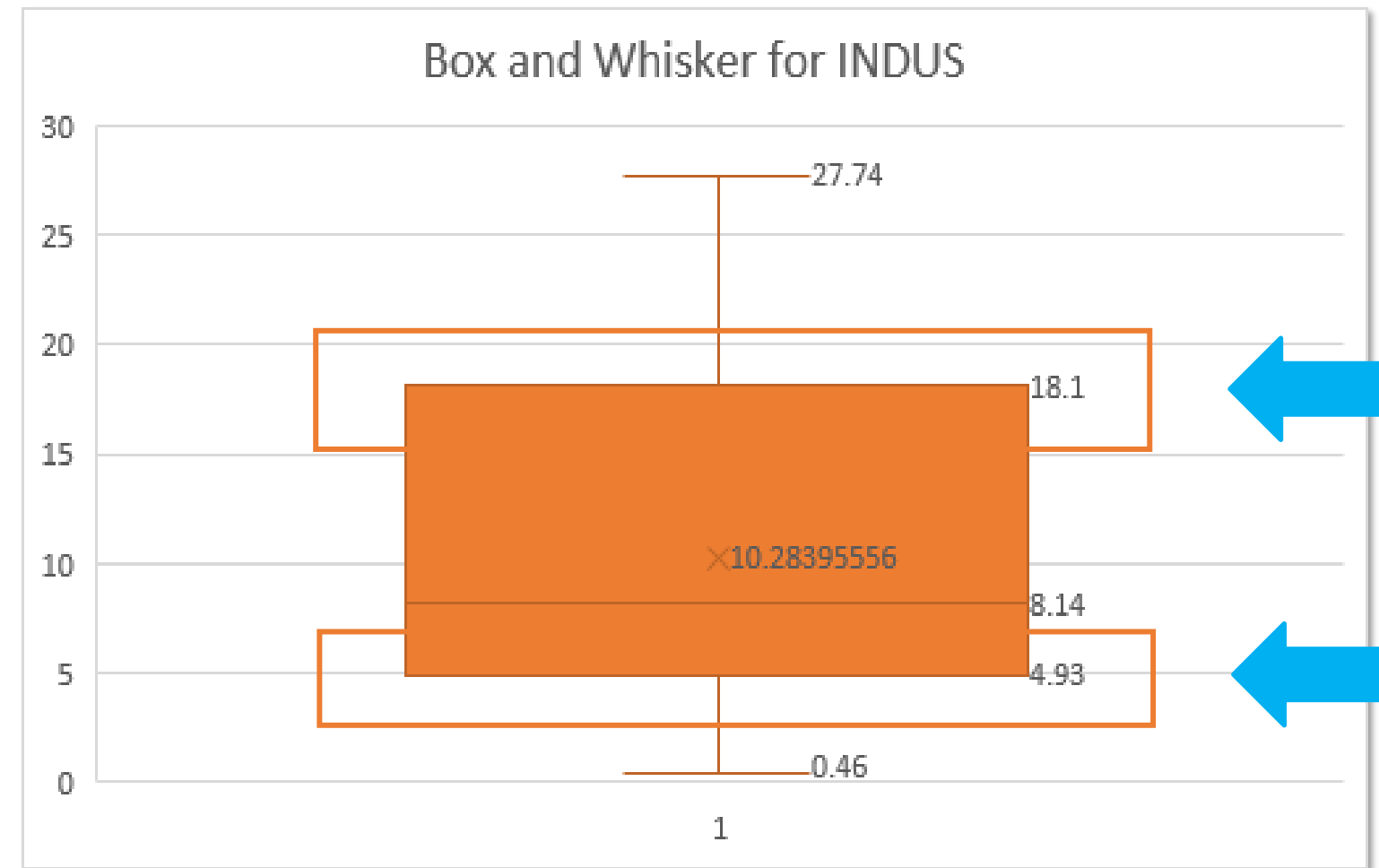
Values of Five Point Summary

The line within the box is Q2 (median).



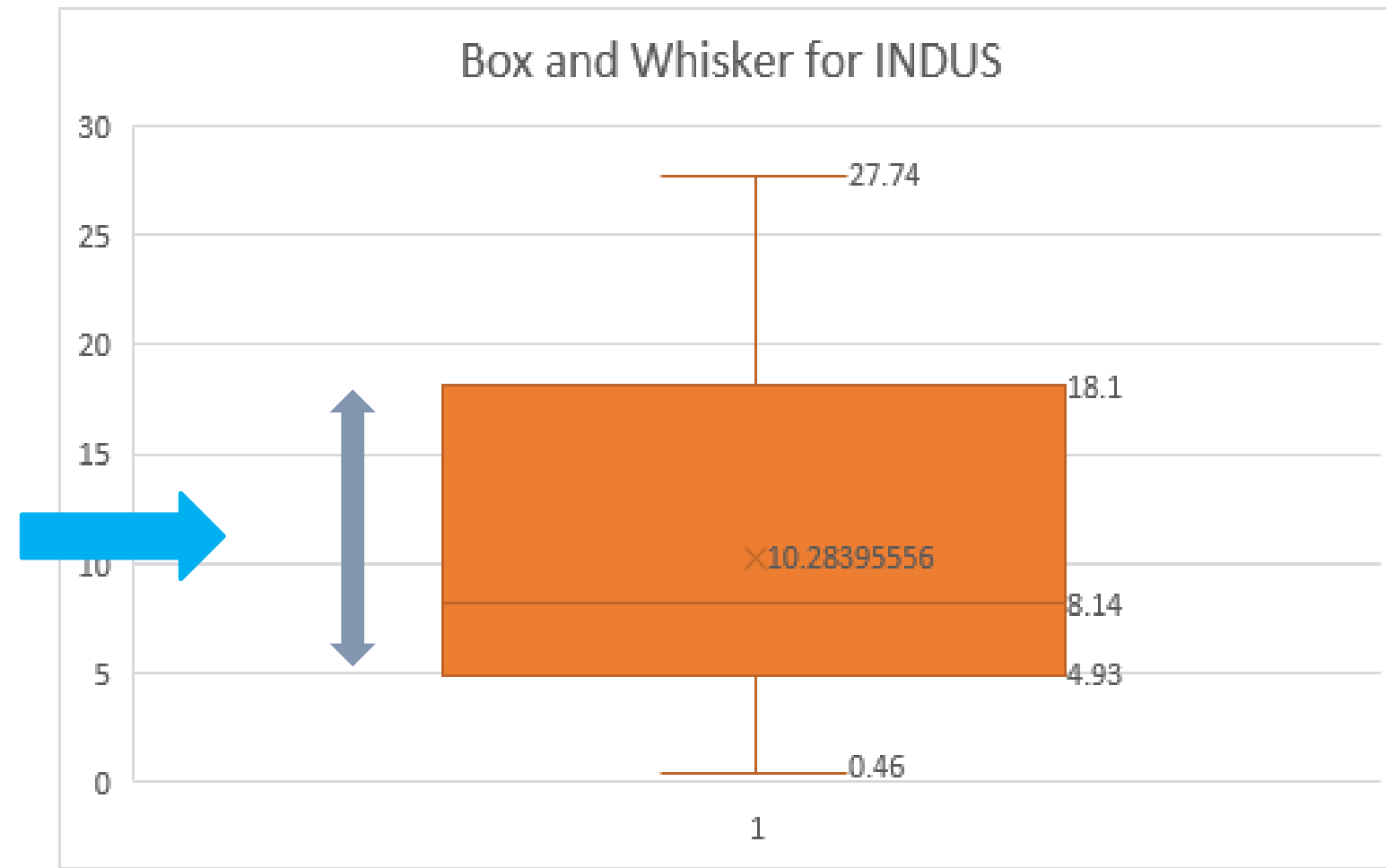
Values of Five Point Summary

Q1 and Q2 are the bottom and top boxes.



Interquartile Range

There is a metric called IQR (Interquartile Range) which is $Q3 - Q1$.



It refers to the height of the box.

Interquartile Range

IQR is used to find outliers in the data set.

Values of a variable more than $Q3 + 1.5 * IQR$ and less than $Q1 - 1.5 * IQR$ are considered as **suspected** outliers.



Calculate Five Point Summary

Press Alt+F11 on the Excel sheet where we have the data set of Boston_housing

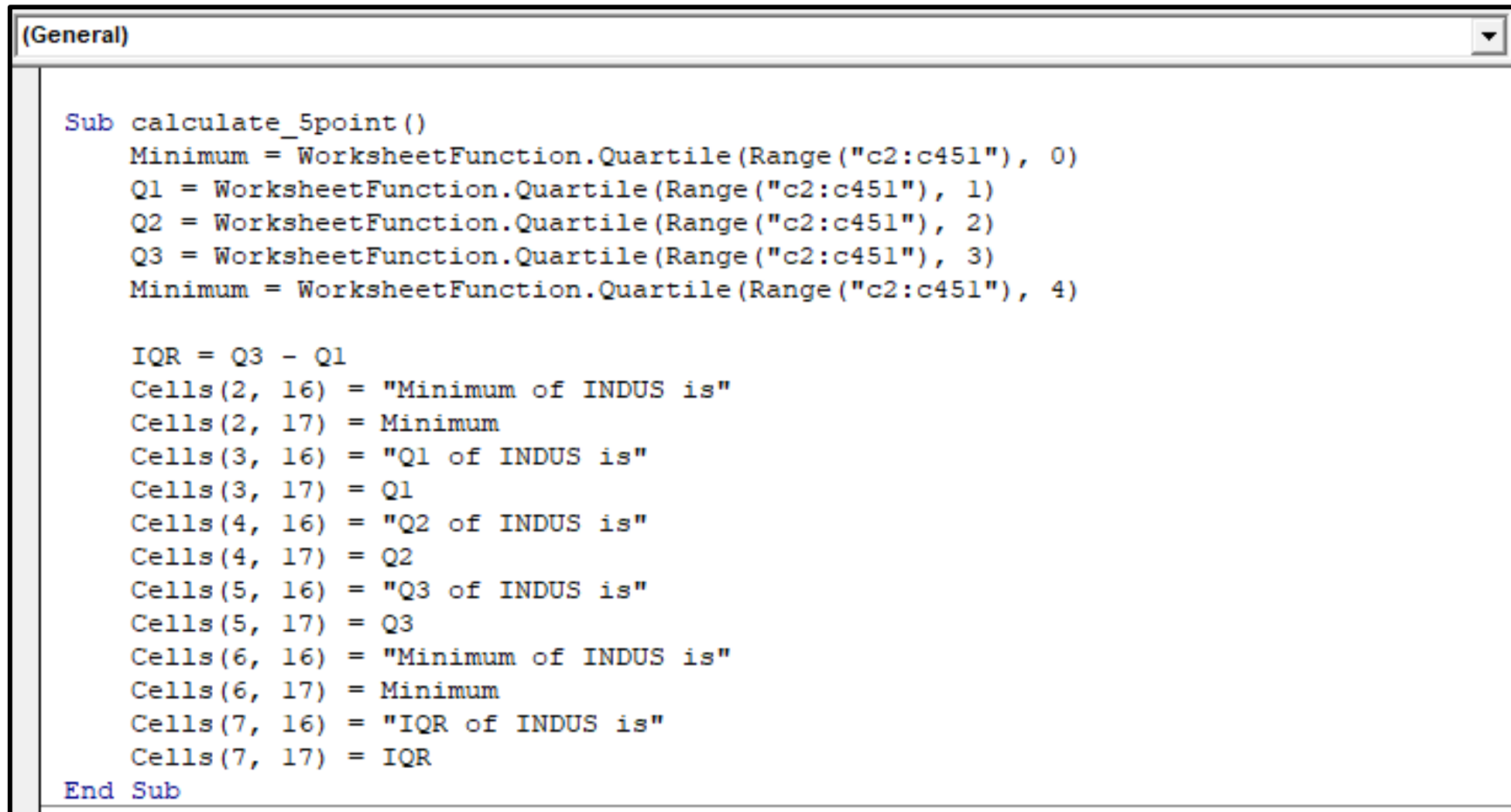


Boston_housing



Calculate Five Point Summary

Copy the following code in macros



```
(General)

Sub calculate_5point()
    Minimum = WorksheetFunction.Quartile(Range("c2:c451"), 0)
    Q1 = WorksheetFunction.Quartile(Range("c2:c451"), 1)
    Q2 = WorksheetFunction.Quartile(Range("c2:c451"), 2)
    Q3 = WorksheetFunction.Quartile(Range("c2:c451"), 3)
    Minimum = WorksheetFunction.Quartile(Range("c2:c451"), 4)

    IQR = Q3 - Q1
    Cells(2, 16) = "Minimum of INDUS is"
    Cells(2, 17) = Minimum
    Cells(3, 16) = "Q1 of INDUS is"
    Cells(3, 17) = Q1
    Cells(4, 16) = "Q2 of INDUS is"
    Cells(4, 17) = Q2
    Cells(5, 16) = "Q3 of INDUS is"
    Cells(5, 17) = Q3
    Cells(6, 16) = "Minimum of INDUS is"
    Cells(6, 17) = Minimum
    Cells(7, 16) = "IQR of INDUS is"
    Cells(7, 17) = IQR
End Sub
```

Calculate Five Point Summary

The function of quartile is used from the **WorksheetFunction** object.

- The function takes 0-4 values to each of the five point metrics.
- The values are stored in columns P and Q.
- The macro is executed.



Calculate Five Point Summary

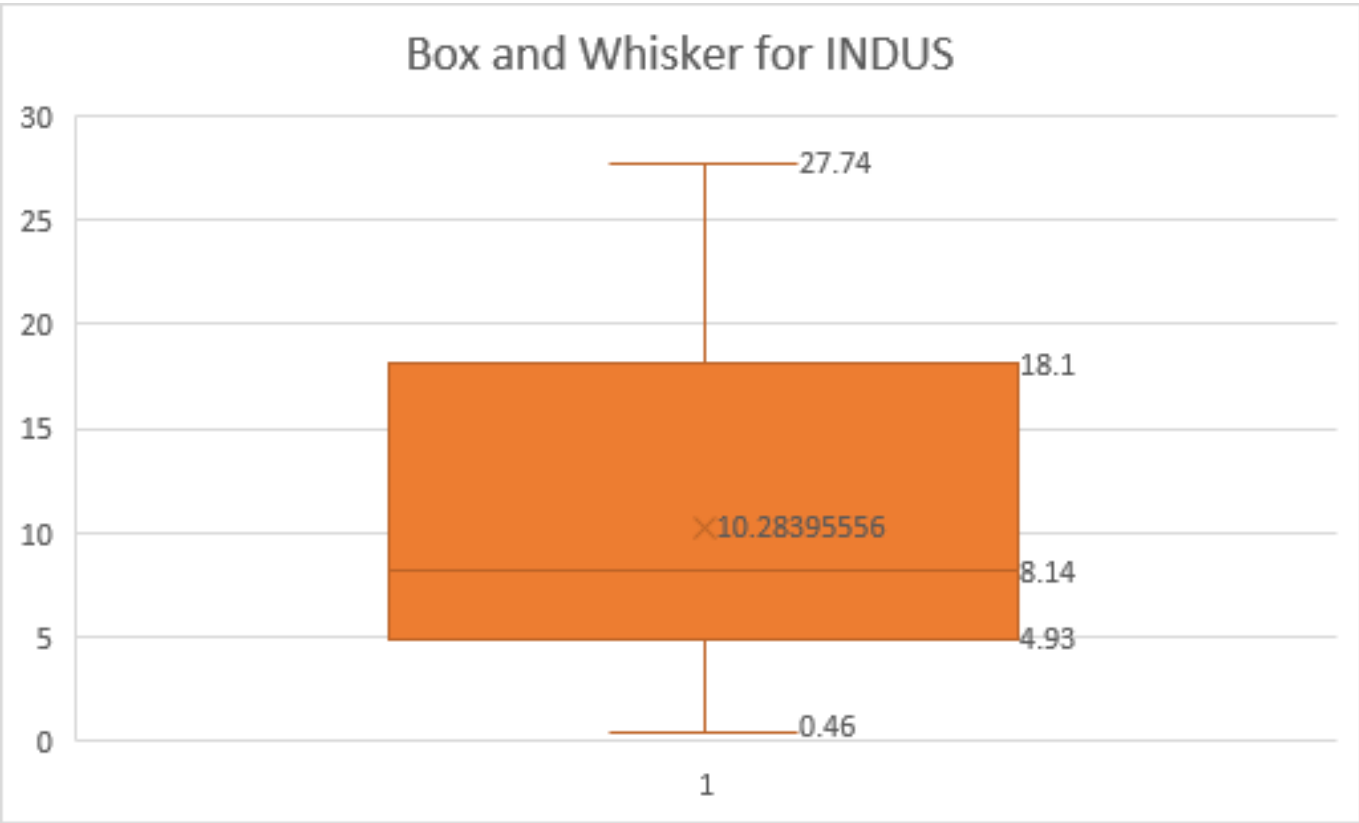
This will be the output.

Minimum of INDUS is	27.74
Q1 of INDUS is	4.93
Q2 of INDUS is	8.14
Q3 of INDUS is	18.1
Maximum of INDUS is	27.74
IQR of INDUS is	13.17



Calculate Five Point Summary

Macro values are the same as in the box and whisker plot.

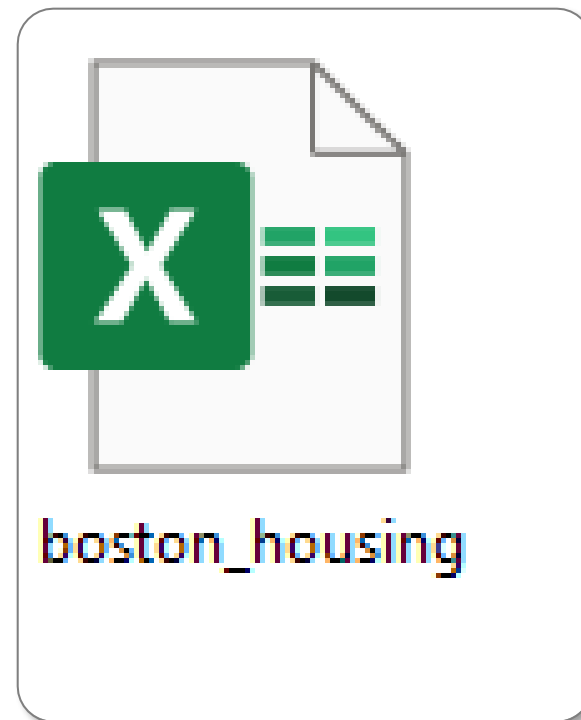


Minimum of INDUS is	27.74
Q1 of INDUS is	4.93
Q2 of INDUS is	8.14
Q3 of INDUS is	18.1
Minimum of INDUS is	27.74
IQR of INDUS is	13.17

Correlation Coefficient Using Macros

Correlation Coefficient Using Macros

Let us consider an example: For the Boston housing data, we can implement a macro to calculate the correlation coefficient between 'INDUS' and 'MEDV' using a macro.



Correlation Coefficient Using Macros

The mathematical equation for the correlation coefficient is:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$



Correlation Coefficient Using Macros

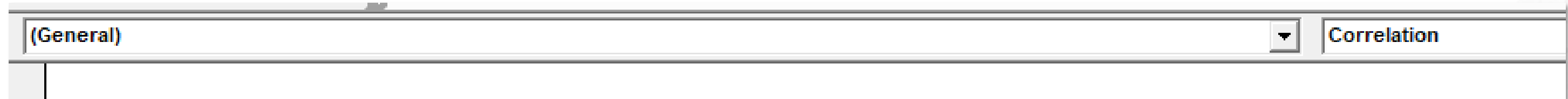
This is implemented as a macro function.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

When we verify the results with the CORREL function, both are the same.

Steps to Find Correlation Coefficient

The following are the steps to find the correlation coefficient.



Step 1: Open the macro editor using Alt+F11

Steps to Find Correlation Coefficient

Step 2: Copy the following code into the editor

```
(General) Correlation

Sub Correlation()
    n = 451

    Dim x(451)
    Dim y(451)

    For i = 1 To n
        x(i) = Cells(i + 1, 3)
        y(i) = Cells(i + 1, 14)
    Next i

    sum_x = 0
    sum_y = 0

    For i = 1 To n
        sum_x = sum_x + x(i)
        sum_y = sum_y + y(i)
    Next i

    mean_x = sum_x / 451
    mean_y = sum_y / 451

    num = 0
    s1 = 0
    s2 = 0

    For i = 1 To n
        num = num + ((x(i) - mean_x) * (y(i) - mean_y))
        s1 = s1 + (x(i) - mean_x) * (x(i) - mean_x)
        s2 = s2 + (y(i) - mean_y) * (y(i) - mean_y)
    Next i

    r = num / Sqr(s1 * s2)
    Cells(9, 16) = "Correlation coefficient of INDUS and MEDV is"
    Cells(9, 17) = r

    Cells(10, 16) = "Correlation coefficient using CORREL function"
    Cells(10, 17) = "=CORREL(C2:C451,N2:N451)"

End Sub
```



Steps to Find Correlation Coefficient

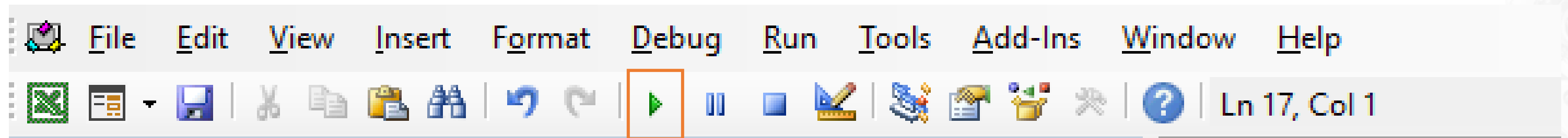
Step 3: The code calculates the correlation coefficient using the mathematical formula for 'INDUS' and 'MEDV' columns

Correlation coefficient of INDUS and MEDV is



Steps to Find Correlation Coefficient

Step 4: The results are stored in columns P and Q



Step 5: Run the macro using the F5 or run button

Steps to Find Correlation Coefficient

Step 6: The results are stored in the same Excel in columns P and Q

Correlation coefficient of INDUS and MEDV is	-0.41035
Correlation coefficient using CORREL function	-0.41035



We can see that the calculated correlation coefficient and CORREL functions are the same.

Steps to Find Correlation Coefficient

Step 7: Any formula can be assigned to a cell value using macros. This is done by the following command:

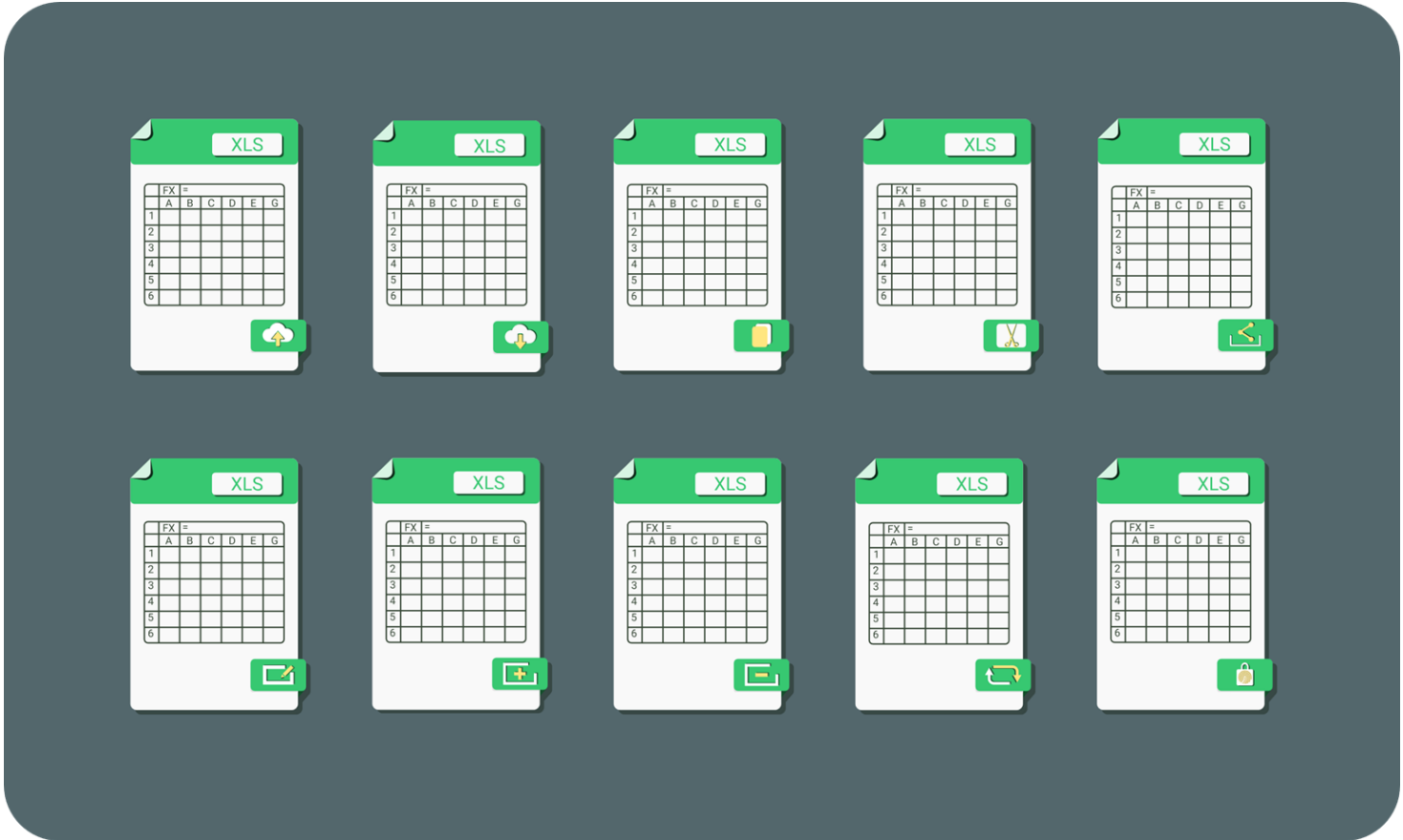
```
Cells(10, 17) = "=CORREL(C2:C451,N2:N451)"
```



Removing Duplicates Using Macros

Removing Duplicates

To remove duplicates using macros in a data set within a range, we can use the **RemoveDuplicates** function.



Removing Duplicates

Use this command to remove duplicates with **RemoveDuplicates** function.

```
Range("A2:A451").RemoveDuplicates Columns:=1, Header:=xlNo
```



Removing Duplicates

Columns:=1 specifies that the first column must be used for checking duplicates.

```
Range("A2:A451").RemoveDuplicates Columns:=1, Header:=xlNo
```

Header:=x | No specifies that there is no header in the data range.

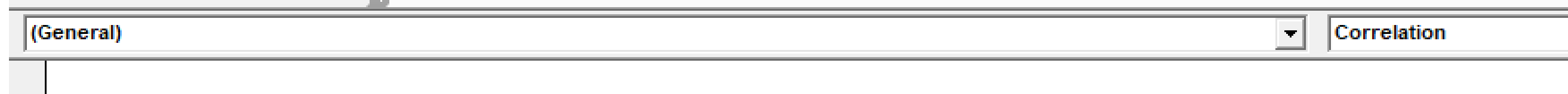
Removing Duplicates

We will try to remove duplicates from a range of rows and columns for our entire data set of Boston housing.

```
Range("A2:N451").RemoveDuplicates Header:=xlNo
```


Steps to Remove Duplicates

Open the Macro editor by pressing Alt+F11



Steps to Remove Duplicates

Create a function on columns A2:N451 to remove the duplicates

```
Sub remove_duplicates()  
    Range("A2:N451").RemoveDuplicates Header:=xlNo  
End Sub
```



Steps to Remove Duplicates

Run the function using the F5 or run button



Steps to Remove Duplicates

The macro automatically removes duplicates from column A.

Duplicates
1
2
3
5
4
8
6
7
24

Any duplicates in column A will be presented as shown.

Steps to Remove Duplicates

If we don't have duplicates in column A, nothing is removed.

435	0.15086
436	0.18337
437	0.20746
438	0.10574
439	0.11132
440	0.17331
441	0.27957
442	0.17899
443	0.2896
444	0.26838
445	0.23912
446	0.17783
447	0.22438
448	0.06263
449	0.04527
450	0.06076
451	0.10959



Key Takeaways

- Macros is an important feature of Excel which allows VBA programming within any Excel workbook.
- We can find the mean of a column of values using Macros for Excel.
- The five point summary in statistics specifies five values to describe a set of numeric values.
- The calculated correlation coefficient and CORREL functions are the same.
- To remove duplicates using macros in a data set within a range, we can use the **RemoveDuplicates** function.





Knowledge Check

Knowledge Check

1

Which of the following functionalities in Excel do Macros allow?

- A. Programming
- B. Formula writing
- C. Charting
- D. All the above



Knowledge Check

1

Which of the following functionalities in Excel do Macros allow?

- A. Programming
- B. Formula writing
- C. Charting
- D. All the above



The correct answer is **D**

All of the above.

Knowledge Check

2

Macros are based on which programming language?

- A. Visual Basic
- B. VC++
- C. VBA



Knowledge
Check

2

Macros are based on which programming language?

- A. Visual Basic
- B. VC++
- C. VBA



The correct answer is **C**

Visual Basic for Applications is used to program macros in Excel.

**Knowledge
Check**

3

VBA in Excel macros stands for?

- A. Visual Basic for Automation
- B. Visual Basic for Applications
- C. Visual Basic Application



**Knowledge
Check**

3

VBA in Excel macros stands for?

- A. Visual Basic for Automation
- B. Visual Basic for Applications
- C. Visual Basic Application



The correct answer is **B**

VBA stands for Visual Basic for Applications.

Knowledge Check

4

What is the mathematical formula for mean of a set of values?

- A. Sum of all values
- B. Sum of all values/number of values
- C. Sum of all values/number of values -1
- D. Number of values/Sum of values



Knowledge
Check

4

What is the mathematical formula for mean of a set of values?

- A. Sum of all values
- B. Sum of all values/number of values
- C. Sum of all values/number of values -1
- D. Number of values/Sum of values



The correct answer is **B**

Average is defined as sum of values/number of values.

**Knowledge
Check**

5

What is the macro way to set the value of C11 to mean of cells A1:A24?

- A. `Cells(11,3).values=="MEAN(A1:A24)"`
- B. `Cells(11,3)=="MEAN(A1:A24)"`
- C. `Cells(11,3)=="AVERAGE(A1:A24)"`
- D. `Cells(11,3).values=="AVERAGE(A1:A24)"`



Knowledge
Check

5

What is the macro way to set the value of C11 to mean of cells A1:A24?

- A. `Cells(11,3).values=="MEAN(A1:A24)"`
- B. `Cells(11,3)=="MEAN(A1:A24)"`
- C. `Cells(11,3)=="AVERAGE(A1:A24)"`
- D. `Cells(11,3).values=="AVERAGE(A1:A24)"`



The correct answer is **C**

AVERAGE function gives mean in Excel.

**Knowledge
Check**

6

For finding median of a dataset using macros, the data is ordered in ascending and middle value is found programmatically? True or False.

- A. True
- B. False



**Knowledge
Check**

6

For finding median of a dataset using macros, the data is ordered in ascending and middle value is found programmatically? True or False.

- A. True
- B. False



The correct answer is **A**

True. By definition of median, the data is ordered in ascending order and the middle value(s) is/are the median(s).

**Knowledge
Check**

7

Which of the following defines Interquartile range?

- A. Q1-Q2
- B. Q2-Q3
- C. Q3-Q1



Knowledge
Check

7

Which of the following defines Interquartile range?

- A. Q1-Q2
- B. Q2-Q3
- C. Q3-Q1



The correct answer is **C**

IQR is 3rd quartile minus 1st quartile.

**Knowledge
Check**

8

Which of the following is NOT a part of the 5-point summary?

- A. Mean
- B. Median
- C. Maximum
- D. Minimum



Knowledge
Check

8

Which of the following is NOT a part of the 5-point summary?

- A. Mean
- B. Median
- C. Maximum
- D. Minimum



The correct answer is **A**

Mean is not a part of the 5-point summary.

Knowledge Check

9

Which plot is based on the 5-point summary?

- A. Bar chart
- B. Box-and-whisker
- C. Line graph
- D. Histogram



Knowledge Check

9

Which plot is based on the 5-point summary?

- A. Bar chart
- B. Box-and-whisker
- C. Line graph
- D. Histogram



The correct answer is **B**

Box-and-Whisker plot shows the 5-point summary

**Knowledge
Check**

10

What is the maximum value of the correlation coefficient?

- A. 0
- B. 1
- C. 2



Knowledge
Check

10

What is the maximum value of the correlation coefficient?

- A. 0
- B. 1
- C. 2



The correct answer is **B**

Maximum value of correlation coefficient is 1.

**Knowledge
Check**

11

If two variables are non-correlated, the value of the correlation coefficient is around which value?

- A. -1
- B. 0
- C. 1



**Knowledge
Check**

11

If two variables are non-correlated, the value of the correlation coefficient is around which value?

- A. -1
- B. 0
- C. 1



The correct answer is **B**

If two variables are non-correlated, the value of the correlation coefficient is 0.

**Knowledge
Check**

12

What is typically done with highly correlated variables for data analytics?

- A. They are retained for the model
- B. They are removed for the model



Knowledge
Check

12

What is typically done with highly correlated variables for data analytics?

- A. They are retained for the model
- B. They are removed for the model



The correct answer is **B**

Highly correlated variables are removed and only one of them is retained for the model.

**Knowledge
Check**

13

What is the range of Correlation values?

- A. 0 and 1
- B. 1 and 2
- C. -1 and +1
- D. -1 and 0



Knowledge
Check

13

What is the range of Correlation values?

- A. 0 and 1
- B. 1 and 2
- C. -1 and +1
- D. -1 and 0



The correct answer is **C**

Correlation coefficient is between -1 and +1

**Knowledge
Check**

14

Duplicates cannot be removed using macros. True or False.

- A. True
- B. False



Knowledge
Check

14

Duplicates cannot be removed using macros. True or False.

- A. True
- B. False



The correct answer is **B**

False. Duplicates can be removed using macros.

**Knowledge
Check**

15

How is the function for RemoveDuplicates used to specify that there is no header in the data?

- A. Header:=xlNo
- B. Header:=No
- C. Header:=FALSE
- D. Header:=None



**Knowledge
Check**

15

How is the function for RemoveDuplicates used to specify that there is no header in the data?

- A. Header:=xlNo
- B. Header:=No
- C. Header:=FALSE
- D. Header:=None



The correct answer is **A**

Header:=xlNo is used to specify that the dataset has no headers.

**Knowledge
Check**

16

Columns keyword is used to specify the column number to check for duplicates in RemoveDuplicates macro function. True or False.

- A. True
- B. False



**Knowledge
Check**

16

Columns keyword is used to specify the column number to check for duplicates in RemoveDuplicates macro function. True or False.

- A. True
- B. False



The correct answer is **A**

True. Columns:=1 is specified to remove duplicates using column 1.